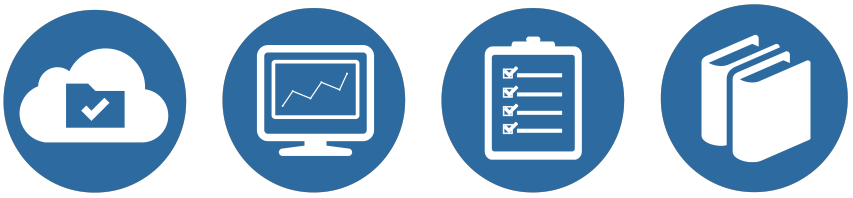
RPS6KA2 and DUSP9 are promising druggable targets for treating Hypertension that control activity of ETS1, ZBTB33 and BCL6 transcription factor on of highly methylated genes

Demo User
geneXplain GmbH
info@genexplain.com
Data received on 08/04/2022 ; Run on 05/11/2024 ; Report generated on 05/11/2024

Genome Enhancer release 3.5 (TRANSFAC®, TRANSPATH® and HumanPSD™ release 2024.2)



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 highly methylated genes: ETS1, ZBTB33 and BCL6. The subsequent network analysis suggested

- RSK
- MKP-4

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: Imatinib, Flavopiridol, Moexipril and 3-METHYL-1,6,8-TRIHYDROXYANTHRAQUINONE.

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 highly methylated 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".

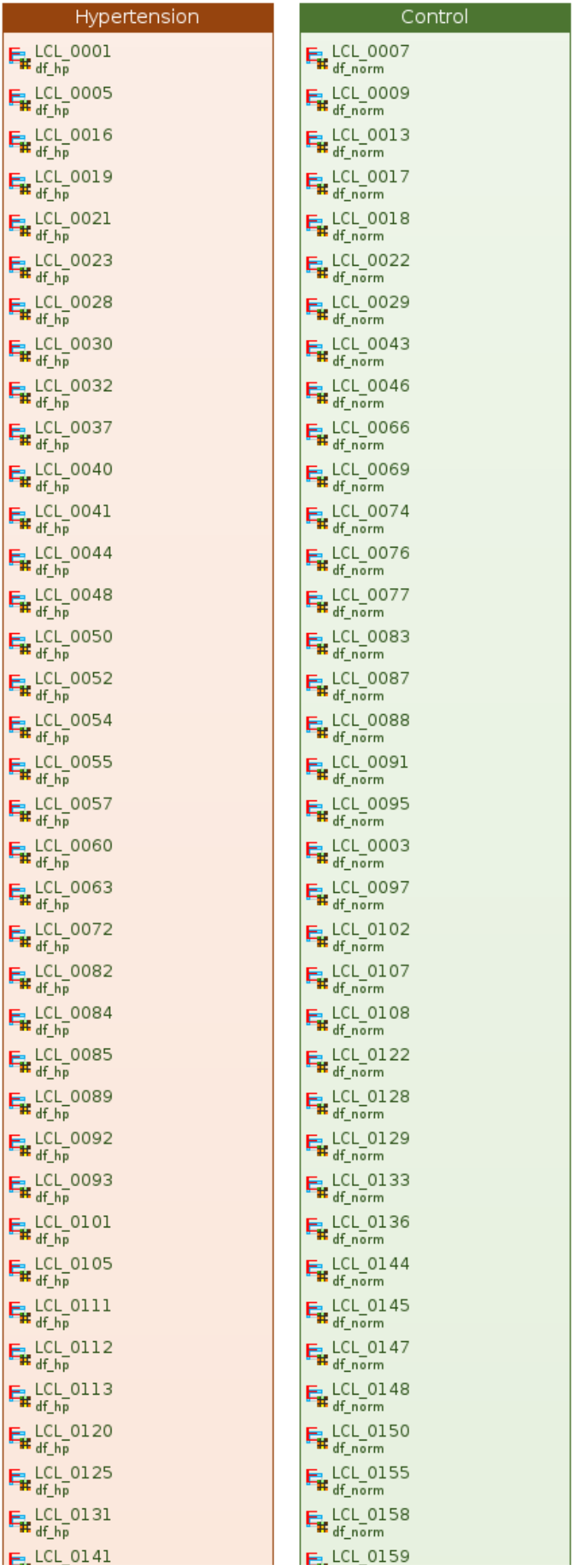
The "upstream analysis" approach has now been extended by a third step that reveals known drugs suitable to inhibit (or activate) the identified molecular targets in the context of the disease under study. This step is performed by using information from HumanPSD™ database [5]. In addition, some known 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

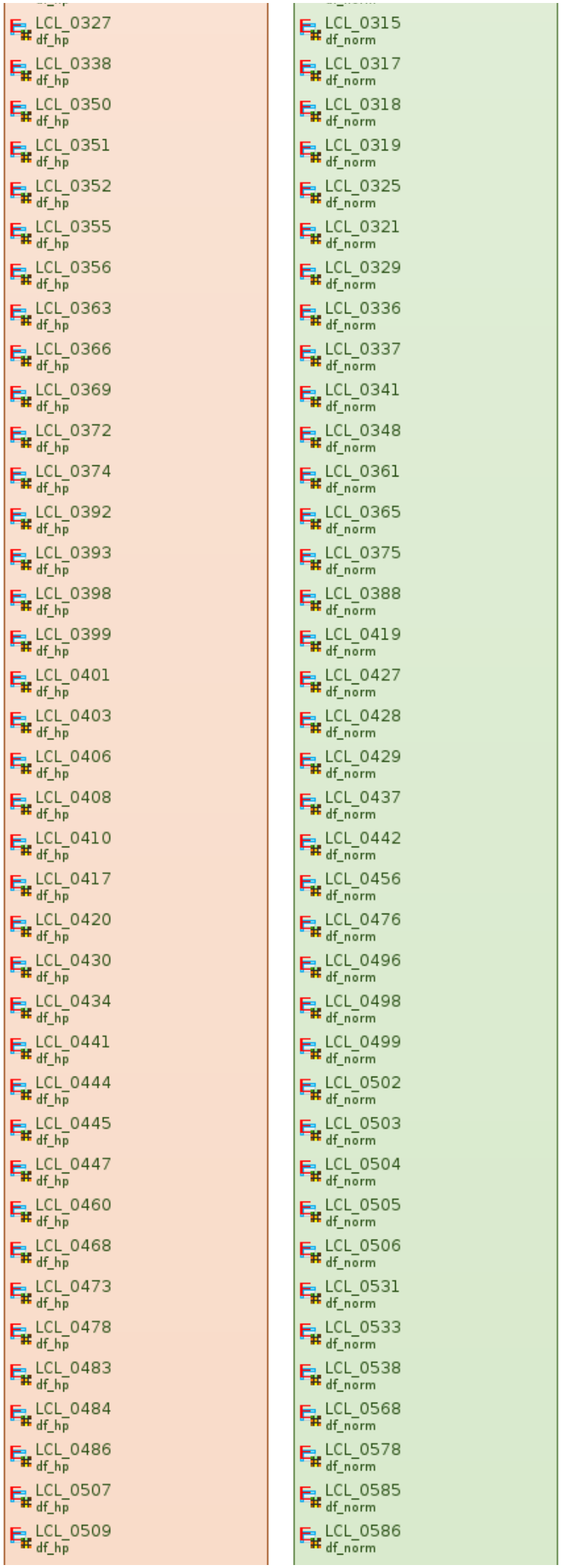
For this study the following experimental data was used:

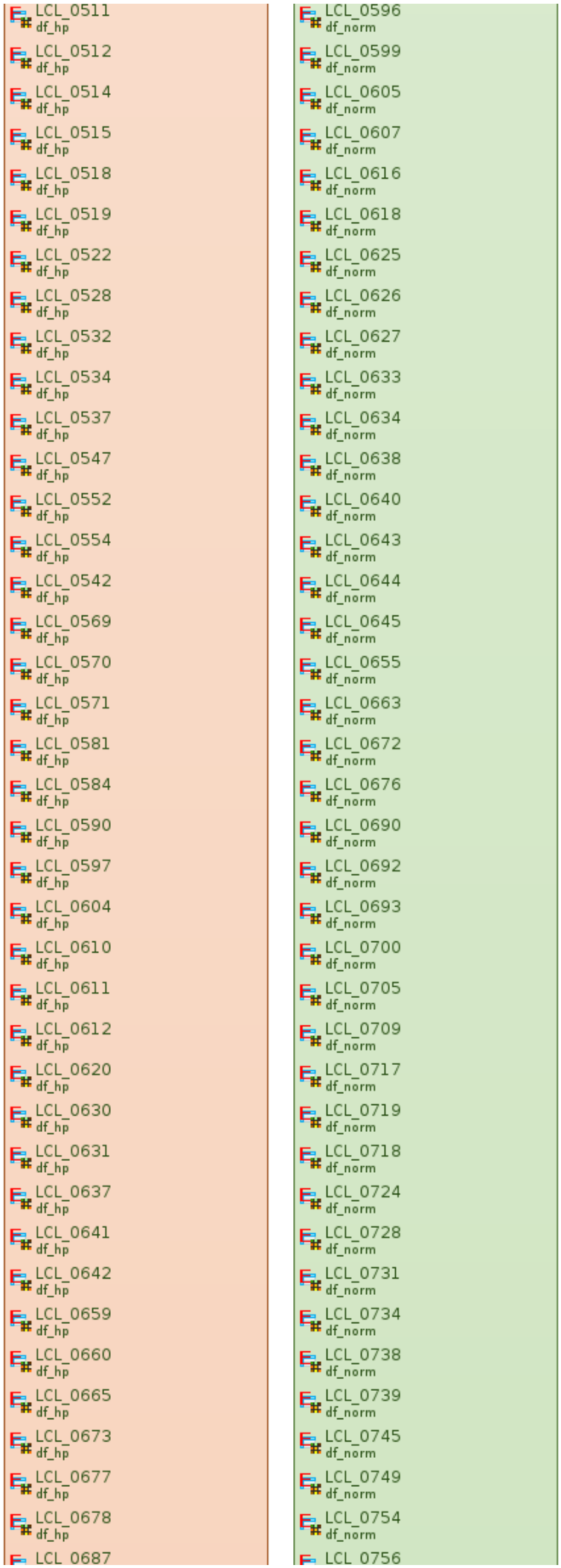
Table 1. Experimental datasets used in the study

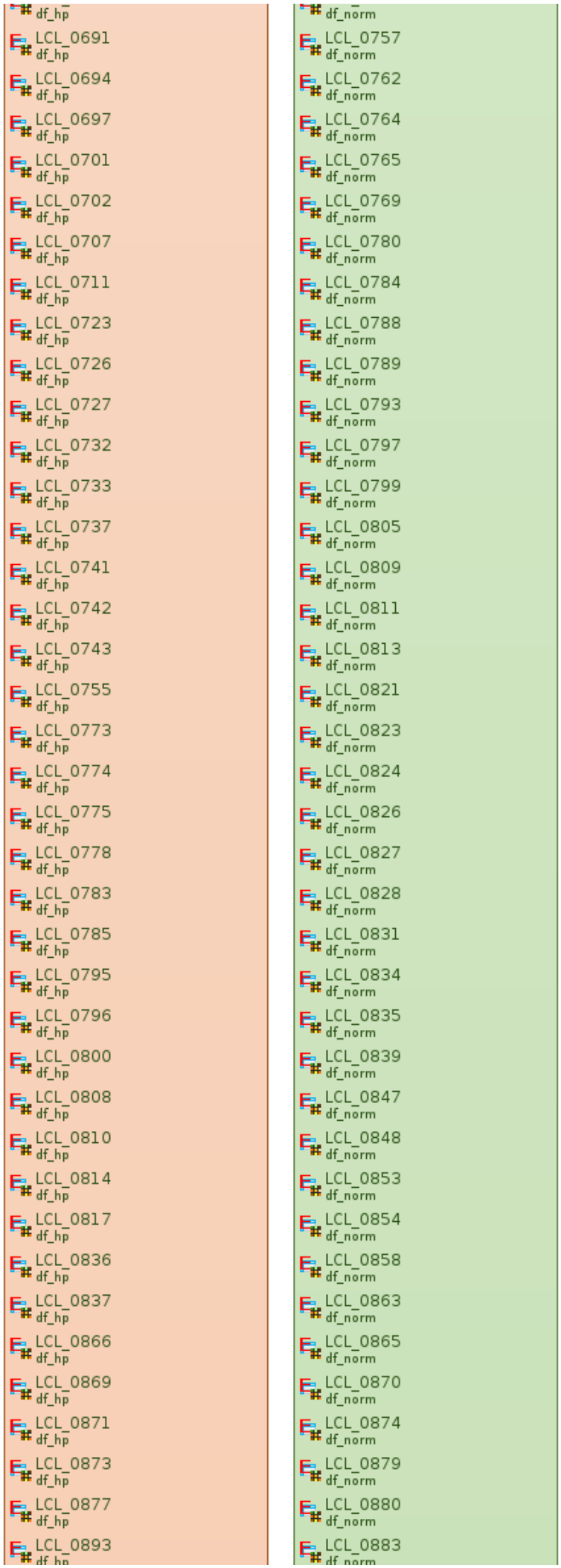
File name	Data type
df_hp	Epigenomics
df_norm	Epigenomics

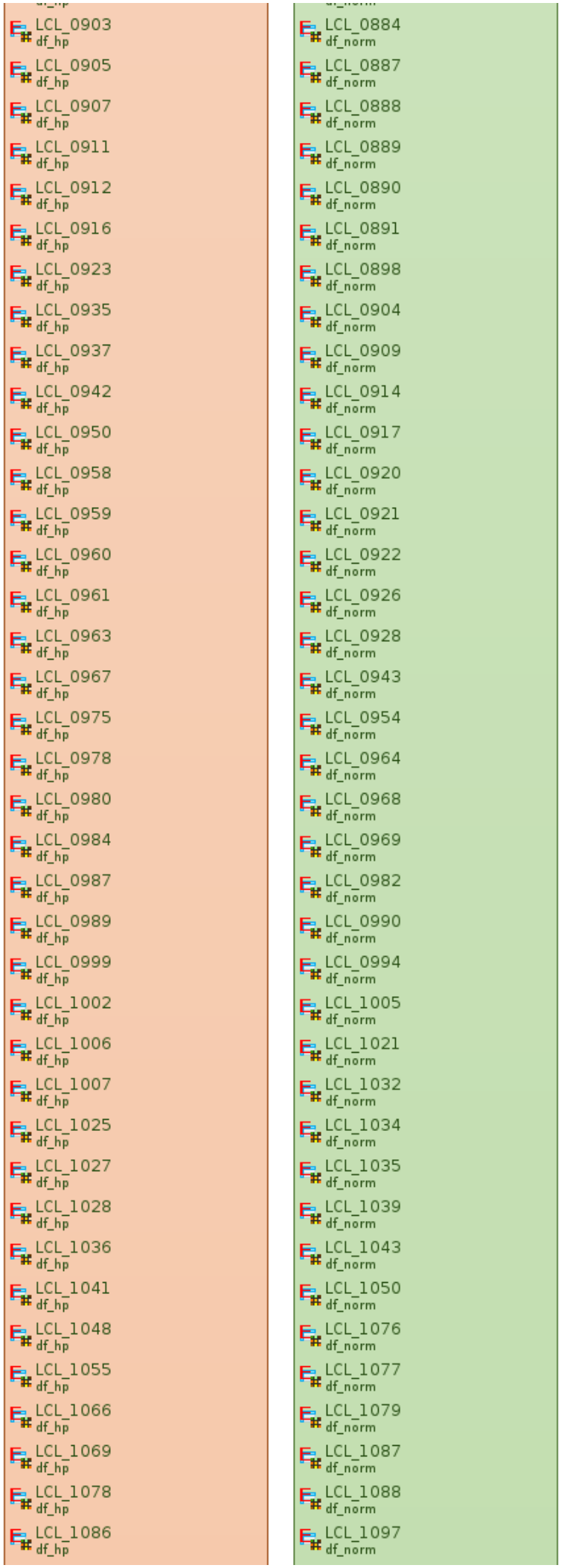

Hypertension
LCL_0001 df_hp
LCL_0005 df_hp
LCL_0016 df_hp
LCL_0019 df_hp
LCL_0021 df_hp
LCL_0023 df_hp
LCL_0028 df_hp
LCL_0030 df_hp
LCL_0032 df_hp
LCL_0037 df_hp
LCL_0040 df_hp
LCL_0041 df_hp
LCL_0044 df_hp
LCL_0048 df_hp
LCL_0050 df_hp
LCL_0052 df_hp
LCL_0054 df_hp
LCL_0055 df_hp
LCL_0057 df_hp
LCL_0060 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
LCL_0131 df_hp
LCL_0141
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_0158 df_norm
LCL_0159

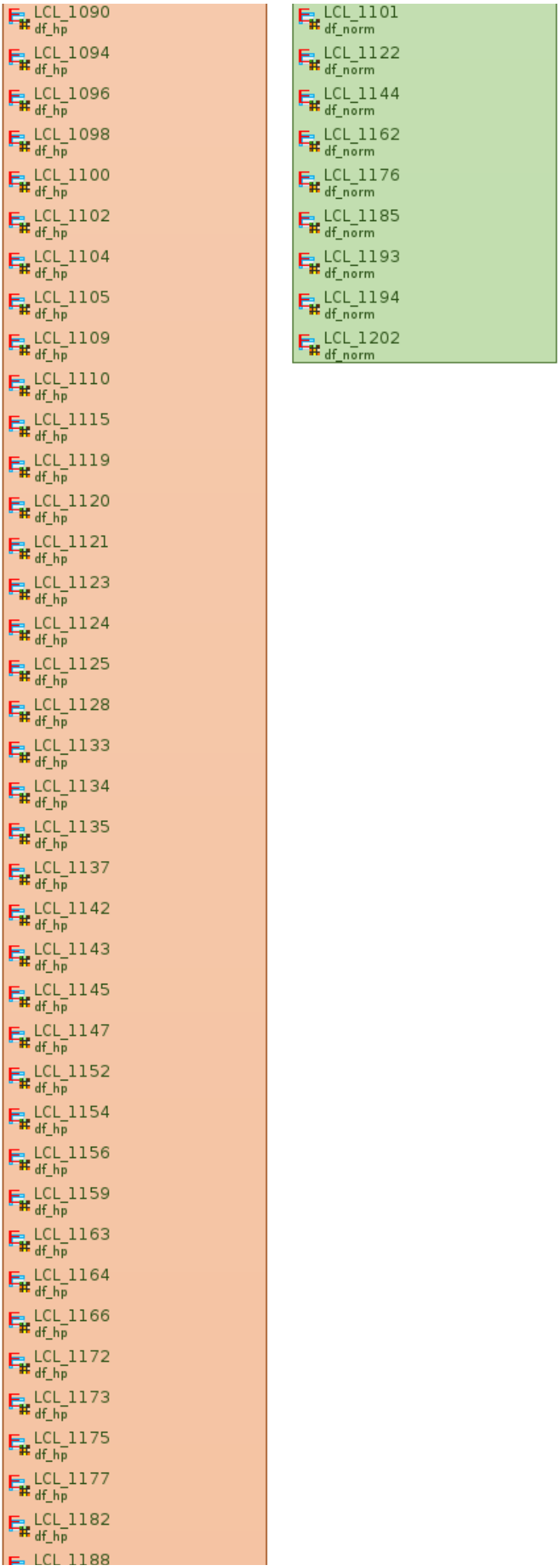
df_hp	df_norm
LCL_0162	LCL_0135
LCL_0171	LCL_0164
LCL_0175	LCL_0165
LCL_0177	LCL_0167
LCL_0189	LCL_0168
LCL_0190	LCL_0174
LCL_0192	LCL_0176
LCL_0193	LCL_0179
LCL_0194	LCL_0183
LCL_0198	LCL_0188
LCL_0199	LCL_0196
LCL_0200	LCL_0197
LCL_0205	LCL_0204
LCL_0211	LCL_0209
LCL_0213	LCL_0219
LCL_0214	LCL_0220
LCL_0217	LCL_0221
LCL_0233	LCL_0223
LCL_0235	LCL_0224
LCL_0236	LCL_0232
LCL_0237	LCL_0234
LCL_0241	LCL_0244
LCL_0246	LCL_0245
LCL_0255	LCL_0253
LCL_0256	LCL_0258
LCL_0257	LCL_0260
LCL_0262	LCL_0264
LCL_0265	LCL_0274
LCL_0266	LCL_0281
LCL_0272	LCL_0290
LCL_0280	LCL_0291
LCL_0285	LCL_0293
LCL_0298	LCL_0294
LCL_0299	LCL_0295
LCL_0310	LCL_0301
LCL_0314	LCL_0303
LCL_0322	LCL_0306
LCL_0324	LCL_0307


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_0484 df_hp
LCL_0486 df_hp
LCL_0507 df_hp
LCL_0509 df_hp
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_0568 df_norm
LCL_0578 df_norm
LCL_0585 df_norm
LCL_0586 df_norm


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 df_hp
LCL_0660 df_hp
LCL_0665 df_hp
LCL_0673 df_hp
LCL_0677 df_hp
LCL_0678 df_hp
LCL_0687
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_norm
LCL_0738 df_norm
LCL_0739 df_norm
LCL_0745 df_norm
LCL_0749 df_norm
LCL_0754 df_norm
LCL_0756


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_hp
LCL_0837 df_hp
LCL_0866 df_hp
LCL_0869 df_hp
LCL_0871 df_hp
LCL_0873 df_hp
LCL_0877 df_hp
LCL_0893 df_hp
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_norm
LCL_0863 df_norm
LCL_0865 df_norm
LCL_0870 df_norm
LCL_0874 df_norm
LCL_0879 df_norm
LCL_0880 df_norm
LCL_0883 df_norm


LCL_0903 df_hp
LCL_0905 df_hp
LCL_0907 df_hp
LCL_0911 df_hp
LCL_0912 df_hp
LCL_0916 df_hp
LCL_0923 df_hp
LCL_0935 df_hp
LCL_0937 df_hp
LCL_0942 df_hp
LCL_0950 df_hp
LCL_0958 df_hp
LCL_0959 df_hp
LCL_0960 df_hp
LCL_0961 df_hp
LCL_0963 df_hp
LCL_0967 df_hp
LCL_0975 df_hp
LCL_0978 df_hp
LCL_0980 df_hp
LCL_0984 df_hp
LCL_0987 df_hp
LCL_0989 df_hp
LCL_0999 df_hp
LCL_1002 df_hp
LCL_1006 df_hp
LCL_1007 df_hp
LCL_1025 df_hp
LCL_1027 df_hp
LCL_1028 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_0884 df_norm
LCL_0887 df_norm
LCL_0888 df_norm
LCL_0889 df_norm
LCL_0890 df_norm
LCL_0891 df_norm
LCL_0898 df_norm
LCL_0904 df_norm
LCL_0909 df_norm
LCL_0914 df_norm
LCL_0917 df_norm
LCL_0920 df_norm
LCL_0921 df_norm
LCL_0922 df_norm
LCL_0926 df_norm
LCL_0928 df_norm
LCL_0943 df_norm
LCL_0954 df_norm
LCL_0964 df_norm
LCL_0968 df_norm
LCL_0969 df_norm
LCL_0982 df_norm
LCL_0990 df_norm
LCL_0994 df_norm
LCL_1005 df_norm
LCL_1021 df_norm
LCL_1032 df_norm
LCL_1034 df_norm
LCL_1035 df_norm
LCL_1039 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_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
LCL_1156
df_hp
LCL_1159
df_hp
LCL_1163
df_hp
LCL_1164
df_hp
LCL_1166
df_hp
LCL_1172
df_hp
LCL_1173
df_hp
LCL_1175
df_hp
LCL_1177
df_hp
LCL_1182
df_hp
LCL_1188
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

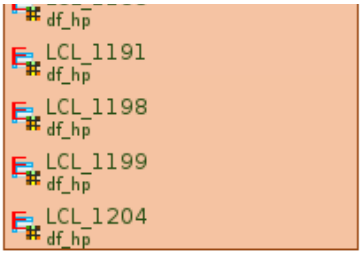


Figure 1. Annotation diagram of experimental data used in this study. With the colored boxes we show those sub-categories of the data that are compared in our analysis.

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		41	5	0
ENSG00000250349	novel proline rich Gla (G-carboxyglutamic acid) 1 (PRRG1) and tetraspanin 7 (TSPAN7) protein	ENSG00000250349		38	2	2
ENSG00000081853	protocadherin gamma subfamily A, 2	PCDHGA2		37	5	0
ENSG00000204970	protocadherin alpha 1	PCDHA1		34	0	0
ENSG00000254245	protocadherin gamma subfamily A, 3	PCDHGA3		34	5	0
ENSG00000204969	protocadherin alpha 2	PCDHA2		33	3	0
ENSG00000254221	protocadherin gamma subfamily B, 1	PCDHGB1		31	5	0
ENSG00000255408	protocadherin alpha 3	PCDHA3		30	2	0
ENSG00000204967	protocadherin alpha 4	PCDHA4		28	1	0
ENSG00000204965	protocadherin alpha 5	PCDHA5		27	3	1

3.2. Functional classification of genes

A functional analysis of highly methylated 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 2-4 show the most significant categories.

Highly methylated genes in Hypertension vs. Control:

300 top methylated genes were taken for the mapping.

GO (biological process)

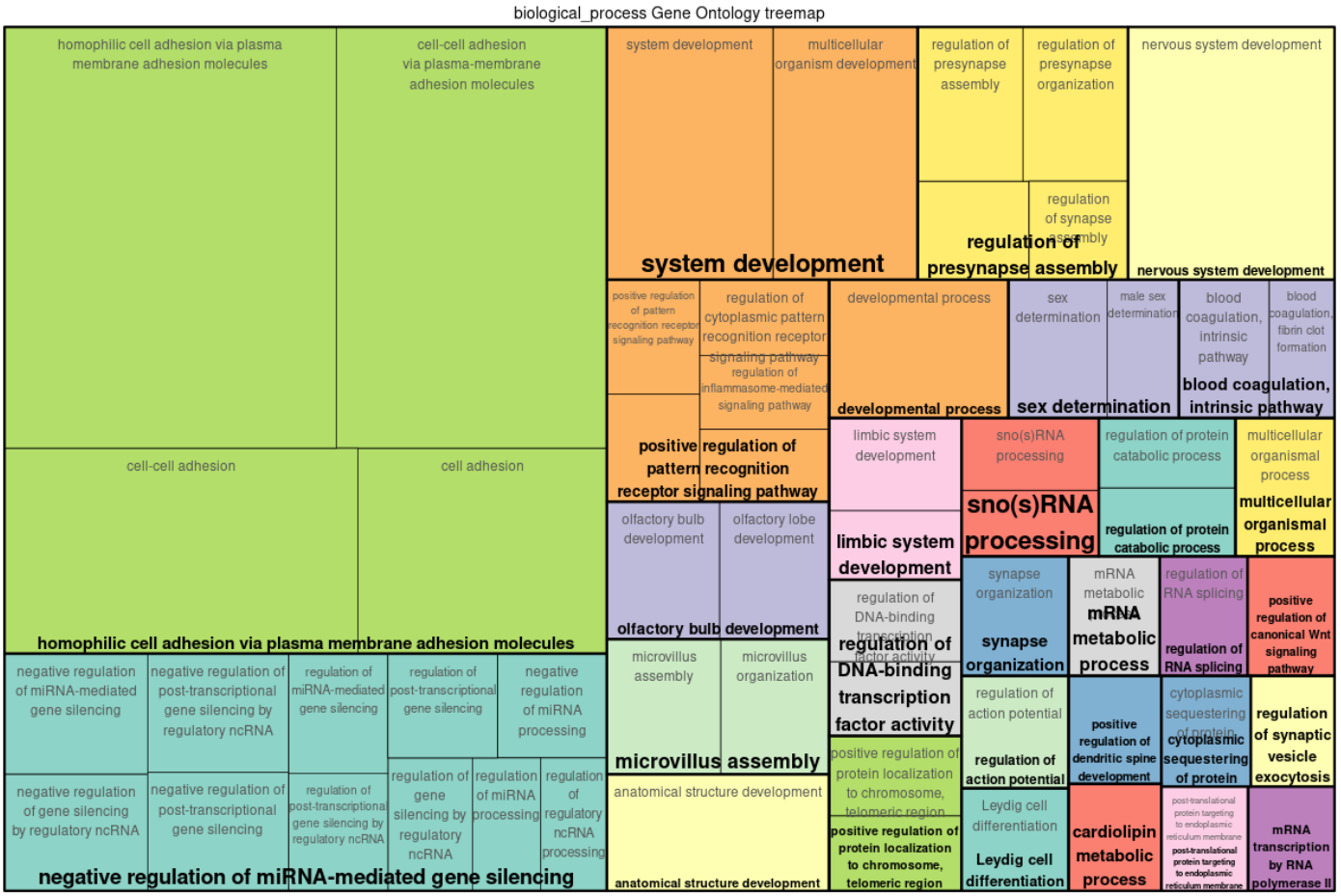


Figure 2. Enriched GO (biological process) of highly methylated genes in Hypertension vs. Control.
Full classification →

TRANSPATH® Pathways (2024.2)

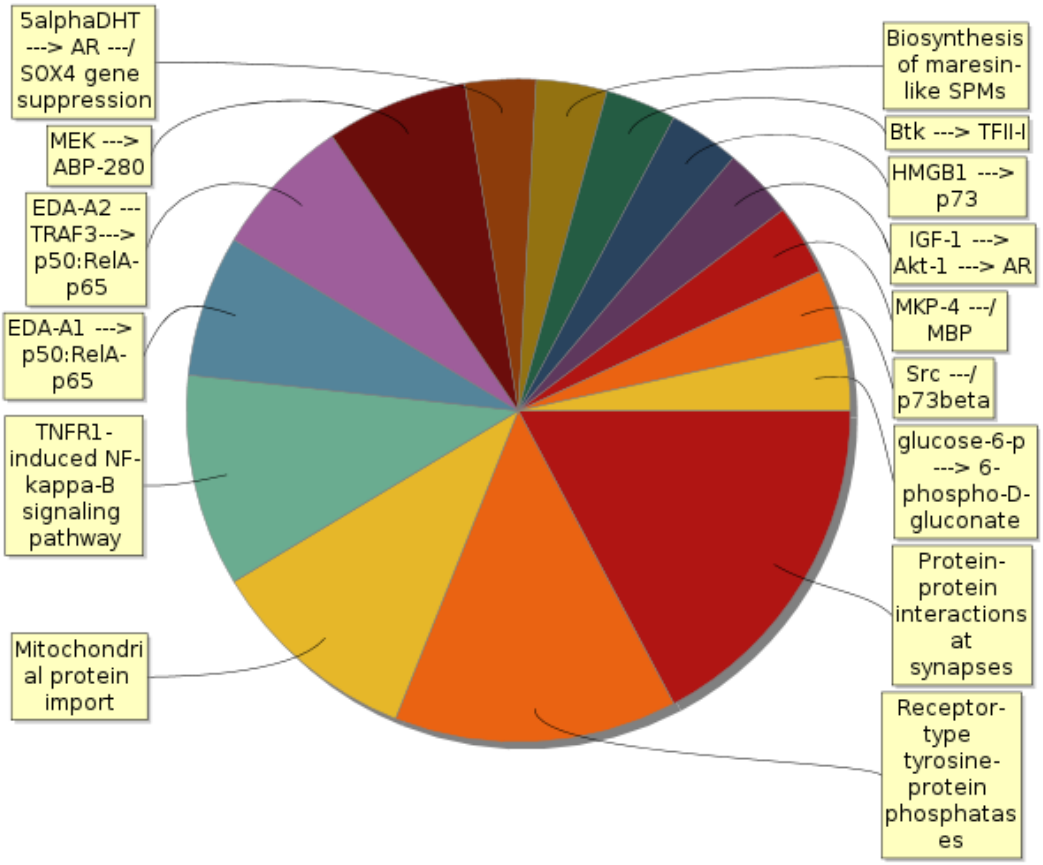


Figure 3. Enriched TRANSPATH® Pathways (2024.2) of highly methylated genes in Hypertension vs. Control.
Full classification →

HumanPSD(TM) disease (2024.2)

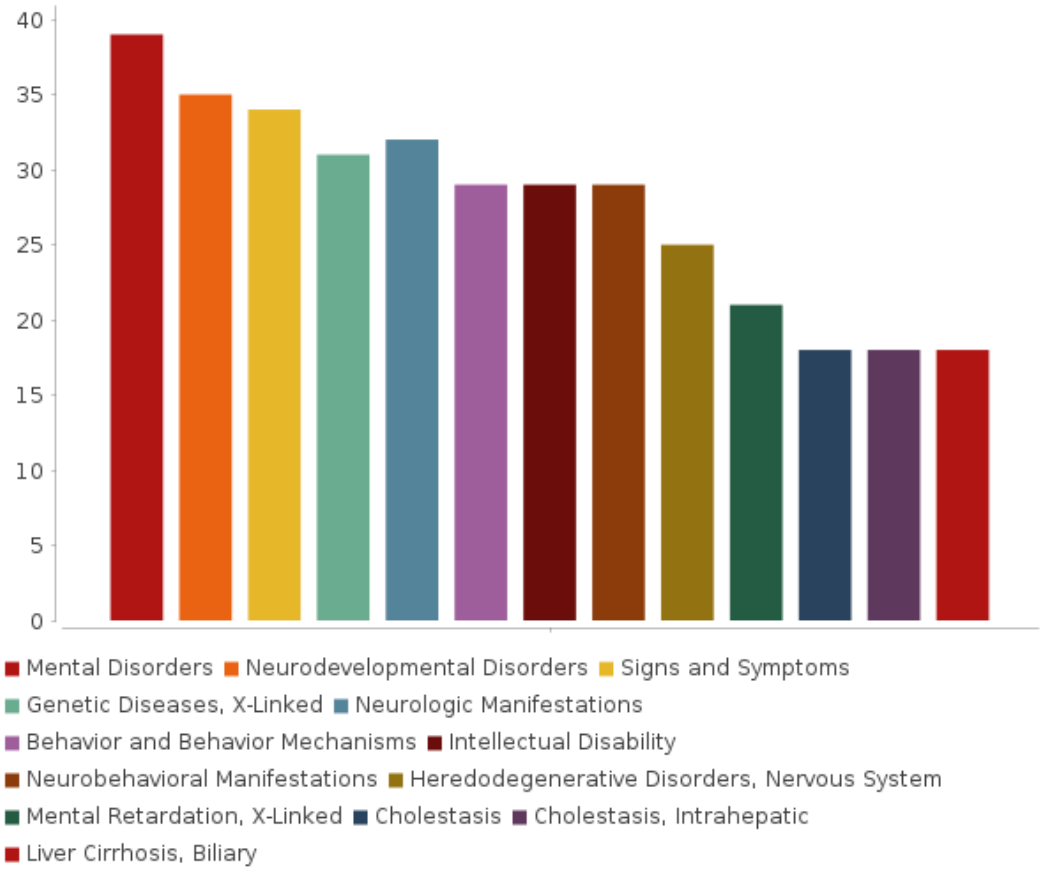
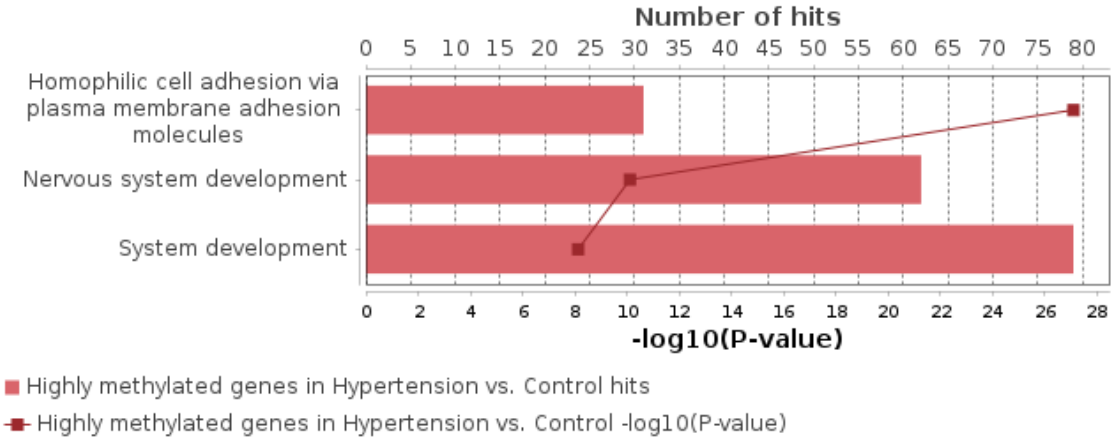


Figure 4. Enriched HumanPSD(TM) disease (2024.2) of highly methylated genes in Hypertension vs. Control. The size of the bars correspond to the number of biomarkers of the given disease found among the input set.

Full classification →

The result of overall Gene Ontology (GO) analysis of the highly methylated genes of the studied pathology can be summarized by the following diagram, revealing the most significant functional categories overrepresented among the observed (highly methylated 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 [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 Epigenomics data from the track(s) "Methylation track" to predict positions of potential ***enhancers*** regulating the highly methylated 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 highly methylated 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) [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 (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$BCL6_10	V$NF1B_Q6_01	V$TIEG1_Q6	V$ETS1_06
0.73; N=3	0.96; N=3	0.98; N=3	0.94; N=3

Module width: 129

Module 2:

V$NF1A_Q6_01	V$ZBTB7B_06	V$ZBTB33_10	V$BEN_01	V$LEF1_Q5
0.99; N=3	0.88; N=3	0.96; N=3	0.96; N=3	0.96; N=3

Module width: 155

Model score (-p*log10(pval)): 27.81
Wilcoxon p-value (pval): 1.73e-54
Penalty (p): 0.517
Average yes-set score: 8.69
Average no-set score: 6.92
AUC: 0.73
Separation point: 7.53
False-positive: 35.65%
False-negative: 26.84%

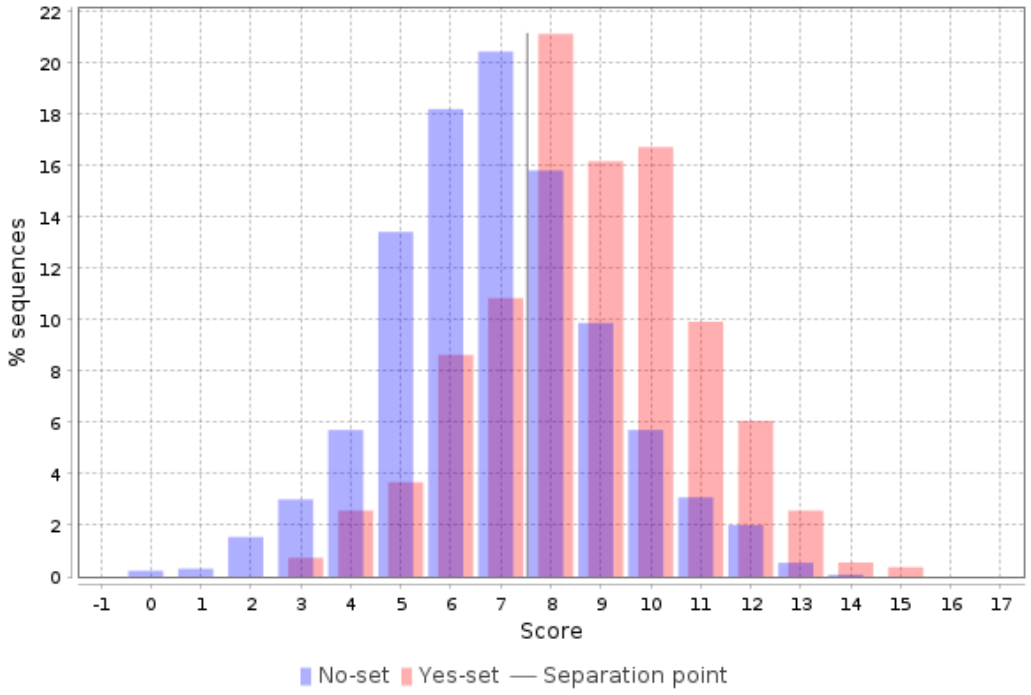


Table 3. List of top ten highly methylated genes in Hypertension vs. Control with identified enhancers in their regulatory regions. ***CMA score*** *- the score of the CMA model of the enhancer identified in the regulatory region.*

See full table →

Ensembl IDs	Gene symbol	Gene description	CMA score	Factor names
ENSG00000112541	PDE10A	phosphodiesterase 10A	15.95	NF-1A(h), GTF2IRD1(h), LEF-1(h), KLF10(h), ZBTB7B(h), c-Ets-1(h), ZBTB33(h)...
ENSG00000284779		novel protein	15.85	LEF-1(h), NF-1A(h), GTF2IRD1(h), ZBTB33(h), ZBTB7B(h), NF-1B(h), c-Ets-1(h)...
ENSG00000205116	TMEM88B	transmembrane protein 88B	15.53	NF-1A(h), ZBTB7B(h), BCL-6(h), NF-1B(h), GTF2IRD1(h), KLF10(h), ZBTB33(h)...
ENSG00000168267	PTF1A	pancreas associated transcription factor 1a	15.15	GTF2IRD1(h), KLF10(h), ZBTB33(h), LEF-1(h), ZBTB7B(h), NF-1A(h), NF-1B(h)...
ENSG00000155269	GPR78	G protein-coupled receptor 78	14.9	GTF2IRD1(h), NF-1A(h), ZBTB33(h), ZBTB7B(h), c-Ets-1(h), KLF10(h), NF-1B(h)...
ENSG00000178947	SMIM10L2A	small integral membrane protein 10 like 2A	14.69	LEF-1(h), ZBTB7B(h), NF-1A(h), NF-1B(h), BCL-6(h), ZBTB33(h), c-Ets-1(h)...
ENSG00000171631	P2RY6	pyrimidinergic receptor P2Y6	14.57	ZBTB7B(h), NF-1A(h), BCL-6(h), KLF10(h), ZBTB33(h), LEF-1(h), NF-1B(h)...
ENSG00000275832	ARHGAP23	Rho GTPase activating protein 23	14.56	ZBTB7B(h), GTF2IRD1(h), ZBTB33(h), NF-1A(h), c-Ets-1(h), BCL-6(h), KLF10(h)...
ENSG00000248479		novel transcript	14.53	KLF10(h), GTF2IRD1(h), ZBTB7B(h), NF-1A(h), LEF-1(h), c-Ets-1(h), ZBTB33(h)...
ENSG00000070729	CNGB1	cyclic nucleotide gated channel subunit beta 1	14.48	BCL-6(h), NF-1B(h), LEF-1(h), KLF10(h), ZBTB33(h), c-Ets-1(h), NF-1A(h)...

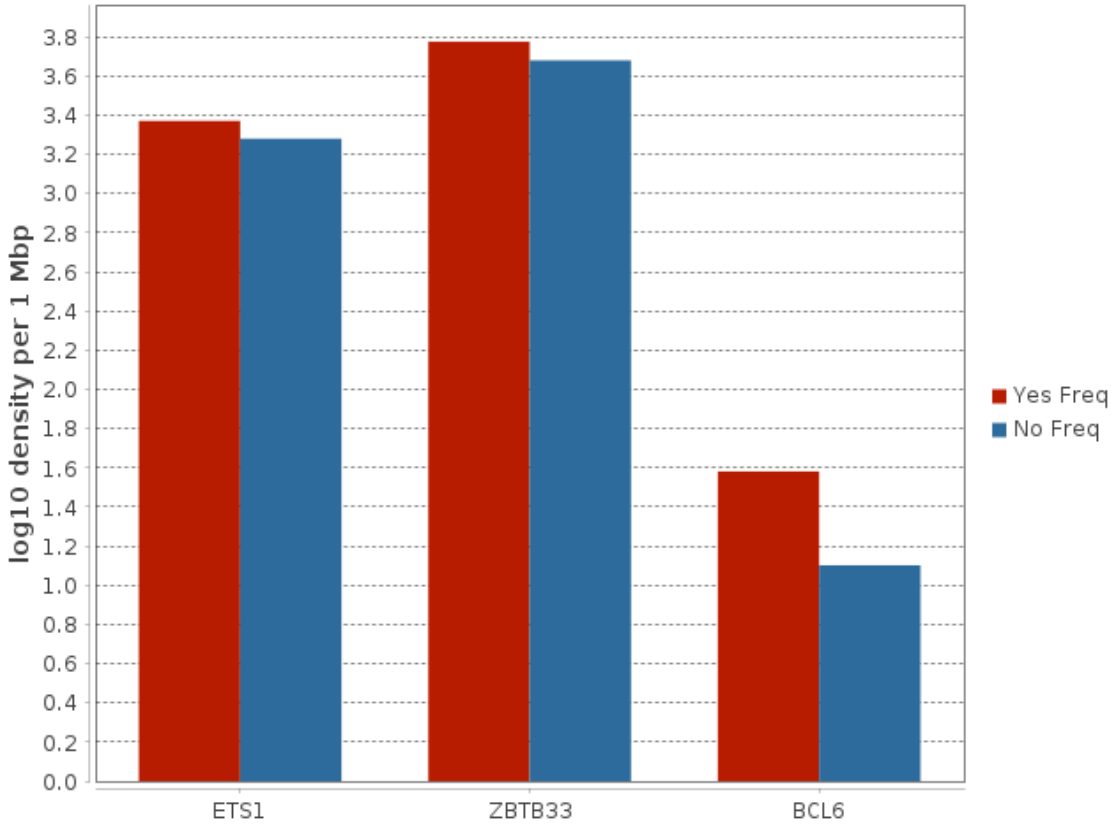
On the basis of the enhancer models we identified transcription factors potentially regulating the ***target genes*** of our interest. We found 9 transcription factors controlling expression of highly methylated genes in Hypertension vs. Control (see Table 4).

Table 4. Transcription factors of the predicted enhancer model potentially regulating the highly methylated 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
MO000059013	ETS1	ETS proto-oncogene 1, transcription factor	2.78	1.23
MO000056592	ZBTB33	zinc finger and BTB domain containing 33	2.32	1.25
MO000026319	BCL6	BCL6 transcription repressor	2.25	3.01
MO000159782	LEF1	lymphoid enhancer binding factor 1	2.16	1.69
MO000070133	GTF2IRD1	GTF2I repeat domain containing 1	2.09	1.28
MO000028709	NFIB	nuclear factor I B	2.05	1.46
MO000095587	ZBTB7B	zinc finger and BTB domain containing 7B	2.05	1.48
MO000028708	NFIA	nuclear factor I A	1.79	1.46
MO000028727	KLF10	KLF transcription factor 10	0.33	1.81

The following diagram represents the key transcription factors, which were predicted to be potentially regulating highly methylated genes in the analyzed pathology: ETS1, ZBTB33 and BCL6.



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 Table 5.

Table 5. 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
MO000032712	MKP-4(h)	DUSP9	dual specificity phosphatase 9	56
MO000058267	RSK(h){p}	RPS6KA1, RPS6KA2, RPS6KA3	ribosomal protein S6 kinase A1, ribosomal protein S6 kinase A2, ribosomal protein S6 kinase A3	98
MO000022013	ABP-280(h)	FLNA	filamin A	115
MO000042811	ABP-280(h){pS2152}	FLNA	filamin A	115
MO000083277	ABP-280-isoform1(h)	FLNA	filamin A	115
MO000255314	ABP-280-isoform2(h)	FLNA	filamin A	115
MO001093954	TLE tetramer:XIAP	TLE1, TLE2, TLE3, TLE4, XIAP	TLE family member 1, transcriptional corepressor, TLE family member 2, transcriptional corepressor, ...	121
MO001094109	XIAP:ub-TLE	TLE1, TLE2, TLE3, TLE4, XIAP	TLE family member 1, transcriptional corepressor, TLE family member 2, transcriptional corepressor, ...	133
MO000329204	Cdk6(h):cyclinD3-isoform1(h)	CCND3, CDK6	cyclin D3, cyclin dependent kinase 6	154
MO000016677	EGFR(h)	EGFR	epidermal growth factor receptor	176

The intracellular regulatory pathways controlled by the above-mentioned master regulators are depicted in Figure 5. This diagram displays the connections between identified transcription factors, which play important roles in the regulation of highly methylated genes, and selected master regulators, which are responsible for the regulation of these TFs.

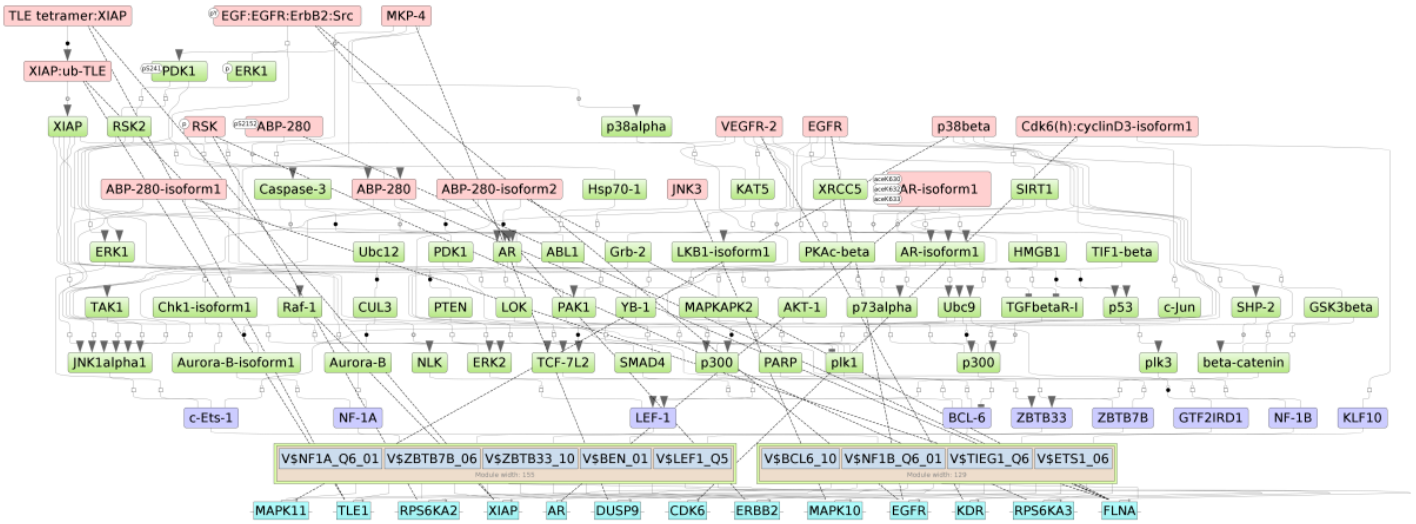


Figure 5. Diagram of intracellular regulatory signal transduction pathways of highly methylated genes in Hypertension vs. Control. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate molecules, which have been added to the network during the search for master regulators from selected TFs.

See full diagram →

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 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 6. 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
RPS6KA2	ribosomal protein S6 kinase A2	27	98
RPS6KA3	ribosomal protein S6 kinase A3	37	98
XIAP	X-linked inhibitor of apoptosis	21	121
CDK6	cyclin dependent kinase 6	16	154
EGFR	epidermal growth factor receptor	95	176
VHL	von Hippel-Lindau tumor suppressor	2	179

*Table 7. 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.*

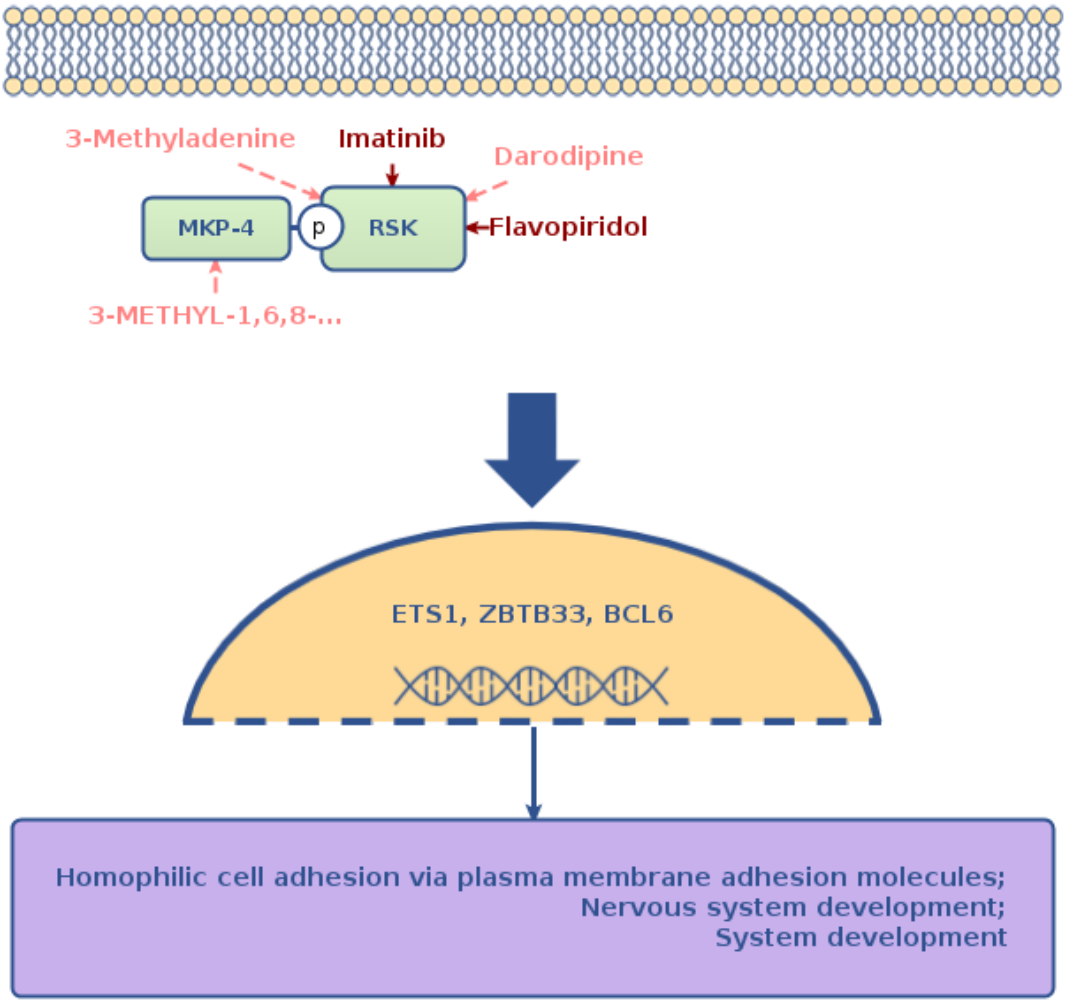
See full table →

Gene symbol	Gene Description	Druggability score	Total rank
DUSP9	dual specificity phosphatase 9	35.02	56
RPS6KA2	ribosomal protein S6 kinase A2	9.78	98
RPS6KA3	ribosomal protein S6 kinase A3	19.33	98
CDK6	cyclin dependent kinase 6	22.91	154
EGFR	epidermal growth factor receptor	23.3	176
MAPK11	mitogen-activated protein kinase 11	97.83	211

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:

- RSK
- MKP-4

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, Flavopiridol, Darodipine, 3-METHYL-1,6,8-TRIHYDROXYANTHRAQUINONE and 3-Methyladenine, 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 8 and 9), 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 8. 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	MAPK10, RPS6KA3, BCR, MAP4K4, BMPR1A, MARK3, IRAK3, NTRK1, NTRK3, PRKAA2, KDR, FLT4, INSR, LTK, EGFR, PRKCH, ACVR2A, PRKG1, ERBB2, CAMK2B, PRKD1, PKMYT1, PIP5K1A, IGF1R, BMX, LCK, ALK, PIM2, EPHB2, RPS6KA2, MAPK4, CSNK1G2, PAK3, INSRR, BTK, CAMKK1, EPHA6, MAPK11, EPHA5, TGFBR2, EPHA8, FYN, FES, RET, ABL2, STK3	94	2	small molecule,approved
Sirolimus	MAPK10, RPS6KA3, PIM2, WT1, MARK3, MTOR, FGF2, PRKCZ, PGR, CAMKK1, MAPK11, PRKD1, STK3	92	2	small molecule,approved,investigational
Pazopanib	MAPK10, RPS6KA3, MAP4K4, BMPR1A, MARK3, IRAK3, NTRK1, NTRK3, PRKAA2, KDR, FLT4, INSR, LTK, EGFR, PRKCH, ACVR2A, PRKG1, ERBB2, CAMK2B, PRKD1, PKMYT1, PIP5K1A, IGF1R, BMX, LCK, ALK, PIM2, EPHB2, RPS6KA2, MAPK4, CSNK1G2, PAK3, INSRR, BTK, CAMKK1, EPHA6, MAPK11, EPHA5, TGFBR2, EPHA8, FYN, FES, RET, ABL2, STK3	92	1	small molecule,approved
Sorafenib	MAPK10, RPS6KA3, MAP4K4, BMPR1A, MARK3, IRAK3, NTRK1, NTRK3, PRKAA2, KDR, FLT4, INSR, LTK, EGFR, PRKCH, ACVR2A, PRKG1, ERBB2, CAMK2B, PRKD1, PKMYT1, PIP5K1A, IGF1R, BMX, LCK, ALK, PIM2, EPHB2, RPS6KA2, MAPK4, CSNK1G2, PAK3, INSRR, PRKCZ, BTK, CAMKK1, EPHA6, MAPK11, EPHA5, TGFBR2, EPHA8, FYN, FES, RET, ABL2, STK3	91	1	small molecule,approved,investigational
Curcumin	MAPK10, HDAC4, WNT3A, WT1, CDK6, SMAD2, HDAC6, CDH1, HK2, AR, CCNA1, EGFR, POU5F1, MMP9, CXCR4, NOTCH1, TAFAZZIN, NANOG, AGAP2, MTOR, MAPK4, DNMT3A, EPAS1, MMP14, SKP2, YAP1, SQSTM1, XIAP, MAPK11, TP73, DLC1	91	1	small molecule, approved, experimental, investigational

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

Repurposing drugs

Table 9. 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
Flavopiridol	MAPK10, RPS6KA3, CDK6, MAP4K4, BMPR1A, MARK3, IRAK3, NTRK1, NTRK3, PRKAA2, KDR, FLT4, INSR, LTK, EGFR, PRKCH, ACVR2A, PRKG1, ERBB2, CAMK2B, PRKD1, PKMYT1, PIP5K1A, IGF1R, BMX, LCK, ALK, PIM2, EPHB2, RPS6KA2, MAPK4, CSNK1G2, PAK3, INSRR, BTK, CAMKK1, EPHA6, XIAP, MAPK11, EPHA5, TGFBR2, EPHA8, FYN, FES, RET, ABL2, STK3	87	PHASE1: Brain Abscess, Carcinoma, Carcinoma, Non-Small-Cell Lung, Carcinoma, Small Cell, Cytopenia, Esophageal Neoplasms, Intestinal Neoplasms, Leukemia, Leukemia, Lymphocytic, Chronic, B-Cell, Leukemia, Lymphoid, Leukemia, Prolymphocytic, Lymphoma, Lymphoma, B-Cell, Lymphoma, B-Cell, Marginal Zone, Lymphoma, Follicular, Lymphoma, Large B-Cell, Diffuse, Lymphoma, Mantle-Cell, Lymphoma, Non-Hodgkin, Mesothelioma, Mesothelioma, Malignant, Multiple Myeloma, Neoplasms, Prostatic Neoplasms, Prostatic Neoplasms, Castration-Resistant, Recurrence, Thrombocytopenia, Waldenstrom Macroglobulinemia
Dasatinib	MAPK10, RPS6KA3, BCR, MAP4K4, BMPR1A, MARK3, IRAK3, NTRK1, NTRK3, PRKAA2, KDR, FLT4, INSR, LTK, EGFR, PRKCH, ACVR2A, PRKG1, ERBB2, CAMK2B, PRKD1, PKMYT1, PIP5K1A, IGF1R, BMX, LCK, ALK, PIM2, EPHB2, RPS6KA2, MAPK4, CSNK1G2, PAK3, INSRR, PRKCZ, BTK, CAMKK1, EPHA6, MAPK11, EPHA5, TGFBR2, EPHA8, FYN, FES, RET, ABL2, STK3	87	EARLY_PHASE1: Glioblastoma, Recurrence
CI-1033	MAPK10, RPS6KA3, MAP4K4, BMPR1A, MARK3, IRAK3, NTRK1, NTRK3, PRKAA2, KDR, FLT4, INSR, LTK, EGFR, PRKCH, ACVR2A, PRKG1, ERBB2, CAMK2B, PRKD1, PKMYT1, PIP5K1A, IGF1R, BMX, LCK, ALK, PIM2, EPHB2, RPS6KA2, MAPK4, CSNK1G2, ERBB4, PAK3, INSRR, BTK, CAMKK1, EPHA6, MAPK11, EPHA5, TGFBR2, EPHA8, FYN, FES, RET, ABL2, STK3	86	PHASE2: Breast Neoplasms, Carcinoma, Non-Small-Cell Lung, Lung Neoplasms, Neoplasms
pelitinib	MAPK10, RPS6KA3, MAP4K4, BMPR1A, MARK3, IRAK3, NTRK1, NTRK3, PRKAA2, KDR, FLT4, INSR, LTK, EGFR, PRKCH, ACVR2A, PRKG1, ERBB2, CAMK2B, PRKD1, PKMYT1, PIP5K1A, IGF1R, BMX, LCK, ALK, PIM2, EPHB2, RPS6KA2, MAPK4, CSNK1G2, ERBB4, PAK3, INSRR, BTK, CAMKK1, EPHA6, MAPK11, EPHA5, TGFBR2, EPHA8, FYN, FES, RET, ABL2, STK3	86	PHASE2: Carcinoma, Non-Small-Cell Lung, Colonic Neoplasms, Colorectal Neoplasms, Neoplasms, Rectal Neoplasms
Erlotinib	MAPK10, RPS6KA3, MAP4K4, BMPR1A, MARK3, IRAK3, NTRK1, NTRK3, PRKAA2, KDR, FLT4, INSR, LTK, EGFR, PRKCH, ACVR2A, PRKG1, ERBB2, CAMK2B, PRKD1, PKMYT1, PIP5K1A, IGF1R, BMX, LCK, ALK, PIM2, EPHB2, RPS6KA2, MAPK4, CSNK1G2, ERBB4, PAK3, INSRR, BTK, CAMKK1, EPHA6, MAPK11, EPHA5, TGFBR2, EPHA8, FYN, FES, RET, ABL2, STK3	86	NA: Carcinoma, Non-Small-Cell Lung, Carcinoma, Squamous Cell, Head and Neck Neoplasms, Lung Neoplasms, Neoplasms

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

Table 10. 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	GDNF, BDNF, NTF3, ITGB3	96	0.59
Quinapril	GDNF, BDNF, NTF3, ITGB3	94	0.45
Progesterone	WWOX, NR1H4, GDNF, BDNF, NTF3, G6PD, PIP5K1A, AR	91	0.19
Hesperetin	MAPK10, GDNF, BDNF, CYP1B1, MAPK4, MAPK11, NTF3, PIP5K1A	90	0.18
(2s,3s)-Trans-Dihydroquercetin	MAPK10, GDNF, BDNF, CYP1B1, MAPK4, MAPK11, NTF3, PIP5K1A	90	0.16

Table 11. 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, DUSP22, MAPK4, CDC14B, MAPK11, PLEC, STAT1, DUSP14, DUPD1, DUSP9, PIP5K1A, BRCA1	91	0.35
2,5,7-Trihydroxynaphthoquinone	MAPK10, DUSP22, MAPK4, CDC14B, MAPK11, STAT1, DUSP14, DUPD1, DUSP9, PIP5K1A, BRCA1	91	0.35
Estrone	MAPK10, WWOX, GDNF, BDNF, MAPK4, MAPK11, NTF3, AR, PIP5K1A	90	0.23
Naringenin	MAPK10, GDNF, BDNF, CYP1B1, MAPK4, MAPK11, NTF3, PIP5K1A	90	0.18
(2S)-5-hydroxy-2-(4-hydroxyphenyl)-7-methoxy-2,3-dihydro-4H-chromen-4-one	MAPK10, GDNF, BDNF, CYP1B1, MAPK4, MAPK11, NTF3, PIP5K1A	89	0.18

As the result of drug search we propose the following drugs as most promising candidates for treating the pathology under study: Imatinib, Flavopiridol, Moexipril and 3-METHYL-1,6,8-TRIHYDROXYANTHRAQUINONE. These drugs were selected for acting on the following targets: RPS6KA2, ITGB3 and DUSP9, 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 highly methylated 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, Flavopiridol, Moexipril and 3-METHYL-1,6,8-TRIHYDROXYANTHRAQUINONE

These drugs were selected for acting on the following targets: RPS6KA2, ITGB3 and DUSP9, 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:

RSK and MKP-4

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, Flavopiridol, Darodipine, 3-METHYL-1,6,8-TRIHYDROXYANTHRAQUINONE and 3-Methyladenine. 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 highly methylated 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.

- RSK
- MKP-4

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 highly methylated genes were analyzed using known DNA-binding motifs described in the TRANSFAC® library, release 2024.2 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transfac).
The master regulator search uses the TRANSPATH® database (BIOBASE), release 2024.2 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transpath). A comprehensive signal transduction network of human cells is built by the software on the basis of reactions annotated in TRANSPATH®.
The information about drugs corresponding to identified drug targets and clinical trials references were extracted from HumanPSD™ database, release 2024.2 (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 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 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 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 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. 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
10. Boyarskikh U, Pintus S, Mandrik N, Stelmashenko D, Kiselev I, Evshin I, Sharipov R, Stegmaier P, Kolpakov F, Filipenko M, Kel A. Computational master-regulator search reveals mTOR and PI3K pathways responsible for low sensitivity of NCI-H292 and A427 lung cancer cell lines to cytotoxic action of p53 activator Nutlin-3. *BMC Med Genomics.* **2018**;11(1):12. doi:10.1186/1471-2105-7-s2-s13
11. 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.
12. Filimonov DA, Poroikov VV. Prognosis of specters of biological activity of organic molecules. *Russian chemical journal.* **2006**;50(2):66-75 (russ)
13. 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 - Detailed report. Composite modules and master regulators (highly methylated genes in Hypertension vs. Control).
2. Supplementary table 2 - 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.