New features of the geneXplain platform release 3.0, March 2014

1. New Start page

The new start page of our platform has been designed to better guide users to their topics of interest. After login to the geneXplain platform 3.0 the new start page opens in the Work Space. It shows you all those research areas that are supported, each by a number of bioinformatic workflows. Clicking on one of these tiles opens a detailed list of functions and pre-composed workflows, which you can directly launch from that list.

2. New Methods

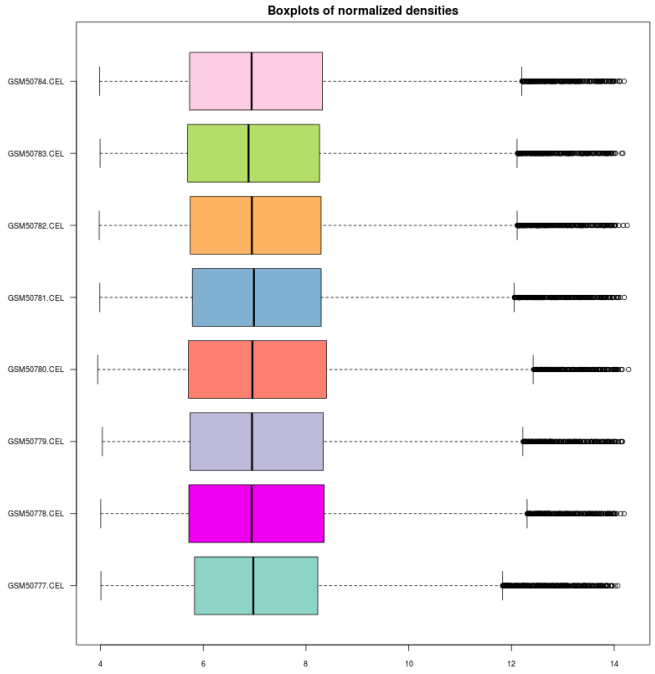
2.1. Expression data analysis/ Statistical analysis

2.1.1. **Limma –** This tool allows for differential expression analysis comparing up to five conditions based on the Bioconductor package limma (linear models for microarray data).

2.1.2. **EBarrays** – This tool allows for differential expression analysis comparing up to five conditions based on the Bioconductor package EBarrays (empirical Bayes methodology for gene expression data).

Both the **Limma** and **EBarrays** methods can be applied to calculate differentially expressed genes for different experimental methods, e.g. microarrays and RNA-seq.

2.1.3. **Normalization quality plots** – Produces box and line plots to visually inspect the results of normalization.



2.1.4. **Normalization support of new Affymetrix platforms** – HGU133+PM and Hugene-2_0-st.

2.1.5. **Principal Components Analysis (PCA)** – Calculates principal components for columns of a data table, showing the importance of each component, the transformation/rotation of each column according to the components as well as a data plot for the first two principal components.

2.2. Binding site analysis

2.2.1. **MEALR –** Analyzes binding site enrichment using sparse logistic regression. The resulting model is discriminative and contains a small subset of motifs from a possibly large database. Motifs are ranked by their importance as predictors.

2.2.2. **Composite Module Analyst (CMA)** – Composite modules are combinations of several TFBSs that are found together in a set of regulatory sequences. The in-house genetic algorithm searches for such combinations of TF binding sites that are overrepresented in the regulatory sequences under study compared to a background set of sequences. As input for the genetic algorithm we take the output of a site search analysis. Here is a list of our improvements:

- Site search performs better with custom sequences and custom intervals.
- Summary, reports can be generated, etc.
- New mode (switched on by default), which requires that sites of at least two different site models from the module are present on track.
- Histogram and model picture can be exported.
- Yes/No tracks open and export properly.

2.2.3. **New TRANSFAC profiles** with cut-offs for the number of sites per 1000 or 10K bases.

3. New Workflows

3.1. 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. 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 by how many reads each one is supported. In the first part of the workflow, the Cufflinks program is called from the Galaxy section of the geneXplain platform.

3.2. Analyze any DNA sequence with TRANSFAC® or GTRD

These workflows have been developed to search for putative transcription factor binding sites, TFBSs, in any input DNA sequence in EMBL, Fasta or GenBank formats. Using these workflows you can analyze DNA sequences of any (eukaryotic) species and of any genomic regions with the TRANSFAC® or GTRD collections of positional weight matrices.

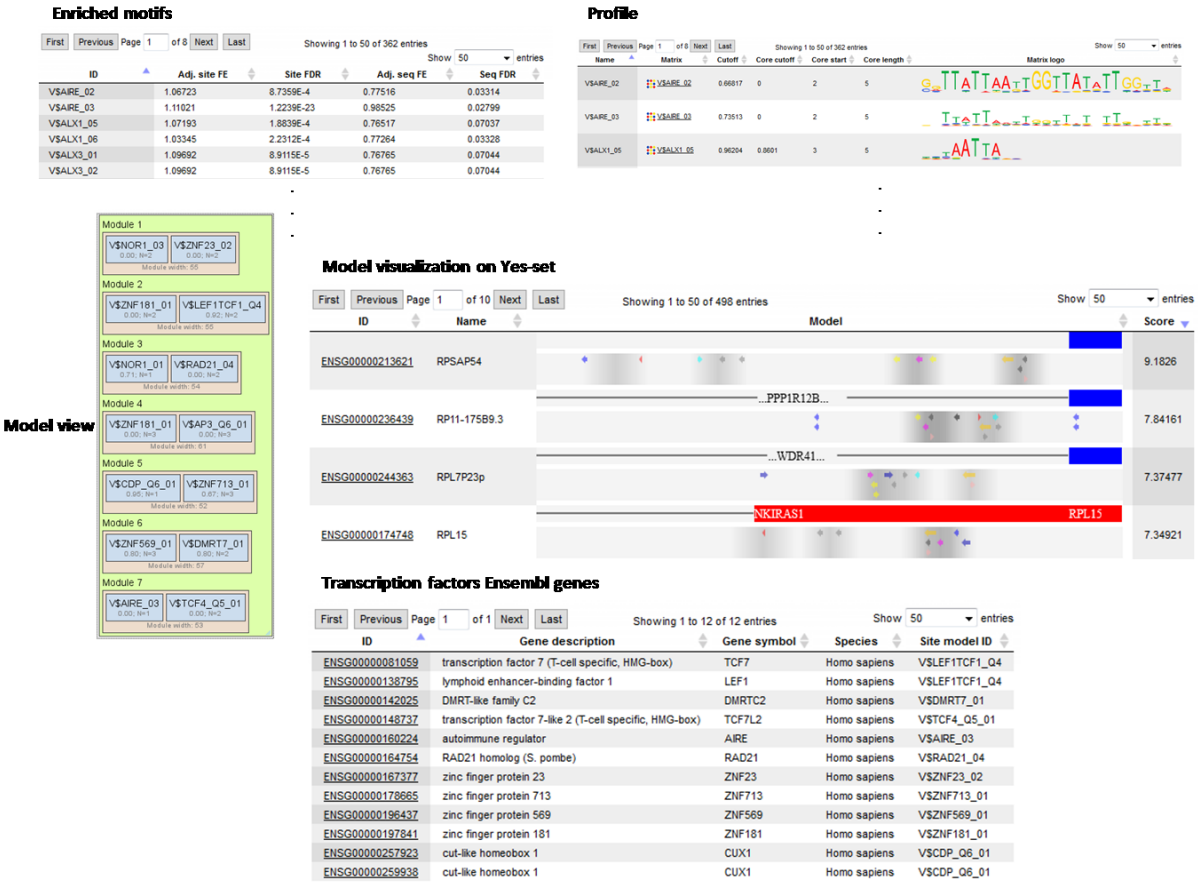
3.3. Identify enriched motifs in promoters with TRANSFAC® or GTRD

These workflows have been created to find individual motifs enriched in the promoters of the input gene set as compared with a background set (No set). The workflows contain the

new and highly efficient method ‘Search for enriched TFBSs’ [Version 2.0 (Adjusted p-values) with TRANSFAC®].

3.4. Identify enriched composite modules in promoters (TRANSFAC®)

This workflow enables the identification of combinations of several enriched TFBSs (= composite modules) in the promoters of the genes under study (Yes-set). The first step of the analysis is to find enriched motifs in the promoters of the input gene set with the new method ‘Search for enriched TFBSs’ [Version 2.0 (Adjusted p-values) with TRANSFAC®].
Also a filtering for highly overrepresented motifs is now possible.
The result of enriched motifs is used to create a Profile. A second motif search is performed with this matrix collection. The resulting composite modules differentiate the Yes-set from a background set (No-set). The Model visualization of the Yes-set and the list of transcription factors (Ensembl IDs) are the final results of the workflow.



3.5. Convert identifiers for multiple gene sets

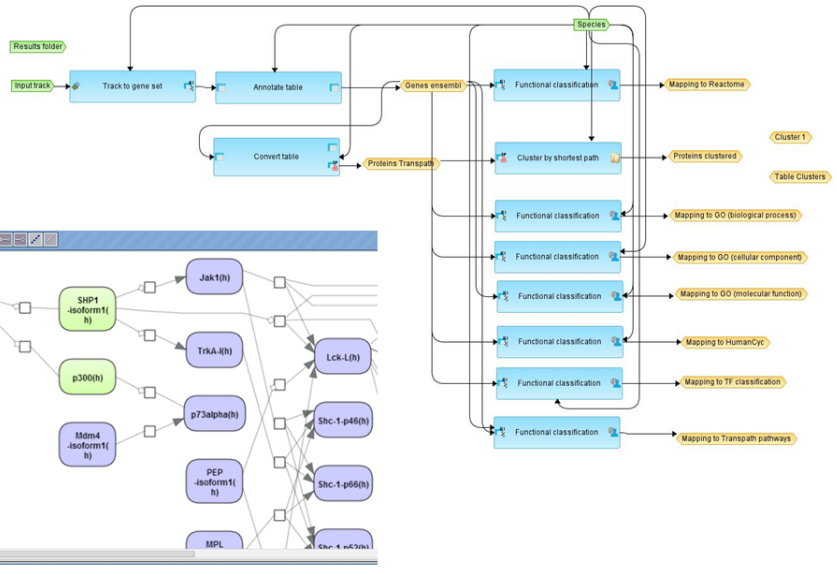
This workflow is designed to convert several gene or protein tables into a list with Ensembl IDs. The input is a folder containing several gene or protein tables, and these tables are processed by the workflow, one input table at a time. The first input table is converted to a list with Ensembl IDs using Convert table analysis and this step is repeated for the next

input table; several cycles are performed automatically corresponding to the number of tables in the input folder. The output is a new folder with tables having Ensembl IDs corresponding to each input table.

3.6. ChIP-Seq - Identify and classify target genes (TRANSPATH®)

This workflow is designed to identify and classify target genes using positional information of peaks found by the ChIP-seq approach. The classification includes enrichment analysis using the Reactome, GeneWays, GO, HumanCyc, TF classification and TRANSPATH® databases; it also comprises a Clustering of proteins with a redesigned analysis.

Workflow overview



Clustering example

Molecules shown in blue color are coming from the input gene/protein list, and those in green are added by the algorithm (intermediate molecules).

4. Enhanced Workflows

4.1. Compute differentially expressed genes (for Affymetrix probes, Agilent probes and Illumina probes)

Output file names are improved.

4.2. Compute differentially expressed genes using Hypergeometric test (for Affymetrix probes, Agilent probes, Illumina probes)

Additional output: non-changed genes are now calculated along with up-regulated and down-regulated ones.

4.3. ChIP-Seq - Identify and classify target genes (PROTEOME, TRANSPATH®)

ChIP-Seq - Identify and classify target genes

Mapping to Reactome pathway is improved.

Clustering by shortest path is enhanced.

4.4. Analyze any DNA sequence for site enrichment (TRANSFAC® and GTRD)

Adjustment of input formats.

4.5. ChIP-Seq - Identify TF binding sites on peaks (TRANSFAC®)

ChIP-Seq - Identify TF binding sites on peaks for multiple datasets (TRANSFAC®)

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

Enhancements for the input form.

5. Genome tracks - Phylogenetic footprinting and DNAse I HS sites

In the Public folder, the platform contains a new folder named Genome tracks with the following items:

5.1. DNAse I HS sites clustered ENCODE-UCSC hg19 – DNAse I hypersensitive sites clustered over multiple cell types in the human genome version hg19/GRCh37. DNAse I hypersensitivity often indicates the presence of bound proteins such as transcription regulators. The track can therefore be used to filter binding site predictions for experimentally determined possible locations of TFs.

5.2. PhastCons 46-way 75 non-CDS UCSC hg19 – These sites in the human genome have a conservation probability of over 75% in over 90% of the covered bases and have been filtered to remove CDS overlaps.

5.3. PhastCons 46-way 90 non-CDS UCSC hg19 – These sites in the human genome have a conservation probability of over 90% in over 90% of the covered bases and have been filtered to remove CDS overlaps.

5.4. PhastCons 60-way 75 non-CDS UCSC mm10 – These sites in the murine genome have a conservation probability of over 75% in over 90% of the covered bases and have been filtered to remove CDS overlaps.

5.5. PhastCons 60-way 90 non-CDS UCSC mm10 – These sites in the murine genome have a conservation probability of over 90% in over 90% of the covered bases and have been filtered to remove CDS overlaps.

6. Updated databases

TRANSFAC® and TRANSPATH® releases 2013.4 have been integrated. Many older releases are still available as well.

7. Other additions

7.1. Improved Javascript editor

7.2. Better R intercommunication – R-based analyses display messages from R.

7.3. Accurate sites/features display in genome browser

7.4. Venn diagram analysis – has been substantially enhanced. The subset tables corresponding to all sections of the Venn diagrams are automatically created and color and name for the resulting diagram can be customized.

7.5. Improved image handling – Images generated by analyses (Venn diagrams, histograms of the fold-change distributions, histograms of the Yes_No distributions in the CMA analysis,

enrichment plots in GSEA, etc.) can be resized and exported into different image formats with customizable scale.

7.6. **Improved appearance of the Reactome diagrams**

7.7. **Adjusted P-values/FDR** - Functional classification outputs contain adjusted P-values / False Discovery Rates according to the Benjamini-Hochberg procedure.