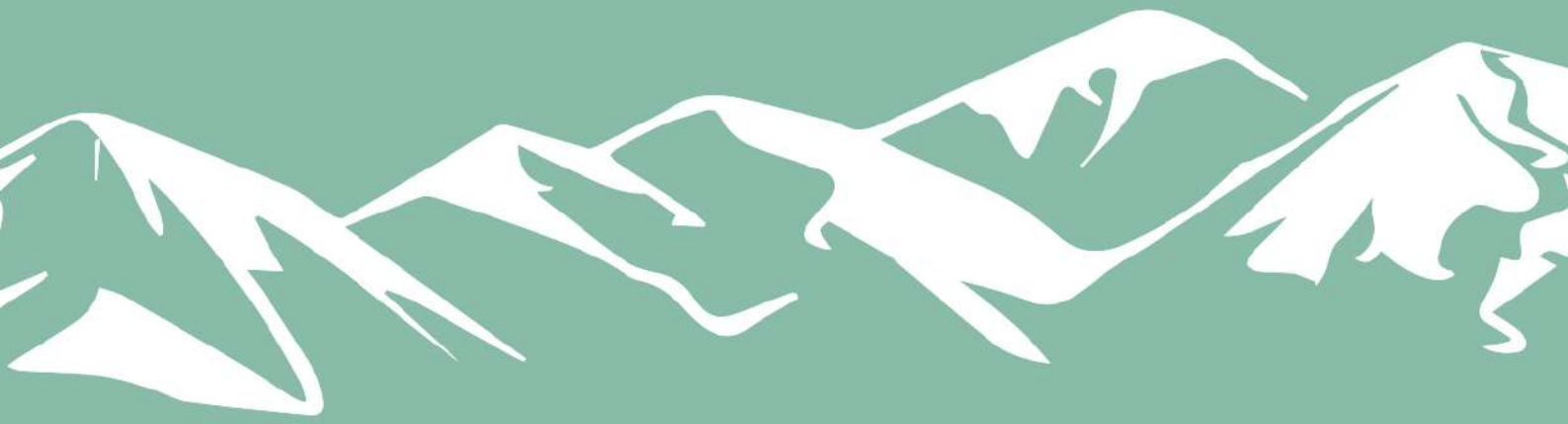
October 2023 ISSN: 1938-1158 01 59 2| Volume 59(2)

BIOMEDICAL SCIENCES INSTRUMENTATION

An international journal for the study of biomedical engineering, technology, & education



Oldest biomedical engineering journal | Published since 1963
IAE Publishing Windsor, Canada
Editor, Michelle Tucci, PhD, FAIMBE

IAE International Academic Express

Aims and Scope

Biomedical Sciences Instrumentation publishes peer-reviewed scientific articles for the advancement of biomedical engineering in relationship to patient safety, patient care, automated instrumentation for clinical decision making, and rehabilitation. It is the oldest engineering journal that encompasses the individual and collaborative efforts of scientists in clinical medicine, dentistry, basic and applied sciences, engineering, and bioethics. The journal is dedicated to the publication of outstanding articles of interest in the biomedical engineering research community.

Society Information

Beginning in 1963, the Rocky Mountain Bioengineering Symposium is the oldest, continuously held biomedical engineering symposium in the United States. It was founded by a group of the most visionary and historical individuals at the US Air Force Academy in the engineering field to promote dialog and the exchange of ideas and experiences between attendees, including between professionals and students.

From its beginning as a regional meeting, it has grown to a global event regularly attracting attendees from across the world. Since 1970, it has merged with the International Society of Automation Biomedical Sciences Instrumentation Symposium. Submitted papers are peer-reviewed, and those accepted for presentation and publication appear in the yearly issue of Biomedical Sciences Instrumentation journal, an internationally distributed publication by International Academic Express Company Ltd (iaexpress.ca).

Education

Biomaterials

Machine Learning

Clinical Rehabilitation

PRIORITIZING COMPLEX DISEASE GENES FROM HETEROGENOUS PUBLIC DATABASES: A CASE STUDY

Eric Gong[1] and Jake Y. Chen[1]

[1]The AI.MED Lab, Informatics Institute, University of Alabama at Birmingham, Birmingham, AL

Corresponding Author: Jake Y. Chen
Email:
Doi: 10.34107/DFYZ4711.10185

ABSTRACT

Background: Complex human diseases are defined not only by sophisticated patterns of genetic variants/mutations upstream but also by many interplaying genes, RNAs, and proteins downstream. Analyzing multiple genomic and functional genomic data types to determine a short list of genes or molecules of interest is a common task called "gene prioritization" in biology. Many statistical, biological, and bioinformatic methods are developed to perform gene prioritization tasks. However, little research has examined the relationships among the technique used, merged/separate use of each data modality, the gene list's network/pathway context, and various gene ranking/expansions.

Methods: We introduce a new analytical framework called "Gene Ranking and Iterative Prioritization based on Pathways" (GRIPP) to prioritize genes derived from different modalities. Multiple data sources, such as CBioPortal, PAGER, and COSMIC were used to compile the initial gene list. We used the PAGER software to expand the gene list based on biological pathways and the BEERE software to construct protein-protein interaction networks that include the gene list to rank order genes. We produced a final gene list for each data modality iteratively from an initial draft gene list, using glioblastoma multiform (GBM) as a case study.

Conclusion: We demonstrated that GBM gene lists obtained from three modalities (differential gene expressions, gene mutations, and copy number alterations) and several data sources could be iteratively expanded and globally ranked using GRIPP. While integrating various modalities of data can be useful to generate an integrated ranked gene list related to any specific disease, the integration may also decrease the overall significance of ranked genes derived from specific data modalities. Therefore, we recommend carefully sorting and integrating gene lists according to each modality, such as gene mutations, epigenetic controls, or differential expressions, to procure modality-specific biological insights into the prioritized genes.

Keywords: Bioinformatics, GRIPP, Gene, Disease, Gene prioritization

INTRODUCTION

Gene prioritization has been an important research topic due to the rapid accumulation of experimental genomics data and the challenge of interpreting them in various biological contexts [1, 2]. For example, querying the gene signature database MSigDB with the term "breast cancer" can retrieve over 2000 statistically significant genes [3]. Creating a single therapeutic solution targeting more than a few target genes/proteins would still be technically challenging for drug development. Similarly, while pathway biomarkers have been proposed to monitor disease prognosis [4], it would be costly to test for all of them in clinical practice. Therefore, gene prioritization is critical for the feasibility of biomarker studies, drug development, or disease mechanism determination [5, 6]. Many approaches have been developed to prioritize genes specific to human diseases. One single method of prioritizing genes is based on measured gene expression changes, e.g., by "fold change", which may be arbitrary to the choice of threshold values and the fact that some genes may change transiently or subtly to produce profound biological effects downstream [7]. Another method is to use p-value or other statistical significance measures to sort genes that changed under different conditions [8, 9]. However, p-value analysis can depend on the choice of samples; therefore, the results may not be transferable to new samples due to inherent sample biases [10]. Literature citation-based and text-mining approaches have also been used to curate significant genes related to a biological condition [1, 11, 12]. However, results may be biased toward literature studies that are "popular" and not necessarily

directly indicative of biological importance and significance [1, 8]. Network biology-based approaches, e.g., prioritizing genes based on gene network centrality measures, may overcome many inherent biases of pure statistical or literature-based gene prioritization techniques [13-16].

Network-based gene prioritization tools today are developed with inherent assumptions and limitations, therefore having to be applied in practice cautiously. Most prioritization tools, for example, may not provide necessary statistical details regarding the generated network. Instead, they may only provide a binary label [17]. Similarly, many network tools directly create the network without consideration for data quality. Thus, the responsibility of avoiding small sample coverage and filtering out noisy data from the pipeline falls to the users. For example, tools such as MGOGP [11] or KOBAS-i [12] are limited by their inability to expand gene lists. Thus, the output of these tools is restricted entirely by the quality of the input data; insufficient input data will lead to low quality or, even worse, biased results [18]. While the BEERE software tool [19] can expand gene lists, it fails to consider pathways and other biological contexts. Correctly placing genes into a biological context can aid the determination of significant genes not found within the initial query list [20]. The ability to parameterize search results—possible in tools such as BEERE and WINNER [8]—is also critical, allowing users to fine-tune results to their specific needs. Finally, almost no network tools implement a continuous, iterative refinement process. Iterative processes can help optimize the biological significance of results from network tools. The potential of many network-based prioritization tools may be hidden if multiple iterations are not conducted.

Here we introduce a new analytical framework called "Gene Ranking and Iterative Prioritization based on Pathways" (**GRIPP**) to address many of the limitations of current network prioritization methods. This framework addresses network biology challenges: providing quantitative statistical values to reflect biological significance, shifting the burden of data quality control away from users, and implementing an iterative process. In this study, GBM is used to illustrate the proposed workflow. However, the process can be easily generalized to any disease. We obtained the initial disease gene list (candidate genes) from several modalities, i.e., gene expressions, mutations, and copy number alterations, from several major biological databases. We then describe how to iterate through a network-based gene prioritization software tool (BEERE) and pathway enrichment/expansion software tool (PAGER) [21-23] to create a biologically relevant, high-quality, rank-ordered disease gene list.

METHODS

The process of procuring a high-quality disease gene list has four steps: compiling candidate gene subsets, gene prioritization based on pathway and network context, evaluation through literature co-citations, and evaluation through modality-specific gene ranks. The network-based gene prioritization and gene-set expansion tasks are iteratively repeated (as seen in Fig. 1 below) to refine the initial gene list into a highly query-specific and biologically significant gene list. To demonstrate this technique, we built a case study using glioblastoma, one of the most aggressive and widespread brain tumors in adults [24]. However, the proposed methodology can easily be generalized to all diseases. To freely access the processed data, results, and a detailed step-by-step description of the technique in this work, one can visit https://github.com/aimed-lab/gene-prioritization

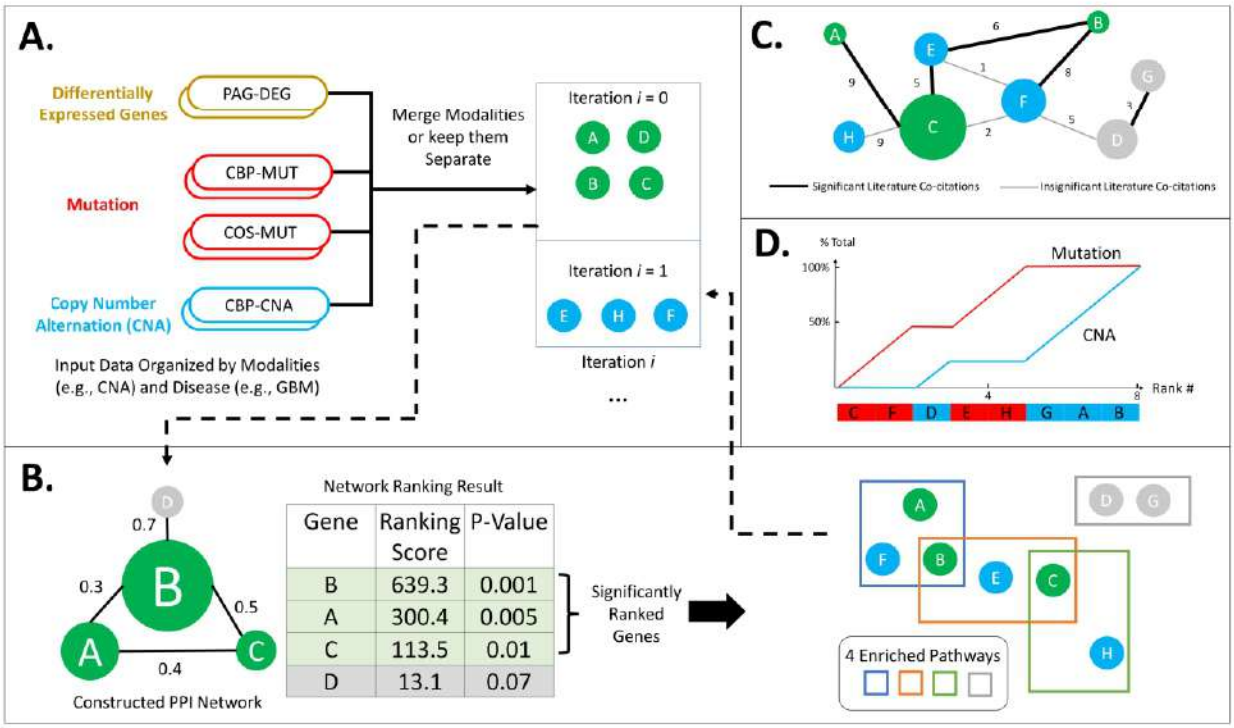


Figure 1. An overview of the Gene Ranking and Iterative Prioritization based on Pathways (GRIPP) framework. A) Compile seed gene lists (iteration $i = 0$), from different modalities such as Differentially Expressed Genes (DEG), Mutation (MUT), and Copy Number Alteration (CNA). They can be merged or kept separate as inputs to the gene prioritization process. B) First, construct a network, and pick any network gene prioritization software to generate significantly ranked genes; then, expand the resulting genes in their enriched pathway context to derive additional genes that will be added back to the next iteration (iteration $i = 1, 2\ldots$) C) Overlay a literature co-citation network on top of the final PPI network involving all the resulting genes from all iterations. Node color indicates the batch of genes, whether significant (green or blue) or insignificant (grey), derived from the iterative process. D) A systematic evaluation of gene lists built from different modalities and their impact on gene rank orders.

Compiling Candidate Gene Lists from Various Data Modalities

We created an initial disease gene list of five data subsets: PAG-DEG, CBP-CNA, CBP-MUT, COS-MUT, and COS-CTL. The COSMIC [25], CBioPortal [26], and PAGER [21] databases below are used to produce the initial disease gene list (candidate genes). Genes related to the query disease "Glioblastoma" were retrieved manually from databases, and a heuristic score was used to limit the number of genes retrieved for each data subset to less than 1200 for convenient downstream processing. Candidate genes from PAGER are differentially expressed genes and are represented by the symbol "**PAG-DEG**." We selected 663 PAG-DEG genes heuristically based on the PAGER nCoCo scores, which signify the biological relatedness of the gene sets included in PAGER. Candidate genes from CBioPortal are copy number alteration genes and mutated genes and are represented separately as "**CBP-CNA**" and "**CBP-MUT**", respectively. We selected 653 CBP-MUT genes and 391 CBP-CNA genes heuristically based on the frequency of copy number alteration and mutation occurrence, respectively. Candidate genes from COSMIC are mutated genes known to have a relation to the query disease and non-mutated control genes known to have no relation to the query disease and are represented separately as "**COS-MUT**" and "**COS-CTL**" respectively. We selected 570 COS-MUT genes and 439 COS-CTL genes heuristically based on the frequency of mutation, and lack of mutation frequency, respectively. To focus on only the highest-quality genes for downstream processing, BEERE was used to determine genes with significant p-value ($p < 0.05$). Here we differentiate between two combined gene lists. The first gene list was created from all five data subsets,

used to test the efficacy of BEERE gene prioritization, and called ALL-CTL. The second gene list was created from all data subsets except COS-CTL, used in the gene prioritization and refinement process, and called ALL-EXP.

Iterative Ranking and Expansion of Genes Based on their Network and Pathway Context

We refined the disease gene list using an iterative process involving a gene expansion and gene prioritization tool. Although currently existing tools are utilized, a unique pipeline is developed to coordinate these two tools to yield biologically significant results. We began the iterative prioritization and expansion process by using BEERE to prioritize the disease gene list, determining genes with significant p-value ($p < 0.05$). BEERE is a user-friendly implementation of the WINNER software [8, 19]. Although BEERE only outputs one p-value per gene [19], it is sufficient for the prioritization methods employed within this methodology. We used Prioritization to select the top genes amongst the candidate and expand genes for further processing, yielding a more specific, biologically significant gene list. BEERE allows for the configuration of a direct p-value cut-off and direct access to a protein-protein interaction database with quality control measures. Thus, it is easy to parameterize the Prioritization process, sensitively controlling the size of the candidate or processed gene list, and as a result, the scope of the study.

Next, we queried the prioritized gene list in PAGER to determine the related "pathway type" PAGs. The top upstream PAGs for each pathway PAG are then retrieved. The genes contained within each of the top upstream PAGs are then used to expand and enrich the gene list. The same heuristic score from the candidate gene compilation step – considering mutation frequency of the gene, frequency of overlap between different data sets, and the PAGER nCoCo score – was used to limit the number of expanded genes for convenient downstream processing before the expanded genes are combined with the original disease gene list. The gene enrichment step allows for significant genes not included in the initial candidate gene list to be added. This increases the overall quality of the gene list, and also aids in mitigating biases potentially present in the initial candidate gene list. The newly formed list was once again prioritized in BEERE, beginning another iteration of Prioritization, expansion, and refinement. The iteration was continued until less than 1% of the total expanded genes found through PAGER are not already present in the candidate gene list. In other words, the intersection of the candidate gene list and the expanded genes contains at least 99% of the expanded genes. The final high-quality gene list created from numerous iterations will be called ALL-FNL.

Evaluation of Prioritized Genes in the Literature Co-citation Network

To assess the effectiveness of the methodology, we performed an evaluation of the ALL-FNL genes. One criterion used to evaluate the workflow is the biological significance of the ALL-FNL genes. To determine the biological significance of the ALL-FNL genes, we constructed a network of statistically significant literature co-citations. This co-citation network was used to create a literature-based gene ranking, which is then compared against the ALL-FNL ranking generated through the iterative Prioritization process. To approximate the number of co-citations between glioblastoma and each retrieved gene, the two terms "Glioblastoma" and the gene name were queried in PubMed.

Evaluation of Prioritized Genes through Modality-specific Gene Ranks

The second main criterion to judge the methodology is its ability to remove insignificant genes. The distribution and enrichment of the control genes within the gene list must be determined to evaluate the methodology's ability to remove genes that do not have biological significance. Specifically, the Gene Set Enrichment Analysis (GSEA) software was utilized [27]. The ALL-CTL gene set – which contains the COS-CTL gene set known to have no significance to glioblastoma – is passed through one iteration of the methodology, with no cut-off parameter in BEERE, and then inputted into GSEA to determine where the COS-CTL genes would be placed in the gene ranking relative to the rank of

the other modalities, demonstrating whether or not the methodology can indeed filter out non-significant genes.

RESULTS

In **Figures 2 and 3**, we can observe that the expanded genes deduced from the pathway and network-based biological context are highly ranked alongside and share many interactions with the well-established Glioblastoma genes. From **Figure 3**, we see that literature popularity is not directly correlated with gene ranking. Thus, looking at both literature and experimental data when deriving biological insight is necessary. For example, while there are 611 PubMed co-citations between SRC and RAC1, there are only 17 co-citations between SUMO1 and TP53. Nonetheless, both of these co-citations are considered significant. Thus, the number of co-citations and the prevalence of co-citations are not directly related to the degree to which the co-citation is biologically significant.

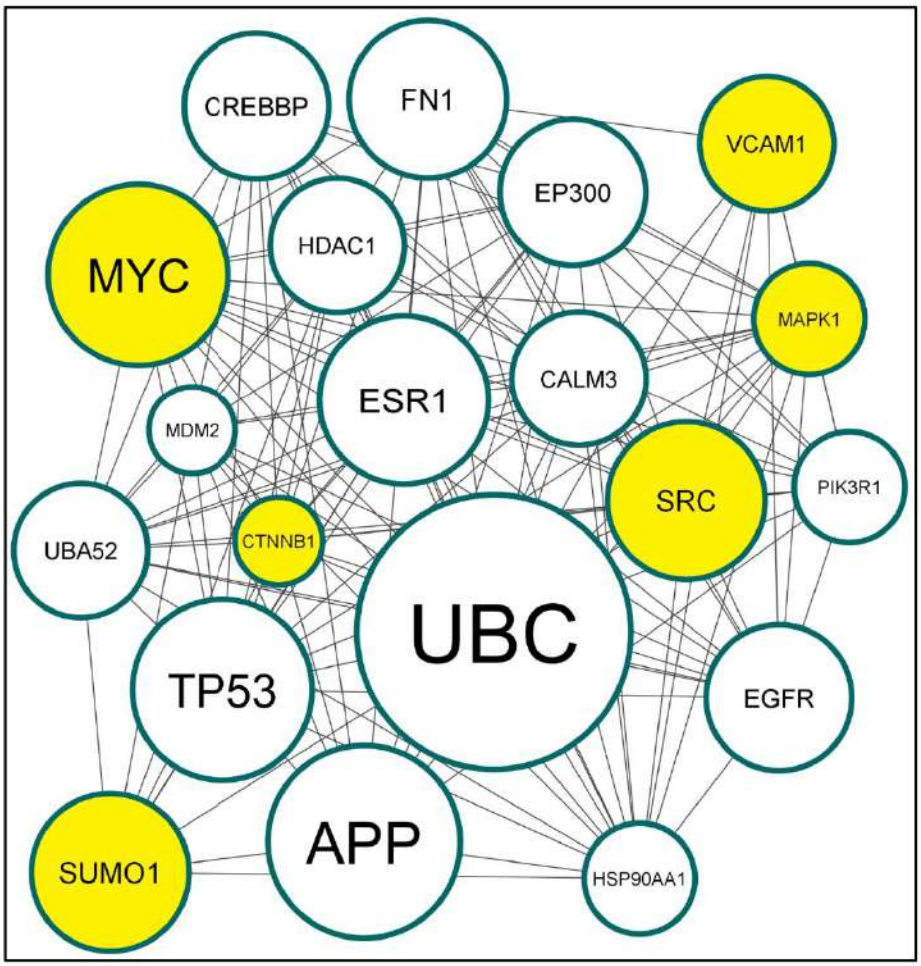


Figure 2. Gene refinement and prioritization using the query "Glioblastoma". Each node indicates a gene and each edge indicates a protein-protein interaction relationship in the glioblastoma gene network. The gene network contains the top 20 ALL-FNL genes. Node size represents gene ranking score obtained by BEERE. Highlighted nodes are genes added through expansion based on biological pathway and network context.

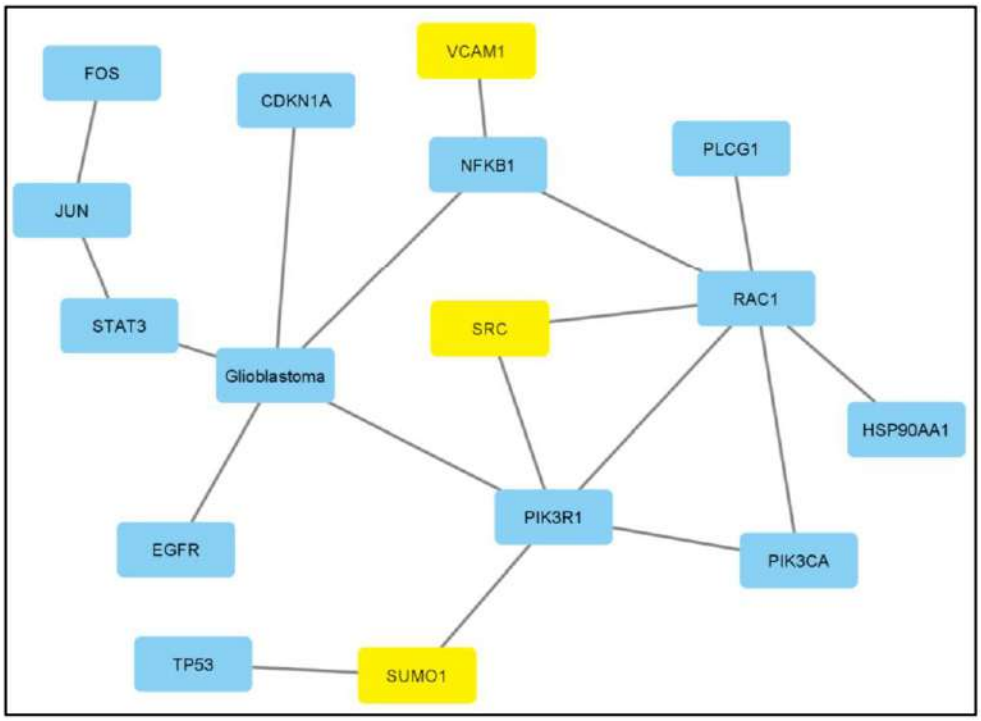


Figure 3. Statistically Significant Literature Co-Citation Network. In this network, each node indicates a biomedical entity term and each edge indicates a significant co-citation relationship determined by the BEERE software. The significant literature co-citation network is generated from the top 40 ALL-FNL genes. Highlighted nodes are genes added through the expansion based on a biological pathway and network context. Edges indicate statistically significant (p-value < 0.05) enriched co-citations in the literature of the two linked terms.

The COS-CTL genes are concentrated towards the lowest ranking, demonstrating the efficacy of BEERE in gene prioritization (**Figure 4**). However, we note that there is a preference for differentially expressed genes when compared to the mutated and CNA genes. Thus, there may be bias when integrating and combining multiple modalities.

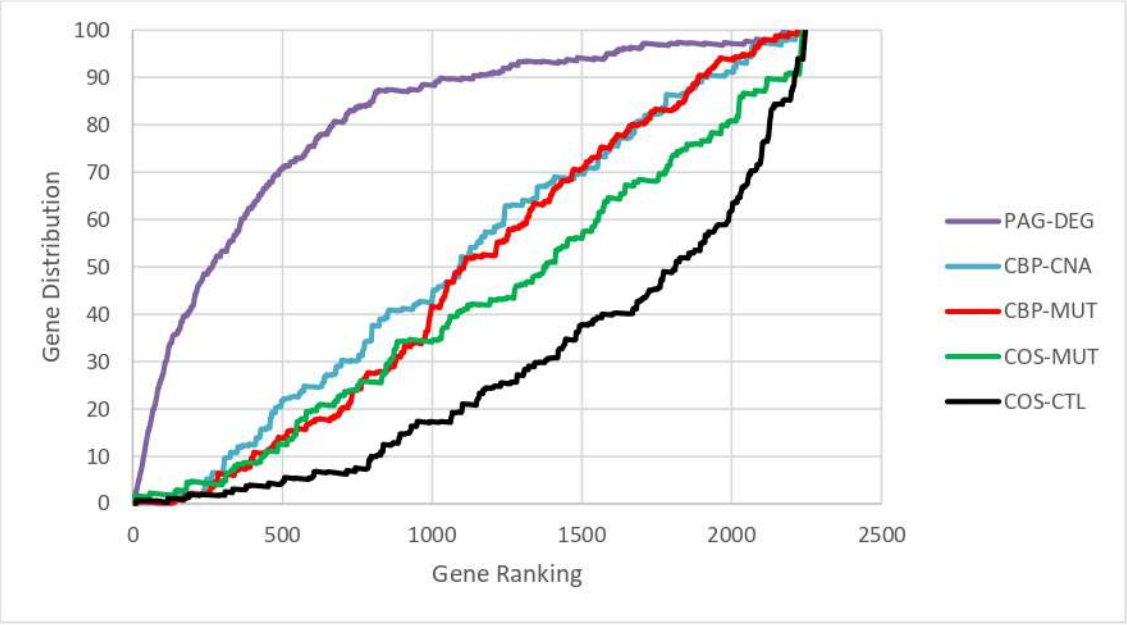


Figure 4. ALL-CTL gene distribution based on each underlying dataset analyzed by GSEA. The gene distribution is the percentage of genes found along the ranked list for each of the five data subsets, including the control data subset COS-CTL.

DISSCUSSION AND CONCLUSIONS

We determine three key conclusions. First, we demonstrate that an iterative network-based gene ranking on different -omic data modalities is biologically meaningful and that traditional network gene ranking can be refined through an iterative process. Our workflow case study using glioblastoma shows how users can combine or divide data according to modalities to generate combined or separate rankings. As seen in **Figure 2**, numerous genes discovered through the iterative Prioritization with pathway and network context are supported by statistically significant literature co-citations with glioblastoma. In addition, some genes added through the expansion process (not found in the initially compiled data subsets) are also significant in the literature, such as VCAM1, SRC, and SUMO1. As seen in **Suppl. Table 1**, the third-ranked gene, MYC, is an expanded gene not found in the initial data subsets, demonstrating the capacity of the iterative expansion process to determine significant genes not already present in the initial gene subsets. Second, gene rankings can reveal essential biology not reflected in the popularity of literature references. For example, there are only 22 co-citations between glioblastoma and the gene UBC (**Suppl. Table 1**).

Furthermore, UBC is not identified in BEERE's statistically significant co-citation network, as seen in **Figure 2**. However, based on PAGER analysis, UBC is present in various biological pathways, such as the RAC1 pathway and ErbB signaling pathway, and is therefore closely linked to Glioblastoma genes established in the literature, such as MYC, TP53, EGFR, and MDM2. Third, we acknowledge that integrating datasets of the same modality from various different sources can yield biologically significant results. However, incorporating other modalities of data can cause ranking signals to be diluted. Therefore, to discover subtle signals, we recommend separating each modality, for example, by gene mutations, epigenetic controls, or differential expressions, to procure modality-specific biological insight. In **Suppl. Table 1**, both COS-MUT and CBP-MUT are data subsets containing mutated genes significant to glioblastoma. However, it is clear that each data subset contains different genes, and therefore both contribute towards the final gene list. In **Figure 4**, we show that although gene prioritization can concentrate the COS-CTL genes – known to have no relation to glioblastoma – near the lower end of the ranking, it is indeed the case that the differentially expressed genes PAG-DEG are diluting the signal of the other data subsets. This can also be seen in **Suppl. Table 1**, the number of PAG-DEG genes in the top 20 ALL-FNL exceeds those in the top 20 ALL-FNL ranking from the other data subsets.

Using glioblastoma as a case study, we demonstrate how to extract biologically significant genes related to complex human diseases with GRIPP. To generalize our approach, one can compile a database containing refined gene lists for all human diseases, thus aiding researchers in future applications, such as prioritizing genes in drug discovery or choosing biomarker candidates. The generalization of our workflow could be accomplished by developing an API involving the PAGER Web APP [28] in combination with WINNER – a streamlined implementation of BEERE – allowing for iterative gene expansion and prioritization in a high-throughput fashion.

In the context of our study, we posit that while network-based gene ranking methodologies may introduce certain intrinsic biases, these biases are notably lesser in magnitude compared to approaches that do not employ network-based gene ranking. Although incorporating diverse data modalities into the ranking process can introduce some bias, it is essential to recognize that the top-ranking genes in these analyses are predominantly upstream regulators or signaling molecules. These hub genes significantly impact downstream genetic activity and are, therefore, more pertinent to the biological phenomena under investigation than their downstream counterparts.

There are limited reports on mitigating factors that may cause gene prioritization biases. For example, [29] used normalized gene ranking scores based on phenotype term popularity, ensuring

genes with less data set coverage are not penalized in their ranking score. However, since our method is based on high-throughput protein-protein interaction (PPI) data sets—with high coverage and incorporated PPI quality as weights—and not based on literature or data set coverage, the reported bias mitigation method does not apply to our work.

Different methods and data modalities may be necessary depending on how a user wishes to use prioritized gene data. For applications geared toward drug development, we advocate for integrating multiple data modalities to enhance the comprehensiveness and robustness of the analysis. On the other hand, in biomarker development, we recommend aligning the modality of the assay platforms with that of the gene sets under consideration to maximize contextual relevance.

While our reported case study in building and refining candidate gene list is the most comprehensive description of addressing network-based gene ranking problem so far, we recognize that the method may have limited utility when the coverage of protein-to-protein interaction coverage is sparse, e.g., for infectious diseases where host-pathogen interactions data are unknown for COVID-19, or the interaction data quality is poorly described, e.g., the use of binary protein-protein interaction representation. The limitation may be mitigated with the experimental collection of high-quality protein/gene association data in collaboration with experimental biologists.

ACKNOWLEDGEMENTS

Both authors thank the generous partial support of the UAB startup fund for this work. EG performed the data analysis using GSEA, drafted the manuscript, and conducted the data processing with PAGER and BEERE. JYC conceived the study, developed the data analysis plan, oversaw the implementation process by providing feedback throughout the research process, and revised the final manuscript before publication.

CONFLICT OF INTEREST

The authors have no financial conflicts of interest to disclose for this work.

REFERENCES

1. Luo, Y., G. Riedlinger, and P. Szolovits, *Text mining in cancer gene and pathway prioritization.* Cancer Inform, 2014. **13**(Suppl 1): p. 69-79.
2. Huang, H., J. Li, and J.Y. Chen, *Disease gene-fishing in molecular interaction networks: a case study in colorectal cancer.* Annu Int Conf IEEE Eng Med Biol Soc, 2009. **2009**: p. 6416-9.
3. Liberzon, A., et al., *The Molecular Signatures Database (MSigDB) hallmark gene set collection.* Cell Syst, 2015. **1**(6): p. 417-425.
4. Zhang, F. and J.Y. Chen, *Discovery of pathway biomarkers from coupled proteomics and systems biology methods.* BMC Genomics, 2010. **11 Suppl 2**: p. S12.
5. Wolf, D.M., et al., *Redefining breast cancer subtypes to guide treatment prioritization and maximize response: Predictive biomarkers across 10 cancer therapies.* Cancer Cell, 2022. **40**(6): p. 609-623 e6.
6. Behan, F.M., et al., *Prioritization of cancer therapeutic targets using CRISPR-Cas9 screens.* Nature, 2019. **568**(7753): p. 511-516.
7. Failli, M., J. Paananen, and V. Fortino, *ThETA: transcriptome-driven efficacy estimates for gene-based TArget discovery.* Bioinformatics, 2020. **36**(14): p. 4214-4216.
8. Nguyen, T., et al., *WINNER: A network biology tool for biomolecular characterization and prioritization.* Front Big Data, 2022. **5**: p. 1016606.
9. Gill, N., S. Singh, and T.C. Aseri, *Computational disease gene prioritization: an appraisal.* J Comput Biol, 2014. **21**(6): p. 456-65.
10. Sjogren, A., E. Kristiansson, M. Rudemo, and O. Nerman, *Weighted analysis of general microarray experiments.* BMC Bioinformatics, 2007. **8**: p. 387.
11. Li, J., X. Zhu, and J.Y. Chen, *Discovering breast cancer drug candidates from biomedical literature.* Int J Data Min Bioinform, 2010. **4**(3): p. 241-55.

12. Li, J., X. Zhu, and J.Y. Chen. *Mining disease-specific molecular association profiles from biomedical literature: a case study*. in *Proceedings of the 2008 ACM symposium on Applied computing*. 2008.
13. Chen, J.Y. and A.Y. Sivachenko, *Data mining in protein interactomics.* Engineering in Medicine and Biology Magazine, IEEE, 2005. **24**(3): p. 95-102.
14. Chen, J.Y., C. Shen, and A.Y. Sivachenko, *Mining Alzheimer disease relevant proteins from integrated protein interactome data.* Pac Symp Biocomput, 2006: p. 367-78.
15. Chen, J.Y., M. Piquette-Miller, and B.P. Smith, *Network medicine: finding the links to personalized therapy*. Clin Pharmacol Ther, 2013. **94**(6): p. 613-6.
16. Huang, H., J. Li, and J.Y. Chen, *Disease gene-fishing in molecular interaction networks: a case study in colorectal cancer.* Conference proceedings : ... Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, 2009. **2009**: p. 6416-9.
17. Frasca, M., et al., *A GPU-based algorithm for fast node label learning in large and unbalanced biomolecular networks.* BMC Bioinformatics, 2018. **19**(Suppl 10): p. 353.
18. Anderson, C.A., et al., *Data quality control in genetic case-control association studies*. Nat Protoc, 2010. **5**(9): p. 1564-73.
19. Yue, Z., C.D. Willey, A.B. Hjelmeland, and J.Y. Chen, *BEERE: a web server for biomedical entity expansion, ranking and explorations.* Nucleic Acids Res, 2019. **47**(W1): p. W578-W586.
20. Marx, D., et al., *Proteomics and Metabolomics for AKI Diagnosis.* Semin Nephrol, 2018. **38**(1): p. 63-87.
21. Yue, Z., et al., *PAGER 2.0: an update to the pathway, annotated-list and gene-signature electronic repository for Human Network Biology.* Nucleic Acids Res, 2018. **46**(D1): p. D668-D676.
22. Yue, Z., et al., *PAGER: constructing PAGs and new PAG-PAG relationships for network biology.* Bioinformatics, 2015. **31**(12): p. i250-7.
23. Yue, Z., et al., *PAGER-CoV: a comprehensive collection of pathways, annotated gene-lists and gene signatures for coronavirus disease studies.* Nucleic Acids Res, 2021. **49**(D1): p. D589-D599.
24. Wirsching, H.G., E. Galanis, and M. Weller, *Glioblastoma.* Handb Clin Neurol, 2016. **134**: p. 381-97.
25. Forbes, S.A., et al., *COSMIC: somatic cancer genetics at high-resolution.* Nucleic Acids Res, 2017. **45**(D1): p. D777-D783.
26. Gao, J., et al., *Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal.* Sci Signal, 2013. **6**(269): p. pl1.
27. Powers, R.K., et al., *GSEA-InContext: identifying novel and common patterns in expression experiments.* Bioinformatics, 2018. **34**(13): p. i555-i564.
28. Yue, Z., R. Slominski, S. Bharti, and J.Y. Chen, *PAGER Web APP: An Interactive, Online Gene Set and Network Interpretation Tool for Functional Genomics*. Front Genet, 2022. **13**: p. 820361.
29. Cornish, A.J., A. David, and M.J.E. Sternberg, *PhenoRank: reducing study bias in gene prioritization through simulation*. Bioinformatics, 2018. **34**(12): p. 2087-2095.

Supplemental Files

Suppl. Table 1. Top 20 GBM network-ranked genes sorted by the overall ranking score and their ranks in each GBM data subsets.

Suppl. Table 2: A summary of the number of genes throughout each step of the GRIPP workflow based on network and pathway context.

Note: Supplemental Excel Files containing the raw data from the data subsets, the iterative prioritization based on network and pathway context, and enrichment score results from GSEA can be found in our GitHub: https://github.com/aimed-lab/gene-prioritization

Suppl. Table 1. Top 20 GBM network-ranked genes sorted by the overall ranking score and their ranks in each GBM data subsets. Ranking scores were generated from the BEERE software (see methods) and the PubMed co-citation statistics was also listed.

Gene Name	BEERE Ranking Score	Rank for each Gene Set					PubMed Co-citation with GBM
		ALL-FNL	PAG-DEG	COS-MUT	CBP-MUT	CBP-CNA	
UBC	626.9	1	1	*	*	*	22
APP	352.6	2	2	*	*	*	44
MYC[+]	204.9	3	*	*	*	*	496
TP53	198.9	4	*	1	*	2	796
ESR1	173.0	5	*	2	*	*	15
FN1	155.3	6	3	3	*	1	29
SRC[+]	137.1	7	*	*	*	*	282
SUMO1[+]	133.9	8	*	*	*	*	7
EP300	127.9	9	*	*	*	8	16
EGFR	127.4	10	5	4	1	3	2894
CREBBP	126.0	11	*	6	*	10	11
VCAM1[+]	124.8	12	*	*	*	*	32
HDAC1	124.7	13	*	*	*	4	40
CALM3	123.6	14	4	*	*	*	2
UBA52	117.8	15	6	*	*	*	2
HSP90AA1	110.8	16	8	*	*	*	5
MAPK1[+]	110.7	17	*	*	*	*	45
PIK3R1	109.0	18	*	5	2	5	30
CTNNB1[+]	104.9	19	*	*	*	*	92
MDM2	103.2	20	7	*	3	*	232

[+]Genes not present in the initial data subsets, and were added through the iterative prioritization based on pathway and network context.

Suppl. Table 2: A summary of the number of genes throughout each step of the GRIPP workflow based on network and pathway context.

Description	Number of Genes
Number of initial candidate genes	1857
Significant genes ($p < 0.05$) after BEERE ranking	242
Genes discovered through expansion (1st iteration)	591
Significant expanded genes after BEERE prioritization (1st iteration)	80
Significant original genes after BEERE prioritization (1st iteration)	214
Total significant genes after 1st iteration	294
Genes discovered through expansion (2nd iteration)	591
Significant expanded genes after BEERE prioritization (2nd iteration)	290
Significant original genes after BEERE prioritization (2nd iteration)	13
Total significant genes after 2nd iteration	303