

SOFTWARE

Facilitating pathway and network based analysis of RNA-Seq data with *pathlinkR*Travis M. Blimkie¹, Andy An¹, Robert E. W. Hancock^{1*}

REW Hancock Laboratory, Center for Microbial Diseases and Immunity Research, Department of Microbiology and Immunology, University of British Columbia, Vancouver, British Columbia, Canada

* bob@hancocklab.com

OPEN ACCESS

Citation: Blimkie TM, An A, Hancock REW (2024) Facilitating pathway and network based analysis of RNA-Seq data with *pathlinkR*. PLoS Comput Biol 20(9): e1012422. <https://doi.org/10.1371/journal.pcbi.1012422>

Editor: Stacey D. Finley, University of Southern California, UNITED STATES OF AMERICA

Received: March 15, 2024

Accepted: August 19, 2024

Published: September 16, 2024

Copyright: © 2024 Blimkie et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: *pathlinkR* is available to install from the software repository Bioconductor at <https://bioconductor.org/packages/pathlinkR/>. The code and example data is available on the GitHub repository at <https://github.com/hancockinformatics/pathlinkR>, under the GPL 3.0 license.

Funding: This work was supported by funding from the Canadian Institutes for Health Research [grant number FDN 154287] to RH. RH holds a UBC Killam Professorship and previously held a Canada Research Chair. AA is funded by a Canada

Abstract

R package *pathlinkR* is designed to aid transcriptomic analyses by streamlining and simplifying the process of analyzing and interpreting differentially expressed genes derived from human RNA-Seq data. It provides an integrated approach to performing pathway enrichment and network-based analyses, while also producing publication-quality figures to summarize these results, allowing users to more efficiently interpret their findings and extract biological meaning from large amounts of data. *pathlinkR* is available to install from the software repository Bioconductor at <https://bioconductor.org/packages/pathlinkR/>, with support available through the Bioconductor forums. The code, example, and supporting data is available on the GitHub repository at <https://github.com/hancockinformatics/pathlinkR>, under the GPL-3.0 license, where users may report problems or make suggestions using GitHub's issue system.

Introduction

RNA-Seq is one of the most prevalent types of omics methodologies, since it provides detailed insights into the biology underlying a variety of studies [1] through analysis of gene expression changes, and is applicable to any organism with a reference genome. Comparing one or more conditions of interest to a reference or control sample allows for the determination of genes that are undergoing changes in their expression specifically due to the condition or treatment. However, lists of differentially expressed genes (DEGs) by themselves are not sufficient for drawing meaningful conclusions about the mechanisms at work and they must be further analyzed. There are several paths one can choose with two of the most common approaches being pathway enrichment [2] and network analysis [3]; these can be combined to further extract biological meaning. While there are several tools and resources to perform these tasks [4,5], they tend to exist in separate packages, making their integration difficult, particularly for those not sufficiently trained in bioinformatics techniques.

Here we present *pathlinkR*, an R [6] package that provides a streamlined and unified interface to perform some of the most common analyses and visualizations of DEGs, with functions that can assist users to easily perform pathway enrichment and network analysis. It is designed to be straightforward to use and accepts input from the most commonly used primary tools in

Graduate Scholarships Doctoral program. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

a typical analysis pipeline, *DESeq2* [7] and *edgeR* [8]. Both packages produce a table of DEGs, which *pathlinkR* can use without modification for input to its core functions, including visualizations, pathway enrichment, and construction of Protein-Protein Interaction (PPI) networks. These analyses can be performed on a number of databases, including Reactome, MSigDB, KEGG, and InnateDB (previously not available within an R package), which are each consistently formatted for simple incorporation into new analysis pipelines. For each of these analysis types supported by *pathlinkR*, users can generate publication-quality figures to summarize their findings.

Design and Implementation

Direct visualization of DEGs

Two functions in *pathlinkR* can be applied to a table of DEGs, enabling simple visualization and comparison of results; both can take standard data objects produced by *DESeq2* [7] and *edgeR* [8] as input. The first is “*eruption*”, which creates a Volcano plot mapping fold change values on the x axis and statistical significance (negative log of the p value) on the y axis. Multiple refinements allow the resulting plot to be tweaked as desired, such as changing colours, adding a summary subtitle, and plotting fold changes as log-transformed or not. Selected genes can be highlighted by providing a list of gene names/IDs, or even the name of a pathway from which constituent genes are automatically highlighted. Pathways accessed by *pathlinkR* are from Reactome [9], a curated and peer-reviewed database of >2,600 pathways in *Homo sapiens*, as well as thousands of pathways from other model organisms, organized hierarchically based on cell-specific processes.

The second method for DEG visualization is “*plotFoldChange*,” which creates a heatmap with genes as rows and DE comparisons as the columns. This allows comparison of how the same set of genes behave in different comparisons of interest. The genes included in the rows can also be selected by providing a list of gene names/IDs, or the name of a Reactome pathway.

Network analysis

A DEG table can be used as input to “*ppiBuildNetwork*” to create a PPI network, using curated molecular human interaction data from InnateDB [10] (previously not available within the R framework), an IMEX consortium database, and leveraging functions from the *igraph* [11] and *tidygraph* [12] packages. Options for network order including zero, first, or minimum are available to suit the number of genes used as a seed for the network. A zero order network contains only genes or nodes that are input by the user, referred to as seed nodes—no additional genes are added via interactions. First order networks include genes which interact with the seed nodes, called interactors, growing and expanding the network using the available data. A minimum order network starts as first order, but any interactors with only a single connection to a seed node are removed, trimming down the overall network. Once built, this network can be visualized directly within R via “*ppiPlotNetwork*,” which can easily map a variable of interest (e.g. fold change directionality, statistical significance, or node degree) to node colour, and highlight and label key “hub” nodes which are integral in binding the network together. PPI networks can be easily tested for enriched pathways using “*ppiEnrichNetwork*” (it uses the same methodology as the main pathway enrichment function; see the next section for details) and a sub-network consisting of nodes present in a pathway of interest can be plotted by using “*ppiExtractSubnetwork*,” followed by the same “*ppiPlotNetwork*” function to create an updated graph highlighting nodes from the pathway. These methods are specifically designed to work with the results from human RNA-Seq experiments, providing previously unavailable network functionality in the context of analyzing DEGs.

Pathway enrichment

The “*pathwayEnrichment*” function can be used on a list of one or more DEG tables, testing each for enriched pathways, and returning a single table to facilitate simple comparison of results, making pathlinkR inherently scalable to complex experiments. Three methodologies for enrichment are included: The first is simple over-representation analysis, utilized by the popular *ReactomePA* [4] package, which treats genes in a pathway individually when calculating the statistical significance of enrichment. The second is gene-pair over-representation analysis, utilized by the *Sigora* package [13], which treats genes in a pathway as unique pairs to provide less redundant pathway enrichment results. Both options are available when analyzing the Reactome [9] or KEGG [14] human pathway databases, while a second option uses simple over-representation analysis with the Molecular Signatures Database (MSigDB) Hallmark Collection [15] for humans, which are gene sets that represent “specific, well-defined biological states or processes with coherent expression”. Finally, users may also perform gene set enrichment analysis (GSEA) built on the *fgsea* [16] package, on either Reactome pathways or MSigDB Hallmark terms. These methods have been provided to enable users to have more freedom in determining which methodology and database suits their analysis. Unlike most pathway enrichment tools, for “*pathwayEnrichment*” the input DEGs from a single comparison can be treated as a complete set, or split based on the directionality of each gene’s fold change, resulting in pathways assigned as “upregulated” (i.e. enriched by upregulated genes, meaning the pathway overall can be thought of as being activated) or “downregulated.” The output format is shared for all available methods and databases, making it simple to incorporate any of these tests into a larger pipeline.

The table of results from “*pathwayEnrichment*” can be visualized with “*pathwayPlots*” to create a compelling figure which summarizes and compares the enrichment results across multiple DEG comparisons, scaling to experiments with multiple conditions of interest. Each comparison (e.g. Treatment A vs. Control and Treatment B vs. Control) is represented as a column on the x axis, while significant pathways are placed alphabetically as the rows on the y axis. A triangle indicates a given pathway was significantly enriched in one comparison, while its direction (pointing up or down) denotes if the pathway was derived from up- or down-regulated genes; or if there is no separation by directionality, circles are used instead. When using any of the methods or databases available, the pathways on the y axis can be grouped according to higher level categories, such as “Immune System” or “Metabolism,” to increase interpretability. A number of options can be used to tweak and customize the final output, such as only showing pathways from certain categories, or adding the number of DEGs below each comparisons’ name on the x axis.

Pathway-based networks

The pathway enrichment results from a single DE comparison can be visualized as a network using the “*pathnet*” family of functions. Each pathway becomes a node in a network, with interactions or edges defined by the number of shared genes between two pathways. Since many genes can have multiple functions and play a role in different pathways, this type of analysis provides an alternative method of visualization that groups seemingly unrelated pathways based on their shared constituent genes, identifying similarities that may otherwise be overlooked. To determine if two pathways are connected, a number of standard distance measures are supported (e.g. Jaccard, Euclidean, and Manhattan), and a threshold is set to define a pair of pathways as being interconnected. By default, the Jaccard Index is used to measure the distance of pathway pairs, and a threshold of 0.8 is used to define a pathway pair as connected. Communities or subgraphs of nodes are not strictly defined, but rather arise organically based

on the methods and options set by the user. Two options are provided for visualizing these pathway-based networks within R. The first produces a static plot of the network using the ggraph [17] library; the second uses visNetwork [18] and is interactive, allowing nodes to be selected, dragged, and grouped together, or categories of nodes (e.g. “Immune System”) can be highlighted—all within the R interface. This dynamic plot can be saved as a standard image file for inclusion in a summary presentation or publication.

Results

The example data set provided by pathlinkR is derived from An et al. [19], which details the study of COVID-19 positive and negative sepsis patients at admission compared to one week later in the ICU. The analysis begins by creating a volcano plot of the DEGs from the COVID-19 positive group (one week vs. admission, Fig 1A). Here it can be observed that a substantial

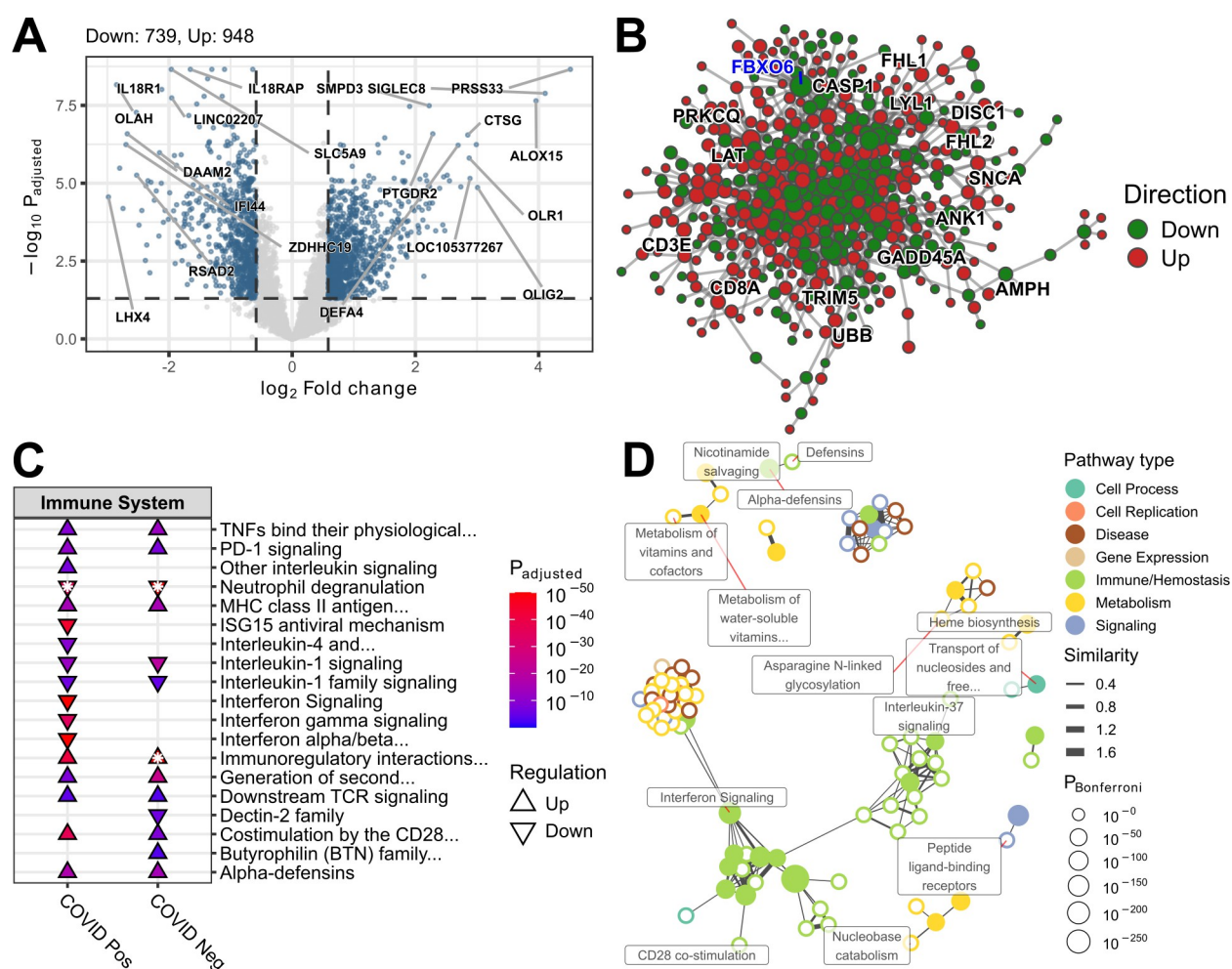


Fig 1. A selection of figures summarizing results generated by pathlinkR. A: Volcano plot created by “eruption” showing the fold changes and significance of DEGs, automatically labelling the top genes. B: A zero order PPI network created with pathlinkR using PPI data from InnateDB, with the node colour indicating fold change direction, and key hub nodes highlighted with blue or black labels. C: Plotting the results of “pathwayEnrichment” for immune pathways, with triangles to indicate the presence and directionality of pathways for each condition on the x axis, and pathways sorted alphabetically on the y-axis. An asterisk indicates a pathway was significant in both directions; only one triangle is depicted, corresponding to the direction with the lower p value. D: A network constructed from pathway enrichment results, with nodes representing single pathways, and edges drawn for pathways with similarity above a specified threshold.

<https://doi.org/10.1371/journal.pcbi.1012422.g001>

number of genes are differentially expressed, primarily upregulated, as indicated in the automatically included title above the plot. By default a selection of top genes ranked by absolute fold change and adjusted p-value are labeled using their HGNC symbols.

To further explore these genes and their potential role in disease, a PPI network can be created using InnateDB interaction data supplied by *pathlinkR*, and providing the DEGs from COVID-19 positive sepsis patients as seed nodes for a zero order network (Fig 1B). The resultant network is densely connected and composed of a mixture of up and down regulated genes, with key hubs highlighted in blue or black, including immune genes CDH1 and HECW2.

To compare the transcriptional changes in COVID-19 positive and negative sepsis patients over time, pathway enrichment can be performed on these two lists of DEGs using “pathwayEnrichment” and the *Sigora* method of gene-pair analysis on the Reactome pathway database. The default when performing pathway enrichment is to treat up and downregulated genes independently for each comparison, and recombine the results at the end. Then the significantly enriched pathways are visualized using “plotPathways” to produce the plot shown in Fig 1C. Attention can be focused by e.g. specifying that only immune-related pathways should be included, making it easier to identify differences in the two results, such as the lack of enriched interferon pathways in the COVID-19 negative sepsis patients compared to their upregulation in the COVID-19 positive sepsis patients.

Lastly, the enriched pathways derived from the COVID-19 positive patients can be used to create a pathway network or “pathnet” using the provided functionality to generate a pathway distance matrix, create the pathway network “foundation” (i.e. interacting pathways), then creating the pathway network itself. Two means of visualizing these pathway networks are provided, one static and one interactive. Fig 1D shows the static pathway network, where each node is a pathway, and edges are drawn if the distance between two pathways (defined by their overlapping genes) is below a defined threshold. In this example, Jaccard distance was used to define pathway interactions, and a distance threshold of 0.8 was set to determine pairs of interacting pathways. Filled nodes represent pathways that were significantly enriched, while empty nodes are those interactors that did not meet statistical significance. Each node is coloured according to its category, the same as those used in Fig 1C to plot only immune pathways. This type of visualization allows the identification of groups of similar pathways that may be overlooked, such as MAP kinase cascade pathways and disease and immune pathways, including interleukin signaling.

Supporting information

S1 SourceCode. Source package of *pathlinkR*, version 1.1.13, representing the “devel” Bioconductor branch 3.20 and including changes requested by reviewers during the revision process.
(TAR.GZ)

Author Contributions

Conceptualization: Travis M. Blimkie, Andy An, Robert E. W. Hancock.

Formal analysis: Andy An.

Funding acquisition: Robert E. W. Hancock.

Investigation: Travis M. Blimkie.

Methodology: Travis M. Blimkie, Andy An.

Project administration: Robert E. W. Hancock.

Software: Travis M. Blimkie, Andy An.

Supervision: Robert E. W. Hancock.

Visualization: Travis M. Blimkie, Andy An.

Writing – original draft: Travis M. Blimkie.

Writing – review & editing: Travis M. Blimkie, Andy An, Robert E. W. Hancock.

References

1. Ozsolak F, Milos PM. RNA sequencing: advances, challenges and opportunities. *Nat Rev Genet.* 2011; 12: 87–98. <https://doi.org/10.1038/nrg2934> PMID: 21191423
2. Zhao K, Rhee SY. Interpreting omics data with pathway enrichment analysis. *Trends Genet.* 2023. <https://doi.org/10.1016/j.tig.2023.01.003> PMID: 36750393
3. Yan J, Risacher SL, Shen L, Saykin AJ. Network approaches to systems biology analysis of complex disease: integrative methods for multi-omics data. *Brief Bioinform.* 2018; 19: 1370–1381. <https://doi.org/10.1093/bib/bbx066> PMID: 28679163
4. Yu G, He Q-Y. ReactomePA: an R/Bioconductor package for Reactome pathway analysis and visualization. *Mol Biosyst.* 2016; 12: 477–479. <https://doi.org/10.1039/c5mb00663e> PMID: 26661513
5. Zhou G, Soufan O, Ewald J, Hancock RE, Basu N, Xia J. NetworkAnalyst 3.0: a visual analytics platform for comprehensive gene expression profiling and meta-analysis. *Nucleic Acids Res.* 2019; 47: W234–W241. <https://doi.org/10.1093/nar/gkz240> PMID: 30931480
6. R Core Team. R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing; 2021. Available: <https://www.R-project.org>
7. Love MI, Huber W, Anders S. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biol.* 2014; 15: 550. <https://doi.org/10.1186/s13059-014-0550-8> PMID: 25516281
8. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics.* 2010; 26: 139–140. <https://doi.org/10.1093/bioinformatics/btp616> PMID: 19910308
9. Milacic M, Beavers D, Conley P, Gong C, Gillespie M, Griss J, et al. The Reactome Pathway Knowledgebase 2024. *Nucleic Acids Res.* 2024; 52: D672–D678. <https://doi.org/10.1093/nar/gkad1025> PMID: 37941124
10. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, et al. InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. *Nucleic Acids Res.* 2013; 41: D1228–D1233. <https://doi.org/10.1093/nar/gks1147> PMID: 23180781
11. Csárdi G, Nepusz T, Traag V, Horvát S, Zanini F, Noom D, et al. igraph: Network Analysis and Visualization in R. 2024. <https://doi.org/10.5281/zenodo.7682609>
12. Pedersen TL. tidygraph: A Tidy API for Graph Manipulation. 2024. Available: <https://tidygraph.data-imaginist.com>
13. Foroushani AB, Brinkman FS, Lynn DJ. Pathway-GPS and SIGORA: identifying relevant pathways based on the over-representation of their gene-pair signatures. *PeerJ.* 2013; 1: e229. <https://doi.org/10.7717/peerj.229> PMID: 24432194
14. Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* 2000; 28: 27–30. <https://doi.org/10.1093/nar/28.1.27> PMID: 10592173
15. Liberzon A, Birger C, Thorvaldsdóttir H, Ghandi M, Mesirov JP, Tamayo P. The molecular signatures database hallmark gene set collection. *Cell Syst.* 2015; 1: 417–425.
16. Korotkevich G, Sukhov V, Budin N, Shpak B, Artyomov MN, Sergushichev A. Fast gene set enrichment analysis. 2016. <https://doi.org/10.1101/060012>
17. Pedersen TL. ggraph: An Implementation of Grammar of Graphics for Graphs and Networks. 2024. Available: <https://ggraph.data-imaginist.com>
18. Almende B.V. and Contributors, Thieurm B. visNetwork: Network Visualization using “vis.js” Library. 2022. Available: <https://datastorm-open.github.io/visNetwork/>
19. An AY, Baghela A, Zhang P, Falsafi R, Lee AH, Trahtemberg U, et al. Severe COVID-19 and non-COVID-19 severe sepsis converge transcriptionally after a week in the intensive care unit, indicating common disease mechanisms. *Front Immunol.* 2023; 14: 1167917. <https://doi.org/10.3389/fimmu.2023.1167917> PMID: 37090709