Master Degree: Data Science 27 November 2024
a.y. 2024/2025

BIOINFORMATICS AND NETWORK MEDICINE

Putative disease gene identification and drug repurposing for Renal cell carcinoma

A. Arman Feili, B. Milad Torabi

GROUP 13

Delivered: 18 Jan 2025

ABSTRACT

Renal cell carcinoma (RCC) is a type of kidney cancer with high mortality. This study uses network medicine to identify new RCC-related genes and explore drug repurposing. By reconstructing a human protein-protein interaction (PPI) network and integrating RCC-related gene-disease associations (GDAs), we identified 179 RCC genes, with 169 mapped to the interactome (the whole set of molecular interactions in a cell). The RCC disease network's largest component has 9,729 nodes and 32,164 edges. Centrality measures like degree and betweenness revealed key genes, including TP53 and EGLN3, highlighting potential roles in cancer progression. This framework aids in discovering novel disease genes and drug targets.

INTRODUCTION

Renal cell carcinoma (RCC) originates from renal epithelial cells and often has a poor prognosis in advanced stages. Network medicine integrates protein interaction data, gene-disease associations, and algorithms to identify disease drivers and drug targets. This project constructs a human interactome, incorporates RCC-related gene-disease associations to define the disease network, and computes centrality measures to identify key RCC genes. Algorithms are evaluated through cross-validation to predict novel RCC genes and identify potential drug repurposing options. This report covers interactome construction, RCC gene verification, and centrality analysis, setting the stage for further gene prediction, validation, and drug analysis.

MATERIALS AND METHODS

1. PPI and GDA data gathering and interactome reconstruction

PPI Data: We started by downloading the latest BioGRID release as a tab-delimited file containing interactions for all organisms. The data was filtered to retain only interactions where both "organism A" and "organism B" were Homo sapiens (taxon ID 9606). We kept only interactions classified as “physical” and removed self-loops and redundant edges to clean the dataset. From this refined protein-protein interaction (PPI) network, we extracted the largest connected component, which we refer to as the "human interactome."

Gene–Disease Associations (GDAs): To incorporate disease-specific data, we obtained curated renal cell carcinoma (RCC) gene–disease associations from the provided *DISEASES_Summary_GDA_CURATED_C0007134.tsv* file. Each gene symbol was verified against the HGNC database for accuracy, and any duplicates or misannotated entries were removed. This process resulted in a total of 179 RCC-associated genes being identified in the GDA file.

Disease Subnetwork Construction and Characterization: We mapped the 179 RCC-associated genes to the human interactome, finding 169 genes present in the network. Using these genes, we constructed a subnetwork specific to RCC, referred to as the "disease interactome." From this, we extracted the largest connected component of the disease interactome, termed the "disease LCC," and analyzed its basic topological measures, including node degree, betweenness, eigenvector centrality, closeness centrality, and the ratio of betweenness to degree. Following these steps, the final human interactome contained 9,729 nodes and 32,164 edges, with 169 RCC-associated genes included in the network.

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
Renal Cell Carcinoma	C0007134	Neoplasms, Urogenital Diseases	179	169	9729

We calculated five key network metrics for each gene in the RCC disease LCC: node degree, betweenness centrality, eigenvector centrality, closeness centrality, and the ratio of betweenness to node degree. Among the top 50 genes ranked by node degree, TP53 stood out as the highest-degree node with a degree of 2180 and a notable betweenness centrality of 0.2024. Other genes, such as EGLN3 and BAP1, also displayed significant connectivity and centrality, suggesting their critical roles in regulating RCC pathogenesis.

Table 2 Main network metrics of disease LCC genes

Ranking	Gene name	Degree	Betweenness	Eigenvector centrality	Closeness centrality	ratio Betw./Degree
1	TP53	2180	0.232813	0.352125	0.498386	0.000107
2	EGLN3	1291	0.102219	0.170427	0.436253	0.000079
3	BAP1	1278	0.076792	0.234356	0.439168	0.000060
4	HSPD1	985	0.080534	0.159718	0.436507	0.000082
5	NR3C1	943	0.083234	0.116472	0.430214	0.000088
6	CUL7	840	0.042759	0.153625	0.411558	0.000051
7	BSG	790	0.057597	0.095130	0.405942	0.000073
8	CDK2	760	0.037538	0.155976	0.419890	0.000049
9	CDH1	718	0.054608	0.101896	0.414964	0.000076
10	HSPB1	698	0.038715	0.130443	0.411384	0.000055
…	…	…	…	…	…	…
50	GSTP1	201	0.008610	0.040405	0.372221	0.000043

2. Comparative analysis of the disease genes identification algorithms

The primary goal of this comparative study was to evaluate three algorithms—DIAMOnD, DiaBLE, and Diffusion—for their effectiveness in identifying genes related to renal cell carcinoma (RCC). Each method uses a different strategy for ranking candidate genes, and their performance was measured using common metrics such as precision, recall, F1-score, and accuracy at multiple ranking thresholds. By systematically testing a range of parameter configurations for each algorithm, we aimed to determine which approach best prioritizes true disease genes and to highlight the importance of parameter tuning for achieving optimal results.

- **DIAMOnD:** DIAMOnD (Disease Module Detection) works by iteratively adding genes to a “disease module” based on their hypergeometric p-value enrichment with known disease-associated genes. In this study, DIAMOnD was tested with varying numbers of steps (max_steps = 5, 10, 15, 20) and different p-value thresholds (pval_threshold = 0.05, 0.01, 0.1, 0.001). Despite exploring these parameter spaces, the algorithm did not correctly identify any RCC genes, consistently yielding zero precision, recall, F1-score, and accuracy. These results suggest that the DIAMOnD approach, at least with the tested configurations, was not well-suited for this dataset or may require additional tuning tailored specifically to RCC.

- **DiaBLE:** DiaBLE modifies DIAMOnD by restricting the hypergeometric test’s “universe size” to the top-degree nodes in the protein-protein interaction network. This constraint focuses on highly connected regions with the expectation of better capturing disease modules. Various universe_size values (500, 1000, 2000, 3000, 4000) were tested in combination with DIAMOnD’s parameters, but DiaBLE, like DIAMOnD, did not produce any true positives for RCC. This outcome indicates that either the approach was too stringent or it was not adequately adapted to the characteristics of the RCC gene set. Consequently, DiaBLE also yielded zero metrics across all evaluated thresholds.

- **Diffusion:** Diffusion uses a random-walk-with-restart (RWR) method to propagate “heat” from known seed genes through the network, continuously restarting at the seeds with a given probability (restart_prob). Multiple restart_prob values (0.0005, 0.002, 0.005, 0.01, 0.02), along with parameters like max_iter, tol, and ensemble_runs, were explored. Unlike DIAMOnD and DiaBLE, Diffusion achieved meaningful results, including non-zero precision, recall, and F1-scores at various ranking thresholds. Notably, for smaller gene sets—such as the top 17—it attained perfect precision by selecting only true RCC genes without false positives. This success highlights the value of a global network traversal for identifying disease-relevant genes.

2.1. Performance:

Overall performance was assessed by how well each algorithm ranked known RCC genes among the top candidates. While DIAMOnD and DiaBLE failed to identify any correct RCC genes under all parameter configurations, Diffusion consistently outperformed them, showing a broad range of viable thresholds where precision, recall, and F1-score remained high. One of the best Diffusion setups (restart_prob = 0.01, max_iter = 200, tol = 1e-08, ensemble_runs = 10) achieved a balanced precision of about 0.548 and recall of 0.783 at the top-50 cutoff. By adjusting the threshold, it could emphasize either precision (top 17 genes) or recall (top 179 genes), indicating that Diffusion offers flexibility depending on the user’s priorities.

2.2. Computational Validation

Cross-Validation: A 5-fold cross-validation strategy ensured that each algorithm was rigorously tested. The RCC-associated genes were split into five subsets, with four subsets used for training (or seeding) and one subset for testing in each iteration. Training genes were considered “seeds” for DIAMOnD and DiaBLE, while Diffusion assigned a heat value of 1 to these training genes and zero to the rest. Importantly, when evaluating Diffusion’s performance, seed genes were excluded from the ranked list to focus on newly identified candidates. This process provided a robust assessment of how well each algorithm generalized to unseen data within the RCC context.

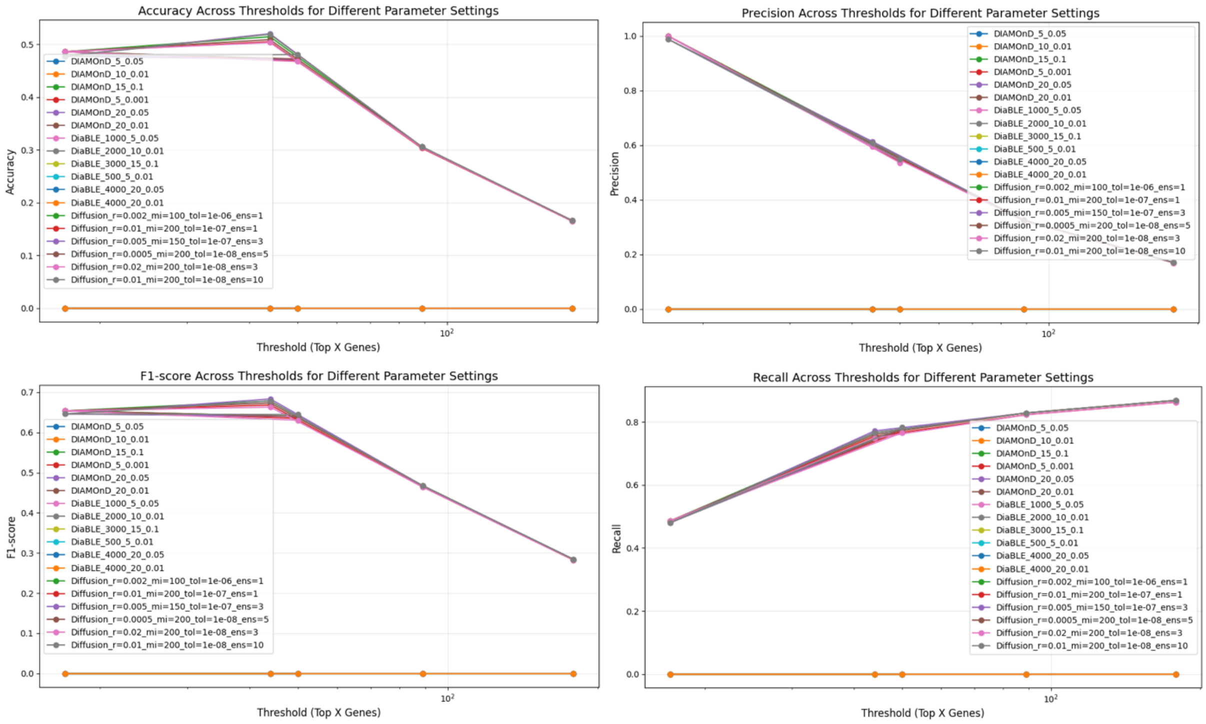
Performance Metrics: We assessed each algorithm's performance by analyzing the top-ranked genes (X = 17, 44, 50, 89, 179) and calculated the following metrics:

- Precision: The proportion of true positives (TP) among all predicted positives (TP + FP).
- Recall: The proportion of true positives among all relevant genes (TP + FN).
- F1-Score: The harmonic mean of precision and recall, representing their balance.
- Accuracy: The proportion of correctly identified genes (TP + TN) among all genes.

Diffusion Algorithm Demonstrated superior performance, achieving high precision, recall, and F1-scores.

- **Top 50 genes:** Precision = 0.544, Recall = 0.777, F1-Score = 0.640.

- **Lower thresholds (e.g., Top 17 genes):** Precision reached 1.0, indicating perfect performance for this smaller, highly relevant subset.
- **Higher thresholds (e.g., Top 179 genes):** Recall improved as more relevant genes were included, but precision decreased due to the inclusion of less relevant genes.



3. Putative Disease Gene Identification

Best-performing Algorithm: The best-performing algorithm from our analysis was the Random Walk with Restart (RWR) diffusion method, which utilized a restart probability of 0.01, a maximum of 200 iterations, a tolerance of 1e-08 for convergence, and 10 ensemble runs for stability. All known RCC genes were included as seed genes, each assigned a heat value of 1, while all other genes in the network started with a heat value of 0, ensuring the algorithm prioritized network proximity to known disease genes.

Obtain a list of 100 putative disease genes: The diffusion algorithm generated a global ranking of all genes based on their closeness to the seed genes, and after excluding the seed genes themselves, the top 100 new candidates were selected as putative disease genes, emphasizing those with the strongest potential links to RCC.

3.2. Enrichment Analysis

Gene Lists: Two gene lists were prepared for analysis: the original disease genes identified in Task 1.2 and the top 100 putative disease genes identified in Section 3.1. The original list consisted of known RCC genes, while the putative list contained new candidates predicted using the Random Walk with Restart diffusion algorithm.

Enrichr Analysis: Both lists were uploaded to Enrichr to identify enrichment in five categories: GO Biological Process (GO-BP), GO Molecular Function (GO-MF), GO Cellular Component (GO-CC), Reactome pathways, and KEGG pathways. Enrichr provided adjusted p-values for the significance of each term, allowing us to filter the results for those with p-values below 0.05.

Finding Shared Terms (Overlap Results): We compared the enriched terms between the original disease genes and the putative genes, focusing on shared biological pathways and functions. This step highlighted terms that were common between the two lists, indicating potential biological relevance of the new candidates.

Results: Using the best diffusion parameters (restart probability = 0.01, max_iter = 200, tol = 1e−08, ensemble_runs = 10), we ranked all genes and removed the known RCC genes from the list. The top 100 remaining genes were taken as putative disease genes, which included candidates like TRIM67, PARK2, MYC, and HSP90AA1. These genes are not part of the original RCC-associated gene set but are strong candidates for further study due to their proximity to seed genes in the network.

Enrichment analysis revealed significant overlaps between the original RCC genes and the putative genes in key biological categories:

- **GO Biological Process (BP):** 370 overlapping terms
- **GO Molecular Function (MF):** 22 overlapping terms
- **GO Cellular Component (CC):** 23 overlapping terms
- **Reactome pathways:** 123 overlapping terms
- **KEGG pathways:** 139 overlapping terms

These overlaps suggest that the putative genes are involved in similar or related biological processes as the original RCC genes. Notable pathways, such as HIF-1 signaling and apoptotic regulation, were enriched in both sets, supporting the biological relevance of the top 100 candidates. This overlap strengthens the case for these putative genes being connected to mechanisms involved in RCC and highlights their potential for further experimental validation.

4. Drug Identification

The top 20 putative RCC-related genes, including TRIM67, PARK2, MYC, HSPA8, HSP90AA1, and CTNNB1, were selected as the highest-ranked candidates from the previous analysis. Using the DGIdb dataset, drugs targeting these genes were identified and analyzed, with details such as gene names, drug names, and approval statuses. The drugs were ranked based on the number of target genes they interacted with, revealing CISPLATIN as the top candidate, targeting three genes, followed by DEXAMETHASONE and HEXACHLOROPHENE, each targeting two genes.

4.2. Clinical Trials Validation

Selecting the Top 3 Drugs: The top three ranked drugs were CISPLATIN, DEXAMETHASONE, and HEXACHLOROPHENE. CISPLATIN targets 3 of the top 20 genes, while the other two drugs each target 2 genes.

Rank	Drug name	Target count
1	CISPLATIN	3
2	DEXAMETHASONE	2
3	HEXACHLOROPHENE	2
…	…	…
100	DAUNORUBICIN LIPOSOMAL	1
101	WARFARIN	1

ClinicalTrials.gov Check:

A search on ClinicalTrials.gov was conducted to assess the relevance of CISPLATIN, DEXAMETHASONE, and HEXACHLOROPHENE for RCC. CISPLATIN was linked to 46 trials, DEXAMETHASONE to 3, and HEXACHLOROPHENE to none, indicating that CISPLATIN and DEXAMETHASONE are more promising for drug repurposing. For validation, an example trial, "Sorafenib Combined With Chemotherapy for Renal Collecting Duct Carcinoma" (NCT01762150), highlighted the process of testing drugs for RCC. Overall, the analysis identified CISPLATIN and DEXAMETHASONE as the most relevant candidates for repurposing, supported by clinical trial data.

5. OPTIONAL TASK - PROCONSUL

The PROCONSUL algorithm was used to predict and rank the top 20 RCC-related genes, which were then compared with the results from another leading method to validate their accuracy. The process involved organizing parameters, reusing prior results, and streamlining the PPI network to focus on key columns. The comparison revealed two overlapping genes, HSP90AA1 and CTNNB1, as significant candidates for further study. These findings underline the reliability of PROCONSUL and its consistency with other methods in identifying critical RCC genes, with results saved for future analysis.

RESULTS AND DISCUSSION

This research identified potential RCC-related genes and explored drug repurposing using a network-based approach. By analyzing a protein-protein interaction (PPI) network and integrating RCC gene-disease associations, key genes like TP53 and EGLN3 were highlighted. The Random Walk with Restart (RWR) diffusion method performed best, accurately prioritizing RCC genes through optimized parameters. Enrichment analysis confirmed the relevance of the top 100 candidate genes, linking them to pathways like HIF-1 signaling. Drug repurposing identified CISPLATIN and DEXAMETHASONE as promising options, supported by clinical trial data. These findings provide a foundation for further RCC research and treatment development.

AUTHOR CONTRIBUTIONS:

A. **Arman Feili:**, algorithm implementation, computational validation, task execution (putative gene identification, enrichment analysis, drug identification, and PROCONSUL application), and writing—original draft preparation.

B**. Milad Torabi:** data gathering algorithm implementation, cross-validation, enrichment analysis, task execution (putative gene identification and drug repurposing), and writing, reviewing and editing the report.

Programming Language and Tools

This project was developed using Python, leveraging libraries like NetworkX for network analysis, Pandas for handling tabular data, NumPy for numerical computations, and Scikit-learn for performance metrics. Matplotlib and Seaborn were used for visualizations, while external tools included Enrichr for enrichment analysis, ClinicalTrials.gov for drug validation, and PROCONSUL for gene prediction. Key data sources were BioGRID, DGIdb, and the DISEASES database.

REFERENCES

1. 1. Jensen, L. J., et al. DISEASES: Gene-Disease Associations database. Available at: https://diseases.jensenlab.org. Accessed: January 2025.
2. Stark, C., et al. BioGRID: A resource for studying protein and genetic interactions. Available at: https://thebiogrid.org. Accessed: January 2025.
3. Wagner, A. H., et al. DGIdb: The Drug-Gene Interaction Database. Available at: https://www.dgidb.org. Accessed: January 2025.
4. Kuleshov, M. V., et al. Enrichr: A comprehensive gene set enrichment analysis web server. Available at: https://maayanlab.cloud/Enrichr. Accessed: January 2025.
5. The Gene Ontology Consortium. Gene Ontology: Tool for the unification of biology. Available at: http://geneontology.org. Accessed: January 2025.
6. Jassal, B., et al. Reactome: A curated knowledgebase of biological pathways. Available at: https://reactome.org. Accessed: January 2025.
7. Kanehisa, M., and Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. Available at: https://www.genome.jp/kegg. Accessed: January 2025.
8. U.S. National Library of Medicine. ClinicalTrials.gov: A database of privately and publicly funded clinical studies. Available at: https://clinicaltrials.gov. Accessed: January 2025.