

Differential analyses with DSS

Hao Wu

Department of Biostatistics and Bioinformatics

Emory University

Atlanta, GA 303022

`hao.wu@emory.edu`

October 14, 2013

Contents

1	Introduction	1
2	Using DSS for differential expression analysis	2
2.1	Single factor experiment	2
2.2	Multifactor experiment	4
3	Using DSS for differential methylation analysis	5
4	Session Info	6

Abstract

This vignette introduces the use of the Bioconductor package DSS (Dispersion Shrinkage for Sequencing data), which is designed for differential analysis based on high-throughput sequencing data. It performs differential expression analyses for RNA-seq, and differential methylation analyses for bisulfite sequencing (BS-seq) data. The core of DSS is a new procedure to estimate and shrink gene- or CpG site-specific dispersions, then conduct Wald tests for differential expression/methylation. Compared with existing methods, DSS provides excellent statistical and computational performance.

1 Introduction

Recent advances in high-throughput sequencing technology have revolutionized genomic research. For example, RNA-seq is a new technology for measuring the abundance of RNA products. Compared to gene expression microarrays, it provides a better dynamic range and lower signal-to-noise ratio. Bisulfite sequencing (BS-seq) is a new technology for measuring DNA methylation. Compared to capture-based methods such as MeDIP-seq, it provides single-base resolution and eliminates biases associated with CpG density.

Fundamental questions for RNA-seq or BS-seq data analyses are whether gene expression regulation or DNA methylation dynamics vary under different biological contexts. Identifying sites or regions exhibiting differential expression (DE) or differential methylation (DM) are thus key tasks in functional genomics research.

RNA- or BS-seq experiments typically have a limited number of biological replicates due to cost constraints. This can lead to unstable estimation of within group variance, and subsequently undesirable results from hypothesis testing. Variance shrinkage methods have been widely used in DE analyses based on microarray data. The methods are typically based on a Bayesian hierarchical model, with a prior imposed on the gene-specific variances to provide a basis for information sharing across all genes/CpG sites. In these models, shrinkage is achieved for variance estimation. Using shrunk variance in hypothesis tests has been shown to provide better results.

A distinct feature of RNA-seq or BS-seq data is that the measurements are in the form of counts. These data are often assumed to be from the Poisson (for RNA-seq) or Binomial (for BS-seq) distributions. Unlike continuous distributions such as the Gaussian distribution, the variances depend on means in these discrete distributions. This implies that the sample variances do not account for biological variation between replicates, and shrinkage cannot be applied on variances directly.

In contrast, we assume that our count data come from the Gamma-Poisson (RNA-seq) or Beta-Binomial (BS-seq) distribution. These distributions can be parameterized by a mean and an over dispersion parameter. The over dispersion parameters, which represent the biological variation for replicates within a treatment group, play a central role in the differential analyses.

Here we present a new DE/DM detection algorithm, where shrinkage is performed on the dispersion parameters. We first impose a log-normal prior on the dispersions, and then combine data from all genes/CpG sites to shrink dispersions through a penalized likelihood approach. Finally, we construct Wald tests to test each gene/site for differential expression/methylation. Our results show that the new method provides excellent performance compared to existing methods, especially when the overall dispersion level is high or the number of replicates is small.

Currently DSS only supports comparison of expression or methylation from two treatment groups. Methods for more advanced study designs are under development and will be implemented soon.

2 Using DSS for differential expression analysis

2.1 Single factor experiment

Required inputs for DSS are (1) gene expression values 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 expression values, experimental designs, and additional annotations.

A typical DE analysis contains the 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 in the simple simulation below.

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.
3. Estimate and shrink gene-wise dispersions
4. With the normalization factors and dispersions ready, the 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
73	73	5.641639	47.97247	-2.065904	-42.33083	-5.228211	1.711579e-07
87	87	4.982167	44.11108	-2.096482	-39.12891	-5.111175	3.201620e-07
43	43	7.958803	56.08815	-1.900592	-48.12935	-5.105646	3.296665e-07
48	48	5.982999	45.56272	-1.960821	-39.57972	-5.103779	3.329368e-07
9	9	5.978512	51.74425	-2.087439	-45.76574	-4.992589	5.957525e-07
		local.fdr	fdr				
73		7.150085e-06	7.150085e-06				
87		9.822037e-06	8.377943e-06				
43		1.005399e-05	8.377943e-06				
48		1.013231e-05	8.377943e-06				
9		1.548878e-05	1.181803e-05				

2.2 Multifactor experiment

DSS provides functionalities for dispersion shrinkage for multifactor experimental designs. Downstream model fitting (through genealized linear model) and hypothesis testing can be performed using other packages such as `edgeR`, with the dispersions estimated from DSS.

Below is an example, based a simple simulation, to illustrate the DE analysis of a crossed design.

1. First simulate data for a 2x2 crossed experiments. Note the counts are randomly generated.

```
> library(DSS)
> library(edgeR)
> counts=matrix(rpois(800, 10), ncol=8)
> design=data.frame(gender=c(rep("M",4), rep("F",4)), strain=rep(c("WT", "Mutant"),4))
> X=model.matrix(~gender+strain, data=design)
```

2. make `SeqCountSet`, then estimate size factors and dispersion

```
> seqData=newSeqCountSet(counts, as.data.frame(X))
> seqData=estNormFactors(seqData)
> seqData=estDispersion(seqData)
```

3. Using `edgeR`'s function to do glm model fitting, but plugging in the estimated size factors and dispersion from DSS.

```
> fit.edgeR <- glmFit(counts, X, lib.size=normalizationFactor(seqData),
+                     dispersion=dispersion(seqData))
```

4. Using `edgeR`'s function to do hypothesis testing on the second parameter of the model (gender).

```
> lrt.edgeR <- glmLRT(fit.edgeR, coef=2)
> head(lrt.edgeR$table)
```

	logFC	logCPM	LR	PValue
1	-0.03339248	21.16591	0.008657057	0.9258692

```

2 -0.01408807 21.28035 0.001658002 0.9675203
3 -0.19637867 21.28005 0.320211463 0.5714806
4 -0.58411149 21.09838 2.457641756 0.1169543
5 -0.55162190 20.82086 1.821212739 0.1771690
6 0.17642641 21.31012 0.269490754 0.6036736

```

3 Using DSS for differential methylation analysis

For BS-seq experiments, after sequence alignment and proper processing, the BS-seq data can be summarized with following information for each C position (mostly CpG sites, with the occasional CH): chromosome number, genomic coordinate, total number of reads covering the position, and number of reads showing methylation at this position. For a sample, this information needs to be saved in a simple text file, with each row representing a CpG site.

DML detection using DSS starts from several such text files. A typical DML detection contains two simple steps. Below we will use files distributed with the package to illustrate the usage of the package.

1. Load in library. Read in text files and create **BSseq** objects for two conditions. This step requires **bsseq** Bioconductor package. **BSseq** class is defined in that package.

```

> library(DSS)
> require(bsseq)
> path <- file.path(system.file(package="DSS"), "extdata")
> dat1.1 <- read.table(file.path(path, "cond1_1.txt"), header=TRUE)
> dat1.2 <- read.table(file.path(path, "cond1_2.txt"), header=TRUE)
> dat2.1 <- read.table(file.path(path, "cond2_1.txt"), header=TRUE)
> dat2.2 <- read.table(file.path(path, "cond2_2.txt"), header=TRUE)
> BS1 <- makeBSseqData( list(dat1.1, dat1.2), paste("cond1",1:2,sep=".") )
> BS2 <- makeBSseqData( list(dat2.1, dat2.2), paste("cond2",1:2,sep=".") )
> BS1

```

```

An object of type 'BSseq' with
  28913 methylation loci
  2 samples
has not been smoothed

```

2. Detect DML by calling **callDML** function.

```

> result <- callDML(BS1, BS2)
> head(result)

```

	chr	pos	mu1	mu2	diff	diff.se	stat
1	chr18	3014740	0.8750000	0.5000000	0.37500000	NaN	NaN
2	chr18	3014769	0.0625000	0.5000000	-0.43750000	NaN	NaN
3	chr18	3014904	0.5056180	0.4705882	0.03502974	0.12361857	0.283369577

```

4 chr18 3031032 0.3421053 0.1739130 0.16819222 0.11131261 1.510989785
5 chr18 3031044 0.3466667 0.3478261 -0.00115942 0.13574931 -0.008540893
6 chr18 3031065 0.4358974 0.3750000 0.06089744 0.09967494 0.610960343

      pval      fdr
1      NaN      NaN
2      NaN      NaN
3 0.7768935 1.0000000
4 0.1307911 0.9121284
5 0.9931854 1.0000000
6 0.5412258 1.0000000

```

Currently DSS only works for detecting **differentially methylated loci (DML)**. This means that a hypothesis test is conducted at each CpG site, and the tests are performed independently. The spatial correlations among CpG sites are ignored in the current version. For whole genome BS-seq data, it might be better to apply a smoothing technique to combine information across nearby CpG sites and improve the estimation of mean methylation levels. This is under development and will be implemented in the near future.

4 Session Info

```
> sessionInfo()
```

```
R version 3.0.2 (2013-09-25)
```

```
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=en_US.UTF-8     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] bsseq_0.10.0      matrixStats_0.8.12  GenomicRanges_1.14.0
[4] XVector_0.2.0     IRanges_1.20.0      edgeR_3.4.0
[7] limma_3.18.0      DSS_1.8.0           locfdr_1.1-7
[10] Biobase_2.22.0    BiocGenerics_0.8.0

```

```
loaded via a namespace (and not attached):
```

```
[1] R.methodsS3_1.5.2 RColorBrewer_1.0-5 colorspace_1.2-4 dichromat_2.0-0
[5] grid_3.0.2        labeling_0.2        lattice_0.20-24    locfit_1.5-9.1
[9] munsell_0.4.2     plyr_1.8            scales_0.2.3       stats4_3.0.2
[13] stringr_0.6.2     tools_3.0.2
```