bioRxiv preprint doi: https://doi.org/10.1101/326223; this version posted May 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

GREIN: An interactive web platform for re-analyzing GEO RNA-seq data

Naim Al Mahi[1], Mehdi Fazel Najafabadi[1], Marcin Pilarczyk[1], Michal Kouril[2] and Mario Medvedovic[1*]

[1] Division of Biostatistics and Bioinformatics, Department of Environmental Health, University of Cincinnati, 3223 Eden Avenue, Cincinnati, OH 45220, USA

[2] Division of Biomedical Informatics, Cincinnati Children's Hospital Medical Center, Cincinnati, OH

*To whom correspondence should be addressed.

Abstract

The vast amount of RNA-seq data deposited in Gene Expression Omnibus (GEO) and Sequence Read Archive (SRA) is still a grossly underutilized resource for biomedical research. To remove technical roadblocks for re-using these data, we have developed a web-application GREIN (GEO RNA-seq Experiments Interactive Navigator) which provides simple user-friendly interfaces for manipulating and analyses of GEO RNA-seq data. GREIN is powered by the back-end computational pipeline for uniform processing of RNA-seq data and the large number (>5,500) of already processed datasets. The front-end user interfaces provide a wealth of user-analytics options including sub-setting and downloading processed data, interactive visualization, statistical power analyses, construction of differential gene expression signatures and their comprehensive functional characterization, connectivity analysis with LINCS L1000 data, etc. The combination of the massive amount of back-end data and front-end analytics options driven by user-friendly interfaces makes GREIN a unique open-source resource for re-using GEO RNA-seq data. GREIN is freely accessible at: https://shiny.ilincs.org/grein, the source code is available at: https://github.com/uc-bd2k/grein, and the Docker container is available at: https://hub.docker.com/r/ucbd2k/grein.

Contact: Mario.Medvedovic@uc.edu

Introduction

Re-analysis of GEO RNA-seq datasets can lead to novel scientific insights (Plocik and Graveley, 2013), and data has been routinely used to inform the design of new studies (Hart *et al.*, 2013). However, re-use of RNA-seq data is made difficult by the complexity of the processing protocols which are often inaccessible to biomedical scientists not specializing in bioinformatics. Recent efforts at re-processing GEO/SRA RNA-seq data (Collado-Torres *et al*., 2017; Lachmann *et al.*, 2018) alleviate this problem by providing access to processed and per-transcript summarized RNA-seq data which significantly simplifies its use. Other resources provide access and analysis tools for specific datasets (e.g. Papatheodorou *et al*., 2017). However, open-source user-friendly tools with comprehensive analytical toolbox for re-analysis of all GEO RNA-seq data are still lacking. We addressed this problem by developing and deploying GEO RNA-seq Experiments Interactive Navigator (GREIN) web tool for analysis of GEO RNA-seq data. GREIN provides access to all human, mouse, and rat GEO RNA-seq data. More than 5,500 datasets are already processed and ready for analysis. Other can be processed in the real time. We also release the back-end GEO RNA-seq experiments processing pipeline (GREP2), that can be used to reproduce GREIN results off-line, as well as a Docker container for easy local deployment of the complete infrastructure.

bioRxiv preprint doi: https://doi.org/10.1101/326223; this version posted May 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

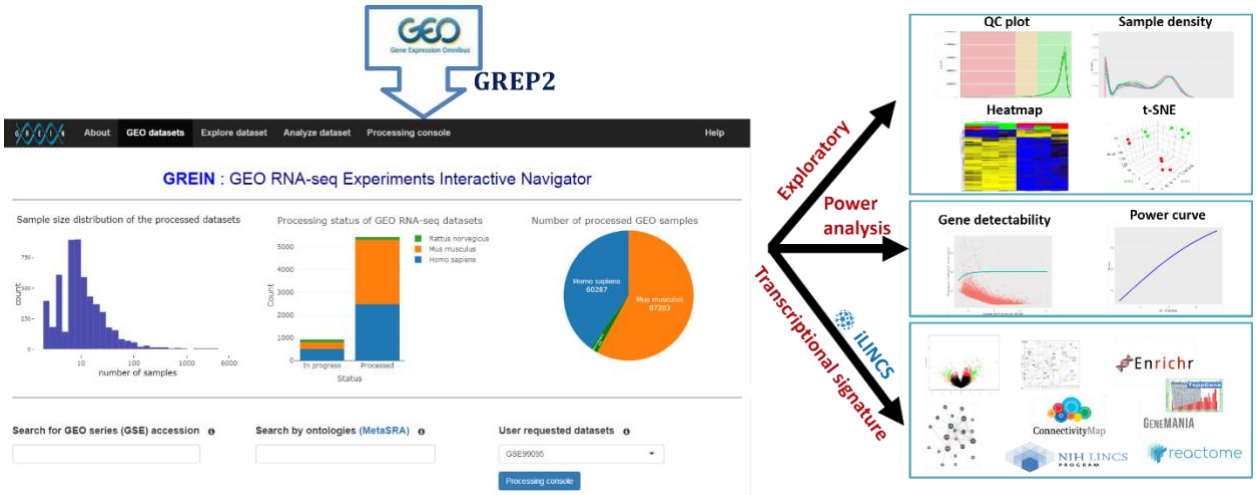

Fig. 1. Brief overview of GREIN infrastructure. Raw data from GEO is processed using GREP2 pipeline and uploaded to GREIN. GUI-driven GREIN workflows facilitate comprehensive analysis and visualization of processed datasets.

Materials and Methods

The conceptual outline of GREIN is showed in Figure 1. Individual RNA-seq datasets are processed by the GREP2 pipeline and stored locally as R Expression Sets. User can access and analyze pre-processed datasets via GREIN graphical user interface (GUI), or submit for processing datasets that have not yet been processed. GUI-driven workflows facilitate examination and visualization of data, statistical analysis, transcriptional signature construction and systems biology interpretation of differentially expressed genes. Both, GREIN and the backend pipeline (GREP2) are written in R (R Core Team, 2016) and released as R packages. Graphical user interfaces for GREIN are implemented in Shiny (Chang *et al*., 2015), a web framework for building dynamic web applications in R. The web instances at https://shiny.ilincs.org/grein is deployed via Shiny server. The complete GREIN infrastructure, including processing pipelines is deployed via Docker container.

GREP2 pipeline proceeds by 1) Retrieving metadata from GEO and the MetaSRA project (Bernstein, *et al.* 2017); 2) Downloading and processing corresponding experiment run files from the SRA; 3) Quantifying transcript abundances; 4) Compiling and organizing QC reports. The detailed description of the pipeline is provided in the Supplemental Methods.

GUI based workflows in GREIN facilitate typical re-use scenario for RNA-seq data such as examination of quality control measures and visualization of expression patterns in the whole dataset, power and sample size analysis for the purpose of informing experimental design of future studies, statistical differential gene expression analysis. The differential gene expression analysis supports standard two-group comparison, but also facilitates fitting of a generalized linear model with two factors where one factor is treated as a covariate, or a batch-effect. The interactive visualization and exploration tools implemented include cluster analysis, interactive heatmaps, principal component, t-SNE, interactive scatter plots, etc.

Biological interpretation of differential gene expressions is aided by direct links to other online tools for performing typical post-hoc analyses such as the gene list and pathway enrichment analysis and the network analysis of differentially expressed genes. The connection to these analytical web services is implemented by submitting the differential gene expression signature (i.e., the list of average changes in gene expression and associated p-values for all genes analyzed) to iLINCS (http://www.ilincs.org/) via API. iLINCS also provides the signatures connectivity analysis for recently released Connectivity Map L1000 signatures (Subramanian *et al*., 2017). For details about the GREIN analytical toolbox please see Supplemental Methods. Step-by-step instructions about GREIN analysis workflows are also provided in the Supplemental Methods.

Conclusions

The combination of the access to a vast number of RNA-seq datasets and the diversity of the analytical tools implemented makes GREIN a unique new resource for re-use of public domain RNA-seq data (for the detailed comparison with other resources please

bioRxiv preprint doi: https://doi.org/10.1101/326223; this version posted May 19, 2018. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

see Supplemental Table 1). GREIN removes technical barriers in re-using GEO RNA-seq data for biomedical scientists. The web instance deployed on our servers provides no-overhead use and requires no technical expertise. Open source R packages and the Docker container provide a flexible and transparent way for computationally savvy users to deploy the complete infrastructure locally with very little effort. The GUI-driven analysis workflows implemented by GREIN cover a large portion of use cases for RNA-seq data analysis, making it the only tool that a scientist may need to meaningfully re-analyze GEO RNA-seq data. In addition, the power analysis workflow provides means to assess the statistical reasons for detecting or not-detecting specific differentially expressed genes (Supplemental Fig 12b). This kind of analyses is not common in the standard RNA-seq pipelines, but the results can be extremely useful when assessing false negative results. Off-line use of GREP2 and GREIN packages, as well as flexible export options enables users to also apply additional tools on GREIN-processed and pre-analyzed datasets.

Funding

This work was supported by the National Institutes of Health [U54HL127624].
Conflict of Interest: none declared.

References

Bernstein,M.N. *et al.* (2017) MetaSRA: normalized human sample-specific metadata for the Sequence Read Archive. *Bioinformatics*, **33**, 2914-2923.

Chang,W. *et al.* (2015) Shiny: web application framework for R. *R package version 0.11*, **1**, 106.

Collardo-Torres,L. *et al.* (2017) Reproducible RNA-seq analysis using recount2. *Nature biotechnology*, **35**, 319.

Hart,S.N. *et al*. (2013) Calculating sample size estimates for RNA sequencing data. *Journal of computational biology*, **20**, 970-978.

Lachmann,A. *et al*. (2018) Massive mining of publicly available RNA-seq data from human and mouse. *Nature communications*, **9**, 1366.

Papatheodorou,I. *et al*. (2017) Expression Atlas: gene and protein expression across multiple studies and organisms. *Nucleic acids research*, **46**, D246-D251.

Plocik,A.M. and Graveley,B.R. (2013) New insights from existing sequence data: generating breakthroughs without a pipette. *Molecular cell*, **49**, 605-617.

R Core Team. (2016) R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

Subramanian,A. *et al*. (2017) A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*, **171**, 1437-1452.