Single gene analysis

Methods description release 2.2

Table of contents

Preliminary steps 1

Finding regulatory regions 1

Estimating cumulative binding affinity of transcription factors to the analyzed regulatory regions 1

Profile optimization by GO terms 1

Orthologous and paralogous extension 4

Redundancy filtering 4

Optimized profile construction 5

Searching for TFBSs 5

Search for TFBSs using the constructed profile 5

Identifying transcription factors regulating the input gene 6

References 9

User guide 10

Note 10

Preliminary steps

Finding regulatory regions

The promoter of the input gene is 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 promoter region of the input gene by using tissue-specific FANTOM5 promoter (best supported / most 5 prime / most 3 prime), if available. If tissue specific FANTOM5 promoter was not found for the input gene, the respective Ensembl promoter will be used for the analysis depending on the parameter selection upon the analysis launch. 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 enhancers and silencers of the input gene are being extracted from the track of all known enhancers and silencers contained in the geneXplain databases using the Filter track by condition method of the geneXplain platform. This track was optimized to contain the enhancers and silences of proper lengths for performing the further analysis, for this all enhancers/silencers with initial lengths less than 300 bp were equally extended from their flanks to the length of 300 bp, and those enhancers/silencers the lengths of which were more than 2500 bp were equally cut to the core of their middle 2500 bp.

If tissue was selected upon the analysis launch, only these enhancers and silencers of the input gene, which are known to be active in the selected tissue, will be extracted. If tissue was not selected, all known enhancers and silencers of the input gene will be considered in further analysis. If there were no enhancers or silencers found for the input gene, only its promoter region will be taken for further analysis.

If the input gene appeared to have the enhancers or silencers, and they have succeeded the tissue filter in case tissue was specified during the analysis launch, these enhancers/silencers will be checked for their overlap with the help of the Jaccard measure [1]. Enhancers or silencers with the minimum overlap of 200 bp and Jaccard index higher than 0.5 will be merged into one regulatory region. The resulting track of all regulatory regions of the input gene, that will be taken to further analysis, will include the promoter and all enhancers and silencers that fulfilled the abovementioned checks.

Estimating cumulative binding affinity of transcription factors to the analyzed regulatory regions

Profile optimization by GO terms

The matrix profile, i.e. a set of positional weight matrices (PWMs) from TRANSFAC® library, used to identify potential TFBSs in the MATCH Suite, is a collection of 5752 TRANSFAC® vertebrate matrices [2,3], carefully selected for the purpose of binding affinity estimation. If Gene Ontology (GO) categories were selected during the analysis launch, this profile will be optimized by narrowing it only to those matrices, the matching transcription factors of which are encoded by the genes that belong to the selected GO categories. The link to the constructed profile, optimized by GO terms, is provided in Step 1 of the analysis report.

Estimation of transcription factors cumulative binding affinity to the analyzed regulatory regions

The constructed track of regulatory regions of the input gene is submitted to the Affinity Match for Tracks [4] method of the TRANSFAC 2.0 for calculation of affinity scores and corresponding p-values for the given regulatory regions using the standard MATCH Suite matrix profile consisting of 5752 TRANSFAC® vertebrate matrices [2,3], or the customized matrix profile optimized by the selected GO terms, as described above.

The affinity score estimates the overall affinity of a binding specificity to each of the studied regulatory regions of the input gene and has the advantage of assigning one single quantity per PWM to the studied sequence without requiring the choice of a PWM score cutoff.

The affinity score for the regulatory region x is calculated by the following equation where W denotes the number of windows of the sliding PWM through positions w, (w=1…W). The number of windows W equals L + 1 - M where M is the length of the PWM.

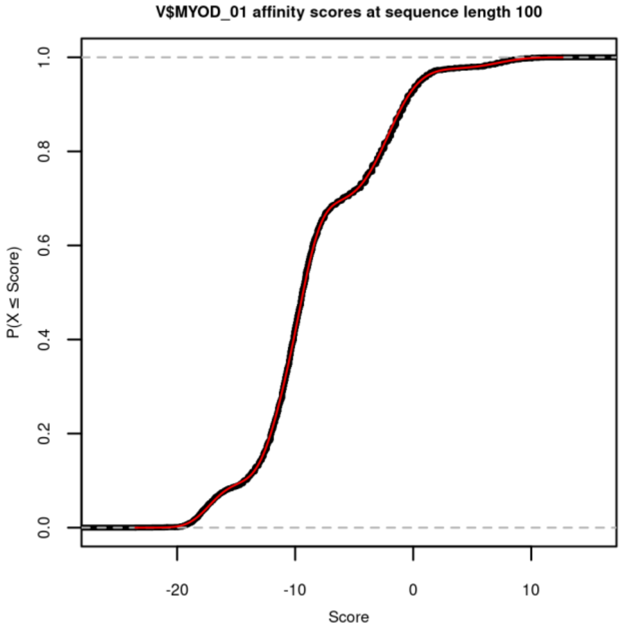
$$f(x) = log(\sum_{w} exp(LS_{max}(x_w))/W)$$

Here, $LS_{max}(x_w)$ is the maximum between two log-odds scores $LS(x_w)$ of the PWM at the position w of regulatory region x in the plus and in the minus strand of DNA*.*

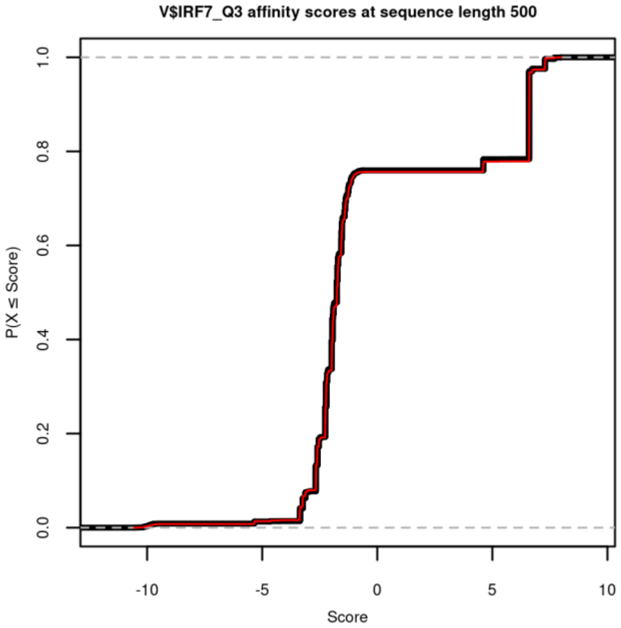
$$LS(x_w) = \log\left(\frac{\prod_{i=1}^{M} p(i, k_i\ )}{\prod_{i=1}^{M} b(\ k_i)}\right)$$

where, index $i=1,\ldots M$ goes through the full length of the PWM; k_i – is the nucleotide letter in the regulatory region x at the position of $w+i-1$; $p(i,\ k_i)$ is the element of the PWM at the *i-th* position of the matrix and for the letter k_i of the matrix; $b(k_i)$ is the background probability of the letter k_i in the considered regulatory regions (e.g for the model of equal distribution of frequencies of all four letters $b(k_i)$= ¼).

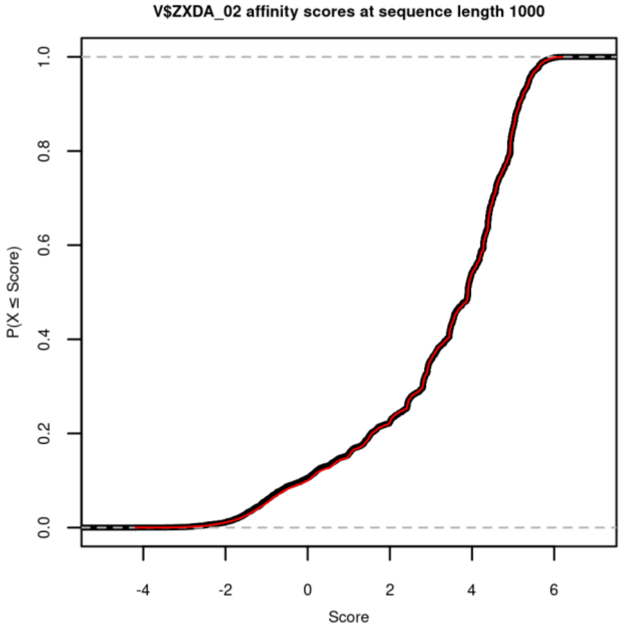
The statistical significance (p-value) of affinity scores is estimated for large sets of random sequences with the dinucleotide composition of human promoters and sequence lengths ranging from 100 to 2500. The score distributions are modeled using mixtures of Gaussian distributions. Score significance for sequence lengths not addressed by a specific mixture model is determined by spline-based interpolation over the available sequence length range. Several examples of affinity score distributions and corresponding mixture model estimates are shown below (black line: empirical distribution, red line: mixture model).



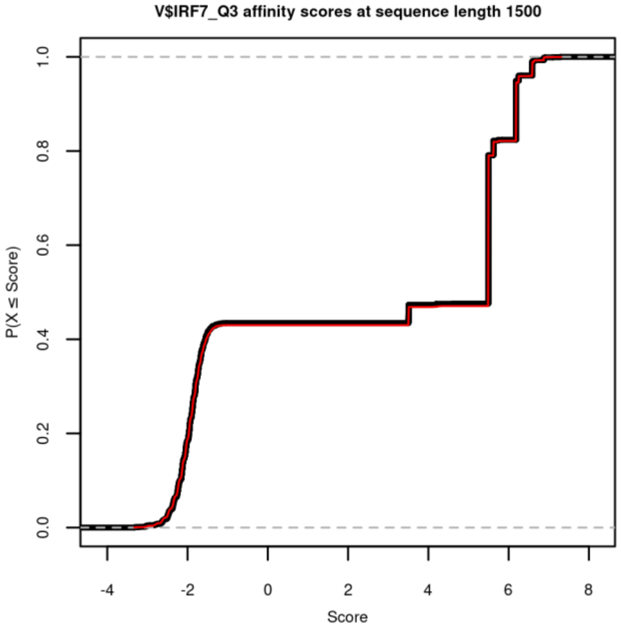
V$MYOD_01 affinity scores at sequence length 100



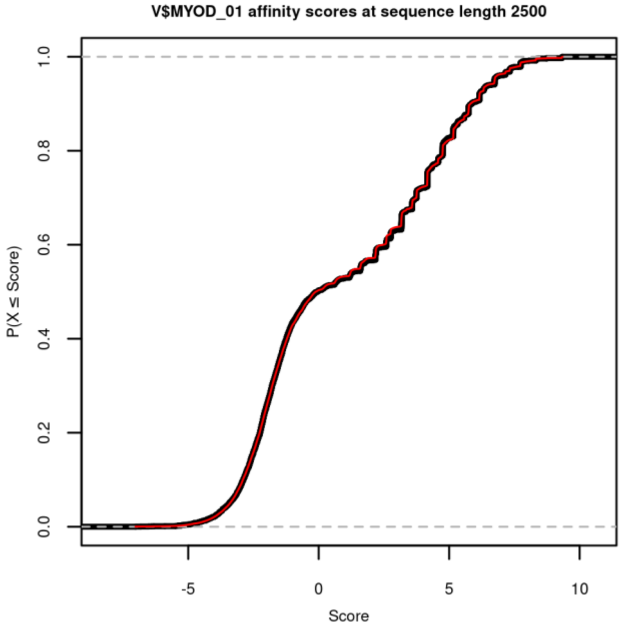
V$IRF7_Q3 affinity scores at sequence length 500



V$ZXDA_02 affinity scores at sequence length 1000

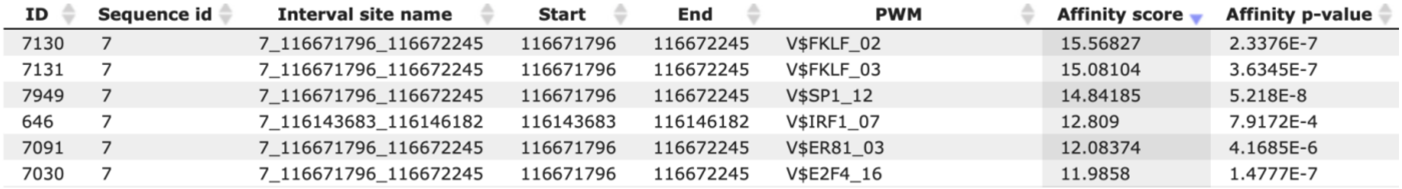


V$IRF7_Q3 affinity scores at sequence length 1500



V$MYOD_01 affinity scores at sequence length 2500

The output table of the Affinity Match for Tracks is presented in the analysis report as Supplementary Table 1 (Table S1). This table reports affinity scores and p-values of PWMs for each of the studied regulatory regions of the input gene:



ID	Sequence id	Interval site name	Start	End	PWM	Affinity score	Affinity p-value
7130	7	7_116671796_116672245	116671796	116672245	V$FKLF_02	15.56827	2.3376E-7
7131	7	7_116671796_116672245	116671796	116672245	V$FKLF_03	15.08104	3.6345E-7
7949	7	7_116671796_116672245	116671796	116672245	V$SP1_12	14.84185	5.218E-8
646	7	7_116143683_116146182	116143683	116146182	V$IRF1_07	12.809	7.9172E-4
7091	7	7_116671796_116672245	116671796	116672245	V$ER81_03	12.08374	4.1685E-6
7030	7	7_116671796_116672245	116671796	116672245	V$E2F4_16	11.9858	1.4777E-7

Filter by Affinity p-value < 0.05 is then applied to Table S1. After that, for each matrix, that remained to be present in Table S1 after filtering, the best affinity p-value among all examined regulatory regions is chosen. The respective affinity score value is taken accordingly. Supplementary Table 2 (Table S2) of the analysis report includes these best affinity p-values and respective affinity score values for each of the PWMs in each of the studied regulatory regions.

Orthologous and paralogous extension

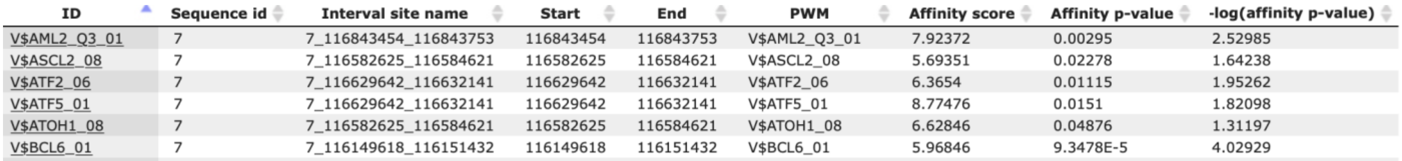
The list of TFs (transcription factors) originally associated with each PWM from the results of Affinity Match (on the basis of TRANSFAC(R) curation) is extended to other TFs that are orthologs and paralogs to the ones already associated with the listed matrices. For this purpose, factor clusters have 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 [5]. 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) [6].

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 minimizes the affinity p-value. The resulting list of matrices after the orthologous and paralogous extension and redundancy filtering is shown in the Supplementary Table 3 (Table S3) of the analysis report:



ID	Sequence id	Interval site name	Start	End	PWM	Affinity score	Affinity p-value	-log(affinity p-value)
V$AML2_Q3_01	7	7_116843454_116843753	116843454	116843753	V$AML2_Q3_01	7.92372	0.00295	2.52985
V$ASCL2_08	7	7_116582625_116584621	116582625	116584621	V$ASCL2_08	5.69351	0.02278	1.64238
V$ATF2_06	7	7_116629642_116632141	116629642	116632141	V$ATF2_06	6.3654	0.01115	1.95262
V$ATF5_01	7	7_116629642_116632141	116629642	116632141	V$ATF5_01	8.77476	0.0151	1.82098
V$ATOH1_08	7	7_116582625_116584621	116582625	116584621	V$ATOH1_08	6.62846	0.04876	1.31197
V$BCL6_01	7	7_116149618_116151432	116149618	116151432	V$BCL6_01	5.96846	9.3478E-5	4.02929

The structure of Table S3 is the same as for Table S1, except that it includes only the matrices which matching transcription factors appeared to have high cumulative binding affinity to the analyzed regulatory regions of the input gene.

Optimized profile construction

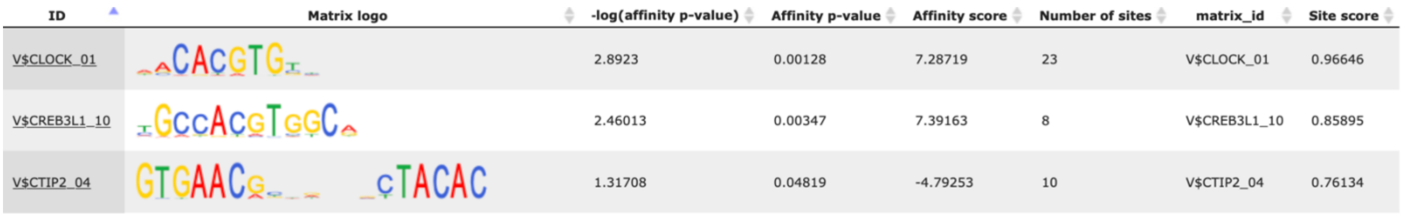
The filtered list of matrices presented in Supplementary Table 3 of the analysis report is then used to construct a new matrix profile that is specific for the analysis of the input gene in the selected conditions. This profile will be further used in the next steps of analysis for running the site search algorithm (MATCH). The profile is constructed using the Create profile from site model table method of the geneXplain platform using the cut-offs from the standard MATCH Suite profile consisting of 5752 TRANSFAC® vertebrate matrices [3,4]. The constructed profile is provided in the analysis report as Supplementary Table 4 (Table S4). 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 TFBSs

Search for TFBSs using the constructed profile

The constructed profile (see above) is used to find potential binding sites by applying the MATCH algorithm [7]. The respective method in the geneXplain platform is called TRANSFAC® Match™ for tracks. This tool predicts binding sites in regulatory regions of the studied gene using the constructed profile and the MATCH algorithm and outputs the track of found sites, which can be then viewed in genome browser interactive visualization of MATCH Suite single gene analysis results.

As a result of this step, the previously constructed table of identified matrices presented in Supplementary Table 3 of the analysis report, is extended with two additional columns: 'Site score' and 'Number of sites':



ID	Matrix logo	-log(affinity p-value)	Affinity p-value	Affinity score	Number of sites	matrix_id	Site score
V$CLOCK_01		2.8923	0.00128	7.28719	23	V$CLOCK_01	0.96646
V$CREB3L1_10		2.46013	0.00347	7.39163	8	V$CREB3L1_10	0.85895
V$CTIP2_04		1.31708	0.04819	-4.79253	10	V$CTIP2_04	0.76134

Number of sites is referring to the number of TFBS found for the respective matrix in the studied regulatory regions with TRANSFAC® Match™ for tracks. The Site score refers on the best MATCH score among all sites of the respective matrix. The matrix logo is displayed in addition. The respective table is presented in the analysis report as Supplementary table 5 (Table S5).

Identifying transcription factors regulating the input gene

Based on the matrices identified on the previous step of analysis (matrices from Supplementary Table 5 in the analysis report), the factor table is constructed, which lists all transcription factors (TFs) associated with the respective matrices (Supplementary table 6 (Table S6) 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 minimum affinity p-value coming from the best matrix of the factor
- 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.

If tissue was selected upon the analysis launch, factors with expression values less then 0.05 in the selected tissue are discarded from the constructed factors table, granting the inclusion only of those factors that are substantially associated with the selected tissue. The expression values used for this filtering were taken from Human Protein Atlas [8].

The columns denominations of Supplementary table 6 are as follows:

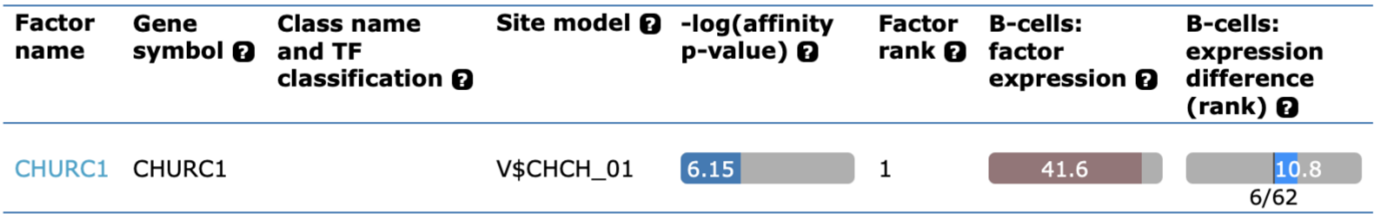
ID – factor ID
Genes: Ensembl ID –ID of gene corresponding to the respective TF
-**log(affinity p-value)** – -log(affinity p-value) value, where affinity p-value is taken from the best matrix, referring to the respective factor (matrix with minimum p-value)
Matrix_id – the ID of the best matrix, referring to the respective factor (matrix with minimum p-value)
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_id – factor ID according to TRANSFAC database
Factor name – name of TF
Gene symbol –symbol of the gene corresponding to the respective TF
Factor classification – as provided by TF Class
Family name – name of TF family as provided by TF Class
Tissue specificity – provides the value of general factor expression specificity described below
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

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
Affinity rank – ranking by affinity p-value (best rank 1 goes to factor with minimum p-value assigned to it)
Ranking by expression – ranking by Factor expression in tissue column
Rank tissue specificity – rank of factor expression in the selected tissue compared to all other supported tissues, if tissue was selected upon the analysis launch, or rank of factor average expression compared to all other supported tissues, if no tissue was selected upon the analysis launch.
Factor summed rank – provides the rank sum of the factor (calculation described above)

Having applied the factor ranking in respect to the abovementioned criteria, the top 40 factors by rank are selected for the factor table (Table 1 in the analysis report).

The example below shows part of the Factor view table (Table 1 in the analysis report) with B-cells selected as tissue in the analysis:

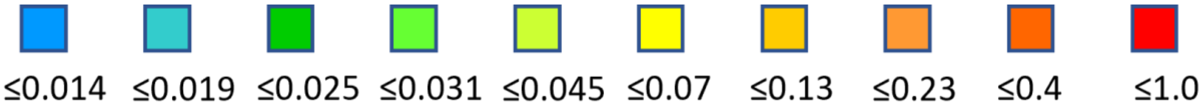


Factor name	Gene symbol ?	Class name and TF classification ?	Site model ?	-log(affinity p-value) ?	Factor rank ?	B-cells: factor expression ?	B-cells: expression difference (rank) ?
CHURC1	CHURC1		V$CHCH_01	6.15	1	41.6	10.8 6/62

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

The 'B-cells: expression difference (rank)' column shows three values:

(1) The expression deviation for the selected tissue from average (10.8 in the given example)
(2) The expression rank of the factor in the selected tissue in comparison to other supported tissues for this factor (6/62 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.

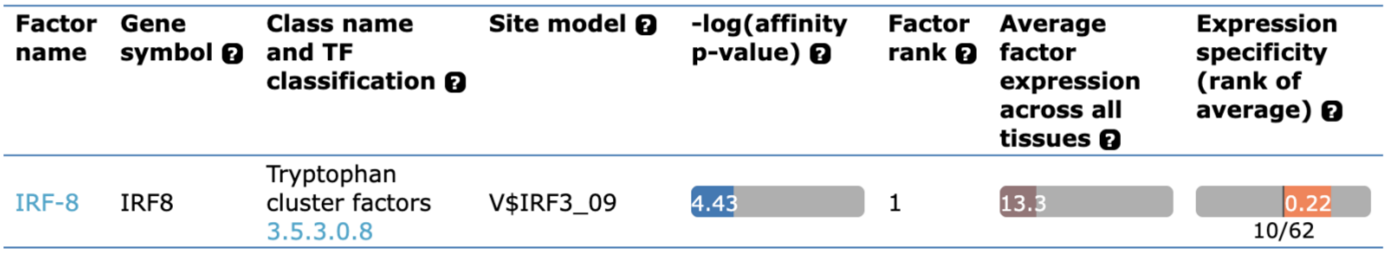
Expression rank

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. [9]. 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 table (Table 1 in the analysis report), when no tissue was specified in the analysis:

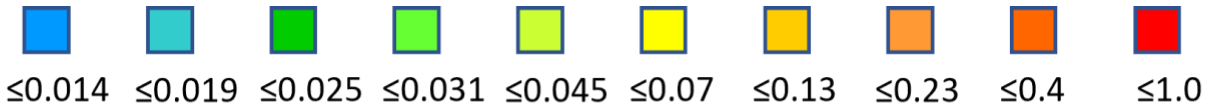


Factor name	Gene symbol	Class name and TF classification	Site model	-log(affinity p-value)	Factor rank	Average factor expression across all tissues	Expression specificity (rank of average)
IRF-8	IRF8	Tryptophan cluster factors 3.5.3.0.8	V$IRF3_09	4.43	1	13.3	0.22 10/62

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 (10/62 in the given example)
(2) The general factor expression specificity level represented by color (🟧) and value (0,22) using the following color-value code:



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

The site model column shows the best matrix corresponding to the given factor (the matrix with minimum affinity p-value.

The -log(affinity p-value) value column refers to the respective value, where affinity p-value is taken from the best matrix of the factor.

Class name and TF classification column shows the respective values for the given transcription factor as provided by the TF Class [6].

The matrices corresponding to factors predicted to be regulating the input gene in the specified conditions (factors from Table 1 in the analysis report) are then listed in Table 2 of the analysis report. These are the most relevant matrices (site models) that determine the regulation of the analyzed gene.

References

[1] Wikipedia entry for the Jaccard index. Link

[2] 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.

[3] 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.

[4] Lloyd, K., Papoutsopoulou, S., Smith, E., Stegmaier, P., Bergey, F., Morris, L., Kittner, M., England, H., Spiller, D., White, M.H. and Duckworth, C.A. (2020) Using systems medicine to identify a therapeutic agent with potential for repurposing in inflammatory bowel disease. Disease models & mechanisms, 13(11), p.dmm044040. PubMed.

[5] 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.

[6] 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.

[7] 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.

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

[9] 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 single gene 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).