Differential analyses with DSS

Hao Wu

Department of Biostatistics and Bioinformatics
Emory University
Atlanta, GA 303022

hao.wu@emory.edu

April 16, 2015

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 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.

For details of the hierarchical model, the shrinkage method and test procedure, please read [2] for differential expression from RNA-seq, and [1] for differential methylation from BS-seq.

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.

```
> 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
29        29 9.000000 73.71429 -2.055665 -64.71429 -5.517181 3.444811e-08 6.708566e-05
5          5 6.666667 58.82540 -2.113597 -52.15873 -5.431086 5.601203e-08 8.140575e-05
80        80 6.666667 54.48016 -2.037532 -47.81349 -5.315136 1.065778e-07 1.258168e-04
96        96 8.666667 61.00000 -1.903463 -52.33333 -5.278586 1.301843e-07 1.448197e-04
89        89 6.333333 49.47222 -1.989655 -43.13889 -5.254035 1.488028e-07 1.575847e-04
            fdr
```

```
29 6.708566e-05
5  7.875000e-05
80 9.310724e-05
96 1.108230e-04
89 1.108230e-04
```

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
17        17 13.333333  5.042903  0.9145628  8.290430  3.021383 0.002516232 0.3599973
25        25 10.333333  4.912871  0.6938482  5.420462  2.218661 0.026509793 1.0000000
24        24  5.666667 11.201978 -0.6405995 -5.535312 -2.148014 0.031712605 0.4075462
76        76 11.333333  5.825742  0.6262931  5.507591  2.108601 0.034979024 1.0000000
82        82  8.333333 14.564355 -0.5337989 -6.231021 -2.021689 0.043208524 0.4914115
3          3  6.666667 11.926731 -0.5504092 -5.260064 -1.931947 0.053365987 0.5708152
         fdr
17 0.3599973
25 0.6821120
24 0.5172543
76 0.7282260
82 0.5592795
3  0.5903341
```

2.2 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  0.30843570 21.46165 0.92899995 0.3351229
2 -0.17926576 21.21919 0.25999065 0.6101266
3  0.09716211 21.21826 0.07720362 0.7811234
4 -0.07388551 21.17219 0.04284388 0.8360195
5 -0.07001393 21.25957 0.03810107 0.8452403
6  0.30216514 20.88942 0.59507396 0.4404637
```

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 coverage 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 coverage depth. Sometimes arbitrary filtering are recommended: loci with coverages lower than an arbitrary threshold are filtered out.

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 (captured by dispersion parameter) as well as the coverage 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].

A great advantage of DSS is that the test can be performed even when there is no biological replicates. That's because by smoothing procedure, the neighboring CpG sites can be viewed as "pseudo-replicates", and the dispersion can still be estimated with reasonable precision.

This package 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

For one BS-seq experiments, after sequence alignment and some processing, the BS-seq data are usually 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
chr18   3031069 17      4
chr18   3031082 93      37
chr18   3031089 76      25
chr18   3031092 76      28
```

DML/DMR detection using DSS starts from several such text files. 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. Below we will use example data distributed with DSS to illustrate these steps.

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
3 chr18 3014904 0.5056818 0.4696970  0.035984848 0.23249555  0.15477650 0.287926584
4 chr18 3031032 0.3400000 0.1590909  0.180909091 0.11782933  1.53534850 0.009589628
5 chr18 3031044 0.3445946 0.3409091  0.003685504 0.13210417  0.02789847 0.011041061
6 chr18 3031065 0.4353448 0.3723404  0.063004402 0.10361769  0.60804676 0.010908903
7 chr18 3031069 0.3000000 0.5370370 -0.237037037 0.13990809 -1.69423398 0.013059144
8 chr18 3031082 0.3506787 0.3950000 -0.044321267 0.07762943 -0.57093380 0.008283076
        phi2       pval       fdr
3 0.01992909 0.87699752 1.0000000
4 0.05528545 0.12469825 0.8061525
5 0.02341821 0.97774313 1.0000000
6 0.01877165 0.54315646 1.0000000
7 0.02732521 0.09022083 0.6975466
8 0.01352427 0.56804452 1.0000000
```

To perform statistical test for DML with smoothing, do:

```
> dmlTest.sm <- DMLtest(BSobj, group1=c("C1", "C2"), group2=c("N1", "N2"), smoothing=TRUE)
Smoothing, this will take a while ...
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
507 chr18 3976129 0.006849315 0.95161290 -0.9447636 0.05519960 -17.11541 0.051293128
508 chr18 3976138 0.006849315 0.95161290 -0.9447636 0.05519960 -17.11541 0.051293128
709 chr18 4431501 0.006944444 0.97368421 -0.9667398 0.05818870 -16.61387 0.052066186
710 chr18 4431511 0.006849315 0.97368421 -0.9668349 0.05812148 -16.63473 0.052003498
789 chr18 4564237 0.918478261 0.02083333  0.8976449 0.05836718  15.37928 0.010057974
863 chr18 4657576 0.987179487 0.06250000  0.9246795 0.05876625  15.73487 0.009968422
          phi2 pval fdr postprob.overThreshold
507 0.02531726    0   0                     1
508 0.02531726    0   0                     1
709 0.07025639    0   0                     1
710 0.07025639    0   0                     1
```

```
789 0.06975075    0   0                       1
863 0.07423324    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
507 chr18 3976129 0.006849315 0.95161290 -0.9447636 0.05519960 -17.11541 0.051293128
508 chr18 3976138 0.006849315 0.95161290 -0.9447636 0.05519960 -17.11541 0.051293128
709 chr18 4431501 0.006944444 0.97368421 -0.9667398 0.05818870 -16.61387 0.052066186
710 chr18 4431511 0.006849315 0.97368421 -0.9668349 0.05812148 -16.63473 0.052003498
789 chr18 4564237 0.918478261 0.02083333  0.8976449 0.05836718  15.37928 0.010057974
863 chr18 4657576 0.987179487 0.06250000  0.9246795 0.05876625  15.73487 0.009968422
          phi2 pval fdr postprob.overThreshold
507 0.02531726    0   0                       1
508 0.02531726    0   0                       1
709 0.07025639    0   0                       1
710 0.07025639    0   0                       1
789 0.06975075    0   0                       1
863 0.07423324    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)
NULL
```

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:

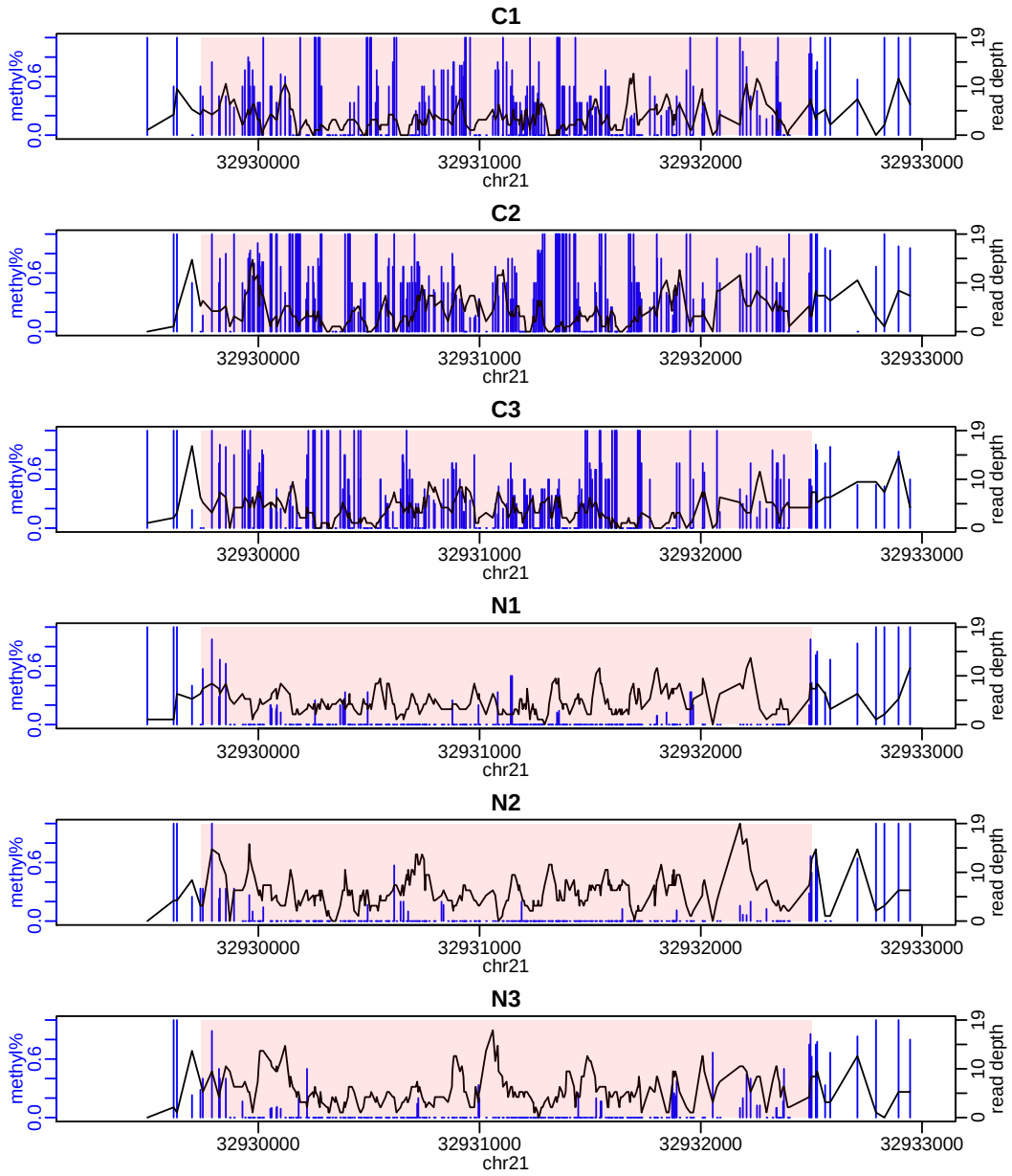
```
> dmrs2 <- callDMR(dmlTest, delta=0.1, p.threshold=0.05)
> head(dmrs2)
      chr    start      end length nCG meanMethy1 meanMethy2 diff.Methy   areaStat
22  chr18  5114580  5114712    133   5  0.9921612  0.4121153  0.5800459 18.5069895
160 chr18 19526612 19529997   3386   4  0.5477310  0.2919271  0.2558039 15.3140699
14  chr18  4222533  4222608     76   4  0.7874547  0.3784722  0.4089825 12.3371481
9   chr18  3542620  3542732    113   4  0.7666223  0.7077564  0.0588659  0.9855913
```

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.



4 Session Info

```
> sessionInfo()

R version 3.2.0 RC (2015-04-08 r68161)
Platform: x86_64-apple-darwin13.4.0 (64-bit)
Running under: OS X 10.9.5 (Mavericks)

locale:
[1] C/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8

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

other attached packages:
 [1] edgeR_3.10.0         limma_3.24.0         DSS_2.6.0            bsseq_1.4.0
 [5] matrixStats_0.14.0   GenomicRanges_1.20.0 GenomeInfoDb_1.4.0   IRanges_2.2.0
 [9] S4Vectors_0.6.0      Biobase_2.28.0       BiocGenerics_0.14.0

loaded via a namespace (and not attached):
 [1] Rcpp_0.11.5      locfit_1.5-9.1   lattice_0.20-31  gtools_3.4.2     plyr_1.8.1
 [6] grid_3.2.0       scales_0.2.4     XVector_0.8.0    BiocStyle_1.6.0  tools_3.2.0
[11] munsell_0.4.2    colorspace_1.2-6
```

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, 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.