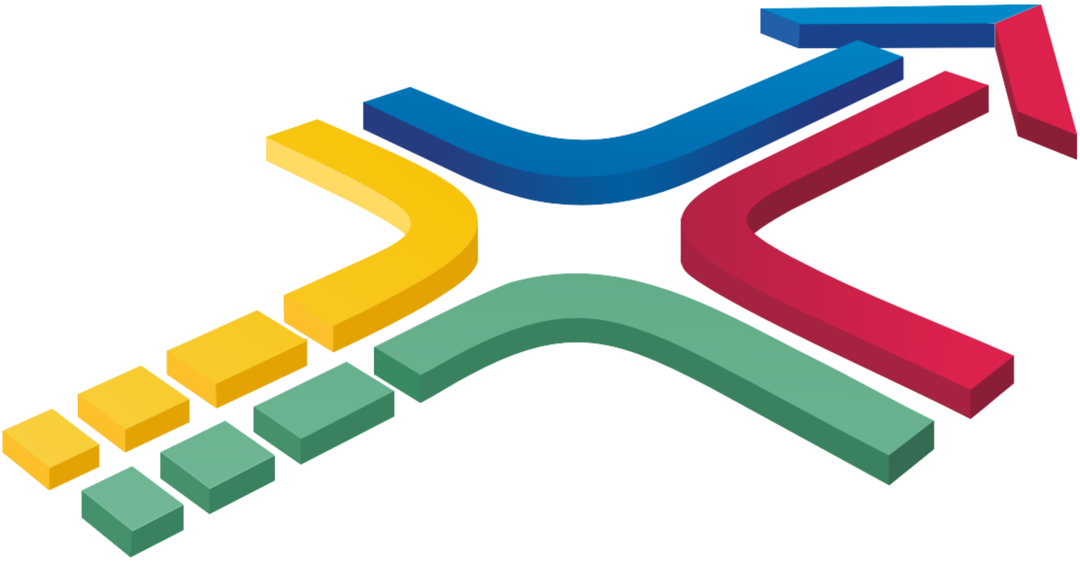
The geneXplain platform



User guide

4.0

July 2016

©2011-2016, geneXplain GmbH, Wolfenbüttel, Germany

geneXplain GmbH
Am Exer 10b
D-38302 Wolfenbüttel
Germany

E-mail: info@genexplain.com

URL: http://www.genexplain.com

Contents

1. Introduction

The geneXplain platform is an online workbench to assist in operating the daily computer applications in life sciences. It comprises a number of bioinformatics and systems biological modules, or BRICKS, which are unified under a standardized interface, with a consistent look-and-feel. These Bricks can be put together to comprehensive workflows using a workflow management system, which is intuitively handled through a simple drag-and-drop system. With this systems, the user can edit the predefined as well as compose own workflows. Own Bricks can easily be added as JavaScript or R scripts and incorporated into workflows as well.

The whole system aims at covering, with time, all areas of computational applications. The community is invited to contribute Bricks, either as public-domain or as a commercial part of the platform. We are confident that this way, an extremely powerful system will grow in which a user-driven selection will ensure the best tools being the most successful ones.

Besides providing a wide range of sophisticated Bricks (more than 60), the geneXplain platform also facilitates standard analyses through a number (61) of pre-composed workflows presently subsumed under 8 categories.

The technology behind the geneXplain platform is BioUML, which has been developed for many years at the Institute of Systems Biology in Novosibirsk. The architecture is open, so that own scripts can be easily loaded into the system, and new Bricks can be programmed and added by each skilled person.

In the following, we will guide you first through the pre-defined workflows, before we are going to introduce you to the different parts of the system in a more systematic way.

Have fun with the system, and don’t hesitate to come back to us with any suggestion for improvements (info@genexplain.com), we know that there is still ample space to make the system better.

2. Organization of the geneXplain platform

2.1. The user interface

When you open the geneXplain platform for the first time, a window opens that contains the following areas:

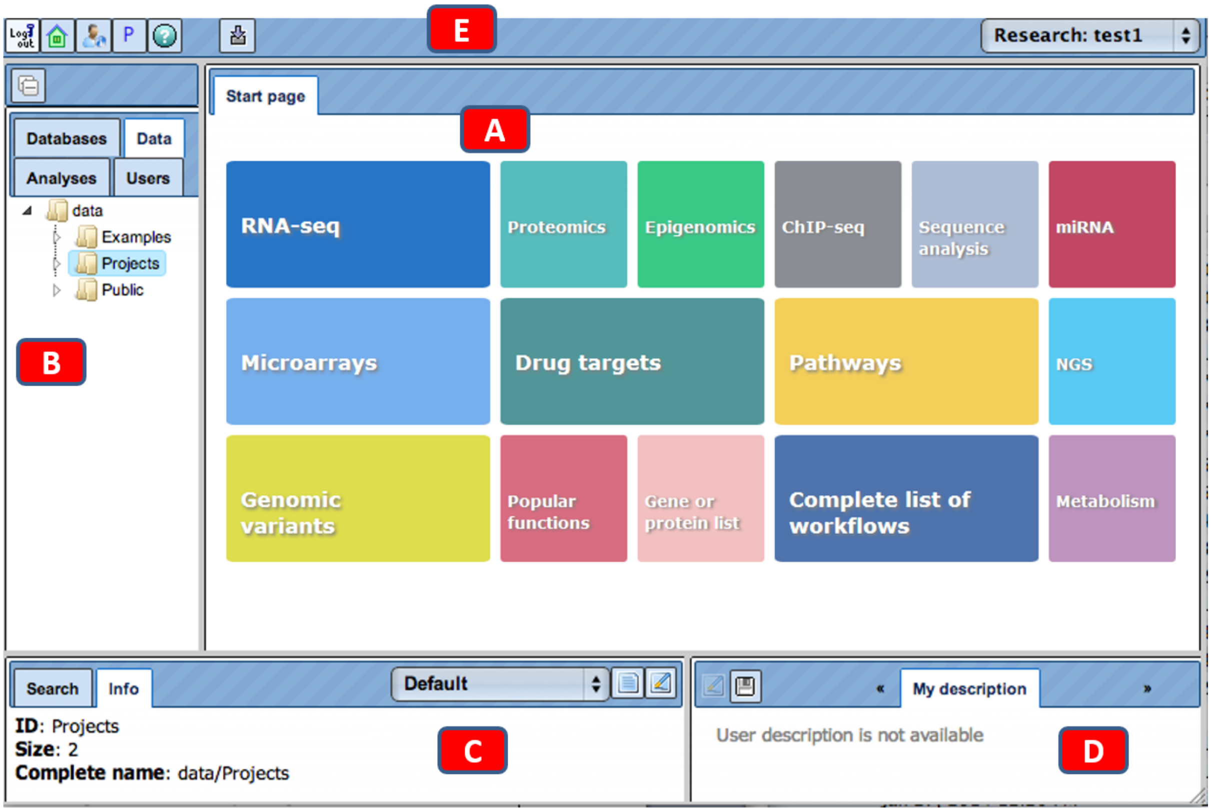
A - The **Work Space**, which is the main part of the window.

B - The **Tree Area** (to the left of the Work Space), where you find the collection of Databases, the uploaded Data files and the available Analyses methods under the corresponding tabs, organized in a hierarchical tree structure.

C - The **Info Box** (in the left lower part), where you can select the data resource to Search in, or where you will get Information about the data file or analysis method that you select with a single click in the Tree Area.

D - The **Operations Field** (right lower part), providing a number of options under the different tabs in a context-dependent manner.

E - The general **Control Panel**, on top of the different areas, showing a context-dependent set of icons for the available operations.



ID: Projects
Size: 2
Complete name: data/Projects

User description is not available

2.1.1. Work Space

When you open the platform, the Work Space will show you all those research areas that are supported by a number of bioinformatic workflows each. Clicking on one of these tiled buttons will open a detailed listing of functions and pre-composed workflows, which you can launch by directly activating the given hyperlinks.

The listings of each Area introduce each Chapter of this User Guide that explains the functionality provided:

Area	**Chapter**	**Area**	**Chapter**
RNA-seq	4	Pathways	12
Proteomics	5	NGS	13
Epigenomics	6	Genomic variants	14
ChIP-seq	7	Metabolism	15
Sequence analysis	8	Popular functions	16
miRNA	9	Gene or protein list	17
Microarrays	10	Complete list of workflows	18
Drug targets	11	Working with different databases	19

The first workflow in nearly any Area is Load data, which is separately described in Chapter 3, before the individual Areas are being depicted.

Also in the Work Space, all major input requirements will have to be done, the progress of a workflow and the results output will appear here as well.

2.1.2. Tree Area

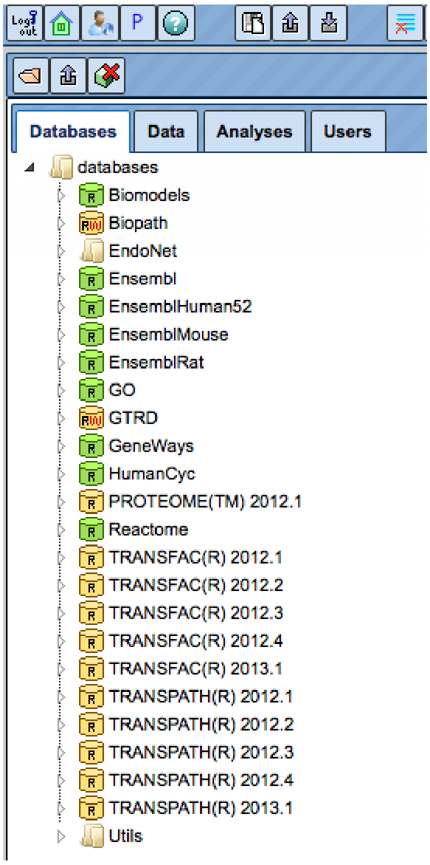
In the Tree Area (**B** in the figure of Section 2.1), you find three tabs: **Databases**, **Data**, and **Analyses**. Under each of these tabs, you find a tree-structured listing of contents: the available databases, the data files to work with, and the analysis methods you can apply. You can open any subdirectory by a single mouse click on the respective item and using the button ⊟, or by a double-click on the name of the folder to be opened.

2.1.2.1. Databases

Under this tab, you find all databases that are at your disposal: a number of public databases (yellow symbols) as well as commercial ones that you have a valid license for (green symbols). Write access may be additionally indicated by a black "W" within the symbol, which may be red in case that write access, although principally possible for this database, has not been activated in your case.

Select one of the databases in the Tree Area by a single click with the left mouse button; double click will open the subdirectory. For instance, when you wish to go to GeneWays, you have to click once on the little white triangle next to the term GeneWays, or double-click on the term itself to open the directory underneath; either operation, or already a single click onto the name of the database, suffice to indicate this data resource in the Info Box, along with a short description.

Presently, by default, you should see the following listing of available databases:



databases
- Biomodels
- Biopath
- EndoNet
- Ensembl
- EnsemblHuman52
- EnsemblMouse
- EnsemblRat
- GO
- GTRD
- GeneWays
- HumanCyc
- PROTEOME(TM) 2012.1
- Reactome
- TRANSFAC(R) 2012.1
- TRANSFAC(R) 2012.2
- TRANSFAC(R) 2012.3
- TRANSFAC(R) 2012.4
- TRANSFAC(R) 2013.1
- TRANSPATH(R) 2012.1
- TRANSPATH(R) 2012.2
- TRANSPATH(R) 2012.3
- TRANSPATH(R) 2012.4
- TRANSPATH(R) 2013.1
- Utils

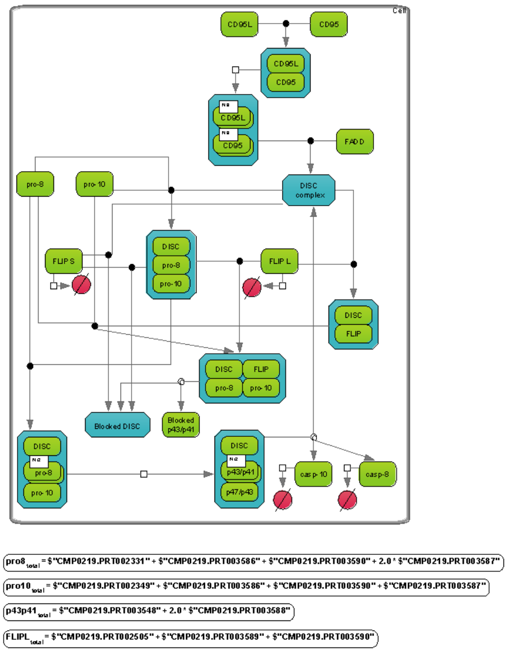
How you can work with the different databases will be explained in Chapter 19.

2.1.2.2. Data

On the tab Data, you will find your own datasets as well as predefined ones in the directories **Examples**, **Projects**, and **Public**.

In the directory **Examples**, we provide you with a number of sample data you may use for getting familiar with the system. Presently, these are some optimization examples as well as one application example on expression data of Psoriasis patient samples.

The optimization example refers to the CD95L pathway (*Reactome entry REACT_900*), shown as diagram *CD95 signaling pathway* under Diagrams. Under Plots, you find an example for a successful parameter optimization (*C3C8_plot*), with the dots giving the experimental values for four different molecules, and the straight lines showing the simulation results.

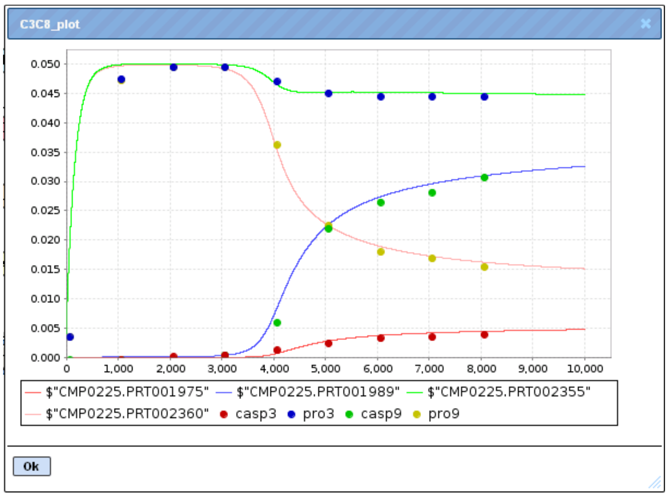


$\mathrm{pro8}_{total} = \$\text{"CMP0219.PRT002331"} + \$\text{"CMP0219.PRT003586"} + \$\text{"CMP0219.PRT003590"} + 2.0 * \$\text{"CMP0219.PRT003587"}$

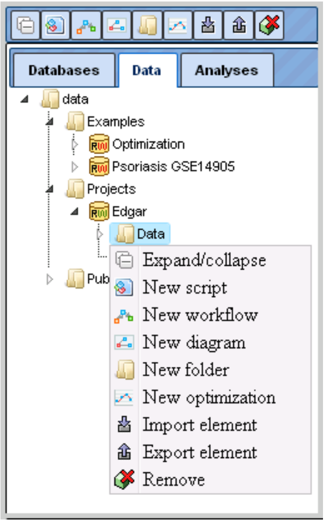
$\mathrm{pro10}_{total} = \$\text{"CMP0219.PRT002349"} + \$\text{"CMP0219.PRT003586"} + \$\text{"CMP0219.PRT003590"} + \$\text{"CMP0219.PRT003587"}$

$\mathrm{p43p41}_{total} = \$\text{"CMP0219.PRT003548"} + 2.0 * \$\text{"CMP0219.PRT003588"}$

$\mathrm{FLIPL}_{total} = \$\text{"CMP0219.PRT002505"} + \$\text{"CMP0219.PRT003589"} + \$\text{"CMP0219.PRT003590"}$



Under the heading **Projects**, you should initially find only one directory with your username, with the two subfolders **Data** and **Journal**. **Data** will be the place where all your own data will be deposited. Here, you can define own subdirectories by opening a context menu with the right mouse button and selecting the option New folder:



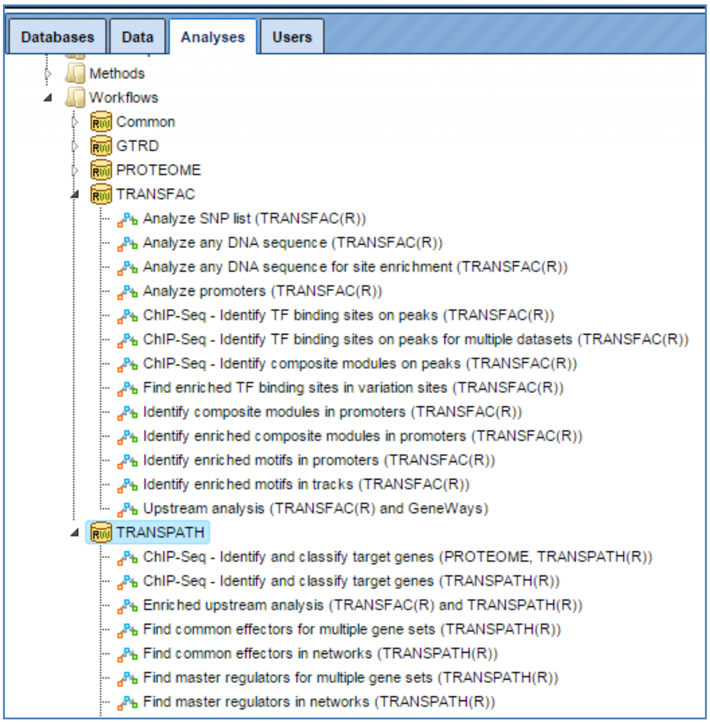
Each option in this menu can also be selected by the corresponding icon on top of the tabs.

The sub-directory **Journal** stores a history of your activities. Opening it in the Work Space shows you a list of all methods that you have previously launched, including time stamp.

The root directory **Public** is presently not used yet; it is planned to provide a platform for sharing own results with a wider community at a later stage.

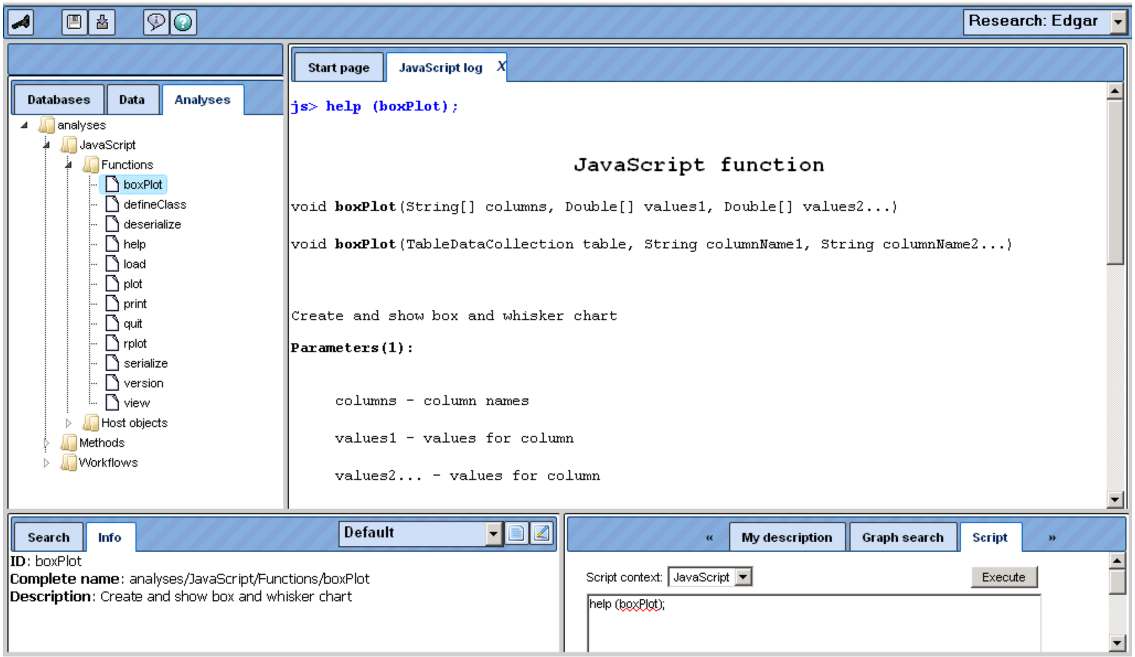
2.1.2.3. Analyses

On this tab, you find all methods you may apply using the geneXplain platform, including the pre-defined workflows:



- Methods
- Workflows
  - Common
  - GTRD
  - PROTEOME
  - TRANSFAC
    - Analyze SNP list (TRANSFAC(R))
    - Analyze any DNA sequence (TRANSFAC(R))
    - Analyze any DNA sequence for site enrichment (TRANSFAC(R))
    - Analyze promoters (TRANSFAC(R))
    - ChIP-Seq - Identify TF binding sites on peaks (TRANSFAC(R))
    - ChIP-Seq - Identify TF binding sites on peaks for multiple datasets (TRANSFAC(R))
    - ChIP-Seq - Identify composite modules on peaks (TRANSFAC(R))
    - Find enriched TF binding sites in variation sites (TRANSFAC(R))
    - Identify composite modules in promoters (TRANSFAC(R))
    - Identify enriched composite modules in promoters (TRANSFAC(R))
    - Identify enriched motifs in promoters (TRANSFAC(R))
    - Identify enriched motifs in tracks (TRANSFAC(R))
    - Upstream analysis (TRANSFAC(R) and GeneWays)
  - TRANSPATH
    - ChIP-Seq - Identify and classify target genes (PROTEOME, TRANSPATH(R))
    - ChIP-Seq - Identify and classify target genes (TRANSPATH(R))
    - Enriched upstream analysis (TRANSFAC(R) and TRANSPATH(R))
    - Find common effectors for multiple gene sets (TRANSPATH(R))
    - Find common effectors in networks (TRANSPATH(R))
    - Find master regulators for multiple gene sets (TRANSPATH(R))
    - Find master regulators in networks (TRANSPATH(R))

In the directory **JavaScript**, a collection of scripts have been gathered that all can be called by your own scripts. The geneXplain platform provides you with the possibility to create your own scripts on the Script tab in the Operations Field. For instance, enter there the command "`help (boxPlot);`" (boxPlot is the first JavaScript in the subdirectory **Functions**) and press Execute (or just Enter): Some help text about boxPlot will be printed to the Work Space:



```
js> help (boxPlot);

                    JavaScript function

void boxPlot(String[] columns, Double[] values1, Double[] values2...)

void boxPlot(TableDataCollection table, String columnName1, String columnName2...)


Create and show box and whisker chart

Parameters(1):

     columns - column names

     values1 - values for column

     values2... - values for column
```

ID: boxPlot
Complete name: analyses/JavaScript/Functions/boxPlot
Description: Create and show box and whisker chart

```
help (boxPlot);
```

Under **Functions**, a number of standard JavaScript functions are listed, whereas in the subdirectory **Host object**, a collection of scripts have been put together that were specifically developed for the gXp platform.

The directory **Methods** contains all individual tools, or Bricks, that perform specific tasks in one of the listed areas. A full listing of all analysis methods that are presently available is given in the Help texts. Each method can be launched either by double-clicking on the respective item, which will open the corresponding input mask in the Work Space, or through the context menu that you can open with a right-button mouse click on the method of interest.

In the directory **Workflows**, you find basically the same functional groups of pre-defined workflows that are also listed on the Start page of the platform (Work Space). Please note that the first two items on the Start page are important preparatory tasks, but not workflows according to the definition of the geneXplain platform. Please, refer to the following Chapters for operating the pre-defined workflows and to Chapter 22 for creating your own workflows.

2.1.3. The Info Box

The Info Box shows two tabs, **Info** and **Search**.

2.1.3.1. Info tab

Under the Info tab, you will find short descriptions about the database, data file or analysis method that you have clicked in the Tree Area. In many cases, you will just see the name of the directory or file as **ID**, or the complete path displayed in the field **Complete name**, and the number of subdirectories and files right under the activated node as **Size**.

The information shown on the Info tab can be viewed in a separate window by clicking the button 📄.

In case the activated file is a table (icon ☰ or derived images; see 25.2.3), the number of rows in the table will be indicated as Size. Optionally, a **Description** is shown as well, depending on whether it has been entered when the file was generated. By clicking on the icon ✎, you may edit some of the information in a newly opened window. Depending on the file type, the edit window may look like this:

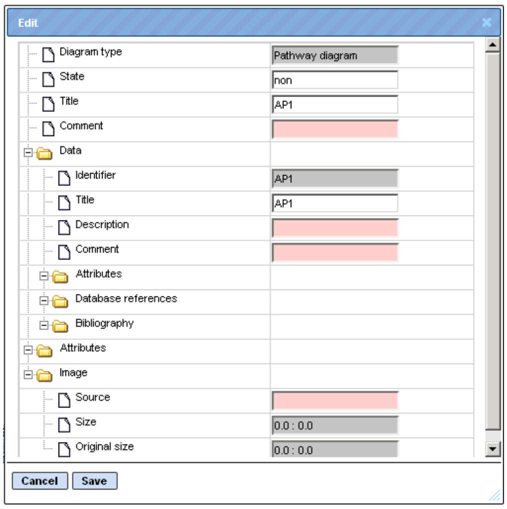
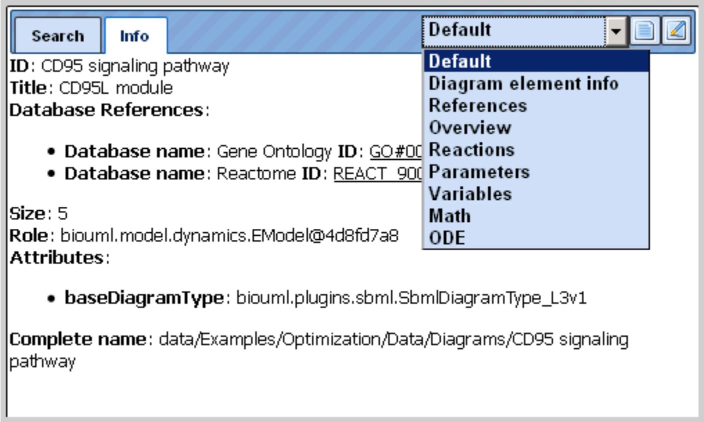


Diagram type	Pathway diagram
State	non
Title	AP1
Comment	
Data	
Identifier	AP1
Title	AP1
Description	
Comment	
Attributes	
Database references	
Bibliography	
Attributes	
Image	
Source	
Size	0.0 : 0.0
Original size	0.0 : 0.0

Gray fields cannot be edited, contents of fields with white background can be changed manually, additional information can be entered in the pink fields which will then appear on the Info tab, if the changes made are saved afterwards with the [Save] button of the Edit window.

Depending on the type of directory or file that you have activated in the Tree Area, different types of information can be selected for the display on the Info tab. This can be done by the selection box right to the tabs in the Info Box, initially always showing "Default":



ID: CD95 signaling pathway
Title: CD95L module
Database References:

- **Database name**: Gene Ontology **ID**: GO#00
- **Database name**: Reactome **ID**: REACT_900

Size: 5
Role: biouml.model.dynamics.EModel@4d8fd7a8
Attributes:

- **baseDiagramType**: biouml.plugins.sbml.SbmlDiagramType_L3v1

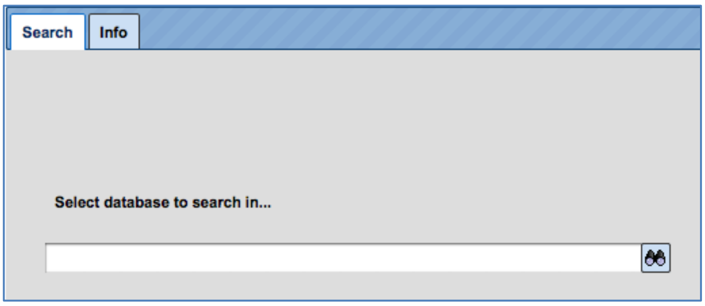
Complete name: data/Examples/Optimization/Data/Diagrams/CD95 signaling pathway

Please, feel encouraged to find out the effects of the different views offered here.

When you have a network diagram opened in the Work Space (**A** in the figure of chapter 2) and you select a molecule or reaction with one click, you find information about the selected element in the Info box.

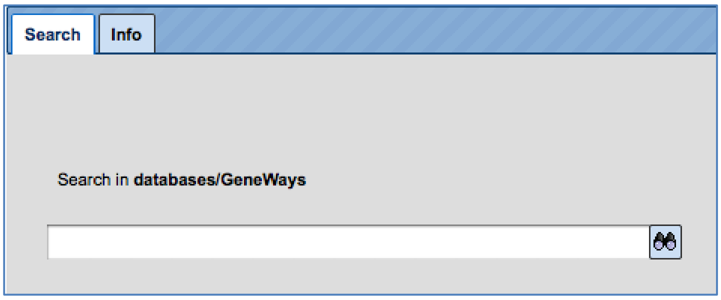
2.1.3.2. Search

As long as you are under the tabs Data or Analyses in the Tree Area, you see here only the message "Select database to search in …, maybe along with fields and data from an earlier search.



Select database to search in...

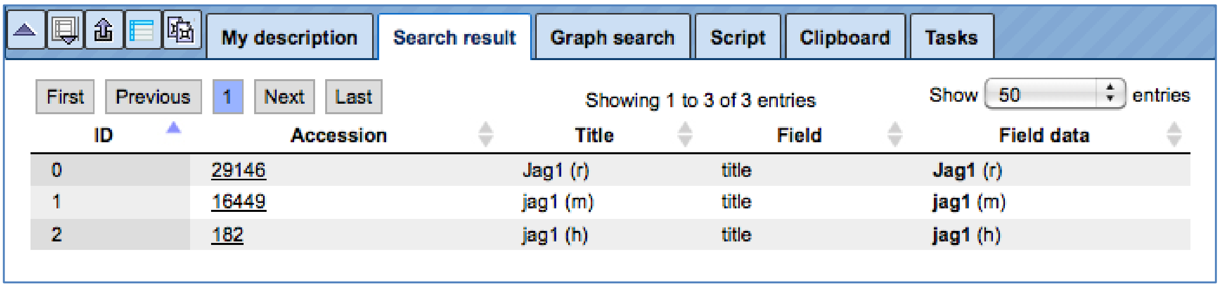
This field gets activated as soon as you go to the tab Databases in the Tree Area, and then click on one of the databases listed there. For instance, when you wish to go to GeneWays, you have to click once on the little white triangle next to the term GeneWays, or double-click on the term itself to open the directory underneath; either operation, or already a single click onto the name of the database, suffices to indicate this data resource in the Info Box:



Search in **databases/GeneWays**

You can insert your search term (e.g., a gene symbol) into the field underneath. Clicking on the icon launches the search. The search routine scans for exact matches, but use of wildcards is possible. Thus, searching for elk* returns results for elk1, elk2p1, elk3, and elk4.

The results will be shown in the Operations Field, under the tab Search result. For instance, when searching in GeneWays for JAG1, the following result table will be displayed:



ID	Accession	Title	Field	Field data
0	29146	Jag1 (r)	title	**Jag1** (r)
1	16449	jag1 (m)	title	**jag1** (m)
2	182	jag1 (h)	title	**jag1** (h)

The search term is highlighted in bold.

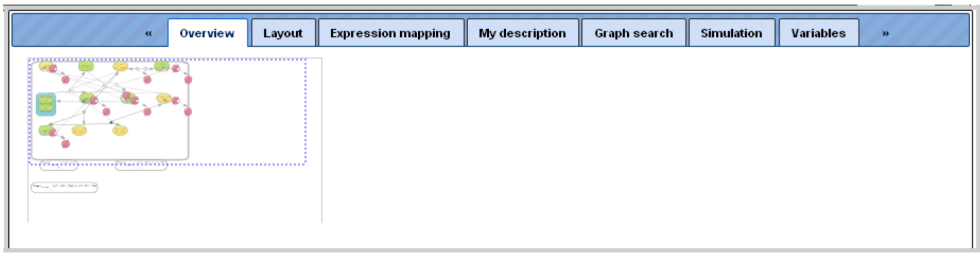
In some cases, the search results can be narrowed down by further specifying the search space in the Info Box.

More about how to operate the individual databases will be explained in Chapter 19.

2.1.4. The Operations Field

In the **Operations Field** (**D** in the figure of the Section introduction 2.1) a number of essential functions to operate the geneXplain platform are provided on a number of tabs. How many and which tabs are shown depends very much on the context.

Please note that not all tabs are always visible due to space constraints. In these cases, double arrowheads left and right of the tabs indicate that there are additional ones, reachable by clicking on these double arrowheads.



The function of the individual tabs will be explained in more detail in those sections where their effect is part of a certain operation. In general, the icon ▷ initiates the corresponding activity within the Operations Field, whereas △ applies to the results generated in the Operations Field of the Work Space.

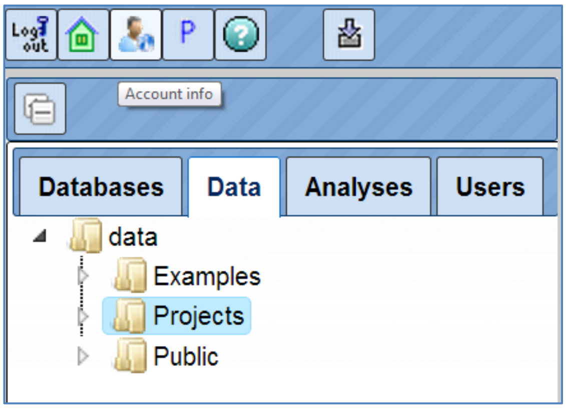
The full range of functions that you can make use of in the Operations Field is explained in greater detail in Chapter 21.

2.2. How to organize the user work space

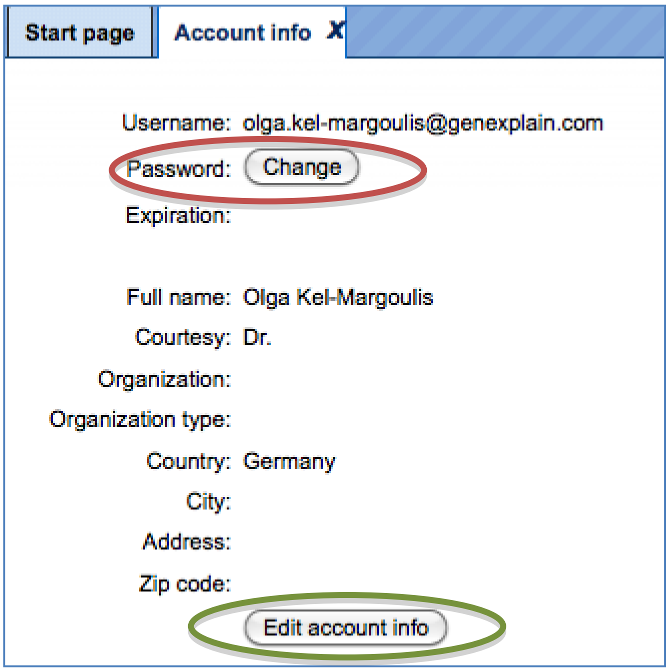
2.2.1. Changing user password and personal data

Your first password for the geneXplain platform is automatically generated and sent to you by e-mail. Once you enter the platform you can change your password and also have an option to edit your personal information.

To change password and edit the personal data, select the *Account info* button (👤) on the top menu control panel:



Your account information will be displayed in the work space as shown below:

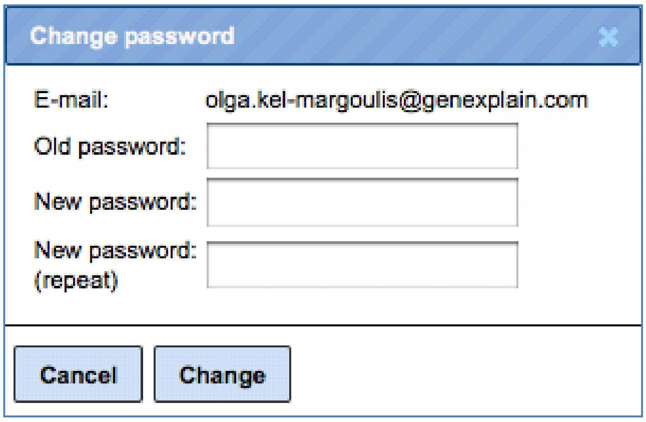


Username: olga.kel-margoulis@genexplain.com
Password: Change
Expiration:

Full name: Olga Kel-Margoulis
Courtesy: Dr.
Organization:
Organization type:
Country: Germany
City:
Address:
Zip code:

Edit account info

The option to change the password is highlighted by the red oval. Once you press [Change], you get an option to enter the new password:



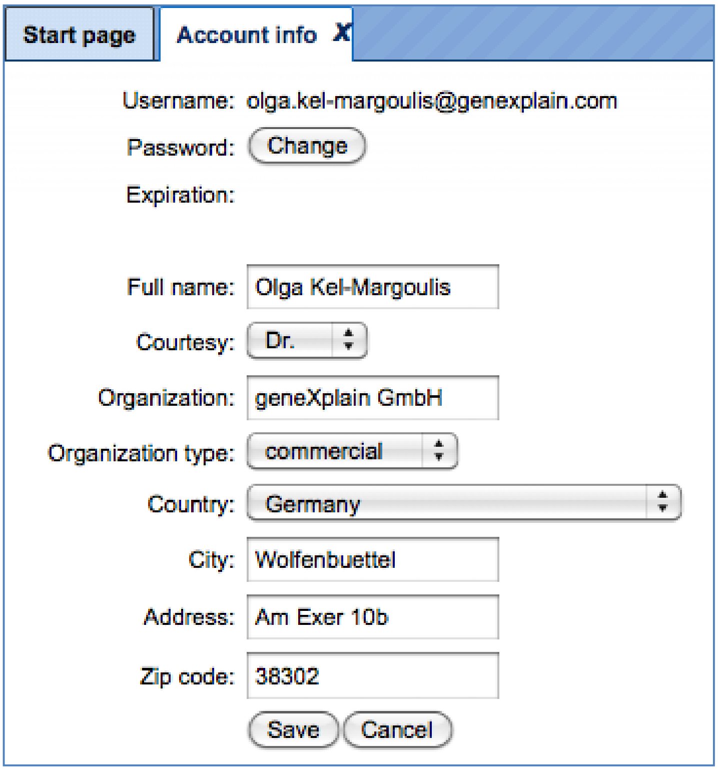
E-mail: olga.kel-margoulis@genexplain.com

Old password:

New password:

New password: (repeat)

To change the account details, press the button [Edit account info] in the form above, highlighted by the green oval. The edit form looks like this:



Username: olga.kel-margoulis@genexplain.com

Password:

Expiration:

Full name: Olga Kel-Margoulis

Courtesy: Dr.

Organization: geneXplain GmbH

Organization type: commercial

Country: Germany

City: Wolfenbuettel

Address: Am Exer 10b

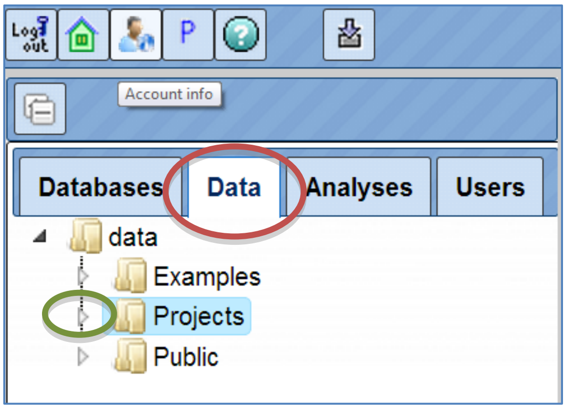
Zip code: 38302

Fill-in your details and press [Save]. Before saving, the system verifies your password to enable the changes made in the form. After entering the password the changes are saved.

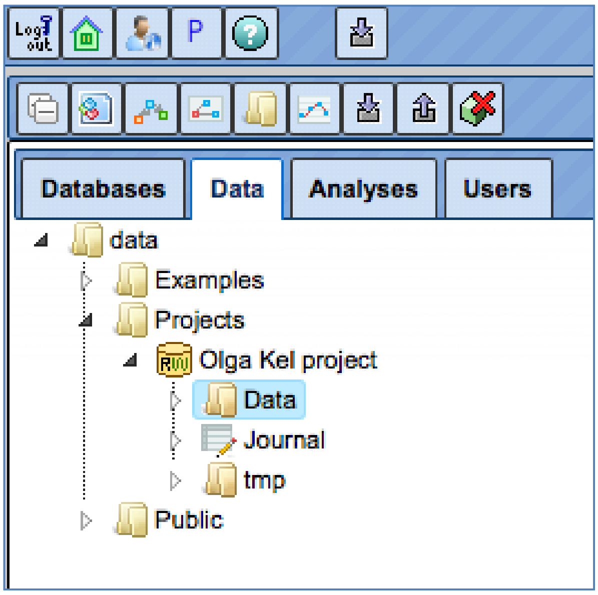
2.2.2. User Project, Data folder, creating new folders

When you enter your account for the first time, you can see the following three folders in the Tree Area under the *Data* tab (red oval on the picture below): Examples, Projects and Public. Each of these folders can be expanded by clicking on the small triangle on the left side of the folder name (green oval on the picture below).

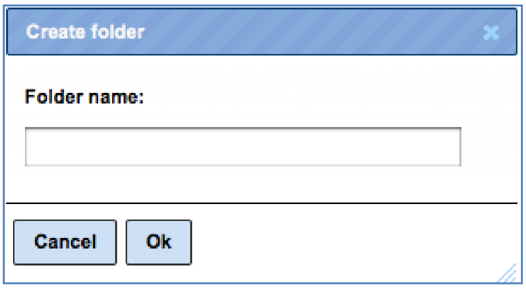
The folder **Projects** is your folder in the tree where you are going to make all the analyses.



If you expand the Project folder, you can see the project that was created upon registration of your account (R/W), and if you expand it further, you can see the folder Data. This location, *data/Projects/User project/Data/* is exactly the location where you can upload your data, and save all the analysis results.

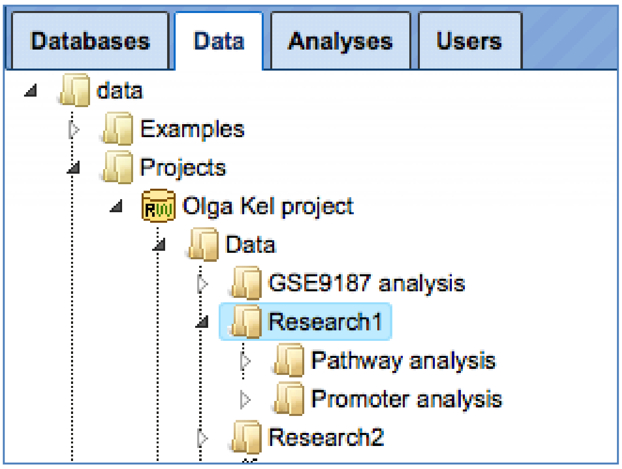


Upon one mouse click on the folder Data, as in the picture above, you can apply the button from the top control menu to create a new folder within the selected one.



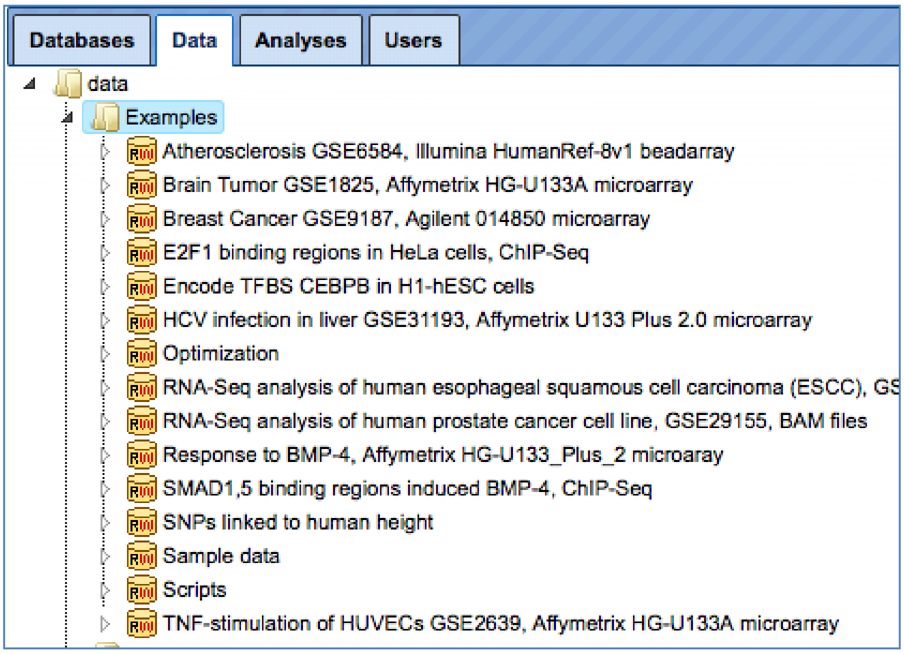
In this way, you can define the hierarchical organization of your folders and sub-folders within your project, for example as shown below.

Every time when you run a workflow, you need to specify a location of the results folder, and you can specify any particular location within your project area.



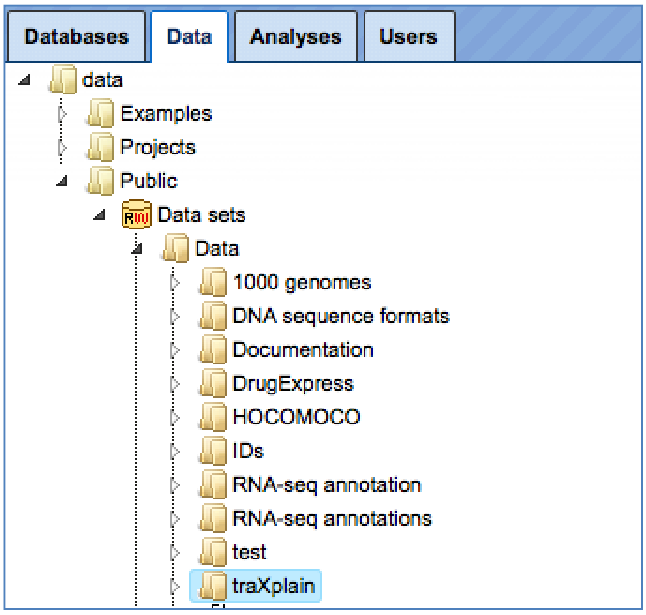
- data
  - Examples
  - Projects
    - Olga Kel project
      - Data
        - GSE9187 analysis
        - Research1
          - Pathway analysis
          - Promoter analysis
        - Research2

There are two other folders available for you initially, Examples and Public. The folder **Examples** contains pre-analyzed publicly available data sets, which you are welcome to have a look through our examples. You can copy tables or tracks from these two folders into your project area and use them, for example, for test runs.



- data
  - Examples
    - Atherosclerosis GSE6584, Illumina HumanRef-8v1 beadarray
    - Brain Tumor GSE1825, Affymetrix HG-U133A microarray
    - Breast Cancer GSE9187, Agilent 014850 microarray
    - E2F1 binding regions in HeLa cells, ChIP-Seq
    - Encode TFBS CEBPB in H1-hESC cells
    - HCV infection in liver GSE31193, Affymetrix U133 Plus 2.0 microarray
    - Optimization
    - RNA-Seq analysis of human esophageal squamous cell carcinoma (ESCC), GS
    - RNA-Seq analysis of human prostate cancer cell line, GSE29155, BAM files
    - Response to BMP-4, Affymetrix HG-U133_Plus_2 microaray
    - SMAD1,5 binding regions induced BMP-4, ChIP-Seq
    - SNPs linked to human height
    - Sample data
    - Scripts
    - TNF-stimulation of HUVECs GSE2639, Affymetrix HG-U133A microarray

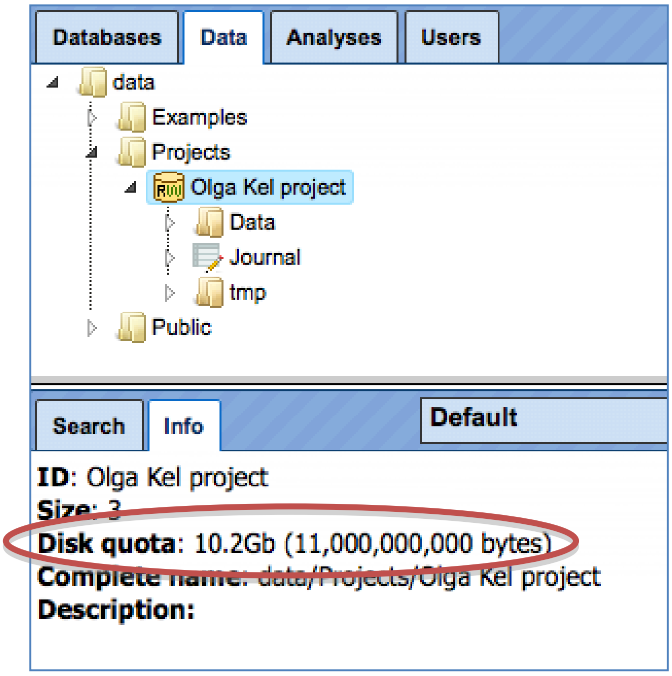
The Public folder contains publicly available data sets which might be useful to apply for various analysis purposes.



All users have read access to the folders Examples and Public, but no write access. You cannot save any files directly in the Examples or Public folder. However, you can copy tables and tracks from these two folders into your project area and then modify and work with them as you like.

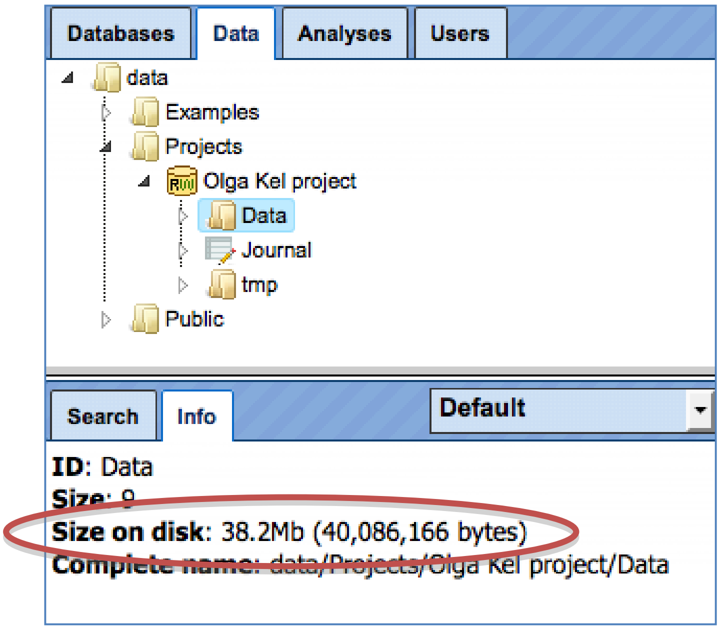
2.2.3. How to check information about the available work space

To check your totally available work space, go with mouse over to the project name, so that it is highlighted in blue as shown below. In the Info Box you can see information about this project including disk quota. This is the space available for you. If you plan to upload large files, please make sure you have enough work space available.



ID: Olga Kel project
Size: 3
Disk quota: 10.2Gb (11,000,000,000 bytes)
Complete name: data/Projects/Olga Kel project
Description:

To check how much space out of your quota is already occupied, go with mouse over to the *Data* folder within your project, so that it is highlighted in blue as shown below. In the Info Box you can see information about this folder including its size on the disk.



ID: Data
Size: 9
Size on disk: 38.2Mb (40,086,166 bytes)
Complete name: data/Projects/Olga Kel project/Data

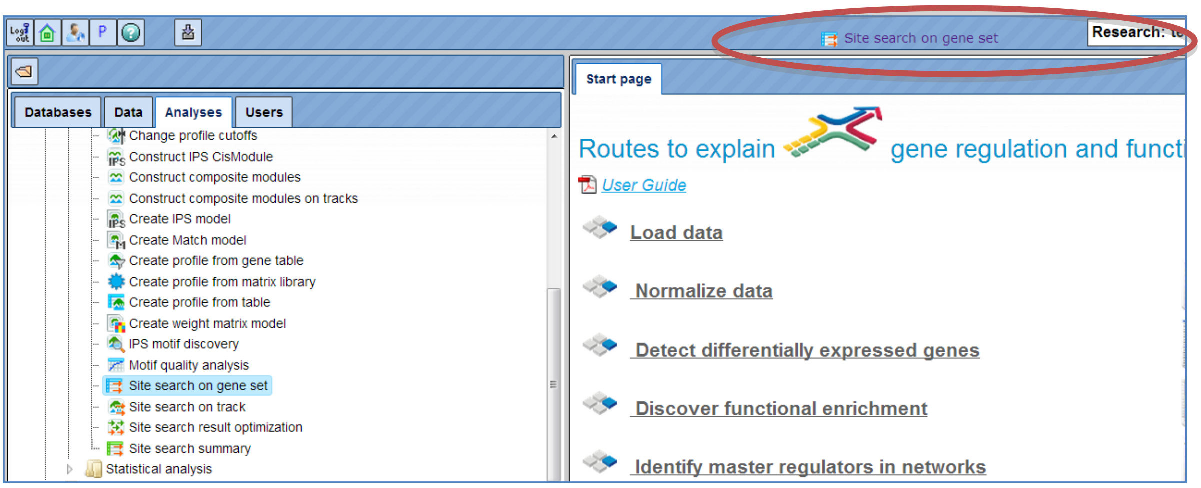
In the same way, you can check the size for every individual folder in your project.

Note. If additional work space or storage space is required, especially if you plan to upload and analyze large data files, please feel free to ask for details (info@genexplain.com).

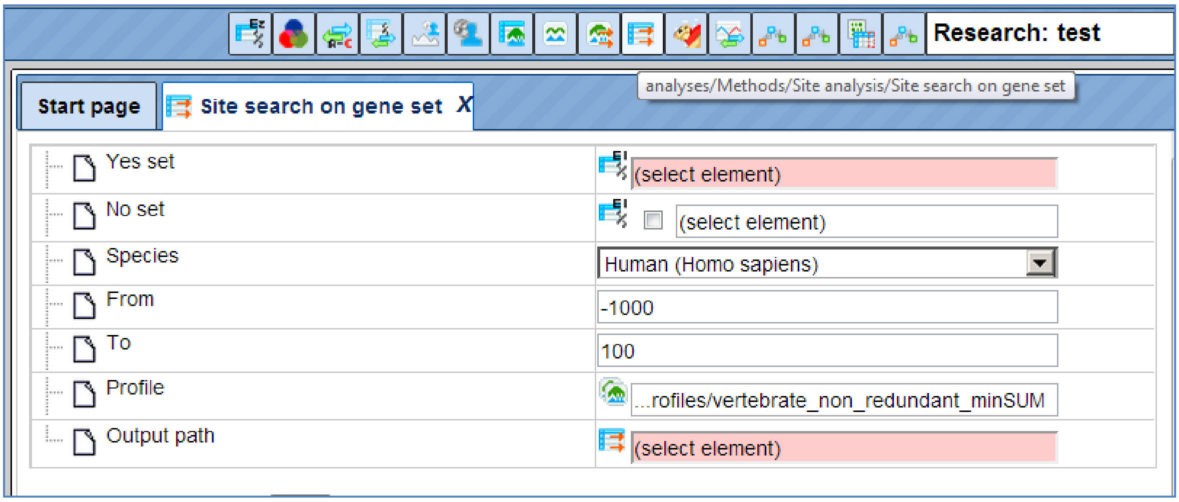
2.2.4. User toolbar

In the Control Panel (see **E** in the figure of Section 2.1), the set of icons on the left side is fixed by default, whereas the right side is customizable. Here, you can create a user-specific toolbar with your most frequently used analysis methods, workflows and datasets. To create your own toolbar you can drag and drop your favorite workflows and files for which you often need a quick access into the Control Panel. The User tool bar will then be located at the top right top corner, highlighted by a red oval in the screenshot below.

To add any analysis method/workflow/gene set, open the respective method in the Tree Area or in the Work Space and drag & drop it onto the user toolbar as shown below.


Site search on gene set
Research: te
Start page
Databases
Data
Analyses
Users
Change profile cutoffs
Construct IPS CisModule
Construct composite modules
Construct composite modules on tracks
Create IPS model
Create Match model
Create profile from gene table
Create profile from matrix library
Create profile from table
Create weight matrix model
IPS motif discovery
Motif quality analysis
Site search on gene set
Site search on track
Site search result optimization
Site search summary
Statistical analysis
Routes to explain gene regulation and functi
User Guide
Load data
Normalize data
Detect differentially expressed genes
Discover functional enrichment
Identify master regulators in networks

You can quickly open your favorite items from this icon menu. To open any method, place the cursor on the symbol and you get the complete name of the method which can be clicked to open in the workspace as shown below:


Research: test
Start page
Site search on gene set X
analyses/Methods/Site analysis/Site search on gene set

Yes set	(select element)
No set	□ (select element)
Species	Human (Homo sapiens)
From	-1000
To	100
Profile	...rofiles/vertebrate_non_redundant_minSUM
Output path	(select element)

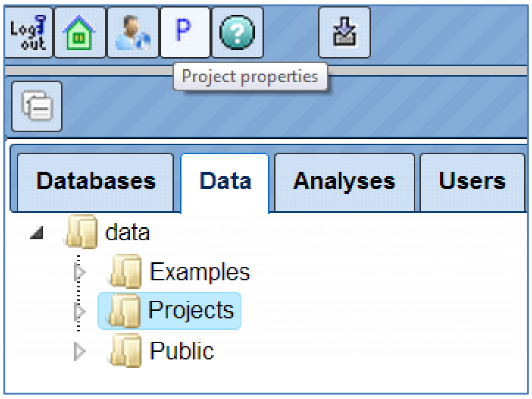
In the above screenshot, the method 'Site search on gene set' has been opened through the user toolbar. If you want to remove any method from the user toolbar, right click the symbol and you will get an option 'REMOVE'.

2.2.5. Project properties, or how to fix the releases of the databases to be applied in user's projects

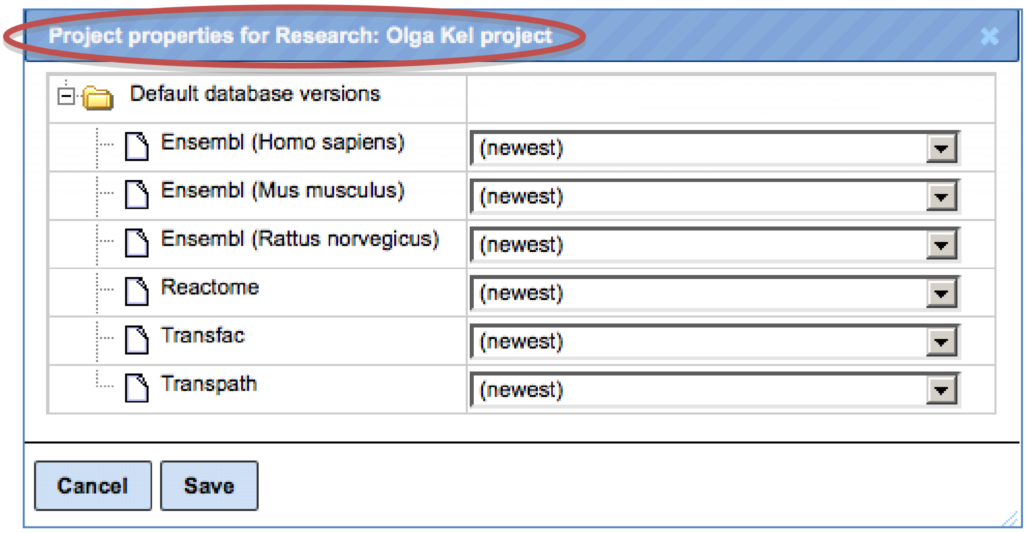
The geneXplain platform provides access to several versions of the databases installed.

You can select your desired version of the database and fix it for your project using the *Project properties* feature. By default the latest version of all databases are applied.

Project Properties form can be opened via the button P in the control panel:

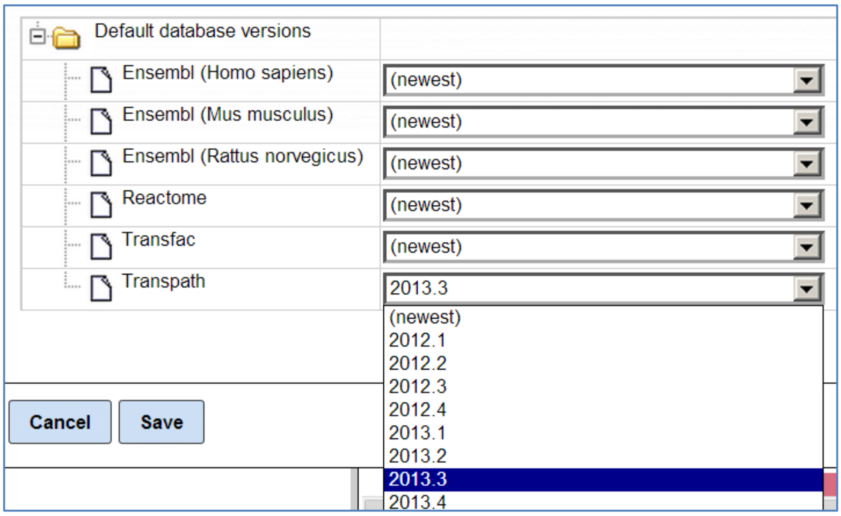


The form opens as shown below:


Project properties for Research: Olga Kel project

Default database versions	
Ensembl (Homo sapiens)	(newest)
Ensembl (Mus musculus)	(newest)
Ensembl (Rattus norvegicus)	(newest)
Reactome	(newest)
Transfac	(newest)
Transpath	(newest)

For each database shown in the form, the available versions can be selected using the drop down menu. As shown below, TRANSPATH® 2013.3 database version is selected. After selection press [Save] and the selected version of the database will be used in this project for all the analyses.


Default database versions	
Ensembl (Homo sapiens)	(newest)
Ensembl (Mus musculus)	(newest)
Ensembl (Rattus norvegicus)	(newest)
Reactome	(newest)
Transfac	(newest)
Transpath	2013.3

Important to note, this change is applied for one selected project, the project name is shown on the top of the form, highlighted by the red oval on the picture above. If you have several projects, you can fix database versions for each project individually.

2.3. How to handle tables, tracks, diagrams: basics

2.3.1. File handling

2.3.1.1. File selection

All analysis tools, and likewise all prepared workflows, require input data from a file in the Tree Area. The respective file can be loaded into an analysis tool by

- *simple dragging-and-dropping of the file from the Tree Area into the corresponding field in the Work Space, or*
- *by clicking into this field and making the selection from the directory which opens.*

Multiple selections can be done only in the second way.

2.3.1.2. File handling in the Tree Area

On mouse-over, the file name will gain a faint-bluish background. Upon a single click on a file name, its background will turn into a persistent light blue, and information about this file will be displayed in the Info Box. Double-clicking on a file will open it under a new tab in the Work Space, and the file name will be additionally emphasized by bold lettering as long as the corresponding tab is in the Work Space's foreground. Only files with an icon attached can be opened this way.

Files can also be opened by right-clicking on them and selecting the "Open table" option. You can also delete files ("Remove"; default value is "No") this way. Both functions are also available through the corresponding icons (and , resp.) in the Control Panel on top of the different frames.

While opening and deleting files works for nearly any file, the third option is different among distinct file types:

- Tables (and derivatives) can be exported. When activating this function, a selection of different formats is provided for the file to be generated. It is also available through the icon on top of the Tree Area.
- Diagrams () can be expanded/collapsed to show/hide their components (nodes and edges); this function can also be accessed by clicking on the white arrowhead next to the diagram icon.

Plots () can be edited. Only under this option, their complete deletion is possible as well.

2.3.2. Tables

Any table may be opened by double-clicking the corresponding name in the Tree Area. It will open under a new tab in the Work Space.

The contents of the table are sorted according to the values in one of its columns. Being opened for the first time, a default column is defined for sorting, usually the ID column. This default column is indicated by a blue arrowhead. If this arrowhead points upwards, the table rows are sorted in ascending order of this column's values. Clicking on this arrowhead will change it into a downwards pointing one, while the values are sorted in descending order. Correspondingly, you may sort the table according to the values of any other column in ascending or descending order by clicking on the up- or downwards pointing gray arrowhead on top of this column, respectively.

On top of the table, you can navigate between the individual pages of the table; it is also shown on which page out of how many pages in total you are, and in the right top corner, the page size in terms of number of entries (rows) is shown and can be adjusted.

You can edit the contents of a table by pressing the [Edit] button in the right upper corner. Now, you can manually edit the contents of each cell in the table. With the [Apply] option, you will save this change, while [Cancel] quits it.

Even without activating the Edit function, you can select

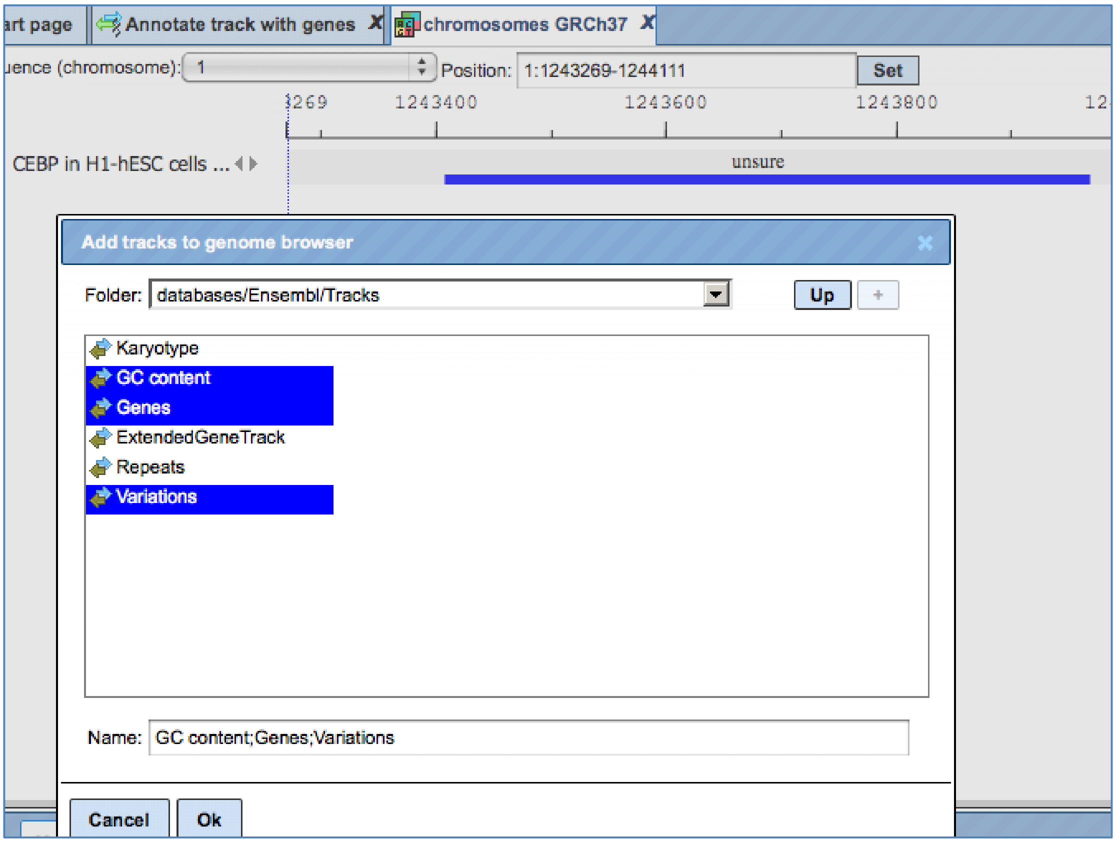
- individual rows with a left-mouse click,
- several ones by keeping the Ctrl key pressed,
- a range of rows with the Shift key pressed when clicking on the last row of the range to be selected, or
- [Select all] by clicking on the corresponding button.

The selected rows can be saved as a separate file, which by default is given the name *<original file name> subset*, but you can change this name.

2.3.3. Basic operations with tracks

View track in genome browser

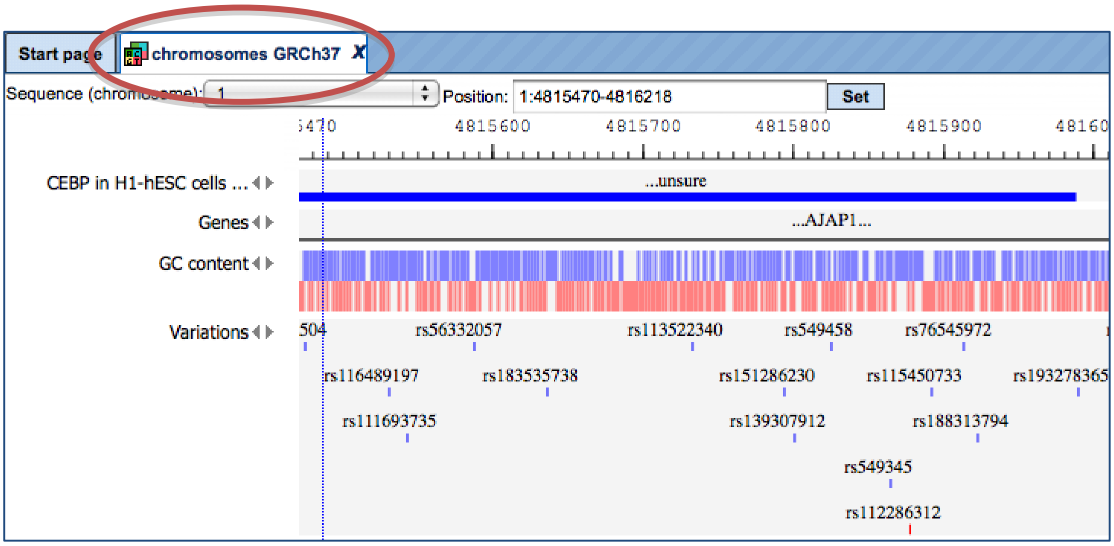
Upon double-clicking on a track name in the tree area the track will be opened in the work space in the genome browser.



Folder: databases/Ensembl/Tracks

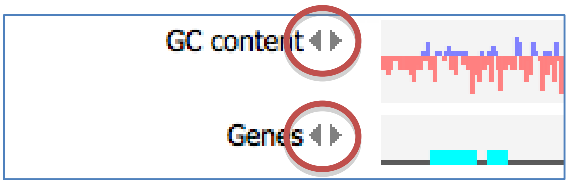
Name: GC content;Genes;Variations

In the pop-up window *Add tracks to genome browser* you can select which tracks, among those available in Ensembl, should be opened together with your track. Here, three tracks are selected, *GC-content*, *Genes*, and *Variations*. When the selection is ready, push the [Ok] to get the following view with your track on top:



On the tab name you can see genome, species and build information for this track, highlighted by the red oval.

The small triangles on the right side of the track name can be used to jump to the next or previous site of this track:



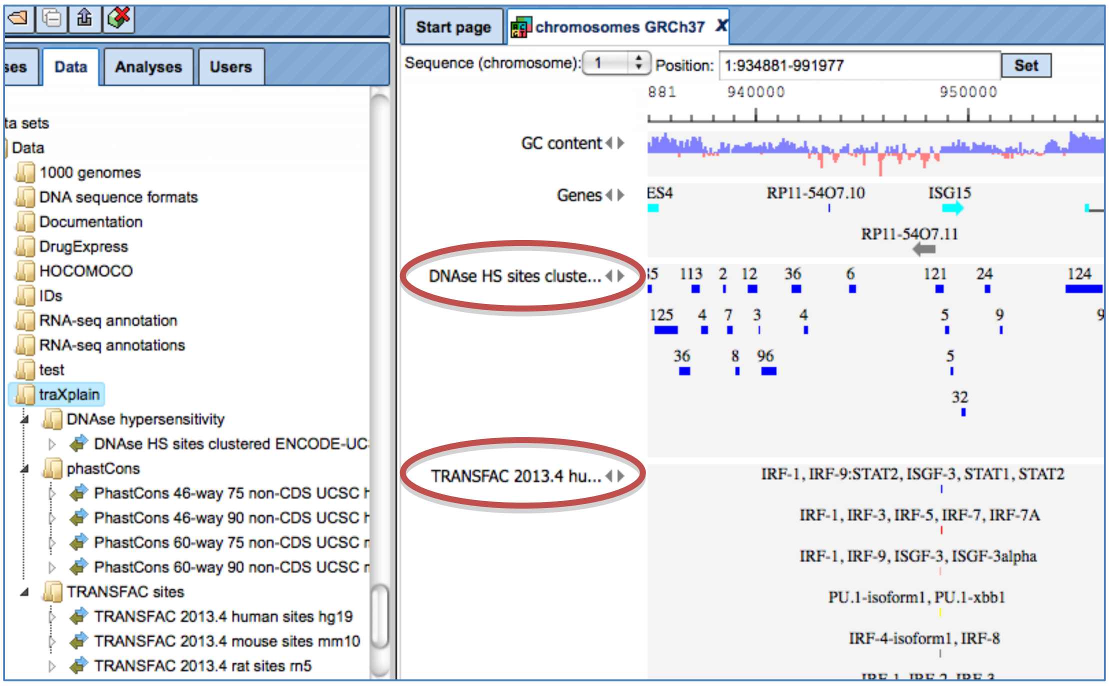
Use the buttons in the top control menu to zoom in and out.

The buttons help to shift the visible part to the left or to the right.

The same effect can be also achieved by dragging the picture with the mouse.

You can also jump between different chromosomes by selecting the chromosome number in the field *Sequence (chromosome)*.

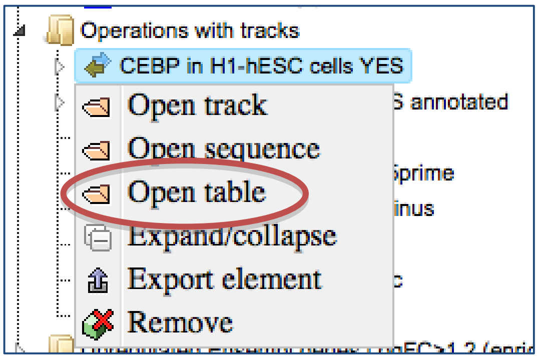
As the next step to enrich the visualization, and to gather more information about your track of interest, you can add the additional pre-existing tracks from the folder *Public*. There, you can find a sub-folder traXplain with several tracks. This sub-folder is highlighted in blue in the screenshot below. As an example, two tracks have been added, the track with DNAse hypersensitive sites from the ENCODE project (*DNAse HS sites clustered ENCODE UCSC hg19*), and the track with the experimentally proven transcription factor binding sites from the TRANSFAC® database (*TRANSFAC 2013.4 human sites hg19*).


Sequence (chromosome): 1 Position: 1:934881-991977 Set

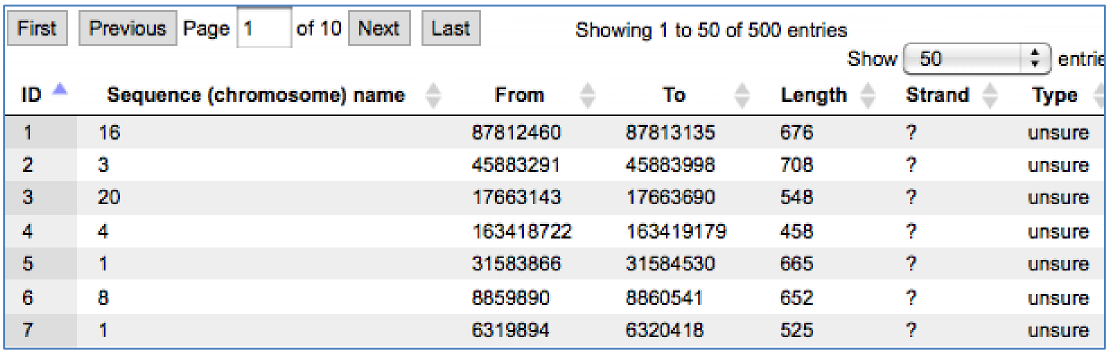
Open track as a table

To open a track as a table, use a right mouse click on the track name in the tree area or [Ctrl + mouse click] for Mac users.

Using the same menu, you can apply other functions to the selected track, e.g. export it in available formats or delete it.



A tabular view for the same track is shown below. Each row corresponds to one fragment. For each fragment, you can see the column **ID** with the fragment number, **Chromosome**, positions in the respective columns **From** and **To**, and several additional columns, three in this case.

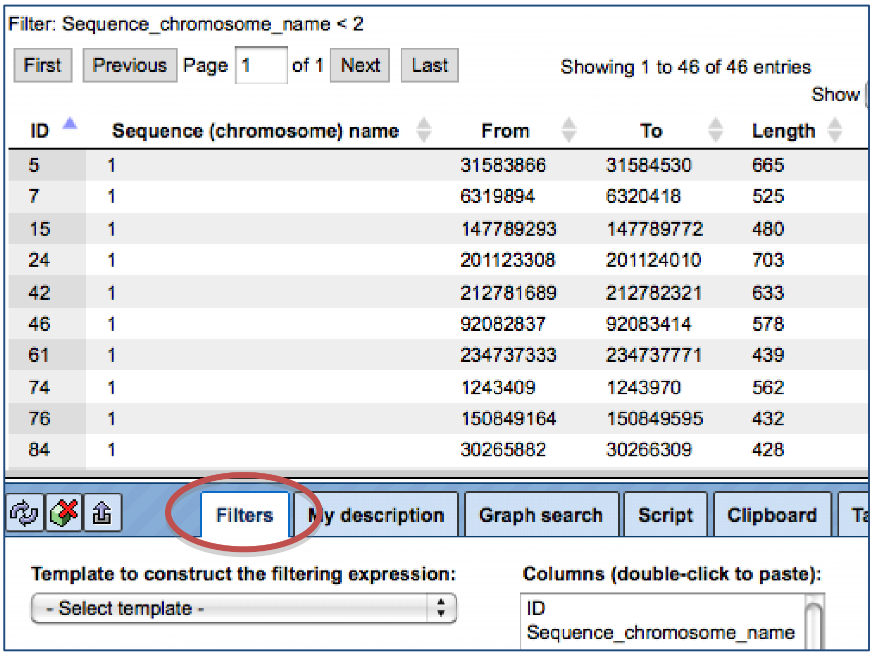


Showing 1 to 50 of 500 entries

ID	Sequence (chromosome) name	From	To	Length	Strand	Type
1	16	87812460	87813135	676	?	unsure
2	3	45883291	45883998	708	?	unsure
3	20	17663143	17663690	548	?	unsure
4	4	163418722	163419179	458	?	unsure
5	1	31583866	31584530	665	?	unsure
6	8	8859890	8860541	652	?	unsure
7	1	6319894	6320418	525	?	unsure

Filter track by condition.

When any track is opened as a table, it can be filtered by any condition in the specified columns. Filter options are available under the tab Filters in the operations field.



Filter: Sequence_chromosome_name < 2

Showing 1 to 46 of 46 entries

ID	Sequence (chromosome) name	From	To	Length
5	1	31583866	31584530	665
7	1	6319894	6320418	525
15	1	147789293	147789772	480
24	1	201123308	201124010	703
42	1	212781689	212782321	633
46	1	92082837	92083414	578
61	1	234737333	234737771	439
74	1	1243409	1243970	562
76	1	150849164	150849595	432
84	1	30265882	30266309	428

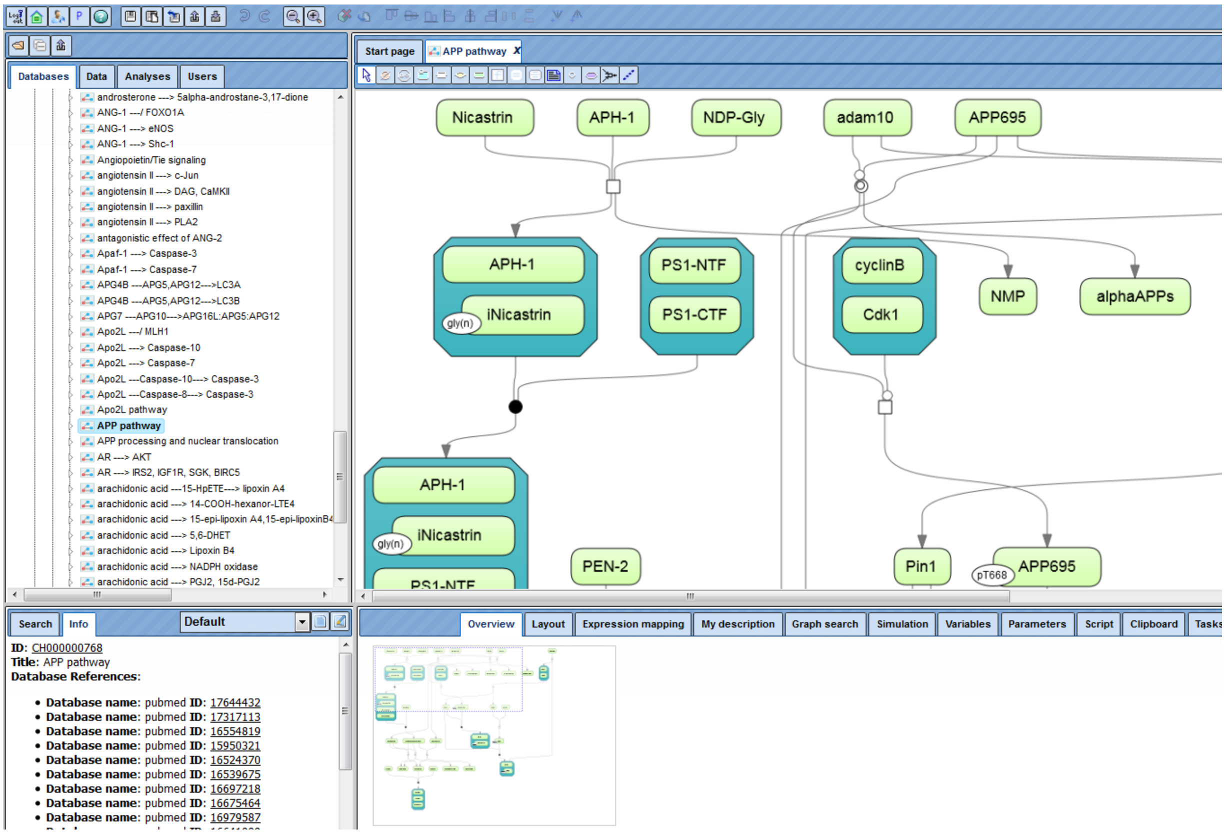
Template to construct the filtering expression: - Select template -

Columns (double-click to paste):
ID
Sequence_chromosome_name

Here, the filter is applied to select the fragments located on chromosome 1.

2.3.4. Diagram handling

Diagrams are provided by a number of databases and tools in the platform. The general schema of their use is that an overview of the corresponding graph is shown in the Operation Field, under the tab Overview, while a full-sized picture is shown in the Work Space, where usually only a part of the whole diagram fits on the screen. The diagrams exhibit components as ovals, reactions as squares, and links as lines or arrows.



androsterone ---> 5alpha-androstane-3,17-dione
ANG-1 ---/ FOXO1A
ANG-1 ---> eNOS
ANG-1 ---> Shc-1
Angiopoietin/Tie signaling
angiotensin II ---> c-Jun
angiotensin II ---> DAG, CaMKII
angiotensin II ---> paxillin
angiotensin II ---> PLA2
antagonistic effect of ANG-2
Apaf-1 ---> Caspase-3
Apaf-1 ---> Caspase-7
APG4B ---APG5,APG12--->LC3A
APG4B ---APG5,APG12--->LC3B
APG7 ---APG10--->APG16L:APG5:APG12
Apo2L ---/ MLH1
Apo2L ---> Caspase-10
Apo2L ---> Caspase-7
Apo2L ---Caspase-10---> Caspase-3
Apo2L ---Caspase-8---> Caspase-3
Apo2L pathway
APP pathway
APP processing and nuclear translocation
AR ---> AKT
AR ---> IRS2, IGF1R, SGK, BIRC5
arachidonic acid ---15-HpETE---> lipoxin A4
arachidonic acid ---> 14-COOH-hexanor-LTE4
arachidonic acid ---> 15-epi-lipoxin A4,15-epi-lipoxinB4
arachidonic acid ---> 5,6-DHET
arachidonic acid ---> Lipoxin B4
arachidonic acid ---> NADPH oxidase
arachidonic acid ---> PGJ2, 15d-PGJ2

ID: CH000000768
Title: APP pathway
Database References:

- **Database name**: pubmed **ID**: 17644432
- **Database name**: pubmed **ID**: 17317113
- **Database name**: pubmed **ID**: 16554819
- **Database name**: pubmed **ID**: 15950321
- **Database name**: pubmed **ID**: 16524370
- **Database name**: pubmed **ID**: 16539675
- **Database name**: pubmed **ID**: 16697218
- **Database name**: pubmed **ID**: 16675464
- **Database name**: pubmed **ID**: 16979587

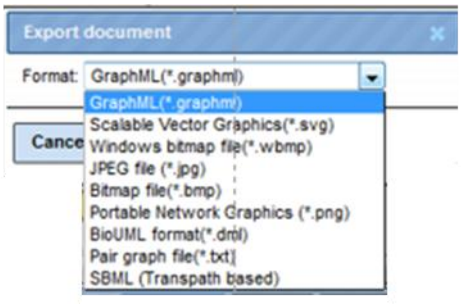
The section displayed can be shifted either in the Work Space by moving the mouse pointer (hand symbol) keeping the left mouse button pressed, or by shifting the blue-dotted rectangle in the Operation Field (mouse over: pointer symbol turns into a four-arrow plus sign, shifting can be performed while keeping the left mouse button pressed).

To facilitate orientation in large diagrams, individual edges are highlighted (turning from a thin, usually black arrow into a thick light-blue one).

Double-click on an individual node will show information about this component under the Info tab of the Info Box (this works for BioModels).

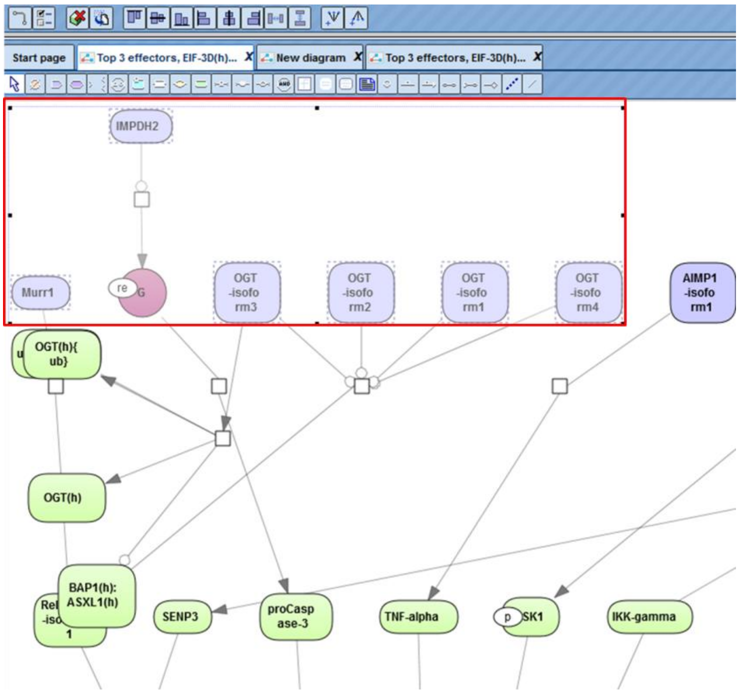
Diagrams can be zoomed in and zoomed out, by click on the buttons 🔍 or 🔍, correspondingly.

They can be exported in several formats by clicking the button ⇪.

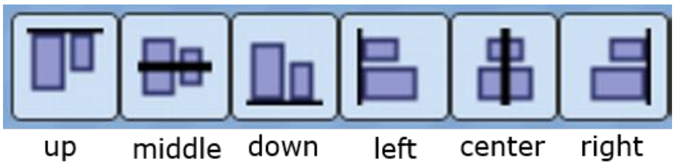


Diagrams can be zoomed in and zoomed out, by clicking on the buttons or , respectively.

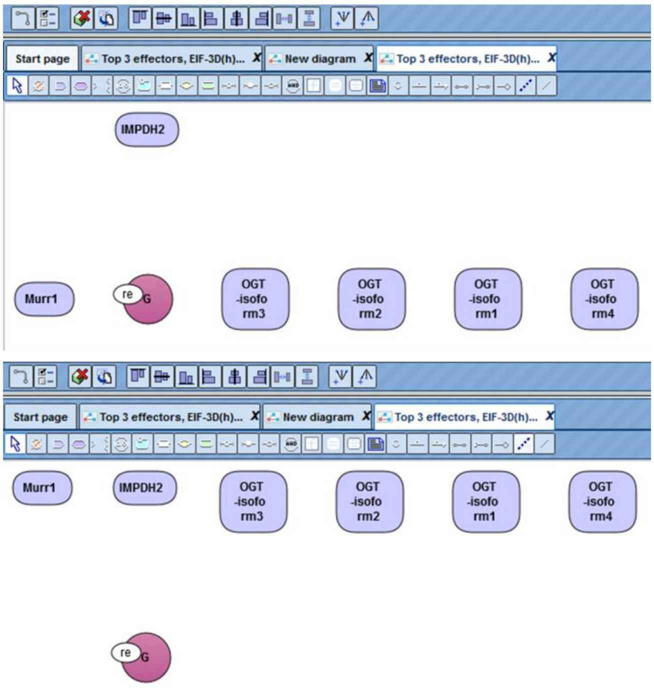
Diagram nodes can be multi-selected via the [Ctrl] button (picture below). Selected nodes can be used for *Alignments* or *Distribution* editing within the diagram or can be saved as a subset.



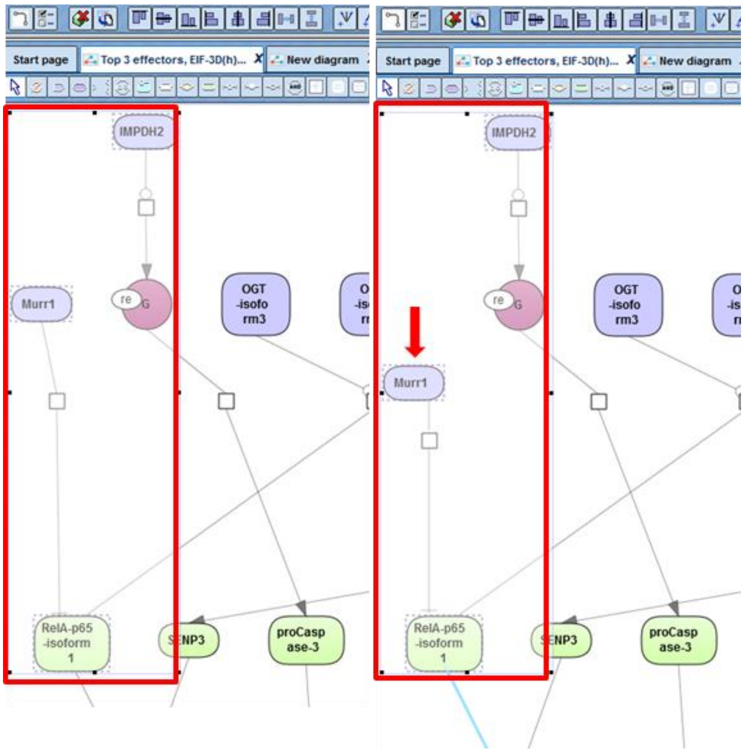
Five alignment methods are available in the tool bar (see picture below) after opening a diagram in the workspace.



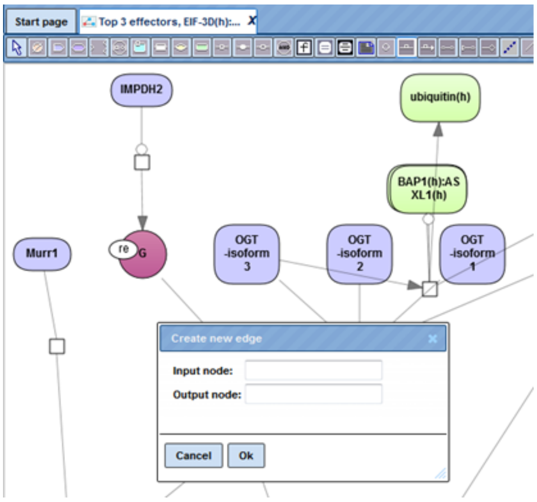
After pushing one of the buttons in the toolbar, the selected nodes in the diagram are aligned accordingly. The picture below shows an *Up alignment* of all nodes.

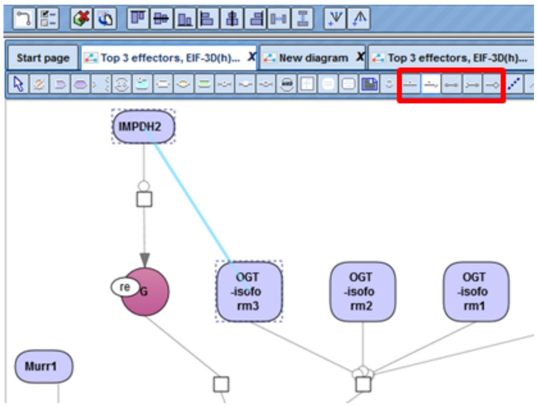


Two distribution methods horizontally and vertically are available in the tool bar after opening a diagram in the workspace. After pushing one of the buttons in the toolbar, the selected nodes in the diagram are positioned accordingly. The picture below shows a *Vertically distribution* of one node.



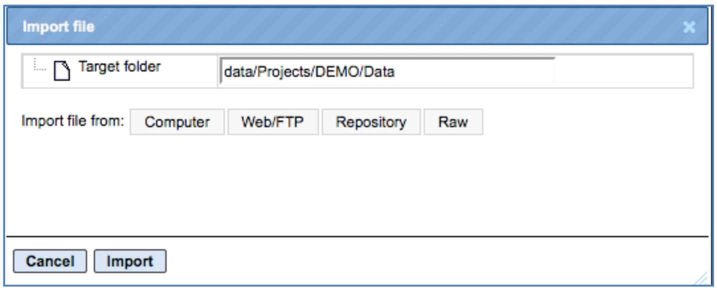
For edge editing you can use either the dialog box or a clickable connecting line to add new edges to a diagram.


Start page
Top 3 effectors, EIF-3D(h):...
IMPDH2
ubiquitin(h)
BAP1(h):ASXL1(h)
Murr1
re
G
OGT-isoform 3
OGT-isoform 2
OGT-isoform 1
Create new edge
Input node:
Output node:
Cancel
Ok


Start page
Top 3 effectors, EIF-3D(h)...
New diagram
Top 3 effectors, EIF-3D(h)...
IMPDH2
re
G
OGT-isoform3
OGT-isoform2
OGT-isoform1
Murr1

3. Load data

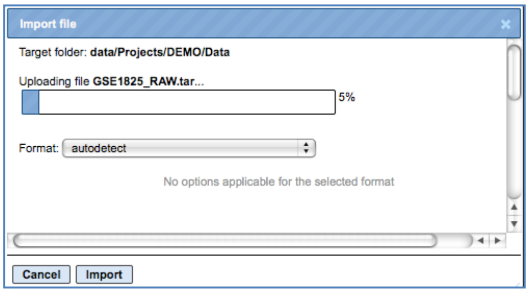
The first step is to load your data into the system. This is facilitated by the first topic listed on the Start page. When you click on the term "Load data", another window with the title "Import file" is opened asking you for the file to upload. By clicking into the topmost field (Target folder), you can select the place where you want to save the file that you are going to upload. This file you can select by several options: upload from your computer, import directly from an FTP address, choose an item from the tree in the geneXplain platform, or paste data, e.g. a DNA sequence. Use one of the buttons [Computer], [Web/FTP], [Repository] or [Raw] to start the import.



Target folder: data/Projects/DEMO/Data

3.1.1. Import a file from a local computer

Let us assume you wish to detect differentially expressed genes in your microarray experiments, which is one of the following workflow groups on the Start page. In this case, you might be interested to upload an archive (ZIP, TAR, GZ,…) of several CEL files (when you worked with an Affymetrix platform). In case of importing raw microarray results, we recommend to create a single ZIP archive of all files that you plan to normalize together. Please note that uploading is a lengthy process, and it may take some time. However, if in the uploading bar you see no progress and just 0% for about 10-15 minutes, please cancel and start again, it might be that the server was overloaded at that moment. Wait till the progress bar is completed. The picture below shows an ongoing uploading.



Target folder: **data/Projects/DEMO/Data**

Uploading file **GSE1825_RAW.tar...**

5%

Format: autodetect

No options applicable for the selected format

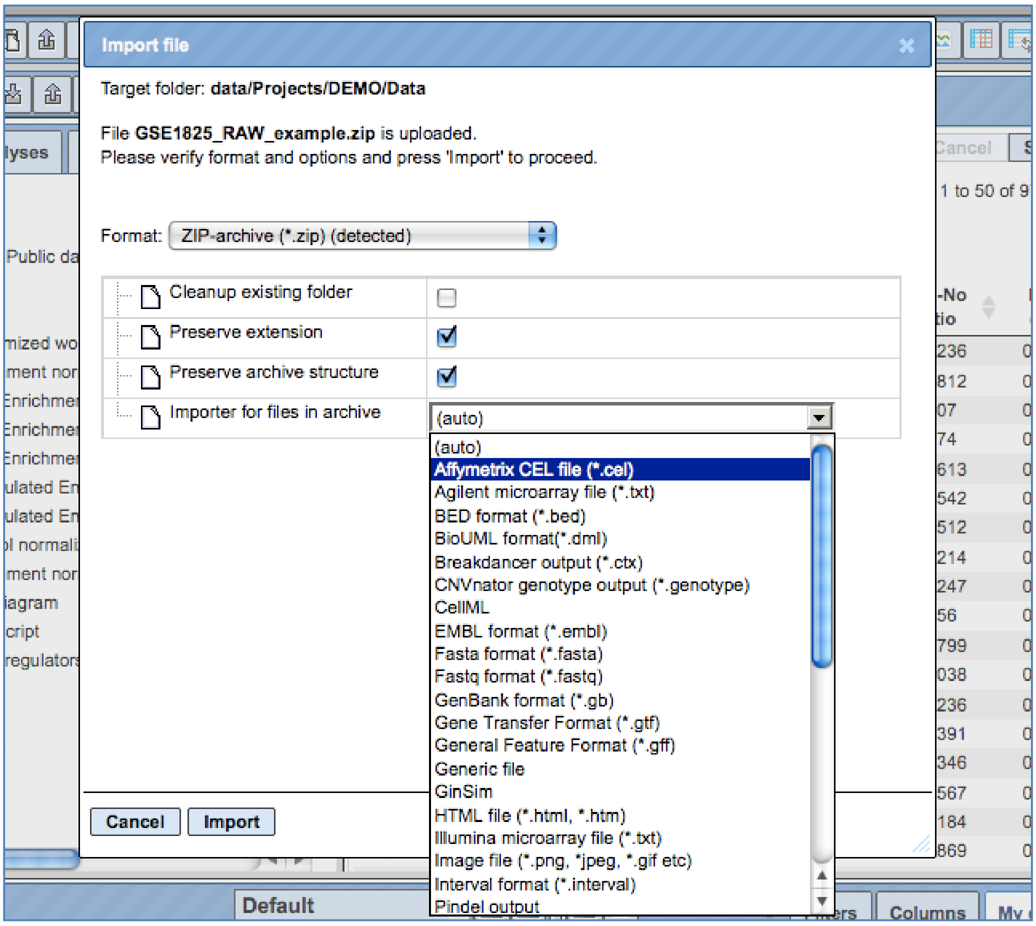
The uploading step is similar for all file formats. After the uploading step has been completed, the next step is specific for the file format, and different forms are displayed for different formats. Here, let's consider two examples, loading a ZIP archive with Affymetrix CEL files and loading an Excel table.

Loading a ZIP archive with Affymetrix CEL files

When the uploading step is complete, additional options will be displayed, as shown below. The format of the imported file is detected automatically and will be indicated in field “Format”, in the example below it is ZIP-archive. Verify the automatic detection and if necessary refine it manually using the drop-down menu. Next, verify four fields as follows.

In case that here exists already a folder with the same name as the imported file, it will be replaced by imported data if you activate the check box “Cleanup existing folder”. Checking the box “Preserve extension” means that individual CEL files will keep the extension ”.CEL” after importing.

Checking the box “Preserve archive structure” allows you to keep the archive structure (folders and subfolders) after importing. Next, choose the type of the files within the imported archive (if an archive is imported), here “Affymetrix CEL files” row is selected, indicated by the dark blue background in the picture below. Finally, press the [Import] button.



Import file

Target folder: **data/Projects/DEMO/Data**

File **GSE1825_RAW_example.zip** is uploaded.
Please verify format and options and press 'Import' to proceed.

Format: ZIP-archive (*.zip) (detected)

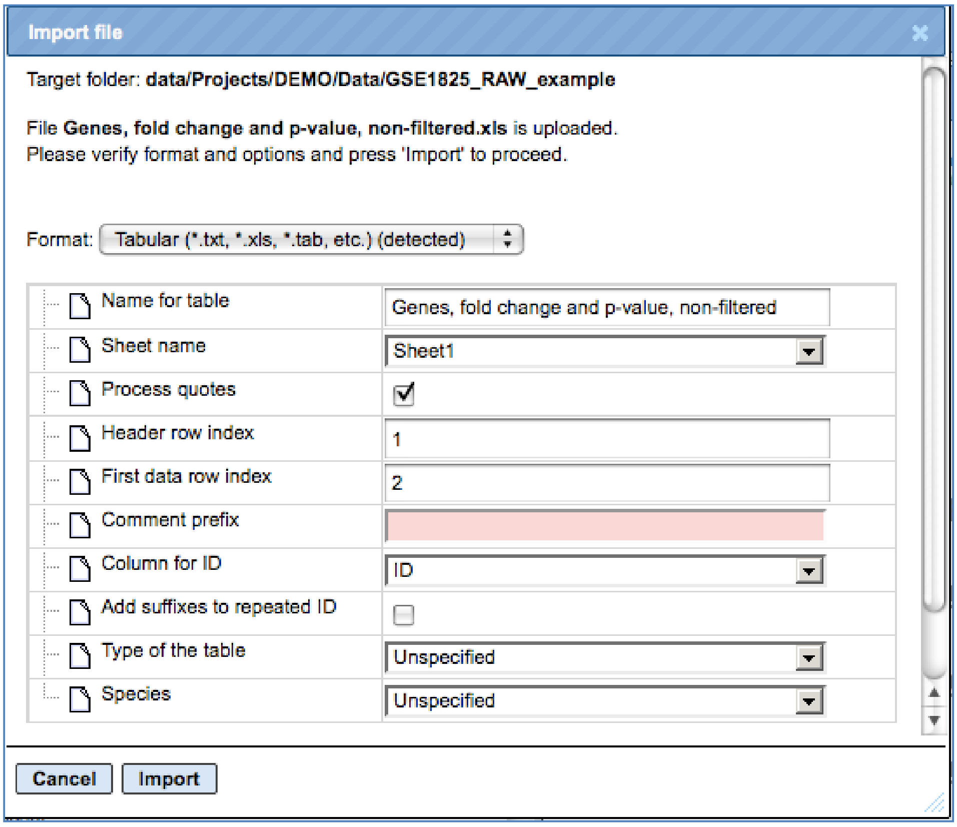
Cleanup existing folder	☐
Preserve extension	☑
Preserve archive structure	☑
Importer for files in archive	(auto)

(auto)
Affymetrix CEL file (*.cel)
Agilent microarray file (*.txt)
BED format (*.bed)
BioUML format(*.dml)
Breakdancer output (*.ctx)
CNVnator genotype output (*.genotype)
CellML
EMBL format (*.embl)
Fasta format (*.fasta)
Fastq format (*.fastq)
GenBank format (*.gb)
Gene Transfer Format (*.gtf)
General Feature Format (*.gff)
Generic file
GinSim
HTML file (*.html, *.htm)
Illumina microarray file (*.txt)
Image file (*.png, *jpeg, *.gif etc)
Interval format (*.interval)
Pindel output
Cancel
Import
Default
Columns

Each imported archive or individual file will appear in the Data tree of the Tree Area when import is complete.

Loading an Excel table

After the uploading step of an Excel table is completed, the following additional options will be displayed as shown below:



Target folder: **data/Projects/DEMO/Data/GSE1825_RAW_example**

File **Genes, fold change and p-value, non-filtered.xls** is uploaded.
Please verify format and options and press 'Import' to proceed.

Format: Tabular (*.txt, *.xls, *.tab, etc.) (detected)

Name for table	Genes, fold change and p-value, non-filtered
Sheet name	Sheet1
Process quotes	☑
Header row index	1
First data row index	2
Comment prefix	
Column for ID	ID
Add suffixes to repeated ID	□
Type of the table	Unspecified
Species	Unspecified

The format of the imported file is detected automatically and is indicated in the field "Format". In the figure above it is the Tabular format, which includes the XLS format among others. Verify the automatically detected format and, if necessary, refine it manually using the drop-down menu. Next, verify the following fields.

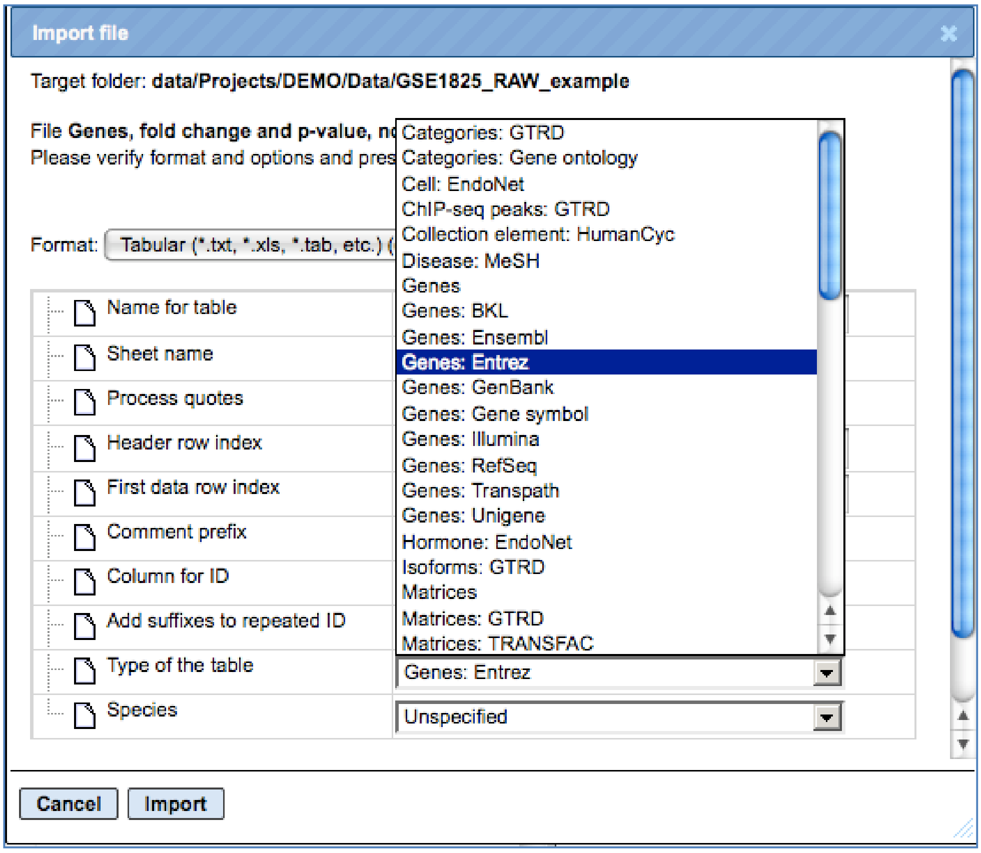
"Name for table" – the name of the uploaded Excel table is shown automatically; you can modify it, and under this name the table will be shown in the tree after import is completed. "Sheet name" – with the help of the drop-down menu specify which page of the Excel file should be imported (for Excel files with multiple pages).

"Header row index" – the number of header rows in the Excel table is detected automatically, in this example there is one header row. Verify automatic detection and if necessary correct the number. "First data row index" – the number of the row where the data start is detected automatically, in this example the data start in the second row and in the field figure "2" is shown. Verify automatic detection and, if necessary, correct the number.

"Column for ID" – use the drop-down menu to indicate which column in the imported table contains unique identifiers. The column "ID" is suggested automatically. Verify automatic suggestion and refine it as necessary.

"Type of the table" – this field aims to specify what type of IDs are used as identifiers in the imported Excel table. Use the drop-down menu to specify. In this example IDs in the Excel table correspond to Entrez gene IDs, and correspondingly in the figure below "Genes: Entrez" is specified.

"Species" – use the drop down menu to specify the biological species that corresponds to the IDs.



Import file

Target folder: **data/Projects/DEMO/Data/GSE1825_RAW_example**

File **Genes, fold change and p-value, n**
Please verify format and options and pres

Format: Tabular (*.txt, *.xls, *.tab, etc.) (

Categories: GTRD
Categories: Gene ontology
Cell: EndoNet
ChIP-seq peaks: GTRD
Collection element: HumanCyc
Disease: MeSH
Genes
Genes: BKL
Genes: Ensembl
Genes: Entrez
Genes: GenBank
Genes: Gene symbol
Genes: Illumina
Genes: RefSeq
Genes: Transpath
Genes: Unigene
Hormone: EndoNet
Isoforms: GTRD
Matrices
Matrices: GTRD
Matrices: TRANSFAC

Name for table	
Sheet name	
Process quotes	
Header row index	
First data row index	
Comment prefix	
Column for ID	
Add suffixes to repeated ID	
Type of the table	Genes: Entrez
Species	Unspecified

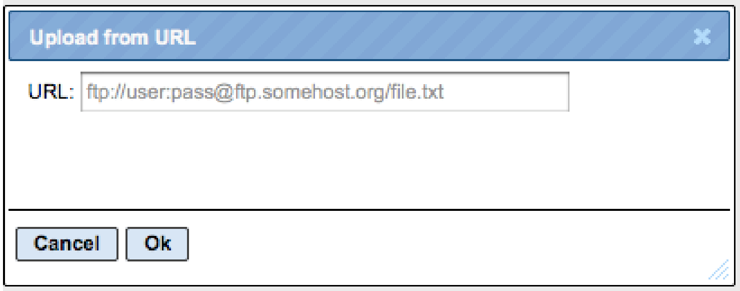
After that, press the [Import] button, and you can find the imported table in the tree, under the location specified in the “Import file” form in the field “Target folder”.

A number of file formats can be defined for import. You find them listed when opening the “Format” selection list. If you can’t find the required format explicitly listed, select “Generic file”.

3.1.2. Uploading data from URL

This option allows you to upload data directly from the web, e.g. by indicating an FTP address, which is significantly faster in comparison with downloading to a local computer and then uploading into the geneXplain platform. If your data are located on a web server, this would be a preferable way of importing.

First, press [Web/FTP] button on the “Import file” form, and indicate the ftp address in the newly opened form as shown below:

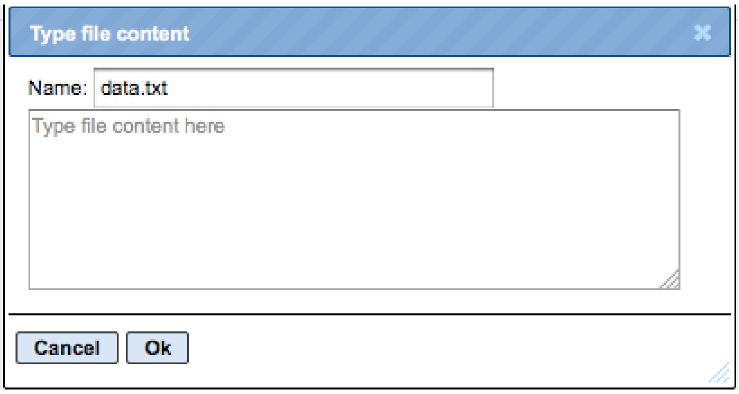


Upload from URL

URL: ftp://user:pass@ftp.somehost.org/file.txt

3.1.3. Paste plain DNA sequence

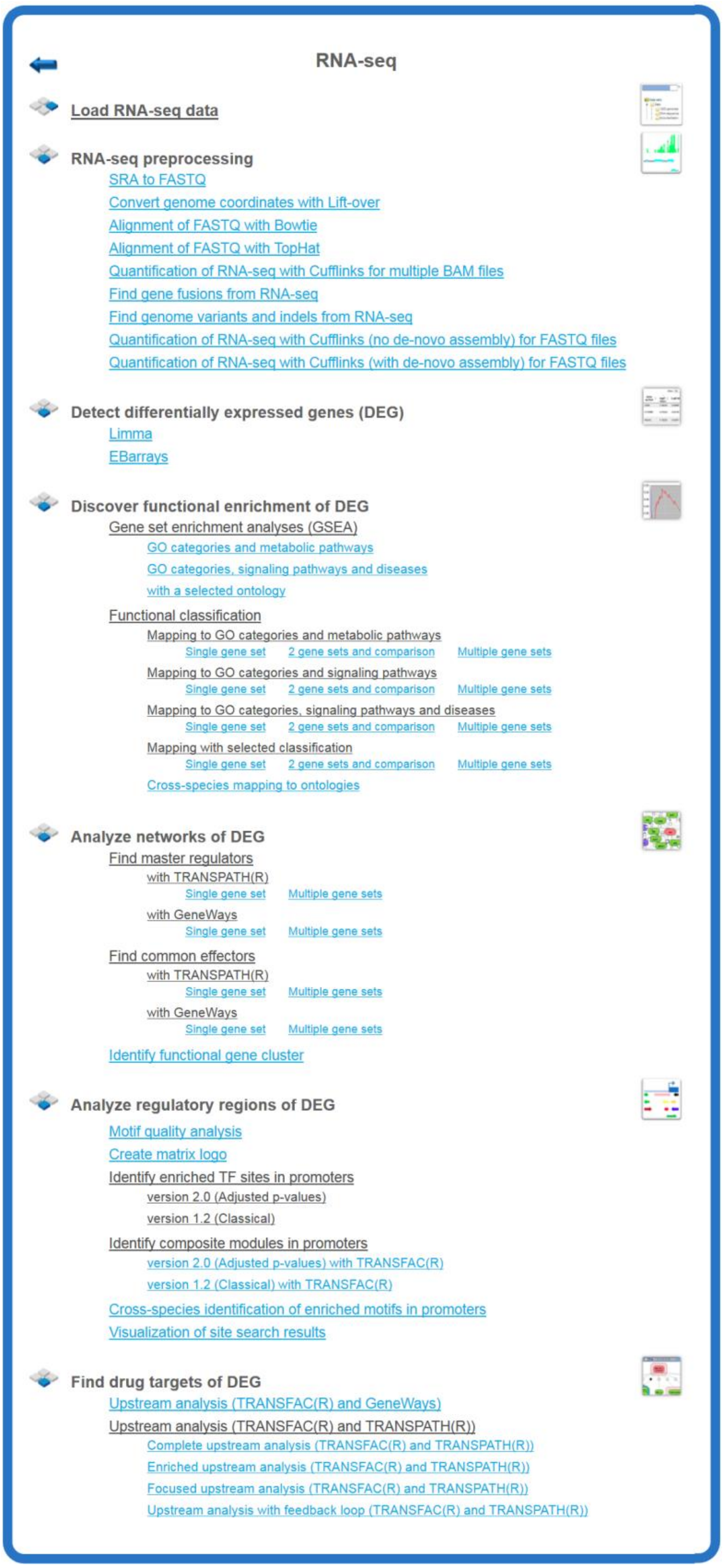
If you would like to import a plain DNA sequence, not saved in any file before, first, choose the button [Raw] on the form "Import file", and a new form will be opened as shown below:



Paste the DNA sequence in the window "Type file content here", and in the field "Name" specify a name as it should appear in the tree after importing.

4. RNA-seq

When you click on the tiled button “RNA-seq” in the Work Space of the on the start page, the following listing will appear:



RNA-seq

Load RNA-seq data

RNA-seq preprocessing
- SRA to FASTQ
- Convert genome coordinates with Lift-over
- Alignment of FASTQ with Bowtie
- Alignment of FASTQ with TopHat
- Quantification of RNA-seq with Cufflinks for multiple BAM files
- Find gene fusions from RNA-seq
- Find genome variants and indels from RNA-seq
- Quantification of RNA-seq with Cufflinks (no de-novo assembly) for FASTQ files
- Quantification of RNA-seq with Cufflinks (with de-novo assembly) for FASTQ files

Detect differentially expressed genes (DEG)
- Limma
- EBarrays

Discover functional enrichment of DEG
- Gene set enrichment analyses (GSEA)
  - GO categories and metabolic pathways
  - GO categories, signaling pathways and diseases
  - with a selected ontology
- Functional classification
  - Mapping to GO categories and metabolic pathways
    - Single gene set
    - 2 gene sets and comparison
    - Multiple gene sets
  - Mapping to GO categories and signaling pathways
    - Single gene set
    - 2 gene sets and comparison
    - Multiple gene sets
  - Mapping to GO categories, signaling pathways and diseases
    - Single gene set
    - 2 gene sets and comparison
    - Multiple gene sets
  - Mapping with selected classification
    - Single gene set
    - 2 gene sets and comparison
    - Multiple gene sets
  - Cross-species mapping to ontologies

Analyze networks of DEG
- Find master regulators
  - with TRANSPATH(R)
    - Single gene set
    - Multiple gene sets
  - with GeneWays
    - Single gene set
    - Multiple gene sets
- Find common effectors
  - with TRANSPATH(R)
    - Single gene set
    - Multiple gene sets
  - with GeneWays
    - Single gene set
    - Multiple gene sets
- Identify functional gene cluster

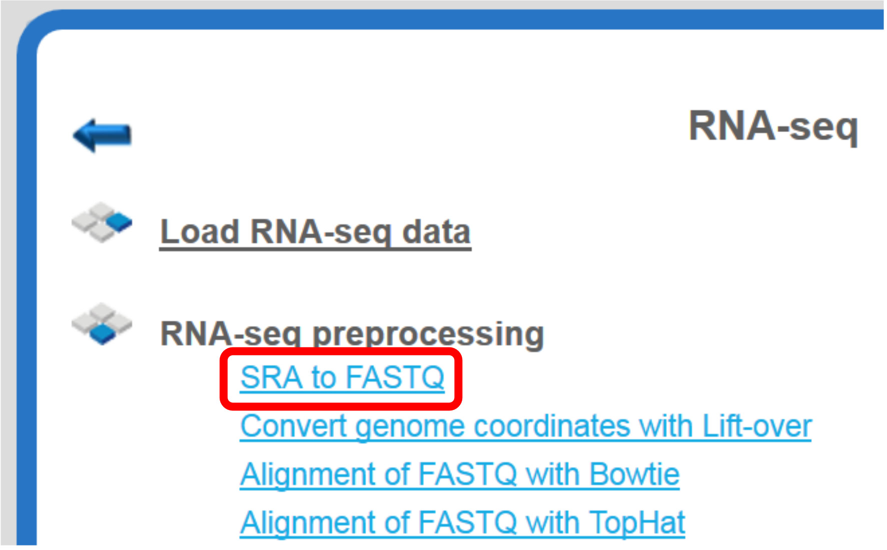
Analyze regulatory regions of DEG
- Motif quality analysis
- Create matrix logo
- Identify enriched TF sites in promoters
  - version 2.0 (Adjusted p-values)
  - version 1.2 (Classical)
- Identify composite modules in promoters
  - version 2.0 (Adjusted p-values) with TRANSFAC(R)
  - version 1.2 (Classical) with TRANSFAC(R)
- Cross-species identification of enriched motifs in promoters
- Visualization of site search results

Find drug targets of DEG
- Upstream analysis (TRANSFAC(R) and GeneWays)
- Upstream analysis (TRANSFAC(R) and TRANSPATH(R))
  - Complete upstream analysis (TRANSFAC(R) and TRANSPATH(R))
  - Enriched upstream analysis (TRANSFAC(R) and TRANSPATH(R))
  - Focused upstream analysis (TRANSFAC(R) and TRANSPATH(R))
  - Upstream analysis with feedback loop (TRANSFAC(R) and TRANSPATH(R))

4.1. RNA-seq preprocessing

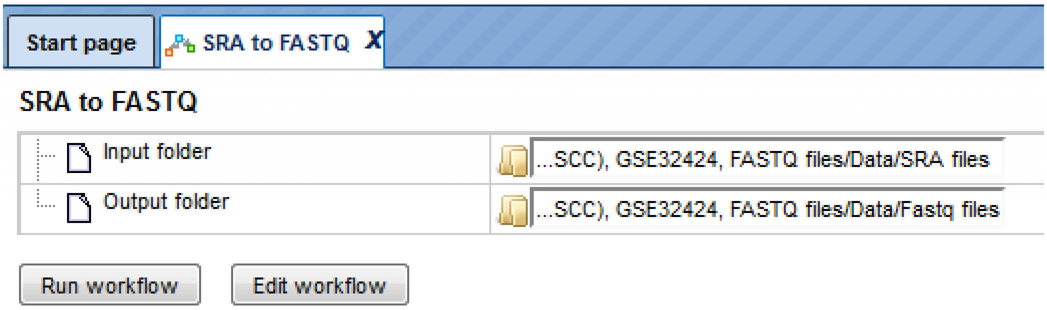
4.1.1. SRA to FASTQ

This workflow can be used to convert SRA data files (e. g. from NGS/RNA-seq experiments) into FASTQ files. The FASTQ format is widely used by a number of tools and the geneXplain platform is among them; on the other hand, NGS data are often collected in SRA format, thus the conversion of SRA format into FASTQ format is an important function. An example of public data stored in SRA format can be found here (http://www.ncbi.nlm.nih.gov/sra?term=SRP051443) and can be uploaded directly via FTP import into the geneXplain platform. The workflow "SRA to FASTQ" can be found on the Start page, under the NGS/RNA-seq button.


RNA-seq
Load RNA-seq data
RNA-seq preprocessing
SRA to FASTQ
Convert genome coordinates with Lift-over
Alignment of FASTQ with Bowtie
Alignment of FASTQ with TopHat

To launch the workflow, follow these steps:

Step1. Open the workflow input form from the Start page. It looks as shown below:



SRA to FASTQ

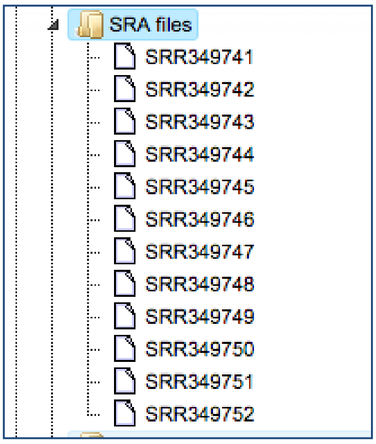
Input folder	...SCC), GSE32424, FASTQ files/Data/SRA files
Output folder	...SCC), GSE32424, FASTQ files/Data/Fastq files

Step 2. Specify the folder with the SRA files in the field **Input folder**. You can drag it from your project within the tree area and drop it in the box beside the folder pictogram. Alternatively, you may click on the field *(select element)* and a new window will be opened, where you can select the input folder.

Example of an input folder:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/RNA-Seq analysis of human esophageal squamous cell carcinoma (ESCC)%2C GSE32424%2C FASTQ files/Data/SRA%2files/

It contains 12 files in SRA format as shown below. Please note, this folder occupies **1.5 GB** work space.



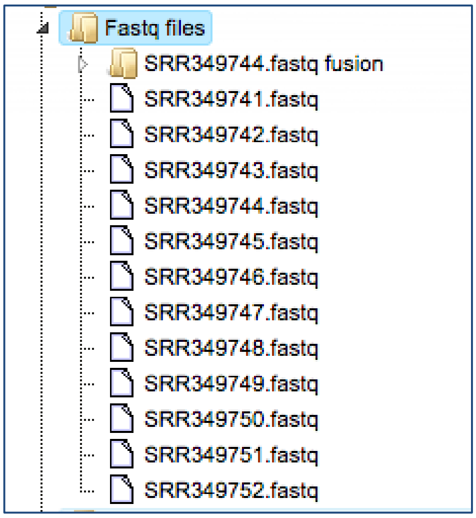
SRA files
SRR349741
SRR349742
SRR349743
SRR349744
SRR349745
SRR349746
SRR349747
SRR349748
SRR349749
SRR349750
SRR349751
SRR349752

The output folder name and folder path is automatically created, but can be changed in a user-specific way.

Press [Run workflow] and wait till the workflow is completed.

Example of an output folder:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/RNA-Seq analysis of human esophageal squamous cell carcinoma (ESCC)%2C GSE32424%2C FASTQ files/Data/Fastq%2files/



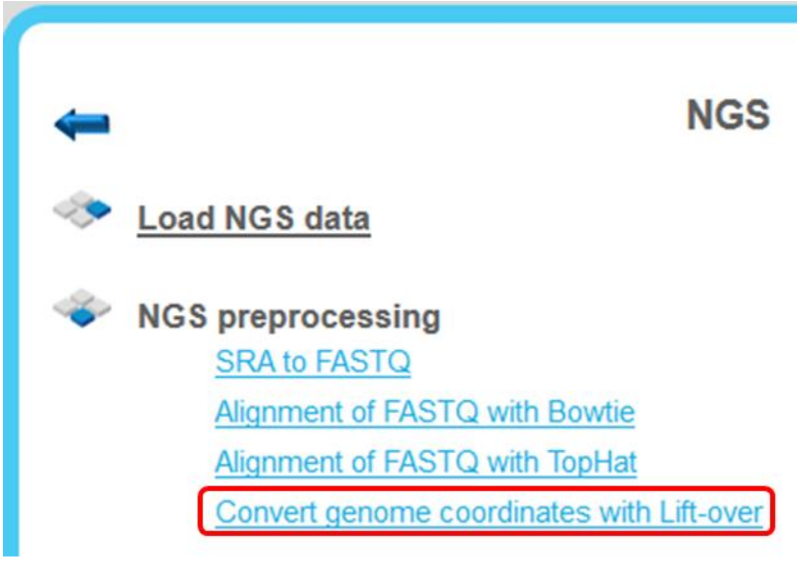
Fastq files
SRR349744.fastq fusion
SRR349741.fastq
SRR349742.fastq
SRR349743.fastq
SRR349744.fastq
SRR349745.fastq
SRR349746.fastq
SRR349747.fastq
SRR349748.fastq
SRR349749.fastq
SRR349750.fastq
SRR349751.fastq
SRR349752.fastq

The output folder contains 12 files with the same names and now with the extension *fastq*. Please note, the size of the output folder is **16.6 GB**.

Note. Working with NGS data in SRA and FASTQ formats requires substantial work space available in your user account. Feel free to contact us (info@genexplain.com) to upgrade your account with additional disk space.

4.1.2. Convert genome coordinates with Lift-Over

While working with NGS data, quite often it might quite often be required to convert positions from one genome assembly to another one, and The Lift-Over program is widely can be applied for this task. Within the geneXplain platform, you can find it on the Start page under the NGS button, as highlighted below.

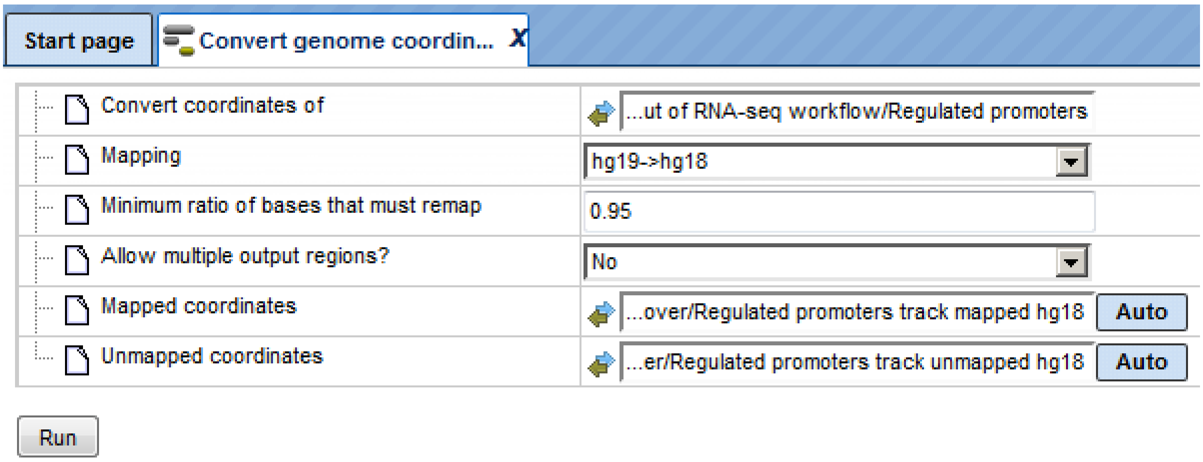

NGS
Load NGS data
NGS preprocessing
SRA to FASTQ
Alignment of FASTQ with Bowtie
Alignment of FASTQ with TopHat
Convert genome coordinates with Lift-over

This tool is based on the LiftOver utility and Chain track from the UC Santa Cruz Genome Browser.

It converts coordinates and annotations between assemblies and genomes. The input is a track with genomic positions according to a particular genome assembly, and the output is a track with positions according to another genome assembly.

To launch the workflow, follow these steps:

Step1. Open the input form from Start page. It looks as shown below:



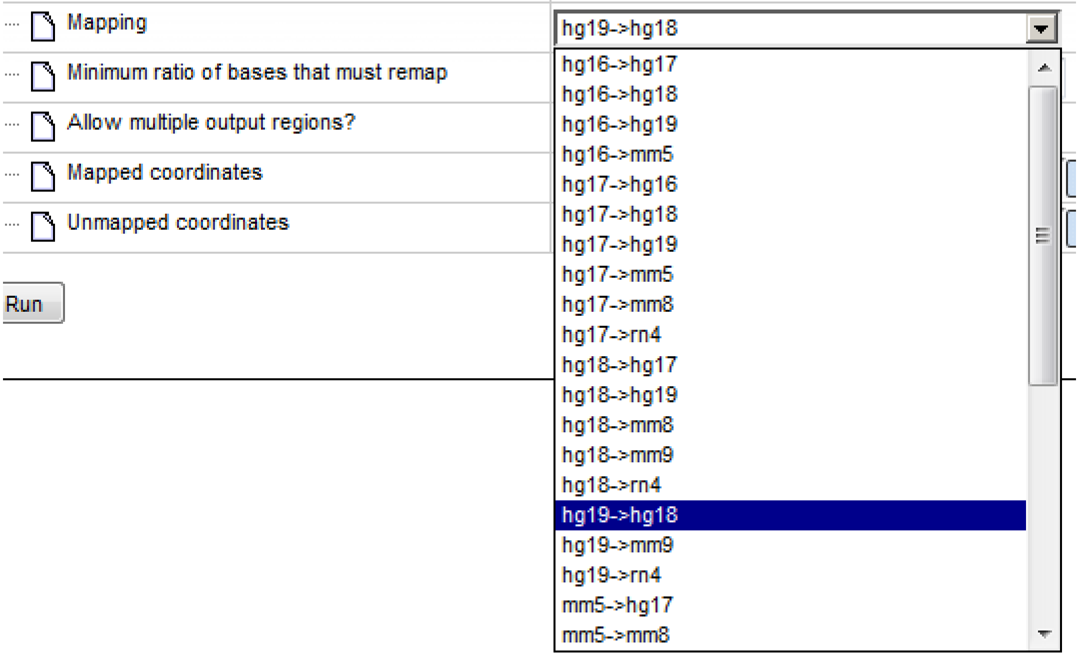
Convert coordinates of	...ut of RNA-seq workflow/Regulated promoters
Mapping	hg19->hg18
Minimum ratio of bases that must remap	0.95
Allow multiple output regions?	No
Mapped coordinates	...over/Regulated promoters track mapped hg18 Auto
Unmapped coordinates	...er/Regulated promoters track unmapped hg18 Auto

Step 2. Specify the input track in the field **Convert coordinates of**. You can drag it from your project within the tree area and drop it in the box of the field.

The further steps of the workflow are demonstrated by means of the tables in one of the pre-prepared examples. You can find these tables in the *Examples* folder, under

http://genexplain-platform.com/bioumlweb/#de=data/Examples/RNA-Seq analysis of human esophageal squamous cell carcinoma (ESCC)%2C GSE32424%2C FASTQ files/Data/Lift_over/

Step 3. Specify the **Mapping** of the input track by selecting the desired genome conversion from the drop-down menu.



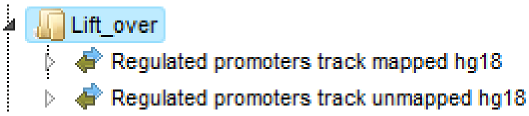
In the majority of cases not everything can be re-mapped to another assembly. The minimum ratio of bases that must be re-mapped is by default 0.95, you can change this by typing in this field.

Step 4. Allow multiple output regions: choose Yes or No.

Step 5. Define where the output tracks should be located in the tree. The method produces two output tracks, one containing all the mapped coordinates and the other containing the unmapped coordinates (if existing).

After filling out all input fields press [Run] and wait till the method is completed.

The output is a folder with two tracks as shown below:



- Lift_over
  - Regulated promoters track mapped hg18
  - Regulated promoters track unmapped hg18

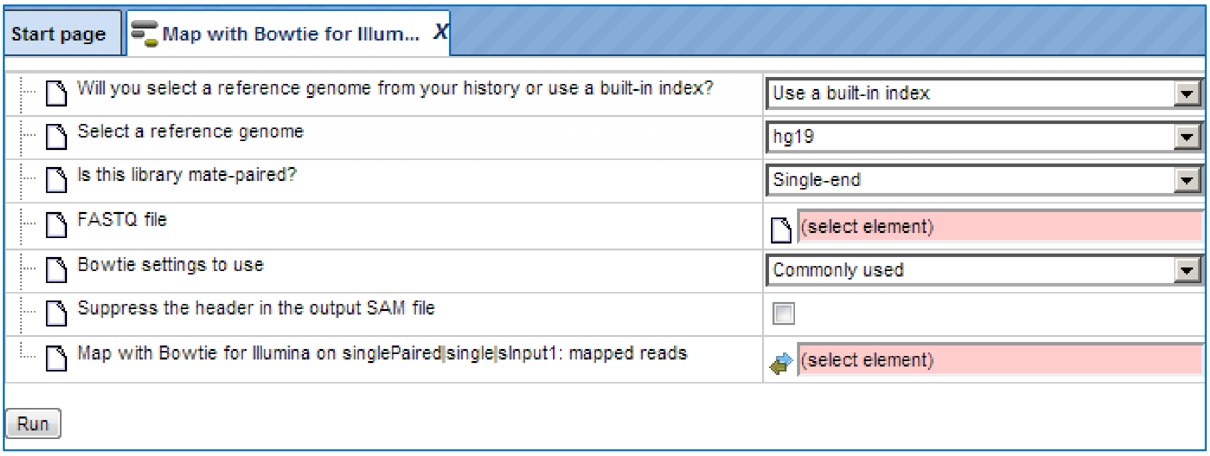
4.1.3. Alignment of FASTQ with Bowtie

Bowtie is a short-read aligner designed to be ultrafast and memory-efficient. It was developed by Ben Langmead and Cole Trapnell (Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology 10:R25). The Bowtie method can be found in the Galaxy section of the platform (analyses/Galaxy/solexa_tools/bowtie_wrapper).

Input format: Bowtie accepts files in Sanger FASTQ format. Often, the sequence files are represented in SRA format (Sequence Read Archives). This format is used to deposit sequences at the Sequence Read Archives at NCBI, EBI, and DDBJ. Please consider, converting the SRA files into FASTQ files before starting Bowtie. You can use the conversion tool *Convert Files in SRA format to FASTQ* or *Convert Files in SRA format to paired FASTQ* in the Galaxy section of the platform (analyses/Galaxy/sra_toolkit/fastq-dump).

To launch the Bowtie tool, follow these steps:

Step1. Open the Bowtie input form from the Start page by clicking on *Alignment of FASTQ with Bowtie* option in the *RNA-seq preprocessing* subsection. It will open in the main Work Space and looks as shown below:

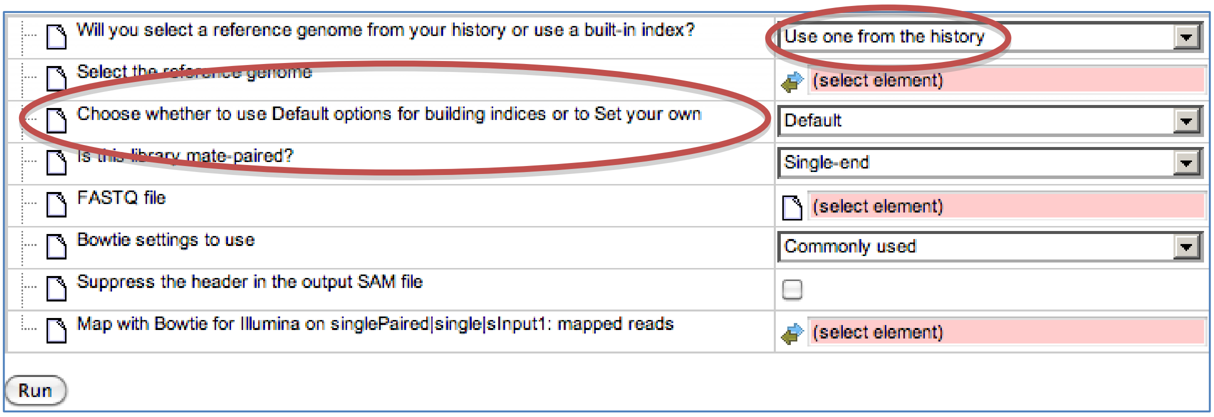


Will you select a reference genome from your history or use a built-in index?	Use a built-in index
Select a reference genome	hg19
Is this library mate-paired?	Single-end
FASTQ file	(select element)
Bowtie settings to use	Commonly used
Suppress the header in the output SAM file	□
Map with Bowtie for Illumina on singlePaired|single|sInput1: mapped reads	(select element)

Step 2. Specify the reference genome.

The field **Will you select a reference genome from your history or use a built-in index?** defines the reference genome. If you keep the default "Use a built-in index" the program makes an alignment to the reference genome which is provided as part of the platform (the genome builds for human, mouse and rat are provided). Depending on the species used, please specify hg19 for human, mm9 for mouse and rn4 for rat in the next field **Select a reference genome**.

If you select the "Use one from the history" option in the 1st field, two new fields will appears in the form: **Select the reference genome** and **Choose whether to use Default options for building indices or to Set your own**, as shown in the screenshot below, highlighted by the red ovals. In this case you should provide a preloaded reference genome (preloaded in Fasta, EMBL or Genebank formats) and choose the way how to build sequence indices which will be used by the alignment algorithm of Bowtie.



Will you select a reference genome from your history or use a built-in index?	Use one from the history
Select the reference genome	(select element)
Choose whether to use Default options for building indices or to Set your own	Default
Is this library mate-paired?	Single-end
FASTQ file	(select element)
Bowtie settings to use	Commonly used
Suppress the header in the output SAM file	□
Map with Bowtie for Illumina on singlePaired|single|sInput1: mapped reads	(select element)

Step 3. The field **Is this library mate-paired?** defines the type of the sequence library which was used in NGS sequencing. The default is *Single-end*, the alternative is *Paired-end*. You should know this details about your library of short reads.

Step 4. Specify the input file in the field **FASTQ file**. You can either drag-and-drop or select the file name from the Tree area. Here, as an example, we use data from a published RNA-seq experiment analyzing the human esophageal squamous cell carcinoma (ESCC), GSE32424. FASTQ files can be found in the following *Examples* folder:

data/Examples/RNA-Seq analysis of human esophageal squamous cell carcinoma (ESCC), GSE32424, FASTQ files

This example contains results of an RNA-seq Illumina NGS sequencing of twelve clinical samples from human esophageal squamous cell carcinoma (ESCC) (seven tumors and five non-tumors). The authors provided sequences as so called *non-aligned* BAM files. We loaded these BAM files directly from GEO as one archive using the ftp uploading function of the geneXplain platform. After that, we converted the non-aligned BAM files into FASTQ files using the tool: *SAM to FASTQ* from the *NGS: Picard (beta)* subsection of the Galaxy section of the platform (*analyses/Galaxy/picard_beta/picard_SamToFastq*).

Step 5. The field **Bowtie settings to use** is set by default to *Commonly used*. If you change it to the *Full parameter list* from the drop-down menu, a full list of parameters of the alignment algorithm is enabled for editing. The full description of all these parameters is given in the original paper of Langmead et al. mentioned above which describes the algorithm of Bowtie.

Step 6. Please keep the box **Suppress the header in the output SAM file** unchecked (as it is by default) to generate SAM/BAM output files suitable for further use by Cufflinks tool.

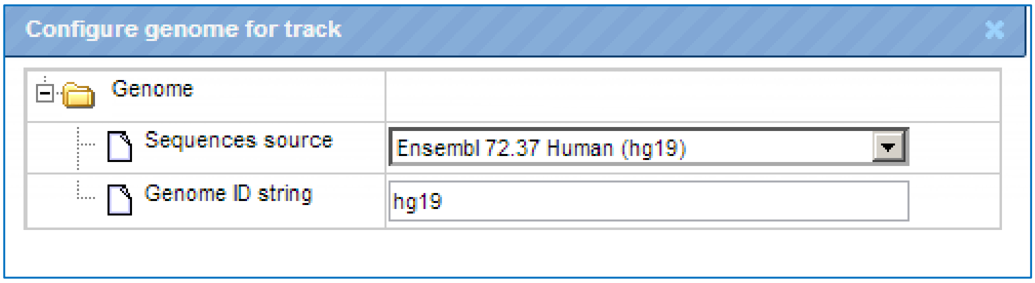
Step 7. Set the output file name (for the output BAM file) in the field **Map with Bowtie for Illumina…** and press the button [Run].

Tip It is recommended to save the output file into a separate folder containing all BAM output files from one particular experiment. This will allow you to run the next workflow for the quantification of all SAM/BAM files from this defined folder.

Results

The result of this method is one BAM file which is generated by the Bowtie program as the result of the alignment of the sequence reads from the input FASTQ file to the reference genome.

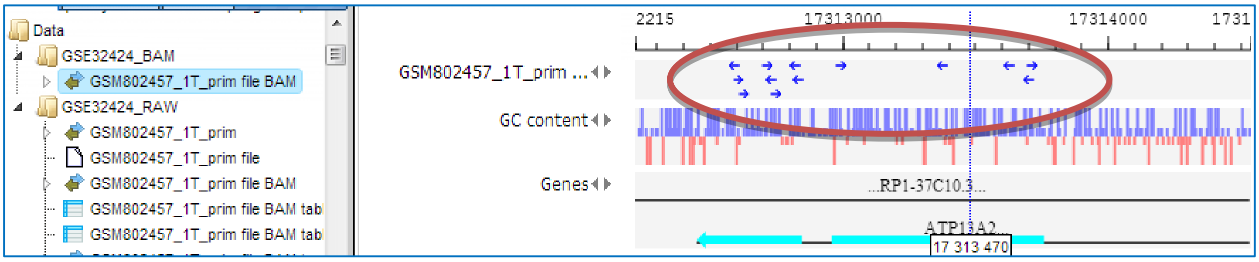
At the end of the Bowtie run, the platform requests to specify the genome build again in order to link the output BAM file to the respective genome files in the Ensembl database installed in the platform. When you get such a pop-up form, choose the sequence score from the drop-down menu and specify the genome ID.



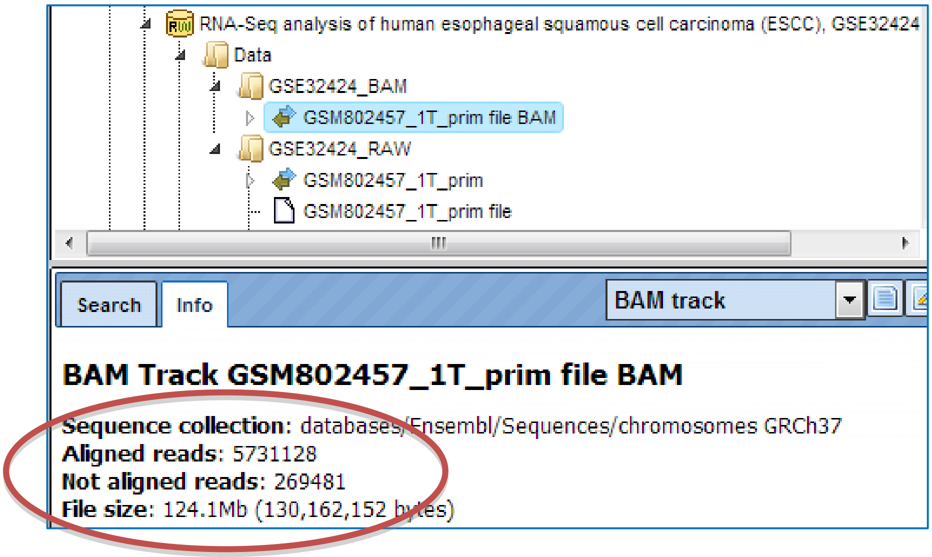
Configure genome for track

Genome	
Sequences source	Ensembl 72.37 Human (hg19)
Genome ID string	hg19

This enables a visualization of the BAM file information in the genome browser. Generated BAM file is a track and has the (◆) icon in the tree. As usual for all tracks, it with a double click opens in the Work Space. You can see the positions of each aligned read in the genome and upon zooming-in you get all detailed information about each read complete with sequence, length, and quality.



In the info box you can see information about the output BAM file. The number of aligned and not aligned reads and overall file size is shown.


RNA-Seq analysis of human esophageal squamous cell carcinoma (ESCC), GSE32424
Data
GSE32424_BAM
GSM802457_1T_prim file BAM
GSE32424_RAW
GSM802457_1T_prim
GSM802457_1T_prim file
Search
Info
BAM track

BAM Track GSM802457_1T_prim file BAM

Sequence collection: databases/Ensembl/Sequences/chromosomes GRCh37
Aligned reads: 5731128
Not aligned reads: 269481
File size: 124.1Mb (130,162,152 bytes)

Note. The input FASTQ and output BAM files of the Bowtie tool require a considerable amount of working space. One FASTQ file can occupy several GB of space. If you need more space for storage and work with your FASTQ and BAM files, please feel free to ask for details (info@genexplain.com).

4.1.4. Quantification of RNA-seq with Cufflinks for multiple BAM files

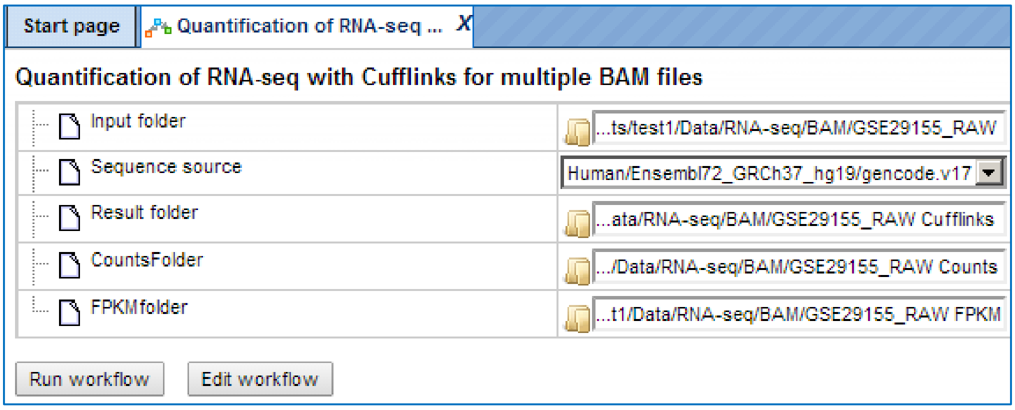
This workflow is designed to estimate abundances of transcripts in several RNA-Seq samples using the Cufflinks method (published in *Trapnell C, Williams BA, Pertea G, Mortazavi AM, Kwan G, van Baren MJ, Salzberg SL, Wold B, Pachter L. Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nature Biotechnology doi:10.1038/nbt.1621*).

The Cufflinks method accepts aligned RNA-Seq reads (in "aligned" BAM files) and assembles the alignments into a set of transcripts using a reference annotation of transcripts and genes. Cufflinks then estimates the relative abundances of these transcripts and genes based on how many reads support each one.

In the first part of the workflow, the Cufflinks program is called from the Galaxy section of the geneXplain platform (*analyses/Galaxy/ngs-rna-tools/cufflinks.*

To launch the workflow, follow these steps:

Step1. Open the workflow input form from the Start page. It will open in the main Work Space and looks as shown below:



Start page | Quantification of RNA-seq ...

Quantification of RNA-seq with Cufflinks for multiple BAM files

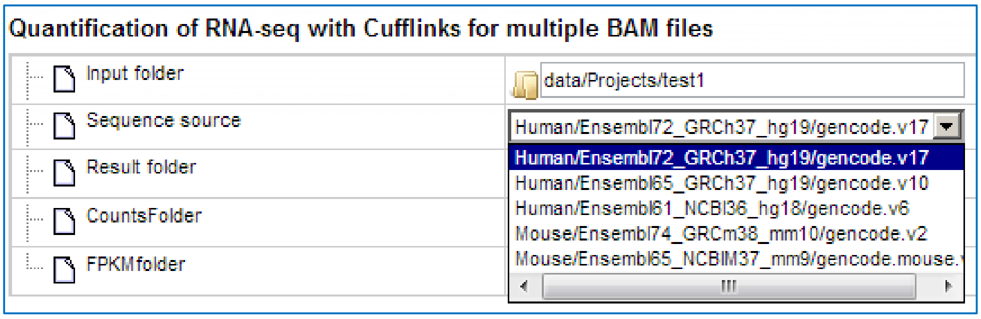
Input folder	...ts/test1/Data/RNA-seq/BAM/GSE29155_RAW
Sequence source	Human/Ensembl72_GRCh37_hg19/gencode.v17
Result folder	...ata/RNA-seq/BAM/GSE29155_RAW Cufflinks
CountsFolder	.../Data/RNA-seq/BAM/GSE29155_RAW Counts
FPKMfolder	...t1/Data/RNA-seq/BAM/GSE29155_RAW FPKM

Step 2. Input a folder containing BAM file(s) from the tree. You can either drag-and-drop or select the folder name from the Tree area. Here, as an example, we use data from a published RNA-seq experiment on Colon cancer retrieved from GEO, accession number GSE29155. The following folder from the *Examples* folder is used:

data/Examples/RNA-Seq analysis of human prostate cancer cell line, GSE29155, BAM files/Data/

The input folder in this example contains 11 BAM files of RNA-Seq reads obtained from 7 colon cancer samples and 4 normal samples using the Illumina NGS sequencer. The reads were aligned by the authors of the data to the human genome build hg18. We loaded these BAM files directly from GEO as once archive using the ftp uploading function of geneXplain platform.

Step 3. Choose an appropriate sequence source from the drop-down menu:



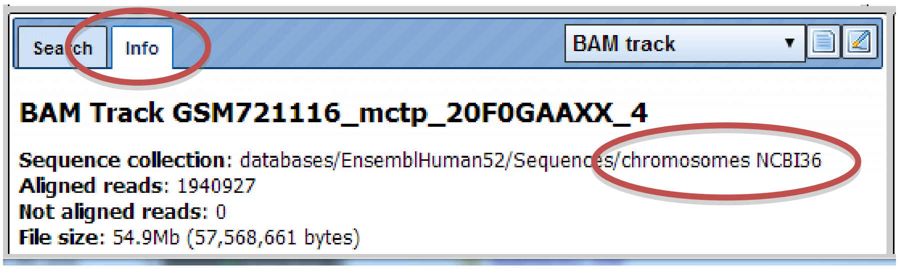
Quantification of RNA-seq with Cufflinks for multiple BAM files

Input folder	data/Projects/test1
Sequence source	Human/Ensembl72_GRCh37_hg19/gencode.v17
Result folder	
CountsFolder	
FPKMfolder	

Human/Ensembl72_GRCh37_hg19/gencode.v17
Human/Ensembl65_GRCh37_hg19/gencode.v10
Human/Ensembl61_NCBI36_hg18/gencode.v6
Mouse/Ensembl74_GRCm38_mm10/gencode.v2
Mouse/Ensembl65_NCBIM37_mm9/gencode.mouse.

The drop-down menu presents in each line: Species, Ensembl build number, Genome build code (e.g. GRCh37_hg19) and the corresponding version of the Genecode derived gene annotation. Select an option which corresponds to the RNA-seq data analyzed and stored in the input BAM files.

The Genome build code refers to the genome build which was used for the alignment of RNA-Seq data when producing the BAM files. You can check the genome build information by clicking on the individual BAM file in the tree area and getting information in the Info box. For example:



BAM Track GSM721116_mctp_20F0GAAXX_4

Sequence collection: databases/EnsemblHuman52/Sequences/chromosomes NCBI36
Aligned reads: 1940927
Not aligned reads: 0
File size: 54.9Mb (57,568,661 bytes)

This BAM file was produced by alignment of RNA-Seq data to the NCBI36_hg18 genome build (in the "Sequence collection" field you can see the "chromosomes NCBI36")

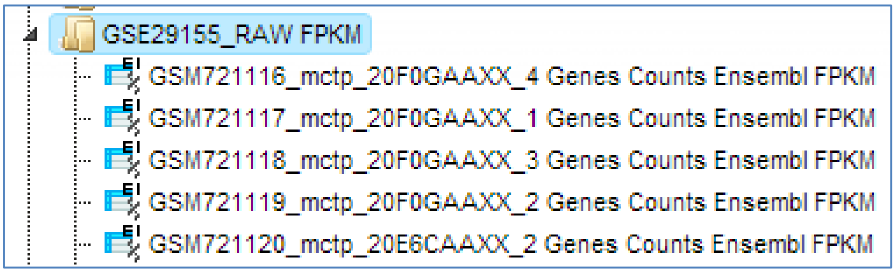
The choice of the Genecode version depends on the particular needs of the user. The higher version of Genecode corresponds to the most up-to-date gene annotation, whereas the earlier Genecode version may correspond to the gene annotation used in other types of data in the same study, and may therefore be chosen for consistency with other data. (e.g. genecode.v10 corresponds to the most recent annotation done in the ENCODE project)

Step 4. Specify the output folder names. The *Result folder* and *CountsFolder* are created temporarily for storing Cufflinks outputs and intermediate quantification outputs. **The FPKMfolder** defines the folder for the final output of the workflow. The abbreviation FPKM stays for **F**ragments **P**er **K**ilobase of transcript per **M**illion mapped reads, and is a commonly accepted standard measure for this kind of data.

Step 5. Press [Run workflow] and wait till the workflow is completed.

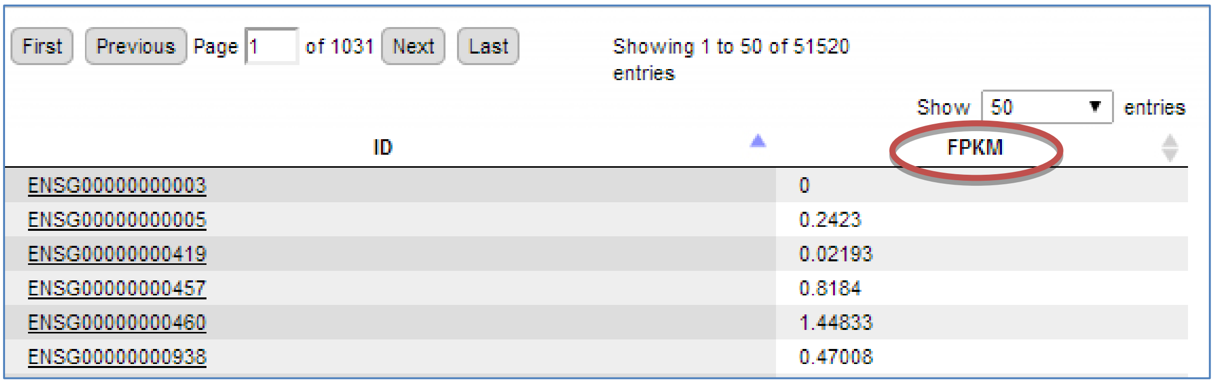
Results.

The results folder consists of several tables of Ensembl type containing the results of quantification of every BAM file from the input folder:



GSE29155_RAW FPKM
GSM721116_mctp_20F0GAAXX_4 Genes Counts Ensembl FPKM
GSM721117_mctp_20F0GAAXX_1 Genes Counts Ensembl FPKM
GSM721118_mctp_20F0GAAXX_3 Genes Counts Ensembl FPKM
GSM721119_mctp_20F0GAAXX_2 Genes Counts Ensembl FPKM
GSM721120_mctp_20E6CAAXX_2 Genes Counts Ensembl FPKM

By double-clicking on each table you can see the result of the quantification.



First Previous Page 1 of 1031 Next Last Showing 1 to 50 of 51520 entries

Show 50 entries

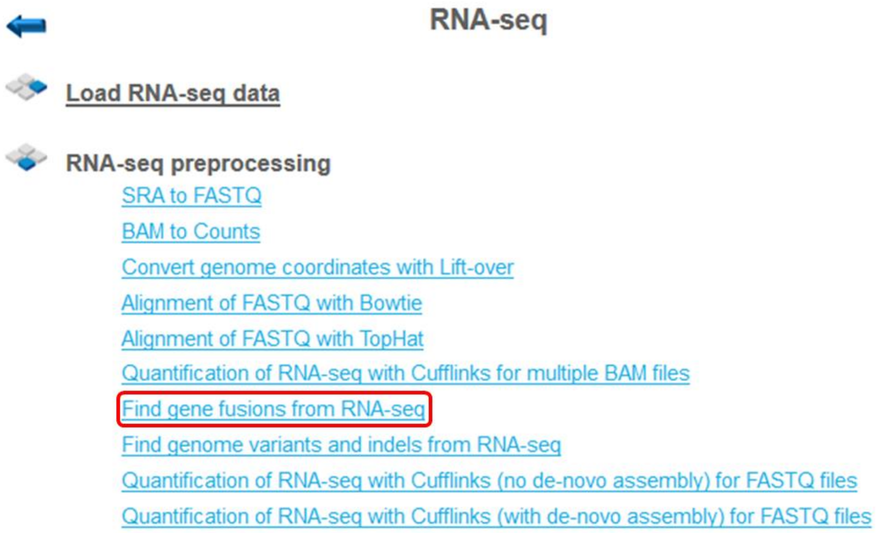
ID	FPKM
ENSG00000000003	0
ENSG00000000005	0.2423
ENSG00000000419	0.02193
ENSG00000000457	0.8184
ENSG00000000460	1.44833
ENSG00000000938	0.47008

For each Ensembl gene out of 51,520 Genecode annotated genes the FPKM value was computed. The FPKM value corresponds to the expression value of this gene. For RNA-Seq data, the relative expression of a transcript is proportional to the number of cDNA fragments that originated from it.

Note. This workflow may take several hours to complete. You can start this workflow and even switch off your computer, e.g. overnight, while the computation will be running on the server. After several hours you can check the results. In case of any questions, please feel free to ask for details (info@genexplain.com).

4.1.5. Find gene fusions from RNA-seq

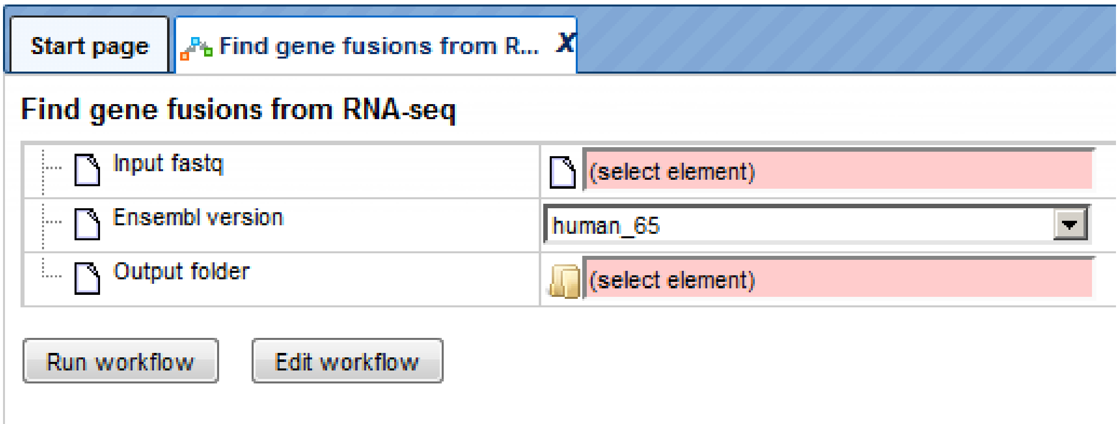
Recently, next-generation sequencing techniques at the transcriptome level (RNA-Seq) have been used to verify known and discover novel transcribed gene fusions. This workflow offers the ability to discover gene fusions from RNA-seq data (single-end (SE) or paired-end (PE) RNA-Seq read data) based on the fast FusionFinder program published in 2012 (Francis et al., PLoS ONE 7:e39987, 2012) . It accepts raw RNA-seq reads (fastq format) and produces a table with found gene fusions. The workflow can be found under the section "RNA-seq preprocessing".



RNA-seq
Load RNA-seq data
RNA-seq preprocessing
SRA to FASTQ
BAM to Counts
Convert genome coordinates with Lift-over
Alignment of FASTQ with Bowtie
Alignment of FASTQ with TopHat
Quantification of RNA-seq with Cufflinks for multiple BAM files
Find gene fusions from RNA-seq
Find genome variants and indels from RNA-seq
Quantification of RNA-seq with Cufflinks (no de-novo assembly) for FASTQ files
Quantification of RNA-seq with Cufflinks (with de-novo assembly) for FASTQ files

To launch the workflow, follow these steps:

Step 1. Open the workflow input form from the Start page. It will open in the main Work Space and looks as shown below:



Find gene fusions from RNA-seq

Input fastq	(select element)
Ensembl version	human_65
Output folder	(select element)

Step 2. Specify the input file in FASTQ format in the field **Input fastq**. It contains data from your RNA-seq study. To specify the input fastq file, you can drag & drop it from your project within the tree area. Alternatively, you may click on the pink field "select element" and a new window will open, where you select the input file. After having selected the file, press the [Ok] button.

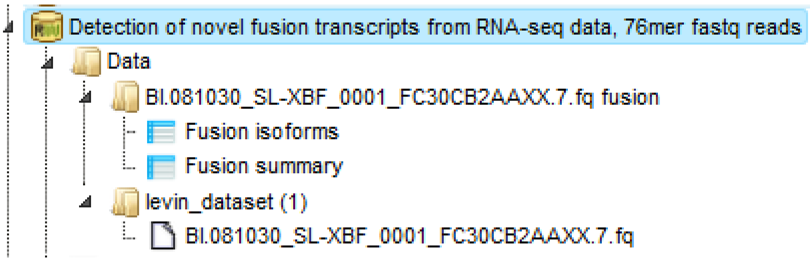
Step 3. Specify the **Ensembl version** from the drop-down menu. By default, the most recent, human_65, is selected.

Step 4. Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink field "select element" in the field **Output folder**, and a new window will be opened, where you can select the location of the results folder and define its name.

Step 5. Press the [Run workflow] button.

Wait until the workflow is completed.

The **Output folder** contains the two tables, *Fusion summary* and *Fusion isoforms*; for this example, let's consider the results folder located under Examples (data/Examples/Detection of novel fusion transcripts from RNA-seq data, 76mer fastq reads/Data/BI.081030_SL-XBF_0001_FC30CB2AAXX.7.fq (Gene fusions from RNA-seq)/). It is highlighted in blue in the figure below:

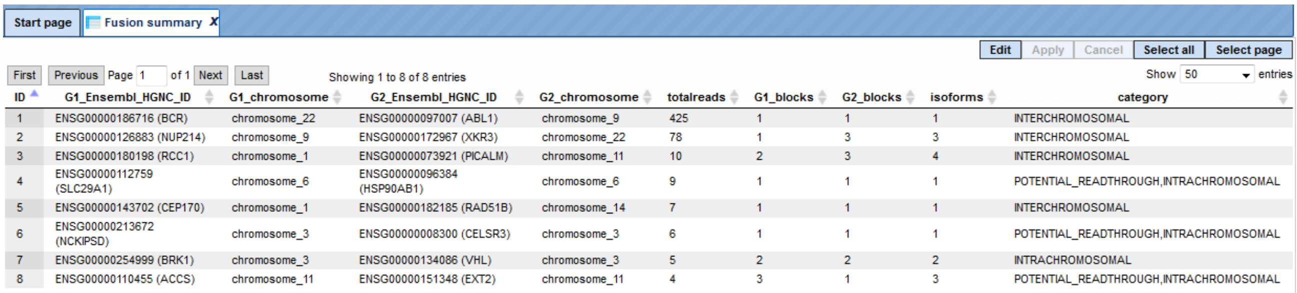


- Detection of novel fusion transcripts from RNA-seq data, 76mer fastq reads
  - Data
    - BI.081030_SL-XBF_0001_FC30CB2AAXX.7.fq fusion
      - Fusion isoforms
      - Fusion summary
    - levin_dataset (1)
      - BI.081030_SL-XBF_0001_FC30CB2AAXX.7.fq

The FusionFinder program analyses FASTQ read data (reads must be of at least 50 nucleotides long; see input example) to identify gene fusion candidates. This is achieved by performing an integrated analysis, which is illustrated in the original paper of Francis et al.

The first step is to align the full length reads against a normal coding reference transcriptome. After creation of pseudo paired-end reads (PE), these PE reads are aligned against the coding reference transcriptome. A further step is to analyze the results and filter false-positives. The last step consists of a block filtering and identification of fused exons and isoforms from candidate fusion transcripts.

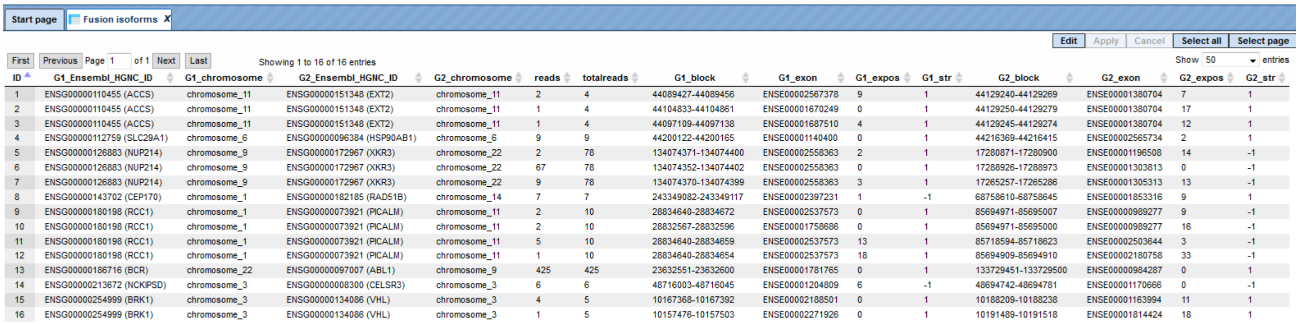
The output table **Fusion summary** is a ranked list of fusion candidates based on their evidence strength (total number of sequence reads = **total reads**). The file provides the Ensembl and HUGO (Human Genome Organization) Gene Nomenclature Committee (HGNC) common name identifiers for G1 and G2 (**G1_Ensembl_HGNC_ID** and **G2_Ensembl_HGNC_ID**), the number of blocks on each gene (**G1_blocks** and **G2_blocks**), an indication of how many **isoforms** exist for each G1:G2 pair and the **category** of fusion indicated by the pair.



Showing 1 to 8 of 8 entries

ID	G1_Ensembl_HGNC_ID	G1_chromosome	G2_Ensembl_HGNC_ID	G2_chromosome	totalreads	G1_blocks	G2_blocks	isoforms	category
1	ENSG00000186716 (BCR)	chromosome_22	ENSG00000097007 (ABL1)	chromosome_9	425	1	1	1	INTERCHROMOSOMAL
2	ENSG00000126883 (NUP214)	chromosome_9	ENSG00000172967 (XKR3)	chromosome_22	78	1	3	3	INTERCHROMOSOMAL
3	ENSG00000180198 (RCC1)	chromosome_1	ENSG00000073921 (PICALM)	chromosome_11	10	2	3	4	INTERCHROMOSOMAL
4	ENSG00000112759 (SLC29A1)	chromosome_6	ENSG00000096384 (HSP90AB1)	chromosome_6	9	1	1	1	POTENTIAL_READTHROUGH,INTRACHROMOSOMAL
5	ENSG00000143702 (CEP170)	chromosome_1	ENSG00000182185 (RAD51B)	chromosome_14	7	1	1	1	INTERCHROMOSOMAL
6	ENSG00000213672 (NCKIPSD)	chromosome_3	ENSG00000008300 (CELSR3)	chromosome_3	6	1	1	1	POTENTIAL_READTHROUGH,INTRACHROMOSOMAL
7	ENSG00000254999 (BRK1)	chromosome_3	ENSG00000134086 (VHL)	chromosome_3	5	2	2	2	INTRACHROMOSOMAL
8	ENSG00000110455 (ACCS)	chromosome_11	ENSG00000151348 (EXT2)	chromosome_11	4	3	1	3	POTENTIAL_READTHROUGH,INTRACHROMOSOMAL

The output table **Fusion isoforms** gives the full details for each isoform of G1 and G2 and includes the genomic coordinates of the alignment blocks on G1 and G2, and their respective corresponding Ensembl exon IDs.



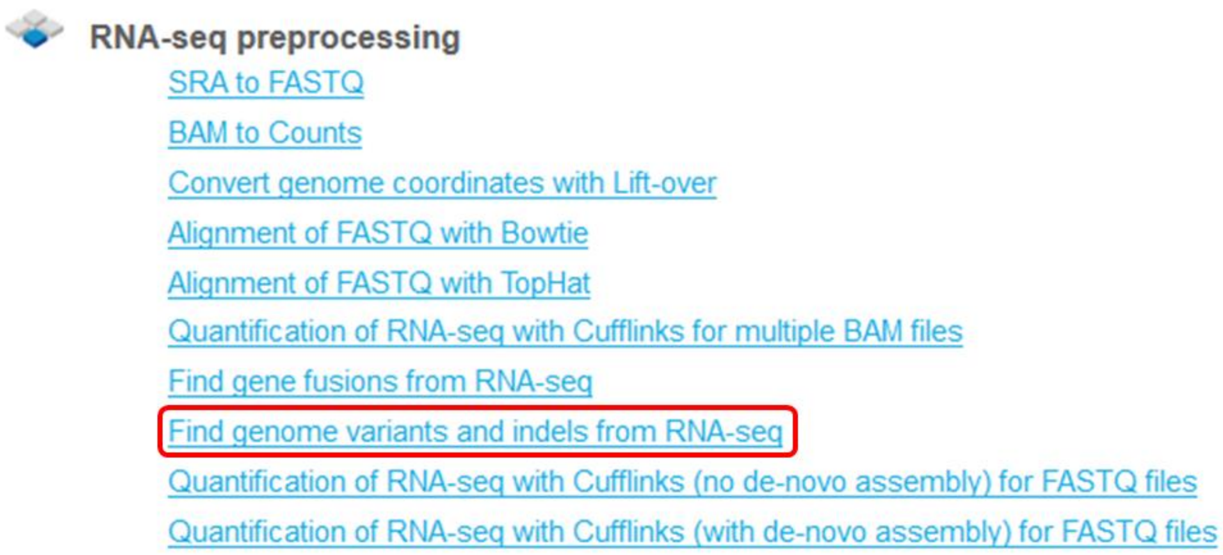
Showing 1 to 16 of 16 entries

ID	G1_Ensembl_HGNC_ID	G1_chromosome	G2_Ensembl_HGNC_ID	G2_chromosome	reads	totalreads	G1_block	G1_exon	G1_expos	G1_str	G2_block	G2_exon	G2_expos	G2_str
1	ENSG00000110455 (ACCS)	chromosome_11	ENSG00000151348 (EXT2)	chromosome_11	2	4	44089427-44089456	ENSE00002567378	9	1	44129240-44129269	ENSE00001380704	7	1
2	ENSG00000110455 (ACCS)	chromosome_11	ENSG00000151348 (EXT2)	chromosome_11	1	4	44104833-44104861	ENSE00001670249	0	1	44129250-44129279	ENSE00001380704	17	1
3	ENSG00000110455 (ACCS)	chromosome_11	ENSG00000151348 (EXT2)	chromosome_11	1	4	44097109-44097138	ENSE00001687510	4	1	44129245-44129274	ENSE00001380704	12	1
4	ENSG00000112759 (SLC29A1)	chromosome_6	ENSG00000096384 (HSP90AB1)	chromosome_6	9	9	44200122-44200165	ENSE00001140400	0	1	44216369-44216415	ENSE00002565734	2	1
5	ENSG00000126883 (NUP214)	chromosome_9	ENSG00000172967 (XKR3)	chromosome_22	2	78	134074371-134074400	ENSE00002558363	2	1	17280871-17280900	ENSE00001196508	14	-1
6	ENSG00000126883 (NUP214)	chromosome_9	ENSG00000172967 (XKR3)	chromosome_22	67	78	134074352-134074402	ENSE00002558363	0	1	17288926-17288973	ENSE00001303813	0	-1
7	ENSG00000126883 (NUP214)	chromosome_9	ENSG00000172967 (XKR3)	chromosome_22	9	78	134074370-134074399	ENSE00002558363	3	1	17265257-17265286	ENSE00001305313	13	-1
8	ENSG00000143702 (CEP170)	chromosome_1	ENSG00000182185 (RAD51B)	chromosome_14	7	7	243349082-243349117	ENSE00002397231	1	-1	68758610-68758645	ENSE00001853316	9	1
9	ENSG00000180198 (RCC1)	chromosome_1	ENSG00000073921 (PICALM)	chromosome_11	2	10	28834640-28834672	ENSE00002537573	0	1	85694971-85695007	ENSE00000989277	9	-1
10	ENSG00000180198 (RCC1)	chromosome_1	ENSG00000073921 (PICALM)	chromosome_11	2	10	28832567-28832596	ENSE00001758686	0	1	85694971-85695000	ENSE00000989277	16	-1
11	ENSG00000180198 (RCC1)	chromosome_1	ENSG00000073921 (PICALM)	chromosome_11	5	10	28834640-28834659	ENSE00002537573	13	1	85718594-85718623	ENSE00002503644	3	-1
12	ENSG00000180198 (RCC1)	chromosome_1	ENSG00000073921 (PICALM)	chromosome_11	1	10	28834640-28834654	ENSE00002537573	18	1	85694909-85694910	ENSE00002180758	33	-1
13	ENSG00000186716 (BCR)	chromosome_22	ENSG00000097007 (ABL1)	chromosome_9	425	425	23632551-23632600	ENSE00001781765	0	1	133729451-133729500	ENSE00000984287	0	1
14	ENSG00000213672 (NCKIPSD)	chromosome_3	ENSG00000008300 (CELSR3)	chromosome_3	6	6	48716003-48716045	ENSE00001204809	6	-1	48694742-48694781	ENSE00001170666	0	-1
15	ENSG00000254999 (BRK1)	chromosome_3	ENSG00000134086 (VHL)	chromosome_3	4	5	10167368-10167392	ENSE00002188501	0	1	10188209-10188238	ENSE00001163994	11	1
16	ENSG00000254999 (BRK1)	chromosome_3	ENSG00000134086 (VHL)	chromosome_3	1	5	10157476-10157503	ENSE00002271926	0	1	10191489-10191518	ENSE00001814424	18	1

4.1.6. Find genome variations and indels from RNA-seq

The challenge of obtaining accurate variant calls from RNA-seq data is substantial. The workflow is based on a framework to discover genotype variations published by De Pristo et al., Nature Genetics 43:491-498, 2011. The process applied includes initial read mapping, local realignment around indels, base quality score recalibration, SNP discovery and genotyping to find all potential variants.

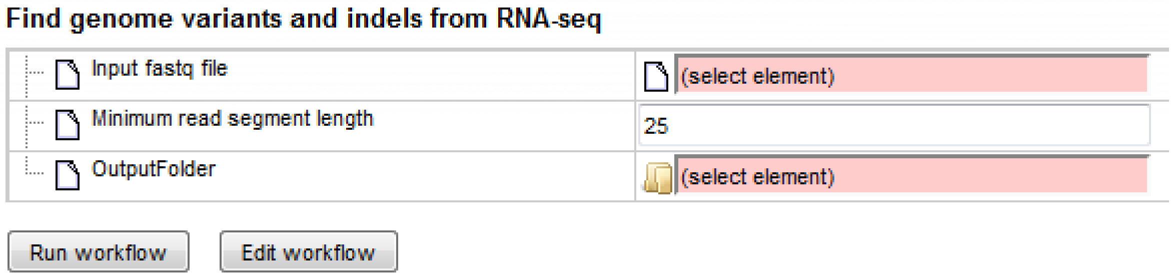
The workflow can be found in the section "RNA-seq Preprocessing".



RNA-seq preprocessing

- SRA to FASTQ
- BAM to Counts
- Convert genome coordinates with Lift-over
- Alignment of FASTQ with Bowtie
- Alignment of FASTQ with TopHat
- Quantification of RNA-seq with Cufflinks for multiple BAM files
- Find gene fusions from RNA-seq
- Find genome variants and indels from RNA-seq
- Quantification of RNA-seq with Cufflinks (no de-novo assembly) for FASTQ files
- Quantification of RNA-seq with Cufflinks (with de-novo assembly) for FASTQ files

Step 1. Open the workflow input form from the Start page. It will open in the main Work Space and looks as shown below:



Find genome variants and indels from RNA-seq

Input fastq file	(select element)
Minimum read segment length	25
OutputFolder	(select element)

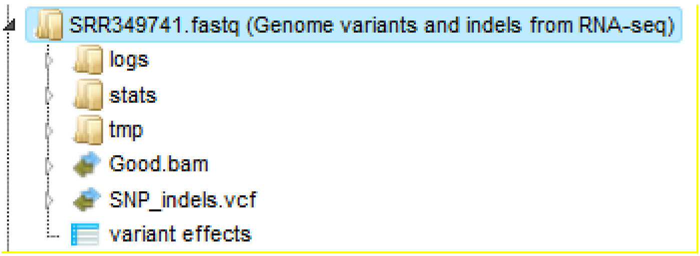
Step 2. Specify the input file in FASTQ format in the field **Input fastq file**. To specify the fastq file, you can drag & drop it from your project within the tree area. Alternatively, you may click on the pink field "select element" and a new window will open, where you select the input track. After having selected the track, press the [Ok] button.

Step 3. Specify the **Minimum read segment length**. By default a minimum length as 25 is given.

Step 4. Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink field "select element" in the field **OutputFolder**, and a new window will be opened, where you can select the location of the results folder and define its name.

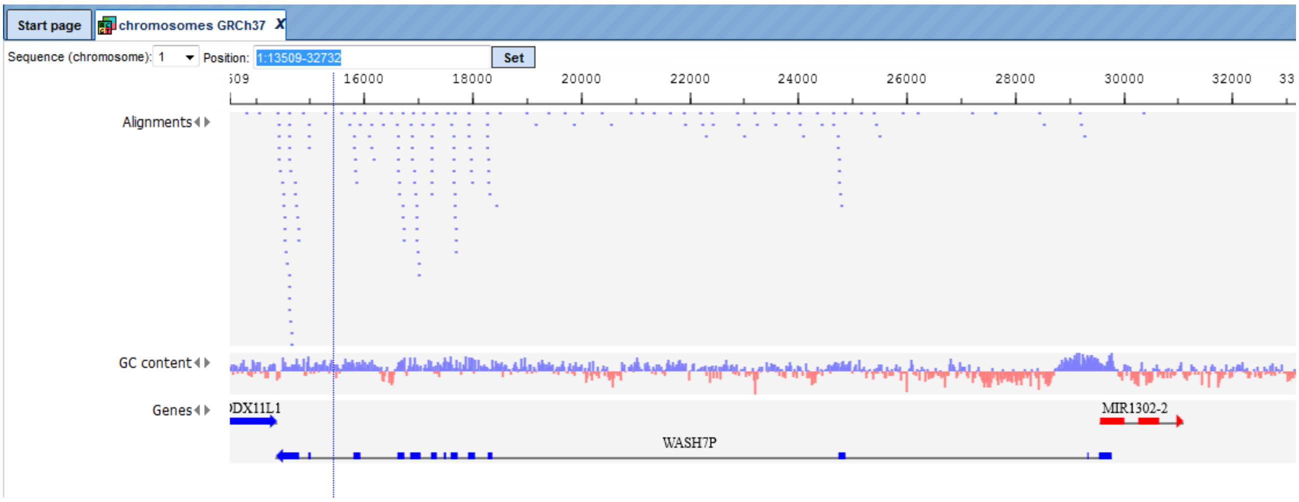
Start the workflow by pressing the [Run workflow] button.

In the following example we took as input the fastq file SRR349741.fastq (data/Examples/RNA-Seq analysis of human esophageal squamous cell carcinoma (ESCC), GSE32424, FASTQ files/Data/Fastq files/SRR349741.fastq). Below you can see the result folder (data/Examples/RNA-Seq analysis of human esophageal squamous cell carcinoma (ESCC), GSE32424, FASTQ files/Data/SRR349741.fastq (Genome variants and indels from RNA-seq)) for the example. The output folder contains several files and sub-folders with all results of the analysis.

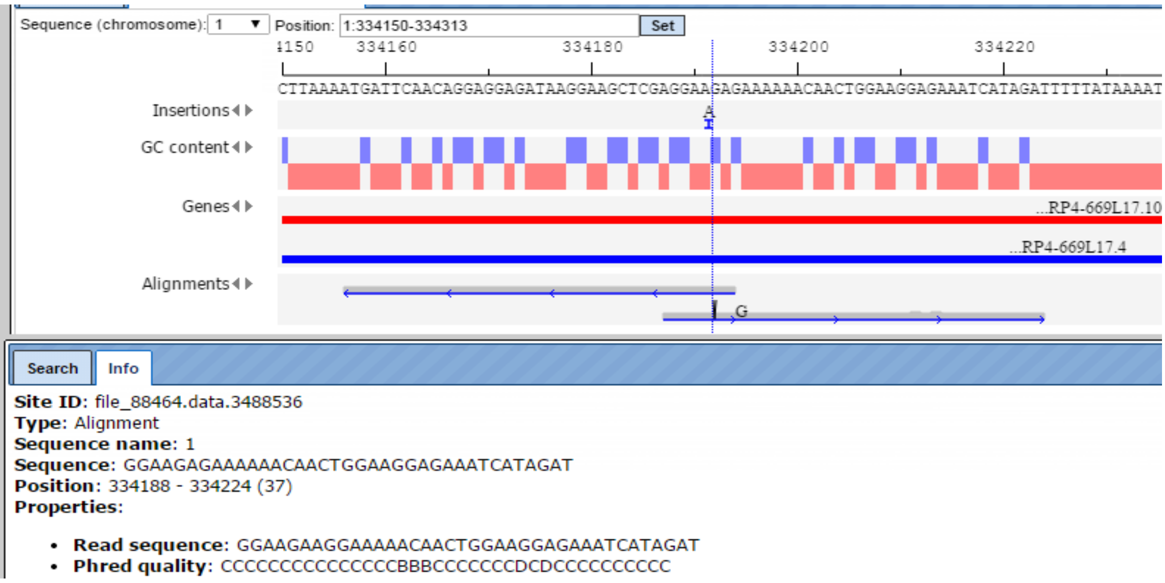


SRR349741.fastq (Genome variants and indels from RNA-seq)
- logs
- stats
- tmp
- Good.bam
- SNP_indels.vcf
- variant effects

The first step of the workflow is an alignment of all reads of fastq file using the TopHat2 tool. In the result folder one can see a sub-folder "tmp" which contains all found Deletions, Insertions, Splice junctions and Alignments. They are stored as tracks and can be opened in the genome browser by double-click on each of the tracks. Each short line (arrow in the higher zoom) represents an aligned "read" from the fastq file.



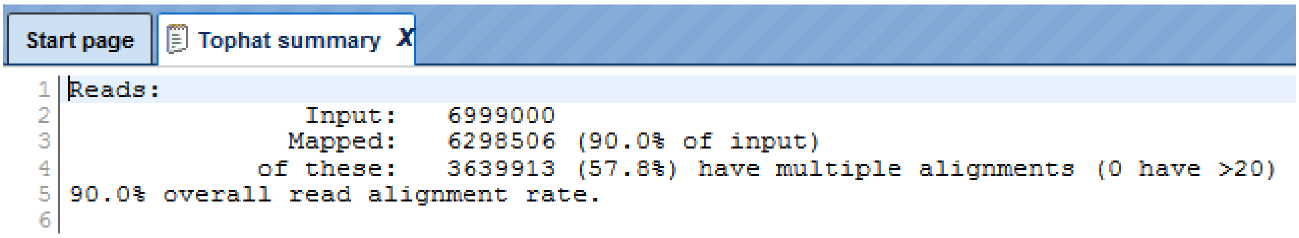
After zooming into each individual aligned read the insertions and deletions in the respective tracks of the browser become visible.



Site ID: file_88464.data.3488536
Type: Alignment
Sequence name: 1
Sequence: GGAAGAGAAAAAACAACTGGAAGGAGAAATCATAGAT
Position: 334188 - 334224 (37)
Properties:

- **Read sequence**: GGAAGAAGGAAAAACAACTGGAAGGAGAAATCATAGAT
- **Phred quality**: CCCCCCCCCCCCCCBBBCCCCCCCDCDCCCCCCCCCC

The *Tophat summary* file shows the total numbers of input reads, mapped reads, reads with multiple alignments and the overall read alignment rate.



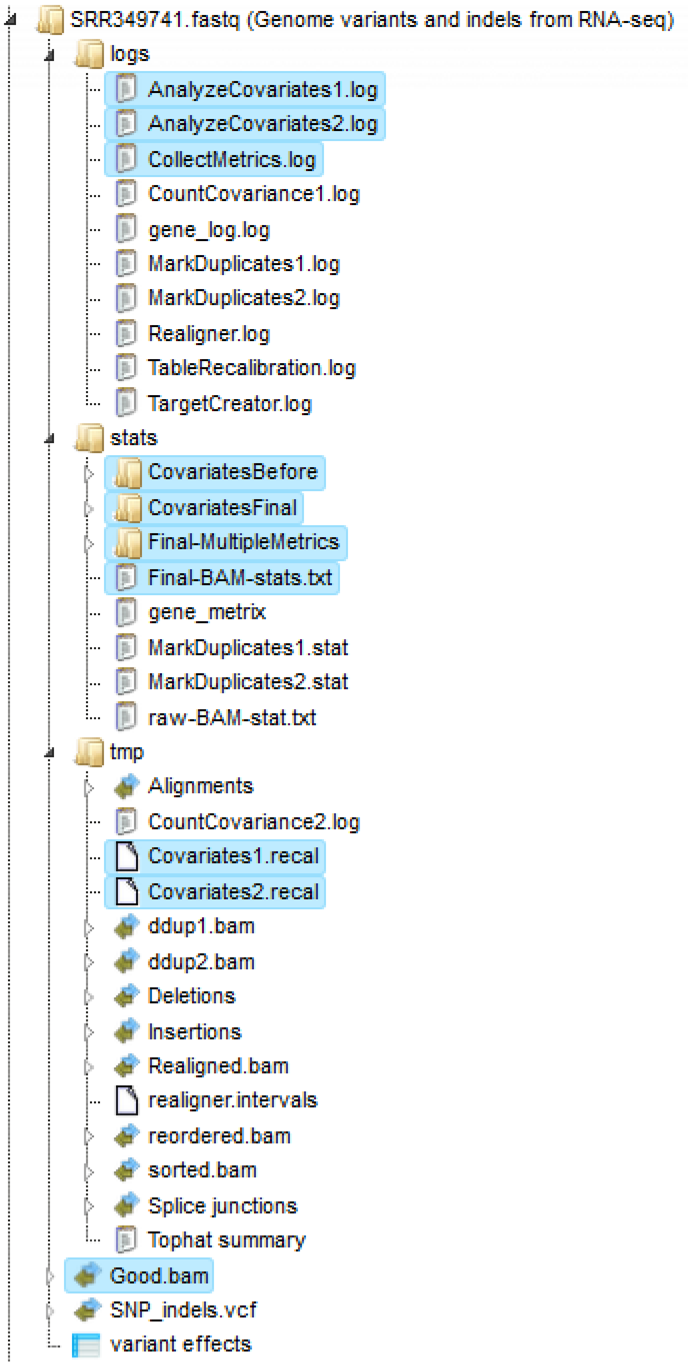
```
Reads:
                    Input:    6999000
                   Mapped:    6298506 (90.0% of input)
                of these:    3639913 (57.8%) have multiple alignments (0 have >20)
90.0% overall read alignment rate.

```

The initial alignments are sorted and reordered to prepare the next quality checking steps. The results of these two steps are stored in the folder tmp as the two files *reorder.bam* and *sorted.bam*.

The next step removes duplicates. The purpose is to mitigate the effects of PCR amplification bias introduced during library construction. Two read pairs are considered duplicate if they align to the same genomic position. The resulting *MarkDuplikates1.log* file is stored in the log folder and the *MarkDuplikates1.stat* file is stored in the stat folder.

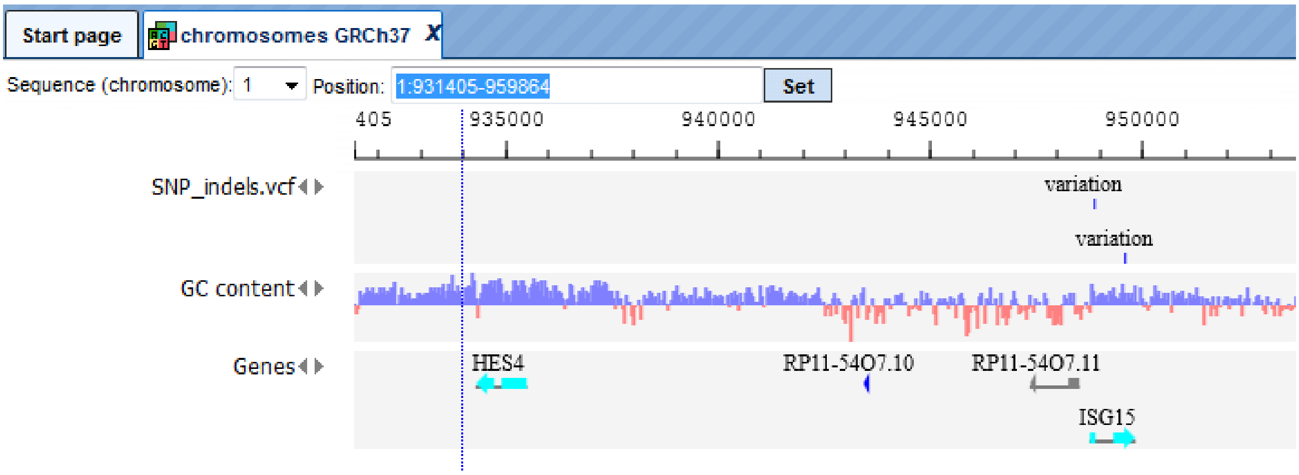
The next step is a local realignment. Read mapping algorithms operate on each read independently, locally realign reads such that the number of mismatching bases is minimized across all the reads. Output files are *Realigner.log* and *TargetCreator.log* in the log folder, *ddup1.bam*, *Realigned.bam* and *realigner.intervals* in the tmp folder.

The realigned BAM file is used again to remove duplicates (output *MarkDuplicates2.log* and *MarkDuplicates2.stat*), because realignment may change genomic positions of read pairs. After this step additional duplicates can be identified. The next step is a recalibration of base quality values. For each base in each read various covariates (such as reported quality score, position in read, dinucleotide, read GC-content) are calculated. Using these values the algorithm builds the model that predicts sequencing errors. Then it applies this model to calculate an empirical base quality score and overwrites the phred quality score currently in the read. Output is a new BAM file (*Good.bam*).



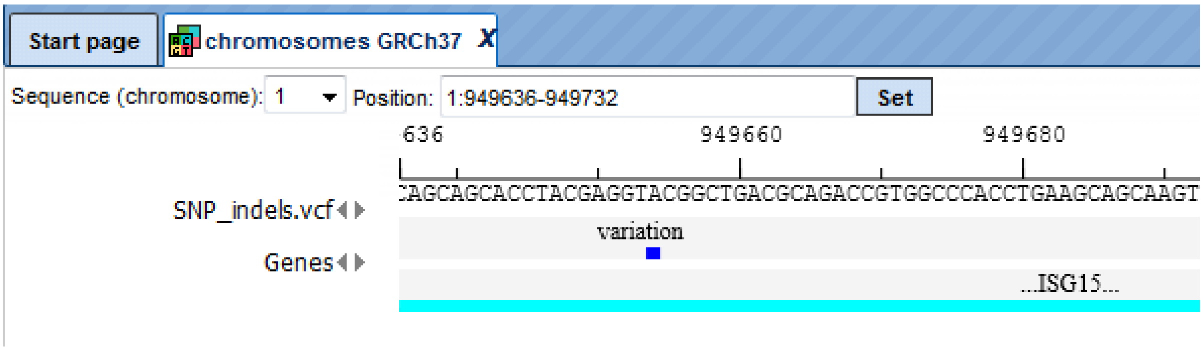
```
SRR349741.fastq (Genome variants and indels from RNA-seq)
    logs
        AnalyzeCovariates1.log
        AnalyzeCovariates2.log
        CollectMetrics.log
        CountCovariance1.log
        gene_log.log
        MarkDuplicates1.log
        MarkDuplicates2.log
        Realigner.log
        TableRecalibration.log
        TargetCreator.log
    stats
        CovariatesBefore
        CovariatesFinal
        Final-MultipleMetrics
        Final-BAM-stats.txt
        gene_metrix
        MarkDuplicates1.stat
        MarkDuplicates2.stat
        raw-BAM-stat.txt
    tmp
        Alignments
        CountCovariance2.log
        Covariates1.recal
        Covariates2.recal
        ddup1.bam
        ddup2.bam
        Deletions
        Insertions
        Realigned.bam
        realigner.intervals
        reordered.bam
        sorted.bam
        Splice junctions
        Tophat summary
    Good.bam
    SNP_indels.vcf
    variant effects
```

This file is used for the unified GATK (Genome Analysis Toolkit) genotyper method to detect the *SNP-indels* (table in VCF format) which the user can visualize by double click.



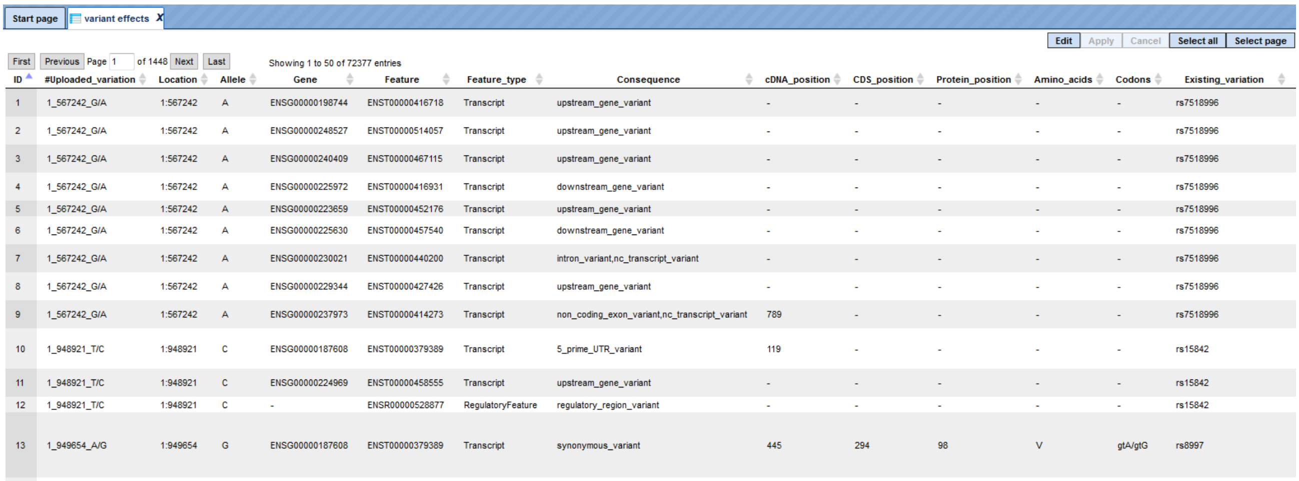
Sequence (chromosome): 1 Position: 1:931405-959864

After Zooming in information of variation on nucleotide basis is shown.



Sequence (chromosome): 1 Position: 1:949636-949732

In the next step each identified variation (SNP_indels) is analysed with the help of the "variant_effect_predictor" algorithm (http://genexplain-platform.com/bioumlweb/#de=analyses/Galaxy/ensembl/variant_effect_predictor). As a result it creates a final variant effects table that gives detailed information about each variation.



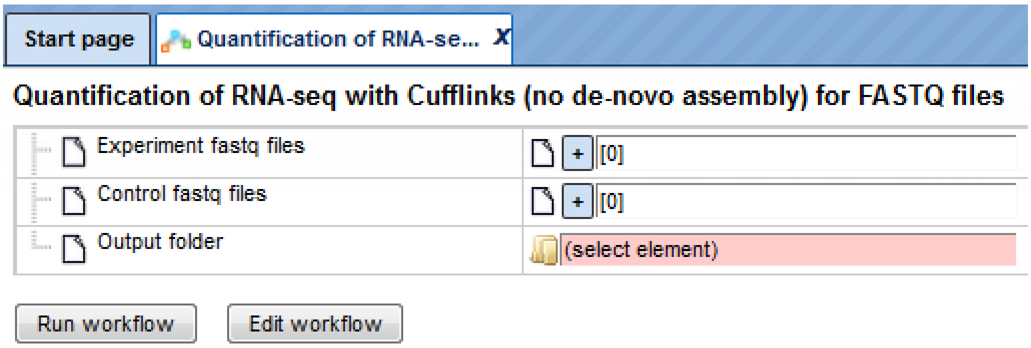
Showing 1 to 50 of 72377 entries

ID	#Uploaded_variation	Location	Allele	Gene	Feature	Feature_type	Consequence	cDNA_position	CDS_position	Protein_position	Amino_acids	Codons	Existing_variation
1	1_567242_G/A	1:567242	A	ENSG00000198744	ENST00000416718	Transcript	upstream_gene_variant	-	-	-	-	-	rs7518996
2	1_567242_G/A	1:567242	A	ENSG00000248527	ENST00000514057	Transcript	upstream_gene_variant	-	-	-	-	-	rs7518996
3	1_567242_G/A	1:567242	A	ENSG00000240409	ENST00000467115	Transcript	upstream_gene_variant	-	-	-	-	-	rs7518996
4	1_567242_G/A	1:567242	A	ENSG00000225972	ENST00000416931	Transcript	downstream_gene_variant	-	-	-	-	-	rs7518996
5	1_567242_G/A	1:567242	A	ENSG00000223659	ENST00000452176	Transcript	upstream_gene_variant	-	-	-	-	-	rs7518996
6	1_567242_G/A	1:567242	A	ENSG00000225630	ENST00000457540	Transcript	downstream_gene_variant	-	-	-	-	-	rs7518996
7	1_567242_G/A	1:567242	A	ENSG00000230021	ENST00000440200	Transcript	intron_variant,nc_transcript_variant	-	-	-	-	-	rs7518996
8	1_567242_G/A	1:567242	A	ENSG00000229344	ENST00000427426	Transcript	upstream_gene_variant	-	-	-	-	-	rs7518996
9	1_567242_G/A	1:567242	A	ENSG00000237973	ENST00000414273	Transcript	non_coding_exon_variant,nc_transcript_variant	789	-	-	-	-	rs7518996
10	1_948921_T/C	1:948921	C	ENSG00000187608	ENST00000379389	Transcript	5_prime_UTR_variant	119	-	-	-	-	rs15842
11	1_948921_T/C	1:948921	C	ENSG00000224969	ENST00000458555	Transcript	upstream_gene_variant	-	-	-	-	-	rs15842
12	1_948921_T/C	1:948921	C	-	ENSR00000528877	RegulatoryFeature	regulatory_region_variant	-	-	-	-	-	rs15842
13	1_949654_A/G	1:949654	G	ENSG00000187608	ENST00000379389	Transcript	synonymous_variant	445	294	98	V	gtA/gtG	rs8997

4.1.7. Quantification of RNA-seq with Cufflinks (no de-novo assembly) for FASTQ files

This workflow offers a possibility to discover new genes and transcripts (splice variants) and measure transcript expression in a single assay from RNA-seq data. This workflow is described in “Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks”, Nat. Protoc. 7:562-578, 2012.

Step 1. Open the workflow input form from the Start page. It will open in the main Work Space and looks as shown below:



Quantification of RNA-seq with Cufflinks (no de-novo assembly) for FASTQ files

Experiment fastq files	+ [0]
Control fastq files	+ [0]
Output folder	(select element)

Step 2. Specify the **Experiment fastq files** and the **Control fastq files**. To specify the input files in Sanger FASTQ format, you can drag & drop it from your project within the tree area.

Step 3. Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink field “select element” in the field **Output folder**, and a new window will be opened, where you can select the location of the results folder and define its name.

Start the workflow by pressing the [Run workflow] button.

All results are saved in the result folder:

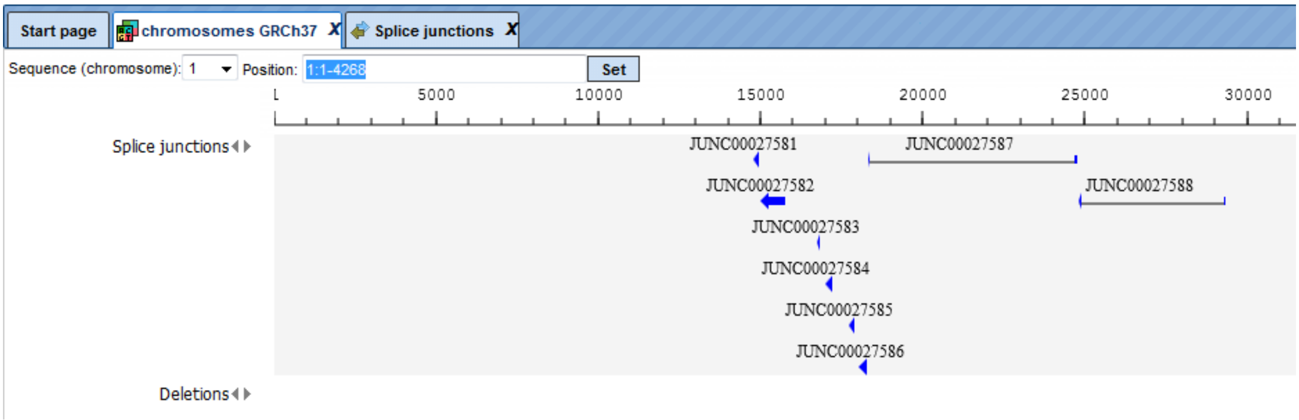
data/Examples/RNA-Seq analysis of human esophageal squamous cell carcinoma (ESCC), GSE32424, FASTQ files/Data/Fastq files (Quantification of RNA-seq (no de-novo assembly))

Read alignment with TopHat

The first step of the workflow is alignment of sequence reads with TopHat (http://tophat.cbcb.umd.edu/). TopHat aligns reads to the genome and discovers transcript splice sites. TopHat uses Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) as an alignment ‘engine’ and breaks up reads that Bowtie cannot align on its own into smaller pieces called segments.

Output files are tables and tracks with insertions, deletions, splice junctions and the alignments.

Example output of splice junctions opened as track:

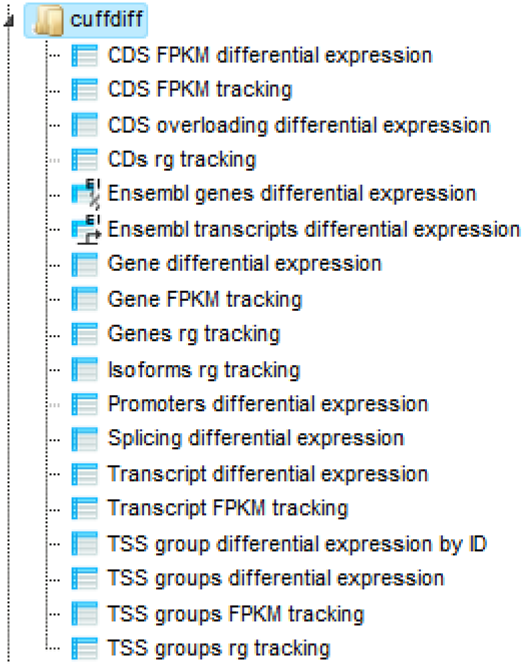


Mismatches, insertions and deletions in the alignments can identify polymorphisms between the sequenced sample and the reference genome, or even pinpoint gene fusion events in tumor samples. Reads that align outside annotated genes are often strong evidence of new protein-coding genes and noncoding RNAs. RNA-seq read alignments can reveal new alternative splicing events and isoforms. Alignments can also be used to accurately quantify gene and transcript expression, because the number of reads produced by a transcript is proportional to its abundance.

Differential analysis with Cuffdiff

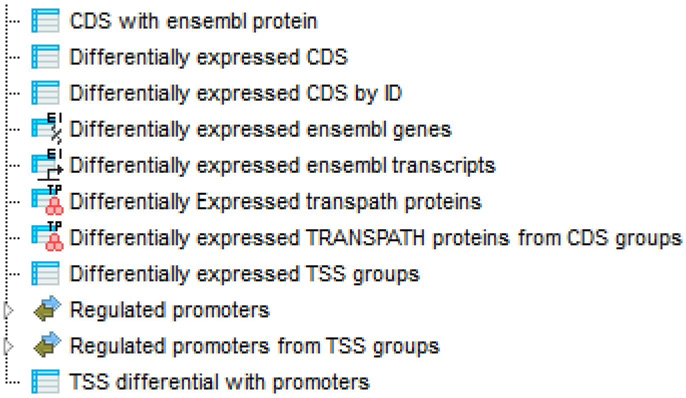
The second step of the workflow is performed by Cuffdiff, part of the Cufflinks package, which calculates expression in two or more samples and tests the statistical significance of each observed change in expression between them. Cuffdiff allows for supplying multiple technical or biological replicate sequencing libraries per condition. With multiple replicates, Cuffdiff learns how read counts vary for each gene across the replicates and uses these variance estimates to calculate the significance of observed changes in expression.

Cuffdiff reports 15 output files:



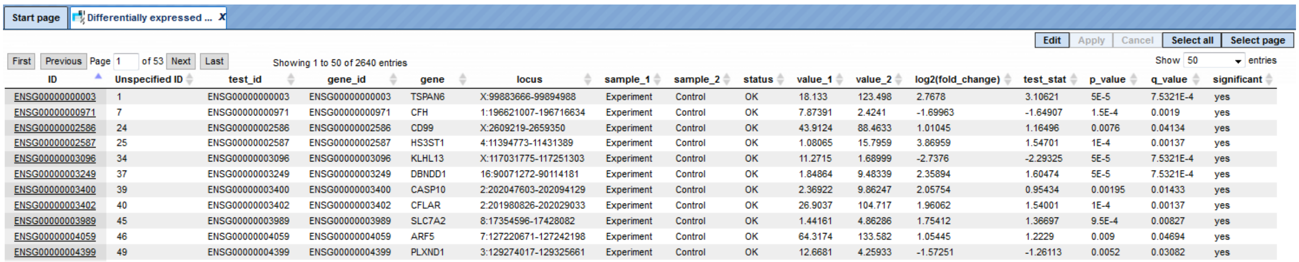
cuffdiff
- CDS FPKM differential expression
- CDS FPKM tracking
- CDS overloading differential expression
- CDs rg tracking
- Ensembl genes differential expression
- Ensembl transcripts differential expression
- Gene differential expression
- Gene FPKM tracking
- Genes rg tracking
- Isoforms rg tracking
- Promoters differential expression
- Splicing differential expression
- Transcript differential expression
- Transcript FPKM tracking
- TSS group differential expression by ID
- TSS groups differential expression
- TSS groups FPKM tracking
- TSS groups rg tracking

The last step of the workflow comprises several conversions and filtering steps of some Cuffdiff output files. The final table and tracks are in the result folder:



- CDS with ensembl protein
- Differentially expressed CDS
- Differentially expressed CDS by ID
- Differentially expressed ensembl genes
- Differentially expressed ensembl transcripts
- Differentially Expressed transpath proteins
- Differentially expressed TRANSPATH proteins from CDS groups
- Differentially expressed TSS groups
- Regulated promoters
- Regulated promoters from TSS groups
- TSS differential with promoters

The table **Differentially expressed Ensembl genes** contains all identified differentially expressed genes (Ensembl IDs) (also converted into a table of transcripts and a table of TRANSPATH® proteins).



Showing 1 to 50 of 2640 entries

Show 50 entries

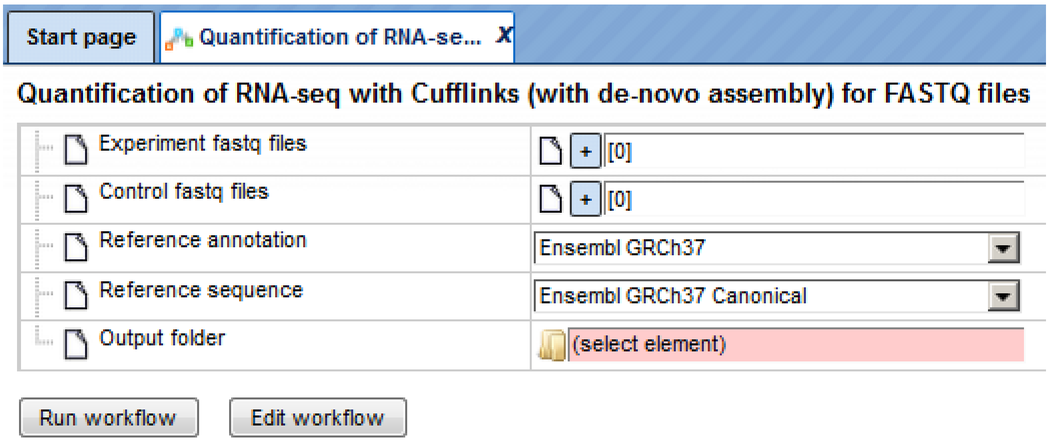
ID	Unspecified ID	test_id	gene_id	gene	locus	sample_1	sample_2	status	value_1	value_2	log2(fold_change)	test_stat	p_value	q_value	significant
ENSG00000000003	1	ENSG00000000003	ENSG00000000003	TSPAN6	X:99883666-99894988	Experiment	Control	OK	18.133	123.498	2.7678	3.10621	5E-5	7.5321E-4	yes
ENSG00000000971	7	ENSG00000000971	ENSG00000000971	CFH	1:196621007-196716634	Experiment	Control	OK	7.87391	2.4241	-1.69963	-1.64907	1.5E-4	0.0019	yes
ENSG00000002586	24	ENSG00000002586	ENSG00000002586	CD99	X:2609219-2659350	Experiment	Control	OK	43.9124	88.4633	1.01045	1.16496	0.0076	0.04134	yes
ENSG00000002587	25	ENSG00000002587	ENSG00000002587	HS3ST1	4:11394773-11431389	Experiment	Control	OK	1.08065	15.7959	3.86959	1.54701	1E-4	0.00137	yes
ENSG00000003096	34	ENSG00000003096	ENSG00000003096	KLHL13	X:117031775-117251303	Experiment	Control	OK	11.2715	1.68999	-2.7376	-2.29325	5E-5	7.5321E-4	yes
ENSG00000003249	37	ENSG00000003249	ENSG00000003249	DBNDD1	16:90071272-90114181	Experiment	Control	OK	1.84864	9.48339	2.35894	1.60474	5E-5	7.5321E-4	yes
ENSG00000003400	39	ENSG00000003400	ENSG00000003400	CASP10	2:202047603-202094129	Experiment	Control	OK	2.36922	9.86247	2.05754	0.95434	0.00195	0.01433	yes
ENSG00000003402	40	ENSG00000003402	ENSG00000003402	CFLAR	2:201980826-202029033	Experiment	Control	OK	26.9037	104.717	1.96062	1.54001	1E-4	0.00137	yes
ENSG00000003989	45	ENSG00000003989	ENSG00000003989	SLC7A2	8:17354596-17428082	Experiment	Control	OK	1.44161	4.86286	1.75412	1.36697	9.5E-4	0.00827	yes
ENSG00000004059	46	ENSG00000004059	ENSG00000004059	ARF5	7:127220671-127242198	Experiment	Control	OK	64.3174	133.582	1.05445	1.2229	0.009	0.04694	yes
ENSG00000004399	49	ENSG00000004399	ENSG00000004399	PLXND1	3:129274017-129325661	Experiment	Control	OK	12.6681	4.25933	-1.57251	-1.26113	0.0052	0.03082	yes

The **Regulated promoters** are extracted from the table Differentially expressed Ensembl transcripts. Transcripts with the same transcription start site (TSS) are merged into the single 'TSS group' and **Differentially expressed TSS groups** are also identified. **Regulated promoters from TSS groups** are identified using those TSS groups. Similarly, **Differentially expressed CDS groups** are identified (CDS group is the group of transcripts with the same coding sequence; they produce exactly the same protein). And CDS group table is used to find **Differentially expressed TRANSPATH proteins from CDS groups**.

4.1.8. Quantification of RNA-seq with Cufflinks (with de-novo assembly) for FASTQ files

This workflow offers the ability to discover new genes and transcripts (splice variants) and measure transcript expression in a single assay from RNA-seq data. This workflow is described in "Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks", Nat. Protoc. 7:562-578, 2012.

Step 1. Open the workflow input form from the Start page. It will open in the main Work Space and looks as shown below:



Quantification of RNA-seq with Cufflinks (with de-novo assembly) for FASTQ files

Experiment fastq files	[0]
Control fastq files	[0]
Reference annotation	Ensembl GRCh37
Reference sequence	Ensembl GRCh37 Canonical
Output folder	(select element)

Step 2. Specify the **Experiment fastq files** and the **Control fastq files**. To specify the input files in Sanger FASTQ format, you can drag & drop it from your project within the tree area.

Step 3. Specify the **Reference annotation**.

Step 4. Specify the **Reference sequence**.

Step 5. Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink field "select element" in the field **Output folder**, and a new window will be opened, where you can select the location of the results folder and define its name.

Start the workflow by pressing the [Run workflow] button.

All results are saved in the result folder:

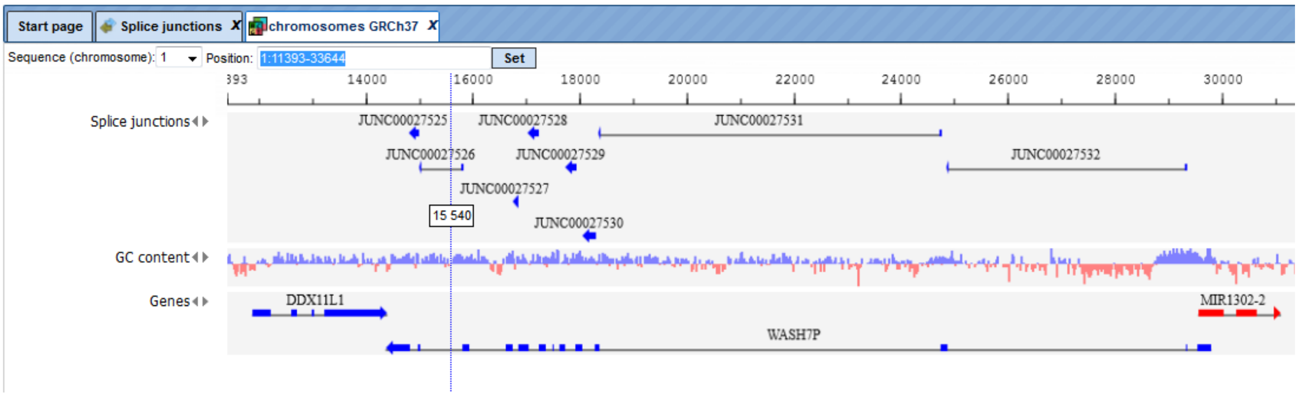
data/Examples/RNA-Seq analysis of human esophageal squamous cell carcinoma (ESCC), GSE32424, FASTQ files/Data/Fastq files (Quantification of RNA-seq (with de novo assembly))

Read alignment with TopHat

The first step of the workflow is a read alignment with TopHat (http://tophat.cbcb.umd.edu/). TopHat aligns reads to the genome and discovers transcript splice sites. TopHat uses Bowtie (http://bowtie-bio.sourceforge.net/index.shtml) as an alignment 'engine' and breaks up reads that Bowtie cannot align on its own into smaller pieces called segments.

Output files are tables and tracks with insertions, deletions, splice junctions and the alignments.

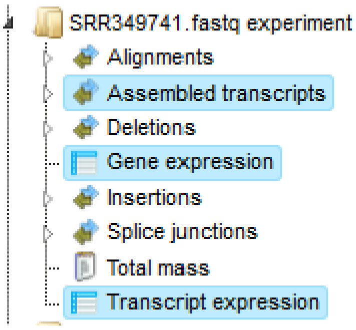
Example output of splice junctions opened as track:



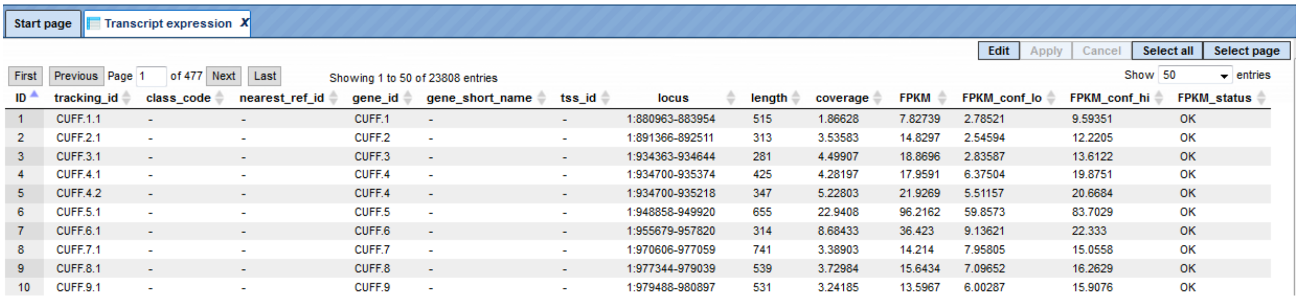
Mismatches, insertions and deletions in the alignments can identify polymorphisms between the sequenced sample and the reference genome, or even pinpoint gene fusion events in tumor samples. Reads that align outside annotated genes are often strong evidence of new protein-coding genes and noncoding RNAs. RNA-seq read alignments can reveal new alternative splicing events and isoforms. Alignments can also be used to accurately quantify gene and transcript expression, because the number of reads produced by a transcript is proportional to its abundance.

Transcript assembly with Cufflinks

Cufflinks uses the alignments to map reads against the genome and to assemble the reads into transcripts. Cufflinks assembles individual transcripts from RNA-seq reads that have been aligned to the genome. Because a sample may contain reads from multiple splice variants for a given gene, Cufflinks must be able to infer the splicing structure of each gene. Thus, Cufflinks reports a parsimonious transcriptome assembly of the data. The algorithm reports as few full-length transcript fragments or 'transfrags' as are needed to 'explain' all the splicing event outcomes in the input data. Output tracks of Cufflinks is the **Assembled transcripts** track, output tables of Cufflinks are **Gene expression** and **Transcript expression** tables.


SRR349741.fastq experiment
- Alignments
- Assembled transcripts
- Deletions
- Gene expression
- Insertions
- Splice junctions
- Total mass
- Transcript expression

This step distinguishes this workflow from the workflow called "**Quantification of RNA-seq with Cufflinks (no de-novo assembly) for FASTQ files".** In the current workflow the transcripts are assembled "de-novo", whereas in that other workflow the transcripts are taken from the reference Ensembl transcript annotation. Since here it is a "de-novo" reconstruction of exon-intron structure, no known gene or transcript names are given. All transcripts are defined by the tracking_id, like Cuff.1.1 and so on. This allows us to find new transcripts that were not yet discovered and annotated in the reference genome.



Showing 1 to 50 of 23808 entries

ID	tracking_id	class_code	nearest_ref_id	gene_id	gene_short_name	tss_id	locus	length	coverage	FPKM	FPKM_conf_lo	FPKM_conf_hi	FPKM_status
1	CUFF.1.1	-	-	CUFF.1	-	-	1:880963-883954	515	1.86628	7.82739	2.78521	9.59351	OK
2	CUFF.2.1	-	-	CUFF.2	-	-	1:891366-892511	313	3.53583	14.8297	2.54594	12.2205	OK
3	CUFF.3.1	-	-	CUFF.3	-	-	1:934363-934644	281	4.49907	18.8696	2.83587	13.6122	OK
4	CUFF.4.1	-	-	CUFF.4	-	-	1:934700-935374	425	4.28197	17.9591	6.37504	19.8751	OK
5	CUFF.4.2	-	-	CUFF.4	-	-	1:934700-935218	347	5.22803	21.9269	5.51157	20.6684	OK
6	CUFF.5.1	-	-	CUFF.5	-	-	1:948858-949920	655	22.9408	96.2162	59.8573	83.7029	OK
7	CUFF.6.1	-	-	CUFF.6	-	-	1:955679-957820	314	8.68433	36.423	9.13621	22.333	OK
8	CUFF.7.1	-	-	CUFF.7	-	-	1:970606-977059	741	3.38903	14.214	7.95805	15.0558	OK
9	CUFF.8.1	-	-	CUFF.8	-	-	1:977344-979039	539	3.72984	15.6434	7.09652	16.2629	OK
10	CUFF.9.1	-	-	CUFF.9	-	-	1:979488-980897	531	3.24185	13.5967	6.00287	15.9076	OK

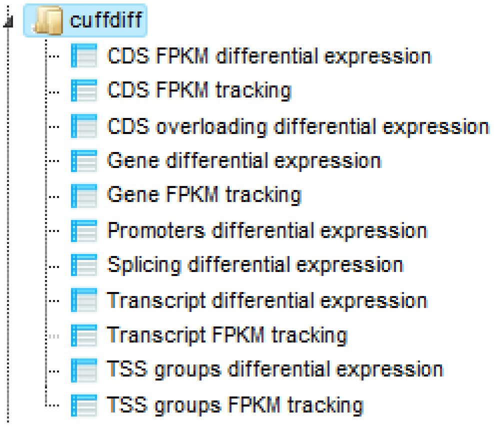
Assembled transcripts merging with Cuffmerge

When you are working with several RNA-seq samples, it becomes necessary to pool the data and assemble them into a comprehensive set of transcripts before proceeding to differential analysis. Cuffmerge, part of the Cufflinks package (http://cufflinks.cbcb.umd.edu/) is essentially a 'meta-assembler' — it treats the assembled transfrags the way Cufflinks treats reads, merging them together parsimoniously. Output is a **Merged assembly** track.

Differential analysis with Cuffdiff

The next step of the workflow is performed by Cuffdiff, part of the Cufflinks package (http://cufflinks.cbcb.umd.edu/), which calculates expression in two or more samples and tests the statistical significance of each observed change in expression between them. Cuffdiff allows supplying multiple technical or biological replicate sequencing libraries per condition. With multiple replicates, Cuffdiff learns how read counts vary for each gene across the replicates and uses these variance estimates to calculate the significance of observed changes in expression.

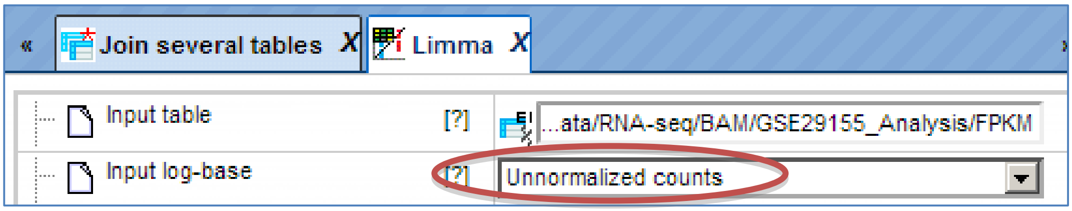
Cuffdiff reports the following 11 output files:



- cuffdiff
  - CDS FPKM differential expression
  - CDS FPKM tracking
  - CDS overloading differential expression
  - Gene differential expression
  - Gene FPKM tracking
  - Promoters differential expression
  - Splicing differential expression
  - Transcript differential expression
  - Transcript FPKM tracking
  - TSS groups differential expression
  - TSS groups FPKM tracking

4.2. Detect differentially expressed gene (DEG)

In order to perform further analyses of the results of the workflow **Quantification of RNA-seq with Cufflinks for multiple BAM files** it is recommended to join all resulting gene tables into one table using the function "Join several tables" of the platform. The joint table can be used for detection of differentially expressed genes using the Limma or EBarrays functions.



Input table	[?]	...ata/RNA-seq/BAM/GSE29155_Analysis/FPKM
Input log-base	[?]	Unnormalized counts

It should be noted here that to perform a Limma or EBarray analysis of the RNA-Seq data you should select the option *Unnormalized counts* in the input field **Input log-base** for either of these two methods, as shown below for Limma.

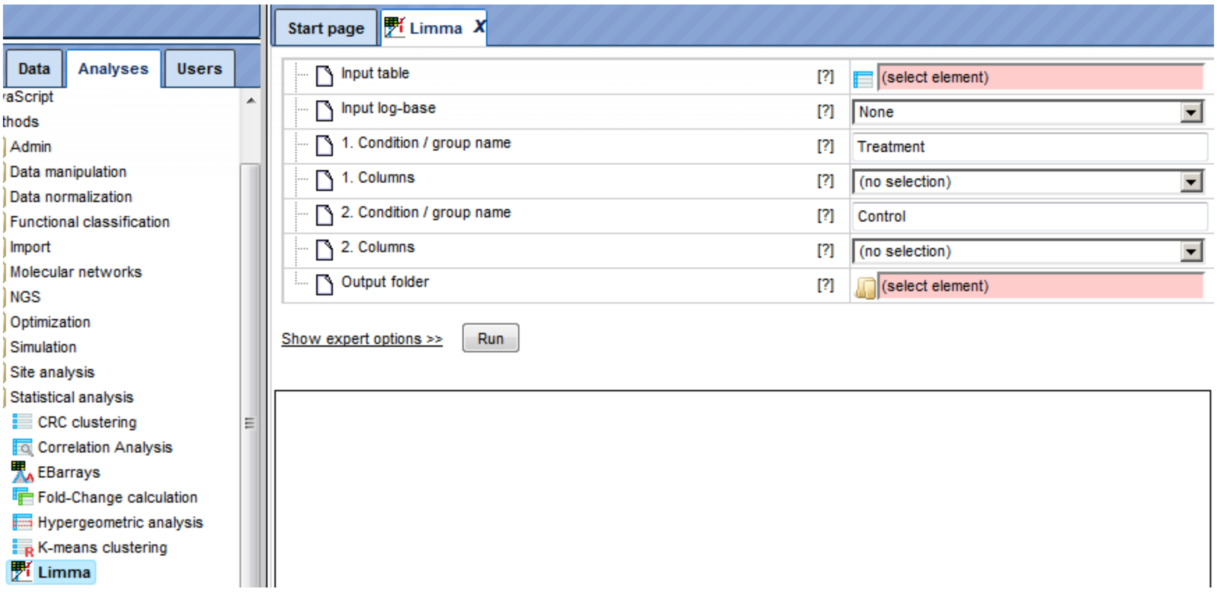
4.2.1. Estimate differential expression using Linear Models for MicroArrays (LIMMA)

Limma estimates differential expression between specified conditions / groups.

This tool provides an interface for the popular and comprehensive Limma package. The platform tool computes differential expression between up to five conditions / groups. The groups consist of columns of a data table that contains normalized measurement values, e.g. from a normalized microarray experiment. Furthermore, one can estimate differential expression for normalized or un-normalized count data as derived from RNA-seq experiments.

All possible contrasts between groups are considered and their output is stored in a common folder. Conditions are compared in the specified order from first to fifth. E.g. given conditions named A, B and C, the output will contain the contrasts AvsB, AvsC and BvsC.

It is necessary to provide a unique name for each group. Also, at least two data columns are required per group.



Input table	[?]	(select element)
Input log-base	[?]	None
1. Condition / group name	[?]	Treatment
1. Columns	[?]	(no selection)
2. Condition / group name	[?]	Control
2. Columns	[?]	(no selection)
Output folder	[?]	(select element)

The input parameters for Limma are described in the following.

Input table: This table contains the columns to analyze.

Input log-base: Here you can specify the scale of the input data. If the log-base is log2, the tool will use the data values as is. If your data are from RNA-seq, you can select *Normalized counts* or *Unnormalized counts*.

1-5. Condition / group name: One can specify up to five groups of columns. Please note that unnamed groups are not considered; a name is not assigned automatically.

1-5. Columns: These fields contain the selected columns. Please note that column selections are not considered without a corresponding name. Columns can only be specified once and there need to be two columns per group.

Output folder: The output folder will contain one output files for each pair of conditions.

An example output table is shown at the end of this section. Its columns are explained in the following. Those highlighted in bold are shown in the default view. The other columns can be included on demand via the Columns tab of the lower right panel (available with opened output table).

logFC: Fold change (log)

CI.025: Fold change (Lower confidence interval)

CI.975: Fold change (Upper confidence interval)

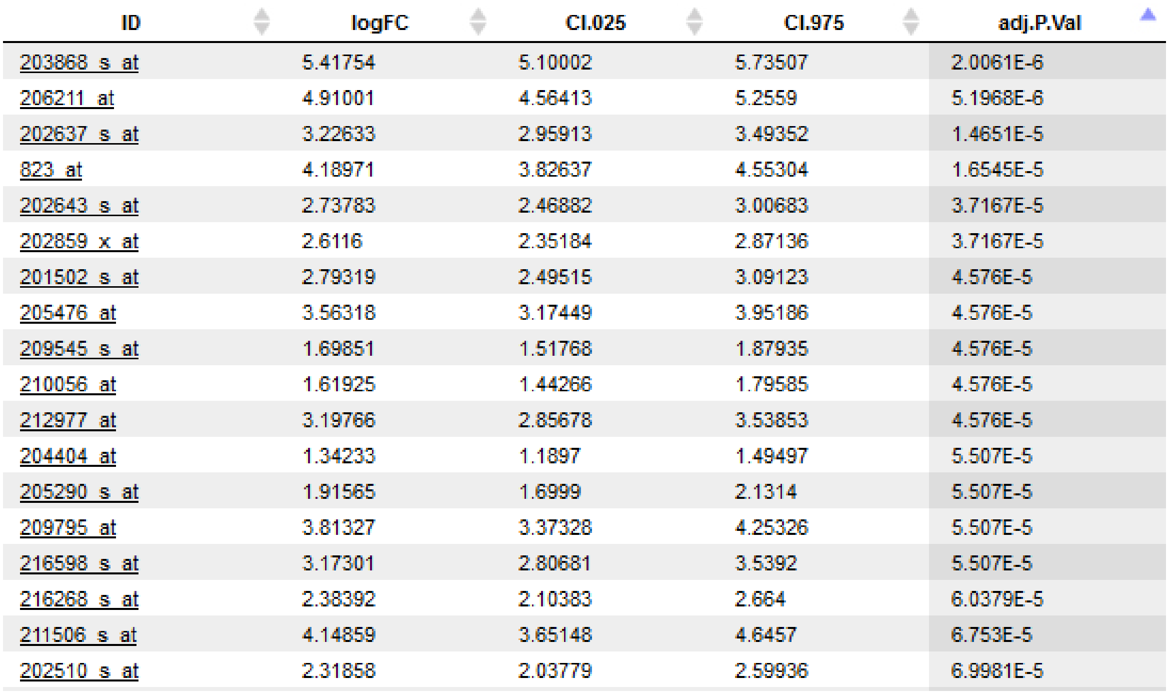
AveExpr: Average log2-expression for the probe over all arrays

t: Moderated T-statistic

P.Value: P-value Differential expression

adj.P.Val: Adjusted P-value (Benjamini-Hochberg)

B: Log-odds that the gene / probe presents differential expression



ID	logFC	CI.025	CI.975	adj.P.Val
203868_s_at	5.41754	5.10002	5.73507	2.0061E-6
206211_at	4.91001	4.56413	5.2559	5.1968E-6
202637_s_at	3.22633	2.95913	3.49352	1.4651E-5
823_at	4.18971	3.82637	4.55304	1.6545E-5
202643_s_at	2.73783	2.46882	3.00683	3.7167E-5
202859_x_at	2.6116	2.35184	2.87136	3.7167E-5
201502_s_at	2.79319	2.49515	3.09123	4.576E-5
205476_at	3.56318	3.17449	3.95186	4.576E-5
209545_s_at	1.69851	1.51768	1.87935	4.576E-5
210056_at	1.61925	1.44266	1.79585	4.576E-5
212977_at	3.19766	2.85678	3.53853	4.576E-5
204404_at	1.34233	1.1897	1.49497	5.507E-5
205290_s_at	1.91565	1.6999	2.1314	5.507E-5
209795_at	3.81327	3.37328	4.25326	5.507E-5
216598_s_at	3.17301	2.80681	3.5392	5.507E-5
216268_s_at	2.38392	2.10383	2.664	6.0379E-5
211506_s_at	4.14859	3.65148	4.6457	6.753E-5
202510_s_at	2.31858	2.03779	2.59936	6.9981E-5

Reference:

Smyth, G. K. (2005). Limma: linear models for microarray data. In: Bioinformatics and

Computational Biology Solutions using R and Bioconductor. R. Gentleman, V. Carey, S. Dudoit, R. Irizarry, W. Huber (eds), Springer, New York, 2005.

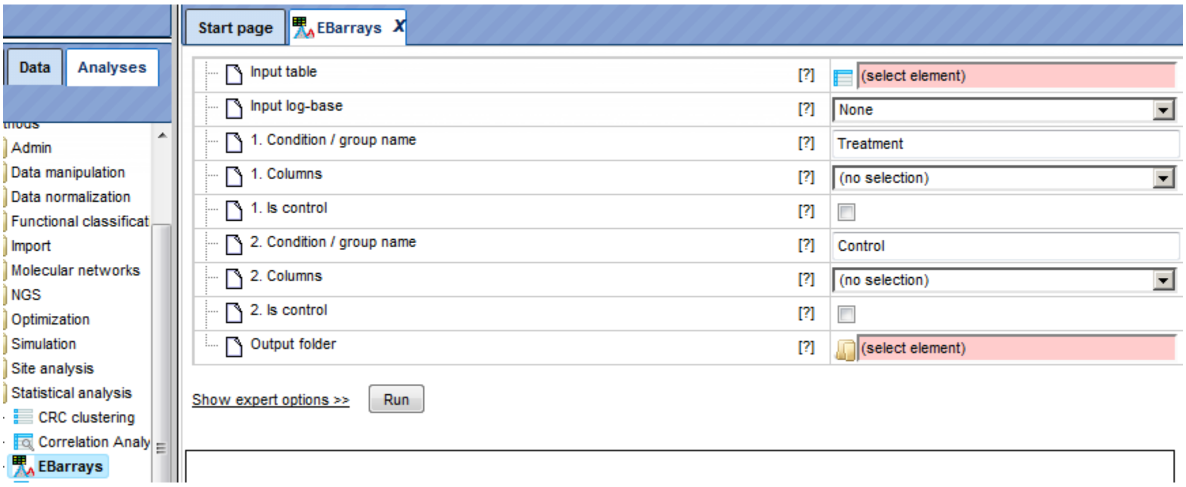
4.2.2. Estimate differential expression by the gene expression mixture model of EBarrays

EBarrays estimates differential expression between specified conditions / groups.

This tool provides for differential expression analysis using the EBarrays package. The platform tool can compare up to five conditions / groups. The groups consist of columns of a data table that contains normalized measurement values, e.g. from a normalized microarry experiment.

EBarrays sets up a mixture model matching the specified groups. Differential expression is identified when components for a pattern describe the distribution of measurement values well. Then probe / gene values in the corresponding group were significantly different from their values in the other groups. This is reflected by high posterior probabilities in the column named after that group.

The package estimates a critical posterior probability cutoff for the given FDR level on the basis of the fitted mixture model. Probes / genes exceeding this cutoff in some condition / group are indicated by a value of 1 (instead of -1) in the output column named "*condition name* Sig". Hence, to isolate the targets differentially expressed in a condition of interest, e.g. condition named "treatment", filter the table for all rows with a value of 1 in the column "treatment Sig". The direction of differential expression can be derived from the fold change column "condition name FC", which contains the log2-fold changes.



Input table	[?]	(select element)
Input log-base	[?]	None
1. Condition / group name	[?]	Treatment
1. Columns	[?]	(no selection)
1. Is control	[?]	□
2. Condition / group name	[?]	Control
2. Columns	[?]	(no selection)
2. Is control	[?]	□
Output folder	[?]	(select element)

The input parameters for EBarrays are described in the following.

Input table: This table contains the columns to analyze.

Input log-base: Here you can specify the scale of the input data. If the log-base is *none*, the tool will use the data values as is. If your data are from RNA-seq, you can select *Normalized counts* or *Unnormalized counts*.

1-5. Condition / group name: One can specify up to five groups of columns. Please note that unnamed groups are not considered; a name is not assigned automatically. Fields for the first two groups need to be set.

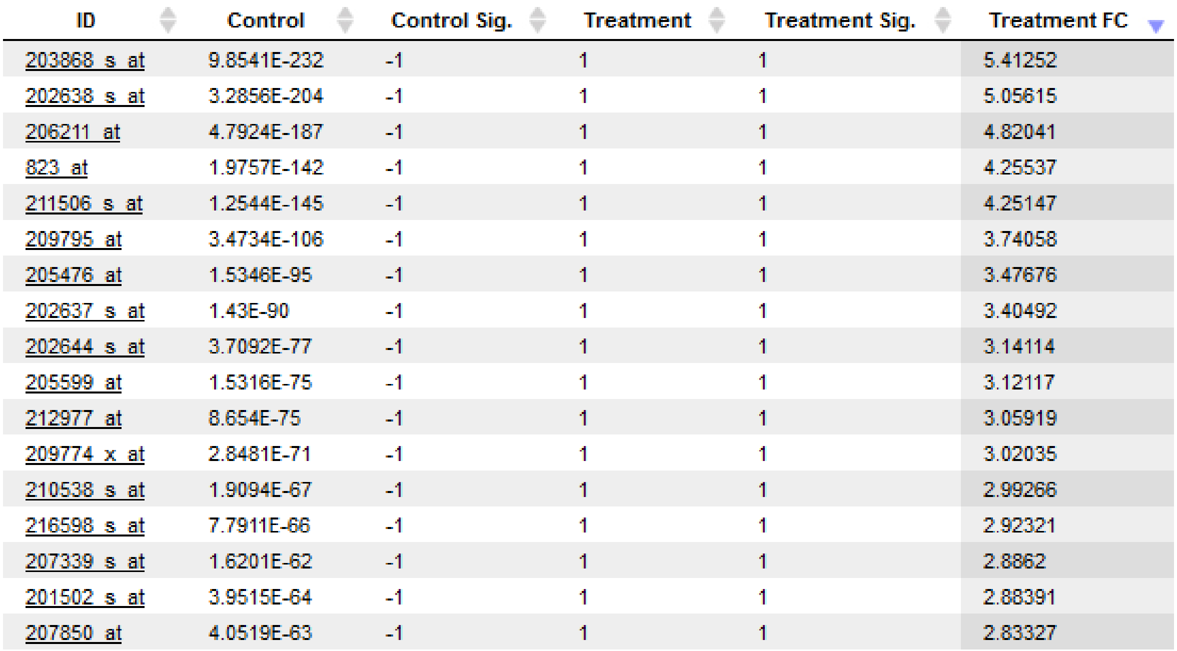
1-5. Columns: These fields contain the selected columns. Please note that column selections are not considered without a corresponding name. Columns can only be specified once and there need to be two columns per group. Fields for the first two groups need to be set.

1-5. Is control: Use this field to mark the control column group. One such group is required.

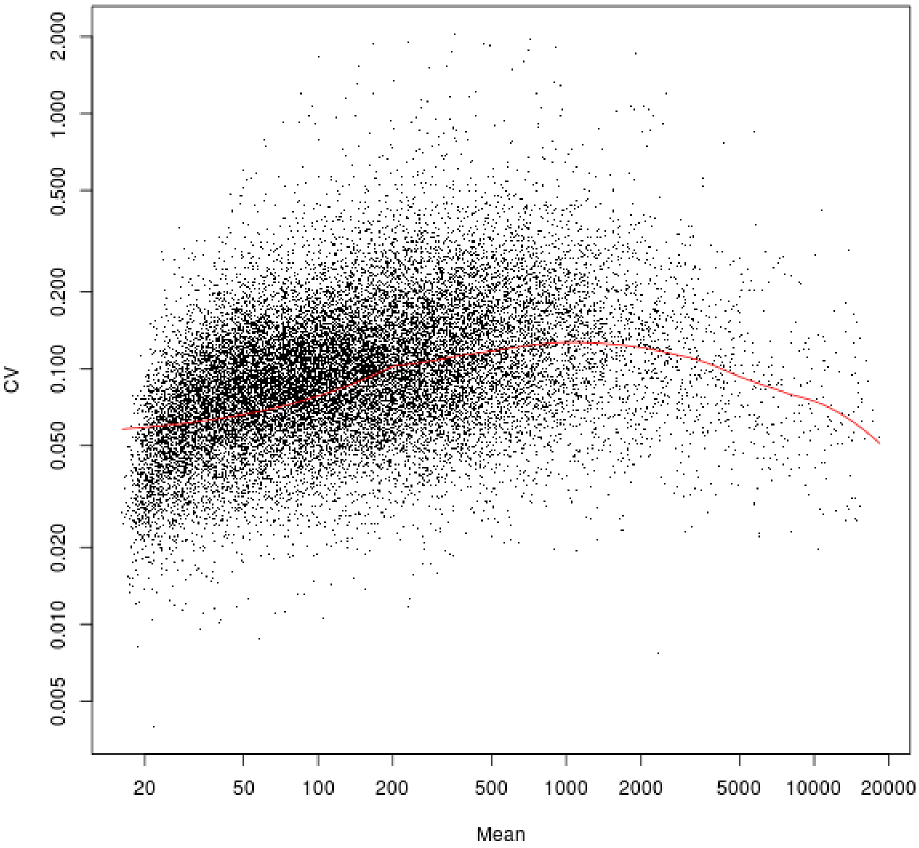
Output folder: The output folder will contain one output files for each pair of conditions.

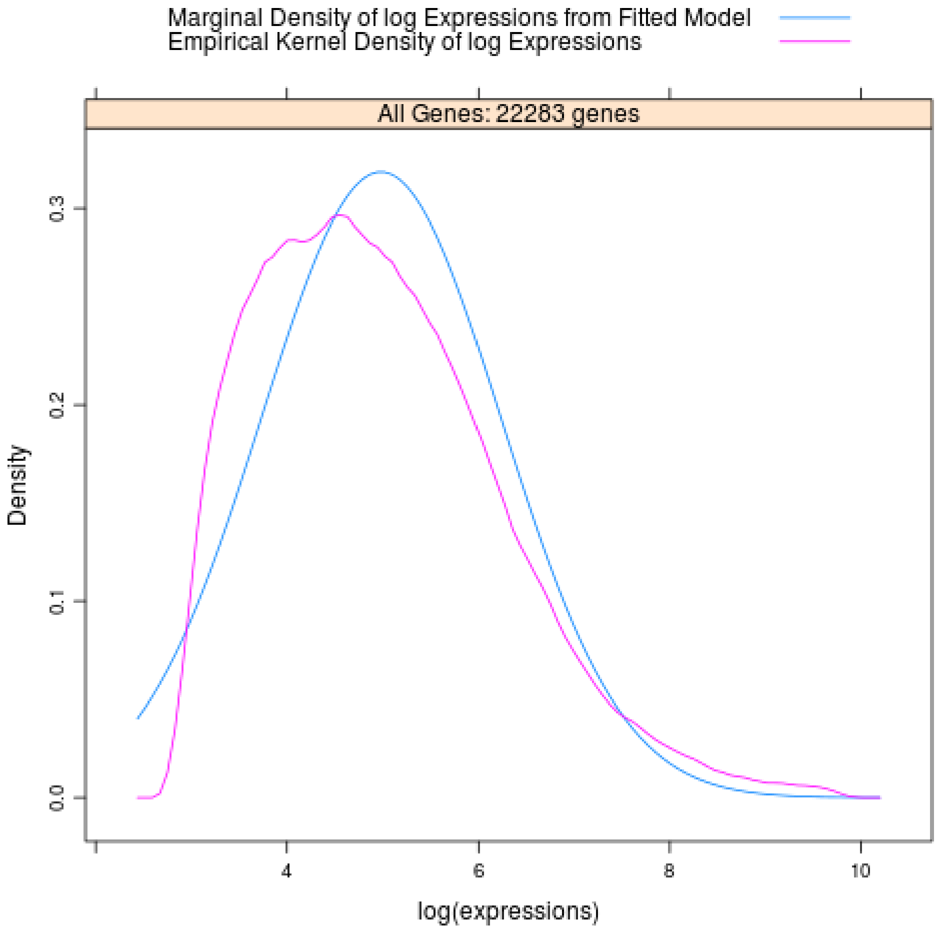
It is necessary to provide a unique name for each group. Also, at least two data columns are required per group and one group needs to be marked as control group.

Besides the main output table containing differential expression estimates for each probe / gene, EBarrays provides two diagnostic plots named EBarrays CCV and EBarrays Marginal fit. These plots enable a judgment about whether assumptions of the approach hold and how well the fitted model represents the data (please refer to the documentation of the EBarrays Bioconductor package for further details). Examples of an output table, a CCV plot and a Marginal fit plot are shown at the end of this section.



ID	Control	Control Sig.	Treatment	Treatment Sig.	Treatment FC
203868_s_at	9.8541E-232	-1	1	1	5.41252
202638_s_at	3.2856E-204	-1	1	1	5.05615
206211_at	4.7924E-187	-1	1	1	4.82041
823_at	1.9757E-142	-1	1	1	4.25537
211506_s_at	1.2544E-145	-1	1	1	4.25147
209795_at	3.4734E-106	-1	1	1	3.74058
205476_at	1.5346E-95	-1	1	1	3.47676
202637_s_at	1.43E-90	-1	1	1	3.40492
202644_s_at	3.7092E-77	-1	1	1	3.14114
205599_at	1.5316E-75	-1	1	1	3.12117
212977_at	8.654E-75	-1	1	1	3.05919
209774_x_at	2.8481E-71	-1	1	1	3.02035
210538_s_at	1.9094E-67	-1	1	1	2.99266
216598_s_at	7.7911E-66	-1	1	1	2.92321
207339_s_at	1.6201E-62	-1	1	1	2.8862
201502_s_at	3.9515E-64	-1	1	1	2.88391
207850_at	4.0519E-63	-1	1	1	2.83327


CV
2.000
1.000
0.500
0.200
0.100
0.050
0.020
0.010
0.005
20
50
100
200
500
1000
2000
5000
10000
20000
Mean


Marginal Density of log Expressions from Fitted Model
Empirical Kernel Density of log Expressions
All Genes: 22283 genes
Density
0.3
0.2
0.1
0.0
4
6
8
10
log(expressions)

Reference:

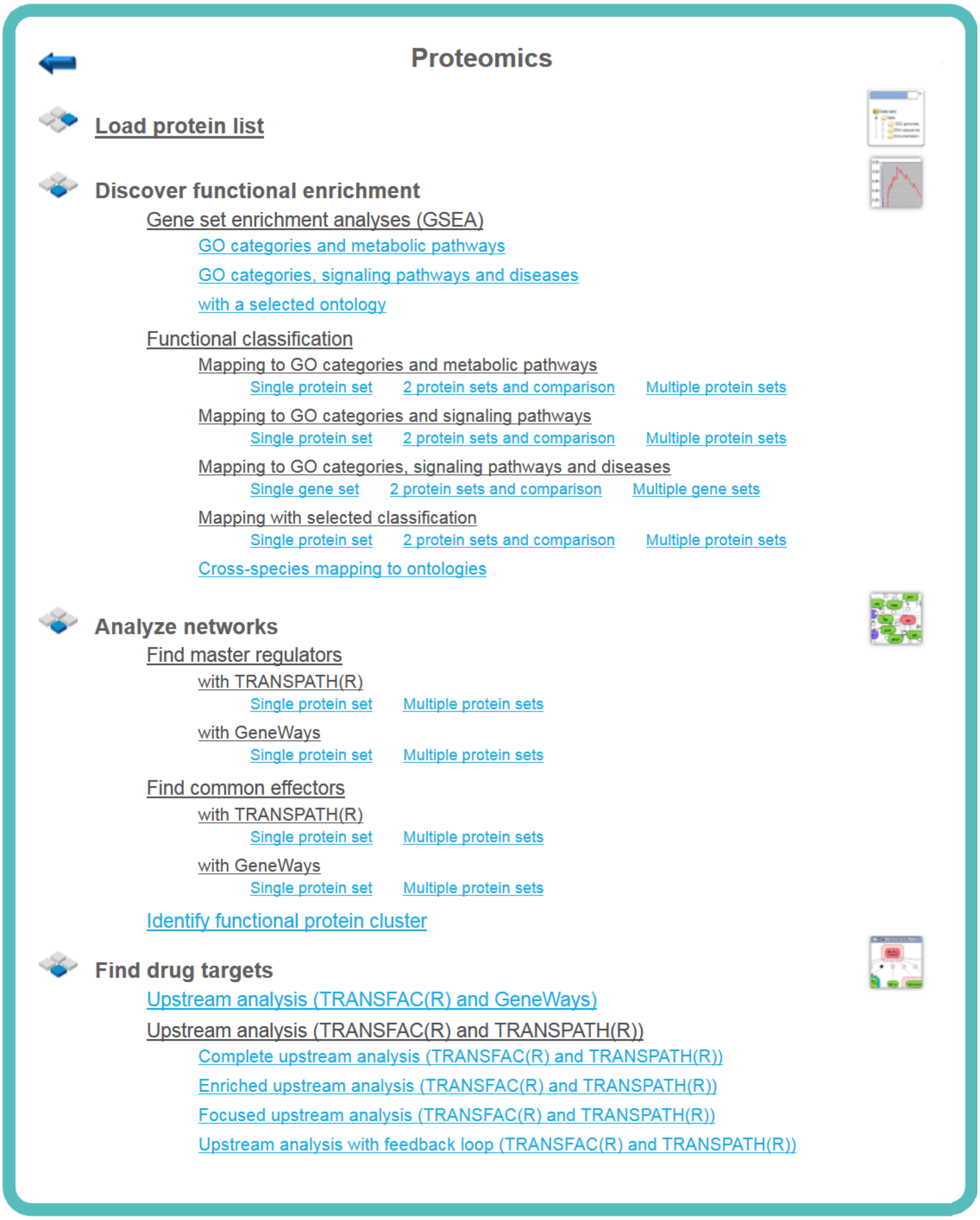
Kendziorski, C.M., Newton, M.A., Lan, H., Gould, M.N. (2003). On parametric empirical Bayes methods for comparing multiple groups using replicated gene expression profiles. Statistics in Medicine 22:3899-3914.

4.3. Further workflows in this area

For the other workflows that you can find in the area ***RNA-seq***, please refer to the following Sections:

Load data	See Chapter 3
Discover functional enrichment of DEG	See Section 10.3
Analyze networks of DEG	See Section 5.1
Analyze regulatory regions of DEG	See Section 10.4
Find drug targets	See Chapter 11

5. Proteomics



Proteomics

Load protein list

Discover functional enrichment

Gene set enrichment analyses (GSEA)

GO categories and metabolic pathways

GO categories, signaling pathways and diseases

with a selected ontology

Functional classification

Mapping to GO categories and metabolic pathways

Single protein set   2 protein sets and comparison   Multiple protein sets

Mapping to GO categories and signaling pathways

Single protein set   2 protein sets and comparison   Multiple protein sets

Mapping to GO categories, signaling pathways and diseases

Single gene set   2 protein sets and comparison   Multiple gene sets

Mapping with selected classification

Single protein set   2 protein sets and comparison   Multiple protein sets

Cross-species mapping to ontologies

Analyze networks

Find master regulators

with TRANSPATH(R)

Single protein set   Multiple protein sets

with GeneWays

Single protein set   Multiple protein sets

Find common effectors

with TRANSPATH(R)

Single protein set   Multiple protein sets

with GeneWays

Single protein set   Multiple protein sets

Identify functional protein cluster

Find drug targets

Upstream analysis (TRANSFAC(R) and GeneWays)

Upstream analysis (TRANSFAC(R) and TRANSPATH(R))

Complete upstream analysis (TRANSFAC(R) and TRANSPATH(R))

Enriched upstream analysis (TRANSFAC(R) and TRANSPATH(R))

Focused upstream analysis (TRANSFAC(R) and TRANSPATH(R))

Upstream analysis with feedback loop (TRANSFAC(R) and TRANSPATH(R))

5.1. Analyze networks

5.1.1. Find master regulators

Potential master regulators of the processes analyzed in a typical proteomics experiment can be identified with the aid of pathway databases. The geneXplain platform support working with the TRANSPATH database (5.1.1.1; license required) or with the public

GeneWays database (5.1.1.2). More details about both these database can be found in the corresponding Sections 19.7 and 19.5.

5.1.1.1. Find master regulators with TRANSPATH®

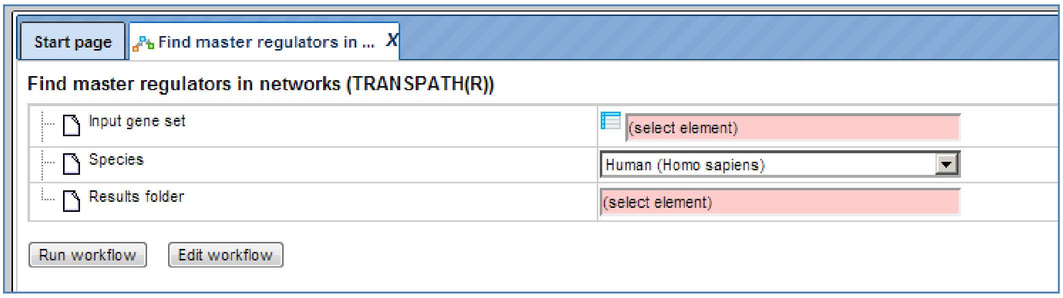
As elsewhere, these workflows can be used to analyze data of a **single protein table** or to mine **multiple protein sets**. These two options will be explained in the following, complemented by a more detailed explanation how the **interpretation of the results** should be done.

Analyze a single gene table

This workflow is designed to find important master regulators in signal transduction pathways. The search is done based on the network of the TRANSPATH® database with a maximum radius of 10 steps upstream of an input gene set, a default cutoff for Score at 0.2, for FDR at 0.05, and for Z-score at 1.0.

To launch the workflow, follow these steps:

Step 1. Open the workflow input form from the Start page. It looks as shown below:

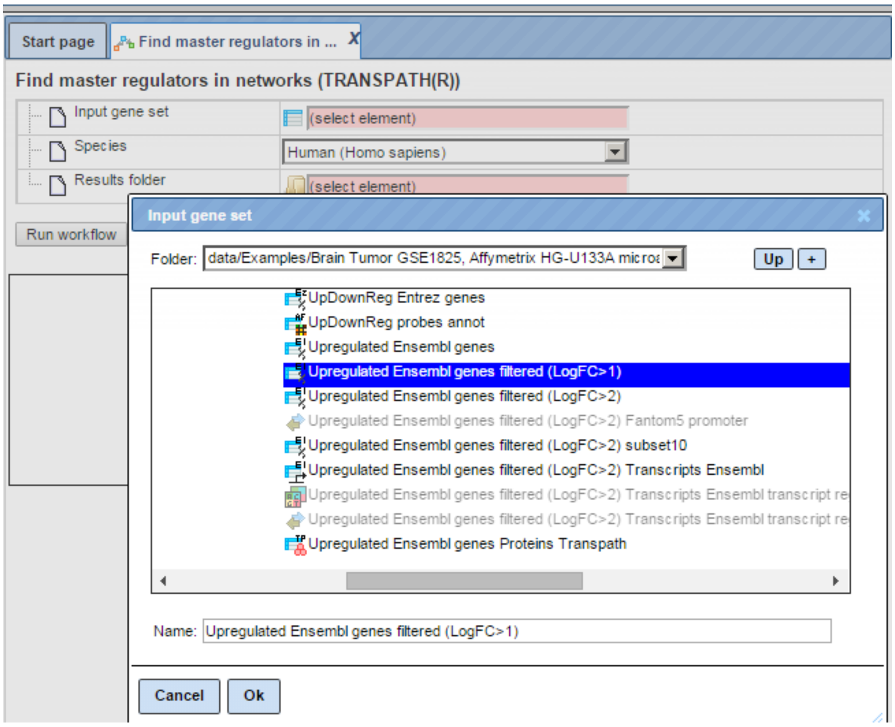


Find master regulators in networks (TRANSPATH(R))

Input gene set	(select element)
Species	Human (Homo sapiens)
Results folder	(select element)

Step2. Specify input gene set. The input gene set might be a list of differentially regulated genes or any gene or protein list of interest. You can drag it from your project within the tree area and drop it in the pink box of the field **Input gene set**. Alternatively, you may click on the pink field "select element" and a new window will be opened, where you can select the input gene set as shown below.

The further steps of the workflow are demonstrated by means of the genes shown to be up-regulated in one of the pre-prepared examples. The pertinent example file can be found in the geneXplain platform online under the path: http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Experiment%20normalized%20(RMA)%20(Differentially%20expressed%20genes%20Affy)/Upregulated%20Ensembl%20genes%20filtered%20(LogFC%3E1)



Find master regulators in networks (TRANSPATH(R))

Input gene set: (select element)
Species: Human (Homo sapiens)
Results folder: (select element)

Input gene set

Folder: data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A micro

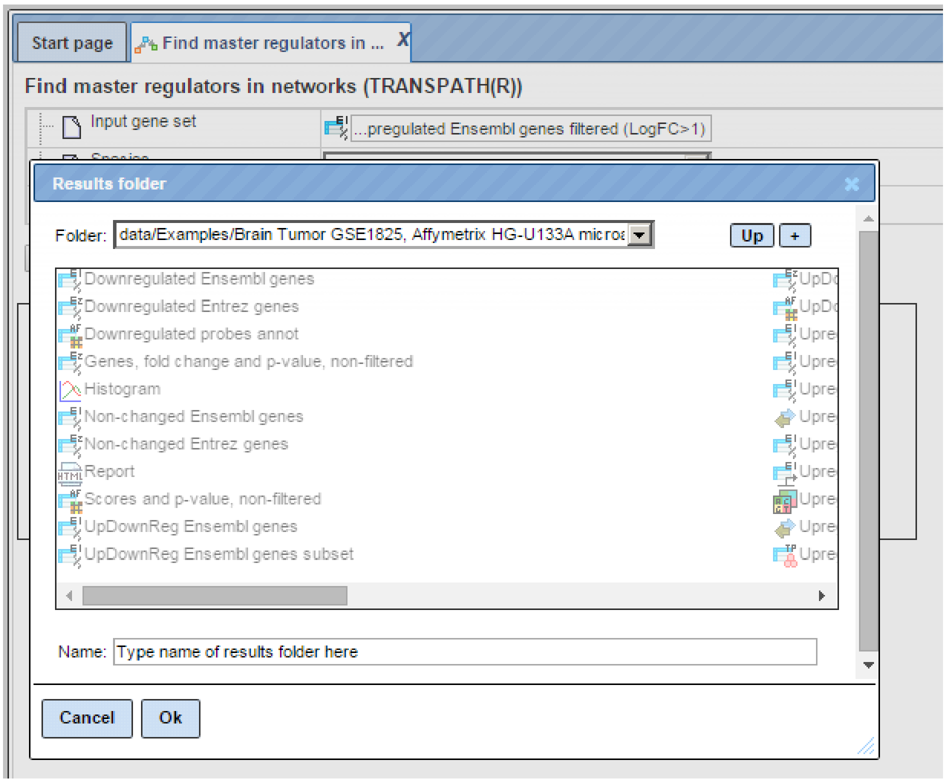
- UpDownReg Entrez genes
- UpDownReg probes annot
- Upregulated Ensembl genes
- Upregulated Ensembl genes filtered (LogFC>1)
- Upregulated Ensembl genes filtered (LogFC>2)
- Upregulated Ensembl genes filtered (LogFC>2) Fantom5 promoter
- Upregulated Ensembl genes filtered (LogFC>2) subset10
- Upregulated Ensembl genes filtered (LogFC>2) Transcripts Ensembl
- Upregulated Ensembl genes filtered (LogFC>2) Transcripts Ensembl transcript re
- Upregulated Ensembl genes filtered (LogFC>2) Transcripts Ensembl transcript re
- Upregulated Ensembl genes Proteins Transpath

Name: Upregulated Ensembl genes filtered (LogFC>1)

When you have selected the gene set, press [Ok].

Step 3. Specify the biological species of the input set in the field **Species** by selecting the required biological species from the drop-down menu.

Step 4. Define where the folder with the results should be located in the tree. You can do so by clicking on the pink field "select element" in the field "Results folder", and a new window will be opened, where you can select the location of the results folder and define its name as shown below.



Find master regulators in networks (TRANSPATH(R))

Input gene set: ...pregulated Ensembl genes filtered (LogFC>1)

Results folder

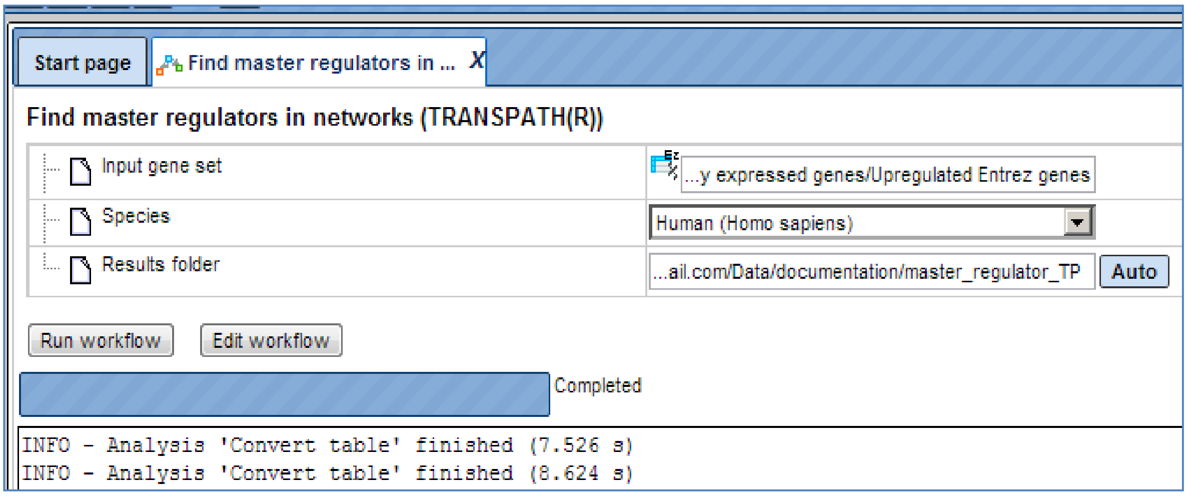
Folder: data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A micro

- Downregulated Ensembl genes
- Downregulated Entrez genes
- Downregulated probes annot
- Genes, fold change and p-value, non-filtered
- Histogram
- Non-changed Ensembl genes
- Non-changed Entrez genes
- Report
- Scores and p-value, non-filtered
- UpDownReg Ensembl genes
- UpDownReg Ensembl genes subset
- UpDo
- UpDo
- Upre
- Upre
- Upre
- Upre
- Upre
- Upre
- Upre
- Upre
- Upre

Name: Type name of results folder here

After you have given the name, press [Ok].

Step5. Press the [Run workflow] button. Wait until the workflow is completed, which is shown below:



Find master regulators in networks (TRANSPATH(R))

Input gene set: ...y expressed genes/Upregulated Entrez genes

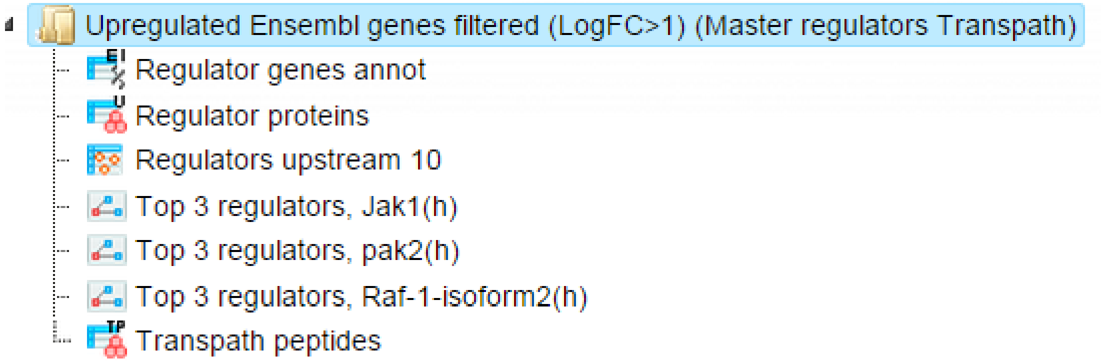
Species: Human (Homo sapiens)

Results folder: ...ail.com/Data/documentation/master_regulator_TP

Completed

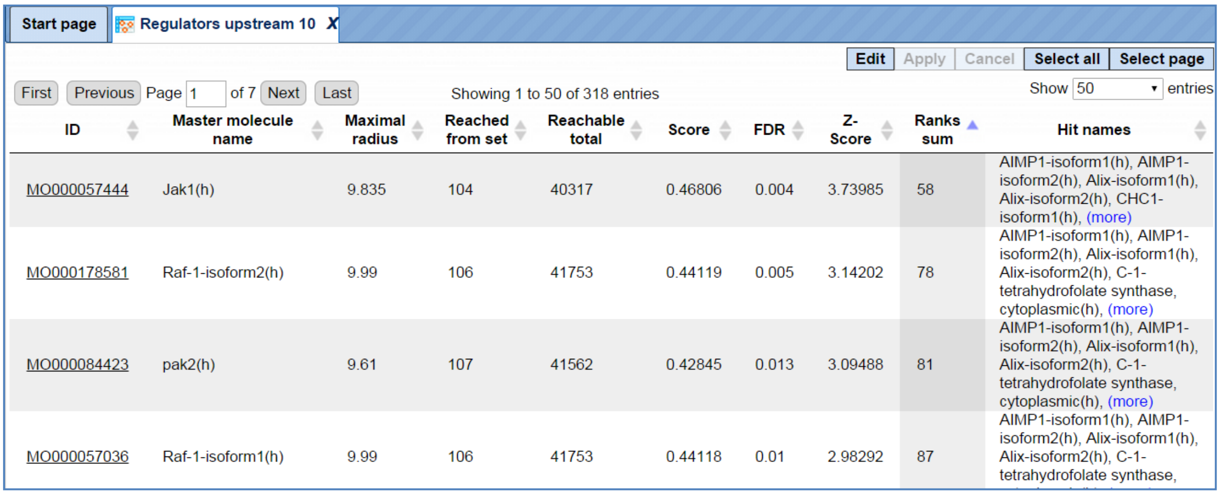
```
INFO - Analysis 'Convert table' finished (7.526 s)
INFO - Analysis 'Convert table' finished (8.624 s)
```

The results folder contains several files.



```
Upregulated Ensembl genes filtered (LogFC>1) (Master regulators Transpath)
    Regulator genes annot
    Regulator proteins
    Regulators upstream 10
    Top 3 regulators, Jak1(h)
    Top 3 regulators, pak2(h)
    Top 3 regulators, Raf-1-isoform2(h)
    Transpath peptides
```

The primary result table *Regulators upstream 10* ( ) is a list of master regulatory molecules that were identified at the distance up to 10 steps upstream of the input molecules. Each master regulatory molecule is characterized by a Score, Z-score, FDR, and Ranks Sum. Further details about these parameters can be found below, under “Interpretation of the results”.



Showing 1 to 50 of 318 entries

ID	Master molecule name	Maximal radius	Reached from set	Reachable total	Score	FDR	Z-Score	Ranks sum	Hit names
MO000057444	Jak1(h)	9.835	104	40317	0.46806	0.004	3.73985	58	AIMP1-isoform1(h), AIMP1-isoform2(h), Alix-isoform1(h), Alix-isoform2(h), CHC1-isoform1(h), (more)
MO000178581	Raf-1-isoform2(h)	9.99	106	41753	0.44119	0.005	3.14202	78	AIMP1-isoform1(h), AIMP1-isoform2(h), Alix-isoform1(h), Alix-isoform2(h), C-1-tetrahydrofolate synthase, cytoplasmic(h), (more)
MO000084423	pak2(h)	9.61	107	41562	0.42845	0.013	3.09488	81	AIMP1-isoform1(h), AIMP1-isoform2(h), Alix-isoform1(h), Alix-isoform2(h), C-1-tetrahydrofolate synthase, cytoplasmic(h), (more)
MO000057036	Raf-1-isoform1(h)	9.99	106	41753	0.44118	0.01	2.98292	87	AIMP1-isoform1(h), AIMP1-isoform2(h), Alix-isoform1(h), Alix-isoform2(h), C-1-tetrahydrofolate synthase,

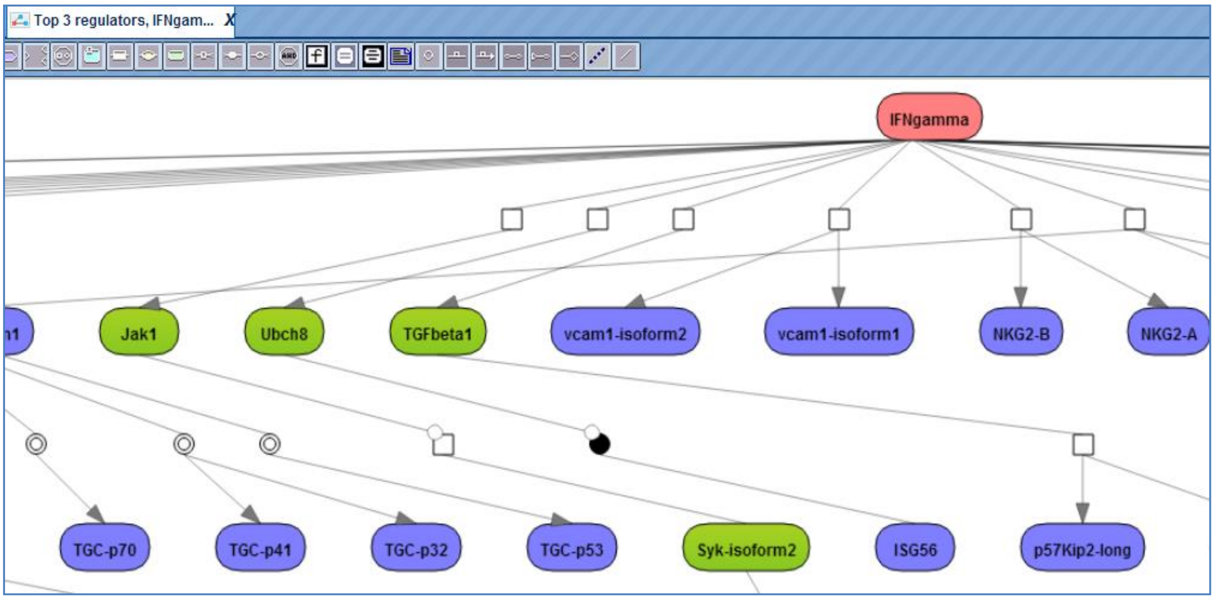
The column “Reached from set” shows the number of the molecules from the input set that is reached from the respective master regulator, and these molecules are explicitly listed in the column **Hit names**. The column **Reachable total** presents the total number of molecules that can be reached from the master regulator in the network, independent of the input list. Details about Score, Z-score, FDR and Ranks sum columns are given below, under “Interpretation of the results”..

Having this table opened in the Work Space you can find additional options available, specific for this kind of table. Select one or several rows in the table “Regulators upstream 10” by mouse click, and you can visualize the network of the selected master regulators ( ), save the network as a list of genes in the Tree Area ( ), or save the hits of this network, listed in the column **Hit names** as a list of genes in the Tree Area ( ).

The table *Regulator proteins* ( ) corresponds to the table *Regulators upstream 10* converted into the UniProtKB/Swiss-Prot IDs.

The table *Regulator genes* ( ) corresponds to the table *Regulators upstream 10* converted into the Ensembl IDs and in the table *Regulator genes annot* ( ) the same genes are additionally annotated with gene symbols and gene descriptions.

The three diagrams *Top 3 regulators* ( ) visualize the networks individually for each of three top master regulators. By default, the top regulators are identified upon sorting the table *Regulators upstream 10* ( ) by the column **Ranks sum** with the lowest rank on top.



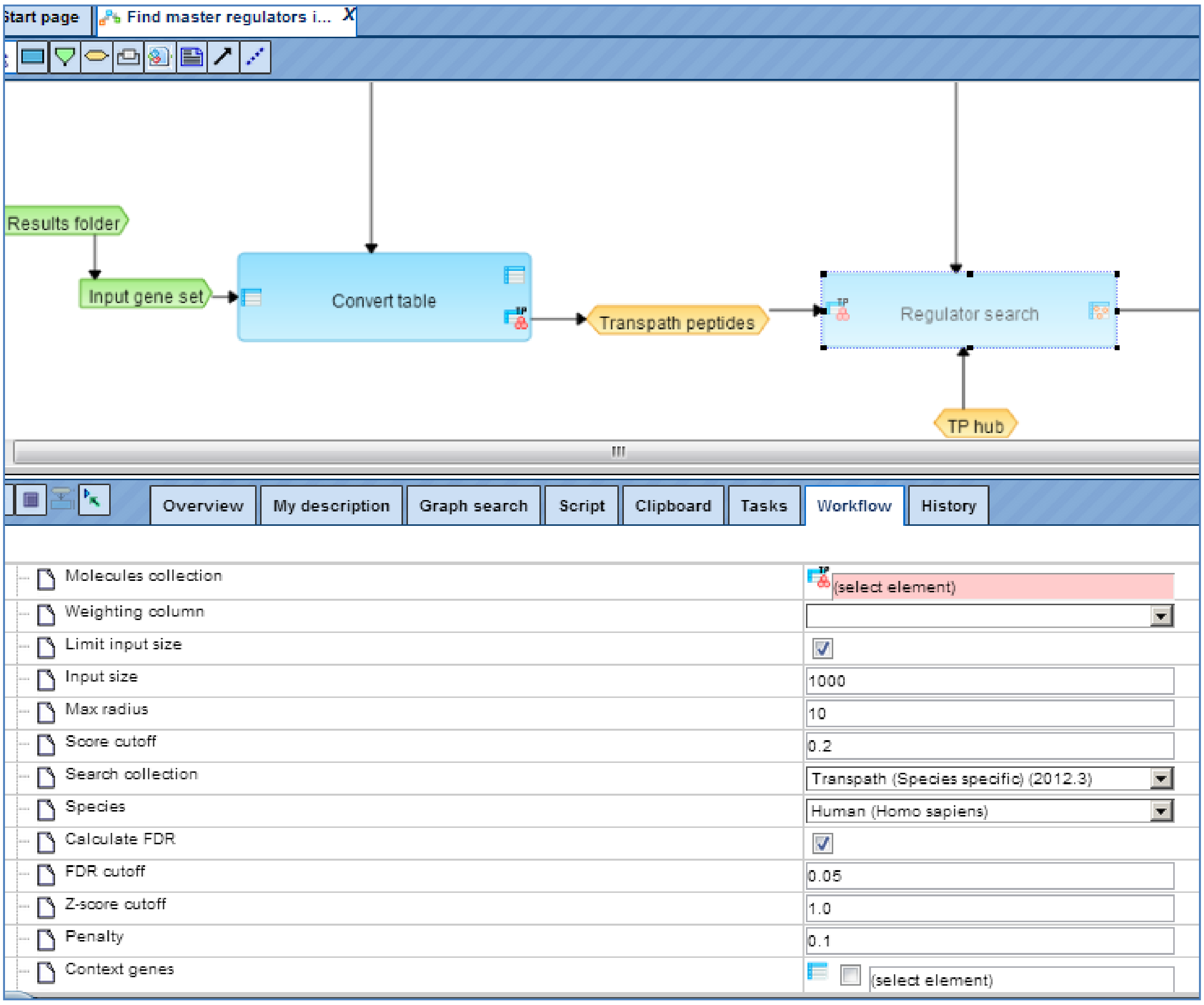
The default color code for the molecules is the following:
blue: molecules from the input list
red: master regulatory molecules;
green: connecting molecules considered by the graph-analyzing algorithm to find the path from input list to the master molecule.

If you are interested in visualizing the network for any other master regulator, you can do this in the following way. Open the table *Regulators upstream 10* and select the master regulator by a single mouse click, then click the button to visualize the selected row and save the new diagram.

Tip for the workflow editing

You can easily create a similar workflow with parameter values adjusted to your needs. For example, you might be interested to change the number of steps used for the regulator search. By default, 10 steps are applied.

To make a change, you need first to open the workflow under the "Edit workflow" mode, and save its copy in your project area. The [Edit workflow] button is located near the button [Run workflow] (see above, Step 1). Upon clicking on "Edit workflow", the workflow diagram will be opened in the Work Space, and you can select the analysis box you would like to modify. On the screenshot below "Regulator search" analysis was selected, and in the Operations Field, on the tab "Workflow", all the parameters are visible. Under this mode, you can modify default parameters and then save the workflow.



Molecules collection	(select element)
Weighting column	
Limit input size	☑
Input size	1000
Max radius	10
Score cutoff	0.2
Search collection	Transpath (Species specific) (2012.3)
Species	Human (Homo sapiens)
Calculate FDR	☑
FDR cutoff	0.05
Z-score cutoff	1.0
Penalty	0.1
Context genes	□ (select element)

In this way you will get a customized workflow, with the parameters specified according to your needs.

Note. This workflow is available together with a valid TRANSPATH® license. Please, feel free to ask for details (info@genexplain.com).

Analyze multiple gene sets

The input is a folder with several gene/protein tables. The steps of this workflow for each individual gene/protein table are the same as described in the section above. The same steps are performed iteratively for each of the gene/protein tables in the input folder.

The output is a folder which contains subfolders with the results for each individual gene/protein table.

Interpretation of the results

Score

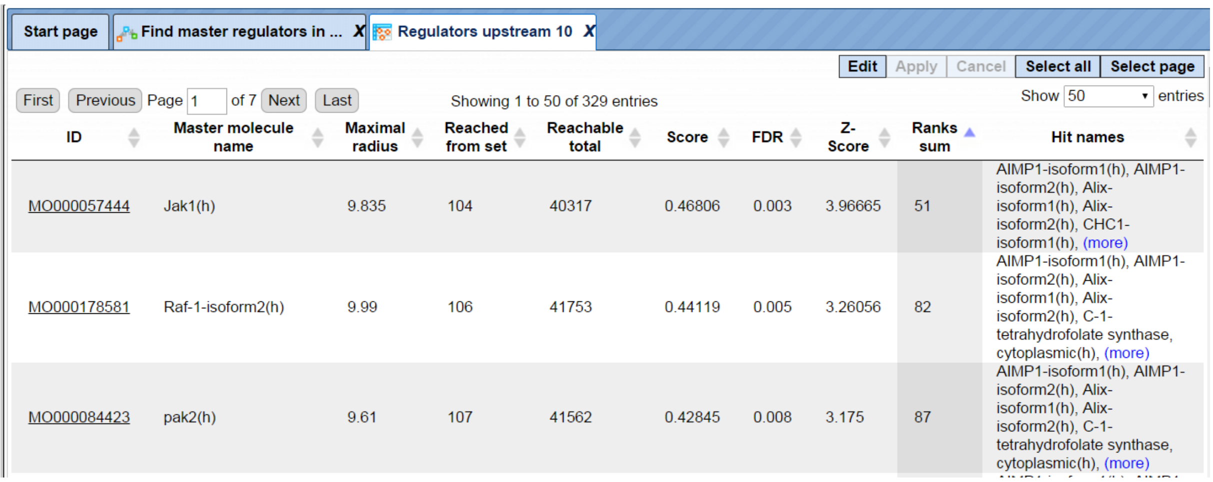
The score value of each master regulatory molecule reflects how well this molecule is connected with other molecules in the database, and how many molecules from the input list are present in the network of this master molecule. The higher the Score value, the better is this molecule connected in the database, and the more “Hits” from the input list are present in the network of this molecule. By default, only the molecules with Score > 0.2 are shown in the output.

Because molecules with high Scores are well connected in the database, they are being suggested quite often by the tool as potential master regulators even with different input lists, and sometimes such molecules are also expected to be found *a priori*. It is possible to say that the molecules with the highest Score values are a kind of "trivial" and expected solutions. At the same time, and also because of their good connectivity, they are well studied and published. Therefore the molecules with high Score values might be biologically interesting as known "hubs" in a network.

Let's have a look at the table *Regulators Upstream 10*, available in the geneXplain platform online under the path:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Upregulated%20Ensembl%20genes%20filtered%20(LogFC%3E1)%20(Master%20regulators%20Transpath)/Regulators%20upstream%2010

In the screenshot below, the table has been sorted by Score, and we can see PKCalpha, PDK1, Cdk5 as the three top molecules with highest Score values. These are well-studied molecules, and in many cases there is no surprise for the researcher to find such molecules as master regulators; it is a kind of expected result. However, master regulators with high Scores might be of interest if you are looking for well-studied reliable molecules, and would like to see many of the input molecules connected by such master regulators.



Showing 1 to 50 of 329 entries

Show 50 entries

ID	Master molecule name	Maximal radius	Reached from set	Reachable total	Score	FDR	Z-Score	Ranks sum	Hit names
MO000057444	Jak1(h)	9.835	104	40317	0.46806	0.003	3.96665	51	AIMP1-isoform1(h), AIMP1-isoform2(h), Alix-isoform1(h), Alix-isoform2(h), CHC1-isoform1(h), (more)
MO000178581	Raf-1-isoform2(h)	9.99	106	41753	0.44119	0.005	3.26056	82	AIMP1-isoform1(h), AIMP1-isoform2(h), Alix-isoform1(h), Alix-isoform2(h), C-1-tetrahydrofolate synthase, cytoplasmic(h), (more)
MO000084423	pak2(h)	9.61	107	41562	0.42845	0.008	3.175	87	AIMP1-isoform1(h), AIMP1-isoform2(h), Alix-isoform1(h), Alix-isoform2(h), C-1-tetrahydrofolate synthase, cytoplasmic(h), (more)

Z-score

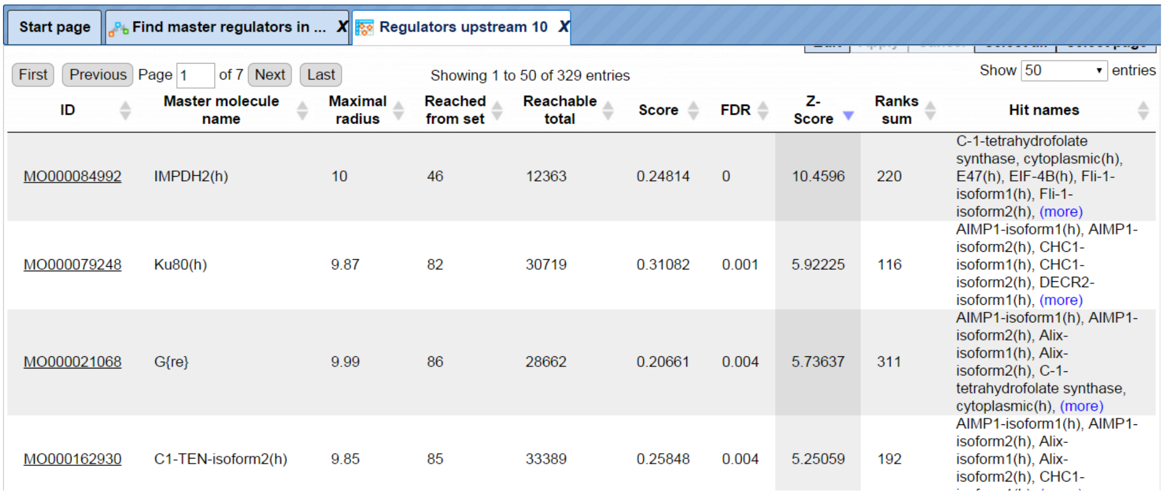
The Z-score value reflects how specific each master molecule is for the input list. The higher the Z-score value for a molecule, the more specific this molecule is for the input list, and the lesser is the probability to find such a molecule as master regulator in another analysis. Z-score and FDR are calculated based on 1000 random results, for which 1000 random input sets of the same size were generated by the algorithm.

Importantly, Score and Z-score reflect different characteristics of the suggested master regulators in the networks. Molecules with high Score values are well connected in the database, and therefore not very specific for the input list, and correspondingly they have quite moderate Z-score values.

Molecules with highest Z-scores are very specific for the input list, probably because of a few connections that are specific for the input list, but generally they are not so well connected within the database and therefore have quite low Score values.

Sorting by Z-score and considering top molecules might be helpful if you are interested in finding novel master regulators which are specific for your input list and generally are not well studied yet. By default, only the molecules with Z-score > 1.0 are shown in the output.

On the screenshot below the same table as above is sorted here by Z-score, and we can see different molecules on top. Even by the names of these molecules the expert can see that they are not coming up so often in the literature, and might represent interesting novel candidates.



Showing 1 to 50 of 329 entries

ID	Master molecule name	Maximal radius	Reached from set	Reachable total	Score	FDR	Z-Score	Ranks sum	Hit names
MO000084992	IMPDH2(h)	10	46	12363	0.24814	0	10.4596	220	C-1-tetrahydrofolate synthase, cytoplasmic(h), E47(h), EIF-4B(h), Fli-1-isoform1(h), Fli-1-isoform2(h), (more)
MO000079248	Ku80(h)	9.87	82	30719	0.31082	0.001	5.92225	116	AIMP1-isoform1(h), AIMP1-isoform2(h), CHC1-isoform1(h), CHC1-isoform2(h), DECR2-isoform1(h), (more)
MO000021068	G{re}	9.99	86	28662	0.20661	0.004	5.73637	311	AIMP1-isoform1(h), AIMP1-isoform2(h), Alix-isoform1(h), Alix-isoform2(h), C-1-tetrahydrofolate synthase, cytoplasmic(h), (more)
MO000162930	C1-TEN-isoform2(h)	9.85	85	33389	0.25848	0.004	5.25059	192	AIMP1-isoform1(h), AIMP1-isoform2(h), Alix-isoform1(h), Alix-isoform2(h), CHC1-

Ranks sum

This column helps to suggest molecules for which both values, Score and Z-score, are quite good. The column **Ranks sum** reflects a combination of sorting by Score and by Z-score in the following way.

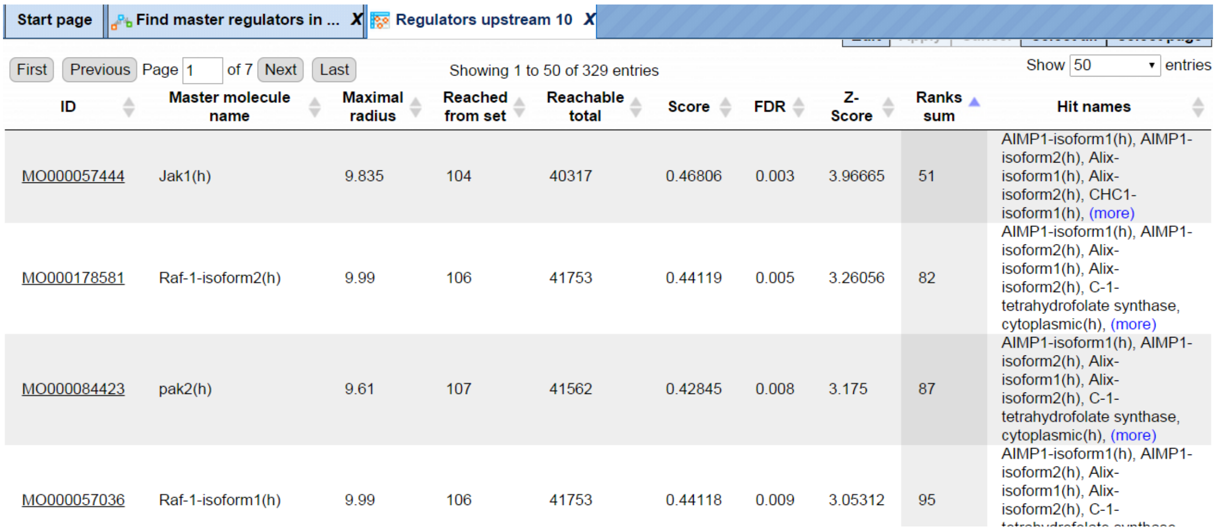
Upon sorting by Score from biggest values to the lowest, a rank is assigned to the molecules; the molecule with the highest Score has rank 1, etc.

Upon independent sorting by Z-Score from biggest values to lowest, a rank is assigned to the molecules; the molecule with the highest Z-score has rank 1, etc.

Next, for each molecule, the ranks upon sorting by Score and upon sorting by Z-Score are summed up in the column **Ranks Sum**. The lower the Ranks sum, the more interesting the candidate molecule is, with good Score and good Z-score values.

On the screenshot below the table from above is sorted by Ranks sum, and we can see different molecules on top. Upon such sorting, on top there are molecules with a good connection in the database, and simultaneously to a quite good extent specific for the input list. In this example, Score values for the top molecules are between 0.34 and 0.45

(moderate), Z-score values vary between 4.7 and 8.4 (very good Z-score values, but not the best in this table).



Showing 1 to 50 of 329 entries

ID	Master molecule name	Maximal radius	Reached from set	Reachable total	Score	FDR	Z-Score	Ranks sum	Hit names
MO000057444	Jak1(h)	9.835	104	40317	0.46806	0.003	3.96665	51	AIMP1-isoform1(h), AIMP1-isoform2(h), Alix-isoform1(h), Alix-isoform2(h), CHC1-isoform1(h), (more)
MO000178581	Raf-1-isoform2(h)	9.99	106	41753	0.44119	0.005	3.26056	82	AIMP1-isoform1(h), AIMP1-isoform2(h), Alix-isoform1(h), Alix-isoform2(h), C-1-tetrahydrofolate synthase, cytoplasmic(h), (more)
MO000084423	pak2(h)	9.61	107	41562	0.42845	0.008	3.175	87	AIMP1-isoform1(h), AIMP1-isoform2(h), Alix-isoform1(h), Alix-isoform2(h), C-1-tetrahydrofolate synthase, cytoplasmic(h), (more)
MO000057036	Raf-1-isoform1(h)	9.99	106	41753	0.44118	0.009	3.05312	95	AIMP1-isoform1(h), AIMP1-isoform2(h), Alix-isoform1(h), Alix-isoform2(h), C-1-

By default, the table *Regulators Upstream 10* are sorted by the **Ranks sum** column, to suggest molecules with a balance between their well-studied status and high connectivity (reflected by Score), and novelty and specificity for the input list (reflected by the Z-score).

Suggestion for sorting master regulatory molecules

It might be very helpful to find out which of the suggested master regulators are expressed in your experiment, and especially which are up-regulated. Such molecules might be promising candidates for further experimental examinations. To do this, you can take the table in the result folder "Regulator genes annot" and take it as input for the analysis "Annotate table".

As annotation source, you can select the table of genes expressed in the same experiment, e.g. the table of all expressed genes that resulted from the workflow "Detect differentially expressed genes"; in this example the path is:

data/Examples/Breast Cancer GSE9187, Agilent 014850 microarray/Data/metadherin gene knockdown cells and control cells/Deferentially expressed genes/Genes, Fold Genes, fold change and p-value, non-filtered

As "Annotation column" you can select **LogFoldChange**, and as a result the suggested master regulators are annotated by their expression.

If you are interested in finding reliable well-studied master regulators, e.g. to confirm already known ones, and would like a master regulator network to contain as many molecules from the input list as possible, you might be interested to sort by Score, and consider master molecules with the highest Score values.

If you are looking for novel master regulators that are very specific for your input list, even when they are not well studied yet, you might be interested to sort by Z-score, and consider master molecules with highest Z-score values.

If you are looking for a good balance between well-connected molecules and novel ones specific for your input list, you might be interested to stay with the default sorting by Ranks sum, and consider master molecules with the lowest Ranks sum values.

5.1.1.2. Find master regulators with GeneWays

This workflow is designed to find important master regulators in the signal transduction pathways. Here, a search for master regulators is done based on the network of the GeneWays database with a maximum radius of 4 steps upstream of an input gene set, a default cutoff for Score at 0.2, for FDR at 0.05, and for Z-score at 1.0. The input form and the resulting tables are very similar to the workflow described above, "Find master regulators in networks (TRANSPATH®)", please refer to Section 5.1.1.1.

The major difference between these two workflows is the underlying database applied for the network analysis, either TRANSPATH® or GeneWays.

More details about the GeneWays and TRANSPATH® databases can be found in Sections 19.5 and 19.7, respectively.

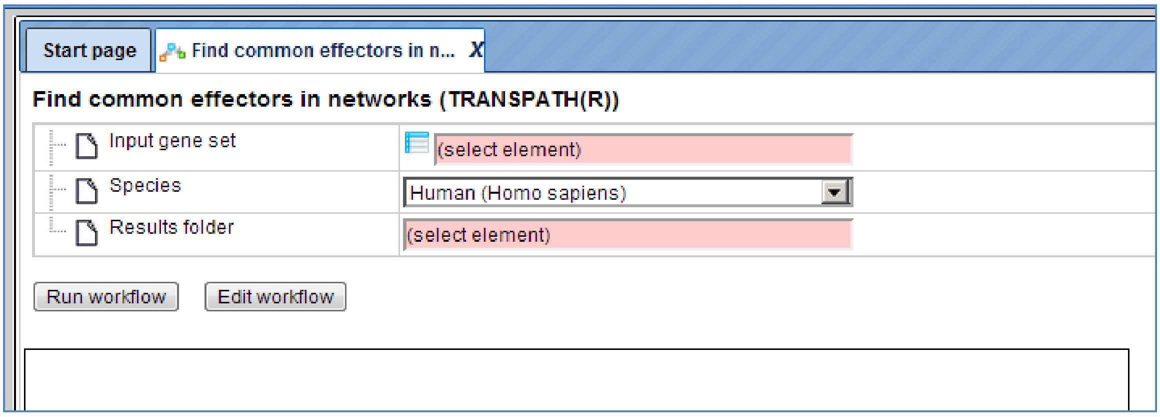
5.1.2. Find common effectors

5.1.2.1. Find common effectors with TRANSPATH®

This workflow is designed to find important effectors in signal transduction pathways. With this workflow, the effector search is done based on the network of the TRANSPATH® database with a maximum radius of 10 steps, FDR cutoff at 0.05, Score cutoff at 0.2 and a Z-score cutoff at 1.0. You have an option to edit the default parameters using the button [Edit workflow].

To launch the workflow, follow these steps:

Step 1. Open the workflow input form via the Start page. It looks as shown below:



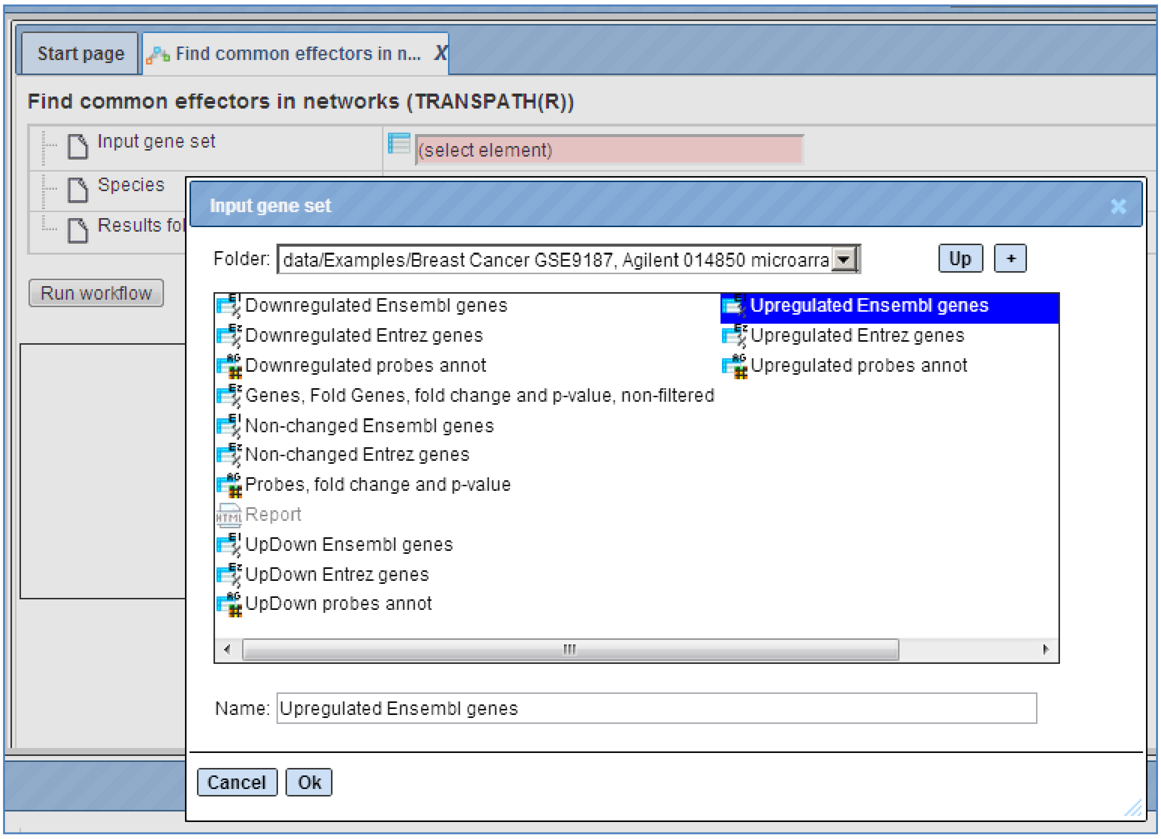
Find common effectors in networks (TRANSPATH(R))

Input gene set	(select element)
Species	Human (Homo sapiens)
Results folder	(select element)

Step2. Specify the input gene set. The input gene set might be a list of differentially regulated genes or any gene or protein list of interest. You can drag & drop it from your project within the tree area and drop it in the pink box of the field **Input gene set**. Alternatively, you may click on the pink field "select element" and a new window will be opened, where you can select the input gene set as shown below.

The further steps of this workflow are demonstrated with genes shown to be up-regulated in one of the examples. The example file can be accessed using the URL:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Experiment%20normalized%20(RMA)%20(Differentially%20expressed%20genes%20Affy)/Upregulated%20Ensembl%20genes%20filtered%20(LogFC%3E2)



Find common effectors in networks (TRANSPATH(R))

Input gene set (select element)
Species
Results fol

Input gene set

Folder: data/Examples/Breast Cancer GSE9187, Agilent 014850 microarra

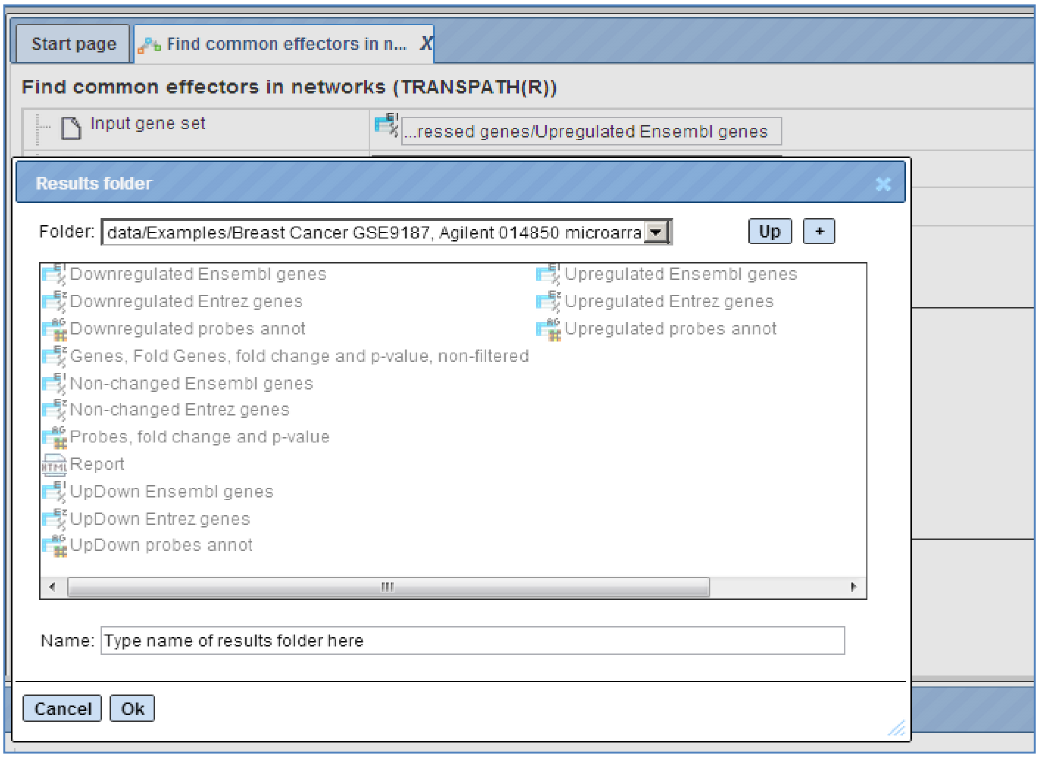
Downregulated Ensembl genes
Downregulated Entrez genes
Downregulated probes annot
Genes, Fold Genes, fold change and p-value, non-filtered
Non-changed Ensembl genes
Non-changed Entrez genes
Probes, fold change and p-value
Report
UpDown Ensembl genes
UpDown Entrez genes
UpDown probes annot
Upregulated Ensembl genes
Upregulated Entrez genes
Upregulated probes annot

Name: Upregulated Ensembl genes

After you have selected the gene set, press [Ok].

Step3. Specify the biological species of the input set in the field **Species** by selecting the required species from the drop-down menu.

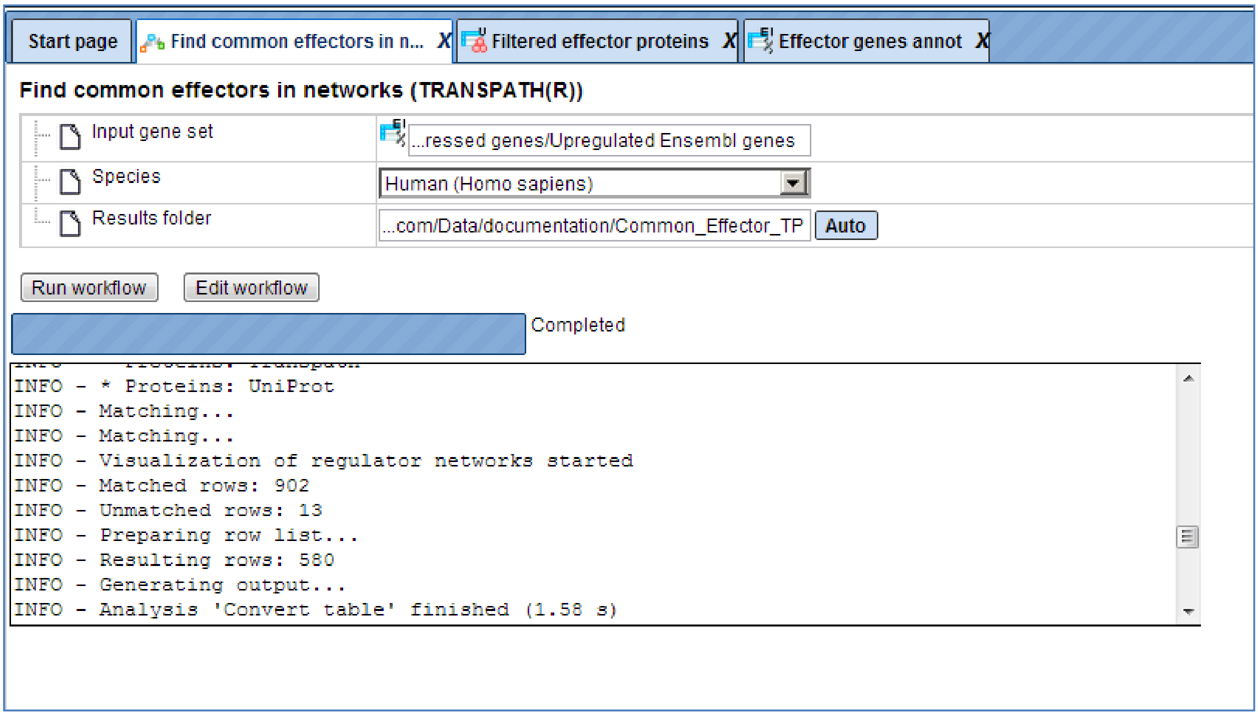
Step 4. Define where the folder with the results should be located in the tree. You can do so by clicking on the pink field “select element” in the field **Results folder**, and a new window will be opened where you can select the location of the results folder and define its name as shown below.



Start page
Find common effectors in n...
Find common effectors in networks (TRANSPATH(R))
Input gene set
...ressed genes/Upregulated Ensembl genes
Results folder
Folder: data/Examples/Breast Cancer GSE9187, Agilent 014850 microarra
Up
+
Downregulated Ensembl genes
Downregulated Entrez genes
Downregulated probes annot
Genes, Fold Genes, fold change and p-value, non-filtered
Non-changed Ensembl genes
Non-changed Entrez genes
Probes, fold change and p-value
Report
UpDown Ensembl genes
UpDown Entrez genes
UpDown probes annot
Upregulated Ensembl genes
Upregulated Entrez genes
Upregulated probes annot
Name: Type name of results folder here
Cancel
Ok

After you have specified the name, press [Ok].

Step 5. Press the button [Run workflow]. Wait until the workflow is completed, which is shown below:

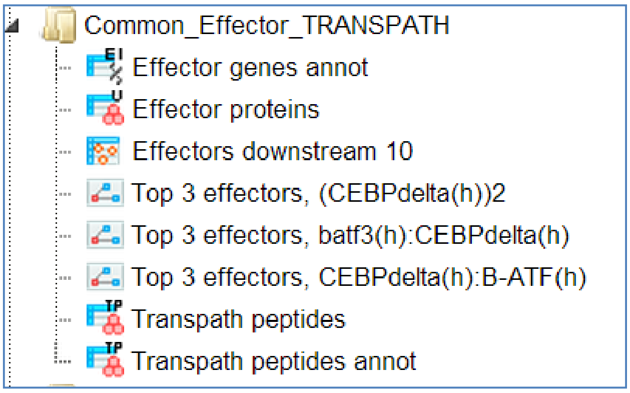


Start page
Find common effectors in n...
Filtered effector proteins
Effector genes annot
Find common effectors in networks (TRANSPATH(R))
Input gene set
...ressed genes/Upregulated Ensembl genes
Species
Human (Homo sapiens)
Results folder
...com/Data/documentation/Common_Effector_TP
Auto
Run workflow
Edit workflow
Completed

```
INFO - * Proteins: UniProt
INFO - Matching...
INFO - Matching...
INFO - Visualization of regulator networks started
INFO - Matched rows: 902
INFO - Unmatched rows: 13
INFO - Preparing row list...
INFO - Resulting rows: 580
INFO - Generating output...
INFO - Analysis 'Convert table' finished (1.58 s)
```

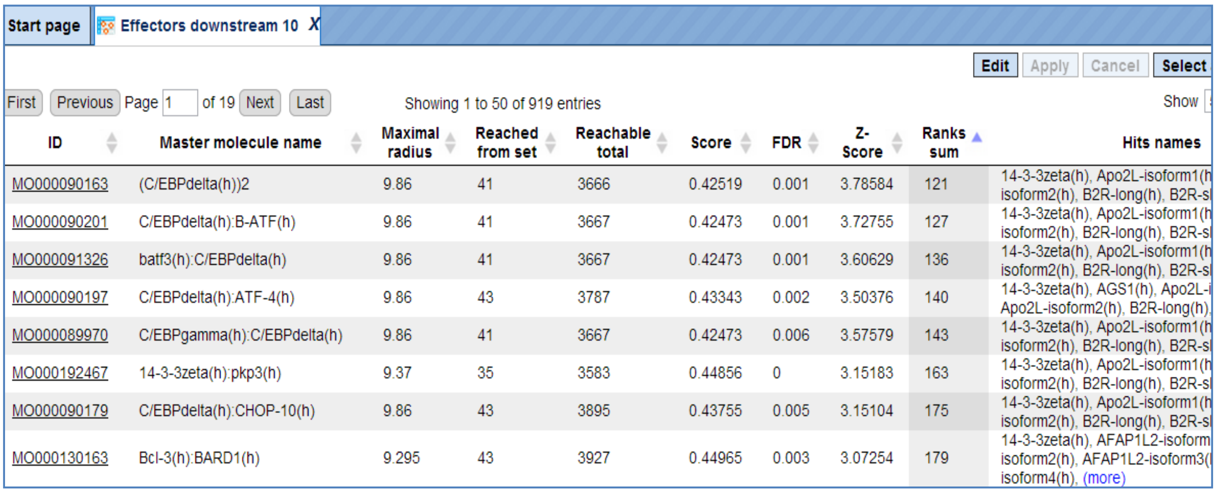
The results folder contains several files; in the example given the path is:

data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/Common Effectors_TP



Common_Effector_TRANSPATH
- Effector genes annot
- Effector proteins
- Effectors downstream 10
- Top 3 effectors, (CEBPdelta(h))2
- Top 3 effectors, batf3(h):CEBPdelta(h)
- Top 3 effectors, CEBPdelta(h):B-ATF(h)
- Transpath peptides
- Transpath peptides annot

The primary result table *Effectors downstream 10* ( ) is the list of effector molecules that were identified at the distance up to 10 steps downstream of the input molecules. Each effector molecule is characterized by Score, Z-score, FDR, and Ranks Sum. Further details about these parameters can be found in the Section 5.1.1.1, under "Interpretation of the results".



Showing 1 to 50 of 919 entries

ID	Master molecule name	Maximal radius	Reached from set	Reachable total	Score	FDR	Z-Score	Ranks sum	Hits names
MO000090163	(C/EBPdelta(h))2	9.86	41	3666	0.42519	0.001	3.78584	121	14-3-3zeta(h), Apo2L-isoform1(h isoform2(h), B2R-long(h), B2R-s
MO000090201	C/EBPdelta(h):B-ATF(h)	9.86	41	3667	0.42473	0.001	3.72755	127	14-3-3zeta(h), Apo2L-isoform1(h isoform2(h), B2R-long(h), B2R-s
MO000091326	batf3(h):C/EBPdelta(h)	9.86	41	3667	0.42473	0.001	3.60629	136	14-3-3zeta(h), Apo2L-isoform1(h isoform2(h), B2R-long(h), B2R-s
MO000090197	C/EBPdelta(h):ATF-4(h)	9.86	43	3787	0.43343	0.002	3.50376	140	14-3-3zeta(h), AGS1(h), Apo2L-i Apo2L-isoform2(h), B2R-long(h),
MO000089970	C/EBPgamma(h):C/EBPdelta(h)	9.86	41	3667	0.42473	0.006	3.57579	143	14-3-3zeta(h), Apo2L-isoform1(h isoform2(h), B2R-long(h), B2R-s
MO000192467	14-3-3zeta(h):pkp3(h)	9.37	35	3583	0.44856	0	3.15183	163	14-3-3zeta(h), Apo2L-isoform1(h isoform2(h), B2R-long(h), B2R-s
MO000090179	C/EBPdelta(h):CHOP-10(h)	9.86	43	3895	0.43755	0.005	3.15104	175	14-3-3zeta(h), Apo2L-isoform1(h isoform2(h), B2R-long(h), B2R-s
MO000130163	Bcl-3(h):BARD1(h)	9.295	43	3927	0.44965	0.003	3.07254	179	14-3-3zeta(h), AFAP1L2-isoform isoform2(h), AFAP1L2-isoform3(isoform4(h), (more)

The column **Reached from set** shows the number of molecules from the input list from which the respective effector molecule can be reached.

The column **Reachable total** gives the total number of molecules from which the respective effector molecule can be reached, independent of the input list.

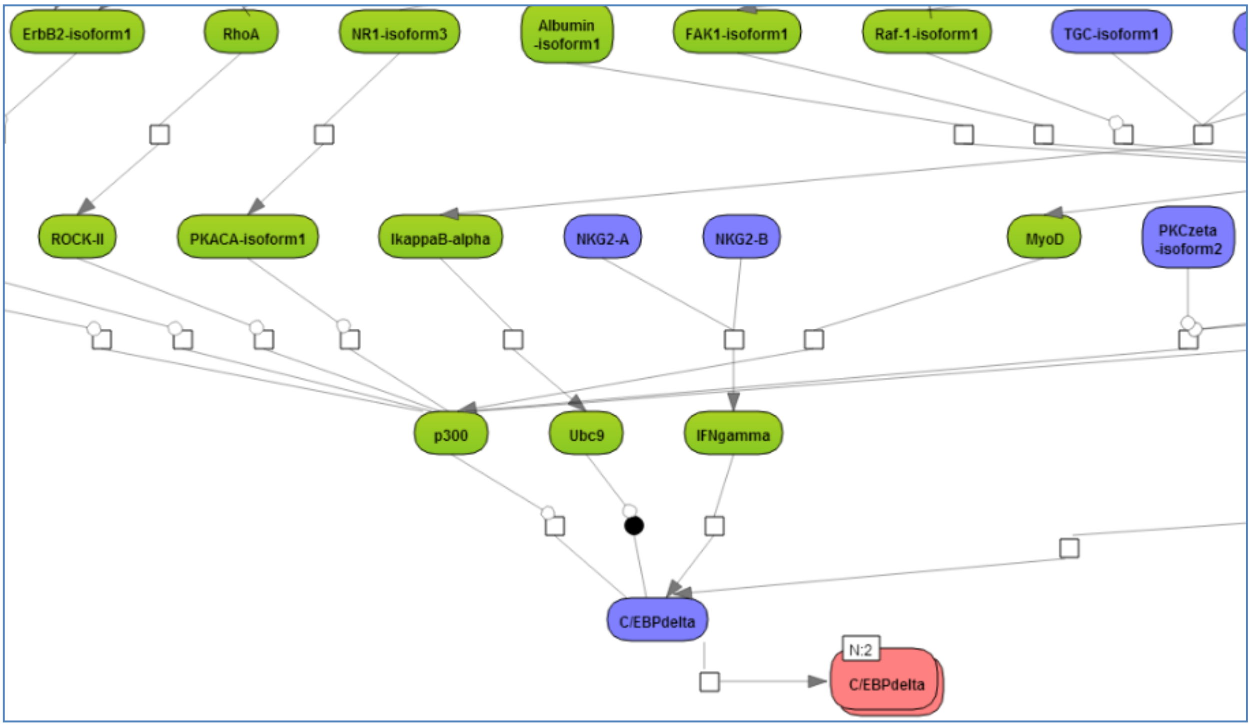
Having this table opened in the Work Space you can find additional options available, specific for this kind of table. Select one or several rows in the table *Effectors downstream 10* by mouse click, and you can visualize the network of the selected Effectors ( ), save the network as a list of genes in the Tree Area ( ), or save hits of this network from the column **Hits** as a list of genes in the Tree Area ( ).

The table *Effector genes annot* ( ) corresponds to the table *Effectors downstream 10* converted into Ensembl IDs and additionally annotated with gene symbols and gene descriptions.

The table *Effector proteins* ( ) corresponds to the table *Effectors downstream 10* converted into the UniProt IDs.

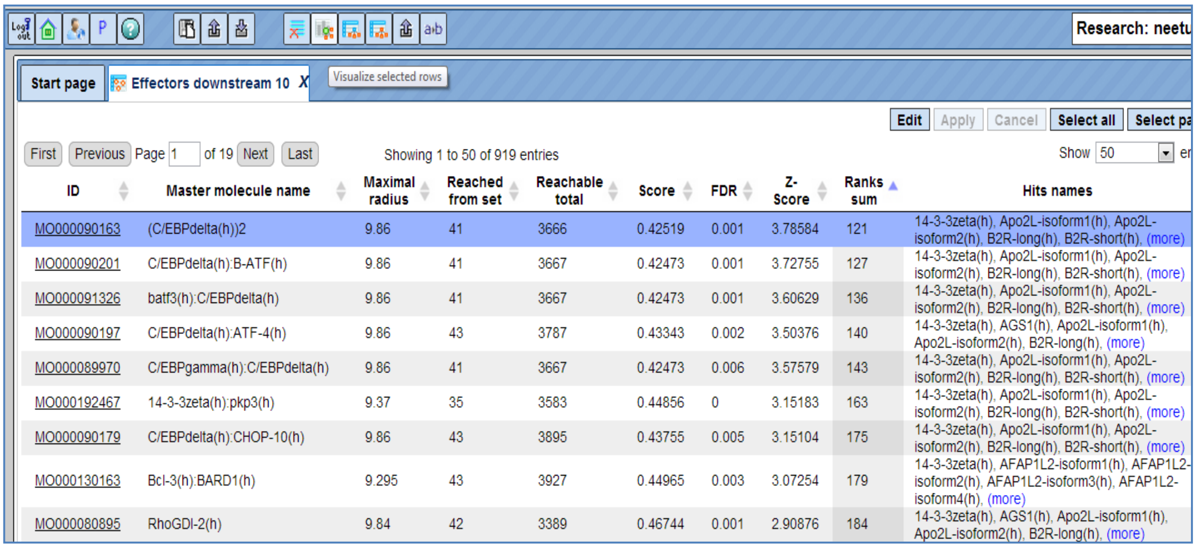
The table *Transpath peptides* ( ) corresponds to the table *Effectors downstream 10* converted into TRANSPATH® molecule IDs, and in the table *Transpath peptides annot* they are further annotated with gene symbols and gene descriptions.

The three diagrams *Top 3 effectors* ( ) visualize networks individually for each of the three top effector molecules. By default, the top effectors are identified upon sorting the table *Effectors downstream 10* ( ) by the column **Ranks sum** with the lowest rank on top.



The default color code for the molecules is the following:
blue: molecules from the input list
red: master regulatory molecules;
green: connecting molecules considered by the graph-analyzing algorithm to find the path from input list to the master molecule.

If you are interested in visualizing the network for any other effector molecule, you may do so in the following way. Open the table *Effectors downstream 10* and select a row with a single mouse click as shown below. Click on menu button to visualize the selected row and save the new diagram into the tree.



Showing 1 to 50 of 919 entries

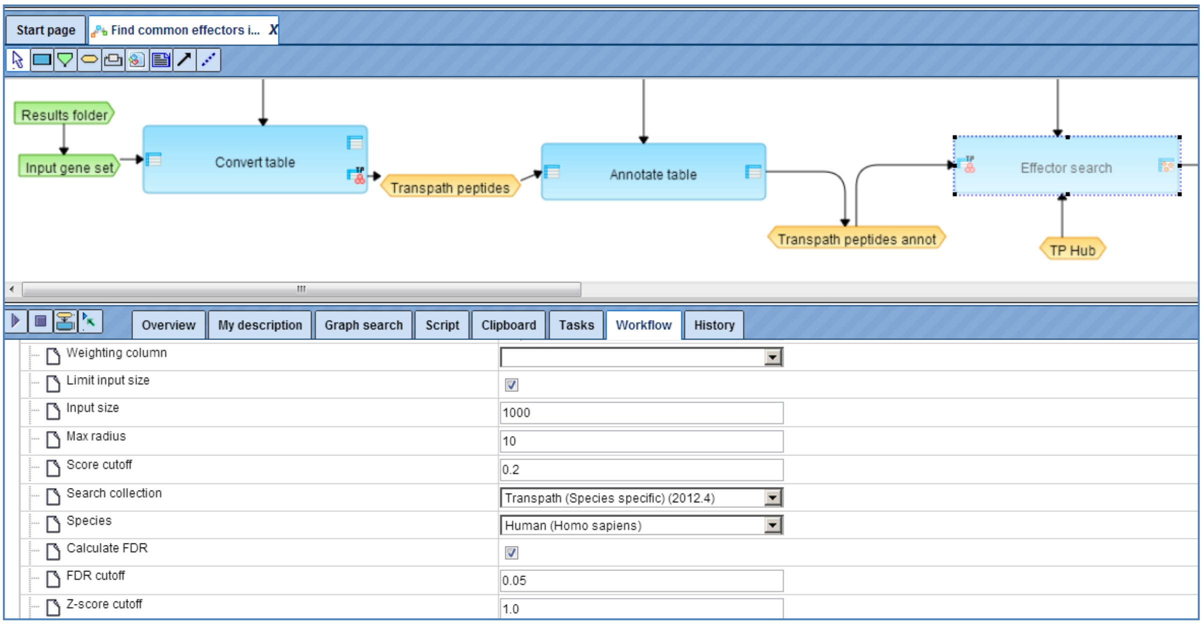
ID	Master molecule name	Maximal radius	Reached from set	Reachable total	Score	FDR	Z-Score	Ranks sum	Hits names
MO000090163	(C/EBPdelta(h))2	9.86	41	3666	0.42519	0.001	3.78584	121	14-3-3zeta(h), Apo2L-isoform1(h), Apo2L-isoform2(h), B2R-long(h), B2R-short(h), (more)
MO000090201	C/EBPdelta(h):B-ATF(h)	9.86	41	3667	0.42473	0.001	3.72755	127	14-3-3zeta(h), Apo2L-isoform1(h), Apo2L-isoform2(h), B2R-long(h), B2R-short(h), (more)
MO000091326	batf3(h):C/EBPdelta(h)	9.86	41	3667	0.42473	0.001	3.60629	136	14-3-3zeta(h), Apo2L-isoform1(h), Apo2L-isoform2(h), B2R-long(h), B2R-short(h), (more)
MO000090197	C/EBPdelta(h):ATF-4(h)	9.86	43	3787	0.43343	0.002	3.50376	140	14-3-3zeta(h), AGS1(h), Apo2L-isoform1(h), Apo2L-isoform2(h), B2R-long(h), (more)
MO000089970	C/EBPgamma(h):C/EBPdelta(h)	9.86	41	3667	0.42473	0.006	3.57579	143	14-3-3zeta(h), Apo2L-isoform1(h), Apo2L-isoform2(h), B2R-long(h), B2R-short(h), (more)
MO000192467	14-3-3zeta(h):pkp3(h)	9.37	35	3583	0.44856	0	3.15183	163	14-3-3zeta(h), Apo2L-isoform1(h), Apo2L-isoform2(h), B2R-long(h), B2R-short(h), (more)
MO000090179	C/EBPdelta(h):CHOP-10(h)	9.86	43	3895	0.43755	0.005	3.15104	175	14-3-3zeta(h), Apo2L-isoform1(h), Apo2L-isoform2(h), B2R-long(h), B2R-short(h), (more)
MO000130163	Bcl-3(h):BARD1(h)	9.295	43	3927	0.44965	0.003	3.07254	179	14-3-3zeta(h), AFAP1L2-isoform1(h), AFAP1L2-isoform2(h), AFAP1L2-isoform3(h), AFAP1L2-isoform4(h), (more)
MO000080895	RhoGDI-2(h)	9.84	42	3389	0.46744	0.001	2.90876	184	14-3-3zeta(h), AGS1(h), Apo2L-isoform1(h), Apo2L-isoform2(h), B2R-long(h), (more)

Tip for the workflow editing

You can easily create a similar workflow with parameter values adjusted to your needs. For example, you might be interested in changing the number of steps used for the effector search. By default, 10 steps are applied.

To make a change, you need first to open the workflow in the “Edit workflow” mode, and save its copy in your project area. The [Edit workflow] button is located near the button [Run workflow] (see above, Step 1). Upon clicking on [Edit workflow], the workflow diagram will be opened in the work area, and you can select one of the analyses you would like to modify. For the screenshot below "Effector Search" analysis was selected, and in the Operations Field, on the tab "Workflow", all parameters are visible. Under this mode, you can modify default parameters and then save the workflow.

In this way you will get a customized workflow, with the parameters adapted to your needs.



Weighting column	
Limit input size	☑
Input size	1000
Max radius	10
Score cutoff	0.2
Search collection	Transpath (Species specific) (2012.4)
Species	Human (Homo sapiens)
Calculate FDR	☑
FDR cutoff	0.05
Z-score cutoff	1.0

Note. This workflow is available together with a valid TRANSPATH® license. Please, feel free to ask for details (info@genexplain.com).

5.1.2.2. Find common effectors with GeneWays

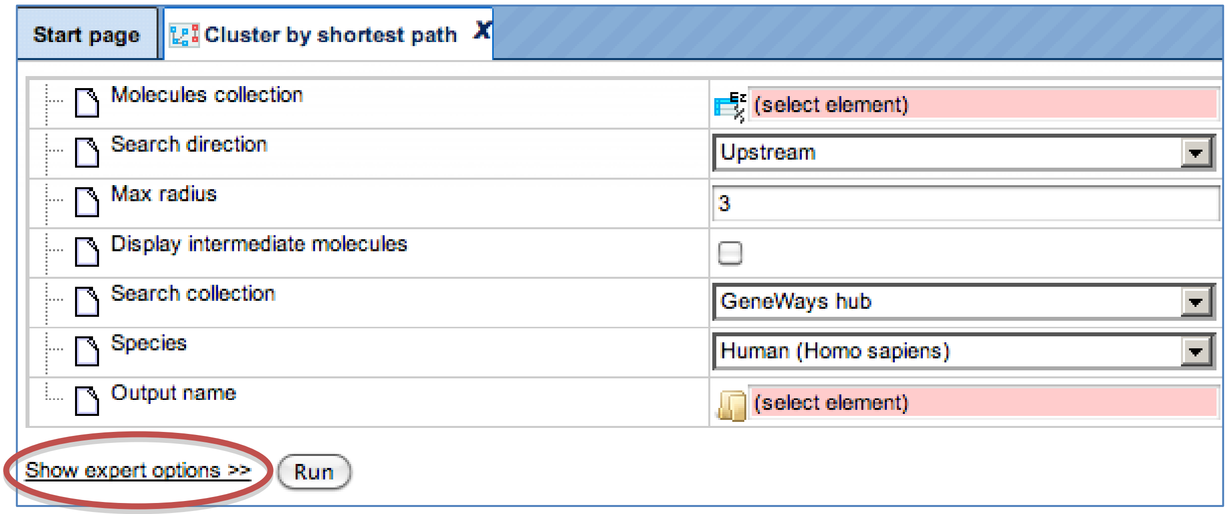
This workflow is designed to find important effector molecules in signal transduction pathways. Here, a search for effector molecules is done based on the network of the GeneWays database with a maximum radius of 4 steps upstream of an input gene set, default cutoffs for Score at 0.2, for FDR at 0.05, and for Z-score at 1.0. The Input form and the resulting tables are very similar to the workflow described above, "Find common effectors in networks (TRANSPATH®)", please refer to Section 5.1.2.1.

The major difference between these two workflows is the underlying database applied for the network analysis, either TRANSPATH® or GeneWays.

More details about the GeneWays and TRANSPATH® databases can be found in the Sections 19.5 and 19.7, respectively.

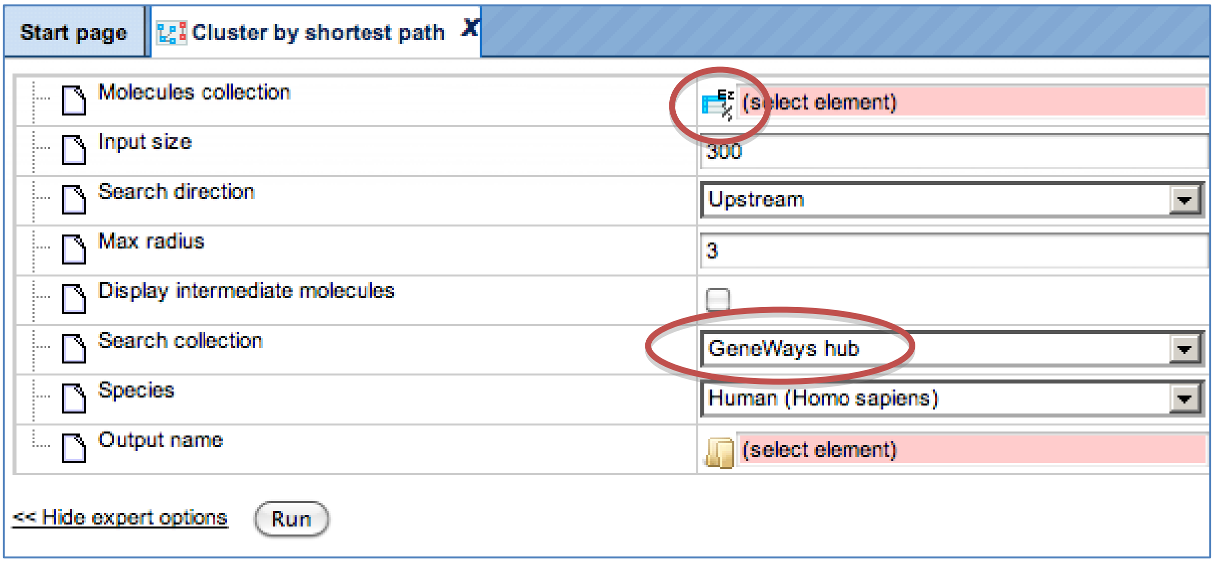
5.1.3. Identify functional protein cluster by shortest path analysis

This analysis finds functional clusters in any input table of genes or proteins. It can be found under the tab *Analyses*, in the folder Methods/Data manipulation/ Molecular networks/Cluster by shortest path ( ). Here the default input form is shown:



Molecules collection	(select element)
Search direction	Upstream
Max radius	3
Display intermediate molecules	□
Search collection	GeneWays hub
Species	Human (Homo sapiens)
Output name	(select element)

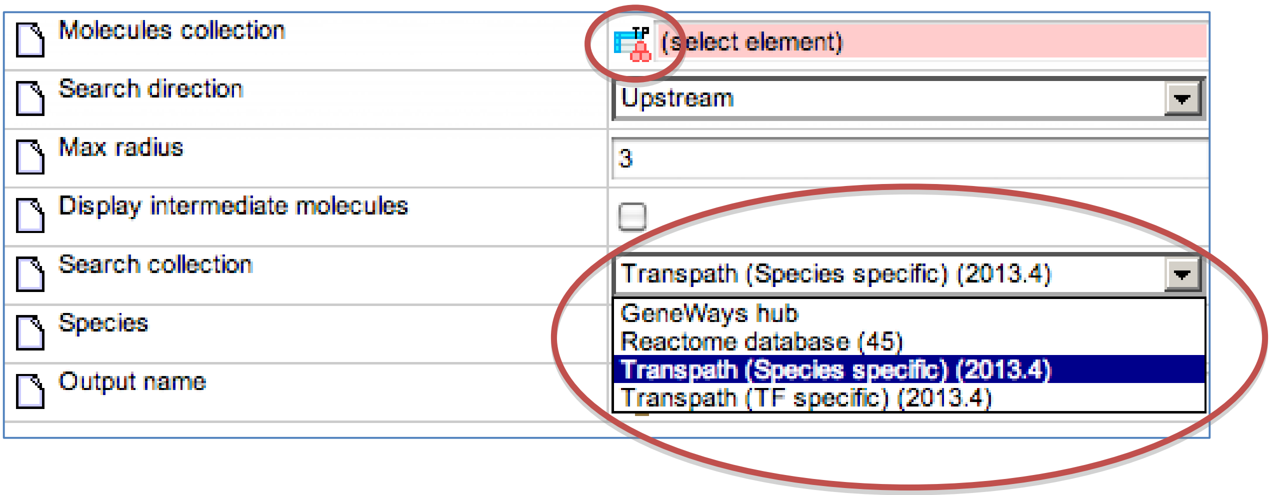
When the expert options are opened, an additional field *Input size* appears, and the form looks like:



Molecules collection	(select element)
Input size	300
Search direction	Upstream
Max radius	3
Display intermediate molecules	□
Search collection	GeneWays hub
Species	Human (Homo sapiens)
Output name	(select element)

In the following, we will consider the input fields one by one.

Search Collection. First, decide which database/search collection you want to use. The connections between the molecules from the specified database will be considered by the clustering algorithm to find the clusters in the input table. You can choose a search collection from the drop-down menu, as shown on the screenshot below. Four search collections are available: GeneWays, Reactome, TRANSPATH® (Species specific) and TRANSPATH® (TF specific).



Molecules collection	(select element)
Search direction	Upstream
Max radius	3
Display intermediate molecules	□
Search collection	Transpath (Species specific) (2013.4)
Species	
Output name	

GeneWays hub
Reactome database (45)
Transpath (Species specific) (2013.4)
Transpath (TF specific) (2013.4)

If you are interested in applying TRANSPATH®, you have to choose either the *Species specific* or the *TF specific* collection. This choice depends on the input table. If you search for clusters among transcription factors, and your input table is a table of transcription factors, it is recommended to choose TRANSPATH® (TF specific). If your input table contains different genes/proteins, TFs and/or other functional groups, it is recommended to choose TRANSPATH® (Species specific).

By default the GeneWays database is applied. Here, the TRANSPATH® (Species specific) collection is chosen.

Molecules collection. Input the collection of molecules/genes for which you wish to find clusters. The input table type depends on the specified search collection. In case of the Geneways database as search collection, the input type should be table with Entrez gene

IDs (EZ). In case of TRANSPATH®, the input table should be a table with TRANSPATH® peptides (TP). In case of Reactome, the input table should be a table having Reactome protein IDs (R). As soon as the search collection is specified, the icon for the required table type is automatically shown in the field **Molecules collection**, as shown in the two screenshots above for TRANSPATH® and GeneWays, respectively.

Tip Before input your table, check which kind of IDs this table has. If necessary, convert your table into any of these formats. You can use the *Convert table* method as mentioned in Section 16.1.2.

Search Direction. Select the direction which the algorithm should consider to find connections between the input molecules. It can be upstream of your input molecules, or downstream, or in both directions. By default the analysis searches in the upstream direction.

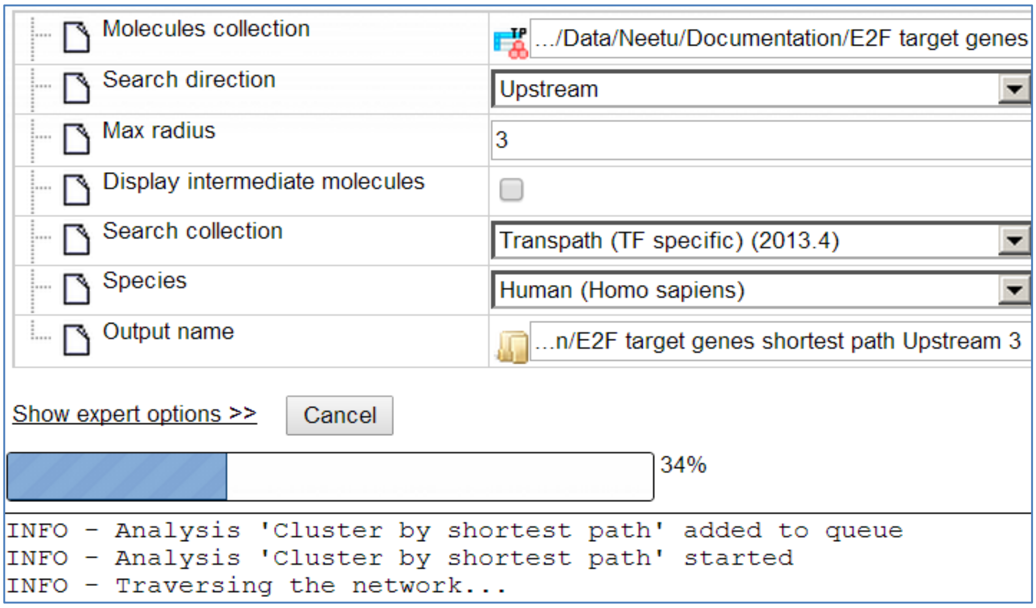
Max radius. Maximum number of steps which the algorithm should consider in the specified direction. By default the number is 3.

Display intermediate molecules. By default this option is unchecked, and in the resulting clusters only the molecules from the input set will be shown. If you prefer the intermediate molecules to be displayed as well, check this box.

Species. Specify human, mouse or rat species corresponding to the input table.

Output name. Specify the path to store the results and the name of the output folder.

Having filled the input form, launch the analysis with the [Run] button. Analysis progresses as shown below. Wait till the analysis is completed.

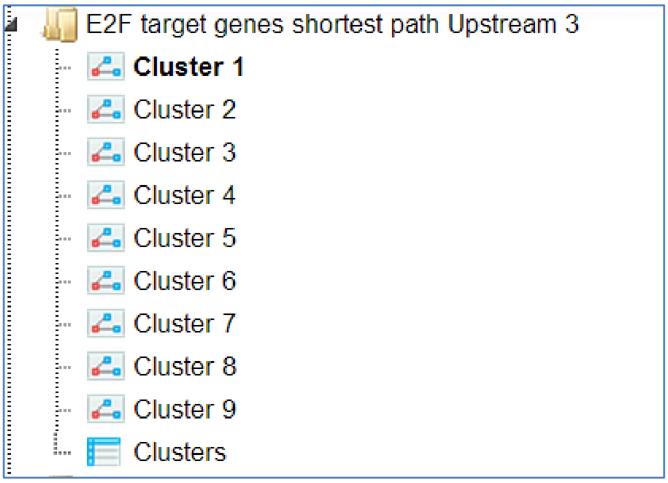


Molecules collection	TP .../Data/Neetu/Documentation/E2F target genes
Search direction	Upstream
Max radius	3
Display intermediate molecules	□
Search collection	Transpath (TF specific) (2013.4)
Species	Human (Homo sapiens)
Output name	...n/E2F target genes shortest path Upstream 3

```
INFO - Analysis 'Cluster by shortest path' added to queue
INFO - Analysis 'Cluster by shortest path' started
INFO - Traversing the network...
```

Results

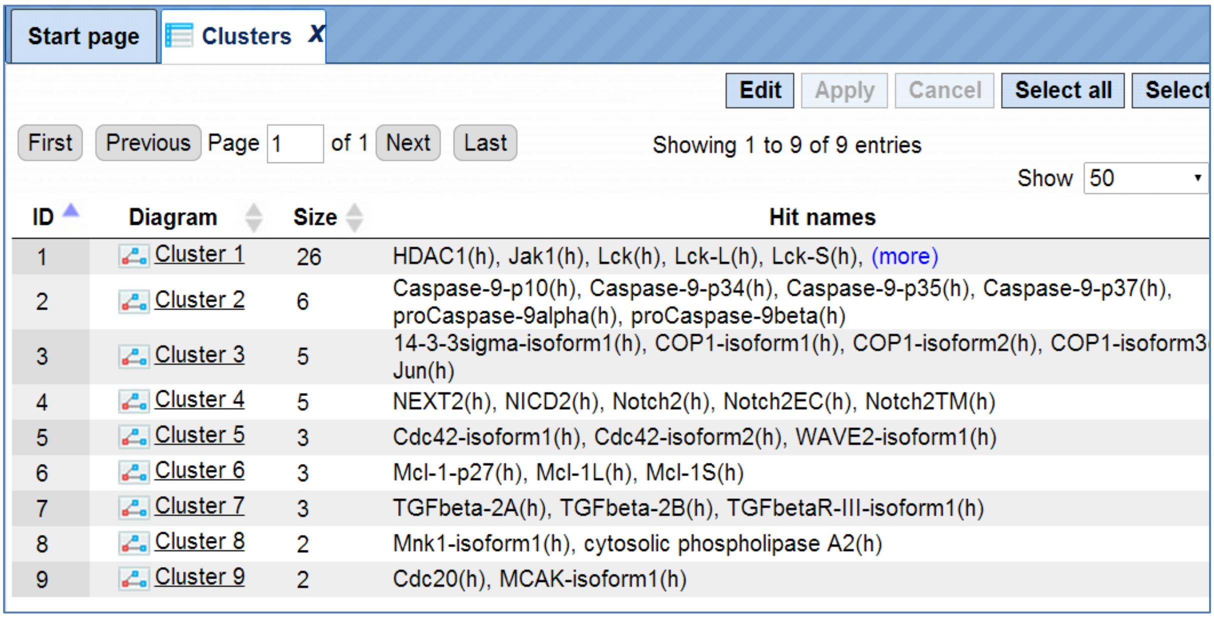
As a result of this analysis, a folder with a specified name is formed, in this example: E2F target genes shortest path Upstream 3, shown below. This folder contains one table, which represents the list of all identified clusters, and several diagrams corresponding to the number of the identified clusters.



E2F target genes shortest path Upstream 3

- **Cluster 1**
- Cluster 2
- Cluster 3
- Cluster 4
- Cluster 5
- Cluster 6
- Cluster 7
- Cluster 8
- Cluster 9
- Clusters

The table *Clusters* (▤) contains a list of all identified clusters, here 9, shown below. Each row shows details for one cluster. The clusters are sorted by their size with the largest cluster on top. The symbol ⧉ next to each cluster name in the column **Diagram** can be used for visualization. The column **Hit names** contains the names of the TRANSPATH® proteins in each cluster.

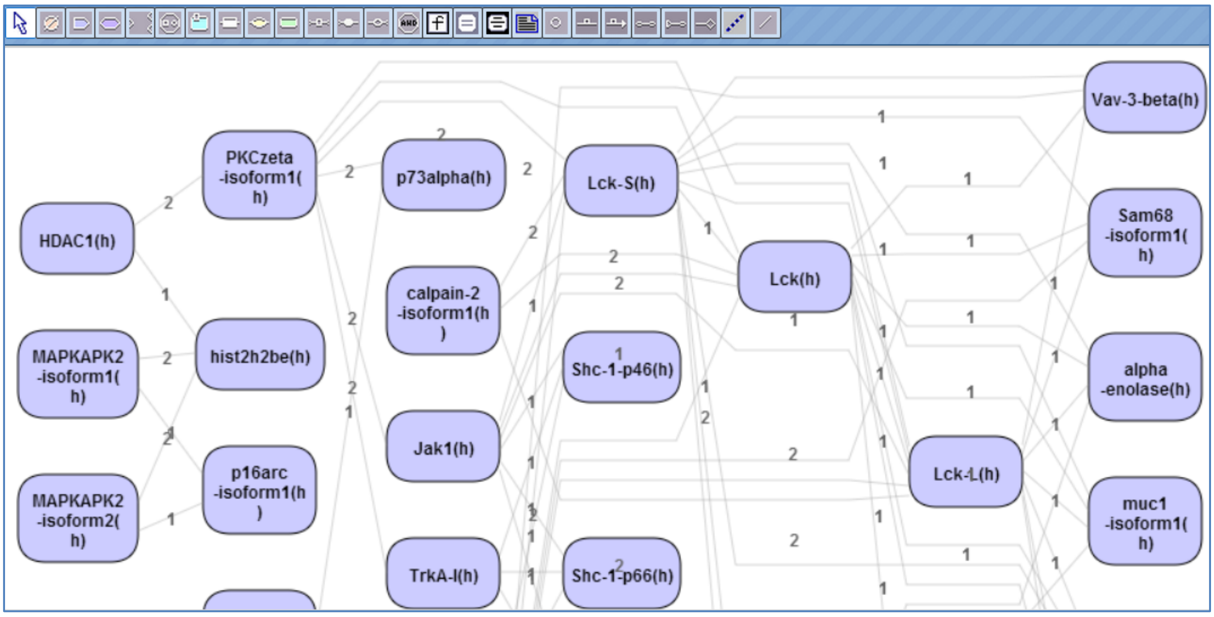


Showing 1 to 9 of 9 entries

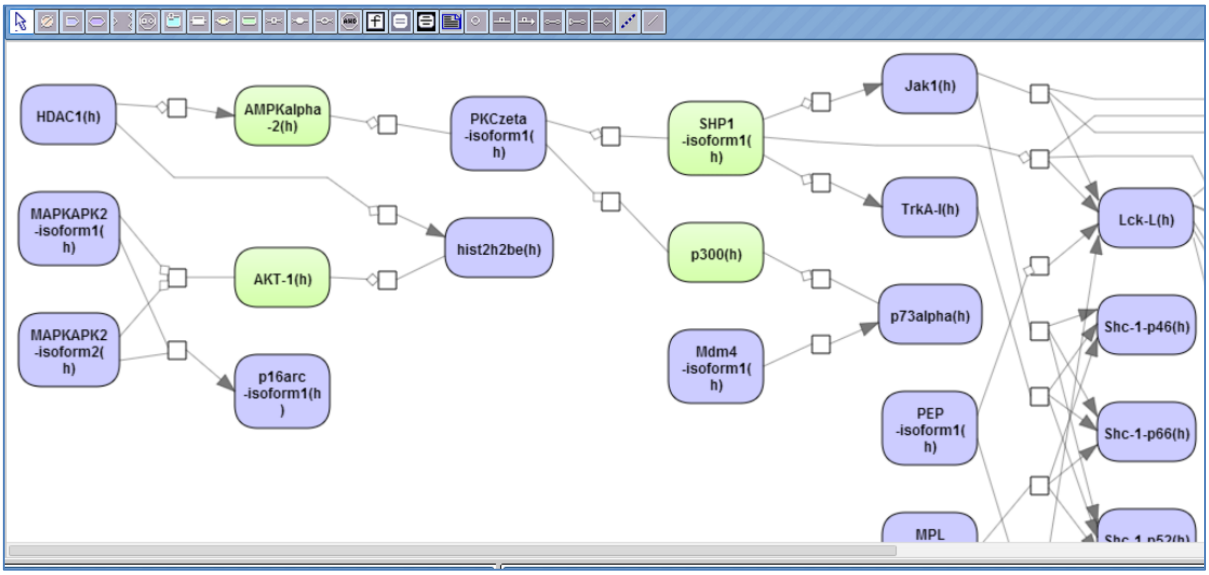
Show 50

ID	Diagram	Size	Hit names
1	Cluster 1	26	HDAC1(h), Jak1(h), Lck(h), Lck-L(h), Lck-S(h), (more)
2	Cluster 2	6	Caspase-9-p10(h), Caspase-9-p34(h), Caspase-9-p35(h), Caspase-9-p37(h), proCaspase-9alpha(h), proCaspase-9beta(h)
3	Cluster 3	5	14-3-3sigma-isoform1(h), COP1-isoform1(h), COP1-isoform2(h), COP1-isoform3 Jun(h)
4	Cluster 4	5	NEXT2(h), NICD2(h), Notch2(h), Notch2EC(h), Notch2TM(h)
5	Cluster 5	3	Cdc42-isoform1(h), Cdc42-isoform2(h), WAVE2-isoform1(h)
6	Cluster 6	3	Mcl-1-p27(h), Mcl-1L(h), Mcl-1S(h)
7	Cluster 7	3	TGFbeta-2A(h), TGFbeta-2B(h), TGFbetaR-III-isoform1(h)
8	Cluster 8	2	Mnk1-isoform1(h), cytosolic phospholipase A2(h)
9	Cluster 9	2	Cdc20(h), MCAK-isoform1(h)

The visualization of *Cluster 1* is shown below. The box **Display intermediate molecules** was unchecked, default setting. All the molecules shown are coming from the input gene/protein set. The numbers on the arrows correspond to the number of steps between two molecules.



When the box **Display intermediate molecules** is checked, the cluster is displayed as shown below. Molecules shown in blue color are coming from the input gene/protein list, and those in green are added by the algorithm when necessary for the connectivity between the input molecules. These green molecules are so-called intermediate molecules.

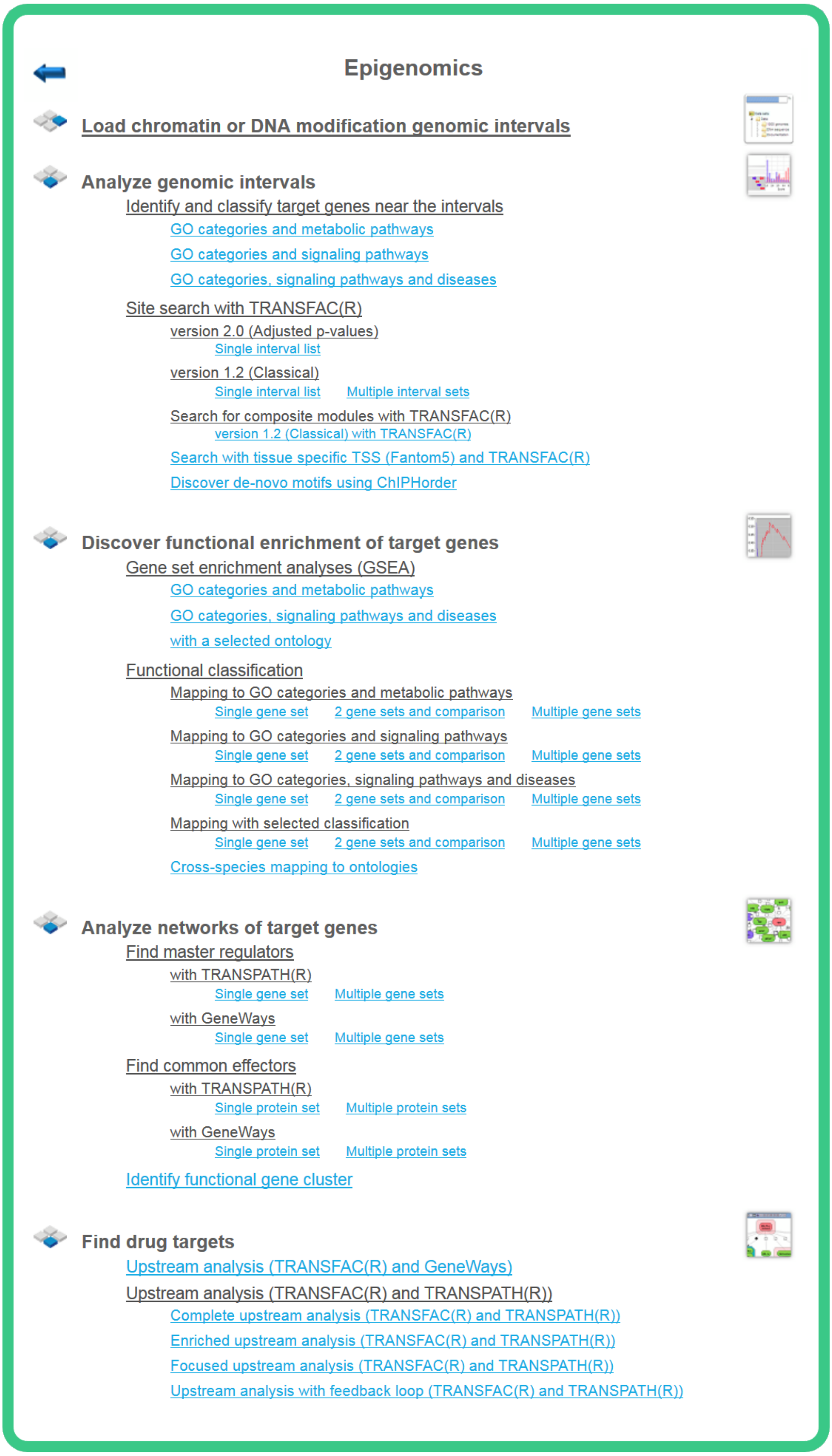


5.2. Further workflows in this area

For the other workflows that you can find in the area ***Proteomics***, please refer to the following Sections:

Load protein list	See Chapter 3
Discover functional enrichment of DEG	See Section 10.3
Find drug targets	See Chapter 11

6. Epigenomics



Epigenomics

Load chromatin or DNA modification genomic intervals

Analyze genomic intervals

- Identify and classify target genes near the intervals
  - GO categories and metabolic pathways
  - GO categories and signaling pathways
  - GO categories, signaling pathways and diseases
- Site search with TRANSFAC(R)
  - version 2.0 (Adjusted p-values)
    - Single interval list
  - version 1.2 (Classical)
    - Single interval list
    - Multiple interval sets
  - Search for composite modules with TRANSFAC(R)
    - version 1.2 (Classical) with TRANSFAC(R)
  - Search with tissue specific TSS (Fantom5) and TRANSFAC(R)
  - Discover de-novo motifs using ChIPHorder

Discover functional enrichment of target genes

- Gene set enrichment analyses (GSEA)
  - GO categories and metabolic pathways
  - GO categories, signaling pathways and diseases
  - with a selected ontology
- Functional classification
  - Mapping to GO categories and metabolic pathways
    - Single gene set
    - 2 gene sets and comparison
    - Multiple gene sets
  - Mapping to GO categories and signaling pathways
    - Single gene set
    - 2 gene sets and comparison
    - Multiple gene sets
  - Mapping to GO categories, signaling pathways and diseases
    - Single gene set
    - 2 gene sets and comparison
    - Multiple gene sets
  - Mapping with selected classification
    - Single gene set
    - 2 gene sets and comparison
    - Multiple gene sets
  - Cross-species mapping to ontologies

Analyze networks of target genes

- Find master regulators
  - with TRANSPATH(R)
    - Single gene set
    - Multiple gene sets
  - with GeneWays
    - Single gene set
    - Multiple gene sets
- Find common effectors
  - with TRANSPATH(R)
    - Single protein set
    - Multiple protein sets
  - with GeneWays
    - Single protein set
    - Multiple protein sets
- Identify functional gene cluster

Find drug targets

- Upstream analysis (TRANSFAC(R) and GeneWays)
- Upstream analysis (TRANSFAC(R) and TRANSPATH(R))
  - Complete upstream analysis (TRANSFAC(R) and TRANSPATH(R))
  - Enriched upstream analysis (TRANSFAC(R) and TRANSPATH(R))
  - Focused upstream analysis (TRANSFAC(R) and TRANSPATH(R))
  - Upstream analysis with feedback loop (TRANSFAC(R) and TRANSPATH(R))

6.1. Analyze genomic intervals

6.1.1. Identify and classify target genes near the intervals

For the three workflows subsumed under this topic, please refer to Section 7.2.1 and apply the steps explained there correspondingly.

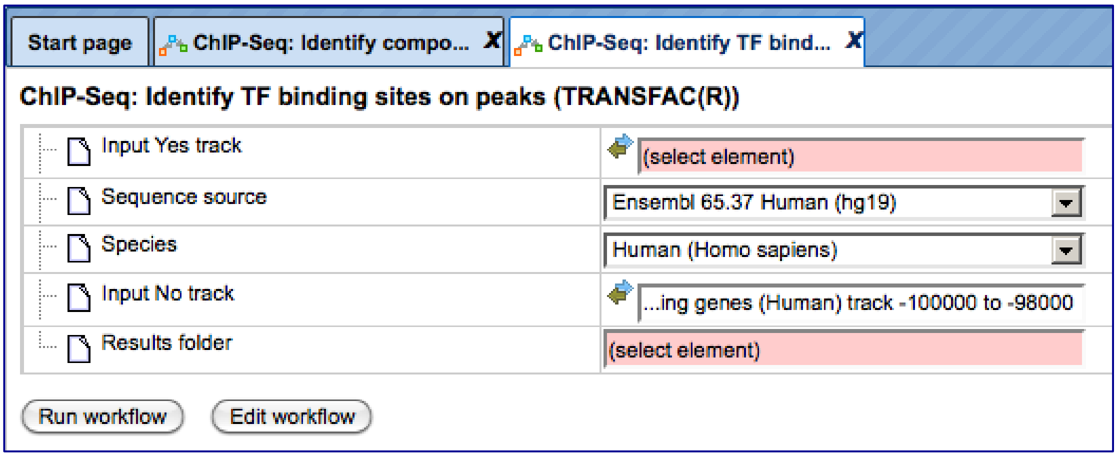
6.1.2. Site search with TRANSFAC®

6.1.2.1. Site search in a single interval list

This workflow helps to map putative TFBSs on peaks calculated from your ChIP-seq data. Site search is done with the help of the TRANSFAC® library of positional weight matrices, PWMs, using the pre-computed profile vertebrate_non_redundant_minSUM.

The few steps to launch the workflow are described in the following.

Step 1. Open workflow input form from the Start page, it will be opened in the main Work Space and looks as it is shown below:

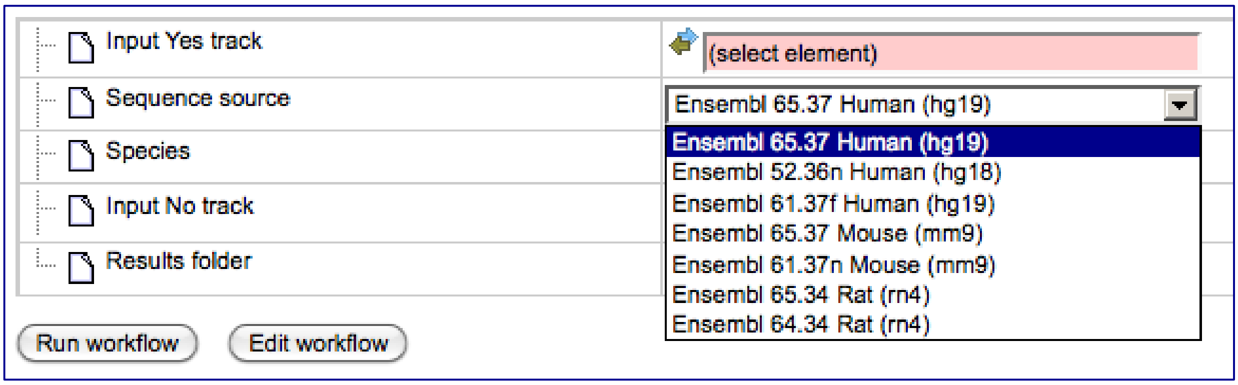


ChIP-Seq: Identify TF binding sites on peaks (TRANSFAC(R))

Input Yes track	(select element)
Sequence source	Ensembl 65.37 Human (hg19)
Species	Human (Homo sapiens)
Input No track	...ing genes (Human) track -100000 to -98000
Results folder	(select element)

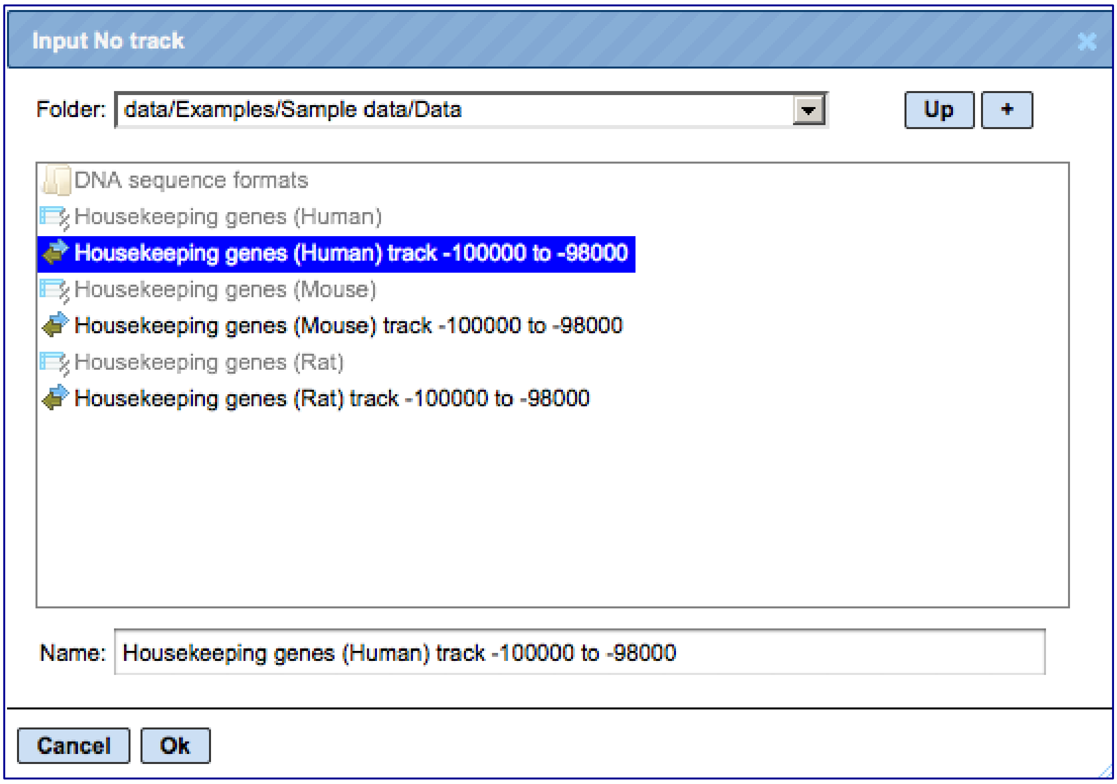
Step 2. Specify the input track in BED format in the field **Input Yes track**. The input Yes track contains peaks from your ChIP-seq study. To specify the Yes track, you can drag & drop it from your project within the tree area. Alternatively, you may click on the pink field “select element” and a new window will open, where you select the input track. After having selected the track, press the [Ok] button.

Step 3. Specify the sequence source from the drop-down menu. Several human, mouse and rat sequence builds are available in the platform, as shown below. By default, the most recent Ensembl human genome, hg19, is specified. Make sure you selected the sequence source (the genome build) that corresponds to your input set, to get correct and meaningful results.


Input Yes track
(select element)
Sequence source
Ensembl 65.37 Human (hg19)
Species
Input No track
Results folder
Ensembl 65.37 Human (hg19)
Ensembl 52.36n Human (hg18)
Ensembl 61.37f Human (hg19)
Ensembl 65.37 Mouse (mm9)
Ensembl 61.37n Mouse (mm9)
Ensembl 65.34 Rat (rn4)
Ensembl 64.34 Rat (rn4)
Run workflow
Edit workflow

Step 4. Specify the biological species of the input set in the field **Species** by selecting the required species from the drop-down menu.

Step 5. Specify No track in BED format in the field **Input No track**. Upon clicking on this field, a supplementary window will open, where you can select the No track from your project tree, or use one of our default No tracks for human, mouse or rat, respectively.


Input No track
Folder: data/Examples/Sample data/Data
Up
+
DNA sequence formats
Housekeeping genes (Human)
Housekeeping genes (Human) track -100000 to -98000
Housekeeping genes (Mouse)
Housekeeping genes (Mouse) track -100000 to -98000
Housekeeping genes (Rat)
Housekeeping genes (Rat) track -100000 to -98000
Name: Housekeeping genes (Human) track -100000 to -98000
Cancel
Ok

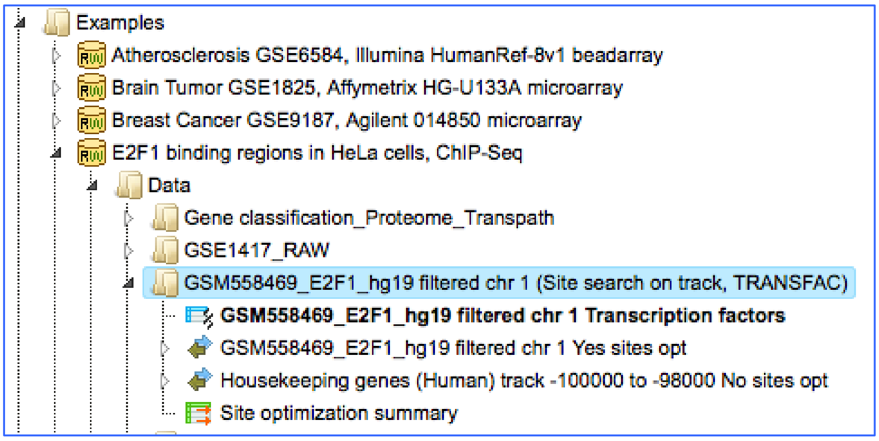
Step 6. Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink field "select element" in the field **Results folder**, and a new window will be opened, where you can select the location of the results folder and define its name.

Step 7. Press the [Run workflow] button.

Ready!

Wait until the workflow is completed.

The results folder contains two tables and two tracks; for this example, let's consider the results folder located under "Examples". It is highlighted by blue in the figure below:



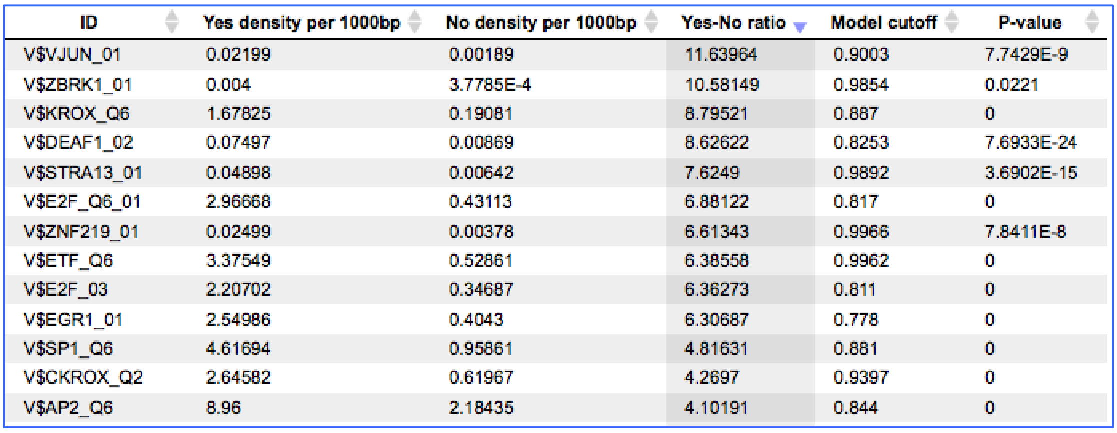
Examples
Atherosclerosis GSE6584, Illumina HumanRef-8v1 beadarray
Brain Tumor GSE1825, Affymetrix HG-U133A microarray
Breast Cancer GSE9187, Agilent 014850 microarray
E2F1 binding regions in HeLa cells, ChIP-Seq
Data
Gene classification_Proteome_Transpath
GSE1417_RAW
GSM558469_E2F1_hg19 filtered chr 1 (Site search on track, TRANSFAC)
GSM558469_E2F1_hg19 filtered chr 1 Transcription factors
GSM558469_E2F1_hg19 filtered chr 1 Yes sites opt
Housekeeping genes (Human) track -100000 to -98000 No sites opt
Site optimization summary

The tables *Site optimization summary* (▦) and *Transcription factors* (▦) are opened automatically in the Work Space as soon as the workflow is completed.

The table *Site optimization summary* includes the matrices the hits of which are over-represented in the Yes track *versus* the No track.

Please note that only the matrices with Yes-No ratio higher than 1 are included in this output table. The hits of these matrices can be interpreted as over-represented in the Yes set *versus* No set.

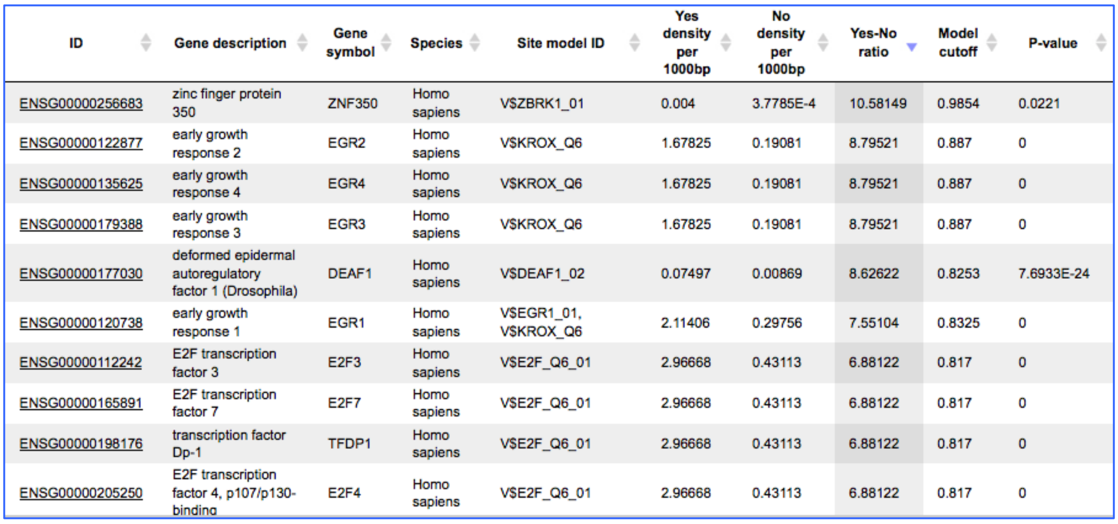
The table *Site optimization summary* shown below has been sorted by the values in the **Yes-No ratio** column.



ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
V$VJUN_01	0.02199	0.00189	11.63964	0.9003	7.7429E-9
V$ZBRK1_01	0.004	3.7785E-4	10.58149	0.9854	0.0221
V$KROX_Q6	1.67825	0.19081	8.79521	0.887	0
V$DEAF1_02	0.07497	0.00869	8.62622	0.8253	7.6933E-24
V$STRA13_01	0.04898	0.00642	7.6249	0.9892	3.6902E-15
V$E2F_Q6_01	2.96668	0.43113	6.88122	0.817	0
V$ZNF219_01	0.02499	0.00378	6.61343	0.9966	7.8411E-8
V$ETF_Q6	3.37549	0.52861	6.38558	0.9962	0
V$E2F_03	2.20702	0.34687	6.36273	0.811	0
V$EGR1_01	2.54986	0.4043	6.30687	0.778	0
V$SP1_Q6	4.61694	0.95861	4.81631	0.881	0
V$CKROX_Q2	2.64582	0.61967	4.2697	0.9397	0
V$AP2_Q6	8.96	2.18435	4.10191	0.844	0

Each row summarizes the information for one PWM. For each selected matrix, the columns **Yes density per 1000bp** and **No density per 1000bp** show the number of matches normalized per 1000 bp length for the sequences in the input Yes set and input No set, respectively. The Column **Yes-No ratio** is the ratio of the first two columns. Only matrices with a Yes-No ratio higher than 1 are included in the *summary* table. The higher the Yes-No ratio, the higher is the enrichment of matches for the respective matrix in the Yes set. The matrix cutoff values as they are calculated by the program at the optimization step are shown in the column **Model cutoff**, and the last column shows the **P-value** of the corresponding event.

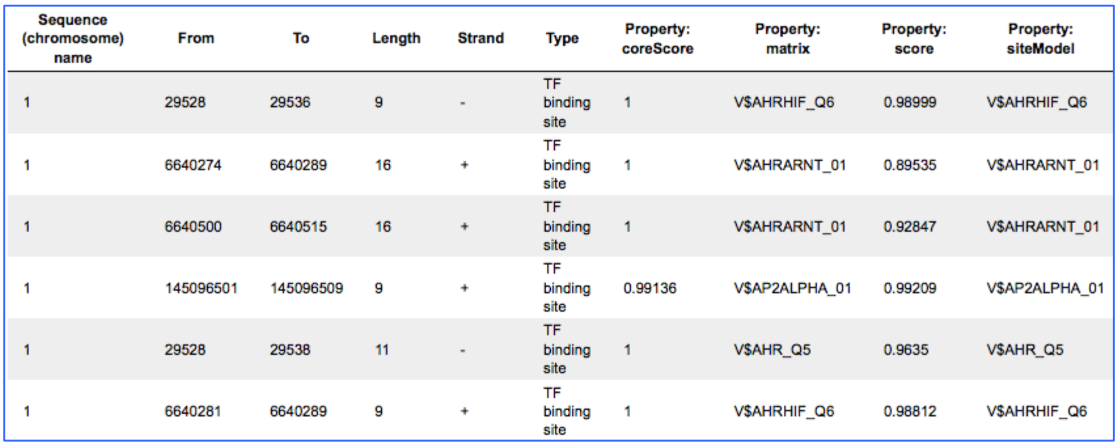
Table Transcription factors:



ID	Gene description	Gene symbol	Species	Site model ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
ENSG00000256683	zinc finger protein 350	ZNF350	Homo sapiens	V$ZBRK1_01	0.004	3.7785E-4	10.58149	0.9854	0.0221
ENSG00000122877	early growth response 2	EGR2	Homo sapiens	V$KROX_Q6	1.67825	0.19081	8.79521	0.887	0
ENSG00000135625	early growth response 4	EGR4	Homo sapiens	V$KROX_Q6	1.67825	0.19081	8.79521	0.887	0
ENSG00000179388	early growth response 3	EGR3	Homo sapiens	V$KROX_Q6	1.67825	0.19081	8.79521	0.887	0
ENSG00000177030	deformed epidermal autoregulatory factor 1 (Drosophila)	DEAF1	Homo sapiens	V$DEAF1_02	0.07497	0.00869	8.62622	0.8253	7.6933E-24
ENSG00000120738	early growth response 1	EGR1	Homo sapiens	V$EGR1_01, V$KROX_Q6	2.11406	0.29756	7.55104	0.8325	0
ENSG00000112242	E2F transcription factor 3	E2F3	Homo sapiens	V$E2F_Q6_01	2.96668	0.43113	6.88122	0.817	0
ENSG00000165891	E2F transcription factor 7	E2F7	Homo sapiens	V$E2F_Q6_01	2.96668	0.43113	6.88122	0.817	0
ENSG00000198176	transcription factor Dp-1	TFDP1	Homo sapiens	V$E2F_Q6_01	2.96668	0.43113	6.88122	0.817	0
ENSG00000205250	E2F transcription factor 4, p107/p130-binding	E2F4	Homo sapiens	V$E2F_Q6_01	2.96668	0.43113	6.88122	0.817	0

This table includes transcription factors (TFs) that are associated with the PWMs that are listed in the table *Site optimization summary*, and each row shows details for one TF, including its Ensembl gene ID (column **ID**), gene symbol, gene description and biological species of the corresponding TF (columns **Gene description**, **Gene symbol**, and **Species**). The column **Site model ID** shows the identifier of the PWM associated with this TF, and several further columns repeat information that is also shown in the table *Site optimization summary*.

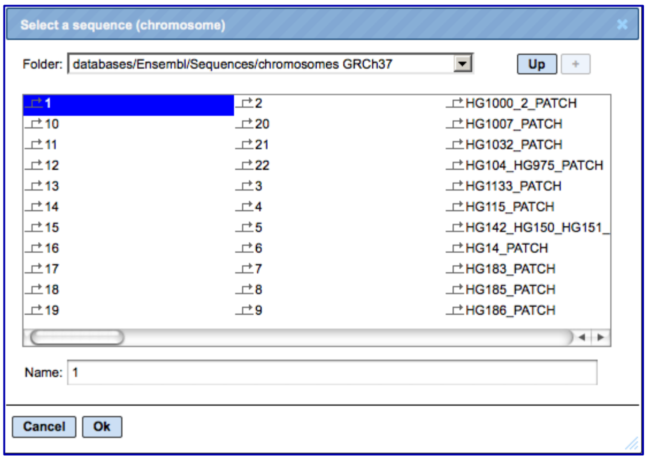
Tracks "Yes sites opt" and "No sites opt" (⇄ *).*



Sequence (chromosome) name	From	To	Length	Strand	Type	Property: coreScore	Property: matrix	Property: score	Property: siteModel
1	29528	29536	9	-	TF binding site	1	V$AHRHIF_Q6	0.98999	V$AHRHIF_Q6
1	6640274	6640289	16	+	TF binding site	1	V$AHRARNT_01	0.89535	V$AHRARNT_01
1	6640500	6640515	16	+	TF binding site	1	V$AHRARNT_01	0.92847	V$AHRARNT_01
1	145096501	145096509	9	+	TF binding site	0.99136	V$AP2ALPHA_01	0.99209	V$AP2ALPHA_01
1	29528	29538	11	-	TF binding site	1	V$AHR_Q5	0.9635	V$AHR_Q5
1	6640281	6640289	9	+	TF binding site	1	V$AHRHIF_Q6	0.98812	V$AHRHIF_Q6

Each row presents details for each individual match for every PWM. Columns **Sequence (chromosome) name**, **From**, **To**, **Length** and **Strand** show the genomic location of the match including chromosome number, start and end positions, strand and length of the match, respectively. The column **Type** contains information about the type of the elements; in this case all matches are considered as "TF binding site". Further columns keep information about PWM producing each match (column **Property:matrix**) as well as a score of the core (column **Property:coreScore**) and a score for the whole matrix (column **Property:score**). The column **Property: siteModel** contains an identifier for the site model, which is the matrix together with the cutoff applied (for details about these scores, please see Kel et al., Nucleic Acids Res. 31:3576-3579, 2003).

Tip. Further visualization of track files in the genome browser: Having tracks "Yes sites opt" and "No sites opt" opened in the Work Space, the menu button can be applied to get a visualization. First, a supplementary window is opened where you can select one chromosome and press [Ok], as shown below.



Select a sequence (chromosome)

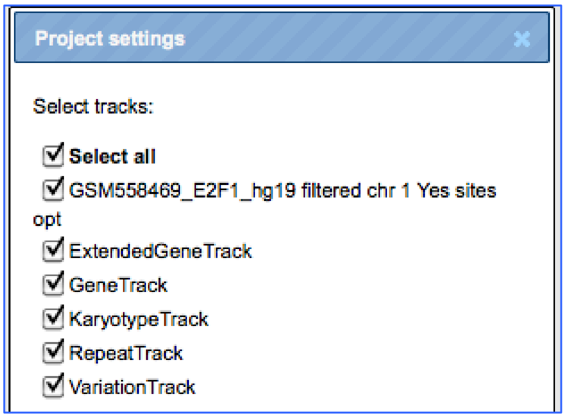
Folder: databases/Ensembl/Sequences/chromosomes GRCh37 Up +

1
10
11
12
13
14
15
16
17
18
19
2
20
21
22
3
4
5
6
7
8
9
HG1000_2_PATCH
HG1007_PATCH
HG1032_PATCH
HG104_HG975_PATCH
HG1133_PATCH
HG115_PATCH
HG142_HG150_HG151_
HG14_PATCH
HG183_PATCH
HG185_PATCH
HG186_PATCH

Name: 1

Cancel Ok

In the second pop-up window, you can select tracks that can be visualized together with your track, e.g. "Yes sites opt" (see above), and press [Ok].

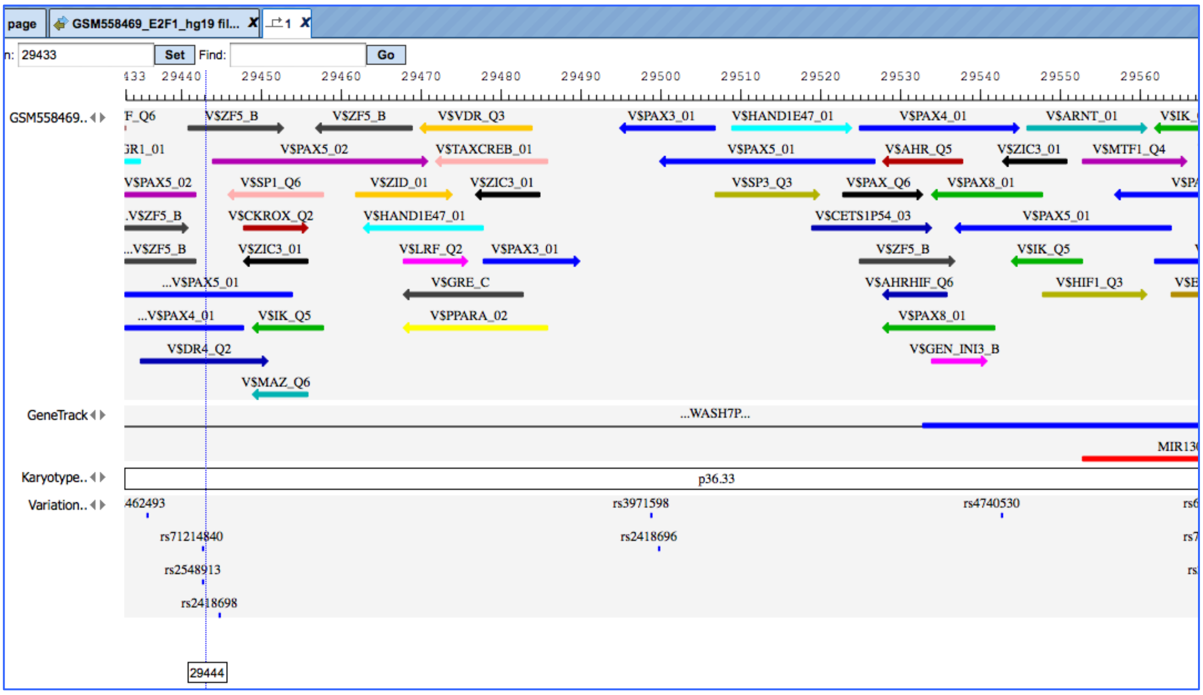


Project settings

Select tracks:

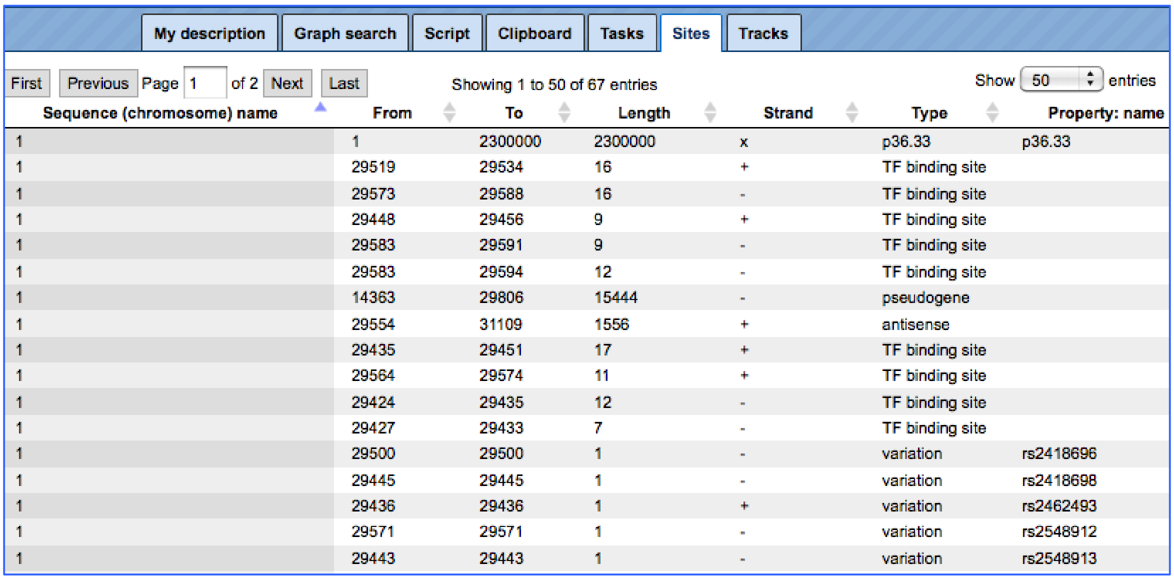
☑ **Select all**
☑ GSM558469_E2F1_hg19 filtered chr 1 Yes sites opt
☑ ExtendedGeneTrack
☑ GeneTrack
☑ KaryotypeTrack
☑ RepeatTrack
☑ VariationTrack

The resulting visualization, after applying the "zoom in" button , looks like it is shown below. Matches for different matrices are shown in colors, and the color schema can be customized.



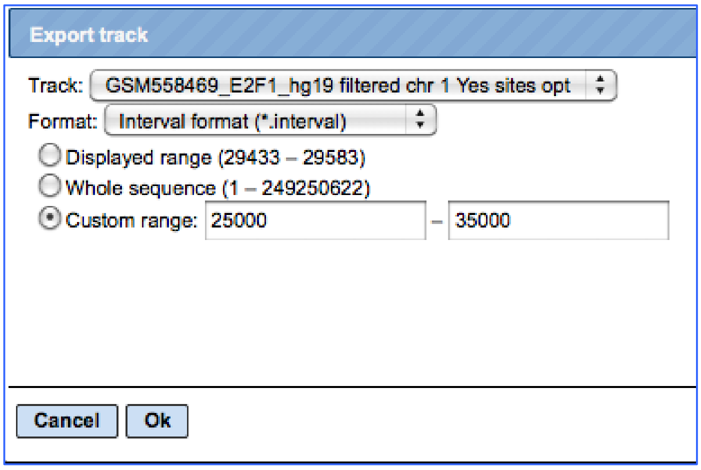
Such a view may help to visually co-localize information on different tracks, e.g. putative TFBS with variations, repeats and genes. In the figure above, the cursor shows position 29444, and two variations are located at this position. You can immediately recognize that these variations are located within particular putative binding sites in the intron region of the WASH7P gene.

The same information is available not just as a picture, but also as a table under the tab “Sites” (shown below). For each element information is shown on chromosome, positions, length, strand, type of the track, and name of the element.



Sequence (chromosome) name	From	To	Length	Strand	Type	Property: name
1	1	2300000	2300000	x	p36.33	p36.33
1	29519	29534	16	+	TF binding site	
1	29573	29588	16	-	TF binding site	
1	29448	29456	9	+	TF binding site	
1	29583	29591	9	-	TF binding site	
1	29583	29594	12	-	TF binding site	
1	14363	29806	15444	-	pseudogene	
1	29554	31109	1556	+	antisense	
1	29435	29451	17	+	TF binding site	
1	29564	29574	11	+	TF binding site	
1	29424	29435	12	-	TF binding site	
1	29427	29433	7	-	TF binding site	
1	29500	29500	1	-	variation	rs2418696
1	29445	29445	1	-	variation	rs2418698
1	29436	29436	1	+	variation	rs2462493
1	29571	29571	1	-	variation	rs2548912
1	29443	29443	1	-	variation	rs2548913

This table can be exported as a track, in several different formats including intervals, bed, wig, gff, gtf and more.



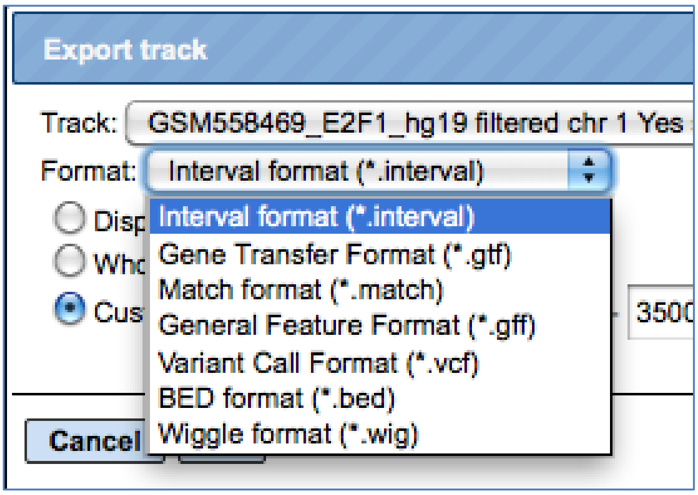
Track: GSM558469_E2F1_hg19 filtered chr 1 Yes sites opt

Format: Interval format (*.interval)

○ Displayed range (29433 – 29583)

○ Whole sequence (1 – 249250622)

● Custom range: 25000 – 35000


Export track
Interval format (*.interval)
Gene Transfer Format (*.gtf)
Match format (*.match)
General Feature Format (*.gff)
Variant Call Format (*.vcf)
BED format (*.bed)
Wiggle format (*.wig)
Disp
Whc
Cus
3500
Cancel

Track: GSM558469_E2F1_hg19 filtered chr 1 Yes

Format: Interval format (*.interval)

Note. This workflow is available together with a valid TRANSFAC® license. Please, feel free to ask for details (info@genexplain.com).

6.1.2.2. Site search in multiple interval sets

This workflow is designed to search for TFBSs in DNA sequences identified by the ChIP-seq approach, for multiple datasets.

In the field **Input Yes tracks**, several different tracks can be simultaneously submitted. The same background dataset, **Input No track**, is used for comparison with each of the submitted Yes tracks. The default No track corresponds to far upstream regions of the house keeping genes, where no functional TFBSs are expected.

The steps of this workflow for a single input Yes track are described in Section 6.1.2.1. In this workflow, the same steps are performed next time for the 2nd Yes track, and so on iteratively for each of the input Yes tracks.

This workflow helps to save time and efforts, especially when you have several sets of ChIP-seq data, e.g. the peaks for a number of different TFs.

Note. This workflow is available together with a valid TRANSFAC® license. Please, feel free to ask for details (info@genexplain.com).

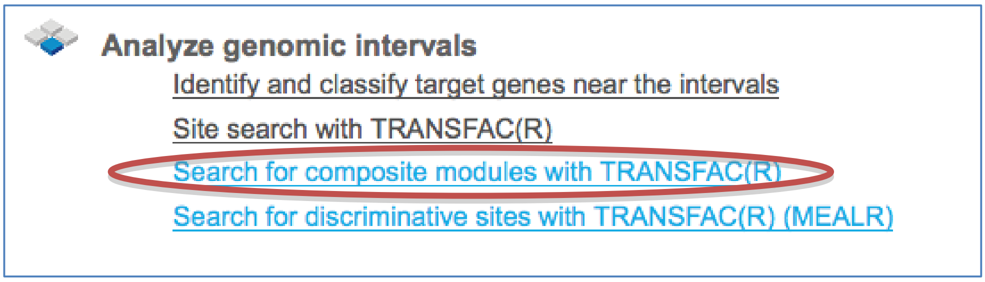
6.1.3. Search for composite modules

6.1.3.1. Search for composite modules with TRANSFAC®

This workflow finds pairs of TFBSs that discriminate between two tracks, the *Yes* and the *No* track. As the Yes track, ChIP-seq peaks or intervals identified in analyses for histone modifications, or any other genomic fragments, can be considered.

In the first part of the workflow, the overrepresented TFBSs are identified by the method *analyses/Methods/Site analysis/Site search on gene set* ( ). Details about this individual method are given in Section 20.1.2. In the second part of this workflow, composite modules are identified by a genetic algorithm based on over-represented TFBSs. For more details about this CMA analysis refer to Section 20.1.5.

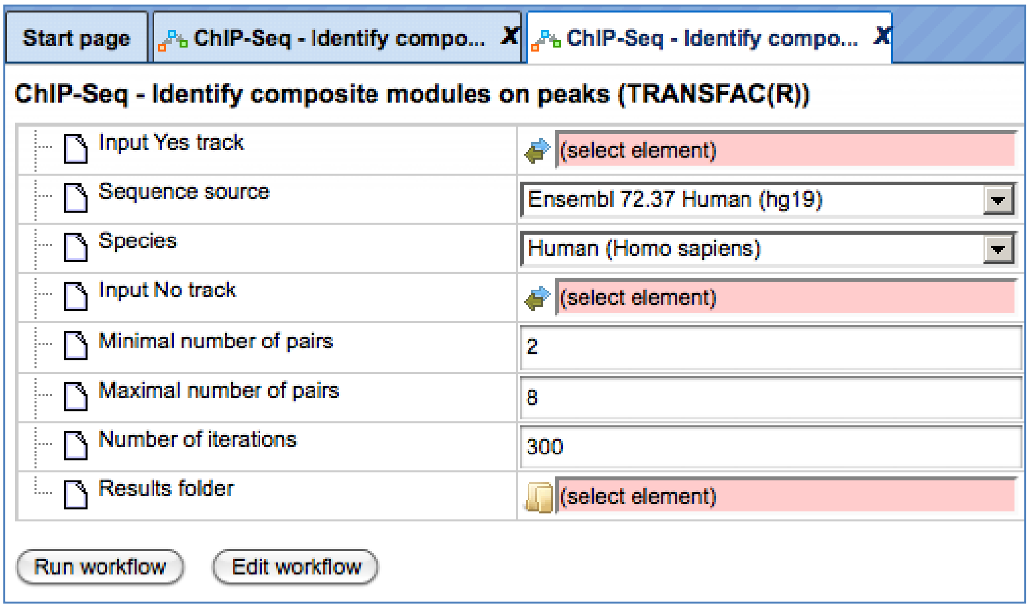
To launch the workflow, follow these steps:



Analyze genomic intervals

- Identify and classify target genes near the intervals
- Site search with TRANSFAC(R)
- Search for composite modules with TRANSFAC(R)
- Search for discriminative sites with TRANSFAC(R) (MEALR)

Step1. Open the workflow input form. It will open in the main Work Space and looks as shown below:



ChIP-Seq - Identify composite modules on peaks (TRANSFAC(R))

Input Yes track	(select element)
Sequence source	Ensembl 72.37 Human (hg19)
Species	Human (Homo sapiens)
Input No track	(select element)
Minimal number of pairs	2
Maximal number of pairs	8
Number of iterations	300
Results folder	(select element)

Step 2. Input the Yes track (the track under study) from the tree. You can either drag-and-drop or select it from the Tree Area. Here the track published in the Gene Expression Omnibus, *GSE54909*, is used as an example.

In this study the HDAC inhibitors TSA and SAHA were used to treat primary human vascular endothelial cells. The effects on genome-wide histone acetylation were studied applying chromatin immunoprecipitation (ChIP) and then deep sequencing (ChIP-seq). Histone acetylation ChIP-seq profiles in SAHA-treated versus control samples were calculated and the peaks were published. The published track contains over 10000 fragments.

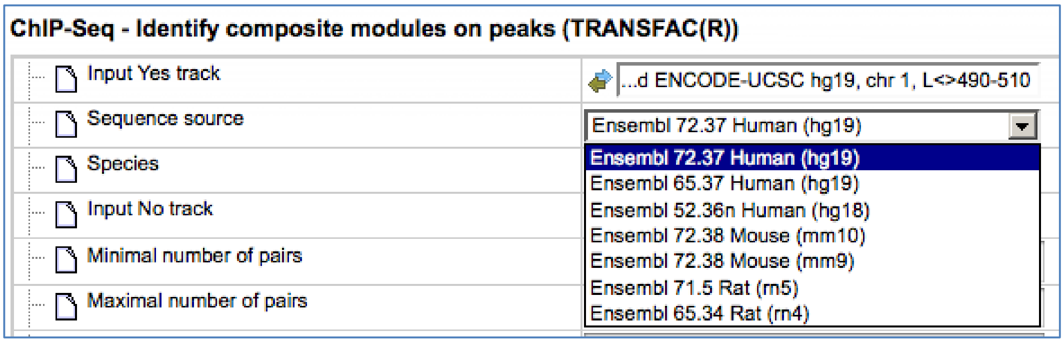
Please consider that such huge tracks cannot be submitted into the workflow. One cannot expect to find common composite modules in 10000 fragments.

Tip We recommend to filter the original track by several conditions, for example by the score of the calculated peaks, by the length of the fragments, by genes located near the peaks, etc. The recommended optimal number of the fragments in the input track is 200-400. Please consider that the time required for the workflow run directly depends on the size of the input tracks.

Here, the original published track was first filtered by the length of the fragments 500-600 bp, which resulted in 816 fragments. These fragments were sorted by the score of the peaks, and the top 200 fragments were selected.

The resulting track of 200 fragments is used as the input Yes track.

Step 3. From the drop-down menu in the field Sequence source choose the Ensembl species and build corresponding to your input track, here hg19.



ChIP-Seq - Identify composite modules on peaks (TRANSFAC(R))

Input Yes track: ...d ENCODE-UCSC hg19, chr 1, L<>490-510

Sequence source: Ensembl 72.37 Human (hg19)

Species

Input No track

Minimal number of pairs

Maximal number of pairs

Ensembl 72.37 Human (hg19)
Ensembl 65.37 Human (hg19)
Ensembl 52.36n Human (hg18)
Ensembl 72.38 Mouse (mm10)
Ensembl 72.38 Mouse (mm9)
Ensembl 71.5 Rat (rn5)
Ensembl 65.34 Rat (rn4)

Step 4. Input the No track from the tree area.

Tip for creating a No track specific for your track under study:

We recommend to apply the method *Create random track* as described in Section 16.2.6. This method helps to create a track with the same number of fragments and similar distribution of the fragment's length. Important especially for the purposes of this workflow: The method *Create random track* ensures that the fragments in the No track are not overlapping with the fragments in the Yes track.

Here, a random track comprising 300 fragments was created and used as the No track.

Step 5. After input of the Yes and No sets, verify the species shown in the species field.

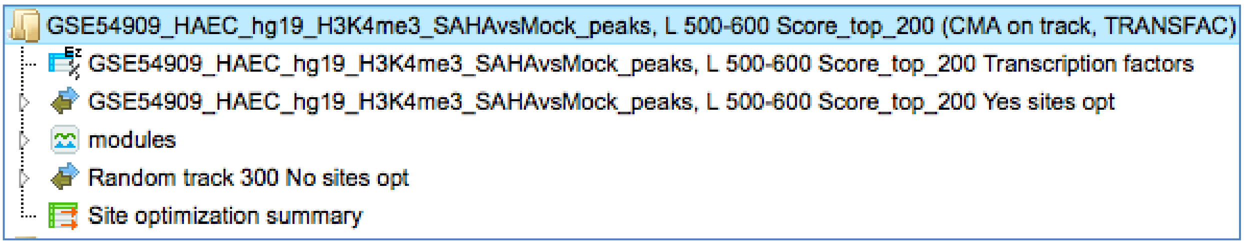
Step 6. Set up parameters for the composite module search. This workflow identifies pairs of sites. By default, the minimum and maximum numbers of pairs are given as 2 and 8. You can change these parameters according to the number of pairs you aim to identify. The number of iterations of the genetic algorithm is 300 by default, and can be adapted as required.

Step 8. Specify the result folder location and name and Press the button [Run workflow]. Wait till the workflow is completed.

> **Note**. This workflow may take time depending on the size of the Yes and No tracks and on the number of iterations. The recommended size of the input tracks is 200-400 fragments with the length of the individual fragments not exceeding 1000 bp. The maximum recommended number of iterations is 300.

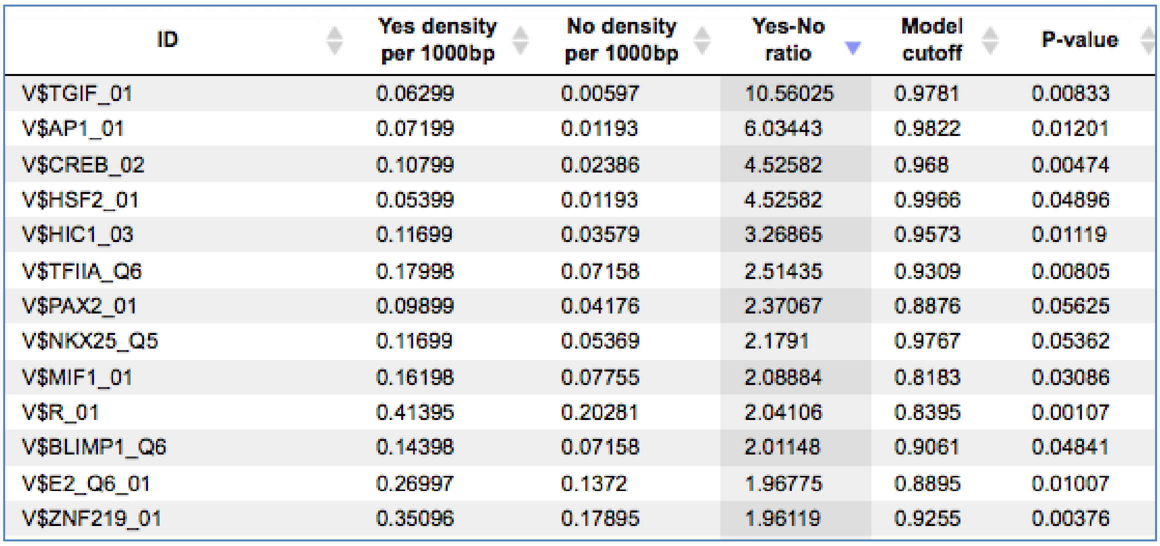
Results

The results folder contains the tables Transcription factors ( ) and Site optimization summary ( ), the tracks Yes sites opt and no sites opt ( ), and the folder modules ( ).



GSE54909_HAEC_hg19_H3K4me3_SAHAvsMock_peaks, L 500-600 Score_top_200 (CMA on track, TRANSFAC)
- GSE54909_HAEC_hg19_H3K4me3_SAHAvsMock_peaks, L 500-600 Score_top_200 Transcription factors
- GSE54909_HAEC_hg19_H3K4me3_SAHAvsMock_peaks, L 500-600 Score_top_200 Yes sites opt
- modules
- Random track 300 No sites opt
- Site optimization summary

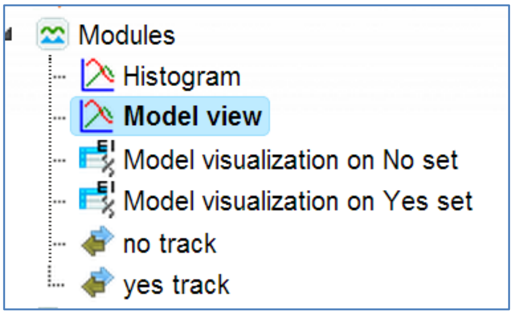
The table **Site optimization summary** ( ) contains those site models, here TRANSFAC® matrices, that are over-represented in the Yes track as compared to the No track.



ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
V$TGIF_01	0.06299	0.00597	10.56025	0.9781	0.00833
V$AP1_01	0.07199	0.01193	6.03443	0.9822	0.01201
V$CREB_02	0.10799	0.02386	4.52582	0.968	0.00474
V$HSF2_01	0.05399	0.01193	4.52582	0.9966	0.04896
V$HIC1_03	0.11699	0.03579	3.26865	0.9573	0.01119
V$TFIIA_Q6	0.17998	0.07158	2.51435	0.9309	0.00805
V$PAX2_01	0.09899	0.04176	2.37067	0.8876	0.05625
V$NKX25_Q5	0.11699	0.05369	2.1791	0.9767	0.05362
V$MIF1_01	0.16198	0.07755	2.08884	0.8183	0.03086
V$R_01	0.41395	0.20281	2.04106	0.8395	0.00107
V$BLIMP1_Q6	0.14398	0.07158	2.01148	0.9061	0.04841
V$E2_Q6_01	0.26997	0.1372	1.96775	0.8895	0.01007
V$ZNF219_01	0.35096	0.17895	1.96119	0.9255	0.00376

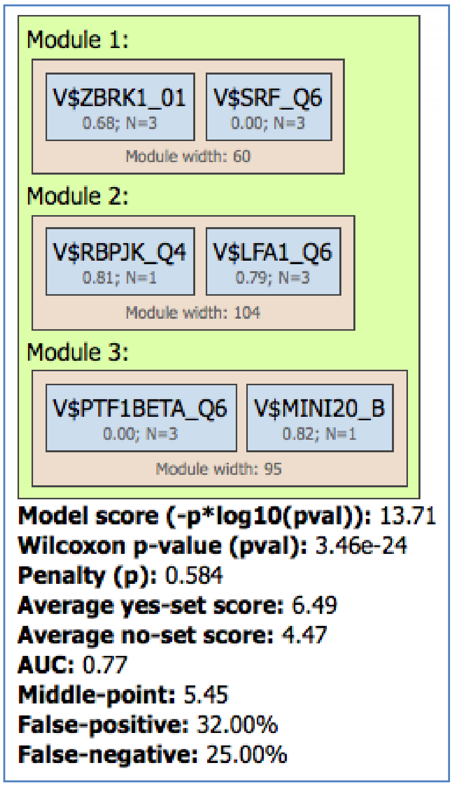
Each row of the table represents the result for one PWM from the input profile. Only those PWMs with Yes-No ratio >1 are included in the output. For details of the output columns please refer to Section 20.1.2.

The **Modules** folder ( ) is a result of the analysis *Construct composite modules*. It contains two tables, two tracks, one histogram, and one model view as shown below:



- Modules
  - Histogram
  - Model view
  - Model visualization on No set
  - Model visualization on Yes set
  - no track
  - yes track

The Model View is a graphical summary for the hierarchically organized composite elements generated as a result of the CMA analysis. As mentioned above, this workflow is designed to identify pairs of sites, and we asked to identify 2 to 8 pairs. The composite module found here contains three pairs, and we can see by exactly which site models (matrices) these pairs are formed as well as the statistical parameters of the overall model.



Module 1:

V$ZBRK1_01 0.68; N=3 | V$SRF_Q6 0.00; N=3

Module width: 60

Module 2:

V$RBPJK_Q4 0.81; N=1 | V$LFA1_Q6 0.79; N=3

Module width: 104

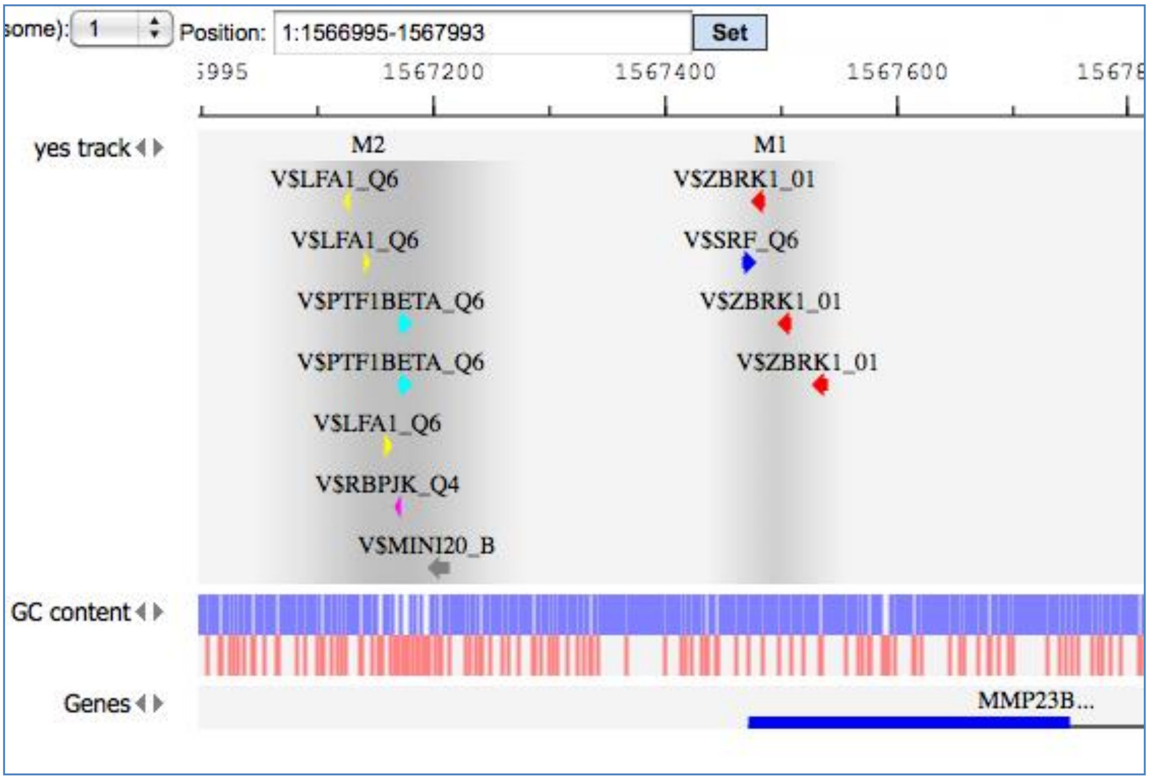
Module 3:

V$PTF1BETA_Q6 0.00; N=3 | V$MINI20_B 0.82; N=1

Module width: 95

Model score (-p*log10(pval)): 13.71
Wilcoxon p-value (pval): 3.46e-24
Penalty (p): 0.584
Average yes-set score: 6.49
Average no-set score: 4.47
AUC: 0.77
Middle-point: 5.45
False-positive: 32.00%
False-negative: 25.00%

Each of the tracks, **yes track** and **no track**, can be directly opened in the genome browser by double-clicking. Visualization of the composite module within one fragment of the input *Yes track*, is shown below.

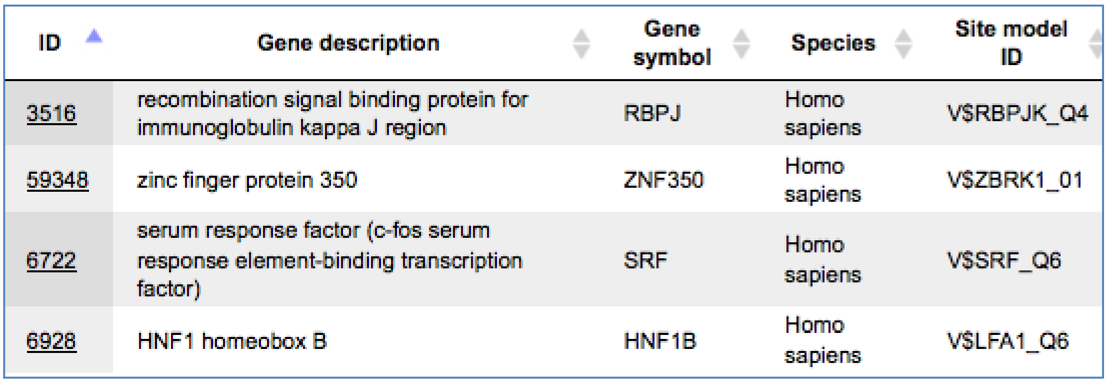


some): 1 Position: 1:1566995-1567993 Set

Upon simultaneous visualization of this track together with default tracks provided by the platform, you can see that this composite module is actually located in the promoter region of MMP23B gene.

For more details on the individual output tables and tracks as well as for **visualization** of the identified composite modules in the genome browser please refer to Section 20.1.5.

The output table **Transcription factors** is a list of transcription factors linked to the site models in the composite module identified by the workflow. For each transcription factor, the Ensembl gene ID is provided, as well as a gene description, the HGNC gene symbol, species, and site model (TRANSFAC® PWM name).



ID	Gene description	Gene symbol	Species	Site model ID
3516	recombination signal binding protein for immunoglobulin kappa J region	RBPJ	Homo sapiens	V$RBPJK_Q4
59348	zinc finger protein 350	ZNF350	Homo sapiens	V$ZBRK1_01
6722	serum response factor (c-fos serum response element-binding transcription factor)	SRF	Homo sapiens	V$SRF_Q6
6928	HNF1 homeobox B	HNF1B	Homo sapiens	V$LFA1_Q6

Four TFs are found to be candidates to specifically bind the fragments in the input track of histone acetylation peaks in SAHA-treated human vascular endothelial cells. The binding motifs for these transcription factors are parts of the identified enriched composite modules.

The same workflow can be applied to find composite modules in the ChIP-seq peaks identified as binding profiles for particular transcription factors. The example of such application for E2F1 binding peaks is described in Section 7.2.3.

Note. This workflow is available together with a valid TRANSFAC® license.
Please, feel free to ask for details (info@genexplain.com).

6.1.4. Search for discriminative sites with TRANSFAC® (MEALR)

The tool MEALR finds combinations of TFBS matrices that discriminate between two sets of sequences (denoted as *Yes* and *No* sets). The *Yes* set may consist of genomic regions identified in a ChIP-seq experiment. *No* sequences are often other non-coding genomic regions not overlapping with the peaks.

MEALR differs from other tools in the following points.

No cutoff or threshold is used on matrix scores to determine potential binding sites. Instead, MEALR calculates threshold-free sequence scores.

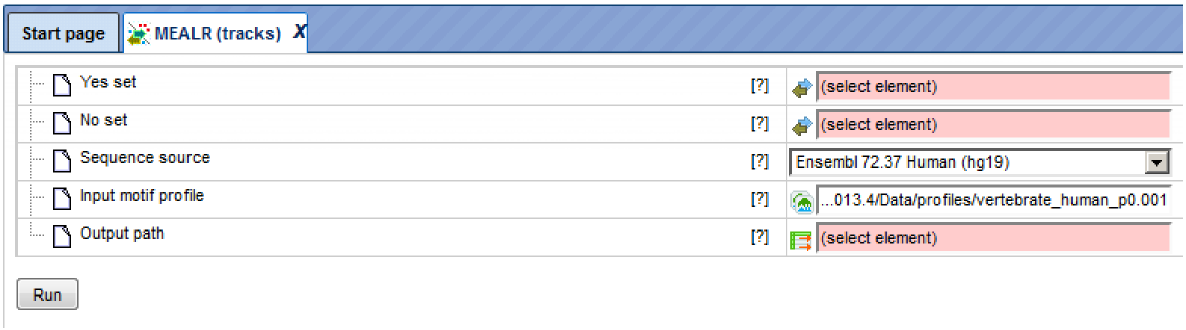
MEALR builds a discriminative model for classification which is well-established and widely applied in statistical analysis called Sparse Logistic Regression. The model consists of a linear model that estimates the probability that a sequence belongs to the Yes set based on its binding site features.

The sparseness constraint enables MEALR to select a subset of matrices relevant for classification of Yes and No sequences from a possibly large matrix library. Therefore MEALR's output differs from other tools by presenting a focused set of matrices.

While other site enrichment tools provided in the platform evaluate enrichment separately for each matrix, the model used in MEALR assesses the importance of matrices for discrimination in combination with other matrices of the library. Therefore, MEALR suggests (linear) combinations of transcription factor motifs.

MEALR calculates the score x of the i^th sequence according to the k^th matrix as $x_{ik} = log(\frac{1}{L_i}\sum_{w_i} exp\,(S_w))$, where S_w is the log-odds score of the w^th window of matrix length. Each sequence is therefore associated with a vector of scores, one from each matrix, and a class (Yes, No).

Let us present an example analysis for a ChIP-seq data set consisting of 500 peak regions and 1000 sequences randomly sampled from regulatory regions across the human genome. The figure below depicts the input mask of the analysis tool.



Yes set [?] (select element)
No set [?] (select element)
Sequence source [?] Ensembl 72.37 Human (hg19)
Input motif profile [?] ...013.4/Data/profiles/vertebrate_human_p0.001
Output path [?] (select element)

Yes set: This is the set of sequence intervals that you want to analyze, for example these can be ChIP-seq peak regions.

No set: This is the set of background intervals (control set).

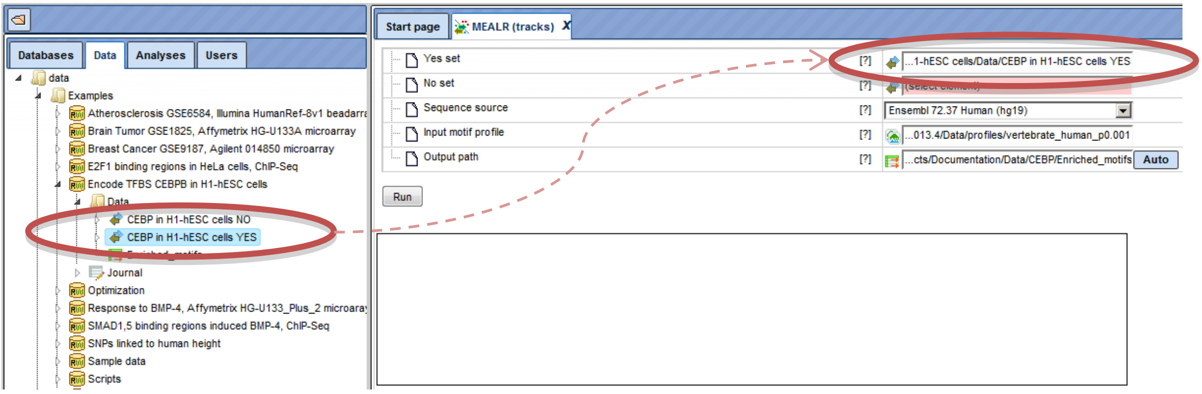
Sequence source: Both Yes and No track need to refer to a common source, such as a genome, as specified by this parameter. Note that you can apply a custom source, e.g. a specifically uploaded genome. Clicking on the "Custom" option will open a new field to choose the custom sequence source.

Input motif profile: The profile lists the PWMs (motifs) that are used to assign scores to Yes and No sequences. By default, this field is set to the profile last applied in your workspace. Note that cutoffs in the profile are ignored, because MEALR calculates whole sequence scores.

Output path: In this field you select a path in the workspace to store the output table.

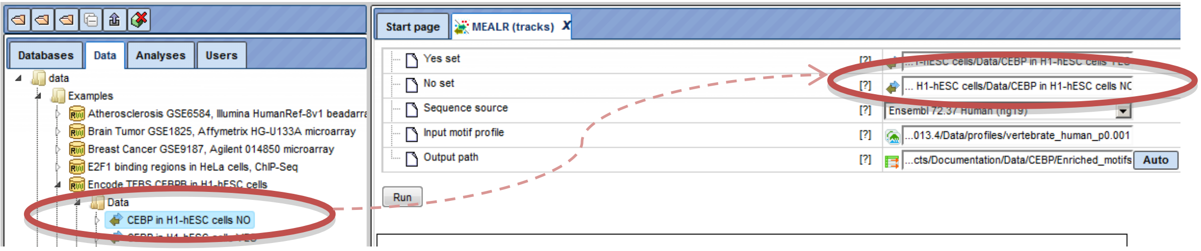
The steps of an analysis can be described as follows:

Step 1. Input Yes set from the tree. As usual, you can drag-and-drop. Here, the set of YES intervals from the Example folder is used as input, highlighted blue on the screenshot below:


Databases
Data
Analyses
Users
data
Examples
Atherosclerosis GSE6584, Illumina HumanRef-8v1 beadarra
Brain Tumor GSE1825, Affymetrix HG-U133A microarray
Breast Cancer GSE9187, Agilent 014850 microarray
E2F1 binding regions in HeLa cells, ChIP-Seq
Encode TFBS CEBPB in H1-hESC cells
Data
CEBP in H1-hESC cells NO
CEBP in H1-hESC cells YES
Journal
Optimization
Response to BMP-4, Affymetrix HG-U133_Plus_2 microarray
SMAD1,5 binding regions induced BMP-4, ChIP-Seq
SNPs linked to human height
Sample data
Scripts
Start page
MEALR (tracks)
Run

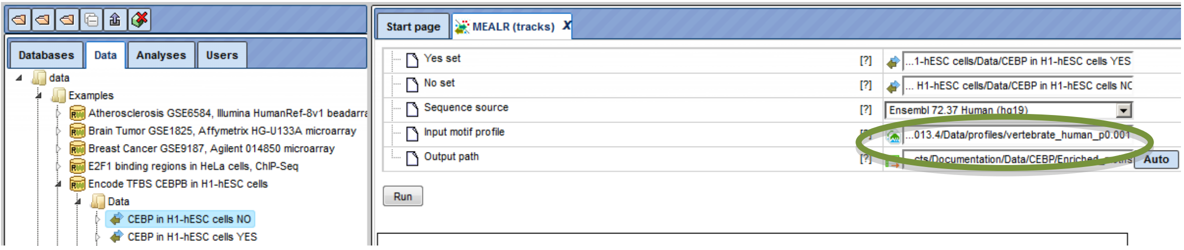
Yes set [?] ...1-hESC cells/Data/CEBP in H1-hESC cells YES
No set [?] (select element)
Sequence source [?] Ensembl 72.37 Human (hg19)
Input motif profile [?] ...013.4/Data/profiles/vertebrate_human_p0.001
Output path [?] ...cts/Documentation/Data/CEBP/Enriched_motifs Auto

Step 2. Input No set (drag-and-drop). Our example uses the set of NO intervals:


Databases
Data
Analyses
Users
data
Examples
Atherosclerosis GSE6584, Illumina HumanRef-8v1 beadarra
Brain Tumor GSE1825, Affymetrix HG-U133A microarray
Breast Cancer GSE9187, Agilent 014850 microarray
E2F1 binding regions in HeLa cells, ChIP-Seq
Encode TFBS CEBPB in H1-hESC cells
Data
CEBP in H1-hESC cells NO
Start page
MEALR (tracks)
Run

Yes set [?] ...1-hESC cells/Data/CEBP in H1-hESC cells YES
No set [?] ... H1-hESC cells/Data/CEBP in H1-hESC cells NC
Sequence source [?] Ensembl 72.37 Human (hg19)
Input motif profile [?] ...013.4/Data/profiles/vertebrate_human_p0.001
Output path [?] ...cts/Documentation/Data/CEBP/Enriched_motifs Auto

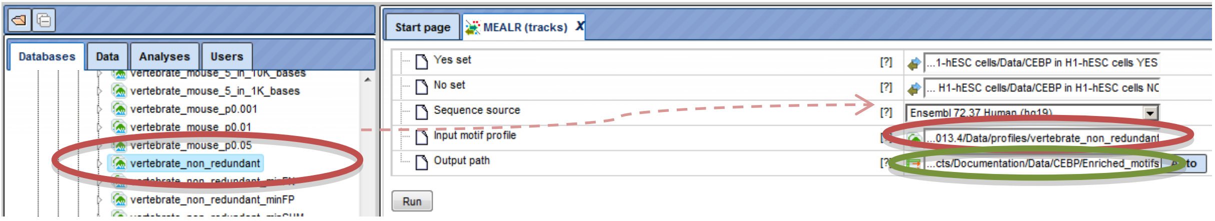
Step 3. The sequence source should be set automatically upon specifying the interval sets. If not select the corresponding sequence source from the pull-down list:



- data
  - Examples
    - Atherosclerosis GSE6584, Illumina HumanRef-8v1 beadarra
    - Brain Tumor GSE1825, Affymetrix HG-U133A microarray
    - Breast Cancer GSE9187, Agilent 014850 microarray
    - E2F1 binding regions in HeLa cells, ChIP-Seq
    - Encode TFBS CEBPB in H1-hESC cells
      - Data
        - CEBP in H1-hESC cells NO
        - CEBP in H1-hESC cells YES

Yes set [?] ...1-hESC cells/Data/CEBP in H1-hESC cells YES
No set [?] ... H1-hESC cells/Data/CEBP in H1-hESC cells NC
Sequence source [?] Ensembl 72.37 Human (hg19)
Input motif profile [?] ...013.4/Data/profiles/vertebrate_human_p0.001
Output path [?] ...cts/Documentation/Data/CEBP/Enriched_... Auto

Step 4. Select the TRANSFAC® or GTRD profile from the available profiles. In this example, we select the TRANSFAC® 2013.1 profile named “vertebrate_non_redundant”:

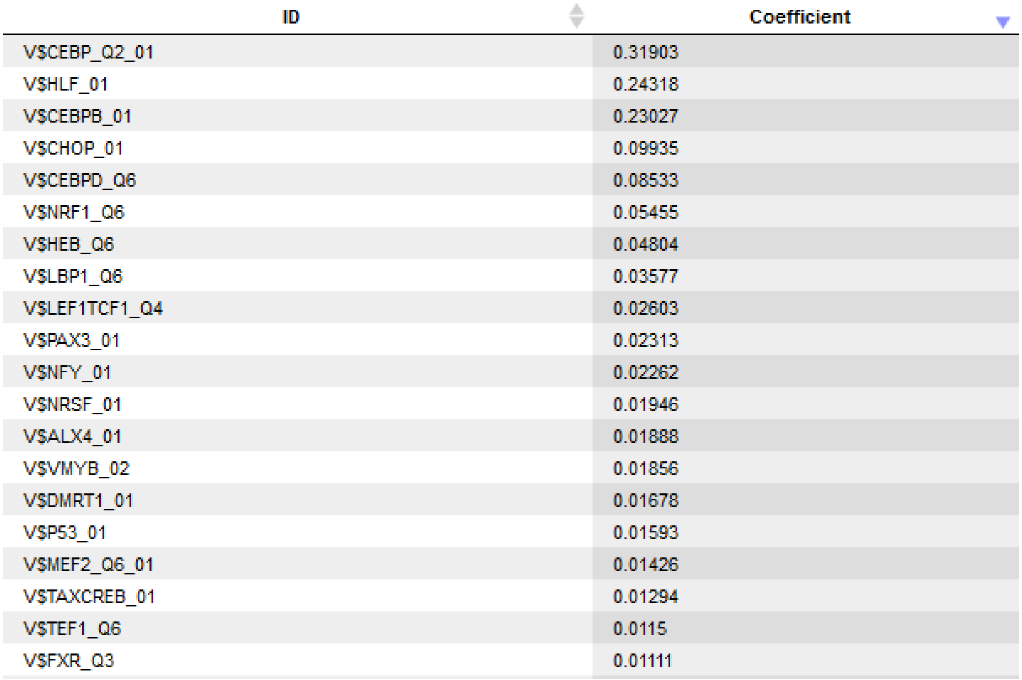


- vertebrate_mouse_5_in_10K_bases
- vertebrate_mouse_5_in_1K_bases
- vertebrate_mouse_p0.001
- vertebrate_mouse_p0.01
- vertebrate_mouse_p0.05
- vertebrate_non_redundant
- vertebrate_non_redundant_minFN
- vertebrate_non_redundant_minFP
- vertebrate_non_redundant_minSUM

Yes set [?] ...1-hESC cells/Data/CEBP in H1-hESC cells YES
No set [?] ... H1-hESC cells/Data/CEBP in H1-hESC cells NC
Sequence source [?] Ensembl 72.37 Human (hg19)
Input motif profile [?] ...013.4/Data/profiles/vertebrate_non_redundant
Output path [?] ...cts/Documentation/Data/CEBP/Enriched_motifs Auto

Step 4. Edit the output path (highlighted green in the figure above). After setting the Yes set, a default output path is suggested. The Example folder may not be writable for your account requiring selection of an alternative such as one of your own projects. A different selection can be made easily by clicking on the field.

Clicking the [Run] button will invoke the analysis. The *summary* table is automatically opened in a new tab when the analysis is completed. Here is a part of the output for our example:



ID	Coefficient
V$CEBP_Q2_01	0.31903
V$HLF_01	0.24318
V$CEBPB_01	0.23027
V$CHOP_01	0.09935
V$CEBPD_Q6	0.08533
V$NRF1_Q6	0.05455
V$HEB_Q6	0.04804
V$LBP1_Q6	0.03577
V$LEF1TCF1_Q4	0.02603
V$PAX3_01	0.02313
V$NFY_01	0.02262
V$NRSF_01	0.01946
V$ALX4_01	0.01888
V$VMYB_02	0.01856
V$DMRT1_01	0.01678
V$P53_01	0.01593
V$MEF2_Q6_01	0.01426
V$TAXCREB_01	0.01294
V$TEF1_Q6	0.0115
V$FXR_Q3	0.01111

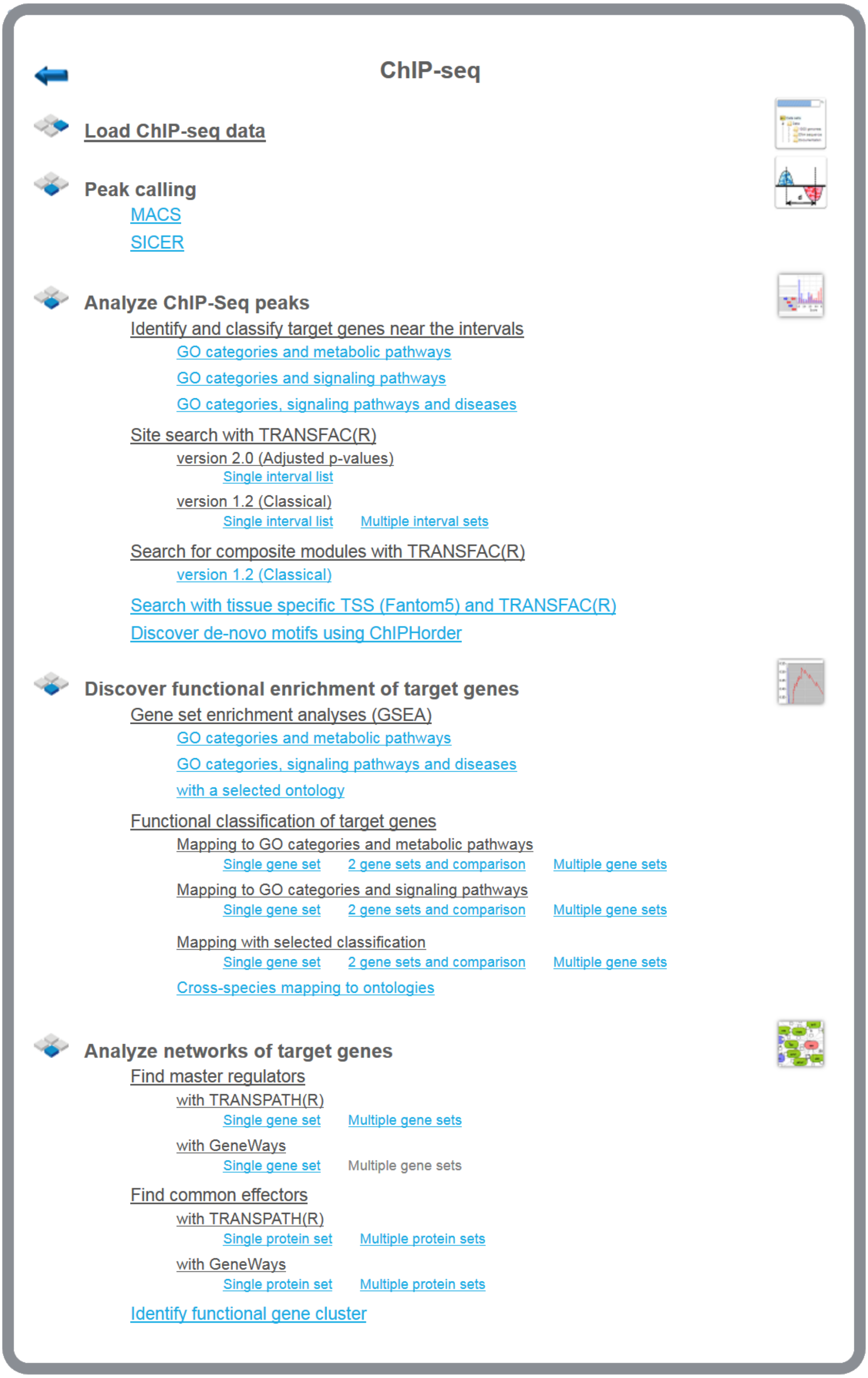
A row of the output table contains matrix identifier and its logistic regression coefficient. The larger the coefficient value, the more important the corresponding matrix was for discriminating between Yes and No sequences. In our example, three of the five top matrices represent members of the transcription factor subfamily C/EBP.

6.2. Further workflows in this area

For the other workflows that you can find in the area ***Epigenomics***, please refer to the following Sections:

Load chromatin or DNA modification genomic intervals	See Chapter 3
Discover functional enrichment of target genes	See Section 10.3
Analyze network of target genes	See Section 5.1
Find drug targets	See Chapter 11

7. ChIP-seq



ChIP-seq

Load ChIP-seq data

Peak calling
- MACS
- SICER

Analyze ChIP-Seq peaks
- Identify and classify target genes near the intervals
  - GO categories and metabolic pathways
  - GO categories and signaling pathways
  - GO categories, signaling pathways and diseases
- Site search with TRANSFAC(R)
  - version 2.0 (Adjusted p-values)
    - Single interval list
  - version 1.2 (Classical)
    - Single interval list   Multiple interval sets
- Search for composite modules with TRANSFAC(R)
  - version 1.2 (Classical)
- Search with tissue specific TSS (Fantom5) and TRANSFAC(R)
- Discover de-novo motifs using ChIPHorder

Discover functional enrichment of target genes
- Gene set enrichment analyses (GSEA)
  - GO categories and metabolic pathways
  - GO categories, signaling pathways and diseases
  - with a selected ontology
- Functional classification of target genes
  - Mapping to GO categories and metabolic pathways
    - Single gene set   2 gene sets and comparison   Multiple gene sets
  - Mapping to GO categories and signaling pathways
    - Single gene set   2 gene sets and comparison   Multiple gene sets
  - Mapping with selected classification
    - Single gene set   2 gene sets and comparison   Multiple gene sets
  - Cross-species mapping to ontologies

Analyze networks of target genes
- Find master regulators
  - with TRANSPATH(R)
    - Single gene set   Multiple gene sets
  - with GeneWays
    - Single gene set   Multiple gene sets
- Find common effectors
  - with TRANSPATH(R)
    - Single protein set   Multiple protein sets
  - with GeneWays
    - Single protein set   Multiple protein sets
- Identify functional gene cluster

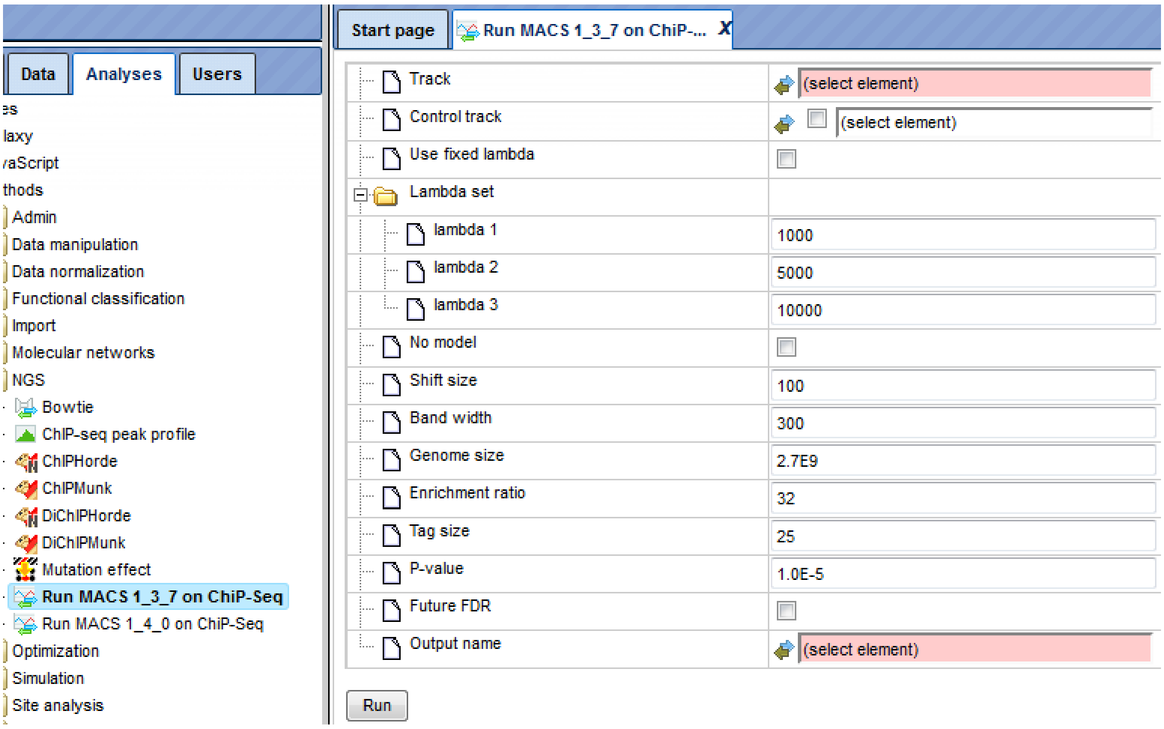
7.1. Peak calling

7.1.1. MACS

MACS is a tool to identify peaks, regions likely bound by targeted protein, in ChIP-seq data. It empirically models the length of the sequenced ChIP fragments, which tends to be shorter than sonication or library construction size estimates, and uses it to improve the spatial resolution of predicted binding sites. Details about the MACS method can also be found here.

The platform provides two versions of the MACS algorithms, 1_3_7 and 1_4_0, whose parameters are explained in the following.

MACS 1_3_7



Track	(select element)
Control track	□ (select element)
Use fixed lambda	□
Lambda set	
lambda 1	1000
lambda 2	5000
lambda 3	10000
No model	□
Shift size	100
Band width	300
Genome size	2.7E9
Enrichment ratio	32
Tag size	25
P-value	1.0E-5
Future FDR	□
Output name	(select element)

Parameters have the following meanings:

Track: Input track to search for peaks enriched with sequencing tags

Control track: A track that can be used as background (optional)

Use fixed lambda: Use a fixed local lambda for all peak regions

Lambda set: Three scopes of surrounding base pairs to calculate dynamic lambda

lambda 1: close surrounding region

lambda 2: medium width surrounding region

lambda 3: wide surrounding region

No model: Do not build the shifting model. In this mode (no model) a fixed shift size parameter is used.

Shift size: The custom shift size in bp. Used in "no model" mode.

Band width: Expected size of sonicated DNA fragments

Genome size: Effective genome size

Enrichment ratio: Cutoff for the high-confidence enrichment ratio against background

Tag size: Size of sequence tags / read length

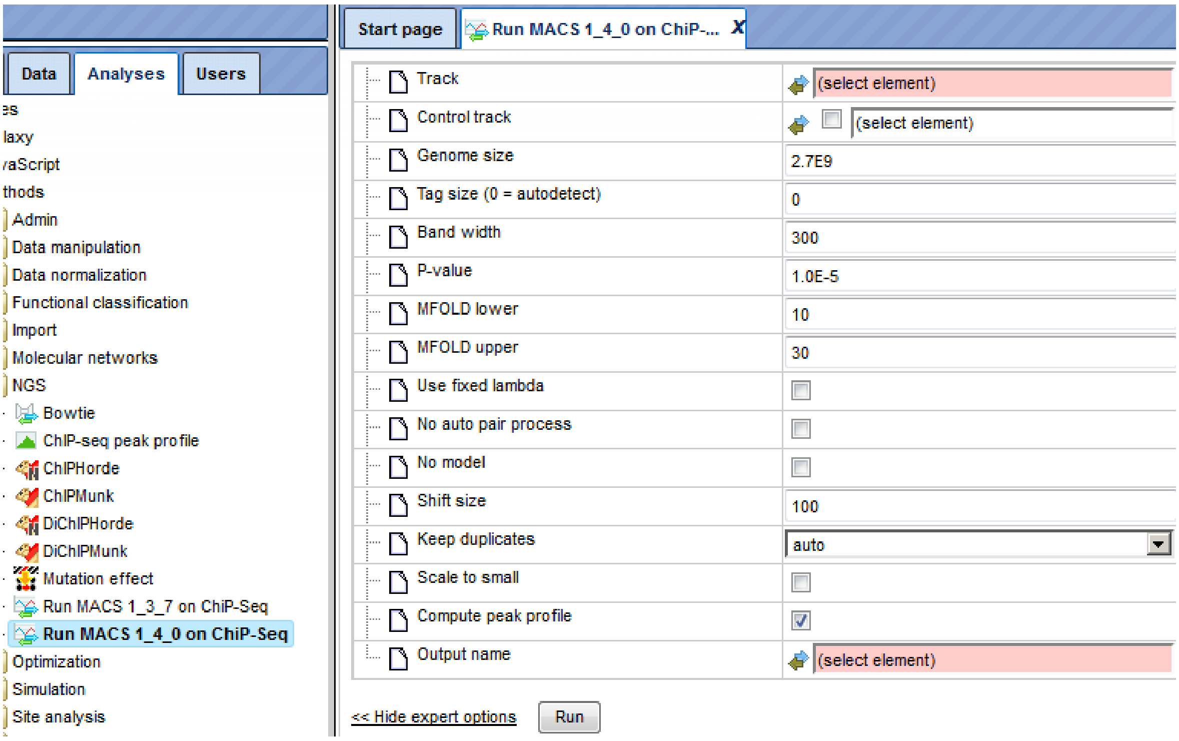
P-value: P-value cutoff for peak detection

Future FDR: Adopt the new peak detection method. The default method only considers the peak location in the 1k, 5k, or 10kb regions of the control data. In contrast, the new method also considers the 5k or 10k regions of the test data to calculate the local bias.

Output name: Name of the output track with MACS peaks.

MACS 1_4_0

This is an advanced version of MACS 1.3.7. During model building, the new algorithm selects regions within a certain range of enrichment. By default, permissible enrichment values range from 10 to 30. If MACS fails to build the model, it will use resort to "no model"-settings with a shift size=100bps, to shift and extend each tags.



Track	(select element)
Control track	☐ (select element)
Genome size	2.7E9
Tag size (0 = autodetect)	0
Band width	300
P-value	1.0E-5
MFOLD lower	10
MFOLD upper	30
Use fixed lambda	☐
No auto pair process	☐
No model	☐
Shift size	100
Keep duplicates	auto
Scale to small	☐
Compute peak profile	☑
Output name	(select element)

Track: Input track to search for peaks enriched with sequencing tags

Control track: A track that can be used as background (optional)

Genome size: Effective genome size

Tag size (0 = autodetect) (expert): Length of the tag sequences / read length in bp. If set to 0, it will be inferred from average read lengths over several first reads.

Band width: Expected size of sonicated DNA fragments

P-value: P-value cutoff for peak detection

MFOLD lower: Lower bound for high-confidence enrichment ratio used in building the paired-peak model

MFOLD upper: Upper bound for high-confidence enrichment ratio used in building the paired-peak model

Use fixed lambda (expert): Use a fixed local lambda for all peak regions

Small region for dynamic lambda: The close surrounding region (in base pairs) to calculate dynamic lambda. This is used to capture the bias near the peak summit region. Only applied with control data.

Large region for dynamic lambda: The large nearby region (in base pairs) to calculate dynamic lambda. Only applied with control data.

No auto pair process (expert): Whether to turn off the auto pair model process. If marked MACS will exit with an error message if the model building fails. If not marked, it will resourt to "no model"-settings

No model (expert): Do not build the shifting model. In this mode (no model) a fixed shift size parameter is used.

Shift size (expert): The custom shift size in bp. Used in "no model" mode.

Keep duplicates (expert): Consideration of replication tags which were assigned to the same genomic location.

Auto: value will be calculated based on binomial distribution using 1e-5 as p-value cutoff

All: Consider all tags

Number: Consider tags at most the specified number of replicated tags

Scale to small: Mark to scale larger dataset down to smaller one

Compute peak profile (expert): Mark to compute a peak profile

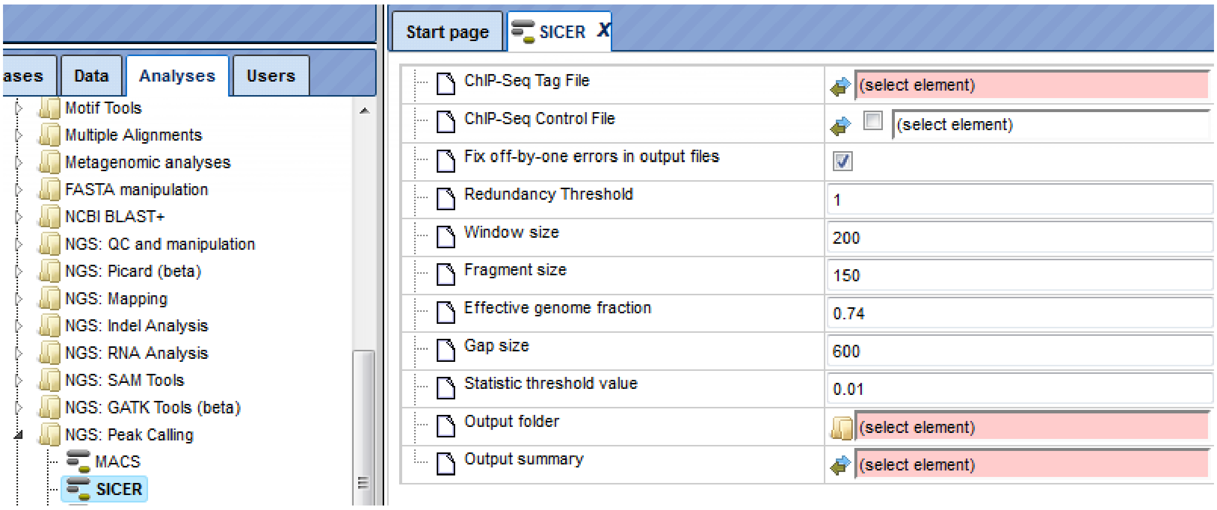
Output name: Name of the output track with MACS peaks

7.1.2. SIZER

The Galaxy tool SICER fulfills two main purposes:

1. Delineation of significantly ChIP-enriched regions, which can be used to associate with other genomic landmarks.

2. Identification of reads on the ChIP-enriched regions, which can be used for profiling and other quantitative analysis.



ChIP-Seq Tag File	(select element)
ChIP-Seq Control File	□ (select element)
Fix off-by-one errors in output files	☑
Redundancy Threshold	1
Window size	200
Fragment size	150
Effective genome fraction	0.74
Gap size	600
Statistic threshold value	0.01
Output folder	(select element)
Output summary	(select element)

Parameters for SICER should be set as described in the following.

ChIP-Seq Tag File: Input track to search for peaks

ChIP-Seq Control File: A track that can be used as background (optional)

Fix off-by-one errors in output files: SICER creates non-standard output files, this option will fix these coordinates

Redundancy Threshold: The number of copies of identical reads allowed in a library

Window size: Resolution of SICER algorithm. For histone modifications, one can use 200 bp.

Fragment size: For determination of the amount of shift from the beginning of a read to the center of the DNA fragment represented by the read. FRAGMENT_SIZE=150 means the shift is 75.

Effective genome fraction: Effective Genome as fraction of the genome size. It depends on read length.

Gap size: Needs to be multiples of window size. Namely if the window size is 200, the gap size should be 0, 200, 400, 600, ...

Statistic threshold value: FDR (with control) or E-value (without control)

Output folder: The output folder contains

- “Output summary” table with the fields ChIP_island_read_count, CONTROL_island_read_count, p_value, fold_change and FDR_threshold
- A non-redundant version of the input track
- A non-redundant version of the control track
- A summary file in bedgraph format
- A normalized output in Wig format
- A track with significant islands
- A normalized Wig file with islands
- A track with island scores

Output summary: A summary output track with islands

7.2. Analyze ChIP-seq peaks

7.2.1. Identify and classify target genes near the peaks

This group of workflows helps to identify genes located near the ChIP-seq peaks or near other genomic intervals. The input can be any track, and the output contains a table of genes overlapping with the fragments of the input track. By default, the gene bound extensions are 10,000 bp 5' relative to TSS and 10,000 bp 3' relative to the last exon.

The three workflows in this group have a very similar structure. In the first step, the input track ( ) is converted into a gene set using the *Track to gene set* analysis ( ), Section 16.2.4. The resulting Ensembl gene list is then submitted to *Functional classification* by several ontologies. In parallel, the same Ensembl gene list is subjected to *Cluster by shortest path* ( ) analysis, Section 5.1.3. The difference between the workflows within this group is in the ontologies applied for functional classification as well as in a database used to find gene/protein clusters, either TRANSPATH® or GeneWays. In the three sections below, three individual workflows are demonstrated for the same input track available in one of the pre-prepared examples present in the *Examples* folder: http://genexplain-platform.com/bioumlweb/#de=data/Examples/E2F1%20binding%20regions%20in%20HeLa%20cells%2C%20ChIP-Seq/Data/GSM558469_E2F1_hg19%20filtered%20chr%201

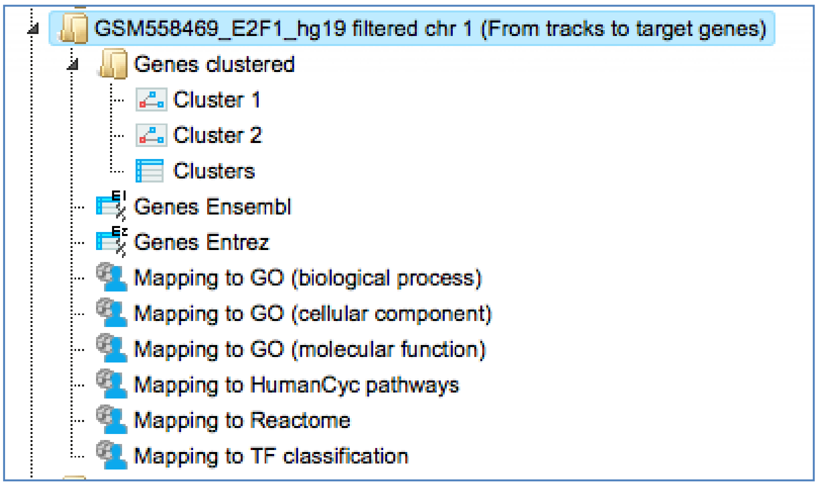
The results for each of the three workflows can be found in the folder:

data/Examples/E2F1 binding regions in HeLa cells, ChIP-Seq/Data

7.2.1.1. Classification by GO categories and metabolic pathways

In this workflow, the functional classification is done using the following ontologies: GO biological processes, GO cellular components, GO molecular function, TF classification, Reactome pathways, and HumanCyc pathways. In parallel, the same Ensembl gene list is subjected to *Cluster by shortest path* analysis. Gene/protein clusters are calculated based on the GeneWays interaction network.

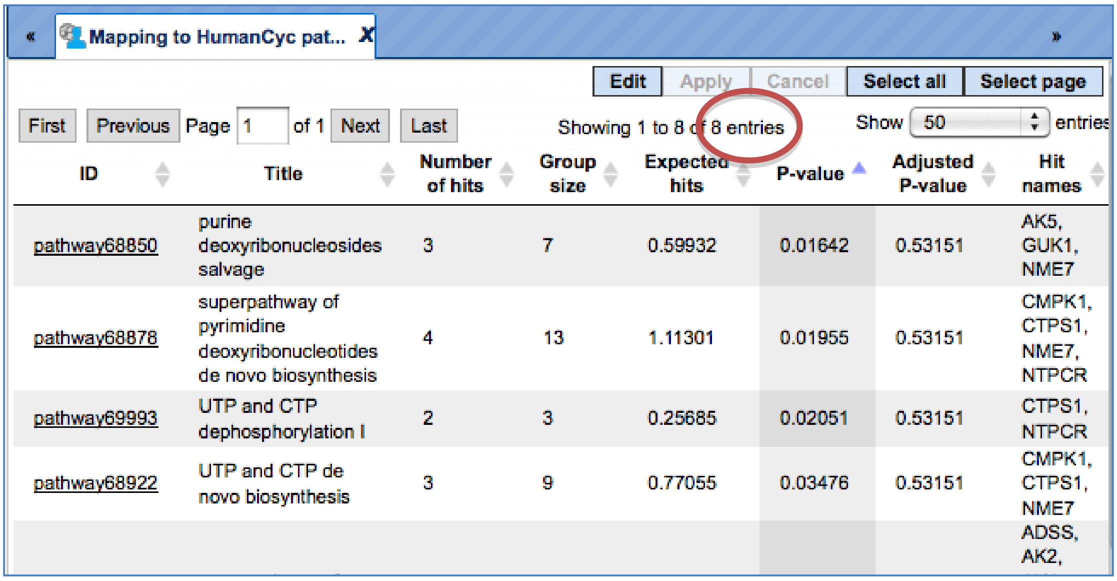
For details, how to launch this workflow, please refer to Section 7.2.1.3. The results folder looks like this:



- GSM558469_E2F1_hg19 filtered chr 1 (From tracks to target genes)
  - Genes clustered
    - Cluster 1
    - Cluster 2
    - Clusters
  - Genes Ensembl
  - Genes Entrez
  - Mapping to GO (biological process)
  - Mapping to GO (cellular component)
  - Mapping to GO (molecular function)
  - Mapping to HumanCyc pathways
  - Mapping to Reactome
  - Mapping to TF classification

The input track contains 1889 in vivo binding fragments for E2F1 transcription factor. These fragments are found to overlap with 2187 Ensembl genes that are shown in the resulting table *Genes Ensembl* ( ). This table is shown in Section 7.2.1.3.

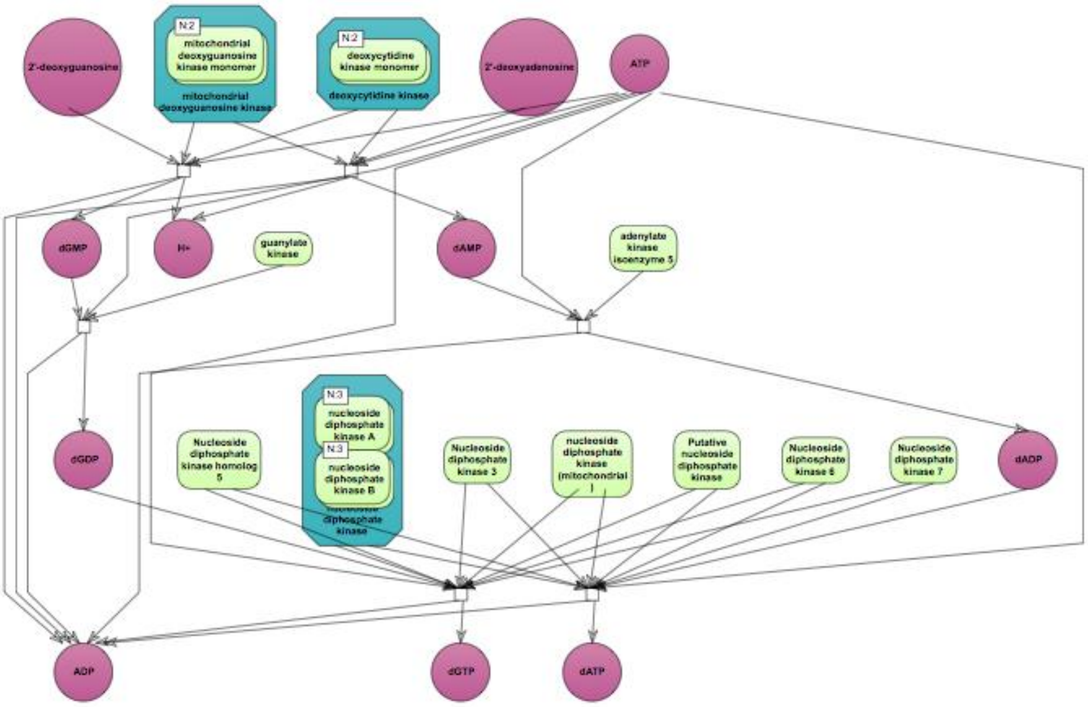
Functional classification by the HumanCyc pathways has found 8 metabolic pathways:



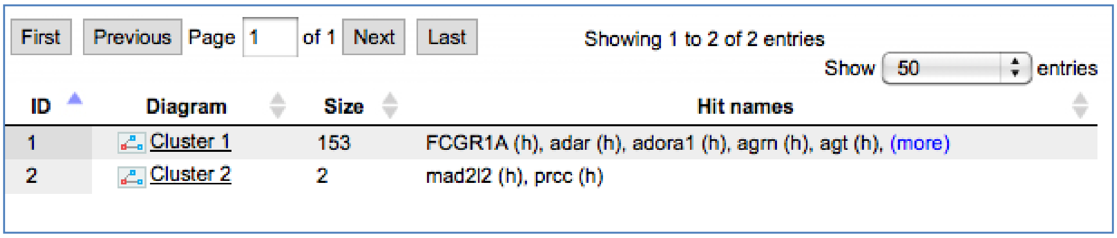
Showing 1 to 8 of 8 entries

ID	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
pathway68850	purine deoxyribonucleosides salvage	3	7	0.59932	0.01642	0.53151	AK5, GUK1, NME7
pathway68878	superpathway of pyrimidine deoxyribonucleotides de novo biosynthesis	4	13	1.11301	0.01955	0.53151	CMPK1, CTPS1, NME7, NTPCR
pathway69993	UTP and CTP dephosphorylation I	2	3	0.25685	0.02051	0.53151	CTPS1, NTPCR
pathway68922	UTP and CTP de novo biosynthesis	3	9	0.77055	0.03476	0.53151	CMPK1, CTPS1, NME7
							ADSS, AK2,

The top pathway visualization diagram can be opened in the work space upon a mouse click to the pathway ID:


2'-deoxyguanosine
N:2
mitochondrial deoxyguanosine kinase monomer
mitochondrial deoxyguanosine kinase
N:2
deoxycytidine kinase monomer
deoxycytidine kinase
2'-deoxyadenosine
ATP
dGMP
H+
guanylate kinase
dAMP
adenylate kinase isoenzyme 5
dGDP
Nucleoside diphosphate kinase homolog 5
N:3
nucleoside diphosphate kinase A
N:3
nucleoside diphosphate kinase B
nucleoside diphosphate kinase
Nucleoside diphosphate kinase 3
nucleoside diphosphate kinase (mitochondrial)
Putative nucleoside diphosphate kinase
Nucleoside diphosphate kinase 6
Nucleoside diphosphate kinase 7
dADP
ADP
dGTP
dATP

The GeneWays clusters are calculated considering upstream direction from the identified with the maximal radius of 2 steps. The following GeneWays clusters are identified:

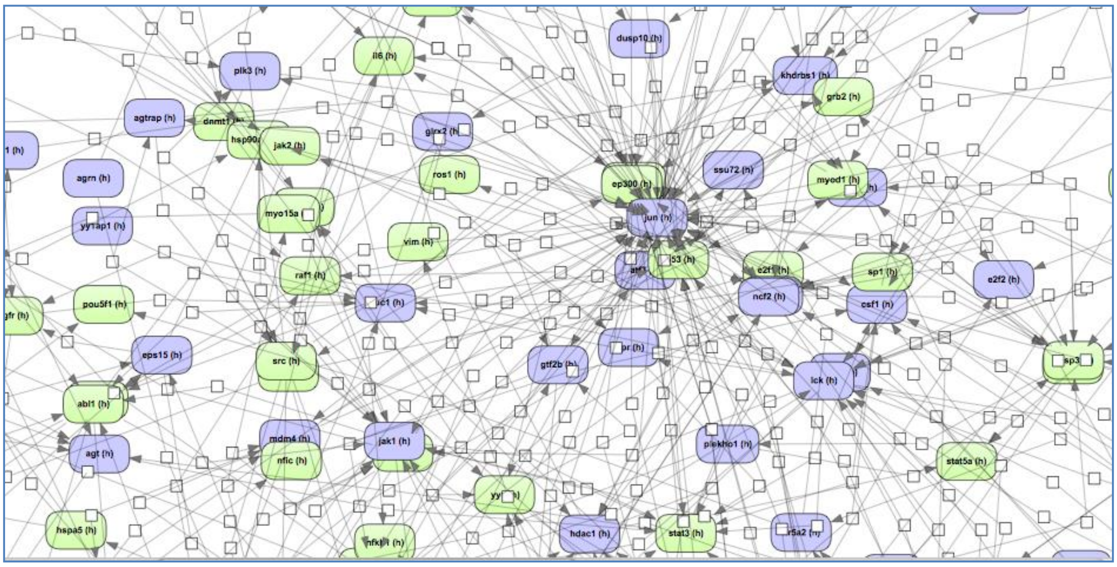


Showing 1 to 2 of 2 entries

Show 50 entries

ID	Diagram	Size	Hit names
1	Cluster 1	153	FCGR1A (h), adar (h), adora1 (h), agrn (h), agt (h), (more)
2	Cluster 2	2	mad2l2 (h), prcc (h)

The picture below presents the fragment of the cluster 1.



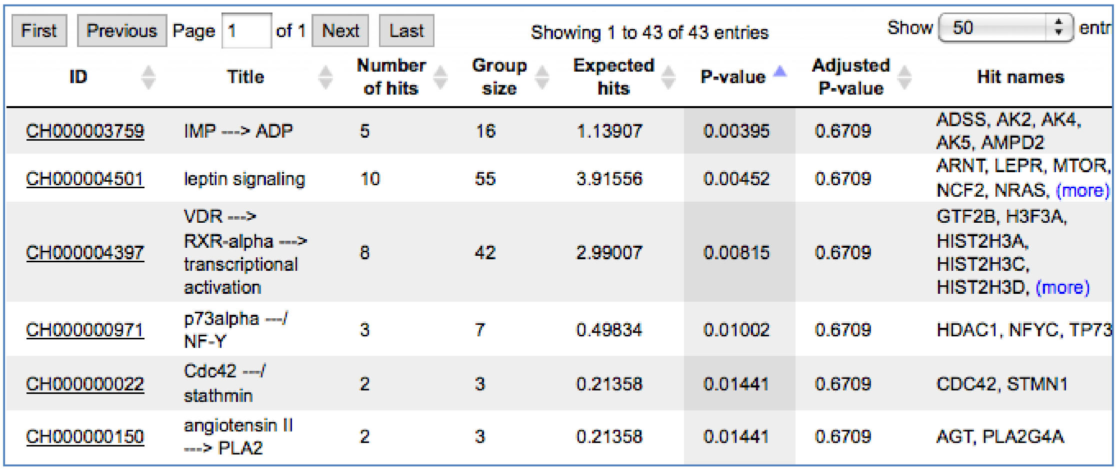
Molecules shown in blue color are coming from the input protein list, and those in green are added by the algorithm when necessary for the connectivity between the elements of the cluster.

7.2.1.2. Classification by GO categories and signaling pathways

In this workflow, the functional classification is done using the following ontologies: GO biological processes, GO cellular components, GO molecular function, TF classification, Reactome pathways, and TRANSPATH® pathways. In parallel, the same Ensembl gene list is subjected to *Cluster by shortest path* analysis. Gene/protein clusters are calculated based on the TRANSPATH® signaling network.

For details, how to launch this workflow and look into the results, please refer to Section 7.2.1.1.

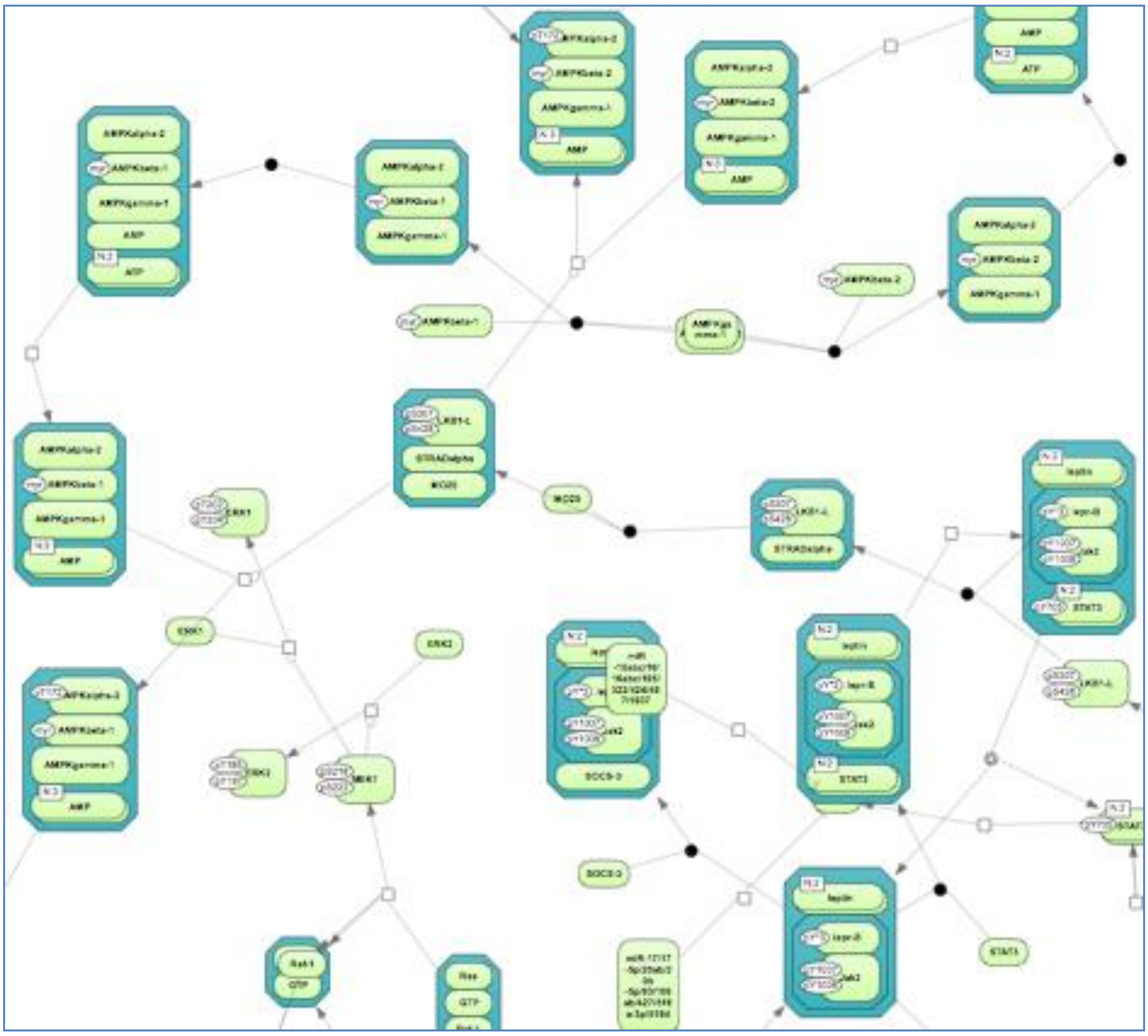
Functional classification by the TRANSPATH® pathways has found 43 signaling pathways and chains:



First Previous Page 1 of 1 Next Last Showing 1 to 43 of 43 entries Show 50 entr

ID	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
CH000003759	IMP ---> ADP	5	16	1.13907	0.00395	0.6709	ADSS, AK2, AK4, AK5, AMPD2
CH000004501	leptin signaling	10	55	3.91556	0.00452	0.6709	ARNT, LEPR, MTOR, NCF2, NRAS, (more)
CH000004397	VDR ---> RXR-alpha ---> transcriptional activation	8	42	2.99007	0.00815	0.6709	GTF2B, H3F3A, HIST2H3A, HIST2H3C, HIST2H3D, (more)
CH000000971	p73alpha ---/ NF-Y	3	7	0.49834	0.01002	0.6709	HDAC1, NFYC, TP73
CH000000022	Cdc42 ---/ stathmin	2	3	0.21358	0.01441	0.6709	CDC42, STMN1
CH000000150	angiotensin II ---> PLA2	2	3	0.21358	0.01441	0.6709	AGT, PLA2G4A

The pathway visualization diagrams can be opened in the work space upon a mouse click to the pathway ID. The fragment of the second top pathway, leptin signaling, is shown in force directed layout on the picture below. Important to mention, you can see protein complexes and modified forms on the TRANSPATH® diagrams.

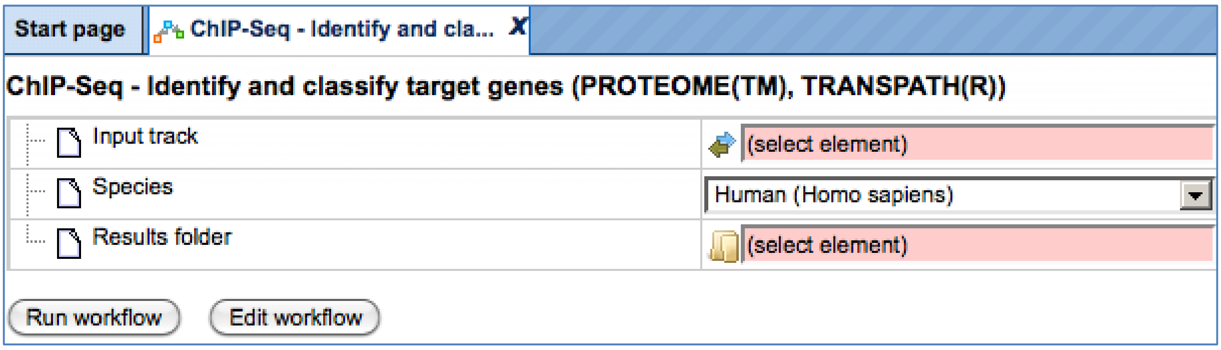


Note. This workflow is available together with a valid TRANSPATH® license. Please, feel free to ask for details (info@genexplain.com).

7.2.1.3. Classification by GO categories, signaling pathway, and diseases

In the first step of this workflow, the input track ( ) is converted into a gene set using the *Track to gene set* analysis ( ), Section 16.2.4. The resulting Ensembl gene list is then submitted to *Functional classification* using the following ontologies: PROTEOME™ GO biological processes, PROTEOME™ GO cellular components, PROTEOME™ GO molecular function, PROTEOME™ disease, TRANSPATH® pathways, TF classification, Reactome pathways, and HumanCyc pathways. In parallel, the same Ensembl gene list is subjected to *Cluster by shortest path* analysis. Gene/protein clusters are calculated based on the TRANSPATH® network.

The input form when opened in the work space is shown below:



ChIP-Seq - Identify and classify target genes (PROTEOME(TM), TRANSPATH(R))

Input track	(select element)
Species	Human (Homo sapiens)
Results folder	(select element)

Step 1. Specify input track in BED format in the field **Input track**.

You can drag & drop it from your project within the tree area. Alternatively, you may click on the pink field *select element* and a new window will be opened, where you can select the input track.

Here, further steps are demonstrated with the track available in one of the pre-prepared examples present in the Tree Area: http://genexplain-platform.com/bioumlweb/#de=data/Examples/E2F1%20binding%20regions%20in%20HeLa%20cells%2C%20ChIP-Seq/Data/GSM558469_E2F1_hg19%20filtered%20chr%201

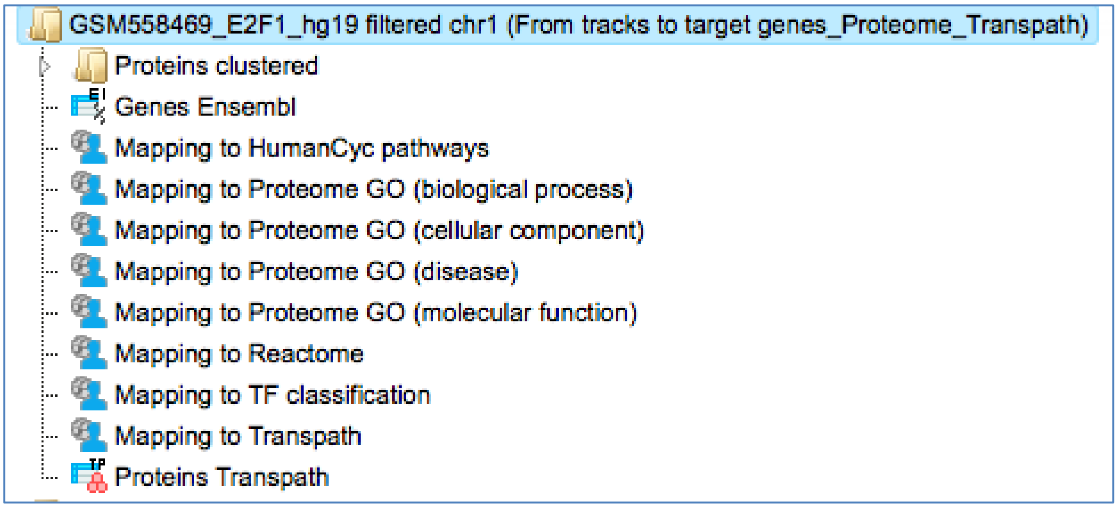
Step 2. After input of the track, the species (human, mouse or rat) is adjusted automatically. Verify the species shown in the **Species** field.

Step 3. Specify the path to store the results and the name of the output folder.

Step 4. Having filled in the input form, launch the analysis with the [Run] button. Wait till the workflow is completed.

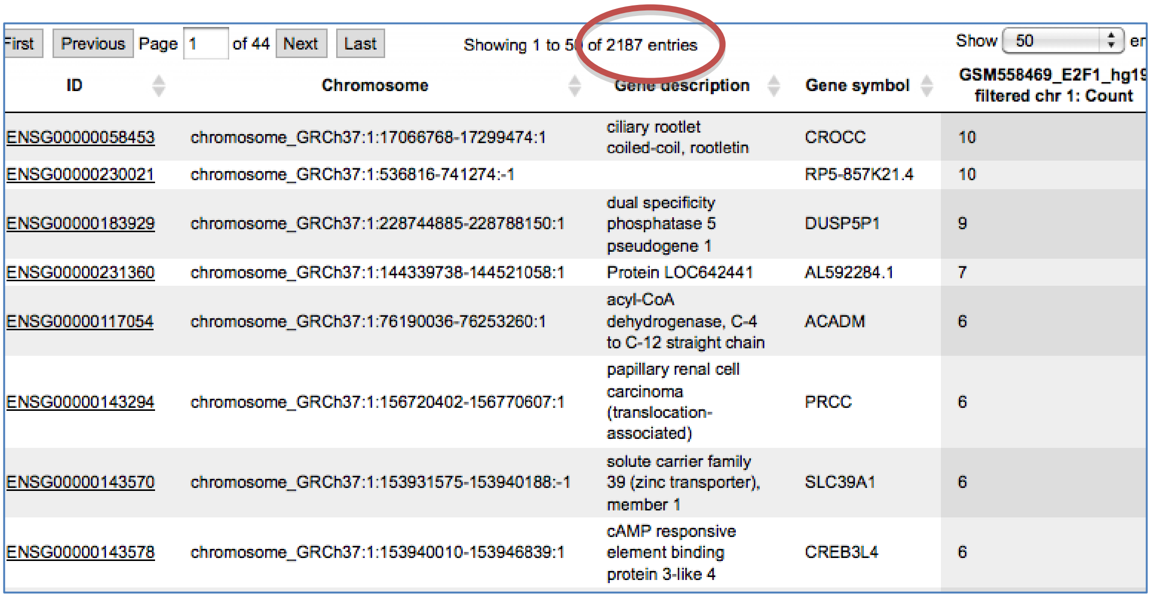
Results

The results folder contains several files as shown below. All tables with the resulting classifications as well as the table with clusters and the diagram of the largest cluster are opened by default in the work space.



- GSM558469_E2F1_hg19 filtered chr1 (From tracks to target genes_Proteome_Transpath)
  - Proteins clustered
  - Genes Ensembl
  - Mapping to HumanCyc pathways
  - Mapping to Proteome GO (biological process)
  - Mapping to Proteome GO (cellular component)
  - Mapping to Proteome GO (disease)
  - Mapping to Proteome GO (molecular function)
  - Mapping to Reactome
  - Mapping to TF classification
  - Mapping to Transpath
  - Proteins Transpath

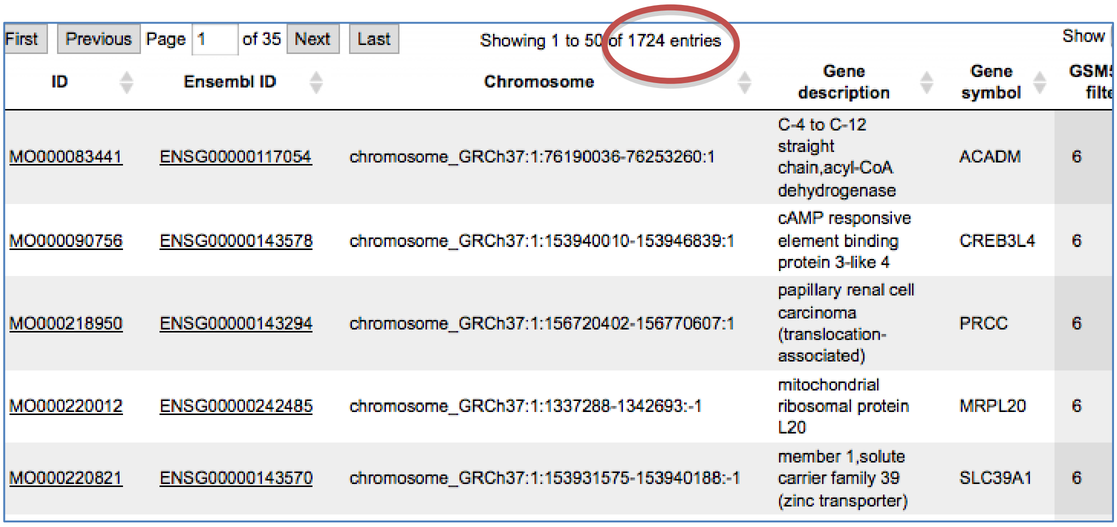
The table *Genes Ensembl* ( ) contains those genes that are identified as located in the regions around the input peaks or fragments. By default this workflow considers the following regions around Ensembl genes: 10000 bp in 5’ direction from TSS and 10000 bp in 3’ direction from the last exon. The positions of each fragment on the input track are compared with positions of the extended gene regions. Genes overlapping with at least one input fragment are considered as resulting target genes. For the input track in this example, 2187 Ensembl genes are identified. The resulting table *Genes Ensembl* is shown below, sorted by the column *Count*.



ID	Chromosome	Gene description	Gene symbol	GSM558469_E2F1_hg19 filtered chr 1: Count
ENSG00000058453	chromosome_GRCh37:1:17066768-17299474:1	ciliary rootlet coiled-coil, rootletin	CROCC	10
ENSG00000230021	chromosome_GRCh37:1:536816-741274:-1		RP5-857K21.4	10
ENSG00000183929	chromosome_GRCh37:1:228744885-228788150:1	dual specificity phosphatase 5 pseudogene 1	DUSP5P1	9
ENSG00000231360	chromosome_GRCh37:1:144339738-144521058:1	Protein LOC642441	AL592284.1	7
ENSG00000117054	chromosome_GRCh37:1:76190036-76253260:1	acyl-CoA dehydrogenase, C-4 to C-12 straight chain	ACADM	6
ENSG00000143294	chromosome_GRCh37:1:156720402-156770607:1	papillary renal cell carcinoma (translocation-associated)	PRCC	6
ENSG00000143570	chromosome_GRCh37:1:153931575-153940188:-1	solute carrier family 39 (zinc transporter), member 1	SLC39A1	6
ENSG00000143578	chromosome_GRCh37:1:153940010-153946839:1	cAMP responsive element binding protein 3-like 4	CREB3L4	6

Each row in this table contains information about one identified gene including Ensembl gene ID, chromosome, exact genomic positions and strand (1 or -1), gene symbol, and description. The column **Count** shows how many fragments on the input track are overlapping with each gene.

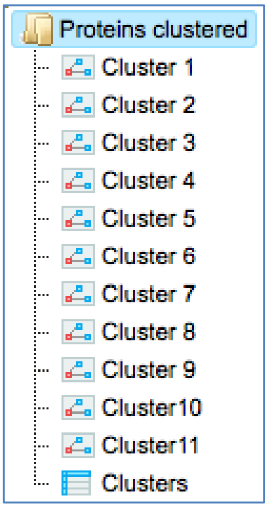
These genes are then converted into TRANSPATH® proteins, the output table Proteins Transpath ( ), shown below.



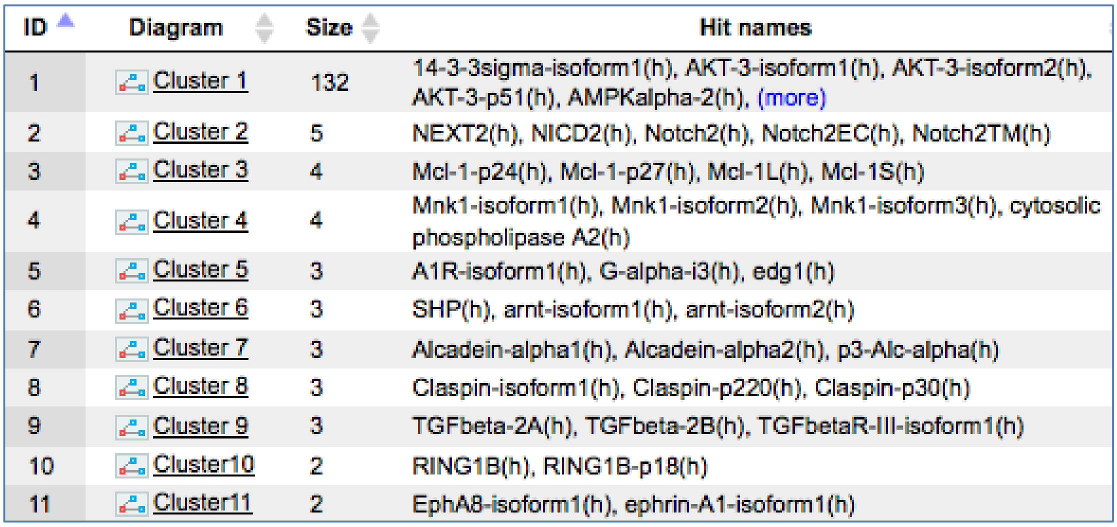
ID	Ensembl ID	Chromosome	Gene description	Gene symbol	GSM filte
MO000083441	ENSG00000117054	chromosome_GRCh37:1:76190036-76253260:1	C-4 to C-12 straight chain,acyl-CoA dehydrogenase	ACADM	6
MO000090756	ENSG00000143578	chromosome_GRCh37:1:153940010-153946839:1	cAMP responsive element binding protein 3-like 4	CREB3L4	6
MO000218950	ENSG00000143294	chromosome_GRCh37:1:156720402-156770607:1	papillary renal cell carcinoma (translocation-associated)	PRCC	6
MO000220012	ENSG00000242485	chromosome_GRCh37:1:1337288-1342693:-1	mitochondrial ribosomal protein L20	MRPL20	6
MO000220821	ENSG00000143570	chromosome_GRCh37:1:153931575-153940188:-1	member 1,solute carrier family 39 (zinc transporter)	SLC39A1	6

The structure of this table is very similar to that of Genes Ensembl, the critical difference is the column **ID**, which represents TRANSPATH® molecule IDs; Ensembl gene IDs are given in the second column. Here, 2187 Ensembl genes are converted into 1724 TRANSPATH® proteins.

Resulting TRANSPATH® proteins are clustered to get functional connections between the gene products. By default, the workflow considers upstream direction from the input TRANSPATH® proteins with a radius of 3. Resulting clusters are present in the folder *Proteins clustered*, shown below.

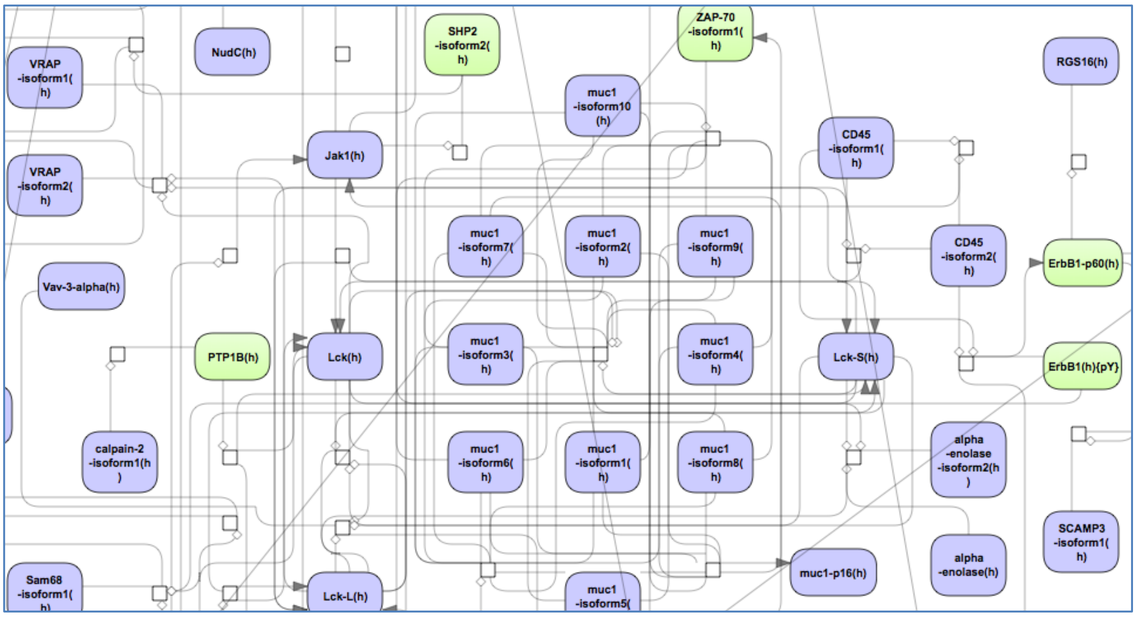


The table *Clusters* (▤) contains a list of all identified clusters, here 11. Each row shows details for one cluster. The clusters are sorted by their size with the largest cluster on top. Symbol next to each cluster name in the column **Diagram**, can be used for visualization. The column **Hit names** contains names of TRANSPATH® proteins in each cluster.



ID	Diagram	Size	Hit names
1	Cluster 1	132	14-3-3sigma-isoform1(h), AKT-3-isoform1(h), AKT-3-isoform2(h), AKT-3-p51(h), AMPKalpha-2(h), (more)
2	Cluster 2	5	NEXT2(h), NICD2(h), Notch2(h), Notch2EC(h), Notch2TM(h)
3	Cluster 3	4	Mcl-1-p24(h), Mcl-1-p27(h), Mcl-1L(h), Mcl-1S(h)
4	Cluster 4	4	Mnk1-isoform1(h), Mnk1-isoform2(h), Mnk1-isoform3(h), cytosolic phospholipase A2(h)
5	Cluster 5	3	A1R-isoform1(h), G-alpha-i3(h), edg1(h)
6	Cluster 6	3	SHP(h), arnt-isoform1(h), arnt-isoform2(h)
7	Cluster 7	3	Alcadein-alpha1(h), Alcadein-alpha2(h), p3-Alc-alpha(h)
8	Cluster 8	3	Claspin-isoform1(h), Claspin-p220(h), Claspin-p30(h)
9	Cluster 9	3	TGFbeta-2A(h), TGFbeta-2B(h), TGFbetaR-III-isoform1(h)
10	Cluster10	2	RING1B(h), RING1B-p18(h)
11	Cluster11	2	EphA8-isoform1(h), ephrin-A1-isoform1(h)

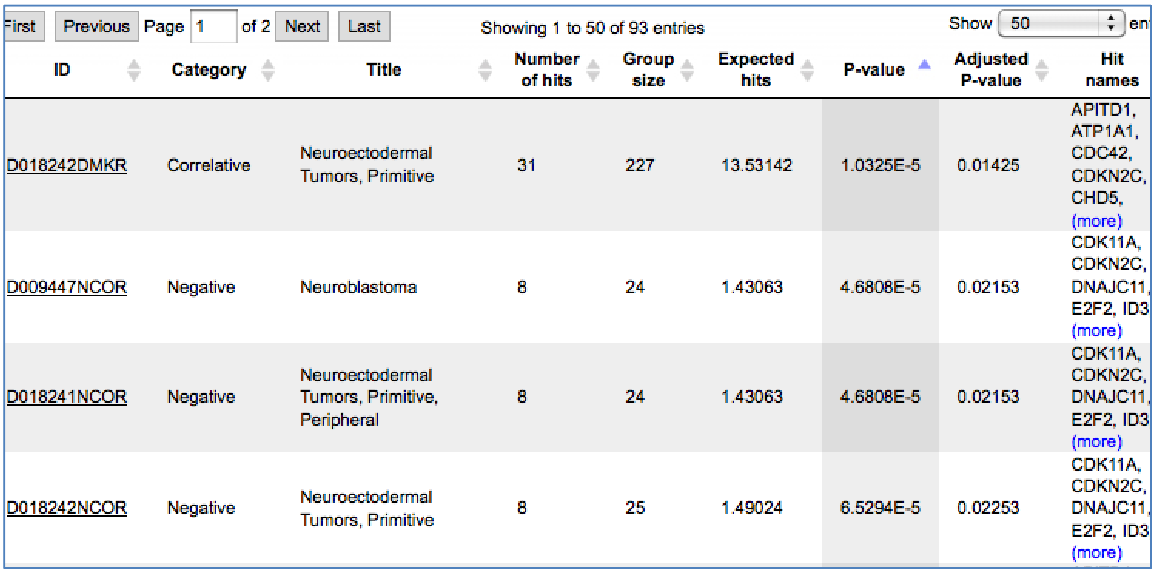
A visualization of the largest cluster, *Cluster 1*, is opened automatically when the workflow is completed. The visualization of *Cluster 1* in orthogonal layout is shown below:



Molecules shown in blue color are coming from the input protein list, and those in green are added by the algorithm when necessary for the connectivity between the elements of the cluster.

In parallel with clustering, the table Genes Ensembl undergoes to *Functional classification.* The results are shown in eight tables with the icon . Each table corresponds to a separate ontological category used for the classification in this workflow.

The resulting table, e.g. *Mapping to Proteome GO (disease)*, looks like this:



ID	Category	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
D018242DMKR	Correlative	Neuroectodermal Tumors, Primitive	31	227	13.53142	1.0325E-5	0.01425	APITD1, ATP1A1, CDC42, CDKN2C, CHD5, (more)
D009447NCOR	Negative	Neuroblastoma	8	24	1.43063	4.6808E-5	0.02153	CDK11A, CDKN2C, DNAJC11, E2F2, ID3 (more)
D018241NCOR	Negative	Neuroectodermal Tumors, Primitive, Peripheral	8	24	1.43063	4.6808E-5	0.02153	CDK11A, CDKN2C, DNAJC11, E2F2, ID3 (more)
D018242NCOR	Negative	Neuroectodermal Tumors, Primitive	8	25	1.49024	6.5294E-5	0.02253	CDK11A, CDKN2C, DNAJC11, E2F2, ID3 (more)

Each row corresponds to one ontological category, which in this case is one of the diseases as they are annotated in the PROTEOME™ database. Commonly accepted disease identifiers are shown in the **ID** column. The disease names are shown in the column **Title**. The column **Group size** represents the number of genes linked to this disease in PROTEOME™, and the column **Category** demonstrates the functional type of the link between genes and disease; it can be causal, correlative or negative. For each row several parameters are calculated, the expected number of hits (**Expected hits**), the actual number of hits (**Number of hits**), **P-value**, as well as **Hit names**. IDs are hyperlinked to an external web page of CTD, the Comparative Toxicogenomics Database. With a click on each ID, a new tab will be opened displaying additional information about the disease.

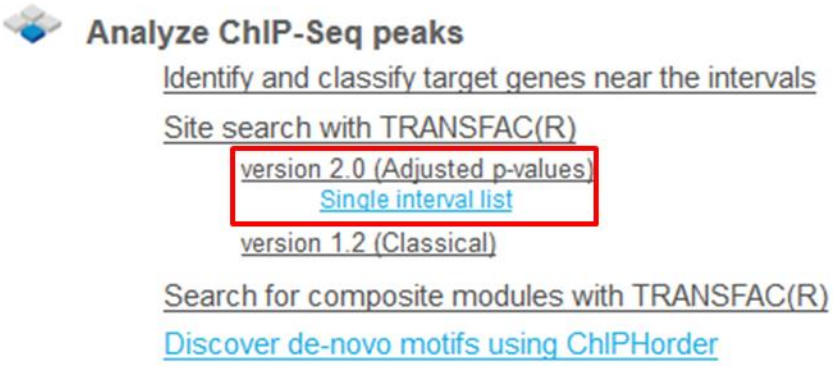
Note. This workflow is available together with a valid PROTEOME™ license. Please, feel free to ask for details (info@genexplain.com).

7.2.2. Site search with TRANSFAC®

7.2.2.1. Version 2.0 (Adjusted p-values, site search on track)

Single interval list

This workflow "Identify enriched motifs in tracks (TRANSFAC®)" is designed to map putative enriched TFBSs on peaks calculated from your ChIP-seq data (Yes set) as compared to a random background set (No set). Importantly, the No set is created automatically and contains by default 1000 intervals. In the first part of the workflow, the enriched motifs are identified by our proprietary MEALR approach (*analyses/Methods/Site analysis/MEALR (tracks)*, icon). Please refer to section 6.1.4 for details of this particular analysis method. Enriched motifs serve as a basis to construct a specific profile. At the next step this newly generated profile is run on the same list of input peaks applying the method *analyses/Methods/Site analysis/Search for enriched TFBSs (tracks)*, icon . Please refer to section 20.1.4.2 for details of this particular analysis method. The workflow can be found under the section “Analyze ChIP-seq peaks” → Site search with TRANSFAC(R) → version 2.0 (Adjusted p-values).

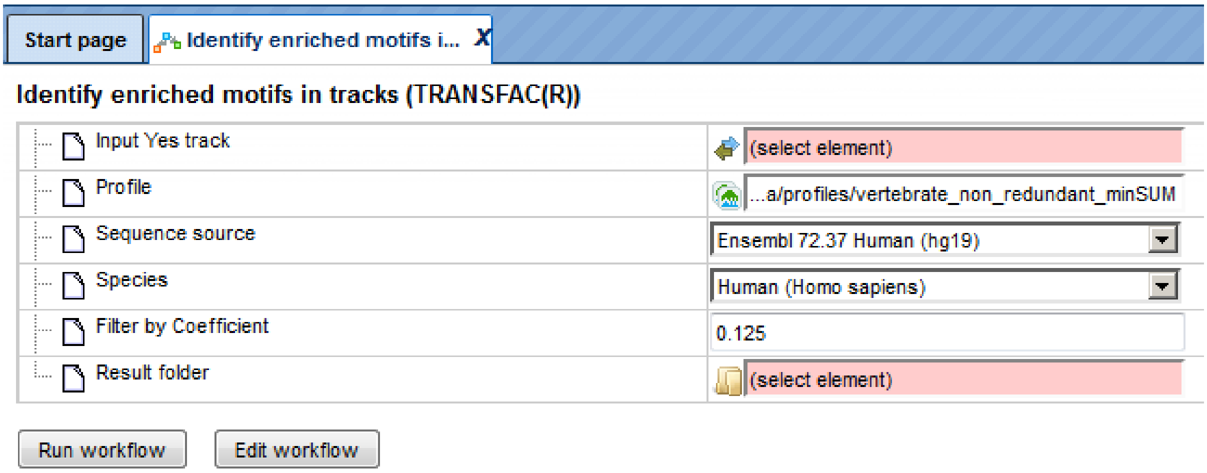


Analyze ChIP-Seq peaks

- Identify and classify target genes near the intervals
- Site search with TRANSFAC(R)
  - version 2.0 (Adjusted p-values)
    - Single interval list
  - version 1.2 (Classical)
- Search for composite modules with TRANSFAC(R)
- Discover de-novo motifs using ChIPHorder

To launch the workflow, follow these steps:

Step 1. Open the workflow input form from the Start page. It will open in the main Work Space and looks as shown below:



Identify enriched motifs in tracks (TRANSFAC(R))

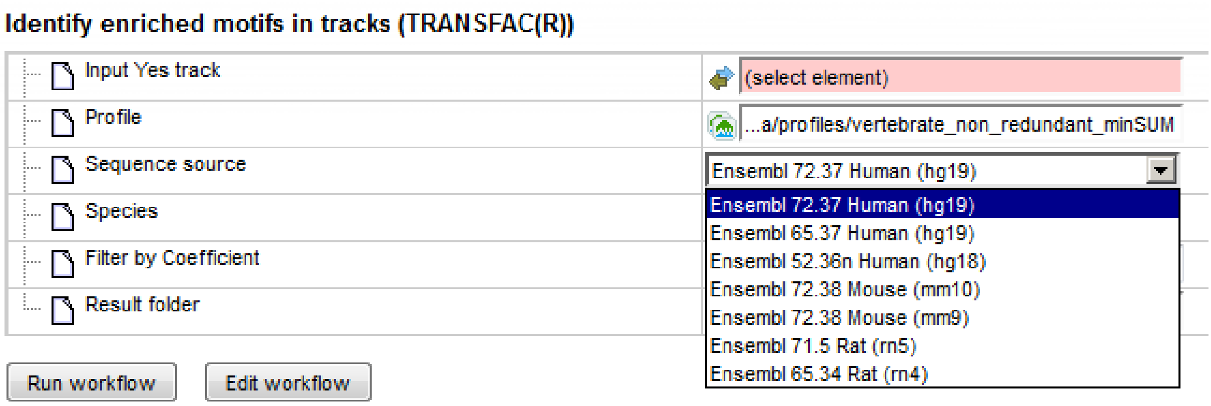
Input Yes track	(select element)
Profile	...a/profiles/vertebrate_non_redundant_minSUM
Sequence source	Ensembl 72.37 Human (hg19)
Species	Human (Homo sapiens)
Filter by Coefficient	0.125
Result folder	(select element)

Step 2. Specify the input track in BED format in the field **Input Yes track**. The input Yes track contains peaks from your ChIP-seq study. To specify the Yes track, you can drag & drop it from your project within the tree area. Alternatively, you may click on the pink field “select element” and a new window will open, where you select the input track. After having selected the track, press the [Ok] button.

Step 3. Select the profile. This profile will be applied at the first part of the workflow for identification of the enriched motifs. The default profile is *vertebrate_non_redundant_minSUM* from the most recent TRANSFAC® release available.

Any other TRANSFAC® profile or user-specific profile can be selected. With a mouse click on the field **Profile**, a pop-up window will open, where a profile can be selected.

Step 4. Specify the sequence source from the drop-down menu. Several human, mouse and rat sequence builds are available in the platform, as shown below. By default, the most recent Ensembl human genome, hg19, is specified. Make sure you selected the sequence source (species and the genome build) that corresponds to your input set, to get correct and meaningful results.



Identify enriched motifs in tracks (TRANSFAC(R))

- Input Yes track: (select element)
- Profile: ...a/profiles/vertebrate_non_redundant_minSUM
- Sequence source: Ensembl 72.37 Human (hg19)
- Species
- Filter by Coefficient
- Result folder

Ensembl 72.37 Human (hg19)
Ensembl 65.37 Human (hg19)
Ensembl 52.36n Human (hg18)
Ensembl 72.38 Mouse (mm10)
Ensembl 72.38 Mouse (mm9)
Ensembl 71.5 Rat (rn5)
Ensembl 65.34 Rat (rn4)
Run workflow
Edit workflow

Step 5. Specify the biological species of the input set in the field **Species** by selecting it from the drop-down menu.

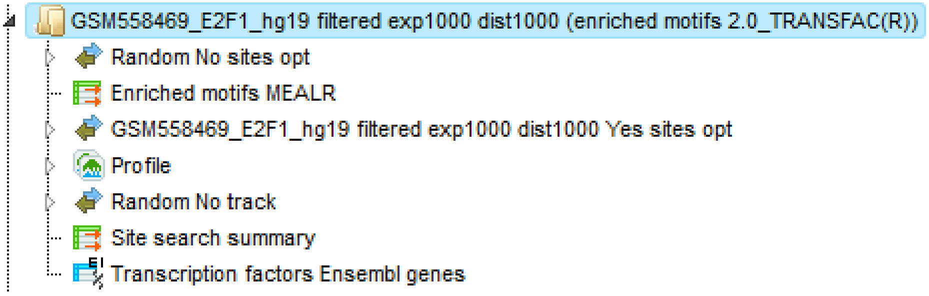
Step 6. Select a filter for the coefficient of the MEALR method. The default filter is set as >0.125 to have 75% or more of true discovery rate, TDR. For 90% TDR, you can type 0.270 in this field and for 50% TDR - 0.05593. The filtered motifs are included in the output as *enriched motifs*. At the later step, PWMs corresponding to the enriched motifs are used to make a new profile.

Step 7. Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink field "select element" in the field **Result folder**, and a new window will be opened, where you can select the location of the results folder and define its name.

Step 8. Press the [Run workflow] button.

Wait until the workflow is completed.

The **Result folder** contains several tables and three tracks; for this example, let's consider the results folder located under Examples. It is highlighted in blue in the figure below:



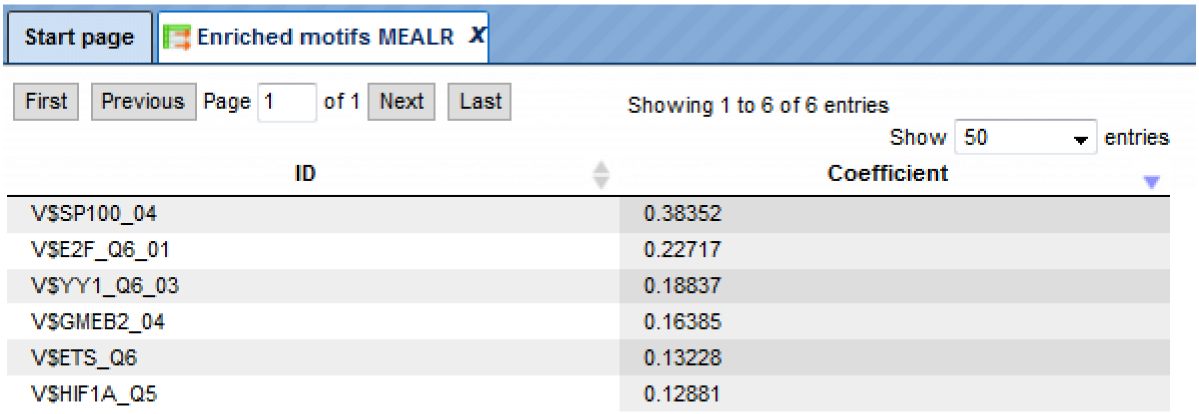
- GSM558469_E2F1_hg19 filtered exp1000 dist1000 (enriched motifs 2.0_TRANSFAC(R))
  - Random No sites opt
  - Enriched motifs MEALR
  - GSM558469_E2F1_hg19 filtered exp1000 dist1000 Yes sites opt
  - Profile
  - Random No track
  - Site search summary
  - Transcription factors Ensembl genes

The tables *Enriched motifs MEALR* (▤) and *Transcription factors* (▤) are opened automatically in the Work Space as soon as the workflow is completed.

The table **Enriched motifs MEALR** includes enriched motifs in the Yes track *versus* the No track, filtered by the coefficient as specified.

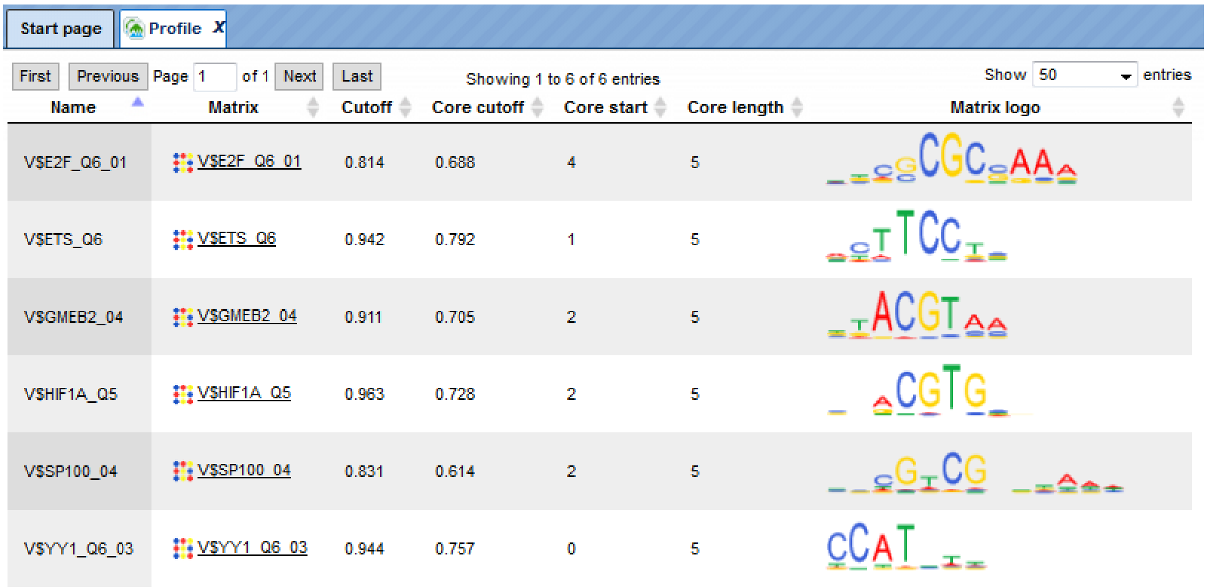
Please note that by default only the matrices with a **Coefficient** >0.125 (75% **T**rue **D**iscovery **R**ate) are included in this output table. These motifs can be interpreted as the best discriminating motifs between the Yes and NO sets.

The table **Enriched motifs MEALR** shown below has been sorted by the values in the **Coefficient** column. The larger the coefficient, the more important the corresponding motif was for discriminating between Yes and No sequences.



ID	Coefficient
V$SP100_04	0.38352
V$E2F_Q6_01	0.22717
V$YY1_Q6_03	0.18837
V$GMEB2_04	0.16385
V$ETS_Q6	0.13228
V$HIF1A_Q5	0.12881

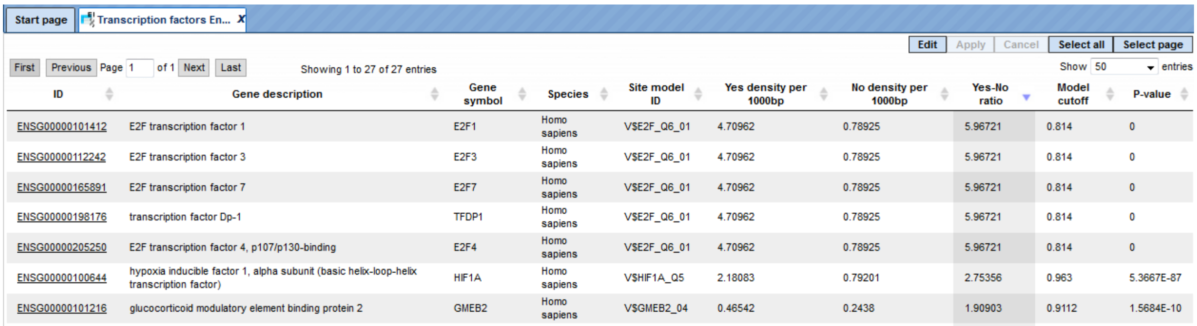
The table **Profile** is opened automatically and is an input-specific profile, based on the filtered *enriched motifs MEALR* from the first part of the workflow.



Name	Matrix	Cutoff	Core cutoff	Core start	Core length	Matrix logo
V$E2F_Q6_01	V$E2F_Q6_01	0.814	0.688	4	5	
V$ETS_Q6	V$ETS_Q6	0.942	0.792	1	5	
V$GMEB2_04	V$GMEB2_04	0.911	0.705	2	5	
V$HIF1A_Q5	V$HIF1A_Q5	0.963	0.728	2	5	
V$SP100_04	V$SP100_04	0.831	0.614	2	5	
V$YY1_Q6_03	V$YY1_Q6_03	0.944	0.757	0	5	

This profile is an intermediate result of the workflow and is used further for *Site search on gene set* analysis in the second part of the workflow.

Table Transcription factors Ensembl:



Showing 1 to 27 of 27 entries

ID	Gene description	Gene symbol	Species	Site model ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
ENSG00000101412	E2F transcription factor 1	E2F1	Homo sapiens	V$E2F_Q6_01	4.70962	0.78925	5.96721	0.814	0
ENSG00000112242	E2F transcription factor 3	E2F3	Homo sapiens	V$E2F_Q6_01	4.70962	0.78925	5.96721	0.814	0
ENSG00000165891	E2F transcription factor 7	E2F7	Homo sapiens	V$E2F_Q6_01	4.70962	0.78925	5.96721	0.814	0
ENSG00000198176	transcription factor Dp-1	TFDP1	Homo sapiens	V$E2F_Q6_01	4.70962	0.78925	5.96721	0.814	0
ENSG00000205250	E2F transcription factor 4, p107/p130-binding	E2F4	Homo sapiens	V$E2F_Q6_01	4.70962	0.78925	5.96721	0.814	0
ENSG00000100644	hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)	HIF1A	Homo sapiens	V$HIF1A_Q5	2.18083	0.79201	2.75356	0.963	5.3667E-87
ENSG00000101216	glucocorticoid modulatory element binding protein 2	GMEB2	Homo sapiens	V$GMEB2_04	0.46542	0.2438	1.90903	0.9112	1.5684E-10

This table includes transcription factors (TFs) that are associated with the PWMs listed in the table *Site search summary*. Each row shows details for one TF, including its Ensembl gene ID (column **ID**), gene symbol, gene description and biological species of the corresponding TF (columns **Gene description**, **Gene symbol**, and **Species**). The column **Site model ID** shows the identifier of the PWM associated with this TF, and several further columns repeat information that is also shown in the table *Site search summary*.

For further visualization of resulting *Yes sites opt* track please refer to section 6.1.2.1.

> **Note**. This workflow is available together with a valid TRANSFAC® license.
> Please, feel free to ask for details (info@genexplain.com).

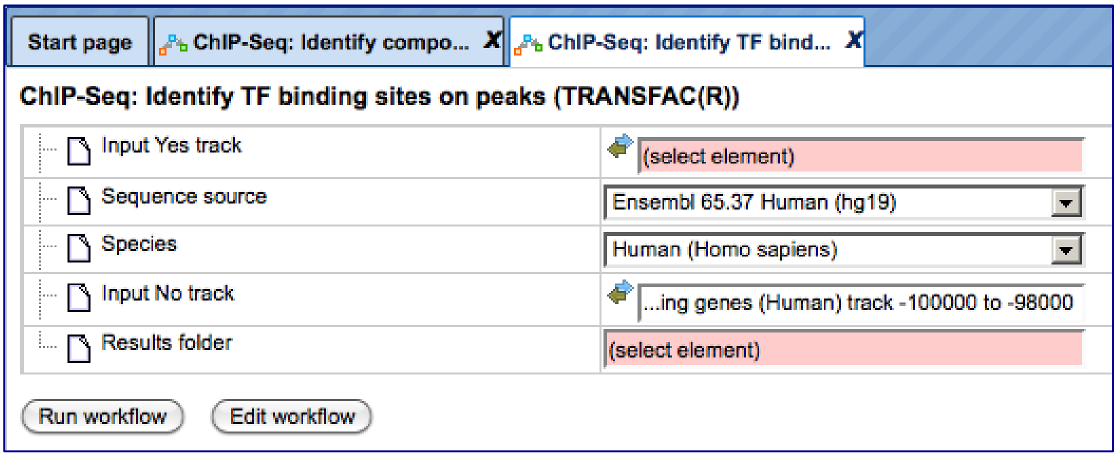
7.2.2.2. Version 1.2 (Classical)

Single interval list

This workflow helps to map putative TFBSs on peaks calculated from your ChIP-seq data. Site search is done with the help of the TRANSFAC® library of positional weight matrices, PWMs, using the pre-computed profile vertebrate_non_redundant_minSUM.

The few steps to launch the workflow are described in the following.

Step 1. Open workflow input form from the Start page, it will be opened in the main Work Space and looks as it is shown below:

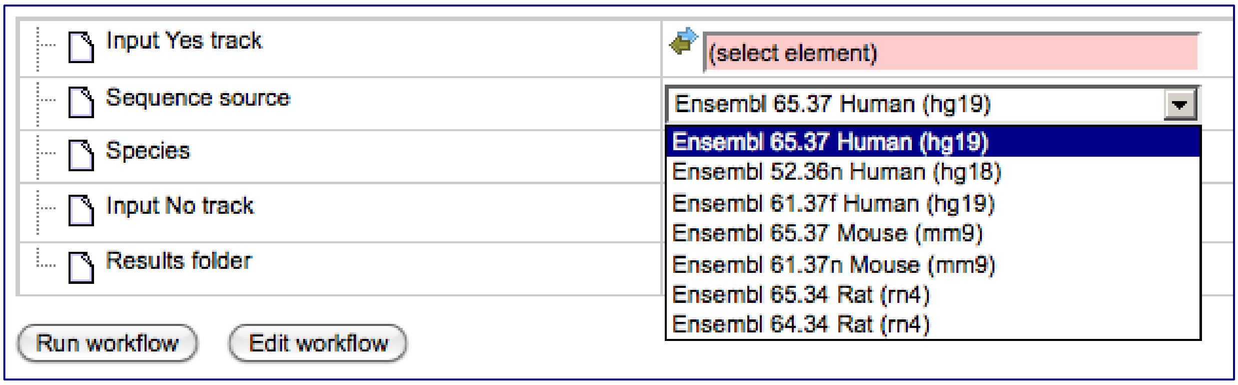


ChIP-Seq: Identify TF binding sites on peaks (TRANSFAC(R))

Input Yes track	(select element)
Sequence source	Ensembl 65.37 Human (hg19)
Species	Human (Homo sapiens)
Input No track	...ing genes (Human) track -100000 to -98000
Results folder	(select element)

Step 2. Specify the input track in BED format in the field **Input Yes track**. The input Yes track contains peaks from your ChIP-seq study. To specify the Yes track, you can drag & drop it from your project within the tree area. Alternatively, you may click on the pink field "select element" and a new window will open, where you select the input track. After having selected the track, press the [Ok] button.

Step 3. Specify the sequence source from the drop-down menu. Several human, mouse and rat sequence builds are available in the platform, as shown below. By default, the most recent Ensembl human genome, hg19, is specified. Make sure you selected the sequence source (the genome build) that corresponds to your input set, to get correct and meaningful results.



Input Yes track: (select element)

Sequence source: Ensembl 65.37 Human (hg19)

Species

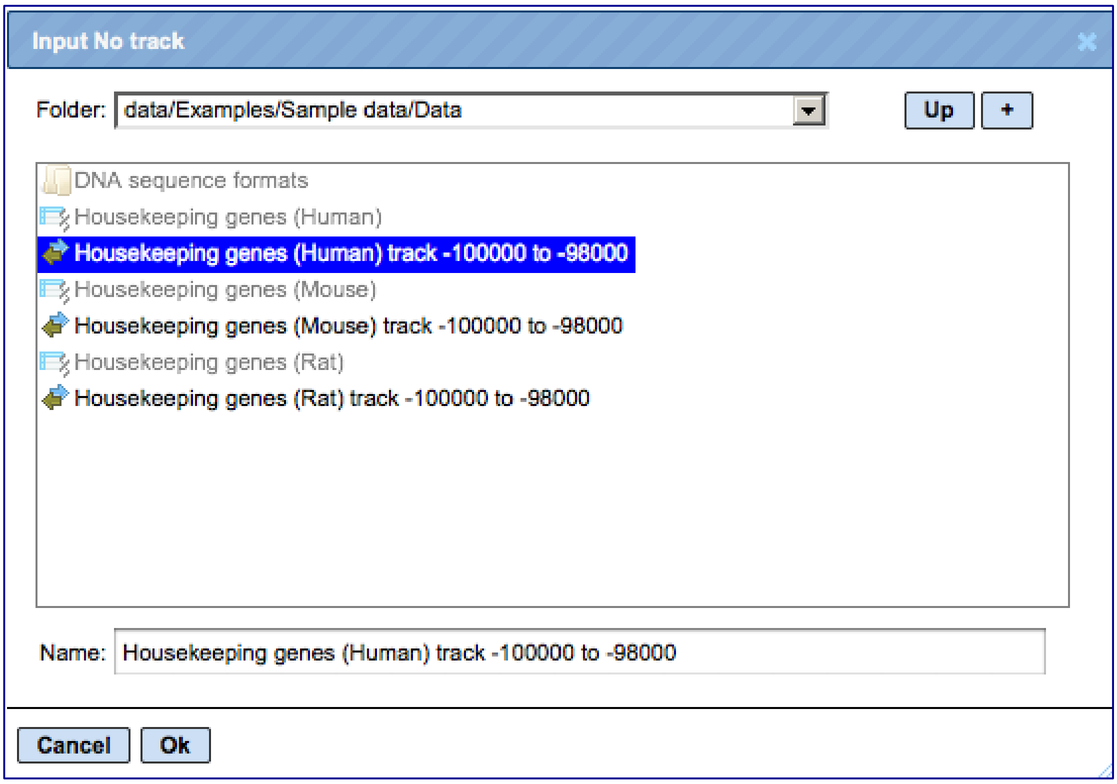
Input No track

Results folder

Ensembl 65.37 Human (hg19)
Ensembl 52.36n Human (hg18)
Ensembl 61.37f Human (hg19)
Ensembl 65.37 Mouse (mm9)
Ensembl 61.37n Mouse (mm9)
Ensembl 65.34 Rat (rn4)
Ensembl 64.34 Rat (rn4)
Run workflow
Edit workflow

Step 4. Specify the biological species of the input set in the field **Species** by selecting the required species from the drop-down menu.

Step 5. Specify No track in BED format in the field **Input No track**. Upon clicking on this field, a supplementary window will open, where you can select the No track from your project tree, or use one of our default No tracks for human, mouse or rat, respectively.



data/Examples/Sample data/Data

- DNA sequence formats
- Housekeeping genes (Human)
- Housekeeping genes (Human) track -100000 to -98000
- Housekeeping genes (Mouse)
- Housekeeping genes (Mouse) track -100000 to -98000
- Housekeeping genes (Rat)
- Housekeeping genes (Rat) track -100000 to -98000

Name: Housekeeping genes (Human) track -100000 to -98000

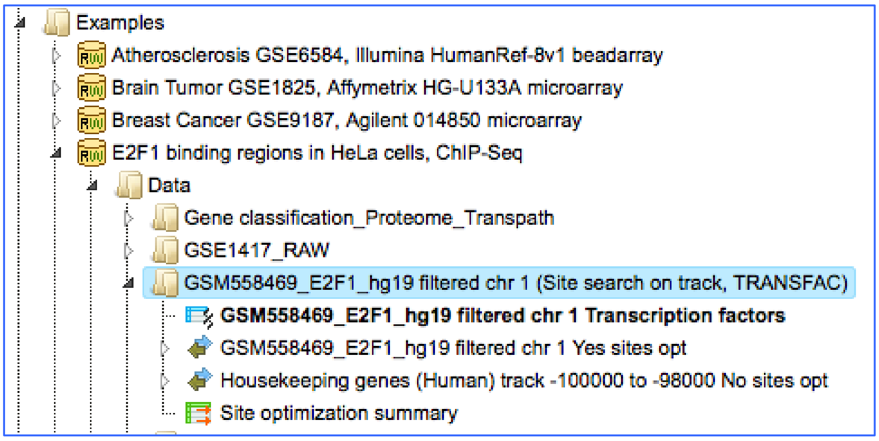
Step 6. Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink field "select element" in the field **Results folder**, and a new window will be opened, where you can select the location of the results folder and define its name.

Step 7. Press the [Run workflow] button.

Ready!

Wait until the workflow is completed.

The results folder contains two tables and two tracks; for this example, let's consider the results folder located under "Examples". It is highlighted by blue in the figure below:



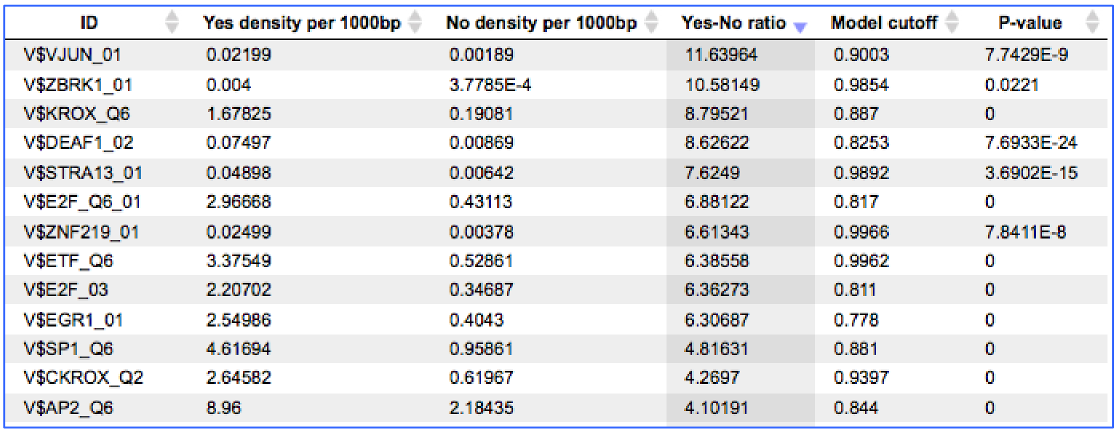
- Examples
  - Atherosclerosis GSE6584, Illumina HumanRef-8v1 beadarray
  - Brain Tumor GSE1825, Affymetrix HG-U133A microarray
  - Breast Cancer GSE9187, Agilent 014850 microarray
  - E2F1 binding regions in HeLa cells, ChIP-Seq
    - Data
      - Gene classification_Proteome_Transpath
      - GSE1417_RAW
      - GSM558469_E2F1_hg19 filtered chr 1 (Site search on track, TRANSFAC)
        - GSM558469_E2F1_hg19 filtered chr 1 Transcription factors
        - GSM558469_E2F1_hg19 filtered chr 1 Yes sites opt
        - Housekeeping genes (Human) track -100000 to -98000 No sites opt
        - Site optimization summary

The tables *Site optimization summary* (▦) and *Transcription factors* (▦) are opened automatically in the Work Space as soon as the workflow is completed.

The table *Site optimization summary* includes the matrices the hits of which are over-represented in the Yes track *versus* the No track.

Please note that only the matrices with Yes-No ratio higher than 1 are included in this output table. The hits of these matrices can be interpreted as over-represented in the Yes set *versus* No set.

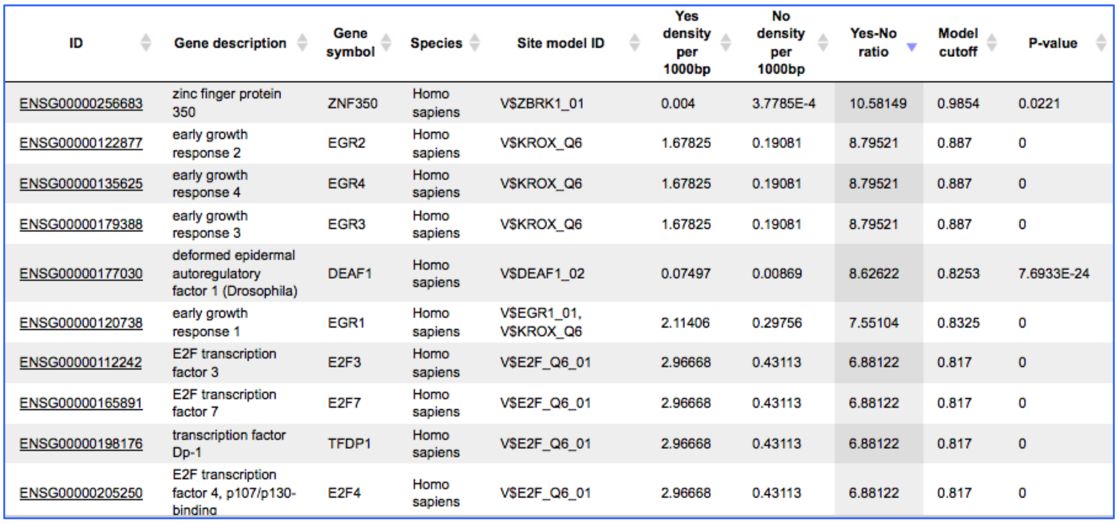
The table *Site optimization summary* shown below has been sorted by the values in the **Yes-No ratio** column.



ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
V$VJUN_01	0.02199	0.00189	11.63964	0.9003	7.7429E-9
V$ZBRK1_01	0.004	3.7785E-4	10.58149	0.9854	0.0221
V$KROX_Q6	1.67825	0.19081	8.79521	0.887	0
V$DEAF1_02	0.07497	0.00869	8.62622	0.8253	7.6933E-24
V$STRA13_01	0.04898	0.00642	7.6249	0.9892	3.6902E-15
V$E2F_Q6_01	2.96668	0.43113	6.88122	0.817	0
V$ZNF219_01	0.02499	0.00378	6.61343	0.9966	7.8411E-8
V$ETF_Q6	3.37549	0.52861	6.38558	0.9962	0
V$E2F_03	2.20702	0.34687	6.36273	0.811	0
V$EGR1_01	2.54986	0.4043	6.30687	0.778	0
V$SP1_Q6	4.61694	0.95861	4.81631	0.881	0
V$CKROX_Q2	2.64582	0.61967	4.2697	0.9397	0
V$AP2_Q6	8.96	2.18435	4.10191	0.844	0

Each row summarizes the information for one PWM. For each selected matrix, the columns **Yes density per 1000bp** and **No density per 1000bp** show the number of matches normalized per 1000 bp length for the sequences in the input Yes set and input No set, respectively. The Column **Yes-No ratio** is the ratio of the first two columns. Only matrices with a Yes-No ratio higher than 1 are included in the *summary* table. The higher the Yes-No ratio, the higher is the enrichment of matches for the respective matrix in the Yes set. The matrix cutoff values as they are calculated by the program at the optimization step are shown in the column **Model cutoff**, and the last column shows the **P-value** of the corresponding event.

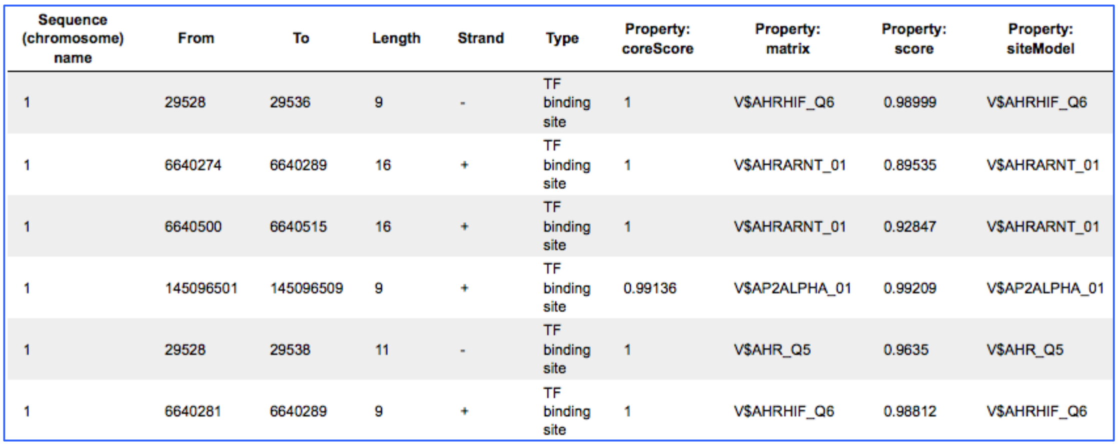
Table Transcription factors:



ID	Gene description	Gene symbol	Species	Site model ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
ENSG00000256683	zinc finger protein 350	ZNF350	Homo sapiens	V$ZBRK1_01	0.004	3.7785E-4	10.58149	0.9854	0.0221
ENSG00000122877	early growth response 2	EGR2	Homo sapiens	V$KROX_Q6	1.67825	0.19081	8.79521	0.887	0
ENSG00000135625	early growth response 4	EGR4	Homo sapiens	V$KROX_Q6	1.67825	0.19081	8.79521	0.887	0
ENSG00000179388	early growth response 3	EGR3	Homo sapiens	V$KROX_Q6	1.67825	0.19081	8.79521	0.887	0
ENSG00000177030	deformed epidermal autoregulatory factor 1 (Drosophila)	DEAF1	Homo sapiens	V$DEAF1_02	0.07497	0.00869	8.62622	0.8253	7.6933E-24
ENSG00000120738	early growth response 1	EGR1	Homo sapiens	V$EGR1_01, V$KROX_Q6	2.11406	0.29756	7.55104	0.8325	0
ENSG00000112242	E2F transcription factor 3	E2F3	Homo sapiens	V$E2F_Q6_01	2.96668	0.43113	6.88122	0.817	0
ENSG00000165891	E2F transcription factor 7	E2F7	Homo sapiens	V$E2F_Q6_01	2.96668	0.43113	6.88122	0.817	0
ENSG00000198176	transcription factor Dp-1	TFDP1	Homo sapiens	V$E2F_Q6_01	2.96668	0.43113	6.88122	0.817	0
ENSG00000205250	E2F transcription factor 4, p107/p130-binding	E2F4	Homo sapiens	V$E2F_Q6_01	2.96668	0.43113	6.88122	0.817	0

This table includes transcription factors (TFs) that are associated with the PWMs that are listed in the table *Site optimization summary*, and each row shows details for one TF, including its Ensembl gene ID (column **ID**), gene symbol, gene description and biological species of the corresponding TF (columns **Gene description**, **Gene symbol**, and **Species**). The column **Site model ID** shows the identifier of the PWM associated with this TF, and several further columns repeat information that is also shown in the table *Site optimization summary*.

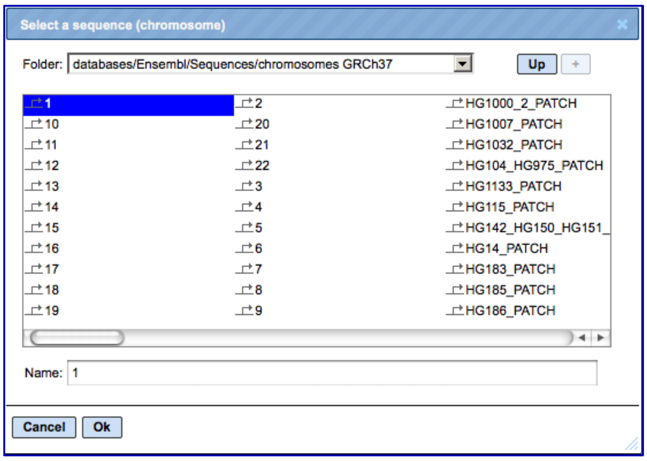
Tracks "Yes sites opt" and "No sites opt" (⇦).



Sequence (chromosome) name	From	To	Length	Strand	Type	Property: coreScore	Property: matrix	Property: score	Property: siteModel
1	29528	29536	9	-	TF binding site	1	V$AHRHIF_Q6	0.98999	V$AHRHIF_Q6
1	6640274	6640289	16	+	TF binding site	1	V$AHRARNT_01	0.89535	V$AHRARNT_01
1	6640500	6640515	16	+	TF binding site	1	V$AHRARNT_01	0.92847	V$AHRARNT_01
1	145096501	145096509	9	+	TF binding site	0.99136	V$AP2ALPHA_01	0.99209	V$AP2ALPHA_01
1	29528	29538	11	-	TF binding site	1	V$AHR_Q5	0.9635	V$AHR_Q5
1	6640281	6640289	9	+	TF binding site	1	V$AHRHIF_Q6	0.98812	V$AHRHIF_Q6

Each row presents details for each individual match for every PWM. Columns **Sequence (chromosome) name**, **From**, **To**, **Length** and **Strand** show the genomic location of the match including chromosome number, start and end positions, strand and length of the match, respectively. The column **Type** contains information about the type of the elements; in this case all matches are considered as "TF binding site". Further columns keep information about PWM producing each match (column **Property:matrix**) as well as a score of the core (column **Property:coreScore**) and a score for the whole matrix (column **Property:score**). The column **Property: siteModel** contains an identifier for the site model, which is the matrix together with the cutoff applied (for details about these scores, please see Kel et al., Nucleic Acids Res. 31:3576-3579, 2003).

Tip. Further visualization of track files in the genome browser: Having tracks "Yes sites opt" and "No sites opt" opened in the Work Space, the menu button ⇦ can be applied to get a visualization. First, a supplementary window is opened where you can select one chromosome and press [Ok], as shown below.



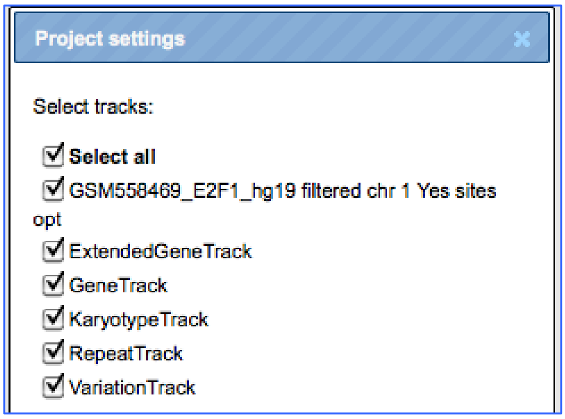
Select a sequence (chromosome)

Folder: databases/Ensembl/Sequences/chromosomes GRCh37

1
10
11
12
13
14
15
16
17
18
19
2
20
21
22
3
4
5
6
7
8
9
HG1000_2_PATCH
HG1007_PATCH
HG1032_PATCH
HG104_HG975_PATCH
HG1133_PATCH
HG115_PATCH
HG142_HG150_HG151_
HG14_PATCH
HG183_PATCH
HG185_PATCH
HG186_PATCH

Name: 1

In the second pop-up window, you can select tracks that can be visualized together with your track, e.g. "Yes sites opt" (see above), and press [Ok].

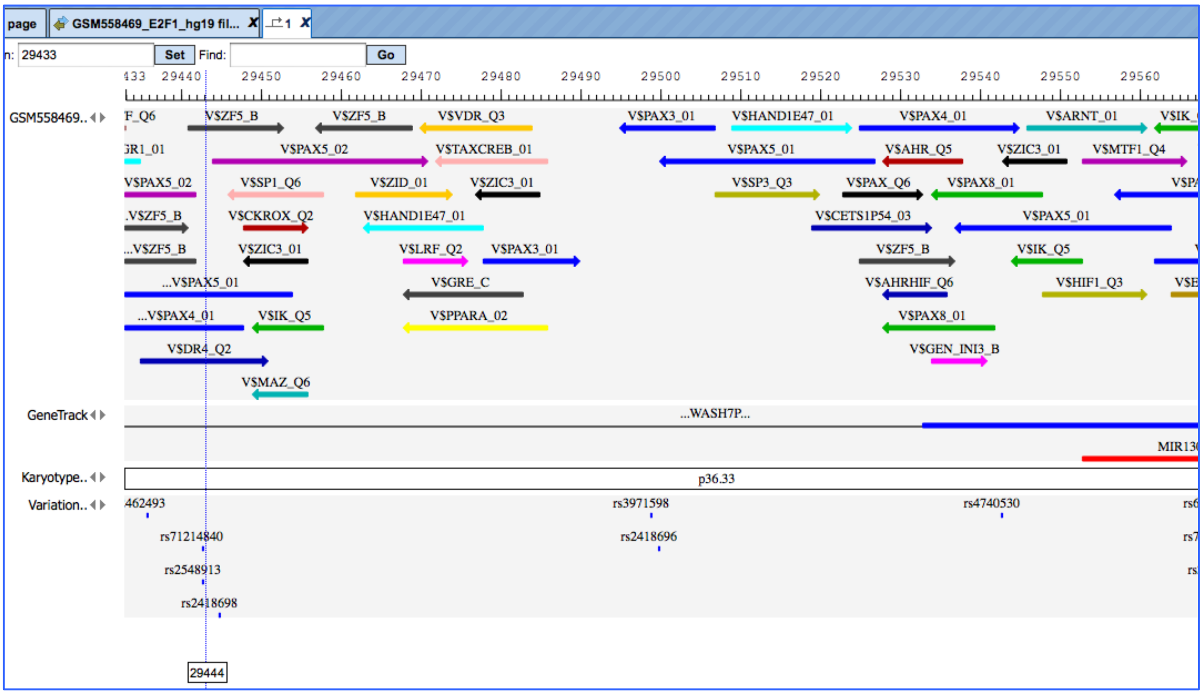


Project settings

Select tracks:

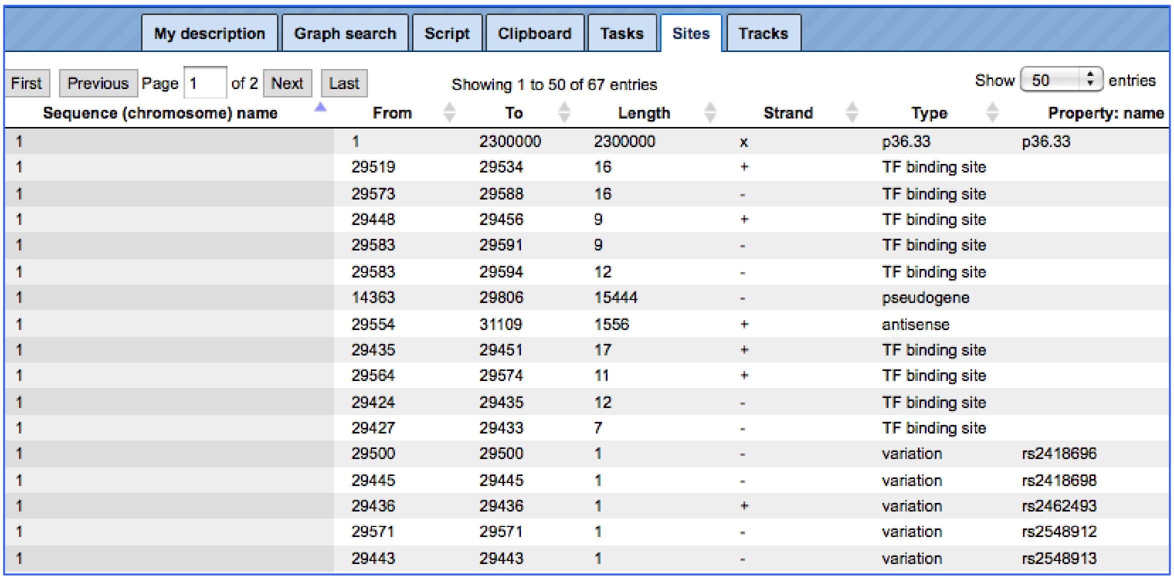
☑ Select all
☑ GSM558469_E2F1_hg19 filtered chr 1 Yes sites opt
☑ ExtendedGeneTrack
☑ GeneTrack
☑ KaryotypeTrack
☑ RepeatTrack
☑ VariationTrack

The resulting visualization, after applying the "zoom in" button , looks like it is shown below. Matches for different matrices are shown in colors, and the color schema can be customized.



Such a view may help to visually co-localize information on different tracks, e.g. putative TFBS with variations, repeats and genes. In the figure above, the cursor shows position 29444, and two variations are located at this position. You can immediately recognize that these variations are located within particular putative binding sites in the intron region of the WASH7P gene.

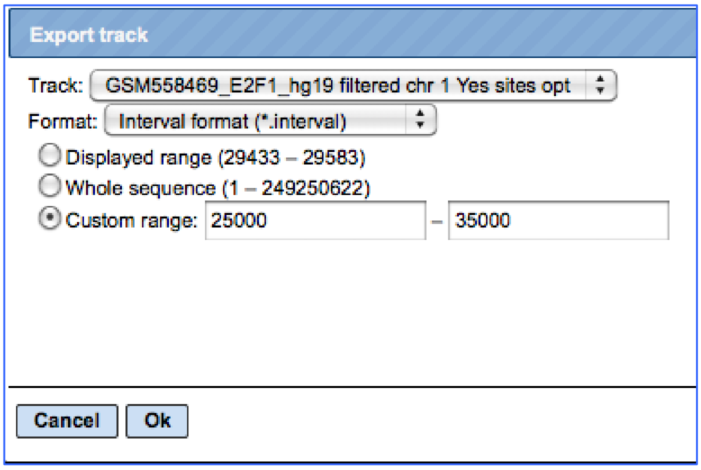
The same information is available not just as a picture, but also as a table under the tab “Sites” (shown below). For each element information is shown on chromosome, positions, length, strand, type of the track, and name of the element.



Showing 1 to 50 of 67 entries

Sequence (chromosome) name	From	To	Length	Strand	Type	Property: name
1	1	2300000	2300000	x	p36.33	p36.33
1	29519	29534	16	+	TF binding site	
1	29573	29588	16	-	TF binding site	
1	29448	29456	9	+	TF binding site	
1	29583	29591	9	-	TF binding site	
1	29583	29594	12	-	TF binding site	
1	14363	29806	15444	-	pseudogene	
1	29554	31109	1556	+	antisense	
1	29435	29451	17	+	TF binding site	
1	29564	29574	11	+	TF binding site	
1	29424	29435	12	-	TF binding site	
1	29427	29433	7	-	TF binding site	
1	29500	29500	1	-	variation	rs2418696
1	29445	29445	1	-	variation	rs2418698
1	29436	29436	1	+	variation	rs2462493
1	29571	29571	1	-	variation	rs2548912
1	29443	29443	1	-	variation	rs2548913

This table can be exported as a track, in several different formats including intervals, bed, wig, gff, gtf and more.



Export track

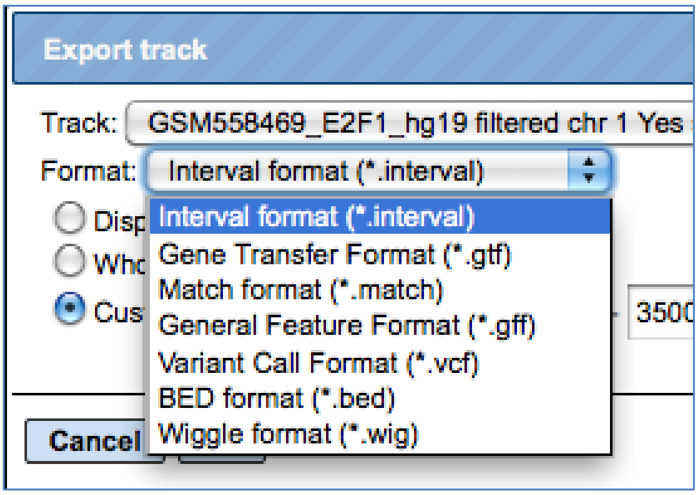
Track: GSM558469_E2F1_hg19 filtered chr 1 Yes sites opt

Format: Interval format (*.interval)

○ Displayed range (29433 – 29583)

○ Whole sequence (1 – 249250622)

● Custom range: 25000 – 35000



Export track

Track: GSM558469_E2F1_hg19 filtered chr 1 Yes

Format: Interval format (*.interval)

Interval format (*.interval)
Gene Transfer Format (*.gtf)
Match format (*.match)
General Feature Format (*.gff)
Variant Call Format (*.vcf)
BED format (*.bed)
Wiggle format (*.wig)
Disp
Whc
Cus
3500
Cancel

Note. This workflow is available together with a valid TRANSFAC® license. Please, feel free to ask for details (info@genexplain.com).

Multiple interval sets

This workflow is designed to search for TFBSs in DNA sequences identified by the ChIP-seq approach, for multiple datasets.

In the field **Input Yes tracks**, several different tracks can be simultaneously submitted. The same background dataset, **Input No track**, is used for comparison with each of the submitted Yes tracks. The default No track corresponds to far upstream regions of the house keeping genes, where no functional TFBSs are expected.

The steps of this workflow for a single input Yes track are described in the previous section. In this workflow, the same steps are performed next time for the 2nd Yes track, and so on iteratively for each of the input Yes tracks.

This workflow helps to save time and efforts, especially when you have several sets of ChIP-seq data, e.g. the peaks for a number of different TFs.

Note. This workflow is available together with a valid TRANSFAC® license. Please, feel free to ask for details (info@genexplain.com).

7.2.3. Search for composite modules with TRANSFAC®

This workflow finds pairs of TFBSs that discriminate between two tracks, the *Yes* and the *No* tracks. As the Yes track, the ChIP-seq peaks identified as binding profiles for particular transcription factors can be considered.

The ChIP-seq experimental technology is widely applied to a variety of biological problems, in particular to study genome-wide histone modification profiles, e.g. histone methylation and histone acetylation profiles. Correspondingly, the same workflow in the platform can be used to analyze histone modification profiles as well. The example of such an application to histone acetylation peaks in SAHA-treated human vascular endothelial cells is described in Section 6.1.3. Please refer to this section for details how to launch the workflow.

Here, let's consider the results of the workflow application to find composite modules in the ChIP-seq peaks identified for in-vivo-bound fragments of transcription factor E2F1 in HeLa cells, published in Gene Expression Omnibus, *GSM558469*.

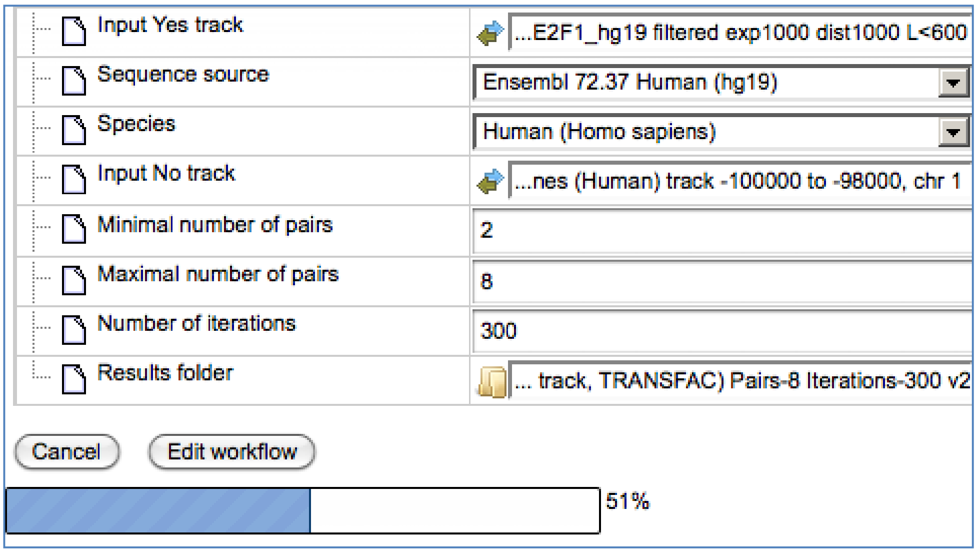
Input Yes track. The original track of genome-wide E2F1 binding fragments was filtered by the length shorter than 600 bp, which resulted in 249 fragments. This track of 249 fragments is used as the input Yes track. It can be found in the *Examples* folder under:

data/Examples/E2F1 binding regions in HeLa cells, ChIP-Seq/Data/GSM558469_E2F1_hg19 filtered exp1000 dist1000 L<600

Input No track. A track of the far upstream fragments of the human housekeeping genes located on chromosome 1 is taken as the No track. It can be found in the *Examples* folder under:

data/Examples/E2F1 binding regions in HeLa cells, ChIP-Seq/Data/ Housekeeping genes (Human) track -100000 to -98000, chr 1

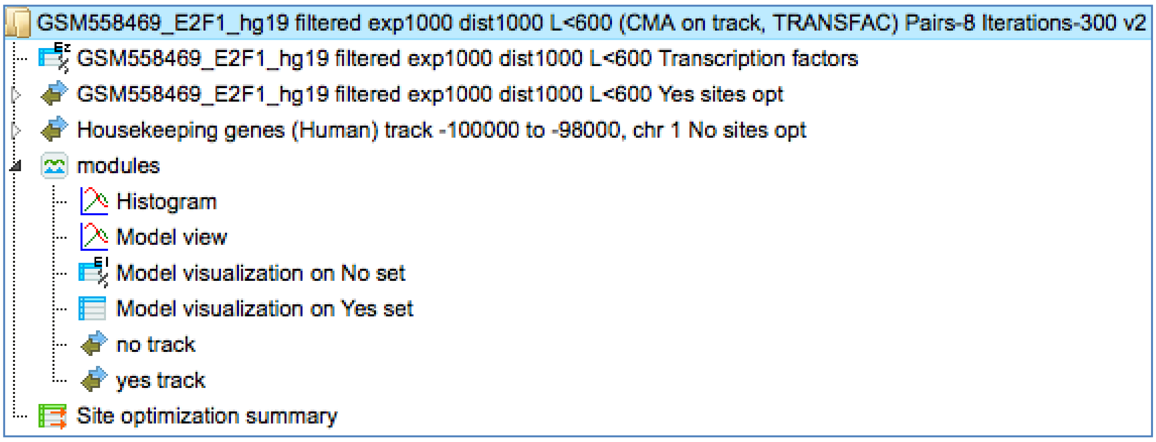
The workflow input form is completed and the run is in progress:



Input Yes track	...E2F1_hg19 filtered exp1000 dist1000 L<600
Sequence source	Ensembl 72.37 Human (hg19)
Species	Human (Homo sapiens)
Input No track	...nes (Human) track -100000 to -98000, chr 1
Minimal number of pairs	2
Maximal number of pairs	8
Number of iterations	300
Results folder	... track, TRANSFAC) Pairs-8 Iterations-300 v2

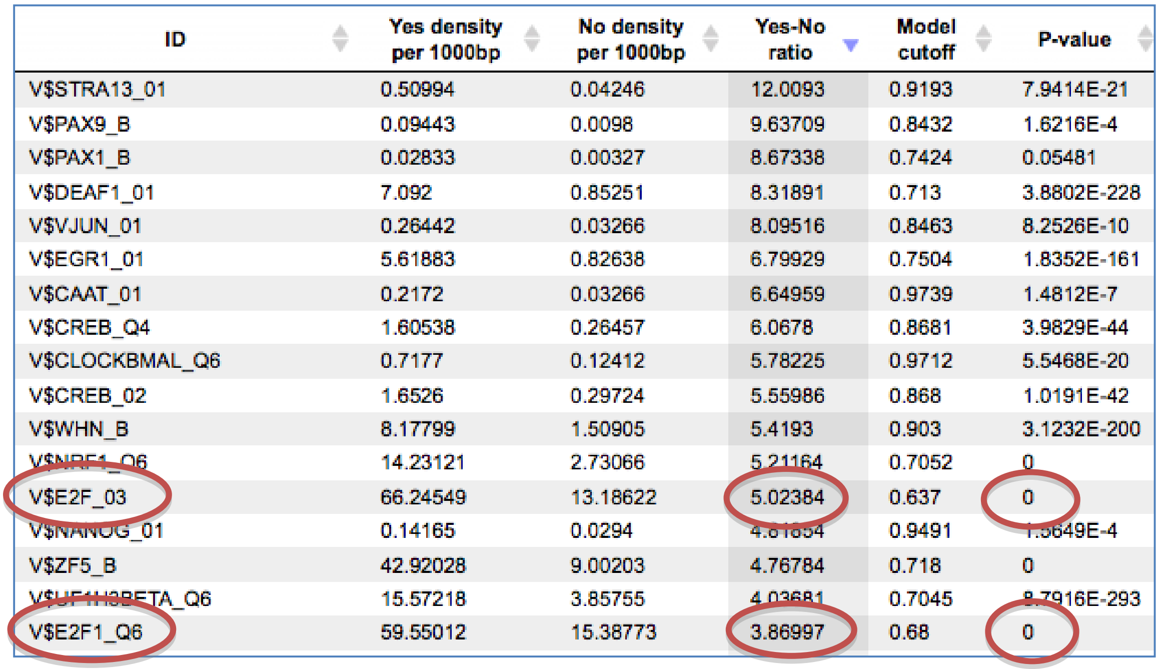
The resulting folder can be found under:

data/Examples/E2F1 binding regions in HeLa cells, ChIP-Seq/Data/GSM558469_E2F1_hg19 filtered exp1000 dist1000 L<600 (CMA on track, TRANSFAC) Pairs-8 Iterations-300 v2



- GSM558469_E2F1_hg19 filtered exp1000 dist1000 L<600 (CMA on track, TRANSFAC) Pairs-8 Iterations-300 v2
  - GSM558469_E2F1_hg19 filtered exp1000 dist1000 L<600 Transcription factors
  - GSM558469_E2F1_hg19 filtered exp1000 dist1000 L<600 Yes sites opt
  - Housekeeping genes (Human) track -100000 to -98000, chr 1 No sites opt
  - modules
    - Histogram
    - Model view
    - Model visualization on No set
    - Model visualization on Yes set
    - no track
    - yes track
  - Site optimization summary

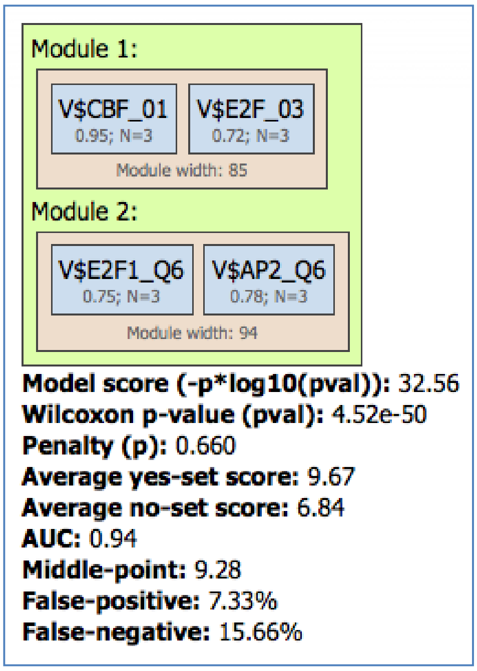
The table **Site optimization summary** ( ) contains those site models, here TRANSFAC® matrices, that are over-represented in the Yes track as compared to the No track.



ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
V$STRA13_01	0.50994	0.04246	12.0093	0.9193	7.9414E-21
V$PAX9_B	0.09443	0.0098	9.63709	0.8432	1.6216E-4
V$PAX1_B	0.02833	0.00327	8.67338	0.7424	0.05481
V$DEAF1_01	7.092	0.85251	8.31891	0.713	3.8802E-228
V$VJUN_01	0.26442	0.03266	8.09516	0.8463	8.2526E-10
V$EGR1_01	5.61883	0.82638	6.79929	0.7504	1.8352E-161
V$CAAT_01	0.2172	0.03266	6.64959	0.9739	1.4812E-7
V$CREB_Q4	1.60538	0.26457	6.0678	0.8681	3.9829E-44
V$CLOCKBMAL_Q6	0.7177	0.12412	5.78225	0.9712	5.5468E-20
V$CREB_02	1.6526	0.29724	5.55986	0.868	1.0191E-42
V$WHN_B	8.17799	1.50905	5.4193	0.903	3.1232E-200
V$NRF1_Q6	14.23121	2.73066	5.21164	0.7052	0
V$E2F_03	66.24549	13.18622	5.02384	0.637	0
V$NANOG_01	0.14165	0.0294	4.81854	0.9491	1.5649E-4
V$ZF5_B	42.92028	9.00203	4.76784	0.718	0
V$UF1H3BETA_Q6	15.57218	3.85755	4.03681	0.7045	8.7916E-293
V$E2F1_Q6	59.55012	15.38773	3.86997	0.68	0

Each row of the table represents the result for one PWM from the input profile. Only those PWMs with Yes-No ratio >1 are included in the output. Upon sorting by the Yes-No ratio, matrices for E2F factors are among top 20 lines. Please note that the p-values of E2F matrices are extremely low, which demonstrates highest statistical significance of the results.

The **Modules** folder ( ). The composite module found contains two pairs, and we can see by exactly which site models (matrices) these pairs are formed as well as the statistical parameters of the overall model.


Module 1:

V$CBF_01 0.95; N=3

V$E2F_03 0.72; N=3

Module width: 85

Module 2:

V$E2F1_Q6 0.75; N=3

V$AP2_Q6 0.78; N=3

Module width: 94

Model score (-p*log10(pval)): 32.56
Wilcoxon p-value (pval): 4.52e-50
Penalty (p): 0.660
Average yes-set score: 9.67
Average no-set score: 6.84
AUC: 0.94
Middle-point: 9.28
False-positive: 7.33%
False-negative: 15.66%

Both pairs contain matrices for E2F factors.

For more details on the individual output tables and tracks as well as for **visualization** of the identified composite modules in the genome browser please refer to Section 20.1.5.

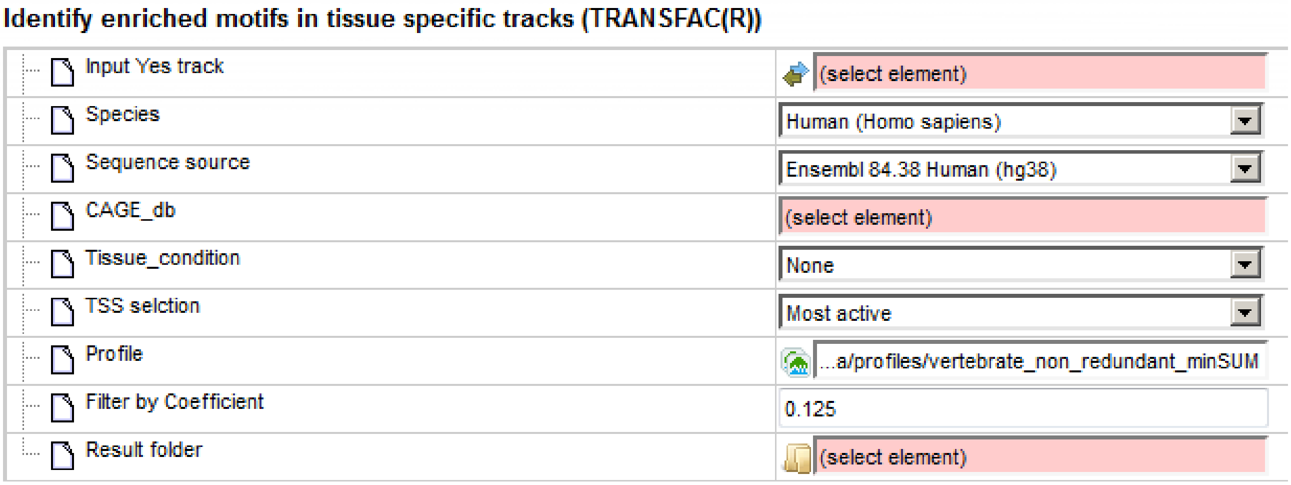
Note. This workflow is available together with a valid TRANSFAC® license.
Please, feel free to ask for details (info@genexplain.com).

7.2.4. Search for discriminative sites with TRANSFAC® (MEALR)

Please, refer to Section 6.1.4 for detailed explanation how to use this function.

7.2.5. Search for enriched TF sites from tissue specific promoter tracks

This workflow searches for enriched transcription factor binding sites (TFBSs) in input tracks versus a random created track. The input track is converted to a gene set, which is used to extract promoter regions by mapping it against the TSS locations defined in CAGE data in the Fantom5 (Nature 507:462–470) database (see also 19.10). The over-represented sites identified with the MEALR method are converted into a profile, which is used for a second round of site search, and ends up with the identification of transcription factors. To launch the workflow, open the workflow input form from the Start page:



Identify enriched motifs in tissue specific tracks (TRANSFAC(R))

Input Yes track	(select element)
Species	Human (Homo sapiens)
Sequence source	Ensembl 84.38 Human (hg38)
CAGE_db	(select element)
Tissue_condition	None
TSS selction	Most active
Profile	...a/profiles/vertebrate_non_redundant_minSUM
Filter by Coefficient	0.125
Result folder	(select element)

Step 1: To specify the **Input Yes track**, you can drag & drop it from your project within the tree area. Alternatively, you may click on the pink field "select element" and a new window will open, where you select the input track. After having selected the track, press the [Ok] button.

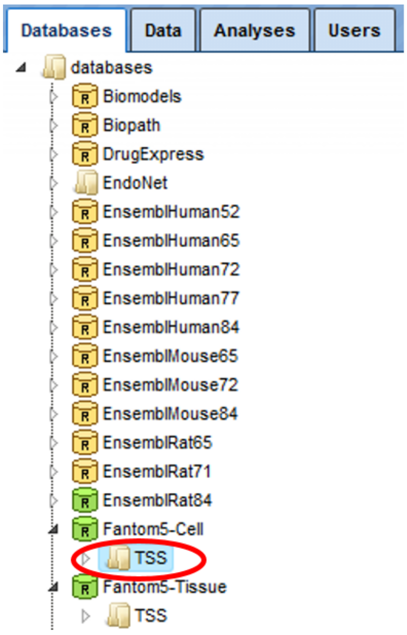
For this example, all further steps are demonstrated with the following input set:

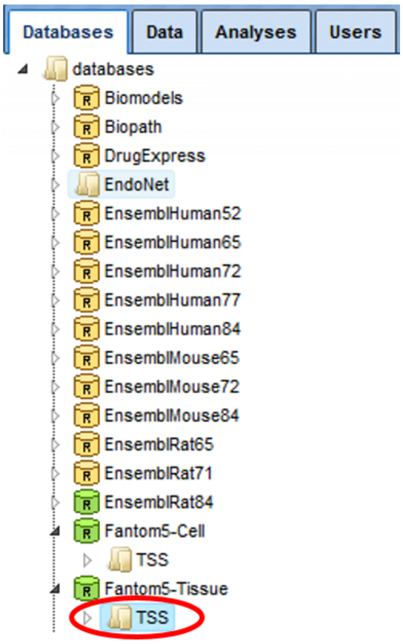
http://genexplain-platform.com:8080/bioumlweb/#de=data/Examples/E2F1%20binding%20regions%20in%20HeLa%20cells%2C%20ChIP-Seq/Data/GSM558469_E2F1_hg19%20filtered%20exp1000%20dist1000

Step 2: Specify the biological species of the input set in the field **Species** by selecting the desired species from the drop-down menu.

Step 3: Specify the **Sequence source** from the drop-down menu. Several human, mouse and rat sequence builds are available in the platform, as shown below. By default, the most recent Ensembl human genome, hg38, is specified. Make sure you select the sequence source (the genome build) that corresponds to your input set in order to get correct and meaningful results.

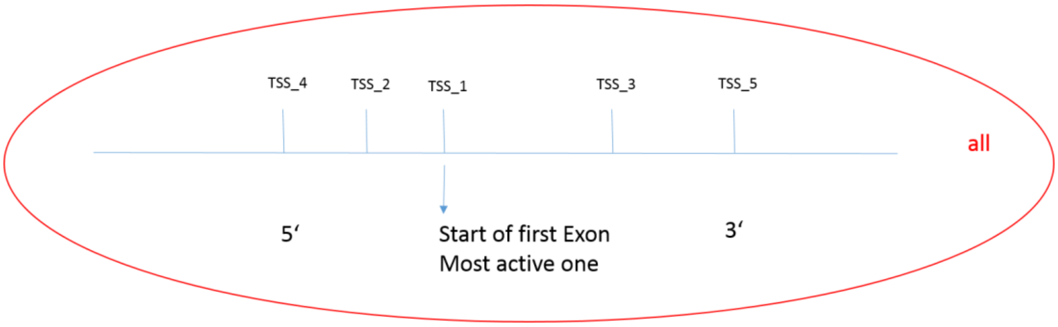
Step 4: Specify the path of the **CAGE_db**. Select either the TSS folder from Fantom5-Cell database or the TSS folder from Fantom5-Tissue database. You can drag & drop it from the Databases within the tree area, as shown below.





Step 5: **Tissue condition**, select the cells/tissues for which you want to create the promoter track from the drop-down menu.

Step 6: The **TSS selection** should be performed if there are multiple transcription start sites. By default, the most active site is considered as TSS. You can select between most active, 5’ active, 3’ active and all (see below) from the drop-down menu.



Step 7: Define a TRANSFAC® profile. The default profile is vertebrate_non_redundant_minSUM. Any other TRANSFAC® profile or user-specific profile can be chosen. With a mouse click on the field **Profile**, a pop-up window will open, where a profile can be selected.

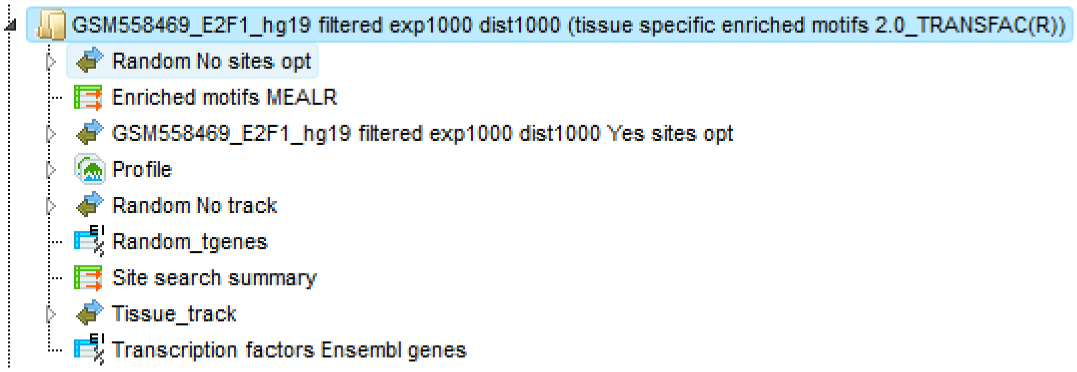
Step 8: Select a filter for the coefficient of the MEALR method. The default filter is set as >0.125 to have 75% or more of true discovery rate, TDR. For a 90% TDR, you can type 0.270 in this field and for 50% TDR - 0.05593. The filtered motifs are included in the output as *enriched motifs*. At the later step, PWMs corresponding to the enriched motifs are used to create a new profile.

Step 9: Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink box (select element) in the field **Results folder**, and a new window will open, where you can select the location of the results folder and define its name.

Step 10: Press the [Run workflow] button. Wait until the workflow is completed, and take a look at the results.

Interpretation of results

The result folder contains several files and one profile (collection of matrices):



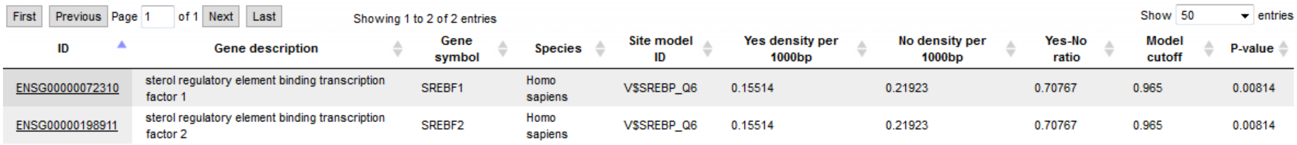
GSM558469_E2F1_hg19 filtered exp1000 dist1000 (tissue specific enriched motifs 2.0_TRANSFAC(R))
- Random No sites opt
- Enriched motifs MEALR
- GSM558469_E2F1_hg19 filtered exp1000 dist1000 Yes sites opt
- Profile
- Random No track
- Random_tgenes
- Site search summary
- Tissue_track
- Transcription factors Ensembl genes

Tracks

The Random No track includes 1000 sequences from the same sequence source as the input track. The random no track and the input track are converted to genes, which are used as input for the tissue-specific promoter track from selected database, tissue and TSS. The resulting tracks are *Tissue_track* (from Yes track) and *Random_tissue_track* (from random track).

Enriched motifs

The list of motifs, which were found during the first part of the workflow, and were filtered by the coefficient >0.05, can be found in the table *Enriched_motifs MEALR* ( ). It contains those site models, here TRANSFAC® matrices, which are enriched in the *Tissue_track* in comparison with the *Random_tissue_track*. The example has x detected motifs with coefficient >0.05. The *Profile* contains the matrix collection of converted and filtered site models. The table *Transcription factors Ensembl genes* includes the corresponding x TFs from the second *Site search summary* and is shown below:



First Previous Page 1 of 1 Next Last Showing 1 to 2 of 2 entries Show 50 entries

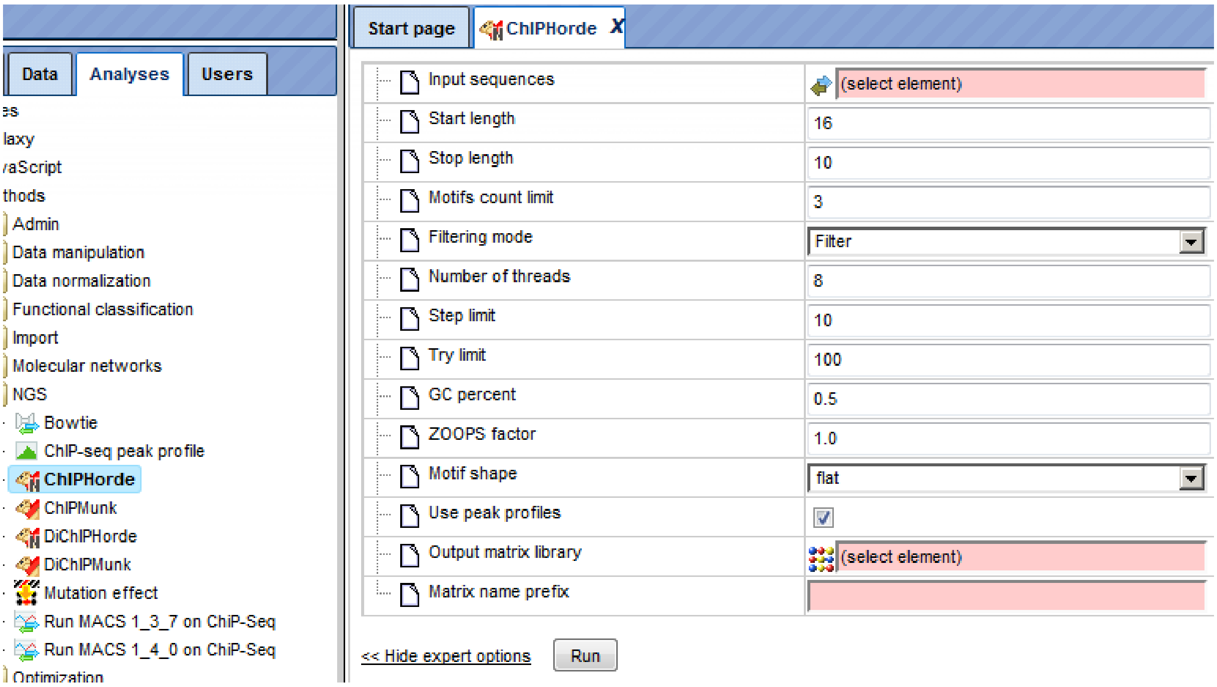
ID	Gene description	Gene symbol	Species	Site model ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
ENSG00000072310	sterol regulatory element binding transcription factor 1	SREBF1	Homo sapiens	V$SREBP_Q6	0.15514	0.21923	0.70767	0.965	0.00814
ENSG00000198911	sterol regulatory element binding transcription factor 2	SREBF2	Homo sapiens	V$SREBP_Q6	0.15514	0.21923	0.70767	0.965	0.00814

> **Note**. This workflow is available together with valid TRANSFAC® and TRANSPATH® licenses. Please feel free to ask for details (info@genexplain.com).

7.2.6. Discover de-novo motifs using ChIPHorde and DiChIPHorde

ChIPHorde uses the fast heuristic of ChIPMunk, which is based on a greedy approach accompanied by bootstrapping, to search for multiple significant motifs in a given dataset using two independent filtering strategies. DiChIPHorde searches for di-nucleotide motifs that capture dependencies between neighboring motif positions.

The image below shows the ChIPHorde interface. The input mask of DiChIPHorde features only slight differences as pointed out in the section describing input parameters.



Input sequences	(select element)
Start length	16
Stop length	10
Motifs count limit	3
Filtering mode	Filter
Number of threads	8
Step limit	10
Try limit	100
GC percent	0.5
ZOOPS factor	1.0
Motif shape	flat
Use peak profiles	☑
Output matrix library	(select element)
Matrix name prefix	

The parameters are described in the following. For further details, please also refer to the ChIPMunk manual.

Input sequences: Track with input reads

Start length: Start length of the matrix

Stop length: Stop length of the matrix

Motifs count limit: Maximum number of motifs to discover

Filtering mode: Whether to mask polyN ("Mask") or to drop entire sequence ("Filter")

Number of threads (expert): Number of concurrent threads when processing

Step limit (expert): The number of bootstrapping runs. With a large data set or if unsure which computational time to expect, try a small number, e.g. 10.

Try limit (expert): This is a number of general optimization runs. For a random seeding this is equal to the number of seeds. As with “step limit”, apply the parameter with the size of your data set in mind. It is always advisable to test with a small number first.

Local background (expert): ***DiChIPHorde only*** If checked, local background estimation is used. Otherwise uniform background estimation is used

GC percent (expert): ***ChIPHorde only*** Relative GC content to be used as nucleotide background (0..1). Set to -1 to use the observed GC content.

ZOOPS factor (expert): Sets the preference for ZOOPS (Zero-or-one[-motif occurrences]-per-sequence) versus OOPS (Only-one-per-sequence) mode.

Motif shape (expert): The type of motif shape prior to use. The parameter corresponds a prior on the number of informative regions within the motif.

Use peak profiles (expert): Whether to apply peak profiles (if available).

Output matrix library: Path to the matrix library to be extended or created.

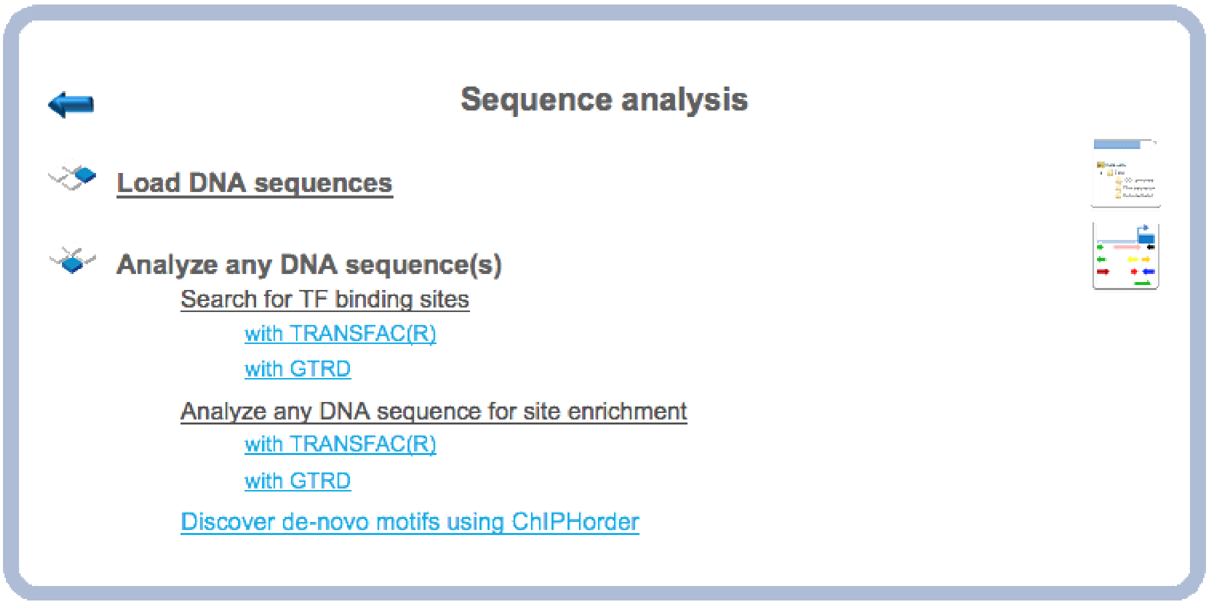
Matrix name prefix: Prefix for matrix names. It will be appended with a number.

7.3. Further workflows in this area

For the other workflows that you can find in the area ***ChIP-seq***, please refer to the following Sections:

Load data	See Chapter 3
Discover functional enrichment of target genes	See Section 10.3
Analyze networks of target genes	See Section 5.1

8. Sequence analysis



Sequence analysis

Load DNA sequences

Analyze any DNA sequence(s)

Search for TF binding sites

with TRANSFAC(R)

with GTRD

Analyze any DNA sequence for site enrichment

with TRANSFAC(R)

with GTRD

Discover de-novo motifs using ChIPHorder

8.1. Analyze any DNA sequence

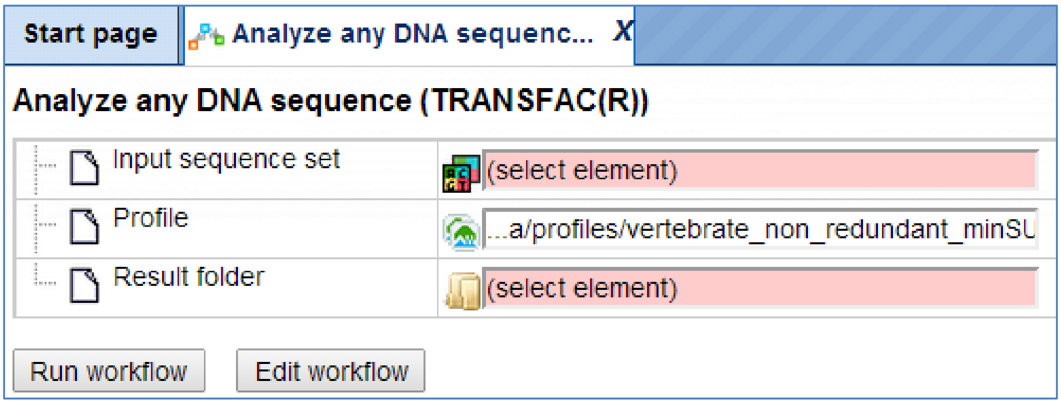
8.1.1. Search for TF binding sites

8.1.1.1. Search for TF binding sites with TRANSFAC®

This workflow is designed to search for putative transcription factor binding sites, TFBS, in any input DNA sequence in EMBL, Fasta or Genbank formats. Using this workflow you can analyze DNA sequences of any species and of any genomic regions.

The steps to launch the workflow are as follows.

Step 1. Open the workflow input form from the Start page, it will open in the main work area and looks as it is shown below:

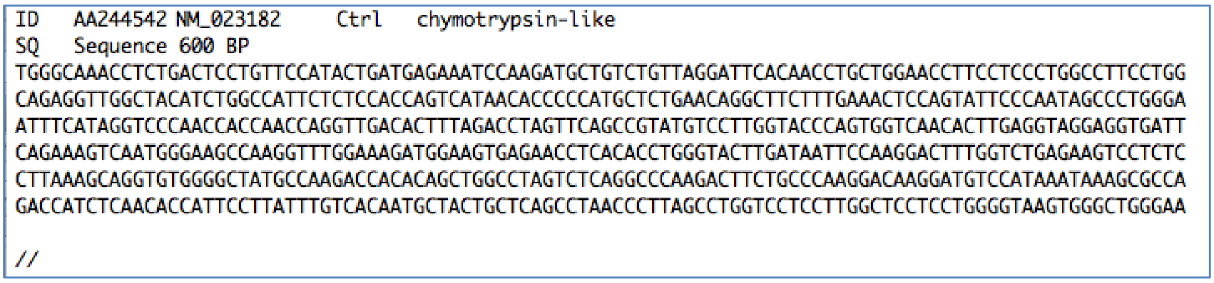


Analyze any DNA sequence (TRANSFAC(R))

Input sequence set	(select element)
Profile	...a/profiles/vertebrate_non_redundant_minSU
Result folder	(select element)

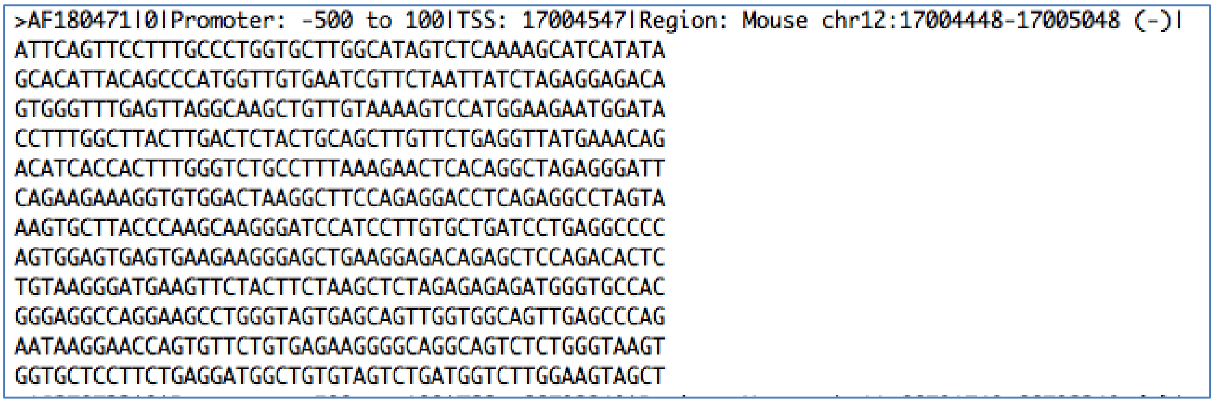
Step 2. Specify an input file in the field **Input sequence set.** The input sequence set can be any sequence file having an EMBL, Genbank, or Fasta file extension. Sample EMBL, FASTA and Genbank sequences are as shown below:

Sample EMBL file:



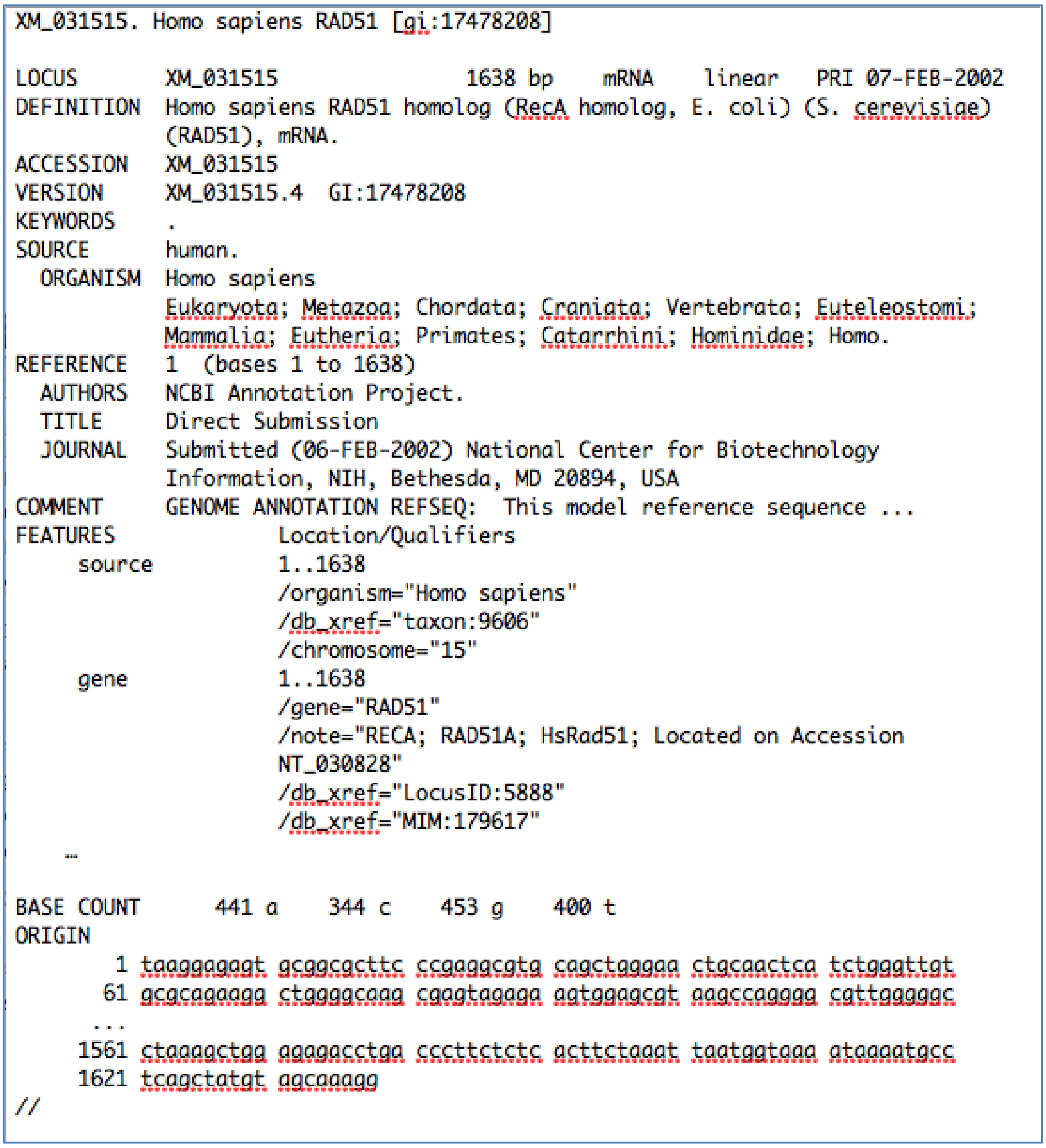
```
ID   AA244542 NM_023182     Ctrl   chymotrypsin-like
SQ   Sequence 600 BP
TGGGCAAACCTCTGACTCCTGTTCCATACTGATGAGAAATCCAAGATGCTGTCTGTTAGGATTCACAACCTGCTGGAACCTTCCTCCCTGGCCTTCCTGG
CAGAGGTTGGCTACATCTGGCCATTCTCTCCACCAGTCATAACACCCCCATGCTCTGAACAGGCTTCTTTGAAACTCCAGTATTCCCAATAGCCCTGGGA
ATTTCATAGGTCCCAACCACCAACCAGGTTGACACTTTAGACCTAGTTCAGCCGTATGTCCTTGGTACCCAGTGGTCAACACTTGAGGTAGGAGGTGATT
CAGAAAGTCAATGGGAAGCCAAGGTTTGGAAAGATGGAAGTGAGAACCTCACACCTGGGTACTTGATAATTCCAAGGACTTTGGTCTGAGAAGTCCTCTC
CTTAAAGCAGGTGTGGGGCTATGCCAAGACCACACAGCTGGCCTAGTCTCAGGCCCAAGACTTCTGCCCAAGGACAAGGATGTCCATAAATAAAGCGCCA
GACCATCTCAACACCATTCCTTATTTGTCACAATGCTACTGCTCAGCCTAACCCTTAGCCTGGTCCTCCTTGGCTCCTCCTGGGGTAAGTGGGCTGGGAA

//
```

Sample Fasta file:



```
>AF180471|0|Promoter: -500 to 100|TSS: 17004547|Region: Mouse chr12:17004448-17005048 (-)|
ATTCAGTTCCTTTGCCCTGGTGCTTGGCATAGTCTCAAAAGCATCATATA
GCACATTACAGCCCATGGTTGTGAATCGTTCTAATTATCTAGAGGAGACA
GTGGGTTTGAGTTAGGCAAGCTGTTGTAAAAGTCCATGGAAGAATGGATA
CCTTTGGCTTACTTGACTCTACTGCAGCTTGTTCTGAGGTTATGAAACAG
ACATCACCACTTTGGGTCTGCCTTTAAAGAACTCACAGGCTAGAGGGATT
CAGAAGAAAGGTGTGGACTAAGGCTTCCAGAGGACCTCAGAGGCCTAGTA
AAGTGCTTACCCAAGCAAGGGATCCATCCTTGTGCTGATCCTGAGGCCCC
AGTGGAGTGAGTGAAGAAGGGAGCTGAAGGAGACAGAGCTCCAGACACTC
TGTAAGGGATGAAGTTCTACTTCTAAGCTCTAGAGAGAGATGGGTGCCAC
GGGAGGCCAGGAAGCCTGGGTAGTGAGCAGTTGGTGGCAGTTGAGCCCAG
AATAAGGAACCAGTGTTCTGTGAGAAGGGGCAGGCAGTCTCTGGGTAAGT
GGTGCTCCTTCTGAGGATGGCTGTGTAGTCTGATGGTCTTGGAAGTAGCT
```

Sample Genbank file:



```
XM_031515. Homo sapiens RAD51 [gi:17478208]

LOCUS       XM_031515               1638 bp    mRNA    linear   PRI 07-FEB-2002
DEFINITION  Homo sapiens RAD51 homolog (RecA homolog, E. coli) (S. cerevisiae)
            (RAD51), mRNA.
ACCESSION   XM_031515
VERSION     XM_031515.4  GI:17478208
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1638)
  AUTHORS   NCBI Annotation Project.
  TITLE     Direct Submission
  JOURNAL   Submitted (06-FEB-2002) National Center for Biotechnology
            Information, NIH, Bethesda, MD 20894, USA
COMMENT     GENOME ANNOTATION REFSEQ:  This model reference sequence ...
FEATURES             Location/Qualifiers
     source          1..1638
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /chromosome="15"
     gene            1..1638
                     /gene="RAD51"
                     /note="RECA; RAD51A; HsRad51; Located on Accession
                     NT_030828"
                     /db_xref="LocusID:5888"
                     /db_xref="MIM:179617"
    …

BASE COUNT      441 a    344 c    453 g    400 t
ORIGIN
        1 taaggaggagt gcggcacttc ccgaggcgta cagctaggaa ctgcaactca tctggattgt
       61 gcgcagaggg ctggggcaag cgagtagaga agtggagcgt aagccagggg cgttggggac
      ...
     1561 ctaaagctgg aggagcctga cccttctctc acttctaaat taatggtaaa ataaaatgcc
     1621 tcagctatgt agcaaagg
//
```

After loading into the platform, the sequences in these formats are shown in the tree area with the icon (ACGT).

In this example, sample sequences of *Arabidopsis thaliana* are used, which were downloaded from the TAIR database in FASTA format (http://www.arabidopsis.org/tools/bulk/sequences/index.jsp). The example input sequence set contains a set of ten promoter sequences 500 bp upstream of the TSS, located on chromosome 1.

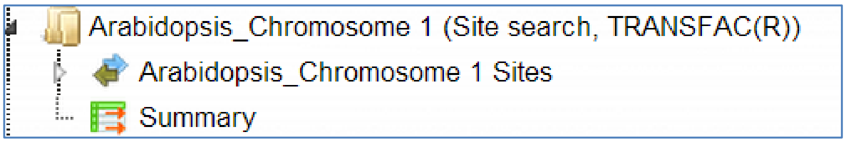
Step3. Select the TRANSFAC® profile from the available profiles. The default profile is *vertebrate_non_redundant_minSUM*. In this example we use the profile called *Plants*. It can be found here:

http://genexplain-platform.com/bioumlweb/#de=databases/TRANSFAC(R)%202014.4/Data/profiles/plants

Step 4. Specify the result folder location and name in your *Project* area. Then press the button [Run workflow]. Wait till the workflow is completed.

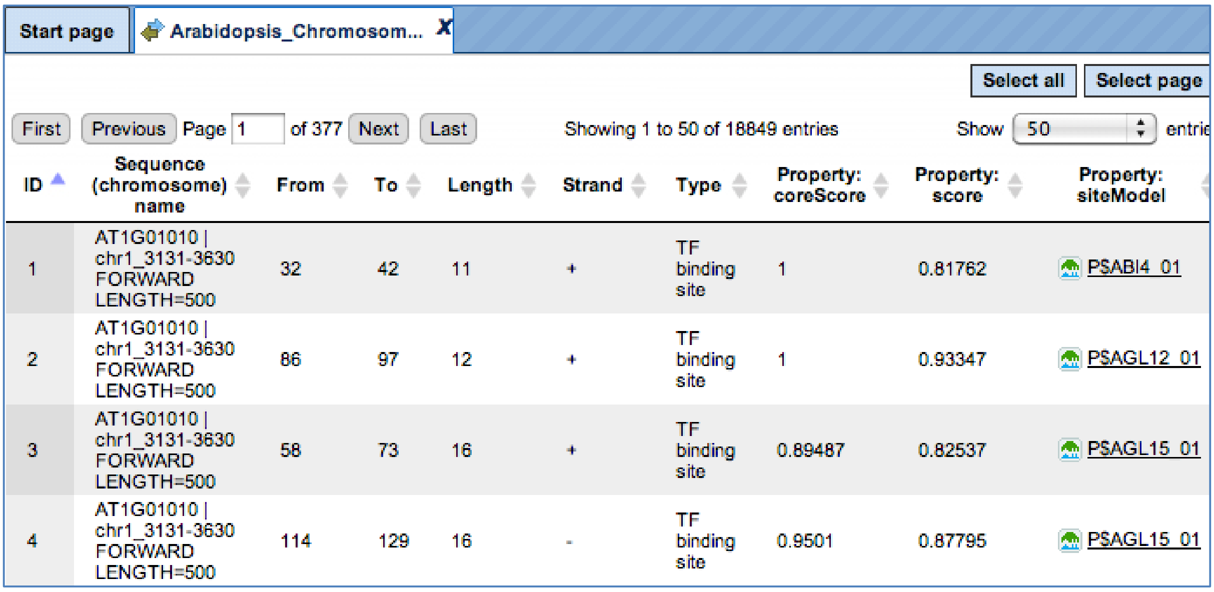
Results

The results folder consists of a summary table and a track with sites as shown below:



Arabidopsis_Chromosome 1 (Site search, TRANSFAC(R))
- Arabidopsis_Chromosome 1 Sites
- Summary

Input sequence Sites: This track ( ) shows TFBSs that are found in the input sequences. As the input sequence set in this example is called *Arabidopsis_Chromosome 1*, the resulting track is called *Arabidopsis_Chromosome 1 Sites.* When opened as a table, this track looks like:

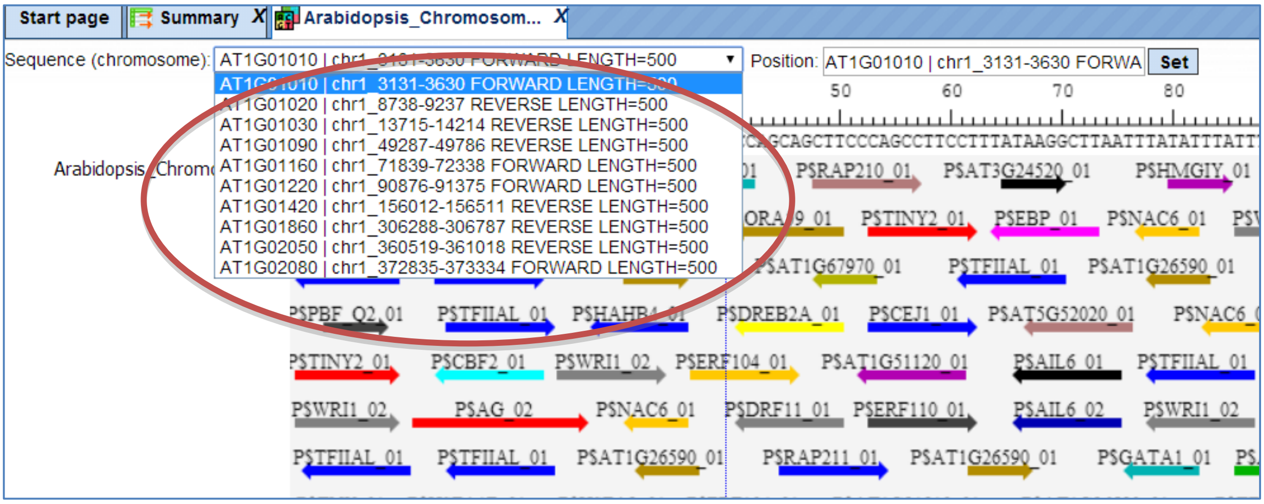


Showing 1 to 50 of 18849 entries

ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: coreScore	Property: score	Property: siteModel
1	AT1G01010 | chr1_3131-3630 FORWARD LENGTH=500	32	42	11	+	TF binding site	1	0.81762	PSABI4_01
2	AT1G01010 | chr1_3131-3630 FORWARD LENGTH=500	86	97	12	+	TF binding site	1	0.93347	PSAGL12_01
3	AT1G01010 | chr1_3131-3630 FORWARD LENGTH=500	58	73	16	+	TF binding site	0.89487	0.82537	PSAGL15_01
4	AT1G01010 | chr1_3131-3630 FORWARD LENGTH=500	114	129	16	-	TF binding site	0.9501	0.87795	PSAGL15_01

Each row corresponds to one resulting TFBS and includes sequence names, site positions (the columns **From** and **To**), site **Length** and **Strand**, **score** calculated by the algorithm and a site model (here, TRANSFAC® matrix). This table can be exported as a track in several different formats including intervals, bed, wig and more. DNA sequences can be exported in multi-FASTA format.

The same track, when opened in the genome browser, looks as shown below:



Start page | Summary X | Arabidopsis_Chromosom... X

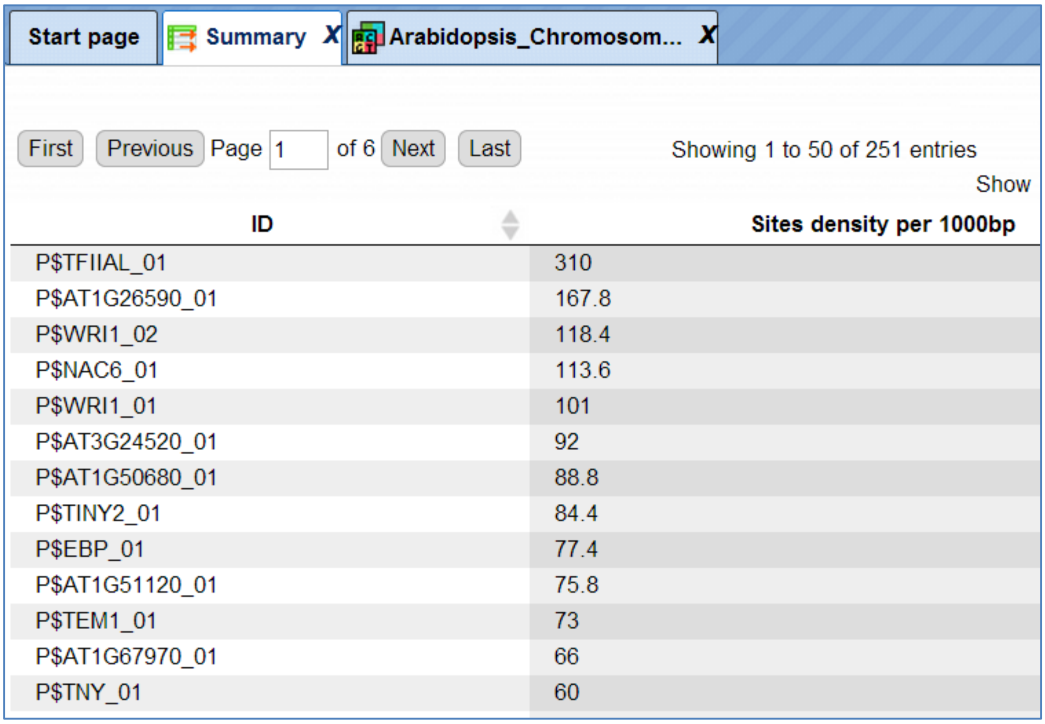
Sequence (chromosome): AT1G01010 | chr1_3131-3630 FORWARD LENGTH=500

- AT1G01010 | chr1_3131-3630 FORWARD LENGTH=500
- AT1G01020 | chr1_8738-9237 REVERSE LENGTH=500
- AT1G01030 | chr1_13715-14214 REVERSE LENGTH=500
- AT1G01090 | chr1_49287-49786 REVERSE LENGTH=500
- AT1G01160 | chr1_71839-72338 FORWARD LENGTH=500
- AT1G01220 | chr1_90876-91375 FORWARD LENGTH=500
- AT1G01420 | chr1_156012-156511 REVERSE LENGTH=500
- AT1G01860 | chr1_306288-306787 REVERSE LENGTH=500
- AT1G02050 | chr1_360519-361018 REVERSE LENGTH=500
- AT1G02080 | chr1_372835-373334 FORWARD LENGTH=500

Position: AT1G01010 | chr1_3131-3630 FORWA Set

In the field *Sequence (chromosome)* you can find a dropdown menu, highlighted by the red oval. This feature helps to easily switch between visualizations of the sequences in the input set. In this particular example the input sequence set comprises ten individual promoter sequences, and each individual promoter can be visualized in the genome browser.

The table Summary (📊) gives the site density per thousand bp for each matrix in the input sequence. When opened in the work space looks as shown below:



Start page | Summary X | Arabidopsis_Chromosom... X

First | Previous | Page 1 of 6 | Next | Last

Showing 1 to 50 of 251 entries

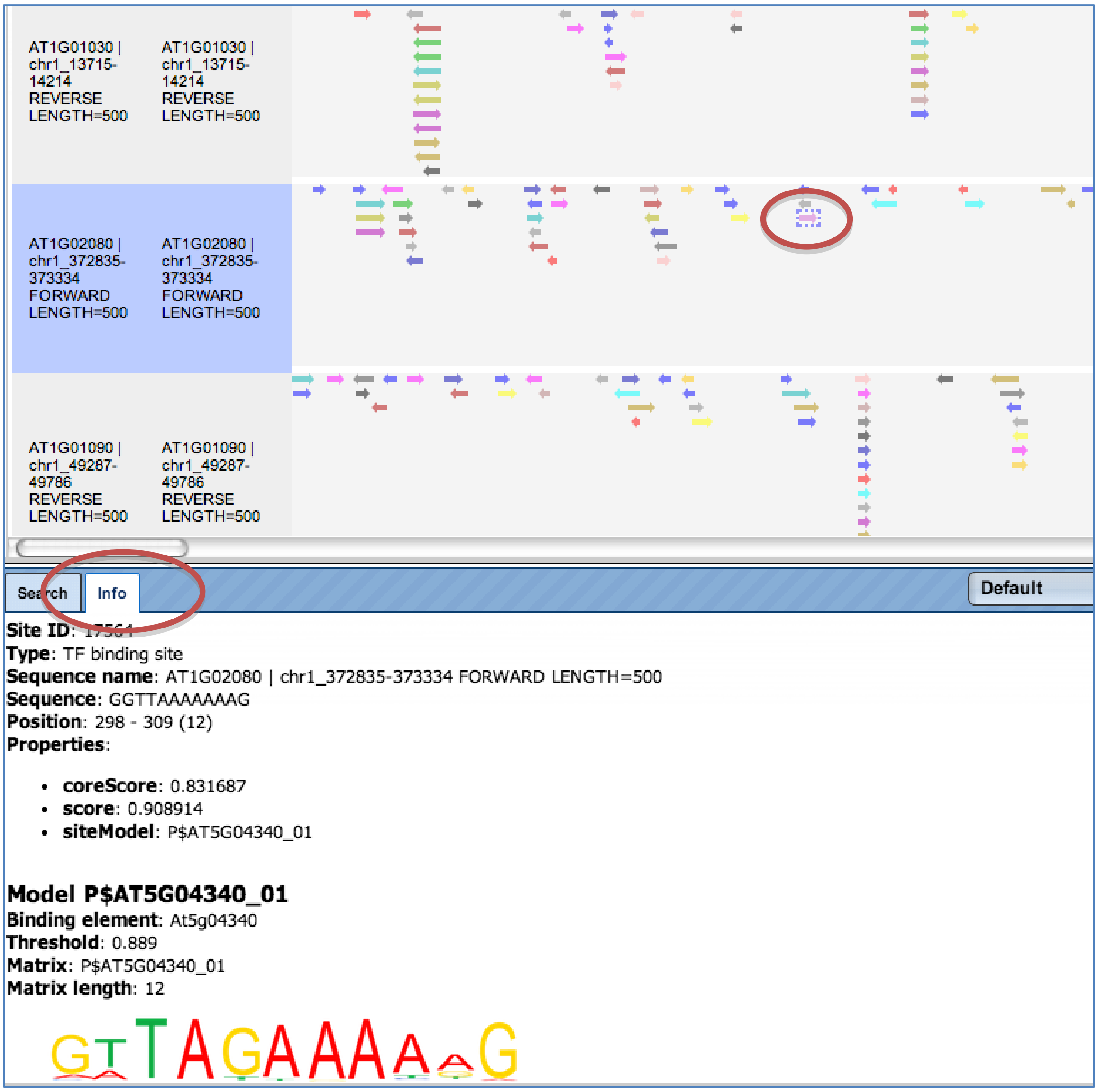
Show

ID	Sites density per 1000bp
P$TFIIAL_01	310
P$AT1G26590_01	167.8
P$WRI1_02	118.4
P$NAC6_01	113.6
P$WRI1_01	101
P$AT3G24520_01	92
P$AT1G50680_01	88.8
P$TINY2_01	84.4
P$EBP_01	77.4
P$AT1G51120_01	75.8
P$TEM1_01	73
P$AT1G67970_01	66
P$TNY_01	60

Each row summarizes the information for one site model (PWM, matrix).

For each row, the column **Site density per 1000bp** shows the number of matches normalized per 1000 bp length for the sequences in the input set.

TFBSs can be further visualized in the input sequences. For this, having the Summary table opened, select one or several rows, and then click the report on selected matrices button ( ) on the control panel. In this example, all matrices with a site density <5 were selected. The visualization results are shown below:



AT1G01030 | chr1_13715-14214 REVERSE LENGTH=500
AT1G01030 | chr1_13715-14214 REVERSE LENGTH=500
AT1G02080 | chr1_372835-373334 FORWARD LENGTH=500
AT1G02080 | chr1_372835-373334 FORWARD LENGTH=500
AT1G01090 | chr1_49287-49786 REVERSE LENGTH=500
AT1G01090 | chr1_49287-49786 REVERSE LENGTH=500
Search
Info
Default

Site ID: 17564
Type: TF binding site
Sequence name: AT1G02080 | chr1_372835-373334 FORWARD LENGTH=500
Sequence: GGTTAAAAAAAG
Position: 298 - 309 (12)
Properties:

- **coreScore**: 0.831687
- **score**: 0.908914
- **siteModel**: P$AT5G04340_01

Model P$AT5G04340_01
Binding element: At5g04340
Threshold: 0.889
Matrix: P$AT5G04340_01
Matrix length: 12

There are ten rows corresponding to the individual sequences in the input set. The column Sites view schematically represents the sequence length with mapped TFBSs. Matches for different matrices are shown in different colors. You can select individual matches by mouse click and get additional information in the Info box.

Note. This workflow is available together with a valid TRANSFAC® license. Please, feel free to ask for details (info@genexplain.com).

8.1.1.2. Search for TF binding sites with GTRD

This workflow is similar to the one described above. The difference is in the default profile applied for the TFBS search; in this workflow it is the profile from the GTRD database called *moderate threshold*. It can be found under the tab *Databases*, in the folder GTRD/Data/profiles/

Correspondingly, the site search results from these two workflows are different.

8.1.2. Analyze any DNA sequence for site enrichment

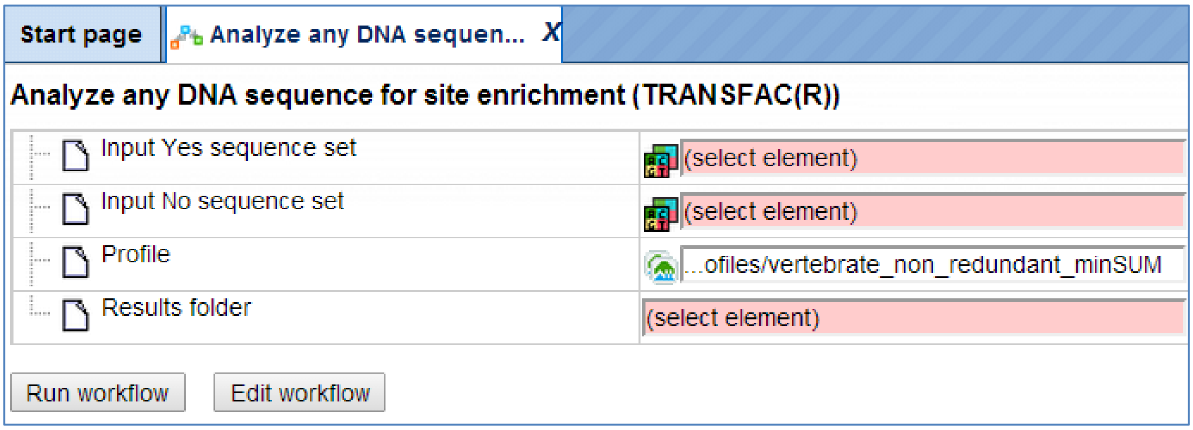
8.1.2.1. Analyze any DNA sequence for site enrichment with TRANSFAC®

This workflow is designed to search for enriched transcription factor binding sites, TFBSs, in any input DNA sequence as compared to a background DNA sequence. The central part of this workflow is performed by two individual methods, *Site search on track* ( ), and *Site search result optimization* ( ). Both individual methods can be found under the tab *Analyses* in the folder Methods/Site analysis/.

With this workflow you can analyze sequences of any species and any genomic region.

The few steps to launch the workflow are as follows.

Step 1. Open the workflow input form from the Start page, it will be opened in the main Work Space and looks as it is shown below:



Analyze any DNA sequence for site enrichment (TRANSFAC(R))

Input Yes sequence set	(select element)
Input No sequence set	(select element)
Profile	...ofiles/vertebrate_non_redundant_minSUM
Results folder	(select element)

Step 2. Specify the input Yes and No sequence sets. The Yes and No sequence sets can be in EMBL, FASTA or GenBank format. After loading into the platform, these sequences are shown in the tree area with the icon ( ).

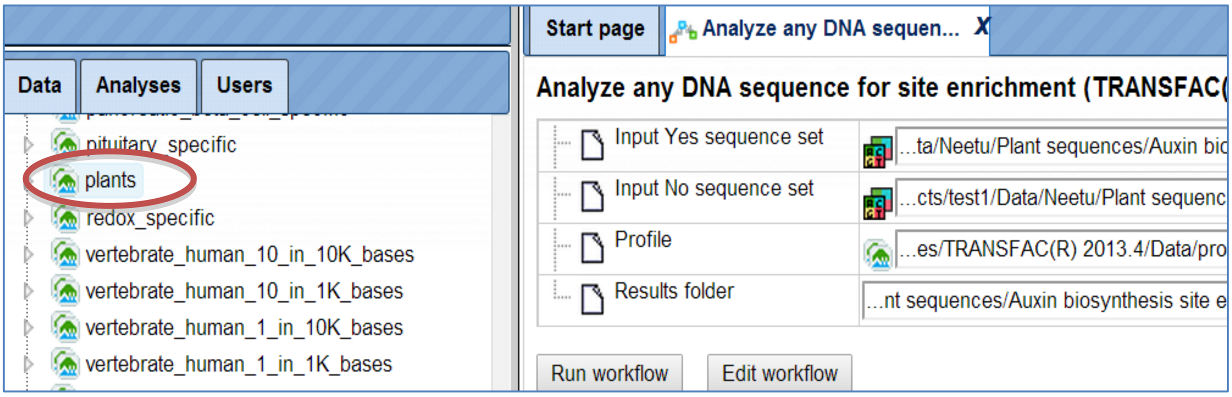
To specify the Input sequence sets, you can drag & drop it from your project within the tree area. Alternatively, you may click on the pink field "select element" and a new window will open, where you can select the input sequence.

In this example, the sample sequences of *Arabidopsis thaliana* are used, which were downloaded from TAIR database http://www.arabidopsis.org/tools/bulk/sequences/index.jsp.

Here, the Yes sequence collection contains a set of four promoter sequences 1000 bp upstream of TSS. These genes play an important role in auxin biosynthesis. The No sequence collection contains a set of four promoter sequences 1000 bp upstream of TSS. These genes are involved in different functions.

Step3. Select the TRANSFAC® profile from the available profiles. The default profile is *vertebrate_non_redundant_minSUM*. In this example we use the profile called *Plants*. It can be found here:

http://genexplain-platform.com/bioumlweb/#de=databases/TRANSFAC(R)%202014.4/Data/profiles/plants



pituitary_specific
plants
redox_specific
vertebrate_human_10_in_10K_bases
vertebrate_human_10_in_1K_bases
vertebrate_human_1_in_10K_bases
vertebrate_human_1_in_1K_bases

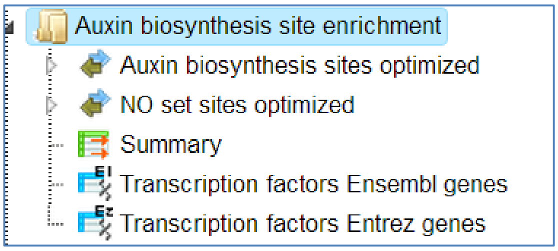
Analyze any DNA sequence for site enrichment (TRANSFAC(

Input Yes sequence set	...ta/Neetu/Plant sequences/Auxin bic
Input No sequence set	...cts/test1/Data/Neetu/Plant sequenc
Profile	...es/TRANSFAC(R) 2013.4/Data/pro
Results folder	...nt sequences/Auxin biosynthesis site e

Step 4. Specify the result folder location and name in your *Project* area. Then press the button [Run workflow]. Wait till the workflow is completed.

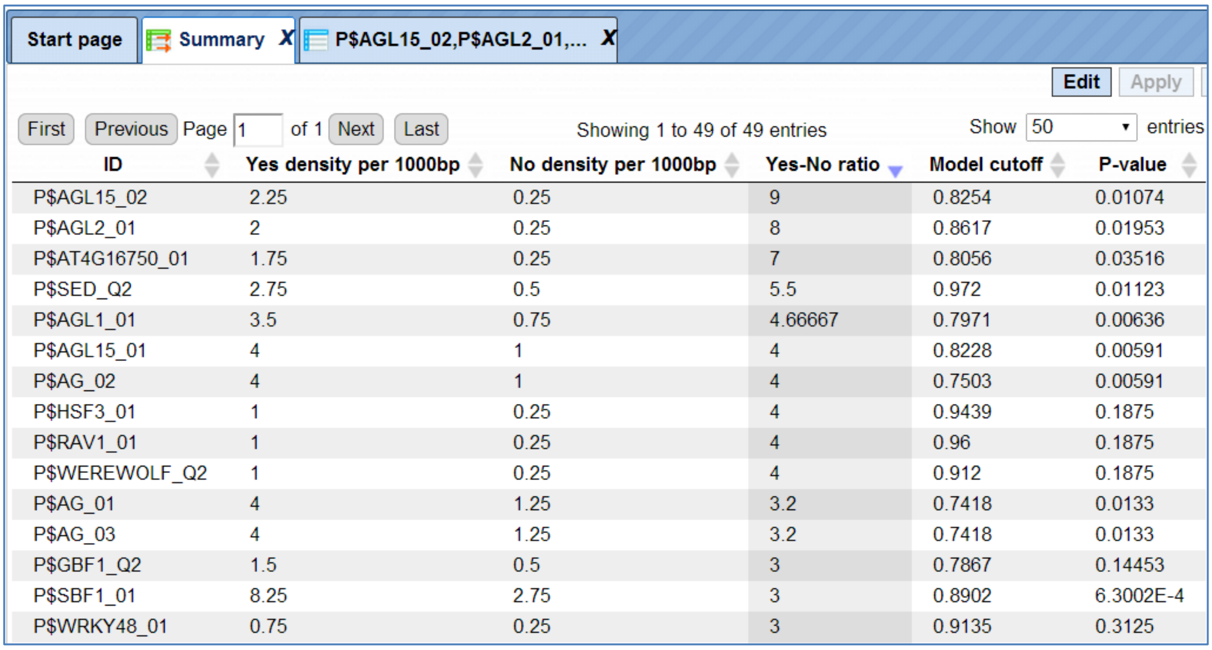
Results

The results folder consists of several tables and tracks as shown below:



Auxin biosynthesis site enrichment
- Auxin biosynthesis sites optimized
- NO set sites optimized
- Summary
- Transcription factors Ensembl genes
- Transcription factors Entrez genes

The table Summary ( ) gives the TFBSs enriched in the Yes set as compared with the No set. It looks as shown below:



Showing 1 to 49 of 49 entries

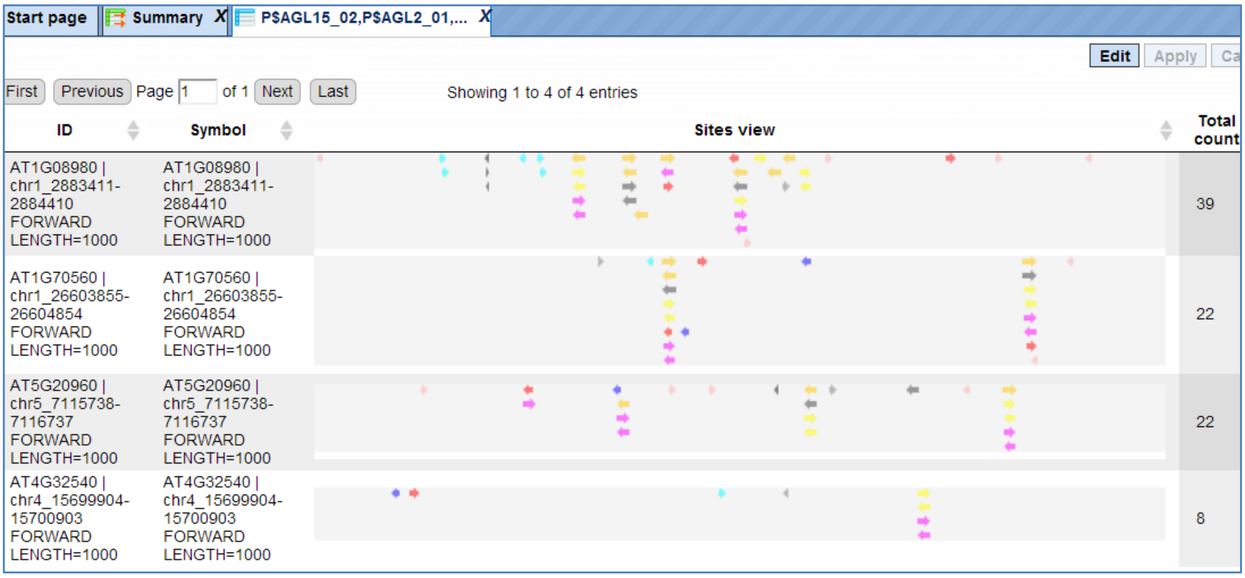
ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
P$AGL15_02	2.25	0.25	9	0.8254	0.01074
P$AGL2_01	2	0.25	8	0.8617	0.01953
P$AT4G16750_01	1.75	0.25	7	0.8056	0.03516
P$SED_Q2	2.75	0.5	5.5	0.972	0.01123
P$AGL1_01	3.5	0.75	4.66667	0.7971	0.00636
P$AGL15_01	4	1	4	0.8228	0.00591
P$AG_02	4	1	4	0.7503	0.00591
P$HSF3_01	1	0.25	4	0.9439	0.1875
P$RAV1_01	1	0.25	4	0.96	0.1875
P$WEREWOLF_Q2	1	0.25	4	0.912	0.1875
P$AG_01	4	1.25	3.2	0.7418	0.0133
P$AG_03	4	1.25	3.2	0.7418	0.0133
P$GBF1_Q2	1.5	0.5	3	0.7867	0.14453
P$SBF1_01	8.25	2.75	3	0.8902	6.3002E-4
P$WRKY48_01	0.75	0.25	3	0.9135	0.3125

Each row summarizes the information for one site model (PWM, matrix).

For each row, the columns **Yes density per 1000bp** and **No density per 1000bp** show the number of matches normalized per 1000 bp length for the sequences in the input Yes set and input No set, respectively. The Column **Yes-No ratio** is the ratio of the first two columns. The higher the Yes-No ratio, the higher is the enrichment of matches for the respective matrix in the Yes set. The matrix cutoff values as they are calculated by the program at the optimization step are shown in the column **Model cutoff**, and the last column shows the p-value of the corresponding event.

TFBSs can be further visualized in the Yes sequences. For this, having the Summary table opened, select one or several rows, and then click the report on selected matrices button () on the control panel.

In this example, all matrices having a Yes-No ratio>3 were selected. The visualization results are shown below:

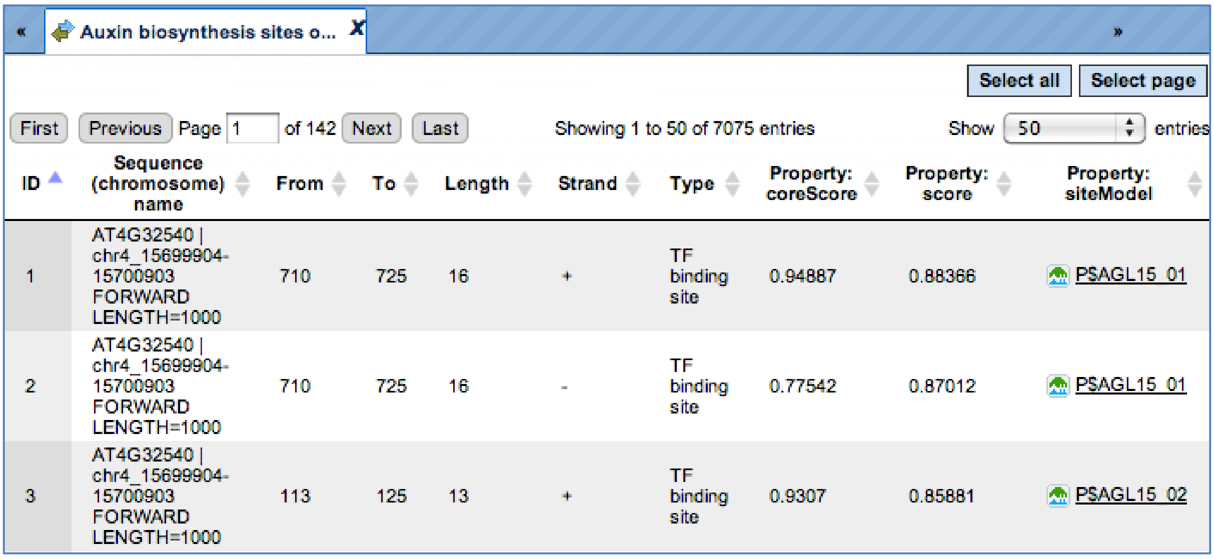


Showing 1 to 4 of 4 entries

ID	Symbol	Sites view	Total count
AT1G08980 | chr1_2883411-2884410 FORWARD LENGTH=1000	AT1G08980 | chr1_2883411-2884410 FORWARD LENGTH=1000		39
AT1G70560 | chr1_26603855-26604854 FORWARD LENGTH=1000	AT1G70560 | chr1_26603855-26604854 FORWARD LENGTH=1000		22
AT5G20960 | chr5_7115738-7116737 FORWARD LENGTH=1000	AT5G20960 | chr5_7115738-7116737 FORWARD LENGTH=1000		22
AT4G32540 | chr4_15699904-15700903 FORWARD LENGTH=1000	AT4G32540 | chr4_15699904-15700903 FORWARD LENGTH=1000		8

There are four rows corresponding to the individual sequences in the input Yes set. The column Sites view schematically represents the sequence length with mapped TFBSs. Matches for different matrices are shown in different colors. You can select individual matches by mouse click and get additional information in the Info box.

Yes (No) sites optimized tracks. These optimized tracks ( ) present those TFBSs that are over-represented in the Yes sequences versus the No sequences. Scores of the putative sites are optimized by the algorithm. As the Yes set in this example is called *Auxin biosynthesis*, the resulting track is called *Auxin biosynthesis sites optimized.* When opened as a table, this track looks like:



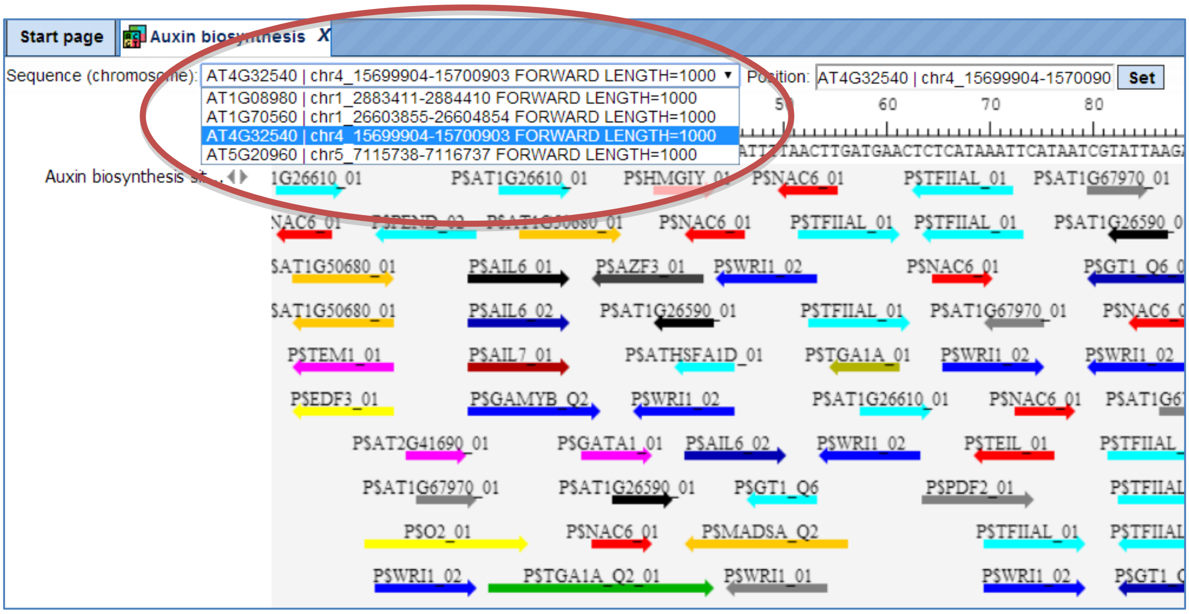
Showing 1 to 50 of 7075 entries

Show 50 entries

ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: coreScore	Property: score	Property: siteModel
1	AT4G32540 | chr4_15699904-15700903 FORWARD LENGTH=1000	710	725	16	+	TF binding site	0.94887	0.88366	PSAGL15_01
2	AT4G32540 | chr4_15699904-15700903 FORWARD LENGTH=1000	710	725	16	-	TF binding site	0.77542	0.87012	PSAGL15_01
3	AT4G32540 | chr4_15699904-15700903 FORWARD LENGTH=1000	113	125	13	+	TF binding site	0.9307	0.85881	PSAGL15_02

Each row corresponds to one resulting TFBS, and includes its position in the Yes sequences (the columns **From** and **To**), length and strand, as well as a score calculated by the algorithm and a site model (matrix). This table can be exported as a track, in several different formats including intervals, bed, wig and more. DNA sequences can be exported in multi-FASTA format.

The same track, when opened in the genome browser, looks as shown below:



Sequence (chromosome): AT4G32540 | chr4_15699904-15700903 FORWARD LENGTH=1000

AT1G08980 | chr1_2883411-2884410 FORWARD LENGTH=1000
AT1G70560 | chr1_26603855-26604854 FORWARD LENGTH=1000
AT4G32540 | chr4_15699904-15700903 FORWARD LENGTH=1000
AT5G20960 | chr5_7115738-7116737 FORWARD LENGTH=1000

Position: AT4G32540 | chr4_15699904-1570090

In the field *Sequence (chromosome)* you can find a drop down menu, highlighted by the red oval. This feature helps to easily switch visualization between the sequences in the input set. In this particular example the Yes sequence set comprises four individual promoter sequences, and each individual promoter can be visualized in the genome browser.

The tables *Transcription factors Ensembl* ( ) and *Transcription factors Entrez* ( ) aim at showing transcription factors linked to the identified site models (matrices). These are potential candidate regulators of genes in the input Yes set. They are supposed to regulate transcription of Yes-genes via the identified enriched TFBSs.

Currently this feature is available for human, mouse and rat. For all other species, this table will be empty, as in this example for *Arabidopsis* sequences.

Note. This workflow is available together with a valid TRANSFAC® license.
Please, feel free to ask for details (info@genexplain.com).

8.1.2.2. Analyze any DNA sequence for site enrichment (GTRD)

This workflow is similar to the one described above. The difference is in the default profile applied for the TFBS search; in this workflow it is the profile from the GTRD database called *moderate threshold*. It can be found under the tab *Databases*, in the folder GTRD/Data/profiles/

Correspondingly, the enriched motifs resulting from these two workflows are different.

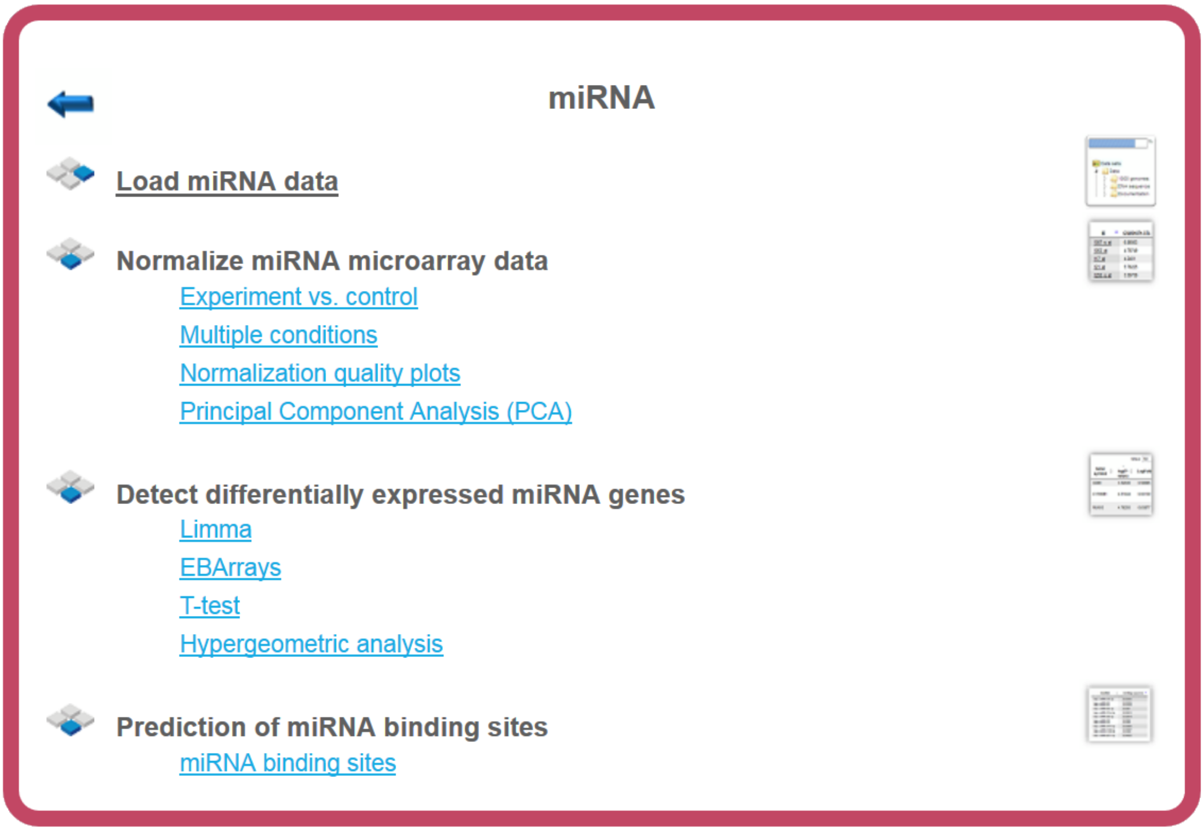
8.2. Further workflows in this area

For the other workflows that you can find in the area ***Sequence analysis***, please refer to the following Sections:

Load data	See Chapter 3
Discover de-novo motifs using ChIPHorde and DiChIPHorde	See Section 7.2.6

9. miRNA

When you open this Area, a list of workflows optimized for working with microRNA data will show up, which looks as follows:



miRNA

Load miRNA data

Normalize miRNA microarray data

- Experiment vs. control
- Multiple conditions
- Normalization quality plots
- Principal Component Analysis (PCA)

Detect differentially expressed miRNA genes

- Limma
- EBArrays
- T-test
- Hypergeometric analysis

Prediction of miRNA binding sites

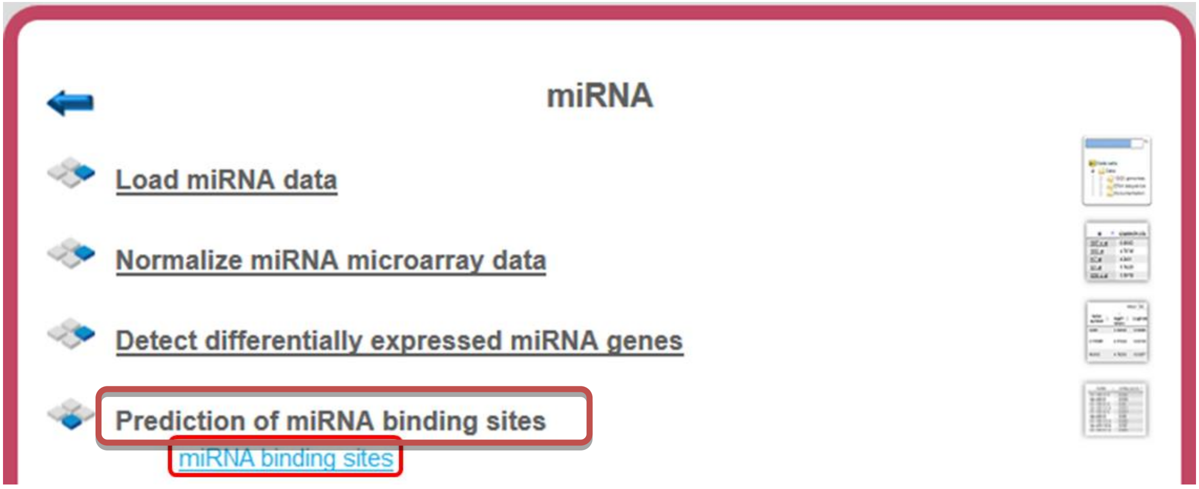
- miRNA binding sites

For detailed explanation of their function and how to operate them, please refer to the following chapters:

Load gene or protein list	See Chapter 3
Detect differentially expressed miRNA genes	See Chapter 10.2

9.1. Prediction of miRNA binding sites

MicroRNAs, or miRNAs, post-transcriptionally affect (mostly: repress) the expression of protein-coding genes. The human genome encodes over 1000 miRNA genes that collectively target the vast majority of messenger RNAs (mRNAs). This workflow can help to predict miRNA binding sites. Starting from a gene list, first a collection of 3' untranslated regions (3' UTRs) is created. This collection is used to map against a miRmap library, which derived from the mirBase database (release 20; http://www.mirbase.org). The last step of the workflow is based on the miRmap method (*analyses/Galaxy/microRNA/mirmap*; http://mirmap.ezlab.org), published by Vejnar & Zdobnov, Nucleic Acids Res. 40:11673-11683, 2012. The result of the workflow is a table with all predicted miRNA binding sites and a track for visualization in the created sequence collection. The workflow can be found on the Start page via the miRNA button.



miRNA

Load miRNA data

Normalize miRNA microarray data

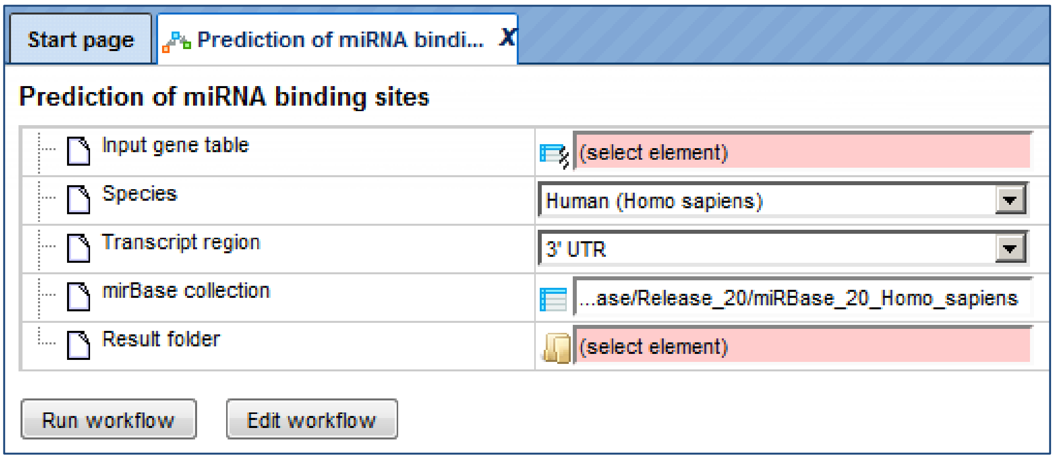
Detect differentially expressed miRNA genes

Prediction of miRNA binding sites

miRNA binding sites

To launch the workflow, follow these steps:

Step 1. Open the workflow input form from the Start page. It looks as shown below:



Prediction of miRNA binding sites

Input gene table	(select element)
Species	Human (Homo sapiens)
Transcript region	3' UTR
mirBase collection	...ase/Release_20/miRBase_20_Homo_sapiens
Result folder	(select element)

Step 2. Specify the input gene table. The input gene sets might be lists of differentially regulated genes or any gene or protein list of interest. You can drag them from your project within the tree area and drop them in the pink box of the field **Input table**. Alternatively, you may click on the pink field "select element" and a new window will be opened, where you can select the input gene set as shown below.

The further steps of the workflow are demonstrated on the example file, which can be found in the geneXplain platform under:

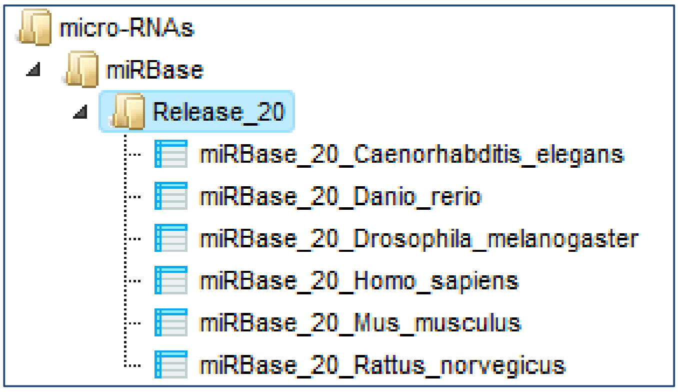
data/Examples/miRNA binding site prediction (miRNA-155 target genes)/Data/hsa-miR-155-5p published target genes

This table contains four genes that are already published targets for one particular miRNA, hsa-miR-155-5p, for which several target genes have been published as well; we have derived this information from the TRANSFAC® database.

Step 3. Specify the biological species of the input set in the field **Species** by selecting it from the drop-down menu.

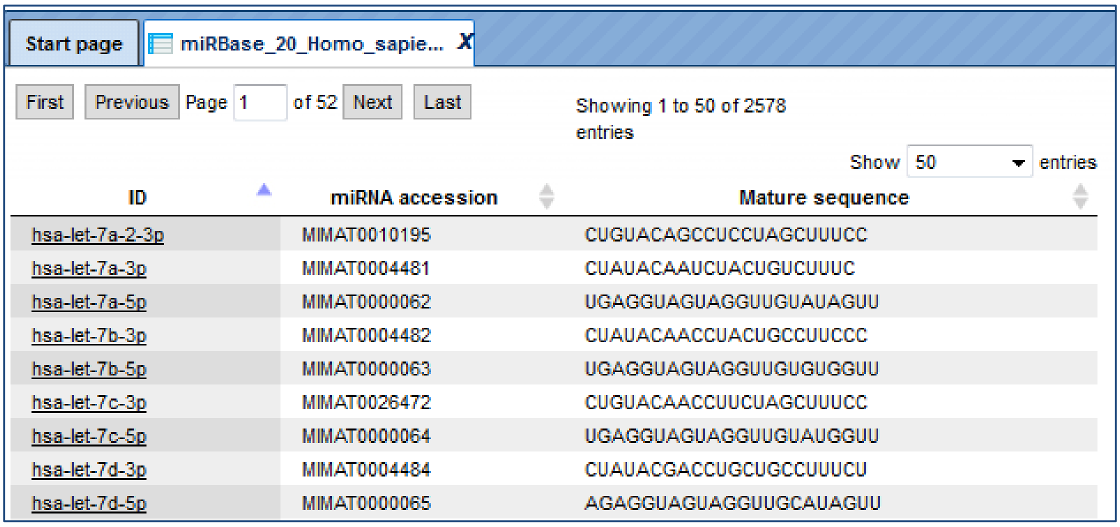
Step 4. In the field **Transcript region**, choose the gene region from the drop-down menu. Here, the 3' UTR is selected as default with a fixed length of 300bp. The CDS information is ignored.

Step 5. Selection of **miRBASE collection**. By clicking in the this field, six different miRBase collections from release 20 are available for *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Drosophila melanogaster*, *Danio rerio* and *Caenorhabditis elegans*.



micro-RNAs
miRBase
Release_20
miRBase_20_Caenorhabditis_elegans
miRBase_20_Danio_rerio
miRBase_20_Drosophila_melanogaster
miRBase_20_Homo_sapiens
miRBase_20_Mus_musculus
miRBase_20_Rattus_norvegicus

The miRBase library can be opened as a table as shown below.



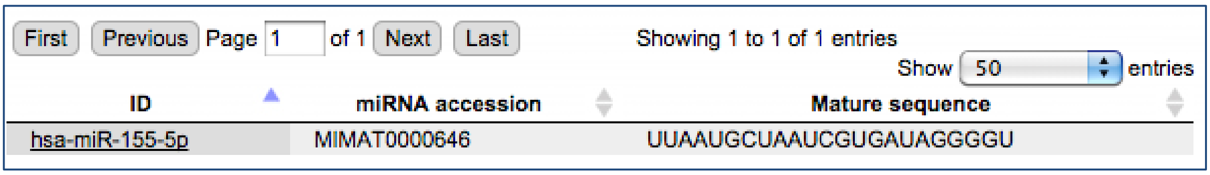
Showing 1 to 50 of 2578 entries

ID	miRNA accession	Mature sequence
hsa-let-7a-2-3p	MIMAT0010195	CUGUACAGCCUCCUAGCUUUCC
hsa-let-7a-3p	MIMAT0004481	CUAUACAAUCUACUGUCUUUC
hsa-let-7a-5p	MIMAT0000062	UGAGGUAGUAGGUUGUAUAGUU
hsa-let-7b-3p	MIMAT0004482	CUAUACAACCUACUGCCUUCCC
hsa-let-7b-5p	MIMAT0000063	UGAGGUAGUAGGUUGUGUGGUU
hsa-let-7c-3p	MIMAT0026472	CUGUACAACCUUCUAGCUUUCC
hsa-let-7c-5p	MIMAT0000064	UGAGGUAGUAGGUUGUAUGGUU
hsa-let-7d-3p	MIMAT0004484	CUAUACGACCUGCUGCCUUUCU
hsa-let-7d-5p	MIMAT0000065	AGAGGUAGUAGGUUGCAUAGUU

The miRBase library table has three columns for each miRNA. Identifiers in the column ID are hyperlinked to miRBase (http://www.mirbase.org/). Two other columns are the miRNA accession numbers and the mature sequence, which is used for binding site identification.

You can use the whole library or you can create a subset of several or even just one miRNA of interest to focus on predicting binding sites for this particular miRNA.

In this example, let's cerate a subset containing just one miRNA, hsa-miR-155-5p, for which several target genes have been already published. To create a subset from the table, apply Filter tab in the operations field. The resulting customized miRBase contains just one row.



First | Previous | Page 1 of 1 | Next | Last

Showing 1 to 1 of 1 entries

Show 50 entries

ID	miRNA accession	Mature sequence
hsa-miR-155-5p	MIMAT0000646	UUAAUGCUAAUCGUGAUAGGGGU

Step 6. Define where the folder with the results should be located in the tree. You can do so by clicking on the pink field *select element* in the field **Result folder**, and a new window will be opened, where you can select the location of the result folder and define its name.

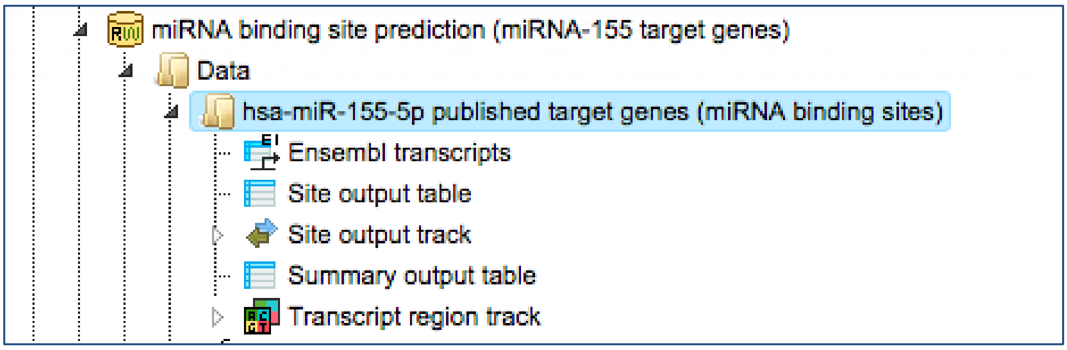
Step 7. Press the [Run workflow] button.

When the workflow is completed, the result folder is opened by default.

Let’s consider the results.

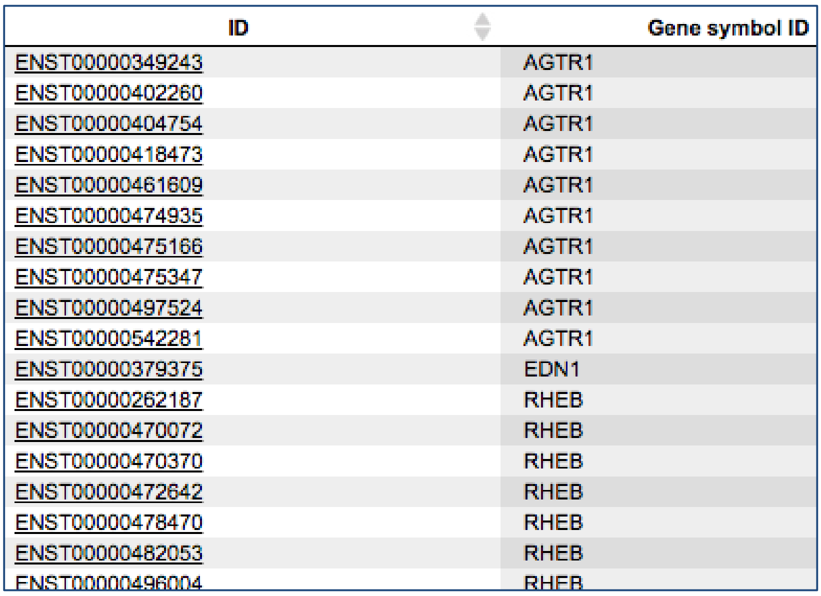
data/Examples/miRNA binding site prediction (miRNA-155 target genes)/Data/hsa-miR-155-5p published target genes (miRNA binding sites)

The result folder contains the following tables: Ensembl transcript ( ),Site output ( ) with all predicted miRNA binding sites, Site output track ( ) for visualization of predicted miRNA sites, Summary output ( ) and Transcript region track .



- miRNA binding site prediction (miRNA-155 target genes)
  - Data
    - hsa-miR-155-5p published target genes (miRNA binding sites)
      - Ensembl transcripts
      - Site output table
      - Site output track
      - Summary output table
      - Transcript region track

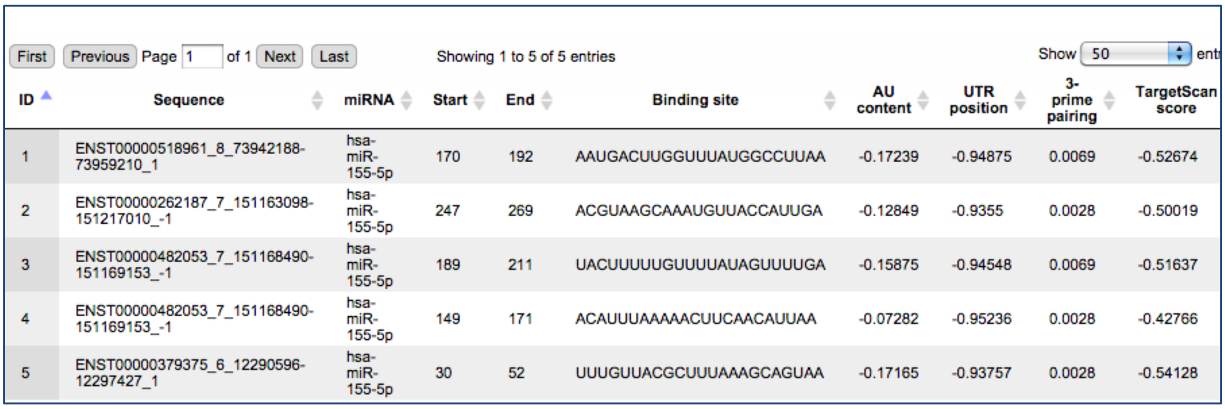
The table **Ensembl transcripts** corresponds to the input gene table converted to the Ensembl transcript IDs (first column **ID**). The other columns of this table are the same as they are in the input table. In this example, 4 Ensembl genes are converted into 26 Ensembl transcripts.



ID	Gene symbol ID
ENST00000349243	AGTR1
ENST00000402260	AGTR1
ENST00000404754	AGTR1
ENST00000418473	AGTR1
ENST00000461609	AGTR1
ENST00000474935	AGTR1
ENST00000475166	AGTR1
ENST00000475347	AGTR1
ENST00000497524	AGTR1
ENST00000542281	AGTR1
ENST00000379375	EDN1
ENST00000262187	RHEB
ENST00000470072	RHEB
ENST00000470370	RHEB
ENST00000472642	RHEB
ENST00000478470	RHEB
ENST00000482053	RHEB
ENST00000496004	RHEB

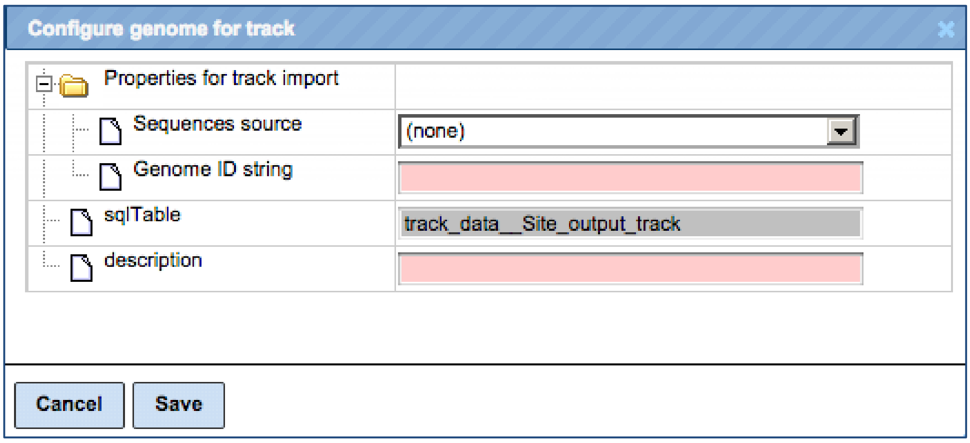
These Ensembl transcript IDs are needed to construct the **Transcript region track, in this workflow especially containing** 3′ UTRs of genes given in the input table. The workflow takes 300 bp regions as 3′ UTRs.

The **Site output table** contains all identified miRNA binding sites; each row in this table corresponds to one identified miRNA binding site, and each site/row has an ID assigned. The column **Sequence** corresponds to the 300 bp 3′ UTR of the indicated Ensembl transcript, where the search for miRNA sites has been done. The column **miRNA** contains the name of **miRNA** binding to this site, in this example hsa-miR-155-5p. The columns **Start** and **End** are positions within the input sequence and the **column Binding site contains the sequence of the site**.



ID	Sequence	miRNA	Start	End	Binding site	AU content	UTR position	3-prime pairing	TargetScan score
1	ENST00000518961_8_73942188-73959210_1	hsa-miR-155-5p	170	192	AAUGACUUGGUUUAUGGCCUUAA	-0.17239	-0.94875	0.0069	-0.52674
2	ENST00000262187_7_151163098-151217010_-1	hsa-miR-155-5p	247	269	ACGUAAGCAAAUGUUACCAUUGA	-0.12849	-0.9355	0.0028	-0.50019
3	ENST00000482053_7_151168490-151169153_-1	hsa-miR-155-5p	189	211	UACUUUUUGUUUUAUAGUUUUGA	-0.15875	-0.94548	0.0069	-0.51637
4	ENST00000482053_7_151168490-151169153_-1	hsa-miR-155-5p	149	171	ACAUUUAAAAACUUCAACAUUAA	-0.07282	-0.95236	0.0028	-0.42766
5	ENST00000379375_6_12290596-12297427_1	hsa-miR-155-5p	30	52	UUUGUUACGCUUUAAAGCAGUAA	-0.17165	-0.93757	0.0028	-0.54128

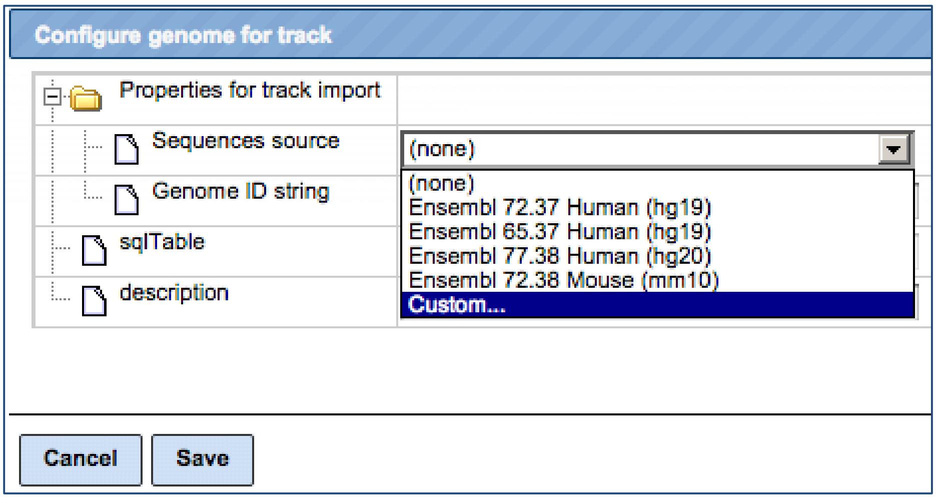
The **Site output track** enables the visualization of the predicted sites in the 3′ UTRs of the input gene list. With a double-click on (Site output track) in the results folder, the dialog box *Configure genome for track* is opened.



Configure genome for track

Properties for track import	
Sequences source	(none)
Genome ID string	
sqlTable	track_data__Site_output_track
description	

Within this dialog box, first set the field **Sequence source** to *Custom...*

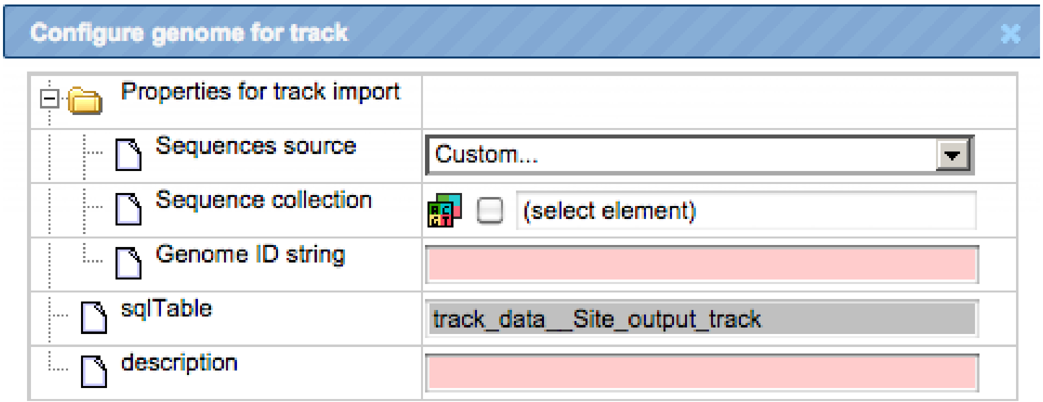


Configure genome for track

Properties for track import	
Sequences source	(none)
Genome ID string	
sqlTable	
description	

(none)
Ensembl 72.37 Human (hg19)
Ensembl 65.37 Human (hg19)
Ensembl 77.38 Human (hg20)
Ensembl 72.38 Mouse (mm10)
Custom...
Cancel
Save

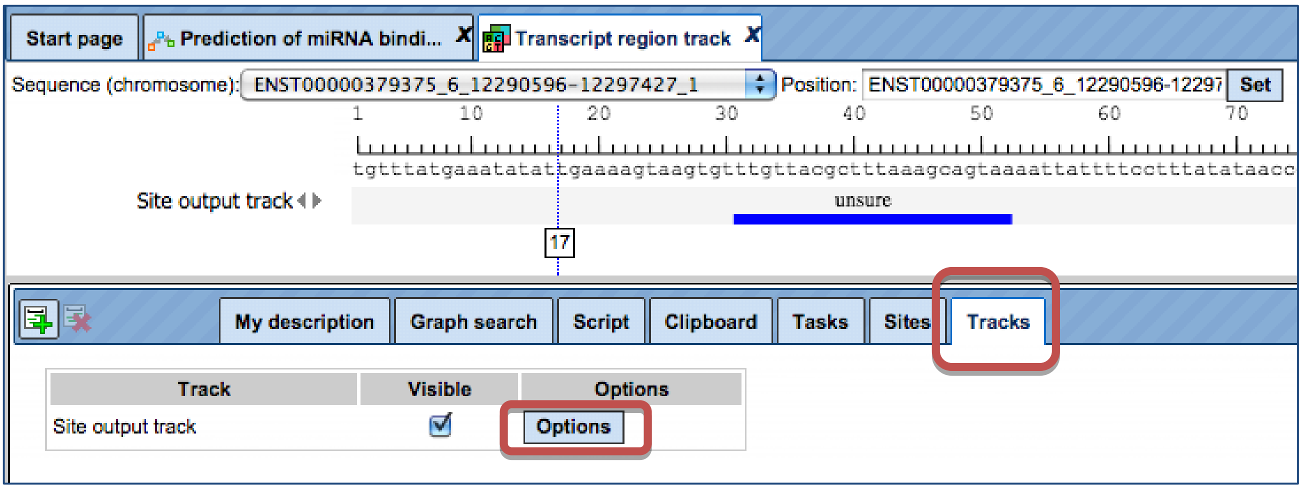
and [Save] it. The dialog box is automatically adjusted, and the field *Sequence collection* becomes available.



Configure genome for track

Properties for track import	
Sequences source	Custom...
Sequence collection	□ (select element)
Genome ID string	
sqlTable	track_data__Site_output_track
description	

You can drag and drop the sequence collection Transcript region track from the workflow results folder and [Save] it. The visualization is opened automatically in the work area.



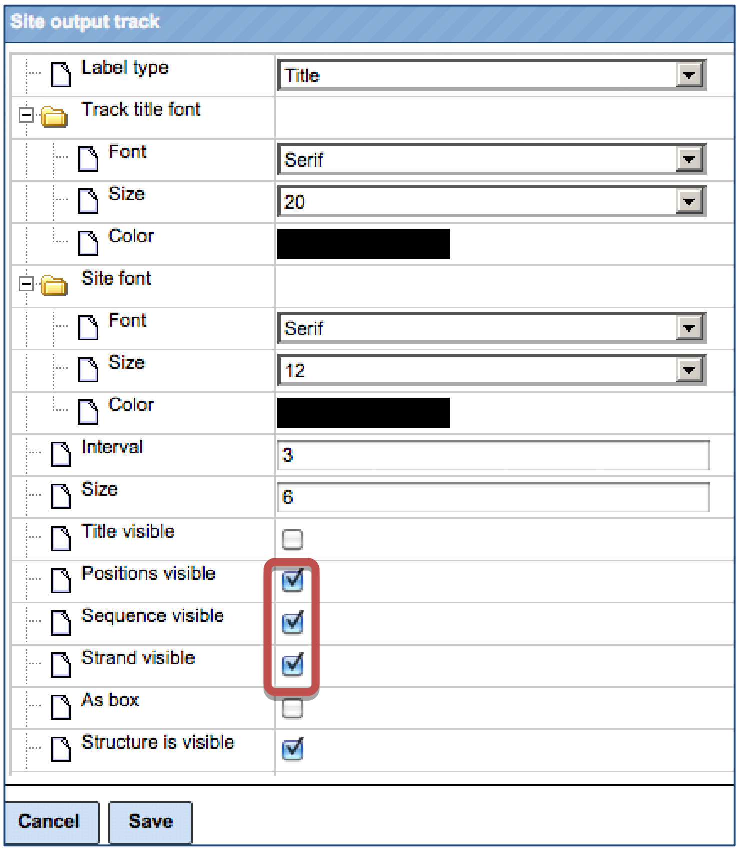
Sequence (chromosome): ENST00000379375_6_12290596-12297427_1 Position: ENST00000379375_6_12290596-12297 Set

tgtttatgaaatatattgaaaagtaagtgtttgttacgctttaaagcagtaaaattattttcctttatataacc

Site output track

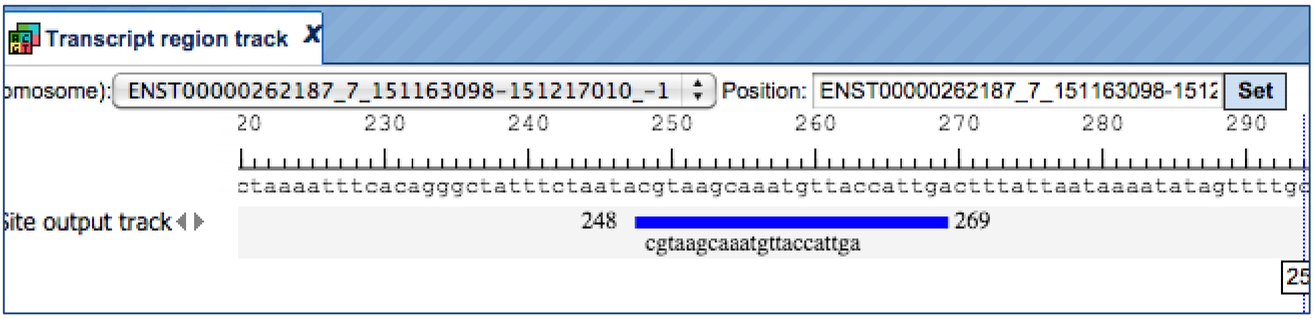
Track	Visible	Options
Site output track	☑	Options

You can modify visualization settings by opening the tab *Tracks* in the *Operations Field,* highlighted by the red boxes on the screenshot above, and press the button [Options]. The dialog box *Site output track* will be opened. You can modify different settings according to your preferences and [Save] them. Here, the settings in the highlighted check-boxes were changed as shown below.


Site output track

Label type	Title
Track title font	
Font	Serif
Size	20
Color	■
Site font	
Font	Serif
Size	12
Color	■
Interval	3
Size	6
Title visible	☐
Positions visible	☑
Sequence visible	☑
Strand visible	☑
As box	☐
Structure is visible	☑

These changes resulted in the following visualization.

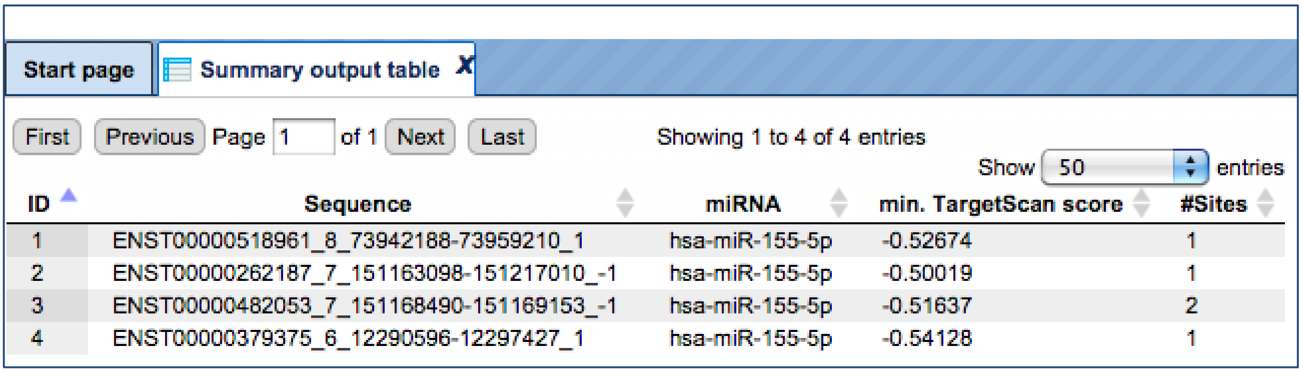


Transcript region track

omosome): ENST00000262187_7_151163098-151217010_-1 Position: ENST00000262187_7_151163098-1512 Set

20
230
240
250
260
270
280
290
ctaaaatttcacagggctatttctaatacgtaagcaaatgttaccattgactttattaataaaatatagttttgc
ite output track
248
269
cgtaagcaaatgttaccattga
25

The **Summary output table** presents those transcripts where miRNA binding sites are identified, in this example four transcripts.



Start page | Summary output table

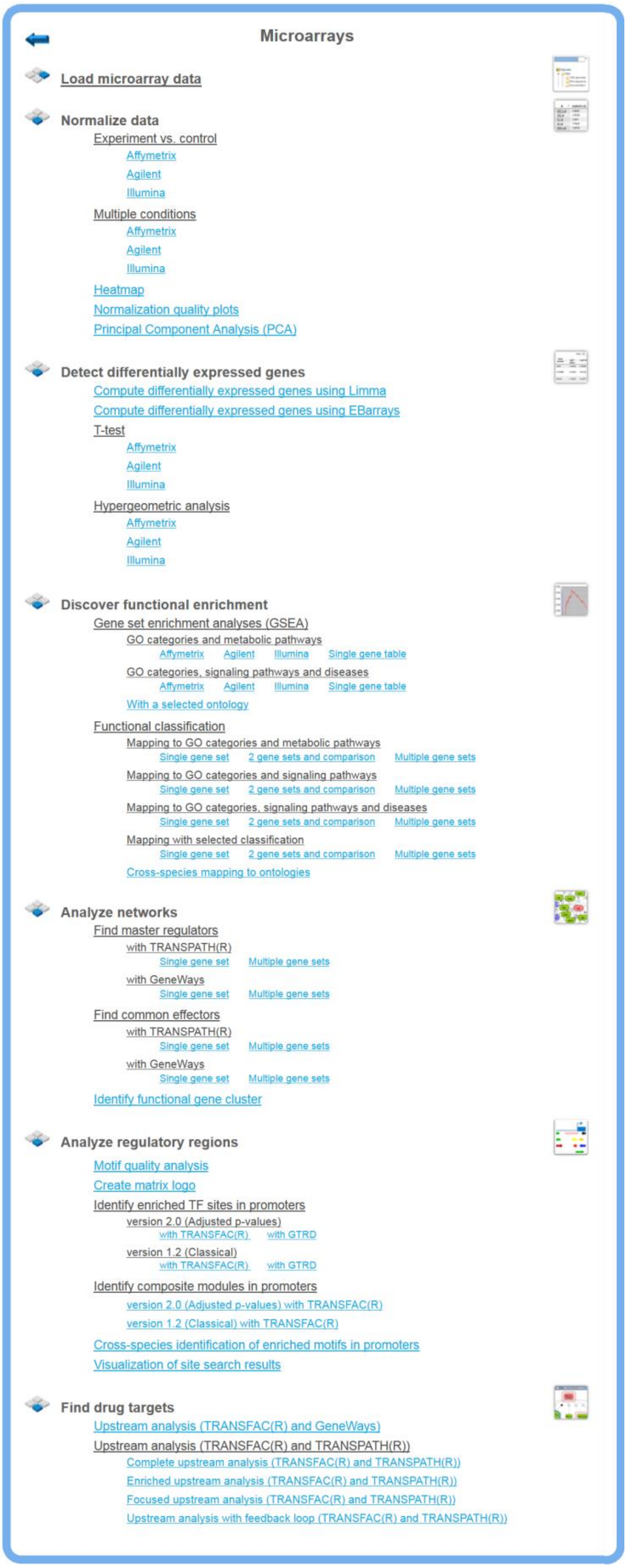
First Previous Page 1 of 1 Next Last Showing 1 to 4 of 4 entries

Show 50 entries

ID	Sequence	miRNA	min. TargetScan score	#Sites
1	ENST00000518961_8_73942188-73959210_1	hsa-miR-155-5p	-0.52674	1
2	ENST00000262187_7_151163098-151217010_-1	hsa-miR-155-5p	-0.50019	1
3	ENST00000482053_7_151168490-151169153_-1	hsa-miR-155-5p	-0.51637	2
4	ENST00000379375_6_12290596-12297427_1	hsa-miR-155-5p	-0.54128	1

Each row corresponds to one transcript. The column **Sequence** contains the titles of the Ensembl transcripts, and the column **miRNA** presents name of the miRNA. The number of binding sites for a particular miRNA in each transcript is shown in the column **#Sites**.

10. Microarrays



Microarrays

- Load microarray data
- Normalize data
  - Experiment vs. control
    - Affymetrix
    - Agilent
    - Illumina
  - Multiple conditions
    - Affymetrix
    - Agilent
    - Illumina
  - Heatmap
  - Normalization quality plots
  - Principal Component Analysis (PCA)
- Detect differentially expressed genes
  - Compute differentially expressed genes using Limma
  - Compute differentially expressed genes using EBarrays
  - T-test
    - Affymetrix
    - Agilent
    - Illumina
  - Hypergeometric analysis
    - Affymetrix
    - Agilent
    - Illumina
- Discover functional enrichment
  - Gene set enrichment analyses (GSEA)
    - GO categories and metabolic pathways
      - Affymetrix Agilent Illumina Single gene table
    - GO categories, signaling pathways and diseases
      - Affymetrix Agilent Illumina Single gene table
    - With a selected ontology
  - Functional classification
    - Mapping to GO categories and metabolic pathways
      - Single gene set 2 gene sets and comparison Multiple gene sets
    - Mapping to GO categories and signaling pathways
      - Single gene set 2 gene sets and comparison Multiple gene sets
    - Mapping to GO categories, signaling pathways and diseases
      - Single gene set 2 gene sets and comparison Multiple gene sets
    - Mapping with selected classification
      - Single gene set 2 gene sets and comparison Multiple gene sets
    - Cross-species mapping to ontologies
- Analyze networks
  - Find master regulators
    - with TRANSPATH(R)
      - Single gene set Multiple gene sets
    - with GeneWays
      - Single gene set Multiple gene sets
  - Find common effectors
    - with TRANSPATH(R)
      - Single gene set Multiple gene sets
    - with GeneWays
      - Single gene set Multiple gene sets
  - Identify functional gene cluster
- Analyze regulatory regions
  - Motif quality analysis
  - Create matrix logo
  - Identify enriched TF sites in promoters
    - version 2.0 (Adjusted p-values)
      - with TRANSFAC(R) with GTRD
    - version 1.2 (Classical)
      - with TRANSFAC(R) with GTRD
  - Identify composite modules in promoters
    - version 2.0 (Adjusted p-values) with TRANSFAC(R)
    - version 1.2 (Classical) with TRANSFAC(R)
  - Cross-species identification of enriched motifs in promoters
  - Visualization of site search results
- Find drug targets
  - Upstream analysis (TRANSFAC(R) and GeneWays)
  - Upstream analysis (TRANSFAC(R) and TRANSPATH(R))
    - Complete upstream analysis (TRANSFAC(R) and TRANSPATH(R))
    - Enriched upstream analysis (TRANSFAC(R) and TRANSPATH(R))
    - Focused upstream analysis (TRANSFAC(R) and TRANSPATH(R))
    - Upstream analysis with feedback loop (TRANSFAC(R) and TRANSPATH(R))

10.1. Normalize data

10.1.1. Experiment vs. control or multiple conditions

If your expression data haven't been normalized yet, as we assume in this example, you have to go now to the second group of options, "Normalize data". Make your choice according to the experimental platform you have used (Affymetrix, Agilent or Illumina). The next form will ask you to define the data file(s). For this, you have two options:

When you click in either field, a window with the title "Select data element" opens allowing you to select a file or, more likely at this step, a number of files by mouse click or by typing their names. Selection by mouse click works as usual for a range of files (keep the Shift key pressed when selecting the last file of the range), or for a number of distinct files (keep the Ctrl key pressed when selecting the second and further files from the list).

Make sure that all hybridizations (i.e., all CEL files) of your experiment are included into one normalization procedure. It should comprise all CEL files, including all multiple repetitions, of all conditions to be compared with each other at a later step, i.e. all tests and controls, at least those that you want to compare later on.

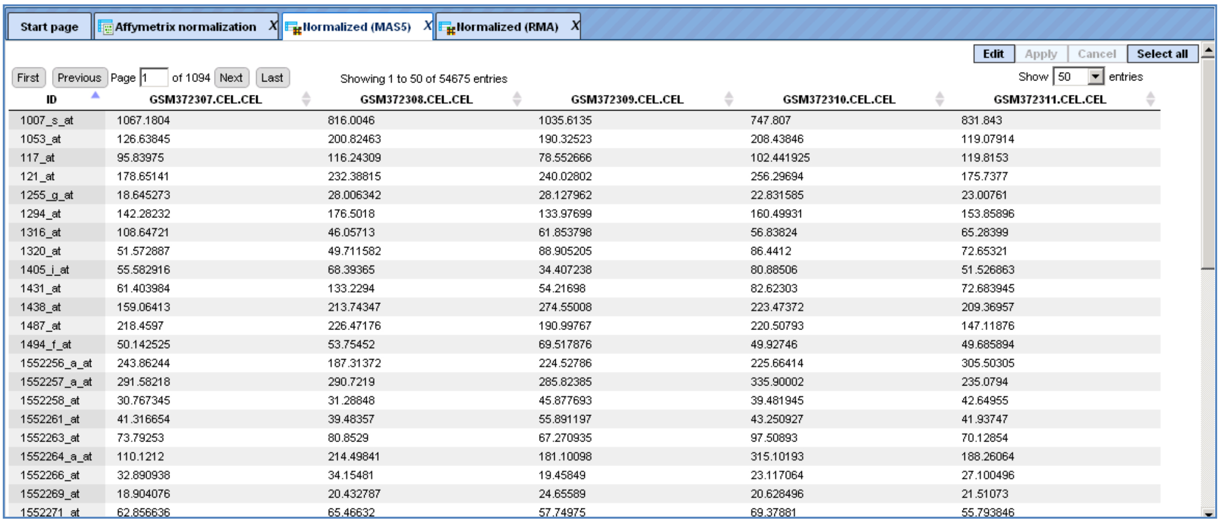
If your readings were from a dual-channel experiment, please tick the checkbox (Agilent only).

In the field Output name, you find a suggested name for the output file. The default name is "Normalized (<*Normalization_method*>)", which you can edit (just click into this field and change the default name). An accordingly named file with the icon ▦ will appear in the Tree Area after the procedure run successfully.

You may also have noticed that some further information about the analyses to be employed is displayed in both the Info Box as well as in the Operations Field ("My description" tab). Sometimes, they are identical by default, but in the latter field, you can edit the contents and add your own comments.

Now launch the normalization routine by pressing the [Run] button. The program will now normalize all your data across all experiments done, i.e. through all CEL files selected. The results will be stored in two different tables, one with the experimental, the other with the control values.

To have a closer look into the full content of one of the results tables, just click with the right mouse button onto the respective file name in the Tree Area, and choose "Open table" from the little menu that appears, or double-click on the file name. The table will open under a new tab in the Work Space. It should look like this, with the probeset IDs in the first and the normalized expression values from the different CEL files in the following columns, each hybridization being represented in one column (picture below).



Showing 1 to 50 of 54675 entries

ID	GSM372307.CEL.CEL	GSM372308.CEL.CEL	GSM372309.CEL.CEL	GSM372310.CEL.CEL	GSM372311.CEL.CEL
1007_s_at	1067.1804	816.0046	1035.6135	747.807	831.843
1053_at	126.63845	200.82463	190.32523	208.43846	119.07914
117_at	95.83975	116.24309	78.552666	102.441925	119.8153
121_at	178.65141	232.38815	240.02802	256.29694	175.7377
1255_g_at	18.645273	28.006342	28.127962	22.831585	23.00761
1294_at	142.28232	176.5018	133.97699	160.49931	153.85896
1316_at	108.64721	46.05713	61.853798	56.83824	65.28399
1320_at	51.572887	49.711582	88.905205	86.4412	72.65321
1405_i_at	55.582916	68.39365	34.407238	80.88506	51.526863
1431_at	61.403984	133.2294	54.21698	82.62303	72.683945
1438_at	159.06413	213.74347	274.55008	223.47372	209.36957
1487_at	218.4597	226.47176	190.99767	220.50793	147.11876
1494_f_at	50.142525	53.75452	69.517876	49.92746	49.685894
1552256_a_at	243.86244	187.31372	224.52786	225.66414	305.50305
1552257_a_at	291.58218	290.7219	285.82385	335.90002	235.0794
1552258_at	30.767345	31.28848	45.877693	39.481945	42.64955
1552261_at	41.316654	39.48357	55.891197	43.250927	41.93747
1552263_at	73.79253	80.8529	67.270935	97.50893	70.12854
1552264_a_at	110.1212	214.49841	181.10098	315.10193	188.26064
1552266_at	32.890938	34.15481	19.45849	23.117064	27.100496
1552269_at	18.904076	20.432787	24.65589	20.628496	21.51073
1552271_at	62.856636	65.46632	57.74975	69.37881	55.793846

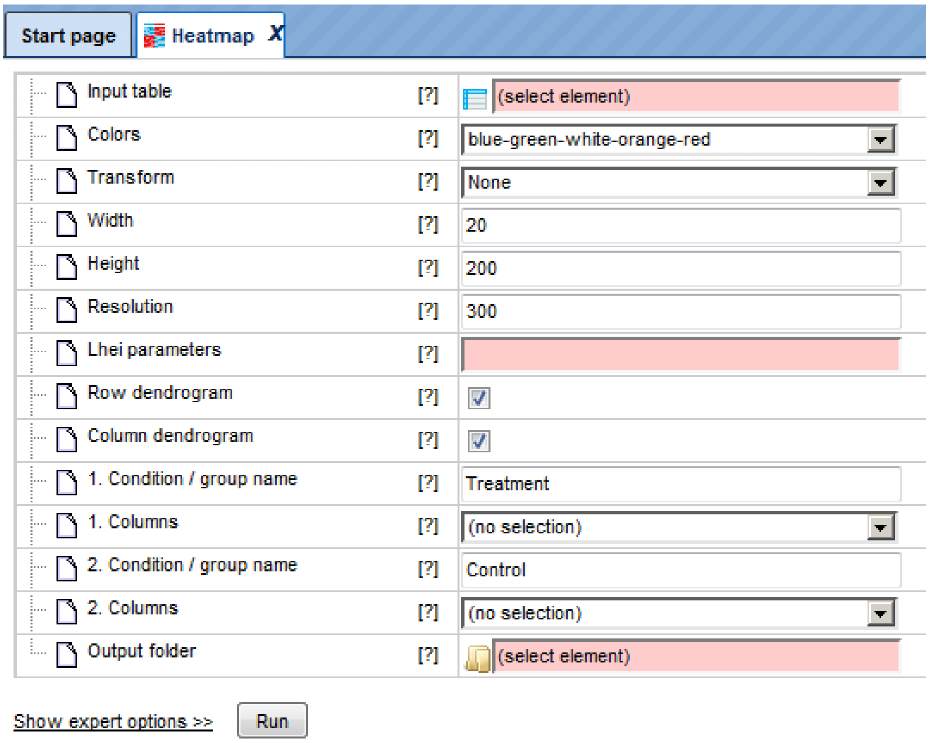
Important note. In the geneXplain platform, the probeset IDs are mapped to genes based on the Ensembl database. If some of the probeset IDs are not annotated in Ensembl, they cannot be mapped to genes and cannot be used for further analysis. That means, you can normalize data and calculate differentially expressed probes. These steps can be done on the probeset ID level, before conversion to genes. The step of converting probeset IDs into genes is depending on the annotation provided by the Ensembl database.

10.1.2. Heatmap

This tool creates a heatmap for the numerical data matrix provided with the input data table. The heatmap is limited to input tables with at most 5000 input rows. It is a graphical representation of data where the individual values contained in a matrix are represented as colors. The output folder contains a TIFF image of the heatmap as well as the ordered lists of row ids (e.g. RNA or gene ids) and column ids. The output tables can be used to extract subsets of correlated rows or columns revealed by the hierarchical clustering and/or the heatmap presentation.

An example can be found here: data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/Normalized (RMA) subset_heatmap

Step 1: Open the workflow input form from the Start page. It will open in the main Work Space and looks as shown below:



Input table	[?]	(select element)
Colors	[?]	blue-green-white-orange-red
Transform	[?]	None
Width	[?]	20
Height	[?]	200
Resolution	[?]	300
Lhei parameters	[?]	
Row dendrogram	[?]	☑
Column dendrogram	[?]	☑
1. Condition / group name	[?]	Treatment
1. Columns	[?]	(no selection)
2. Condition / group name	[?]	Control
2. Columns	[?]	(no selection)
Output folder	[?]	(select element)

Step 2: Specify the **Input table** with e.g. normalized data for different experimental conditions. To specify the input table, you can drag & drop it from your project within the tree area.

Step 3: Specify the Transformation to be applied to data values from the drop-down menu. Possible transformations are Log or Rank; the default is None.

Step 4: Select **Width** and **Height** as well the **Resolution** of the output image.

Step 5: Specify layout heights (Lhei). This should be a list of 2 (when no column groups are specified) or 3 (with column groups) float values separated by comma that adjust the heights of the layout parts. Please refer to the documentation of R's heatmap.2 {gplots} tool for details.

Step 6: Please check Row dendrogram and/or Column dendrogram if you want to have a raw dendogram and/or a dendrogram with column in the output image.

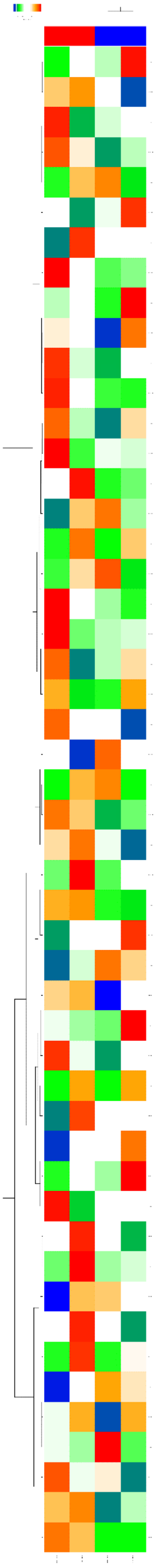
Step 7: Specify the conditions/groups names for up to five groups (expert level).

Step 8: Specify the conditions/groups. They are shown as columns of the input table. You can select the column names for each condition/group via the drop-down menu.

Step 9: Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink field “select element” in the field **Output folder**, and a new window will open, where you can select the location of the results folder and define its name.

Start the method by pressing the [Run workflow] button.

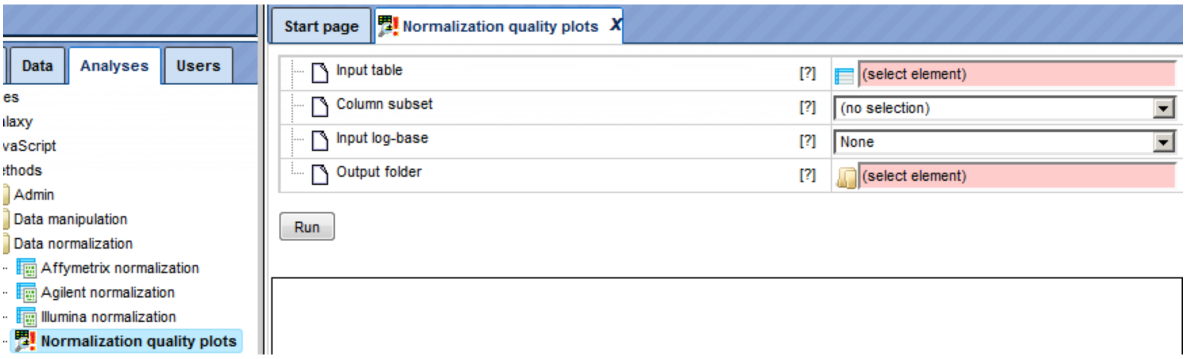
Output are two tables (heatmap_columns and heatmap_genes) and one image file in .tiff format. This image file needs to be downloaded and can be used in any graphical program/presentation.



10.1.3. Normalization quality plots

This tool can be applied plot densities of columns of a data table. As its name implies the intended use case is to inspect the quality of results of normalization as conducted in microarray experiments.

Example outputs, a box plot and a density plot, are shown at the end of this section. Colors were automatically assigned to selected columns.



Input table [?] (select element)
Column subset [?] (no selection)
Input log-base [?] None
Output folder [?] (select element)

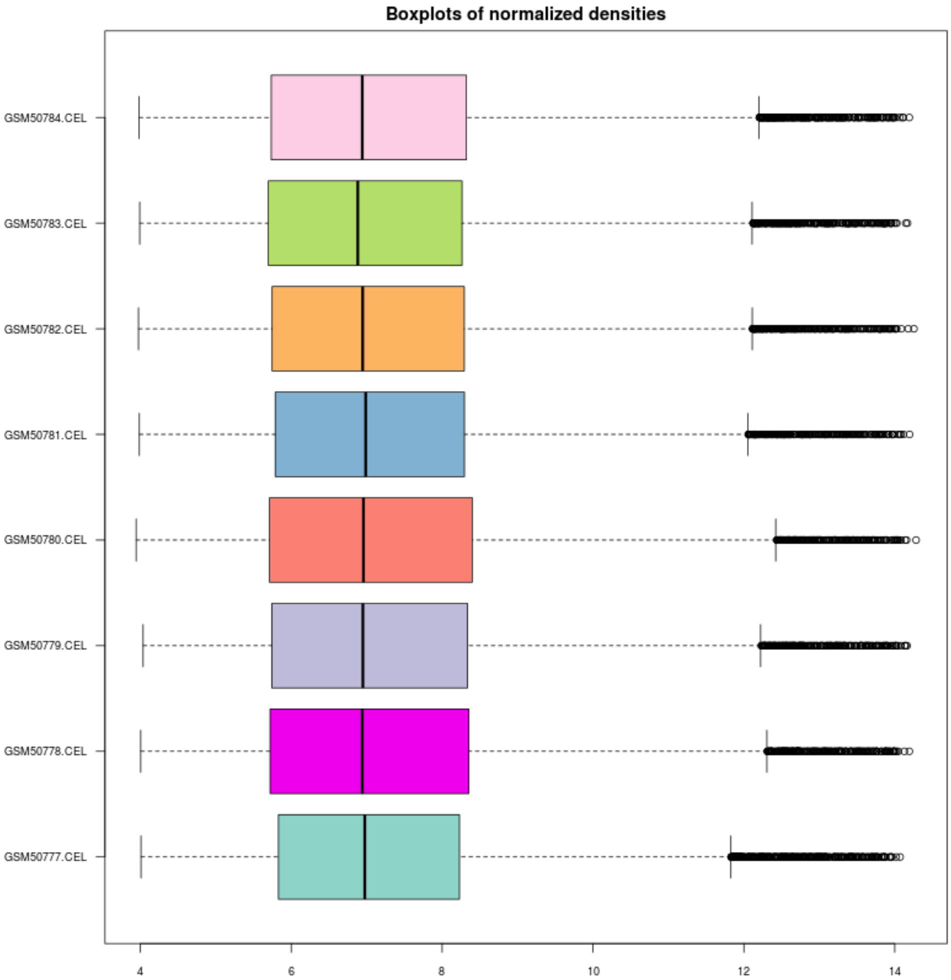
The input parameters are described in the following.

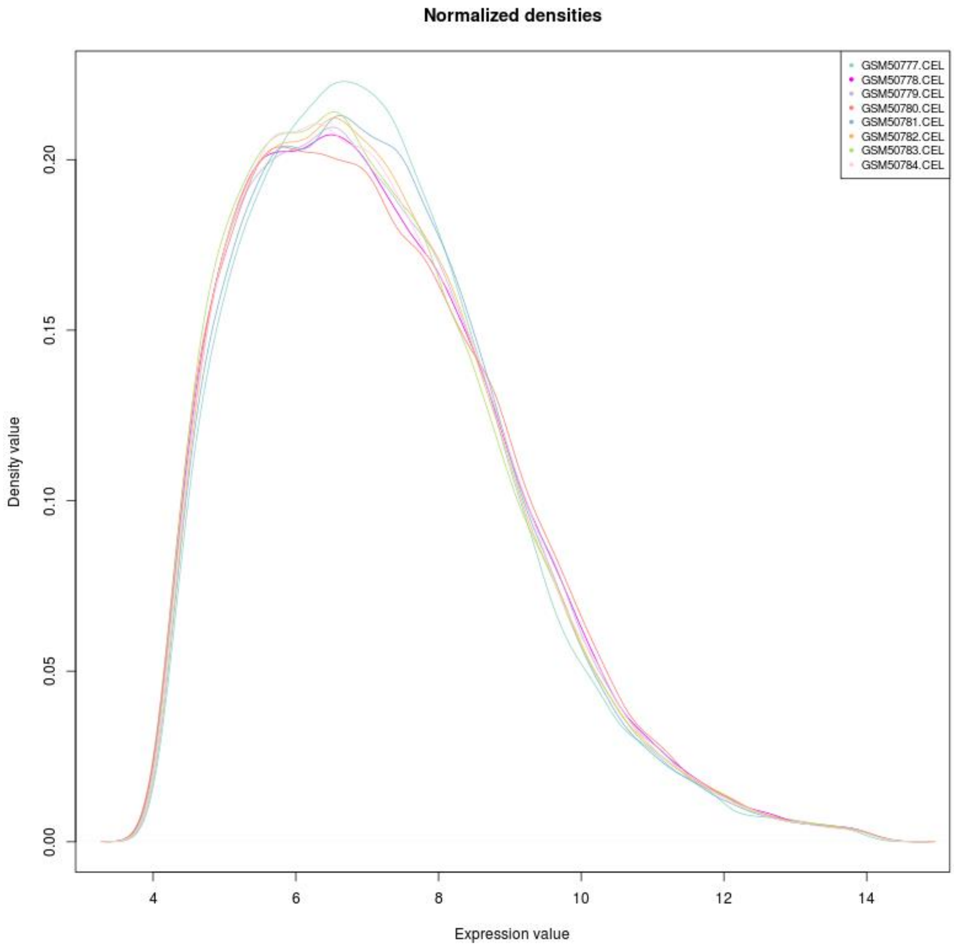
Input table: This table contains the numerical columns to analyze.

Column subset: Here you can select the set of columns to show in plots.

Input log-base: Densities and box plots will be computed for data on the log$_2$ scale. Here you can specify the actual scale of the input data. If the log-base is log$_2$, the tool will use the data values as is.

Output folder: The output folder will contain a density plot and a box plot for the specified columns.


Boxplots of normalized densities
GSM50784.CEL
GSM50783.CEL
GSM50782.CEL
GSM50781.CEL
GSM50780.CEL
GSM50779.CEL
GSM50778.CEL
GSM50777.CEL
4
6
8
10
12
14

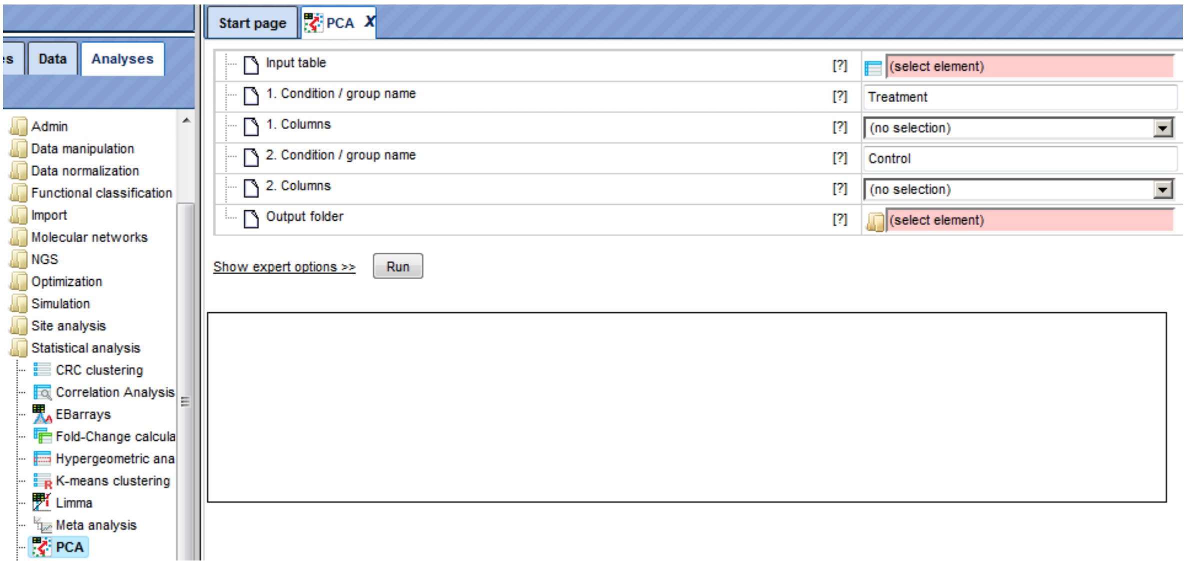


10.1.4. Principal Component Analysis (PCA)

PCA is a statistical method that transforms data in a way, so that a maximum amount of variance within the data can be expressed in fewer or, at most, as many dimensions as the original data. The new dimensions onto which data are projected are the principal components. They capture the original variance in decreasing order, so that the first principal component presents most of the variance. PCA is often used reduce the complexity of (to compress) or to identify groups in high-dimensional data.

This tool applies PCA to a table of numerical data, e.g. to normalized microarray measurements. For visualization purposes one can assign columns to one of up to five groups, which will be differentially colored in the generated output (scatter plot).

The output is stored in a specified folder and consists of three files. The **PCA Scatter plot** shows the items of specified groups at their transformed coordinates according to the first two principal components. The entire set of coordinates is available in **PCA Transformed coordinates**. Finally, the table **PCA Component importance** provides information about the relative importance of each principal component with respect to the proportion of explained variance.



Input table	[?]	(select element)
1. Condition / group name	[?]	Treatment
1. Columns	[?]	(no selection)
2. Condition / group name	[?]	Control
2. Columns	[?]	(no selection)
Output folder	[?]	(select element)

The input parameters for PCA are described in the following.

Input table: This table contains the numerical columns to analyze.

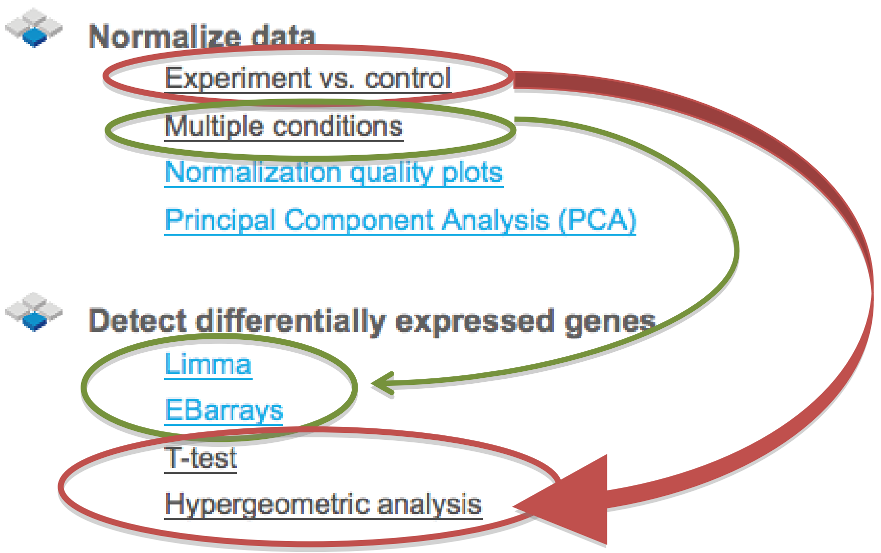
1-5. Condition / group name: One can specify up to five groups of columns. These fields contain the names that will be shown in outputs. Please note that unnamed groups are not considered, a name is not assigned automatically.

1-5. Columns: These fields contain the selected columns. Please note that column selections are not considered without a corresponding name. Columns can only be specified once.

Output folder: The output folder will contain the described output files.

10.2. Detect differentially expressed genes

After the microarray results are normalized, the next step is to compute differentially expressed genes (DEG).There are four different statistical methods provided for DEG calculation in the platform: T-test, hypergeometric analysis, Limma, and EBarrays. For DEG calculation by T-test and hypergeometric analysis, there are predefined workflows, which take as input two tables with the normalized data, for two different conditions, referred to as *Experiment normalized* and *Control normalized*. The methods Limma and EBarrays require one input table with all the conditions, and you can specify up to five different conditions for one run of each of these two methods.


Normalize data
Experiment vs. control
Multiple conditions
Normalization quality plots
Principal Component Analysis (PCA)
Detect differentially expressed genes
Limma
EBarrays
T-test
Hypergeometric analysis

If you applied the normalization method “Experiment vs. control”, you can detect DEGs applying T-test and/or hypergeometric analysis to the workflows, highlighted in green in the picture above. If you applied the normalization method “Multiple conditions”, you can detect DEGs with Limma and/or EBarrays, highlighted in red.

In this chapter, the predefined workflows for DEG calculation with T-test and with hypergeometric analysis are described in detail. For a description of Limma and EBarray please refer to the sections 4.2.1 and 4.2.2, respectively.

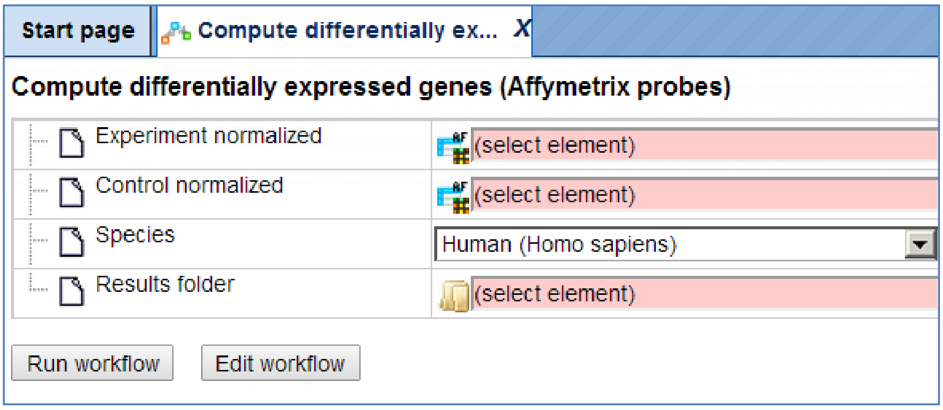
10.2.1. Detect differentially expressed genes with T-test

This workflow is designed to find the set of up-regulated and down-regulated genes applying Student’s T-test. There are three workflows designed for different experimental platforms (Affymetrix, Agilent and Illumina).

In the first step p-values for normalized files are calculated for all probes using the “Up and Down Identification” analysis. This analysis applies Student’s T-test for p-value calculation, thus the number of data points should be at least three for each experiment and control.

To launch the workflow, follow these steps:

Step1. Open the workflow input form from the Start page. It looks as shown below:



Compute differentially expressed genes (Affymetrix probes)

Experiment normalized	(select element)
Control normalized	(select element)
Species	Human (Homo sapiens)
Results folder	(select element)

Step 2. Specify the tables with normalized data in the fields **Experiment normalized** and **Control normalized**. You can drag it from your project within the tree area and drop it in the pink box of the fields. Alternatively, you may click on the pink field *(select element)* and a new window will be opened, where you can select the input tables.

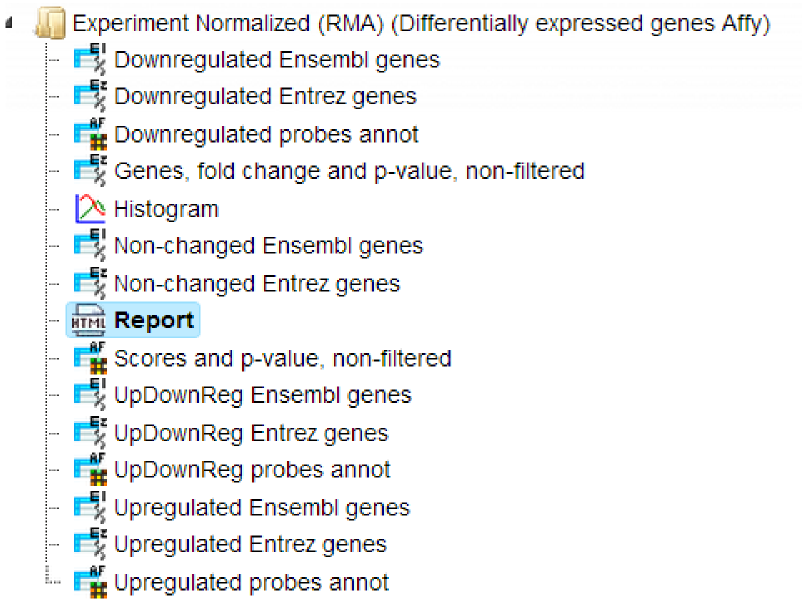
The further steps of the workflow are demonstrated by means of the tables in one of the pre-prepared examples. You can find these tables in the *Examples* folder, under data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma

Step 3. Specify the biological species of the input sets in the field **Species** by selecting the required biological species from the drop-down menu.

Step 4. Define where the folder with the results should be located in the tree. You can do so by clicking on the pink field *(select element)* in the field **Results folder**, and a new window will be opened, where you can select the location of the results folder and define its name.

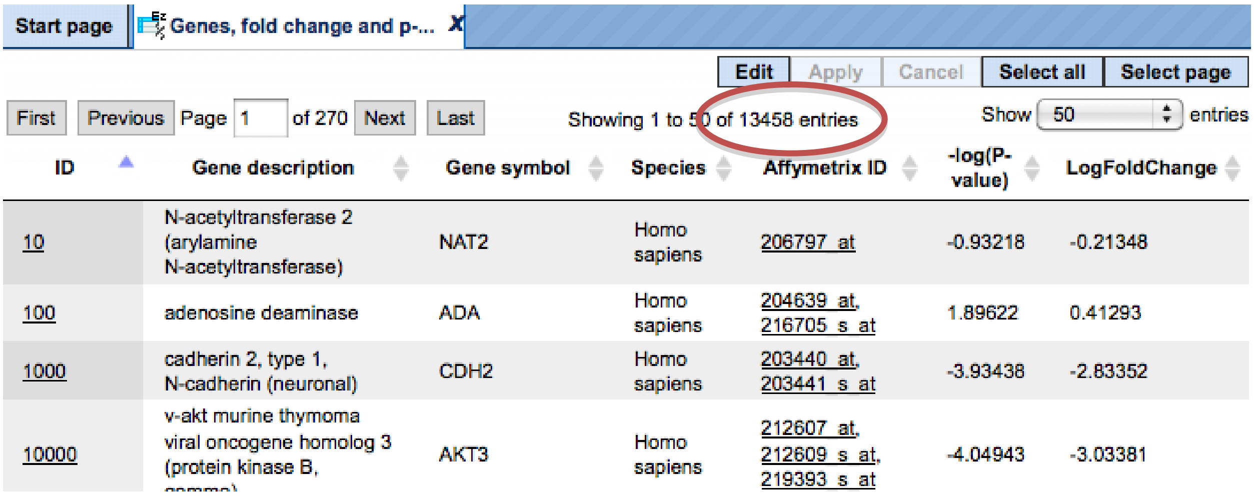
After entering all input fields press [Run workflow] and wait till the workflow is completed.

The output is a folder with several files as shown below:



Experiment Normalized (RMA) (Differentially expressed genes Affy)
- Downregulated Ensembl genes
- Downregulated Entrez genes
- Downregulated probes annot
- Genes, fold change and p-value, non-filtered
- Histogram
- Non-changed Ensembl genes
- Non-changed Entrez genes
- Report
- Scores and p-value, non-filtered
- UpDownReg Ensembl genes
- UpDownReg Entrez genes
- UpDownReg probes annot
- Upregulated Ensembl genes
- Upregulated Entrez genes
- Upregulated probes annot

The table **Genes, fold change and P-value, non-filtered**. This table contains all genes with LogFoldChange and p-value calculated; each row corresponds to one gene.



Page 1 of 270

Showing 1 to 50 of 13458 entries

Show 50 entries

ID	Gene description	Gene symbol	Species	Affymetrix ID	-log(P-value)	LogFoldChange
10	N-acetyltransferase 2 (arylamine N-acetyltransferase)	NAT2	Homo sapiens	206797_at	-0.93218	-0.21348
100	adenosine deaminase	ADA	Homo sapiens	204639_at, 216705_s_at	1.89622	0.41293
1000	cadherin 2, type 1, N-cadherin (neuronal)	CDH2	Homo sapiens	203440_at, 203441_s_at	-3.93438	-2.83352
10000	v-akt murine thymoma viral oncogene homolog 3 (protein kinase B, gamma)	AKT3	Homo sapiens	212607_at, 212609_s_at, 219393_s_at	-4.04943	-3.03381

The columns **ID**, **Gene description** and **Gene symbol** present Entrez identifiers for the genes, a full name for each gene, and a standard gene symbol, respectively. The column **Species** shows the corresponding taxonomic species. The column **AffymetrixID** contains the probe set IDs corresponding to each gene, and you can see sometimes more than one Affymetrix probe corresponding to one gene. The column **LogFoldChange** shows the base 2 logarithm of the ratio between expression value in experiment vs. control. The column **–log(P-value)** shows the negative base 10 logarithm of the *p*-value.

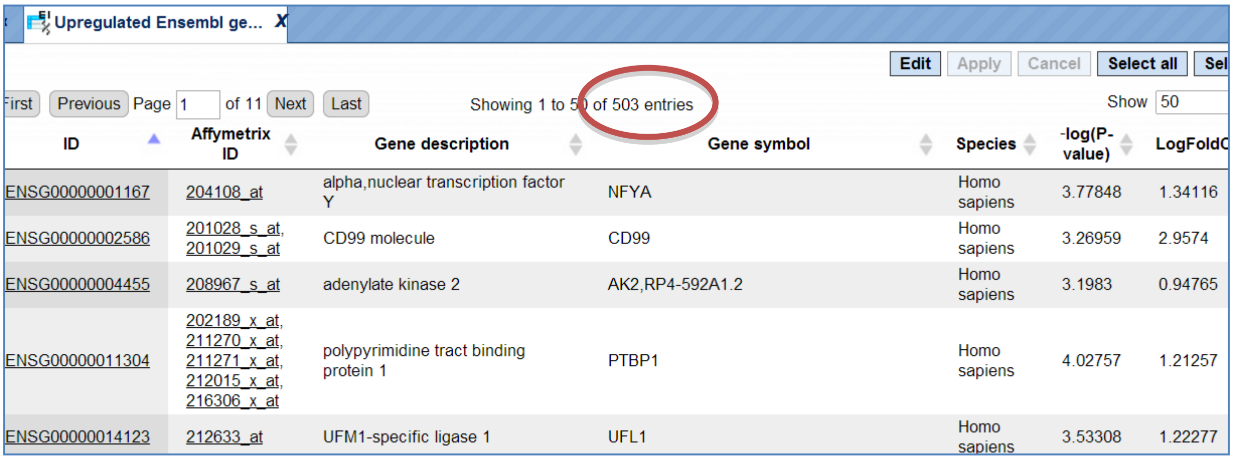
Please note that the column **–log(P-value)**, according to a widely accepted convention, has algebraic signs according to being up- (positive values) or down-regulated (negative values).

In the course of worklow progression, this table has been filtered by several conditions in parallel to identify up-regulated, down-regulated, and non-changed Affymetrix probeset IDs and genes.

The filtering criteria used are:

- *For up-regulated probes: LogFoldChange>0.5 and -log_P_value_>3*
- *For down- regulated probes: LogFoldChange<-0.5 and -log_P_value_<-3*
- *For non-changed genes : LogFoldChange<0.002 and LogFoldChange>-0.002*

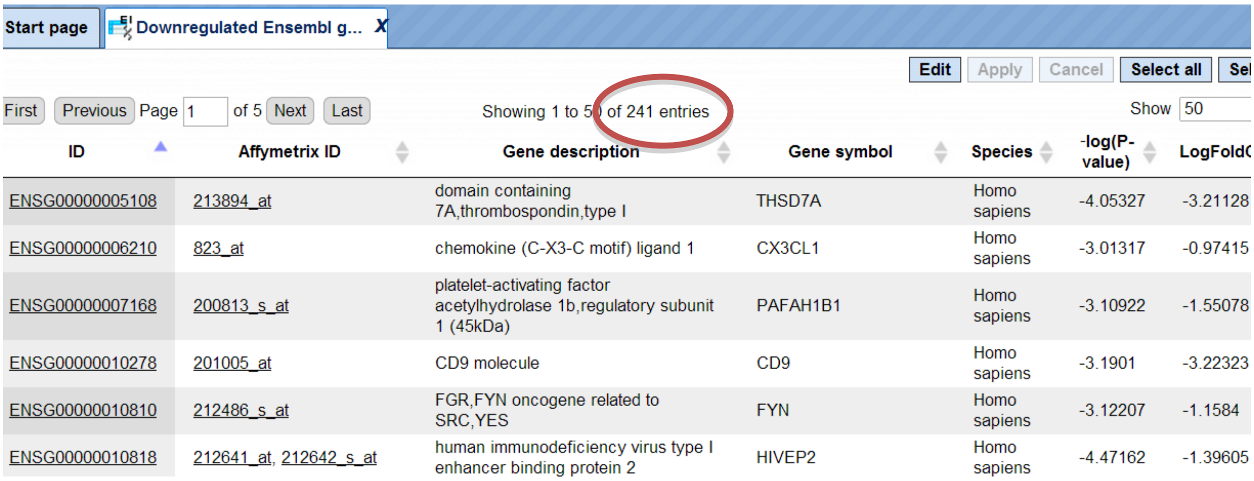
The table **Upregulated Ensembl genes**. You can find the number of the resulting up-regulated genes written on top of each output table (highlighted by the red circle):



Showing 1 to 50 of 503 entries

ID	Affymetrix ID	Gene description	Gene symbol	Species	-log(P-value)	LogFoldC
ENSG00000001167	204108_at	alpha,nuclear transcription factor Y	NFYA	Homo sapiens	3.77848	1.34116
ENSG00000002586	201028_s_at, 201029_s_at	CD99 molecule	CD99	Homo sapiens	3.26959	2.9574
ENSG00000004455	208967_s_at	adenylate kinase 2	AK2,RP4-592A1.2	Homo sapiens	3.1983	0.94765
ENSG00000011304	202189_x_at, 211270_x_at, 211271_x_at, 212015_x_at, 216306_x_at	polypyrimidine tract binding protein 1	PTBP1	Homo sapiens	4.02757	1.21257
ENSG00000014123	212633_at	UFM1-specific ligase 1	UFL1	Homo sapiens	3.53308	1.22277

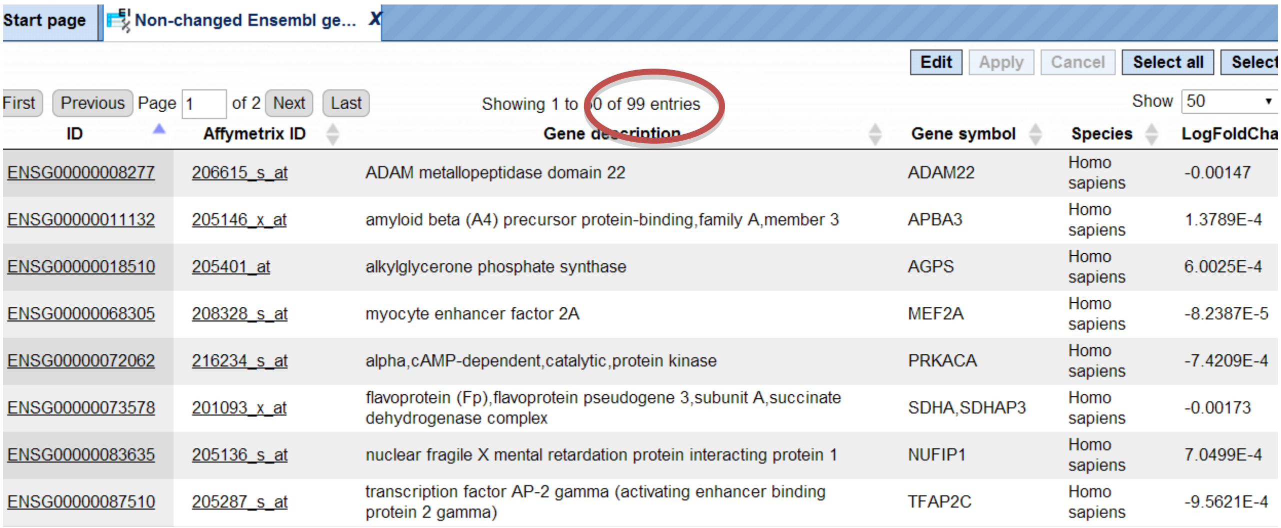
The table **Downregulated Ensembl genes**. The structure and the meaning of the columns in the tables are the same as in the Upregulated Ensembl genes table.



Showing 1 to 50 of 241 entries

ID	Affymetrix ID	Gene description	Gene symbol	Species	-log(P-value)	LogFoldC
ENSG00000005108	213894_at	domain containing 7A,thrombospondin,type I	THSD7A	Homo sapiens	-4.05327	-3.21128
ENSG00000006210	823_at	chemokine (C-X3-C motif) ligand 1	CX3CL1	Homo sapiens	-3.01317	-0.97415
ENSG00000007168	200813_s_at	platelet-activating factor acetylhydrolase 1b,regulatory subunit 1 (45kDa)	PAFAH1B1	Homo sapiens	-3.10922	-1.55078
ENSG00000010278	201005_at	CD9 molecule	CD9	Homo sapiens	-3.1901	-3.22323
ENSG00000010810	212486_s_at	FGR,FYN oncogene related to SRC,YES	FYN	Homo sapiens	-3.12207	-1.1584
ENSG00000010818	212641_at, 212642_s_at	human immunodeficiency virus type I enhancer binding protein 2	HIVEP2	Homo sapiens	-4.47162	-1.39605

The table **Non-changed Ensembl genes**.



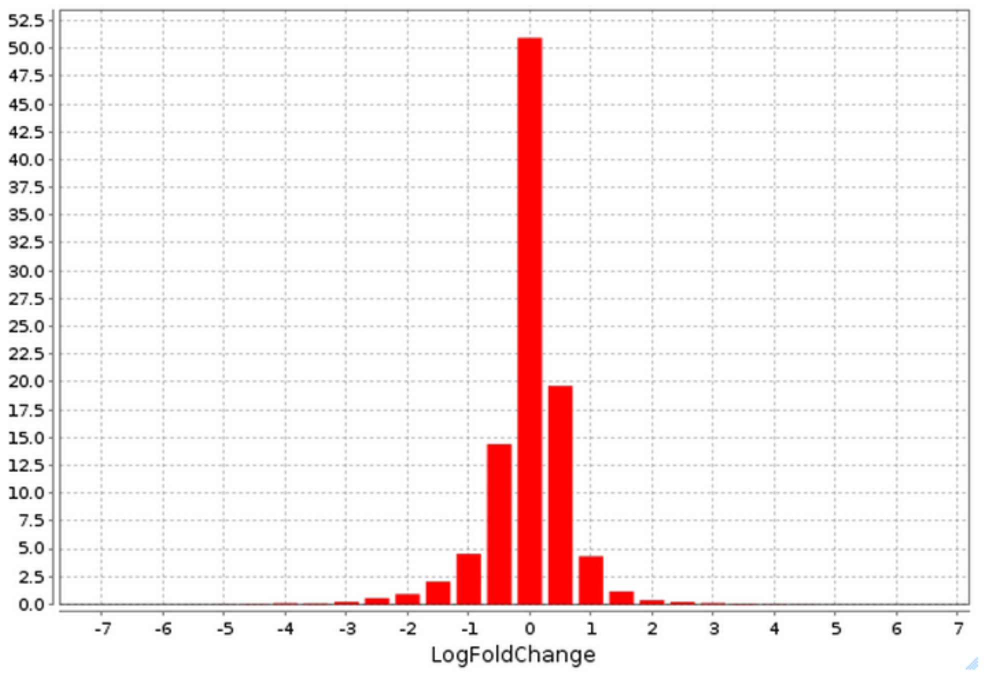
Showing 1 to 50 of 99 entries

ID	Affymetrix ID	Gene description	Gene symbol	Species	LogFoldCha
ENSG00000008277	206615_s_at	ADAM metallopeptidase domain 22	ADAM22	Homo sapiens	-0.00147
ENSG00000011132	205146_x_at	amyloid beta (A4) precursor protein-binding,family A,member 3	APBA3	Homo sapiens	1.3789E-4
ENSG00000018510	205401_at	alkylglycerone phosphate synthase	AGPS	Homo sapiens	6.0025E-4
ENSG00000068305	208328_s_at	myocyte enhancer factor 2A	MEF2A	Homo sapiens	-8.2387E-5
ENSG00000072062	216234_s_at	alpha,cAMP-dependent,catalytic,protein kinase	PRKACA	Homo sapiens	-7.4209E-4
ENSG00000073578	201093_x_at	flavoprotein (Fp),flavoprotein pseudogene 3,subunit A,succinate dehydrogenase complex	SDHA,SDHAP3	Homo sapiens	-0.00173
ENSG00000083635	205136_s_at	nuclear fragile X mental retardation protein interacting protein 1	NUFIP1	Homo sapiens	7.0499E-4
ENSG00000087510	205287_s_at	transcription factor AP-2 gamma (activating enhancer binding protein 2 gamma)	TFAP2C	Homo sapiens	-9.5621E-4

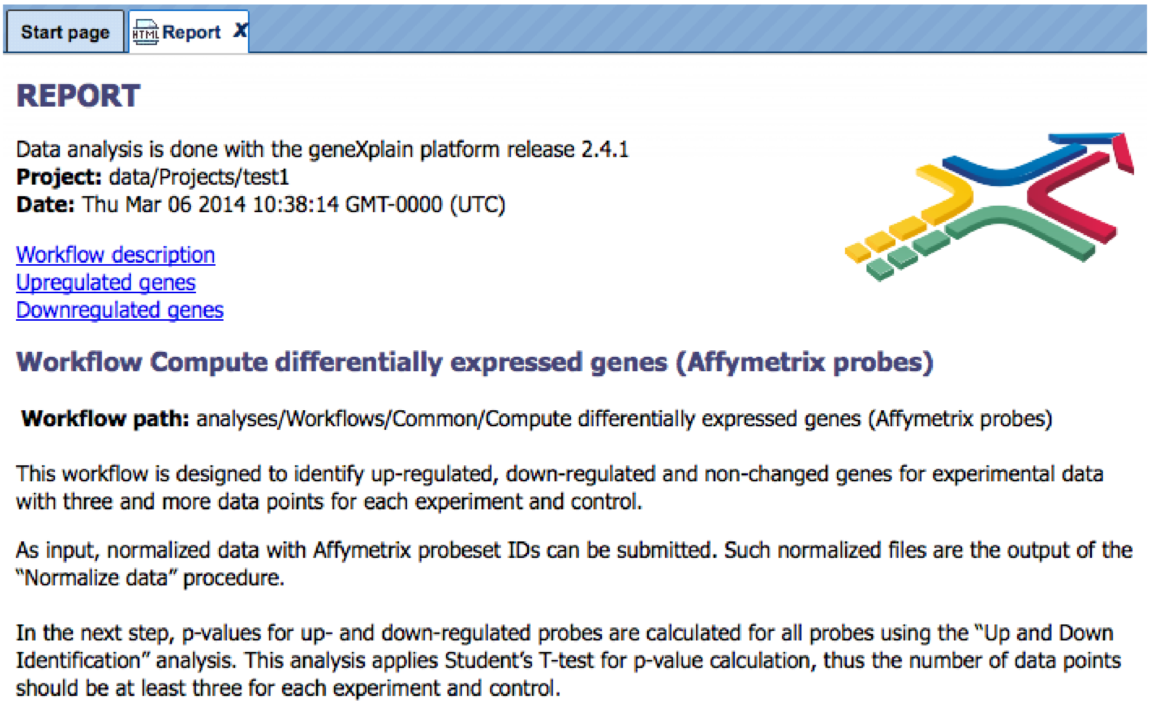
In this example the number of up-regulated, down-regulated and non-changed genes are 503, 241, and 99, respectively.

These individual output files can be used further as input for running other workflows as described in the following sections.

The plot (📈) contains a histogram of the log fold change distribution for all genes:



The Report. The workflow summarizes all results and automatically produces a report. In addition you can have a look at the list of both up-regulated and down-regulated genes.



REPORT

Data analysis is done with the geneXplain platform release 2.4.1
Project: data/Projects/test1
Date: Thu Mar 06 2014 10:38:14 GMT-0000 (UTC)

Workflow description
Upregulated genes
Downregulated genes

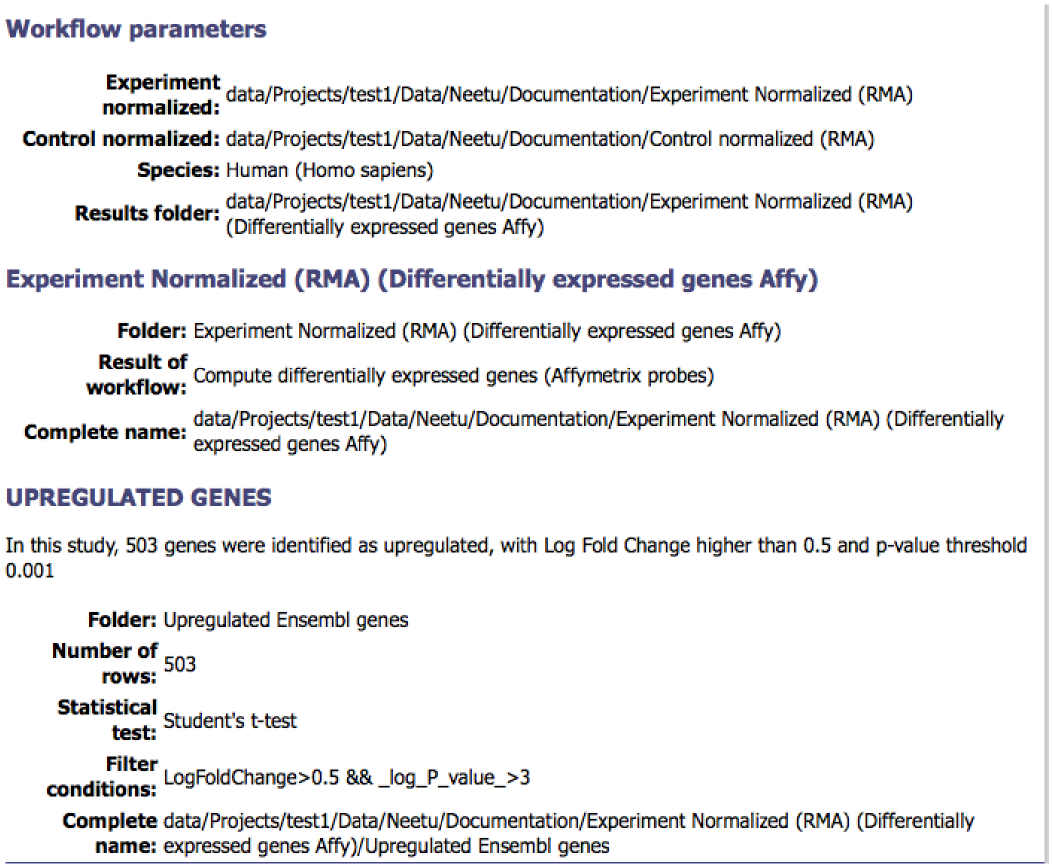
Workflow Compute differentially expressed genes (Affymetrix probes)

Workflow path: analyses/Workflows/Common/Compute differentially expressed genes (Affymetrix probes)

This workflow is designed to identify up-regulated, down-regulated and non-changed genes for experimental data with three and more data points for each experiment and control.

As input, normalized data with Affymetrix probeset IDs can be submitted. Such normalized files are the output of the "Normalize data" procedure.

In the next step, p-values for up- and down-regulated probes are calculated for all probes using the "Up and Down Identification" analysis. This analysis applies Student's T-test for p-value calculation, thus the number of data points should be at least three for each experiment and control.



Workflow parameters

Experiment normalized:	data/Projects/test1/Data/Neetu/Documentation/Experiment Normalized (RMA)
Control normalized:	data/Projects/test1/Data/Neetu/Documentation/Control normalized (RMA)
Species:	Human (Homo sapiens)
Results folder:	data/Projects/test1/Data/Neetu/Documentation/Experiment Normalized (RMA) (Differentially expressed genes Affy)

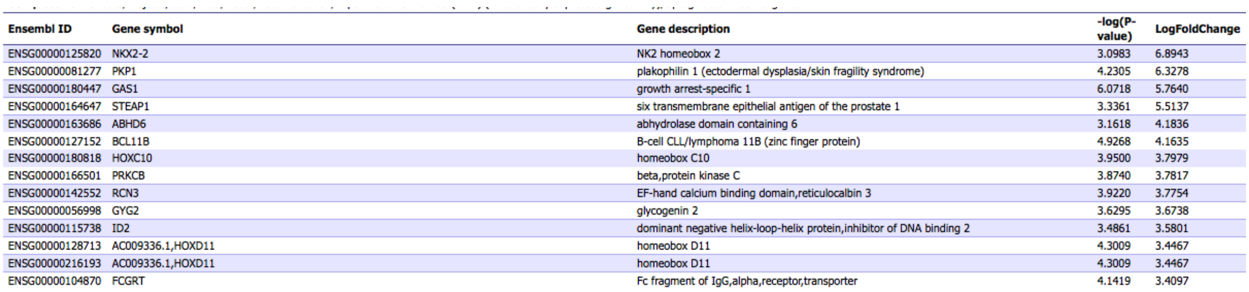
Experiment Normalized (RMA) (Differentially expressed genes Affy)

Folder:	Experiment Normalized (RMA) (Differentially expressed genes Affy)
Result of workflow:	Compute differentially expressed genes (Affymetrix probes)
Complete name:	data/Projects/test1/Data/Neetu/Documentation/Experiment Normalized (RMA) (Differentially expressed genes Affy)

UPREGULATED GENES

In this study, 503 genes were identified as upregulated, with Log Fold Change higher than 0.5 and p-value threshold 0.001

Folder:	Upregulated Ensembl genes
Number of rows:	503
Statistical test:	Student's t-test
Filter conditions:	LogFoldChange>0.5 && _log_P_value_>3
Complete name:	data/Projects/test1/Data/Neetu/Documentation/Experiment Normalized (RMA) (Differentially expressed genes Affy)/Upregulated Ensembl genes



Ensembl ID	Gene symbol	Gene description	-log(P-value)	LogFoldChange
ENSG00000125820	NKX2-2	NK2 homeobox 2	3.0983	6.8943
ENSG00000081277	PKP1	plakophilin 1 (ectodermal dysplasia/skin fragility syndrome)	4.2305	6.3278
ENSG00000180447	GAS1	growth arrest-specific 1	6.0718	5.7640
ENSG00000164647	STEAP1	six transmembrane epithelial antigen of the prostate 1	3.3361	5.5137
ENSG00000163686	ABHD6	abhydrolase domain containing 6	3.1618	4.1836
ENSG00000127152	BCL11B	B-cell CLL/lymphoma 11B (zinc finger protein)	4.9268	4.1635
ENSG00000180818	HOXC10	homeobox C10	3.9500	3.7979
ENSG00000166501	PRKCB	beta,protein kinase C	3.8740	3.7817
ENSG00000142552	RCN3	EF-hand calcium binding domain,reticulocalbin 3	3.9220	3.7754
ENSG00000056998	GYG2	glycogenin 2	3.6295	3.6738
ENSG00000115738	ID2	dominant negative helix-loop-helix protein,inhibitor of DNA binding 2	3.4861	3.5801
ENSG00000128713	AC009336.1,HOXD11	homeobox D11	4.3009	3.4467
ENSG00000216193	AC009336.1,HOXD11	homeobox D11	4.3009	3.4467
ENSG00000104870	FCGRT	Fc fragment of IgG,alpha,receptor,transporter	4.1419	3.4097

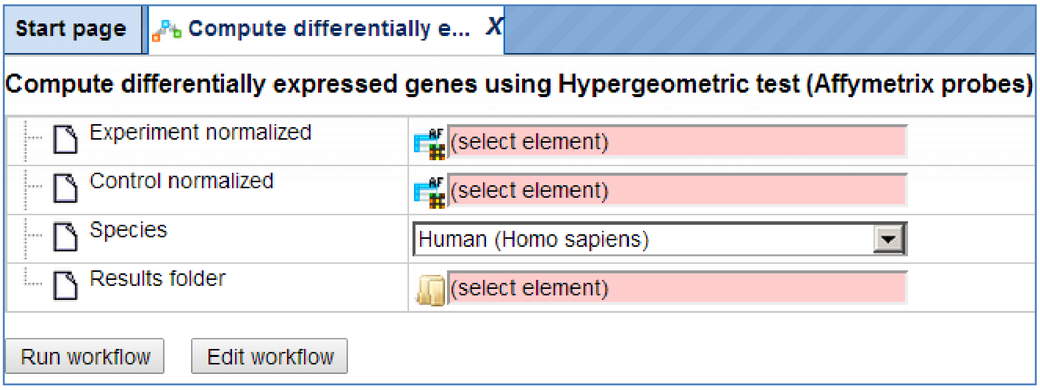
This report can be exported in html format.

10.2.2. Detect differentially expressed genes by hypergeometric analysis

This workflow is very similar to the one described above in section 9.2.1. The principal difference is in the statistical method for calculation of the p-value. In this workflow, the p-value is calculated by hypergeometric analysis (Y.V.Kondrakhin, R.N.Sharipov, A.E.Kel, F.A.Kolpakov. (2008) Identification of Differentially Expressed Genes by Meta-Analysis of Microarray Data on Breast Cancer, *In Silico Biology*, 8: 383-411).

Tip If you have just two or even one data point in each experiment and control (e.g. one CEL file in experiment and one CEL file in control), you can apply hypergeometric analysis to calculate DEGs. In contrast to the T-test which requires at least three data points, hypergeometric analysis can make calculations for two and even one data point in each normalized experiment and normalized control files. This allows to calculate DEGs to compare, for instance, one patient data set with one healthy data set.

The workflow input form looks as shown below:



Compute differentially expressed genes using Hypergeometric test (Affymetrix probes)

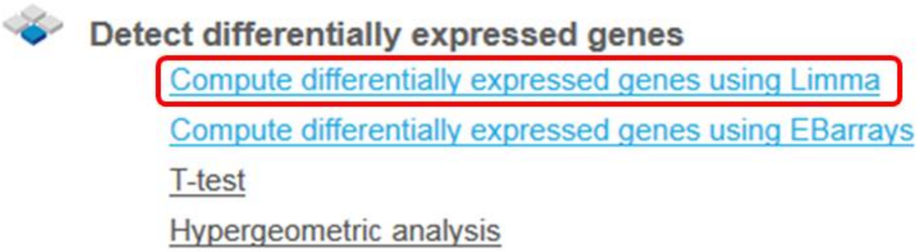
Experiment normalized	(select element)
Control normalized	(select element)
Species	Human (Homo sapiens)
Results folder	(select element)

The output folder and the structure of the individual tables, as well as the report, are similar to those described in 9.2.1.

10.2.3. Detect differentially expressed genes with Limma

This workflow is designed to find sets of up-regulated and down-regulated genes starting with the normalized table of your expression data. Please refer to section 4.2.1 for details on this particular analysis method. This workflow is designed for different experimental platforms (Affymetrix, Agilent and Illumina).

In the first step this workflow computes the differential expression between up to five conditions / groups. Each group corresponds to one experimental condition (time point, treatment, cell type, etc.) or control. You can specify 2 to 5 conditions. An input table is a data table that contains several columns with normalized measurement values, e.g. from a normalized microarray experiment. All possible contrasts between groups are considered and their output is stored in a common folder. Conditions are compared in the specified order from first to fifth; e.g. for the given conditions named 1, 2 and 3, the output will contain the contrasts "Condition 1 versus Condition 2", Condition 1 versus Condition 3" and "Condition 2 versus Condition 3". ". The workflow can be found on the Start page, under the button Microarrays, under the section "Detect differentially expressed genes".



Detect differentially expressed genes

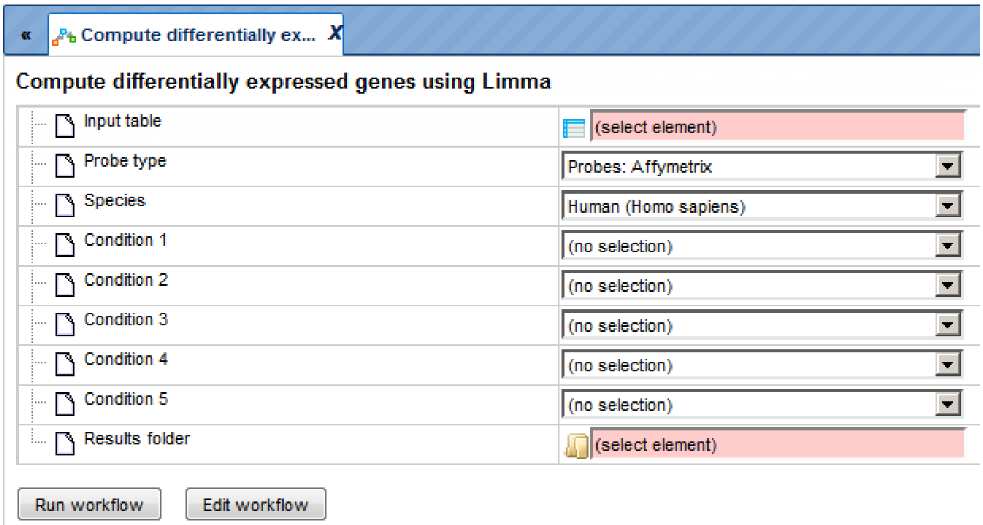
Compute differentially expressed genes using Limma

Compute differentially expressed genes using EBarrays

T-test

Hypergeometric analysis

Step1. Open the workflow input form from the Start page. It looks as shown below:



Compute differentially expressed genes using Limma

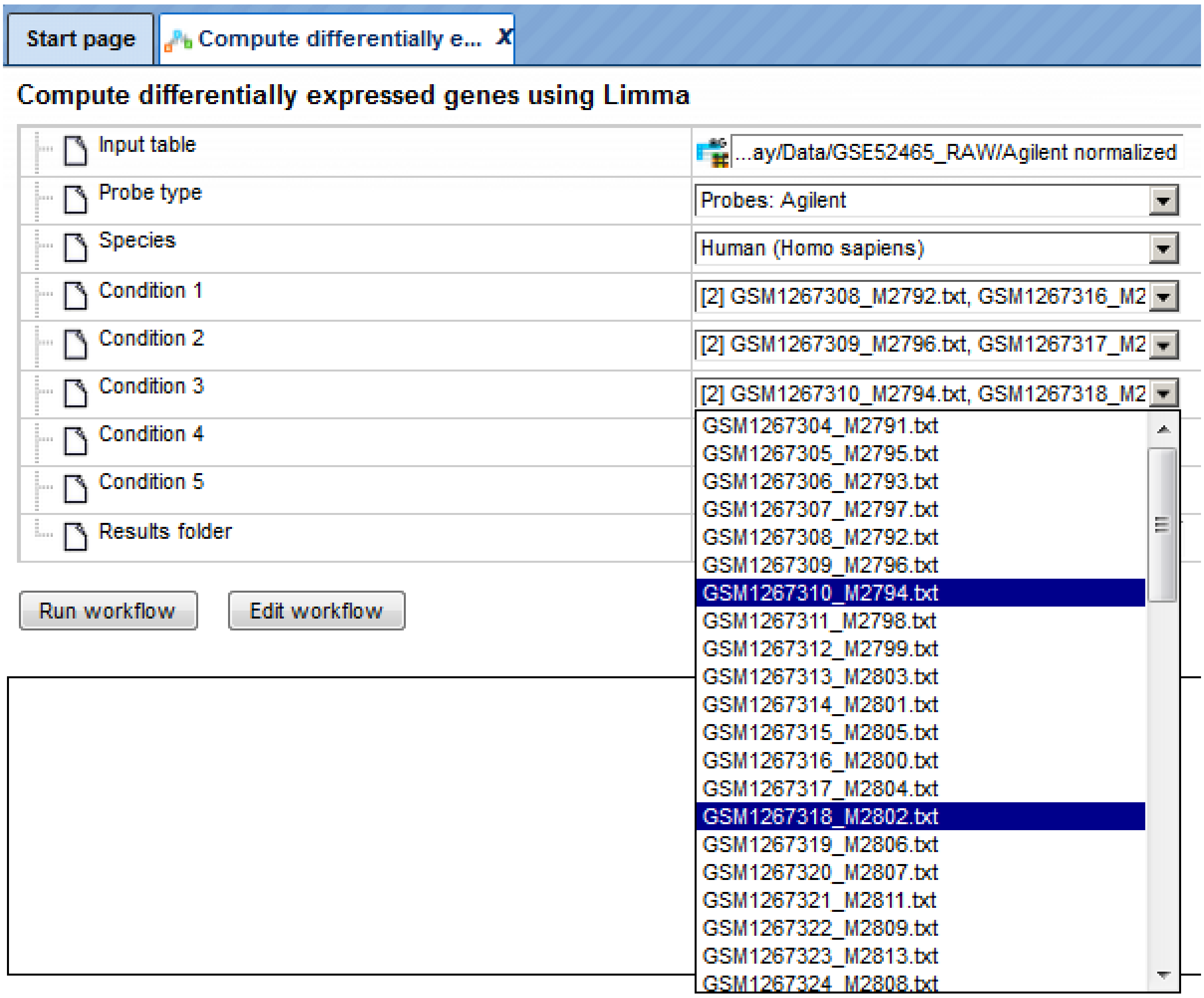
Input table	(select element)
Probe type	Probes: Affymetrix
Species	Human (Homo sapiens)
Condition 1	(no selection)
Condition 2	(no selection)
Condition 3	(no selection)
Condition 4	(no selection)
Condition 5	(no selection)
Results folder	(select element)

Step 2. Specify the table with normalized data in the field **Input table**. You can drag it from your project within the tree area and drop it in the pink box of the fields. Alternatively, you may click on the pink field *(select element)* and a new window will be opened, where you can select the input table.

The further steps of the workflow are demonstrated by means of the tables in one of the pre-prepared examples. You can find these tables in the *Examples* folder, under data/Examples/Cytokine-triggered gene expression in cell cycle stages, GSE52465, Agilent-014850 microarray/Data/Agilent normalized DEGs with limma

Step 3. Specify the biological species of the input table in the field **Species** by selecting it from the drop-down menu.

Step 4. Specify the conditions / groups. They are shown as columns of the input table. You can select the column names for each condition/group via drop-down menu.



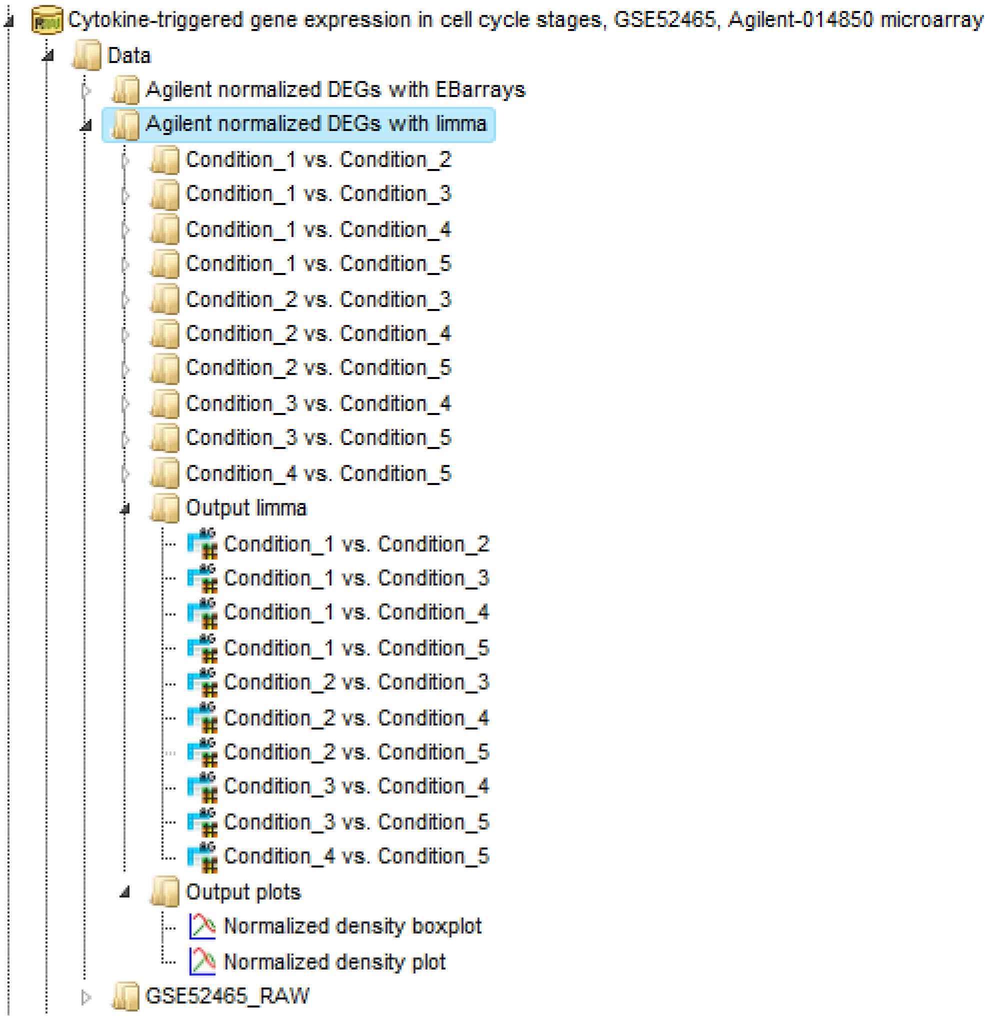
Compute differentially expressed genes using Limma

Input table	...ay/Data/GSE52465_RAW/Agilent normalized
Probe type	Probes: Agilent
Species	Human (Homo sapiens)
Condition 1	[2] GSM1267308_M2792.txt, GSM1267316_M2
Condition 2	[2] GSM1267309_M2796.txt, GSM1267317_M2
Condition 3	[2] GSM1267310_M2794.txt, GSM1267318_M2
Condition 4	
Condition 5	
Results folder	

GSM1267304_M2791.txt
GSM1267305_M2795.txt
GSM1267306_M2793.txt
GSM1267307_M2797.txt
GSM1267308_M2792.txt
GSM1267309_M2796.txt
GSM1267310_M2794.txt
GSM1267311_M2798.txt
GSM1267312_M2799.txt
GSM1267313_M2803.txt
GSM1267314_M2801.txt
GSM1267315_M2805.txt
GSM1267316_M2800.txt
GSM1267317_M2804.txt
GSM1267318_M2802.txt
GSM1267319_M2806.txt
GSM1267320_M2807.txt
GSM1267321_M2811.txt
GSM1267322_M2809.txt
GSM1267323_M2813.txt
GSM1267324_M2808.txt

Step 5. Define where the folder with the results should be located in the tree. You can do so by clicking on the pink field (select element) in the field **Results folder,** and a new window will be opened, where you can select the location of the results folder and define its name.

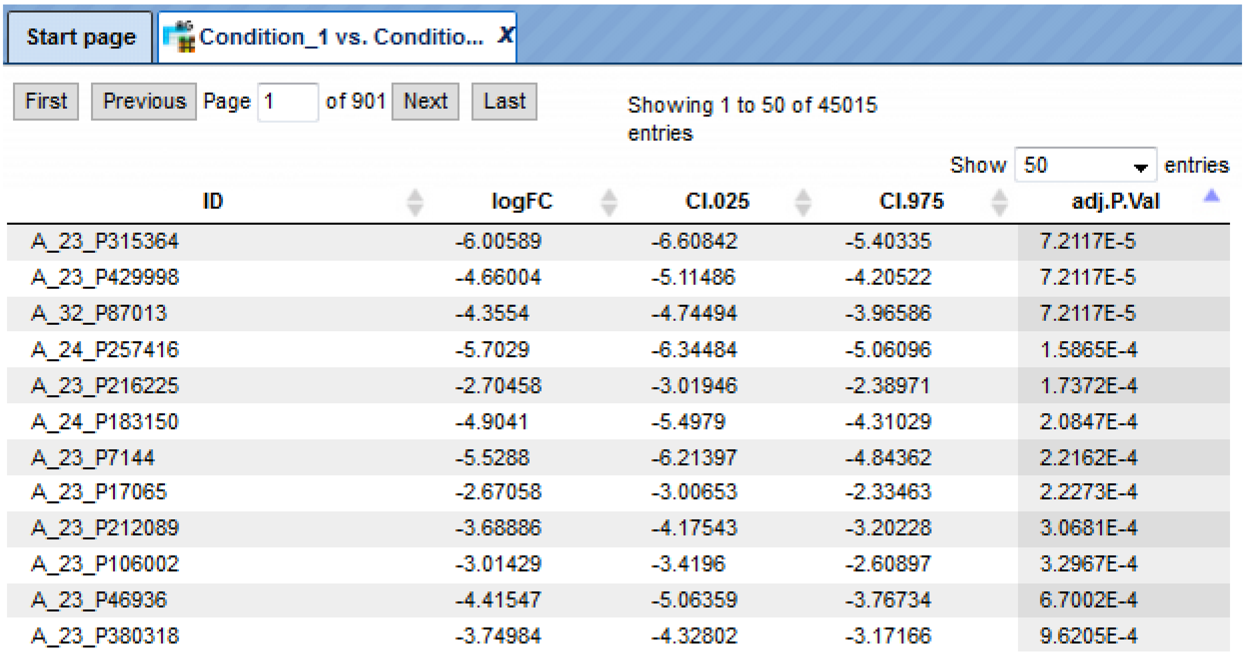
After entering all input fields press [Run workflow] and wait till the workflow is completed. The output is a folder with 10 result folders (e.g. *Condition_1 vs. Condition_2* for DEG calculation), a folder for each individual comparison (*Output limma*) and one *Output plots* folder, as shown below:



- Cytokine-triggered gene expression in cell cycle stages, GSE52465, Agilent-014850 microarray
  - Data
    - Agilent normalized DEGs with EBarrays
    - Agilent normalized DEGs with limma
      - Condition_1 vs. Condition_2
      - Condition_1 vs. Condition_3
      - Condition_1 vs. Condition_4
      - Condition_1 vs. Condition_5
      - Condition_2 vs. Condition_3
      - Condition_2 vs. Condition_4
      - Condition_2 vs. Condition_5
      - Condition_3 vs. Condition_4
      - Condition_3 vs. Condition_5
      - Condition_4 vs. Condition_5
      - Output limma
        - Condition_1 vs. Condition_2
        - Condition_1 vs. Condition_3
        - Condition_1 vs. Condition_4
        - Condition_1 vs. Condition_5
        - Condition_2 vs. Condition_3
        - Condition_2 vs. Condition_4
        - Condition_2 vs. Condition_5
        - Condition_3 vs. Condition_4
        - Condition_3 vs. Condition_5
        - Condition_4 vs. Condition_5
      - Output plots
        - Normalized density boxplot
        - Normalized density plot
    - GSE52465_RAW

The **Normalized density boxplot** and the **Normalized density plot** in the folder *Output plots* show a quality control of the input normalized data table. Please refer to section 10.1.3 for details on this particular analysis method.

The tables in the folder *Output limma* are the output tables from the limma method, sorted via adjusted p-values.

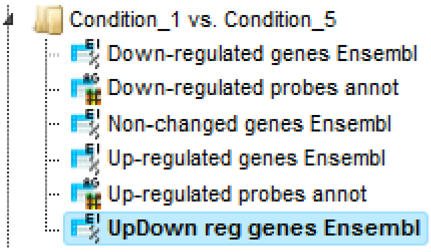


Showing 1 to 50 of 45015 entries

Show 50 entries

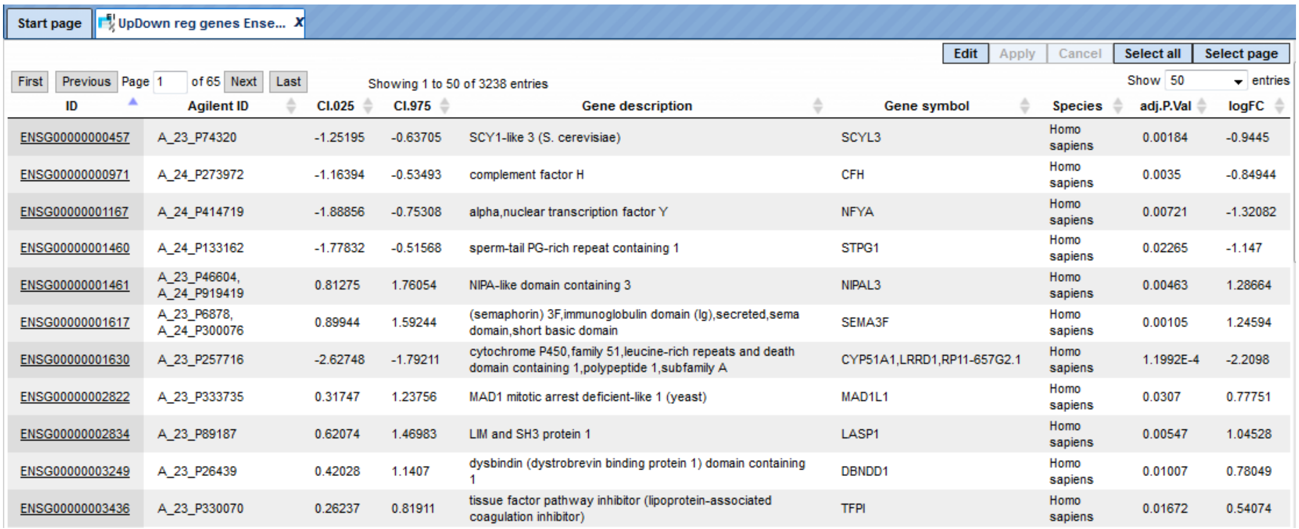
ID	logFC	CI.025	CI.975	adj.P.Val
A_23_P315364	-6.00589	-6.60842	-5.40335	7.2117E-5
A_23_P429998	-4.66004	-5.11486	-4.20522	7.2117E-5
A_32_P87013	-4.3554	-4.74494	-3.96586	7.2117E-5
A_24_P257416	-5.7029	-6.34484	-5.06096	1.5865E-4
A_23_P216225	-2.70458	-3.01946	-2.38971	1.7372E-4
A_24_P183150	-4.9041	-5.4979	-4.31029	2.0847E-4
A_23_P7144	-5.5288	-6.21397	-4.84362	2.2162E-4
A_23_P17065	-2.67058	-3.00653	-2.33463	2.2273E-4
A_23_P212089	-3.68886	-4.17543	-3.20228	3.0681E-4
A_23_P106002	-3.01429	-3.4196	-2.60897	3.2967E-4
A_23_P46936	-4.41547	-5.06359	-3.76734	6.7002E-4
A_23_P380318	-3.74984	-4.32802	-3.17166	9.6205E-4

The 10 output folders for each comparison e.g. Condition_1 vs. Condition_5 contain the results of the identified up-, down- and non-regulated Ensembl genes.



Condition_1 vs. Condition_5
- Down-regulated genes Ensembl
- Down-regulated probes annot
- Non-changed genes Ensembl
- Up-regulated genes Ensembl
- Up-regulated probes annot
- UpDown reg genes Ensembl

The table **UpDown reg genes Ensembl** in the Folder *Condition_1 vs. Condition_5* contains all differentially expressed genes filtered by LogFoldChange and p-value for up- and down-regulated genes; each row corresponds to one gene.



Showing 1 to 50 of 3238 entries

Show 50 entries

ID	Agilent ID	CI.025	CI.975	Gene description	Gene symbol	Species	adj.P.Val	logFC
ENSG00000000457	A_23_P74320	-1.25195	-0.63705	SCY1-like 3 (S. cerevisiae)	SCYL3	Homo sapiens	0.00184	-0.9445
ENSG00000000971	A_24_P273972	-1.16394	-0.53493	complement factor H	CFH	Homo sapiens	0.0035	-0.84944
ENSG00000001167	A_24_P414719	-1.88856	-0.75308	alpha,nuclear transcription factor Y	NFYA	Homo sapiens	0.00721	-1.32082
ENSG00000001460	A_24_P133162	-1.77832	-0.51568	sperm-tail PG-rich repeat containing 1	STPG1	Homo sapiens	0.02265	-1.147
ENSG00000001461	A_23_P46604, A_24_P919419	0.81275	1.76054	NIPA-like domain containing 3	NIPAL3	Homo sapiens	0.00463	1.28664
ENSG00000001617	A_23_P6878, A_24_P300076	0.89944	1.59244	(semaphorin) 3F,immunoglobulin domain (Ig),secreted,sema domain,short basic domain	SEMA3F	Homo sapiens	0.00105	1.24594
ENSG00000001630	A_23_P257716	-2.62748	-1.79211	cytochrome P450,family 51,leucine-rich repeats and death domain containing 1,polypeptide 1,subfamily A	CYP51A1,LRRD1,RP11-657G2.1	Homo sapiens	1.1992E-4	-2.2098
ENSG00000002822	A_23_P333735	0.31747	1.23756	MAD1 mitotic arrest deficient-like 1 (yeast)	MAD1L1	Homo sapiens	0.0307	0.77751
ENSG00000002834	A_23_P89187	0.62074	1.46983	LIM and SH3 protein 1	LASP1	Homo sapiens	0.00547	1.04528
ENSG00000003249	A_23_P26439	0.42028	1.1407	dysbindin (dystrobrevin binding protein 1) domain containing 1	DBNDD1	Homo sapiens	0.01007	0.78049
ENSG00000003436	A_23_P330070	0.26237	0.81911	tissue factor pathway inhibitor (lipoprotein-associated coagulation inhibitor)	TFPI	Homo sapiens	0.01672	0.54074

The columns **ID, Gene description** and **Gene symbol** represent Ensembl identifiers for the genes, a full name for each gene, and a standard gene symbol, respectively. The column **Species** shows the corresponding taxonomic species. The column **AffymetrixID** contains the probe set IDs corresponding to each gene, and you can see sometimes more than one Affymetrix probe corresponding to one gene. The column **logFC** shows the base 2 logarithm of the ratio between expression values in experiment vs. control. The column **adj.P.Val** shows the adjusted *p*-value (Benjamini-Hochberg).

In the course of workflow progression, this table has been filtered by several conditions in parallel to identify up-regulated, down-regulated, and non-changed probeset IDs that were then converted into Ensembl gene identifiers.

The filtering criteria used are:

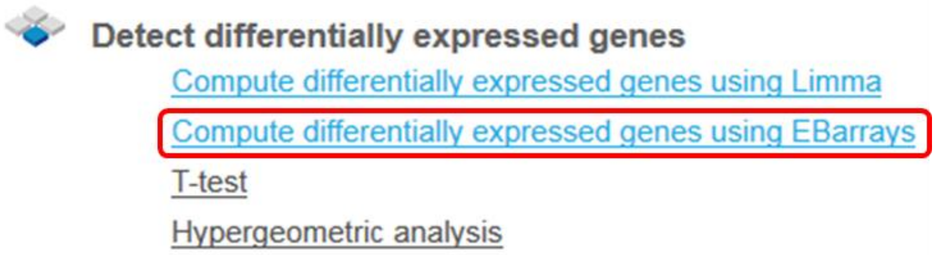
- *For up-regulated genes: logFC >0.5 and adj.P.Val <0.05*
- *For down- regulated genes: logFC <-0.5 and adj.P.Val <0.05*
- *For non-changed genes :logFC <0.002 and logFC >-0.002*

10.2.4. Detect differentially expressed genes with EBarrays

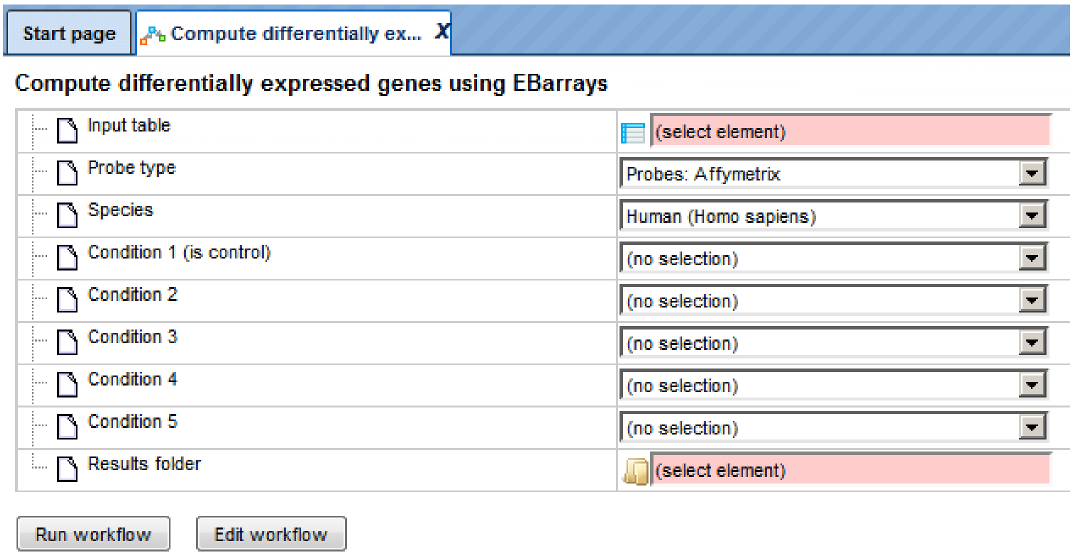
Similarly to the workflow described above, this workflow is designed to find the set of up-regulated and down-regulated genes starting with a normalized table of your expression data, but using a different statistical method, EBarrays. Please refer to section 4.2.2 for details on this particular analysis method. This workflow is designed for different experimental platforms (Affymetrix, Agilent and Illumina).

In the first step the workflow computes the differential expression between up to five conditions/groups. Each group corresponds to one experimental condition (time point, treatment, cell type, etc.) or control. You can specify 2 to 5 conditions. An input table is a data table that contains normalized measurement values, e.g. from a normalized microarray experiment. All possible contrasts between groups are considered and their output is stored in a common folder. Conditions are compared in the specified order from first to fifth. E.g. given conditions named 1, 2 and 3, the output will contain the contrasts "Condition 1 versus Condition 2", Condition 1 versus Condition 3" and "Condition 2 versus Condition 3".

The workflow can be found on the Start page, under the button Microarrays, under the section "Detect differentially expressed genes".


Detect differentially expressed genes
Compute differentially expressed genes using Limma
Compute differentially expressed genes using EBarrays
T-test
Hypergeometric analysis

Step1. Open the workflow input form from the Start page. It looks as shown below:



Compute differentially expressed genes using EBarrays

Input table	(select element)
Probe type	Probes: Affymetrix
Species	Human (Homo sapiens)
Condition 1 (is control)	(no selection)
Condition 2	(no selection)
Condition 3	(no selection)
Condition 4	(no selection)
Condition 5	(no selection)
Results folder	(select element)

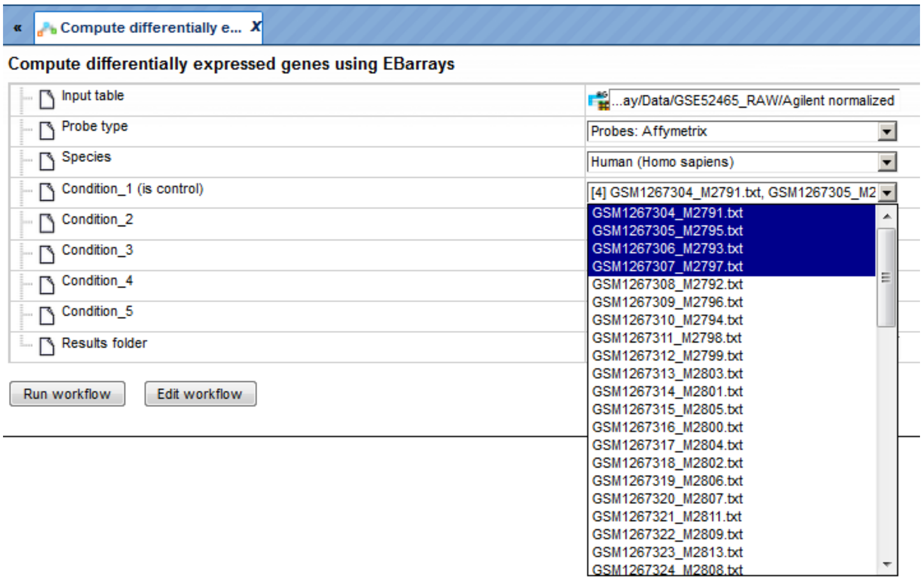
Step 2. Specify the table with normalized data in the field **Input table**. You can drag it from your project within the tree area and drop it in the pink box of the fields. Alternatively, you may click on the pink field *(select element)* and a new window will be opened, where you can select the input table.

The further steps of the workflow are demonstrated by means of the tables in one of the pre-prepared examples. You can find these tables in the *Examples* folder, under

data/Examples/Cytokine-triggered gene expression in cell cycle stages, GSE52465, Agilent-014850 microarray/Data/Agilent normalized DEGs with EBarrays

Step 3. Specify the biological species of the input table in the field **Species** by selecting it from the drop-down menu.

Step 4. Specify the conditions/groups of the input table. You can select the column names for each condition/group via drop-down menu.



Compute differentially expressed genes using EBarrays

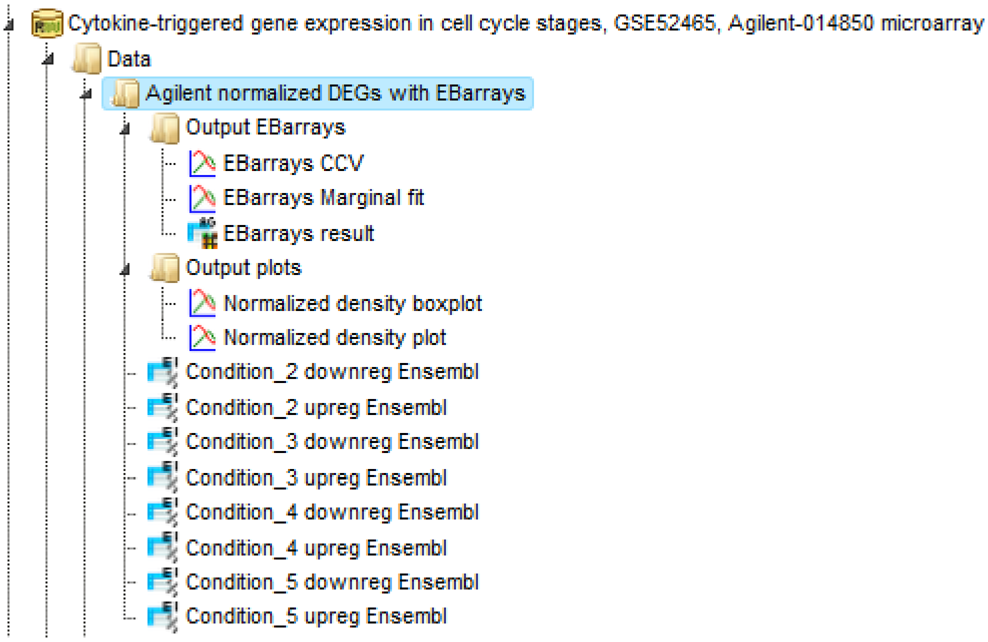
Input table	...ay/Data/GSE52465_RAW/Agilent normalized
Probe type	Probes: Affymetrix
Species	Human (Homo sapiens)
Condition_1 (is control)	[4] GSM1267304_M2791.txt, GSM1267305_M2
Condition_2	
Condition_3	
Condition_4	
Condition_5	
Results folder	

GSM1267304_M2791.txt
GSM1267305_M2795.txt
GSM1267306_M2793.txt
GSM1267307_M2797.txt
GSM1267308_M2792.txt
GSM1267309_M2796.txt
GSM1267310_M2794.txt
GSM1267311_M2798.txt
GSM1267312_M2799.txt
GSM1267313_M2803.txt
GSM1267314_M2801.txt
GSM1267315_M2805.txt
GSM1267316_M2800.txt
GSM1267317_M2804.txt
GSM1267318_M2802.txt
GSM1267319_M2806.txt
GSM1267320_M2807.txt
GSM1267321_M2811.txt
GSM1267322_M2809.txt
GSM1267323_M2813.txt
GSM1267324_M2808.txt

Step 5. Define where the folder with the results should be located in the tree. You can do so by clicking on the pink field *(select element)* in the field **Results folder**, and a new window will be opened, where you can select the location of the results folder and define its name.

After entering all input fields press [Run workflow] and wait till the workflow is completed.

The output is a folder with several result folders and files as shown below:

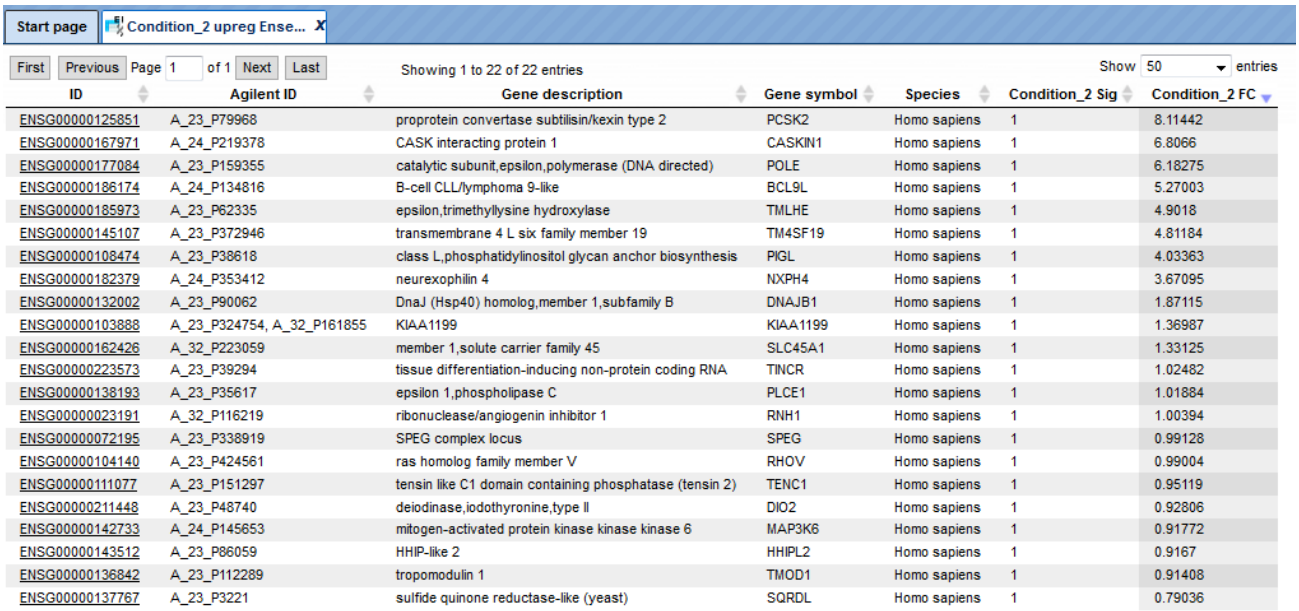


Cytokine-triggered gene expression in cell cycle stages, GSE52465, Agilent-014850 microarray
Data
Agilent normalized DEGs with EBarrays
Output EBarrays
EBarrays CCV
EBarrays Marginal fit
EBarrays result
Output plots
Normalized density boxplot
Normalized density plot
Condition_2 downreg Ensembl
Condition_2 upreg Ensembl
Condition_3 downreg Ensembl
Condition_3 upreg Ensembl
Condition_4 downreg Ensembl
Condition_4 upreg Ensembl
Condition_5 downreg Ensembl
Condition_5 upreg Ensembl

The **Normalized density boxplot** and the **Normalized density plot** in the folder *Output plots* show a quality control of the input normalized data table. Please refer to section 10.1.3 for details on this particular analysis method.

The table EBarrays result and two plots in the folder *Output EBarrays* are the output table and plots from the EBarrays method.

The tables for each condition (except control) e.g. **Condition_2 downreg Ensembl** and **Condition_2 upreg Ensembl** contain all differentially expressed genes with filtered LogFoldChange and p-value for up- and down-regulated genes; each row corresponds to one gene.



Showing 1 to 22 of 22 entries

ID	Agilent ID	Gene description	Gene symbol	Species	Condition_2 Sig	Condition_2 FC
ENSG00000125851	A_23_P79968	proprotein convertase subtilisin/kexin type 2	PCSK2	Homo sapiens	1	8.11442
ENSG00000167971	A_24_P219378	CASK interacting protein 1	CASKIN1	Homo sapiens	1	6.8066
ENSG00000177084	A_23_P159355	catalytic subunit,epsilon,polymerase (DNA directed)	POLE	Homo sapiens	1	6.18275
ENSG00000186174	A_24_P134816	B-cell CLL/lymphoma 9-like	BCL9L	Homo sapiens	1	5.27003
ENSG00000185973	A_23_P62335	epsilon,trimethyllysine hydroxylase	TMLHE	Homo sapiens	1	4.9018
ENSG00000145107	A_23_P372946	transmembrane 4 L six family member 19	TM4SF19	Homo sapiens	1	4.81184
ENSG00000108474	A_23_P38618	class L,phosphatidylinositol glycan anchor biosynthesis	PIGL	Homo sapiens	1	4.03363
ENSG00000182379	A_24_P353412	neurexophilin 4	NXPH4	Homo sapiens	1	3.67095
ENSG00000132002	A_23_P90062	DnaJ (Hsp40) homolog,member 1,subfamily B	DNAJB1	Homo sapiens	1	1.87115
ENSG00000103888	A_23_P324754, A_32_P161855	KIAA1199	KIAA1199	Homo sapiens	1	1.36987
ENSG00000162426	A_32_P223059	member 1,solute carrier family 45	SLC45A1	Homo sapiens	1	1.33125
ENSG00000223573	A_23_P39294	tissue differentiation-inducing non-protein coding RNA	TINCR	Homo sapiens	1	1.02482
ENSG00000138193	A_23_P35617	epsilon 1,phospholipase C	PLCE1	Homo sapiens	1	1.01884
ENSG00000023191	A_32_P116219	ribonuclease/angiogenin inhibitor 1	RNH1	Homo sapiens	1	1.00394
ENSG00000072195	A_23_P338919	SPEG complex locus	SPEG	Homo sapiens	1	0.99128
ENSG00000104140	A_23_P424561	ras homolog family member V	RHOV	Homo sapiens	1	0.99004
ENSG00000111077	A_23_P151297	tensin like C1 domain containing phosphatase (tensin 2)	TENC1	Homo sapiens	1	0.95119
ENSG00000211448	A_23_P48740	deiodinase,iodothyronine,type II	DIO2	Homo sapiens	1	0.92806
ENSG00000142733	A_24_P145653	mitogen-activated protein kinase kinase kinase 6	MAP3K6	Homo sapiens	1	0.91772
ENSG00000143512	A_23_P86059	HHIP-like 2	HHIPL2	Homo sapiens	1	0.9167
ENSG00000136842	A_23_P112289	tropomodulin 1	TMOD1	Homo sapiens	1	0.91408
ENSG00000137767	A_23_P3221	sulfide quinone reductase-like (yeast)	SQRDL	Homo sapiens	1	0.79036

The columns **ID, Gene description** and **Gene symbol** present Ensembl identifiers for the genes, a full name for each gene, and a standard gene symbol, respectively. The column **Species** shows the corresponding taxonomic species. The column **Agilent ID** contains the probe set IDs corresponding to each gene, and you can see sometimes more than one Affymetrix probe corresponding to one gene. The direction of differential expression can be derived from the fold change column e.g. **Condition_2 FC**, which contains the log2-fold changes. The EBarrays method estimates a critical posterior probability cutoff for the given FDR level on the basis of the fitted mixture model. Probes / genes exceeding this cutoff in some treatment are indicated by a value of 1 (instead of -1) in the output column named e.g. **Condition_2 Sig**.

In the course of workflow progression, this table has been filtered by several conditions in parallel to identify up-regulated and down-regulated genes.

The filtering criteria used are:

- *For up-regulated genes: Condition_x FC >0.5 and Condition_x Sig = 1*
- *For down- regulated genes: Condition_x FC <-0.5 and Condition_x Sig = 1*

In the folder *Output plots* there are two diagnostic plots named **EBarrays CCV** and **EBarrays Marginal fit**. These plots enable a judgment about whether the assumptions of the approach hold and how well the fitted model represents the data (please refer to the documentation of the EBarrays Bioconductor package for further details).

> **Note.** In the input field Condition 1 always the control condition should be specified. Each of the input conditions 2, 3, 4, and 5 are compared with condition 1.

10.3. Discover functional enrichment

This set of workflows helps to identify certain functional groups in your input list of genes or proteins, namely those that are particularly affected in a statistically significant manner.

The first approach to do this is the Gene Set Enrichment Analysis (GSEA). Applying this method to genes/protein in focus, the workflows will find out whether any category of Gene Ontology (GO), Reactome pathways, TRANSPATH® pathways or PROTEOME™ disease terms are statistically overrepresented among them, and if so, whether this overrepresentation is valid for the up- or down-regulated genes if expression values are present in your input table.

An alternative approach to GSEA is Functional classification, or mapping to ontologies. As input, you can use a table of genes/proteins. The difference of this option to GSEA is that no enrichment of the categories is calculated, but that all genes in the list are mapped to GO categories or other ontologies. For example, you can use tables with up-regulated or down-regulated genes coming from the previous analysis steps of microarray or RNA-seq experiments, or proteins identified in proteomics experiments, the lists of genes located nearby of ChIP_seq peaks, etc.

10.3.1. Gene Set Enrichment Analysis (GSEA)

There are two types of workflows, depending on the format of the input tables. You can start the GSEA with the normalized microarray tables, before calculating DEGs. The program first computes fold changes, which are then used to dynamically detect functional groups of genes that are differentially affected by the experimental conditions.

Alternatively, you can start the GSEA with any gene or protein table having a numerical column that can be used for enrichment calculations, e.g. expression fold change after calculating DEGs.

10.3.1.1. GSEA by GO categories and metabolic pathways

Gene set enrichment analysis (GSEA) with GO categories or with metabolic pathway annotation in REACTOME can be done by either starting from raw data of any of the widely used experimental platforms (**Affymetrix, Agilent, or Illumina**), or from a **single gene table** that you may have composed yourself.

Affymetrix, Agilent, or Illumina data sets

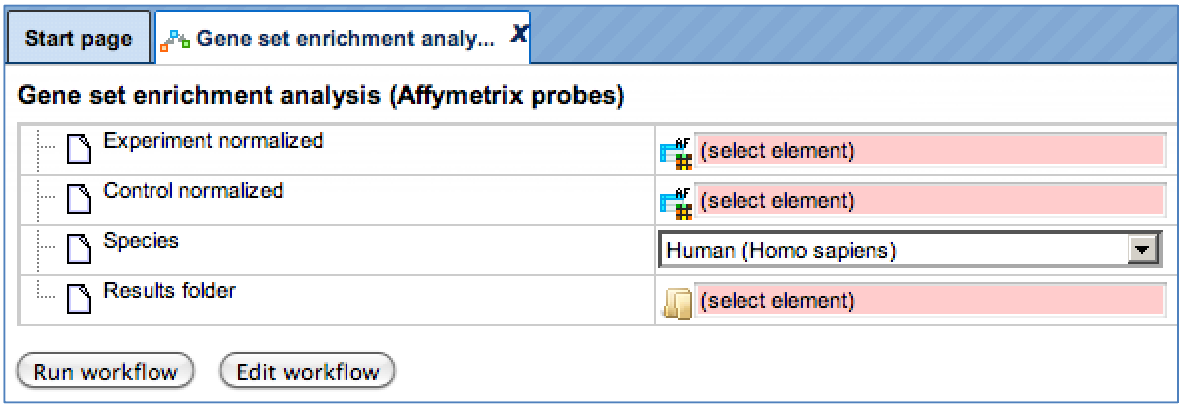
The three workflows under this category are designed to perform the GSEA by the three branches of Gene Ontology, biological process, molecular function and cellular component as well as by the Reactome pathways:

- ❖ *Gene Set Enrichment Analysis (Affymetrix probes),*
- ❖ *Gene Set Enrichment Analysis (Agilent probes),*
- ❖ *Gene Set Enrichment Analysis (Illumina probes).*

Each of the three workflows differs in the format of the input data, for Affymetrix, Agilent or Illumina microarray platforms, respectively. The analysis performed by these workflows and the interpretation of the results are the same. As an example, let’s consider an Affymetrix-specific workflow.

To launch the workflow, follow these steps:

Step1. Open the workflow input form from the Start page. It opens in the main Work Space and looks as shown below:

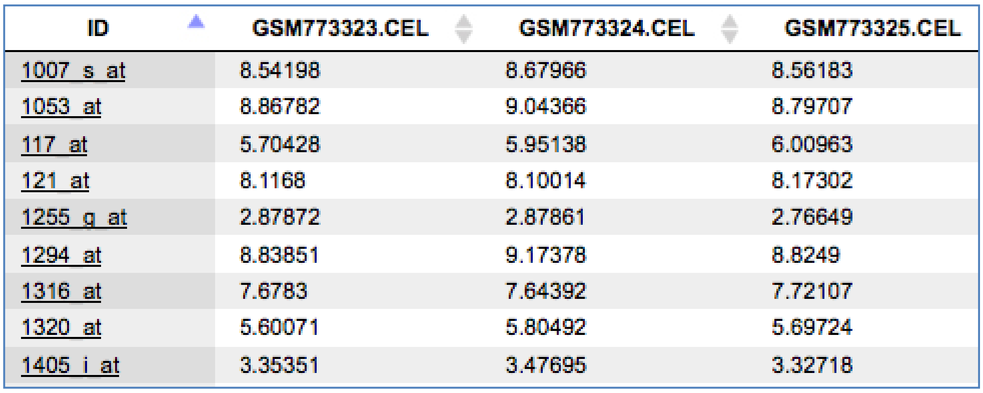


Gene set enrichment analysis (Affymetrix probes)

Experiment normalized	(select element)
Control normalized	(select element)
Species	Human (Homo sapiens)
Results folder	(select element)

Step 2. Input the **Experiment normalized** and **Control normalized** tables from the tree. You can either drag-and-drop or click on the select element box to specify the tables in the Tree area. Here, the tables from the Example folder/ HCV infection in liver GSE31193, Affymetrix U133 Plus 2.0 are used. We aim to find out the enriched functional categories of gene expression upon treatment with IFN type III after 24 hours versus non-treated cells.

It is important to note that for this workflow, the input tables should have Affymetrix probeset IDs in the ID column. Such tables have an (AF) icon in the tree area and look like:



ID	GSM773323.CEL	GSM773324.CEL	GSM773325.CEL
1007_s_at	8.54198	8.67966	8.56183
1053_at	8.86782	9.04366	8.79707
117_at	5.70428	5.95138	6.00963
121_at	8.1168	8.10014	8.17302
1255_g_at	2.87872	2.87861	2.76649
1294_at	8.83851	9.17378	8.8249
1316_at	7.6783	7.64392	7.72107
1320_at	5.60071	5.80492	5.69724
1405_i_at	3.35351	3.47695	3.32718

You can see Affymetrix probeset IDs in the **ID** column, and several columns with the normalized values; each column corresponds to one CEL file.

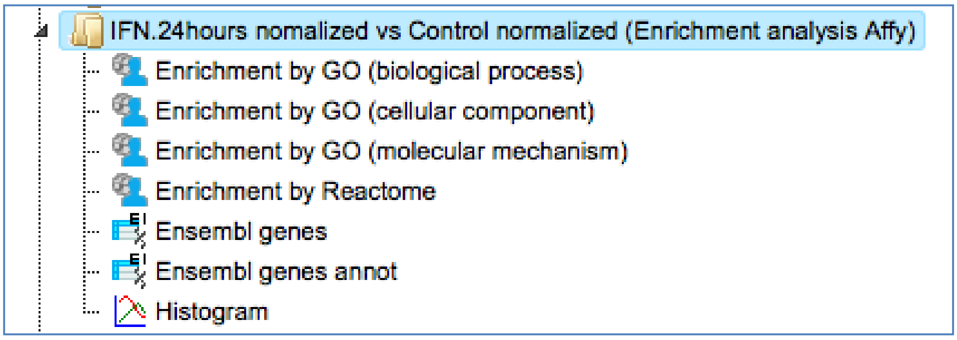
Step 3. Choose human, mouse, or rat **species** from the drop-down menu.

Step 4. Specify location and name of the **Results folder**. Important: the results folder should be located in your *Project* in the tree.

Step 5. Press the button [Run workflow] and wait till the workflow is completed.

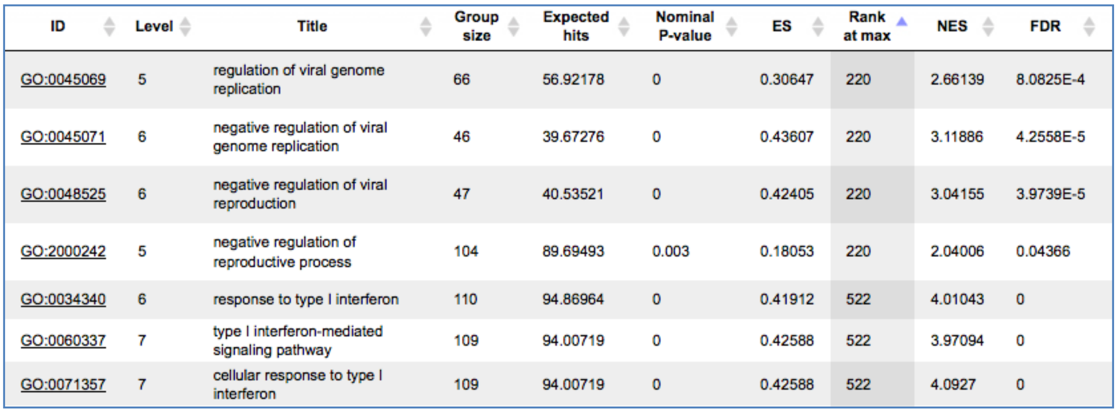
Results

The results folder contains four tables with the results of the enrichment analysis ( ) divided by the three branches of Gene Ontology, biological process, molecular function and cellular component, as well as by the Reactome pathways.



- IFN.24hours nomalized vs Control normalized (Enrichment analysis Affy)
  - Enrichment by GO (biological process)
  - Enrichment by GO (cellular component)
  - Enrichment by GO (molecular mechanism)
  - Enrichment by Reactome
  - Ensembl genes
  - Ensembl genes annot
  - Histogram

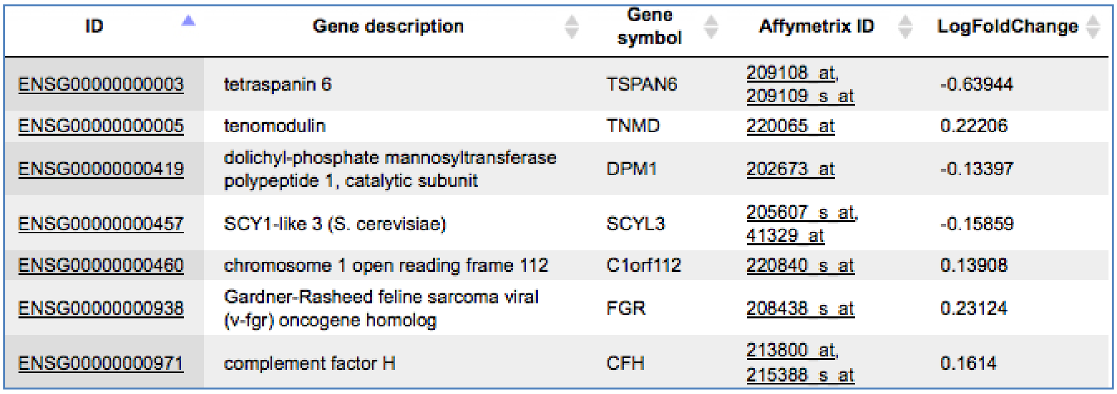
The tables with the enriched categories look like:



ID	Level	Title	Group size	Expected hits	Nominal P-value	ES	Rank at max	NES	FDR
GO:0045069	5	regulation of viral genome replication	66	56.92178	0	0.30647	220	2.66139	8.0825E-4
GO:0045071	6	negative regulation of viral genome replication	46	39.67276	0	0.43607	220	3.11886	4.2558E-5
GO:0048525	6	negative regulation of viral reproduction	47	40.53521	0	0.42405	220	3.04155	3.9739E-5
GO:2000242	5	negative regulation of reproductive process	104	89.69493	0.003	0.18053	220	2.04006	0.04366
GO:0034340	6	response to type I interferon	110	94.86964	0	0.41912	522	4.01043	0
GO:0060337	7	type I interferon-mediated signaling pathway	109	94.00719	0	0.42588	522	3.97094	0
GO:0071357	7	cellular response to type I interferon	109	94.00719	0	0.42588	522	4.0927	0

The GSEA results are described in details in a separate section below.

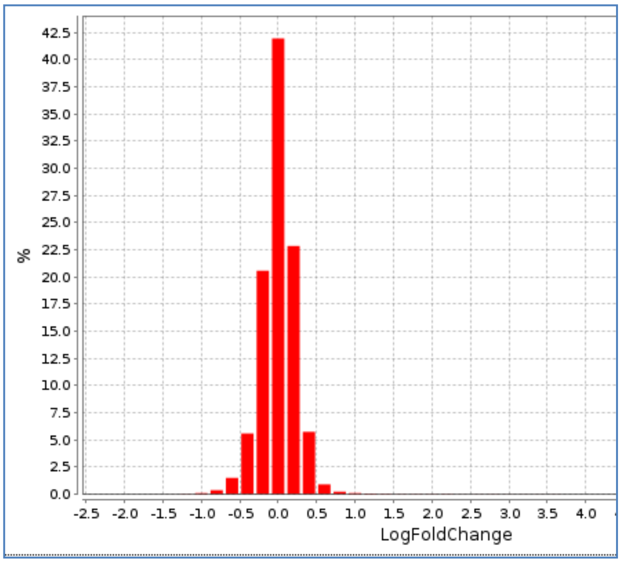
The table *Ensembl genes annot* contains Ensembl genes as a result of Affymetrix IDs convertion into Ensembl gene IDs:



ID	Gene description	Gene symbol	Affymetrix ID	LogFoldChange
ENSG00000000003	tetraspanin 6	TSPAN6	209108_at, 209109_s_at	-0.63944
ENSG00000000005	tenomodulin	TNMD	220065_at	0.22206
ENSG00000000419	dolichyl-phosphate mannosyltransferase polypeptide 1, catalytic subunit	DPM1	202673_at	-0.13397
ENSG00000000457	SCY1-like 3 (S. cerevisiae)	SCYL3	205607_s_at, 41329_at	-0.15859
ENSG00000000460	chromosome 1 open reading frame 112	C1orf112	220840_s_at	0.13908
ENSG00000000938	Gardner-Rasheed feline sarcoma viral (v-fgr) oncogene homolog	FGR	208438_s_at	0.23124
ENSG00000000971	complement factor H	CFH	213800_at, 215388_s_at	0.1614

For each gene, gene symbol, gene description, and Affymetrix probeset ID are shown. Additionally, the LogFoldChange value is calculated for each gene.

The distribution of LogFoldChange values is shown in the Histogram:

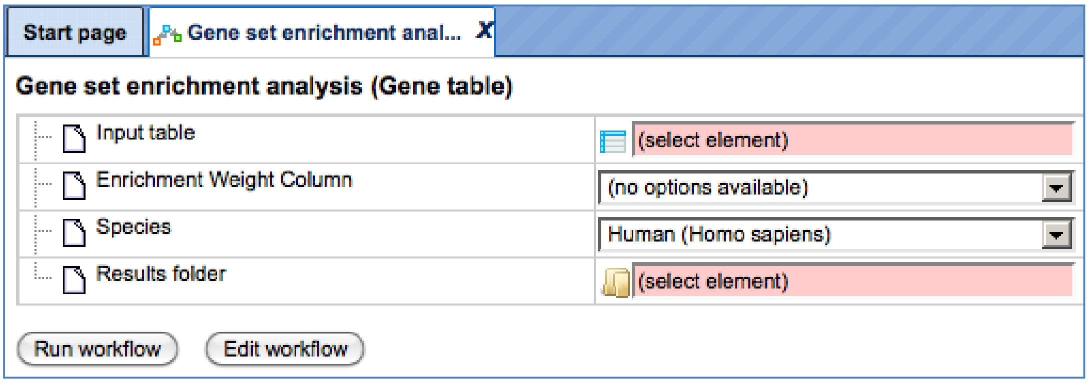


GSEA by GO categories and metabolic pathways for a single gene table

This workflow performs the GSEA divided by the three branches of Gene Ontology, biological process, molecular function and cellular component, as well as by the Reactome pathways, for any input gene or protein table. It is important to note that such a table should have a column which can be used as a weight column for enrichment, e.g. expression value.

To launch the workflow, follow these steps:

Step1. Open the workflow input form from the Start page. It opens in the main Work Space and looks as shown below:



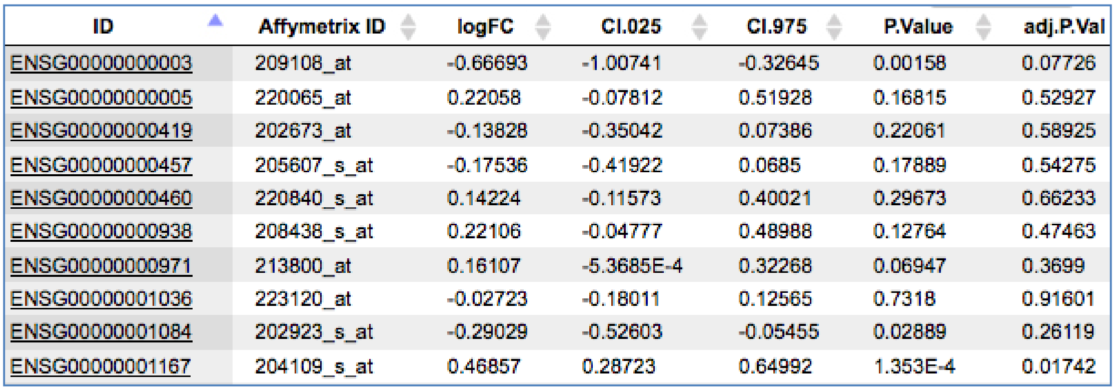
Gene set enrichment analysis (Gene table)

Input table	(select element)
Enrichment Weight Column	(no options available)
Species	Human (Homo sapiens)
Results folder	(select element)

Step 2. Input a gene table with FoldChange (LogFoldChange) calculated. You can either drag-and-drop or click on the select element box to specify the table in the Tree area. Here, the table from the Example folder/HCV infection in liver GSE31193, Affymetrix U133 Plus 2.0 is used. We aim to find out the enriched functional categories of gene expression upon treatment with IFN type III after 24 hours versus non-treated cells.

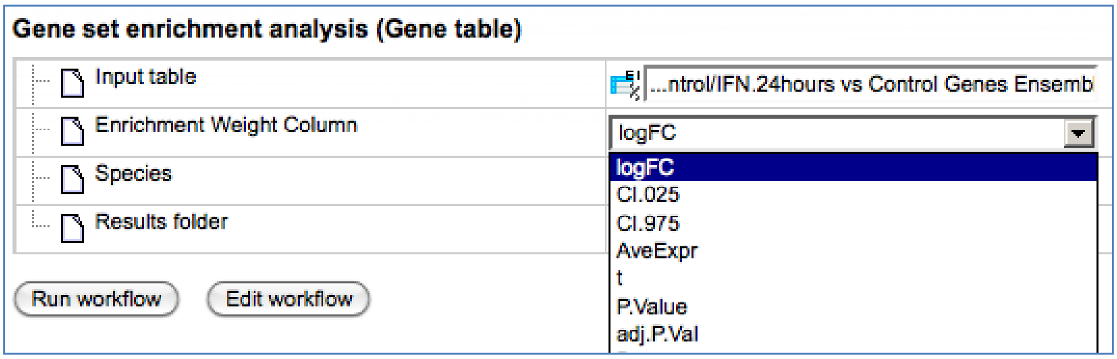
The input table may look like the one shown below. This table contains the column logFC (LogFoldChange). This table is an output of the Limma method (Section 4.2.1).

For best GSEA results, input the table with all genes analyzed, e.g. all genes present on the chip in the microarray experiment.



ID	Affymetrix ID	logFC	CI.025	CI.975	P.Value	adj.P.Val
ENSG00000000003	209108_at	-0.66693	-1.00741	-0.32645	0.00158	0.07726
ENSG00000000005	220065_at	0.22058	-0.07812	0.51928	0.16815	0.52927
ENSG00000000419	202673_at	-0.13828	-0.35042	0.07386	0.22061	0.58925
ENSG00000000457	205607_s_at	-0.17536	-0.41922	0.0685	0.17889	0.54275
ENSG00000000460	220840_s_at	0.14224	-0.11573	0.40021	0.29673	0.66233
ENSG00000000938	208438_s_at	0.22106	-0.04777	0.48988	0.12764	0.47463
ENSG00000000971	213800_at	0.16107	-5.3685E-4	0.32268	0.06947	0.3699
ENSG00000001036	223120_at	-0.02723	-0.18011	0.12565	0.7318	0.91601
ENSG00000001084	202923_s_at	-0.29029	-0.52603	-0.05455	0.02889	0.26119
ENSG00000001167	204109_s_at	0.46857	0.28723	0.64992	1.353E-4	0.01742

Step 3. As soon as you specified the input table, the drop-down menu in the field **Enrichment Weight Column** becomes active. It presents all numerical columns in the input table. Select which column should be used for enrichment calculations. Here, the column *logFC* is selected.



Gene set enrichment analysis (Gene table)

Input table: ...ntrol/IFN.24hours vs Control Genes Ensembl

Enrichment Weight Column: logFC

Species

Results folder

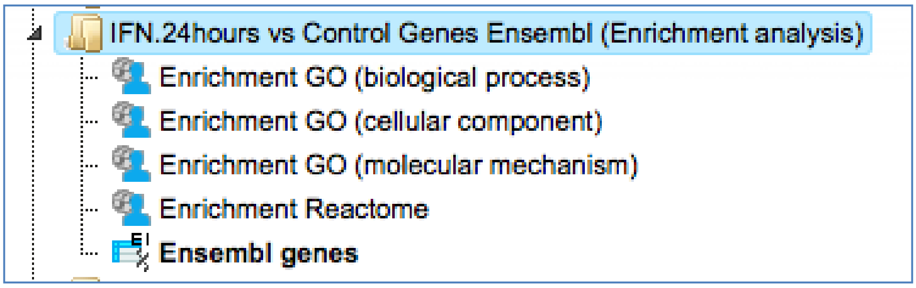
Step 4. Choose human, mouse, or rat **species** from the drop-down menu.

Step 5. Specify location and name of the **Results folder**. It is important to note that the results folder should be located in your *Project* in the tree.

Step 5. Press the button [Run workflow] and wait till the workflow is completed.

Results

The results folder contains four tables with the results of the enrichment analysis () corresponding to the three branches of Gene Ontology, biological process, molecular function and cellular component as well as by the Reactome pathways.



IFN.24hours vs Control Genes Ensembl (Enrichment analysis)
- Enrichment GO (biological process)
- Enrichment GO (cellular component)
- Enrichment GO (molecular mechanism)
- Enrichment Reactome
- Ensembl genes

The GSEA results are described in detail in a separate section below.

10.3.1.2. GSEA by GO categories, signaling pathways and diseases

Also this type of gene set enrichment analysis (GSEA) can be done by either starting from raw data of any of the widely used experimental platforms (**Affymetrix, Agilent, or Illumina**), or from a **single gene table** that you may have composed yourself.

Affymetrix, Agilent and Illumina microarrays

The three workflows under this category are similar to those described under 10.3.1.1, 1st part, requiring exactly the same two normalized input tables, and the same steps to launch these workflows. The difference is in the ontologies applied. The three workflows under this category are designed to perform a GSEA utilizing the three branches of the PROTEOME™-curated gene ontology, PROTEOME™ biological process, PROTEOME™ molecular function and PROTEOME™ cellular component as well as the TRANSPATH® pathways:

- *Gene Set Enrichment Analysis PROTEOME (Affymetrix probes),*
- *Gene Set Enrichment Analysis PROTEOME (Agilent probes),*
- *Gene Set Enrichment Analysis PROTEOME (Illumina probes).*

The GSEA results are described in detail in a separate section below.

> **Note.** This workflow is available together with a valid PROTEOME™/TRANSPATH® license. Please feel free to ask for details (info@genexplain.com).

Single gene table

This workflow is similar to the one described under 10.3.1.1, 2nd part, it requires exactly the same format of the input table, and the steps to launch this workflow are the same. The difference is in the ontologies applied. This workflow is designed to perform a GSEA utilizing the three branches of the PROTEOME™-curated gene ontology, PROTEOME™ biological process, PROTEOME™ molecular function and PROTEOME™ cellular component, as well as the TRANSPATH® pathways.

The GSEA results are described in detail in a separate section below.

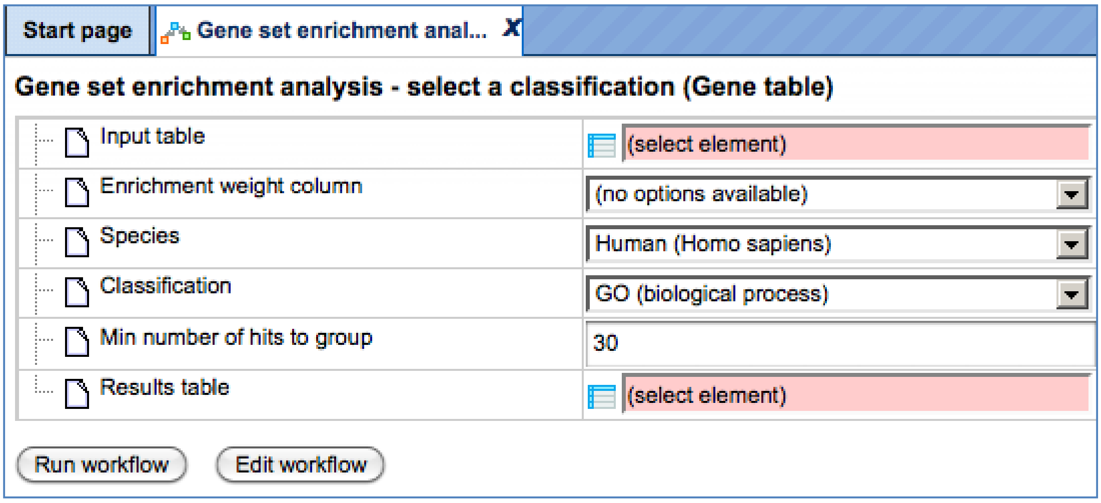
> **Note.** This workflow is available together with a valid PROTEOME™/TRANSPATH® license. Please feel free to ask for details (info@genexplain.com).

10.3.1.3. GSEA with a selected ontology

This workflow performs a GSEA with one selected ontology for an input gene or protein table. It is important to note that such a table should have a numerical column which can be used as a weight column for enrichment, e.g. expression value or fold change.

To launch the workflow, follow these steps:

Step1. Open the workflow input form from the Start page. It opens in the main Work Space and looks as shown below:



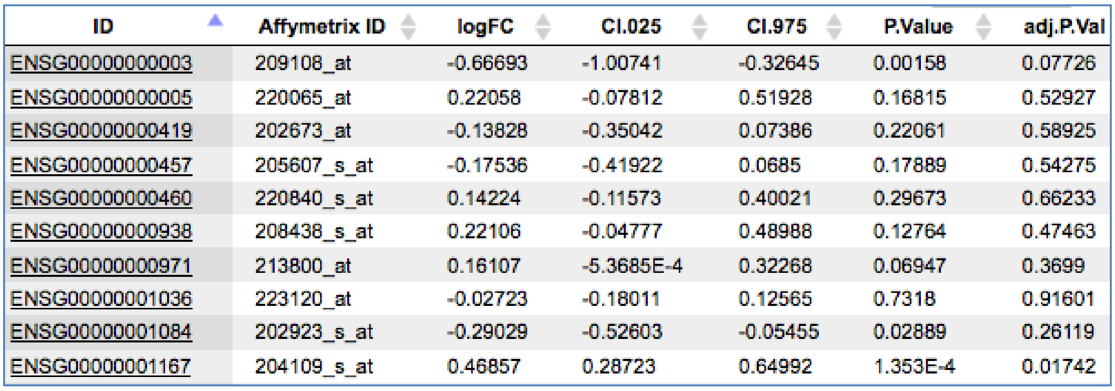
Gene set enrichment analysis - select a classification (Gene table)

Input table	(select element)
Enrichment weight column	(no options available)
Species	Human (Homo sapiens)
Classification	GO (biological process)
Min number of hits to group	30
Results table	(select element)

Step 2. Input a gene table with FoldChange (LogFoldChange) calculated. You can either drag-and-drop or click on the select element box to specify the table in the Tree area. Here, the table from the Example folder/ HCV infection in liver GSE31193, Affymetrix U133 Plus 2.0 is used. We aim to find out the enriched functional categories of gene expression upon treatment with IFN type III after 24 hours versus non-treated cells.

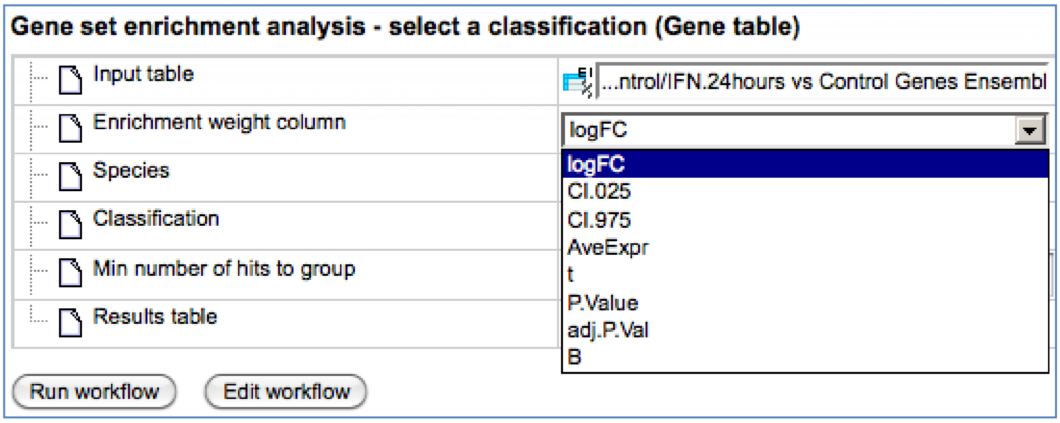
The input table may look like the one shown below. This table contains the column logFC (LogFoldChange). This table is an output of the Limma method (Section 4.2.1).

For the best GSEA results, input the table with all genes analyzed, e.g. all genes present on the chip in the microarray experiment.



ID	Affymetrix ID	logFC	CI.025	CI.975	P.Value	adj.P.Val
ENSG00000000003	209108_at	-0.66693	-1.00741	-0.32645	0.00158	0.07726
ENSG00000000005	220065_at	0.22058	-0.07812	0.51928	0.16815	0.52927
ENSG00000000419	202673_at	-0.13828	-0.35042	0.07386	0.22061	0.58925
ENSG00000000457	205607_s_at	-0.17536	-0.41922	0.0685	0.17889	0.54275
ENSG00000000460	220840_s_at	0.14224	-0.11573	0.40021	0.29673	0.66233
ENSG00000000938	208438_s_at	0.22106	-0.04777	0.48988	0.12764	0.47463
ENSG00000000971	213800_at	0.16107	-5.3685E-4	0.32268	0.06947	0.3699
ENSG00000001036	223120_at	-0.02723	-0.18011	0.12565	0.7318	0.91601
ENSG00000001084	202923_s_at	-0.29029	-0.52603	-0.05455	0.02889	0.26119
ENSG00000001167	204109_s_at	0.46857	0.28723	0.64992	1.353E-4	0.01742

Step 3. As soon as you specified the input table, the drop-down menu in the field **Enrichment Weight Column** becomes active. It presents all numerical columns in the input table. Select which column should be used for enrichment calculations. Here, the column *logFC* is selected.



Gene set enrichment analysis - select a classification (Gene table)

Input table: ...ntrol/IFN.24hours vs Control Genes Ensembl

Enrichment weight column: logFC

Species

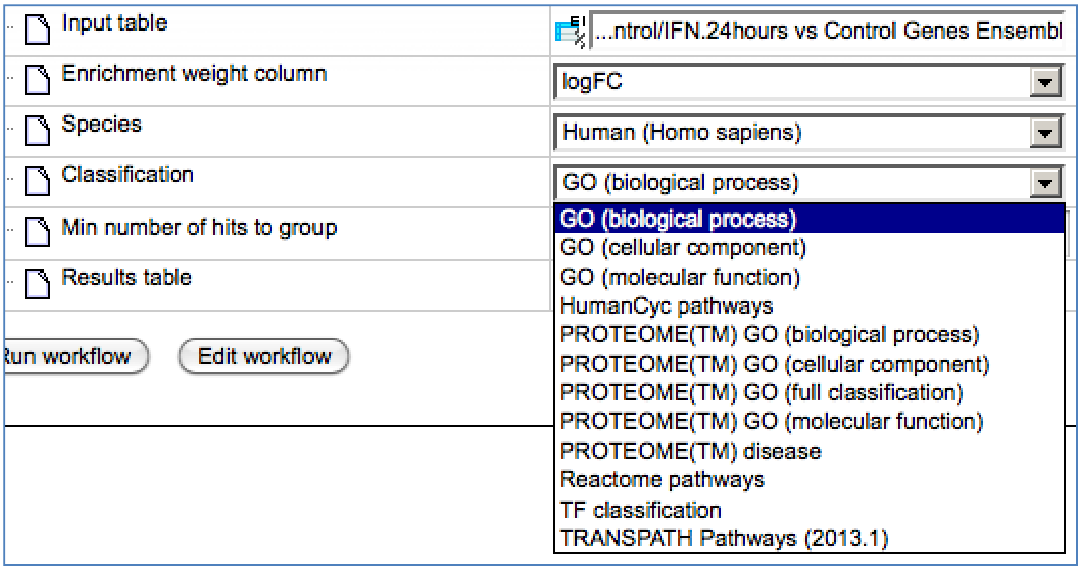
Classification

Min number of hits to group

Results table

Step 4. Choose human, mouse, or rat **species** from the drop-down menu.

Step 5. In the field **Classification**, choose the ontology from the drop-down menu. Here, GO biological process is selected.



Input table: ...ntrol/IFN.24hours vs Control Genes Ensembl

Enrichment weight column: logFC

Species: Human (Homo sapiens)

Classification: GO (biological process)

Min number of hits to group

Results table

GO (biological process)
GO (cellular component)
GO (molecular function)
HumanCyc pathways
PROTEOME(TM) GO (biological process)
PROTEOME(TM) GO (cellular component)
PROTEOME(TM) GO (full classification)
PROTEOME(TM) GO (molecular function)
PROTEOME(TM) disease
Reactome pathways
TF classification
TRANSPATH Pathways (2013.1)
Run workflow
Edit workflow

Step 6. Define a minimal number of hits in one group which you would like to consider in the field **Min number of hits to group**. By default it is 30.

Step 7. Specify location and name of the **Results folder**. Please note that the results folder should be located in your *Project* in the tree.

Step 5. Press the button [Run workflow] and wait till the workflow is completed.

Results

The results folder contains one table with the results of the enrichment analysis ( ) by the selected ontology.

The results are described in detail in a separate section below.

10.3.2. Functional classification

An alternative approach to GSEA is another group of workflows, Functional classification, which comprises several "Mapping to ontologies" workflows. The difference of this option to GSEA is that no enrichment of the categories is calculated, but that all genes in the list are mapped to GO categories or other ontologies. For example, you can use tables with pre-calculated up-regulated or down-regulated genes, as they are obtained as output of the workflow "Detect differentially expressed genes", and use these as input into the workflows „Mapping to ontologies".

The output tabulates which and how many genes from your list ("hits") fall into which category, how many known genes are in this category, how many hits would have been expected by chance, and what the P-value for the found number of hits being obtained by chance is.

The difference between the workflows within this group is in the ontologies applied as well as in the number of input tables.

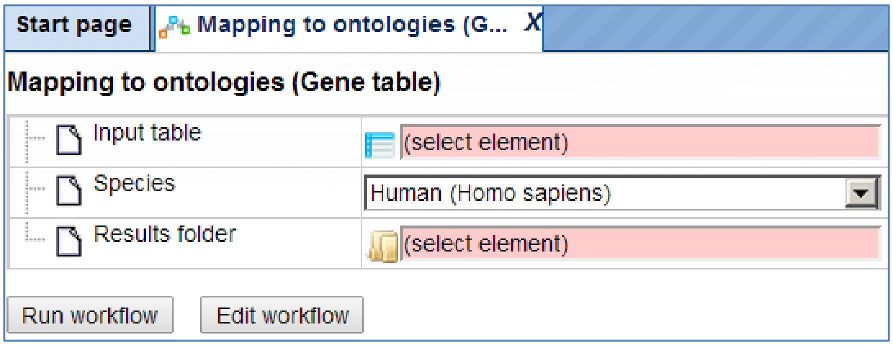
10.3.2.1. Mapping to GO categories and metabolic pathways

Single gene or protein table

This workflow is designed to classify an input gene set based on several ontologies, and to identify terms hits for which are overrepresented in the input set. The input file can be any gene or protein table. There is only one obligatory column, the column with gene or protein IDs; all other columns are optional.
In the first step, the input table is converted into a table with Ensembl Gene IDs. This table with Ensembl Gene IDs is subjected to a functional classification.

To launch the workflow, follow these steps:

Step 1. Open the workflow input form from the Start page. It looks as shown below:



Mapping to ontologies (Gene table)

Input table	(select element)
Species	Human (Homo sapiens)
Results folder	(select element)

Step 2. Specify the input table. The input gene set might be a list of differentially regulated genes or any gene or protein list of interest. You can drag it from your project within the tree area and drop it in the pink box of the field **Input table**. Alternatively, you may click on the pink field "select element" and a new window will be opened, where you can select the input gene set as shown below.

The further steps of the workflow are demonstrated for the genes shown to be up-regulated in one of the pre-prepared examples. The pertinent example file can be found in the geneXplain platform under:

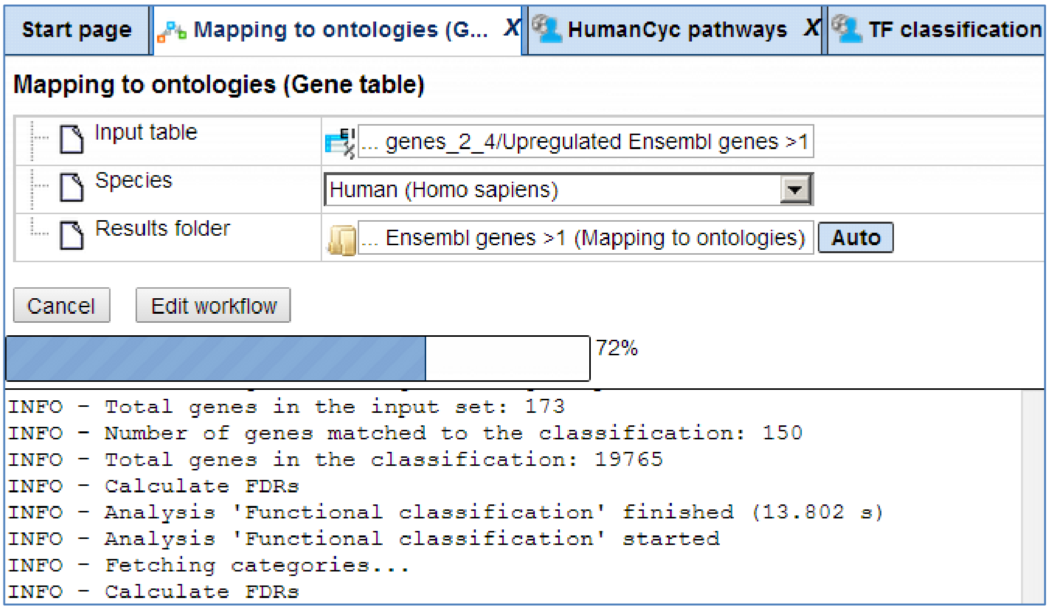
data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/detect deferentially expressed genes

Step 3. Specify the biological species of the input set in the field **Species** by selecting the required biological species from the drop-down menu.

Step 4. Define where the folder with the results should be located in the tree. You can do so by clicking on the pink field *select element* in the field **Results folder**, and a new window will be opened, where you can select the location of the results folder and define its name.

Step 5. Press the [Run workflow] button.

The workflow is running as shown below, wait till it is completed.



Mapping to ontologies (Gene table)

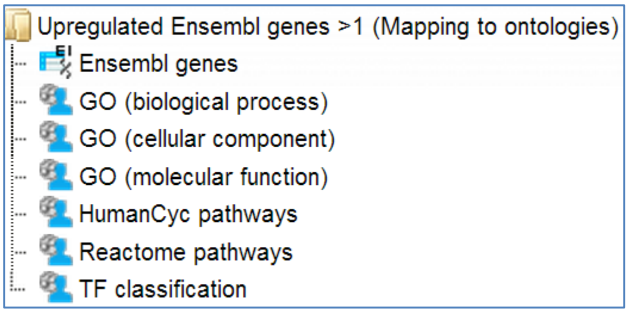
Input table: ... genes_2_4/Upregulated Ensembl genes >1

Species: Human (Homo sapiens)

Results folder: ... Ensembl genes >1 (Mapping to ontologies)

```
INFO - Total genes in the input set: 173
INFO - Number of genes matched to the classification: 150
INFO - Total genes in the classification: 19765
INFO - Calculate FDRs
INFO - Analysis 'Functional classification' finished (13.802 s)
INFO - Analysis 'Functional classification' started
INFO - Fetching categories...
INFO - Calculate FDRs
```

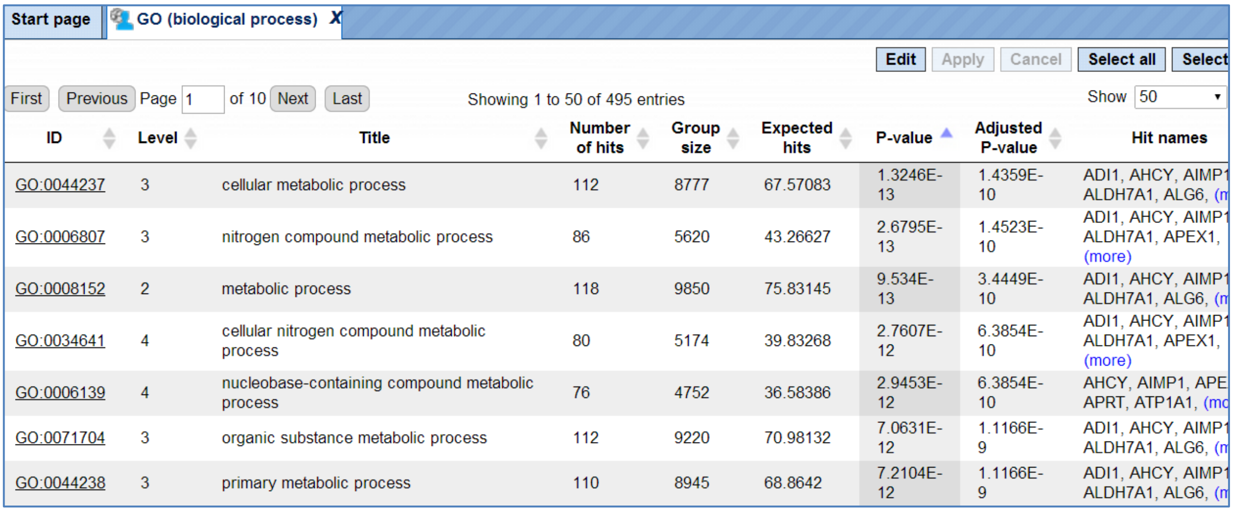
The results folder contains several tables with the resulting mapping, one table each for the applied ontological groups ( ), as well as one gene table ( ) as shown below.


Upregulated Ensembl genes >1 (Mapping to ontologies)
- Ensembl genes
- GO (biological process)
- GO (cellular component)
- GO (molecular function)
- HumanCyc pathways
- Reactome pathways
- TF classification

When the workflow is completed, all output tables are opened by default.

Let’s consider the output tables.

Mapping to the three GO branches, biological processes, cellular components, and molecular functions ( ). The tables with the enriched categories look like:



Showing 1 to 50 of 495 entries

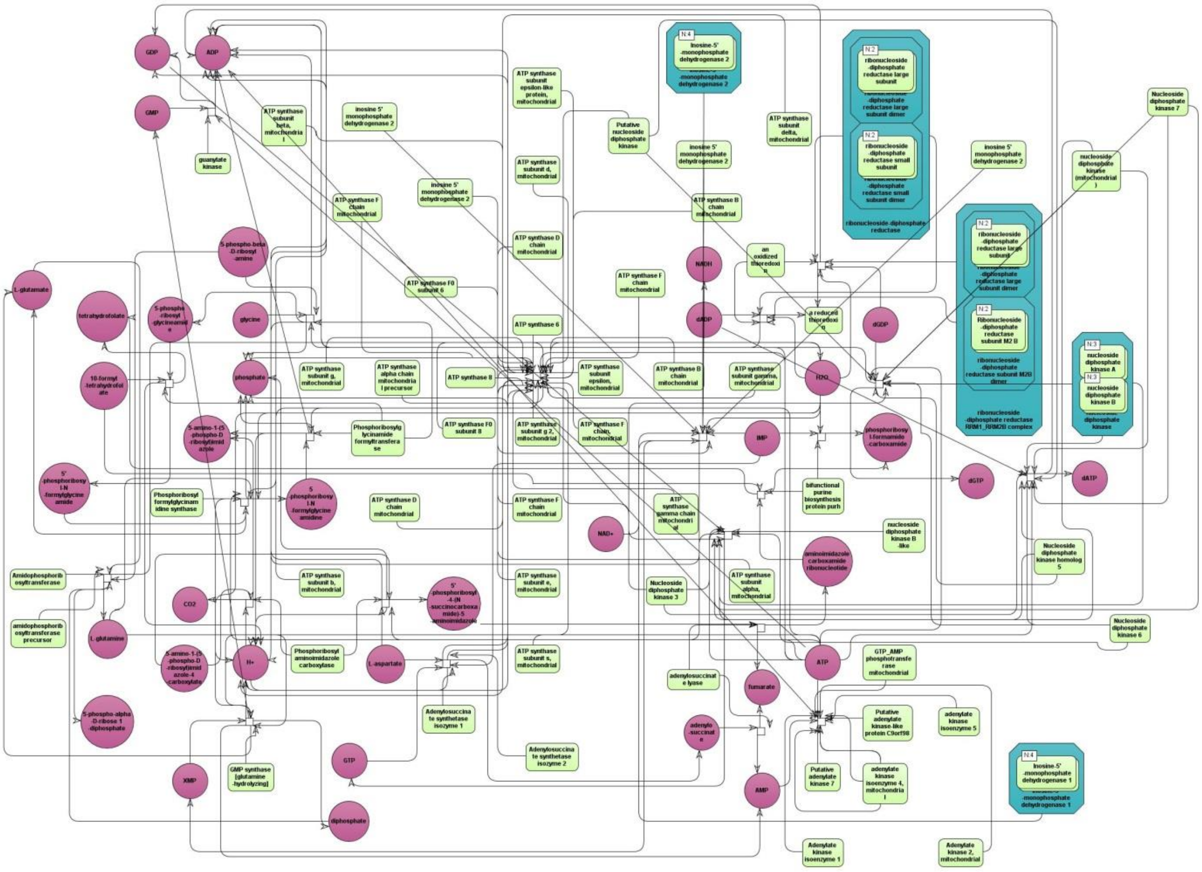
ID	Level	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
GO:0044237	3	cellular metabolic process	112	8777	67.57083	1.3246E-13	1.4359E-10	ADI1, AHCY, AIMP1 ALDH7A1, ALG6, (m
GO:0006807	3	nitrogen compound metabolic process	86	5620	43.26627	2.6795E-13	1.4523E-10	ADI1, AHCY, AIMP1 ALDH7A1, APEX1, (more)
GO:0008152	2	metabolic process	118	9850	75.83145	9.534E-13	3.4449E-10	ADI1, AHCY, AIMP1 ALDH7A1, ALG6, (m
GO:0034641	4	cellular nitrogen compound metabolic process	80	5174	39.83268	2.7607E-12	6.3854E-10	ADI1, AHCY, AIMP1 ALDH7A1, APEX1, (more)
GO:0006139	4	nucleobase-containing compound metabolic process	76	4752	36.58386	2.9453E-12	6.3854E-10	AHCY, AIMP1, APE APRT, ATP1A1, (mo
GO:0071704	3	organic substance metabolic process	112	9220	70.98132	7.0631E-12	1.1166E-9	ADI1, AHCY, AIMP1 ALDH7A1, ALG6, (m
GO:0044238	3	primary metabolic process	110	8945	68.8642	7.2104E-12	1.1166E-9	ADI1, AHCY, AIMP1 ALDH7A1, ALG6, (m

Each row presents details about one ontological term. The column **ID** comprises the identifier of the ontological category, here identifiers of Gene Ontology biological process terms. These identifiers are hyperlinked to the page http://www.ebi.ac.uk/QuickGO/ where you can get further information about this ontological term.

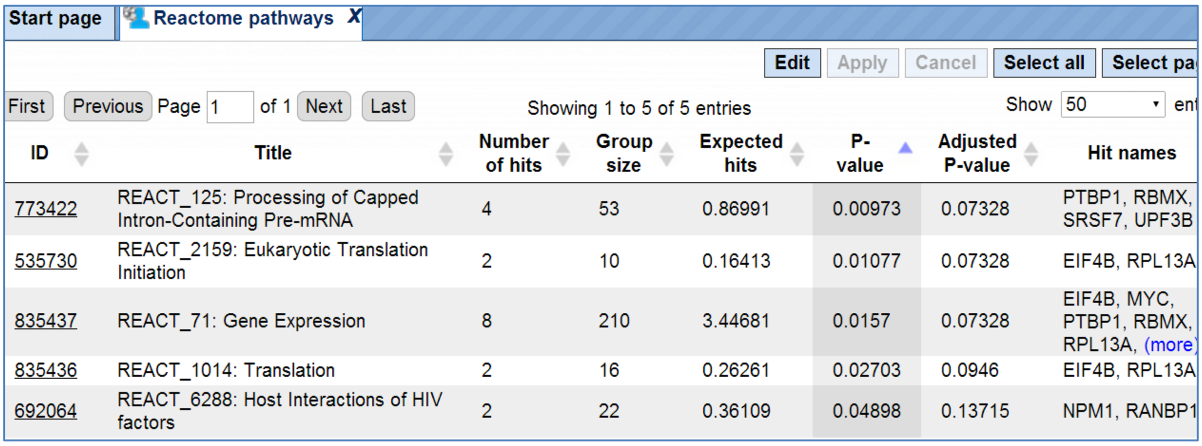
The columns **Title** and **Group size** contain further details about the ontological terms, its title and the number of genes linked to this term in the corresponding database, here in GO. The column **Expected hits** shows the number of genes expected to fall into this category by random chance, based on the size of the input set and the size of the category. The column **Number of hits** shows how many genes from the input table fall into this category. **P-value** and **Adjusted p-value** are calculated for the difference between expected and real numbers of hits. The genes mapped to each category are explicitly listed in the column **Hit names**. As the lists can get quite long, only a few genes are shown by default in each row. To get the full list, press [more].

Tip The hits for one or several selected rows can be saved as a separate gene table. This can be done with the button *Save hits* in the top control menu. Such genes tables can be analyzed further, e.g. to find master regulatory molecules in the networks, and to identify transcription factors that might commonly regulate these genes.

The table HumanCyc pathways ( ). In the column **ID** the identifiers of the HumanCyc pathways are given. Upon a mouse click, a diagram of the corresponding metabolic pathway opens in the workspace:


GDP
ADP
GMP
guanylate kinase
ATP synthase subunit beta, mitochondrial
inosine 5' monophosphate dehydrogenase 2
inosine 5' monophosphate dehydrogenase 2
ATP synthase F chain mitochondrial
ATP synthase subunit epsilon-like protein, mitochondrial
ATP synthase subunit d, mitochondrial
ATP synthase D chain mitochondrial
ATP synthase F0 subunit 6
ATP synthase 6
5-phospho-beta-D-ribosyl-amine
L-glutamate
tetrahydrofolate
5-phospho-ribosyl-glycineamide
glycine
10-formyl-tetrahydrofolate
phosphate
ATP synthase subunit g, mitochondrial
ATP synthase alpha chain mitochondrial precursor
ATP synthase 8
ATP synthase subunit epsilon, mitochondrial
5-amino-1-(5-phospho-D-ribosyl)imidazole
Phosphoribosylglycinamide formyltransferase
ATP synthase F0 subunit 8
ATP synthase subunit g 2, mitochondrial
ATP synthase F chain, mitochondrial
5'-phosphoribosyl-N-formylglycineamide
Phosphoribosylformylglycinamidine synthase
5-phosphoribosyl-N-formylglycineamidine
ATP synthase D chain mitochondrial
ATP synthase F chain mitochondrial
NAD+
Amidophosphoribosyltransferase
amidophosphoribosyltransferase precursor
CO2
L-glutamine
ATP synthase subunit b, mitochondrial
5'-phosphoribosyl-4-(N-succinocarboxamide)-5-aminoimidazole
ATP synthase subunit e, mitochondrial
5-amino-1-(5-phospho-D-ribosyl)imidazole-4-carboxylate
H+
Phosphoribosylaminoimidazole carboxylase
L-aspartate
ATP synthase subunit s, mitochondrial
5-phospho-alpha-D-ribose 1-diphosphate
Adenylosuccinate synthetase isozyme 1
Adenylosuccinate synthetase isozyme 2
GTP
XMP
GMP synthase [glutamine-hydrolyzing]
diphosphate
N:4
Inosine-5'-monophosphate dehydrogenase 2
Inosine-5'-monophosphate dehydrogenase 2
Putative nucleoside diphosphate kinase
inosine 5' monophosphate dehydrogenase 2
ATP synthase subunit delta, mitochondrial
N:2
ribonucleoside-diphosphate reductase large subunit
ribonucleoside-diphosphate reductase large subunit dimer
N:2
ribonucleoside-diphosphate reductase small subunit
ribonucleoside-diphosphate reductase small subunit dimer
ribonucleoside-diphosphate reductase
inosine 5' monophosphate dehydrogenase 2
Nucleoside diphosphate kinase 7
nucleoside diphosphate kinase (mitochondrial)
ATP synthase B chain mitochondrial
NADH
an oxidized thioredoxin
ATP synthase F chain mitochondrial
dADP
a reduced thioredoxin
dGDP
N:2
ribonucleoside-diphosphate reductase large subunit
ribonucleoside-diphosphate reductase large subunit dimer
N:2
Ribonucleoside-diphosphate reductase subunit M2 B
ribonucleoside-diphosphate reductase subunit M2B dimer
ribonucleoside-diphosphate reductase RRM1_RRM2B complex
N:3
nucleoside diphosphate kinase A
N:3
nucleoside diphosphate kinase B
nucleoside diphosphate kinase
ATP synthase subunit epsilon, mitochondrial
ATP synthase B chain mitochondrial
ATP synthase subunit gamma, mitochondrial
H2O
IMP
phosphoribosyl-formamido-carboxamide
dGTP
dATP
bifunctional purine biosynthesis protein purh
ATP synthase gamma chain mitochondrial
nucleoside diphosphate kinase B-like
aminoimidazole carboxamide ribonucleotide
Nucleoside diphosphate kinase homolog 5
Nucleoside diphosphate kinase 3
ATP synthase subunit alpha, mitochondrial
Nucleoside diphosphate kinase 6
ATP
GTP_AMP phosphotransferase mitochondrial
adenylosuccinate lyase
fumarate
Putative adenylate kinase-like protein C9orf98
adenylate kinase isoenzyme 5
adenylo-succinate
N:4
Inosine-5'-monophosphate dehydrogenase 1
Inosine-5'-monophosphate dehydrogenase 1
AMP
Putative adenylate kinase 7
adenylate kinase isoenzyme 4, mitochondrial
Adenylate kinase isoenzyme 1
Adenylate kinase 2, mitochondrial

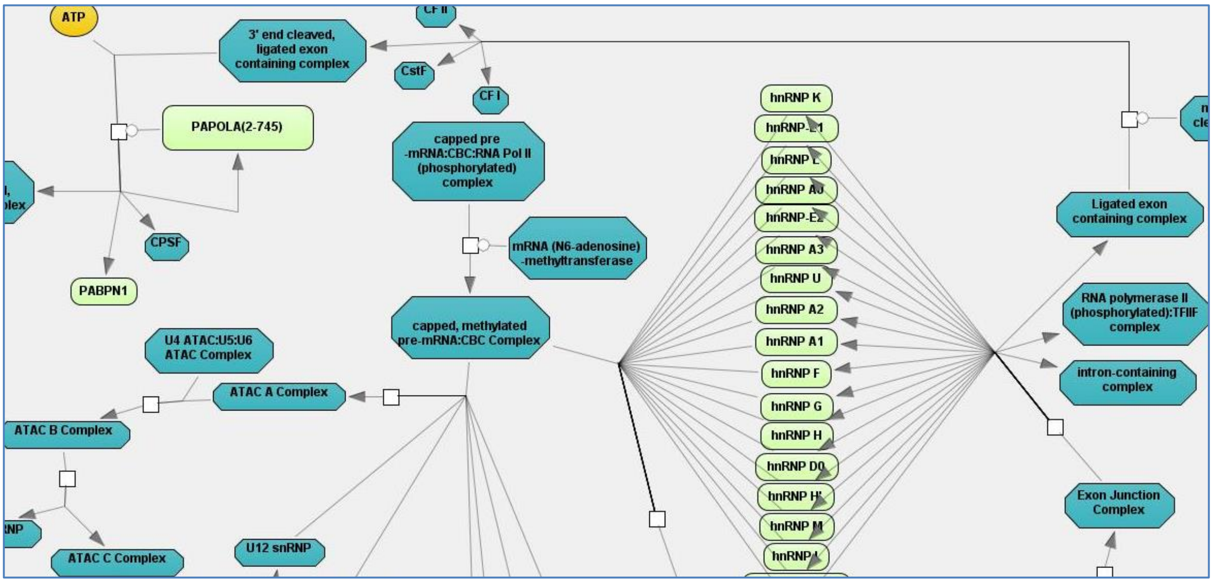
The table Reactome pathways ( ). In the column **ID** you can find the identifiers of the Reactome pathways.



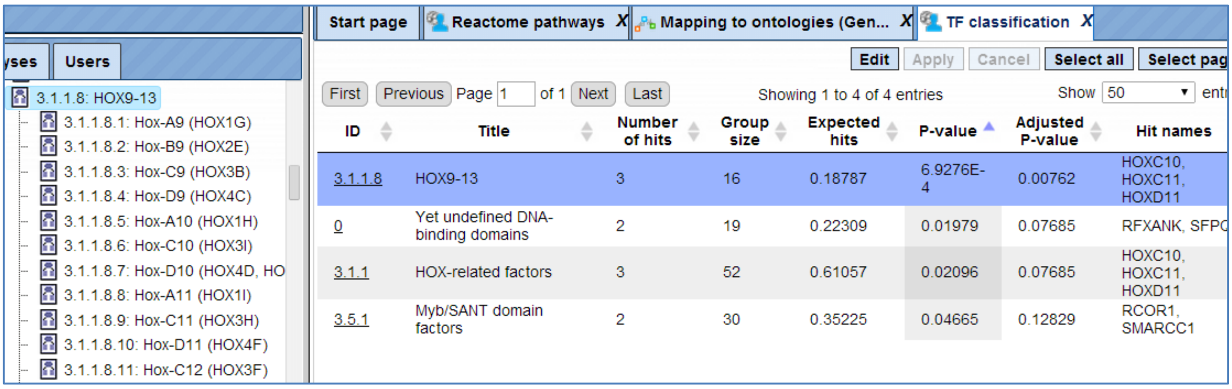
Showing 1 to 5 of 5 entries

ID	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
773422	REACT_125: Processing of Capped Intron-Containing Pre-mRNA	4	53	0.86991	0.00973	0.07328	PTBP1, RBMX, SRSF7, UPF3B
535730	REACT_2159: Eukaryotic Translation Initiation	2	10	0.16413	0.01077	0.07328	EIF4B, RPL13A
835437	REACT_71: Gene Expression	8	210	3.44681	0.0157	0.07328	EIF4B, MYC, PTBP1, RBMX, RPL13A, (more)
835436	REACT_1014: Translation	2	16	0.26261	0.02703	0.0946	EIF4B, RPL13A
692064	REACT_6288: Host Interactions of HIV factors	2	22	0.36109	0.04898	0.13715	NPM1, RANBP1

Upon a mouse click, a diagram of the corresponding pathway opens in the workspace.


ATP
3' end cleaved, ligated exon containing complex
CF II
CstF
CF I
PAPOLA(2-745)
capped pre -mRNA:CBC:RNA Pol II (phosphorylated) complex
CPSF
PABPN1
mRNA (N6-adenosine) -methyltransferase
capped, methylated pre-mRNA:CBC Complex
U4 ATAC:U5:U6 ATAC Complex
ATAC A Complex
ATAC B Complex
ATAC C Complex
U12 snRNP
hnRNP K
hnRNP-L1
hnRNP L
hnRNP A0
hnRNP-E2
hnRNP A3
hnRNP U
hnRNP A2
hnRNP A1
hnRNP F
hnRNP G
hnRNP H
hnRNP D0
hnRNP H'
hnRNP M
hnRNP A
Ligated exon containing complex
RNA polymerase II (phosphorylated):TFIIF complex
intron-containing complex
Exon Junction Complex

The table TF classification ( ). Your input table is mapped to the classification of Transcription factors (*Nucleic Acids Res. 41, D165-D170 (2013)*), which is also integrated in the platform. In the column **ID** the identifiers of the TF classification are shown. They are hyperlinked to the corresponding classification categories.



- 3.1.1.8: HOX9-13
  - 3.1.1.8.1: Hox-A9 (HOX1G)
  - 3.1.1.8.2: Hox-B9 (HOX2E)
  - 3.1.1.8.3: Hox-C9 (HOX3B)
  - 3.1.1.8.4: Hox-D9 (HOX4C)
  - 3.1.1.8.5: Hox-A10 (HOX1H)
  - 3.1.1.8.6: Hox-C10 (HOX3I)
  - 3.1.1.8.7: Hox-D10 (HOX4D, HO
  - 3.1.1.8.8: Hox-A11 (HOX1I)
  - 3.1.1.8.9: Hox-C11 (HOX3H)
  - 3.1.1.8.10: Hox-D11 (HOX4F)
  - 3.1.1.8.11: Hox-C12 (HOX3F)

Showing 1 to 4 of 4 entries

ID	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
3.1.1.8	HOX9-13	3	16	0.18787	6.9276E-4	0.00762	HOXC10, HOXC11, HOXD11
0	Yet undefined DNA-binding domains	2	19	0.22309	0.01979	0.07685	RFXANK, SFPQ
3.1.1	HOX-related factors	3	52	0.61057	0.02096	0.07685	HOXC10, HOXC11, HOXD11
3.5.1	Myb/SANT domain factors	2	30	0.35225	0.04665	0.12829	RCOR1, SMARCC1

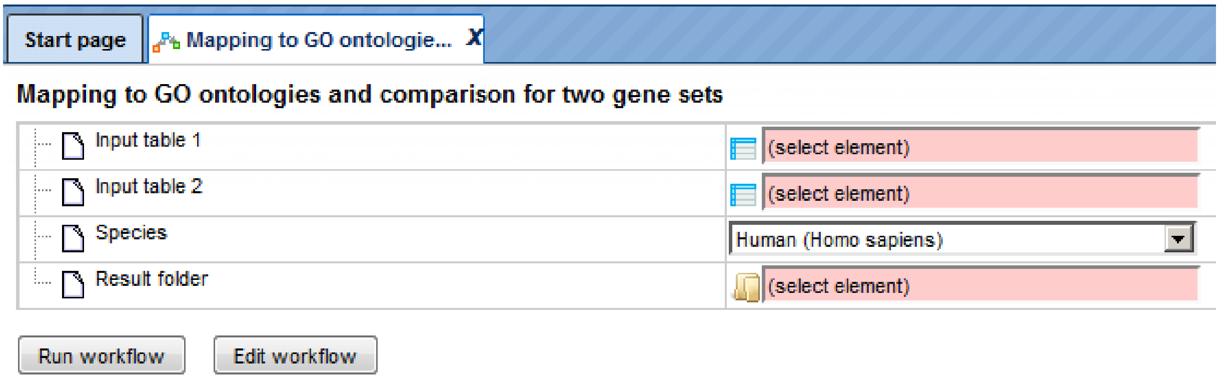
The table Ensembl genes ( ). The input gene or protein table is converted to a table with Ensembl gene IDs, and the result is shown in this table. For example, if your input was a table with UniProt IDs, it is converted into Ensembl gene IDs and included in the results folder of this workflow.

2 Gene sets and comparison

This workflow is designed to map two input tables to all **Gene Ontology** categories (*biological process*, *molecular function* and *cellular component*) to identify terms hits and to compare the results. While the workflow described above allows for selecting one particular ontology, this workflow runs three branches of the GO ontology in parallel, and the comparison between two gene sets is done regarding all three GO branches at once. The input files can be any gene or protein table. In the first step, the input tables are converted into two tables with Ensembl Gene IDs. These new tables are subjected to a **Gene Ontology** mapping. As result two mapped tables (for each GO category) are stored and further compared via P-values. This last comparison step is based on the method *analyses/Methods/Statistical analysis/Compare analysis results*, icon . Please refer to section 16.3 for details on this particular analysis method. The comparison can help to reveal items that show different enrichment across certain conditions.

To launch the workflow, follow these steps:

Step 1. Open the workflow input form from the Start page. It looks as shown below:



Mapping to GO ontologies and comparison for two gene sets

Input table 1	(select element)
Input table 2	(select element)
Species	Human (Homo sapiens)
Result folder	(select element)

Step 2. Specify the input tables 1 and 2. The input gene sets might be lists of differentially regulated genes or any gene or protein list of interest. You can drag it from your project within the tree area and drop it in the pink box of the field **Input table**. Alternatively, you may click on the pink field "select element" and a new window will be opened, where you can select the input gene set as shown below.

The further steps of the workflow are demonstrated for the genes shown to be up-regulated (Top100) and down-regulated (Top100) in one of the pre-prepared examples. The pertinent example file can be found in the geneXplain platform under:

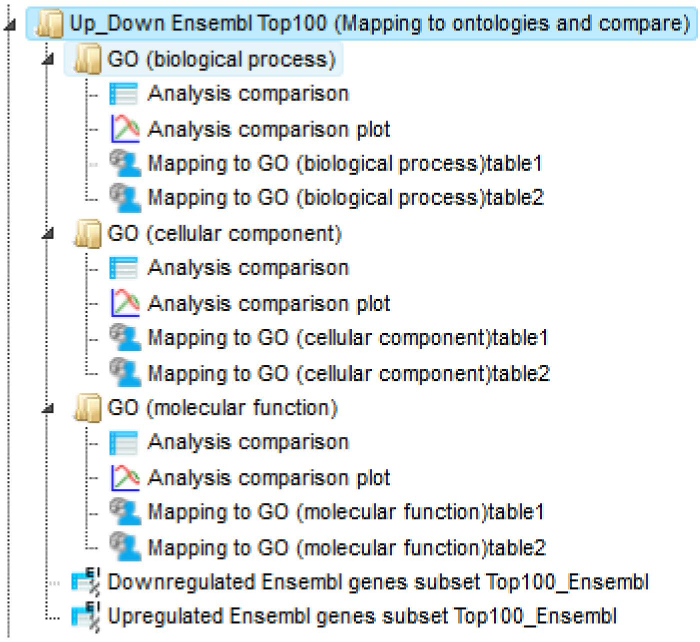
data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/Mapping to ontologies and compare/Up_Down Ensembl Top100 (Mapping to ontologies and compare)

Step 3. Specify the biological species of the input set in the field **Species** by selecting it from the drop-down menu.

Step 4. Define where the folder with the results should be located in the tree. You can do so by clicking on the pink field *select element* in the field **Result folder**, and a new window will be opened, where you can select the location of the result folder and define its name.

Step 5. Press the [Run workflow] button.

When the workflow is completed, the result folder is opened by default.



- Up_Down Ensembl Top100 (Mapping to ontologies and compare)
  - GO (biological process)
    - Analysis comparison
    - Analysis comparison plot
    - Mapping to GO (biological process)table1
    - Mapping to GO (biological process)table2
  - GO (cellular component)
    - Analysis comparison
    - Analysis comparison plot
    - Mapping to GO (cellular component)table1
    - Mapping to GO (cellular component)table2
  - GO (molecular function)
    - Analysis comparison
    - Analysis comparison plot
    - Mapping to GO (molecular function)table1
    - Mapping to GO (molecular function)table2
  - Downregulated Ensembl genes subset Top100_Ensembl
  - Upregulated Ensembl genes subset Top100_Ensembl

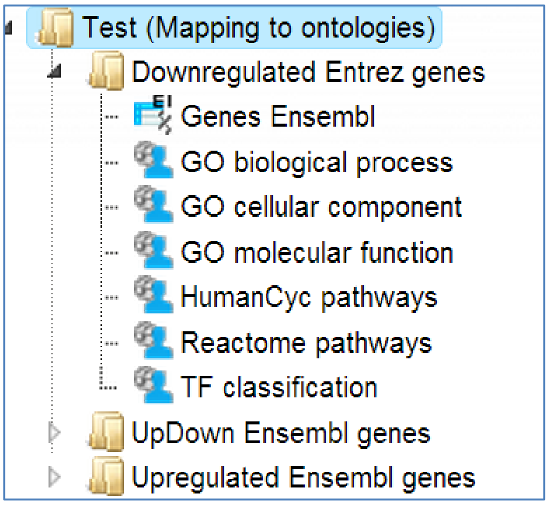
The result folder contains the three subfolders *GO* for all three GO categories applied in this workflow and two tables ( ). The two tables correspond to the input tables with the identifiers converted into Ensembl gene IDs. Each subfolder contains two tables ( ) with the mapped ontology results, one table ( ) with the analysis comparison result and one plot ( ). Please refer to section 10.3.2.1 for the description of the resulting tables.

Multiple gene sets

This workflow is designed to classify several sets of genes or proteins based on the three GO branches, Reactome and HumanCyc pathways and TF classification. Several input gene or protein tables should be located in one folder.

The input is a folder with several gene or protein tables. The steps of this workflow for each individual gene or protein table are the same as described in the section above. The same steps are performed iteratively for each of the gene or protein tables in the input folder.

The output is a folder which contains subfolders with the results for each individual input table. The subfolders are automatically given the same names as the input tables.



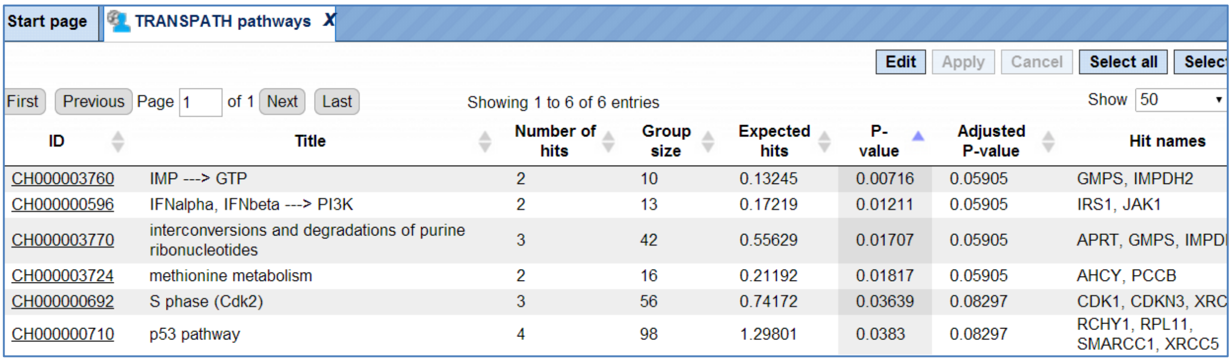
- Test (Mapping to ontologies)
  - Downregulated Entrez genes
    - Genes Ensembl
    - GO biological process
    - GO cellular component
    - GO molecular function
    - HumanCyc pathways
    - Reactome pathways
    - TF classification
  - UpDown Ensembl genes
  - Upregulated Ensembl genes

10.3.2.2. Mapping to GO categories and signaling pathways

Single gene or protein table

The steps of this workflow are the same as described above in the section 9.3.2.1. The difference is in the ontologies applied. In this workflow, your input table is mapped to GO biological processes, GO cellular components, GO molecular functions, Reactome, HumanCyc, TF classification and TRANSPATH® pathways.

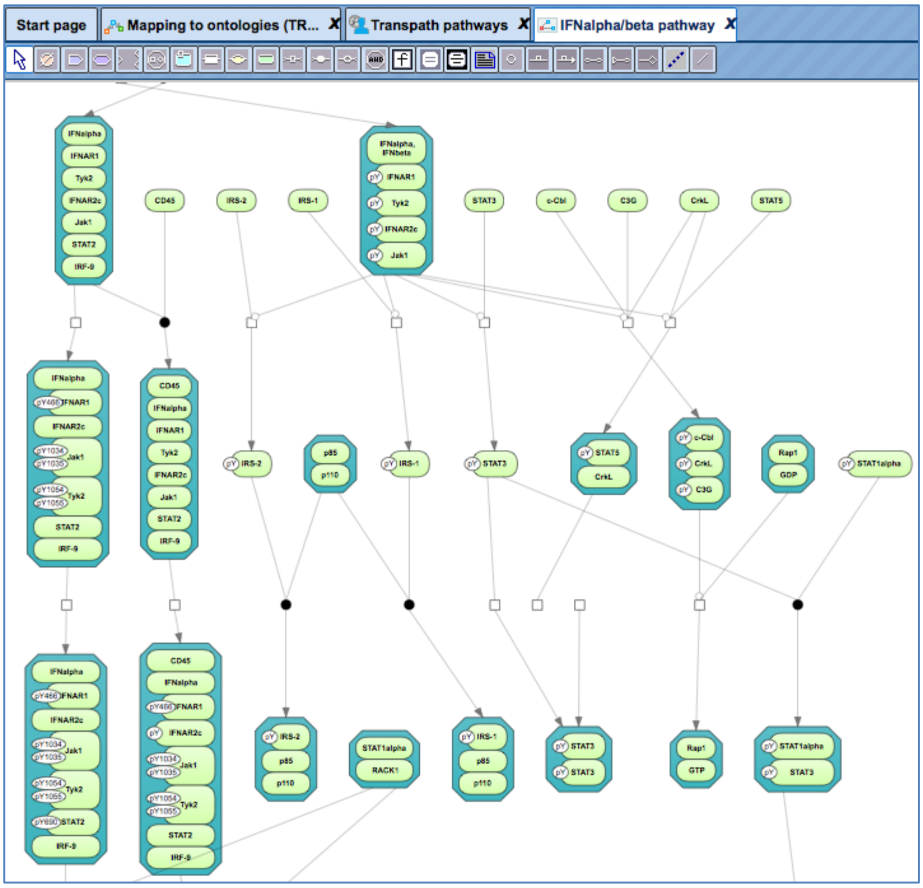
The genes in the input table are mapped to the TRANSPATH® pathways using the latest TRANSPATH® release available in the platform. The columns **ID**, **Title** and **Group size** present information about the TRANSPATH® pathways significantly enriched among your input genes.



Showing 1 to 6 of 6 entries

ID	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
CH000003760	IMP ---> GTP	2	10	0.13245	0.00716	0.05905	GMPS, IMPDH2
CH000000596	IFNalpha, IFNbeta ---> PI3K	2	13	0.17219	0.01211	0.05905	IRS1, JAK1
CH000003770	interconversions and degradations of purine ribonucleotides	3	42	0.55629	0.01707	0.05905	APRT, GMPS, IMPD
CH000003724	methionine metabolism	2	16	0.21192	0.01817	0.05905	AHCY, PCCB
CH000000692	S phase (Cdk2)	3	56	0.74172	0.03639	0.08297	CDK1, CDKN3, XRC
CH000000710	p53 pathway	4	98	1.29801	0.0383	0.08297	RCHY1, RPL11, SMARCC1, XRCC5

The pathway identifiers provide a link to open the corresponding pathway as a diagram in the work space. In the screenshot below a fragment of the IFN alpha/beta pathway is opened.



Tip You can work with the pathway diagrams as with other diagrams in the platform. For example, you can map expression data, save a copy, and export in a number of different formats. To save all genes linked to this diagram as a separate gene table, you need to select the corresponding row in the table with the classification results and apply the *Save hits* button from the top control menu.

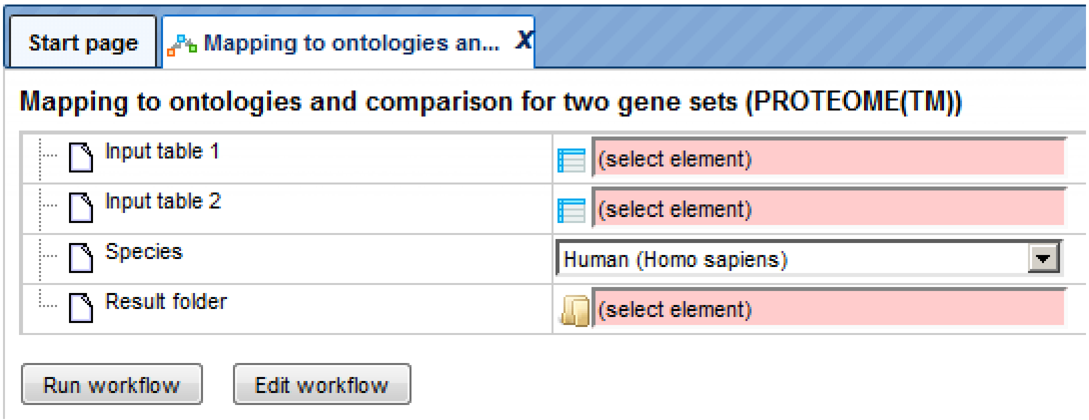
Note. This workflow is available together with a valid TRANSPATH® license. Please feel free to ask for details (info@genexplain.com).

2 Gene sets and comparison

Mapping to GO ontologies and comparison for two gene sets (PROTEOME™):

The overall idea of this workflow is similar to that the one described above. However, this workflow is designed to map two input tables, to identify hits and to compare the results according to the eight particular ontologies. These are five proprietary ontologies (BIOBASE GmbH), namely the PROTEOME™-curated Gene Ontology categories (PROTEOME™ biological process, PROTEOME™ molecular function and PROTEOME™ cellular component), PROTEOME™ disease and TRANSPATH® pathways, as well as three public ontologies, Reactome pathways, HumanCyc metabolic pathways and the transcription factor classification. Similarly to the workflows described above, the input can be any gene or protein tables. In the first step, the input tables are converted into two tables with Ensembl Gene IDs. These tables are then subjected to a functional mapping to the eight listed ontologies in parallel. The last comparison step is based on the method *analyses/Methods/Statistical analysis/Compare analysis results*, icon . Please refer to Section 16.3.1 for details on this particular analysis method. The comparison helps to reveal ontological categories that are different between two input data sets. To launch the workflow, follow these steps:

Step 1. Open the workflow input form from the Start page. It looks as shown below:



Mapping to ontologies and comparison for two gene sets (PROTEOME(TM))

Input table 1	(select element)
Input table 2	(select element)
Species	Human (Homo sapiens)
Result folder	(select element)

Step 2. Specify the input tables 1 and 2. The input gene sets might be lists of differentially regulated genes or any gene or protein list of interest. You can drag it from your project within the tree area and drop it in the pink box of the field **Input table**. Alternatively, you may click on the pink field "select element" and a new window will be opened, where you can select the input gene set as shown below.

The further steps of the workflow are demonstrated for the genes shown to be up-regulated (Top100) and down-regulated (Top100) in one of the pre-prepared examples. The pertinent example file can be found in the geneXplain platform under:

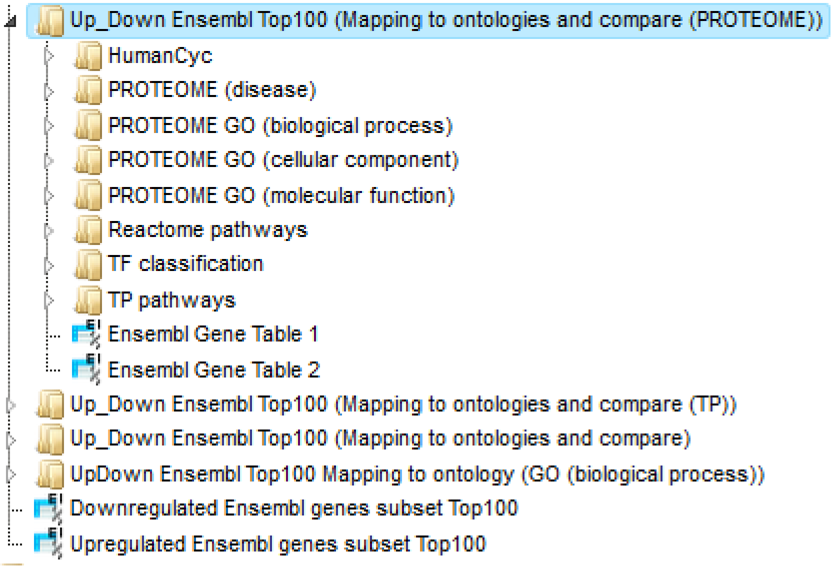
data/Examples/Brain Tumor GSE1825%2C Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/Mapping to ontologies and compare/Up_Down Ensembl Top100 (Mapping to ontologies and compare (PROTEOME))

Step 3. Specify the biological species of the input set in the field **Species** by selecting it from the drop-down menu.

Step 4. Define where the folder with the results should be located in the tree. You can do so by clicking on the pink field *select element* in the field **Result folder**, and a new window will be opened, where you can select the location of the result folder and define its name.

Step 5. Press the [Run workflow] button.

When the workflow is completed, the result folder is opened by default.



- Up_Down Ensembl Top100 (Mapping to ontologies and compare (PROTEOME))
  - HumanCyc
  - PROTEOME (disease)
  - PROTEOME GO (biological process)
  - PROTEOME GO (cellular component)
  - PROTEOME GO (molecular function)
  - Reactome pathways
  - TF classification
  - TP pathways
  - Ensembl Gene Table 1
  - Ensembl Gene Table 2
- Up_Down Ensembl Top100 (Mapping to ontologies and compare (TP))
- Up_Down Ensembl Top100 (Mapping to ontologies and compare)
- UpDown Ensembl Top100 Mapping to ontology (GO (biological process))
- Downregulated Ensembl genes subset Top100
- Upregulated Ensembl genes subset Top100

The result folder contains eight subfolders; one subfolder for each applied ontology and two tables ( ). The two tables correspond to the input tables with the identifiers converted into Ensembl gene IDs. Each subfolder contains two tables ( ) with the mapped ontology/pathway/classification results, one table ( ) with the analysis comparison result and one plot ( ). Please refer to section 9.3.2.4. for the description of the resulting tables.

Note. This workflow is available together with a valid PROTEOME™ license. Please, feel free to ask for details (info@genexplain.com).

Multiple gene sets

This workflow is designed to classify several sets of genes or proteins based on the three GO branches, Reactome and HumanCyc pathways, TF classification, and TRANSPATH® pathways. Several input gene or protein tables should be located in one folder.

The input is a folder with several gene or protein tables. The steps of this workflow for each individual gene or protein table are the same as described in 10.3.2.1, last section The same steps are performed iteratively for each of the gene or protein tables in the input folder.

The output is a folder which contains subfolders with the results for each individual input table. The subfolders are automatically given the same names as the input tables.

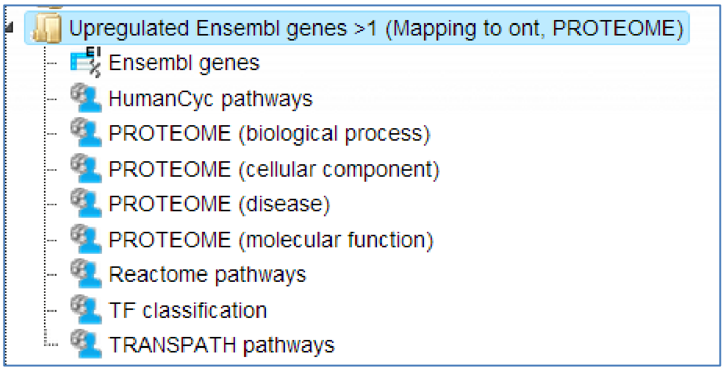
Note. This workflow is available together with a valid TRANSPATH® license. Please, feel free to ask for details (info@genexplain.com).

10.3.2.3. Mapping to GO categories, signaling pathways and diseases

Single gene or protein table

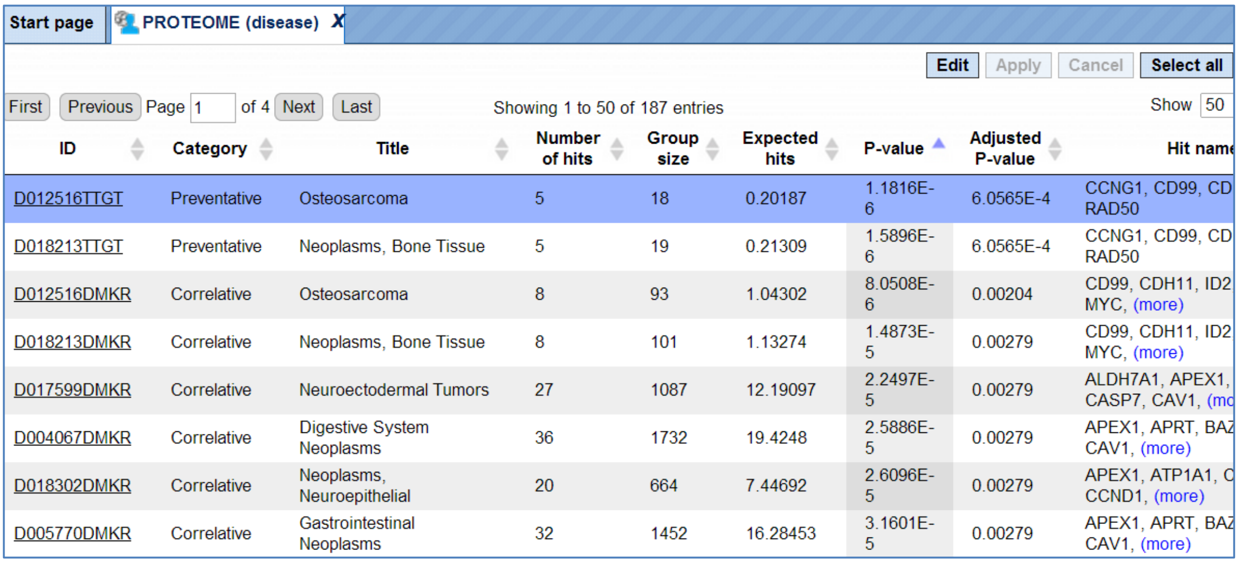
The steps of this workflow are the same as described above in the sections 9.3.2.1 and 9.3.2.3. The difference is in the ontologies applied. In this workflow, your input table is mapped to PROTEOME™ biological processes, PROTEOME™ cellular components, PROTEOME™ molecular functions, PROTEOME™ disease, Reactome, HumanCyc, TF classification and TRANSPATH® pathways.

The results folder contains several tables with the resulting mapping, one table each for each applied ontological group (), as well as one gene table () as shown below.



Upregulated Ensembl genes >1 (Mapping to ont, PROTEOME)
- Ensembl genes
- HumanCyc pathways
- PROTEOME (biological process)
- PROTEOME (cellular component)
- PROTEOME (disease)
- PROTEOME (molecular function)
- Reactome pathways
- TF classification
- TRANSPATH pathways

When the workflow is completed, all the output files are opened by default. The output file **PROTEOME (disease)** when opened in the work space looks is shown here:



Showing 1 to 50 of 187 entries

ID	Category	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit name
D012516TTGT	Preventative	Osteosarcoma	5	18	0.20187	1.1816E-6	6.0565E-4	CCNG1, CD99, CD RAD50
D018213TTGT	Preventative	Neoplasms, Bone Tissue	5	19	0.21309	1.5896E-6	6.0565E-4	CCNG1, CD99, CD RAD50
D012516DMKR	Correlative	Osteosarcoma	8	93	1.04302	8.0508E-6	0.00204	CD99, CDH11, ID2 MYC, (more)
D018213DMKR	Correlative	Neoplasms, Bone Tissue	8	101	1.13274	1.4873E-5	0.00279	CD99, CDH11, ID2 MYC, (more)
D017599DMKR	Correlative	Neuroectodermal Tumors	27	1087	12.19097	2.2497E-5	0.00279	ALDH7A1, APEX1, CASP7, CAV1, (mo
D004067DMKR	Correlative	Digestive System Neoplasms	36	1732	19.4248	2.5886E-5	0.00279	APEX1, APRT, BAZ CAV1, (more)
D018302DMKR	Correlative	Neoplasms, Neuroepithelial	20	664	7.44692	2.6096E-5	0.00279	APEX1, ATP1A1, C CCND1, (more)
D005770DMKR	Correlative	Gastrointestinal Neoplasms	32	1452	16.28453	3.1601E-5	0.00279	APEX1, APRT, BAZ CAV1, (more)

The columns **ID**, **Category**, **Title** and **Group size** present information about the diseases, as they are curated in the PROTEOME™ database, significantly enriched among your input genes.

Each disease identifier is hyperlinked to an external web page, the Comparative Toxicogenomic Database, where you can find more details about this disease:

http://ctdbase.org/

Note. This workflow is available together with a valid PROTEOME™/TRANSPATH® license. Please, feel free to ask for details (info@genexplain.com).

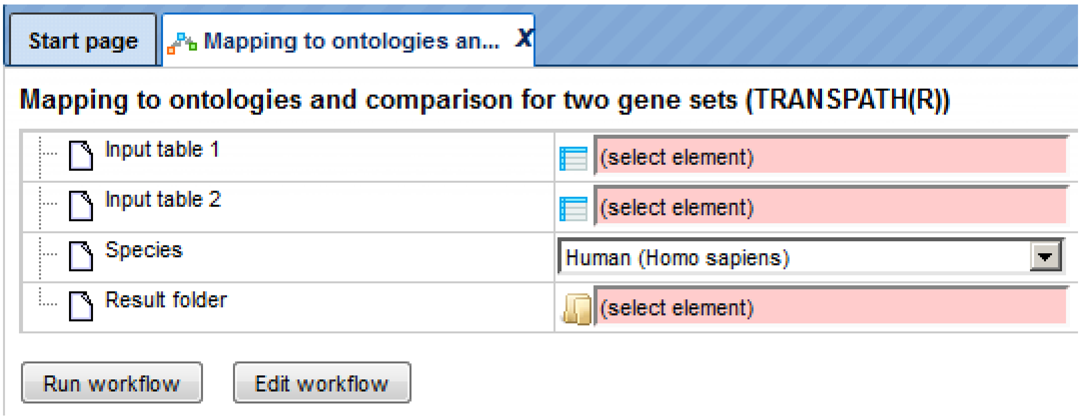
2 Gene sets and comparison

Mapping to ontologies and comparison for two gene sets (TRANSPATH®)

The specialty of this workflow in comparison with the one described above is in the ontologies applied. This workflow is designed to map two input tables to the seven following ontologies, the public **Gene Ontology** categories (*biological process*, *molecular function* and *cellular component*), TRANSPATH®, Reactome and HumanCyc pathways as well as transcription factor classification to identify hits and to compare the results. Similar to the workflows described above, the input can be any gene or protein tables. In the first step, the input tables are converted into two tables with Ensembl Gene IDs. These tables are subjected to a functional mapping. The last comparison step is based on the method *analyses/Methods/Statistical analysis/Compare analysis results*, icon . Please refer to section 16.3 for details on this particular analysis method. The comparison can help to reveal items that show different enrichment across certain conditions.

To launch the workflow, follow these steps:

Step 1. Open the workflow input form from the Start page. It looks as shown below:



Mapping to ontologies and comparison for two gene sets (TRANSPATH(R))

Input table 1	(select element)
Input table 2	(select element)
Species	Human (Homo sapiens)
Result folder	(select element)

Step 2. Specify the input tables 1 and 2. The input gene sets might be lists of differentially regulated genes or any gene or protein list of interest. You can drag it from your project within the tree area and drop it in the pink box of the field **Input table**. Alternatively, you may click on the pink field “select element” and a new window will be opened, where you can select the input gene set as shown below.

The further steps of the workflow are demonstrated for the genes shown to be up-regulated (Top100) and down-regulated (Top100) in one of the pre-prepared examples. The pertinent example file can be found in the geneXplain platform under:

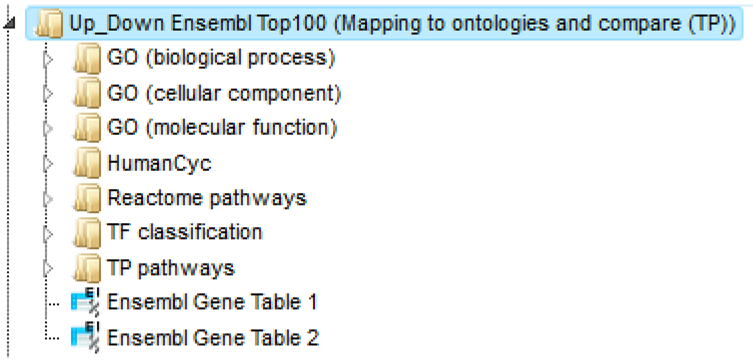
data/Examples/Brain Tumor GSE1825%2C Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/Mapping to ontologies and compare/Up Down Ensembl Top100 (Mapping to ontologies and compare (TP))

Step 3. Specify the biological species of the input set in the field **Species** by selecting it from the drop-down menu.

Step 4. Define where the folder with the results should be located in the tree. You can do so by clicking on the pink field *select element* in the field **Result folder**, and a new window will be opened, where you can select the location of the result folder and define its name.

Step 5. Press the [Run workflow] button.

When the workflow is completed, the result folder is opened by default.



Up_Down Ensembl Top100 (Mapping to ontologies and compare (TP))
- GO (biological process)
- GO (cellular component)
- GO (molecular function)
- HumanCyc
- Reactome pathways
- TF classification
- TP pathways
- Ensembl Gene Table 1
- Ensembl Gene Table 2

The result folder contains the seven subfolders; one subfolder for each applied ontology and two tables ( ). The two tables correspond to the input tables with the identifiers converted into Ensembl gene IDs. Each subfolder contains two tables ( ) with the mapped ontology/pathway/classification results, one table ( ) with the analysis comparison result and one plot ( ). Please refer to section 9.3.2.4. for the description of the resulting tables.

Note. This workflow is available together with a valid TRANSPATH® license. Please, feel free to ask for details (info@genexplain.com).

Multiple gene sets

Please refer to 10.3.1.1, last section, for the description of the resulting tables.

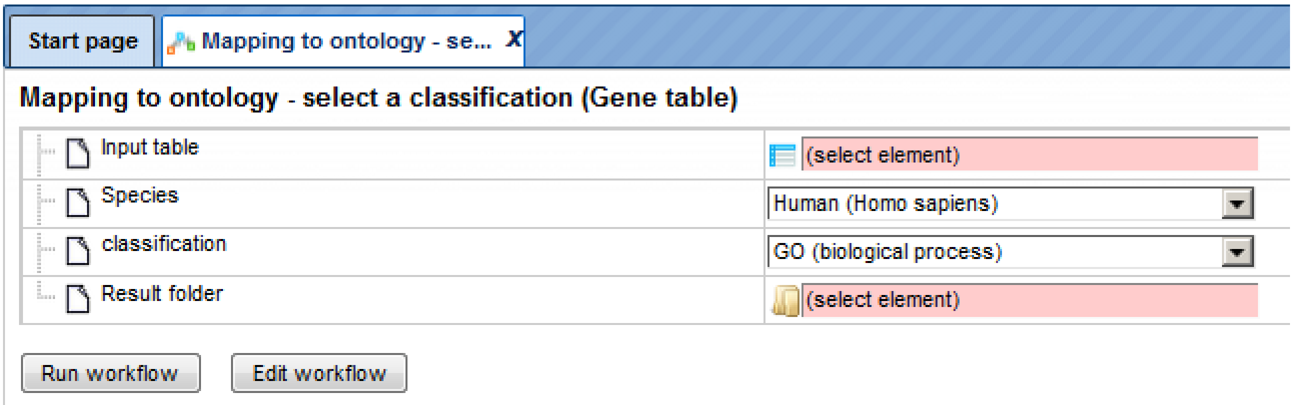
10.3.2.4. Mapping with selected classification

Single gene set

This workflow is designed to map one input tables to one selected ontology classification. The input can be any gene or protein table. In the first step, the input table is converted into one table with Ensembl Gene IDs. The table with Ensembl Gene ID is subjected to a functional classification. As result the mapped table is stored.

To launch the workflow, follow these steps:

Step 1. Open the workflow input form from the Start page. It looks as shown below:



Start page | Mapping to ontology - se... X

Mapping to ontology - select a classification (Gene table)

Input table	(select element)
Species	Human (Homo sapiens)
classification	GO (biological process)
Result folder	(select element)

Run workflow | Edit workflow

Step 2. Specify the input table. The input gene set might be the list of differentially regulated genes or any gene or protein list of interest. You can drag it from your project within the tree area and drop it in the pink box of the field **Input table**. Alternatively, you may click on the pink field “select element” and a new window will be opened, where you can select the input gene set as shown below.

The pertinent example file can be found in the geneXplain platform under:

data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/Experiment normalized (RMA) (Differentially expressed genes Affy)/Upregulated Ensembl genes filtered (LogFC>2)

Step 3. Specify the biological species of the input set in the field **Species** by selecting it from the drop-down menu.

Step 4. In the field **Classification**, choose the ontology from the drop-down menu. Here, GO biological process is selected.

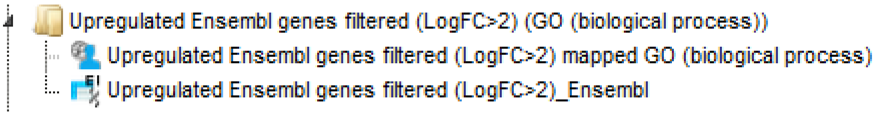
Step 5. Define where the folder with the results should be located in the tree. You can do so by clicking on the pink field *select element* in the field **Result folder**, and a new window will be opened, where you can select the location of the result folder and define its name.

Step 6. Press the [Run workflow] button.

When the workflow is completed, the result folder is opened by default.

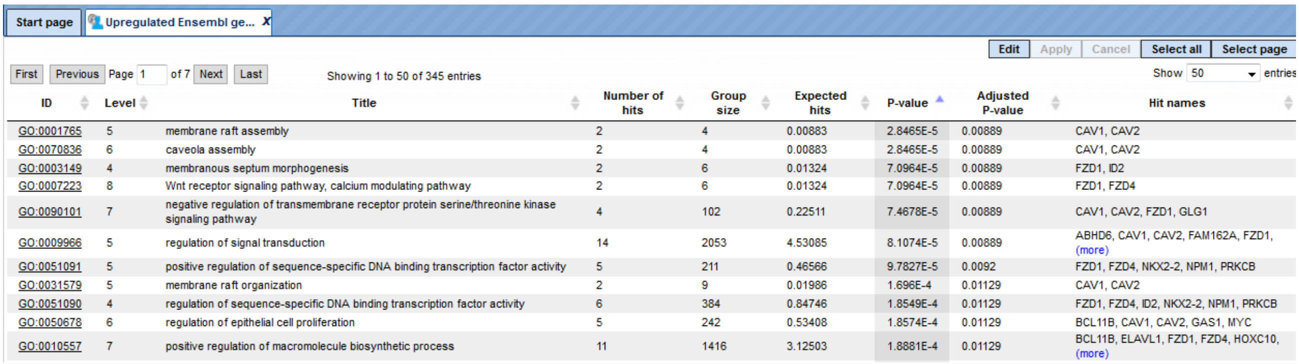
The result folder contains 1 tables ( ) with the converted Ensembl table and one table ( ) with the mapped ontology results.

data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/Upregulated Ensembl genes filtered (LogFC>2) (GO (biological process))



Upregulated Ensembl genes filtered (LogFC>2) (GO (biological process))
- Upregulated Ensembl genes filtered (LogFC>2) mapped GO (biological process)
- Upregulated Ensembl genes filtered (LogFC>2)_Ensembl

Let’s consider the output table with the mapping results. The tables with the **mapped** selected category ( ) look like:



Showing 1 to 50 of 345 entries

ID	Level	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
GO:0001765	5	membrane raft assembly	2	4	0.00883	2.8465E-5	0.00889	CAV1, CAV2
GO:0070836	6	caveola assembly	2	4	0.00883	2.8465E-5	0.00889	CAV1, CAV2
GO:0003149	4	membranous septum morphogenesis	2	6	0.01324	7.0964E-5	0.00889	FZD1, ID2
GO:0007223	8	Wnt receptor signaling pathway, calcium modulating pathway	2	6	0.01324	7.0964E-5	0.00889	FZD1, FZD4
GO:0090101	7	negative regulation of transmembrane receptor protein serine/threonine kinase signaling pathway	4	102	0.22511	7.4678E-5	0.00889	CAV1, CAV2, FZD1, GLG1
GO:0009966	5	regulation of signal transduction	14	2053	4.53085	8.1074E-5	0.00889	ABHD6, CAV1, CAV2, FAM162A, FZD1, (more)
GO:0051091	5	positive regulation of sequence-specific DNA binding transcription factor activity	5	211	0.46566	9.7827E-5	0.0092	FZD1, FZD4, NKX2-2, NPM1, PRKCB
GO:0031579	5	membrane raft organization	2	9	0.01986	1.696E-4	0.01129	CAV1, CAV2
GO:0051090	4	regulation of sequence-specific DNA binding transcription factor activity	6	384	0.84746	1.8549E-4	0.01129	FZD1, FZD4, ID2, NKX2-2, NPM1, PRKCB
GO:0050678	6	regulation of epithelial cell proliferation	5	242	0.53408	1.8574E-4	0.01129	BCL11B, CAV1, CAV2, GAS1, MYC
GO:0010557	7	positive regulation of macromolecule biosynthetic process	11	1416	3.12503	1.8881E-4	0.01129	BCL11B, ELAVL1, FZD1, FZD4, HOXC10, (more)

Each row presents details about one ontological term. The column **ID** comprises the identifiers of the ontological categories, here identifiers of Gene Ontology biological process terms. These identifiers are hyperlinked to the page http://www.ebi.ac.uk/QuickGO/ where you can get further information about this ontological term.

The columns **Title** and **Group size** contain further details about the ontological terms, its title and the number of genes linked to this term in the corresponding database, here in GO. The column **Expected hits** shows the number of genes expected to fall into this category by random chance, based on the size of the input set and the size of the category. The column **Number of hits** shows how many genes from the input table fall into this category. **P-value** and **Adjusted P-value** are calculated for the difference between expected and real numbers of hits. The genes mapped to each category are explicitly listed in the column **Hit names**. As the lists can get quite long, only a few genes are shown by default in each row. To get the full list, press [more].

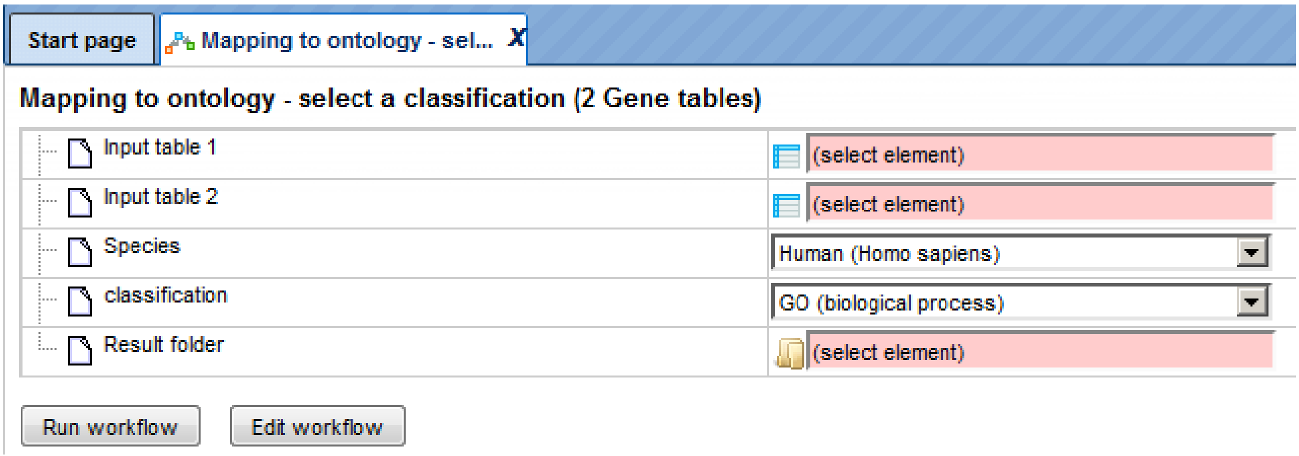
2 Gene sets and comparison

Mapping to ontology - select a classification (2 Gene tables)

This workflow is designed to map each of the two input tables to one selected ontology classification, to identify term hits and to compare the results. The input can be any gene or protein table. In the first step, the input tables are converted into two tables with Ensembl Gene IDs. These tables with Ensembl Gene IDs are subjected to a functional classification. As result two mapped tables are stored and further compared via P-values. This final comparison step is based on the method *analyses/Methods/Statistical analysis/Compare analysis results*, icon . Please refer to section 13.3 for details on this particular analysis method. The comparison can help to reveal terms that show different enrichment across certain conditions.

To launch the workflow, follow these steps:

Step 1. Open the workflow input form from the Start page. It looks as shown below:



Mapping to ontology - select a classification (2 Gene tables)

Input table 1	(select element)
Input table 2	(select element)
Species	Human (Homo sapiens)
classification	GO (biological process)
Result folder	(select element)

Step 2. Specify the input tables 1 and 2. The input gene sets might be the lists of differentially regulated genes or any gene or protein list of interest. You can drag it from your project within the tree area and drop it in the pink box of the field **Input table**. Alternatively, you may click on the pink field "select element" and a new window will be opened, where you can select the input gene set as shown below.

The further steps of the workflow are demonstrated for the genes shown to be up-regulated (Top100) and down-regulated (Top100) in one of the pre-prepared examples. The pertinent example file can be found in the geneXplain platform under:

data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/Mapping to ontologies and compare/Up Down Ensembl Top100 (Mapping to ontology (GO (biological process))

Step 3. Specify the biological species of the input set in the field **Species** by selecting it from the drop-down menu.

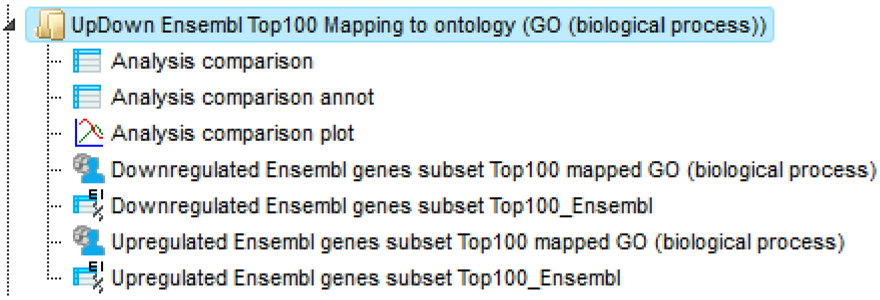
Step 4. In the field **Classification**, choose the ontology from the drop-down menu. Here, GO biological process is selected.

Step 5. Define where the folder with the results should be located in the tree. You can do so by clicking on the pink field *select element* in the field **Result folder**, and a new window will be opened, where you can select the location of the result folder and define its name.

Step 6. Press the [Run workflow] button.

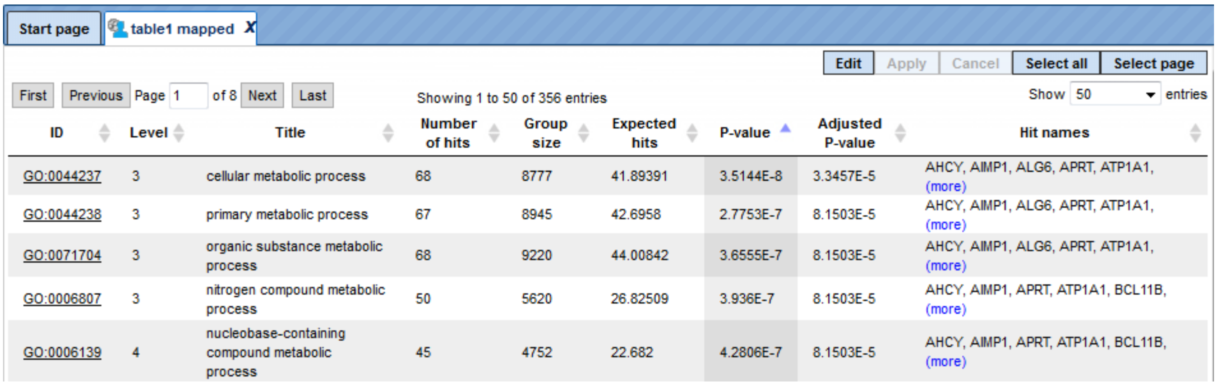
When the workflow is completed, the result folder is opened by default.

The result folder contains 2 tables ( ) with the converted Ensembl tables, two tables with the mapped ontology results, two tables ( ) with the analysis comparison result annotated and one plot ( ).



UpDown Ensembl Top100 Mapping to ontology (GO (biological process))
- Analysis comparison
- Analysis comparison annot
- Analysis comparison plot
- Downregulated Ensembl genes subset Top100 mapped GO (biological process)
- Downregulated Ensembl genes subset Top100_Ensembl
- Upregulated Ensembl genes subset Top100 mapped GO (biological process)
- Upregulated Ensembl genes subset Top100_Ensembl

Let's consider the output tables. The tables with the **mapped** selected category ( ) look like:



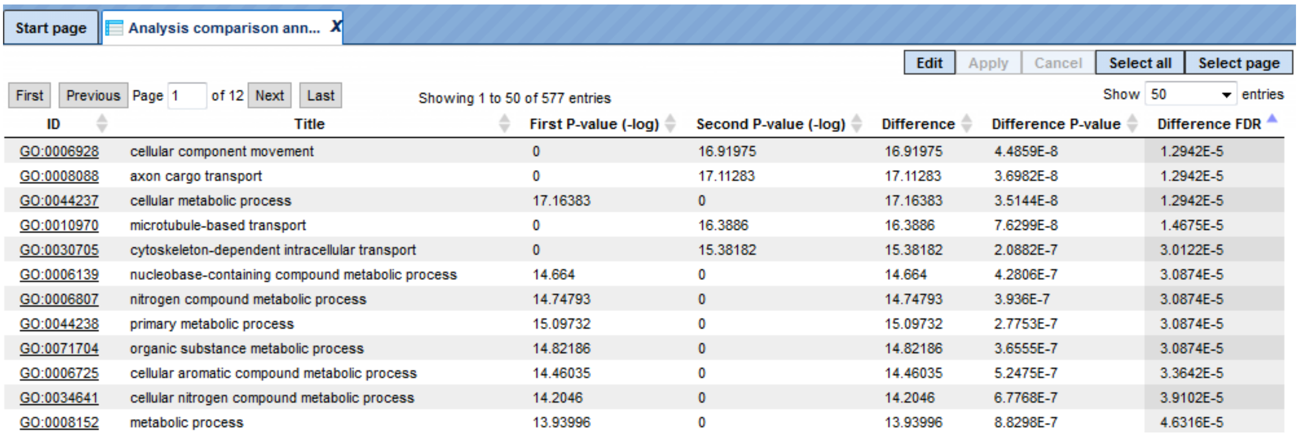
Showing 1 to 50 of 356 entries

ID	Level	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
GO:0044237	3	cellular metabolic process	68	8777	41.89391	3.5144E-8	3.3457E-5	AHCY, AIMP1, ALG6, APRT, ATP1A1, (more)
GO:0044238	3	primary metabolic process	67	8945	42.6958	2.7753E-7	8.1503E-5	AHCY, AIMP1, ALG6, APRT, ATP1A1, (more)
GO:0071704	3	organic substance metabolic process	68	9220	44.00842	3.6555E-7	8.1503E-5	AHCY, AIMP1, ALG6, APRT, ATP1A1, (more)
GO:0006807	3	nitrogen compound metabolic process	50	5620	26.82509	3.936E-7	8.1503E-5	AHCY, AIMP1, APRT, ATP1A1, BCL11B, (more)
GO:0006139	4	nucleobase-containing compound metabolic process	45	4752	22.682	4.2806E-7	8.1503E-5	AHCY, AIMP1, APRT, ATP1A1, BCL11B, (more)

Each row presents details about one ontological term. The column **ID** comprises the identifiers of the ontological categories, here identifiers of Gene Ontology biological process terms. These identifiers are hyperlinked to the page http://www.ebi.ac.uk/QuickGO/ where you can get further information about this ontological term.

The columns **Title** and **Group size** contain further details about the ontological terms, its title and the number of genes linked to this term in the corresponding database, here in GO. The column **Expected hits** shows the number of genes expected to fall into this category by random chance, based on the size of the input set and the size of the category. The column **Number of hits** shows how many genes from the input table fall into this category. **P-value** and **Adjusted P-value** are calculated for the difference between expected and real numbers of hits. The genes mapped to each category are explicitly listed in the column **Hit names**. As the lists can get quite long, only a few genes are shown by default in each row. To get the full list, press [more].

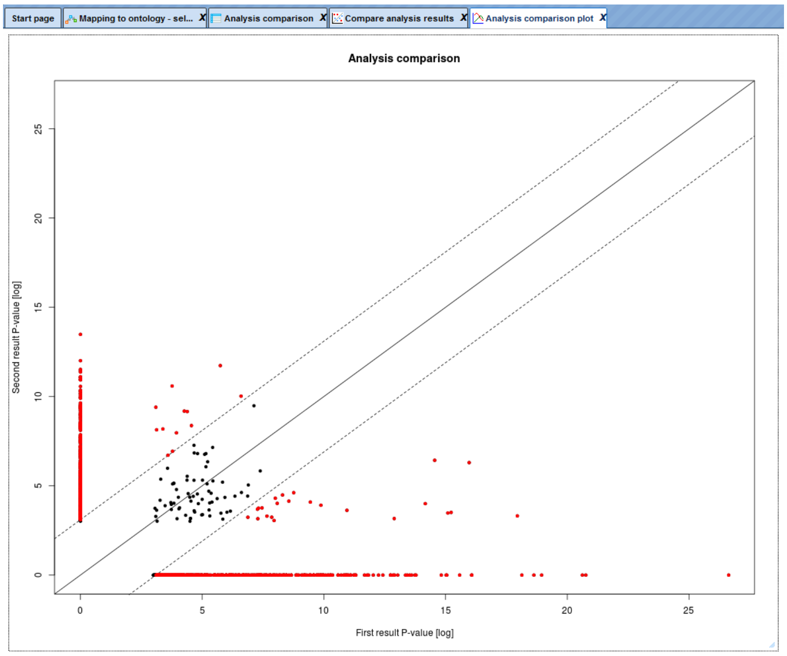
The **Analysis comparison annot** table lists identifiers, annotation of identifiers, P-values for the first input set of genes and P-values for the second input set of genes (-log). The column Difference shows the absolute difference between two P-values. The columns **Difference P-value** and **Difference FDR** show the statistical significance of the absolute difference and upon sorting by one of these two columns on top you can see those ontology terms that are statistically most significantly different between two input gene sets. The comparison reveals GO terms that show different enrichment across the two input gene lists.



Showing 1 to 50 of 577 entries

ID	Title	First P-value (-log)	Second P-value (-log)	Difference	Difference P-value	Difference FDR
GO:0006928	cellular component movement	0	16.91975	16.91975	4.4859E-8	1.2942E-5
GO:0008088	axon cargo transport	0	17.11283	17.11283	3.6982E-8	1.2942E-5
GO:0044237	cellular metabolic process	17.16383	0	17.16383	3.5144E-8	1.2942E-5
GO:0010970	microtubule-based transport	0	16.3886	16.3886	7.6299E-8	1.4675E-5
GO:0030705	cytoskeleton-dependent intracellular transport	0	15.38182	15.38182	2.0882E-7	3.0122E-5
GO:0006139	nucleobase-containing compound metabolic process	14.664	0	14.664	4.2806E-7	3.0874E-5
GO:0006807	nitrogen compound metabolic process	14.74793	0	14.74793	3.936E-7	3.0874E-5
GO:0044238	primary metabolic process	15.09732	0	15.09732	2.7753E-7	3.0874E-5
GO:0071704	organic substance metabolic process	14.82186	0	14.82186	3.6555E-7	3.0874E-5
GO:0006725	cellular aromatic compound metabolic process	14.46035	0	14.46035	5.2475E-7	3.3642E-5
GO:0034641	cellular nitrogen compound metabolic process	14.2046	0	14.2046	6.7768E-7	3.9102E-5
GO:0008152	metabolic process	13.93996	0	13.93996	8.8298E-7	4.6316E-5

The **Analysis comparison plot** is a scatter plot of P-values on the log-scale together with the diagonal and the difference cutoffs at FDR < 0.05. Every dot corresponds to one particular GO term. On the X-axis, the –log(p-value) for this GO term in the first input table is shown, and on the Y-axis, the –log(p-value) for the same GO term in the second input table is shown. The red dots correspond to those GO terms that are statistically significantly different between two input tables, with FDR<0.05. The black dots located close to the diagonal, between two dotted lines, are not significantly different between two input datasets.

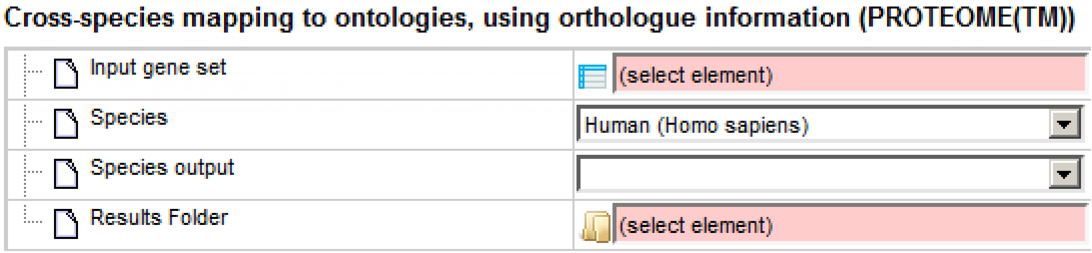


Multiple gene sets

Please refer to 10.3.1.1, last section, for the description of the resulting tables.

10.3.2.5. Cross-species mapping to ontologies, using ortholog information (PROTEOME™)

This workflow is very similar to the workflows described above in sections 10.3.2.1 and 10.3.2.3. The Input can be any gene or protein table for mouse or rat. The workflow will convert the list to desired species output and mapp it to various ontologies. In this workflow, your input table is mapped to PROTEOME™ biological processes, PROTEOME™ cellular components, PROTEOME™ molecular functions, PROTEOME™ disease, Reactome, HumanCyc, TF classification and TRANSPATH® pathways. It can be found under the tab Workflows, in the folder PROTEOME™/Cross-species mapping to ontologies, using ortholog information (PROTEOME™). The input form of the workflow looks as shown below:



Cross-species mapping to ontologies, using orthologue information (PROTEOME(TM))

Input gene set	(select element)
Species	Human (Homo sapiens)
Species output	
Results Folder	(select element)

Step 1: Input the gene or protein table of any species for which you wish to map gene ontologies. You can drag & drop it from your project within the tree area. Alternatively, you may click on the pink field "select element" and a new window will open, where you can select the input table.

Here further steps are demonstrated with the track available in one of the pre-prepared examples present in the Tree Area:

data/Examples/Transcriptional biomarkers to predict mouse liver tumors, GSE18858/Data/Normalized (RMA) DEGs with EBarrays/Naphthalene_20ppm upreg Ensembl Select the species of the input table

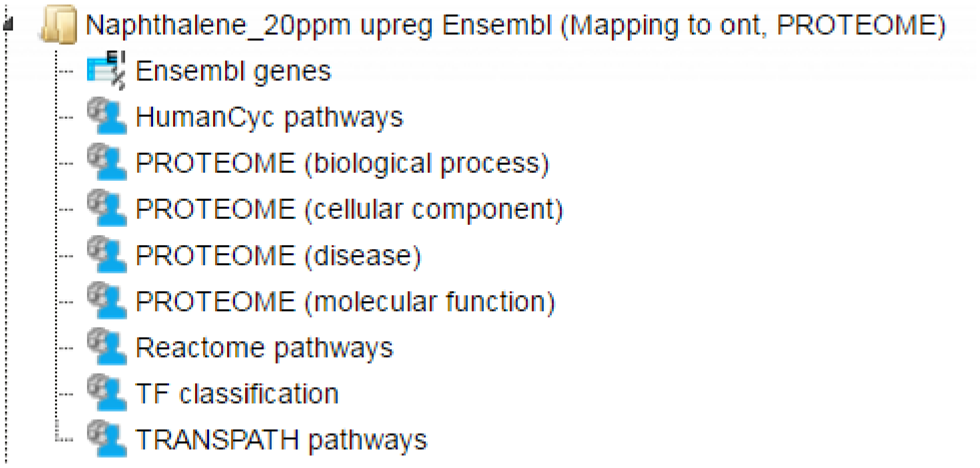
Step 2: Select the desired species of the output table.

Specify the path to store the results and the name of the output folder.

Having filled in the input form, launch the workflow with the [Run] button. Wait till the workflow is completed.

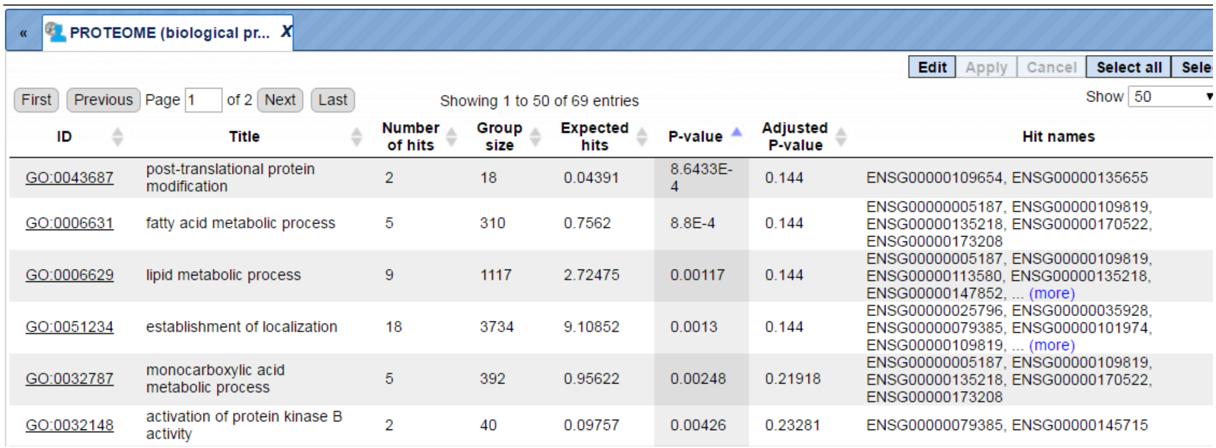
Here the input mouse data is functionally classified and mapped to human data.

In the first step of the workflow, the input gene table is converted to the desired Ensemble gene table using the method ‚convert table via homology‘. In the next step the converted Ensembl gene table is functionally classified using the Proteome database with a p_value threshold of 0.05. The output results folder contains diverse files as shown below:



Naphthalene_20ppm upreg Ensembl (Mapping to ont, PROTEOME)
- Ensembl genes
- HumanCyc pathways
- PROTEOME (biological process)
- PROTEOME (cellular component)
- PROTEOME (disease)
- PROTEOME (molecular function)
- Reactome pathways
- TF classification
- TRANSPATH pathways

Mapping to the three GO branches, biological processes, cellular components, and molecular functions ( ). The tables with the enriched categories look like:



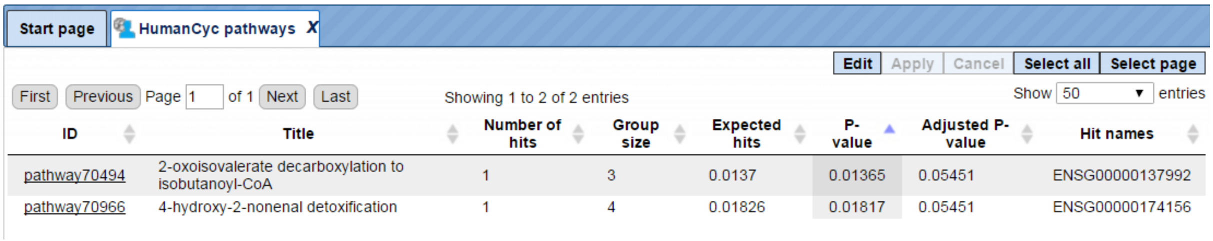
« PROTEOME (biological pr... X
Edit Apply Cancel Select all Sele
First Previous Page 1 of 2 Next Last
Show 50

Showing 1 to 50 of 69 entries

ID	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
GO:0043687	post-translational protein modification	2	18	0.04391	8.6433E-4	0.144	ENSG00000109654, ENSG00000135655
GO:0006631	fatty acid metabolic process	5	310	0.7562	8.8E-4	0.144	ENSG00000005187, ENSG00000109819, ENSG00000135218, ENSG00000170522, ENSG00000173208
GO:0006629	lipid metabolic process	9	1117	2.72475	0.00117	0.144	ENSG00000005187, ENSG00000109819, ENSG00000113580, ENSG00000135218, ENSG00000147852, ... (more)
GO:0051234	establishment of localization	18	3734	9.10852	0.0013	0.144	ENSG00000025796, ENSG00000035928, ENSG00000079385, ENSG00000101974, ENSG00000109819, ... (more)
GO:0032787	monocarboxylic acid metabolic process	5	392	0.95622	0.00248	0.21918	ENSG00000005187, ENSG00000109819, ENSG00000135218, ENSG00000170522, ENSG00000173208
GO:0032148	activation of protein kinase B activity	2	40	0.09757	0.00426	0.23281	ENSG00000079385, ENSG00000145715

For each ontological term several parameters are calculated, including expected number of hits, actual number of hits, p-value, as well as hit names and the link to the corresponding ontological term. For more details on the details of each column please refer to section 10.3.2.1.

The Tables Reactome pathways, Transpath pathways, and humanCyc pathways give the results of the mapping of the input gene set to each of these pathways. Each table has a unique identifier for the corresponding pathway; upon a mouse click, a diagram opens in the workspace.

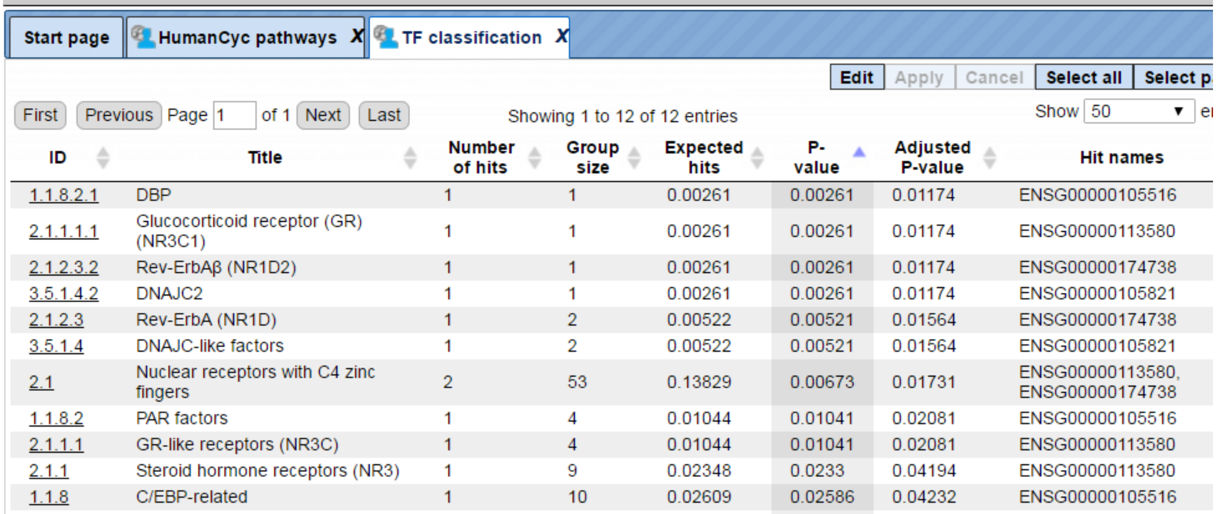


Start page
HumanCyc pathways X
Edit Apply Cancel Select all Select page
First Previous Page 1 of 1 Next Last
Show 50 entries

Showing 1 to 2 of 2 entries

ID	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
pathway70494	2-oxoisovalerate decarboxylation to isobutanoyl-CoA	1	3	0.0137	0.01365	0.05451	ENSG00000137992
pathway70966	4-hydroxy-2-nonenal detoxification	1	4	0.01826	0.01817	0.05451	ENSG00000174156

The table TF classification ( ). Your input table is mapped to the classification of Transcription factors (*Nucleic Acids Res. 41, D165-D170 (2013)*), which is also integrated in the platform. In the column **ID** the identifiers of the TF classification are shown. They are hyperlinked to the corresponding classification categories:



ID	Title	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
1.1.8.2.1	DBP	1	1	0.00261	0.00261	0.01174	ENSG00000105516
2.1.1.1.1	Glucocorticoid receptor (GR) (NR3C1)	1	1	0.00261	0.00261	0.01174	ENSG00000113580
2.1.2.3.2	Rev-ErbAβ (NR1D2)	1	1	0.00261	0.00261	0.01174	ENSG00000174738
3.5.1.4.2	DNAJC2	1	1	0.00261	0.00261	0.01174	ENSG00000105821
2.1.2.3	Rev-ErbA (NR1D)	1	2	0.00522	0.00521	0.01564	ENSG00000174738
3.5.1.4	DNAJC-like factors	1	2	0.00522	0.00521	0.01564	ENSG00000105821
2.1	Nuclear receptors with C4 zinc fingers	2	53	0.13829	0.00673	0.01731	ENSG00000113580, ENSG00000174738
1.1.8.2	PAR factors	1	4	0.01044	0.01041	0.02081	ENSG00000105516
2.1.1.1	GR-like receptors (NR3C)	1	4	0.01044	0.01041	0.02081	ENSG00000113580
2.1.1	Steroid hormone receptors (NR3)	1	9	0.02348	0.0233	0.04194	ENSG00000113580
1.1.8	C/EBP-related	1	10	0.02609	0.02586	0.04232	ENSG00000105516

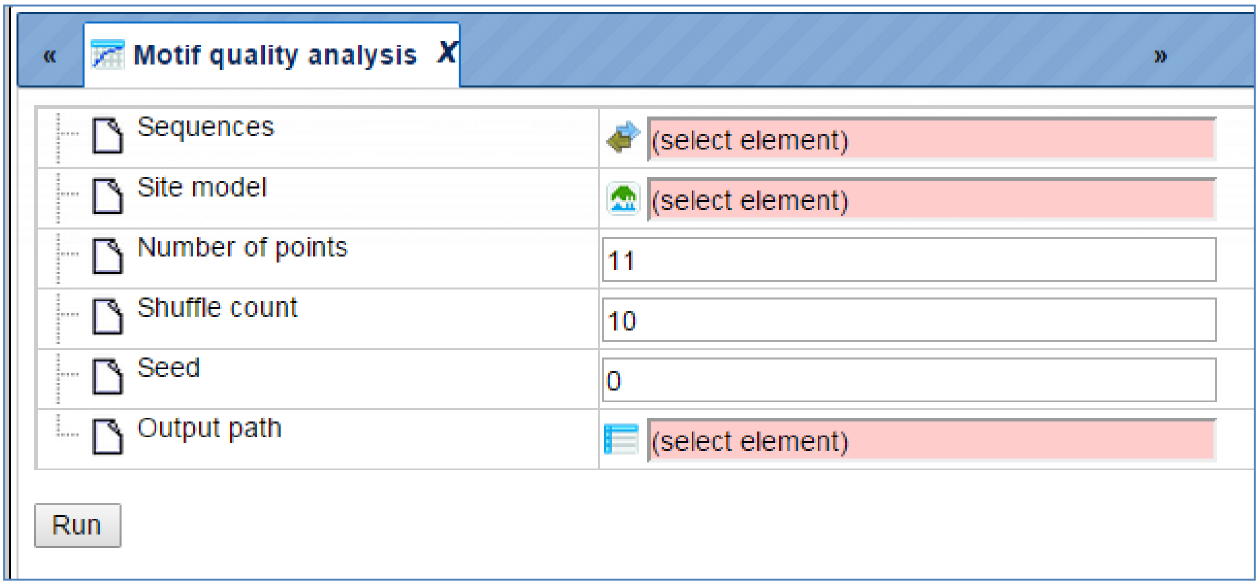
10.4. Analyze regulatory regions

This set of workflows helps to find putative TF binding sites in the DNA sequences under study. There are several workflows in this group that perform searches in different genomic regions, either in promoters, in the peaks calculated from ChIP-seq data, or in any input DNA sequences. This group of workflows is designed using the core functionality of a "site search on gene set" analysis as described in Section 20.1.2.

10.4.1. Motif quality analysis

This tool analyzes the quality of a motif model. The "Motif quality analysis" item is located in the NGS folder of the analysis methods (analyses/Methods/Site analysis/Motif quality analysis) and in the start page group 'Microarrays' under section 'Analyze regulatory regions'.

Step 1. Open the analysis form from the Start page. It will open in the main Work Space and looks as shown below:



Motif quality analysis

Sequences	(select element)
Site model	(select element)
Number of points	11
Shuffle count	10
Seed	0
Output path	(select element)

Step 2. The Sequences input is a track file with sequences containing the motif.

The following link directs to an example input:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Encode%20TFBS%20CEBPB%20in%20H1-hESC%20cells/Data/CEBP%20in%20H1-hESC%20cells%20YES

Step 3. Select a Site model from a profile which can be used to compare the input motif. The model can result from a workflow generated 'Profile', can be selected from the TRANSFAC® database or can be built from the 'Create profile from matrix library' method (input is ChIPHorde or DiChIPHorde motif).

For this example we selected the profile: http://genexplain-platform.com/bioumlweb/#de=data/Examples/Encode%20TFBS%20CEBPB%20in%20H1-hESC%20cells/Data/ChIPMunk/CEBP%20H1-hESC%20cells%20motif%20profile

CEBP H1-hESC cells motif profile

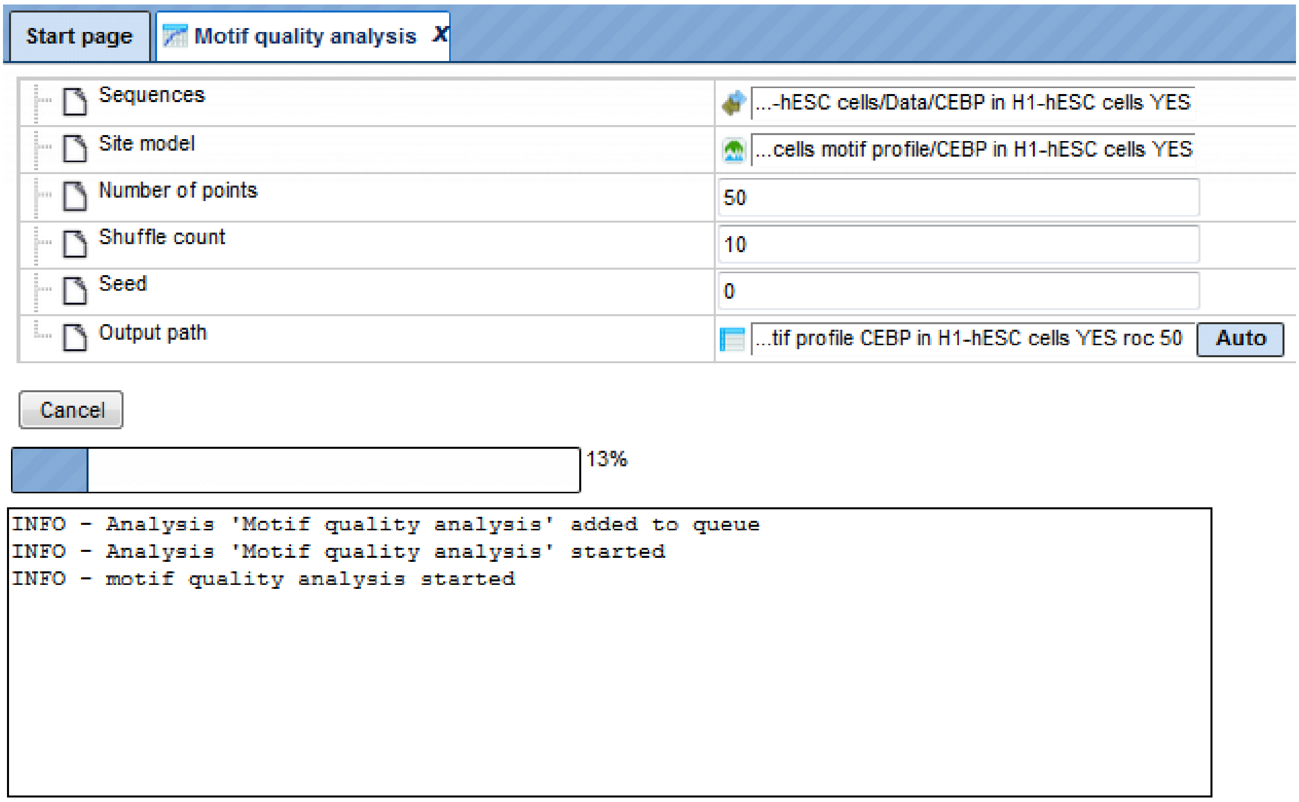
Step 4. Specify the total **Number of points** for sensitivity and FDR calculation. By default, the analysis uses 11 points. For the example we use 50 points.

Step 5. Specify the number of **Shuffle counts**. This is the number of times sequence characters are shuffled to generate random sequences for FDR estimation. By default this number is 10.

Step 6. Select a **Seed** for the random number generator or keep the default of 0.

Step 7. Declare the **Output path** to store results in the tree area.

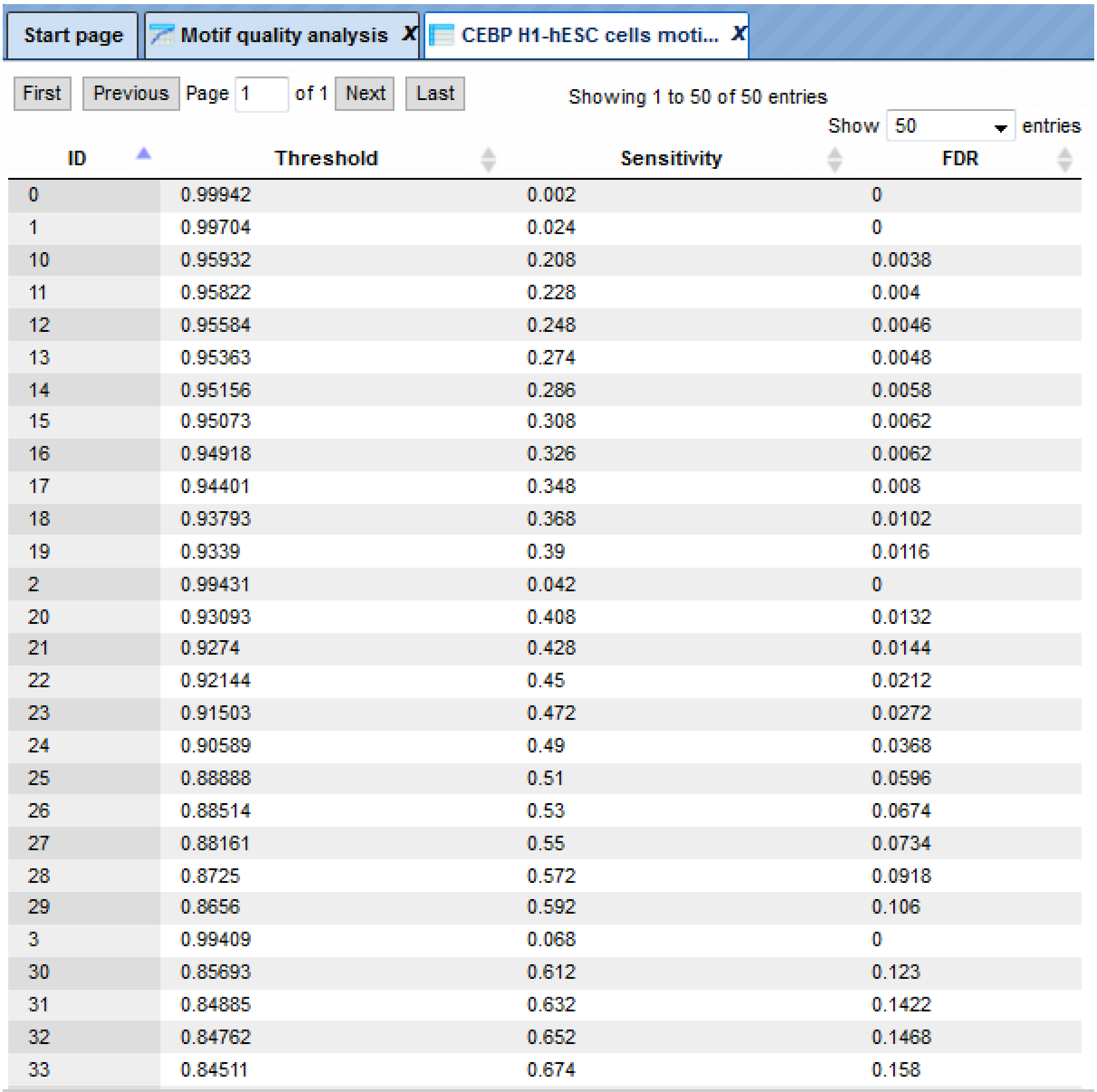
After entering the input parameters, press 'RUN'. The method starts as shown below:



Sequences	...-hESC cells/Data/CEBP in H1-hESC cells YES
Site model	...cells motif profile/CEBP in H1-hESC cells YES
Number of points	50
Shuffle count	10
Seed	0
Output path	...tif profile CEBP in H1-hESC cells YES roc 50

```
INFO - Analysis 'Motif quality analysis' added to queue
INFO - Analysis 'Motif quality analysis' started
INFO - motif quality analysis started
```

Post completion the output table is opened in the work space in a new tab and consists of a table like the one shown below.

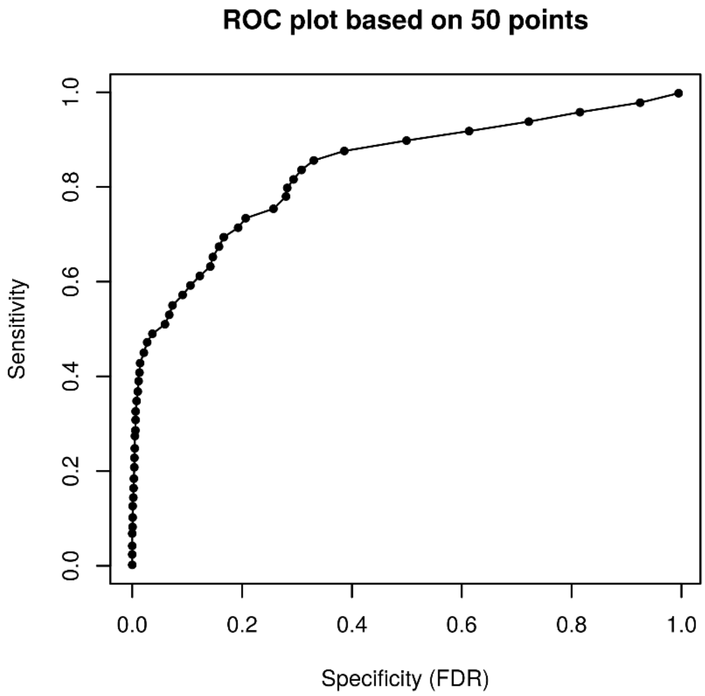


Showing 1 to 50 of 50 entries

Show 50 entries

ID	Threshold	Sensitivity	FDR
0	0.99942	0.002	0
1	0.99704	0.024	0
10	0.95932	0.208	0.0038
11	0.95822	0.228	0.004
12	0.95584	0.248	0.0046
13	0.95363	0.274	0.0048
14	0.95156	0.286	0.0058
15	0.95073	0.308	0.0062
16	0.94918	0.326	0.0062
17	0.94401	0.348	0.008
18	0.93793	0.368	0.0102
19	0.9339	0.39	0.0116
2	0.99431	0.042	0
20	0.93093	0.408	0.0132
21	0.9274	0.428	0.0144
22	0.92144	0.45	0.0212
23	0.91503	0.472	0.0272
24	0.90589	0.49	0.0368
25	0.88888	0.51	0.0596
26	0.88514	0.53	0.0674
27	0.88161	0.55	0.0734
28	0.8725	0.572	0.0918
29	0.8656	0.592	0.106
3	0.99409	0.068	0
30	0.85693	0.612	0.123
31	0.84885	0.632	0.1422
32	0.84762	0.652	0.1468
33	0.84511	0.674	0.158

The output table can be used to create a ROC curve for the visualization of the motif quality and for comparison of different motifs.

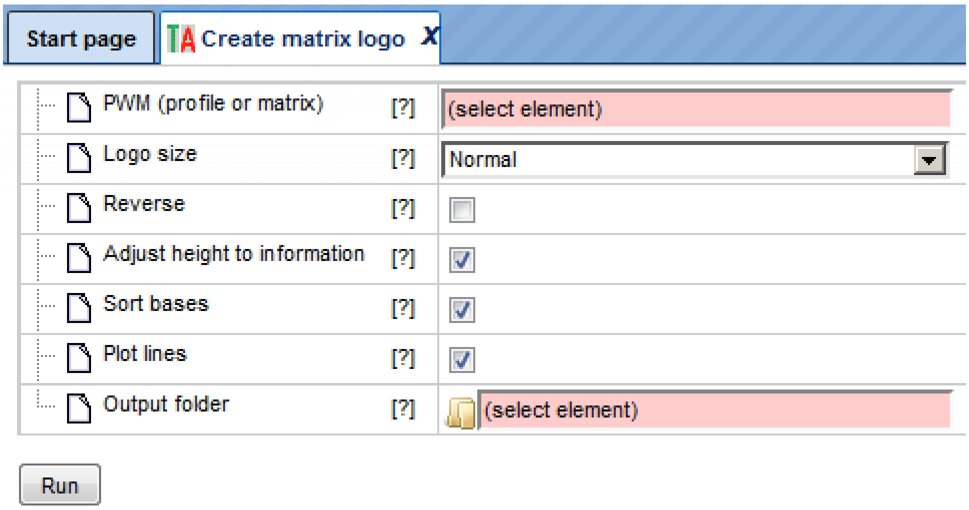


10.4.2. Create matrix logo

This tool creates logo representations for position weight or frequency matrices of transcription factor binding sites.

The input can be a profile with a set of matrices or a single matrix.

The input form is as shown below:



PWM (profile or matrix)	[?]	(select element)
Logo size	[?]	Normal
Reverse	[?]	□
Adjust height to information	[?]	☑
Sort bases	[?]	☑
Plot lines	[?]	☑
Output folder	[?]	(select element)

Each individual parameter is described below:

PWM (profile or matrix) – Specify the input profile or a single matrix. You can drag it from your project within the tree area and drop it in the pink box of the field **PWM**. Alternatively, you may click on the pink field "select element" and a new window will be opened, where you can select the input gene set as shown below.

Logo size – The method gives an option to select one of four different sizes for the Matrix logo image. It ranges from small to extra-large.

Reverse – Check this box to create logos for the reverse orientation. By default this box is unchecked.

Adjust height to information –Check this box to adjust total height of bases to information content of position.

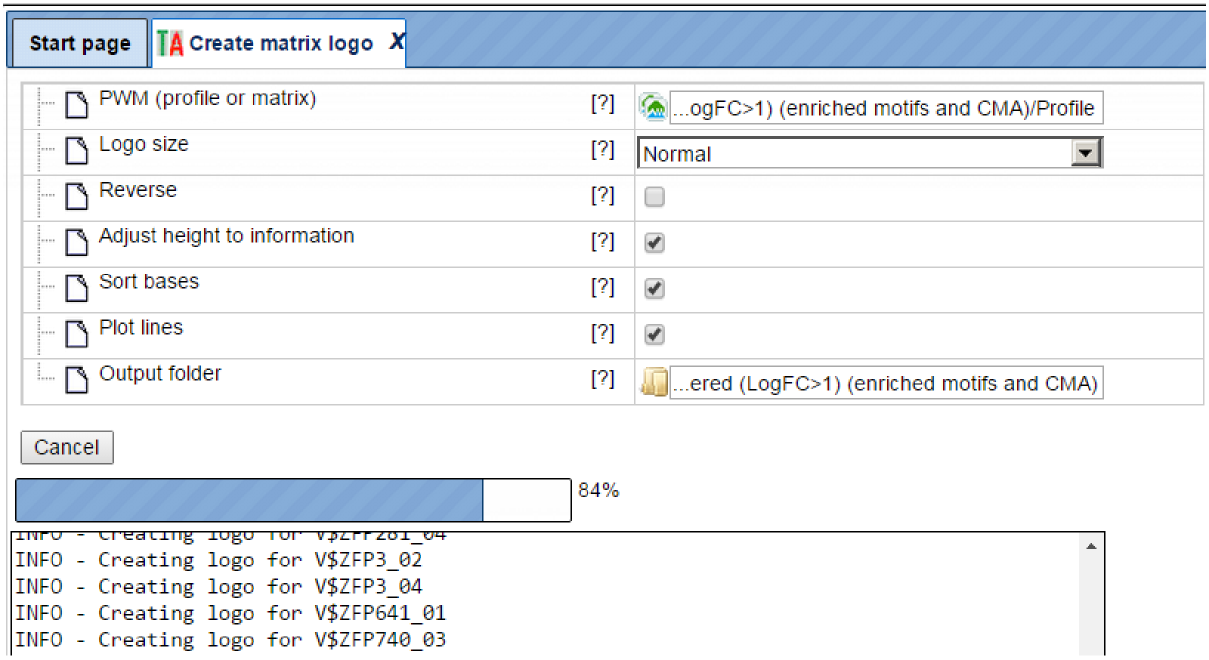
Sort bases –Check this box to sort bases, the most important on top.

Plot lines - Checked this box to draw lines behind bases partitioning plot region into four sections.

Output folder – Specify the name and path of the output folder for the created logos.

Here, we take a profile created by the workflow 'Identify enriched composite modules in promoters (TRANSFAC®)' as input.

Keeping all other parameters as default, the method runs as shown below:

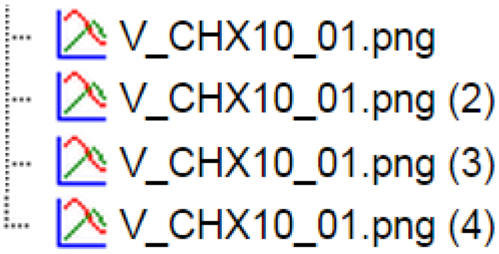


PWM (profile or matrix)	...ogFC>1) (enriched motifs and CMA)/Profile
Logo size	Normal
Reverse	□
Adjust height to information	☑
Sort bases	☑
Plot lines	☑
Output folder	...ered (LogFC>1) (enriched motifs and CMA)

84%

```
INFO - Creating logo for V$ZFP281_04
INFO - Creating logo for V$ZFP3_02
INFO - Creating logo for V$ZFP3_04
INFO - Creating logo for V$ZFP641_01
INFO - Creating logo for V$ZFP740_03
```

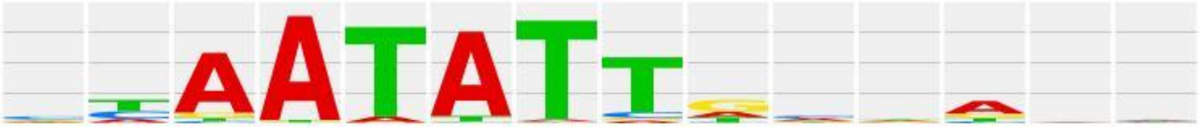
The output folder contains one PNG image for each matrix of the specified input. Existing files in the output folder are not overwritten. In case of name conflicts the tool suffixes a number to the file name as shown below:



V_CHX10_01.png
V_CHX10_01.png (2)
V_CHX10_01.png (3)
V_CHX10_01.png (4)

:

The matrix logo output image is as shown below:



Each matrix image can be exported in either .jpeg, .png, or .bmp file formats using the 'Export document' button.

10.4.3. Identify enriched TF sites in promoters

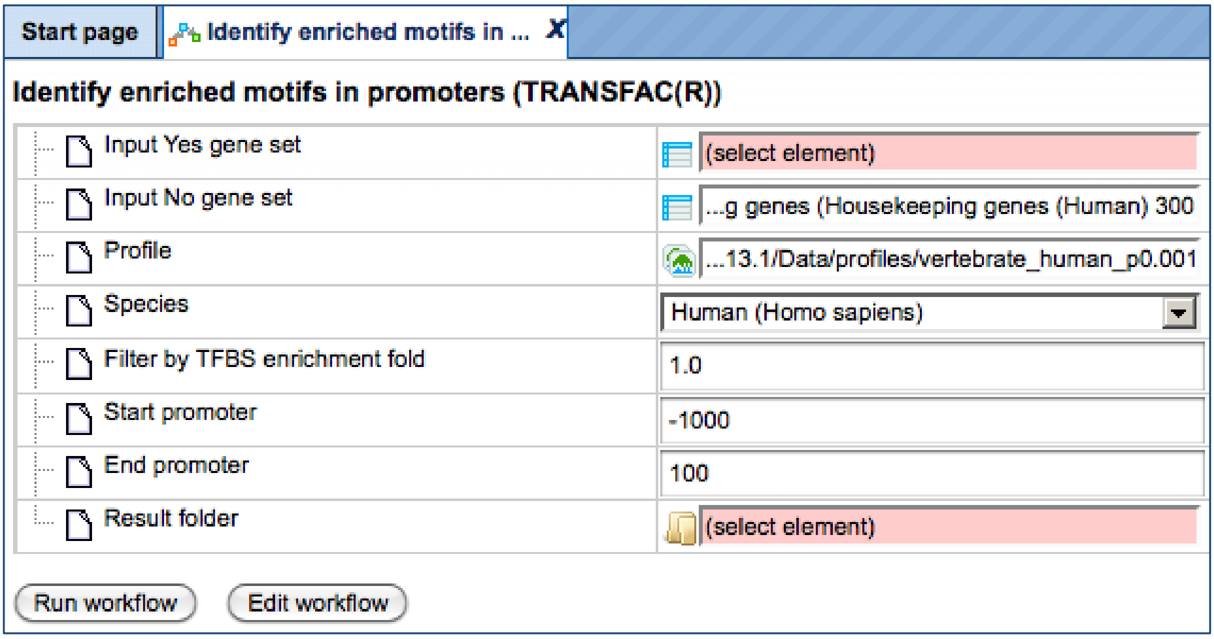
10.4.3.1. Version 2.0 (Adjusted p-values)

TRANSFAC®

This workflow is designed to find individual motifs enriched in the promoters of the input gene set as compared with a background set (No set). In the first part of the workflow, the enriched motifs are identified by the method *analyses/Methods/Site analysis/Search for enriched TFBSs (genes)*, icon . Please refer to section 20.1.4 for details on this particular analysis method. Filtered enriched motifs serve as a basis to construct a specific profile, and this profile is run on the promoters of the input gene set, method *analyses/Methods/Site analysis/Site search on gene set*, icon . Details about this individual method are given in section 6.1.2.

To launch the workflow, follow these steps:

Step1. Open the workflow input form from the Start page. It will open in the main Work Space and looks as shown below:



Identify enriched motifs in promoters (TRANSFAC(R))

Input Yes gene set	(select element)
Input No gene set	...g genes (Housekeeping genes (Human) 300
Profile	...13.1/Data/profiles/vertebrate_human_p0.001
Species	Human (Homo sapiens)
Filter by TFBS enrichment fold	1.0
Start promoter	-1000
End promoter	100
Result folder	(select element)

Step 2. Input the Yes set from the tree. You can either drag-and-drop or select the Yes set from the Tree area. Here, the set of up-regulated genes from the following *Examples* folder is used:

data/Examples/HCV infection in liver GSE31193, Affymetrix U133 Plus 2.0/Data/

The Yes set in this example contains 125 genes up-regulated in human liver cells treated with interferon-γ (IFNγ) as compared with non-treated cells.

Step 3. Similarly input the NO set from the tree area. By default the workflow uses a subset fo 300 genes randomly taken out of the human housekeeping genes. The default NO set can be found here:

http://genexplain-platform.com/bioumlweb/#de=data/Public/Data%20sets/Data/Housekeeping%20genes%20(Human)%20300

Here, the set of non-changed genes from the *Examples* folder is used:

data/Examples/HCV infection in liver GSE31193, Affymetrix U133 Plus 2.0/Data/

The No set in this example contains 237 genes whose expression was unchanged in human liver cells treated with IFNγ as compared with non-treated cells.

Step 4. Select the profile. This profile will be applied at the first part of the workflow for identification of the enriched motifs. The default profile is *vertebrate_human_p0.001* from the most recent TRANSFAC® release available. It can be found here:

http://genexplain-platform.com/bioumlweb/#de=databases/TRANSFAC(R)%202014.4/Data/profiles/vertebrate_human_p0.001

The number of matrices in the profile shown here is 4321.

Any other TRANSFAC® profile or user-specific profile can be selected. With a mouse click on the field **Profile**, a pop-up window will open, where a profile can be selected.

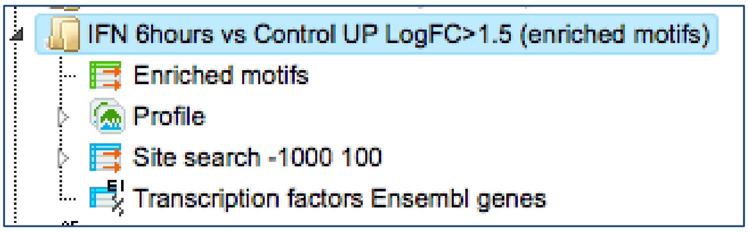
Step 5. After input of the Yes and No sets, the species (human, mouse or rat) is adjusted automatically. Verify the species shown in the species field.

Step 6. Filter by TFBS enrichment fold: In this field you can specify the enrichment fold (FE) to filter the motifs. By default it is 1.0, which means all motifs with FE>1.0 will be reported in the resulting table and the same motifs will serve to create a specific profile. If you want to use highly-enriched motifs, you can specify higher thresholds, e.g. 1.1, 1.2 etc, or even 2.0 or 3.0 depending on your Yes and No sets. It is recommended that you run it with default parameters first, check the results, and then run again with the desired filter value.

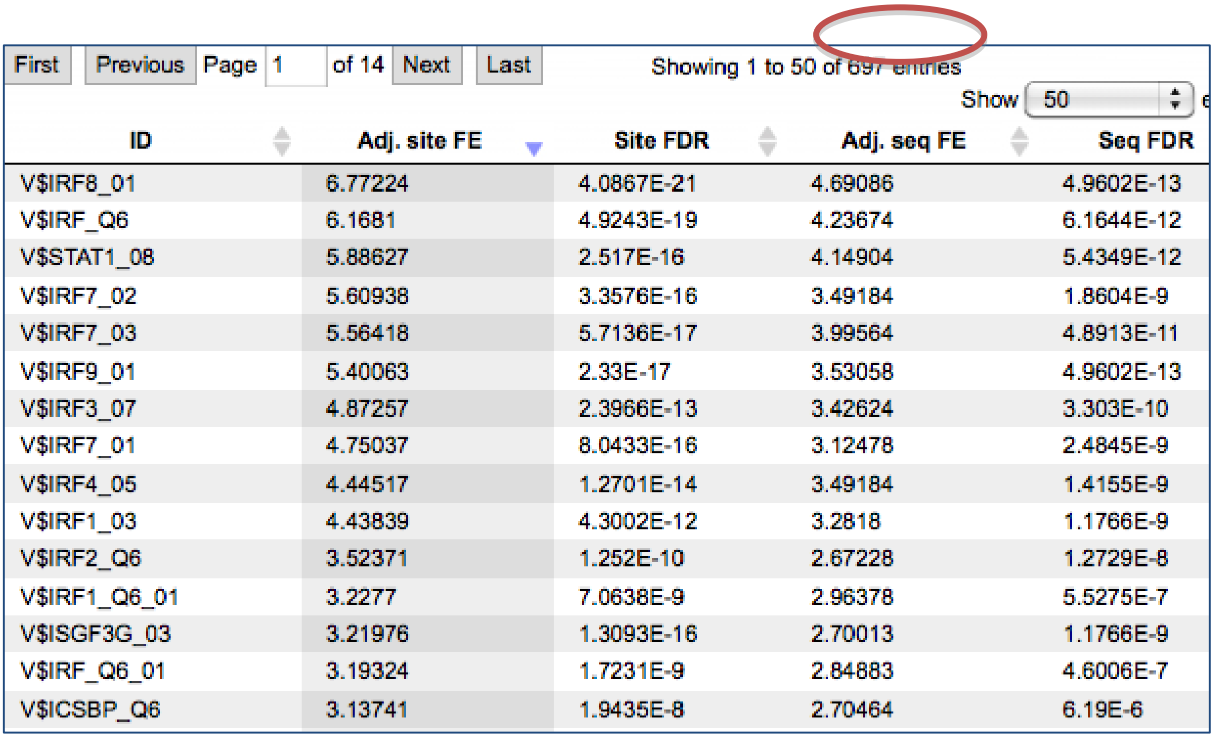
Step 7. Specify the promoter region relative to TSS as they are annotated in Ensembl. The default promoter region is -1000 to +100 relative to the TSS. You can edit the fields *Start promoter* and *End promoter* as required.

Step 8. Specify the result folder location and name and Press the button [Run workflow]. Wait till the workflow is completed.

Results.
The results folder consists of several files and folders as shown below:


- IFN 6hours vs Control UP LogFC>1.5 (enriched motifs)
  - Enriched motifs
  - Profile
  - Site search -1000 100
  - Transcription factors Ensembl genes

The table **Enriched Motifs** ( ) contains those site models, here TRANSFAC® matrices, which are enriched in the Yes set in comparison with the No set as shown below.



Showing 1 to 50 of 697 entries

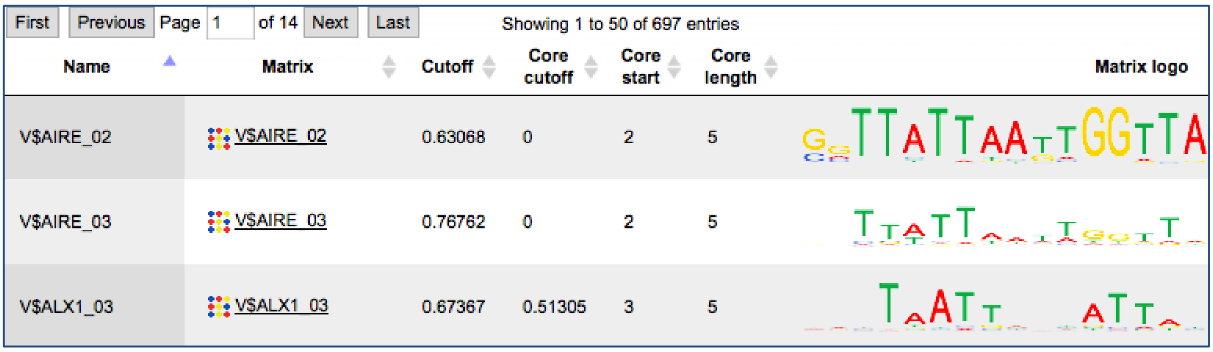
ID	Adj. site FE	Site FDR	Adj. seq FE	Seq FDR
V$IRF8_01	6.77224	4.0867E-21	4.69086	4.9602E-13
V$IRF_Q6	6.1681	4.9243E-19	4.23674	6.1644E-12
V$STAT1_08	5.88627	2.517E-16	4.14904	5.4349E-12
V$IRF7_02	5.60938	3.3576E-16	3.49184	1.8604E-9
V$IRF7_03	5.56418	5.7136E-17	3.99564	4.8913E-11
V$IRF9_01	5.40063	2.33E-17	3.53058	4.9602E-13
V$IRF3_07	4.87257	2.3966E-13	3.42624	3.303E-10
V$IRF7_01	4.75037	8.0433E-16	3.12478	2.4845E-9
V$IRF4_05	4.44517	1.2701E-14	3.49184	1.4155E-9
V$IRF1_03	4.43839	4.3002E-12	3.2818	1.1766E-9
V$IRF2_Q6	3.52371	1.252E-10	2.67228	1.2729E-8
V$IRF1_Q6_01	3.2277	7.0638E-9	2.96378	5.5275E-7
V$ISGF3G_03	3.21976	1.3093E-16	2.70013	1.1766E-9
V$IRF_Q6_01	3.19324	1.7231E-9	2.84883	4.6006E-7
V$ICSBP_Q6	3.13741	1.9435E-8	2.70464	6.19E-6

Each row of the output table represents the result for one PWM from the input profile. Only those PWMs with adj. site FE >1 are included in the output. For details on the output columns please refer to section 20.1.4.1. Recommended sorting, as shown on the screenshot above, is by column *Adj. site FE* (adjusted fold enrichment for sites) with the highest values on top.

Please note that out of 4307 matrices in the initial profile, hits for 697 matrices are enriched with adj. site FE >1. These matrices are considered to create profiles specific for the input Yes and No sets.

Motifs for IRF, STAT, ICSBP transcription factors are highly enriched, with adj. site FE >2, as shown in the screenshot above. This is a very relevant result considering that here the effect of IFNγ on liver cells is studied.

The table **Profile** ( ) presents details for PWMs with *adj. site FE >1*.

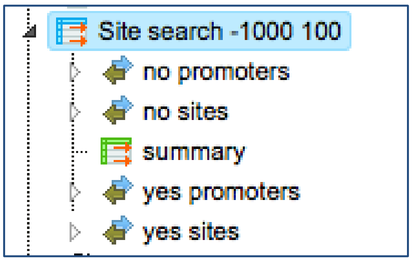


First | Previous | Page 1 of 14 | Next | Last   Showing 1 to 50 of 697 entries

Name	Matrix	Cutoff	Core cutoff	Core start	Core length	Matrix logo
V$AIRE_02	V$AIRE_02	0.63068	0	2	5	GTTATTAATTGGTTA
V$AIRE_03	V$AIRE_03	0.76762	0	2	5	TTATTATTGTT
V$ALX1_03	V$ALX1_03	0.67367	0.51305	3	5	TAATT ATTA

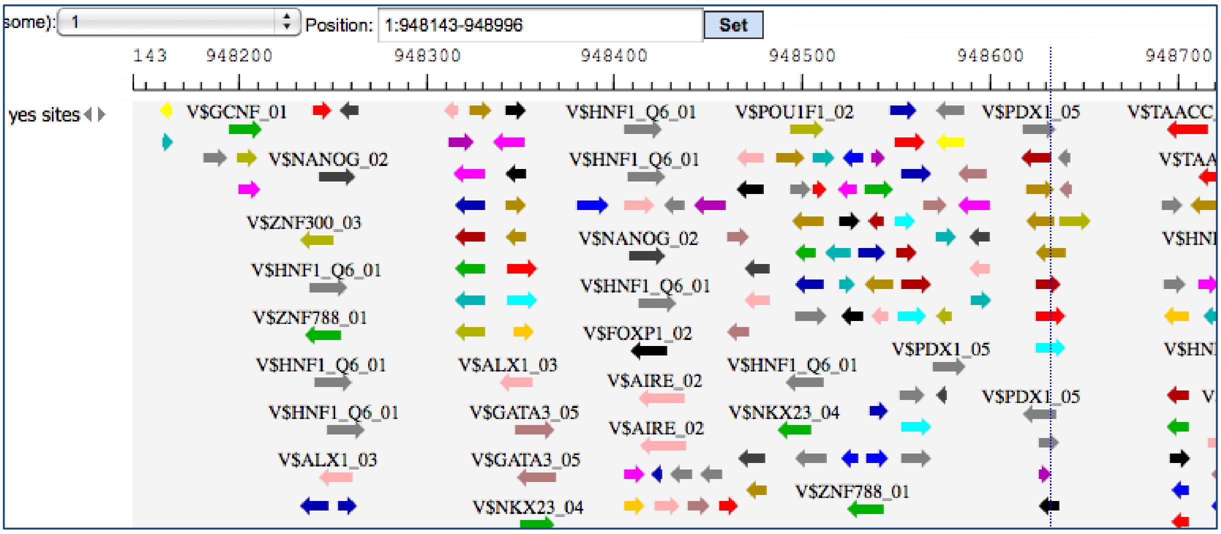
This profile is an intermediate result of the workflow and is used further for *Site search on gene set* analysis.

Site search analysis output ( ) serves to visualize enriched motifs in the promoters. This folder contains four tracks ( ):



- Site search -1000 100
  - no promoters
  - no sites
  - summary
  - yes promoters
  - yes sites

Each track can be opened in the genome browser by double-clicking. A visualization of the track *yes sites* is shown below:



some): 1   Position: 1:948143-948996   Set

The same track can be opened as a table; for this use right mouse click on the track name in the tree area or Ctrl +mouse click for Mac users.

With the same menu, you can apply other functions to the selected track, e.g. export it in available formats or delete.

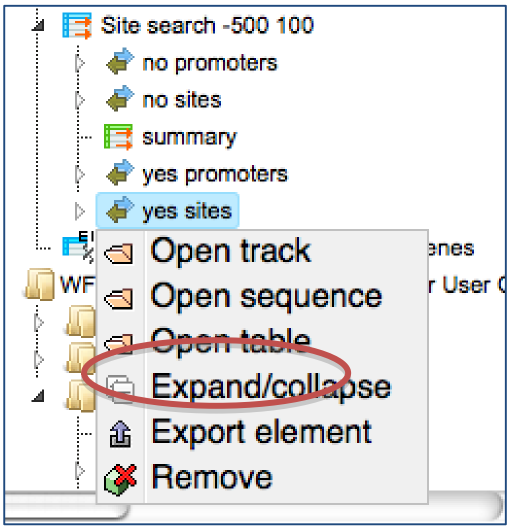
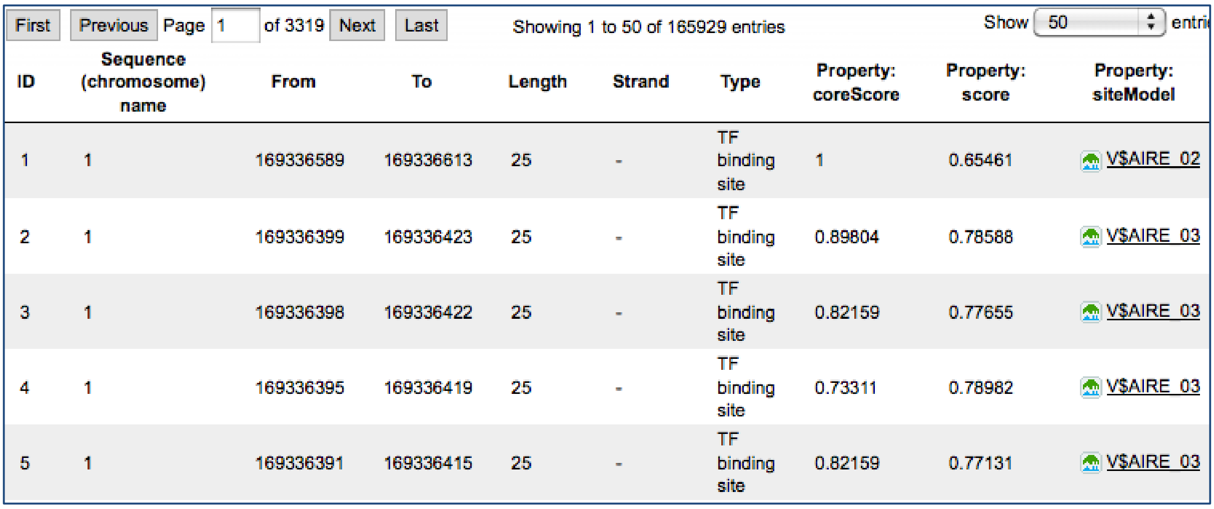


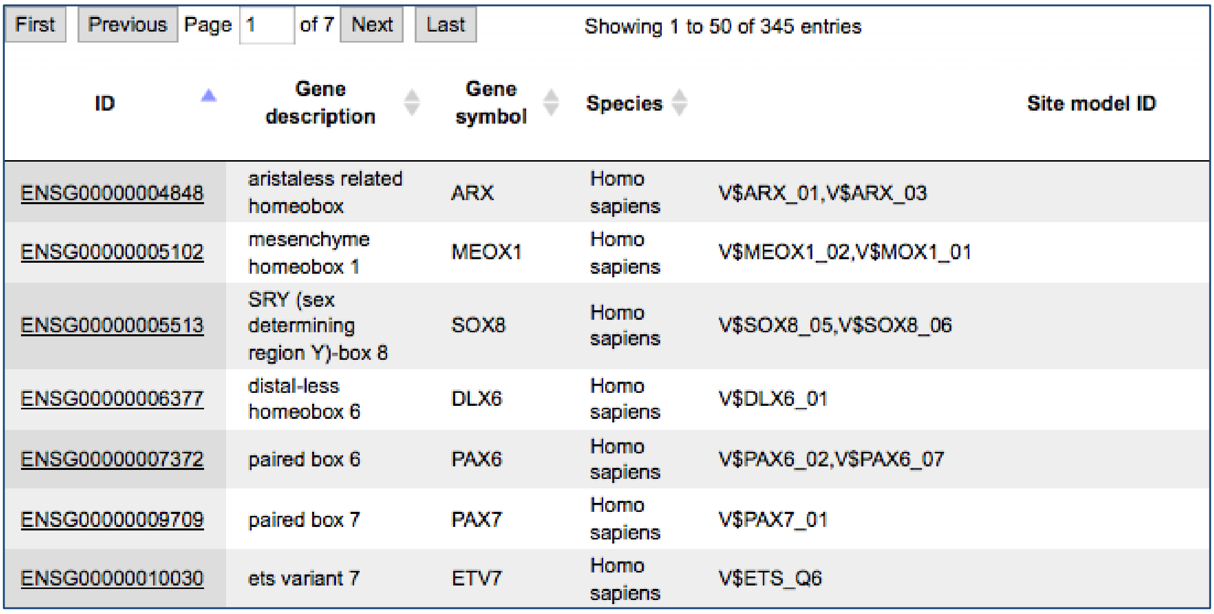
Table view on the track *yes sites* is the following:



First Previous Page 1 of 3319 Next Last Showing 1 to 50 of 165929 entries Show 50 entri

ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: coreScore	Property: score	Property: siteModel
1	1	169336589	169336613	25	-	TF binding site	1	0.65461	V$AIRE_02
2	1	169336399	169336423	25	-	TF binding site	0.89804	0.78588	V$AIRE_03
3	1	169336398	169336422	25	-	TF binding site	0.82159	0.77655	V$AIRE_03
4	1	169336395	169336419	25	-	TF binding site	0.73311	0.78982	V$AIRE_03
5	1	169336391	169336415	25	-	TF binding site	0.82159	0.77131	V$AIRE_03

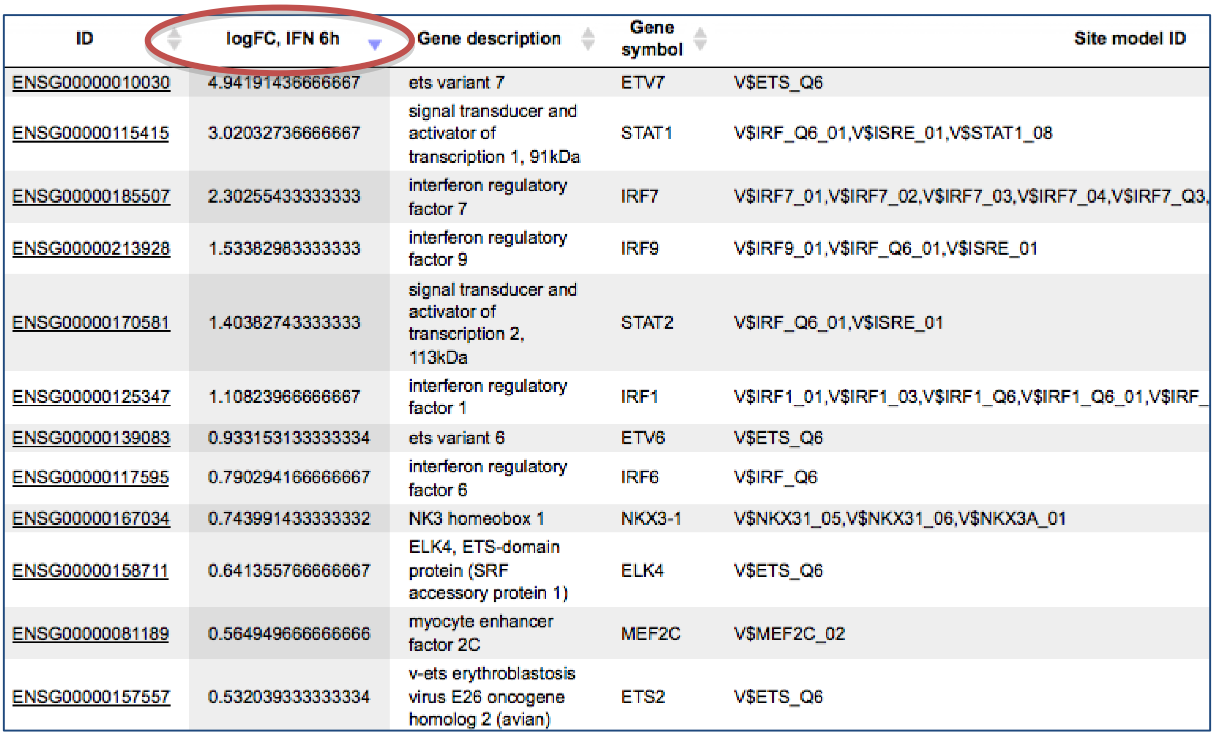
The output table *Transcription factors Ensembl genes* is a list of transcription factors linked to the enriched motifs. For each transcription factor, the Ensembl gene ID is provided, as well gene description, HGNC gene symbol, species, and site model (TRANSFAC® PWM name).



Showing 1 to 50 of 345 entries

ID	Gene description	Gene symbol	Species	Site model ID
ENSG00000004848	aristaless related homeobox	ARX	Homo sapiens	VARX_01,VARX_03
ENSG00000005102	mesenchyme homeobox 1	MEOX1	Homo sapiens	V$MEOX1_02,V$MOX1_01
ENSG00000005513	SRY (sex determining region Y)-box 8	SOX8	Homo sapiens	V$SOX8_05,V$SOX8_06
ENSG00000006377	distal-less homeobox 6	DLX6	Homo sapiens	V$DLX6_01
ENSG00000007372	paired box 6	PAX6	Homo sapiens	V$PAX6_02,V$PAX6_07
ENSG00000009709	paired box 7	PAX7	Homo sapiens	V$PAX7_01
ENSG00000010030	ets variant 7	ETV7	Homo sapiens	V$ETS_Q6

This table can be further annotated to add a column with expression values, as shown below. Details for annotation of the tables are given in the section 16.1.1.



ID	logFC, IFN 6h	Gene description	Gene symbol	Site model ID
ENSG00000010030	4.94191436666667	ets variant 7	ETV7	V$ETS_Q6
ENSG00000115415	3.02032736666667	signal transducer and activator of transcription 1, 91kDa	STAT1	VIRF_Q6_01,VISRE_01,V$STAT1_08
ENSG00000185507	2.30255433333333	interferon regulatory factor 7	IRF7	V$IRF7_01,V$IRF7_02,V$IRF7_03,V$IRF7_04,V$IRF7_Q3,
ENSG00000213928	1.53382983333333	interferon regulatory factor 9	IRF9	V$IRF9_01,V$IRF_Q6_01,V$ISRE_01
ENSG00000170581	1.40382743333333	signal transducer and activator of transcription 2, 113kDa	STAT2	VIRF_Q6_01,VISRE_01
ENSG00000125347	1.10823966666667	interferon regulatory factor 1	IRF1	V$IRF1_01,V$IRF1_03,V$IRF1_Q6,V$IRF1_Q6_01,V$IRF_
ENSG00000139083	0.933153133333334	ets variant 6	ETV6	V$ETS_Q6
ENSG00000117595	0.790294166666667	interferon regulatory factor 6	IRF6	V$IRF_Q6
ENSG00000167034	0.743991433333332	NK3 homeobox 1	NKX3-1	V$NKX31_05,V$NKX31_06,V$NKX3A_01
ENSG00000158711	0.641355766666667	ELK4, ETS-domain protein (SRF accessory protein 1)	ELK4	V$ETS_Q6
ENSG00000081189	0.564949666666666	myocyte enhancer factor 2C	MEF2C	V$MEF2C_02
ENSG00000157557	0.532039333333334	v-ets erythroblastosis virus E26 oncogene homolog 2 (avian)	ETS2	V$ETS_Q6

Twelve TFs are found to be highly up-regulated under the same conditions, all members of Ets, STAT, IRF, MEF2 families. The role of these TFs in regulation of the input *Yes genes* is suggested by two independent lines of evidence: first, genes encoding these TFs are highly up-regulated, and second, their binding motifs are significantly enriched in the promoters of Yes genes.

Note. This workflow is available together with a valid TRANSFAC® license. Please, feel free to ask for details (info@genexplain.com).

GTRD

This workflow is designed to search for putative transcription factor binding sites, TFBS, in the promoters of an input gene set. It is very similar to the workflow described above in Section 10.4. The only difference is in the PWM library applied. Here, site search is done with the help of the GTRD library (see 19.8 for further details about this library).

For the input form and description of the results folder, please refer to Section 10.4.

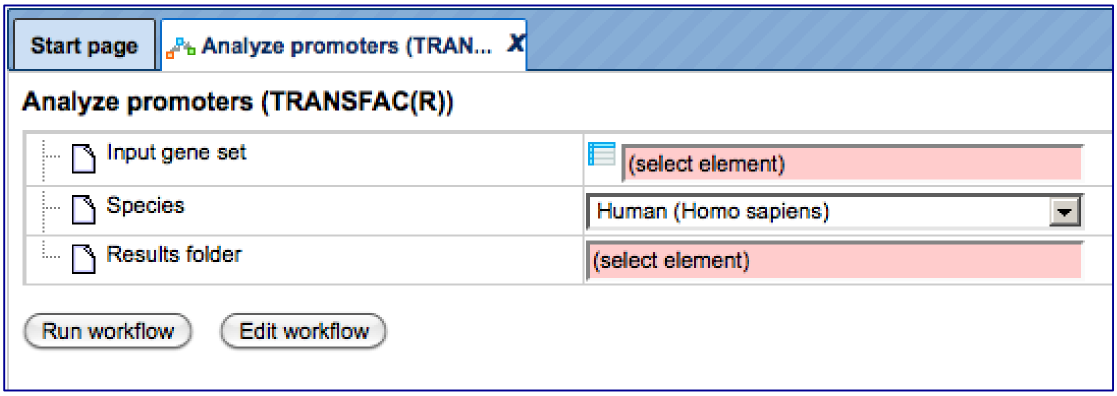
10.4.3.2. Version 1.2 (Classical)

TRANSFAC®

This workflow is designed to search for putative transcription factor binding sites, TFBS, in the promoters of an input gene set. Site search is done with the help of the TRANSFAC® library of positional weight matrices, PWMs, namely with the profile vertebrate_non_redundant_minSUM.

To launch the workflow, follow these steps:

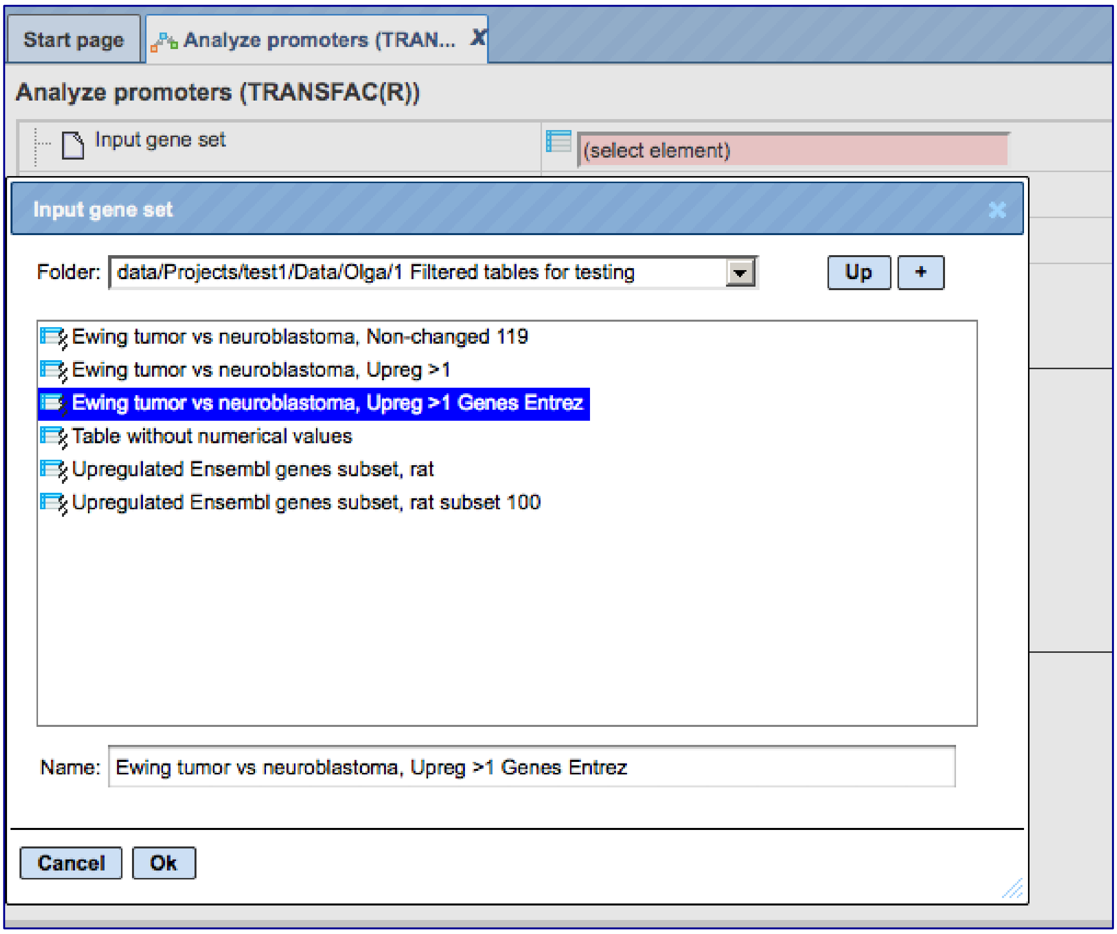
Step 1. Open the workflow input form from the Start page. It will open in the main Work Space and looks as shown below:



Analyze promoters (TRANSFAC(R))

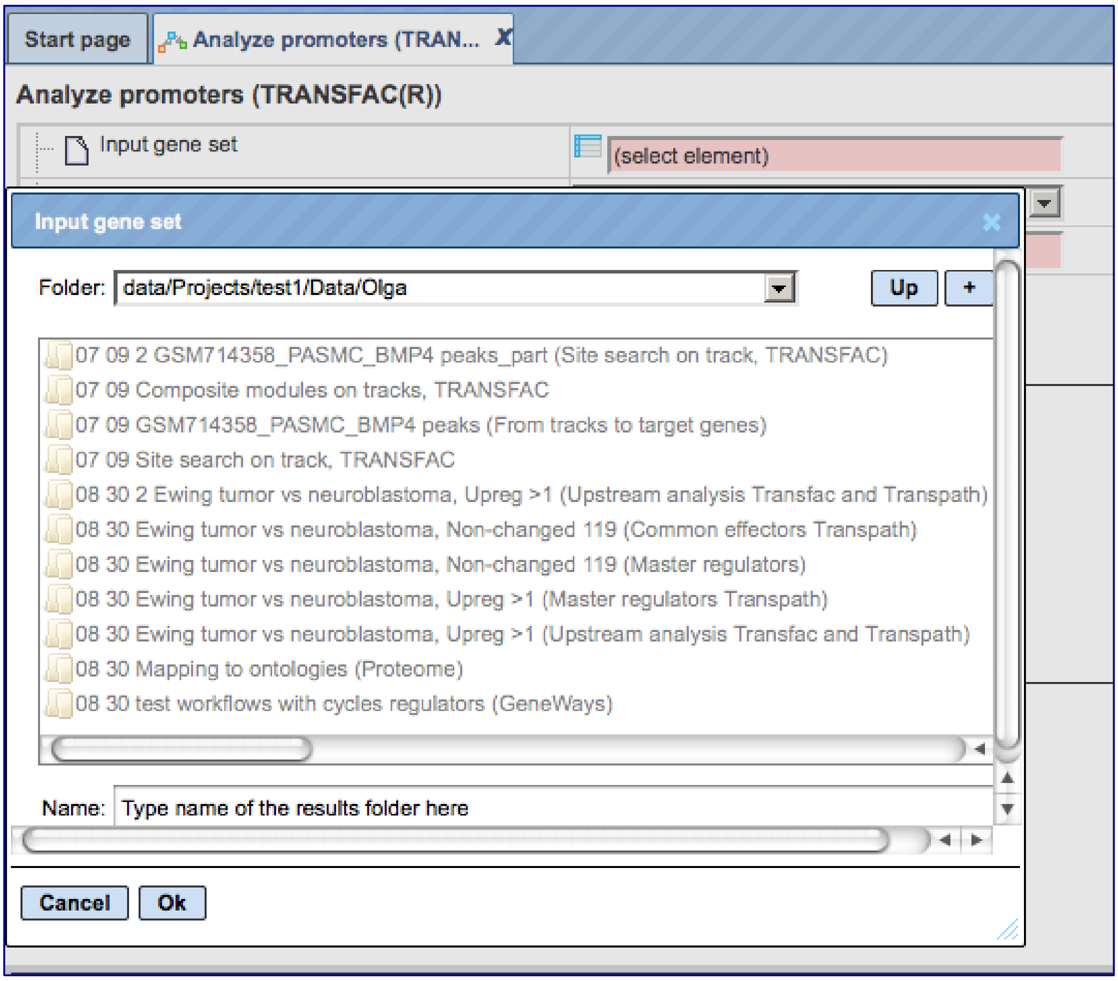
Input gene set	(select element)
Species	Human (Homo sapiens)
Results folder	(select element)

Step 2. Specify input gene or protein set in the field **Input gene set**. The input table contains the genes under study, and it is also called the 'Yes' set. To specify a gene set, you can drag & drop it from your project within the Tree Area, and drop it in the pink box of the field **Input gene set**. Alternatively, you may click on the pink field "select element", and a new window will open, where you can select the input gene set as shown below. After you have selected the gene set, press [Ok].


Start page
Analyze promoters (TRAN...
Analyze promoters (TRANSFAC(R))
Input gene set
(select element)
Input gene set
Folder: data/Projects/test1/Data/Olga/1 Filtered tables for testing
Up
+
Ewing tumor vs neuroblastoma, Non-changed 119
Ewing tumor vs neuroblastoma, Upreg >1
Ewing tumor vs neuroblastoma, Upreg >1 Genes Entrez
Table without numerical values
Upregulated Ensembl genes subset, rat
Upregulated Ensembl genes subset, rat subset 100
Name: Ewing tumor vs neuroblastoma, Upreg >1 Genes Entrez
Cancel
Ok

Step 3. Specify the biological species of the input set in the field **Species** by selecting the required species from the drop-down menu.

Step 4. Define where the folder with the results should be located in the tree. You can do so by clicking on the pink field "select element" in the field **Results folder**, and a new window will open where you can select the location of the results folder and define its name as shown below.



Analyze promoters (TRANSFAC(R))

Input gene set (select element)

Input gene set

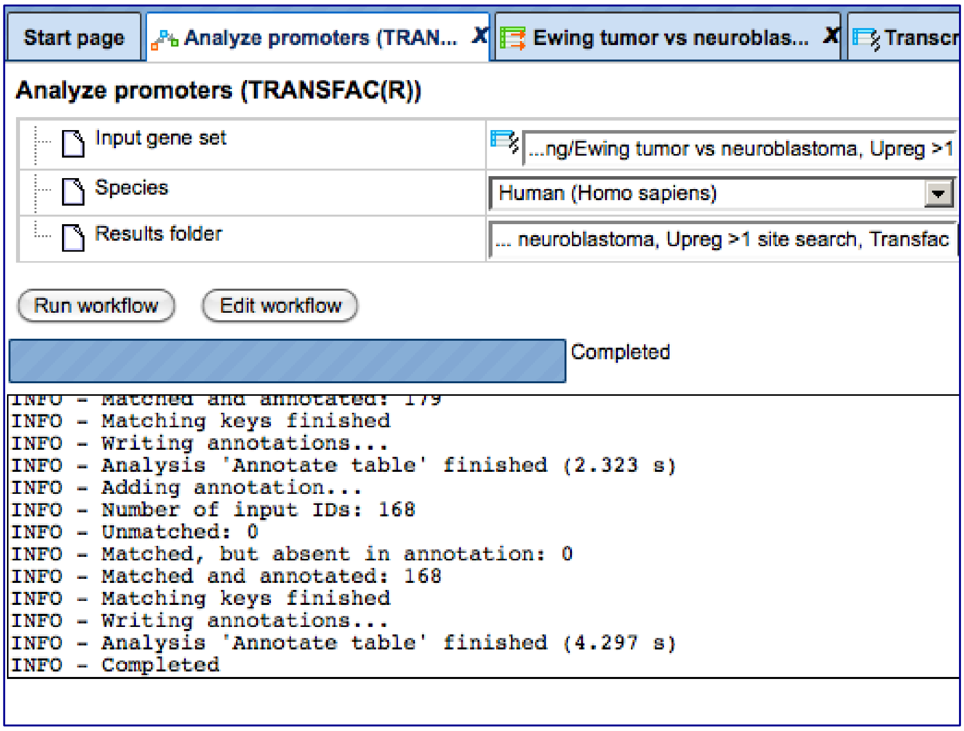
Folder: data/Projects/test1/Data/Olga

07 09 2 GSM714358_PASMC_BMP4 peaks_part (Site search on track, TRANSFAC)
07 09 Composite modules on tracks, TRANSFAC
07 09 GSM714358_PASMC_BMP4 peaks (From tracks to target genes)
07 09 Site search on track, TRANSFAC
08 30 2 Ewing tumor vs neuroblastoma, Upreg >1 (Upstream analysis Transfac and Transpath)
08 30 Ewing tumor vs neuroblastoma, Non-changed 119 (Common effectors Transpath)
08 30 Ewing tumor vs neuroblastoma, Non-changed 119 (Master regulators)
08 30 Ewing tumor vs neuroblastoma, Upreg >1 (Master regulators Transpath)
08 30 Ewing tumor vs neuroblastoma, Upreg >1 (Upstream analysis Transfac and Transpath)
08 30 Mapping to ontologies (Proteome)
08 30 test workflows with cycles regulators (GeneWays)

Name: Type name of the results folder here

Step 5. Press the [Run workflow] button.

Wait until the workflow is completed, which is shown below:



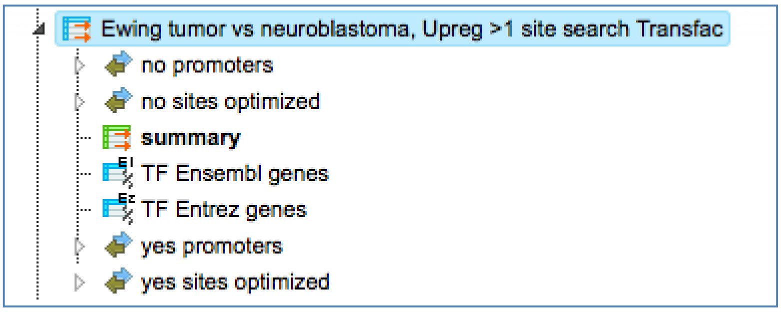
Analyze promoters (TRANSFAC(R))

Input gene set ...ng/Ewing tumor vs neuroblastoma, Upreg >1
Species Human (Homo sapiens)
Results folder ... neuroblastoma, Upreg >1 site search, Transfac

Completed

```
INFO - Matched and annotated: 179
INFO - Matching keys finished
INFO - Writing annotations...
INFO - Analysis 'Annotate table' finished (2.323 s)
INFO - Adding annotation...
INFO - Number of input IDs: 168
INFO - Unmatched: 0
INFO - Matched, but absent in annotation: 0
INFO - Matched and annotated: 168
INFO - Matching keys finished
INFO - Writing annotations...
INFO - Analysis 'Annotate table' finished (4.297 s)
INFO - Completed
```

The results folder contains several files:



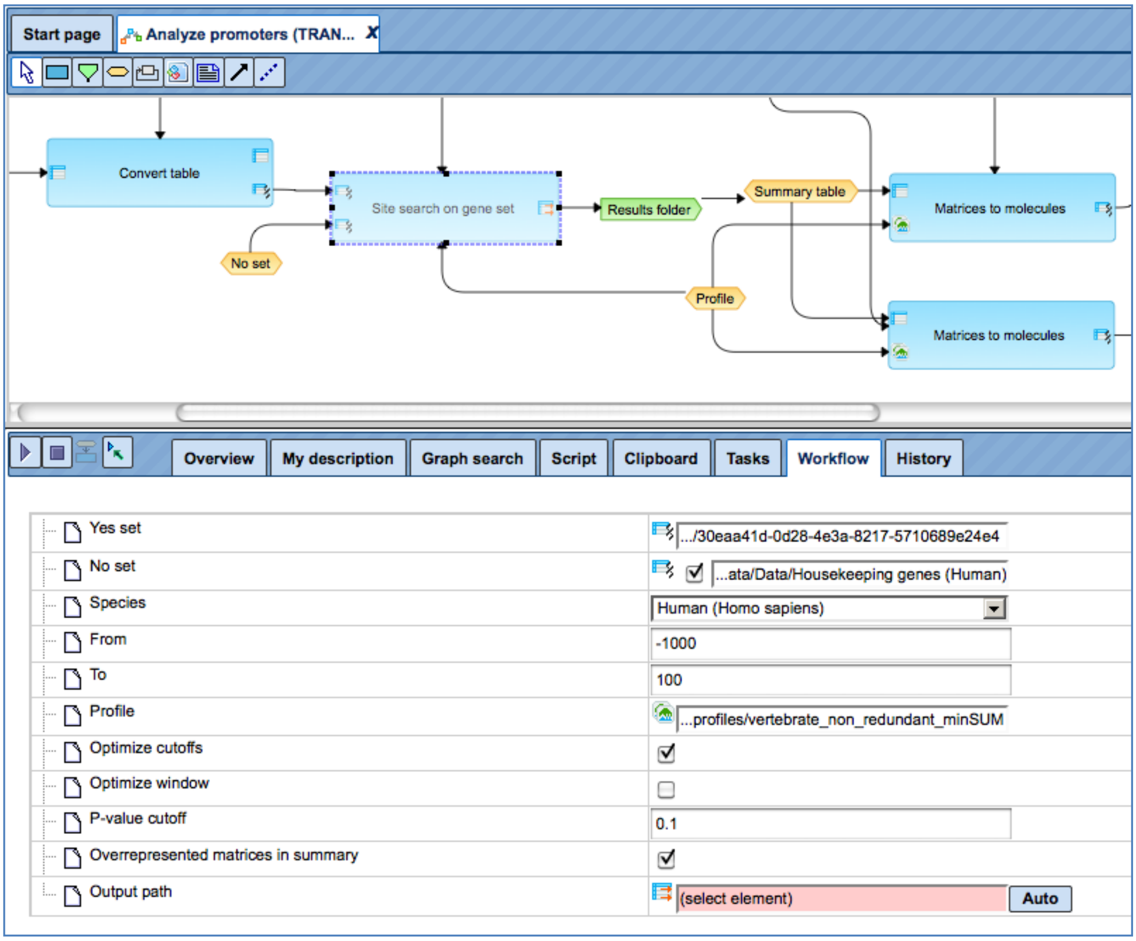
Ewing tumor vs neuroblastoma, Upreg >1 site search Transfac
- no promoters
- no sites optimized
- summary
- TF Ensembl genes
- TF Entrez genes
- yes promoters
- yes sites optimized

The tables called *summary* (▤), *TF Ensembl genes*, and *TF Entrez genes* are opened automatically in the work area as soon as the workflow is completed.

For more details about the results, please refer to Section 20.1.4.

Tip. You can easily create a similar workflow with parameter values adjusted to your needs. For example, you can select another profile from the list of available TRANSFAC® profiles, or specify different promoter positions relative to the TSS (default is -1000 to +100).

To do this, you need first to open the workflow in the “Edit workflow” mode, and save a copy in your project area. The [Edit workflow] button is located near the button [Run workflow] (see above, Step 1). Upon clicking on [Edit workflow], the workflow diagram will open in the Work Space, and you can select one of the analyses you would like to modify. For the screenshot below the "Site search on gene set" analysis was selected, and in the Operations Field, on the tab "Workflow", all the parameters are visible. Under this mode, you can either check what the default parameters are, or modify them according to your needs.



Yes set	.../30eaa41d-0d28-4e3a-8217-5710689e24e4
No set	☑ ...ata/Data/Housekeeping genes (Human)
Species	Human (Homo sapiens)
From	-1000
To	100
Profile	...profiles/vertebrate_non_redundant_minSUM
Optimize cutoffs	☑
Optimize window	□
P-value cutoff	0.1
Overrepresented matrices in summary	☑
Output path	(select element)

Note. This workflow is available together with a valid TRANSFAC® license. Please, feel free to ask for details (info@genexplain.com).

GTRD

This workflow is designed to search for putative transcription factor binding sites, TFBS, in the promoters of an input gene set. It is very similar to the workflow described above in Section 10.4. The only difference is in the PWM library applied. Here, site search is done with the help of the GTRD library (see 19.8 for further details about this library).

For the input form and description of the results folder, please refer to Section 10.4.

10.4.4. Identify composite modules in promoters

10.4.4.1. Version 2.0 (Adjusted p-values) with TRANSFAC®

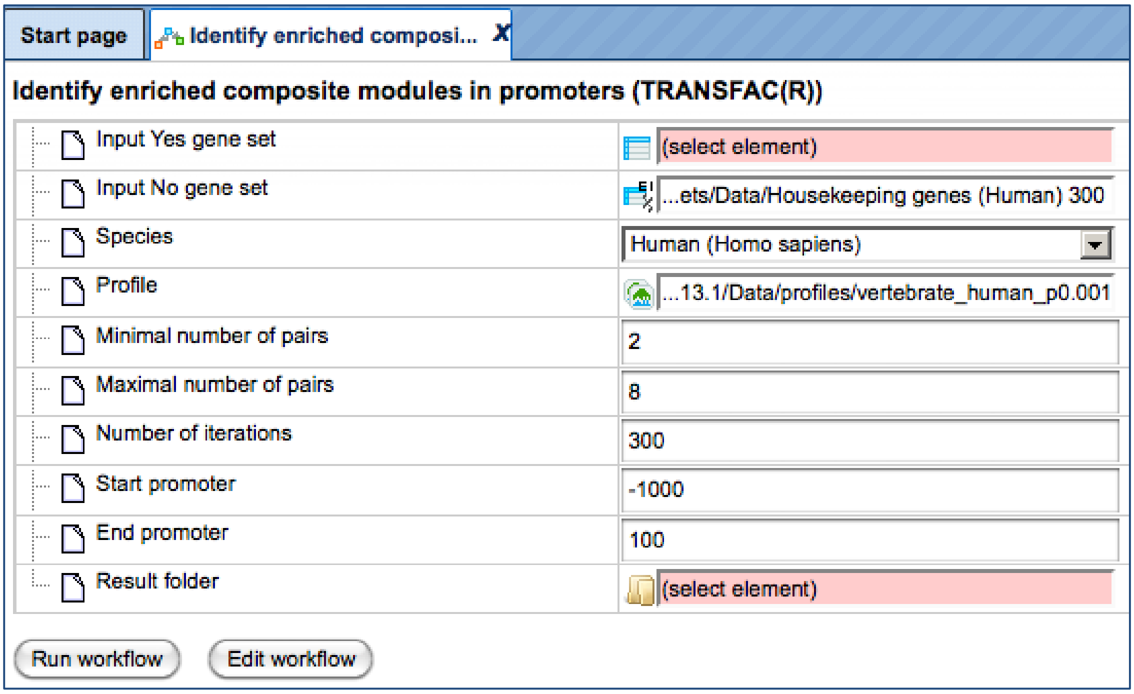
This workflow is designed to find pairs of sites in the promoters of the input gene set. This workflow enables the identification of combinations of several enriched TFBSs in the promoters of the genes under study (Yes-set). The resulting composite module differentiates the Yes-set from a background set (No-set).

In the first part of the workflow, the enriched motifs are identified by the method *analyses/Methods/Site analysis/Search for enriched TFBSs (genes)*, icon . Please refer to Section 20.1.4.1 for details on this individual analysis method.

Motifs with an enrichment of >1.0 fold serve as a basis for constructing a specific profile, and this profile is run on the promoters of the input gene set, method *analyses/Methods/Site analysis/Site search on gene set*, icon . Details about this individual method are given in the section 20.1.2. In the second part of this workflow, composite modules are identified based on the enriched TFBSs. For more details about CMA analysis refer to section 20.1.5.

To launch the workflow, follow these steps:

Step1. Open the workflow input form from the Start page. It will open in the main Work Space and looks as shown below:



Identify enriched composite modules in promoters (TRANSFAC(R))

Input Yes gene set	(select element)
Input No gene set	...ets/Data/Housekeeping genes (Human) 300
Species	Human (Homo sapiens)
Profile	...13.1/Data/profiles/vertebrate_human_p0.001
Minimal number of pairs	2
Maximal number of pairs	8
Number of iterations	300
Start promoter	-1000
End promoter	100
Result folder	(select element)

Step 2. Input the Yes set from the tree. You can either drag-and-drop or select the Yes set from the Tree area. Here, the set of up-regulated genes from the following *Examples* folder is used:

data/Examples/HCV infection in liver GSE31193, Affymetrix U133 Plus 2.0/Data/

The Yes set in this example contains 125 genes up-regulated in human liver cells treated with interferon-γ (IFNγ) as compared with non-treated cells.

Step 3. Similarly input the NO set from the tree area. By default the workflow uses a subset of 300 genes randomly taken out of the human housekeeping genes. Here, the set of non-changed genes from the same *Examples* folder is used.

Step 4. After input of the Yes and No sets, the species (human, mouse or rat) is adjusted automatically. Verify the species shown in the species field.

Step 5. Select the profile. The selected profile will be applied at the first part of the workflow for identification of enriched motifs. The default profile is *vertebrate_human_p0.001* from the most recent TRANSFAC® release available. Any other TRANSFAC® profile or user-specific profile can be chosen. With a mouse click on the field *Profile*, a pop-up window will open, where a profile can be selected. The number of matrices in the default profile, which is used here, is 4307.

Step 6. Set up parameters for the composite module search. This workflow identifies pairs of sites. By default, the minimum and maximum numbers of pairs are given as 2 and 8. You can change these parameters according to the number of pairs you aim to identify. The number of iterations of the genetic algorithm is 300 by default, and can be adapted as required.

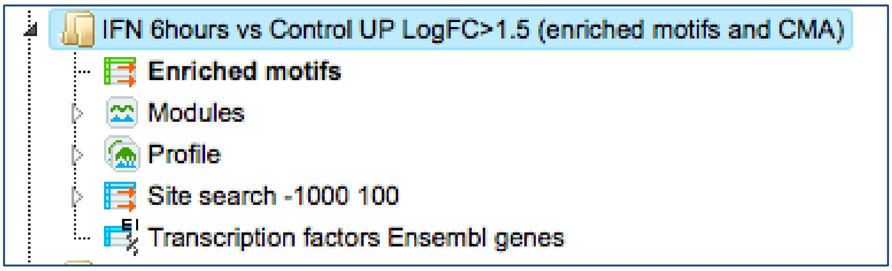
Step 7. Specify the promoter region relative to TSS as they are annotated in Ensembl. The default promoter region is -1000 to +100 relative to the TSS. You can edit the fields *Start promoter* and *End promoter* as required.

Step 8. Specify the result folder location and name and Press the button [Run workflow]. Wait till the workflow is completed.

Note. This workflow may take more time depending on the size of the Yes and No sets and on the number of iterations. The recommended size of the Yes set is 150 genes maximum, and the recommended size of the No set is 300 genes maximum. The maximum recommended number of iterations is 300.

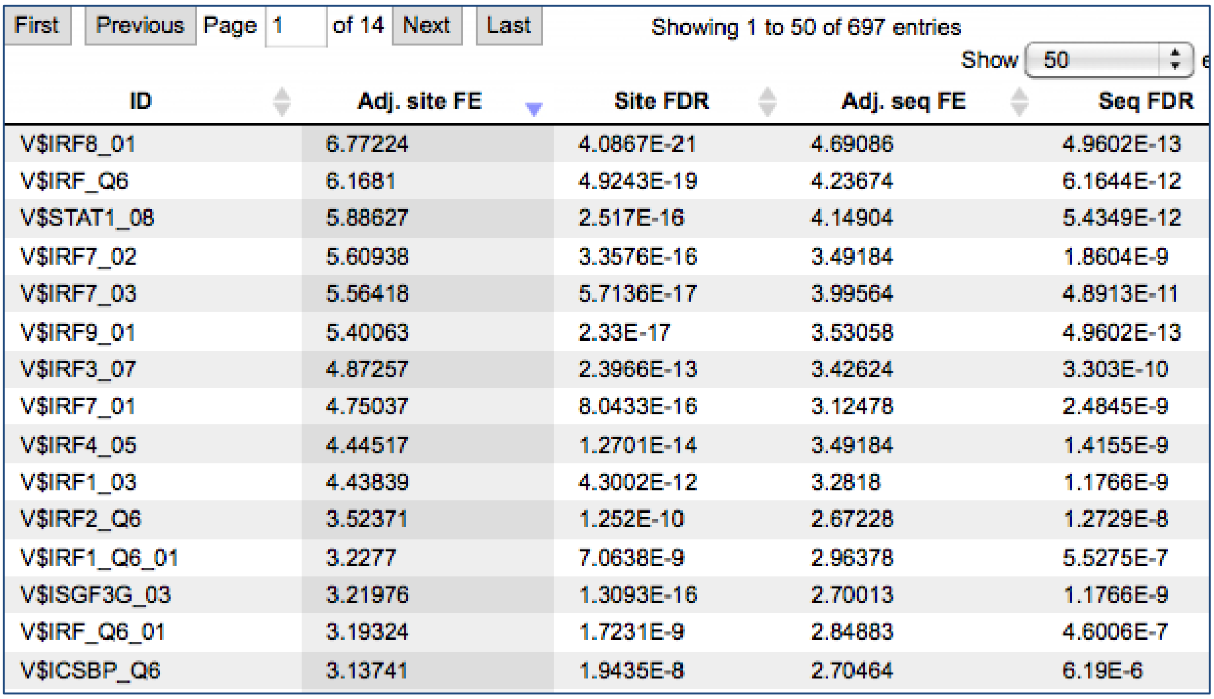
Results

The results folder consists of several folders and files as shown below:



IFN 6hours vs Control UP LogFC>1.5 (enriched motifs and CMA)
- Enriched motifs
- Modules
- Profile
- Site search -1000 100
- Transcription factors Ensembl genes

The table **Enriched Motifs** ( ) contains those site models, here TRANSFAC® matrices, which are enriched in the Yes set as compared to the No set as shown below.



First | Previous | Page 1 of 14 | Next | Last   Showing 1 to 50 of 697 entries   Show 50

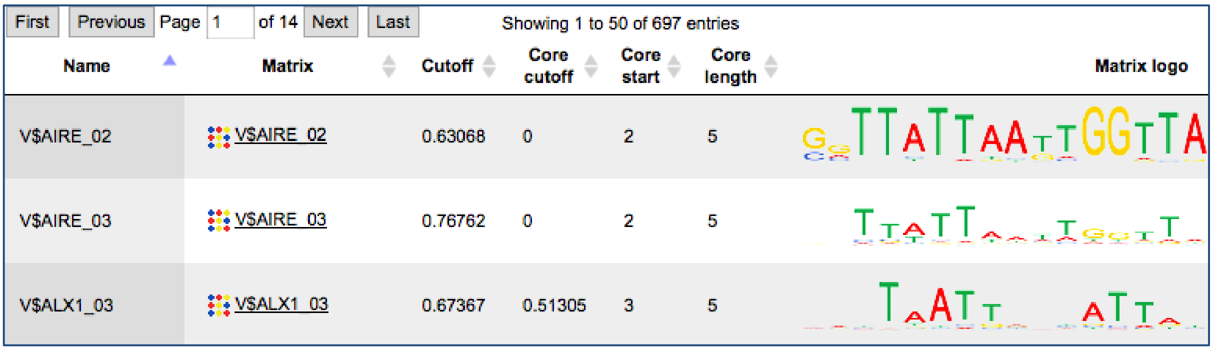
ID	Adj. site FE	Site FDR	Adj. seq FE	Seq FDR
V$IRF8_01	6.77224	4.0867E-21	4.69086	4.9602E-13
V$IRF_Q6	6.1681	4.9243E-19	4.23674	6.1644E-12
V$STAT1_08	5.88627	2.517E-16	4.14904	5.4349E-12
V$IRF7_02	5.60938	3.3576E-16	3.49184	1.8604E-9
V$IRF7_03	5.56418	5.7136E-17	3.99564	4.8913E-11
V$IRF9_01	5.40063	2.33E-17	3.53058	4.9602E-13
V$IRF3_07	4.87257	2.3966E-13	3.42624	3.303E-10
V$IRF7_01	4.75037	8.0433E-16	3.12478	2.4845E-9
V$IRF4_05	4.44517	1.2701E-14	3.49184	1.4155E-9
V$IRF1_03	4.43839	4.3002E-12	3.2818	1.1766E-9
V$IRF2_Q6	3.52371	1.252E-10	2.67228	1.2729E-8
V$IRF1_Q6_01	3.2277	7.0638E-9	2.96378	5.5275E-7
V$ISGF3G_03	3.21976	1.3093E-16	2.70013	1.1766E-9
V$IRF_Q6_01	3.19324	1.7231E-9	2.84883	4.6006E-7
V$ICSBP_Q6	3.13741	1.9435E-8	2.70464	6.19E-6

Each row of the output table represents the result for one PWM from the input profile. Only those PWMs with adj. site FE >1 are included in the output. For details on the output columns please refer to section 9.5.1. Recommended sorting, as shown in the screenshot above, is done by highest *Adj. site FE* (adjusted fold enrichment for sites).

Please note that out of 4307 matrices in the initial profile, hits for 697 matrices are enriched with adj. site FE >1. These matrices are considered to create profiles specific for the input Yes and No sets.

Motifs for IRF, STAT, ICSBP transcription factors are highly enriched, with adj. site FE >2, as shown in the screenshot above. This is a very relevant result considering that here the effect of IFNγ on liver cells is studied.

The table **Profile** ( ) presents details for PWMs with *adj. site FE >1*.



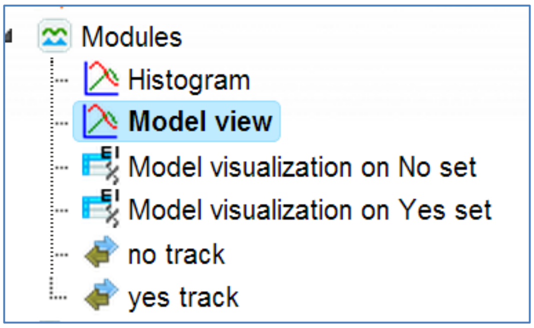
First | Previous | Page 1 of 14 | Next | Last   Showing 1 to 50 of 697 entries

Name	Matrix	Cutoff	Core cutoff	Core start	Core length	Matrix logo
V$AIRE_02	V$AIRE_02	0.63068	0	2	5	
V$AIRE_03	V$AIRE_03	0.76762	0	2	5	
V$ALX1_03	V$ALX1_03	0.67367	0.51305	3	5	

This profile is an intermediate result of the workflow, and it is used further for *Site search on gene set* analysis.

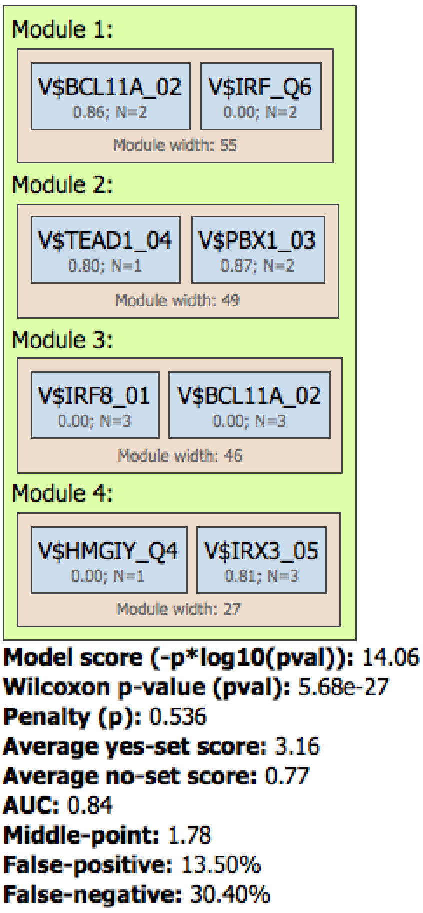
Site search analysis output ( ) is an intermediate result of the workflow, and its results are used further for the identification of composite modules. Details about the individual output files in this folder can be found in Section 20.1.4.

The **Modules** folder ( ) is a result of the analysis *Construct composite modules*. It contains two tables, two tracks, one histogram, and one model view as shown below:



Modules
- Histogram
- Model view
- Model visualization on No set
- Model visualization on Yes set
- no track
- yes track

The Model View is a graphical summary for the hierarchically organized composite elements generated as a result of the CMA analysis. As mentioned above, this workflow is designed to identify pairs of sites, and we asked to identify 2 to 5 pairs. The Model view presents four pairs, and we can see by exactly which site models (matrices) these pairs are formed as well as statistical parameters of the overall model.



Module 1:
V$BCL11A_02 0.86; N=2
V$IRF_Q6 0.00; N=2
Module width: 55

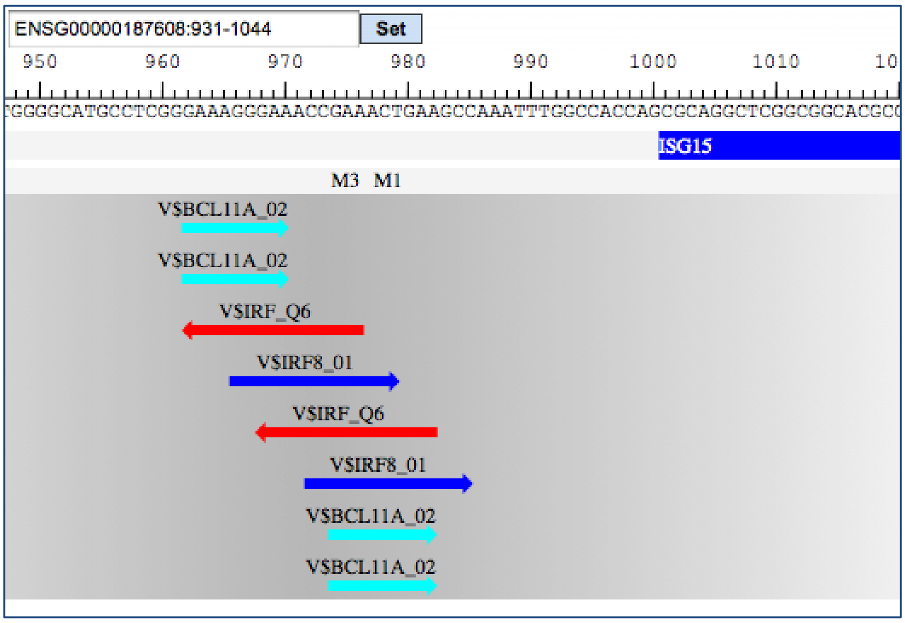
Module 2:
V$TEAD1_04 0.80; N=1
V$PBX1_03 0.87; N=2
Module width: 49

Module 3:
V$IRF8_01 0.00; N=3
V$BCL11A_02 0.00; N=3
Module width: 46

Module 4:
V$HMGIY_Q4 0.00; N=1
V$IRX3_05 0.81; N=3
Module width: 27

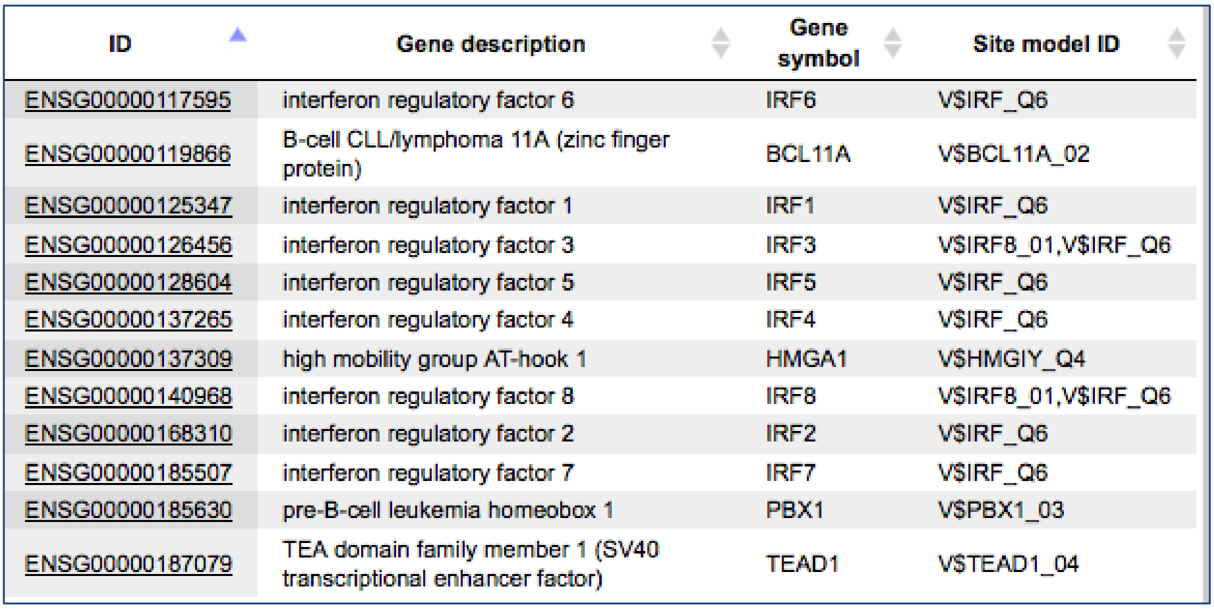
Model score (-p*log10(pval)): 14.06
Wilcoxon p-value (pval): 5.68e-27
Penalty (p): 0.536
Average yes-set score: 3.16
Average no-set score: 0.77
AUC: 0.84
Middle-point: 1.78
False-positive: 13.50%
False-negative: 30.40%

Each track can be opened in the genome browser by double-clicking. Visualization of the composite modules on the promoter of ISG15, one of the *Yes set* genes, is shown below.



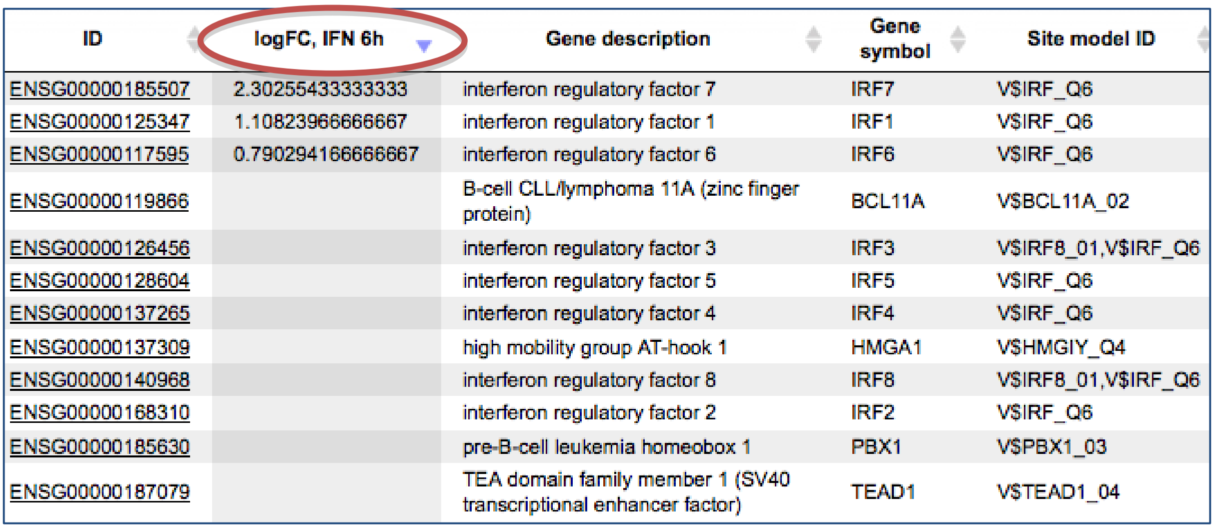
For more details on the individual output tables and tracks as well as for **visualization** of the identified composite modules in the genome browser please refer to Section 20.1.4.

The output table *Transcription factors Ensembl genes* is a list of transcription factors linked to the site models in the composite module identified by the workflow. For each transcription factor, the Ensembl gene ID is provided, as well as a gene description, the HGNC gene symbol, species, and site model (TRANSFAC® PWM name).



ID	Gene description	Gene symbol	Site model ID
ENSG00000117595	interferon regulatory factor 6	IRF6	V$IRF_Q6
ENSG00000119866	B-cell CLL/lymphoma 11A (zinc finger protein)	BCL11A	V$BCL11A_02
ENSG00000125347	interferon regulatory factor 1	IRF1	V$IRF_Q6
ENSG00000126456	interferon regulatory factor 3	IRF3	V$IRF8_01,V$IRF_Q6
ENSG00000128604	interferon regulatory factor 5	IRF5	V$IRF_Q6
ENSG00000137265	interferon regulatory factor 4	IRF4	V$IRF_Q6
ENSG00000137309	high mobility group AT-hook 1	HMGA1	V$HMGIY_Q4
ENSG00000140968	interferon regulatory factor 8	IRF8	V$IRF8_01,V$IRF_Q6
ENSG00000168310	interferon regulatory factor 2	IRF2	V$IRF_Q6
ENSG00000185507	interferon regulatory factor 7	IRF7	V$IRF_Q6
ENSG00000185630	pre-B-cell leukemia homeobox 1	PBX1	V$PBX1_03
ENSG00000187079	TEA domain family member 1 (SV40 transcriptional enhancer factor)	TEAD1	V$TEAD1_04

Twelve transcription factors shown in the table above are candidate regulators of genes in the input *Yes set*. They are suggested to regulate transcription of Yes-genes via the identified composite elements. This table can be further annotated to add a column with expression values, as shown below. Details for annotation of the tables are given in Section 16.1.1.



ID	logFC, IFN 6h	Gene description	Gene symbol	Site model ID
ENSG00000185507	2.30255433333333	interferon regulatory factor 7	IRF7	V$IRF_Q6
ENSG00000125347	1.10823966666667	interferon regulatory factor 1	IRF1	V$IRF_Q6
ENSG00000117595	0.790294166666667	interferon regulatory factor 6	IRF6	V$IRF_Q6
ENSG00000119866		B-cell CLL/lymphoma 11A (zinc finger protein)	BCL11A	V$BCL11A_02
ENSG00000126456		interferon regulatory factor 3	IRF3	V$IRF8_01,V$IRF_Q6
ENSG00000128604		interferon regulatory factor 5	IRF5	V$IRF_Q6
ENSG00000137265		interferon regulatory factor 4	IRF4	V$IRF_Q6
ENSG00000137309		high mobility group AT-hook 1	HMGA1	V$HMGIY_Q4
ENSG00000140968		interferon regulatory factor 8	IRF8	V$IRF8_01,V$IRF_Q6
ENSG00000168310		interferon regulatory factor 2	IRF2	V$IRF_Q6
ENSG00000185630		pre-B-cell leukemia homeobox 1	PBX1	V$PBX1_03
ENSG00000187079		TEA domain family member 1 (SV40 transcriptional enhancer factor)	TEAD1	V$TEAD1_04

Three TFs are found to be highly up-regulated under the same conditions, IRF1, 6 and 7. The role of these TFs in regulation of the input *Yes genes* is suggested by two independent lines of evidence: genes encoding these TFs are highly up-regulated, and their motifs are parts of the identified enriched composite modules.

Note*.* This workflow is available together with a valid TRANSFAC® license. Please, feel free to ask for details (info@genexplain.com).

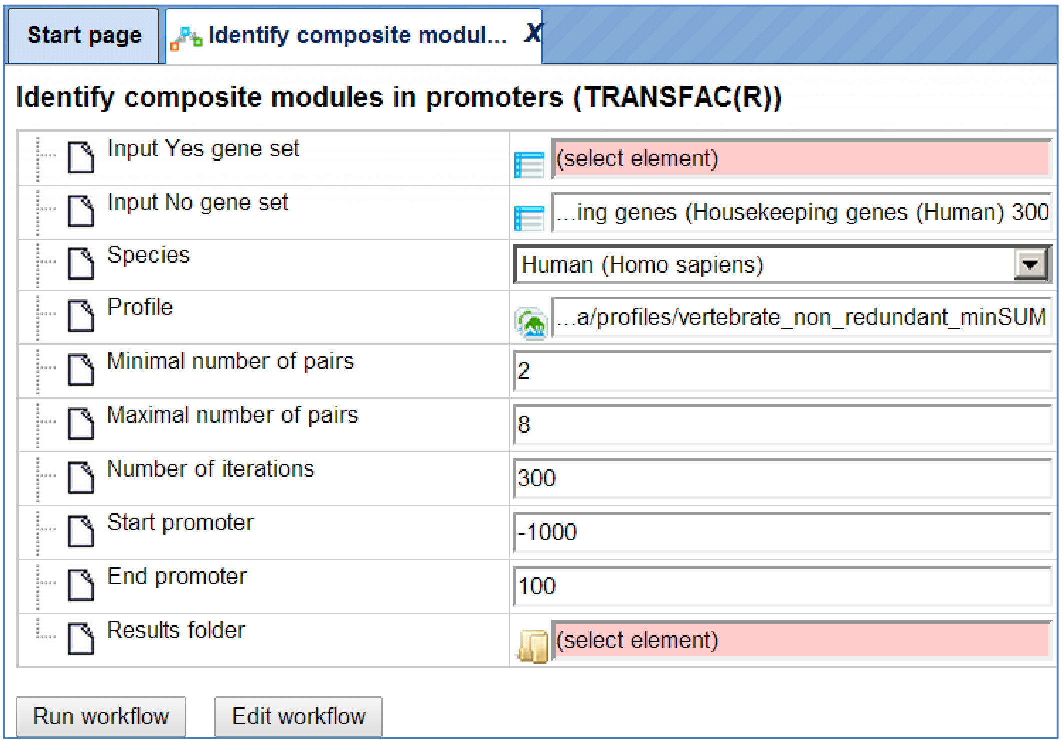
10.4.4.2. Version 1.2 (Classical) with TRANSFAC®

This workflow enables the identification of combinations of several TFBSs in the promoters of the genes under study (Yes-set). Such combinations of sites are referred to as composite modules. The resulting composite module differentiates the Yes-set from a background set (No-set).

In the first part of the workflow a *Site search on gene set (* *)* is performed with your selected Yes-set, No-set and a specified profile of matrices. You can refer to section 9.3 for details of this method. In the second part of this workflow, composite modules are identified ( ) based on the identified single sites in the Yes and No sets. For more details about the hierarchical structure of the composite modules, search for composite modules, visualization and interpretation of the results refer to section 20.1.5.4.

To launch the workflow, follow these steps:

Step1. Open the workflow input form from the Start page. It opens in the main Work Space and looks as shown below:



Identify composite modules in promoters (TRANSFAC(R))

Input Yes gene set	(select element)
Input No gene set	...ing genes (Housekeeping genes (Human) 300
Species	Human (Homo sapiens)
Profile	...a/profiles/vertebrate_non_redundant_minSUM
Minimal number of pairs	2
Maximal number of pairs	8
Number of iterations	300
Start promoter	-1000
End promoter	100
Results folder	(select element)

Step 2. Input the Yes set from the tree. You can either drag-and-drop or select the Yes set from the Tree area. Here, the set of genes from the Example folder is used as input Yes set.

Step 3. Similarly input the No set from the tree area. By default the workflow uses a subset fo 300 genes randomly taken out of the human housekeeping genes. The default NO set can be found here:

http://genexplain-platform.com/bioumlweb/#de=data/Public/Data%20sets/Data/Housekeeping%20genes%20(Human)%20300

Step 4. After input of the Yes and No sets, the species (human, mouse or rat) is adjusted automatically. Verify the species shown in the species field.

Step 5. Select the profile. This profile will be applied in the first part of the workflow for identification of TFBSs. The default profile is *vertebrate_non_redundant_minSUM* from the most recent TRANSFAC® release available. Any other TRANSFAC® profile or user-specific profile created with TRANSFAC® matrices can be chosen. With a mouse click on the field **Profile**, a pop-up window opens, where a profile can be selected. The profile used in this example is:

http://genexplain-platform.com/bioumlweb/#de=databases/TRANSFAC(R)%202014.4/Data/profiles/vertebrate_human_p0.001

Tip If you are interested in finding site models for particular TFs, and see them eventually in the resulting composite modules, you need to be sure that such matrices are present in the selected profile.

Step 6. Set up parameters for the composite module search. This workflow identifies pairs of sites. By default, the minimum and maximum number of pairs are fixed as 2 and

8, respectively. You can change these parameters according to the number of pairs you want to identify. The number of iterations of the genetic algorithm is 300 by default, and can be adapted as required.

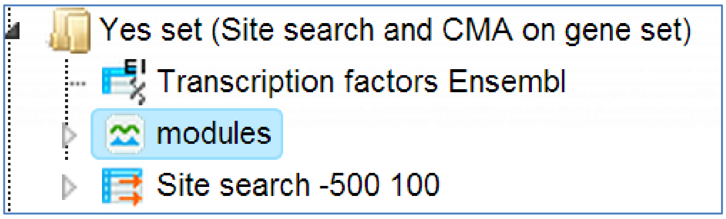
Step 7. Specify promoter regions relative to the TSS as they are annotated in Ensembl. The default promoter region is -1000 to +100 relative to the TSS. You can edit the fields **Start promoter** and **End promoter** as required.

Step 8. Specify the result folder location and name and press the button [Run workflow].

> **Note***.* This workflow may take more time depending on the size of the Yes and No sets and on the number of iterations. The recommended size of the Yes set is 150 genes maximum, and the recommended size of the No set is 300 genes maximum.

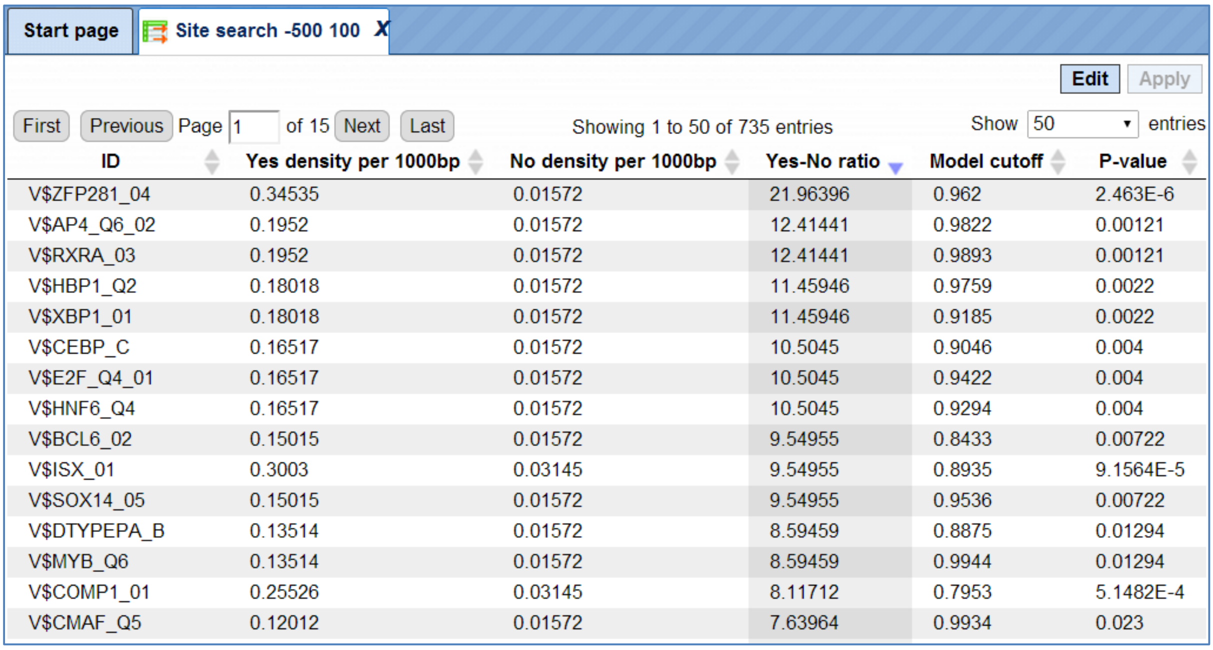
Results

The results folder consists of two folders and one table as shown below:



Yes set (Site search and CMA on gene set)
- Transcription factors Ensembl
- modules
- Site search -500 100

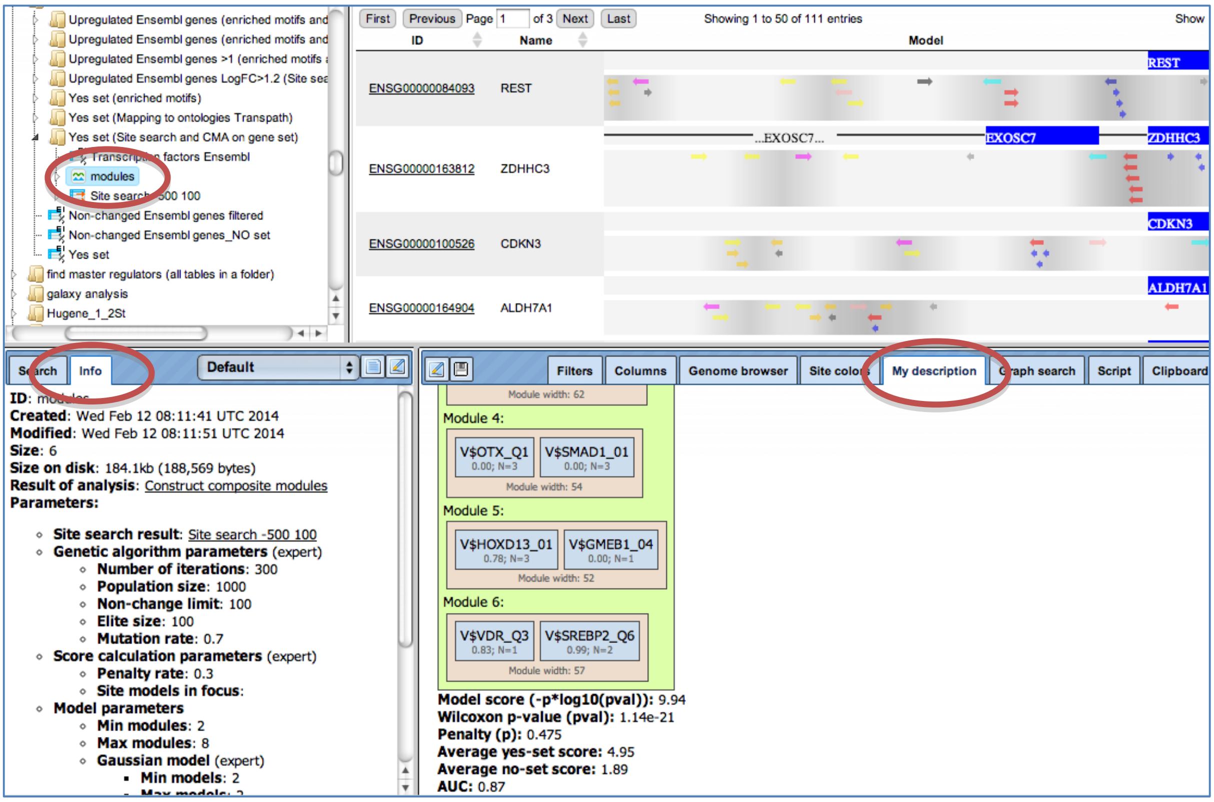
Site search analysis output ( ) The summary table of the site search is shown below. Identified TFBSs are used further for the identification of composite modules. Details about the individual output files can be found in Section 20.1.2.



Showing 1 to 50 of 735 entries

ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
V$ZFP281_04	0.34535	0.01572	21.96396	0.962	2.463E-6
V$AP4_Q6_02	0.1952	0.01572	12.41441	0.9822	0.00121
V$RXRA_03	0.1952	0.01572	12.41441	0.9893	0.00121
V$HBP1_Q2	0.18018	0.01572	11.45946	0.9759	0.0022
V$XBP1_01	0.18018	0.01572	11.45946	0.9185	0.0022
V$CEBP_C	0.16517	0.01572	10.5045	0.9046	0.004
V$E2F_Q4_01	0.16517	0.01572	10.5045	0.9422	0.004
V$HNF6_Q4	0.16517	0.01572	10.5045	0.9294	0.004
V$BCL6_02	0.15015	0.01572	9.54955	0.8433	0.00722
V$ISX_01	0.3003	0.03145	9.54955	0.8935	9.1564E-5
V$SOX14_05	0.15015	0.01572	9.54955	0.9536	0.00722
V$DTYPEPA_B	0.13514	0.01572	8.59459	0.8875	0.01294
V$MYB_Q6	0.13514	0.01572	8.59459	0.9944	0.01294
V$COMP1_01	0.25526	0.03145	8.11712	0.7953	5.1482E-4
V$CMAF_Q5	0.12012	0.01572	7.63964	0.9934	0.023

With a double-click on the folder **modules**, the visualization of the composite modules in the promoters of the *Yes* set will be opened in the work space. Simultaneously, in the operations field, under the tab My description, a plot with a schematic representation of the composite modules and statistical parameters are shown. In the Info box you can see the list of parameters this particular run of the workflow was done with.


Upregulated Ensembl genes (enriched motifs and
Upregulated Ensembl genes (enriched motifs and
Upregulated Ensembl genes >1 (enriched motifs
Upregulated Ensembl genes LogFC>1.2 (Site sea
Yes set (enriched motifs)
Yes set (Mapping to ontologies Transpath)
Yes set (Site search and CMA on gene set)
Transcription factors Ensembl
modules
Site search -500 100
Non-changed Ensembl genes filtered
Non-changed Ensembl genes_NO set
Yes set
find master regulators (all tables in a folder)
galaxy analysis
Hugene_1_2St
First
Previous
Page 1 of 3
Next
Last
Showing 1 to 50 of 111 entries
Show
REST
...EXOSC7...
EXOSC7
ZDHHC3
CDKN3
ALDH7A1
Search
Info
Default
Filters
Columns
Genome browser
Site colors
My description
Graph search
Script
Clipboard

ID	Name	Model
ENSG00000084093	REST	
ENSG00000163812	ZDHHC3	
ENSG00000100526	CDKN3	
ENSG00000164904	ALDH7A1	

ID: modules
Created: Wed Feb 12 08:11:41 UTC 2014
Modified: Wed Feb 12 08:11:51 UTC 2014
Size: 6
Size on disk: 184.1kb (188,569 bytes)
Result of analysis: Construct composite modules
Parameters:

- **Site search result**: Site search -500 100
- **Genetic algorithm parameters** (expert)
  - **Number of iterations**: 300
  - **Population size**: 1000
  - **Non-change limit**: 100
  - **Elite size**: 100
  - **Mutation rate**: 0.7
- **Score calculation parameters** (expert)
  - **Penalty rate**: 0.3
  - **Site models in focus**:
- **Model parameters**
  - **Min modules**: 2
  - **Max modules**: 8
  - **Gaussian model** (expert)
    - **Min models**: 2
    - **Max models**: 2

Module width: 62

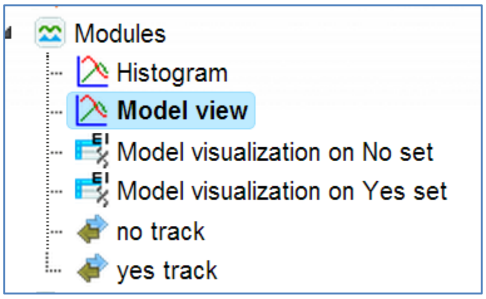
Module 4:
V$OTX_Q1 0.00; N=3
V$SMAD1_01 0.00; N=3
Module width: 54

Module 5:
V$HOXD13_01 0.78; N=3
V$GMEB1_04 0.00; N=1
Module width: 52

Module 6:
V$VDR_Q3 0.83; N=1
V$SREBP2_Q6 0.99; N=2
Module width: 57

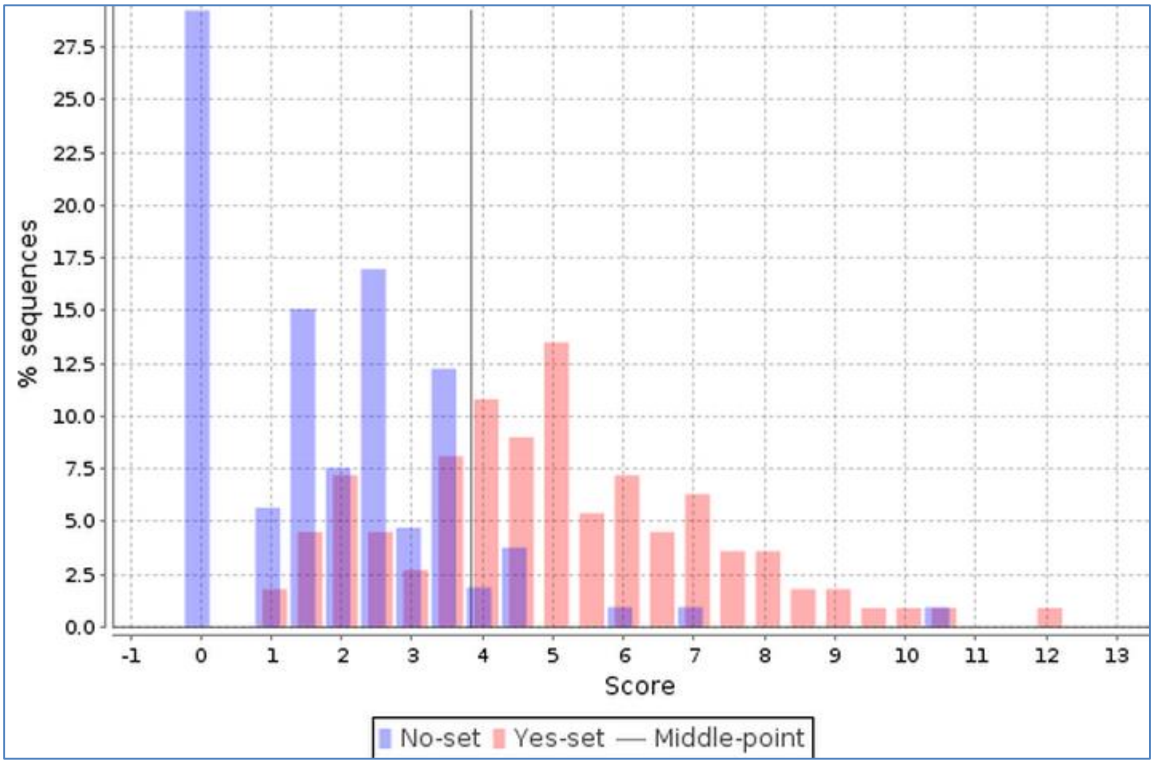
Model score (-p*log10(pval)): 9.94
Wilcoxon p-value (pval): 1.14e-21
Penalty (p): 0.475
Average yes-set score: 4.95
Average no-set score: 1.89
AUC: 0.87

The **modules** folder can be expanded in the tree area. It contains two tables (), two tracks (), and two plots ():



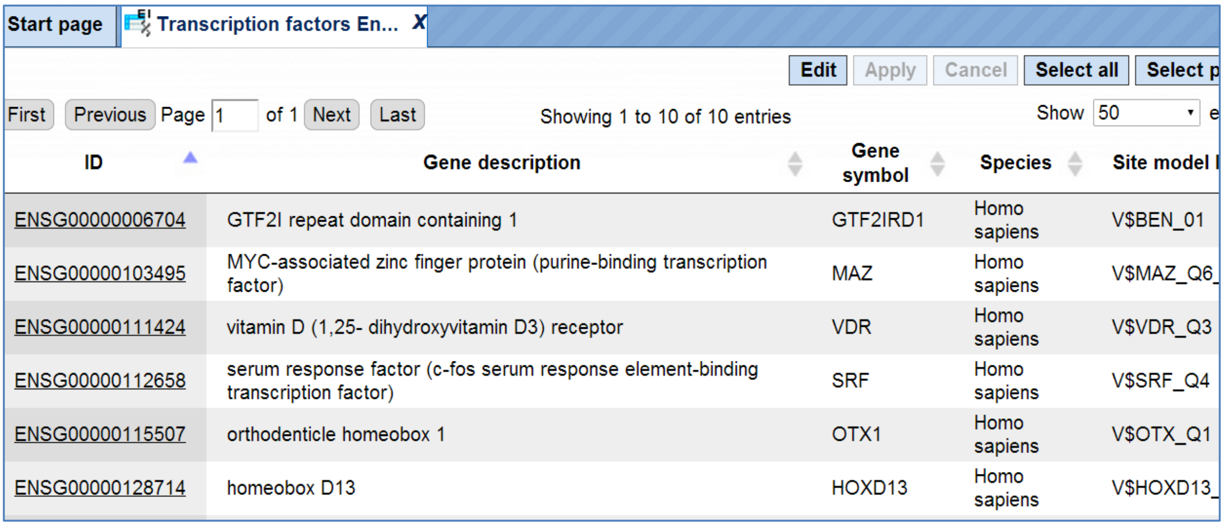
The plot **Model View** is a graphical summary for all composite modules generated as a result of the CMA analysis, and it can be opened in the work space.

The plot **Histogram** is a distribution of scores for individual promoters:



For a detailed interpretation of the histogram as well as for a visualization of the identified composite modules in the genome browser, please refer to section 20.1.5.5.

The output table **Transcription factors Ensembl genes** is a list of transcription factors linked to the site models in the composite module. For each transcription factor, the Ensembl gene ID is provided, as well a gene description, HGNC gene symbol, species, and site model (TRANSFAC® PMW name):



Showing 1 to 10 of 10 entries

ID	Gene description	Gene symbol	Species	Site model I
ENSG00000006704	GTF2I repeat domain containing 1	GTF2IRD1	Homo sapiens	V$BEN_01
ENSG00000103495	MYC-associated zinc finger protein (purine-binding transcription factor)	MAZ	Homo sapiens	V$MAZ_Q6_
ENSG00000111424	vitamin D (1,25- dihydroxyvitamin D3) receptor	VDR	Homo sapiens	V$VDR_Q3
ENSG00000112658	serum response factor (c-fos serum response element-binding transcription factor)	SRF	Homo sapiens	V$SRF_Q4
ENSG00000115507	orthodenticle homeobox 1	OTX1	Homo sapiens	V$OTX_Q1
ENSG00000128714	homeobox D13	HOXD13	Homo sapiens	V$HOXD13_

> **Note**. This workflow is available together with a valid TRANSPATH® license. Please, feel free to ask for details (info@genexplain.com).

10.4.5. Cross-species identification of enriched motifs in promoters using ortholog information (TRANSFAC®)

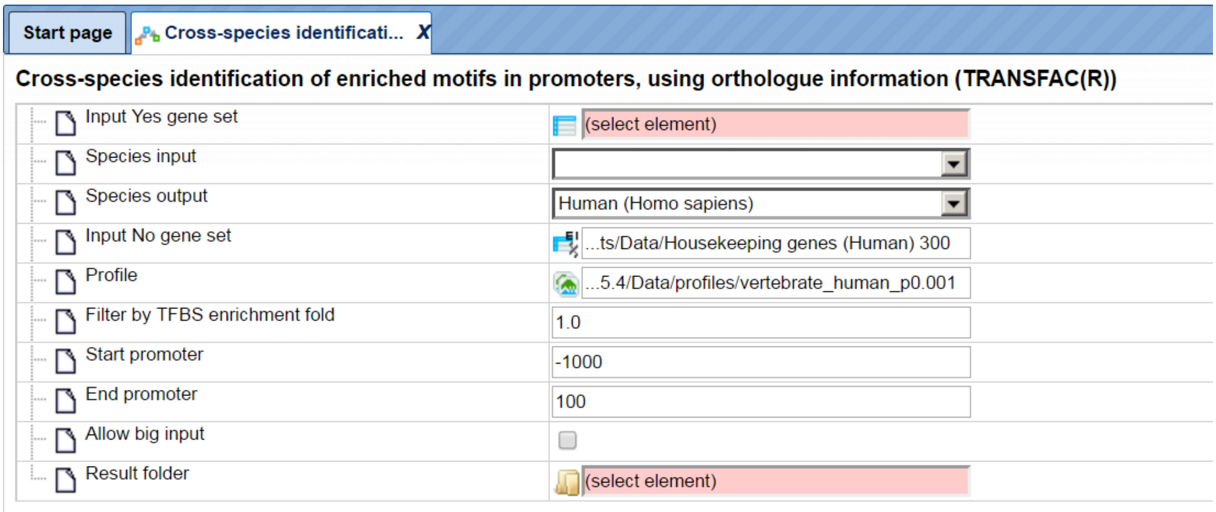
This workflow is designed to find individual motifs enriched in the promoters of the input gene set as compared with a background set (No set). It is very similar to the workflow described in section 10.4.3.1 except that here you can use an input table for any species and get the output for the desired species.

The workflow can be accessed from the start page here:

analyses/Workflows/TRANSFAC/Cross-species identification of enriched motifs in promoters, using ortholog information (TRANSFAC(R))

In the first part of the workflow, enriched motifs are identified by the method *analyses/Methods/Site analysis/Search for enriched TFBSs (genes)*, icon . Please refer to section 20.1.4 for details on this particular analysis method. Filtered enriched motifs serve as a basis to construct a specific profile, and this profile is applied to the promoters of the input gene set, method *analyses/Methods/Site analysis/Site search on gene set*. Details about this individual method are given in section 6.1.2. The last step is a conversion to homology transcription factors.

The input form looks as shown below:



Cross-species identification of enriched motifs in promoters, using orthologue information (TRANSFAC(R))

Input Yes gene set	(select element)
Species input	
Species output	Human (Homo sapiens)
Input No gene set	...ts/Data/Housekeeping genes (Human) 300
Profile	...5.4/Data/profiles/vertebrate_human_p0.001
Filter by TFBS enrichment fold	1.0
Start promoter	-1000
End promoter	100
Allow big input	□
Result folder	(select element)

Step 1: Select an **Input Yes gene set** from the tree. You can either drag-and-drop or select the Yes set from the Tree area.

Here, the set of up-regulated genes from the following *Examples* folder is used:

data/Examples/Transcriptional biomarkers to predict mouse liver tumors, GSE18858/Data/Normalized (RMA) DEGs with EBarrays/Naphthalene_20ppm upreg Ensembl Select the species of the input table

Step 2: Specify the **Species** of the **input** set and the Species of the **output** set.

Step 3: Select Input No gene set from the tree area. By default, the workflow uses a subset from 300 genes randomly taken out of the human housekeeping genes.

Step 4: The profile will be applied in the first part of the workflow for the identification of enriched motifs. The default profile is *vertebrate_human_p0.001* from the most recent

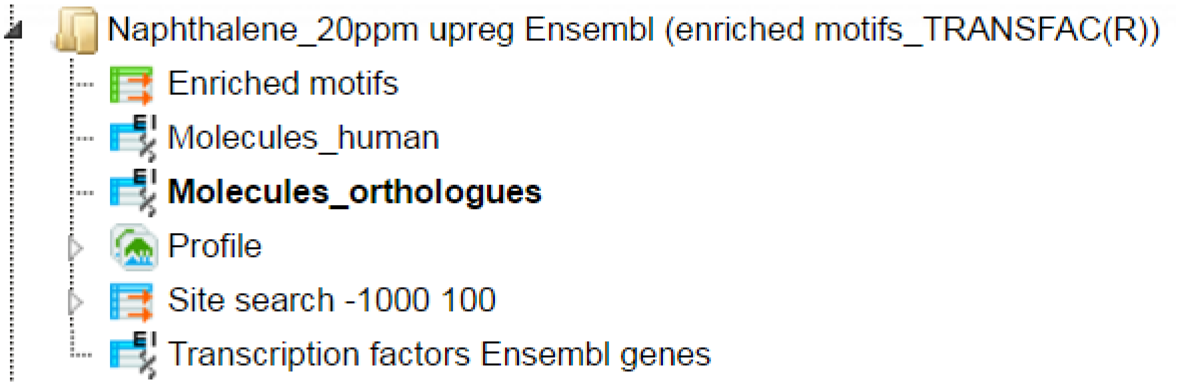
TRANSFAC® release available. Any other TRANSFAC® profile or user-specific profile can be selected. With a mouse click on the field **Profile**, a pop-up window will open, where a profile can be selected.

Step 5: Filter by TFBS enrichment fold: In this field you can specify the enrichment fold (FE) to filter the motifs. By default, it is 1.0, which means all motifs with FE>1.0 will be reported in the resulting table and the same motifs will serve to create a specific profile. If you want to use highly-enriched motifs, you can specify higher thresholds, e.g. 1.1, 1.2 etc., or even 2.0 or 3.0 depending on your Yes and No sets. It is recommended that you run it with default parameters first, check the results, and then repeat with the desired filter value.

Step 6: Specify the promoter region relative to TSS as they are annotated in Ensembl. The default promoter region is -1000 to +100 relative to the TSS. You can edit the fields **Start promoter** and **End promoter** as required.

Step 7: Specify the **Result folder** location and name and Press the button [Run workflow]. Wait till the workflow is completed.

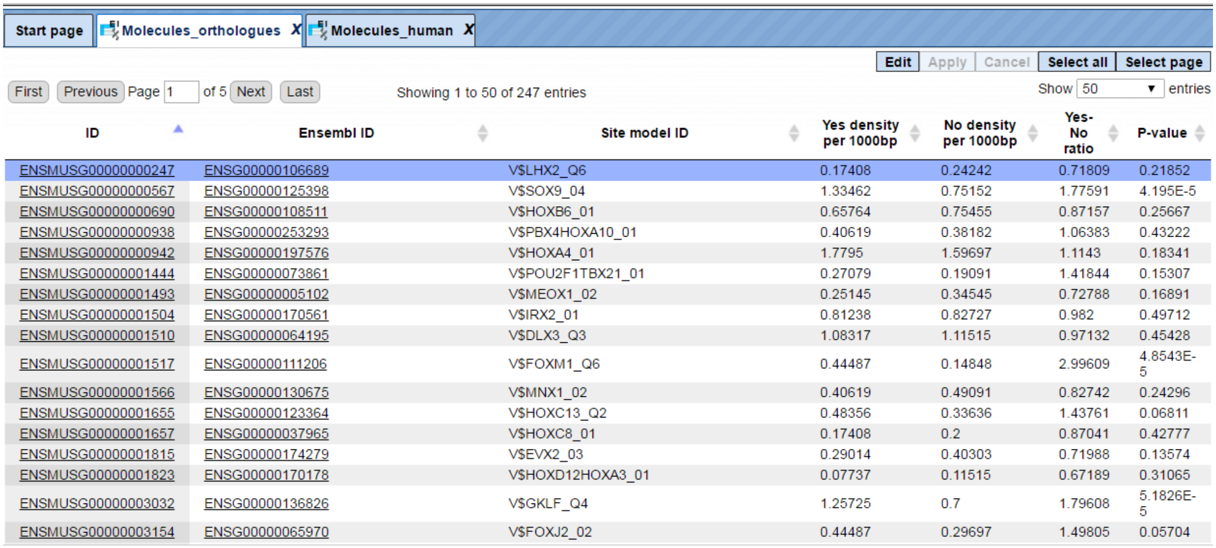
The **result folder** consists of several files as shown below:



Naphthalene_20ppm upreg Ensembl (enriched motifs_TRANSFAC(R))
- Enriched motifs
- Molecules_human
- **Molecules_orthologues**
- Profile
- Site search -1000 100
- Transcription factors Ensembl genes

The table **Enriched Motifs** ( ) contains those site models, here TRANSFAC® matrices, which are enriched in the Yes set in comparison with the No set. More details on the result can be found in section 10.4.3.1.

The table **Molecules Orthologs** have a list of molecules from the input gene set with their site search results as shown below:



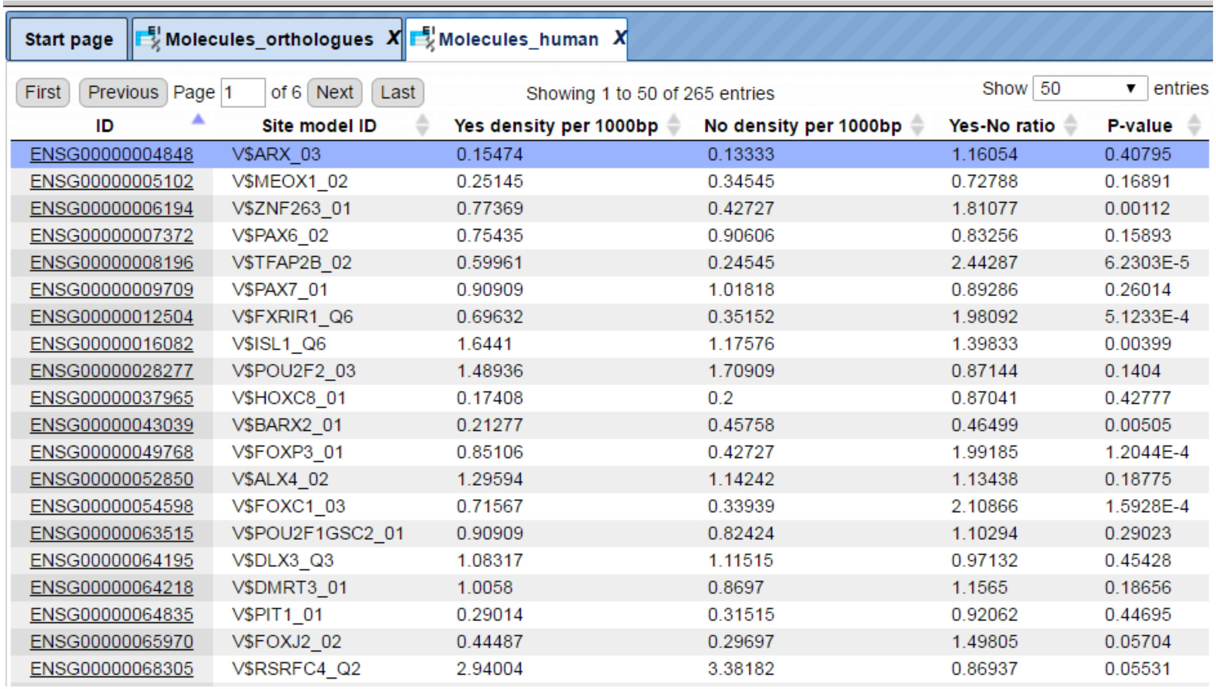
Showing 1 to 50 of 247 entries

ID	Ensembl ID	Site model ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	P-value
ENSMUSG00000000247	ENSG00000106689	V$LHX2_Q6	0.17408	0.24242	0.71809	0.21852
ENSMUSG00000000567	ENSG00000125398	V$SOX9_04	1.33462	0.75152	1.77591	4.195E-5
ENSMUSG00000000690	ENSG00000108511	V$HOXB6_01	0.65764	0.75455	0.87157	0.25667
ENSMUSG00000000938	ENSG00000253293	V$PBX4HOXA10_01	0.40619	0.38182	1.06383	0.43222
ENSMUSG00000000942	ENSG00000197576	V$HOXA4_01	1.7795	1.59697	1.1143	0.18341
ENSMUSG00000001444	ENSG00000073861	V$POU2F1TBX21_01	0.27079	0.19091	1.41844	0.15307
ENSMUSG00000001493	ENSG00000005102	V$MEOX1_02	0.25145	0.34545	0.72788	0.16891
ENSMUSG00000001504	ENSG00000170561	V$IRX2_01	0.81238	0.82727	0.982	0.49712
ENSMUSG00000001510	ENSG00000064195	V$DLX3_Q3	1.08317	1.11515	0.97132	0.45428
ENSMUSG00000001517	ENSG00000111206	V$FOXM1_Q6	0.44487	0.14848	2.99609	4.8543E-5
ENSMUSG00000001566	ENSG00000130675	V$MNX1_02	0.40619	0.49091	0.82742	0.24296
ENSMUSG00000001655	ENSG00000123364	V$HOXC13_Q2	0.48356	0.33636	1.43761	0.06811
ENSMUSG00000001657	ENSG00000037965	V$HOXC8_01	0.17408	0.2	0.87041	0.42777
ENSMUSG00000001815	ENSG00000174279	V$EVX2_03	0.29014	0.40303	0.71988	0.13574
ENSMUSG00000001823	ENSG00000170178	V$HOXD12HOXA3_01	0.07737	0.11515	0.67189	0.31065
ENSMUSG00000003032	ENSG00000136826	V$GKLF_Q4	1.25725	0.7	1.79608	5.1826E-5
ENSMUSG00000003154	ENSG00000065970	V$FOXJ2_02	0.44487	0.29697	1.49805	0.05704

Every gene is linked to the corresponding matrix molecule by the Yes-No ratio. More details on each column of the above results can be found in section 6.1.2.

The table **Molecules_human** contain the site models of the converted input table. In this case the output species is Human hence this table is Molecules_human with mapping to human Ensembl genes with corresponding matrices. If the output species is mouse, then this table will have mouse Ensembl genes.

The table **Transcription factor Ensemble genes** contains



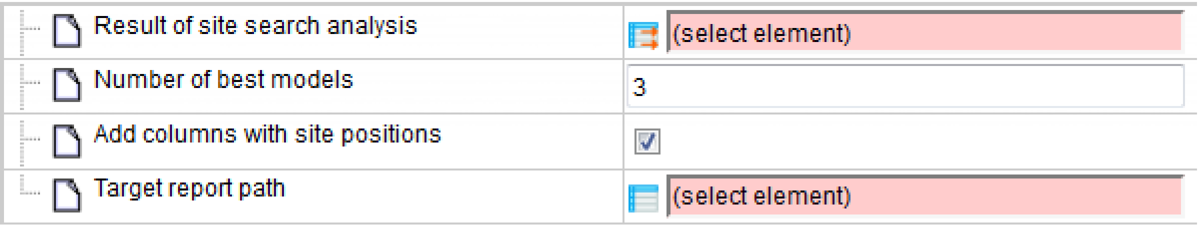
Showing 1 to 50 of 265 entries

ID	Site model ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	P-value
ENSG00000004848	V$ARX_03	0.15474	0.13333	1.16054	0.40795
ENSG00000005102	V$MEOX1_02	0.25145	0.34545	0.72788	0.16891
ENSG00000006194	V$ZNF263_01	0.77369	0.42727	1.81077	0.00112
ENSG00000007372	V$PAX6_02	0.75435	0.90606	0.83256	0.15893
ENSG00000008196	V$TFAP2B_02	0.59961	0.24545	2.44287	6.2303E-5
ENSG00000009709	V$PAX7_01	0.90909	1.01818	0.89286	0.26014
ENSG00000012504	V$FXRIR1_Q6	0.69632	0.35152	1.98092	5.1233E-4
ENSG00000016082	V$ISL1_Q6	1.6441	1.17576	1.39833	0.00399
ENSG00000028277	V$POU2F2_03	1.48936	1.70909	0.87144	0.1404
ENSG00000037965	V$HOXC8_01	0.17408	0.2	0.87041	0.42777
ENSG00000043039	V$BARX2_01	0.21277	0.45758	0.46499	0.00505
ENSG00000049768	V$FOXP3_01	0.85106	0.42727	1.99185	1.2044E-4
ENSG00000052850	V$ALX4_02	1.29594	1.14242	1.13438	0.18775
ENSG00000054598	V$FOXC1_03	0.71567	0.33939	2.10866	1.5928E-4
ENSG00000063515	V$POU2F1GSC2_01	0.90909	0.82424	1.10294	0.29023
ENSG00000064195	V$DLX3_Q3	1.08317	1.11515	0.97132	0.45428
ENSG00000064218	V$DMRT3_01	1.0058	0.8697	1.1565	0.18656
ENSG00000064835	V$PIT1_01	0.29014	0.31515	0.92062	0.44695
ENSG00000065970	V$FOXJ2_02	0.44487	0.29697	1.49805	0.05704
ENSG00000068305	V$RSRFC4_Q2	2.94004	3.38182	0.86937	0.05531

The output table Transcription factors Ensembl genes is a list of transcription factors linked to the enriched motifs. For each transcription factor, the Ensembl gene ID is provided, as well as gene description, HGNC gene symbol, species, and site model (TRANSFAC® PWM name). This table can be further annotated to add a column with expression values, as shown below. Details for annotation of the tables are given in the section 16.1.1.

10.4.6. Visualization of site search results

This method visualizes results of the site search analyses. It can be found under the tab *Analyses*, in the folder Methods/Site analysis/Site search report (). Here the default input form is shown:



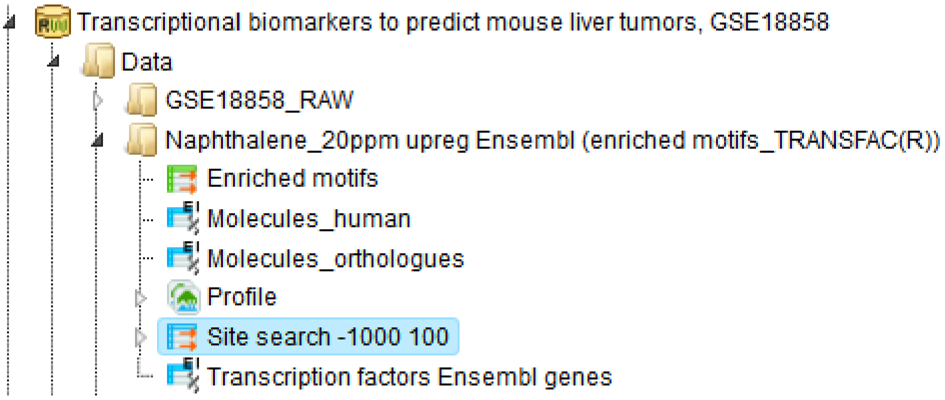
Result of site search analysis	(select element)
Number of best models	3
Add columns with site positions	☑
Target report path	(select element)

In the following, we will consider the input fields one by one.

Result of site search analysis. You can drag & drop the site search result (must contain summary table with p-value column) from your project within the tree area. Alternatively, you may click on the pink field "select element", and a new window will open, where you select the site search result. After having selected the result, press the [Ok] button.

For this example, all further steps are demonstrated with the following input set:

data/Examples/Transcriptional biomarkers to predict mouse liver tumors, GSE18858/Data/Naphthalene_20ppm upreg Ensembl (enriched motifs_TRANSFAC(R))/Site search -1000 100



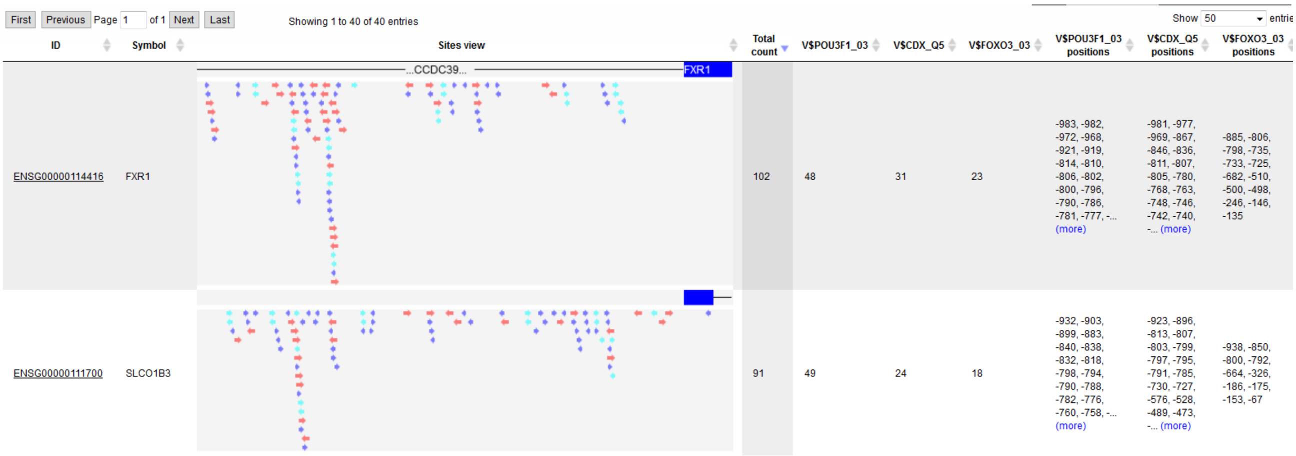
Transcriptional biomarkers to predict mouse liver tumors, GSE18858
- Data
  - GSE18858_RAW
  - Naphthalene_20ppm upreg Ensembl (enriched motifs_TRANSFAC(R))
    - Enriched motifs
    - Molecules_human
    - Molecules_orthologues
    - Profile
    - Site search -1000 100
    - Transcription factors Ensembl genes

Number of best models. Choose the number of best models (according to p-value) to include in the report. The default value are the three top models.

Add columns with site positions. If you are interested in the promoter positions of the single sites, please check the appropriate box.

Target report path. Define where the table with the results should be located in your project tree. You can do so by clicking on the pink box (select element) in the field, and a new window will open, where you can select the location of the resulting table and define its name.

Press the [Run] button and wait until the method is completed. The result opens automatically and looks like this:



ID	Symbol	Sites view	Total count	V$POU3F1_03	V$CDX_Q5	V$FOXO3_03	V$POU3F1_03 positions	V$CDX_Q5 positions	V$FOXO3_03 positions
ENSG00000114416	FXR1	...CCDC39... FXR1	102	48	31	23	-983, -982, -972, -968, -921, -919, -814, -810, -806, -802, -800, -796, -790, -786, -781, -777, -... (more)	-981, -977, -969, -867, -846, -836, -811, -807, -805, -780, -768, -763, -748, -746, -742, -740, -... (more)	-885, -806, -798, -735, -733, -725, -682, -510, -500, -498, -246, -146, -135
ENSG00000111700	SLCO1B3		91	49	24	18	-932, -903, -899, -883, -840, -838, -832, -818, -798, -794, -790, -788, -782, -776, -760, -758, -... (more)	-923, -896, -813, -807, -803, -799, -797, -795, -791, -785, -730, -727, -576, -528, -489, -473, -... (more)	-938, -850, -800, -792, -664, -326, -186, -175, -153, -67

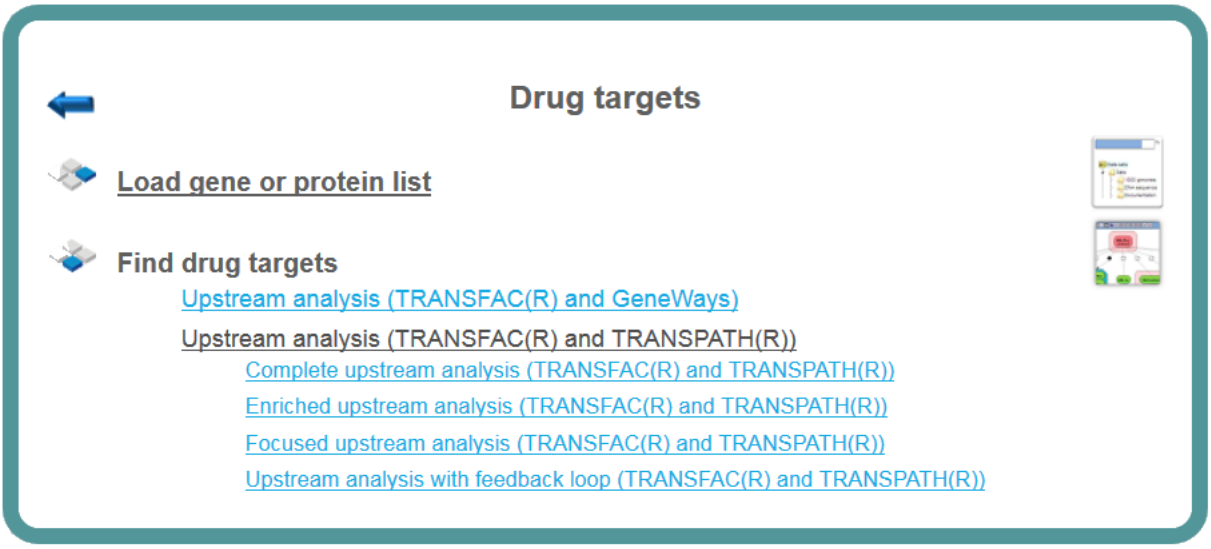
We see a visualization of two promoters from the genes FXR1 and SLCO1B3. The TSS is on the right side, where the blue box is located. All arrows are identified sites of the best models and marked with different colors. Also the orientation of the individual sites is shown by the arrow head. The number **Total count** gives the number of all sites according to the three best models. The numbers of sites for every model (1-3) are given in the next three columns. As an example, model **V$POU3F1_03** finds 49 sites in the promoter of FXR1. The column **V$POU3F1_03 positions** contains all single site positions for the first model in the promoter of FXR1, e.g. -994 from TSS.

10.5. Further workflows in this area

For the other workflows that you can find in the area ***Microarrays***, please refer to the following Sections:

Load data	See Chapter 3
Analyze networks	See Section 5.1
Find drug targets	See Chapter 11

11. Drug targets



Drug targets

Load gene or protein list

Find drug targets

Upstream analysis (TRANSFAC(R) and GeneWays)

Upstream analysis (TRANSFAC(R) and TRANSPATH(R))

Complete upstream analysis (TRANSFAC(R) and TRANSPATH(R))

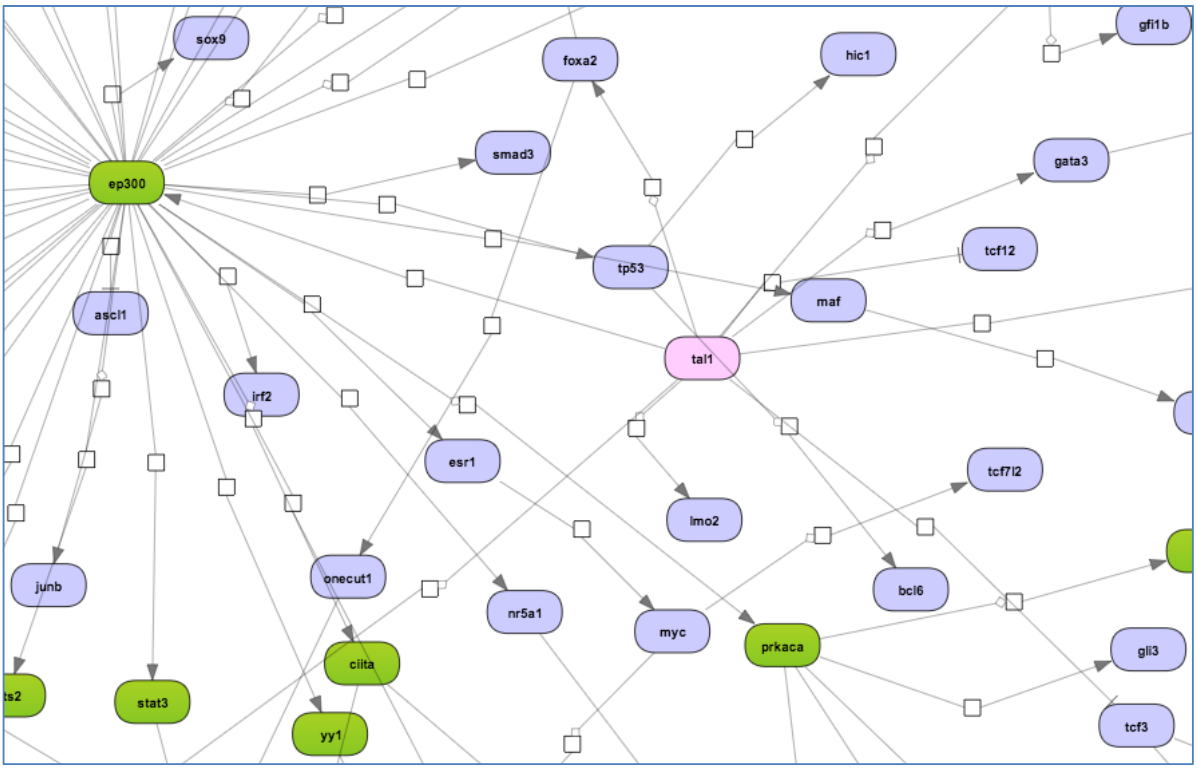
Enriched upstream analysis (TRANSFAC(R) and TRANSPATH(R))

Focused upstream analysis (TRANSFAC(R) and TRANSPATH(R))

Upstream analysis with feedback loop (TRANSFAC(R) and TRANSPATH(R))

11.1. Find drug targets with TRANSFAC® and GeneWays

This is very similar to the workflow described in greater detail below, Section 11.2, with the only difference being the pathway database applied for network analysis. Here the search for master regulatory molecules is performed with the protein-protein interaction network of the GeneWays database. Because of the high connectivity between molecules in GeneWays, by default 4 steps upstream of the input list of TFs are considered for the search for master regulators, as compared with 10 steps when applying the TRANSPATH® database. A diagram for the top master regulator as suggested from a GeneWays-based analysis is shown below in a force-directed layout.



For the input form of this workflow and for the description of the corresponding results folder, please refer to Section 11.2.

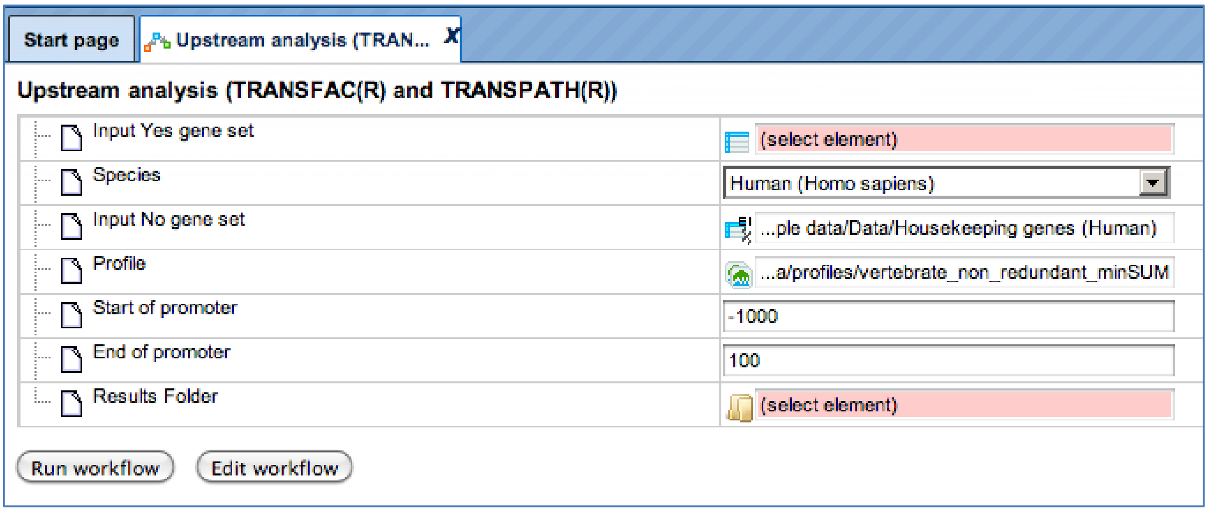
Note. This workflow is available together with a valid TRANSFAC® license.
Please feel free to ask for details (info@genexplain.com).

11.2. Find drug targets with TRANSFAC® and TRANSPATH®

The geneXplain upstream analysis is an integrated promoter (TRANSFAC®) and pathway (TRANSPATH®) analysis to discover unanticipated causal relationships in your data.

11.2.1. Complete upstream analysis (TRANSFAC® and TRANSPATH®)

To launch the workflow, open the workflow input form from the Start page:



Start page | Upstream analysis (TRAN...

Upstream analysis (TRANSFAC(R) and TRANSPATH(R))

Input Yes gene set	(select element)
Species	Human (Homo sapiens)
Input No gene set	...ple data/Data/Housekeeping genes (Human)
Profile	...a/profiles/vertebrate_non_redundant_minSUM
Start of promoter	-1000
End of promoter	100
Results Folder	(select element)

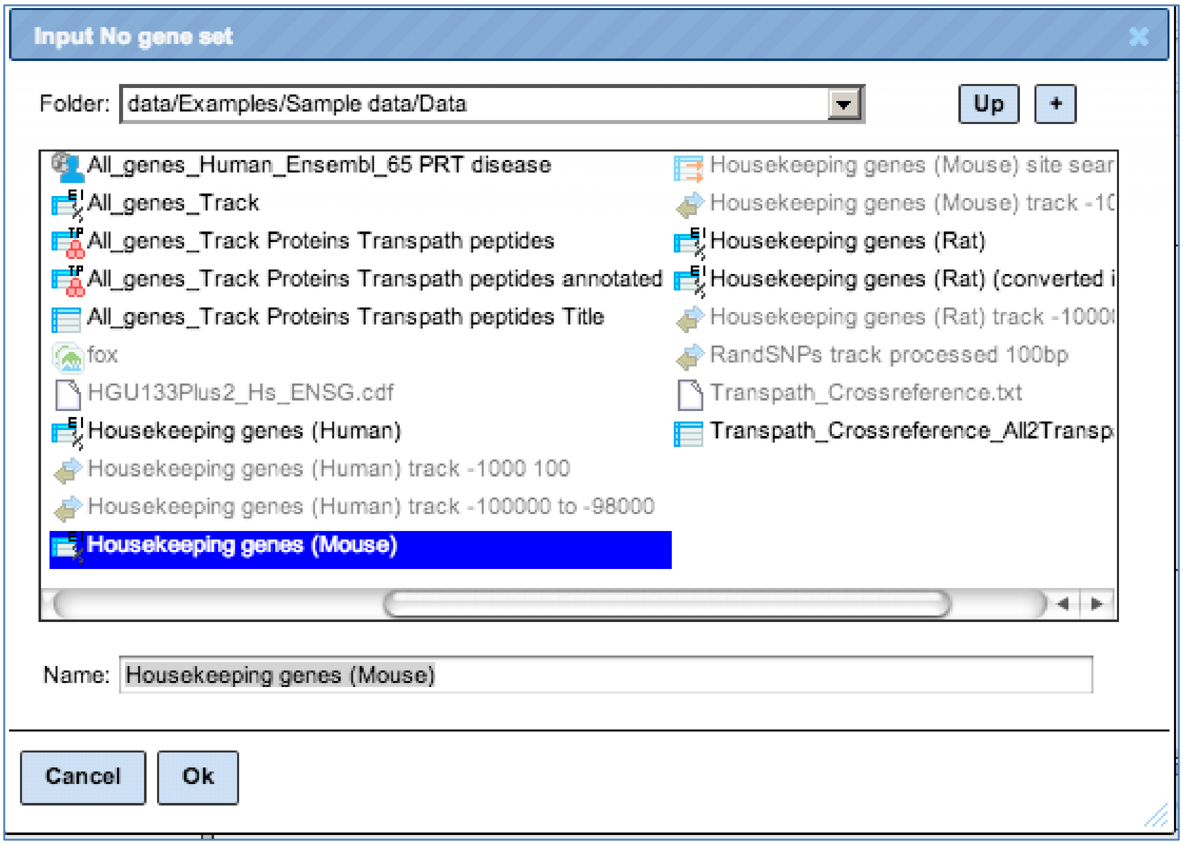
Step 1: Specify a gene set under study, e.g. a list of differentially regulated genes, as the Input Yes gene set. You can drag & drop it from your project within the Tree Area and drop in the pink box of the field **Input gene set**.

For this example, the further steps are demonstrated with the following input set:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Experiment%20normalized%20(RMA)%20(Differentially%20expressed%20genes%20Affy)/Upregulated%20Ensembl%20genes%20filtered%20(LogFC%3E1)

Step 2: Specify the biological species of the input set in the field **Species** by selecting the required species from the drop-down menu.

Step 3: Input a No gene set. This is the set of background genes or control set. The default No set used for this workflow is data/Examples/Sample data/Data/Housekeeping genes (Human). If your Yes set is from mouse or rat, you may wish to adjust the No set accordingly. With a mouse click on the field **Input No gene set**, a pop-up window will be opened as shown below. You can select mouse or rat housekeeping genes from these pre-compiled sample sets, or you can alternatively select any of your specific gene sets from your project. When selection is done, press [Ok].


Input No gene set
Up
+
All_genes_Human_Ensembl_65 PRT disease
All_genes_Track
All_genes_Track Proteins Transpath peptides
All_genes_Track Proteins Transpath peptides annotated
All_genes_Track Proteins Transpath peptides Title
fox
HGU133Plus2_Hs_ENSG.cdf
Housekeeping genes (Human)
Housekeeping genes (Human) track -1000 100
Housekeeping genes (Human) track -100000 to -98000
Housekeeping genes (Mouse)
Housekeeping genes (Mouse) site sear
Housekeeping genes (Mouse) track -1C
Housekeeping genes (Rat)
Housekeeping genes (Rat) (converted i
Housekeeping genes (Rat) track -1000
RandSNPs track processed 100bp
Transpath_Crossreference.txt
Transpath_Crossreference_All2Transp
Cancel
Ok

Folder: data/Examples/Sample data/Data

Name: Housekeeping genes (Mouse)

Step 4: Define a TRANSFAC® profile. The default profile is vertebrate_non_redundant_minSUM. Any other TRANSFAC® profile or user-specific profile can be chosen. With a mouse click on the field **Profile**, a pop-up window will open, where a profile can be selected.

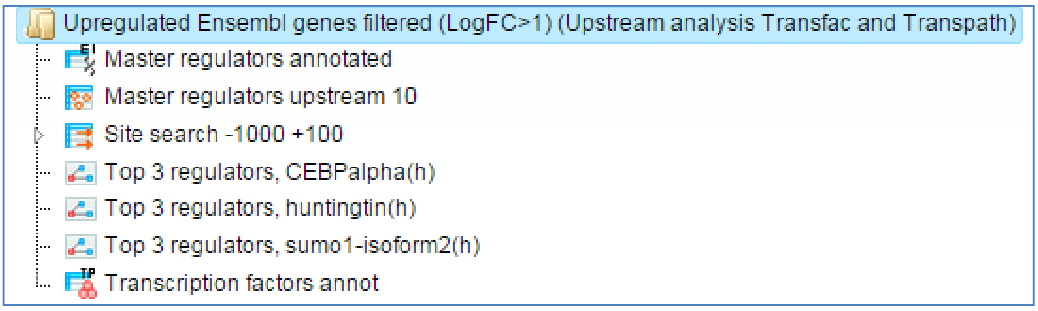
Step 5: Define the promoter regions to be analyzed. The default promoter region is from-1000 to 100 relative to the TSS as annotated in the Ensembl database. You can adjust **Start of promoter** and **End of promoter** by typing in the corresponding fields.

Step 6: Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink box (select element) in the field **Results folder**, and a new window will open, where you can select the location of the results folder and define its name.

Step 7: Press the [Run workflow] button. Wait until the workflow is completed, and take a look at the results.

Visualization and interpretation of results

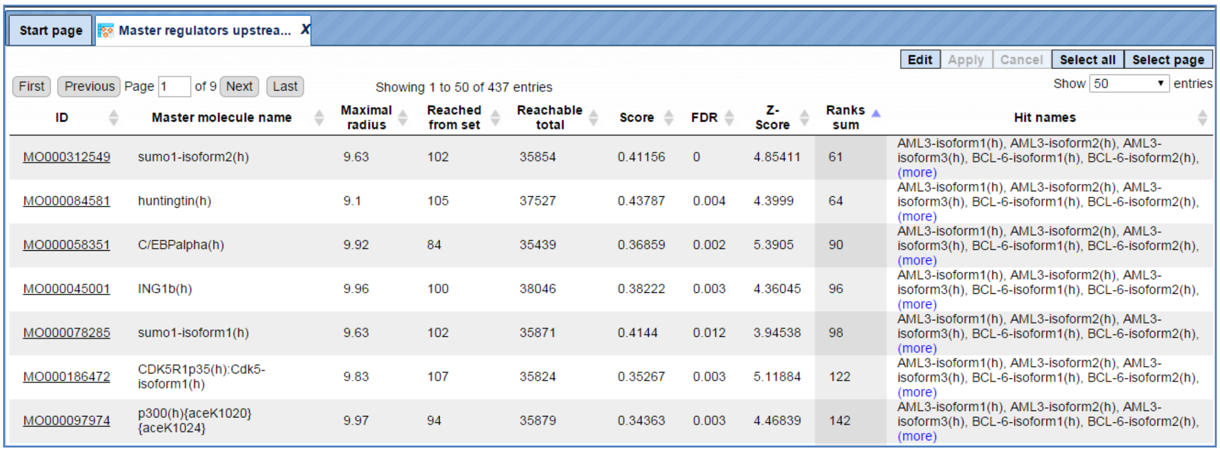
The result folder contains several files and one subfolder:



Upregulated Ensembl genes filtered (LogFC>1) (Upstream analysis Transfac and Transpath)
- Master regulators annotated
- Master regulators upstream 10
- Site search -1000 +100
- Top 3 regulators, CEBPalpha(h)
- Top 3 regulators, huntingtin(h)
- Top 3 regulators, sumo1-isoform2(h)
- Transcription factors annot

Master regulators

The primary result table *Master regulators upstream 10* (icon) is a list of master regulatory molecules that were identified at a distance of up to 10 steps upstream of the input TFs. Each master regulatory molecule is characterized by a Score, Z-score, FDR, and Ranks Sum.



Page 1 of 9

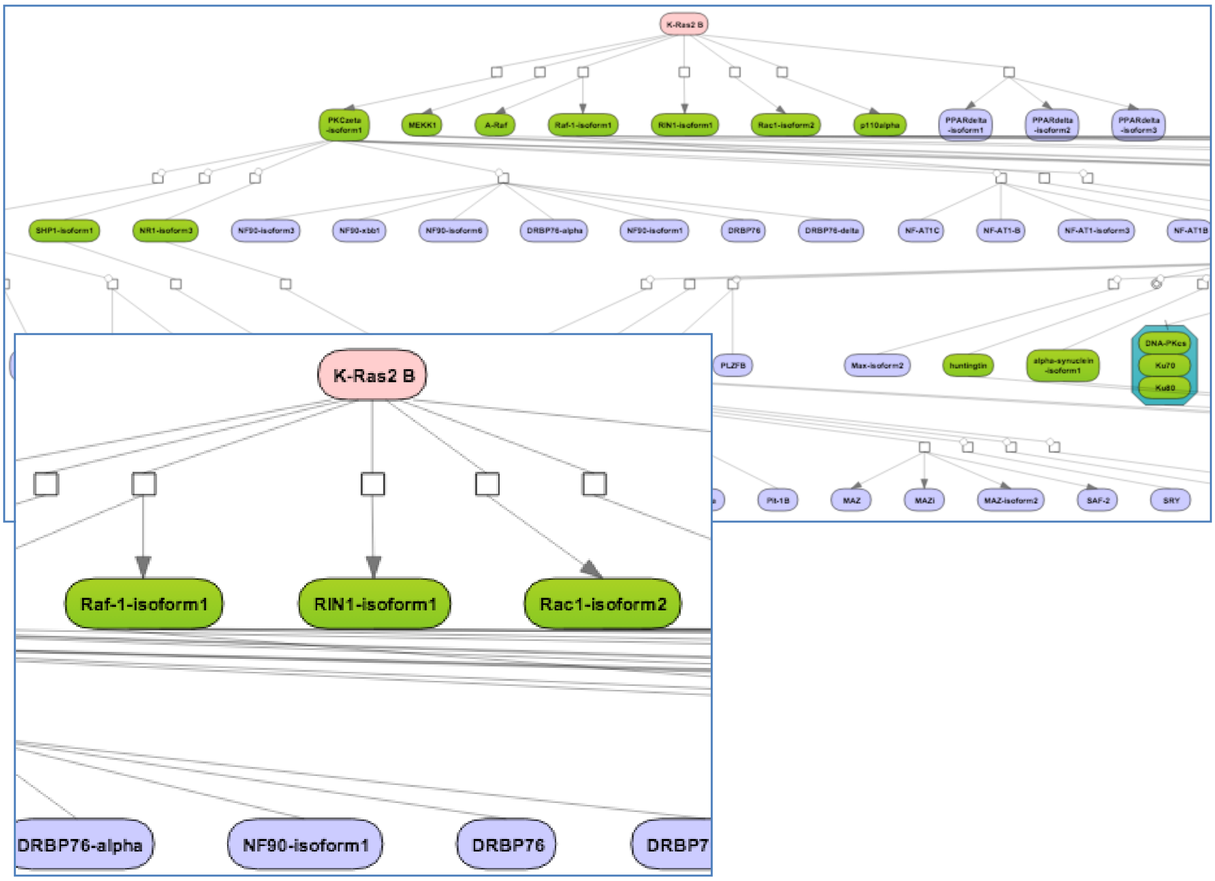
Showing 1 to 50 of 437 entries

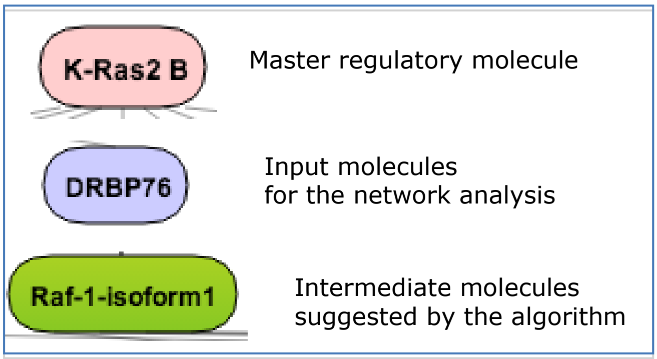
Show 50 entries

ID	Master molecule name	Maximal radius	Reached from set	Reachable total	Score	FDR	Z-Score	Ranks sum	Hit names
MO000312549	sumo1-isoform2(h)	9.63	102	35854	0.41156	0	4.85411	61	AML3-isoform1(h), AML3-isoform2(h), AML3-isoform3(h), BCL-6-isoform1(h), BCL-6-isoform2(h), (more)
MO000084581	huntingtin(h)	9.1	105	37527	0.43787	0.004	4.3999	64	AML3-isoform1(h), AML3-isoform2(h), AML3-isoform3(h), BCL-6-isoform1(h), BCL-6-isoform2(h), (more)
MO000058351	C/EBPalpha(h)	9.92	84	35439	0.36859	0.002	5.3905	90	AML3-isoform1(h), AML3-isoform2(h), AML3-isoform3(h), BCL-6-isoform1(h), BCL-6-isoform2(h), (more)
MO000045001	ING1b(h)	9.96	100	38046	0.38222	0.003	4.36045	96	AML3-isoform1(h), AML3-isoform2(h), AML3-isoform3(h), BCL-6-isoform1(h), BCL-6-isoform2(h), (more)
MO000078285	sumo1-isoform1(h)	9.63	102	35871	0.4144	0.012	3.94538	98	AML3-isoform1(h), AML3-isoform2(h), AML3-isoform3(h), BCL-6-isoform1(h), BCL-6-isoform2(h), (more)
MO000186472	CDK5R1p35(h):Cdk5-isoform1(h)	9.83	107	35824	0.35267	0.003	5.11884	122	AML3-isoform1(h), AML3-isoform2(h), AML3-isoform3(h), BCL-6-isoform1(h), BCL-6-isoform2(h), (more)
MO000097974	p300(h){aceK1020}{aceK1024}	9.97	94	35879	0.34363	0.003	4.46839	142	AML3-isoform1(h), AML3-isoform2(h), AML3-isoform3(h), BCL-6-isoform1(h), BCL-6-isoform2(h), (more)

Further details about the columns of this table and how to work further with it are given in Section 5.1.1. The selection of the best master regulatory molecules based on Score, Z-score and Ranks sum is explained in there under "Interpretation of the results".

The three *Top 3 regulators* diagrams (icon)visualize the networks for each of the three top master regulators. By default, the top regulators are identified upon sorting the *Master regulators upstream 10* table (icon) by the column **Ranks sum** with the lowest rank on top.



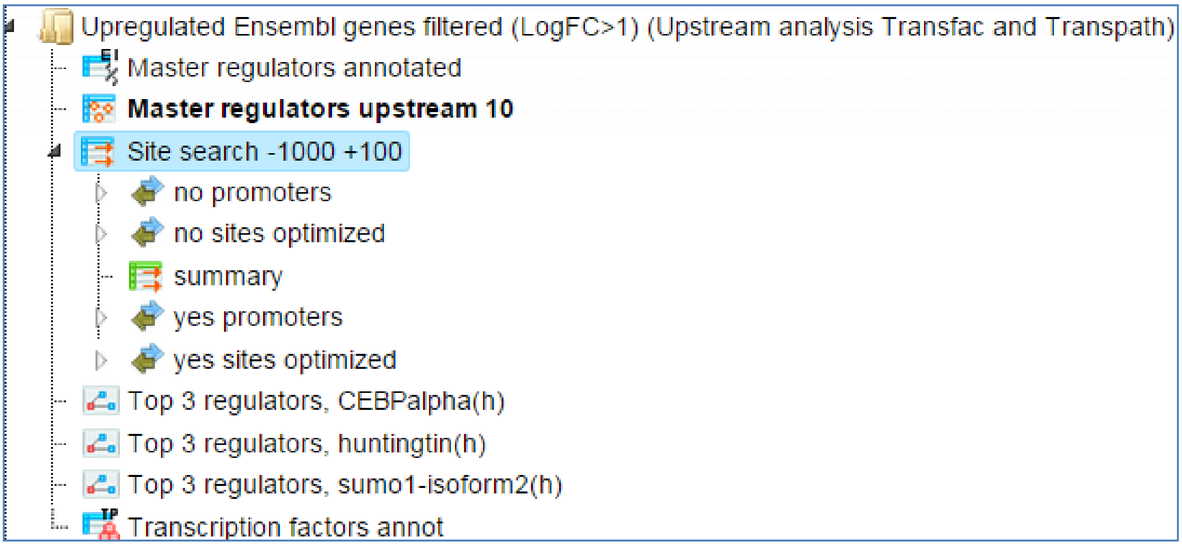


Tip for working with the diagrams

By default network diagrams are shown in the vertical hierarchical layout. The layout can be interactively changed into horizontal hierarchical, or force directed, or orthogonal layouts as described in Chapter 23. Expression data can be mapped on the diagrams as described in the Section 21.3.

Results of the promoter analysis

Along with the master regulatory molecules, this workflow returns the results of the promoter analysis, including TFBSs enriched in the promoters of the Yes set as compared with the No set, see *summary* ( ). The tracks with the Yes and No promoters and with the TF binding sites (⇦) are also included in the output. In the screenshot below the results of this workflow are shown with the subfolder Site search -1000 +100 opened:



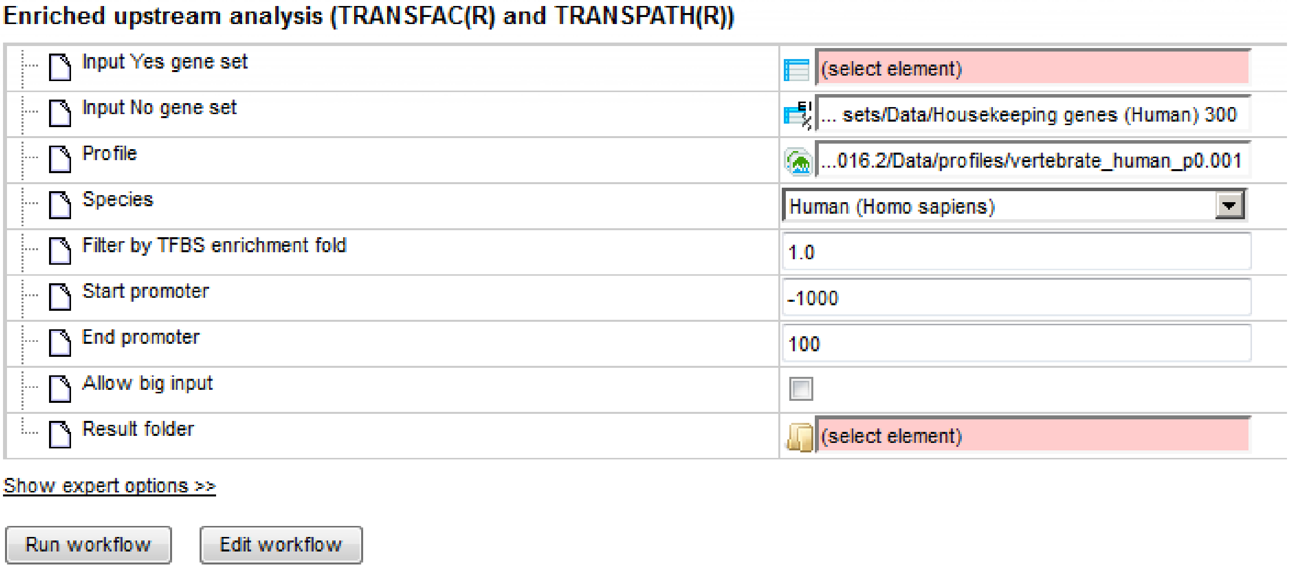
Upregulated Ensembl genes filtered (LogFC>1) (Upstream analysis Transfac and Transpath)
- Master regulators annotated
- **Master regulators upstream 10**
- Site search -1000 +100
  - no promoters
  - no sites optimized
  - summary
  - yes promoters
  - yes sites optimized
- Top 3 regulators, CEBPalpha(h)
- Top 3 regulators, huntingtin(h)
- Top 3 regulators, sumo1-isoform2(h)
- Transcription factors annot

The corresponding tables and tracks are described in detail in Section 16.2.5.

> **Note**. This workflow is available together with valid TRANSFAC® and TRANSPATH® licenses. Please feel free to ask for details (info@genexplain.com).

11.2.2. Enriched upstream analysis

This workflow enables a complete upstream analysis using the newest algorithm to detect enriched transcription factor binding sites (version 2.0), resulting in the identification of master regulators upstream from the transcriptional key molecules. To launch the workflow, open the workflow input form from the Start page:



Enriched upstream analysis (TRANSFAC(R) and TRANSPATH(R))

Input Yes gene set	(select element)
Input No gene set	... sets/Data/Housekeeping genes (Human) 300
Profile	...016.2/Data/profiles/vertebrate_human_p0.001
Species	Human (Homo sapiens)
Filter by TFBS enrichment fold	1.0
Start promoter	-1000
End promoter	100
Allow big input	□
Result folder	(select element)

Step 1: Specify a gene set under study, e.g. a list of differentially regulated genes, as the Input Yes gene set. You can drag & drop it from your project within the Tree Area and drop in the pink box of the field **Input gene set**.

For this example, all further steps are demonstrated by means of the following input set:

http://genexplain-platform.com:8080/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/New%20workflow%203.1.1/Upregulated%20Ensembl%20genes%20filtered_2

Step 2: Input a No gene set. This is the set of background genes or control set. The default No set used for this workflow is data/Examples/Sample data/Data/Housekeeping genes (Human). If your Yes set is from mouse or rat, you may wish to adjust the No set accordingly.

Step 3: Define a TRANSFAC® profile. The default profile is vertebrate_human_p0.001. Any other TRANSFAC® profile or user-specific profile can be chosen. With a mouse click on the field **Profile**, a pop-up window will open, where a profile can be selected.

Step 4: Specify the biological species of the input set in the field **Species** by selecting the required species from the drop-down menu.

Step 5: Filter by TFBS enrichment fold: In this field you can specify the enrichment fold (FE) to filter the motifs. By default, FE is 1.0, which means all motifs with FE>1.0 will be reported in the resulting table and the same motifs will serve to create a specific profile. If you want to use highly-enriched motifs, you can specify higher thresholds, e.g. 1.1, 1.2, or even 2.0 or 3.0 depending on your Yes and No sets. It is recommended that you run the workflow with default parameters first, check the results, and then run again with the desired filter value.

Step 6: Define the length of the promoter regions to be analyzed. The default promoter region is from-1000 to 100 relative to the TSS as annotated in the Ensembl database. You can adjust **Start of promoter** and **End of promoter** by typing in the corresponding fields.

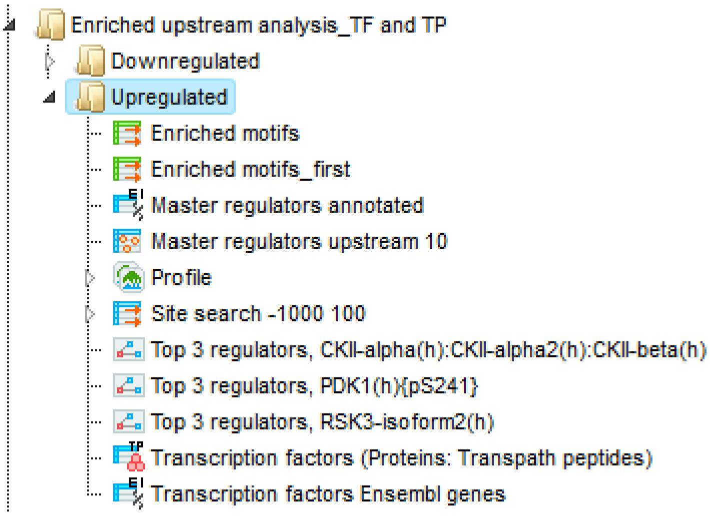
Step 7: Checking **Allow big input** enables analysis of more than 500 promoters.

Step 8: Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink box (select element) in the field **Results folder**, and a new window will open, where you can select the location of the results folder and define its name.

Step 9: Press the [Run workflow] button. Wait until the workflow is completed, and take a look at the results.

Visualization and interpretation of results

The result example folder can be found under data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/New workflow 3.1.1/Enriched upstream analysis_TF and TP/Upregulated; itcontains several files:

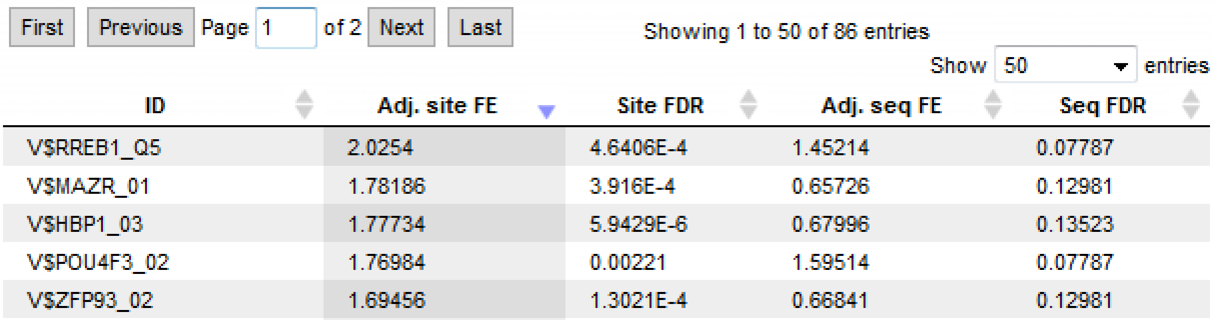


- Enriched upstream analysis_TF and TP
  - Downregulated
  - Upregulated
    - Enriched motifs
    - Enriched motifs_first
    - Master regulators annotated
    - Master regulators upstream 10
    - Profile
    - Site search -1000 100
    - Top 3 regulators, CKII-alpha(h):CKII-alpha2(h):CKII-beta(h)
    - Top 3 regulators, PDK1(h){pS241}
    - Top 3 regulators, RSK3-isoform2(h)
    - Transcription factors (Proteins: Transpath peptides)
    - Transcription factors Ensembl genes

Enriched motifs

The list of motifs, which were identified during the first part of the workflow and filtered with enrichment fold >1 can be found in the table **Enriched_motifs** ( ). It contains those site models, here TRANSFAC® matrices, which are enriched in the Yes set in comparison with the No set as shown below.

The example has 86 detected motifs with enrichment fold >1.



ID	Adj. site FE	Site FDR	Adj. seq FE	Seq FDR
V$RREB1_Q5	2.0254	4.6406E-4	1.45214	0.07787
V$MAZR_01	1.78186	3.916E-4	0.65726	0.12981
V$HBP1_03	1.77734	5.9429E-6	0.67996	0.13523
V$POU4F3_02	1.76984	0.00221	1.59514	0.07787
V$ZFP93_02	1.69456	1.3021E-4	0.66841	0.12981

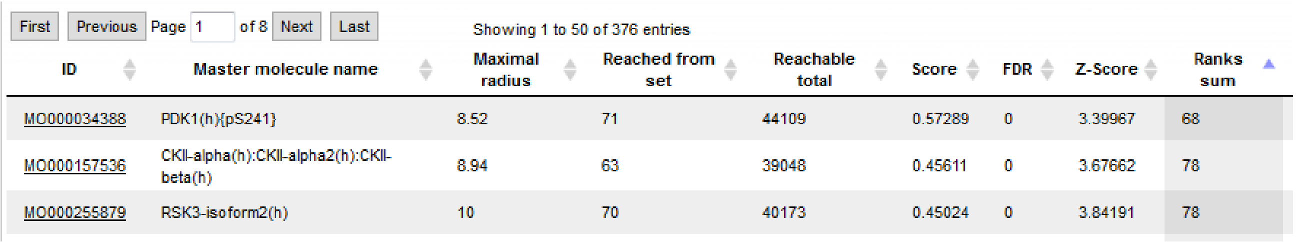
Please refer to section 10.4.3 to learn more about the site enrichment results. The table **Profile** ( ) presents details for PWMs with *adj. site FE >1*. This profile is an intermediate result of the workflow and is used further for *Site search on gene set* analysis.

Site search analysis output ( ) serves to visualize enriched motifs in the promoters. This folder contains four tracks ( ). The output table *Transcription factors Ensembl genes* is a list of transcription factors linked to the enriched motifs. For each transcription factor, the Ensembl gene ID is provided, as well gene description, HGNC gene symbol, species, and site model (TRANSFAC® PWM name).

This list of transcription factors is the input for the second part of the workflow, the master regulator search.

Master regulators

The primary result table *Master regulators upstream 10* ( ) is a list of master regulatory molecules that were identified at a distance of up to 10 steps upstream of the input TFs. Each master regulatory molecule is characterized by a Score, Z-score, FDR, and Ranks Sum.



Showing 1 to 50 of 376 entries

ID	Master molecule name	Maximal radius	Reached from set	Reachable total	Score	FDR	Z-Score	Ranks sum
MO000034388	PDK1(h){pS241}	8.52	71	44109	0.57289	0	3.39967	68
MO000157536	CKII-alpha(h):CKII-alpha2(h):CKII-beta(h)	8.94	63	39048	0.45611	0	3.67662	78
MO000255879	RSK3-isoform2(h)	10	70	40173	0.45024	0	3.84191	78

Further details about the columns of this table and how to work further with it are given in Section 5.1.1. The selection of the best master regulatory molecules based on Score, Z-score and Ranks sum is explained therein under "Interpretation of the results".

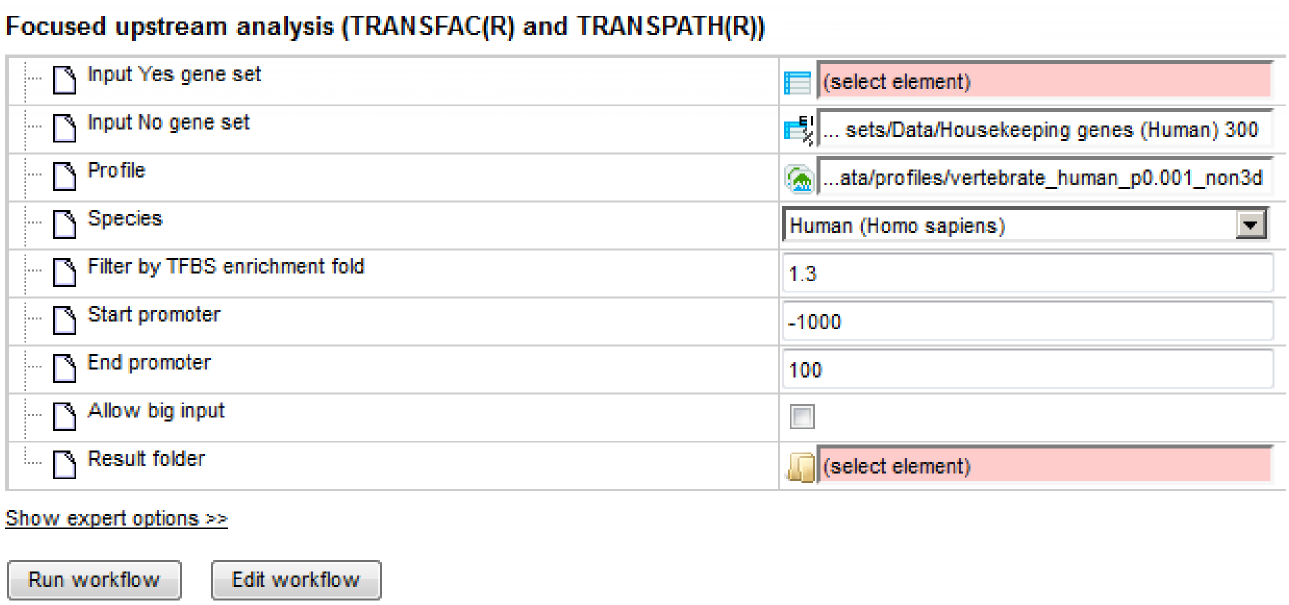
The three *Top 3 regulators* diagrams ( ) visualize the networks for each of the three top master regulators. By default, the top regulators are identified upon sorting the *Master regulators upstream 10* table ( ) by the column **Ranks sum** with the lowest rank on top.

Please refer to section 5.1.1 for more details about master regulator results.

Note*.* This workflow is available together with valid TRANSFAC® and TRANSPATH® licenses. Please feel free to ask for details (info@genexplain.com).

11.2.3. Focused upstream analysis

This workflow searches for enriched transcription factor binding sites (TFBSs), and selects those transcription factors (TFs), which were detected via direct (from input genes) regulator search method. To launch the workflow, open the workflow input form from the Start page:



Focused upstream analysis (TRANSFAC(R) and TRANSPATH(R))

Input Yes gene set	(select element)
Input No gene set	... sets/Data/Housekeeping genes (Human) 300
Profile	...ata/profiles/vertebrate_human_p0.001_non3d
Species	Human (Homo sapiens)
Filter by TFBS enrichment fold	1.3
Start promoter	-1000
End promoter	100
Allow big input	□
Result folder	(select element)

Step 1: Specify a gene set under study, e.g. a list of differentially regulated genes, as the Input Yes gene set. You can drag & drop it from your project within the Tree Area and drop in the pink box of the field **Input Yes gene set**.

For this example, all further steps are demonstrated with the following input set:

http://genexplain-platform.com:8080/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Experiment%20normalized%20(RMA)%20(Differentially%20expressed%20genes%20Affy)/UpDownReg%20Ensembl%20genes

Step 2: Input an **Input No gene set**. This is the set of background genes or control set. The default No set used for this workflow is data/Examples/Sample data/Data/Housekeeping genes (Human). If your Yes set is from mouse or rat, you may wish to adjust the No set accordingly.

Step 3: Define a TRANSFAC® profile. The default profile is vertebrate_human_p0.001_non3d. Any other TRANSFAC® profile or user-specific profile can be chosen. With a mouse click on the field **Profile**, a pop-up window will open, where a profile can be selected.

Step 4: Specify the biological species of the input set in the field **Species** by selecting the desired species from the drop-down menu.

Step 5: Filter by TFBS enrichment fold: In this field you can specify the enrichment fold (FE) to filter the motifs. By default, it is 1.3, which means all motifs with FE>1.3 will be reported in the resulting table, and the same motifs will serve to create a specific profile. If you want to use highly-enriched motifs, you can specify higher thresholds, e.g. 1.5, 1.6 etc., or even 2.0 or 3.0 depending on your Yes and No sets. It is recommended that you run it with default parameters first, check the results, and then repeat with the desired filter value.

Step 6: Define the promoter regions to be analyzed. The default promoter region is from -1000 to 100 relative to the TSS as annotated in the Ensembl database. You can adjust **Start of promoter** and **End of promoter** by typing in the corresponding fields.

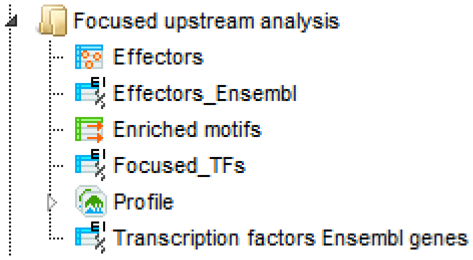
Step 7: Checking **Allow big input** enables analysis of more than 500 promoters.

Step 8: Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink box (select element) in the field **Results folder**, and a new window will open, where you can select the location of the results folder and define its name.

Step 9: Press the [Run workflow] button. Wait until the workflow is completed, and take a look at the results.

Interpretation of results

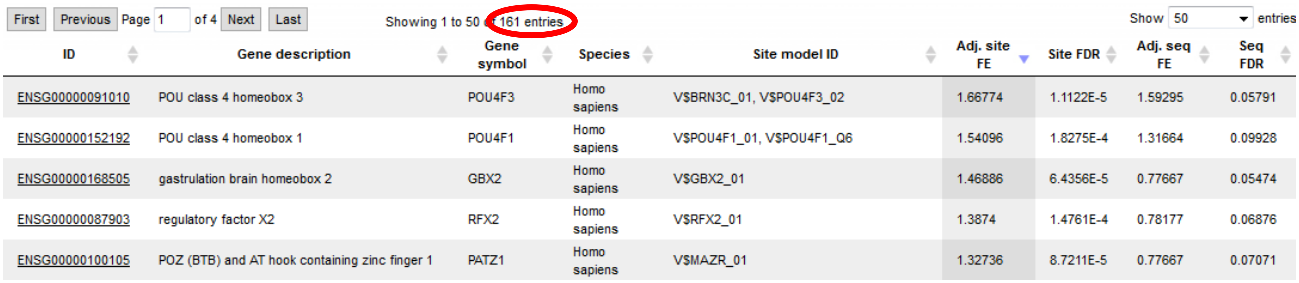
The result folder contains several files and one profile (collection of matrices):



- Focused upstream analysis
  - Effectors
  - Effectors_Ensembl
  - Enriched motifs
  - Focused_TFs
  - Profile
  - Transcription factors Ensembl genes

Enriched motifs

The list of motifs, which were found during the first part of the workflow, and were filtered with enrichment fold >1.3, can be found in the table *Enriched_motifs* ( ). It contains those site models, here TRANSFAC® matrices, which are enriched in the Yes set in comparison with the No set. The example has 260 detected motifs with an enrichment fold >1.3. The table *Transcription factors Ensembl genes* includes the corresponding 161 TFs from the site models as shown below. The *Profile* contains the matrix collection of converted and filtered site models.



Showing 1 to 50 of 161 entries

Show 50 entries

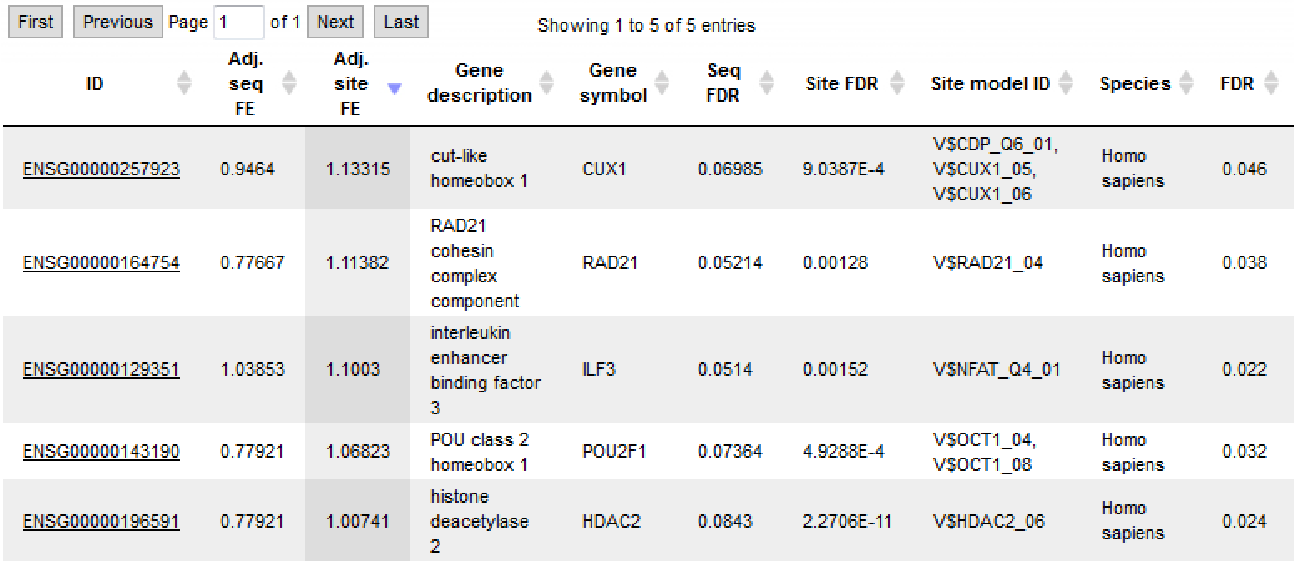
ID	Gene description	Gene symbol	Species	Site model ID	Adj. site FE	Site FDR	Adj. seq FE	Seq FDR
ENSG00000091010	POU class 4 homeobox 3	POU4F3	Homo sapiens	V$BRN3C_01, V$POU4F3_02	1.66774	1.1122E-5	1.59295	0.05791
ENSG00000152192	POU class 4 homeobox 1	POU4F1	Homo sapiens	V$POU4F1_01, V$POU4F1_Q6	1.54096	1.8275E-4	1.31664	0.09928
ENSG00000168505	gastrulation brain homeobox 2	GBX2	Homo sapiens	V$GBX2_01	1.46886	6.4356E-5	0.77667	0.05474
ENSG00000087903	regulatory factor X2	RFX2	Homo sapiens	V$RFX2_01	1.3874	1.4761E-4	0.78177	0.06876
ENSG00000100105	POZ (BTB) and AT hook containing zinc finger 1	PATZ1	Homo sapiens	V$MAZR_01	1.32736	8.7211E-5	0.77667	0.07071

Effectors

The result table *Effectors* ( ) is a list of identified regulatory molecules found with effector search method from the input gene list. Each effector molecule is characterized by a Score, Z-score, FDR, and Ranks Sum.

Focused transcription factors

The final table of *focused_TFs* is a list of 5 linking molecules between detected TFs and the regulatory elements from effector search analysis. The final list contains CUX1, RAD21, ILF3, POU2F1 and HDAC2 as upstream regulators of the identified TFs.



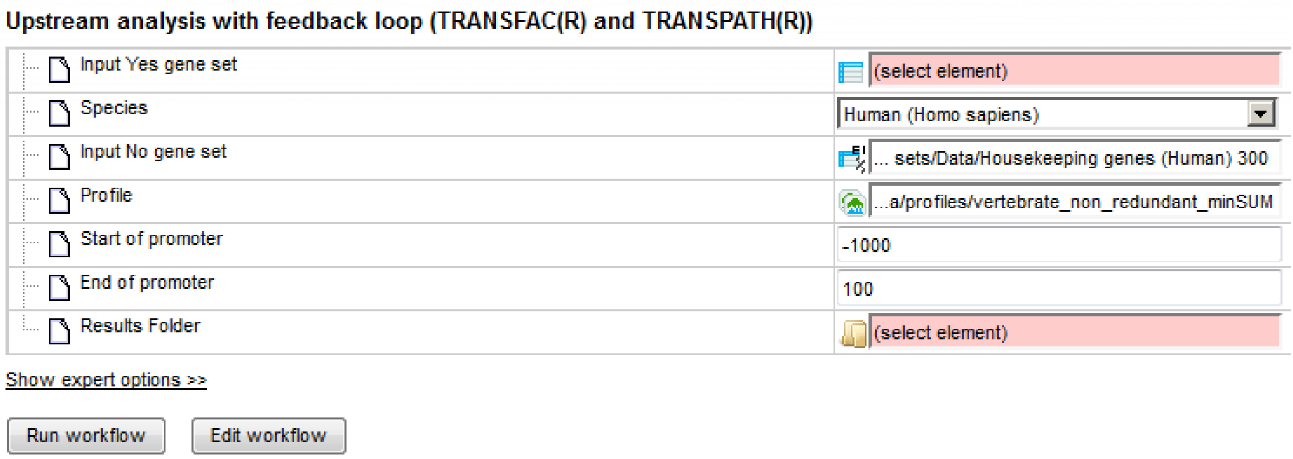
First | Previous | Page 1 of 1 | Next | Last   Showing 1 to 5 of 5 entries

ID	Adj. seq FE	Adj. site FE	Gene description	Gene symbol	Seq FDR	Site FDR	Site model ID	Species	FDR
ENSG00000257923	0.9464	1.13315	cut-like homeobox 1	CUX1	0.06985	9.0387E-4	VCDP_Q6_01, VCUX1_05, V$CUX1_06	Homo sapiens	0.046
ENSG00000164754	0.77667	1.11382	RAD21 cohesin complex component	RAD21	0.05214	0.00128	V$RAD21_04	Homo sapiens	0.038
ENSG00000129351	1.03853	1.1003	interleukin enhancer binding factor 3	ILF3	0.0514	0.00152	V$NFAT_Q4_01	Homo sapiens	0.022
ENSG00000143190	0.77921	1.06823	POU class 2 homeobox 1	POU2F1	0.07364	4.9288E-4	V$OCT1_04, V$OCT1_08	Homo sapiens	0.032
ENSG00000196591	0.77921	1.00741	histone deacetylase 2	HDAC2	0.0843	2.2706E-11	V$HDAC2_06	Homo sapiens	0.024

Note. This workflow is available together with valid TRANSFAC® and TRANSPATH® licenses. Please feel free to ask for details (info@genexplain.com).

11.2.4. Upstream analysis with feedback loop

This workflow enables a complete upstream analysis, detecting enriched transcription factor binding sites and resulting in the identification of master regulators upstream from the transcriptional regulators. Some master regulators with expression values (fold changes) from the input set are identified (with feedback loop). The results of this workflow include master regulators from all transcription factors and master regulators only with expression values from the input set (=feedback loop). To launch the workflow, open the workflow input form from the Start page:



Upstream analysis with feedback loop (TRANSFAC(R) and TRANSPATH(R))

Input Yes gene set	(select element)
Species	Human (Homo sapiens)
Input No gene set	... sets/Data/Housekeeping genes (Human) 300
Profile	...a/profiles/vertebrate_non_redundant_minSUM
Start of promoter	-1000
End of promoter	100
Results Folder	(select element)

Step 1: Specify a gene set under study, e.g. a list of differentially regulated genes, as the Input Yes gene set. You can drag & drop it from your project within the Tree Area and drop in the pink box of the field **Input gene set**.

For this example, all further steps are demonstrated with the following input set:

http://genexplain-platform.com:8080/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Experiment%20normalized%20(RMA)%20(Differentially%20expressed%20genes%20Affy)/UpDownReg%20Ensembl%20genes

Step 2: Specify the biological species of the input set in the field **Species** by selecting the required species from the drop-down menu.

Step 3: Input a No gene set. This is the set of background genes or control set. The default No set used for this workflow is data/Examples/Sample data/Data/Housekeeping genes (Human). If your Yes set is from mouse or rat, you may wish to adjust the No set accordingly.

Step 4: Define a TRANSFAC® profile. The default profile is vertebrate_non_redundant_minSUM. Any other TRANSFAC® profile or user-specific profile can be chosen. With a mouse click on the field **Profile**, a pop-up window will open, where a profile can be selected.

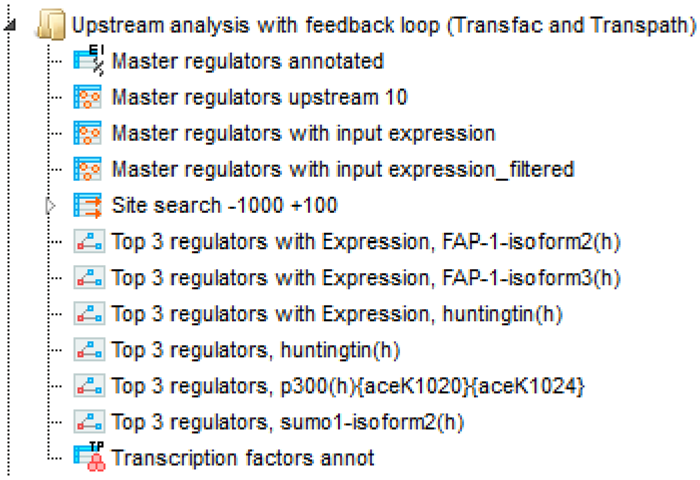
Step 5: Define the promoter regions to be analyzed. The default promoter region is from -1000 to 100 relative to the TSS as annotated in the Ensembl database. You can adjust **Start of promoter** and **End of promoter** by typing in the corresponding fields.

Step 6: Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink box (select element) in the field **Results folder**, and a new window will open, where you can select the location of the results folder and define its name.

Step 7: Press the [Run workflow] button. Wait until the workflow is completed, and take a look at the results.

Visualization and interpretation of results

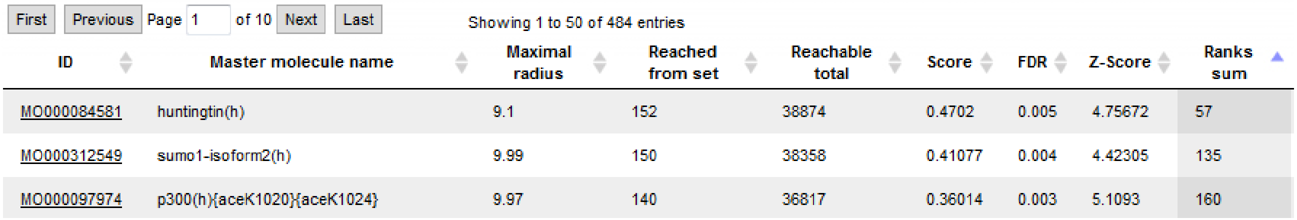
The example result folder is here: data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/Upstream analysis with feedback loop (Transfac and Transpath) and contains several files.



Upstream analysis with feedback loop (Transfac and Transpath)
- Master regulators annotated
- Master regulators upstream 10
- Master regulators with input expression
- Master regulators with input expression_filtered
- Site search -1000 +100
- Top 3 regulators with Expression, FAP-1-isoform2(h)
- Top 3 regulators with Expression, FAP-1-isoform3(h)
- Top 3 regulators with Expression, huntingtin(h)
- Top 3 regulators, huntingtin(h)
- Top 3 regulators, p300(h){aceK1020}{aceK1024}
- Top 3 regulators, sumo1-isoform2(h)
- Transcription factors annot

Master regulators

The primary result table *Master regulators upstream 10* (▦) is a list of master regulatory molecules that were identified at a distance of up to 10 steps upstream of the input TFs. Each master regulatory molecule is characterized by a Score, Z-score, FDR, and Ranks Sum.



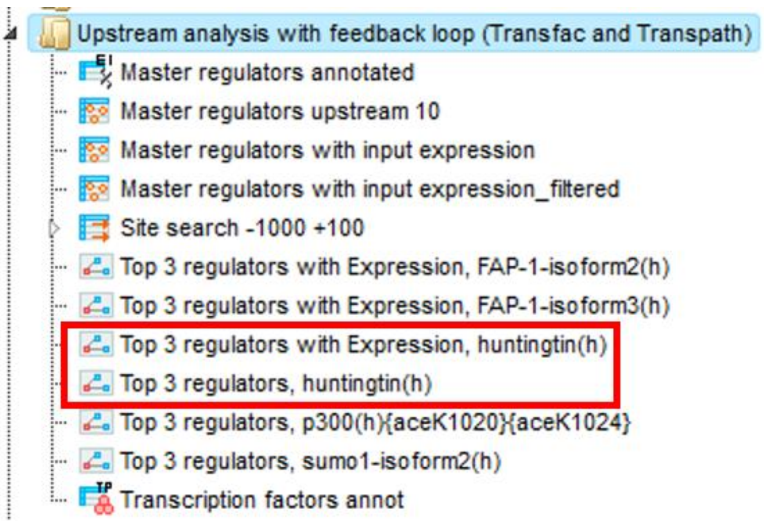
First Previous Page 1 of 10 Next Last Showing 1 to 50 of 484 entries

ID	Master molecule name	Maximal radius	Reached from set	Reachable total	Score	FDR	Z-Score	Ranks sum
MO000084581	huntingtin(h)	9.1	152	38874	0.4702	0.005	4.75672	57
MO000312549	sumo1-isoform2(h)	9.99	150	38358	0.41077	0.004	4.42305	135
MO000097974	p300(h){aceK1020}{aceK1024}	9.97	140	36817	0.36014	0.003	5.1093	160

Further details about the columns of this table and how to work further with it are given in Section 5.1.1. The selection of the best master regulatory molecules based on Score, Z-score and Ranks sum is explained therein under “Interpretation of the results”.

The three *Top 3 regulators* diagrams (⧉)visualize the networks for each of the three top master regulators. By default, the top regulators are identified upon sorting the *Master regulators upstream 10* table (▦) by the column **Ranks sum** with the lowest rank on top.

The three *Top 3 regulators* diagrams with Expression (⧉) visualize the networks for each of the three top master regulators, which have expression values from the input table. These master regulators with expression values are the results of a feedback loop; they are regulated by themselves. By default, the top regulators with expression values are identified upon sorting the *Master regulators with input expression_filtered* (▦) by the column **Ranks sum** with the lowest rank on top.



Upstream analysis with feedback loop (Transfac and Transpath)
- Master regulators annotated
- Master regulators upstream 10
- Master regulators with input expression
- Master regulators with input expression_filtered
- Site search -1000 +100
- Top 3 regulators with Expression, FAP-1-isoform2(h)
- Top 3 regulators with Expression, FAP-1-isoform3(h)
- Top 3 regulators with Expression, huntingtin(h)
- Top 3 regulators, huntingtin(h)
- Top 3 regulators, p300(h){aceK1020}{aceK1024}
- Top 3 regulators, sumo1-isoform2(h)
- Transcription factors annot

The output of the workflow shows the top master regulator huntingtin, which was found as a feedback-loop regulated molecule according to the input parameters.

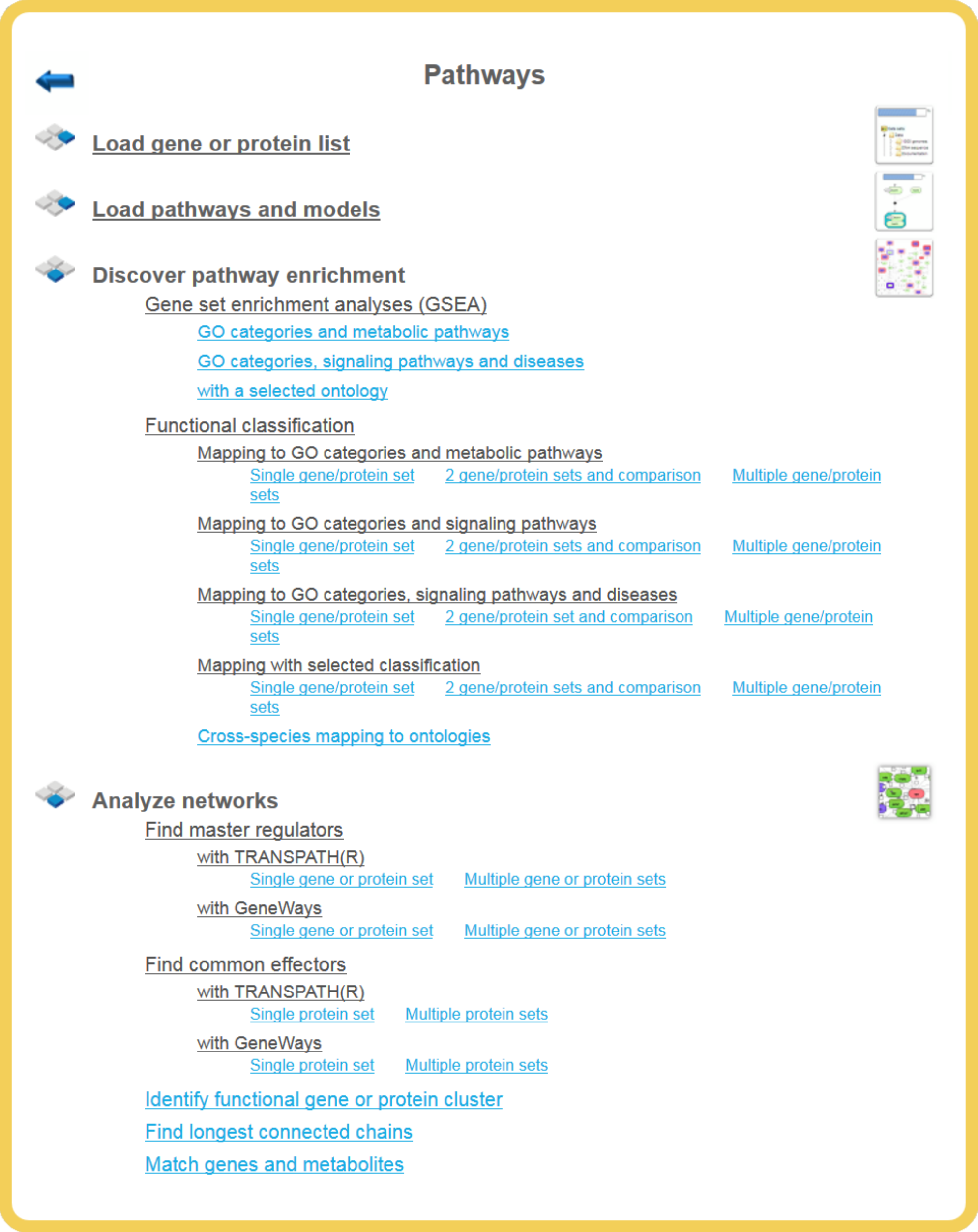
Note. This workflow is available together with valid TRANSFAC® and TRANSPATH® licenses. Please feel free to ask for details (info@genexplain.com).

11.3. Further workflows in this area

Load gene or protein list See Chapter 3

12. Pathways

All the workflows that constitute this Area have already been explained in other Chapters, where pathway-related functions contribute. Therefore, please refer to the Chapters and Section listed below.



Pathways

Load gene or protein list

Load pathways and models

Discover pathway enrichment

Gene set enrichment analyses (GSEA)

GO categories and metabolic pathways

GO categories, signaling pathways and diseases

with a selected ontology

Functional classification

Mapping to GO categories and metabolic pathways

Single gene/protein set 2 gene/protein sets and comparison Multiple gene/protein sets

Mapping to GO categories and signaling pathways

Single gene/protein set 2 gene/protein sets and comparison Multiple gene/protein sets

Mapping to GO categories, signaling pathways and diseases

Single gene/protein set 2 gene/protein set and comparison Multiple gene/protein sets

Mapping with selected classification

Single gene/protein set 2 gene/protein sets and comparison Multiple gene/protein sets

Cross-species mapping to ontologies

Analyze networks

Find master regulators

with TRANSPATH(R)

Single gene or protein set Multiple gene or protein sets

with GeneWays

Single gene or protein set Multiple gene or protein sets

Find common effectors

with TRANSPATH(R)

Single protein set Multiple protein sets

with GeneWays

Single protein set Multiple protein sets

Identify functional gene or protein cluster

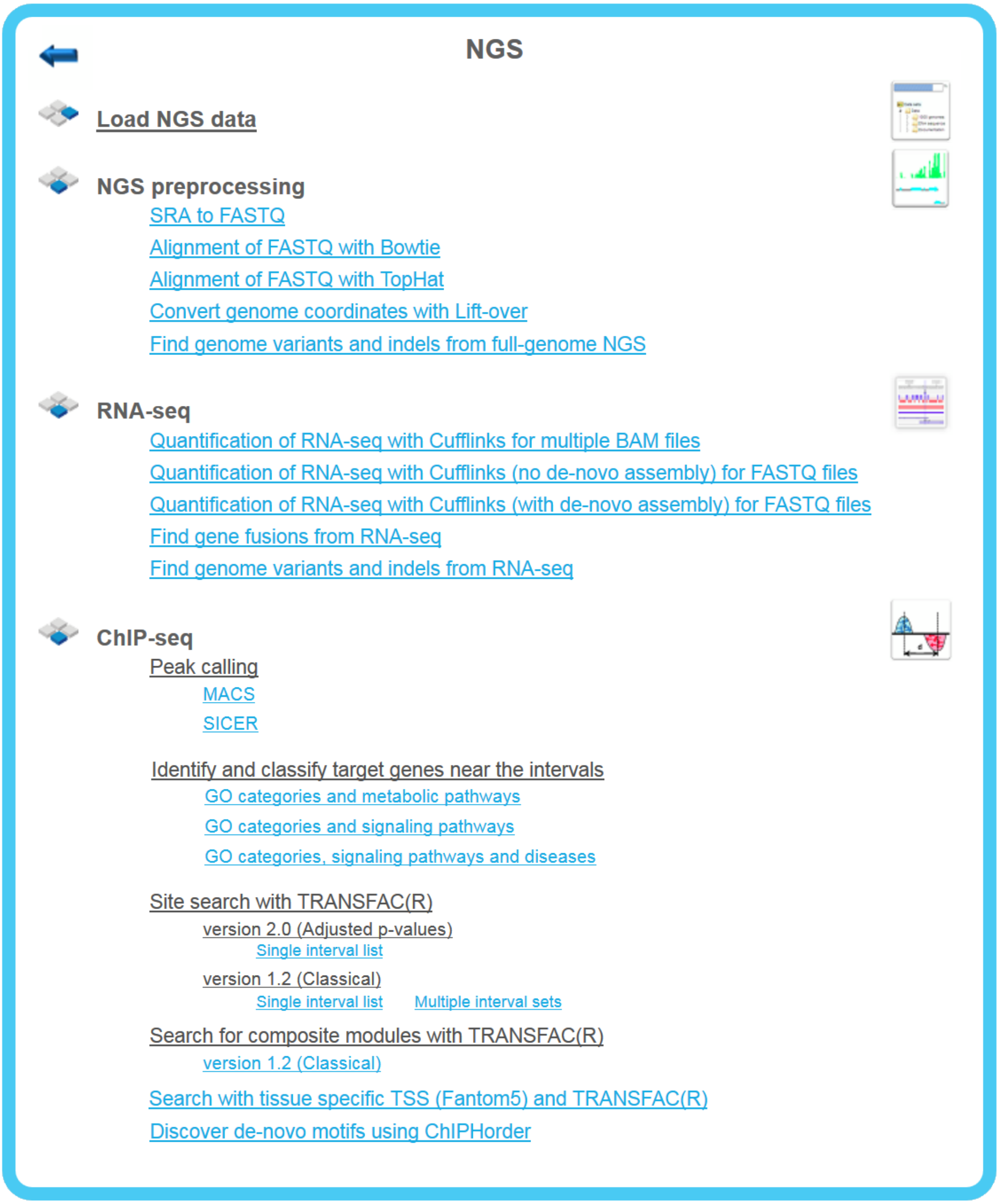
Find longest connected chains

Match genes and metabolites

Load gene or protein list,	See Chapter 3
Load pathways and models	
Discover pathway enrichment	See Section 10.3
Analyze networks	See Section 5.1

13. NGS

Many of the workflows that constitute this Area have already been explained in other Chapters, please refer to the Chapters and Section listed below.



NGS

Load NGS data

NGS preprocessing

- SRA to FASTQ
- Alignment of FASTQ with Bowtie
- Alignment of FASTQ with TopHat
- Convert genome coordinates with Lift-over
- Find genome variants and indels from full-genome NGS

RNA-seq

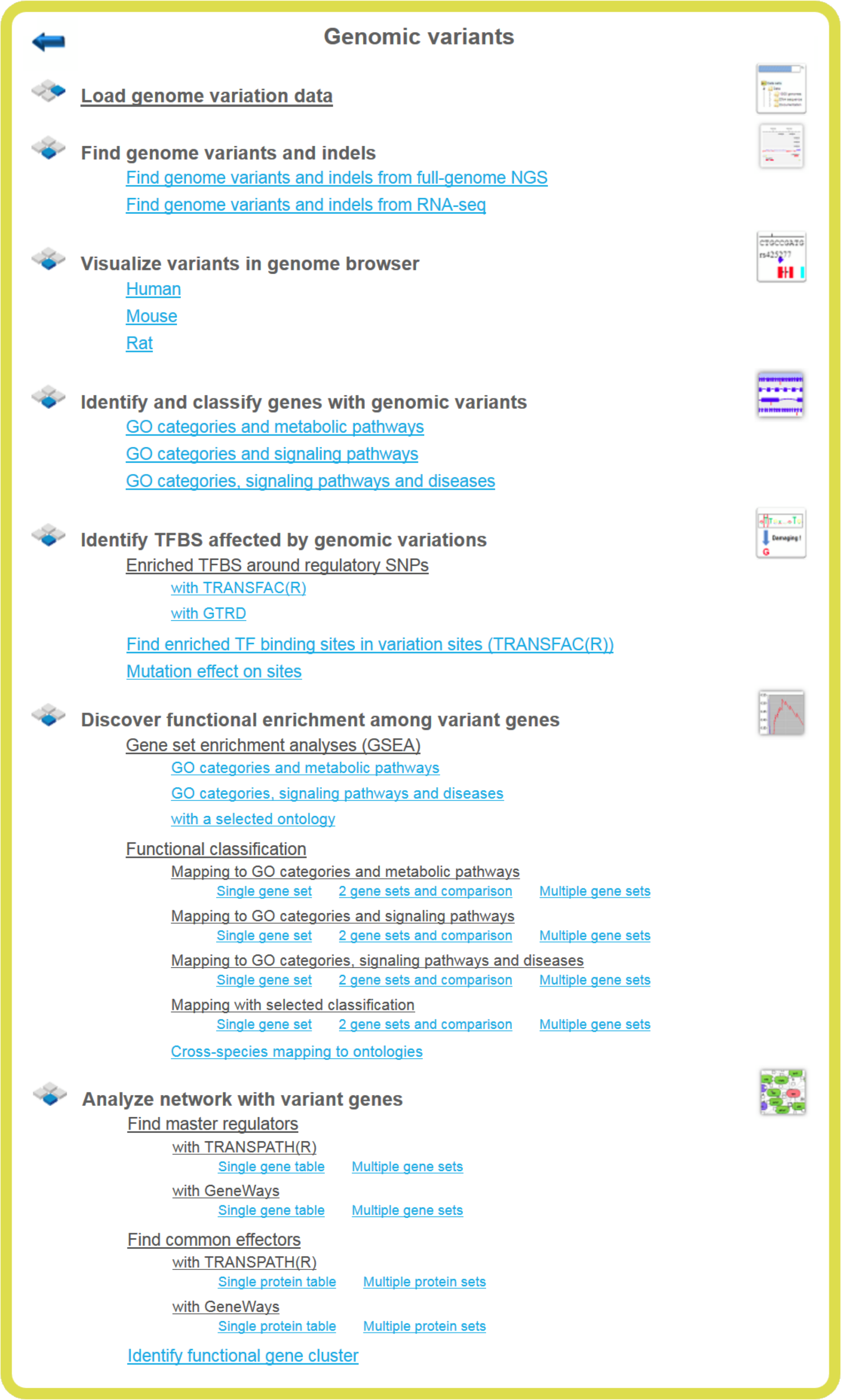
- Quantification of RNA-seq with Cufflinks for multiple BAM files
- Quantification of RNA-seq with Cufflinks (no de-novo assembly) for FASTQ files
- Quantification of RNA-seq with Cufflinks (with de-novo assembly) for FASTQ files
- Find gene fusions from RNA-seq
- Find genome variants and indels from RNA-seq

ChIP-seq

- Peak calling
  - MACS
  - SICER
- Identify and classify target genes near the intervals
  - GO categories and metabolic pathways
  - GO categories and signaling pathways
  - GO categories, signaling pathways and diseases
- Site search with TRANSFAC(R)
  - version 2.0 (Adjusted p-values)
    - Single interval list
  - version 1.2 (Classical)
    - Single interval list
    - Multiple interval sets
- Search for composite modules with TRANSFAC(R)
  - version 1.2 (Classical)
- Search with tissue specific TSS (Fantom5) and TRANSFAC(R)
- Discover de-novo motifs using ChIPHorder

Load gene or protein list	See Chapter 3
Discover pathway enrichment	See Section 10.3
ChIP-seq	See Chapter 7

14. Genomic variants



Genomic variants

Load genome variation data

Find genome variants and indels
- Find genome variants and indels from full-genome NGS
- Find genome variants and indels from RNA-seq

Visualize variants in genome browser
- Human
- Mouse
- Rat

Identify and classify genes with genomic variants
- GO categories and metabolic pathways
- GO categories and signaling pathways
- GO categories, signaling pathways and diseases

Identify TFBS affected by genomic variations
- Enriched TFBS around regulatory SNPs
  - with TRANSFAC(R)
  - with GTRD
- Find enriched TF binding sites in variation sites (TRANSFAC(R))
- Mutation effect on sites

Discover functional enrichment among variant genes
- Gene set enrichment analyses (GSEA)
  - GO categories and metabolic pathways
  - GO categories, signaling pathways and diseases
  - with a selected ontology
- Functional classification
  - Mapping to GO categories and metabolic pathways
    - Single gene set   2 gene sets and comparison   Multiple gene sets
  - Mapping to GO categories and signaling pathways
    - Single gene set   2 gene sets and comparison   Multiple gene sets
  - Mapping to GO categories, signaling pathways and diseases
    - Single gene set   2 gene sets and comparison   Multiple gene sets
  - Mapping with selected classification
    - Single gene set   2 gene sets and comparison   Multiple gene sets
  - Cross-species mapping to ontologies

Analyze network with variant genes
- Find master regulators
  - with TRANSPATH(R)
    - Single gene table   Multiple gene sets
  - with GeneWays
    - Single gene table   Multiple gene sets
- Find common effectors
  - with TRANSPATH(R)
    - Single protein table   Multiple protein sets
  - with GeneWays
    - Single protein table   Multiple protein sets
- Identify functional gene cluster

Genomic variants can be uploaded in the platform in different formats. The files uploaded in *bed* format are shown in the tree area as tracks ( ). Genomic variants can be also uploaded as a table, where the ID column contains standard SNP IDs (e.g. rs10010325). When imported into the platform, the tables with this type of ID have a special icon ( ) in the tree area. An example of such a table can be found in the *Examples* folder:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/SNPs%20linked%20to%20human%20height/Data/SNP_height_hg19

14.1. Find genome variants and indels from full-genome NGS

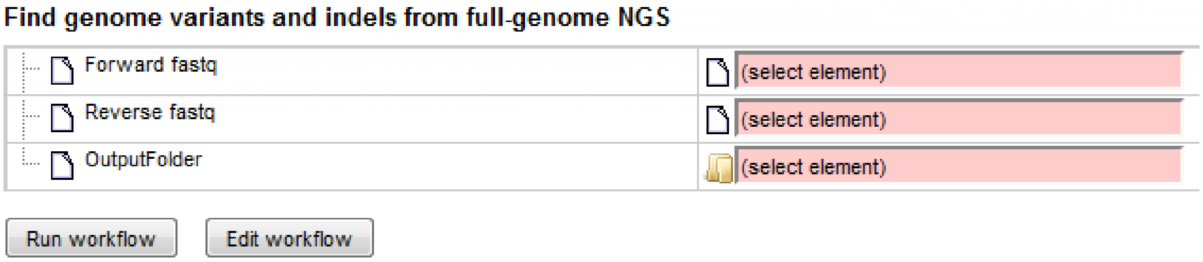
This workflow is based on a framework to discover genotype variations in full-genome NGS data by De Pristo et al., Nature Genetics 43:491-498, 2011. The process includes initial read mapping, local realignment around indels, base quality score recalibration, SNP discovery and genotyping to find all potential variants.

In the first part of the workflow the input sequences are mapped using the BWA tool (Galaxy). BWA is a fast light-weight tool that aligns relatively short sequences to a sequence database, such as the human reference genome (published by Li & Durbin, Bioinformatics 25:1754-1760, 2009).

The second part includes local realignment around indels, base quality score recalibration, SNP discovery and genotyping to find all potential variants. After the first part, and after identification of duplicates and covariates, the workflow creates a first output as a new BAM file. Then the recalibrated BAM file is used as an input for SNP discovery and genotyping to find all potential variants by GATK (Genome Analysis Toolkit).

To launch the workflow, follow these steps:

Step 1. Open the workflow input form from the Start page. It will open in the main Work Space and looks as shown below:



Find genome variants and indels from full-genome NGS

Forward fastq	(select element)
Reverse fastq	(select element)
OutputFolder	(select element)

Step 2. Input the **Forward** and **Reverse** fastq files. You can either drag&drop or select the files from the Tree area. Here, a set of files from the Example folder is used as input.

data/Examples/Chronic Myeloid Leukemia Patient Genotyping/Data/SRR944150 forward.fastq

data/Examples/Chronic Myeloid Leukemia Patient Genotyping/Data/SRR944150 reverse.fastq

Step 3. Specify the **OutputFolder** location and name and press the button [Run workflow].

All results are saved in the result folder:
data/Examples/Chronic Myeloid Leukemia Patient Genotyping/Data/SRR944150 forward.fastq (Genome variants and indels from RNA-seq)

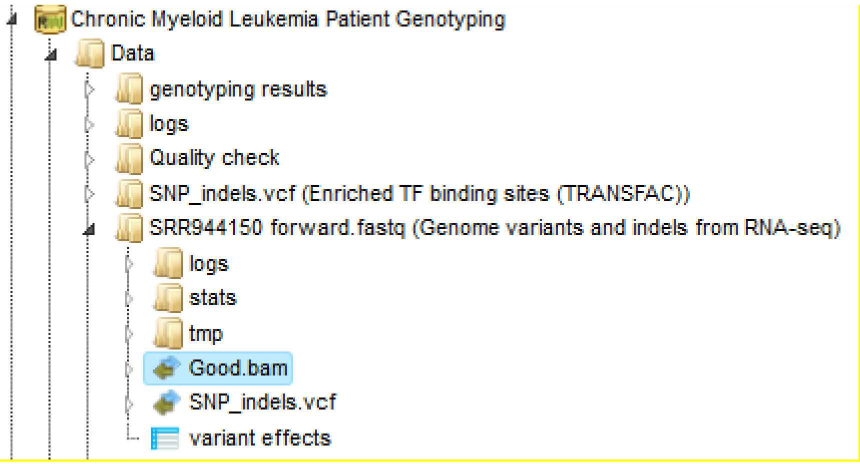
In the first step the input fastq sequences are subjected to the BWA method from Illumina. BWA is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome. It consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM. The first algorithm is designed for Illumina sequence reads up to 100bp, while the other two algorithms are designed for longer sequences ranging from 70bp to 1Mbp. BWA-MEM and BWA-SW share similar features such as long-read support and split alignment, but BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.

In the next step the sorted files are subjected to the Mark Duplicates method.

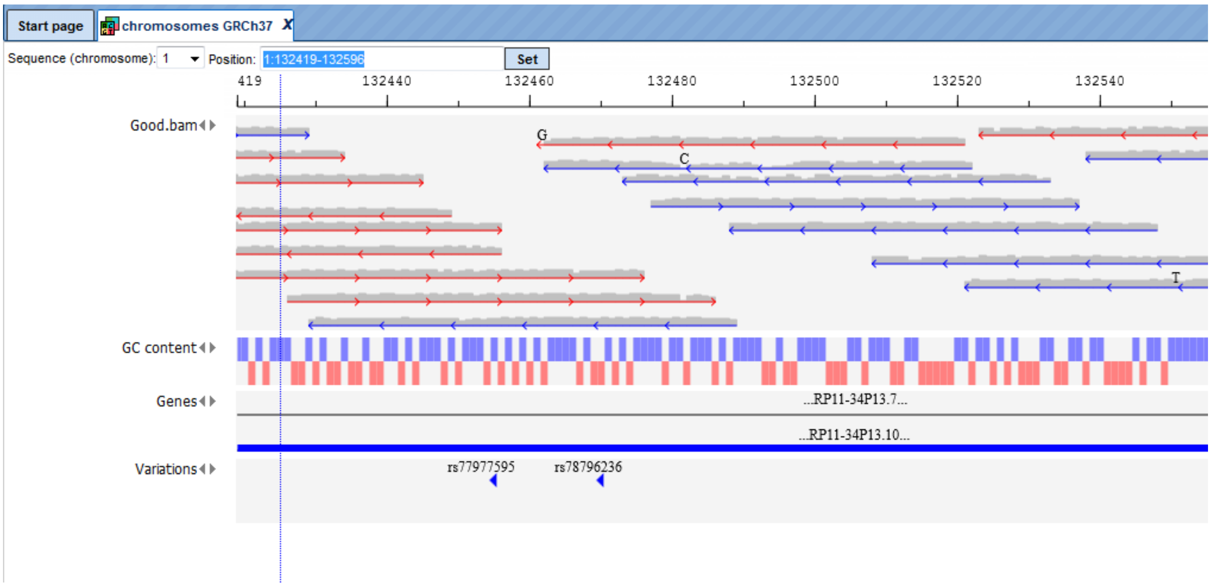
This method removes duplicates. The purpose is to mitigate the effects of PCR amplification bias introduced during library construction. Two read pairs are considered duplicate if they align to the same genomic position. The resulting MarkDuplikates1.log file is stored in the log folder and the MarkDuplikates1.stat file is stored in the stat folder.

The next step is a local realignment. Read mapping algorithms operate on each read independently, locally realigning reads such that the number of mismatching bases is minimized across all reads. Output files are Realigner.log and TargetCreator.log in the log folder, ddup1.bam, Realigned.bam and realigner.intervals in the tmp folder.

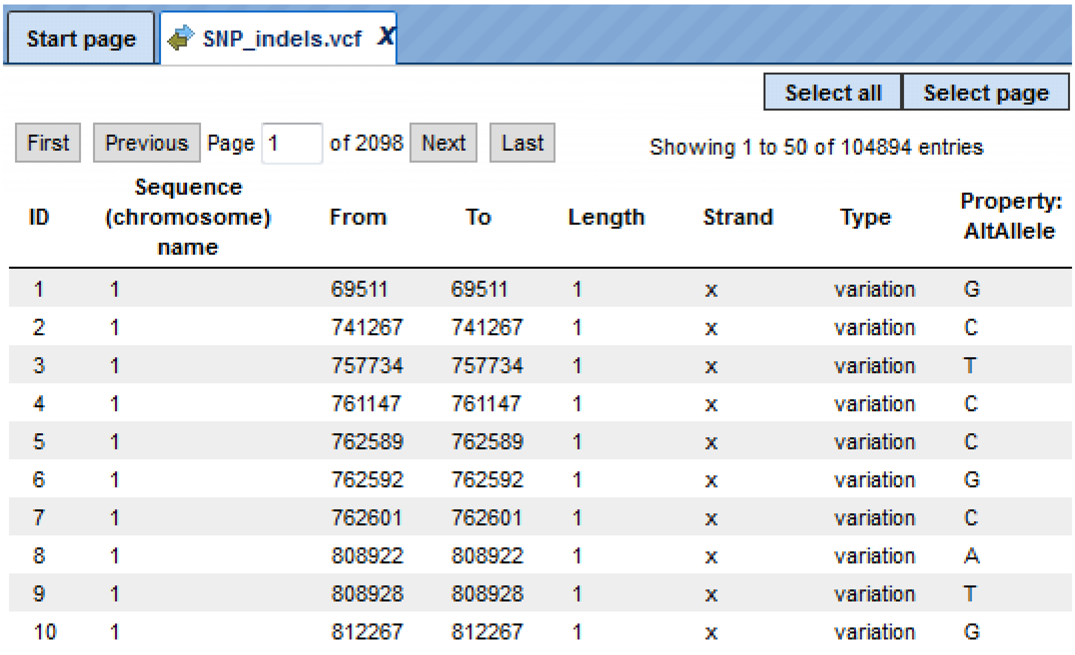
The realigned BAM file is used again to remove duplicates (output MarkDuplicates2.log and MarkDuplicates2.stat), because realignment may change genomic positions of read pairs, after this step additional duplicates can be identified. The next step is a recalibration of base quality values. For each base in each read the method calculates various covariates (such as reported quality score, position in read, dinucleotide, read GC-content). Using these values it builds the model that predicts sequencing errors. Then it applies this model to calculate an empirical base quality score and overwrites the phred quality score in the read. Output is a new BAM file (Good.bam).


Chronic Myeloid Leukemia Patient Genotyping
Data
genotyping results
logs
Quality check
SNP_indels.vcf (Enriched TF binding sites (TRANSFAC))
SRR944150 forward.fastq (Genome variants and indels from RNA-seq)
logs
stats
tmp
Good.bam
SNP_indels.vcf
variant effects

The user can view the Good.bam files in the genome browser by double-clicking on it. The browser shows each aligned read and also shows nucleotides mismatching between the reads and the reference genome sequence.

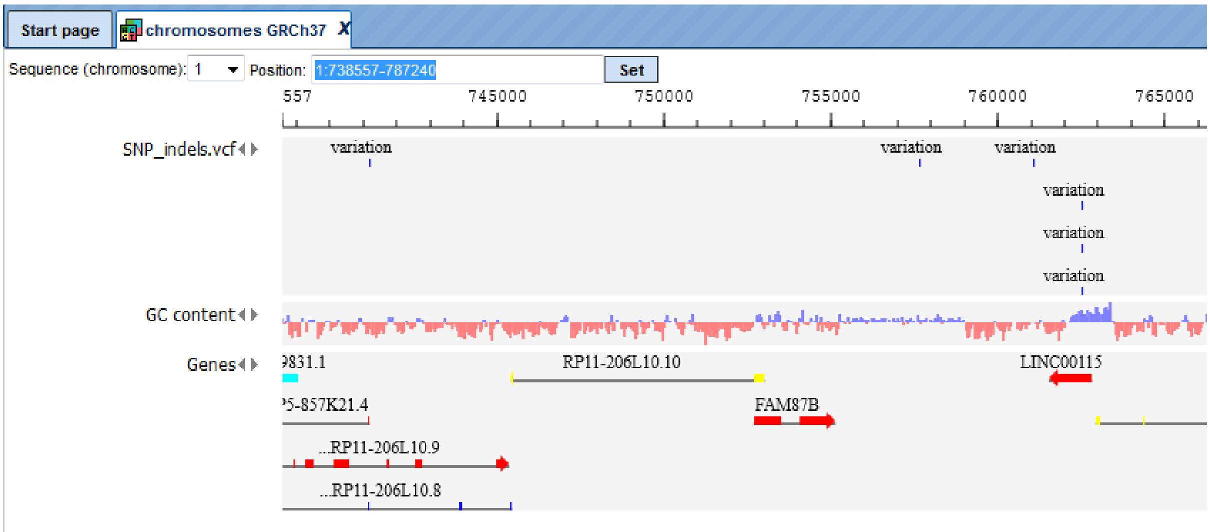


This file is used then for the unified GATK (Genome Analysis Toolkit) genotyper method to detect the SNP-indels. It generates a table in VCF format, which can be viewed either as a table or as a track in the genome browser (right mouse button click and select either "Open track" or "Open table").



Showing 1 to 50 of 104894 entries

ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: AltAllele
1	1	69511	69511	1	x	variation	G
2	1	741267	741267	1	x	variation	C
3	1	757734	757734	1	x	variation	T
4	1	761147	761147	1	x	variation	C
5	1	762589	762589	1	x	variation	C
6	1	762592	762592	1	x	variation	G
7	1	762601	762601	1	x	variation	C
8	1	808922	808922	1	x	variation	A
9	1	808928	808928	1	x	variation	T
10	1	812267	812267	1	x	variation	G

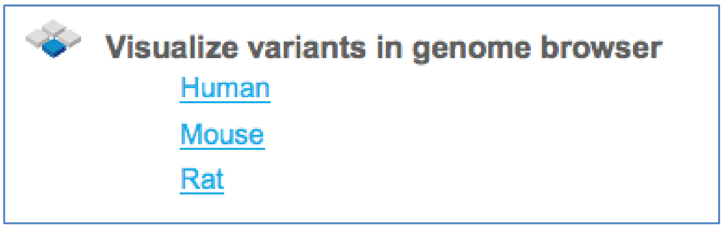


Sequence (chromosome): 1 Position: 1:738557-787240

In the track visualization the information about each variation (either a base substitution or an indel) is shown in the info box when clicking on each variation.

14.2. Visualize variants in genome browser

The genomic variants shown in the tree area as tracks ( ) can be directly visualized in the genome browser. Tables with SNP IDs ( ) should be first processed into the tracks. For this, you can apply the method called SNP matching ( ); for details please refer to the section 17.1.1.



Visualize variants in genome browser

Human

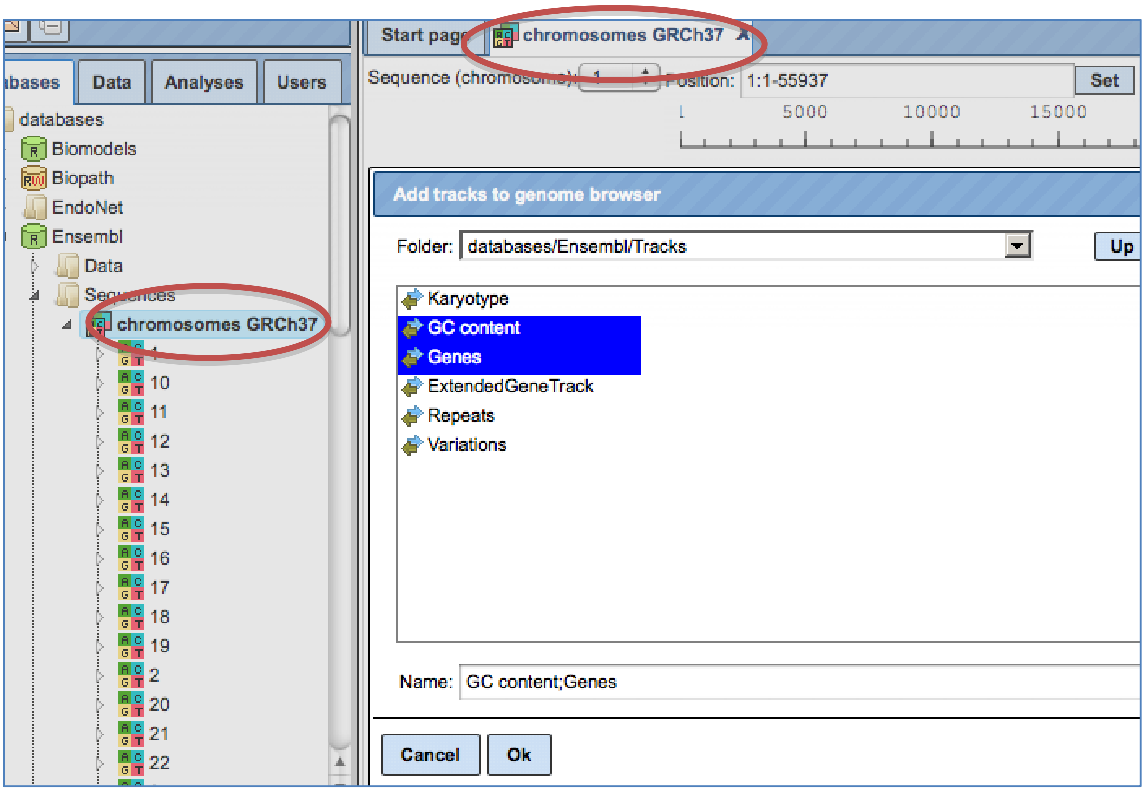
Mouse

Rat

A mouse click on the links Human, Mouse or Rat immediately opens up a genome browser for the corresponding species in the work space, and the corresponding Ensembl database appears in the tree area.

14.2.1. Human

When *Human* is selected, the genome browser opens up the latest Ensembl build, hg19 chromosomes GRCh37, highlighted by the red oval.

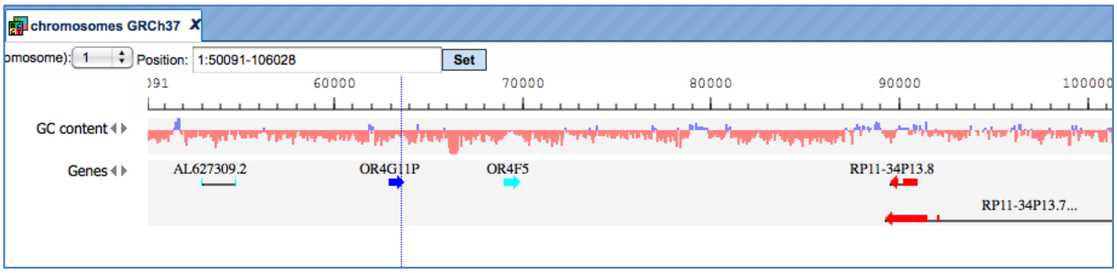


Sequence (chromosome): 1 Position: 1:1-55937

Folder: databases/Ensembl/Tracks

Name: GC content;Genes

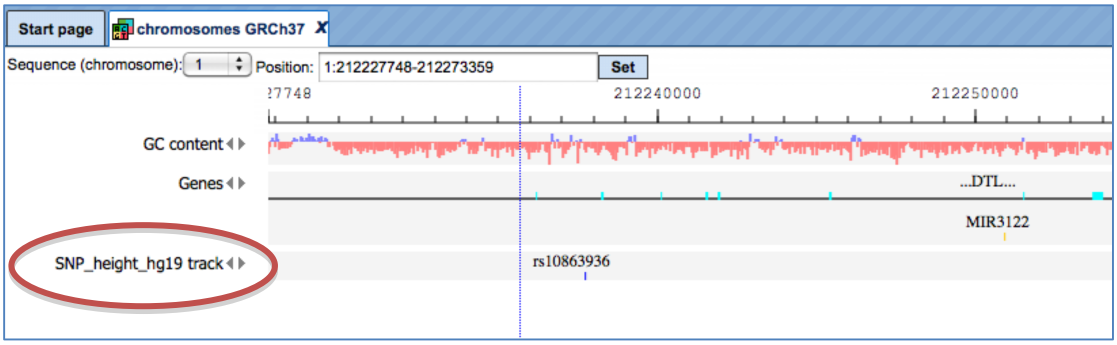
In the pop-up window *Add tracks to genome browser* you can select which tracks among those available in Ensembl should be opened together with your track of the genomic variants. Two tracks are selected by default, *GC-content* and *Genes*. When the selection is ready, push the [Ok] to get the following view:



omosome): 1 Position: 1:50091-106028

Now you can drag & drop your track with the genomic variants on the genome browser to add it to the default tracks. As an example, the following track is shown here, in the screenshot below:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/SNPs%20linked%20to%20human%20height/Data/SNP_height_hg19%20(Analyse%20SNP%20list%20(TRANSFAC))/All%20SNPs/SNP_height_hg19%20track

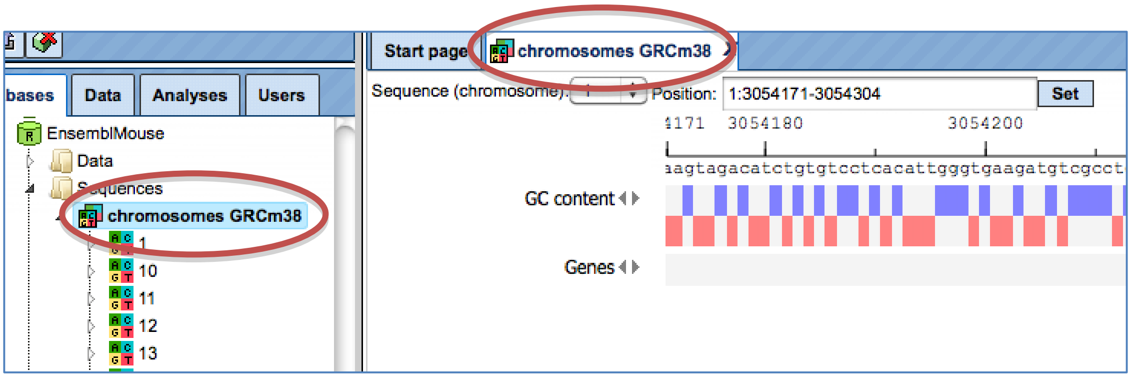


Sequence (chromosome): 1 Position: 1:212227748-212273359 Set

For further details regarding visualizations, please refer to the basic operations with tracks, Section 2.3.3.

14.2.2. Mouse

When *Mouse* is selected, the genome browser opens up the latest Ensembl build for mouse, mm10 chromosomes GRCm38

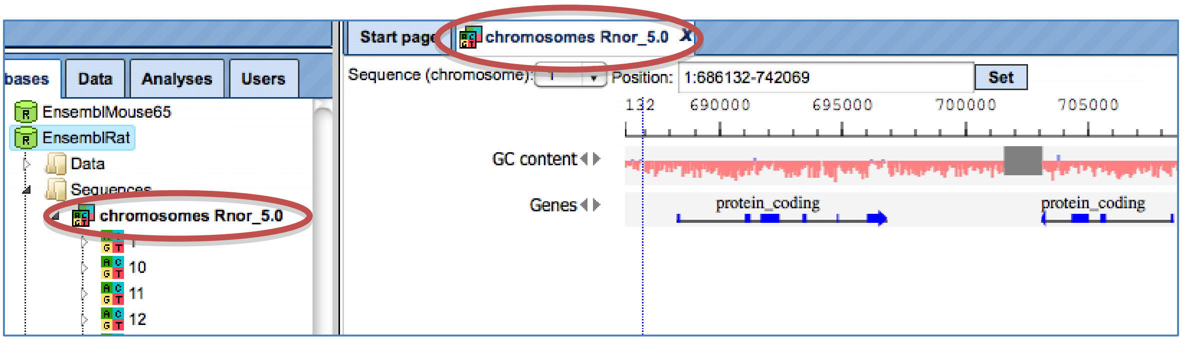


Sequence (chromosome): 1 Position: 1:3054171-3054304 Set

The further steps of the visualization are similar to the human tracks.

14.2.3. Rat

When *Rat* is selected, the genome browser opens up the latest Ensembl build for rat, rn5 chromosomes Rnor_5.0



Sequence (chromosome): 1 Position: 1:686132-742069 Set

The further steps of the visualization are similar to the human tracks.

14.3. Identify and classify genes with genomic variants

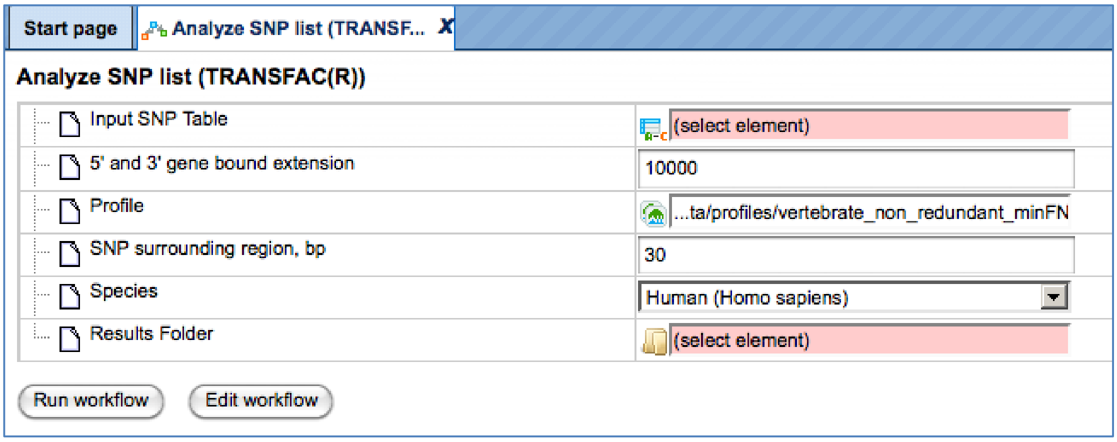
Genomic variants are represented in the same format as genome intervals or ChIP-seq peaks, i.e. with their absolute chromosomal positional locations. Therefore, please apply correspondingly the workflows explained in detail under 6.1.1 or 7.2.1.

14.4. Identify TFBS affected by genomic variations

14.4.1. Enriched TF sites around regulatory SNPs and SIFT analysis

Analysis with TRANSFAC®

The input form of this workflow, when opened form the Start page, is the following:



Analyze SNP list (TRANSFAC(R))

Input SNP Table	(select element)
5' and 3' gene bound extension	10000
Profile	...ta/profiles/vertebrate_non_redundant_minFN
SNP surrounding region, bp	30
Species	Human (Homo sapiens)
Results Folder	(select element)

Step 1. Specify an input table in the field **Input SNP table**. A table with standard SNP IDs in the format like *rs10010325* can be used as an input. The tables with this type of IDs have a special icon (A-c) in the tree area. In this example the following input table with 180 SNPs is used:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/SNPs%20linked%20to%20human%20height/Data/SNP_height_hg19

Step 2. Specify the region around each SNP in the field **5' and 3' gene bound extension**. By default this region is 10000 bp long. Genes located within the region of 10000 bp around each SNP in the input list will be considered as SNP target genes.

Step 3. Specify a TRANSFAC® profile in the field **Profile**. The workflow uses the default profile vertebrate_non_redundant_minFN from the TRANSFAC® library, but another TRANSFAC profile can be chosen as needed.

Step 4. Specify the region around each SNP that will be analyzed for potential TFBSs in the field **SNP surrounding region.** The default length of this region is 30 bp on each flank.

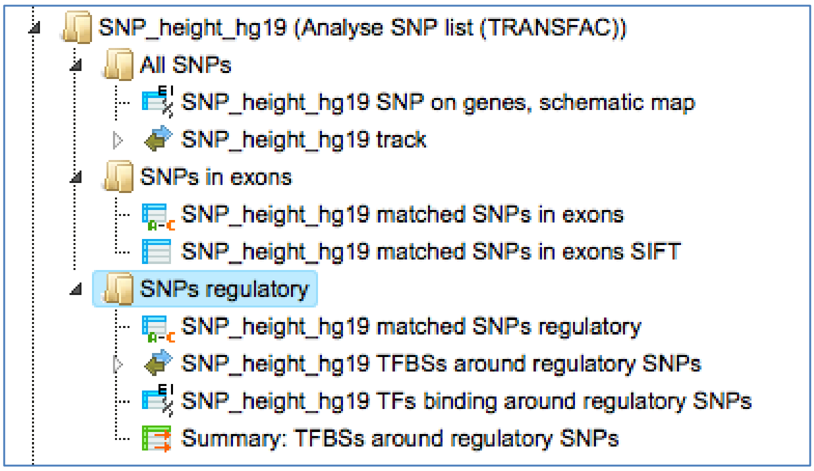
Step 5. Select a species corresponding to your input table from the drop-down menu in the field **Species**.

Step 6. Specify the path to store the results and the name of the output folder in the field **Results folder**.

Step 7. Having filled the input form, launch the analysis with the [Run] button. Wait till the workflow is completed.

Results

The output is a result folder with three subfolders named *all SNPs*, *SNPs in exons* and *SNPs regulatory*, respectively, containing all resulting tables and tracks:



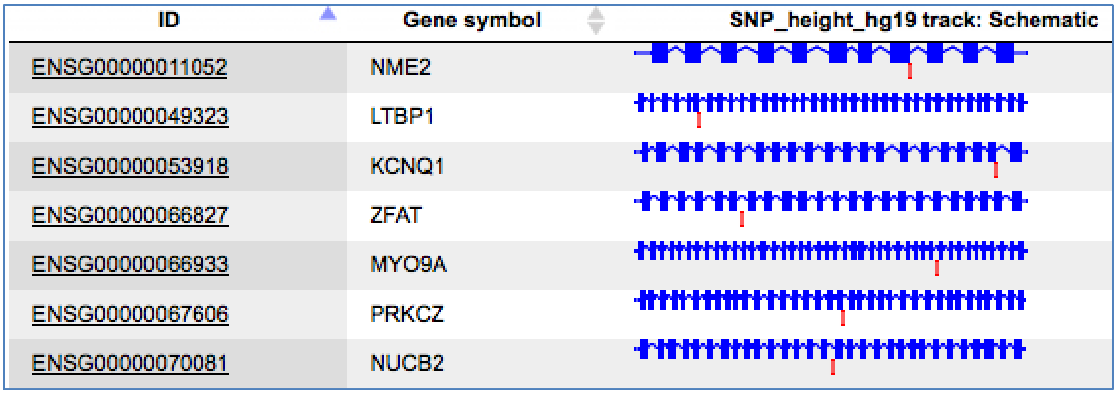
SNP_height_hg19 (Analyse SNP list (TRANSFAC))
- All SNPs
  - SNP_height_hg19 SNP on genes, schematic map
  - SNP_height_hg19 track
- SNPs in exons
  - SNP_height_hg19 matched SNPs in exons
  - SNP_height_hg19 matched SNPs in exons SIFT
- SNPs regulatory
  - SNP_height_hg19 matched SNPs regulatory
  - SNP_height_hg19 TFBSs around regulatory SNPs
  - SNP_height_hg19 TFs binding around regulatory SNPs
  - Summary: TFBSs around regulatory SNPs

The results shown here can be found in the trea area: data/Examples/SNPs linked to human height/Data/SNP_height_hg19 (Analyze SNP list (TRANSFAC))

Subfolder *All SNPs*

This folder includes one gene table and one track.

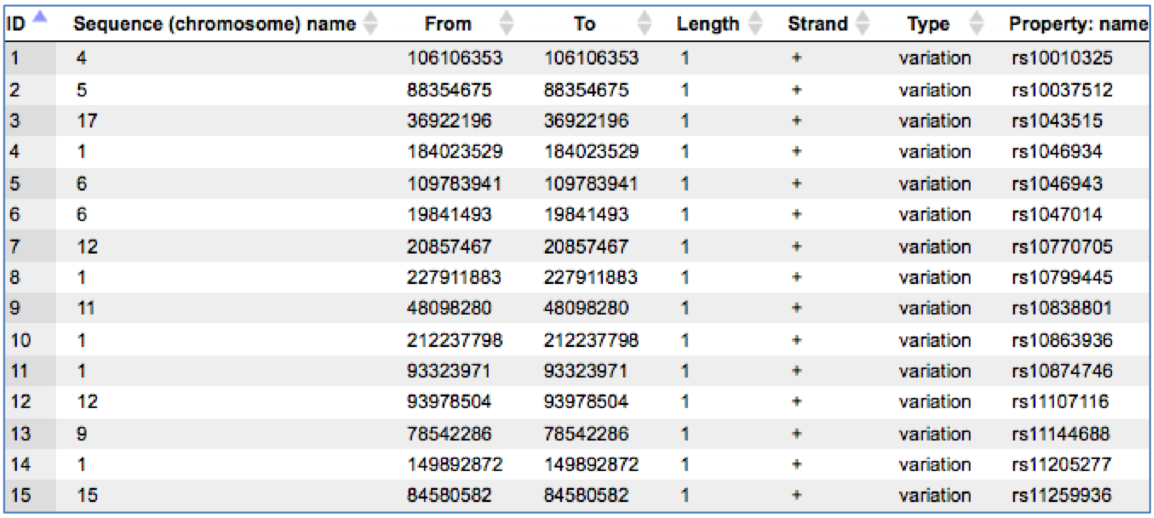
The table *SNPs on genes, schematic map* ( ) contains all genes that were identified in the region of 10000 bp on both flanks around each SNP, in this example 119 genes.



ID	Gene symbol	SNP_height_hg19 track: Schematic
ENSG00000011052	NME2	
ENSG00000049323	LTBP1	
ENSG00000053918	KCNQ1	
ENSG00000066827	ZFAT	
ENSG00000066933	MYO9A	
ENSG00000067606	PRKCZ	
ENSG00000070081	NUCB2	

Each row of this table corresponds to one gene; the column **ID** presents Ensembl gene IDs, and HGNC gene symbols are listed in the column **Gene symbol**. The title of the last column **Schematic** also contains the name of the input table. This column represents a schematic view for each gene, where blue boxes correspond to exons, and the lines between exons symbolize introns, drawn in logarithmic scale. SNPs are shown by vertical red lines. This schema provides an overview of SNP location within genes.

A track ( ) represents the results of the SNP mapping to genomic positions. In this example, out of 180 input SNPs 148 were mapped to the genome, and for them the following information is shown:



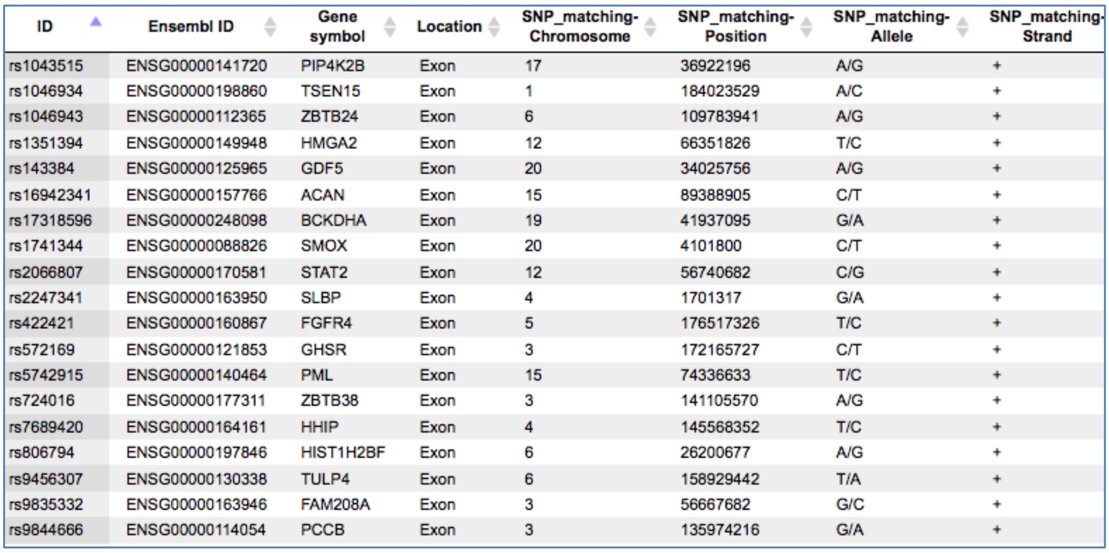
ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: name
1	4	106106353	106106353	1	+	variation	rs10010325
2	5	88354675	88354675	1	+	variation	rs10037512
3	17	36922196	36922196	1	+	variation	rs1043515
4	1	184023529	184023529	1	+	variation	rs1046934
5	6	109783941	109783941	1	+	variation	rs1046943
6	6	19841493	19841493	1	+	variation	rs1047014
7	12	20857467	20857467	1	+	variation	rs10770705
8	1	227911883	227911883	1	+	variation	rs10799445
9	11	48098280	48098280	1	+	variation	rs10838801
10	1	212237798	212237798	1	+	variation	rs10863936
11	1	93323971	93323971	1	+	variation	rs10874746
12	12	93978504	93978504	1	+	variation	rs11107116
13	9	78542286	78542286	1	+	variation	rs11144688
14	1	149892872	149892872	1	+	variation	rs11205277
15	15	84580582	84580582	1	+	variation	rs11259936

For each SNP the tabulated view of the track contains information about chromosomal location, absolute positions, length, and strand. In the column **Type** the value *variation* is shown for all SNPs, and in the column **Property: name** SNP IDs are shown.

Subfolder SNPs in exons

This folder includes two tables, both present information for those SNPs that are located in exons. In our example, 19 out of 148 SNPs mapped to the genome are located in exons.

One of the tables contains standard SNP IDs in the **ID** column, and has the same icon as the input SNP table ( ).



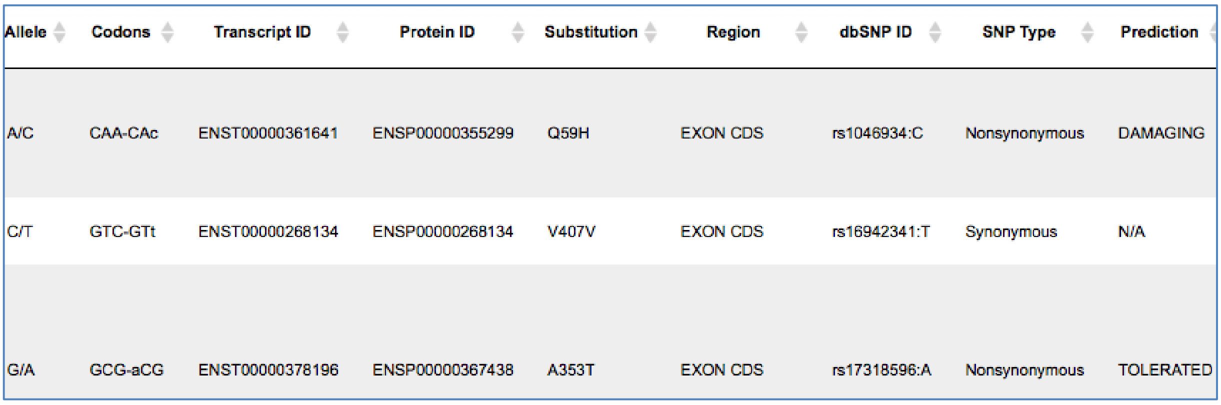
ID	Ensembl ID	Gene symbol	Location	SNP_matching-Chromosome	SNP_matching-Position	SNP_matching-Allele	SNP_matching-Strand
rs1043515	ENSG00000141720	PIP4K2B	Exon	17	36922196	A/G	+
rs1046934	ENSG00000198860	TSEN15	Exon	1	184023529	A/C	+
rs1046943	ENSG00000112365	ZBTB24	Exon	6	109783941	A/G	+
rs1351394	ENSG00000149948	HMGA2	Exon	12	66351826	T/C	+
rs143384	ENSG00000125965	GDF5	Exon	20	34025756	A/G	+
rs16942341	ENSG00000157766	ACAN	Exon	15	89388905	C/T	+
rs17318596	ENSG00000248098	BCKDHA	Exon	19	41937095	G/A	+
rs1741344	ENSG00000088826	SMOX	Exon	20	4101800	C/T	+
rs2066807	ENSG00000170581	STAT2	Exon	12	56740682	C/G	+
rs2247341	ENSG00000163950	SLBP	Exon	4	1701317	G/A	+
rs422421	ENSG00000160867	FGFR4	Exon	5	176517326	T/C	+
rs572169	ENSG00000121853	GHSR	Exon	3	172165727	C/T	+
rs5742915	ENSG00000140464	PML	Exon	15	74336633	T/C	+
rs724016	ENSG00000177311	ZBTB38	Exon	3	141105570	A/G	+
rs7689420	ENSG00000164161	HHIP	Exon	4	145568352	T/C	+
rs806794	ENSG00000197846	HIST1H2BF	Exon	6	26200677	A/G	+
rs9456307	ENSG00000130338	TULP4	Exon	6	158929442	T/A	+
rs9835332	ENSG00000163946	FAM208A	Exon	3	56667682	G/C	+
rs9844666	ENSG00000114054	PCCB	Exon	3	135974216	G/A	+

This table contains general information about SNPs that are located in exons. Each row in this table corresponds to one SNP. The columns **Ensembl ID** and **Gene symbol** refer to the gene in which this particular SNP is located. The column **Location** confirms that all SNPs are located in exons. The absolute genomic positions of the SNPs are shown in the columns **SNP_matching-Chromosome,** **SNP_matching-Position** and **SNP_matching-Strand**. The column **SNP_matching-Allele** shows which nucleotide exactly varies at the listed position.

If your input SNP table contains more columns in addition to IDs, all these columns will be preserved and will be added to the right side of this table.

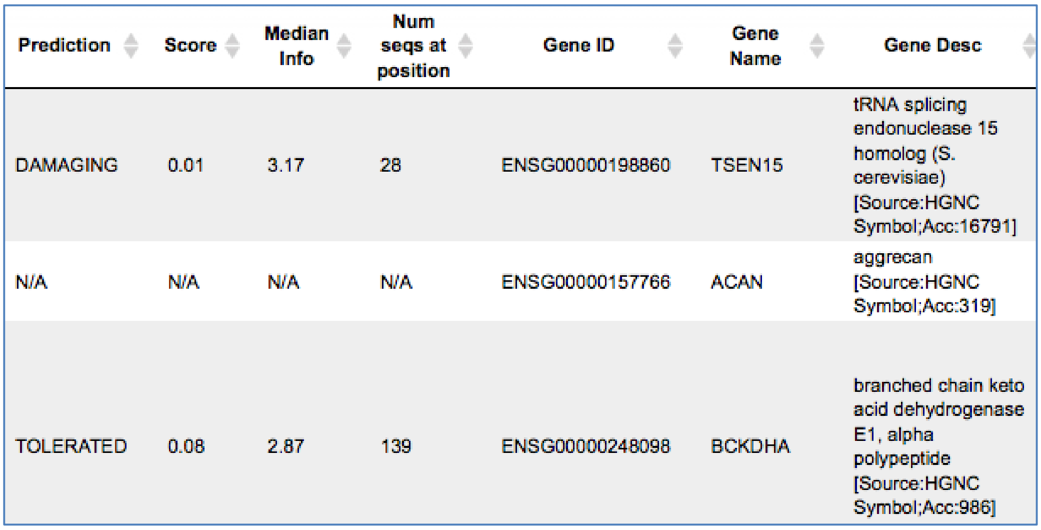
The other table in this subfolder results from the *SIFT analysis,* and is represented by an icon for a general table (▤). SIFT is a widely accepted method to check whether a particular variation is synonymous or non-synonymous, and in case of a non-synonymous variation whether it is damaging or tolerated. More details about SIFT can be found under http://sift.jcvi.org/www/SIFT_help.html.

This table also has 19 rows according to the number of SNPs identified in exons. There are many columns in this table, we will consider the most important ones in the following.



Allele	Codons	Transcript ID	Protein ID	Substitution	Region	dbSNP ID	SNP Type	Prediction
A/C	CAA-CAc	ENST00000361641	ENSP00000355299	Q59H	EXON CDS	rs1046934:C	Nonsynonymous	DAMAGING
C/T	GTC-GTt	ENST00000268134	ENSP00000268134	V407V	EXON CDS	rs16942341:T	Synonymous	N/A
G/A	GCG-aCG	ENST00000378196	ENSP00000367438	A353T	EXON CDS	rs17318596:A	Nonsynonymous	TOLERATED

The columns **Codons** and **Substitution** show which nucleotide in a codon varies and which amino acid is substituted by which. The column **SNP Type** shows if it is a synonymous or a non-synonymous variation, and in case of non-synonymous variations the column **Prediction** shows if it is damaging or tolerated. An extension of this table to its right side is shown below, starting with the column **Prediction**:

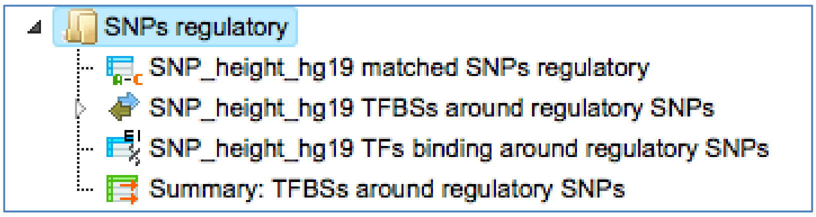


Prediction	Score	Median Info	Num seqs at position	Gene ID	Gene Name	Gene Desc
DAMAGING	0.01	3.17	28	ENSG00000198860	TSEN15	tRNA splicing endonuclease 15 homolog (S. cerevisiae) [Source:HGNC Symbol;Acc:16791]
N/A	N/A	N/A	N/A	ENSG00000157766	ACAN	aggrecan [Source:HGNC Symbol;Acc:319]
TOLERATED	0.08	2.87	139	ENSG00000248098	BCKDHA	branched chain keto acid dehydrogenase E1, alpha polypeptide [Source:HGNC Symbol;Acc:986]

The columns **Gene ID**, **Gene Name**, **Gene Desc** show information about which genes and gene products are affected and might be even damaged by a given variation.

Subfolder SNPs regulatory

This subfolder contains three tables and one track.

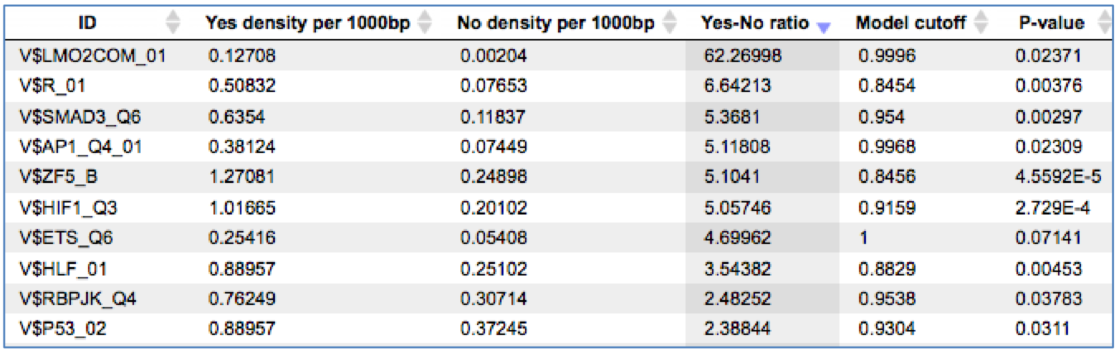


SNPs regulatory
- SNP_height_hg19 matched SNPs regulatory
- SNP_height_hg19 TFBSs around regulatory SNPs
- SNP_height_hg19 TFs binding around regulatory SNPs
- Summary: TFBSs around regulatory SNPs

In this example, 129 out of 148 SNPs mapped to the genome are located in introns or gene flanking regions. The table ( ) contains standard SNP IDs in the **ID** column, and has the same icon as the input SNP table, and as the table with SNPs in exons. The structure of the latter was described above in detail, under the subheading *Subfolder SNPs in exons*.

The other two tables and one track in this subfolder present the results of the TFBS search in the SNP surrounding regions.

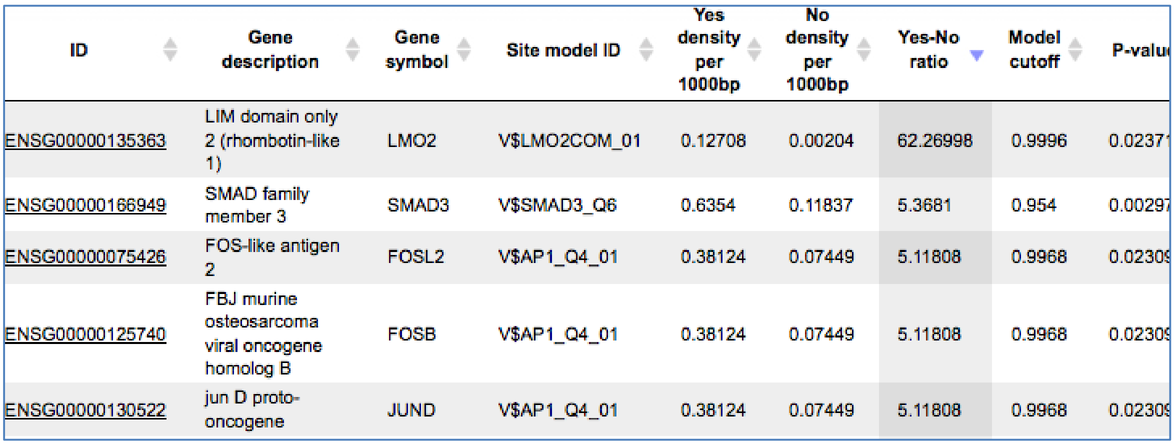
The table *Summary: TFBSs around regulatory SNPs* ( ) shown below has been sorted by the values in the **Yes-No ratio** column.



ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
V$LMO2COM_01	0.12708	0.00204	62.26998	0.9996	0.02371
V$R_01	0.50832	0.07653	6.64213	0.8454	0.00376
V$SMAD3_Q6	0.6354	0.11837	5.3681	0.954	0.00297
V$AP1_Q4_01	0.38124	0.07449	5.11808	0.9968	0.02309
V$ZF5_B	1.27081	0.24898	5.1041	0.8456	4.5592E-5
V$HIF1_Q3	1.01665	0.20102	5.05746	0.9159	2.729E-4
V$ETS_Q6	0.25416	0.05408	4.69962	1	0.07141
V$HLF_01	0.88957	0.25102	3.54382	0.8829	0.00453
V$RBPJK_Q4	0.76249	0.30714	2.48252	0.9538	0.03783
V$P53_02	0.88957	0.37245	2.38844	0.9304	0.0311

Each row summarizes the information for one PWM. The columns **Yes density per 1000bp** and **No density per 1000bp** show the number of matches normalized per 1000 bp length for the sequences around SNPs and in the sequences around random genomic positions, respectively. The column **Yes-No ratio** is the ratio of the first two columns. Only matrices with a Yes-No ratio higher than 1 are included in the *Summary* table. The higher the Yes-No ratio, the higher the enrichment of matches is for the respective matrix in the sequences around regulatory SNPs. The matrix cutoff values calculated by the program at the optimization step are shown in the column **Model cutoff**, and the last column shows the **P-value** of the corresponding event.

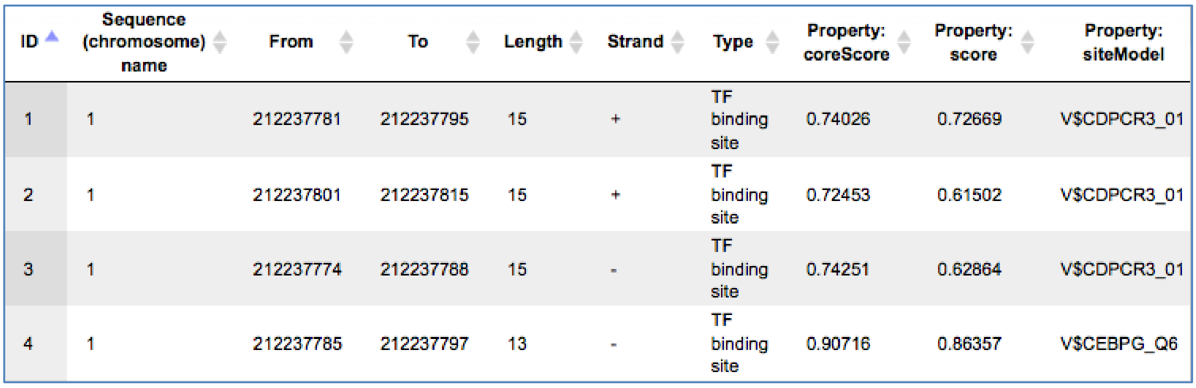
The table *TFs binding around regulatory SNPs* ( ) includes transcription factors (TFs) that are associated with the PWMs listed in the table above, and each row shows details for one TF, including its Ensembl gene ID (column **ID**), gene symbol, gene description of the corresponding TF (columns **Gene description**, **Gene symbol**). The column **Site model ID** shows the identifier of the PWM associated with this TF, and several further columns repeat information that is also shown in the table above.



ID	Gene description	Gene symbol	Site model ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-valu
ENSG00000135363	LIM domain only 2 (rhombotin-like 1)	LMO2	V$LMO2COM_01	0.12708	0.00204	62.26998	0.9996	0.02371
ENSG00000166949	SMAD family member 3	SMAD3	V$SMAD3_Q6	0.6354	0.11837	5.3681	0.954	0.00297
ENSG00000075426	FOS-like antigen 2	FOSL2	V$AP1_Q4_01	0.38124	0.07449	5.11808	0.9968	0.02309
ENSG00000125740	FBJ murine osteosarcoma viral oncogene homolog B	FOSB	V$AP1_Q4_01	0.38124	0.07449	5.11808	0.9968	0.02309
ENSG00000130522	jun D proto-oncogene	JUND	V$AP1_Q4_01	0.38124	0.07449	5.11808	0.9968	0.02309

These TFs are suggested to have their binding sites in close proximity or even overlapping with SNPs, and their binding might be affected by a given SNP.

The track *TFBSs around regulatory SNPs* ( ) gives information about the genomic positions of the identified TFBSs.

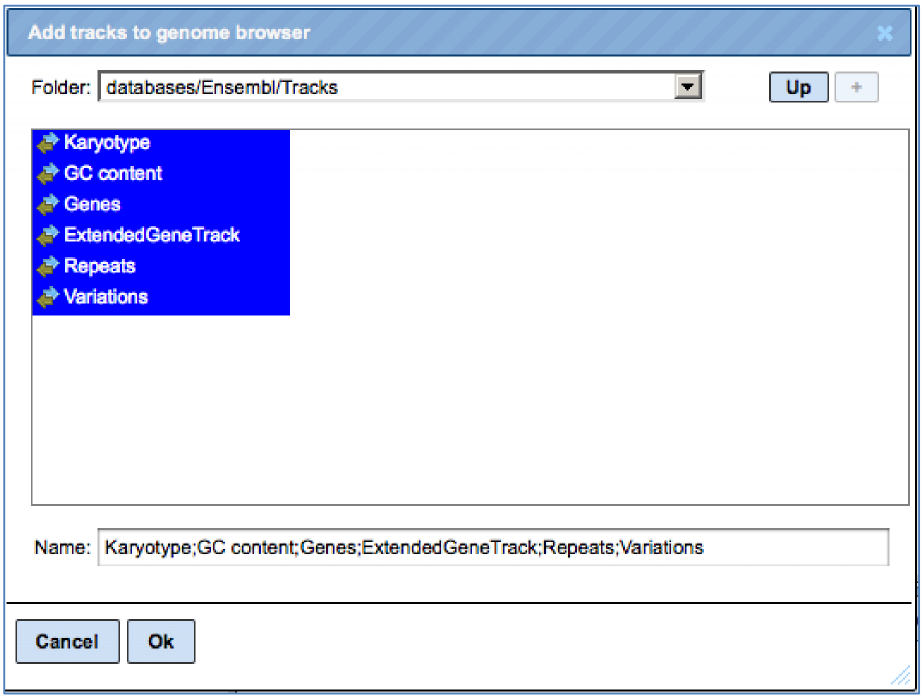


ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: coreScore	Property: score	Property: siteModel
1	1	212237781	212237795	15	+	TF binding site	0.74026	0.72669	V$CDPCR3_01
2	1	212237801	212237815	15	+	TF binding site	0.72453	0.61502	V$CDPCR3_01
3	1	212237774	212237788	15	-	TF binding site	0.74251	0.62864	V$CDPCR3_01
4	1	212237785	212237797	13	-	TF binding site	0.90716	0.86357	V$CEBPG_Q6

Each row presents details for an individual TFBS. The columns **Sequence (chromosome) name**, **From**, **To**, **Length** and **Strand** show the genomic location of the match including chromosome number, start and end positions, strand, and length of the match. The column **Type** contains information about the type of the elements; in this case all matches are assigned the type *TF binding site*. Further columns keep information about the site model producing each match (column **Property:siteModel**) as well as a score of the core (column **Property:coreScore**), and a score for the whole site model (column **Property:score**).

Tip. Further visualization of tracks in the genome browser:

When having tracks opened in the Work Space, the menu button ⇄ on the top panel can be applied to visualize it. A supplementary window *Add tracks to genome browser* will open. Here you can select tracks that can be visualized together with your track (shown below).

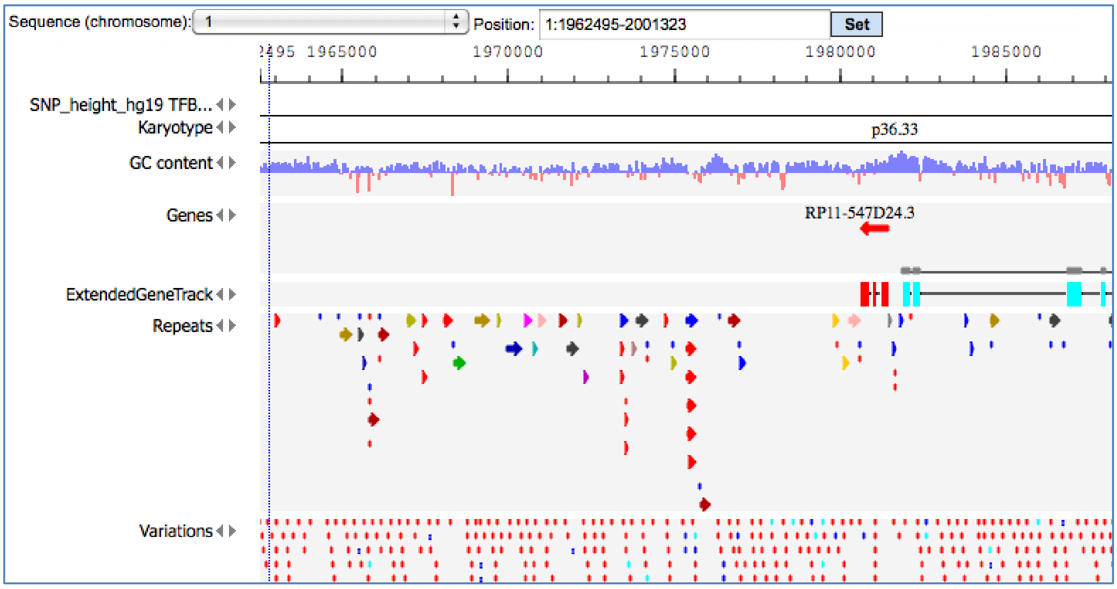


Folder: databases/Ensembl/Tracks

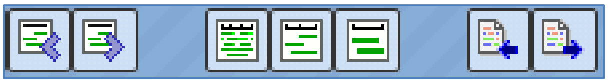
- Karyotype
- GC content
- Genes
- ExtendedGeneTrack
- Repeats
- Variations

Name: Karyotype;GC content;Genes;ExtendedGeneTrack;Repeats;Variations

After pressing [Ok] you will get the picture shown below with your track in focus at the top position.

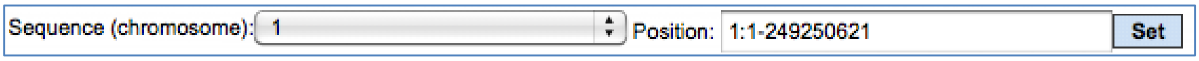


A number of options are available to navigate the browser and get the desired view. You can use several buttons at the top panel to zoom in and out, and to shift the visible part of the map left or right.



With the help of the small triangles next to the track names you can jump to the next or previous element of this track.

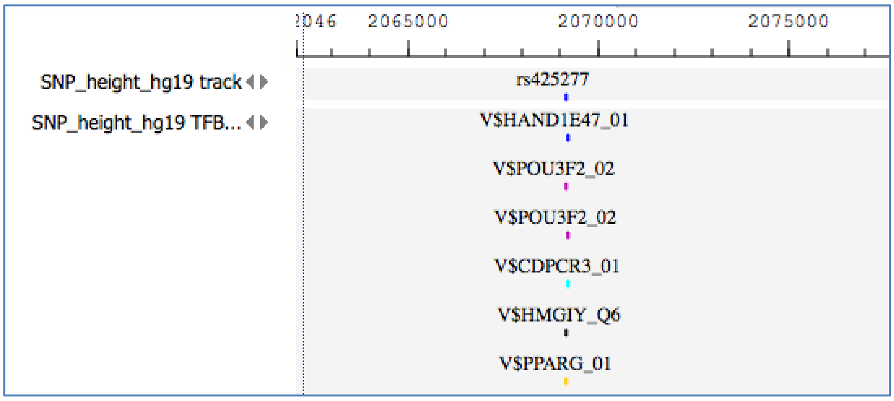
With the drop-down menu shown below, you can jump between different chromosomes, or specify exact positions in the *Position* window.



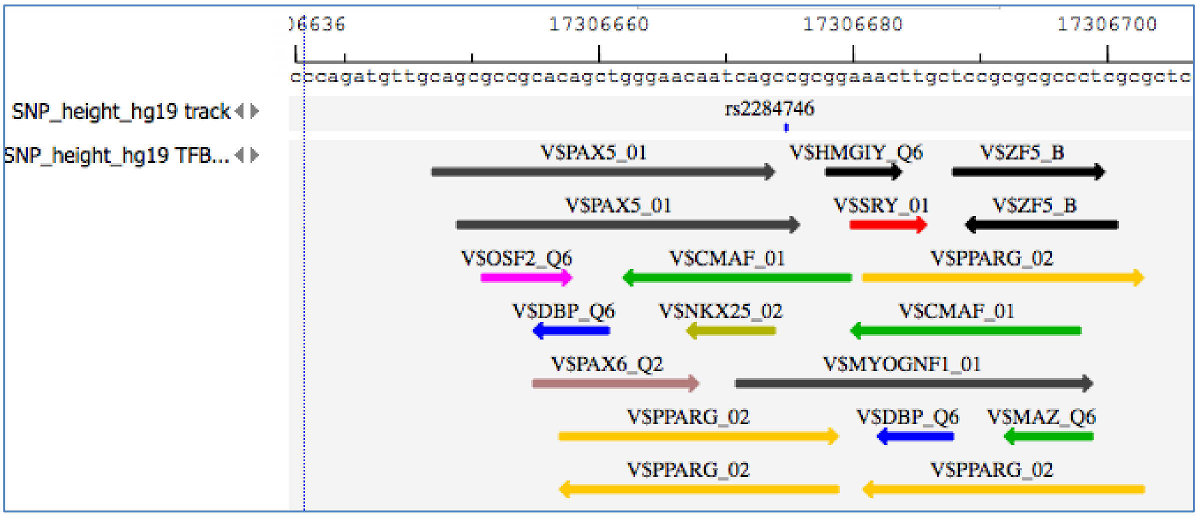
Sequence (chromosome): 1 Position: 1:1-249250621 Set

When having a track opened in the genome browser, you can drag & drop any other track over the same picture to add it to the browser, where you can find it below the bottom-most track. You can drag & drop it then to any track into desired position.

For example, you can add the track with all input SNPs from the folder *All SNPs*, and shift it to the top position, and then jump to the position of the 1st SNP along the chromosome. The resulting picture is shown below.



Next, you can zoom in down to the nucleotide level, and get the following picture, where one of the identified regulatory SNPs, rs2284746, is overlapping with several TFBSs, e.g. with the binding sites for c-Maf and PPARγ.



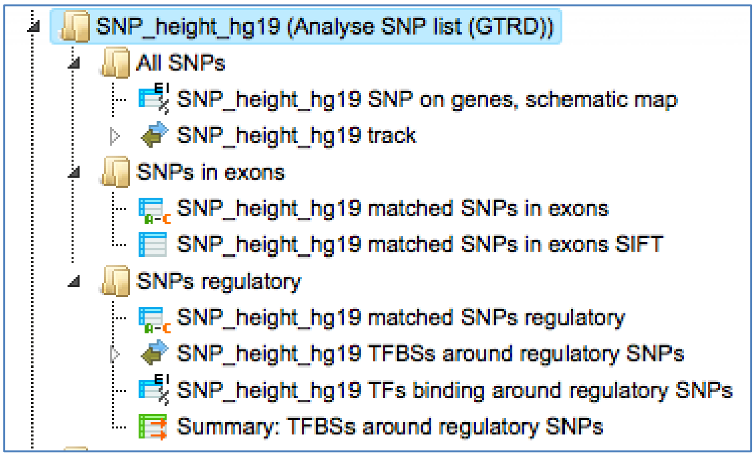
Note. This workflow is available together with a valid TRANSFAC® license.
Please feel free to ask for details (info@genexplain.com).

Analysis with GTRD

This workflow is similar to the one described above. The difference is in the database applied for the TFBS search; in this workflow it is the GTRD database. Correspondingly, the results for regulatory SNPs overlapping with TFBSs might be different.

The results of this workflow can be found under:

data/Examples/SNPs linked to human height/Data/SNP_height_hg19 (Analyze SNP list (GTRD))



SNP_height_hg19 (Analyse SNP list (GTRD))
- All SNPs
  - SNP_height_hg19 SNP on genes, schematic map
  - SNP_height_hg19 track
- SNPs in exons
  - SNP_height_hg19 matched SNPs in exons
  - SNP_height_hg19 matched SNPs in exons SIFT
- SNPs regulatory
  - SNP_height_hg19 matched SNPs regulatory
  - SNP_height_hg19 TFBSs around regulatory SNPs
  - SNP_height_hg19 TFs binding around regulatory SNPs
  - Summary: TFBSs around regulatory SNPs

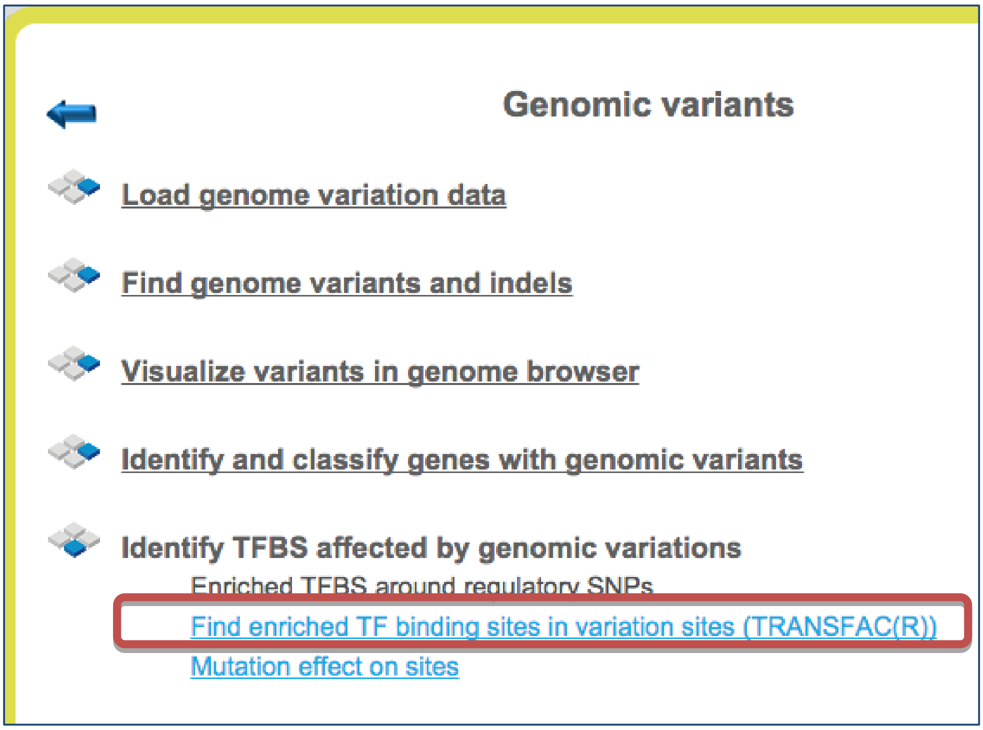
14.4.2. Find enriched TF binding sites in variation sites

This workflow is designed to study variations (mutations) located especially within promoter regions. It helps to address the questions, which TFBSs are enriched around the variations, and which TFs are responsible for the regulation of the corresponding promoters.

As input, you submit two tracks, one track in vcf format with variations (mutations), and another track with the promoters in focus; by default all promoters from Ensembl 65.37 are used. You also need to specify a profile (a set of PWMs), and a window (region) around each variation point, e.g. 10-20 bp, where TFBS analysis is to be performed.

As output you get a summary table with enriched TFBSs, a table with transcription factors as well as a track with the enriched TFBSs, which can be used for visualization in the genome browser.

The workflow can be found on the Start page, in the section “Genomic variants”.



Genomic variants

Load genome variation data

Find genome variants and indels

Visualize variants in genome browser

Identify and classify genes with genomic variants

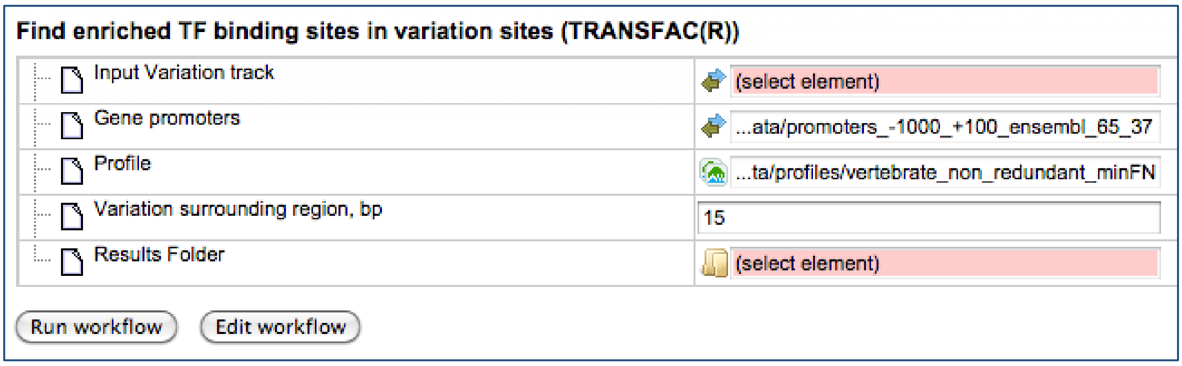
Identify TFBS affected by genomic variations

Enriched TFBS around regulatory SNPs

Find enriched TF binding sites in variation sites (TRANSFAC(R))

Mutation effect on sites

Step 1. Open the workflow input form from the Start page. It will open in the main Work Space and looks as shown below:



Find enriched TF binding sites in variation sites (TRANSFAC(R))

Input Variation track	(select element)
Gene promoters	...ata/promoters_-1000_+100_ensembl_65_37
Profile	...ta/profiles/vertebrate_non_redundant_minFN
Variation surrounding region, bp	15
Results Folder	(select element)

Step 2. Specify the **Input Variation track** in vcf format. To specify the track, you can drag & drop it from your project within the tree area. Alternatively, you may click on the pink field “select element” and a new window will open, where you can select the input track. After having selected the track, press the [Ok] button. You can use either a newly imported track in vcf format, or a track that has been calculated within the platform, e.g. as a result of the workflow *Find genome variants and indels from RNA-seq* or the workflow *Find genome variants and indels from full-genome NGS*. Both workflows can be found under http://genexplain-platform.com/bioumlweb/#de=analyses/Workflows/Common/

In the following example we took as input the track SNP_indels.vcf, which can be found at: data/Examples/Chronic Myeloid Leukemia Patient Genotyping/Data/SNP_indels.vcf

This vcf file was produced by the workflow *Find genome variants and indels from full-genome NGS*.

Step 3. Specify the **Gene promoters**. By default all promoters -1000 to +100 bp relative to the TSS from Ensembl 65.37 genome version are used.

Step 4. Select the **Profile**. This profile will be applied for the identification of the enriched motifs around variation sites. The default profile is *vertebrate_non_redundant_minSUM* from the most recent TRANSFAC® release available.

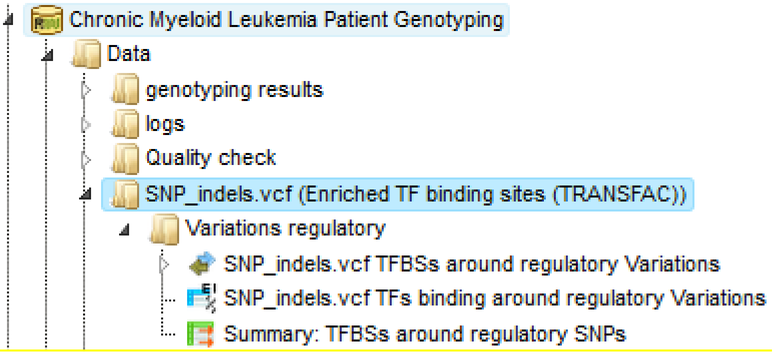
Step 5. Specify the **Variation surrounding region** in base pairs. By default 15 bp are used. Within these region/window the search for enriched TFBSs will be performed.

Step 6. Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink field “select element” in the field **Results Folder**, and a new window will be opened, where you can select the location of the results folder and define its name.

Start the workflow by pressing the [Run workflow] button.

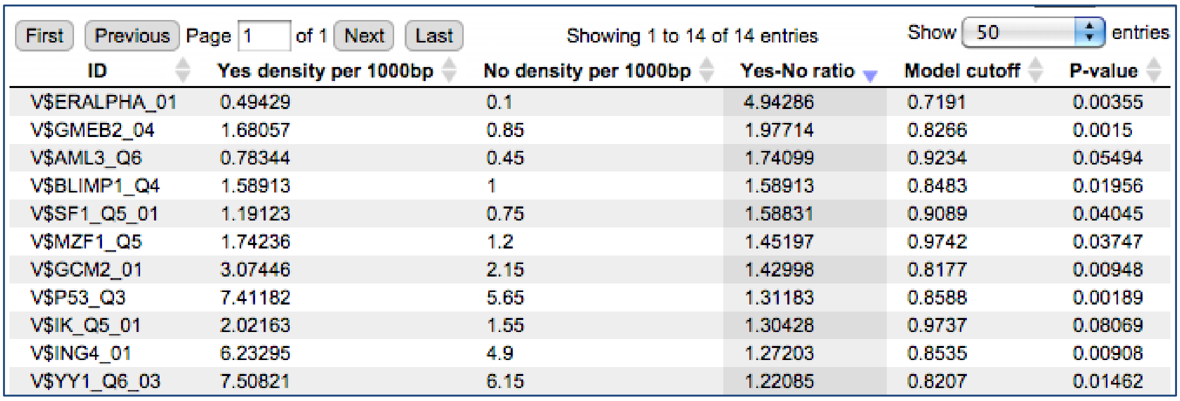
Below you can see the result folder for the example: data/Examples/Chronic Myeloid Leukemia Patient Genotyping/Data/SNP_indels.vcf (Enriched TF binding sites (TRANSFAC))/

The output folder contains on sub-folder with a track and two tables.



- Chronic Myeloid Leukemia Patient Genotyping
  - Data
    - genotyping results
    - logs
    - Quality check
    - SNP_indels.vcf (Enriched TF binding sites (TRANSFAC))
      - Variations regulatory
        - SNP_indels.vcf TFBSs around regulatory Variations
        - SNP_indels.vcf TFs binding around regulatory Variations
        - Summary: TFBSs around regulatory SNPs

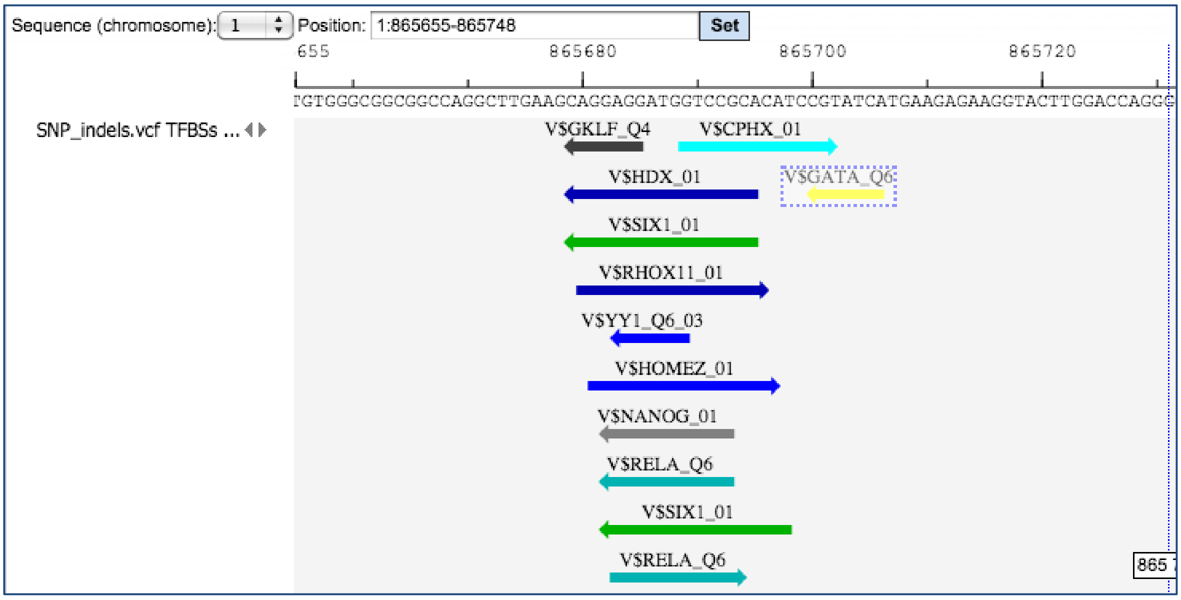
The Summary table: This table contains a list of site models (PWMs) over-represented around variation (mutation) sites in promoters, in this case 14 site models. By default this table is sorted by the Yes-No ratio, with the most over-represented model on top.



First
Previous
Page 1 of 1
Next
Last
Showing 1 to 14 of 14 entries
Show 50 entries

ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
V$ERALPHA_01	0.49429	0.1	4.94286	0.7191	0.00355
V$GMEB2_04	1.68057	0.85	1.97714	0.8266	0.0015
V$AML3_Q6	0.78344	0.45	1.74099	0.9234	0.05494
V$BLIMP1_Q4	1.58913	1	1.58913	0.8483	0.01956
V$SF1_Q5_01	1.19123	0.75	1.58831	0.9089	0.04045
V$MZF1_Q5	1.74236	1.2	1.45197	0.9742	0.03747
V$GCM2_01	3.07446	2.15	1.42998	0.8177	0.00948
V$P53_Q3	7.41182	5.65	1.31183	0.8588	0.00189
V$IK_Q5_01	2.02163	1.55	1.30428	0.9737	0.08069
V$ING4_01	6.23295	4.9	1.27203	0.8535	0.00908
V$YY1_Q6_03	7.50821	6.15	1.22085	0.8207	0.01462

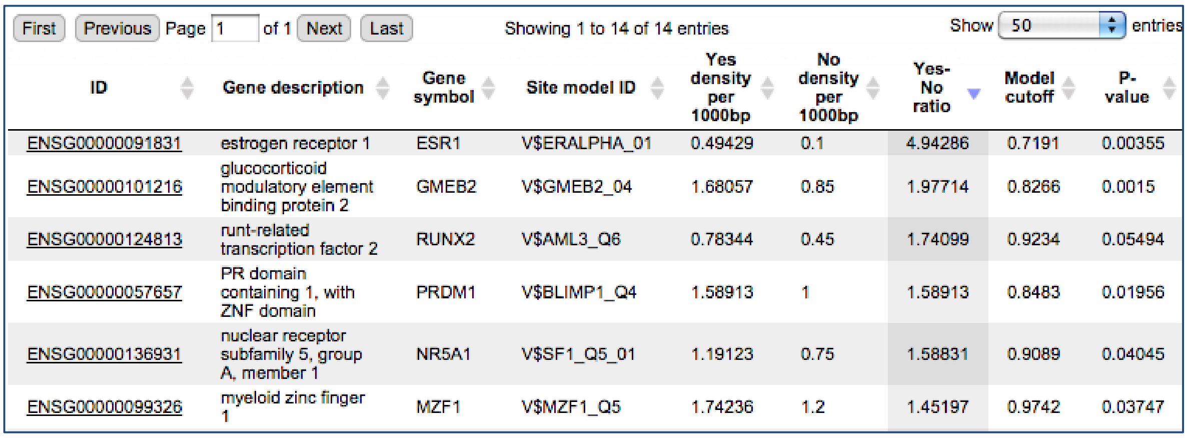
For a visualization of the over-represented TF binding sites in the variation sites you can open the *SNP_indels.vcf TFBS around regulatory Variations* track in the genome browser.



Sequence (chromosome): 1 Position: 1:865655-865748 Set

The table *SNP_indels.vcf TFBS around regulatory Variations* presents the list of TFs corresponding to the over-represented site models, in this case 14 TFs. This table contains the **Site model IDs** of over-represented binding sites, as well as the **Gene**

description and **Gene symbol** for each transcription factor. The results can be sorted by **Yes-No ratio**, as it is shown on the screenshot below.



Showing 1 to 14 of 14 entries

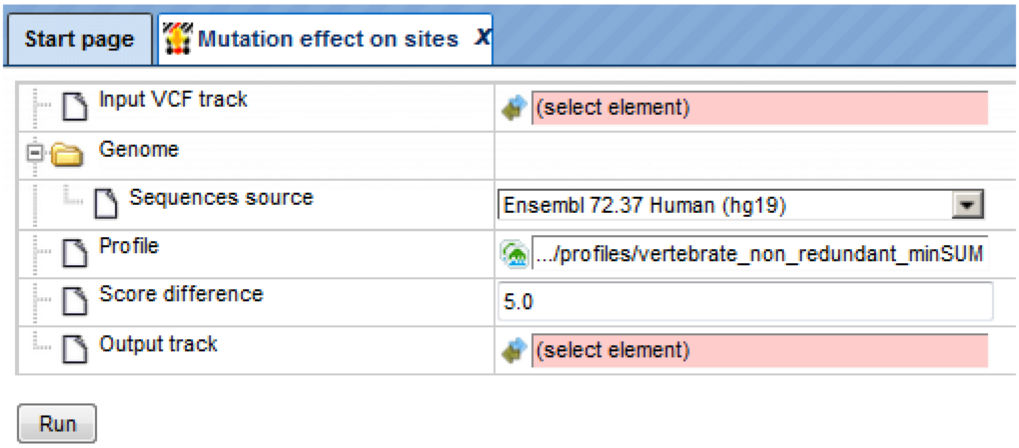
ID	Gene description	Gene symbol	Site model ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
ENSG00000091831	estrogen receptor 1	ESR1	V$ERALPHA_01	0.49429	0.1	4.94286	0.7191	0.00355
ENSG00000101216	glucocorticoid modulatory element binding protein 2	GMEB2	V$GMEB2_04	1.68057	0.85	1.97714	0.8266	0.0015
ENSG00000124813	runt-related transcription factor 2	RUNX2	V$AML3_Q6	0.78344	0.45	1.74099	0.9234	0.05494
ENSG00000057657	PR domain containing 1, with ZNF domain	PRDM1	V$BLIMP1_Q4	1.58913	1	1.58913	0.8483	0.01956
ENSG00000136931	nuclear receptor subfamily 5, group A, member 1	NR5A1	V$SF1_Q5_01	1.19123	0.75	1.58831	0.9089	0.04045
ENSG00000099326	myeloid zinc finger 1	MZF1	V$MZF1_Q5	1.74236	1.2	1.45197	0.9742	0.03747

14.4.3. Mutation effect on sites analysis

This method allows to find transcription factor binding sites (TFBSs) affected by variations or mutations.

The analysis "Mutation effect on sites" can be found in the NGS folder of the analysis methods (analyses/Methods/NGS/Mutation effect on sites) or under the start page button 'Genomic variants' under section 'Identify TFBS affected by genomic variations'.

Step 1. Open the analysis form from the Start page. It will open in the main Work Space and looks as shown below:



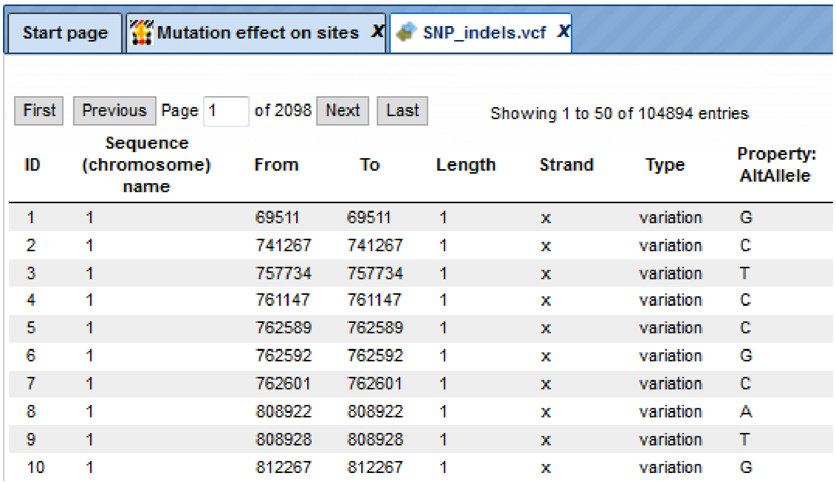
Input VCF track	(select element)
Genome	
Sequences source	Ensembl 72.37 Human (hg19)
Profile	.../profiles/vertebrate_non_redundant_minSUM
Score difference	5.0
Output track	(select element)

Step 2. The Input **VCF track** is a track file with mutations and should be in vcf format.

One input example is here on the platform:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Chronic%20Myeloid%20Leukemia%20Patient%20Genotyping/Data/SNP_indels.vcf

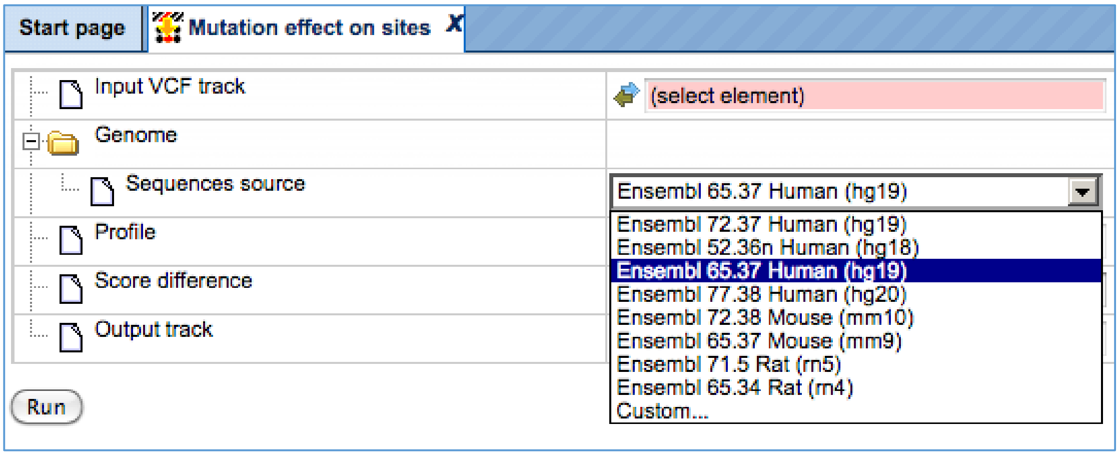
Open the track file as a table, and for each variation point you can see several columns with genomic position, chromosome, alternative nucleotide, etc., as shown below.



Showing 1 to 50 of 104894 entries

ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: AltAllele
1	1	69511	69511	1	x	variation	G
2	1	741267	741267	1	x	variation	C
3	1	757734	757734	1	x	variation	T
4	1	761147	761147	1	x	variation	C
5	1	762589	762589	1	x	variation	C
6	1	762592	762592	1	x	variation	G
7	1	762601	762601	1	x	variation	C
8	1	808922	808922	1	x	variation	A
9	1	808928	808928	1	x	variation	T
10	1	812267	812267	1	x	variation	G

Step 3. Verify the **Sequences source** and use the drop-down menu for different Ensembl genome annotations of human, mouse and rat, as shown below.

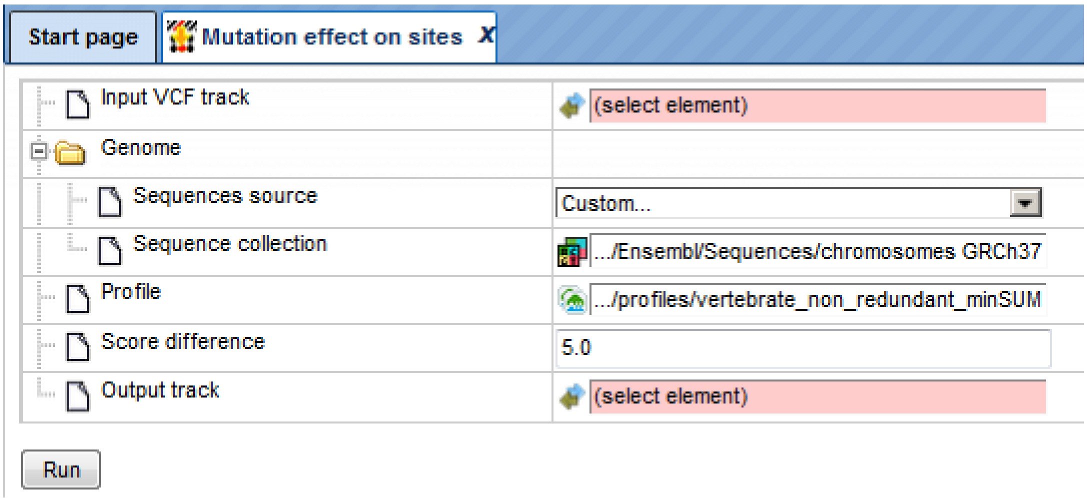


Start page
Mutation effect on sites
Run

Input VCF track	(select element)
Genome	
Sequences source	Ensembl 65.37 Human (hg19)
Profile	
Score difference	
Output track	

Ensembl 72.37 Human (hg19)
Ensembl 52.36n Human (hg18)
Ensembl 65.37 Human (hg19)
Ensembl 77.38 Human (hg20)
Ensembl 72.38 Mouse (mm10)
Ensembl 65.37 Mouse (mm9)
Ensembl 71.5 Rat (rn5)
Ensembl 65.34 Rat (rn4)
Custom...

Alternatively, you can choose 'Custom' from the same menu, if you would like specify another genome, e.g. a particular patient genome imported into the platform before. As soon as the option 'Custom' is chosen, an additional field, Sequence collection, automatically appears on the input form (screenshot below), and you can specify the sequences location manually.



Input VCF track	(select element)
Genome	
Sequences source	Custom...
Sequence collection	.../Ensembl/Sequences/chromosomes GRCh37
Profile	.../profiles/vertebrate_non_redundant_minSUM
Score difference	5.0
Output track	(select element)

Step 4. Select the **Profile**. This profile will be applied for the identification of transcription factor binding sites overlapping with the variation positions. The default profile is *vertebrate_non_redundant_minSUM* from the most recent TRANSFAC® release available.

Step 5. The **Score difference** from unaffected to affected site is per default 5.0. This parameter is a threshold for the difference between the TFBS score in the reference genome and the TFBS score at the same position with a variation in the alternative sequence. For TRANSFAC matrices adjust this parameter between 0.1 and 0.5.

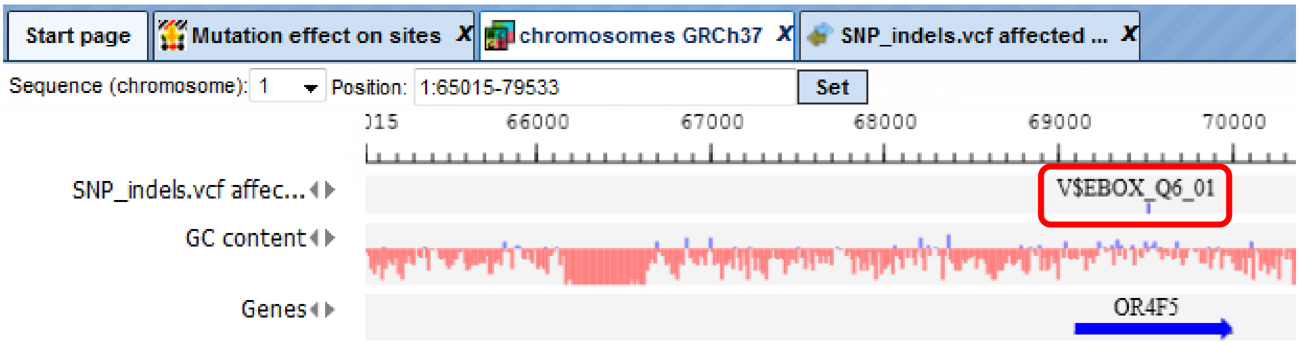
All TFBSs with score differences above this specified value will be reported in the output track.

The lower this value, the more TFBS will be reported as the result, because even a small change in the score will be considered. If you are interested in those TFBSs that are more strongly affected by a variation, set this parameter to 0.5 (see also example output with score difference = 0.5).

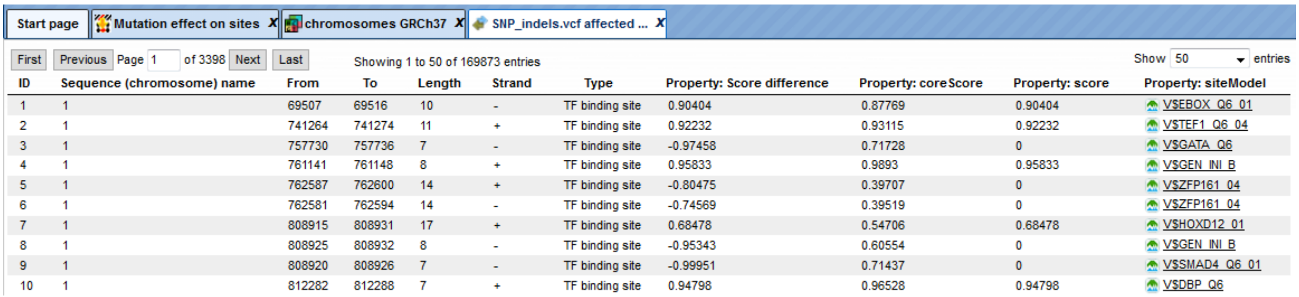
Step 5. Specify the path and name of the **Output track**.

After completion the output track file (SNP_indels.vcf affected sites) is opened by default in the work space. One example of the affected identified site is shown in the red box (V$EBOX_Q6_01).

This resulting track can be found in the Examples folder under the URL: http://genexplain-platform.com/bioumlweb/#de=data/Examples/Chronic%20Myeloid%20Leukemia%20Patient%20Genotyping/Data/Affected%20binding%20sites



Opening the track as a table shows all affected TFBSs in table format.



Showing 1 to 50 of 169873 entries

ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: Score difference	Property: coreScore	Property: score	Property: siteModel
1	1	69507	69516	10	-	TF binding site	0.90404	0.87769	0.90404	V$EBOX_Q6_01
2	1	741264	741274	11	+	TF binding site	0.92232	0.93115	0.92232	V$TEF1_Q6_04
3	1	757730	757736	7	-	TF binding site	-0.97458	0.71728	0	V$GATA_Q6
4	1	761141	761148	8	+	TF binding site	0.95833	0.9893	0.95833	V$GEN_INI_B
5	1	762587	762600	14	+	TF binding site	-0.80475	0.39707	0	V$ZFP161_04
6	1	762581	762594	14	-	TF binding site	-0.74569	0.39519	0	V$ZFP161_04
7	1	808915	808931	17	+	TF binding site	0.68478	0.54706	0.68478	V$HOXD12_01
8	1	808925	808932	8	-	TF binding site	-0.95343	0.60554	0	V$GEN_INI_B
9	1	808920	808926	7	-	TF binding site	-0.99951	0.71437	0	V$SMAD4_Q6_01
10	1	812282	812288	7	+	TF binding site	0.94798	0.96528	0.94798	V$DBP_Q6

The upper example highlighted by the red box has ID=1 in the table. The columns **From** and **To** define the positions of the affected site within the genome on chromosome 1 (**Sequence (chromosome) name**). The column **Length** shows the length of the binding motif, here 10. The **Type** TF binding site shows that a transcription factor binding site is affected with a score difference of 0.90404.

The column **Property: Score difference** shows the arithmetical difference between TFBS score in the reference genome and the TFBS score at the same position with a variation (in the alternative sequence). The score difference can be positive or negative. A positive score indicates a disrupted site and a negative score predicts a new site (site appears).

A positive score difference means that the given TFBS had a better score in the reference sequence, and it was decreased by the variation. In the other words, the given variation disrupted a TFBS which occurred in the reference sequence.

A negative score means that the given TFBS has a better score in the sequence with the variation as compared to the reference sequence. A given TFBS became stronger or even appears after the variation. The conclusion can be made that a given variation created or enhanced corresponding TFBS.

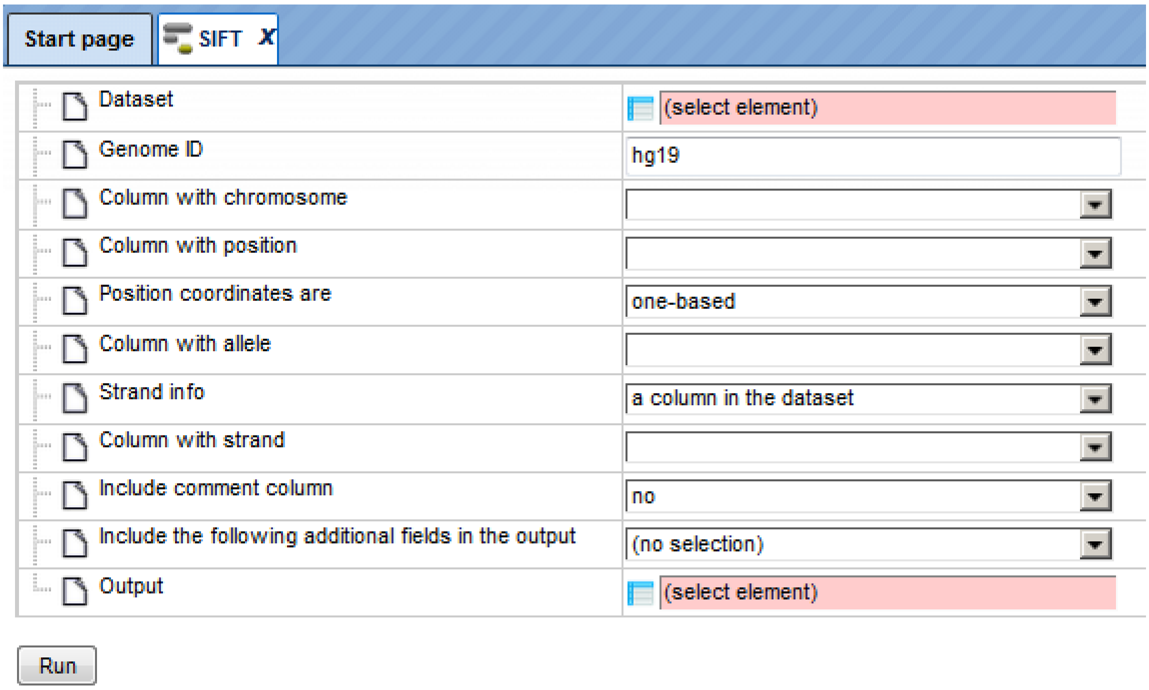
The last column **Property: siteModel** gives a link to the matrix model and can be opened in the workspace to view the matrix logo.

14.4.4. SIFT (Sorting Tolerant From Intolerant) analysis

The SIFT analysis tool predicts whether a single amino acid substitution (AAS) affects protein function, based on sequence homology and the physical properties of amino acids. SIFT can be applied to naturally occurring non-synonymous single nucleotide polymorphisms (nsSNP) and laboratory-induced missense mutations. This tool uses a SQLite databases containing pre-computed SIFT scores and annotations for all possible nucleotide substitutions at each position in the human exome. Allele frequency data are from the HapMap frequency database, and additional transcript and gene-level data are from Ensembl BioMart. The updated version of SIFT is published in Nat Protoc. 4:1073-1081, 2009.

The tool can be found in the Galaxy section of the geneXplain platform (analyses/Galaxy/Human Genome Variation/SIFT) or on the start page button 'Genomic variants' under section 'Identify TFBS affected by genomic variations'.

Step 1. Open the analysis form from the Start page. It will open in the main Work Space and looks as shown below:



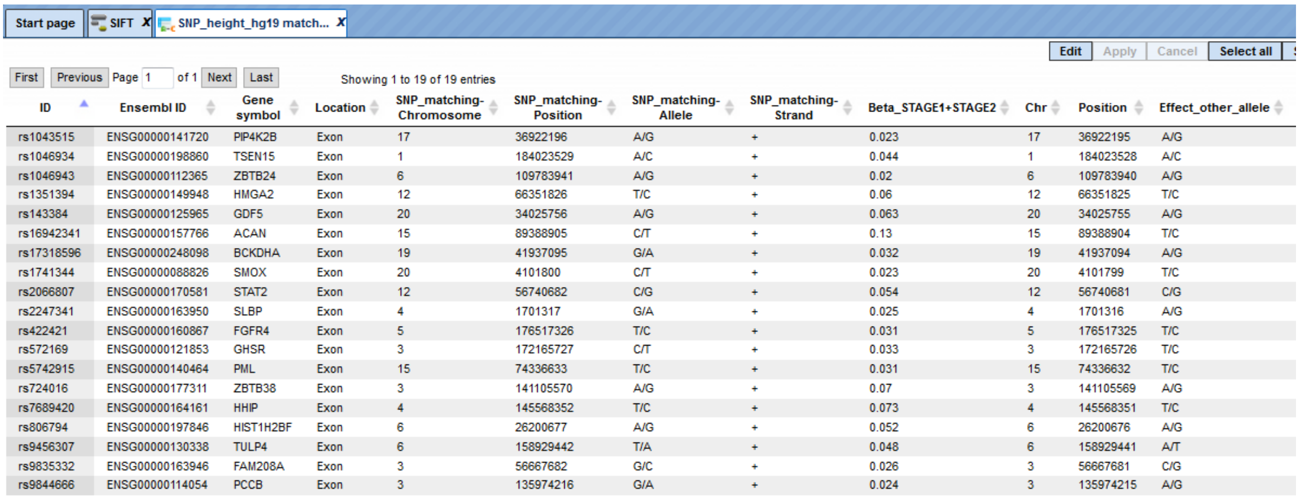
Dataset	(select element)
Genome ID	hg19
Column with chromosome	
Column with position	
Position coordinates are	one-based
Column with allele	
Strand info	a column in the dataset
Column with strand	
Include comment column	no
Include the following additional fields in the output	(no selection)
Output	(select element)

Step 2. The input **Dataset** must contain columns for the chromosome, position, and alleles. The alleles must be two nucleotides separated by '/', usually the reference allele and the allele of interest. The strand must either be in another column.

Input example format:

chr3	81780820	+	T/C
chr2	230341630	+	G/A
chr2	43881517	+	A/T
chr2	43857514	+	T/C
chr6	88375602	+	G/A
chr22	29307353	-	T/A
chr10	115912482	-	G/T
chr10	115900918	-	C/T
chr16	69875502	+	G/T

One example input table can be found here on the platform: data/Examples/SNPs linked to human height/Data/SNP height hg19 (Analyse SNP list (TRANSFAC))/SNPs in exons/SNP height hg19 matched SNPs in exons



Showing 1 to 19 of 19 entries

ID	Ensembl ID	Gene symbol	Location	SNP_matching-Chromosome	SNP_matching-Position	SNP_matching-Allele	SNP_matching-Strand	Beta_STAGE1+STAGE2	Chr	Position	Effect_other_allele
rs1043515	ENSG00000141720	PIP4K2B	Exon	17	36922196	A/G	+	0.023	17	36922195	A/G
rs1046934	ENSG00000198860	TSEN15	Exon	1	184023529	A/C	+	0.044	1	184023528	A/C
rs1046943	ENSG00000112365	ZBTB24	Exon	6	109783941	A/G	+	0.02	6	109783940	A/G
rs1351394	ENSG00000149948	HMGA2	Exon	12	66351826	T/C	+	0.06	12	66351825	T/C
rs143384	ENSG00000125965	GDF5	Exon	20	34025756	A/G	+	0.063	20	34025755	A/G
rs16942341	ENSG00000157766	ACAN	Exon	15	89388905	C/T	+	0.13	15	89388904	T/C
rs17318596	ENSG00000248098	BCKDHA	Exon	19	41937095	G/A	+	0.032	19	41937094	A/G
rs1741344	ENSG00000088826	SMOX	Exon	20	4101800	C/T	+	0.023	20	4101799	T/C
rs2066807	ENSG00000170581	STAT2	Exon	12	56740682	C/G	+	0.054	12	56740681	C/G
rs2247341	ENSG00000163950	SLBP	Exon	4	1701317	G/A	+	0.025	4	1701316	A/G
rs422421	ENSG00000160867	FGFR4	Exon	5	176517326	T/C	+	0.031	5	176517325	T/C
rs572169	ENSG00000121853	GHSR	Exon	3	172165727	C/T	+	0.033	3	172165726	T/C
rs5742915	ENSG00000140464	PML	Exon	15	74336633	T/C	+	0.031	15	74336632	T/C
rs724016	ENSG00000177311	ZBTB38	Exon	3	141105570	A/G	+	0.07	3	141105569	A/G
rs7689420	ENSG00000164161	HHIP	Exon	4	145568352	T/C	+	0.073	4	145568351	T/C
rs806794	ENSG00000197846	HIST1H2BF	Exon	6	26200677	A/G	+	0.052	6	26200676	A/G
rs9456307	ENSG00000130338	TULP4	Exon	6	158929442	T/A	+	0.048	6	158929441	A/T
rs9835332	ENSG00000163946	FAM208A	Exon	3	56667682	G/C	+	0.026	3	56667681	C/G
rs9844666	ENSG00000114054	PCCB	Exon	3	135974216	G/A	+	0.024	3	135974215	A/G

Step 3. Determine the **Genome ID**. The tool currently works only for genome builds hg18 or hg19.

Step 4. Define the **Column with chromosome**. In our example above it is column 5 (SNP_matching-Chromosome).

Step 5. Define the **Column with position**. In our example above it is column 6 (SNP_matching-Position).

Step 6. Selection if **Position coordinates are** one-based (default) or zero-based counted.

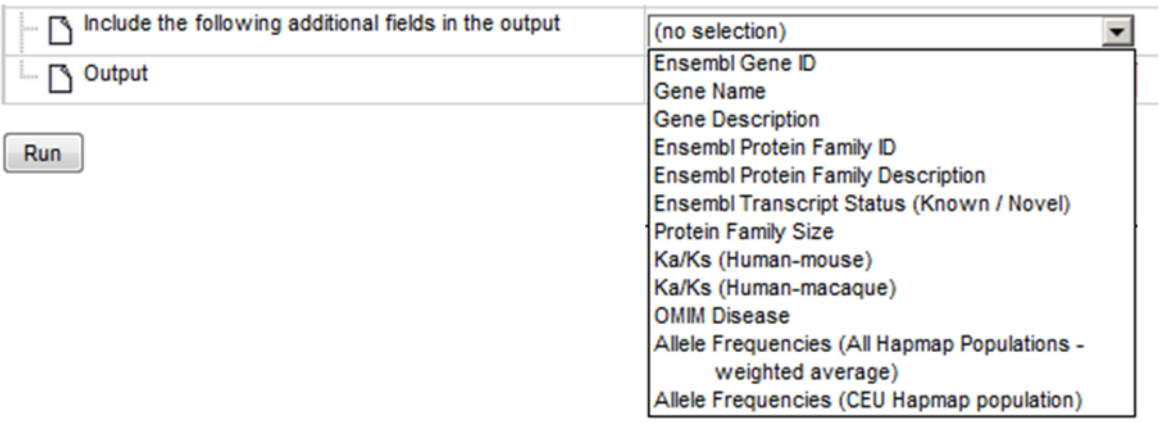
Step 7. Define the **Column with allele**. In our example above it is column 7 (SNP_matching-Allele).

Step 8. Define whether the **Strand info** is a column in the dataset (default).

Step 9. Define the **Column with strand**. In our example above it is column 8 (SNP_matching-Strand).

Step 10. Select **Include comment column** for additional comments.

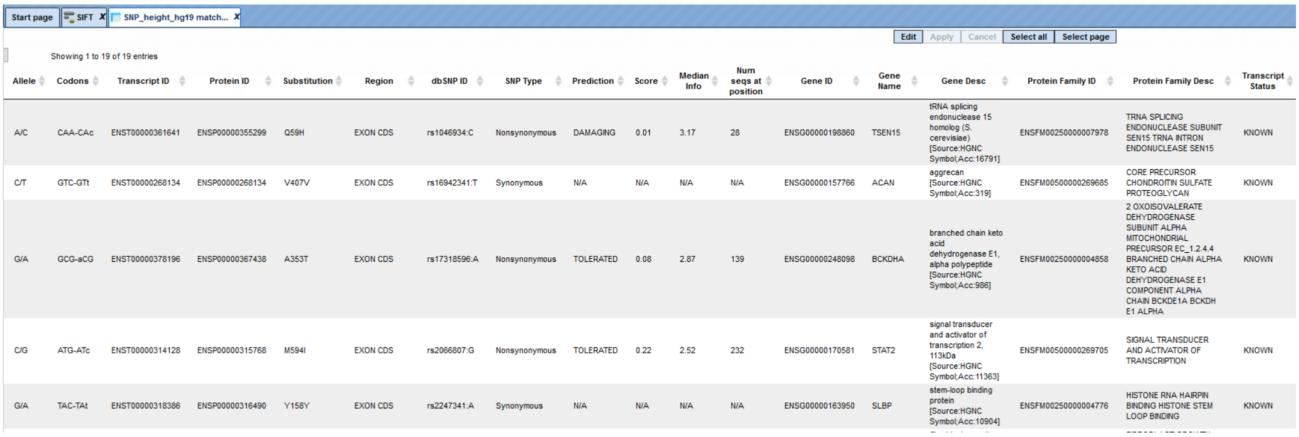
Step 11. Possibility to select multiple output columns and **Include the following additional fields in the output** table.


Include the following additional fields in the output
(no selection)
Output
Run
Ensembl Gene ID
Gene Name
Gene Description
Ensembl Protein Family ID
Ensembl Protein Family Description
Ensembl Transcript Status (Known / Novel)
Protein Family Size
Ka/Ks (Human-mouse)
Ka/Ks (Human-macaque)
OMIM Disease
Allele Frequencies (All Hapmap Populations - weighted average)
Allele Frequencies (CEU Hapmap population)

Step 3. Define where the output table should be located in your project tree. You can do so by clicking on the pink field “select element” in the field **Output**, and a new window will be opened, where you can select the location of the table and define its name.

Start the SIFT analysis by pressing the [Run workflow] button.

An example output table can be found here: data/Examples/SNPs linked to human height/Data/SNP height hg19 (Analyse SNP list (TRANSFAC))/SNPs in exons/SNP height hg19 matched SNPs in exons SIFT



Showing 1 to 19 of 19 entries

Allele	Codons	Transcript ID	Protein ID	Substitution	Region	dbSNP ID	SNP Type	Prediction	Score	Median Info	Num seqs at position	Gene ID	Gene Name	Gene Desc	Protein Family ID	Protein Family Desc	Transcript Status
A/C	CAA-CAc	ENST00000361641	ENSP00000355299	Q59H	EXON CDS	rs1046934:C	Nonsynonymous	DAMAGING	0.01	3.17	28	ENSG00000198860	TSEN15	tRNA splicing endonuclease 15 homolog (S. cerevisiae) [Source:HGNC Symbol;Acc:16791]	ENSFM00250000007978	TRNA SPLICING ENDONUCLEASE SUBUNIT SEN15 TRNA INTRON ENDONUCLEASE SEN15	KNOWN
C/T	GTC-GTt	ENST00000268134	ENSP00000268134	V407V	EXON CDS	rs16942341:T	Synonymous	N/A	N/A	N/A	N/A	ENSG00000157766	ACAN	aggrecan [Source:HGNC Symbol;Acc:319]	ENSFM00500000269685	CORE PRECURSOR CHONDROITIN SULFATE PROTEOGLYCAN	KNOWN
G/A	GCG-aCG	ENST00000378196	ENSP00000367438	A353T	EXON CDS	rs17318596:A	Nonsynonymous	TOLERATED	0.08	2.87	139	ENSG00000248098	BCKDHA	branched chain keto acid dehydrogenase E1, alpha polypeptide [Source:HGNC Symbol;Acc:986]	ENSFM00250000004858	2 OXOISOVALERATE DEHYDROGENASE SUBUNIT ALPHA MITOCHONDRIAL PRECURSOR EC_1.2.4.4 BRANCHED CHAIN ALPHA KETO ACID DEHYDROGENASE E1 COMPONENT ALPHA CHAIN BCKDE1A BCKDH E1 ALPHA	KNOWN
C/G	ATG-ATc	ENST00000314128	ENSP00000315768	M594I	EXON CDS	rs2066807:G	Nonsynonymous	TOLERATED	0.22	2.52	232	ENSG00000170581	STAT2	signal transducer and activator of transcription 2, 113kDa [Source:HGNC Symbol;Acc:11363]	ENSFM00500000269705	SIGNAL TRANSDUCER AND ACTIVATOR OF TRANSCRIPTION	KNOWN
G/A	TAC-TAt	ENST00000318386	ENSP00000316490	Y158Y	EXON CDS	rs2247341:A	Synonymous	N/A	N/A	N/A	N/A	ENSG00000163950	SLBP	stem-loop binding protein [Source:HGNC Symbol;Acc:10904]	ENSFM00250000004776	HISTONE RNA HAIRPIN BINDING HISTONE STEM LOOP BINDING	KNOWN

The column **Codons** in the output table shows the originally and changed codon of the input position coordinates. The **Transcript ID** and **Protein ID** give information about the affected gene/protein. If the **SNP Type** is Nonsynonymous, the **Prediction** of the protein function is DAMAGING, which means that the functionality of the protein is predicted as being compromised.

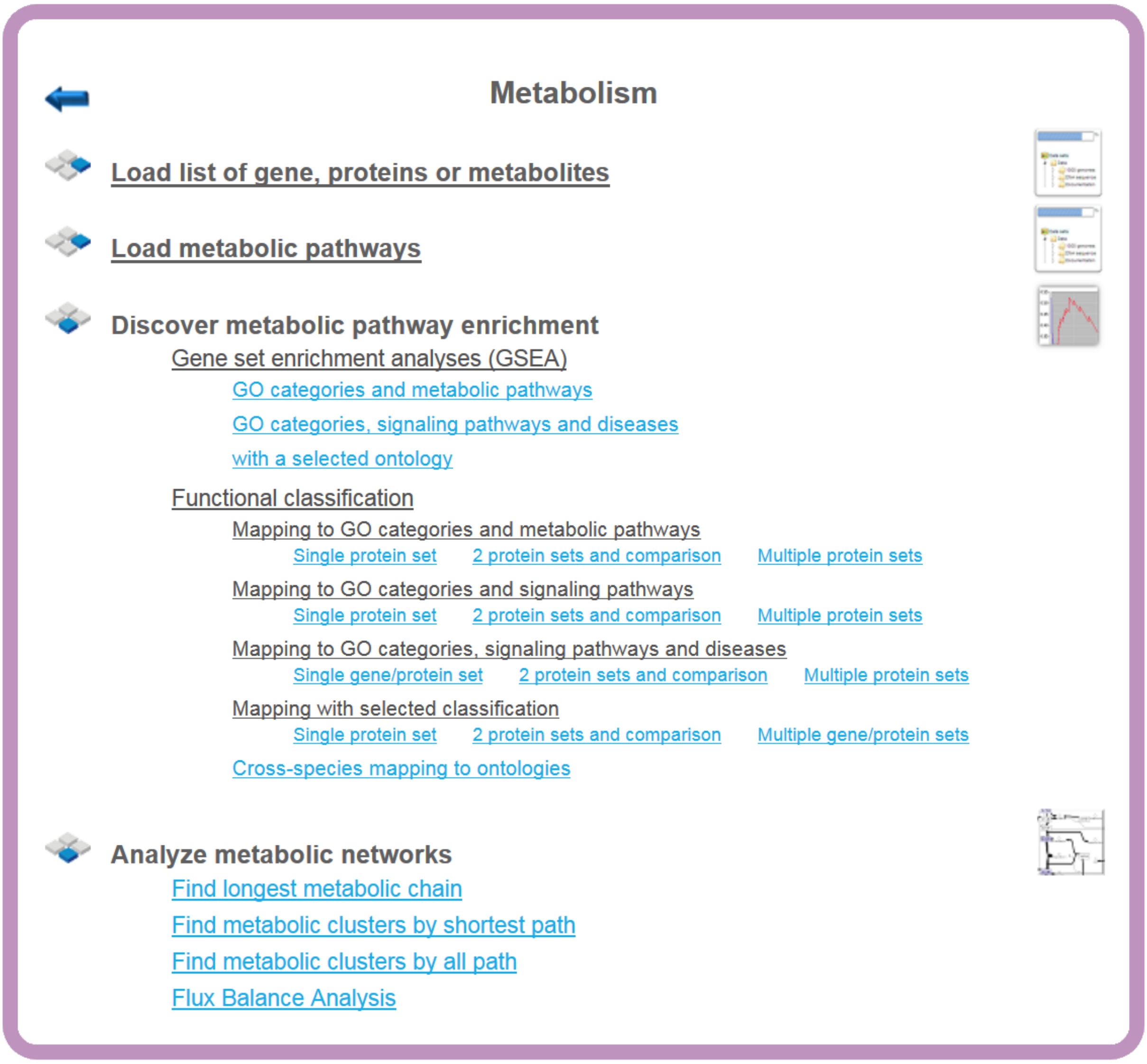
Selected additional fields like Gene ID, Gene Name and others are shown in the output table.

14.5. Further workflows in this area

For the other workflows that you can find in the area ***Genomic variants***, please refer to the following Sections:

Load genome variation data	See Chapter 3
Identify and classify variant genes	See Sections 6.1.1, 7.2.1
Discover functional enrichment among variant genes	See Section 10.3
Analyze networks of with variant genes	See Section 5.1

15. Metabolism



Metabolism

Load list of gene, proteins or metabolites

Load metabolic pathways

Discover metabolic pathway enrichment

Gene set enrichment analyses (GSEA)

GO categories and metabolic pathways

GO categories, signaling pathways and diseases

with a selected ontology

Functional classification

Mapping to GO categories and metabolic pathways

Single protein set 2 protein sets and comparison Multiple protein sets

Mapping to GO categories and signaling pathways

Single protein set 2 protein sets and comparison Multiple protein sets

Mapping to GO categories, signaling pathways and diseases

Single gene/protein set 2 protein sets and comparison Multiple protein sets

Mapping with selected classification

Single protein set 2 protein sets and comparison Multiple gene/protein sets

Cross-species mapping to ontologies

Analyze metabolic networks

Find longest metabolic chain

Find metabolic clusters by shortest path

Find metabolic clusters by all path

Flux Balance Analysis

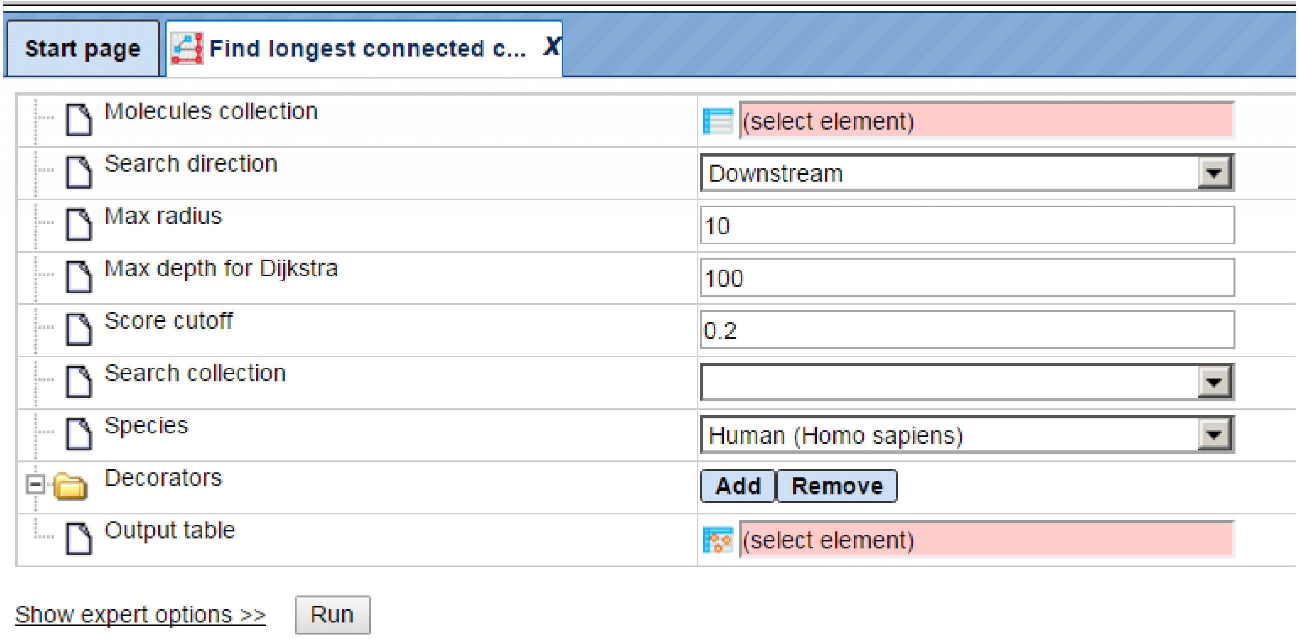
15.1. Analyze metabolic networks

15.1.1. Find longest metabolic chain

The goal of this analysis is to find longest chains which contain as many elements from the input collection as possible. Here chain means a path which starts and ends with the elements from the input collection. In this path the length between two elements from the input collection is limited by the maximum search radius.

This method can be found under the analyses tab using the path analyses/Methods/Molecular networks/Find longest connected chains.

The input form of the method looks as shown below:



Start page | Find longest connected c...

Molecules collection	(select element)
Search direction	Downstream
Max radius	10
Max depth for Dijkstra	100
Score cutoff	0.2
Search collection	
Species	Human (Homo sapiens)
Decorators	Add Remove
Output table	(select element)

Show expert options >> Run

Step 1: Specify the **Molecules collection**, which can be any molecule, protein or gene list.

To specify the input table, you can drag & drop it from your project within the tree area.

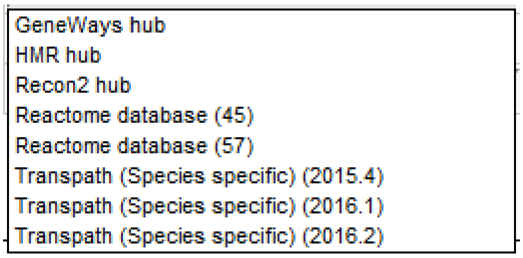
Step 2: Specify the **Search direction**, either upstream, downstream reactions or both directions.

Step 3: Selection of **Max**imal search **radius**, the default is 10.

Step 4: Selection of **Max**imal **depth** which will be used **by the Dijkstra** search algorithm, the default is 100.

Step 5: Specify the **Score cutoff** – Molecules with a Score lower than specified will be excluded from the result.

Step 6: Specify the **Search collection**, which can be one of the drop-down menu shown below.



GeneWays hub
HMR hub
Recon2 hub
Reactome database (45)
Reactome database (57)
Transpath (Species specific) (2015.4)
Transpath (Species specific) (2016.1)
Transpath (Species specific) (2016.2)

Step 7: Specify the biological species of the input set in the field **Species** by selecting the desired species from the drop-down menu.

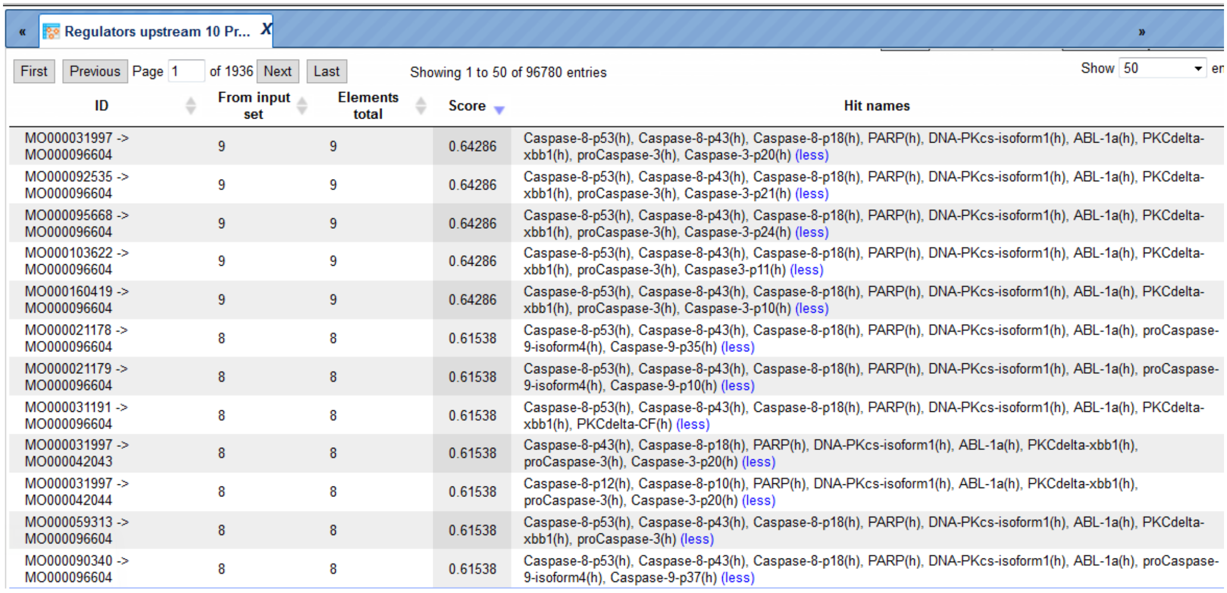
Step 8: Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink box (select element) in the field **Output name**, and a new window will open, where you can select the location of the results folder and define its name.

Example:

data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/Find Master regulator_Transpath/Upregulated_Ensembl/Regulators upstream 10 Proteins Transpath peptides

Using all other default parameters, press run and wait for the method to complete.

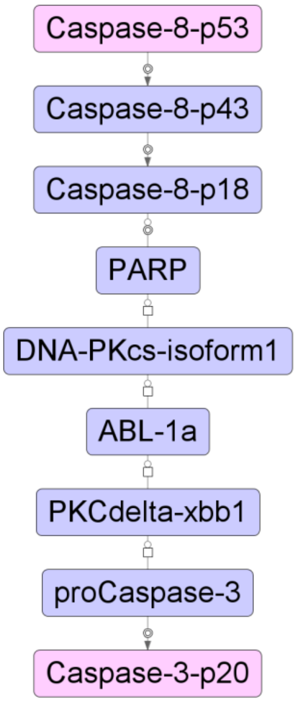
The result is a table which opens by default as shown below:



Showing 1 to 50 of 96780 entries

ID	From input set	Elements total	Score	Hit names
MO000031997 -> MO000096604	9	9	0.64286	Caspase-8-p53(h), Caspase-8-p43(h), Caspase-8-p18(h), PARP(h), DNA-PKcs-isoform1(h), ABL-1a(h), PKCdelta-xbb1(h), proCaspase-3(h), Caspase-3-p20(h) (less)
MO000092535 -> MO000096604	9	9	0.64286	Caspase-8-p53(h), Caspase-8-p43(h), Caspase-8-p18(h), PARP(h), DNA-PKcs-isoform1(h), ABL-1a(h), PKCdelta-xbb1(h), proCaspase-3(h), Caspase-3-p21(h) (less)
MO000095668 -> MO000096604	9	9	0.64286	Caspase-8-p53(h), Caspase-8-p43(h), Caspase-8-p18(h), PARP(h), DNA-PKcs-isoform1(h), ABL-1a(h), PKCdelta-xbb1(h), proCaspase-3(h), Caspase-3-p24(h) (less)
MO000103622 -> MO000096604	9	9	0.64286	Caspase-8-p53(h), Caspase-8-p43(h), Caspase-8-p18(h), PARP(h), DNA-PKcs-isoform1(h), ABL-1a(h), PKCdelta-xbb1(h), proCaspase-3(h), Caspase3-p11(h) (less)
MO000160419 -> MO000096604	9	9	0.64286	Caspase-8-p53(h), Caspase-8-p43(h), Caspase-8-p18(h), PARP(h), DNA-PKcs-isoform1(h), ABL-1a(h), PKCdelta-xbb1(h), proCaspase-3(h), Caspase-3-p10(h) (less)
MO000021178 -> MO000096604	8	8	0.61538	Caspase-8-p53(h), Caspase-8-p43(h), Caspase-8-p18(h), PARP(h), DNA-PKcs-isoform1(h), ABL-1a(h), proCaspase-9-isoform4(h), Caspase-9-p35(h) (less)
MO000021179 -> MO000096604	8	8	0.61538	Caspase-8-p53(h), Caspase-8-p43(h), Caspase-8-p18(h), PARP(h), DNA-PKcs-isoform1(h), ABL-1a(h), proCaspase-9-isoform4(h), Caspase-9-p10(h) (less)
MO000031191 -> MO000096604	8	8	0.61538	Caspase-8-p53(h), Caspase-8-p43(h), Caspase-8-p18(h), PARP(h), DNA-PKcs-isoform1(h), ABL-1a(h), PKCdelta-xbb1(h), PKCdelta-CF(h) (less)
MO000031997 -> MO000042043	8	8	0.61538	Caspase-8-p43(h), Caspase-8-p18(h), PARP(h), DNA-PKcs-isoform1(h), ABL-1a(h), PKCdelta-xbb1(h), proCaspase-3(h), Caspase-3-p20(h) (less)
MO000031997 -> MO000042044	8	8	0.61538	Caspase-8-p12(h), Caspase-8-p10(h), PARP(h), DNA-PKcs-isoform1(h), ABL-1a(h), PKCdelta-xbb1(h), proCaspase-3(h), Caspase-3-p20(h) (less)
MO000059313 -> MO000096604	8	8	0.61538	Caspase-8-p53(h), Caspase-8-p43(h), Caspase-8-p18(h), PARP(h), DNA-PKcs-isoform1(h), ABL-1a(h), PKCdelta-xbb1(h), proCaspase-3(h) (less)
MO000090340 -> MO000096604	8	8	0.61538	Caspase-8-p53(h), Caspase-8-p43(h), Caspase-8-p18(h), PARP(h), DNA-PKcs-isoform1(h), ABL-1a(h), proCaspase-9-isoform4(h), Caspase-9-p37(h) (less)

All chains within the radius 10 are included in the results. You can click on each row and visualize the results as shown below:



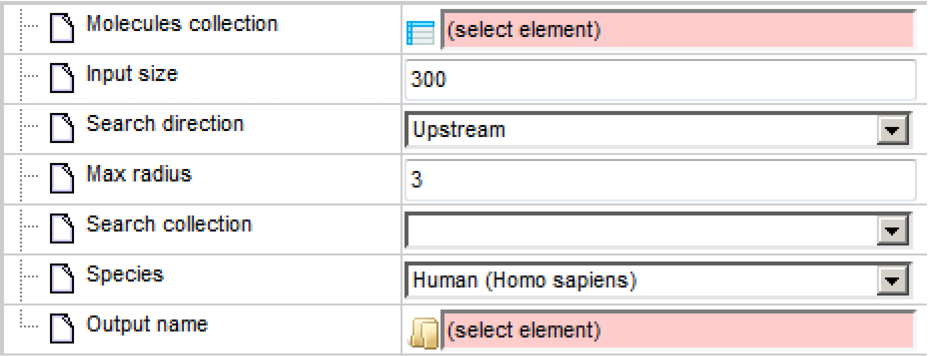
Similarly other long chains can be visualized from the result table.

15.1.2. Find metabolic clusters by shortest path

Please refer to Section 3 and apply the steps explained there correspondingly.

15.1.3. Find metabolic clusters by all path

This analysis allows you to generate a cluster of genes/molecules upstream or downstream or both by taking reactions and all intermediate molecules from a specified search collection. To launch the analysis, open the method form from the Start page:



Molecules collection	(select element)
Input size	300
Search direction	Upstream
Max radius	3
Search collection	
Species	Human (Homo sapiens)
Output name	(select element)

Step 1: Specify the **Molecules collection**, which can be any molecule, protein or gene list.

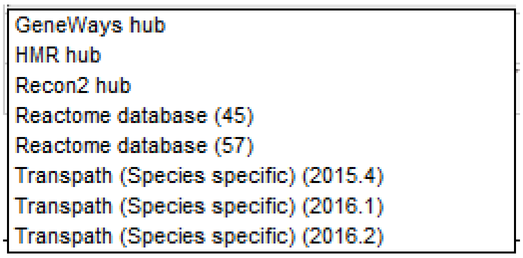
To specify the input table, you can drag & drop it from your project within the tree area.

Step 2: Select the maximum **Input size** (expert modus).

Step 3: Specify the **Search direction**, either upstream, downstream reactions or both directions.

Step 4: Selection of **Maximal search radius**, the default is 3.

Step 5: Specify the **Search collection**, which can be one of the drop-down menu shown below.



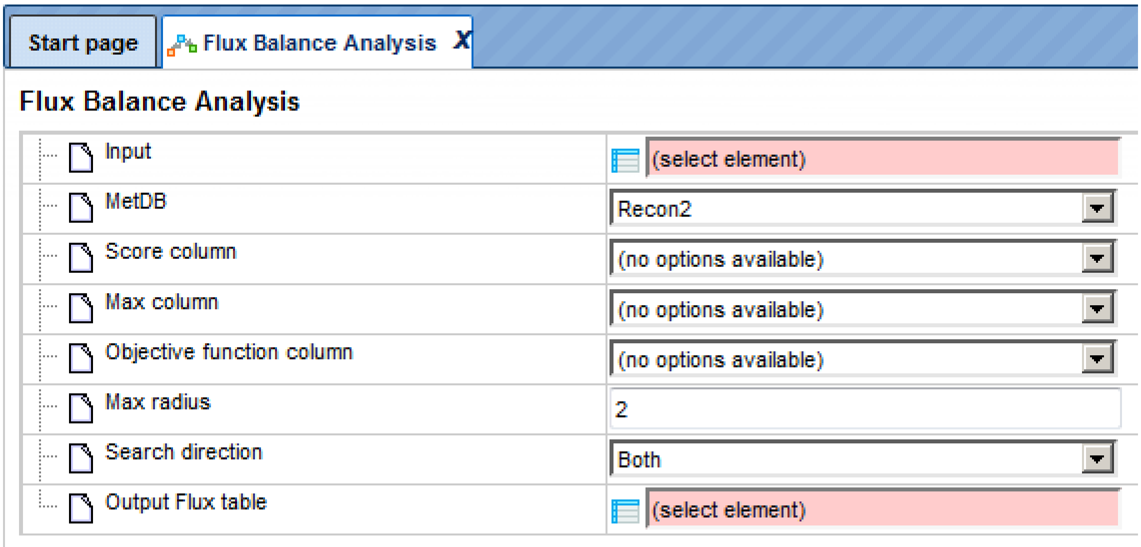
GeneWays hub
HMR hub
Recon2 hub
Reactome database (45)
Reactome database (57)
Transpath (Species specific) (2015.4)
Transpath (Species specific) (2016.1)
Transpath (Species specific) (2016.2)

Step 6: Specify the biological species of the input set in the field **Species** by selecting the desired species from the drop-down menu.

Step 7: Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink box (select element) in the field **Output name**, and a new window will open, where you can select the location of the results folder and define its name.

15.1.4. Flux Balance Analysis

To launch this workflow, open the workflow input form from the Start page:



Flux Balance Analysis

Input	(select element)
MetDB	Recon2
Score column	(no options available)
Max column	(no options available)
Objective function column	(no options available)
Max radius	2
Search direction	Both
Output Flux table	(select element)

Step 1: Specify a gene set under study, e.g. a list of differentially regulated genes, as the Input Yes gene set. You can drag & drop it from your project within the Tree Area and drop in the pink box of the field **Input gene set**.

For this example, all further steps are demonstrated with the following input set:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Experiment%20normalized%20(RMA)%20(Differentially%20expressed%20genes%20Affy)/Upregulated%20Ensembl%20genes%20filtered%20(LogFC%3E1)

Step 2: Specify the metabolism database (**MetDB**) by selecting Recon2 or HMR from the drop-down menu. As default the Recon2 database is selected.

Step 3: Define the **Score column, Max column** and **Objective function column** by selecting one numerical column from your input table from the drop down menu.

Step 4: Define the **Max radius**, which is the maximal number of steps within the selected metabolism database. The default value is 2.

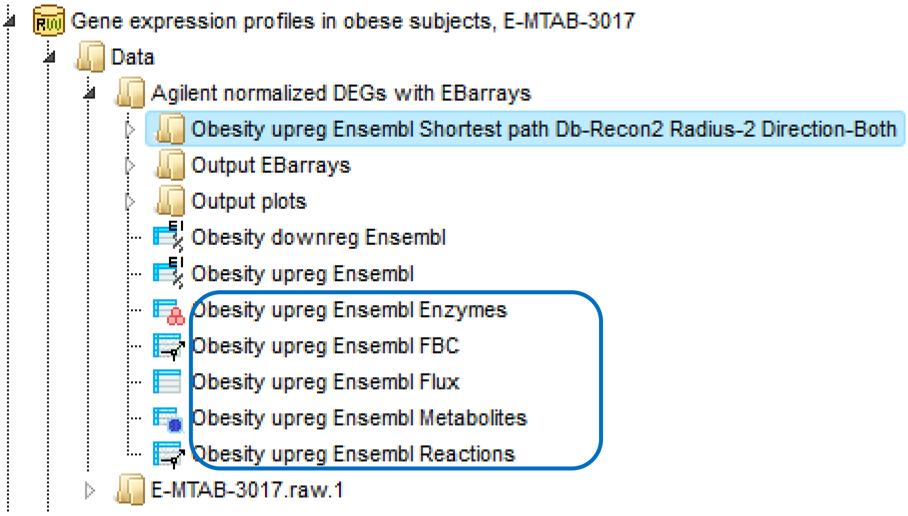
Step 5: Specify the **Search direction** by selecting from the drop-down menu; default is both.

Step 6: Define where the output table should be located in your project tree. You can do so by clicking on the pink box (select element) in the field **Output flux table**, and a new window will open, where you can select the location of the results table and define its name.

Step 7: Press the [Run workflow] button. Wait until the workflow is completed, and take a look at the results.

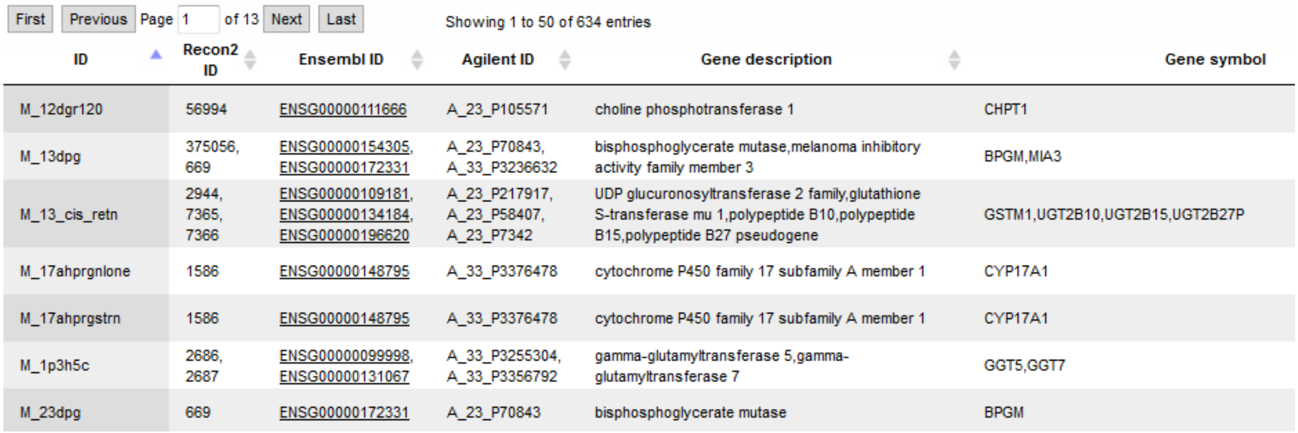
Visualization and interpretation of results

The results consist of several files and one folder:



- Gene expression profiles in obese subjects, E-MTAB-3017
  - Data
    - Agilent normalized DEGs with EBarrays
      - Obesity upreg Ensembl Shortest path Db-Recon2 Radius-2 Direction-Both
      - Output EBarrays
      - Output plots
      - Obesity downreg Ensembl
      - Obesity upreg Ensembl
      - Obesity upreg Ensembl Enzymes
      - Obesity upreg Ensembl FBC
      - Obesity upreg Ensembl Flux
      - Obesity upreg Ensembl Metabolites
      - Obesity upreg Ensembl Reactions
    - E-MTAB-3017.raw.1

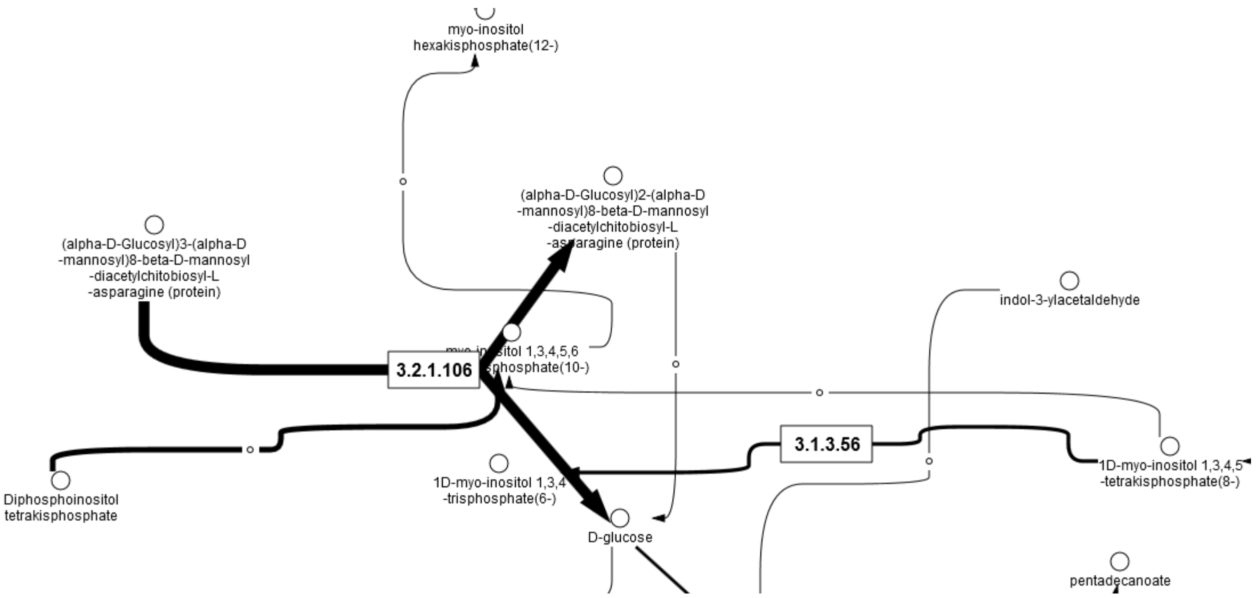
The first step of the workflow is to convert the input genes into enzymes (output example: Obesity upreg Ensembl Enzymes). Next step is to match the enzymes with metabolites (output example: Obesity upreg Ensembl Metabolites). The 634 metabolites are listed in table format with Recon2 IDs.



Showing 1 to 50 of 634 entries

ID	Recon2 ID	Ensembl ID	Agilent ID	Gene description	Gene symbol
M_12dgr120	56994	ENSG00000111666	A_23_P105571	choline phosphotransferase 1	CHPT1
M_13dpg	375056, 669	ENSG00000154305, ENSG00000172331	A_23_P70843, A_33_P3236632	bisphosphoglycerate mutase,melanoma inhibitory activity family member 3	BPGM,MIA3
M_13_cis_retn	2944, 7365, 7366	ENSG00000109181, ENSG00000134184, ENSG00000196620	A_23_P217917, A_23_P58407, A_23_P7342	UDP glucuronosyltransferase 2 family,glutathione S-transferase mu 1,polypeptide B10,polypeptide B15,polypeptide B27 pseudogene	GSTM1,UGT2B10,UGT2B15,UGT2B27P
M_17ahprgnlone	1586	ENSG00000148795	A_33_P3376478	cytochrome P450 family 17 subfamily A member 1	CYP17A1
M_17ahprgstrn	1586	ENSG00000148795	A_33_P3376478	cytochrome P450 family 17 subfamily A member 1	CYP17A1
M_1p3h5c	2686, 2687	ENSG00000099998, ENSG00000131067	A_33_P3255304, A_33_P3356792	gamma-glutamyltransferase 5,gamma-glutamyltransferase 7	GGT5,GGT7
M_23dpg	669	ENSG00000172331	A_23_P70843	bisphosphoglycerate mutase	BPGM

Parallel to the described conversion and match steps, the enzyme list is converted into an enzyme reaction table (output example: *Obesity upreg Ensembl Reactions*), and 663 reactions are listed in table format with Recon2 IDs. Then a Flux Balance Table (*Obesity upreg Ensembl FBC*) is calculated, where the input parameters **Score column, Max column and Objective function column** are used. After that, a table with flux data is constructed. From this *Obesity upreg Ensembl Flux* table, together with the first cluster, a new diagram (*Cluster 1 flux*) is created, which shows the input proteins from cluster 1 with added and calculated flux information. A screen from the output is shown below:



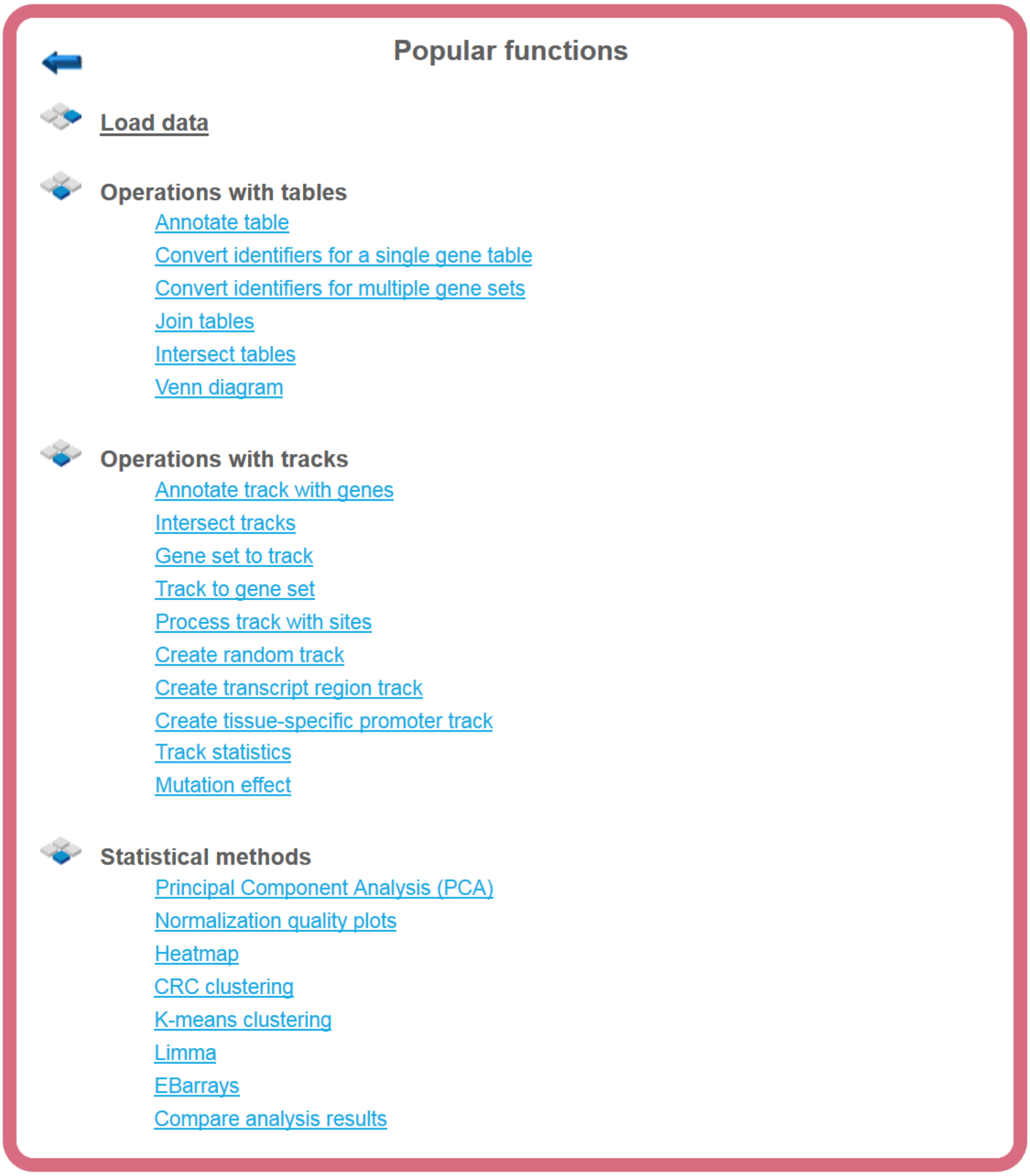
The folder *Obesity upreg Ensembl Shortest path Db-Recon2 Radius-2 Direction-Both* contains all results from the cluster analysis within the workflow.

15.2. Further workflows in this area

For the other workflows that you can find in the area ***Metabolism***, please refer to the following Sections:

Load list of gene, proteins or metabolites	See Chapter 3
Load metabolic pathways	See Chapter 3
Discover metabolic pathway enrichment	See Section 10.3

16. Popular functions



Popular functions

Load data

Operations with tables
- Annotate table
- Convert identifiers for a single gene table
- Convert identifiers for multiple gene sets
- Join tables
- Intersect tables
- Venn diagram

Operations with tracks
- Annotate track with genes
- Intersect tracks
- Gene set to track
- Track to gene set
- Process track with sites
- Create random track
- Create transcript region track
- Create tissue-specific promoter track
- Track statistics
- Mutation effect

Statistical methods
- Principal Component Analysis (PCA)
- Normalization quality plots
- Heatmap
- CRC clustering
- K-means clustering
- Limma
- EBarrays
- Compare analysis results

16.1. Operations with tables

Any table may be opened by double-clicking the corresponding name in the Tree Area. It will open under a new tab in the Work Space.

The contents of the table are sorted according to the values in one of its columns. Being opened for the first time, a default column is defined for sorting, usually the ID column. This default column is indicated by a blue arrowhead. If this arrowhead points upwards, the table rows are sorted in ascending order of this column’s values. Clicking on this arrowhead will change it into a downwards pointing one, while the values are sorted in descending order. Correspondingly, you may sort the table according to the values of any other column in ascending or descending order by clicking on the up- or downwards pointing gray arrowhead on top of this column, respectively.

On top of the table, you can navigate between the individual pages of the table; it is also shown on which page out of how many pages in total you are, and in the right top corner, the page size in terms of number of entries (rows) is shown and can be adjusted.

You can edit the contents of a table by pressing the [Edit] button in the right upper corner. Now, you can manually edit the contents of each cell in the table. With the [Apply] option, you will save this change, while [Cancel] quits it.

Even without activating the Edit function, you can select

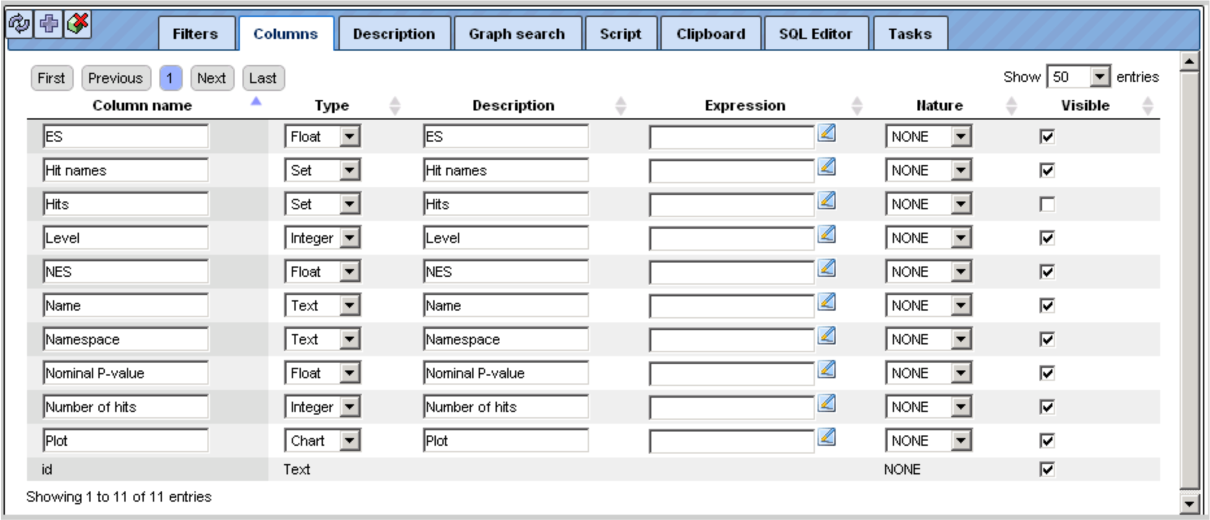
- individual rows with a left-mouse click,
- several ones by keeping the Ctrl key pressed,
- a range of rows with the Shift key pressed when clicking on the last row of the range to be selected, or
- [Select all] by clicking on the corresponding button.

The selected rows can be saved as a separate file, which by default is given the name *<original file name> subset*, but you can change this name.

Changing the table structure in the Operations Field

Having opened a table in the Work Space, e.g. by double clicking on its name in the Tree Area, it is possible to edit its structure in the Operations Field under the tab *Columns*.

For instance, if you have opened a table with data about Enrichment GO Molecular Mechanism (resulting from having run a GSEA), this field may look like this:



Column name	Type	Description	Expression	Nature	Visible
ES	Float	ES		NONE	☑
Hit names	Set	Hit names		NONE	☑
Hits	Set	Hits		NONE	☐
Level	Integer	Level		NONE	☑
NES	Float	NES		NONE	☑
Name	Text	Name		NONE	☑
Namespace	Text	Namespace		NONE	☑
Nominal P-value	Float	Nominal P-value		NONE	☑
Number of hits	Integer	Number of hits		NONE	☑
Plot	Chart	Plot		NONE	☑
id	Text			NONE	☑

Showing 1 to 11 of 11 entries

Recognizably, you can change the column headers, the data type in the column, or its (usually hidden) descriptions. You may add an Expression, which may be a mathematical formula, formulated in Java script; you find detailed explanations for this when you press the Edit key (✎) next to this field. In the last column, you can specify which columns are visible or shall be hidden (unmarking a column here does NOT delete it, it hides it from the currently displayed table).

If you hide a column by unmarking it, you have to refresh the Work Space by pressing the button 🔄 in the control panel right on top of the Operations Field. Here, you can also add new (➕) columns. Before removing a column with the button ✖, you have to mark it by clicking somewhere in the background of the line specifying this column; the selected item will be highlighted in blue. But be careful: Deleting it from the table will irrevocably erase the column including all its contents!

16.1.1. Annotate table

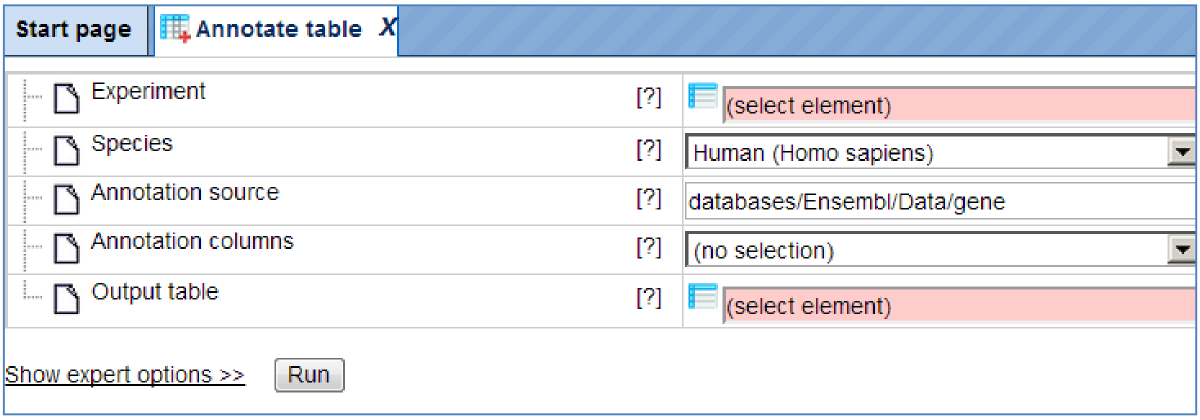
The analysis method *Annotate table* ( ) can be found in the Tree Area, under the Analyses tab in the folder *Methods*, subfolder *Data manipulation*. The complete path to this method is:

http://genexplain-platform.com/bioumlweb/#de=analyses/Methods/Data%20manipulation/Annotate%20table

Applying the *Annotate table* method, you can add columns to any gene or protein table in the tree. The source for annotation columns can be a database or any other table.

The input table will not be changed. As a result of the analysis, a new table with additional columns will be created.

The input form of this method, when opened in the work space, is shown below:



Experiment	[?]	(select element)
Species	[?]	Human (Homo sapiens)
Annotation source	[?]	databases/Ensembl/Data/gene
Annotation columns	[?]	(no selection)
Output table	[?]	(select element)

In the following, we will consider the input fields one by one:

Experiment – Input a table for which you wish to add annotation. In order for this analysis to work properly, the ID column of this table should contain recognizable biological identifiers that can be mapped to the annotation source identifiers.

Species - Species corresponding to the input table. By default, human is selected. If your input table corresponds to a mouse or rat dataset, please specify it.

Annotation source – Select the data collection, a database or any table in the tree area, which you plan to use as a source of additional columns. Below two examples are given, with the Ensembl database and with a user-specific table as possible annotation sources.

Annotation Columns - As soon as the annotation source is specified, all columns of this table are visible in the drop-down menu. There you can select one or several columns from the drop-down menu, which will be added to the input table.

Output table – Select the location in the tree area where the resulting table will be stored, and define a name for the new table. If a table with the same name already exists at the same location, it will be replaced.

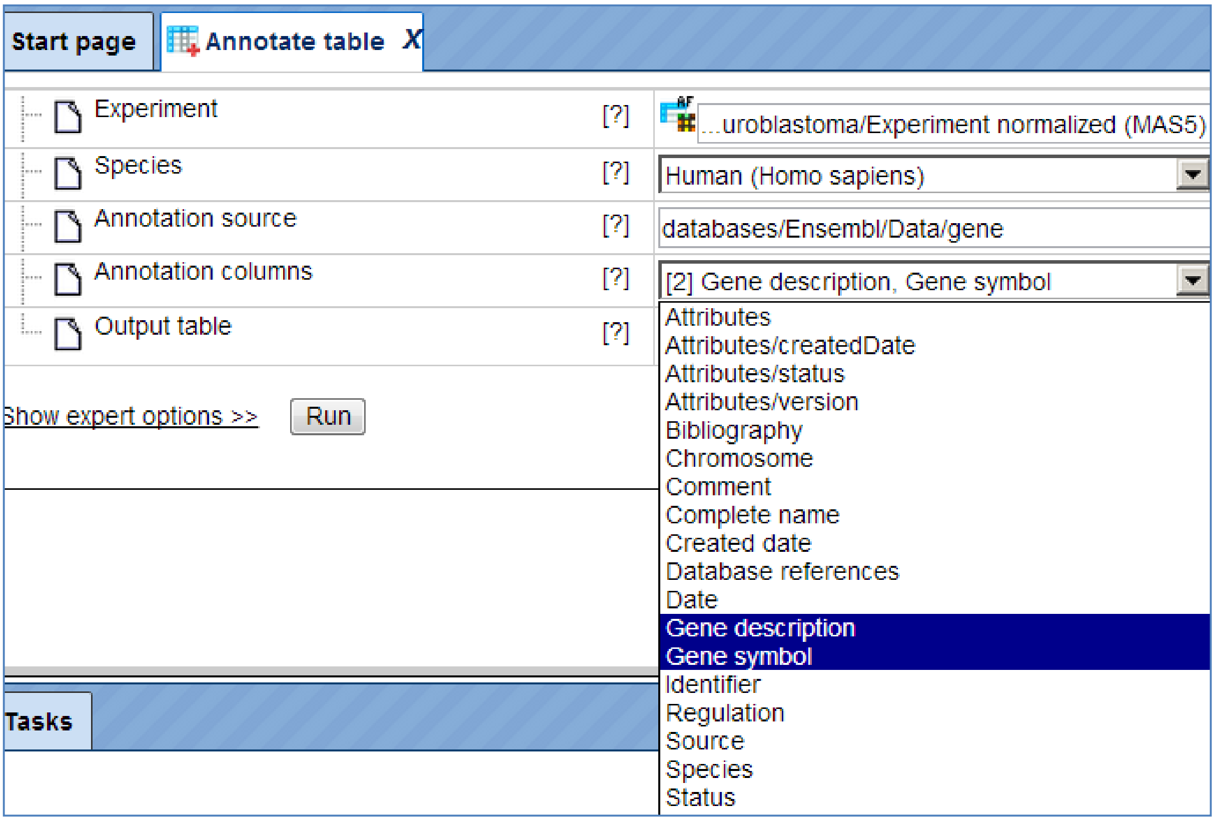
Annotate table of Affymetrix probe IDs with gene description and gene symbol

As input, a table with normalized Affymetrix probes is selected. This input file can be accessed with the URL:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Experiment%20normalized%20(RMA)

In the field **Species** *Human* is selected, because the input table corresponds to a human dataset.

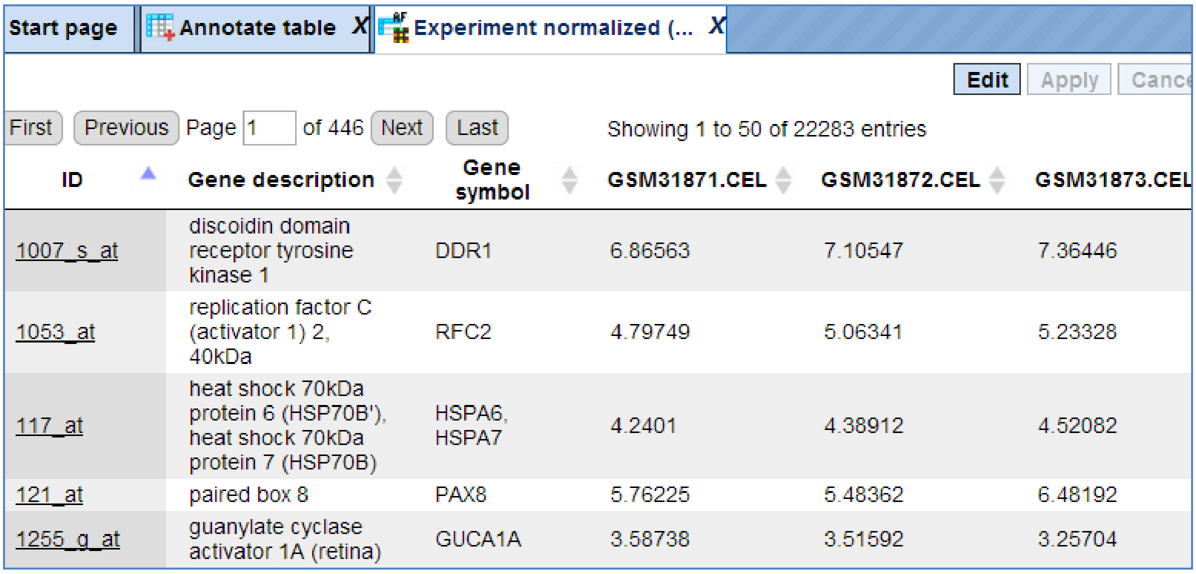
By default, the Ensembl database, namely its gene table, is selected in the field **Annotation source**. All columns present in this table are available in the drop-down menu of the field Annotation columns. As shown in the screenshot below, two columns are selected, **Gene description** and **Gene symbol**:



Experiment [?] ...uroblastoma/Experiment normalized (MAS5)
Species [?] Human (Homo sapiens)
Annotation source [?] databases/Ensembl/Data/gene
Annotation columns [?] [2] Gene description, Gene symbol
Output table [?]

Next, the output path is defined, and you can press the [Run] button.

After completion of the analysis the output file is opened automatically in the work space as shown below:



Showing 1 to 50 of 22283 entries

ID	Gene description	Gene symbol	GSM31871.CEL	GSM31872.CEL	GSM31873.CEL
1007_s_at	discoidin domain receptor tyrosine kinase 1	DDR1	6.86563	7.10547	7.36446
1053_at	replication factor C (activator 1) 2, 40kDa	RFC2	4.79749	5.06341	5.23328
117_at	heat shock 70kDa protein 6 (HSP70B'), heat shock 70kDa protein 7 (HSP70B)	HSPA6, HSPA7	4.2401	4.38912	4.52082
121_at	paired box 8	PAX8	5.76225	5.48362	6.48192
1255_g_at	guanylate cyclase activator 1A (retina)	GUCA1A	3.58738	3.51592	3.25704

In this result table two new columns are added, **Gene description** and **Gene symbol**, to the right of the **ID** column. The **ID** column itself, and all the other columns are exactly the same as they were in the input table.

Annotate a gene or protein table with expression values

You may wish to see the expression values in any gene or protein table. In this example, let's consider the annotation of the master regulatory molecules table with fold change values. Such a table can be generated, e.g. by the workflows *Find master regulators in networks,* described in Section 5.1.1.

Further steps are shown with the following input table:

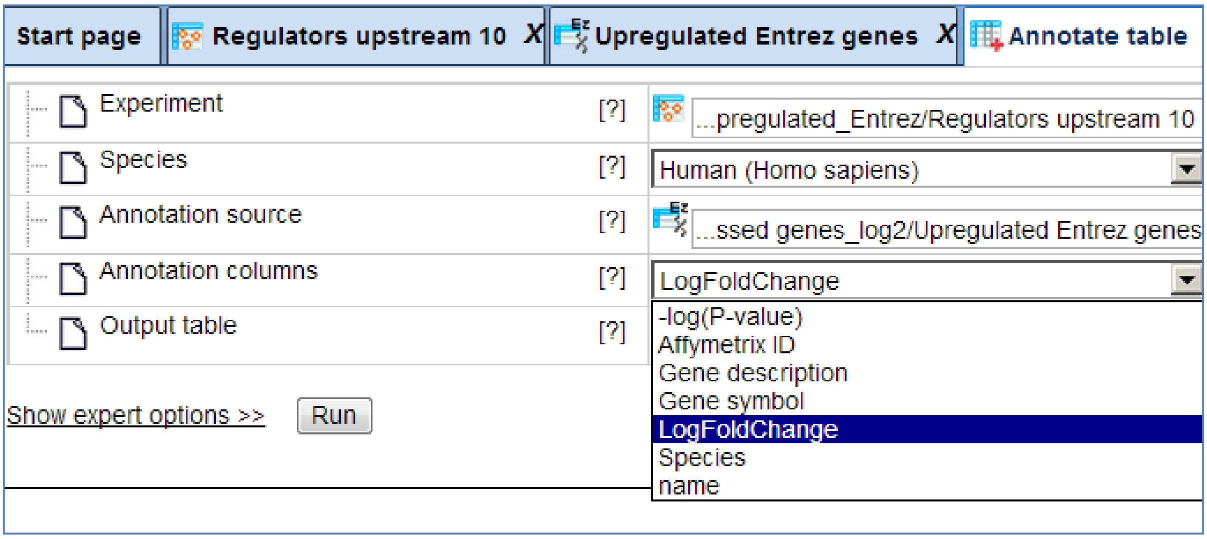
http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Upregulated%20Ensembl%20genes%20filtered%20(LogFC%3E1)%20(Master%20regulators%20Transpath)/Regulators%20upstream%2010

In the field **Species** *Human* is selected, because the input table corresponds to a human dataset.

As **Annotation source** you can use a table with expression values corresponding to this dataset. You may have such a table in your tree area, e.g. a table with differentially expressed genes. In this example, the following table is used as annotation source:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Experiment%20normalized%20(RMA)%20(Differentially%20expressed%20genes%20Affy)/Upregulated%20Ensembl%20genes%20filtered%20(LogFC%3E1)

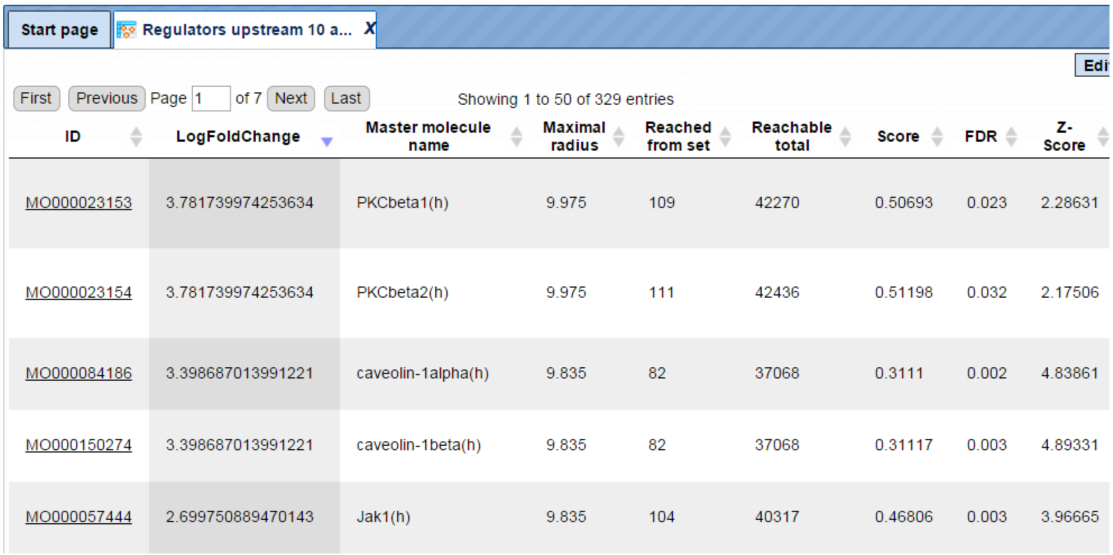
As soon as the annotation source table is selected in the field **Annotation source,** you can see all available columns in the drop-down menu of the **Annotation columns** field. The **LogFoldchange** column is selected as shown below:



Experiment	[?]	...pregulated_Entrez/Regulators upstream 10
Species	[?]	Human (Homo sapiens)
Annotation source	[?]	...ssed genes_log2/Upregulated Entrez genes
Annotation columns	[?]	LogFoldChange
Output table	[?]	

Next, the output path is defined, and you can press the [Run] button.

After completion of the analysis the output file is opened automatically in the work space as shown below:



Showing 1 to 50 of 329 entries

ID	LogFoldChange	Master molecule name	Maximal radius	Reached from set	Reachable total	Score	FDR	Z-Score
MO000023153	3.781739974253634	PKCbeta1(h)	9.975	109	42270	0.50693	0.023	2.28631
MO000023154	3.781739974253634	PKCbeta2(h)	9.975	111	42436	0.51198	0.032	2.17506
MO000084186	3.398687013991221	caveolin-1alpha(h)	9.835	82	37068	0.3111	0.002	4.83861
MO000150274	3.398687013991221	caveolin-1beta(h)	9.835	82	37068	0.31117	0.003	4.89331
MO000057444	2.699750889470143	Jak1(h)	9.835	104	40317	0.46806	0.003	3.96665

In this result table one new column is added, **LogFoldChange**, to the right of the **ID** column. The **ID** column, and all the other columns are exactly the same as they were in the input table.

16.1.2. Convert identifiers for a gene table

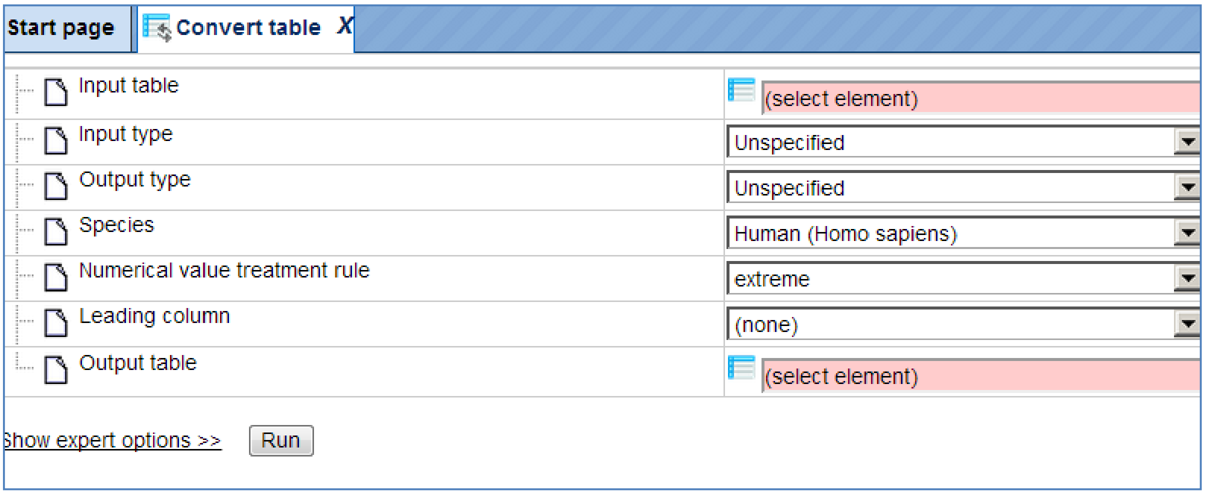
Single gene table

The analysis method *Convert table* () can be found in the tree area, on the Analyses tab in the folder *Methods*, subfolder *Data manipulation*. The complete path to this method is:

http://genexplain-platform.com/bioumlweb/#de=analyses/Methods/Data%20manipulation/Convert%20table

This method changes the type of identifiers using the internal chain of BioHubs. BioHubs is an internal, proprietary database that maps the IDs of a wealth of data source to each other. For example, this method converts the *Genes: Ensembl* type of identifiers into *Proteins: Ensembl*. If a direct conversion between two selected types is impossible, this analysis will create an optimal chain of several BioHubs and use them subsequently.

The analysis input form when opened in the work space is shown below:



Input table	(select element)
Input type	Unspecified
Output type	Unspecified
Species	Human (Homo sapiens)
Numerical value treatment rule	extreme
Leading column	(none)
Output table	(select element)

In the following, we will consider the input fields one by one:

Input table: Input the data table for which you wish to convert the identifiers.

Input type: Type of identifiers in the input table. This is automatically detected in the majority of cases. However, if there are two columns with different identifiers in the input table, you can manually select the identifier you wish to convert.

Output type: Type of identifiers into which you wish to convert the input type.

Species: Select human, mouse or rat, corresponding to the input table.

Numerical value treatment rule: Select one of the rules for treating the values in the numerical columns of the input table. Rule selection is important, when several rows are merged into a single one. We have to take into account that one identifier of a given type may correspond to several identifiers of another type, each of which is associated with a numerical value in the Leading Column (for this, see below). To choose which of these numerical values has to be taken into the merged row, a rule has to be defined. It is to be chosen from a drop-down menu. By default the “extreme” rule is selected, which is equivalent to the maximal value in case of positive numbers, but corresponds to the minimal value in case of negative numbers. In cases of "average", "average w/o 20% outliers" and "sum", the selected rule is applied to all numerical columns of the table.

In case of the "minimum", "maximum" and "extreme" rules a new option appears below which requests the user to select a **Leading column**. The chosen rule is applied then to the values in the selected Leading column (e.g. in the Leading column the maximum value is computed among all merged rows). All other numerical values of the table will be taken from that row which corresponds to the selected value in the leading column.

Output table: Path to store the resulting table in the tree.

Note that several non-trivial situations might occur during conversion:

- A single source ID matches to several target IDs. In this case the source row will be copied several times, one copy per target ID.
- A source ID doesn't match to any target ID. In this case the source row will be removed from the result.
- Several source IDs match to a single target ID. In this case two options are available:

If you have specified the leading column, only one out of all suitable source rows will be shown in the resulting table, based on the specified rule. For example, if you specified 'maximum' as a rule, the source row with maximal value in the main column will be selected from suitable rows.

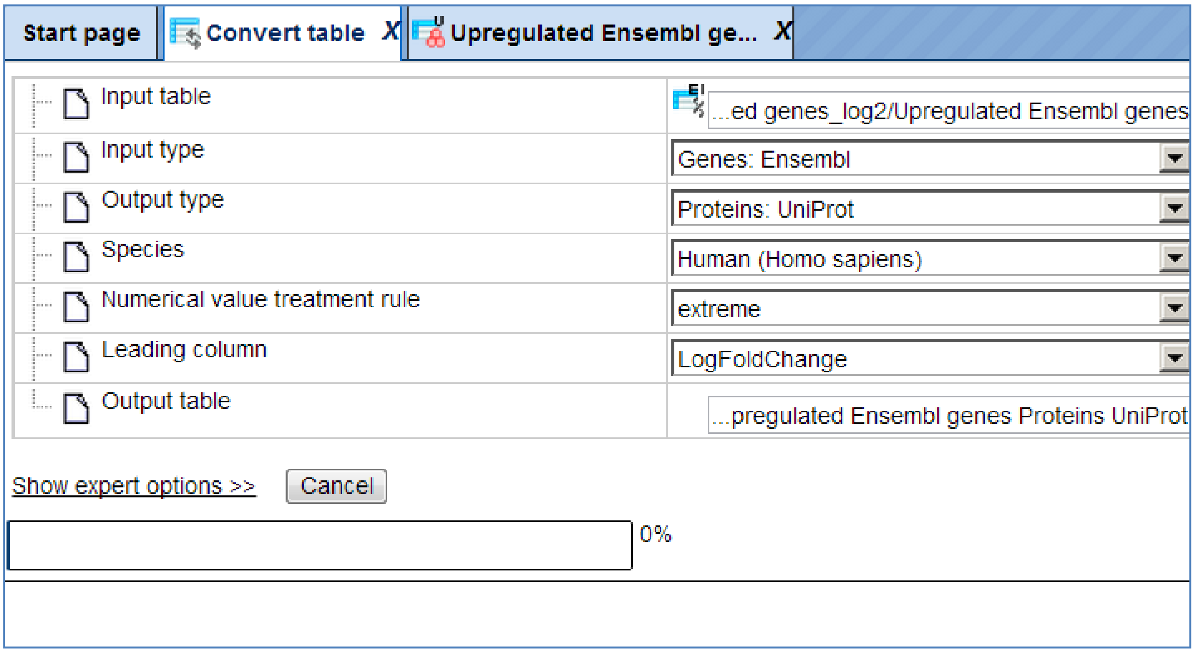
If you have not specified a leading column, all the corresponding source rows will be merged together using merging rules. Non-trivial columns like 'Graph' will not be shown in the resulting table. Text columns will have all values joined into a sorted comma-separated list with duplicates removed. Numerical columns will be merged based on the selected rule. For example, if you select 'average' as a rule, then the mean value will appear in the resulting table. If your source column has an integer type, it might be changed into float.

Example: Conversion of Ensembl gene IDs to UniProt IDs

The input table with Ensembl gene IDs can be accessed via URL:

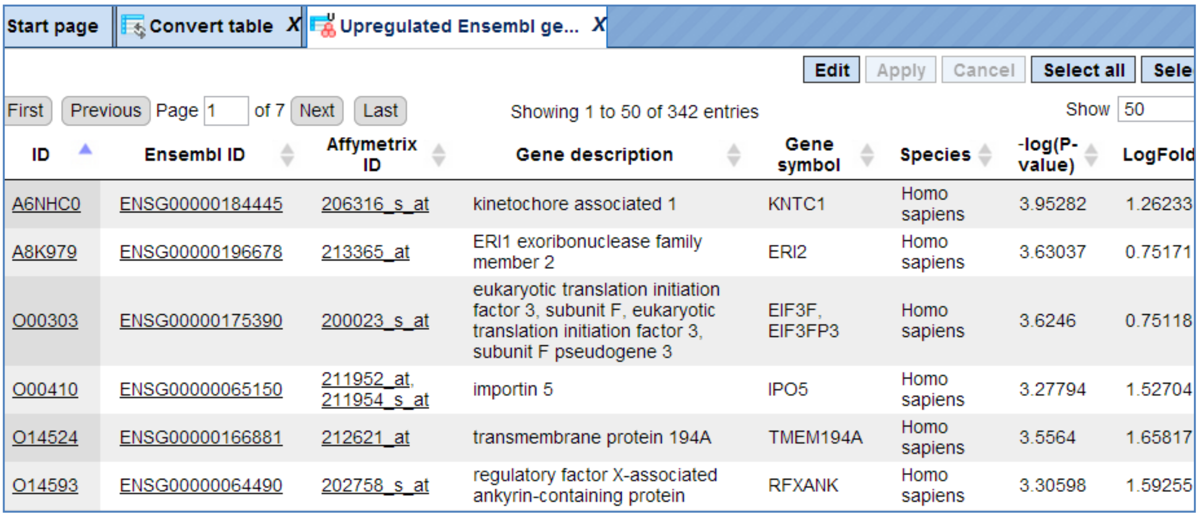
http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Experiment%20normalized%20(RMA)%20(Differentially%20expressed%20genes%20Affy)/Upregulated%20Ensembl%20genes

Fill the input fields and press [Run] as shown below:



Input table	...ed genes_log2/Upregulated Ensembl genes
Input type	Genes: Ensembl
Output type	Proteins: UniProt
Species	Human (Homo sapiens)
Numerical value treatment rule	extreme
Leading column	LogFoldChange
Output table	...pregulated Ensembl genes Proteins UniProt

Upon completion of the analysis the output file is opened by default as shown below:



Showing 1 to 50 of 342 entries

ID	Ensembl ID	Affymetrix ID	Gene description	Gene symbol	Species	-log(P-value)	LogFold
A6NHC0	ENSG00000184445	206316_s_at	kinetochore associated 1	KNTC1	Homo sapiens	3.95282	1.26233
A8K979	ENSG00000196678	213365_at	ERI1 exoribonuclease family member 2	ERI2	Homo sapiens	3.63037	0.75171
O00303	ENSG00000175390	200023_s_at	eukaryotic translation initiation factor 3, subunit F, eukaryotic translation initiation factor 3, subunit F pseudogene 3	EIF3F, EIF3FP3	Homo sapiens	3.6246	0.75118
O00410	ENSG00000065150	211952_at, 211954_s_at	importin 5	IPO5	Homo sapiens	3.27794	1.52704
O14524	ENSG00000166881	212621_at	transmembrane protein 194A	TMEM194A	Homo sapiens	3.5564	1.65817
O14593	ENSG00000064490	202758_s_at	regulatory factor X-associated ankyrin-containing protein	RFXANK	Homo sapiens	3.30598	1.59255

The column **ID** now contains UniProt IDs. The column **Ensembl ID**, which was the ID column in the input table, is also present in the output table, as the second column to the right of the new **ID** column.

All the other columns of the input table are included in the output table as well. Numerical values are calculated according to the selected rule.

Similarly, any other table in the tree area with gene or protein identifiers can be converted into the desired type of identifiers.

Multiple gene sets

The input is a folder with several gene tables. The steps of this workflow for each individual gene table are the same as described in the section above. The same steps are performed iteratively for each of the gene tables in the input folder.

The output is a folder which contains subfolders with the results for each individual input table. The subfolders are automatically given the same names as the input tables.

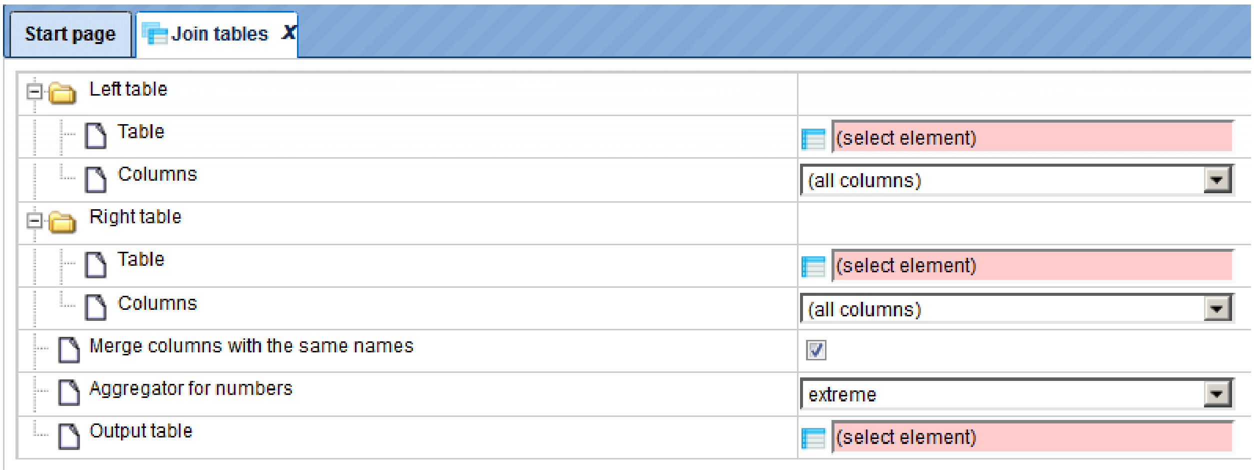
16.1.3. Join tables

The analysis method *Join tables* ( ) can be found in the Tree Area under the Analyses tab in the folder *Methods*, subfolder *Data manipulation*. The complete path to this method is:

http://genexplain-platform.com/bioumlweb/#de=analyses/Methods/Data%20manipulation/Join%20tables

Applying the *Join table* method, you can join two tables together in one new table containing selected columns. Joining is performed according to ID matching from left and right tables. The result table will contain IDs present in at least one table.

The input form of this method, when opened in the work space, is shown below:



Left table	
Table	(select element)
Columns	(all columns)
Right table	
Table	(select element)
Columns	(all columns)
Merge columns with the same names	☑
Aggregator for numbers	extreme
Output table	(select element)

In the following, we will consider the input fields one by one:

Left Table – left (first table) for join.

Right Table – right (second table) for join.

Aggregator for numbers (expert) – Function to be used for numerical columns when several rows are merged into a single one, if the merge columns option is selected.

Output table - Name of the table where results will be saved. If a table with that name already exists it will be replaced.

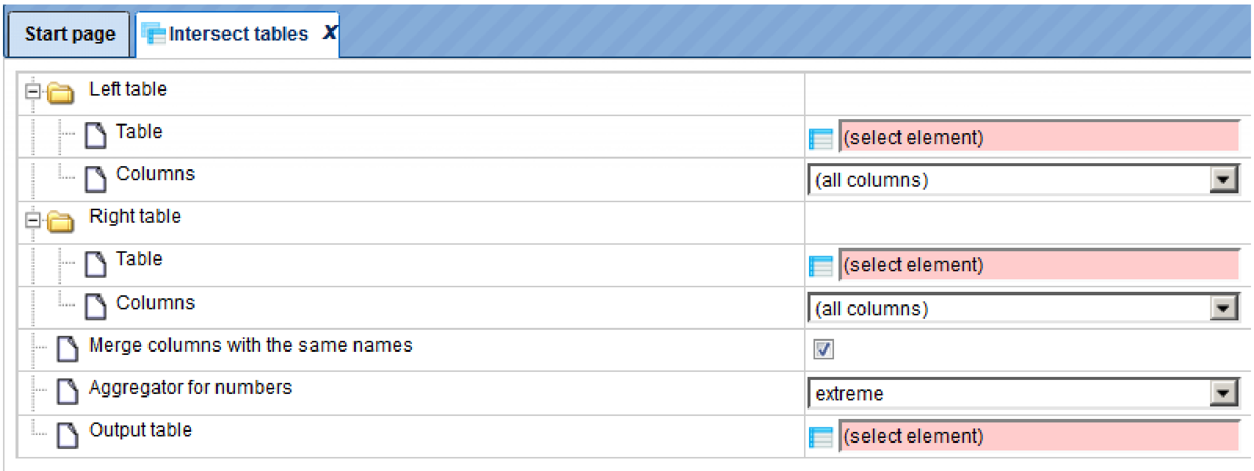
If you like to join more than two tables, please see method *Join several tables* ( ).

16.1.4. Intersect tables

The analysis method *Intersect tables* ( ) can be found in the Tree Area, under the Analyses tab in the folder *Methods*, subfolder *Data manipulation*. The complete path to this method is:

http://genexplain-platform.com/bioumlweb/#de=analyses/Methods/Data%20manipulation/Intersect%20tables

This function allows for the identification of common rows between two tables. The intersection is performed according to IDs matching from "left" table to "right". The result is a single table that contains rows common for both input tables.



Left table	
Table	(select element)
Columns	(all columns)
Right table	
Table	(select element)
Columns	(all columns)
Merge columns with the same names	☑
Aggregator for numbers	extreme
Output table	(select element)

In the following, we will consider the input fields one by one:

Left Table – left (first table) for intersection.

Right Table – right (second table) for intersection.

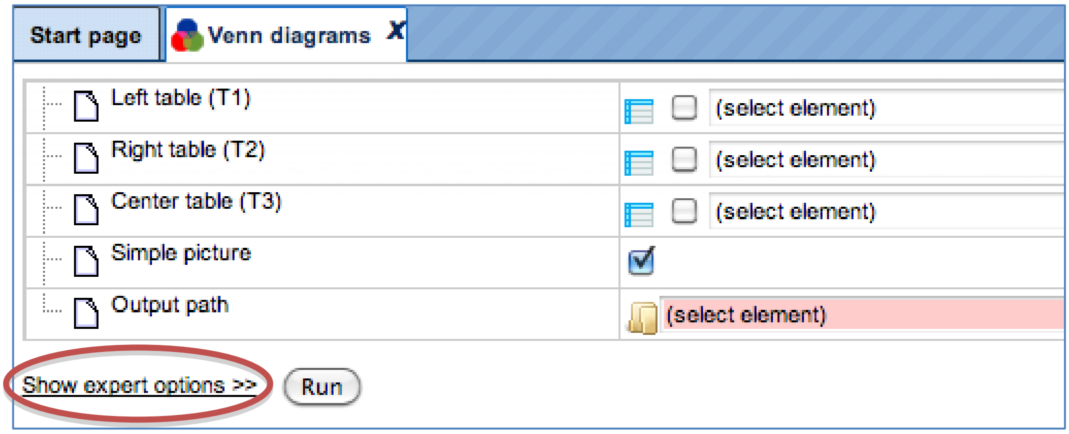
Aggregator for numbers (expert) – Function to be used for numerical columns when several rows are merged into a single one if merge columns option is selected.

Output table - Name of the table where the results will be saved. If a table with that name already exists it will be replaced.

16.1.5. Venn diagram

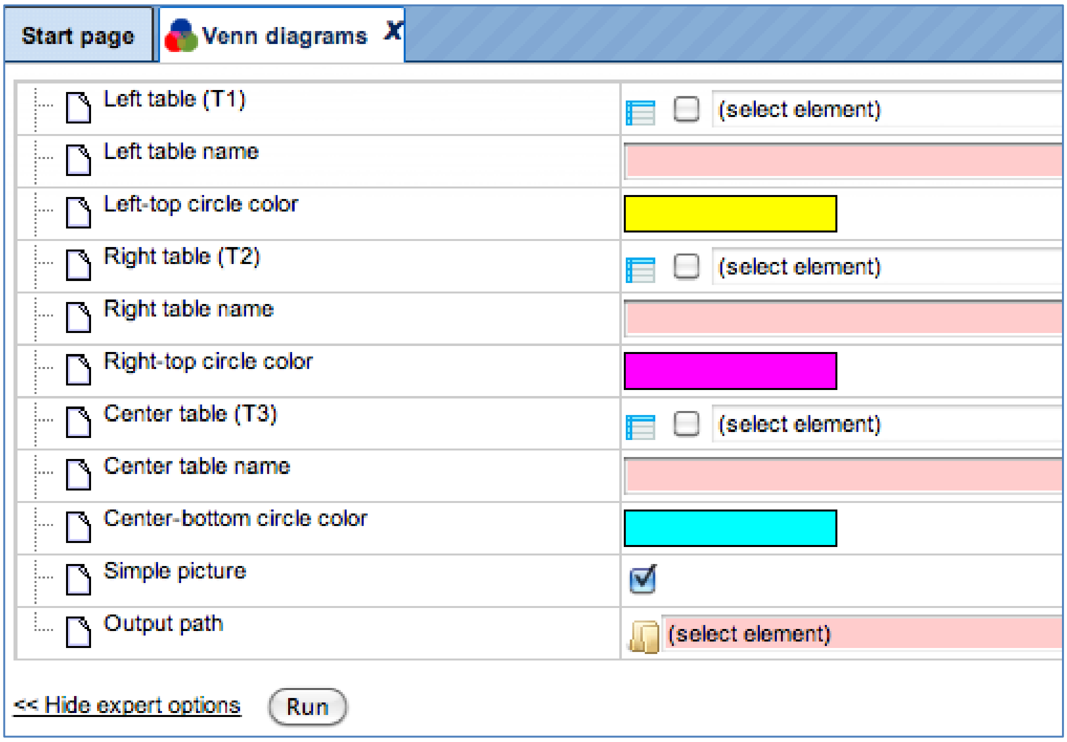
With this feature you can create VENN diagrams from the input tables as well as to get tables of common and unique genes according to the sections of the VENN diagrams. VENN diagrams are images that show all possible logical relations between the input tables. As input, two or three gene tables can be provided for which you wish to know common and unique genes. These input tables can be in any format (gene or protein IDs). The VENN diagram function can be found under the tab *Analyses*, in the folder Methods/Data manipulation/Venn diagrams ( ).

The initial form of this analysis looks as it is shown below:



Left table (T1)	☐ (select element)
Right table (T2)	☐ (select element)
Center table (T3)	☐ (select element)
Simple picture	☑
Output path	(select element)

When the expert options are opened, the form looks like:



Left table (T1)	☐ (select element)
Left table name	
Left-top circle color	🟨
Right table (T2)	☐ (select element)
Right table name	
Right-top circle color	🟪
Center table (T3)	☐ (select element)
Center table name	
Center-bottom circle color	🟦
Simple picture	☑
Output path	(select element)

To perform this analysis you can input two or three tables.

Left Table (T1), **Right Table(T2)** and **Center table (T3)**. You can drag and drop the input tables which you wish to add to the VENN diagram.

Left table name, Right table name, and **Center table name**. These are expert options. You can specify the names of the input tables as you want to see them in the output diagram, if you want them to be different from the names of the input tables. By default the original names of the input tables will be shown in the resulting diagram.

Left-top circle color, **Right-top circle color**, and **Center-bottom circle color**. These are expert options. In these fields you can specify the colors you wish to see in the diagram. The default colors are displayed in the input form. To change them just click on the colored boxes.

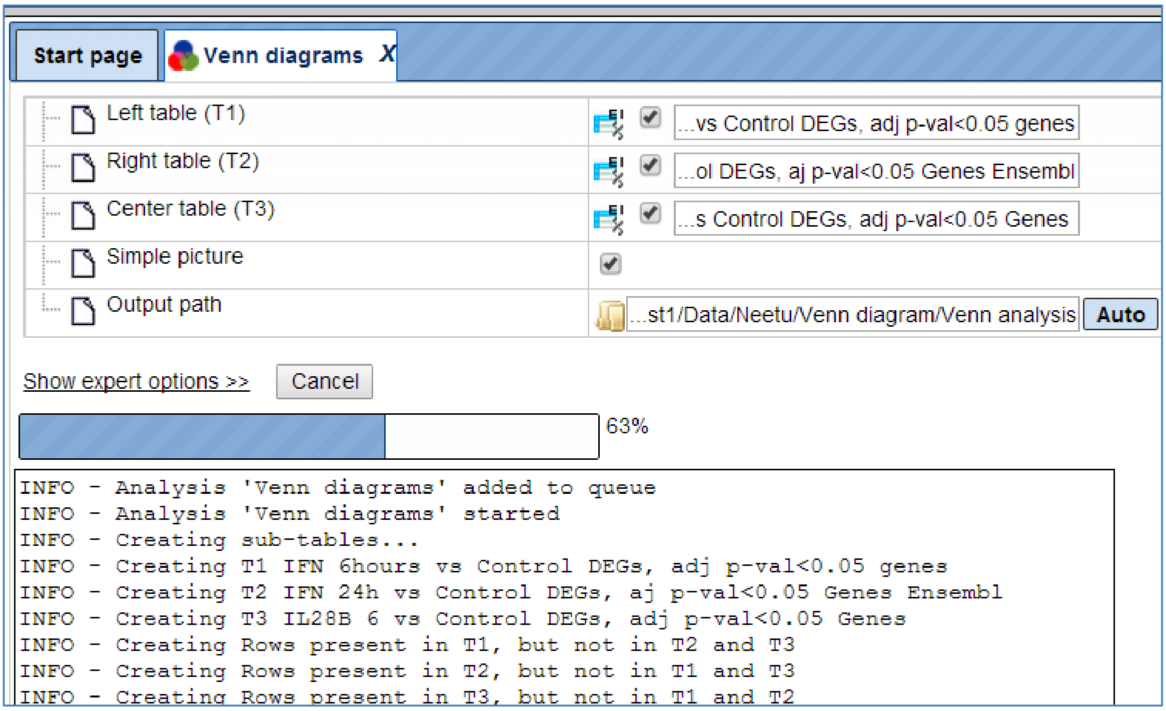
Simple picture. When this box is checked-, all three circles in the resulting diagram will have the same size, no matter whether the input tables are of the same size or not. By default this option is checked. When this box is unchecked, the size of the circles will be proportional to the size of the input tables in the resulting diagram.

Output path. Specify the output path. Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink field "select element" in the field **Output path**, and a new window will be opened, where you can select the location of the results folder and define its name.

Important: please define the output path within one of your Project folders.

Press [Run].

The analysis will start as shown below:



Left table (T1) ☑ ...vs Control DEGs, adj p-val<0.05 genes

Right table (T2) ☑ ...ol DEGs, aj p-val<0.05 Genes Ensembl

Center table (T3) ☑ ...s Control DEGs, adj p-val<0.05 Genes

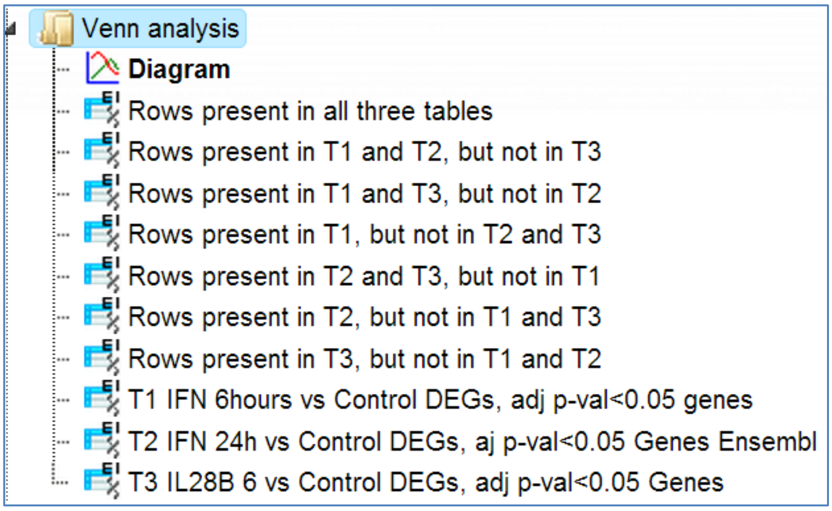
Simple picture ☑

Output path ...st1/Data/Neetu/Venn diagram/Venn analysis

```
INFO - Analysis 'Venn diagrams' added to queue
INFO - Analysis 'Venn diagrams' started
INFO - Creating sub-tables...
INFO - Creating T1 IFN 6hours vs Control DEGs, adj p-val<0.05 genes
INFO - Creating T2 IFN 24h vs Control DEGs, aj p-val<0.05 Genes Ensembl
INFO - Creating T3 IL28B 6 vs Control DEGs, adj p-val<0.05 Genes
INFO - Creating Rows present in T1, but not in T2 and T3
INFO - Creating Rows present in T2, but not in T1 and T3
INFO - Creating Rows present in T3, but not in T1 and T2
```

Wait till the analysis is completed.

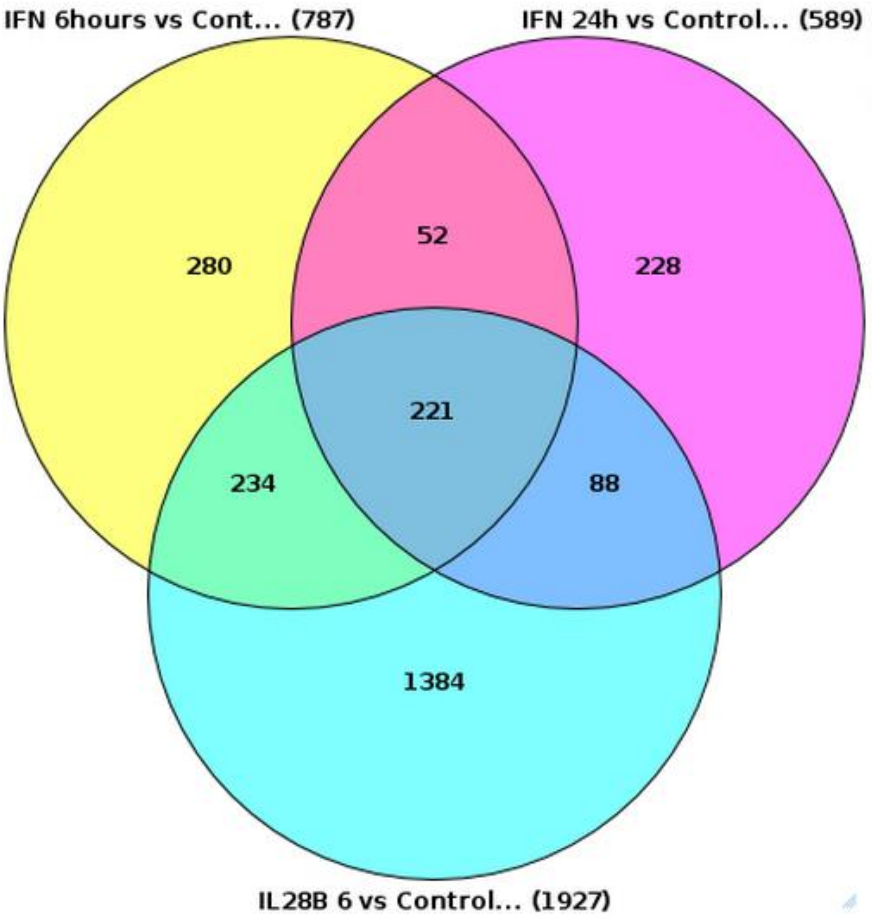
The output is a folder comprising several tables and one diagram as shown below:



Venn analysis
- Diagram
- Rows present in all three tables
- Rows present in T1 and T2, but not in T3
- Rows present in T1 and T3, but not in T2
- Rows present in T1, but not in T2 and T3
- Rows present in T2 and T3, but not in T1
- Rows present in T2, but not in T1 and T3
- Rows present in T3, but not in T1 and T2
- T1 IFN 6hours vs Control DEGs, adj p-val<0.05 genes
- T2 IFN 24h vs Control DEGs, aj p-val<0.05 Genes Ensembl
- T3 IL28B 6 vs Control DEGs, adj p-val<0.05 Genes

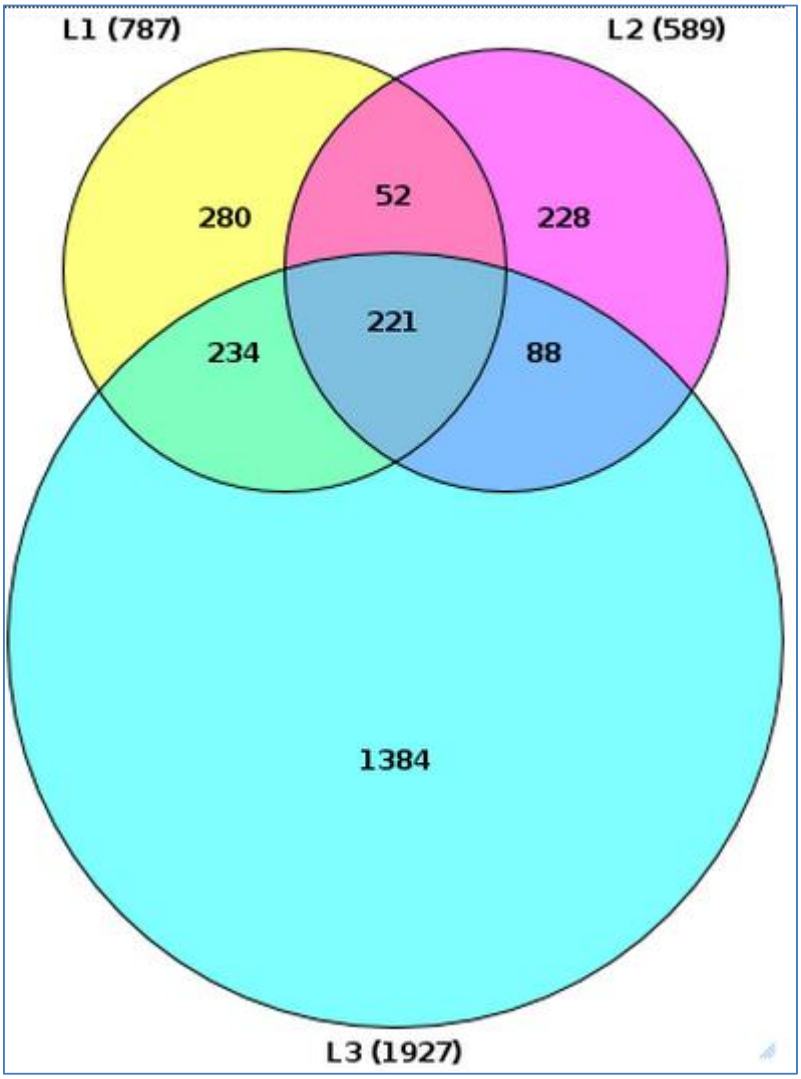
Results

Diagram opens automatically in the work space when the analysis is completed:



This picture can be exported in png, jpg, or bmp formats with the help of the Export (⇪) button on the top control panel.

For the same input files, if you specify the input table names under the expert options as L1, L2, and L3 respectively, and un-check the *Simple picture* box, the resulting diagram looks different, as shown below:



Please note that the size of the circles is proportional to the number of genes in the input tables.

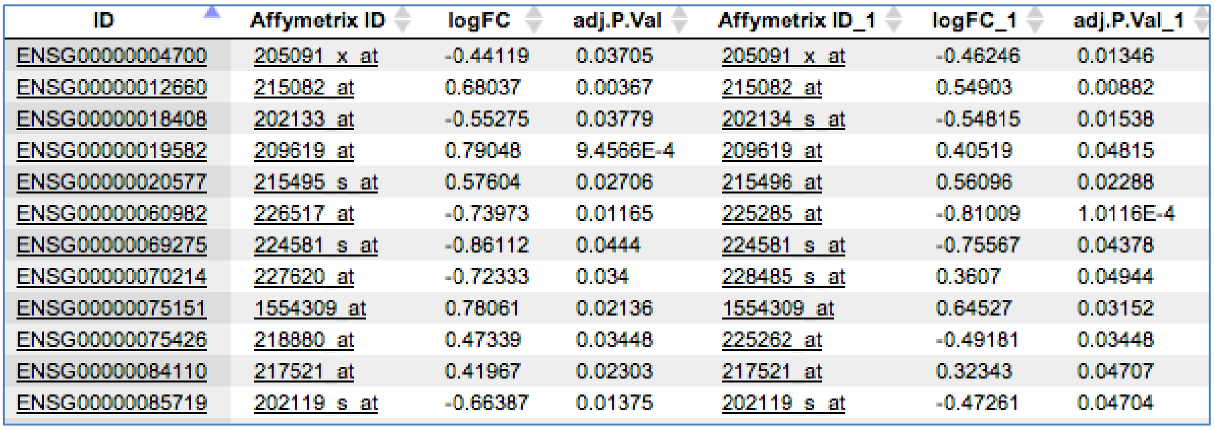
Along with the diagram, there are several tables in the output folder.

T1, T2 and *T3* are the three original input tables.

Each individual intersection is shown as a separate gene table; correspondingly, the following tables are displayed:

Rows present in all three tables,
Rows present in T1 and T2, but not in T3,
Rows present in T2 and T3, *but not in T1,*
Rows present in T1, but not in T2 and T3,
Rows present in T2 and T3, but not in T1,
Rows present in T2, but not in T1 and T3,
Rows present in T3, but not in T1 and T2

The tables containing rows present in all three or in two tables, also contain columns from three or two input tables, respectively. As an example, the table *Rows present in T2 and T3*, *but not in T1,* contains the columns from T2 and from T3 tables*,* as shown below. If the column names are the same in two input tables, "_1" is automatically added to the name for the second column.



ID	Affymetrix ID	logFC	adj.P.Val	Affymetrix ID_1	logFC_1	adj.P.Val_1
ENSG00000004700	205091_x_at	-0.44119	0.03705	205091_x_at	-0.46246	0.01346
ENSG00000012660	215082_at	0.68037	0.00367	215082_at	0.54903	0.00882
ENSG00000018408	202133_at	-0.55275	0.03779	202134_s_at	-0.54815	0.01538
ENSG00000019582	209619_at	0.79048	9.4566E-4	209619_at	0.40519	0.04815
ENSG00000020577	215495_s_at	0.57604	0.02706	215496_at	0.56096	0.02288
ENSG00000060982	226517_at	-0.73973	0.01165	225285_at	-0.81009	1.0116E-4
ENSG00000069275	224581_s_at	-0.86112	0.0444	224581_s_at	-0.75567	0.04378
ENSG00000070214	227620_at	-0.72333	0.034	228485_s_at	0.3607	0.04944
ENSG00000075151	1554309_at	0.78061	0.02136	1554309_at	0.64527	0.03152
ENSG00000075426	218880_at	0.47339	0.03448	225262_at	-0.49181	0.03448
ENSG00000084110	217521_at	0.41967	0.02303	217521_at	0.32343	0.04707
ENSG00000085719	202119_s_at	-0.66387	0.01375	202119_s_at	-0.47261	0.04704

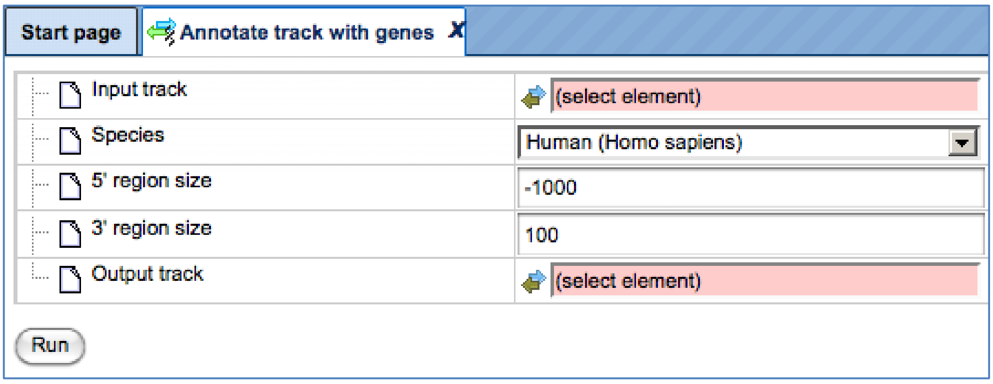
Each of the resulting tables can be used for all operations with tables and serve as input for a number of workflows, e.g. for functional classifications, promoter analysis, pathway analysis and more.

16.2. Operations with tracks

The folder *Data manipulation* contains several methods allowing useful operations with tracks, among other methods. This folder can be found on the *Analyses* tab. Track is a set of DNA fragments or intervals with obligatory information about their chromosomal location and absolute positions of the beginning and of the end. Optionally, any additional information about the fragments can be included. Tracks are very often available in BED format. Within the geneXplain platform, tracks in the tree area are shown as ( ). As for the basic operations with tracks, kindly refer to the previous descriptions (2.3.3).

16.2.1. Annotate track with genes

The method *Annotate track with genes* ( ) helps to add information about nearby located genes to each fragment. The input form is shown below:



Input track	(select element)
Species	Human (Homo sapiens)
5' region size	-1000
3' region size	100
Output track	(select element)

Input track. Specify the input track. You can drag & drop it from your project within the tree area. Alternatively, you may click on the pink field *select element* and a new window will be opened, where you can select the input track.

- *The input track used in this example can be found under location:*

- *http://genexplain-platform.com/bioumlweb/#de=data/Examples/Encode%20TFBS%20CEBPB%20in%20H1-hESC%20cells/Data/CEBP%20in%20H1-hESC%20cells%20YES*
- *This track contains 500 in vivo binding fragments for C/EBP transcription factor (Encode project).*

Species. Choose human, mouse or rat from the drop-down menu.

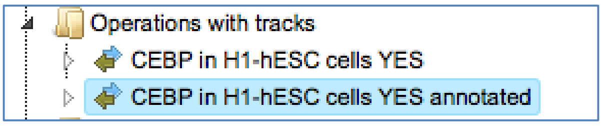
5′ region size and **3′ region size**. By default this method considers the following regions around Ensembl genes: 1000 bp in 5′ direction from TSS and 100 bp in 3′ direction from the last exon. The positions of each fragment on the input track are compared with the positions of the extended gene regions. Genes overlapping with an input fragment are considered for annotation of this fragment.

Output track. Specify the path and name to store the output track.

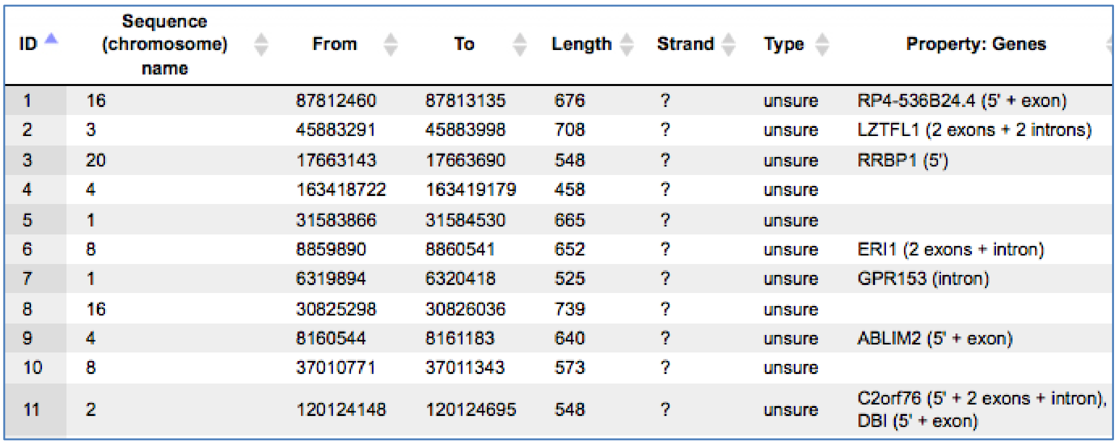
Having filled in the input form, launch the analysis with the [Run] button. Wait till the analysis is completed.

Results. The resulting track is automatically opened in genome browser in the work space.

In the tree area, at the location specified in the input form, you can find the resulting track, highlighted in blue on the screenshot below:


Operations with tracks
CEBP in H1-hESC cells YES
CEBP in H1-hESC cells YES annotated

When opened as a table, it looks like this:



ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: Genes
1	16	87812460	87813135	676	?	unsure	RP4-536B24.4 (5' + exon)
2	3	45883291	45883998	708	?	unsure	LZTFL1 (2 exons + 2 introns)
3	20	17663143	17663690	548	?	unsure	RRBP1 (5')
4	4	163418722	163419179	458	?	unsure	
5	1	31583866	31584530	665	?	unsure	
6	8	8859890	8860541	652	?	unsure	ERI1 (2 exons + intron)
7	1	6319894	6320418	525	?	unsure	GPR153 (intron)
8	16	30825298	30826036	739	?	unsure	
9	4	8160544	8161183	640	?	unsure	ABLIM2 (5' + exon)
10	8	37010771	37011343	573	?	unsure	
11	2	120124148	120124695	548	?	unsure	C2orf76 (5' + 2 exons + intron), DBI (5' + exon)

All columns of the input track are present, and one column is added, called **Property:Genes**. This newly added column is a result of an annotation of the input track with genes, and for each fragment it contains gene symbols of overlapping genes. As you can see, some of the fragments are not overlapping with any genes, and some of the fragments may be overlapping with two or even more genes. It depends on the particular fragments, their length and location as well as on the length of the gene-bound extension regions specified in the input form.

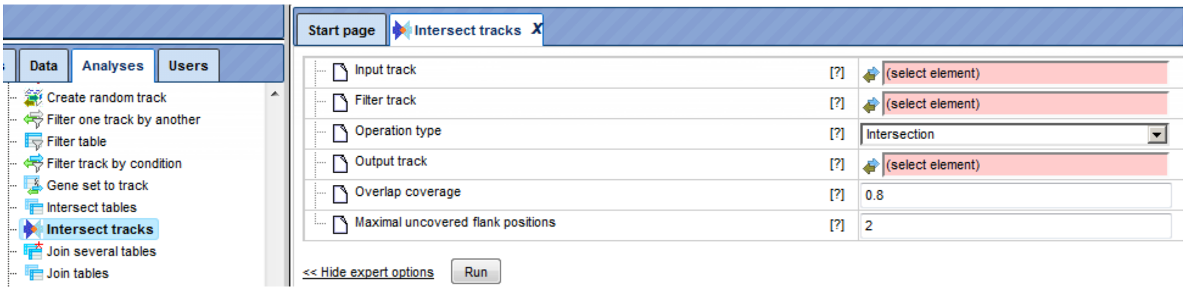
Next to each gene symbol there are gene regions specified, for example *ERI1 (2 exons + intron)*. This means, a particular fragment overlaps two exons and one intron of the ERI1 gene.

Tip If you would like to annotate overlapping genes for all fragments in the input track, you might be interested to increase the gene-bound extension regions in the input form, and run the analysis again.

16.2.2. Intersect tracks

Track intersection provides two types of operations whose results are either the intersection itself or the difference of two tracks. In the first case, the output track consists of intervals two tracks have in common (which overlap). In the second, the output contains those intervals uniquely found in the input track.

This analysis can be used, for instance, to filter predicted binding sites for conserved regions.



Input track [?] (select element)

Filter track [?] (select element)

Operation type [?] Intersection

Output track [?] (select element)

Overlap coverage [?] 0.8

Maximal uncovered flank positions [?] 2

The parameters can be described as follows.

Input track: The input track contains the intervals which will be available in or omitted from the output track if they overlap with intervals of the filter track.

Filter track: The filter track contains the intervals against which input intervals are tested for overlap.

Operation type: Here one can select the desired input intervals, intersection or difference.

Output track: The output track will contain the input intervals contained in the intersection of difference set.

Overlap coverage: The overlap coverage is the relative proportion of an input interval that needs to overlap with a filter interval.

Maximal uncovered flank positions: This parameter limits the number uncovered positions (sometimes called "overhanging ends"). Note that this limit is applied to each side of an input interval, not to the total number of uncovered end positions.

16.2.3. Gene set to track

The method *Gene set to track* ( ) aims at creating a track corresponding to any table with Ensembl gene IDs. As with the fragments of the output track, this method takes the gene regions around TSS (transcription start sites). It is a useful method to create a track of gene promoters or upstream regions for any input gene table. The input form is shown below:

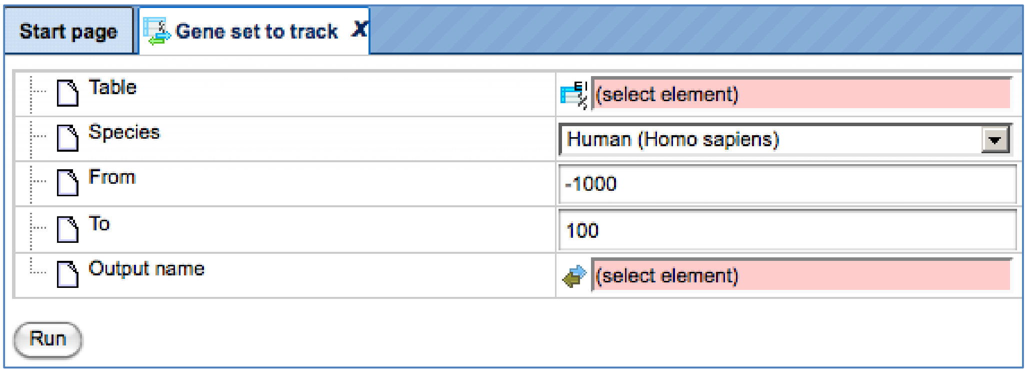
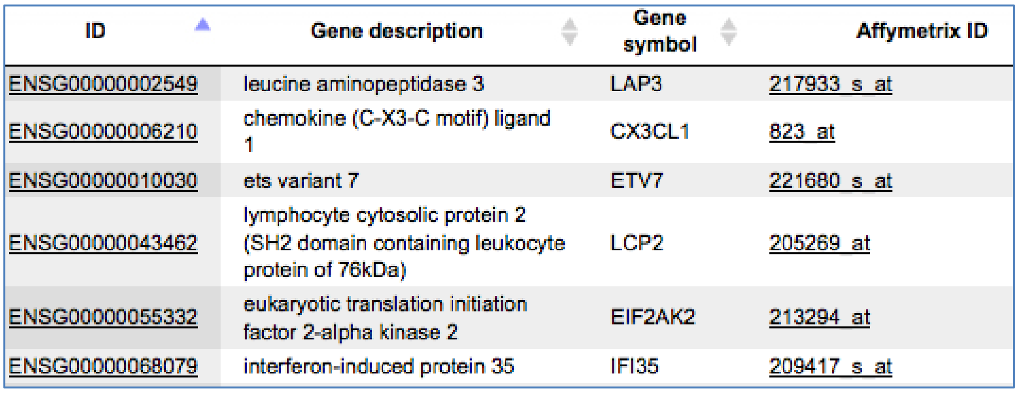


Table	(select element)
Species	Human (Homo sapiens)
From	-1000
To	100
Output name	(select element)

Input table. Specify the input table with Ensembl gene IDs. If your table has different IDs, you need to convert it first. Details on how to convert table identifiers are given in the *Section 11.3.3*. You can drag & drop the table from your project within the tree area. Alternatively, you may click on the pink field *select element* and a new window will be opened, where you can select the table. Here, the following table is taken as input.



ID	Gene description	Gene symbol	Affymetrix ID
ENSG00000002549	leucine aminopeptidase 3	LAP3	217933_s_at
ENSG00000006210	chemokine (C-X3-C motif) ligand 1	CX3CL1	823_at
ENSG00000010030	ets variant 7	ETV7	221680_s_at
ENSG00000043462	lymphocyte cytosolic protein 2 (SH2 domain containing leukocyte protein of 76kDa)	LCP2	205269_at
ENSG00000055332	eukaryotic translation initiation factor 2-alpha kinase 2	EIF2AK2	213294_at
ENSG00000068079	interferon-induced protein 35	IFI35	209417_s_at

You can see Ensembl gene IDs in the column **ID**. Such a table may contain any number of additional columns. Here, three additional columns are present, **Gene description**, **Gene symbol** and **Affymetrix ID**.

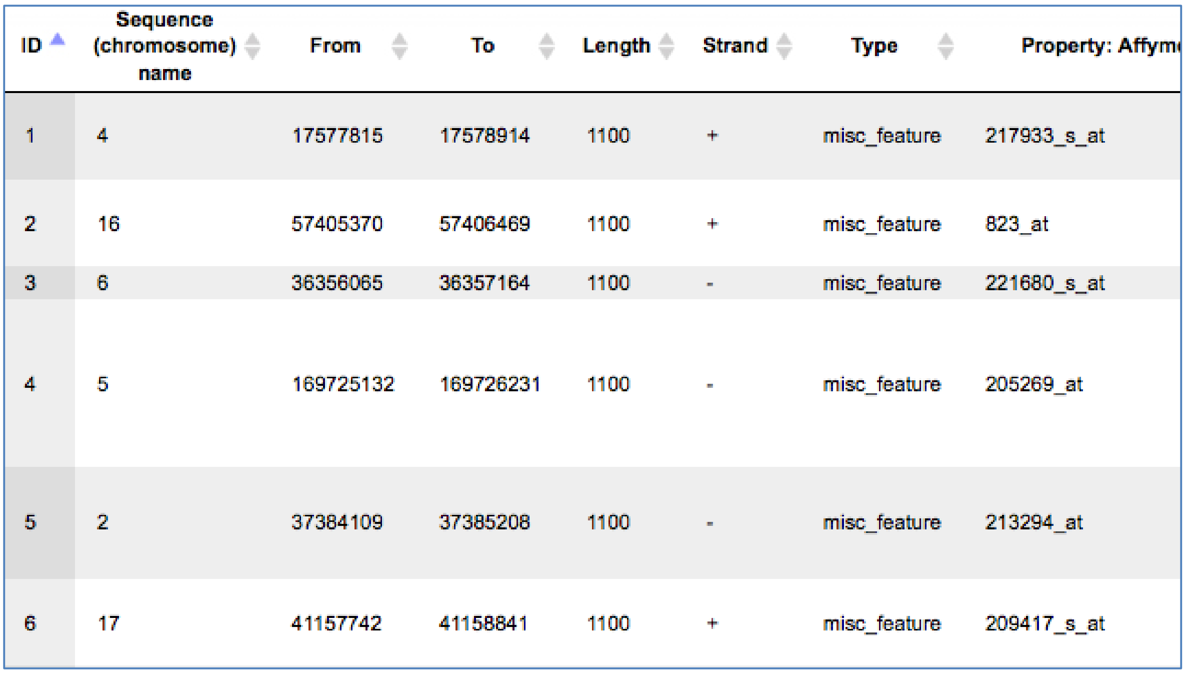
Species. After input of the table, the species (human, mouse or rat) is adjusted automatically. Verify the species shown in the species field.

From and **To**. By default this method considers the following regions around the TSS of the input genes: 1000 bp in 5' direction and 100 bp in 3' direction.

Output track. Specify the path and name to store the output track.

Having filled in the input form, launch the analysis with the [Run] button. Wait till the analysis is completed. The resulting track is automatically opened in genome browser in your work space.

The output track when opened as a table, is shown below:

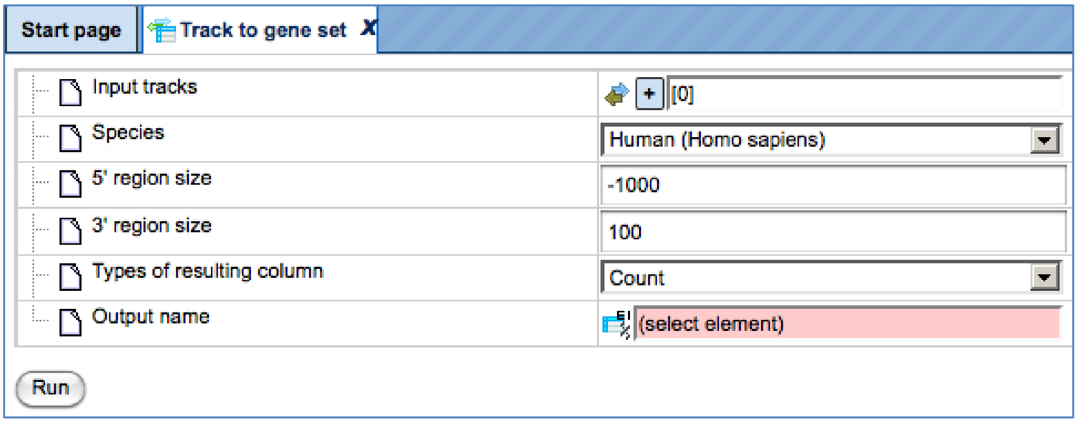


ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: Affym
1	4	17577815	17578914	1100	+	misc_feature	217933_s_at
2	16	57405370	57406469	1100	+	misc_feature	823_at
3	6	36356065	36357164	1100	-	misc_feature	221680_s_at
4	5	169725132	169726231	1100	-	misc_feature	205269_at
5	2	37384109	37385208	1100	-	misc_feature	213294_at
6	17	41157742	41158841	1100	+	misc_feature	209417_s_at

This table contains exactly the same number of the fragments (rows) as the number of Ensembl genes in the input table. There are columns for **chromosome**, positions **From** and **To**, **Length**, **Strand**, and **Type**. The type of the fragments after this conversion is automatically assigned as *misc_feature*. Other columns present in the input table are all added on the right side of this table, e.g. here **Affymetrix ID** column.

16.2.4. Track to gene set

The method *Track to gene set* ( ) aims at identifying genes located close to the fragments in the input track. Genes overlapping with at least one input fragment are considered resulting target genes. The input form is shown below:



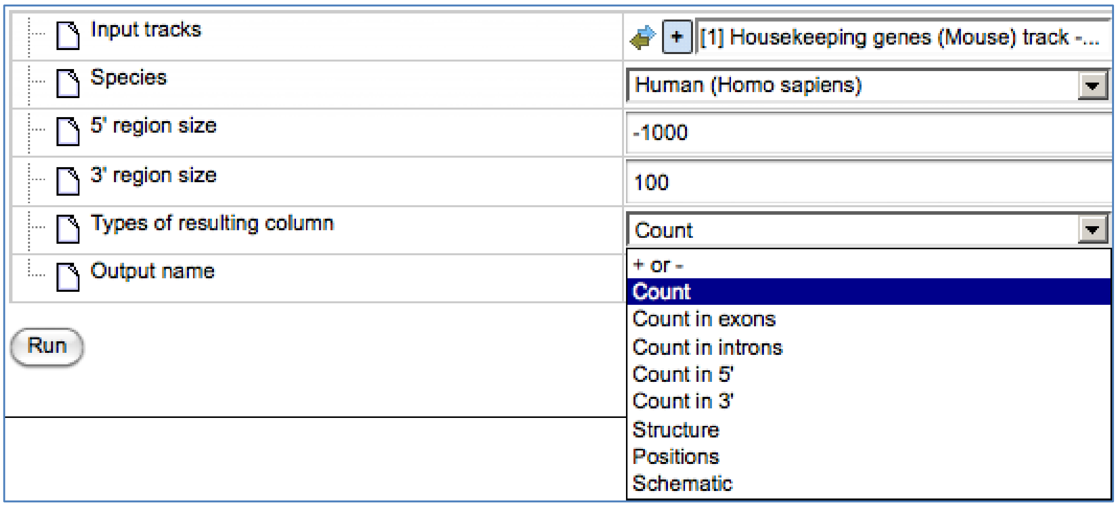
Input tracks	[0]
Species	Human (Homo sapiens)
5' region size	-1000
3' region size	100
Types of resulting column	Count
Output name	(select element)

Input track. Specify input track. You can drag & drop it from your project within the tree area. Alternatively, you may click on the pink field *select element* and a new window will be opened, where you can select the input track. With the [track Plus] button (+) you can add another track to have several tracks analyzed together.

Species. Choose human, mouse or rat species from the drop-down menu.

5' region size and **3' region size**. By default this method considers the following regions around Ensembl genes: 1000 bp in 5' direction from the TSS and 100 bp in 3' direction from the last exon. The positions of each fragment on the input track are compared with the positions of the extended gene regions. Genes overlapping with at least one input fragment are considered resulting target genes.

Types of resulting columns. This analysis specifies the overlap between the extended gene regions and the fragments; such a specification can be given in several different waysanda desirable way of representation can be specified in this field. Let's consider the options available in the drop-down list:



Input tracks	+ [1] Housekeeping genes (Mouse) track -...
Species	Human (Homo sapiens)
5' region size	-1000
3' region size	100
Types of resulting column	Count
Output name	

+ or -. This option shows the presence or absence of overlap between any of the input tracks and the specified gene regions.

Count. Returns the number of fragments overlapping with each gene.

Count in exons, Count in introns, Count in 5', Count in 3'. When any of these options is chosen, the number of fragments overlapping with the corresponding gene regions is shown for each gene in the resulting table.

Structure. This option returns the names of gene regions overlapping with the fragment(s).

Positions. With this option, you can see the position of the 5' end of the fragment relative to the TSS of the overlapping gene.

Schematic. The gene structure is shown schematically with exons and introns, and the overlapping fragments are displayed.

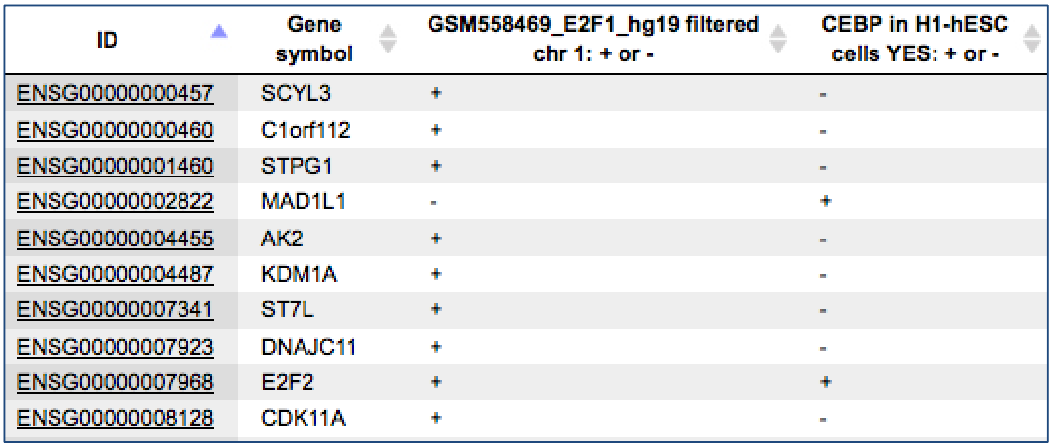
The resulting tables with all available types of representation are shown below.

Output name. Specify the path and name to store the output table with Ensembl genes.

Having filled in the input form, launch the analysis with the [Run] button. Wait till the analysis is completed. The resulting table is opened automatically in the work space. Let’s consider different variants of the resulting gene tables depending on the selected option in the input field **Types of resulting column**.

+ or -.

This option is especially useful when two or more tracks are added as input. In the resulting table (below) you can see the columns corresponding to each of the input tracks, here two columns.



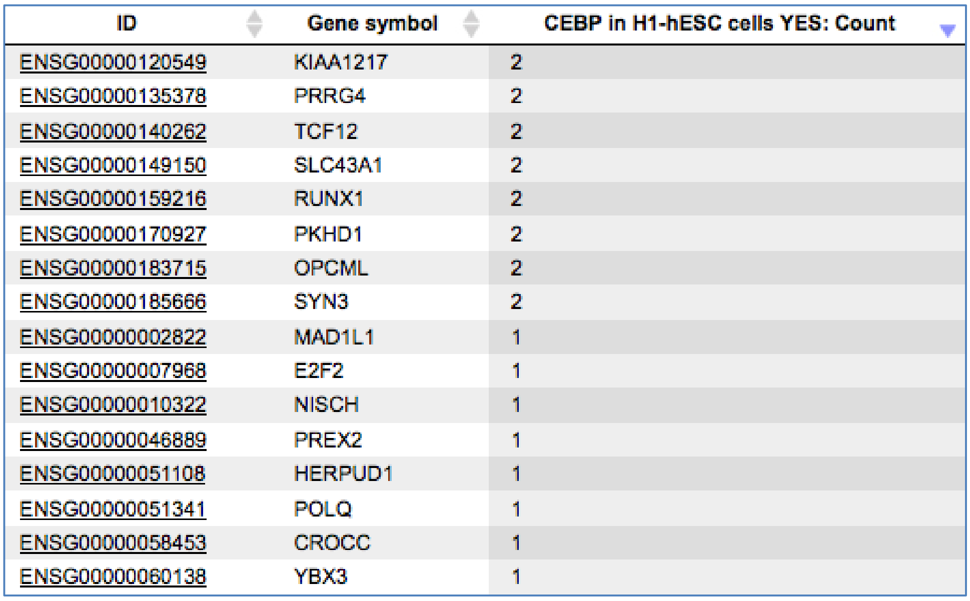
ID	Gene symbol	GSM558469_E2F1_hg19 filtered chr 1: + or -	CEBP in H1-hESC cells YES: + or -
ENSG00000000457	SCYL3	+	-
ENSG00000000460	C1orf112	+	-
ENSG00000001460	STPG1	+	-
ENSG00000002822	MAD1L1	-	+
ENSG00000004455	AK2	+	-
ENSG00000004487	KDM1A	+	-
ENSG00000007341	ST7L	+	-
ENSG00000007923	DNAJC11	+	-
ENSG00000007968	E2F2	+	+
ENSG00000008128	CDK11A	+	-

Each row corresponds to one gene overlapping with at least one fragment in at least one of the input tracks. For example, in the table above, the gene SCYL3 is overlapping with at least one fragment in the track *GSM558469_E2F1_hg19 filtered chr 1*, and is not overlapping with any fragment in the track *CEBP in H1-hESC cells YES*.

Tip If you would like to find overlapping genes for all fragments in the input track(s), you might be interested in increasing the gene-bound extension regions on the input form, and run the analysis again.

To learn more details, e.g. how many fragments are overlapping with gene regions and with exactly which parts of particular genes, you might be interested to choose other types of the output, as shown below.

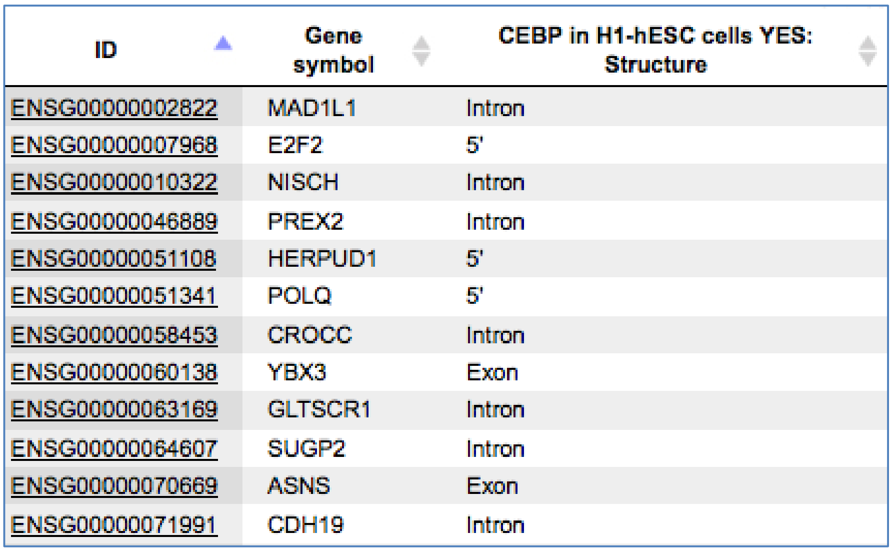
Count



ID	Gene symbol	CEBP in H1-hESC cells YES: Count
ENSG00000120549	KIAA1217	2
ENSG00000135378	PRRG4	2
ENSG00000140262	TCF12	2
ENSG00000149150	SLC43A1	2
ENSG00000159216	RUNX1	2
ENSG00000170927	PKHD1	2
ENSG00000183715	OPCML	2
ENSG00000185666	SYN3	2
ENSG00000002822	MAD1L1	1
ENSG00000007968	E2F2	1
ENSG00000010322	NISCH	1
ENSG00000046889	PREX2	1
ENSG00000051108	HERPUD1	1
ENSG00000051341	POLQ	1
ENSG00000058453	CROCC	1
ENSG00000060138	YBX3	1

For each gene, a gene symbol is given, and in the column **Count** you can see a number of the fragments overlapping with each gene. Here, the sorting is done by this column.

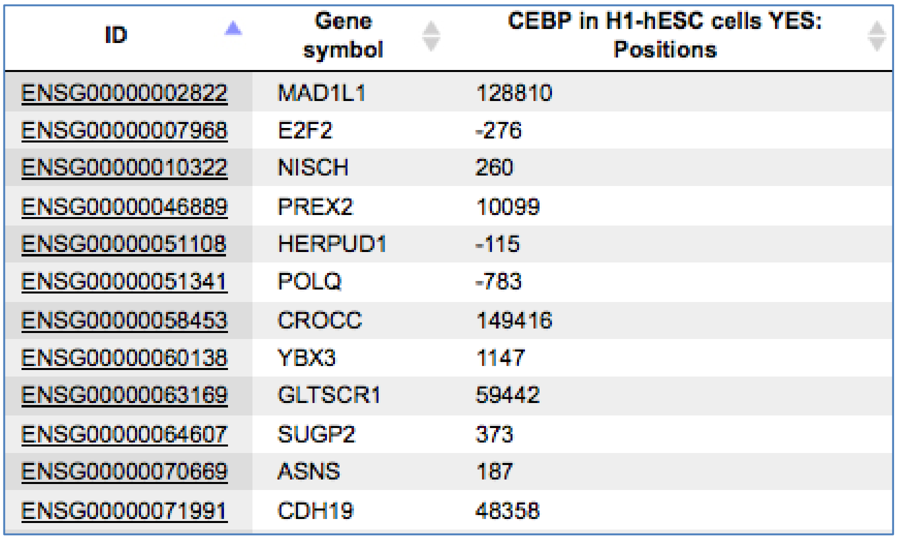
Structure



ID	Gene symbol	CEBP in H1-hESC cells YES: Structure
ENSG00000002822	MAD1L1	Intron
ENSG00000007968	E2F2	5'
ENSG00000010322	NISCH	Intron
ENSG00000046889	PREX2	Intron
ENSG00000051108	HERPUD1	5'
ENSG00000051341	POLQ	5'
ENSG00000058453	CROCC	Intron
ENSG00000060138	YBX3	Exon
ENSG00000063169	GLTSCR1	Intron
ENSG00000064607	SUGP2	Intron
ENSG00000070669	ASNS	Exon
ENSG00000071991	CDH19	Intron

The column **Structure** contains the names of gene regions overlapping with the fragment(s). The table can be sorted by this column to get all genes where the fragments overlap the gene regions in focus.

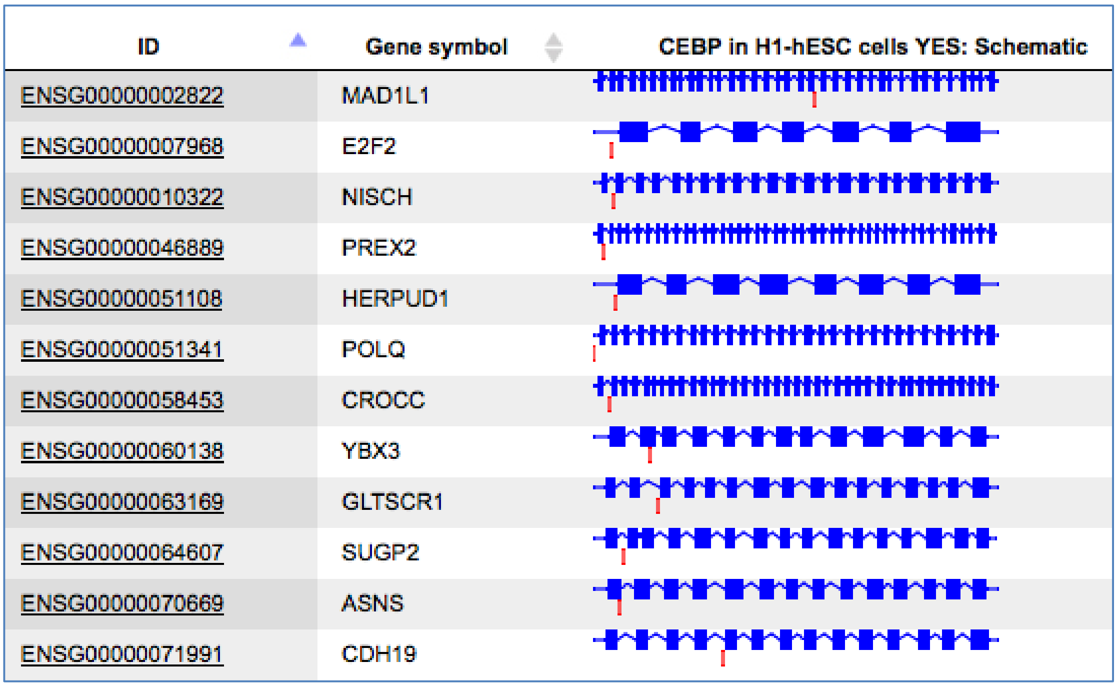
Positions



ID	Gene symbol	CEBP in H1-hESC cells YES: Positions
ENSG00000002822	MAD1L1	128810
ENSG00000007968	E2F2	-276
ENSG00000010322	NISCH	260
ENSG00000046889	PREX2	10099
ENSG00000051108	HERPUD1	-115
ENSG00000051341	POLQ	-783
ENSG00000058453	CROCC	149416
ENSG00000060138	YBX3	1147
ENSG00000063169	GLTSCR1	59442
ENSG00000064607	SUGP2	373
ENSG00000070669	ASNS	187
ENSG00000071991	CDH19	48358

The column **Positions** presents positions at the 5′ end of the fragment overlapping with this gene. Positions are shown relative to the TSS of the gene in each row.

Schematic



ID	Gene symbol	CEBP in H1-hESC cells YES: Schematic
ENSG00000002822	MAD1L1	
ENSG00000007968	E2F2	
ENSG00000010322	NISCH	
ENSG00000046889	PREX2	
ENSG00000051108	HERPUD1	
ENSG00000051341	POLQ	
ENSG00000058453	CROCC	
ENSG00000060138	YBX3	
ENSG00000063169	GLTSCR1	
ENSG00000064607	SUGP2	
ENSG00000070669	ASNS	
ENSG00000071991	CDH19	

The column **Schematic** presents a gene schema with depicted as blue boxes. Introns, 5′ regions and 3′ regions are represented by blue lines, and the fragments on the input track by red short vertical lines. The length of the introns is calculated in logarithmic scale relative to the length of the exons, to allow for a reasonable schematic representation.

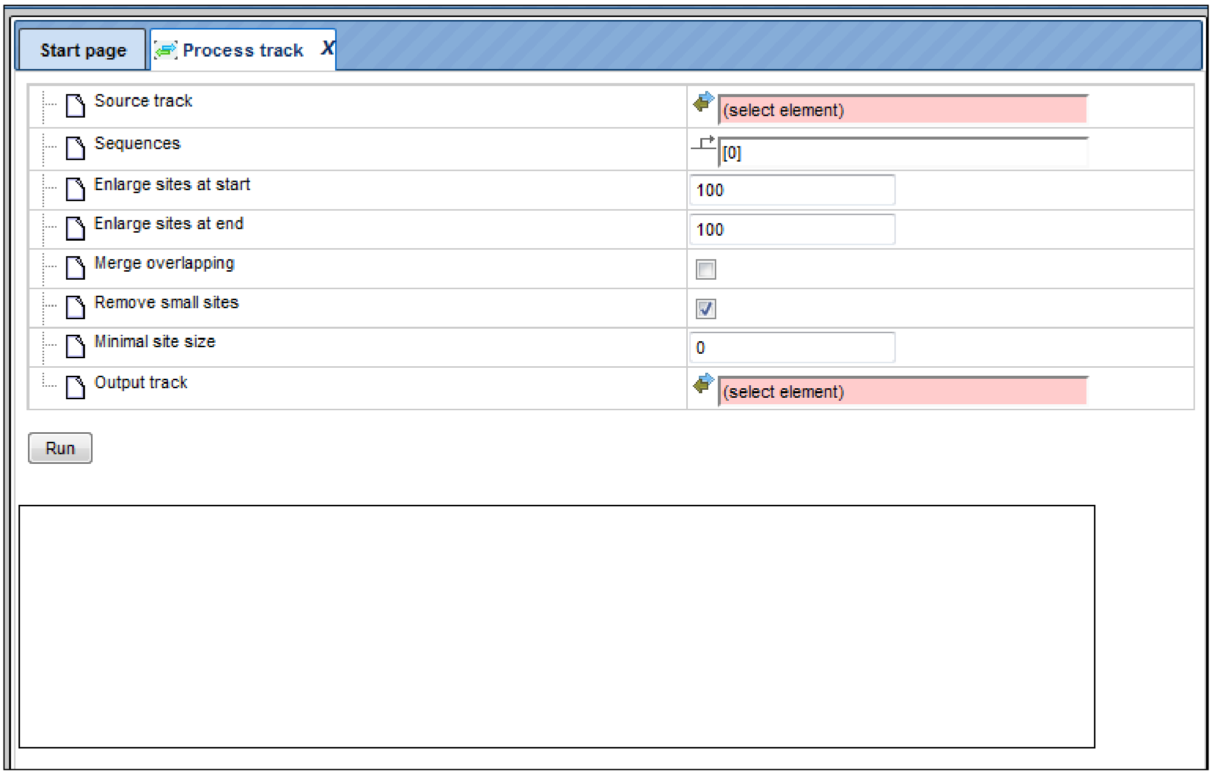
All Ensembl gene IDs are hyperlinked, and upon click on them the corresponding Ensembl gene page is opened in a new tab of the browser.

16.2.5. Process track with sites

In general, a track is a set of intervals where positions are specified that we can map on a chromosome. These track files can be visualized in a genome browser and can be used as input for various site analysis functions.

The geneXplain platform provides you with an option to modify these track files. "Process track with Sites" is a function which enables the user to enlarge/shrink sites on the track, merge overlapping sites or remove too short sites. For example an already saved track in the repository can be processed by adding sequences from Ensembl or some other database.

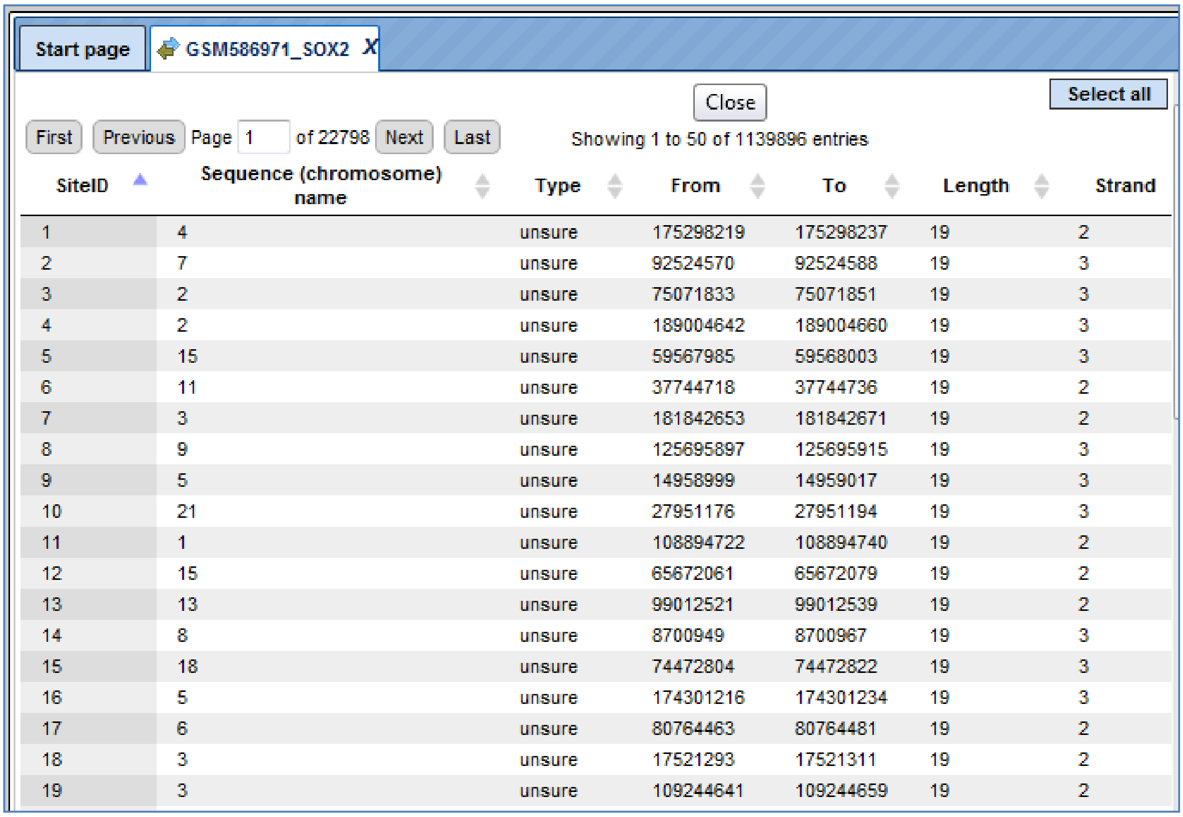
The initial form of this analysis looks as shown below:



Source track	(select element)
Sequences	[0]
Enlarge sites at start	100
Enlarge sites at end	100
Merge overlapping	□
Remove small sites	☑
Minimal site size	0
Output track	(select element)

Source track: Track you want to process
Sequences: Sequences to use
Enlarge sites at start: Use positive numbers to enlarge and negative to shrink
Enlarge sites at end: Use positive numbers to enlarge and negative to shrink
Merge overlapping: Checking this box merges overlapping sites into a single site. Site annotations will be lost!
Remove small sites: If checked, sites smaller then **Minimal site size** will be removed, otherwise they will be expanded to **Minimal site size**
Minimal site size: Sites shorter than the specified size will be removed from output
Output track: You should specify the path for the processed track here.

An example source track file saved in the repository to which you want to add sequences may look like this:

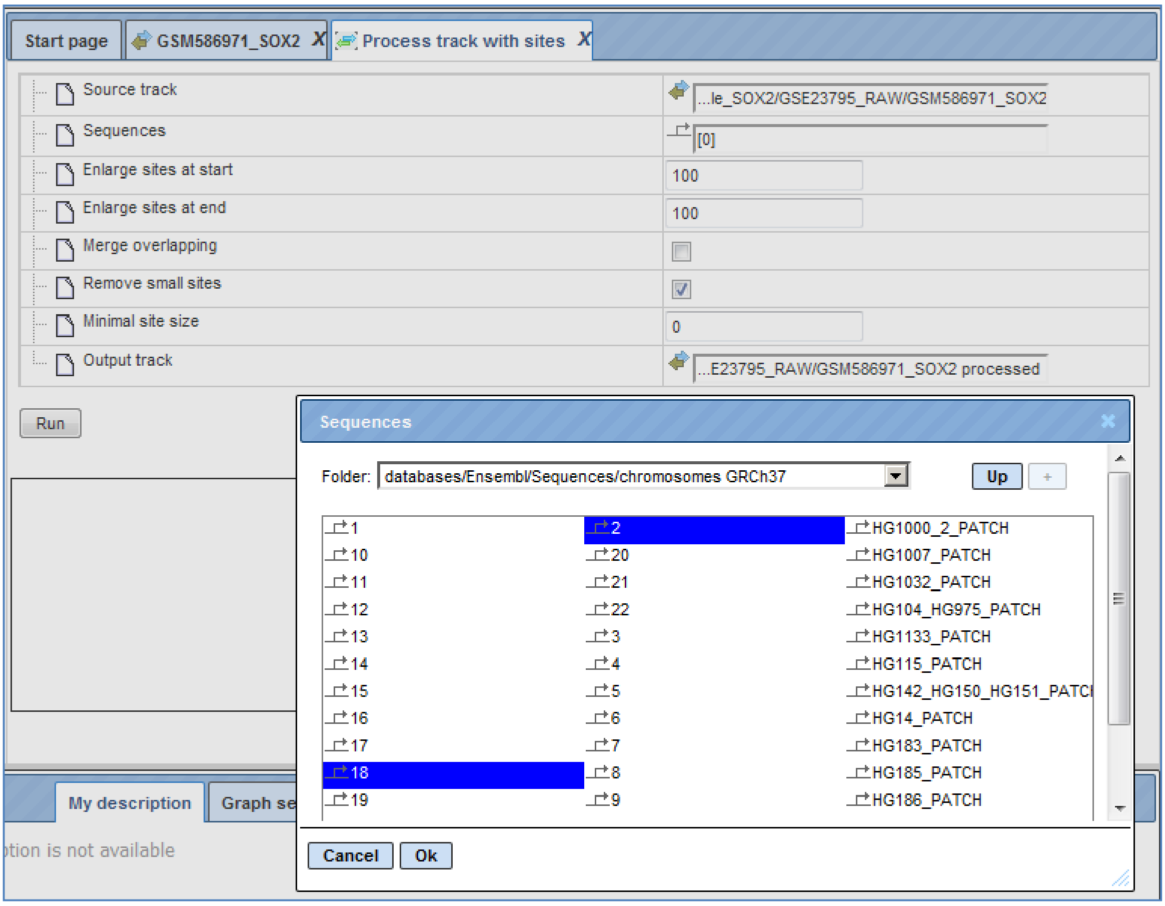


Showing 1 to 50 of 1139896 entries

SiteID	Sequence (chromosome) name	Type	From	To	Length	Strand
1	4	unsure	175298219	175298237	19	2
2	7	unsure	92524570	92524588	19	3
3	2	unsure	75071833	75071851	19	3
4	2	unsure	189004642	189004660	19	3
5	15	unsure	59567985	59568003	19	3
6	11	unsure	37744718	37744736	19	2
7	3	unsure	181842653	181842671	19	2
8	9	unsure	125695897	125695915	19	3
9	5	unsure	14958999	14959017	19	3
10	21	unsure	27951176	27951194	19	3
11	1	unsure	108894722	108894740	19	2
12	15	unsure	65672061	65672079	19	2
13	13	unsure	99012521	99012539	19	2
14	8	unsure	8700949	8700967	19	3
15	18	unsure	74472804	74472822	19	3
16	5	unsure	174301216	174301234	19	3
17	6	unsure	80764463	80764481	19	2
18	3	unsure	17521293	17521311	19	2
19	3	unsure	109244641	109244659	19	2

The track file shown provides you with the positions of promoter areas selected for analysis, as shown in columns **From** and **To**. The column **Strand** shows the strand of the chromosome where these promoters are located, where 1 means strand not applicable, 2 means forward strand, 3 means reverse strand, 4 means both strands. This file can be dragged and dropped on a particular chromosome opened in the genome browser to visualize its positions (see Section 16.2.3).

This Source track file can be selected as an input to “Process track with Sites”. The sequences we want to map are selected from the Ensembl database as shown below:



Source track	...le_SOX2/GSE23795_RAW/GSM586971_SOX2
Sequences	[0]
Enlarge sites at start	100
Enlarge sites at end	100
Merge overlapping	□
Remove small sites	☑
Minimal site size	0
Output track	...E23795_RAW/GSM586971_SOX2 processed

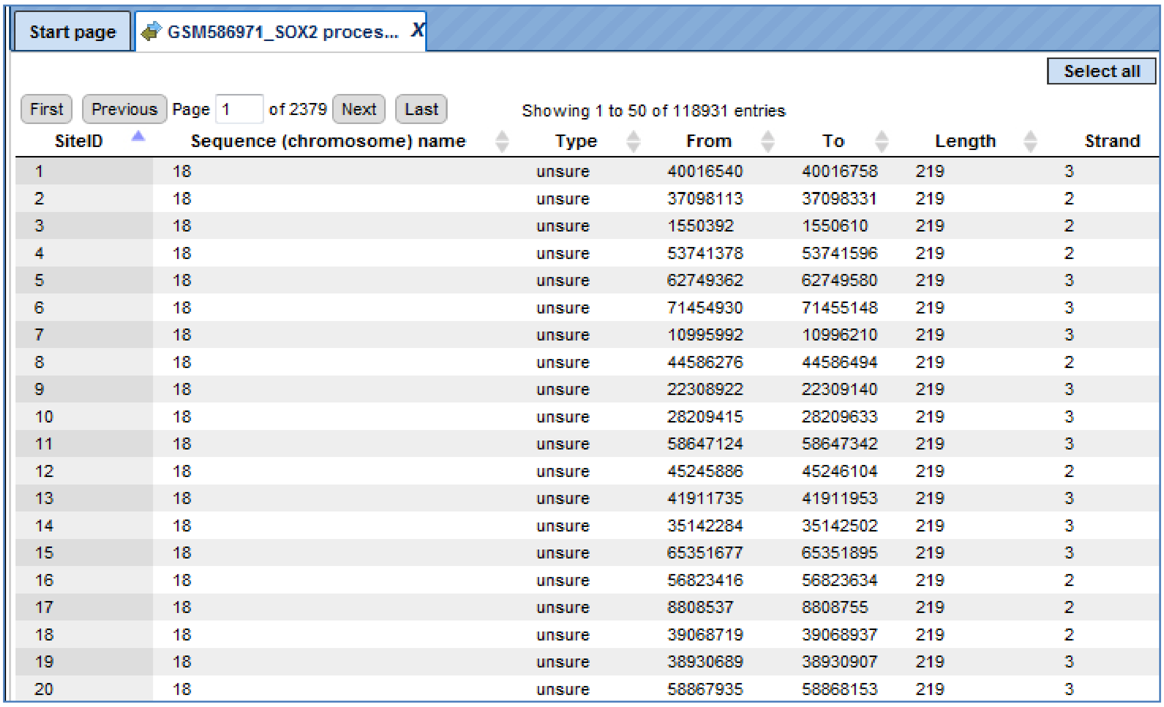
Sequences

Folder: databases/Ensembl/Sequences/chromosomes GRCh37

1	2	HG1000_2_PATCH
10	20	HG1007_PATCH
11	21	HG1032_PATCH
12	22	HG104_HG975_PATCH
13	3	HG1133_PATCH
14	4	HG115_PATCH
15	5	HG142_HG150_HG151_PATC
16	6	HG14_PATCH
17	7	HG183_PATCH
18	8	HG185_PATCH
19	9	HG186_PATCH

Using default conditions for the other parameters you can now press [Run].

The output track looks like shown below:

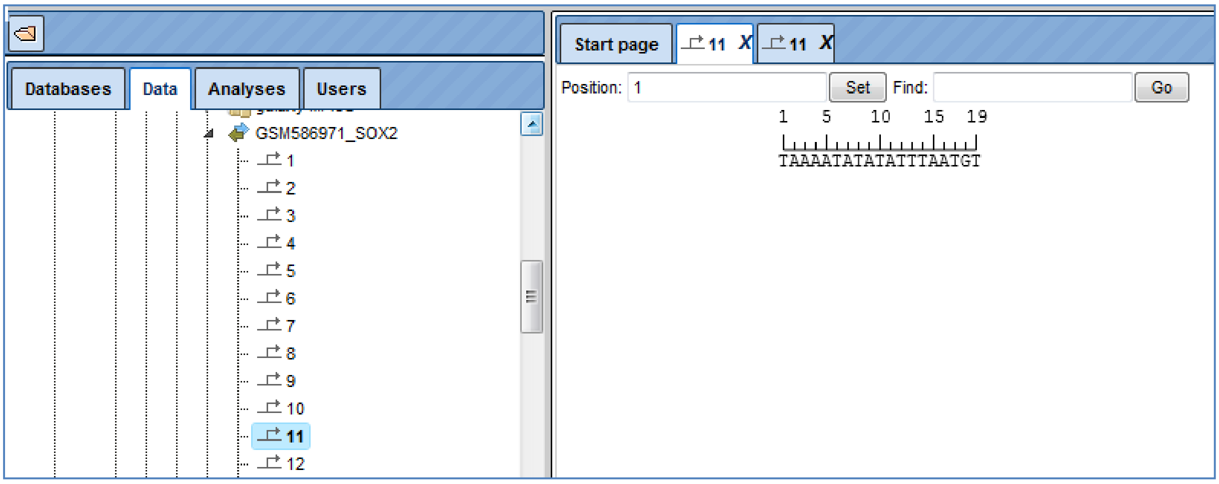


Page 1 of 2379

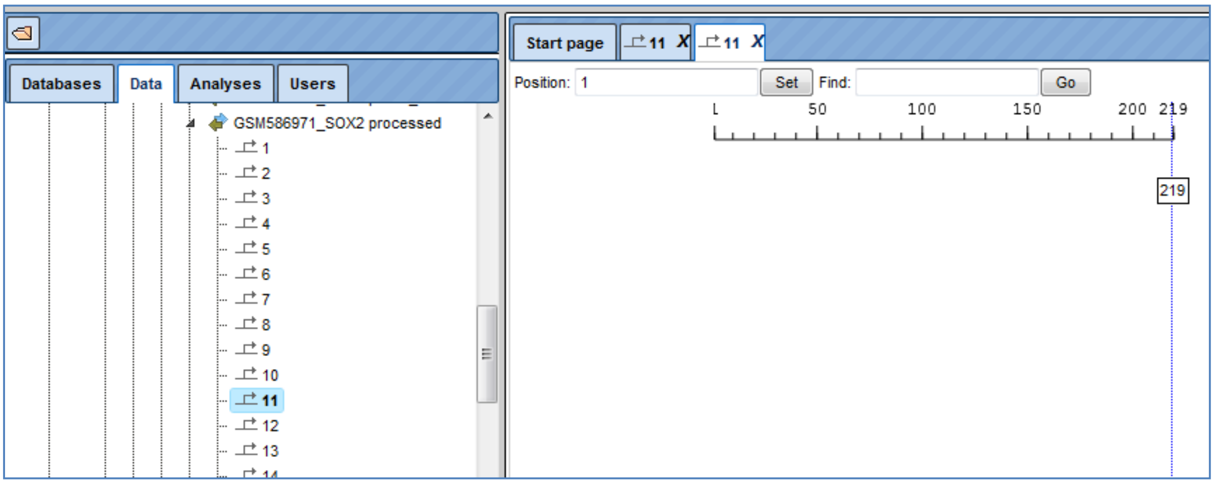
Showing 1 to 50 of 118931 entries

SiteID	Sequence (chromosome) name	Type	From	To	Length	Strand
1	18	unsure	40016540	40016758	219	3
2	18	unsure	37098113	37098331	219	2
3	18	unsure	1550392	1550610	219	2
4	18	unsure	53741378	53741596	219	2
5	18	unsure	62749362	62749580	219	3
6	18	unsure	71454930	71455148	219	3
7	18	unsure	10995992	10996210	219	3
8	18	unsure	44586276	44586494	219	2
9	18	unsure	22308922	22309140	219	3
10	18	unsure	28209415	28209633	219	3
11	18	unsure	58647124	58647342	219	3
12	18	unsure	45245886	45246104	219	2
13	18	unsure	41911735	41911953	219	3
14	18	unsure	35142284	35142502	219	3
15	18	unsure	65351677	65351895	219	3
16	18	unsure	56823416	56823634	219	2
17	18	unsure	8808537	8808755	219	2
18	18	unsure	39068719	39068937	219	2
19	18	unsure	38930689	38930907	219	3
20	18	unsure	58867935	58868153	219	3

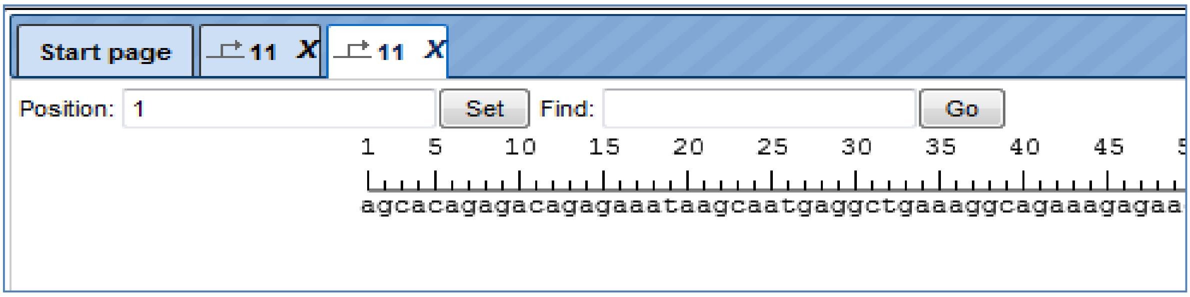
For comparison of the results you can click on individual chromosome sequences from both the original track and the Processed track from the Tree Area as shown below:



```
TAAAATATATATTTAATGT
```



The detailed view of the processed track is as shown below:



```
agcacagagacagagaaataagcaatgaggctgaaaggcagaaagagaa
```

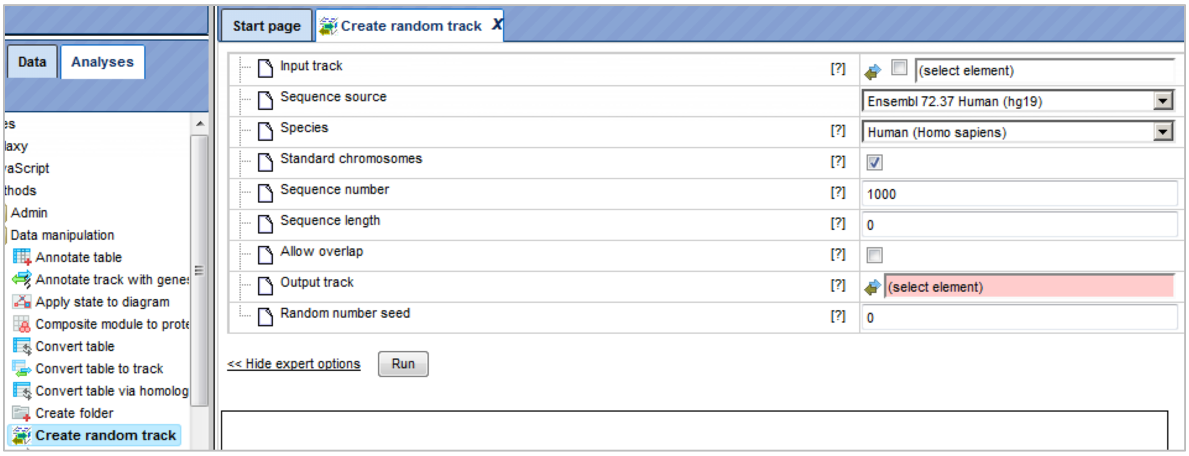
100bp are added to both the sides and thus from original 19bp track, you now have a track with 219bp. This processed track can be used further for other site analysis functions.

16.2.6. Create random track

This method creates a track of randomly sampled sequence regions, also denoted as intervals, segments or subsequences. Upstream regions of genes serve as source for the random segments.

Sampling can take into account an input track in two ways. First, the lengths of output regions are sampled from lengths observed in the input track, so that the output track has a similar length distribution. This functionality can be overridden by specifying a common sequence length, in which case all sampled sequences will have the same length. Second, gene upstream regions that overlap with segments in the input track can be omitted from the sampling. Omission of overlapping upstream regions is active by default and can be switched off (see parameter description).

Specification of an input track is optional. Random seed and sequence length arguments with values less than or equal to 0 are ignored. However, if no input track is provided, the sequence length argument is required.



Input track	[?]	□ (select element)
Sequence source		Ensembl 72.37 Human (hg19)
Species	[?]	Human (Homo sapiens)
Standard chromosomes	[?]	☑
Sequence number	[?]	1000
Sequence length	[?]	0
Allow overlap	[?]	□
Output track	[?]	(select element)
Random number seed	[?]	0

The input mask of the tool is shown above. The parameters are described in the following.

Input track: This is argument is optional. The input track can be supplied to obtain a random track with a similar length distribution and/or void of segments overlapping with input intervals.

Sequence source: The sequence source specifies which sequences are associated with intervals. Note that you can apply a custom source, e.g. a specifically uploaded genome. Clicking on the "Custom" option will open a new field to choose the custom sequence source.

Species: Upstream regions of genes will be compiled from the annotation for the specified species.

Standard chromosomes: If marked (default), sampling will only take into account standard chromosomes. As non-standard chromosomes, this analysis considers for instance haplotype segments.

Sequence number: This is the number of sequence regions to sample.

Sequence length: If greater than 0, this value specifies one length for all sampled sequence regions. Otherwise, an input track must be provided and random interval length will be sampled from length observed in the input track.

Allow overlap: If marked, sampled intervals are allowed to overlap with input intervals.

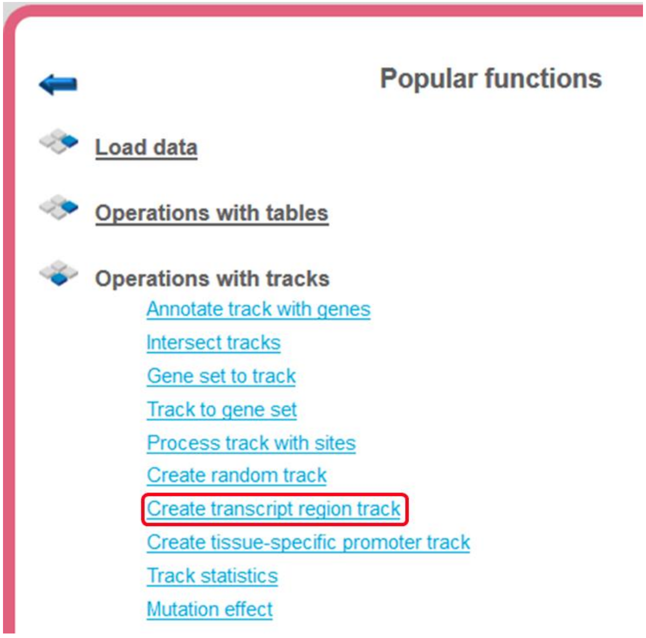
Output track: The path of the track with random track to be created.

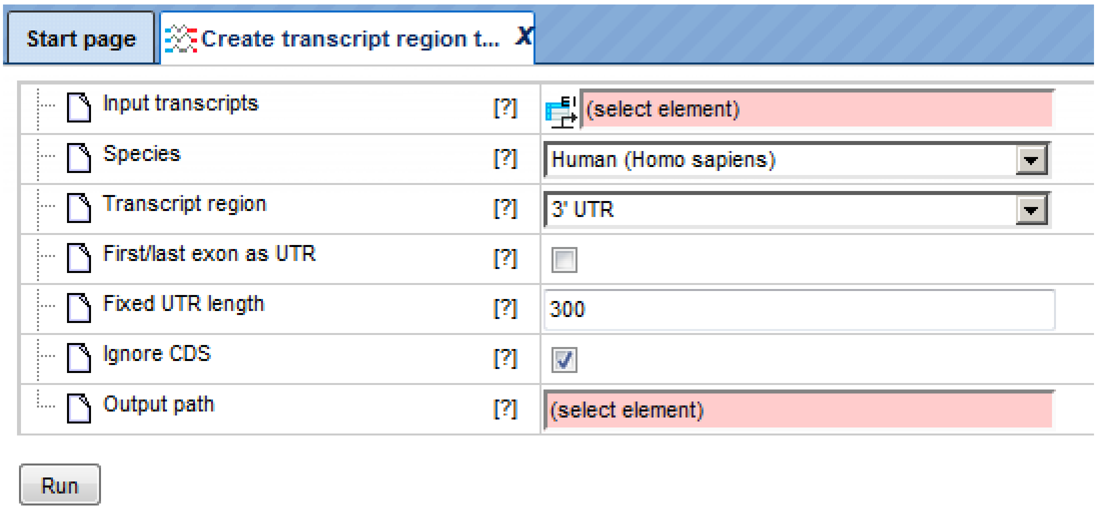
Random number seed: If greater than 0, this number will be supplied as seed for the random number generator in order to be able to reproduce the sampling result.

16.2.7. Create transcript region track

This method allows for creating tracks specific for particular transcript regions, e.g. promoters, 5’ UTRs, 3’ UTRs, exons, introns.

The analysis method ‘Create transcript region track’ can be found on the Start page, under the button ‘Popular functions’.





Input transcripts	[?]	(select element)
Species	[?]	Human (Homo sapiens)
Transcript region	[?]	3' UTR
First/last exon as UTR	[?]	□
Fixed UTR length	[?]	300
Ignore CDS	[?]	☑
Output path	[?]	(select element)

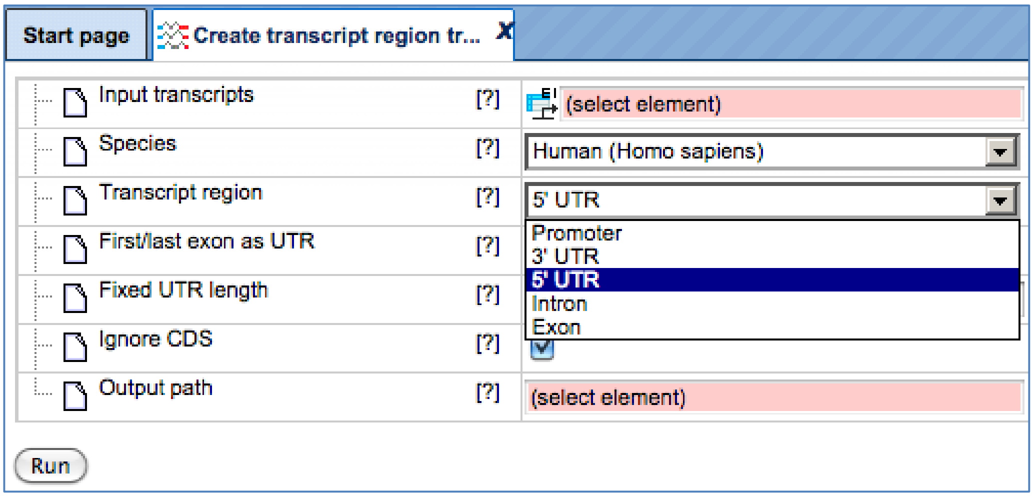
Here it is shown how to create a track with 5' UTR sequences starting from input transcripts. The input should be a table with Ensemble transcripts.

In the following, the input fields are shown one by one:

Input transcripts – Enter a table of Ensemble transcripts in this field. You can either drag and drop the file from the tree area or select it from the drop-down menu. In case your gene table does not have transcript information, use the *'convert table'* function(section XXX) to convert any gene or protein table into Ensembl transcripts.

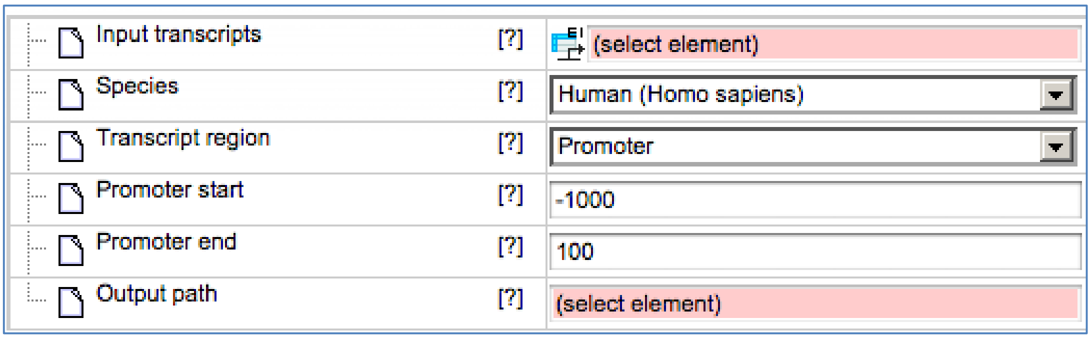
Species – Select the species of the input transcripts.

Transcript region – You can select the part of a transcript region which you wish to include in the output track. The region can be selected from 3' UTR, 5' UTR, promoter, intron, and exon.



Input transcripts	[?]	(select element)
Species	[?]	Human (Homo sapiens)
Transcript region	[?]	5' UTR
First/last exon as UTR	[?]	
Fixed UTR length	[?]	
Ignore CDS	[?]	☑
Output path	[?]	(select element)

As soon as you have chosen a transcript region from the drop-down menu, the following input fields are adjusted. If **Promoter** is selected as the transcript region, the input form becomes the following:

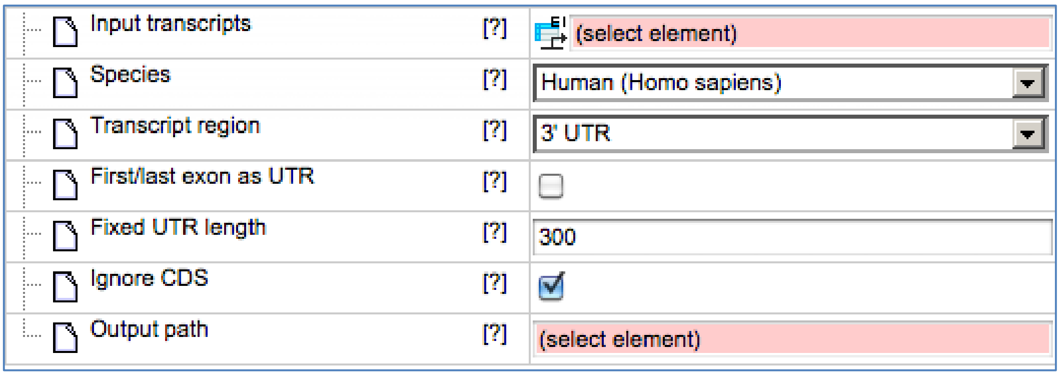


Input transcripts	[?]	(select element)
Species	[?]	Human (Homo sapiens)
Transcript region	[?]	Promoter
Promoter start	[?]	-1000
Promoter end	[?]	100
Output path	[?]	(select element)

Promoter start – You should specify the first base of the promoter relative to the TSS.

Promoter end – Here you should specify the last base of the promoter relative to the TSS.

If **3' UTR** or **5' UTR** are selected as the transcript region, the input form looks as follows.



Input transcripts	[?]	(select element)
Species	[?]	Human (Homo sapiens)
Transcript region	[?]	3' UTR
First/last exon as UTR	[?]	☐
Fixed UTR length	[?]	300
Ignore CDS	[?]	☑
Output path	[?]	(select element)

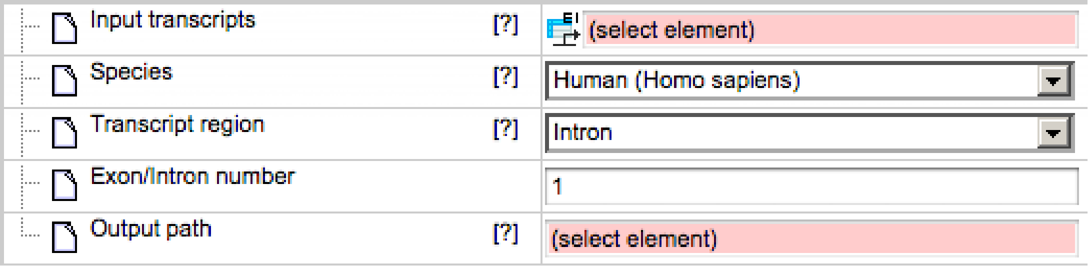
First/last exon as UTR - This check box can be used to select the first or last exon as UTR if it is not defined in the input transcript. By default this box is unchecked.

Fixed UTR length – The column can be used to create UTRs of fixed length. By default the method uses 300bp as the track length. Please note that actual UTRs can be very long.

Ignore CDS information – This box is checked to ignore CDS information and create fixed length UTRs. By default this box is checked.

Output path - Specify the path to store the result and indicate the name of the output track or sequences.

If **Exon** or **Intron** are selected as the transcript region, the input form is adjusted as follows:



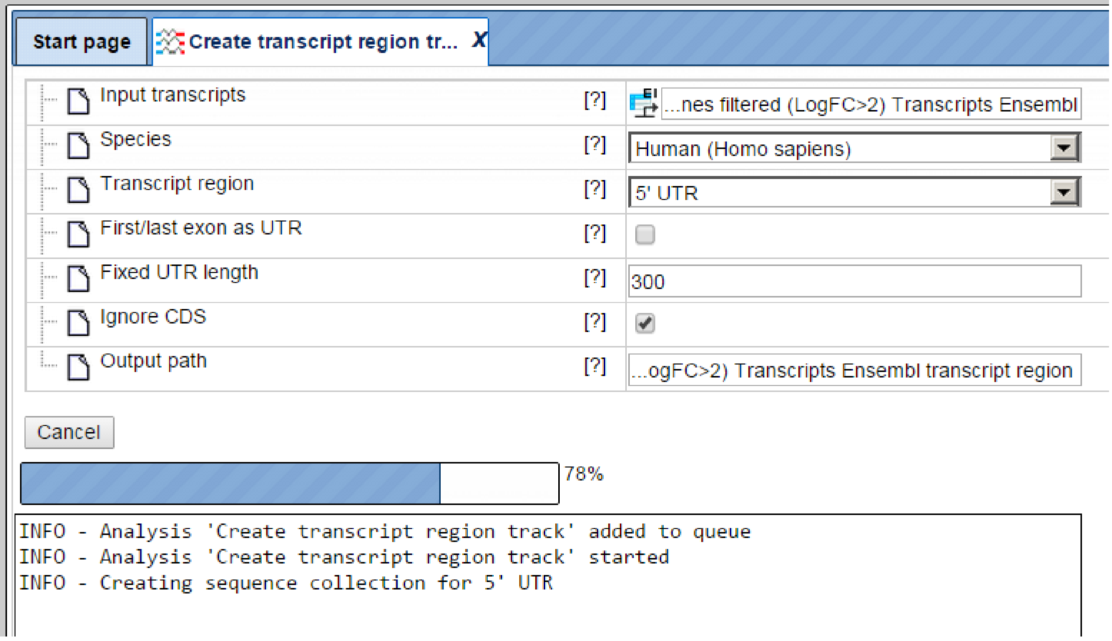
Input transcripts	[?]	(select element)
Species	[?]	Human (Homo sapiens)
Transcript region	[?]	Intron
Exon/Intron number		1
Output path	[?]	(select element)

Exon/Intron number - This field becomes active when you select Intron or Exon as the transcript region. You should specify 1, 2 ... for first, second, ... exon/intron, or -1, -2, ... for last, second last, ... exon/intron.

Output path - Specify the path to store the result and indicate the name of the output track or sequences.

Below it is shown how to create a track with 5' UTR sequences starting from the input transcripts. The input should be a table with Ensemble transcripts.

The analysis will start as shown below:



Input transcripts [?] ...nes filtered (LogFC>2) Transcripts Ensembl

Species [?] Human (Homo sapiens)

Transcript region [?] 5' UTR

First/last exon as UTR [?] ☐

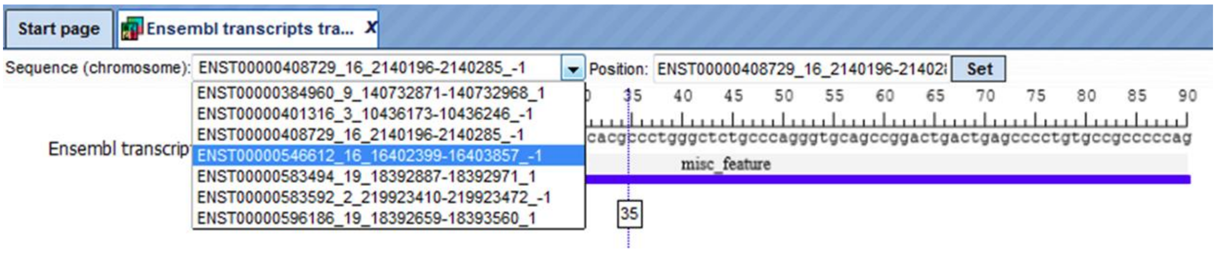
Fixed UTR length [?] 300

Ignore CDS [?] ☑

Output path [?] ...ogFC>2) Transcripts Ensembl transcript region

```
INFO - Analysis 'Create transcript region track' added to queue
INFO - Analysis 'Create transcript region track' started
INFO - Creating sequence collection for 5' UTR
```

After the run is completed, the output track is opened automatically in the work space as shown below:

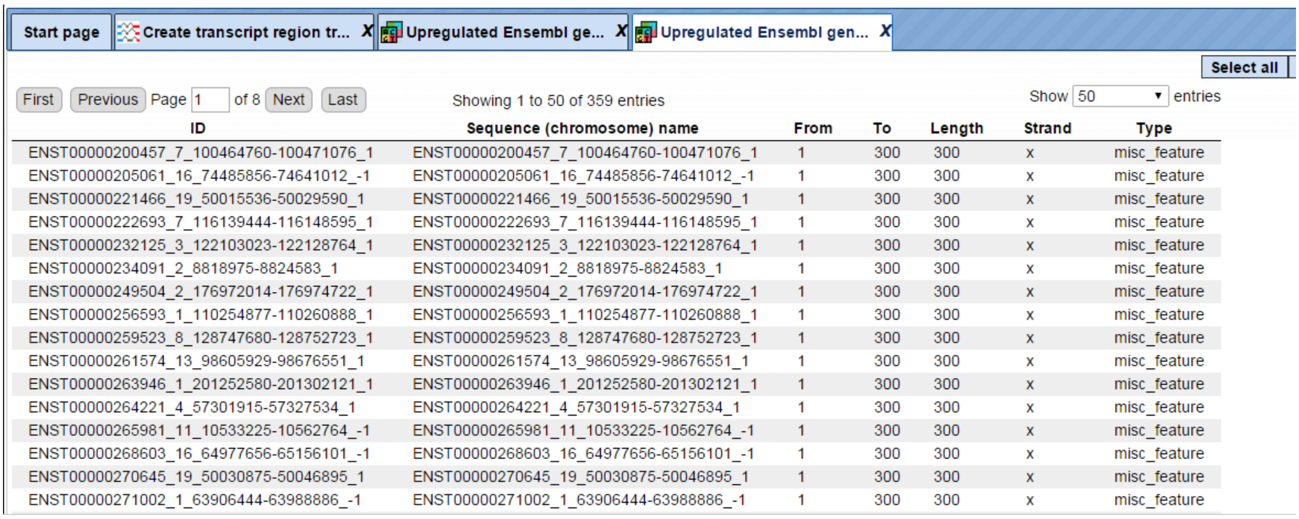

Start page
Ensembl transcripts tra...
Set
Ensembl transcrip
ENST00000384960_9_140732871-140732968_1
ENST00000401316_3_10436173-10436246_-1
ENST00000408729_16_2140196-2140285_-1
ENST00000546612_16_16402399-16403857_-1
ENST00000583494_19_18392887-18392971_1
ENST00000583592_2_219923410-219923472_-1
ENST00000596186_19_18392659-18393560_1
35 40 45 50 55 60 65 70 75 80 85 90
cacgcctgggctctgcccagggtgcagccggactgactgagcccctgtgccgccccag
misc_feature
35

Sequence (chromosome): ENST00000408729_16_2140196-2140285_-1

Position: ENST00000408729_16_2140196-21402

You can select the sequence (chromosome) number from the drop-down arrow menu and view the corresponding track file.

The track file when viewed as a table looks like as shown below:


Start page
Create transcript region tr...
Upregulated Ensembl ge...
Upregulated Ensembl gen...
Select all
First
Previous
Page 1 of 8
Next
Last
Show 50 entries

Showing 1 to 50 of 359 entries

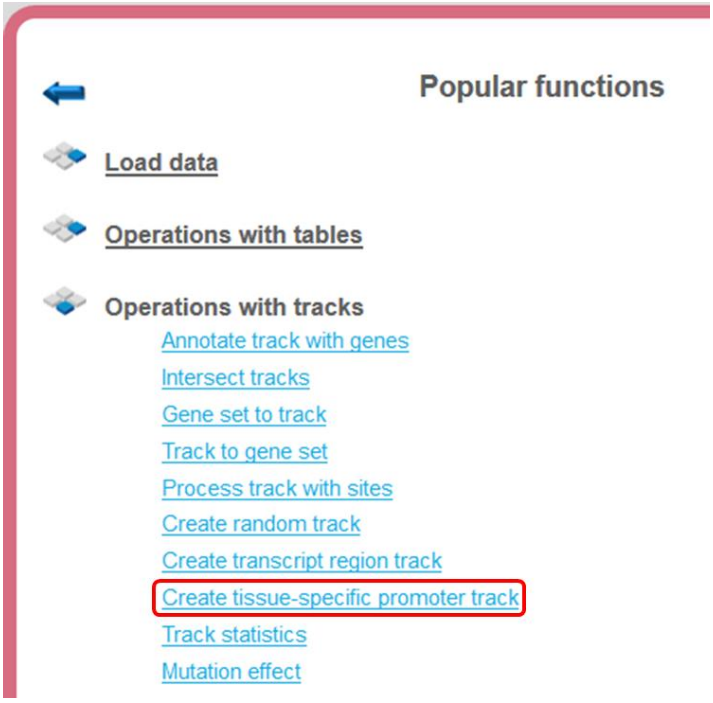
ID	Sequence (chromosome) name	From	To	Length	Strand	Type
ENST00000200457_7_100464760-100471076_1	ENST00000200457_7_100464760-100471076_1	1	300	300	x	misc_feature
ENST00000205061_16_74485856-74641012_-1	ENST00000205061_16_74485856-74641012_-1	1	300	300	x	misc_feature
ENST00000221466_19_50015536-50029590_1	ENST00000221466_19_50015536-50029590_1	1	300	300	x	misc_feature
ENST00000222693_7_116139444-116148595_1	ENST00000222693_7_116139444-116148595_1	1	300	300	x	misc_feature
ENST00000232125_3_122103023-122128764_1	ENST00000232125_3_122103023-122128764_1	1	300	300	x	misc_feature
ENST00000234091_2_8818975-8824583_1	ENST00000234091_2_8818975-8824583_1	1	300	300	x	misc_feature
ENST00000249504_2_176972014-176974722_1	ENST00000249504_2_176972014-176974722_1	1	300	300	x	misc_feature
ENST00000256593_1_110254877-110260888_1	ENST00000256593_1_110254877-110260888_1	1	300	300	x	misc_feature
ENST00000259523_8_128747680-128752723_1	ENST00000259523_8_128747680-128752723_1	1	300	300	x	misc_feature
ENST00000261574_13_98605929-98676551_1	ENST00000261574_13_98605929-98676551_1	1	300	300	x	misc_feature
ENST00000263946_1_201252580-201302121_1	ENST00000263946_1_201252580-201302121_1	1	300	300	x	misc_feature
ENST00000264221_4_57301915-57327534_1	ENST00000264221_4_57301915-57327534_1	1	300	300	x	misc_feature
ENST00000265981_11_10533225-10562764_-1	ENST00000265981_11_10533225-10562764_-1	1	300	300	x	misc_feature
ENST00000268603_16_64977656-65156101_-1	ENST00000268603_16_64977656-65156101_-1	1	300	300	x	misc_feature
ENST00000270645_19_50030875-50046895_1	ENST00000270645_19_50030875-50046895_1	1	300	300	x	misc_feature
ENST00000271002_1_63906444-63988886_-1	ENST00000271002_1_63906444-63988886_-1	1	300	300	x	misc_feature

The resulting tracks file can be used as input for various other workflows, for example to search for TF binding sites, enriched motifs and composite modules, and others.

16.2.8. Create tissue-specific promoter track

The analysis method 'Create tissue-specific promoter track' can be found on the Start page, under the button 'Popular functions'.

This method uses a set of Ensembl genes as input and extracts promoter regions by mapping it against the TSS locations defined in CAGE data in the Fantom5 (Nature 507:462–470) database (see also 19.10).



Popular functions

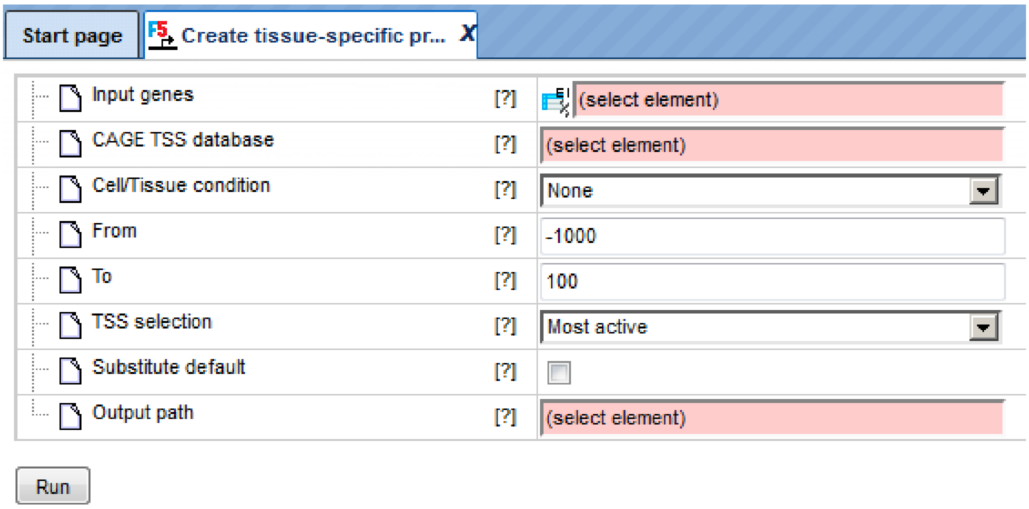
Load data

Operations with tables

Operations with tracks

- Annotate track with genes
- Intersect tracks
- Gene set to track
- Track to gene set
- Process track with sites
- Create random track
- Create transcript region track
- Create tissue-specific promoter track
- Track statistics
- Mutation effect

The input form is as shown below:



Start page | Create tissue-specific pr... X

Input genes	[?]	(select element)
CAGE TSS database	[?]	(select element)
Cell/Tissue condition	[?]	None
From	[?]	-1000
To	[?]	100
TSS selection	[?]	Most active
Substitute default	[?]	□
Output path	[?]	(select element)

Run

Input genes: Enter the set of genes or a gene table to extract transcription start sites (TSSs).

CAGE TSS database: Specify the path of the Fantom database.

Cell/Tissue condition: Once you specify the database, select the cells/tissues for which you want to create the promoter track from the drop-down menu.

From/To: Specify the promoter length relative to the TSS; by default the promoter length is from -1000 to +100 bp

TSS selection: The TSS should be selected if there are multiple TSS. By default the most active site is considered as TSS.

Substitute default: By default this box is unchecked. If checked it will substitute the gene promoter by default, if the promoter is missing in the selected condition.

Output path: Define the output file name and path in the tree area where you wish to save the Fantom5 promoter track.

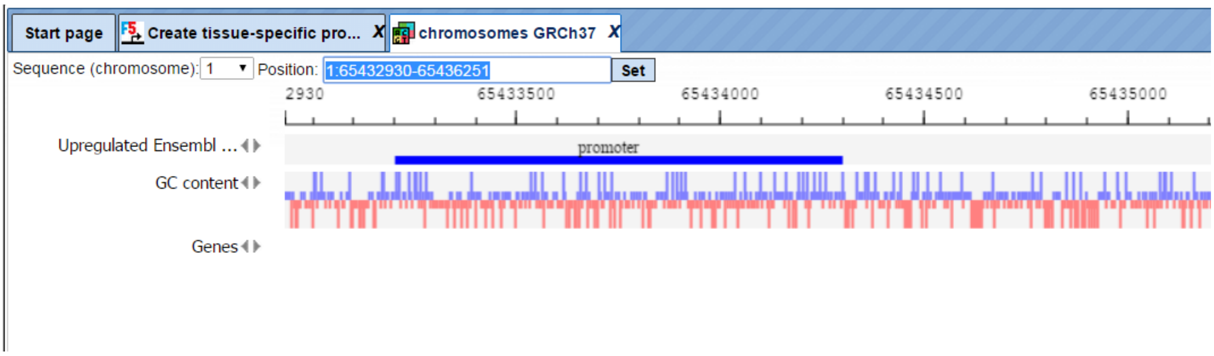
For example:

The method is run using a set of upregulated genes from brain tumor as input, specifying cerebellum→adult as cell tissue condition, and keeping all other conditions as default.

The input dataset can be found here:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Experiment%20normalized%20(RMA)%20(Differentially%20expressed%20genes%20Affy)/Upregulated%20Ensembl%20genes%20filtered%20(LogFC%3E2)

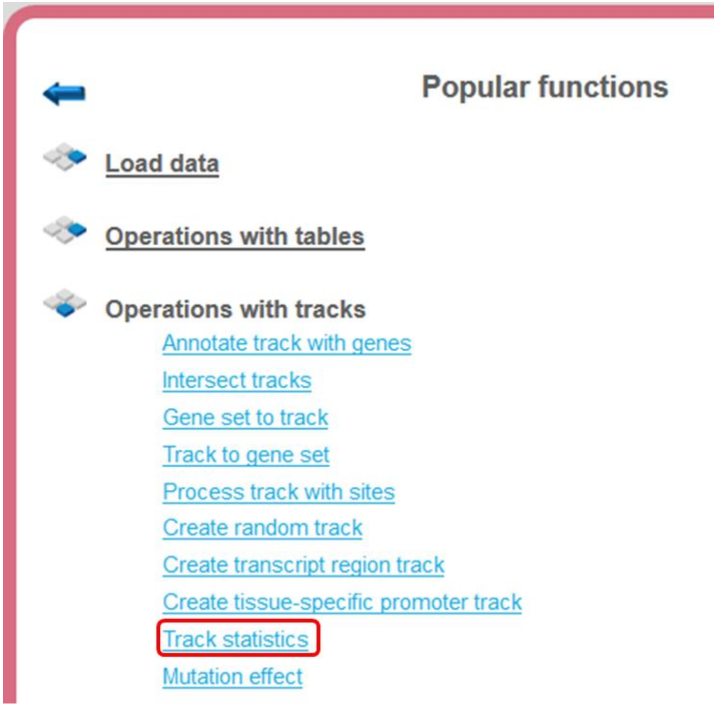
Upon completion the output file is opened in the work space and looks as shown below:



These set of promoters can be used as an input for other site analysis workflows.

16.2.9. Track statistics

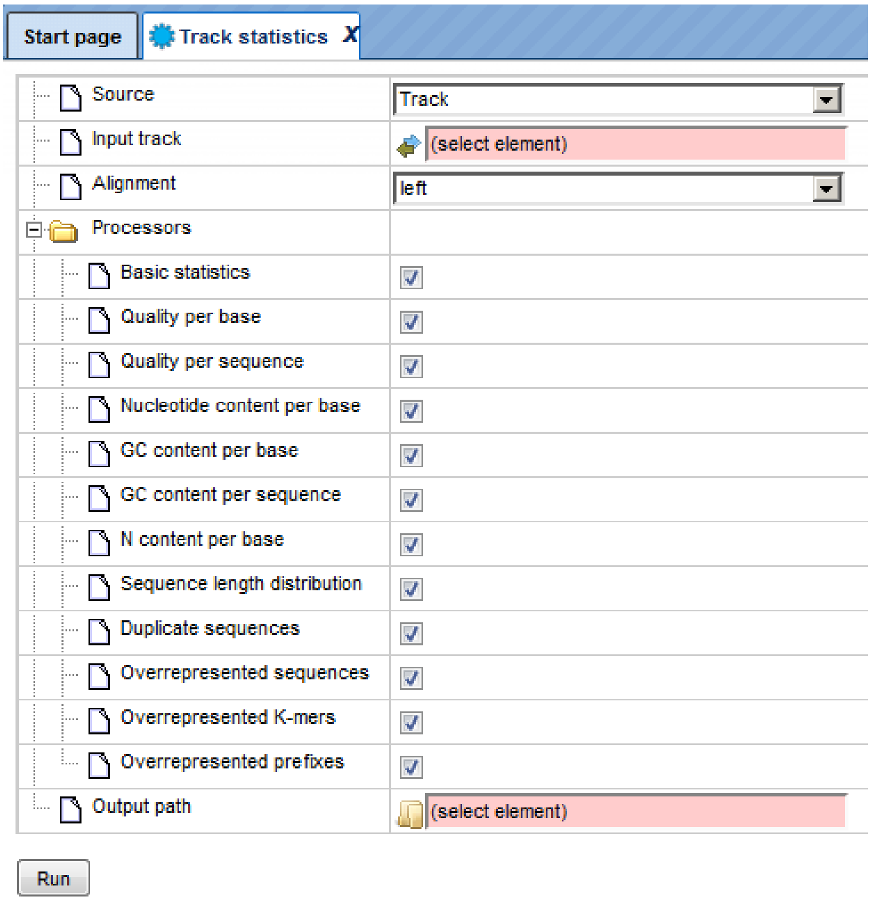
The analysis method *Track Statistics* can be found on the Start page, under the button 'Popular functions' as shown below:



Popular functions

- Load data
- Operations with tables
- Operations with tracks
  - Annotate track with genes
  - Intersect tracks
  - Gene set to track
  - Track to gene set
  - Process track with sites
  - Create random track
  - Create transcript region track
  - Create tissue-specific promoter track
  - Track statistics
  - Mutation effect

This method gathers various statistical information about any input track or Fastq file. This information is helpful to calculate the number of reads in a particular input file which is a pre-requisite for many workflows. The input form is shown below:



Source	Track
Input track	(select element)
Alignment	left
Processors	
Basic statistics	☑
Quality per base	☑
Quality per sequence	☑
Nucleotide content per base	☑
GC content per base	☑
GC content per sequence	☑
N content per base	☑
Sequence length distribution	☑
Duplicate sequences	☑
Overrepresented sequences	☑
Overrepresented K-mers	☑
Overrepresented prefixes	☑
Output path	(select element)

The input form parameters are as follows:

Source –Specify the type of input track that you wish to process using this method. The source can vary from Track, FastQ, Solid and CSFastQ.

FastQ file – This is a sequence file with reads in FastQ format.

CSfastq file – This is a file containing reads in color space.

If the source is **Track**, you have to specify the **Input track.** Based on the specified source, you should input the track to process using this method.

Alignment – Specify whether to align sites on the left or on the right.

In case the source is a Fastq or CSFasta file, you need to specify:

Quality encoding – This specifies how phred quality values are encoded in the FASTQ file. In most of the cases the system detects this value automatically. You may change it manually if the auto-detection worked incorrectly.

Alignment – Specify whether to align sites on the left or on the right.

Processors – This is a list of methods to gather diverse statistics:

Basic statistics – Gathers basic statistics like reads count and average read length.

Quality per base – Distribution of phred quality score along the bases.

Quality per sequence – Distribution of phred quality score among the sequences.

Nucleotide content per base – Distribution of individual nucleotides along the bases.

GC content per base – Distribution of GC along the bases.

GC content per sequence – Draws a distribution of GC content among reads.

N content per base – Distribution of 'N' along the bases.

Sequence length distribution – Calculates a distribution of read lengths and outputs them as a table and a chart.

Duplicate sequences – Calculates the rate of sequences duplication: how many sequences occur 2, 3 etc. times relative to unique sequences. This statistic is based on the first 200000 reads.

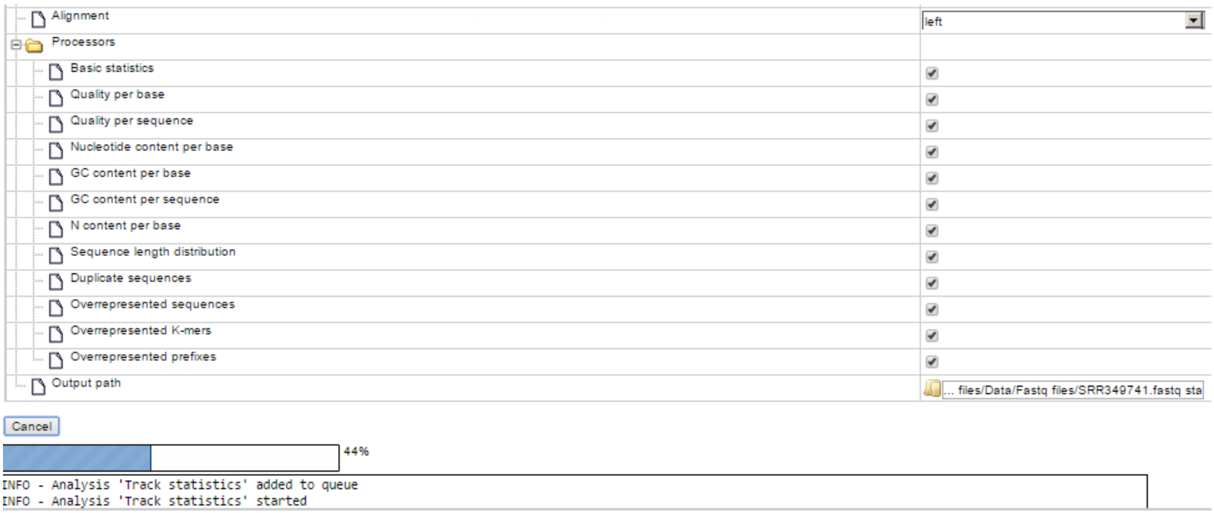
Overrepresented sequences – Looks for sequences which appear in more than 0.1% cases.

Overrepresented K-mers – Search for K-mers which are represented 3x times per sequence or 5x times per position.

Overrepresented prefixes – Search for read prefixes (starting from the read start) up toa length od 15 bp which are overrepresented in the set.

Output path – Specify the output file name and path in the tree where you want to save the output file.

After pressing ‘Run’ the method runs as shown below:



Alignment	left
Processors	
Basic statistics	☑
Quality per base	☑
Quality per sequence	☑
Nucleotide content per base	☑
GC content per base	☑
GC content per sequence	☑
N content per base	☑
Sequence length distribution	☑
Duplicate sequences	☑
Overrepresented sequences	☑
Overrepresented K-mers	☑
Overrepresented prefixes	☑
Output path	... files/Data/Fastq files/SRR349741.fastq sta

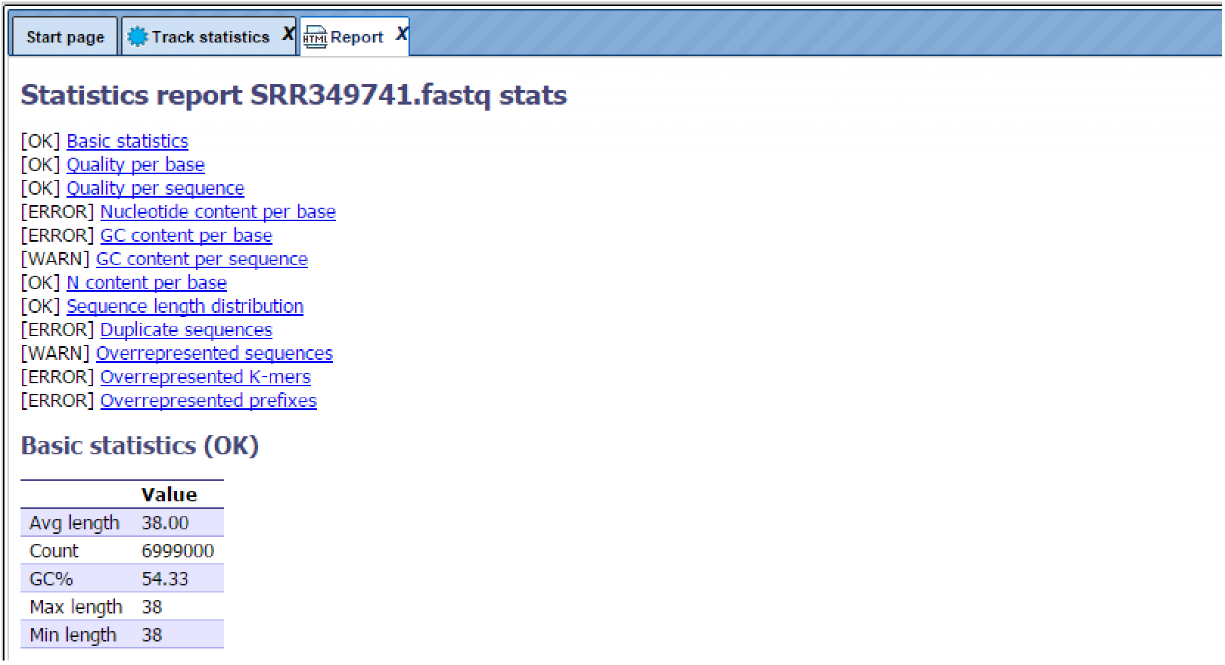
```
INFO - Analysis 'Track statistics' added to queue
INFO - Analysis 'Track statistics' started
```

After completion of the method, the output folder is created and an HTML report opens in the workspace.

The link to an example HTML output report for an input FastQ file is here:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/RNA-Seq%20analysis%20of%20human%20esophageal%20squamous%20cell%20carcinoma%20(ESCC)%2C%20GSE32424%2C%20FASTQ%20files/Data/Fastq%20files/SRR349741.fastq%20stats/Report

The report looks as shown below:



Statistics report SRR349741.fastq stats

[OK] Basic statistics
[OK] Quality per base
[OK] Quality per sequence
[ERROR] Nucleotide content per base
[ERROR] GC content per base
[WARN] GC content per sequence
[OK] N content per base
[OK] Sequence length distribution
[ERROR] Duplicate sequences
[WARN] Overrepresented sequences
[ERROR] Overrepresented K-mers
[ERROR] Overrepresented prefixes

Basic statistics (OK)

	Value
Avg length	38.00
Count	6999000
GC%	54.33
Max length	38
Min length	38

It gives a summary of all the parameters at first glance; details can be retrieved by clicking on the respective hyperlinks.

The tool calculates each parameter and lets you know if the particular value looks 'OK' in the input sequence, else it gives a WARNING or ERROR

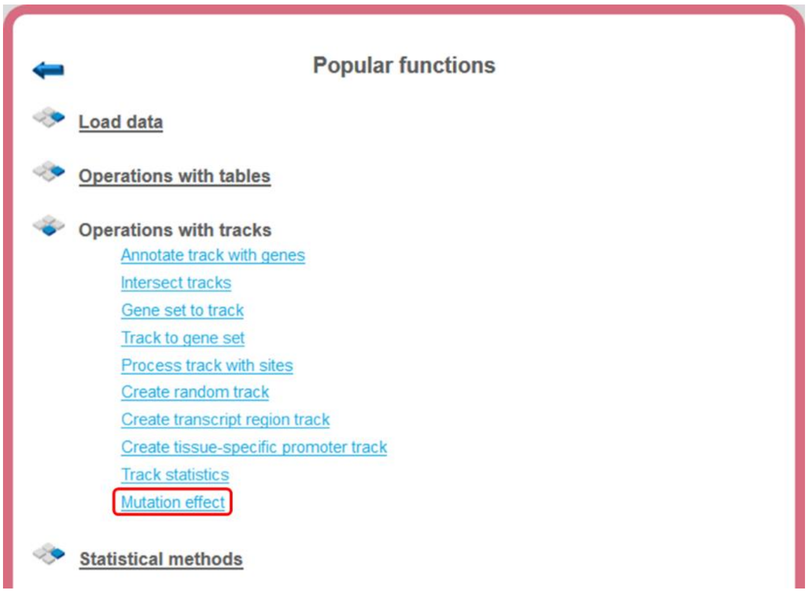
The link to an output HTML file using a track file as input is here:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/E2F1%20binding%20regions%20in%20HeLa%20cells%2C%20ChIP-Seq/Data/GSM558469_E2F1_hg19%20filtered%20chr%201%20stats/Report

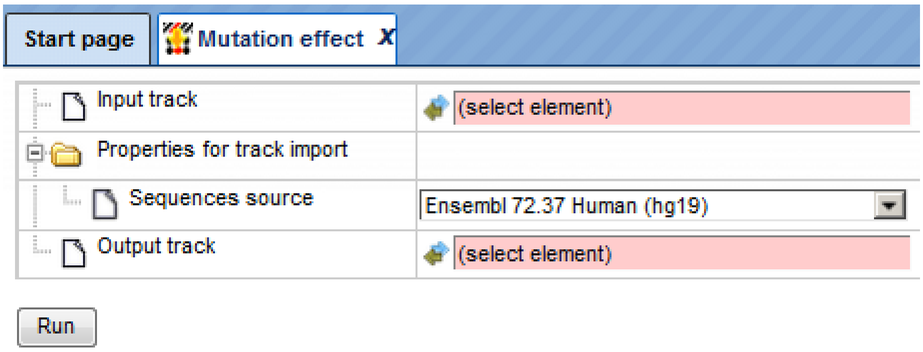
16.2.10. Mutation effect analysis

This tool allows to find proteins affected by mutations. The mutation effect analysis determines the effect of a certain genomic mutation on a protein, such as synonymous, gain/loss of stop codon, frameshift or others. It accepts a list of Single Nucleotide Variations (mutations), and determines the type for each mutation.

The analysis "Mutation effect" can be found in the NGS folder of analysis methods (analyses/Methods/NGS/Mutation_effect) or on the start page button 'Popular functions' under the section 'Operations with tracks'.



Step 1. Open the analysis form from the Start page. It will open in the main Work Space and looks as shown below:



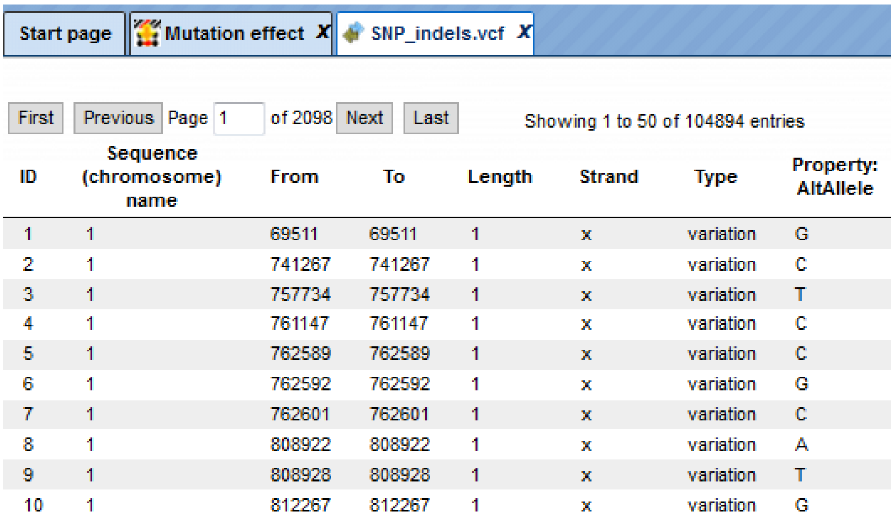
Input track	(select element)
Properties for track import	
Sequences source	Ensembl 72.37 Human (hg19)
Output track	(select element)

Step 2. The **Input track** is a track file with a list of single nucleotide variations (mutations) and should be in vcf format.

One input example is here on the platform:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Chronic%20Myeloid%20Leukemia%20Patient%20Genotyping/Data/SNP_indels.vcf

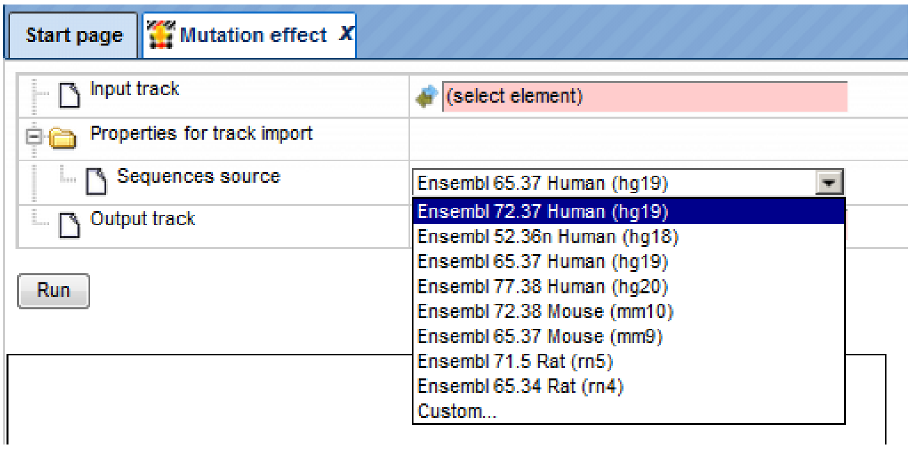
Open the track file as a table, and for each variation point you can see several columns with genomic position, chromosome, alternative nucleotide etc., as shown below.



Showing 1 to 50 of 104894 entries

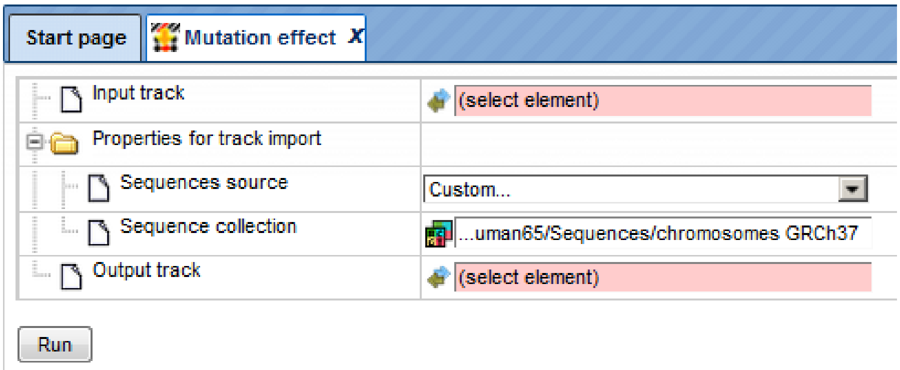
ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: AltAllele
1	1	69511	69511	1	x	variation	G
2	1	741267	741267	1	x	variation	C
3	1	757734	757734	1	x	variation	T
4	1	761147	761147	1	x	variation	C
5	1	762589	762589	1	x	variation	C
6	1	762592	762592	1	x	variation	G
7	1	762601	762601	1	x	variation	C
8	1	808922	808922	1	x	variation	A
9	1	808928	808928	1	x	variation	T
10	1	812267	812267	1	x	variation	G

Step 2. Verify the **Sequences source** and use the drop-down menu for different Ensembl genome annotations of human, mouse and rat, as shown below.


Start page
Mutation effect
Run
Ensembl 72.37 Human (hg19)
Ensembl 52.36n Human (hg18)
Ensembl 65.37 Human (hg19)
Ensembl 77.38 Human (hg20)
Ensembl 72.38 Mouse (mm10)
Ensembl 65.37 Mouse (mm9)
Ensembl 71.5 Rat (rn5)
Ensembl 65.34 Rat (rn4)
Custom...

Input track	(select element)
Properties for track import	
Sequences source	Ensembl 65.37 Human (hg19)
Output track	

Alternatively, you can choose 'Custom' from the same menu if you would like to specify another genome, e.g. a particular patient genome imported into the platform before. As soon as the option 'Custom' is chosen, an additional field, Sequence collection, automatically appears on the input form (screenshot below), and you can specify the sequences location manually.



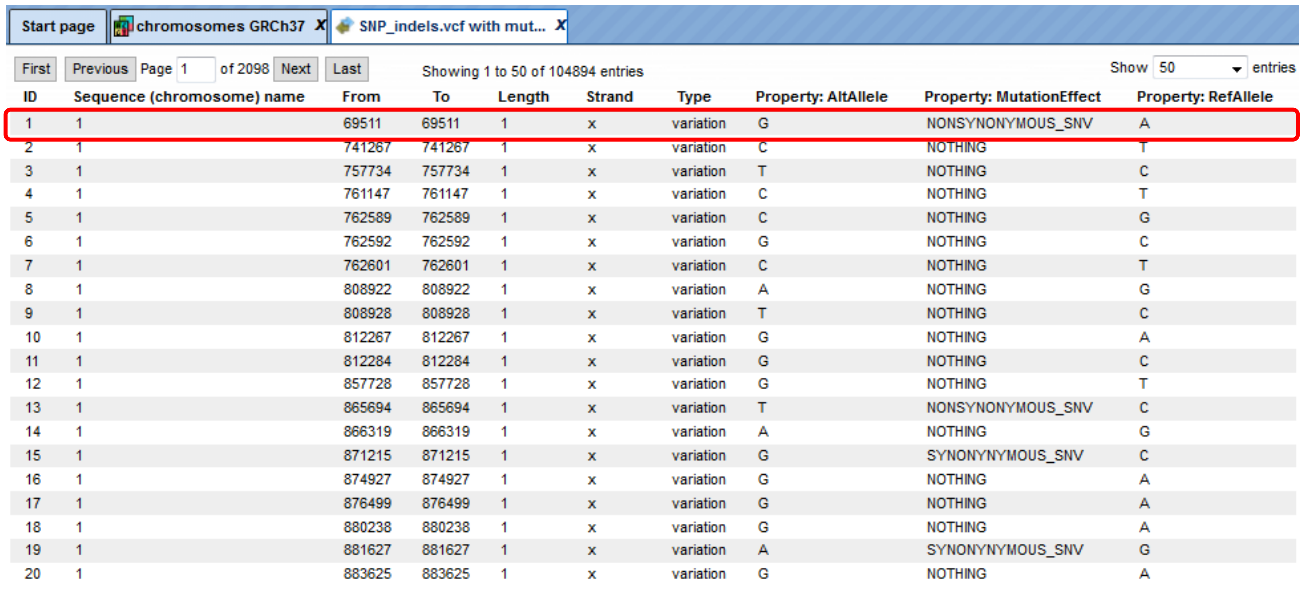
Input track	(select element)
Properties for track import	
Sequences source	Custom...
Sequence collection	...uman65/Sequences/chromosomes GRCh37
Output track	(select element)

Step 3. Specify the path and name of the **Output track**.

After completion the output track file (SNP_indels.vcf with mutation effect) is opened by default in the work space.

This resulting track can be found in the Examples folder under the URL: http://genexplain-platform.com/bioumlweb/#de=data/Examples/Chronic%20Myeloid%20Leukemia%20Patient%20Genotyping/Data/SNP_indels.vcf%20with%20mutation%20effect

The output track is created from the input track by adding the single column 'MutationEffect' with the determined mutation type. Note that a single mutation can affect multiple proteins with distinct consequences. In this case the MutationEffect column contains a list of mutation types separated by comma. This analysis uses the Ensembl database for protein genomic annotations.



Showing 1 to 50 of 104894 entries

ID	Sequence (chromosome) name	From	To	Length	Strand	Type	Property: AltAllele	Property: MutationEffect	Property: RefAllele
1	1	69511	69511	1	x	variation	G	NONSYNONYMOUS_SNV	A
2	1	741267	741267	1	x	variation	C	NOTHING	T
3	1	757734	757734	1	x	variation	T	NOTHING	C
4	1	761147	761147	1	x	variation	C	NOTHING	T
5	1	762589	762589	1	x	variation	C	NOTHING	G
6	1	762592	762592	1	x	variation	G	NOTHING	C
7	1	762601	762601	1	x	variation	C	NOTHING	T
8	1	808922	808922	1	x	variation	A	NOTHING	G
9	1	808928	808928	1	x	variation	T	NOTHING	C
10	1	812267	812267	1	x	variation	G	NOTHING	A
11	1	812284	812284	1	x	variation	G	NOTHING	C
12	1	857728	857728	1	x	variation	G	NOTHING	T
13	1	865694	865694	1	x	variation	T	NONSYNONYMOUS_SNV	C
14	1	866319	866319	1	x	variation	A	NOTHING	G
15	1	871215	871215	1	x	variation	G	SYNONYNYMOUS_SNV	C
16	1	874927	874927	1	x	variation	G	NOTHING	A
17	1	876499	876499	1	x	variation	G	NOTHING	A
18	1	880238	880238	1	x	variation	G	NOTHING	A
19	1	881627	881627	1	x	variation	A	SYNONYNYMOUS_SNV	G
20	1	883625	883625	1	x	variation	G	NOTHING	A

The upper example highlighted by the red box has ID=1 in the track. The columns **From** and **To** define the positions of the affected position within the genome on chromosome 1 (**Sequence (chromosome) name**). The column **Length** shows the length of the position, here 1. The **Property:AltAllele** exhibits the nucleotide in the mutated sequence and **Property:RefAllele** gives the nucleotide of the reference genome at the indicated position. The **Property:MutationEffect** shows NONSYNOMYMOUS and means a single nucleotide change which will cause an amino acid change.

Possible Mutation effect types are:

1. SYNONYNYMOUS_SNV - a single nucleotide change that does not cause an amino acid change

2. NONSYNONYMOUS_SNV - a single nucleotide change that causes an amino acid change

3. STOP_GAIN - a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that leads to the creation of a stop codon at the variant site. For frameshift mutations, the creation of a stop codon downstream of the variant will not be counted as "stopgain"!

4. STOP_LOSS - a nonsynonymous SNV, frameshift insertion/deletion, nonframeshift insertion/deletion or block substitution that leads to the immediate elimination of a stop codon at the variant site

5. NONFRAMESHIFT_INSERTION - an insertion of 3 or multiples of 3 nucleotides that does not cause frameshift changes in the protein coding sequence

6. NONFRAMESHIFT_DELETION - a deletion of 3 or mutliples of 3 nucleotides that does not cause frameshift changes in the protein coding sequence

7. NONFRAMESHIFT_BLOCK_SUBSTITUTION - a block substitution of one or more nucleotides that does not cause frameshift changes in the protein coding sequence

8. FRAMESHIFT_INSERTION - an insertion of one or more nucleotides that causes frameshift changes in the protein coding sequence

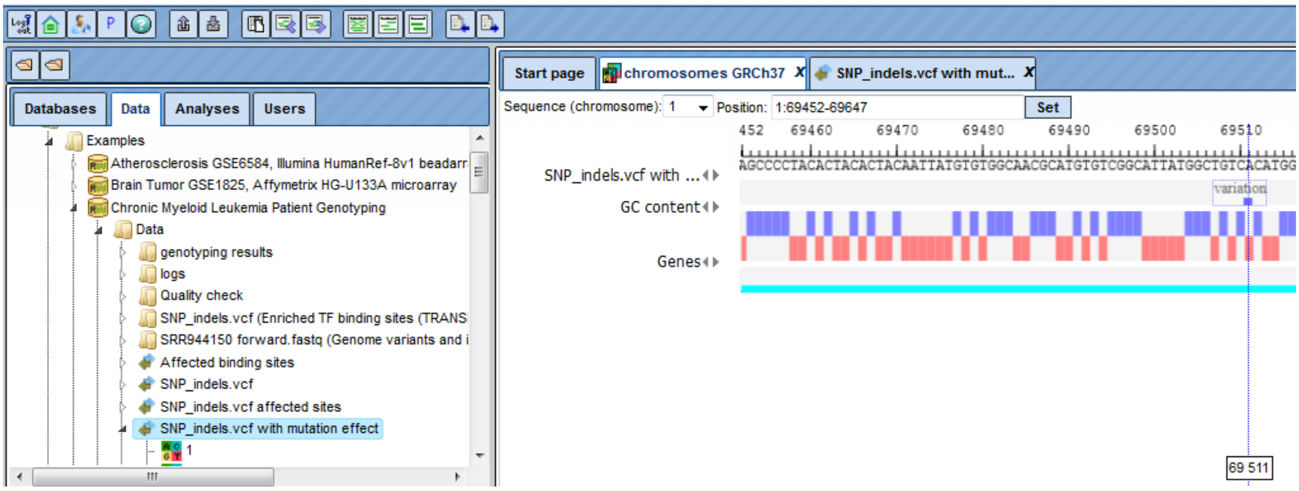
9. FRAMESHIFT_DELETION - a deletion of one or more nucleotides that causes frameshift changes in the protein coding sequence

10. FRAMESHIFT_BLOCK_SUBSTITUTION - a block substitution of one or more nucleotides that causes frameshift changes in the protein coding sequence

11. NOTHING - coding sequence is not changed

Tip. If you would like to not show the single nucleotide changes that do not cause an amino acid change, use Property_MutationEffect != 'NOTHING' for filtering.

The output track can be opened in the genome browser as shown in the picture below.

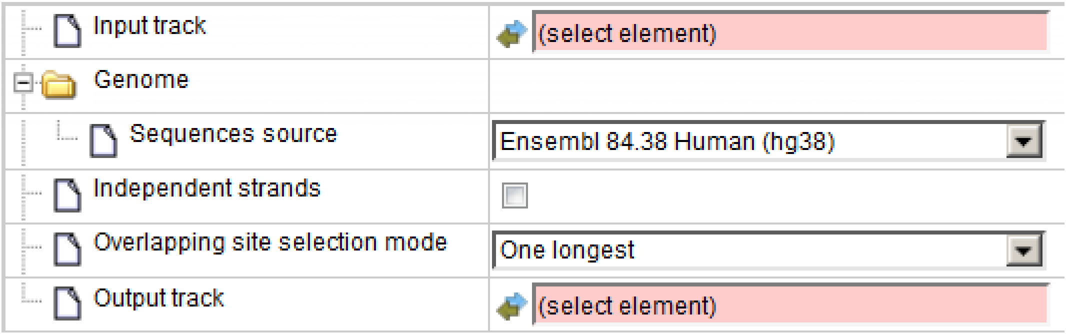


Sequence (chromosome): 1 Position: 1:69452-69647

Databases
Data
Analyses
Users
Examples
Atherosclerosis GSE6584, Illumina HumanRef-8v1 beadarr
Brain Tumor GSE1825, Affymetrix HG-U133A microarray
Chronic Myeloid Leukemia Patient Genotyping
Data
genotyping results
logs
Quality check
SNP_indels.vcf (Enriched TF binding sites (TRANS
SRR944150 forward.fastq (Genome variants and i
Affected binding sites
SNP_indels.vcf
SNP_indels.vcf affected sites
SNP_indels.vcf with mutation effect
1
Start page
chromosomes GRCh37
SNP_indels.vcf with mut...
Set
452
69460
69470
69480
69490
69500
69510
AGCCCCTACACTACACTACAATTATGTGTGGCAACGCATGTGTCGGCATTATGGCTGTCACATGG
variation
SNP_indels.vcf with ...
GC content
Genes
69 511

16.2.11. Remove overlapping sites

This method removes overlapping sites from any track and constructs a subset of the input track with no sites overlap in the output track. It can be found under the tab *Analyses*, in the folder Methods/Data manipulation/Remove overlapping sites. Here the default input form is shown:



Input track	(select element)
Genome	
Sequences source	Ensembl 84.38 Human (hg38)
Independent strands	□
Overlapping site selection mode	One longest
Output track	(select element)

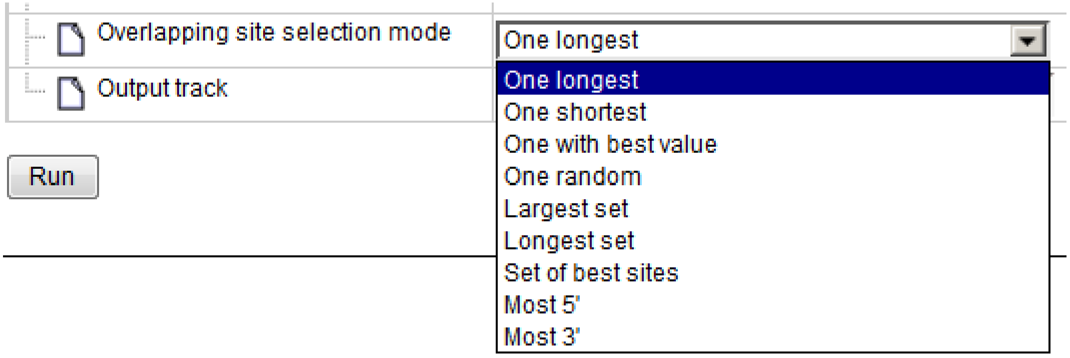
In the following, we will consider the input fields one by one.

Input track. You can drag & drop the input track from your project within the tree area. Alternatively, you may click on the pink field "select element" and a new window will open, where you can select the track.

Sequence source. Specify the reference genome or select *Custom* to specify the sequences location manually.

Independent strands. Select whether you want to handle sites from different strands independently; by default it is unchecked, and strands are not handled independently.

Overlapping site selection mode. Choose one mode from the drop-down menu for the selection of sites.



Overlapping site selection mode	One longest
Output track	

The *One longest*, *One shortest*, *One with best value* and *One random* modes will select a single site from the set of overlapping sites. Other modes can select more than one site from the set of overlapping sites, but the resulting sites will not overlap with each other. The *Largest set* mode constructs a non-overlapping set of sites with the maximum number of sites. The *Longest set* mode constructs a non-overlapping set with the largest total length of sites. The *Set of best sites* mode iteratively selects the best site and removes sites that overlap with the best site, then selects the best from the remaining sites and removes those which overlap with the best site, and so on until no sites remain. The *Most 5′* and *Most 3′* select the site located at the 5′ or 3′ end.

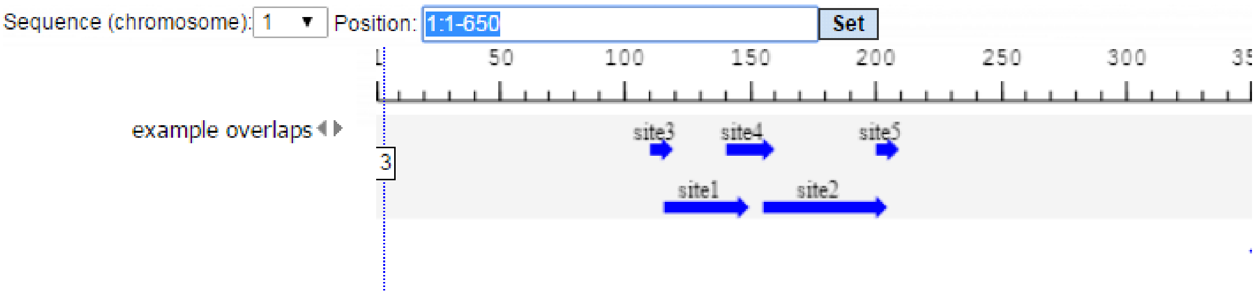
Output track. Define where the track with the result should be located in your project tree. You can do so by clicking on the pink box (select element) in the field, and a new window will open, where you can select the location of the resulting track and define its name.

Press the [Run] button and wait until the method is completed.

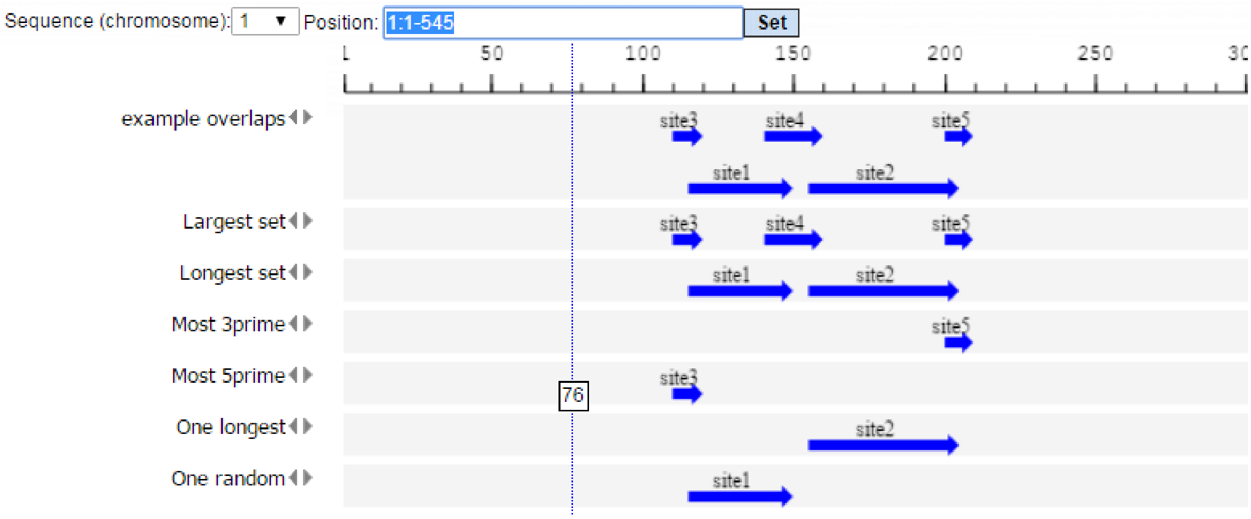
For this example, all further steps are demonstrated with the following input track:

data/Examples/Sample data/Data/Overlapping sites/example overlaps

The track looks as shown below:



We perform several runs with different site selection modes. All different output tracks are shown below:



16.3. Statistical methods

These methods have been described in detail in other sections, where they contribute essentially to certain workflows. Please, refer to these sections as specified below.

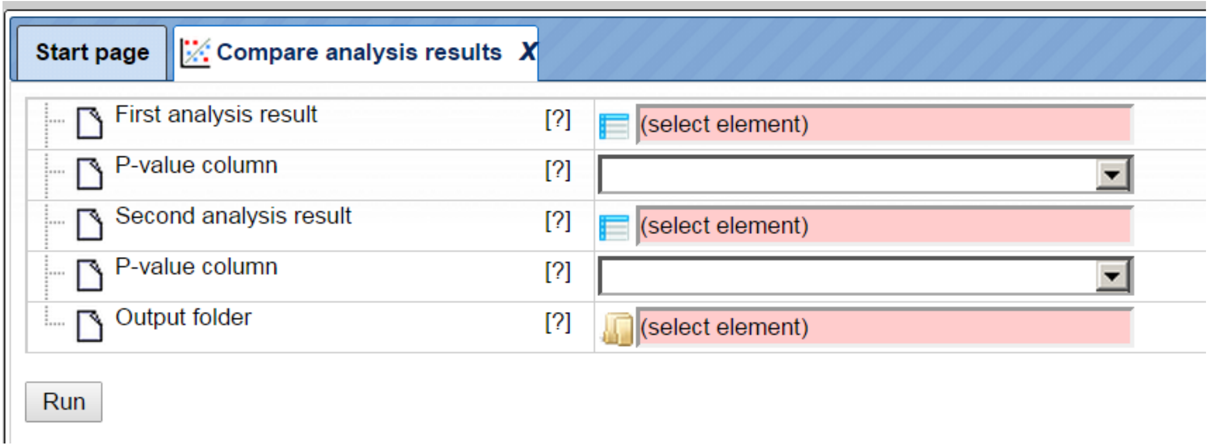
Principal Component Analysis (PCA)	See Section 10.1.4
LIMMA (Linear Models for MicroArrays)	See Section 4.2.1
EBarrays	See Section 4.2.2

16.3.1. Compare analysis results

This tool compares P-values in two analysis results. Analyses of interest are, for instance, binding sites or GO term enrichment results. The comparison can help to reveal items that show different enrichment across certain conditions.

This analysis method can be found on the Start page under the button “Popular functions”.

The input form looks as shown below:



First analysis result	[?]	(select element)
P-value column	[?]	
Second analysis result	[?]	(select element)
P-value column	[?]	
Output folder	[?]	(select element)

In the following, the input fields are shown one by one:

First analysis result - Enter the first input table which you wish to compare having a P-value column.

Note: Currently the method does not compare FDR or log (P-value) columns.

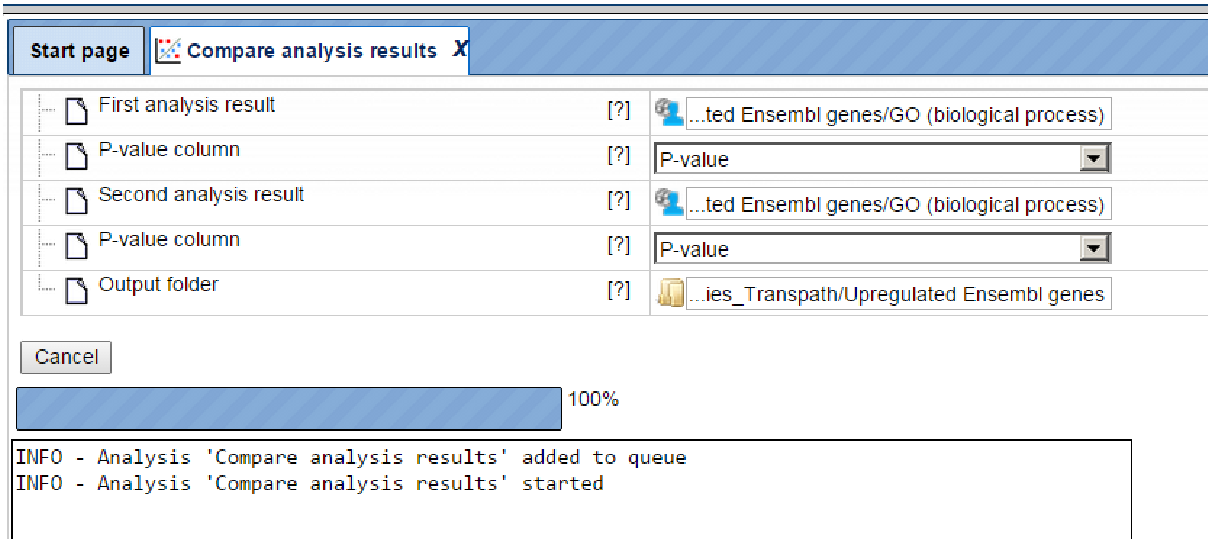
P-value column - From the drop-down menu select the P-value column which will be used for comparison.

Second analysis result - Enter the second input table which you wish to compare having a P-value column.

P-value column - Select the P-value column from the drop-down menu.

Output folder - Specify the path to store the result and indicate the name of the output folder.

Here, two enrichment results are taken for comparison from the Examples folder; the analysis will start as shown below:



First analysis result: ...ted Ensembl genes/GO (biological process)

P-value column: P-value

Second analysis result: ...ted Ensembl genes/GO (biological process)

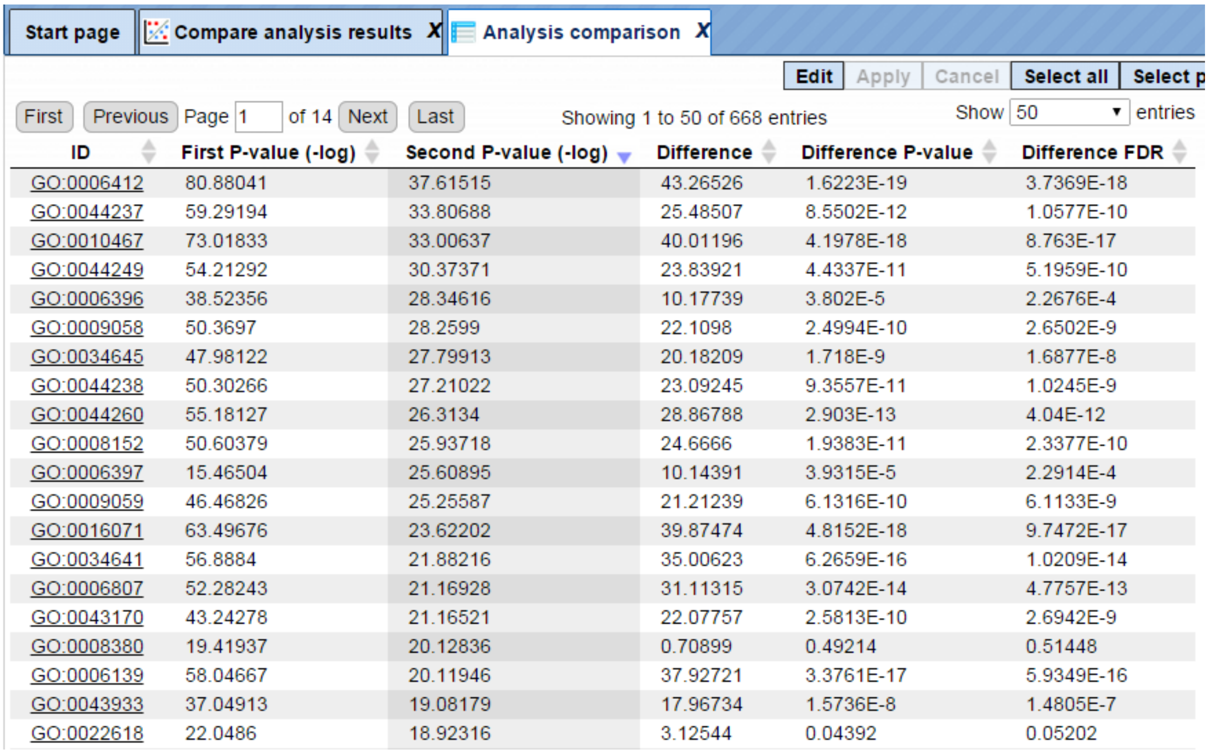
P-value column: P-value

Output folder: ...ies_Transpath/Upregulated Ensembl genes

```
INFO - Analysis 'Compare analysis results' added to queue
INFO - Analysis 'Compare analysis results' started
```

The output consists of two files: the Analysis comparison plot and the **Analysis comparison Table**.

The **Analysis comparison table** as shown below lists all P-values, absolute differences, difference P-values and estimated FDR.



ID	First P-value (-log)	Second P-value (-log)	Difference	Difference P-value	Difference FDR
GO:0006412	80.88041	37.61515	43.26526	1.6223E-19	3.7369E-18
GO:0044237	59.29194	33.80688	25.48507	8.5502E-12	1.0577E-10
GO:0010467	73.01833	33.00637	40.01196	4.1978E-18	8.763E-17
GO:0044249	54.21292	30.37371	23.83921	4.4337E-11	5.1959E-10
GO:0006396	38.52356	28.34616	10.17739	3.802E-5	2.2676E-4
GO:0009058	50.3697	28.2599	22.1098	2.4994E-10	2.6502E-9
GO:0034645	47.98122	27.79913	20.18209	1.718E-9	1.6877E-8
GO:0044238	50.30266	27.21022	23.09245	9.3557E-11	1.0245E-9
GO:0044260	55.18127	26.3134	28.86788	2.903E-13	4.04E-12
GO:0008152	50.60379	25.93718	24.6666	1.9383E-11	2.3377E-10
GO:0006397	15.46504	25.60895	10.14391	3.9315E-5	2.2914E-4
GO:0009059	46.46826	25.25587	21.21239	6.1316E-10	6.1133E-9
GO:0016071	63.49676	23.62202	39.87474	4.8152E-18	9.7472E-17
GO:0034641	56.8884	21.88216	35.00623	6.2659E-16	1.0209E-14
GO:0006807	52.28243	21.16928	31.11315	3.0742E-14	4.7757E-13
GO:0043170	43.24278	21.16521	22.07757	2.5813E-10	2.6942E-9
GO:0008380	19.41937	20.12836	0.70899	0.49214	0.51448
GO:0006139	58.04667	20.11946	37.92721	3.3761E-17	5.9349E-16
GO:0043933	37.04913	19.08179	17.96734	1.5736E-8	1.4805E-7
GO:0022618	22.0486	18.92316	3.12544	0.04392	0.05202

The output columns are explained below:

First P-value (-log): This column contains the –log P-values calculated from the P-values of the first input table.

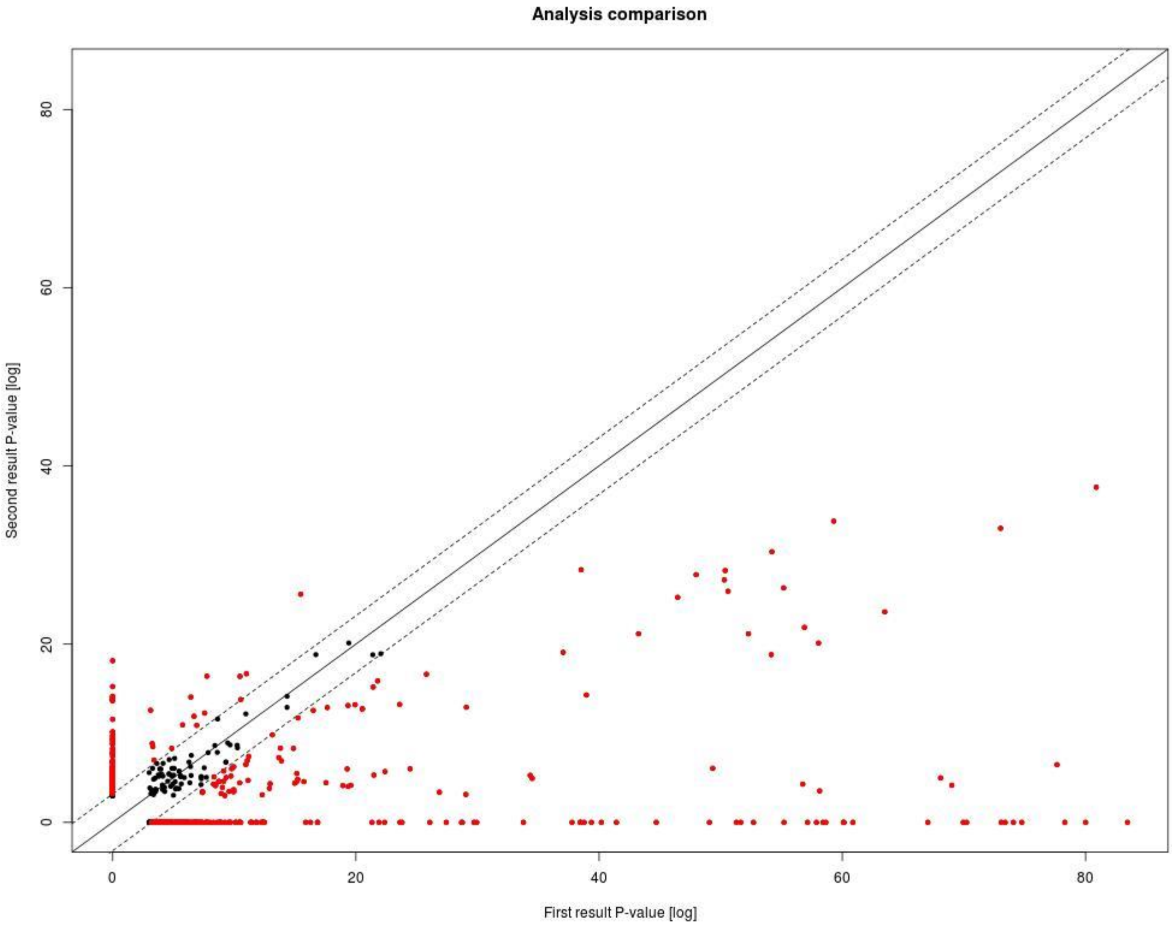
Second P-value (-log): This column contains the –log p-values calculated from the P-values of the second input table

Difference: Column 2 – Column 1

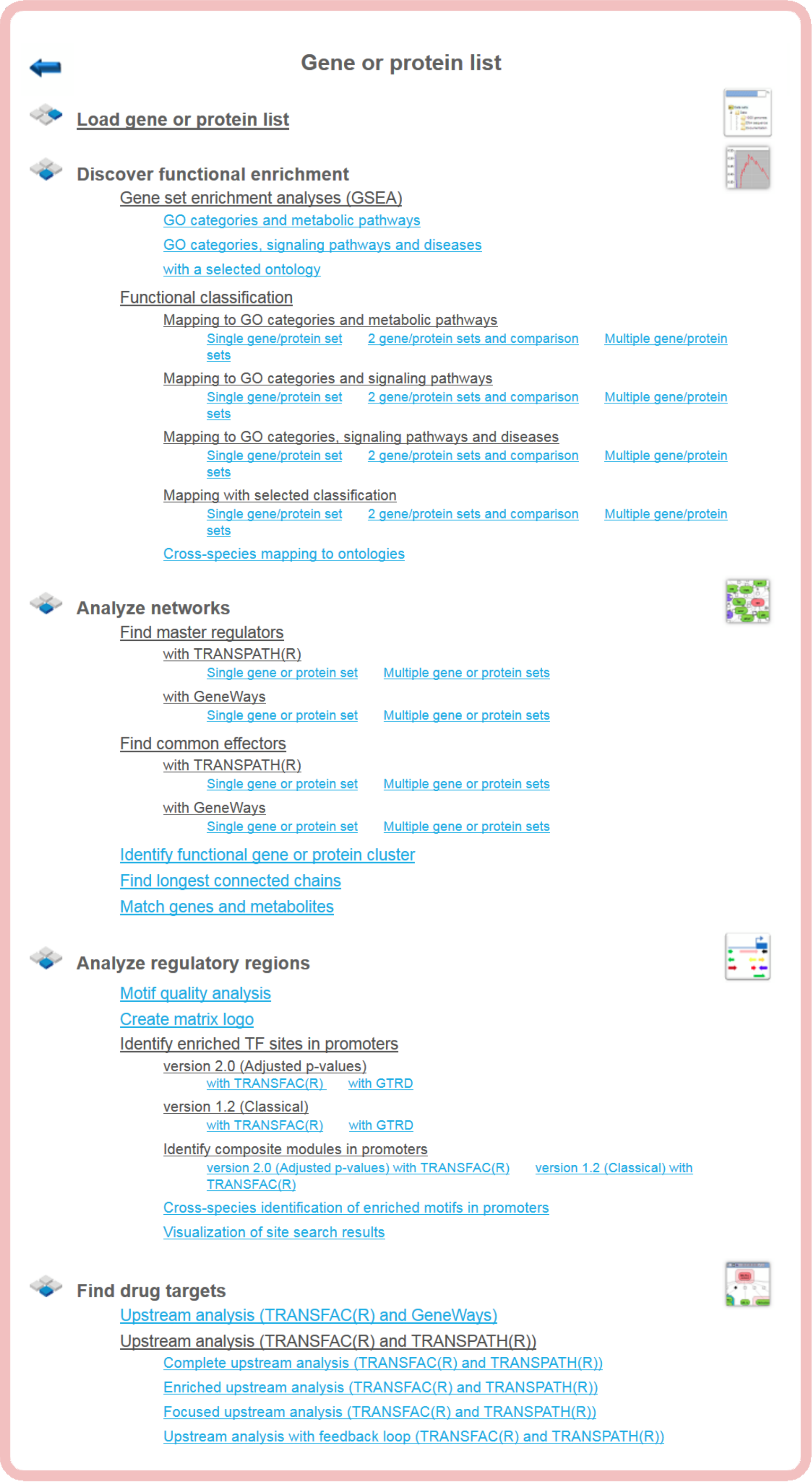
Difference P-value: This column contains the calculated P-values for the results in column 3 (Difference)

Difference FDR: This column contains the calculated FDR values for the results in column 3 (Difference)

The analysis comparison plot as shown below is a scatter plot of P-values on the log-scale together with the diagonal and the difference cutoffs at FDR < 0.05.



17. Gene or protein lists



Gene or protein list

Load gene or protein list

Discover functional enrichment

Gene set enrichment analyses (GSEA)

- GO categories and metabolic pathways
- GO categories, signaling pathways and diseases
- with a selected ontology

Functional classification

Mapping to GO categories and metabolic pathways

Single gene/protein set   2 gene/protein sets and comparison   Multiple gene/protein sets

Mapping to GO categories and signaling pathways

Single gene/protein set   2 gene/protein sets and comparison   Multiple gene/protein sets

Mapping to GO categories, signaling pathways and diseases

Single gene/protein set   2 gene/protein sets and comparison   Multiple gene/protein sets

Mapping with selected classification

Single gene/protein set   2 gene/protein sets and comparison   Multiple gene/protein sets

Cross-species mapping to ontologies

Analyze networks

Find master regulators

with TRANSPATH(R)

Single gene or protein set   Multiple gene or protein sets

with GeneWays

Single gene or protein set   Multiple gene or protein sets

Find common effectors

with TRANSPATH(R)

Single gene or protein set   Multiple gene or protein sets

with GeneWays

Single gene or protein set   Multiple gene or protein sets

Identify functional gene or protein cluster

Find longest connected chains

Match genes and metabolites

Analyze regulatory regions

Motif quality analysis

Create matrix logo

Identify enriched TF sites in promoters

version 2.0 (Adjusted p-values)

with TRANSFAC(R)   with GTRD

version 1.2 (Classical)

with TRANSFAC(R)   with GTRD

Identify composite modules in promoters

version 2.0 (Adjusted p-values) with TRANSFAC(R)   version 1.2 (Classical) with TRANSFAC(R)

Cross-species identification of enriched motifs in promoters

Visualization of site search results

Find drug targets

Upstream analysis (TRANSFAC(R) and GeneWays)

Upstream analysis (TRANSFAC(R) and TRANSPATH(R))

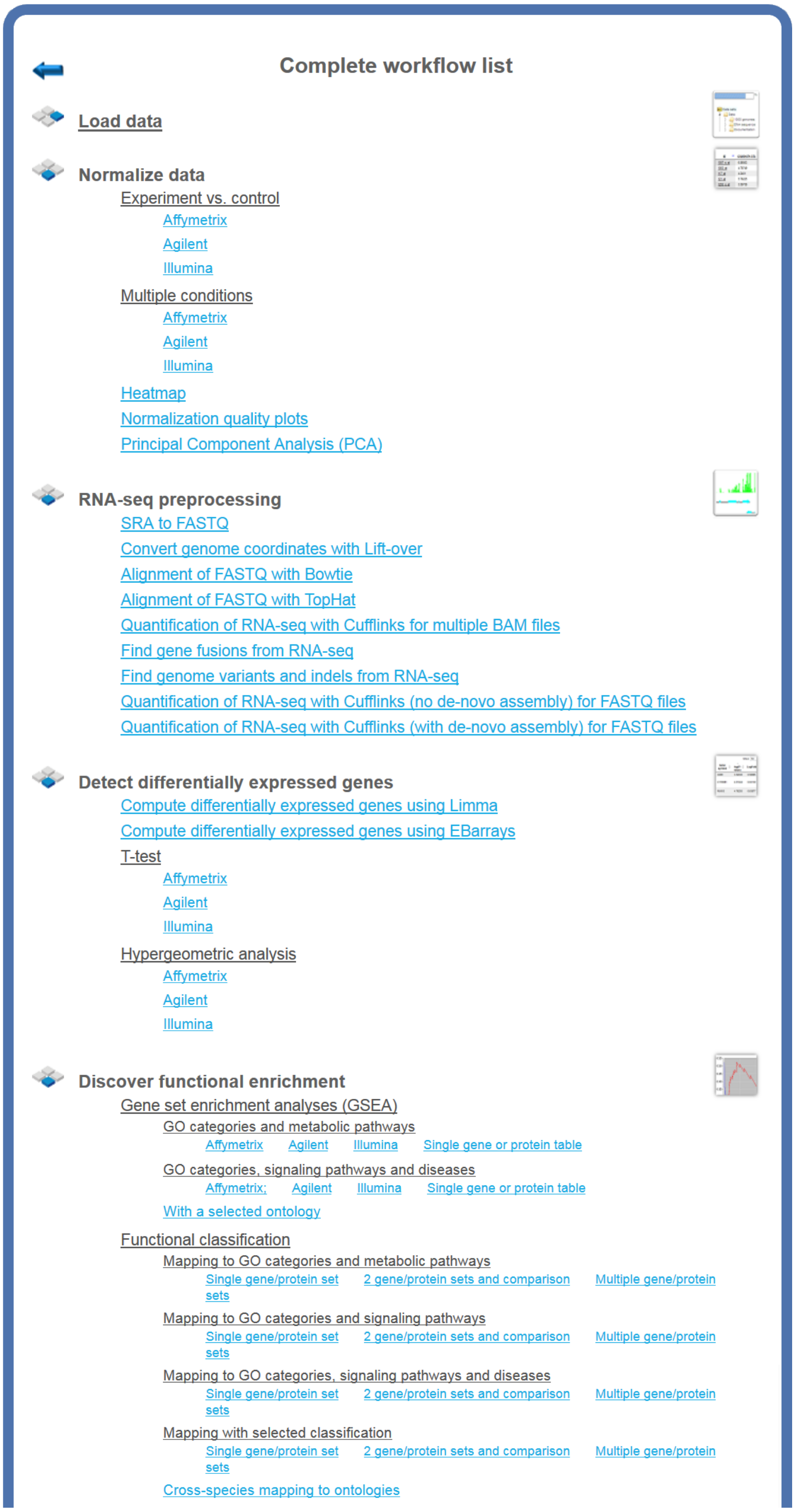
- Complete upstream analysis (TRANSFAC(R) and TRANSPATH(R))
- Enriched upstream analysis (TRANSFAC(R) and TRANSPATH(R))
- Focused upstream analysis (TRANSFAC(R) and TRANSPATH(R))
- Upstream analysis with feedback loop (TRANSFAC(R) and TRANSPATH(R))

All functions collected in this area have already been described in other sections. Kindly refer to those parts of this User Guide, as specified in the following:

Load gene or protein list	See Chapter 3
Discover functional enrichment	See Section 10.3
Analyze networks	See Section 5.1
Analyze regulatory regions	See Section 10.4
Find drug targets	See Chapter 11

18. Complete list of workflows

When you open this Area, the following complete listing of workflows available in the geneXplain platform will show up:

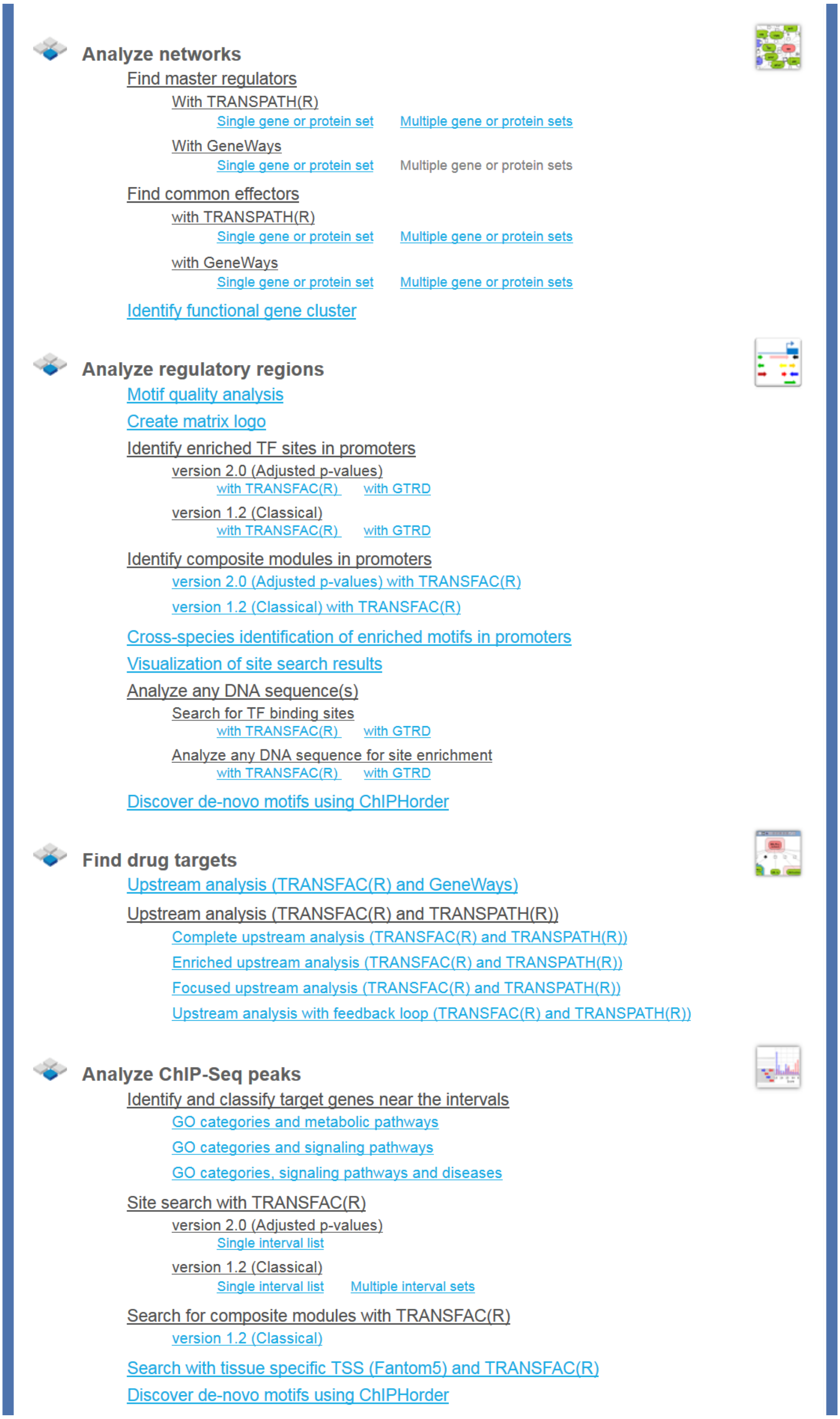


Complete workflow list

- Load data
- Normalize data
  - Experiment vs. control
    - Affymetrix
    - Agilent
    - Illumina
  - Multiple conditions
    - Affymetrix
    - Agilent
    - Illumina
  - Heatmap
  - Normalization quality plots
  - Principal Component Analysis (PCA)
- RNA-seq preprocessing
  - SRA to FASTQ
  - Convert genome coordinates with Lift-over
  - Alignment of FASTQ with Bowtie
  - Alignment of FASTQ with TopHat
  - Quantification of RNA-seq with Cufflinks for multiple BAM files
  - Find gene fusions from RNA-seq
  - Find genome variants and indels from RNA-seq
  - Quantification of RNA-seq with Cufflinks (no de-novo assembly) for FASTQ files
  - Quantification of RNA-seq with Cufflinks (with de-novo assembly) for FASTQ files
- Detect differentially expressed genes
  - Compute differentially expressed genes using Limma
  - Compute differentially expressed genes using EBarrays
  - T-test
    - Affymetrix
    - Agilent
    - Illumina
  - Hypergeometric analysis
    - Affymetrix
    - Agilent
    - Illumina
- Discover functional enrichment
  - Gene set enrichment analyses (GSEA)
    - GO categories and metabolic pathways
      - Affymetrix   Agilent   Illumina   Single gene or protein table
    - GO categories, signaling pathways and diseases
      - Affymetrix;   Agilent   Illumina   Single gene or protein table
    - With a selected ontology
  - Functional classification
    - Mapping to GO categories and metabolic pathways
      - Single gene/protein set   2 gene/protein sets and comparison   Multiple gene/protein sets
    - Mapping to GO categories and signaling pathways
      - Single gene/protein set   2 gene/protein sets and comparison   Multiple gene/protein sets
    - Mapping to GO categories, signaling pathways and diseases
      - Single gene/protein set   2 gene/protein sets and comparison   Multiple gene/protein sets
    - Mapping with selected classification
      - Single gene/protein set   2 gene/protein sets and comparison   Multiple gene/protein sets
    - Cross-species mapping to ontologies

(continues next page)

(continued)



Analyze networks

- Find master regulators
  - With TRANSPATH(R)
    - Single gene or protein set
    - Multiple gene or protein sets
  - With GeneWays
    - Single gene or protein set
    - Multiple gene or protein sets
- Find common effectors
  - with TRANSPATH(R)
    - Single gene or protein set
    - Multiple gene or protein sets
  - with GeneWays
    - Single gene or protein set
    - Multiple gene or protein sets
- Identify functional gene cluster

Analyze regulatory regions

- Motif quality analysis
- Create matrix logo
- Identify enriched TF sites in promoters
  - version 2.0 (Adjusted p-values)
    - with TRANSFAC(R)
    - with GTRD
  - version 1.2 (Classical)
    - with TRANSFAC(R)
    - with GTRD
- Identify composite modules in promoters
  - version 2.0 (Adjusted p-values) with TRANSFAC(R)
  - version 1.2 (Classical) with TRANSFAC(R)
- Cross-species identification of enriched motifs in promoters
- Visualization of site search results
- Analyze any DNA sequence(s)
  - Search for TF binding sites
    - with TRANSFAC(R)
    - with GTRD
  - Analyze any DNA sequence for site enrichment
    - with TRANSFAC(R)
    - with GTRD
- Discover de-novo motifs using ChIPHorder

Find drug targets

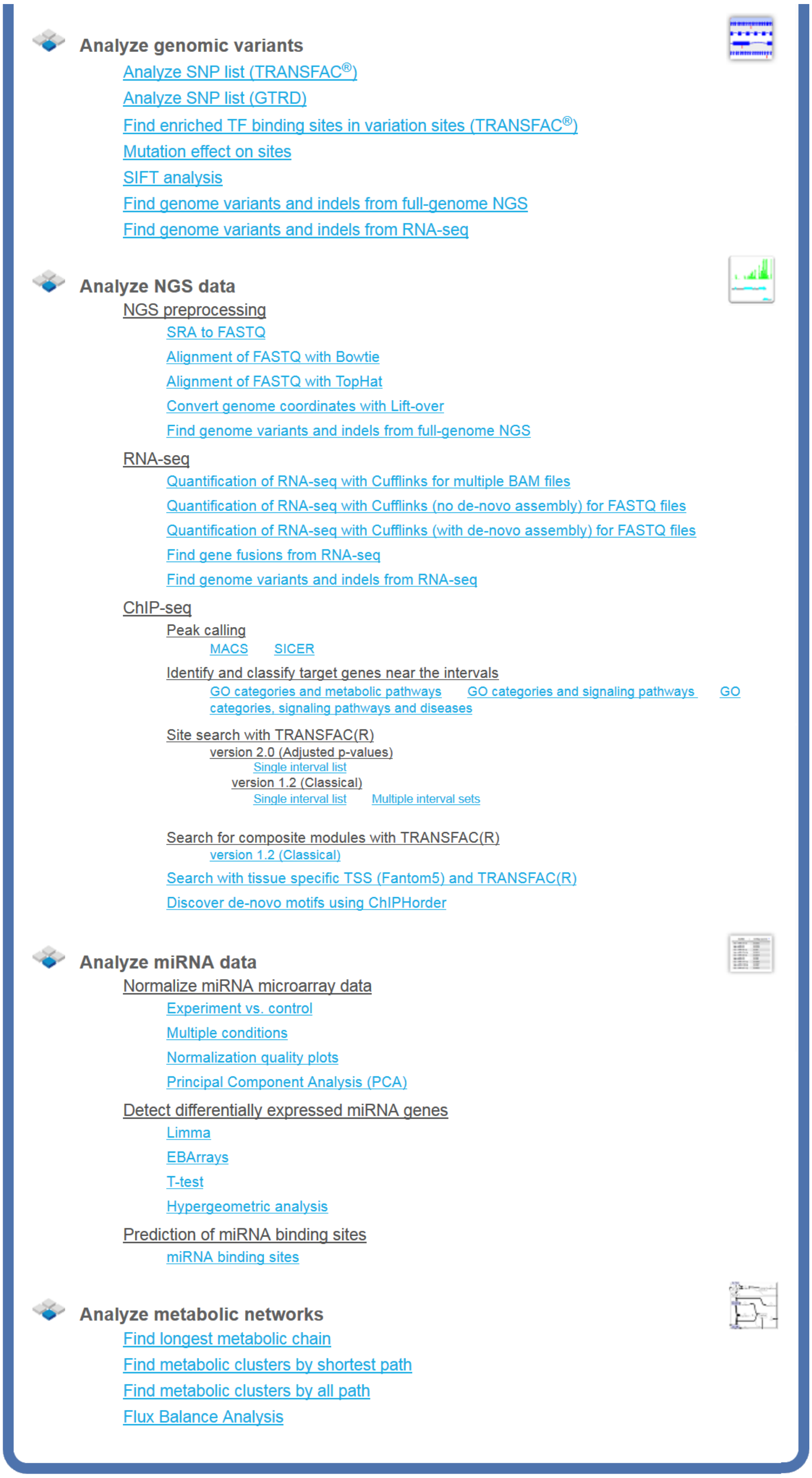
- Upstream analysis (TRANSFAC(R) and GeneWays)
- Upstream analysis (TRANSFAC(R) and TRANSPATH(R))
  - Complete upstream analysis (TRANSFAC(R) and TRANSPATH(R))
  - Enriched upstream analysis (TRANSFAC(R) and TRANSPATH(R))
  - Focused upstream analysis (TRANSFAC(R) and TRANSPATH(R))
  - Upstream analysis with feedback loop (TRANSFAC(R) and TRANSPATH(R))

Analyze ChIP-Seq peaks

- Identify and classify target genes near the intervals
  - GO categories and metabolic pathways
  - GO categories and signaling pathways
  - GO categories, signaling pathways and diseases
- Site search with TRANSFAC(R)
  - version 2.0 (Adjusted p-values)
    - Single interval list
  - version 1.2 (Classical)
    - Single interval list
    - Multiple interval sets
- Search for composite modules with TRANSFAC(R)
  - version 1.2 (Classical)
- Search with tissue specific TSS (Fantom5) and TRANSFAC(R)
- Discover de-novo motifs using ChIPHorder

(continues next page)

(continued)



Analyze genomic variants

- Analyze SNP list (TRANSFAC®)
- Analyze SNP list (GTRD)
- Find enriched TF binding sites in variation sites (TRANSFAC®)
- Mutation effect on sites
- SIFT analysis
- Find genome variants and indels from full-genome NGS
- Find genome variants and indels from RNA-seq

Analyze NGS data

- NGS preprocessing
  - SRA to FASTQ
  - Alignment of FASTQ with Bowtie
  - Alignment of FASTQ with TopHat
  - Convert genome coordinates with Lift-over
  - Find genome variants and indels from full-genome NGS
- RNA-seq
  - Quantification of RNA-seq with Cufflinks for multiple BAM files
  - Quantification of RNA-seq with Cufflinks (no de-novo assembly) for FASTQ files
  - Quantification of RNA-seq with Cufflinks (with de-novo assembly) for FASTQ files
  - Find gene fusions from RNA-seq
  - Find genome variants and indels from RNA-seq
- ChIP-seq
  - Peak calling
    - MACS SICER
  - Identify and classify target genes near the intervals
    - GO categories and metabolic pathways GO categories and signaling pathways GO categories, signaling pathways and diseases
  - Site search with TRANSFAC(R)
    - version 2.0 (Adjusted p-values)
      - Single interval list
    - version 1.2 (Classical)
      - Single interval list Multiple interval sets
  - Search for composite modules with TRANSFAC(R)
    - version 1.2 (Classical)
  - Search with tissue specific TSS (Fantom5) and TRANSFAC(R)
  - Discover de-novo motifs using ChIPHorder

Analyze miRNA data

- Normalize miRNA microarray data
  - Experiment vs. control
  - Multiple conditions
  - Normalization quality plots
  - Principal Component Analysis (PCA)
- Detect differentially expressed miRNA genes
  - Limma
  - EBArrays
  - T-test
  - Hypergeometric analysis
- Prediction of miRNA binding sites
  - miRNA binding sites

Analyze metabolic networks

- Find longest metabolic chain
- Find metabolic clusters by shortest path
- Find metabolic clusters by all path
- Flux Balance Analysis

For detailed explanation of their function and how to operate them, please refer to the previous chapters.

19. Working with the different databases

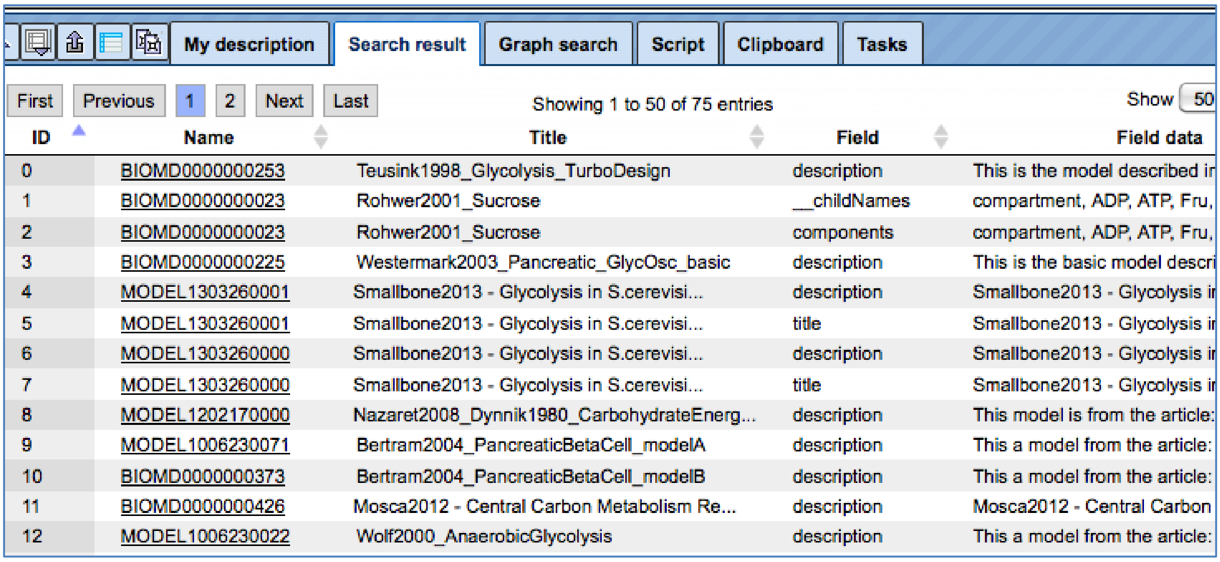
Working with one of the databases listed in the Tree Area involves several, frequently all four areas of your screen. In this chapter, we will briefly demonstrate how to operate the individual databases.

19.1. Biomodels

Biomodels is a source of mathematical models suitable for simulating biological processes. They have been compiled by the BioModels.Net project (http://www.biomodels.net/). These models are stored in the geneXplain platform as database Biomodels, subdirectory Diagrams, along with their graphical representations, parameters, simulation default values, etc. The present version is release 25, comprising 426 curated and 522 non-curated diagrams

You can access the Biomodels contents by browsing or searching. When you browse the list of models in the Tree Area (Databases > Biomodels > Diagrams), please note that only 50 of either 421 curated or 433 non-curated diagrams are displayed at once. Selecting one of the models is by double-clicking on the respective name.

Searching for a model starts from the Search tab in the Info Box. You may enter, for instance, *glycolysis* and start the search with the button 🔍. You will get back 75 entries, the contents displayed under the tab Search result in the Operations Field, the information given being essentially self-explaining: A short description of the model including references is given. You may open any model by a click on the corresponding Accession number in the last column.



Showing 1 to 50 of 75 entries

ID	Name	Title	Field	Field data
0	BIOMD0000000253	Teusink1998_Glycolysis_TurboDesign	description	This is the model described i
1	BIOMD0000000023	Rohwer2001_Sucrose	__childNames	compartment, ADP, ATP, Fru,
2	BIOMD0000000023	Rohwer2001_Sucrose	components	compartment, ADP, ATP, Fru,
3	BIOMD0000000225	Westermark2003_Pancreatic_GlycOsc_basic	description	This is the basic model descri
4	MODEL1303260001	Smallbone2013 - Glycolysis in S.cerevisi...	description	Smallbone2013 - Glycolysis i
5	MODEL1303260001	Smallbone2013 - Glycolysis in S.cerevisi...	title	Smallbone2013 - Glycolysis i
6	MODEL1303260000	Smallbone2013 - Glycolysis in S.cerevisi...	description	Smallbone2013 - Glycolysis i
7	MODEL1303260000	Smallbone2013 - Glycolysis in S.cerevisi...	title	Smallbone2013 - Glycolysis i
8	MODEL1202170000	Nazaret2008_Dynnik1980_CarbohydrateEnerg...	description	This model is from the article:
9	MODEL1006230071	Bertram2004_PancreaticBetaCell_modelA	description	This a model from the article:
10	BIOMD0000000373	Bertram2004_PancreaticBetaCell_modelB	description	This a model from the article:
11	BIOMD0000000426	Mosca2012 - Central Carbon Metabolism Re...	description	Mosca2012 - Central Carbon
12	MODEL1006230022	Wolf2000_AnaerobicGlycolysis	description	This a model from the article:

Having opened one of the diagrams by either method, the network schema will appear in the Work Space as well as in the Operations Field, tab Overview; see Section 23.3 for editing these diagrams. The tab Layout provides a number of options for changing the layout style of the diagram (see 21.2); none of these options will change the diagram contents. Clicking on an individual node will show information about this component and its role in this model within the Info Box (tab Info).

Many Biomodels have been optimized for dynamic simulation. Open, for instance, the subdirectory Diagrams and double-click *Goldbeter1991_MinMitOscil_ExplInact*. The diagram will open in the Work Space, the tab Info in the Info Box will show some details about the model, database links for the individual components. In the Operations Field, under the tab "My description", a detailed description of the model is given; it shows the original reference, its abstract and the PubMed link. Further information about how the model was constructed is added.

Under the tabs Simulation, you will find the default settings for the simulation, which can be changed before launching the simulation (▶). The simulation results will be graphically displayed in a new window, which can be saved as image using the Export button (⇧) in the Control Panel (E, see Chapter 2). For this, a number of formats are available.

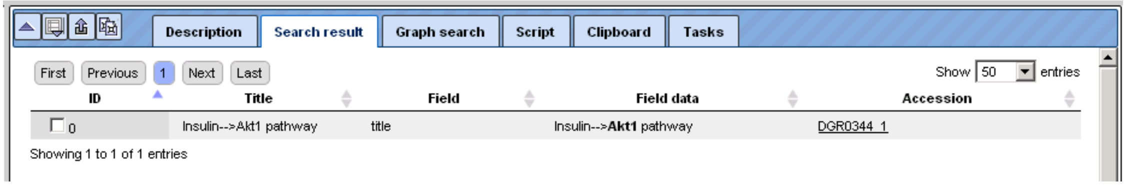
19.2. Biopath

Biopath is a collection of molecular pathways, manually annotated from original scientific publications and comprising biological models and diagrams. A more detailed description can be found in the Info Box, tab Info, after clicking on the term Biopath in the Tree Area, Databases.

User accessible entities are stored in the subdirectory *Diagrams* (presently 571).

This way, you may browse the database contents, receiving information about the individual entities in the Info Box, after a single click with the left mouse button on the corresponding entity. The Diagrams can be opened by double-click with the left mouse button, or by opening a context menu with the right mouse button and selecting "Open diagram". These diagram contain their components listed underneath as nodes (✳) and edges (↗), for which additional information can be retrieved in the same way as for the whole diagram. Many of them may be assigned to a hierarchy of compartments (🟨).

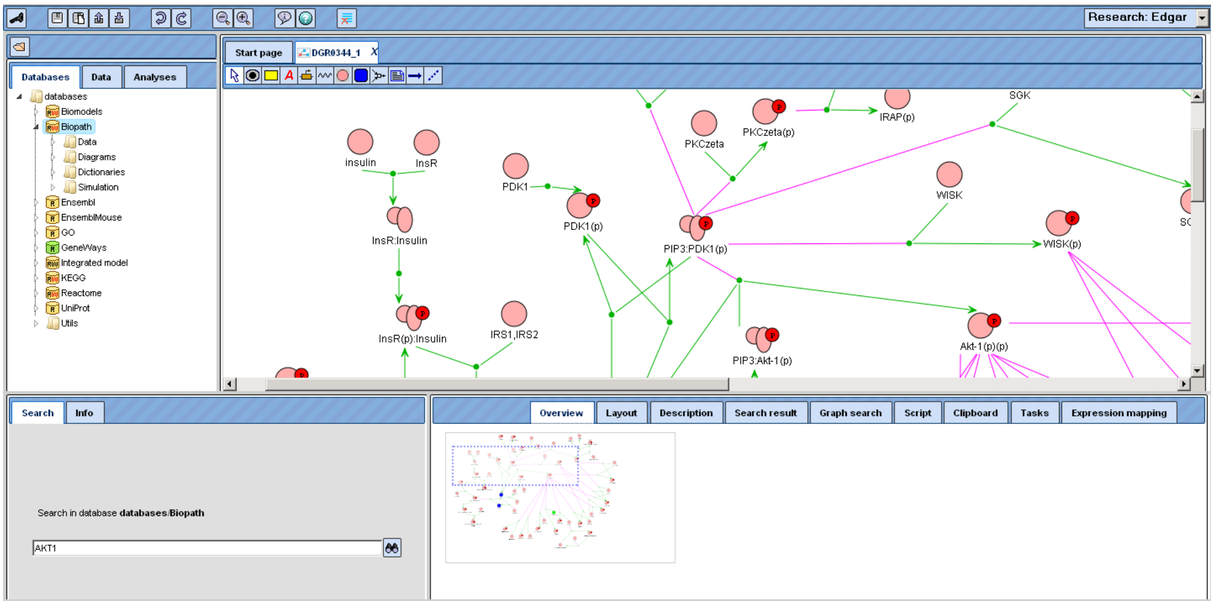
If you select the Biopath database for a search, you may enter a gene name into the field of the Search tab in the Info Box (try, for instance, AKT1). In the Operations Field, under the tab Search result, you will find the following hit:



ID	Title	Field	Field data	Accession
□ 0	Insulin-->Akt1 pathway	title	Insulin-->**Akt1** pathway	DGR0344_1

Showing 1 to 1 of 1 entries

It shows the molecule encoded by the gene AKT1 being involved in the insulin pathway. The hyperlinked Accession number here opens a pathway diagram from the Biopath database, which includes insulin and AKT1 as well as the path between them. The pathway is shown in full in the Work Space; an overview is depicted under the "Overview" tab of the Operations Field:



Search in database databases/Biopath

AKT1

The part of the diagram that is displayed in the Work Space is framed by the dotted blue line in the Overview, which can be shifted with the mouse, the Work Space adapting accordingly. You may also shift your mouse pointer over the Work Space: It turns into a hand, indicating that you can now shift the diagram section in the display by pressing the left mouse button and moving the mouse accordingly. The dotted blue rectangle in the "Overview" tab of the Operations Field will move accordingly.

Information about an individual node can be obtained in the Info Box after double-click on the corresponding symbol in the Work Space.

In the diagram, green arrows represent conversions, magenta edges catalytic effects. To facilitate overview in complex diagrams, individual edges are highlighted in light blue on mouse-over.

19.3. Ensembl

The Ensembl database provides annotation of genes from the Ensembl genome databases (http://www.ensembl.org/index.html).

Presently, there are human, mouse and rat Ensembl data included in the geneXplain platform, separately listed on the Databases tab. The following versions are provided:

Human:

- *Genome build hg19, version 72.37*
- *Genome build hg19, version 65.37*
- *Genome build hg18, version 61.37f*

Mouse:

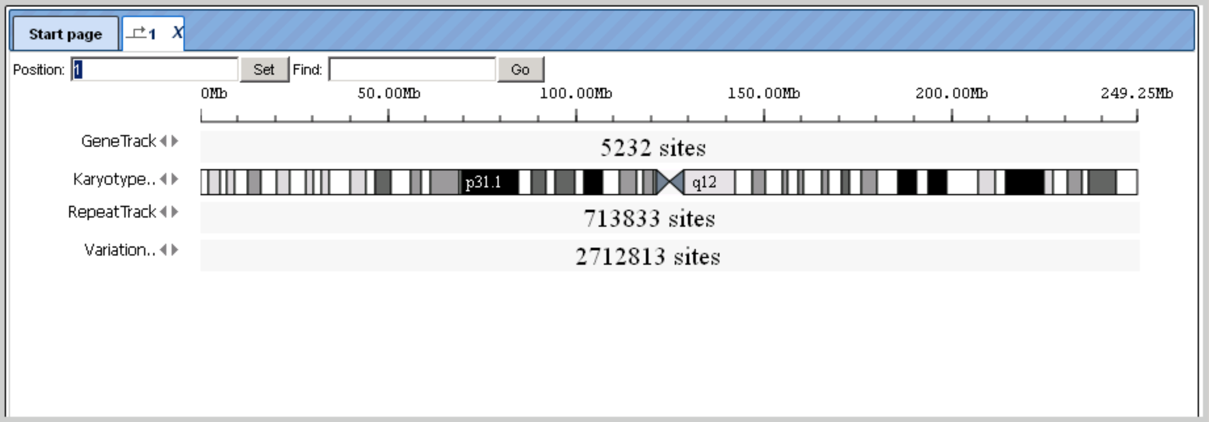
- *Genome build mm10, version 72.38*
- *Genome build mm9, version 65.37*

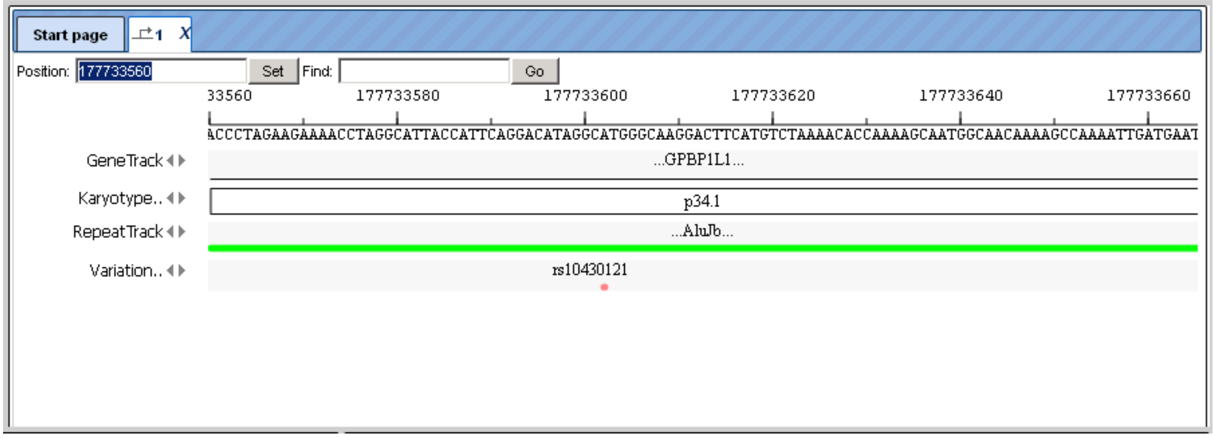
Rat:

- *Genome build rn5, version 71.5*

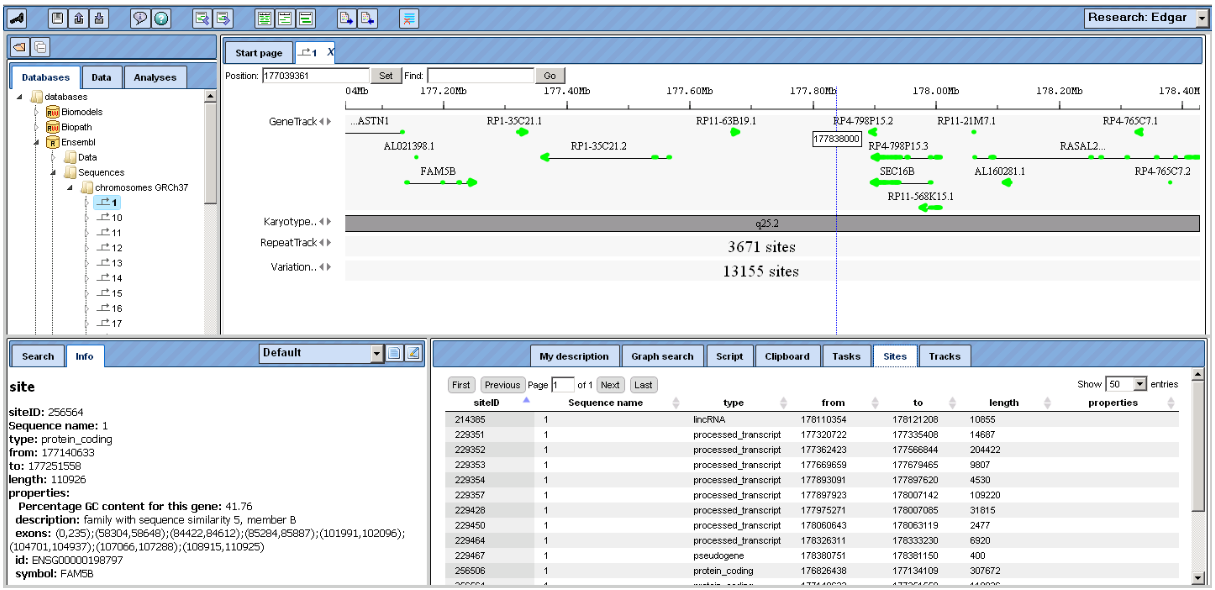
- *Genome build rn4, version 65.34*

In the subdirectories "Sequences/Chromosomes …" the individual chromosomes of the corresponding genome are stored. A double-click on an individual chromosome symbol opens the corresponding object with its annotation in the genomes browser of the geneXplain platform. Here, you can move the mouse pointer along the sequence to select a certain position. With the buttons and (zoom-out and zoom-in, resp.), you can go up to a level where you have the whole chromosome displayed or down to a level where you see the individual nucleotides. The most extreme views can also be directly selected with the buttons (overview) and (detail; see below). A moderate view is provided as "default" ().


Position: 1


Position: 177733560

At an intermediate resolution of, e.g., human chromosome 1, you see individual genes highlighted in the Work Space:



site

siteID: 256564
Sequence name: 1
type: protein_coding
from: 177140633
to: 177251558
length: 110926
properties:
Percentage GC content for this gene: 41.76
description: family with sequence similarity 5, member B
exons: (0,235);(58304,58648);(84422,84612);(85284,85887);(101991,102096);(104701,104937);(107066,107288);(108915,110925)
id: ENSG00000198797
symbol: FAM5B

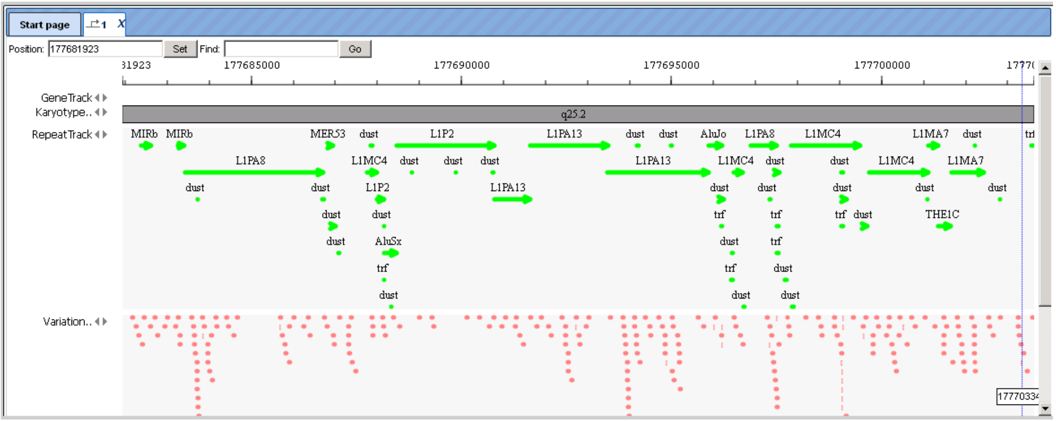
siteID	Sequence name	type	from	to	length	properties
214385	1	lincRNA	178110354	178121208	10855	
229351	1	processed_transcript	177320722	177335408	14687	
229352	1	processed_transcript	177362423	177566844	204422	
229353	1	processed_transcript	177669659	177679465	9807	
229354	1	processed_transcript	177893091	177897620	4530	
229357	1	processed_transcript	177897923	178007142	109220	
229428	1	processed_transcript	177975271	178007085	31815	
229450	1	processed_transcript	178060643	178063119	2477	
229464	1	processed_transcript	178326311	178333230	6920	
229467	1	pseudogene	178380751	178381150	400	
256506	1	protein_coding	176826438	177134109	307672	

The alignment of genes and other genomic data against a reference genome can be viewed as data tracks in genome browser. The GeneTrack shows the localization of primary transcripts including intron/exon structure and direction of transcription. Scrolling backwards and forward through the genome can be done using the arrowheads next to the track name, or with the buttons “Page backward” () and “Page forward” (), respectively.

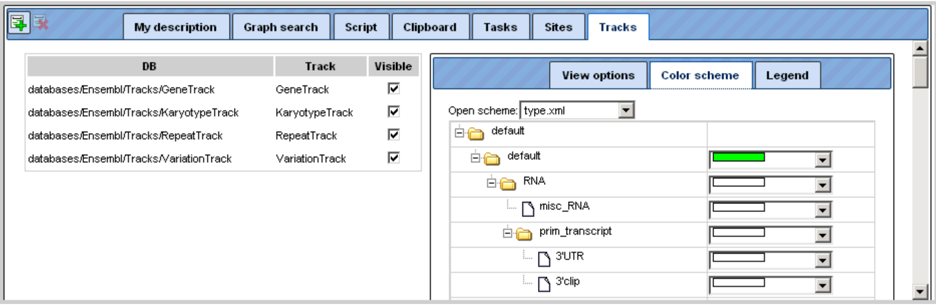
Information about individual genes is displayed in the Info Box after clicking on the respective gene symbol in the Work Space; on the Sites tab of the Operations Field, a detailed list of all functional sites

The KaryotypeTrack shows on which arm and in which karyotypic band(s) of the chromosome the present view is located.

RepeatTrack and VariationTrack comprise large numbers of sites scattered all over the chromosome. In a resolution like the one given above, only summarizing figures can be given. When zooming in, at a certain level of resolution, localization of individual repeats and their names will appear, and similarly sequence variations will show up in the Work Space, and they are listed in the Operations Field, under the Sites tab. Note that information about individual sites can be invoked in the Info Box only after a resolution has been adjusted that allows the display of their names in the Work Space, since the names are the clickable items.



The color scheme for the display in the Work Space can be changed in the Operations Field; here, on the Tracks tab, you can adjust the settings of the genome browser:



DB	Track	Visible
databases/Ensembl/Tracks/GeneTrack	GeneTrack	☑
databases/Ensembl/Tracks/KaryotypeTrack	KaryotypeTrack	☑
databases/Ensembl/Tracks/RepeatTrack	RepeatTrack	☑
databases/Ensembl/Tracks/VariationTrack	VariationTrack	☑

Open scheme: type.xml

- default
  - default 🟩
    - RNA
      - misc_RNA
      - prim_transcript
        - 3'UTR
        - 3'clip

Individual tracks can be removed from the display in the Work Space (e.g. by de-selecting them on this tab). The same way, they can be brought back. Another possibility is to open the subdirectory Tracks in the Tree Area, where all available tracks are listed with the symbol . Just drag-and-drop the track of interest to the Work Space will render the corresponding data amenable to the browser.

The Ensembl database sections are searchable in the same way as described before: Just click on the name of the respective database in the Tree Area, so that it receives a light-blue background and its path appears on the Search tab of the Info Box. Enter your search term, launch the search by pressing , and find your Search results on the respective tab on the Operations Field. Among the multiple hits, the one with the perfect match will be highlighted in bold.

19.4. Gene Ontology (GO)

Contents from Gene Ontology are imported into the geneXplain platform and are updated regularly. The present version is 06.2013.

GO is searchable in the same way as described before: Just click on the name GO in the Tree Area, so that it receives a light-blue background and its path appears on the Search tab of the Info Box. Enter your search term, launch the search by pressing , and find your Search results on the respective tab on the Operations Field. When clicking on linked descriptions, additional information will appear in the Info Box, Info tab.

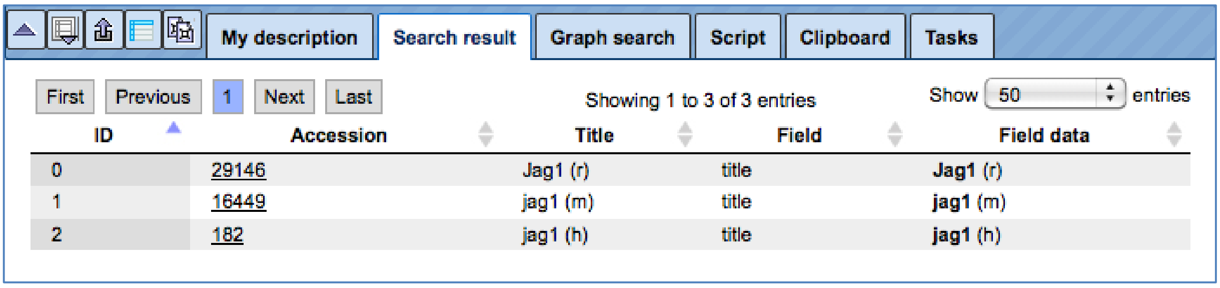
19.5. GeneWays

GeneWays is a database about genes and their functional interactions. The underlying data (version 7.0) have been retrieved from the original scientific literature by a sophisticated text mining system applied to more than 360,000 full text papers and of more than eight million publication abstracts [Iossifov I., Rodriguez-Esteban R., Mayzus I., Millen K.J., and Rzhetsky A. Looking at cerebellar malformations through text-mined interactomes of mice and humans. PLoS Comput Biol. 2009, 5:e1000559. PubMed PMID: 19893633].

The directory GeneWays/Data has two subdirectories: Genes and Reactions. When you click on any gene entry, information about the gene will be retrieved from the Entrez database. When you click on any reaction entry, you will find reaction title, links to Entrez for both incoming and outgoing molecules and the link to the PubMed entries of the corresponding publication in the Info Box.

A single click on the term GeneWays in the Tree Area / Databases suffices to indicate this data source in the Info Box, tab Search. You can insert your search term (e.g., a gene symbol) into the field underneath. Clicking on the icon 🔍 launches the search. The search routine scans for exact matches, but use of wildcards is possible. Thus, searching for elk* returns results for elk1, elk2p1, elk3, and elk4.

The results will be shown in the Operations Field, under the tab Search result. For instance, when searching in GeneWays for JAG1, the following result table will be displayed:



Showing 1 to 3 of 3 entries

ID	Accession	Title	Field	Field data
0	29146	Jag1 (r)	title	**Jag1** (r)
1	16449	jag1 (m)	title	**jag1** (m)
2	182	jag1 (h)	title	**jag1** (h)

The search term is highlighted in bold.

You may recognize that the numbers in the last column (**Accession**) are hyperlinked. When you click on them, information about the gene will be retrieved from the Entrez database at NCBI, displayed in a new window or tab, depending on your browser settings. Rows referring to reactions show a different type of accession number, they are linked to a PubMed entry of the corresponding publication.

19.6. Reactome

Reactome is a database that provides information on biological objects such as proteins, protein complexes, reactions etc. It is an open-source, open access, manually curated and peer-reviewed pathway database. Pathway annotations are authored by expert biologists in collaboration with the Reactome editorial staff, and cross-referenced to many bioinformatics databases. The contents of Reactome are copyright © 2003-2010

Cold Spring Harbor Laboratory (CSHL), Ontario Institute for Cancer Research (OICR) and the European Bioinformatics Institute (EBI).

So far 5222 diagrams have been imported from version 45 of this database into the geneXplain platform. They can be used as graphical and editable schemata. In the Tree Area, all components are listed in each diagram subdirectory.

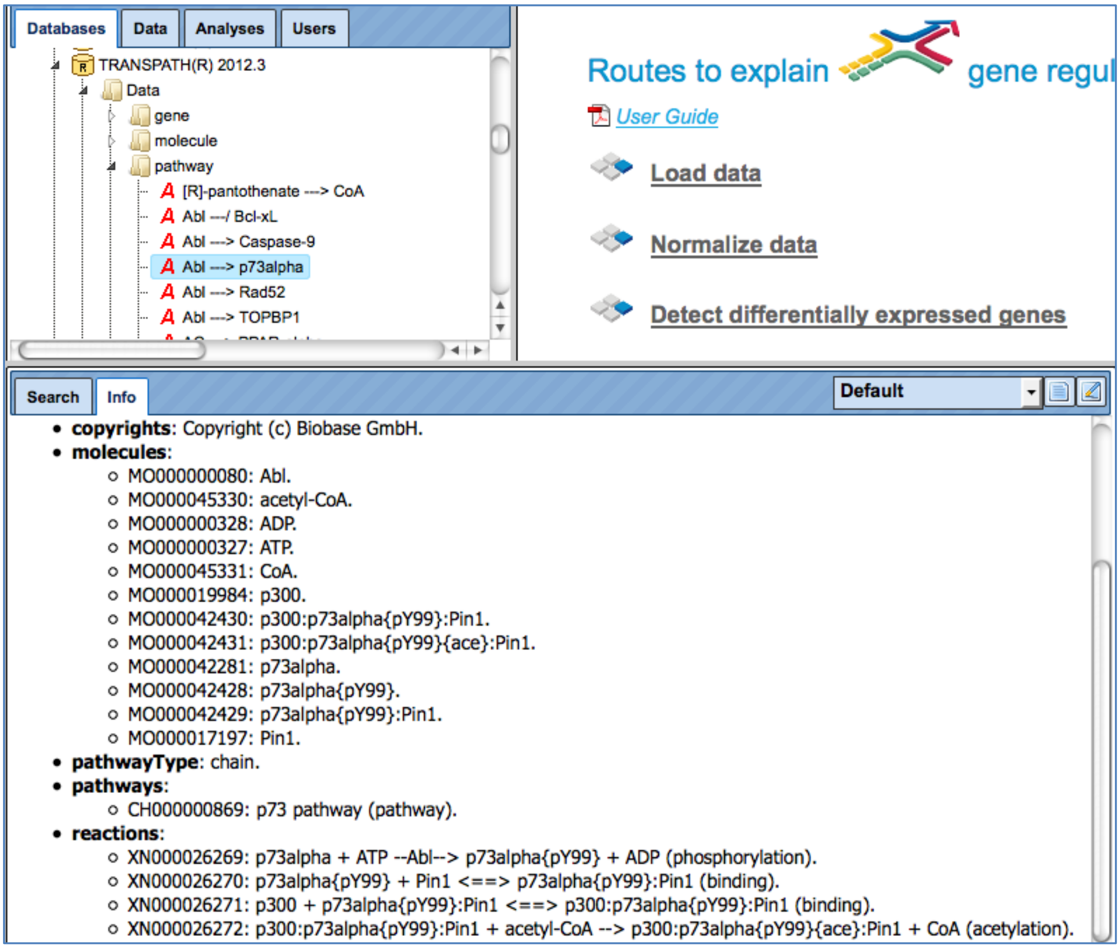
19.7. TRANSPATH®

TRANSPATH® is BIOBASE's database about signal transduction and metabolic pathways in human and the most important model organisms, mouse and rat. Its particular structure, which models signaling components and reactions in multiple hierarchical ways, makes TRANSPATH® a unique resource for both encyclopedic and modeling purposes.

Under the geneXplain platform, the TRANSPATH directory contains the following folders:

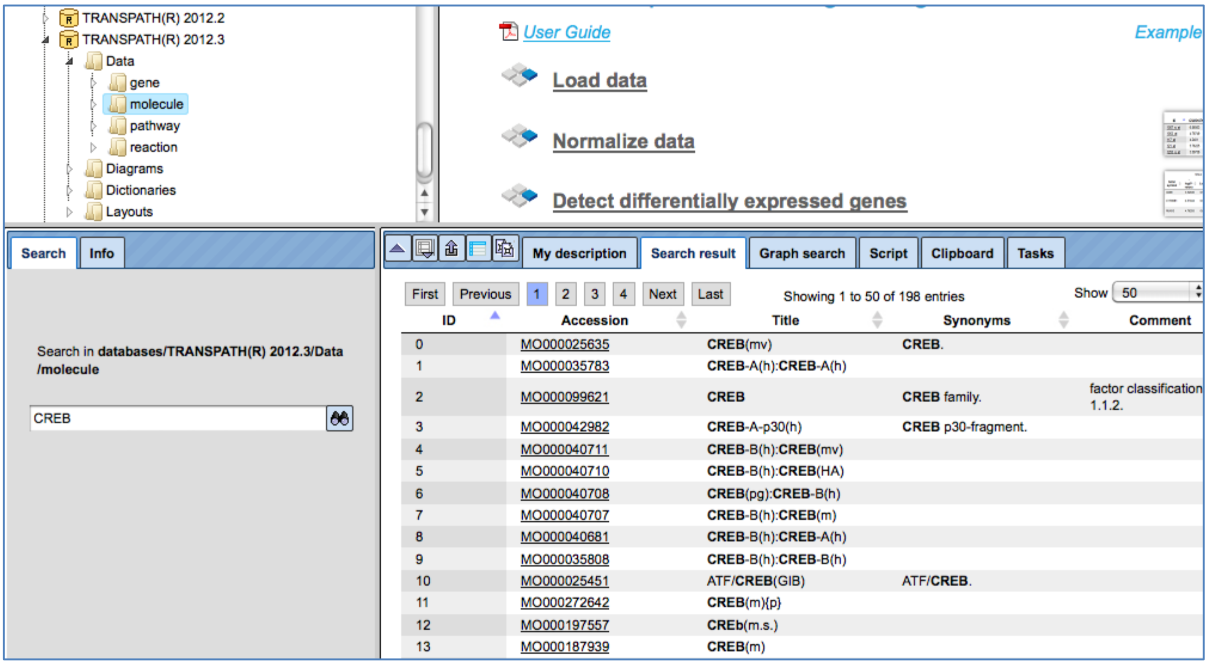
- *Data*
- *Diagrams*
- *Dictionaries*
- *Layouts*

Under "Data", browsable lists of objects, like genes and molecules, and of processes, like reactions and pathways, are stored. As usual, detailed database contents are displayed in the Info Box upon selecting individual entities from the lists by mouse click.



- **copyrights**: Copyright (c) Biobase GmbH.
- **molecules**:
  - MO000000080: Abl.
  - MO000045330: acetyl-CoA.
  - MO000000328: ADP.
  - MO000000327: ATP.
  - MO000045331: CoA.
  - MO000019984: p300.
  - MO000042430: p300:p73alpha{pY99}:Pin1.
  - MO000042431: p300:p73alpha{pY99}{ace}:Pin1.
  - MO000042281: p73alpha.
  - MO000042428: p73alpha{pY99}.
  - MO000042429: p73alpha{pY99}:Pin1.
  - MO000017197: Pin1.
- **pathwayType**: chain.
- **pathways**:
  - CH000000869: p73 pathway (pathway).
- **reactions**:
  - XN000026269: p73alpha + ATP --Abl--> p73alpha{pY99} + ADP (phosphorylation).
  - XN000026270: p73alpha{pY99} + Pin1 <==> p73alpha{pY99}:Pin1 (binding).
  - XN000026271: p300 + p73alpha{pY99}:Pin1 <==> p300:p73alpha{pY99}:Pin1 (binding).
  - XN000026272: p300:p73alpha{pY99}:Pin1 + acetyl-CoA --> p300:p73alpha{pY99}{ace}:Pin1 + CoA (acetylation).

As for other databases, the Search tab in the Info Box provides an easy search function to retrieve individual entities from the resource. The search results are displayed in the Operations Field, amenable to further activities.



Search in **databases/TRANSPATH(R) 2012.3/Data /molecule**

CREB

Showing 1 to 50 of 198 entries

ID	Accession	Title	Synonyms	Comment
0	MO000025635	**CREB**(mv)	**CREB**.	
1	MO000035783	**CREB**-A(h):**CREB**-A(h)		
2	MO000099621	**CREB**	**CREB** family.	factor classification 1.1.2.
3	MO000042982	**CREB**-A-p30(h)	**CREB** p30-fragment.	
4	MO000040711	**CREB**-B(h):**CREB**(mv)		
5	MO000040710	**CREB**-B(h):**CREB**(HA)		
6	MO000040708	**CREB**(pg):**CREB**-B(h)		
7	MO000040707	**CREB**-B(h):**CREB**(m)		
8	MO000040681	**CREB**-B(h):**CREB**-A(h)		
9	MO000035808	**CREB**-B(h):**CREB**-B(h)		
10	MO000025451	ATF/**CREB**(GIB)	ATF/**CREB**.	
11	MO000272642	**CREB**(m){p}		
12	MO000197557	**CREb**(m.s.)		
13	MO000187939	**CREB**(m)		

19.8. GTRD

GTRD (Gene Transcription Regulation Database) is a database of weight matrices to recognize transcription factor binding sites. The GTRD library of weight matrices consists of new matrices derived from ChIP-seq data by applying our multiple alignment method IPSmatrix. The IPSmatrix represents a modification of the previously published alignment method [Ananko E.A. *et al.* Recognition of interferon-inducible sites, promoters, and enhancers. BMC Bioinformatics 2007, 8:56. PMID: 17309789]. Each individual matrix is derived from an analysis of the corresponding set of TF-binding regions obtained from a set of raw ChIP-seq data. It is assumed that:

some TF-binding regions do not contain TF-binding sites;

strand orientation and exact location of TF-binding sites in TF-binding regions are unknown.

The IPSmatrix relies on the Gibbs sampling principle and our conception of the Individual Probability Score (IPS) where IPS represents the extension of the common matrix score. In addition to the cores of TF-binding sites, the IPSs take advantage of nucleotide contents of both flanks of site cores.

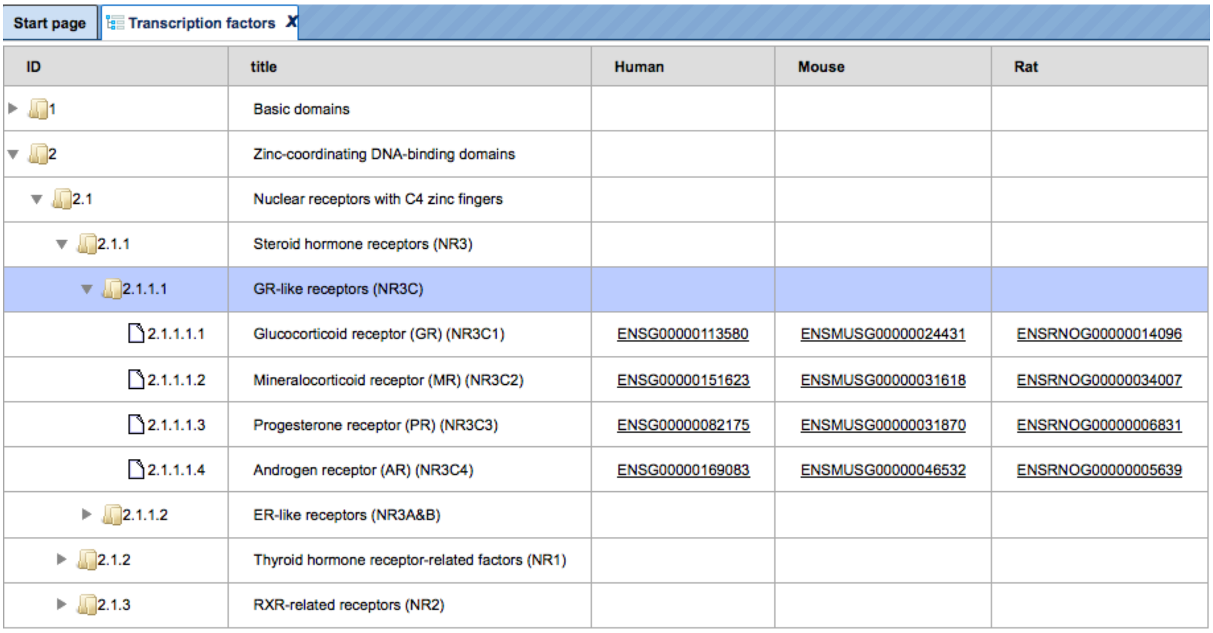
When several ChIP-seq datasets are available for a single TF, several matrices were calculated from which the optimal matrix was selected by comparing the corresponding Receiver Operator Characteristic (ROC) curves.

In the GTRD directory of the geneXplain platform, the folder "matrices" provides the list of matrices generated in the way described. You may also wish to include at this point a matrix collection provided by the UniProbe database.

Under "profiles", the matrices are stored along with thresholds that are to be applied when using the matrices for TFBS recognition. All these "profiles" are grouped according to three different thresholds (strong, moderate and weak).

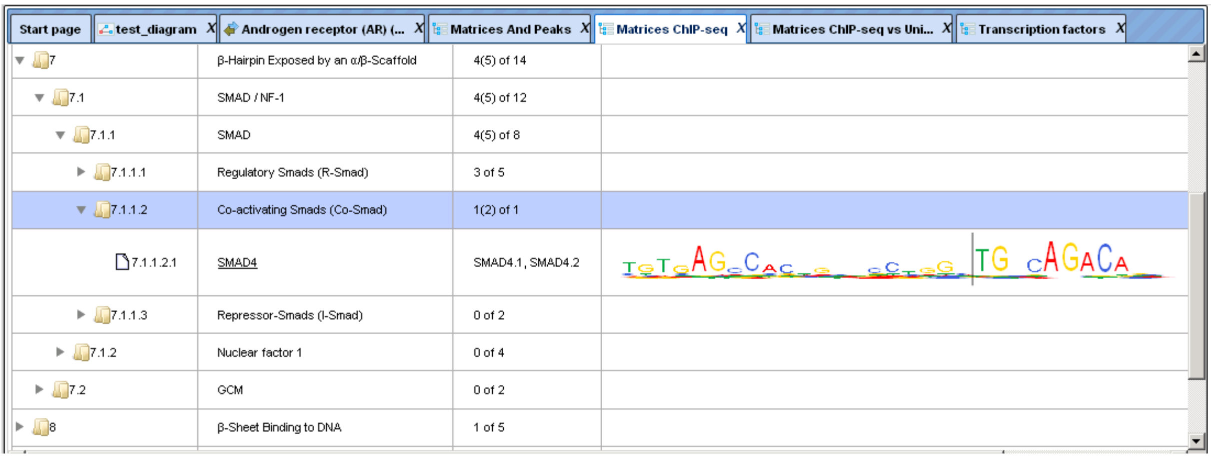
The directory "tracks" provides ChIP-seq (or ChIP-chip) data, e.g. from GEO (Gene Expression Omnibus database at NCBI) in a format that renders them suitable for visualization in the Genome Browser.

Under "views", available matrices (PWMs), underlying ChIP-seq experiments or the TFs themselves can be browsed. They are arranged according to the most recent classification of human TFs (http://www.edgar-wingender.de/huTF_classification.html) and extended to mouse and rat homologs.



ID	title	Human	Mouse	Rat
1	Basic domains			
2	Zinc-coordinating DNA-binding domains			
2.1	Nuclear receptors with C4 zinc fingers			
2.1.1	Steroid hormone receptors (NR3)			
2.1.1.1	GR-like receptors (NR3C)			
2.1.1.1.1	Glucocorticoid receptor (GR) (NR3C1)	ENSG00000113580	ENSMUSG00000024431	ENSRNOG00000014096
2.1.1.1.2	Mineralocorticoid receptor (MR) (NR3C2)	ENSG00000151623	ENSMUSG00000031618	ENSRNOG00000034007
2.1.1.1.3	Progesterone receptor (PR) (NR3C3)	ENSG00000082175	ENSMUSG00000031870	ENSRNOG00000006831
2.1.1.1.4	Androgen receptor (AR) (NR3C4)	ENSG00000169083	ENSMUSG00000046532	ENSRNOG00000005639
2.1.1.2	ER-like receptors (NR3A&B)			
2.1.2	Thyroid hormone receptor-related factors (NR1)			
2.1.3	RXR-related receptors (NR2)			

In cases where several ChIP-seq experiments are available to deduce matrices, all respective logo plots are given when navigating down to the level of individual TFs.

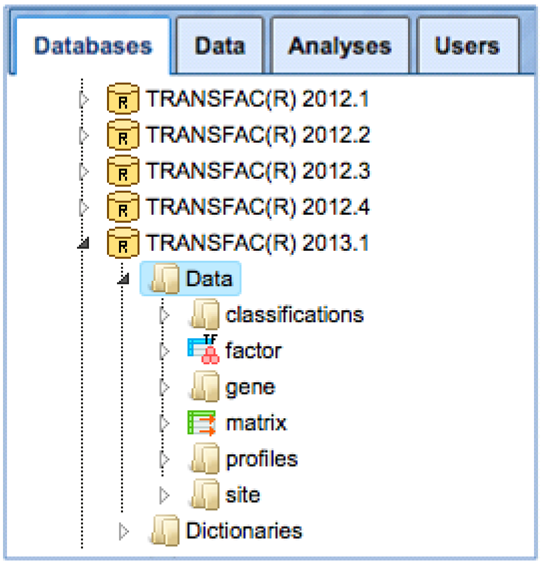


7	β-Hairpin Exposed by an α/β-Scaffold	4(5) of 14	
7.1	SMAD / NF-1	4(5) of 12	
7.1.1	SMAD	4(5) of 8	
7.1.1.1	Regulatory Smads (R-Smad)	3 of 5	
7.1.1.2	Co-activating Smads (Co-Smad)	1(2) of 1	
7.1.1.2.1	SMAD4	SMAD4.1, SMAD4.2	
7.1.1.3	Repressor-Smads (I-Smad)	0 of 2	
7.1.2	Nuclear factor 1	0 of 4	
7.2	GCM	0 of 2	
8	β-Sheet Binding to DNA	1 of 5	

19.9. TRANSFAC®

TRANSFAC® is the most comprehensive database on eukaryotic transcription factors (TFs), their genomic DNA-binding sites and DNA-binding profiles. It is a commercial database, maintained and distributed by BIOBASE GmbH as well as, in most countries, also offered for licensing by geneXplain. A valid license is thus required to access the contents of this database.

In the TRANSFAC folder labeled according to the respective release, there is the data subdirectory with the areas “classifications”, “factor”, “gene”, “matrix”, “profiles” and “site”:



TRANSFAC(R) 2012.1
TRANSFAC(R) 2012.2
TRANSFAC(R) 2012.3
TRANSFAC(R) 2012.4
TRANSFAC(R) 2013.1
Data
classifications
factor
gene
matrix
profiles
site
Dictionaries

Under “classifications”, the class definitions of DNA-binding domains of eukaryotic TFs are listed with their accession numbers (such as C0001 for “zinc finger”). Clicking on any item invokes further information in the Info Box. The same holds for taxa (of biological species), and for the TF classification as it was basically established in 1999.

By clicking on the folder “factor”, you can browse all TF accession numbers, with detailed information about any selected TF appearing in the Info Box. The same applies to the folder “matrix”, which lists all PWMs available in the actual database release, and the Info Box displaying the matrix and the corresponding logo plot.

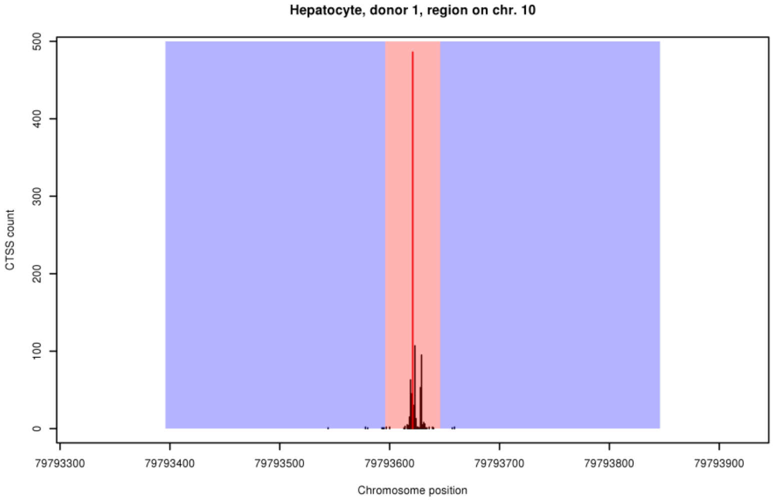
Under “profiles”, matrix collections are given for TFs that are known to play a role in a certain biological context, as defined by the profile name.

The TRANSFAC folder “Dictionaries” contains mostly files with links to other databases for internal use of the program.

19.10. Fantom5 cell-type and tissue-specific transcription start sites

The Fantom5 databases (Fantom5-Cell and Fantom5-Tissue) provide cell- and tissue-type specific transcription start site (TSS) annotations derived from CAGE measurements of the functional annotation of the mammalian genome 5 (FANTOM5) project (FANTOM Consortium, Nature 507:462-470, 2014). The databases cover 171 cell types and 121 tissue types (plus one type of *default TSSs* from Ensembl), respectively.

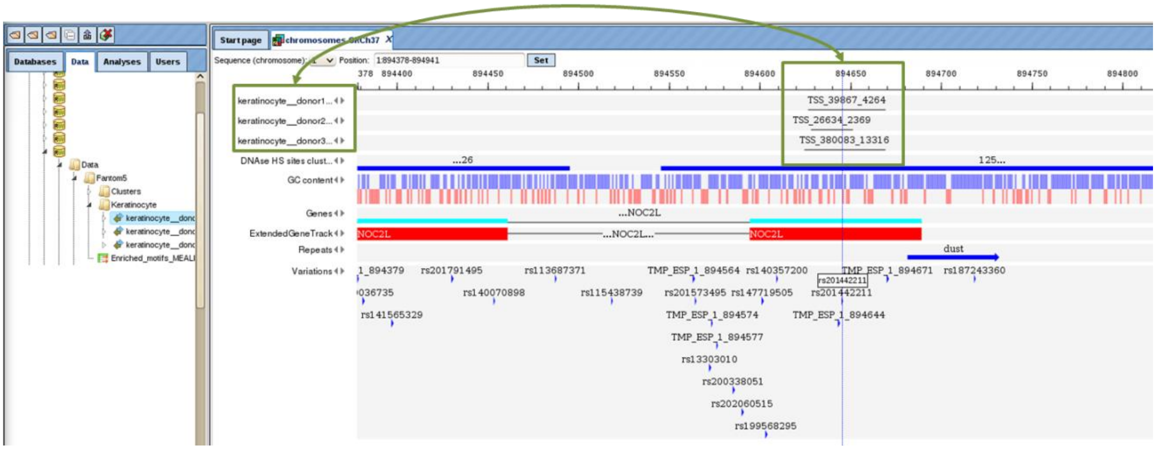
TSSs were inferred by a two-step process. The data for inference were the number of Cage TSSs (CTSSs) that had been mapped experimentally to genomic locations. First, a set of CTSS clusters was obtained by a sliding-window method similar to the one described by Strbenac et al. (BMC Genomics 5:S9, 2013). As illustrated in the figure below, this method used estimates of the rate of CTSS “hits” per nucleotide to calculate a statistical significance for the putative TSS (red line) on the basis of the Poisson distribution. For this, local background CTSS rates were calculated from flanking regions (blue areas) of a putative TSS (red line), excluding its direct vicinity (red area).



In the second step, we further refined the initial CTSS clusters using a Hidden Markov Model (HMM). The HMM modelled CTSS hits in TSS regions as well as in the genomic background by negative binomial distributions whose parameters, as well as their transitions, were estimated from the first set of CTSS clusters.

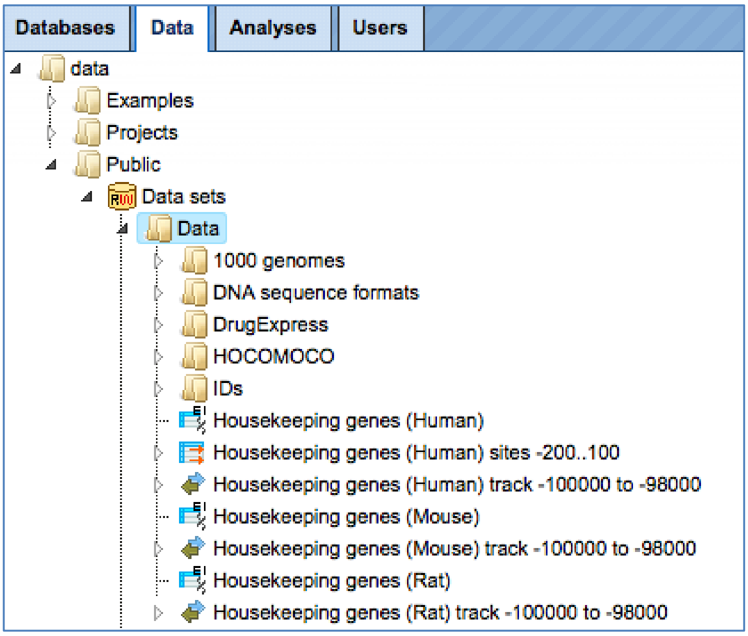
The HMM-based CTSS clusters from corresponding cell or tissue samples (the Fantom5 project collected tissue or cell samples from several donors) were eventually grouped into sets of overlapping clusters and annotated with high CTSS mark as consensus TSS.

The derived databases are now available in the geneXplain platform denoted as Fantom5-Cell and Fantom5-Tissue. Condition-specific TSSs can be extracted for gene sets using the tool named "Create tissue-specific promoter track".



19.11. Other data sets

In addition to databases, there are several datasets available within the platform that might be interesting to make use of in particular analyses. The datasets are available under the tab *Data* in the folder *Public*, within the project *Data sets*.



data
- Examples
- Projects
- Public
  - Data sets
    - Data
      - 1000 genomes
      - DNA sequence formats
      - DrugExpress
      - HOCOMOCO
      - IDs
      - Housekeeping genes (Human)
      - Housekeeping genes (Human) sites -200..100
      - Housekeeping genes (Human) track -100000 to -98000
      - Housekeeping genes (Mouse)
      - Housekeeping genes (Mouse) track -100000 to -98000
      - Housekeeping genes (Rat)
      - Housekeeping genes (Rat) track -100000 to -98000

19.11.1. DrugExpress – genome-wide transcriptional signatures of drug response

The folder DrugExpress contains sets of genes that significantly change their expression in response to treatment by different drugs. It originates from the Connectivity Map (also known as cmap) project developed at the Broad Institute, USA, http://www.broadinstitute.org/cmap/.

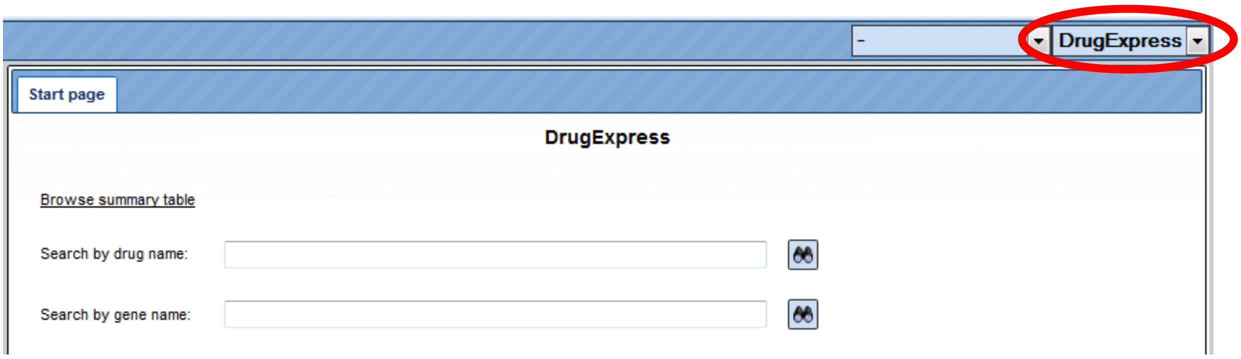
In this collection, we identified 321 compounds that could be mapped to Drugbank (http://www.drugbank.ca/), and also have known target genes in Drugbank. The following steps were performed:

- *Data normalization with affy package in R with parameters: method="quantiles", bgcorrect.method="rma", pmcorrect.method="pmonly", summary.method="liwong".*
- *Gene IDs were converted into Ensembl IDs using Convert table analysis.*
- *Up- and down-regulated genes were identified applying the following criteria: (LogFoldChange >= 0.7) and (p_value <= 0.05) for up-regulated genes, and (LogFoldChange <= -0.7) and (p_value <= 0.05) for down-regulated genes. LogFoldChange is a logarithm of fold change with base 2 of gene expression in treated versus untreated cells.*

The DrugExpress folder contains 321 gene tables for up-regulated and 321 gene tables for down-regulated genes for each chemical compound or drug, altogether 642 gene tables. The names of the files correspond to the names of the chemical compounds. Along with the individual files, there are 17 subfolders, in which the files are grouped according to the classification of the respective drugs (for example, “adrenergic antagonist”).

Search with DrugExpress Database

To search within the DrugExpress database you have to switch from **Default** mode to **DrugExpress,** mode using the drop-down menu at the top right corner (see picture below). The input mask of the search page appears automatically in the main Workspace.



DrugExpress

Browse summary table

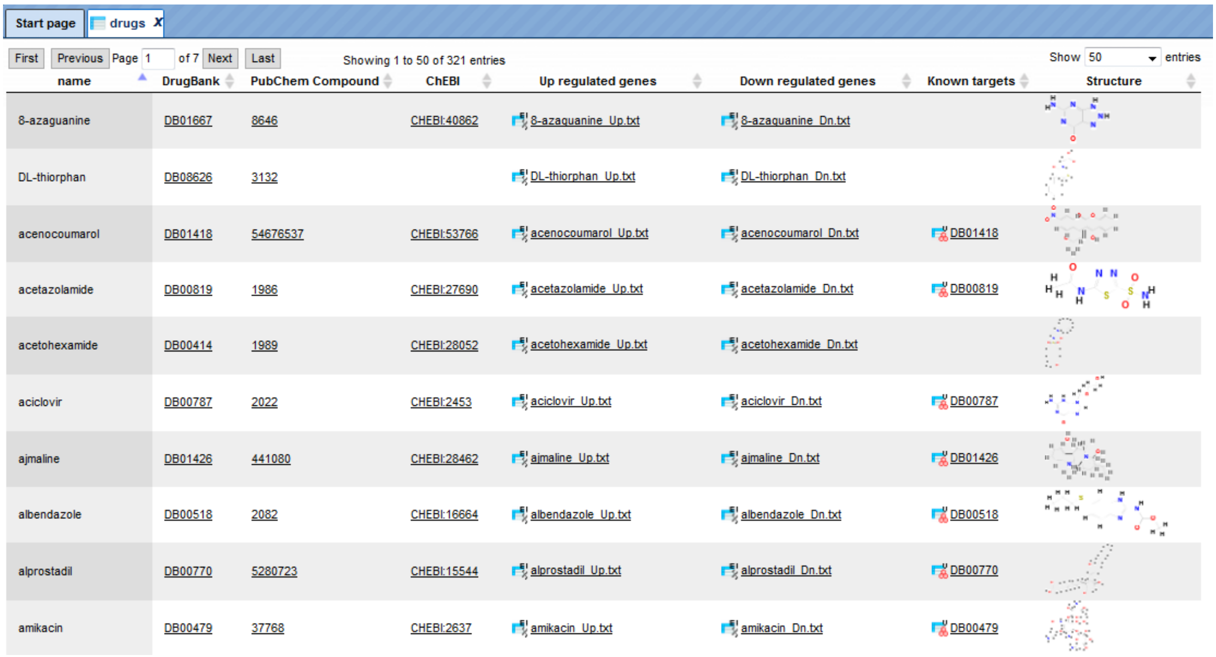
Search by drug name:

Search by gene name:

Browsing the **summary** drug **table** can be done by clicking the link "Browse summary table". The table with 321 entries, corresponding to the represented drugs, will be automatically opened as shown below. The columns of the table are sorted alphabetically. The links to the following external databases are provided:

Drug Bank (http://www.drugbank.ca/drugs),
PubChem Compound (https://www.ncbi.nlm.nih.gov/pccompound),
ChEBI (http://www.ebi.ac.uk/chebi/init.do).

The next columns provide links to the gene lists of all **up-** and **down-regulated genes** in response to drug treatment. The **Known targets** for every drug is shown in the next column. A **Structure** of the drug is visible in the last column of the table.



Page 1 of 7

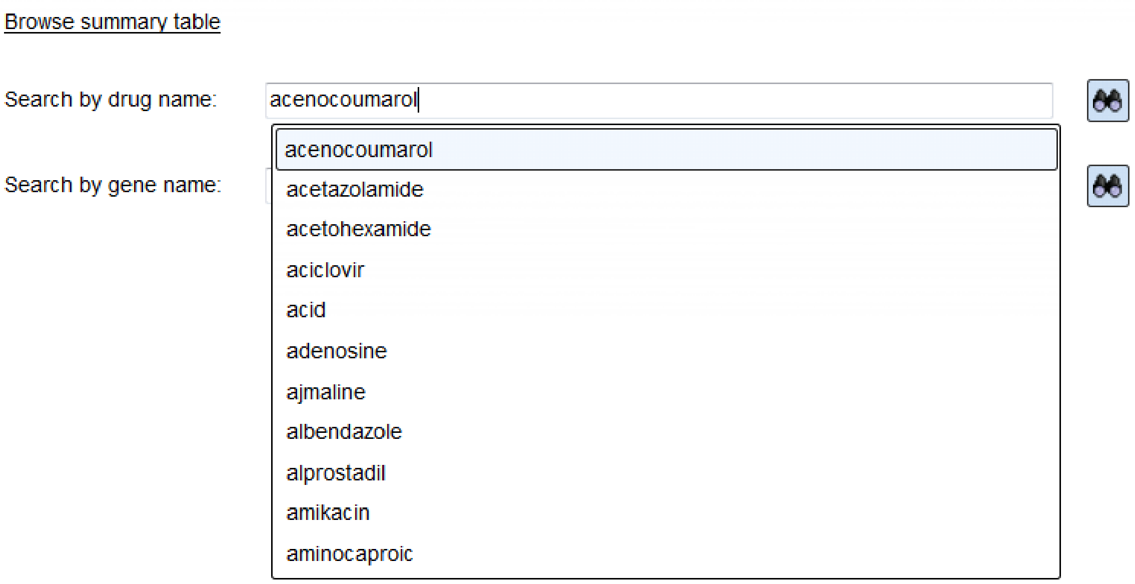
Showing 1 to 50 of 321 entries

Show 50 entries

name	DrugBank	PubChem Compound	ChEBI	Up regulated genes	Down regulated genes	Known targets	Structure
8-azaguanine	DB01667	8646	CHEBI:40862	8-azaguanine_Up.txt	8-azaguanine_Dn.txt		
DL-thiorphan	DB08626	3132		DL-thiorphan_Up.txt	DL-thiorphan_Dn.txt		
acenocoumarol	DB01418	54676537	CHEBI:53766	acenocoumarol_Up.txt	acenocoumarol_Dn.txt	DB01418	
acetazolamide	DB00819	1986	CHEBI:27690	acetazolamide_Up.txt	acetazolamide_Dn.txt	DB00819	
acetohexamide	DB00414	1989	CHEBI:28052	acetohexamide_Up.txt	acetohexamide_Dn.txt		
aciclovir	DB00787	2022	CHEBI:2453	aciclovir_Up.txt	aciclovir_Dn.txt	DB00787	
ajmaline	DB01426	441080	CHEBI:28462	ajmaline_Up.txt	ajmaline_Dn.txt	DB01426	
albendazole	DB00518	2082	CHEBI:16664	albendazole_Up.txt	albendazole_Dn.txt	DB00518	
alprostadil	DB00770	5280723	CHEBI:15544	alprostadil_Up.txt	alprostadil_Dn.txt	DB00770	
amikacin	DB00479	37768	CHEBI:2637	amikacin_Up.txt	amikacin_Dn.txt	DB00479	

Let's close the summary table to return to the search form of the DrugExpress database.

To **search by the drug name,** you can type in your search term and press .



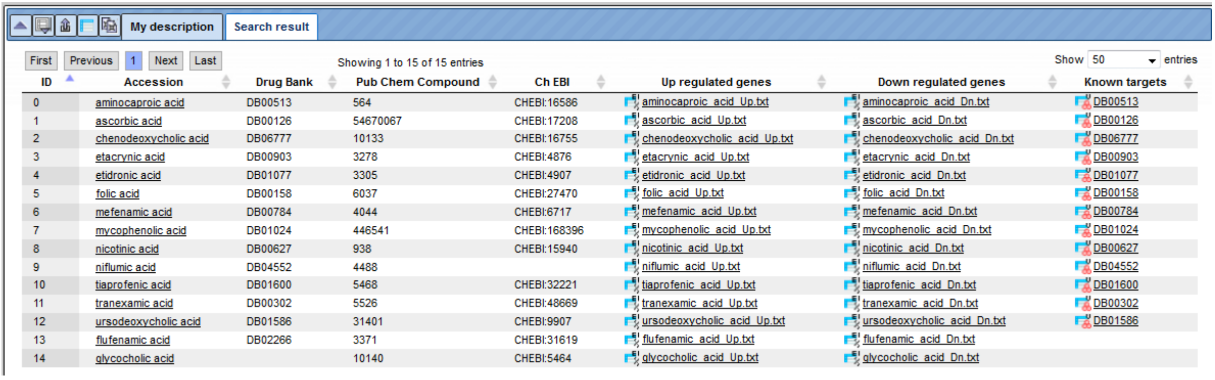
Browse summary table

Search by drug name: acenocoumarol

- acenocoumarol
- acetazolamide
- acetohexamide
- aciclovir
- acid
- adenosine
- ajmaline
- albendazole
- alprostadil
- amikacin
- aminocaproic

Search by gene name:

Let’s consider this example. Searching for “acid” returns 15 search results in the database with all columns described above.

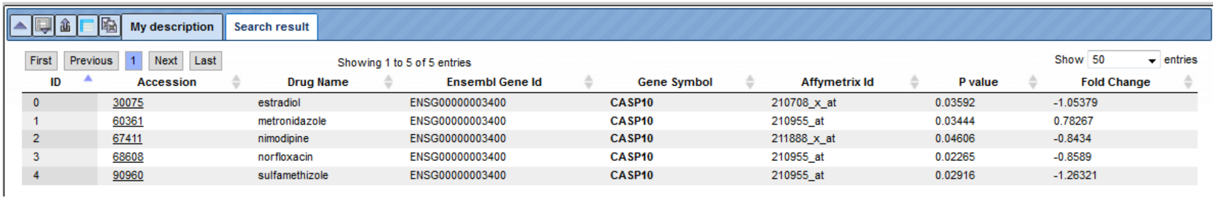


Showing 1 to 15 of 15 entries

ID	Accession	Drug Bank	Pub Chem Compound	Ch EBI	Up regulated genes	Down regulated genes	Known targets
0	aminocaproic acid	DB00513	564	CHEBI:16586	aminocaproic_acid_Up.txt	aminocaproic_acid_Dn.txt	DB00513
1	ascorbic acid	DB00126	54670067	CHEBI:17208	ascorbic_acid_Up.txt	ascorbic_acid_Dn.txt	DB00126
2	chenodeoxycholic acid	DB06777	10133	CHEBI:16755	chenodeoxycholic_acid_Up.txt	chenodeoxycholic_acid_Dn.txt	DB06777
3	etacrynic acid	DB00903	3278	CHEBI:4876	etacrynic_acid_Up.txt	etacrynic_acid_Dn.txt	DB00903
4	etidronic acid	DB01077	3305	CHEBI:4907	etidronic_acid_Up.txt	etidronic_acid_Dn.txt	DB01077
5	folic acid	DB00158	6037	CHEBI:27470	folic_acid_Up.txt	folic_acid_Dn.txt	DB00158
6	mefenamic acid	DB00784	4044	CHEBI:6717	mefenamic_acid_Up.txt	mefenamic_acid_Dn.txt	DB00784
7	mycophenolic acid	DB01024	446541	CHEBI:168396	mycophenolic_acid_Up.txt	mycophenolic_acid_Dn.txt	DB01024
8	nicotinic acid	DB00627	938	CHEBI:15940	nicotinic_acid_Up.txt	nicotinic_acid_Dn.txt	DB00627
9	niflumic acid	DB04552	4488		niflumic_acid_Up.txt	niflumic_acid_Dn.txt	DB04552
10	tiaprofenic acid	DB01600	5468	CHEBI:32221	tiaprofenic_acid_Up.txt	tiaprofenic_acid_Dn.txt	DB01600
11	tranexamic acid	DB00302	5526	CHEBI:48669	tranexamic_acid_Up.txt	tranexamic_acid_Dn.txt	DB00302
12	ursodeoxycholic acid	DB01586	31401	CHEBI:9907	ursodeoxycholic_acid_Up.txt	ursodeoxycholic_acid_Dn.txt	DB01586
13	flufenamic acid	DB02266	3371	CHEBI:31619	flufenamic_acid_Up.txt	flufenamic_acid_Dn.txt	
14	glycocholic acid		10140	CHEBI:5464	glycocholic_acid_Up.txt	glycocholic_acid_Dn.txt	

To **search** within the DrugExpress database **by the gene name** please type in your search term and press button.

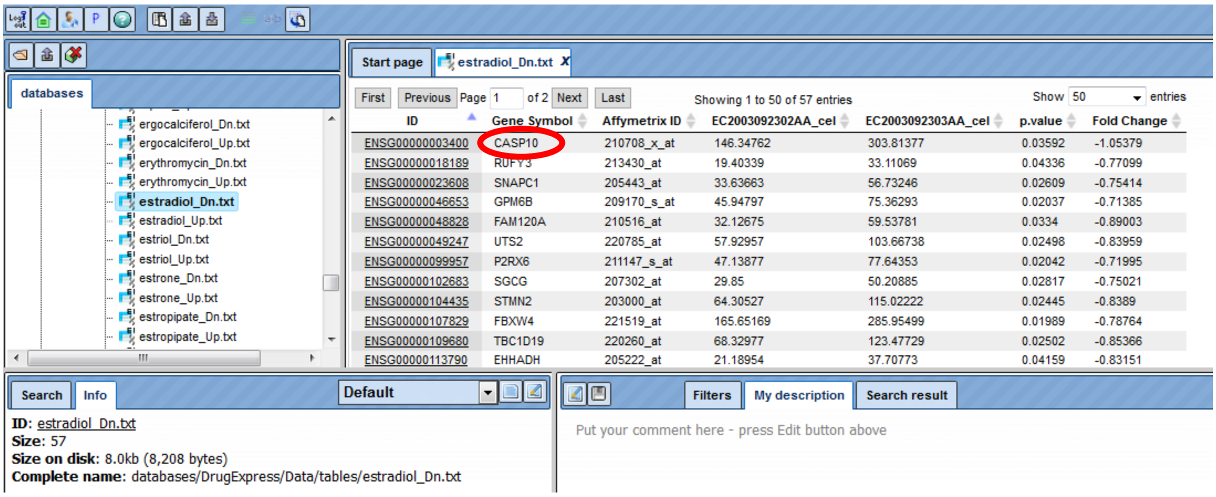
Example. Searching for Caspase10 (casp10) returns 5 entries. The resulting table contains the following columns, the **Accession** numbers of the linked drug, **Drug Name**, **ENSEMBL Gene Id**, **Gene Symbol**, **Affymetrix Id**, **P_value** and the **Fold Change** of gene expression in the drug treatment experiment.



Showing 1 to 5 of 5 entries

ID	Accession	Drug Name	Ensembl Gene Id	Gene Symbol	Affymetrix Id	P value	Fold Change
0	30075	estradiol	ENSG00000003400	**CASP10**	210708_x_at	0.03592	-1.05379
1	60361	metronidazole	ENSG00000003400	**CASP10**	210955_at	0.03444	0.78267
2	67411	nimodipine	ENSG00000003400	**CASP10**	211888_x_at	0.04606	-0.8434
3	68608	norfloxacin	ENSG00000003400	**CASP10**	210955_at	0.02265	-0.8589
4	90960	sulfamethizole	ENSG00000003400	**CASP10**	210955_at	0.02916	-1.26321

To find out which other genes are regulated in response to the same drug, follow the link from the drug accession number. This opens the whole table with genes that significantly change their expression in response to the treatment.



Showing 1 to 50 of 57 entries

ID	Gene Symbol	Affymetrix ID	EC2003092302AA_cel	EC2003092303AA_cel	p.value	Fold Change
ENSG00000003400	CASP10	210708_x_at	146.34762	303.81377	0.03592	-1.05379
ENSG00000018189	RUFY3	213430_at	19.40339	33.11069	0.04336	-0.77099
ENSG00000023608	SNAPC1	205443_at	33.63663	56.73246	0.02609	-0.75414
ENSG00000046653	GPM6B	209170_s_at	45.94797	75.36293	0.02037	-0.71385
ENSG00000048828	FAM120A	210516_at	32.12675	59.53781	0.0334	-0.89003
ENSG00000049247	UTS2	220785_at	57.92957	103.66738	0.02498	-0.83959
ENSG00000099957	P2RX6	211147_s_at	47.13877	77.64353	0.02042	-0.71995
ENSG00000102683	SGCG	207302_at	29.85	50.20885	0.02817	-0.75021
ENSG00000104435	STMN2	203000_at	64.30527	115.02222	0.02445	-0.8389
ENSG00000107829	FBXW4	221519_at	165.65169	285.95499	0.01989	-0.78764
ENSG00000109680	TBC1D19	220260_at	68.32977	123.47729	0.02502	-0.85366
ENSG00000113790	EHHADH	205222_at	21.18954	37.70773	0.04159	-0.83151

ID: estradiol_Dn.txt
Size: 57
Size on disk: 8.0kb (8,208 bytes)
Complete name: databases/DrugExpress/Data/tables/estradiol_Dn.txt

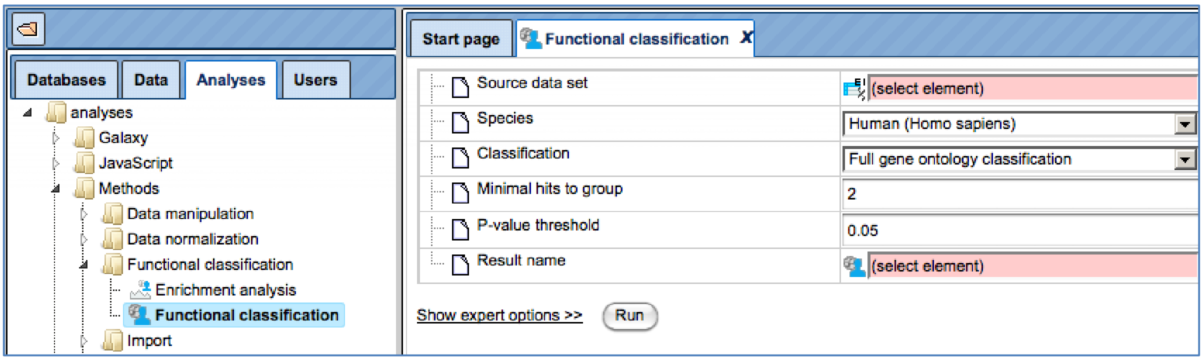
Put your comment here - press Edit button above

Functional analysis applying DrugExpress, the input form

The classification of drugs can be also applied to any gene signatures, for example to the genes differentially expressed under certain disease conditions. Mapping of disease-specific gene signatures to the gene signatures of drug responses and the similarities identified may result in intriguing suggestions which drugs can be potentially used for disease treatment.

The DrugExpress dataset can be used by two methods, *Functional classification* and *Enrichment analysis*, as a user-specific ontology, the so-called **Repository folder**. Any input gene or protein table will be classified using the DrugExpress data. 17 subfolders and 642 gene tables will be used as categories for the classification of the input table. We will exemplify this in the following.

Step 1. Open Functional classification analysis ( ), under the tab Analyses, folder Methods, subfolder Functional classification.



Source data set	(select element)
Species	Human (Homo sapiens)
Classification	Full gene ontology classification
Minimal hits to group	2
P-value threshold	0.05
Result name	(select element)

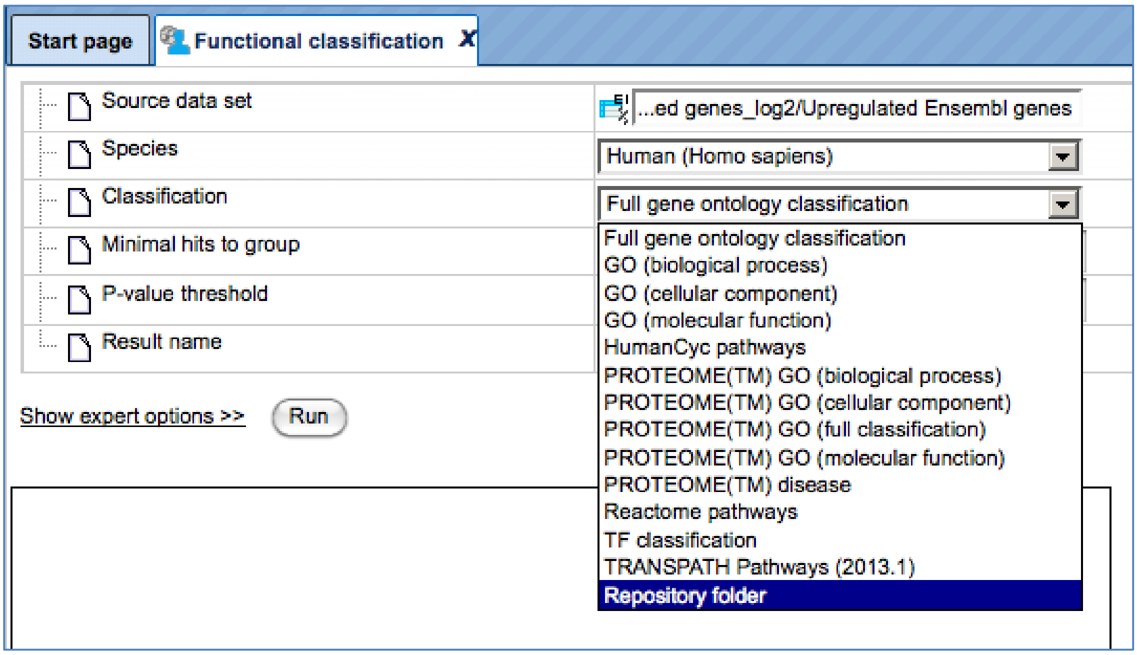
Step 2. Select your input gene table with Ensembl IDs and the corresponding species in the input fields, **Source data set** and **Species**, respectively.

Please note that if the table you plan to classify has other IDs, you first need to convert it into a table with Ensembl IDs. This can be done using the *Convert table* ( ) function located at analyses/Methods/Data manipulation/Convert table.

For this example, the following input table is used:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Experiment%20normalized%20(RMA)%20(Differentially%20expressed%20genes%20Affy)/Upregulated%20Ensembl%20genes

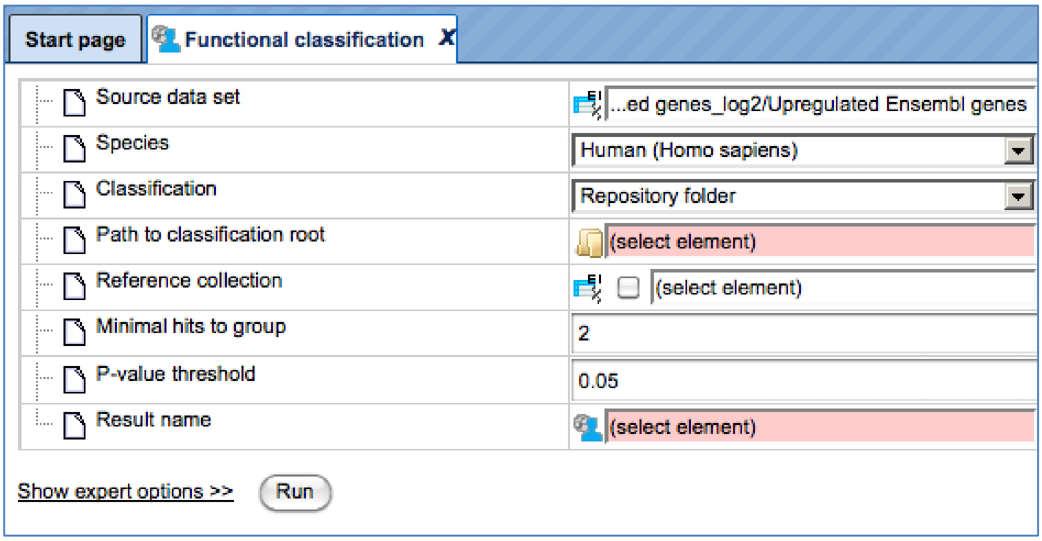
Step 3. In the **Classification** field, select the option **Repository folder** from the drop-down menu, as shown below.



Source data set	...ed genes_log2/Upregulated Ensembl genes
Species	Human (Homo sapiens)
Classification	Full gene ontology classification
Minimal hits to group	
P-value threshold	
Result name	

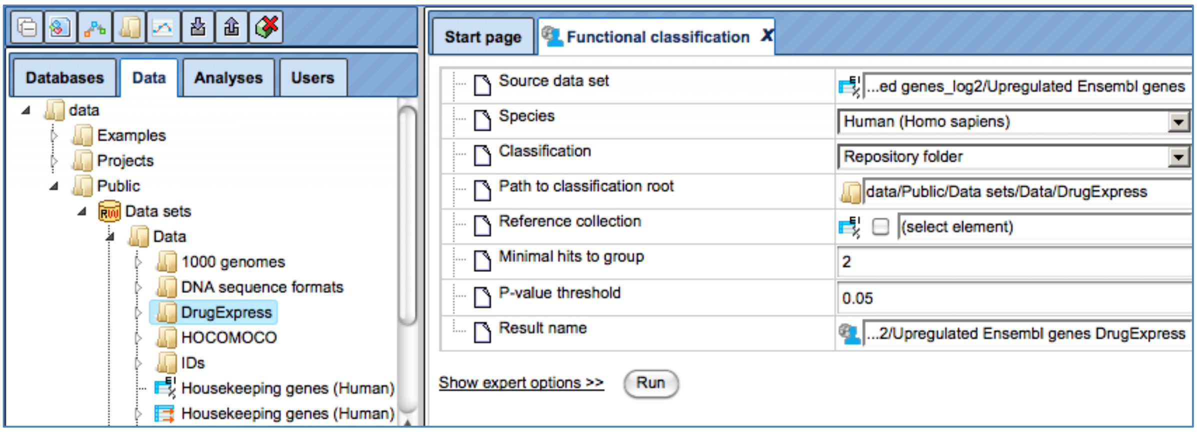
Start page
Functional classification
Full gene ontology classification
GO (biological process)
GO (cellular component)
GO (molecular function)
HumanCyc pathways
PROTEOME(TM) GO (biological process)
PROTEOME(TM) GO (cellular component)
PROTEOME(TM) GO (full classification)
PROTEOME(TM) GO (molecular function)
PROTEOME(TM) disease
Reactome pathways
TF classification
TRANSPATH Pathways (2013.1)
Repository folder
Show expert options >>
Run

Step 4. As soon as the option **Repository folder** is selected, two additional fields will automatically appear in the input form, **Path to classification root** and **Reference collection**. The input form with the added fields is shown below.



Source data set	...ed genes_log2/Upregulated Ensembl genes
Species	Human (Homo sapiens)
Classification	Repository folder
Path to classification root	(select element)
Reference collection	□ (select element)
Minimal hits to group	2
P-value threshold	0.05
Result name	(select element)

Step 5. Mark the DrugExpress folder in the field **Path to classification root**.



Source data set	...ed genes_log2/Upregulated Ensembl genes
Species	Human (Homo sapiens)
Classification	Repository folder
Path to classification root	data/Public/Data sets/Data/DrugExpress
Reference collection	□ (select element)
Minimal hits to group	2
P-value threshold	0.05
Result name	...2/Upregulated Ensembl genes DrugExpress

Step 6. Leave the fields **Minimal hits to group** and **P-value threshold** as per default, and specify location and name for the output table in the field **Result name**.

Step 7. Press the [Run] button and wait till the analysis is completed.

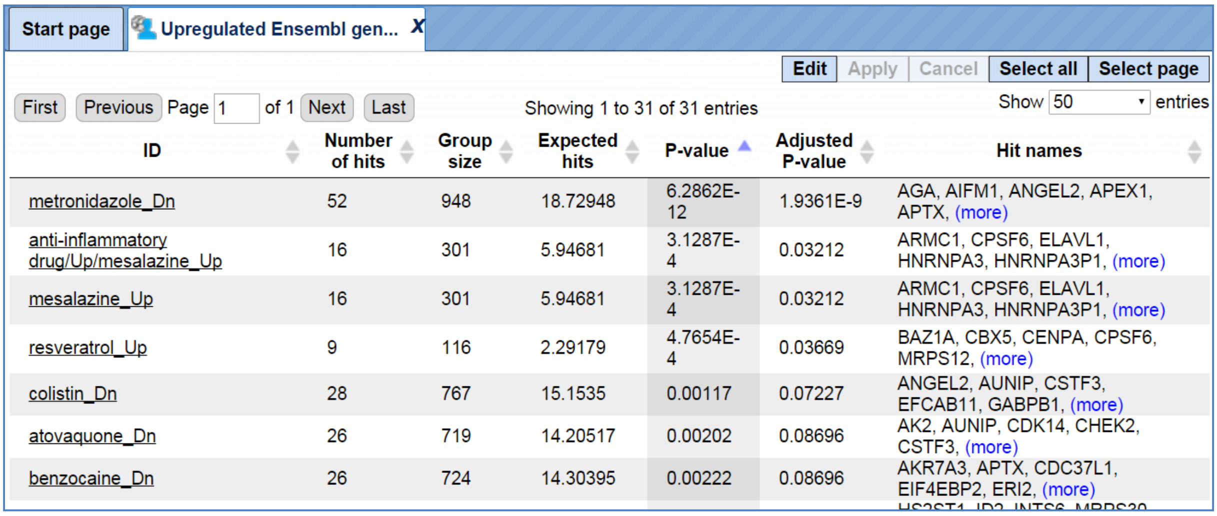
Results of the functional classification using DrugExpress

As a result, a table is generated with all columns as usual for *Functional classification* results, shown below. You can find the resulting table at:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Upregulated%20Ensembl%20genes%20DrugExpress

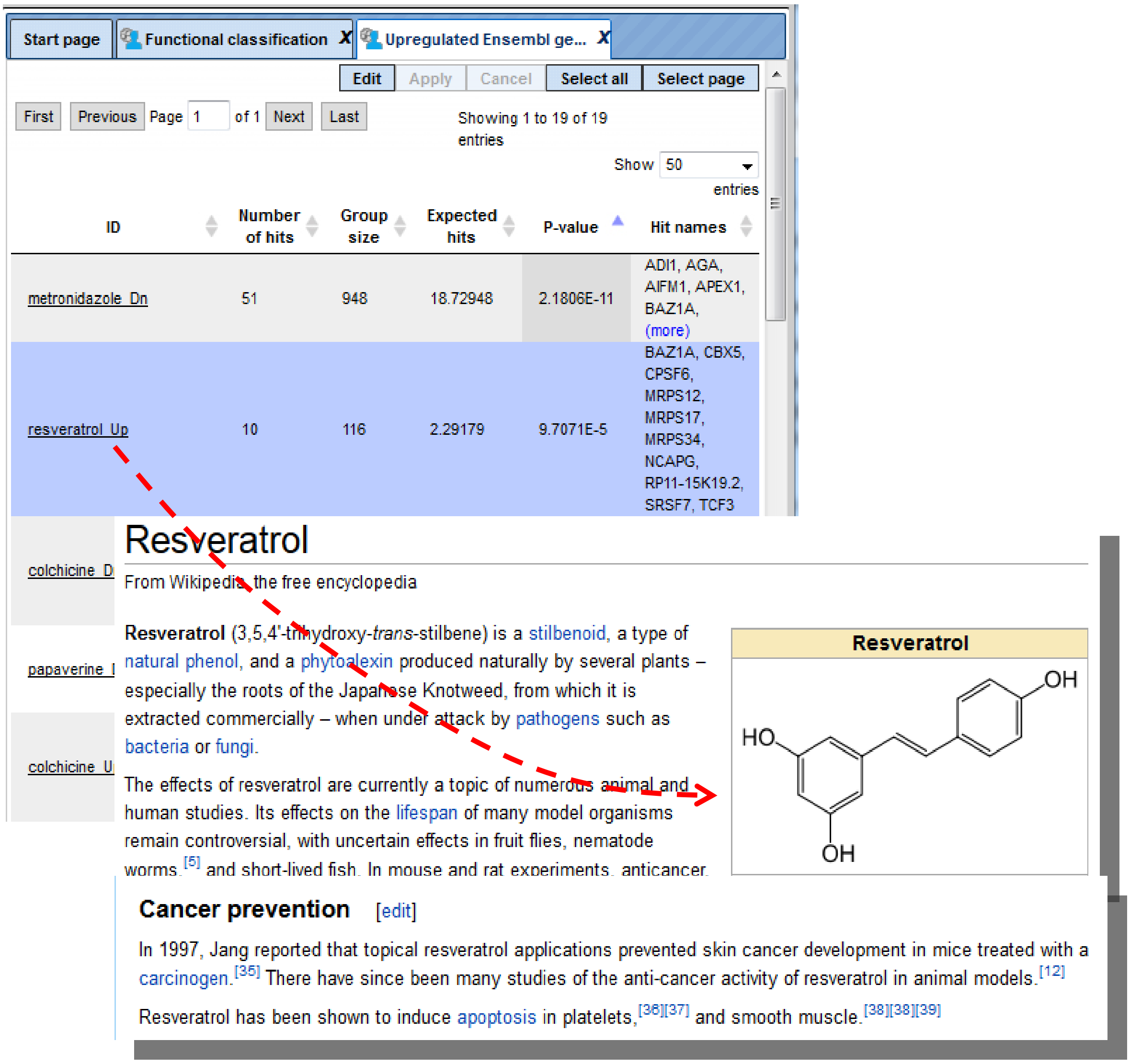
Each row corresponds to one ontological category, which in this case is one of the subfolders or tables from the DrugExpress folder. The names of the tables and subfolders are shown in the ID column. For each row several parameters are calculated, including the expected number of hits, actual number of hits, p-values, as well as hit names.

All IDs are internally hyperlinked, and with a click on each name the corresponding gene table will be opened.



ID	Number of hits	Group size	Expected hits	P-value	Adjusted P-value	Hit names
metronidazole_Dn	52	948	18.72948	6.2862E-12	1.9361E-9	AGA, AIFM1, ANGEL2, APEX1, APTX, (more)
anti-inflammatory drug/Up/mesalazine_Up	16	301	5.94681	3.1287E-4	0.03212	ARMC1, CPSF6, ELAVL1, HNRNPA3, HNRNPA3P1, (more)
mesalazine_Up	16	301	5.94681	3.1287E-4	0.03212	ARMC1, CPSF6, ELAVL1, HNRNPA3, HNRNPA3P1, (more)
resveratrol_Up	9	116	2.29179	4.7654E-4	0.03669	BAZ1A, CBX5, CENPA, CPSF6, MRPS12, (more)
colistin_Dn	28	767	15.1535	0.00117	0.07227	ANGEL2, AUNIP, CSTF3, EFCAB11, GABPB1, (more)
atovaquone_Dn	26	719	14.20517	0.00202	0.08696	AK2, AUNIP, CDK14, CHEK2, CSTF3, (more)
benzocaine_Dn	26	724	14.30395	0.00222	0.08696	AKR7A3, APTX, CDC37L1, EIF4EBP2, ERI2, (more)

For this example, we classified genes up-regulated in the Ewing brain tumor versus neuroblastoma, using DrugExpress gene signatures. Each resulting line identifies a statistically significant similarity of genes up-regulated in Ewing brain tumor versus neuroblastoma, with the gene signature in response to a given drug or chemical substance. For example, 52 genes from the input list are classified into the group *metronidazole Dn*, which means that these 52 genes are known to be down-regulated by metronidazole; the p-value of this classification is 6.2862E-12. In the next lines, 9 genes from the input list are classified into the group *resveratrol Up*, which means these 9 genes are known to be up-regulated by resveratrol; the p-value of this classification is 4.7654E-4 This might be an interesting hint, because resveratrol has a potential anticancer activity.



Showing 1 to 19 of 19 entries

Show 50 entries

ID	Number of hits	Group size	Expected hits	P-value	Hit names
metronidazole_Dn	51	948	18.72948	2.1806E-11	ADI1, AGA, AIFM1, APEX1, BAZ1A, (more)
resveratrol_Up	10	116	2.29179	9.7071E-5	BAZ1A, CBX5, CPSF6, MRPS12, MRPS17, MRPS34, NCAPG, RP11-15K19.2, SRSF7, TCF3
colchicine_D					
papaverine_					
colchicine_U					

Resveratrol

From Wikipedia, the free encyclopedia

Resveratrol (3,5,4'-trihydroxy-*trans*-stilbene) is a stilbenoid, a type of natural phenol, and a phytoalexin produced naturally by several plants – especially the roots of the Japanese Knotweed, from which it is extracted commercially – when under attack by pathogens such as bacteria or fungi.

The effects of resveratrol are currently a topic of numerous animal and human studies. Its effects on the lifespan of many model organisms remain controversial, with uncertain effects in fruit flies, nematode worms,[5] and short-lived fish. In mouse and rat experiments, anticancer

Cancer prevention [edit]

In 1997, Jang reported that topical resveratrol applications prevented skin cancer development in mice treated with a carcinogen.[35] There have since been many studies of the anti-cancer activity of resveratrol in animal models.[12]

Resveratrol has been shown to induce apoptosis in platelets,[36][37] and smooth muscle.[38][38][39]

Tip. Any user-created ontology or collection of tables can be used for the classification of input gene lists, in a similar way as it has been shown for DrugExpress. To do this, you need to create a corresponding folder in your project, which can contain a hierarchy of subfolders as well. Importantly, each table in such a folder should have Ensembl IDs. Each subfolder and individual table will be used as a separate classification category, and will result in a separate line in the resulting table.

20. Description of analysis methods

20.1. Sequence analysis methods

20.1.1. SNP matching

NGS DNA sequencing is a powerful analytical method to discover novel SNPs and detect known SNPs. The geneXplain platform provides a unique analysis method termed SNP matching. Using this method and an SNP table (derived after sequencing) as input, the corresponding SNP loci are mapped to Ensembl genes so that you can get an annotated SNP table as output.

By default the SNP matching tool looks as shown below:

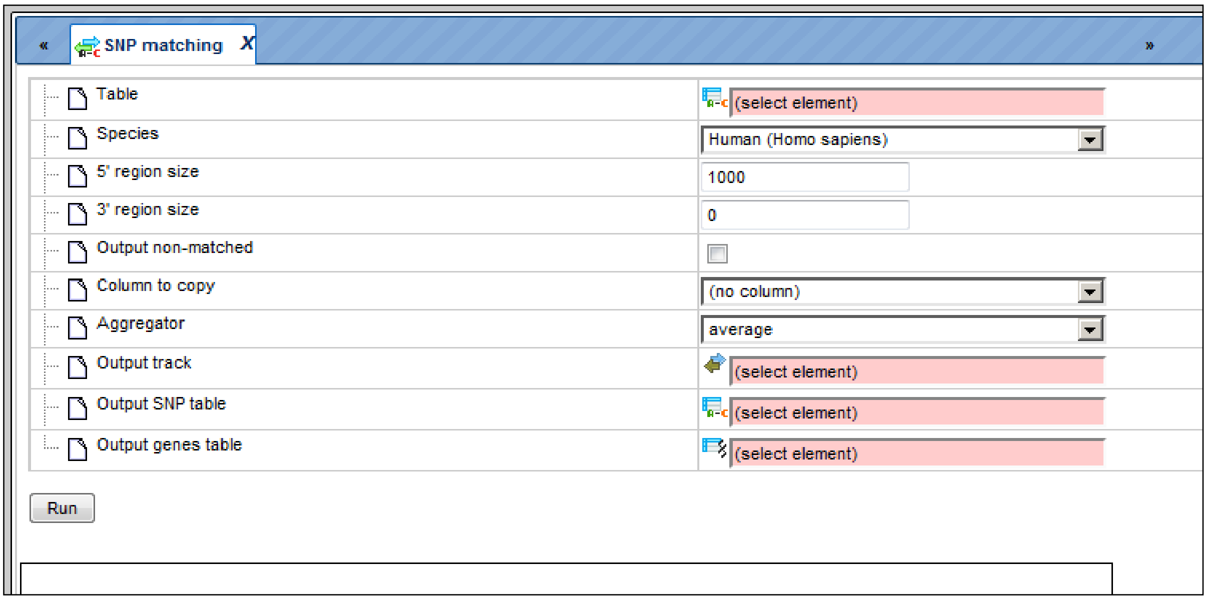
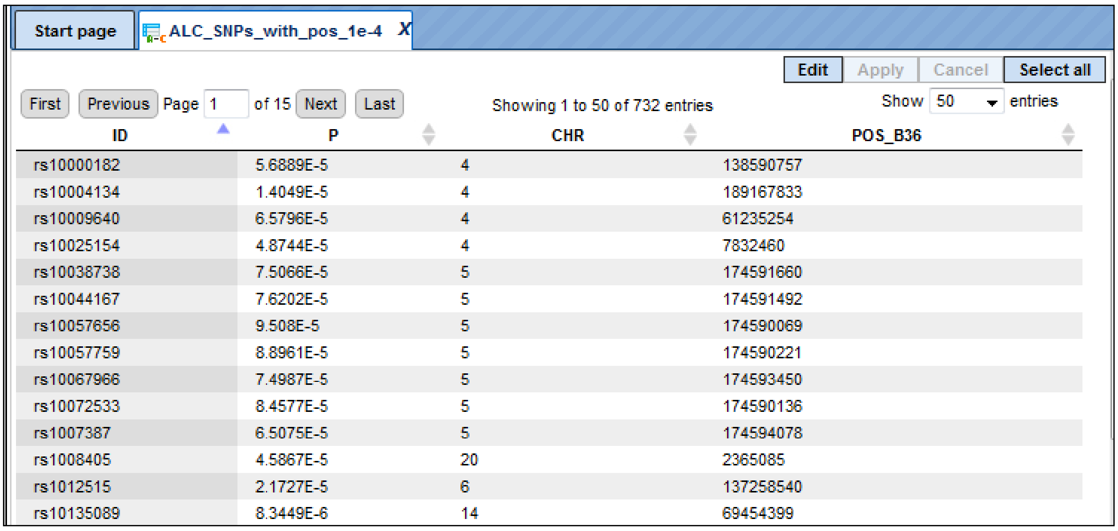


Table	(select element)
Species	Human (Homo sapiens)
5' region size	1000
3' region size	0
Output non-matched	□
Column to copy	(no column)
Aggregator	average
Output track	(select element)
Output SNP table	(select element)
Output genes table	(select element)

Example:

An example data file to be used as an input table for SNP matching may look like this:



Page 1 of 15

Showing 1 to 50 of 732 entries

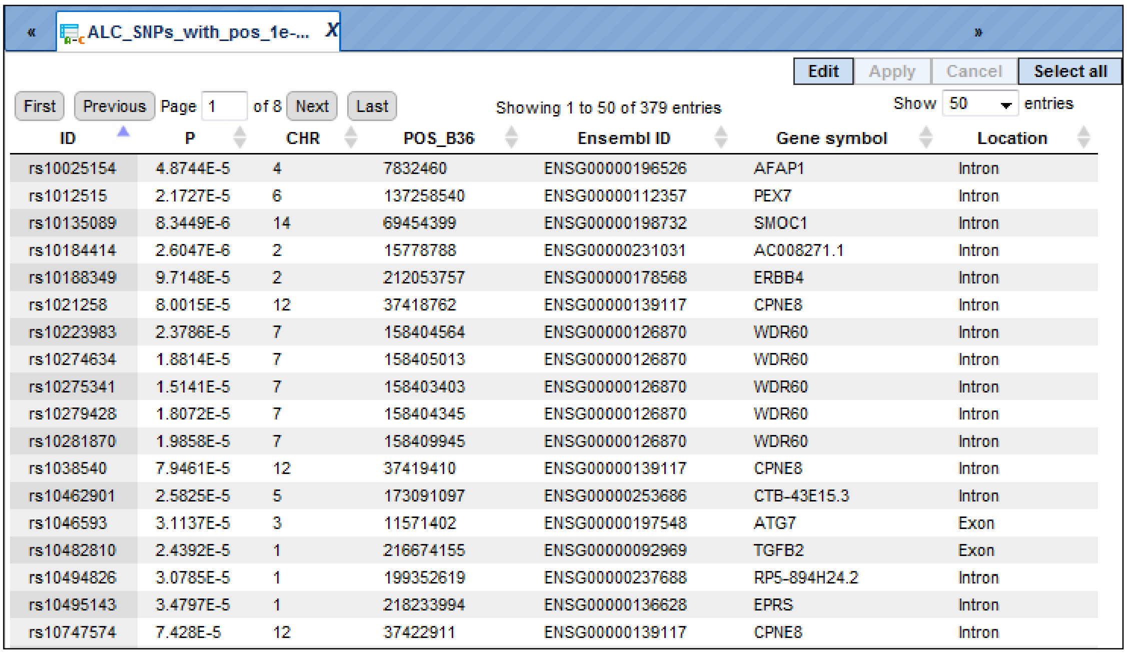
Show 50 entries

ID	P	CHR	POS_B36
rs10000182	5.6889E-5	4	138590757
rs10004134	1.4049E-5	4	189167833
rs10009640	6.5796E-5	4	61235254
rs10025154	4.8744E-5	4	7832460
rs10038738	7.5066E-5	5	174591660
rs10044167	7.6202E-5	5	174591492
rs10057656	9.508E-5	5	174590069
rs10057759	8.8961E-5	5	174590221
rs10067966	7.4987E-5	5	174593450
rs10072533	8.4577E-5	5	174590136
rs1007387	6.5075E-5	5	174594078
rs1008405	4.5867E-5	20	2365085
rs1012515	2.1727E-5	6	137258540
rs10135089	8.3449E-6	14	69454399

In this table, **ID** is the SNP identifier, **CHR** is the chromosome, **P** is some p-value, and **POS_B36** is the genomic position of a given SNP in NCBI-build36 human genome.

You can save this table in the repository and input the saved table in the SNP matching tool. The tool yields three output files as shown below:

Output table 1 (ALC_SNPs_annotated)



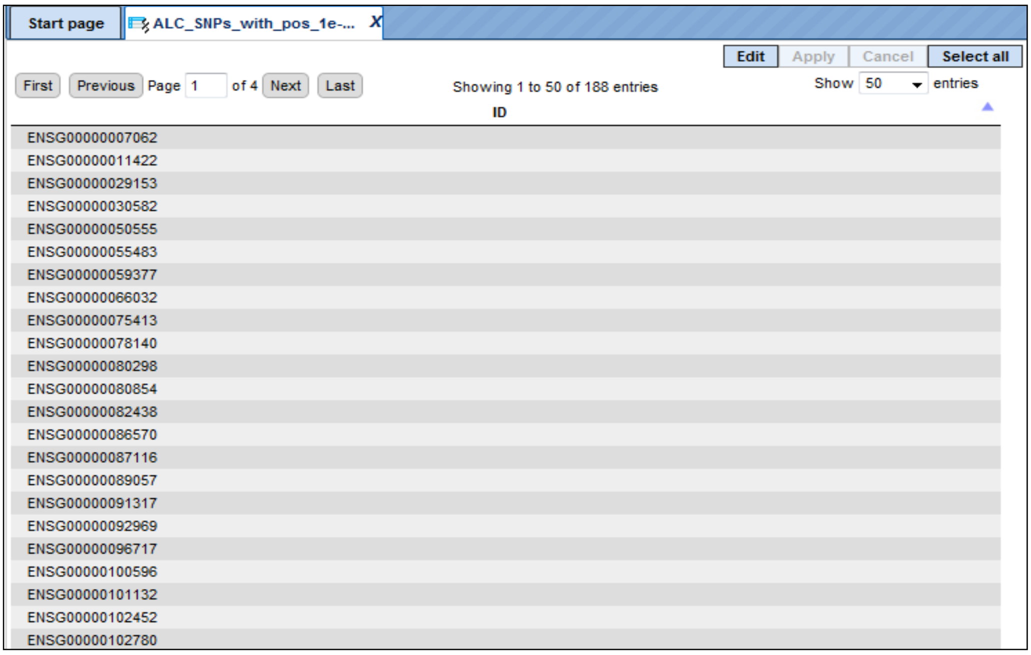
Page 1 of 8

Showing 1 to 50 of 379 entries

Show 50 entries

ID	P	CHR	POS_B36	Ensembl ID	Gene symbol	Location
rs10025154	4.8744E-5	4	7832460	ENSG00000196526	AFAP1	Intron
rs1012515	2.1727E-5	6	137258540	ENSG00000112357	PEX7	Intron
rs10135089	8.3449E-6	14	69454399	ENSG00000198732	SMOC1	Intron
rs10184414	2.6047E-6	2	15778788	ENSG00000231031	AC008271.1	Intron
rs10188349	9.7148E-5	2	212053757	ENSG00000178568	ERBB4	Intron
rs1021258	8.0015E-5	12	37418762	ENSG00000139117	CPNE8	Intron
rs10223983	2.3786E-5	7	158404564	ENSG00000126870	WDR60	Intron
rs10274634	1.8814E-5	7	158405013	ENSG00000126870	WDR60	Intron
rs10275341	1.5141E-5	7	158403403	ENSG00000126870	WDR60	Intron
rs10279428	1.8072E-5	7	158404345	ENSG00000126870	WDR60	Intron
rs10281870	1.9858E-5	7	158409945	ENSG00000126870	WDR60	Intron
rs1038540	7.9461E-5	12	37419410	ENSG00000139117	CPNE8	Intron
rs10462901	2.5825E-5	5	173091097	ENSG00000253686	CTB-43E15.3	Intron
rs1046593	3.1137E-5	3	11571402	ENSG00000197548	ATG7	Exon
rs10482810	2.4392E-5	1	216674155	ENSG00000092969	TGFB2	Exon
rs10494826	3.0785E-5	1	199352619	ENSG00000237688	RP5-894H24.2	Intron
rs10495143	3.4797E-5	1	218233994	ENSG00000136628	EPRS	Intron
rs10747574	7.428E-5	12	37422911	ENSG00000139117	CPNE8	Intron

Output table 2 (ALC_SNPs_genes)



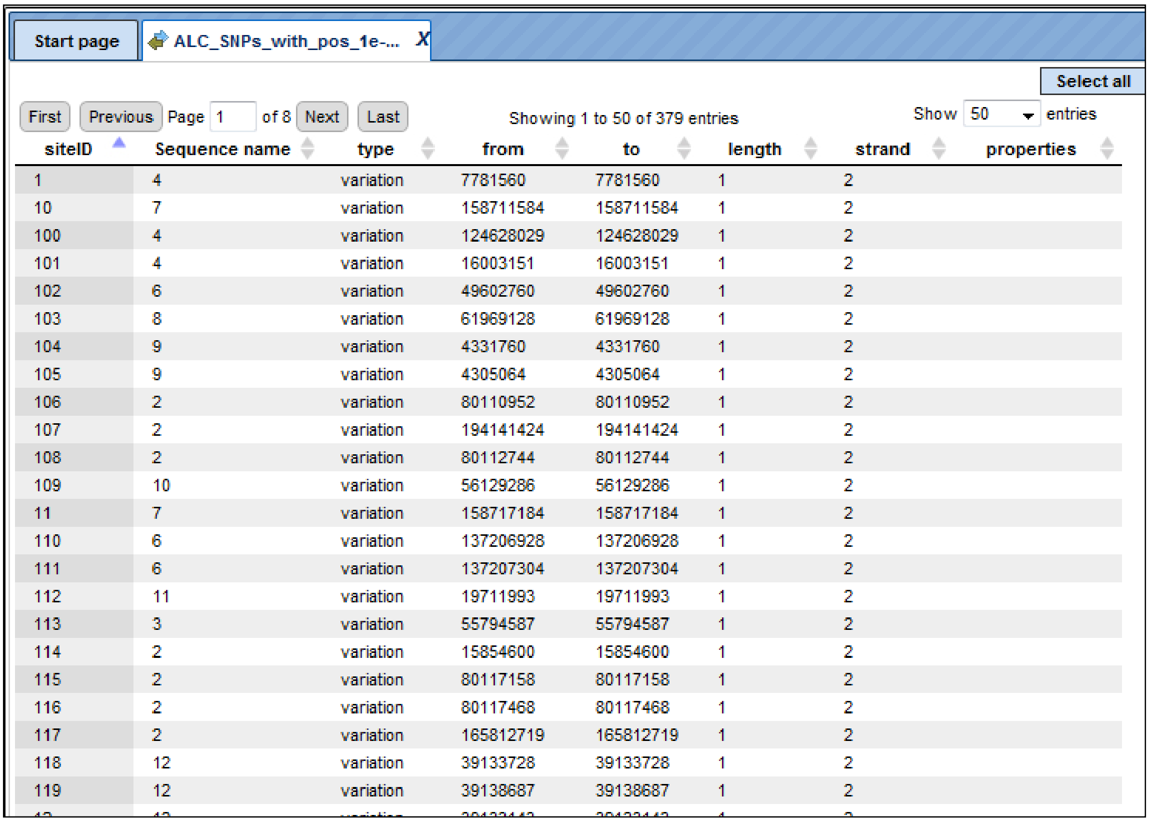
Page 1 of 4

Showing 1 to 50 of 188 entries

Show 50 entries

ID
ENSG00000007062
ENSG00000011422
ENSG00000029153
ENSG00000030582
ENSG00000050555
ENSG00000055483
ENSG00000059377
ENSG00000066032
ENSG00000075413
ENSG00000078140
ENSG00000080298
ENSG00000080854
ENSG00000082438
ENSG00000086570
ENSG00000087116
ENSG00000089057
ENSG00000091317
ENSG00000092969
ENSG00000096717
ENSG00000100596
ENSG00000101132
ENSG00000102452
ENSG00000102780

Output table 3 (ALC_SNPs_track)



Page 1 of 8

Showing 1 to 50 of 379 entries

Show 50 entries

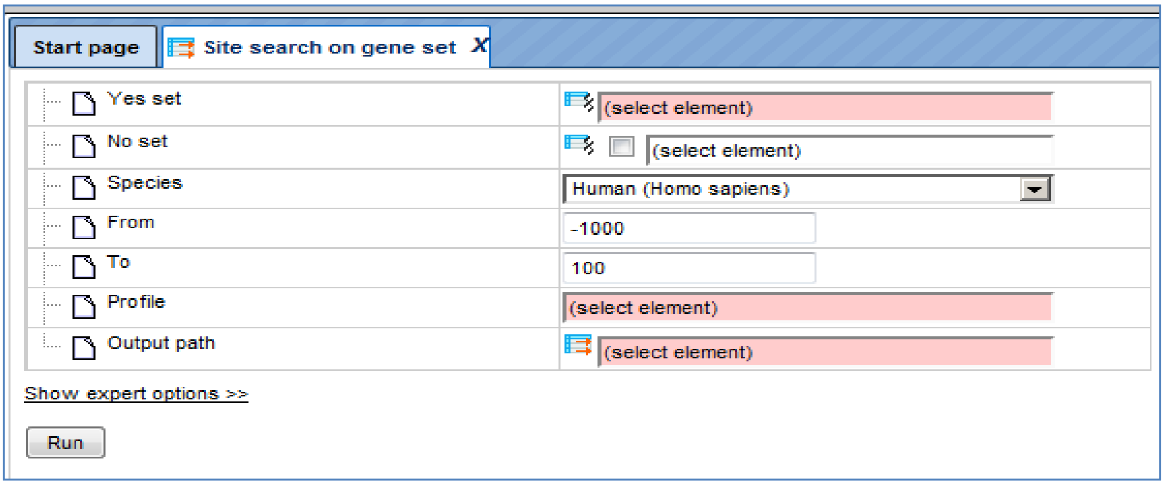
siteID	Sequence name	type	from	to	length	strand	properties
1	4	variation	7781560	7781560	1	2	
10	7	variation	158711584	158711584	1	2	
100	4	variation	124628029	124628029	1	2	
101	4	variation	16003151	16003151	1	2	
102	6	variation	49602760	49602760	1	2	
103	8	variation	61969128	61969128	1	2	
104	9	variation	4331760	4331760	1	2	
105	9	variation	4305064	4305064	1	2	
106	2	variation	80110952	80110952	1	2	
107	2	variation	194141424	194141424	1	2	
108	2	variation	80112744	80112744	1	2	
109	10	variation	56129286	56129286	1	2	
11	7	variation	158717184	158717184	1	2	
110	6	variation	137206928	137206928	1	2	
111	6	variation	137207304	137207304	1	2	
112	11	variation	19711993	19711993	1	2	
113	3	variation	55794587	55794587	1	2	
114	2	variation	15854600	15854600	1	2	
115	2	variation	80117158	80117158	1	2	
116	2	variation	80117468	80117468	1	2	
117	2	variation	165812719	165812719	1	2	
118	12	variation	39133728	39133728	1	2	
119	12	variation	39138687	39138687	1	2	

The output tables can be further used for any other analysis of the geneXplain platform.

20.1.2. Site search on gene set

This feature provides you with an option to search for putative transcription factor binding sites (TFBS) in a set of genes. As input for the analysis you are supposed to indicate two gene sets, Yes (e.g. differentially expressed in an experiment, test set) and No (set of background genes, control set) as well as positional range relative to the TSS and a collection of predefined weight matrices with a particular threshold (profile).

The initial form of this analysis looks as it is shown below:



Yes set	(select element)
No set	□ (select element)
Species	Human (Homo sapiens)
From	-1000
To	100
Profile	(select element)
Output path	(select element)

To perform this analysis you have to input two datasets:

Yes set: This is the set of genes that you want to analyze, for example these can be genes the expression of which has changed in an experiment (test set). Tables with Ensembl gene IDs can be used as input in this column. In case you have a file with different identifiers, you can first use the Convert table function in the Data manipulation folder.

No set: This is the set of background genes (control set). This again should be a gene table with Ensembl gene IDs as input.

These two datasets might be taken from the output tables of previous analyses; see, e.g., Detect differentially expressed genes (Section 10.2).

Species: A pull down menu allows you to select the biological species according to the species of your Yes and No gene sets. Currently, the analysis can be done for human, mouse and rat genes.

From: You can indicate the gene region where the search for putative TFBS should be done. Here, you enter the 5' border of the region, relative to the transcription start site (TSS) as it is annotated in the current version of Ensembl.

To: Here, you enter the 3' border of the region relative to the TSS.

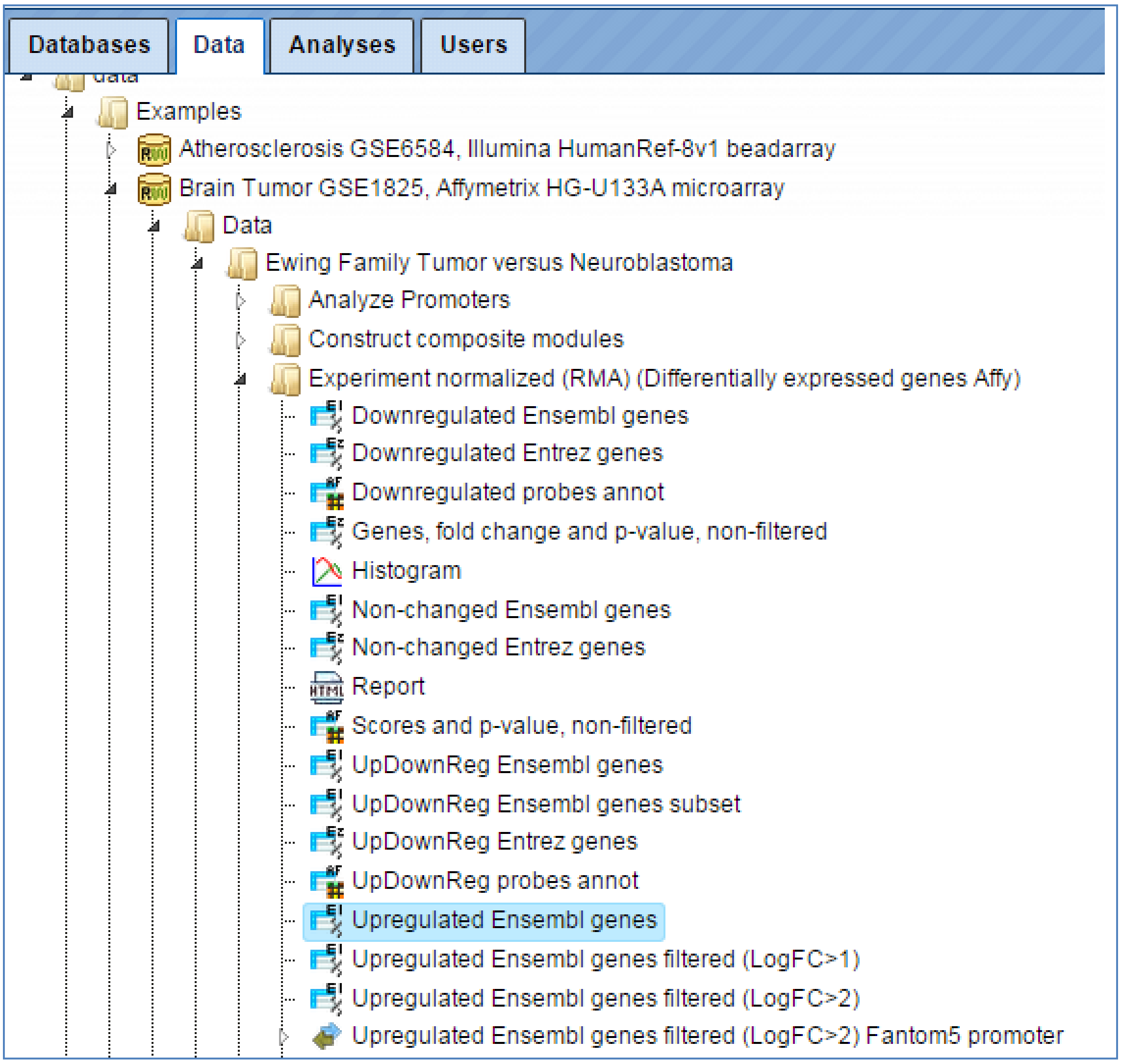
Profile: This is a predefined set of positional weight matrices with a particular threshold. By default, the TRANSFAC® profile *vertebrate_non_redundant_minSUM* is applied. You can use other available TRANSFAC® profiles. Alternatively, you can apply profiles in the GTRD database. Currently, there are three profiles included in the GTRD database; they depend on the threshold for matrices: strong threshold (with an Individual

Probability Score (IPS) higher than 6), moderate threshold (with IPS higher than 5), and weak threshold (with IPS higher than 4).

You also have the option to import matrices and profiles into the geneXplain platform and use them for your further analysis.

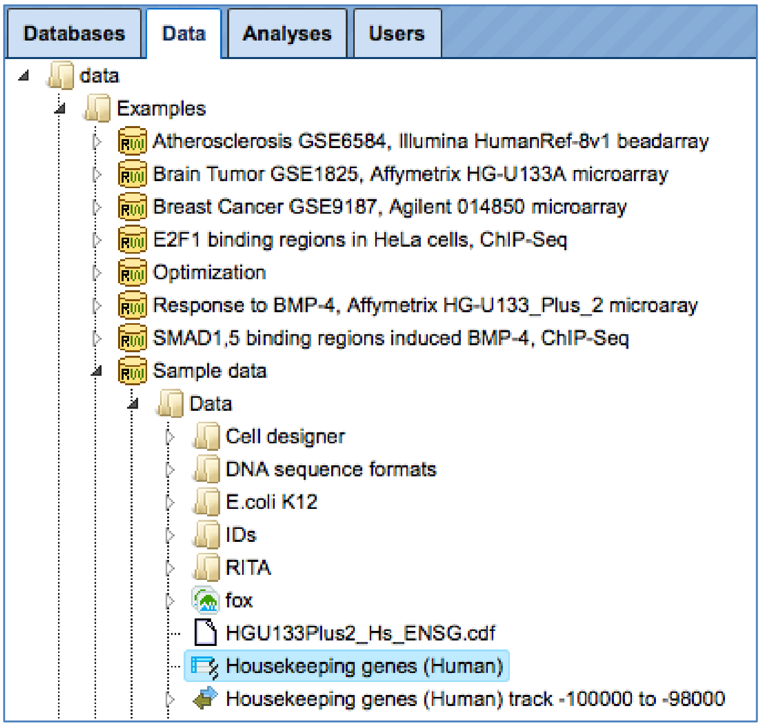
To perform site search following steps are recommended:

Step 1. Input Yes set from the tree. You can drag-and-drop as usual. Here, the set of genes from the Example folder is used as input Yes set, highlighted blue on the screenshot below:



- data
  - Examples
    - Atherosclerosis GSE6584, Illumina HumanRef-8v1 beadarray
    - Brain Tumor GSE1825, Affymetrix HG-U133A microarray
      - Data
        - Ewing Family Tumor versus Neuroblastoma
          - Analyze Promoters
          - Construct composite modules
          - Experiment normalized (RMA) (Differentially expressed genes Affy)
            - Downregulated Ensembl genes
            - Downregulated Entrez genes
            - Downregulated probes annot
            - Genes, fold change and p-value, non-filtered
            - Histogram
            - Non-changed Ensembl genes
            - Non-changed Entrez genes
            - Report
            - Scores and p-value, non-filtered
            - UpDownReg Ensembl genes
            - UpDownReg Ensembl genes subset
            - UpDownReg Entrez genes
            - UpDownReg probes annot
            - Upregulated Ensembl genes
            - Upregulated Ensembl genes filtered (LogFC>1)
            - Upregulated Ensembl genes filtered (LogFC>2)
            - Upregulated Ensembl genes filtered (LogFC>2) Fantom5 promoter

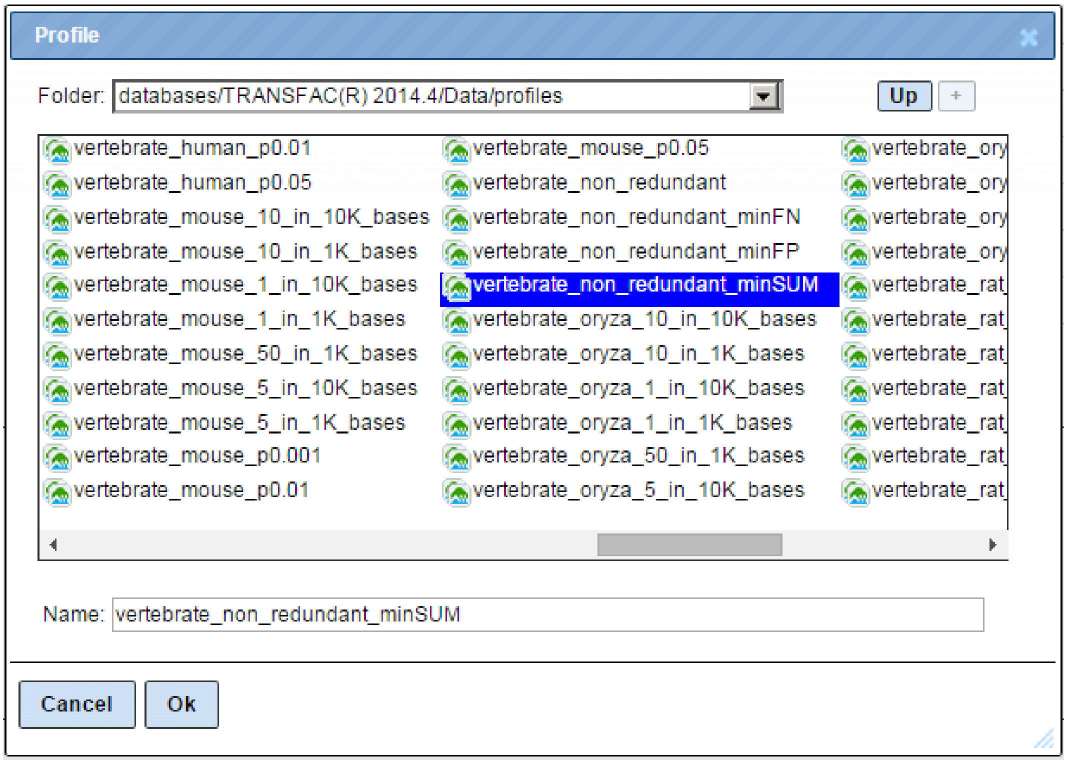
Step 2. Input pre-saved No set. You can drag-and-drop as usual. Here, the set of human housekeeping genes from the Example folder is used as input No set, highlighted blue on the screenshot below:



- data
  - Examples
    - Atherosclerosis GSE6584, Illumina HumanRef-8v1 beadarray
    - Brain Tumor GSE1825, Affymetrix HG-U133A microarray
    - Breast Cancer GSE9187, Agilent 014850 microarray
    - E2F1 binding regions in HeLa cells, ChIP-Seq
    - Optimization
    - Response to BMP-4, Affymetrix HG-U133_Plus_2 microaray
    - SMAD1,5 binding regions induced BMP-4, ChIP-Seq
    - Sample data
      - Data
        - Cell designer
        - DNA sequence formats
        - E.coli K12
        - IDs
        - RITA
        - fox
        - HGU133Plus2_Hs_ENSG.cdf
        - Housekeeping genes (Human)
        - Housekeeping genes (Human) track -100000 to -98000

Step 3. Select the species and the promoter length. By default the promoter length is set as -1000 to +100 relative to the TSS. In this example, the range -500 to 100 is selected.

Step 4. Select the TRANSFAC® or GTRD profile from the pre-saved profiles in the tool. In this example the default TRANSFAC® profile *vertebrate_non_redundant_minSUM* is used:



Folder: databases/TRANSFAC(R) 2014.4/Data/profiles

vertebrate_human_p0.01
vertebrate_human_p0.05
vertebrate_mouse_10_in_10K_bases
vertebrate_mouse_10_in_1K_bases
vertebrate_mouse_1_in_10K_bases
vertebrate_mouse_1_in_1K_bases
vertebrate_mouse_50_in_1K_bases
vertebrate_mouse_5_in_10K_bases
vertebrate_mouse_5_in_1K_bases
vertebrate_mouse_p0.001
vertebrate_mouse_p0.01
vertebrate_mouse_p0.05
vertebrate_non_redundant
vertebrate_non_redundant_minFN
vertebrate_non_redundant_minFP
vertebrate_non_redundant_minSUM
vertebrate_oryza_10_in_10K_bases
vertebrate_oryza_10_in_1K_bases
vertebrate_oryza_1_in_10K_bases
vertebrate_oryza_1_in_1K_bases
vertebrate_oryza_50_in_1K_bases
vertebrate_oryza_5_in_10K_bases
vertebrate_ory
vertebrate_ory
vertebrate_ory
vertebrate_ory
vertebrate_rat
vertebrate_rat
vertebrate_rat
vertebrate_rat
vertebrate_rat
vertebrate_rat
vertebrate_rat

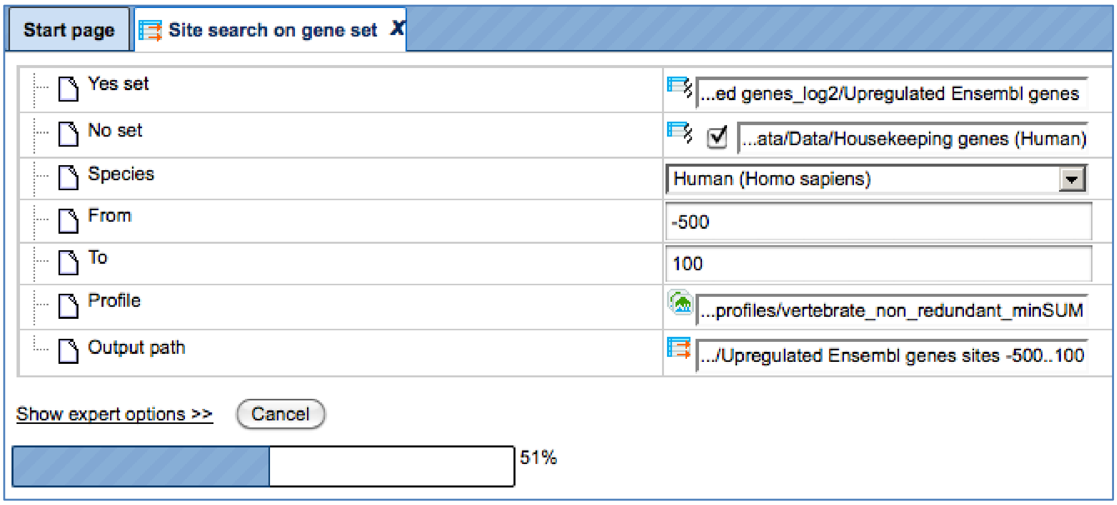
Name: vertebrate_non_redundant_minSUM

Step 5. Identify the output path. Define where the folder with the results should be located in your project tree. You can do so by clicking on the pink field “select element” in the field **Output path**, and a new window will be opened, where you can select the location of the results folder and define its name.

Important: please define the output path within one of your Project folders.

Press [Run].

The analysis will start as shown below:

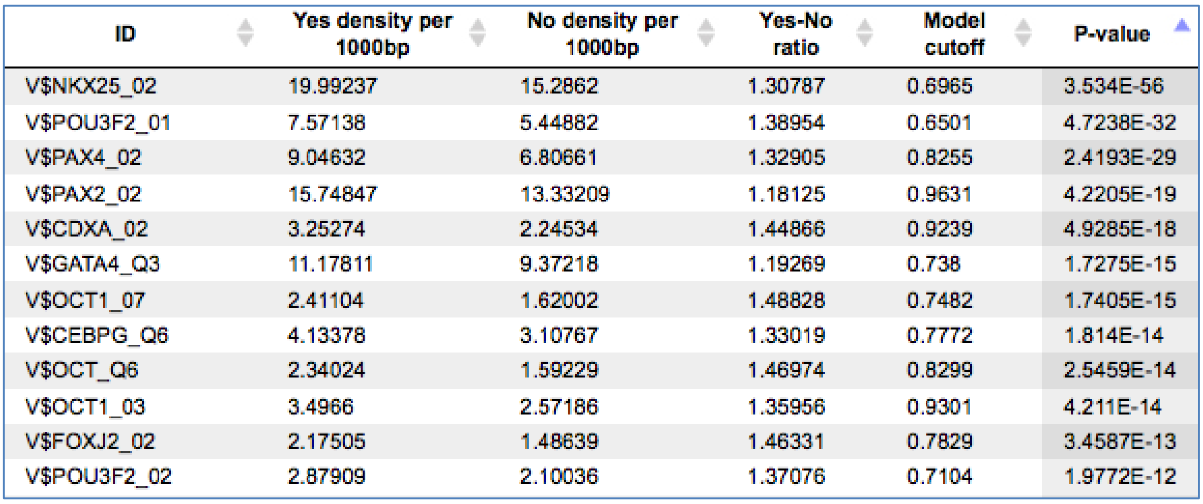


Yes set	...ed genes_log2/Upregulated Ensembl genes
No set	☑ ...ata/Data/Housekeeping genes (Human)
Species	Human (Homo sapiens)
From	-500
To	100
Profile	...profiles/vertebrate_non_redundant_minSUM
Output path	.../Upregulated Ensembl genes sites -500..100

Wait until the analysis is complete as shown by the progress bar. The output of Site search on gene set contains one table and six tracks:

summary table ( ), yes promoters ( ), no promoters ( ), yes sites ( ), no sites ( ), yes sites optimized ( ) and no sites optimized ( ).

The *summary* table is automatically opened in a new tab of the geneXplain platform when the analysis is completed. An example of the summary table is shown below:



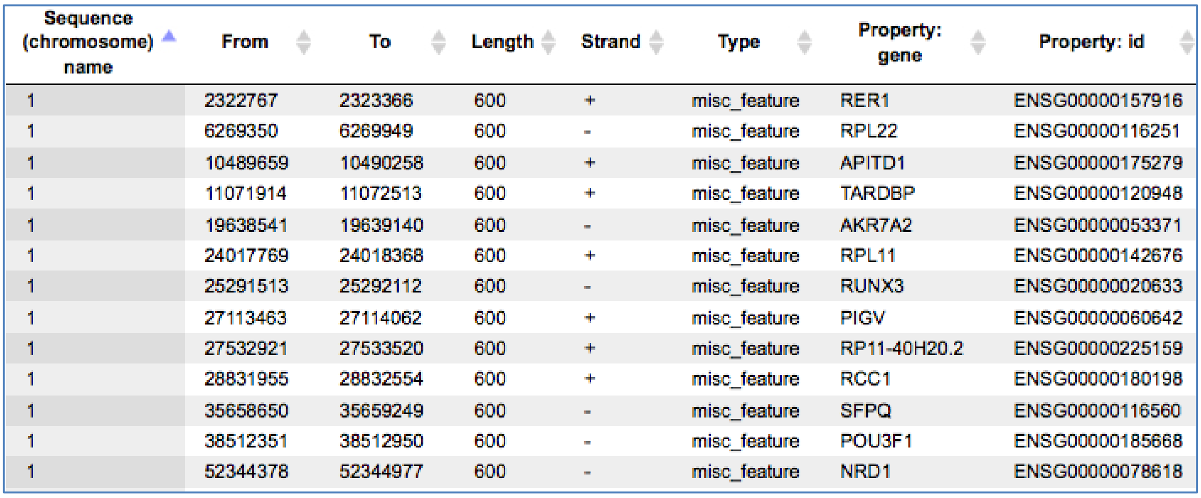
ID	Yes density per 1000bp	No density per 1000bp	Yes-No ratio	Model cutoff	P-value
V$NKX25_02	19.99237	15.2862	1.30787	0.6965	3.534E-56
V$POU3F2_01	7.57138	5.44882	1.38954	0.6501	4.7238E-32
V$PAX4_02	9.04632	6.80661	1.32905	0.8255	2.4193E-29
V$PAX2_02	15.74847	13.33209	1.18125	0.9631	4.2205E-19
V$CDXA_02	3.25274	2.24534	1.44866	0.9239	4.9285E-18
V$GATA4_Q3	11.17811	9.37218	1.19269	0.738	1.7275E-15
V$OCT1_07	2.41104	1.62002	1.48828	0.7482	1.7405E-15
V$CEBPG_Q6	4.13378	3.10767	1.33019	0.7772	1.814E-14
V$OCT_Q6	2.34024	1.59229	1.46974	0.8299	2.5459E-14
V$OCT1_03	3.4966	2.57186	1.35956	0.9301	4.211E-14
V$FOXJ2_02	2.17505	1.48639	1.46331	0.7829	3.4587E-13
V$POU3F2_02	2.87909	2.10036	1.37076	0.7104	1.9772E-12

Each row summarizes the information for one PWM. For each selected matrix, the columns **Yes density per 1000bp** and **No density per 1000bp** show the number of matches normalized per 1000 bp length for the sequences in the input Yes set and input No set, respectively. The Column **Yes-No ratio** is the ratio of the first two columns. Only matrices with a Yes-No ratio higher than 1 are included in the summary table. The higher the Yes-No ratio, the higher is the enrichment of matches for the respective matrix in the Yes set. The matrix cutoff values as they are calculated by the program at the optimization step are shown in the column **Model cutoff**, and the last column shows the p-value of the corresponding event.

Tracks Yes promoters and No promoters

These two files resulting from the "Site search on gene set" analysis represent promoters of the corresponding set, as a track (). The track files might be used for site visualization or for further analyses, e.g. Construct Composite Modules, Construct IPS CisModule.

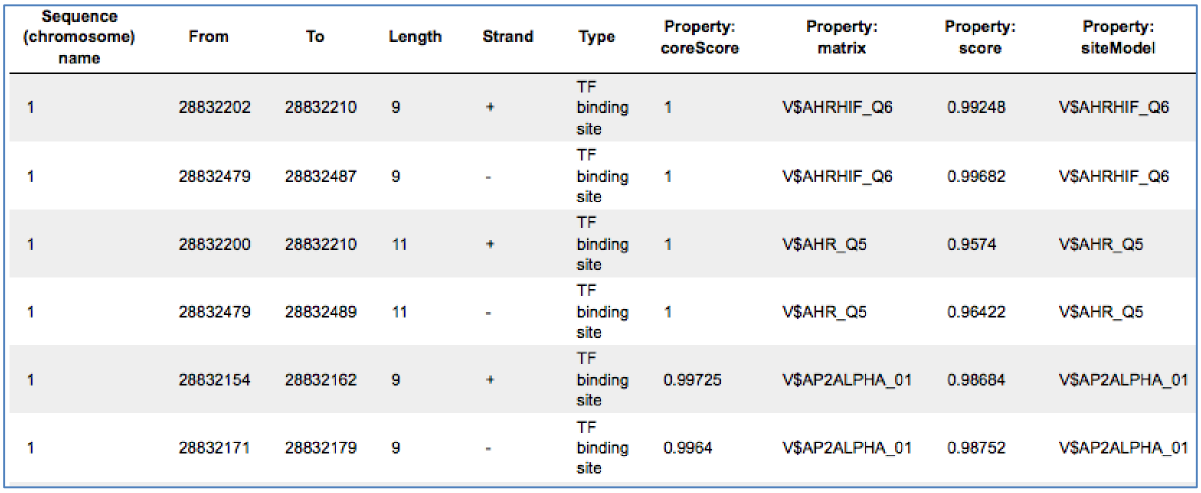
The track file *Yes promoters* when opened in the work space is shown below. The track file *No promoters* has a similar structure.



Sequence (chromosome) name	From	To	Length	Strand	Type	Property: gene	Property: id
1	2322767	2323366	600	+	misc_feature	RER1	ENSG00000157916
1	6269350	6269949	600	-	misc_feature	RPL22	ENSG00000116251
1	10489659	10490258	600	+	misc_feature	APITD1	ENSG00000175279
1	11071914	11072513	600	+	misc_feature	TARDBP	ENSG00000120948
1	19638541	19639140	600	-	misc_feature	AKR7A2	ENSG00000053371
1	24017769	24018368	600	+	misc_feature	RPL11	ENSG00000142676
1	25291513	25292112	600	-	misc_feature	RUNX3	ENSG00000020633
1	27113463	27114062	600	+	misc_feature	PIGV	ENSG00000060642
1	27532921	27533520	600	+	misc_feature	RP11-40H20.2	ENSG00000225159
1	28831955	28832554	600	+	misc_feature	RCC1	ENSG00000180198
1	35658650	35659249	600	-	misc_feature	SFPQ	ENSG00000116560
1	38512351	38512950	600	-	misc_feature	POU3F1	ENSG00000185668
1	52344378	52344977	600	-	misc_feature	NRD1	ENSG00000078618

This table lists the positions of the promoter areas selected for the analysis on particular chromosomes, as shown in the columns **From** and **To**. The column **Strand** shows the strand on which each particular promoter is located. This track can be dragged and dropped on a particular chromosome opened in the genome browser to visualize the localizations of the promoters as discussed in Section 19.3.

***Yes (No) sites optimized* track.** The file *Yes sites* visualizes those putative sites that are over-represented in the promoters of the Yes set versus the No set as they are located in the promoters of the Yes set. Putative TFBSs are shown as a track . Scores of the putative sites are optimized by the algorithm.



Sequence (chromosome) name	From	To	Length	Strand	Type	Property: coreScore	Property: matrix	Property: score	Property: siteModel
1	28832202	28832210	9	+	TF binding site	1	V$AHRHIF_Q6	0.99248	V$AHRHIF_Q6
1	28832479	28832487	9	-	TF binding site	1	V$AHRHIF_Q6	0.99682	V$AHRHIF_Q6
1	28832200	28832210	11	+	TF binding site	1	V$AHR_Q5	0.9574	V$AHR_Q5
1	28832479	28832489	11	-	TF binding site	1	V$AHR_Q5	0.96422	V$AHR_Q5
1	28832154	28832162	9	+	TF binding site	0.99725	V$AP2ALPHA_01	0.98684	V$AP2ALPHA_01
1	28832171	28832179	9	-	TF binding site	0.9964	V$AP2ALPHA_01	0.98752	V$AP2ALPHA_01

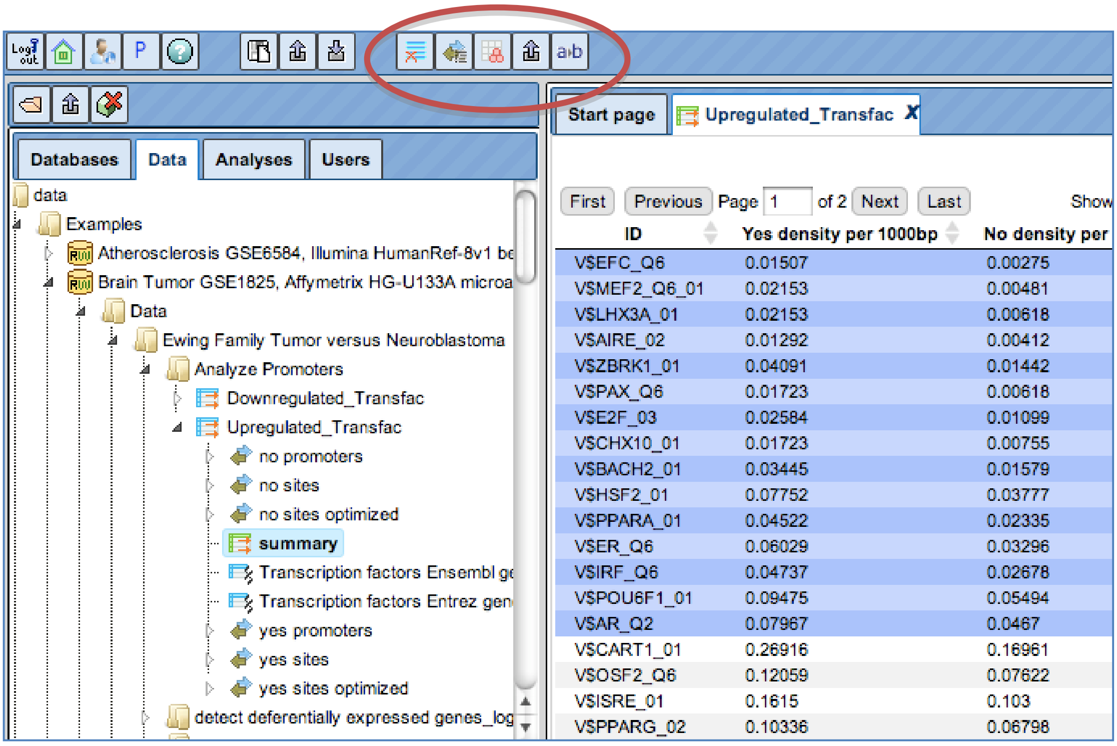
This track is a list of all putative TFBS found in one analysis. Each row presents details for each individual match for every PWM. The columns **Sequence (chromosome) name**, **From**, **To**, **Length** and **Strand** show, correspondingly, genomic location of the match including chromosome number, start and end positions, strand and length of the match. The column **Type** contains information about the type of the elements, in this case all matches are considered as "TF binding site". Further columns keep information about PWM producing each match (column **Property: matrix**) as well as score for the whole matrix (column **Property: score**). The column **Property: siteModel** contains the identifier for the corresponding site model, which is the matrix together with a cutoff applied (and in the example shown is identical to the matrix identifier).

***Yes (No) sites* tracks** are very similar in structure. The major difference is that these tracks include putative binding sites before the cutoff optimization, and thus they contain more sites.

These track files can be used as an input for other functions, for example sites can be visualized on chromosomes. For visualization details please refer to the tip in the Section 7.2.2.

Visualization of TFBS for individual genes and for individual matrices

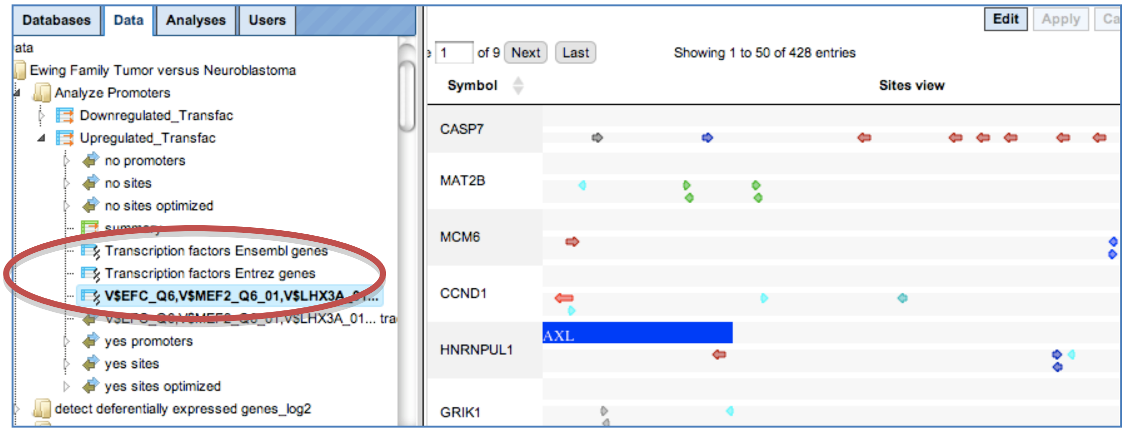
Having the summary table opened in the work space, you can select different rows and apply a couple of different functions with the help of the buttons on the top menu bar. These five buttons are marked by red oval in the figure below.



data
- Examples
  - Atherosclerosis GSE6584, Illumina HumanRef-8v1 be
  - Brain Tumor GSE1825, Affymetrix HG-U133A microa
    - Data
      - Ewing Family Tumor versus Neuroblastoma
        - Analyze Promoters
          - Downregulated_Transfac
          - Upregulated_Transfac
            - no promoters
            - no sites
            - no sites optimized
            - summary
            - Transcription factors Ensembl ge
            - Transcription factors Entrez gen
            - yes promoters
            - yes sites
            - yes sites optimized
        - detect deferentially expressed genes_log

ID	Yes density per 1000bp	No density per
V$EFC_Q6	0.01507	0.00275
V$MEF2_Q6_01	0.02153	0.00481
V$LHX3A_01	0.02153	0.00618
V$AIRE_02	0.01292	0.00412
V$ZBRK1_01	0.04091	0.01442
V$PAX_Q6	0.01723	0.00618
V$E2F_03	0.02584	0.01099
V$CHX10_01	0.01723	0.00755
V$BACH2_01	0.03445	0.01579
V$HSF2_01	0.07752	0.03777
V$PPARA_01	0.04522	0.02335
V$ER_Q6	0.06029	0.03296
V$IRF_Q6	0.04737	0.02678
V$POU6F1_01	0.09475	0.05494
V$AR_Q2	0.07967	0.0467
V$CART1_01	0.26916	0.16961
V$OSF2_Q6	0.12059	0.07622
V$ISRE_01	0.1615	0.103
V$PPARG_02	0.10336	0.06798

To get a visualization of TFBS for individual genes, the button should be applied on the selected matrices. After click on this button, two new files will be saved in the tree area, a table and a track, as shown below. The table is automatically opened in the Work Space.

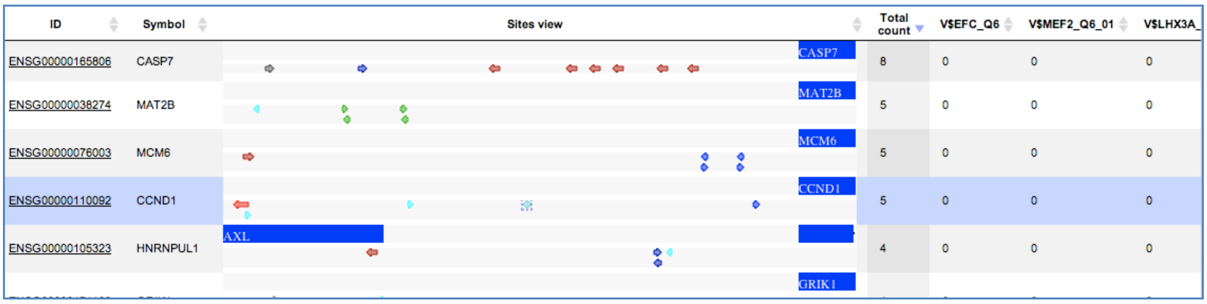


ata
- Ewing Family Tumor versus Neuroblastoma
  - Analyze Promoters
    - Downregulated_Transfac
    - Upregulated_Transfac
      - no promoters
      - no sites
      - no sites optimized
      - summary
      - Transcription factors Ensembl genes
      - Transcription factors Entrez genes
      - VEFC_Q6,VMEF2_Q6_01,V$LHX3A_01...
      - VEFC_Q6,VMEF2_Q6_01,V$LHX3A_01... tra
      - yes promoters
      - yes sites
      - yes sites optimized
  - detect deferentially expressed genes_log2

Showing 1 to 50 of 428 entries

Symbol	Sites view
CASP7	
MAT2B	
MCM6	
CCND1	
HNRNPUL1	
GRIK1	

Let’s consider this table in more detail. Each row corresponds to one individual gene.

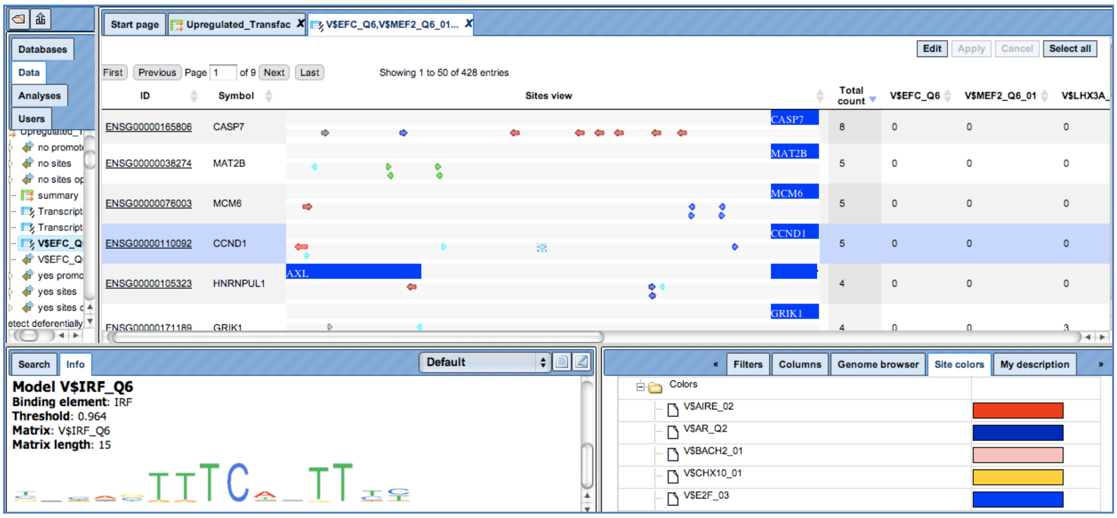


ID	Symbol	Sites view	Total count	V$EFC_Q6	V$MEF2_Q6_01	V$LHX3A_
ENSG00000165806	CASP7		8	0	0	0
ENSG00000038274	MAT2B		5	0	0	0
ENSG00000076003	MCM6		5	0	0	0
ENSG00000110092	CCND1		5	0	0	0
ENSG00000105323	HNRNPUL1		4	0	0	0

The column **ID** presents the Ensembl ID for each gene, and the gene symbol is shown in the column **Symbol**. The column **Sites view** shows a schematic representation for each gene, where blue bars correspond to gene starts and coding regions, and TFBSs for different matrices are shown by arrows of different colors. The column **Total count** shows the number of TFBSs for all matrices together in the promoter of each particular gene. The next columns are named as matrices in the summary table and represent the number of TFBSs for each matrix in each particular gene.

On the picture above the table is sorted by the column **Total count**, and on the top we can see those genes that contain the highest total number of sites. This table can be sorted by different columns corresponding to individual matrices, and then on the top you will see those genes that contain the highest number of sites for the matrix in focus.

The TFBS color schema can be customized. For this, open the tab „Site colors" in the Operation Field in the bottom right area of the tool (figure below). You can see the default colors for different matrices, and can adjust them by clicking on each color box.



Showing 1 to 50 of 428 entries

ID	Symbol	Sites view	Total count	V$EFC_Q6	V$MEF2_Q6_01	V$LHX3A_
ENSG00000165806	CASP7		8	0	0	0
ENSG00000038274	MAT2B		5	0	0	0
ENSG00000076003	MCM6		5	0	0	0
ENSG00000110092	CCND1		5	0	0	0
ENSG00000105323	HNRNPUL1		4	0	0	0
ENSG00000171189	GRIK1		4	0	0	3

Model V$IRF_Q6
Binding element: IRF
Threshold: 0.964
Matrix: V$IRF_Q6
Matrix length: 15

Colors

V$AIRE_02	🟥
V$AR_Q2	🟦
V$BACH2_01	🟥
V$CHX10_01	🟨
V$E2F_03	🟦

This table can be exported in tab-separated format (txt) or comma-separated format (csv).

The second file, a track, has the same structure as described above for other track files. For visualization details please refer to the tip in Section 7.2.2.

20.1.3. Creation of customized profiles

"Profile" is a term used for a collection of positional weight matrices (PWMs) with a particular threshold (or cutoff), also referred to as "site models". The geneXplain platform provides an option to create profiles from a table of site models or from any gene table that contains some transcription factor genes. The newly created customized profiles can be further used for analyses of regulatory regions.

There are two possibilities to create a user-specific profile.

Create profile from gene table

This option supports you in creating a new profile from any gene table. The resulting profile contains site models linked to the genes encoding the corresponding transcription factors in the input gene set. This option can be accessed via the URL:

http://genexplain-platform.com/bioumlweb/#de=analyses/Methods/Site%20analysis/Create%20profile%20from%20gene%20table

Create profile from table

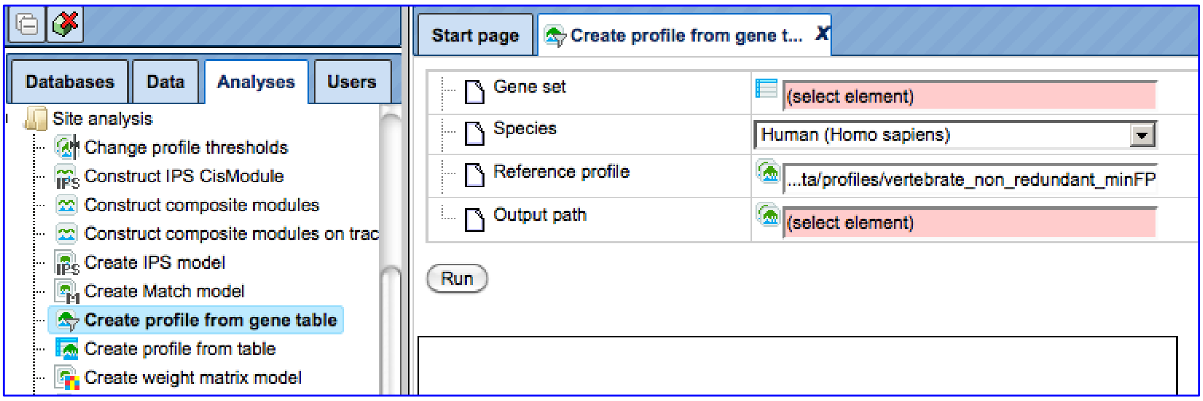
Here you can create a profile from a table of matrices. This option can be accessed via the URL:

http://genexplain-platform.com/bioumlweb/#de=analyses/Methods/Site%20analysis/Create%20profile%20from%20site%20model%20table

20.1.3.1. Create profile from gene table

This function creates a new profile from any gene table. The resulting profile contains site models linked to the transcription factors in the input gene set.

The input form when opened in the work space is shown below:


Databases
Data
Analyses
Users
Site analysis
Change profile thresholds
Construct IPS CisModule
Construct composite modules
Construct composite modules on trac
Create IPS model
Create Match model
Create profile from gene table
Create profile from table
Create weight matrix model
Start page
Create profile from gene t...
Run

Gene set	(select element)
Species	Human (Homo sapiens)
Reference profile	...ta/profiles/vertebrate_non_redundant_minFP
Output path	(select element)

In the following, we will consider each of the input fields.

Gene set: Input any gene table here for which you intend to design a profile. The algorithm automatically identifies transcription factor genes in the input table and the matrices linked to these factors in the TRANSFAC® database. These matrices will be put into the newly created profile.

Further steps are demonstrated with the table of genes available in one of the platform examples present in the Tree Area. The example file can be accessed using the URL:

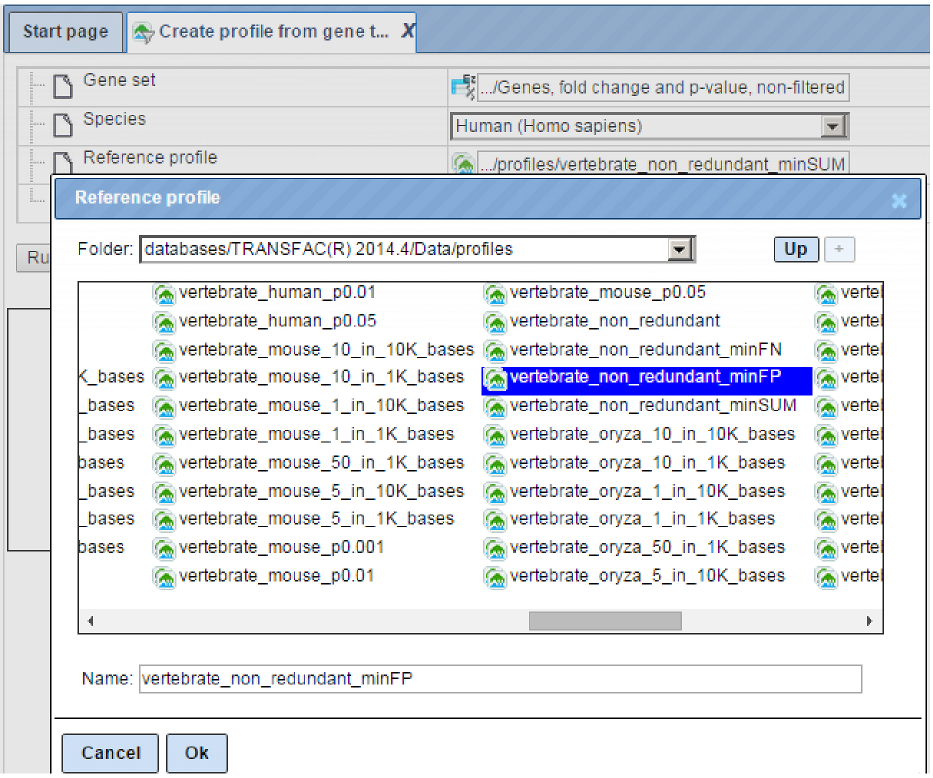
http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Experiment%20normalized%20(RMA)%20(Differentially%20expressed%20genes%20Affy)/Genes%2C%20fold%20change%20and%20p-value%2C%20non-filtered

Species: Select the biological species corresponding to the input gene set from the drop-down menu.

Reference profile: Specify one of the TRANSFAC® profiles available. Cutoffs for the corresponding matrices will be copied from the selected reference TRANSFAC® profile into the newly created profile.

To select the reference profile click on the box of this field and a new window will be opened. Scroll towards the desired profile and click on it so that the name of the selected profile is displayed in the field "Name". When the selection is done, press [Ok].

Please note that this analysis method works only for the TRANSFAC® profiles.



Gene set: .../Genes, fold change and p-value, non-filtered

Species: Human (Homo sapiens)

Reference profile: .../profiles/vertebrate_non_redundant_minSUM

Reference profile

Folder: databases/TRANSFAC(R) 2014.4/Data/profiles

K_bases
_bases
_bases
bases
_bases
_bases
bases

vertebrate_human_p0.01
vertebrate_human_p0.05
vertebrate_mouse_10_in_10K_bases
vertebrate_mouse_10_in_1K_bases
vertebrate_mouse_1_in_10K_bases
vertebrate_mouse_1_in_1K_bases
vertebrate_mouse_50_in_1K_bases
vertebrate_mouse_5_in_10K_bases
vertebrate_mouse_5_in_1K_bases
vertebrate_mouse_p0.001
vertebrate_mouse_p0.01

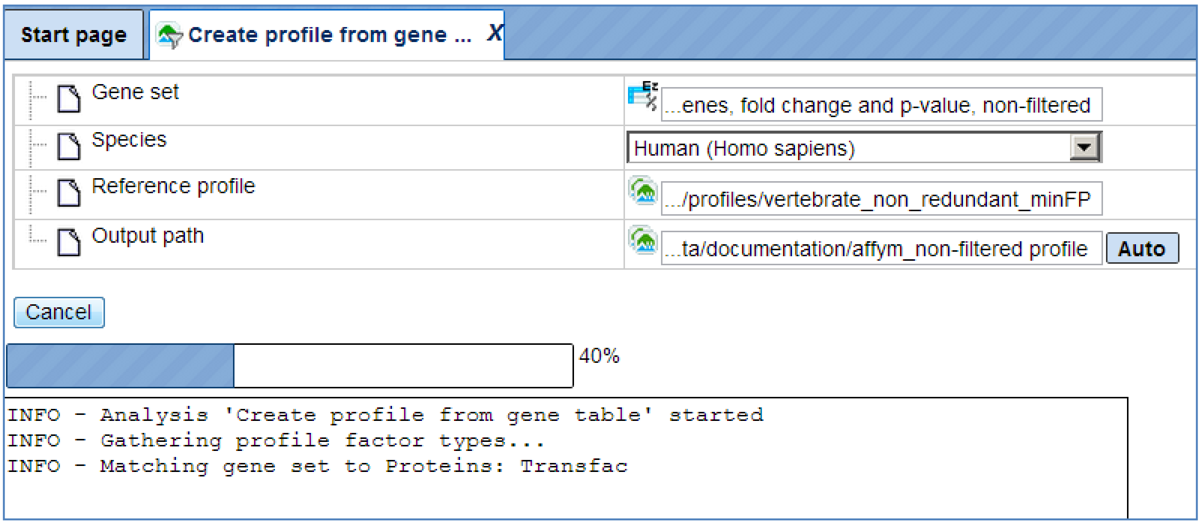
vertebrate_mouse_p0.05
vertebrate_non_redundant
vertebrate_non_redundant_minFN
vertebrate_non_redundant_minFP
vertebrate_non_redundant_minSUM
vertebrate_oryza_10_in_10K_bases
vertebrate_oryza_10_in_1K_bases
vertebrate_oryza_1_in_10K_bases
vertebrate_oryza_1_in_1K_bases
vertebrate_oryza_50_in_1K_bases
vertebrate_oryza_5_in_10K_bases

verte
verte
verte
verte
verte
verte
verte
verte
verte
verte
verte

Name: vertebrate_non_redundant_minFP

Output path: Specify the path to store the result and indicate the name for the new profile.

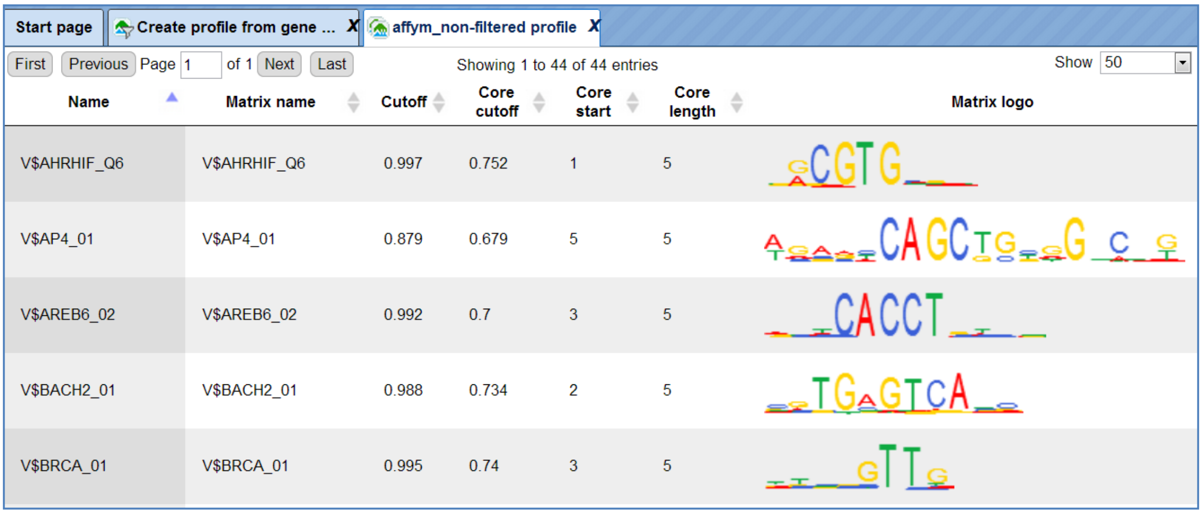
Having filled all the input fields, launch the analysis with the [Run] button. The process will start as shown below:



Gene set	...enes, fold change and p-value, non-filtered
Species	Human (Homo sapiens)
Reference profile	.../profiles/vertebrate_non_redundant_minFP
Output path	...ta/documentation/affym_non-filtered profile Auto

```
INFO - Analysis 'Create profile from gene table' started
INFO - Gathering profile factor types...
INFO - Matching gene set to Proteins: Transfac
```

After completion of the analysis the output profile is opened automatically as shown below:


Start page
Create profile from gene ... X
affym_non-filtered profile X
First Previous Page 1 of 1 Next Last
Showing 1 to 44 of 44 entries
Show 50

Name	Matrix name	Cutoff	Core cutoff	Core start	Core length	Matrix logo
V$AHRHIF_Q6	V$AHRHIF_Q6	0.997	0.752	1	5	
V$AP4_01	V$AP4_01	0.879	0.679	5	5	
V$AREB6_02	V$AREB6_02	0.992	0.7	3	5	
V$BACH2_01	V$BACH2_01	0.988	0.734	2	5	
V$BRCA_01	V$BRCA_01	0.995	0.74	3	5	

Let us now have a look into the newly created profile.

Each row presents the information for one site model. In the column **Name** the name of the site model is given which here is the same as for the matrix. In the column **Matrix name** the name for the positional weight matrix is present. For each site model, a cutoff adapted from the "Reference profile" is shown in the column **Cutoff**. According to the TRANSFAC® standard, the core part is specified for each matrix. The core is represented by the 5 consecutive most conserved nucleotides. The columns **Core cutoff**, **Core start** and **Core length** provide details about the core of each matrix. In the last column the matrix logo is shown.

In the Tree Area, the newly created profile has the symbol , the same symbol as for all other profiles, and is ready to use for the analysis of regulatory regions.

Suggestion when to use "Create profile from gene table"

If you plan to do a promoter analysis for differentially expressed genes from a particular experiment, you might be interested in creating a profile from a table of genes expressed under the same conditions. For example, you might take one of the tables resulting from the workflows "Detect differentially expressed genes…" as input gene table for profile creation.

The example table is a list of non-filtered genes from the workflow "Detect differentially expressed genes…":
http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Experiment%20normalized%20(RMA)%20(Differentially%20expressed%20genes%20Affy)/Genes%2C%20fold%20change%20and%20p-value%2C%20non-filtered

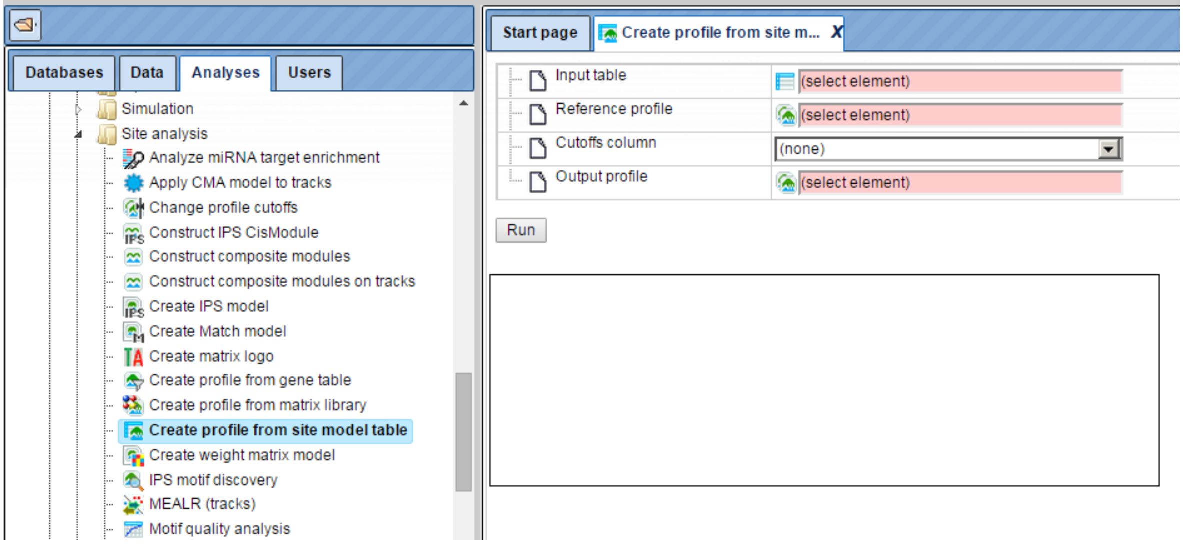
A profile created from this gene table can be found here:
http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Genes%2C%20fold%20change%20and%20p-value%2C%20non-filtered%20profile

Such a profile contains matrices for the transcription factors expressed under the same experimental conditions, and thus it might be reasonable to apply it for the promoter analysis of genes up-regulated or down-regulated in this experiment.

20.1.3.2. Create profile from site model table

This function creates a profile from the table of site models. For example, a new profile can be created from the *summary* table resulting from the workflow "Analyze promoters (TRANSFAC®)" or "Upstream analysis (TRANSFAC® and TRANSPATH®)".

The analysis input form can be found in the Tree Area, under the Analyses tab, in the folder Methods/Site analysis. The opened input form is shown below.


Databases
Data
Analyses
Users
Simulation
Site analysis
Analyze miRNA target enrichment
Apply CMA model to tracks
Change profile cutoffs
Construct IPS CisModule
Construct composite modules
Construct composite modules on tracks
Create IPS model
Create Match model
Create matrix logo
Create profile from gene table
Create profile from matrix library
Create profile from site model table
Create weight matrix model
IPS motif discovery
MEALR (tracks)
Motif quality analysis
Start page
Create profile from site m...
Run

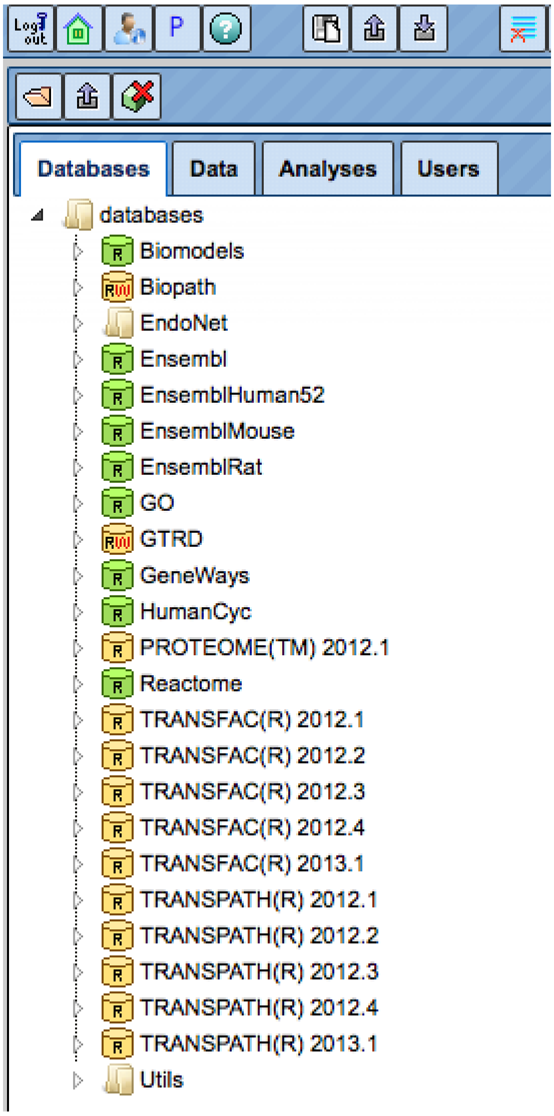
Input table	(select element)
Reference profile	(select element)
Cutoffs column	(none)
Output profile	(select element)

Let us consider the individual input fields:

Input table: Input a table with matrices for which you intend to design a profile. The input table should contain TRANSFAC® matrix IDs as row names. Such a table can be the result of a site search analysis.

In the following, further steps are demonstrated with a table of site models from one of the examples. The example file can be accessed using the URL:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Analyze%20Promoters/Upregulated%20(filtered)_Transfac/Site%20search%20-1000%20100/summary

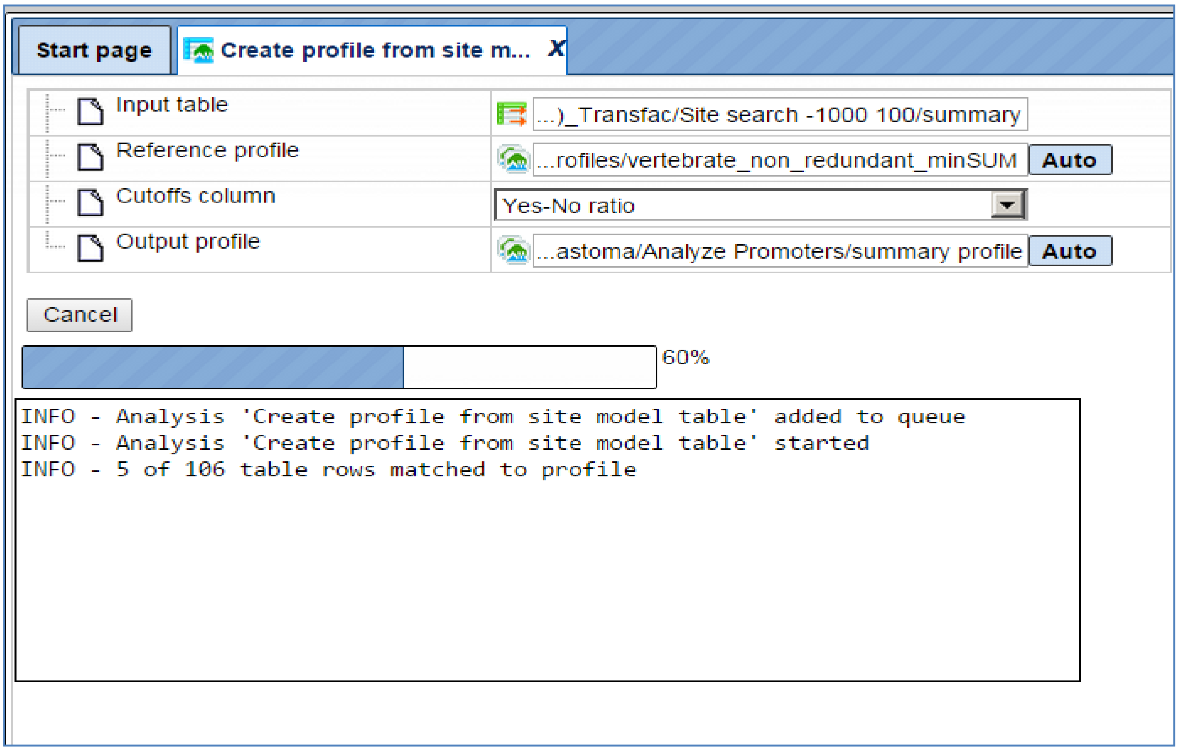


Reference profile: In this field, you indicate the profile where the cutoffs for the new profile should be taken from. It is filled automatically if the input table is a result of a site search analysis or derived from it, and it corresponds to the profile used for site search. You also have an option to manually select the reference profile from the available TRANSFAC® profiles.

Cutoffs column: In this field you specify which column of the input table should be considered for the cutoff values. It is filled automatically if the input table is the result of a site search optimization analysis. It is advisable to use "(none)" to leave the cutoffs as in the reference profile.

Output profile: Specify the path to store the result and indicate the name for the new profile.

To launch the analysis, fill the input fields and press [Run]. The process will start as shown below:

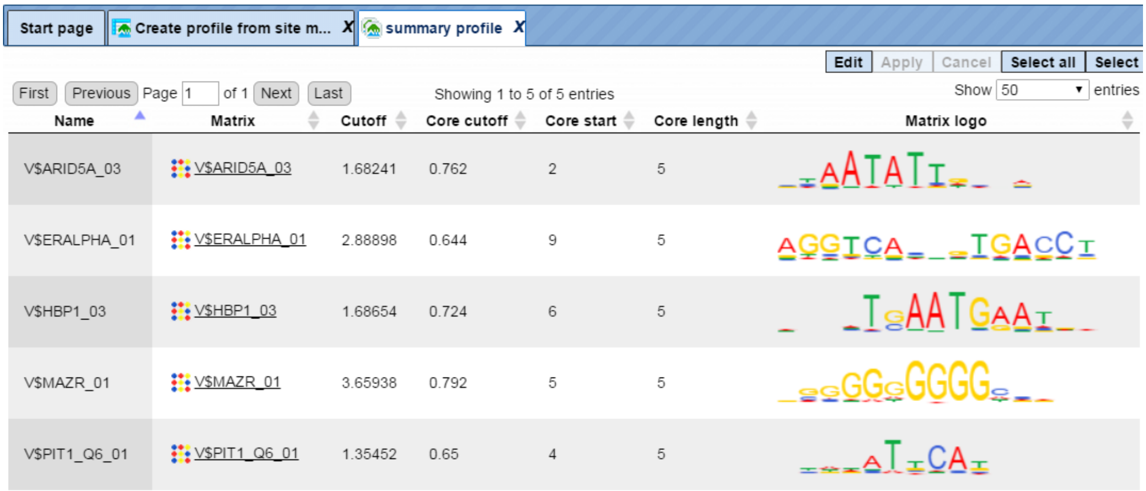


Input table	...)_Transfac/Site search -1000 100/summary
Reference profile	...rofiles/vertebrate_non_redundant_minSUM
Cutoffs column	Yes-No ratio
Output profile	...astoma/Analyze Promoters/summary profile

60%

```
INFO - Analysis 'Create profile from site model table' added to queue
INFO - Analysis 'Create profile from site model table' started
INFO - 5 of 106 table rows matched to profile
```

After completion of the analysis the output profile is opened automatically as shown below:



Showing 1 to 5 of 5 entries

Name	Matrix	Cutoff	Core cutoff	Core start	Core length	Matrix logo
V$ARID5A_03	V$ARID5A_03	1.68241	0.762	2	5	
V$ERALPHA_01	V$ERALPHA_01	2.88898	0.644	9	5	
V$HBP1_03	V$HBP1_03	1.68654	0.724	6	5	
V$MAZR_01	V$MAZR_01	3.65938	0.792	5	5	
V$PIT1_Q6_01	V$PIT1_Q6_01	1.35452	0.65	4	5	

An example profile created from this gene table can be found here:

http://genexplain-platform.com/bioumlweb/#de=data/Examples/Brain%20Tumor%20GSE1825%2C%20Affymetrix%20HG-U133A%20microarray/Data/Ewing%20Family%20Tumor%20versus%20Neuroblastoma/Analyze%20Promoters/summary%20profile

Each row summarizes the information for one site model. In the column **Name** the name of the site model is given, which in the majority of cases is the same as for the matrix. In the column **Matrix name** the name for the positional weight matrix is present. For each site model, cutoff is shown in the column **Cutoff**. According to the TRANSFAC® standard, the core part is specified for each matrix. The core is represented by the 5 consecutive most conserved nucleotides. The columns **Core cutoff**, **Core start** and **Core length** give details about the core of each matrix. In the last column the matrix logo of each matrix is shown.

The new created profile has the symbol , the same symbol as for all other profiles, and is ready to use for the analysis of regulatory regions.

Suggestion when to use "Create profile from table"

Creating a new profile from a table of site models (matrices) might be very useful e.g. if you plan to focus the analysis on those matrices shown to be over-represented upon the first run of promoter analysis and plan further to construct composite promoter models.

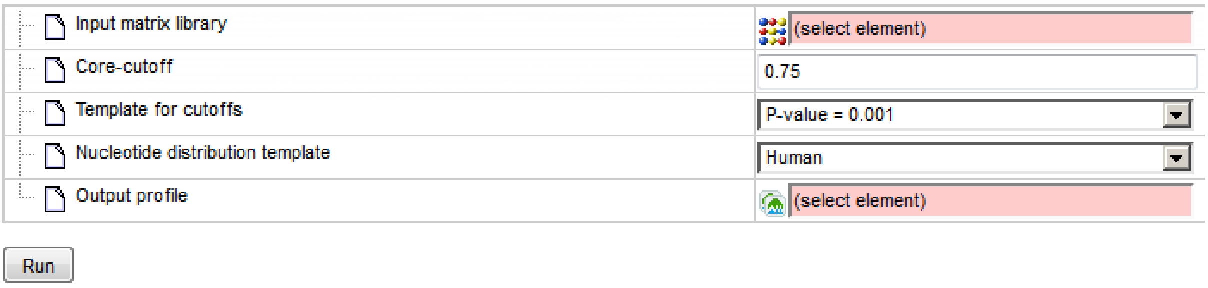
For this, you first run a site search with optimization, e.g. the workflow "Analyze promoters (TRANSFAC®)" or "Upstream analysis (TRANSFAC® and TRANSPATH®)". Then, you can use the table *summary* as an input to construct your specific profile.

It will contain only those site models (matrices) for which hits are over-represented in your genes of interest versus the background set, which is a specific subset of the TRANSFAC® matrix library. Having such a profile specific for your genes of interest, you can run it on the same set of genes without optimization, and then use the results as input to construct composite promoter models.

20.1.3.3. Create profile from matrix library

This tool can set score cutoffs for an entire matrix library and store a corresponding profile, which in turn can be applied binding site analyses.

The figure below shows the default input mask. There are two general ways of setting score cutoffs, either by P-value (the default) or as one cutoff value that will be applied to all matrices.



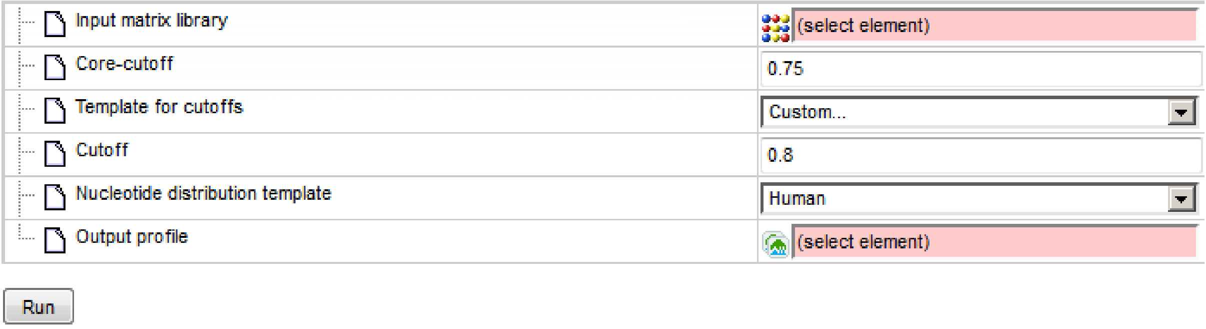
Input matrix library	(select element)
Core-cutoff	0.75
Template for cutoffs	P-value = 0.001
Nucleotide distribution template	Human
Output profile	(select element)

The input parameters are as follows.

Input matrix library: Selection of the matrix library. All matrices of the library will be present in the profile ("Output profile").

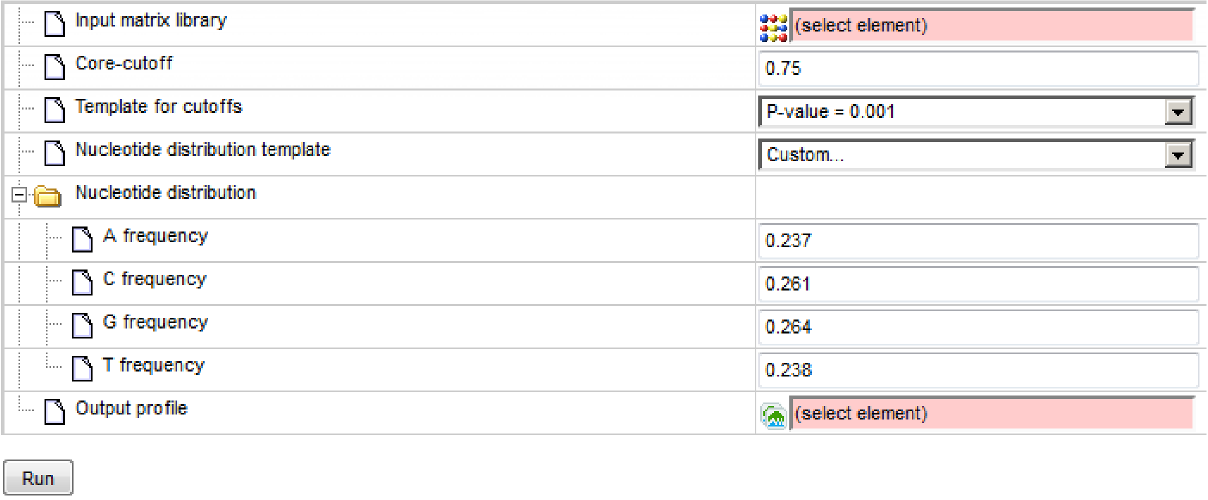
Core-cutoff: The cutoff for the matrix core, which is the five consecutive matrix position which are most selective for certain nucleotides. When the matrix cutoff is defined by a P-value, the core cutoff will be 0 for all matrices.

Template for cutoffs: The P-value threshold according to which the score cutoff for each matrix will be set. Selecting the "Custom…" option presents a new field to set one common cutoff for all matrices of the library.



Input matrix library	(select element)
Core-cutoff	0.75
Template for cutoffs	Custom...
Cutoff	0.8
Nucleotide distribution template	Human
Output profile	(select element)

Nucleotide distribution template: The distribution of nucleotides on which P-value calculations are based. Selecting the "Custom…" option enables custom setting of individual base frequencies.

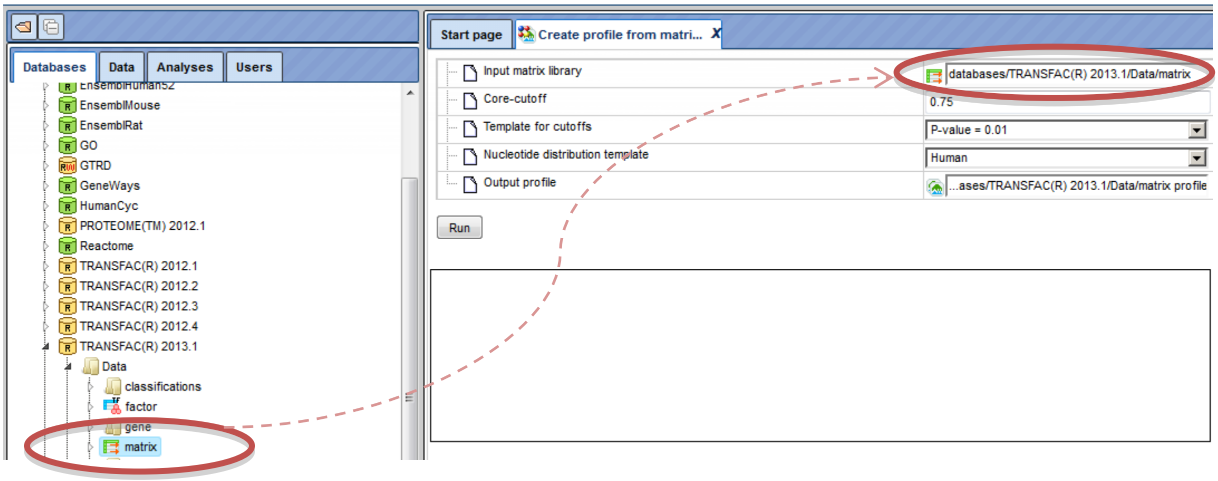


Input matrix library	(select element)
Core-cutoff	0.75
Template for cutoffs	P-value = 0.001
Nucleotide distribution template	Custom...
Nucleotide distribution	
A frequency	0.237
C frequency	0.261
G frequency	0.264
T frequency	0.238
Output profile	(select element)

Output profile: The path for the output profile.

An profile can be created as follows.

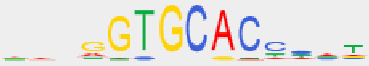
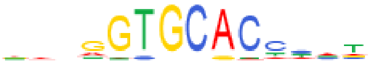
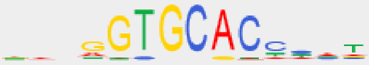
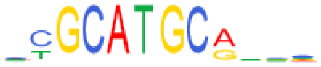
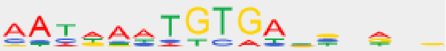
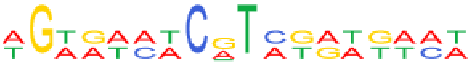
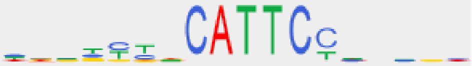
Step 1. Input the matrix library. As usual, you can drag-and-drop. Here we use the TRANSFAC® 2013.1 library:



Input matrix library	databases/TRANSFAC(R) 2013.1/Data/matrix
Core-cutoff	0.75
Template for cutoffs	P-value = 0.01
Nucleotide distribution template	Human
Output profile	...ases/TRANSFAC(R) 2013.1/Data/matrix profile

Step 2. Edit other input parameters as shown in the figure above. The output profile path needs to be chosen in a writable directory, e.g. one of your projects.

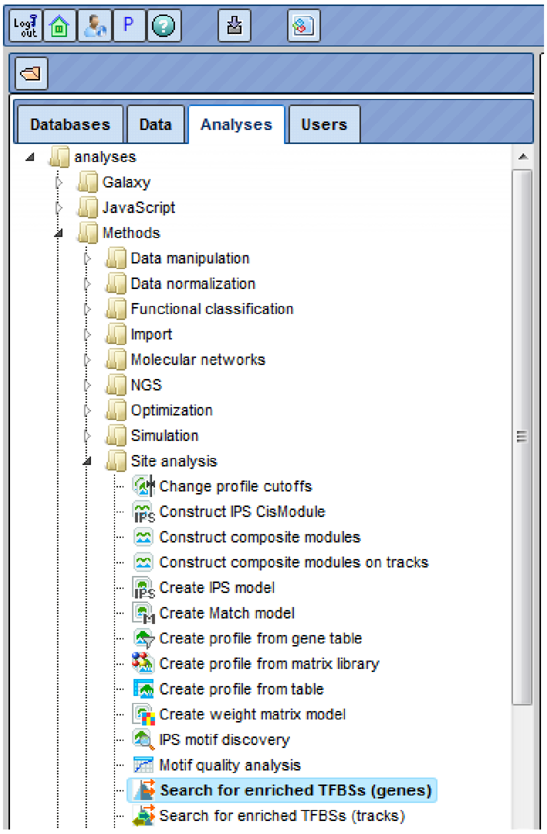
Clicking the [Run] button will invoke the analysis. A part of the resulting profile is shown below:

Name	Matrix name	Cutoff	Core cutoff	Core start	Core length	Matrix logo
A$APIAP2_01	A$APIAP2_01	0.7	0	5	5	
A$APIAP2_02	A$APIAP2_02	0.7	0	5	5	
A$APIAP2_03	A$APIAP2_03	0.7	0	5	5	
A$CGD2	A$CGD2	0.75	0	4	5	
B$CRP_C	B$CRP_C	0.62069	0	7	5	
B$TRAM_01	B$TRAM_01	0.73333	0	5	5	
F$ABAA_01	F$ABAA_01	0.75	0	7	5	

20.1.4. Search for enriched TFBSs

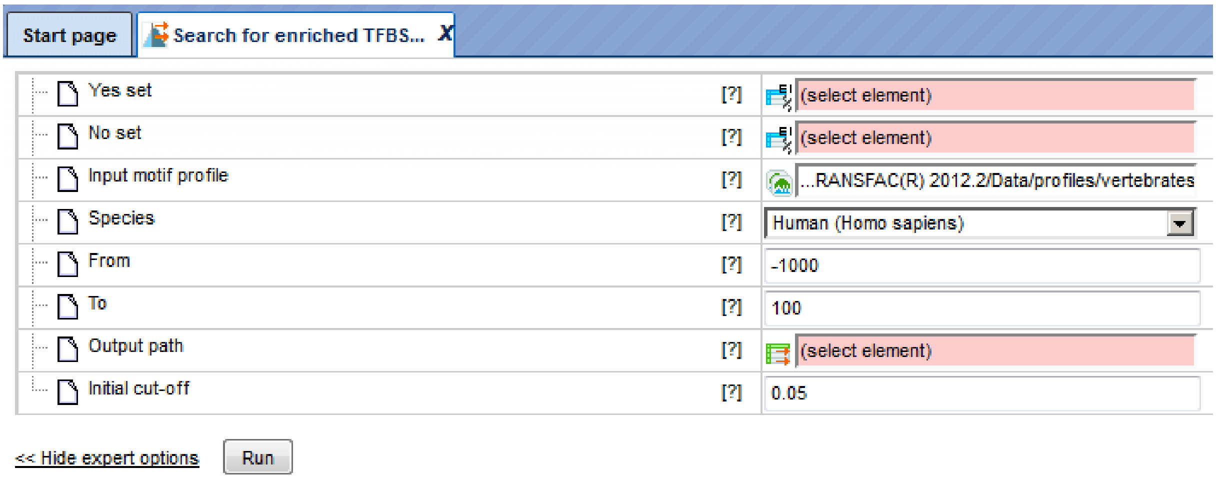
The platform comprises tools that are dedicated to discovering types of binding sites enriched in a set of sequences. They are named “Search for enriched TFBSs (genes)” and “Search for enriched TFBSs (tracks)”. Both apply a common core algorithm to gene promoters or tracks, respectively. The “Search for enriched TFBSs”-tools are accessible in the “Analysis” tab under “analyses/Methods/Site Analysis”.

Dedication to discovering enriched types of binding sites means that these methods won't bother with storing predicted binding sites and they also do not require predictions as input. Their sole output is a relatively small table summarizing the enrichment detected for PWM models in a sequence set of interest. As a result, the "*Search for enriched TFBSs*"-tools are often considerably faster than solutions that have to read and write binding site information. In addition, their outputs consume only the disk space needed to present the results of analyzing binding site enrichment. Hence, a "Search for enriched TFBSs" is for you if you wish to gain insight into enriched binding sites quickly and in an uncomplicated way.



20.1.4.1. Search for enriched TFBSs (genes)

Let us begin with the analysis for gene promoters. The figure below depicts the input mask of the analysis tool. The parameters are similar to those used by "Site search on gene set" and are described in the following.



Yes set	[?]	(select element)
No set	[?]	(select element)
Input motif profile	[?]	...RANSFAC(R) 2012.2/Data/profiles/vertebrates
Species	[?]	Human (Homo sapiens)
From	[?]	-1000
To	[?]	100
Output path	[?]	(select element)
Initial cut-off	[?]	0.05

Yes set: This is the set of genes that you want to analyze, for example these can be genes with altered expression. The program accepts genes specified by Ensembl gene identifiers. Note that the "Convert table" functionality in the "Data manipulation" folder can map other identifiers to the required Ensembl genes.

No set: This is the set of background genes (control set), which also need to be specified by Ensemble gene IDs.

These two datasets might be taken from the output tables of previous analyses; see, e.g., "Detect differentially expressed genes" (Section 10.2).

Species: This option specifies the biological species of Yes and No gene sets. Currently, the method is applicable to human, mouse or rat genes.

From and **To**: These values define the length and location of promoter regions (upstream and downstream) relative to the transcription start site (TSS). Respective sequence regions are extracted for each Yes and No gene according to Ensembl annotation.

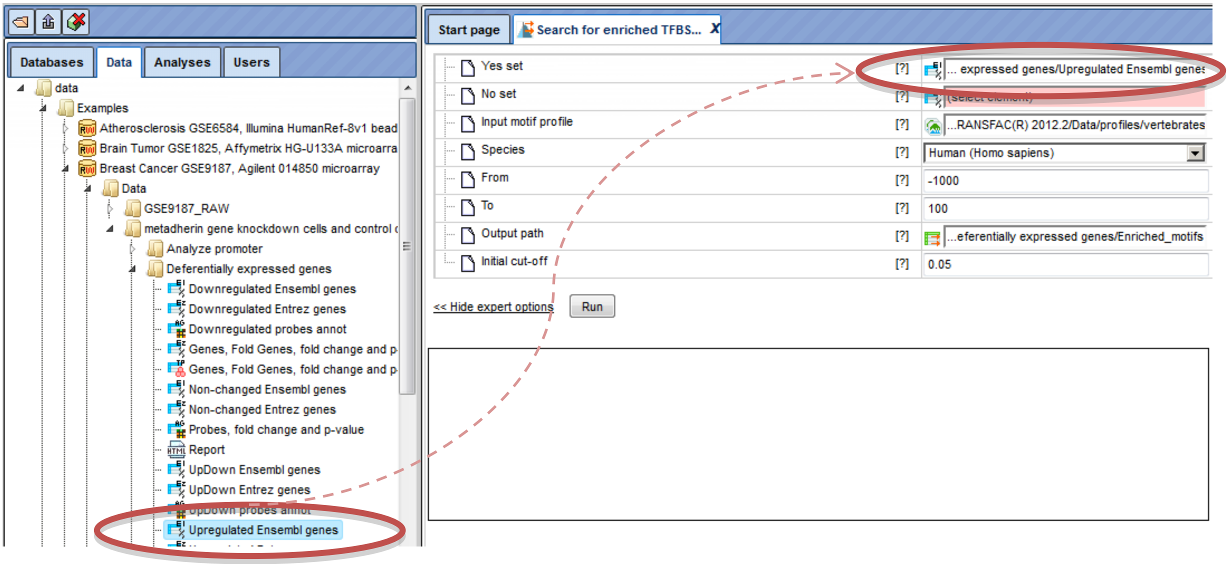
Input motif profile: The profile lists the PWMs (motifs) to be used for binding site prediction together with a score cutoff. By default, this field is set to the last profiled set in your workspace. Note that cutoffs in the profile are ignored, because the "Search for enriched TFBSs" determines a starting threshold specified in the Initial cutoff field.

Output path: In this field you select a path in the workspace to store the output table.

Initial cutoff: This cutoff determines the starting point of the analysis with respect to the score threshold to predict binding sites. The choice is expressed as frequency of predicted sites per base. By default, the analysis begins with a frequency of 5 sites in 100 bases (0.05). From the initial cutoff the algorithm iterates over higher cutoffs and eventually reports the one that resulted in optimal enrichment of binding sites in the Yes set. This is done separately for each PWM.

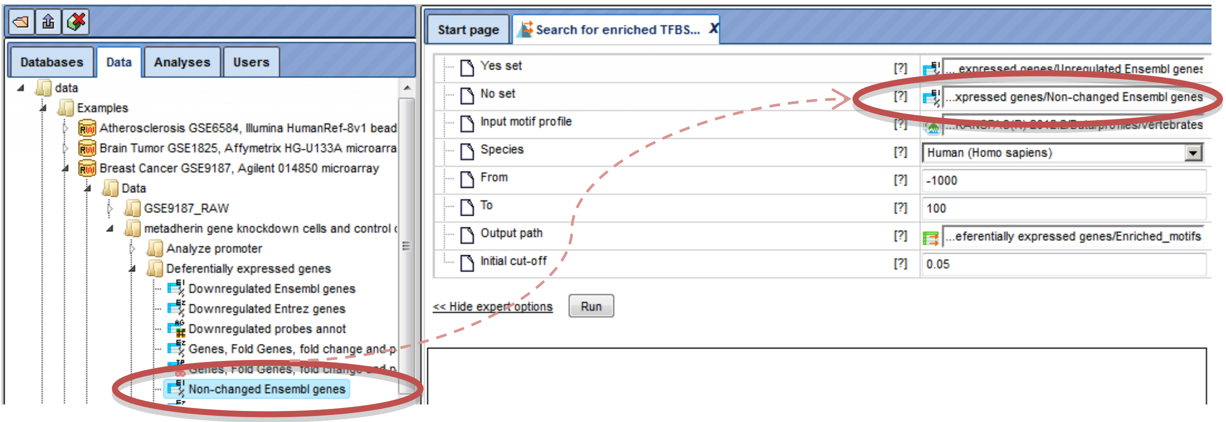
The steps of an analysis can be described as follows:

Step 1. Input Yes set from the tree. As usual, you can drag-and-drop. Here, the set of genes from the Example folder is used as input Yes set, highlighted blue on the screenshot below:


Databases
Data
Analyses
Users
data
Examples
Atherosclerosis GSE6584, Illumina HumanRef-8v1 bead
Brain Tumor GSE1825, Affymetrix HG-U133A microarra
Breast Cancer GSE9187, Agilent 014850 microarray
Data
GSE9187_RAW
metadherin gene knockdown cells and control c
Analyze promoter
Deferentially expressed genes
Downregulated Ensembl genes
Downregulated Entrez genes
Downregulated probes annot
Genes, Fold Genes, fold change and p
Genes, Fold Genes, fold change and p
Non-changed Ensembl genes
Non-changed Entrez genes
Probes, fold change and p-value
Report
UpDown Ensembl genes
UpDown Entrez genes
UpDown probes annot
Upregulated Ensembl genes
Start page
Search for enriched TFBS...
<< Hide expert options
Run

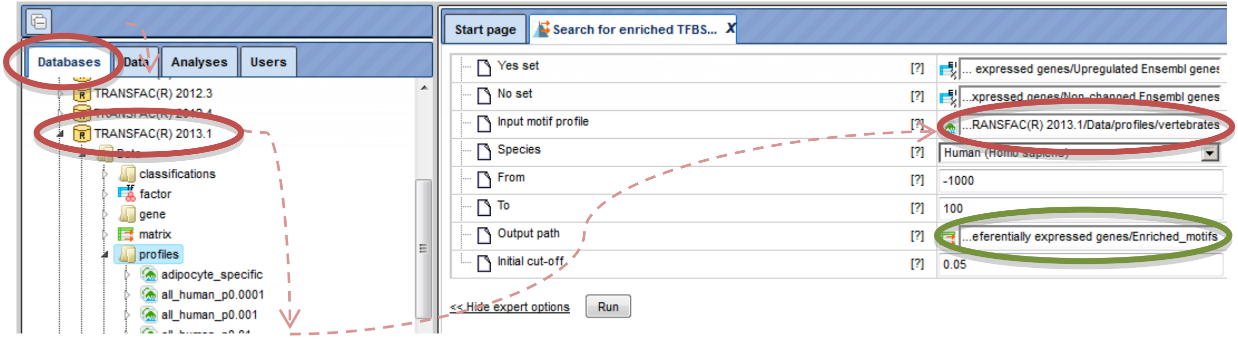
Yes set [?] ... expressed genes/Upregulated Ensembl genes
No set [?] (select element)
Input motif profile [?] ...RANSFAC(R) 2012.2/Data/profiles/vertebrates
Species [?] Human (Homo sapiens)
From [?] -1000
To [?] 100
Output path [?] ...eferentially expressed genes/Enriched_motifs
Initial cut-off [?] 0.05

Step 2. Input No set (drag-and-drop). Our example uses the set of "Non-changed Ensembl genes":


Databases
Data
Analyses
Users
data
Examples
Atherosclerosis GSE6584, Illumina HumanRef-8v1 bead
Brain Tumor GSE1825, Affymetrix HG-U133A microarra
Breast Cancer GSE9187, Agilent 014850 microarray
Data
GSE9187_RAW
metadherin gene knockdown cells and control c
Analyze promoter
Deferentially expressed genes
Downregulated Ensembl genes
Downregulated Entrez genes
Downregulated probes annot
Genes, Fold Genes, fold change and p
Genes, Fold Genes, fold change and p
Non-changed Ensembl genes
Start page
Search for enriched TFBS...
<< Hide expert options
Run

Yes set [?] ...expressed genes/Upregulated Ensembl genes
No set [?] ...xpressed genes/Non-changed Ensembl genes
Input motif profile [?] ...RANSFAC(R) 2013.1/Data/profiles/vertebrates
Species [?] Human (Homo sapiens)
From [?] -1000
To [?] 100
Output path [?] ...eferentially expressed genes/Enriched_motifs
Initial cut-off [?] 0.05

Step 3. Select the TRANSFAC® or GTRD profile from the available profiles. In this example, we select the TRANSFAC® 2013.1 profile named "vertebrates":

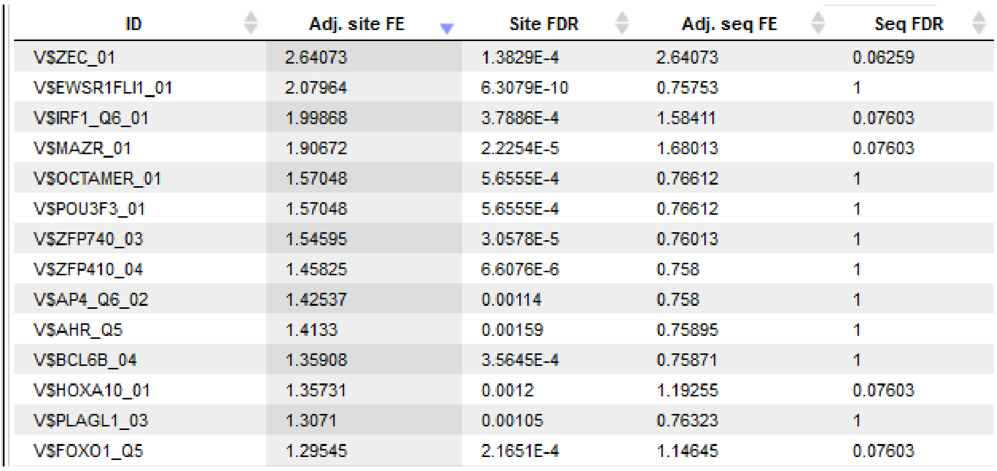


Yes set [?] ... expressed genes/Upregulated Ensembl genes
No set [?] ...xpressed genes/Non-changed Ensembl genes
Input motif profile [?] ...RANSFAC(R) 2013.1/Data/profiles/vertebrates
Species [?] Human (Homo sapiens)
From [?] -1000
To [?] 100
Output path [?] ...eferentially expressed genes/Enriched_motifs
Initial cut-off [?] 0.05

Step 4. Edit the output path (highlighted green in the figure above). After setting the Yes gene set, a default output path is suggested. The Example folder may however not be writable for your account requiring selection of an alternative such as one of your own projects. A different selection can be made easily by clicking on the field.

We keep the defaults for promoter range (starting from 1000th base upstream to the 100th base downstream of the TSS) and initial cutoff.

Clicking the [Run] button will invoke the analysis. The *summary* table ▦ is automatically opened in a new tab when the analysis is completed. Here is a part of the output for our example:

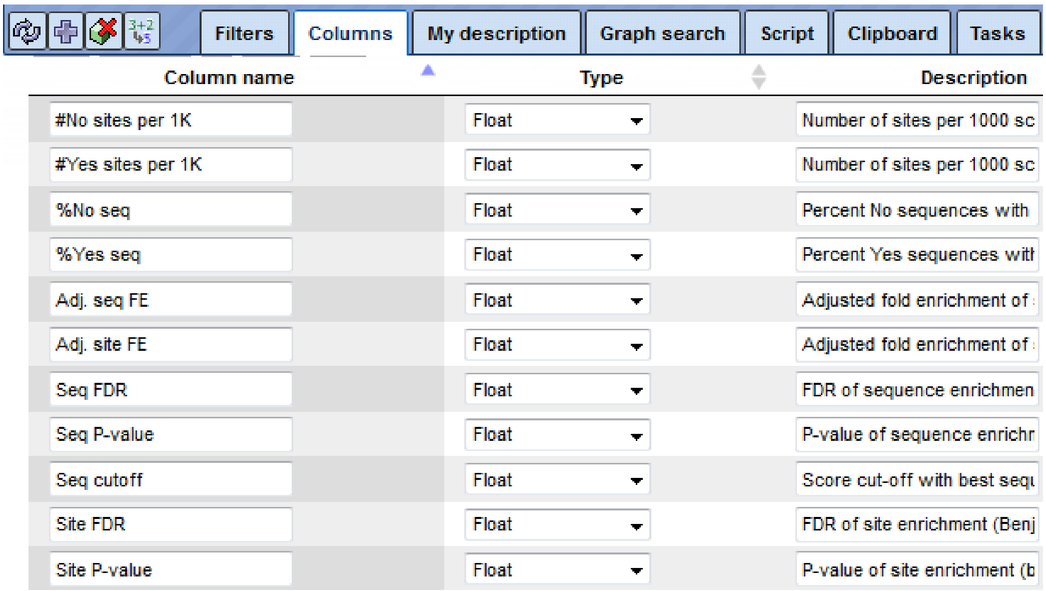


ID	Adj. site FE	Site FDR	Adj. seq FE	Seq FDR
V$ZEC_01	2.64073	1.3829E-4	2.64073	0.06259
V$EWSR1FLI1_01	2.07964	6.3079E-10	0.75753	1
V$IRF1_Q6_01	1.99868	3.7886E-4	1.58411	0.07603
V$MAZR_01	1.90672	2.2254E-5	1.68013	0.07603
V$OCTAMER_01	1.57048	5.6555E-4	0.76612	1
V$POU3F3_01	1.57048	5.6555E-4	0.76612	1
V$ZFP740_03	1.54595	3.0578E-5	0.76013	1
V$ZFP410_04	1.45825	6.6076E-6	0.758	1
V$AP4_Q6_02	1.42537	0.00114	0.758	1
V$AHR_Q5	1.4133	0.00159	0.75895	1
V$BCL6B_04	1.35908	3.5645E-4	0.75871	1
V$HOXA10_01	1.35731	0.0012	1.19255	0.07603
V$PLAGL1_03	1.3071	0.00105	0.76323	1
V$FOXO1_Q5	1.29545	2.1651E-4	1.14645	0.07603

Each row of the output table represents the result for one PWM from the input profile. Fold enrichment (FE) values quantify the enrichment of binding sites in the Yes sequence set as a whole (**Adj. site FE**) and of Yes sequences with at least one binding site (**Adj. seq FE**). The FE is an odds ratio that compares the ratio of Yes sites to No sites with the ratio of Yes sequences to No sequences. Unlike in other tools the FE values are statistically corrected (adjusted) quantities that report a value below the actually observed ratio according to the 99% confidence interval. This correction takes into account the underlying site numbers and sequence numbers and penalizes enrichment values that are based on only few binding site occurrences, e.g. at a high score threshold. The output table is sorted by the adjusted Site FE by default.

The statistical significance of enrichment is further assessed by one-tailed binomial test and Fisher tests P-values, for which False Discovery Rates (FDRs) are reported. The **Site FDR** is based on binomial test P-values calculated for the number of binding site in Yes and No set. The **Seq FDR** column contains FDRs for Fisher test P-values for the number of sequences with at one site in Yes and No sets.

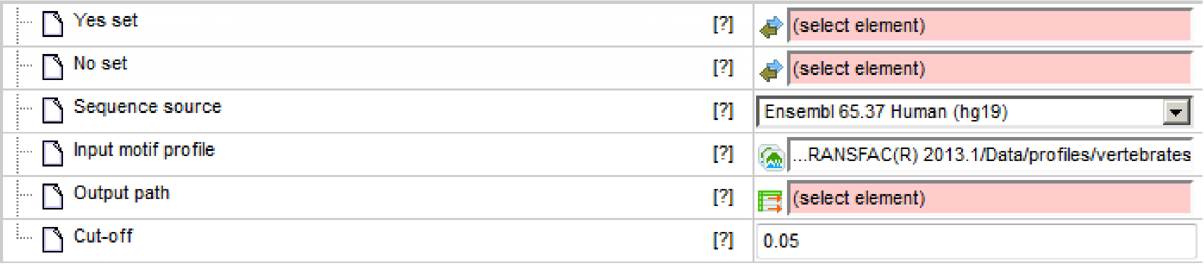
Additional columns are available via the "Columns" tab of the lower-right panel. Each column is also accompanied by a concise description and can be included into the table presentation upon demand.



Column name	Type	Description
#No sites per 1K	Float	Number of sites per 1000 sc
#Yes sites per 1K	Float	Number of sites per 1000 sc
%No seq	Float	Percent No sequences with
%Yes seq	Float	Percent Yes sequences wit
Adj. seq FE	Float	Adjusted fold enrichment of
Adj. site FE	Float	Adjusted fold enrichment of
Seq FDR	Float	FDR of sequence enrichmen
Seq P-value	Float	P-value of sequence enrichr
Seq cutoff	Float	Score cut-off with best sequ
Site FDR	Float	FDR of site enrichment (Benj
Site P-value	Float	P-value of site enrichment (b

20.1.4.2. Search for enriched TFBSs (tracks)

The "tracks" variant of the "Search for enriched TFBSs" can be applied to sequence tracks signified by the symbol (). For instance, a track may contain genomic intervals identified by a ChIP-seq experiment.



Yes set	[?]	(select element)
No set	[?]	(select element)
Sequence source	[?]	Ensembl 65.37 Human (hg19)
Input motif profile	[?]	...RANSFAC(R) 2013.1/Data/profiles/vertebrates
Output path	[?]	(select element)
Cut-off	[?]	0.05

The analysis uses the following parameters:

Yes set: This is the track that you want to analyze, for example these can be ChIP-seq intervals bound a transcription factor.

No set: This is the set of background intervals (control set).

Sequence source: Both Yes and No track need to refer to a common source, such as a genome, as specified by this parameter. Note that you can apply a custom source, e.g. a specifically uploaded genome. Clicking on the "Custom" option will open a new field to choose the custom sequence source.

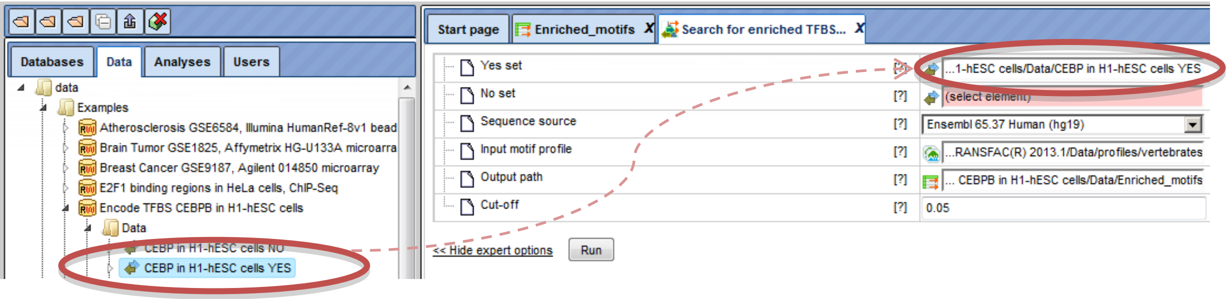
Input motif profile: The profile lists the PWMs (motifs) to be used for binding site prediction together with a score cutoff. By default, this field is set to the last profiled set in your workspace. Note that cutoffs in the profile are ignored, because the "Search for enriched TFBSs" determines a starting threshold specified in the Initial cutoff field.

Output path: In this field you select a path in the workspace to store the output table.

Initial cutoff: This cutoff determines the starting point of the analysis with respect to the score threshold to predict binding sites. The choice is expressed as frequency of predicted sites per base. By default, the analysis begins with a frequency of 5 sites in 100 bases (0.05). From the initial cutoff the algorithm iterates over higher cutoffs and eventually reports the one that resulted in optimal enrichment of binding sites in the Yes set. This is done separately for each PWM.

The platform is provides an out-of-the-box example for this tool under "*data/Examples/ Encode TFBS CEBPB in H1-hESC cells*". The ChIP-seq experiment targeted CEBPB binding sites in H1-hESC cells. The steps of an analysis can be described as follows:

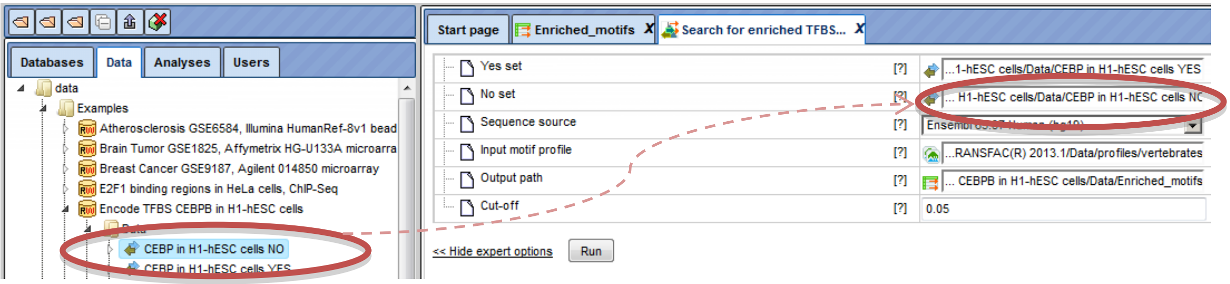
Step 1. Input the Yes set from the tree. As usual, you can drag-and-drop. The YES set contains the 500 most significant peaks:



data
Examples
Atherosclerosis GSE6584, Illumina HumanRef-8v1 bead
Brain Tumor GSE1825, Affymetrix HG-U133A microarra
Breast Cancer GSE9187, Agilent 014850 microarray
E2F1 binding regions in HeLa cells, ChIP-Seq
Encode TFBS CEBPB in H1-hESC cells
Data
CEBP in H1-hESC cells NO
CEBP in H1-hESC cells YES

Yes set	[?]	...1-hESC cells/Data/CEBP in H1-hESC cells YES
No set	[?]	(select element)
Sequence source	[?]	Ensembl 65.37 Human (hg19)
Input motif profile	[?]	...RANSFAC(R) 2013.1/Data/profiles/vertebrates
Output path	[?]	... CEBPB in H1-hESC cells/Data/Enriched_motifs
Cut-off	[?]	0.05

Step 2. Input the No set (drag-and-drop). The NO set contains 1000 random intervals from promoter regions with the same length distribution as the 500 YES intervals:



- data
  - Examples
    - Atherosclerosis GSE6584, Illumina HumanRef-8v1 bead
    - Brain Tumor GSE1825, Affymetrix HG-U133A microarra
    - Breast Cancer GSE9187, Agilent 014850 microarray
    - E2F1 binding regions in HeLa cells, ChIP-Seq
    - Encode TFBS CEBPB in H1-hESC cells
      - Data
        - CEBP in H1-hESC cells NO
        - CEBP in H1-hESC cells YES

Yes set [?] ...1-hESC cells/Data/CEBP in H1-hESC cells YES

No set [?] ... H1-hESC cells/Data/CEBP in H1-hESC cells NC

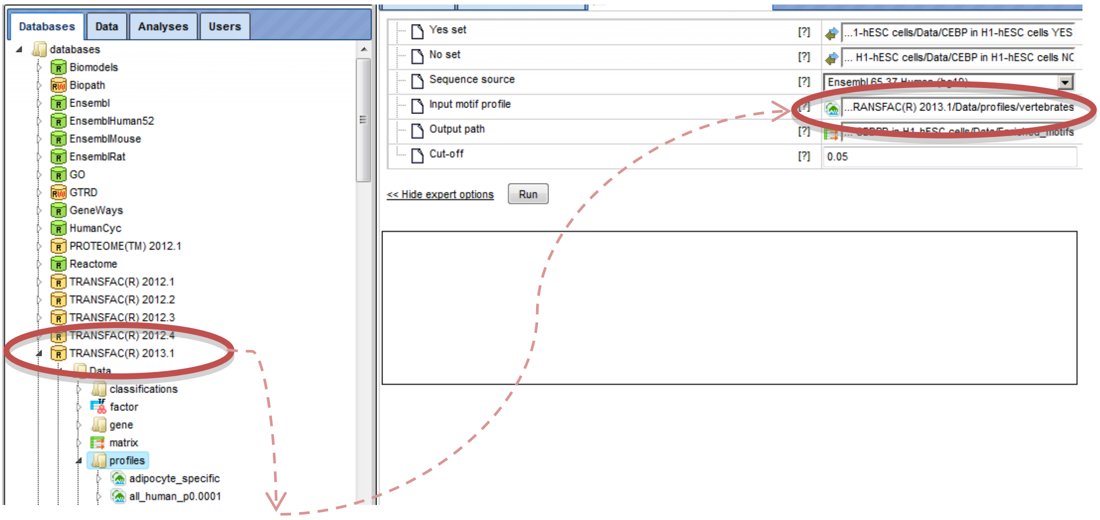
Sequence source [?] Ensembl 65.37 Human (hg19)

Input motif profile [?] ...RANSFAC(R) 2013.1/Data/profiles/vertebrates

Output path [?] ... CEBPB in H1-hESC cells/Data/Enriched_motifs

Cut-off [?] 0.05

Step 3. Select the TRANSFAC® or GTRD profile from the available profiles. In this example, we select the TRANSFAC® 2013.1 profile named "vertebrates":



- databases
  - Biomodels
  - Biopath
  - Ensembl
  - EnsemblHuman52
  - EnsemblMouse
  - EnsemblRat
  - GO
  - GTRD
  - GeneWays
  - HumanCyc
  - PROTEOME(TM) 2012.1
  - Reactome
  - TRANSFAC(R) 2012.1
  - TRANSFAC(R) 2012.2
  - TRANSFAC(R) 2012.3
  - TRANSFAC(R) 2012.4
  - TRANSFAC(R) 2013.1
    - Data
      - classifications
      - factor
      - gene
      - matrix
      - profiles
        - adipocyte_specific
        - all_human_p0.0001

Yes set [?] ...1-hESC cells/Data/CEBP in H1-hESC cells YES

No set [?] ... H1-hESC cells/Data/CEBP in H1-hESC cells NC

Sequence source [?] Ensembl 65.37 Human (hg19)

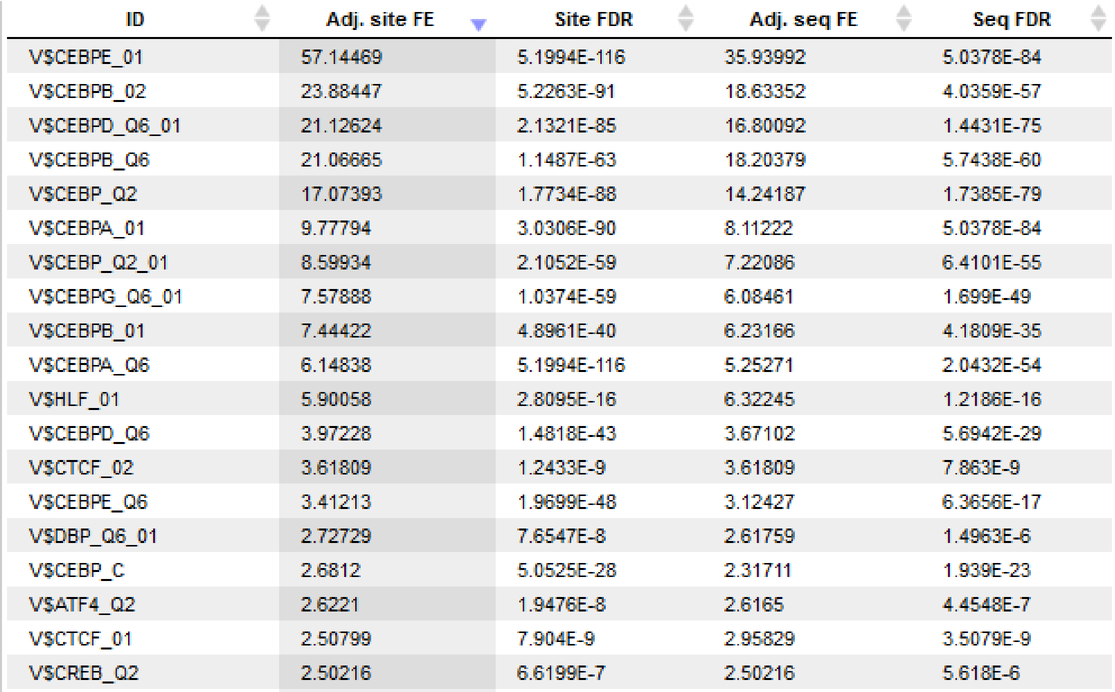
Input motif profile [?] ...RANSFAC(R) 2013.1/Data/profiles/vertebrates

Output path [?] ... CEBPB in H1-hESC cells/Data/Enriched_motifs

Cut-off [?] 0.05

Step 4. Edit the output path. After setting the Yes gene set, a default output path is suggested. The Example folder may however not be writable for your account requiring selection of an alternative such as one of your own projects. A different selection can be made easily by clicking on the field.

Clicking the [Run] button will invoke the analysis. The *summary* table is automatically opened in a new tab when the analysis is completed. A part of the output for our example is shown below. Please refer to "Search for enriched TFBSs (genes)" for further description.



ID	Adj. site FE	Site FDR	Adj. seq FE	Seq FDR
V$CEBPE_01	57.14469	5.1994E-116	35.93992	5.0378E-84
V$CEBPB_02	23.88447	5.2263E-91	18.63352	4.0359E-57
V$CEBPD_Q6_01	21.12624	2.1321E-85	16.80092	1.4431E-75
V$CEBPB_Q6	21.06665	1.1487E-63	18.20379	5.7438E-60
V$CEBP_Q2	17.07393	1.7734E-88	14.24187	1.7385E-79
V$CEBPA_01	9.77794	3.0306E-90	8.11222	5.0378E-84
V$CEBP_Q2_01	8.59934	2.1052E-59	7.22086	6.4101E-55
V$CEBPG_Q6_01	7.57888	1.0374E-59	6.08461	1.699E-49
V$CEBPB_01	7.44422	4.8961E-40	6.23166	4.1809E-35
V$CEBPA_Q6	6.14838	5.1994E-116	5.25271	2.0432E-54
V$HLF_01	5.90058	2.8095E-16	6.32245	1.2186E-16
V$CEBPD_Q6	3.97228	1.4818E-43	3.67102	5.6942E-29
V$CTCF_02	3.61809	1.2433E-9	3.61809	7.863E-9
V$CEBPE_Q6	3.41213	1.9699E-48	3.12427	6.3656E-17
V$DBP_Q6_01	2.72729	7.6547E-8	2.61759	1.4963E-6
V$CEBP_C	2.6812	5.0525E-28	2.31711	1.939E-23
V$ATF4_Q2	2.6221	1.9476E-8	2.6165	4.4548E-7
V$CTCF_01	2.50799	7.904E-9	2.95829	3.5079E-9
V$CREB_Q2	2.50216	6.6199E-7	2.50216	5.618E-6

20.1.5. Construct composite modules

Composite modules are combinations of several TFBSs that are found together in a set of regulatory sequences. We search for such combinations of TF binding sites that are overrepresented in the regulatory sequences under study compared to a background set of sequences. The search for composite modules is performed using an in-house implementation of a genetic algorithm. As input for the genetic algorithm we take the output of a site search analysis.

There are two individual analysis functions available with the same symbol ; they are different with respect to the type of sequences where the search for composite modules is done, and correspondingly with respect to the format of the input data.

- ***Construct composite modules*** *analysis works on the promoter sequences specified relative to TSS in the set of genes. As input, it takes the results of the Site search on gene set analysis function.*
- ***Construct composite modules on tracks*** *works with any DNA sequences specified by their absolute genomic positions, and is very often applied for the analysis of ChIP-seq fragments. As input, it takes the results of the* ***Site search on track analysis*** *function.*

Both analysis functions can be found in the geneXplain platform online under the path

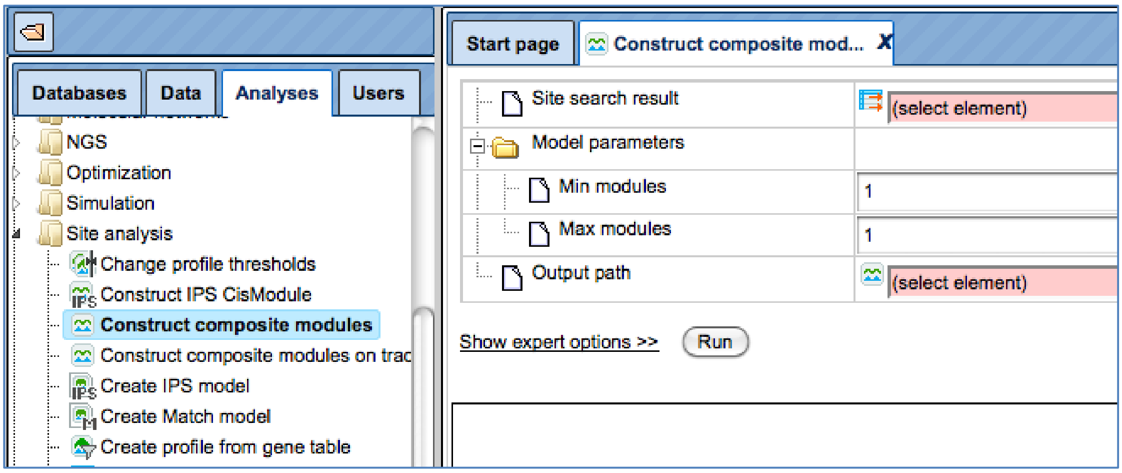
http://genexplain-platform.com/bioumlweb/#de=analyses/Methods/Site%20analysis

20.1.5.1. Construct composite modules

This analysis function enables the identification of combinations of several TFBSs in the promoters of the genes under study (Yes-set). The resulting composite module differentiates the Yes-set from a background set (No-set).

Before starting this analysis, you need to perform *Site search on gene set* with your selected Yes-set, No-set and a specified profile of matrices. If you are interested in finding site models for particular TFs, and see them eventually in the resulting composite modules, you need to be sure that such matrices are present in the selected profile. You can use one of the available TRANSFAC® profiles, or alternatively you can construct a customized profile; for details please see Section 20.1.3.3*.*

The input form for this analysis is shown below:



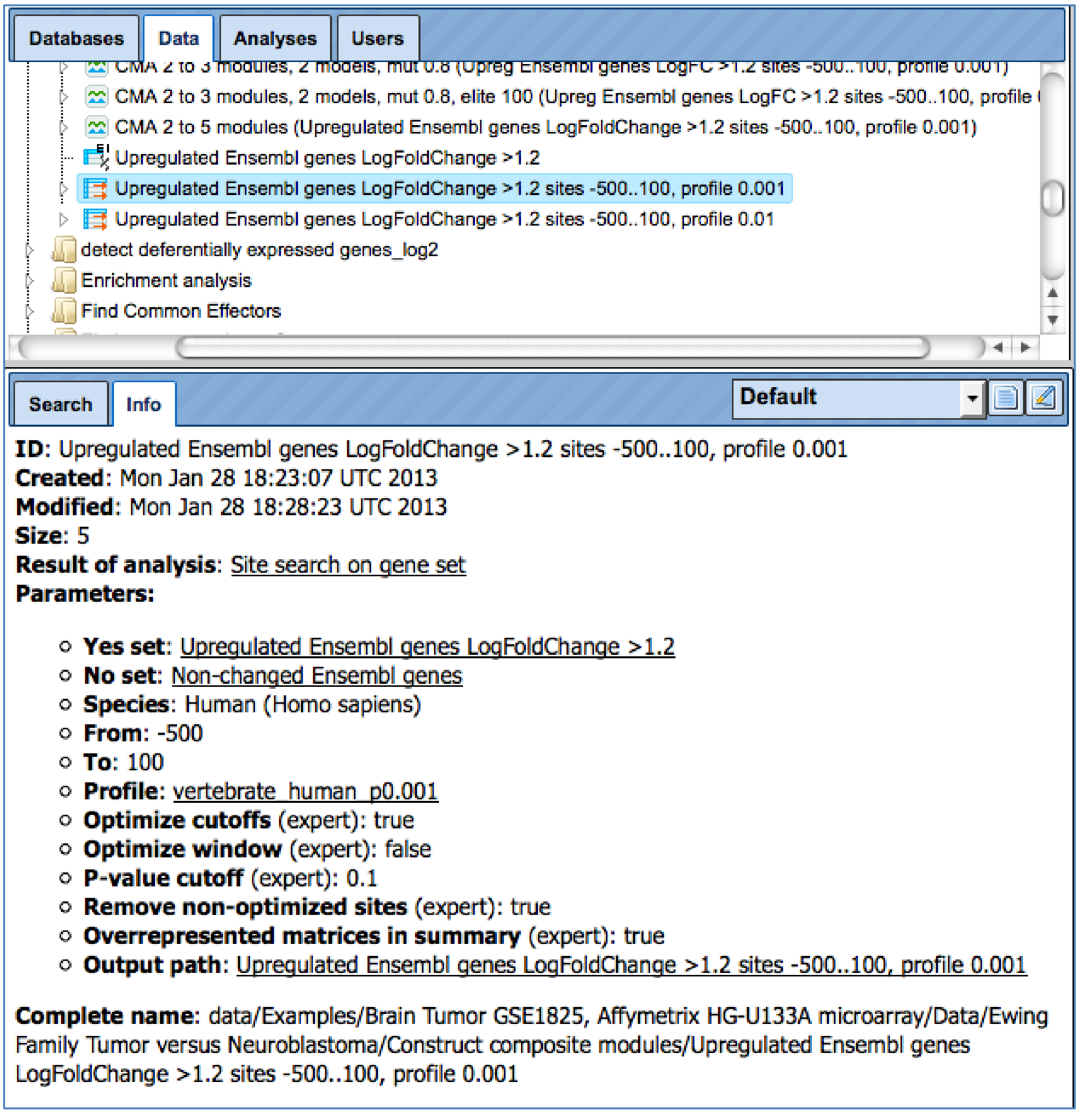
Site search result	(select element)
Model parameters	
Min modules	1
Max modules	1
Output path	(select element)

Step 1. Specify **Site search result.** This is the input field for the analysis, where you specify the results of the site search. The input field is marked by the symbol , which means that the input data set should have the same symbol.

In the example below, we will use an input data set, which you can find in the geneXplain platform online under the following path:

data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/Construct composite modules/Upregulated Ensembl genes LogFoldChange >1.2 sites -500..100, profile 0.001

As you can see, this is a folder with the results of the **Site search on gene set.** If you single-click on this folder in the tree area it will be highlighted in blue, and in the Info box you can find a description of all the details about Yes-set, No-set, profile, promoter regions applied to get these results (cf. screenshot below).


Databases
Data
Analyses
Users
CMA 2 to 3 modules, 2 models, mut 0.8 (Upreg Ensembl genes LogFC >1.2 sites -500..100, profile 0.001)
CMA 2 to 3 modules, 2 models, mut 0.8, elite 100 (Upreg Ensembl genes LogFC >1.2 sites -500..100, profile
CMA 2 to 5 modules (Upregulated Ensembl genes LogFoldChange >1.2 sites -500..100, profile 0.001)
Upregulated Ensembl genes LogFoldChange >1.2
Upregulated Ensembl genes LogFoldChange >1.2 sites -500..100, profile 0.001
Upregulated Ensembl genes LogFoldChange >1.2 sites -500..100, profile 0.01
detect deferentially expressed genes_log2
Enrichment analysis
Find Common Effectors
Search
Info
Default

ID: Upregulated Ensembl genes LogFoldChange >1.2 sites -500..100, profile 0.001
Created: Mon Jan 28 18:23:07 UTC 2013
Modified: Mon Jan 28 18:28:23 UTC 2013
Size: 5
Result of analysis: Site search on gene set
Parameters:

- **Yes set**: Upregulated Ensembl genes LogFoldChange >1.2
- **No set**: Non-changed Ensembl genes
- **Species**: Human (Homo sapiens)
- **From**: -500
- **To**: 100
- **Profile**: vertebrate_human_p0.001
- **Optimize cutoffs** (expert): true
- **Optimize window** (expert): false
- **P-value cutoff** (expert): 0.1
- **Remove non-optimized sites** (expert): true
- **Overrepresented matrices in summary** (expert): true
- **Output path**: Upregulated Ensembl genes LogFoldChange >1.2 sites -500..100, profile 0.001

Complete name: data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/Construct composite modules/Upregulated Ensembl genes LogFoldChange >1.2 sites -500..100, profile 0.001

You can drag & drop the name of the folder *Upregulated Ensembl genes LogFoldChange >1.2 sites -500..100, profile 0.001* into the input field, or you can select it in the tree in the pop-up window if you click on the pink box of the input field.

Step 2. Specify **Model parameters**. You can specify the number of elements in the hierarchical structure of the desired composite module. Details and explanations on how to do this are explained below in Section 9.5.3 *Hierarchical structure of the composite modules*.

Step 3. Specify **Output path**. Specify a location for the results in your project in the tree area. The resulting folder will be marked by the same icon as the analysis, .

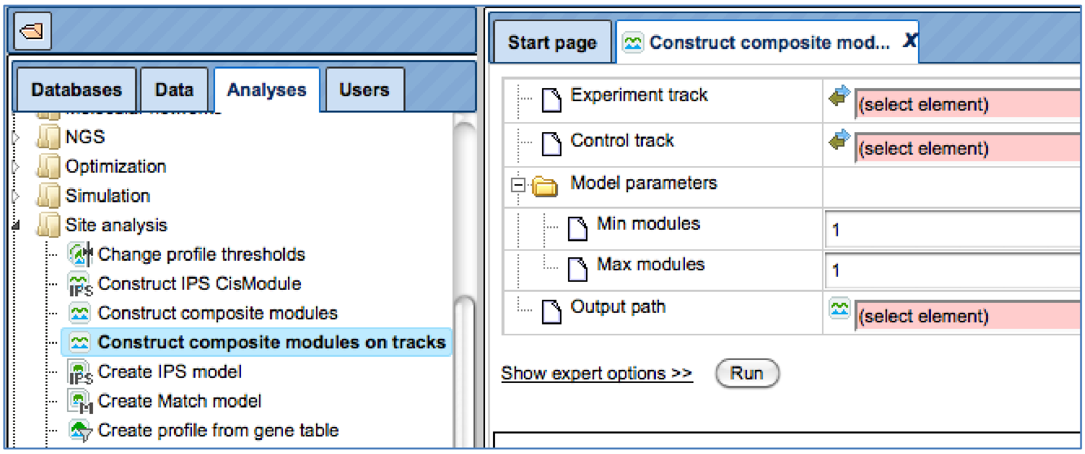
Results are described below in Section 20.1.5.4.

20.1.5.2. Construct composite modules on tracks

This analysis is designed for identifiying combinations of several TFBSs in DNA sequences specified by their genomic positions (tracks). An example of a track that is very often used is a set of the ChIP-seq data. The resulting composite module differentiates between a Yes-track and a background (No-track).

Before starting this analysis, you need to perform *Site search on track* with your selected Yes-set, No-set and specified profile of matrices. If you are interested in seeing the site models for particular TFs in the resulting composite modules, you have to make sure that such matrices are present in the selected profile. You can use one of the available TRANSFAC® profiles, or alternatively you can construct a customized profile; for details please see Section 20.1.3.

The input form for this analysis is shown below:



Experiment track	(select element)
Control track	(select element)
Model parameters	
Min modules	1
Max modules	1
Output path	(select element)

Step 1. Experiment track is the input field for the Yes-track, or track under study.

Step 2. Control track is the input field for the No-track, or background track.

Step 3. Model parameters. You can specify the number of elements in the hierarchical structure of the desired composite module. You may find details and explanations on how to do this below in Section 9.5.3 "Hierarchical structure of the composite modules".

Step 4. Output path. Specify a location for the results in your project in the tree area. The resulting folder will be marked by the same icon as the analysis: .

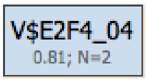
Results are described below in Section 20.1.5.4.

Note. This analysis is the central part the workflow ChIP-Seq - Identify composite modules on peaks (TRANSFAC®), and it might be more convenient to use the workflow instead of the individual analysis.

20.1.5.3. Hierarchical structure of the composite modules

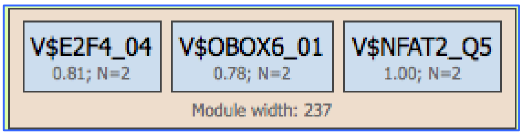
Composite modules may have a complex hierarchical structure consisting of two levels: site models and modules. The highest hierarchical level contains several modules and corresponds to the promoter model.

The first level, **site model**, corresponds to the individual site model, often based on one PWM. Names of the site models are often the same as the matrix names (in case the site models are based on a library of matrices). The site models are taken from the profile that was used in the site search. In the resulting schemas the site models are shown by blue boxes, for instance:



Within these boxes, there are two values below the site model. The first value is the threshold value for the score of the respective site model, which is determined by the genetic algorithm during the optimization process (here it is equal to 0.81); in some cases this value is equal to 0.0, which means that the original threshold value given in the profile was found by the algorithm to be the optimal one. The second value, in this example N=2, is the maximum number of best found individual matches (sites) for this site model which are taken into account for calculating the score of the module.

The next level, **module**, may contain several site models, shown within the light brown boxes:

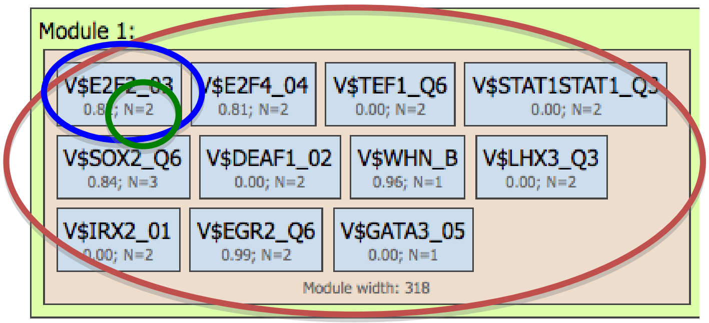


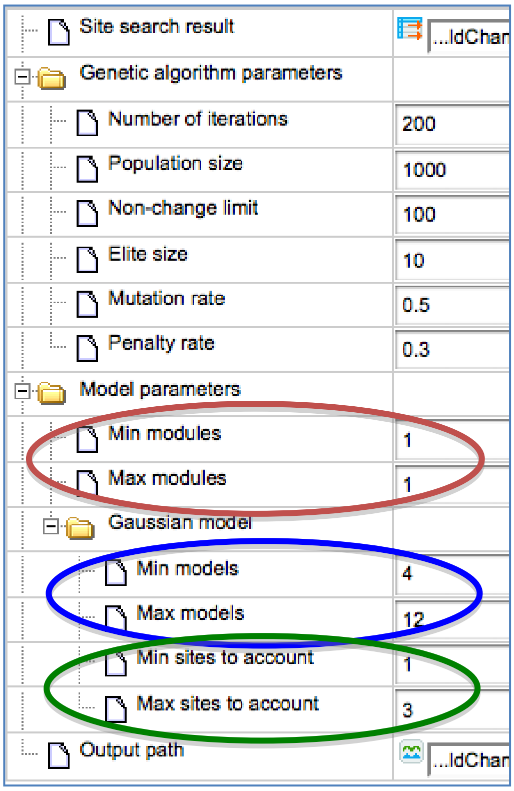
The module is characterized by its width, the average length of DNA window containing matches for the mentioned site models. In the example, the module width is 237 bp. In the resulting schemas modules are shown in green boxes, and they are numbered, e.g. Module 1, Module 2, ….

In the input form you can define the complexity of *the promoter model* to be constructed by specifying the number of units of each level: number of modules, number of site models, and also the minimum and maximum numbers of individual sites to be considered. In order to illustrate how to specify these parameters, let's consider three examples of resulting modules depending on the input parameters.

Example 1.

In the picture below (left part) you can see the composite module resulting from the performed analysis. On the right side of the picture the input form with specified parameters is shown.





Site search result	...ldChar
Genetic algorithm parameters	
Number of iterations	200
Population size	1000
Non-change limit	100
Elite size	10
Mutation rate	0.5
Penalty rate	0.3
Model parameters	
Min modules	1
Max modules	1
Gaussian model	
Min models	4
Max models	12
Min sites to account	1
Max sites to account	3
Output path	...ldChar

We can see that **Min modules** (minimum number of modules) and **Max modules** (maximum number of modules) is 1, and correspondingly there is just one module in the resulting picture, Module 1, highlighted by the red circle on both the resulting schema and the input parameters.

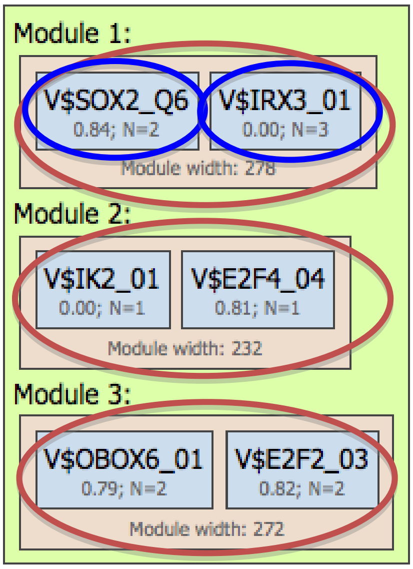
The blue circle highlights the parameters **Min models** (minimum number of site models within one module) and **Max models** (maximum number of site models within one module) in the input form. As we can see, the number of site models is set to vary from 4 to 12. Correspondingly, in the resulting schema (left part above) there are 11 site models (blue boxes) identified by the algorithm.

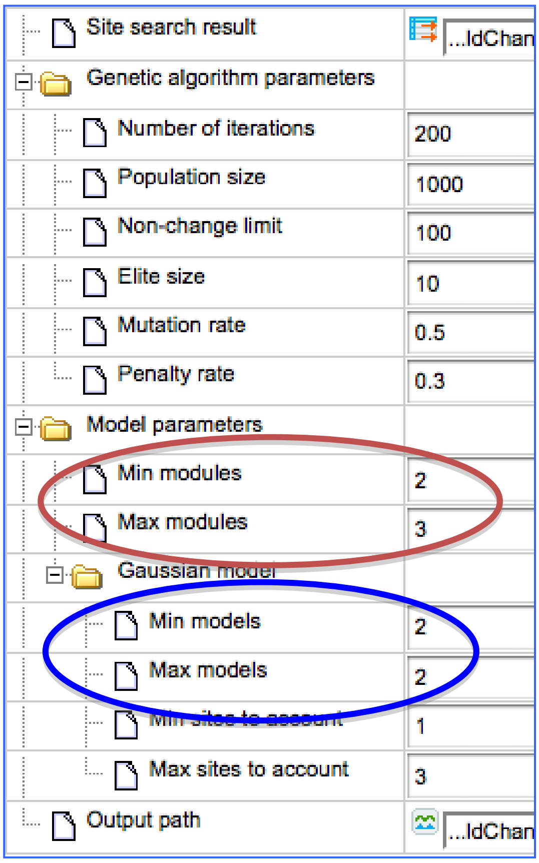
The third parameter, **Min sites to account** (minimum number of individual sites for each site model to be considered) and **Max sites to account** (maximum number of individual sites for each site model to be considered), is highlighted by a green circle. As we can see, this parameter is set to vary from 1 to 3, and correspondingly in the resulting schema, for the different matrices we can see N=1 or N=2 or N=3.

Example 2.

In this example, the number of modules (red circles) is specified from 2 to 3, and correspondingly the resulting promoter model contains three modules, Module 1, Module 2, and Module 3 (picture below, left part, red circles).

The number of site models is specified in the input form as from 2 to 2, which means that we are going to search for pairs of individual site models. In agreement with the input parameters, in the resulting schema we can see each module containing two site models highlighted by blue circles.



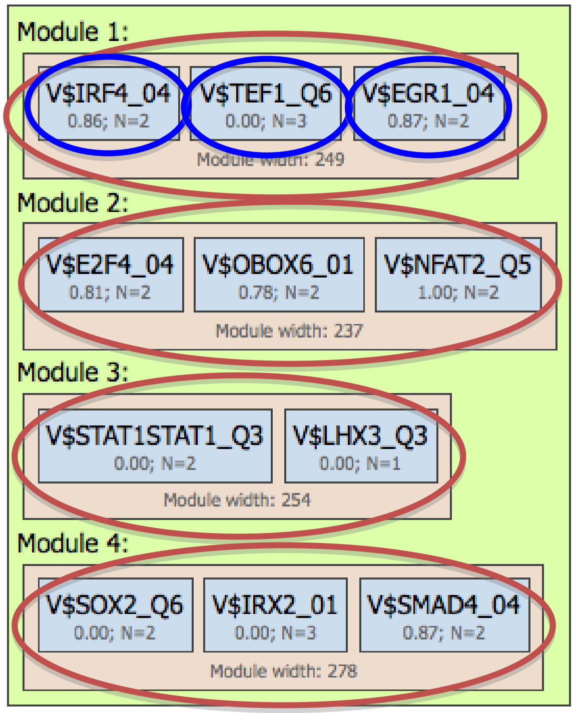


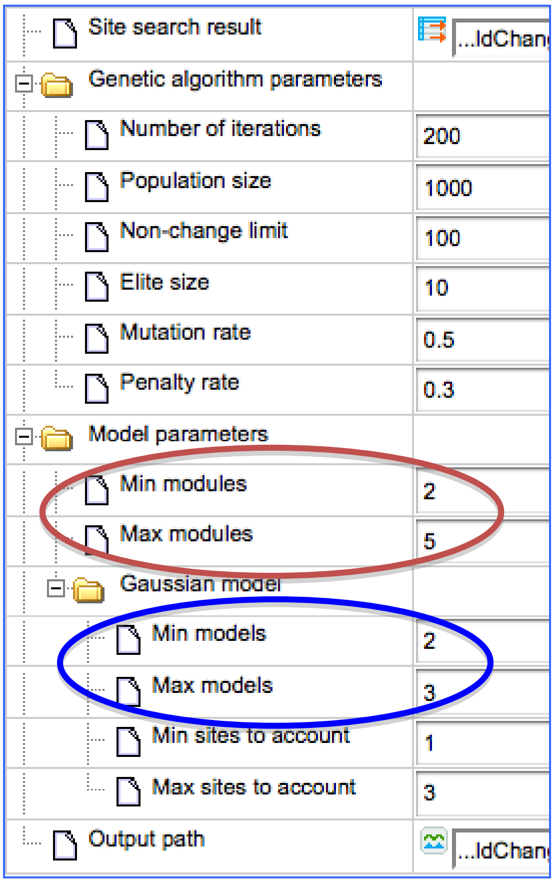
Site search result	...ldChan
Genetic algorithm parameters	
Number of iterations	200
Population size	1000
Non-change limit	100
Elite size	10
Mutation rate	0.5
Penalty rate	0.3
Model parameters	
Min modules	2
Max modules	3
Gaussian model	
Min models	2
Max models	2
Min sites to account	1
Max sites to account	3
Output path	...ldChan

Example 3.

In this example, the number of modules (red circles) is specified from 2 to 5, and correspondingly the resulting promoter model consists of four modules selected by the algorithm, Module 1, Module 2, Module 3 and Module 4 (picture below, left part, red circles).

The number of site models is specified in the input form as from 2 to 3, which means that we would like to find either pairs or triplets of individual site models. In agreement with the input parameters, three out of four modules contain three site models (highlighted by blue circles within Module 1), and one module, Module 3, contains two site models.



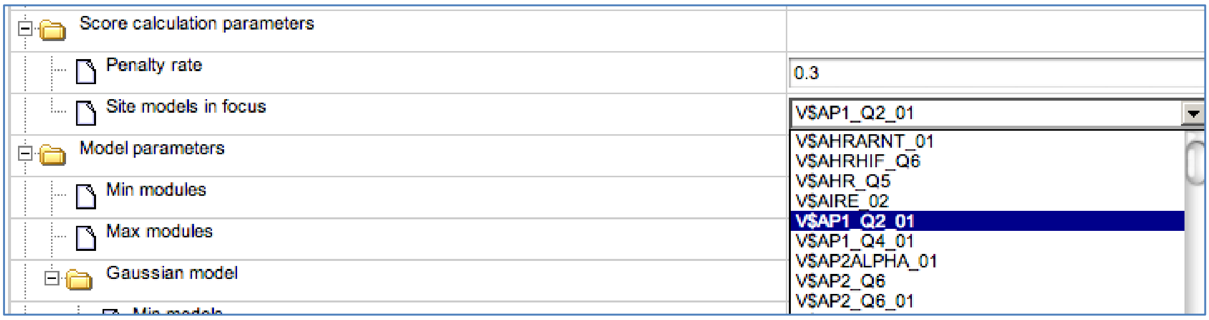


Site search result	...ldChan
Genetic algorithm parameters	
Number of iterations	200
Population size	1000
Non-change limit	100
Elite size	10
Mutation rate	0.5
Penalty rate	0.3
Model parameters	
Min modules	2
Max modules	5
Gaussian model	
Min models	2
Max models	3
Min sites to account	1
Max sites to account	3
Output path	...ldChan

Based on these three examples you can specify input parameters for the number of modules, site models, and individual sites, depending on what resulting promoter model you would like to get.

Site models in focus

There are situations when researchers would like to focus on particular TFs and would like to find out with what other TFs they may form composite modules. Site models that must be present in the resulting modules, are referred to as site models in focus. In the *expert options* menu, under *Score calculation parameters*, there is a field **Site models in focus**. As soon as the site search results are submitted to the input field, all site models from the profile used before for site search are now available for selection via the drop-down menu, as shown on the screenshot below.



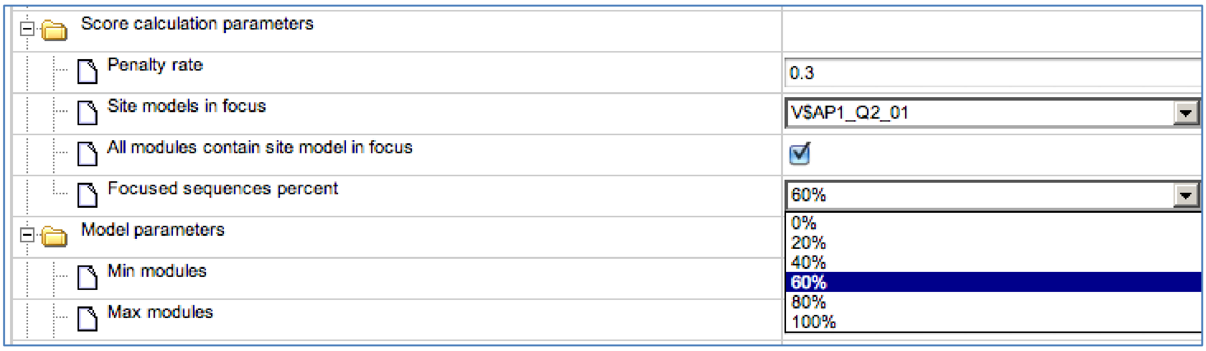
Score calculation parameters
Penalty rate: 0.3
Site models in focus: V$AP1_Q2_01
Model parameters
Min modules
Max modules
Gaussian model

In this example, we can see that V$AP1_Q2_01 is selected. Using the Ctrl button, several matrices from the list can be selected.

As soon as the site model is selected, two new fields will appear just below the field Site models in focus (screenshot below).

All modules contain site model in focus. When this box is checked, all the resulting modules must contain specified site models. When it is unchecked, at least one of the resulting modules must have a specified site model, and other may have them too, but not necessarily.

Focused sequences percent. In the drop-down menu, you can specify the minimum percentage of the Yes sequences (genes, promoters) that must contain modules with the specified site models in focus.



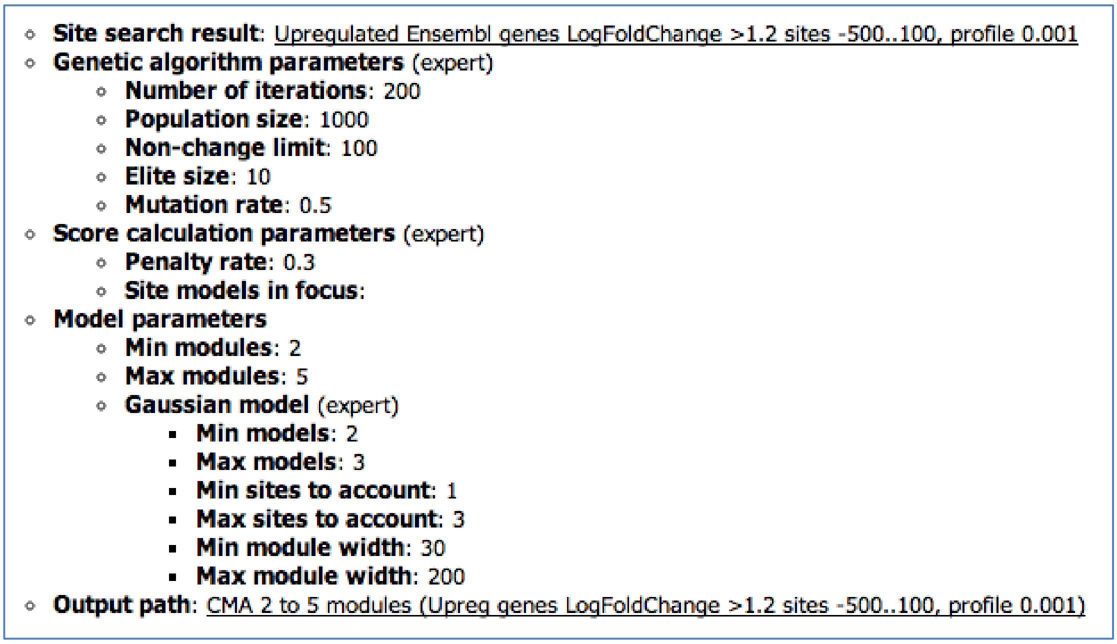
Score calculation parameters
Penalty rate: 0.3
Site models in focus: V$AP1_Q2_01
All modules contain site model in focus: ☑
Focused sequences percent: 60%
Model parameters
Min modules
Max modules

20.1.5.4. Visualization and interpretation of the results

Let us consider the results of the *Construct composite module* analysis obtained for the following input data set, which you can find in the geneXplain platform online under the following path:

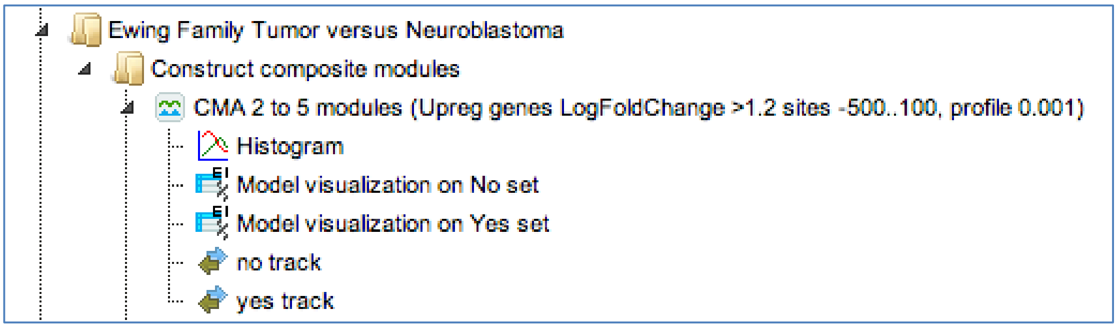
data/Examples/Brain Tumor GSE1825, Affymetrix HG-U133A microarray/Data/Ewing Family Tumor versus Neuroblastoma/Construct composite modules/Upregulated Ensembl genes LogFoldChange >1.2 sites -500..100, profile 0.001

Input parameters used were the following:



- **Site search result**: Upregulated Ensembl genes LogFoldChange >1.2 sites -500..100, profile 0.001
- **Genetic algorithm parameters** (expert)
  - **Number of iterations**: 200
  - **Population size**: 1000
  - **Non-change limit**: 100
  - **Elite size**: 10
  - **Mutation rate**: 0.5
- **Score calculation parameters** (expert)
  - **Penalty rate**: 0.3
  - **Site models in focus**:
- **Model parameters**
  - **Min modules**: 2
  - **Max modules**: 5
  - **Gaussian model** (expert)
    - **Min models**: 2
    - **Max models**: 3
    - **Min sites to account**: 1
    - **Max sites to account**: 3
    - **Min module width**: 30
    - **Max module width**: 200
- **Output path**: CMA 2 to 5 modules (Upreg genes LogFoldChange >1.2 sites -500..100, profile 0.001)

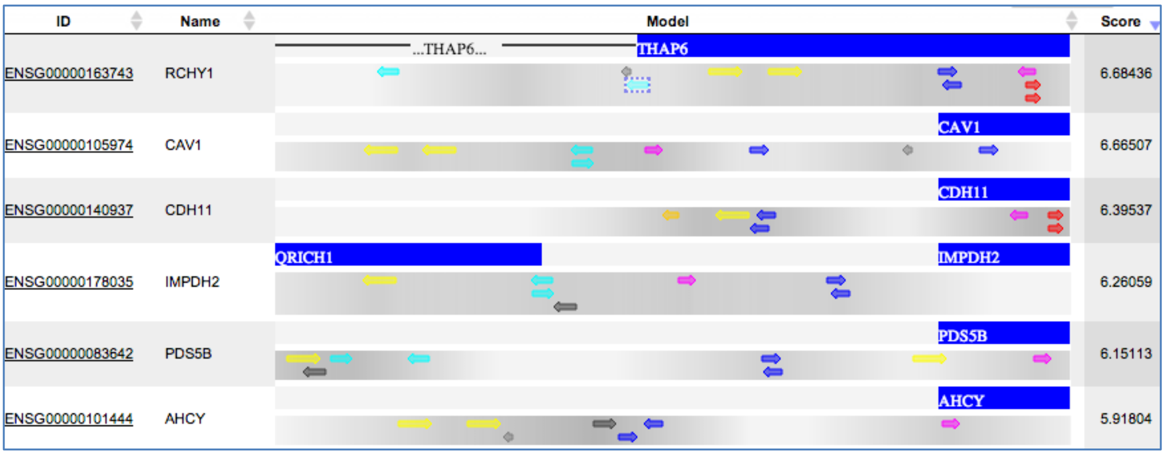
As result, a new folder is generated containing two tables, two tracks, and one histogram, as shown below.



- Ewing Family Tumor versus Neuroblastoma
  - Construct composite modules
    - CMA 2 to 5 modules (Upreg genes LogFoldChange >1.2 sites -500..100, profile 0.001)
      - Histogram
      - Model visualization on No set
      - Model visualization on Yes set
      - no track
      - yes track

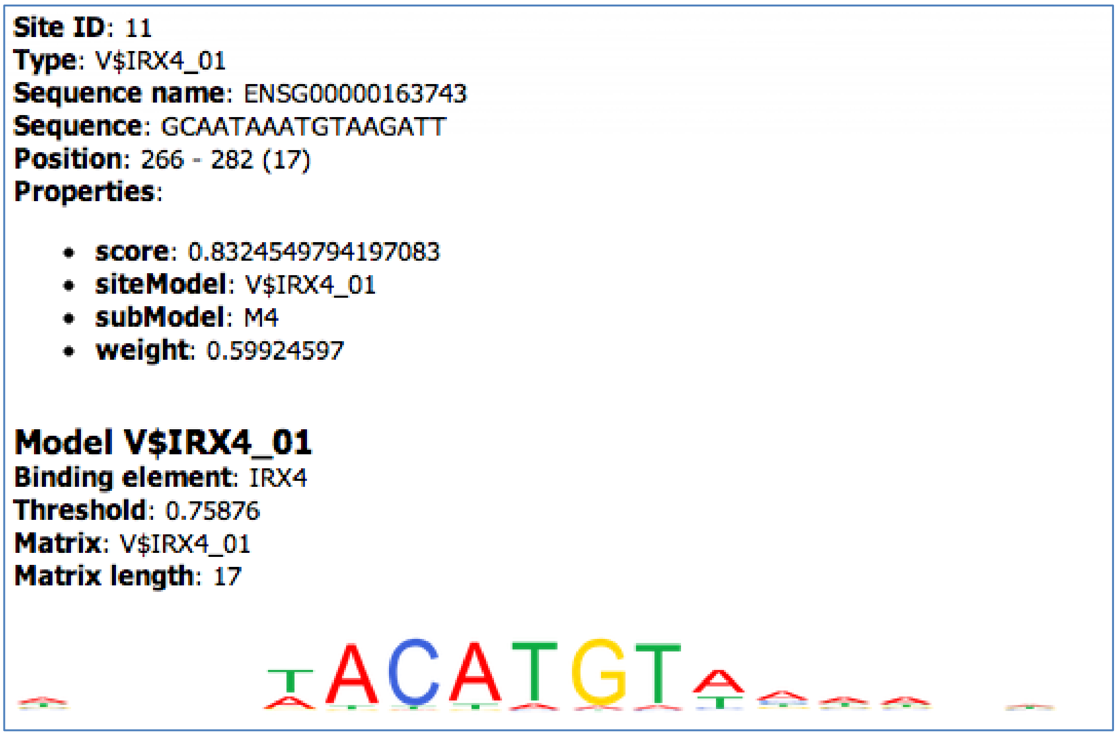
Model visualization in Yes set

This table represents the primary results of the analysis, and shows the visualization of the identified composite modules in the promoters of the Yes set.



ID	Name	Model	Score
ENSG00000163743	RCHY1	...THAP6... THAP6	6.68436
ENSG00000105974	CAV1	CAV1	6.66507
ENSG00000140937	CDH11	CDH11	6.39537
ENSG00000178035	IMPDH2	QRICH1 IMPDH2	6.26059
ENSG00000083642	PDS5B	PDS5B	6.15113
ENSG00000101444	AHCY	AHCY	5.91804

Each row in this table corresponds to one gene of the Yes set, and for each gene the Ensembl ID and the gene symbol are shown in the two first columns. The column **Model** displays a symbolic map of the gene promoter taken for the analysis, in this case -500/+100 relative to the TSS. Arrows of different colors correspond to individual TFBSs, and a gradient in grey corresponds to the statistical density of the identified composite modules. The most intensive grey color corresponds to the center of a composite module. Each individual TFBS on this map is clickable, and upon a click information is displayed in the Info box (bottom left corner in the tool). As an example, one blue arrow is selected on the promoter of the top gene in the screenshot above, and for this selected TFBS the following details are shown in Info box:



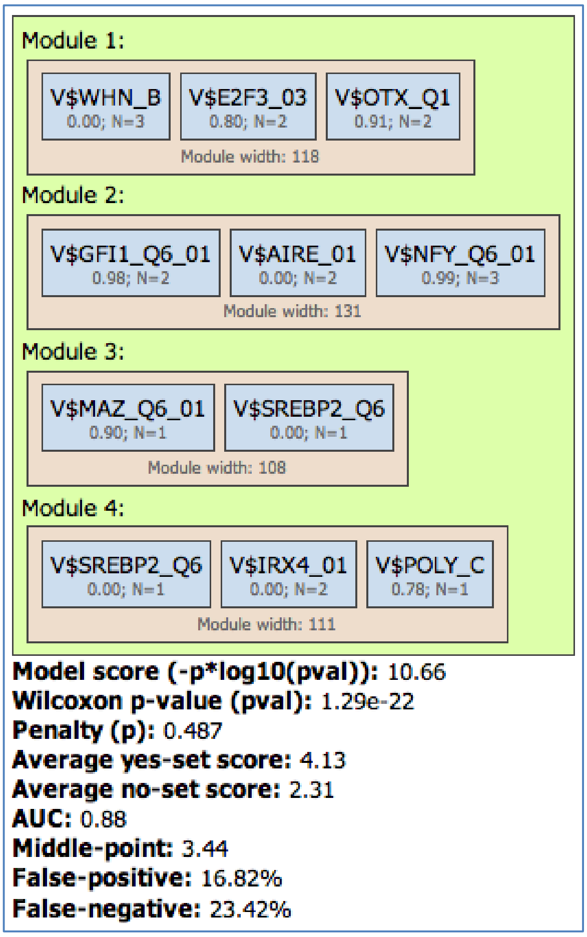
Site ID: 11
Type: V$IRX4_01
Sequence name: ENSG00000163743
Sequence: GCAATAAATGTAAGATT
Position: 266 - 282 (17)
Properties:

- **score**: 0.8324549794197083
- **siteModel**: V$IRX4_01
- **subModel**: M4
- **weight**: 0.59924597

Model V$IRX4_01
Binding element: IRX4
Threshold: 0.75876
Matrix: V$IRX4_01
Matrix length: 17

The last column in the table, **Score**, shows a score calculated for each promoter depending on the number of modules, site models, sites, their scores and other statistical parameters. The higher the score for a promoter, the better the differentiation of this promoter from the promoters of the No set. The column **Score** is used for default sorting of the table, with the highest scores on top.

Having opened the table **Model visualization on Yes set**, you can see the schematic representation of the hierarchical structure of the identified composite module as well as a comprehensive set of its statistical characteristics at the bottom part of the tool, under ***My description*** tab, as shown on the screenshot below.



Module 1:

V$WHN_B 0.00; N=3 | V$E2F3_03 0.80; N=2 | V$OTX_Q1 0.91; N=2

Module width: 118

Module 2:

V$GFI1_Q6_01 0.98; N=2 | V$AIRE_01 0.00; N=2 | V$NFY_Q6_01 0.99; N=3

Module width: 131

Module 3:

V$MAZ_Q6_01 0.90; N=1 | V$SREBP2_Q6 0.00; N=1

Module width: 108

Module 4:

V$SREBP2_Q6 0.00; N=1 | V$IRX4_01 0.00; N=2 | V$POLY_C 0.78; N=1

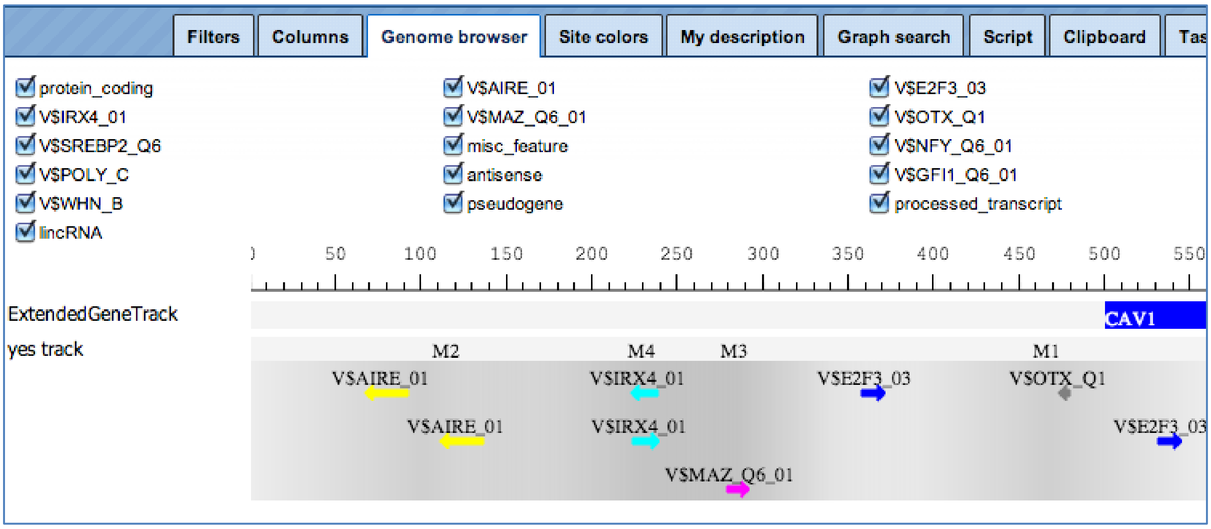
Module width: 111

Model score (-p*log10(pval)): 10.66
Wilcoxon p-value (pval): 1.29e-22
Penalty (p): 0.487
Average yes-set score: 4.13
Average no-set score: 2.31
AUC: 0.88
Middle-point: 3.44
False-positive: 16.82%
False-negative: 23.42%

Yes track

The Yes track provides essential information about the regulation of individual promoters and is therefore important to be included in the visualization of individual promoters by the genome browser.

The schematic visualization can be comfortably extended to a more detailed visualization for each individual promoter.

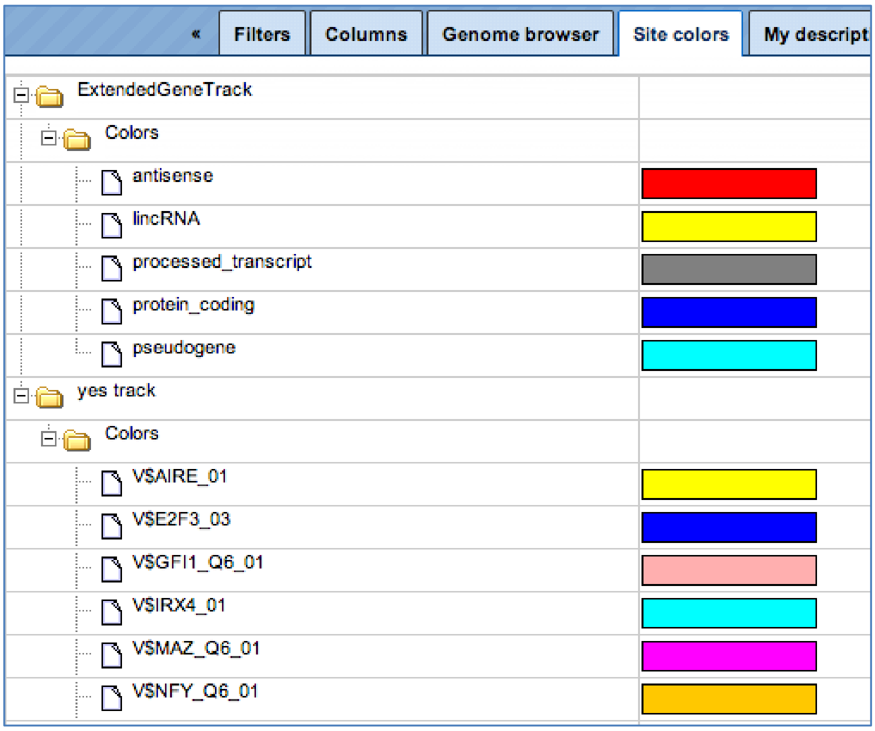


☑ protein_coding
☑ V$IRX4_01
☑ V$SREBP2_Q6
☑ V$POLY_C
☑ V$WHN_B
☑ lincRNA
☑ V$AIRE_01
☑ V$MAZ_Q6_01
☑ misc_feature
☑ antisense
☑ pseudogene
☑ V$E2F3_03
☑ V$OTX_Q1
☑ V$NFY_Q6_01
☑ V$GFI1_Q6_01
☑ processed_transcript

To do this, open the table *Model visualization on Yes set*, select one row in the table with a mouse click, and open the tab *Genome browser* in the *Operation Field* (bottom right part of the tool),.

For a selected promoter, you can see a more detailed map, including the names of the matrices and the numbers of individual modules, M1 through M4. Each element of this interactive map has a corresponding check box. Unchecked elements will not be displayed on the map. De-selection is applied simultaneously to both the detailed view of one promoter, and the table with the schematic representation of all promoters.

To adjust colors for individual matrices, you can open the next tab to the right, called *Site colors*, and change the colors to your liking by clicking on the individual colored box, as shown below.



ExtendedGeneTrack	
Colors	
antisense	🟥
lincRNA	🟨
processed_transcript	⬛
protein_coding	🟦
pseudogene	🟦
yes track	
Colors	
V$AIRE_01	🟨
V$E2F3_03	🟦
V$GFI1_Q6_01	🟥
V$IRX4_01	🟦
V$MAZ_Q6_01	🟪
V$NFY_Q6_01	🟧

Model visualization on No set and No track

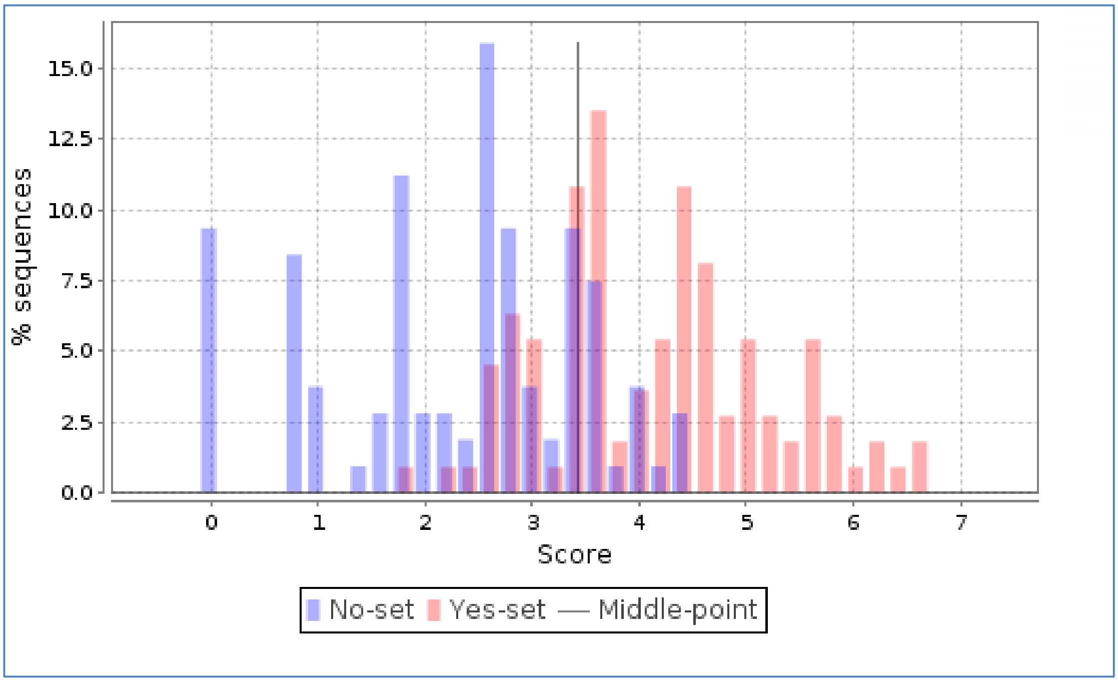
The table *Model visualization on No set* shows a visualization of the identified composite modules in the promoters of the No set.

The structure of this table is the same as that of the *Model visualization on Yes set* table, described above.

The function of the **No track** is to provide a possibility for a detailed visualization of no promoters in a way similar to that of the **Yes track**.

Histogram

The distribution of scores for individual promoters is shown as a histogram, where the promoter score value is shown on X axis and the percentage of promoters (% sequences) having this score is shown on the Y axis.



This histogram can be further interpreted applying the statistical characteristics described above.

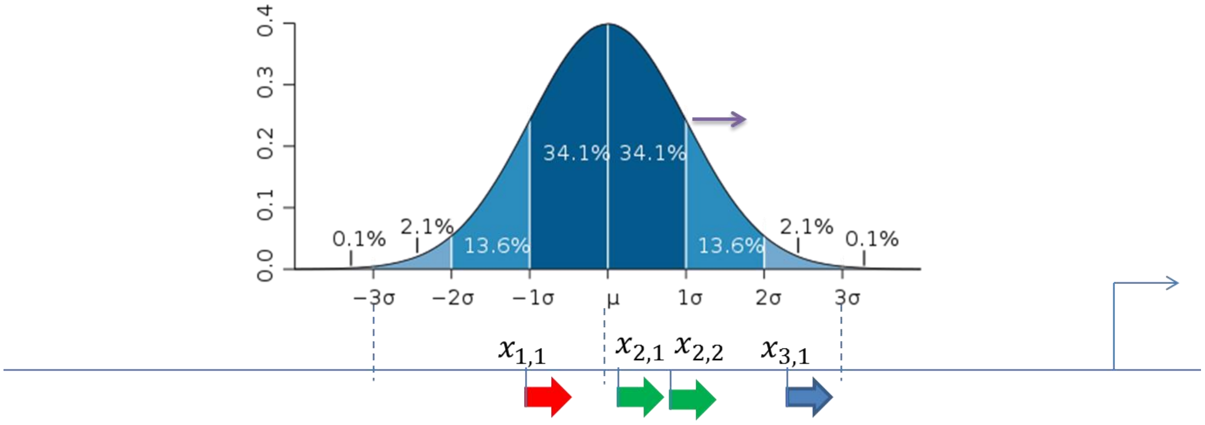
The center, a vertical grey line, corresponds to the average score value and is equal to 3.44 in this example. Promoters from the No set with a score above 3.44 are shown in the histogram as blue bars to the right of the center, and they are referred to as false positives. In this example, the false positive rate is 16.82 %.

Promoters from the Yes set with a score below 3.44 are shown in the histogram as red bars to the left of the center, and they are referred to as false negatives. In this example, the false negative rate is 23.42 %.

A visual analysis of the histogram suggests that the Yes promoters with a score above 4.5 are very well separated from the No promoters, which means that for this part of the promoters the composite model constructed is most suitable. In this example there are 38 promoters with the score value >4.5; they can be saved as a separate gene set, and for them the model obtained works best.

20.1.5.5. Score calculation of the composite models

The figure below demonstrates the calculation of the score value for the composite modules in the promoter sequences. The TSS is shown as a thin arrow on the right side of the figure. Four thick arrows exemplify four sites found in this promoter. The color of the arrows exemplifies the site model which these sites belong to (three site models – red, green and blue).



A promoter model consists of *K* modules. The score of each module *M_k* (*Score(M_k)*, *k* = 1, …, *K*) is calculated according to this formula:

$$Score\ (M_k) = \max_{\mu=1,L} \sum_{t=1}^{T_k} \sum_{i=1}^{m_t} SiteScore(t,i) \times f(x_{t,i},\mu,\sigma^2)$$

Here,

Site Score (t,i) is the site score for the sites found in the promoter, which is calculated by the Match algorithm.

m_t – the number of sites of the site model *t* found in the promoter.

T_k – the number of site models in the module M_k, and

$$f(x;\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The final promoter score is calculated as the sum of the module scores M_k.

Standard deviation (σ) of the normal distribution is subject of optimization by the genetic algorithm and represents the width of the module in the output of the composite module analysis.

20.2. About the GSEA analysis and the interpretation of the results

Schematic description of the GSEA algorithm

The Gene Set Enrichment Analysis (GSEA) is a method that determines whether an independently defined set of genes shows a statistically significant enrichment either in up-regulated or down-regulated genes http://www.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html

Independently defined sets of genes might be groups of genes linked to Gene Ontology terms, or to TRANSPATH® pathways, or to Reactome pathways, etc.

As input for the GSEA, we can use either normalized tables after normalization of the microarray raw data, or a table with genes that contains a column with pre-calculated expression values. The algorithm of the GSEA can be schematically presented in three steps:

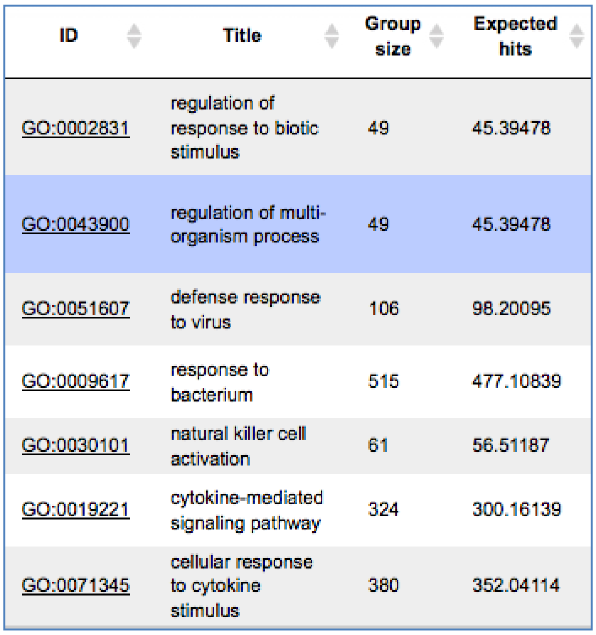
All genes are sorted by Fold Change, from highly up-regulated to highly down-regulated. In such a sorted table, each gene is given a rank, which is just the number of the line where this gene is located in the sorted table. The most highly up-regulated gene has the rank "1", and so on.

The algorithm takes a set of genes linked to one biological term (e.g. to a particular GO category) and maps each gene in this set to the sorted (ranked) table. Genes in some categories might be distributed randomly over the ranked table; some are among the up-regulated, some among the down-regulated, and some among the non-changed genes. For some of the GO categories, a majority of the genes are among the up-regulated. These GO categories are the most interesting for a biological interpretation.

Several statistical parameters, ES (Enrichment Score), NES (Normalized Enrichment Score), Rank at max, nominal p-value, FDR, for each GO category are calculated. ES and NES reflect how significantly genes in each particular GO category are enriched (over-represented) among up-regulated genes.

Structure of the resulting tables with enriched ontological terms

The resulting tables with enriched ontological categories contain twelve columns. Here, to have a better resolution of the screenshots, one table is shown in two parts. This is the result of a GSEA using the PROTEOME™ biological process.

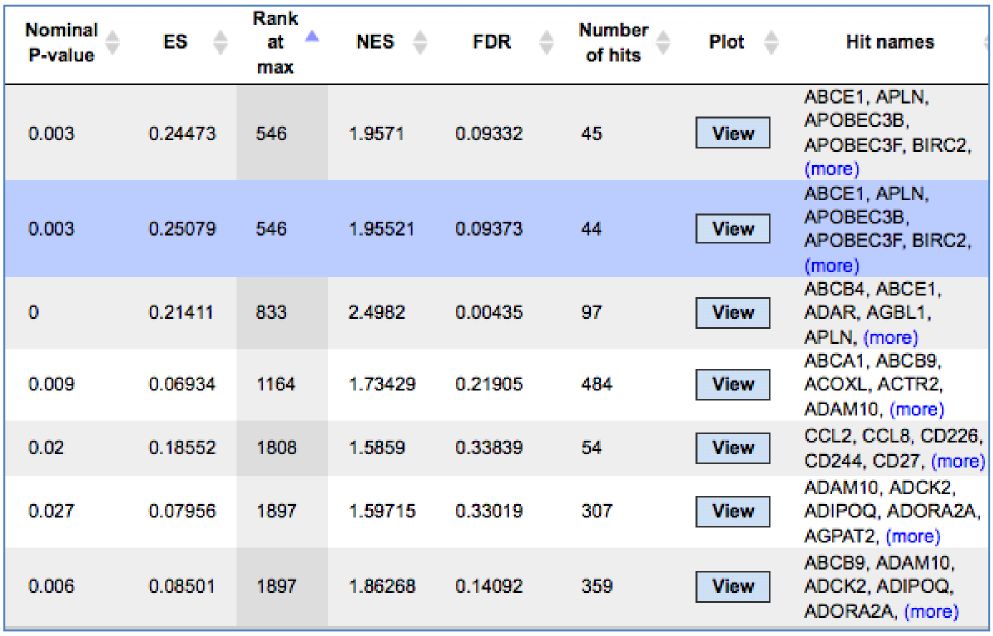


ID	Title	Group size	Expected hits
GO:0002831	regulation of response to biotic stimulus	49	45.39478
GO:0043900	regulation of multi-organism process	49	45.39478
GO:0051607	defense response to virus	106	98.20095
GO:0009617	response to bacterium	515	477.10839
GO:0030101	natural killer cell activation	61	56.51187
GO:0019221	cytokine-mediated signaling pathway	324	300.16139
GO:0071345	cellular response to cytokine stimulus	380	352.04114

Each row presents details about one enriched ontological term. The column **ID** comprises the identifier of the ontological category, here identifiers of Gene Ontology biological process terms. These identifiers are hyperlinked to the page http://www.ebi.ac.uk/QuickGO/ where you can get further information about this ontological term. In case of enrichment by PROTEOME™ diseases the disease identifiers are hyperlinked to http://ctdbase.org/

The columns **Title** and **Group size** contain further details about the ontological terms, the title and the number of genes linked to this term in the corresponding database, here in PROTEOME™.

The column **Expected hits** shows the number of genes expected to fall into this category by random chance, based on the size of the input set and the size of the category.



Nominal P-value	ES	Rank at max	NES	FDR	Number of hits	Plot	Hit names
0.003	0.24473	546	1.9571	0.09332	45	View	ABCE1, APLN, APOBEC3B, APOBEC3F, BIRC2, (more)
0.003	0.25079	546	1.95521	0.09373	44	View	ABCE1, APLN, APOBEC3B, APOBEC3F, BIRC2, (more)
0	0.21411	833	2.4982	0.00435	97	View	ABCB4, ABCE1, ADAR, AGBL1, APLN, (more)
0.009	0.06934	1164	1.73429	0.21905	484	View	ABCA1, ABCB9, ACOXL, ACTR2, ADAM10, (more)
0.02	0.18552	1808	1.5859	0.33839	54	View	CCL2, CCL8, CD226, CD244, CD27, (more)
0.027	0.07956	1897	1.59715	0.33019	307	View	ADAM10, ADCK2, ADIPOQ, ADORA2A, AGPAT2, (more)
0.006	0.08501	1897	1.86268	0.14092	359	View	ABCB9, ADAM10, ADCK2, ADIPOQ, ADORA2A, (more)

Five columns, **Nominal P-value**, **ES** (Enrichment Score), **Rank at max**, **NES** (Normalized Enrichment Score), and **FDR**, show corresponding statistics of the results. For more details about each of these statistics, please refer to http://www.broadinstitute.org/gsea/doc/GSEAUserGuideFrame.html, under the section *Interpreting GSEA results/GSEA statistics*.

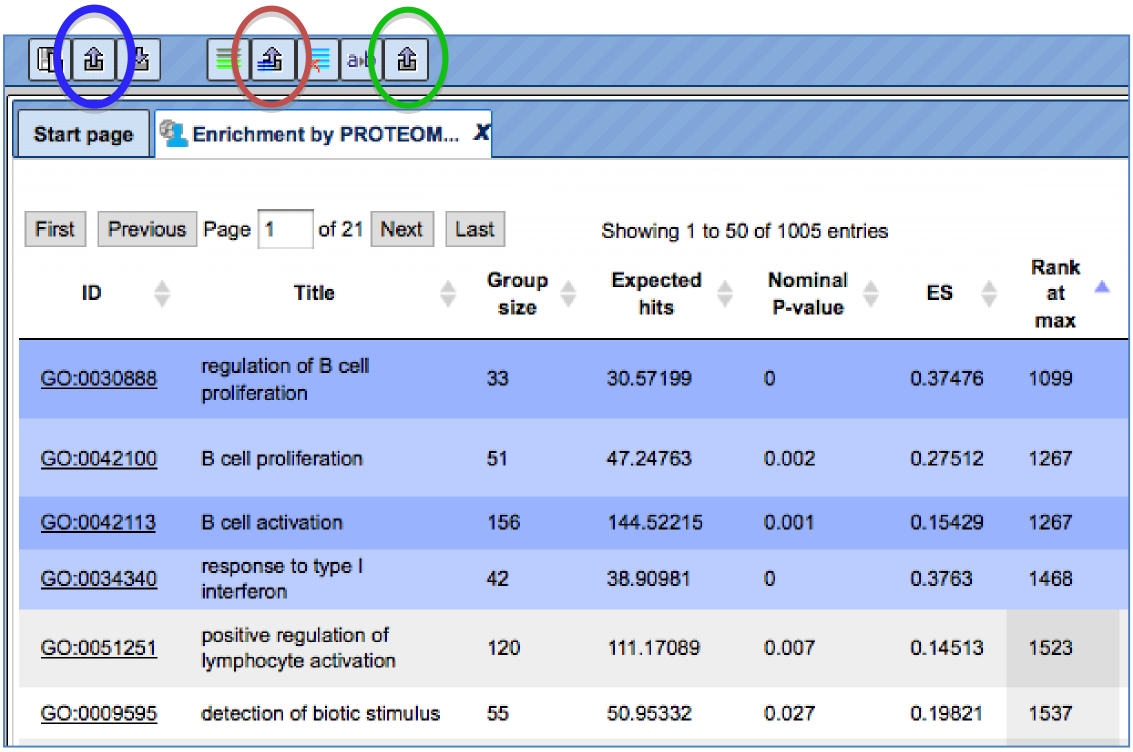
The column **Number of hits** shows how many genes from the input set fall into the enrichment group; these genes are explicitly listed in the column **Hit names**. As the lists can get quite long, only a few genes are shown by default in each row. To get the full list, press **(more)**. The column **Plot** and **View** diagrams are described in detail in a separate sub-section below.

Further possible actions with the enrichment results.

The tables with the GSEA results can be exported in txt, csv or html formats. To export, you can apply the *Export* button in the top control menu, highlighted by the dark-blue oval.

One or several rows of the table can be selected with shift-click , shown in blue color in the screenshot below. Having selected one or several rows in this way, you can save hits of these rows in a separate gene table, with the button *Save hits* in the top control menu, highlighted by the red oval. Such genes tables can be analyzed further, e.g. to find master regulatory molecules in the networks, and to identify transcription factors that might commonly regulate these genes.

You can save the selected rows in a separate table, with the button *Save selected rows as…* in the top control menu, highlighted by the green oval.



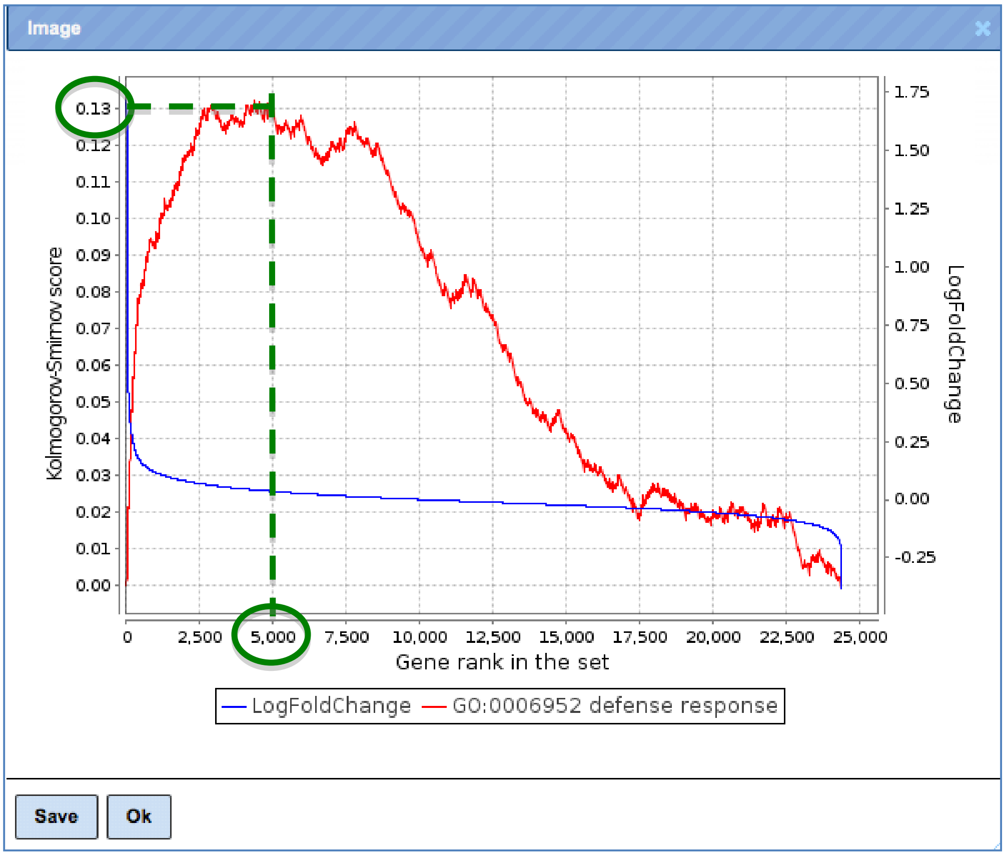
Showing 1 to 50 of 1005 entries

ID	Title	Group size	Expected hits	Nominal P-value	ES	Rank at max
GO:0030888	regulation of B cell proliferation	33	30.57199	0	0.37476	1099
GO:0042100	B cell proliferation	51	47.24763	0.002	0.27512	1267
GO:0042113	B cell activation	156	144.52215	0.001	0.15429	1267
GO:0034340	response to type I interferon	42	38.90981	0	0.3763	1468
GO:0051251	positive regulation of lymphocyte activation	120	111.17089	0.007	0.14513	1523
GO:0009595	detection of biotic stimulus	55	50.95332	0.027	0.19821	1537

Enrichment plots

As described above, there is a column **Plot** with the buttons **View** in each row in the GSEA resulting tables.

Let's open a visualization plot for the category defense response.



X-axis

On this line, all (human in this example) genes present in the input table are ranked from 1 to approximately 23,000 according to their LogFoldChange. The numbers we see on the X-axis, 2,500; 5,000 etc., are the ranks of the genes.

At the left end of the line X, up-regulated genes are located, and down-regulated genes at the right end . The smaller the rank of a gene is, the higher is its up-regulation.

Y-axis, right

This axis shows the values of the LogFoldChange. Positive values correspond to up-regulated genes, negative values to down-regulated genes.

Blue plot

This is a distribution of LogFoldChange values among all (human) genes.

Y-axis, left

This axis shows the values of the ES parameter (also known in statistics as Kolmogorov-Smirnov score). Positive values correspond to up-regulated genes, negative values to down-regulated genes.

Red plot

This is a distribution of ES values for a particular GO category over all (human) genes. ES is calculated by summing up on the previous values, so the most interesting parts for us are those where the plot is coming up.

Interpretation of the plots

For interpretation, let’s look at the red plot. The most interesting parts are those where the red line is ascending. In this example plot, let’s follow the red line from its left end to its maximum.

The starting point for the red line is 0 on the Kolmogorov-Smirnov score axis. Next, the algorithm takes each gene within the category "defense response", and looks what the LogFoldChange for this gene is. If gene is up-regulated, the red line grows.Then the algorithm takes the next gene from "GDP binding" and again looks what the LogFoldChange for the 2d gene is. If it is also up-regulated, the red line grows further. This is called incremental growth: the closer the next gene is to the previous one on the X-axis, the more pronounced is the growth of the red line.

Let’s add an auxiliary line (here, the green dotted line) from the maximum point of the red plot down to the X-axis. The intersection of the green dotted line with X-axis corresponds to the rank of the gene at maximum enrichment. This value is exactly what in statistics is called **Rank at max**, and it is shown in a dedicated column in the tables with the GSEA results. In the plot shown above, the green line crosses the X-axis at approximately 4,500. To know this value exactly, look at the table, and in the column **Rank at max** for this category, you can find the number 4328.

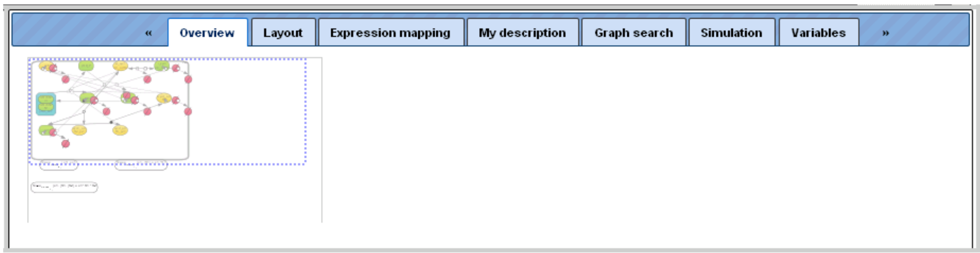
To describe this plot, we can say:

“Genes belonging to the GO category *defense response* are enriched among the top 4328 up-regulated genes”.

21. Operations field

In the **Operations Field** (**D** in the figure of Section 2.1) a number of essential functions to operate the geneXplain platform are provided on a number of tabs. How many and which tabs are shown depends very much on the context.

Please note that not all tabs are always visible due to space constraints. In these cases, double arrowheads left and right of the tabs indicate that there are additional ones, reachable by clicking on these double arrowheads.

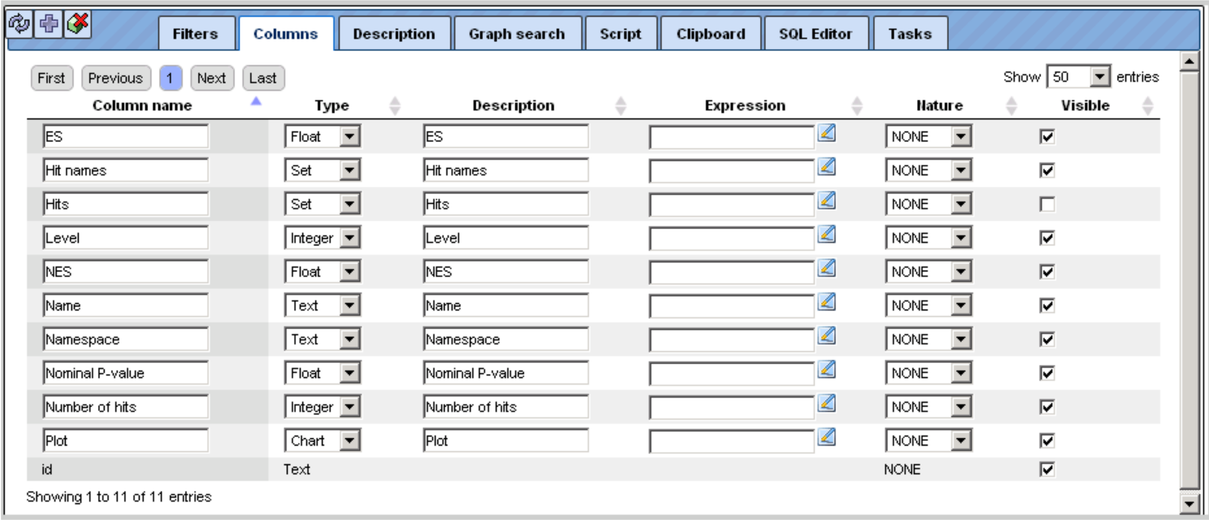


The function of the individual tabs will be explained in more detail in those sections where their effect is part of a certain operation. In general, the icon ▶ initiates the corresponding activity within the Operations Field, whereas ▲ applies to the results generated in the Operations Field of the Work Space.

21.1. Changing the table structure in the Operations Field

Having opened a table in the Work Space, e.g. by double clicking on its name in the Tree Area, it is possible to edit its structure in the Operations Field under the tab *Columns*.

For instance, if you have opened a table with data about Enrichment GO Molecular Mechanism (resulting from having run a GSEA), this field may look like this:



Column name	Type	Description	Expression	Nature	Visible
ES	Float	ES		NONE	☑
Hit names	Set	Hit names		NONE	☑
Hits	Set	Hits		NONE	□
Level	Integer	Level		NONE	☑
NES	Float	NES		NONE	☑
Name	Text	Name		NONE	☑
Namespace	Text	Namespace		NONE	☑
Nominal P-value	Float	Nominal P-value		NONE	☑
Number of hits	Integer	Number of hits		NONE	☑
Plot	Chart	Plot		NONE	☑
id	Text			NONE	☑

Showing 1 to 11 of 11 entries

Recognizably, you can change the column headers, the data type in the column, or its (usually hidden) descriptions. You may add an Expression, which may be a mathematical formula, formulated in Java script; you find detailed explanations for this when you press the Edit key (✎) next to this field. In the last column, you can specify which columns are visible or shall be hidden (unmarking a column here does NOT delete it, it hides it from the currently displayed table).

If you hide a column by unmarking it, you have to refresh the Work Space by pressing the button ⟳ in the control panel right on top of the Operations Field. Here, you can also add new (✚) columns. Before removing a column with the button ✖, you have to mark it by clicking somewhere in the background of the line specifying this column; the selected item will be highlighted in blue. But be careful: Deleting it from the table will irrevocably erase the column including all its contents!

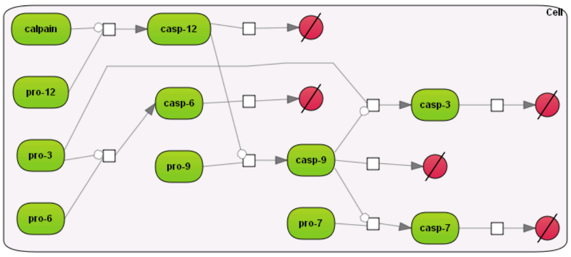
21.2. Changing the layout

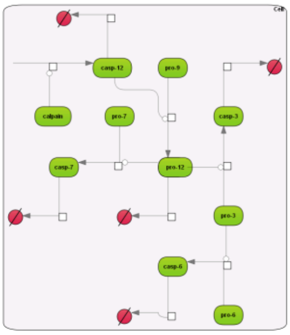
Under the Layout tab, you can change the layout of the diagram. First, you can select one of the following layout schemes:

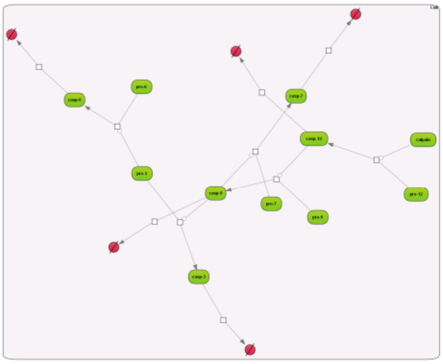
- *Hierarchical layout (default)*
- *Orthogonal layout*
- *Force directed layout*
- *Cross cost grid layout*
- *Grid layout*

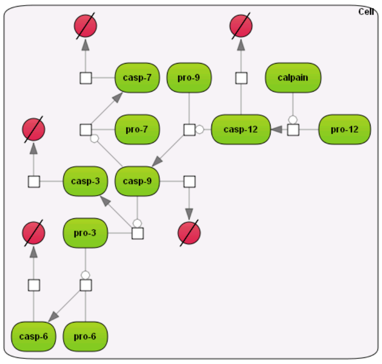
When you have selected another layout type, you have to press the "Prepare layout" button (▷), showing the new layout at the right of the same window in the Operation Field. Pressing "Apply layout" (△) transfer the new layout to the Work Space.

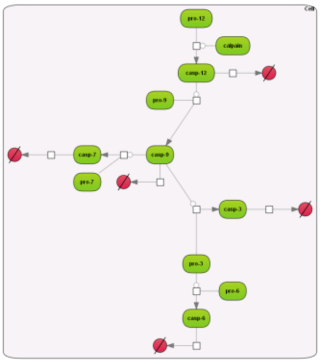
Here are examples how the different layouts look like; the example is the Caspase 12 pathway and has been taken from the database Integrated models (*Int_casp12_module12*):











Some of the re-layouting options may take considerable time. If you want to interrupt the process, press the "Stop layout" button (■).

The layout that has been applied to the Work Space can be further edited manually. A single click on a node (component or reaction node) highlights it; it can be shifted now by mouse movement, or can be deleted (right mouse button click opens a context menu with the option "*Remove*"). The results of this manual editing can be saved to the Operation Field (▽) so that they will be retained for the following work.

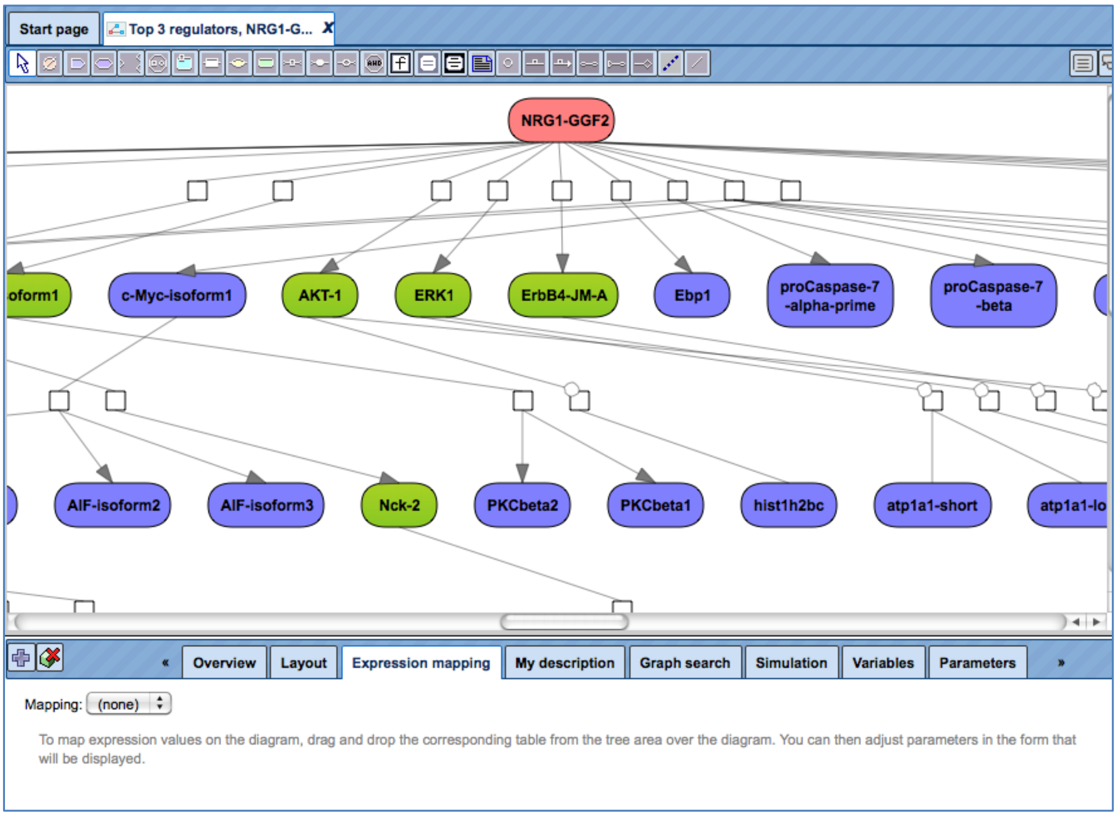
> **Please, be careful**: changes in your own diagrams are automatically saved! You can even close the diagram in the Work Space, but still can undo your last changes with the Undo button (↺). After logout, all your changes will be automatically fixed.

Further editing of the layout schemes can be done by parameter settings in the Operation Field, Layout tab. A detailed description of these layout schemes, the underlying algorithms and parameterization can be found in the Help texts.

21.3. Expression Mapping

This function enables highlighting of up-regulated and down-regulated genes in the network diagrams.

The Expression mapping tab can be found in the Operations Field after having a network diagram opened in the work area. Initially, the expression mapping form is empty as shown in the figure below.



Mapping: (none)

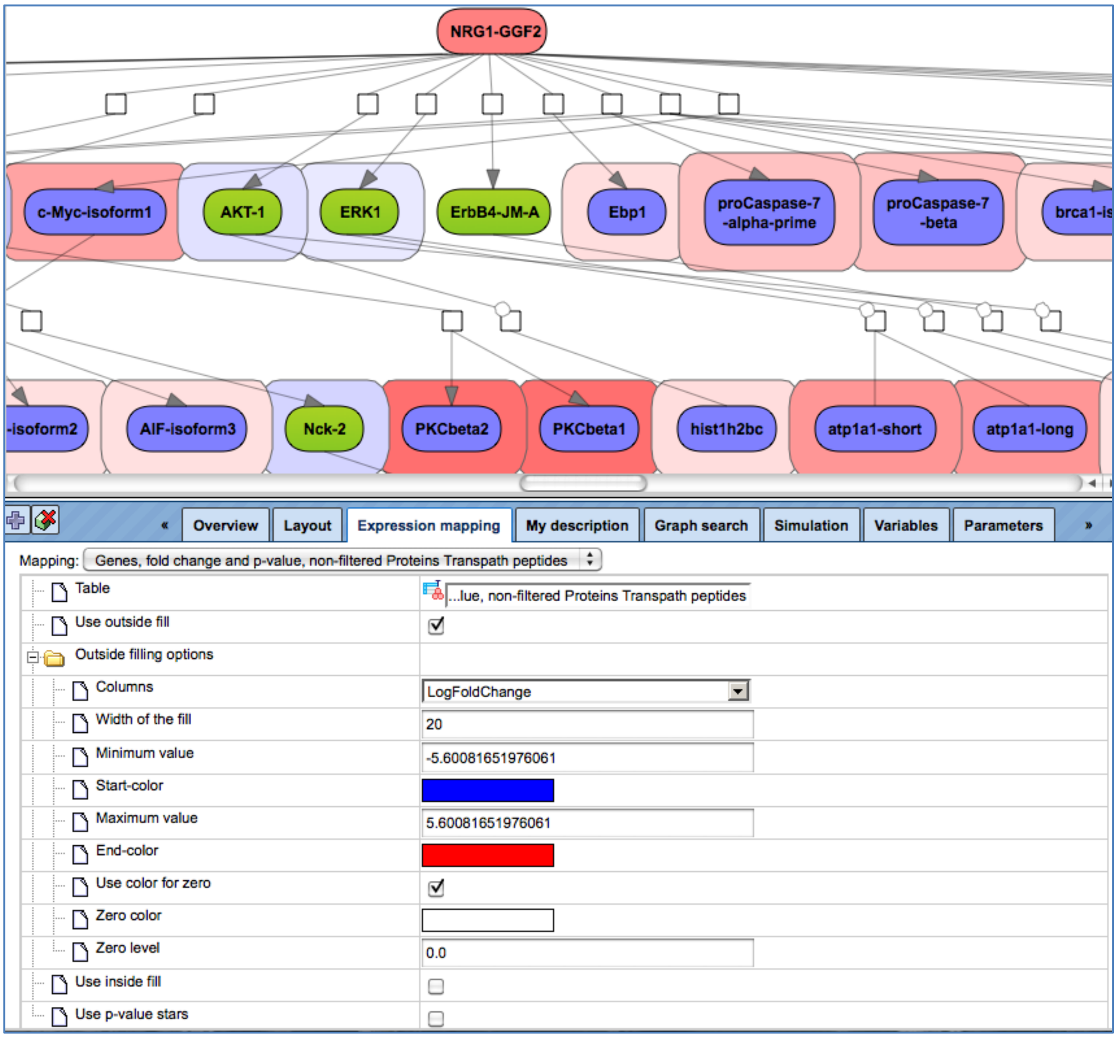
To map expression values on the diagram, drag and drop the corresponding table from the tree area over the diagram. You can then adjust parameters in the form that will be displayed.

First, drag and drop the table with expression data from the Tree Area over the diagram. You might be interested to use the table with identified differentially expressed genes and calculated fold change values for expression mapping.

Important note. The format of the table with expression data that can be used depends on the format of the diagram. If the diagram was constructed based on the TRANSPATH® database, the format of the table to drag and drop should be "Proteins: Transpath peptides", and the table should have the symbol . If the diagram was constructed based on the GeneWays database, the format of the table to drag and drop should be "Genes: Entrez", and the table should have the symbol .

Please check the format of the table with expression data before dragging and dropping it over the diagram, and if necessary, convert it to the required format with the function "Convert table" that can be found at the analyses/Methods/Data manipulation/Convert table.

After the table with expression data is dragged and dropped, the up- and down-regulated genes are automatically highlighted, and the expression mapping tab looks as shown below:

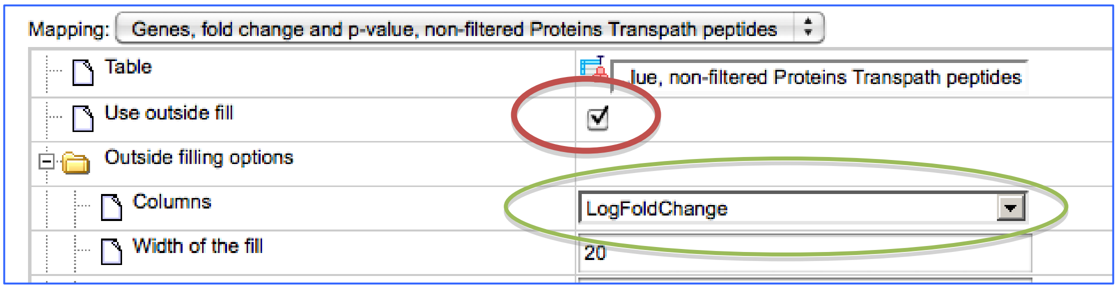


Mapping: Genes, fold change and p-value, non-filtered Proteins Transpath peptides

Table	...lue, non-filtered Proteins Transpath peptides
Use outside fill	☑
Outside filling options	
Columns	LogFoldChange
Width of the fill	20
Minimum value	-5.60081651976061
Start-color	🟦
Maximum value	5.60081651976061
End-color	🟥
Use color for zero	☑
Zero color	⬜
Zero level	0.0
Use inside fill	☐
Use p-value stars	☐

Let's consider the main options/fields of the expression mapping form.

The default type of mapping is "outside fill", and the corresponding check-box is checked, highlighted by the red oval below.

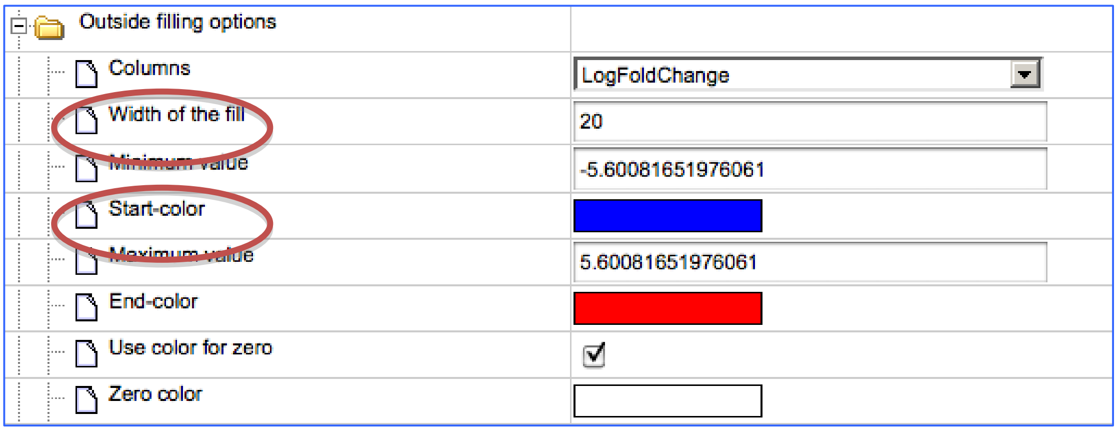


Mapping: Genes, fold change and p-value, non-filtered Proteins Transpath peptides

Table	lue, non-filtered Proteins Transpath peptides
Use outside fill	☑
Outside filling options	
Columns	LogFoldChange
Width of the fill	20

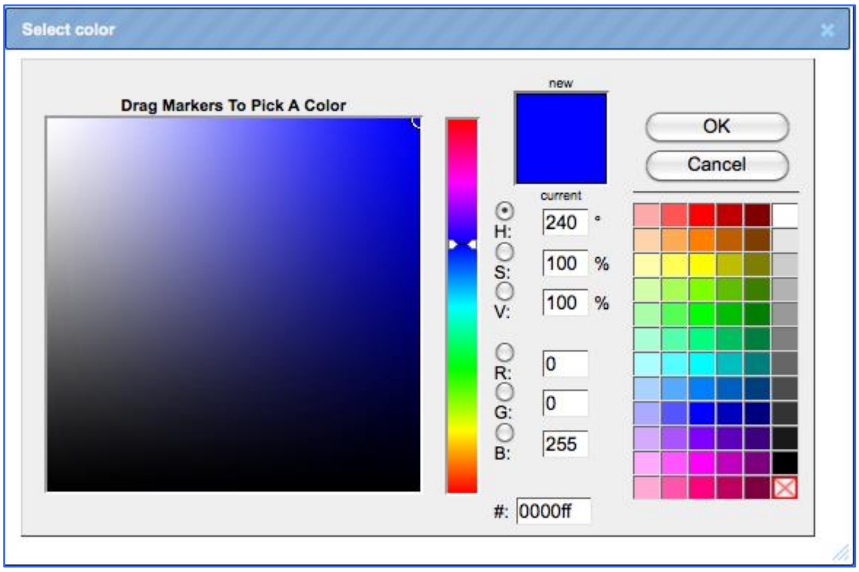
If the selected table contains a column **LogFoldChange**, this column is automatically chosen in the field "Columns", highlighted by the green oval above. Other numerical columns in the selected table are available under the drop-down menu, and can be chosen instead of the default column to be applied for mapping expression data.

The fields **Minimum value** and **Maximum value** display the corresponding values of the selected column with expression data, (see highlighted red ovals in the figure below). These values are used to calculate the intensity of the colors for expression.



Outside filling options	
Columns	LogFoldChange
Width of the fill	20
Minimum value	-5.60081651976061
Start-color	🟦
Maximum value	5.60081651976061
End-color	🟥
Use color for zero	☑
Zero color	⬜

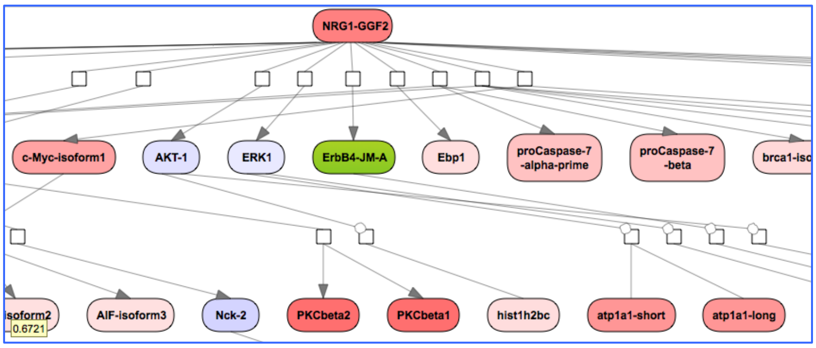
You can select colors to indicate up- and down-regulation by mouse click over the colored boxes, and the following toolbox will be displayed.



○ H: 240 °
○ S: 100 %
○ V: 100 %
○ R: 0
○ G: 0
○ B: 255
#: 0000ff

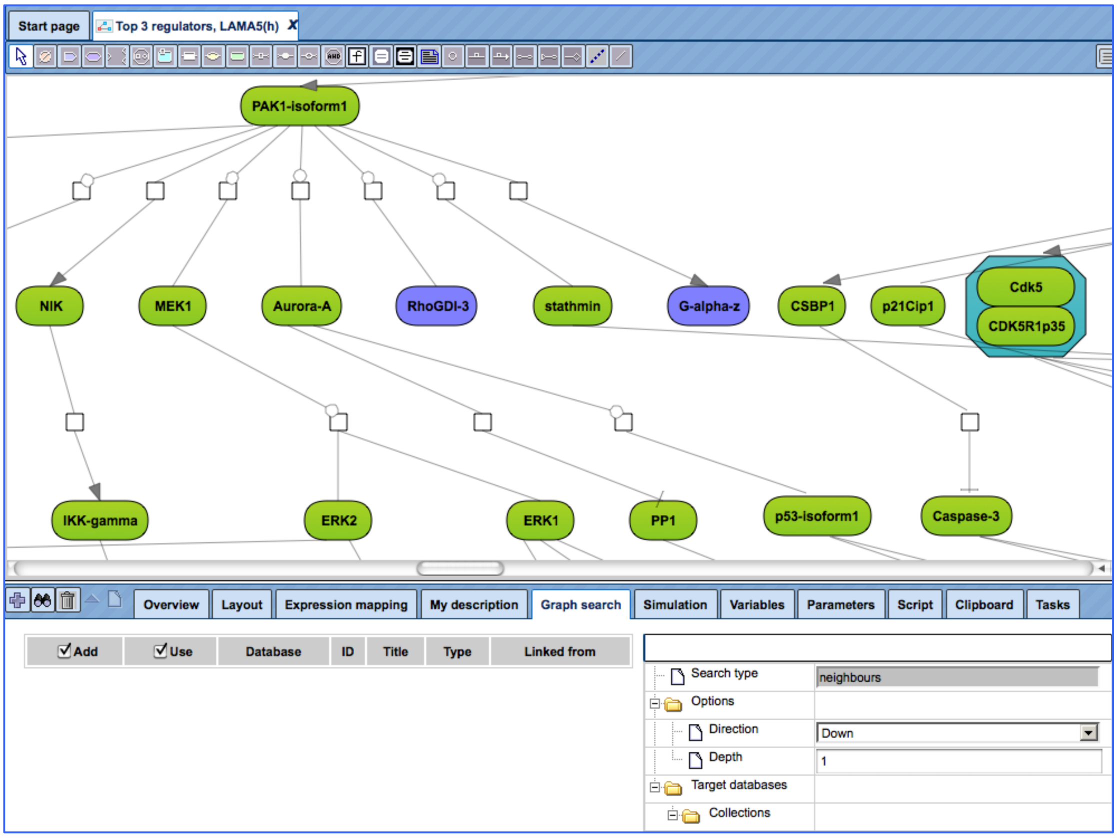
The expression values can be inferred by the color gradation. A more intensive color corresponds to a bigger fold change value whereas the lighter shade corresponds to a smaller fold change value. In this example all the up-regulated genes are shown with a color gradient from white to red whereas down-regulated genes are shown in with a color gradient from white to blue.

Check-in the box **Use inside fill** and check-out the box **Use outside fill** result in the following picture.



21.4. Graph Search

The "Graph search" option allows to extend diagrams already saved in the tree, e.g. to add interactions around the molecules in focus. Whenever a network diagram is opened in the Work Space, the "Graph search" tab can be found in the Operations Field as it is shown below.



☑ Add	☑ Use	Database	ID	Title	Type	Linked from

Search type	neighbours
Options	
Direction	Down
Depth	1
Target databases	
Collections	

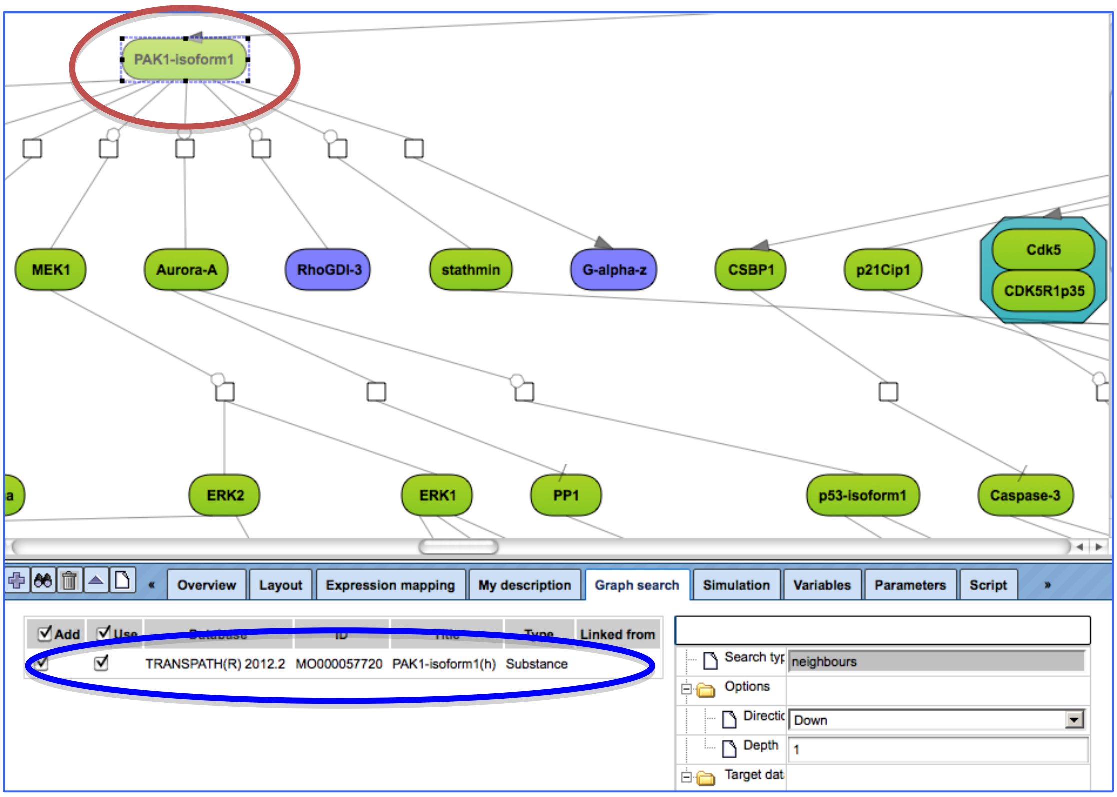
The gray fields in the graph search form cannot be edited. The user can change the search direction (upstream of the molecule in focus, or downstream, or both directions) and number of steps (depth) in the selected direction.

Graph search can be done in several iterations.

First iteration of a graph search.

To perform a graph search the following steps are recommended:

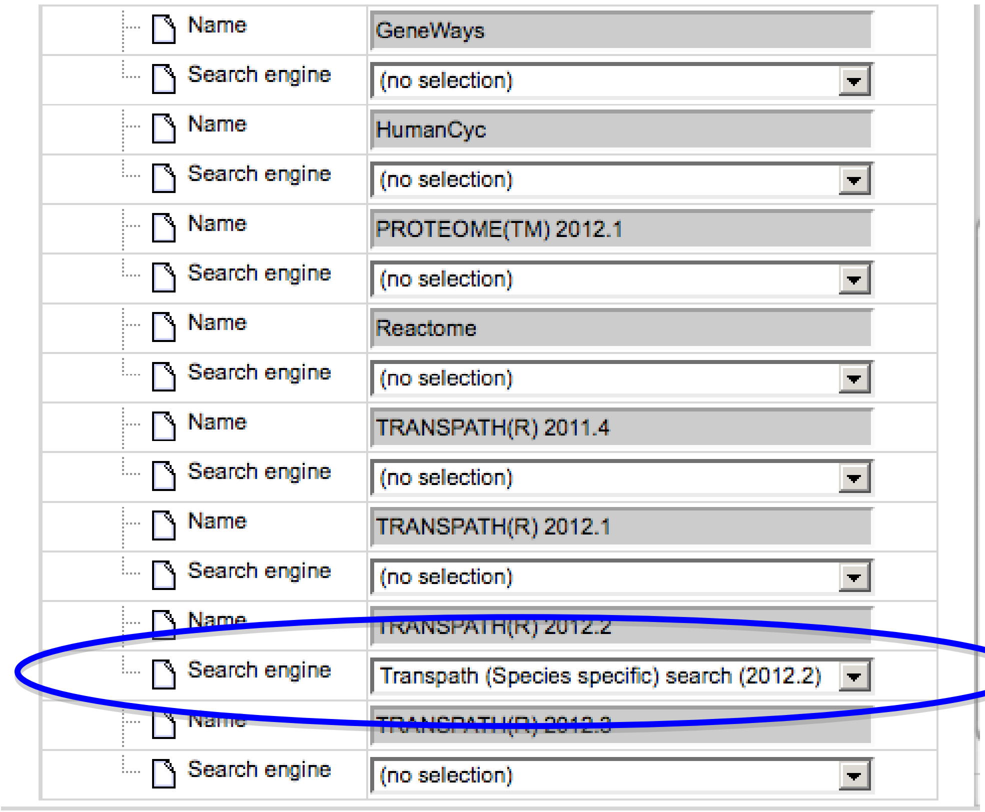
- *Open a diagram in the Work Space and select one element (gene or protein or reaction) for which you want to perform search by mouse-clicking over that element. On the picture below, the molecule "PAK1-isoform1" has been selected, marked by the red oval.*
- *Use the button* ✚ *to add the selected molecule to the elements pane. The added element is marked by the blue oval on the picture below.*



☑ Add	☑ Use	Database	ID	Title	Type	Linked from
☑	☑	TRANSPATH(R) 2012.2	MO000057720	PAK1-isoform1(h)	Substance	

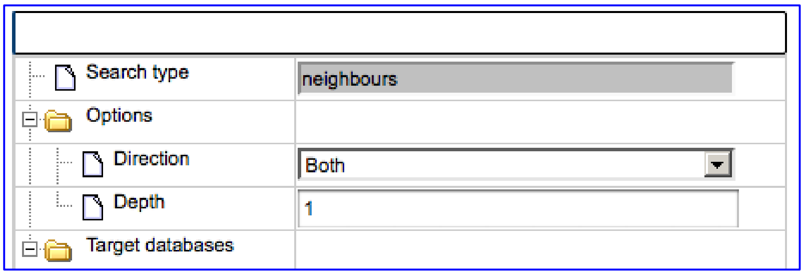
Search typ	neighbours
Options	
Directic	Down
Depth	1
Target dat	

- *Simultaneously, the database to search for additional interactions is automatically identified as the database based on which the network diagram is constructed. For the identified database the **Search engine** field is specified, marked by the blue oval on the picture below.*



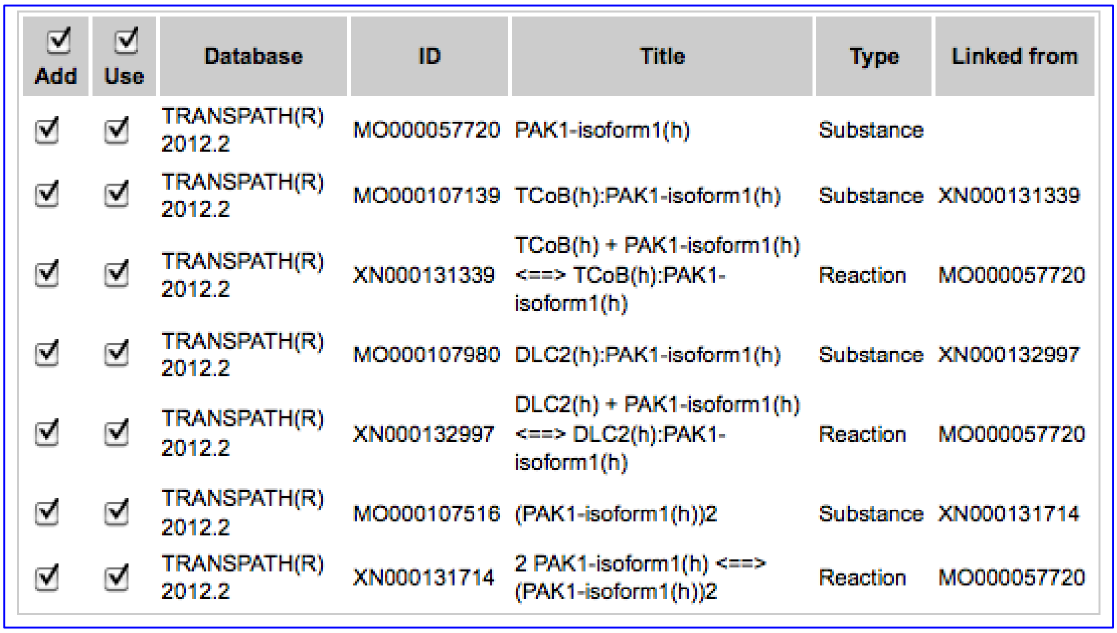
Name	GeneWays
Search engine	(no selection)
Name	HumanCyc
Search engine	(no selection)
Name	PROTEOME(TM) 2012.1
Search engine	(no selection)
Name	Reactome
Search engine	(no selection)
Name	TRANSPATH(R) 2011.4
Search engine	(no selection)
Name	TRANSPATH(R) 2012.1
Search engine	(no selection)
Name	TRANSPATH(R) 2012.2
Search engine	Transpath (Species specific) search (2012.2)
Name	TRANSPATH(R) 2012.3
Search engine	(no selection)

- *Next, choose the direction and depth of search.*



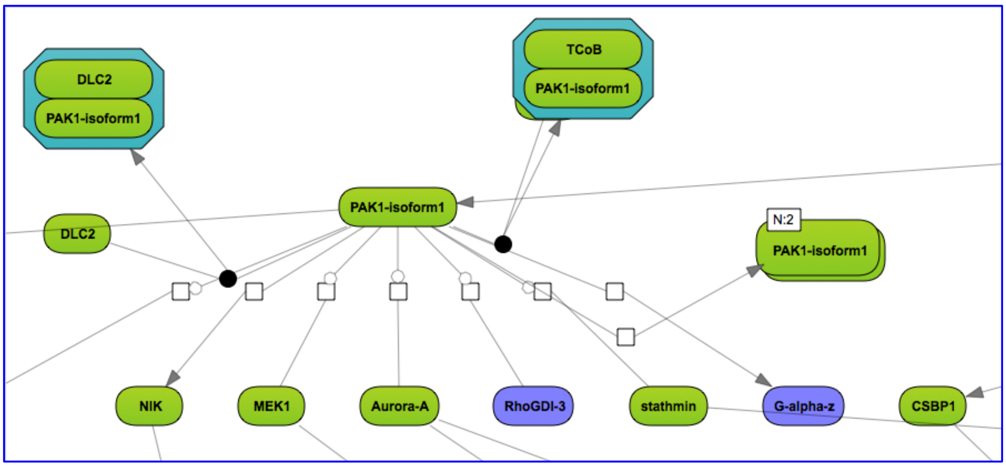
Search type	neighbours
Options	
Direction	Both
Depth	1
Target databases	

- *Use to start the search. Once the search is finished you get the search results in the elements pane as shown in the screenshot below.*



☑ Add	☑ Use	Database	ID	Title	Type	Linked from
☑	☑	TRANSPATH(R) 2012.2	MO000057720	PAK1-isoform1(h)	Substance	
☑	☑	TRANSPATH(R) 2012.2	MO000107139	TCoB(h):PAK1-isoform1(h)	Substance	XN000131339
☑	☑	TRANSPATH(R) 2012.2	XN000131339	TCoB(h) + PAK1-isoform1(h) <==> TCoB(h):PAK1-isoform1(h)	Reaction	MO000057720
☑	☑	TRANSPATH(R) 2012.2	MO000107980	DLC2(h):PAK1-isoform1(h)	Substance	XN000132997
☑	☑	TRANSPATH(R) 2012.2	XN000132997	DLC2(h) + PAK1-isoform1(h) <==> DLC2(h):PAK1-isoform1(h)	Reaction	MO000057720
☑	☑	TRANSPATH(R) 2012.2	MO000107516	(PAK1-isoform1(h))2	Substance	XN000131714
☑	☑	TRANSPATH(R) 2012.2	XN000131714	2 PAK1-isoform1(h) <==> (PAK1-isoform1(h))2	Reaction	MO000057720

- *When the **Add** box is checked, the corresponding molecules will be added to the diagram.*
  *The column with the check box **Use** means that selected molecules can be in turn used for searching, e.g. in the next search iteration. Direction of the search (up, down or both) and depth will be applied to the molecules with checked boxes in the **Use** column. By default, all boxes are checked. If you like to add to the diagram only a few molecules, you need to uncheck the others.*
- *In the next step, found molecules can be added to the current diagram by using the icon. By default all found molecules are checked. The user can uncheck some of the molecules in the column **Add** before adding to the diagram. Only the molecules checked in the column **Add** are added to the diagram. Here, all molecules are added, as shown on the picture below.*



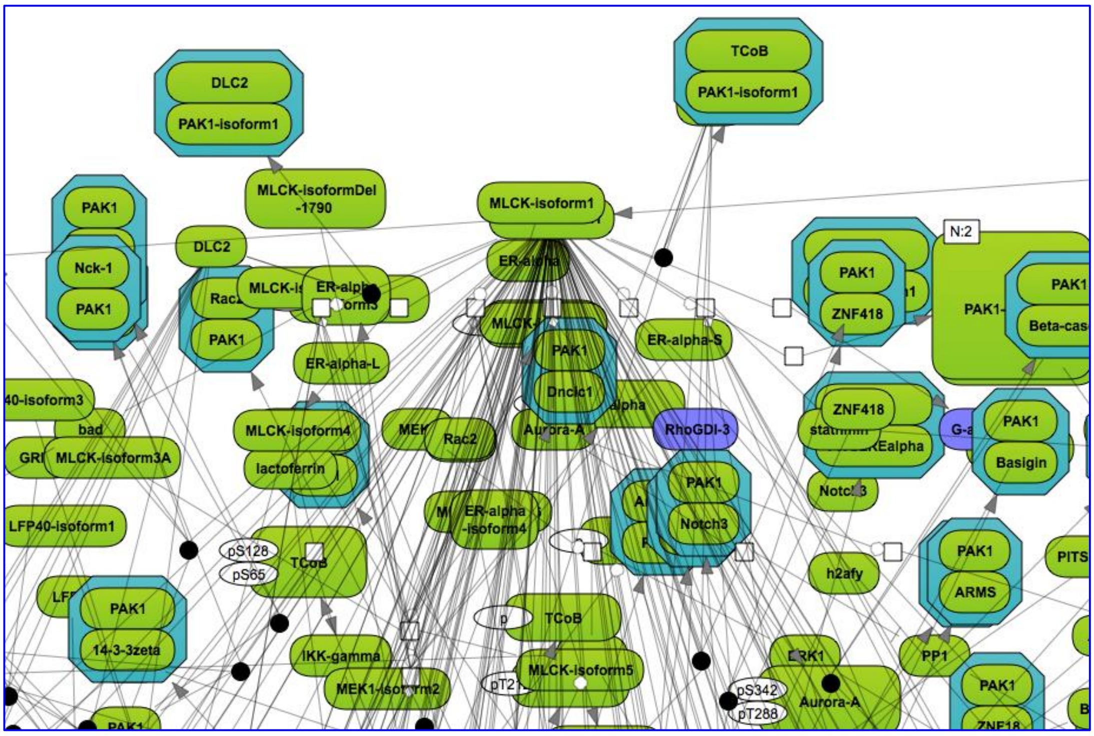
Now, the layout for the extended diagram can be specified, details of the layouts are described in Section "Changing the layout".

If you would like to remove the search results and select another molecule, the button 🗑 can be applied.

Second and next iterations of the graph search.

- *Having several molecules in the element pane at the previous iteration of a graph search, the user can start searching again in the specified direction from these molecules.*
- *The user can uncheck some of the molecules in the column "Use" before starting the search again. Only the molecules checked-in in the column "Use" are used for the second search.*

After the search is completed, all found molecules can be again added to the diagram, which may result in a picture, like shown below:



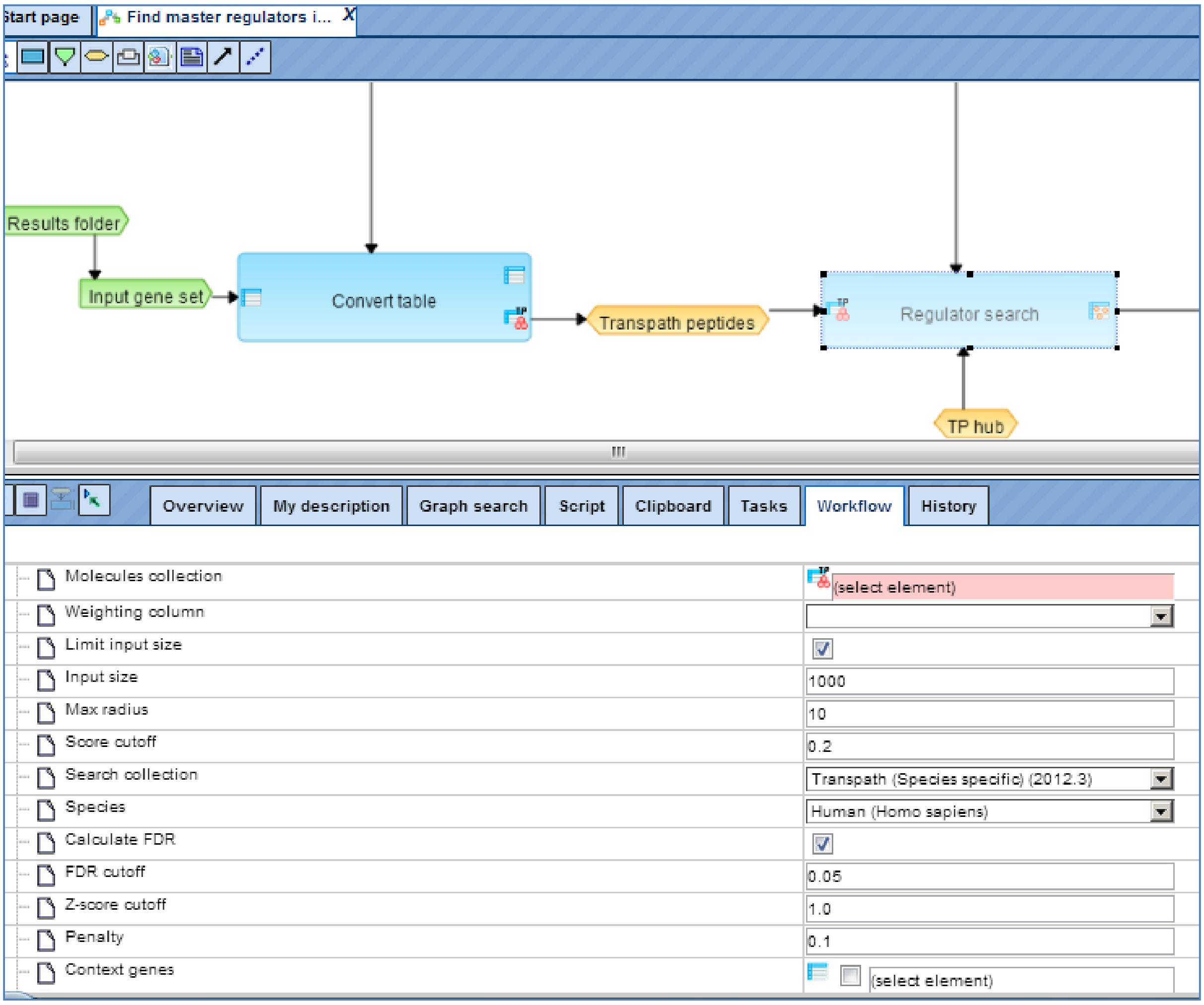
22. Editing and creating workflows

22.1. Edit a pre-existing workflow

Tip for the workflow editing

You can easily create a similar workflow with parameter values adjusted to your needs. For example, you might be interested to change the number of steps used for the regulator search. By default, 10 steps are applied.

To make a change, you need first to open the workflow under the “Edit workflow” mode, and save its copy in your project area. The [Edit workflow] button is located near the button [Run workflow]. Upon clicking on "Edit workflow", the workflow diagram will be opened in the Work Space, and you can select the analysis box you would like to modify. On the screenshot below "Regulator search" analysis was selected, and in the Operations Field, on the tab "Workflow", all the parameters are visible. Under this mode, you can modify default parameters and then save the workflow.

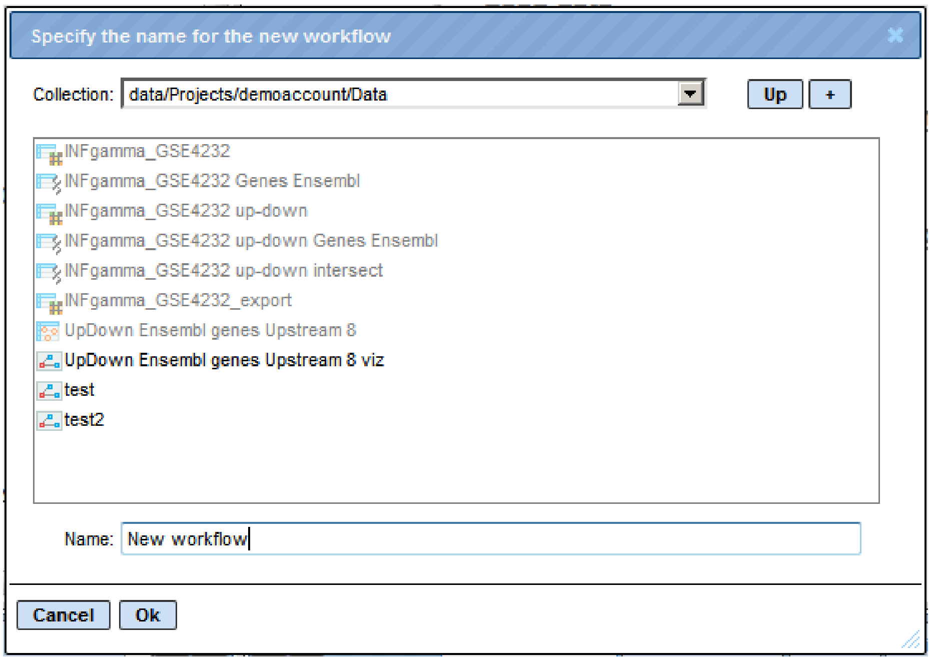


Molecules collection	(select element)
Weighting column	
Limit input size	☑
Input size	1000
Max radius	10
Score cutoff	0.2
Search collection	Transpath (Species specific) (2012.3)
Species	Human (Homo sapiens)
Calculate FDR	☑
FDR cutoff	0.05
Z-score cutoff	1.0
Penalty	0.1
Context genes	☐ (select element)

In this way you will get a customized workflow, with the parameters specified according to your needs.

22.2. Create a new workflow

To create your own workflow, please go to the Start page and click "Create your own workflow" under the list of pre-defined workflow groups. You will be asked to specify the name of the new workflow.



Specify the name for the new workflow

Collection: data/Projects/demoaccount/Data

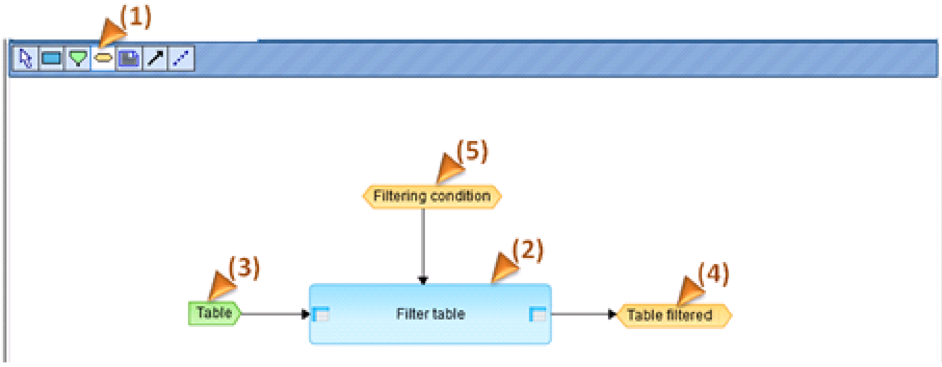
- INFgamma_GSE4232
- INFgamma_GSE4232 Genes Ensembl
- INFgamma_GSE4232 up-down
- INFgamma_GSE4232 up-down Genes Ensembl
- INFgamma_GSE4232 up-down intersect
- INFgamma_GSE4232_export
- UpDown Ensembl genes Upstream 8
- UpDown Ensembl genes Upstream 8 viz
- test
- test2

Name: New workflow

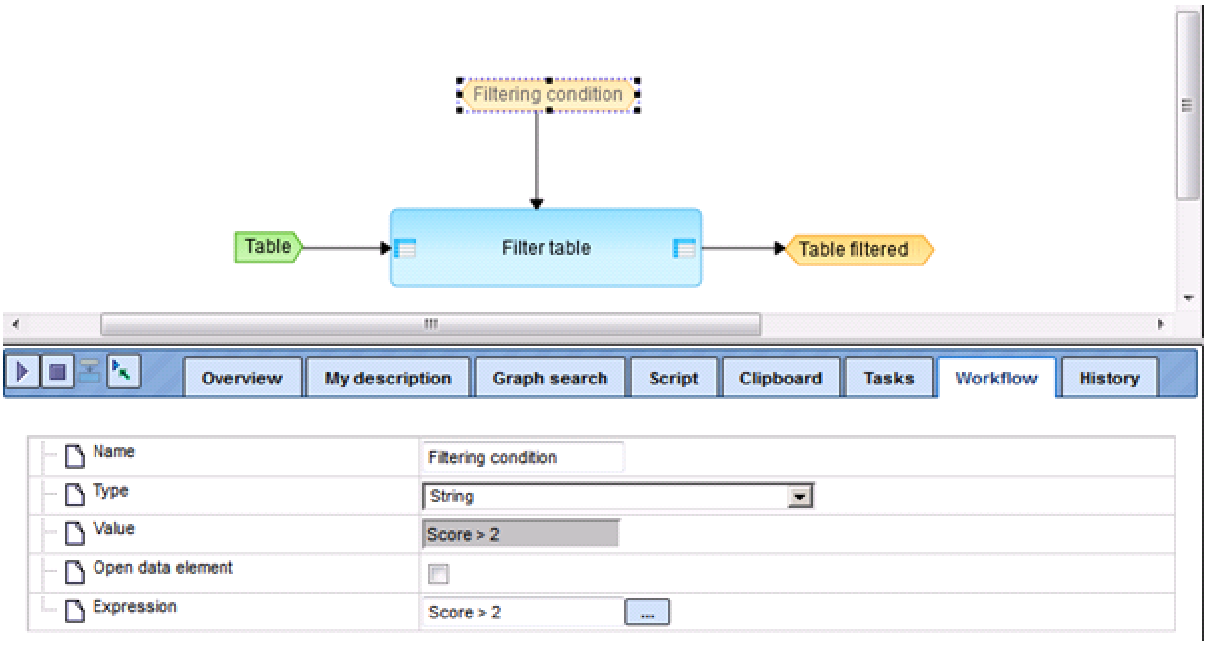
In this example, an entry *New workflow* (default name) will be created.

You will see now a new tab being opened in the Work Space, where you can design a new workflow diagram. The workflow diagram represents different analysis functions being connected by input and output files. The resulting directed graph visualizes the sequence of analysis steps in the workflow. The diagram also may contain parameters, which are to be defined by the user.

Below you can see an example of one step of a new workflow. At the top of the Work Space there is a toolbox (1) to create all the elements of the workflow. The light blue rectangle in the center (2) is the analysis function (Filter table) to be used in this step. The green box at the left (3) stands for an input table, the right yellow box (4) for the output table of this analysis function. The yellow box above (5) represents a parameter of the analysis function, which in this case (Filter table function) defines a filtering condition.



Upon clicking on any component of the workflow you can see the information about this particular element in the Operations Field. For instance, in the figure below, we can see the content of the parameter "Filtering condition", which was set as "Score > 2" (so the input table will be filtered by this condition in the table column "Score").

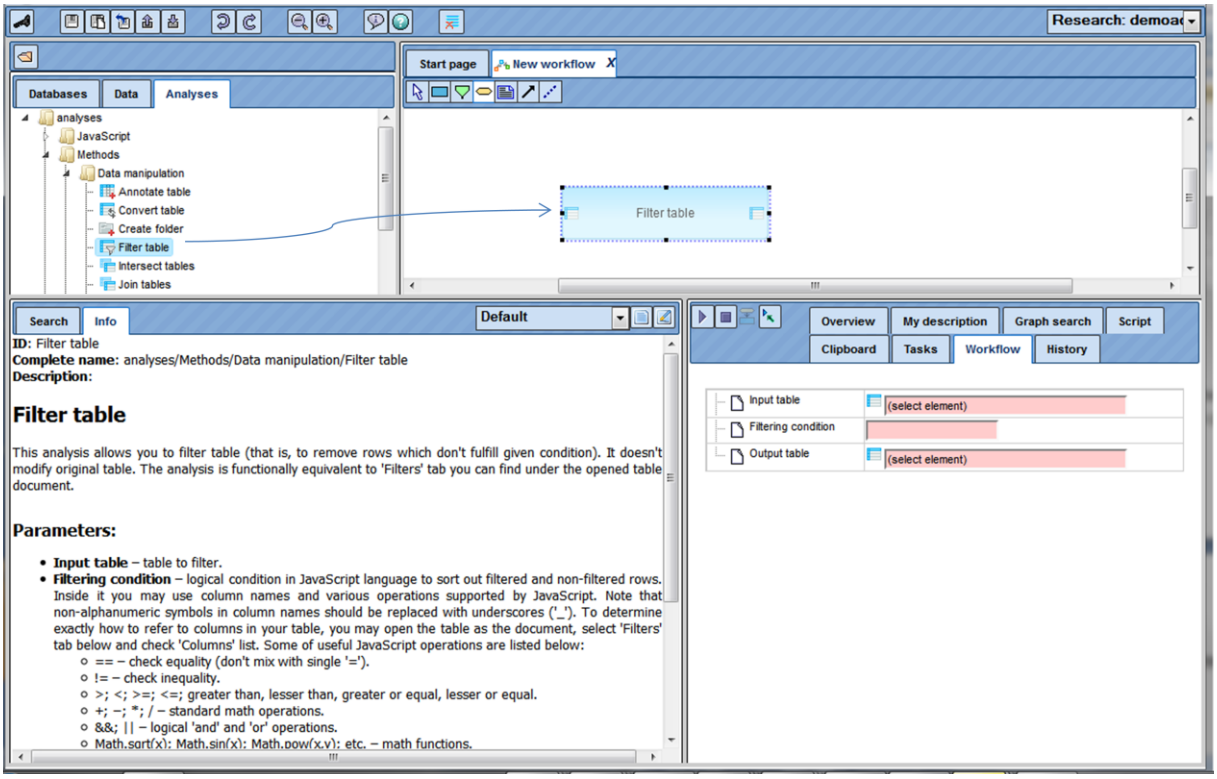


Name	Filtering condition
Type	String
Value	Score > 2
Open data element	□
Expression	Score > 2

22.2.1. Six steps to create a simple workflow

Let us go now through the steps of creating this workflow.

Step 1. The elements of the workflow can be placed into the Work Space by using the toolbar or by drag and drop from the Tree Area. Let us place the Filter table function by drag and drop from the subdirectory Methods/Data manipulation on the "Analysis" tab of the Tree Area:



ID: Filter table
Complete name: analyses/Methods/Data manipulation/Filter table
Description:

Filter table

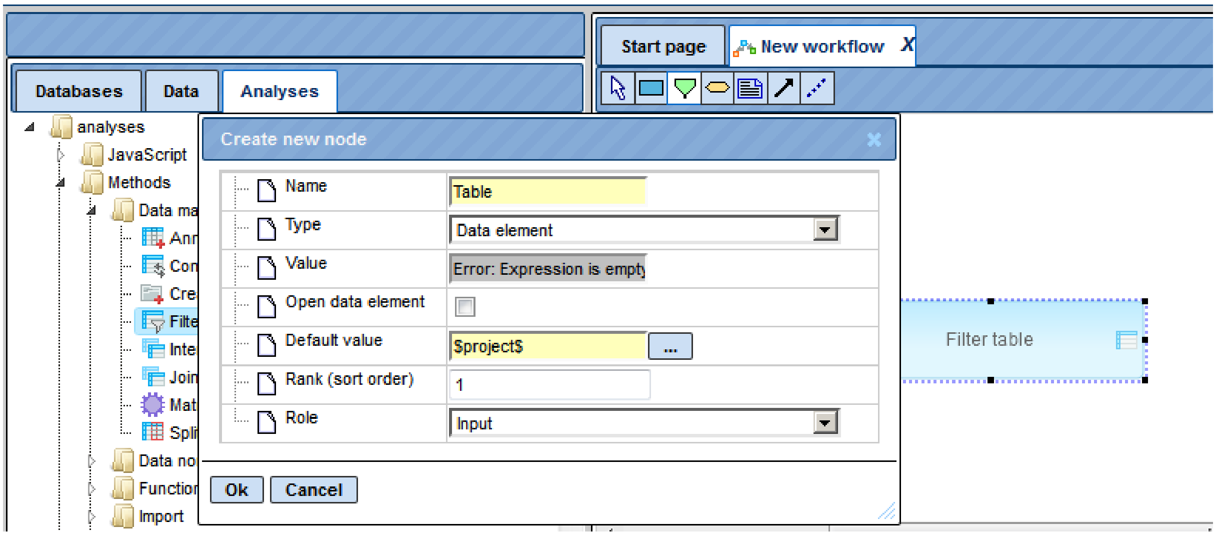
This analysis allows you to filter table (that is, to remove rows which don't fulfill given condition). It doesn't modify original table. The analysis is functionally equivalent to 'Filters' tab you can find under the opened table document.

Parameters:

- **Input table** – table to filter.
- **Filtering condition** – logical condition in JavaScript language to sort out filtered and non-filtered rows. Inside it you may use column names and various operations supported by JavaScript. Note that non-alphanumeric symbols in column names should be replaced with underscores ('_'). To determine exactly how to refer to columns in your table, you may open the table as the document, select 'Filters' tab below and check 'Columns' list. Some of useful JavaScript operations are listed below:
  - == – check equality (don't mix with single '=').
  - != – check inequality.
  - >; <; >=; <=; greater than, lesser than, greater or equal, lesser or equal.
  - +; –; *; / – standard math operations.
  - &&; || – logical 'and' and 'or' operations.
  - Math.sqrt(x); Math.sin(x); Math.pow(x,y); etc. – math functions.

Input table	(select element)
Filtering condition	
Output table	(select element)

Step 2. For creating the input table you should click on the green element in the tool bar, locate the cursor in the Work Space where you would like to put this element and click. A new window "Create new node" will pop up, where you can define the parameters of the element as will be explained in the following.

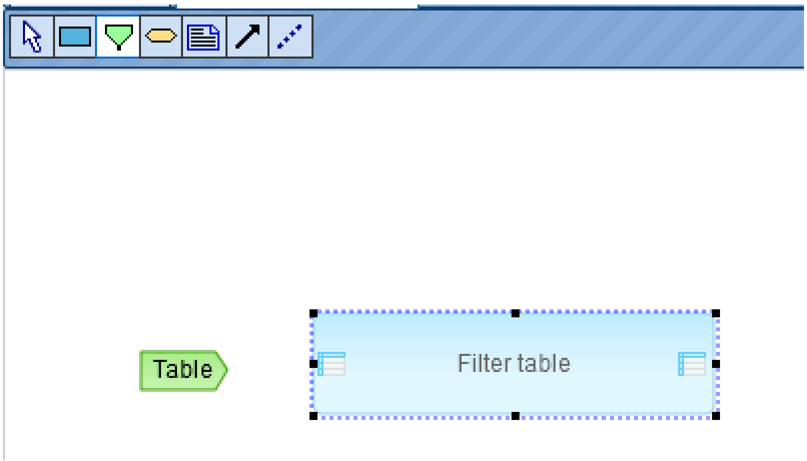


Create new node

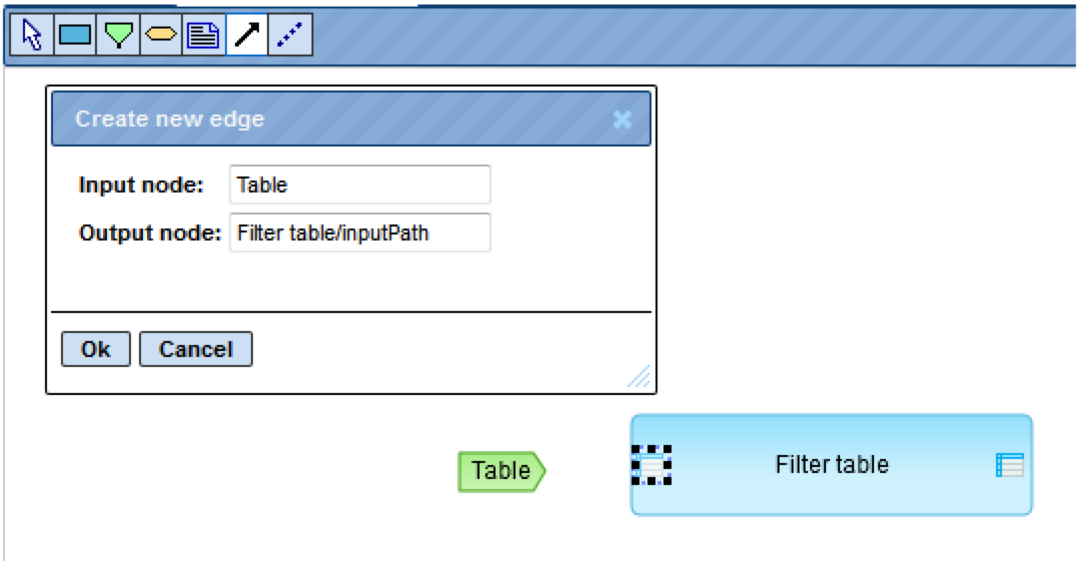
Name	Table
Type	Data element
Value	Error: Expression is empty
Open data element	□
Default value	$project$
Rank (sort order)	1
Role	Input

Name field: the title of the element; **Type**: please, select "Data element", for any objects like tables. In the field **Default value** you can type a full folder path where the table is located. You can also use some global variables, like "$project$" that already contain the full path (by clicking on the "…" button you can access all the global variables defined for this workflow. **Rank (sort order)**: this number gives the position of this input element in the list of all elements upon starting the workflow. **Role**: Input, since we are using this element for inputting a table into the workflow.

We have created now an input table element:



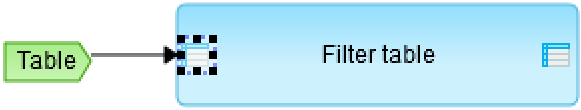
Step 3. Connecting elements on the diagram is done by clicking on the arrow symbol in the toolbar. A new window "Create new edge" will pop up. By clicking on the "Table" element in the Work Space you select it as **Input node** for this edge. Similarly, you click on the left table symbol in the "Filter table" element to select it as an **Output node** of this new edge.


Create new edge

Input node: Table

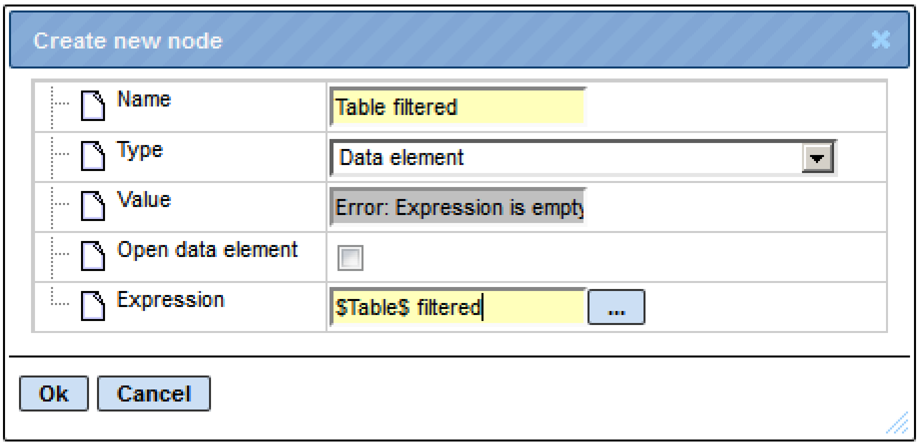
Output node: Filter table/inputPath

After pressing "OK" a new connection is created:



Step 4. The same way you can create now an output table element on the diagram by selecting the yellow element in the tool bar (since it is going to be an intermediate table

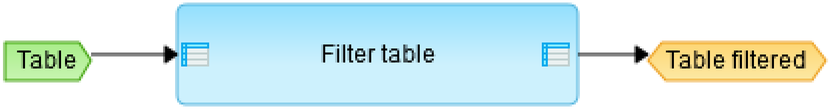
for further use in the next steps of the workflow) and connecting it with the output icon of the “Filter table” element.


Create new node

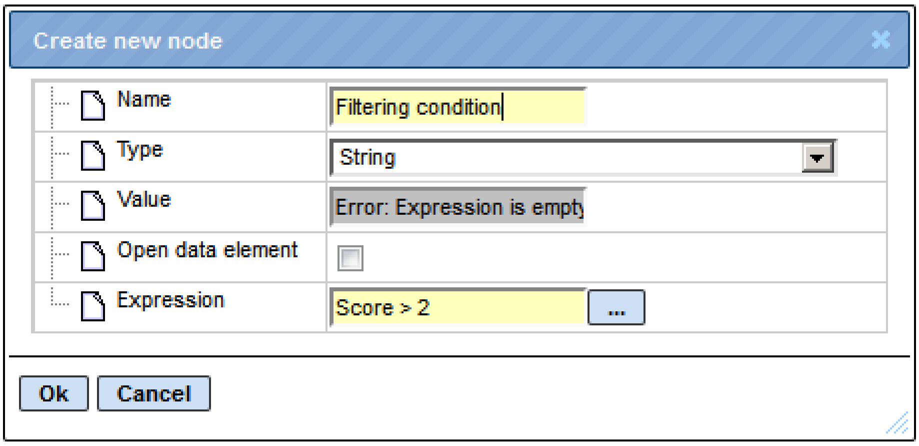
Name	Table filtered
Type	Data element
Value	Error: Expression is empty
Open data element	□
Expression	$Table$ filtered

In the field “Expression” you can now use a new global variable “$Table$” which will contain, during the run of the workflow, the name of the table which you have entered. So in this case we are creating a new name for the future output table “$Table$ filtered” by adding to the name of the input table an ending “filtered”.

As a result, we have now created one step of the workflow.

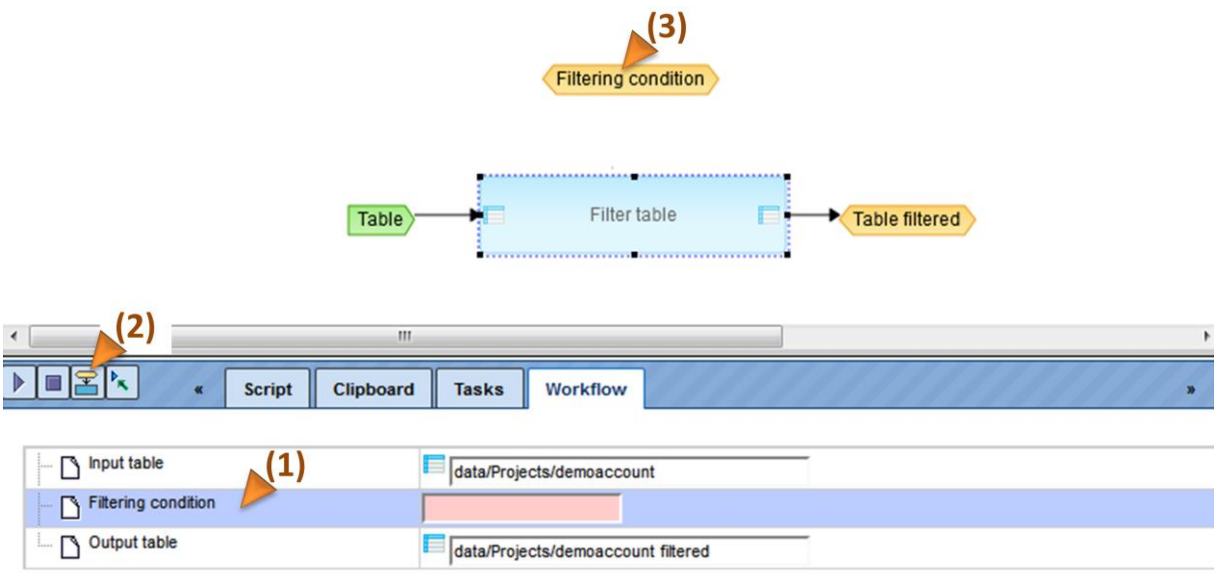


Step 5. To filter data, we have to define a filtering condition. To do this we have to create a new element “Filtering condition” (yellow element in the tool bar), which will be now of simple “String” type and which contains a filtering condition “Score > 2” in the field **Expression**.


Create new node

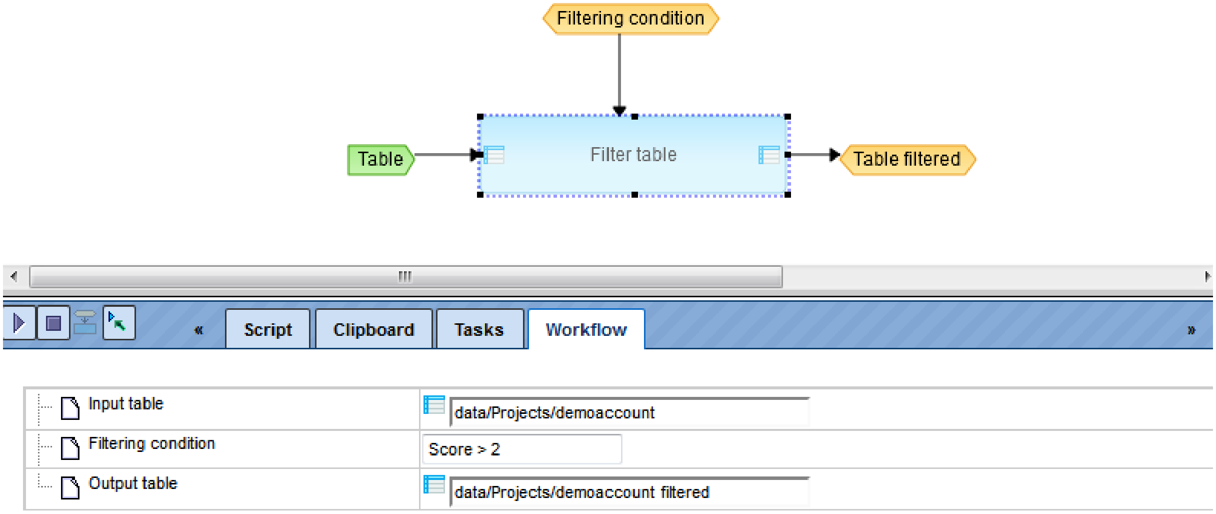
Name	Filtering condition
Type	String
Value	Error: Expression is empty
Open data element	□
Expression	Score > 2

A new element “Filtering condition” is created. This element should now be connected to the analysis function “Filter table” in order to define the filtering condition that is going to be applied in this step of the workflow. To do that, please click first on the "Filter table" element and open the parameters of this element in the Operations Field. After that, click on the field "Filtering condition" (1) in the parameter list and select it (a blue background color indicates that the field is selected). Click on the "Bind property to variable" button (2) in the toolbar of the Operations Field. And after that, move the cursor to the Work Space and click on the "Filtering condition" element on the diagram (3).



Input table	data/Projects/demoaccount
Filtering condition	
Output table	data/Projects/demoaccount filtered

So, a filtering condition parameter is now connected to the corresponding field in the "Filter table" function.



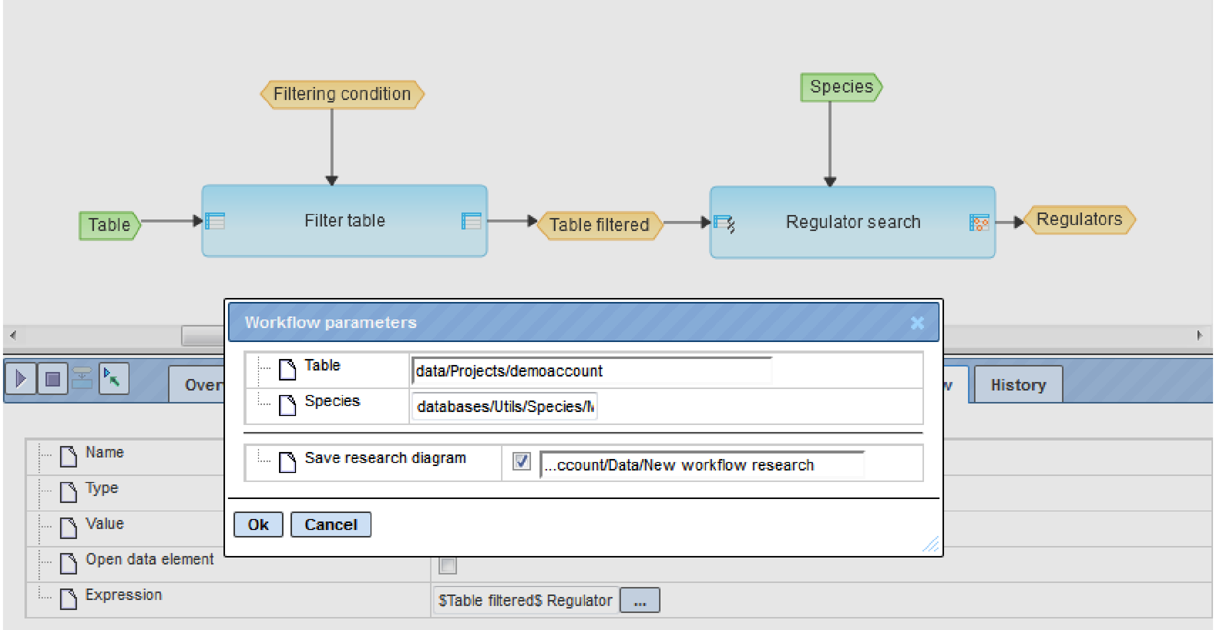
Input table	data/Projects/demoaccount
Filtering condition	Score > 2
Output table	data/Projects/demoaccount filtered

Step 6. The workflow is now ready to be executed. To start the workflow please click on the "Run workflow" button (▷) in the toolbar of the Operations Field.

In the pop-up menu "Workflow parameters" you should specify the input table. Please navigate to the folder with your tables and select a table which has a column **Score** and press [Ok]. The workflow will be executed and a new table with a new name and the appendix “filtered” will be created in the same folder as the input table.

22.2.2. Complex workflows

More complex workflows are created by adding further workflow steps and by connecting them through a common data element. As in the example below, the output element “Table filtered” of the first step is used as an input element of the second step of the workflow, “Regulator search”. You can also see that we have added a new input parameter “Species”, which appears now among the workflow parameters upon starting the workflow. With this, you can select a taxonomic species (presently human, mouse or rat) for the table you are going to run through the workflow.



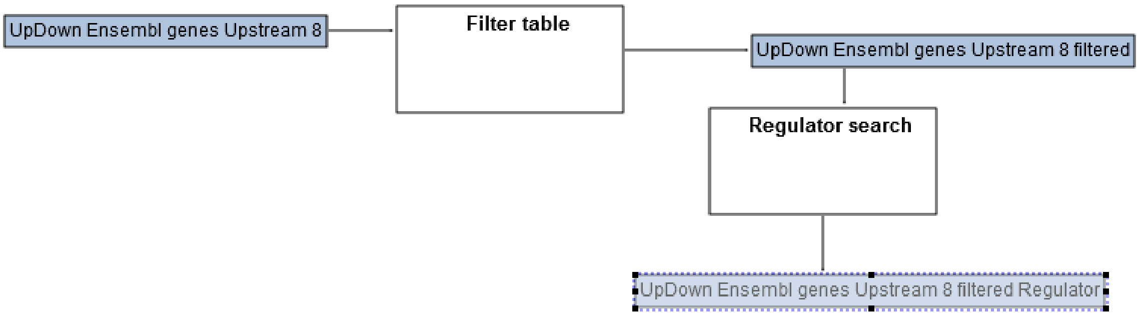
Workflow parameters

Table	data/Projects/demoaccount
Species	databases/Utils/Species/I
Save research diagram	☑ ...ccount/Data/New workflow research

Ok Cancel

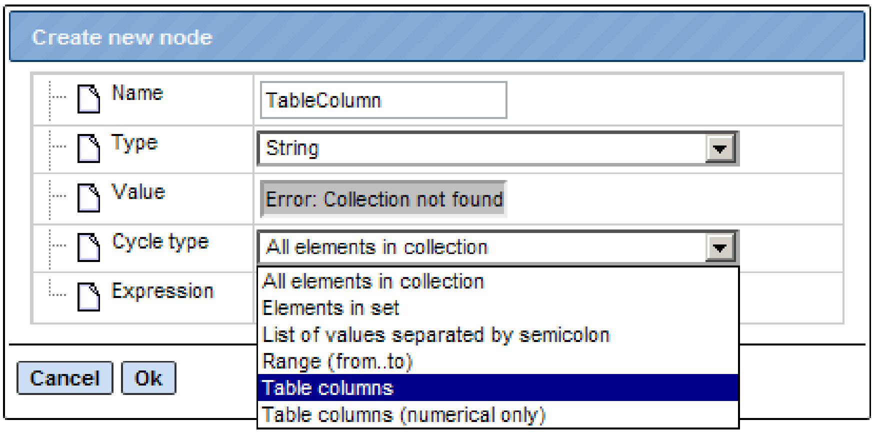
Name	
Type	
Value	
Open data element	□
Expression	$Table filtered$ Regulator ...

Note: During execution of the workflow a *research diagram* is saved (you can specify the name of this diagram before starting of the workflow). The research diagram (see figure below for an example) contains the history of the workflow execution with the names of all input and output files. It also contains all the links to these tables, so that you can easily open them by clicking on the respective element in the diagram.


UpDown Ensembl genes Upstream 8
Filter table
UpDown Ensembl genes Upstream 8 filtered
Regulator search
UpDown Ensembl genes Upstream 8 filtered Regulator

22.2.3. Cycles and scripts

One more element available in the workflow is *cycle*. It can be created using the “cycle” button in the tool bar. It is necessary to specify the **Name** of the cycle (the “cycle variable”), and to choose the appropriate **Type** and **Cycle type**.

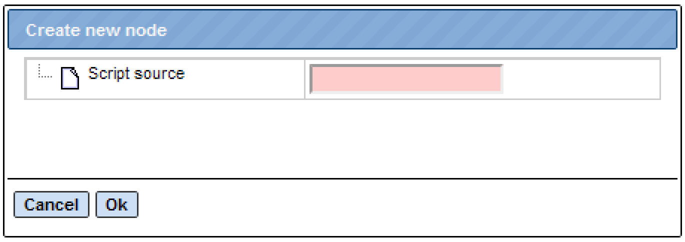

Create new node

Name	TableColumn
Type	String
Value	Error: Collection not found
Cycle type	All elements in collection
Expression	

All elements in collection
Elements in set
List of values separated by semicolon
Range (from..to)
Table columns
Table columns (numerical only)
Cancel
Ok

The option “All elements in collection” (**Cycle type**) together with the **Type** “Data element” and some folder name in the field **Expression** means that all data elements from that folder will be taken one by one as cycle variable values. E.g., when selecting **Cycle type** “Table columns”, **Expression** should specify the name of this table. Or when choosing **Cycle type** “Range (from ...to)” for **Type** “Integer number” and **Expression** as “2 ... 6”, the cycle will be executed by assigning the values 2,3,4,5 and 6 to the cycle variable.

The *Script* workflow element represents a code written in JavaScript, which can be executed during the workflow run. To add a script the user should press the tool bar button “analysis-script”, click the proper place in the workflow diagram area and type the JavaScript code in the **Script source** field.


Create new node

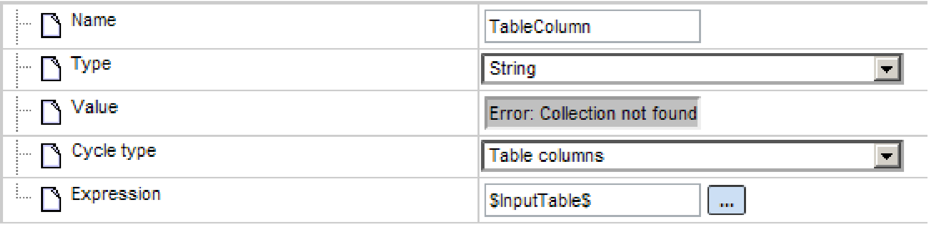
Script source	

Another way to add a script to the workflow is to drag and drop some script data element right from the project tree.

All variables defined in a workflow (green and orange boxes) are available inside the script as JavaScript variables. If the name doesn’t contain spaces, it can be used as is, the name should be put into $[“…”] otherwise. For example, the variable “TableColumn” can be accessed from a script by name TableColumn, but “Table column N1” should be called as $[“Table column N1”].

Example 1. Print column names

In the example below the data element parameter InputTable (green box) is first added to the workflow. Then a cycle with the following settings is added:



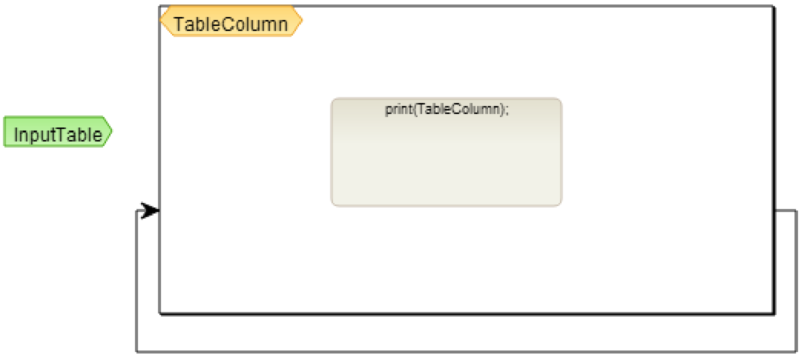
Name	TableColumn
Type	String
Value	Error: Collection not found
Cycle type	Table columns
Expression	$InputTable$

Note that expression is set to $InputTable$ for a cycle.

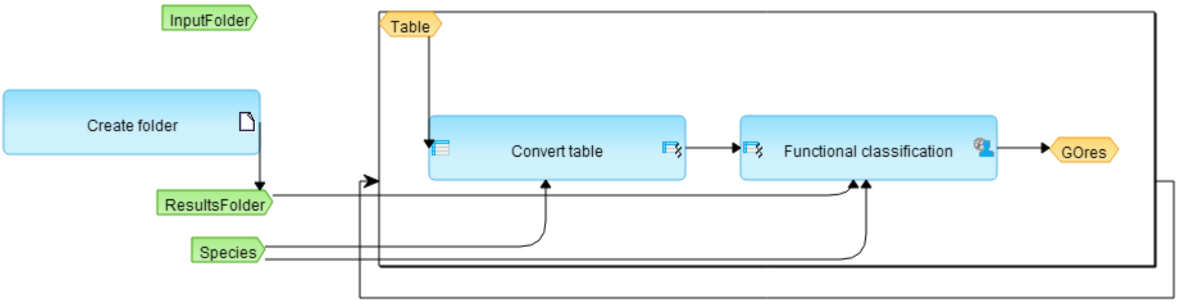
Then a script element with the following code will be created inside a cycle:

print(TableColumn);

When the [Run] button is pressed, the workflow will ask for a table path, and then will print all column names into the workflow output log.

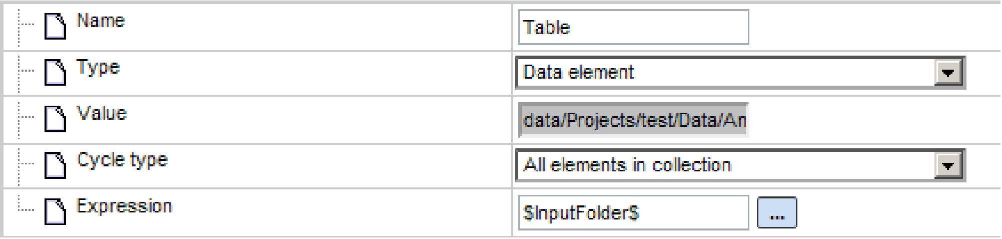


Example 2. Run GO classification for all tables in a folder



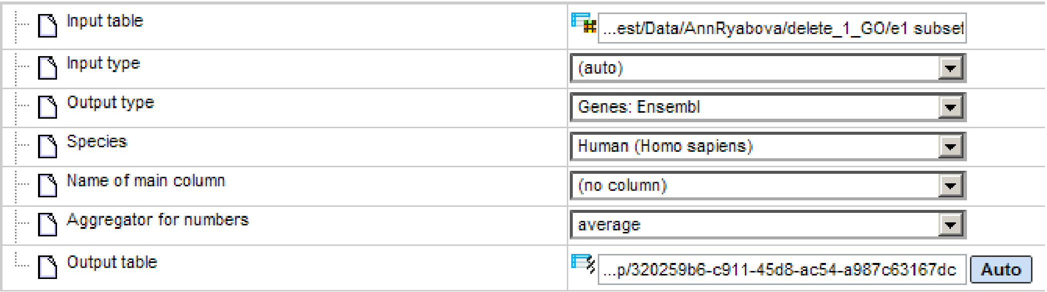
This workflow contains 3 input elements: InputFolder, where one or more Tables should be placed; ResultFolder that will be created by the method "Create folder", and the data element Species, required for table conversion and functional classification.

The cycle has following settings:



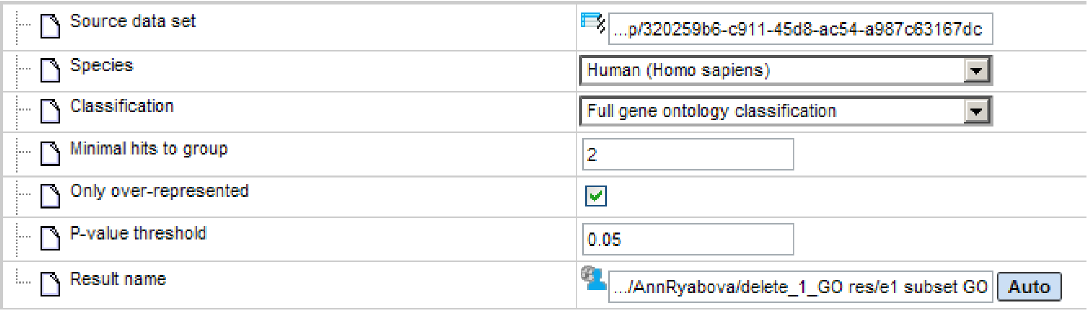
Name	Table
Type	Data element
Value	data/Projects/test/Data/An
Cycle type	All elements in collection
Expression	$InputFolder$ ...

Here the cycle variable named Table will adopt the names of the tables in the InputFolder. Then it goes to the method "Convert table", and identifiers are converted to Ensembl genes, according to the analysis settings below:



Input table	...est/Data/AnnRyabova/delete_1_GO/e1 subset
Input type	(auto)
Output type	Genes: Ensembl
Species	Human (Homo sapiens)
Name of main column	(no column)
Aggregator for numbers	average
Output table	...p/320259b6-c911-45d8-ac54-a987c63167dc Auto

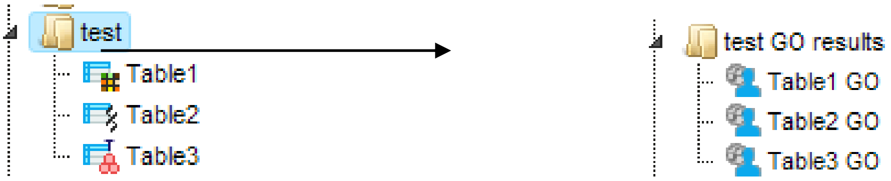
The conversion result is taken as input set in the Functional classification analysis:



Source data set	...p/320259b6-c911-45d8-ac54-a987c63167dc
Species	Human (Homo sapiens)
Classification	Full gene ontology classification
Minimal hits to group	2
Only over-represented	☑
P-value threshold	0.05
Result name	.../AnnRyabova/delete_1_GO res/e1 subset GO Auto

The output of the functional classification is a data element named *GOres* with Expression $ResultsFolder$/$Table/name$ GO.

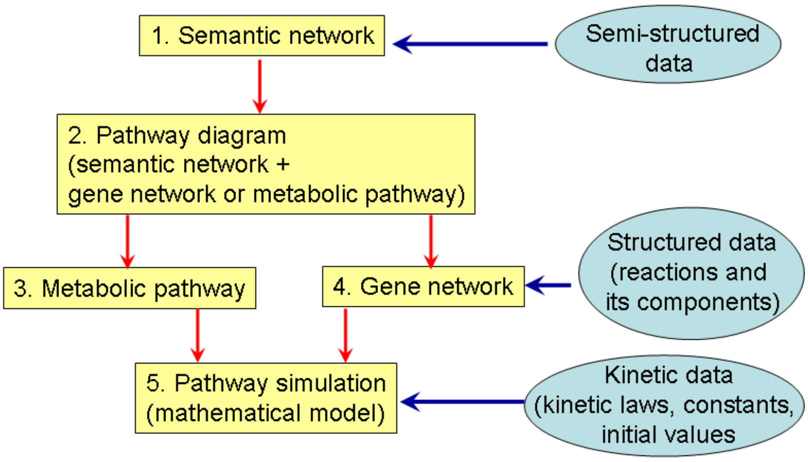
When the [Run] button is pressed, the workflow will start. If, for instance, the folder *test* with some tables is defined as input, and the result folder name is *test result GO*, we get after completion of the workflow:



23. Editing and creating diagrams

23.1. Diagram types

Diagram types are hierarchically organized as shown in the following scheme:


1. Semantic network
Semi-structured data
2. Pathway diagram (semantic network + gene network or metabolic pathway)
3. Metabolic pathway
4. Gene network
Structured data (reactions and its components)
5. Pathway simulation (mathematical model)
Kinetic data (kinetic laws, constants, initial values

There are 5 types of diagrams in the geneXplain platform:

- *Pathway diagram*
- *Pathway simulation*
- *Composite diagram*
- *Pathway simulation (SBGN)*
- *SBML simulation (SBGN)*

You find them listed for selection when pressing the button to create a new diagram. This button appears on top of the Tree Area when you go into the Data folder of your project directory (usually under your user name).

The *Pathway type* diagram is used for formalized description of biological pathway structure (metabolic pathway, gene network, etc).

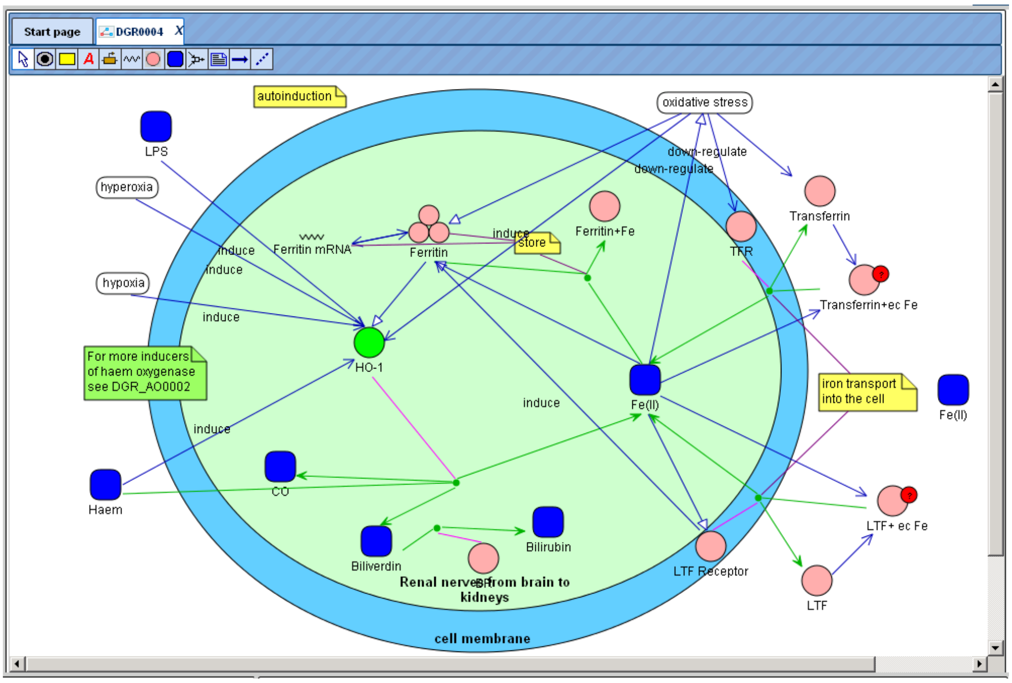
The *Pathway simulation* type diagram is an extension of pathway type, where variables are associated with graph nodes and differential equations with graph edges. This allows automatic generation of the mathematical model of the system and simulation of its dynamics.

The *Composite diagram* may contain several pathway simulation diagrams as well as some types of links between them to join separate simulation modules into one composite model.

The *Pathway simulation SBGN* diagram type is the same as pathway simulation, and visualized according to the rules of System Biology Graphic Notation.

SBML-SBGN is a specific type to wrap an SBML diagram in the SBGN view.

Please note that diagrams of Semantic type can’t be constructed under general user account.



Entity types and proteins are represented as follows:

Entity types

nucleus	Compartment
●	Cell
🟦	Substance
⊐	Gene
ᴡᴡ	RNA

Proteins

●	Active monomer
🟢	Inactive monomer
●●	Heterodimer
●●	Homodimer
●●●	Multimer
●	Phosphorylated protein

The following relations can be defined:

Reactions

—•→	binary
≻•→	complex

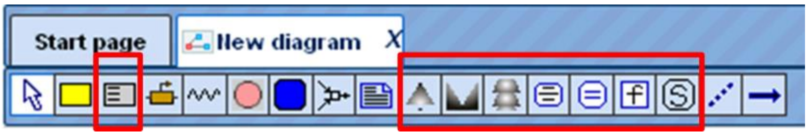
Interactions

—•→	direct
----•---→	indirect

Regulatory events

—•→	switch on
—•→	increase
—•→	decrease
—•→	switch off

When you choose the option "Pathway simulation", additional functions are available which you can see from the extended series of icons on top of the Work Space:



The seven new icons on the right (red frame) are specifically introduced for simulation functions, the one on the left replaces the icon with the capital A in the "Pathway diagram" option.

23.2. Creation of a new diagram

A new diagram can be made in the Work Space by two methods:

You can edit pre-composed diagrams and create a new diagram by saving it in your own project area (see below, 23.3).

You can start a new diagram in the work space by using graph search. The details of this method are further described below.

To start with the creation of a new diagram, please select the correct folder in your project area and press the icon in the Control Panel. The tool will ask you to select the type of diagram you want to create in the work space. In the following, the creation of a new "Pathway diagram" will be described.

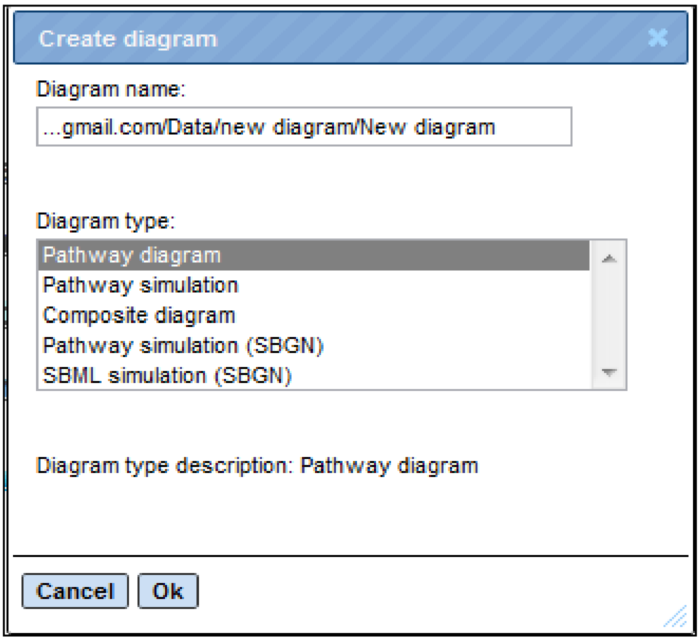

Create diagram

Diagram name:

...gmail.com/Data/new diagram/New diagram

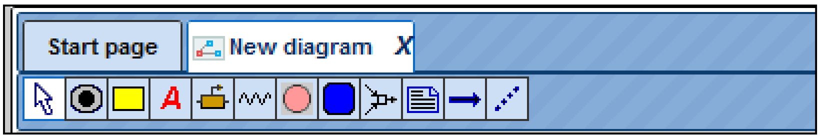
Diagram type:

Pathway diagram
Pathway simulation
Composite diagram
Pathway simulation (SBGN)
SBML simulation (SBGN)

Diagram type description: Pathway diagram

The description for diagram types has been mentioned above. Upon selecting the type of a diagram and clicking [Ok] a new node *New Diagram* will appear in the Tree and the corresponding tab is opened in the Work Space to start the diagram creation.

A pathway diagram opened in the work space has a tool bar as shown below:



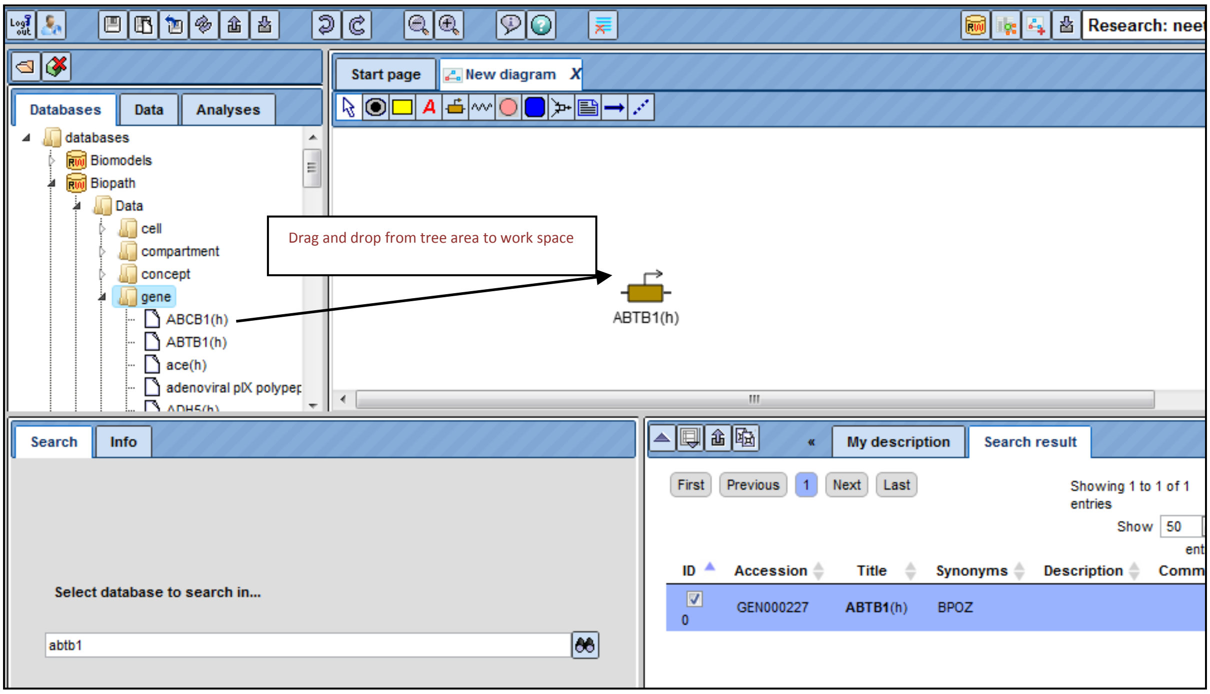
Using these icons, you can manually add a node that represents a cell (◉), a cellular compartment (🟨), a gene (⊟), an RNA (〰), a protein (🔴) or a substance (or small metabolite, 🟦). It may also something as abstract as a concept (A). These nodes may be related to each other with any kind of link (→), or , more specifically and if you wish to model your network as a bipartite diagram, with a reaction (⊳).

In case you wish to link two already existing nodes by a new edge, click on the icon →, which will open a new mask where you can define the input and the output node. Just move the mouse cursor into the respective field and click on the node that should serve as input or output node, respectively. The corresponding name or accession number will appear in the field. The same works for linking a note (to be defined before with 📄) to a node by introducing an edge of the type ⋰ .

To add an element to the diagram, several ways are available:

i) Search for the element using the search tab in the databases folder (as described in the Section 2.1.3.2) and then add this element to the diagram using the icon ⏏.

ii) Drag and drop the element from the Tree Area to the Work Space as shown below.



Drag and drop from tree area to work space

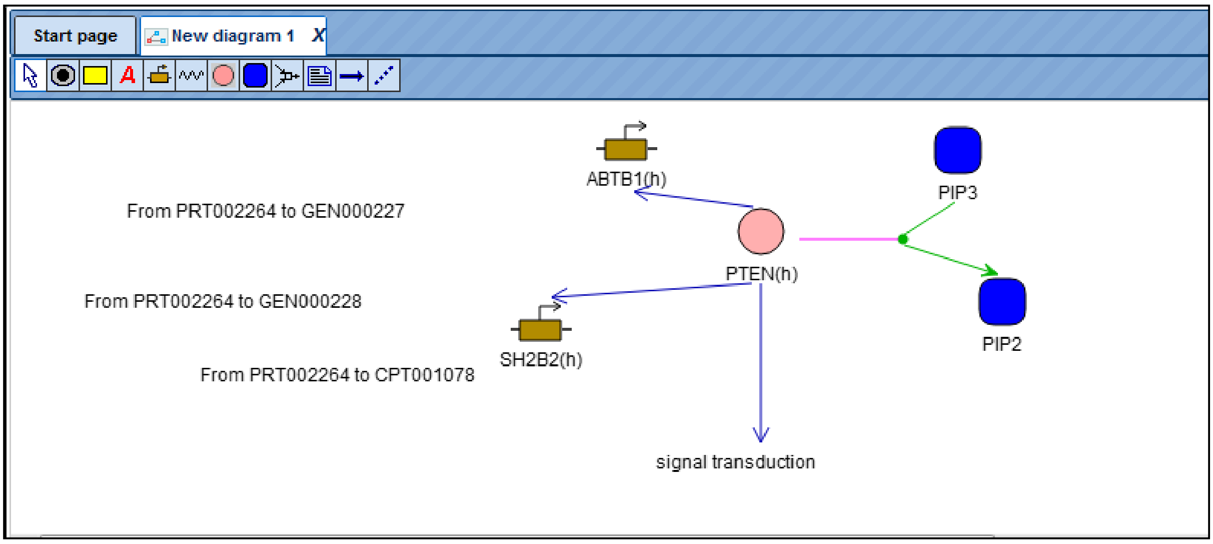
abtb1

ID	Accession	Title	Synonyms	Description	Comm
☑ 0	GEN000227	ABTB1(h)	BPOZ		

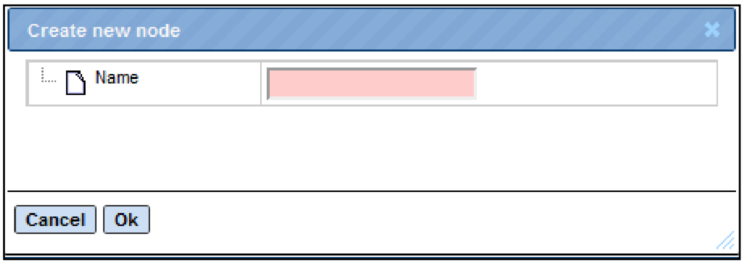
Please note that double clicking on the gene gives the information about that gene in the Info box.

iii) You can then use the Graph search option as described in Section 21.4 to further search for other elements upstream or downstream of those that are already included in the diagram. You can add elements to the new diagram as per your requirement as shown below.

You can specify a layout for the extended diagram, details of the layouts are described under "Changing the layout".

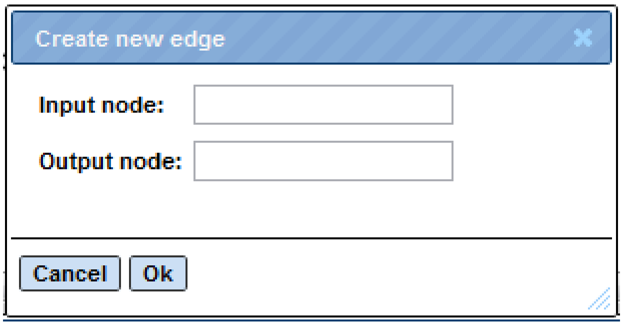

Start page
New diagram 1 X
ABTB1(h)
PIP3
From PRT002264 to GEN000227
PTEN(h)
From PRT002264 to GEN000228
PIP2
SH2B2(h)
From PRT002264 to CPT001078
signal transduction

iv) To create an entirely new node on the same diagram, click on the icon on the pathway diagram menu bar and click on the diagram where you want to add a node. You will get a message as shown below:

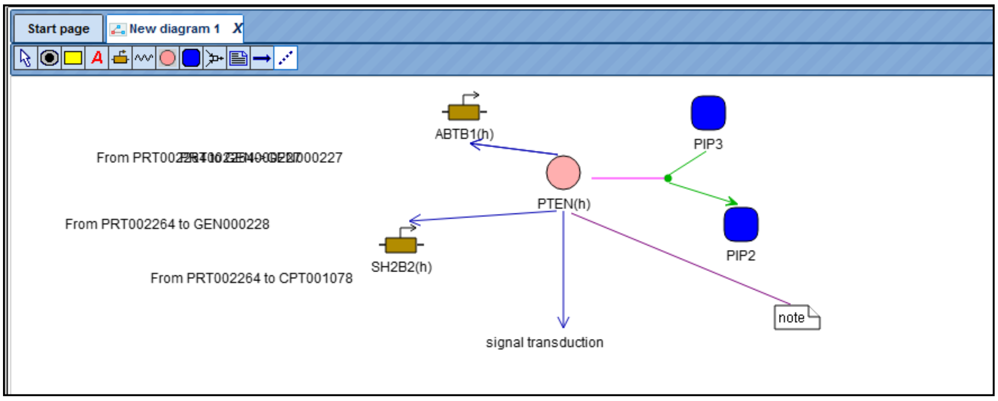


You can add the node/Name and press [Ok]; the node will be added to the diagram.

You can also link this node to another element by clicking on the icon in the pathway diagram menu bar. The tool asks the user to select input and output nodes.

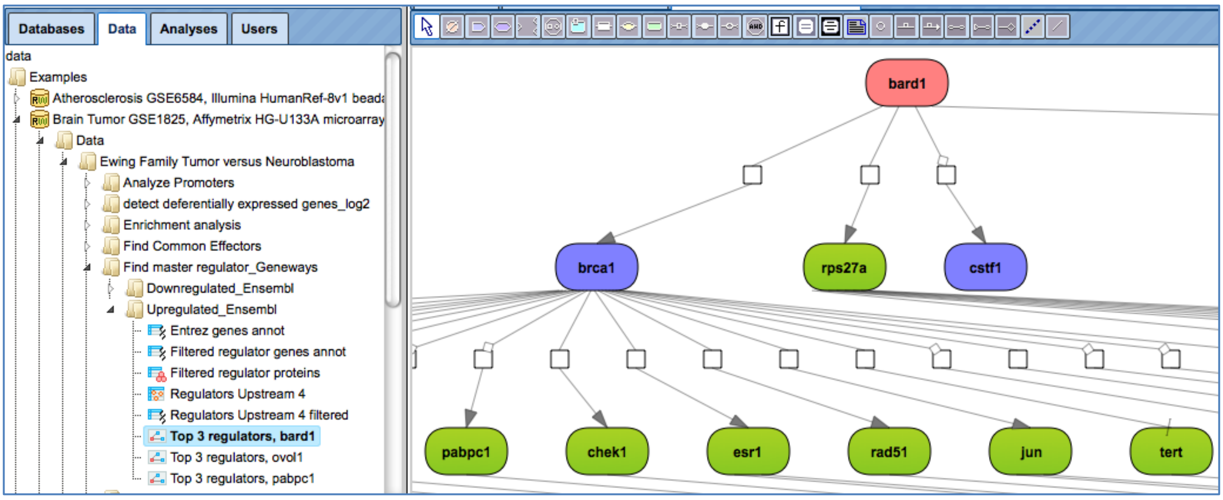


You can select input and output nodes easily by clicking first on the element you want to select as input node, and second on the element you want to be the output node. The diagram after adding the node and linking it to a protein looks like shown below.


Start page
New diagram 1
ABTB1(h)
PIP3
From PRT002264 to GEN000227
PTEN(h)
From PRT002264 to GEN000228
SH2B2(h)
PIP2
From PRT002264 to CPT001078
note
signal transduction

23.3. Editing a pre-composed diagram

Diagrams saved in the tree area can be edited in several ways. For this, please first copy one of the diagrams, e.g. located in the Examples folder, in your own project area. For this, you open the diagram in work space by double-clicking on its name. On the picture below, the diagram which is opened in the work space, highlighted blue.


Databases
Data
Analyses
Users
data
Examples
Atherosclerosis GSE6584, Illumina HumanRef-8v1 bead
Brain Tumor GSE1825, Affymetrix HG-U133A microarray
Data
Ewing Family Tumor versus Neuroblastoma
Analyze Promoters
detect deferentially expressed genes_log2
Enrichment analysis
Find Common Effectors
Find master regulator_Geneways
Downregulated_Ensembl
Upregulated_Ensembl
Entrez genes annot
Filtered regulator genes annot
Filtered regulator proteins
Regulators Upstream 4
Regulators Upstream 4 filtered
Top 3 regulators, bard1
Top 3 regulators, ovol1
Top 3 regulators, pabpc1
bard1
brca1
rps27a
cstf1
pabpc1
chek1
esr1
rad51
jun
tert

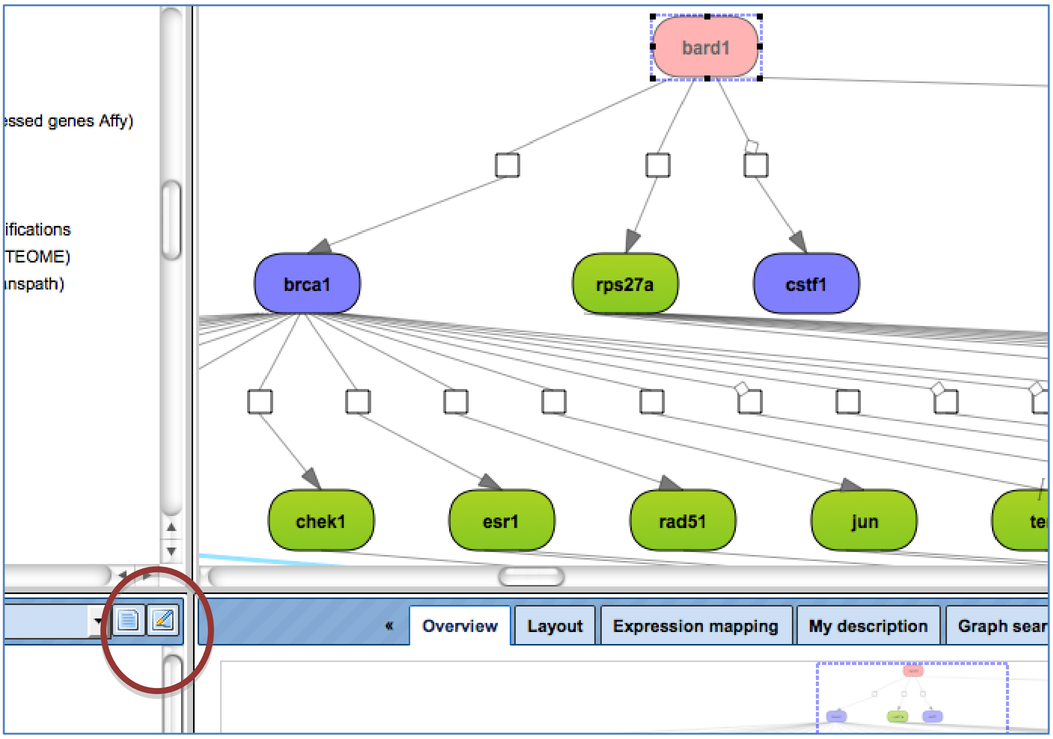
Then choose the option "Save as" (button in the Control Panel). A new mask opens ("Save document as"), where you will find your project and a folder labeled "data"; you should save this copy under any new name.

Changing location, color and title of the nodes on a diagram

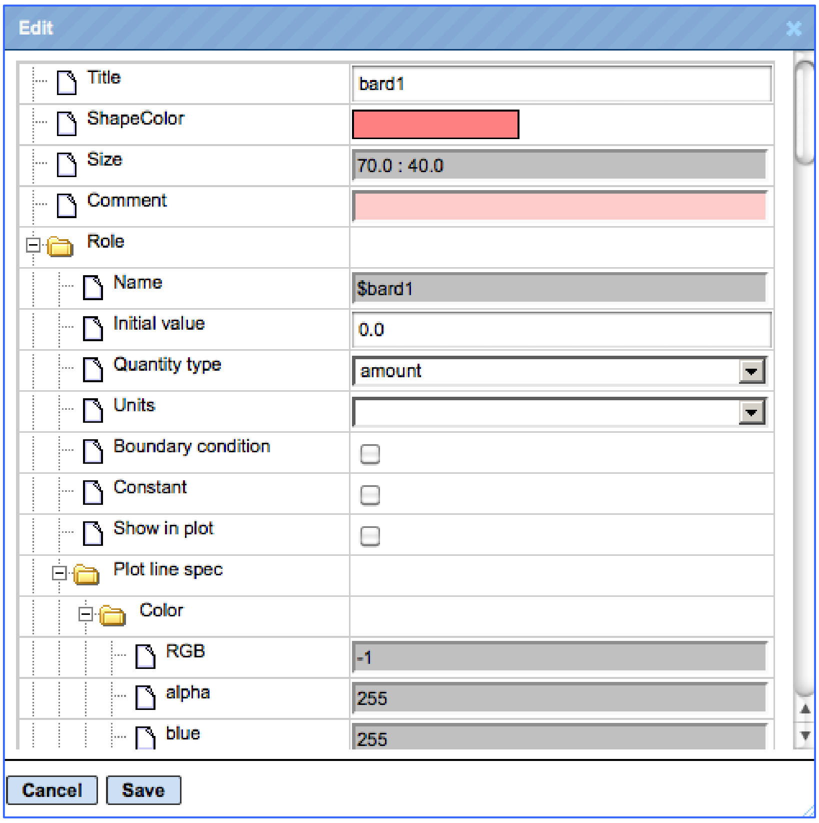
When you open a diagram saved in your project, you can change color and edit the title of the nodes, shift one node relative to others, remove or add nodes. Adding new nodes to a diagram is described in the previous section, Section 23.2.

To edit a node, first, select it by a mouse click. Below, the molecule bard1 is selected.

Apply the Edit button at the top right corner of the Info box to open the form for editing the selected node. The Edit button is highlighted by a red oval on the picture below.



The Edit form is opened as shown below.

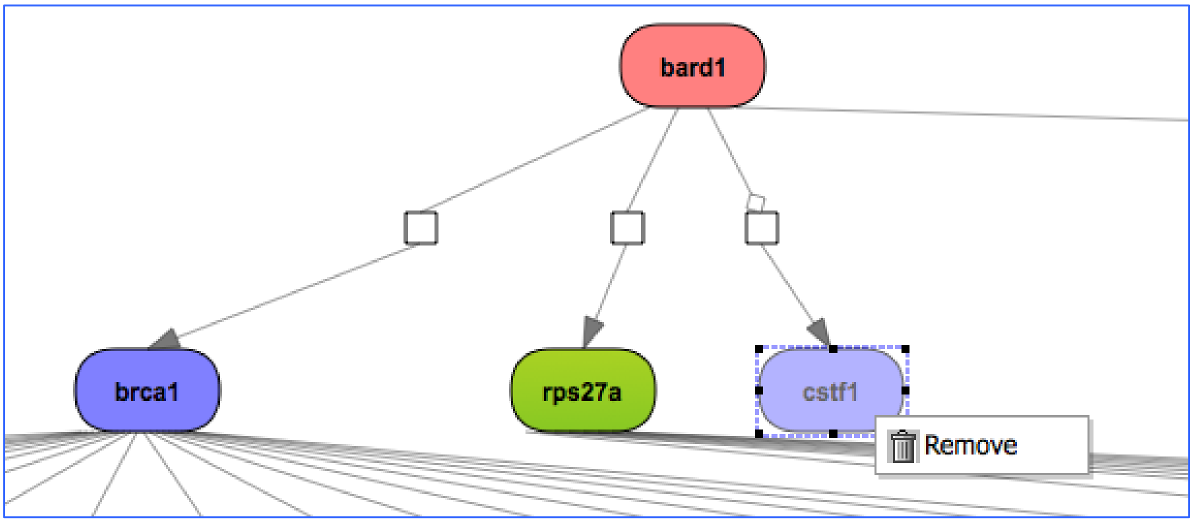


Title	bard1
ShapeColor	🟥
Size	70.0 : 40.0
Comment	
Role	
Name	$bard1
Initial value	0.0
Quantity type	amount
Units	
Boundary condition	□
Constant	□
Show in plot	□
Plot line spec	
Color	
RGB	-1
alpha	255
blue	255

In the Title field you can edit the name of the selected node. In the ShapeColor field the current color of the selected node is shown. By clicking into this field you can open the Select color form and change the color.

After editing is done, press Save on the Edit form.

To remove a node from the diagram, first select this node and then right-click (see figure below).



To shift a node relative to other nodes, select it and drag to a desired position.

Saving changes and undo & redo functions.

After editing of a diagram is complete, you can either save changes or undo them. Saving a diagram under the same path and name is done by the button on the top menu. If you would like to save a diagram to a different folder or project, or under different name, use the Save button .

In case you would like to return to the previous variant, use the Undo button .

Important: undo and redo functions work for one previous step, you need to click the same button again to undo or redo the next previous step. Please note, undo and redo functions work only before saving.

To return to the previously saved version of a diagram, use the Revert button .

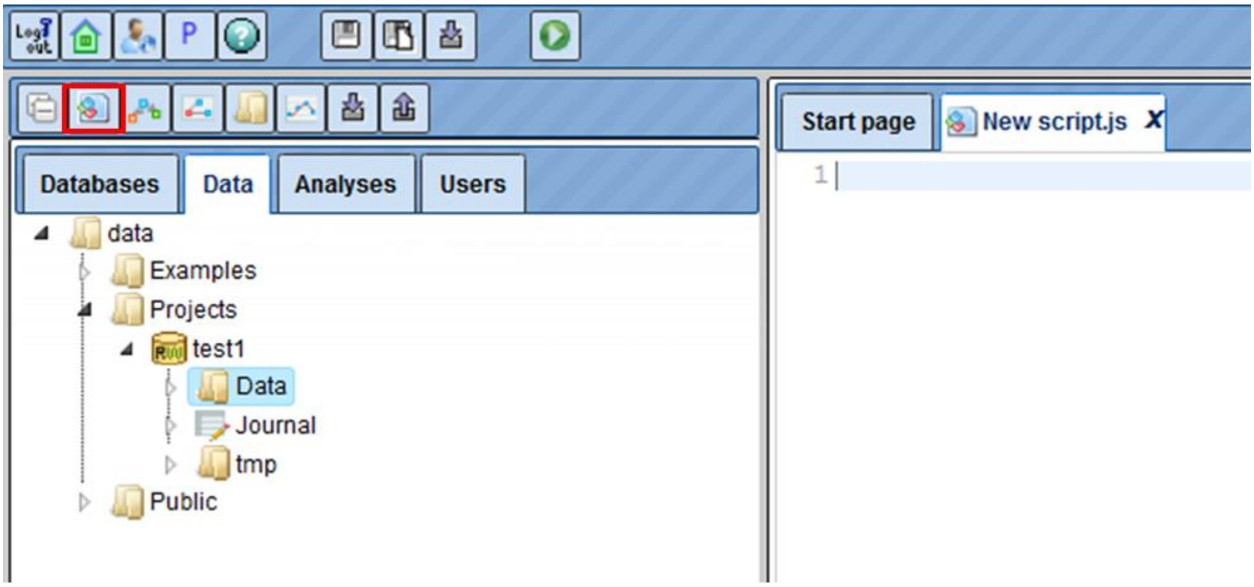
Further editing of diagrams includes mapping available expression data on the molecules and adding interactions around the selected molecules. For details on these two functions please refer to the sections 21.3 "Expression mapping", and 21.4, "Graph search".

24. Editing and creating JS scripts

User-specific scripts in JavaScript (JS) can be added directly into the platform, and immediately executed. They can be combined with pre-existing analyses and can be part of the workflows; existing codes in workflows can be edited.

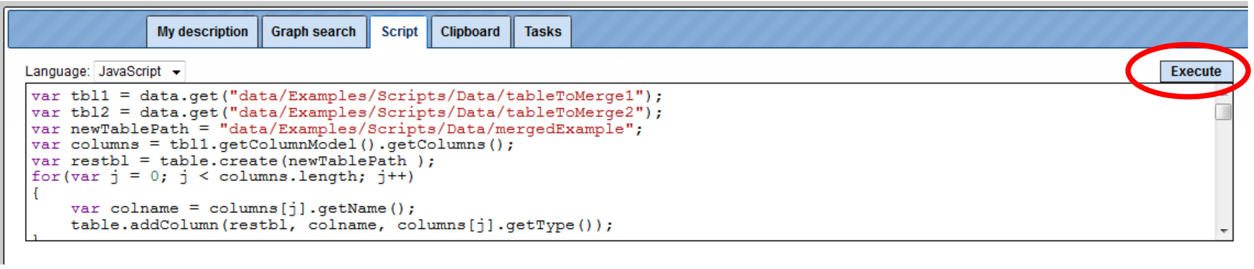
24.1. Creating new JS scripts

For writing a new JS script, go to any of your data folder (blue marked) and click on the "New JS script" button () in the toolbar. After pressing a new tab "New script.js**"** opens in the Workspace. You can write your script and save it for later execution or incorporation into a workflow.



24.2. Executing JS scripts

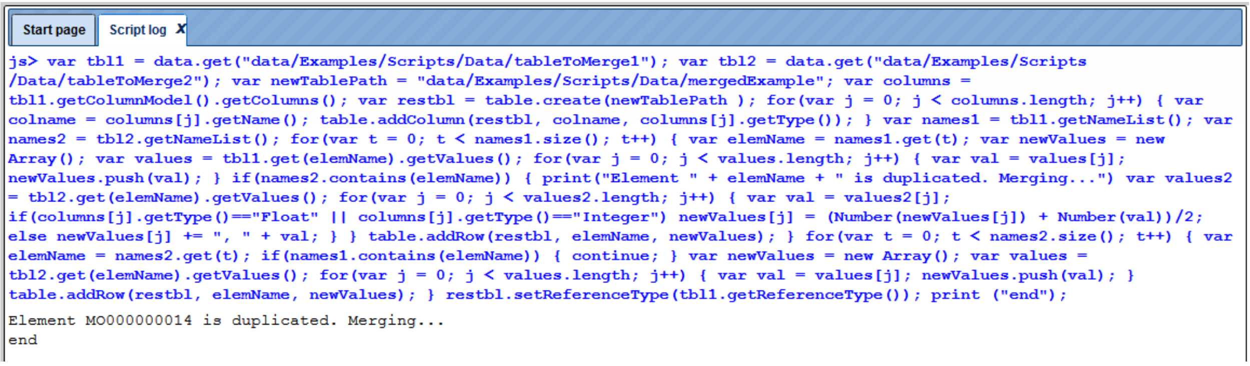
To execute a JS code directly, click the "Script" tab in the Operation Field. Simply write or copy and paste your script code in the box. To run the script, press the button [Execute].



Language: JavaScript

```
var tbl1 = data.get("data/Examples/Scripts/Data/tableToMerge1");
var tbl2 = data.get("data/Examples/Scripts/Data/tableToMerge2");
var newTablePath = "data/Examples/Scripts/Data/mergedExample";
var columns = tbl1.getColumnModel().getColumns();
var restbl = table.create(newTablePath );
for(var j = 0; j < columns.length; j++)
{
    var colname = columns[j].getName();
    table.addColumn(restbl, colname, columns[j].getType());
}
```

After pressing [Execute] the new tab "Script log" opens in the Workspace. Here you can find information about the success and the output of the script.



```
js> var tbl1 = data.get("data/Examples/Scripts/Data/tableToMerge1"); var tbl2 = data.get("data/Examples/Scripts
/Data/tableToMerge2"); var newTablePath = "data/Examples/Scripts/Data/mergedExample"; var columns =
tbl1.getColumnModel().getColumns(); var restbl = table.create(newTablePath ); for(var j = 0; j < columns.length; j++) { var
colname = columns[j].getName(); table.addColumn(restbl, colname, columns[j].getType()); } var names1 = tbl1.getNameList(); var
names2 = tbl2.getNameList(); for(var t = 0; t < names1.size(); t++) { var elemName = names1.get(t); var newValues = new
Array(); var values = tbl1.get(elemName).getValues(); for(var j = 0; j < values.length; j++) { var val = values[j];
newValues.push(val); } if(names2.contains(elemName)) { print("Element " + elemName + " is duplicated. Merging...") var values2
= tbl2.get(elemName).getValues(); for(var j = 0; j < values2.length; j++) { var val = values2[j];
if(columns[j].getType()=="Float" || columns[j].getType()=="Integer") newValues[j] = (Number(newValues[j]) + Number(val))/2;
else newValues[j] += ", " + val; } } table.addRow(restbl, elemName, newValues); } for(var t = 0; t < names2.size(); t++) { var
elemName = names2.get(t); if(names1.contains(elemName)) { continue; } var newValues = new Array(); var values =
tbl2.get(elemName).getValues(); for(var j = 0; j < values.length; j++) { var val = values[j]; newValues.push(val); }
table.addRow(restbl, elemName, newValues); } restbl.setReferenceType(tbl1.getReferenceType()); print ("end");
Element MO000000014 is duplicated. Merging...
end
```

To familiarize yourself with the handling of JavaScript, you may use the following example script:

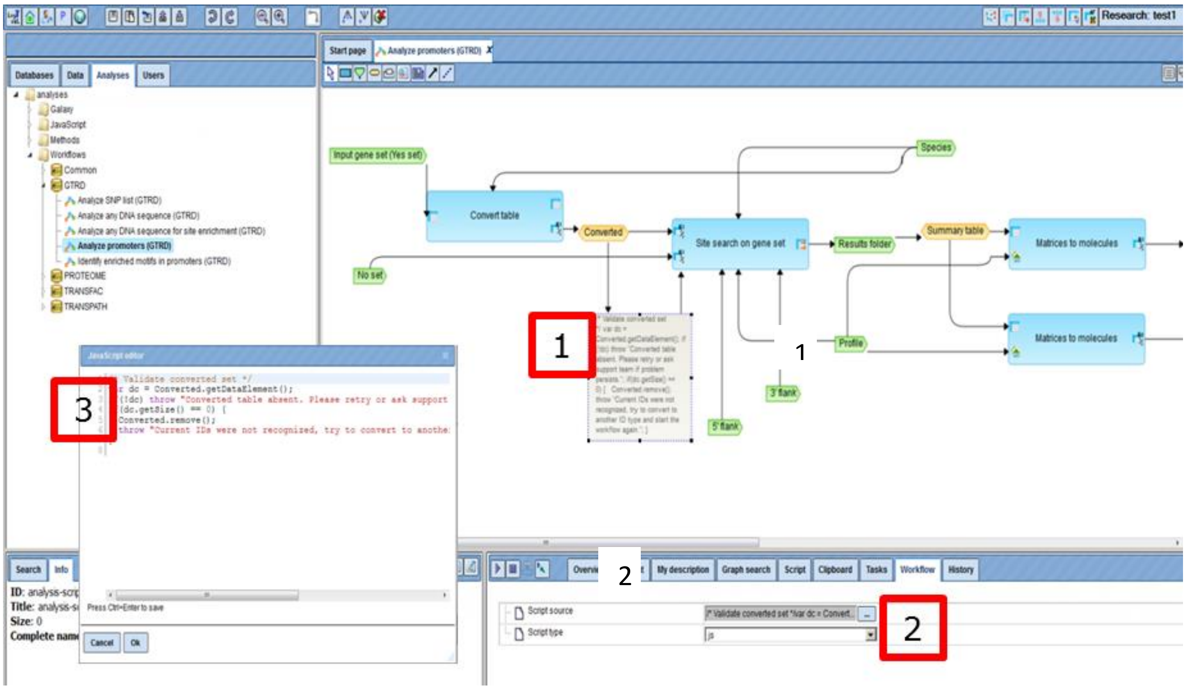
http://platform.genexplain.com/bioumlweb/#de=data/Examples/Scripts/Data/mergeTables.js

24.3. Editing JS scripts in workflows

To edit a JavaScript code in an existing workflow, you have to open the workflow in Edit mode and copy it in your data folder. An example is given by the following workflow:

http://platform.genexplain.com/bioumlweb/#de=analyses/Workflows/GTRD/Analyze%20promoters%20%28GTRD%29

To change the JavaScript code, the grey **ScriptBox** must be clicked in the workflow overview (1), whereupon the **Script source** appears in the "Workflow" tab of the Operation field (2). Press the button [...] to open the **JavaScript editor** (3). In this window, the code can now be changed directly. The confirmation of the editing is done using the [Ok] button.

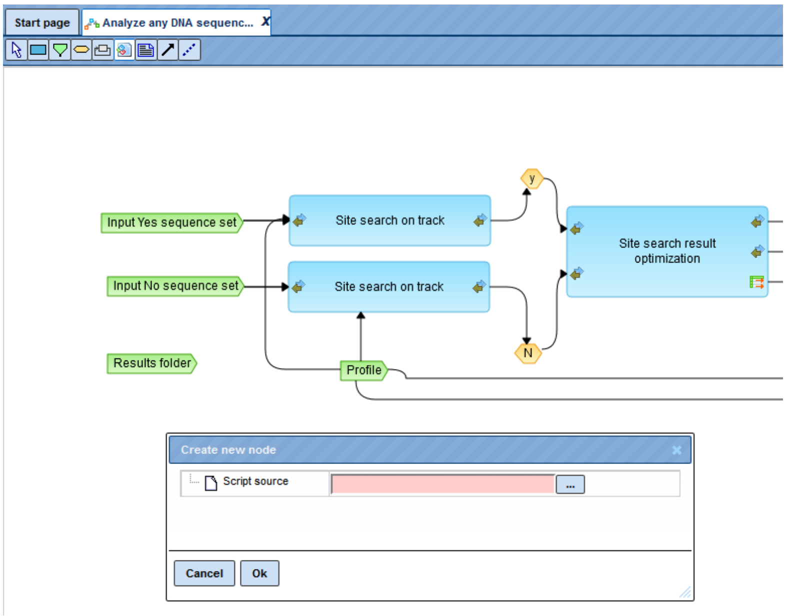


```
/* Validate converted set */
r dc = Converted.getDataElement();
(!dc) throw "Converted table absent. Please retry or ask support
(dc.getSize() == 0) {
Converted.remove();
throw "Current IDs were not recognized, try to convert to anothe
```

Script source: /* Validate converted set */var dc = Convert...
Script type: js

24.4. Inserting JS scripts in workflows

To insert a script in an existing or self-created workflow, press the “Analysis –script” button () in the toolbar. After clicking on button [...] of the **Script source**, the **JavaScript editor** opens. The script can be inserted.



25. List of icons and their meaning

25.1. General (Control Panel)

Icon	Function
Log out	Log-out
⌂	Home
👤	Account info
P	Project Properties
💾	Save
💾	Save as, applied to the element (table or diagram) opened in the Work Space
↶	Revert to the previously stored version, to discard the changes introduced and revert to the default diagram
📥	Import a file
📤	Export the element (table or diagram) opened in the Work Space
↶	Undo previous action
↷	Redo previous action
🔍+	Zoom in
🔍−	Zoom out
i	Information about geneXplain platform
?	Help
✕	Remove selected rows from a table

25.2. Tree Area Panel

25.2.1. General

Note that the appearance of these icons on top of the directories in the Tree Area is context-sensitive.

⊟	Expand / collapse directory
📜	Invoke a script
⚙	Compose a new workflow
▨	Compose a new diagram

	Create a new directory
	Start an parameter optimization process
	Import a file into the selected (sub-)directory
	Export the element selected in the (sub-)directory of the Tree Area
	Remove
	Open the selected file (highlighted by a light-blue background in the Tree Area)

25.2.2. Database types

Icon	Function
R	Protected database with read access enabled
R	Public database with read access
RW	Database with read & write access
RW	Database with read access, write access enabled
RW	Database with read access, write access disabled

25.2.3. Data types

Icon	Function
	Table (general)
	Table of probes and properties of their signals
AF	Table of probes with Affymetrix IDs
AG	Table of probes with Agilent IDs
Il	Table of probes with Illumina IDs
	Table of genes
EI	Table of genes with Ensembl IDs
Ez	Table of genes with Entrez IDs
Gb	Table of genes with GeneBank IDs
G	Table of genes with Gene Symbol as IDs

RS	Table of genes with RefSeq IDs
U	Table of genes with UniGene IDs
Il	Table of genes with Illumina IDs
P	Table of genes with Proteome IDs
Tp	Table of genes with Transpath IDs
	Table of proteins (general)
R	Table of proteins (Reactome database)
TP	Table of proteins (TRANSPATH database)
El	Table of proteins (Ensembl)
Gb	Table of proteins (GeneBank)
I	Table of proteins (IPI)
RS	Table of proteins (RefSeq)
TF	Table of proteins (TRANSFAC)
U	Table of proteins (UniProt)
A-C	Table of SNPs
	Table of transcripts
El	Transcripts Ensembl
A C G T	DNA sequences in EMBL, FASTA, GenBank format
	Workflow
	Diagram, with nodes and (optionally) edges as components
🟨	Compartment, with nodes and edges hierarchically assigned
	Node in a diagram
🔴	Protein
	Gene
	Edge (relation) in a diagram
	Plot
	Reaction
	Set of genomic intervals (Track)

25.2.4. Types of analyses

Icon	Function
Data manipulation	
	Annotate a table (add additional columns)
	Annotate track with genes
	Apply State to Diagram
	Composite modules to Proteins
	Convert table identifiers using BioHub(s)
	Convert table to track
	Create a folder
	Create Random track
	Filter one track by another
	Filter table
	Filter track by condition
	Gene set to track
	Intersect / join tables
	Intersect Tracks
	Join several Tables
	Join Tracks
	Matrices to molecules
	Process track with Sites
	SNP matching
	Split table by columns
	Track to gene sets
	Venn diagram
Data normalization	
	Normalize expression data
	Normalize data and split between experiment and control
	Normalization Quality Plots
Functional classification	
	Gene Set Enrichment Analysis (GSEA)
	Functional classification

Import	
	Import a file
Molecular networks	
	Add expression values to a network / diagram
	Cluster by shortest path/Join Diagrams
	Search for a common effector
	Extend network to all reachable nodes
	Search for a common regulator
	Save hits
	Save network
	Visualize results as diagram
Optimization	
	Parameter optimization
Sequence manipulation	
	Bowtie
	Chip-Seq peak Profile
	Model-based analysis of ChiP-Seq data (MACS)
	ChipHorde/DiChipHorde
	ChipMunk/DiChipMunk
	Mutation Effect
Site analysis	
	Create Profile cutoffs
	Construct IPS cisModule
	Construct Composite Modules
	Create IPS model
	Create Match model
	Create profile from gene table
	Create profile from the table of site models (matrices)
	Create profile from matrix library
	Create weight matrix model
	IPS motif discovery

	MEALR(tracks)
	Motif quality analysis
	Search for enriched TFBSs (genes)
	Seach for enriched TFBS (tracks)
	Site search on gene set
	Site search on track
	Site search result optimization
Statistical analysis	
	K-means cluster analysis
	Chinese restaurant cluster analysis (CRC)
	Correlation analysis
	EBarrays
	Fold change calculation
	Hypergeometric analysis
	LIMMA
	Meta-analysis
	PCA
	Polynomial regression analysis
	Identification of up- and down-regulated genes

25.3. Info Box

📄	View information in a separate window
✎	Edit description on the Info tab (only for own files; editing possible in pink fields)
🔭	Launch search for the entered term in the selected database

25.4. Operations Field

▶	Execute (e.g., prepare layout as defined)
▲	Apply layout as generated in the Operations Field to the diagram in the Work Space
■	Stop process (e.g., stop preparing new layout)
▼	Accept layout edited in the Work Space to the Operations Field
✚	Add new (e.g., new expression mapping)
✖	Remove
✎	Edit description
💾	Save
🔭	Launch graph search
🗑	Discard changes / clear elements
■	Stop Task

25.5. Work Space

25.5.1. Diagram editing

◉	Insert a node representing a cell
🟨	Insert a node representing a cell compartment
A	Insert a node representing a concept
⊥	Insert a node representing a gene
∿∿	Insert a node representing an RNA
🔴	Insert a node representing a protein
🟦	Insert a node representing a substance
⋟	Define a reaction
🗎	Add a note
➡	Add an edge representing a relation
⋰	Add an edge to a note