bulkAnalyseR: An accessible, interactive pipeline for analysing and sharing bulk sequencing results

Ilias Moutsopoulos *, Eleanor C Williams * and Irina I Mohorianu @

Wellcome - MRC Cambridge Stem Cell Institute, Jeffrey Cheah Biomedical Centre, University of Cambridge, CB2 0AW, UK
* Joint first authors, @To whom correspondence should be addressed.

Abstract

Motivation: Bulk sequencing experiments are essential for exploring a wide range of biological questions. To bring data analysis closer to its interpretation, and facilitate both interactive, exploratory tasks and the sharing of easily accessible information, we present **bulkAnalyseR**, an R package that offers a seamless, customisable solution for most bulk RNAseq datasets.
Results: In **bulkAnalyseR**, we integrate state-of-the-art approaches, without relying on extensive computational support. We replace static summary images with interactive panels to further strengthen the usability and interpretability of data. The package enables standard analyses on bulk sequencing output, using an expression matrix as the starting point (with the added flexibility of choosing subsets of samples). In an interactive web-based interface, steps such as quality checking, noise detection, inference of differential expression and expression patterns, and biological interpretation (enrichment analyses and identification of regulatory interactions), can be customised, easing the exploration and testing of hypotheses.
Availability: **bulkAnalyseR** is available on GitHub, along with extensive documentation and usage examples (`https://github.com/Core-Bioinformatics/bulkAnalyseR`)
Contact: Irina Mohorianu iim22@cam.ac.uk

1 Introduction

Bulk sequencing experiments play a key role in modern biomedical research, being a primary resource for generating hypotheses [1]. However, the wide diversity of analyses approaches, each with specific trade-offs on speed, robustness, and reproducibility, hinder swift exploration of datasets. Current options include bespoke analyses (through dedicated bioinformatics support), commercially available solutions (often expensive, and with limited flexibility for customisation) or fragmented pipeline-components relying on intermediary conversions of inputs [2, 3, 4].
Most commonly used, publicly available tools have drawbacks that prevent them from being self-sufficient, from taking over the analyses of datasets end-to-end, and being adopted by researchers with restricted bioinformatics resources. VIPER [5] relies on computing skills for setting up analyses; moreover the options for output customisation are limited for user-controlled parameters (e.g. thresholds) or plotting. BioJupies [6] is much simpler to run and boasts a larger selection of outputs. However, it significantly restricts the customisation of individual pipeline components. Critically, neither VIPER, nor BioJupies offer support for more complex differential expression (DE) tasks, beyond simple pair-wise comparisons. This limits the biological interpretations from more complex experimental designs. Searchlight [7] alleviates some of these issues, focusing on an interactive exploration of results; however, it requires as input tables of differentially expressed genes, implying that several steps e.g. QCs and DE inference need to be performed *a priori*. It also relies on computational expertise for setting up the environment and retrieving the outputs. In addition, the resulting plots, while suitable for hypothesis generation and mining of results, require additional tweaking to meet publication requirements.
bulkAnalyseR enables the analysis of bulk sequencing data in an interactive interface, with numerous customisable components (from noise detection [8] to the identification of patterns [9] and GRN prediction [4]). The setup facilitates a seamless transition to publication-ready figures without needing additional bioinformatics support. The resulting web-app can be easily published, incentivising research reproducibility.

2 Features and Usage

bulkAnalyseR provides an accessible, yet flexible framework for the analysis of bulk sequencing data without relying on prior programming expertise (Fig S1). The users can create a shareable shiny app in two lines of code,

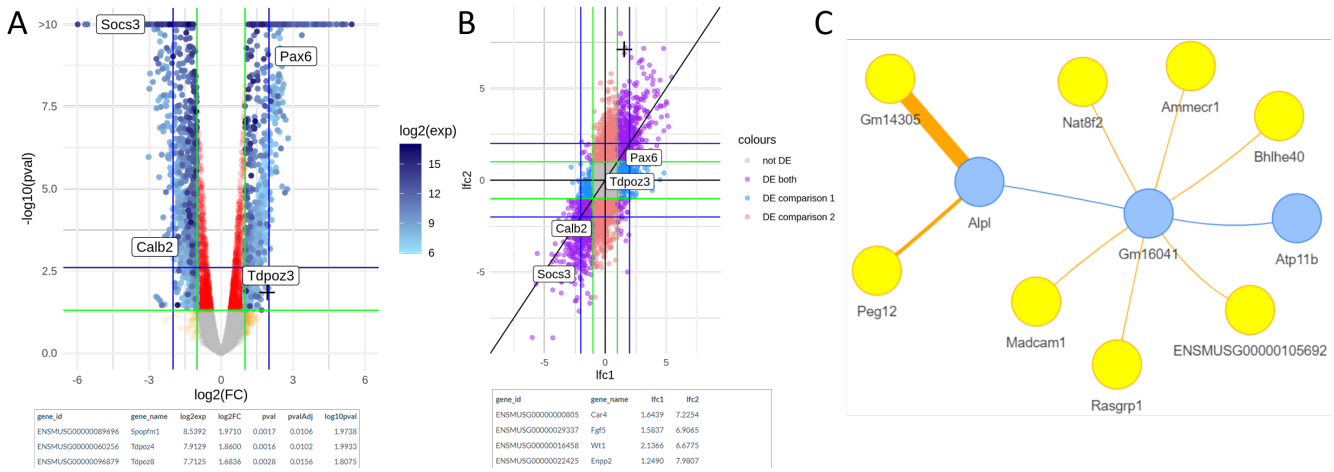

Figure 1: **Examples of bulkAnalyseR summary plots on bulk mRNAseq data [10]** A) Volcano plot showcasing DE genes between the 0h and 12h samples. B) Cross-plot comparing two DE outputs (log_2FC from 0h vs 12h on the x-axis and and 0h vs 36h on the y-axis). C) Visualisation of top 8 connections in an inferred GRN starting from Gm16041, Atp11b and Alpl (DE for the 0h and 12h samples); the edge widths are proportional to the weights from the adjacency matrix.

from an expression matrix and a metadata table. The pre-processing step `preprocessExpressionMatrix` handles the counts-based noise detection [8] and normalisation of expression levels (Fig S2). Next, `generateShinyApp` checks the compatibility of inputs and a shiny app is created. Subsequent analyses can be performed interactively (Fig S3). QC plots highlight outlier samples; differential expression (DE) analyses (and comparison of predictions, Fig 1A,B) can be performed and visualised. The interpretation of results relies on enrichment analyses for summarising the biological significance of DE genes, identification of patterns and localised Gene Regulatory Networks (GRNs), Fig 1C. All plots and tables can be downloaded individually; the app itself can be shared with ease.

Inputs and pre-processing The inputs for **bulkAnalyseR** are an expression matrix and a metadata table. The former contains counts per genes across samples (usually generated from BAM alignments e.g. with featureCounts). The metadata table contains additional experimental information and represents the starting point for complex DE analyses and comparisons. The users can access the full noise detection pipeline, noisyR [8] which outputs a denoised, un-normalised expression matrix (Fig S2). Next, several options are available for normalisation (quantile normalisation, as default); the integration with other preprocessing steps is possible. We recommend a thorough, systematic assessment of these steps, not included in the app due to potentially long processing times. The shiny app is generated using `generateShinyApp` on the denoised, normalised expression matrix. The resulting app is standalone and shareable e.g. via online platforms like shinyapps.io. This setup facilitates the mining of the same instance of a dataset by the community, increasing access and promoting reproducibility of bioinformatics analyses.

Interactive Visualisation for exploratory analyses The shiny app comprises several tabs: [a] the Quality Checks (QCs) include a Jaccard Similarity Index (JSI) heatmap and a PCA dimensionality reduction, with groups based on the metadata summary (Fig S3A,B). This enables a high-level overview of the similarity across samples, reflecting the experimental design. [b] The differential expression (DE) tab includes DE summaries across selected comparisons performed using edgeR [2] and DESeq2 [3] with customisable parameters; the outputs are lists of DE genes (Fig S3C). [c] The DE results can be visualised in the Volcano/MA and DE summary panels which promote an interactive exploration of the data (Fig 1A, Fig S3E,F,H). [d] the interpretation of the DE genes commences with the enrichment tab that overviews a gene set enrichment analysis using g:profiler [11]. The user selects DE genes of interest; the GSEA is performed on GO terms, KEGG and Reactome pathway terms, and regulatory features (miRNAs and transcription factors), Fig S3G. Further information can be obtained by [f] comparing DE lists in cross-plots (Fig 1B, Fig S3D) and [e] identifying simplified patterns that may underline regulatory interactions (Fig S3I). The visualisation of the latter is included (Fig 1C, Fig S3J), with a GRN generated using the GENIE3 approach [4].

All outputs can be downloaded directly from the app (as high resolution png files); the embedded customisation leads to publication-ready figures. **bulkAnalyseR** includes highly customisable plotting functions for JSI heatmaps, PCA, MA and Volcano plots that enhance the visualisation flexibility. Lists of differentially expressed genes and enriched terms can be downloaded through the app in **.csv** format.

3 Discussion

The input based on an expression matrix confers flexibility to the **bulkAnalyseR** package. Other sequencing datasets, which can be summarised in an expression matrix, can be processed on this pipeline (we illustrate a case study of H3K4me3 ChIPseq data, Fig S4). This versatility is a promising first (interactive) step to integrating multiple sources of information, and guiding mechanistic studies (e.g. for multi-omics GRN inference).

Author Contributions and Funding

IM and ECW implemented the **bulkAnalyseR** package; IIM supervised the project. All authors wrote and revised the manuscript. This research was funded by the Wellcome Trust [203151/Z/16/Z] and the UKRI Medical Research Council [MC_PC_17230].

References

[1] R Stark et al. RNA sequencing: the teenage years. *Nature Reviews Genetics*, 20, 07 2019.

[2] Mark D. Robinson et al. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140, 11 2009.

[3] M Love et al. Moderated estimation of fold change and dispersion for RNA-Seq data with DESeq2. *Genome Biology*, 15:550, 12 2014.

[4] Vân Anh Huynh-Thu et al. Inferring regulatory networks from expression data using tree-based methods. *PloS one*, 5, 09 2010.

[5] Macintosh Cornwell et al. VIPER: Visualization Pipeline for RNA-seq, a Snakemake workflow for efficient and complete RNA-seq analysis. *BMC Bioinformatics*, 19, 04 2018.

[6] Denis Torre et al. BioJupies: automated generation of interactive notebooks for RNA-Seq data analysis in the cloud. *Cell Systems*, 7(5):556–561.e3, 2018.

[7] John Cole et al. Searchlight: automated bulk RNA-seq exploration and visualisation using dynamically generated R scripts. *BMC Bioinformatics*, 22, 08 2021.

[8] Ilias Moutsopoulos et al. noisyR: enhancing biological signal in sequencing datasets by characterizing random technical noise. *Nucleic Acids Research*, 49(14):e83–e83, 06 2021.

[9] Sara Lopez-Gomollon et al. Diverse correlation patterns between micrornas and their targets during tomato fruit development indicates different modes of microrna actions. *Planta*, 236, 08 2012.

[10] P Yang et al. Multi-omic profiling reveals dynamics of the phased progression of pluripotency. *Cell Systems*, 8(5):427–445.e10, 2019.

[11] Uku Raudvere et al. g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Research*, 47(W1):W191–W198, 05 2019.