

## **PRKCD and DUSP5 are promising druggable targets for treating neoplasm metastasis and osteosarcoma that control activity of SMAD2, SMAD1 and STAT3 transcription factors on promoters of differentially expressed genes**

Demo User

geneXplain GmbH

info@genexplain.com

Data received on 07/09/2019 ; Run on 09/09/2019 ; Report generated on 09/09/2019

---



### **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, SMAD1, POU5F1, STAT3 and SMAD3. The subsequent network analysis suggested PRKCD, PDK1, DUSP5, DUSP2 and BMP2 as the most promising and druggable molecular targets. Finally, the following drugs were identified as the most promising treatment candidates: 13-Acetylphorbol, 4-[(3-CHLORO-4-

## 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 [1-5] applied here has been devised 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 [11]. 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 biological 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 [12-14].

## 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

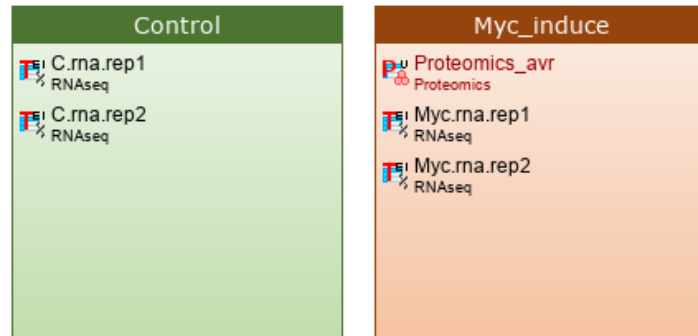


Figure 1. Annotation diagram of experimental data used in this study. With the colored boxes we show those sub-categories of the data that are compared in our analysis.

### 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 501 upregulated ([Supplementary table 1](#)) and 645 downregulated ([Supplementary table 2](#)) genes (p-value<0.1, LogFC>0.75 for up-regulated and LogFC<-0.75 for down-regulated). For further upstream analysis we ranked all genes according to their LogFC and chose the 300 most up-regulated genes and the 300 most downregulated genes (see Table 2 and Table 3 for the top ten up- and down-regulated genes, resp).

Table 2. Top ten 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 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 target genes**

A functional analysis of differentially expressed genes was done by mapping the up- and 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.

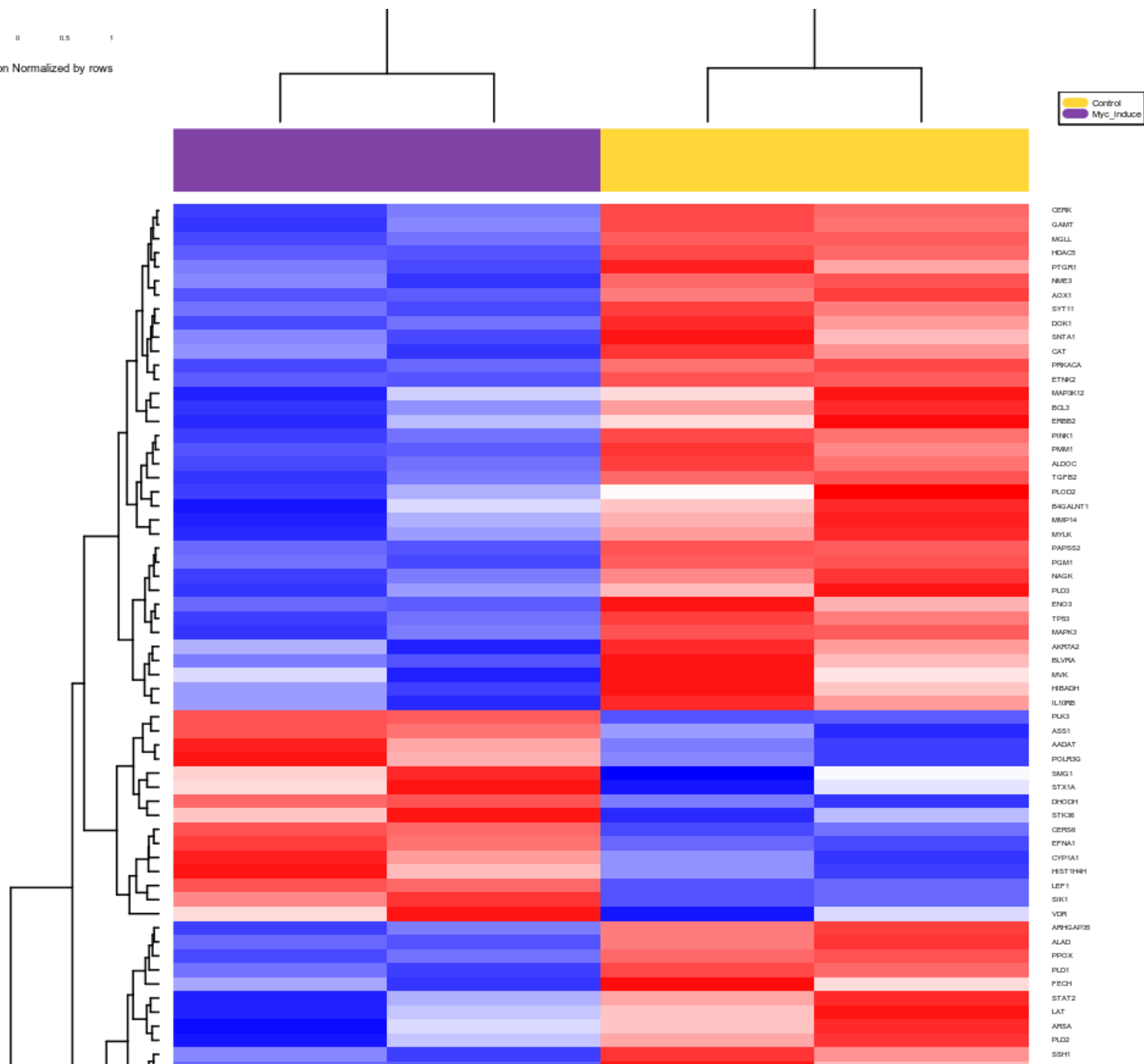
### **Heatmap of differentially expressed genes in *Myc\_induce* vs. *Control***

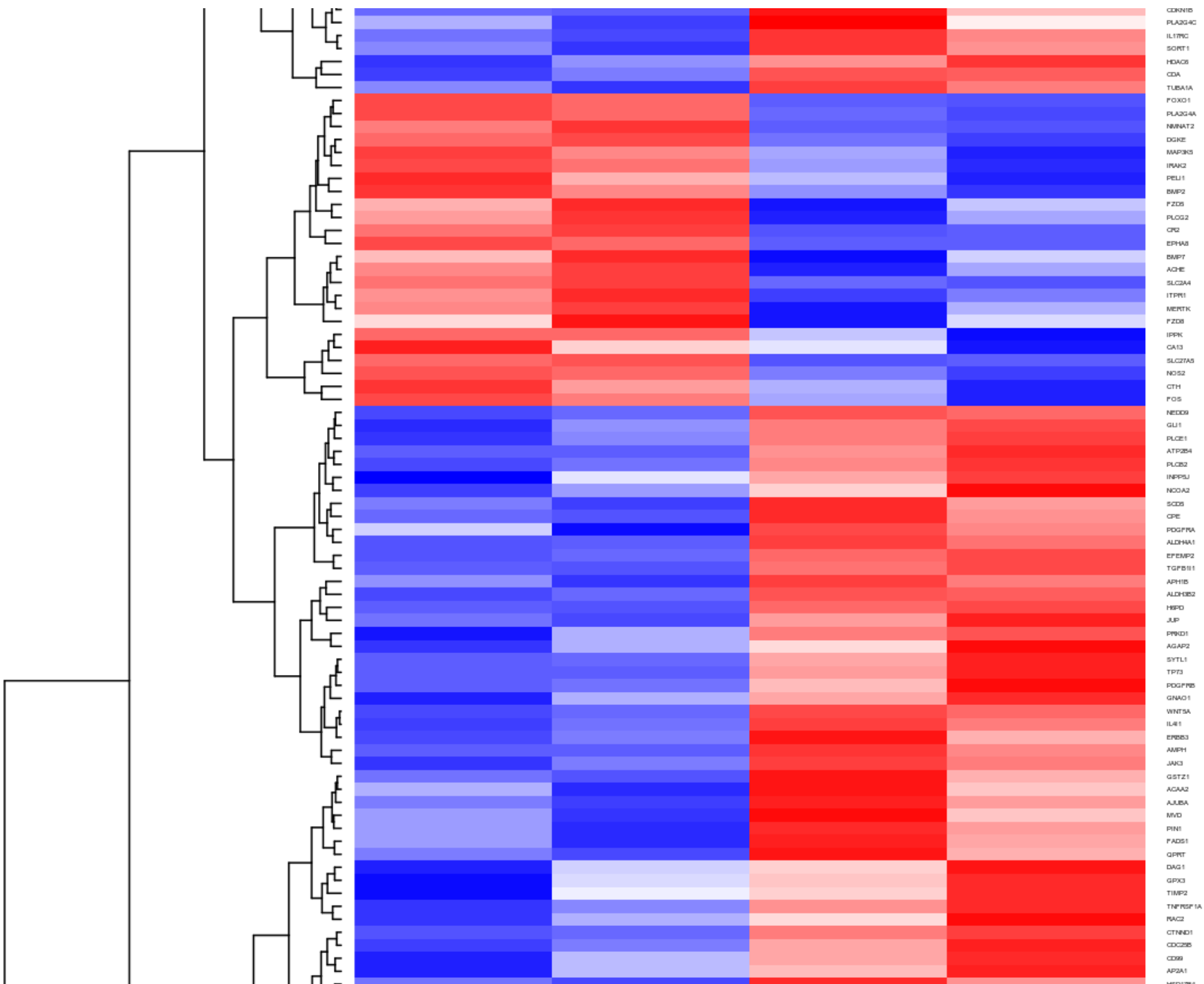
A heatmap of all differentially expressed genes playing a potential regulatory role in the system (enriched in TRANSPATH® pathways) is presented in Figure 2.



-1   -0.5   0   0.5   1

Gene Expression Normalized by rows





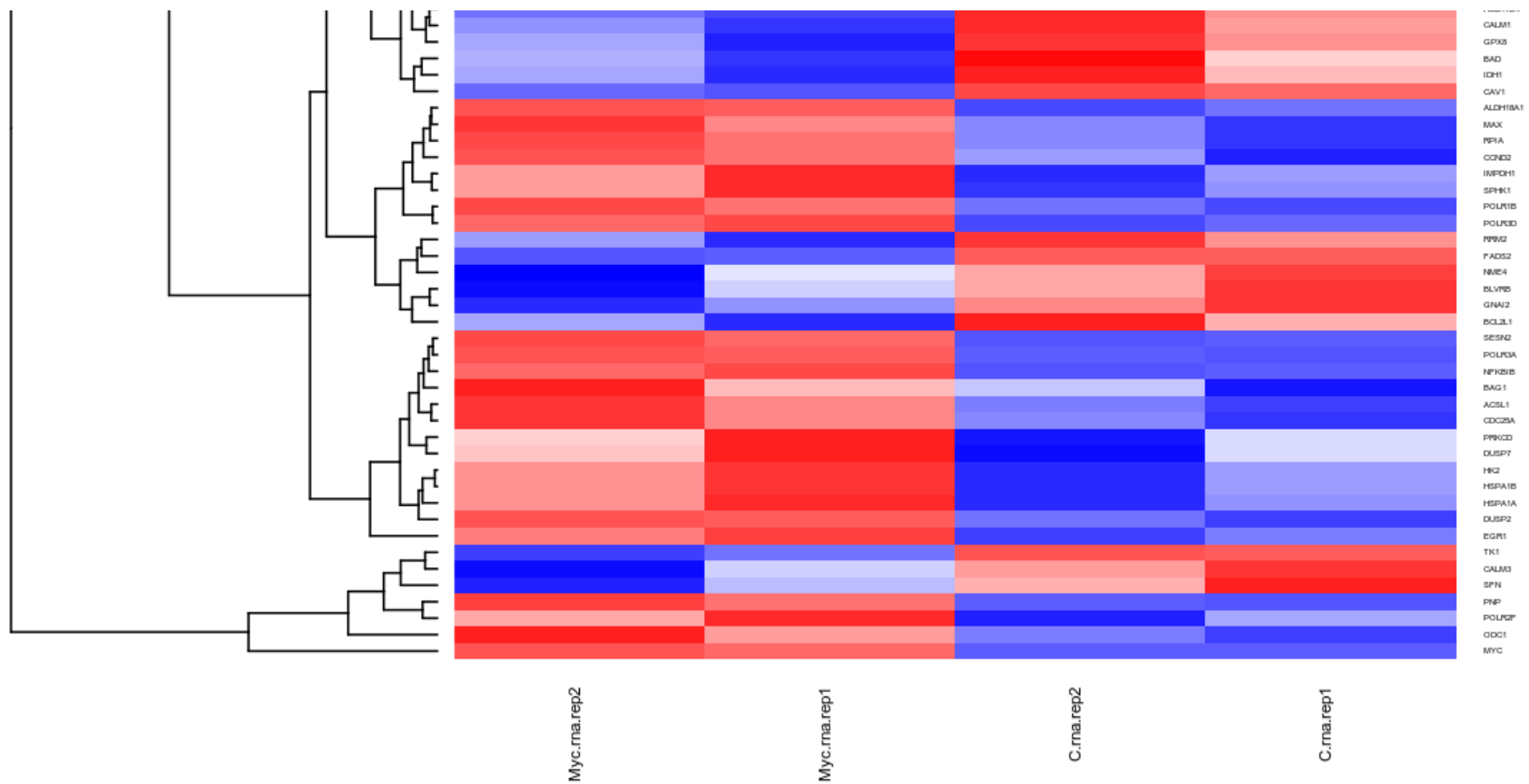


Figure 2. Heatmap of genes enriched in Transpath categories. The colored bar at the top shows the types of the samples according to the legend in the upper right corner.

[See full diagram →](#)

## Up-regulated genes:

GO (biological process)

biological\_process Gene Ontology treemap

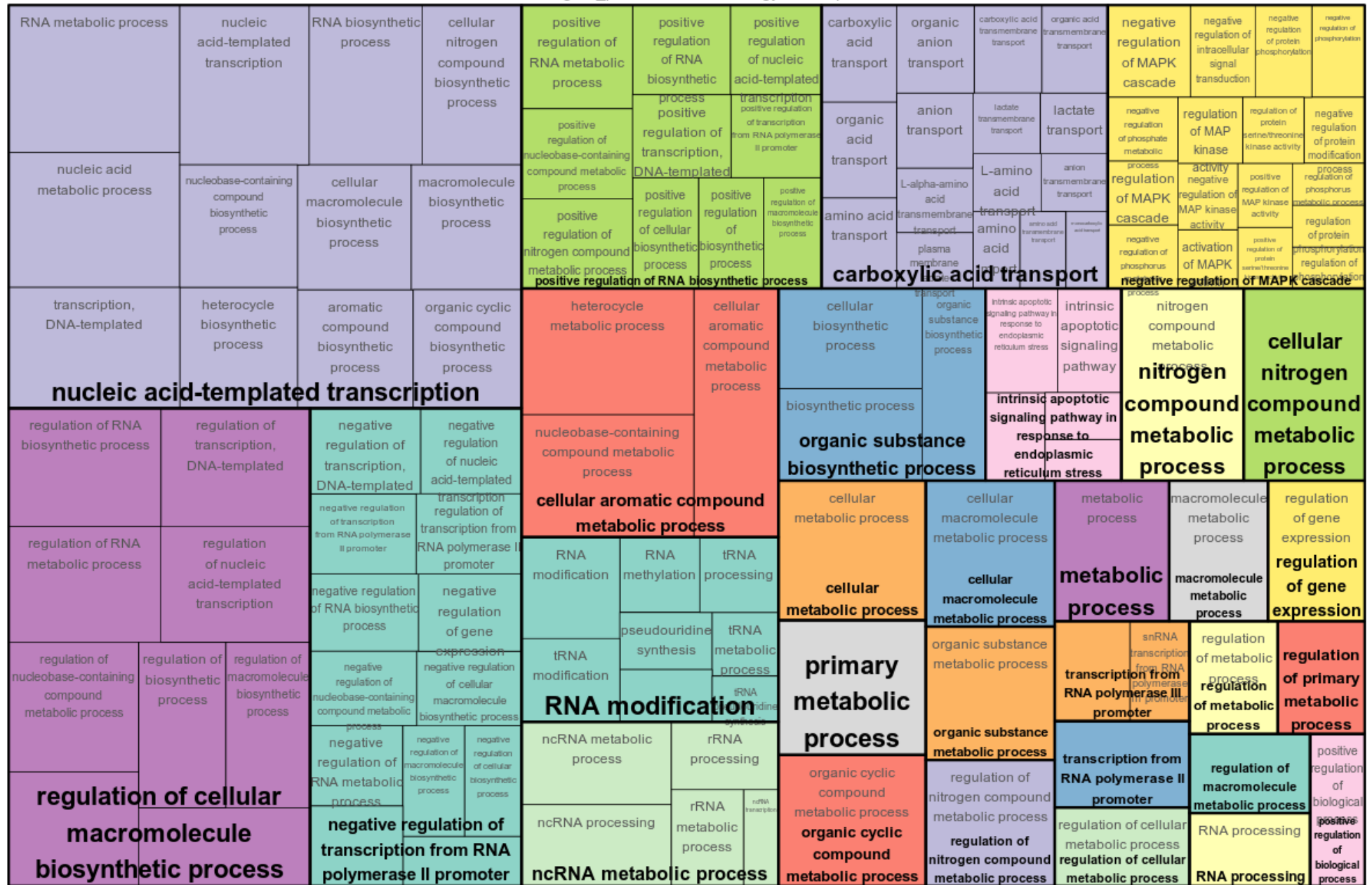


Figure 3. Enriched GO (biological process) of up-regulated genes.

### Full classification →



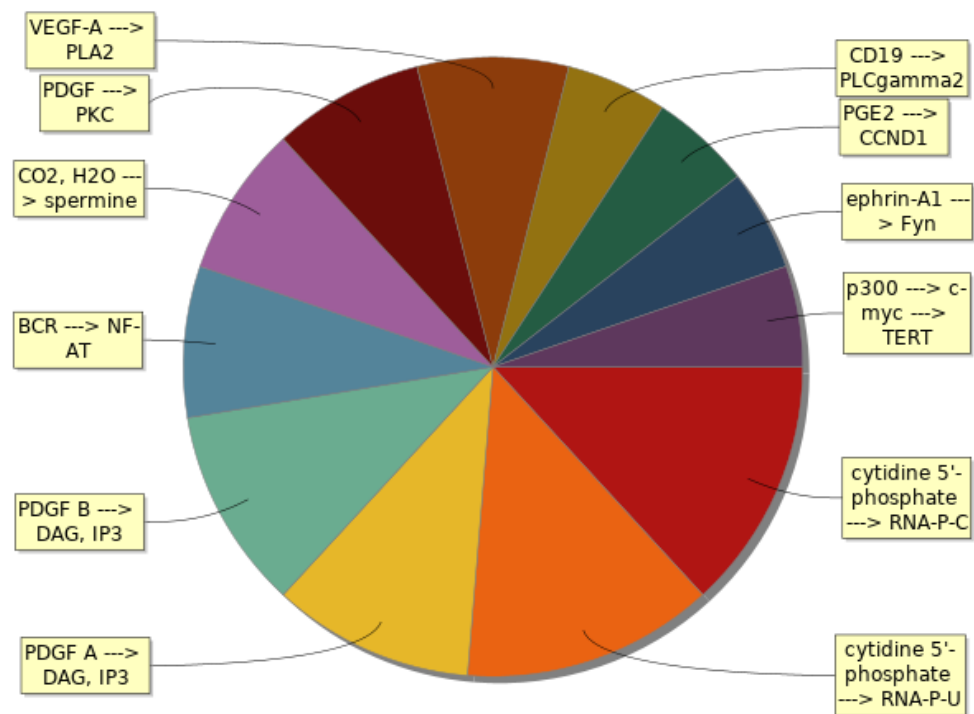


Figure 4. Enriched TRANSPATH® Pathways (2019.2) of up-regulated genes.

[Full classification →](#)

**HumanPSD(TM) disease (2019.2)**

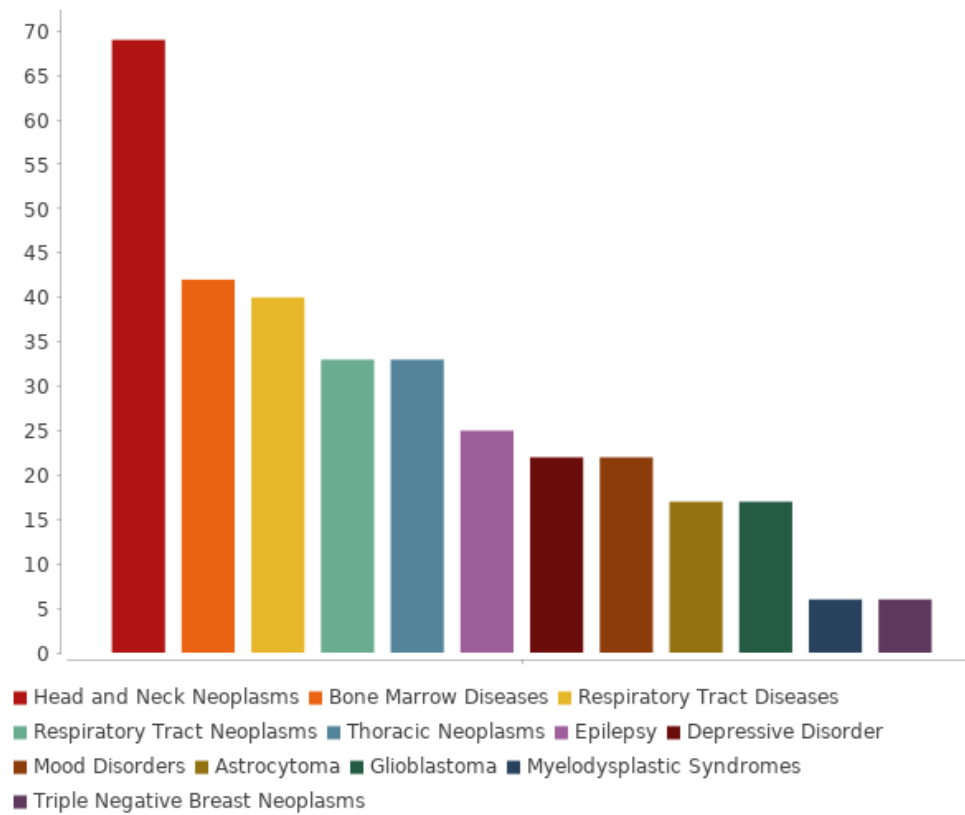


Figure 5. Enriched HumanPSD(TM) disease (2019.2) of up-regulated genes. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

[Full classification →](#)

## Down-regulated genes:

GO (biological process)

biological\_process Gene Ontology treemap

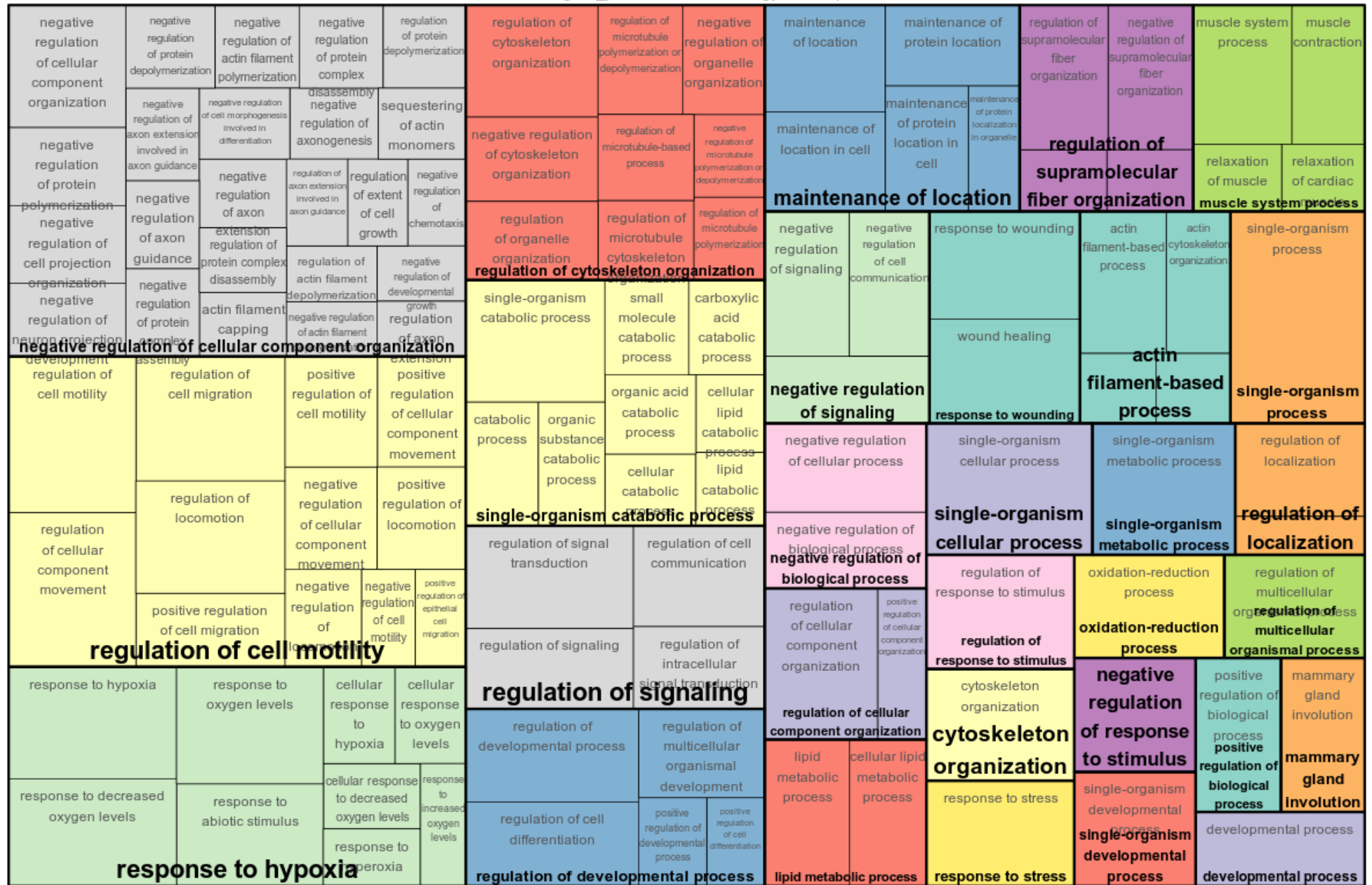


Figure 6. Enriched GO (biological process) of down-regulated genes.

### Full classification →

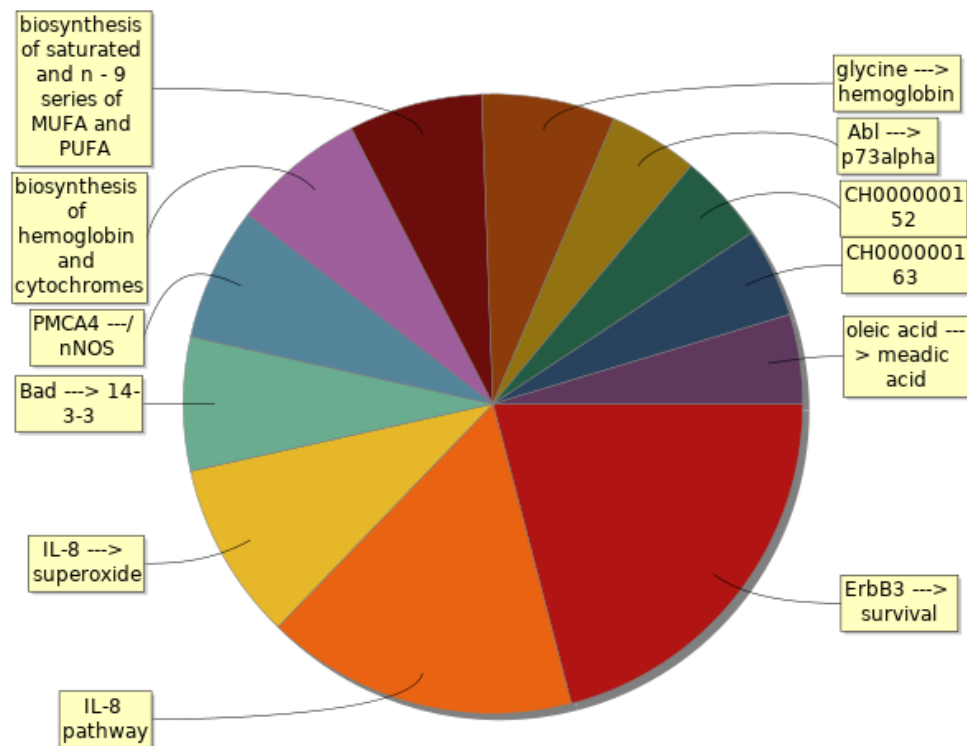


Figure 7. Enriched TRANSPATH® Pathways (2019.2) of down-regulated genes.

[Full classification →](#)

**HumanPSD(TM) disease (2019.2)**

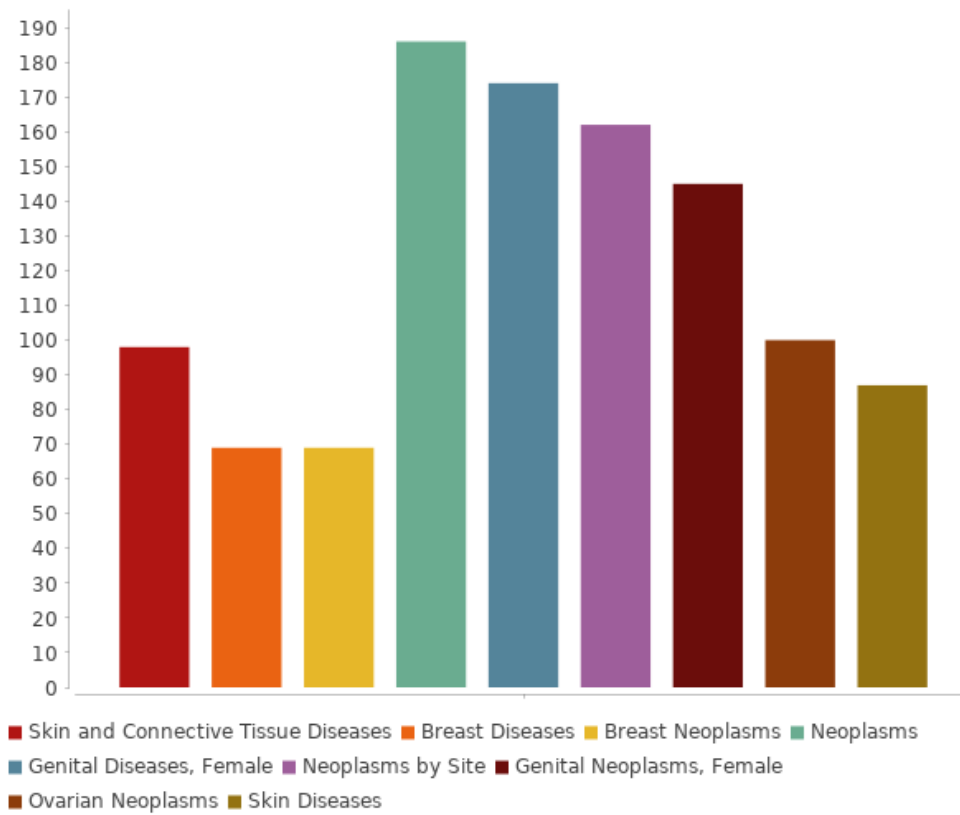


Figure 8. Enriched HumanPSD(TM) disease (2019.2) of down-regulated genes. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

[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
<a href="#">ENSG00000173598</a>	nudix hydrolase 4	NUDT4	4.36
<a href="#">ENSG00000100335</a>	mitochondrial elongation factor 1	MIEF1	3.8
<a href="#">ENSG00000115884</a>	syndecan 1	SDC1	3.62
<a href="#">ENSG00000102910</a>	lon peptidase 2, peroxisomal	LONP2	3.3
<a href="#">ENSG00000179046</a>	tripartite motif family like 2	TRIML2	2.87
<a href="#">ENSG00000114648</a>	kelch like family member 18	KLHL18	2.76
<a href="#">ENSG00000170525</a>	6-phosphofructo-2-kinase/fructose-2,6-biphosphatase 3	PFKFB3	2.69
<a href="#">ENSG00000120949</a>	TNF receptor superfamily member 8	TNFRSF8	2.46
<a href="#">ENSG00000188158</a>	NHS actin remodeling regulator	NHS	2.46
<a href="#">ENSG00000119599</a>	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:**

##### **GO (biological process)**

biological\_process Gene Ontology treemap



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

### Full classification →

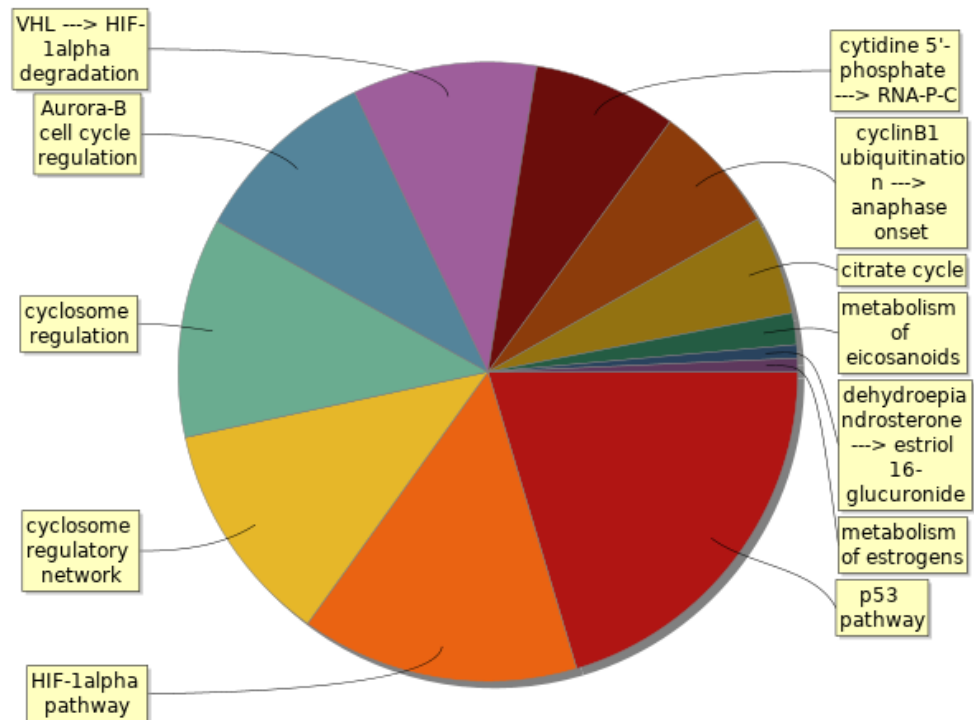


Figure 10. Enriched TRANSPATH® Pathways (2019.2) of the list of proteins provided as input.

[Full classification →](#)

**HumanPSD(TM) disease (2019.2)**



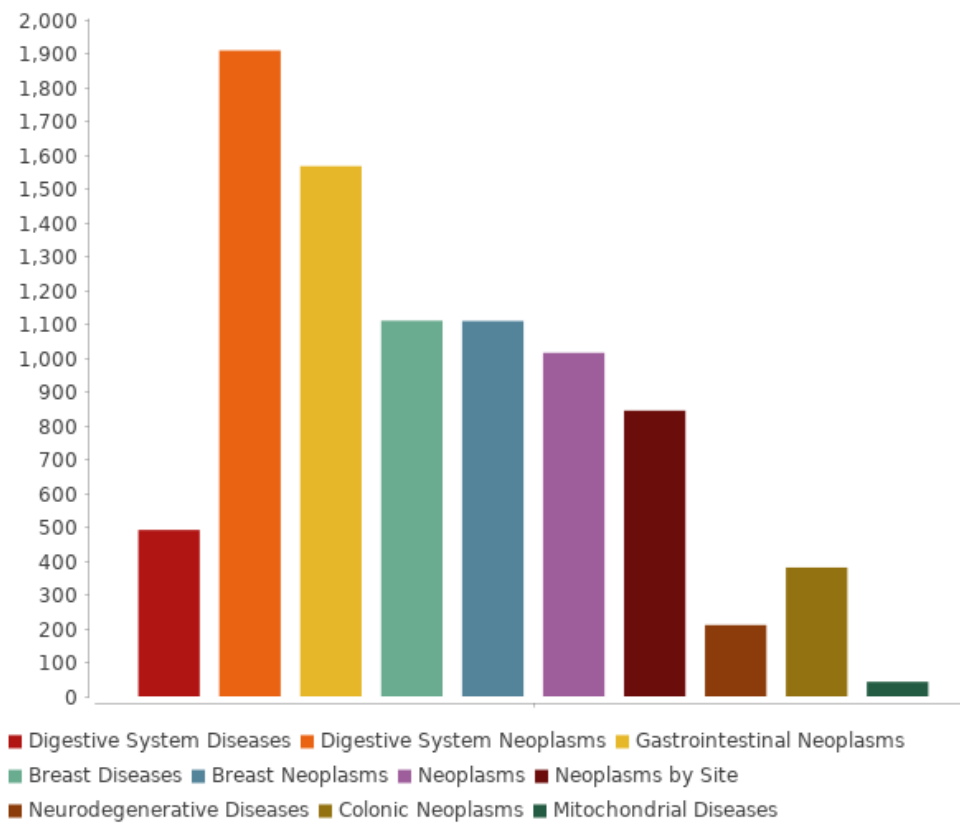


Figure 11. Enriched HumanPSD(TM) disease (2019.2) of the list of proteins provided as input. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

[Full classification →](#)

### 3.5. Comparison plot of transcriptome and proteome

After the analysis of transcriptome and proteome data they were compared with each other.

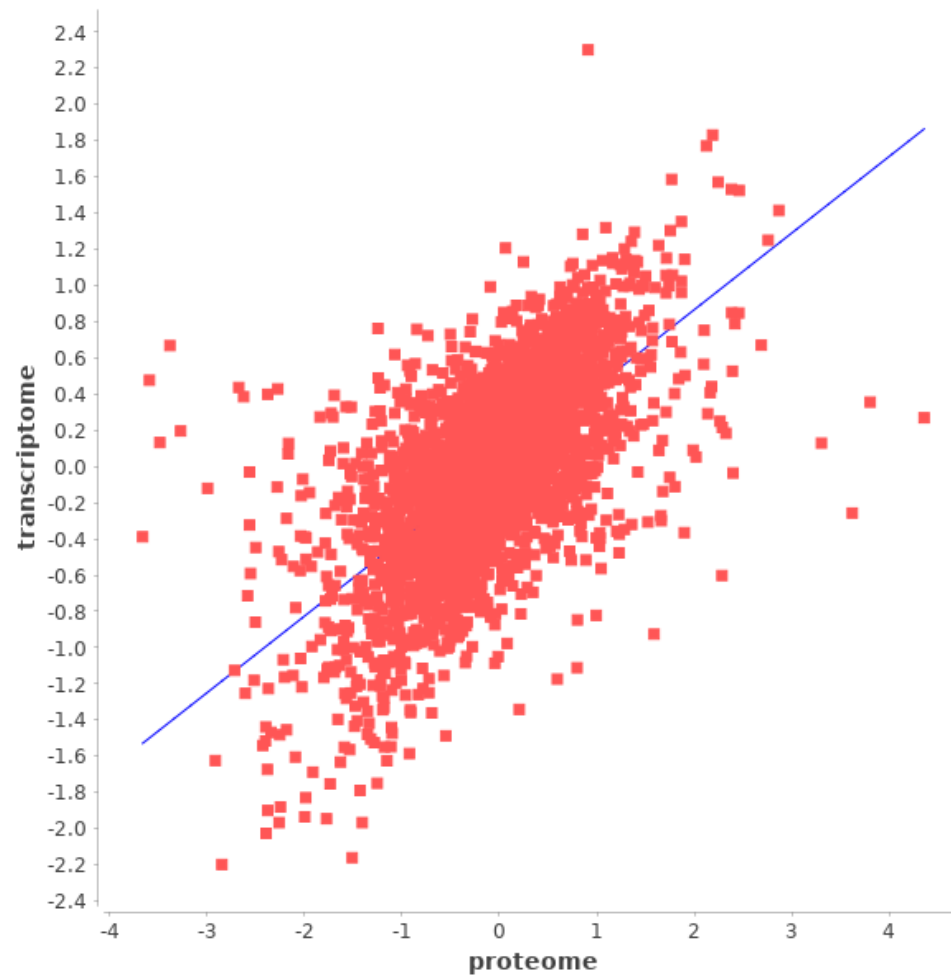
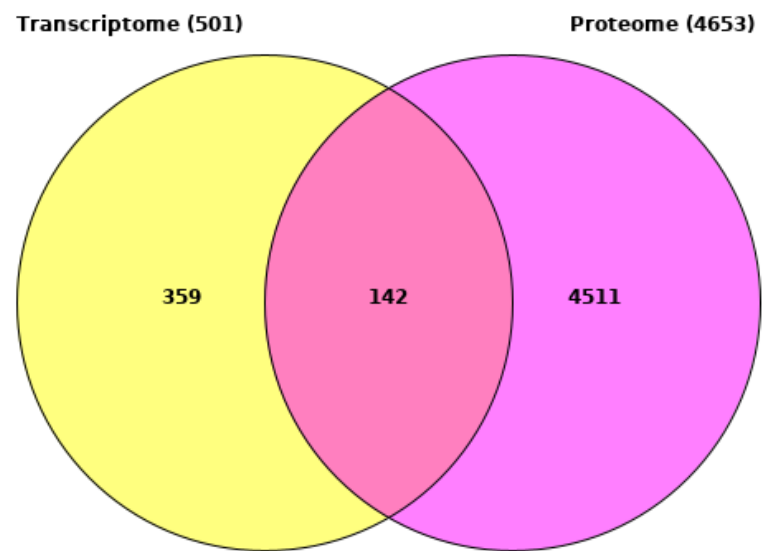


Figure 12. Comparison plot of comparison proteome vs transcriptome. X axis: protein expression value - Proteomics\_avr. Y axis: LogFC of differential gene expression.

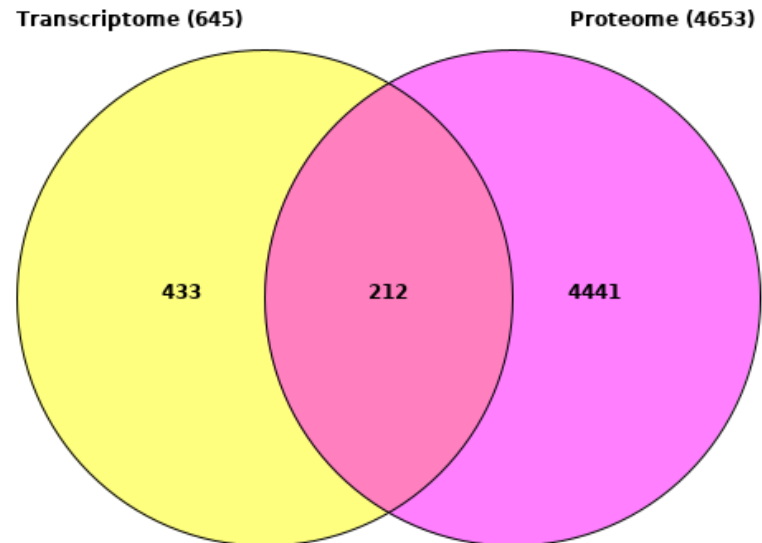
[Full comparison →](#)

**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).**

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\$NFKAPPAB50_01 0.00; N=2	V\$POU5F1_04 0.00; N=2	V\$COUPTF_Q6 0.00; N=3	V\$SMAD1_Q6 0.00; N=2	V\$SMAD2_Q6 0.00; N=1
-------------------------------	---------------------------	---------------------------	--------------------------	--------------------------

Module width: 96

##### Module 2:

V\$TFAP2A_02 0.89; N=2	V\$TATA_01 0.00; N=2	V\$SMAD2_Q6 0.00; N=1	V\$GR_Q6 0.87; N=3	V\$CP2_Q4 0.98; N=2	V\$HIF2A_01 0.00; N=1
---------------------------	-------------------------	--------------------------	-----------------------	------------------------	--------------------------

Module width: 110

**Model score (-p\*log10(pval)):** 13.87

**Wilcoxon p-value (pval):** 3.38e-29

**Penalty (p):** 0.487

**Average yes-set score:** 8.61

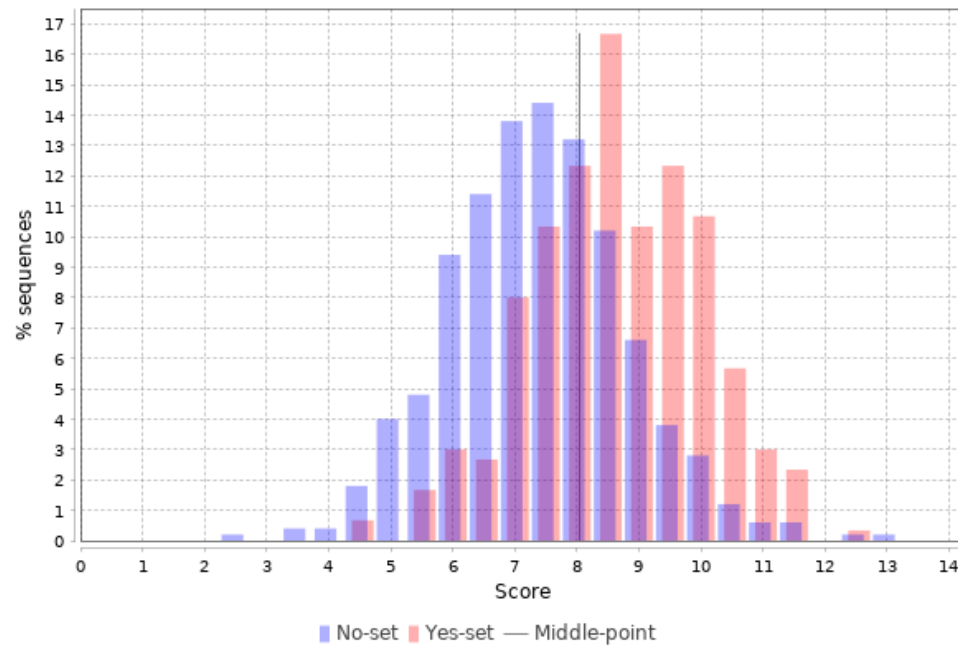
**Average no-set score:** 7.39

**AUC:** 0.74

**Middle-point:** 8.04

**False-positive:** 30.80%

**False-negative:** 32.67%



[See model visualization table](#) →

**Enhancer model potentially involved in regulation of target genes (down-regulated genes in Myc\_induce vs. Control).**

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\$PU1\_Q4  
0.87; N=2

V\$GR\_Q6  
0.88; N=3

V\$STAT3\_Q8  
0.00; N=3

V\$PAX2\_Q2  
0.00; N=3

V\$SMAD4\_Q6  
0.82; N=1

V\$PPARGRXRA\_Q2  
0.78; N=2

Module width: 117

#### Module 2:

V\$TAL1BETAITF2\_Q1  
0.00; N=3

V\$MZF1\_Q5\_Q1  
0.00; N=3

V\$BTEB2\_Q3\_Q1  
0.96; N=3

V\$SMAD2\_Q6  
0.00; N=3

V\$SMAD3\_Q6  
0.00; N=2

V\$SMAD1\_Q1  
0.89; N=3

Module width: 115

**Model score (-p\*log10(pval)): 20.04**

**Wilcoxon p-value (pval): 5.84e-43**

**Penalty (p): 0.475**

**Average yes-set score: 12.32**

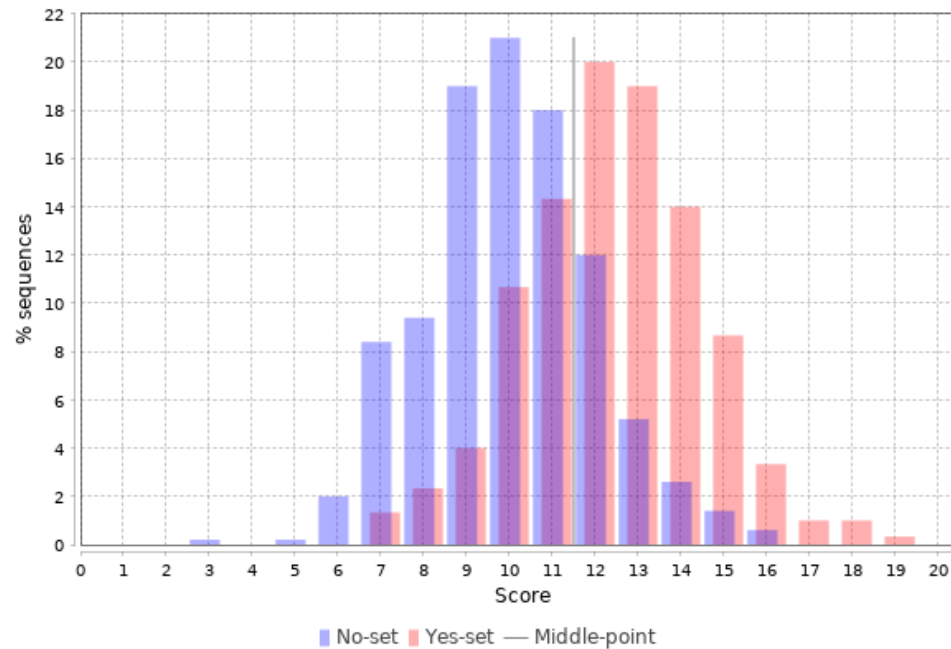
**Average no-set score: 10.05**

**AUC: 0.79**

**Middle-point: 11.50**

**False-positive: 21.80%**

**False-negative: 32.67%**



[See model visualization table →](#)

On the basis of the enhancer models we identified the following transcription factors potentially regulating the **target genes** of our interest. We found 10 and 12 transcription factors controlling expression of up- and down-regulated genes respectively (see Tables 5-6).

Table 5. 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
<a href="#">MO000057829</a>	SMAD2	SMAD family member 2	2.9	1.21
<a href="#">MO000019609</a>	SMAD1	SMAD family member 1	2.46	1.21
<a href="#">MO000056618</a>	POU5F1	POU class 5 homeobox 1	2.24	1.74
<a href="#">MO000031266</a>	NR3C1	nuclear receptor subfamily 3 group C member 1	2.21	1.11
<a href="#">MO000117988</a>	TFCP2	transcription factor CP2	1.97	1.33
<a href="#">MO000024736</a>	NR2F1	nuclear receptor subfamily 2 group F member 1	1.9	8.33
<a href="#">MO000026694</a>	EPAS1	endothelial PAS domain protein 1	1.79	1.25
<a href="#">MO000021896</a>	TBP	TATA-box binding protein	1.78	1.23
<a href="#">MO000019356</a>	NFKB1	nuclear factor kappa B subunit 1	1.72	1.32
<a href="#">MO000024664</a>	NR2F2	nuclear receptor subfamily 2 group F member 2	0	8.33

Table 6. 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
<a href="#">MO000013123</a>	STAT3	signal transducer and activator of transcription 3	1.89	1.23
<a href="#">MO000019609</a>	SMAD1	SMAD family member 1	1.83	1.52
<a href="#">MO000057832</a>	SMAD3	SMAD family member 3	1.77	1.72
<a href="#">MO000057829</a>	SMAD2	SMAD family member 2	1.74	1.52
<a href="#">MO000085616</a>	SPI1	Spi-1 proto-oncogene	1.57	1.34
<a href="#">MO000032489</a>	TAL1	TAL bHLH transcription factor 1, erythroid differentiation factor	1.54	1.41
<a href="#">MO000020402</a>	SMAD4	SMAD family member 4	1.39	1.67
<a href="#">MO000031266</a>	NR3C1	nuclear receptor subfamily 3 group C member 1	1.29	1.6
<a href="#">MO000025957</a>	PAX2	paired box 2	1.19	3.01
<a href="#">MO000024921</a>	TCF4	transcription factor 4	0.69	1.31

### **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" [5] 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 7-8.

Table 7. 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)
<a href="#">MO000031101</a>	<a href="#">plk3(h)</a>	PLK3	polo like kinase 3	0	20	1.21
<a href="#">MO000138699</a>	<a href="#">plk3(h)</a>	PLK3	polo like kinase 3	0	28	1.21
<a href="#">MO000031189</a>	<a href="#">PKCdelta(h)</a>	PRKCD	protein kinase C delta	1	29	0.86
<a href="#">MO000033396</a>	<a href="#">DUSP5(h)</a>	DUSP5	dual specificity phosphatase 5	0	29	1.21
<a href="#">MO000137304</a>	<a href="#">DUSP5(h)</a>	DUSP5	dual specificity phosphatase 5	0	30	1.21
<a href="#">MO000039099</a>	<a href="#">IL-1beta-p17:IL-1RI:IL-1RAcP:MyD88:tollip:IRAK-1{pS376}{pT387}:IRAK-4:IRAK-2</a>	IL1B, IL1R1, IL1RAP, IRAK1, IRAK2, IRAK4, MYD88, TOLLIP	interleukin 1 beta, interleukin 1 receptor accessory protein, interleukin 1 receptor associated kina...	1	33	1.33
<a href="#">MO000059577</a>	<a href="#">PKCdelta(h)</a>	PRKCD	protein kinase C delta	1	35	0.86
<a href="#">MO000038316</a>	<a href="#">LPS:lbp:CD14:TLR4:MD-2:TIRAP:IRAK-2</a>	CD14, IRAK2, LBP, LY96, TIRAP, TLR4	CD14 molecule, TIR domain containing adaptor protein, interleukin 1 receptor associated kinase 2, li...	0	39	1.33
<a href="#">MO000021128</a>	<a href="#">Hsp70-1(h)</a>	HSPA1A	heat shock protein family A (Hsp70) member 1A	0	40	1.7
<a href="#">MO000022223</a>	<a href="#">PAC-1(h)</a>	DUSP2	dual specificity phosphatase 2	0	44	1.77



Table 8. 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)
<a href="#">MO000021305</a>	TGFbeta2(h)	TGFB2	transforming growth factor beta 2	1	52	-1.48
<a href="#">MO000033313</a>	PKACA(h)	PRKACA	protein kinase cAMP-activated catalytic subunit alpha	1	61	-1.18
<a href="#">MO000021274</a>	caveolin-1(h)	CAV1	caveolin 1	1	64	-2.03
<a href="#">MO000017291</a>	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	66	-1.41
<a href="#">MO000038590</a>	Rac1:GTP:MEKK4	CYBA, CYBB, MAP3K4, NCF1, NCF2, NCF4, RAC1, SYTL1	cytochrome b-245 alpha chain, cytochrome b-245 beta chain, mitogen-activated protein kinase kinase k...	1	69	-1.02
<a href="#">MO000022340</a>	IL-8(h):CXCR2(h):G-alpha-i2(h)	CXCL8, CXCR2, GNAI2	C-X-C motif chemokine ligand 8, C-X-C motif chemokine receptor 2, G protein subunit alpha i2	1	76	-1.47
<a href="#">MO000102457</a>	PKACA-isoform1(h)	PRKACA	protein kinase cAMP-activated catalytic subunit alpha	1	76	-1.18
<a href="#">MO000022339</a>	IL-8(h):CXCR1(h):G-alpha-i2(h)	CXCL8, CXCR1, GNAI2	C-X-C motif chemokine ligand 8, C-X-C motif chemokine receptor 1, G protein subunit alpha i2	1	77	-1.47
<a href="#">MO000102458</a>	PKACA-isoform2(h)	PRKACA	protein kinase cAMP-activated catalytic subunit alpha	1	77	-1.18
<a href="#">MO000279336</a>	Rac1:GTP:pak2	CYBA, CYBB, NCF1, NCF2, NCF4, PAK2, RAC1, SYTL1	cytochrome b-245 alpha chain, cytochrome b-245 beta chain, neutrophil cytosolic factor 1, neutrophil...	1	78	-1.02

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.





Table 9. 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)
<a href="#">ENSG00000163932</a>	PRKCD	protein kinase C delta	2	1	51	0.86
<a href="#">ENSG00000152256</a>	PDK1	pyruvate dehydrogenase kinase 1	2	0	66	0.94

Table 10. 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.

[See full table →](#)

ID	Name	Target names	Target activity score	NA	Phase 1	Phase 2	Phase 3	Phase 4	Drug rank
<a href="#">DB04376</a>	13-Acetylphorbol	PRKCD	0.12						4
<a href="#">DB05013</a>	Ingenol Mebutate	PRKCD	9.12E-2	Keratosis, Keratosis, Actinic	Keratosis, Keratosis, Actinic, Warts	Carcinoma, Basal Cell, Keratosis, Keratosis, Actinic, Keratosis, Seborrheic, Noma, Sunburn	Keratosis, Keratosis, Actinic	Keratosis, Keratosis, Actinic	4
<a href="#">DB07403</a>	4-[(3-CHLORO-4-[(2R)-3,3,3-TRIFLUORO-2-HYDROXY-2-METHYLPROPANOYL]AMINO}PHENYL)SULFONYL]-N,N-DIMETHY...	PDK1	6.53E-3						6
<a href="#">DB08809</a>	Dichloroacetic Acid	PDK1	6.53E-3						6

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 [12-14] on the basis of a (Q)SAR approach. Table 11 shows the resulting list of druggable master regulators, which represent the predicted drug targets of the studied pathology. Table 12 lists chemical compounds and known drugs potentially acting on the corresponding master regulators.

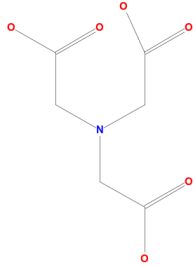
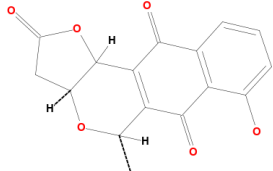
Table 11. Extended list of drug targets revealed in this study (targets that are predicted by PASS program potentially targeted by an extended list of known drugs and pharmaceutically active chemical compounds). The column **Druggability score** contains a numeric value which indicates how suitable this target is to be inhibited (or activated) by a drug. See [Methods](#) section for details.

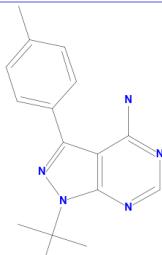
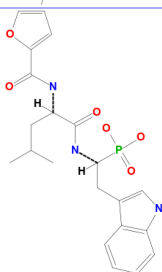
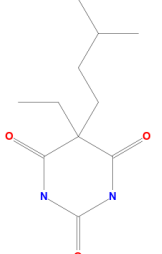
[See full table →](#)

ID	Name	Gene symbol	Gene description	Druggability score	Contained in proteome set	Total rank	logFC (transcriptome)
<a href="#">ENSG00000138166</a>	DUSP5	DUSP5	dual specificity phosphatase 5	13.33	0	30	1.21
<a href="#">ENSG00000158050</a>	DUSP2	DUSP2	dual specificity phosphatase 2	13.33	0	44	1.77
<a href="#">ENSG00000125845</a>	BMP2	BMP2	bone morphogenetic protein 2	0.45	0	50	1.79
<a href="#">ENSG00000163932</a>	PRKCD	PRKCD	protein kinase C delta	1.97	1	51	0.86
<a href="#">ENSG00000197442</a>	MAP3K5	MAP3K5	mitogen-activated protein kinase kinase kinase 5	1.4	1	66	1.13
<a href="#">ENSG00000198355</a>	PIM3	PIM3	Pim-3 proto-oncogene, serine/threonine kinase	0.83	0	73	0.89
<a href="#">ENSG00000164086</a>	DUSP7	DUSP7	dual specificity phosphatase 7	13.33	0	75	0.87
<a href="#">ENSG00000164045</a>	CDC25A	CDC25A	cell division cycle 25A	0.9	0	86	0.85

Table 12. 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. **Toxicity score** column contains maximal value of probability to be active for all toxicities corresponding to the given compound. **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.

[See full table →](#)

Name	Structure	Target names	Target activity score	Toxicity score	Disease activity score	Drug rank
Nitrilotriacetic Acid		DUSP2, DUSP5, MAP3K5, DUSP7	8.6E-2	0.97	0.25	12
Nanaomycin D		DUSP2, DUSP5, DUSP7	5.28E-2	0.86	0.55	17

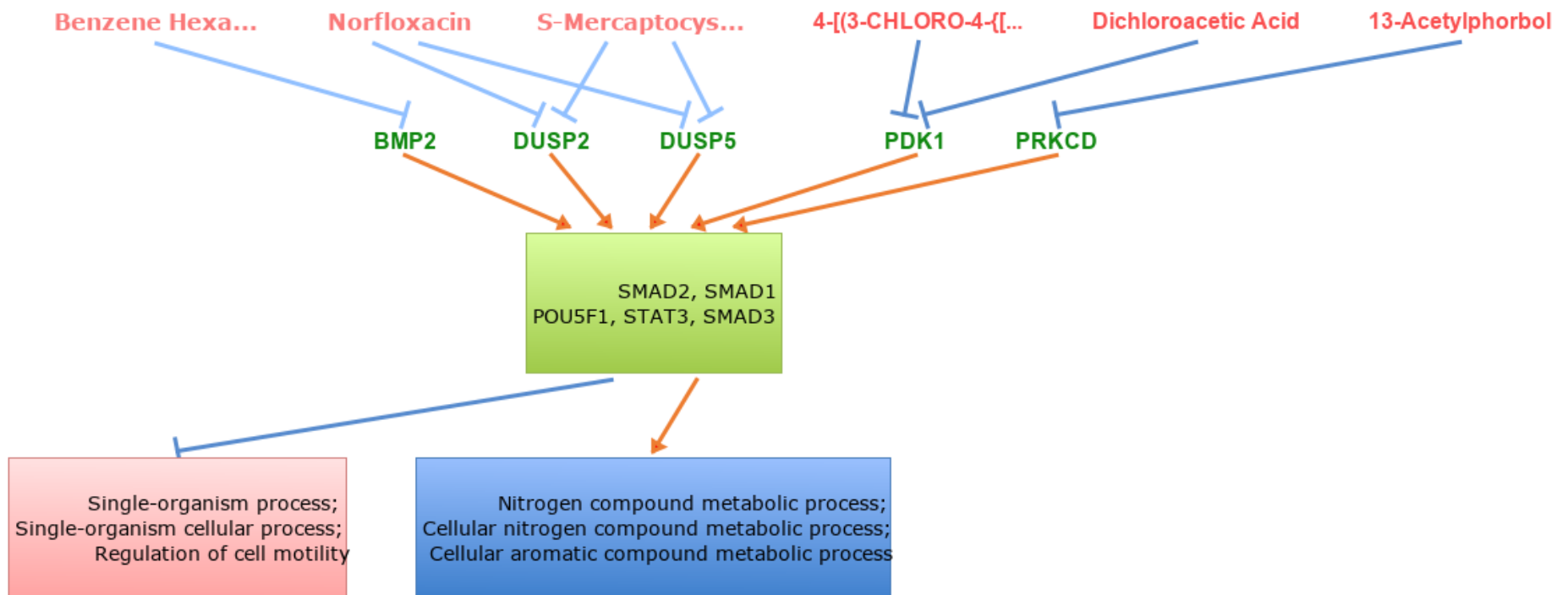
1-Ter-Butyl-3-P-Tolyl-1h-Pyrazolo[3,4-D]Pyrimidin-4-Ylamine		DUSP2, DUSP5, DUSP7	7.31E-2	0.85	0.27	17
N-[(Furan-2-Yl)Carbonyl]-(S)-Leucyl-(R)-[1-Amino-2(1h-Indol-3-Yl)Ethyl]-Phosphonic Acid		DUSP2, DUSP5, DUSP7	6.19E-2	0.61	0.39	18
Amobarbital		CDC25A, PRKCD	6.55E-2	0.98	0.23	20

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 [6], release 2019.2 (geneXplain GmbH, Wolfenbüttel, Germany) (<http://genexplain.com/transfac>).

The master regulator search uses the TRANSPATH® database (BIOBASE) [9]. A comprehensive signal transduction network of human cells is built by the software on the basis of reactions annotated in TRANSPATH®.

### 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 [4,5]. 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:

- ranking by "Target activity score" ( $T\text{-score}_{PSD}$ ),
- ranking by "Disease activity score" ( $D\text{-score}_{PSD}$ ),
- 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" ( $T\text{-score}_{PSD}$ ) 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" ( $D\text{-score}_{PSD}$ ):

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

where  $D$  is the set of selected diseases, and if  $D$  is empty set,  $D\text{-score}_{PSD}=0$ .  $P$  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:

- Toxicity below a chosen toxicity threshold (defines as  $Pa$ , probability to be active as toxic substance).
- For all predicted pharmacological effects that correspond to a set of user selected disease(s)  $Pa$  is greater than a chosen effect threshold.
- 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. 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
3. 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
4. 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.
5. 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
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
10. 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
11. Michael H, Hogan J, Kel A, Kel-Margoulis O, Schacherer F, Voss N. Building a knowledge base for systems pathology. *Brief Bioinformatics*. **2008**;9(6):518-531. doi:10.1093/bib/bbn038
12. 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.
13. Filimonov DA, Poroikov VV. Prognosis of specters of biological activity of organic molecules. *Russian chemical journal*. **2006**;50(2):66-75 (russ)
14. Filimonov D, Poroikov V, Borodina Y, Glorizova 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](mailto:support@genexplain.com)

### Supplementary material

- [Supplementary table 1 - Significant up-regulated genes](#)
- [Supplementary table 2 - Significant down-regulated genes](#)
- [Supplementary table 3 - Detailed report. Composite modules and master-regulators \(up-regulated genes in Myc\\_induce vs. Control\).](#)
- [Supplementary table 4 - Detailed report. Composite modules and master-regulators \(down-regulated genes in Myc\\_induce vs. Control\).](#)
- [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.