segmentSeq: methods for detecting methylation loci and differential methylation

Thomas J. Hardcastle

October 17, 2016

1 Introduction

This vignette introduces analysis methods for data from high-throughput sequencing of bisulphite treated DNA to detect cytosine methylation. The segmentSeq package was originally designed to detect siRNA loci [1] and many of the methods developed for this can be used to detect loci of cytosine methylation from replicated (or unreplicated) sequencing data.

2 Preparation

Preparation of the segmentSeq package proceeds as in siRNA analysis. We begin by loading the segmentSeq package.

```
>   library(segmentSeq)
```

Note that because the experiments that segmentSeq is designed to analyse are usually massive, we should use (if possible) parallel processing as implemented by the parallel package. If using this approach, we need to begin by define a *cluster*. The following command will use eight processors on a single machine; see the help page for 'makeCluster' for more information. If we don't want to parallelise, we can proceed anyway with a NULL cluster. Results may be slightly different depending on whether or not a cluster is used owing to the non-deterministic elements of the method.

```
> if(require("parallel"))
+ {
+     numCores <- min(8, detectCores())
+     cl <- makeCluster(numCores)
+ } else {
+     cl <- NULL
+ }
```

The segmentSeq package is designed to read in output from the YAMA (Yet Another Methylome Aligner) program. This is a perl-based package using either bowtie or bowtie2 to align bisulphite treated reads (in an unbiased manner) to a reference and identify the number of times each cytosine is identified as methylated or unmethylated. Unlike most other aligners, YAMA does not require that reads that map to more than one location are discarded, instead it reports the number of alternate matches to the reference for each cytosine. This is then used by segmentSeq to weight the observed number of methylated/un-methylated cytosines at a location. The files used here have been compressed to save space.

```
> datadir <- system.file("extdata", package = "segmentSeq")
> files <- c("short_18B_C24_C24_trim.fastq_CG_methCalls.gz",
+ "short_Sample_17A_trimmed.fastq_CG_methCalls.gz",
+ "short_13_C24_col_trim.fastq_CG_methCalls.gz",
+ "short_Sample_28_trimmed.fastq_CG_methCalls.gz")
> mD <- readMeths(files = files, dir = datadir,
+ libnames = c("A1", "A2", "B1", "B2"), replicates = c("A","A","B","B"),
+ nonconversion = c(0.004777, 0.005903, 0.016514, 0.006134))
```

We can begin by plotting the distribution of methylation for these samples. The distribution can be plotted for each sample individually, or as an average across multiple samples. We can also subtract one distribution from another to visualise patterns of differential methylation on the genome.

```
> par(mfrow = c(2,1))
> dists <- plotMethDistribution(mD, main = "Distributions of methylation", chr = "Chr1")
> plotMethDistribution(mD, subtract = rowMeans(sapply(dists, function(x) x[,2])), main = "Differences b
```

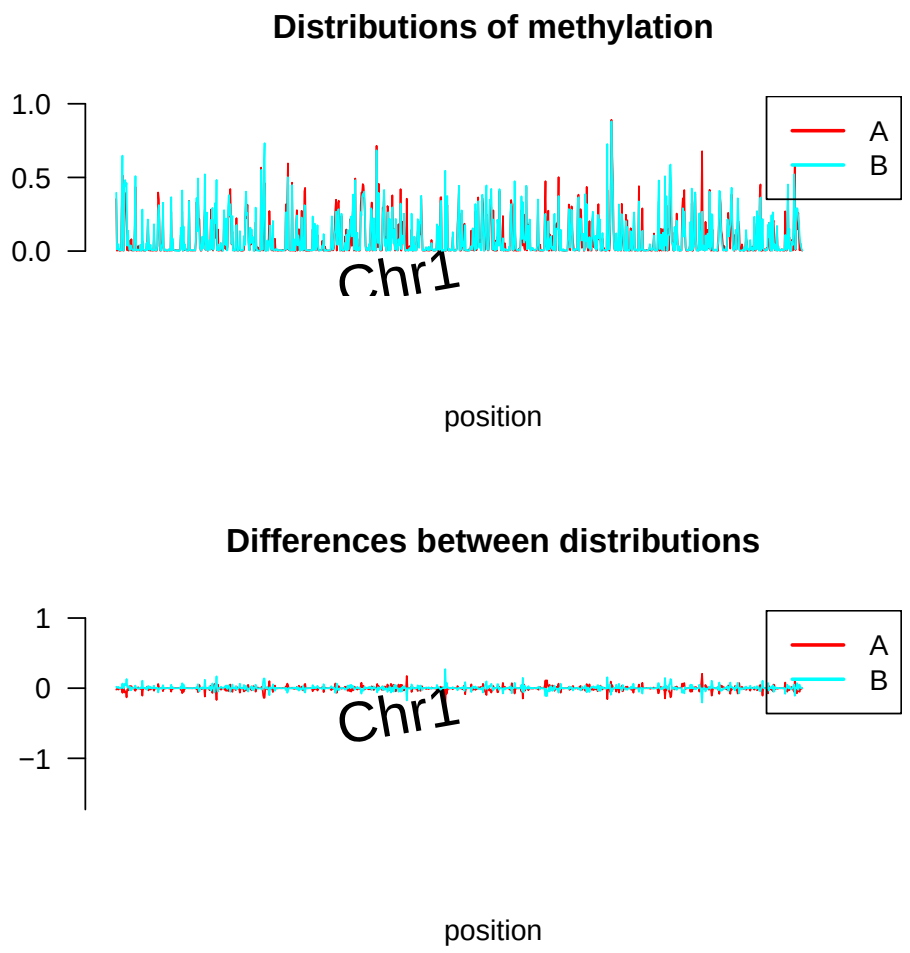

Figure 1: Distributions of methylation on the genome (first two million bases of chromosome 1.

Next, we process this `alignmentData` object to produce a `segData` object. This `segData` object contains a set of potential segments on the genome defined by the start and end points of regions of overlapping alignments in the `alignmentData` object. It then evaluates the number of tags that hit in each of these segments.

```
> sD <- processAD(mD, gap = 300, squeeze = 10, filterProp = 0.05, verbose = TRUE, strandSplit = TRUE, c
```

We can now construct a segment map from these potential segments.

Segmentation by heuristic Bayesian methods

A fast method of segmentation can be achieved by assuming a binomial distribution on the data with an uninformative beta prior, and identifying those potential segments which have a sufficiently large posterior likelihood that the proportion of methylation exceeds some critical value.

```
> hS <- heuristicSeg(sD = sD, aD = mD, prop = "auto", cl = cl, gap = 100, getLikes = FALSE)
> hS
```

```
Slot "coordinates"
GRanges object with 2739 ranges and 0 metadata columns:
         seqnames             ranges strand
            <Rle>          <IRanges>  <Rle>
     [1]     Chr1     [  108,   790]      +
     [2]     Chr1     [  648,   648]      -
     [3]     Chr1     [  809,   948]      +
     [4]     Chr1     [17541, 18715]      +
     [5]     Chr1     [17545, 18541]      -
     ...      ...                ...    ...
  [2735]     Chr1 [1987825, 1987825]      -
  [2736]     Chr1 [1988049, 1988049]      -
  [2737]     Chr1 [1990160, 1990160]      +
  [2738]     Chr1 [1990186, 1990298]      +
  [2739]     Chr1 [1993335, 1993335]      +
  -------
  seqinfo: 1 sequence from an unspecified genome; no seqlengths
An object of class "methData"
2739 rows and 4 columns

Slot "replicates"
A A B B
Slot "groups":
list()

Slot "data":
     A.1      A.2     B.1    B.2
[1,] 156:53   138:79  108:46 22:8
[2,] 24:1     12:2    13:1   2:0
[3,] 10:14    7:10    16:3   0:3
[4,] 1808:309 852:184 387:31 31:6
[5,] 1039:150 588:166 214:30 25:7
2734 more rows...

Slot "annotation":
data frame with 0 columns and 2739 rows

Slot "locLikelihoods" (stored on log scale):
Matrix with  2739  rows.
       A   B
1      1   1
2      1   1
3      0   1
4      1   1
5      1   1
...  ... ...
2735   1   1
2736   1   0
2737   0   1
2738   0   1
2739   1   0
```

Within a methylation locus, it is not uncommon to find completely unmethylated cytosines. If the coverage of these cytosines is too high, it is possible that these will cause the locus to be split into two or more fragments. The mergeMethSegs function can be used to overcome this splitting by merging loci with identical patterns of expression that are not separated by too great a gap. Merging in this manner is optional, but recommended.

```
> hS <- mergeMethSegs(hS, mD, gap = 5000, cl = cl)
```

We can then estimate posterior likelihoods on the defined loci by applying empirical Bayesian methods. These will

not change the locus definition, but will assign likelihoods that the identified loci represent a true methylation locus in each replicate group.

```
> hSL <- lociLikelihoods(hS, mD, cl = cl)
```

```
......
```

Visualising loci

By one of these methods, we finally acquire an annotated `methData` object, with the annotations describing the co-ordinates of each segment.

We can use this `methData` object, in combination with the `alignmentMeth` object, to plot the segmented genome.

```
> plotMeth(mD, hSL, chr = "Chr1", limits = c(1, 50000), cap = 10)
```

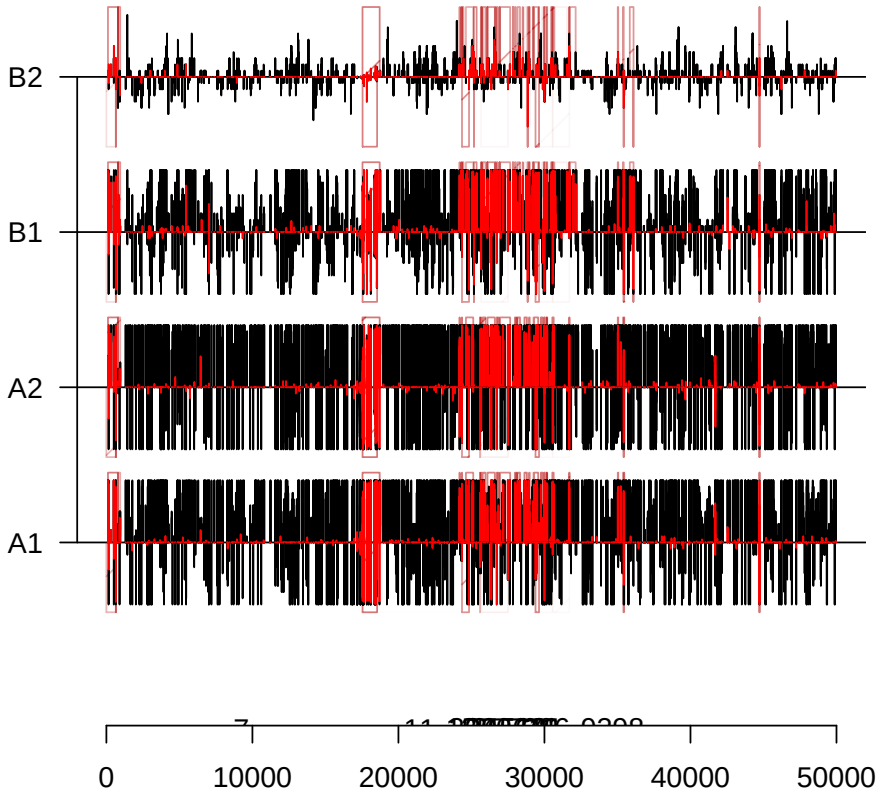

Figure 2: Methylation and identified loci on the first ten thousand bases of chromosome 1.

Differential Methylation analysis

We can also examine the `methData` object for differentially methylated regions using the beta-binomial methods [2] implemented in baySeq. We first define a group structure on the data.

```
> groups(hSL) <- list(NDE = c(1,1,1,1), DE = c("A", "A", "B", "B"))
```

The methObservables function pre-calculates a set of data to improve the speed of prior and posterior estimation (at some minor memory cost).

```
> hSL <- methObservables(hSL)
```

The density function used here is a composite of the beta-binomial and a binomial distribution that accounts for the reported non-conversion rates.

```
> densityFunction(hSL) <- bbNCDist
```

We can then determine a prior distribution on the parameters of the model for the data.

```
> hSL <- getPriors(hSL, cl = cl)
```

We can then find the posterior likelihoods of the models defined in the groups structure.

```
> hSL <- getLikelihoods(hSL, cl = cl)
```

.

We can then retrieve the data for the top differentially methylated regions.

```
> topCounts(hSL, "DE")
```

```
   seqnames   start     end width strand     A.1     A.2      B.1   B.2 Likelihood
1      Chr1  846832  847647   816      + 404:188 666:333    9:615  0:40  0.9999543
2      Chr1  958535  959092   558      +   0:275   0:695 1109:329 83:33  0.9999146
3      Chr1 1764051 1764186   136      +  129:10    72:8     0:35   0:3  0.9998595
4      Chr1  122485  122485     1      -    83:7    20:2     0:31   0:6  0.9998401
5      Chr1 1733868 1733987   120      +    0:23    0:85     48:5  10:1  0.9998344
6      Chr1 1402274 1402297    24      +    0:62    0:60     53:5   7:1  0.9998198
7      Chr1  888548  888604    57      + 630:237  277:93     1:44  0:13  0.9997983
8      Chr1   25168   25354   187      +   0:197   1:119   197:64  10:4  0.9997770
9      Chr1 1655890 1656076   187      +   0:154    0:81    28:13   7:3  0.9997532
10     Chr1   25580   25583     4      +    16:2    27:3     0:55   0:3  0.9997481
   ordering       FDR.DE      FWER.DE
1       A>B 4.565606e-05 4.565606e-05
2       B>A 6.550599e-05 1.310081e-04
3       A>B 9.048853e-05 2.714433e-04
4       A>B 1.078530e-04 4.313461e-04
5       B>A 1.194098e-04 5.969119e-04
6       B>A 1.295354e-04 7.769675e-04
7       A>B 1.398381e-04 9.784651e-04
8       B>A 1.502327e-04 1.201242e-03
9       B>A 1.609578e-04 1.447704e-03
10      A>B 1.700511e-04 1.699231e-03
```

Finally, to be a good citizen, we stop the cluster we started earlier:

```
> if(!is.null(cl))
+     stopCluster(cl)
```

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] parallel  stats4    stats     graphics  grDevices utils     datasets  methods
[9] base

other attached packages:
 [1] segmentSeq_2.8.0           ShortRead_1.32.0           GenomicAlignments_1.10.0
 [4] SummarizedExperiment_1.4.0 Biobase_2.34.0             Rsamtools_1.26.0
 [7] Biostrings_2.42.0          XVector_0.14.0             BiocParallel_1.8.0
[10] baySeq_2.8.0               abind_1.4-5                GenomicRanges_1.26.0
[13] GenomeInfoDb_1.10.0        IRanges_2.8.0              S4Vectors_0.12.0
[16] BiocGenerics_0.20.0

loaded via a namespace (and not attached):
 [1] edgeR_3.16.0        zlibbioc_1.20.0     lattice_0.20-34     hwriter_1.3.2
 [5] tools_3.3.1         grid_3.3.1          latticeExtra_0.6-28 Matrix_1.2-7.1
 [9] RColorBrewer_1.1-2  bitops_1.0-6        limma_3.30.0        locfit_1.5-9.1
[13] BiocStyle_2.2.0
```

References

[1] Thomas J. Hardcastle and Krystyna A. Kelly and David C. Baulcombe. *Identifying small RNA loci from high-throughput sequencing data.* Bioinformatics (2012).

[2] Thomas J. Hardcastle and Krystyna A. Kelly. *Empirical Bayesian analysis of paired high-throughput sequencing data with a beta-binomial distribution.* BMC Bioinformatics (2013).