Sequence and Pathway analysis

IL2RB and PTPN22 are promising druggable targets for treating diabetes mellitus that control activity of CEBPB, CEBPA and CEBPD transcription factor on promoters of genes carrying sequence variations

Demo User
geneXplain GmbH
info@genexplain.com

Data received on 19/02/2020 ; Run on 19/02/2020 ; Report generated on 19/02/2020

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

Abstract

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

From the data set analyzed in this study, we found the following TFs to be potentially involved in the regulation of the genes carrying sequence variations: CEBPB, CEBPA and CEBPD. The subsequent network analysis suggested IL2RB, FGG, TYK2, PTPN22 and ERBB3 as the most promising and druggable molecular targets. Finally, the following drugs were identified as the most promising treatment candidates: Sucralfate, Aldesleukin, 3-{(3R,4R)-4-methyl-3-[methyl(7H-pyrrolo[2,3-d]pyrimidin-4-yl)amino]piperidin-1-yl}-3-oxopropanenitrile, Enprofylline, N6-(2,5-Dimethoxy-Benzyl)-N6-Methyl-Pyrido[2,3-D]Pyrimidine-2,4,6-Triamine and Lipoic Acid.

1. Introduction

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

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

The "upstream analysis" approach has now been extended by a third step that reveals known drugs suitable to inhibit (or activate) the identified molecular targets in the context of the disease under study. This step is performed by using information from HumanPSD™ database [5]. In addition, new potential small molecular ligands are subsequently predicted for the revealed targets. A general druggability check is performed using a precomputed database of biologcal activities of chemical compounds from a library of about 13000 pharmaceutically most active compounds. The spectra of biological activities are computed using the program PASS on the basis of a (Q)SAR approach [11-13].

2. Data

For this study the following experimental data was used:

Table 1. Experimental datasets used in the study

File name	Data type
E04_Genomics_SNP_diabetes	Genomics

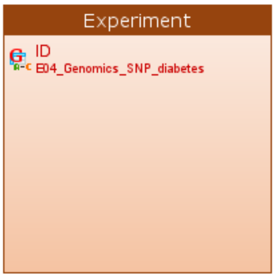


Figure 1. Annotation diagram of experimental data used in this study. With the colored boxes we show those sub-categories of the data that are compared in our analysis.

3. Results

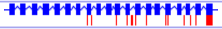
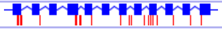
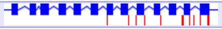
We have analysed the following condition: Experiment.

3.1. Identification of target genes

In the first step of the analysis ***target genes*** were identified from the uploaded experimental data. 300 genes with maximal number of SNPs were used as target genes.

Table 2. Top ten genes carrying SNP variations in Experiment.

See full table →

ID	Gene description	Gene symbol	Gene schematic representation	Number of variations
ENSG00000196735	major histocompatibility complex, class II, DQ alpha 1	HLA-DQA1		82
ENSG00000130164	low density lipoprotein receptor	LDLR		31
ENSG00000165029	ATP binding cassette subfamily A member 1	ABCA1		30
ENSG00000169174	proprotein convertase subtilisin/kexin type 9	PCSK9		22
ENSG00000175445	lipoprotein lipase	LPL		18
ENSG00000084674	apolipoprotein B	APOB		16
ENSG00000161888	SPC24, NDC80 kinetochore complex component	SPC24		16
ENSG00000196301	major histocompatibility complex, class II, DR beta 9 (pseudogene)	HLA-DRB9		15
ENSG00000128918	aldehyde dehydrogenase 1 family member A2	ALDH1A2		14
ENSG00000166035	lipase C, hepatic type	LIPC		12

3.2. Functional classification of genes

A functional analysis of genes carrying sequence variations 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.

Genes carrying SNP variations in Experiment:

300 top carrying SNP variation genes were taken for the mapping.

GO (biological process)

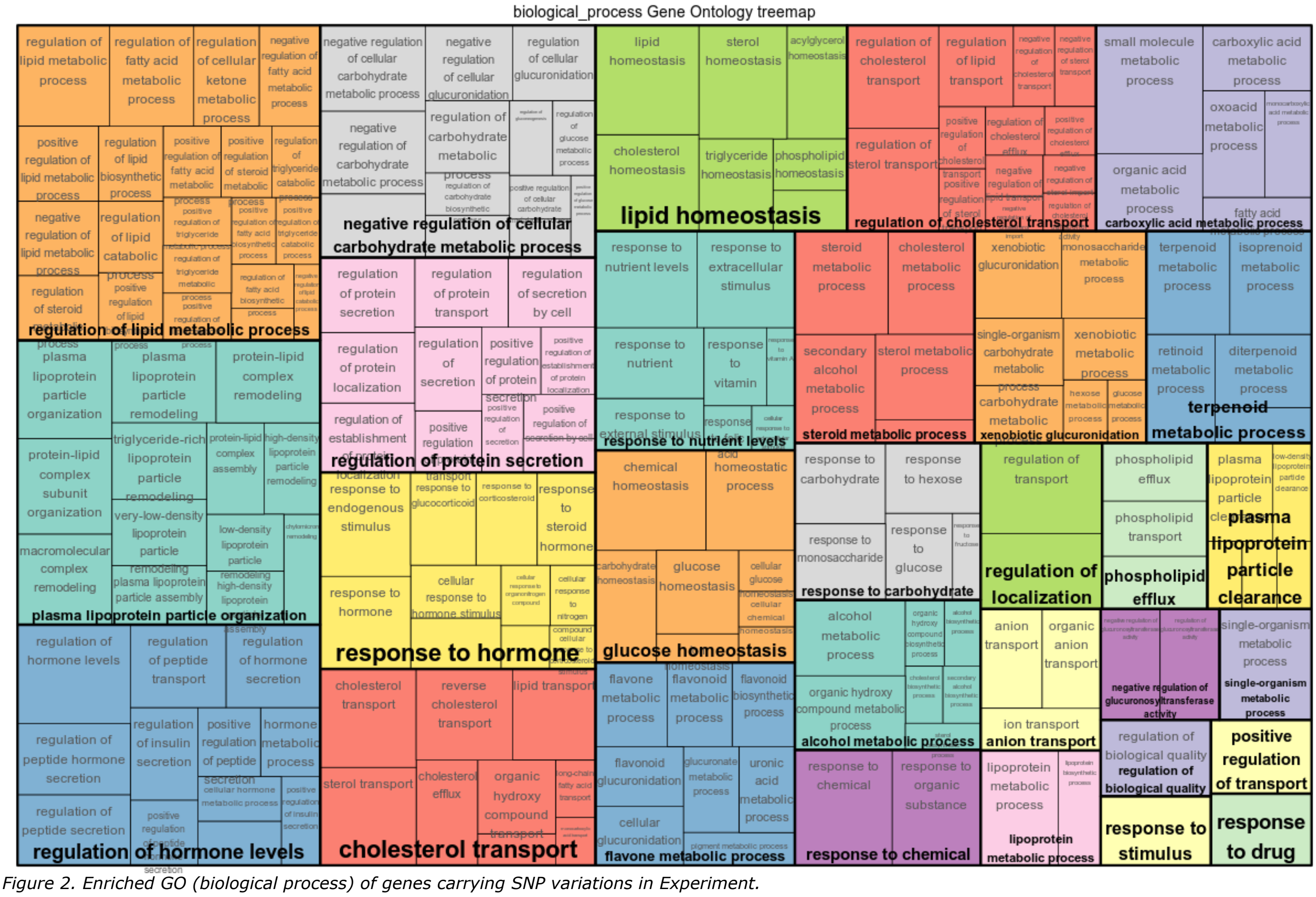


Figure 2. Enriched GO (biological process) of genes carrying SNP variations in Experiment.

Full classification →

TRANSPATH® Pathways (2020.1)

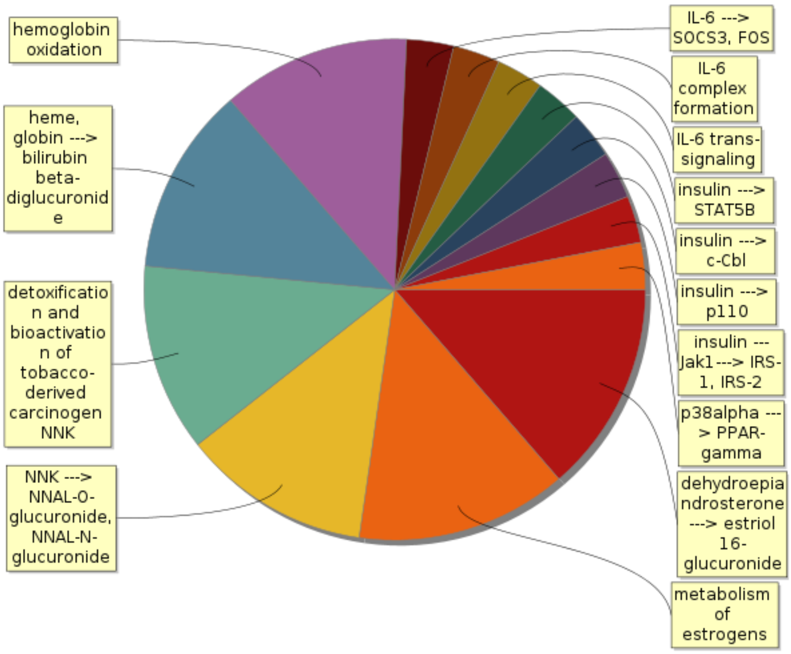


Figure 3. Enriched TRANSPATH® Pathways (2020.1) of genes carrying SNP variations in Experiment.

Full classification →

HumanPSD(TM) disease (2020.1)

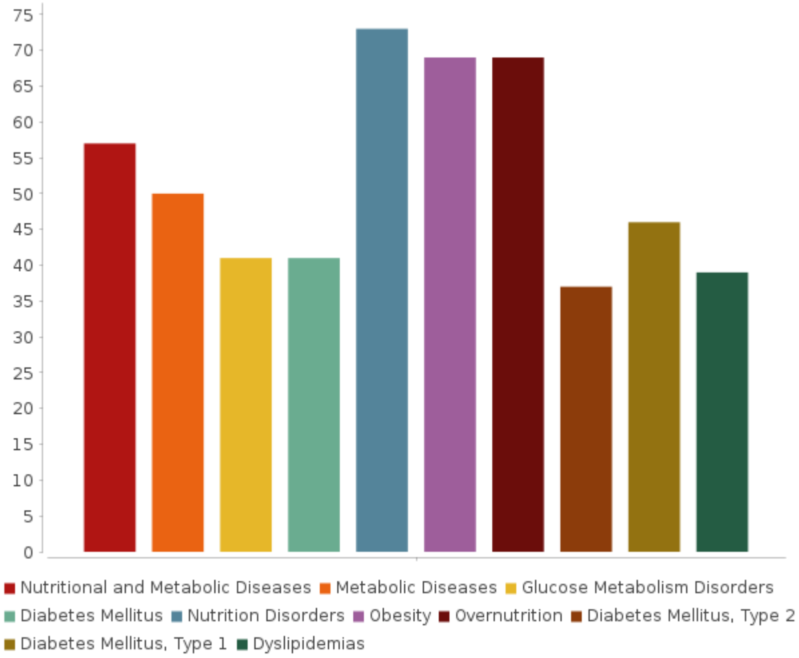


Figure 4. Enriched HumanPSD(TM) disease (2020.1) of genes carrying SNP variations in Experiment. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

Full classification →

3.3. Analysis of enriched transcription factor binding sites and composite modules

In the next step a search for transcription factors binding sites (TFBS) was performed in the regulatory regions of the ***target genes*** by using the TF binding motif library of the TRANSFAC® database. We searched for so called **composite-modules** that act as potential condition-specific **enhancers** of the ***target genes*** in their upstream regulatory regions (-1000 bp upstream of transcription start site (TSS)) and identify transcription factors regulating activity of the genes through such **enhancers**.
Classically, **enhancers** are defined as regions in the genome that increase transcription of one or several genes when inserted in either orientation at various distances upstream or downstream of the gene [8]. Enhancers typically have a length of several hundreds of nucleotides and are bound by multiple transcription factors in a cooperative manner [9].

We analysed mutations that were revealed in the potential enhancers located upstream, downstream or inside the ***target genes*** (see Table 3). We identified 1075 mutations potentially affecting gene regulation. Table 4 shows the following lists of PWMs whose sites were lost or gained due to these mutations. These PWMs were put in focus of the CMA algorithm that constructs the model of the enhancers by specifying combinations of TF motifs (see more details of the algorithm in the Method section).

Table 3. Mutations revealed in genes in genes carrying SNP variations
See full table →

ID	Gene symbol	Gene schematic representation	Number of variations
ENSG00000196735	HLA-DQA1		82
ENSG00000130164	LDLR		31
ENSG00000165029	ABCA1		30
ENSG00000169174	PCSK9		22
ENSG00000175445	LPL		18
ENSG00000084674	APOB		16
ENSG00000161888	SPC24		16
ENSG00000196301	HLA-DRB9		15
ENSG00000128918	ALDH1A2		14
ENSG00000166035	LIPC		12

Table 4. PWMs whose sites were lost or gained due to mutations in genes carrying SNP variations
See full table →

ID	P-value (gains)	P-value (losses)	yesCount (gains)	yesCount (losses)
V$TTF1_Q5_01	4.64E-2	8.25E-3	16	25
V$BCL6_Q3_01	4.03E-2	4.18E-3	2	80
V$PIT1_Q6_01	1.21E-2	3.46E-2	23	161
V$HOMEZ_01	1E-2	3.36E-2	24	42
V$ERALPHA_01	8.61E-3	4.52E-3	35	48
V$CREBP1_01	6.9E-3		80	null
V$ZFP206_01	5.32E-3	2.4E-2	0	110
V$RELA_Q6	4.82E-3	7.86E-3	56	75
V$HNF4A_Q3	4.45E-3		290	null
V$RORALPHA_Q4	3.83E-3	2.84E-2	29	64
V$SF1_Q5_01	3.16E-3		30	null
V$REVERBALPHA_Q6	8.25E-4	2.06E-2	37	60
V$EGR1_Q6		6.88E-4	null	94
V$LEF1_Q5_01		9.38E-4	null	48
V$POU6F1_02		8.57E-3	null	140
V$TEF1_Q6_04		7.19E-3	null	61
V$ZFP105_04		4.85E-3	null	47
V$ZIC1_05		1.38E-3	null	266

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 (genes carrying SNP variations in Experiment).

To build the most specific composite modules we choose top carrying SNP variation genes as the input of CMA algorithm. The obtained CMA model is then applied to compute CMA score for all genes carrying SNP variations.

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$POU6F1_02	V$MZF1_Q5	V$HNF1A_Q4	V$SNA_Q6	V$HNF1A_01
0.00; N=3	0.00; N=1	0.93; N=2	0.00; N=2	0.86; N=2

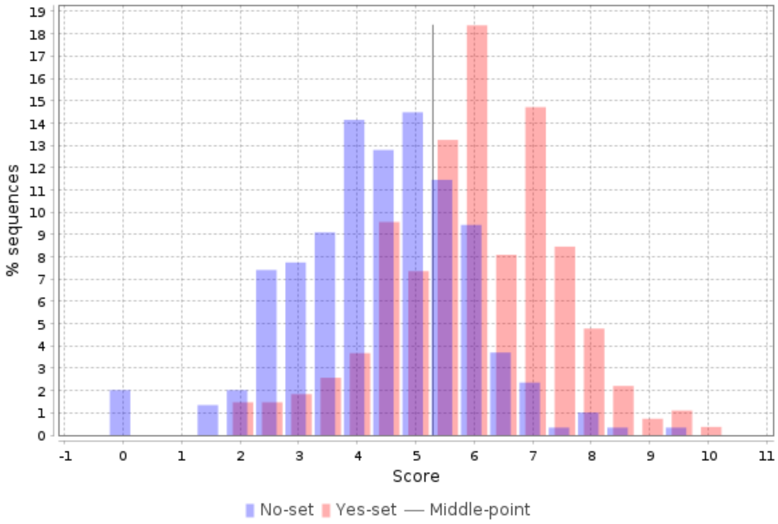
Module width: 129

Module 2:

V$RARG_Q3	V$CEBP_Q2_01	V$GCM_Q4	V$EVI1_04	V$TRIM28_02
0.89; N=2	0.98; N=2	0.00; N=3	0.77; N=2	0.89; N=3

Module width: 144

Model score (-p*log10(pval)): 14.98
Wilcoxon p-value (pval): 1.29e-30
Penalty (p): 0.501
Average yes-set score: 5.97
Average no-set score: 4.43
AUC: 0.78
Middle-point: 5.30
False-positive: 26.94%
False-negative: 27.94%



See model visualization table →

Table 5. List of top ten genes carrying SNP variations in Experiment 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
ENSG00000243135	UGT1A3	UDP glucuronosyltransferase family 1 member A3	11.61	Evi-1(h), POU6F1(h), SNA(h), RNF96(h), GCMa(h),GCMb(h), RAR-gamma(h), MZF-1(h)...
ENSG00000257138	TAS2R38	taste 2 receptor member 38	11.03	RNF96(h), GCMa(h),GCMb(h), RAR-gamma(h), Evi-1(h), C/EBPalpha(h),C/EBPbeta(h),C/EBPdelta(h),C/EBPepsilon(h),C/EBPgamma(h), HNF-1alpha(h), POU6F1(h)...
ENSG00000145321	GC	GC, vitamin D binding protein	10.94	SNA(h), GCMa(h),GCMb(h), POU6F1(h), HNF-1alpha(h), MZF-1(h), C/EBPalpha(h),C/EBPbeta(h),C/EBPdelta(h),C/EBPepsilon(h),C/EBPgamma(h), Evi-1(h)...
ENSG00000180525	PRR26	proline rich 26	10.92	GCMa(h),GCMb(h), RAR-gamma(h), RNF96(h), C/EBPalpha(h),C/EBPbeta(h),C/EBPdelta(h),C/EBPepsilon(h),C/EBPgamma(h), HNF-1alpha(h), POU6F1(h), Evi-1(h)...
ENSG00000152253	SPC25	SPC25, NDC80 kinetochore complex component	10.88	C/EBPalpha(h),C/EBPbeta(h),C/EBPdelta(h),C/EBPepsilon(h),C/EBPgamma(h), Evi-1(h), RAR-gamma(h), GCMa(h),GCMb(h), SNA(h), HNF-1alpha(h), POU6F1(h)
ENSG00000255518	RP11-148O21.4		10.69	C/EBPalpha(h),C/EBPbeta(h),C/EBPdelta(h),C/EBPepsilon(h),C/EBPgamma(h), RAR-gamma(h), Evi-1(h), RNF96(h), GCMa(h),GCMb(h), MZF-1(h), SNA(h)...
ENSG00000240224	UGT1A5	UDP glucuronosyltransferase family 1 member A5	10.63	Evi-1(h), C/EBPalpha(h),C/EBPbeta(h),C/EBPdelta(h),C/EBPepsilon(h),C/EBPgamma(h), GCMa(h),GCMb(h), SNA(h), RNF96(h), POU6F1(h), RAR-gamma(h)...
ENSG00000157017	GHRL	ghrelin and obestatin prepropeptide	10.61	MZF-1(h), RAR-gamma(h), POU6F1(h), SNA(h), C/EBPalpha(h),C/EBPbeta(h),C/EBPdelta(h),C/EBPepsilon(h),C/EBPgamma(h), RNF96(h), GCMa(h),GCMb(h)...
ENSG00000254275	LINC00824	long intergenic non-protein coding RNA 824	10.53	POU6F1(h), HNF-1alpha(h), SNA(h), RNF96(h), MZF-1(h), Evi-1(h), RAR-gamma(h)...
ENSG00000198099	ADH4	alcohol dehydrogenase 4 (class II), pi polypeptide	10.46	MZF-1(h), HNF-1alpha(h), RNF96(h), GCMa(h),GCMb(h), POU6F1(h), SNA(h), C/EBPalpha(h),C/EBPbeta(h),C/EBPdelta(h),C/EBPepsilon(h),C/EBPgamma(h)...

On the basis of the enhancer models we identified the following transcription factors potentially regulating the ***target genes*** of our interest. We found 14 transcription factors controlling expression of the genes associated with genomic variations (see Table 6).

Table 6. Transcription factors of the predicted enhancer model potentially regulating the genes carrying sequence variations (genes carrying SNP variations in Experiment). ***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 TFin 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
MO000019381	CEBPB	CCAAT/enhancer binding protein beta	3.28	1.39
MO000019418	CEBPA	CCAAT/enhancer binding protein alpha	3.05	2.58
MO000002641	CEBPD	CCAAT/enhancer binding protein delta	3.01	1.27
MO000044348	SNAI1	snail family transcriptional repressor 1	2.59	1.31
MO000069886	TRIM28	tripartite motif containing 28	2.32	1.38
MO000033253	MECOM	MDS1 and EVI1 complex locus	2.29	1.54
MO000028320	POU6F1	POU class 6 homeobox 1	2.12	1.93
MO000026306	GCM1	glial cells missing homolog 1	1.92	1.28
MO000028673	CEBPE	CCAAT/enhancer binding protein epsilon	1.75	1.15
MO000082618	HNF1A	HNF1 homeobox A	1.61	3.19

3.4. Finding master regulators in networks

In the second step of the upstream analysis common regulators of the revealed TFs were identified. We identified 3 signaling proteins whose structure and function is highly damaged by the mutations (see Table 7).

Table 7. Signaling proteins whose structure and function is damaged by the mutations in genes carrying SNP variations

See full table →

ID	Title	Mutation count	Consequence	Codons
MO000104653	alcohol dehydrogenase 1C(h)	3	stop_gained	Gga/Tga
MO000036095	LpL(h)	1	stop_gained	tCa/tGa
MO000078586	LpL-p20(h)	1	stop_gained	tCa/tGa

Top 3 mutated proteins for genes carrying SNP variations were used in the algorithm of master regulator search as a list of nodes of the signal transduction network that are removed from the network during the search of master regulators (see more details in of the algorithm in the Method section). These master regulators appear to be the key candidates for therapeutic targets as they have a master effect on regulation of intracellular pathways that activate the pathological process of our study. The identified master regulators are shown in Table 8.

Table 8. Master regulators that may govern the regulation of genes carrying SNP variations in Experiment. ***Total rank*** *is the sum of the ranks of the master molecules sorted by keynode score, CMA score, genomics data.*

See full table →

ID	Master molecule name	Gene symbol	Gene description	Total rank
MO000039099	IL-1beta-p17:IL-1RI:IL-1RAcP:MyD88:tollip:IRAK-1{pS376}{pT387}:IRAK-4:IRAK-2	IL1B, IL1R1, IL1RAP, IRAK1, IRAK2, IRAK4, MYD88, TOLLIP	interleukin 1 beta, interleukin 1 receptor accessory protein, interleukin 1 receptor associated kina...	15
MO000009410	MKK5(h)	MAP2K5	mitogen-activated protein kinase kinase 5	16
MO000058229	MEK(h){p}	MAP2K1, MAP2K2, MAP2K5	mitogen-activated protein kinase kinase 1, mitogen-activated protein kinase kinase 2, mitogen-activa...	22
MO000078407	MKK5-isoform2(h)	MAP2K5	mitogen-activated protein kinase kinase 5	24
MO000130227	PEP(h)	PTPN22	protein tyrosine phosphatase, non-receptor type 22	24
MO000007346	IL-2Rbeta(h)	IL2RB	interleukin 2 receptor subunit beta	25
MO000121450	MKK5-isoform1(h)	MAP2K5	mitogen-activated protein kinase kinase 5	25
MO000014070	Tyk2(h)	TYK2	tyrosine kinase 2	26
MO000038143	Fibrinogen(h)	FGA, FGB, FGG	fibrinogen alpha chain, fibrinogen beta chain, fibrinogen gamma chain	26
MO000281381	(angiotensin II)2:(AT2 receptor)2:(ATIP-isoform3)2:SHP-1	AGT, AGTR2, MTUS1, PTPN6	angiotensin II receptor type 2, angiotensinogen, microtubule associated tumor suppressor 1, protein ...	29

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 genes carrying sequence variations, and selected master regulators, which are responsible for the regulation of these TFs.

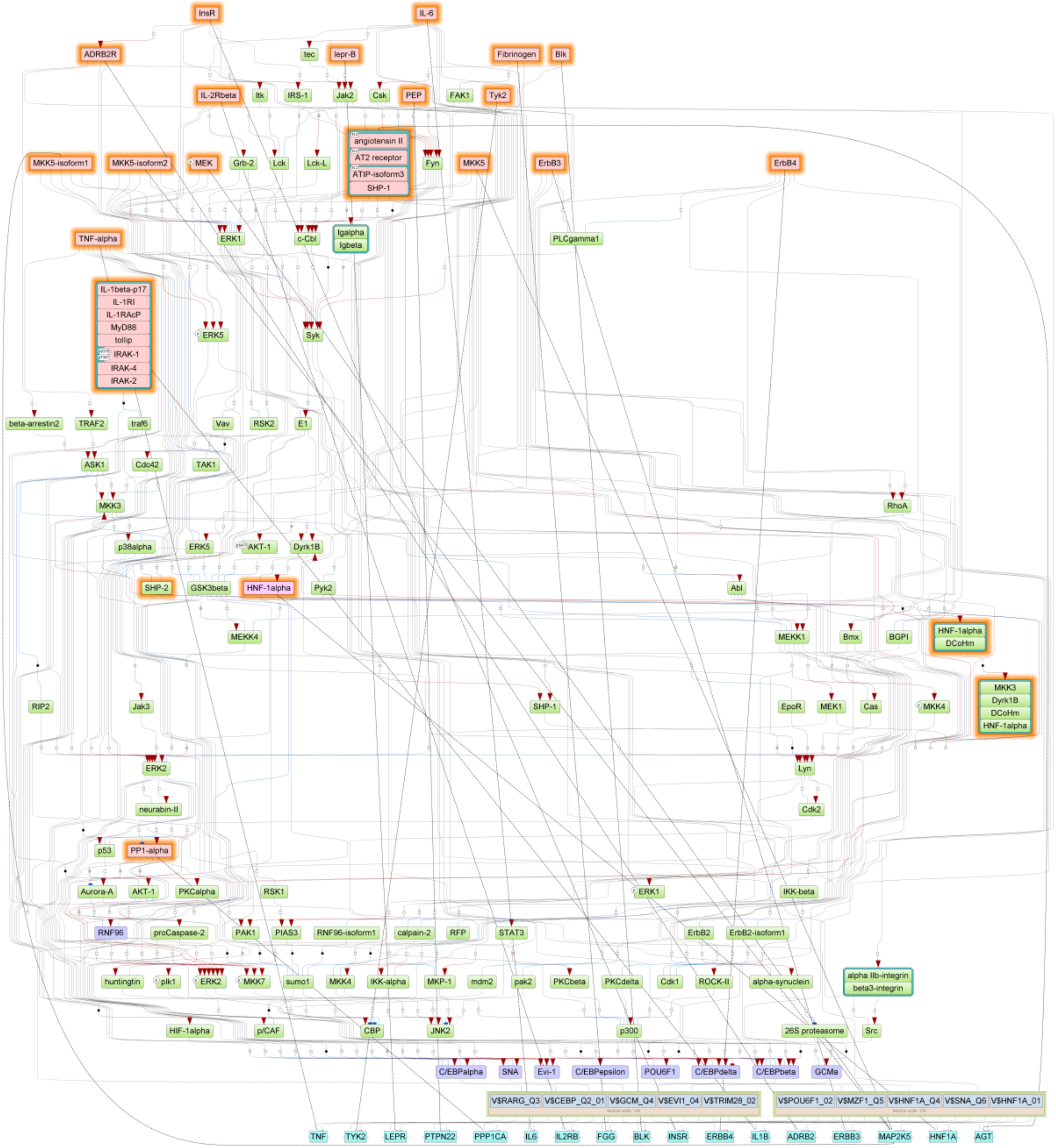


Figure 5. Diagram of intracellular regulatory signal transduction pathways of genes carrying SNP variations in Experiment. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate molecules, which have been added to the network during the search for master regulators from selected TFs. Orange frames highlight molecules presented in original mapping.

See full diagram →

4. Identification of potential drugs

In the last step of the analysis we strived to identify known drugs as well as new potentially active chemical compounds that are potentially suitable for inhibition (or activation) of the identified molecular targets in the context of specified human disease.
First, we identify known drugs using information from HumanPSD™ database [5] about their targets and about clinical trials where the drugs have been tested for the treatment of various human diseases. Table 9 shows the resulting list of druggable master regulators that represent the predicted drug targets of the studied pathology. Table 10 lists chemical compounds and known drugs (from the HumanPSD™ database) potentially acting on corresponding master regulators.

*Table 9. Known drug targets for known drugs revealed in this study.The column **Druggability score** contains the number of drugs that are potentially suitable for inhibition (or activation) of the target. **Total rank** is the sum of the ranks of the master molecules sorted by keynode score, CMA score, genomics data.*

See full table →

ID	Gene symbol	Gene description	Druggability score	Total rank
ENSG00000100385	IL2RB	interleukin 2 receptor subunit beta	4	25
ENSG00000171557	FGG	fibrinogen gamma chain	1	26
ENSG00000105397	TYK2	tyrosine kinase 2	2	36
ENSG00000125538	IL1B	interleukin 1 beta	13	40
ENSG00000179295	PTPN11	protein tyrosine phosphatase, non-receptor type 11	1	42
ENSG00000136244	IL6	interleukin 6	8	47
ENSG00000169252	ADRB2	adrenoceptor beta 2	57	53
ENSG00000232810	TNF	tumor necrosis factor	30	53
ENSG00000135100	HNF1A	HNF1 homeobox A	1	57
ENSG00000171105	INSR	insulin receptor	14	66

*Table 10. The list of drugs (from Human PSD) approved or used in clinical trials for the application in diabetes mellitus and acting on master regulators revealed in our study. The column **Target activity score** contains the value of numeric function that depends on ranks of all targets that were found for the drug. The column **Disease activity score** contains the weighted sum of user selected diseases where the drug is known to be applied. We use sum of clinical trials phases as the weight of the disease. **Drug rank** column contains total rank of given drug among all found. See Methods section for details.*

See full table →

ID	Name	Target names	Target activity score	NA	Phase 1	Phase 2	Phase 3	Phase 4	Disease activity score	Drug rank
DB06372	Rilonacept	IL1B	0.31	Bursitis, Cardiovascular Diseases, Gout, Kidney Diseases, Renal Insufficiency, Chronic, Urticaria, Vascular Diseases	Diabetes Mellitus, Arthritis, Arthritis, Juvenile, Diabetes Mellitus, Type 1, Hearing Loss, ST Elevation Myocardial Infarction, Scleroderma, Diffuse...	Anemia, Atherosclerosis, Coronary Artery Disease, Cryopyrin-Associated Periodic Syndromes, Familial Mediterranean Fever, Hepatitis, Hepatitis, Alcoholic...	Cryopyrin-Associated Periodic Syndromes, Genetic Diseases, Inborn, Gout, Urticaria	Renal Insufficiency, Renal Insufficiency, Chronic	1	33
DB00612	Bisoprolol	ADRB2	0.15	Aortic Valve Stenosis, Atrial Fibrillation, Bone Diseases, Metabolic, Constriction, Pathologic, Cysts, Heart Diseases, Heart Failure...	Aneurysm, Aortic Aneurysm, Aortic Aneurysm, Abdominal, Bites and Stings, Breast Neoplasms, Coronary Artery Disease, Familial Primary Pulmonary Hypertension...	Diabetes Mellitus, Breast Neoplasms, Coronary Artery Disease, Familial Primary Pulmonary Hypertension, Heart Failure, Hypertension, Lung Diseases...	Diabetes Mellitus, Breast Neoplasms, Cardiomyopathies, Chagas Cardiomyopathy, Coronary Artery Disease, Heart Failure, Hypertension...	Diabetes Mellitus, Aortic Valve Stenosis, Atherosclerosis, Atrial Fibrillation, Constriction, Pathologic, Coronary Artery Disease, Diabetes Mellitus, Type 1...	9	55
DB00264	Metoprolol	ADRB2	0.15	Atrial Fibrillation, Atrophy, Autonomic Nervous System Diseases, Brain Abscess, Cerebral Hemorrhage, Character, Coronary Disease...	Diabetes Mellitus, Angina Pectoris, Atrial Fibrillation, Bites and Stings, Dermatitis, Dermatitis, Atopic, Diabetes Mellitus, Type 2...	Diabetes Mellitus, Acute Coronary Syndrome, Aortic Valve Stenosis, Apnea, Arrhythmias, Cardiac, Atrial Fibrillation, Brain Abscess...	Albuminuria, Apnea, Arrhythmias, Cardiac, Atrial Fibrillation, Breast Neoplasms, Cardiomyopathies, Cardiovascular Diseases...	Diabetes Mellitus, Acute Coronary Syndrome, Angina Pectoris, Aortic Valve Insufficiency, Arrhythmias, Cardiac, Arthritis, Asthma...	7	56
DB00335	Atenolol	ADRB2	0.15	Atrial Fibrillation, Cachexia, Cardiovascular Diseases, Coronary Disease, Heart Diseases, Hypertension, Marfan Syndrome...	Eclampsia, Hypertension, Obesity, Obesity, Morbid, Orthostatic Intolerance, Postural Orthostatic Tachycardia Syndrome, Pre-Eclampsia...	Angina Pectoris, Hypertension, Marfan Syndrome, Orthostatic Intolerance, Postural Orthostatic Tachycardia Syndrome, Stress Disorders, Post-Traumatic, Tachycardia...	Cardiovascular Diseases, Cerebral Small Vessel Diseases, Cerebrovascular Disorders, Constriction, Pathologic, Coronary Disease, Heart Diseases, Hemangioma...	Diabetes Mellitus, Aneurysm, Angina Pectoris, Aortic Aneurysm, Aortic Aneurysm, Abdominal, Atherosclerosis, Cognitive Dysfunction...	4	58
DB04861	Nebivolol	ADRB2	0.15	Atherosclerosis, Atrophy, Autonomic Nervous System Diseases, Brain Abscess, Glaucoma, Heart Failure, Heart Failure, Diastolic...	Heart Diseases, Hypertension	Hypertension	Cardiomyopathies, Heart Failure, Hypertension, Marfan Syndrome, Muscular Diseases, Muscular Dystrophies, Muscular Dystrophy, Duchenne...	Diabetes Mellitus, Apnea, Atherosclerosis, Coronary Artery Disease, Diabetes Mellitus, Type 2, Erectile Dysfunction, Heart Failure...	4	58

Table 11. The list of drugs (from HumanPSD) known to be acting on master regulators revealed in our study that can be proposed as a drug repurposing initiative for the treatment of diabetes mellitus. ***Target activity score*** *column contains value of numeric function that depends on ranks of all targets that were found for the drug.* ***Drug rank*** *column contains total rank of given drug among all found. See Methods section for details.*

ID	Name	Target names	Target activity score	NA	Phase 1	Phase 2	Phase 3	Phase 4	Drug rank
DB00364	Sucralfate	FGG	0.23	Dyspepsia, Gastroesophageal Reflux	Eosinophilic Esophagitis, Esophagitis	Hand, Foot and Mouth Disease, Herpangina, Mouth Diseases, Stomatitis	Head and Neck Neoplasms, Mucositis, Neoplasms	Gastritis, Pneumonia, Pneumonia, Ventilator-Associated, Proctitis, Ulcer	47
DB00065	Infliximab	TNF	0.19	Arteritis, Arthritis, Arthritis, Psoriatic, Arthritis, Rheumatoid, Churg-Strauss Syndrome, Colitis, Colitis, Ulcerative...	Aneurysm, Berylliosis, Blindness, Colitis, Colitis, Ulcerative, Corneal Diseases, Depression...	Aneurysm, Arteritis, Arthritis, Arthritis, Juvenile, Arthritis, Reactive, Arthritis, Rheumatoid, Berylliosis...	Arthritis, Arthritis, Juvenile, Arthritis, Psoriatic, Arthritis, Rheumatoid, Behcet Syndrome, Colitis, Colitis, Ulcerative...	Arthritis, Arthritis, Psoriatic, Arthritis, Rheumatoid, Colitis, Colitis, Ulcerative, Crohn Disease, Depression...	48
DB00867	Ritodrine	ADRB2	0.19	Obstetric Labor, Premature				Obstetric Labor, Premature, Premature Birth	48
DB00871	Terbutaline	ADRB2	0.19	Asthma, Diabetes Mellitus, Type 1, Fatigue, Fetal Distress, Heart Failure, Hypertrophy	Diabetes Mellitus, Type 1	Asthma, Asthma, Exercise-Induced, Neuralgia	Asthma, Lung Diseases, Lung Diseases, Obstructive, Pulmonary Disease, Chronic Obstructive	Asthma, Heart Failure, Status Asthmaticus	48
DB00938	Salmeterol	ADRB2	0.19	Acute Lung Injury, Airway Obstruction, Asthma, Bronchiectasis, Bronchitis, Bronchitis, Chronic, Cough...	Asthma, Lung Diseases, Lung Diseases, Obstructive, Malaria, Malaria, Falciparum, Pulmonary Disease, Chronic Obstructive, Spinal Cord Injuries...	Asthma, Lung Diseases, Lung Diseases, Obstructive, Pulmonary Disease, Chronic Obstructive	Asthma, Asthma, Exercise-Induced, Bronchitis, Bronchitis, Chronic, Emphysema, Lung Diseases, Lung Diseases, Obstructive...	Asthma, Asthma, Exercise-Induced, Bronchial Spasm, Bronchitis, Bronchitis, Chronic, Candidiasis, Candidiasis, Oral...	48

Next, new potential small molecular ligands were predicted for the revealed targets and a general druggability check was run using a pre-computed database of spectra of biological activities of chemical compounds from a library of 13040 most pharmaceutically active known compounds. The spectra of biological activities has been computed using the program PASS [11-13] on the basis of a (Q)SAR approach. Table 12 shows the resulting list of druggable master regulators, which represent the predicted drug targets of the studied pathology. Table 13 lists chemical compounds and known drugs potentially acting on the corresponding master regulators.

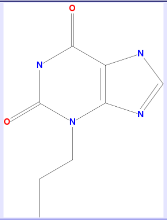
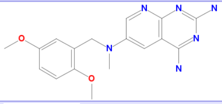
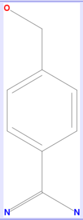
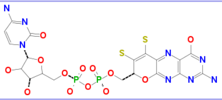
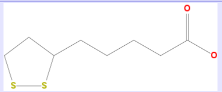
Table 12. Extended list of drug targets revealed in this study (targets that are predicted by PASS program potentially targeted by an extended list of known drugs and pharmaceutically active chemical compounds). The column ***Druggability score*** *contains a numeric value which indicates how suitable this target is to be inhibited (or activated) by a drug. See Methods section for details.*

See full table →

ID	Name	Gene symbol	Gene description	Druggability score	Total rank
ENSG00000134242	PTPN22	PTPN22	protein tyrosine phosphatase, non-receptor type 22	0.93	24
ENSG00000100385	IL2RB	IL2RB	interleukin 2 receptor subunit beta	4.91	25
ENSG00000065361	ERBB3	ERBB3	erb-b2 receptor tyrosine kinase 3	19.19	31
ENSG00000105397	TYK2	TYK2	tyrosine kinase 2	4.02	36
ENSG00000125538	IL1B	IL1B	interleukin 1 beta	41.17	40
ENSG00000136244	IL6	IL6	interleukin 6	41.9	47
ENSG00000136573	BLK	BLK	BLK proto-oncogene, Src family tyrosine kinase	1.52	53
ENSG00000232810	TNF	TNF	tumor necrosis factor	1.6	53
ENSG00000135100	HNF1A	HNF1A	HNF1 homeobox A	0.87	57
ENSG00000178568	ERBB4	ERBB4	erb-b2 receptor tyrosine kinase 4	19.19	57

Table 13. The chemical compounds and known drugs identified by the PASS program as potentially acting on master regulators revealed in our study. Based on the revealed mechanism of action these compounds can be proposed for the treatment of diabetes mellitus in the current pathological case. ***Disease activity score*** *column contains maximal value of probability to be active for all activities corresponding to the selected diseases for the given compound or 0 if no diseases were selected (in this case column will be hidden).* ***Target activity score*** *column contains value of numeric function which depends on all activity-mechanisms correspondent to the drug.* ***Drug rank*** *column contains total rank of given drug among all found. See Methods section for details.*

See full table →

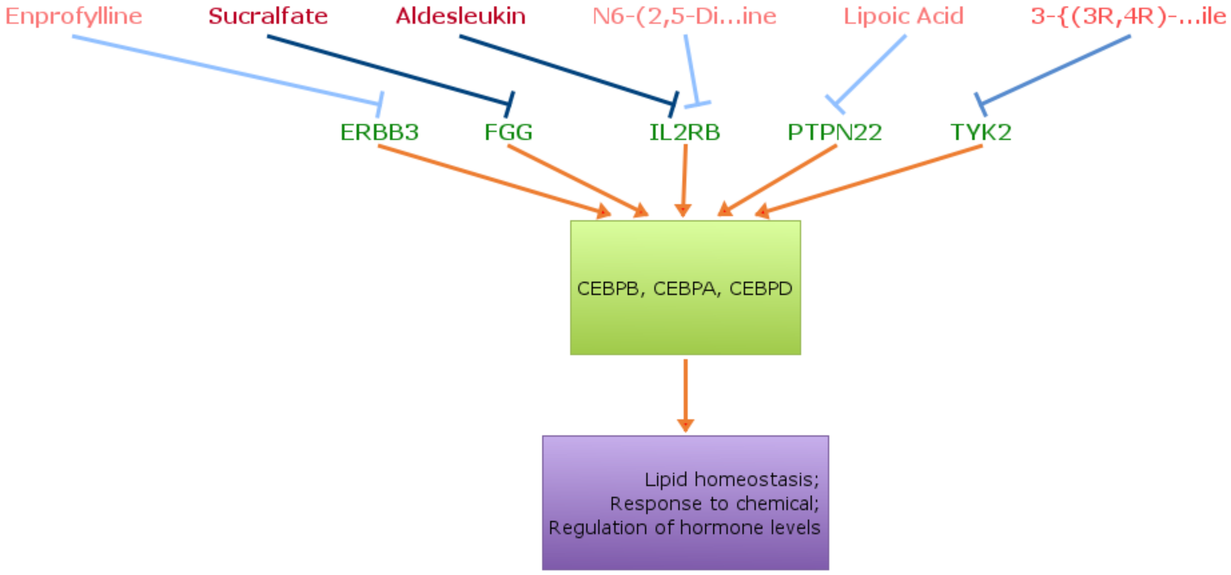
Name	Structure	Target names	Target activity score	Disease activity score	Drug rank
Enprofylline		ERBB3, BLK, ERBB4, TYK2, INSR	0.13	0	2
N6-(2,5-Dimethoxy-Benzyl)-N6-Methyl-Pyrido[2,3-D]Pyrimidine-2,4,6-Triamine		IL2RB, TYK2	0.13	0	3
4-(Hydroxymethyl)Benzamidine		IL2RB, TYK2	0.13	0	3
Pterin Cytosine Dinucleotide		ERBB3, ERBB4, INSR	0.12	0	5
Lipoic Acid		HNF1A, PTPN22	0.11	0	6

As a result of the drug search we came up with two lists of chemical compounds potentially applicable to the targets of our interest. The first list is based on drugs that are known as ligands for the revealed targets in the context of the diseases in our focus as well as in other disease conditions. The second list of identified compounds is based on the prediction of their potential biological activities, which was done using the program PASS. Such computational predictions should be taken as mere suggestions and should be used with care in further experiments.

5. Conclusion

We applied the software package "Genome Enhancer" to a data set that contains *genomics* data. The study is done in the context of *diabetes mellitus*. The data were pre-processed, statistically analyzed and genes carrying sequence variations were identified. Also checked was the enrichment of GO or disease categories among the studied gene sets.

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



6. Methods

Databases used in the study

Transcription factor binding sites in promoters and enhancers of differentially expressed genes were analyzed using known DNA-binding motifs described in the TRANSFAC® library, release 2020.1 (geneXplain GmbH, Wolfenbüttel, Germany) (http://genexplain.com/transfac).
The master regulator search uses the TRANSPATH® database (BIOBASE), release 2020.1 (geneXplain GmbH, Wolfenbüttel, Germany) (http://genexplain.com/transpath). A comprehensive signal transduction network of human cells is built by the software on the basis of reactions annotated in TRANSPATH®.
The information about drugs corresponding to identified drug targets and clinical trials references were extracted from HumanPSD™ database, release 2020.1 (http://genexplain.com/humanpsd).
The Ensembl database release Human88.38 (hg38) (http://www.ensembl.org) was used for gene IDs representation and Gene Ontology (GO) (http://geneontology.org) was used for functional classification of the studied gene set.

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

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

Methods for finding master regulators in networks

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

Methods for analysis of pharmaceutical compounds

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

Method for analysis of known pharmaceutical compounds

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

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

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

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

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

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

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

Method for prediction of pharmaceutical compounds

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

1. Toxicity below a chosen toxicity threshold (defines as *Pa*, probability to be active as toxic substance).

2. For all predicted pharmacological effects that correspond to a set of user selected disease(s) *Pa* is greater than a chosen effect threshold.
3. There are at least 2 targets (corresponding to the predicted activity-mechanisms) with predicted *Pa* greater than a chosen target threshold.

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

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

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

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

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

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

7. References

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

0. Boyarskikh U, Pintus S, Mandrik N, Stelmashenko D, Kiselev I, Evshin I, Sharipov R, Stegmaier P, Kolpakov F, Filipenko M, Kel A. Computational master-regulator search reveals mTOR and PI3K pathways responsible for low sensitivity of NCI-H292 and A427 lung cancer cell lines to cytotoxic action of p53 activator Nutlin-3. *BMC Med Genomics.* **2018**;11(1):12. doi:10.1186/1471-2105-7-s2-s13

1. Filimonov D, Poroikov V. Probabilistic Approaches in Activity Prediction. Varnek A, Tropsha A. *Chemoinformatics Approaches to Virtual Screening.* Cambridge (UK): RSC Publishing. **2008**;:182-216.

2. Filimonov DA, Poroikov VV. Prognosis of specters of biological activity of organic molecules. *Russian chemical journal.* **2006**;50(2):66-75 (russ)

3. Filimonov D, Poroikov V, Borodina Y, Gloriozova T. Chemical Similarity Assessment Through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *ChemInform.* **1999**;39(4):666-670. doi:10.1002/chin.199940210

Thank you for using the Genome Enhancer!

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

Supplementary material

1. Supplementary table 1 - Detailed report. Composite modules and master-regulators (genes carrying SNP variations in Experiment).
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.