Putative disease gene identification and drug repurposing for Steatohepatitis

Milad Torabi, Possenti Francesca Group 9
Sapienza University, Bioinformatics and Network Medicine

January 16, 2025

Abstract

Non-alcoholic steatohepatitis (NASH) is a complex liver disorder characterized by the accumulation of fat in the liver, inflammation, and hepatocellular injury. As the prevalence of NASH continues to rise globally, identifying putative disease genes and exploring potential drug repurposing strategies are critical for advancing therapeutic interventions. This report presents an analysis of the genetics behind NASH, leveraging network science methods and algorithms (DIAMOnD, DIAble, Diffusion) to identify putative disease-associated genes in the context of a PPI network.

Our approach involves integrative analysis of multi-omics data, including interactomics - proteomics and genomics, to determine key genetic factors contributing to NASH pathogenesis. Through the identification of new disease-associated genes, we aim to understand the mechanisms involved in disease development and progression.

Furthermore, the study extends its focus to drug repurposing, capitalizing on already known therapies to simplify the drug development process. By screening approved drugs against the identified putative disease genes and pathways, we seek to uncover potential therapeutic candidates that may be efficacious in mitigating NASH-related pathology.

The findings of this report might be helpful in advancing our understanding of the molecular landscape of NASH and providing novel insights into potential therapeutic avenues. The integration of bioinformatics analyses with drug repurposing strategies offers a comprehensive approach to accelerate the development of targeted therapies for NASH, addressing the urgent need for effective treatments in the face of the growing global burden of liver diseases.

1 Introduction

1.1 Background

Non-alcoholic steatohepatitis (NASH) represents a significant and increasingly prevalent form of liver disease characterized by hepatic steatosis, inflammation, and hepatocellular injury. Unlike simple fatty liver disease, NASH involves inflammation and cellular

damage, making it a more severe and potentially progressive condition. This liver disorder has emerged as a major public health concern, closely associated with the rising global incidence of obesity, metabolic syndrome, and type 2 diabetes ([PC20]).

NASH poses a multifaceted challenge, encompassing diverse factors such as genetic predisposition, lifestyle choices, and metabolic abnormalities. Its insidious progression often leads to advanced liver fibrosis, cirrhosis, and an elevated risk of hepatocellular carcinoma. The intricate interplay of various molecular and cellular events underlying NASH remains a subject of intense research, as elucidating these mechanisms is crucial for the development of targeted therapeutic interventions.

In this context, understanding the pathophysiology of NASH is essential for clinicians, researchers, and policymakers alike. As the burden of liver diseases continues to escalate globally, it is evermore necessary to identify therapeutic targets, and potentially repurpose existing drugs to mitigate disease progression. This introduction sets the stage for a comprehensive exploration into the multifaceted landscape of NASH, emphasizing the urgency of addressing this silent epidemic through innovative research and therapeutic strategies.

2 Materials and Methods

2.1 Data

The analysis was performed leveraging on two different DBs, BioGRID and DisGeNET.

To build the human interactome, the archive of files related to all organisms was downloaded from the BioGRID latest releases section. Among all the organisms files, the one associated to the organism Homo-sapiens was selected. Each line in the file represents an interaction between two proteins/genes. Initially, 1203844 proteins interactions were present in the human file as rows, each described by 37 columns, containing information about the interaction. Out of those 37 columns, only a few were considered for the analysis.

As regards the gene-disease associations (GDAs), the data related to Steatohepatitis (code : C2711227) were gathered from DisGeNET database following the manual procedure from the website to download the file. The data consisted in 88 genes, each described by 18 columns.

The Python programming language has been chosen to perform the tasks of the project.

2.2 The PPI network and its LCC

To build the PPI network from the interactome, the raw data from the BioGRID DB was analysed and filtered. A data-frame was created from the downloaded file, where each row contained data related to the interaction it encodes. The first step was to build the PPI network, where each node is represented by a protein as protein's ID, while an edge is defined when two proteins are connected by an interaction row in the data-frame. The data were filtered to isolate the human physical interactions: only the physical interactions (the ones observed during a physical experiment) where both interactors are humans were kept. After the filtering, 825997 interactions remained.

The PPI network was built using the NetworkX Python library : nodes were defined as proteins and edges as interactions. Self loops and duplicates were removed, obtaining 822762 final edges and 19838 nodes. From the graph, the connected components were extracted to isolate the LCC, resulting in 6 CCs: a LCC of 822762 edges and 19833 nodes and 5 others CCs of 5 isolated nodes (the ones involved only in self loops).

The data exploration was performed using the Pandas library.

2.3 The disease LCC

To build the disease LCC, the list of known disease genes associated to NASH was extracted from the DisGeNET data. The intersection between the disease genes and the genes in the (human, physical) interactome was computed. Later, the interactome was filtered to maintain only the interactions involving disease genes. As a result, a network was built using the NetworkX library: the output was a disease graph with 84 edges and 46 nodes. After removing self-loops, in the disease graph 61 edges and 46 nodes remained. From this network, the LCC was extracted, getting a graph with 59 edges and 40 nodes. Some metrics were computed with respect to the graph, as shown in Table 1. In particular, the analysis was performed to check if the disease LCC was a topological module: its internal density was compared to the external density of the PPI graph, showing that the sub-graph is somehow isolated from the rest of the graph, as indicated by the lower density of connections with external nodes (internal density: 0.075641 , external density: 0.004198).

Graph	PPI	LCC	Disease LCC
Edgecount	809943	809943	40
Size	19823	19818	59
Density	0.004122	0.004125	0.07564
Clustercoefficient	0.130564	0.130597	0.252253

Table 1: Graphs Metrics

The following metrics were computed on the LCC Disease Network:

- Node degree;
- Betweenness centrality;
- Eigenvector centrality;
- Closeness Centrality;
- Ratio Betweenness/Node Degree.

A scatter-plot was created to analyze the relationship between Node degree and Betweenness centrality.The Node Degree is set on the x-axis, ranging from 0 to 14. Such low degrees, if compared to the overall degrees of each node, show how the disease LCC is a very small sub-graph in a much bigger context (see 'Node degree' and 'Node degree in LCC-ppi' columns in Figure 4). The Betweenness Centrality, set on the y-axis and ranging from 0 to 0.6, measures the extent to which a vertex lies on paths

between other vertices. It was computed with respect to the disease LCC only. The color intensity of each point corresponds to its betweenness value, with darker colors representing lower values and lighter colors representing higher values.

From the scatter-plot in Figure 1, it is possible to observe that:

- most points are clustered towards the lower end of both axes, indicating nodes with low degree and low betweenness centrality: these nodes might be peripheral or less influential in the network;
- there is one point near (14,0.6) that stands out for having both high node degree and high betweenness centrality: this node, with ID 4609, is a hub and highly influential in the network, as it has many connections and also acts as a bridge in many shortest paths.

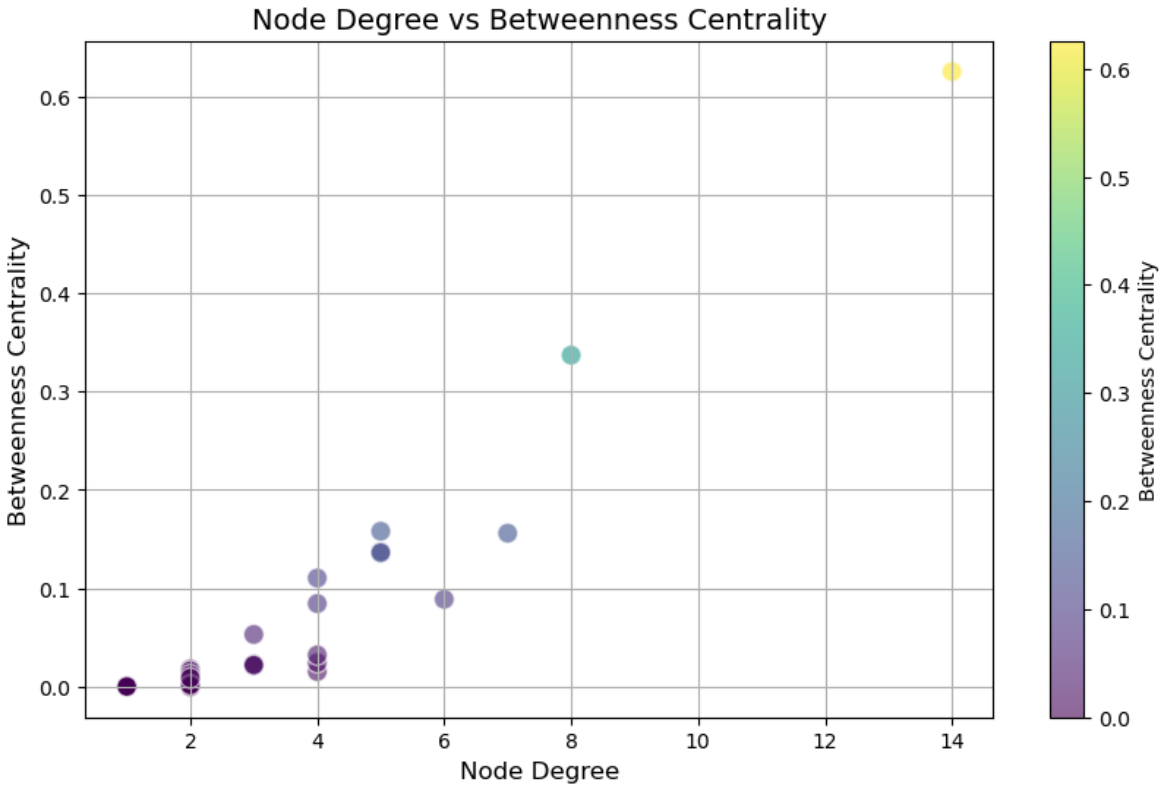

Figure 1: Node Degree vs Betweenness Centrality in LCC Disease Network

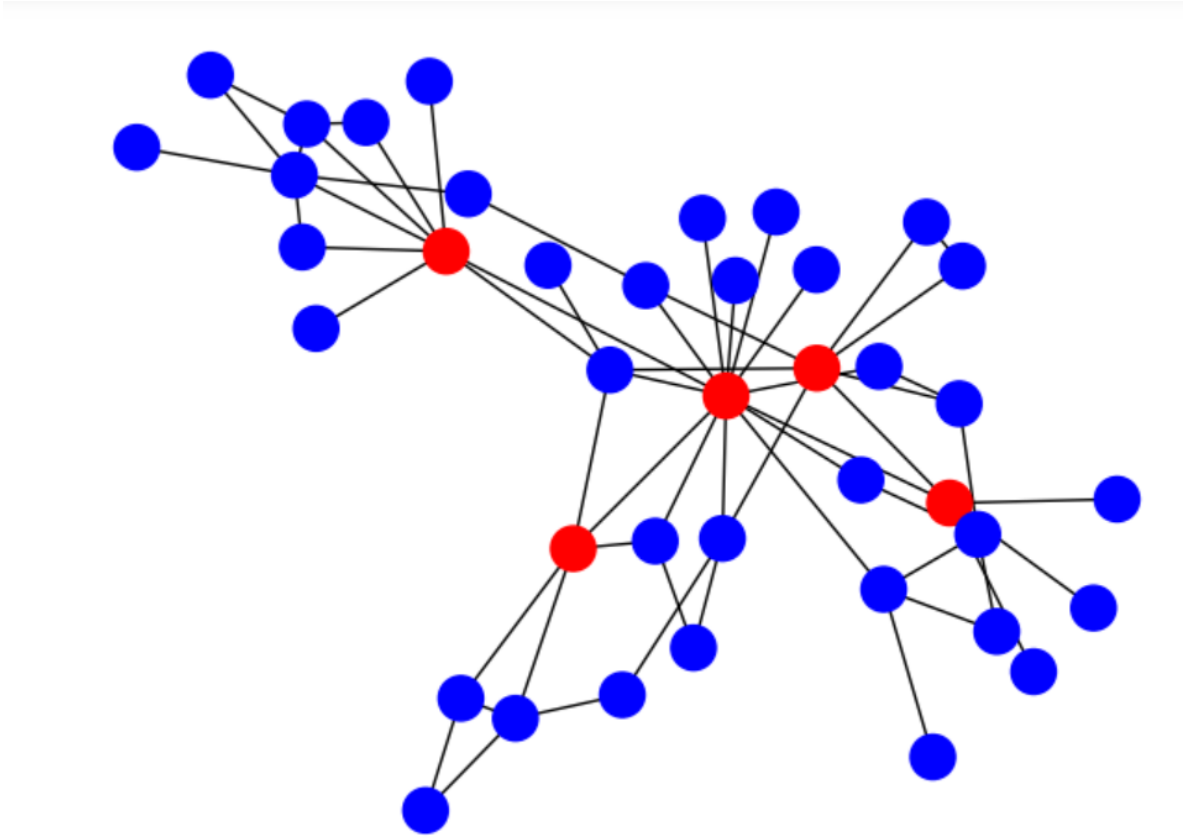

Figure 2: LCC Disease Network with top 5 central nodes is red

	disease name	UMLS disease ID	MeSH disease class	number of associated genes	number of genes present in the interactome	LCC size of the disease interactome
0	Steatohepatitis	C2711227	C02	88	76	40

Figure 3: Disease LCC Summary

	Gene	Node Degree	Node Degree in LCC_ppi	Betweenness Centrality	Eigenvector Centrality	Closeness Centrality	Ratio Betweenness/Node Degree
2	4609	14	2941	0.625686	0.537770	0.565217	0.044692
13	23411	8	379	0.337112	0.309770	0.458824	0.042139
12	330	7	1321	0.155960	0.234980	0.410526	0.022280
9	8431	6	66	0.088754	0.150694	0.342105	0.014792
0	338	5	130	0.157962	0.191828	0.414894	0.031592
29	54205	5	156	0.136100	0.300792	0.458824	0.027220
7	847	5	152	0.137157	0.257971	0.419355	0.027431
17	4780	4	144	0.110301	0.153860	0.390000	0.027575
6	1649	4	117	0.015250	0.094697	0.309524	0.003812
25	6647	4	403	0.032074	0.094665	0.312000	0.008018
10	10062	4	51	0.024157	0.131780	0.333333	0.006039
21	3032	4	422	0.084368	0.196411	0.414894	0.021092
33	5728	3	630	0.021682	0.187991	0.386139	0.007227
26	6648	3	87	0.022582	0.083675	0.307087	0.007527
20	5686	3	324	0.052969	0.180636	0.410526	0.017656
4	2147	3	35	0.021368	0.099811	0.339130	0.007123
15	7494	2	44	0.011966	0.070940	0.327731	0.005983
34	3630	2	506	0.001799	0.082302	0.304688	0.000900
3	1051	2	108	0.017859	0.135413	0.378641	0.008929
28	468	2	87	0.000000	0.053217	0.295455	0.000000

Figure 4: Disease LCC top-20 nodes metrics

2.4 Comparative analysis of the disease genes identification algorithms

Out of 88 disease genes, 76 genes were found in the interactome. These genes were used as seed genes input for the disease genes identification algorithms:

- DIAMOnD algorithm was run using the module downloaded from https://github.com/dinaghiassian/DIAMOnD;
- DIAble algorithm was run leveraging on the DIAMOnD module and manually modifying as necessary the above mentioned module;
- Cytoscape Diffusion algorithm was run directly connecting to the Diffusion service at http://v3.heat-diffusion.cytoscape.io, following the tutorial at https://github.com/idekerlab/heat-diffusion/blob/master/README.md via POST request. The algorithm was run three times with different time parameters set to 0.1, 0.002 and 0.005, passed in the post request as query parameters. The request payload was build using the ndex library to create a Nice CX object from Networkx graph. The request was performed using the requests library.

To select the best algorithm, a 4-Fold Cross Validation was performed, choosing a smaller number for the fold since the available seed genes were only 76. The algorithms were run multiple times to get the performance measures selecting the top 50, 76/10, 76/4, 76/2 and 76 putative genes. Based on the results in Figure 5, the DIAble algorithm was chosen for the next steps for its better performances. In general, it is supposed that the poor performances are due to the fact that the graph has a quite high average clustering coefficient. The size of the universe for each iteration of the DIAble algorithm is always almost identical to the size of the entire LCC-PPI graph. This suggests that the seed genes are well integrated in the context of the whole graph, making it harder to identify new putative disease genes.

3 Results and Discussion

3.1 Enrichment Analysis

Enrichment analysis is a powerful tool since it decodes the functional context of gene sets and helps understanding the functional interconnections among genes.

Firstly, an enrichment analysis was performed on the Putative Disease Genes. To begin, protein IDs (used as nodes' labels) were converted into genes IDs and mapped with their symbols using the service MyGene.info at http://mygene.info/v3/gene, which provides simple-to-use REST web services to query/retrieve gene annotation data. This analysis aimed to identify over-represented biological themes, including pathways and processes, among the putative disease genes. The top 5 enriched terms based on their adjusted P-values were visualized using a bar plot. For the enrichment function application a cutoff of 0.05 was set, resulting in fewer but more significant enriched terms. As shown in Figure 6 each term on the y-axis has a corresponding bar indicating its adjusted P-value.

The terms "Hepatocellular carcinoma" and "Transcriptional misregulation in cancer" had very low Adjusted P-values, indicating a strong association with the putative

	precision	recall	f1_score	source
7	0.14286	0.05263	0.07692	diamond
7	0.14286	0.05263	0.07692	diable
19	0.08985	0.08985	0.26653	diamond
19	0.08985	0.08985	0.26653	diable
38	0.06890	0.13779	0.26669	diamond
38	0.06890	0.13779	0.26669	diable
50	0.05414	0.14248	0.07847	diamond
50	0.05414	0.14248	0.07847	diable
76	0.04394	0.17577	0.07031	diamond
76	0.03562	0.14248	0.05699	diable
76	0.01557	0.06226	0.20792	cyto005
76	0.01316	0.05263	0.02105	cyto002
76	0.01316	0.05263	0.02105	cyto01

Figure 5: 4-Fold Cross Validation for all algorithms

disease genes. It could mean that the genes associated with these terms are more likely to be found in our set of putative disease genes than would be expected by chance. This could suggest a potential functional relationship or shared biological pathways. The x-axis limits were adjusted to ensure all bars were visible on the plot.

While on the first visual representation the focus is strictly on the adjusted P-values, now also the Combined Values are taken into account, showing how a term is relevant to the gene set. In the scatter-plot in Figure 7 most data points are clustered near the bottom of the plot, indicating low combined scores across various adjusted p-values. This suggests that while many terms may be statistically significant (low p-value), their overall relevance to the gene set (combined score) may be low.

However, there is one outlier data point with a high combined score and low adjusted p-value. This indicates a term that is both statistically significant and highly relevant to the gene set. This term has been of particular interest for further investigation.

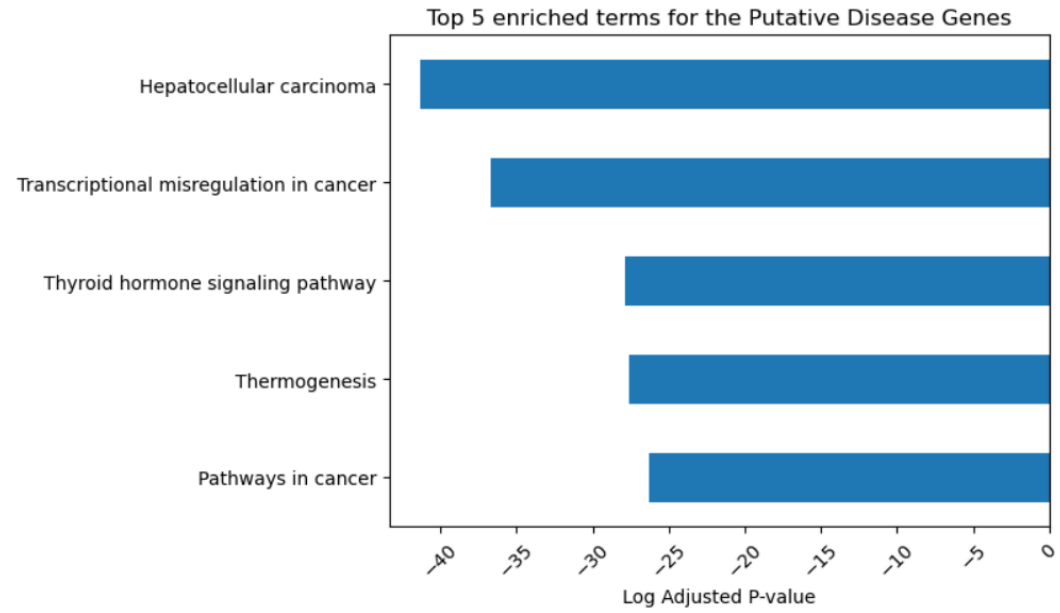

Figure 6: Enrichment analysis for Putative Genes.

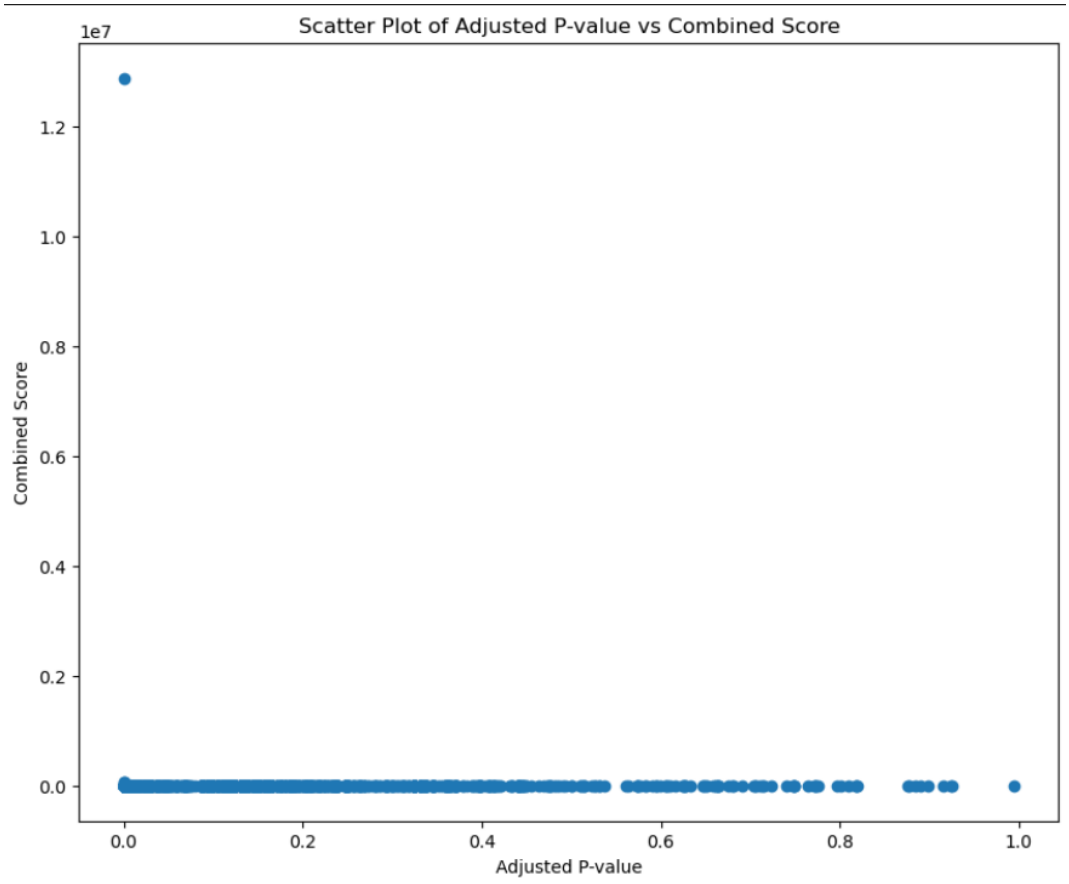

Figure 7: Adjusted p-value vs Combined Score for Putative Genes

As can be seen in Table 2, for the term npBAF complex (GO:0071564) from the

"GO_Cellular_Component_2018" category, the "Overlap" value of "11/11" indicates that all 11 genes in this gene set were found in our data. The "Adjusted P-value" and "Combined Score" are statistical measures used to assess the significance of this overlap. The extremely low "Adjusted P-value" (3.039838e-25) suggests that the overlap is statistically significant, and not likely due to random chance. The high "Combined Score" (1.288242e+07) also indicates a strong association.

Term	Overlap	Adj P-value	Combined Score
npBAF complex (GO:0071564)	11/11	3.039838e-25	1.288242e+07

Table 2: Outlier term highlight in the case of Putative Disease Genes

As stated in [npB], the npBAF complex is a type of SWI/SNF-type complex found in neural progenitor cells. While the direct role of the npBAF complex in NASH is not clear from the current literature, it's known that the SWI/SNF complexes are involved in chromatin remodeling (see [PPM08]), which is a crucial process in the regulation of gene expression.

Given that NASH involves changes in the expression of many genes related to inflammation, lipid metabolism, and fibrosis ([DC19]), it's plausible that a complex involved in gene regulation could play a role in this disease. The significant representation of the npBAF complex in the data could suggests that the genes in this complex are being differentially expressed in NASH, potentially contributing to the disease process. However, this is a hypothesis that would need to be validated with further experiments.

The second part of the Enrichment Analysis was performed for the Original Disease Genes. As in Figure 8, the results show some big differences with respect to the terms associated to the set of putative genes.

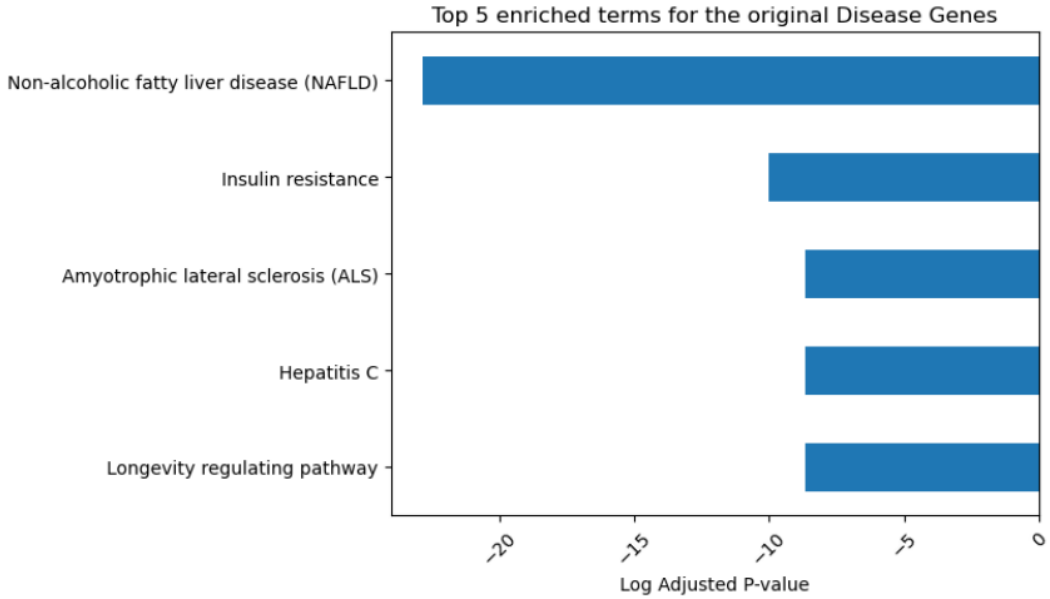

Figure 8: Enrichment analysis for known Disease Genes

While the known disease genes are associated with metabolic and neurological conditions (such as NAFLD, Insulin resistance, ALS, Hepatitis C, and Longevity regulating

pathway), the putative genes identified by DIAble are more associated with cancer and hormonal signaling pathways (such as Hepatocellular carcinoma, Transcriptional misregulation in cancer and thyroid hormone signaling pathway). This results could be uncovering potential new associations not previously considered or it may indicate a bias towards certain types of diseases or pathways (since it considers as gene universe the smallest local expansion of the current seeds set at each iteration step).

From Figure 9, it is observed that there are several data points with high combined scores and low adjusted p-values. These could be considered as outliers as they deviate significantly from the majority of the data points.

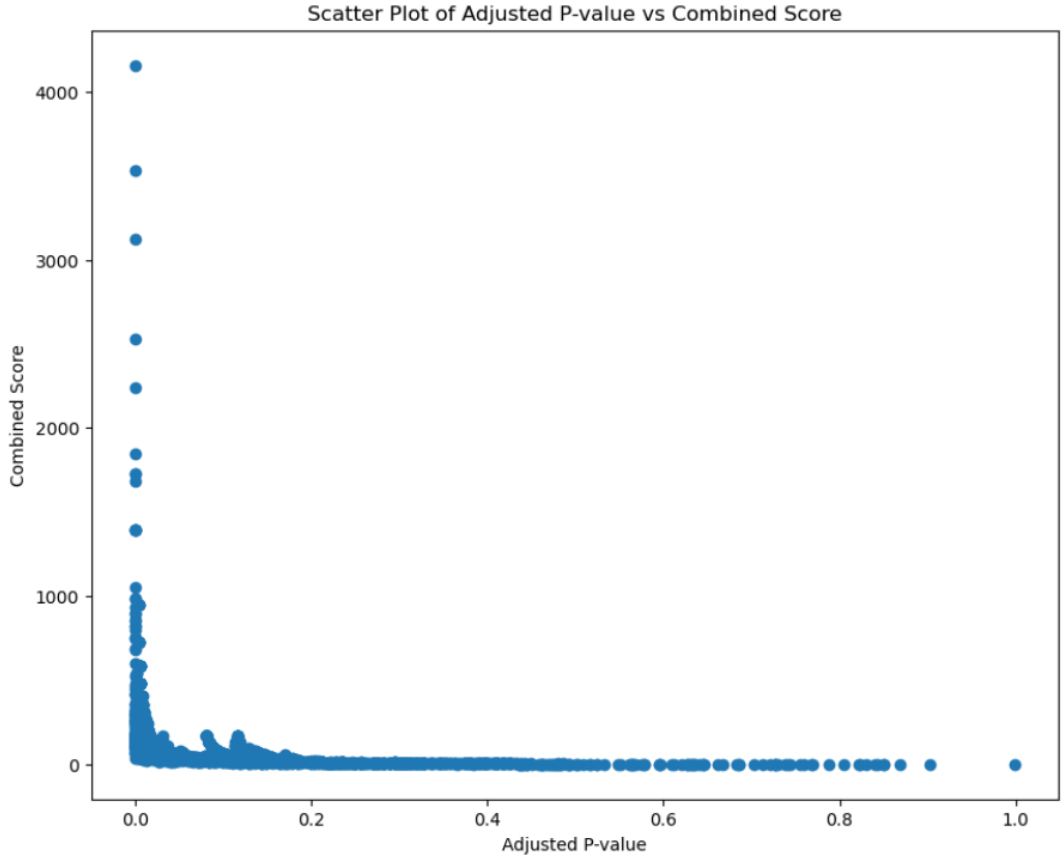

Figure 9: Adjusted p-value vs Combined Score for the known Disease Genes

Term	Overlap	Adj P-value	Combined Score
PERK-mediated unfolded protein response	5/12	2.169719e-07	3528.044
response to leptin	4/8	2.637012e-06	4156.629
cellular triglyceride homeostasis	3/6	6.556150e-05	3121.415
cellular response to leptin stimulus	3/7	1.059995e-04	2243.201
RNA polymerase II transcription factor	6/19	1.438670e-08	2527.638

Table 3: Outliers in the case of known Disease Genes

These associations could provide valuable insights into the biological processes and molecular functions related to the diseases under study combined with the one of the

putative disease genes. However, further validation and investigation would be needed to confirm these findings and interpret the outliers.

For the overlap evaluation, different methods have been adopted:

- Sets intersection: from the intersection of the Enrichment results 693 common terms between the putative and original disease genes were obtained. However, this would only give a rough idea of the similarity, as it does not take into account the statistical significance of the overlap.

- Fisher's exact Test: for a more rigorous comparison, the Fisher's exact test was used to compute the p-value for the difference in overlap between the two sets, resulting to be 6.7074e-10. Being extremely small and less than 0.05, it indicates that the overlap between the enriched functions of original disease genes and putative disease genes is statistically significant. However, while a low p-value indicates statistical significance, it does not measure the size of the effect or the importance of a result.

- Jaccard Similarity: defined as the size of the intersection divided by the size of the union of the sets, the Jaccard Index was computed to compare the similarity and the diversity of the sample sets. The result, equal to 0.27018, indicates that approximately 27% of the total unique enriched functions in both the original disease genes and putative disease genes are common to both sets. This suggests a modest overlap between the two sets of enriched functions.

3.2 Drug repurposing

For the drugs identification, the interactions file from `https://dgidb.org/downloads` was downloaded and used. This file was transformed in a data frame and a filtering procedure (based on the top-20 putative genes obtained in point 3.1) on the rows was performed.

After getting all the interactions, a top-3 ranking was obtained by starting from the drugs mostly associated with the 20 genes.

Finally, for the three drugs in Table 5, a validation procedure was performed by searching at `https://www.clinicaltrials.gov` for any clinical trial that tested them for the disease. The results show that TRETINOIN has been tested, as stated in [soh].

Gene name	Drug name
CREBBP	ROLIPRAM
CREBBP	ETAZOLATE
CREBBP	TRIAZOLAM
CREBBP	CHEMBL1530911
CREBBP	ESTAZOLAM
CREBBP	CHEMBL1797712
CREBBP	PAPAVERINE
CREBBP	CHEMBL257748
CREBBP	PRI-724
CREBBP	NOCODAZOLE
CREBBP	CHEMBL1797707
CREBBP	CHEMBL1797708
CREBBP	CHEMBL1797711
CREBBP	MIDAZOLAM
CREBBP	CHEMBL1797713
CREBBP	TRACAZOLATE
CREBBP	ISCHEMIN
CREBBP	COLCHICINE
CREBBP	ALPRAZOLAM
CREBBP	OXOGLAUCINE

Table 4: Interactions between gene CREBBP and drugs

Drug	Number of occurencies
ACITRETIN	7
ALITRETINOIN	7
TRETINOIN	7

Table 5: Top 3 of the ranking

References

[DC19] Chiappini F. Desterke C. Lipid related genes altered in nash connect inflammation in liver pathogenesis progression to hcc: A canonical pathway. Int J Mol Sci., 2019.

[npB]

[PC20] Woodman OL Ritchie RH Qin CX. Peng C, Stewart AG. Non-alcoholic steatohepatitis: A review of its mechanism, models and medical treatments. Front Pharmacol., 2020.

[PPM08] Montse Sanchez-Cespedes Pedro P Medina. Involvement of the chromatin-remodeling factor brg1/smarca4 in human cancer. Epigenetics, doi: 10.4161/epi.3.2.6153. Epub, 2008.

[soh] Effect of vitamin a and calcium in patients with non-alcoholic fatty liver disease.