BIOINFORMATICS & NETWORK MEDICINE

Presented by
Francesco Mari, Livia Oddi and Lorenzo Pannacci
Group 06

OVERVIEW

- Introduction
- Data gathering and interactome reconstruction
- Comparative analysis of the identification algorithms
- Putative disease gene identification
- Drug repurposing
- Proconsul
- Conclusions

INTRODUCTION

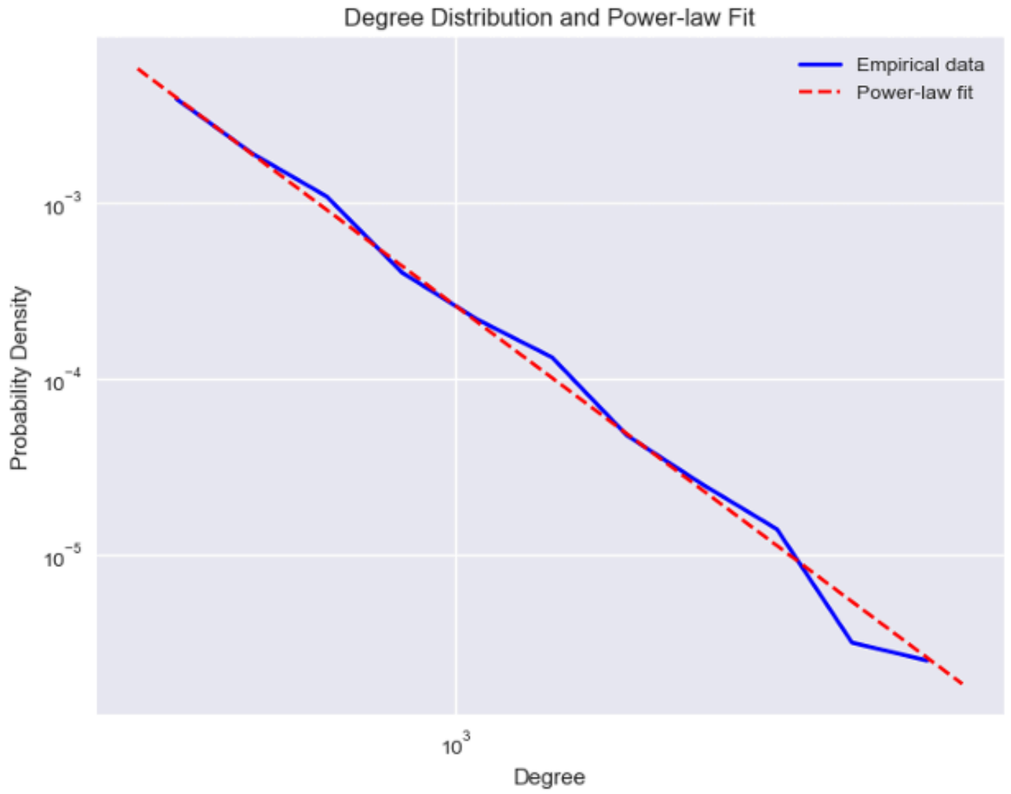
Background

Retinal dystrophies are hereditary and degenerative conditions affecting the retina and the choroid, causing symptoms like visual field contraction, night blindness, and alteration in color perception, often leading to blindness. They exhibit **high clinical** and **genetic heterogeneity**, with multiple genes and mutations causing varied phenotypes. Currently, there is no cure, but **gene therapy shows potential** as a **promising treatment approach**.

Objectives

Our study has the aim of firstly **find new putative disease genes** for retinal dystrophy using some disease-gene identification algorithms, infer from them pathways and molecular functions via enrichment analysis and finally **propose** the use of **new drugs** on **the basis of the newfound knowledge**.

DATA GATHERING (PPI)



01

The PPI database records 1.265.586 interactions of which 94.104 include **non-human proteins** and **are removed,** ulterior entries are removed due to being **non-physical** or **self-loops**.

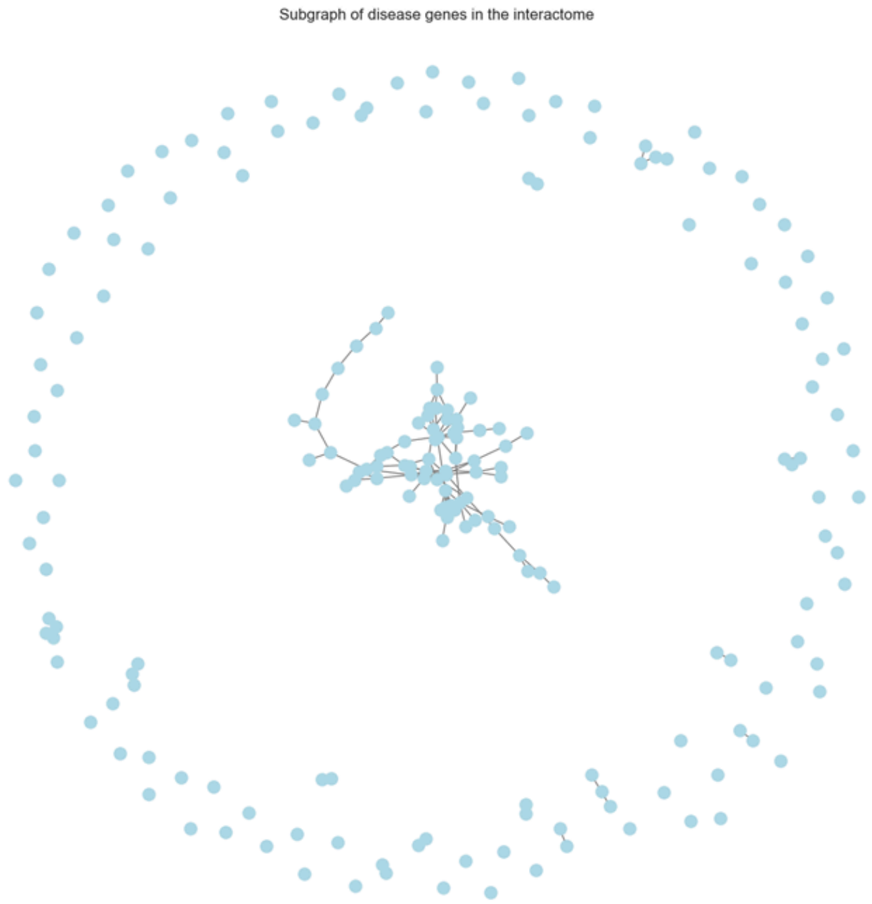
02

It ends up with an interactome of 19.972 nodes (genes) and 861.240 edges (physical interactions). Interestingly enough the **interactome is already** a **single connected component**.

03

To analyze the topology we perform a **power-law fit** to check wether the network follows a scale-free structure. The goodness-of-fit is measured with a **Kolmogorov-Smirnov test.**

DATA GATHERING (GDA)



01

The GDA database for retinal dystrophy records 263 genes of which 47 are **discarded** due to being **not protein-coding** and for the rest symbol checking is performed via HGNC's multi-symbol checker tool.

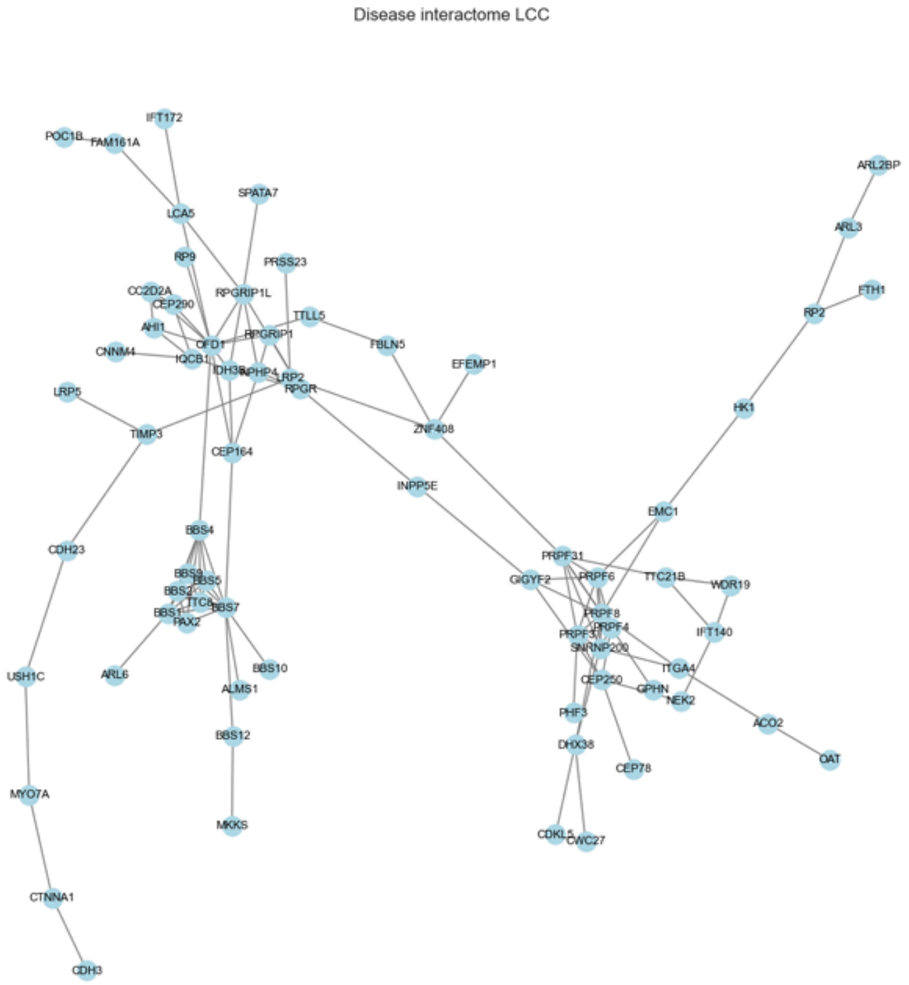
02

While all inputted symbols are 'Approved' symbols, **some** are **also** '**Alias**' or '**Previous**' symbols of other genes and for them the **manual check** has been performed using mostly the **gene full name** or the **GeneEnsembleID**.

03

Finally the disease interactome is identified by getting the **interactions among disease genes only**. The disease interactome is composed of a single large connected component, few little components with 2-3 nodes and a majority of unconnected nodes.

DATA GATHERING (DISEASE LCC)



04

Selecting the LCC we end up with a graph of **72** nodes and **120** edges. We obtained a small LCC of about 1/3 of the size of the GDA.

05

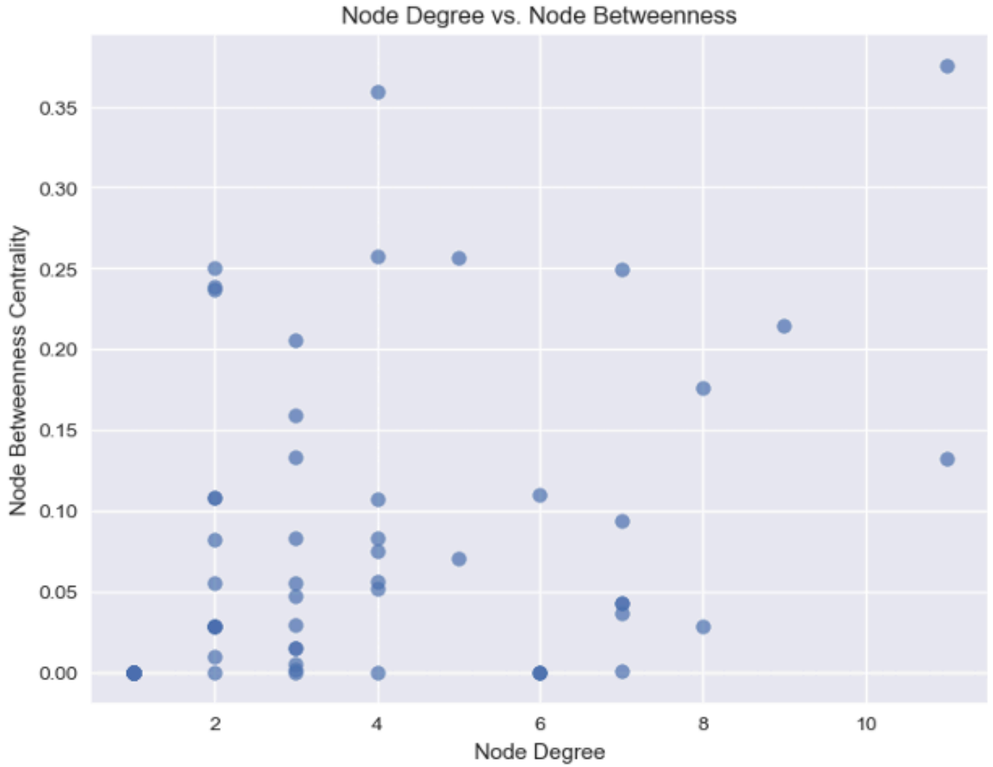
Confronting this value to what is obtained using the datasets assigned to other teams a **large difference** is **evident**, with their LCC being always a larger proportion of the dataset, usually about 150-200 genes and also better connected.

NETWORK ANALYSIS

Basic Network Measures

Network measures of the first 50 disease genes in the disease LCC ordered for node degree from higher to lower.

Ranking	Gene name	Degree	Betweenness	Eigenvector centrality	Closeness centrality	Ratio Betw./Degree
1	OFD1	11	0.0342	0.3758	0.2367	0.0031
2	BBS7	11	0.0120	0.1322	0.1888	0.0011
3	PRPF8	9	0.0238	0.2146	0.2351	0.0026
4	BBS4	8	0.0220	0.1760	0.2052	0.0028
5	BBS1	8	0.0036	0.0285	0.1766	0.0004
6	PRPF4	7	0.0061	0.0427	0.2139	0.0009
7	PRPF6	7	0.0133	0.0933	0.2305	0.0019
8	PRPF3	7	0.0061	0.0427	0.2139	0.0009
9	PRPF31	7	0.0356	0.2495	0.2399	0.0051
10	SNRNP200	7	0.0053	0.0369	0.2152	0.0008
11	BBS2	7	0.0	0.0	0.1762	0.0
12	BBS5	6	0.0	0.0	0.1757	0.0
13	BBS9	6	0.0	0.0	0.1757	0.0



Degree vs Betweenness scatterplot

As consequence of the structure of our GDA, both the metrics recorded in the table and the scatterplot are different from what could be expected.

E.g. node degree and betweenness centrality seem **not positively correlated**.

COMPARATIVE ANALYSIS OF THE GENE IDENTIFICATION ALGORITHMS - PART 1

DIseAse MOdule Detection (DIAMOnD):

Assumption that **connectivity significance** is the best predictor for finding putative disease genes. Iteratively computes **hypergeometric tests** for all candidate genes and chooses the one with the lowest p-value.

$$\text{p-value} = \sum_{k_i=k_s}^{k} p(k, k_i)$$

$$p(k, k_s) = \frac{\binom{s_0}{k_s}\binom{N-s_0}{k-k_s}}{\binom{N}{k}}$$

DIAMOnD Background Local Expansion (DiaBLE):

Introduces in the DIAMOnD algorithm the concept of **dynamic gene universe**. Instead a fixed universe made of the whole interactome uses just a selection made of the seed set, the candidates and their neighbors.

COMPARATIVE ANALYSIS OF THE GENE IDENTIFICATION ALGORITHMS - PART 2

Heat diffusion-based:

Uses **Fourier's laws of heat conduction** to test distance between candidates and seed set by setting seed genes heat value to 1 and make it diffuse through the network. Formula obtained solving a **differential equation**:

$$h(t) = e^{-Lt} h(0)$$

Computational Validation:

Best algorithm is selected by measuring their capability of **reconstructing a missing part of the seed set** with 5-fold cross-validation.

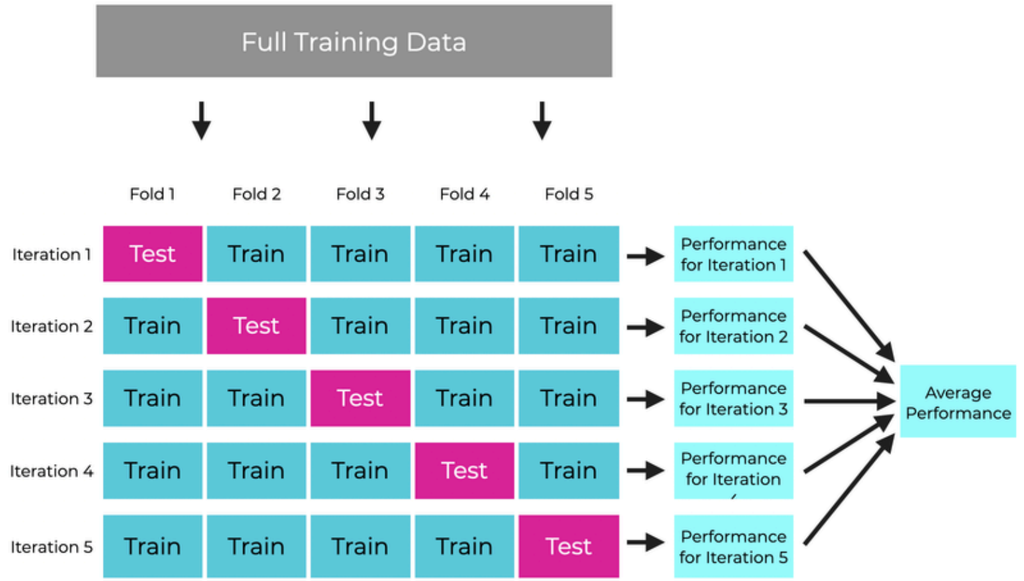
COMPARATIVE ANALYSIS OF THE GENE IDENTIFICATION ALGORITHMS - PART 3

Setup of cross-validation:

- All algorithms have been run with **default parameters** and a **varying amount of selected genes**: 50, n/10, n/4, n/2, n with n the amount of disease seed genes in the LCC;

- Also **different diffusion times** for the third algorithm are tested: 0.002, 0.005, 0.010;

- The metrics that we will take in most consideration for choosing the best performing algorithm are those for **n = 100%**, since it is the **closest to the 100 putative disease genes** that will be retrieved in point 3.1.



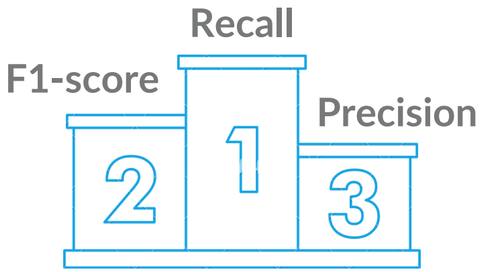
COMPARATIVE ANALYSIS OF THE GENE IDENTIFICATION ALGORITHMS - PART 4

Metrics discussion:

The comparison was done according to an order of metrics: first Recall, then F1-score, and finally Precision.

- **Recall**: important for capturing as many true disease-related genes as possible;
- **F1-score**: a tradeoff between recall and precision.
- **Precision:** important for ensuring that the identified genes are truly disease-related and minimizing the inclusion of false positives.

The choice was **context-related**, since in this area we want to find the greatest number of putative disease genes.



	POSITIVE	NEGATIVE
POSITIVE	TP	FN
NEGATIVE	FP	TN

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN}$$

$$Accuracy = \frac{TP + TN}{TP + FP + FN + TN}$$

$$F1\ Score = 2 \times \frac{Precision \times Recall}{Precision + Recall}$$

COMPARATIVE ANALYSIS OF THE GENE IDENTIFICATION ALGORITHMS - PART 5

Observations on results:

n	Diffusion time	Precision	Recall	F1-score
50	0.002	0.052 ± 0.016	0.179 ± 0.049	0.081 ± 0.024
50	0.005	0.052 ± 0.016	0.179 ± 0.049	0.081 ± 0.024
50	0.010	0.052 ± 0.016	0.179 ± 0.049	0.081 ± 0.024
10%	0.002	0.229 ± 0.069	0.110 ± 0.032	0.149 ± 0.044
10%	0.005	0.229 ± 0.069	0.110 ± 0.032	0.149 ± 0.044
10%	0.010	0.229 ± 0.069	0.110 ± 0.032	0.149 ± 0.044
25%	0.002	0.100 ± 0.042	0.124 ± 0.048	0.111 ± 0.045
25%	0.005	0.100 ± 0.042	0.124 ± 0.048	0.111 ± 0.045
25%	0.010	0.100 ± 0.042	0.124 ± 0.048	0.111 ± 0.045
50%	0.002	0.061 ± 0.021	0.151 ± 0.047	0.087 ± 0.029
50%	0.005	0.061 ± 0.021	0.151 ± 0.047	0.087 ± 0.029
50%	0.010	0.061 ± 0.021	0.151 ± 0.047	0.087 ± 0.029
100%	0.002	0.039 ± 0.010	0.193 ± 0.046	0.065 ± 0.017
100%	0.005	0.039 ± 0.010	0.193 ± 0.046	0.065 ± 0.017

Table: Results of cross-validation for heat diffusion

n	Precision	Recall	F1-score
50	0.032 ± 0.016	0.109 ± 0.051	0.049 ± 0.024
10%	0.143 ± 0.128	0.068 ± 0.059	0.153 ± 0.041
25%	0.089 ± 0.044	0.109 ± 0.051	0.098 ± 0.048
50%	0.044 ± 0.022	0.109 ± 0.051	0.063 ± 0.031
100%	0.025 ± 0.014	0.123 ± 0.063	0.042 ± 0.022

Table: Results of cross-validation for DiaBLE

n	Precision	Recall	F1-score
50	0.024 ± 0.008	0.083 ± 0.025	0.037 ± 0.012
10%	0.114 ± 0.107	0.054 ± 0.050	0.123 ± 0.042
25%	0.067 ± 0.022	0.083 ± 0.025	0.074 ± 0.024
50%	0.033 ± 0.011	0.083 ± 0.025	0.048 ± 0.015
100%	0.019 ± 0.011	0.096 ± 0.052	0.032 ± 0.018

Table: Results of cross-validation for DIAMOnD

- DiaBLE is better than DIAMOnD in every metric using the disease LCC but using all disease genes makes them having identical results. Exploring the DiaBLE universe we notice that **even at the first iteration it is 98% of the whole interactome**.

- Another interesting behavior is that the **ranking derived from heat diffusion is irrelevant of the diffusion time** (but the heat scores themselves vary).

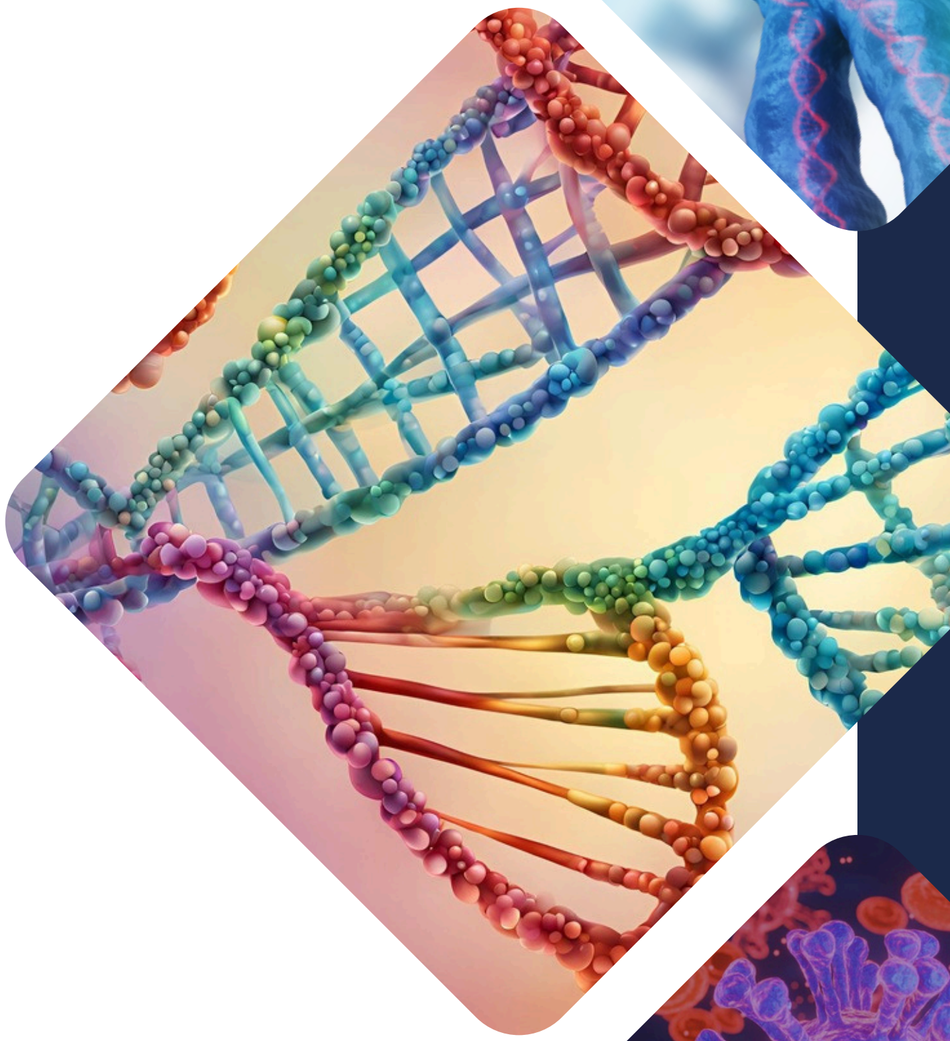
- We believe both these phenomena can be **attribuited to the network topology**.

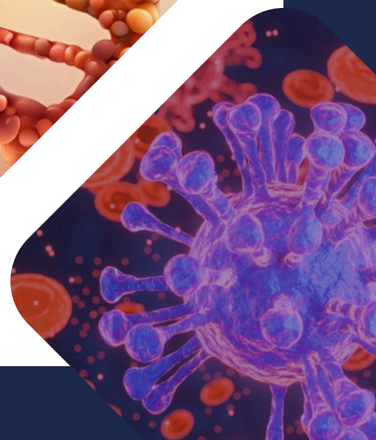
PUTATIVE DISEASE GENE IDENTIFICATION

Best performing Algorithm

- The **best performing** algorithm was chosen according to the 3 metrics cited before;

- The best in all 3 was the **diffusion-based algorithm**.





DIFFUSION-BASED ALGORITHM

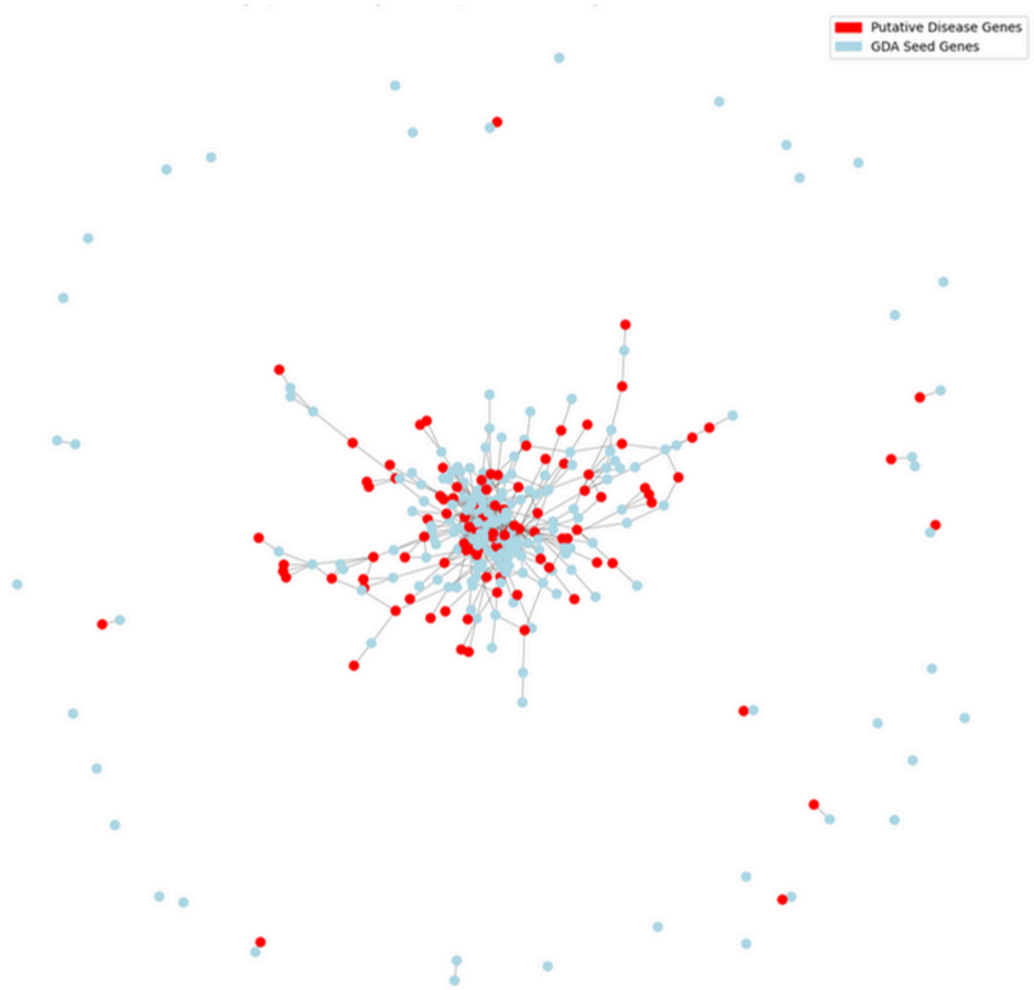


Figure: plot of GDA seed genes and putative disease genes in the interactome.

01

We used all GDAs as seed genes and we ran the diffusion algorithm using diffusion time of t = 0.002;

02

Finally we retrieved the list of the top 100 putative disease genes ranked with respect to the heat value;

03

It is interesting to notice that the found putative disease genes have managed to “attach” many of the components to the disease LCC.

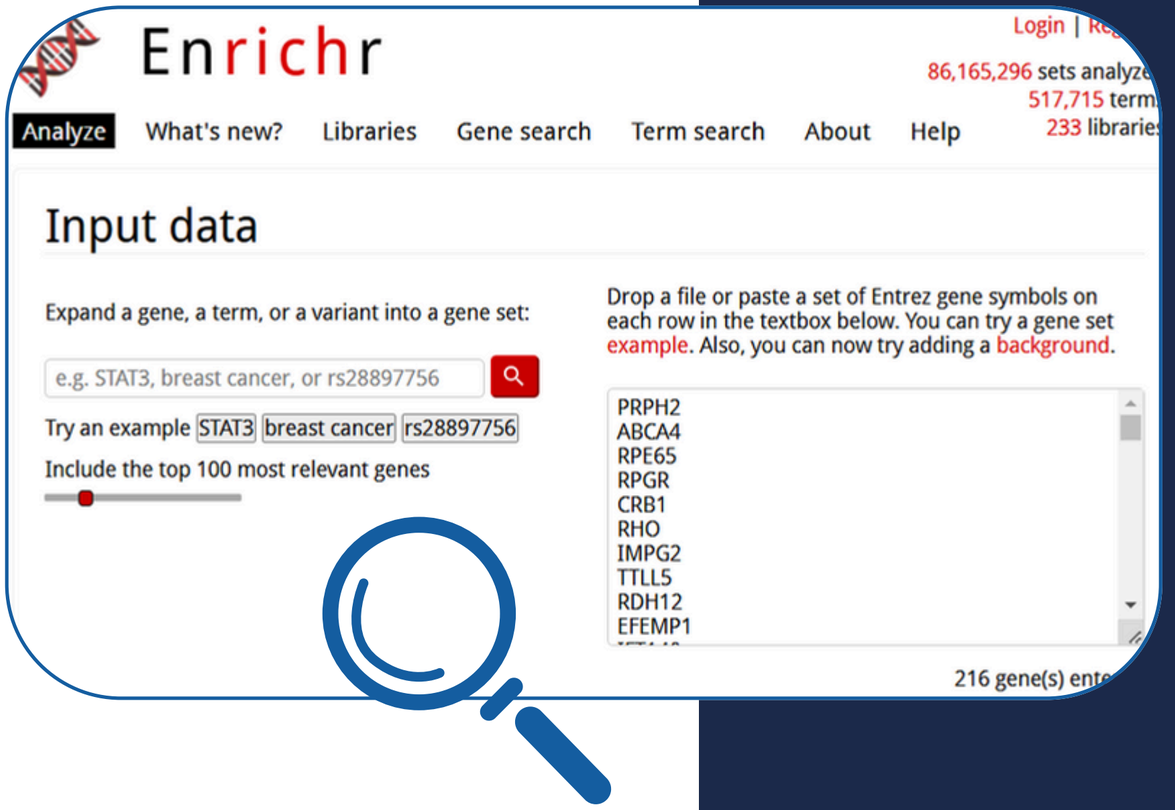
ENRICHMENT ANALYSIS

Enrichment analysis identifies overrepresented biological processes, pathways, or functions in a gene set compared to a background set.

Validation and discovery

We evaluated the **overlap between** enriched functions (with adjusted p-value<0.05) of **original** disease genes and **putative** disease genes. The results are the following:

- The **highest amount** of overlapping terms was found in the **Reactome Pathways** with **7** terms, corresponding to the 2% of the total.
- We also found **3** overlapping terms with on **Cellular Components** (1% in proportion).
- Nothing was found with respect to the KEGG Pathways, Molecular Function, or Biological Process



DRUG REPURPOSING

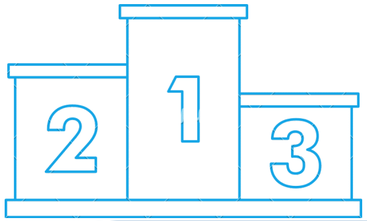
Drug identification

To identify potential drug candidates for retinal dystrophy we selected the first **20 putative disease genes** from the list of genes **predicted by** the **diffusion-based algorithm**.
We then used DGIdb latest ***interactions.tsv*** file to associate such 20 genes to ***approved drugs***

gene_name	*drug_name*	*Approved*
BHMT	Betaine Hydrochloride	True
NRXN1	Duloxetine Hydrochloride	True
BHMT	Betaine	True
NRXN1	Nicotine Polacrilex	True

retained drugs where ranked based on the number of disease genes they targeted

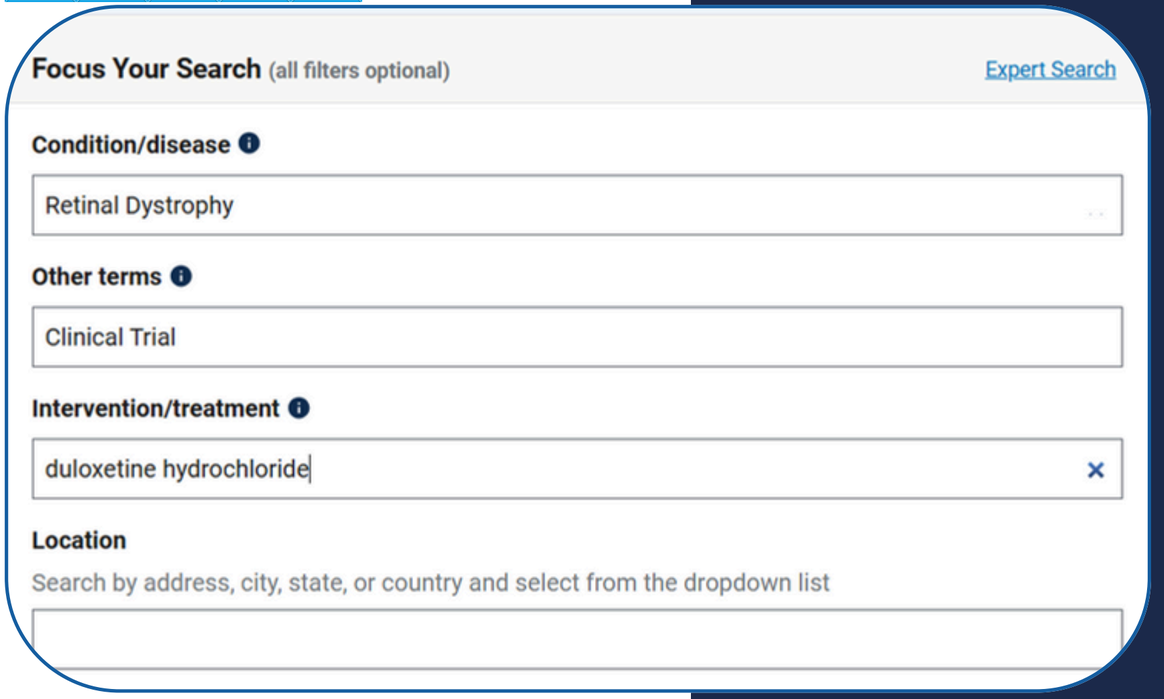
DRUG REPURPOSING



Validation and discovery

The ***predicted genes*** had **limited evidence in existing databases (GDAdb) connecting them to retinal dystrophy**, possibly due to :

- A lack of studies exploring their connection to retinal dystrophy.
- Emerging research suggesting relevance, though insufficient for database inclusion.
- Potential involvement in retinal dystrophy pathways without direct evidence.



Pharmacological properties of drugs like ***Duloxetine*** and ***Betaine*** might provide **indirect benefits**

modulates neuroinflammation

plays a role in processes linked to oxidative stress and cell death

PROCONSUL

PROCONSUL is a probabilistic modification of DIAMOnD that explores putative disease genes by converting p-values into probability distributions using softmax, enabling the selection of genes that may lead to better overall results. It reduces statistical fluctuations with techniques like temperature scaling and multiple runs.

PROCONSUL ALGORITHM

Putative disease genes
LZTFL1
PCM1
SSX2IP
KIAA0753
HAUS8...

Weighted Kendall's Tau

$$\tau_w = \frac{\sum_{i<j} w_{ij} \cdot \text{concordant}(i,j)}{\sum_{i<j} w_{ij}}$$

Spearman's Footnote Distance

$$H(R_1, R_2) = \sum_{i=1}^{n} |R_1(i) - R_2(i)|$$

DIFFUSION-BASED ALGORITHM

Putative disease genes
NKX6-2
GPR65
BBIP1
DEFB118
LZTFL1 ...

PROCONSUL

Results

Common putative disease gene
LZTFL1

- ***Minimal overlapping***: just one common gene between the 2 lists of 20 putative disease genes
- No meaning in comparing the lists with the previously mentioned metrics

The ranked list of genes produced by PROCONSUL was **analyzed for drug repurposing** opportunities, but **no approved drugs** targeting the PROCONSUL-identified genes were found **in the DGI database**

gene_name	drug_name	approved

Disease complexity influenced algorithm performance and overlap between rankings.

Despite inconclusive results for drug repurposing, these methods might hold potential for diseases with less genetic complexity.

CONCLUSIONS

While the evidence gathered is weak probably due to the structure of the GDA, which itself is most likely caused by the well-known genetic heterogeneity of the disease, we managed to find two drug candidates which influenced pathways seems to be indeed related to retinal conditions, even if no clinical trials involving them and retinal dystrohpy currently exist.

THANK YOU!

For your attention