Master Degree: Data Science 05 December 2023
a.y. 2023/2024

BIOINFORMATICS AND NETWORK MEDICINE

# Putative disease gene identification and drug repurposing for Ciliopathies - C4277690

O. Panneh, M. Razavi, J. Qian

GROUP 07

**ABSTRACT**

Ciliopathies comprise a group of inherited diseases caused by mutations in genes encoding proteins that localize to cilia or centrosomes. They afflict multiple organs and are one of the most frequent monogenic causes of kidney failure in adults, adolescents and children. Primary cilia play diverse roles in cell signalling, cell cycle regulation, planar cell polarity and mechanosensing. The use of patient-derived cells possessing endogenous disease-causing mutations enables the study of these processes and their dysregulation in disease. Here, we aim to describe methods of identifying putative disease genes and drug repurposing for Ciliopathies using network algorithms.

**INTRODUCTION**

Ciliopathies, a group of inherited disorders caused by defects in cilia, pose significant challenges in genetic medicine, particularly affecting the kidneys. While our understanding of these disorders has advanced, effective treatments are still lacking, highlighting the need for continued research.

Drug repurposing (DR) offers a viable alternative in drug development, utilizing approved drugs for new therapeutic purposes. This approach is cost-effective and time-efficient compared to traditional drug discovery.

Our study employs a network-based approach to identify potential disease genes and repurposable drugs for ciliopathies, using algorithms like DIAMOnD, DiaBLE, and Diffusion-based methods. This report details our methodology, algorithmic comparisons, and the potential of our findings in the context of drug repurposing for ciliopathies, aiming to contribute to the development of effective treatments for these disorders.

**MATERIALS AND METHODS**

**PPI and GDA Data Gathering and Interactome Reconstruction:**

PPI and GDA Data Collection:

Source: BioGRID database, DisGeNET database.

Selection Criteria: Human protein interactions, GDAs related to the specified disease.

Data Extraction Method: Utilization of the latest BioGRID release, Employing the most recent DisGeNET release.

Network Construction:

Nodes Representation: Each protein in the PPI dataset.

Edges Representation: Interactions between proteins.

Integration of Data: Merging PPI network with GDA data.

Largest Connected Component (LCC) Analysis:

Identification: Focusing on the largest subset of interconnected nodes.

Characterization: Computing node degree, clustering coefficients, and path lengths.

**Comparative Analysis of Disease Genes Identification Algorithms:**

Algorithm Implementation:

DIAMOnD, DiaBLE, and Diffusion-based algorithms.

Customization: Adjusting parameters specific to our disease model.

Validation and Performance Analysis: 5-fold Cross-Validation

Metrics: Precision, recall, and F1-score.

**Putative Disease Gene Identification:**

Algorithm Selection: Based on the comparative analysis results.

Prediction Process:

Input: All known GDAs as seed genes.

Output: New putative disease genes list.

Enrichment Analysis: Conducted on the predicted genes.

**Drug Repurposing Analysis:**

Identification of Candidate Drugs:

Targeting: Top 20 putative disease genes.

Drug-Gene Interaction Data Source: DrugBank or similar databases.

Clinical Trial Check: For the top three identified drugs.

**Optional Analysis:**

Disease Module Identification:

Algorithms: MCL and Louvain for community detection.

Enrichment Analysis: Of the identified disease modules.

Data Analysis Tools and Software:

Bioinformatics Tools: Cytoscape for network visualization and analysis.

**RESULTS AND DISCUSSION**

Table 1 Summary of GDAs and basic network data

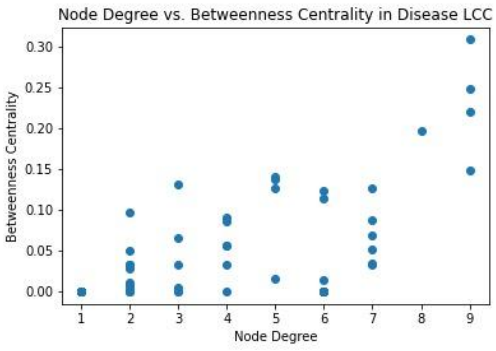
| disease name | UMLS disease ID | MeSH disease class | number of associated genes | number of genes present in the interactome | LCC size of the disease interactome |
|---|---|---|---|---|---|
| Ciliopathies | C4277690 | D000072661 | 110 | 104 | 120 |

Interactome network size after filtering = 822762.
We get the first 50 disease genes in the disease LCC ordered for node degree from higher to lower. We have listed the Top 10 in Table 2 below. The complete Top 50 can be viewed in the file 'top_50_disease_genes_metrics.csv'.

Table 2 Main network metrics of disease LCC genes

| Ranking | Gene name | Degree | Betweenness | Eigenvector centrality | Closeness centrality | ratio Betw./Degree |
|---|---|---|---|---|---|---|
| 1 | CC2D2A | 9 | 0.247988297 | 0.193626872 | 0.335195531 | 0.027554255 |
| 2 | OFD1 | 9 | 0.3079179 | 0.144808172 | 0.375 | 0.0342131 |
| 3 | BBS1 | 9 | 0.219598691 | 0.336135576 | 0.326086957 | 0.024399855 |
| 4 | BBS7 | 9 | 0.14775917 | 0.333996271 | 0.3125 | 0.016417686 |
| 5 | BBS4 | 8 | 0.196034661 | 0.333430729 | 0.342857143 | 0.024504333 |
| 6 | RPGRIP1L | 7 | 0.06907116 | 0.057273711 | 0.320855615 | 0.009867309 |
| 7 | B9D1 | 7 | 0.050823917 | 0.163812348 | 0.288461538 | 0.00726056 |
| 8 | TCTN2 | 7 | 0.033333333 | 0.159023624 | 0.269058296 | 0.004761905 |
| 9 | NPHP1 | 7 | 0.087825756 | 0.055617863 | 0.315789474 | 0.012546537 |
| 10 | BBS2 | 7 | 0.033722984 | 0.320446711 | 0.315789474 | 0.004817569 |

We plotted the node degree and node betweenness in a scatterplot which you can see below:



**PERFORMANCE COMPARISON**

We applied the diamond algo (Assuming 'seed_genes' is your list of known disease genes and 'network' is your interactive network) and saved the top 100 genes.

The first 5 top genes are

| Ranking | DIAMOnD_node | p_hyper |
|---|---|---|
| 1 | CEP72 | 8.186954e-29 |
| 2 | POC5 | 9.318644e-29 |
| 3 | KRT19 | 2.696095e-28 |
| 4 | FAM184A | 5.820018e-28 |
| 5 | HAUS1 | 7.641349e-28 |

Starting from the DIAMOnD code, we change the universe size used in the hypergeometric function to perform diable algo then we saved the top 100 genes:

Top 5 genes are:

| Ranking | DIAMOnD_node | p_hyper |
|---|---|---|
| 1 | POC5 | 1.948972e-28 |
| 2 | KRT19 | 7.105536e-28 |
| 3 | HAUS1 | 1.794255e-27 |
| 4 | CENPJ | 1.020523e-26 |
| 5 | HAUS5 | 1.490605e-26 |

Then we use Diffusion-based algorithm (available on Cytoscape), diffusion times (arbitrary unit): t=0.002, 0.005, 0.01

t = 0.01 results:

| | | | | |
|---|---|---|---|---|
| 0.990096499 | 0 | 6 | FALSE | ARL9 |
| 0.009061901 | 1 | 5 | FALSE | CKAP2L |
| 3.83E-05 | 2 | 5 | FALSE | SIRT5 |
| 3.12E-05 | 3 | 4 | FALSE | TRIM36 |
| 2.97E-05 | 4 | 4 | FALSE | CAMSAP2 |

Then we performed the algo for t = 0.002 and t = 0.005, and the ranking of the genes were the same you can see the results in the files "diffusion1", ""diffusion002", "diffusion005"

For computational validation we performed a 5-fold cross validation and compute the performance metrics for the diamond and diable algo the results are:

Diamond:

average Precision: 0.08 ± 0.08

average Recall: 0.12 ± 0.08

average F1 Score: 0.08 ± 0.07

Diable:

average Precision: 0.02 ± 0.06

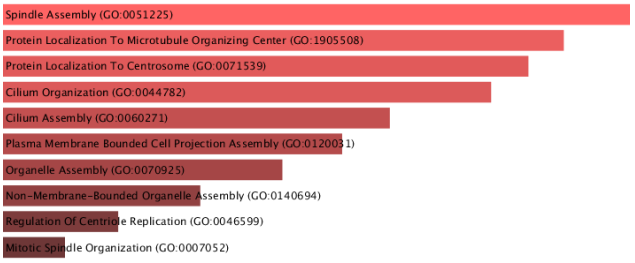
average Recall: 0.03 ± 0.06

average F1 Score: 0.02 ± 0.05

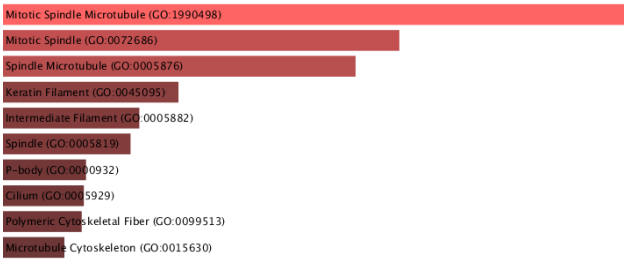
So we can see that the Diamond results are more accurate and we will use the Diamond in the next sections.

According to the performance metrics obtained in the validation phase at point 2, we performed the Diamond algo on the whole seed genes and saved the top 100 putative disease genes and the Enrichment analysis on the top 100 genes and also on the whole gene set you can see the results below.

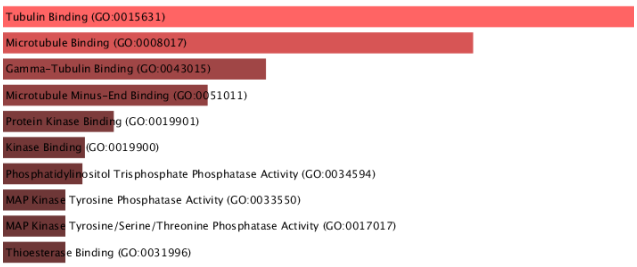
On the top 100 genes:



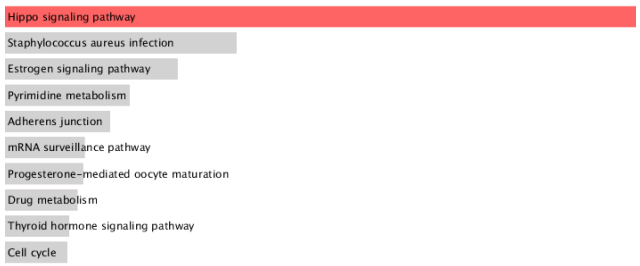

GO_Biological_Process




GO_Cellular_Component

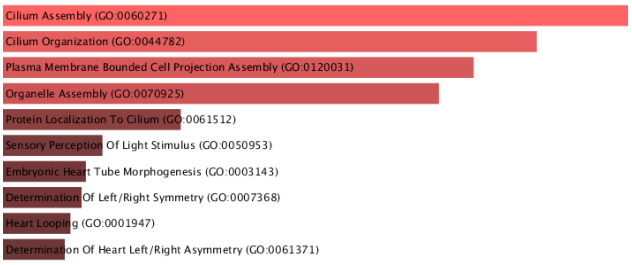



GO_Molecular_Function

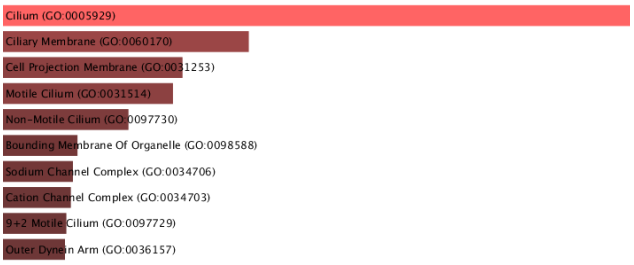



KEGG_Pahways

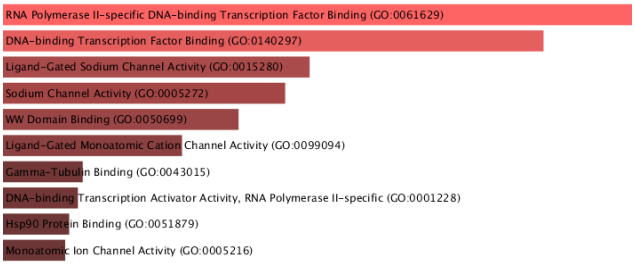
On the whole gene set:



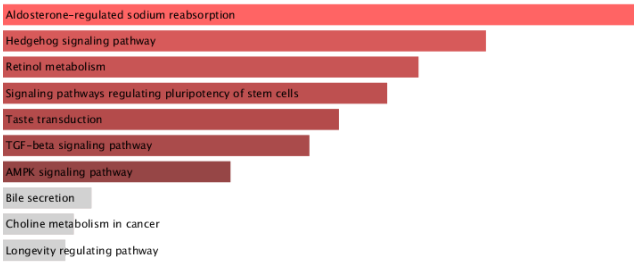

GO_Biological_Process




GO_Cellular_Component




GO_Molecular_Function




KEGG_Pahways

Here we find the drugs associated with the top 20 putative disease genes.

| | | | | | |
|---|---|---|---|---|---|
| KRT18 | KRT18 | MITOMYCIN | Approved | | 1.22 |
| CEP72 | CEP72 | VINCRISTINE | Approved | antineoplastic agent | 0.83 |
| KRT18 | KRT18 | RIBAVIRIN | Approved | | 0.79 |

Upon reviewing the top three drugs identified in Part 4.1 through the ClinicalTrials.gov database, it was determined that none of these drugs are currently involved in any clinical trials for the treatment of our targeted disease. This absence highlights a notable gap in existing clinical research and suggests an opportunity for future trials exploring their potential efficacy in this context.

**Bonus Part: Putative disease genes identification via clustering:** we used the MCL and Louvain community detection algorithms and found 6 clusters using MCL and 8 clusters using Louvain.it took about 6 hours to complete the MCL algorithm. Then with the use of hypergeom we tried to find if it is a putative disease module or not but unfortunatly there were no significant ones.

```
Elapsed time:1:10:55.956601
Elapsed time:0:53:33.073668
Elapsed time:0:55:05.922628
Elapsed time:0:56:35.993539
Elapsed time:0:54:19.067377
Elapsed time:0:55:13.553476
```

## AUTHOR CONTRIBUTIONS

Here is a short description of the contribution to the project of each author:

M.R..:data gathering; M.R.: algorithm implementation; M.R.,J.Q.,O.P.: tasks; M.R.: cross-validation; O.P.,J.Q.: writing—original draft preparation; J.Q.,O.P.,M.R.: writing—review & editing.

## REFERENCES


I. Hildebrandt, F., Benzing, T., & Katsanis, N. (2011). Ciliopathies. The New England Journal of Medicine, 364(16), 1533-1543. DOI: 10.1056/NEJMra1010172.

II. Waters, A. M., & Beales, P. L. (2011). Ciliopathies: an expanding disease spectrum. Pediatric Nephrology, 26(7), 1039-1056. DOI: 10.1007/s00467-010-1731-7.

III. Pushpakom, S., Iorio, F., Eyers, P. A., Escott, K. J., Hopper, S., Wells, A., ... & Pirmohamed, M. (2019). Drug repurposing: progress, challenges and recommendations. Nature Reviews Drug Discovery, 18(1), 41-58. DOI: 10.1038/nrd.2018.168.

IV. Zhou, X., Menche, J., Barabási, A.-L., & Sharma, A. (2014). Human symptoms–disease network. Nature Communications, 5, 4212. DOI: 10.1038/ncomms5212.

V. Guala, D., & Sonnhammer, E. L. L. (2017). A large-scale benchmark of gene prioritization methods. Scientific Reports, 7, 46598. DOI: 10.1038/srep46598.

VI. ChatGPT