Gene set analysis

Methods description release 2.0

Table of contents

Preliminary steps

Data preparation

Depending on the type of the input gene list, the list is first prepared for further analysis. Gene lists submitted in the form of Entrez gene IDs or gene symbols are automatically converted to Ensembl IDs for further processing using the Convert table method of the geneXplain platform. Gene lists containing Ensembl IDs are directly taken for analysis.

All input gene lists are annotated inside the system with Gene description and Gene symbol columns for easy input overview. If you have selected to optimize your gene list by GO categories, the respective subset of your input genes belonging to the selected categories will be used as eventual input for the MATCH Suite analysis. The GO categorization is performed by the Functional classification method of the geneXplain platform. Further tree map visualization is done using the TreeMap on Functional Classification method. The tree map categorization is based on the REVIGO procedure of Gene Ontology visualization reduction [1]. The categories from the result of GO functional classification are clustered by REVIGO algorithm. The name of each cluster representative chosen by the method is shown in bold. The tree map is based on the maximum of top 30 clusters with best p-values in the results of gene set functional classification (general p-value threshold is set to the minimum of 0.05 with not less than 2 gene hits from the input set inside one GO category). The clusters similarity border used for running the REVIGO algorithm is set to 0.7.

Finding regulatory regions (YES set)

The promoters of genes from the input list are extracted using the Find regulatory regions method of the geneXplain platform. In case the studied tissue was specified during the analysis launch, this method creates a track of regulatory regions for input genes by using tissue-specific FANTOM5 promoters, where available. For other genes without tissue specific FANTOM5 promoters the system uses Ensembl promoter annotation instead. If tissue is not specified by the user, Ensembl promoters are used for all genes from the input list. The TSSs are taken from Fantom/TSS database (CAGE TSS database parameter is set to databases/Fantom5-Tissue-hg38). The Ensembl database used is Human104.38 (hg38). The promoter range to be analyzed can be specified during the analysis launch, [-500,100] by default.

The promoters of the input genes are compiled to form the YES set that is used for the site analysis.

Creating NO set

To identify transcription factor binding sites that regulate the genes from the input gene list, control sets (NO sets) of regulatory regions are required. They are generated using the Create random track method of the geneXplain platform. This method creates 5000 randomly sampled promoter regions from the human genome of the same length and promoter range as the promoters of the YES set. Random gene promoters that overlap with segments in the YES set are omitted from the sampling.

Constructing optimized profile distinguishing YES and NO sets

Site enrichment analysis

Site enrichment analysis is performed using the TRANSFAC® library of positional weight matrices (PWMs). It is done according to the PWM cut-off optimization approach published earlier [2,3]. The geneXplain platform method Search for enriched TFBS on tracks [3] is run on the YES and NO tracks to identify the binding sites enriched in the sequences of the YES track. To handle possible incidental enrichment of PWMs in some YES and NO set combinations, 5 iterations of site enrichment algorithm are run on 5 different NO sets, 1000 promoters each, randomly sampled from the 5000 random promoters described above. The summary output includes only those matrices that satisfied the given score cut-off thresholds in all 5 runs (Supplementary table 1 in the analysis report). Fold enrichment of sites (Site enrichment) as well as of sequences with at least one site (Sequence enrichment) are optimized and reported as statistically corrected odds ratios (99% confidence interval). The reported values are corrected for small site or sequence numbers, taking into account possible variability, and are therefore more suitable for ranking PWMs by their fold enrichment in the YES track promoters. The algorithm seeks optimal score thresholds for each type of enrichment separately and reports False Discovery Rates (FDRs, Benjamini-Hochberg method) in addition to uncorrected P-values (site enrichment p-value, binomial test and sequence enrichment p-value, Fisher test).

The matrix profile used to identify potential TFBSs is a collection of 5643 TRANSFAC® vertebrate matrices [4,5], carefully selected for the purpose of enrichment analysis. For each matrix, the cutoff value of the Match Site Score (MSS) is optimized so that an optimal enrichment of sites in the YES compared to the NO set is achieved. From the 5 independent runs, the median of these site cutoff values is computed [3]. The site enrichment cutoff is set to 1.0, and the site FDR cutoff is set to 0.05. For the Sequence enrichment initially no cutoff is set.

The results of the site enrichment analysis are presented in the Supplementary table 1, for which a link is given in the analysis report. The matrices in this table are sorted by site enrichment. If #Accepted is equal to 5 and the matrix rate (%) equal to 100, the selected matrices have fulfilled the thresholds in all 5 runs.

Orthologous and paralogous extension

The list of TFs (transcription factors) originally associated with each PWM (on the basis of TRANSFAC(R) curation) is extended to other TFs that are orthologs and paralogs to the ones already associated with the listed matrices. For this purpose, factor clusters have been defined based on geneXplain's expert knowledge; the corresponding table of factor clusters, with grouped matrices, can be found here.

Orthologous extension is done in all cases where a matrix was derived for a defined TF from one particular species. Among mammals, and even among vertebrates, the DNA-binding domains of orthologous TFs are (nearly) identical, so that the same DNA-binding specificity can be reasonably inferred for all orthologs.

Similarly, paralogous extension means to infer the DNA-binding specificity of a TF's paralogs [6]. The term "paralog" is not used here *sensu stricto*, that is we do not claim that all TFs called "paralogs" here were really generated by gene duplication events. However, they also exhibit

(nearly) identical DNA-binding domains and were therefore defined as Subfamily in the Transcription Factor Classification (TFClass) [7].

Factor expression filtering

The list of matrices (PWMs) received on the previous step of analysis is then filtered by factor expression for the tissue selected during the analysis launch (if applicable). Matrices that do not represent any factors (after orthologous and paralogous extension) that are known to be expressed in the selected tissue are excluded from further consideration. If no tissue was selected during the analysis launch, this step is skipped.

The expression values used for this filtering were taken from Human Protein Atlas [8]. A TF is considered to be expressed in a certain tissue if its expression value is higher than 1.

Redundancy filtering

For each factor cluster defined by the table described above, only one matrix is left for further consideration, which is the one that maximizes the adjusted site enrichment value. The resulting list of matrices after the orthologous and paralogous extension, tissue filtering and redundancy filtering is shown in the Supplementary table 2 of the analysis report. Its columns are a subset of Supplementary table 1 and have the same denominations and values as in Supplementary table 1 described above.

Optimized profile distinguishing YES from NO sets

The filtered list of matrices presented in Supplementary table 2 in the analysis report is used to construct a new matrix profile that is specific for the input gene set. This profile will be further used in the next steps of analysis for running the site search algorithm (MATCH). The profile is constructed using the Create profile from site model method of the geneXplain platform. The cutoffs for the profile are selected as the median site cutoff of matrices from the Supplementary table 2 linked to the analysis report. The constructed profile is provided in the analysis report as Supplementary table 3. Each row of this table summarizes the information for one site model. For each site model, the cutoff is shown in the column Cutoff. According to the TRANSFAC® standard, the core of each matrix is specified. 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.

Searching for over-represented TFBSs and their combinations

Search for TFBSs using the constructed profile

The constructed profile (see above) is used to find potential binding sites by applying the MATCH algorithm [9]. The respective method in the geneXplain platform is called TRANSFAC® Match™ for tracks. It produces a track of sites, which were found to be overrepresented in the YES set compared to the NO set. The produced track is then taken for the combinatorial analysis.

Combinatorial analysis of TFBSs

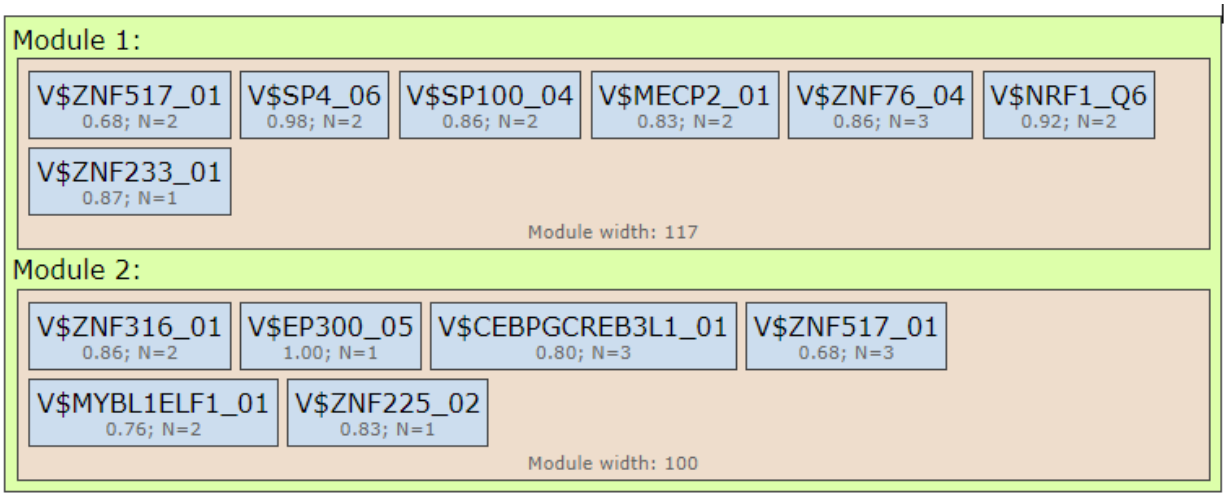
Composite modules are combinations of several TFBSs that are found together in a set of regulatory sequences. The search for composite modules is performed using geneXplain's in-house implementation of a genetic algorithm, the Composite Module Analyst (CMA) [10]. This genetic algorithm takes the output from the site search on track analysis as input. As result, it produces composite modules that differentiate the YES set from the background NO set. CMA constructs a generalized model of the regulatory regions of the studied genes by specifying combinations of potential TFBSs that are most frequently clustered together. CMA identifies the transcription factors that through their cooperation are able to provide a synergistic effect and, thus, are likely to have a great influence on the gene regulation process.

CMA is provided by the Construct composite modules on tracks method of the geneXplain platform. The following parameters are used for running the CMA algorithm:

Number of iterations - 500
Population size - 1000
Non-change limit (number of iterations to stop after if best score is not improved) - 50
Elite size (number of elite organisms, i.e. best organisms to survive unconditionally) - 50
Mutation rate (how often mutations occur and how significant they are) - 0.9
Penalty rate - 0.3
Min modules - 2
Max modules - 2
Min models - 6
Max models - 8
Min sites to account - 1
Max sites to account - 3
Min module width - 10
Max module width - 200

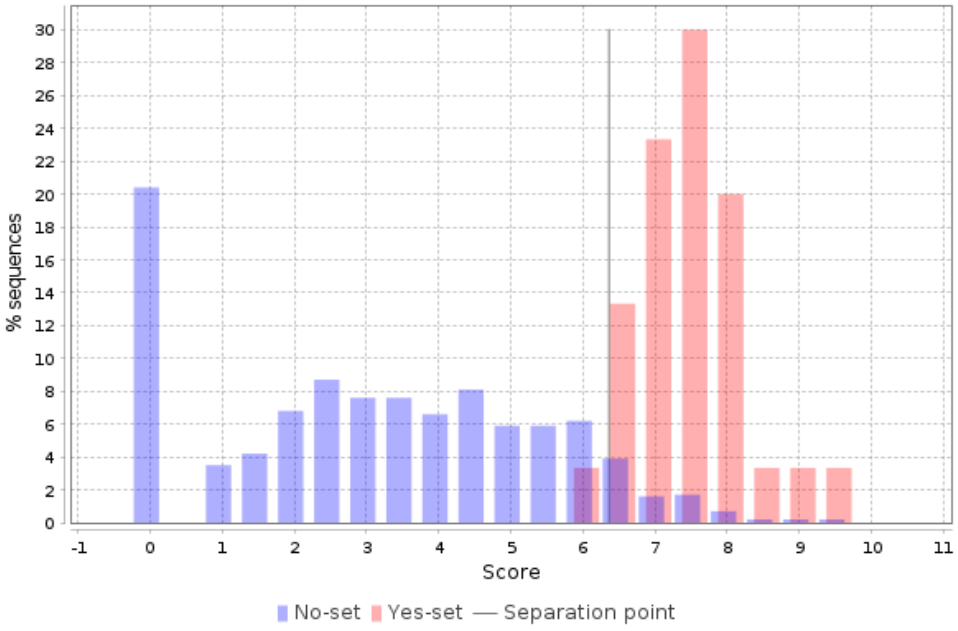
Composite modules can have a complex hierarchical structure consisting of two levels: site models and modules. The highest hierarchical level contains two modules and corresponds to the whole promoter model. The first level, site model, corresponds to the individual site model, based on one PWM. Names of the site models are the same as the matrix names (site models are taken from the profile that was used in the site search). In the resulting schema (see example below) the site models are shown by blue boxes. Within these boxes, there are two values below each 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. The second value is the maximum number of best individual matches (sites) found for this site model and taken into account for calculating the score of the module.

Promoter model example:



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 resulting schemas modules are shown in green boxes, and they are numbered: Module 1 and Module 2. The complexity of the promoter model to be constructed is defined by the number of units of each level: number of modules, number of site models, as well as the maximum number of individual sites to be considered.

The CMA score is calculated for each promoter depending on the number of modules, site models, sites, their scores and other statistical parameters. The higher the CMA score of a promoter, the better is the differentiation of this promoter from the promoters of the NO set. The distribution of scores for individual promoters is shown as a histogram in Figure 3 of the analysis report. An example is given below:

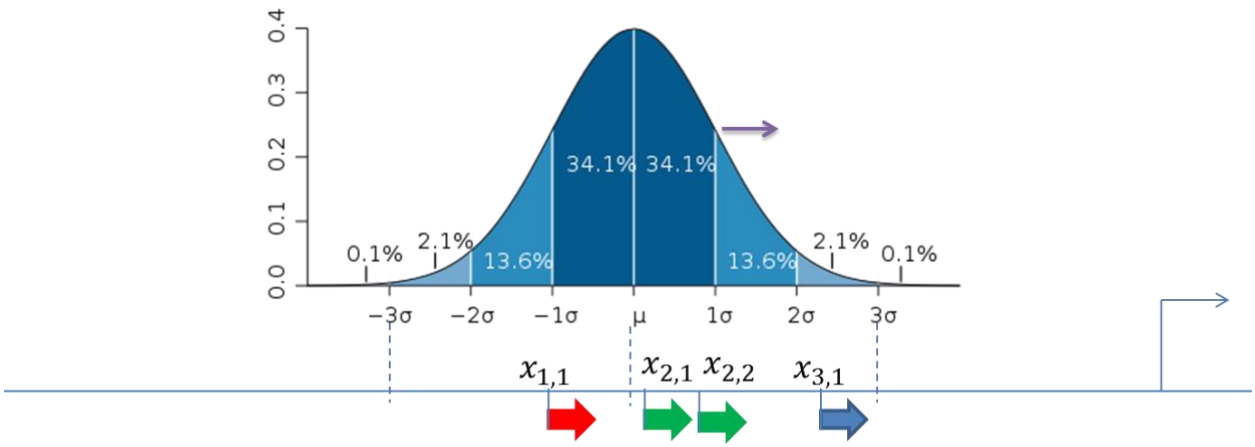


The CMA score of the promoters is shown on the X axis of the histogram and the percentage of promoters (% sequences) having this score is shown on the Y axis. The separation point shown in gray corresponds to the average value. Promoters from the NO set with a score above the separation point (blue bars to the right of the separation point) are referred to as false positives.

Promoters from the YES set with a score below the separation point (red bars to the left of the separation point) are referred to as false negatives. The YES promoters with a score above the separation point are well separated from the NO promoters, meaning that for these promoters the constructed composite model is most suitable. In Table 5 of the analysis report the CMA score is used for sorting the analyzed genes.

Score calculation of the composite modules

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 *Mk* (*Score(Mk)*, *k* = 1, …, *K*) is calculated according to the following 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 *Mk*, 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 *Mk*.

Standard deviation (σ) of the normal distribution is subject to optimization by the genetic algorithm and represents the width of the module in the result of the composite module analysis.

Identifying transcription factors regulating the input gene set

Based on the matrices identified as significantly enriched in the promoters of the input gene set (matrices from Supplementary table 2 in the analysis report) and the results of performed MATCH and CMA analyses, the factor table is constructed, which lists all transcription factors (TFs) associated with the respective matrices (Supplementary table 4 in the analysis report). The factors are ranked by an integrative ranking procedure, based on the sum of the ranks calculated for the following factor parameters:

- the maximum value of adjusted site enrichment of the factor's matrices
- the presence of the factor's matrices in the constructed CMA model
- the level of factor expression in the tissue selected during the analysis launch or, when no tissue was selected, the average factor expression value across all supported tissues, (values of relative factor expression in a given tissue are taken from the Protein Atlas)
- the rank of factor expression in a given tissue compared to all other supported tissues or the rank of average factor expression among all supported tissues.

The rank of factor expression levels is based on the 'rank / number of supported tissues', i. e. up to 61. But as the ranking includes the values of average factor expression across all tissues, the maximum rank can be 62.

The columns denominations of Supplementary table 4 are as follows:

ID – factor ID
Genes: Ensembl ID –ID of gene corresponding to the respective TF
Adjusted factor enrichment – maximum adjusted site enrichment value among the factor's matrices
Adjusted sequence enrichment - sequence enrichment of the factor's matrix with the highest value of adjusted site enrichment
Factor classification – as provided by TF Class
Factor expression in tissue – factor expression value in the tissue selected for the analysis launch as provided by Protein Atlas or the value of average factor expression across all supported tissues
Factor name – name of TF
Family name – name of TF family as provided by TF Class
Gene symbol –symbol of the gene corresponding to the respective TF
Sequence enrichment FDR - sequence enrichment FDR of the factor's matrix with the highest value of adjusted site enrichment
Site enrichment FDR – site enrichment FDR of the factor's matrix having the maximum adjusted site enrichment value
Site model – ID of the factor's matrix with maximum adjusted site enrichment
Difference of tissue – difference of factor expression in the tissue selected for the analysis launch from the average expression value of factor across all supported tissues; if no tissue was selected for the analysis launch, this column duplicates the factor expression specificity value provided in the 'Tissue specificity' column. See explanation of 'expression deviation from average' below for more info
Rank of tissue – the expression rank of the tissue selected for the analysis launch out of all tissues supported for the current factor; if no tissue was selected for the analysis launch, the rank of average factor expression value out of all tissue expression values available for the current factor is shown
Tissue specificity – provides the value of general factor expression specificity described below

Composite model component – provides the ID of the factor's matrix that was included in the constructed CMA model (if applicable)
Factor rank – provides the rank sum of the factor (calculation described above)

Having applied the factor ranking in respect to the abovementioned criteria, the top 30 factors by rank are selected for the factor table (Table 3 of the analysis report) together with all those factors whose matrices were included in the CMA model.

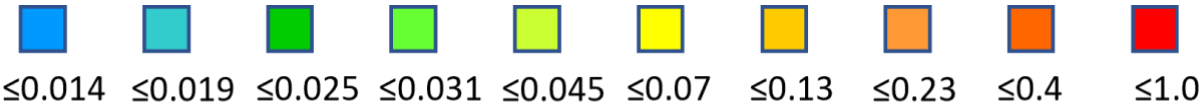
The example below shows part of the Factor view Pro (Table 3 of the analysis report) with amygdala selected as tissue in the analysis:

Factor name	Gene symbol	Class name and TF classification	Site model	Adjusted factor enrichment	Factor rank	Amygdala: factor expression	Amygdala: expression difference (rank)
BTEB4	KLF16	C2H2 zinc finger factors 2.3.1.2.16	V$KLF7_03	2.14	1	19.5	10.7 5/62

The 'Amygdala: factor expression' column shows the value of factor expression in the selected tissue as provided by the Protein Atlas.

The 'Amygdala: expression difference (rank)' column shows three values:

(1) The expression deviation for the selected tissue from average (10.7 in the given example)
(2) The expression rank of the factor in the selected tissue in comparison to other supported tissues for this factor (5/62 in the given example)
(3) The general factor expression specificity level represented by color (■) using the following color code:



These values were calculated as follows:

Expression deviation from average

For each expression value of a factor in a given tissue its difference to the average expression of this factor across all supported tissues was calculated. The difference can be either positive or negative depending on whether the factor expression level in the inspected tissue is higher or lower than its average expression level across all tissues.

Expression rank

All available factor expression values across all supported tissues were sorted in a decreasing order and ranked respectively (rank 1 refers to tissue of maximum expression). The rank of factor expression in a given tissue is provided in relation to the supported number of tissues (61) together with the average value of factor expression across all tissues (resulting in the maximum possible rank of 62 for the tissue of lowest expression).

General factor expression specificity

On a scale from 0 (blue, lowest specificity) to 1 (red, highest specificity) the specificity of the factor expression profile among all supported tissues is shown. The expression values taken from Human Protein Atlas [8] were used to calculate for each TF the entropy of its expression distribution as defined by Schug et al. [11]. To convert it into a metric for expression specificity, it was subtracted from the maximal value possible ($\log_2 N$, with N the number of tissues considered) and scaled to a range between 0 and 1, so that a value of 0 indicates equal expression of a TF in all tissues analyzed, and 1 for exclusive expression of a TF in one tissue only.

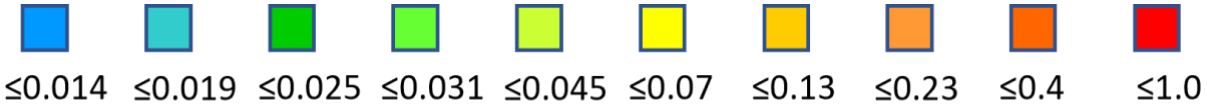
The example below shows part of the Factor view Pro (Table 3 of the analysis report), when no tissue was specified in the analysis:

Factor name	Gene symbol	Class name and TF classification	Site model	Adjusted factor enrichment	Factor rank	Average factor expression across all tissues	Expression specificity (rank of average)
STAT1	STAT1	STAT domain factors 6.2.1.0.1	**V$STAT5B_01**	1.54	1	26.01	0.02 20/62

The 'Average factor expression across all tissues' is calculated from the expression values provided by the Protein Atlas.

The 'Expression specificity (rank of average)' column shows two values:

(1) The rank of average factor expression in comparison to supported tissues for this factor (20/62 in the given example)
(2) The general factor expression specificity level represented by color (■) and value (0,02) using the following color-value code:



Denominations of these values are the same as abovementioned, when a tissue was selected during the analysis.

The Adjusted factor enrichment column of the factor table refers to the maximum value of adjusted site enrichment among the factor's matrices identified in the MATCH analysis.

The site model column shows the matrices corresponding to the given factor that were identified by MATCH. In bold are given the matrix IDs of matrices that got into the constructed CMA model.

Class name and TF classification column shows the respective values for the given transcription factor as provided by the TF Class [7].

References

[1] Supek, F., Bošnjak, M., Škunca, N., & Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. PloS one, 6(7), PubMed.

[2] Kel A, Voss N, Jauregui R, Kel-Margoulis O, Wingender E. (2006) Beyond microarrays: find key transcription factors controlling signal transduction pathways. BMC Bioinformatics. 7 Suppl 2(Suppl 2), S13. PubMed.

[3] Koschmann J, Bhar A, Stegmaier P, Kel AE, Wingender E. "Upstream analysis": an integrated promoter-pathway analysis approach to causal interpretation of microarray data. Microarrays. 2015;4:270–86. Pubmed.

[4] Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., Land, S., Barre-Dirrie, A., Reuter, I., Chekmenev, D., Krull, M., Hornischer, K., Voss, N., Stegmaier, P., Lewicki-Potapov, B., Saxel, H., Kel, A.E., Wingender, E. (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. Nucleic Acids Res. 34:D108-D110. PubMed.

[5] Wingender, E. (2008) The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief. Bioinform. 9:326-332. PubMed.

[6] Haubrock , M., Li, J., Wingender, E. (2012) Using potential master regulator sites and paralogous expansion to construct tissue-specific transcriptional networks. BMC Syst. Biol. 6 (Suppl. 2):S15. PubMed.

[7] Wingender, E., Schoeps, T., Haubrock, M., Krull, M., Dönitz, J. (2018) TFClass: expanding the classification of human transcription factors to their mammalian orthologs. Nucleic Acids Res. 46:D343-D347. PubMed.

[8] Uhlén, M. et al. (2015) Tissue-based map of the human proteome. Science 347:1260419. PubMed.

[9] Kel, A.E., Gössling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V., Wingender, E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences. Nucleic Acids Res. 31:3576-3579. PubMed.

[10] Waleev, T., Shtokalo, D., Konovalova, T., Voss, N., Cheremushkin, E., Stegmaier, P., Kel-Margoulis, O., Wingender, E., Kel, A. (2006). Composite Module Analyst: identification of transcription factor binding site combinations using genetic algorithm. Nucleic Acids Research, 34(suppl_2):W541-W545. PubMed.

[11] Schug, J., Schuller, W.-P., Kappen, C., Salbaum, J. M., Bucan, M., Stoeckert, C. J. Jr (2005) Promoter features related to tissue specificity as measured by Shannon entropy. Genome Biol. 6:R33. PubMed.

User guide

This document is intended to explain the analysis process underlying the MATCH Suite gene set analysis pipeline and is not aiming to provide any instructions on how to use the system. For MATCH Suite interface description and any further assistance on how to operate in the system, please refer to the MATCH Suite User guide.

Note

Please note that all methods of the geneXplain platform have extended descriptions accessible upon viewing the info box with method information from the geneXplain platform perspective (open the method of your interest by the link in this document, switch to *Platform* perspective in the right upper corner of the system and click on the *Toggle UI mode*  button at the top menu panel to see the method description in the info box located in the bottom left corner of the screen).