

NTRK2 and ITGA3 are promising druggable targets for treating squamous cell carcinoma that control activity of TP53, E2F1 and ETS1 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 17/02/2020 ; Report generated on 17/02/2020

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



Abstract

In the present study we applied the software package "Genome Enhancer" to a 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) 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: TP53, E2F1, TAL1, ETS1, RXRA and RARA. The subsequent network analysis suggested NTRK2, ROCK2, PSMA7, ITGA3, ITGB5 and ITGA6 as the most promising and druggable molecular targets. Finally, the following drugs were identified as the most promising treatment candidates: 5-(1,4-DIAZEPAN-1-SULFONYL)ISOQUINOLINE, 2-ACETYLAMINO-4-METHYL-PENTANOIC ACID [1-(1-FORMYL-PENTYL-CARBAMOYL)-3-METHYL-BUTYL]-AMIDE, Amitriptyline, Propylthiouracil, L-2-Amino-6-Methylene-Pimelic Acid and L-2-Amino-4-(Guanidinoxy)Butyric Acid.

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 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 [5]. In addition, new potential small molecular ligands are subsequently predicted for the revealed targets. A general druggability check is performed using a precomputed database of 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 [11-13].

2. Data

For this study the following experimental data was used:

Table 1. Experimental datasets used in the study

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

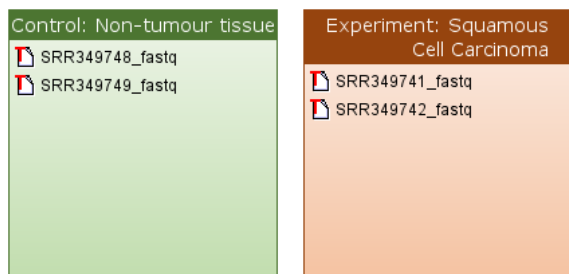


Figure 1. Annotation diagram of experimental data used in this study. With the colored boxes we show those sub-categories of the data that are compared in our analysis.

3. Results

We have compared the following conditions: 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 3. Top ten significant **down-regulated** genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue.

[See full table](#) →

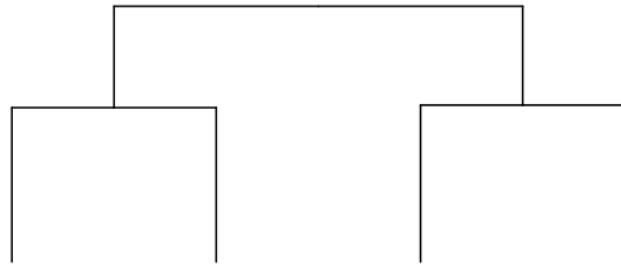
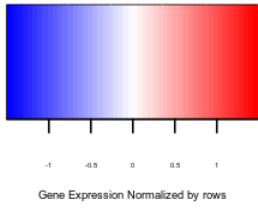
ID	Gene symbol	Gene description	logFC	logCPM	PValue	FDR
ENSG00000136155	SCEL	scellin	-7.36	10.74	2.01E-12	1.76E-8
ENSG00000163209	SPRR3	small proline rich protein 3	-6.39	14.08	2.27E-5	2E-3
ENSG00000143369	ECM1	extracellular matrix protein 1	-6.04	10.66	2.28E-9	1.82E-6
ENSG00000189334	S100A14	S100 calcium binding protein A14	-6	10.05	7.93E-10	6.95E-7
ENSG00000229732	AC019349.5		-5.88	12.56	3.53E-9	2.03E-6
ENSG00000086548	CEACAM6	carcinoembryonic antigen related 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 protease, serine 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. Functional classification of genes

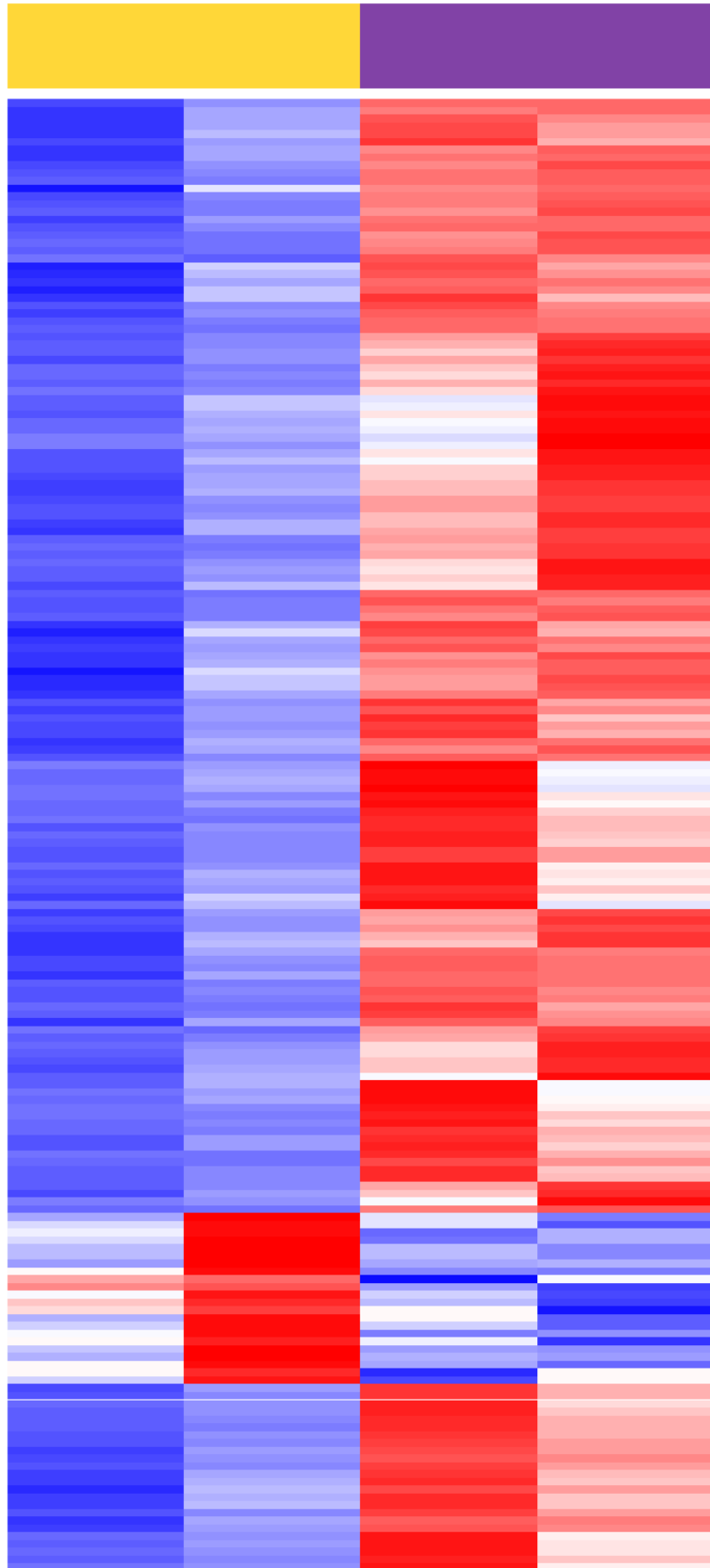
A functional analysis of differentially expressed genes was done by mapping the significant up-regulated and significant down-regulated genes to several known ontologies, such as Gene Ontology (GO), disease ontology (based on HumanPSD™ database) and the ontology of signal transduction and metabolic pathways 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

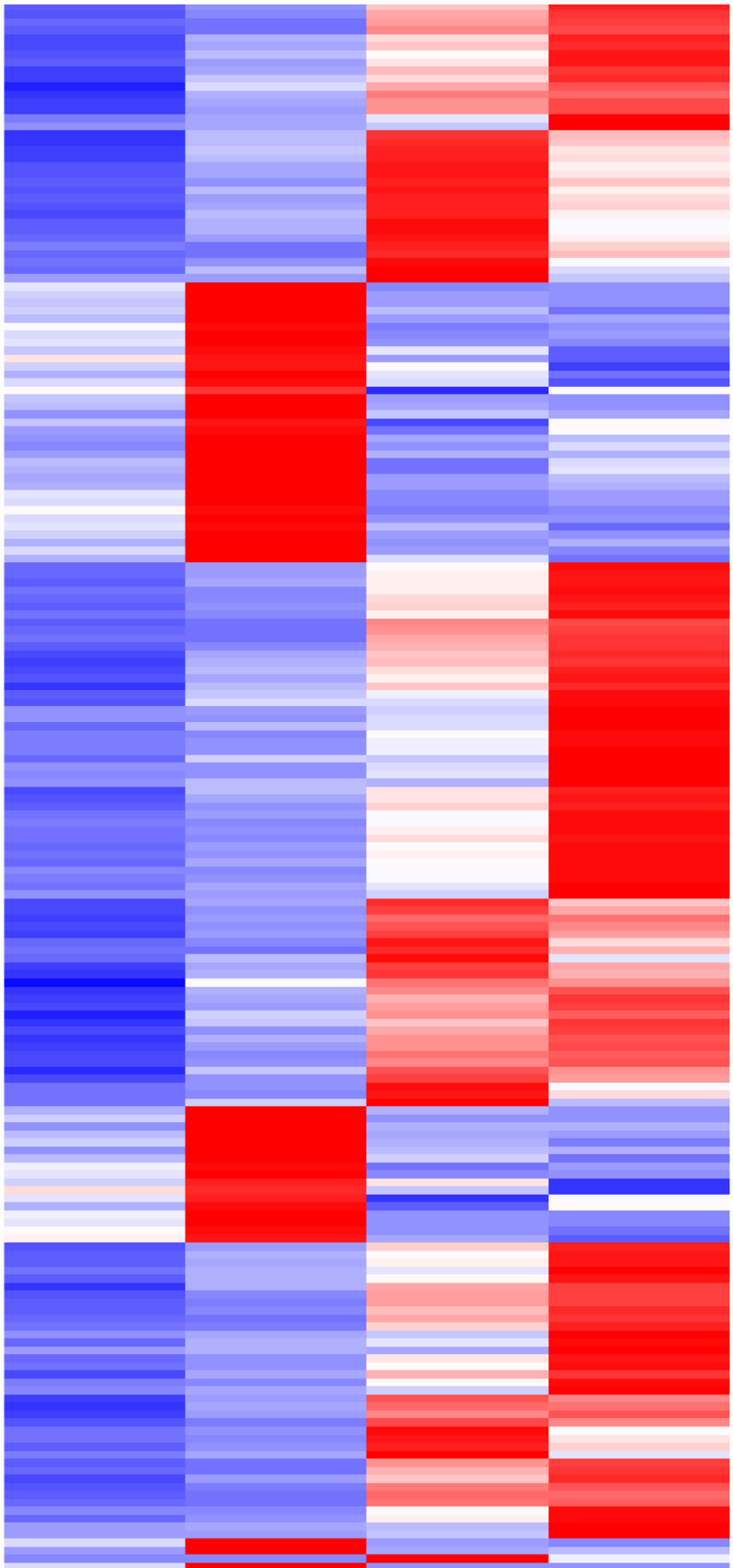
A heatmap of all differentially expressed genes playing a potential regulatory role in the system (enriched in [TRANSPATH®](#) pathways) is presented in Figure 2.



Control: Non-tumour tissue
 Experiment: Squamous Cell Carcinoma



- CSNK3A2
- USP1
- PRM2
- LINC
- GPR11
- CRK
- ATP2B
- TBR1
- PLP3
- MASP1K11
- NAMA1
- PGM3
- CDNF
- BRCS
- SOX3
- KIDINS220
- PRDM4
- ITPR1
- CEBPB
- GPRD
- IFI1
- MARLU2
- MORF2
- USP21
- PTPRC
- MGCC1
- PAN2
- HEX
- SMAD1
- NEO2
- HST1H4
- CCND1
- LUCIF4
- YORW4
- BLB1
- HST1H4H
- HST1H4L
- EDTRD
- LYN
- STN4
- ADSS
- TSF1
- DIAPH3
- TRIO
- HST1H3A
- MAPP1
- PRM2
- GDR
- PLD1
- ASHL
- PRM1
- PTPRN2
- PCYT1A
- STAT2
- LATS1
- BRUCE
- LPCAT2
- CTSD
- KPNA2
- AP3B1
- CYBP3
- EPN1
- BLB3
- UKDH
- PRKCH2
- DAL3
- PLD3
- POCB
- SMAD3
- MATH3
- GSK3B
- CLL1
- PTK2
- TRPM2
- CDCC27
- MDC1
- PLP3
- CHD4
- DDX51
- PLRN1
- SMO1
- PANCS
- BLP1
- HEPK6A1
- ATR
- ACACA
- PHK2GA
- HTT
- ATM
- ITPR2
- KITN1
- EDB1
- TMEM
- CASK
- PRK4
- ACLY
- PRF23A
- GMP5
- PRK6
- POLR2B
- TAK1
- LUC7B
- SPSM
- NLRP6
- PPP2R3D
- ASPM1
- SALNF2
- ADSL
- DTN1A
- TRK2
- PRKAB2
- APK2Z1
- APK2Z2
- CTNNA3
- CHRT1
- ABL1
- AKK
- PRK1
- NCK1
- PSAT1
- PRK2D
- CASP2
- CDCC3
- TRK1
- PTPRB
- MED17
- TRP3
- SH2D1
- HR1
- DDK
- MTOR
- WDR5
- MARPLAP3
- KLC1
- ZFP94
- NBE2
- PRF1
- MPRL1
- GRAN1
- POLR2J
- COX8B
- SH2L1
- MTK
- ACAA1
- CTNNA
- TAM1
- MECOM
- TRK
- AKK
- DAL2
- DUSP16
- SH2D3
- KSR1
- AKK
- HST1H4H
- ZFP74
- PTSD
- SCAP3
- PRAD3A
- IL13
- DIAT1
- TSG1
- ALDH3A1
- SHANK2
- REV3L
- POLR2A
- CSF1R
- SMD
- DAL2
- GRN2
- APK2AP3
- CD52
- HES6
- CNEBP
- MED1
- TRF3
- TUBGCP4
- ILK
- MORF1
- CAD
- GASPR1
- GBA1
- UNC119
- PRK2D
- TUBA1A
- NCOR1
- GRB2



- CCNA2
- PP5P1
- COM2
- GLB1
- ESM2T
- HSPF1
- HFA
- HEPT1
- SPIC
- HNG21
- DUT
- GA2
- SAE1
- PRKX2
- NA1
- HST19A
- NECD4L
- PRKNC2B
- KODR
- ELP2
- NECA
- SPAG9
- PCN1
- UBC2K
- BGAL2
- ALBA
- NEO3
- CEL
- PP59A3
- HAT2A
- HQHR
- CF3
- SGH1
- SETD7
- CHP
- GALE
- ATPKAP10
- TGAM1
- PRAG27A
- PNM1
- IMP1
- ITFC
- ESPL1
- DCAT2
- DUXK1
- CKK
- PLD2
- ELDL2B
- GF2
- MAPK3
- PLD3
- PNP2
- SLTB1
- CYP2E1
- PEL1
- UBC2B
- CCNT
- ELVNS
- GATM
- DCAL
- NFSC2
- AGC2B
- CLTB
- GRD5
- SESN2
- POB2
- ALPHA3
- NAK
- PRK3
- SAAP2
- LSA11
- PRKCA
- ELVLS
- TNAP
- PAK2
- GRK4
- PNK1
- EPH2C
- EPH4
- NET
- ATP2B4
- UBA2
- IMPAD1
- PCNA
- CRN1
- LPA
- PRK21
- PRKAA1
- HST19C
- GLM
- APPC1A
- HST19B
- HST19H
- DIS3H
- RYR3
- HST19U
- CHML
- MA2SYN1
- STAT1
- KPNA1
- LRWD1
- HST19K
- HST19H
- FZD6
- HST19D
- GLC
- ADAM17
- CDH1
- HST19G
- APR
- IL1RAP
- HST19P
- ITGB2
- MAP3K2
- NR1
- YOMG
- PRKPH2
- PRKDC
- PRK
- CDH1
- JAG1
- SMAD2
- HSP90A1
- YKL
- CALM2
- YORAC
- PSMA7
- NDCA3
- GR1
- PRK3
- SPC1
- CDK6
- TSPY1
- PRKAR1A
- BSIP2
- ITGB2
- DYNCH1
- TRAF2
- CRN1
- LRWD1
- ELVLS
- PRKX1
- ALPHA1
- LRP1
- GRD5
- SCN1A
- KAT5
- EPH2
- NSL1
- SCN1B
- EPH2
- SORT1
- EPH2
- PTG2
- GFY3
- GRK1
- WRSA
- TRKB
- TRAF1
- SELENK
- MOXA
- MSN
- PTT1
- CDR5
- PRK2
- LAMA1
- CTTN
- CTP-PT1
- CDR1
- TRAF2
- HST19B
- NSL1
- TNAP1
- SGR1
- SPK2
- PRKDC
- ITGB1
- ANK2
- MMP14
- HFA
- ALPHA1
- ITGAV
- ANK2
- ND1
- GFY2
- ANK2
- PRK
- NTRC2
- AP
- DKOR1
- EGFR
- DLG1
- POS
- OCI
- 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 (2020.1)

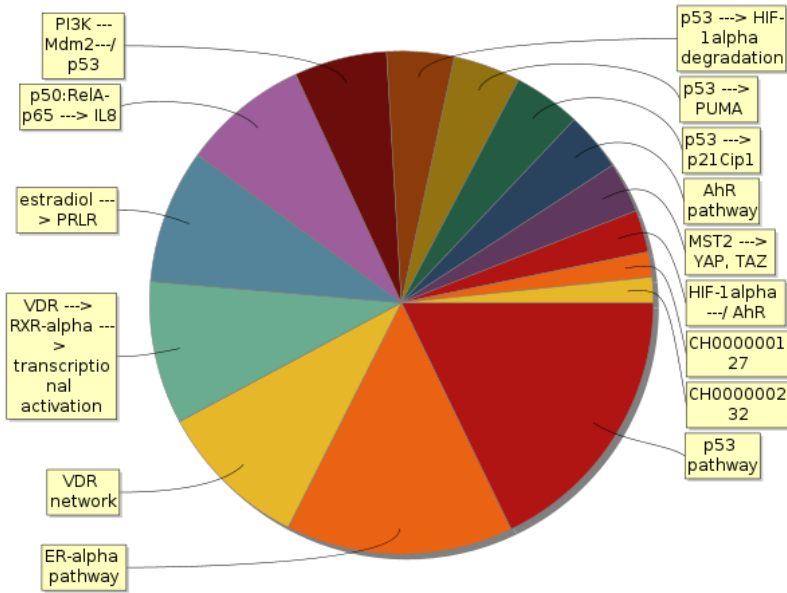


Figure 4. Enriched TRANSPATH® Pathways (2020.1) of up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. [Full classification →](#)

HumanPSD(TM) disease (2020.1)

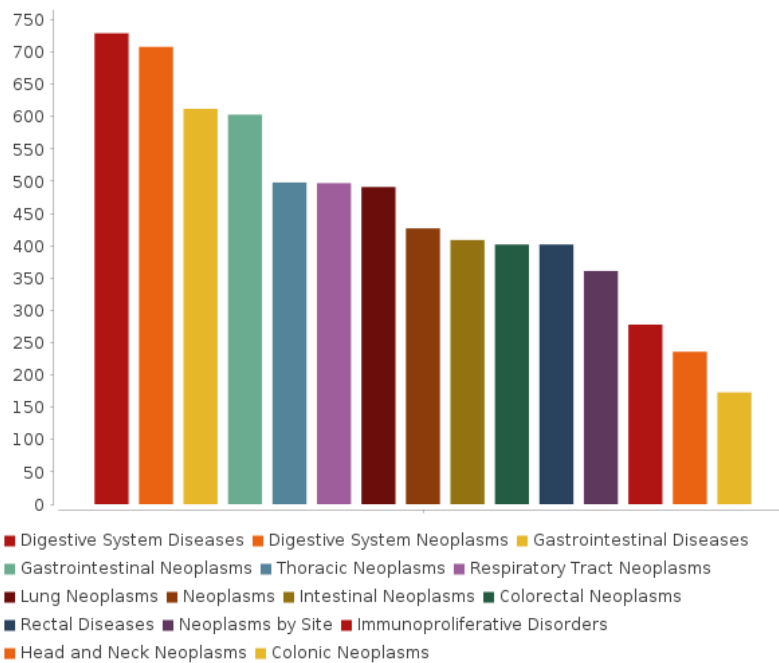


Figure 5. Enriched HumanPSD(TM) disease (2020.1) of up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. The size of the bars correspond to the number of bio-markers of the given disease found among the input set. [Full classification →](#)

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)

biological_process Gene Ontology treemap



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

TRANSPATH® Pathways (2020.1)

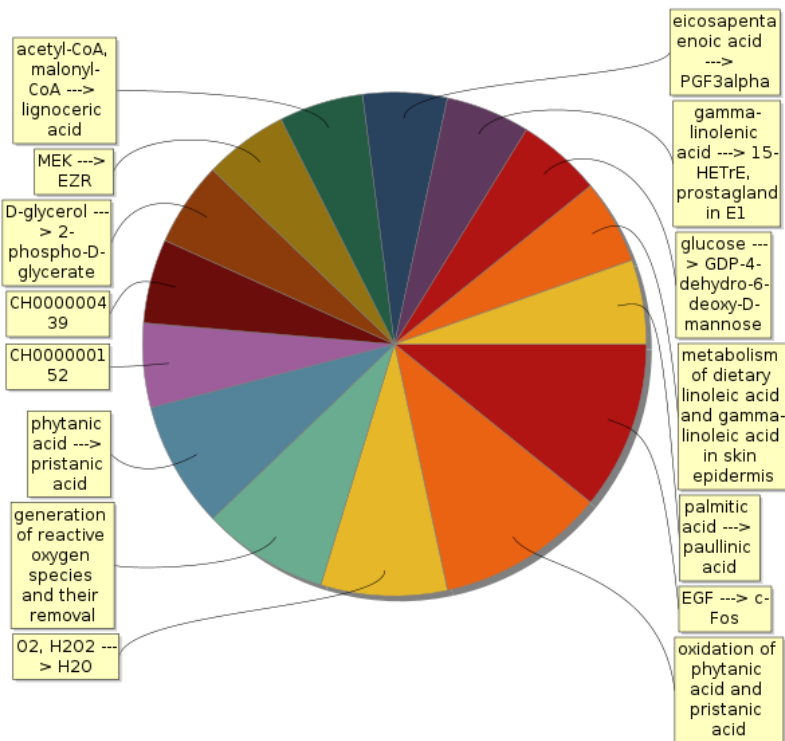


Figure 7. Enriched TRANSPATH® Pathways (2020.1) of down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. Full classification →

HumanPSD(TM) disease (2020.1)

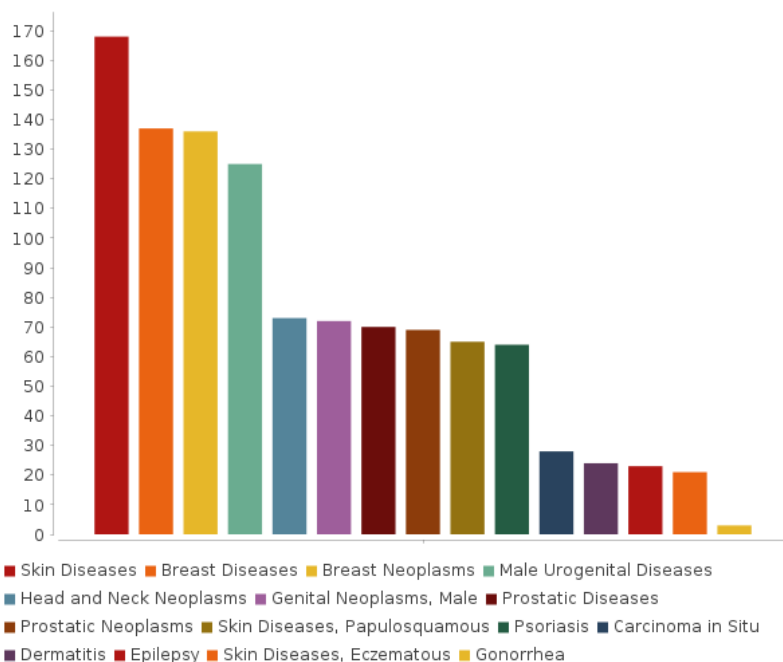


Figure 8. Enriched HumanPSD(TM) disease (2020.1) of down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

[Full classification](#) →

3.3. 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 analysed mutations that were revealed in the potential enhancers located upstream, downstream or inside the **target genes** (see Table 4). We identified 1119 mutations potentially affecting gene regulation. Table 5 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 4. Mutations revealed in genes in Experiment: Squamous Cell Carcinoma and Control: Non-tumour tissue

[See full table](#) →

ID	Gene symbol	Gene schematic representation	Number of variations
ENSG00000146648	EGFR		33
ENSG00000134871	COL4A2		25
ENSG00000083857	FAT1		23
ENSG00000186340	THBS2		19
ENSG00000226445	XXyac-YX65C7_A.2		18
ENSG00000114999	TTL		16
ENSG00000152291	TGOLN2		14
ENSG00000113140	SPARC		12
ENSG00000134247	PTGFRN		12
ENSG00000142173	COL6A2		12

Table 5. 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\$BBX_04	4.66E-2	1.97E-3	4	43
V\$POU6F1_02	3.8E-2	1.54E-3	6	14
V\$MAFA_Q4	2.84E-2	1.17E-3	72	55
V\$DUXL_01	2.74E-2	4.71E-4	2	7
V\$REST_Q5	1.72E-2	7.36E-4	213	14
V\$ZBRK1_01	1.12E-2	1.04E-3	10	17
V\$ARID5A_03	6.08E-3	1.54E-3	6	20
V\$RFX_Q6	3.25E-3	3.34E-2	17	136
V\$HES1_Q6	3.07E-3		7	null
V\$SIX1_01	2.62E-3		4	null
V\$OSR1_03	2.57E-3	4.1E-3	143	1
V\$NKX25_Q6	2.48E-3	1.56E-2	30	126
V\$ZBTB12_03	2.06E-3	1.84E-3	84	42
V\$MRF2_01	1.54E-3	4.3E-2	64	26
V\$RNF96_01	1.22E-3	5.58E-3	4	38
V\$LRH1_Q5_01	9.67E-4	3.55E-2	3	27
V\$FREAC3_01	5.72E-4	1.54E-2	16	3
V\$MEIS1_01		2.65E-4	null	12
V\$TBX5_Q2		3.95E-4	null	1

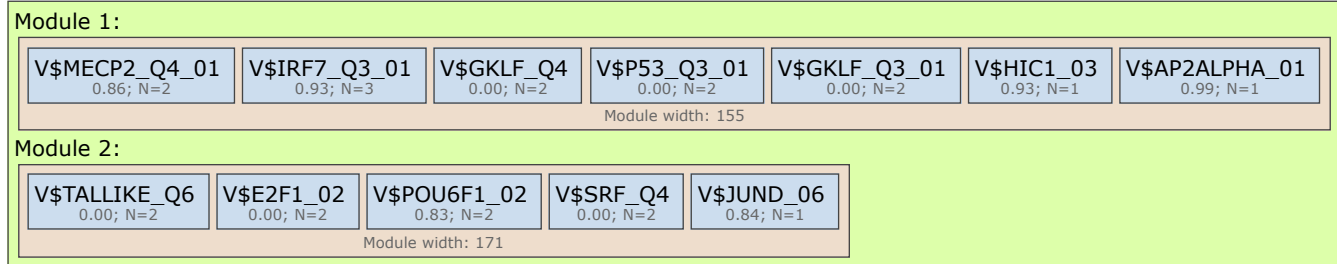
We applied the Composite Module Analyst (CMA) [8] method to detect such potential enhancers, as targets of multiple TFs bound in a cooperative manner to the regulatory regions of the genes of interest. CMA applies a genetic algorithm to construct a generalized model of the enhancers by specifying combinations 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.



Model score (-p*log10(pval)): 15.04

Wilcoxon p-value (pval): 2.06e-32

Penalty (p): 0.475

Average yes-set score: 7.32

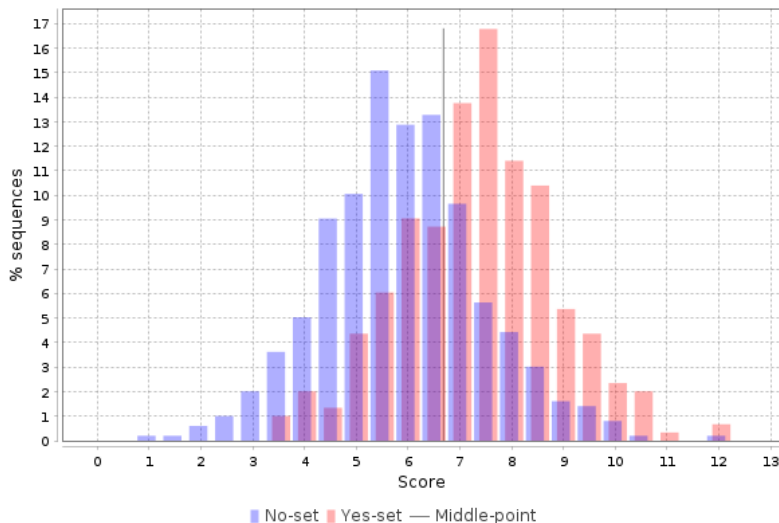
Average no-set score: 5.92

AUC: 0.75

Middle-point: 6.68

False-positive: 28.17%

False-negative: 30.54%



[See model visualization table →](#)

Table 6. 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
ENSG00000108021	FAM208B	family with sequence similarity 208 member B	15.67	E2F-1(h), SRF(h), POU6F1(h), JunD(h), HEN2(h), Lyl-1(h), Tal-1(h), AP-2alpha(h), IRF-7(h)...
ENSG00000067066	SP100	SP100 nuclear antigen	15.11	IRF-7(h), E2F-1(h), AP-2alpha(h), SRF(h), p53(h), GKLF(h), MECP-2(h)...
ENSG00000198380	GFPT1	glutamine--fructose-6-phosphate transaminase 1	14.95	GKLF(h), IRF-7(h), p53(h), HIC-1(h), AP-2alpha(h), MECP-2(h), POU6F1(h)...
ENSG00000182326	C1S	complement C1s	14.84	POU6F1(h), E2F-1(h), SRF(h), HEN2(h), Lyl-1(h), Tal-1(h), JunD(h), IRF-7(h), AP-2alpha(h)...
ENSG00000159335	PTMS	parathyrosin	14.73	SRF(h), E2F-1(h), POU6F1(h), HEN2(h), Lyl-1(h), Tal-1(h), JunD(h), GKLF(h), IRF-7(h)...
ENSG00000163714	U2SURP	U2 snRNP associated SURP domain containing	14.67	HEN2(h), Lyl-1(h), Tal-1(h), JunD(h), SRF(h), E2F-1(h), POU6F1(h), MECP-2(h), GKLF(h)...
ENSG0000010270	STARD3NL	STARD3 N-terminal like	14.52	POU6F1(h), SRF(h), E2F-1(h), HEN2(h), Lyl-1(h), Tal-1(h), JunD(h), MECP-2(h), IRF-7(h)...
ENSG00000140854	KATNB1	katanin regulatory subunit B1	14.46	MECP-2(h), HEN2(h), Lyl-1(h), Tal-1(h), POU6F1(h), SRF(h), GKLF(h), E2F-1(h), p53(h)...
ENSG00000002834	LASP1	LIM and SH3 protein 1	14.44	MECP-2(h), GKLF(h), POU6F1(h), IRF-7(h), E2F-1(h), p53(h), JunD(h)...
ENSG00000143622	RIT1	Ras like without CAAX 1	14.4	AP-2alpha(h), GKLF(h), SRF(h), MECP-2(h), p53(h), IRF-7(h), JunD(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\$SNAI2_01 0.00; N=3	V\$HMG1Y_01 0.96; N=3	V\$TR4_Q3 0.00; N=3	V\$ETS1_B 0.00; N=3	V\$FOSJUN_02 0.87; N=3
Module width: 113				

Module 2:

V\$TFIII_Q6_01 0.91; N=3	V\$POU6F1_02 0.80; N=3	V\$RAD21_10 0.00; N=3	V\$NR1NR2_Q3 0.00; N=3	V\$POU6F1_07 0.84; N=3
Module width: 126				

Model score (-p*log10(pval)): 15.23

Wilcoxon p-value (pval): 4.10e-31

Penalty (p): 0.501

Average yes-set score: 10.91

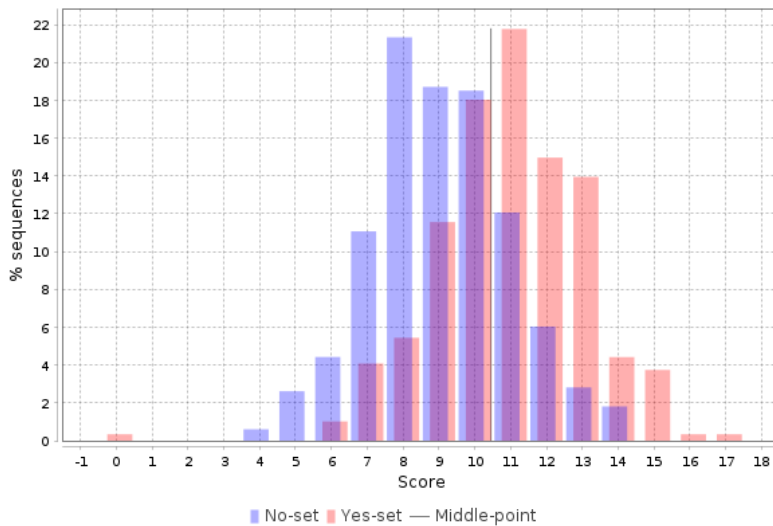
Average no-set score: 9.12

AUC: 0.75

Middle-point: 10.45

False-positive: 23.54%

False-negative: 38.10%



[See model visualization table](#) →

Table 7. 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
ENSG00000166289	PLEKHF1	pleckstrin homology and FYVE domain containing 1	18.48	slug(h), TR4(h), c-Ets-1(h), HMG1Y(h), TFII-I(h), CAR(h), COUP-TF1(h), COUP-TF2(h), LXR-alpha(h), LXR-beta(h), NR1B1(h), NR1B2(h), PXR(h), RAR-gamma(h), RXR-alpha(h), RXR-beta(h), POU6F1(h)...
ENSG00000130703	OSBPL2	oxysterol binding protein like 2	17.33	CAR(h), COUP-TF1(h), COUP-TF2(h), LXR-alpha(h), LXR-beta(h), NR1B1(h), NR1B2(h), PXR(h), RAR-gamma(h), RXR-alpha(h), RXR-beta(h), Rad21(h), TFII-I(h), POU6F1(h), c-Ets-1(h), slug(h), TR4(h)
ENSG00000132109	TRIM21	tripartite motif containing 21	17.17	Rad21(h), POU6F1(h), HMG1Y(h), c-Ets-1(h), TFII-I(h), slug(h), TR4(h)...
ENSG00000254003	CTB-167B5.1		17.16	POU6F1(h), HMG1Y(h), TFII-I(h), c-Ets-1(h), slug(h), Rad21(h), TR4(h)...
ENSG00000227543	SPAG5-AS1	SPAG5 antisense RNA 1	17.06	POU6F1(h), HMG1Y(h), c-Ets-1(h), slug(h), TR4(h), CAR(h), COUP-TF1(h), COUP-TF2(h), LXR-alpha(h), LXR-beta(h), NR1B1(h), NR1B2(h), PXR(h), RAR-gamma(h), RXR-alpha(h), RXR-beta(h), TFII-I(h)...
ENSG00000163435	ELF3	E74 like ETS transcription factor 3	16.69	slug(h), CAR(h), COUP-TF1(h), COUP-TF2(h), LXR-alpha(h), LXR-beta(h), NR1B1(h), NR1B2(h), PXR(h), RAR-gamma(h), RXR-alpha(h), RXR-beta(h), TR4(h), c-Ets-1(h), HMG1Y(h), TFII-I(h), POU6F1(h)...
ENSG00000169469	SPRR1B	small proline rich protein 1B	16.68	TFII-I(h), POU6F1(h), Rad21(h), CAR(h), COUP-TF1(h), COUP-TF2(h), LXR-alpha(h), LXR-beta(h), NR1B1(h), NR1B2(h), PXR(h), RAR-gamma(h), RXR-alpha(h), RXR-beta(h), TR4(h), HMG1Y(h), c-Ets-1(h)...
ENSG00000188505	NCCRP1	non-specific cytotoxic cell receptor protein 1 homolog (zebrafish)	16.41	slug(h), c-Ets-1(h), CAR(h), COUP-TF1(h), COUP-TF2(h), LXR-alpha(h), LXR-beta(h), NR1B1(h), NR1B2(h), PXR(h), RAR-gamma(h), RXR-alpha(h), RXR-beta(h), TR4(h), Rad21(h), POU6F1(h), TFII-I(h)
ENSG00000130649	CYP2E1	cytochrome P450 family 2 subfamily E member 1	16.39	HMG1Y(h), c-Ets-1(h), TR4(h), slug(h), TFII-I(h), Rad21(h), CAR(h), COUP-TF1(h), COUP-TF2(h), LXR-alpha(h), LXR-beta(h), NR1B1(h), NR1B2(h), PXR(h), RAR-gamma(h), RXR-alpha(h), RXR-beta(h)...
ENSG00000196754	S100A2	S100 calcium binding protein A2	16.39	slug(h), CAR(h), COUP-TF1(h), COUP-TF2(h), LXR-alpha(h), LXR-beta(h), NR1B1(h), NR1B2(h), PXR(h), RAR-gamma(h), RXR-alpha(h), RXR-beta(h), POU6F1(h), TFII-I(h), Rad21(h), c-Ets-1(h), TR4(h)...

On the basis of the enhancer models we identified the following transcription factors potentially regulating the **target genes** of our interest. We found 13 and 18 transcription factors controlling expression of up- and down-regulated genes respectively (see Tables 8-9).

Table 8. 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	7.97	1.22
MO000004274	E2F1	E2F transcription factor 1	5.7	1.2
MO000032489	TAL1	TAL bHLH transcription factor 1, erythroid differentiation factor	5.28	1.39
MO000013015	SRF	serum response factor	5.06	4.18
MO000007703	IRF7	interferon regulatory factor 7	5.04	1.36
MO000125561	KLF4	Kruppel like factor 4	4.59	1.3
MO000007834	JUND	JunD proto-oncogene, AP-1 transcription factor subunit	4.53	3.34
MO000028711	MECP2	methyl-CpG binding protein 2	4.08	1.16
MO000001275	TFAP2A	transcription factor AP-2 alpha	3.97	2.64
MO000028320	null	null	3.8	1.29

Table 9. 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
MO000059013	ETS1	ETS proto-oncogene 1, transcription factor	2.09	1.25
MO000019619	RXRA	retinoid X receptor alpha	1.94	1.9
MO000033904	RARA	retinoic acid receptor alpha	1.7	1.9
MO000019622	GTF2I	general transcription factor Iii	1.69	1.28
MO000028320	null	null	1.62	1.34
MO000028767	SNAI2	snail family transcriptional repressor 2	1.51	1.46
MO000042938	RAD21	RAD21 cohesin complex component	1.45	1.22
MO000026358	HMGA1	high mobility group AT-hook 1	1.35	1.13
MO000019618	RARB	retinoic acid receptor beta	1.27	1.9
MO000025593	RXRB	retinoid X receptor beta	1.25	1.9

3.4. Finding master regulators in networks

In the second step of the upstream analysis common regulators of the revealed TFs were identified. We identified 14 signaling proteins whose structure and function is highly damaged by the mutations (see Table 10).

Table 10. 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
MO000068933	HLA-G(h)	6	NMD_transcript_variant,splice_region_variant,stop_lost	Tga/Aga
MO000212079	HBS1L(h)	5	stop_gained	taC/taG
MO000176885	ZNF117(h)	2	stop_gained	Cga/Tga
MO000189841	ZSWIM1(h)	2	stop_gained	tGg/tAg
MO000208420	GJB3(h)	2	stop_gained	tGg/tAg
MO000109306	PSMA4(h)	1	stop_lost	Tga/Cga
MO000119745	SNX1(h)	1	stop_gained	tgG/tgA
MO000144222	APT2(h)	1	stop_lost	Tag/Cag
MO000172130	c3orf1(h)	1	NMD_transcript_variant,stop_lost	tGa/tCa
MO000175986	oas2(h)	1	stop_lost	tAg/tGg

Top 14 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 in of 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 11-12.

Table 11. 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	195
MO000118076	EGF:ErbB1{pY}:ErbB2{pY}:Src	EGF, EGFR, ERBB2, SRC	SRC proto-oncogene, non-receptor tyrosine kinase, epidermal growth factor, epidermal growth factor r...	4.92	240
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	248
MO000039099	IL-1beta-p17:IL-1RI:IL-1RAcP:MyD88:tollip:IRAK-1{pS376}{pT387}:IRAK-4:IRAK-2	IL1B, IL1R1, IL1RAP, IRAK1, IRAK2, IRAK4, MYD88, TOLLIP	interleukin 1 beta, interleukin 1 receptor accessory protein, interleukin 1 receptor associated kina...	1.93	265
MO000019548	p53(h)	TP53	tumor protein p53	1.43	282
MO000031094	MTA1L1(h):Mi2-BETA(h):HDAC1(h):RbAp48(h):mbd3(h)	CHD4, HDAC1, MBD3, MTA2, RBBP4	RB binding protein 4, chromatin remodeling factor, chromodomain helicase DNA binding protein 4, hist...	1.73	292
MO000033272	SGK-1(h)	SGK1	serum/glucocorticoid regulated kinase 1	1.45	293
MO000018901	CKII-alpha(h):CKII-alpha2(h):(CKII-beta(h))2	CSNK2A1, CSNK2A2, CSNK2B	casein kinase 2 alpha 1, casein kinase 2 alpha 2, casein kinase 2 beta	1.46	317
MO000157536	CKII-alpha(h):CKII-alpha2(h):CKII-beta(h)	CSNK2A1, CSNK2A2, CSNK2B	casein kinase 2 alpha 1, casein kinase 2 alpha 2, casein kinase 2 beta	1.46	324
MO000021478	Src(h)	SRC	SRC proto-oncogene, non-receptor tyrosine kinase	1.55	365

Table 12. 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	84
MO000036550	MKP-7(h)	DUSP16	dual specificity phosphatase 16	-1.71	120
MO000020367	Caspase-9(h)	CASP9	caspase 9	-0.89	124
MO000033396	DUSP5(h)	DUSP5	dual specificity phosphatase 5	-4.43	129
MO000176198	JKAP(h)	DUSP22	dual specificity phosphatase 22	-0.99	149
MO000022227	MKK7(h)	MAP2K7	mitogen-activated protein kinase kinase 7	-0.56	152
MO000103285	MKP-7-isoform1(h)	DUSP16	dual specificity phosphatase 16	-1.71	152
MO000019948	E1(h)	UBA1	ubiquitin like modifier activating enzyme 1	-0.69	157
MO000022228	MKK7(h){p}	MAP2K7	mitogen-activated protein kinase kinase 7	-0.56	163
MO000056883	ERK1-isoform1(h)	MAPK3	mitogen-activated protein kinase 3	-1.85	165

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.

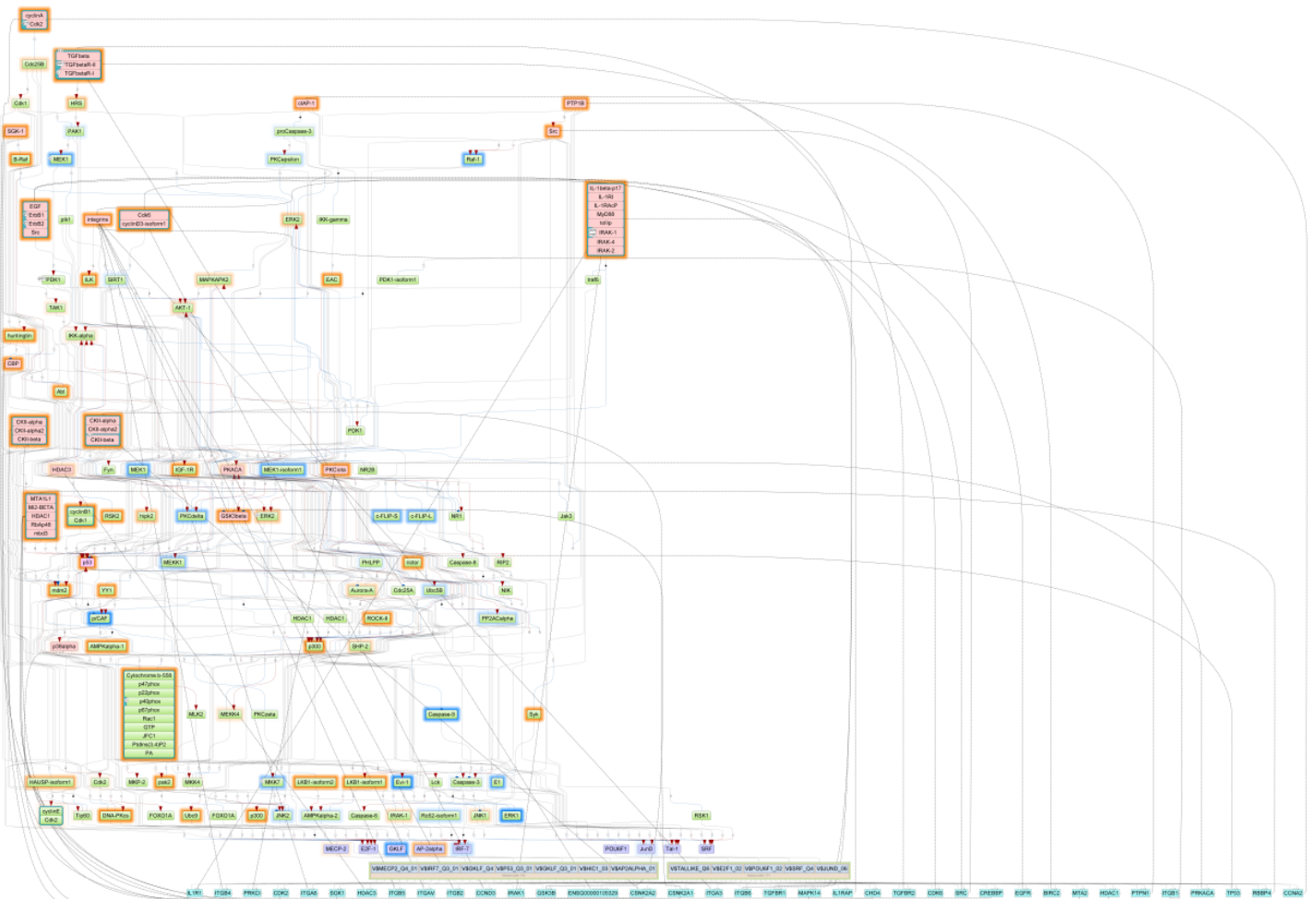


Figure 9. Diagram of intracellular regulatory signal transduction pathways of up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate molecules,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

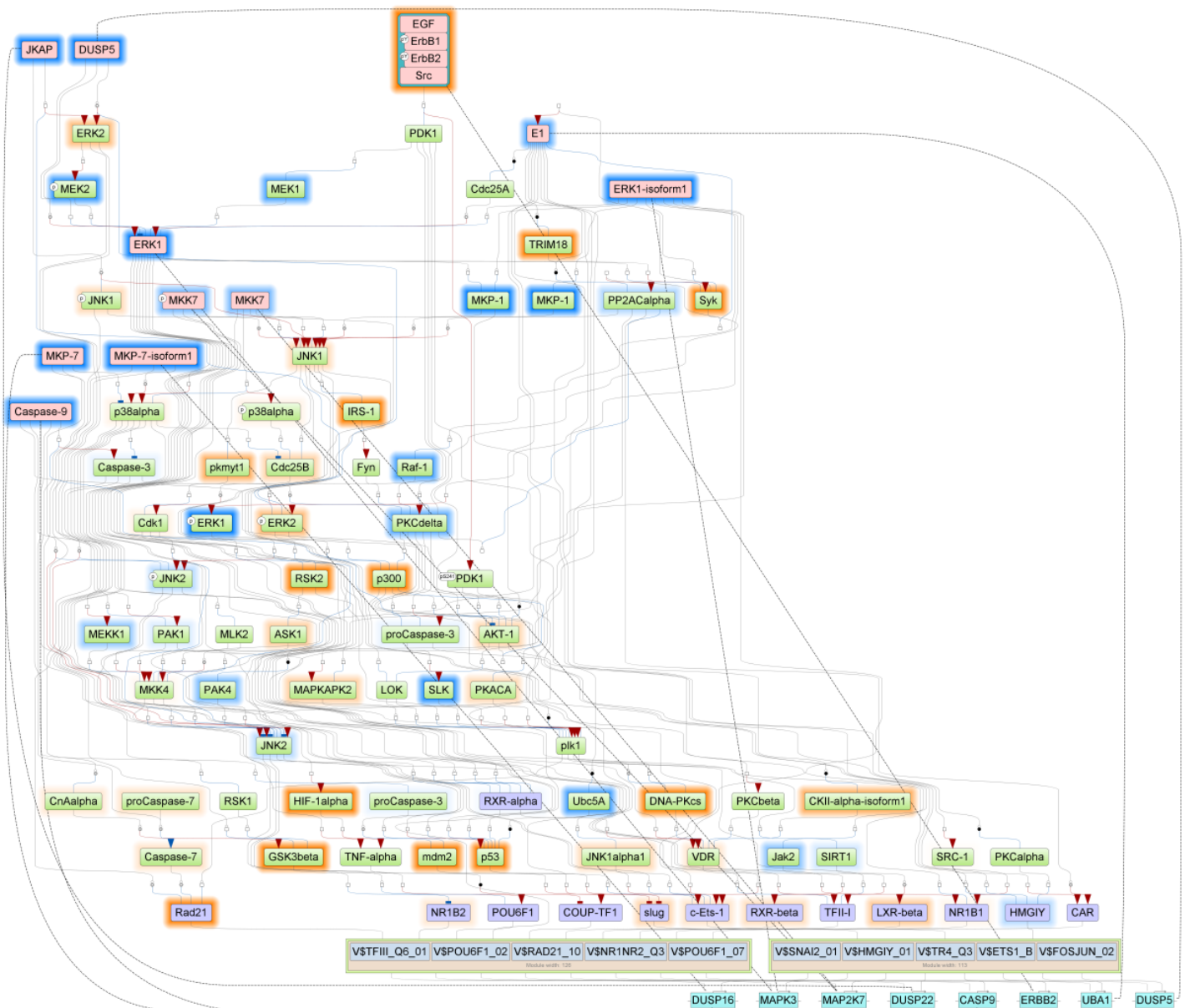


Figure 10. Diagram of intracellular regulatory signal transduction pathways of down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate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

In the last step of the analysis we strived to identify known drugs as well as new potentially active chemical compounds that are potentially suitable for inhibition (or activation) of the identified molecular targets in the context of specified human disease.

First, we identify known drugs using information from HumanPSD™ database [5] about their targets and about clinical trials where the drugs have been tested for the treatment of various human diseases. Table 13 shows the resulting list of druggable master regulators that represent the predicted drug targets of the studied pathology. Table 14 lists chemical compounds and known drugs (from the HumanPSD™ database) potentially acting on corresponding master regulators.

Table 13. 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 data.

[See full table →](#)

ID	Gene symbol	Gene description	Druggability score	logFC	Total rank
ENSG00000148053	NTRK2	neurotrophic receptor tyrosine kinase 2	1	6.48	393
ENSG00000134318	ROCK2	Rho associated coiled-coil containing protein kinase 2	2	2.61	426
ENSG00000101182	PSMA7	proteasome subunit alpha 7	3	1.71	526
ENSG00000197122	SRC	SRC proto-oncogene, non-receptor tyrosine kinase	46	4.92	658
ENSG00000160255	ITGB2	integrin subunit beta 2	1	3.47	811
ENSG00000139687	RB1	RB transcriptional corepressor 1	2	0.76	820
ENSG00000118515	SGK1	serum/glucocorticoid regulated kinase 1	3	1.45	836
ENSG00000138448	ITGAV	integrin subunit alpha V	1	3.47	849
ENSG00000145391	SETD7	SET domain containing lysine methyltransferase 7	1	1.63	866
ENSG00000101966	XIAP	X-linked inhibitor of apoptosis	2	0.39	909

Table 14. The list of drugs (from Human PSD) approved or used in clinical trials for the application in squamous cell carcinoma and acting on master regulators revealed in our study. The column **Target activity score** contains the value of numeric function that depends on ranks of all targets that were found for the drug. The column **Disease activity score** contains the weighted sum of user selected diseases where the drug is known to be applied. We use sum of clinical trials phases as the weight of the disease. **Drug rank** column contains total rank of given drug among all found. See [Methods](#) section for details.

[See full table](#) →

ID	Name	Target names	Target activity score	NA	Phase 1	Phase 2	Phase 3	Phase 4	Disease activity score	Drug rank
DB01254	Dasatinib	SRC, ABL1, YES1, ABL2	0.94	Carcinoma, Squamous Cell, Brain Neoplasms, Carcinoma, Transitional Cell, Gastrointestinal Stromal Tumors, Glioblastoma, Leukemia, Lymphoid...	Carcinoma, Squamous Cell, Adenocarcinoma, Clear Cell, Adenocarcinoma, Mucinous, Brain Abscess, Brain Diseases, Breast Neoplasms...	Carcinoma, Squamous Cell, Adenocarcinoma, Clear Cell, Blast Crisis, Brain Abscess, Brain Diseases, Brain Neoplasms...	Leukemia, Leukemia, Lymphoid, Leukemia, Myelogenous, Chronic, BCR-ABL Positive, Leukemia, Myeloid, Leukemia, Myeloid, Accelerated Phase, Leukemia, Myeloid, Acute, Leukemia, Myeloid, Chronic-Phase...	Leukemia, Leukemia, Lymphoid, Chronic, BCR-ABL Positive, Leukemia, Myeloid, Precursor Cell Lymphoblastic Leukemia-Lymphoma	4	16
DB01169	Arsenic trioxide	CCND1, MAPK1, AKT1, JUN	0.66	Carcinoma, Basal Cell, Leukemia, Lymphocytic, Chronic, B-Cell, Leukemia, Lymphoid, Leukemia, Myeloid, Leukemia, Prolymphocytic, Leukemia, Promyelocytic, Acute...	Brain Abscess, Brain Neoplasms, Carcinoma, Non-Small-Cell Lung, Carcinoma, Small Cell, Central Nervous System Neoplasms, Colorectal Neoplasms, Glioblastoma...	Carcinoma, Squamous Cell, Adenocarcinoma, Brain Abscess, Breast Neoplasms, Carcinoma, Hepatocellular, Carcinoma, Small Cell, Transitional Cell...	Carcinoma, Hepatocellular, Leukemia, Leukemia, Myeloid, Acute, Leukemia, Promyelocytic, Acute, Myelodysplastic Syndromes, Neoplasms...	Leukemia, Leukemia, Myeloid, Leukemia, Promyelocytic, Acute	2	39
DB09079	Nintedanib	FGFR3, SRC, LYN	0.65	Carcinoma, Carcinoma, Non-Small-Cell Lung, Colorectal Neoplasms, Endometrial Neoplasms, Fallopian Tube Neoplasms, Idiopathic Pulmonary Fibrosis, Lung Diseases...	Adenocarcinoma, Breast Neoplasms, Carcinoma, Hepatocellular, Carcinoma, Non-Small-Cell Lung, Carcinoma, Renal Cell, Carcinoma, Small Cell, Colonic Neoplasms...	Carcinoma, Squamous Cell, Adenocarcinoma, Adenocarcinoma, Clear Cell, Adenocarcinoma, Mucinous, Angiomyoma, Appendiceal Neoplasms, Breast Neoplasms...	Carcinoma, Non-Small-Cell Lung, Colorectal Neoplasms, Idiopathic Pulmonary Fibrosis, Lung Diseases, Lung Diseases, Interstitial, Mesothelioma, Neoplasms...	Idiopathic Pulmonary Fibrosis, Pulmonary Fibrosis	2	42
DB09073	Palbociclib	CDK6, CDK4	0.54	Carcinoma, Squamous Cell, Breast Neoplasms, Carcinoma, Non-Small-Cell Lung, Carcinoma, Small Cell, Colorectal Neoplasms, Glioma, Lung Neoplasms...	Adenocarcinoma, Astrocytoma, Behavior, Brain Abscess, Breast Diseases, Breast Neoplasms, Carcinoma...	Carcinoma, Squamous Cell, Adenocarcinoma, Astrocytoma, Brain Abscess, Breast Diseases, Breast Neoplasms, Breast Neoplasms, Male...	Breast Neoplasms, Neoplasms, Noma	Breast Neoplasms, Neoplasms	3	73
DB05294	Vandetanib	VEGFA, EGFR	0.48	Biliary Tract Neoplasms, Breast Neoplasms, Carcinoma, Non-Small-Cell Lung, Carcinoma, Small Cell, Gastrointestinal Neoplasms, Glioma, Intestinal Neoplasms...	Adenocarcinoma, Brain Abscess, Breast Neoplasms, Carcinoma, Hepatocellular, Carcinoma, Non-Small-Cell Lung, Carcinoma, Renal Cell...	Carcinoma, Squamous Cell, Astrocytoma, Biliary Tract Neoplasms, Brain Abscess, Breast Neoplasms, Carcinoma, Hepatocellular, Carcinoma, Non-Small-Cell Lung...	Carcinoma, Non-Small-Cell Lung, Carcinoma, Small Cell, Lung Neoplasms, Neoplasms, Noma, Small Cell Lung Carcinoma, Thyroid Neoplasms	Neoplasms, Thyroid Neoplasms	2	99

Table 15. 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 squamous cell carcinoma. **Target activity score** column contains value of numeric function that depends on ranks of all targets that were found for the drug. **Drug rank** column contains total rank of given drug among all found. See [Methods](#) section for details.

ID	Name	Target names	Target activity score	NA	Phase 1	Phase 2	Phase 3	Phase 4	Drug rank
DB06616	Bosutinib	CAMK2G, SRC, ABL1, HCK, LYN, CDK2	1.54	Breast Neoplasms, Leukemia, Leukemia, Lymphoid, Leukemia, Myelogenous, Chronic, BCR-ABL Positive, Leukemia, Myeloid, Neoplasms, Precursor Cell Lymphoblastic Leukemia-Lymphoma	Acute Kidney Injury, Breast Neoplasms, Carcinoma, Non-Small-Cell Lung, Cholangiocarcinoma, Cognitive Dysfunction, Colorectal Neoplasms, Dementia...	Brain Abscess, Breast Neoplasms, Cholangiocarcinoma, Colorectal Neoplasms, Cysts, Glioblastoma, Kidney Diseases, Cystic...	Leukemia, Leukemia, Myelogenous, Chronic, BCR-ABL Positive, Leukemia, Myeloid	Leukemia, Myeloid	41
DB00030	Insulin Regular	IGFBP7, RB1, IGF1R	0.6		Sarcopenia			Diabetes Mellitus, Diabetes Mellitus, Type 1	90
DB06603	Panobinostat	HDAC6, HDAC7, HDAC3, HDAC1	0.53	Anemia, Refractory, with Excess of Blasts, Brain Abscess, Breast Neoplasms, HIV Infections, Leukemia, Leukemia, Myeloid, Leukemia, Myeloid, Acute...	Adenocarcinoma, Anemia, Sickle Cell, Brain Abscess, Breast Neoplasms, Burkitt Lymphoma, Carcinoma, Non-Small-Cell Lung, Carcinoma, Renal Cell...	Brain Abscess, Breast Neoplasms, Burkitt Lymphoma, Carcinoma, Renal Cell, Glioblastoma, Glioma, Graft vs Host Disease...	Brain Abscess, Hodgkin Disease, Leukemia, Myelogenous, Chronic, BCR-ABL Positive, Leukemia, Lymphoma, T-Cell...	Brain Abscess, Multiple Myeloma	97
DB00098	Anti-thymocyte Globulin (Rabbit)	ITGB1, ITGAV	0.44	Arthritis, Osteoarthritis		Anemia, Aplastic, Hemoglobinuria, Hemoglobinuria, Paroxysmal, Hodgkin Disease, Leukemia, Leukemia, Lymphoid, Leukemia, Myelogenous, Chronic, BCR-ABL Positive...	Sepsis, Shock, Shock, Septic	Anemia, Anemia, Aplastic, Leukemia, Liver Diseases	118
DB00641	Simvastatin	ITGB2	0.41	Acute Coronary Syndrome, Acute Lung Injury, Affect, Alzheimer Disease, Aneurysm, Aortic Aneurysm, Asthma...	Acute Coronary Syndrome, Acute Kidney Injury, Affect, Anemia, Sickle Cell, Angiomyoma, Arthritis, Arthritis, Rheumatoid...	Adenocarcinoma, Affect, Alzheimer Disease, Anemia, Sickle Cell, Angiomyoma, Asthma, Atherosclerosis...	Acute Coronary Syndrome, Affect, Alzheimer Disease, Aortic Valve Stenosis, Asthma, Atherosclerosis, Brain Abscess...	Acute Coronary Syndrome, Affect, Alzheimer Disease, Atherosclerosis, Bipolar Disorder, Brain Abscess, Brain Injuries...	144

Next, new potential small molecular ligands were predicted for the revealed targets and a general druggability check was run using a pre-computed database of spectra of biological activities of chemical compounds from a library of 13040 most pharmaceutically active known compounds. The spectra of biological activities has been computed using the program PASS [11-13] on the basis of a (Q)SAR approach. Table 16 shows the resulting list of druggable master regulators, which represent the predicted drug targets of the studied pathology. Table 17 lists chemical compounds and known drugs potentially acting on the corresponding master regulators.

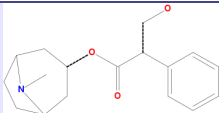
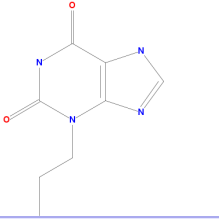
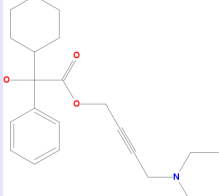
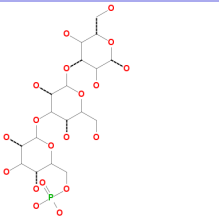
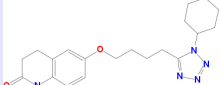
Table 16. Extended list of drug targets revealed in this study (targets that are predicted by PASS program potentially targeted by an extended list of known drugs and pharmaceutically active chemical compounds). The column **Druggability score** contains a numeric value which indicates how suitable this 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	logFC	Total rank
ENSG0000005884	ITGA3	ITGA3	integrin subunit alpha 3	3.31	3.47	248
ENSG00000082781	ITGB5	ITGB5	integrin subunit beta 5	0.65	3.47	248
ENSG00000091409	ITGA6	ITGA6	integrin subunit alpha 6	3.31	3.47	248
ENSG00000115221	ITGB6	ITGB6	integrin subunit beta 6	7.63	3.47	248
ENSG00000111642	CHD4	CHD4	chromodomain helicase DNA binding protein 4	15.7	1.73	292
ENSG00000149480	MTA2	MTA2	metastasis associated 1 family member 2	15.7	1.73	292
ENSG00000162521	RBBP4	RBBP4	RB binding protein 4, chromatin remodeling factor	17.43	1.73	292
ENSG00000148053	NTRK2	NTRK2	neurotrophic receptor tyrosine kinase 2	20.54	6.48	393
ENSG00000134318	ROCK2	ROCK2	Rho associated coiled-coil containing protein kinase 2	0.11	2.61	426
ENSG00000087191	PSMC5	PSMC5	proteasome 26S subunit, ATPase 5	0.85	1.71	526

Table 17. 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 squamous cell carcinoma in the current pathological case. **Disease activity score** column contains maximal value of probability to be active for all activities corresponding to the selected diseases for the given compound or 0 if no diseases were selected (in this case column will be hidden). **Target activity score** column contains value of numeric function which depends on all activity-mechanisms correspondent to the drug. **Drug rank** column contains total rank of given drug among all found. See [Methods](#) section for details.

[See full table](#) →

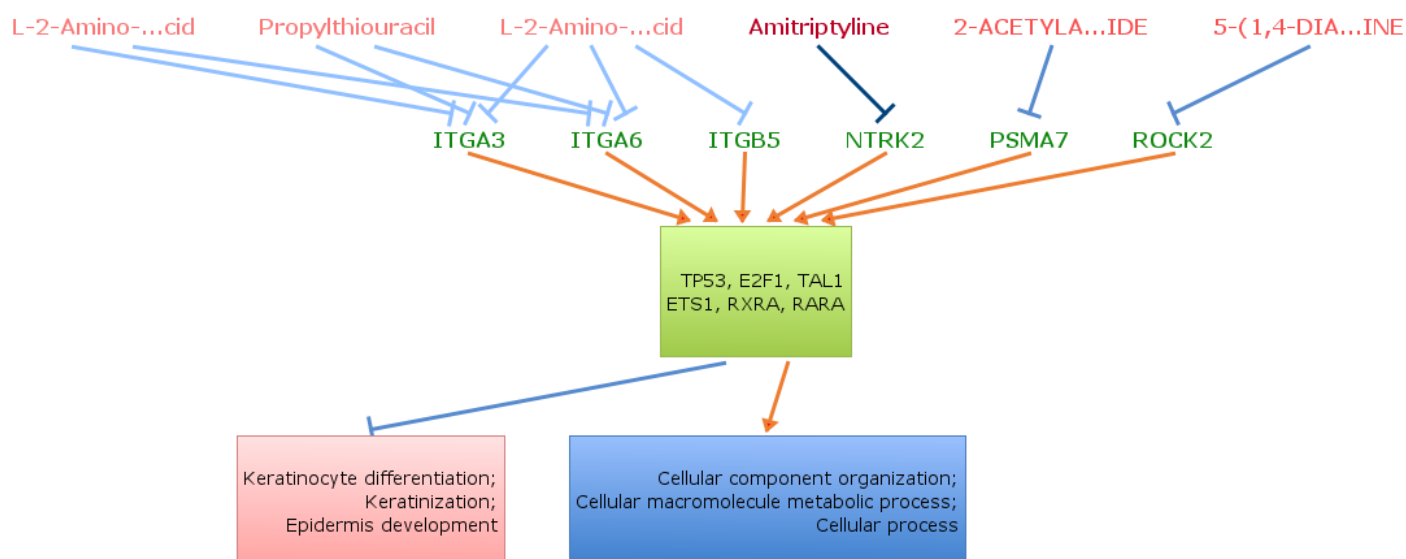
Name	Structure	Target names	Target activity score	Disease activity score	Drug rank
Hyoscyamine		CCND1, FKBP1A, CDK6, CCNH, CCND3, CCNB1, CCNA2...	3.35	0	2
Enprofylline		RIPK2, FGFR3, SRC, NTRK2, ABL1, PIM3, MERTK...	2.74	0	3
Oxybutynin		CDK6, CCND3, CCNB1, CDK1, CCNB2, CDK4, CCND1...	2.34	0	4
1-3 Sugar Ring of Pentamannosyl 6-Phosphate		TRAF4, TBL1XR1, UBR4, HLTf, PRPF19, MNAT1, HTRA2...	2.3	0	5
Cilostazol		CCND1, CDK6, CCNH, CCND3, PRKCI, CCNB1, AURKA...	2.21	0	6

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



6. Methods

Databases used in the study

Transcription factor binding sites in promoters and enhancers of differentially expressed genes were analyzed using known DNA-binding motifs described in the TRANSFAC® library, release 2020.1 (geneXplain GmbH, Wolfenbüttel, Germany) (<http://genexplain.com/transfac>).

The master regulator search uses the TRANSPATH® database (BIOBASE), release 2020.1 (geneXplain GmbH, Wolfenbüttel, Germany) (<http://genexplain.com/transpath>). A comprehensive signal transduction network of human cells is built by the software on the basis of reactions annotated in TRANSPATH®.

The information about drugs corresponding to identified drug targets and clinical trials references were extracted from HumanPSD™ database, release 2020.1 (<http://genexplain.com/humanpsd>).

The Ensembl database release Human88.38 (hg38) (<http://www.ensembl.org>) was used for gene IDs representation and Gene Ontology (GO) (<http://geneontology.org>) was used for functional classification of the studied gene set.

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

Transcription factor binding sites in promoters and enhancers of differentially expressed genes were analyzed using known DNA-binding motifs. The motifs are specified using position weight matrices (PWMs) that give weights to each nucleotide in each position of the DNA binding motif for a transcription factor or a group of them.

We search for transcription factor binding sites (TFBS) that are enriched in the promoters and enhancers under study as compared to a background sequence set such as promoters of genes that were not differentially regulated under the condition of the experiment. We denote study and background sets briefly as Yes and No sets. In the current work we used a workflow considering promoter sequences of a standard length of 1100 bp (-1000 to +100). The error rate in this part of the pipeline is controlled by estimating the adjusted p-value (using the Benjamini-Hochberg procedure) in comparison to the TFBS frequency found in randomly selected regions of the human genome (adj.p-value < 0.01).

We have applied the CMA algorithm (Composite Module Analyst) for searching composite modules [7] in the promoters and enhancers of the Yes and No sets. We searched for a composite module consisting of a cluster of 10 TFs in a sliding window of 200-300 bp that statistically significantly separates sequences in the Yes and No sets (minimizing Wilcoxon p-value).

Methods for finding master regulators in networks

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

Methods for analysis of pharmaceutical compounds

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

Method for analysis of known pharmaceutical compounds

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

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

To calculate clinical trials phase for the given compound we select the maximum phase of all diseases that are known to have clinical trials with this compound. "Target activity score" ($T\text{-score}_{PSD}$) is calculated as follows:

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

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 of all targets related to the compound and number of elements in it, w is weight multiplier, $\text{rank}(t)$ is rank of given target, $\text{maxRank}(T)$ equals $\text{max}(\text{rank}(t))$ for all targets t in T .

We use following formula to calculate "Disease activity score" ($D\text{-score}_{PSD}$):

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

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, $\text{phase}(p,d)$ equals to the phase number if there are known clinical trials for the selected disease on this phase and zero otherwise.

Method for prediction of pharmaceutical compounds

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

We selected compounds that satisfied the following conditions:

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

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

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

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

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

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

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

7. References

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

Thank you for using the Genome Enhancer!

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

Supplementary material

1. [Supplementary table 1 - Up-regulated genes](#)
2. [Supplementary table 2 - Down-regulated genes](#)
3. [Supplementary table 3 - Detailed report. Composite modules and master-regulators \(up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue\).](#)
4. [Supplementary table 4 - Detailed report. Composite modules and master-regulators \(down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue\).](#)
5. [Supplementary table 5 - Detailed report. Pharmaceutical compounds and drug targets.](#)

Disclaimer

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

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

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

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

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

Note that the text given in the report is not unique and can be fully or partially repeated in other Genome Enhancer analysis reports, including reports of other users. This should be considered when publishing any results or excerpts from the report. This restriction refers only to the general description of analysis methods used for generating the report. All data and graphics referring to the concrete set of input data, including lists of mutated genes, differentially expressed genes/proteins/metabolites, functional classifications, identified transcription factors and master regulators, constructed molecular networks, lists of chemical compounds and reconstructed model of molecular mechanisms of the studied pathology are unique in respect to the used input data set and Genome Enhancer pipeline parameters used for the current run.