Differential analyses with DSS

Hao Wu

Department of Biostatistics and Bioinformatics
Emory University
Atlanta, GA 303022

hao.wu@emory.edu

October 17, 2016

Contents

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 procedure based on Bayesian hierarchical model to estimate and shrink gene- or CpG site-specific dispersions, then conduct Wald tests for detecting differential expression/methylation.

1 Introduction

Recent advances in various high-throughput sequencing technologies have revolutionized genomics research. Among them, RNA-seq is designed to measure the the abundance of RNA products, and Bisulfite sequencing (BS-seq) is for measuring DNA methylation. A fundamental question in functional genomics research is whether gene expression or DNA methylation vary under different biological contexts. Thus, identifying differential expression genes (DEGs) or differential methylation loci/regions (DML/DMRs) are key tasks in RNA-seq or BS-seq data analyses.

The differential expression (DE) or differential methylation (DM) analyses are often based on gene- or CpG-specific statistical test. A key limitation in RNA- or BS-seq experiments is that the number of biological replicates is usually limited 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 applied in DE analyses in microarray data to improve the estimation of gene-specific within group variances. These 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.

A distinct feature of RNA-seq or BS-seq data is that the measurements are in the form of counts and have to be modeld by discrete distributions. Unlike continuous distributions (such as Gaussian), the variances depend on means in these discrete distributions. This implies that the sample variances do not account for biological variation, and shrinkage cannot be applied on variances directly. In DSS, we assume that the count data are from the Gamma-Poisson (for RNA-seq) or Beta-Binomial (for BS-seq) distribution. We then parameterize the distributions by a mean and a dispersion parameters. The dispersion parameters, which represent the biological variation for replicates within a treatment group, play a central role in the differential analyses.

DSS implements a series of DE/DM detection algorithms based on the dispersion shrinkage method followed by Wald statistical test to test each gene/CpG site for differential expression/methylation. It provides functions for RNA-seq DE analysis for both two group comparision and multi-factor design, BS-seq DM analysis for two group comparision, multi-factor design, and data without biological replicate. Simulation and real data results show that the methods provides excellent performance compared to existing methods, especially when the overall dispersion level is high or the number of replicates is small.

For more details of the data model, the shrinkage method, and test procedures, please read [4] for differential expression from RNA-seq, [1] for differential methylation for two-group comparison from BS-seq, [2] for differential methylation for data without biological replicate, and [3] for differential methylation for general experimental design.

2 Using DSS for differential expression analysis

2.1 Input data preparation

DSS requires a count table (a matrix of **integers**) for gene expression values (rows are for genes and columns are for samples). This is different from the isoform expression based analysis such as in cufflink/cuffdiff, where the gene expressions are represented as non-integers values. There are a number of ways to obtain the count table from raw

sequencing data (fastq file), here we provide some example codes using several Bioconductor packages (the codes require installation of `GenomicFeatures`, `Rsamtools`, and `GenomicRanges` packages).

1. Sequence alignment. There are several RNA-seq aligner, for example, `tophat` or `STAR`. Assume the alignment result is saved in a BAM file `data.bam`.
2. Choose a gene annotation. `GenomicFeatures` package provides a convenient way to access different gene annotations. For example, if one wants to use RefSeq annotation for human genome build hg19, one can use following codes:

   ```
   > library(GenomicFeatures)
   > txdb = makeTranscriptDbFromUCSC(genom="hg19",tablename="refGene")
   > genes = genes(txdb)
   ```
3. Obtain count table based on the alignment results and gene annotation. This can be done in several steps. First read in the BAM file using the `Rsamtools` package:

   ```
   > bam=scanBam("data.bam")
   ```

   Next, create `GRanges` object for the aligned sequence reads.

   ```
   > IRange.reads=GRanges(seqnames=Rle(bam$rname), ranges=IRanges(bam$pos, width=bam$qwidth))
   ```

   Finally, use the `countOverlaps` function in `GenomicRanges` function to obtain the read counts overlap each gene.

   ```
   > counts = countOverlaps(genes, IRange.reads)
   ```

There are other ways to obtain the counts, for example, using `QuasR` or `easyRNASeq` Bioconductor package. Please refer to the package vignettes for more details.

2.2 Single factor experiment

In single factor RNA-seq experiment, DSS also requires a vector representing experimental designs. The length of the design vector must match the number of columns of the count table. 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.

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

3. Estimate and shrink gene-wise dispersions

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

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    local.fdr
46        46 8.333333 69.77778 -2.073923 -61.44444 -5.745911 9.142750e-09 6.311273e-05
54        54 7.666667 65.90303 -2.095682 -58.23636 -5.731582 9.949833e-09 6.311273e-05
66        66 5.333333 53.70505 -2.229185 -48.37172 -5.728543 1.012967e-08 6.311273e-05
5          5 4.333333 49.05455 -2.327538 -44.72121 -5.705159 1.162344e-08 6.365709e-05
95        95 3.666667 44.35152 -2.376241 -40.68485 -5.671783 1.413185e-08 7.183876e-05
            fdr
46 6.311273e-05
54 6.311273e-05
66 6.311273e-05
5  6.311273e-05
95 6.311273e-05
```

A higher level wrapper function `DSS.DE` is provided for simple RNA-seq DE analysis in a two-group comparison. User only needs to provide a count matrix and a vector of 0's and 1's representing the design, and get DE test results in one line. A simple example is listed below:

```
> counts = matrix(rpois(600, 10), ncol=6)
> designs = c(0,0,0,1,1,1)
> result = DSS.DE(counts, designs)
```

```
> head(result)
```

```
   geneIndex       muA       muB        lfc   difExpr     stats       pval local.fdr
66        66 12.666667  5.982193  0.7086294  6.684473  2.407593 0.01605807 0.3575555
96        96  8.666667 14.555080 -0.4961417 -5.888413 -1.873248 0.06103413 0.6955445
23        23 13.000000  7.984330  0.4644688  5.015670  1.695238 0.09003029 1.0000000
54        54 13.666667  8.542497  0.4489565  5.124170  1.674618 0.09400919 1.0000000
90        90 11.000000  6.631054  0.4778880  4.368946  1.636723 0.10168835 1.0000000
44        44  7.333333 11.774690 -0.4491513 -4.441357 -1.590703 0.11167642 0.6869399
         fdr
66 0.3575555
96 0.7329983
23 0.7642020
54 0.7725718
90 0.7739964
44 0.7725718
```

2.3 Multifactor experiment

DSS provides functionalities for dispersion shrinkage for multifactor experimental designs. Downstream model fitting (through genearlized 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(glmfit=fit.edgeR, coef=2)
   > head(lrt.edgeR$table)
            logFC   logCPM           LR    PValue
   ```

```
1  0.178865387 21.25400 2.498957e-01 0.6171485
2 -0.178334349 21.02475 2.114130e-01 0.6456622
3 -0.478846041 21.13138 1.564911e+00 0.2109475
4 -0.172443332 20.86399 1.773499e-01 0.6736606
5 -0.098105851 21.08778 6.474474e-02 0.7991480
6  0.001844588 21.29552 2.627074e-05 0.9959105
```

3 Using DSS for differential methylation analysis

3.1 Overview

To detect differential methylation, statistical tests are conducted at each CpG site, and then the differential methylation loci (DML) or differential methylation regions (DMR) are called based on user specified threshold. A rigorous statistical tests should account for biological variations among replicates and the sequencing depth. Most existing methods for DM analysis are based on *ad hoc* methods. For example, using Fisher's exact ignores the biological variations, using t-test on estimated methylation levels ignores the sequencing depth. Sometimes arbitrary filtering are implemented: loci with depth lower than an arbitrary threshold are filtered out, which results in information loss

The DM detection procedure implemented in DSS is based on a rigorous Wald test for beta-binomial distributions. The test statistics depend on the biological variations (characterized by dispersion parameter) as well as the sequencing depth. An important part of the algorithm is the estimation of dispersion parameter, which is achieved through a shrinkage estimator based on a Bayesian hierarchical model [1]. An advantage of DSS is that the test can be performed even when there is no biological replicates. That's because by smoothing, the neighboring CpG sites can be viewed as "pseudo-replicates", and the dispersion can still be estimated with reasonable precision.

DSS also works for general experimental design, based on a beta-binomial regression model with "arcsine" link function. Model fitting is performed on transformed data with generalized least square method, which achieves much improved computational performance compared with methods based on generalized linear model.

DSS depends on `bsseq` Bioconductor package, which has neat definition of data structures and many useful utility functions. In order to use the DM detection functionalities, `bsseq` needs to be pre-installed.

3.2 Input data preparation

DSS requires data from each BS-seq experiment to be summarized into following information for each CG position: chromosome number, genomic coordinate, total number of reads, and number of reads showing methylation. For a sample, this information are saved in a simple text file, with each row representing a CpG site. Below shows an example of a small part of such a file:

```
chr     pos     N       X
chr18   3014904 26      2
chr18   3031032 33      12
chr18   3031044 33      13
```

```
chr18   3031065 48      24
```

One can follow below steps to obtain such data from raw sequence file (fastq file), using `bismark` (version 0.10.0, commands for newer versions could be different) for BS-seq alignment and count extraction. These steps require installation of `bowtie` or `bowtie2`, `bismark`, and the fasta file for reference genome.

1. Prepare Bisulfite reference genome. This can be done using the `bismark_genome_preparation` function (details in bismark manual). Example command is:
   `bismark_genome_preparation -path_to_bowtie /usr/local/bowtie/ -verbose /path/to/refgenomes/`
2. BS-seq alignment. Example command is:
   `bismark -q -n 1 -l 50 -path_to_bowtie /path/bowtie/ BS-refGenome reads.fastq`
   This step will produce two text files `reads.fastq_bismark.sam` and `reads.fastq_bismark_SE_report.txt`.
3. Extract methylation counts using `bismark_methylation_extractor` function:
   `bismark_methylation_extractor -s -bedGraph reads.fastq_bismark.sam`. This will create multiple txt files to summarize methylation call and cytosine context, a bedGraph file to display methylation percentage, and a coverage file containing counts information. The count file contain following columns:`chr, start, end, methylation%, count methylated, count unmethylated`. This file can be modified to make the input file for DSS.

A typical DML detection contains two simple steps. First one conduct DM test at each CpG site, then DML/DMR are called based on the test result and user specified threshold.

3.3 DML/DMR detection from two-group comparison

Below are the steps to call DML or DMR for BS-seq data in two-group comparison setting.

1. Load in library. Read in text files and create an object of `BSseq` class, which is defined in `bsseq` Bioconductor package. 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)
> BSobj <- makeBSseqData( list(dat1.1, dat1.2, dat2.1, dat2.2),
+       c("C1","C2", "N1", "N2") )[1:10000,]
> BSobj
An object of type 'BSseq' with
  10000 methylation loci
  4 samples
has not been smoothed
```

2. Perform statistical test for DML by calling `DMLtest` function. This function basically performs following steps: (1) estimate mean methylation levels for all CpG site; (2) estimate dispersions at each CpG sites; (3) conduct Wald

test. For the first step, there's an option for smoothing or not. Because the methylation levels show strong spatial correlations, smoothing can help obtain better estimates of mean methylation when the CpG sites are dense in the data (such as from the whole-genome BS-seq). However for data with sparse CpG, such as from RRBS or hydroxyl-methylation, smoothing is not recommended.

To perform DML test without smoothing, do:

```
> dmlTest <- DMLtest(BSobj, group1=c("C1", "C2"), group2=c("N1", "N2"))
Estimating dispersion for each CpG site, this will take a while ...
> head(dmlTest)
    chr     pos       mu1       mu2        diff    diff.se       stat        phi1
1 chr18 3014904 0.3850276 0.4623677 -0.07734011 0.24899738 -0.3106061 0.286610813
2 chr18 3031032 0.3384423 0.1416667  0.19677555 0.11402507  1.7257217 0.009525212
3 chr18 3031044 0.3436302 0.3299846  0.01364560 0.12508174  0.1090935 0.010959806
4 chr18 3031065 0.4372540 0.3646882  0.07256585 0.10435917  0.6953471 0.010846613
5 chr18 3031069 0.2939132 0.5397749 -0.24586172 0.13554689 -1.8138499 0.012965376
6 chr18 3031082 0.3529066 0.3903499 -0.03744333 0.08197055 -0.4567900 0.008246131
        phi2       pval       fdr
1 0.01964971 0.75610007 0.9999077
2 0.05282909 0.08439748 0.7175525
3 0.02285714 0.91312835 0.9999077
4 0.01849672 0.48683778 0.9999077
5 0.02658296 0.06970083 0.6313442
6 0.01338446 0.64782202 0.9999077
```

To perform statistical test for DML with smoothing, do:

```
> dmlTest.sm <- DMLtest(BSobj, group1=c("C1", "C2"), group2=c("N1", "N2"), smoothing=TRUE)
Smoothing ...
Estimating dispersion for each CpG site, this will take a while ...
```

There are two options for smoothing: a simple moving average, or the BSmooth method implemented in bsseq package. The BSmooth method produces much smoother curve, which is good for visualization purpose. However, it is very computationally intensive, and the results are not very different from moving average in terms of DMR calling. So we recommend using moving average. Smoothing span is an important parameter in smoothing procedure and have non-trivial impact on DMR calling. We use 500 bp as default, and think that it performs well in real data tests.

3. With the test results, one can call DML by using `callDML` function. The results DMLs are sorted by the significance.

```
>   dmls <- callDML(dmlTest, p.threshold=0.001)
>   head(dmls)
      chr     pos        mu1        mu2       diff    diff.se       stat       phi1
450 chr18 3976129 0.01048590 0.93773387 -0.9272480 0.06785915 -13.664303 0.04319795
451 chr18 3976138 0.01048590 0.93773387 -0.9272480 0.06785915 -13.664303 0.04319795
582 chr18 4340682 0.96365233 0.03162952  0.9320228 0.09947722   9.369209 0.06181184
583 chr18 4340709 0.96365233 0.03162952  0.9320228 0.09947722   9.369209 0.06181184
638 chr18 4431501 0.01325735 0.94195478 -0.9286974 0.09390549  -9.889704 0.04380187
639 chr18 4431511 0.01321013 0.94195478 -0.9287446 0.09387324  -9.893604 0.04377619
```

```
          phi2 pval fdr postprob.overThreshold
450 0.02756948    0   0                     1
451 0.02756948    0   0                     1
582 0.05147878    0   0                     1
583 0.05147878    0   0                     1
638 0.08184949    0   0                     1
639 0.08184949    0   0                     1
```

By default, the test is based on the null hypothesis that the difference in methylation levels is 0. Alternatively, users can specify a threshold for difference. For example, to detect loci with difference greater than 0.1, do:

```
> dmls2 <- callDML(dmlTest, delta=0.1, p.threshold=0.001)
> head(dmls2)
```

```
      chr     pos        mu1        mu2       diff    diff.se       stat       phi1
450 chr18 3976129 0.01048590 0.93773387 -0.9272480 0.06785915 -13.664303 0.04319795
451 chr18 3976138 0.01048590 0.93773387 -0.9272480 0.06785915 -13.664303 0.04319795
582 chr18 4340682 0.96365233 0.03162952  0.9320228 0.09947722   9.369209 0.06181184
583 chr18 4340709 0.96365233 0.03162952  0.9320228 0.09947722   9.369209 0.06181184
638 chr18 4431501 0.01325735 0.94195478 -0.9286974 0.09390549  -9.889704 0.04380187
639 chr18 4431511 0.01321013 0.94195478 -0.9287446 0.09387324  -9.893604 0.04377619
          phi2 pval fdr postprob.overThreshold
450 0.02756948    0   0                     1
451 0.02756948    0   0                     1
582 0.05147878    0   0                     1
583 0.05147878    0   0                     1
638 0.08184949    0   0                     1
639 0.08184949    0   0                     1
```

When delta is specified, the function will compute the posterior probability that the difference of the means is greater than delta. So technically speaking, the threshold for p-value here actually refers to the threshold for 1-posterior probability, or the local FDR. Here we use the same parameter name for the sake of the consistence of function syntax.

4. DMR detection is also Based on the DML test results, by calling `callDMR` function. Regions with many statistically significant CpG sites are identified as DMRs. Some restrictions are provided by users, including the minimum length, minimum number of CpG sites, percentage of CpG site being significant in the region, etc. There are some *post hoc* procedures to merge nearby DMRs into longer ones.

   ```
   > dmrs <- callDMR(dmlTest, p.threshold=0.01)
   > head(dmrs)
   ```

   ```
       chr    start      end length nCG meanMethy1 meanMethy2 diff.Methy  areaStat
   142 chr18 13131705 13131880    176   6  0.9689797 0.06107942  0.9079003 68.741081
   26  chr18  4657576  4657639     64   4  0.5084704 0.31870495  0.1897655 14.421382
   30  chr18  5027578  5027743    166   4  0.7045195 0.38058436  0.3239352  9.109377
   ```

   Here the DMRs are sorted by "areaStat", which is defined in `bsseq` as the sum of the test statistics of all CpG sites within the DMR.

   Similarly, users can specify a threshold for difference. For example, to detect regions with difference greater than

0.1, do:

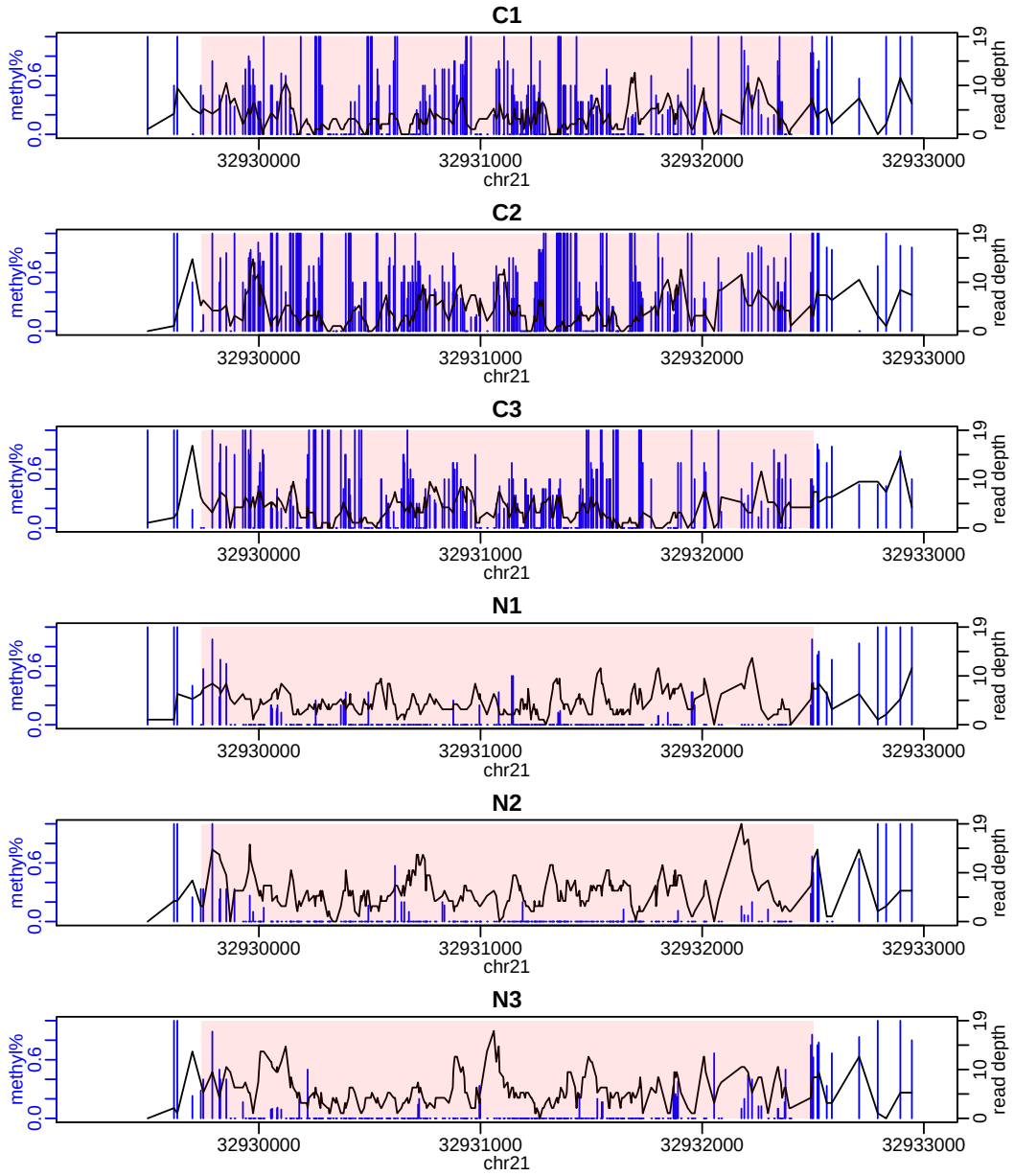
```
> dmrs2 <- callDMR(dmlTest, delta=0.1, p.threshold=0.05)
> head(dmrs2)
      chr    start      end length nCG meanMethy1 meanMethy2 diff.Methy areaStat
175 chr18 13131705 13131880    176   6  0.9689797 0.06107942 0.90790031 68.74108
29  chr18  4657576  4657639     64   4  0.5084704 0.31870495 0.18976548 14.42138
277 chr18 23100427 23100601    175   4  0.5329967 0.43403678 0.09895989 12.76449
18  chr18  4222533  4222608     76   4  0.7882151 0.36126847 0.42694665 12.66181
33  chr18  5027549  5027743    195   5  0.6696348 0.32605410 0.34358066 11.42014
```

Note that the distribution of test statistics (and p-values) depends on the differences in methylation levels and biological variations, as well as technical factors such as coverage depth. It is very difficulty to select a natural and rigorous threshold for defining DMRs. We recommend users try different thresholds in order to obtain satisfactory results.

5. The DMRs can be visualized using `showOneDMR` function, This function provides more information than the `plotRegion` function in `bsseq`. It plots the methylation percentages as well as the coverage depths at each CpG sites, instead of just the smoothed curve. So the coverage depth information will be available in the figure. To use the function, do

   ```
   > showOneDMR(dmrs[1,], BSobj)
   ```

   The result figure looks like the following. **Note that the figure below is not generated from the above example. The example data are from RRBS experiment so the DMRs are much shorter.**



3.4 DML/DMR detection from general experimental design

In DSS, BS-seq data from a general experimental design (such as crossed experiment, or experiment with covariates) is modeled through a generalized linear model framework. We use "arcsine" link function instead of the typical logit link for it better deals with data at boundaries (methylation levels close to 0 or 1). Linear model fitting is done through ordinary least square on transformed methylation levels. Standard errors for the estimates are derived with consideration of count data distribution and transformation. A Wald test is applied to perform hypothesis testing.

1. Load in data distributed with DSS. This is a small portion of a set of RRBS experiments. There are 5000 CpG sites and 16 samples. The experiment is a 2 design (2 cases and 2 cell types). There are 4 replicates in each case:cell combination.

```
> data(RRBS)
> RRBS
An object of type 'BSseq' with
  5000 methylation loci
  16 samples
has not been smoothed
> design
   case cell
1    HC   rN
2    HC   rN
3    HC   rN
4   SLE   aN
5   SLE   rN
6   SLE   aN
7   SLE   rN
8   SLE   aN
9   SLE   rN
10  SLE   aN
11  SLE   rN
12   HC   aN
13   HC   aN
14   HC   aN
15   HC   aN
16   HC   rN
```

2. Fit a linear model using `DMLfit.multiFactor` function, include case, cell, and case:cell interaction.

```
> DMLfit = DMLfit.multiFactor(RRBS, design=design, formula=~case+cell+case:cell)
Fitting DML model for CpG site:
```

3. Use `DMLtest.multiFactor` function to test the cell effect. It is important to note that the `coef` parameter is the index of the coefficient to be tested for being 0. Because the model (as specified by `formula` in `DMLfit.multiFactor`) include intercept, the cell effect is the 3rd column in the design matrix, so we use `coef=3` here.

```
> DMLtest.cell = DMLtest.multiFactor(DMLfit, coef=3)
```

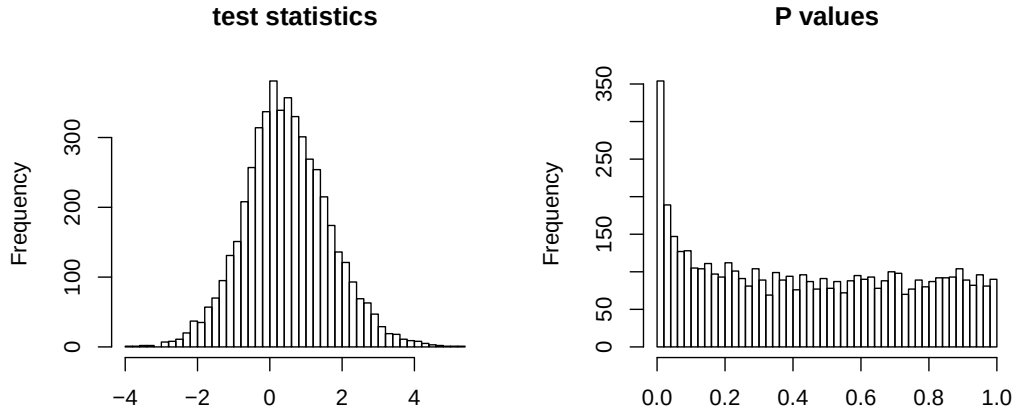
Result from this step is a data frame with chromosome number, CpG site position, test statistics, p-values (from normal distribution), and FDR. Rows are sorted by chromosome/position of the CpG sites. To obtain top ranked CpG sites, one can sort the data frame using following codes:

```
> ix=sort(DMLtest.cell[,"pvals"], index.return=TRUE)$ix
> head(DMLtest.cell[ix,])
      chr     pos     stat        pvals         fdrs
1273 chr1 2930315 5.280301 1.289720e-07 0.0006448599
4706 chr1 3321251 5.037839 4.708164e-07 0.0011770409
3276 chr1 3143987 4.910412 9.088510e-07 0.0015147517
2547 chr1 3069876 4.754812 1.986316e-06 0.0024828953
```

```
3061 chr1 3121473 4.675736 2.929010e-06 0.0029290097
527  chr1 2817715 4.441198 8.945925e-06 0.0065858325
```

Below is a figure showing the distributions of test statistics and p-values from this example dataset

```
> par(mfrow=c(1,2))
> hist(DMLtest.cell$stat, 50, main="test statistics", xlab="")
> hist(DMLtest.cell$pvals, 50, main="P values", xlab="")
```



These procedures are computationally very efficient. For a typical RRBS dataset with 4 million CpG sites, it takes around 20 minutes. In comparison, other methods such as RADMeth or BiSeq takes at least 10 times longer.

4 Session Info

```
> sessionInfo()

R version 3.3.1 (2016-06-21)
Platform: x86_64-pc-linux-gnu (64-bit)
Running under: Ubuntu 16.04.1 LTS

locale:
 [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C               LC_TIME=en_US.UTF-8
 [4] LC_COLLATE=C               LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
 [7] LC_PAPER=en_US.UTF-8       LC_NAME=C                  LC_ADDRESS=C
[10] LC_TELEPHONE=C             LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C

attached base packages:
 [1] splines   stats4    parallel  stats     graphics  grDevices utils     datasets
 [9] methods   base

other attached packages:
 [1] edgeR_3.16.0               DSS_2.14.0                 bsseq_1.10.0
 [4] limma_3.30.0               SummarizedExperiment_1.4.0 GenomicRanges_1.26.0
 [7] GenomeInfoDb_1.10.0        IRanges_2.8.0              S4Vectors_0.12.0
[10] Biobase_2.34.0             BiocGenerics_0.20.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.12.7         XVector_0.14.0      zlibbioc_1.20.0     munsell_0.4.3
 [5] colorspace_1.2-7    lattice_0.20-34     plyr_1.8.4          tools_3.3.1
 [9] grid_3.3.1          data.table_1.9.6    R.oo_1.20.0         gtools_3.5.0
[13] matrixStats_0.51.0 permute_0.9-4       Matrix_1.2-7.1      R.utils_2.4.0
[17] R.methodsS3_1.7.1  scales_0.4.0        locfit_1.5-9.1      BiocStyle_2.2.0
[21] chron_2.3-47
```

References

[1] Hao Feng, Karen Conneely and Hao Wu. (2014). A bayesian hierarchical model to detect differentially methylated loci from single nucleotide resolution sequencing data. *Nucleic Acids Research*. **42**(8), e69–e69.

[2] Hao Wu, Tianlei Xu, Hao Feng, Li Chen, Ben Li, Bing Yao, Zhaohui Qin, Peng Jin and Karen N. Conneely. (2015). Detection of differentially methylated regions from whole-genome bisulfite sequencing data without replicates. *Nucleic Acids Research*. doi: 10.1093/nar/gkv715.

[3] Yongseok Park, Hao Wu. (2016). Differential methylation analysis for BS-seq data under general experimental design. *Bioinformatics*. doi:10.1093/bioinformatics/btw026.

[4] Hao Wu, Chi Wang and Zhijing Wu. (2013). A new shrinkage estimator for dispersion improves differential expression detection in RNA-seq data. *Biostatistics*. **14**(2), 232–243.