SETD7 is promising druggable target for treating Lung Neoplasms that controls activity of CTCF, LEF1 and MECOM transcription factor on promoters of genes encoding enzymes metabolizing given metabolites

Demo User
geneXplain GmbH
info@genexplain.com
Data received on 06/09/2019 ; Run on 11/06/2020 ; Report generated on 11/06/2020

Genome Enhancer release 2.0 (TRANSFAC®, TRANSPATH® and HumanPSD™ release 2020.2)

Abstract

In the present study we applied the software package "Genome Enhancer" to a data set that contains *metabolomics* data. The study is done in the context of *Lung Neoplasms*. 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 encoding enzymes metabolizing given metabolites: CTCF, LEF1 and MECOM. The subsequent network analysis suggested

- setd7
- SUV39H1

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: S-Adenosyl-L-Homocysteine.

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 encoding enzymes metabolizing given metabolites 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 2507 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
TGF_72h vs. NO_TGF_72h	Metabolomics

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: TNF vs NO_TNF 72h.

3.1. Identification of target genes

In the first step of the analysis ***target genes*** were identified from the uploaded experimental data. The metabolites were mapped to Recon2 database. Then, genes encoding enzymes, which are involved in synthesis, degradation or modification of these metabolites were identified in Recon2 database. These genes (***target genes***) were then used for further upstream analysis.

Table 2. Top ten genes encoding enzymes, metabolising target metabolites in TNF vs NO_TNF 72h.
See full table →

ID	Gene description	Gene symbol	Recon2 ID	Title	logFC
ENSG00000109107	aldolase, fructose-bisphosphate C	ALDOC	M_pmtcrn	L-palmitoylcarnitine	12.21
ENSG00000110090	carnitine palmitoyltransferase 1A	CPT1A	M_dgsn,M_pmtcrn	Deoxyguanosine,L-palmitoylcarnitine	12.21
ENSG00000136872	aldolase, fructose-bisphosphate B	ALDOB	M_pmtcrn	L-palmitoylcarnitine	12.21
ENSG00000149925	aldolase, fructose-bisphosphate A	ALDOA	M_pmtcrn	L-palmitoylcarnitine	12.21
ENSG00000157184	carnitine palmitoyltransferase 2	CPT2	M_dgsn,M_pmtcrn	Deoxyguanosine,L-palmitoylcarnitine	12.21
ENSG00000169169	carnitine palmitoyltransferase 1C	CPT1C	M_dgsn,M_pmtcrn	Deoxyguanosine,L-palmitoylcarnitine	12.21
ENSG00000178537	solute carrier family 25 member 20	SLC25A20	M_pmtcrn	L-palmitoylcarnitine	12.21
ENSG00000205560	carnitine palmitoyltransferase 1B	CPT1B	M_dgsn,M_pmtcrn	Deoxyguanosine,L-palmitoylcarnitine	12.21
ENSG00000129673	aralkylamine N-acetyltransferase	AANAT	M_Nacsertn	N-acetylserotonin	11.77
ENSG00000196433	acetylserotonin O-methyltransferase	ASMT	M_Nacsertn,M_ahcys	N-acetylserotonin,S-Adenosyl-L-homocysteine	11.77

3.2. Functional classification of genes

A functional analysis of genes encoding enzymes metabolizing given metabolites 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 encoding enzymes, metabolising target metabolites in TNF vs NO_TNF 72h:

236 genes encoding enzymes, metabolising target metabolites genes were taken for the mapping.

GO (biological process)

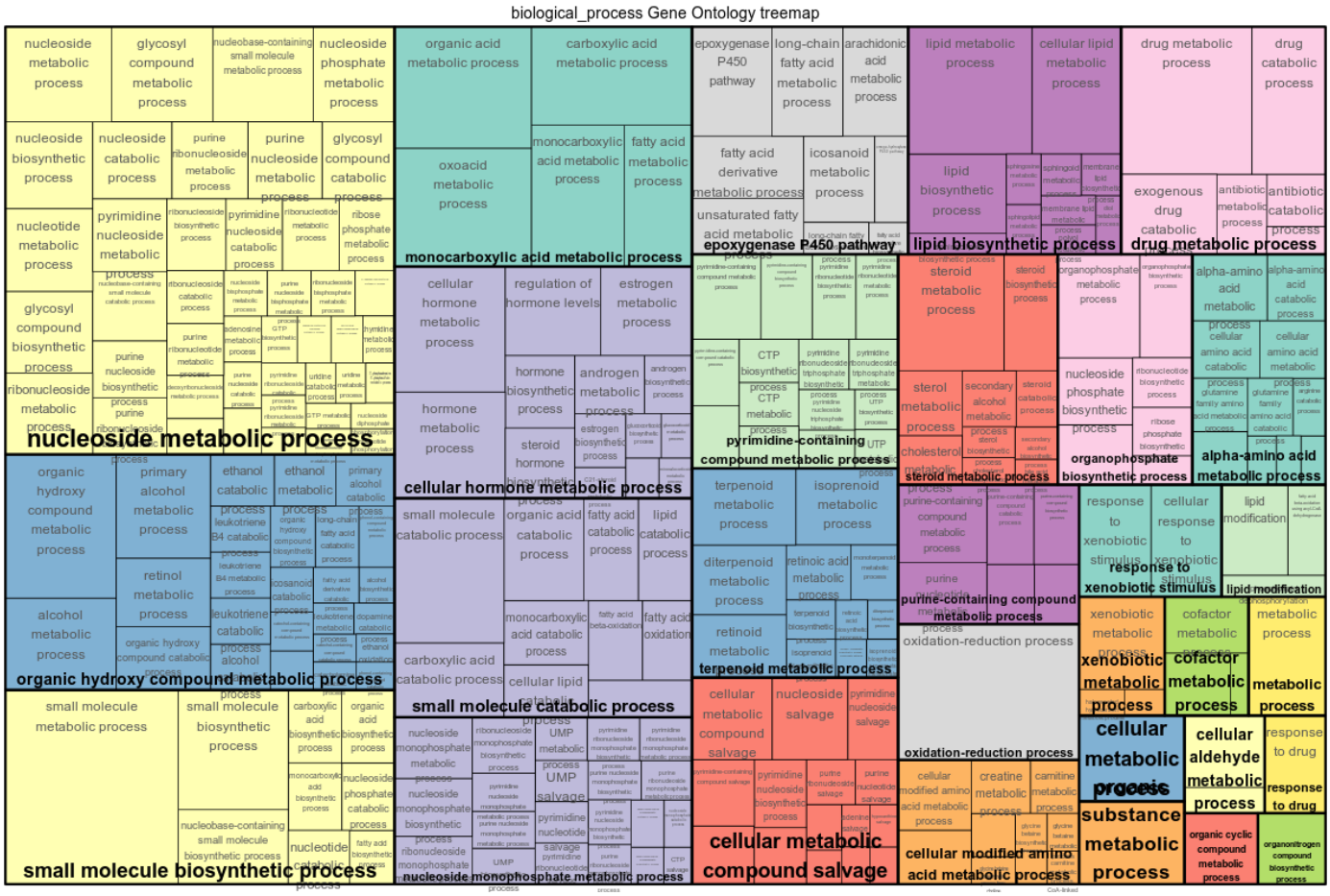


Figure 2. Enriched GO (biological process) of genes encoding enzymes, metabolising target metabolites in TNF vs NO_TNF 72h.

Full classification →

TRANSPATH® Pathways (2020.2)

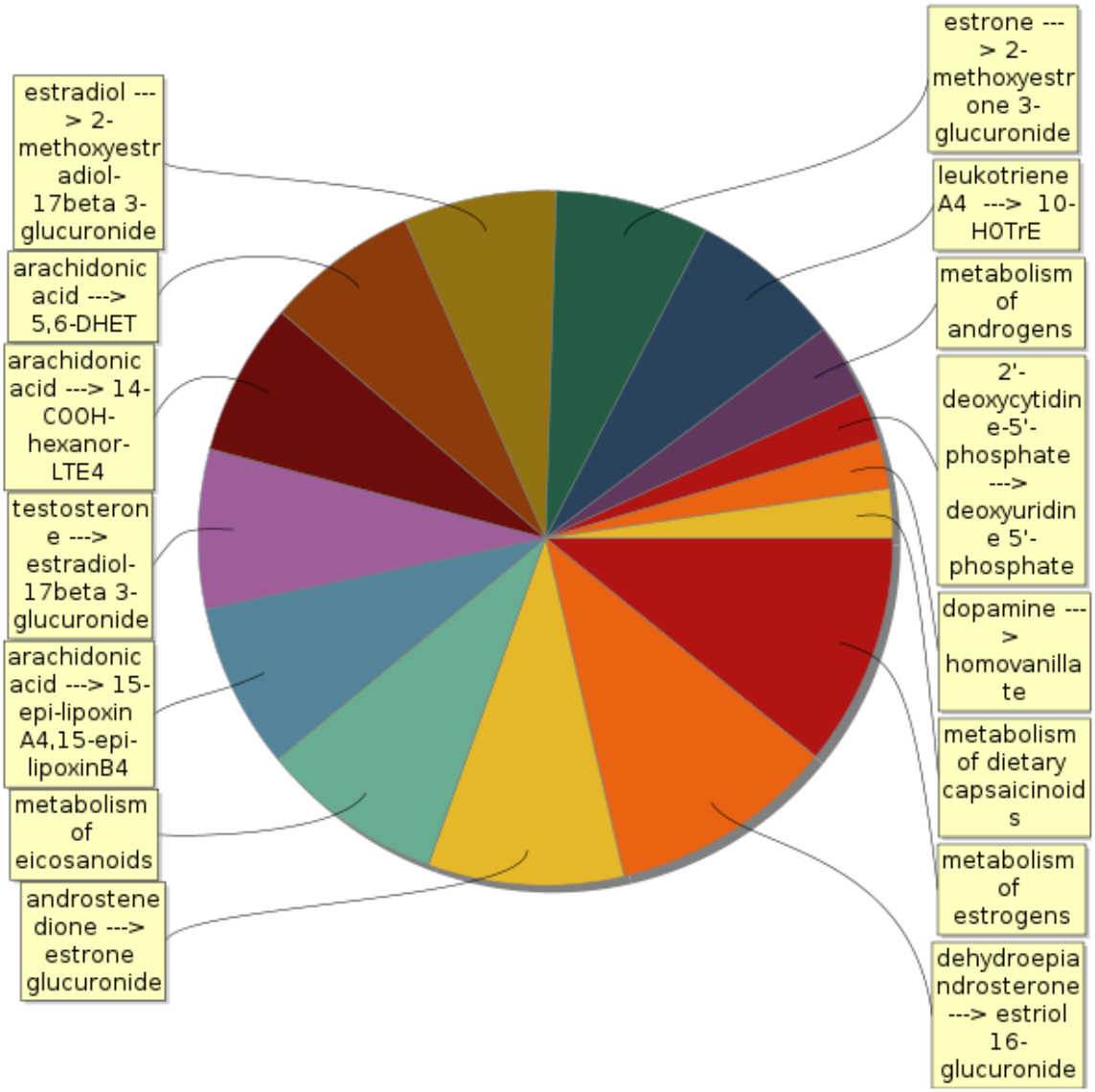


Figure 3. Enriched TRANSPATH® Pathways (2020.2) of genes encoding enzymes, metabolising target metabolites in TNF vs NO_TNF 72h.

Full classification →

HumanPSD(TM) disease (2020.2)

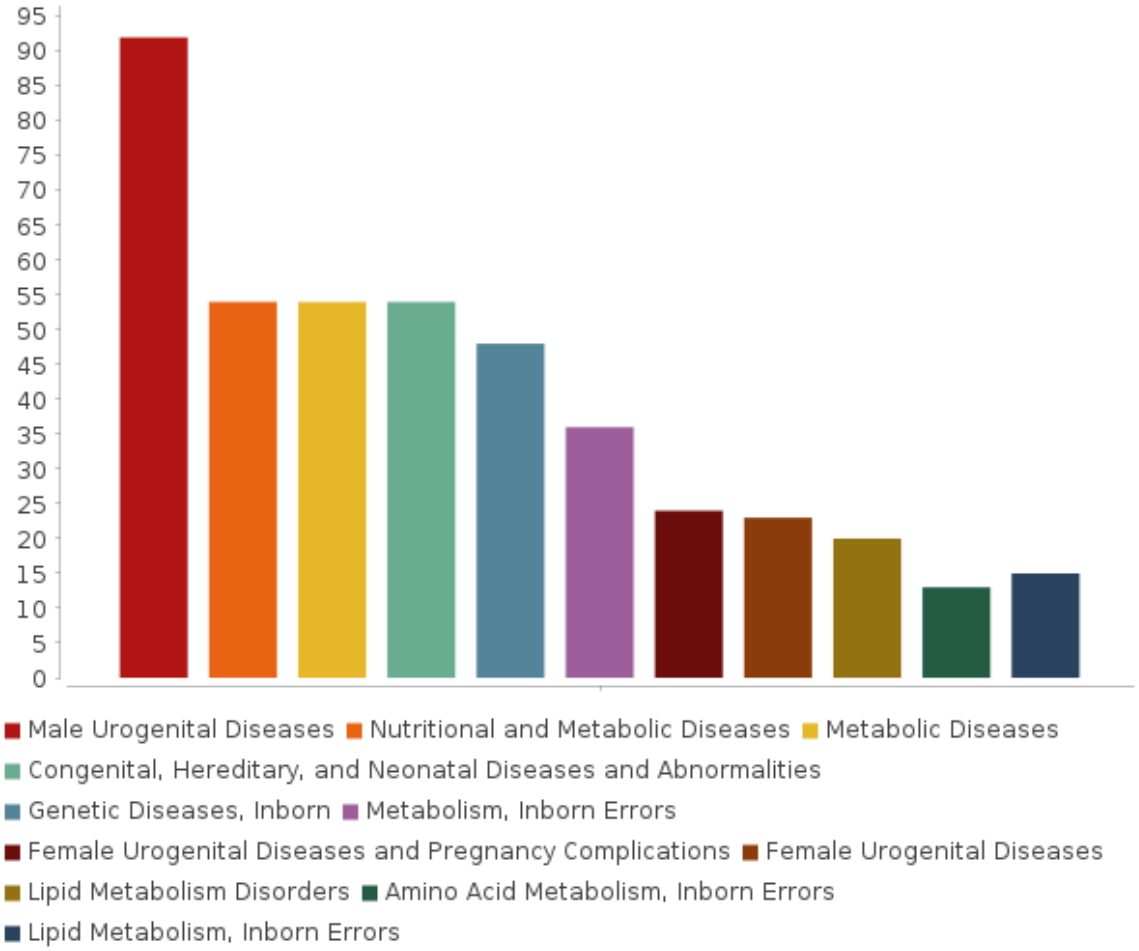
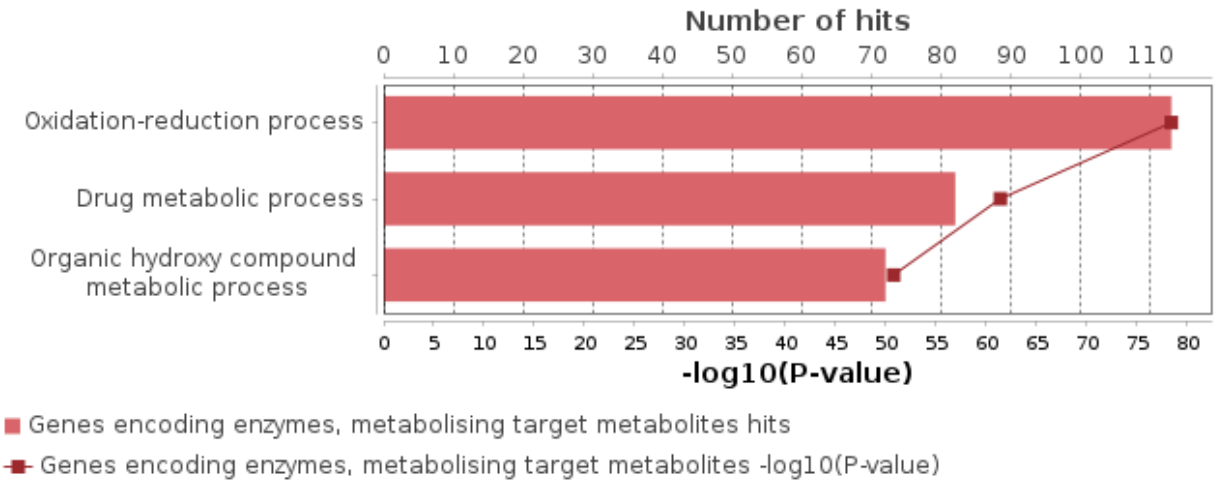


Figure 4. Enriched HumanPSD(TM) disease (2020.2) of genes encoding enzymes, metabolising target metabolites in TNF vs NO_TNF 72h. The size of the bars correspond to the number of bio-markers of the given disease found among the input set.

Full classification →

The result of overall Gene Ontology (GO) analysis of the genes encoding enzymes metabolizing given metabolites of the studied pathology can be summarized by the following diagram, revealing the most significant functional categories overrepresented among the observed (genes encoding enzymes metabolizing given metabolites):



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 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 encoding enzymes, metabolising target metabolites in TNF vs NO_TNF 72h).

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

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

Module 1:

V$RARG_Q3	V$FXR_Q2	V$CTCF_08	V$TCF7L2_06	V$POU6F1_01
0.00; N=3	0.00; N=2	0.00; N=3	0.00; N=2	0.87; N=2

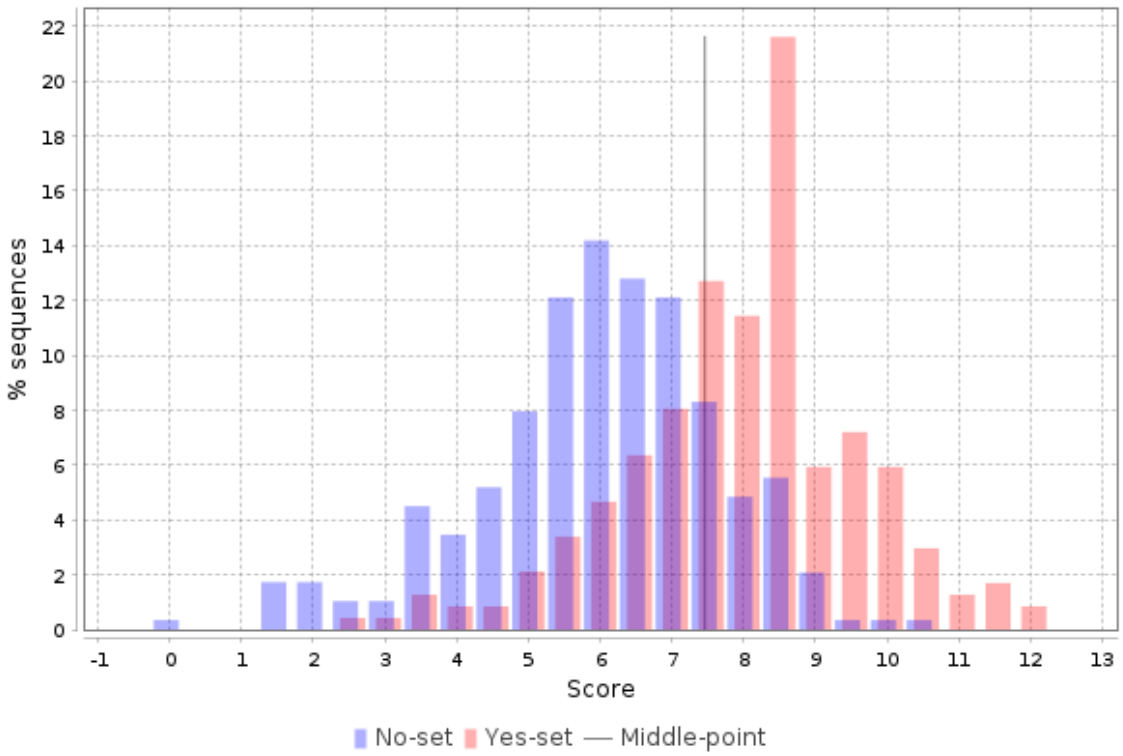
Module width: 120

Module 2:

V$FXR_01	V$EVI_Q6	V$IK_Q5	V$E2A_Q6_01	V$CDP_Q6_01	V$LEF1_Q4
0.92; N=2	0.78; N=3	0.95; N=3	0.97; N=2	0.00; N=3	0.98; N=1

Module width: 126

Model score (-p*log10(pval)): 16.29
Wilcoxon p-value (pval): 3.65e-34
Penalty (p): 0.487
Average yes-set score: 7.96
Average no-set score: 6.02
AUC: 0.81
Middle-point: 7.46
False-positive: 16.61%
False-negative: 32.20%



See model visualization table →

Table 3. List of top ten genes encoding enzymes, metabolising target metabolites in TNF vs NO_TNF 72h 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
ENSG00000135929	CYP27A1	cytochrome P450 family 27 subfamily A member 1	11.88	Evi-1(h), LEF-1(h), E2A(h), FXR(h), RAR-gamma(h), TCF-4(h), ctcf(h)...
ENSG00000138061	CYP1B1	cytochrome P450 family 1 subfamily B member 1	11.8	FXR(h), RAR-gamma(h), ctcf(h), Evi-1(h), CDP(h), E2A(h), Ikaros(h)
ENSG00000151093	OXSM	3-oxoacyl-ACP synthase, mitochondrial	11.61	RAR-gamma(h), TCF-4(h), FXR(h), ctcf(h), Evi-1(h), CDP(h), Ikaros(h)
ENSG00000171903	CYP4F11	cytochrome P450 family 4 subfamily F member 11	11.55	Ikaros(h), E2A(h), TCF-4(h), FXR(h), RAR-gamma(h), ctcf(h), Evi-1(h)
ENSG00000186115	CYP4F2	cytochrome P450 family 4 subfamily F member 2	11.53	Ikaros(h), FXR(h), E2A(h), Evi-1(h), CDP(h), RAR-gamma(h), POU6F1(h)...
ENSG00000142973	CYP4B1	cytochrome P450 family 4 subfamily B member 1	11.25	CDP(h), Evi-1(h), FXR(h), E2A(h), LEF-1(h), Ikaros(h), RAR-gamma(h)...
ENSG00000132423	COQ3	coenzyme Q3, methyltransferase	11.17	RAR-gamma(h), Ikaros(h), TCF-4(h), LEF-1(h), Evi-1(h), FXR(h), E2A(h)
ENSG00000158125	XDH	xanthine dehydrogenase	10.98	CDP(h), Ikaros(h), Evi-1(h), RAR-gamma(h), FXR(h), ctcf(h), TCF-4(h)
ENSG00000186529	CYP4F3	cytochrome P450 family 4 subfamily F member 3	10.98	LEF-1(h), FXR(h), RAR-gamma(h), TCF-4(h), ctcf(h), Ikaros(h), E2A(h)
ENSG00000198246	SLC29A3	solute carrier family 29 member 3	10.64	Ikaros(h), E2A(h), ctcf(h), RAR-gamma(h), FXR(h), TCF-4(h)

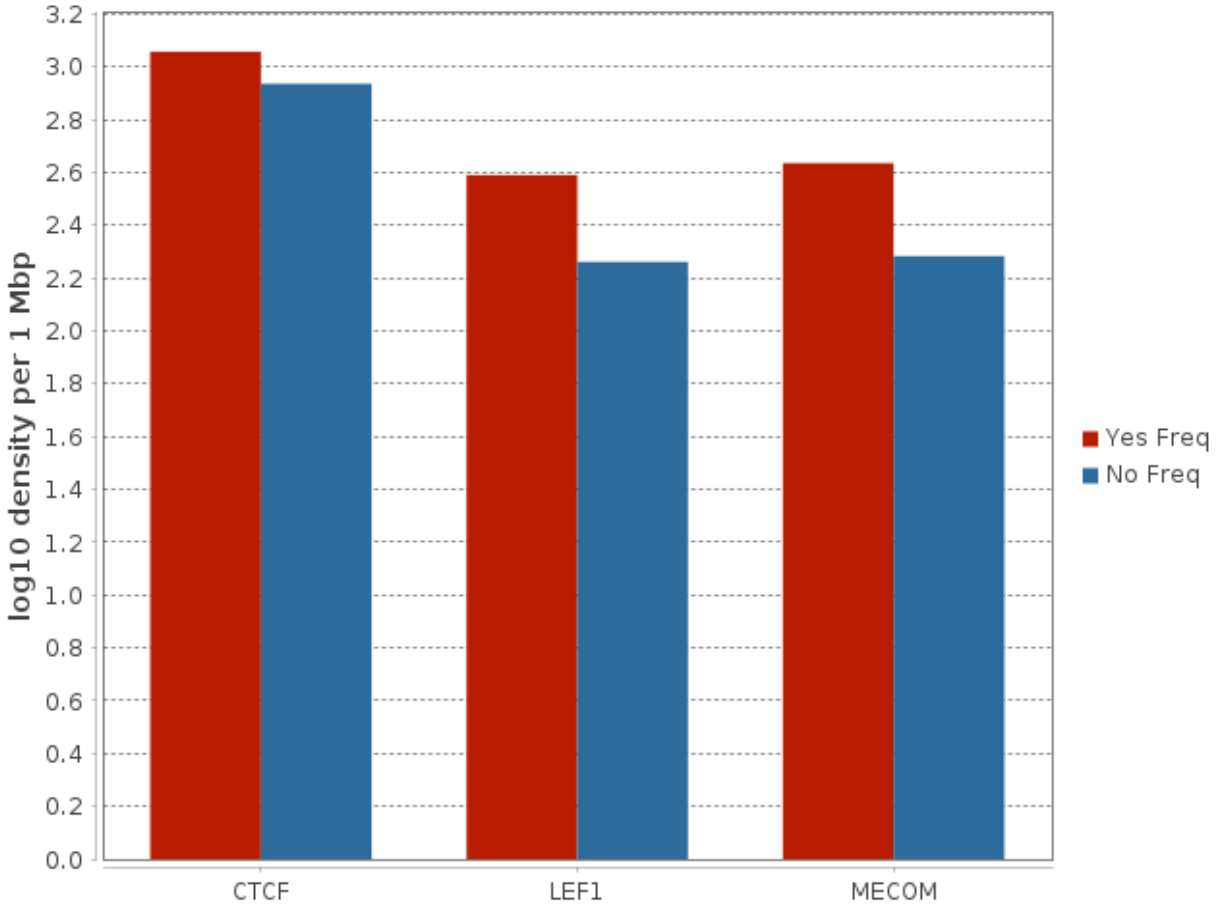
On the basis of the enhancer models we identified transcription factors potentially regulating the ***target genes*** of our interest. We found 10 transcription factors controlling expression of genes encoding enzymes, metabolising target metabolites (see Table 4).

Table 4. Transcription factors of the predicted enhancer model potentially regulating the genes encoding enzymes metabolizing given metabolites (genes encoding enzymes, metabolising target metabolites in TNF vs NO_TNF 72h). ***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
MO000046076	CTCF	CCCTC-binding factor	0.51	1.32
MO000159782	LEF1	lymphoid enhancer binding factor 1	0.37	2.13
MO000033253	MECOM	MDS1 and EVI1 complex locus	0.33	2.25
MO000032492	TCF3	transcription factor 3	0.3	1.38
MO000024708	CUX1	cut like homeobox 1	0.29	1.7
MO000026882	TCF7L2	transcription factor 7 like 2	0.27	1.62
MO000088742	NR1H4	nuclear receptor subfamily 1 group H member 4	0.27	1.71
MO000026678	IKZF1	IKAROS family zinc finger 1	0.24	1.4
MO000028320	null	null	0.24	12.25
MO000025130	RARG	retinoic acid receptor gamma	0	1.43

The following diagram represents the key transcription factors, which were predicted to be potentially regulating genes encoding enzymes metabolizing given metabolites in the analyzed pathology: CTCF, LEF1 and MECOM.



3.4. Finding master regulators in networks

In the second step of the upstream analysis common regulators of the revealed TFs were identified. These master regulators appear to be the key candidates for therapeutic targets as they have a master effect on regulation of intracellular pathways that activate the pathological process of our study. The identified master regulators are shown in Table 5.

Table 5. Master regulators that may govern the regulation of genes encoding enzymes, metabolising target metabolites in TNF vs NO_TNF 72h. ***Total rank*** *is the sum of the ranks of the master molecules sorted by keynode score, CMA score, metabolomics data.*

See full table →

ID	Master molecule name	Gene symbol	Gene description	Total rank
MO000081811	SUV39H1-isoform1(h)	SUV39H1	suppressor of variegation 3-9 homolog 1	3
MO000044865	setd7(h)	SETD7	SET domain containing 7, histone lysine methyltransferase	4
MO000081812	SUV39H1(h)	SUV39H1	suppressor of variegation 3-9 homolog 1	4

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 encoding enzymes metabolizing given metabolites, and selected master regulators, which are responsible for the regulation of these TFs.

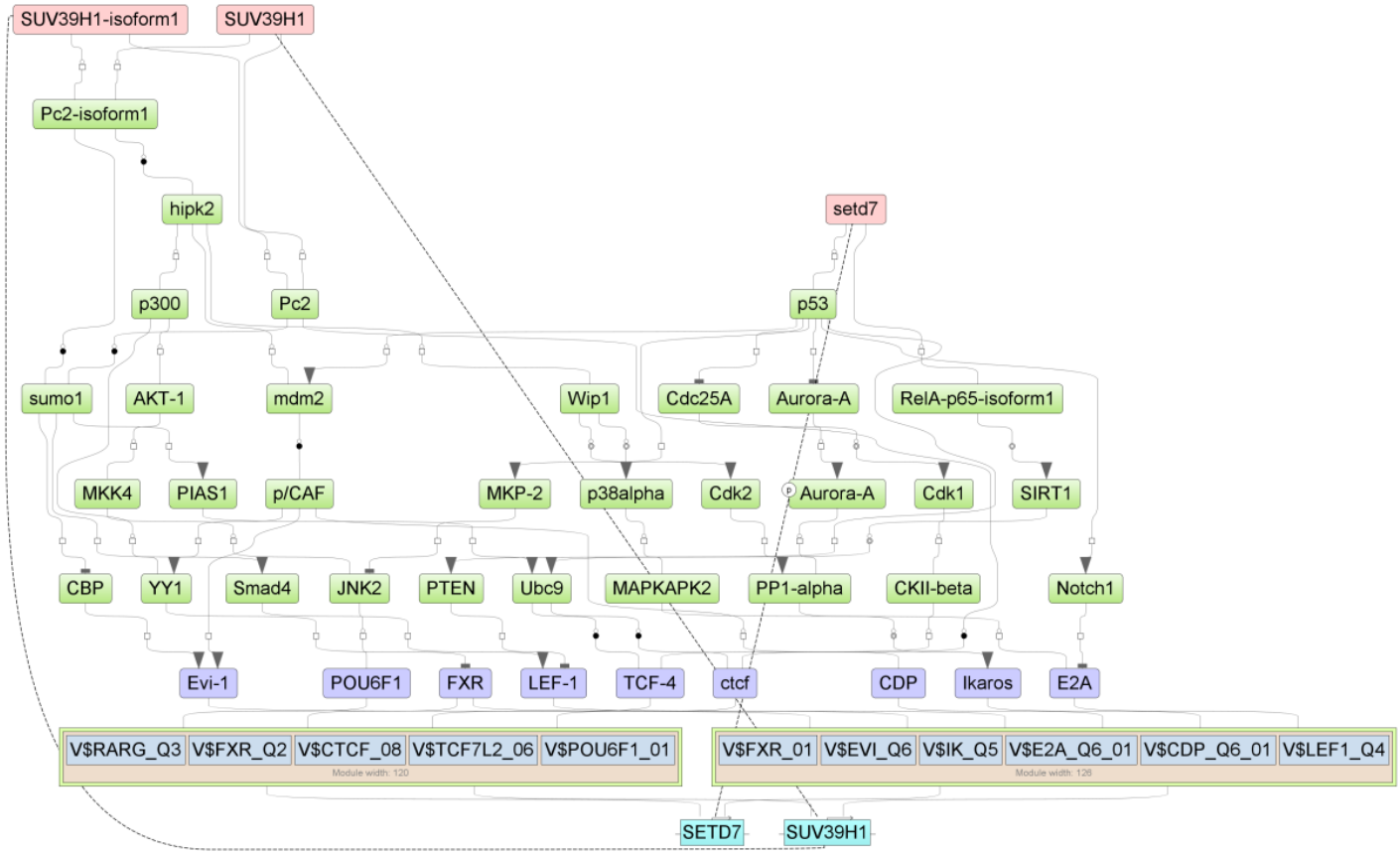


Figure 5. Diagram of intracellular regulatory signal transduction pathways of genes encoding enzymes, metabolising target metabolites in TNF vs NO_TNF 72h. Master regulators are indicated by red rectangles, transcription factors are blue rectangles, and green rectangles are intermediate molecules, which have been added to the network during the search for master regulators from selected TFs. Orange and blue frames highlight molecules that are encoded by up- and downregulated genes, resp.

See full diagram →

4. Finding prospective drug targets

The identified master regulators that may govern pathology associated genes were checked for druggability potential using HumanPSD™ [5] database of gene-disease-drug assignments and PASS [11-13] software for prediction of biological activities of chemical compounds on the basis of a (Q)SAR approach. Respectively, for each master regulator protein we have computed two druggability scores: HumanPSD druggability score and PASS druggability score. Where druggability score represents the number of drugs that are potentially suitable for inhibition (or activation) of the corresponding target either according to the information extracted from medical literature (from HumanPSD™ database) or according to cheminformatics predictions of compounds activity against the examined target (from PASS software).

The cheminformatics druggability check is done using a pre-computed database of spectra of biological activities of chemical compounds from a library of all small molecular drugs from HumanPSD™ database, 2507 pharmaceutically active known chemical compounds in total. The spectra of biological activities has been computed using the program PASS [11-13] on the basis of a (Q)SAR approach.

If both druggability scores were below defined thresholds (see Method section for the details) such master regulator proteins were not used in further analysis of drug prediction.

As a result we created the following two tables of prospective drug targets (top targets are shown here):

Table 6. Prospective drug targets selected from full list of identified master regulators filtered by druggability score from HumanPSD™ database. ***Druggability score*** *contains the number of drugs that are potentially suitable for inhibition (or activation) of the target. The drug targets are sorted according to the* ***Total rank*** *which is the sum of three ranks computed on the basis of the three scores: keynode score, CMA score and expression change score (logFC, if present). See Methods section for details.*
See full table →

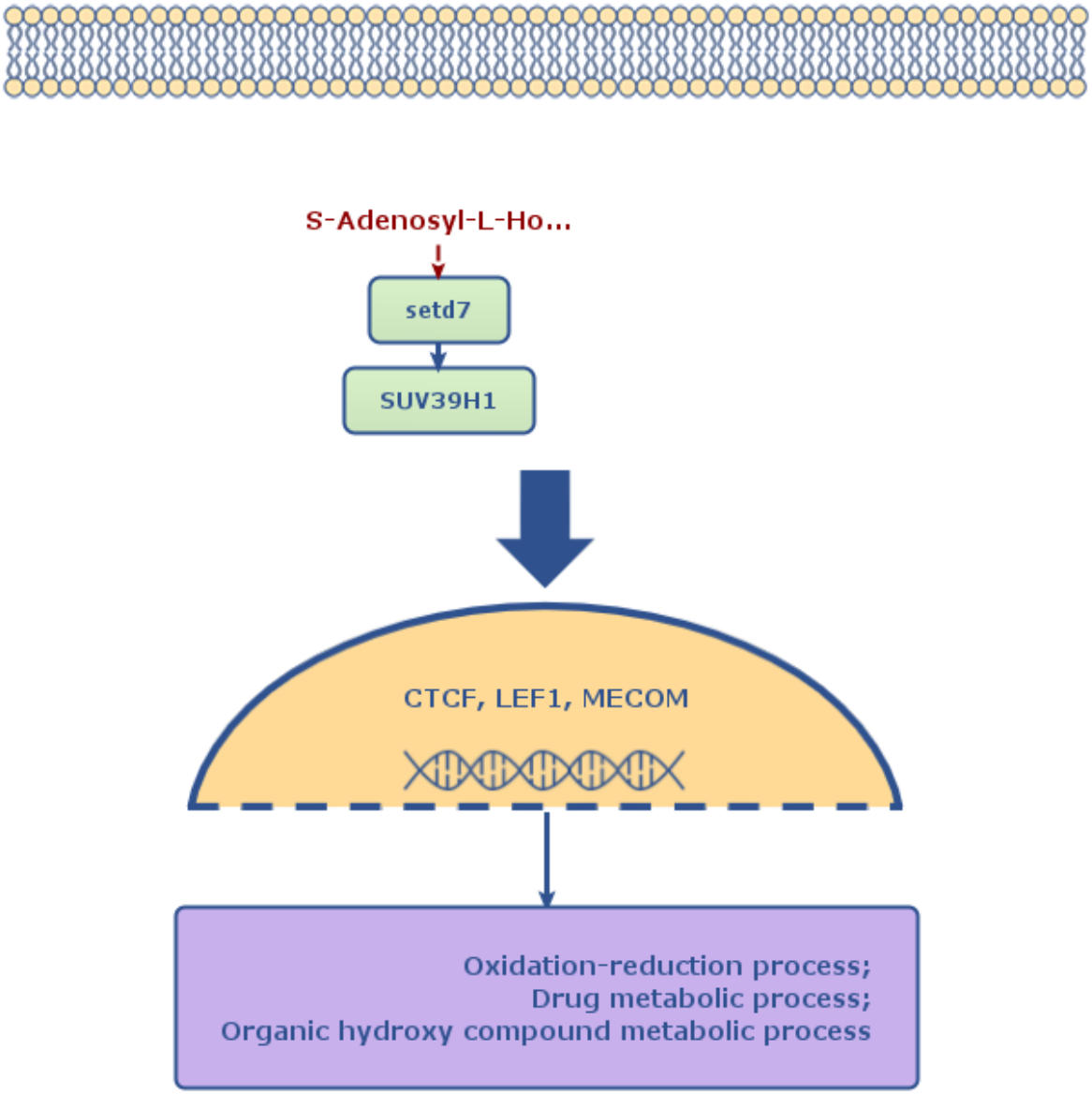
Gene symbol	Gene Description	Druggability score	Total rank
SETD7	SET domain containing 7, histone lysine methyltransferase	1	4

No prospective drug targets were found in the list of identified master regulators after filtering this list by druggability score predicted by PASS software.

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:

- setd7
- SUV39H1

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: S-Adenosyl-L-Homocysteine, should be considered as a prospective research initiative for further drug repurposing and drug development. These drugs were selected as top matching treatments to the most prospective drug targets of the studied pathology, however, these results should be considered with special caution and are to be used for research purposes only, as there is not enough clinical information for adapting these results towards immediate treatment of patients.
The drugs given in dark red color on the schema are FDA approved drugs or drugs which have gone through various phases of clinical trials as active treatments against the selected targets.
The drugs given in pink color on the schema are drugs, which were cheminformatically predicted to be active against the selected targets.

5. Identification of potential drugs

In the last step of the analysis we strived to identify known activities as well as drugs with cheminformatically predicted activities that are potentially suitable for inhibition (or activation) of the identified molecular targets in the context of specified human diseases(s).

Proposed drugs are top ranked drug candidates, that were found to be active on the identified targets and were selected from 4 categories:

1. FDA approved drugs or used in clinical trials drugs for the studied pathology;
2. Repurposing drugs used in clinical trials for other pathologies;
3. Drugs, predicted by PASS to be active against identified drug targets and against the studied pathology;
4. Drugs, predicted by PASS to be active against identified drug targets but for other pathologies.

Proposed drugs were selected on the basis of drug rank which was computed from two scores:

- target activity score (depends on ranks of all targets that were found for the selected drug);
- disease activity score (weighted sum of number of clinical trials on disease(s) under study where the selected drug is known to be applied or PASS disease activity score - cheminformatically predicted property of the compound to be active against the studied disease(s)).

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

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

No FDA approved drugs or drugs used in clinical trials for the studied disease(s) were found to be prospective drug candidates for treatment of the studied pathology.

Repurposing drugs

Table 7. Repurposed drugs used in clinical trials for other pathologies (prospective drugs against the identified drug targets on the basis of literature curation in HumanPSD™ database)
See full table →

Name	Target names	Drug rank	Phase 4	Status (provided by Drugbank)
S-Adenosyl-L-Homocysteine	SETD7	3		small molecule,experimental

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).

No prospective drugs were found, which would be predicted by PASS software to be active against the identified drug targets though without cheminformatically predicted activity against the studied disease(s).

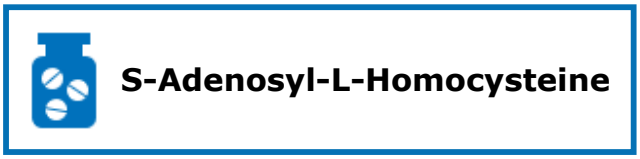
As the result of drug search we propose the following drugs as most promising candidates for treating the pathology under study: S-Adenosyl-L-Homocysteine. These drugs were selected for acting on the following targets: SETD7, which were predicted to be active in the molecular mechanism of the studied 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 *metabolomics* data. The study is done in the context of *Lung Neoplasms*. The data were pre-processed, statistically analyzed and genes encoding enzymes metabolizing given metabolites 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:



These drugs were selected for acting on the following targets: SETD7, 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:

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: S-Adenosyl-L-Homocysteine. 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 encoding enzymes metabolizing given metabolites 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.

- setd7
- SUV39H1

Potential drug compounds which can be affecting these targets can be found in the "Finding prospective drug targets" section.

7. Methods

Databases used in the study

Transcription factor binding sites in promoters and enhancers of differentially expressed genes were analyzed using known DNA-binding motifs described in the TRANSFAC® library, release 2020.2 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transfac).
The master regulator search uses the TRANSPATH® database (BIOBASE), release 2020.2 (geneXplain GmbH, Wolfenbüttel, Germany) (https://genexplain.com/transpath). A comprehensive signal transduction network of human cells is built by the software on the basis of reactions annotated in TRANSPATH®.
The information about drugs corresponding to identified drug targets and clinical trials references were extracted from HumanPSD™ database, release 2020.2 (https://genexplain.com/humanpsd).
The Ensembl database release Human99.38 (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 vcf file or computed by Genome Enhancer from fastq files), first of all, we compute a specific mutation weight (w) 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 = 0.7 for variations in exon area
w = 1.3 for variations in promoter region (-1000bp upstream and 100bp downstream of TSS),
w = 1.0 for variations in other locations.

Total Gene mutation weight is the sum of the weights w of all variations located inside the gene body and in the gene flanking regions.
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 = 1.5 for genes assigned to selected diseases,
In_transpath = 2.0 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) are also used to find the regulatory regions of the genes most affected by the variations. 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 promoters and enhancers under study as compared to a background sequence set such as promoters of genes that were not differentially regulated under the condition of the experiment. We denote study and background sets briefly as Yes and No sets. In the current work we used a workflow considering promoter sequences of a standard length of 1100 bp (-1000 to +100). The error rate in this part of the pipeline is controlled by estimating the adjusted p-value (using the Benjamini-Hochberg procedure) in comparison to the TFBS frequency found in randomly selected regions of the human genome (adj.p-value < 0.01).
We have applied the CMA algorithm (Composite Module Analyst) for searching composite modules [7] in the promoters and enhancers of the Yes and No sets. We searched for a composite module consisting of a cluster of 10 TFs in a sliding window of 200-300 bp that statistically significantly separates sequences in the Yes and No sets (minimizing Wilcoxon p-value).

Methods for finding master regulators in networks

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

Methods for analysis of pharmaceutical compounds

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

Method for analysis of known pharmaceutical compounds

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

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

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

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

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

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

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

Method for prediction of pharmaceutical compounds

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

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

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

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

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

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

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

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

8. References

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

cytotoxic action of p53 activator Nutlin-3. *BMC Med Genomics.* **2018**;11(1):12. doi:10.1186/1471-2105-7-s2-s13
1. Filimonov D, Poroikov V. Probabilistic Approaches in Activity Prediction. Varnek A, Tropsha A. *Cheminformatics Approaches to Virtual Screening.* Cambridge (UK): RSC Publishing. **2008**;:182-216.
2. Filimonov DA, Poroikov VV. Prognosis of specters of biological activity of organic molecules. *Russian chemical journal.* **2006**;50(2):66-75 (russ)
3. Filimonov D, Poroikov V, Borodina Y, Gloriozova T. Chemical Similarity Assessment Through Multilevel Neighborhoods of Atoms: Definition and Comparison with the Other Descriptors. *ChemInform.* **1999**;39(4):666-670. doi:10.1002/chin.199940210

Thank you for using the Genome Enhancer!

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

Supplementary material

1. Supplementary table 1 - Detailed report. Composite modules and master regulators (genes encoding enzymes, metabolising target metabolites in TNF vs NO_TNF 72h).

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.