RNA-Seq: Differential Expression Analysis

BCB 5250 Introduction to Bioinformatics II

Spring 2020

Tae-Hyuk (Ted) Ahn
Department of Computer Science
Program of Bioinformatics and Computational Biology
Saint Louis University

SAINT LOUIS
UNIVERSITY
EST. 1818

Analyzing differential gene expression

- Biological replicates are important!
- Normalization is required in order to compare expression between samples
  - Different library sizes
  - RNA composition bias caused by sampling approach
- Raw counts are needed to assess measurement precision
  - Counts are the ”the units of evidence” for expression
  - Uncertainty information is lost if counts are transformed to FPKM
  - Normalizes for gene length and library size.
  - Example:
    - 20 kb transcript has 400 counts, library size is 20 million reads: FPKM = (400/20) / 20
    - 0.5 kb transcript has 10 counts, library size is 20 million reads: FPKM = (10/0.5) / 20
    - In both cases FPKM =1, but it is less likely to get 400 reads just by chance

Software packages for DE analysis

- edgeR
- DESeq2
- Sleuth
- DRIMSeq
- DEXSeq
- Cuffdiff, Ballgown
- Limma + voom, limma + vst
- ...

Which tool should I use?

Method | Open Access | Published: 10 September 2013

Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data

Franck Rapaport, Raya Khanin, Yupu Liang, Mono Pirun, Azra Krek, Paul Zumbo, Christopher E Mason, Nicholas D Socci & Doron Betel

Genome Biology **14**, Article number: 3158 (2013) | Cite this article

160k Accesses | **350** Citations | **74** Altmetric | Metrics

The Erratum to this article has been published in Genome Biology 2015 16:261

PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

A Comparative Study of Techniques for Differential Expression Analysis on RNA-Seq Data

Zong Hong Zhang, Dhanisha J. Jhaveri, Vikki M. Marshall, Denis C. Bauer, Janette Edson, Ramesh K. Narayanan, Gregory J. Robinson, Andreas E. Lundberg, Perry F. Bartlett, Naomi R. Wray, Qiong-Yi Zhao

Published: August 13, 2014 • https://doi.org/10.1371/journal.pone.0103207 • >> See the preprint

Article | Authors | Metrics | Comments | Media Coverage

Abstract

Abstract

PUBLISH ABOUT BROWSE

PLOS ONE

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

RNA-Seq differential expression analysis: An extended review and a software tool

Juliana Costa-Silva, Douglas Domingues, Fabricio Martins Lopes

Published: December 21, 2017 • https://doi.org/10.1371/journal.pone.0190152

Article | Authors | Metrics | Comments | Media Coverage

Abstract

Abstract

DESeq2

Installation

To install this package, start R (version "3.6") and enter:

```
if (!requireNamespace("BiocManager", quietly = TRUE))
    install.packages("BiocManager")

BiocManager::install("DESeq2")
```

Analyzing RNA-seq data with DESeq2

Michael I. Love, Simon Anders, and Wolfgang Huber

10/29/2019

Abstract

A basic task in the analysis of count data from RNA-seq is the detection of differentially expressed genes. The count data are presented as a table which reports, for each sample, the number of sequence fragments that have been assigned to each gene. Analogous data also arise for other assay types, including comparative ChIP-Seq, HiC, shRNA screening, mass spectrometry. An important analysis question is the quantification and statistical inference of systematic changes between conditions, as compared to within-condition variability. The package DESeq2 provides methods to test for differential expression by use of negative binomial generalized linear models; the estimates of dispersion and logarithmic fold changes incorporate data-driven prior distributions This vignette explains the use of the package and demonstrates typical workflows. An RNA-seq workflow on the Bioconductor website covers similar material to this vignette but at a slower pace, including the generation of count matrices from FASTQ files. DESeq2 package version: 1.26.0

DEseq2 Input

The DESeqDataSet

The object class used by the DESeq2 package to store the read counts and the intermediate estimated quantities during statistical analysis is the *DESeqDataSet*, which will usually be represented in the code here as an object `dds`.

A technical detail is that the *DESeqDataSet* class extends the *RangedSummarizedExperiment* class of the SummarizedExperiment package. The "Ranged" part refers to the fact that the rows of the assay data (here, the counts) can be associated with genomic ranges (the exons of genes). This association facilitates downstream exploration of results, making use of other Bioconductor packages' range-based functionality (e.g. find the closest ChIP-seq peaks to the differentially expressed genes).

A *DESeqDataSet* object must have an associated *design formula*. The design formula expresses the variables which will be used in modeling. The formula should be a tilde (~) followed by the variables with plus signs between them (it will be coerced into an *formula* if it is not already). The design can be changed later, however then all differential analysis steps should be repeated, as the design formula is used to estimate the dispersions and to estimate the log2 fold changes of the model.

Note: In order to benefit from the default settings of the package, you should put the variable of interest at the end of the formula and make sure the control level is the first level.

We will now show 4 ways of constructing a *DESeqDataSet*, depending on what pipeline was used upstream of DESeq2 to generated counts or estimated counts:

1. From transcript abundance files and tximport
2. From a count matrix
3. From htseq-count files
4. From a SummarizedExperiment object

A simple example data

- Copy example file to your work directory:
  hopper:/public/ahnt/courses/bcb5250/rna_seq_lab/count_matrix.txt

```
[ahnt@hopper:/public/ahnt/courses/bcb5250/rna_seq_lab]$ head -10 count_matrix.txt
GENE    ctrl1   ctrl2   ctrl3   treat1  treat2  treat3
0610005C13Rik   1438    1104    1825    1348    1154    1005
0610007N19Rik   1012    1152    1139    878     885     835
0610007P14Rik   704     796     881     826     865     929
0610009B22Rik   757     802     780     885     853     987
0610009D07Rik   1107    1183    1220    1258    1221    1428
0610009L18Rik   129     154     139     138     166     179
0610009O20Rik   1585    1791    1729    2004    1951    2215
0610010B08Rik   0       0       0       0       0       0
0610010F05Rik   401     455     484     499     509     622
```

R script is ready to run

- hopper:/public/ahnt/courses/bcb5250/rna_seq_lab/deseq2.r

RNA-seq count matrix

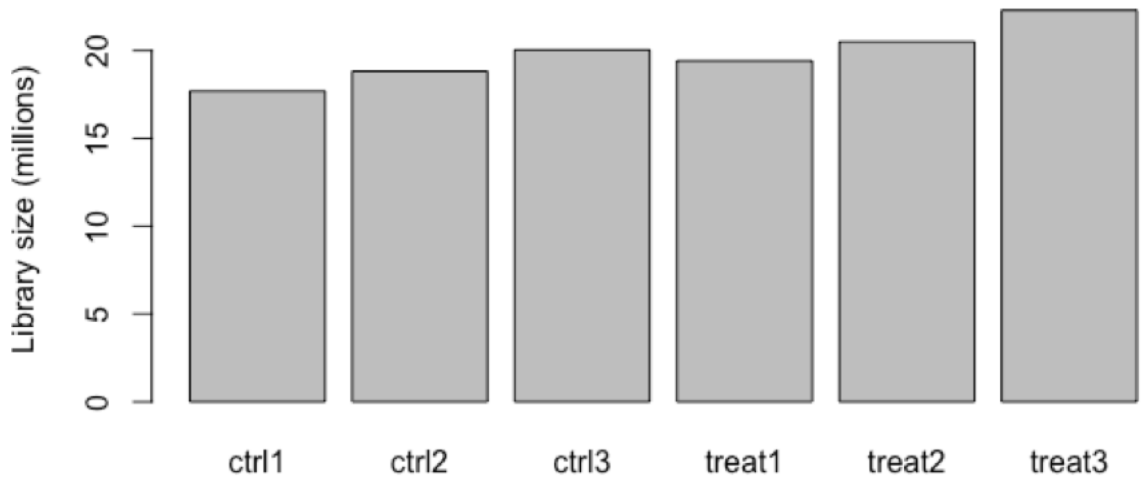
```r
# working directory
getwd()

# read in count matrix
countData <- read.csv("count_matrix.txt", header=T, row.names=1, sep="\t")
dim(countData)
head(countData)
```

```
> getwd()
[1] "/Users/ahnt/Google Drive/Teaching/SLU/2019-20/BCB5250_IntroBioinformaticsII/Labs/rna_seq"
> # read in count matrix
> countData <- read.csv("count_matrix.txt", header=T, row.names=1, sep="\t")
> dim(countData)
[1] 24009     6
> head(countData)
              ctrl1 ctrl2 ctrl3 treat1 treat2 treat3
0610005C13Rik  1438  1104  1825   1348   1154   1005
0610007N19Rik  1012  1152  1139    878    885    835
0610007P14Rik   704   796   881    826    865    929
0610009B22Rik   757   802   780    885    853    987
0610009D07Rik  1107  1183  1220   1258   1221   1428
0610009L18Rik   129   154   139    138    166    179
>
```

Basic QC

```
 9  # basic QC
10  barplot(colSums(countData)*1e-6,
11          names=colnames(countData),
12          ylab="Library size (millions)")
13
```



Run DESeq2

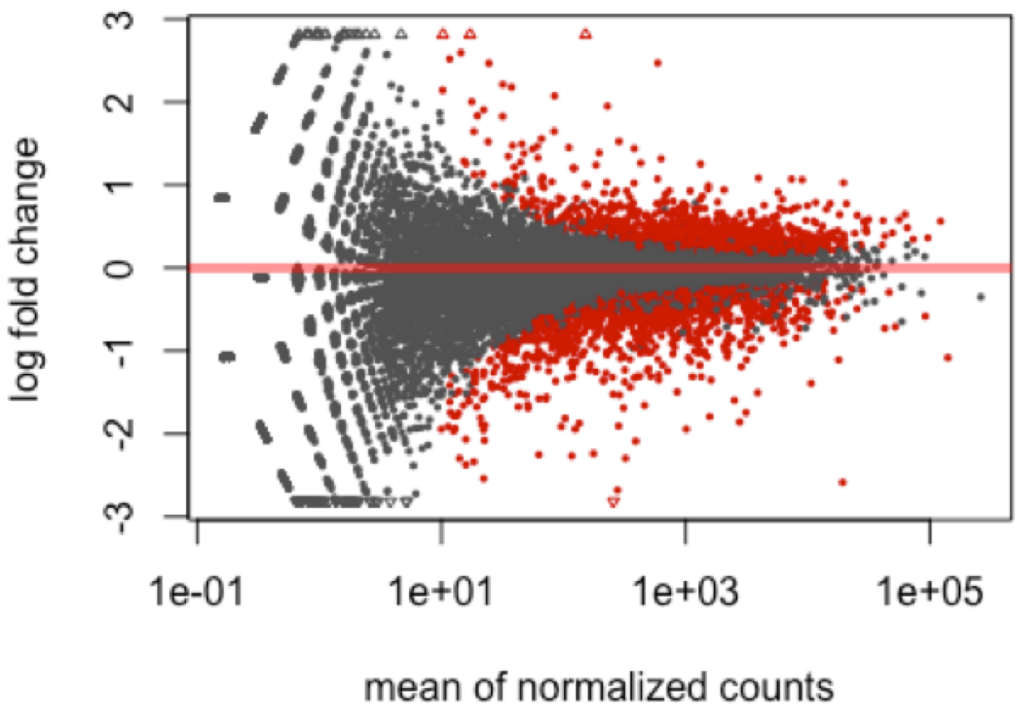
```
# load library
library(DESeq2)

# create experiment labels (two conditions)
colData <- DataFrame(condition=factor(c("ctrl","ctrl","ctrl", "treat", "treat", "treat")))

# create DESeq input matrix
dds <- DESeqDataSetFromMatrix(countData, colData, formula(~ condition))

# run DEseq
dds <- DESeq(dds)

# visualize differentially expressed genes
plotMA(dds)
```



Get the DE genes

```
29  # get differentially expressed genes
30  res <- results(dds)
31
32  # order by BH adjusted p-value
33  resOrdered <- res[order(res$padj),]
34
35  # top of ordered matrix
36  head(resOrdered)
37
38  # how many differentially expressed genes ? FDR=10%, |fold-change|>2 (up and down)
39  # get differentially expressed gene matrix
40  sig <- resOrdered[!is.na(resOrdered$padj) &
41                    resOrdered$padj<0.10 &
42                    abs(resOrdered$log2FoldChange)>=1,]
43
44  # top of the differentially expressed genes
45  head(sig)
```

```
> head(sig)
log2 fold change (MLE): condition treat vs ctrl
Wald test p-value: condition treat vs ctrl
DataFrame with 6 rows and 6 columns
                baseMean     log2FoldChange             lfcSE                stat
               <numeric>          <numeric>         <numeric>           <numeric>
Pck1    19300.0081461269 -2.58567884003822  0.15587324529687 -16.5883428879257
S100a14 590.630493920057  2.47135262182836 0.179673051397215  13.7547206028397
Ugt1a2   2759.7012247113 -1.85763333812422 0.147125409154535 -12.6261897846145
Mlph    1320.64049620762  1.25376185953446 0.119286425402794  10.5105158051378
Ly6e    19570.5876108613  1.02588269075554 0.102962319699055  9.96367111535617
Pklr    787.359243644928 -1.03560914563203 0.109252461308009 -9.47904635953601
                      pvalue                 padj
                   <numeric>            <numeric>
Pck1    8.46231417732845e-62 1.19428639984636e-57
S100a14 4.77120358045434e-43 3.36679980654761e-39
Uat1a2   1.5143514865502e-36 7.12401417656101e-33
```

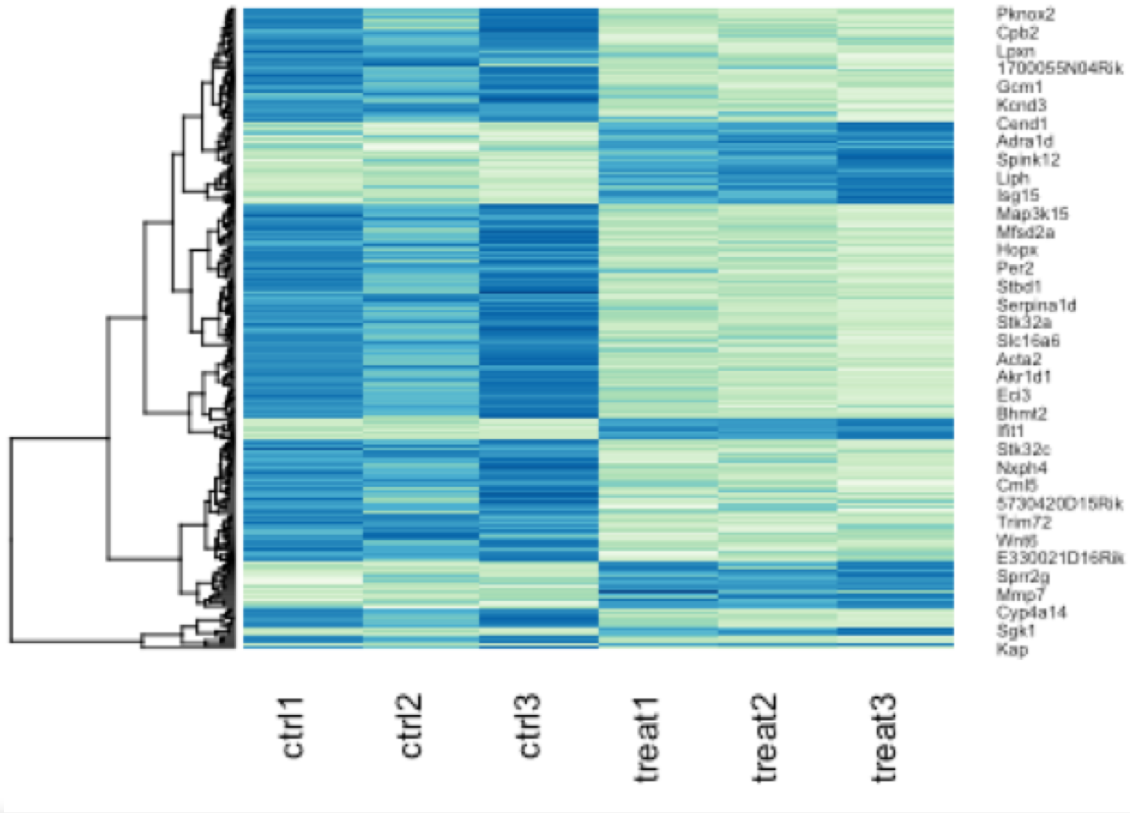
Get Heatmap

```r
# load libraries for the heat map
library("RColorBrewer")
if (!requireNamespace("BiocManager", quietly = TRUE))
  install.packages("BiocManager")
BiocManager::install("gplots")
library("gplots")

# colors of the heat map
hmcol <- colorRampPalette(brewer.pal(9, "GnBu"))(100) ## hmcol <- heat.colors

# heatmap
heatmap.2( log2(counts(dds,normalized=TRUE)[rownames(dds) %in% selected,]),
           col = hmcol, scale="row",
           Rowv = TRUE, Colv = FALSE,
           dendrogram="row",
           trace="none",
           margin=c(4,6), cexRow=0.5, cexCol=1, keysize=1 )

```



What conditions? Work the DESeq2 for both comparisons!

```
[ahnt@hopper:~/Course/bcb5250/2020S/labs/rna_seq/chrX_data]$ cat geuvadis_phenodata.csv
"ids","sex","population"
"ERR188044","male","YRI"
"ERR188104","male","YRI"
"ERR188234","female","YRI"
"ERR188245","female","GBR"
"ERR188257","male","GBR"
"ERR188273","female","YRI"
"ERR188337","female","GBR"
"ERR188383","male","GBR"
"ERR188401","male","GBR"
"ERR188428","female","GBR"
"ERR188454","male","YRI"
"ERR204916","female","YRI"
```