Sequence and Pathway analysis

CDK7 and MNAT1 are promising druggable targets for treating neoplasm metastasis and osteosarcoma that control activity of SMAD2, POU5F1 and TAL1 transcription factors on promoters of differentially expressed genes

Demo User
geneXplain GmbH
info@genexplain.com

Data received on 07/09/2019 ; Run on 19/02/2020 ; Report generated on 19/02/2020

Genome Enhancer release 1.9 (TRANSFAC®, TRANSPATH® and HumanPSD™ release 2020.1)

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) novel biologically 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: SMAD2, POU5F1, IRF2, TAL1 and SMAD3. The subsequent network analysis suggested CDK7, LYN, IKBKB, MNAT1 and CCNH as the most promising and druggable molecular targets. Finally, the following drugs were identified as the most promising treatment candidates: Bosutinib, Mesalazine, Flavopiridol, Streptomycin, 1-3 Sugar Ring of Pentamannosyl 6-Phosphate and Uridine.

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, new potential small molecular ligands are subsequently predicted for the revealed targets. A general druggability check is performed using a precomputed database of biologcal activities of chemical compounds from a library of about 13000 pharmaceutically most active compounds. The spectra of biological activities are computed using the program PASS on the basis of a (Q)SAR approach [11-13].

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

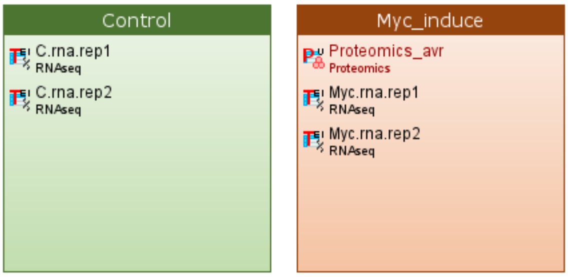


Figure 1. Annotation diagram of experimental data used in this study. With the colored boxes we show those sub-categories of the data that are compared in our analysis.

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	v-myc avian myelocytomatosis viral oncogene homolog	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 →

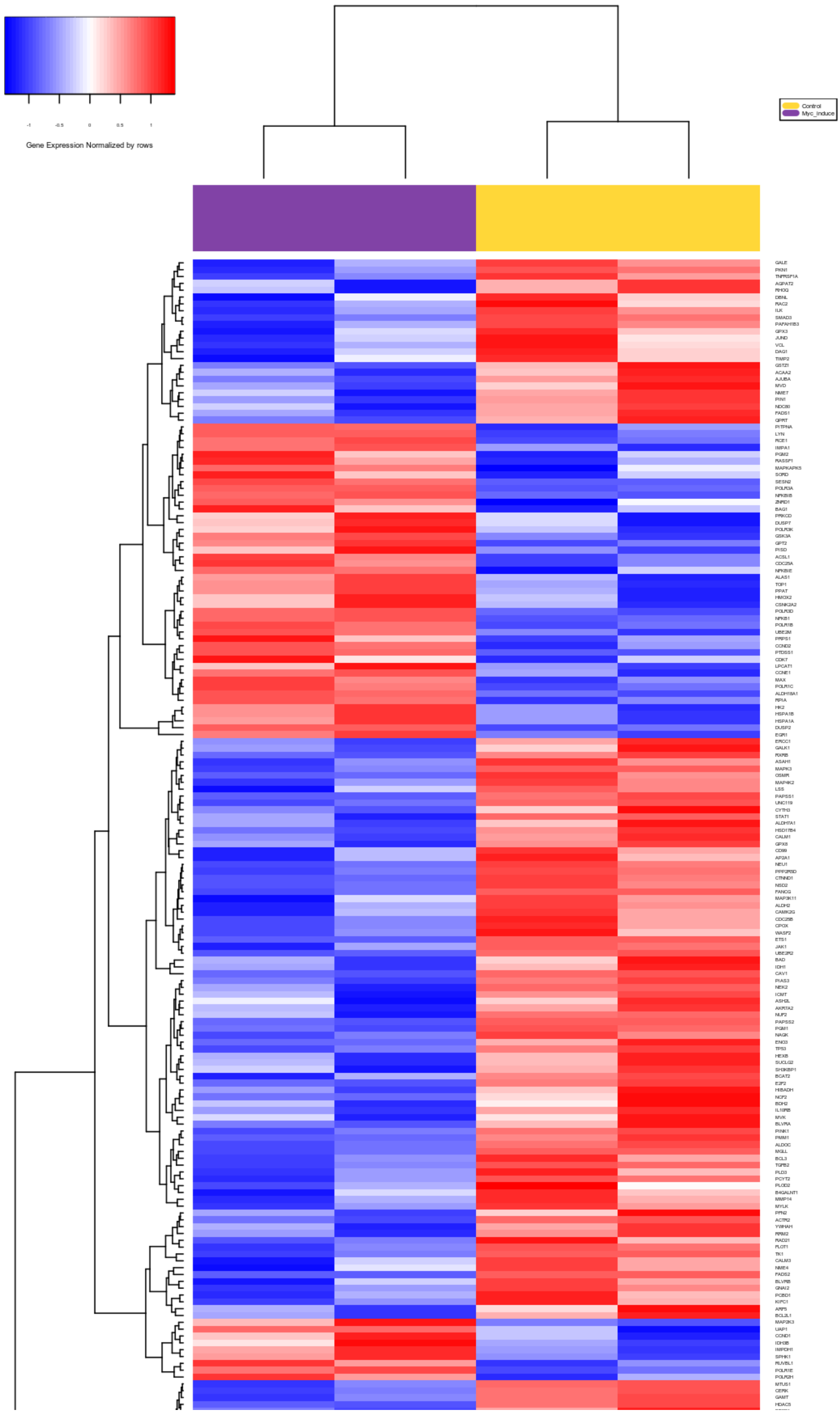
ID	Gene symbol	Gene description	logFC	P.Value	adj.P.Val
ENSG00000116774	OLFML3	olfactomedin like 3	-3.06	1.11E-4	0.13
ENSG00000138131	LOXL4	lysyl oxidase like 4	-2.62	8.88E-4	0.14
ENSG00000187867	PALM3	paralemmin 3	-2.62	2.65E-3	0.14
ENSG00000205542	TMSB4X	thymosin beta 4, X-linked	-2.58	2.22E-4	0.13
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
GALE
PKN1
TNFRSF1A
AGPAT2
RHOQ
DBNL
RAC2
ILK
SMAD3
PAFAH1B3
GPX3
JUND
VCL
DAG1
TIMP2
GSTZ1
ACAA2
AJUBA
MVD
NME7
PIN1
NDC80
FADS1
QPRT
PITPNA
LYN
RCE1
IMPA1
PGM2
RASSF1
MAPKAPK5
SORD
SESN2
POLR3A
NFKBIB
ZNRD1
BAG1
PRKCD
DUSP7
POLR3K
GSK3A
GPT2
PISD
ACSL1
CDC25A
NFKBIE
ALAS1
TOP1
PPAT
HMOX2
CSNK2A2
POLR3D
NFKB1
POLR1B
UBE2M
PRPS1
CCND2
PTDSS1
CDK7
LPCAT1
CCNE1
MAX
POLR1C
ALDH18A1
RPIA
HK2
HSPA1B
HSPA1A
DUSP2
EGR1
ERCC1
GALK1
RXRB
ASAH1
MAPK3
OSMR
MAP4K2
LSS
PAPSS1
UNC119
CYTH3
STAT1
ALDH7A1
HSD17B4
CALM1
GPX8
CD99
AP2A1
NEU1
PPP2R5D
CTNND1
NSD2
FANCG
MAP3K11
ALDH2
CAMK2G
CDC25B
CPOX
WASF2
ETS1
JAK1
UBE2R2
BAD
IDH1
CAV1
PIAS3
NEK2
ICMT
ASH2L
AKR7A2
NUF2
PAPSS2
PGM1
NAGK
ENO3
TP53
HEXB
SUCLG2
SH3KBP1
BCAT2
E2F2
HIBADH
NCF2
BDH2
IL10RB
MVK
BLVRA
PINK1
PMM1
ALDOC
MGLL
BCL3
TGFB2
PLD3
PCYT2
PLOD2
B4GALNT1
MMP14
MYLK
PFN2
ACTR2
YWHAH
RRM2
RAD21
FLOT1
TK1
CALM3
NME4
FADS2
BLVRB
GNAI2
PCBD1
KIFC1
ARF5
BCL2L1
MAP2K3
UAP1
CCND1
IDH3B
IMPDH1
SPHK1
RUVBL1
POLR1E
POLR2H
MTUS1
CERK
GAMT
HDAC5

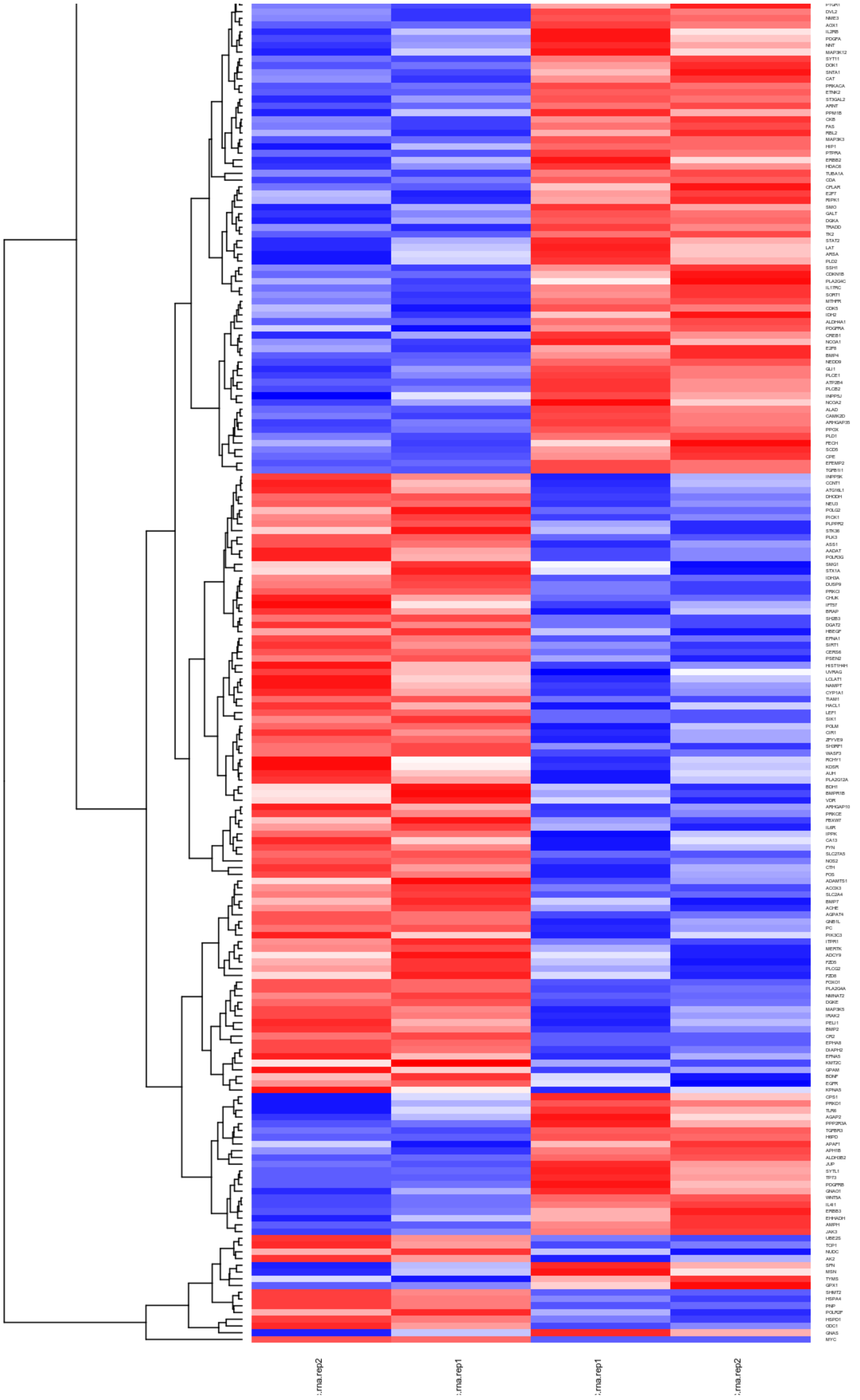

PTGR1
DVL2
NME3
AOX1
IL2RB
PDGFA
NNT
MAP3K12
SYT11
DOK1
SNTA1
CAT
PRKACA
ETNK2
ST3GAL2
ARNT
PPM1B
CKB
FAS
RBL2
MAP3K3
HIP1
PTPRA
ERBB2
HDAC6
TUBA1A
CDA
CFLAR
E2F7
RIPK1
SMO
GALT
DGKA
TRADD
TK2
STAT2
LAT
ARSA
PLD2
SSH1
CDKN1B
PLA2G4C
IL17RC
SORT1
MTHFR
CDK5
IDH2
ALDH4A1
PDGFRA
CREB1
NCOA1
E2F8
BMP4
NEDD9
GLI1
PLCE1
ATP2B4
PLCB2
INPP5J
NCOA2
ALAD
CAMK2D
ARHGAP35
PPOX
PLD1
FECH
SCD5
CPE
EFEMP2
TGFB1I1
INPP5K
CCNT1
ATG16L1
DHODH
NEU3
POLG2
PICK1
PLPPR2
STK36
PLK3
ASS1
AADAT
POLR3G
SMG1
STX1A
IDH3A
DUSP9
PRKCI
CHUK
IFT57
BRAP
SH2B3
DGAT2
HBEGF
EFNA1
SIRT1
CERS6
PSEN2
HIST1H4H
UVRAG
LCLAT1
NAMPT
CYP1A1
TIAM1
HACL1
LEF1
SIK1
POLM
CIR1
ZFYVE9
SH3RF1
WASF3
RCHY1
KDSR
AUH
PLA2G12A
BDH1
BMPR1B
VDR
ARHGAP10
PRKCE
FBXW7
IL6R
IPPK
CA13
FYN
SLC27A5
NOS2
CTH
FOS
ADAMTS1
ACOX3
SLC2A4
BMP7
ACHE
AGPAT4
GNB1L
PC
PIK3C3
ITPR1
MERTK
ADCY9
FZD5
PLCG2
FZD8
FOXO1
PLA2G4A
NMNAT2
DGKE
MAP3K5
IRAK2
PELI1
BMP2
CR2
EPHA8
DIAPH2
EFNA5
KMT2C
GPAM
BDNF
EGFR
KPNA5
CPS1
PRKD1
TLR6
AGAP2
PPP2R3A
TGFBR3
H6PD
APAF1
APH1B
ALDH3B2
JUP
SYTL1
TP73
PDGFRB
GNAO1
WNT5A
IL4I1
ERBB3
EHHADH
AMPH
JAK3
UBE2S
TCP1
NUDC
AK2
SPN
MSN
TYMS
GPX1
SHMT2
HSPA4
PNP
POLR2F
HSPD1
ODC1
GNAS
MYC
.rna.rep2
.rna.rep1
.rna.rep1
.rna.rep2

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)

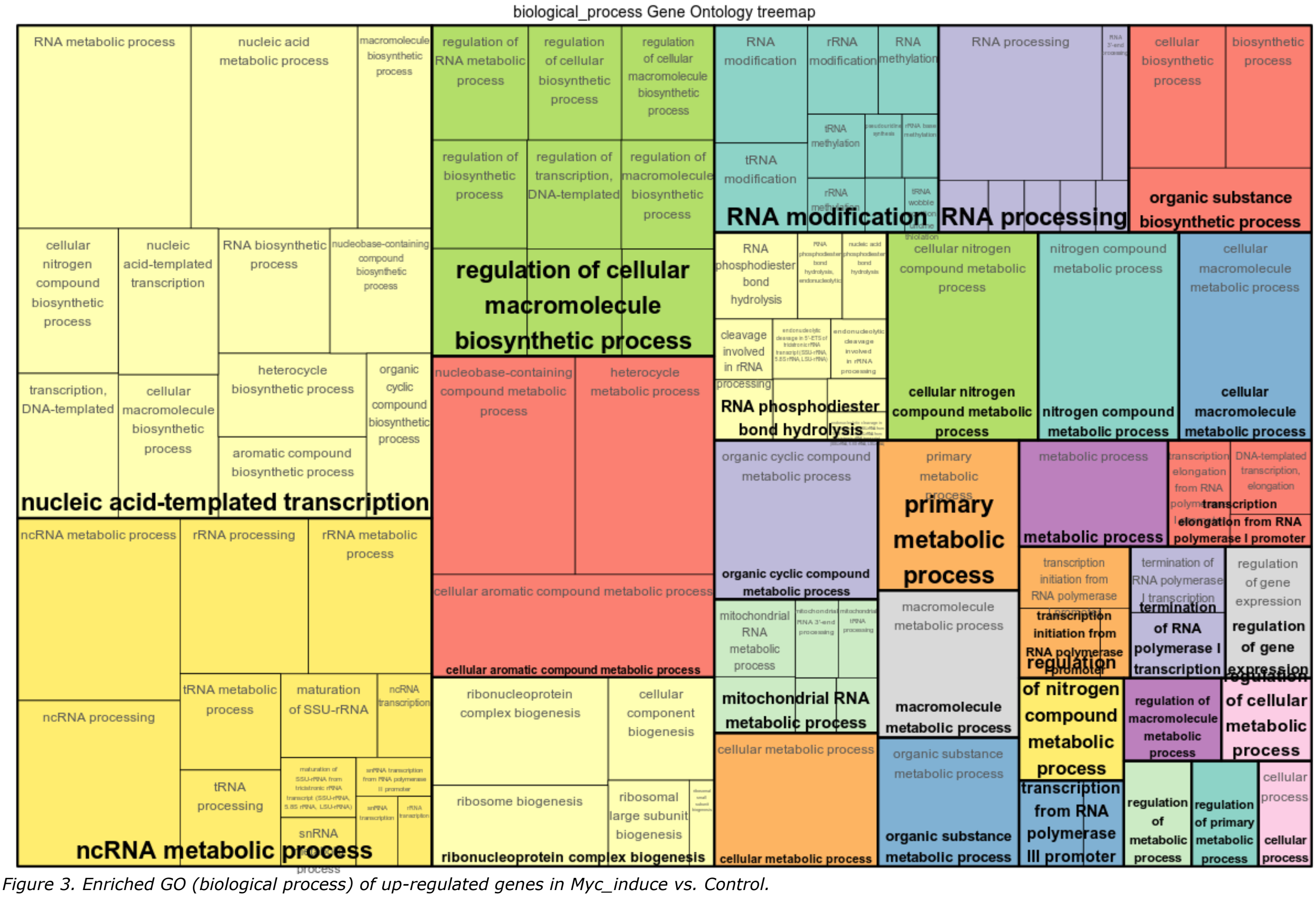


Figure 3. Enriched GO (biological process) of up-regulated genes in Myc_induce vs. Control.

Full classification →

TRANSPATH® Pathways (2020.1)

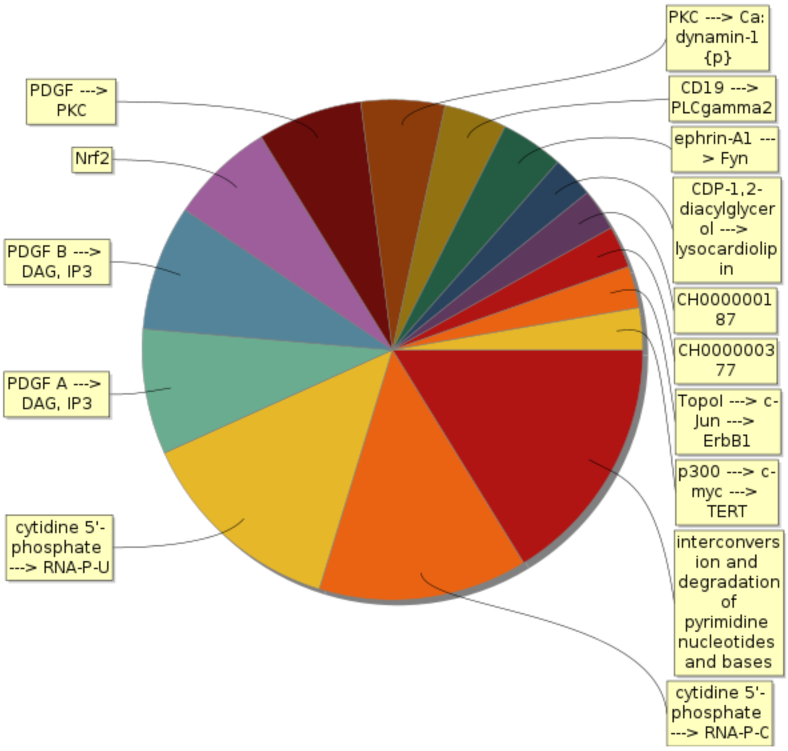


Figure 4. Enriched TRANSPATH® Pathways (2020.1) of up-regulated genes in Myc_induce vs. Control.
Full classification →

HumanPSD(TM) disease (2020.1)

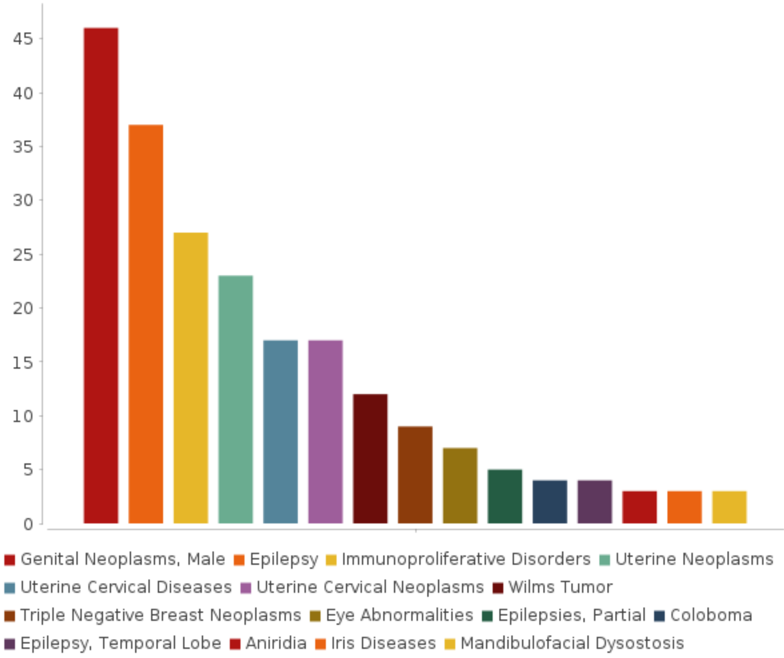


Figure 5. Enriched HumanPSD(TM) disease (2020.1) of up-regulated genes in Myc_induce vs. Control. The size of the bars correspond to the number of biomarkers of the given disease found among the input set.
Full classification →

Down-regulated genes in Myc_induce vs. Control:

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

GO (biological process)

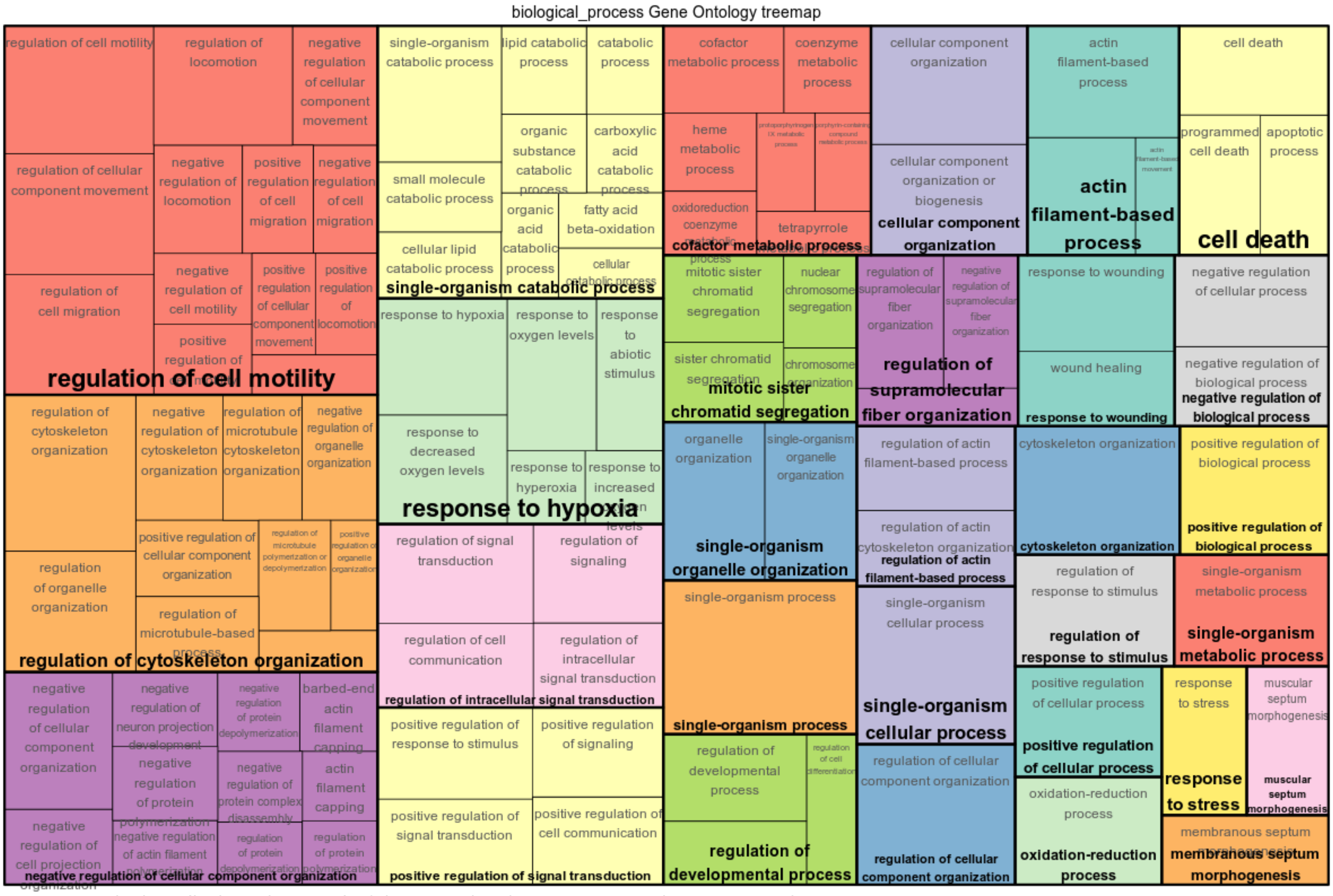


Figure 6. Enriched GO (biological process) of down-regulated genes in Myc_induce vs. Control.

Full classification →

TRANSPATH® Pathways (2020.1)

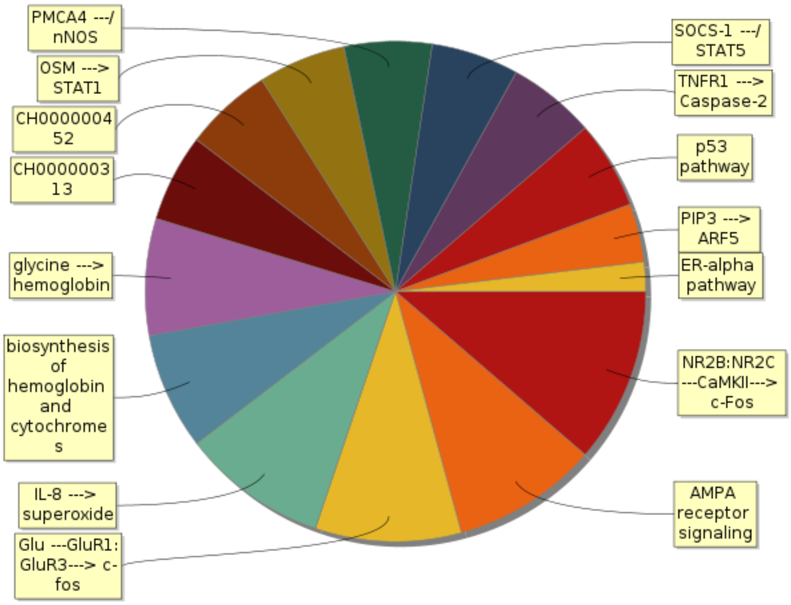


Figure 7. Enriched TRANSPATH® Pathways (2020.1) of down-regulated genes in Myc_induce vs. Control.

Full classification →

HumanPSD(TM) disease (2020.1)

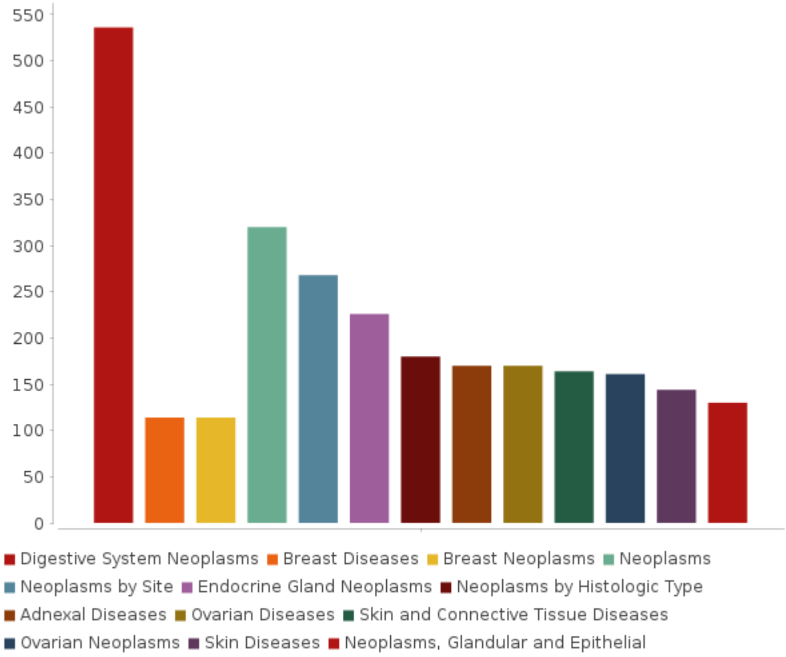


Figure 8. Enriched HumanPSD(TM) disease (2020.1) of down-regulated genes in Myc_induce vs. Control. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

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:

4653 the list of genes provided as input genes were taken for the mapping.

GO (biological process)

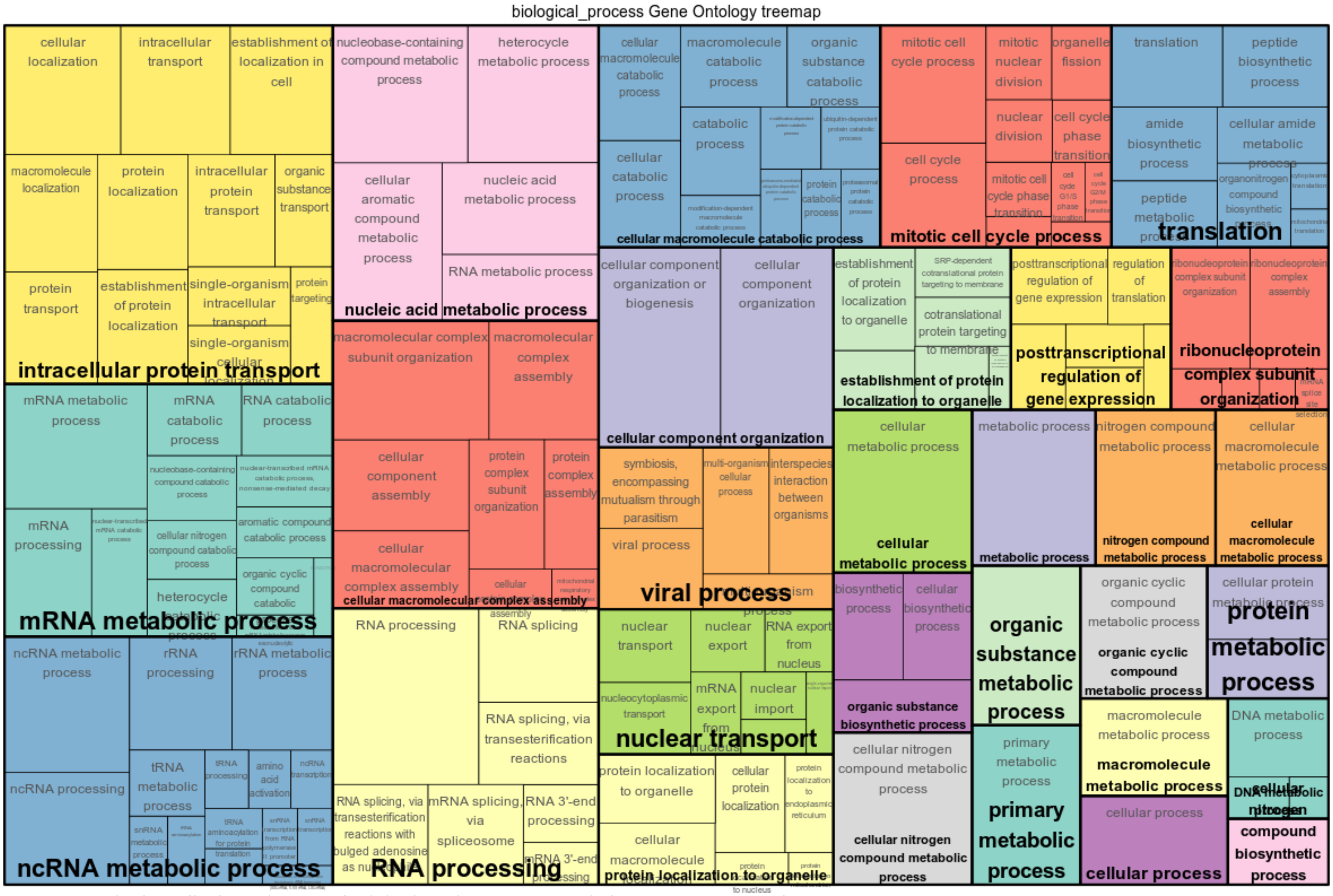


Figure 9. Enriched GO (biological process) of the list of proteins provided as input in Myc_induce.

Full classification →

TRANSPATH® Pathways (2020.1)

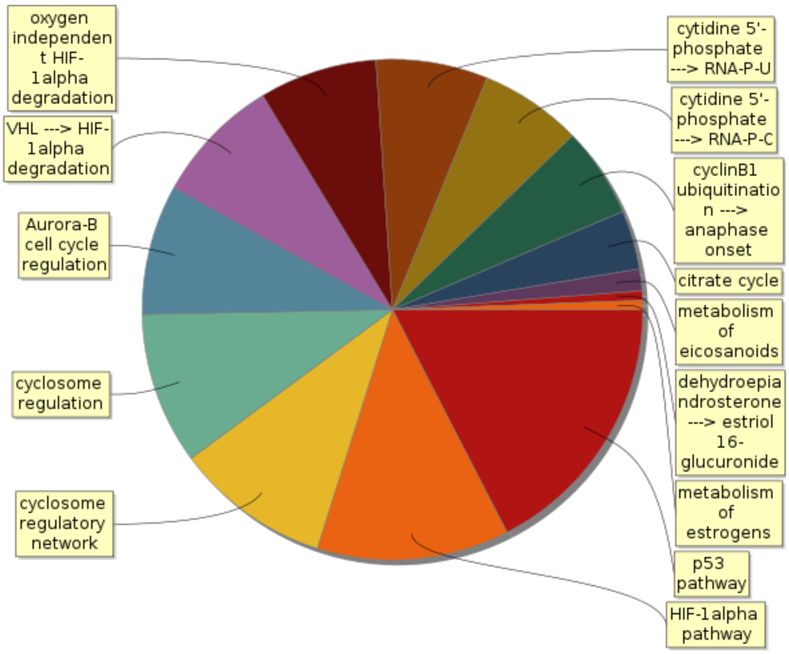


Figure 10. Enriched TRANSPATH® Pathways (2020.1) of the list of proteins provided as input in Myc_induce.

Full classification →

HumanPSD(TM) disease (2020.1)

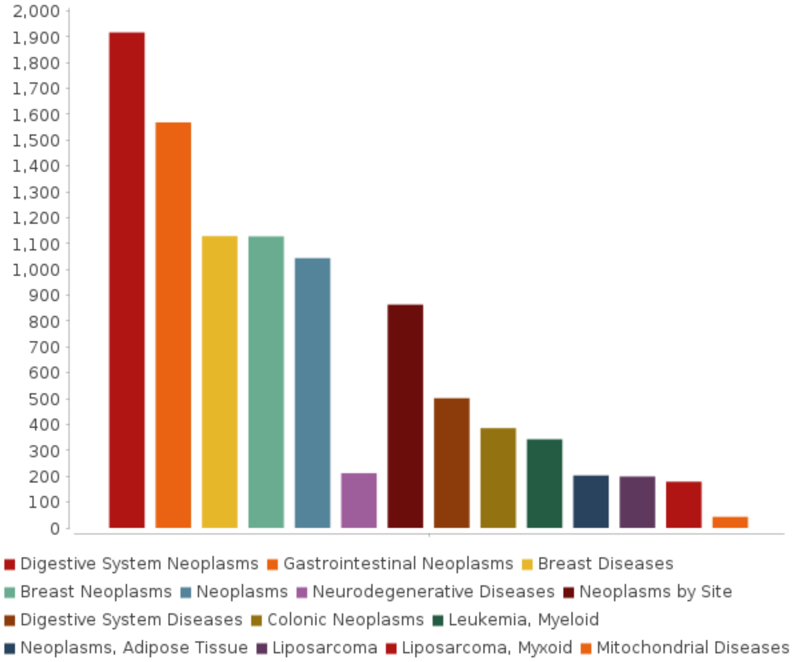


Figure 11. Enriched HumanPSD(TM) disease (2020.1) of the list of proteins provided as input in Myc_induce. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

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 4653 proteins.

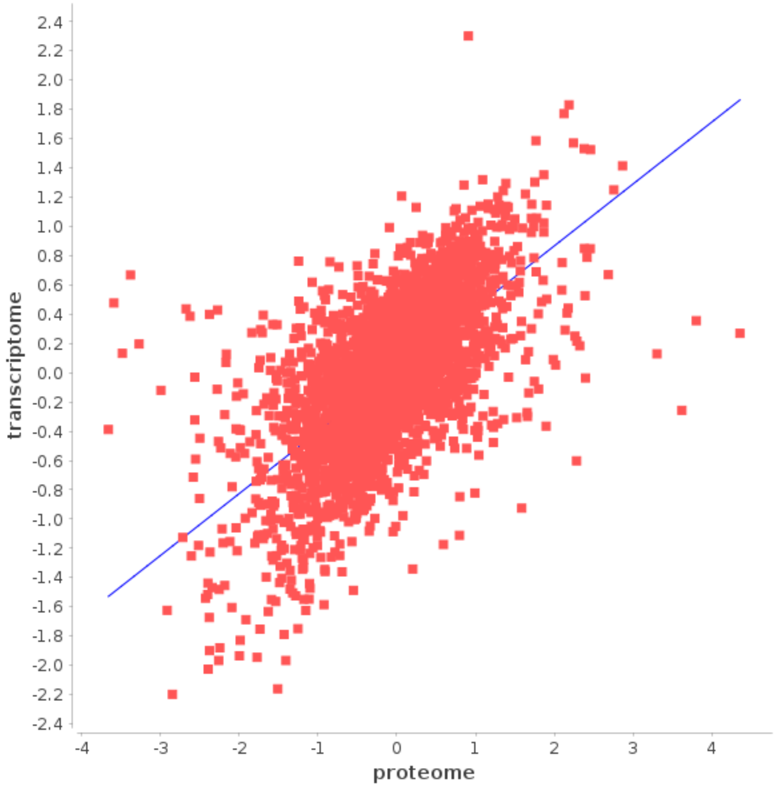


Figure 12. Comparison plot of comparison proteome vs transcriptome. X axis: protein expression value - Proteomics_avr. Y axis: LogFC of differential gene expression.

Full comparison →

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

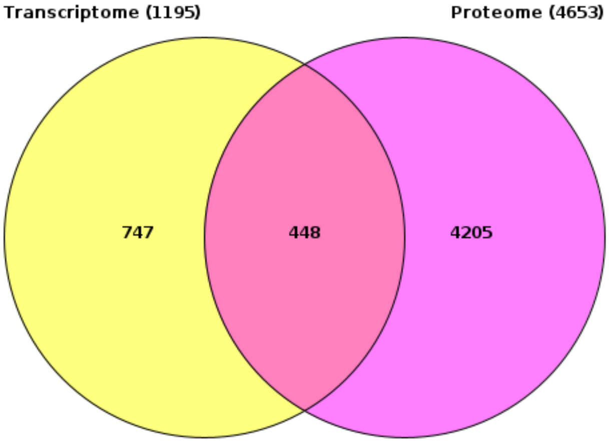


Figure 13. Intersection of up-regulated genes and the list of proteins provided as input

See full diagram →

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

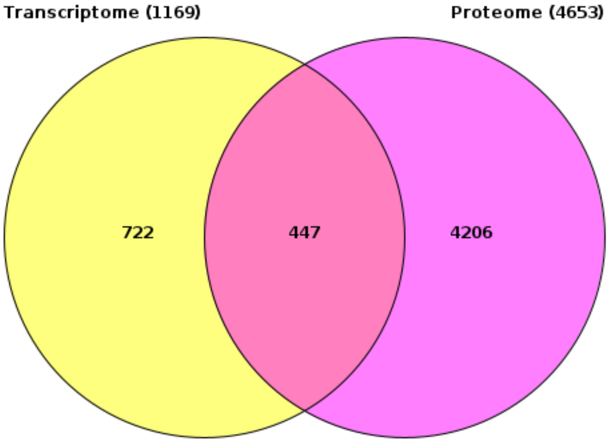


Figure 14. Intersection of down-regulated genes and the list of proteins provided as input

See full diagram →

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 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$SALL2_02	V$CP2_02	V$IRF_Q6	V$COUPTF_Q6	V$OCT4_01
0.75; N=3	0.00; N=2	0.00; N=2	0.79; N=3	0.00; N=2

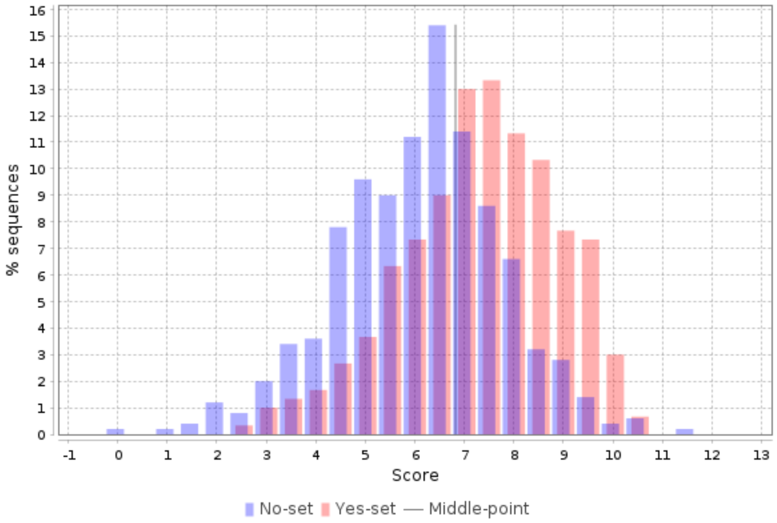
Module width: 131

Module 2:

V$SMAD2_Q6	V$TFAP2A_02	V$MZF1_Q5	V$TATA_01	V$SOX9_09	V$GCM1_06
0.00; N=3	0.88; N=3	0.96; N=2	0.00; N=3	0.87; N=3	0.83; N=2

Module width: 126

Model score (-p*log10(pval)): 10.03
Wilcoxon p-value (pval): 2.52e-21
Penalty (p): 0.487
Average yes-set score: 7.27
Average no-set score: 6.12
AUC: 0.70
Middle-point: 6.82
False-positive: 33.20%
False-negative: 33.67%



See model visualization table →

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
ENSG00000154122	ANKH	ANKH inorganic pyrophosphate transport regulator	14.8	SALL2(h), COUP-TF1(h),COUP-TF2(h), MZF-1(h), GCMa(h), CP2(h), Smad2(h), Oct3(h)...
ENSG00000083857	FAT1	FAT atypical cadherin 1	14.74	CP2(h), COUP-TF1(h),COUP-TF2(h), Oct3(h), IRF-1(h),IRF-2(h),IRF-3(h),IRF-4(h),IRF-5(h),IRF-6(h),IRF-7(h),IRF-8(h), Smad2(h), MZF-1(h), Sox-9(h)...
ENSG00000120071	KANSL1	KAT8 regulatory NSL complex subunit 1	14.67	GCMa(h), MZF-1(h), Smad2(h), COUP-TF1(h),COUP-TF2(h), TBP(h), Oct3(h), Sox-9(h)...
ENSG00000105197	TIMM50	translocase of inner mitochondrial membrane 50	14.5	Sox-9(h), COUP-TF1(h),COUP-TF2(h), MZF-1(h), Smad2(h), SALL2(h), TBP(h), GCMa(h)...
ENSG00000107951	MTPAP	mitochondrial poly(A) polymerase	14.41	Sox-9(h), MZF-1(h), TBP(h), Smad2(h), CP2(h), COUP-TF1(h),COUP-TF2(h), Oct3(h)...
ENSG00000119950	MXI1	MAX interactor 1, dimerization protein	14.39	COUP-TF1(h),COUP-TF2(h), Smad2(h), Sox-9(h), IRF-1(h),IRF-2(h),IRF-3(h),IRF-4(h),IRF-5(h),IRF-6(h),IRF-7(h),IRF-8(h), TBP(h), MZF-1(h), Oct3(h)...
ENSG00000168807	SNTB2	syntrophin beta 2	14.33	TBP(h), Smad2(h), Oct3(h), Sox-9(h), IRF-1(h),IRF-2(h),IRF-3(h),IRF-4(h),IRF-5(h),IRF-6(h),IRF-7(h),IRF-8(h), COUP-TF1(h),COUP-TF2(h), CP2(h)...
ENSG00000138386	NAB1	NGFI-A binding protein 1	14.31	TBP(h), Smad2(h), Sox-9(h), GCMa(h), CP2(h), IRF-1(h),IRF-2(h),IRF-3(h),IRF-4(h),IRF-5(h),IRF-6(h),IRF-7(h),IRF-8(h), Oct3(h)
ENSG00000144228	SPOPL	speckle type BTB/POZ protein like	14.27	COUP-TF1(h),COUP-TF2(h), Smad2(h), Sox-9(h), MZF-1(h), TBP(h), IRF-1(h),IRF-2(h),IRF-3(h),IRF-4(h),IRF-5(h),IRF-6(h),IRF-7(h),IRF-8(h), GCMa(h)...
ENSG00000070444	MNT	MAX network transcriptional repressor	14.25	SALL2(h), Sox-9(h), Smad2(h), COUP-TF1(h),COUP-TF2(h), CP2(h), MZF-1(h), IRF-1(h),IRF-2(h),IRF-3(h),IRF-4(h),IRF-5(h),IRF-6(h),IRF-7(h),IRF-8(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 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$POU6F1_02	V$SMAD2_Q6	V$SIRT6_01	V$POU6F1_03	V$YY1_Q6_03	V$TAL1BETAITF2_01	V$COUPTF_Q6
0.00; N=3	0.99; N=3	0.95; N=3	0.89; N=1	0.00; N=3	0.00; N=3	0.77; N=3

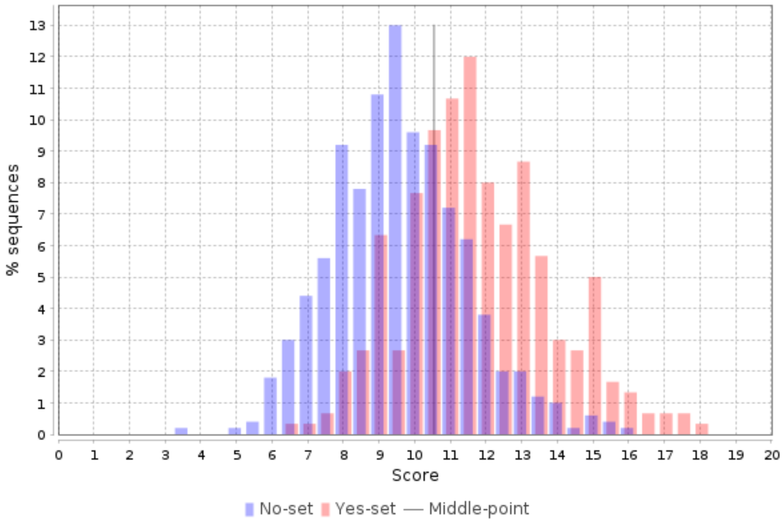
Module width: 136

Module 2:

V$NANOG_02	V$MZF1_Q5	V$BTEB2_Q3	V$AR_Q6	V$SMAD3_Q6	V$HNF4_Q6_03
0.88; N=3	0.96; N=3	0.83; N=3	0.98; N=3	0.00; N=2	0.00; N=3

Module width: 129

Model score (-p*log10(pval)): 19.09
Wilcoxon p-value (pval): 6.24e-42
Penalty (p): 0.463
Average yes-set score: 11.76
Average no-set score: 9.59
AUC: 0.79
Middle-point: 10.54
False-positive: 28.00%
False-negative: 26.33%



See model visualization table →

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
ENSG00000101997	CCDC22	coiled-coil domain containing 22	20.18	AR(h), BTEB2(h), MZF-1(h), Smad3(h), HNF-4alpha(h), SIR2L6(h), COUP-TF1(h),COUP-TF2(h)...
ENSG00000063601	MTMR1	myotubularin related protein 1	19.05	Smad2(h), COUP-TF1(h),COUP-TF2(h), YY1(h), POU6F1(h), ITF-2(h),Tal-1(h), Smad3(h), HNF-4alpha(h)...
ENSG00000105048	TNNT1	troponin T1, slow skeletal type	18.88	AR(h), COUP-TF1(h),COUP-TF2(h), HNF-4alpha(h), ITF-2(h),Tal-1(h), POU6F1(h), YY1(h), SIR2L6(h)...
ENSG00000148337	CIZ1	CDKN1A interacting zinc finger protein 1	18.82	BTEB2(h), MZF-1(h), nanog(h), HNF-4alpha(h), Smad3(h), SIR2L6(h), ITF-2(h),Tal-1(h)...
ENSG00000122861	PLAU	plasminogen activator, urokinase	18.82	COUP-TF1(h),COUP-TF2(h), Smad2(h), POU6F1(h), ITF-2(h),Tal-1(h), SIR2L6(h), YY1(h), nanog(h)...
ENSG00000115461	IGFBP5	insulin like growth factor binding protein 5	18.72	AR(h), Smad3(h), BTEB2(h), MZF-1(h), HNF-4alpha(h), YY1(h), COUP-TF1(h),COUP-TF2(h)...
ENSG00000092051	JPH4	junctophilin 4	18.53	Smad3(h), BTEB2(h), MZF-1(h), ITF-2(h),Tal-1(h), nanog(h), AR(h), POU6F1(h)...
ENSG00000113721	PDGFRB	platelet derived growth factor receptor beta	18.31	Smad3(h), HNF-4alpha(h), MZF-1(h), BTEB2(h), nanog(h), ITF-2(h),Tal-1(h), COUP-TF1(h),COUP-TF2(h)...
ENSG00000169045	HNRNPH1	heterogeneous nuclear ribonucleoprotein H1	18.18	AR(h), MZF-1(h), BTEB2(h), HNF-4alpha(h), ITF-2(h),Tal-1(h), Smad3(h), YY1(h)...
ENSG00000130176	CNN1	calponin 1	18.03	Smad2(h), YY1(h), COUP-TF1(h),COUP-TF2(h), ITF-2(h),Tal-1(h), SIR2L6(h), POU6F1(h), nanog(h)...

On the basis of the enhancer models we identified the following transcription factors potentially regulating the ***target genes*** of our interest. We found 18 and 14 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 TFin 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	5.1	1.21
MO000056618	POU5F1	POU class 5 homeobox 1	4.32	1.74
MO000007691	IRF2	interferon regulatory factor 2	4.29	1.3
MO000024736	NR2F1	nuclear receptor subfamily 2 group F member 1	4.02	8.33
MO000285816	IRF3	interferon regulatory factor 3	3.93	1.3
MO000117988	TFCP2	transcription factor CP2	3.9	1.34
MO000007703	IRF7	interferon regulatory factor 7	3.88	1.3
MO000007686	IRF1	interferon regulatory factor 1	3.75	1.3
MO000021896	TBP	TATA-box binding protein	3.64	1.26
MO000018993	SOX9	SRY-box 9	3.61	1.63

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 TFin 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
MO000032489	TAL1	TAL bHLH transcription factor 1, erythroid differentiation factor	5.14	1.34
MO000057832	SMAD3	SMAD family member 3	5.08	1.69
MO000057829	SMAD2	SMAD family member 2	4.83	1.66
MO000021454	AR	androgen receptor	4.54	11.68
MO000078913	YY1	YY1 transcription factor	4.14	1.16
MO000134485	NANOG	Nanog homeobox	4.1	1.33
MO000027755	HNF4A	hepatocyte nuclear factor 4 alpha	3.87	1.56
MO000028320	POU6F1	POU class 6 homeobox 1	3.78	1.39
MO000142283	SIRT6	sirtuin 6	3.59	1.2
MO000026229	KLF5	Kruppel like factor 5	3.5	1.3

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)
MO000031189	PKCdelta(h)	PRKCD	protein kinase C delta	1	100	0.86
MO000059577	PKCdelta(h)	PRKCD	protein kinase C delta	1	132	0.86
MO000056654	p300(h)	EP300	E1A binding protein p300	1	134	0.26
MO000022058	Lyn(h)	LYN	LYN proto-oncogene, Src family tyrosine kinase	1	147	0.46
MO000117508	TC-PTP(h)	PTPN2	protein tyrosine phosphatase, non-receptor type 2	1	149	0.33
MO000021902	TFIIH-CAK(h)	CCNH, CDK7, MNAT1	MNAT1, CDK activating kinase assembly factor, cyclin H, cyclin dependent kinase 7	1	152	0.63
MO000009386	MEK1(h)	MAP2K1	mitogen-activated protein kinase kinase 1	1	176	0.22
MO000162702	phlpp2(h)	PHLPP2	PH domain and leucine rich repeat protein phosphatase 2	1	182	0.38
MO000020073	Ubc5A(h)	UBE2D1	ubiquitin conjugating enzyme E2 D1	1	186	0.42
MO000022059	LynB(h)	LYN	LYN proto-oncogene, Src family tyrosine kinase	1	191	0.46

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)
MO000032768	ILK(h)	ILK	integrin linked kinase	1	133	-0.75
MO000033313	PKACA(h)	PRKACA	protein kinase cAMP-activated catalytic subunit alpha	1	218	-1.18
MO000021736	Cdk2(h)	CDK2	cyclin dependent kinase 2	1	224	-0.51
MO000021208	Caspase-6(h)	CASP6	caspase 6	1	233	-0.73
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	236	-1.41
MO000099197	ILK-isoform1(h)	ILK	integrin linked kinase	1	240	-0.75
MO000079043	PML-4(h)	PML	promyelocytic leukemia	1	252	-0.51
MO000032135	proCaspase-6(h)	CASP6	caspase 6	1	306	-0.73
MO000124674	EPHB2(h)	EPHB2	EPH receptor B2	1	335	-0.91
MO000118076	EGF:ErbB1{pY}:ErbB2{pY}:Src	EGF, EGFR, ERBB2, SRC	SRC proto-oncogene, non-receptor tyrosine kinase, epidermal growth factor, epidermal growth factor r...	1	336	-0.93

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.

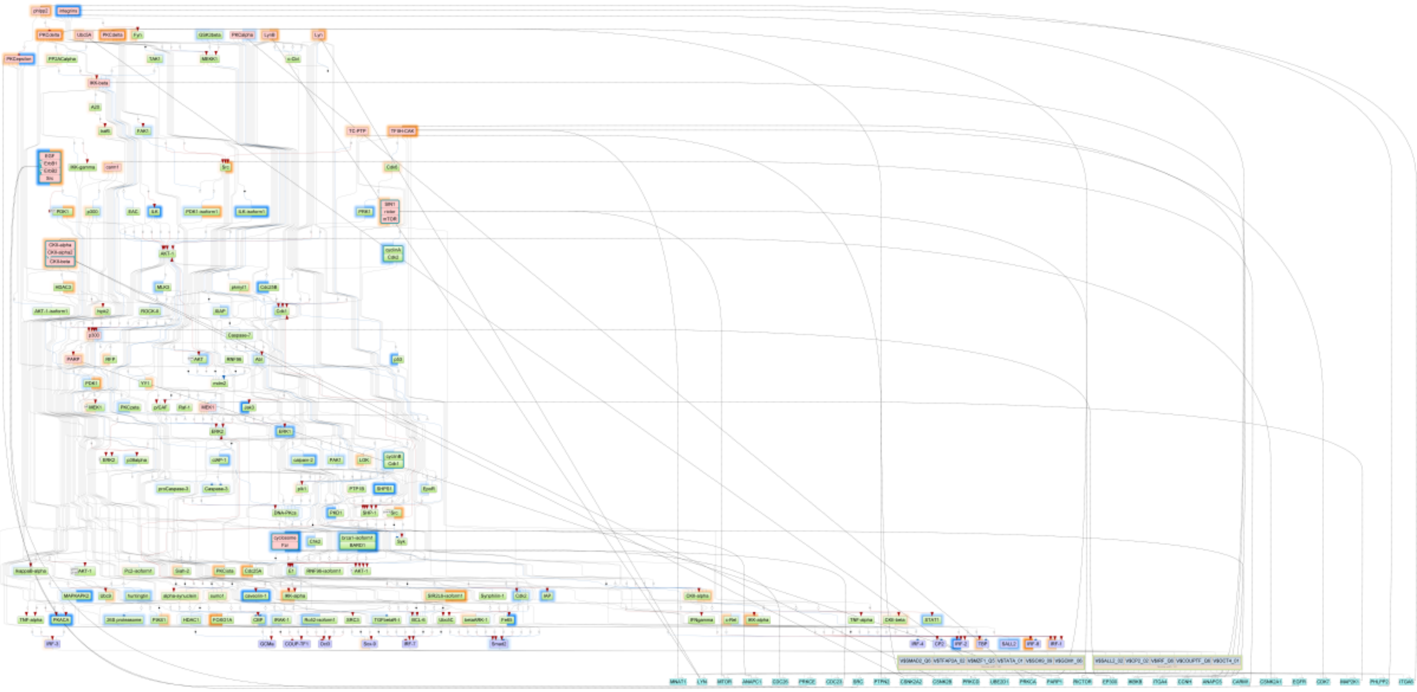


Figure 15. Diagram of intracellular regulatory signal transduction pathways of up-regulated genes in Myc_induce vs. Control. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate molecules, which have been added to the network 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 →

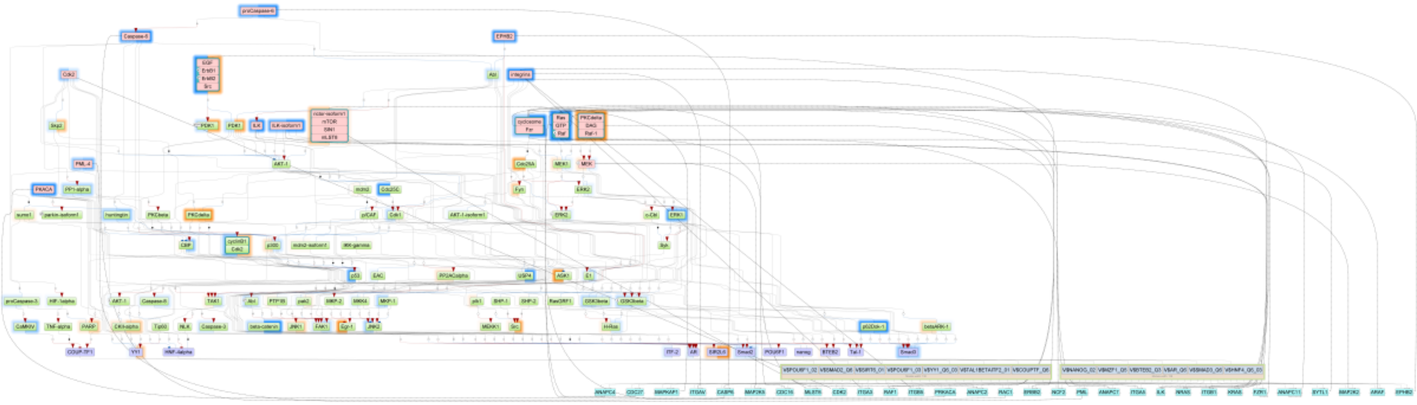


Figure 16. Diagram of intracellular regulatory signal transduction pathways of down-regulated genes in Myc_induce vs. Control. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate molecules, which have been added to the network 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. Identification of potential drugs

In the last step of the analysis we strived to identify known drugs as well as new potentially active chemical compounds that are potentially suitable for inhibition (or activation) of the identified molecular targets in the context of specified human disease.
First, we identify known drugs using information from HumanPSD™ database [5] about their targets and about clinical trials where the drugs have been tested for the treatment of various human diseases. Table 11 shows the resulting list of druggable master regulators that represent the predicted drug targets of the studied pathology. Table 12 lists chemical compounds and known drugs (from the HumanPSD™ database) potentially acting on corresponding master regulators.

Table 11. Known drug targets for known drugs revealed in this study.The column ***Druggability score*** *contains the number of drugs that are potentially suitable for inhibition (or activation) of the target.* ***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	Gene symbol	Gene description	Druggability score	Contained in proteome set	Total rank	logFC (transcriptome)
ENSG00000134058	CDK7	cyclin dependent kinase 7	2	1	152	0.63
ENSG00000254087	LYN	LYN proto-oncogene, Src family tyrosine kinase	4	1	192	0.46
ENSG00000104365	IKBKB	inhibitor of nuclear factor kappa B kinase subunit beta	7	1	211	0.23
ENSG00000104695	PPP2CB	protein phosphatase 2 catalytic subunit beta	1	1	231	0.21
ENSG00000163932	PRKCD	protein kinase C delta	2	1	231	0.86
ENSG00000065613	SLK	STE20 like kinase	2	1	243	0.4
ENSG00000115232	ITGA4	integrin subunit alpha 4	8	1	247	0.51
ENSG00000169032	MAP2K1	mitogen-activated protein kinase kinase 1	10	1	268	0.22
ENSG00000153208	MERTK	MER proto-oncogene, tyrosine kinase	1	0	278	0.95
ENSG00000070770	CSNK2A2	casein kinase 2 alpha 2	1	1	298	0.47

*Table 12. The list of drugs (from Human PSD) approved or used in clinical trials for the application in neoplasm metastasis and osteosarcoma and acting on master regulators revealed in our study. The column **Target activity score** contains the value of numeric function that depends on ranks of all targets that were found for the drug. The column **Disease activity score** contains the weighted sum of user selected diseases where the drug is known to be applied. We use sum of clinical trials phases as the weight of the disease. **Drug rank** column contains total rank of given drug among all found. See Methods section for details.*

See full table →

ID	Name	Target names	Target activity score	NA	Phase 1	Phase 2	Phase 3	Phase 4	D a s
DB06616	Bosutinib	SRC, MAP2K1, LYN	0.96	Breast Neoplasms, Leukemia, Leukemia, Lymphoid, Leukemia, Myelogenous, Chronic, BCR-ABL Positive, Leukemia, Myeloid, Neoplasms, Precursor Cell Lymphoblastic Leukemia-Lymphoma	Acute Kidney Injury, Breast Neoplasms, Carcinoma, Non-Small-Cell Lung, Cholangiocarcinoma, Cognitive Dysfunction, Colorectal Neoplasms, Dementia...	Neoplasm Metastasis, Brain Abscess, Breast Neoplasms, Cholangiocarcinoma, Colorectal Neoplasms, Cysts, Glioblastoma...	Leukemia, Leukemia, Myelogenous, Chronic, BCR-ABL Positive, Leukemia, Myeloid	Leukemia, Myeloid	2
DB01269	Panitumumab	EGFR	0.39	Adenocarcinoma, Carcinoma, Carcinoma, Squamous Cell, Carcinoma, Transitional Cell, Colorectal Neoplasms, Histology, Neoplasms...	Adenocarcinoma, Carcinoma, Non-Small-Cell Lung, Carcinoma, Small Cell, Carcinoma, Squamous Cell, Colonic Neoplasms, Colorectal Neoplasms, Dermoid Cyst...	Neoplasm Metastasis, Adenocarcinoma, Adenoma, Adenoma, Pleomorphic, Biliary Tract Neoplasms, Breast Neoplasms, Carcinoid Tumor...	Colorectal Neoplasms, Neoplasms, Neoplasms, Squamous Cell, Noma, Rectal Neoplasms, Stomach Neoplasms	Neoplasms	2
DB00317	Gefitinib	EGFR	0.39	Adenocarcinoma, Bronchopulmonary Dysplasia, Carcinoma, Carcinoma, Adenoid Cystic, Carcinoma, Basal Cell, Carcinoma, Medullary, Carcinoma, Mucoepidermoid...	Osteosarcoma, Astrocytoma, Breast Neoplasms, Carcinoma, Non-Small-Cell Lung, Carcinoma, Renal Cell, Carcinoma, Small Cell, Carcinoma, Squamous Cell...	Adenocarcinoma, Adenocarcinoma, Mucinous, Adenoma, Astrocytoma, Brain Neoplasms, Breast Neoplasms, Breast Neoplasms, Male...	Adenocarcinoma, Breast Neoplasms, Carcinoma, Carcinoma, Non-Small-Cell Lung, Carcinoma, Small Cell, Carcinoma, Squamous Cell, Colorectal Neoplasms...	Carcinoma, Non-Small-Cell Lung, Carcinoma, Small Cell, Lung Neoplasms, Neoplasms, Noma, Small Cell Lung Carcinoma	1
DB08881	Vemurafenib	BRAF	0.32	Melanoma, Neoplasms, Noma, Thyroid Neoplasms	Colorectal Neoplasms, Glioma, Lymphoma, Melanoma, Neoplasms, Noma, Rectal Neoplasms...	Osteosarcoma, Anger, Brain Abscess, Carcinoma, Non-Small-Cell Lung, Carcinoma, Small Cell, Colorectal Neoplasms, Craniopharyngioma...	Melanoma, Noma	Melanoma, Neoplasms, Noma	2
DB06151	Acetylcysteine	IKBKB, CHUK	0.35	Acute Kidney Injury, Acute Lung Injury, Albuminuria, Aneurysm, Apnea, Bipolar Disorder, Bites and Stings...	Osteosarcoma, Albinism, Albinism, Oculocutaneous, Alcoholism, Altitude Sickness, Amphetamine-Related Disorders, Anemia...	Acute Kidney Injury, Acute Lung Injury, Adrenoleukodystrophy, Albinism, Albinism, Oculocutaneous, Alcohol Drinking, Alcoholism...	Acute Kidney Injury, Anemia, Sickle Cell, Atrial Fibrillation, Atrophy, Bacteremia, Brain Death, Brain Diseases...	Acute Kidney Injury, Alcoholism, Anemia, Atherosclerosis, Atrophy, Bipolar Disorder, Bronchiectasis...	1

Table 13. The list of drugs (from HumanPSD) known to be acting on master regulators revealed in our study that can be proposed as a drug repurposing initiative for the treatment of neoplasm metastasis and osteosarcoma. ***Target activity score*** *column contains value of numeric function that depends on ranks of all targets that were found for the drug.* ***Drug rank*** *column contains total rank of given drug among all found. See Methods section for details.*

ID	Name	Target names	Target activity score	NA	Phase 1	Phase 2	Phase 3	Phase 4	Drug rank
DB05013	Ingenol Mebutate	PRKCD, PRKCA	0.87	Keratosis, Keratosis, Actinic	Keratosis, Keratosis, Actinic, Warts	Carcinoma, Basal Cell, Keratosis, Keratosis, Actinic, Keratosis, Seborrheic, Noma, Sunburn	Keratosis, Keratosis, Actinic	Keratosis, Keratosis, Actinic	24
DB00163	Vitamin E	PPP2CB, PRKCA, PPP2CA	0.61	Apnea, Fatty Liver, Fatty Liver, Alcoholic, Fragile X Syndrome, HIV Infections, Hepatitis, Infection...	Alzheimer Disease, Anemia, Anemia, Iron-Deficiency, Arthritis, Arthritis, Rheumatoid, Asthma, Brain Abscess...	Acute Kidney Injury, Adrenoleukodystrophy, Anemia, Arsenic Poisoning, Arteriosclerosis, Atherosclerosis, Brain Abscess...	Alzheimer Disease, Angina, Unstable, Arterial Occlusive Diseases, Arteriosclerosis, Burns, Cardiovascular Diseases, Cataract...	Angina Pectoris, Variant, Asphyxia, Cicatrix, Cicatrix, Hypertrophic, Diabetes Mellitus, Dyslipidemias, Epilepsy...	28
DB09033	Vedolizumab	ITGA4	0.48	Cholangitis, Cholangitis, Sclerosing, Colitis, Colitis, Ulcerative, Crohn Disease, Graft vs Host Disease, Inflammatory Bowel Diseases...	Colitis, Colitis, Ulcerative, Crohn Disease, Inflammatory Bowel Diseases, Melanoma, Noma, Ulcer	Celiac Disease, Colitis, Colitis, Ulcerative, Crohn Disease, Ulcer	Cholangitis, Cholangitis, Sclerosing, Colitis, Colitis, Ulcerative, Crohn Disease, Inflammatory Bowel Diseases, Ulcer	Colitis, Colitis, Ulcerative, Crohn Disease, Ulcer	32
DB08911	Trametinib	MAP2K1	0.44	Adenocarcinoma, Biliary Tract Neoplasms, Brain Abscess, Carcinoma, Non-Small-Cell Lung, Carcinoma, Small Cell, Gastrointestinal Stromal Tumors, Germinoma...	Adenocarcinoma, Adenocarcinoma, Clear Cell, Behavior, Brain Abscess, Breast Neoplasms, Carcinoma, Carcinoma, Non-Small-Cell Lung...	Adenocarcinoma, Astrocytoma, Bile Duct Neoplasms, Brain Abscess, Breast Neoplasms, Carcinoma, Hepatocellular, Carcinoma, Non-Small-Cell Lung...	Adenocarcinoma, Melanoma, Noma	Carcinoma, Small Cell, Glioma, Lung Neoplasms, Melanoma, Neoplasms, Noma, Small Cell Lung Carcinoma	39
DB00244	Mesalazine	IKBKB, CHUK	0.44	Abdominal Pain, Character, Colitis, Colitis, Collagenous, Colitis, Ulcerative, Crohn Disease, Diarrhea...		Adenoma, Colitis, Colitis, Ulcerative, Colorectal Neoplasms, Crohn Disease, Digestive System Diseases, Gastrointestinal Diseases...	Colitis, Colitis, Collagenous, Colitis, Ulcerative, Crohn Disease, Digestive System Diseases, Diverticulitis, Gastrointestinal Diseases...	Colitis, Colitis, Ulcerative, Diarrhea, Diverticulum, Irritable Bowel Syndrome, Ulcer	47

Next, new potential small molecular ligands were predicted for the revealed targets and a general druggability check was run using a pre-computed database of spectra of biological activities of chemical compounds from a library of 13040 most pharmaceutically active known compounds. The spectra of biological activities has been computed using the program PASS [11-13] on the basis of a (Q)SAR approach. Table 14 shows the resulting list of druggable master regulators, which represent the predicted drug targets of the studied pathology. Table 15 lists chemical compounds and known drugs potentially acting on the corresponding master regulators.

Table 14. Extended list of drug targets revealed in this study (targets that are predicted by PASS program potentially targeted by an extended list of known drugs and pharmaceutically active chemical compounds). The column ***Druggability score*** *contains a numeric value which indicates how suitable this target is to be inhibited (or activated) by a drug. See Methods section for details.*

See full table →

ID	Name	Gene symbol	Gene description	Druggability score	Contained in proteome set	Total rank	logFC (transcriptome)
ENSG00000020426	MNAT1	MNAT1	MNAT1, CDK activating kinase assembly factor	17.43	1	152	0.63
ENSG00000134058	CDK7	CDK7	cyclin dependent kinase 7	11.42	1	152	0.63
ENSG00000134480	CCNH	CCNH	cyclin H	9.81	1	152	0.63
ENSG00000104312	RIPK2	RIPK2	receptor interacting serine/threonine kinase 2	1.41	1	192	0.42
ENSG00000254087	LYN	LYN	LYN proto-oncogene, Src family tyrosine kinase	7.71	1	192	0.46
ENSG00000171132	PRKCE	PRKCE	protein kinase C epsilon	62.09	1	194	0.5
ENSG00000104365	IKBKB	IKBKB	inhibitor of nuclear factor kappa B kinase subunit beta	1.6	1	211	0.23
ENSG00000145335	SNCA	SNCA	synuclein alpha	1.74	1	222	8.69E-2
ENSG00000163932	PRKCD	PRKCD	protein kinase C delta	62.73	1	231	0.86
ENSG00000100393	EP300	EP300	E1A binding protein p300	15.7	1	240	0.26

Table 15. The chemical compounds and known drugs identified by the PASS program as potentially active for the treatment of neoplasm metastasis and osteosarcoma and acting on master regulators revealed in our study. ***Disease activity score*** *column contains maximal value of probability to be active for all activities corresponding to the selected diseases for the given compound.* ***Target activity score*** *column contains value of numeric function which depends on all activity-mechanisms correspondent to the drug.* ***Drug rank*** *column contains total rank of given drug among all found. See Methods section for details.*

See full table →

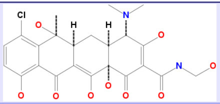
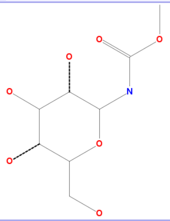
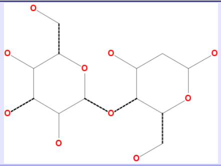
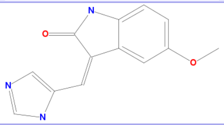
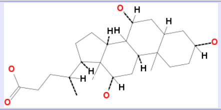
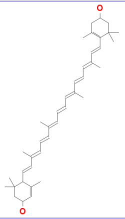
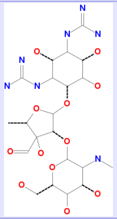
Name	Structure	Target names	Target activity score	Disease activity score	Drug rank
Clomocycline		CSNK1G1, SIRT1	0.13	0.9	201
1-Deoxy-1-Methoxycarbamido-Beta-D-Glucopyranose		CARM1, ITGA4, ITGA6, IGF1R	3.67E-3	0.83	504

Table 16. The chemical compounds and known drugs identified by the PASS program as potentially acting on master regulators revealed in our study. Based on the revealed mechanism of action these compounds can be proposed for the treatment of neoplasm metastasis and osteosarcoma in the current pathological case. ***Disease activity score*** *column contains maximal value of probability to be active for all activities corresponding to the selected diseases for the given compound or 0 if no diseases were selected (in this case column will be hidden).* ***Target activity score*** *column contains value of numeric function which depends on all activity-mechanisms correspondent to the drug.* ***Drug rank*** *column contains total rank of given drug among all found. See Methods section for details.*

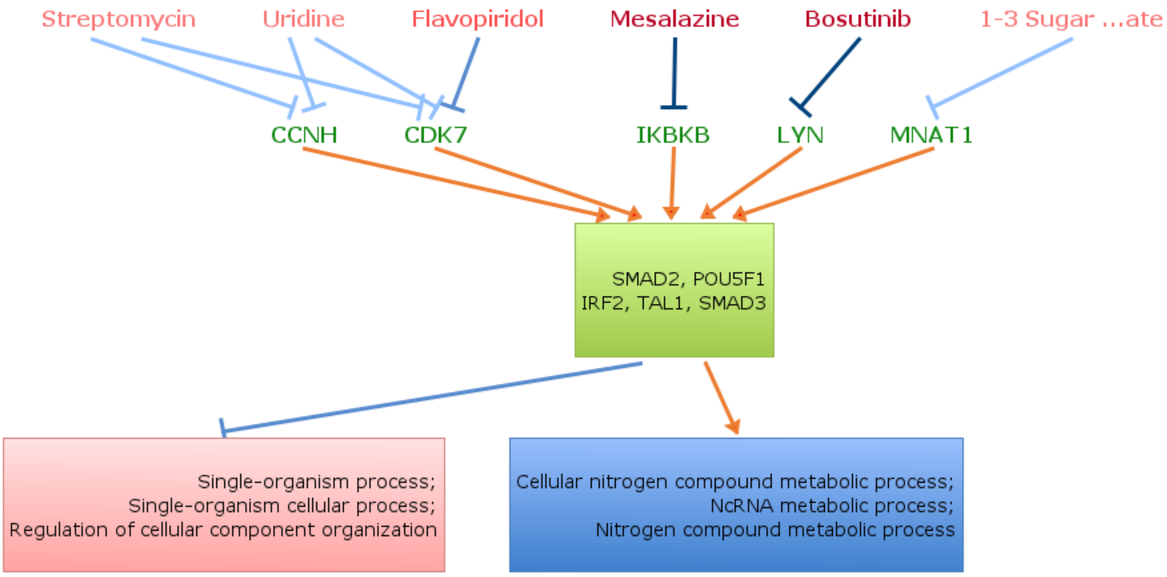
Name	Structure	Target names	Target activity score	Disease activity score	Drug rank
2'-Deoxymaltose		PRKCE, PRKCD, PRKCI, PRKCA	0.38	0.74	55
SU9516		MTOR, PRKCE, PRKCD, PRKCI, PRKCA	0.92	0.5	69
Cholic Acid		EGFR, GRK2, PRKAA2, MERTK, ROCK1, PKMYT1, IGF1R	0.38	0.62	71
Xanthophyll		PRKCE, PRKCD, PRKCI, PRKCA	0.46	0.55	80
Streptomycin		CDK6, CCNH, CCND3, PRKCE, PRKCD, PRKCI, PRKCA...	0.75	0.49	80

As a result of the drug search we came up with two lists of chemical compounds potentially applicable to the targets of our interest. The first list is based on drugs that are known as ligands for the revealed targets in the context of the diseases in our focus as well as in other disease conditions. The second list of identified compounds is based on the prediction of their potential biological activities, which was done using the program PASS. Such computational predictions should be taken as mere suggestions and should be used with care in further experiments.

5. 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 schema of how the selected drugs may interfere with the identified target molecules and pathogenic processes discovered by the study reported here.



6. 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 2020.1 (geneXplain GmbH, Wolfenbüttel, Germany) (http://genexplain.com/transfac).
The master regulator search uses the TRANSPATH® database (BIOBASE), release 2020.1 (geneXplain GmbH, Wolfenbüttel, Germany) (http://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 2020.1 (http://genexplain.com/humanpsd).
The Ensembl database release Human88.38 (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 three other ranks:

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

To calculate clinical trials phase for the given compound we select the maximum phase of all diseases that are known to have clinical trials with this compound. "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*$_{PSD}$):

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

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.

7. 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. *Chemoinformatics 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.