Sequence and Pathway analysis

ROCK2 and ITGA3 are promising druggable targets for treating Squamous Cell Carcinoma that control activity of TP53, SRF and GTF2I transcription factors on promoters of differentially expressed genes in esophagus tissue

Demo User
geneXplain GmbH
info@genexplain.com

Data received on 13/08/2019 ; Run on 11/06/2020 ; Report generated on 11/06/2020

Genome Enhancer release 2.0 (TRANSFAC®, TRANSPATH® and HumanPSD™ release 2020.2)

Abstract

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

From the data set analyzed in this study, we found the following TFs to be potentially involved in the regulation of the differentially expressed genes: TP53, SRF, E2F1, GTF2I and TAL1. The subsequent network analysis suggested

- integrins
- ROCK-II
- Cdk6:cyclinD3-isoform1
- IL-1beta-p17:IL-1RI:IL-1RAcP:MyD88:tollip:IRAK-1{pS376}{pT387}:IRAK-4:IRAK-2
- setd7

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

1. Introduction

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

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

The "upstream analysis" approach has now been extended by a third step that reveals known drugs suitable to inhibit (or activate) the identified molecular targets in the context of the disease under study. This step is performed by using information from HumanPSD™ database [5]. In addition, some known drugs and investigational active chemical compounds are subsequently predicted as potential ligands for the revealed molecular targets. They are predicted using a pre-computed database of spectra of biological activities of chemical compounds of a library of 2507 known drugs and investigational chemical compounds from HumanPSD™ database. The spectra of biological activities for these compounds are computed using the program PASS on the basis of a (Q)SAR approach [11-13]. These predictions can be used for the research purposes - for further drug development and drug repurposing initiatives.

2. Data

For this study the following experimental data was used:

Table 1. Experimental datasets used in the study

File name	Data type
SRR349741.fastq	Transcriptomics
SRR349742.fastq	Transcriptomics
SRR349748.fastq	Transcriptomics
SRR349749.fastq	Transcriptomics

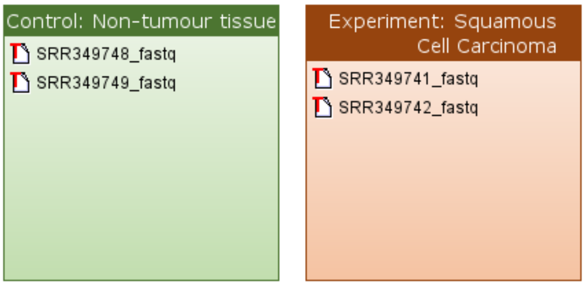

Control: Non-tumour tissue
SRR349748_fastq
SRR349749_fastq
Experiment: Squamous Cell Carcinoma
SRR349741_fastq
SRR349742_fastq

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: Experiment: Squamous Cell Carcinoma *versus* Control: Non-tumour tissue.

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 edgeR tool (R/Bioconductor package integrated into our pipeline) and compared gene expression in the following sets: "Experiment: Squamous Cell Carcinoma" with "Control: Non-tumour tissue". edgeR 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 4994 upregulated genes (LogFC>0) out of which 1436 genes were found as significantly upregulated (p-value<0.1) and 3767 downregulated genes (LogFC<0) out of which 513 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 Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue.*

See full table →

ID	Gene symbol	Gene description	logFC	logCPM	PValue	FDR
ENSG00000115758	ODC1	ornithine decarboxylase 1	7.17	10.32	2.21E-11	6.44E-8
ENSG00000148053	NTRK2	neurotrophic receptor tyrosine kinase 2	6.48	9.32	5.21E-11	1.14E-7
ENSG00000113140	SPARC	secreted protein acidic and cysteine rich	6.14	10.69	2.91E-9	2.03E-6
ENSG00000163359	COL6A3	collagen type VI alpha 3 chain	5.68	9.13	2.4E-8	1E-5
ENSG00000120708	TGFBI	transforming growth factor beta induced	5.24	8.77	6.25E-10	6.08E-7
ENSG00000134871	COL4A2	collagen type IV alpha 2 chain	5.14	7.97	1.36E-10	2.38E-7
ENSG00000186340	THBS2	thrombospondin 2	5.1	8.46	2.19E-7	5.04E-5
ENSG00000146648	EGFR	epidermal growth factor receptor	4.92	9.64	4.36E-6	5.44E-4
ENSG00000144824	PHLDB2	pleckstrin homology like domain family B member 2	4.9	8.29	3.7E-9	2.03E-6
ENSG00000145824	CXCL14	C-X-C motif chemokine ligand 14	4.89	8.54	1.11E-7	3.05E-5

*Table 4. Top ten significant **down-regulated** genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue.*

See full table →

ID	Gene symbol	Gene description	logFC	logCPM	PValue	FDR
ENSG00000136155	SCEL	sciellin	-7.36	10.74	2.01E-12	1.76E-8
ENSG00000163209	SPRR3	small proline rich protein 3	-6.39	14.08	2.27E-5	2E-3
ENSG00000143369	ECM1	extracellular matrix protein 1	-6.04	10.66	2.28E-9	1.82E-6
ENSG00000189334	S100A14	S100 calcium binding protein A14	-6	10.05	7.93E-10	6.95E-7
ENSG00000229732	AC019349.1	novel transcript	-5.88	12.56	3.53E-9	2.03E-6
ENSG00000086548	CEACAM6	CEA cell adhesion molecule 6	-5.82	9.92	2.89E-10	3.61E-7
ENSG00000171401	KRT13	keratin 13	-5.76	14.53	2.55E-8	1.02E-5
ENSG00000087128	TMPRSS11E	transmembrane serine protease 11E	-5.67	9.79	2.03E-8	8.91E-6
ENSG00000197632	SERPINB2	serpin family B member 2	-5.5	8.35	1.72E-10	2.51E-7
ENSG00000165272	AQP3	aquaporin 3 (Gill blood group)	-5.46	10.95	2.63E-6	3.78E-4

3.2. Regulatory regions of target genes

We mapped the uploaded Epigenomic peaks on the ***target genes*** and selected those peaks only that were found located in the body of the gene (in exons or introns of the genes) or in the 5000 nucleotide long flanking regions of the genes. In the tables below we demonstrate localization of such potential regulatory regions in the top up-regulated and down-regulated genes.

*Table 3. Top ten **up-regulated** genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue with epigenomic peaks.*

See full table →

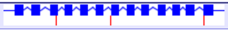
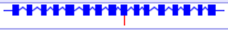
ID	Gene symbol	Gene schematic representation
ENSG00000115758	ODC1	
ENSG00000113140	SPARC	
ENSG00000163359	COL6A3	
ENSG00000120708	TGFBI	
ENSG00000134871	COL4A2	
ENSG00000186340	THBS2	
ENSG00000146648	EGFR	
ENSG00000182326	C1S	
ENSG00000122786	CALD1	
ENSG00000053747	LAMA3	

*Table 5. Top ten **down-regulated** genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue with epigenomic peaks.*

See full table →

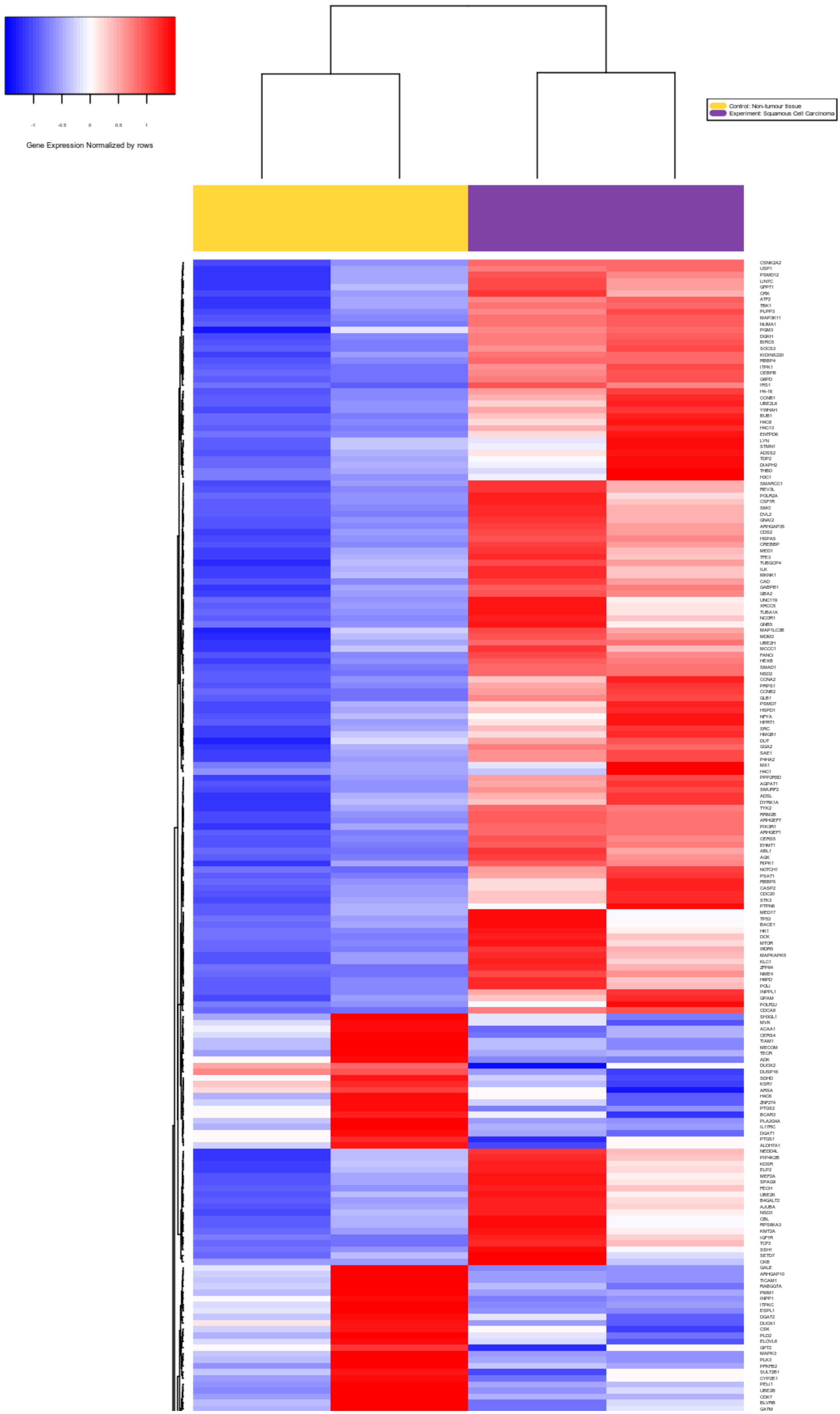
ID	Gene symbol	Gene schematic representation
ENSG00000163209	SPRR3	
ENSG00000244094	SPRR2F	
ENSG00000177191	B3GNT8	
ENSG00000260276	AC022167.2	
ENSG00000124466	LYPD3	
ENSG00000074416	MGLL	
ENSG00000153048	CARHSP1	
ENSG00000170545	SMAGP	
ENSG00000211448	DIO2	
ENSG00000135373	EHF	

3.3. 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 Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue

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: Non-tumour tissue
Experiment: Squamous Cell Carcinoma
CSNK2A2
USP1
PSMD12
LIN7C
GFPT1
CRK
ATF2
TBK1
PLPP3
MAP3K11
NUMA1
PGM3
DGKH
BIRC5
SOCS3
KIDINS220
RBBP4
ITPK1
CEBPB
G6PD
IRS1
H4-16
CCNB1
UBE2L6
YWHAH
BUB1
H4C8
H4C13
ENTPD6
LYN
STMN1
ADSS2
TDP2
DIAPH2
THBD
H3C1
SMARCC1
REV3L
POLR2A
CSF1R
SMO
DVL2
GNAI2
ARHGAP35
CDS2
HSPA5
CREBBP
MED1
TFE3
TUBGCP4
ILK
MKNK1
CAD
GABPB1
GBA2
UNC119
XRCC5
TUBA1A
NCOR1
GNB5
MAP1LC3B
MDM2
UBE2H
MCCC1
FANCI
HEXB
SMAD1
NSD2
CCNA2
PRPS1
CCNB2
GLB1
PSMD7
HSPD1
NFYA
HPRT1
SRC
HMGB1
DUT
GGA2
SAE1
P4HA2
MX1
H4C1
PPP2R5D
AGPAT1
SMURF2
ADSL
DYRK1A
TYK2
RRM2B
ARHGEF7
PIK3R1
ARHGEF1
CERS5
EHMT1
ABL1
AGK
RIPK1
NOTCH1
PSAT1
RBBP5
CASP2
CDC20
STK3
PTPN6
MED17
TP53
BACE1
HK1
DCK
MTOR
WDR5
MAPKAPK5
KLC1
ZFP64
NME4
H6PD
POLI
INPPL1
GPAM
POLR2J
CDCA8
SH3GL1
MVK
ACAA1
CERS4
TIAM1
MECOM
TECR
ADK
DUOX2
DUSP16
SDHD
KSR1
ARSA
H4C6
ZNF274
PTGS2
BCAR3
PLA2G4A
IL17RC
DGAT1
PTGS1
ALDH7A1
NEDD4L
PIP4K2B
KDSR
ELP2
MEF2A
SPAG9
FECH
UBE2K
B4GALT2
AJUBA
NSD3
CBL
RPS6KA3
KMT2A
IGF1R
TCF3
SSH1
SETD7
CKB
GALE
ARHGAP10
TICAM1
RABGGTA
PMM1
INPP1
ITPKC
ESPL1
DGAT2
DUOX1
CSK
PLD2
ELOVL6
GPT2
MAPK3
PLK3
PFKFB2
SULT2B1
CYP2E1
PELI1
UBE2B
CDK7
BLVRB
GATM

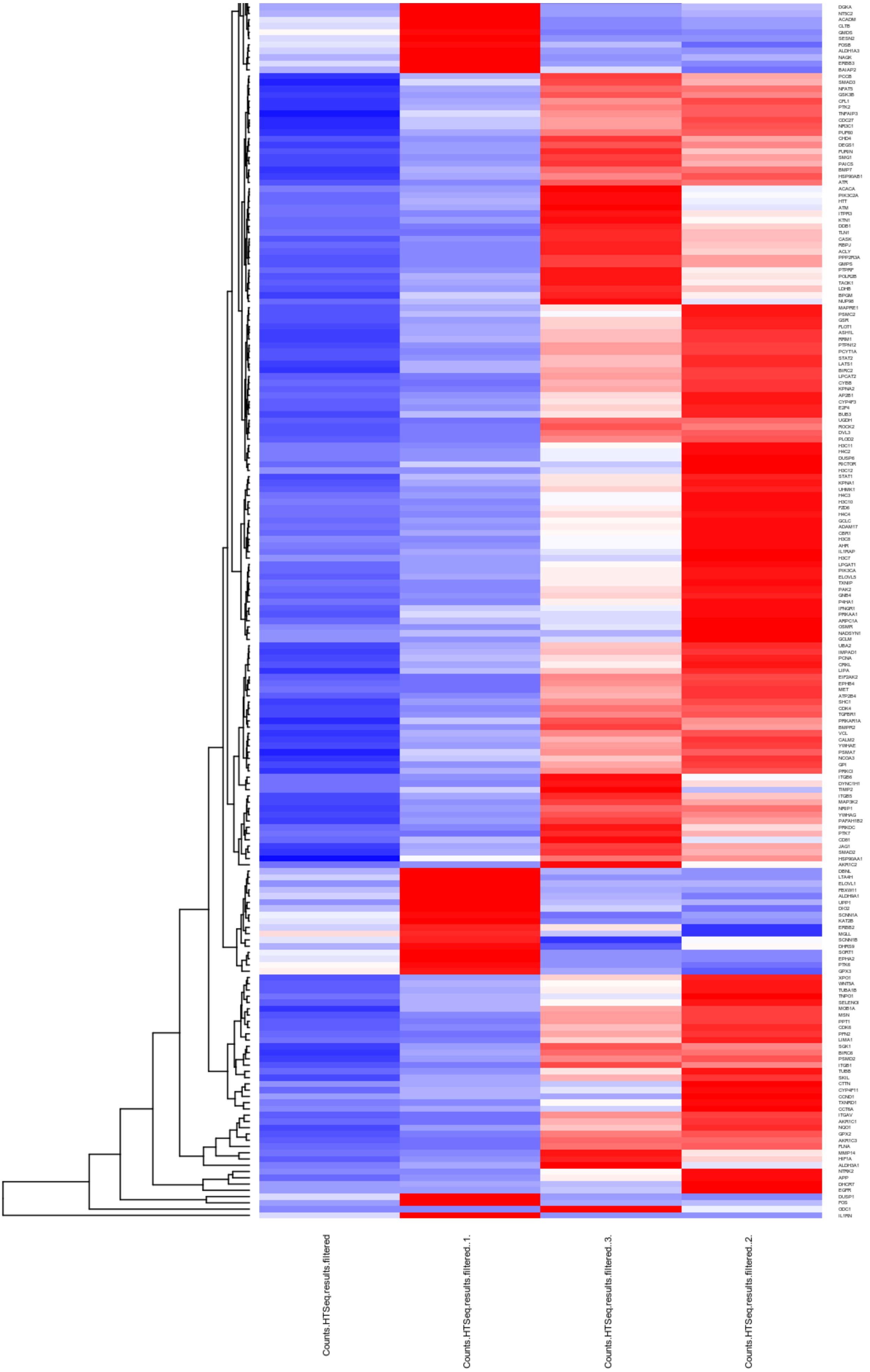

DGKA
NT5C2
ACADM
CLTB
GMDS
SESN2
FOSB
ALDH1A3
NAGK
ERBB3
BAIAP2
PCCB
SMAD3
NFAT5
GSK3B
CPL1
PTK2
TNFAIP3
CDC27
NR3C1
PUF60
CHD4
DEGS1
FURIN
SMG1
PAICS
BMP7
HSP90AB1
ATR
ACACA
PIK3C2A
HTT
ATM
ITPR3
KTN1
DDB1
TLN1
CASK
RBPJ
ACLY
PPP2R3A
GMPS
PTPRF
POLR2B
TAOK1
LDHB
BPGM
NUP98
MAPRE1
PSMC2
GSR
FLOT1
ASH1L
RRM1
PTPN12
PCYT1A
STAT2
LATS1
BIRC2
LPCAT2
CYBB
KPNA2
AP2B1
CYP4F3
E2F4
BUB3
UGDH
ROCK2
DVL3
PLOD2
H3C11
H4C2
DUSP6
RICTOR
H3C12
STAT1
KPNA1
UHMK1
H4C3
H3C10
FZD6
H4C4
GCLC
ADAM17
CBR1
H3C8
AHR
IL1RAP
H3C7
LPGAT1
PIK3CA
ELOVL5
TXNIP
PAK2
GNB4
P4HA1
IFNGR1
PRKAA1
ARPC1A
OSMR
NADSYN1
GCLM
UBA2
IMPAD1
PCNA
CRKL
LIPA
EIF2AK2
EPHB4
MET
ATP2B4
SHC1
CDK4
TGFBR1
PRKAR1A
BMPR2
VCL
CALM2
YWHAE
PSMA7
NCOA3
GPI
PRKCI
ITGB6
DYNC1H1
TIMP2
ITGB5
MAP3K2
NRIP1
YWHAG
PAFAH1B2
PRKDC
PTK7
CD81
JAG1
SMAD2
HSP90AA1
AKR1C2
DBNL
LTA4H
ELOVL1
FBXW11
ALDH9A1
UPP1
DIO2
SCNN1A
KAT2B
ERBB2
MGLL
SCNN1B
DHRS9
SORT1
EPHA2
PTK6
GPX3
XPO1
WNT5A
TUBA1B
TNPO1
SELENOI
MOB1A
MSN
PPT1
CDK6
PFN2
LIMA1
SGK1
BIRC6
PSMD2
ITGB1
TUBB
SKIL
CTTN
CYP4F11
CCND1
TXNRD1
CCT6A
ITGAV
AKR1C1
NQO1
GPX2
AKR1C3
FLNA
MMP14
HIF1A
ALDH3A1
NTRK2
APP
DHCR7
EGFR
DUSP1
FOS
ODC1
IL1RN
Counts.HTSeq.results.filtered
Counts.HTSeq.results.filtered..1.
Counts.HTSeq.results.filtered..3.
Counts.HTSeq.results.filtered..2.

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 Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue:

1436 significant up-regulated genes were taken for the mapping.

GO (biological process)

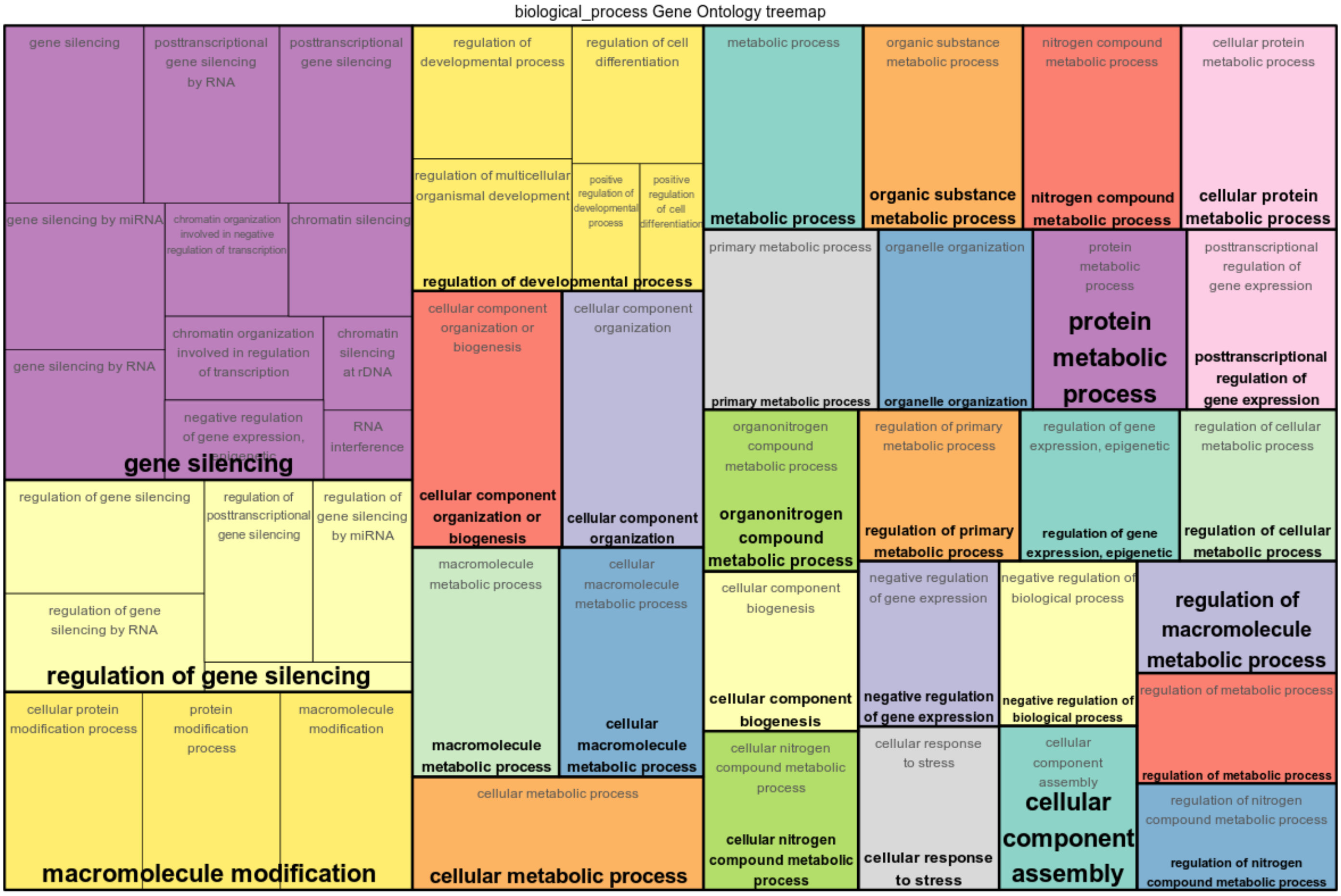


Figure 3. Enriched GO (biological process) of up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue.
Full classification →

TRANSPATH® Pathways (2020.2)

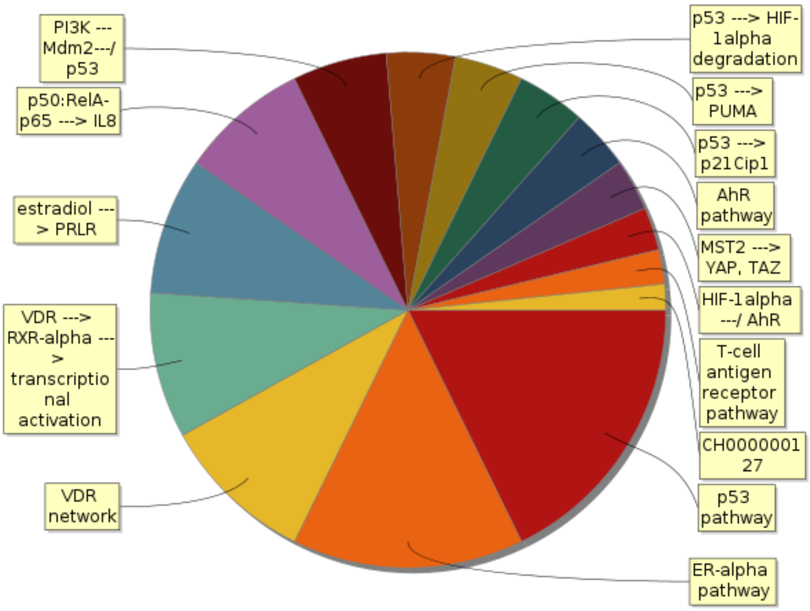


Figure 4. Enriched TRANSPATH® Pathways (2020.2) of up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue.
Full classification →

HumanPSD(TM) disease (2020.2)

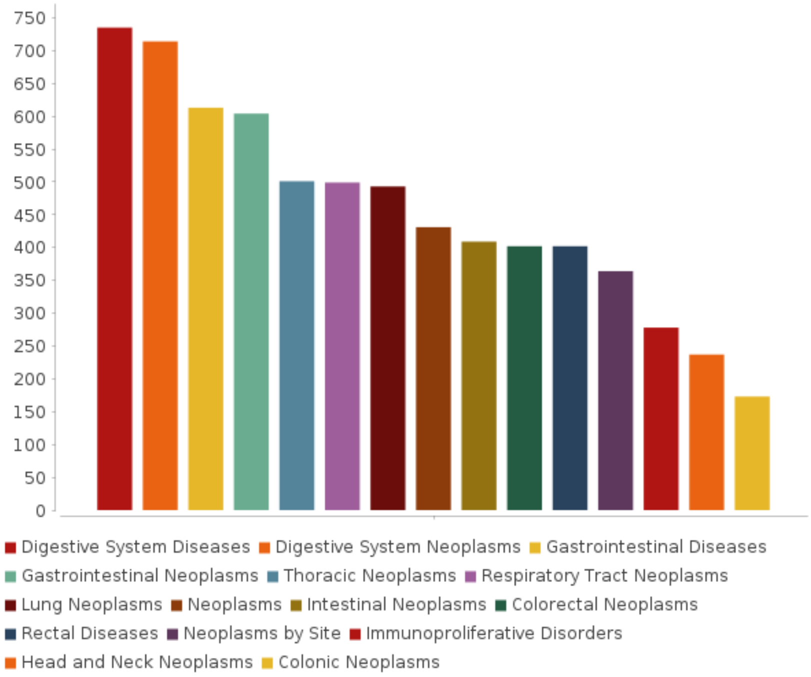


Figure 5. Enriched HumanPSD(TM) disease (2020.2) of up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.
Full classification →

Down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue:

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

GO (biological process)

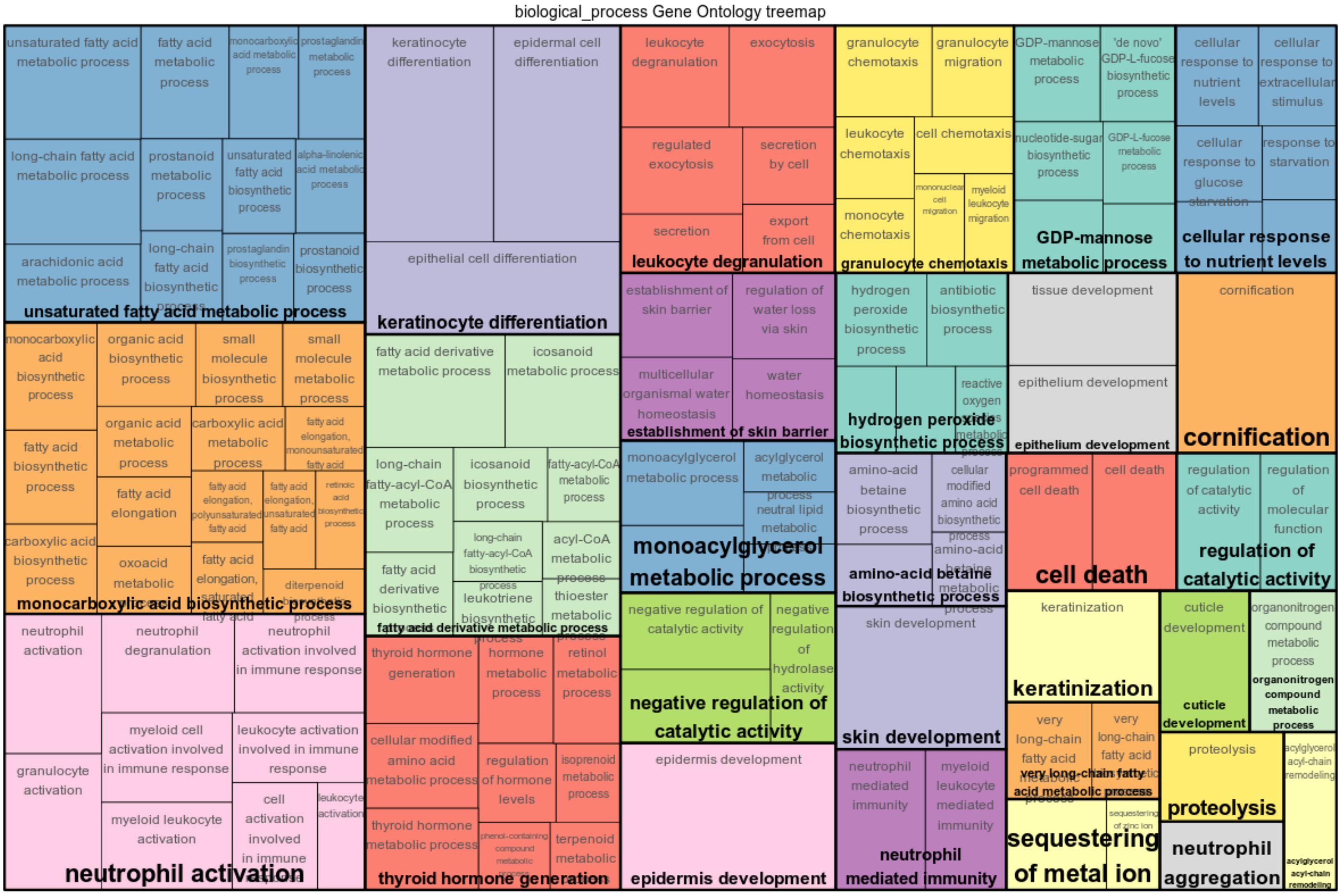


Figure 6. Enriched GO (biological process) of down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue.

Full classification →

TRANSPATH® Pathways (2020.2)

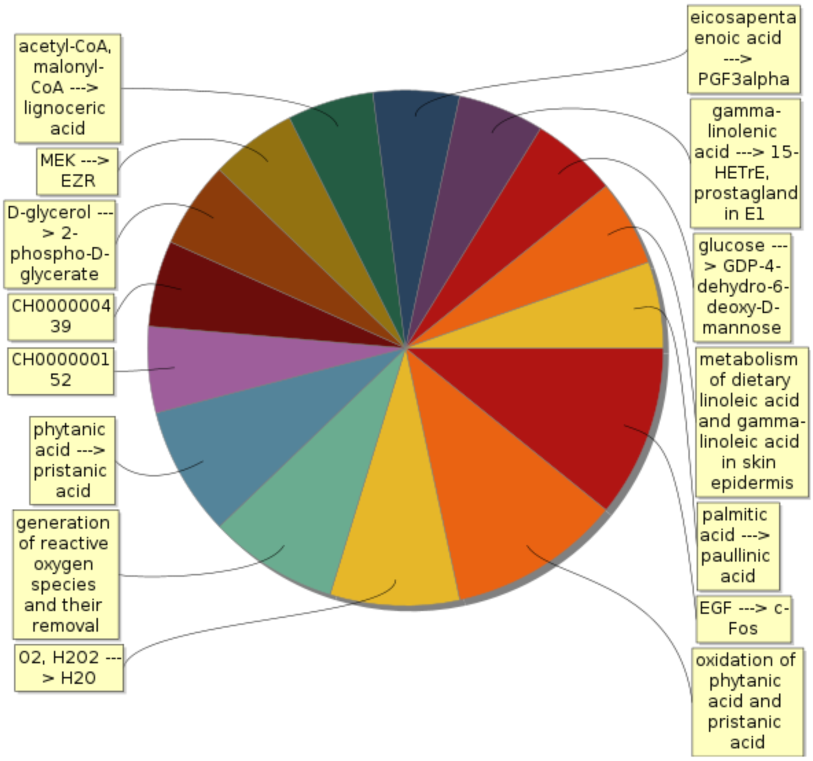


Figure 7. Enriched TRANSPATH® Pathways (2020.2) of down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue.

Full classification →

HumanPSD(TM) disease (2020.2)

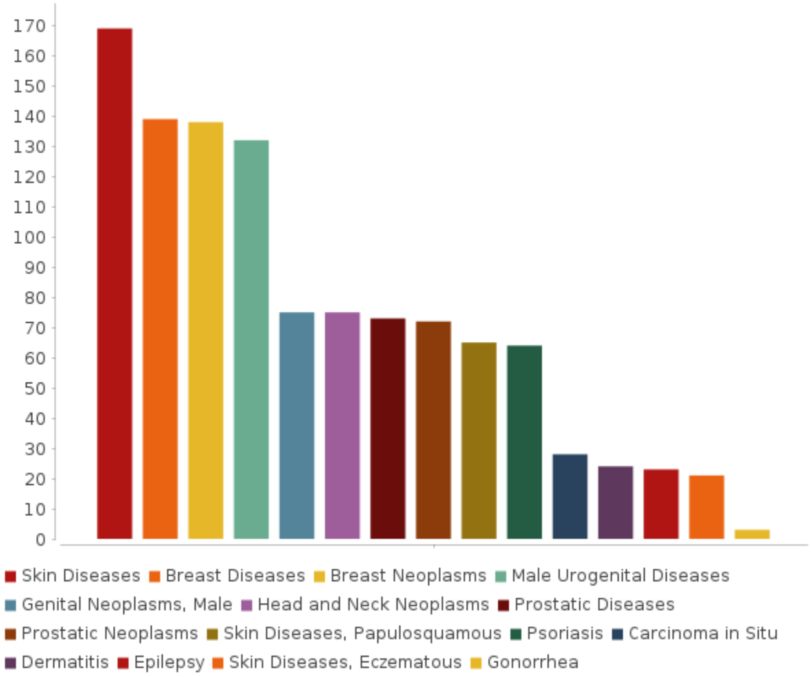
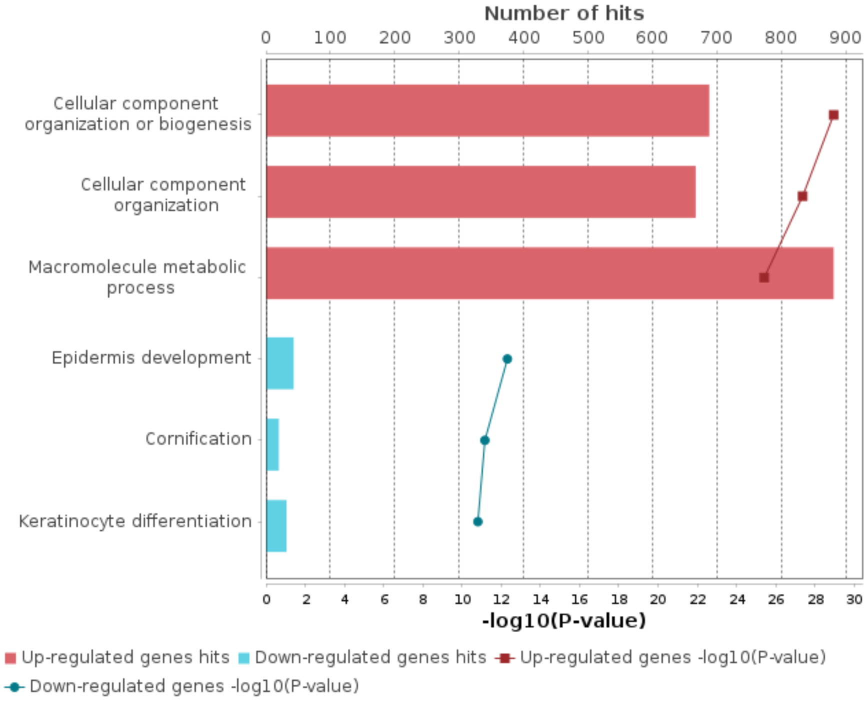


Figure 8. Enriched HumanPSD(TM) disease (2020.2) of down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

Full classification →

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



3.4. 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].

In the current work, we use the Genomics data from the "Yes VCF track" track to predict positions of potential ***enhancers*** where the observed sequence variations may influence the gene expression in the pathology under study. We scan 5kb flanking regions and the body of all genes caring the variations, with a sliding window of 1100bp size and find the position of the window with the maximal sum of the mutation weights, where we then perform the search for potential condition-specific enhancers (CMA model search).

We analyzed mutations that were revealed in the potential enhancers located upstream, downstream or inside the ***target genes*** (see Table 6). We identified 179 mutations potentially affecting gene regulation. Table 7 shows the following lists of PWMs whose sites were lost or gained due to these mutations. These PWMs were put in focus of the CMA algorithm that constructs the model of the enhancers by specifying combinations of TF motifs (see more details of the algorithm in the Method section).

Table 6. Mutations revealed in Experiment: Squamous Cell Carcinoma versus Control: Non-tumour tissue

See full table →

ID	Gene symbol	Gene schematic representation	Number of variations
ENSG00000186340	THBS2		8
ENSG00000226445	BX322234.1		7
ENSG00000142173	COL6A2		5
ENSG00000134871	COL4A2		4
ENSG00000171903	CYP4F11		4
ENSG00000063660	GPC1		3
ENSG00000115758	ODC1		3
ENSG00000139178	C1RL		3
ENSG00000149212	SESN3		3
ENSG00000152291	TGOLN2		3

Table 7. PWMs whose sites were lost or gained due to mutations in Experiment: Squamous Cell Carcinoma and Control: Non-tumour tissue

See full table →

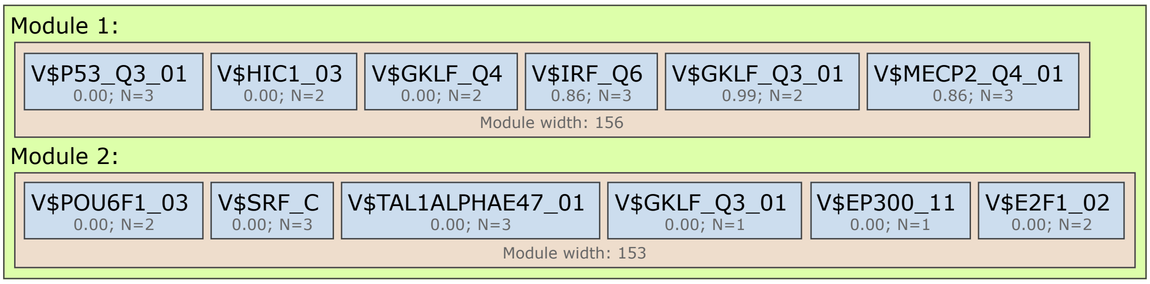
ID	P-value (gains)	P-value (losses)	yesCount (gains)	yesCount (losses)
V$E2F1_05	9.62E-3	1.07E-13	9	233
V$E2F1_Q4_02	8.15E-3	9.15E-13	13	265
V$E2F4_Q6	1.12E-3	1.27E-16	34	237
V$SP1_Q6	6.36E-4	1.33E-12	28	420
V$P53_Q3_01	6.06E-8		457	null
V$OSX_Q3	2.24E-8		56	null
V$E2F4_Q3	4.77E-9	8.53E-11	105	138
V$E2F1_09	2.99E-9	7.36E-12	98	144
V$E2F_Q6_01	7.41E-10	7.19E-3	307	25
V$E2F1_Q6_01	5.58E-11	1.03E-23	207	294
V$E2F4_09	2.57E-11	5.42E-15	154	219
V$E2F_Q3_01	7.67E-13	7.93E-22	165	241
V$TIEG1_02	2.7E-13	5.54E-8	136	147
V$E2F4_05	4.32E-14	7.51E-17	192	255
V$GCM_Q2		1.48E-15	null	328

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 Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue).

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$P53_Q3_01 0.00; N=3
V$HIC1_03 0.00; N=2
V$GKLF_Q4 0.00; N=2
V$IRF_Q6 0.86; N=3
V$GKLF_Q3_01 0.99; N=2
V$MECP2_Q4_01 0.86; N=3

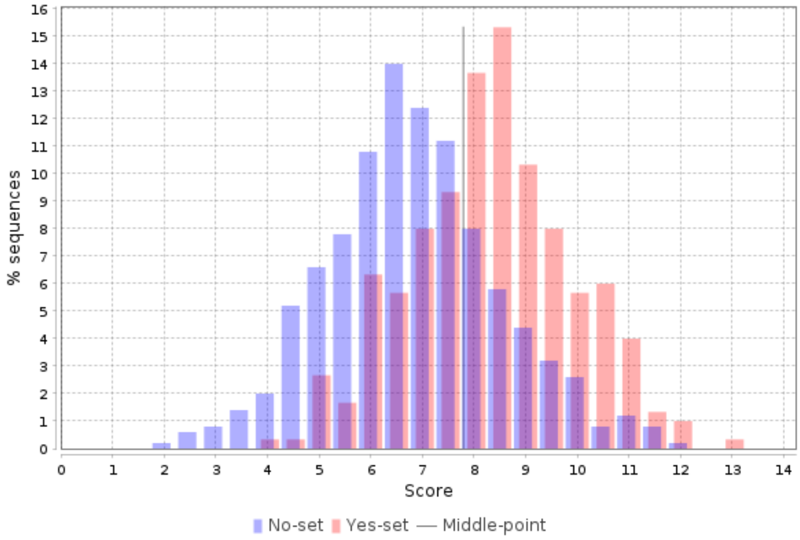
Module width: 156

Module 2:

V$POU6F1_03 0.00; N=2
V$SRF_C 0.00; N=3
V$TAL1ALPHAE47_01 0.00; N=3
V$GKLF_Q3_01 0.00; N=1
V$EP300_11 0.00; N=1
V$E2F1_02 0.00; N=2

Module width: 153

Model score (-p*log10(pval)): 13.88
Wilcoxon p-value (pval): 5.72e-30
Penalty (p): 0.475
Average yes-set score: 8.31
Average no-set score: 6.86
AUC: 0.74
Middle-point: 7.80
False-positive: 25.80%
False-negative: 34.67%



See model visualization table →

Table 8. List of top ten up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue 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
ENSG00000105983	LMBR1	limb development membrane protein 1	18.68	p53(h), MECP-2(h), GKLF(h), HIC-1(h), SRF(h), E2A(h),Tal-1(h), p300(h)...
ENSG00000002834	LASP1	LIM and SH3 protein 1	17.46	POU6F1(h), HIC-1(h), MECP-2(h), GKLF(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), p300(h), E2F-1(h)...
ENSG00000156642	NPTN	neuroplastin	17.32	GKLF(h), p53(h), HIC-1(h), MECP-2(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), E2A(h),Tal-1(h), E2F-1(h)...
ENSG00000087253	LPCAT2	lysophosphatidylcholine acyltransferase 2	17.22	p53(h), POU6F1(h), E2A(h),Tal-1(h), GKLF(h), HIC-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), E2F-1(h)...
ENSG00000171867	PRNP	prion protein	17.15	MECP-2(h), p53(h), HIC-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), SRF(h), GKLF(h), POU6F1(h)...
ENSG00000221963	APOL6	apolipoprotein L6	16.98	E2A(h),Tal-1(h), p300(h), SRF(h), POU6F1(h), E2F-1(h), GKLF(h), MECP-2(h)...
ENSG00000265241	RBM8A	RNA binding motif protein 8A	16.93	E2F-1(h), E2A(h),Tal-1(h), SRF(h), p300(h), GKLF(h), POU6F1(h), MECP-2(h)...
ENSG00000174282	ZBTB4	zinc finger and BTB domain containing 4	16.93	p53(h), GKLF(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), MECP-2(h), SRF(h), p300(h), E2F-1(h)...
ENSG00000071994	PDCD2	programmed cell death 2	16.89	GKLF(h), MECP-2(h), p53(h), HIC-1(h), E2A(h),Tal-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), E2F-1(h)...
ENSG00000153207	AHCTF1	AT-hook containing transcription factor 1	16.83	POU6F1(h), GKLF(h), E2A(h),Tal-1(h), E2F-1(h), SRF(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), MECP-2(h)...

Enhancer model potentially involved in regulation of target genes (down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue).

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$OCT1_06	V$TAL1BETAITF2_01	V$TR4_Q3	V$MAZ_Q6	V$AREB6_01
0.90; N=3	0.84; N=2	0.00; N=3	0.90; N=3	0.90; N=3

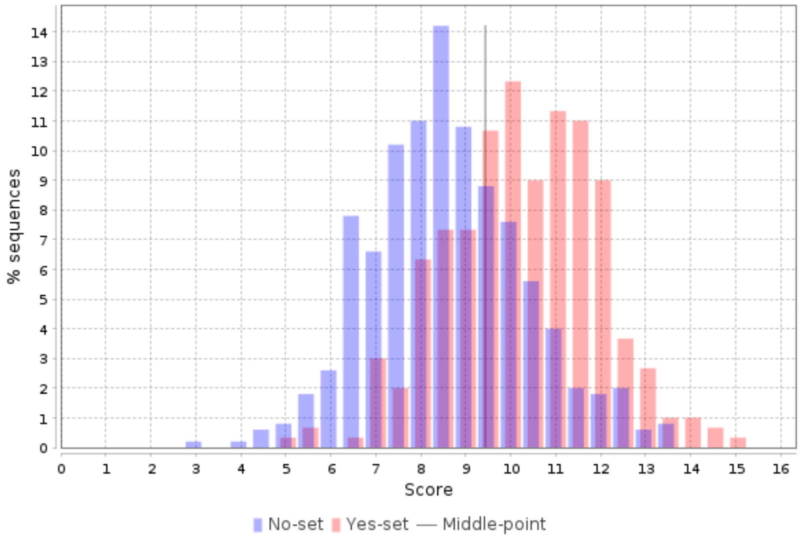
Module width: 99

Module 2:

V$P53_Q3_01	V$POU6F1_07	V$TAL1GATA1_02	V$FRA2_01	V$TFIII_Q6_01	V$ETS1_B
0.89; N=1	0.84; N=3	0.79; N=2	0.89; N=3	0.92; N=2	0.84; N=2

Module width: 135

Model score (-p*log10(pval)): 16.25
Wilcoxon p-value (pval): 4.30e-34
Penalty (p): 0.487
Average yes-set score: 10.25
Average no-set score: 8.63
AUC: 0.76
Middle-point: 9.43
False-positive: 28.40%
False-negative: 30.33%



See model visualization table →

*Table 9. List of top ten down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue 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
ENSG00000196754	S100A2	S100 calcium binding protein A2	17.33	POU2F1(h), TR4(h), ZEB1(h), MAZ(h), Fra-2(h), TFII-I(h), GATA-1(h),Tal-1(h)...
ENSG00000132109	TRIM21	tripartite motif containing 21	17.18	MAZ(h), POU6F1(h), GATA-1(h),Tal-1(h), p53(h), c-Ets-1(h), ITF-2(h),Tal-1(h), TFII-I(h)...
ENSG00000089723	OTUB2	OTU deubiquitinase, ubiquitin aldehyde binding 2	15.26	ZEB1(h), TR4(h), MAZ(h), p53(h), TFII-I(h), ITF-2(h),Tal-1(h), c-Ets-1(h)...
ENSG00000110723	EXPH5	exophilin 5	14.88	Fra-2(h), c-Ets-1(h), ZEB1(h), POU2F1(h), GATA-1(h),Tal-1(h), MAZ(h), TR4(h)...
ENSG00000173786	CNP	2',3'-cyclic nucleotide 3' phosphodiesterase	14.7	TR4(h), ZEB1(h), MAZ(h), GATA-1(h),Tal-1(h), POU2F1(h), ITF-2(h),Tal-1(h), p53(h)...
ENSG00000105701	FKBP8	FKBP prolyl isomerase 8	14.65	ZEB1(h), MAZ(h), TR4(h), ITF-2(h),Tal-1(h), POU2F1(h), GATA-1(h),Tal-1(h), TFII-I(h)...
ENSG00000182585	EPGN	epithelial mitogen	14.46	c-Ets-1(h), TFII-I(h), GATA-1(h),Tal-1(h), Fra-2(h), p53(h), POU6F1(h), ZEB1(h)...
ENSG00000235297	FAUP1	FAU pseudogene 1	14.39	c-Ets-1(h), POU2F1(h), ZEB1(h), TR4(h), Fra-2(h), ITF-2(h),Tal-1(h), p53(h)...
ENSG00000147883	CDKN2B	cyclin dependent kinase inhibitor 2B	14.38	ZEB1(h), POU2F1(h), TFII-I(h), TR4(h), GATA-1(h),Tal-1(h), Fra-2(h), POU6F1(h)...
ENSG00000196597	ZNF782	zinc finger protein 782	14.37	POU2F1(h), ZEB1(h), TR4(h), p53(h), GATA-1(h),Tal-1(h), Fra-2(h), POU6F1(h)...

On the basis of the enhancer models we identified transcription factors potentially regulating the ***target genes*** of our interest. We found 18 and 12 transcription factors controlling expression of up- and down-regulated genes respectively (see Tables 10-11).

Table 10. Transcription factors of the predicted enhancer model potentially regulating the differentially expressed genes (up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue). ***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 →

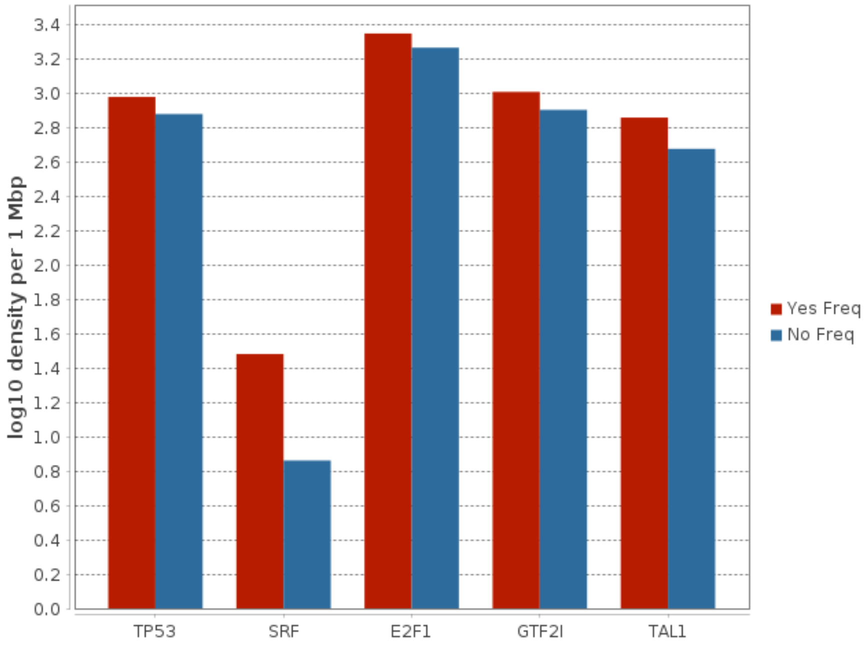
ID	Gene symbol	Gene description	Regulatory score	Yes-No ratio
MO000019548	TP53	tumor protein p53	8.67	1.26
MO000013015	SRF	serum response factor	6.54	4.17
MO000004274	E2F1	E2F transcription factor 1	6.46	1.21
MO000056654	EP300	E1A binding protein p300	6.45	1.23
MO000032492	TCF3	transcription factor 3	6.13	2.23
MO000032489	TAL1	TAL bHLH transcription factor 1, erythroid differentiation factor	6.12	1.39
MO000007691	IRF2	interferon regulatory factor 2	5.71	1.3
MO000007703	IRF7	interferon regulatory factor 7	5.71	1.37
MO000028320	null	null	5.37	1.33
MO000285816	IRF3	interferon regulatory factor 3	5.23	1.3

Table 11. Transcription factors of the predicted enhancer model potentially regulating the differentially expressed genes (down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue). ***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
MO000019548	TP53	tumor protein p53	3.42	1.48
MO000019622	GTF2I	general transcription factor IIi	2.62	1.27
MO000032489	TAL1	TAL bHLH transcription factor 1, erythroid differentiation factor	2.59	1.52
MO000059013	ETS1	ETS proto-oncogene 1, transcription factor	2.4	1.23
MO000139677	ZEB1	zinc finger E-box binding homeobox 1	2.16	1.36
MO000025003	POU2F1	POU class 2 homeobox 1	2	2.05
MO000026074	FOSL2	FOS like 2, AP-1 transcription factor subunit	1.99	2.4
MO000046001	GATA1	GATA binding protein 1	1.82	1.33
MO000028320	null	null	1.68	1.39
MO000105384	MAZ	MYC associated zinc finger protein	1.48	

The following diagram represents the key transcription factors, which were predicted to be potentially regulating differentially expressed genes in the analyzed pathology: TP53, SRF, E2F1, GTF2I and TAL1.



3.5. Finding master regulators in networks

In the second step of the upstream analysis common regulators of the revealed TFs were identified. We identified 2 signaling proteins whose structure and function is highly damaged by the mutations (see Table 12).

Table 12. Signaling proteins whose structure and function is damaged by the mutations in Experiment: Squamous Cell Carcinoma and Control: Non-tumour tissue

See full table →

ID	Title	Mutation count	Consequence	Codons
MO000172130	c3orf1(h)	1	NMD_transcript_variant,stop_lost	tGa/tCa
MO000212738	EMC10(h)	1	stop_lost	taG/taT

Top 2 mutated proteins for Experiment: Squamous Cell Carcinoma and Control: Non-tumour tissue were used in the algorithm of master regulator search as a list of nodes of the signal transduction network that are removed from the network during the search of master regulators (see more details about the algorithm in the Method section). 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 13-14.

*Table 13. Master regulators that may govern the regulation of **up-regulated** genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. **Total rank** is the sum of the ranks of the master molecules sorted by keynode score, CMA score, transcriptomics data.*

See full table →

ID	Master molecule name	Gene symbol	Gene description	logFC	Total rank
MO000329204	Cdk6(h):cyclinD3-isoform1(h)	CCND3, CDK6	cyclin D3, cyclin dependent kinase 6	3.09	182
MO000039099	IL-1beta-p17:IL-1RI:IL-1RAcP:MyD88:tollip:IRAK-1{pS376}{pT387}:IRAK-4:IRAK-2	AC093012.1, IL1B, IL1R1, IL1RAP, IRAK1, IRAK2, MYD88, TOLLIP	MYD88 innate immune signal transduction adaptor, interleukin 1 beta, interleukin 1 receptor accessor...	1.93	198
MO000060292	PKD3(h)	PRKD3	protein kinase D3	1.72	243
MO000018003	PP2A(h)	PPP2CA, PPP2R3A, PPP2R3B, PPP2R5A, PPP2R5B, PPP2R5C, PPP2R5D	protein phosphatase 2 catalytic subunit alpha, protein phosphatase 2 regulatory subunit B''alpha, pr...	1.93	255
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...	3.47	284
MO000161481	Cytochrome b-558:p22phox:p40phox{p}:p67phox:Rac1:GTP:JFC1:PtdIns(3,4)P2:PA:p47phox	CYBA, CYBB, NCF1, NCF2, NCF4, RAC1, SYTL1	Rac family small GTPase 1, cytochrome b-245 alpha chain, cytochrome b-245 beta chain, neutrophil cyt...	2.71	289
MO000042946	pak2(h)	PAK2	p21 (RAC1) activated kinase 2	2.36	294
MO000032352	RSK2(h)	RPS6KA3	ribosomal protein S6 kinase A3	1.56	318
MO000041170	EAC(h)	CYLD	CYLD lysine 63 deubiquitinase	1.14	318
MO000033272	SGK-1(h)	SGK1	serum/glucocorticoid regulated kinase 1	1.45	354

*Table 14. Master regulators that may govern the regulation of **down-regulated** genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. **Total rank** is the sum of the ranks of the master molecules sorted by keynode score, CMA score, transcriptomics data.*

See full table →

ID	Master molecule name	Gene symbol	Gene description	logFC	Total rank
MO000004672	ERK1(h)	MAPK3	mitogen-activated protein kinase 3	-1.85	53
MO000056883	ERK1-isoform1(h)	MAPK3	mitogen-activated protein kinase 3	-1.85	79
MO000031003	ERK1(h){p}	MAPK3	mitogen-activated protein kinase 3	-1.85	86
MO000033299	pim1(h)	PIM1	Pim-1 proto-oncogene, serine/threonine kinase	-2.6	135
MO000022222	MKP-1(h)	DUSP1	dual specificity phosphatase 1	-2.29	157
MO000041952	calpain-1(h)	CAPN1	calpain 1	-1.23	166
MO000036550	MKP-7(h)	DUSP16	dual specificity phosphatase 16	-1.71	183
MO000030911	PIAS1(h)	PIAS1	protein inhibitor of activated STAT 1	-0.77	195
MO000020073	Ubc5A(h)	UBE2D1	ubiquitin conjugating enzyme E2 D1	-0.77	196
MO000176198	JKAP(h)	DUSP22	dual specificity phosphatase 22	-0.99	199

The intracellular regulatory pathways controlled by the above-mentioned master regulators are depicted in Figures 9 and 10. 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.

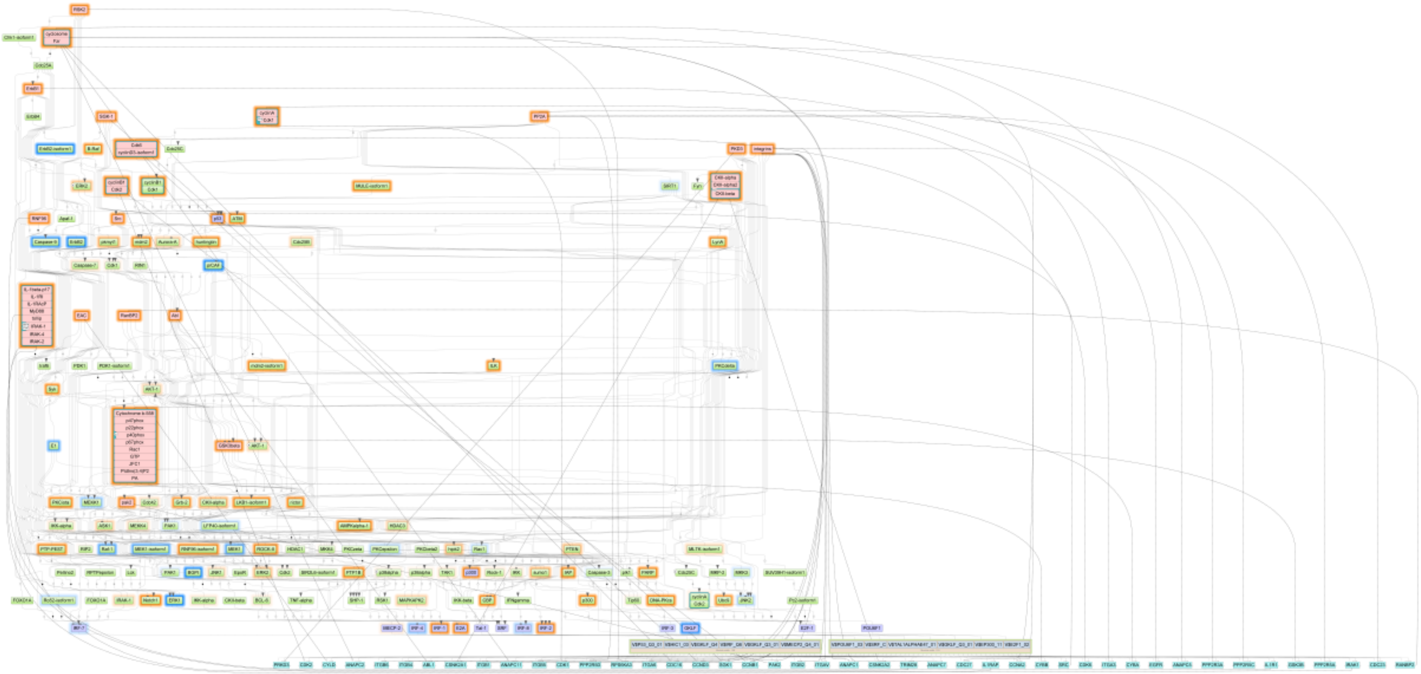


Figure 9. Diagram of intracellular regulatory signal transduction pathways of up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate molecules, 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.

See full diagram →

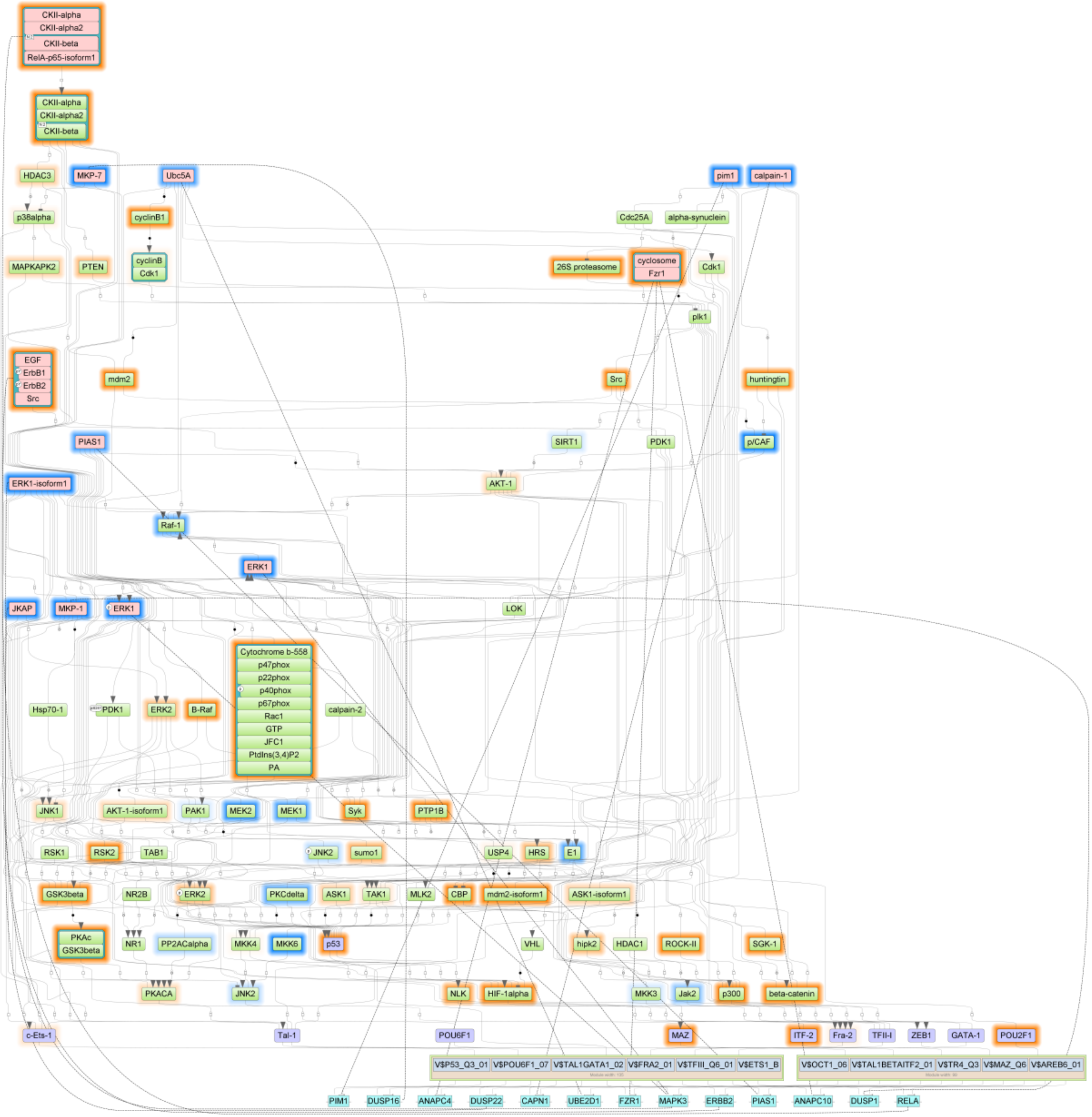


Figure 10. Diagram of intracellular regulatory signal transduction pathways of down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate 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.

See full diagram →

4. Finding prospective drug targets

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

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

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

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

Table 15. Prospective drug targets selected from full list of identified master regulators filtered by druggability score from HumanPSD™ database. ***Druggability score*** *contains the number of drugs that are potentially suitable for inhibition (or activation) of the target. The drug targets are sorted according to the* ***Total rank*** *which is the sum of three ranks computed on the basis of the three scores: keynode score, CMA score and expression change score (logFC, if present). See Methods section for details.*

See full table →

Gene symbol	Gene Description	Druggability score	logFC	Total rank
ROCK2	Rho associated coiled-coil containing protein kinase 2	2	2.61	371
SETD7	SET domain containing 7, histone lysine methyltransferase	1	1.63	496
NTRK2	neurotrophic receptor tyrosine kinase 2	1	6.48	590
PSMA7	proteasome 20S subunit alpha 7	3	1.71	696
CCND1	cyclin D1	1	3.09	715
FURIN	furin, paired basic amino acid cleaving enzyme	2	1.14	800

Table 16. Prospective drug targets selected from full list of identified master regulators filtered by druggability score predicted by PASS software. Here, the ***druggability score*** *for master regulator proteins is computed as a sum of PASS calculated probabilities to be active as a target for various small molecular compounds. The drug targets are sorted according to the* ***Total rank*** *which is the sum of three ranks computed on the basis of the three scores: keynode score, CMA score and expression change score (logFC, if present). See Methods section for details.*

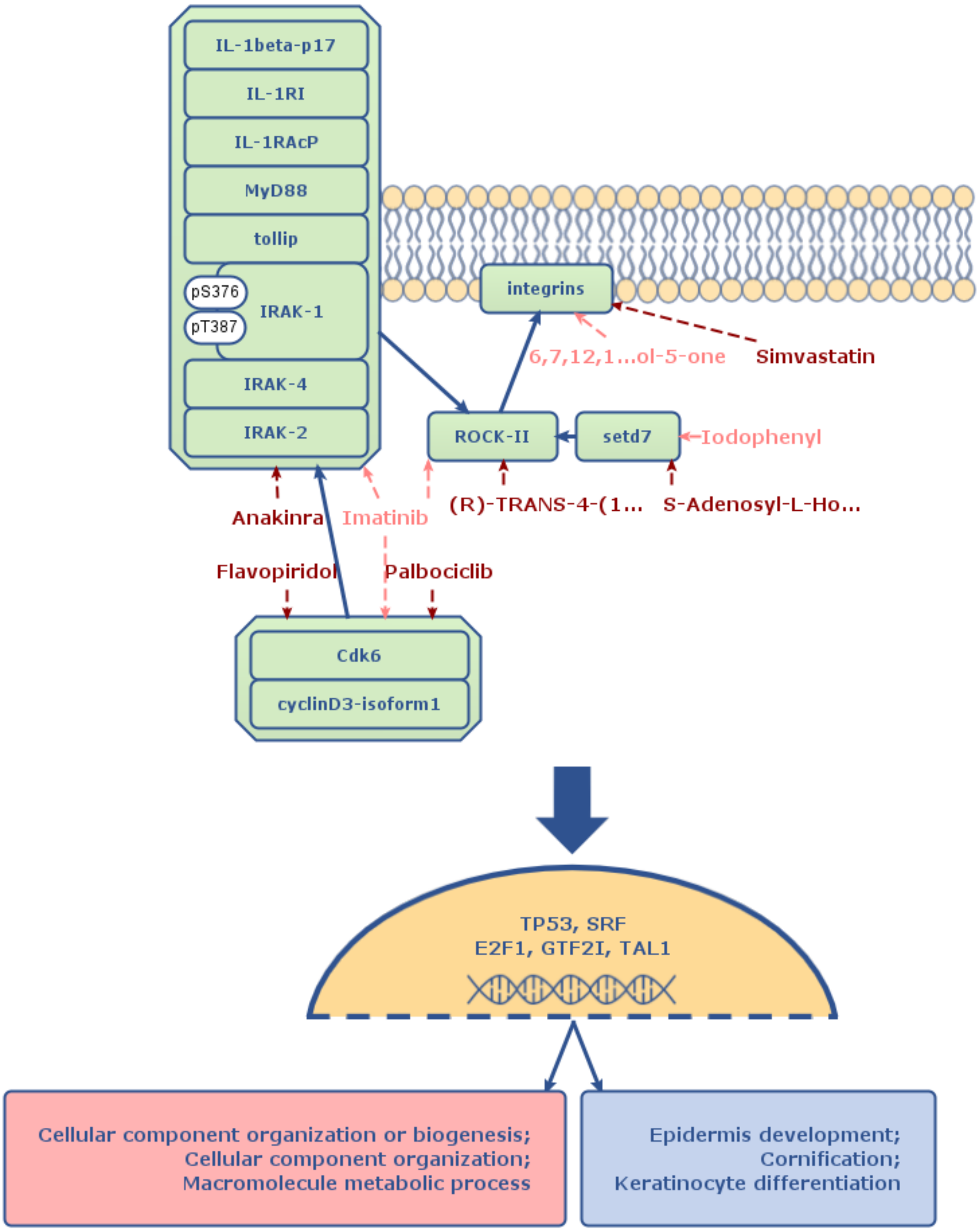
See full table →

Gene symbol	Gene Description	Druggability score	logFC	Total rank
ITGA3	integrin subunit alpha 3	96.96	3.47	284
ITGB5	integrin subunit beta 5	60.63	3.47	284
ITGA6	integrin subunit alpha 6	96.96	3.47	284
ITGB6	integrin subunit beta 6	62	3.47	284
ITGB4	integrin subunit beta 4	96.96	3.47	284
ROCK2	Rho associated coiled-coil containing protein kinase 2	59.99	2.61	371

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

- integrins
- ROCK-II
- Cdk6:cyclinD3-isoform1
- IL-1beta-p17:IL-1RI:IL-1RAcP:MyD88:tollip:IRAK-1{pS376}{pT387}:IRAK-4:IRAK-2
- setd7

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



Drugs which are shown on this schema: S-Adenosyl-L-Homocysteine, Imatinib, Flavopiridol, Simvastatin, Palbociclib, 6,7,12,13-tetrahydro-5H-indolo[2,3-a]pyrrolo[3,4-c]carbazol-5-one, (R)-TRANS-4-(1-AMINOETHYL)-N-(4-PYRIDYL) CYCLOHEXANECARBOXAMIDE, Iodophenyl and Anakinra, should be considered as a prospective research initiative for further drug repurposing and drug development. These drugs were selected as top matching treatments to the most prospective drug targets of the studied pathology, however, these results should be considered with special caution and are to be used for research purposes only, as there is not enough clinical information for adapting these results towards immediate treatment of patients.
The drugs given in dark red color on the schema are FDA approved drugs or drugs which have gone through various phases of clinical trials as active treatments against the selected targets.
The drugs given in pink color on the schema are drugs, which were cheminformatically predicted to be active against the selected targets.

5. Identification of potential drugs

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

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

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

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

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

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

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

Drugs approved in clinical trials

Table 17. FDA approved drugs or drugs used in clinical trials for the studied pathology (most promising treatment candidates selected for the identified drug targets on the basis of literature curation in HumanPSD™ database)
See full table →

Name	Target names	Drug rank	Disease activity score	Phase 4	Status (provided by Drugbank)
Dasatinib	SRC, ABL1, YES1, ABL2	14	4	Leukemia, Leukemia, Lymphoid, Leukemia, Myelogenous, Chronic, BCR-ABL Positive, Leukemia, Myeloid, Precursor Cell Lymphoblastic Leukemia-Lymphoma	small molecule,approved,investigational
Palbociclib	CDK6, CDK4	20	3	Breast Neoplasms, Neoplasms	small molecule,approved
Arsenic trioxide	CCND1, MAPK1, AKT1	34	2	Leukemia, Leukemia, Myeloid, Leukemia, Promyelocytic, Acute	small molecule,approved,investigational
Nintedanib	FGFR3, SRC, LYN	35	2	Idiopathic Pulmonary Fibrosis, Pulmonary Fibrosis	small molecule,approved
Vandetanib	VEGFA, EGFR	118	2	Neoplasms, Thyroid Neoplasms	small molecule,approved

Repurposing drugs

Table 18. Repurposed drugs used in clinical trials for other pathologies (prospective drugs against the identified drug targets on the basis of literature curation in HumanPSD™ database)
See full table →

Name	Target names	Drug rank	Phase 4	Status (provided by Drugbank)
Bosutinib	CAMK2G, SRC, ABL1, HCK, LYN, CDK2	34	Leukemia, Myeloid	small molecule,approved
Pirfenidone	FURIN	58	Acute Kidney Injury, Dermatomyositis, Idiopathic Pulmonary Fibrosis, Lung Diseases, Lung Diseases, Interstitial, Myositis, Polymyositis...	small molecule,investigational
Peginterferon alfa-2a	IFNAR1, IFNAR2	60	HIV Infections, Hemophilia A, Hepatitis, Hepatitis B, Hepatitis B, Chronic, Hepatitis C, Hepatitis C, Chronic...	biotech,approved,investigational
Peginterferon alfa-2b	IFNAR1, IFNAR2	60	Hepatitis, Hepatitis B, Hepatitis B, Chronic, Hepatitis C, Hepatitis C, Chronic, Hepatitis, Chronic	biotech,approved
Interferon beta-1a	IFNAR1, IFNAR2	60	Brain Abscess, Multiple Sclerosis, Multiple Sclerosis, Relapsing-Remitting	biotech,approved,investigational

No prospective drugs were found, which would be predicted by PASS software to be active against the identified drug targets and would be predicted to have biological activity against the studied disease(s).

Table 19. Prospective drugs, predicted by PASS software to be active against the identified drug targets, though without cheminformatically predicted activity against the studied disease(s) (drug candidates predicted with the cheminformatics tool PASS)
See full table →

Name	Target names	Drug rank	Target activity score
Imatinib	ROCK2, MAP4K4, MARK3, ABL1, NEK7, PRKACA, PAK2...	2	7.12
CEP-1347	ROCK2, MAP4K4, MARK3, NEK7, PAK2, PRKACA, GSK3B...	3	6.62
1-(5-OXO-2,3,5,9B-TETRAHYDRO-1H-PYRR...	ROCK2, MAP4K4, MARK3, ABL1, NEK7, PAK2, GSK3B...	4	6.06
(2S)-1-{4-[(4-ANILINO-5-BROMOPYRIMID...	ROCK2, MAP4K4, MARK3, ABL1, NEK7, PRKACA, PAK2...	5	5.95
(2S)-1-[4-({4-[(2,5-DICHLOROPHENYL)A...	ROCK2, MAP4K4, MARK3, ABL1, NEK7, PRKACA, PAK2...	6	5.24

As the result of drug search we propose the following drugs as most promising candidates for treating the pathology under study: Dasatinib, Bosutinib and Imatinib. These drugs were selected for acting on the following targets: SRC and ROCK2, which were predicted to be active in the molecular mechanism of the studied pathology.

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

6. Conclusion

We applied the software package "Genome Enhancer" to a data set that contains *transcriptomics* data obtained from *esophagus* tissue. The study is done in the context of *Squamous Cell Carcinoma*. The data were pre-processed, statistically analyzed and differentially expressed genes were identified. Also checked was the enrichment of GO or disease categories among the studied gene sets.

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

Dasatinib, Bosutinib and Imatinib

These drugs were selected for acting on the following targets: SRC and ROCK2, which were predicted to be involved in the molecular mechanism of the pathology under study.

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

integrins, ROCK-II, Cdk6:cyclinD3-isoform1, IL-1beta-p17:IL-1RI:IL-1RAcP:MyD88:tollip:IRAK-1{pS376}{pT387}:IRAK-4:IRAK-2 and setd7

These potential drug targets should be considered as a prospective research initiative for further drug repurposing and drug development purposes. The following drugs were predicted as, matching those drug targets: S-Adenosyl-L-Homocysteine, Imatinib, Flavopiridol, Simvastatin, Palbociclib, 6,7,12,13-tetrahydro-5H-indolo[2,3-a]pyrrolo[3,4-c]carbazol-5-one, (R)-TRANS-4-(1-AMINOETHYL)-N-(4-PYRIDYL) CYCLOHEXANECARBOXAMIDE, Iodophenyl and Anakinra. These drugs should be considered with special caution for research purposes only.

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

- integrins
- ROCK-II
- Cdk6:cyclinD3-isoform1
- IL-1beta-p17:IL-1RI:IL-1RAcP:MyD88:tollip:IRAK-1{pS376}{pT387}:IRAK-4:IRAK-2
- setd7

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

7. Methods

Databases used in the study

Transcription factor binding sites in promoters and enhancers of differentially expressed genes were analyzed using known DNA-binding motifs described in the TRANSFAC® library, release 2020.2 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transfac).
The master regulator search uses the TRANSPATH® database (BIOBASE), release 2020.2 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transpath). A comprehensive signal transduction network of human cells is built by the software on the basis of reactions annotated in TRANSPATH®.
The information about drugs corresponding to identified drug targets and clinical trials references were extracted from HumanPSD™ database, release 2020.2 (https://genexplain.com/humanpsd).
The Ensembl database release Human99.38 (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.

Genomic data processing

When analyzing a list of genomic variations (from vcf file or computed by Genome Enhancer from fastq files), first of all, we compute a specific mutation weight (w) for each variation depending on it's location in gene body and gene flanking regions (-1000 upstream and +1000 downstream of the gene body).

> w = 0.7 for variations in exon area
> w = 1.3 for variations in promoter region (-1000bp upstream and 100bp downstream of TSS),
> w = 1.0 for variations in other locations.

Total Gene mutation weight is the sum of the weights w of all variations located inside the gene body and in the gene flanking regions.
Next, a weighted score is calculated for all genes with the following formula:
Weighted score = In_disease * In_transpath * Gene mutation weight, where

> In_disease = 1.5 for genes assigned to selected diseases,
> In_transpath = 2.0 for genes mapped to Transpath pathways,
> and In_disease = In_transpath = 1.0 in all other cases.

At the next step, 300 genes with highest weighted score are selected for further CMA model search.
The mutation weights (w) are also used to find the regulatory regions of the genes most affected by the variations. A sliding window of 1100 bp is used to scan through the intronic, 5' and 3' regions of the genes and a region is selected with the highest sum of the mutation weights.

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

Transcription factor binding sites in promoters and enhancers of differentially expressed genes were analyzed using known DNA-binding motifs. The motifs are specified using position weight matrices (PWMs) that give weights to each nucleotide in each position of the DNA binding motif for a transcription factor or a group of them.
We search for transcription factor binding sites (TFBS) that are enriched in the promoters and enhancers under study as compared to a background sequence set such as promoters of genes that were not differentially regulated under the condition of the experiment. We denote study and background sets briefly as Yes and No sets. In the current work we used a workflow considering promoter sequences of a standard length of 1100 bp (-1000 to +100). The error rate in this part of the pipeline is controlled by estimating the adjusted p-value (using the Benjamini-Hochberg procedure) in comparison to the TFBS frequency found in randomly selected regions of the human genome (adj.p-value < 0.01).
We have applied the CMA algorithm (Composite Module Analyst) for searching composite modules [7] in the promoters and enhancers of the Yes and No sets. We searched for a composite module consisting of a cluster of 10 TFs in a sliding window of 200-300 bp that statistically significantly separates sequences in the Yes and No sets (minimizing Wilcoxon p-value).

Methods for finding master regulators in networks

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

Methods for analysis of pharmaceutical compounds

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

Method for analysis of known pharmaceutical compounds

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

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

"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" ($D\text{-}score_{PSD}$):

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

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:

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

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

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

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

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

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

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

8. References

1. Kel A, Voss N, Jauregui R, Kel-Margoulis O, Wingender E. Beyond microarrays: Finding key transcription factors controlling signal transduction pathways. *BMC Bioinformatics.* **2006**;7(S2), S13. doi:10.1186/1471-2105-7-s2-s13
2. Stegmaier P, Voss N, Meier T, Kel A, Wingender E, Borlak J. Advanced Computational Biology Methods Identify Molecular Switches for Malignancy in an EGF Mouse Model of Liver Cancer. *PLoS ONE.* **2011**;6(3):e17738. doi:10.1371/journal.pone.0017738
3. Koschmann J, Bhar A, Stegmaier P, Kel A, Wingender E. "Upstream Analysis": An Integrated Promoter-Pathway Analysis Approach to Causal Interpretation of Microarray Data. *Microarrays.* **2015**;4(2):270-286. doi:10.3390/microarrays4020270.
4. Kel A, Stegmaier P, Valeev T, Koschmann J, Poroikov V, Kel-Margoulis OV, and Wingender E. Multi-omics "upstream analysis" of regulatory genomic regions helps identifying targets against methotrexate resistance of colon cancer. *EuPA Open Proteom.* **2016**;13:1-13. doi:10.1016/j.euprot.2016.09.002
5. Michael H, Hogan J, Kel A et al. Building a knowledge base for systems pathology. *Brief Bioinformatics.* **2008**;9(6):518-531. doi:10.1093/bib/bbn038
6. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E. TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res.* **2006**;34(90001):D108-D110. doi:10.1093/nar/gkj143
7. Kel AE, Gössling E, Reuter I, Cheremushkin E, Kel-Margoulis OV, Wingender E. MATCH: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* **2003**;31(13):3576-3579. doi:10.1093/nar/gkg585
8. Waleev T, Shtokalo D, Konovalova T, Voss N, Cheremushkin E, Stegmaier P, Kel-Margoulis O, Wingender E, Kel A. Composite Module Analyst: identification of transcription factor binding site combinations using genetic algorithm. *Nucleic Acids Res.* **2006**;34(Web Server issue):W541-5.
9. Krull M, Pistor S, Voss N, Kel A, Reuter I, Kronenberg D, Michael H, Schwarzer K, Potapov A, Choi C, Kel-Margoulis O, Wingender E. TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations. *Nucleic Acids Res.* **2006**;34(90001):D546-D551. doi:10.1093/nar/gkj107
0. Boyarskikh U, Pintus S, Mandrik N, Stelmashenko D, Kiselev I, Evshin I, Sharipov R, Stegmaier P, Kolpakov F, Filipenko M, Kel A. Computational master-regulator search reveals mTOR and PI3K pathways responsible for low sensitivity of NCI-H292 and A427 lung cancer cell lines to cytotoxic action of p53 activator Nutlin-3. *BMC Med Genomics.* **2018**;11(1):12. doi:10.1186/1471-2105-7-s2-s13
1. Filimonov D, Poroikov V. Probabilistic Approaches in Activity Prediction. Varnek A, Tropsha A. *Cheminformatics Approaches to Virtual Screening.* Cambridge (UK): RSC Publishing. **2008**;:182-216.
2. Filimonov DA, Poroikov VV. Prognosis of specters of biological activity of organic molecules. *Russian chemical journal.* **2006**;50(2):66-75 (russ)
3. Filimonov D, Poroikov V, Borodina Y, Gloriozova T. Chemical Similarity Assessment Through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *ChemInform.* **1999**;39(4):666-670. doi:10.1002/chin.199940210

Thank you for using the Genome Enhancer!

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

Supplementary material

1. Supplementary table 1 - Up-regulated genes
2. Supplementary table 2 - Down-regulated genes
3. Supplementary table 3 - Detailed report. Composite modules and master regulators (up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue).
4. Supplementary table 4 - Detailed report. Composite modules and master regulators (down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue).

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.