Differential expression with DSS (Dispersion Shrinkage for Sequencing data)

Hao Wu

Department of Biostatistics and Bioinformatics
Emory University
Atlanta, GA 303022

hao.wu@emory.edu

April 4, 2013

Contents

Abstract

This vignette introduces the use of Bioconductor package DSS (Dispersion Shrinkage for Sequencing data), which is designed primarily for differential expression detection for count data from RNA-seq. DSS uses new procedures to estimate and shrink gene-specific dispersions, then conduct Wald test for hypothesis testing. Compared to existing methods (DESeq and edgeR) DSS provides excellent statistical and computational performance, especially when overall dispersion level is high in data.

1 Introdution

RNA-seq is a new technology for measuring the abundance of RNA products in a biological sample. Compared to gene expression microarrays, it provides better dynamic ranges and lower signal-to-noise ratio, so it's quickly becoming the technology of choice for gene expression quantifications. One of the fundamental questions for RNA-seq data analyses is the regulation of gene expression under different biological contexts. Therefore identifying differential expression (DE) remains a key task in studying gene expression.

The major distinction of RNA-seq data compared to microarray is that the expression measurements are counts. Most of the existing statistical methods model the count data as over-dispersed Poisson, or negative binomial. The over dispersion parameters, which represent the biological variations for replicates within a treatment group, play a central role in the DE detection algorithm. There have been several statistical

methods and software tools available to perform DE detection from RNA-seq data, each with different procedures for dispersion estimation and hypothesis testing.

Here we present a new DE detection algorithm. First the gene specific dispersions are estimated through a method of moment estimator. Then data from all genes were combined to shrink dispersions through a penalized likelihood approach. Finally hypothesis testing is conducted using a Wald test. Results showed that the new method provide excellent performance compared to existing method, especially when overall dispersion level is high. The method is implemented in the Bioconductor package DSS, referring to Dispersion Shrinkage for Sequencing data.

Currently DSS only support comparison of expressions from two treatment groups. Methods for more advanced design is under development and will be implemented soon.

2 Getting started to use DSS

Required inputs for DSS are (1) gene expressions as a matrix of integers, rows are for genes and columns are for samples; and (2) a vector representing experimental designs. The length of the design vector must match the number of columns of input counts. Optionally, normalization factors or additional annotation for genes can be supplied.

The basic data container in the package is `SeqCountSet` class, which is directly inherited from `ExpressionSet` class defined in `Biobase`. An object of the class contains all necessary information for a DE analysis: gene expressions, experimental designs, and additional annotations.

A typical DE analysis contain following simple steps.

1. Create a `SeqCountSet` object using `newSeqCountSet`.
2. Estimate normalization factor using `estNormFactors`.
3. Estimate and shrink gene-wise dispersion using `estDispersion`
4. Two group comparison using `waldTest`.

The usage of DSS is demonstrated by below simple simulation.

1. First load in the library, and make a `SeqCountSet` object from some counts for 2000 genes and 6 samples.

```
> library(DSS)
> counts1=matrix(rnbinom(300, mu=10, size=10), ncol=3)
> counts2=matrix(rnbinom(300, mu=50, size=10), ncol=3)
> X1=cbind(counts1, counts2) ## these are 100 DE genes
> X2=matrix(rnbinom(11400, mu=10, size=10), ncol=6)
> X=rbind(X1,X2)
> designs=c(0,0,0,1,1,1)
> seqData=newSeqCountSet(X, designs)
> seqData
```

```
SeqCountSet (storageMode: lockedEnvironment)
assayData: 2000 features, 6 samples
  element names: exprs
protocolData: none
phenoData
  sampleNames: 1 2 ... 6 (6 total)
  varLabels: designs
  varMetadata: labelDescription
featureData: none
experimentData: use 'experimentData(object)'
Annotation:
```

2. Estimate normalization factor.

```
> seqData=estNormFactors(seqData)
```

3. Estimate and shrink gene-wise dispersions

```
> seqData=estDispersion(seqData)
```

4. With normalization factors and dispersions ready, two group comparison can be conducted via a wald test:

```
> result=waldTest(seqData, 0, 1)
> head(result,5)
```

```
   geneIndex      muA      muB       lfc   difExpr     stats         pval
74        74 5.216304 59.24796 -2.346813 -54.03165 -5.689586 1.273480e-08
96        96 6.540956 56.78256 -2.096252 -50.24160 -5.521114 3.368579e-08
81        81 7.182030 59.84519 -2.061197 -52.66316 -5.512709 3.533518e-08
24        24 7.862904 56.38749 -1.917270 -48.52459 -5.277646 1.308540e-07
11        11 3.273917 33.38360 -2.194818 -30.10968 -5.269244 1.369865e-07
      local.fdr          fdr
74 1.002985e-05 1.002985e-05
96 1.389234e-05 1.261873e-05
81 1.416870e-05 1.261873e-05
24 2.510610e-05 1.879196e-05
11 2.559068e-05 1.879196e-05
```

3 Session Info

```
> sessionInfo()
```

```
R version 3.0.0 (2013-04-03)
Platform: x86_64-unknown-linux-gnu (64-bit)
```

```
locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
 [3] LC_TIME=en_US.UTF-8        LC_COLLATE=C
 [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=C                 LC_NAME=C
 [9] LC_ADDRESS=C               LC_TELEPHONE=C
[11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
[1] splines   parallel  stats     graphics  grDevices utils     datasets
[8] methods   base

other attached packages:
[1] DSS_1.4.0          locfdr_1.1-7       Biobase_2.20.0     BiocGenerics_0.6.0

loaded via a namespace (and not attached):
[1] tools_3.0.0
```