A Bioconductor R pipeline for analysis of RNA-seq data

Wei Shi (shi@wehi.edu.au)
Gordon Smyth's Lab, Walter and Eliza Hall Institute, Melbourne

Method Article

Keywords: next-generation sequencing, RNA-seq, read alignment, read summarization, normalization, differential expression

Posted Date: May 18th, 2015

DOI: https://doi.org/10.1038/protex.2015.039

Abstract

We describe a powerful and easy-to-use RNA-seq analysis pipeline that can be used for complete analysis of RNA-seq data. It starts with raw read output of an sequencing instrument and reports lists of genes that are found to be differentially expressed in the comparison of different cell types. It consists of several analysis modules including Subread read alignment \[1], featureCounts read summarization \[2], voom normalization \[3] and statistical testing of differential expression using empirical Bayes moderated t-statistic \[4]. The entire pipeline mainly makes use of two R packages, Rsubread and limma, both available from the popular Bioconductor project.

Procedure

1. Check the sequencing quality of RNA-seq data using 'qualityScores' function in Rsubread package. 2. Use 'align' function in Rsubread to align the reads. 3. Use 'featureCounts' function in Rsubread to assign reads to genes. 4. Use 'voom' function in limma package to normalize read counts and to estimate the mean-variance relationship. 5. Use 'lmFit' function in limma to fit linear models to genes. 6. Use 'treat' \(or 'eBayes') function in limma to compute moderated t statistic for each gene for each comparison. 7. Call differentially expressed genes using the 'decideTests' function in limma. The Rsubread package can be downloaded from http://bioconductor.org/packages/release/bioc/html/Rsubread.html. The limma package can be downloaded from http://bioconductor.org/packages/release/bioc/html/limma.html

Timing

The whole pipeline will take less than 8 hours to complete the analysis of an RNA-seq dataset including 100 million reads in total, on a computer with >=4 CPUs and >=8GB of memory.

Anticipated Results

Lists of genes whose expression levels are found to be statistically significantly changed in different conditions \(eg. different cell types or different treatments).

References

\[1] Liao Y, Smyth GK and Shi W. The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. Nucleic Acids Research, 41\(10):e108, 2013 \[2] Liao Y, Smyth GK and Shi W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics, 30\(7):923-30, 2014 \[3] Law CW, Chen Y, Shi W and Smyth GK. Voom: precision weights unlock linear model analysis tools for RNA-seq read counts. Genome Biology, 15\(2):R29, 2014 \[4] Smyth GK. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. Statistical Applications in Genetics and Molecular Biology, Volume 3, Article 3, 2004