Report on gene regulation analysis

Demo User
geneXplain GmbH
info@genexplain.com
Data received on 04/07/2024; Run on 04/07/2024; Report generated on 04/07/2024
MATCH Suite release 3.0 (TRANSFAC® 2.0 release 2024.1)

Summary

The MATCH Suite software package was applied to the rat input gene list demo-gene-list that contained 144 genes. The goal of the analysis was to identify the transcription factors that regulate the studied set of genes through their −500/+100 promoter region.

The most promising transcription factors regulating the input gene set appeared to be factors: **KLF10, KLF15, C-MYC, CREB, EGR2, ASH-2, CREB-3L2, SMAD3, PAX-6, PAX-4**. The selection of these factors was based on a complex criteria that includes: (1) the statistically significant enrichment of the binding sites for these factors in the promoters of the studied set of genes, (2) combinatorial effect of several TFs acting together binding to co-localized binding sites in these promoters, and (3) GO enrichment of the genes encoding those transcription factors among the following GO terms selected for the analysis launch:

- <GO:0007507> heart development
- <GO:0007517> muscle organ development
- <GO:0009888> tissue development
- <GO:0010720> positive regulation of cell development

At least 104 genes from your input set are regulated by these factors.

Results overview

Transcription Factors Identified

In the promoters of the analyzed genes, potential transcription factor binding sites were identified and checked for (a) statistical enrichment and (b) for overrepresented combinations (see Methods).

From these analyses, the transcription factors presented in the Table 1 were identified as the most probable regulators of the studied gene set.

Table 1. Key transcription factors identified as the potential regulators of the analyzed gene set.

Factor name	Enrichment analysis	Combinatorial analysis	GO enrichment score
KLF10	●	✓	220.12 2/4
KLF15	●	✓	173.97 1/4
C-MYC	●	✓	283.48 3/4
CREB	●	✓	331.23 4/4
EGR2	●	–	283.48 3/4
ASH-2	●	✓	173.97 1/4
CREB-3L2	●	✓	173.97 1/4
SMAD3	●	–	285.08 3/4
PAX-6	●	–	220.12 2/4
PAX-4	●	–	173.97 1/4
HAND-1	●	✓	285.08 3/4
TWIST-1	●	✓	285.08 3/4
EGR3	●	–	46.15 1/4
GCMA	●	–	173.97 1/4
SMAD2	●	–	221.72 2/4
SMAD5	●	–	173.97 1/4
PTF-1A	●	✓	173.97 1/4
RUNX1	●	–	220.12 2/4
RUNX3	●	–	220.12 2/4
RXRGAMMA	●	–	285.08 3/4

View full table →

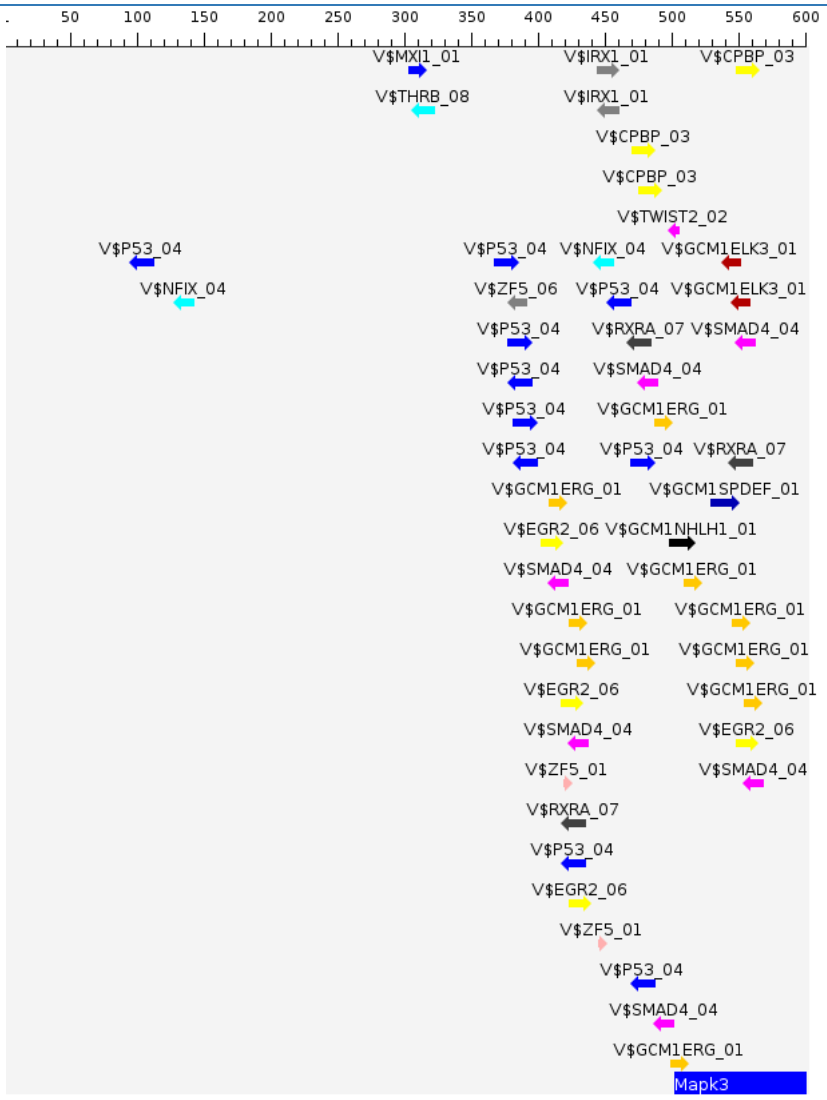
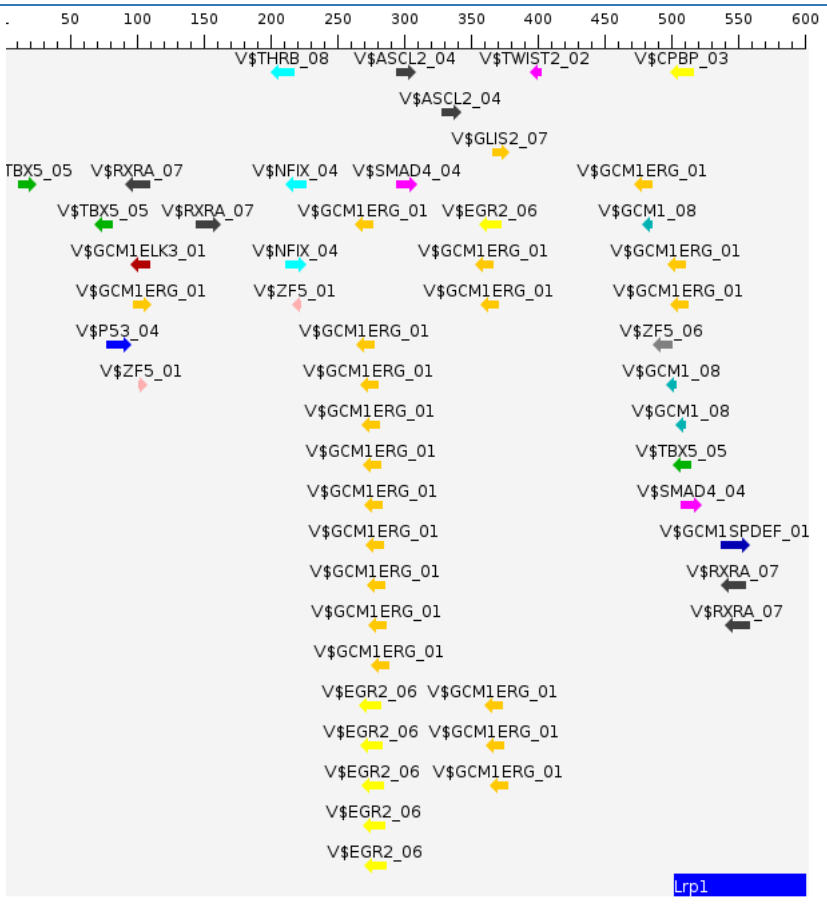
Table 1 shows the transcription factors the binding sites of which are enriched in the analyzed promoters. The sites that have sequence enrichment FDR less than 0.05 are marked with green color in the column 'Enrichment analysis', the rest are marked with blue. Those factors that additionally are part of a combinatory module are marked by a tick in the

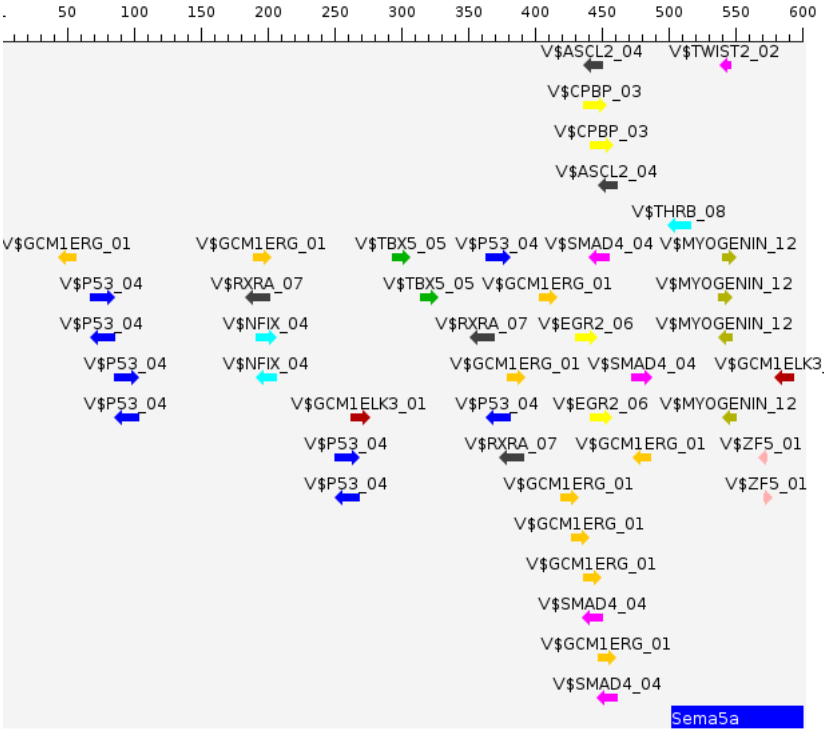
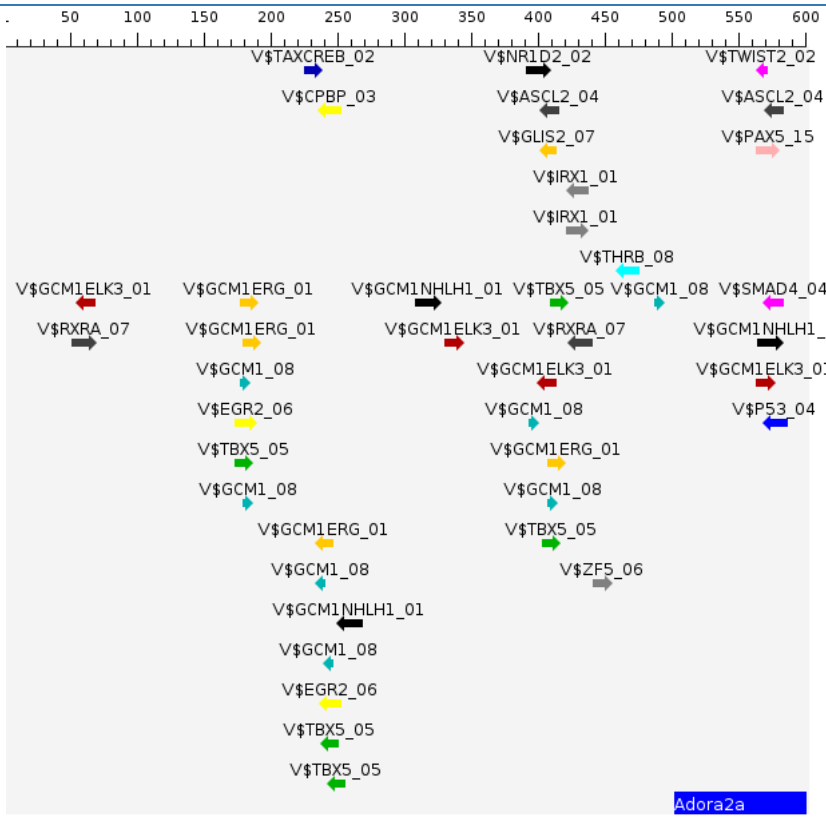
'Combinatorial Analysis' column. The 'GO enrichment' column shows the factor's enrichment in the GO functional categories selected for the analysis launch. The value below represents the number of GO categories to which this factor belongs to among the total number of GO categories selected for the analysis launch.

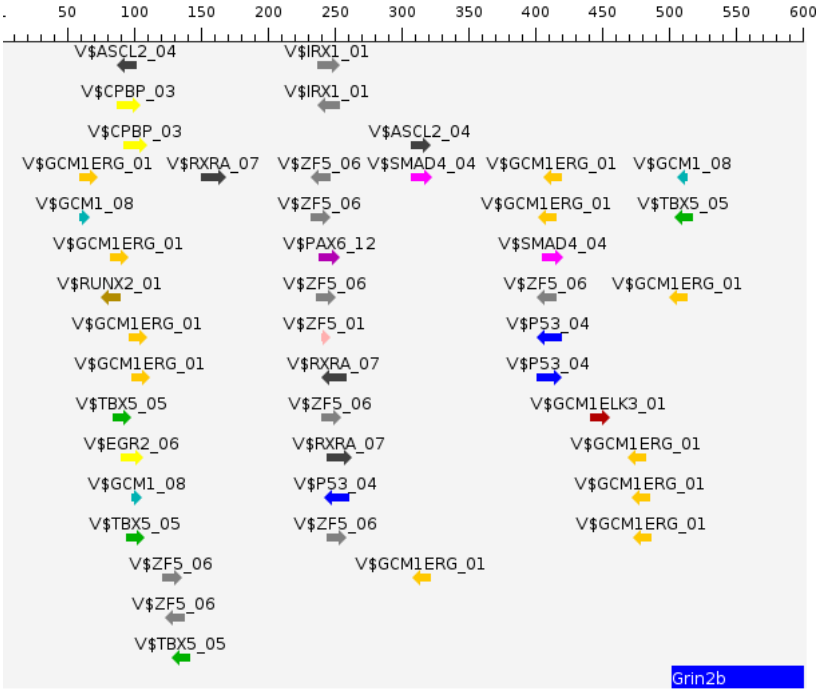
Regulation of Analyzed Genes

Not all transcription factors identified act equally on all genes in the analyzed gene set. Table 2 shows the analyzed genes with respective site models (PWMs) found in their promoters. The target genes are ranked according to the enrichment of their promoters by binding sites of the best ranked transcription factors. The visualization shows the binding sites of matrices from CMA model and all other enriched sites.

Table 2. Top regulated genes from the analyzed gene set

ID	Gene symbol	Sites view	CMA Score
ENSRNOG00000053583	Mapk3		6.23
ENSRNOG00000025053	Lrp1		4.06

ENSRNOG00000011977	Sema5a		5.41
ENSRNOG00000001302	Adora2a		6.16

ENSRNOG00000008766	Grin2b		3.56

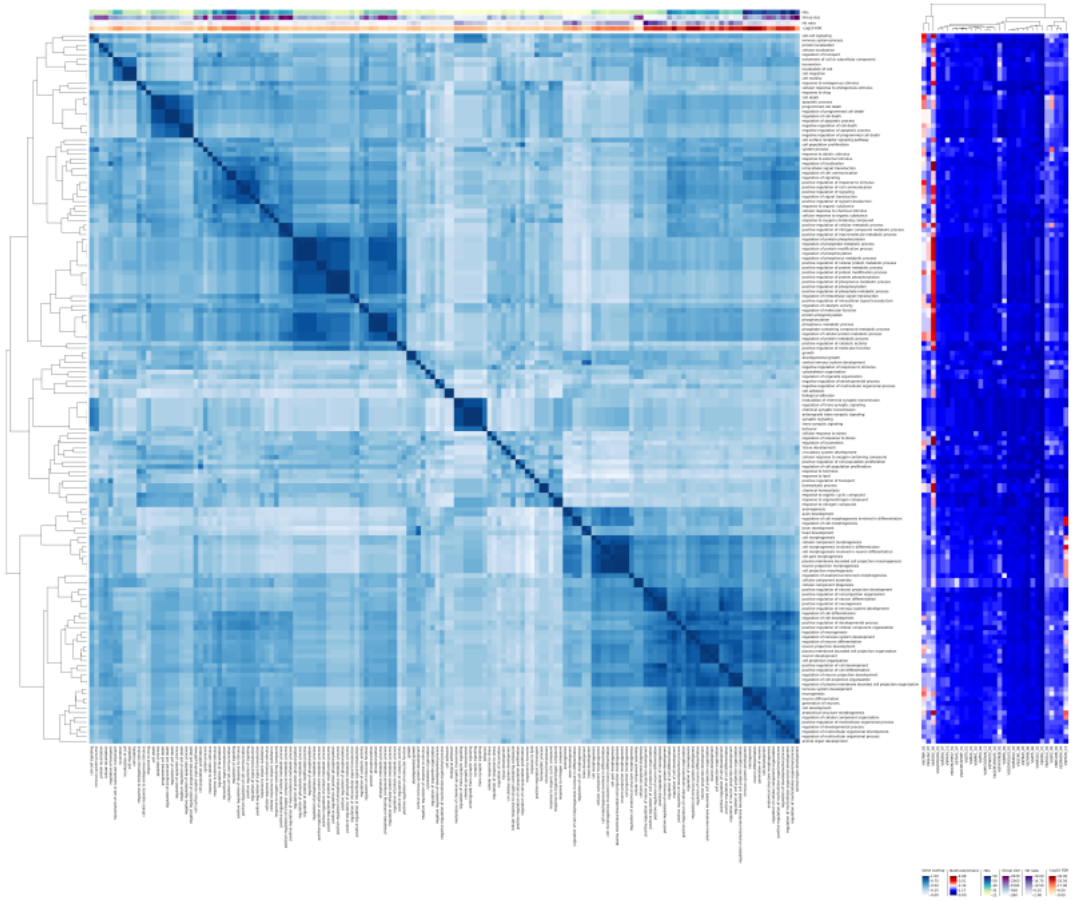
View full table →

Functional Analysis of Gene Regulation

The input gene set was categorized to the top GO terms overrepresented in the studied set of genes.

The promoters of the genes that belong to each of these GO categories were analyzed for enriched transcription factor binding sites. The heatmap on the right side of the Figure 1 represents the spread of the binding site motifs (TRANSFAC® positional weight matrices) that were found to be enriched in the promoters of genes from respective GO category.

Figure 1. On the left: heatmap of GO to GO terms mapping for the GO terms overrepresented among the studied gene set; On the right: heatmap visualizing how enriched motifs are associated with the respective GO categories.

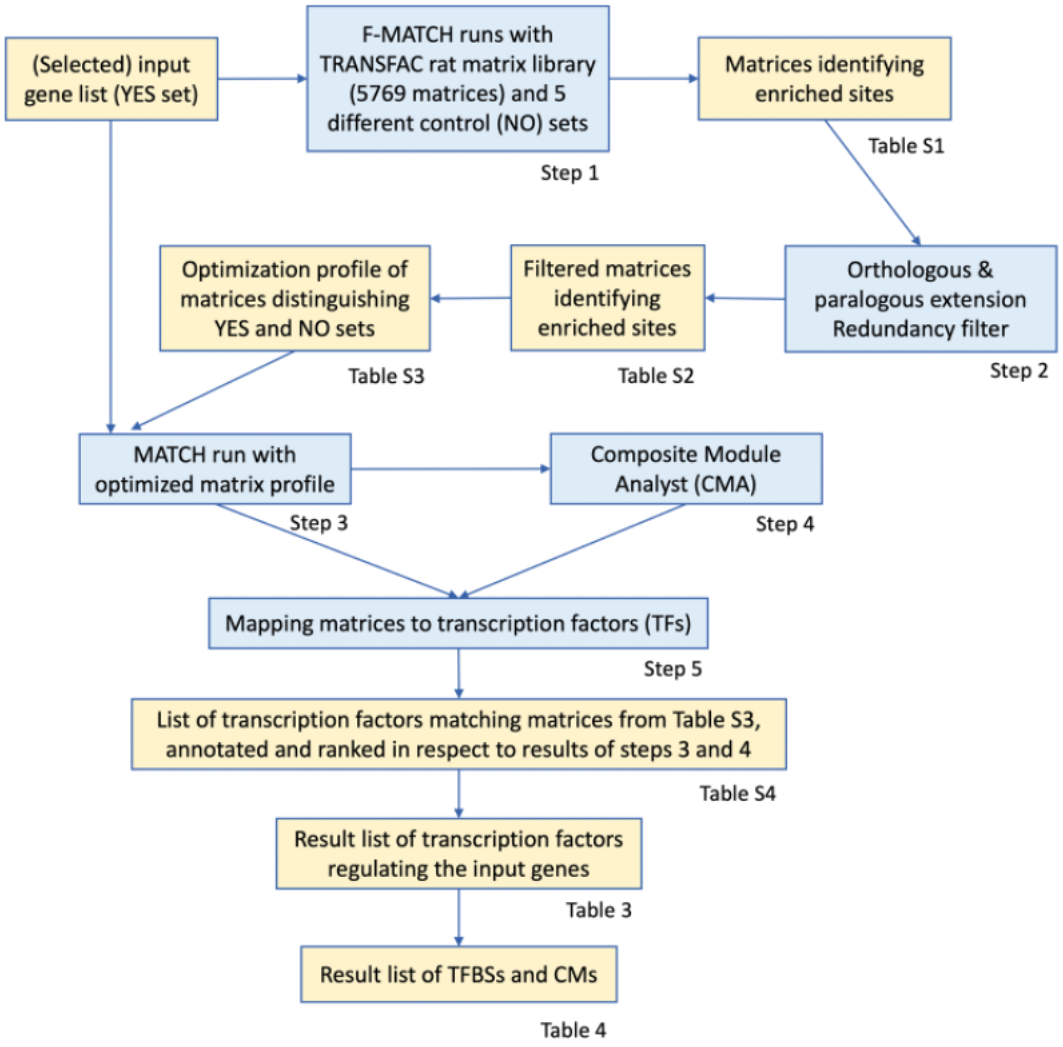


Alternative (vertical) layout of the heatmaps presented on Figure 1 can be found here

Results details

The analysis workflow of the MATCH Suite comprises 5 steps, which are outlined in Figure 2. Each step produces a table either of intermediary or of final results, available as Supplementary table following the links given, or as part of this report, respectively.

Figure 2. Overall schema of the MATCH Suite workflow



Step 1

The promoters of the input gene set selected for specified GO categories were extracted using the Ensembl best supported promoter. The promoter regions used were taken as −500 bp upstream of transcription start site (TSS) and +100 bp downstream the TSS. The extracted promoters were compiled to one track of regulatory regions, which was then used to perform site enrichment analysis using the TRANSFAC® library of positional weight matrices (PWMs). It is done according to the PWM cut-off optimization approach published earlier [1,2]. The geneXplain platform method 'Search for enriched TFBS on tracks' [2] was used here to analyze site enrichment in the promoters of the input gene set.

The YES set (analysis set) applied in this analysis consisted of the promoters of the input genes, while five NO sets (control/ background sets), each comprising 1000 promoters, were randomly sampled from genes not belonging to your input gene set. Transcription factor binding sites (TFBSs) enriched in the YES set were identified by 5 independent F-MATCH runs, each using one of the five NO sets. The matrix profile used for identifying potential TFBSs is the collection of 8067 TRANSFAC® rat matrices [3, 4].

Since the following GO terms were specified upon the analysis launch for the profile optimization:

- <GO:0007507> heart development
- <GO:0007517> muscle organ development
- <GO:0009888> tissue development
- <GO:0010720> positive regulation of cell development

The matrices which do not have corresponding transcription factors belonging to those GO terms are excluded from the used profile.

For each matrix of the profile, 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 [2].

Supplementary table 1 (Table S1) comprises the results of this analysis, by default sorting the matrices according to the enrichment of sites they identified. 494 transcription factor binding site models (matrices) exhibited a site enrichment higher than 1 (lower boundary of the 99% confidence interval of the site enrichment). See Methods for a detailed explanation of the contents of this table.

Step 2

The list of matrices is then extended to matrices associated with 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 can be found here. See Methods for a detailed explanation of the orthologous and paralogous extension applied.

Finally, redundancy elimination is done by selecting just one matrix for each factor cluster - the one that maximizes the adjusted site enrichment value.

The resulting list of matrices after orthologous and paralogous extension and redundancy filtering is shown in Supplementary table 2 (Table S2). It comprises 31 matrices. A detailed explanation of the values shown in Table S2 is given in the Methods section.

The filtered list of matrices (Table S2) is used to construct a new matrix profile that is specific for the analysis of the input gene set and will be used for the next analyses. The cut-offs for the profile are selected as the median site cutoff of matrices from Table S2. The complete profile is shown in Supplementary table 3 (Table S3).

Step 3

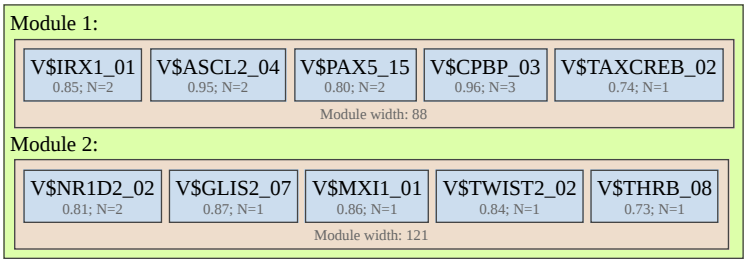
With the next step, the MATCH Suite workflow uses the constructed profile from Step 2 for another search for potential transcription factor binding sites (TFBS) in the set of studied genes. The method uses the MATCH algorithm (see [5] and Methods for further info) and generates tracks of sites found in the YES and NO set.

Step 4

Subsequently, the MATCH Suite workflow searches for composite modules of the predicted TFBSs. Composite modules are combinations of several TFBSs that are found together in a set of regulatory sequences. Combinations of TF binding sites that are overrepresented in the regulatory regions of the genes in the YES compared to those in the NO set are identified by the Composite Module Analyst (CMA - see [6] and the Methods for further info). The genetic algorithm takes the output from the site search in Step 3 as input and comes up with a resulting composite module that differentiates the YES set from the background NO set. CMA identifies the transcription factors that through their cooperation may provide a synergistic effect and thus have a great influence on the gene regulatory process.

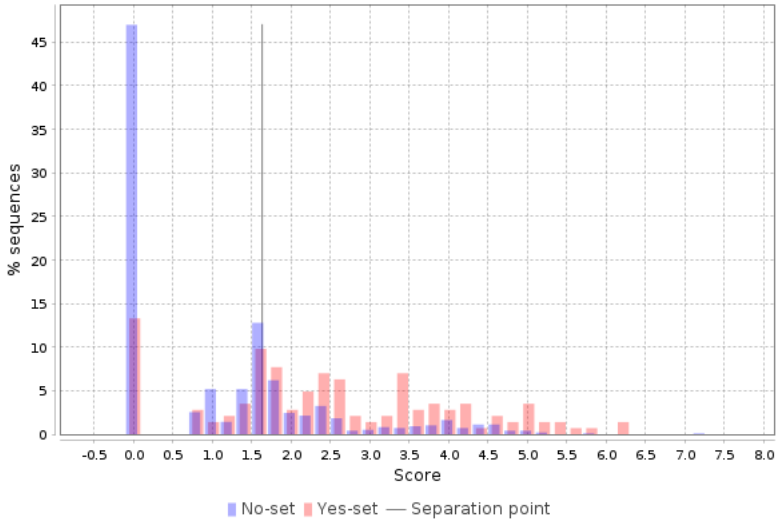
Figure 3 shows the CMA model constructed on the basis of found YES and NO sites tracks. The obtained CMA model is then applied to compute CMA score for all of the genes from the input set.
The model consists of 2 module(s). Below, for each module the following information is shown:
- PWMs producing matches,
- number of individual matches for each PWM,
- score of the best match.

Figure 3. The constructed CMA model



Model score (-p*log10(pval)): 11.47
Wilcoxon p-value (pval): 1.28e-23
Penalty (p): 0.501
Average yes-set score: 2.49
Average no-set score: 1.11
AUC: 0.76
Separation point: 1.64
False-positive: 30.12%
False-negative: 27.27%

Figure 4. Distribution of CMA scores for individual promoters



The histogram from Figure 4 shows the distribution of scores for individual promoters 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. The center, a vertical gray line (separation point), corresponds to the average score value. Promoters from the NO set with a score above the separation line, i.e. blue bars to the right of the gray line, are referred to as false positives. Promoters from the YES set with a score below the separation line (red bars to the left of the gray line), are false negatives.

The YES promoters with a score above the separation line are very well separated from the NO promoters, which means that for this part of the promoters the composite model constructed is most suitable. See Methods section for further info.

Step 5

The transcription factors associated with the matrices received on steps 3 and 4 are given in Supplementary table 4 (Table S4). The sorting of the factors takes into account the combinatorial score and enrichment value of the respective binding sites, as well as factor's enrichment in the GO categories selected for profile optimization upon the analysis launch. Table 3 contains the top 30 factors from Supplementary table 4 together with the factors corresponding to the matrices from the CMA model. These factors are predicted to be regulating the input gene set.

Table 3. Transcription factors that regulate the analyzed gene set.

Factor name	Gene symbol	Class name and TF classification	Site model	Adjusted factor enrichment	GO enrichment score	Factor rank
KLF10	Klf10	C2H2 zinc finger factors 2.3.1.2.10	**V$CPBP_03**	2.13	220.12 2/4	1
KLF15	Klf15	C2H2 zinc finger factors 2.3.1.2.15	**V$CPBP_03**	2.13	173.97 1/4	2
C-MYC	Myc	Basic helix-loop-helix factors (bHLH) 1.2.6.5.1	**V$MXI1_01**	1.50	283.48 3/4	3
CREB	Creb1	Basic leucine zipper factors (bZIP) 1.1.7.1.1	**V$TAXCREB_02**	1.26	331.23 4/4	4
EGR2	Egr2	C2H2 zinc finger factors 2.3.1.3.2	V$EGR2_06	1.78	283.48 3/4	5
ASH-2	Ascl2	Basic helix-loop-helix factors (bHLH) 1.2.2.2.2	**V$ASCL2_04**	1.47	173.97 1/4	6
CREB-3L2	Creb3l2	Basic leucine zipper factors (bZIP) 1.1.7.2.3	**V$TAXCREB_02**	1.26	173.97 1/4	7
SMAD3	Smad3	SMAD/NF-1 DNA-binding domain factors 7.1.1.1.3	V$SMAD4_04	1.34	285.08 3/4	8
PAX-6	Pax6	Paired box factors 3.2.1.2.2	V$PAX6_12	1.73	220.12 2/4	9
PAX-4	Pax4	Paired box factors 3.2.1.2.1	V$PAX6_12	1.73	173.97 1/4	10

View full table →

In addition to the TF name, its gene symbol, its numerical identifier in the TF classification [7] and the description of the class it belongs to, Table 3 also shows all matrices from steps 3 and 4 that refer to the respective factor in the 'Site model' column. The matrices that belong to the CMA combinatorial modules are displayed in bold. The factor enrichment value is derived from the maximum enrichment value among all matrices referring to the factor. The length of the bar in the 'Factor enrichment' column is proportional to the maximum site enrichment value of the factor's matrices. The color of this bar is green if the maximum sequence enrichment FDR of the factor's matrices is less than 0.05 and blue otherwise.

The matrices corresponding to the factors from Table 3 are listed in Table 4. These are the most relevant matrices (site models) that determine the regulation of the analyzed gene set.

Table 4. Resulting matrices (site models) table

ID	Matrix logo	Site enrichment (adjusted enrichment)	Site enrichment FDR	Adjusted sequence enrichment	Sequence enrichment FDR	Composite model	Site rank
V$CPBP_03		3.53 (2.13)	9.78E-10	1.83	7.61E-6	yes	1
V$MXI1_01		3.42 (1.50)	1.46E-3	1.67	3.51E-4	yes	2
V$ASCL2_04		2.33 (1.47)	2.63E-9	1.21	8.65E-5	yes	3
V$TAXCREB_02		2.31 (1.26)	2.23E-4	1.11	2.76E-3	yes	4
V$TWIST2_02		1.51 (1.16)	3.21E-12	0.98	2E-5	yes	5
V$IRX1_01		1.54 (1.15)	1.41E-7	1.2	2.04E-6	yes	6
V$THRB_08		1.54 (1.12)	8.89E-6	0.94	9.56E-4	yes	7
V$NR1D2_02		1.58 (1.07)	9.95E-4	0.9	2.46E-3	yes	8
V$EGR2_06		3.41 (1.78)	1.09E-5	1.3	9.42E-6		9
V$PAX6_12		3.98 (1.73)	4.65E-4	1.74	7.32E-6		10
V$GCM1SPDEF_01		3.99 (1.70)	7.25E-5	1.59	3.76E-5		11
V$SMAD4_04		1.85 (1.34)	8.19E-12	1.17	1.77E-5		12
V$GCM1ERG_01		2.18 (1.33)	1.3E-14	1.08	2.16E-4		13
V$ZF5_01		1.79 (1.32)	2.27E-12	1.2	2.41E-7		14
V$RUNX2_01		2.36 (1.30)	8.66E-4	1.23	1.64E-3		15
V$RXRA_07		1.62 (1.22)	2.96E-12	1.02	1.01E-5		16
V$GCM1ELK3_01		1.63 (1.22)	3.09E-10	0.98	7.75E-4		17
V$NFIX_04		1.92 (1.20)	7.49E-4	0.94	8.83E-3		18
V$GCM1NHLH1_01		1.86 (1.19)	2.14E-4	0.94	2.44E-4		19
V$ZF5_06		1.69 (1.17)	4.07E-5	0.88	6.58E-3		20

View full table →

The list of all genes from the input set, their CMA scores, the total number of identified TFBSs and the hits obtained with each site model are presented in Table 5. Underneath each matrix name, the TFs referring to it are given in the order of their ranks. The ranking of genes in done according to their CMA scores.

Table 5. Gene table with the identified transcription factors and their site models regulating the genes from the analyzed gene set

Ensembl ID	Gene symbol	Gene description	CMA Score	Total number of sites	V$ASCL2_04 ASH-2	V$CPBP_03	V$EGR2_06	V$GCM1ELK3_01 GCMA	V$GCM1ERG_01 GCMA	V$GCM1NHLH1_01 GCMA
ENSRNOG00000053583	Mapk3	mitogen activated protein kinase 3	6.23	51	0	3	4	2	9	1
ENSRNOG00000001302	Adora2a	adenosine A2a receptor	6.16	50	3	1	2	4	4	3
ENSRNOG00000013953	Ntrk1	neurotrophic receptor tyrosine kinase 1	5.84	53	2	2	0	1	4	0
ENSRNOG00000023861	Snap91	synaptosome associated protein 91	5.62	50	2	1	2	1	5	4
ENSRNOG00000011977	Sema5a	semaphorin 5A	5.41	49	2	3	2	2	9	0
ENSRNOG00000007948	Nf2	neurofibromin 2	5.39	23	1	1	0	0	1	0
ENSRNOG00000004621	Rtn4	reticulon 4	5.24	30	1	1	0	2	5	0
ENSRNOG00000001412	Epo	erythropoietin	5.2	40	0	0	1	3	5	1
ENSRNOG00000011475	Srcin1	SRC kinase signaling inhibitor 1	5.06	40	1	2	1	5	6	2
ENSRNOG00000016571	Ngf	nerve growth factor	5.05	31	1	1	1	1	5	0
ENSRNOG00000058898	Nedd4	NEDD4 E3 ubiquitin protein ligase	5.02	37	3	2	0	3	1	0
ENSRNOG00000003029	Calr	calreticulin	4.97	34	1	1	0	0	2	0
ENSRNOG00000005519	Grm3	glutamate metabotropic receptor 3	4.95	33	1	2	0	3	4	1
ENSRNOG00000028156	Pld1	phospholipase D1	4.88	31	0	2	0	1	2	1
ENSRNOG00000001849	Mapk1	mitogen activated protein kinase 1	4.71	35	0	0	1	2	6	0
ENSRNOG00000052296	Shank3	SH3 and multiple ankyrin repeat domains 3	4.6	39	1	1	0	6	10	0
ENSRNOG00000009540	Gpr3	G protein-coupled receptor 3	4.55	46	2	2	0	3	8	0
ENSRNOG00000058202	Ppp2r2c	protein phosphatase 2, regulatory subunit B, gamma	4.54	44	0	0	0	3	4	1
ENSRNOG00000010881	Trak2	trafficking kinesin protein 2	4.49	38	1	0	0	1	3	1
ENSRNOG00000003164	Pla2g10	phospholipase A2, group X	4.3	23	1	1	1	0	2	1

View full table →

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[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.

How to cite

Please use the results received with the MATCH Suite in your publications or presentations with the following reference:

The results were obtained with the MATCH Suite software integrated into the TRANSFAC® 2.0 solution for gene regulation analysis release 3.0 (https://genexplain.com/transfac).

Please also provide reference to the following publication:

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.

Disclaimer