BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Web Server

Volume 7(1)

DEB: A web interface for RNA-seq digital gene expression analysis

Ji Qiang Yao* & Fahong Yu

Interdisciplinary Center for Biotechnology Research, University of Florida, Gainesville, FL 32610; Ji Qiang Yao - Email: jiqiangyao@ufl.edu; *Corresponding author

Received July 30, 2011; Accepted July 31, 2011; Published August 20, 2011

Abstract:
Digital expression (DE) is an important application of RNA-seq technology to quantify the transcriptome. The number of mapped reads to each transcript or gene varies under different conditions and replicates. Currently, three different statistical algorithms (edgeR, DESeq and bayseq) are available as R packages, to compare the reads to identify significantly expressed transcripts or genes. So far, users have to manually install and run each R package separately. It is also of users' interest to compare the results of different approaches. Here, we present a pipeline DEB which automates all the steps in file preparation, computation and result comparison.

Keywords: DEB, edgeR, DESeq, baySeq, RNA-seq, digital expression, nextGen sequencing

Availability: DEB is freely accessed at http://www.ijbcb.org/DEB/php/onlinetool.php

Background:
RNA-seq is a revolutionary tool for transcriptomics **[1]**. This high-throughput sequencing technology quickly becomes valuable for many functional genomics applications such as digital gene expression (DGE) study. Typically, RNA-Seq reads are classified based on their mapping to a common region of the target genome such as exon or transcript. One of the fundamental data analysis tasks for RNA-seq studies is to determine whether there is evidence that read counts for a transcript or gene are significantly different across experimental conditions. At present, there are three major algorithms to address this problem. EdgeR is designed for the analysis of replicated count-based expression data and is an implementation of methodology developed by Robinson and Smyth **[2]**. DESeq is similar to edgeR. Both assume negative binomial distribution model. The difference between the two methods lies on their estimation of the squared coefficient of variation (SCV) **[3]**. Recently, Hardcastle and Kelly developed bayseq, which assumes a negative binomial distribution for the data and derives an empirically determined prior distribution from the entire dataset **[4]**. DEB is a web interface (**Figure 1**) that integrates the three algorithms into one place. The user can select any or all of the algorithms for data analysis. In case more than one algorithm is selected, the shared genes among the algorithms are generated.

Implementation:
DEB was developed using HTML, PHP, Perl scripting language, R programing language and MYSQL as the database backend. The pipeline book-keeps users' input information, initiates job-running process, updates job status, delivers the final results and deletes the results after 48 hours. DEB has been tested on the following browsers, Firefox 4.0, IE 8.0, Chrome 12.0, Safari 5.0 and Opera 11.5.

Software Input:
The input to DEB is a tab-delimited count data file. The format of the input file is as follows: 1) The leftmost column must be the gene list, the rest columns are count data; 2) The first row (header) contains sample names. The samples are categorized to two groups, e.g. patient and control; 3) Sample names within one group are differentiated only by the last suffix, e.g. T1, T2, T3. A demo input file is provided for illustration and test purpose. The user can select one of the following False

Discovery Rate (FDR) level, i.e. 1%, 5%, 10%, 15% and 20%. The default value is 10%. The user can also select any one, two or all of the algorithms provided, i.e. edgeR, DESeq and baySeq. The default is all the three algorithms. It is required that an email address is provided so that the final result is conveniently delivered to the user in case the computation takes long time to complete.

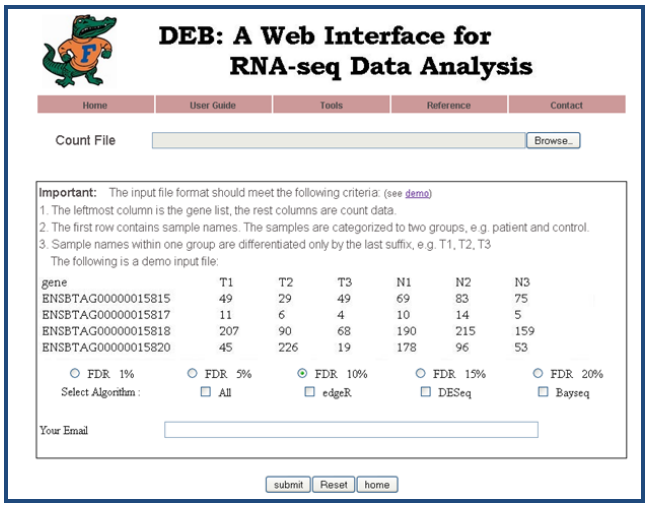

Figure 1: The DEB web-interface.

Software output:

For each algorithm selected, the final result generates a list of genes that are significantly expressed under the user-selected FDR level. If more than two algorithms are selected, the shared gene lists among all the results of the selected algorithms, are also provided. For users' convenience, the gene list files are made in different formats, such as text, excel and html. In addition to the data files, smear plots (**Figure 2**) of significantly expressed genes are also generated for the user to download. We tested the software with the demo file that contains 25,668 genes and a selection of all three algorithms. The total time response is about three minutes.

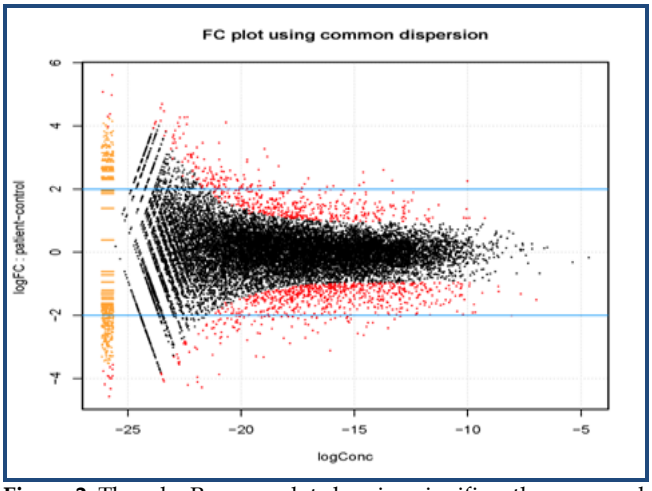

Figure 2: The edgeR smear plot showing significantly expressed genes (red)

Conclusion:

DEB is a convenient web tool to identify significantly expressed genes for RNA-seq data analysis using edgeR, DESeq and baySeq algorithms.

Caveat and future development:

Currently, the program can only accept count data which is generated by users using other bioinformatics tools. It is our plan to develop a pipeline so that the user can submit the raw sequencing data files to the server. The server can automate cleaning, mapping and counting processes to generate the count file.

References:

[1] Wang Z *et al*. *Nat Rev Genet*. 2009 **10**: 57 [PMID: 19015660]
[2] Robinson MD *et al*. *Bioinformatics* 2010 **26**: 139 [PMID: 19910308]
[3] Anders S & Huber W. *Genome Biol*. 2010 **11**: R106 [PMID: 20979621]
[4] Hardcastle TJ & Kelly KA. *BMC Bioinformatics*. 2010 **11**: 422 [PMID: 20698981]

Edited by P Kangueane
Citation: Yao & Yu. Bioinformation 7(1): 44-45 (2011)