PRKAA2 and INSRR are promising druggable targets for treating Hypertension that control activity of ZBTB33, GTF2B and RELA transcription factors on of differentially expressed genes

Demo User
geneXplain GmbH
info@genexplain.com
Data received on 08/04/2022 ; Run on 15/07/2026 ; Report generated on 15/07/2026

Genome Enhancer release 3.8 (TRANSFAC®, TRANSPATH® and HumanPSD™ release 2026.1)

Abstract

In the present study we applied the software package "Genome Enhancer" to a data set that contains *epigenomics* data. The study is done in the context of *Hypertension*. 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: ZBTB33, GTF2B, PAX4, RELA, SMAD1 and RARA. The subsequent network analysis suggested

- IRR
- AMPKalpha-2(h):AMPK beta(h):AMPK gamma(h):adenosine 5'-monophosphate
- AMPKalpha-2:AMPK beta:AMPK gamma:adenosine 5'-monophosphate

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.

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® [5] is employed together with the TF binding site identification algorithms Match [6] and CMA [7]. The second step involves the signal transduction database TRANSPATH® [8] and special graph search algorithms [10-11] implemented in the software "Genome Enhancer".

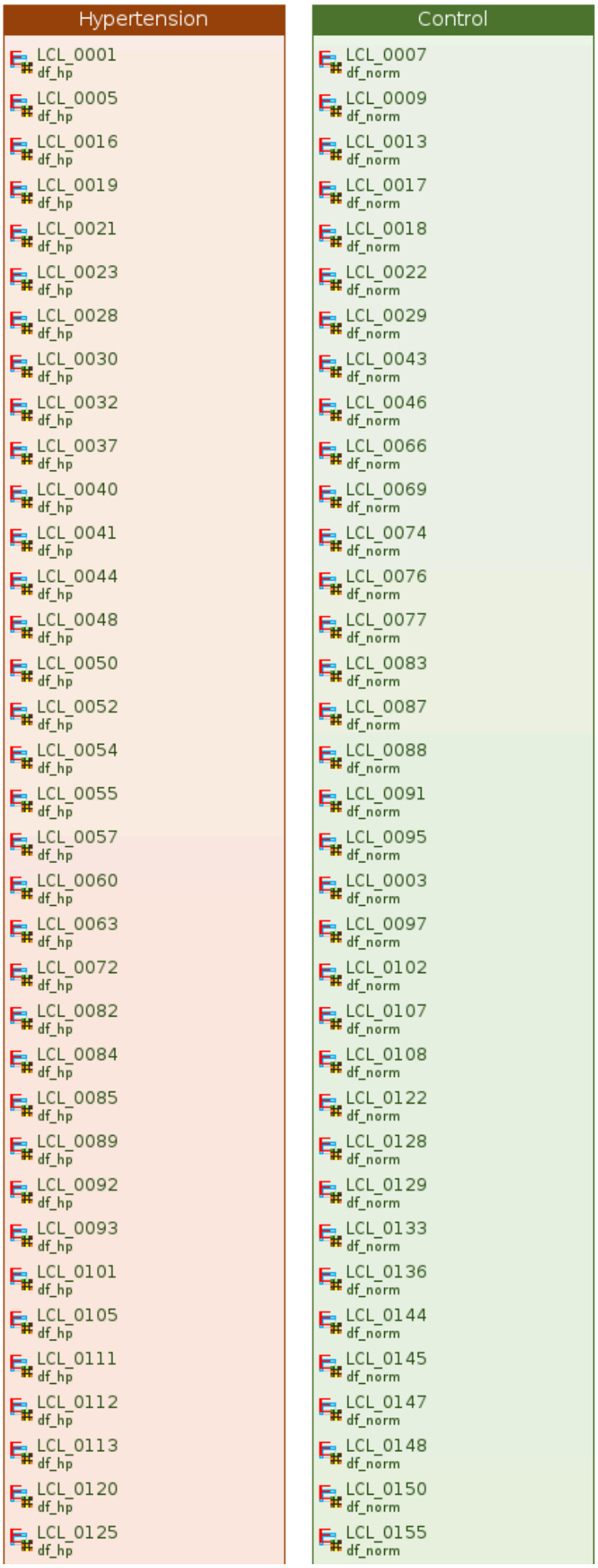
The "upstream analysis" approach has now been extended by a third step that reveals known drugs suitable to inhibit (or activate) the identified molecular targets in the context of the disease under study. This step is performed by using information from HumanPSD™ database [11]. In addition, some known 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 [12-14]. These predictions can be used for the research purposes - for further drug development and drug repurposing initiatives.

2. Data

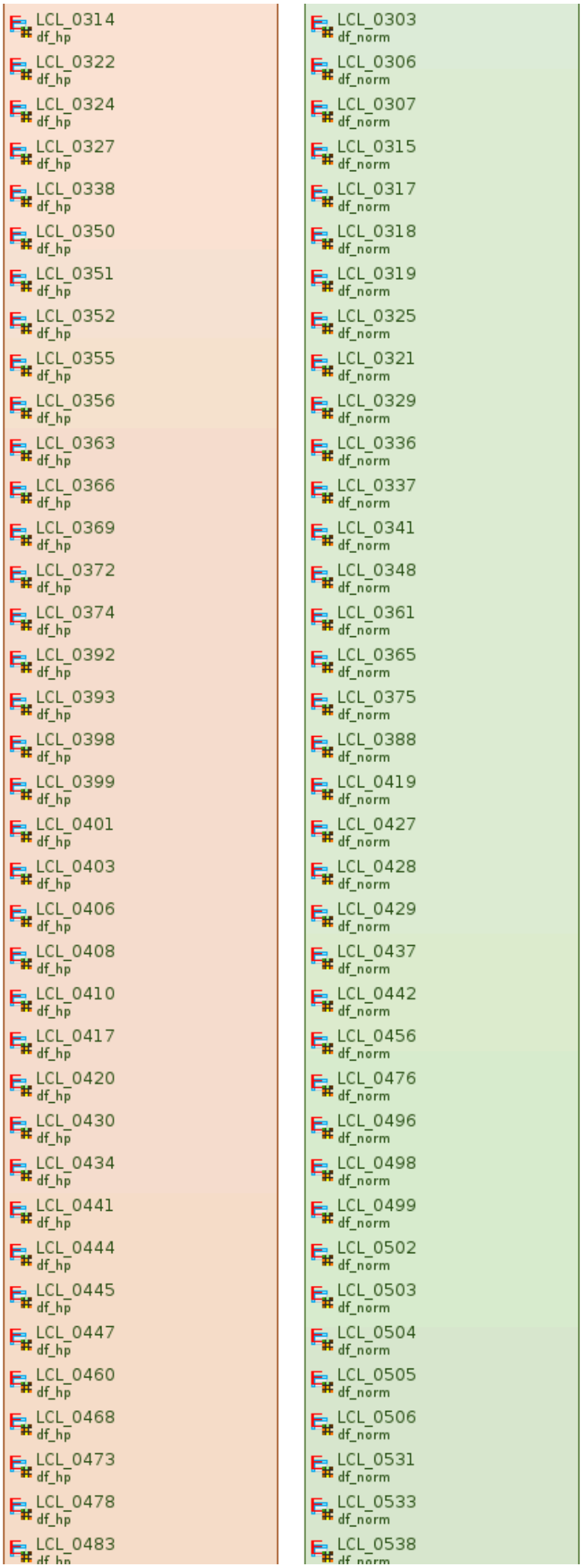
For this study the following experimental data was used:

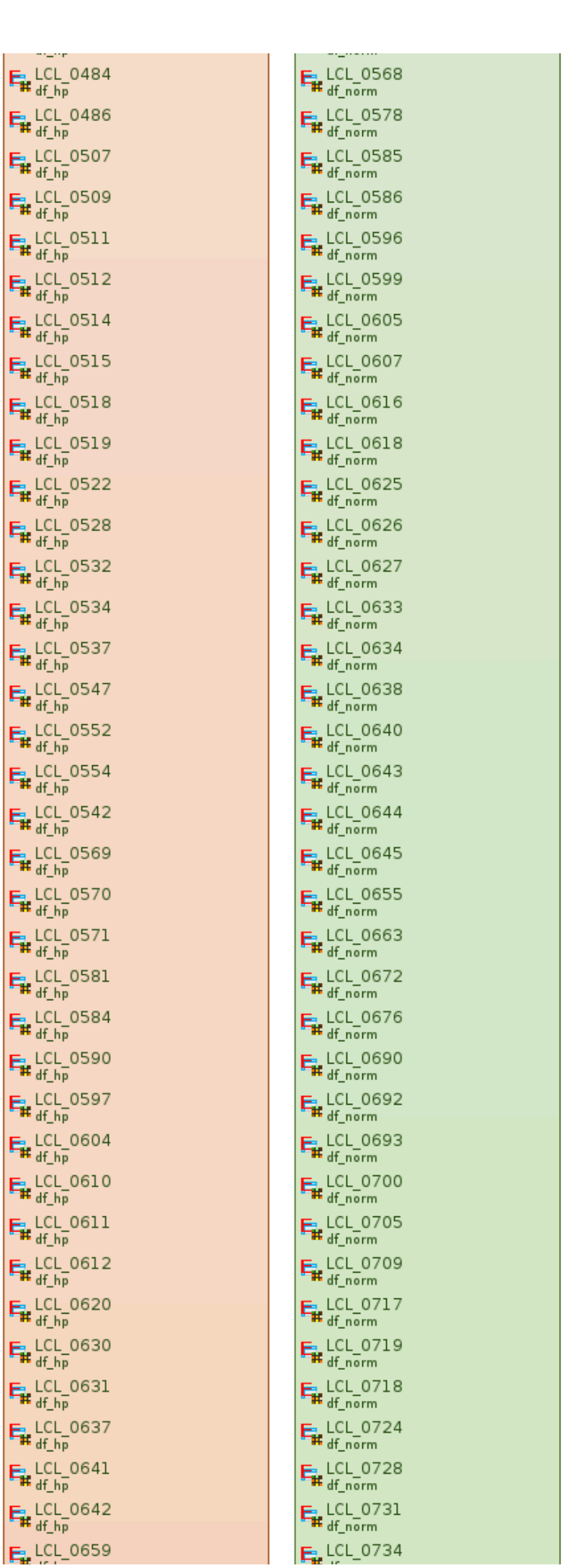
Table 1. Experimental datasets used in the study

File name	Data type
df_hp	Epigenomics
df_norm	Epigenomics


Hypertension
LCL_0001 df_hp
LCL_0005 df_hp
LCL_0016 df_hp
LCL_0019 df_hp
LCL_0021 df_hp
LCL_0023 df_hp
LCL_0028 df_hp
LCL_0030 df_hp
LCL_0032 df_hp
LCL_0037 df_hp
LCL_0040 df_hp
LCL_0041 df_hp
LCL_0044 df_hp
LCL_0048 df_hp
LCL_0050 df_hp
LCL_0052 df_hp
LCL_0054 df_hp
LCL_0055 df_hp
LCL_0057 df_hp
LCL_0060 df_hp
LCL_0063 df_hp
LCL_0072 df_hp
LCL_0082 df_hp
LCL_0084 df_hp
LCL_0085 df_hp
LCL_0089 df_hp
LCL_0092 df_hp
LCL_0093 df_hp
LCL_0101 df_hp
LCL_0105 df_hp
LCL_0111 df_hp
LCL_0112 df_hp
LCL_0113 df_hp
LCL_0120 df_hp
LCL_0125 df_hp
Control
LCL_0007 df_norm
LCL_0009 df_norm
LCL_0013 df_norm
LCL_0017 df_norm
LCL_0018 df_norm
LCL_0022 df_norm
LCL_0029 df_norm
LCL_0043 df_norm
LCL_0046 df_norm
LCL_0066 df_norm
LCL_0069 df_norm
LCL_0074 df_norm
LCL_0076 df_norm
LCL_0077 df_norm
LCL_0083 df_norm
LCL_0087 df_norm
LCL_0088 df_norm
LCL_0091 df_norm
LCL_0095 df_norm
LCL_0003 df_norm
LCL_0097 df_norm
LCL_0102 df_norm
LCL_0107 df_norm
LCL_0108 df_norm
LCL_0122 df_norm
LCL_0128 df_norm
LCL_0129 df_norm
LCL_0133 df_norm
LCL_0136 df_norm
LCL_0144 df_norm
LCL_0145 df_norm
LCL_0147 df_norm
LCL_0148 df_norm
LCL_0150 df_norm
LCL_0155 df_norm

LCL_0131 df_hp	LCL_0158 df_norm
LCL_0141 df_hp	LCL_0159 df_norm
LCL_0162 df_hp	LCL_0135 df_norm
LCL_0171 df_hp	LCL_0164 df_norm
LCL_0175 df_hp	LCL_0165 df_norm
LCL_0177 df_hp	LCL_0167 df_norm
LCL_0189 df_hp	LCL_0168 df_norm
LCL_0190 df_hp	LCL_0174 df_norm
LCL_0192 df_hp	LCL_0176 df_norm
LCL_0193 df_hp	LCL_0179 df_norm
LCL_0194 df_hp	LCL_0183 df_norm
LCL_0198 df_hp	LCL_0188 df_norm
LCL_0199 df_hp	LCL_0196 df_norm
LCL_0200 df_hp	LCL_0197 df_norm
LCL_0205 df_hp	LCL_0204 df_norm
LCL_0211 df_hp	LCL_0209 df_norm
LCL_0213 df_hp	LCL_0219 df_norm
LCL_0214 df_hp	LCL_0220 df_norm
LCL_0217 df_hp	LCL_0221 df_norm
LCL_0233 df_hp	LCL_0223 df_norm
LCL_0235 df_hp	LCL_0224 df_norm
LCL_0236 df_hp	LCL_0232 df_norm
LCL_0237 df_hp	LCL_0234 df_norm
LCL_0241 df_hp	LCL_0244 df_norm
LCL_0246 df_hp	LCL_0245 df_norm
LCL_0255 df_hp	LCL_0253 df_norm
LCL_0256 df_hp	LCL_0258 df_norm
LCL_0257 df_hp	LCL_0260 df_norm
LCL_0262 df_hp	LCL_0264 df_norm
LCL_0265 df_hp	LCL_0274 df_norm
LCL_0266 df_hp	LCL_0281 df_norm
LCL_0272 df_hp	LCL_0290 df_norm
LCL_0280 df_hp	LCL_0291 df_norm
LCL_0285 df_hp	LCL_0293 df_norm
LCL_0298 df_hp	LCL_0294 df_norm
LCL_0299 df_hp	LCL_0295 df_norm
LCL_0310 df_hp	LCL_0301 df_norm


LCL_0314
df_hp
LCL_0322
df_hp
LCL_0324
df_hp
LCL_0327
df_hp
LCL_0338
df_hp
LCL_0350
df_hp
LCL_0351
df_hp
LCL_0352
df_hp
LCL_0355
df_hp
LCL_0356
df_hp
LCL_0363
df_hp
LCL_0366
df_hp
LCL_0369
df_hp
LCL_0372
df_hp
LCL_0374
df_hp
LCL_0392
df_hp
LCL_0393
df_hp
LCL_0398
df_hp
LCL_0399
df_hp
LCL_0401
df_hp
LCL_0403
df_hp
LCL_0406
df_hp
LCL_0408
df_hp
LCL_0410
df_hp
LCL_0417
df_hp
LCL_0420
df_hp
LCL_0430
df_hp
LCL_0434
df_hp
LCL_0441
df_hp
LCL_0444
df_hp
LCL_0445
df_hp
LCL_0447
df_hp
LCL_0460
df_hp
LCL_0468
df_hp
LCL_0473
df_hp
LCL_0478
df_hp
LCL_0483
df_hp
LCL_0303
df_norm
LCL_0306
df_norm
LCL_0307
df_norm
LCL_0315
df_norm
LCL_0317
df_norm
LCL_0318
df_norm
LCL_0319
df_norm
LCL_0325
df_norm
LCL_0321
df_norm
LCL_0329
df_norm
LCL_0336
df_norm
LCL_0337
df_norm
LCL_0341
df_norm
LCL_0348
df_norm
LCL_0361
df_norm
LCL_0365
df_norm
LCL_0375
df_norm
LCL_0388
df_norm
LCL_0419
df_norm
LCL_0427
df_norm
LCL_0428
df_norm
LCL_0429
df_norm
LCL_0437
df_norm
LCL_0442
df_norm
LCL_0456
df_norm
LCL_0476
df_norm
LCL_0496
df_norm
LCL_0498
df_norm
LCL_0499
df_norm
LCL_0502
df_norm
LCL_0503
df_norm
LCL_0504
df_norm
LCL_0505
df_norm
LCL_0506
df_norm
LCL_0531
df_norm
LCL_0533
df_norm
LCL_0538
df_norm


LCL_0484 df_hp
LCL_0486 df_hp
LCL_0507 df_hp
LCL_0509 df_hp
LCL_0511 df_hp
LCL_0512 df_hp
LCL_0514 df_hp
LCL_0515 df_hp
LCL_0518 df_hp
LCL_0519 df_hp
LCL_0522 df_hp
LCL_0528 df_hp
LCL_0532 df_hp
LCL_0534 df_hp
LCL_0537 df_hp
LCL_0547 df_hp
LCL_0552 df_hp
LCL_0554 df_hp
LCL_0542 df_hp
LCL_0569 df_hp
LCL_0570 df_hp
LCL_0571 df_hp
LCL_0581 df_hp
LCL_0584 df_hp
LCL_0590 df_hp
LCL_0597 df_hp
LCL_0604 df_hp
LCL_0610 df_hp
LCL_0611 df_hp
LCL_0612 df_hp
LCL_0620 df_hp
LCL_0630 df_hp
LCL_0631 df_hp
LCL_0637 df_hp
LCL_0641 df_hp
LCL_0642 df_hp
LCL_0659
LCL_0568 df_norm
LCL_0578 df_norm
LCL_0585 df_norm
LCL_0586 df_norm
LCL_0596 df_norm
LCL_0599 df_norm
LCL_0605 df_norm
LCL_0607 df_norm
LCL_0616 df_norm
LCL_0618 df_norm
LCL_0625 df_norm
LCL_0626 df_norm
LCL_0627 df_norm
LCL_0633 df_norm
LCL_0634 df_norm
LCL_0638 df_norm
LCL_0640 df_norm
LCL_0643 df_norm
LCL_0644 df_norm
LCL_0645 df_norm
LCL_0655 df_norm
LCL_0663 df_norm
LCL_0672 df_norm
LCL_0676 df_norm
LCL_0690 df_norm
LCL_0692 df_norm
LCL_0693 df_norm
LCL_0700 df_norm
LCL_0705 df_norm
LCL_0709 df_norm
LCL_0717 df_norm
LCL_0719 df_norm
LCL_0718 df_norm
LCL_0724 df_norm
LCL_0728 df_norm
LCL_0731 df_norm
LCL_0734

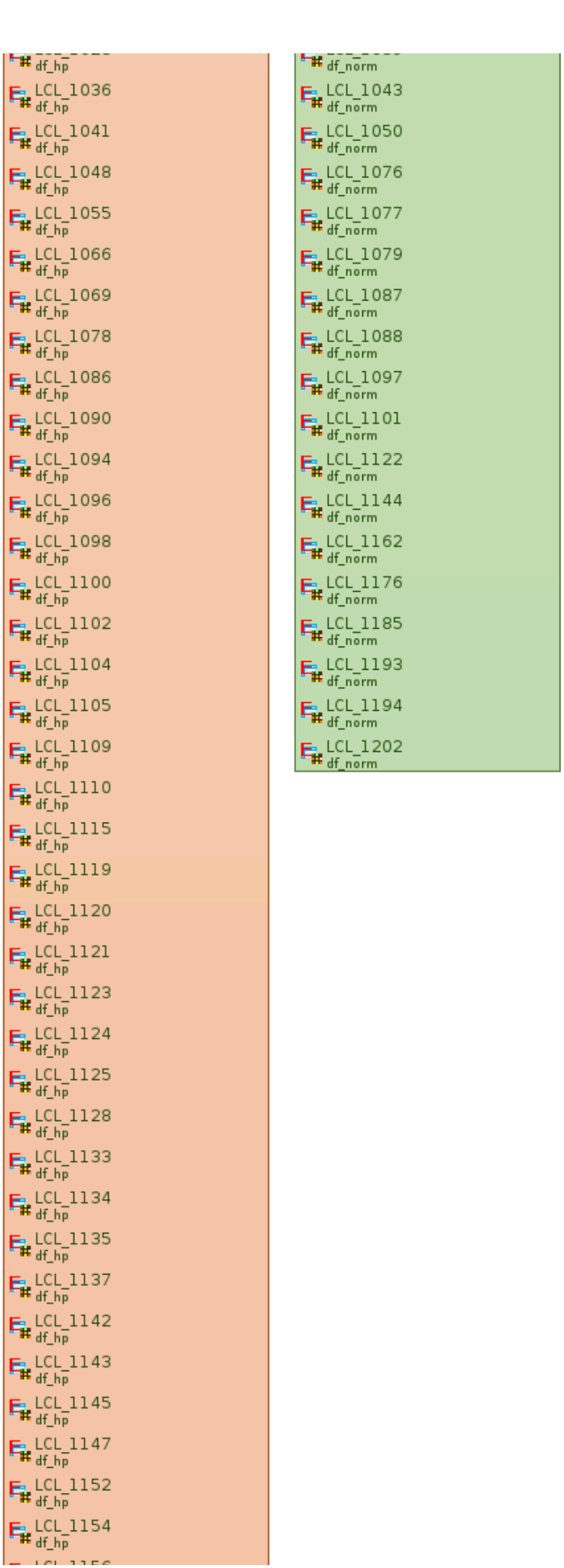
df_hp

LCL_0660 df_hp
LCL_0665 df_hp
LCL_0673 df_hp
LCL_0677 df_hp
LCL_0678 df_hp
LCL_0687 df_hp
LCL_0691 df_hp
LCL_0694 df_hp
LCL_0697 df_hp
LCL_0701 df_hp
LCL_0702 df_hp
LCL_0707 df_hp
LCL_0711 df_hp
LCL_0723 df_hp
LCL_0726 df_hp
LCL_0727 df_hp
LCL_0732 df_hp
LCL_0733 df_hp
LCL_0737 df_hp
LCL_0741 df_hp
LCL_0742 df_hp
LCL_0743 df_hp
LCL_0755 df_hp
LCL_0773 df_hp
LCL_0774 df_hp
LCL_0775 df_hp
LCL_0778 df_hp
LCL_0783 df_hp
LCL_0785 df_hp
LCL_0795 df_hp
LCL_0796 df_hp
LCL_0800 df_hp
LCL_0808 df_hp
LCL_0810 df_hp
LCL_0814 df_hp
LCL_0817 df_hp
LCL_0836

df_norm

LCL_0738 df_norm
LCL_0739 df_norm
LCL_0745 df_norm
LCL_0749 df_norm
LCL_0754 df_norm
LCL_0756 df_norm
LCL_0757 df_norm
LCL_0762 df_norm
LCL_0764 df_norm
LCL_0765 df_norm
LCL_0769 df_norm
LCL_0780 df_norm
LCL_0784 df_norm
LCL_0788 df_norm
LCL_0789 df_norm
LCL_0793 df_norm
LCL_0797 df_norm
LCL_0799 df_norm
LCL_0805 df_norm
LCL_0809 df_norm
LCL_0811 df_norm
LCL_0813 df_norm
LCL_0821 df_norm
LCL_0823 df_norm
LCL_0824 df_norm
LCL_0826 df_norm
LCL_0827 df_norm
LCL_0828 df_norm
LCL_0831 df_norm
LCL_0834 df_norm
LCL_0835 df_norm
LCL_0839 df_norm
LCL_0847 df_norm
LCL_0848 df_norm
LCL_0853 df_norm
LCL_0854 df_norm
LCL_0858

df_hp	df_norm
LCL_0837	LCL_0863
LCL_0866	LCL_0865
LCL_0869	LCL_0870
LCL_0871	LCL_0874
LCL_0873	LCL_0879
LCL_0877	LCL_0880
LCL_0893	LCL_0883
LCL_0903	LCL_0884
LCL_0905	LCL_0887
LCL_0907	LCL_0888
LCL_0911	LCL_0889
LCL_0912	LCL_0890
LCL_0916	LCL_0891
LCL_0923	LCL_0898
LCL_0935	LCL_0904
LCL_0937	LCL_0909
LCL_0942	LCL_0914
LCL_0950	LCL_0917
LCL_0958	LCL_0920
LCL_0959	LCL_0921
LCL_0960	LCL_0922
LCL_0961	LCL_0926
LCL_0963	LCL_0928
LCL_0967	LCL_0943
LCL_0975	LCL_0954
LCL_0978	LCL_0964
LCL_0980	LCL_0968
LCL_0984	LCL_0969
LCL_0987	LCL_0982
LCL_0989	LCL_0990
LCL_0999	LCL_0994
LCL_1002	LCL_1005
LCL_1006	LCL_1021
LCL_1007	LCL_1032
LCL_1025	LCL_1034
LCL_1027	LCL_1035
LCL_1028	LCL_1039


df_hp
LCL_1036
df_hp
LCL_1041
df_hp
LCL_1048
df_hp
LCL_1055
df_hp
LCL_1066
df_hp
LCL_1069
df_hp
LCL_1078
df_hp
LCL_1086
df_hp
LCL_1090
df_hp
LCL_1094
df_hp
LCL_1096
df_hp
LCL_1098
df_hp
LCL_1100
df_hp
LCL_1102
df_hp
LCL_1104
df_hp
LCL_1105
df_hp
LCL_1109
df_hp
LCL_1110
df_hp
LCL_1115
df_hp
LCL_1119
df_hp
LCL_1120
df_hp
LCL_1121
df_hp
LCL_1123
df_hp
LCL_1124
df_hp
LCL_1125
df_hp
LCL_1128
df_hp
LCL_1133
df_hp
LCL_1134
df_hp
LCL_1135
df_hp
LCL_1137
df_hp
LCL_1142
df_hp
LCL_1143
df_hp
LCL_1145
df_hp
LCL_1147
df_hp
LCL_1152
df_hp
LCL_1154
df_hp
df_norm
LCL_1043
df_norm
LCL_1050
df_norm
LCL_1076
df_norm
LCL_1077
df_norm
LCL_1079
df_norm
LCL_1087
df_norm
LCL_1088
df_norm
LCL_1097
df_norm
LCL_1101
df_norm
LCL_1122
df_norm
LCL_1144
df_norm
LCL_1162
df_norm
LCL_1176
df_norm
LCL_1185
df_norm
LCL_1193
df_norm
LCL_1194
df_norm
LCL_1202
df_norm

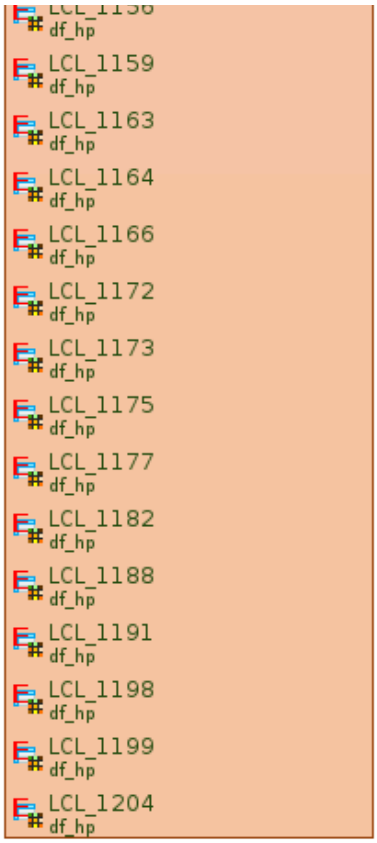


Figure 1. Annotation diagram of experimental data used in this study. With the colored boxes we show those sub-categories of the data that are compared in our analysis.

3. Results

We have compared the following conditions: Hypertension *versus* Control.

3.1. Identification of target genes

In the first step of the analysis ***target genes*** were identified from the uploaded experimental data. The most highly methylated genes were used as target genes.

Table 2. Top ten highly methylated genes in Hypertension vs. Control.
See full table →

ID	Gene description	Gene symbol	Gene schematic representation	Number of methylation sites	Methylation sites in exons	Methylation sites in 5' region
ENSG00000204956	protocadherin gamma subfamily A, 1	PCDHGA1		27	2	0
ENSG00000081853	protocadherin gamma subfamily A, 2	PCDHGA2		26	4	0
ENSG00000152977	Zic family member 1	ZIC1		25	7	3
ENSG00000254245	protocadherin gamma subfamily A, 3	PCDHGA3		24	3	0
ENSG00000254221	protocadherin gamma subfamily B, 1	PCDHGB1		23	4	0
ENSG00000078900	tumor protein p73	TP73		22	5	6
ENSG00000204969	protocadherin alpha 2	PCDHA2		22	1	0
ENSG00000204970	protocadherin alpha 1	PCDHA1		22	0	0
ENSG00000204965	protocadherin alpha 5	PCDHA5		21	0	1
ENSG00000204967	protocadherin alpha 4	PCDHA4		21	0	0

Table 3. Top ten low methylated genes in Hypertension vs. Control.
See full table →

ID	Gene description	Gene symbol	Gene schematic representation	Number of methylation sites	Methylation sites in exons	Methylation sites in 5' region
ENSG00000001630	cytochrome P450 family 51 subfamily A member 1	CYP51A1		1	0	1
ENSG00000004660	calcium/calmodulin dependent protein kinase kinase 1	CAMKK1		1	0	0
ENSG00000004809	solute carrier family 22 member 16	SLC22A16		1	1	0
ENSG00000005001	serine protease 22	PRSS22		1	1	0
ENSG00000005059	mitochondrial calcium uniporter dominant negative subunit beta	MCUB		1	0	0
ENSG00000005108	thrombospondin type 1 domain containing 7A	THSD7A		1	1	0
ENSG00000005884	integrin subunit alpha 3	ITGA3		1	1	0
ENSG00000006071	ATP binding cassette subfamily C member 8	ABCC8		1	1	0
ENSG00000006377	distal-less homeobox 6	DLX6		1	0	1
ENSG00000006576	putative homeodomain transcription factor 2	PHTF2		1	1	0

3.2. Functional classification of genes

A functional analysis of differentially expressed genes was done by mapping the 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 2-7 show the most significant categories.

Highly methylated genes in Hypertension vs. Control:

300 top methylated genes were taken for the mapping.

GO (biological process)

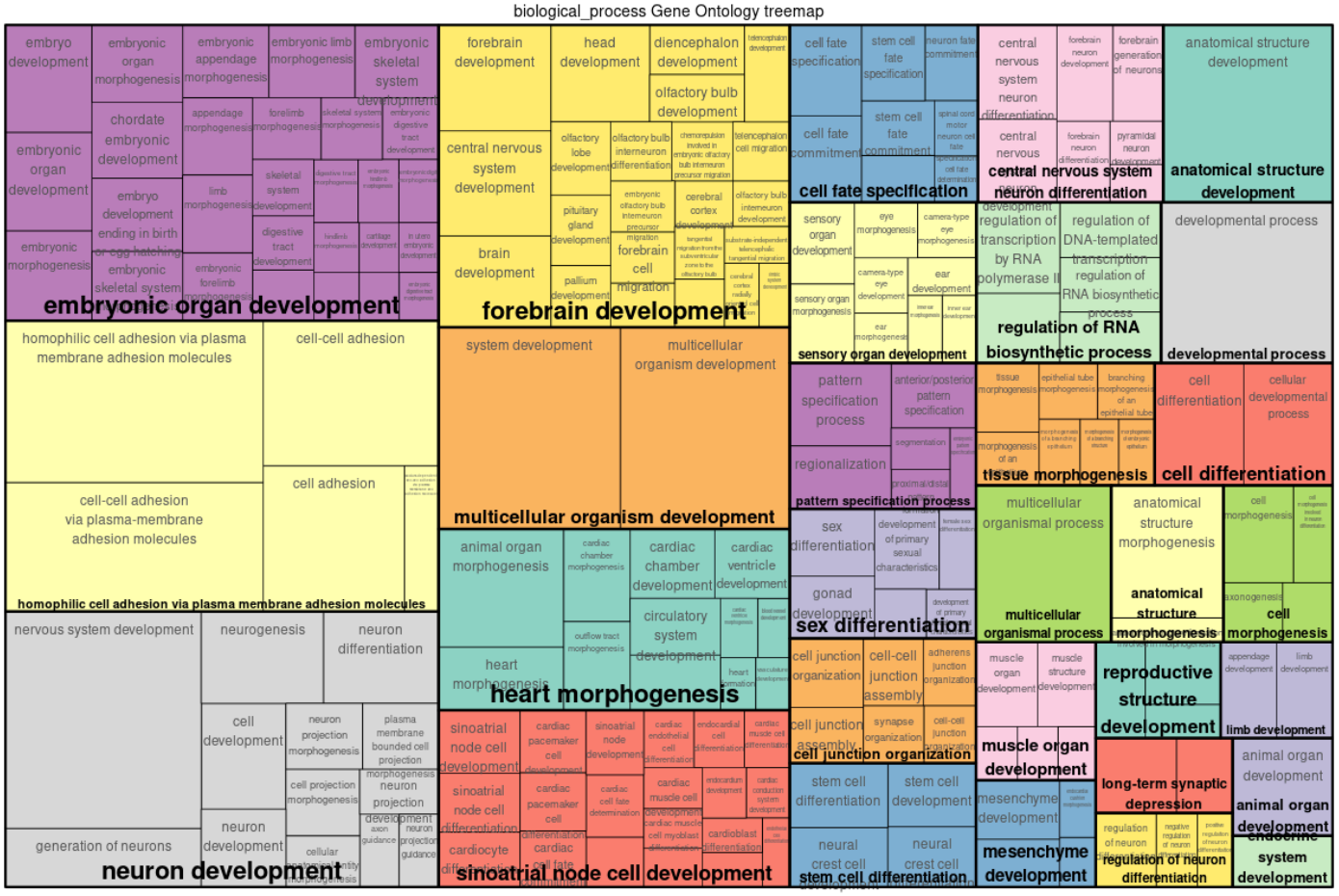


Figure 2. Enriched GO (biological process) of highly methylated genes in Hypertension vs. Control.
Full classification →

TRANSPATH® Pathways (2026.1)

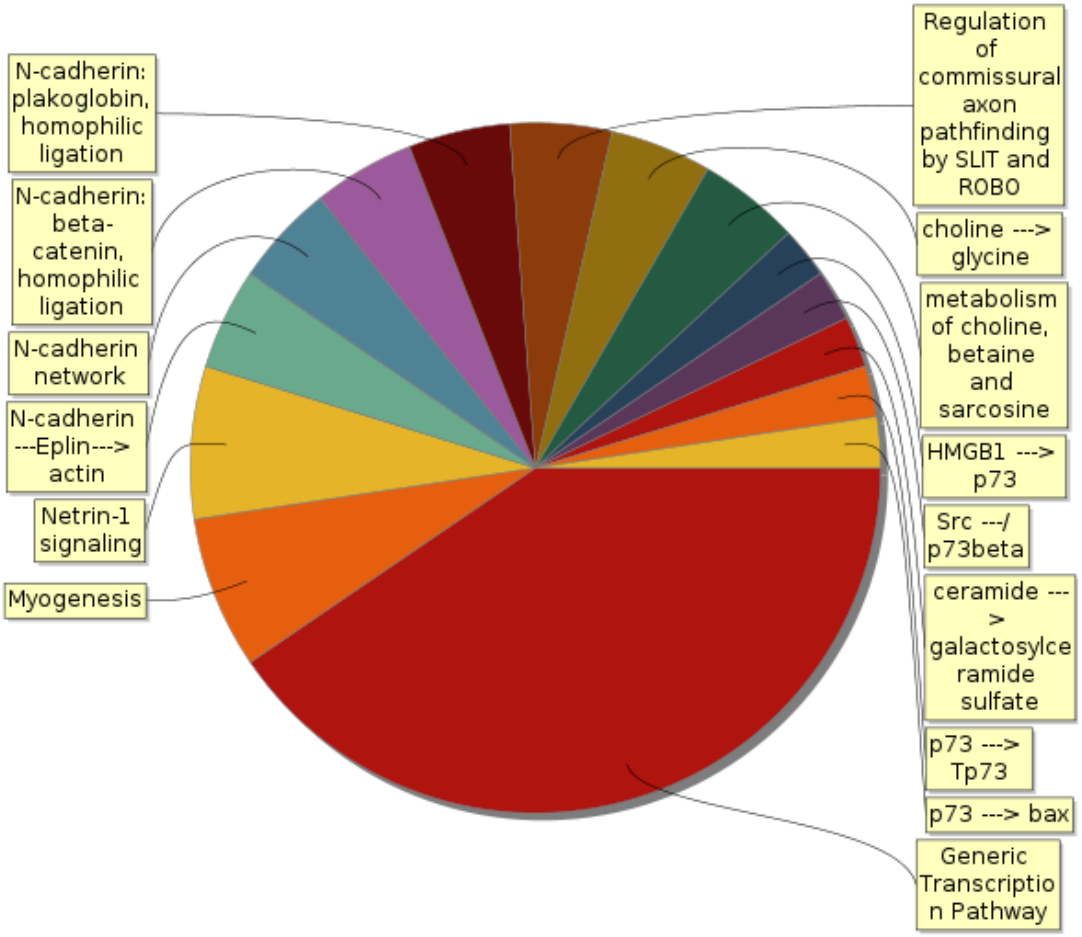


Figure 3. Enriched TRANSPATH® Pathways (2026.1) of highly methylated genes in Hypertension vs. Control.
Full classification →

HumanPSD(TM) disease (2026.1)

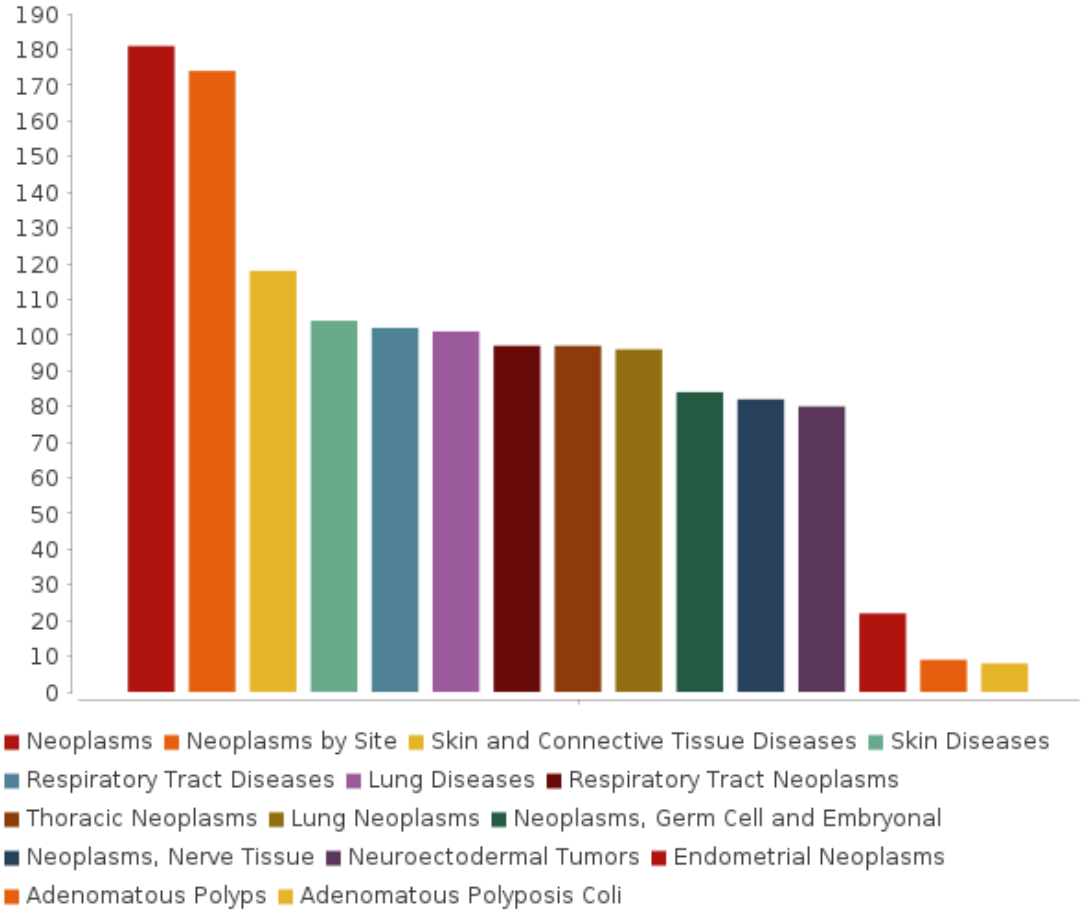


Figure 4. Enriched HumanPSD(TM) disease (2026.1) of highly methylated genes in Hypertension vs. Control. The size of the bars correspond to the number of biomarkers of the given disease found among the input set.

Full classification →

Low methylated genes in Hypertension vs. Control:

300 less methylated genes were taken for the mapping.

GO (biological process)

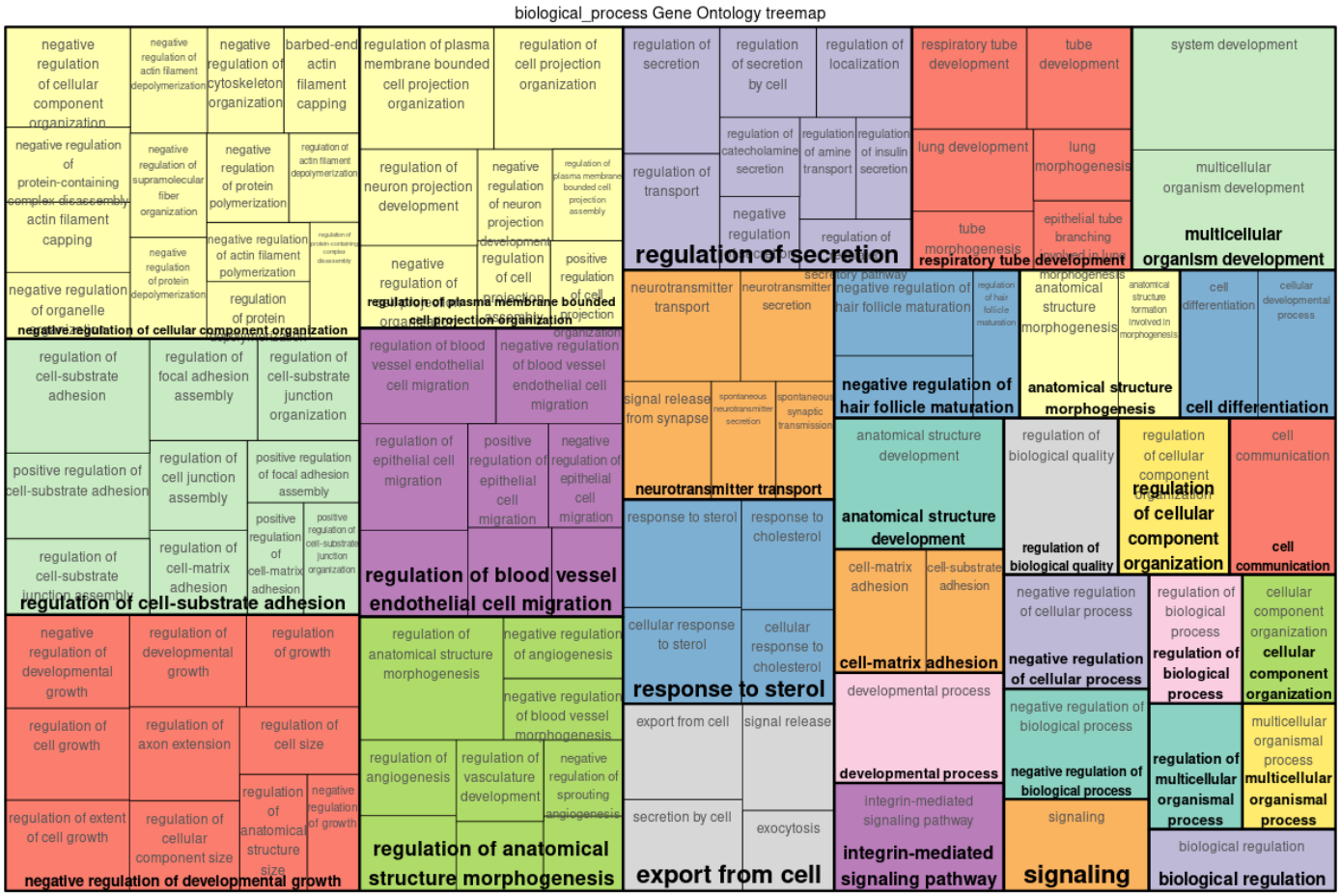


Figure 5. Enriched GO (biological process) of low methylated genes in Hypertension vs. Control.

Full classification →

TRANSPATH® Pathways (2026.1)

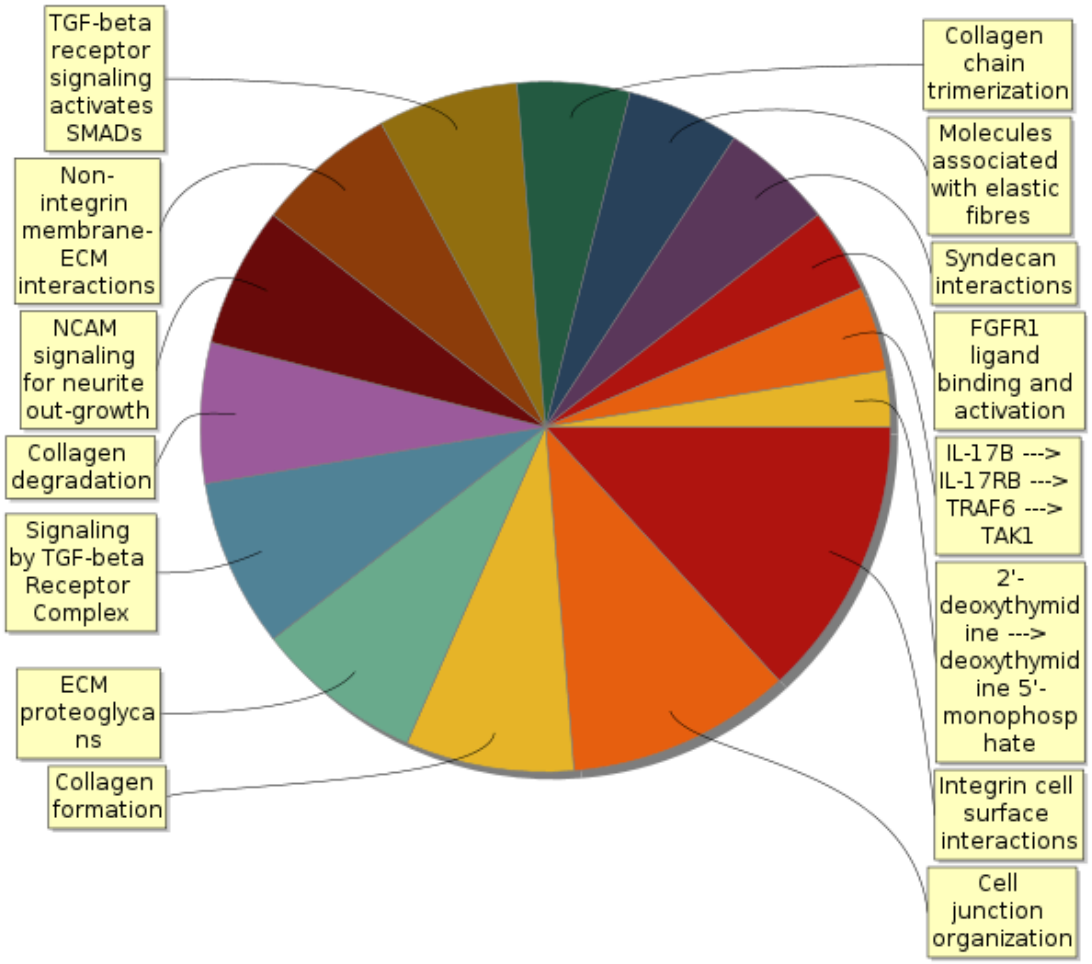


Figure 6. Enriched TRANSPATH® Pathways (2026.1) of low methylated genes in Hypertension vs. Control.
Full classification →

HumanPSD(TM) disease (2026.1)

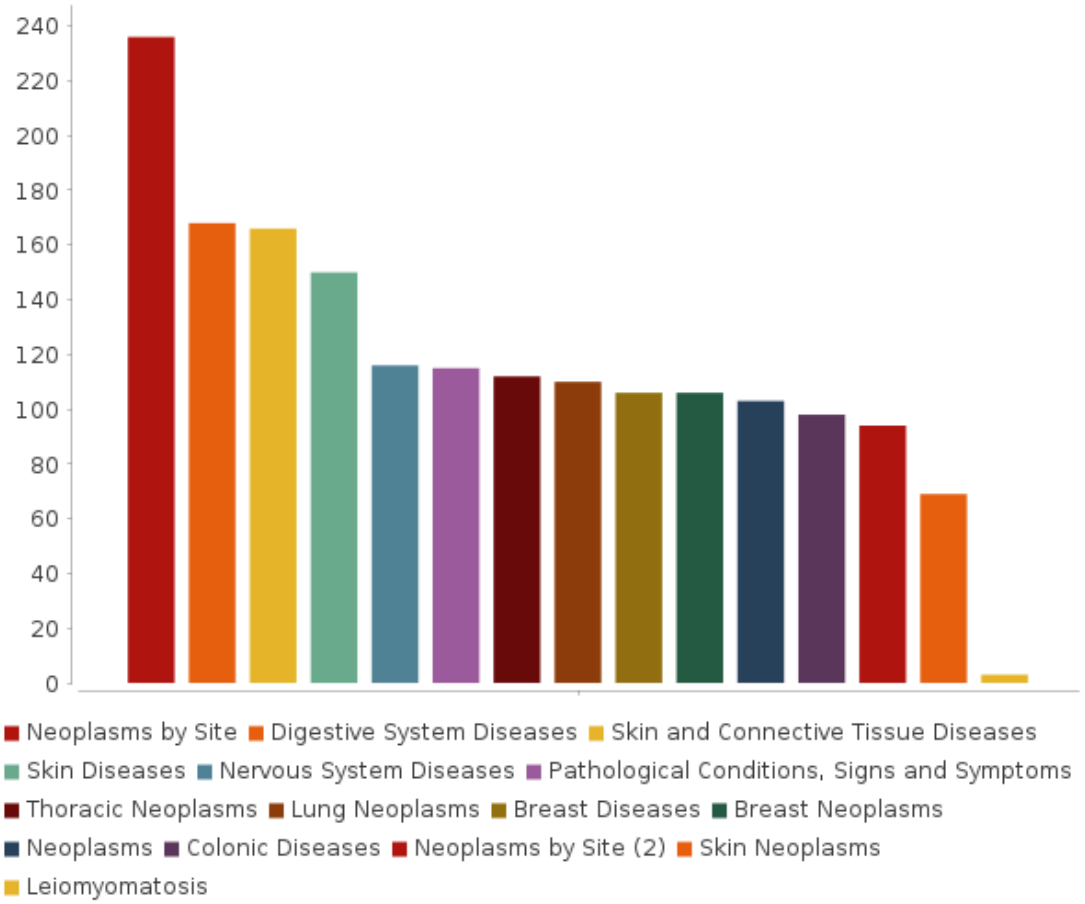
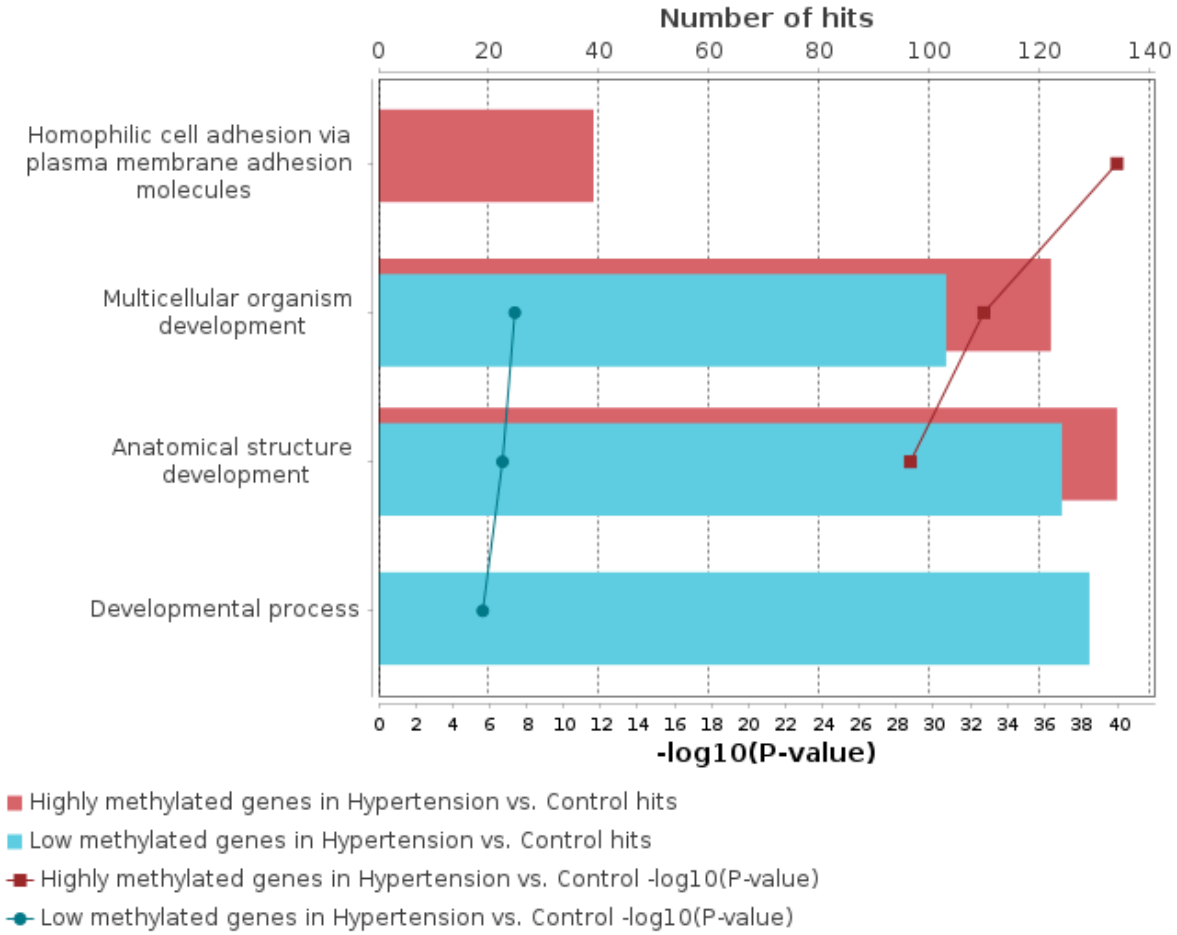


Figure 7. Enriched HumanPSD(TM) disease (2026.1) of low methylated genes in Hypertension vs. Control. The size of the bars correspond to the number of biomarkers of the given disease found among the input set.

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):



3.3. 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 [7]. Enhancers typically have a length of several hundreds of nucleotides and are bound by multiple transcription factors in a cooperative manner [8].

In the current work we use the Epigenomics data from the track(s) "Methylation track" to predict positions of potential ***enhancers*** regulating the differentially expressed genes revealed by comparative epigenomics analysis. We took genomic regions -550bp upstream and 550bp downstream from the middle point of each interval of the track and check if these regions are located inside the 5kb flanking areas of the differentially expressed genes (or inside the body of the genes). In such cases, these genomic regions are used for the search for potential condition-specific enhancers. In all other cases when the differentially expressed genes did not contain epigenomic peaks in their body or in the 5kb flanking regions we used the upstream regulatory regions of these genes (-1000bp upstream and +100bp downstream of TSS) for the search for condition-specific enhancers.

We applied the Composite Module Analyst (CMA) [7] 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 (highly methylated genes in Hypertension vs. Control).

To build the most specific composite modules we choose top methylated genes as the input of CMA algorithm. The obtained CMA model is then applied to compute CMA score for all highly methylated genes in Hypertension vs. Control.

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

Module 1:

V$BEN_01	V$HOX13_02	V$MZF1_01	V$TFIIB_Q6_02	V$BEN_02
0.91; N=3	0.82; N=2	0.99; N=3	0.95; N=3	0.75; N=3

Module width: 144

Module 2:

V$HIC1_03	V$PAX4_Q2	V$TFAP2C_05	V$ZBTB33_10
0.92; N=1	0.85; N=3	0.87; N=2	0.99; N=3

Module width: 162

Model score (-p*log10(pval)): 42.06
Wilcoxon p-value (pval): 4.89e-82
Penalty (p): 0.517
Average yes-set score: 9.01
Average no-set score: 7.07
AUC: 0.77
Separation point: 8.36
False-positive: 26.61%
False-negative: 31.48%
The AUC of the model achieves value significantly higher than expected for a random set of regulatory regions
Z-score = 5.07

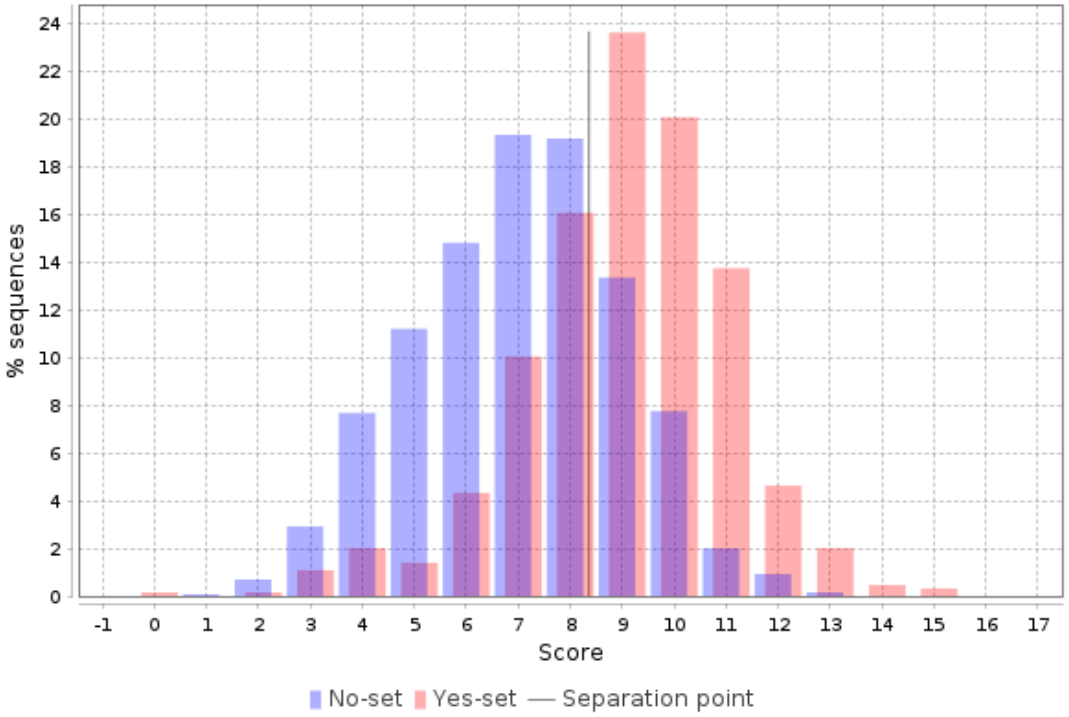


Table 4. List of top ten highly methylated genes in Hypertension vs. Control with identified enhancers in their regulatory regions. ***CMA score*** *- the score of the CMA model of the enhancer identified in the regulatory region.*

See full table →

Ensembl IDs	Gene symbol	Gene description	CMA score	Factor names
ENSG00000242540		novel transcript	15.04	TFIIB(h), ZBTB33(h), Pax-4(h), MZF1(h), HIC1(h), GTF2IRD1(h)
ENSG00000257261	SLC38A4-AS1	SLC38A4 antisense RNA 1	14.58	Pax-4(h), TFIIB(h), AP-2gamma(h), GTF2IRD1(h), ZBTB33(h), MZF1(h), HIC1(h)
ENSG00000273238	TMEM271	transmembrane protein 271	14.46	GTF2IRD1(h), Pax-4(h), MZF1(h), TFIIB(h), AP-2gamma(h), ZBTB33(h), HIC1(h)
ENSG00000125492	BARHL1	BarH like homeobox 1	14.45	Pax-4(h), Hox-A5(h), GTF2IRD1(h), MZF1(h), TFIIB(h), ZBTB33(h), HIC1(h)...
ENSG00000259663	ISL1-DT	ISL1 divergent transcript	14.39	GTF2IRD1(h), Hox-A5(h), ZBTB33(h), TFIIB(h), HIC1(h), AP-2gamma(h), Pax-4(h)
ENSG00000171303	KCNK3	potassium two pore domain channel subfamily K member 3	14.17	GTF2IRD1(h), Pax-4(h), MZF1(h), Hox-A5(h), TFIIB(h), ZBTB33(h), AP-2gamma(h)...
ENSG00000115041	KCNIP3	potassium voltage-gated channel interacting protein 3	14.09	ZBTB33(h), AP-2gamma(h), Pax-4(h), HIC1(h), MZF1(h), TFIIB(h), GTF2IRD1(h)
ENSG00000169902	TPST1	tyrosylprotein sulfotransferase 1	14.06	Hox-A5(h), Pax-4(h), MZF1(h), TFIIB(h), GTF2IRD1(h), AP-2gamma(h), HIC1(h)...
ENSG00000145536	ADAMTS16	ADAM metallopeptidase with thrombospondin type 1 motif 16	13.98	Pax-4(h), AP-2gamma(h), HIC1(h), TFIIB(h), ZBTB33(h), GTF2IRD1(h), MZF1(h)
ENSG00000181234	TMEM132C	transmembrane protein 132C	13.95	MZF1(h), TFIIB(h), GTF2IRD1(h), Pax-4(h), ZBTB33(h), AP-2gamma(h), HIC1(h)

Enhancer model potentially involved in regulation of target genes (low methylated genes in Hypertension vs. Control).

To build the most specific composite modules we choose less methylated genes as the input of CMA algorithm.

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$RARA_17	V$SIN3A_03	V$EGR2_Q6
0.81; N=3	0.74; N=3	0.99; N=3

Module width: 140

Module 2:

V$LYL1_01	V$RELA_03	V$SMAD1_Q6
0.84; N=2	0.94; N=3	0.95; N=2

Module width: 127

Model score (-p*log10(pval)): 26.39
Wilcoxon p-value (pval): 6.64e-46
Penalty (p): 0.584
Average yes-set score: 7.66
Average no-set score: 5.75
AUC: 0.75
Separation point: 6.58
False-positive: 33.08%
False-negative: 26.01%

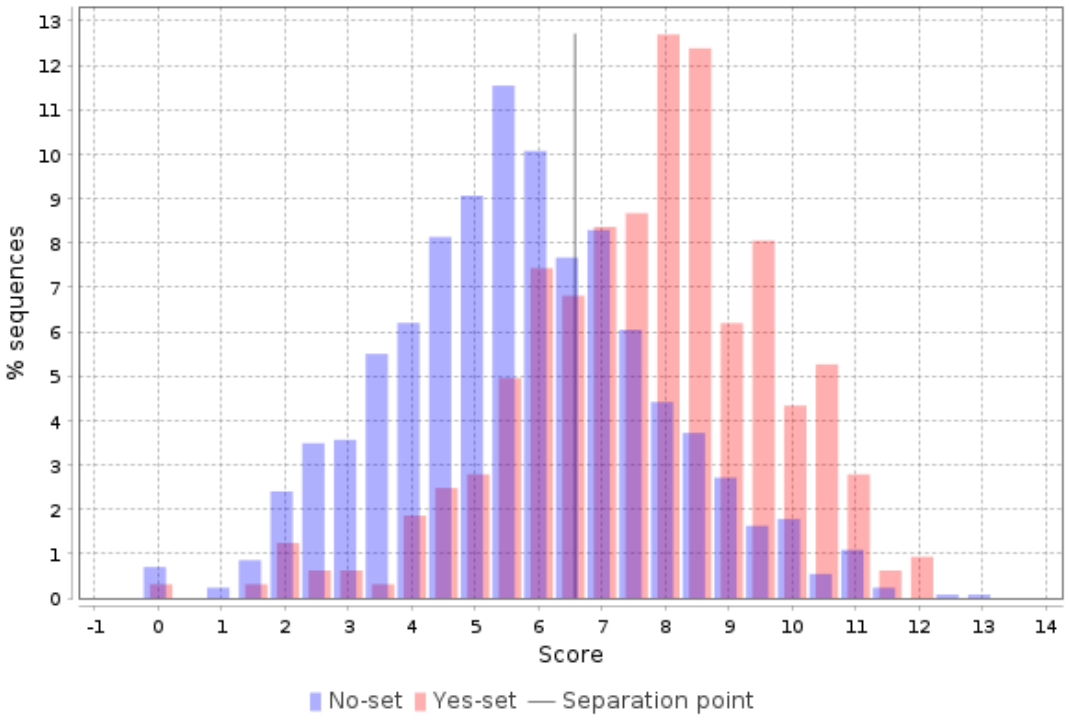


Table 5. List of top ten low methylated genes in Hypertension vs. Control with identified enhancers in their regulatory regions. ***CMA score*** *- the score of the CMA model of the enhancer identified in the regulatory region.*

See full table →

Ensembl IDs	Gene symbol	Gene description	CMA score	Factor names
ENSG00000030110	BAK1	BCL2 antagonist/killer 1	12.25	SMAD1(h), NF-kappaB-p65(h), Lyl-1(h), RAR-alpha(h), SIN3A(h), EGR2(h)
ENSG00000006638	TBXA2R	thromboxane A2 receptor	12.03	SIN3A(h), SMAD1(h), NF-kappaB-p65(h), RAR-alpha(h), Lyl-1(h), EGR2(h)
ENSG00000007516	BAIAP3	BAI1 associated protein 3	11.83	RAR-alpha(h), SMAD1(h), NF-kappaB-p65(h), Lyl-1(h), SIN3A(h), EGR2(h)
ENSG00000063245	EPN1	epsin 1	11.6	RAR-alpha(h), EGR2(h), SIN3A(h), SMAD1(h), NF-kappaB-p65(h), Lyl-1(h)
ENSG00000090971	NAT14	N-acetyltransferase 14 (putative)	11.47	NF-kappaB-p65(h), Lyl-1(h), SMAD1(h), RAR-alpha(h), SIN3A(h), EGR2(h)
ENSG00000102575	ACP5	acid phosphatase 5, tartrate resistant	11.24	SIN3A(h), EGR2(h), RAR-alpha(h), SMAD1(h), NF-kappaB-p65(h), Lyl-1(h)
ENSG00000105325	FZR1	fizzy and cell division cycle 20 related 1	11.24	RAR-alpha(h), EGR2(h), SIN3A(h), Lyl-1(h), NF-kappaB-p65(h), SMAD1(h)
ENSG00000099822	HCN2	hyperpolarization activated cyclic nucleotide gated potassium and sodium channel 2	11.23	SIN3A(h), EGR2(h), RAR-alpha(h), Lyl-1(h), NF-kappaB-p65(h), SMAD1(h)
ENSG00000070388	FGF22	fibroblast growth factor 22	11.07	NF-kappaB-p65(h), Lyl-1(h), SMAD1(h), RAR-alpha(h), SIN3A(h), EGR2(h)
ENSG00000067057	PFKP	phosphofructokinase, platelet	10.96	SMAD1(h), Lyl-1(h), NF-kappaB-p65(h), RAR-alpha(h), SIN3A(h), EGR2(h)

On the basis of the enhancer models we identified transcription factors potentially regulating the ***target genes*** of our interest. We found 8 and 6 transcription factors controlling expression of up- and down-regulated genes respectively (see Tables 6-7).

Table 6. Transcription factors of the predicted enhancer model potentially regulating the differentially expressed genes (highly methylated genes in Hypertension vs. Control). ***Yes-No ratio*** *is the ratio between frequencies of the sites in Yes sequences versus No sequences. It describes the level of the enrichment of binding sites for the indicated TF in the regulatory target regions.* ***Regulatory score*** *is the measure of involvement of the given TF in the controlling of expression of genes that encode master regulators presented below (through positive feedback loops).*

See full table →

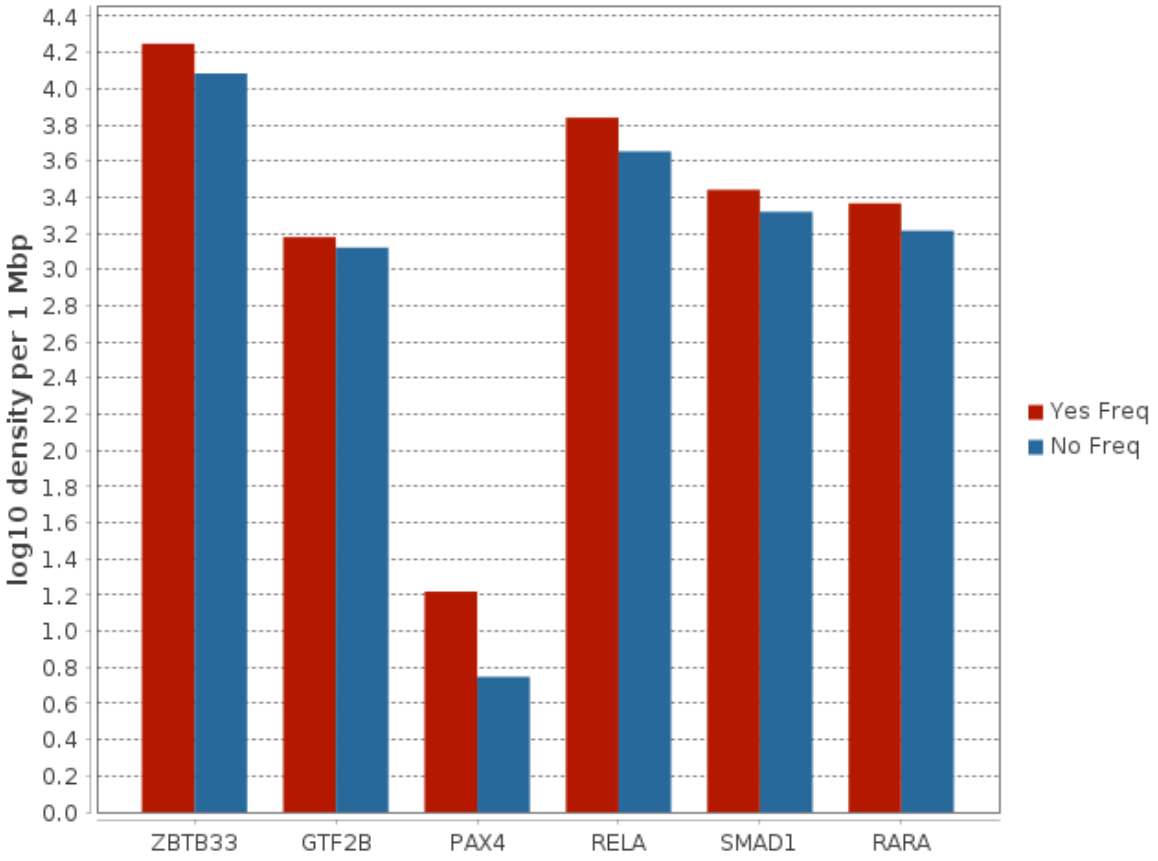
ID	Gene symbol	Gene description	Regulatory score	Yes-No ratio
MO000056592	ZBTB33	zinc finger and BTB domain containing 33	2.13	1.46
MO000013528	GTF2B	general transcription factor IIB	1.9	1.14
MO000026933	PAX4	paired box 4	1.74	2.96
MO000070133	GTF2IRD1	GTF2I repeat domain containing 1	1.74	1.57
MO000117395	HIC1	HIC ZBTB transcriptional repressor 1	1.44	1.72
MO000120562	HOXA5	homeobox A5	1.36	
MO000026449	TFAP2C	transcription factor AP-2 gamma	0	10.53
MO000084468	MZF1	myeloid zinc finger 1	0	1.67

Table 7. Transcription factors of the predicted enhancer model potentially regulating the differentially expressed genes (low methylated genes in Hypertension vs. Control). ***Yes-No ratio*** *is the ratio between frequencies of the sites in Yes sequences versus No sequences. It describes the level of the enrichment of binding sites for the indicated TF in the regulatory target regions.* ***Regulatory score*** *is the measure of involvement of the given TF in the controlling of expression of genes that encode master regulators presented below (through positive feedback loops).*

See full table →

ID	Gene symbol	Gene description	Regulatory score	Yes-No ratio
MO000079319	RELA	RELA proto-oncogene, NF-kB subunit	1.83	1.54
MO000019609	SMAD1	SMAD family member 1	1.76	1.32
MO000033904	RARA	retinoic acid receptor alpha	1.6	1.42
MO000030983	SIN3A	SIN3 transcription regulator family member A	1.48	1.28
MO000025819	LYL1	LYL1 basic helix-loop-helix family member	1.25	1.45
MO000024797	EGR2	early growth response 2	1.16	1.67

The following diagram represents the key transcription factors, which were predicted to be potentially regulating differentially expressed genes in the analyzed pathology: ZBTB33, GTF2B, PAX4, RELA, SMAD1 and RARA.



3.4. Finding master regulators in networks

In the second step of the upstream analysis common regulators of the revealed TFs were identified. 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 8-9.

*Table 8. Master regulators that may govern the regulation of highly methylated genes in Hypertension vs. Control. **Total rank** is the sum of the ranks of the master molecules sorted by keynode score, CMA score, epigenomics data.*

See full table →

ID	Master molecule name	Gene symbol	Gene description	Total rank
MO001090011	AMPKalpha-2{pT172}:AMPK beta:AMPK gamma:adenosine 5'-monophosphate	PRKAA2, PRKAB1, PRKAB2, PRKAG1, PRKAG2, PRKAG3	protein kinase AMP-activated catalytic subunit alpha 2, protein kinase AMP-activated non-catalytic s...	18
MO001090010	AMPKalpha-2(h){pT172}:AMPK beta(h):AMPK gamma(h):adenosine 5'-monophosphate	PRKAA2	protein kinase AMP-activated catalytic subunit alpha 2	19
MO000114254	AMPKalpha-2(h)	PRKAA2	protein kinase AMP-activated catalytic subunit alpha 2	20
MO000114255	AMPKalpha-2(h)	PRKAA2	protein kinase AMP-activated catalytic subunit alpha 2	20
MO001092721	AMPKalpha-2(h):Hsc70(h):Phosphorylated PLINs on lipid droplet surface(h)	HSPA8, PRKAA2	heat shock protein family A (Hsp70) member 8, protein kinase AMP-activated catalytic subunit alpha 2	22
MO001092722	AMPKalpha-2:Hsc70:Phosphorylated PLINs on lipid droplet surface	HSPA8, PLIN2, PLIN3, PRKAA2	heat shock protein family A (Hsp70) member 8, perilipin 2, perilipin 3, protein kinase AMP-activated...	22
MO001093054	Hsc70(h):PLINs on lipid droplet surface(h):AMPKalpha-2(h)	HSPA8, PRKAA2	heat shock protein family A (Hsp70) member 8, protein kinase AMP-activated catalytic subunit alpha 2	24
MO001093055	Hsc70:PLINs on lipid droplet surface:AMPKalpha-2	HSPA8, PLIN2, PLIN3, PRKAA2	heat shock protein family A (Hsp70) member 8, perilipin 2, perilipin 3, protein kinase AMP-activated...	24
MO000007757	IRR(h)	INSRR	insulin receptor related receptor	52
MO000221008	SATB-2-isoform1(h)	SATB2	SATB homeobox 2	64

*Table 9. Master regulators that may govern the regulation of low methylated genes in Hypertension vs. Control. **Total rank** is the sum of the ranks of the master molecules sorted by keynode score, CMA score, epigenomics data.*

See full table →

ID	Master molecule name	Gene symbol	Gene description	Total rank
MO001093132	cyclinG1(h):PP2A-PPP2R5C(h):p-T218,S166,S188-MDM2, p-S166,S188-MDM2(h)	CCNG1, PPP2R5C	cyclin G1, protein phosphatase 2 regulatory subunit B'gamma	18
MO000042160	B56gamma(h)	PPP2R5C	protein phosphatase 2 regulatory subunit B'gamma	19
MO000044269	cyclosome(h):Fzr(h)	ANAPC1, ANAPC10, ANAPC11, ANAPC2, ANAPC4, ANAPC5, ANAPC7, CDC16, CDC23, CDC26, CDC27, FZR1	anaphase promoting complex subunit 1, anaphase promoting complex subunit 10, anaphase promoting comp...	21
MO000007757	IRR(h)	INSRR	insulin receptor related receptor	24
MO000018003	PP2A(h)	PPP2CA, PPP2R3A, PPP2R3B, PPP2R5A, PPP2R5B, PPP2R5C, PPP2R5D	protein phosphatase 2 catalytic subunit alpha, protein phosphatase 2 regulatory subunit B"alpha, pr...	33
MO001087323	PP2A-B56-beta,gamma	PPP2CA, PPP2CB, PPP2R1A, PPP2R1B, PPP2R5B, PPP2R5C	protein phosphatase 2 catalytic subunit alpha, protein phosphatase 2 catalytic subunit beta, protein...	40
MO000020878	VEGFR-1(h)	FLT1	fms related receptor tyrosine kinase 1	41
MO000083006	CaMKK-alpha(h)	CAMKK1	calcium/calmodulin dependent protein kinase kinase 1	45
MO001093133	cyclinG1:PP2A-PPP2R5C:p-T218,S166,S188-MDM2, p-S166,S188-MDM2	CCNG1, MDM2, PPP2CA, PPP2CB, PPP2R1A, PPP2R1B, PPP2R5C	MDM2 proto-oncogene, cyclin G1, protein phosphatase 2 catalytic subunit alpha, protein phosphatase 2...	46
MO000114737	B56gamma-isoform1(h)	PPP2R5C	protein phosphatase 2 regulatory subunit B'gamma	52

The intracellular regulatory pathways controlled by the above-mentioned master regulators are depicted in Figures 8 and 9. 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.

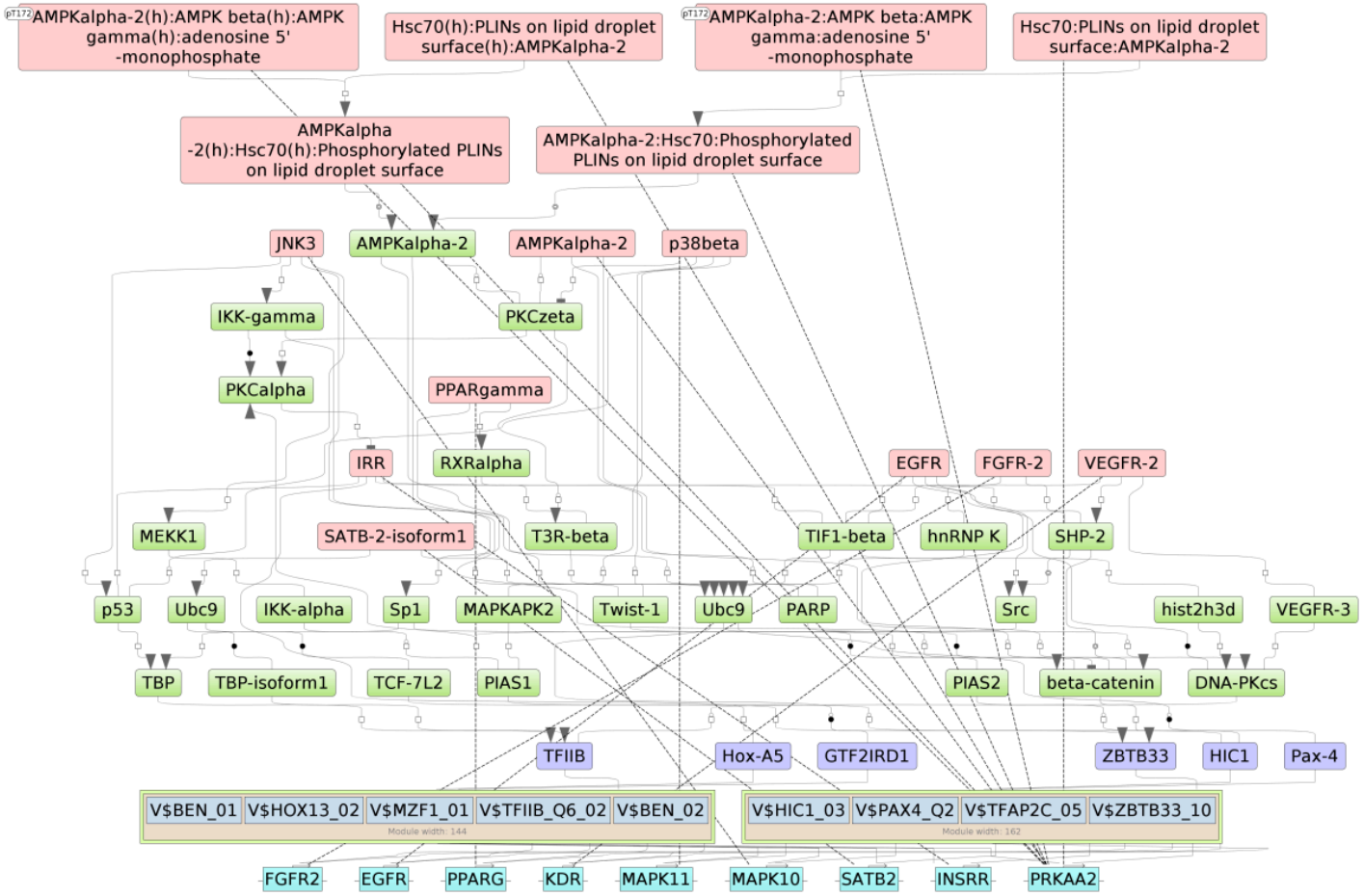


Figure 8. Diagram of intracellular regulatory signal transduction pathways of highly methylated genes in Hypertension vs. Control. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate molecules, which have been added to the network during the search for master regulators from selected TFs.

See full diagram →

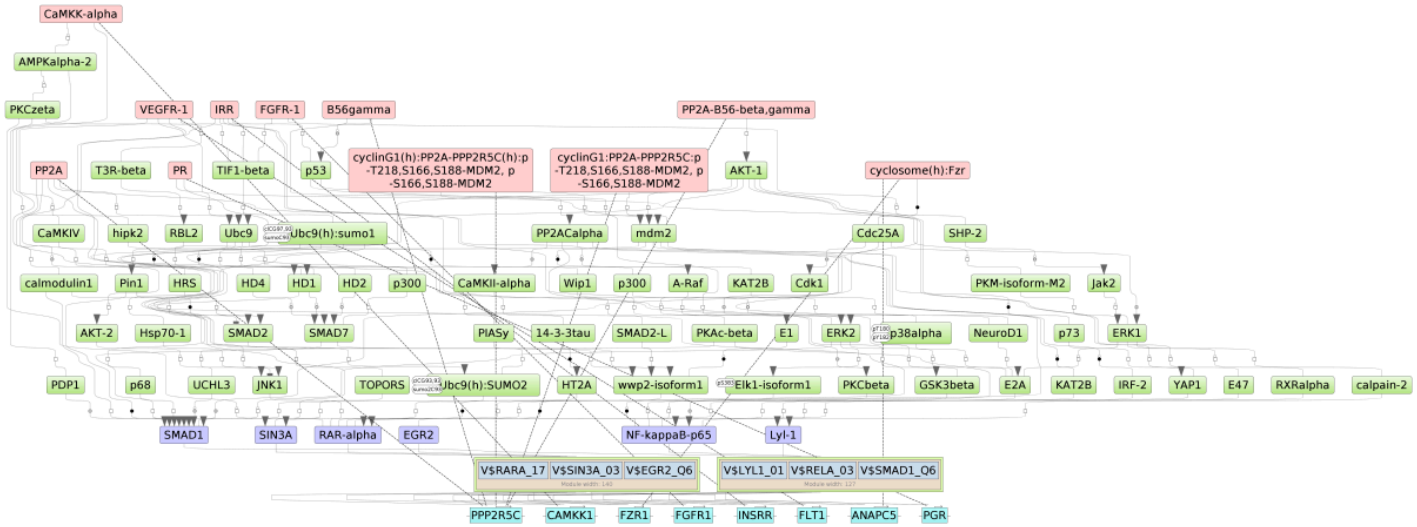


Figure 9. Diagram of intracellular regulatory signal transduction pathways of low methylated genes in Hypertension vs. Control. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate molecules, which have been added to the network during the search for master regulators from selected TFs. Orange and blue frames highlight molecules that are encoded by up- and downregulated genes, resp.

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™ [11] database of gene-disease-drug assignments and PASS [12-14] 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 [12-14] on the basis of a (Q)SAR approach.

If both Druggability scores were below defined thresholds (see Methods 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 10. 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	Total rank
PRKAA2	protein kinase AMP-activated catalytic subunit alpha 2	35	8
INSRR	insulin receptor related receptor	27	16
CDK6	cyclin dependent kinase 6	16	22
EPHB2	EPH receptor B2	32	31
GRM2	glutamate metabotropic receptor 2	2	41
IRS1	insulin receptor substrate 1	6	42

*Table 11. 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.*

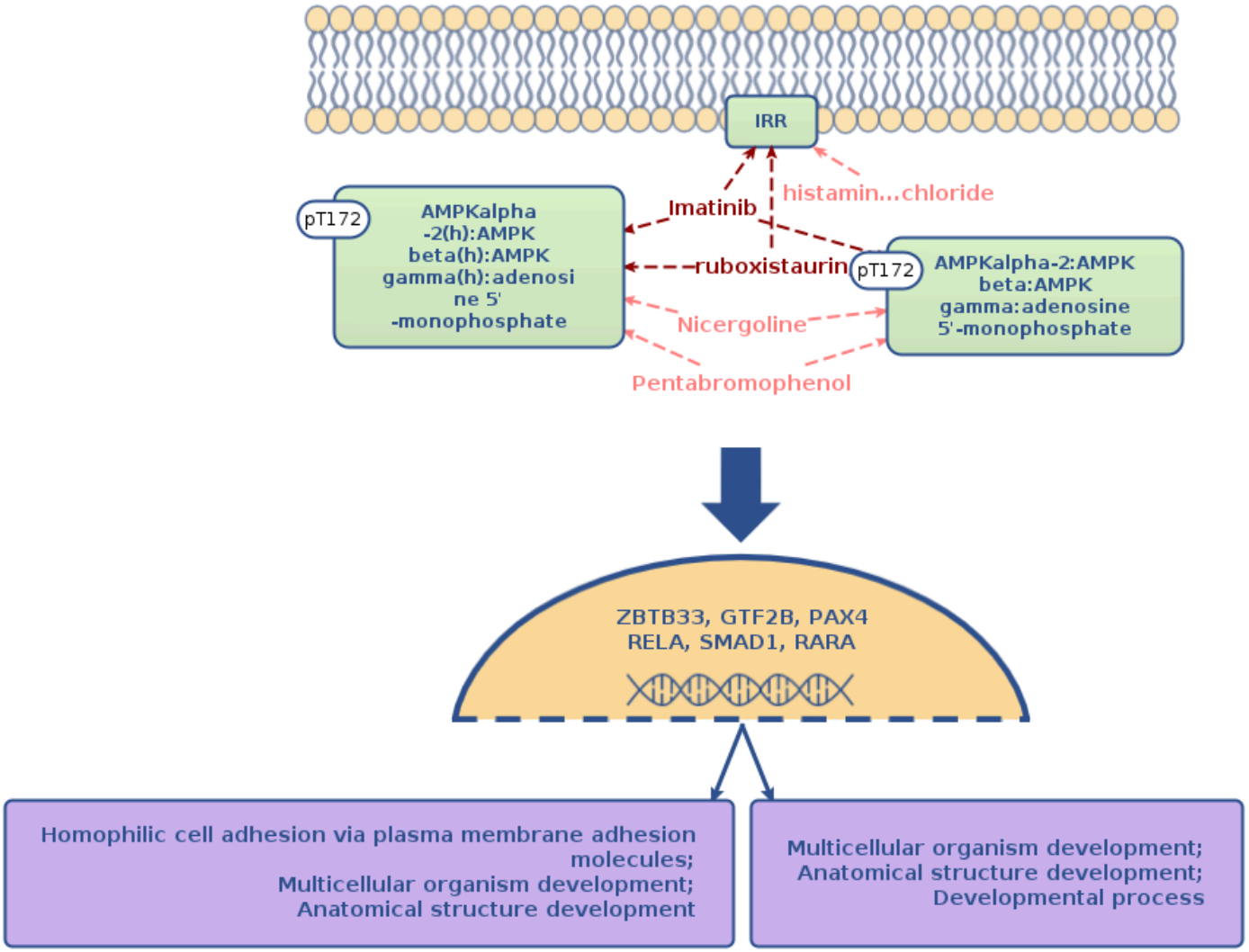
See full table →

Gene symbol	Gene Description	Druggability score	Total rank
PRKAA2	protein kinase AMP-activated catalytic subunit alpha 2	17.12	8
INSRR	insulin receptor related receptor	9.41	16
CDK6	cyclin dependent kinase 6	13.46	22
EPHB2	EPH receptor B2	37.05	31
GRM2	glutamate metabotropic receptor 2	14.42	41
ALK	ALK receptor tyrosine kinase	10.24	52

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:

- IRR
- AMPKalpha-2(h):AMPK beta(h):AMPK gamma(h):adenosine 5'-monophosphate
- AMPKalpha-2:AMPK beta:AMPK gamma:adenosine 5'-monophosphate

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: Imatinib, Nicergoline, Pentabromophenol, ruboxistaurin and histamine dihydrochloride, 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 the ranks sum based on the individual ranks of the following 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));

- Clinical validity score (applicable only for drugs predicted on the basis of literature curation in HumanPSD™ database (Tables 12 and 13), reflects the number of the highest clinical trials phase on which the drug was tested for any pathology).

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

Based on the Drug rank, a numerical value of Drug score was calculated, which reflects the potential activity of the respective drug on the overall molecular mechanism of the studied pathology. Drug score values belong to the range from 1 to 100 and are calculated as a quotient of maximum drug rank and the drug rank of the given drug multiplied by 100.

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

Drugs approved in clinical trials

Table 12. 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 score	Disease activity score	Disease trial phase
Imatinib	MAP4K4, IRAK3, ABL1, FGFR2, PAK2, SYK, EPHA2, ERBB2, CAMK2B, PRKD1, MAP2K6, IGF1R, LCK, CAMKK2, MAPK4, EPHA7, PTK2, MAPK11, FLT1, LYN, CSNK2A1, EPHA8, FES, STK3, MAPK10, ACVR1B, CSNK1A1, NTRK1, PRKAA2, KDR, EPHA4, PDGFRA, FLT4, LTK, EGFR, PRKG1, BMX, KIT, ALK, EPHB2, RPS6KA2, NTRK2, EPHB4, PRKG2, CRK, FGFR1, INSRR, CAMKK1, EPHA6, TGFBR2, ABL2, RPS6KA4	94	5	Phase 3: Hypertension, Adenocarcinoma, Astrocytoma, Bone Marrow Diseases, Brain Neoplasms, COVID-19, Cytokine Release Syndrome, Gastrointestinal Stromal Tumors, Glioblastoma, Hematologic Diseases, Hypertension, Pulmonary, Leukemia, Leukemia, Lymphoid, Leukemia, Myelogenous, Chronic, BCR-ABL Positive, Leukemia, Myeloid, Leukemia, Myeloid, Chronic-Phase, Lymphoma, Neoplasm Metastasis, Neoplasms, Pneumonia, Precursor Cell Lymphoblastic Leukemia-Lymphoma, Pulmonary Arterial Hypertension, Sarcoma, Severe Acute Respiratory Syndrome
Pazopanib	MAP4K4, IRAK3, ABL1, FGFR2, PAK2, SYK, EPHA2, ERBB2, CAMK2B, PRKD1, MAP2K6, IGF1R, LCK, CAMKK2, MAPK4, EPHA7, PTK2, MAPK11, FLT1, LYN, CSNK2A1, EPHA8, FES, STK3, MAPK10, ACVR1B, NTRK1, PRKAA2, KDR, EPHA4, PDGFRA, FLT4, LTK, EGFR, PRKG1, BMX, KIT, ALK, EPHB2, RPS6KA2, NTRK2, EPHB4, PRKG2, FGFR1, INSRR, CAMKK1, EPHA6, TGFBR2, ABL2, RPS6KA4	91	2	Phase 2: Hypertension, Anemia, Brain Abscess, Brain Neoplasms, Breast Neoplasms, Carcinoma, Carcinoma, Non-Small-Cell Lung, Carcinoma, Renal Cell, Carcinoma, Small Cell, Carcinoma, Transitional Cell, Central Nervous System Neoplasms, Cholangiocarcinoma, Chondrosarcoma, Chondrosarcoma, Mesenchymal, Dermatofibrosarcoma, Desmoplastic Small Round Cell Tumor, Dilatation, Pathologic, Disorders of Sex Development, Drug-Related Side Effects and Adverse Reactions, Edema, Epistaxis, Fibrosarcoma, Gallbladder Neoplasms, Gastrointestinal Neoplasms, Gastrointestinal Stromal Tumors, Glioblastoma, Glioma, Gliosarcoma, Glomus Tumor, Glucagonoma, Granular Cell Tumor, Granuloma, Granuloma, Plasma Cell, Hemangioendothelioma, Hemangioendothelioma, Epithelioid, Hemangioma, Hemangiosarcoma, Hemorrhage, Histiocytoma, Histiocytoma, Malignant Fibrous, Insulinoma, Intestinal Neoplasms, Kidney Neoplasms, Leiomyosarcoma, Liposarcoma, Liposarcoma, Myxoid, Lymphedema, Macular Degeneration, Melanoma, Multiple Endocrine Neoplasia, Multiple Endocrine Neoplasia Type 1, Myoepithelioma, Myosarcoma, Neoplasm Metastasis, Neoplasms, Neoplasms, Adipose Tissue, Neoplasms, Plasma Cell, Nerve Sheath Neoplasms, Nervous System Neoplasms, Neurofibrosarcoma, Osteosarcoma, Ovarian Neoplasms, Pancreatic Neoplasms, Paraganglioma, Paraganglioma, Extra-Adrenal, Perivascular Epithelioid Cell Neoplasms, Pheochromocytoma, Ranula, Recurrence, Rhabdomyosarcoma, Rosacea, Sarcoma, Sarcoma, Alveolar Soft Part, Sarcoma, Clear Cell, Sarcoma, Synovial, Small Cell Lung Carcinoma, Smooth Muscle Tumor, Solitary Fibrous Tumors, Somatostatinoma, Telangiectasia, Hereditary Hemorrhagic, Telangiectasis, Urinary Bladder Neoplasms, Uterine Cervical Neoplasms, Vascular Neoplasms, Zollinger-Ellison Syndrome
Sorafenib	MAP4K4, IRAK3, ABL1, FGFR2, PAK2, SYK, EPHA2, ERBB2, CAMK2B, PRKD1, MAP2K6, IGF1R, LCK, CAMKK2, MAPK4, EPHA7, PRKCZ, PTK2, MAPK11, FLT1, LYN,	91	1	Phase 1: Hypertension, Adenoma, Adenoma, Islet Cell, Adenoma, Liver Cell, Anemia, Bile Duct Neoplasms, Biliary Tract Neoplasms, Bites and Stings, Blast Crisis, Brain Abscess, Brain Neoplasms, Breast Neoplasms, Burkitt Lymphoma, Carcinoid Tumor, Carcinoma, Carcinoma, Hepatocellular, Carcinoma, Renal Cell, Cholangiocarcinoma, Colorectal Neoplasms, Gallbladder Neoplasms, Glioblastoma, Gliosarcoma, Head and Neck Neoplasms, Hematologic Neoplasms, Hodgkin Disease, Hypertension, Pulmonary, Kidney Neoplasms, Leukemia, Leukemia, Lymphocytic, Chronic, B-Cell, Leukemia, Lymphoid, Leukemia, Myelogenous, Chronic, BCR-ABL

	CSNK2A1, EPHA8, FES, STK3, MAPK10, RIPK3, ACVR1B, NTRK1, PRKAA2, KDR, EPHA4, PDGFRA, FLT4, LTK, EGFR, PRKG1, BMX, KIT, ALK, HIPK2, EPHB2, RPS6KA2, NTRK2, EPHB4, PRKG2, FGFR1, INSRR, CAMKK1, EPHA6, TGFBR2, ABL2, RPS6KA4			Positive, Leukemia, Myeloid, Leukemia, Myeloid, Accelerated Phase, Leukemia, Myeloid, Acute, Leukemia, Myeloid, Chronic, Atypical, BCR-ABL Negative, Leukemia, Myeloid, Chronic-Phase, Leukemia, Myelomonocytic, Chronic, Liver Neoplasms, Lung Neoplasms, Lymphoma, Lymphoma, B-Cell, Lymphoma, B-Cell, Marginal Zone, Lymphoma, Follicular, Lymphoma, Large B-Cell, Diffuse, Lymphoma, Large-Cell, Immunoblastic, Lymphoma, Mantle-Cell, Lymphoma, Non-Hodgkin, Lymphoma, T-Cell, Lymphoma, T-Cell, Cutaneous, Melanoma, Mesothelioma, Multiple Myeloma, Mycosis Fungoides, Myelodysplastic Syndromes, Myelodysplastic-Myeloproliferative Diseases, Myeloproliferative Disorders, Myosarcoma, Neoplasm Metastasis, Neoplasms, Neoplasms, Plasma Cell, Neuroblastoma, Neuroendocrine Tumors, Osteosarcoma, Pancreatic Neoplasms, Pica, Precursor Cell Lymphoblastic Leukemia-Lymphoma, Primary Myelofibrosis, Prostatic Neoplasms, Pulmonary Arterial Hypertension, Rectal Neoplasms, Recurrence, Rhabdomyosarcoma, Sarcoma, Sezary Syndrome, Thrombocytosis, Urinary Bladder Neoplasms, Waldenstrom Macroglobulinemia
Sunitinib	MAP4K4, IRAK3, ABL1, FGFR2, PAK2, SYK, EPHA2, ERBB2, CAMK2B, PRKD1, MAP2K6, IGF1R, LCK, CAMKK2, MAPK4, EPHA7, PTK2, MAPK11, FLT1, LYN, CSNK2A1, EPHA8, FES, STK3, MAPK10, ACVR1B, NTRK1, PRKAA2, KDR, EPHA4, PDGFRA, FLT4, LTK, EGFR, PRKG1, BMX, KIT, ALK, EPHB2, RPS6KA2, NTRK2, EPHB4, PRKG2, FGFR1, INSRR, CAMKK1, EPHA6, TGFBR2, ABL2, RPS6KA4	88	1	N/A
Tamoxifen	TGM2, PGR, NTRK1, ESR1, ERBB2, CASP3, DUSP6, NRG1, BRCA1, IGF1R	87	5	Phase 3: Hypertension, Bipolar Disorder, Breast Neoplasms, Carcinoma, Carcinoma, Intraductal, Noninfiltrating, Cardiovascular Diseases, Coronary Artery Disease, Coronary Disease, Fatigue, Heart Diseases, Hot Flashes, Infertility, Ischemia, Myocardial Ischemia, Neoplasm Metastasis, Neoplasm Recurrence, Local, Neoplasms, Pain, Pulmonary Fibrosis, Recurrence, Sleep Wake Disorders, Vascular Diseases

The ***Disease trial phase*** column reflects the maximum clinical trials phase in which the drug was studied for the analyzed pathology.

Repurposing drugs

Table 13. 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 score	Maximum trial phase
ruboxistaurin	MAP4K4, IRAK3, ABL1, FGFR2, PAK2, SYK, EPHA2, ERBB2, CAMK2B, PRKD1, MAP2K6, IGF1R, LCK, PRKCG, CAMKK2, MAPK4, EPHA7, PRKCZ, PTK2, MAPK11, FLT1, LYN, CSNK2A1, EPHA8, FES, STK3, MAPK10, ACVR1B, NTRK1, PRKAA2, KDR, EPHA4, PDGFRA, FLT4, LTK, EGFR, PRKG1, BMX, KIT, ALK, HIPK2, EPHB2, RPS6KA2, NTRK2, EPHB4, PRKG2, FGFR1, INSRR, CAMKK1, EPHA6, TGFBR2, ABL2, RPS6KA4	88	Phase 3: Diabetes Mellitus, Diabetes Mellitus, Type 1, Diabetic Retinopathy
Flavopiridol	MAP4K4, IRAK3, ABL1, FGFR2, PAK2, SYK, EPHA2, ERBB2, CAMK2B, PRKD1, MAP2K6, IGF1R, LCK, CAMKK2, MAPK4, EPHA7, PTK2, MAPK11, FLT1, LYN, CSNK2A1, EPHA8, FES, STK3, MAPK10, ACVR1B, CDK6, NTRK1, PRKAA2, KDR, EPHA4, PDGFRA, FLT4, LTK, EGFR, PRKG1, BMX, KIT, ALK, EPHB2, RPS6KA2, NTRK2, EPHB4, PRKG2, FGFR1, INSRR, CAMKK1, EPHA6, TGFBR2, ABL2, RPS6KA4	88	Phase 2: Burkitt Lymphoma, Carcinoma, Squamous Cell, Embolism, Head and Neck Neoplasms, Leukemia, Leukemia, Lymphoid, Lymphoma, Lymphoma, Follicular, Lymphoma, Large B-Cell, Diffuse, Lymphoma, Large-Cell, Immunoblastic, Lymphoma, Mantle-Cell, Lymphoma, Non-Hodgkin, Neoplasms, Precursor Cell Lymphoblastic Leukemia-Lymphoma, Salivary Gland Neoplasms, Sarcoma, Squamous Cell Carcinoma of Head and Neck, Thromboembolism, Tumor Lysis Syndrome
CI-1033	MAP4K4, IRAK3, ABL1, FGFR2, PAK2, SYK, EPHA2, ERBB2, CAMK2B, PRKD1, MAP2K6, IGF1R, LCK, CAMKK2, MAPK4, EPHA7, PTK2, MAPK11, FLT1, LYN, CSNK2A1, EPHA8, FES, STK3, MAPK10, ACVR1B, NTRK1, PRKAA2, KDR, EPHA4, PDGFRA, FLT4, LTK, EGFR, PRKG1, BMX, KIT, ALK, EPHB2, RPS6KA2, NTRK2, EPHB4, PRKG2, FGFR1, ERBB4, INSRR, CAMKK1, EPHA6, TGFBR2, ABL2, RPS6KA4	87	Phase 2: Breast Neoplasms, Carcinoma, Non-Small-Cell Lung, Lung Neoplasms, Neoplasms
pelitinib	MAP4K4, IRAK3, ABL1, FGFR2, PAK2, SYK, EPHA2, ERBB2, CAMK2B, PRKD1, MAP2K6, IGF1R, LCK, CAMKK2, MAPK4, EPHA7, PTK2, MAPK11, FLT1, LYN, CSNK2A1, EPHA8, FES, STK3, MAPK10, ACVR1B, NTRK1, PRKAA2, KDR, EPHA4, PDGFRA, FLT4, LTK, EGFR, PRKG1, BMX, KIT, ALK, EPHB2, RPS6KA2, NTRK2, EPHB4, PRKG2, FGFR1, ERBB4, INSRR, CAMKK1, EPHA6, TGFBR2, ABL2, RPS6KA4	87	Phase 2: Carcinoma, Non-Small-Cell Lung, Colonic Neoplasms, Colorectal Neoplasms, Neoplasms, Rectal Neoplasms
1-(5-Tert-Butyl-2-P-Tolyl-2h-Pyrazol-3-Yl)-3-[4-(2-Morpholin-4-Yl-Ethoxy)-Naphthalen-1-Yl]-Urea	MAP4K4, IRAK3, ABL1, FGFR2, PAK2, SYK, EPHA2, ERBB2, CAMK2B, PRKD1, MAP2K6, IGF1R, LCK, CAMKK2, MAPK4, EPHA7, PRKCZ, PTK2, MAPK11, FLT1, LYN, CSNK2A1, EPHA8, FES, STK3, MAPK10, ACVR1B, NTRK1, PRKAA2, KDR, EPHA4, PDGFRA, FLT4, LTK, EGFR, PRKG1, BMX, KIT, ALK, HIPK2, EPHB2, RPS6KA2, NTRK2, EPHB4, PRKG2, FGFR1, INSRR, CAMKK1, EPHA6, TGFBR2, ABL2, RPS6KA4	87	Phase 2: Arthritis, Arthritis, Rheumatoid, Psoriasis

The ***Maximum trial phase*** column reflects the maximum clinical trials phase in which the drug was studied for any pathology.

Table 14. Prospective drugs, predicted by PASS software to be active against the identified drug targets with predicted activity against the studied disease(s) (drug candidates predicted with the cheminformatics tool PASS)

See full table →

Name	Target names	Drug score	Target activity score
Moexipril	BDNF, NTF3, ITGB3, NGF	94	0.59
Quinapril	BDNF, NTF3, ITGB3, NGF	92	0.45
Hesperetin	MAPK10, BDNF, SMAD3, MAPK4, MAPK11, CASP3, ESR1, NTF3, NGF	90	0.26
(2s,3s)-Trans-Dihydroquercetin	MAPK10, BDNF, SMAD3, MAPK4, MAPK11, CASP3, ESR1, NTF3, NGF	90	0.23
Silibinin	BDNF, ESR1, CASP3, NTF3, NGF	85	0.13

Table 15. 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 score	Target activity score
3-METHYL-1,6,8-TRIHYDROXYANTHRAQUINONE	MAPK10, DUSP26, MAPK4, CDC14B, MAPK11, ESR1, CASP3, DUSP6, DUSP14, DUPD1, NGF, BRCA1	90	0.34
(2S)-5-hydroxy-2-(4-hydroxyphenyl)-7-methoxy-2,3-dihydro-4H-chromen-4-one	MAPK10, BDNF, SMAD3, MAPK4, MAPK11, CASP3, ESR1, NTF3, NGF	89	0.27
Naringenin	MAPK10, BDNF, SMAD3, MAPK4, MAPK11, CASP3, ESR1, NTF3, NGF	89	0.26
2,5,7-Trihydroxynaphthoquinone	MAPK10, DUSP26, SMAD3, MAPK4, CDC14B, MAPK11, DUSP6, DUSP14, DUPD1, NGF, BRCA1	89	0.35
Estrone	MAPK10, BDNF, MAPK4, MAPK11, ESR1, NTF3, NGF	87	0.18

As the result of drug search we propose the following drugs as most promising candidates for treating the pathology under study: Imatinib, ruboxistaurin, Moexipril and 3-METHYL-1,6,8-TRIHYDROXYANTHRAQUINONE. These drugs were selected for acting on the following targets: PRKAA2, BDNF and MAPK11, 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 *epigenomics* data. The study is done in the context of *Hypertension*. 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:

Imatinib, ruboxistaurin, Moexipril and 3-METHYL-1,6,8-TRIHYDROXYANTHRAQUINONE

These drugs were selected for acting on the following targets: PRKAA2, BDNF and MAPK11, 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:

IRR, AMPKalpha-2(h):AMPK beta(h):AMPK gamma(h):adenosine 5'-monophosphate and AMPKalpha-2:AMPK beta:AMPK gamma:adenosine 5'-monophosphate

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: Imatinib, Nicergoline, Pentabromophenol, ruboxistaurin and histamine dihydrochloride. 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.

- IRR
- AMPKalpha-2(h):AMPK beta(h):AMPK gamma(h):adenosine 5'-monophosphate
- AMPKalpha-2:AMPK beta:AMPK gamma:adenosine 5'-monophosphate

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 2026.1 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transfac).
The master regulator search uses the TRANSPATH® database (BIOBASE), release 2026.1 (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 2026.1 (https://genexplain.com/humanpsd).
The Ensembl database release Human112.38 (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.

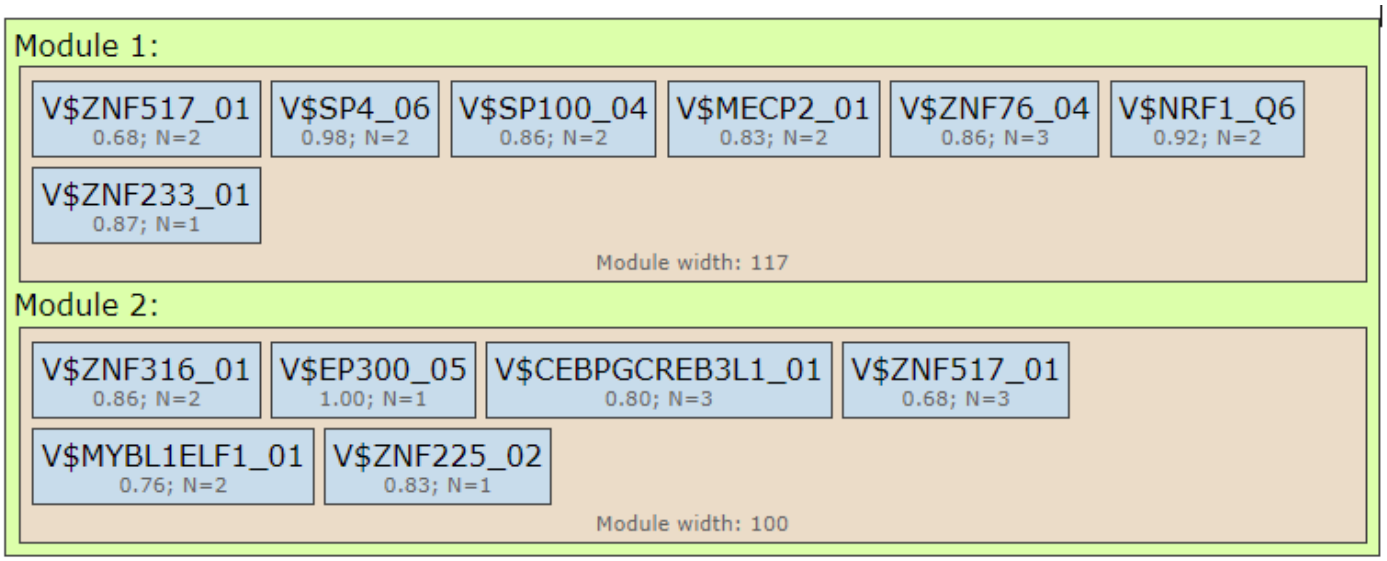
Epigenomics data processing

When analyzing a list of CpG sites, we compute the fold change values between the methylation status in the studied pathology and the control set. Top 10 000 CpG sites with highest logFC values are taken to further analysis. These sites are mapped to corresponding genes, which will be further compared to the list of housekeeping genes at the step of promoter analysis.

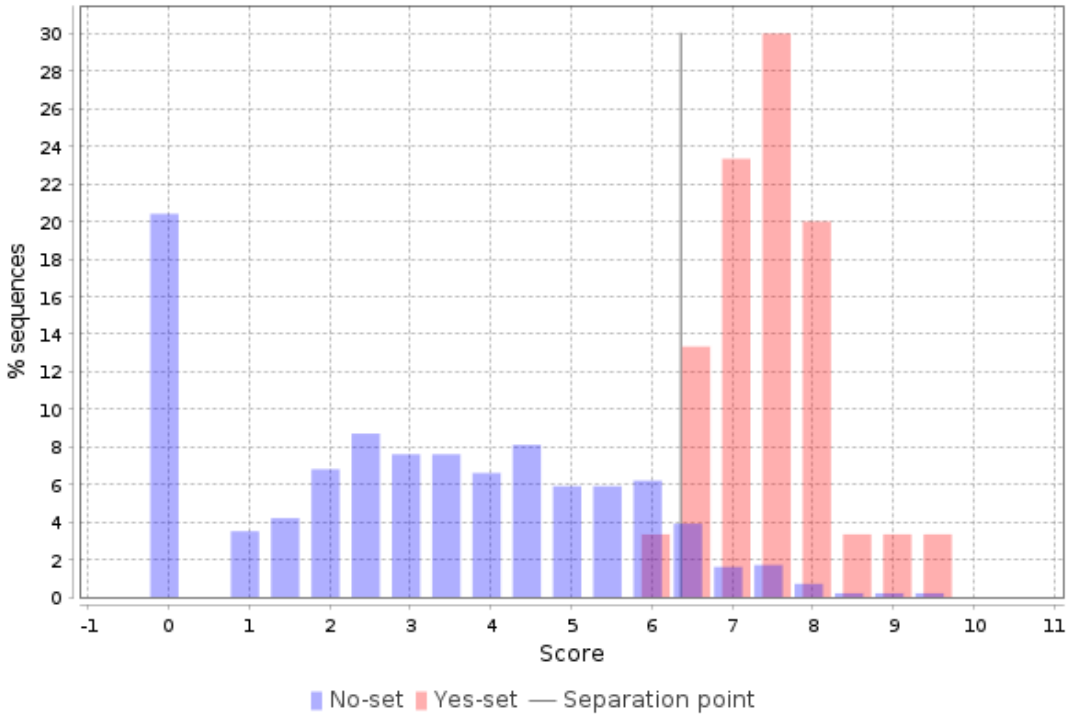
Methods for the analysis of enriched transcription factor binding sites and composite modules

Transcription factor binding sites in promoters and enhancers of disease affected 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. The matrix profile used to identify potential transcription factor binding sites is a collection of 3390 TRANSFAC® vertebrate matrices carefully selected for the purpose of enrichment analysis.
We search for transcription factor binding sites (TFBS) that are enriched in the enhancers under study as compared to a background set of promoters of housekeeping genes. 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 length 1100bp (-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 [6] 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).
CMA constructs a generalized model of the regulatory regions of the studied genes by specifying combinations of potential TFBSs that are most frequently clustered together. CMA identifies the transcription factors that through their cooperation are able to

provide a synergistic effect and, thus, are likely to have a great influence on the gene regulation process.
Composite modules can have a complex hierarchical structure consisting of two levels: site models and modules. The highest hierarchical level contains two modules and corresponds to the whole promoter model. The first level, site model, corresponds to the individual site model, based on one PWM. Names of the site models are the same as the matrix names (site models are taken from the profile that was used in the site search). In the resulting schema (see example below) the site models are shown by blue boxes. Within these boxes, there are two values below each site model. The first value is the threshold value for the score of the respective site model, which is determined by the genetic algorithm during the optimization process. The second value is the maximum number of best individual matches (sites) found for this site model and taken into account for calculating the score of the module.
Promoter model example:

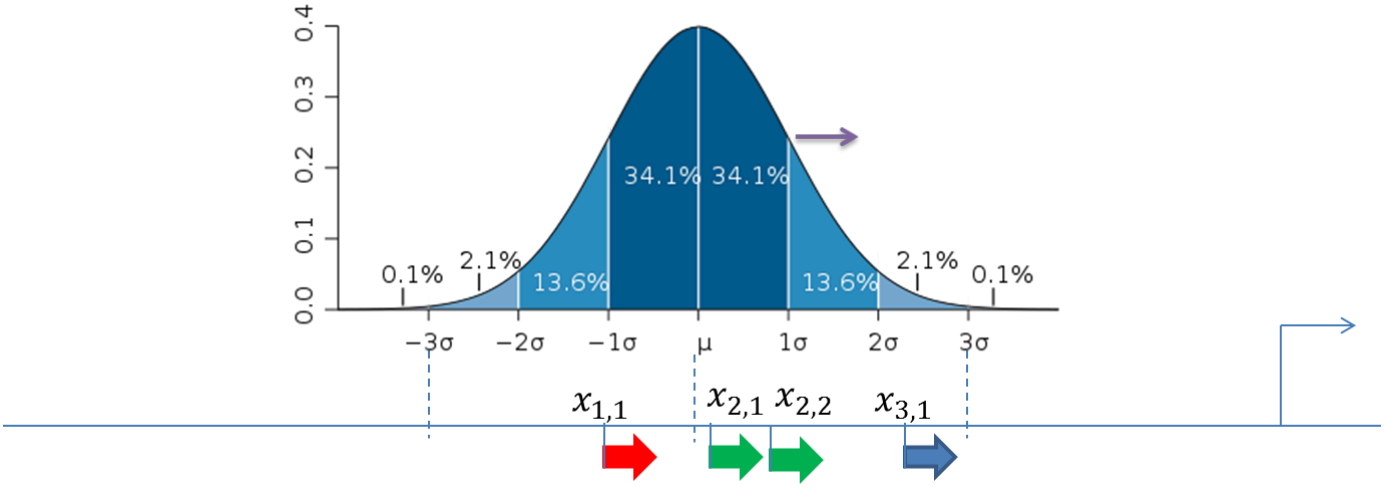


The next level, module, may contain several site models, shown within the light brown boxes. The module is characterized by its width, the average length of DNA window containing matches for the mentioned site models. In the resulting schemas modules are shown in green boxes, and they are numbered: Module 1 and Module 2. The complexity of the promoter model to be constructed is defined by the number of units of each level: number of modules, number of site models, as well as the maximum number of individual sites to be considered. In the Genome Enhancer workflow we limit the number of modules to 2, the number of site models from 6 to 8, and the maximum number of individual sites to be considered from 1 to 3.
The CMA score is calculated for each promoter depending on the number of modules, site models, sites, their scores and other statistical parameters. The higher is the CMA score of a promoter, the better is the differentiation of this promoter from the promoters of the NO set. The distribution of scores for individual promoters is shown on an example histogram below:



The CMA score of the promoters is shown on the X axis of the histogram and the percentage of promoters (% sequences) having this score is shown on the Y axis. The separation point shown in gray corresponds to the average value. Promoters from the NO set with a score above the separation point (blue bars to the right of the separation point) are referred to as false positives. Promoters from the YES set with a score below the separation point (red bars to the left of the separation point) are referred to as false

negatives. The YES promoters with a score above the separation point are well separated from the NO promoters, meaning that for these promoters the constructed composite model is most suitable.
The figure below demonstrates the calculation of the score value for the composite modules in the promoter sequences. The TSS is shown as a thin arrow on the right side of the figure. Four thick arrows exemplify four sites found in this promoter. The color of the arrows exemplifies the site model which these sites belong to (three site models – red, green and blue).



A promoter model consists of K modules. The score of each module $M_{k\ (Score(M_{k),\ k\ =\ 1,\ \ldots,\ K)}}$ is calculated according to the following formula:

$$Score\ (M_k) = \max_{\mu=1,L} \sum_{t=1}^{T_k} \sum_{i=1}^{m_t} SiteScore(t,i) \times f(x_{t,i}, \mu, \sigma^2)$$

Here,

- *SiteScore (t,i)* is the site score for the sites found in the promoter, which is calculated by the MATCH algorithm.
- m_t – the number of sites of the site model t found in the promoter.
- T_k – the number of site models in the module M_k, and
- $$f(x; \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The final promoter score is calculated as the sum of the module scores M_k.
Standard deviation (σ) of the normal distribution is subject to optimization by the genetic algorithm and represents the width of the module in the result of the composite module analysis.
Transcription factors that correspond to the PWMs of the composite modules are further taken to the next step of finding master regulators in networks.

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 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. [3,4]

The signal transduction network is represented as a weighted graph, which is defined as Θ = (M, E, S), where M is the set of nodes (all molecules in the database), E is the set of edges (all reactions between molecules in the database) and S : E → R+ ∪ {0} is the cost function that defines a non-negative value for every edge. In the simplest variant of the algorithm, the initial values of the cost functions for each direct reaction are defined as 1.0. So, the cost of any path through the consequent reactions will be equal to the number of reactions. In our application of the algorithm in Genome Enhancer, we used a weighted graph, which encodes the types of reactions (direct or indirect) from TRANSPATH® database into different edge weights (costs).
The core of the algorithm is the Dijkstra's shortest-path algorithm. The upstream search starts from each molecule of the input set (subset M_x of M; e.g. transcription factors found on the previous step) and constructs the shortest-path to all nodes i of M, limited by a specified radius. After evaluating all nodes of M_x the algorithm calculates the number of visits N_i for each node i of M. This corresponds to the number of molecules of the input set M_x that can receive the signal from the node i. The values N_i are required to calculate the "master-regulator" score. The higher the value N_i the better chances has molecule i to transduce its signal to the

molecules of the initial list M_x. However, if to use the value of N_i as a straightforward "master-regulator" score would be too simplistic since it would rank those molecules high that are in general highly connected within the network and thus are likely to be false positives (so called "hub" molecules). So, in our score we incorporate the full number of TF nodes in the whole network M that can be reached from the molecule i. The score is computed for each potential master regulator and it reflects a certain balance between sensitivity and specificity of signal transduction from this master node to the downstream effector TFs:

$$S(k) = \sum_{k=1}^{k_{\max}} \frac{M_k}{\left(1 + \kappa \cdot \frac{N_k}{N_{\max,k}}\right) \cdot M_{\max,k}}$$

where k is the radius of pathway steps from the master node to the effector nodes, M_k is the number of input TFs reached by a signal from the master node within k steps, and N_k is the total number of all potential TFs in the database reached by a signal from the master node within k steps. $M_{max,k}$ and $N_{max,k}$ are the highest values among all possible master regulator nodes which helps to normalize the score to the (0,1)-interval. The higher this score, the more sensitive and more specific this master regulator is for the set of input TFs. The parameter k is the penalty which is set to 0.1 in the Genome Enhancer workflow.
Incorporation of proteomics data into the analysis of master regulators is done through application of the so called "Context algorithm", which is described below.

When we have additional experimental information about the activity of certain components of the signal transduction network (additional context information) we can use this information to adapt the search for master regulators. To do that, the context algorithm encodes this additional context information as modified edge costs in the signaling network. For instance, the proteomics data gives information about proteins that are expressed in the cell. We call them "context proteins". The idea of the approach is to attract the key node search (e.g. the underlying Dijkstra algortithm for shortest paths) towards context proteins by decreasing the costs of those edges that are close to the context proteins in the network. There are two major aspects of this algorithm:
1. Attraction ("gravity") of the shortest-paths towards context proteins. The gravity is achieved by decreasing the costs of the incoming and outgoing edges in the graph for the nodes corresponding to the context proteins. The zero cost of the edges gives the maximal "gravity" for the nodes. During the search the path that follows through this node will be always "cheaper" than other alternative paths. This way we push the search algorithm to construct its paths maximally through the context proteins when it is possible.
2. "Decay" factor. It is necessary to extend the attraction power of the "gravity" to a wider area of the network around the context proteins in order to attract the search algorithm to go as close as possible to the context proteins in case there is no path that goes through the context gene directly. To enable such "long distance" attraction power we distribute the gravity to edges of the next neighboring nodes. This pushes the shortest path algorithm to find alternative paths still at least close to the context proteins. The gravity strength reduces with increasing distance by a decay factor. The decay factor is set to 0.1 in the Genome Enhancer workflow.
The context algorithm is based on creating a graph G consisting of gravity factors [0..1], which reflect both aspects described above by modifying edge costs. Graph G is used to modify the costs of S of original graph Θ resulting in S' which can be then used for any subsequent key node analysis or possibly other shortest path based analysis.

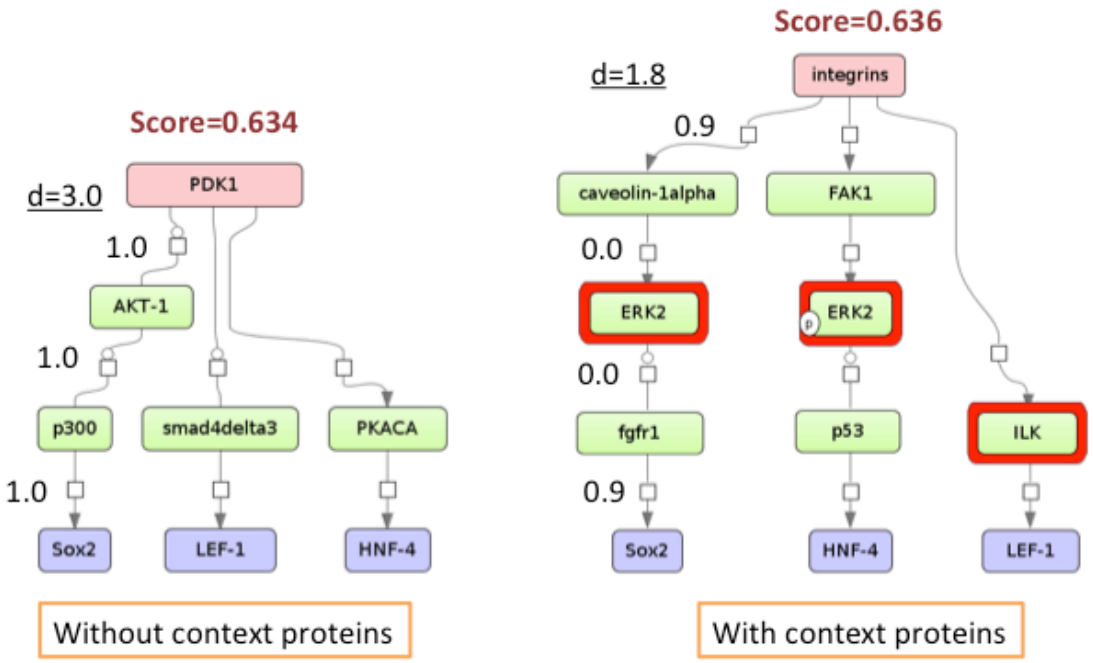
Initially the graph G is created as a copy of original signaling graph Θ, where all edges get initialized with weight zero, while the edges linked to context proteins are initialized by value 1.0. If quantitative proteomics data are available the algorithm allows to use the absolute protein expression values as the graph weights instead of 1.0. In such case the protein expression values are normalized to the range [0..1] (in such case these expression values will represent different context strengths). As the result we obtain a vector $\vec{c}$ representing the "expression" values of context proteins. The initial graph G reflects the gravity of the context proteins in the closest proximity only. The gravity is then propagated through the whole graph in downstream and upstream directions as follows. The gravity for the closest edge subsequently forwarded to the next adjacent edge by multiplying it by a user defined decay factor d.

$$D = \sum_{k=1} d^{k-1} \cdot \vec{c} \cdot S^k \qquad U = \sum_{k=1} d^{k-1} \cdot \vec{c} \cdot \left(S^T\right)^k$$

The first graph D is the result of the downstream direction and U is the result of the upstream direction of the gravity propagation through the graph G. These are two independent feed-forward runs. Both are added together, normalized to [0..1] and each weight $(D+U)_{ij}$ is replaced by $(1-(D+U)_{ij})$, such that the weights become compatible to the cost paradigm. So, the final graph G is constructed as follows:

$$G = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \dots & \vdots \\ 1 & \dots & 1 \end{pmatrix} - \frac{(D+U)}{\|D+U\|_{max}}$$

At the final step, the costs of the edges of the initial signaling graph Θ are modified by the context protein information stored in the gravity graph G by simple multiplication. The factor r scales the new edge costs such that they are equal to the initial costs of graph Θ in average. $S'_{ij} = S_{ij} \cdot G_{ij} \cdot r$ for all i and j.
The Context algorithm is illustrated in the figure below.



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 [8]). We control the error rate by the FDR threshold 0.05.
To identify feed forward loops in transduction network we calculate regulatory scores for transcription factors from CMA result and keynodes analysis. The matrices from the CMA model are converted to TFs. Keynode analysis is then performed on these TFs and resulting keynodes are sorted by keynode score. Then the keynodes, which are regulated by the model, are selected.

$$S_j = \sum_{j=1}^{M_j} \frac{1}{D_{i,j}}; \; D_{i,j} = \sum_{k=1}^{C_{i,j}} d_k$$

Here:

- j – TF (Transcription Factor),
- i – MR (Master Regulator),
- S_j – Regulatory score of the TF,
- $D_{i,j}$– Cumulative distance between MR i and TF j,
- M_j – Number of MRs whose signal can reach the TF j and whose gene is regulated by TF j,
- $C_{i,j}$ – Number of steps in the network from between MR i and TF j (shortest path),
- d_k – Distance of the step k in the shortest path between MR i and TF j.

In unweighted TRANSPATH network $d_k = 1$ for each direct signaling reaction (like A+B-C = D), $d_k >= 3$ for each semantic reaction (like A → B). dk depends on the type of reaction, on basic or orthology level of its components, and on species of its components.
In weighted TRANSPATH network d_k is changed after applying the Context Algorithm as described in [4].

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 the sum of the following ranks:

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

"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. The clinical validity score reflects the number of the highest clinical trials phase (from 1 to 4) on which the drug was ever tested for any pathology.

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 provided by the PASS software version 2020. 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. 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
6. 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
7. 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.
8. 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
9. 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
10. Kel, A., Boyarskikh, U., Stegmaier, P., Leskov, L.S., Sokolov, A.V., Yevshin, I., Mandrik, N., Stelmashenko, D., Koschmann, J., Kel-Margoulis, O. and Krull, M. Walking pathways with positive feedback loops reveal DNA methylation biomarkers of colorectal cancer. *BMC bioinformatics.* Cambridge (UK): RSC Publishing. **2019**;20(Suppl 4):119:1-20. doi:10.1186/s12859-019-2687-7
11. 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
12. 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.
13. Filimonov DA, Poroikov VV. Prognosis of specters of biological activity of organic molecules. *Russian chemical journal*. **2006**;50(2):66-75 (russ)
14. Filimonov D, Poroikov V, Borodina Y, Gloriozova T. Chemical Similarity Assessment Through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *ChemInform.* **1999**;39(4):666-670. doi:10.1021/ci980335o

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 - Detailed report. Composite modules and master regulators (highly methylated genes in Hypertension vs. Control).
2. Supplementary table 2 - Detailed report. Composite modules and master regulators (low methylated genes in Hypertension vs. Control).
3. Supplementary table 3 - 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.