# RITAN: rapid integration of term annotation and network resources

Michael T. Zimmermann[1,2,3], Brian Kabat[3], Diane E. Grill[3], Richard B. Kennedy[4] and Gregory A. Poland[4]

[1] Bioinformatics Research and Development Laboratory, Genomic Sciences and Precision Medicine Center, Medical College of Wisconsin, Milwaukee, WI, USA
[2] Clinical and Translational Sciences Institute, Medical College of Wisconsin, Milwaukee, WI, USA
[3] Division of Biomedical Statistics and Informatics, Department of Health Sciences Research, Mayo clinic, Rochester, MN, USA
[4] Mayo Clinic Vaccine Research Group, Mayo Clinic, Rochester, MN, USA

## ABSTRACT

**Background:** Identifying the biologic functions of groups of genes identified in high-throughput studies currently requires considerable time and/or bioinformatics experience. This is due in part to each resource housed within separate databases, requiring users to know about them, and integrate across them. Time consuming and often repeated for each study, integrating across resources and merging with data under study is an increasingly common bioinformatics task.

**Methods:** We developed an open-source R software package for assisting researchers in annotating their genesets with functions, pathways, and their interconnectivity across a diversity of network resources.

**Results:** We present rapid integration of term annotation and network resources (RITAN) for the rapid and comprehensive annotation of a list of genes using functional term and pathway resources and their relationships among each other using multiple network biology resources. Currently, and to comply with data redistribution policies, RITAN allows rapid access to 16 term annotations spanning gene ontology, biologic pathways, and immunologic modules, and nine network biology resources, with support for user-supplied resources; we provide recommendations for additional resources and scripts to facilitate their addition to RITAN. Having the resources together in the same system allows users to derive novel combinations. RITAN has a growing set of tools to explore the relationships within resources themselves. These tools allow users to merge resources together such that the merged annotations have a minimal overlap with one another. Because we index both function annotation and network interactions, the combination allows users to expand small groups of genes using links from biologic networks—either by adding all neighboring genes or by identifying genes that efficiently connect among input genes—followed by term enrichment to identify functions. That is, users can start from a core set of genes, identify interacting genes from biologic networks, and then identify the functions to which the expanded list of genes contribute.

**Conclusion:** We believe RITAN fills the important niche of bridging the results of high-throughput experiments with the ever-growing corpus of functional annotations and network biology resources.

**Availability:** Rapid integration of term annotation and network resources is available as an R package at github.com/MTZimmer/RITAN and BioConductor.org.

## INTRODUCTION

High-throughput technologies are enabling systems-level assays for an increasing diversity of applications, which is revealing previously unknown genetic underpinnings of many diseases and phenotypes. It is common for researchers to identify a list of potentially interesting genes, many of which have no known relationship to the condition studied. In this situation, it is challenging to identify the functions these genes contribute to, their precise mechanisms, and the role each gene plays in orchestrating those functions. Further, it is necessary to estimate how many of the prioritized genes may be chance associations due to biologic variability. Many data systems and computational tools have been developed to aid researchers in addressing these critical challenges; however, to better understand how genes interact with each other to achieve a common function, the comprehensive integration of multiple annotation and network biology resources typically remains a data wrangling venture re-invented for each research project.

Many resources have been developed to standardize and codify terms or descriptive annotations for the functions of genes or their encoded products. Ontologies such as gene ontology (GO) (*Ashburner et al., 2000*) aim to describe each gene specifically and comprehensively. Pathway databases such as PID (*Schaefer et al., 2009*), KEGG (*Kanehisa et al., 2012*), and others, gather sets of genes participating in a particular function that has defined initiation points and outcomes and the topology that connects them. A third approach is to identify which genes are co-regulated across conditions (functional modules), as exemplified by Hallmarks in MSigDB (*Subramanian et al., 2005*) or blood transcription modules (*Li et al., 2014*). Resources like DisGeNet index relationships between genes and diseases (*Pinero et al., 2015*). Broadening the concept of a pathway, network biology resources including HPRD (*Prasad, Kandasamy & Pandey, 2009*), CCSB (*Rolland et al., 2014*), and STRING (*Szklarczyk et al., 2011*) seek a more comprehensive view—how genes, proteins, and potentially other molecules broadly interact with one another. Each of these resource types has its own strengths and limitations. Therefore, in order to gain the most comprehensive understanding of how a group of genes relate to each other, and what biologic functions they may define or contribute to, data integration is necessary.

We have generated an open-source and publically available software package, RITAN, for the Rapid Integration of Term Annotation and Network resources. For term annotation or enrichment analysis, our system currently indexes 16 resources spanning GO, pathways, and immunologic modules, and facilitates annotation of a list of genes (e.g., those differentially expressed). Users may query one resource at a time or multiple resources together, allowing false discovery to be accounted for across all resources used. For network biology, nine resources are indexed and confidence-score filtering is enabled when available. See our vignettes within the package for the full list of resources. Users may add resources using standard file formats to extend the depth and breadth

of coverage. By using these resources and adding additional resources of interest, analyses can be more efficient, standardized, and reproducible by minimizing manual steps and keeping resources indexed in a common and accessible system. We have used this integrated network resource in our previous studies (*Haralambieva et al., 2016a*, *2016b*; *Kennedy et al., 2016*; *Zimmermann et al., 2016*) to provide an organizational framework to unify results from multiple omics data (mRNA, protein, DNA methylation), yielding an integrated signature that is more informative than each alone.

## METHODS

Rapid integration of term annotation and network resource (RITAN) is implemented as a package in the R programing language (*R Core Team, 2014*). It leverages multiple existing packages, extending their utility, including igraph (*Csardi & Nepusz, 2006*) and STRINGdb. Enrichment analysis currently uses the hypergeometric test.

Reproducible analysis examples are available in package vignettes, which are also available in our Supplemental Data. We also downloaded normalized breast invasive carcinoma mRNA-Seq gene expression data from the Cancer Genome Atlas (TCGA) pan-cancer analysis (*Ding et al., 2018*) for demonstrating selected utilities. We used EdgeR (*McCarthy, Chen & Smyth, 2012*) to calculate differential gene expression. We used annotations for known protein complexes using CORUM, (*Ruepp et al., 2010*) rare diseases using Orphanet, (*INSERM, 1997*) and regulatory programs using NetPath (*Kandasamy et al., 2010*).

We have developed a self-contained executable R-Shiny application to bring some of RITAN's functionality to a broader audience. The app leverages a combination of technologies including the chromium web browser, electron stand-alone server, and RPortable, and built into a self-contained Windows executable using node.js.

## RESULTS

Rapid integration of term annotation and network resource is an enabling tool for identifying the functional associations within datasets generated by high-throughput methods. We have implemented features that help researchers to uncover relationships within their dataset and augmented by closely related genes in biologic networks (Figs. 1A and 1B). We used a study of anti-TREM1 effects on influenza response, as an example (Figs. 1C and 1D). Data for this study is publicly posted on GEO under the accession ID, GSE9988 (*Dower et al., 2008*). The study considered multiple treatments and lists of genes associated with each comparison of those treatments were published. Using RITAN, multiple resources were combined into a single enrichment analysis, allowing more comprehensive false-discovery control. The visualization made by RITAN shows a study design matrix with enrichment heatmap to visually annotate how different treatments activate different responses (Fig. 1C). Differential pathway activity by each treatment is presented visually and numerically. Genes associating with up-regulated responses are further contextualized by identification of their known molecular inter-relationships (Fig. 1D). The network information can be directly mined, used in computation, or visualized. In this case, we have uploaded the integrated network returned by RITAN to Cytoscape for visualization and scaled each gene by the number
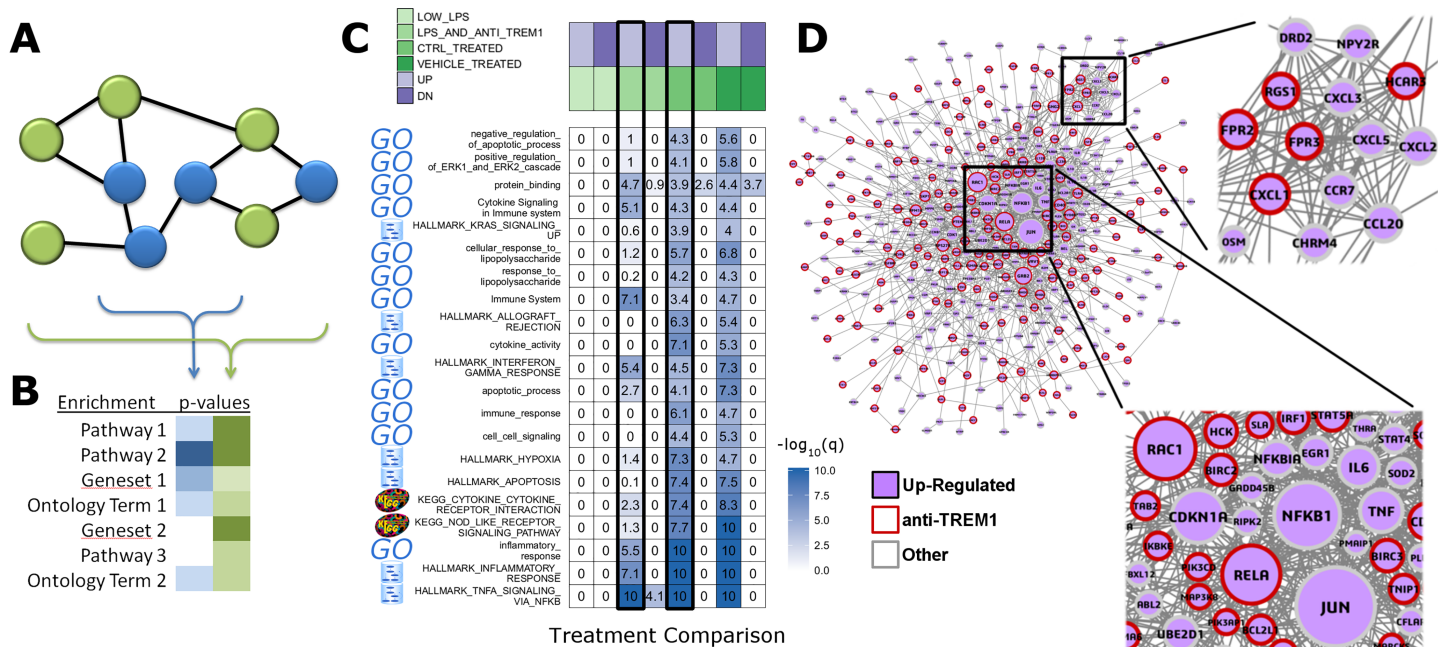
**Figure 1 RITAN facilitates rapid and comprehensive annotation and network integration.** We use the arbitrarily chosen example of GSE9988, a study of anti-TREM1 effects on influenza response. This study considered multiple treatments and lists of genes associated with each comparison of those treatments have been published. (A) From an input gene list (blue nodes), RITAN leverages multiple network resources to identify neighbors (green nodes) and (B) performs integrated term enrichment. We use GSE9988 (*Dower et al., 2008*), a study of anti-TREM1 effects on influenza response, as an example. Using RITAN, multiple resources were combined into a single enrichment analysis with results presented as a heatmap where greater intensity indicates greater statistical significance. The values in the heatmap are the −log$_{10}$($q$-value) with values capped at 10 for visualization. To better use the color scale. Adding a study design matrix above the enrichment heatmap (also by RITAN) visually reveals that significant enrichment scores are defined by the up-regulated gene lists and that they capture aspects of inflammation, innate immunity, cytokine signaling, and interferon signaling. Considering the two comparisons emphasized by a black border, LPS and anti-TREM1-antibody vs. control-treated, there are many terms and pathways exhibiting significant differences in enrichment. (C) RITAN also integrates across multiple network biology resources. The genes associating with up-regulated responses in Fig. 1B are further contextualized by identification of their known inter-relationships. The network information can be directly mined, used in computation, or visualized. In this case, we have uploaded the integrated network returned by RITAN to Cytoscape for visualization. We have scaled each gene by the number of interactions they have as an indication of how "central" they are to response. The network facilitates hypothesis generation by visually displaying the differential association with anti-TREM1 response (indicated by red outline) with different NFkB subunits, differential Akt regulation via JUN, PTEN, GRB2, etc., and different neutrophil chemotactic signals (e.g., FPR3 or CXCL3). Thus, the network is a hypothesis generating tool, focused by first identifying the most relevant subset of results from our integrated enrichment analysis. (D) RITAN integrates network biology resources, which can be directly visualized or imported into Cytoscape. In this example, JUN, NFKB1, TNF, and multiple cytokines were not differentially expressed, but their interactions with the differentially expressed genes were identified, implicating their activity in cellular responses. Full-size ◱ DOI: 10.7717/peerj.6994/fig-1

of interactions they have as an indication of how "central" they are to response. The network facilitates hypothesis generation by visually displaying the differential association with anti-TREM1 response with different NFkB subunits, differential Akt regulation via JUN, PTEN, GRB2, etc., and different neutrophil chemotactic signals (e.g., FPR3 or CXCL3). Thus, the network is a hypothesis generating tool, focused by first identifying the most relevant subset of results from our integrated enrichment analysis. Through data within biologic networks, genes that were not differentially expressed, but closely interact with differentially expressed genes, can be identified, implicating their potential activity change despite no mRNA expression change.

In addition to facilitating rapid functional annotation, RITAN includes tools to analyze the relationships between resources. For example, many diseases are contributed to by
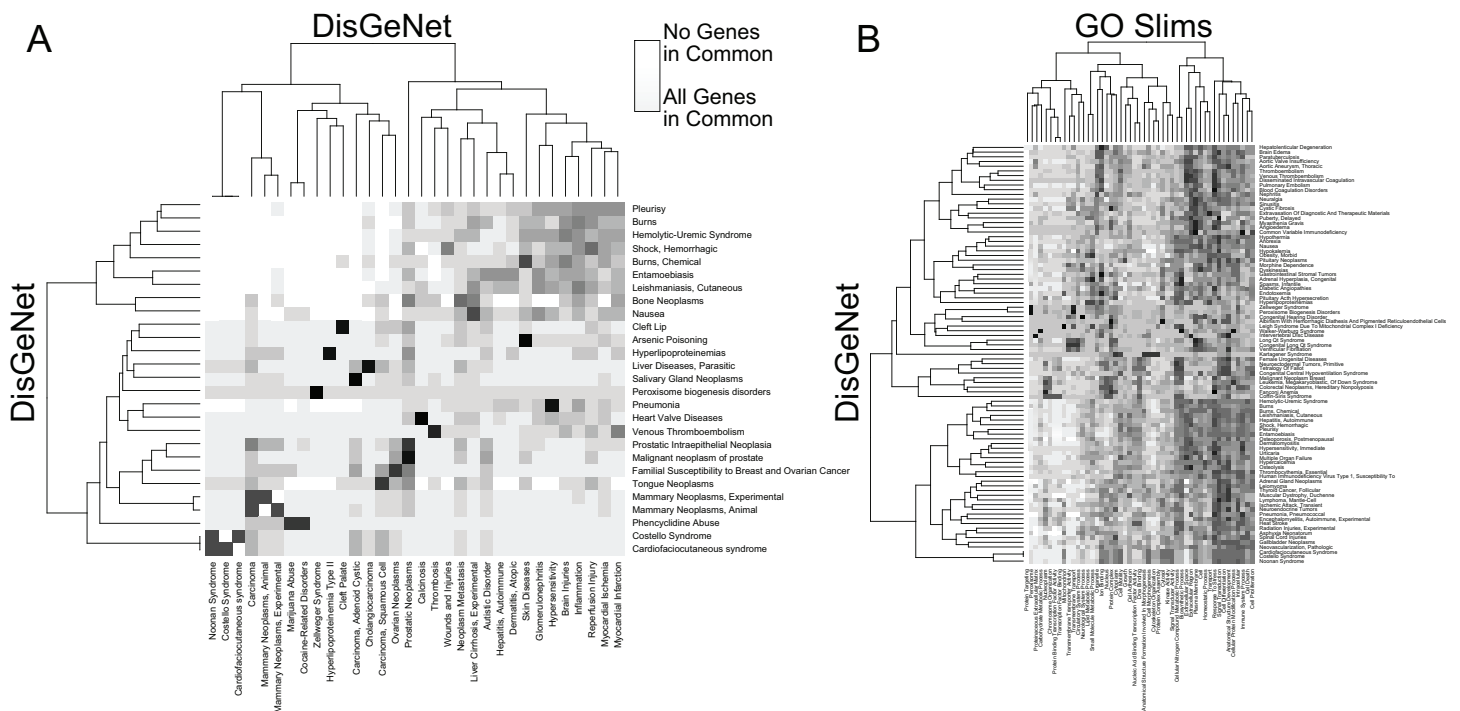
**Figure 2 RITAN facilitates comparing resource similarities.** (A) Within the DisGeNet resource, we compare the overlap of genes annotated to each disease. A disease is shown if it shares at least an 80% overlap with another disease, ignoring self-overlap. (B) Between two resources, DisGeNet, and Go Slim, we show interactions among terms when any pair shares at least 95% overlap. Full-size ◨ DOI: 10.7717/peerj.6994/fig-2

similar sets of genes (Fig. 2A). Additionally, many diseases are characterized by alteration of specific biologic processes or molecular functions (Fig. 2B). By identifying how resources overlap (fraction of shared genes), we are able to combine the resources in a way that reduces the overlap. Once reduced to a set of terms or pathways that are more independent, enrichment analysis can be run on the combined set of terms and using a more accurate false discovery adjustment. Without reducing overlapping terms, false discovery adjustment across resources will be overly conservative. Additionally, having many resources in the same system allows exploration of relationships among them, and making novel combinations.

We use RITAN as a simple knowledge management system that facilitates data annotation and hypothesis exploration—activities that are nor supported by other tools or are challenging to use programmatically. As additional demonstrations of the utility enabled by RITAN, we investigated differential gene expression profiles between breast cancer subtypes from TCGA. This analysis is to demonstrate an approach to data annotation and exploration enabled by RITAN, rather than a definitive analysis of this specific and well-studied dataset. While we distribute many resources with RITAN, other resource licenses prohibit redistribution. We provide recommendations for additional resources (see our GitHub page; we plan to implement an automated download and setup script). Our use of simple and accepted file formats simplifies the process of adding resources. Leveraging additional resources, we annotated the differentially expressed genes between subtypes, first identifying which of them are shared between specific subtypes,
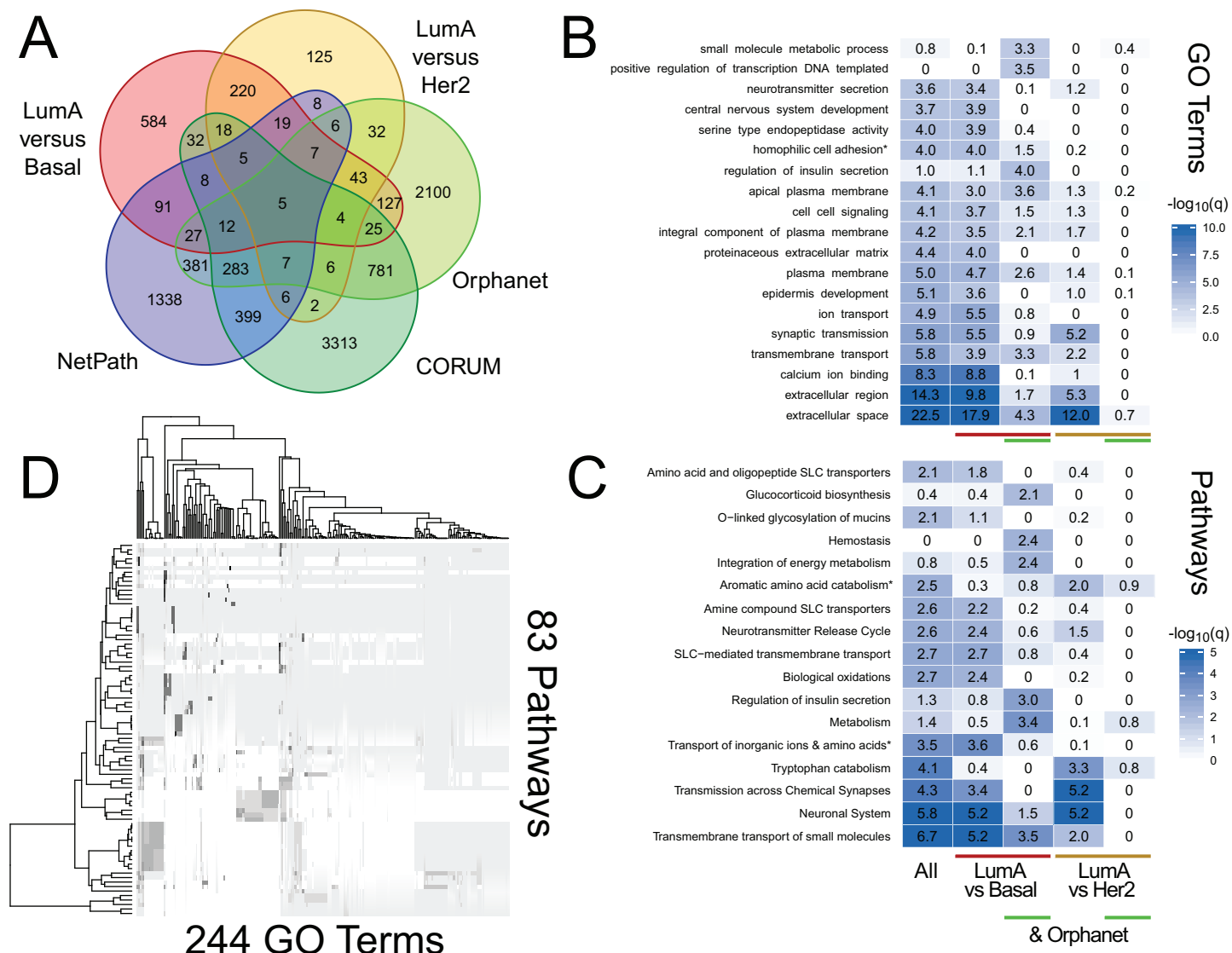
**Figure 3 RITAN facilitates hypothesis generation and exploration.** We used TCGA BRCA differential gene expression data and RITAN to explore specific hypotheses. (A) We annotated differentially expressed genes among luminal A (LumA), Basal, and HER2+ subtypes by their presence in protein complexes, rare disease definitions, and known regulatory programs. The overlapping genes for each combination of annotations indicates additional hypotheses to explore. We used a five-way Venn diagram to show how many genes have each annotation. (B) An example of an additional hypothesis is how functional term enrichment may differ, not only between disease subtypes, but also by contribution to rare disease definitions. We show enrichment plots by subgroup formatted similar to Fig. 1. (C) Additionally, how pathway activation may differ. (D) Similar to Fig. 2, RITAN can be used to explore relationships among resources and to filter each resource to a unique subset. The analysis script for this figure is short (93 lines including comments and plotting; available in Supplemental Data), emphasizing the flexibility and simplicity afforded by RITAN. Full-size ◧ DOI: 10.7717/peerj.6994/fig-3

part of known protein complexes, define rare diseases, or have clearly defined regulatory programs (Fig. 3A). Multiple protein complexes have differentially expressed genes between disease subtypes including SNARE and the chromosomal passenger complex. Both of these complexes have known roles in carcinogenesis and breast cancer (*Kabisch et al., 2015*; *Meng & Wang, 2015*). Multiple rare disease genes are differentially expressed and contribute to precursor B- and T-cell acute lymphoblastic leukemia, among others.

Next, we considered GO term (Fig. 3B) and Reactome (*Croft et al., 2011*) pathway (Fig. 3C) enrichment among genes differentially expressed between disease subtypes, and for the further subset of differentially expressed genes that contribute to rare diseases. These annotations highlight the role of metabolic differences among disease subtypes and that these functional overlap rare disease mechanisms, complementing the use of rare disease annotations for prioritizing functional somatic variants (*Ma et al., 2017*). Finally, we compared GO terms to Reactome pathways to identify which were shared and which were unique. There are 5,231 GO terms with a one-to-one mapping to 1,607 Reactome pathways (100% of genes shared), and 83 GO terms that highly overlap 244 Reactome pathways (95% of genes shared; Fig. 3D). Further vignettes are available within the package and are included in our Supplemental Data. These analyses among gene annotation resources facilitate hypothesis generation and exploration.

## DISCUSSION

Enrichment analysis has become commonplace in bioinformatics and biostatistics; however, current tools have limitations. First, most resources are distributed independently of each other and integrating across them is left to individual users. Second, each enrichment or pathway analysis is typically performed independently, preventing false discovery adjustment across all tests performed. Further, while web servers are convenient for small numbers of queries, programmatic solutions are an advantage for data analysis workflows and scripting for reproducible research.

Enrichment analysis is a statistical method wherein an input list of genes is compared to a term, function, or pathway definition. The definition is usually a pre-existing list of genes curated to describe a specific cellular function or observation. The number of genes in the pathway is recorded and combined with the size of the input list, the size of the term or pathway, and the total number of genes that were assayed (e.g., all human protein coding genes expressed in the tissue of interest), an overrepresentation statistic is computed. When this statistic is significant (e.g., occurring with a probability less than 0.01 by chance and after false discovery correction), the term is typically ascribed to the input list and we describe the function as "enriched for" within the input list. Thus, defining all three areas is critical: the input list, the term definitions, and the universe of possible gene results.

We encourage users to decide upon an analysis strategy prior to running their analysis, including consideration of which resources are most appropriate for their dataset/experiment, statistical significance thresholds, background geneset, etc. The ease with which RITAN facilitates multi-resource query may lead users to "try one more test," leading to an increased number of hypothesis tests made that may not be accounted for by multiple testing correction. To prevent this and appropriately maintain false-discovery correction, we encourage users to "include one more resource"—to add an additional resource to a single query in RITAN.

Other tools that fill similar roles to RITAN exist, but each within a specific niche, using specific data sources, or having additional input requirements. Within the R framework, GAGE (*Luo et al., 2009*) provides access to GO and KEGG database for multiple species. Reactome offers related services, but annotations are specific to their database.

Web servers for enrichment analysis have been developed including WebGestalt (*Wang et al., 2013*) and DAVID (*Da Huang, Sherman & Lempicki, 2009*). The flexibility of RITAN and ability to integrate and explore relationships among both geneset and network resources, distinguishes RITAN from existing software.

RITAN allows annotation integration across many publically available resources; thus, it facilitates rapid development of novel hypotheses about the potential functions achieved by prioritized genes and multiple-testing correction across all resources used. The annotations from RITAN can be leveraged by existing down-stream analysis tools, such as Cytoscape (*Shannon et al., 2003*) or igraph (*Csardi & Nepusz, 2006*), in order to provide greater visualization and network-analytic power. As illustrated in Fig. 1, the two major features of RITAN are an integrated system for term enrichment and connections among genes via network biology, facilitating reproducible research and hypothesis generation. For analyses across resources, it is important to consider how similar the resources are to each other. Similarities may occur within a pathway resource or even between resource types. For example, within the KEGG pathway resource, JAK/STAT, and MAPK signaling share SOS, RAS, RAF, and other genes. Ontology, pathway, and network information can be applied to multiple entities: gene-centric, protein-centric, or term-centric. For example, the GO term GO:1901184 marks the process "regulation of ERBB signaling pathway" and genes are annotated with this term when they are involved in the process, while KEGG's hsa04012 pathway indexes the relationships between genes participating in the "ERBB Signaling Pathway." In this particular case, the GO term is conceptually the same as the pathway, but without topology. Ontologies aim for greater specificity for each gene while pathways aim for greater resolution on a particular process or function. If similar terms, such as the GO term for ERBB signaling and the KEGG pathway for ERBB signaling, are included in the same analysis, false-discovery correction may be more stringent than is needed. Thus, functions within RITAN to identify, and filter resources base on semi-redundant or highly overlapping concepts, will assist the interpretation of data derived from high-throughput sequencing.

We also distinguish between a discovery *p*-value and an annotation *p*-value. For instance, when calculating differentially expressed genes, a discovery *p*-value is generated and is one of the criteria for calling a gene differentially expressed. False-discovery or other statistical controls should be used at this stage. After discovering differentially expressed genes, RITAN can be a powerful tool to assist interpreting those results. The *p*-values generated in annotation have a different interpretation and role from discovery. Researchers should always be aware of weather they are using a statistical test for discovery (inference) or annotation. Iterating between discovery and annotation should be strictly avoided or explicitly identified—iterating can be a powerful discovery approach, but it can quickly lead to over-fitting. Statistical frameworks for combined feature selection and classification exist and could be adapted to implement an iterative approach in this setting.

We have active plans to expand RITAN's capabilities by adding additional statistical testing methods (*Subramanian et al., 2005*), supporting improved resource-reduction methods, and novel metrics for the statistical significance of network overlaps

(*Ding et al., 2016*; *Zimmermann et al., 2016*). Non-programers could benefit from the capabilities of RITAN. Thus, we have developed a self-contained executable application and made the app available for download on RITAN's GitHub page. We plan to further develop both resources—to improve the core features of RITAN and expand our portable application for use by a broader audience. The features currently within RITAN are highly useful and their implementation in a common framework facilitates ease of use and reproducible research.

## CONCLUSION

Rapid integration of term annotation and network resource fills a niche in being a simple gene-centric knowledge management system—to index geneset and network resources—as well as provide multiple tools for resource integration, assessment, enrichment, and exploration. We believe RITAN will be an enabling tool for the interpretation of high-throughput genomic studies.

## ACKNOWLEDGEMENTS

## ADDITIONAL INFORMATION AND DECLARATIONS

### Competing Interests
The authors declare that they have no competing interests.

### Author Contributions
- Michael T. Zimmermann conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, prepared figures and/or tables, authored or reviewed drafts of the paper, approved the final draft.
- Brian Kabat performed the experiments, analyzed the data, contributed reagents/ materials/analysis tools, authored or reviewed drafts of the paper, approved the final draft.
- Diane E. Grill analyzed the data, contributed reagents/materials/analysis tools, authored or reviewed drafts of the paper, approved the final draft.
- Richard B. Kennedy authored or reviewed drafts of the paper, approved the final draft, provided feedback on application areas.
- Gregory A. Poland authored or reviewed drafts of the paper, approved the final draft, provided feedback on application areas.

## Data Availability

The following information was supplied regarding data availability:

Data is available at GitHub: https://github.com/MTZimmer/RITAN.

## Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.6994#supplemental-information.

## REFERENCES

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM, Sherlock G. 2000. Gene ontology: tool for the unification of biology. *Gene Ontology Consortium. Nature genetics* **25(1)**:25–29 DOI 10.1038/75556.

Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L, Caudy M, Garapati P, Gopinath G, Jassal B, Jupe S, Kalatskaya I, Mahajan S, May B, Ndegwa N, Schmidt E, Shamovsky V, Yung C, Birney E, Hermjakob H, D'Eustachio P, Stein L. 2011. Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Research* **39(Database)**:D691–D697 DOI 10.1093/nar/gkq1018.

Csardi G, Nepusz T. 2006. The igraph software package for complex network research. *InterJournal, Complex Systems* **1695(5)**:1–9.

Da Huang W, Sherman BT, Lempicki RA. 2009. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature Protocols* **4(1)**:44–57 DOI 10.1038/nprot.2008.211.

Ding L, Bailey MH, Porta-Pardo E, Thorsson V, Colaprico A, Bertrand D, Gibbs DL, Weerasinghe A, Huang K-L, Tokheim C, Cortés-Ciriano I, Jayasinghe R, Chen F, Yu L, Sun S, Olsen C, Kim J, Taylor AM, Cherniack AD, Akbani R, Suphavilai C, Nagarajan N, Stuart JM, Mills GB, Wyczalkowski MA, Vincent BG, Hutter CM, Zenklusen JC, Hoadley KA, Wendl MC, Shmulevich I, Lazar AJ, Wheeler DA, Getz G, Caesar-Johnson SJ, Demchok JA, Felau I, Kasapi M, Ferguson ML, Hutter CM, Sofia HJ, Tarnuzzer R, Wang Z, Yang L, Zenklusen JC, Zhang J (Julia), Chudamani S, Liu J, Lolla L, Naresh R, Pihl T, Sun Q, Wan Y, Wu Y, Cho J, DeFreitas T, Frazer S, Gehlenborg N, Getz G, Heiman DI, Kim J, Lawrence MS, Lin P, Meier S, Noble MS, Saksena G, Voet D, Zhang H, Bernard B, Chambwe N, Dhankani V, Knijnenburg T, Kramer R, Leinonen K, Liu Y, Miller M, Reynolds S, Shmulevich I, Thorsson V, Zhang W, Akbani R, Broom BM, Hegde AM, Ju Z, Kanchi RS, Korkut A, Li J, Liang H, Ling S, Liu W, Lu Y, Mills GB, Ng K-S, Rao A, Ryan M, Wang J, Weinstein JN, Zhang J, Abeshouse A, Armenia J, Chakravarty D, Chatila WK, De Bruijn I, Gao J, Gross BE, Heins ZJ, Kundra R, La K, Ladanyi M, Luna A, Nissan MG, Ochoa A, Phillips SM, Reznik E, Sanchez-Vega F, Sander C, Schultz N, Sheridan R, Sumer SO, Sun Y, Taylor BS, Wang J, Zhang H, Anur P, Peto M, Spellman P, Benz C, Stuart JM, Wong CK, Yau C, Hayes DN, Parker JS, Wilkerson MD, Ally A, Balasundaram M, Bowlby R, Brooks D, Carlsen R, Chuah E, Dhalla N, Holt R, Jones SJM, Kasaian K, Lee D, Ma Y, Marra MA, Mayo M, Moore RA, Mungall AJ, Mungall K, Robertson AG, Sadeghi S, Schein JE, Sipahimalani P, Tam A, Thiessen N, Tse K, Wong T, Berger AC, Beroukhim R, Cherniack AD, Cibulskis C, Gabriel SB, Gao GF, Ha G, Meyerson M, Schumacher SE, Shih J, Kucherlapati MH, Kucherlapati RS, Baylin S, Cope L, Danilova L, Bootwalla MS, Lai PH, Maglinte DT, Van Den Berg DJ,

Weisenberger DJ, Auman JT, Balu S, Bodenheimer T, Fan C, Hoadley KA, Hoyle AP, Jefferys SR, Jones CD, Meng S, Mieczkowski PA, Mose LE, Perou AH, Perou CM, Roach J, Shi Y, Simons JV, Skelly T, Soloway MG, Tan D, Veluvolu U, Fan H, Hinoue T, et al. 2018. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell* **173(2)**:305–320.e10 DOI 10.1016/j.cell.2018.03.033.

Ding Z, Zhang X, Sun D, Luo B. 2016. Overlapping community detection based on network decomposition. *Scientific Reports* **6(1)**:24115 DOI 10.1038/srep24115.

Dower K, Ellis DK, Saraf K, Jelinsky SA, Lin L-L. 2008. Innate immune responses to TREM-1 activation: overlap, divergence, and positive and negative cross-talk with bacterial lipopolysaccharide. *Journal of Immunology* **180(5)**:3520–3534 DOI 10.4049/jimmunol.180.5.3520.

Haralambieva IH, Ovsyannikova IG, Kennedy RB, Zimmermann MT, Grill DE, Oberg AL, Poland GA. 2016a. Transcriptional signatures of influenza A/H1N1-specific IgG memory-like B cell response in older individuals. *Vaccine* **34(34)**:3993–4002 DOI 10.1016/j.vaccine.2016.06.034.

Haralambieva IH, Zimmermann MT, Ovsyannikova IG, Grill DE, Oberg AL, Kennedy RB, Poland GA. 2016b. Whole transcriptome profiling identifies CD93 and other plasma cell survival factor genes associated with measles-specific antibody response after vaccination. *PLOS ONE* **11(8)**:e0160970 DOI 10.1371/journal.pone.0160970.

INSERM. 1997. *Orphanet: an online database of rare diseases and orphan drugs.* Paris: French National Institute for Health and Medical Research. *Available at* http://www.orpha.net.

Kabisch M, Lorenzo Bermejo J, Dünnebier T, Ying S, Michailidou K, Bolla MK, Wang Q, Dennis J, Shah M, Perkins BJ, Czene K, Darabi H, Eriksson M, Bojesen SE, Nordestgaard BG, Nielsen SF, Flyger H, Lambrechts D, Neven P, Peeters S, Weltens C, Couch FJ, Olson JE, Wang X, Purrington K, Chang-Claude J, Rudolph A, Seibold P, Flesch-Janys D, Peto J, Dos-Santos-Silva I, Johnson N, Fletcher O, Nevanlinna H, Muranen TA, Aittomäki K, Blomqvist C, Schmidt MK, Broeks A, Cornelissen S, Hogervorst FBL, Li J, Brand JS, Humphreys K, Guénel P, Truong T, Menegaux F, Sanchez M, Burwinkel B, Marmé F, Yang R, Bugert P, González-Neira A, Benitez J, Pilar Zamora M, Arias Perez JI, Cox A, Cross SS, Reed MWR, Andrulis IL, Knight JA, Glendon G, Tchatchou S, Sawyer EJ, Tomlinson I, Kerin MJ, Miller N, Haiman CA, Schumacher F, Henderson BE, Le Marchand L, Lindblom A, Margolin S, Hooning MJ, Hollestelle A, Kriege M, Koppert LB, Hopper JL, Southey MC, Tsimiklis H, Apicella C, Slettedahl S, Toland AE, Vachon C, Yannoukakos D, Giles GG, Milne RL, McLean C, Fasching PA, Ruebner M, Ekici AB, Beckmann MW, Brenner H, Dieffenbach AK, Arndt V, Stegmaier C, Ashworth A, Orr N, Schoemaker MJ, Swerdlow A, García-Closas M, Figueroa J, Chanock SJ, Lissowska J, Goldberg MS, Labrèche F, Dumont M, Winqvist R, Pylkäs K, Jukkola-Vuorinen A, Grip M, Brauch H, Brüning T, Ko Y-D, Radice P, Peterlongo P, Scuvera G, Fortuzzi S, Bogdanova N, Dörk T, Mannermaa A, Kataja V, Kosma V-M, Hartikainen JM, Devilee P, Tollenaar RAEM, Seynaeve C, Van Asperen CJ, Jakubowska A, Lubinski J, Jaworska-Bieniek K, Durda K, Zheng W, Shrubsole MJ, Cai Q, Torres D, Anton-Culver H, Kristensen V, Bacot F, Tessier DC, Vincent D, Luccarini C, Baynes C, Ahmed S, Maranian M, Simard J, Chenevix-Trench G, Hall P, Pharoah PDP, Dunning AM, Easton DF, Hamann U. 2015. Inherited variants in the inner centromere protein (INCENP) gene of the chromosomal passenger complex contribute to the susceptibility of ER-negative breast cancer. *Carcinogenesis* **36(2)**:256–271 DOI 10.1093/carcin/bgu326.

Kandasamy K, Mohan S, Raju R, Keerthikumar S, Kumar GS, Venugopal AK, Telikicherla D, Navarro JD, Mathivanan S, Pecquet C, Gollapudi SK, Tattikota SG, Mohan S, Padhukasahasram H, Subbannayya Y, Goel R, Jacob HK, Zhong J, Sekhar R, Nanjappa V, Balakrishnan L, Subbaiah R, Ramachandra YL, Rahiman A, Prasad TSK, Lin J-X, Houtman JCD, Desiderio S, Renauld J-C, Constantinescu SN, Ohara O, Hirano T, Kubo M, Singh S, Khatri P, Draghici S, Bader GD, Sander C, Leonard WJ, Pandey A. 2010. NetPath: a public resource of curated signal transduction pathways. *Genome Biology* **11(1)**:R3 DOI 10.1186/gb-2010-11-1-r3.

Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Research* **40(D1)**:D109–D114 DOI 10.1093/nar/gkr988.

Kennedy RB, Ovsyannikova IG, Haralambieva IH, Oberg AL, Zimmermann MT, Grill DE, Poland GA. 2016. Immunosenescence-related transcriptomic and immunologic changes in older individuals following influenza vaccination. *Frontiers in Immunology* **7(16)**:450 DOI 10.3389/fimmu.2016.00450.

Li S, Rouphael N, Duraisingham S, Romero-Steiner S, Presnell S, Davis C, Schmidt DS, Johnson SE, Milton A, Rajam G, Kasturi S, Carlone GM, Quinn C, Chaussabel D, Palucka AK, Mulligan MJ, Ahmed R, Stephens DS, Nakaya HI, Pulendran B. 2014. Molecular signatures of antibody responses derived from a systems biology study of five human vaccines. *Nature Immunology* **15(2)**:195–204 DOI 10.1038/ni.2789.

Luo W, Friedman MS, Shedden K, Hankenson KD, Woolf PJ. 2009. GAGE: generally applicable gene set enrichment for pathway analysis. *BMC Bioinformatics* **10(1)**:161 DOI 10.1186/1471-2105-10-161.

Ma M, Wang C, Glicksberg BS, Schadt EE, Li SD, Chen R. 2017. Identify cancer driver genes through shared mendelian disease pathogenic variants and cancer somatic mutations. *Pacific Symposium on Biocomputing* **22**:473–484 DOI 10.1142/9789813207813_0044.

McCarthy DJ, Chen Y, Smyth GK. 2012. Differential expression analysis of multifactor RNA-Seq experiments with respect to biological variation. *Nucleic Acids Research* **40(10)**:4288–4297 DOI 10.1093/nar/gks042.

Meng J, Wang J. 2015. Role of SNARE proteins in tumourigenesis and their potential as targets for novel anti-cancer therapeutics. *Biochimica et Biophysica Acta (BBA)—Reviews on Cancer* **1856(1)**:1–12 DOI 10.1016/j.bbcan.2015.04.002.

Pinero J, Queralt-Rosinach N, Bravo A, Deu-Pons J, Bauer-Mehren A, Baron M, Sanz F, Furlong LI. 2015. DisGeNET: a discovery platform for the dynamical exploration of human diseases and their genes. *Database: The Journal of Biological Databases and Curation* **2015**:bav028 DOI 10.1093/database/bav028.

Prasad TS, Kandasamy K, Pandey A. 2009. Human protein reference database and human Proteinpedia as discovery tools for systems biology. *Methods in Molecular Biology* **577(Suppl 5)**:67–79 DOI 10.1007/978-1-60761-232-2_6.

R Core Team. 2014. *R: a language and environment for statistical computing.* Vienna: The R Foundation for Statistical Computing. *Available at* http://www.R-project.org.

Rolland T, Taşan M, Charloteaux B, Pevzner SJ, Zhong Q, Sahni N, Yi S, Lemmens I, Fontanillo C, Mosca R, Kamburov A, Ghiassian SD, Yang X, Ghamsari L, Balcha D, Begg BE, Braun P, Brehme M, Broly MP, Carvunis A-R, Convery-Zupan D, Corominas R, Coulombe-Huntington J, Dann E, Dreze M, Dricot A, Fan C, Franzosa E, Gebreab F, Gutierrez BJ, Hardy MF, Jin M, Kang S, Kiros R, Lin GN, Luck K, MacWilliams A, Menche J, Murray RR, Palagi A, Poulin MM, Rambout X, Rasla J, Reichert P, Romero V, Ruyssinck E,

Sahalie JM, Scholz A, Shah AA, Sharma A, Shen Y, Spirohn K, Tam S, Tejeda AO, Trigg SA, Twizere J-C, Vega K, Walsh J, Cusick ME, Xia Y, Barabási A-L, Iakoucheva LM, Aloy P, De Las Rivas J, Tavernier J, Calderwood MA, Hill DE, Hao T, Roth FP, Vidal M. 2014. A proteome-scale map of the human interactome network. *Cell* 159(5):1212–1226 DOI 10.1016/j.cell.2014.10.050.

Ruepp A, Waegele B, Lechner M, Brauner B, Dunger-Kaltenbach I, Fobo G, Frishman G, Montrone C, Mewes H-W. 2010. CORUM: the comprehensive resource of mammalian protein complexes—2009. *Nucleic Acids Research* 38(suppl_1):D497–D501 DOI 10.1093/nar/gkp914.

Schaefer CF, Anthony K, Krupa S, Buchoff J, Day M, Hannay T, Buetow KH. 2009. PID: the pathway interaction database. *Nucleic Acids Research* 37(suppl_1):D674–D679 DOI 10.1093/nar/gkn653.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research* 13(11):2498–2504 DOI 10.1101/gr.1239303.

Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES, Mesirov JP. 2005. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America* 102(43):15545–15550 DOI 10.1073/pnas.0506580102.

Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, Minguez P, Doerks T, Stark M, Muller J, Bork P, Jensen LJ, Von Mering C. 2011. The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research* 39(Database):D561–D568 DOI 10.1093/nar/gkq973.

Wang J, Duncan D, Shi Z, Zhang B. 2013. WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013. *Nucleic Acids Research* 41(W1):W77–W83 DOI 10.1093/nar/gkt439.

Zimmermann MT, Oberg AL, Grill DE, Ovsyannikova IG, Haralambieva IH, Kennedy RB, Poland GA. 2016. System-wide associations between DNA-Methylation, gene expression, and humoral immune response to influenza vaccination. *PLOS ONE* 11(3):e0152034 DOI 10.1371/journal.pone.0152034.