

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 ?	Skeletal muscle: factor expression ?	Skeletal muscle: expression difference (rank) ?
MEF-2C		✓	154.8	142.2 1/62
MEF-2D		✓	97.8	80.3 1/62
TIEG-1		–	88.6	70.2 1/62
DB1		–	51.7	27.4 1/62
Mef-2a		✓	46.1	26.7 2/62
KLF15		–	41.0	29.2 2/62
BTEB1		–	38.9	23.8 2/62
ZNF511		–	25.9	11.3 2/62
NF-IX		–	139.7	119.7 1/62
NF-1C		–	101.7	85.0 1/62
BCL-6		–	114.3	92.1 1/62
CPBP		–	57.8	30.9 6/62
znf414		–	23.6	10.0 3/62
E2F-6		✓	17.7	12.3 1/62
Sp2		–	30.2	11.9 3/62
LRF		–	32.0	17.2 1/62
SRC-1		–	42.5	31.2 1/62
USF2		✓	108.9	74.3 1/62
LKLF		–	43.7	21.0 9/62
MAZ		–	41.0	6.3 12/62

[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 'skeletal muscle: expression difference (rank)' column shows the difference between factor expression in skeletal muscle tissue and the average value of factor expression among all tissues. The color of the bar refers to the factor's general expression entropy and uses the following color code:



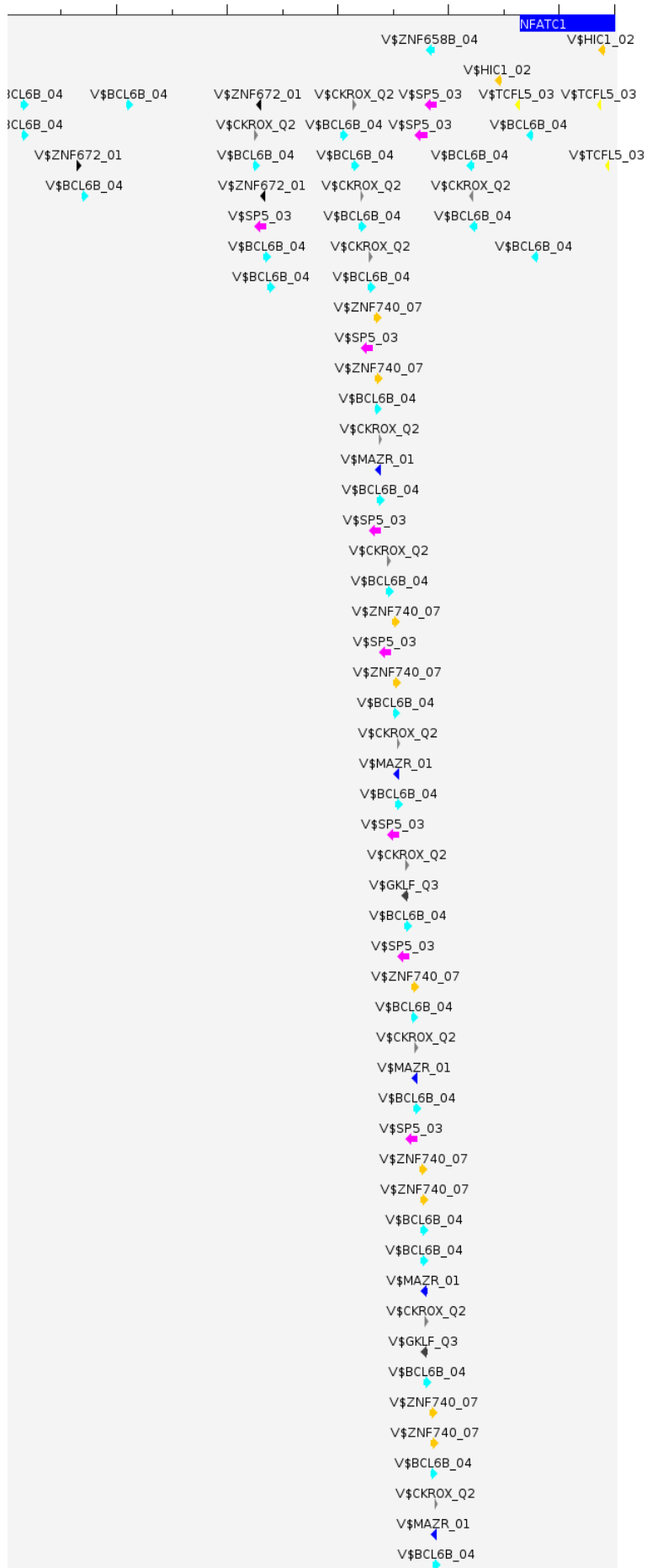
The rank below represents the rank of factor expression in skeletal muscle out of the number of supported tissues for each factor and factor's average expression value in addition to them. The maximum number of currently supported tissues for one factor is 61.

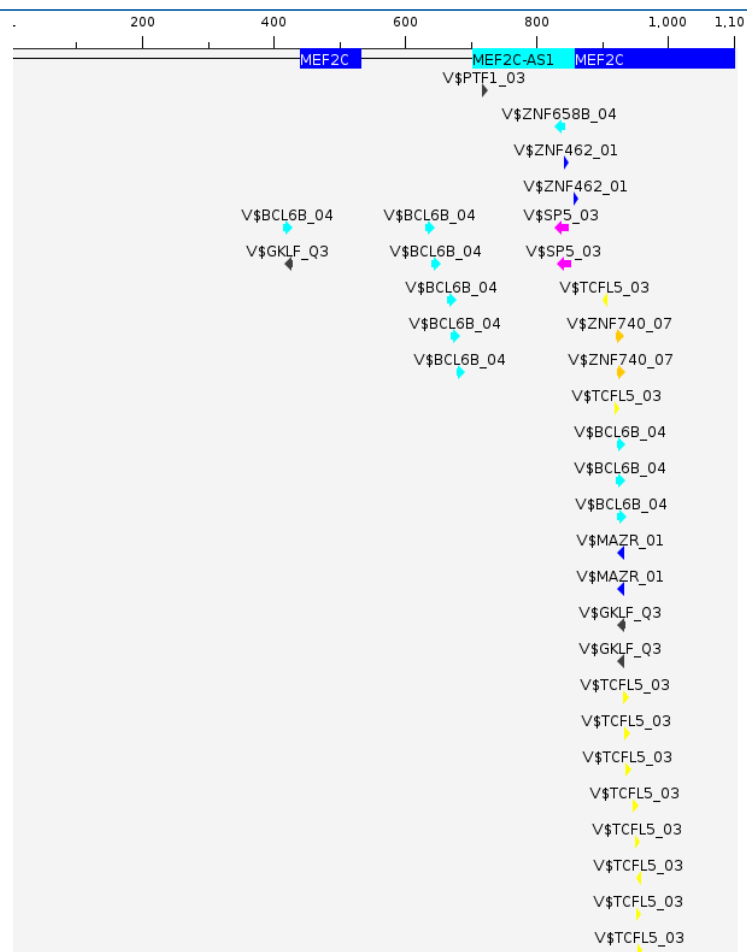
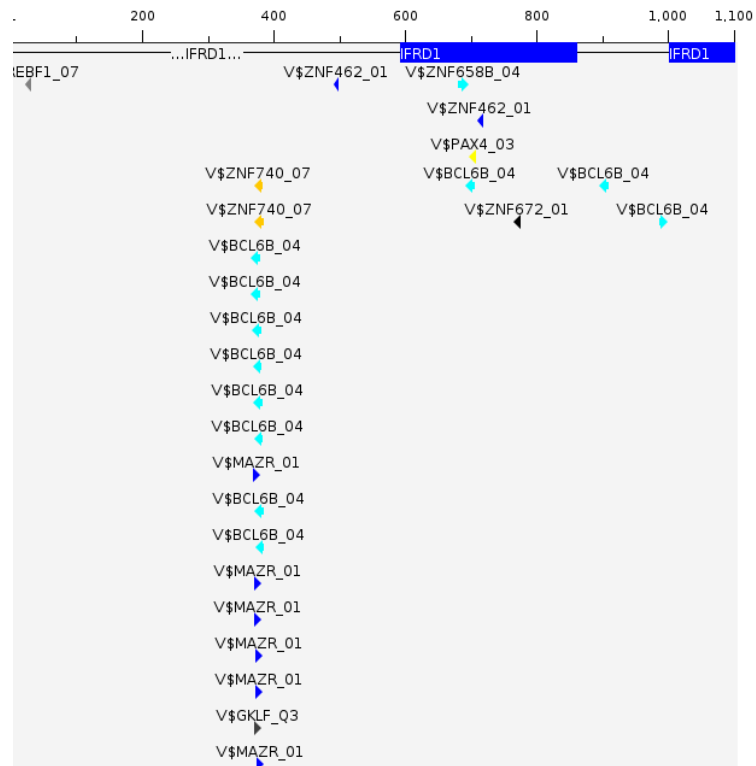
Regulation of Analyzed Genes

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
ENSG00000188130	MAPK12		3.93
ENSG00000131196	NFATC1		1.67

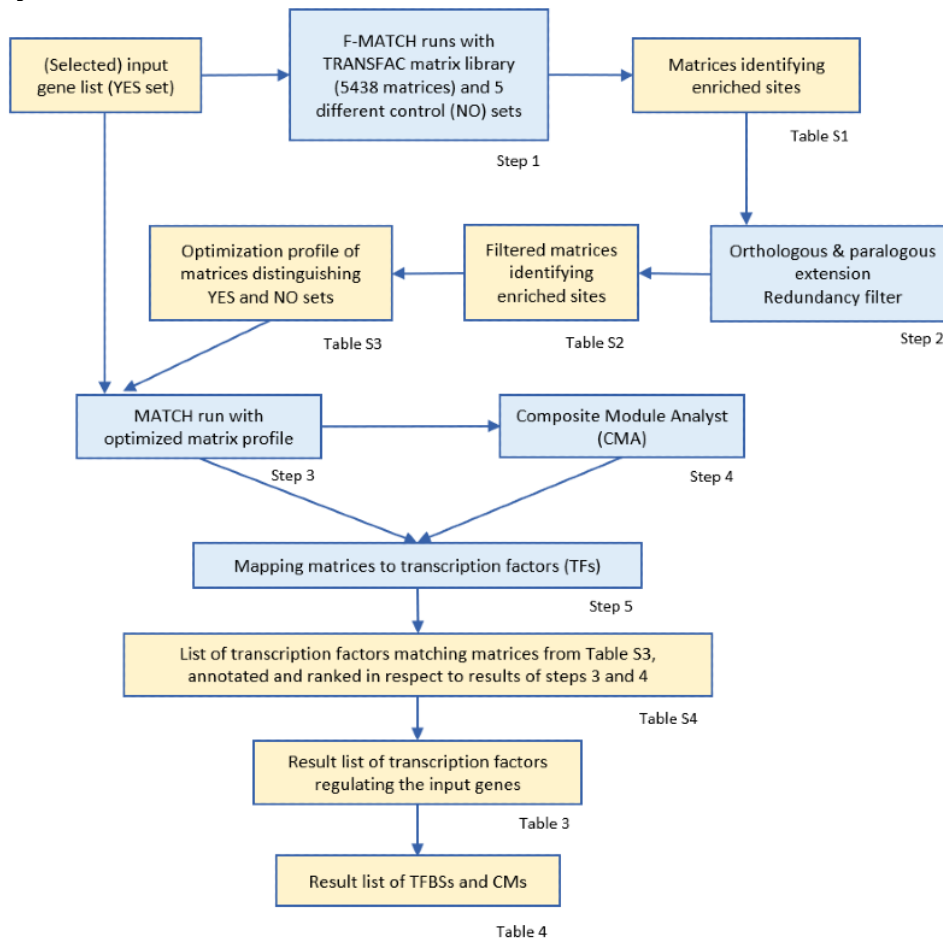






The analysis workflow of the MATCH Suite comprises 5 steps, which are outlined in Figure 2. Each step produces a table either of intermediary or of final results, available as Supplementary table following the links given, or as part of this report, respectively.

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 FANTOM5 database promoters for the skeletal muscle tissue. The promoter regions used were taken as −1000 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 run the F-MATCH analysis [1] (search for enriched TFBS on track).[2]

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 5438 TRANSFAC® vertebrate matrices [3, 4]. 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 [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. 715 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 (PWMs) from Step 1 is filtered by expression in skeletal muscle tissue: matrices that don't represent any factors that are expressed in skeletal muscle tissue are removed from the table.

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 the tissue filtering, orthologous and paralogous extension and redundancy filtering is shown in Supplementary table 2 (Table S2). It comprises 83 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

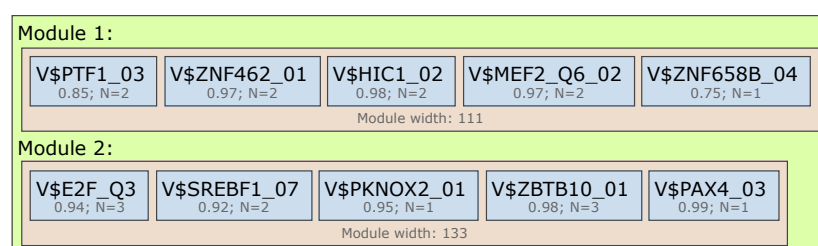
Subsequently, the MATCH Suite workflow searches for composite modules of the predicted TFBSs. Composite modules are combinations of several TFBSs that are found together in a set of regulatory sequences. Combinations of TF binding sites that are overrepresented in the regulatory regions of the genes 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)): 10.53

Wilcoxon p-value (pval): 9.86e-22

Penalty (p): 0.501

Average yes-set score: 2.43

Average no-set score: 0.58

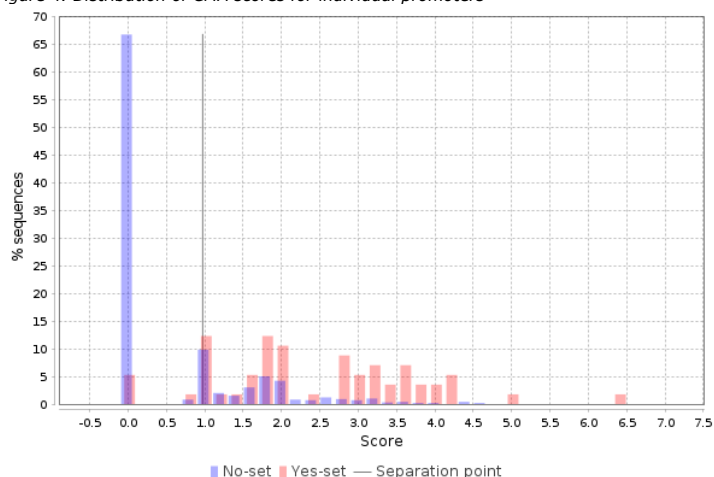
AUC: 0.87

Separation point: 0.97

False-positive: 30.23%

False-negative: 7.02%

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 expression values and expression specificity of factors in the skeletal muscle tissue. 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	Factor rank	Skeletal muscle: factor expression	Skeletal muscle: expression difference (rank)
MEF-2C	MEF2C	MADS box factors 5.1.1.1.3	V\$MEF2_Q6_02	<div><div>2.15</div></div>	1	<div><div>154.8</div></div>	<div><div>142.2</div></div> 1/62
MEF-2D	MEF2D	MADS box factors 5.1.1.1.4	V\$MEF2_Q6_02	<div><div>2.15</div></div>	2	<div><div>97.8</div></div>	<div><div>80.3</div></div> 1/62
TIEG-1	KLF10	C2H2 zinc finger factors 2.3.1.2.10	V\$GKLF_Q3	<div><div>2.82</div></div>	3	<div><div>88.6</div></div>	<div><div>70.2</div></div> 1/62
DB1	VEZF1	C2H2 zinc finger factors 2.3.4.8.2	V\$MAZR_01	<div><div>2.85</div></div>	4	<div><div>51.7</div></div>	<div><div>27.4</div></div> 1/62
Mef-2a	MEF2A	MADS box factors 5.1.1.1.1	V\$MEF2_Q6_02	<div><div>2.15</div></div>	5	<div><div>46.1</div></div>	<div><div>26.7</div></div> 2/62
KLF15	KLF15	C2H2 zinc finger factors 2.3.1.2.15	V\$GKLF_Q3	<div><div>2.82</div></div>	6	<div><div>41.0</div></div>	<div><div>29.2</div></div> 2/62
BTEB1	KLF9	C2H2 zinc finger factors 2.3.1.2.9	V\$GKLF_Q3	<div><div>2.82</div></div>	7	<div><div>38.9</div></div>	<div><div>23.8</div></div> 2/62
ZNF511	ZNF511	C2H2 zinc finger factors 2.3.2.4.2	V\$ZNF740_07	<div><div>3.24</div></div>	8	<div><div>25.9</div></div>	<div><div>11.3</div></div> 2/62
NF-IX	NFIX	SMAD/NF-1 DNA-binding domain factors 7.1.2.0.4	V\$NF1C_03	<div><div>1.54</div></div>	9	<div><div>139.7</div></div>	<div><div>119.7</div></div> 1/62
NF-1C	NFIC	SMAD/NF-1 DNA-binding domain factors 7.1.2.0.3	V\$NF1C_03	<div><div>1.54</div></div>	10	<div><div>101.7</div></div>	<div><div>85.0</div></div> 1/62

[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\$HIC1_02		4.88 (2.16)	5.44E-5	1.73	1.58E-3	yes	1
V\$MEF2_Q6_02		5.51 (2.15)	4.25E-4	1.95	3.38E-3	yes	2
V\$E2F_Q3		4.55 (1.92)	3.98E-4	1.69	4.33E-3	yes	3
V\$ZNF658B_04		2.98 (1.85)	1.09E-12	1.47	2.65E-4	yes	4
V\$PKNOX2_01		3.48 (1.53)	2.12E-3	1.63	2.17E-3	yes	5
V\$ZNF462_01		2.14 (1.32)	1.74E-6	1.11	1.98E-3	yes	6
V\$ZBTB10_01		2.36 (1.30)	3.33E-6	1.07	2.77E-3	yes	7
V\$PTF1_03		1.95 (1.13)	1.23E-3	1.09	2.82E-3	yes	8
V\$PAX4_03		1.97 (1.13)	1.63E-3	1.09	3.7E-3	yes	9
V\$SREBF1_07		1.88 (1.13)	5.42E-4	1.01	2.86E-3	yes	10
V\$ZNF740_07		6.10 (3.24)	2.99E-12	1.6	3.48E-3		11
V\$MAZR_01		5.25 (2.85)	3.2E-11	1.33	6.22E-3		12
V\$GKLF_Q3		5.96 (2.82)	1.66E-7	2.09	2.23E-3		13
V\$SP5_03		5.28 (2.74)	2.36E-9	1.49	4.78E-3		14
V\$TFCP2_08		3.77 (1.95)	3.17E-6	1.3	2.65E-3		15
V\$CKROX_Q2		2.97 (1.69)	6.71E-7	0.89	4.32E-3		16
V\$NF1C_03		3.49 (1.54)	1.86E-3	0.64	3.86E-2		17
V\$BCL6B_04		2.16 (1.46)	5.01E-23	1.21	9.62E-3		18
V\$TCFL5_03		2.33 (1.41)	9.01E-7	0.86	1.38E-2		19
V\$ZNF672_01		2.63 (1.34)	3.67E-4	0.94	1.38E-2		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 significance of expression in the skeletal muscle tissue. The ranking of genes is 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\$BCL6B_04 BCL-6	V\$CKROX_Q2 LRF	V\$E2F_Q3 E2F-4 DP-1 <i>more</i>	V\$GKLF_Q3 TIEG-1 KLF15 BTEB1 <i>more</i>	V\$HIC1_Q2 HIC-1 hic2	V\$MAZR_Q1 DB1 MAZ	V\$MEF2_Q6_Q2 MEF-2C MEF-2D Mef-2a <i>more</i>	V\$MYOD_Q6_Q2 Myf-6 Myogenin	V\$NF1C_Q3 NF-1X NF-1C	V\$PAX4_Q3 pax-6	V\$PKNOX_Q1 PREP-2 PREP-1	V\$PTF1_Q3 EC2 scx Dermo-1 <i>more</i>	V\$SP5_Q3 Sp2	V\$SREBF_Q7 USF2 tfeb MITF <i>more</i>	V\$TCFL_Q3 SRC-1	V\$TFCP_Q8 CP2	V\$ZBTB_Q1 zbtb10	V\$ZNF462_Q1 ZNF462	V\$ZNF658B_Q4 ZNF658	V\$ZNF672_Q1 ZNF672	V\$ZNF740_Q7 ZNF511 znf414
ENSG00000182533	CAV3	caveolin 3	6.48	13	0	0	0	0	0	0	0	2	1	0	1	2	0	3	0	0	2	2	0	0	0
ENSG00000196557	CACNA1H	calcium voltage-gated channel subunit alpha1 H	5.1	67	24	2	1	1	0	2	0	0	0	3	0	0	3	1	13	2	2	1	8	2	2
ENSG00000071564	TCF3	transcription factor 3	4.25	19	10	0	1	0	2	0	0	0	0	1	0	0	0	0	0	0	0	2	2	1	0
ENSG00000129152	MYOD1	myogenic differentiation 1	4.16	18	6	1	1	1	1	0	0	0	0	1	0	1	0	1	2	0	1	1	1	0	0
ENSG00000143632	ACTA1	actin alpha 1, skeletal muscle	4.13	26	8	1	0	0	1	0	0	2	1	2	0	1	1	2	0	1	0	4	2	0	0
ENSG00000010278	CD9	CD9 molecule	4.08	29	8	0	0	0	1	1	0	0	0	3	0	1	0	2	3	3	1	2	2	1	1
ENSG00000188130	MAPK12	mitogen-activated protein kinase 12	3.93	67	21	4	0	8	0	12	0	1	0	1	0	0	3	1	1	0	2	1	2	0	10
ENSG00000162430	SELENON	selenoprotein N	3.88	32	12	1	0	0	1	0	0	0	2	2	1	1	1	1	0	4	0	1	3	2	0
ENSG00000072195	SPEG	striated muscle enriched protein kinase	3.86	14	4	0	0	0	2	0	1	0	0	0	0	1	1	3	0	0	0	1	1	0	0
ENSG00000154358	OBSCN	obscurin, cytoskeletal catmodulin and titin-interacting RhoGEF	3.64	14	6	1	0	0	0	0	0	0	0	2	0	1	0	1	0	2	0	0	1	0	0
ENSG00000017427	IGF1	insulin like growth factor 1	3.53	9	2	1	0	0	0	0	0	0	0	1	0	1	0	1	0	0	0	2	1	0	0
ENSG00000068024	HDAC4	histone deacetylase 4	3.51	18	7	0	1	0	0	0	0	0	0	2	0	1	0	1	0	4	1	0	1	0	0
ENSG00000185386	MAPK11	mitogen-activated protein kinase 11	3.51	8	2	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	3	1	0	0
ENSG00000173991	TCAP	titin-cap	3.48	9	2	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	2	1	1	0	0
ENSG00000109063	MYH3	myosin heavy chain 3	3.31	17	4	1	0	1	0	0	0	1	0	1	0	2	0	1	0	0	0	4	0	1	1
ENSG00000141448	GATA6	GATA binding protein 6	3.28	27	8	3	1	1	0	1	0	0	0	1	0	1	1	1	1	0	0	2	4	0	2
ENSG00000129170	CSRP3	cysteine and glycine rich protein 3	3.26	6	1	0	0	0	0	0	1	0	0	1	1	1	0	1	0	0	0	0	0	0	0
ENSG00000089225	TBX5	T-box transcription factor 5	3.25	10	2	0	2	0	0	0	2	0	2	1	0	0	0	0	0	0	0	0	1	0	0
ENSG00000081189	MEF2C	myocyte enhancer factor 2C	3.21	44	9	0	2	3	0	2	0	0	0	0	0	1	2	0	10	0	0	10	3	0	2
ENSG00000125378	BMP4	bone morphogenetic protein 4	2.94	7	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	1	1	0	0	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 1.0 (<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

The results produced by the MATCH Suite, contained in any of the reports or results visualization produced by this software, are based on the best of geneXplain's knowledge, however, we do not guarantee completeness and reliability of this information. GeneXplain GmbH does not guarantee comprehensiveness, reliability or accuracy of the information contained in the reports generated by the MATCH Suite.