Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Manuscript Category

Systems Biology

leapR: An R Package for Multi-Omic Pathway Analysis

Vincent Danna[1], Hugh Mitchell[1], Lindsey Anderson[1], Iobani Godinez[1], Sara Gosline[1], Song Feng[1], Justin Teeguarden[1], and Jason E. McDermott[1,2*] [final author order TBD]

[1]Computational Biology Group, Pacific Northwest National Laboratory, Richland Washington 99352 USA., [2]Department of Molecular Microbiology and Immunology, Oregon Health & Sciences University, Portland, Oregon 97201 USA.

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Summary: A generalized goal of many high-throughput data studies is to identify functional mechanisms that underlie observed biological phenomena, whether disease outcomes or metabolic output. Increasingly, studies that rely on multiple sources of high-throughput data (genomic, transcriptomic, proteomic, metabolomic) are faced with a challenge of utilizing the data in a way that maximizes utility. However, methods for integration of multiple forms of molecular data into a biologically coherent frameworks are needed. We have developed a framework to assess biological pathway activity that relates to phenotypic outcome using multi-source data.

Availability and implementation: The leapR package with user manual and example workflow is available for download from GitHub (https://github.com/biodataganache/leapR).

Contact: Jason.McDermott@pnnl.gov

1 Introduction

A generalized goal of many high-throughput data studies is to identify functional mechanisms that underlie observed biological phenomena, whether disease outcomes or metabolic output. Increasingly, studies that rely on multiple sources of high-throughput data (genomic, transcriptomic, proteomic, metabolomic) are faced with a challenge of utilizing the data in a way that maximizes utility. However, methods for integration of multiple forms of molecular data into a biologically and mechanistically coherent framework are needed. We have developed a framework to assess biological pathway activity that relates to phenotypic outcome using multi-source data.

Functional enrichment is a common method for analysis of single-source high-throughput 'omics data and to connect the molecular response observed to higher-level biological functions and mechanisms. Because it relies on patterns of many pathway components rather than individual measurements alone, pathway analysis can provide results that are at once more resistant to noise and able to detect more subtle signals. There have been many different approaches taken to pathway enrichment analysis for different purposes and requiring different assumptions about the data (*1*, *2*).

Recently several methods have been developed to perform specific kinds of pathway enrichment on multi-omics data. MGSEA adapts the popular GSEA algorithm (*1*) to handle multi-omic data inputs (*3*). PathwayPCA uses a principal component analysis approach to data integration to map multi-omics data to pathways (*4*). Finally, ActivePathways uses a multi-stage approach to identify pathway enrichment from p-values derived from multiple omics types (*5*). Though these methods all useful for various purposes, they each focus on a specific approach to enrichment and none support data arising from phosphoproteomics data.

To provide a basis for simple pathway analysis of multiple different types of data, including post-translational modification, we have developed a framework, the Layered Enrichment Analysis of Pathways in R (leapR), to represent multiple omics types, perform pathway analysis on

the individual sets or combined sets, and analyze and represent the results in a biologically meaningful manner.

2 Implementation

The leapR package has functions for reading omics data in the form of data matrices, with rows indicating the molecular species, and columns indicating the sample (treatment, condition, subject, etc.). For data types such as phosphoproteomics the matrix contains additional information about the site of the modification. Multiple methods for functional enrichment are represented through a common interface and with a comprehensive vignette demonstrating their use on a previously published multi-omics dataset (*6*); t test, Fisher's exact test, and Kolgomorov-Smirnov tests can be applied in various ways depending on the type of data and problem considered.

Because pathway analysis often requires testing of many hypotheses (one for each pathway) we also include standard multiple hypothesis correction methods for post-processing as well as randomization methods for calculating significance of results empirically. Support for phosphoproteomics incorporation into pathway enrichment is a key distinction of the package, and example uses are provided in the vignette.

Finally, we introduce a number of customized algorithms to examine pathway enrichment based on correlation of pathway members across conditions, enrichment in interactions in pathways (e.g. from protein-protein interactions), and pathway-specific principle component analysis.

The collection of functions and algorithms provides a basis to analyze and interpret complex, multi-omic datasets and link them with phenotypic outputs in the form of functional pathways.

3 Examples

An example of a multi-omics data set focused on a single problem is from our proteomics study of high-grade serous ovarian cancer (HGSOC) (*6*). We use this dataset, which includes transcriptomics, mass-spectrometry assisted proteomics and phosphoproteomics data for 174 tumors (transcriptomics and proteomics) and 69 tumors (phosphoproteomics), to demonstrate the utility of our methods. We appfunctional enrichment methods to discriminate between patients with short and long survival time. The results show that phosphoproteomics provides the most pathway information as judged by the number of significantly enriched pathways and that all data types show a stronger correlation in the platinum resistant tumors than in the platinum sensitive tumors. Data and an Rmarkdown tutorial which recreates this figure are provided as a package vignette.

4 Conclusions

The rise in the ability to quickly and inexpensively assay the same samples using multiple different molecular profiling technologies has driven the need for improved methods for analyzing and integrating such multi-omic data. Additionally, the biological insight provided by such datasets demonstrates the benefits of developing more sophisticated methods. We believe that the leapR package provides a useful and unique platform for representation and analysis of multi-omic data sets.

Acknowledgements

Funding

Analysis was conducted under the Laboratory Directed Research and Development Program at Pacific Northwest National Laboratory, a multiprogram national laboratory operated by Battelle for the U.S. Department of Energy under Contract DE-AC05-76RL01830.

Conflict of Interest: none declared.

References

1. A. Subramanian *et al.*, Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).
2. W. Huang da, B. T. Sherman, R. A. Lempicki, Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* **37**, 1-13 (2009).
3. K. L. Tiong, C. H. Yeang, MGSEA - a multivariate Gene set enrichment analysis. *BMC Bioinformatics* **20**, 145 (2019).
4. G. J. Odom *et al.*, PathwayPCA: an R/Bioconductor Package for Pathway Based Integrative Analysis of Multi-Omics Data. *Proteomics*, e1900409 (2020).
5. M. Paczkowska *et al.*, Integrative pathway enrichment analysis of multivariate omics data. *Nat Commun* **11**, 735 (2020).
6. H. Zhang *et al.*, Integrated Proteogenomic Characterization of Human High-Grade Serous Ovarian Cancer. *Cell* **166**, 755-765 (2016).