The toolbox of TRANSFAC 2.0

User guide release 1.0

Table of contents

Getting started

Welcome to MATCH Suite

Welcome to MATCH Suite – the toolbox of TRANSFAC 2.0, which comprehensively addresses the syntax and semantics of gene regulation and allows you to identify the transcription factors regulating the gene set of your interest.

In its first release the MATCH Suite 1.0 is designed to work with an input list of human genes to identify the activation patterns of these genes on the basis of integrated FMATCH and CMA analyses (see Methods for further details).

Explore the functional enrichment of your gene set, find only transcription factors (TFs) expressed in the tissue of your interest, filter the predicted TFs by those intersecting with enhancers or conservative regions of the genome and much more! Navigate to *Start analysis* to immediately launch the comprehensive and fully automatized study of the gene set of your interest or use the demo input to explore the abilities of the MATCH Suite 1.0.

Basic interface

The MATCH Suite interface is based on two main sections:

(1) **Start analysis** – navigate to this section to launch a new analysis
(2) **Launch history** – navigate to this section to view the history of your previous launches and open their results
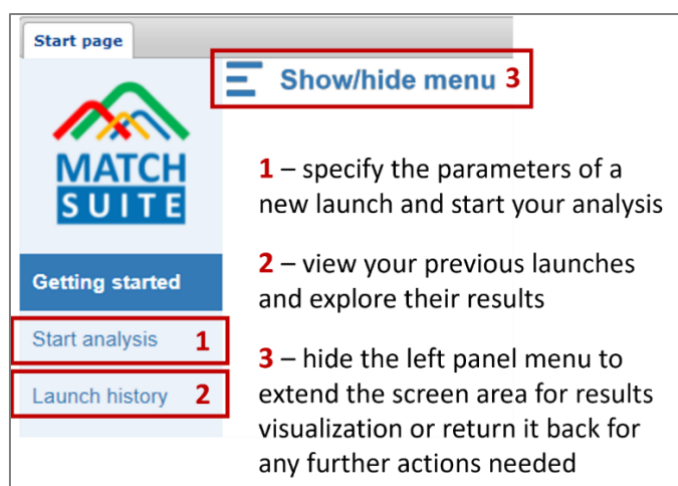
The **Show/hide** menu button at the top of the screen (3) allows you to hide the left panel menu for extending the useful screen area when this is needed (this function is of particular importance when visualizing the analysis results).

Launching the analysis

Selecting the input gene list

Composing an input gene list in the TRANSFAC® database

A particularly convenient way to compose a gene list for analysis is to run the MATCH Suite on your search results in the TRANSFAC® database. For instance, when you search for a certain disease in the standard Search field or select genes belonging to a certain Gene Ontology (GO) category (Search > Ontology search), you end up with a list of genes.
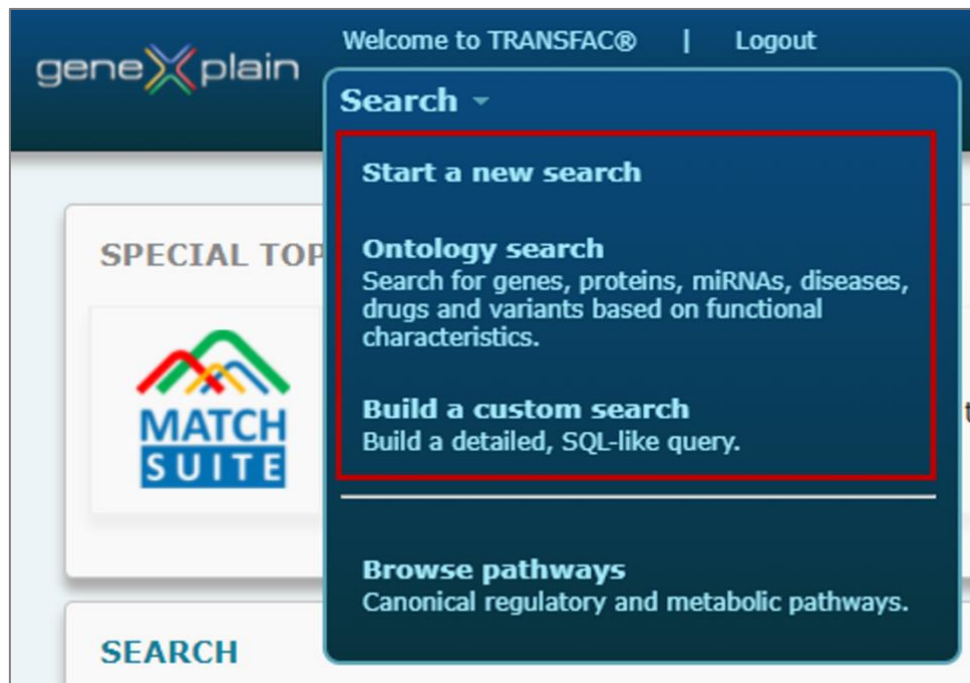
In the search result inside TRANSFAC® database you can view all found entries on one page (select "All" in the "Hits on page" list) and then select the search results of your interest to launch the MATCH Suite on a respective gene list (you can use

the *Mark all on page* option and then only human genes from your selection will be automatically taken to the MATCH Suite analysis upon launching it).

When you have selected the search results of your interest, press the button "Match Suite" and you will be automatically transferred to the MATCH Suite tool, where you will be asked to assign a name to your gene list prior to proceeding to the next step of the analysis launch wizard:

After giving a name to your gene list, the wizard will immediately take you to the *Specifying the launch parameters* step of the MATCH Suite analysis launch wizard.

Selecting input gene list from the MATCH Suite interface

When you click on the *Start analysis* section, the system immediately navigates you to the analysis launch wizard. At the very first step of this wizard, you will be asked to select the gene list which should be used for the analysis.

You can try launching the analysis using the **Use demo gene list** (1) option, or you can specify your own gene list by selecting the **Use my data** (2). The demo gene list contains 50 Ensembl genes, you can view them by clicking on the *demo gene list* blue link. The **Supported input formats** (3) link also is clickable, it provides you with the information about the requested format of the gene list, which can be analyzed with the MATCH Suite 1.0:

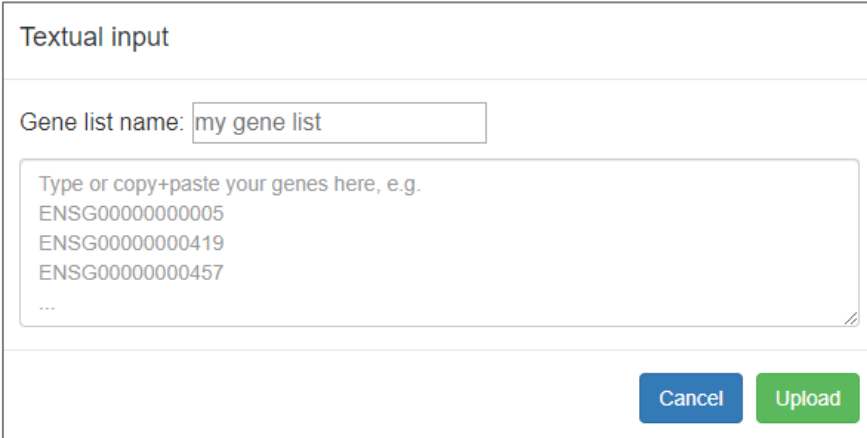Lists of 20 to 1000 genes coming in Ensembl ID, Entrez ID or Gene symbols format are accepted by the MATCH Suite 1.0. The system will select automatically the human genes from your list, only these will be subject to the further analysis.

The gene list for your analysis can be uploaded from your local computer using the **Upload files** (4) option, it can be also selected from the data stored in any geneXplain platform project accessible to you by using the **Browse stored data** (5) option. In such case the selected gene list will be copied to the MATCH Suite project and taken for further analysis. The next option for input gene list specification is **Textual input** (6). This function provides you with an ability to simply copy and paste any gene list of your interest (in Ensembl, Entrez or Gene symbols format) to the dedicated textual input form:

The last option to specify an input gene list is to select it from the gene lists that were previously used by you as inputs for the MATCH Suite analysis (7). The names of the previously used gene lists are clickable and the respective gene list will open upon the click on its name (8). You can manage the gene lists stored in your project and permanently delete the unnecessary ones using the **Delete** option (9). Please note that the maximum size of your MATCH Suite project is 2 GB. You can free up space by deleting the unnecessary gene lists and analysis results if that would be needed (see *Operating in the Launch history* section below for further info)

To select the gene list for the current analysis run, mark it with the radio button in the list of all currently available inputs (7) (any newly uploaded gene list will automatically appear in this list).

Please note that regardless of the input gene list source, the selected gene list will be checked by the MATCH Suite for correspondence to the input gene lists requirements and only human gene lists containing from 20 to 1000 human genes in supported formats will be accepted for further analysis.

TRANSFAC 2.0

Specifying the launch parameters

In the next step of the analysis launch wizard the MATCH Suite will ask you to fill in the following form:

You can give a name to your analysis launch or keep the default name suggested by the system in the **Name of the analysis launch** field (1). This name will be further used in the launches history for easy selection of the run of your interest.

Optionally you can specify the tissue of your interest from the dropdown list of supported tissues in the **Studied tissue** field (2). The provided tissues are the tissues from FANTOM5 database with specific coordinates for the transcription start sites (TSSs) that allow to select tissue-specific promoters for the subsequent analysis. The selected tissue is also used further for identification of transcription factors that are known to be expressed in the tissue of your interest. Please refer to the Methods document for further details.

The promoter range that will be used for the search of TFBS in your analysis run is specified by the **Promoter range** parameter (3). By default the promoter range used by the MATCH Suite is [-500,100] relative to the TSS. You can specify a customized promoter range limited to the maximum of -5000 and +1000 from TSS.

Your input gene set can be optimized by functional enrichment using the **Gene set optimization** option (4). Having clicked on it, you will be offered to select between the supported GO categories:

Depending on your selection, the tree map visualization of functional classification of your input gene set will be built either on biological processes, or on cellular components, or on molecular functions. If you do not wish to optimize your input gene set by functional categories, simply do not tick the *Gene set optimization* option.

When done with parameters selection, click on **Next** (5) to proceed either to GO functional optimization of your input gene set or skip this step and start the analysis immediately by confirming the selected analysis launch parameters.

Optimizing the input gene set by GO functional classification (optional step)

In case you have selected to optimize your input gene list by GO functional classification at the parameters specification step, the tree map of GO biological processes, cellular components or molecular functions will be constructed for your input gene set depending on your selection.

You can click **Back** (1) to return to the parameters specification step and cancel the GO optimization of your input gene list by unticking the *Gene set optimization* option. Otherwise you are requested to select the GO categories of your interest, following the instructions provided at the top of the screen. Only genes belonging to the selected GO categories will be taken by the MATCH Suite for further analysis.

The **Select all** (2) button will allow you to select for further analysis all genes, belonging to all GO categories enriched in your input gene set (genes belonging to all GO categories visualized on the tree map). The **Clear selection** (3) button will deselect all previously selected GO categories and will reset the gene count to 0. The number you will see next to the **Selected gene count** (4) will show you the total amount of genes underlying the currently selected GO categories. Once this number will reach the minimum of 20 genes that are requested for launching the analysis, the **Proceed with selected genes** (5) button will become active and you will be able to launch the analysis using the optimized gene set.

Confirming the launch parameters and starting the analysis

After specifying all parameters for your launch, the MATCH Suite wizard will ask you to confirm your selection:
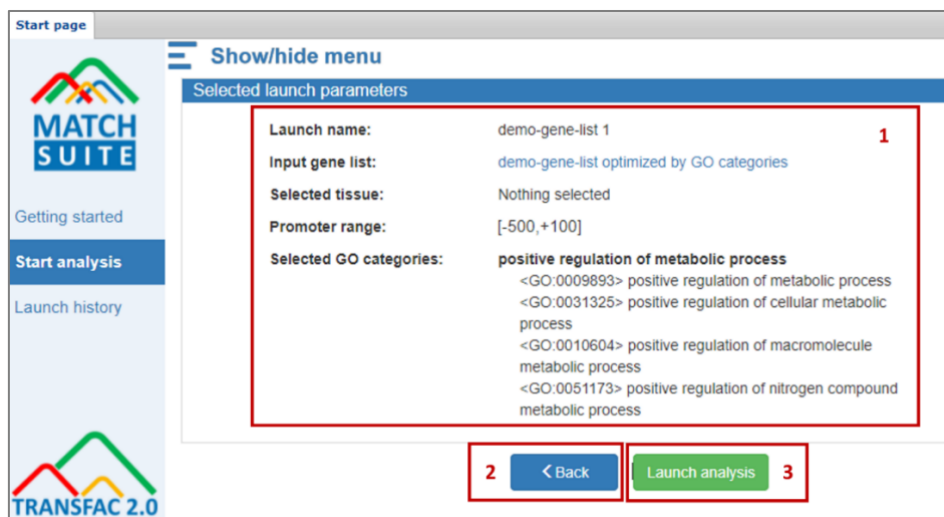
All parameters that you selected for the current launch will be shown on the screen (1). You can check the genes that were eventually selected for the analysis launch after GO functional classification (if it was applied) by clicking on the name of the input gene list.

If you want to apply any changes to the specified parameters of the launch, click on **Back** (2), otherwise you are ready to start your analysis by clicking on the **Launch analysis** button (3).

Viewing the results

Operating in the *Launch history*

Once your analysis was launched, you will be redirected to the *Launch history* section, also accessible by the direct link at the left menu panel. The Launch history allows you to view the results of all your previous analysis runs and to follow the progress of the currently running analyses.

You can check the parameters used for the launch by clicking on the **View input parameters** link (1). The following pop-up form will appear:

The gene set used for the current run can be viewed by clicking on the respective link in the *Input gene list* field. In case you have selected to optimize your input gene list by certain GO categories, you can view them by clicking on the *show* link

next to the *Categories selected to narrow the input gene list field*. If no GO functional categories optimization was done, the respective field will display *none*.

The progress of the currently running analysis will be displayed in the **Status** column (2). Please note that the progress is displayed in percentage of the finished steps of the underlying workflow and it has no direct correlation with the time left for the analysis to finish. Commonly one analysis run will take several hours, but this time interval is highly dependent on the input gene set and other parameters selected for the respective analysis launch.

It is not recommended to have several analysis launches running in parallel. Analysis runs will finish faster when launched consecutively one after another.

If you wish to terminate the launched analysis, you can click on the **Terminate** button (3).

To view the results of a finished analysis please click on the **View results** button (4).

You can manage the stored analysis results and delete the unnecessary data by clicking on the **Delete** icon (5). This action will delete all results of the respective analysis launch. The input gene list used for this analysis launch will still be accessible in the available list of inputs, which is viewed at the very first step of the start analysis wizard (see the *Selecting the input gene list* section of this document). By default, your MATCH Suite account is equipped with 2 GB disk space for storing your analysis results and input gene lists used. You can extend this volume by contacting us via info@genexplain.com with a respective request.

You can launch a new analysis run directly from the *Launch history* section by clicking on the **Start new analysis button** (6).

Results visualization

Having selected in the *Launch history* section the analysis result which you want to view, it will open in the results visualization mode on your screen.

In the results visualization mode the screen will be divided into four different segments with the following functions:

(1)   The **identified transcription factors** regulating your gene set and respective matrices tables

TRANSFAC 2.0

(2)   The **table of your input genes** and found site hits in their promoters

(3)   The **info box** displaying the information about the currently selected object

(4)   The **genome browser**, allowing to visualize the tracks of found sites and additional annotation tracks for further results interpretation

The Factor, Matrix and Gene tables that you see on the screen can be exported with the applied filters (see filtering instructions below) using the **Download tables button** (5). The archive containing these three tables in tab-separated text format will be downloaded to your local computer.

You can open the comprehensive analysis report about the respective run by clicking on the **Report** (6) link in the left menu panel. The self-explaining report will contain the factor, matrix and gene tables you see in the results visualization section along with supplementary tables and analysis steps description.

You can extend or narrow certain segments of the screen by moving the splitter lines. At this point you might want to use the *Show/hide menu* button, which will hide the left panel menu from view and will broaden the results area of the screen.

In the screen segment, which visualizes the tables of predicted factors and respective matrices, you will find three accessible tables: Factor view, Factor view Pro and Matrix view.

Factor view table

In the **Factor view** (1) table you will see a simplified summary of the transcription factors predicted to be regulating your input gene set. By default top 10 factors will be shown, this can be changed in **Show <number> entries** (2) field. The total number of factors identified will be displayed in (3). You can navigate through the predicted factors using the pages in (4). The columns of the *Factor view* table are fully matching the *Table 1* given in the analysis report of the respective run. You can refer to the analysis report for denominations of column names and their contents or use the info hints provided in the results visualization interface as a mouseover message that will appear upon hovering above the **?** sign (5) available at multiple places in the interface.

The columns of tables have in-built sorting option (6) which allows to sort the values within one column by ascending or descending order or by alphabetical order in case of textual contents. Simply click on the gray arrows for the sorting to be applied. By default, best factors (or matrices) are brought to the top (please see the analysis report and the [Methods](#) document for explanations on the factors and matrices ranking procedures). If you want to return to the original order of factors, click on the **Best factors on top** button (7).

Factor view Pro table

The **Factor view Pro** table (1) provides a deeper look into the transcription factors identified to be regulating the input gene set. It fully corresponds to the *Table 3* of the analysis report, where you can find the denominations of its column names and their contents description. Respective info is also summarized in the **?** hints available as mouseover messages upon navigating on them, similar to the *Factor view* table.

The *Factor view Pro* table allows you to apply the **Tissue** filter (2). You can select the tissues(s) of your interest from the dropdown list in (2) and click on the **Apply changes** button (4) to recalculate the factors and genes tables, as well as the track of found sites in the genome browser, leaving only those factors and their sites, which are known to be expressed in the tissue(s) of your selection. Respective columns with expression values of factors in the selected tissue(s) will be added to the table. You can refer to the Methods document for information on the expression values origin. To remove the filtering of factors by selected tissues, click on the **Remove all filters** button (3).

Please note that the tissue expression filter will be also auto applied to all tables and tracks when you will switch from the Factor view Pro table to either of the Factor view or Matrix view tables. To cancel the filters, click on the *Remove all filters* button.

Matrix view table

The **Matrix view** table (1) shows the PWMs (positional weight matrices) of the TRANSFAC® library, the respective sites of which were identified in the promoters of the studied gene set.

Similar to the factors table, the *Best matrices on top* button allows to bring the best matrices to the top of the matrix table in case their order was changed while sorting the values in individual columns. This table fully corresponds to the *Table 4* of the analysis report, where you can find the denominations of its column names and their contents description. Respective info is also summarized in the **?** hints available as mouseover messages upon hovering above them, similar to the *Factor view* and *Factor view Pro* tables.

The **Adjusted site enrichment filter** (2) allows to leave only matrices with the adjusted site enrichment values higher than the threshold specified by the filter. The set filter will be first applied exclusively to the *Matrix view* table. To recalculate the factors and the genes tables, as well as the track of the found sites in the genome browser, only with matrices that were left after applying the filter, you should click on the *Apply changes* button.

The **Adjusted sequence enrichment filter** (3) allows to leave only matrices with the adjusted sequence enrichment values higher than the threshold specified by the filter. The set filter will be first applied exclusively to the *Matrix view* table. To recalculate the factors and the genes tables, as well as the track of the found sites in the genome browser, only with matrices that were left after applying the filter, you should click on the *Apply changes* button.

Filters will be also auto applied to all tables and tracks when you will switch from matrix table to either of the factor tables. To cancel the filters, click on the *Remove all filters* button.

TRANSFAC 2.0

Genes table

The genes table presented in the results visualization section fully corresponds to the gene table provided in the *Table 5* of the analysis report, where you can find the denominations of its column names and their contents description. Respective info is also summarized in the **?** hints available as mouseover messages upon navigating on them, similar to the *Factor view*, *Factor view Pro* and *Matrix view* tables.


To visualize in genome browser the sites found in the promoter of any gene, just click on the line with the gene of your interest inside the genes table and refer to the genome browser below to explore the predicted regulation model of the respective gene.
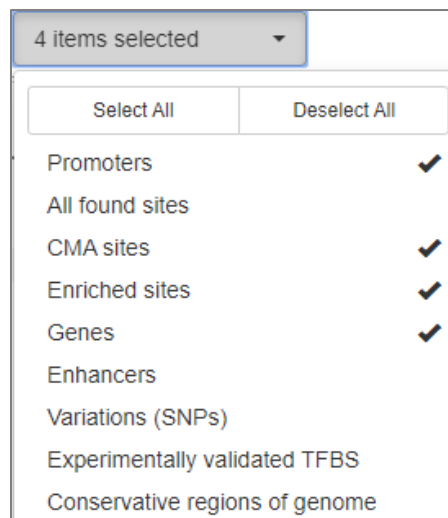
Genome Browser

The genome browser provides you with visualizations of the predicted gene regulation models for each gene from the input gene set. Having clicked on any gene from the *Genes table*, the promoter model of the respective gene will be automatically opened in the genome browser with visualization of the sites found within the respective promoter. By default the displayed tracks include:

- **Promoters** - the track of all promoters that were used for the performed analysis

- **CMA sites** - the track of sites belonging to the combinatorial matrices of the constructed CMA model (see Methods for further info)
- **Enriched sites** - the track of the remaining sites found to be enriched in the promoters of the studied genes
- **Genes** – the track with all Ensembl genes
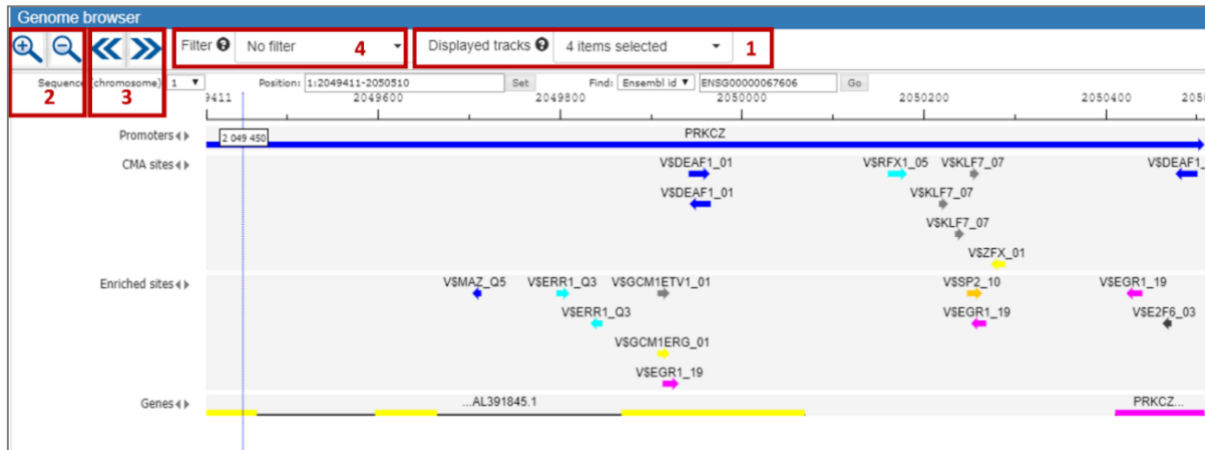
Displayed tracks can be customized using the **Displayed tracks** dropdown list (1)

You can zoom in and zoom out for a more generalized or a more detailed view of the promoter model using the **zoom in** and **zoom out** buttons (2). In case navigation in genome browser is lost due to multiple zoom clicks, you can always return to the promoter of your interest by selecting the respective gene in the genes table once again.

Navigation through the track is supported by the **Shift to the left** and **Shift to the right** buttons (3)

Filtering can be applied to the displayed sites using the **Filter** option (4):
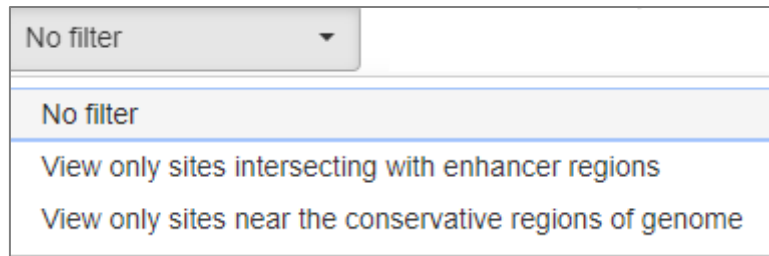
Depending on your selection, only sites intersecting with the conservative regions of the genome or those intersecting with the enhancer regions will be displayed. The genes table will be automatically recalculated leaving only the hits for the sites left after applying the respective filter. To cancel the applied filter, select the *No filter* option in the dropdown list of filters.

Please note that the *CMA sites* track shows only the best sites relevant to the identified CMA combinatorial model. The genes table will contain the full counts for all sites identified, both enriched and combinatorial. To view the respective full track of all found sites, you can add the *All found sites* track to the visualization:
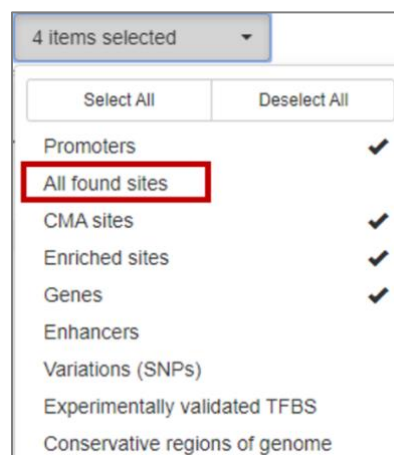
For receiving additional information about any of the found sites, you can click on the site of your interest and explore the contents displayed in the *Info box* section:
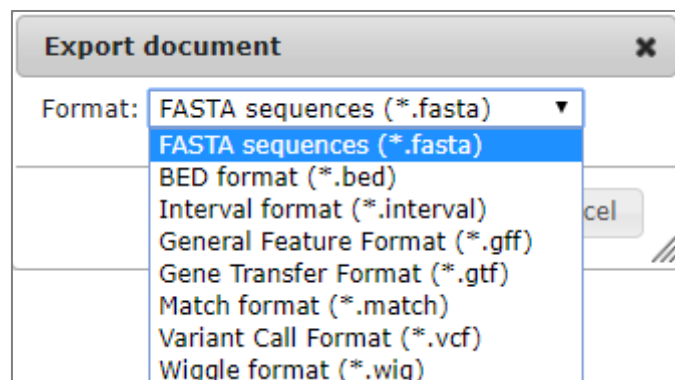
TRANSFAC 2.0

A right click on the track name in genome browser will open a context menu allowing you to remove the track from the visualization or open it as a table (function supported only for tracks of predicted sites or promoter track):
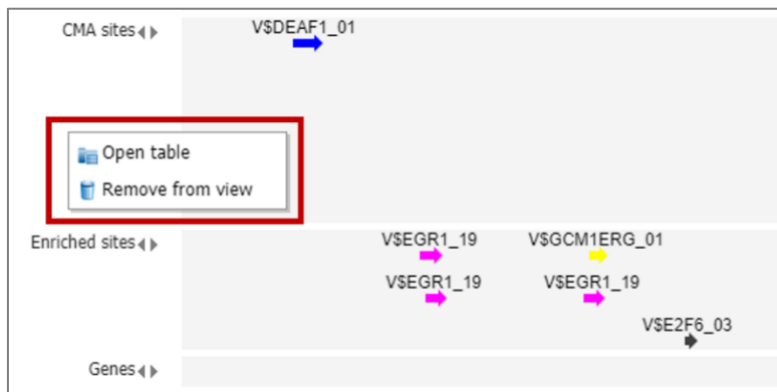
Tracks opened as table can be exported using the *Export* button at the top menu panel or used for further analysis from the geneXplain platform perspective.

The list of available export formats is as follows:

TRANSFAC 2.0

Heatmaps

If the functional classification of your input gene set by GO terms produced reliable results (by p-values and by the number of genes in each functional category), heatmaps for the respective GO categorization will be included in the optional section *Functional Analysis of Gene Regulation* of the analysis report.

The constructed image will contain two parts:

(1)   *The* heatmap of GO to GO terms mapping for the GO terms overrepresented among the studied gene set;

(2)   The heatmap visualizing how enriched motifs are associated with the respective GO categories.

The visualization examples of such heatmaps is given below. Depending on your preference, you can use either of the heatmaps: the one with the horizontal layout (provided in the analysis report as a clickable image), or the one with the vertical layout (provided under the link to alternative view in the report text):

TRANSFAC 2.0
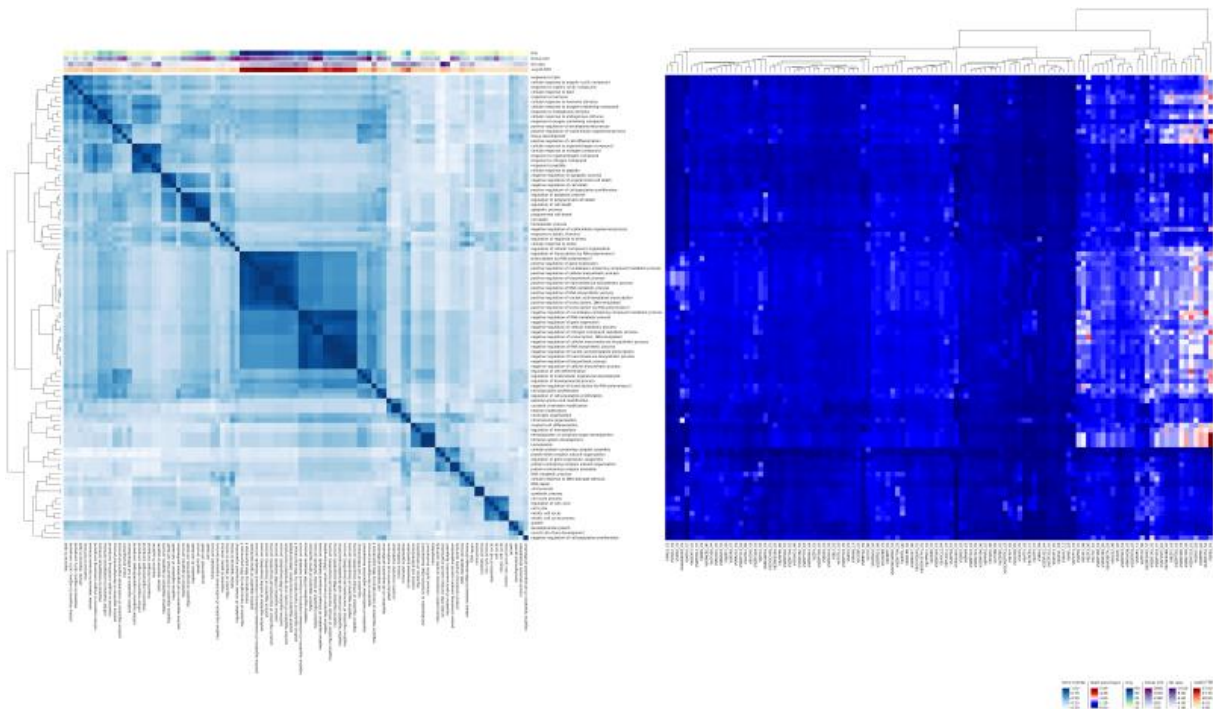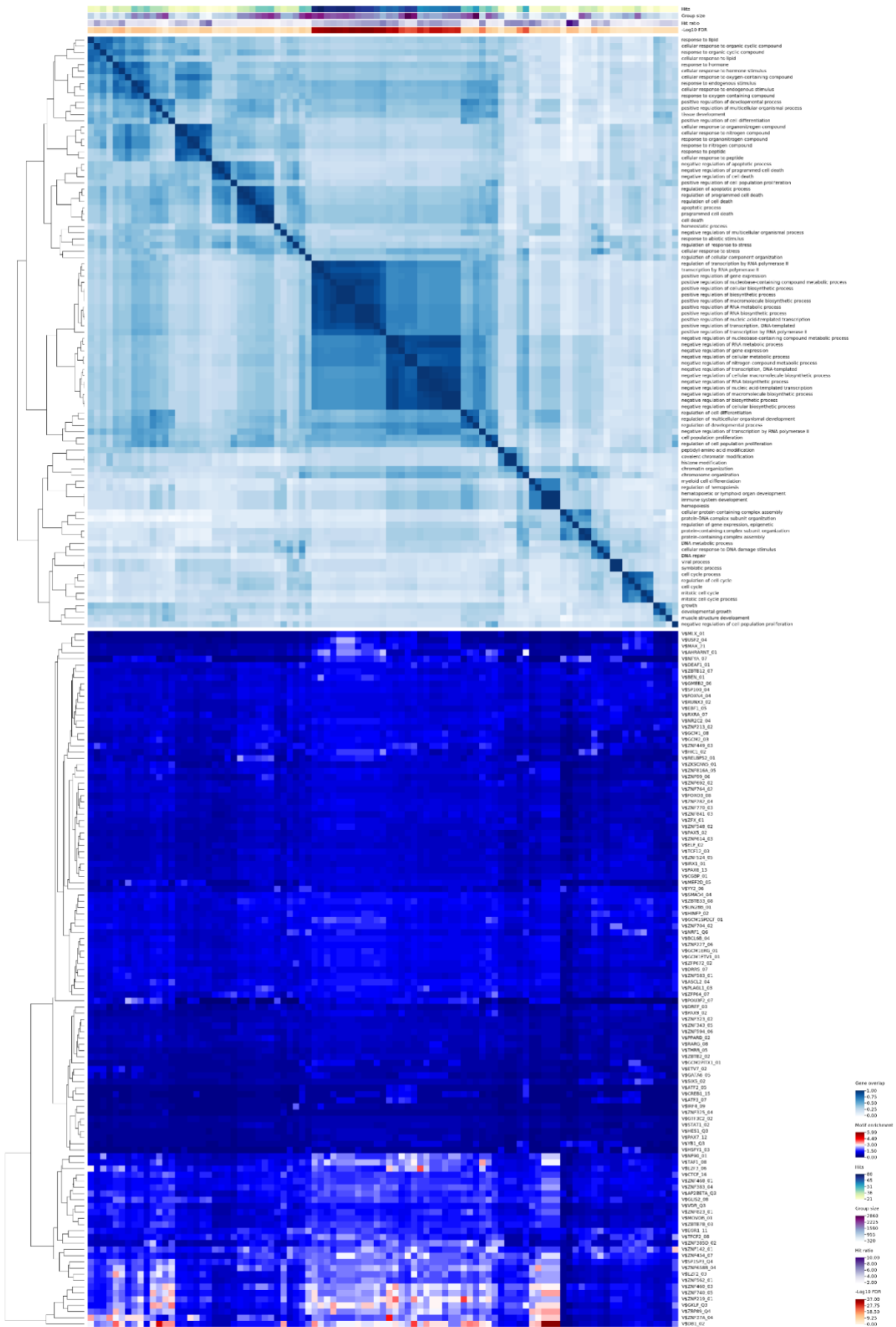
Switching to geneXplain platform

All results received with the MATCH Suite can be further analyzed in the geneXplain platform tool, access to which is provided to all users of the MATCH Suite. GeneXplain platform is a comprehensive online toolbox and workflow management system for a broad range of bioinformatics and systems biology applications.

For switching from the MATCH Suite perspective view to the geneXplain platform view, please use the perspectives switcher located at the top upper corner of the screen:
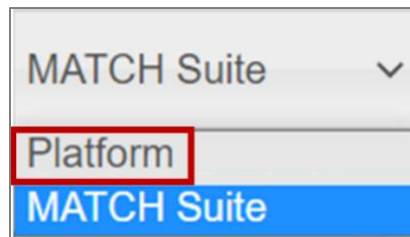
Please, refer to the User Guide of the geneXplain platform for detailed description of its functions.

TRANSFAC 2.0