CCND1 and CCND3 are promising druggable targets for treating Squamous Cell Carcinoma that control activity of E2F1, NFATC1 and TP53 transcription factors on promoters of differentially expressed genes

Demo User
geneXplain GmbH
info@genexplain.com
Data received on 13/08/2019 ; Run on 30/10/2020 ; Report generated on 30/10/2020

Genome Enhancer release 2.2 (TRANSFAC®, TRANSPATH® and HumanPSD™ release 2020.3)

Abstract

In the present study we applied the software package "Genome Enhancer" to a data set that contains *transcriptomics* data. 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: E2F1, NFATC1, EGR1, TP53, TAL1 and TCF3. The subsequent network analysis suggested

- Cdk6:cyclinD3-isoform1
- Cdk4-isoform1:cyclinD1a
- PP2A

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: Palbociclib, Bosutinib and {(2Z)-4-AMINO-2-[(4-METHOXYPHENYL)IMINO]-2,3-DIHYDRO-1,3-THIAZOL-5-YL}(4-METHOXYPHENYL)METHANONE.

1. Introduction

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

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

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

2. Data

For this study the following experimental data was used:

Table 1. Experimental datasets used in the study

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

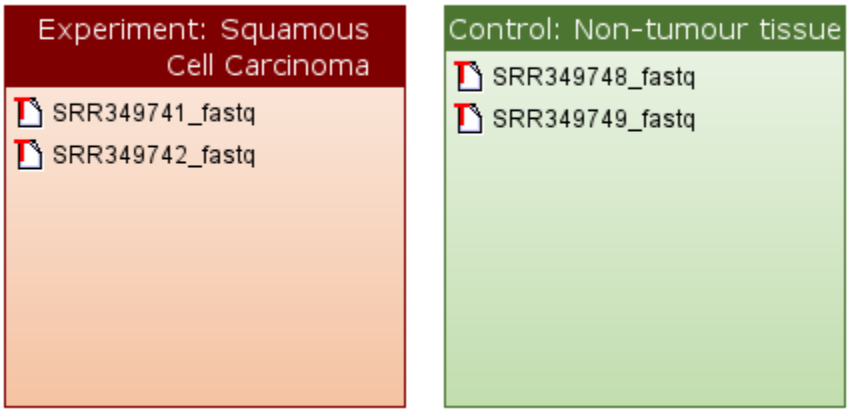


Figure 1. Annotation diagram of experimental data used in this study. With the colored boxes we show those sub-categories of the data that are compared in our analysis.

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 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 →

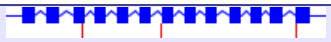
ID	Gene symbol	Gene schematic representation
ENSG00000115758	ODC1	
ENSG00000113140	SPARC	
ENSG00000163359	COL6A3	
ENSG00000120708	TGFBI	
ENSG00000134871	COL4A2	
ENSG00000186340	THBS2	
ENSG00000146648	EGFR	
ENSG00000182326	C1S	
ENSG00000122786	CALD1	
ENSG00000053747	LAMA3	

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

See full table →

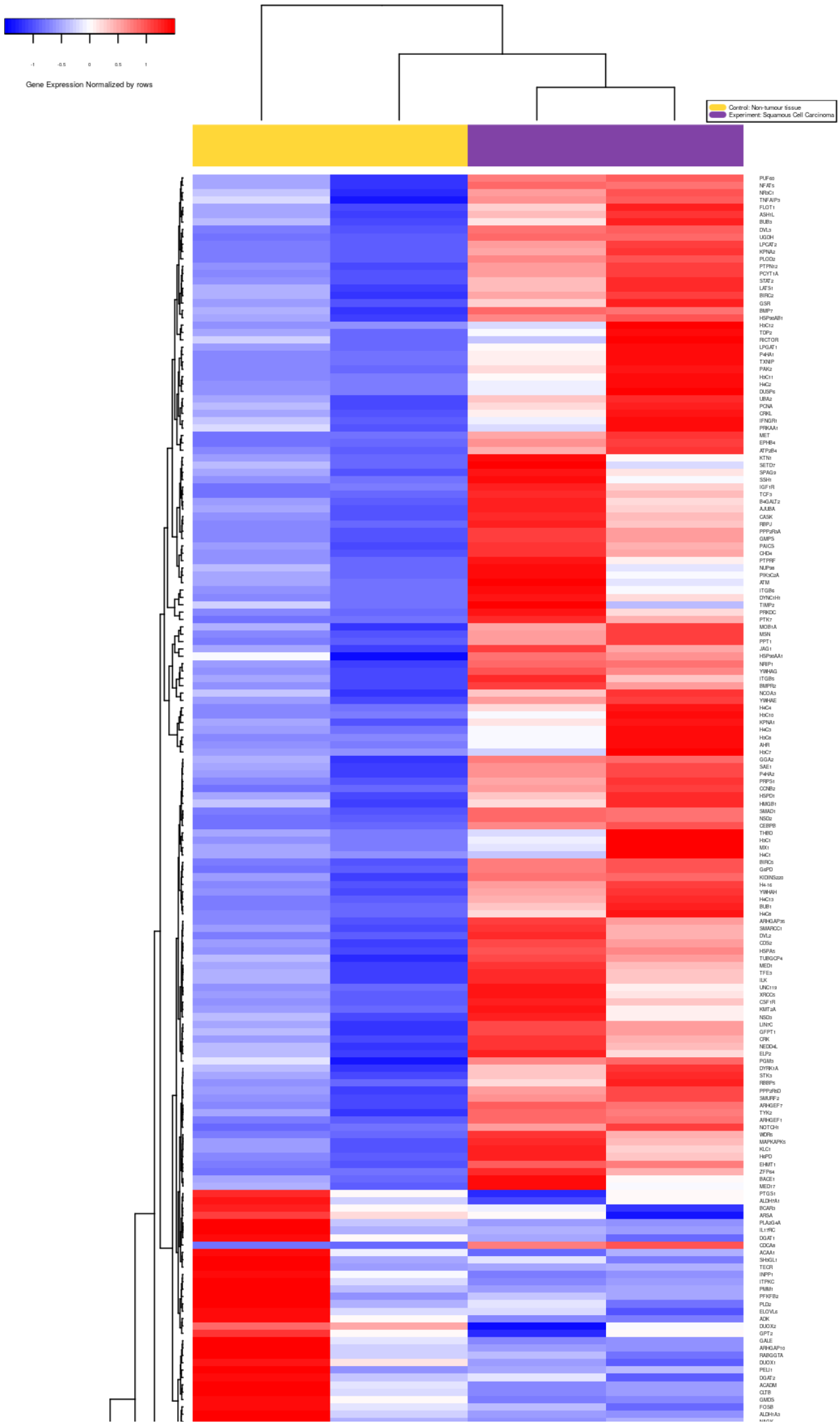
ID	Gene symbol	Gene schematic representation
ENSG00000163209	SPRR3	
ENSG00000244094	SPRR2F	
ENSG00000177191	B3GNT8	
ENSG00000260276	AC022167.2	
ENSG00000124466	LYPD3	
ENSG00000074416	MGLL	
ENSG00000153048	CARHSP1	
ENSG00000170545	SMAGP	
ENSG00000211448	DIO2	
ENSG00000135373	EHF	

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
PUF60
NFATc
NR3C1
TNFAIP3
FLOT1
ASH1L
BUB3
DVL3
UGDH
LPCAT2
KPNA2
PLOD2
PTPN12
PCYT1A
STAT2
LATS1
BIRC2
GSR
BMP7
HSP90AB1
H3C12
TDP2
RICTOR
LPGAT1
P4HA1
TXNIP
PAK2
H3C11
H4C2
DUSP6
UBA2
PCNA
CRNL
IFNGR1
PRKAA1
MET
EPHB4
ATP2B4
KTN1
SETD7
SPAG9
SSH1
IGF1R
TCF3
B4GALT2
AJUBA
CASK
RBPJ
PPP2R5A
GMPS
PAICS
CHD4
PTPRF
NUP98
PIK3CA
ATM
ITGB5
DYNC1H1
TIMP2
PRKDC
PTK7
MOB1A
MSN
PPT1
JAG1
HSP90AA1
NRP1
YWHAG
ITGB1
BMPR2
NCOA3
YWHAE
H4C4
H3C10
KPNA1
H4C3
H3C8
AHR
H3C7
GGA2
SAE1
P4HA2
PRPS1
CCNB2
HSPD1
HMGB1
SMAD1
NSD2
CEBPB
THBD
H3C1
MXI
H4C1
BIRC5
G6PD
KIDINS220
H4-16
YWHAH
H4C13
BUB1
H4C8
ARHGAP35
SMARCC1
DVL2
CDS2
HSPA5
TUBGCP4
MED1
TFE3
ILK
UNC119
XRCC5
CSF1R
KMT2A
NSD3
LRRC
GFPT1
CRK
NEDD4L
ELP2
PGM3
DYRK1A
STK3
RRBP1
PPP4R2
SMURF2
ARHGEF7
TYK2
ARHGEF1
NOTCH1
WDR5
MAPKAPK5
KLC1
H6PD
EHMT1
ZFP64
BACE1
MED17
PTGS1
ALDH7A1
BCAR3
AR5A
PLA2G4A
IL11RC
DGAT1
CDCA8
ACAA1
SHOGL1
TECR
INPP1
ITPKC
PMM1
PFKFB2
PLD2
ELOVL5
AOK
DUOX2
GPT2
GALE
ARHGAP10
RABGGTA
DUOX1
PELI1
DGAT2
ACADM
CLTB
GMDS
FOSB
ALDH1A3

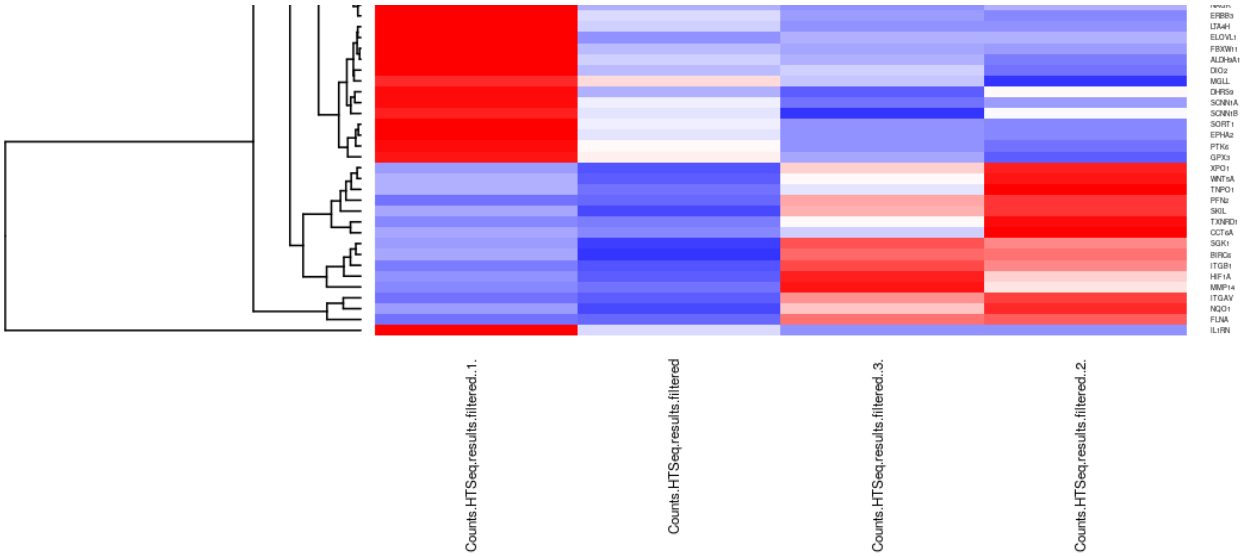


Figure 2. Heatmap of genes enriched in Transpath categories. The colored bar at the top shows the types of the samples according to the legend in the upper right corner.
See full diagram →

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 (2020.3)

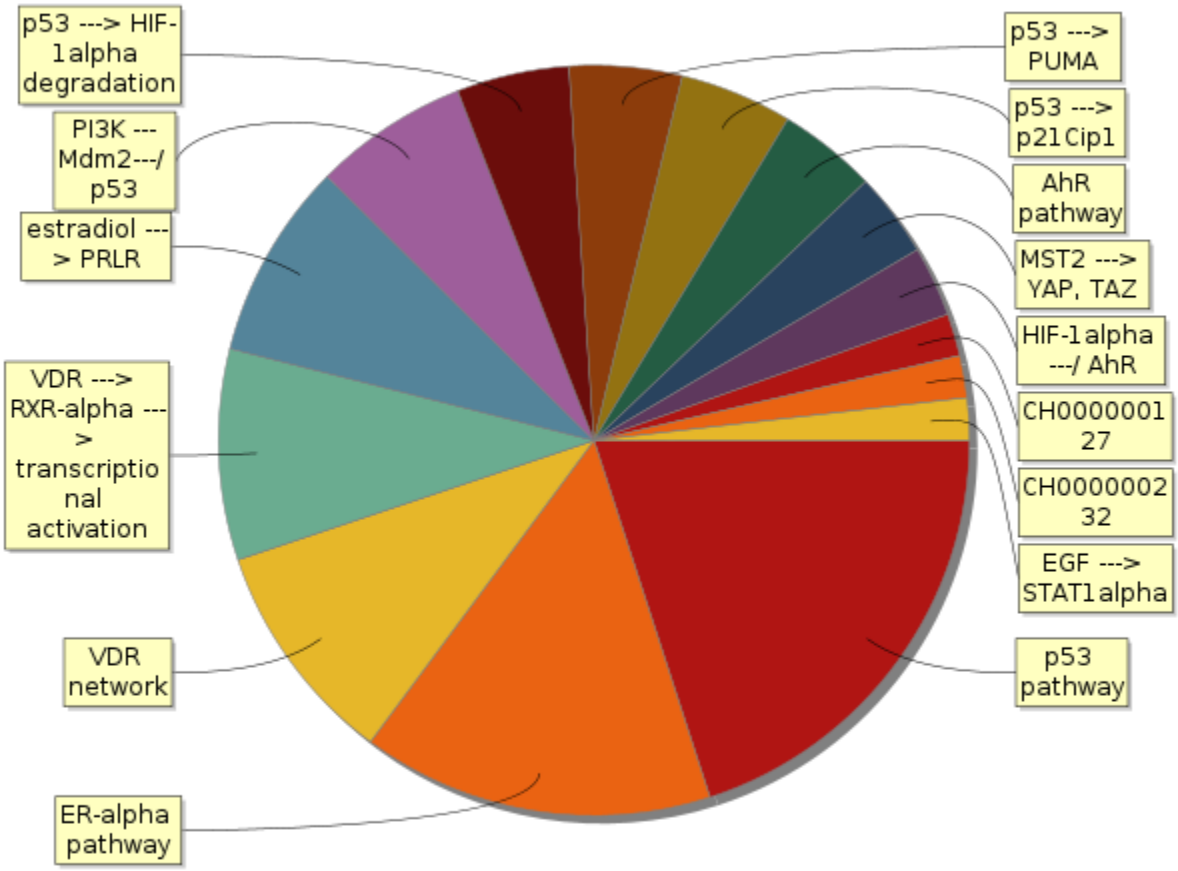


Figure 4. Enriched TRANSPATH® Pathways (2020.3) of up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue.

Full classification →

HumanPSD(TM) disease (2020.3)

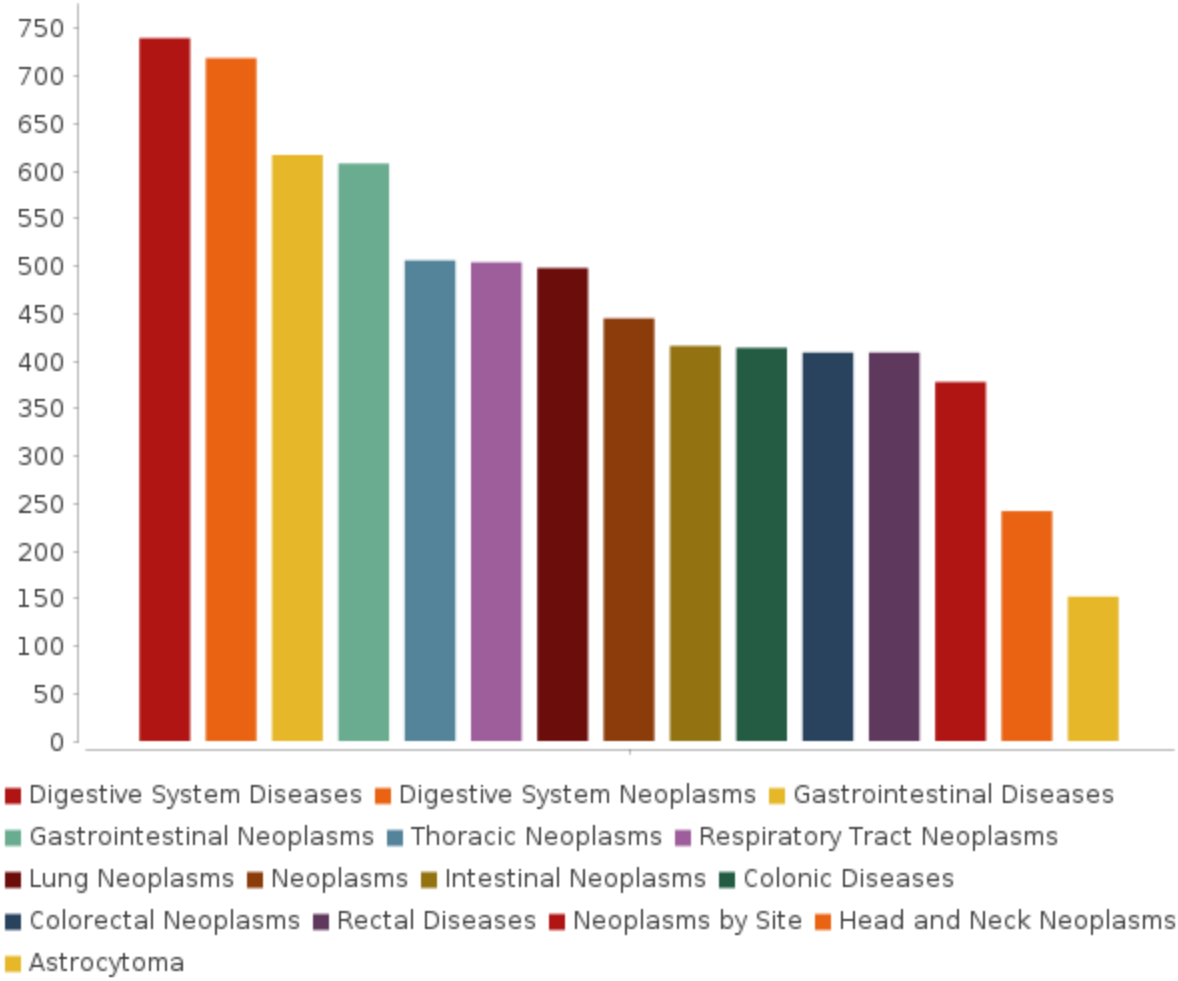


Figure 5. Enriched HumanPSD(TM) disease (2020.3) of up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

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 (2020.3)

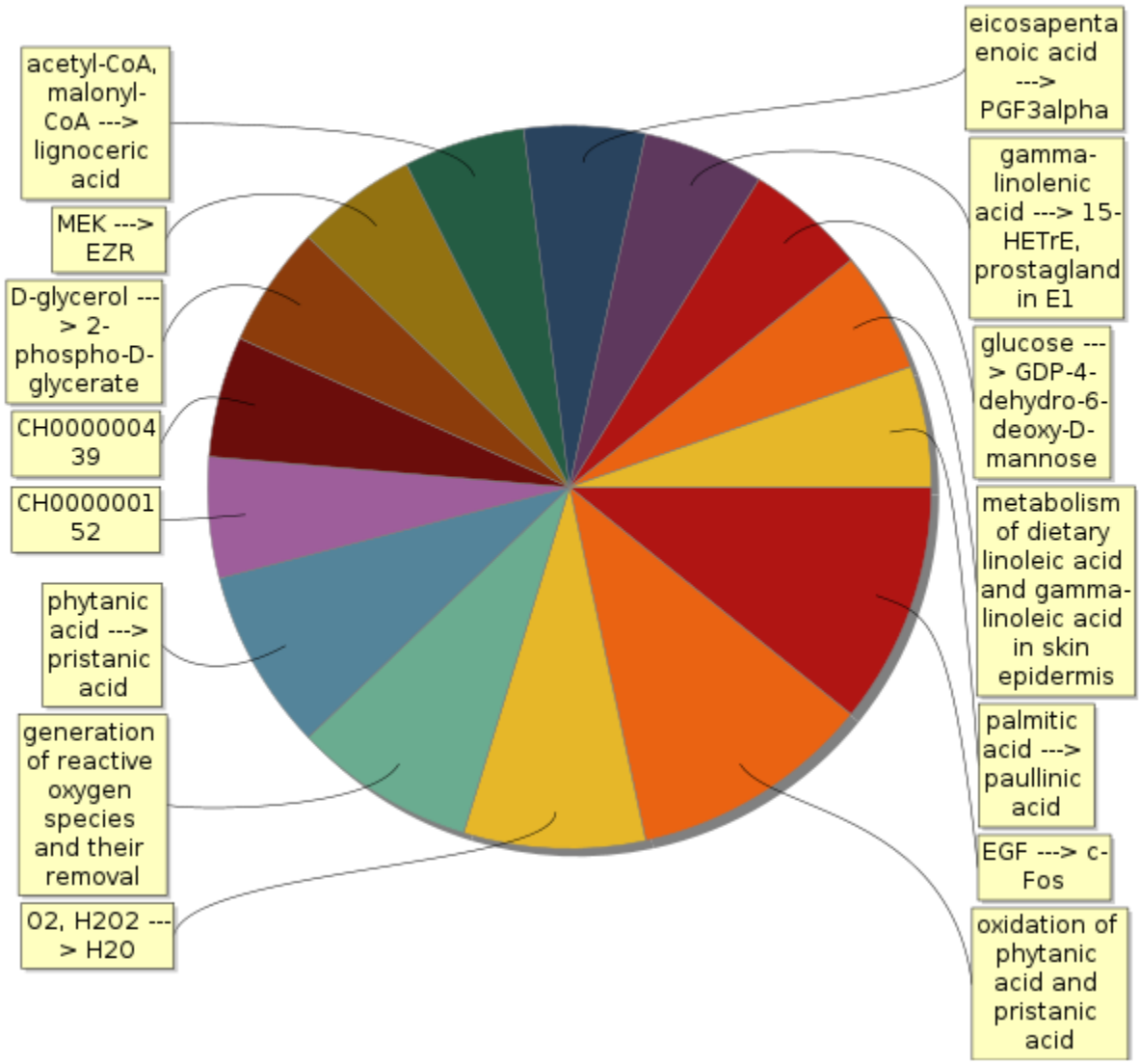


Figure 7. Enriched TRANSPATH® Pathways (2020.3) of down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue.

Full classification →

HumanPSD(TM) disease (2020.3)

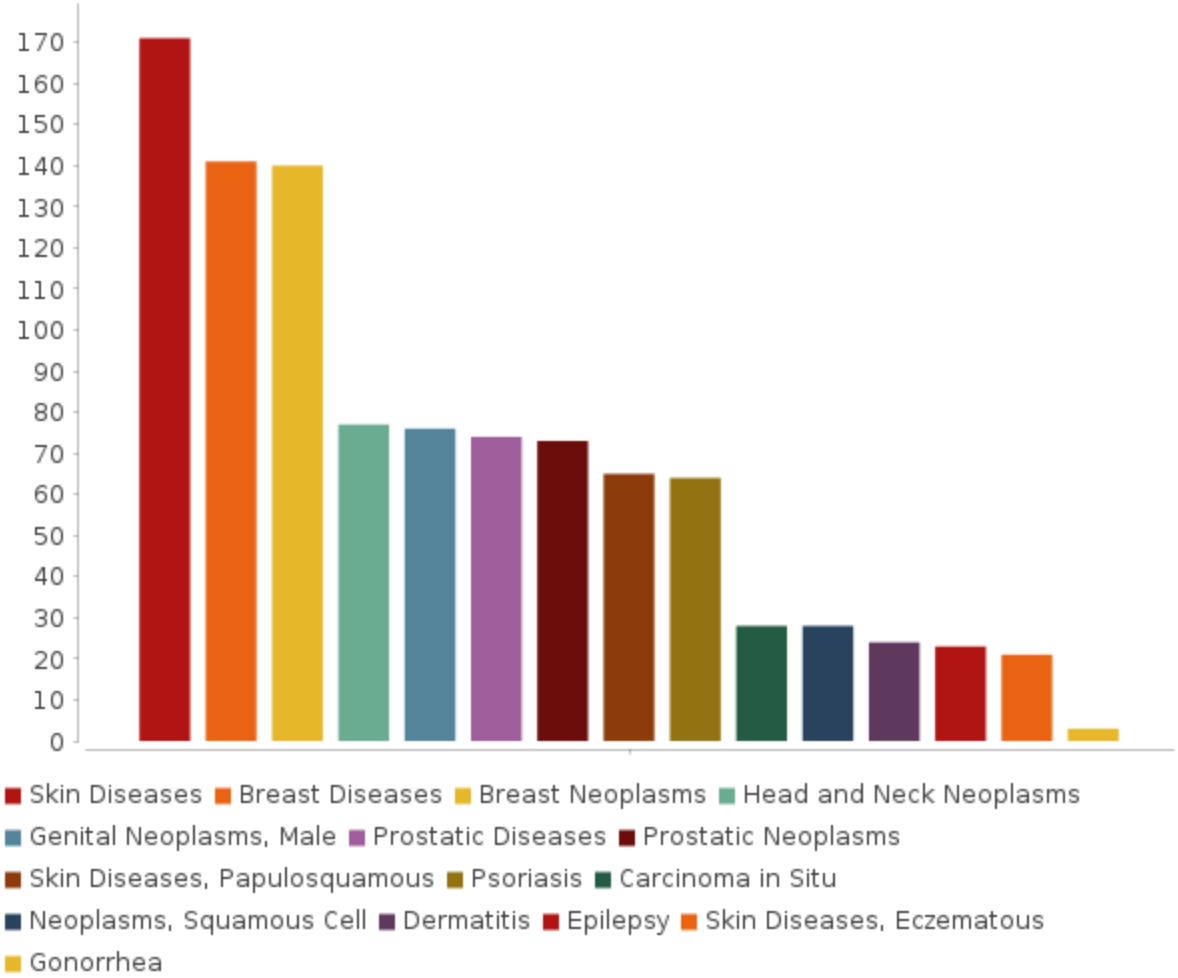
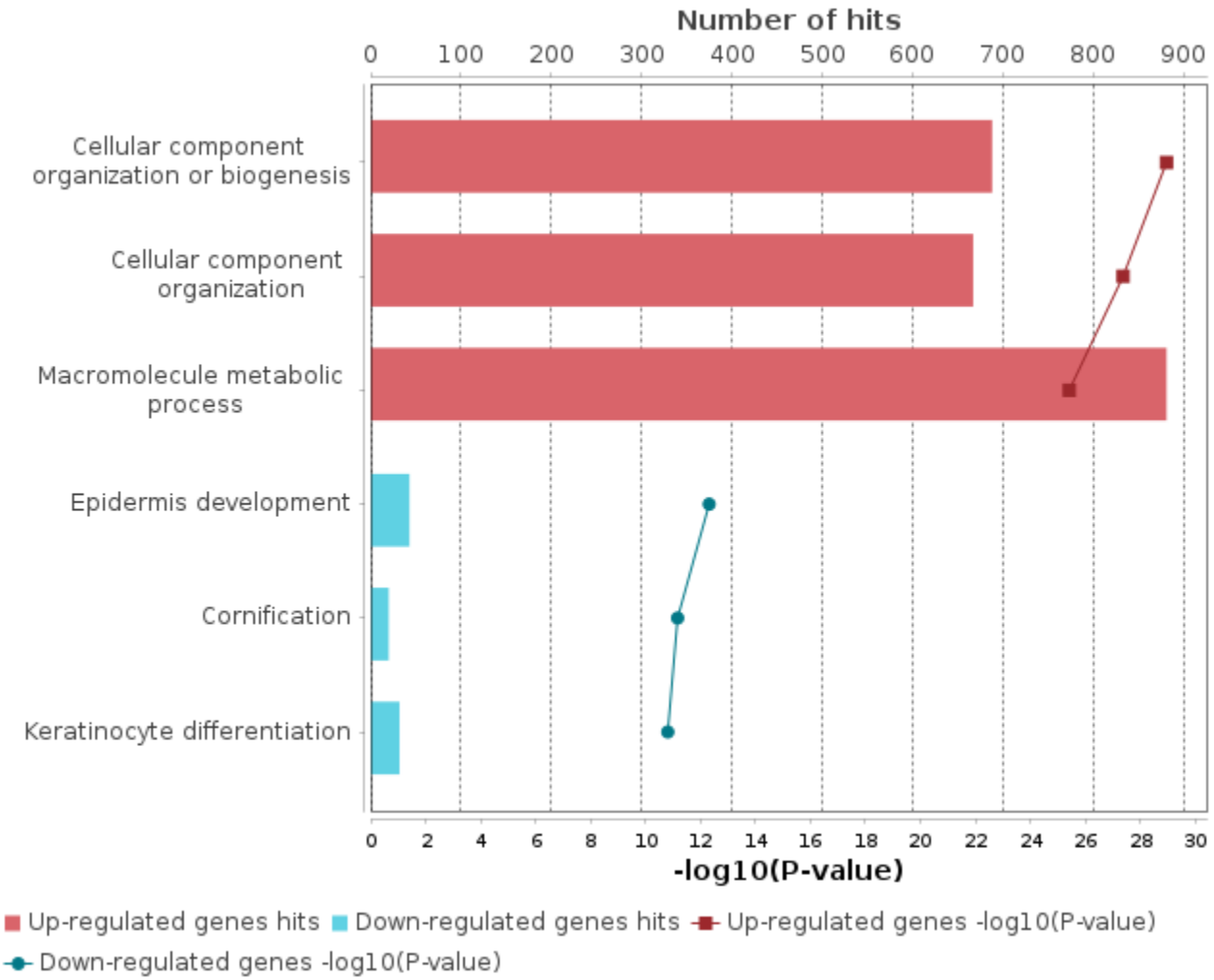


Figure 8. Enriched HumanPSD(TM) disease (2020.3) of down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

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 180 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 top 300 significant up-regulated genes as the input of CMA algorithm. The obtained CMA model is then applied to compute CMA score for all up-regulated genes.

The model consists of 2 module(s). Below, for each module the following information is shown:
- PWMs producing matches,
- number of individual matches for each PWM,
- score of the best match.

Module 1:

V$E2F1_02	V$GCM_Q2	V$GKLF_Q3_01	V$TAL1BETAITF2_01	V$EGR_Q3
0.94; N=2	0.91; N=3	0.99; N=2	0.86; N=3	0.90; N=1

V$MZF1_02
0.92; N=2

Module width: 101

Module 2:

V$EGR1_09	V$CEBPD_Q6	V$BCL11A_02	V$NFAT2_Q5	V$SOX10_Q3
0.87; N=3	0.89; N=3	0.88; N=1	1.00; N=1	0.99; N=3

V$SOX9_08
0.70; N=2

Module width: 137

Model score (-p*log10(pval)): 11.69
Wilcoxon p-value (pval): 2.27e-25
Penalty (p): 0.475
Average yes-set score: 5.86
Average no-set score: 4.51
AUC: 0.75
Middle-point: 5.61
False-positive: 21.60%
False-negative: 37.02%

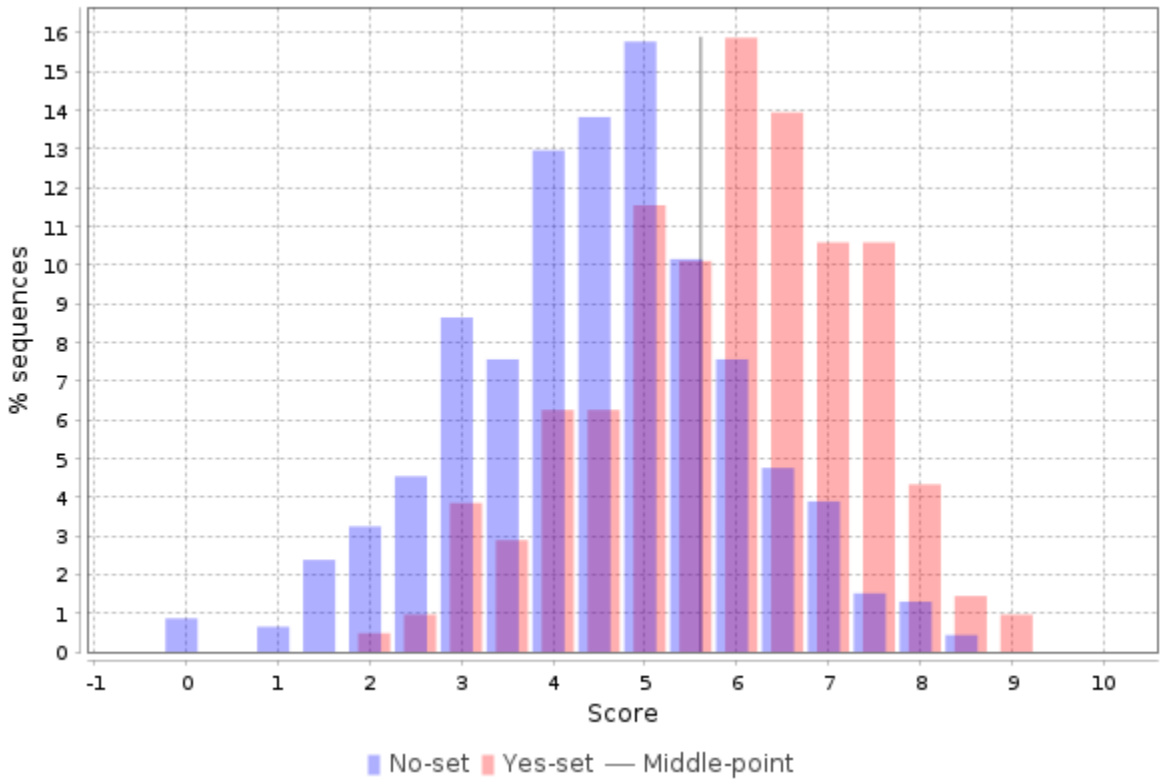


Table 8. List of top ten up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue with identified enhancers in their regulatory regions. ***CMA score*** *- the score of the CMA model of the enhancer identified in the regulatory region.*

See full table →

Ensembl IDs	Gene symbol	Gene description	CMA score	Factor names
ENSG00000166847	DCTN5	dynactin subunit 5	11.12	E2F-1(h), GCMa(h),GCMb(h), GKLF(h), Egr-1(h),Egr-2(h),Egr-4(h),egr-3(h), MZF-1(h), Sox-10(h), C/EBPdelta(h)...
ENSG00000173575	CHD2	chromodomain helicase DNA binding protein 2	10.62	Sox-10(h), C/EBPdelta(h), NFATc1(h), E2F-1(h), Sox-9(h), BCL-11A(h), GKLF(h)...
ENSG00000153250	RBMS1	RNA binding motif single stranded interacting protein 1	10.26	NFATc1(h), Egr-1(h), C/EBPdelta(h), GKLF(h), Egr-1(h),Egr-2(h),Egr-4(h),egr-3(h), MZF-1(h), GCMa(h),GCMb(h)...
ENSG00000140153	WDR20	WD repeat domain 20	10.26	NFATc1(h), Sox-10(h), C/EBPdelta(h), Sox-9(h), BCL-11A(h), MZF-1(h), GKLF(h)...
ENSG00000250317	SMIM20	small integral membrane protein 20	10.13	MZF-1(h), E2F-1(h), ITF-2(h),Tal-1(h), C/EBPdelta(h), Sox-9(h), Sox-10(h), BCL-11A(h)...
ENSG00000146386	ABRACL	ABRA C-terminal like	9.73	E2F-1(h), Egr-1(h),Egr-2(h),Egr-4(h),egr-3(h), GKLF(h), MZF-1(h), Sox-9(h), C/EBPdelta(h), NFATc1(h)...
ENSG00000168438	CDC40	cell division cycle 40	9.62	C/EBPdelta(h), E2F-1(h), MZF-1(h), GKLF(h), GCMa(h),GCMb(h), Egr-1(h),Egr-2(h),Egr-4(h),egr-3(h), BCL-11A(h)...
ENSG00000120008	WDR11	WD repeat domain 11	9.45	E2F-1(h), GKLF(h), Sox-10(h), C/EBPdelta(h), NFATc1(h), Sox-9(h), BCL-11A(h)
ENSG00000278828	H3C10	H3 clustered histone 10	9.41	E2F-1(h), C/EBPdelta(h), GKLF(h), Egr-1(h),Egr-2(h),Egr-4(h),egr-3(h), BCL-11A(h), Sox-9(h), NFATc1(h)...
ENSG00000005020	SKAP2	src kinase associated phosphoprotein 2	9.41	Sox-10(h), Egr-1(h),Egr-2(h),Egr-4(h),egr-3(h), E2F-1(h), C/EBPdelta(h), Sox-9(h), BCL-11A(h), NFATc1(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 top 300 significant down-regulated genes as the input of CMA algorithm. The obtained CMA model is then applied to compute CMA score for all down-regulated genes.

The model consists of 2 module(s). Below, for each module the following information is shown:
- PWMs producing matches,
- number of individual matches for each PWM,
- score of the best match.

Module 1:

V$P53_Q3_01	V$GKLF_Q4	V$TAL1ALPHAE47_01	V$PREP1_01
0.90; N=2	1.00; N=3	0.85; N=2	0.81; N=2

V$POU2F1_02	V$TAL1_Q6
0.85; N=3	0.94; N=1

Module width: 136

Module 2:

V$CP2_Q6	V$FOXO4_05	V$STAT6_02	V$USF2_Q6_01	V$JUNB_02
0.97; N=3	0.78; N=2	1.00; N=3	0.84; N=2	0.87; N=3

Module width: 110

Model score (-p*log10(pval)): 13.44
Wilcoxon p-value (pval): 2.58e-28
Penalty (p): 0.487
Average yes-set score: 7.88
Average no-set score: 6.39
AUC: 0.74
Middle-point: 7.15
False-positive: 29.74%
False-negative: 30.63%

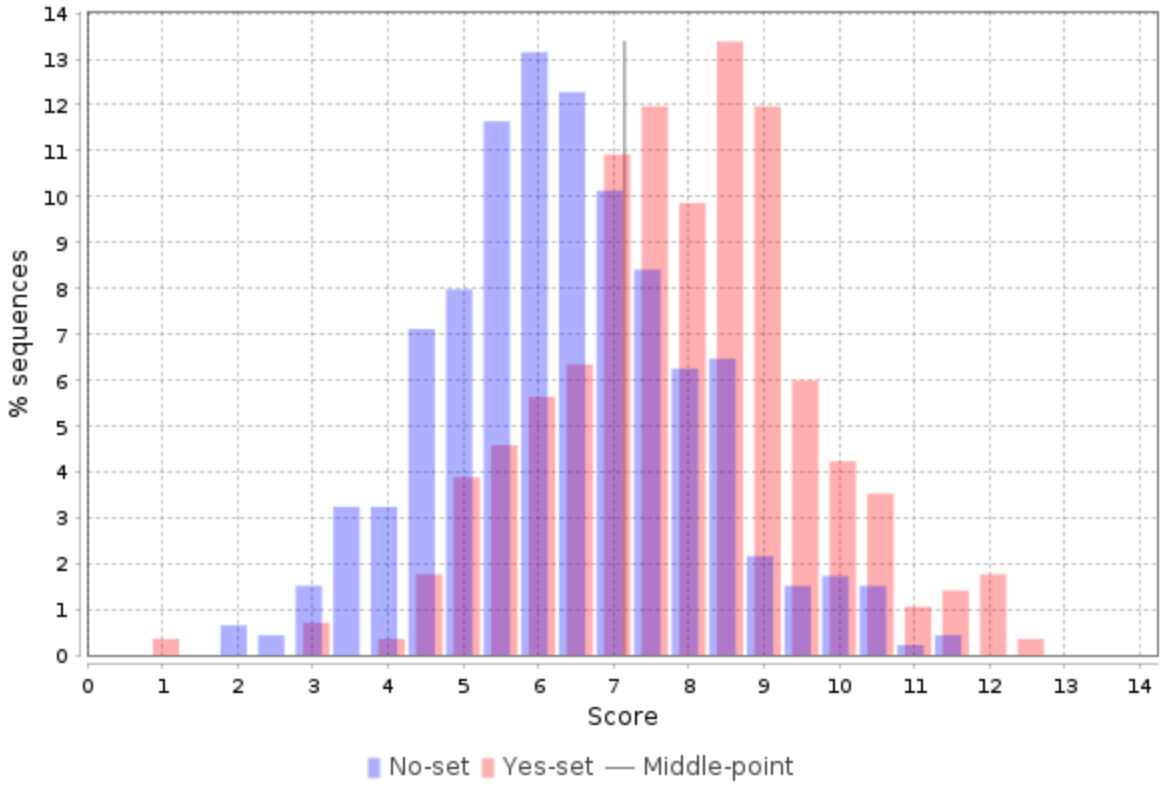


Table 9. List of top ten down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue with identified enhancers in their regulatory regions. ***CMA score*** *- the score of the CMA model of the enhancer identified in the regulatory region.*

See full table →

Ensembl IDs	Gene symbol	Gene description	CMA score	Factor names
ENSG00000156711	MAPK13	mitogen-activated protein kinase 13	13.54	E2A(h),Tal-1(h), CP2(h), PREP-1(h), GKLF(h), POU2F1(h), p53(h), JunB(h)...
ENSG00000173757	STAT5B	signal transducer and activator of transcription 5B	13.27	p53(h), STAT6(h), CP2(h), USF2(h), foxo4(h), JunB(h), POU2F1(h)...
ENSG00000102886	GDPD3	glycerophosphodiester phosphodiesterase domain containing 3	13.26	GKLF(h), foxo4(h), CP2(h), JunB(h), STAT6(h), USF2(h), Tal-1(h)...
ENSG00000196172	ZNF681	zinc finger protein 681	13.18	Tal-1(h), p53(h), PREP-1(h), CP2(h), E2A(h),Tal-1(h), GKLF(h), POU2F1(h)...
ENSG00000142632	ARHGEF19	Rho guanine nucleotide exchange factor 19	12.76	GKLF(h), p53(h), PREP-1(h), E2A(h),Tal-1(h), Tal-1(h), CP2(h), POU2F1(h)...
ENSG00000184574	LPAR5	lysophosphatidic acid receptor 5	12.74	CP2(h), foxo4(h), E2A(h),Tal-1(h), USF2(h), JunB(h), STAT6(h), PREP-1(h)...
ENSG00000124882	EREG	epiregulin	12.62	POU2F1(h), STAT6(h), E2A(h),Tal-1(h), CP2(h), PREP-1(h), JunB(h), p53(h)...
ENSG00000163156	SCNM1	sodium channel modifier 1	12.42	p53(h), JunB(h), USF2(h), foxo4(h), STAT6(h), GKLF(h), PREP-1(h)...
ENSG00000160408	ST6GALNAC6	ST6 N-acetylgalactosaminide alpha-2,6-sialyltransferase 6	12.35	USF2(h), CP2(h), STAT6(h), JunB(h), foxo4(h), E2A(h),Tal-1(h), p53(h)...
ENSG00000201512	SNORA71C	small nucleolar RNA, H/ACA box 71C	12.32	p53(h), GKLF(h), E2A(h),Tal-1(h), PREP-1(h), STAT6(h), Tal-1(h), CP2(h)...

On the basis of the enhancer models we identified transcription factors potentially regulating the ***target genes*** of our interest. We found 16 and 11 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 →

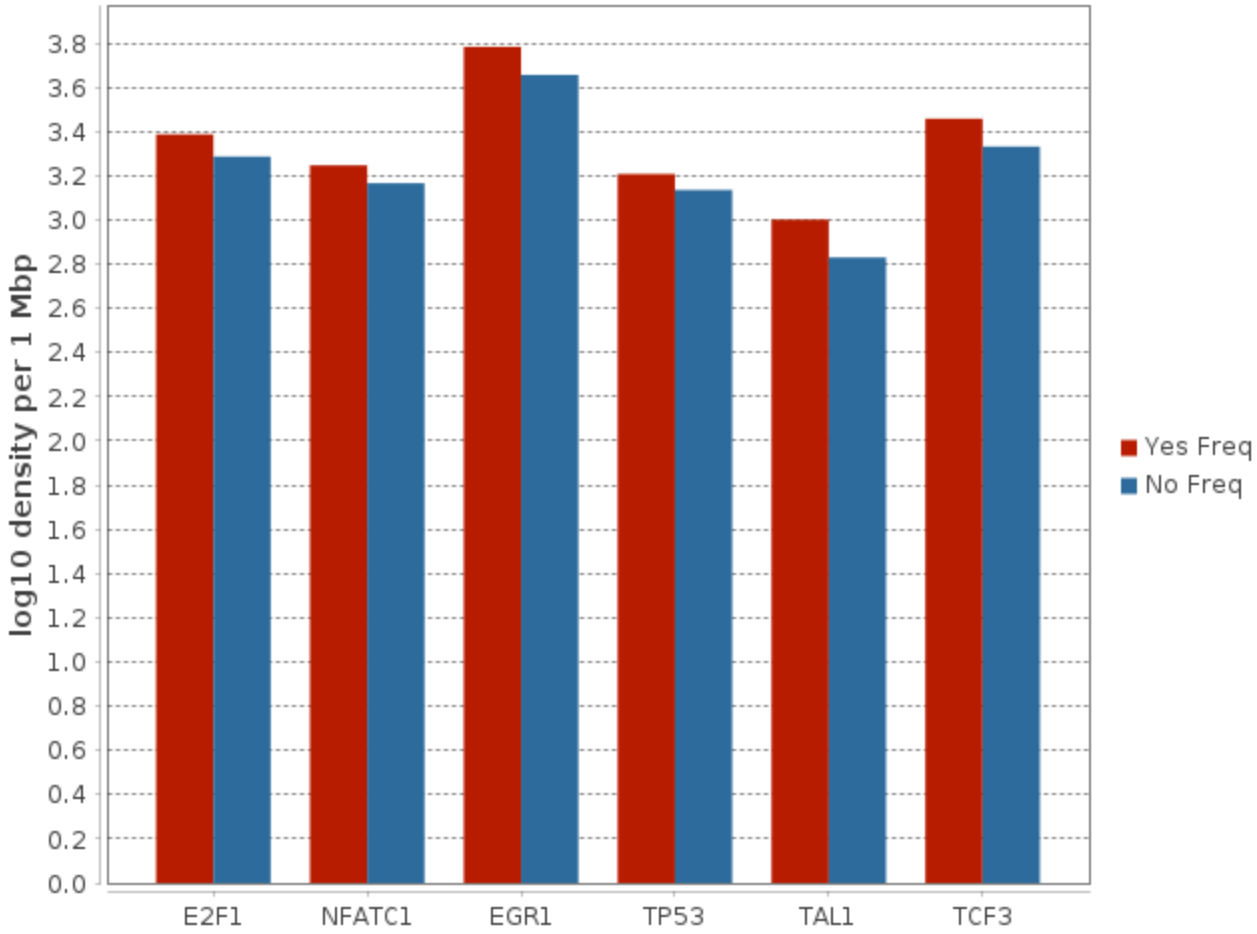
ID	Gene symbol	Gene description	Regulatory score	Yes-No ratio
MO000004274	E2F1	E2F transcription factor 1	3.89	1.26
MO000020760	NFATC1	nuclear factor of activated T cells 1	3.52	1.2
MO000017914	EGR1	early growth response 1	3.42	1.34
MO000032489	TAL1	TAL bHLH transcription factor 1, erythroid differentiation factor	3.14	2.22
MO000002641	CEBPD	CCAAT enhancer binding protein delta	2.87	1.5
MO000125561	KLF4	Kruppel like factor 4	2.6	1.45
MO000018993	SOX9	SRY-box transcription factor 9	2.54	1.49
MO000026306	GCM1	glial cells missing transcription factor 1	2.16	1.37
MO000024921	TCF4	transcription factor 4	1.95	1.46
MO000024797	EGR2	early growth response 2	1.71	1.28

*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	3.44	1.18
MO000032489	TAL1	TAL bHLH transcription factor 1, erythroid differentiation factor	2.87	1.48
MO000032492	TCF3	transcription factor 3	2.25	1.34
MO000025003	POU2F1	POU class 2 homeobox 1	2.22	1.27
MO000000904	FOXO4	forkhead box O4	2.15	1.27
MO000125561	KLF4	Kruppel like factor 4	1.87	11.45
MO000117988	TFCP2	transcription factor CP2	1.79	1.27
MO000031956	STAT6	signal transducer and activator of transcription 6	1.77	3.57
MO000007830	JUNB	JunB proto-oncogene, AP-1 transcription factor subunit	1.75	4.91
MO000025243	USF2	upstream transcription factor 2, c-fos interacting	0	1.29

The following diagram represents the key transcription factors, which were predicted to be potentially regulating differentially expressed genes in the analyzed pathology: E2F1, NFATC1, EGR1, TP53, TAL1 and TCF3.



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	14
MO000329205	Cdk4-isoform1(h):cyclinD1a(h)	CCND1, CDK4	cyclin D1, cyclin dependent kinase 4	2.58	84
MO000030923	ATR(h)	ATR	ATR serine/threonine kinase	2.18	149
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	165
MO000016989	GSK3beta(h)	GSK3B	glycogen synthase kinase 3 beta	1.55	166
MO000032653	MKP3(h)	DUSP6	dual specificity phosphatase 6	2.48	173
MO000032768	ILK(h)	ILK	integrin linked kinase	1.49	184
MO000022058	Lyn(h)	LYN	LYN proto-oncogene, Src family tyrosine kinase	1.2	186
MO000019674	p110alpha(h)	PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	2.32	192
MO000023341	Cdk6(h)	CDK6	cyclin dependent kinase 6	3.09	194

*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
MO000056491	p/CAF(h)	KAT2B	lysine acetyltransferase 2B	-2.74	63
MO000023409	p/CAF(h)	KAT2B	lysine acetyltransferase 2B	-2.74	77
MO000031101	plk3(h)	PLK3	polo like kinase 3	-2.46	105
MO000102190	PTK6-isoform1(h)	PTK6	protein tyrosine kinase 6	-3.89	106
MO000138699	plk3(h)	PLK3	polo like kinase 3	-2.46	147
MO000004672	ERK1(h)	MAPK3	mitogen-activated protein kinase 3	-1.85	170
MO000056883	ERK1-isoform1(h)	MAPK3	mitogen-activated protein kinase 3	-1.85	203
MO000018962	ErbB2(h)	ERBB2	erb-b2 receptor tyrosine kinase 2	-1.16	208
MO000031003	ERK1(h){p}	MAPK3	mitogen-activated protein kinase 3	-1.85	212
MO000128346	kallikrein-6-isoform1(h)	KLK6	kallikrein related peptidase 6	-1.54	220

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.

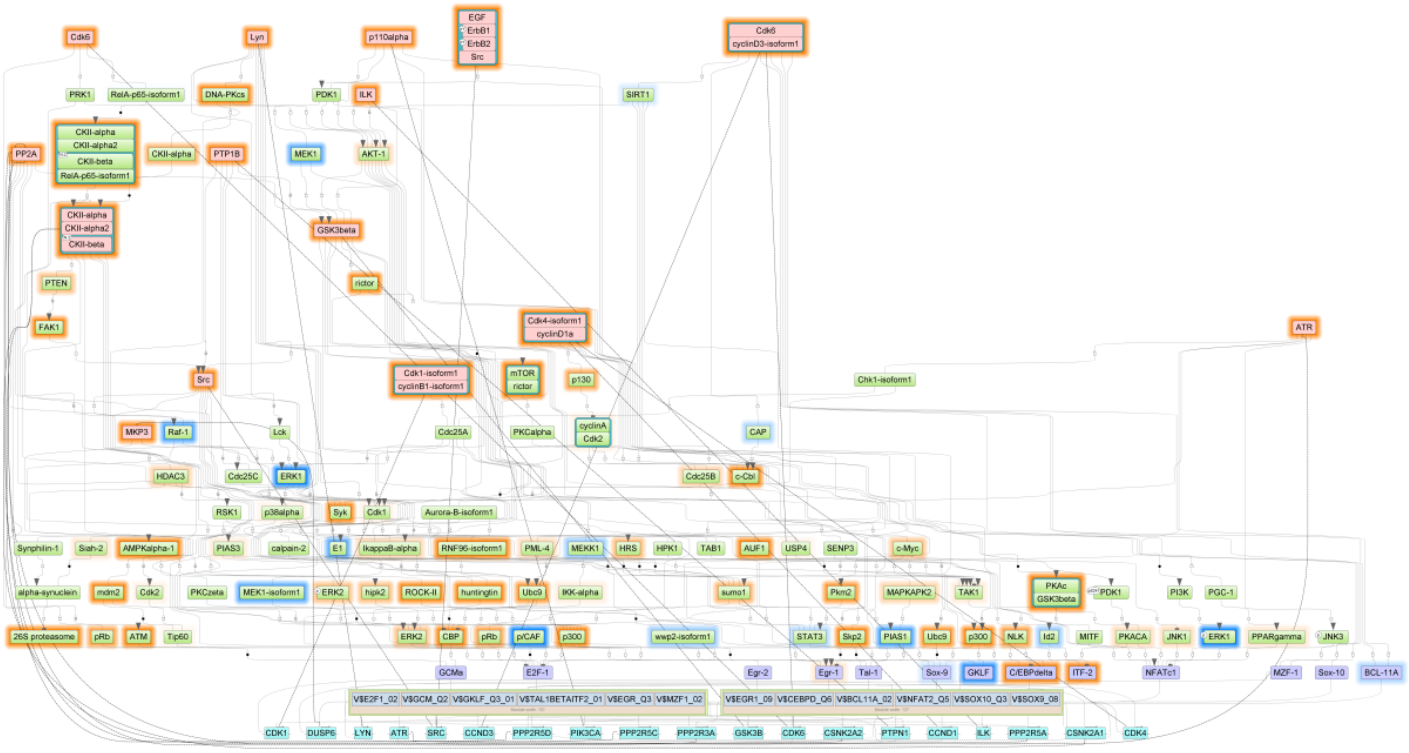


Figure 9. Diagram of intracellular regulatory signal transduction pathways of up-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate molecules,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

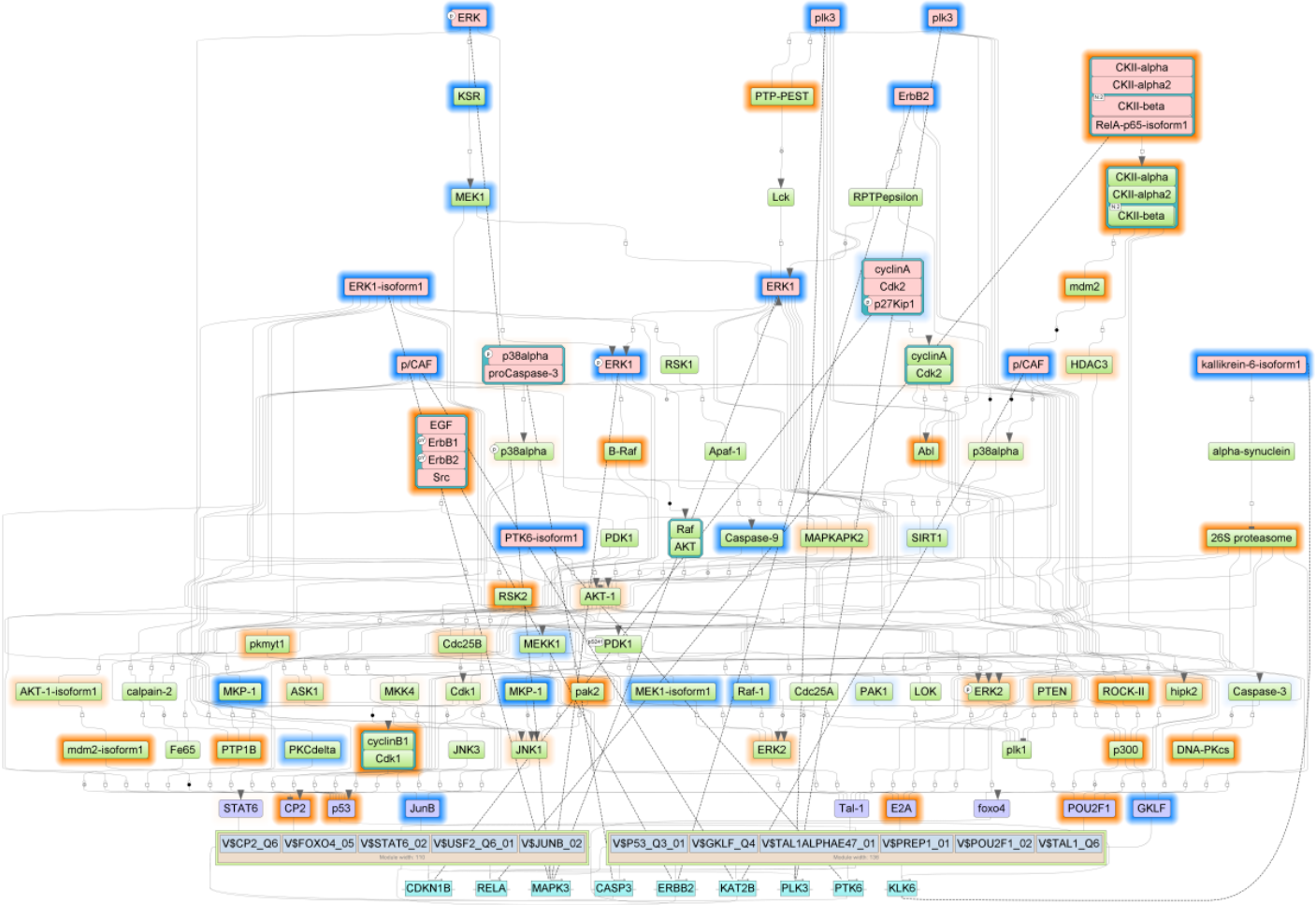


Figure 10. Diagram of intracellular regulatory signal transduction pathways of down-regulated genes in Experiment: Squamous Cell Carcinoma vs. Control: Non-tumour tissue. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate 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
CCND1	cyclin D1	1	2.58	84
PPP2R5C	protein phosphatase 2 regulatory subunit B'gamma	2	1.93	165
CDK6	cyclin dependent kinase 6	4	3.09	194
PIK3CA	phosphatidylinositol-4,5-bisphosphate 3-kinase catalytic subunit alpha	4	2.32	299
PSMA7	proteasome 20S subunit alpha 7	3	1.71	303
LYN	LYN proto-oncogene, Src family tyrosine kinase	4	1.2	351

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.*

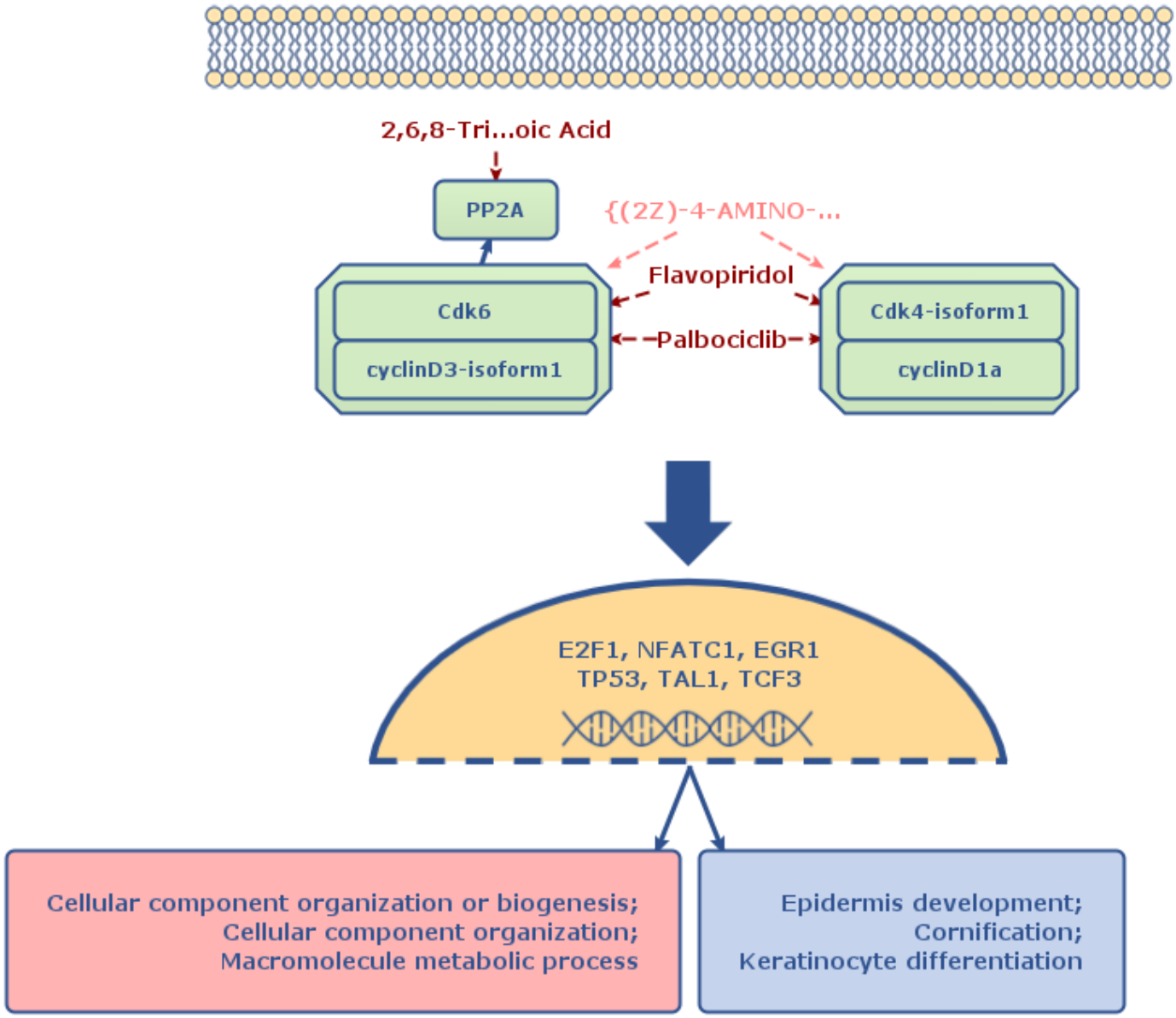
See full table →

Gene symbol	Gene Description	Druggability score	logFC	Total rank
CCND3	cyclin D3	14.31	3.09	14
CCND1	cyclin D1	16.64	2.58	84
DUSP6	dual specificity phosphatase 6	43.57	2.48	173
CDK6	cyclin dependent kinase 6	33.23	3.09	194
ATR	ATR serine/threonine kinase	1.58	2.18	281
ILK	integrin linked kinase	2.94	1.49	293

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:

- Cdk6:cyclinD3-isoform1
- Cdk4-isoform1:cyclinD1a
- PP2A

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: Flavopiridol, Palbociclib, 2,6,8-Trimethyl-3-Amino-9-Benzyl-9-Methoxynonanoic Acid and {(2Z)-4-AMINO-2-[(4-METHOXYPHENYL)IMINO]-2,3-DIHYDRO-1,3-THIAZOL-5-YL}(4-METHOXYPHENYL)METHANONE, 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)
Palbociclib	CDK6, CDK4	18	3	Breast Neoplasms, Neoplasms	small molecule, approved
Arsenic trioxide	CCND1, MAPK1, AKT1	28	2	Leukemia, Leukemia, Myeloid, Leukemia, Promyelocytic, Acute	small molecule, approved, investigational
Dasatinib	SRC, ABL1, YES1	53	4	Leukemia, Leukemia, Lymphoid, Leukemia, Myelogenous, Chronic, BCR-ABL Positive, Leukemia, Myeloid, Precursor Cell Lymphoblastic Leukemia-Lymphoma	small molecule, approved, investigational
Nintedanib	FGFR3, SRC, LYN	58	2	Idiopathic Pulmonary Fibrosis, Pulmonary Fibrosis	small molecule, approved
Vandetanib	VEGFA, EGFR	133	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	36	Leukemia, Myeloid	small molecule, approved
Peginterferon alfa-2a	IFNAR1, IFNAR2	53	HIV Infections, Hemophilia A, Hepatitis, Hepatitis B, Hepatitis B, Chronic, Hepatitis C, Hepatitis C, Chronic...	biotech, approved, investigational
Peginterferon alfa-2b	IFNAR1, IFNAR2	53	Hepatitis, Hepatitis B, Hepatitis B, Chronic, Hepatitis C, Hepatitis C, Chronic, Hepatitis, Chronic	biotech, approved
Interferon beta-1a	IFNAR1, IFNAR2	53	Brain Abscess, Multiple Sclerosis, Multiple Sclerosis, Relapsing-Remitting	biotech, approved, investigational
Interferon beta-1b	IFNAR1, IFNAR2	53	Brain Abscess, Multiple Sclerosis, Multiple Sclerosis, Relapsing-Remitting	biotech, approved

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
{(2Z)-4-AMINO-2-[(4-METHOXYPHENYL)IMINO]-2,3-DIHYDRO-1,3-THIAZOL-5-YL}(4-METHOXYPHENYL)METHANONE	CCND1, CDK6, CCNH, CCND3, CCNB1, CLK1, CCNA2...	2	8.72
3-Bromo-7-Nitroindazole	RPS6KA3, CDK6, GSK3A, CCND3, CCNB1, GSK3B, CDK1...	3	7.3
6-CYCLOHEXYLMETHYLOXY-5-NITROSO-PYRIMIDINE-2,4-DIAMINE	CCND1, CDK6, GSK3A, MTOR, CCNH, CCND3, CCNB1...	4	7.2
2-(2-HYDROXYETHYLAMINO)-6-(3-CHLOROANILINO)-9-ISOPROPYLPURINE	CDK6, SRC, CCND3, CCNB1, CDK5R1, CDK1, CDK4...	5	6.75
O6-CYCLOHEXYLMETHOXY-2-(4'-SULPHAMOYLANILINO) PURINE	CCND1, CDK6, CCNH, CCND3, CCNB1, CCNA2, CDK1...	6	6.58

As the result of drug search we propose the following drugs as most promising candidates for treating the pathology under study: Palbociclib, Bosutinib and {(2Z)-4-AMINO-2-[(4-METHOXYPHENYL)IMINO]-2,3-DIHYDRO-1,3-THIAZOL-5-YL}(4-METHOXYPHENYL)METHANONE. These drugs were selected for acting on the following targets: CDK6, LYN and CCND3, 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. 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:

Palbociclib, Bosutinib and {(2Z)-4-AMINO-2-[(4-METHOXYPHENYL)IMINO]-2,3-DIHYDRO-1,3-THIAZOL-5-YL}(4-METHOXYPHENYL)METHANONE

These drugs were selected for acting on the following targets: CDK6, LYN and CCND3, 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:

Cdk6:cyclinD3-isoform1, Cdk4-isoform1:cyclinD1a and PP2A

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: Flavopiridol, Palbociclib, 2,6,8-Trimethyl-3-Amino-9-Benzyl-9-Methoxynonanoic Acid and {(2Z)-4-AMINO-2-[(4-METHOXYPHENYL)IMINO]-2,3-DIHYDRO-1,3-THIAZOL-5-YL}(4-METHOXYPHENYL)METHANONE. 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.

- Cdk6:cyclinD3-isoform1
- Cdk4-isoform1:cyclinD1a
- PP2A

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 2020.3 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transfac).
The master regulator search uses the TRANSPATH® database (BIOBASE), release 2020.3 (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 2020.3 (https://genexplain.com/humanpsd).
The Ensembl database release Human100.38 (hg38) (http://www.ensembl.org) was used for gene IDs representation and Gene Ontology (GO) (http://geneontology.org) was used for functional classification of the studied gene set.

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

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

Methods for finding master regulators in networks

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

Methods for analysis of pharmaceutical compounds

We seek for the optimal combination of molecular targets (key elements of the regulatory network of the cell) that potentially interact with pharmaceutical compounds from a library of

known drugs and biologically active chemical compounds, using information about known drugs from HumanPSD™ and predicting potential drugs using PASS program.

Method for analysis of known pharmaceutical compounds

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

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

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

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

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

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

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

Method for prediction of pharmaceutical compounds

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

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

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

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

where *M(s)* is the set of activity-mechanisms for the given structure (which passed the chosen

threshold for activity-mechanisms *Pa*); *G(m)* is the set of targets (converted to genes) that corresponds to the given activity-mechanism (m) for the given compound; *pa(m)* is the probability to be active of the activity-mechanism (m), *IAP(g)* is the invariant accuracy of prediction for gene from *G(m)*; *optWeight(g)* is the additional weight multiplier for gene. *T* is set of all targets related to the compound intersected with input list, |*T*| is number of elements in *T*, *AT* and |*AT*| are set set of all targets related to the compound and number of elements in it, *w* is weight multiplier.
"Druggability score" (D-score) is calculated as follows:

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

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

8. References

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

2008;:182-216.
2. Filimonov DA, Poroikov VV. Prognosis of specters of biological activity of organic molecules. *Russian chemical journal.* **2006**;50(2):66-75 (russ)
3. Filimonov D, Poroikov V, Borodina Y, Gloriozova T. Chemical Similarity Assessment Through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *ChemInform.* **1999**;39(4):666-670. doi:10.1002/chin.199940210

Thank you for using the Genome Enhancer!

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

Supplementary material

1. Supplementary table 1 - Up-regulated genes
2. Supplementary table 2 - Down-regulated genes
3. Supplementary table 3 - Detailed report. Composite modules and master regulators (up-regulated genes in 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.