**Using PPI network embeddings to identify functional modules associated with pathogenesis of Nonalcoholic Steatohepatitis**

**Methods:**

1. **Map genes to functional module annotations, vector embeddings, and clusters**

Embeddings: created using [node2vec](https://snap.stanford.edu/node2vec/) and high confidence (> 800) STRING PPI edges

* 14,708 genes embedded
* Node2vec hyperparameters: length of walks=30, number of walks=10, min count=1, batch word=6, window=10

Clusters: assigned using k-means, 160 total

Functional module annotations:

|  |  |  |
| --- | --- | --- |
| *Label* | *Module type* | *Source* |
| ME | Immune response modules | [ImmProt](https://pubmed.ncbi.nlm.nih.gov/28263321/) |
| HM | Hallmark signaling pathways | MSigDB |
| MS | Metabolic subsystems | [Human GEM](https://metabolicatlas.org/) |
|  | Disease | DisGeNET – curated list |

1. **Create module vectors:** sum of the vector embeddings of all genes included in a module
2. **Annotate modules with NASH SNPs, NASH drug targets, PGC1 regulated pathways**
3. **Calculate cosine similarity of module vectors:** created a lookup table of cosine similarity between all functional modules and the NASH disease module using *sklearn.pairwise.cosine\_similarity*
4. **Generate a p-value for each cosine similarity**

Example:

Module m1 with n1 genes

Module m2 with n2 genes

Sreal = cosine similarity of m1 to m2

To calculate p-value:

Randomly sample 1000 subsets of n1 genes from the set of all genes

Sum subsets to form “random module vectors”

Calculate cosine similarity Srandom of m2 to each “random module vector”

p-value = proportion Srandom’s that are greater or equal to Sreal

1. **Calculate overlap in genes between modules**
2. **Generate hypergeometric p-values for enrichment of each module in each cluster**

Used *scipy.stats.hypergeom.sf(x – 1, M, n, N)*

x = module – cluster overlap

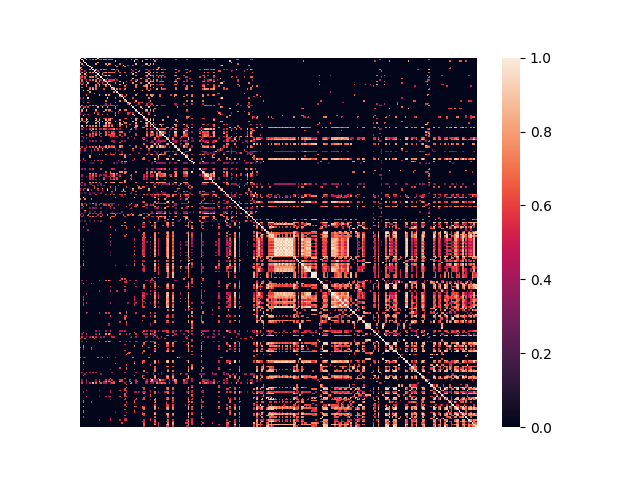
M = total # of genes (14,708)

n = # of genes in cluster

N = # of genes in module

**Figures/Tables:**

**Figure 1:** Heatmaps of module-module cosine similarities and gene overlap.

****

**A picture containing computer, monitor, sitting, table

Description automatically generated**

**Module – module overlap**

**Module – module cosine similarity**\*

\* values with p > .05 are set to 0

**Figure 2:** Graph representation of functional modules and their relation to NASH. Graph vertices represent functional modules and edges represent cosine *distance* between modules. Modules with a path length < .35 from NASH are included in the diagram below. A threshold of .35 was chosen for visibility purposes).

A picture containing light

Description automatically generated

**Figure 3:** Top modules by similarity to NASH

**\*\*p < .01**

**Table 1:** NASH enriched clusters and modules that are enriched in the same cluster (hypergeometric p-value < .01)

|  |  |
| --- | --- |
| Cluster | Modules |
| 26 | BILE\_ACID\_METABOLISM, PEROXISOME, FATTY\_ACID\_METABOLISM, fatty\_acid\_oxidation, pool\_reactions, UV\_RESPONSE\_DN |
| 54 | arachidonic\_acid\_metabolism, prostaglandin\_biosynthesis, metabolism\_of\_xenobiotics\_by\_cytochrome\_p450, leukotriene\_metabolism, estrogen\_metabolism, linoleate\_metabolism, eicosanoid\_metabolism, XENOBIOTIC\_METABOLISM, omega\_3\_fatty\_acid\_metabolism, glutathione\_metabolism, retinol\_metabolism, vitamin\_b6\_metabolism, FATTY\_ACID\_METABOLISM, vitamin\_e\_metabolism, drug\_metabolism, xenobiotics\_metabolism |
| 61 | IL6\_JAK\_STAT3\_SIGNALING, ALLOGRAFT\_REJECTION, INFLAMMATORY\_RESPONSE, INTERFERON\_GAMMA\_RESPONSE, IL2\_STAT5\_SIGNALING, TNFA\_SIGNALING\_VIA\_NFKB, ME 29 |
| 155 | retinol\_metabolism, tyrosine\_metabolism, metabolism\_of\_xenobiotics\_by\_cytochrome\_p450, omega\_6\_fatty\_acid\_metabolism, bile\_acid\_biosynthesis, omega\_3\_fatty\_acid\_metabolism, miscellaneous, leukotriene\_metabolism, XENOBIOTIC\_METABOLISM, PEROXISOME, FATTY\_ACID\_METABOLISM, estrogen\_metabolism, BILE\_ACID\_METABOLISM, fatty\_acid\_oxidation, ME 46, tryptophan\_metabolism |
| 101 | protein\_modification, PI3K\_AKT\_MTOR\_SIGNALING, TNFA\_SIGNALING\_VIA\_NFKB, ESTROGEN\_RESPONSE\_LATE, HYPOXIA, UV\_RESPONSE\_DN |

**Table 2:** Drug target modules

|  |  |
| --- | --- |
| **Drug** | **Modules** |
| ARAMCHOL | carnitine\_shuttle\_endoplasmic\_reticular, CHOLESTEROL\_HOMEOSTASIS, transport\_reactions, ANDROGEN\_RESPONSE, MYOGENESIS, MTORC1\_SIGNALING |
| BMS-986036 | Nonalcoholic Steatohepatitis |
| CENICRIVIROC | ALLOGRAFT\_REJECTION, ME 5, ME 3 |
| ELAFIBRANOR | CHOLESTEROL\_HOMEOSTASIS, FATTY\_ACID\_METABOLISM, Nonalcoholic Steatohepatitis, UV\_RESPONSE\_DN, XENOBIOTIC\_METABOLISM, ADIPOGENESIS, NOTCH\_SIGNALING, WNT\_BETA\_CATENIN\_SIGNALING |
| EMRICASAN | APOPTOSIS, COAGULATION, COMPLEMENT, HYPOXIA, IL2\_STAT5\_SIGNALING, INTERFERON\_GAMMA\_RESPONSE, UV\_RESPONSE\_UP, XENOBIOTIC\_METABOLISM, OXIDATIVE\_PHOSPORYLATION, GLYCOLYSIS, INTERFERON\_GAMMA\_RESPONSE, P53\_PATHWAY, UV\_RESPONSE\_UP, ME 10, ME 1, ME 2, ME 18 |
| GR-MD-02 | APOPTOSIS, CHOLESTEROL\_HOMEOSTASIS, COMPLEMENT, ME 13 |
| OBETICHOLIC ACID | BILE\_ACID\_METABOLISM, KRAS\_SIGNALING\_UP, Nonalcoholic Steatohepatitis |
| SEMAGLUTIDE/ LIRAGLUTIDE | N/A |
| MGL-3196 | KRAS\_SIGNALING\_DN |
| SELONSERTIB | ME 1 |

**Table 3:** NASH SNP modules

|  |  |
| --- | --- |
| **SNP** | **Modules** |
| GCKR | XENOBIOTIC\_METABOLISM |
| MBOAT7 | transport\_reactions, ME 9, leukotriene\_metabolism |
| PNPLA3 | transport\_reactions, Nonalcoholic Steatohepatitis, transport\_reactions, acylglycerides\_metabolism, leukotriene\_metabolism, retinol\_metabolism, triacylglycerol\_synthesis |
| TM6SF2 | Nonalcoholic Steatohepatitis |

**Table 4:** PGC1 regulated transcription factors related to NASH

|  |  |
| --- | --- |
| **Transcription Factor** | **Modules** |
| FOXA2 | PANCREAS\_BETA\_CELLS, ME 16 |
| HNF4A | XENOBIOTIC\_METABOLISM, COAGULATION, COMPLEMENT |
| NRF1 | Isolated, ME 1 |
| PPARA | Nonalcoholic Steatohepatitis, FATTY\_ACID\_METABOLISM, isolated |
| YY1 | ME 4 |

**Table 5:** PGC1 downstream genes

|  |  |
| --- | --- |
| **Downstream Gene** | **Modules** |
| ACACA | MTORC1\_SIGNALING, transport\_reactions, fatty\_acid\_biosynthesis\_even\_chain, PI3K\_AKT\_MTOR\_SIGNALING |
| ACADM | transport\_reactions, ADIPOGENESIS, FATTY\_ACID\_METABOLISM, OXIDATIVE\_PHOSPHORYLATION, transport\_reactions, beta\_oxidation\_of\_branched\_chain\_fatty\_acids\_mitochondrial, butanoate\_metabolism, fatty\_acid\_oxidation, omega\_6\_fatty\_acid\_metabolism, propanoate\_metabolism, purine\_metabolism, valine\_leucine\_and\_isoleucine\_metabolism |
| ACLY | MTORC1\_SIGNALING, ADIPOGENESIS, ME 41, tricarboxylic\_acid\_cycle\_and\_glyoxylate\_dicarboxylate\_metabolism |
| APOC3 | COAGULATION, ME 8, protein\_assembly |
| CPT1A | carnitine\_shuttle\_endoplasmic\_reticular, transport\_reactions, Nonalcoholic Steatohepatitis, FATTY\_ACID\_METABOLISM, OXIDATIVE\_PHOSPHORYLATION, ME 46, carnitine\_shuttle\_cytosolic, carnitine\_shuttle\_peroxisomal, fatty\_acid\_oxidation |
| DGAT1 | transport\_reactions, ADIPOGENESIS, UV\_RESPONSE\_UP, acylglycerides\_metabolism, retinol\_metabolism, triacylglycerol\_synthesis |
| FASN | CHOLESTEROL\_HOMEOSTASIS, transport\_reactions, FATTY\_ACID\_METABOLISM, ME 43, ESTROGEN\_RESPONSE\_EARLY, fatty\_acid\_biosynthesis, fatty\_acid\_biosynthesis\_even\_chain |
| G6PC | XENOBIOTIC\_METABOLISM, glycolysis\_gluconeogenesis |
| PCK2 | ME 3, pyruvate\_metabolism |
| PDK4 | transport\_reactions, Nonalcoholic Steatohepatitis, ME 5, isolated, XENOBIOTIC\_METABOLISM, OXIDATIVE\_PHOSPHORYLATION |
| SCD | ANDROGEN\_RESPONSE, carnitine\_shuttle\_endoplasmic\_reticular, CHOLESTEROL\_HOMEOSTASIS, MTORC1\_SIGNALING, MYOGENESIS, transport\_reactions |
| TFAM | ME 6 |

**Explanation of Top NASH Modules:**

1. KRAS Signaling UP

This hallmark signaling module includes genes that are upregulated by KRAS activation. KRAS is a gene that codes for the protein K-Ras, which signals cells to divide and differentiate.

Gene alterations to the K-Ras pathway are [implicated](https://link.springer.com/article/10.1007/s11901-015-0260-z) in Hepatocellular Carcinoma (HCC). Mutagenic activation of Ras signaling was shown to be significantly enhanced by NAFLD in [mouse models](https://www.sciencedirect.com/science/article/abs/pii/S0304383518306384). Ras activation alleviated hepatic steatosis but accelerated hepatocarcinogenesis. K-Ras signaling may be a [contributing factor](https://pubmed.ncbi.nlm.nih.gov/24177031/) to FXR silencing. FXR deficiency can exacerbate NASH, while FXR activation [can be protective](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5624538/) against liver inflammation. Obeticholic Acid is an FXR agonist and is the drug that is closest to being deployed for treating NASH.

Cosine similarity: 0.78

Annotations: Obeticholic Acid

1. Estrogen Response Late

This hallmark signaling module includes genes with a late response to estrogen.

Nonnuclear estrogen receptor activation improves hepatic steatosis [in mice](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5045504/). An estrogen receptor agonist has shown to inhibit NASH in [preclinical models](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5685260/) by regulating bile acid and xenobiotic receptors. A longer duration of estrogen deficiency [confers fibrosis risk](https://aasldpubs.onlinelibrary.wiley.com/doi/full/10.1002/hep.28514) among postmenopausal women with NAFLD.

Cosine similarity: 0.77

Annotations: N/A

1. TNFA Signaling via NFKB

This hallmark signaling module includes genes that are regulated by NF-kB in response to TNF. TNF cytokines activate NF-kB.

Gene products regulated by NF-kB are [involved in](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4132122/) inflammation of the liver, liver fibrosis, liver regeneration, and apoptosis. NF-kB activation was associated with steatohepatitis in [rat models](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4132122/).

Cosine similarity: 0.76

Annotations: N/A

1. IL2 STAT5 Signaling

This hallmark signaling module includes genes up-regulated by STAT5 in response to IL2 stimulation.

STAT5B knockout mice are deficient in lipogenesis signals from insulin, resulting in hyperinsulinemia. [These mice](https://www.mdpi.com/1422-0067/14/11/21833/htm) present with IR, hyperglycemia, hyperlipidemia, and hepatic steatosis, but not with inflammation or fibrosis. STAT5 [is involved](https://link.springer.com/content/pdf/10.1023/B:APPT.0000045785.65546.a2.pdf) in the regulation of caspases, which are inhibited by the drug Emricasan.

Cosine similarity: 0.76

Annotations: Emricasan

1. Estrogen Response Early

This hallmark signaling module includes genes with an early response to estrogen.

Cosine similarity: 0.75

Annotations: FASN (PGC1 downstream gene)

1. Heme Degradation

This metabolic subsystem module includes genes involved in the degradation of heme.

Heme oxygenase-1 (HO-1) is the rate limiting enzyme in heme catabolism. HO-1 prevents nutritional steatohepatitis through [suppressing hepatocyte apoptosis.](https://lipidworld.biomedcentral.com/articles/10.1186/1476-511X-9-124) The induction of HO-1 is [an adaptive response](https://www.sciencedirect.com/science/article/pii/S0168827805000139?casa_token=UzRXvhKLlOYAAAAA:qd5MEG5zsy5wjaMuNq73V3hTvGR4i2EUyGBcDJ2RiRI0AYRuk5LpR1Lo3Nj-rrJA_OcRIg) against oxidative damage elicited by lipid peroxidation and may be critical in the progression of NASH.

Cosine similarity: 0.75

Annotations: N/A

1. Xenobiotic Metabolism

This hallmark signaling module includes genes encoding proteins involved in processing of drugs or other xenobiotics.

The progression from steatosis to steatohepatitis may be induced by xenobiotics via oxidative stress, inflammation, endoplasmic reticulum stress, and cell death. [Exposure to chemicals](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6062016/) may stimulate cell growth in the liver and increase chances of HCC developing.

Cosine similarity: 0.74

Annotations: Elafibranor, Emricasan, GCKR (SNP), HNF4A (PGC1 regulated transcription factor), G6PC (PGC1 downstream gene), PDK4 (PGC1 downstream gene)

1. Eicosanoid Metabolism

This metabolic subsystem module includes genes involved in the metabolism of eicosanoid signaling molecules formed from the oxidation of polyunsaturated fatty acids.

Eicosanoid plasma levels have [been suggested](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4274066/) for use as a clinical tool to discriminate between NAFL and NASH.

Cosine similarity: 0.73

Annotations: N/A

1. Hypoxia

This hallmark signaling module includes genes up-regulated in response to low oxygen levels.

Hypoxia occurs in the development and progression of [fatty liver disease.](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4223242/)

Cosine similarity: 0.72

Annotations: Emricasan

**Next Steps:**

* Do similar random sampling method against genes from different diseases (rather than all genes) to get p-values
  + Narrow disease from 11,000 by choosing diseases with similar number of genes as NASH (50+)
* Read GEO paper THOROUGHLY and summarize (generally gather information about the biology of the pathogenesis of NASH)

**Useful Literature:**

**Using vector embeddings to represent genes:**

[**Integrating node embeddings and biological annotations for genes to predict disease-gene associations**](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6311944/pdf/12918_2018_Article_662.pdf)

* Binary classification for disease gene prediction using embeddings and annotations as features
* Used node2vec embeddings concatenated with biological annotations from uniprot ("keywords") to predict gene-disease associations from OMIM
* Used feature selection to identify annotations relevant to each disease
  + Different feature selection by disease to predict
* Performed imbalance correction to correct for smaller # of genes associated with each disease
* SVM, random forest, kNN, GLM

[**Predicting Parkinson's Disease Genes Based on Node2vec and Autoencoder**](https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6454041/)

* Used SVM on node2vec embeddings learned from a PPI network to predict new Parkinson’s disease related genes
* Reduced dimension of the embedding vector through use of an autoencoder
* Randomly selected genes not associated with Parkinson’s disease to use as negative set
* New predicted genes were verified through literature search

**[Pathway and network embedding methods for prioritizing psychiatric drugs](https://www.biorxiv.org/content/10.1101/728055v1.full.pdf)**

* Yash and Margaret’s project
* Used gene expression data from a couple difference psychiatric conditions to predict diagnosis
  + PCA and UMAP to predict disease/no disease
* Used PROPS method to calculate pathway importance scores
  + Decision tree, SVM, random forest to predict which disease
  + Added to gene targets for each drug by extracting differentially expressed genes from CMap (expression data before and after treatment)
  + Combined disease gene signatures and drug-gene target lists to recommend drugs for a disease in ranked order
* Used node2vec on String and GNBR
  + Augmented drug recommendation list with mean pairwise cosine similarity of genes in disease module to genes in drug modules

**[Semantic Disease Gene Embeddings (SmuDGE): phenotype-based disease gene prioritization without phenotypes](https://academic.oup.com/bioinformatics/article/34/17/i901/5093225)**

* Created vector embeddings from STRING PPI
* Used cosine similarity of gene/disease embeddings - ranks gene-disease based on pairwise similarity of gene-disease (each disease has 1 embedding)
* Trained artificial neural network to predict gene-disease associations from embedding vectors
* Created a ranking classifier based on the model’s prediction scores, computed AUC

[**SemanticGO: a tool for gene functional similarity analysis in *Arabidopsis thaliana* and rice**](https://www.sciencedirect.com/science/article/pii/S0168945220301321?casa_token=ri5TrUHU4nMAAAAA:7XL8-g8IhTuxVPJWqyLAcgqLmfjgYhUk-MWG6bix-mCSjdBzXxuF00pRLRPXIluWgPV7ug)

* Used latent semantic analysis to create gene vectors from GO ontology labels on genes
* Adds vectors of genes in a pathway to create a pathway representing vector
* Computed the cosine similarity of vectors to determine association

[**Inferring novel genes related to oral cancer with a network embedding method and one-class learning algorithms**](https://www.nature.com/articles/s41434-019-0099-y)

* Embedded STRING PPI network using node2vec
* Used feature selection and built several different one class algorithm based inferring models

**GEO expression data:**

[**A comprehensive bioinformatics analysis on multiple Gene Expression Omnibus datasets of nonalcoholic fatty liver disease and nonalcoholic steatohepatitis**](https://www.nature.com/articles/s41598-018-25658-4.pdf)

* CD24 was the only gene coexpressed in all 3 datasets
* Contains GEO datasets –
  + GSE66676 contained 33 NAFLD or NASH tissues and 34 normal liver tissues. GSE49541 contained 32 advanced NAFLD tissues and 40 mild NAFLD tissues. GSE834521 included 126 NASH tissues and 98 normal liver tissues.