Sequence and Pathway analysis

LDLR and PTPN11 are promising druggable targets for treating Diabetes Mellitus that control activity of NR3C1, POU5F1 and FOXO1 transcription factor on promoters of genes carrying sequence variations

Demo User
geneXplain GmbH
info@genexplain.com

Data received on 19/02/2020 ; Run on 09/04/2022 ; Report generated on 09/04/2022

Genome Enhancer release 3.0 (TRANSFAC®, TRANSPATH® and HumanPSD™ release 2022.1)

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) 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 genes carrying sequence variations: NR3C1, POU5F1 and FOXO1. The subsequent network analysis suggested

- SHP-2
- IGF-1
- LDL receptor
- InsR

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: Insulin Aspart, Porfimer and Lapatinib.

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, 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
E04_Genomics_SNP_diabetes	Genomics

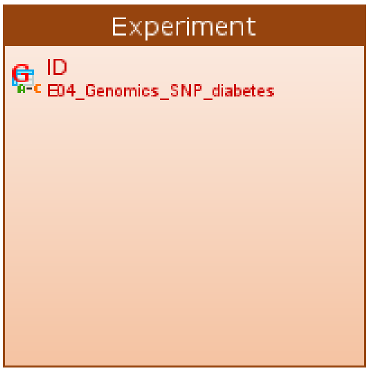


Figure 1. Annotation diagram of experimental data used in this study. With the colored boxes we show those sub-categories of the data that are compared in our analysis.

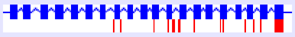
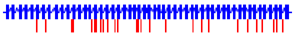
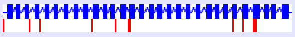
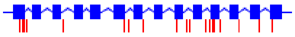
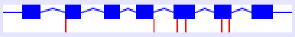
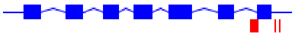
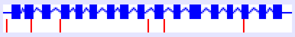
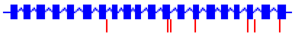
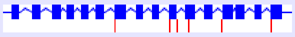
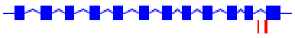
3. Results

We have analyzed 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. The most frequently mutated genes were used as target genes.

Table 2. Top ten the most frequently mutated genes in Experiment.
See full table →

ID	Gene description	Gene symbol	Gene schematic representation	Number of variations	Gene weight	Weighted score
ENSG00000130164	low density lipoprotein receptor	LDLR		30	93.08	186.16
ENSG00000165029	ATP binding cassette subfamily A member 1	ABCA1		30	82.12	164.24
ENSG00000084674	apolipoprotein B	APOB		16	41.69	83.37
ENSG00000169174	proprotein convertase subtilisin/kexin type 9	PCSK9		20	55.09	55.09
ENSG00000135100	HNF1 homeobox A	HNF1A		6	17.35	52.06
ENSG00000161888	SPC24 component of NDC80 kinetochore complex	SPC24		15	49.56	49.56
ENSG00000087237	cholesteryl ester transfer protein	CETP		7	22.05	44.09
ENSG00000160200	cystathionine beta-synthase	CBS		9	28.63	42.94
ENSG00000101076	hepatocyte nuclear factor 4 alpha	HNF4A		7	19.02	38.04
ENSG00000111424	vitamin D receptor	VDR		3	12.07	36.2

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.

The most frequently mutated genes in Experiment:

282 top mutated genes were taken for the mapping.

GO (biological process)

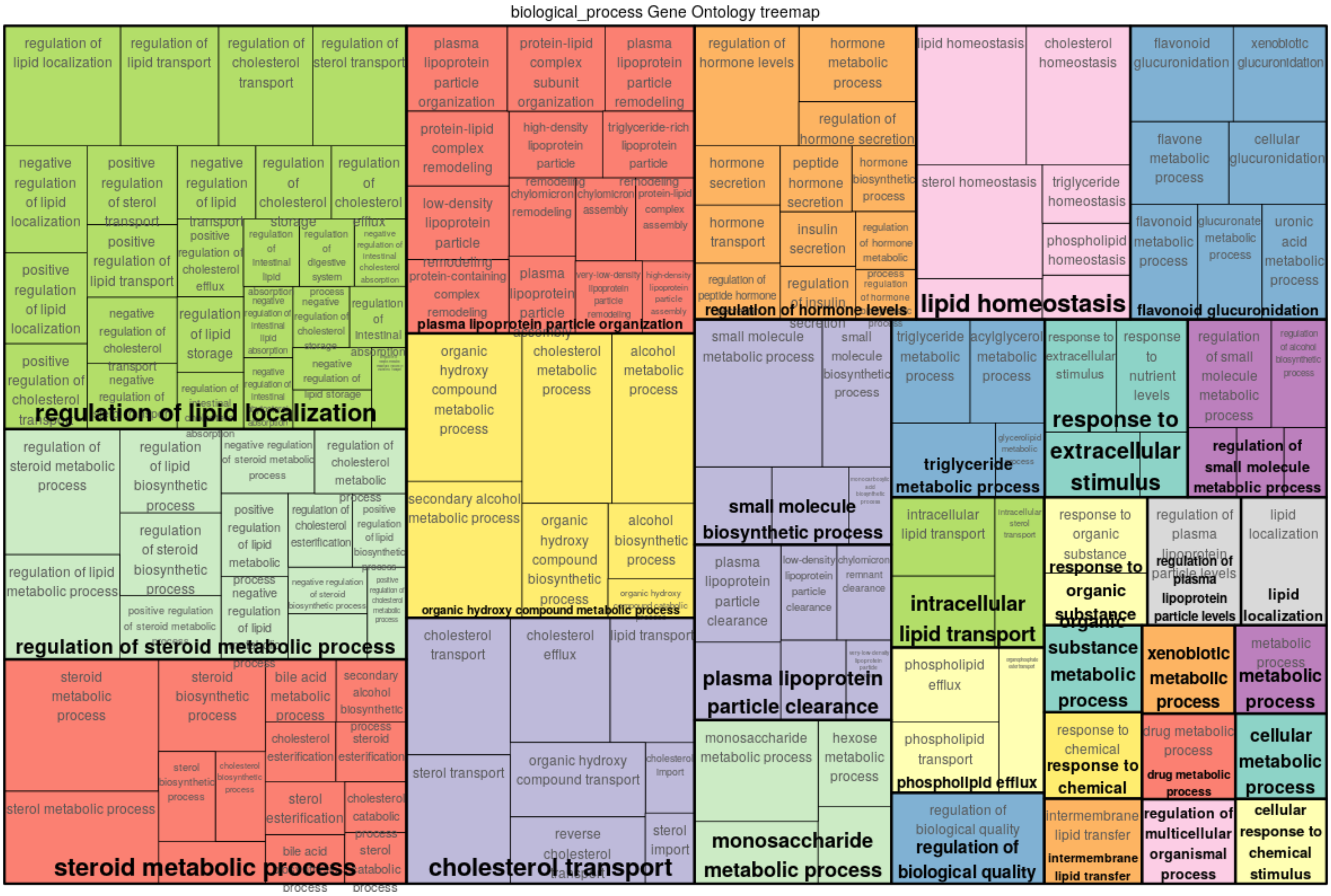


Figure 2. Enriched GO (biological process) of the most frequently mutated genes in Experiment.

Full classification →

TRANSPATH® Pathways (2022.1)

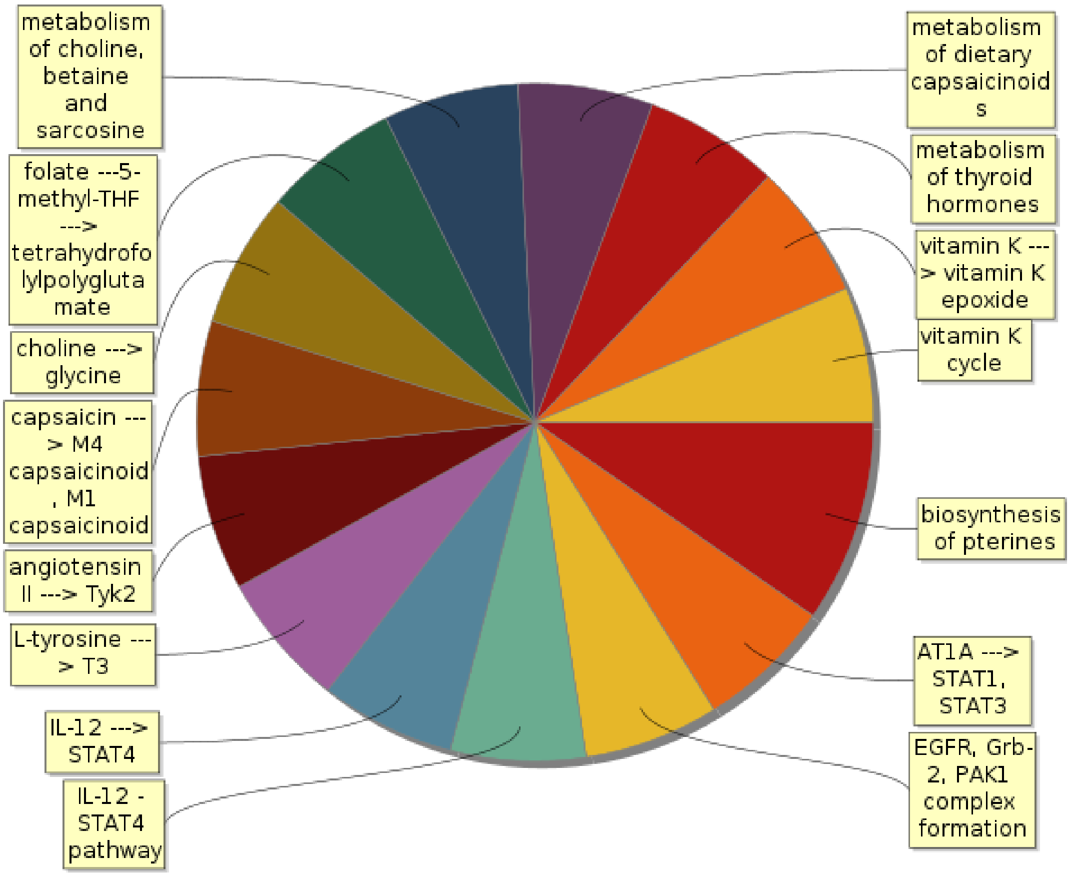


Figure 3. Enriched TRANSPATH® Pathways (2022.1) of the most frequently mutated genes in Experiment.

Full classification →

HumanPSD(TM) disease (2022.1)

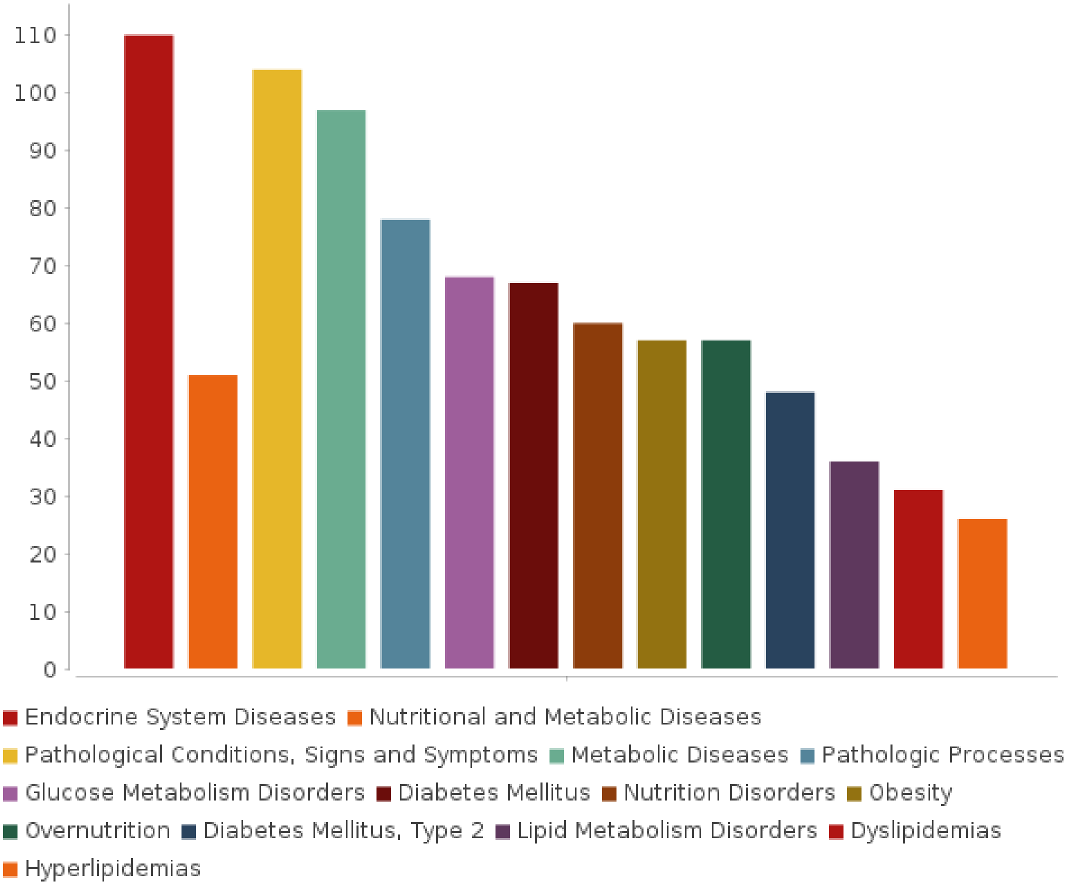
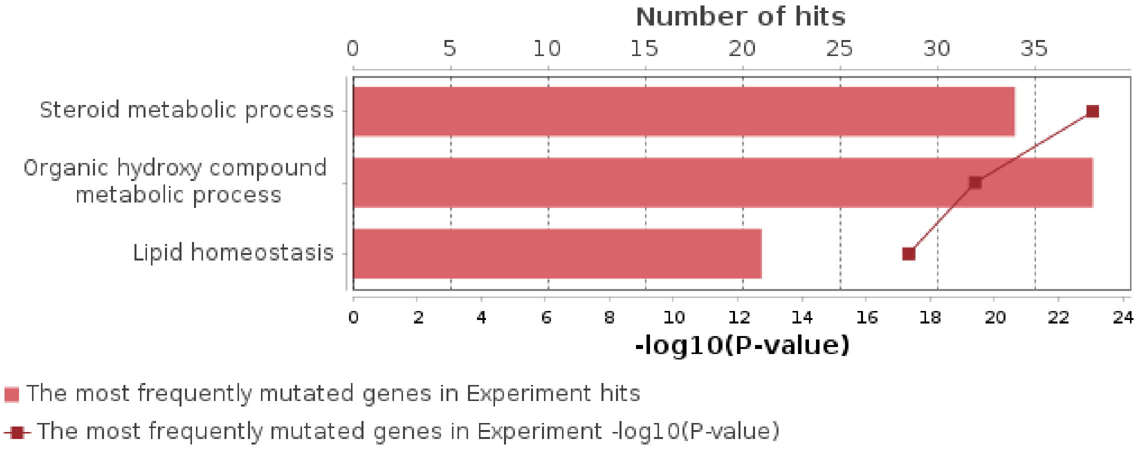


Figure 4. Enriched HumanPSD(TM) disease (2022.1) of the most frequently mutated genes in Experiment. The size of the bars correspond to the number of biomarkers of the given disease found among the input set.
Full classification →

The result of overall Gene Ontology (GO) analysis of the genes carrying sequence variations of the studied pathology can be summarized by the following diagram, revealing the most significant functional categories overrepresented among the observed (genes carrying sequence variations):



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 Genomics data from the "Yes VCF track" track to predict positions of potential ***enhancers*** where the observed sequence variations may influence the gene expression in the pathology under study. We scan 5kb flanking regions and the body of all genes caring the variations, with a sliding window of 1100bp size and find the position of the window with the maximal sum of the mutation weights, where we then perform the search for potential condition-specific enhancers (CMA model search).

We analyzed mutations that were revealed in the potential enhancers located upstream, downstream or inside the ***target genes*** (see Table 3). We identified 960 mutations potentially affecting gene regulation. Table 4 shows the following lists of PWMs whose sites were lost or gained due to these mutations. Weighting of mutations was done in respect to the significance of the change in TF affinity binding to the sequence. Mutations that maximally affected the change of binding affinity received higher weights. 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 Methods section).

Table 3. Mutations revealed in carrying SNP variation

See full table →

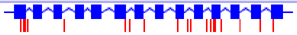
ID	Gene symbol	Gene schematic representation	Number of variations
ENSG00000130164	LDLR		46
ENSG00000161888	SPC24		30
ENSG00000165029	ABCA1		30
ENSG00000169174	PCSK9		20
ENSG00000084674	APOB		16
ENSG00000197114	ZGPAT		12
ENSG00000273154	ENSG00000273154		12
ENSG00000068781	STON1-GTF2A1L		11
ENSG00000140830	TXNL4B		9
ENSG00000157978	LDLRAP1		9

Table 4. PWMs whose sites were lost or gained due to mutations in carrying SNP variation

See full table →

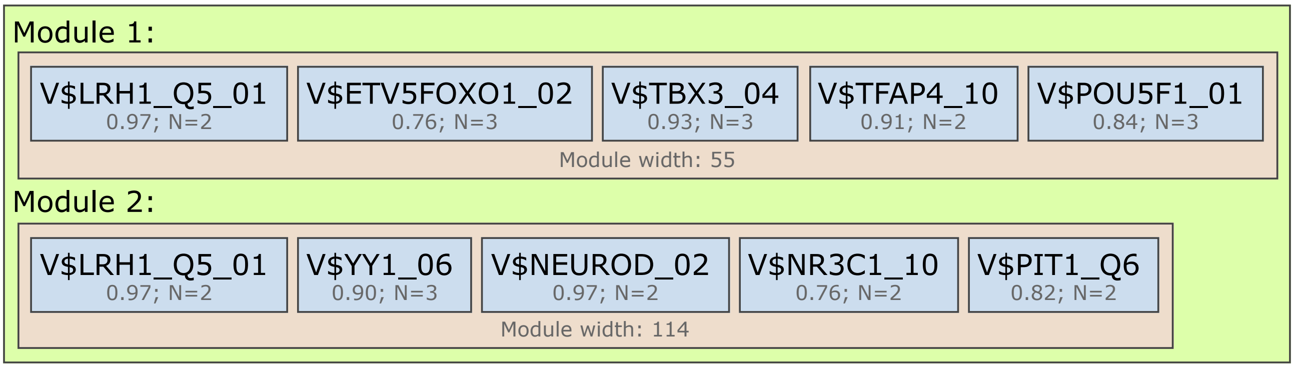
ID	P-value (gains)	P-value (losses)	yesCount (gains)	yesCount (losses)
V$E2F1HES7_02	4.98E-2	9.54E-5	4	10
V$ARNTL_04	2.48E-2	1.71E-5	3	30
V$E2F3HES7_01	2.1E-2	3.53E-4	40	41
V$MYC_07	1.59E-2	2.93E-4	6	5
V$FXR_02	1.11E-2	2.78E-4	4	40
V$RXRB_04	1.11E-2	2.84E-4	4	14
V$HES1_05	7.32E-3	2.04E-5	2	16
V$WT1_Q6_01	7.32E-3	4.73E-4	2	7
V$EGR3_07	2.52E-3	1.27E-4	2	3
V$BTEB2_Q3_01	1.58E-3	1.27E-4	5	3
V$PAX5_11	1.58E-3	3.79E-4	5	4
V$SMAD6_02	1.53E-4	4.1E-2	11	16
V$ZFP281_02	1.23E-4	6.54E-4	6	4
V$ZBTB7B_06	1.19E-4	4.93E-2	19	47
V$E2F1_15	1.17E-4	1.1E-2	40	74
V$EGR2_06	1.15E-4	8.06E-3	5	28
V$SP1_12	1.15E-4	4.25E-3	5	4
V$FKLF_Q5	1.01E-4	4.45E-2	13	5
V$EKLF_02	8.76E-5	1.42E-2	8	2
V$SP1_10	8.76E-5	3.31E-2	8	2

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 (the most frequently mutated genes in Experiment).

To build the most specific composite modules we choose top mutated genes as the input of CMA algorithm. The obtained CMA model is then applied to compute CMA score for all the most frequently mutated genes in Experiment.

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$LRH1_Q5_01 0.97; N=2
V$ETV5FOXO1_02 0.76; N=3
V$TBX3_04 0.93; N=3
V$TFAP4_10 0.91; N=2
V$POU5F1_01 0.84; N=3

Module width: 55

Module 2:

V$LRH1_Q5_01 0.97; N=2
V$YY1_06 0.90; N=3
V$NEUROD_02 0.97; N=2
V$NR3C1_10 0.76; N=2
V$PIT1_Q6 0.82; N=2

Module width: 114

Model score (-p*log10(pval)): 29.31
Wilcoxon p-value (pval): 3.29e-59
Penalty (p): 0.501
Average yes-set score: 9.85
Average no-set score: 8.08
AUC: 0.81
Separation point: 8.92
False-positive: 29.17%
False-negative: 21.63%
The AUC of the model achieves value significantly higher than expected for a random set of regulatory regions
Z-score = 3.51

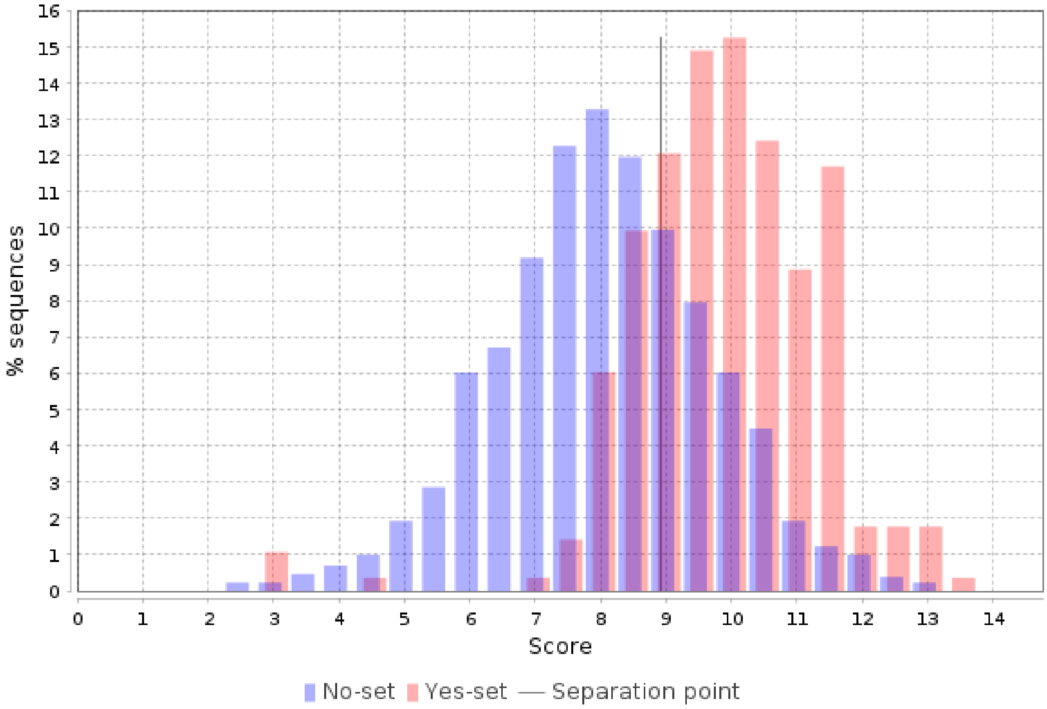


Table 5. List of top ten the most frequently mutated genes in Experiment with identified enhancers in their regulatory regions. ***CMA score*** *- the score of the CMA model of the enhancer identified in the regulatory region.*

See full table →

Ensembl IDs	Gene symbol	Gene description	CMA score	Factor names
ENSG00000089057	SLC23A2	solute carrier family 23 member 2	13.44	POU5F1(h), POU1F1(h), YY1(h), ETV5(h),FOXO1(h), GR(h), TBX3(h), NeuroD1(h)...
ENSG00000179295	PTPN11	protein tyrosine phosphatase non-receptor type 11	12.89	ETV5(h),FOXO1(h), POU5F1(h), YY1(h), LRH-1(h), GR(h), TBX3(h), AP-4(h)...
ENSG00000171435	KSR2	kinase suppressor of ras 2	12.82	POU5F1(h), YY1(h), LRH-1(h), TBX3(h), GR(h), NeuroD1(h), ETV5(h),FOXO1(h)...
ENSG00000101464	PIGU	phosphatidylinositol glycan anchor biosynthesis class U	12.8	POU1F1(h), TBX3(h), POU5F1(h), GR(h), ETV5(h),FOXO1(h), YY1(h), LRH-1(h)...
ENSG00000162551	ALPL	alkaline phosphatase, biomineralization associated	12.76	POU1F1(h), ETV5(h),FOXO1(h), TBX3(h), YY1(h), GR(h), LRH-1(h), AP-4(h)...
ENSG00000169344	UMOD	uromodulin	12.47	POU1F1(h), ETV5(h),FOXO1(h), POU5F1(h), TBX3(h), AP-4(h), GR(h), LRH-1(h)...
ENSG00000176890	TYMS	thymidylate synthetase	12.4	GR(h), NeuroD1(h), LRH-1(h), YY1(h), POU5F1(h), ETV5(h),FOXO1(h), TBX3(h)...
ENSG00000176912	TYMSOS	TYMS opposite strand	12.4	GR(h), NeuroD1(h), LRH-1(h), YY1(h), POU5F1(h), ETV5(h),FOXO1(h), TBX3(h)...
ENSG00000225670	CADM3-AS1	CADM3 antisense RNA 1	12.35	LRH-1(h), GR(h), YY1(h), NeuroD1(h), POU1F1(h), POU5F1(h), ETV5(h),FOXO1(h)...
ENSG00000213088	ACKR1	atypical chemokine receptor 1 (Duffy blood group)	12.35	LRH-1(h), GR(h), YY1(h), NeuroD1(h), POU1F1(h), POU5F1(h), ETV5(h),FOXO1(h)...

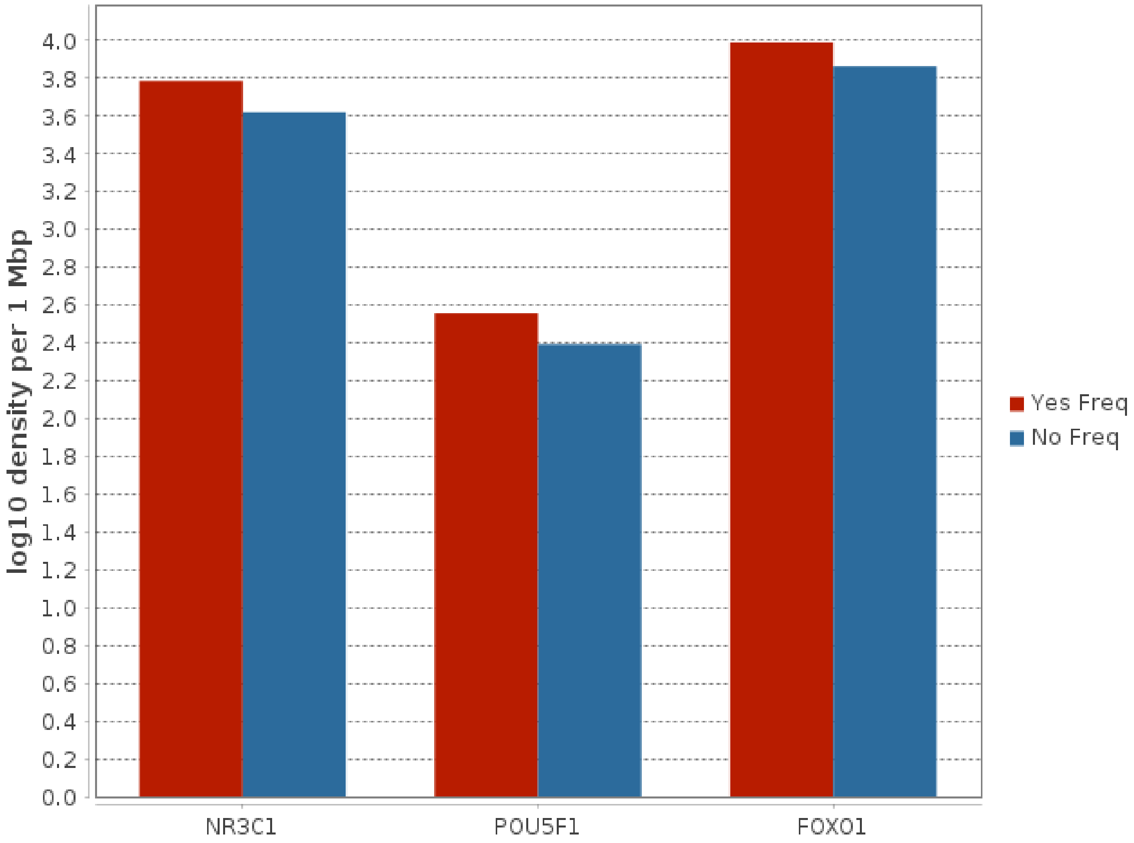
On the basis of the enhancer models we identified transcription factors potentially regulating the ***target genes*** of our interest. We found 10 transcription factors controlling expression of the genes associated with genomic variations (see Table 6).

Table 6. Transcription factors of the predicted enhancer model potentially regulating the genes carrying sequence variations (the most frequently mutated genes 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 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
MO000031266	NR3C1	nuclear receptor subfamily 3 group C member 1	2.75	1.47
MO000056618	POU5F1	POU class 5 homeobox 1	2.15	1.46
MO000034454	FOXO1	forkhead box O1	2.04	1.34
MO000024660	TFAP4	transcription factor AP-4	1.81	1.53
MO000078913	YY1	YY1 transcription factor	1.78	1.33
MO000028384	NEUROD1	neuronal differentiation 1	1.65	3.71
MO000026742	NR5A2	nuclear receptor subfamily 5 group A member 2	1.43	8.92
MO000084573	POU1F1	POU class 1 homeobox 1	1.43	1.45
MO000038600	ETV5	ETS variant transcription factor 5	0	1.13
MO000118269	TBX3	T-box transcription factor 3	0	4.57

The following diagram represents the key transcription factors, which were predicted to be potentially regulating genes carrying sequence variations in the analyzed pathology: NR3C1, POU5F1 and FOXO1.



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 1 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 carrying SNP variation

See full table →

ID	Title	Mutation count	Consequence	Codons
MO000127845	TRPM6(h)	3	stop_gained	Aag/Tag

Top 1 mutated proteins for carrying SNP variation were used in the algorithm of master regulator search as a list of nodes of the signal transduction network that are removed from the network during the search of master regulators (see more details about the algorithm in the Methods 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 the most frequently mutated genes 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	Weighted score
MO000007566	InsR(h)	INSR	insulin receptor	31	23.23
MO000061801	LDL receptor(h)	LDLR	low density lipoprotein receptor	32	186.16
MO000115795	LDL receptor-isoform1(h)	LDLR	low density lipoprotein receptor	32	186.16
MO000271239	LDL receptor-isoform2(h)	LDLR	low density lipoprotein receptor	32	186.16
MO000271240	LDL receptor-isoform3(h)	LDLR	low density lipoprotein receptor	32	186.16
MO000271241	LDL receptor-isoform4(h)	LDLR	low density lipoprotein receptor	32	186.16
MO000286473	LDL receptor-isoform5(h)	LDLR	low density lipoprotein receptor	32	186.16
MO000319009	LDL receptor-isoform6(h)	LDLR	low density lipoprotein receptor	32	186.16
MO000484562	LDL receptor(h){ub}	LDLR	low density lipoprotein receptor	32	186.16
MO000057585	InsR(h){pY}	INSR	insulin receptor	35	23.23

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.

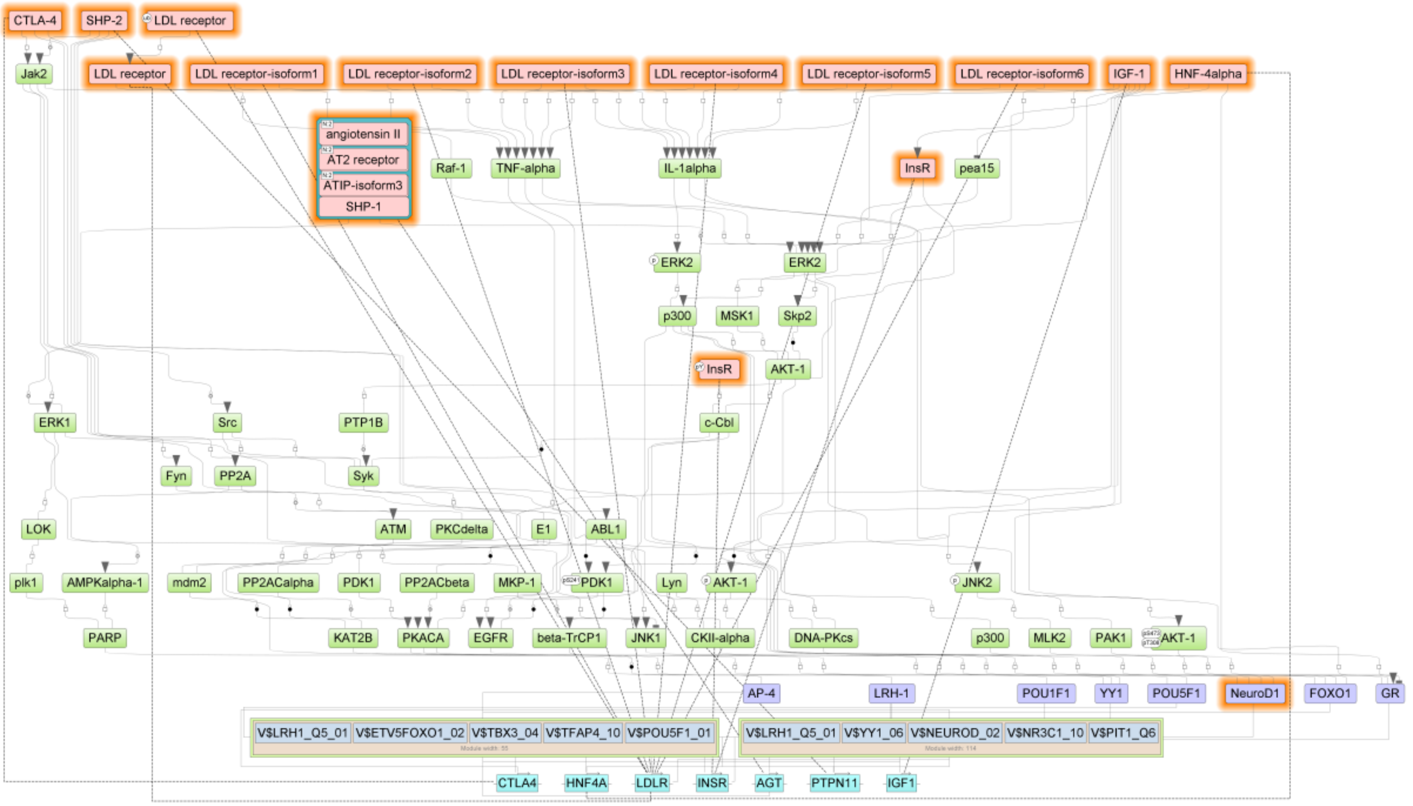


Figure 5. Diagram of intracellular regulatory signal transduction pathways of the most frequently mutated genes in Experiment. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate molecules, which have been added to the network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

4. 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 9. 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	Weighted score
LDLR	low density lipoprotein receptor	12	32	186.16
IGF1	insulin like growth factor 1	2	48	11.43
INSR	insulin receptor	47	58	23.23
CTLA4	cytotoxic T-lymphocyte associated protein 4	7	59	25.94
HNF4A	hepatocyte nuclear factor 4 alpha	7	63	38.04
ADRB2	adrenoceptor beta 2	73	78	22.31

*Table 10. 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.*

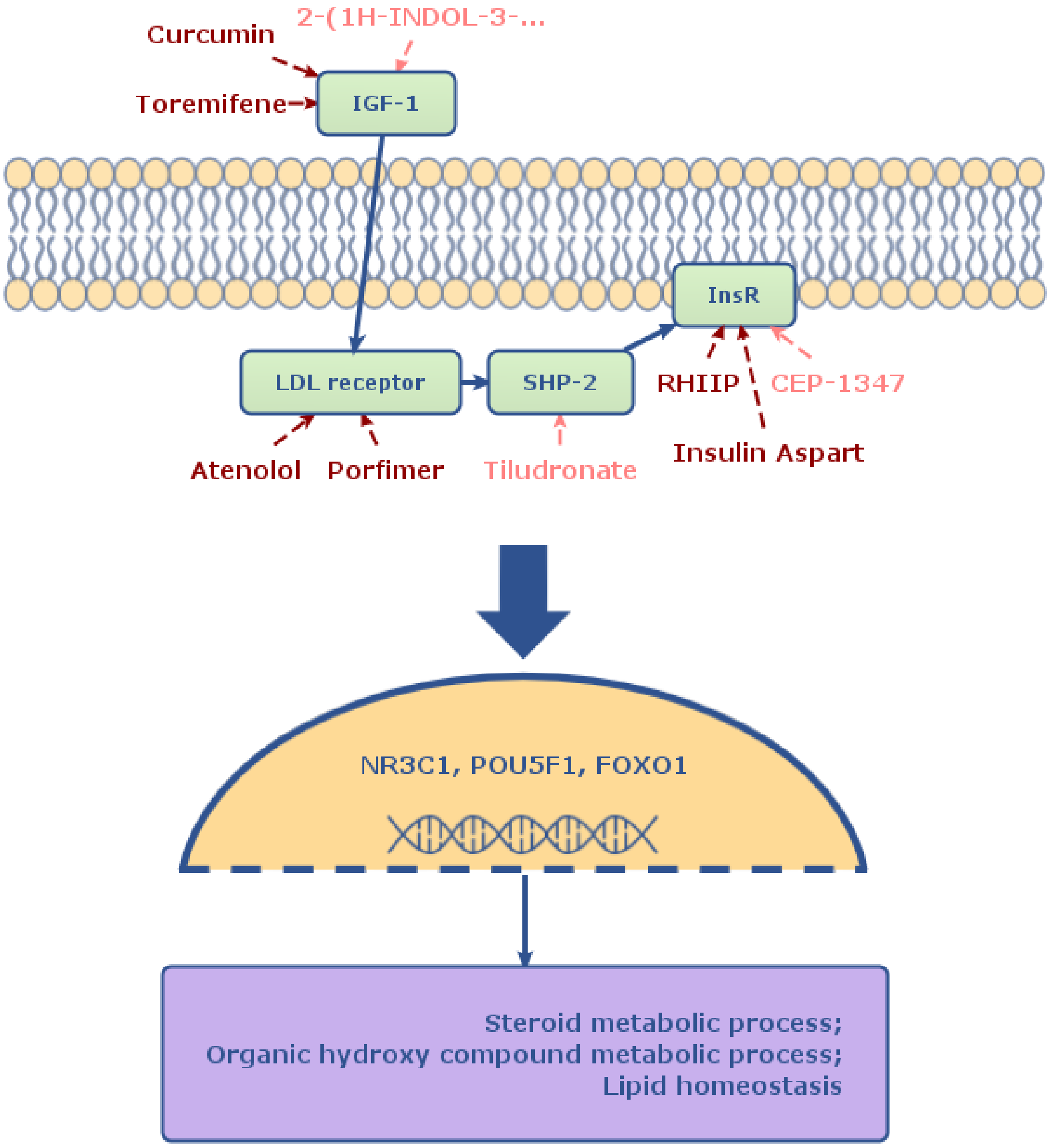
See full table →

Gene symbol	Gene Description	Druggability score	Total rank	Weighted score
PTPN11	protein tyrosine phosphatase non-receptor type 11	21.51	45	5.26
IGF1	insulin like growth factor 1	19.54	48	11.43
INSR	insulin receptor	12.03	58	23.23
HNF4A	hepatocyte nuclear factor 4 alpha	2.42	63	38.04
ADRB2	adrenoceptor beta 2	18.81	78	22.31
TYK2	tyrosine kinase 2	12.38	83	7.08

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:

- SHP-2
- IGF-1
- LDL receptor
- InsR

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: Insulin Aspart, CEP-1347, Toremifene, Curcumin, Atenolol, Porfimer, 2-(1H-INDOL-3-YL)ACETAMIDE, Tiludronate and RHIIP, 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 11 and 12), 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 11. 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
Insulin Aspart	INSR	100	12	Phase 4: Diabetes Mellitus, Arteriosclerosis, Atherosclerosis, Diabetes Mellitus, Type 1, Diabetes Mellitus, Type 2, Hyperglycemia, Hyperkalemia, Insulin Resistance
Insulin Detemir	INSR	100	12	Phase 4: Diabetes Mellitus, Arteriosclerosis, Atherosclerosis, Cardiovascular Diseases, Diabetes Mellitus, Type 1, Diabetes Mellitus, Type 2, Glucose Metabolism Disorders, Hyperglycemia, Vascular Diseases
Insulin Glulisine	INSR	100	12	Phase 4: Diabetes Mellitus, Diabetes Mellitus, Type 1, Diabetes Mellitus, Type 2, Diabetic Nephropathies, Hyperglycemia, Kidney Diseases, Renal Insufficiency
Insulin Lispro	INSR	96	12	Phase 4: Diabetes Mellitus, Diabetes Mellitus, Type 1, Diabetes Mellitus, Type 2, Diabetes, Gestational, Hyperglycemia, Infarction, Myocardial Infarction
Insulin Glargine	INSR	96	12	Phase 4: Diabetes Mellitus, Coronary Artery Disease, Coronary Disease, Diabetes Mellitus, Type 1, Diabetes Mellitus, Type 2, Diabetic Foot, Diabetic Retinopathy, Fatty Liver, Fatty Liver, Alcoholic, Foot Ulcer, Gastrointestinal Neoplasms, Heart Failure, Hyperglycemia, Hypoglycemia, Infarction, Intestinal Neoplasms, Ischemia, Liver Diseases, Myocardial Infarction, Myocardial Ischemia, Neoplasm Metastasis, Neoplasms, Non-alcoholic Fatty Liver Disease, Renal Insufficiency, Retinal Diseases, Ulcer

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

Repurposing drugs

Table 12. 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
Porfimer	LDLR	78	Phase 3: Central Nervous System Neoplasms, Cholangiocarcinoma, Glioma, Klatskin Tumor, Neoplasms, Nervous System Neoplasms
RHIIP	INSR	78	N/A
ipilimumab	CTLA4	76	Phase 4: Carcinoma, Renal Cell, Melanoma
AVI-4557	HNF4A	75	N/A
tributyrin	LDLR	75	Phase 1: Neoplasms, Prostatic Neoplasms

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

No prospective drugs were found, which would be predicted by PASS software to be active against the identified drug targets and would be predicted to have biological activity against the studied disease(s).

Table 13. 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
Lapatinib	ERBB3, ERBB4	100	0.33
N-[4-(3-BROMO-PHENYLAMINO)-QUINAZOLIN-6-YL]-ACRYLAMIDE	ERBB3, GRK5, ERBB4	99	0.16
SB220025	IL1B, GRK5, ERBB4, TYK2	98	0.13
CI-1033	ERBB3, ERBB4	98	0.11
CEP-1347	ERBB3, ERBB4, INSR	97	0.1

As the result of drug search we propose the following drugs as most promising candidates for treating the pathology under study: Insulin Aspart, Porfimer and Lapatinib. These drugs were selected for acting on the following targets: INSR, LDLR and ERBB3, 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 *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 drugs as most promising candidates for treating the pathology under study:

Insulin Aspart, Porfimer and Lapatinib

These drugs were selected for acting on the following targets: INSR, LDLR and ERBB3, 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:

SHP-2, IGF-1, LDL receptor and InsR

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: Insulin Aspart, CEP-1347, Toremifene, Curcumin, Atenolol, Porfimer, 2-(1H-INDOL-3-YL)ACETAMIDE, Tiludronate and RHIIP. 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 genes carrying sequence variations 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.

- SHP-2
- IGF-1
- LDL receptor
- InsR

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 genes carrying sequence variations were analyzed using known DNA-binding motifs described in the TRANSFAC® library, release 2022.1 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transfac).
The master regulator search uses the TRANSPATH® database (BIOBASE), release 2022.1 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transpath). A comprehensive signal transduction network of human cells is built by the software on the basis of reactions annotated in TRANSPATH®.
The information about drugs corresponding to identified drug targets and clinical trials references were extracted from HumanPSD™ database, release 2022.1 (https://genexplain.com/humanpsd).
The Ensembl database release Human104.38 (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.

Genomic data processing

When analyzing a list of genomic variations (from input vcf file or computed by Genome Enhancer from SNP list or from fastq files), first of all, we compute a specific mutation weight (w_1) for each variation depending on it’s location in gene body and gene flanking regions (-1000 upstream and +1000 downstream of the gene body).

w_1 = 0.7 for variations in exon area
w_1 = 1.3 for variations in promoter region (-1000bp upstream and 100bp downstream of TSS),
w_1 = 1.0 for variations in other locations.

Next, VCF track (Yes track), provided as input or created by Genome Enhancer from SNP list or fastq files, is compared to Random VCF track (No track) of 10000 random human variations. On both tracks we calculate the score delta values (differences between PWM score values of the TF sites with the reference or with the alternative allele of the considered variation). For each variation we find then the maximal score delta values at each PWM leading either to the gain or to the loss of TF site (with the alternative allele). For selecting the maximum score delta values we consider both directions of DNA strand. Next, by going through all variations we compute two p-values for each PWM – the p-value of site losses and p-value of site gains. The p-values are computed using cumulative Binomial distribution estimating the random chances to observe the found high number of lost or gained TF sites in Yes track in the comparison to the No track. The PWM cut-offs are optimized to obtain the most extreme p-values. We further take top 20 best matrices by p-value from each: gained and lost sites and calculate the mutation weights on the Yes track on the basis of the obtained 40 matrices. Each mutation is assigned with a respective matrix that got the maximum delta value either for the site gain or for the site loss (changed the binding affinity most significantly). This delta is then compared to other delta values that were computed for the respective matrix on the No track. The eventual weight that reflects the transcription factor binding affinity change caused by the mutation is calculated as follows:

w_2 = -log10(NoGr / NoAll), if NoGr > 0
w_2 = -log10(1.0 / (2.0 * NoAll), if NoGr = 0

where NoGr is the number of deltas from the No track that appeared to be greater than the inspected delta and NoAll is the total number of deltas in the No track. The resulting track is then constructed that contains all sites of the initial Yes track together with the additional weights reflecting the transcription factor binding affinity change caused by the mutation.
The list of 40 matrices most affected by variations will be further used in composite modules search described in the next section.
Total Gene mutation weight is the sum of the weights w_1 of all variations located inside the gene body and in the gene flanking regions summed up with the weight w_2 that reflects the transcription factor binding affinity change caused by the mutation. This weight is calculated by estimating the importance of a certain mutation in terms of gains or losses of binding sites caused by it.
Next, a weighted score is calculated for all genes with the following formula:
Weighted score = In_disease * In_transpath * Gene mutation weight, where

In_disease = 2.0 for genes assigned to selected diseases,
In_transpath = 1.5 for genes mapped to Transpath pathways,
and In_disease = In_transpath = 1.0 in all other cases.

At the next step, 300 genes with highest weighted score are selected for further CMA model search.
The mutation weights ($w = w_1+w_2$) are also used to find the regulatory regions of the genes most affected by the variations/SNP. A sliding window of 1100 bp is used to scan through the intronic, 5′ and 3′ regions of the genes and a region is selected with the highest sum of the mutation weights.

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). Each composite module is forced to include at least one matrix that was identified as matrix causing the significant change in the transcription factor binding affinity as the result of the observed mutation.

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" ($T\text{-}score_{PSD}$),
2. ranking by "Disease activity score" ($D\text{-}score_{PSD}$),
3. ranking by "Clinical validity score".

"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.
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
0. Boyarskikh U, Pintus S, Mandrik N, Stelmashenko D, Kiselev I, Evshin I, Sharipov R, Stegmaier P, Kolpakov F, Filipenko M, Kel A. Computational master-regulator search reveals mTOR and PI3K pathways responsible for low sensitivity of NCI-H292 and A427 lung cancer cell lines to cytotoxic action of p53 activator Nutlin-3. *BMC Med Genomics.* **2018**;11(1):12. doi:10.1186/1471-2105-7-s2-s13
1. Filimonov D, Poroikov V. Probabilistic Approaches in Activity Prediction. Varnek A, Tropsha A. *Cheminformatics Approaches to Virtual Screening.* Cambridge (UK): RSC Publishing. **2008**;:182-216.
2. Filimonov DA, Poroikov VV. Prognosis of specters of biological activity of organic molecules. *Russian chemical journal.* **2006**;50(2):66-75 (russ)
3. Filimonov D, Poroikov V, Borodina Y, Gloriozova T. Chemical Similarity Assessment Through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *ChemInform.* **1999**;39(4):666-670. doi:10.1002/chin.199940210

Thank you for using the Genome Enhancer!

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

Supplementary material

1. Supplementary table 1 - Detailed report. Composite modules and master regulators (the most frequently mutated genes 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-todateness, 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.