ITGA4 and CCND3 are promising druggable targets for treating Neoplasm Metastasis and Osteosarcoma that control activity of ELK1, EP300 and SMAD2 transcription factors on promoters of differentially expressed genes

Demo User
geneXplain GmbH
info@genexplain.com
Data received on 07/09/2019 ; Run on 26/01/2021 ; Report generated on 27/01/2021

Genome Enhancer release 2.3 (TRANSFAC®, TRANSPATH® and HumanPSD™ release 2021.1)

Abstract

In the present study we applied the software package "Genome Enhancer" to a multiomics data set that contains *transcriptomics and proteomics* data. The study is done in the context of *Neoplasm Metastasis and Osteosarcoma*. The goal of this pipeline is to identify potential drug targets in the molecular network that governs the studied pathological process. In the first step of analysis pipeline discovers transcription factors (TFs) that regulate genes activities in the pathological state. The activities of these TFs are controlled by so-called master regulators, which are identified in the second step of analysis. After a subsequent druggability checkup, the most promising master regulators are chosen as potential drug targets for the analyzed pathology. At the end the pipeline comes up with (a) a list of known drugs and (b) investigational active chemical compounds with the potential to interact with selected drug targets.

From the data set analyzed in this study, we found the following TFs to be potentially involved in the regulation of the differentially expressed genes: ELK1, EP300, KLF8, SMAD2, TAL1 and CEBPB. The subsequent network analysis suggested

- integrins
- c-Mer
- Cdk6:cyclinD3-isoform1
- PKCepsilon

as the most promising molecular targets for further research, drug development and drug repurposing initiatives on the basis of identified molecular mechanism of the studied pathology. Having checked the actual druggability potential of the full list of identified targets, both, via information available in medical literature and via cheminformatics analysis of drug compounds, we have identified the following drugs as the most promising treatment candidates for the studied pathology: Bosutinib, Vedolizumab, Risedronate and Erlotinib.

1. Introduction

Recording "-omics" data to measure gene activities, protein expression or metabolic events is becoming a standard approach to characterize the pathological state of an affected organism or tissue. Increasingly, several of these methods are applied in a combined approach leading to large "multiomics" datasets. Still the challenge remains how to reveal the underlying molecular mechanisms that render a given pathological state different from the norm. The disease-causing mechanism can be described by a re-wiring of the cellular regulatory network, for instance as a result of a genetic or epigenetic alterations influencing the activity of relevant genes. Reconstruction of the disease-specific regulatory networks can help identify potential master regulators of the respective pathological process. Knowledge about these master regulators can point to ways how to block a pathological regulatory cascade. Suppression of certain molecular targets as components of these cascades may stop the pathological process and cure the disease.

Conventional approaches of statistical "-omics" data analysis provide only very limited information about the causes of the observed phenomena and therefore contribute little to the understanding of the pathological molecular mechanism. In contrast, the "upstream analysis" method [1-4] applied here has been deviced to provide a casual interpretation of the data obtained for a pathology state. This approach comprises two major steps: (1) analysing promoters and enhancers of differentially expressed genes for the transcription factors (TFs) involved in their regulation and, thus, important for the process under study; (2) re-constructing the signaling pathways that activate these TFs and identifying master regulators at the top of such pathways. For the first step, the database TRANSFAC® [6] is employed together with the TF binding site identification algorithms Match [7] and CMA [8]. The second step involves the signal transduction database TRANSPATH® [9] and special graph search algorithms [10] implemented in the software "Genome Enhancer".

The "upstream analysis" approach has now been extended by a third step that reveals known drugs suitable to inhibit (or activate) the identified molecular targets in the context of the disease under study. This step is performed by using information from HumanPSD™ database [5]. In addition, some known drugs and investigational active chemical compounds are subsequently predicted as potential ligands for the revealed molecular targets. They are predicted using a pre-computed database of spectra of biological activities of chemical compounds of a library of 2245 known drugs and investigational chemical compounds from HumanPSD™ database. The spectra of biological activities for these compounds are computed using the program PASS on the basis of a (Q)SAR approach [11-13]. These predictions can be used for the research purposes - for further drug development and drug repurposing initiatives.

2. Data

For this study the following experimental data was used:

Table 1. Experimental datasets used in the study

File name	Data type
Proteomics	Proteomics
RNAseq	Transcriptomics

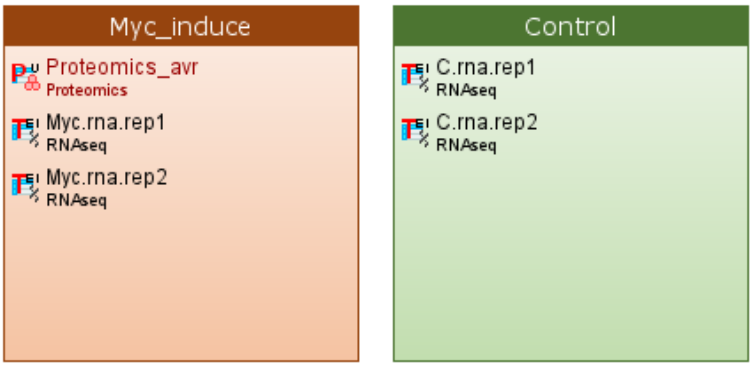


Figure 1. Annotation diagram of experimental data used in this study. With the colored boxes we show those sub-categories of the data that are compared in our analysis.

3. Results

We have compared the following conditions: Myc_induce *versus* Control.

3.1. Identification of target genes

In the first step of the analysis ***target genes*** were identified from the uploaded experimental data. We applied the Limma tool (R/Bioconductor package integrated into our pipeline) and compared gene expression in the following sets: "Myc_induce" with "Control". Limma calculated the LogFC (the logarithm to the base 2 of the fold change between different conditions), the p-value and the adjusted p-value (corrected for multiple testing) of the observed fold change. As a result, we detected 5047 upregulated genes (LogFC>0) out of which 1195 genes were found as significantly upregulated (p-value<0.1) and 4524 downregulated genes (LogFC<0) out of which 1169 genes were significantly downregulated (p-value<0.1). See tables below for the top significantly up- and downregulated genes. Below we call **target genes** the full list of up- and downregulated genes revealed in our analysis (see tables in Supplementary section).

*Table 2. Top ten significant **up-regulated** genes in Myc_induce vs. Control.*
See full table →

ID	Gene symbol	Gene description	logFC	P.Value	adj.P.Val
ENSG00000136997	MYC	MYC proto-oncogene, bHLH transcription factor	5.96	7.45E-6	7.13E-2
ENSG00000164076	CAMKV	CaM kinase like vesicle associated	4.08	8.1E-5	0.13
ENSG00000120738	EGR1	early growth response 1	3.51	5.46E-4	0.14
ENSG00000173110	HSPA6	heat shock protein family A (Hsp70) member 6	3.14	1.66E-4	0.13
ENSG00000123360	PDE1B	phosphodiesterase 1B	2.85	1.08E-4	0.13
ENSG00000137571	SLCO5A1	solute carrier organic anion transporter family member 5A1	2.79	9.53E-5	0.13
ENSG00000078549	ADCYAP1R1	ADCYAP receptor type I	2.69	2.44E-3	0.14
ENSG00000143333	RGS16	regulator of G protein signaling 16	2.69	2.47E-4	0.13
ENSG00000170345	FOS	Fos proto-oncogene, AP-1 transcription factor subunit	2.57	4.12E-3	0.15
ENSG00000117322	CR2	complement C3d receptor 2	2.46	2.57E-4	0.13

*Table 3. Top ten significant **down-regulated** genes in Myc_induce vs. Control.*
See full table →

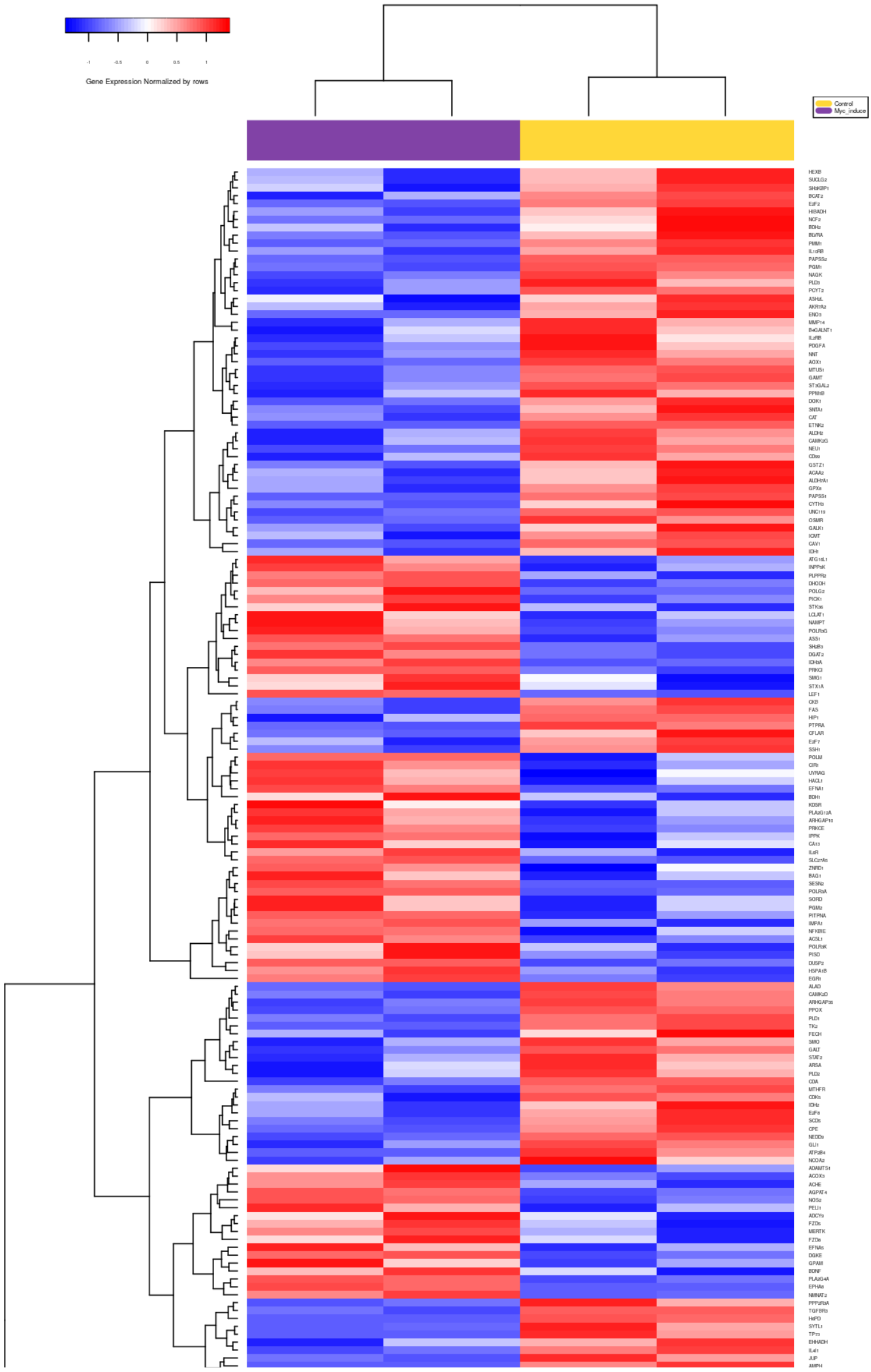
ID	Gene symbol	Gene description	logFC	P.Value	adj.P.Val
ENSG00000116774	OLFML3	olfactomedin like 3	-3.06	1.11E-4	0.13
ENSG00000138131	LOXL4	lysyl oxidase like 4	-2.62	8.88E-4	0.14
ENSG00000187867	PALM3	paralemmin 3	-2.62	2.65E-3	0.14
ENSG00000205542	TMSB4X	thymosin beta 4 X-linked	-2.58	2.22E-4	0.13
ENSG00000158825	CDA	cytidine deaminase	-2.54	3.49E-4	0.13
ENSG00000127129	EDN2	endothelin 2	-2.49	3.28E-4	0.13
ENSG00000182667	NTM	neurotrimin	-2.48	4.08E-4	0.13
ENSG00000114115	RBP1	retinol binding protein 1	-2.46	1.06E-4	0.13
ENSG00000132746	ALDH3B2	aldehyde dehydrogenase 3 family member B2	-2.35	1.93E-4	0.13
ENSG00000188042	ARL4C	ADP ribosylation factor like GTPase 4C	-2.29	1.87E-3	0.14

3.2. Functional classification of genes

A functional analysis of differentially expressed genes was done by mapping the significant up-regulated and significant down-regulated genes to several known ontologies, such as Gene Ontology (GO), disease ontology (based on HumanPSD™ database) and the ontology of signal transduction and metabolic pathways from the TRANSPATH® database. Statistical significance was computed using a binomial test. Figures 3-8 show the most significant categories.

Heatmap of differentially expressed genes in Myc_induce vs. Control

A heatmap of all differentially expressed genes playing a potential regulatory role in the system (enriched in TRANSPATH® pathways) is presented in Figure 2.


-1
-0.5
0
0.5
1
Gene Expression Normalized by rows
Control
Myc_induce
HEXB
SUCLG2
SH3KBP1
BCAT2
EIF2
HIBADH
NCF2
BDH2
RUVBL1
PMM1
IL1RAP
PAPSS2
PGM1
NAGK
PLD3
PCYT2
ASH2L
AKR1A1
ENO3
MMP14
B4GALNT1
IL2RB
PDGFA
NNT
AOX1
MTUS1
GAMT
ST3GAL2
PPM1B
DGKI
SNTA1
CAT
ETNK2
ALDH2
CAMKK2
NEU1
CD99
GSTZ1
ACAA2
ALDH1A1
GPX4
PAPSS1
CYTH3
UNC119
OSMR
GALK1
ICMT
CAV1
IDH1
ATG10L1
INPP5K
PLPPR2
DHODH
POLG2
PICK1
STK36
LCLAT1
NAMPT
POLR3G
ASS1
SH3BP5
DGAT2
IDH3A
PRKCI
SMG1
STX1A
LEF1
CKB
FAS
HIP1
PTPRA
CFLAR
EIF3I
SSH1
POLM
CIR1
UVRAG
HACL1
EFNA1
BDH1
KDSR
PLA2G12A
ARHGAP10
PRKCE
IPPK
CA13
IL6R
SLC2A5
ZNRD1
BAG1
SESN2
POLR3A
SORD
PGM2
PITPNA
IMPA1
NFKBIE
ACSL1
POLR3K
PISD
DUSP2
HSPA1B
EGR1
ALAD
CAMKK2
ARHGAP35
PPOX
PLD1
TK2
FECH
SMO
GALT
STAT2
ARSA
PLD2
CDA
MTHFR
CDK5
IDH2
EIF4
SCD5
CPE
NEDD9
GLI1
ATP2B4
NCOA2
ADAMTS1
ACOX3
ACHE
AGPAT4
NOS2
PELI1
ADCY9
FZD5
MERTK
FZD8
EFNA5
DGKE
GPAM
BDNF
PLA2G4A
EPHA4
NMNAT2
PPP2R5A
TGFBR3
HsPD
SYTL1
TPI1
EHHADH
IL4I1
JUP
AMPH

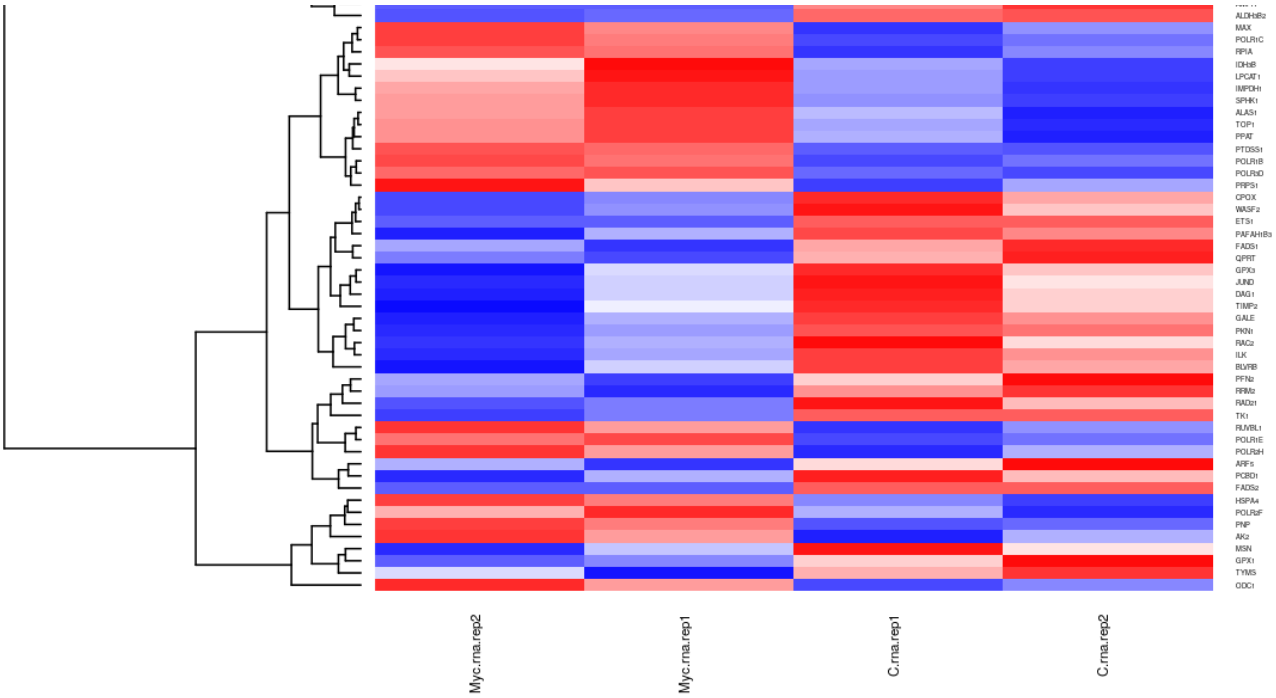


Figure 2. Heatmap of genes enriched in Transpath categories. The colored bar at the top shows the types of the samples according to the legend in the upper right corner.
See full diagram →

Up-regulated genes in Myc_induce vs. Control:

1195 significant up-regulated genes were taken for the mapping.

GO (biological process)

Figure 3. Enriched GO (biological process) of up-regulated genes in Myc_induce vs. Control.

Full classification →

TRANSPATH® Pathways (2021.1)

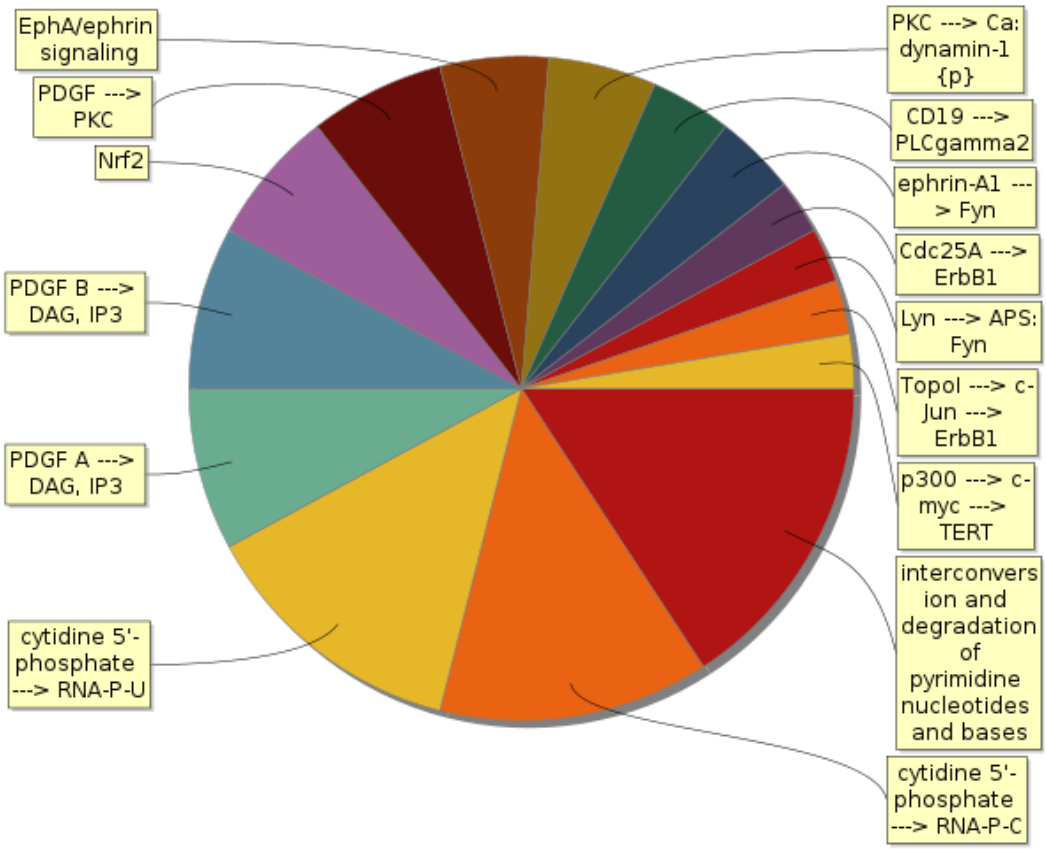


Figure 4. Enriched TRANSPATH® Pathways (2021.1) of up-regulated genes in Myc_induce vs. Control.
Full classification →

HumanPSD(TM) disease (2021.1)

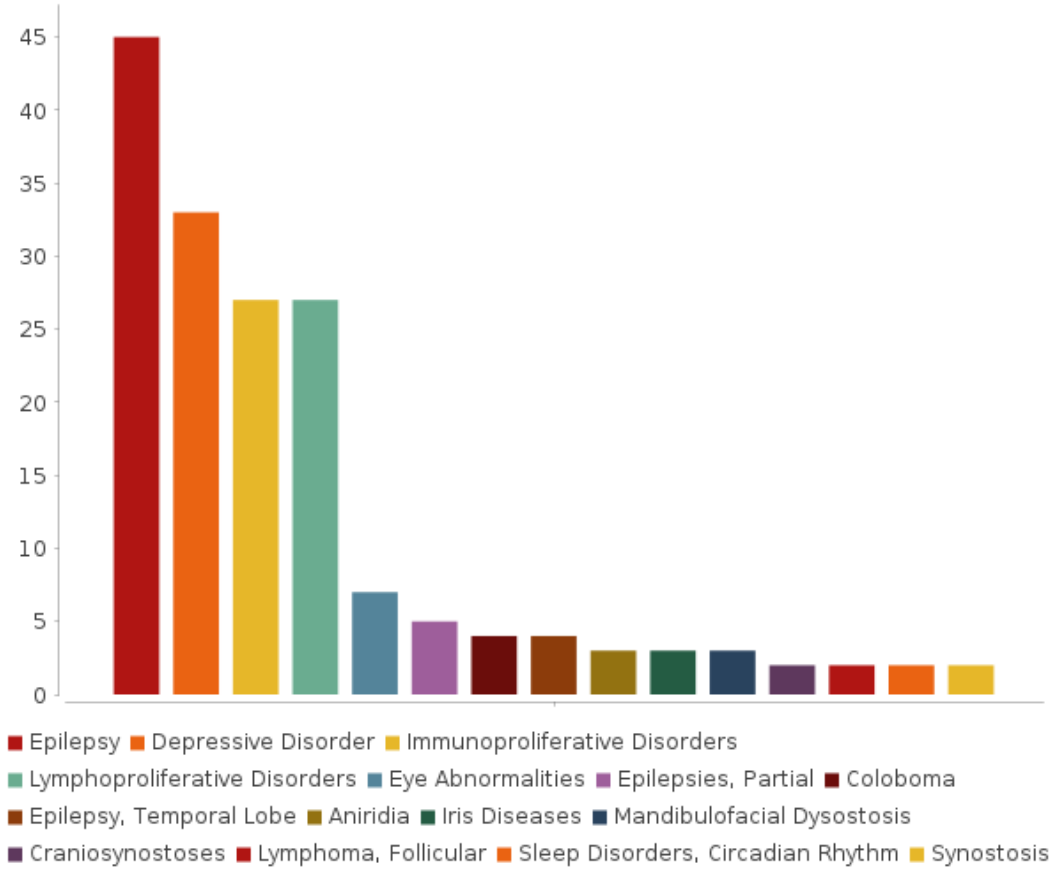


Figure 5. Enriched HumanPSD(TM) disease (2021.1) of up-regulated genes in Myc_induce vs. Control. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

Full classification →

Down-regulated genes in Myc_induce vs. Control:

1169 significant down-regulated genes were taken for the mapping.

GO (biological process)

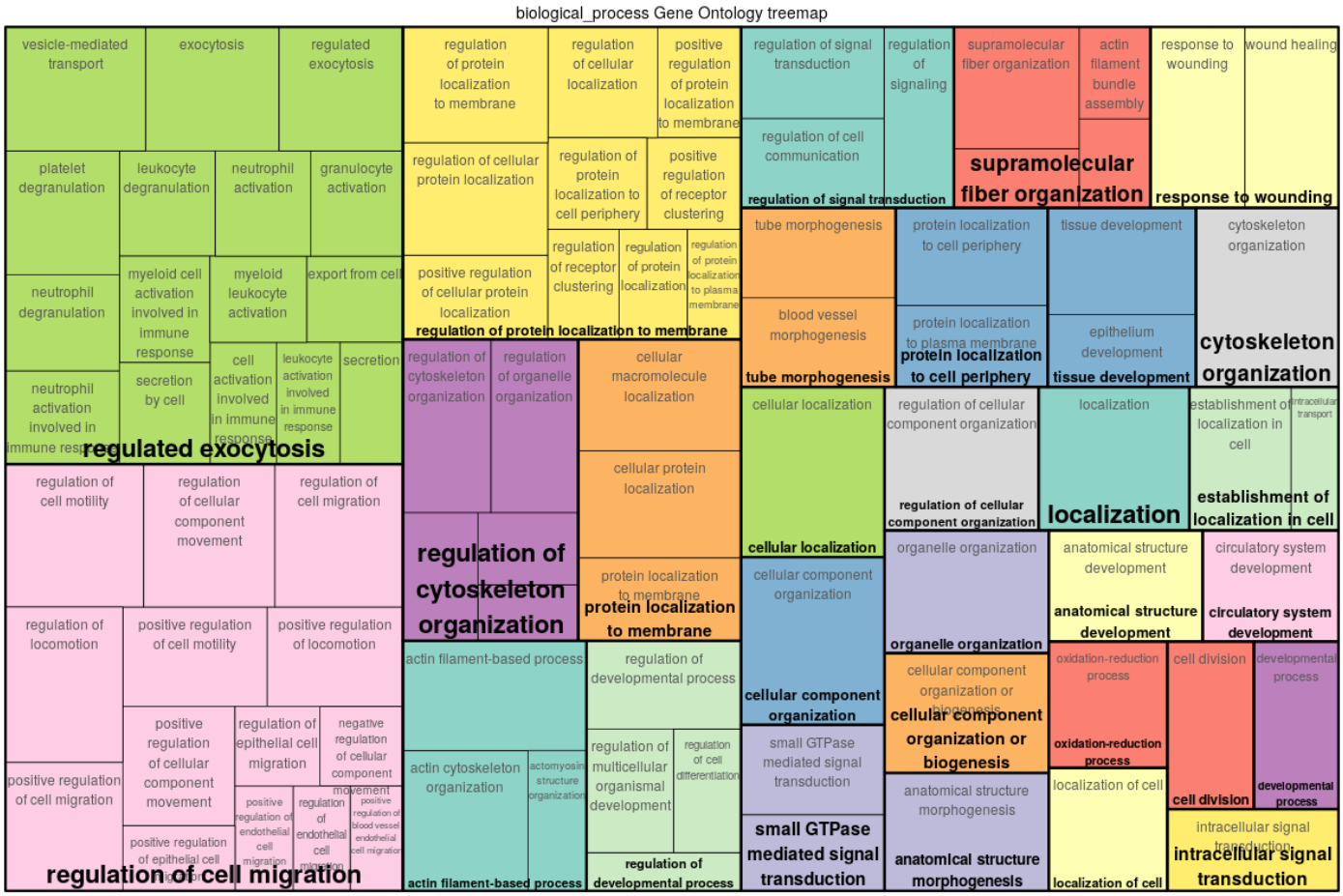


Figure 6. Enriched GO (biological process) of down-regulated genes in Myc_induce vs. Control.

Full classification →

TRANSPATH® Pathways (2021.1)

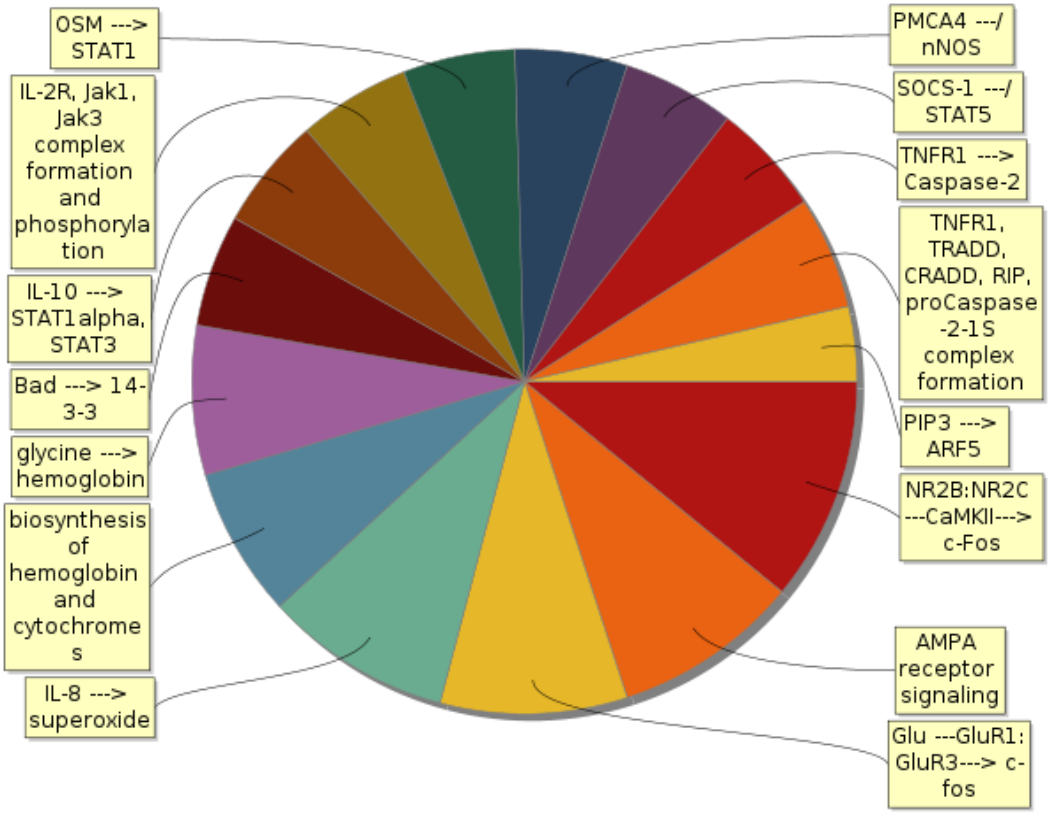


Figure 7. Enriched TRANSPATH® Pathways (2021.1) of down-regulated genes in Myc_induce vs. Control.
Full classification →

HumanPSD(TM) disease (2021.1)

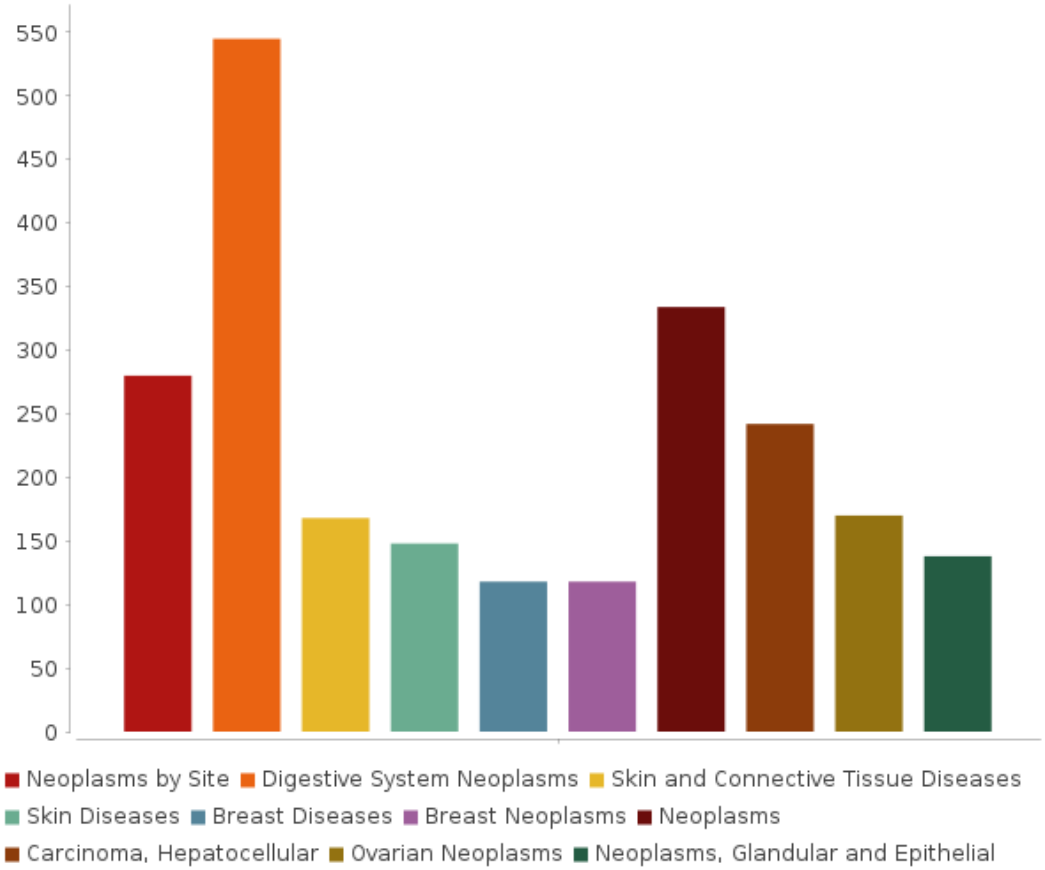


Figure 8. Enriched HumanPSD(TM) disease (2021.1) of down-regulated genes in Myc_induce vs. Control. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

Full classification →

3.3. Identification of proteins

In the first step of the proteome data analysis target proteins were identified from the uploaded experimental data (the list of 4665 proteins) and were converted to corresponding genes. These genes were used in the further steps of analysis.

Table 4. Top ten the list of genes provided as input in Myc_induce.

See full table →

ID	Gene description	Gene symbol	Proteomics_avr
ENSG00000173598	nudix hydrolase 4	NUDT4	4.36
ENSG00000100335	mitochondrial elongation factor 1	MIEF1	3.8
ENSG00000115884	syndecan 1	SDC1	3.62
ENSG00000102910	lon peptidase 2, peroxisomal	LONP2	3.3
ENSG00000179046	tripartite motif family like 2	TRIML2	2.87
ENSG00000114648	kelch like family member 18	KLHL18	2.76
ENSG00000170525	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3	PFKFB3	2.69
ENSG00000120949	TNF receptor superfamily member 8	TNFRSF8	2.46
ENSG00000188158	NHS actin remodeling regulator	NHS	2.46
ENSG00000119599	DDB1 and CUL4 associated factor 4	DCAF4	2.42

3.4. Functional classification of expressed proteins

A functional analysis of expressed proteins was done by mapping the protein IDs to several known ontologies, such as Gene Ontology (GO), disease ontology (based on HumanPSD™ database) and the ontology of signal transduction and metabolic pathways from the TRANSPATH® database. Statistical significance was computed using a binomial test.

Figures 9-11 show the most significant categories.

The list of proteins provided as input in Myc_induce:

4655 the list of genes provided as input genes were taken for the mapping.

GO (biological process)

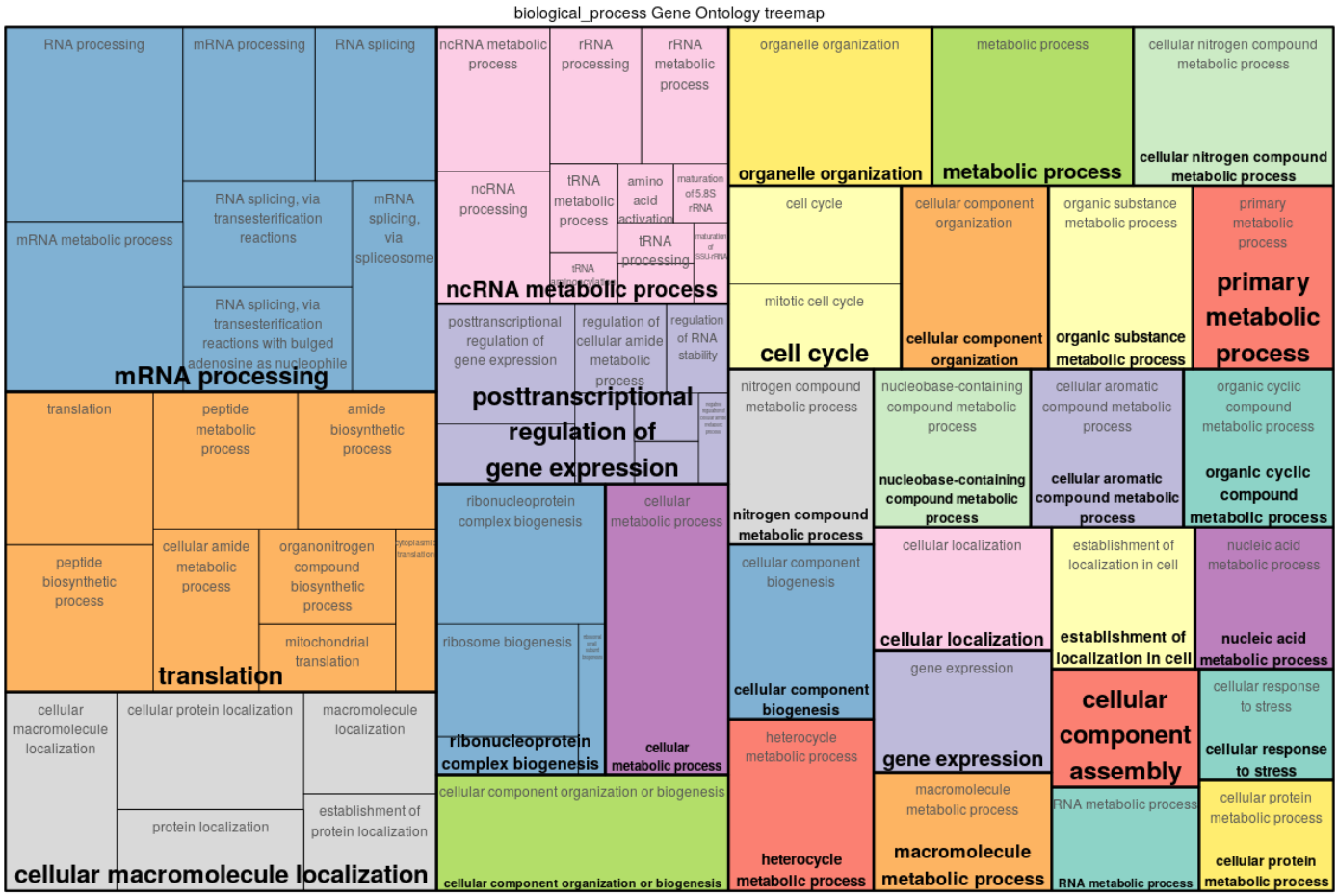


Figure 9. Enriched GO (biological process) of the list of proteins provided as input in Myc_induce.
Full classification →

TRANSPATH® Pathways (2021.1)

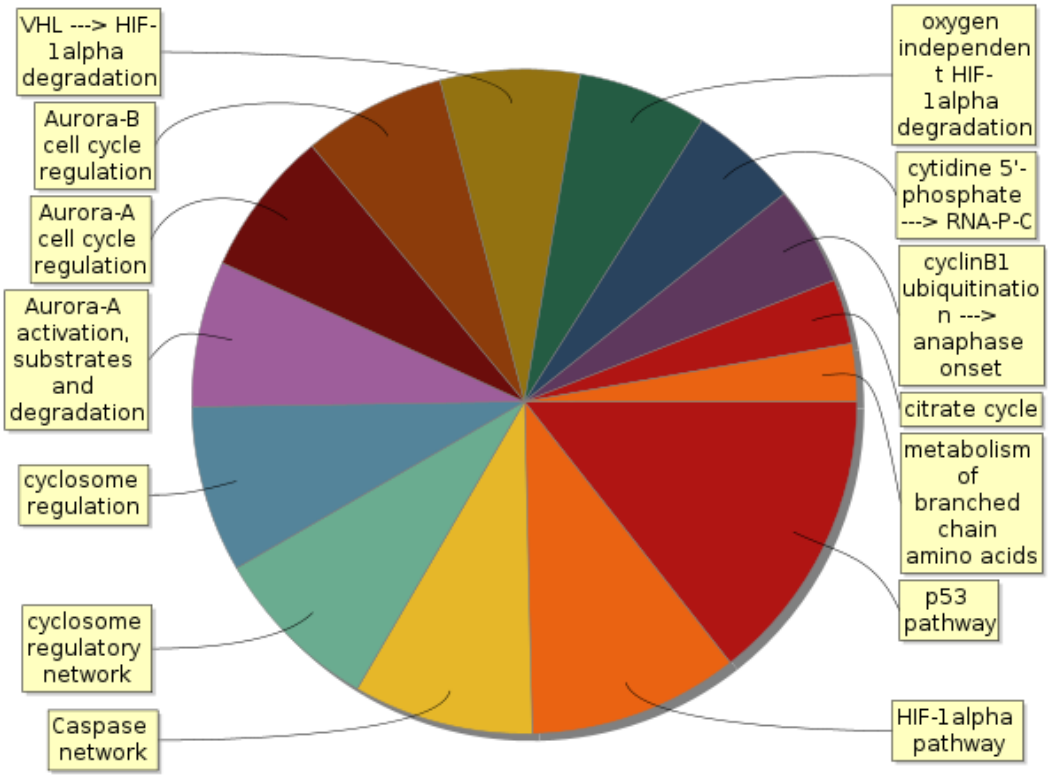


Figure 10. Enriched TRANSPATH® Pathways (2021.1) of the list of proteins provided as input in Myc_induce.
Full classification →

HumanPSD(TM) disease (2021.1)

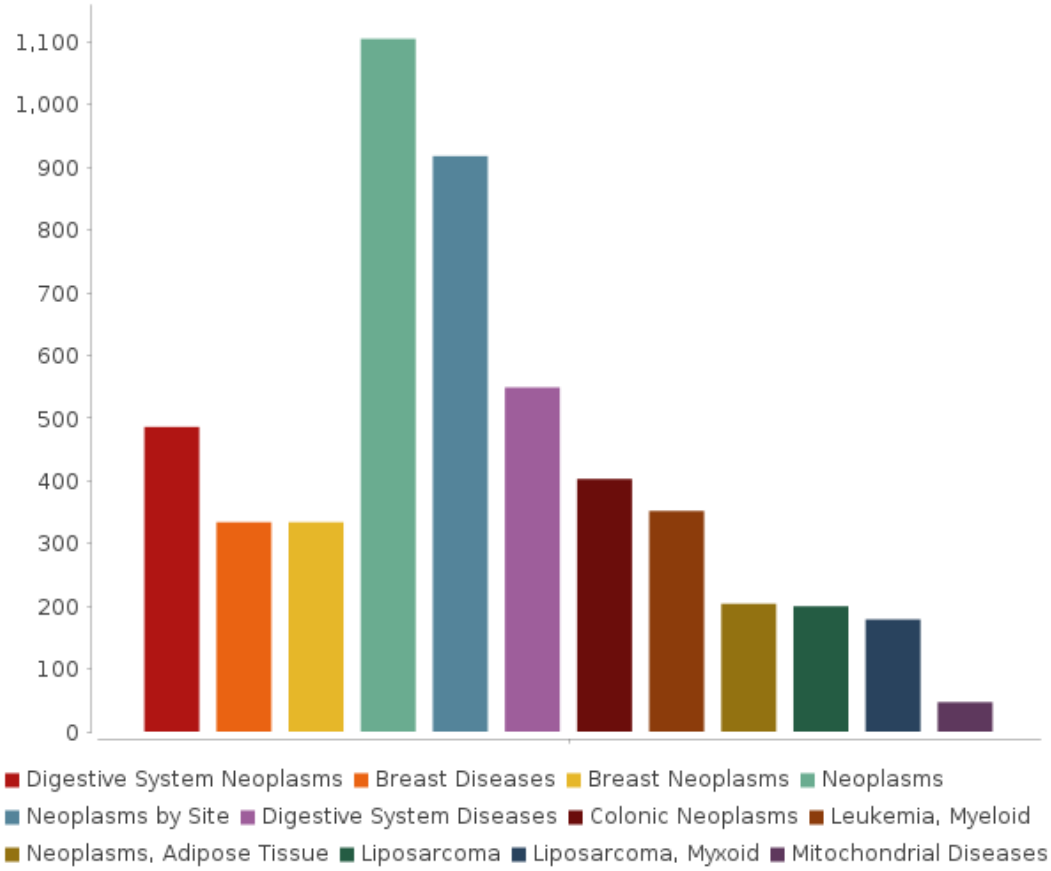


Figure 11. Enriched HumanPSD(TM) disease (2021.1) of the list of proteins provided as input in Myc_induce. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

Full classification →

3.5. Comparison plot of transcriptome and proteome

After the analysis of transcriptome and proteome data they were compared with each other. Below we plot 9578 genes and 4655 proteins.

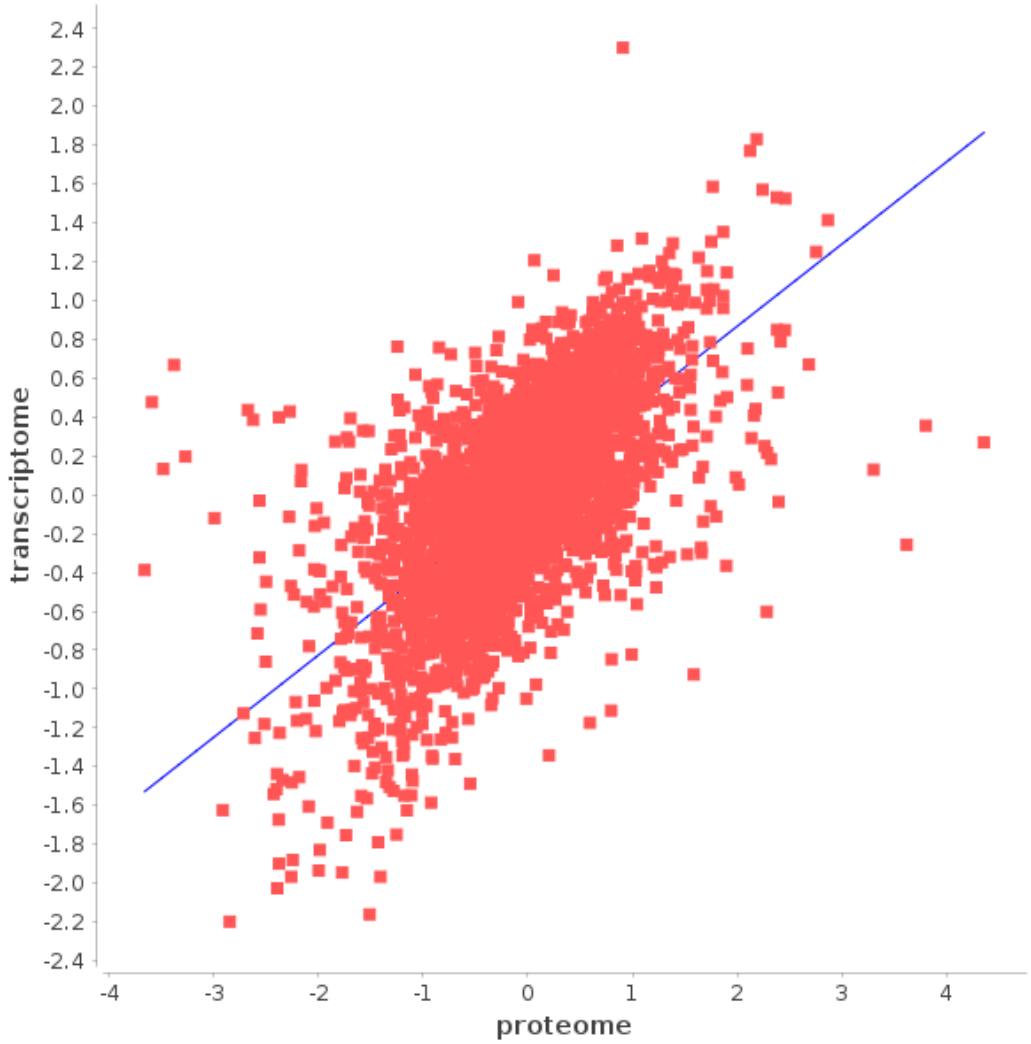


Figure 12. Comparison plot of comparison proteome vs transcriptome. X axis: protein expression value - Proteomics_avr. Y axis: LogFC of differential gene expression.

Full comparison →

Comparison of up-regulated genes (transcriptome data) and the list of proteins provided as input (proteome data)

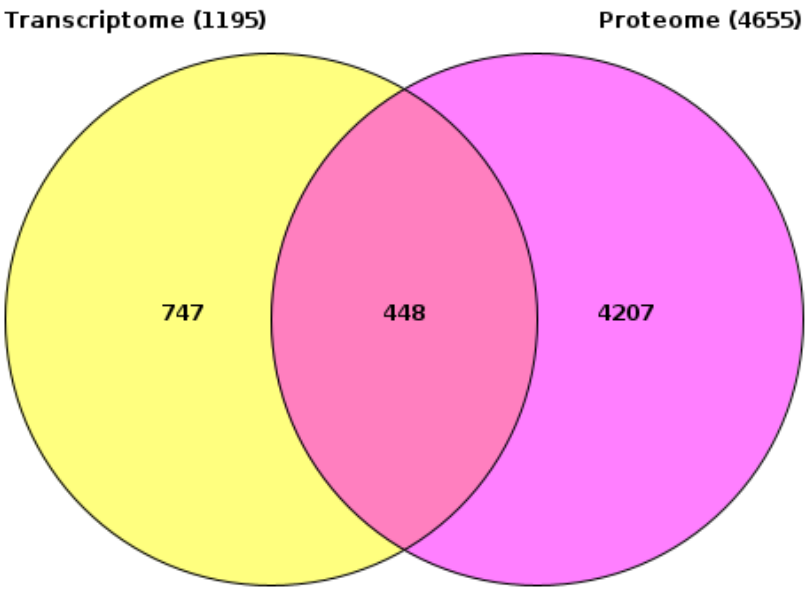


Figure 13. Intersection of up-regulated genes and the list of proteins provided as input

See full diagram →

Comparison of down-regulated genes (transcriptome data) and the list of proteins provided as input (proteome data)

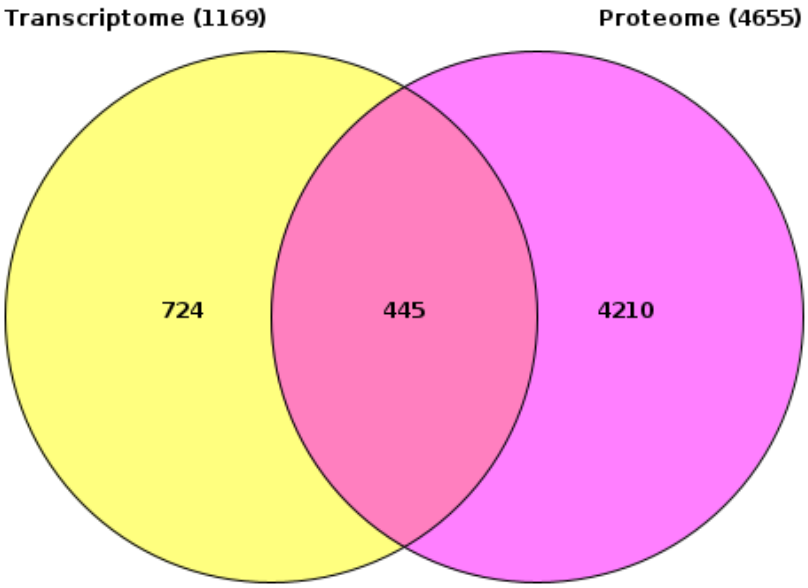
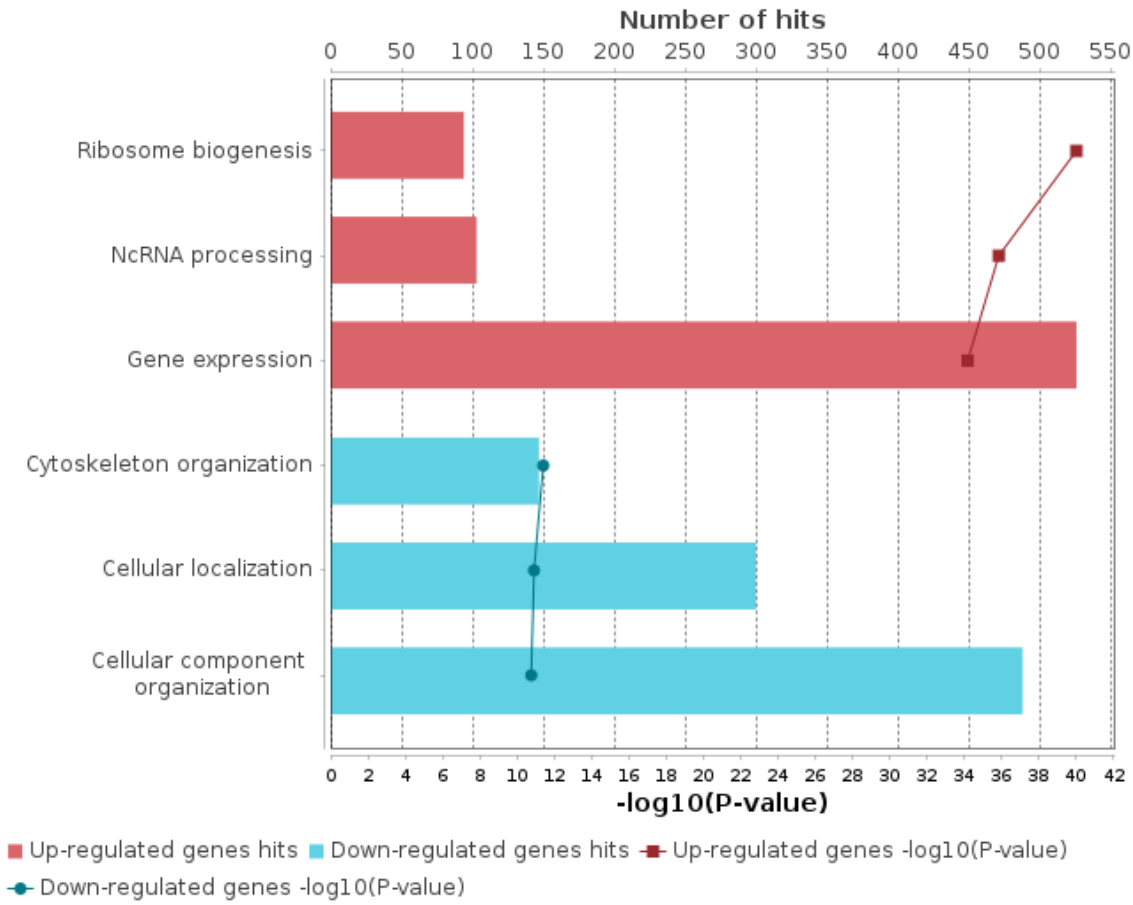


Figure 14. Intersection of down-regulated genes and the list of proteins provided as input
See full diagram →

The result of overall Gene Ontology (GO) analysis of the differentially expressed genes of the studied pathology can be summarized by the following diagram, revealing the most significant functional categories overrepresented among the observed (differentially expressed genes):



3.6. Analysis of enriched transcription factor binding sites and composite modules

In the next step a search for transcription factors binding sites (TFBS) was performed in the regulatory regions of the ***target genes*** by using the TF binding motif library of the TRANSFAC® database. We searched for so called **composite modules** that act as potential condition-specific **enhancers** of the ***target genes*** in their upstream regulatory regions (-1000 bp upstream of transcription start site (TSS)) and identify transcription factors regulating activity of the genes through such **enhancers**.

Classically, **enhancers** are defined as regions in the genome that increase transcription of one or several genes when inserted in either orientation at various distances upstream or downstream of the gene [8]. Enhancers typically have a length of several hundreds of nucleotides and are bound by multiple transcription factors in a cooperative manner [9].

We applied the Composite Module Analyst (CMA) [8] method to detect such potential enhancers, as targets of multiple TFs bound in a cooperative manner to the regulatory regions of the genes of interest. CMA applies a genetic algorithm to construct a generalized model of the enhancers by specifying combinations of TF motifs (from TRANSFAC®) whose sites are most frequently clustered together in the regulatory regions of the studied genes. CMA identifies the transcription factors that through their cooperation provide a synergistic effect and thus have a great influence on the gene regulation process.

Enhancer model potentially involved in regulation of target genes (up-regulated genes in Myc_induce vs. Control).

To build the most specific composite modules we choose top 300 significant up-regulated genes as the input of CMA algorithm. The obtained CMA model is then applied to compute CMA score for all up-regulated genes.

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.

Module 1:

V$AP2ALPHA_02	V$EP300_06	V$ARNT_01	V$TBX2_06	V$KLF8_Q5	V$STAT2_01
0.90; N=3	0.87; N=3	0.84; N=1	0.97; N=3	0.97; N=3	0.84; N=3

Module width: 115

Module 2:

V$CP2_02	V$ELK1SPDEF_01	V$FOXK1_13	V$CLOCK_01	V$HOXA10_13	V$TATA_C
0.85; N=3	0.73; N=3	0.92; N=2	0.94; N=3	0.88; N=3	0.94; N=1

Module width: 191

Model score (-p*log10(pval)): 12.94
Wilcoxon p-value (pval): 5.43e-28
Penalty (p): 0.475
Average yes-set score: 6.59
Average no-set score: 5.08
AUC: 0.73
Middle-point: 6.05
False-positive: 26.81%
False-negative: 36.36%

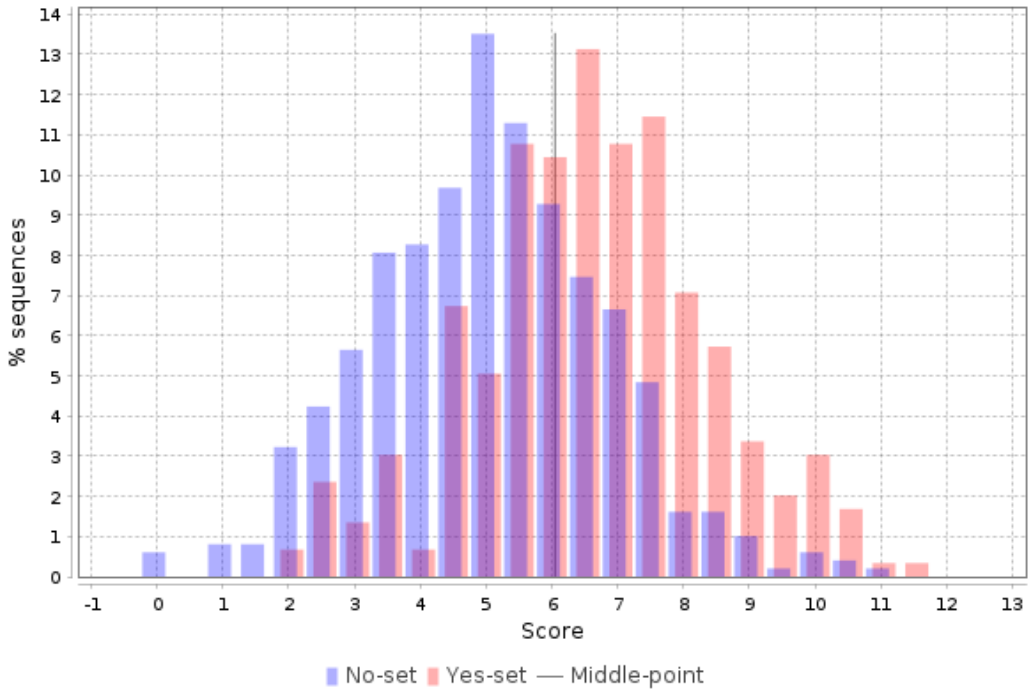


Table 5. List of top ten up-regulated genes in Myc_induce vs. Control with identified enhancers in their regulatory regions. ***CMA score*** *- the score of the CMA model of the enhancer identified in the regulatory region.*
See full table →

Ensembl IDs	Gene symbol	Gene description	CMA score	Factor names
ENSG00000121274	TENT4B	terminal nucleotidyltransferase 4B	12.01	HOXA10(h), CLOCK(h), FOXK1(h), Elk-1(h),PDEF(h), AP-2alpha(h), arnt(h), TBX2(h)...
ENSG00000151014	NOCT	nocturnin	11.74	Elk-1(h),PDEF(h), KLF8(h), FOXK1(h), AP-2alpha(h), TBX2(h), arnt(h), CLOCK(h)...
ENSG00000140474	ULK3	unc-51 like kinase 3	11.68	STAT2(h), CP2(h), HOXA10(h), AP-2alpha(h), KLF8(h), TBX2(h), arnt(h)
ENSG00000126226	PCID2	PCI domain containing 2	11.31	CP2(h), Elk-1(h),PDEF(h), TBX2(h), arnt(h), KLF8(h), AP-2alpha(h)
ENSG00000170525	PFKFB3	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3	11.22	arnt(h), KLF8(h), AP-2alpha(h), TBX2(h), STAT2(h), Elk-1(h),PDEF(h), CP2(h)
ENSG00000196453	ZNF777	zinc finger protein 777	11.19	Elk-1(h),PDEF(h), FOXK1(h), CP2(h), HOXA10(h), STAT2(h), arnt(h), AP-2alpha(h)...
ENSG00000118523	CCN2	cellular communication network factor 2	11.18	CP2(h), TBX2(h), AP-2alpha(h), TBP(h), STAT2(h), FOXK1(h), Elk-1(h),PDEF(h)...
ENSG00000100105	PATZ1	POZ/BTB and AT hook containing zinc finger 1	11.14	AP-2alpha(h), TBX2(h), p300(h), KLF8(h), STAT2(h), arnt(h), Elk-1(h),PDEF(h)...
ENSG00000196368	NUDT11	nudix hydrolase 11	11.12	STAT2(h), CP2(h), Elk-1(h),PDEF(h), AP-2alpha(h), KLF8(h), arnt(h), CLOCK(h)...
ENSG00000167840	ZNF232	zinc finger protein 232	11.03	CP2(h), AP-2alpha(h), CLOCK(h), KLF8(h), TBX2(h), arnt(h), FOXK1(h)

Enhancer model potentially involved in regulation of target genes (down-regulated genes in Myc_induce vs. Control).

To build the most specific composite modules we choose top 300 significant down-regulated genes as the input of CMA algorithm. The obtained CMA model is then applied to compute CMA score for all down-regulated genes.

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.

Module 1:

V$ZNF462_01	V$MZF1_Q5	V$SMAD2_Q6	V$SOX10_01	V$VDR_03
0.93; N=3	0.98; N=2	1.00; N=3	1.00; N=2	0.82; N=3

Module width: 198

Module 2:

V$ZNF462_01	V$TRIM28_02	V$TAL1BETAITF2_01	V$KLF5_01	V$CEBPB_Q6
0.93; N=2	0.83; N=2	0.79; N=3	0.98; N=3	0.96; N=3

V$PAX2_Q2
0.86; N=3

Module width: 151

Model score (-p*log10(pval)): 22.23
Wilcoxon p-value (pval): 2.29e-46
Penalty (p): 0.487
Average yes-set score: 10.34
Average no-set score: 7.89
AUC: 0.80
Middle-point: 9.22
False-positive: 24.40%
False-negative: 27.00%
The AUC of the model achieves value significantly higher than expected for a random set of regulatory regions
Z-score = 3.51

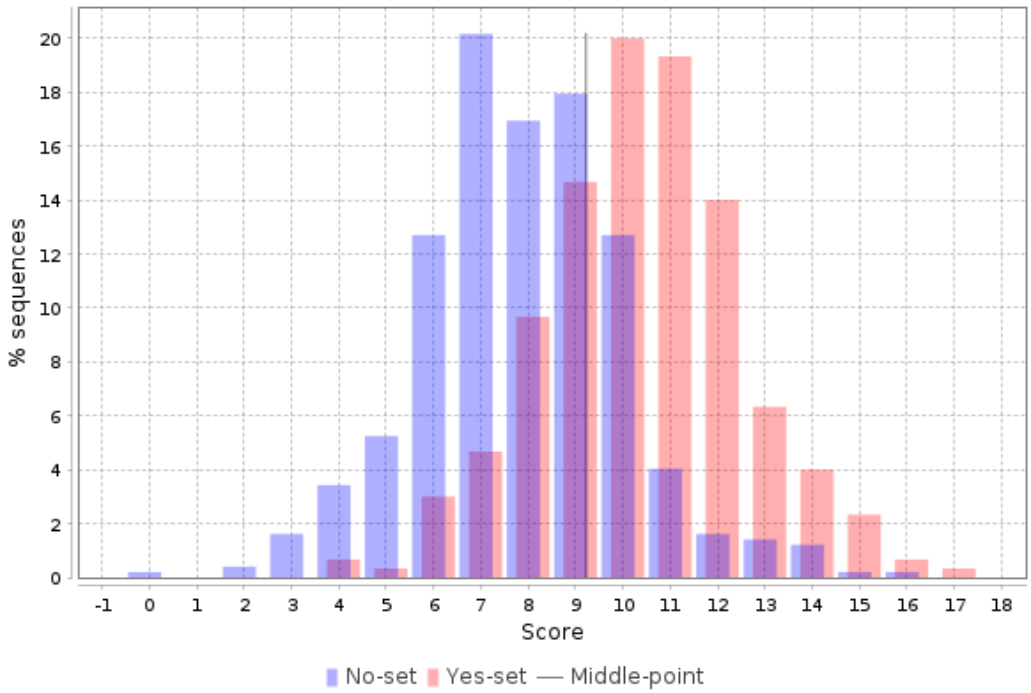


Table 6. List of top ten down-regulated genes in Myc_induce vs. Control with identified enhancers in their regulatory regions. ***CMA score*** *- the score of the CMA model of the enhancer identified in the regulatory region.*

See full table →

Ensembl IDs	Gene symbol	Gene description	CMA score	Factor names
ENSG00000063180	CA11	carbonic anhydrase 11	18.01	BTEB2(h), pax-2(h), ZNF462(h), ITF-2(h),Tal-1(h), RNF96(h), VDR(h), MZF-1(h)...
ENSG00000185133	INPP5J	inositol polyphosphate-5-phosphatase J	17.39	MZF-1(h), pax-2(h), VDR(h), C/EBPbeta(h), ITF-2(h),Tal-1(h), Smad2(h), ZNF462(h)...
ENSG00000105048	TNNT1	troponin T1, slow skeletal type	16.93	pax-2(h), C/EBPbeta(h), ITF-2(h),Tal-1(h), RNF96(h), Smad2(h), ZNF462(h), BTEB2(h)...
ENSG00000147065	MSN	moesin	16.81	Smad2(h), pax-2(h), BTEB2(h), ZNF462(h), RNF96(h), ITF-2(h),Tal-1(h), VDR(h)...
ENSG00000135414	GDF11	growth differentiation factor 11	16.55	ITF-2(h),Tal-1(h), pax-2(h), BTEB2(h), RNF96(h), ZNF462(h), VDR(h), Smad2(h)...
ENSG00000159216	RUNX1	RUNX family transcription factor 1	16.54	BTEB2(h), ZNF462(h), VDR(h), Smad2(h), MZF-1(h), RNF96(h), ITF-2(h),Tal-1(h)...
ENSG00000069399	BCL3	BCL3 transcription coactivator	16.42	pax-2(h), Smad2(h), ZNF462(h), MZF-1(h), RNF96(h), VDR(h), BTEB2(h)
ENSG00000051128	HOMER3	homer scaffold protein 3	16.21	Sox-10(h), BTEB2(h), pax-2(h), ITF-2(h),Tal-1(h), ZNF462(h), MZF-1(h), VDR(h)...
ENSG00000130592	LSP1	lymphocyte specific protein 1	16.02	ZNF462(h), MZF-1(h), VDR(h), pax-2(h), Smad2(h), ITF-2(h),Tal-1(h), BTEB2(h)...
ENSG00000186193	SAPCD2	suppressor APC domain containing 2	15.73	ITF-2(h),Tal-1(h), VDR(h), BTEB2(h), Smad2(h), Sox-10(h), ZNF462(h), MZF-1(h)...

On the basis of the enhancer models we identified transcription factors potentially regulating the ***target genes*** of our interest. We found 13 and 11 transcription factors controlling expression of up- and down-regulated genes respectively (see Tables 7-8).

Table 7. Transcription factors of the predicted enhancer model potentially regulating the differentially expressed genes (up-regulated genes in Myc_induce vs. Control). ***Yes-No ratio*** *is the ratio between frequencies of the sites in Yes sequences versus No sequences. It describes the level of the enrichment of binding sites for the indicated TF in the regulatory target regions.* ***Regulatory score*** *is the measure of involvement of the given TF in the controlling of expression of genes that encode master regulators presented below (through positive feedback loops).*

See full table →

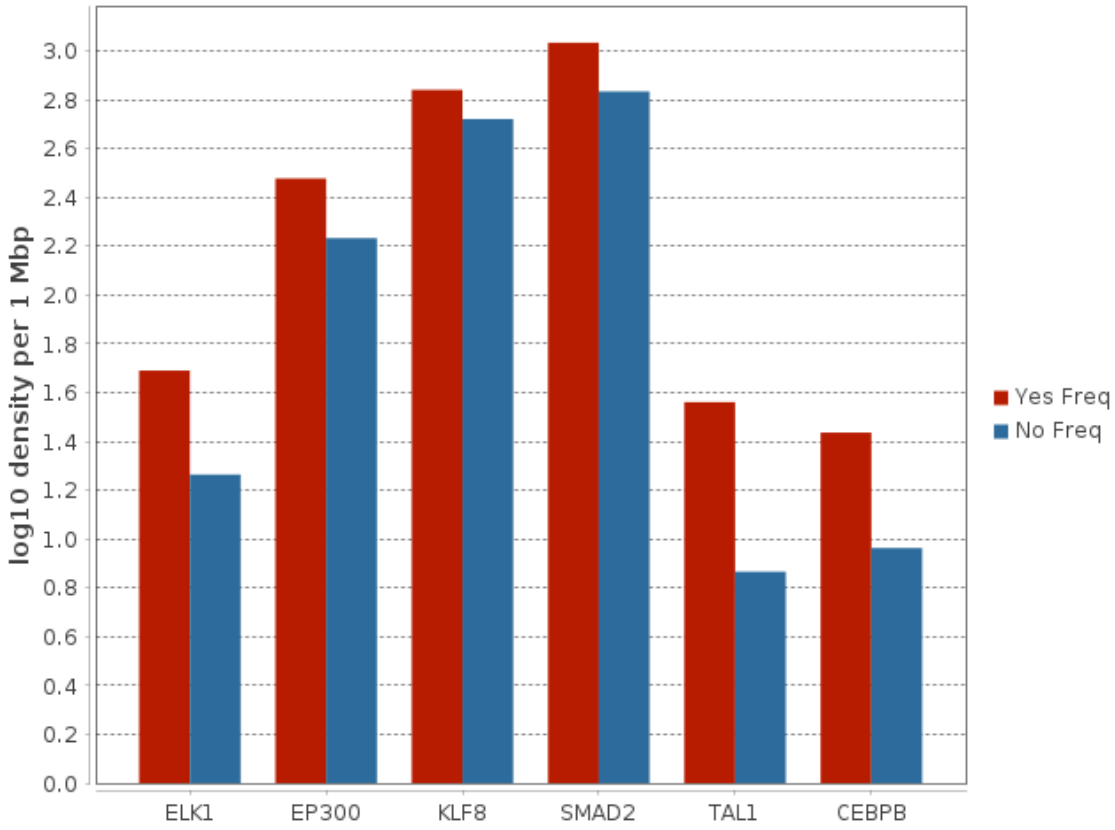
ID	Gene symbol	Gene description	Regulatory score	Yes-No ratio
MO000019544	ELK1	ETS transcription factor ELK1	3.61	2.67
MO000056654	EP300	E1A binding protein p300	3.5	1.76
MO000095459	KLF8	Kruppel like factor 8	2.99	1.32
MO000013121	STAT2	signal transducer and activator of transcription 2	2.88	1.23
MO000117988	TFCP2	transcription factor CP2	2.74	1.38
MO000021896	TBP	TATA-box binding protein	2.47	1.45
MO000114191	ARNT	aryl hydrocarbon receptor nuclear translocator	2.33	1.33
MO000001275	TFAP2A	transcription factor AP-2 alpha	2.3	1.32
MO000089495	HOXA10	homeobox A10	2.2	1.22
MO000028681	CLOCK	clock circadian regulator	2.13	5.83

Table 8. Transcription factors of the predicted enhancer model potentially regulating the differentially expressed genes (down-regulated genes in Myc_induce vs. Control). ***Yes-No ratio*** *is the ratio between frequencies of the sites in Yes sequences versus No sequences. It describes the level of the enrichment of binding sites for the indicated TF in the regulatory target regions.* ***Regulatory score*** *is the measure of involvement of the given TF in the controlling of expression of genes that encode master regulators presented below (through positive feedback loops).*

See full table →

ID	Gene symbol	Gene description	Regulatory score	Yes-No ratio
MO000057829	SMAD2	SMAD family member 2	6.01	1.58
MO000032489	TAL1	TAL bHLH transcription factor 1, erythroid differentiation factor	5.89	4.95
MO000019381	CEBPB	CCAAT enhancer binding protein beta	4.48	2.97
MO000092587	ZNF462	zinc finger protein 462	3.87	1.37
MO000021495	VDR	vitamin D receptor	3.85	1.3
MO000025957	PAX2	paired box 2	3.77	3.71
MO000069886	TRIM28	tripartite motif containing 28	3.66	1.24
MO000026229	KLF5	Kruppel like factor 5	3.26	1.37
MO000024921	TCF4	transcription factor 4	2.95	1.35
MO000028699	SOX10	SRY-box transcription factor 10	0	2.64

The following diagram represents the key transcription factors, which were predicted to be potentially regulating differentially expressed genes in the analyzed pathology: ELK1, EP300, KLF8, SMAD2, TAL1 and CEBPB.



3.7. Finding master regulators in networks

In the second step of the upstream analysis common regulators of the revealed TFs were identified. Using proteomics data we selected differentially expressed proteins that are involved in signal transduction pathways and used these proteins as "context set" [4] in the algorithm of identification of master regulators. These master regulators appear to be the key candidates for therapeutic targets as they have a master effect on regulation of intracellular pathways that activate the pathological process of our study. The identified master regulators are shown in Tables 9-10.

*Table 9. Master regulators that may govern the regulation of **up-regulated** genes in Myc_induce vs. Control. **Total rank** is the sum of the ranks of the master molecules sorted by keynode score, CMA score, transcriptomics and proteomics data.*

See full table →

ID	Master molecule name	Gene symbol	Gene description	Contained in proteome set	Total rank	logFC (transcriptome)
MO0000032632	PKCepsilon(h)	PRKCE	protein kinase C epsilon	1	166	0.5
MO0000329204	Cdk6(h):cyclinD3-isoform1(h)	CCND3, CDK6	cyclin D3, cyclin dependent kinase 6	1	208	0.32
MO0000032561	Yes(h)	YES1	YES proto-oncogene 1, Src family tyrosine kinase	1	259	0.51
MO0000031189	PKCdelta(h)	PRKCD	protein kinase C delta	1	283	0.86
MO0000017291	integrins	ITGA1, ITGA2B, ITGA3, ITGA4, ITGA5, ITGA6, ITGA8, ITGA9, ITGAL, ITGAV, ITGB1, ITGB2, ITGB3, ITGB4, I...	integrin subunit alpha 1, integrin subunit alpha 2b, integrin subunit alpha 3, integrin subunit alph...	1	284	0.51
MO0000019352	traf6(h)	TRAF6	TNF receptor associated factor 6	1	295	0.23
MO0000059577	PKCdelta(h)	PRKCD	protein kinase C delta	1	300	0.86
MO0000022217	MKK3(h){p}	MAP2K3	mitogen-activated protein kinase kinase 3	1	303	0.45
MO0000045420	LOK(h)	STK10	serine/threonine kinase 10	1	303	0.51
MO0000009403	MKK3(h)	MAP2K3	mitogen-activated protein kinase kinase 3	1	316	0.45

*Table 10. Master regulators that may govern the regulation of **down-regulated** genes in Myc_induce vs. Control. **Total rank** is the sum of the ranks of the master molecules sorted by keynode score, CMA score, transcriptomics and proteomics data.*

See full table →

ID	Master molecule name	Gene symbol	Gene description	Contained in proteome set	Total rank	logFC (transcriptome)
MO000033313	PKACA(h)	PRKACA	protein kinase cAMP-activated catalytic subunit alpha	1	60	-1.18
MO000102457	PKACA-isoform1(h)	PRKACA	protein kinase cAMP-activated catalytic subunit alpha	1	194	-1.18
MO000102458	PKACA-isoform2(h)	PRKACA	protein kinase cAMP-activated catalytic subunit alpha	1	209	-1.18
MO000020249	26S proteasome(h)	PSMA7, PSMC2, PSMC3, PSMC5, PSMD4, PSMD5	proteasome 20S subunit alpha 7, proteasome 26S subunit, ATPase 2, proteasome 26S subunit, ATPase 3, ...	1	223	-0.37
MO000021208	Caspase-6(h)	CASP6	caspase 6	1	254	-0.73
MO000038322	LPS:lbp:CD14:TLR4:MD-2:MyD88:IRAK-1{pS376}{pT387}	CD14, IRAK1, LBP, LY96, MYD88, TLR4	CD14 molecule, MYD88 innate immune signal transduction adaptor, interleukin 1 receptor associated ki...	1	295	-0.82
MO000044885	PP1-alpha(h)	PPP1CA	protein phosphatase 1 catalytic subunit alpha	1	315	-0.32
MO000004672	ERK1(h)	MAPK3	mitogen-activated protein kinase 3	1	317	-0.79
MO000017291	integrins	ITGA1, ITGA2B, ITGA3, ITGA4, ITGA5, ITGA6, ITGA8, ITGA9, ITGAL, ITGAV, ITGB1, ITGB2, ITGB3, ITGB4, I...	integrin subunit alpha 1, integrin subunit alpha 2b, integrin subunit alpha 3, integrin subunit alph...	1	335	-1.41
MO000124674	EPHB2(h)	EPHB2	EPH receptor B2	1	337	-0.91

The intracellular regulatory pathways controlled by the above-mentioned master regulators are depicted in Figures 15 and 16. These diagrams display the connections between identified transcription factors, which play important roles in the regulation of differentially expressed genes, and selected master regulators, which are responsible for the regulation of these TFs.

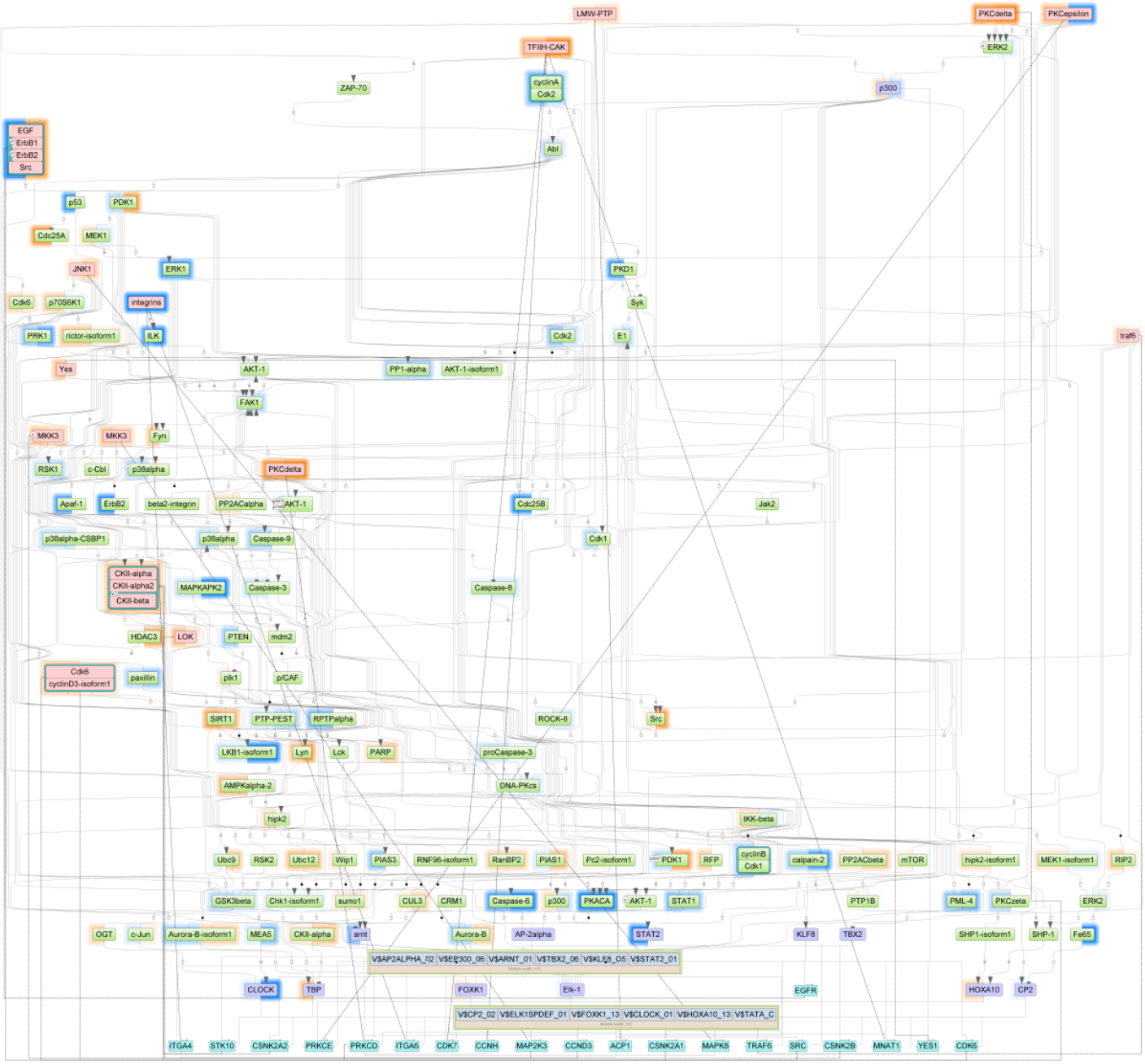


Figure 15. Diagram of intracellular regulatory signal transduction pathways of up-regulated genes in Myc_induce vs. Control. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate molecules, which have been added to the network during the search for master regulators from selected TFs. Orange and blue frames highlight molecules that are encoded by up- and downregulated genes, resp. The left half of a highlighting frame corresponds to transcriptomic data, the right one to proteomic data.

See full diagram →

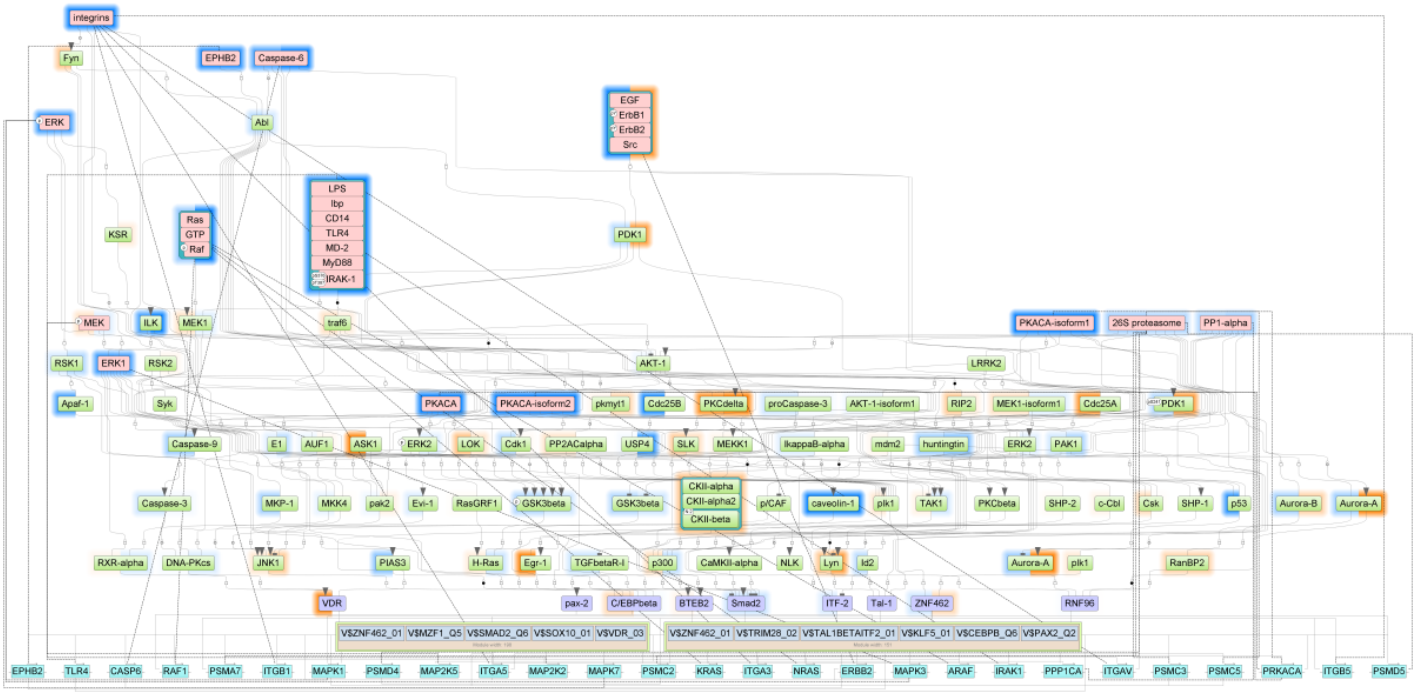


Figure 16. Diagram of intracellular regulatory signal transduction pathways of down-regulated genes in Myc_induce vs. Control. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate molecules, which have been added to the network during the search for master regulators from selected TFs. Orange and blue frames highlight molecules that are encoded by up- and downregulated genes, resp. The left half of a highlighting frame corresponds to transcriptomic data, the right one to proteomic data.

See full diagram →

4. Finding prospective drug targets

The identified master regulators that may govern pathology associated genes were checked for druggability potential using HumanPSD™ [5] database of gene-disease-drug assignments and PASS [11-13] software for prediction of biological activities of chemical compounds on the basis of a (Q)SAR approach. Respectively, for each master regulator protein we have computed two Druggability scores: HumanPSD Druggability score and PASS Druggability score. Where Druggability score represents the number of drugs that are potentially suitable for inhibition (or activation) of the corresponding target either according to the information extracted from medical literature (from HumanPSD™ database) or according to cheminformatics predictions of compounds activity against the examined target (from PASS software).

The cheminformatics druggability check is done using a pre-computed database of spectra of biological activities of chemical compounds from a library of all small molecular drugs from HumanPSD™ database, 2507 pharmaceutically active known chemical compounds in total. The spectra of biological activities has been computed using the program PASS [11-13] on the basis of a (Q)SAR approach.

If both Druggability scores were below defined thresholds (see Method section for the details) such master regulator proteins were not used in further analysis of drug prediction.

As a result we created the following two tables of prospective drug targets (top targets are shown here):

*Table 11. Prospective drug targets selected from full list of identified master regulators filtered by Druggability score from HumanPSD™ database. **Druggability score** contains the number of drugs that are potentially suitable for inhibition (or activation) of the target. The drug targets are sorted according to the **Total rank** which is the sum of three ranks computed on the basis of the three scores: keynode score, CMA score and expression change score (logFC, if present). See Methods section for details.*

See full table →

Gene symbol	Gene Description	Druggability score	Contained in proteome set	Total rank	logFC (transcriptome)
ITGA4	integrin subunit alpha 4	8	1	284	0.51
MERTK	MER proto-oncogene, tyrosine kinase	1	0	347	0.95
CDK7	cyclin dependent kinase 7	2	1	411	0.63
CDK6	cyclin dependent kinase 6	4	1	451	0.32
CSK	C-terminal Src kinase	2	1	466	0.27
LYN	LYN proto-oncogene, Src family tyrosine kinase	4	1	477	0.46

*Table 12. Prospective drug targets selected from full list of identified master regulators filtered by Druggability score predicted by PASS software. Here, the **Druggability score** for master regulator proteins is computed as a sum of PASS calculated probabilities to be active as a target for various small molecular compounds. The drug targets are sorted according to the **Total rank** which is the sum of three ranks computed on the basis of the three scores: keynode score, CMA score and expression change score (logFC, if present). See Methods section for details.*

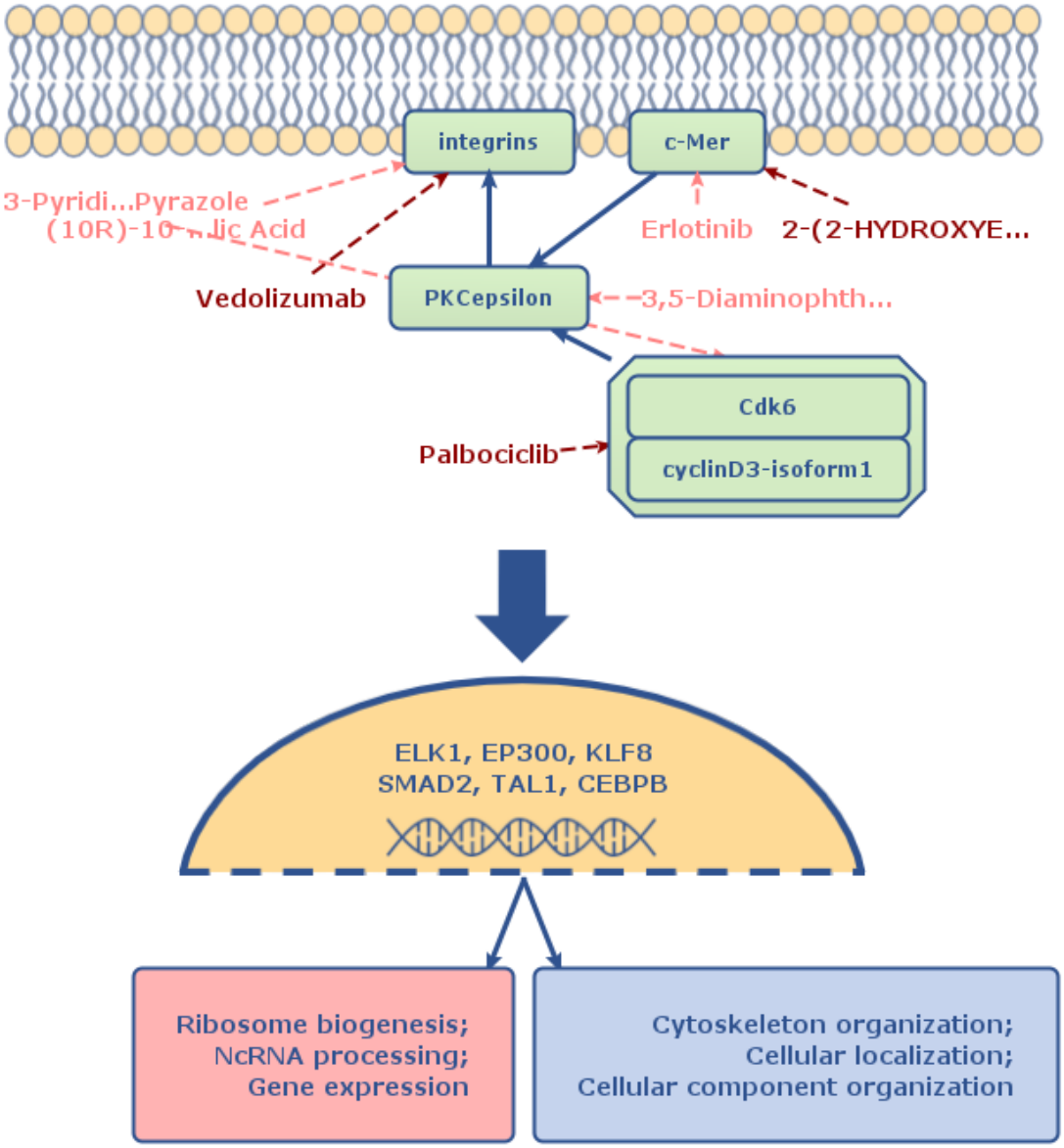
See full table →

Gene symbol	Gene Description	Druggability score	Contained in proteome set	Total rank	logFC (transcriptome)
CCND3	cyclin D3	7.89	1	208	0.32
ITGA6	integrin subunit alpha 6	2.95	1	284	0.51
ITGA4	integrin subunit alpha 4	3.42	1	284	0.51
MERTK	MER proto-oncogene, tyrosine kinase	18.05	0	347	0.95
SIRT1	sirtuin 1	4.27	1	351	0.75
PRKCE	protein kinase C epsilon	14.06	1	361	0.5

Below we represent schematically the main mechanism of the studied pathology. In the schema we considered the top two drug targets of each of the two categories computed above. In addition we have added two top identified master regulators for which no drugs may be identified yet, but that are playing the crucial role in the molecular mechanism of the studied pathology. Thus the molecular mechanism of the studied pathology was predicted to be mainly based on the following key master regulators:

- integrins
- c-Mer
- Cdk6:cyclinD3-isoform1
- PKCepsilon

This result allows us to suggest the following schema of affecting the molecular mechanism of the studied pathology:



Drugs which are shown on this schema: 2-(2-HYDROXYETHYLAMINO)-6-(3-CHLOROANILINO)-9-ISOPROPYLPURINE, 3-Pyridin-4-Yl-2,4-Dihydro-Indeno[1,2-.C.]Pyrazole, (10R)-10-Formyl-5,8,10-Trideazafolic Acid, Erlotinib, Palbociclib, Vedolizumab and 3,5-Diaminophthalhydrazide, should be considered as a prospective research initiative for further drug repurposing and drug development. These drugs were selected as top matching treatments to the most prospective drug targets of the studied pathology, however, these results should be considered with special caution and are to be used for research purposes only, as there is not enough clinical information for adapting these results towards immediate treatment of patients.
The drugs given in dark red color on the schema are FDA approved drugs or drugs which have gone through various phases of clinical trials as active treatments against the selected targets.
The drugs given in pink color on the schema are drugs, which were cheminformatically predicted to be active against the selected targets.

5. Identification of potential drugs

In the last step of the analysis we strived to identify known activities as well as drugs with cheminformatically predicted activities that are potentially suitable for inhibition (or activation) of the identified molecular targets in the context of specified human diseases(s).

Proposed drugs are top ranked drug candidates, that were found to be active on the identified targets and were selected from 4 categories:

1. FDA approved drugs or used in clinical trials drugs for the studied pathology;
2. Repurposing drugs used in clinical trials for other pathologies;
3. Drugs, predicted by PASS to be active against identified drug targets and against the studied pathology;
4. Drugs, predicted by PASS to be active against identified drug targets but for other pathologies.

Proposed drugs were selected on the basis of Drug rank which was computed from two scores:

- Target activity score (depends on ranks of all targets that were found for the selected drug);
- Disease activity score (weighted sum of number of clinical trials on disease(s) under study where the selected drug is known to be applied or PASS Disease activity score - cheminformatically predicted property of the compound to be active against the studied disease(s)).

You can refer to the Methods section for more details on drug ranking procedure.

Top drugs of each category are given in the tables below:

Drugs approved in clinical trials

Table 13. FDA approved drugs or drugs used in clinical trials for the studied pathology (most promising treatment candidates selected for the identified drug targets on the basis of literature curation in HumanPSD™ database)
See full table →

Name	Target names	Drug rank	Disease activity score	Phase 4	Status (provided by Drugbank)
Bosutinib	SRC, MAP2K1, LYN	23	2	Leukemia, Myeloid	small molecule, approved
Temsirolimus	MTOR	57	5	Carcinoma, Renal Cell, Hodgkin Disease, Lymphoma, Lymphoma, Non-Hodgkin, Noma	small molecule, approved
Everolimus	MTOR	57	5	Breast Neoplasms, Carcinoma, Hepatocellular, Carcinoma, Renal Cell, Communicable Diseases, Coronary Artery Disease, Cysts, Cytomegalovirus Infections...	small molecule, approved
Aflibercept	VEGFA	93	6	Central Serous Chorioretinopathy, Choroidal Neovascularization, Cysts, Diabetic Retinopathy, Edema, Macular Degeneration, Macular Edema...	biotech, approved
Acetylcysteine	IKBKB, CHUK	106	1	Acute Kidney Injury, Alcoholism, Anemia, Atherosclerosis, Atrophy, Bipolar Disorder, Bronchiectasis...	small molecule, approved

Repurposing drugs

Table 14. Repurposed drugs used in clinical trials for other pathologies (prospective drugs against the identified drug targets on the basis of literature curation in HumanPSD™ database)
See full table →

Name	Target names	Drug rank	Phase 4	Status (provided by Drugbank)
Vedolizumab	ITGA4	26	Colitis, Colitis, Ulcerative, Crohn Disease, Ulcer	biotech, approved
Dasatinib	SRC, YES1, FYN, ABL2	32	Leukemia, Leukemia, Lymphoid, Leukemia, Myelogenous, Chronic, BCR-ABL Positive, Leukemia, Myeloid, Precursor Cell Lymphoblastic Leukemia-Lymphoma	small molecule, approved, investigational
Tinzaparin	ITGA4	39	Diabetes Mellitus, Embolism, Fetal Growth Retardation, Kidney Failure, Chronic, Pulmonary Embolism, Renal Insufficiency, Thromboembolism...	small molecule, approved
Ingenol Mebutate	PRKCD, PRKCA	40	Keratosis, Keratosis, Actinic	small molecule, approved
Vitamin E	PPP2CB, PRKCA, PPP2CA	41	Angina Pectoris, Variant, Asphyxia, Cicatrix, Cicatrix, Hypertrophic, Diabetes Mellitus, Dyslipidemias, Epilepsy...	small molecule, approved, nutraceutical

Table 15. Prospective drugs, predicted by PASS software to be active against the identified drug targets with predicted activity against the studied disease(s) (drug candidates predicted with the cheminformatics tool PASS)
See full table →

Name	Target names	Drug rank	Target activity score
Risedronate	PTPRR, PTPN3, PTPRF, PTPN1, PTPN2, PTPN13, UBASH3B...	41	0.69
Temozolomide	CSNK1G1, CSNK1D, PLK3, CSNK1A1, CSNK1E	85	0.26
D-Myo-Inositol-Hexasulphate	CDC25A, VEGFA, DUSP2, DUSP5, DUSP7, ACP1, DUSP9...	93	0.22
Sorafenib	MAPK8, CAMKK1, GRK2, MAPK6, BRAF	96	0.17
(CHLOROACETYL)CARBAMIC ACID (3R,4S,5S,5R)-5-METHOXY-4-[(2R,3R)-2-METHYL-3-(3-METHYL-2-BUTENYL)OXIRAN...	VEGFA, MTOR	145	4.7E-2

Table 16. Prospective drugs, predicted by PASS software to be active against the identified drug targets, though without cheminformatically predicted activity against the studied disease(s) (drug candidates predicted with the cheminformatics tool PASS)
See full table →

Name	Target names	Drug rank	Target activity score
Erlotinib	CLK4, VEGFA, EGFR, MAP2K3, SRC, MAP2K7, MERTK...	26	1.91
4-[3-Hydroxyanilino]-6,7-Dimethoxyquinazoline	CLK4, VEGFA, RIPK2, EGFR, SRC, MERTK, LYN...	39	1.38
Dithiane Diol	PTPRR, PTPN3, PTPRF, PTPN1, PTPN2, PTPN13, UBASH3B...	51	1.17
N-[4-(3-BROMO-PHENYLAMINO)-QUINAZOLIN-6-YL]-ACRYLAMIDE	CLK4, VEGFA, RIPK2, EGFR, SRC, GRK2, MERTK...	57	1.56
4-[3-Methylsulfanylanilino]-6,7-Dimethoxyquinazoline	CLK4, VEGFA, EGFR, SRC, MERTK, CLK1, FYN...	64	1.17

As the result of drug search we propose the following drugs as most promising candidates for treating the pathology under study: Bosutinib, Vedolizumab, Risedronate and Erlotinib. These drugs were selected for acting on the following targets: LYN, ITGA4, DUSP2 and MERTK, which were predicted to be active in the molecular mechanism of the studied pathology.

The selected drugs are top ranked drug candidates from each of the four categories of drugs: (1) FDA approved drugs or used in clinical trials drugs for the studied pathology; (2) repurposing drugs used in clinical trials for other pathologies; (3) drugs, predicted by PASS software to be active against the studied pathology; (4) drugs, predicted by PASS software to be repurposed from other pathologies.

6. Conclusion

We applied the software package "Genome Enhancer" to a multi-omics data set that contains *transcriptomics and proteomics* data. The study is done in the context of *Neoplasm Metastasis and Osteosarcoma*. The data were pre-processed, statistically analyzed and differentially expressed genes were identified. Also checked was the enrichment of GO or disease categories among the studied gene sets.

We propose the following drugs as most promising candidates for treating the pathology under study:

Bosutinib, Vedolizumab, Risedronate and Erlotinib

These drugs were selected for acting on the following targets: LYN, ITGA4, DUSP2 and MERTK, which were predicted to be involved in the molecular mechanism of the pathology under study.

The identified molecular mechanism of the studied pathology was predicted to be mainly based on the following key drug targets:

integrins, c-Mer, Cdk6:cyclinD3-isoform1 and PKCepsilon

These potential drug targets should be considered as a prospective research initiative for further drug repurposing and drug development purposes. The following drugs were predicted as, matching those drug targets: 2-(2-HYDROXYETHYLAMINO)-6-(3-CHLOROANILINO)-9-ISOPROPYLPURINE, 3-Pyridin-4-Yl-2,4-Dihydro-Indeno[1,2-.C.]Pyrazole, (10R)-10-Formyl-5,8,10-Trideazafolic Acid, Erlotinib, Palbociclib, Vedolizumab and 3,5-Diaminophthalhydrazide. These drugs should be considered with special caution for research purposes only.

In this study, we came up with a detailed signal transduction network regulating differentially expressed genes in the studied pathology. In this network we have revealed the following top master regulators (signaling proteins and their complexes) that play a crucial role in the molecular mechanism of the studied pathology, which can be proposed as the most promising molecular targets for further drug repurposing and drug development initiatives.

- integrins
- c-Mer
- Cdk6:cyclinD3-isoform1
- PKCepsilon

Potential drug compounds which can be affecting these targets can be found in the "Finding prospective drug targets" section.

7. Methods

Databases used in the study

Transcription factor binding sites in promoters and enhancers of differentially expressed genes were analyzed using known DNA-binding motifs described in the TRANSFAC® library, release 2021.1 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transfac).
The master regulator search uses the TRANSPATH® database (BIOBASE), release 2021.1 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transpath). A comprehensive signal transduction network of human cells is built by the software on the basis of reactions annotated in TRANSPATH®.
The information about drugs corresponding to identified drug targets and clinical trials references were extracted from HumanPSD™ database, release 2021.1 (https://genexplain.com/humanpsd).
The Ensembl database release Human100.38 (hg38) (http://www.ensembl.org) was used for gene IDs representation and Gene Ontology (GO) (http://geneontology.org) was used for functional classification of the studied gene set.

Methods for the analysis of enriched transcription factor binding sites and composite modules

Transcription factor binding sites in promoters and enhancers of differentially expressed genes were analyzed using known DNA-binding motifs. The motifs are specified using position weight matrices (PWMs) that give weights to each nucleotide in each position of the DNA binding motif for a transcription factor or a group of them.
We search for transcription factor binding sites (TFBS) that are enriched in the promoters and enhancers under study as compared to a background sequence set such as promoters of genes that were not differentially regulated under the condition of the experiment. We denote study and background sets briefly as Yes and No sets. In the current work we used a workflow considering promoter sequences of a standard length of 1100 bp (-1000 to +100). The error rate in this part of the pipeline is controlled by estimating the adjusted p-value (using the Benjamini-Hochberg procedure) in comparison to the TFBS frequency found in randomly selected regions of the human genome (adj.p-value < 0.01).
We have applied the CMA algorithm (Composite Module Analyst) for searching composite modules [7] in the promoters and enhancers of the Yes and No sets. We searched for a composite module consisting of a

cluster of 10 TFs in a sliding window of 200-300 bp that statistically significantly separates sequences in the Yes and No sets (minimizing Wilcoxon p-value).

Methods for finding master regulators in networks

We searched for master regulator molecules in signal transduction pathways upstream of the identified transcription factors. The master regulator search uses a comprehensive signal transduction network of human cells. The main algorithm of the master regulator search has been described earlier [3,4]. The goal of the algorithm is to find nodes in the global signal transduction network that may potentially regulate the activity of a set of transcription factors found at the previous step of the analysis. Such nodes are considered as most promising drug targets, since any influence on such a node may switch the transcriptional programs of hundreds of genes that are regulated by the respective TFs. In our analysis, we have run the algorithm with a maximum radius of 12 steps upstream of each TF in the input set. The error rate of this algorithm is controlled by applying it 10000 times to randomly generated sets of input transcription factors of the same set-size. Z-score and FDR value of ranks are calculated then for each potential master regulator node on the basis of such random runs (see detailed description in [9]). We control the error rate by the FDR threshold 0.05.

Methods for analysis of pharmaceutical compounds

We seek for the optimal combination of molecular targets (key elements of the regulatory network of the cell) that potentially interact with pharmaceutical compounds from a library of known drugs and biologically active chemical compounds, using information about known drugs from HumanPSD™ and predicting potential drugs using PASS program.

Method for analysis of known pharmaceutical compounds

We selected compounds from HumanPSD™ database that have at least one target. Next, we sort compounds using "*Drug rank*" that is sum of two other ranks:

1. ranking by "Target activity score" (*T-score*$_{PSD}$),
2. ranking by "Disease activity score" (*D-score*$_{PSD}$).

"Target activity score" (*T-score*$_{PSD}$) is calculated as follows:

$$T\text{-}score_{PSD} = -\frac{|T|}{|T| + w(|AT| - |T|))} \sum_{t \in T} log_{10}\left(\frac{rank(t)}{1 + maxRank(T)}\right),$$

where *T* is set of all targets related to the compound intersected with input list, |*T*| is number of elements in *T*, *AT* and |*AT*| are set set of all targets related to the compound and number of elements in it, *w* is weight multiplier, *rank(t)* is rank of given target, *maxRank(T)* equals *max(rank(t))* for all targets *t* in *T*.
We use following formula to calculate "Disease activity score" (*D-score*$_{PSD}$):

$$D\text{-}score_{PSD} = \begin{cases} \sum_{d \in D} \sum_{p \in P} phase(d,p) \\ 0, \ D = \varnothing \end{cases},$$

where *D* is the set of selected diseases, and if *D* is empty set, *D-score*$_{PSD}$*=0*. *P* is a set of all known phases for each disease, *phase(p,d)* equals to the phase number if there are known clinical trials for the selected disease on this phase and zero otherwise.

Method for prediction of pharmaceutical compounds

In this study, the focus was put on compounds with high pharmacological efficiency and low toxicity. For this purpose, comprehensive library of chemical compounds and drugs was subjected to a SAR/QSAR analysis. This library contains 13040 compounds along with their pre-calculated potential pharmacological activities of those substances, their possible side and toxic effects, as well as the possible mechanisms of action. All biological activities are expressed as probability values for a substance to exert this activity (*Pa*). We selected compounds that satisfied the following conditions:

1. Toxicity below a chosen toxicity threshold (defines as *Pa*, probability to be active as toxic substance).
2. For all predicted pharmacological effects that correspond to a set of user selected disease(s) *Pa* is greater than a chosen effect threshold.
3. There are at least 2 targets (corresponding to the predicted activity-mechanisms) with predicted *Pa* greater than a chosen target threshold.

The maximum *Pa* value for all toxicities corresponding to the given compound is selected as the "Toxicity score". The maximum *Pa* value for all activities corresponding to the selected diseases for the given compound is used as the "Disease activity score". "Target activity score" (T-score) is calculated as follows:

$$T\text{-}score(s) = \frac{|T|}{|T| + w(|AT| - |T|))} \sum_{m \in M(s)} \left(pa(m) \sum_{g \in G(m)} IAP(g) optWeight(g) \right),$$

where *M(s)* is the set of activity-mechanisms for the given structure (which passed the chosen threshold for activity-mechanisms *Pa*); *G(m)* is the set of targets (converted to genes) that corresponds to the given activity-mechanism (m) for the given compound; *pa(m)* is the probability to be active of the activity-mechanism (m), *IAP(g)* is the invariant accuracy of prediction for gene from *G(m)*; *optWeight(g)* is the additional weight multiplier for gene. *T* is set of all targets related to the compound intersected with input list, |*T*| is number of elements in *T*, *AT* and |*AT*| are set set of all targets related to the compound and number of elements in it, *w* is weight multiplier.

"Druggability score" (D-score) is calculated as follows:

$$D\text{-}score(g) = IAP(g) \sum_{s \in S(g)} \sum_{m \in M(s,g)} pa(m),$$

where *S(g)* is the set of structures for which target list contains given target, *M(s,g)* is the set of activity-mechanisms (for the given structure) that corresponds to the given gene, *pa(m)* is the probability to be active of the activity-mechanism (m), *IAP(g)* is the invariant accuracy of prediction for the given gene.

8. References

1. Kel A, Voss N, Jauregui R, Kel-Margoulis O, Wingender E. Beyond microarrays: Finding key transcription factors controlling signal transduction pathways. *BMC Bioinformatics.* **2006**;7(S2), S13. doi:10.1186/1471-2105-7-s2-s13
2. Stegmaier P, Voss N, Meier T, Kel A, Wingender E, Borlak J. Advanced Computational Biology Methods Identify Molecular Switches for Malignancy in an EGF Mouse Model of Liver Cancer. *PLoS ONE.* **2011**;6(3):e17738. doi:10.1371/journal.pone.0017738
3. Koschmann J, Bhar A, Stegmaier P, Kel A, Wingender E. "Upstream Analysis": An Integrated Promoter-Pathway Analysis Approach to Causal Interpretation of Microarray Data. *Microarrays.* **2015**;4(2):270-286. doi:10.3390/microarrays4020270.
4. Kel A, Stegmaier P, Valeev T, Koschmann J, Poroikov V, Kel-Margoulis OV, and Wingender E. Multi-omics "upstream analysis" of regulatory genomic regions helps identifying targets against methotrexate resistance of colon cancer. *EuPA Open Proteom.* **2016**;13:1-13. doi:10.1016/j.euprot.2016.09.002
5. Michael H, Hogan J, Kel A et al. Building a knowledge base for systems pathology. *Brief Bioinformatics.* **2008**;9(6):518-531. doi:10.1093/bib/bbn038
6. Matys V, Kel-Margoulis OV, 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 AE, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **2006**;34(90001):D108-D110. doi:10.1093/nar/gkj143
7. Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **2003**;31(13):3576-3579. doi:10.1093/nar/gkg585
8. Waleev T, Shtokalo D, Konovalova T, Voss N, Cheremushkin E, Stegmaier P, Kel-Margoulis O, Wingender E, Kel A. Composite Module Analyst: identification of transcription factor binding site combinations using genetic algorithm. *Nucleic Acids Res.* **2006**;34(Web Server issue):W541-5.
9. Krull M, Pistor S, Voss N, Kel A, Reuter I, Kronenberg D, Michael H, Schwarzer K, Potapov A, Choi C, Kel-Margoulis O, Wingender E. TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res.* **2006**;34(90001):D546-D551. doi:10.1093/nar/gkj107
0. Boyarskikh U, Pintus S, Mandrik N, Stelmashenko D, Kiselev I, Evshin I, Sharipov R, Stegmaier P, Kolpakov F, Filipenko M, Kel A. Computational master-regulator search reveals mTOR and PI3K pathways responsible for low sensitivity of NCI-H292 and A427 lung cancer cell lines to cytotoxic action of p53 activator Nutlin-3. *BMC Med Genomics.* **2018**;11(1):12. doi:10.1186/1471-2105-7-s2-s13

1. Filimonov D, Poroikov V. Probabilistic Approaches in Activity Prediction. Varnek A, Tropsha A. *Cheminformatics Approaches to Virtual Screening.* Cambridge (UK): RSC Publishing. **2008**;:182-216.
2. Filimonov DA, Poroikov VV. Prognosis of specters of biological activity of organic molecules. *Russian chemical journal.* **2006**;50(2):66-75 (russ)
3. Filimonov D, Poroikov V, Borodina Y, Gloriozova T. Chemical Similarity Assessment Through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *ChemInform.* **1999**;39(4):666-670. doi:10.1002/chin.199940210

Thank you for using the Genome Enhancer!

In case of any questions please contact us at support@genexplain.com

Supplementary material

1. Supplementary table 1 - Up-regulated genes
2. Supplementary table 2 - Down-regulated genes
3. Supplementary table 3 - Detailed report. Composite modules and master regulators (up-regulated genes in Myc_induce vs. Control).
4. Supplementary table 4 - Detailed report. Composite modules and master regulators (down-regulated genes in Myc_induce vs. Control).
5. Supplementary table 5 - Detailed report. Pharmaceutical compounds and drug targets.

Disclaimer

Decisions regarding care and treatment of patients should be fully made by attending doctors. The predicted chemical compounds listed in the report are given only for doctor's consideration and they cannot be treated as prescribed medication. It is the physician's responsibility to independently decide whether any, none or all of the predicted compounds can be used solely or in combination for patient treatment purposes, taking into account all applicable information regarding FDA prescribing recommendations for any therapeutic and the patient's condition, including, but not limited to, the patient's and family's medical history, physical examinations, information from various diagnostic tests, and patient preferences in accordance with the current standard of care. Whether or not a particular patient will benefit from a selected therapy is based on many factors and can vary significantly.

The compounds predicted to be active against the identified drug targets in the report are not guaranteed to be active against any particular patient's condition. GeneXplain GmbH does not give any assurances or guarantees regarding the treatment information and conclusions given in the report. There is no guarantee that any third party will provide a refund for any of the treatment decisions made based on these results. None of the listed compounds was checked by Genome Enhancer for adverse side-effects or even toxic effects.

The analysis report contains information about chemical drug compounds, clinical trials and disease biomarkers retrieved from the HumanPSD™ database of gene-disease assignments maintained and exclusively distributed worldwide by geneXplain GmbH. The information contained in this database is collected from scientific literature and public clinical trials resources. It is updated to the best of geneXplain's knowledge however we do not guarantee completeness and reliability of this information leaving the final checkup and consideration of the predicted therapies to the medical doctor.

The scientific analysis underlying the Genome Enhancer report employs a complex analysis pipeline which uses geneXplain's proprietary Upstream Analysis approach, integrated with TRANSFAC® and TRANSPATH® databases maintained and exclusively distributed worldwide by geneXplain GmbH. The pipeline and the databases are updated to the best of geneXplain's knowledge and belief, however, geneXplain GmbH shall not give a warranty as to the characteristics or to the content and any of the results produced by Genome Enhancer. Moreover, any warranty concerning the completeness, up-to-dateness, correctness and usability of Genome Enhancer information and results produced by it, shall be excluded.

The results produced by Genome Enhancer, including the analysis report, severely depend on the quality of input data used for the analysis. It is the responsibility of Genome Enhancer users to check the input data quality and parameters used for running the Genome Enhancer pipeline.

Note that the text given in the report is not unique and can be fully or partially repeated in other Genome Enhancer analysis reports, including reports of other users. This should be considered when publishing any

results or excerpts from the report. This restriction refers only to the general description of analysis methods used for generating the report. All data and graphics referring to the concrete set of input data, including lists of mutated genes, differentially expressed genes/proteins/metabolites, functional classifications, identified transcription factors and master regulators, constructed molecular networks, lists of chemical compounds and reconstructed model of molecular mechanisms of the studied pathology are unique in respect to the used input data set and Genome Enhancer pipeline parameters used for the current run.