RTrans: a pipeline for multi-way analysis of differential gene expression profiles

Georgy Sergeevich Krasnov
Laboratory of Postgenomic Research
EIMB RAS
Moscow, Russia
gskrasnov@mail.ru

Anastasiya Andreevna Kobelyatskaya
Laboratory of Postgenomic Research
EIMB RAS
Moscow, Russia
kaa.chel@mail.ru

Anastasiya Vladimirovna Snezhkina
Laboratory of Postgenomic Research
EIMB RAS
Moscow, Russia
leftger@rambler.ru

Elena Anatolevna Pudova
Laboratory of Postgenomic Research
EIMB RAS
Moscow, Russia
pudova_elena@inbox.ru

Anna Victorovna Kudryavtseva
Laboratory of Postgenomic Research
EIMB RAS
Moscow, Russia
rhizamoeba@mail.ru

Introduction

RNA-Seq is a widespread technique routinely used to reveal differentially expressed (DE) genes, splicing events and affected signaling pathways. Usually, RNA-Seq data analysis needs skills in programming and bioinformatics. Last years, several pipelines facilitating transcriptome analysis have been developed (DEBrowser, UTAP, iDEP, Chipster, RNACocktail, DEapp) but there is still a need for tools that would be convenient in daily practice [1-6].

Hence, we aimed to develop RNA-Seq data processing pipeline that would satisfy several requirements:

o   include various statistical test, both parametric and non-parametric, and multifactor ANOVA;

o   be convenient to use in everyday practice, require minimal user attention, not require filling up numerous forms at each step of the analysis;

o   quickly make several intergroup comparisons or ANOVA analyses in one run;

o   include pathway enrichment analyses across several databases, and WGCNA;

o   create interactive and flexible reports, including heatmaps, MDS plots, enrichment maps, etc.

METHODS

- We present RTrans, a pipeline for analyzing gene expression data (read counts) that aims to automate the analysis, minimize the needs for user attention, and parallelize the analyses.

- Before RTrans launch, users only need to fill out Excel workbook describing experimental design and provide read counts data.

- RTrans is written in R and Python. It is based on several Bioconductor packages: edgeR and DESeq (differential expression); clusterProfiler, topGO, ReactomePA (enrichment analyses); pathview, ggplot2, pheatmap, etc. (visualization).

- It includes several tests for differential expression analysis: quasi-likelihood F-test, likelihood ratio test, exact Fisher's test for GLM (edgeR); non-parametric Mann-Whitney, Wilcoxon, Kruskal-Wallis test; Student's test; Spearman and Pearson correlations; comparison of distribution densities. RTrans also considers samples pairing (e.g tumor-normal).

- Additionally, RTrans analyzes dependencies between gene expression and transcript lengths; it stratifies genes into several bins depending on transcript lengths and then normalize read count within each bin. This procedure partially eliminates the effect of 3'-bias coming from RIN imbalance.

- RTrans performs enrichment tests on GO, KEGG, Reactome, WikiPathways, Disease Ontology, DisGeNET, Network of Cancer Genes databases.

- It creates GO/KEGG-centric DE expression plots, heatmaps, MDS plots, Venn diagrams, enrichment maps, and other reports. Most of the reports generated by RTrans are formatted Excel workbooks.

Results

The developed pipeline aims to automate the analysis, minimize the needs for user attention, and parallelize the analyses. Before RTrans launch, users only need to fill out Excel workbook describing experimental design and provide read counts data.

RTrans produces Excel-formatted differential gene expression reports with embedded sparklines/heatmaps illustrating gene expression distribution, pathway enrichment reports (Gene Ontology, KEGG, Reactome, WikiPathways, Disease Ontology, DisGeNET, Network of Cancer Genes databases), heatmaps, MDS plots, Venn diagrams, enrichment maps, GO/KEGG-centric DE expression plots, and other reports.

RTrans represents a versatile pipeline facilitating DE analysis based on RNA-Seq and other quantitative data. The pipeline is available at https://github.com/gskrasnov/RTrans

References

[1] Ge SX, Son EW, Yao R. iDEP: an integrated web application for differential expression and pathway analysis of RNA-Seq data. BMC Bioinformatics. 2018 Dec 19;19(1):534. doi: 10.1186/s12859-018-2486-6.

[2] Kallio MA, Tuimala JT, Hupponen T, Klemelä P, Gentile M, Scheinin I, Koski M, Käki J, Korpelainen EI. Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genomics. 2011 Oct 14;12:507. doi: 10.1186/1471-2164-12-507.

[3] Kohen R, Barlev J, Hornung G, Stelzer G, Feldmesser E, Kogan K, Safran M, Leshkowitz D. UTAP: User-friendly Transcriptome Analysis Pipeline. BMC Bioinformatics. 2019 Mar 25;20(1):154. doi: 10.1186/s12859-019-2728-2.

[4] Kucukural A, Yukselen O, Ozata DM, Moore MJ, Garber M. DEBrowser: interactive differential expression analysis and visualization tool for count data. BMC Genomics. 2019 Jan 5;20(1):6. doi: 10.1186/s12864-018-5362-x.

[5] Li Y, Andrade J. DEApp: an interactive web interface for differential expression analysis of next generation sequence data. Source Code Biol Med. 2017 Feb 3;12:2. doi: 10.1186/s13029-017-0063-4. eCollection 2017.

[6] Sahraeian SME, Mohiyuddin M, Sebra R, Tilgner H, Afshar PT, Au KF, Bani Asadi N, Gerstein MB, Wong WH, Snyder MP, Schadt E, Lam HYK. Gaining comprehensive biological insight into the transcriptome by performing a broad-spectrum RNA-seq analysis. Nat Commun. 2017 Jul 5;8(1):59. doi: 10.1038/s41467-017-00050-4.