segmentSeq: methods for identifying small RNA loci from high-throughput sequencing data

Thomas J. Hardcastle

May 1, 2024

1 Introduction

High-throughput sequencing technologies allow the production of large volumes of short sequences, which can be aligned to the genome to create a set of matches to the genome. By looking for regions of the genome which to which there are high densities of matches, we can infer a segmentation of the genome into regions of biological significance. The methods we propose allows the simultaneous segmentation of data from multiple samples, taking into account replicate data, in order to create a consensus segmentation. This has obvious applications in a number of classes of sequencing experiments, particularly in the discovery of small RNA loci and novel mRNA transcriptome discovery.

We approach the problem by considering a large set of potential segments upon the genome and counting the number of tags that match to that segment in multiple sequencing experiments (that may or may not contain replication). We then adapt the empirical Bayesian methods implemented in the baySeq package [1] to establish, for a given segment, the likelihood that the count data in that segment is similar to background levels, or that it is similar to the regions to the left or right of that segment. We then rank all the potential segments in order of increasing likelihood of similarity and reject those segments for which there is a high likelihood of similarity with the background or the regions to the left or right of the segment. This gives us a large list of overlapping segments. We reduce this list to identify non-overlapping loci by choosing, for a set of overlapping segments, the segment which has the lowest likelihood of similarity with either background or the regions to the left or right of that segment and rejecting all other segments that overlap with this segment. For fuller details of the method, see Hardcastle et al. [2].

2 Preparation

We begin by loading the segmentSeq package.

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

The readGeneric function is able to read in tab-delimited files which have appropriate column names, and create an alignmentData object. Alternatively, if the appropriate column names are not present, we can specify which columns to use for the data. In either case, to use this function we pass a character vector of files, together with information on which data are to be treated as replicates to the function. We also need to define the lengths of the chromosome and specify the chromosome names as a character. The data here, drawn from text files in the 'data' directory of the segmentSeq package are taken from the first million bases of an alignment to chromosome 1 and the first five hundred thousand bases of an alignment to chromosome 2 of Arabidopsis thaliana in a sequencing experiment where libraries 'SL9' and 'SL10' are replicates, as are 'SL26' and 'SL32'. Libraries 'SL9' and 'SL10' are sequenced from an Argonaute 6 IP, while 'SL26' and 'SL32' are an Argonaute 4 IP.

A similar function, readBAM performs the same operation on files in the BAM format. Please consult the help page for further details.

> datadir <- system.file("extdata", package = "segmentSeq")
> libnames <- c("SL9", "SL10", "SL26", "SL32")
> replicates <- c("AGO6", "AGO6", "AGO4", "AGO4")
> aD <- readGeneric(files = libfiles, dir = datadir, 
>             replicates = replicates, libnames = libnames, 
>             polyLength = 10, header = TRUE, gap = 200)
> aD

An object of class "alignmentData"
3149 rows and 4 columns

Slot "libnames":
[1] "SL9" "SL10" "SL26" "SL32"

Slot "replicates":
[1] AGO6 AGO6 AGO4 AGO4
Levels: AGO4 AGO6

Slot "alignments":
GRanges object with 3149 ranges and 2 metadata columns:

  seqnames  ranges     strand | tag   multireads
     <Rle> <IRanges> <Rle> | <character> <numeric>
[1] >Chr1 265-284 - | AAATGAAGATAAAACCATCA 1
[2] >Chr1 405-427 - | AAGGATGGATGAAATGACAATA.. 1
[3] >Chr1 406-428 - | AAGAATGACATTA.. 1
[4] >Chr1 600-623 + | AAGATTGCTGGTGTTGAAGA.. 1
[5] >Chr1 665-688 + | ATCCCTGTCAGCAGCATTTT.. 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.

```r
> sD <- processAD(aD, cl = cl)
> sD
```

GRanges object with 1452 ranges and 1 metadata column:

```
  seqnames ranges strand | chunk
<IRanges> <IRanges>     | <IRle>
[1] >Chr1  265-284  * |  1
[2] >Chr1  265-420  * |  1
[3] >Chr1  265-623  * |  1
[4] >Chr1  265-688  * |  1
[5] >Chr1  265-830  * |  1
[1448] >Chr1  992039-992054  * | 171
[1449] >Chr1  995357-995372  * | 172
[1450] >Chr1  995357-995507  * | 172
[1451] >Chr1  995493-995507  * | 172
[1452] >Chr1  995817-995840  * | 173
```

seqinfo: 1 sequence from an unspecified genome; no seqlengths

An object of class "lociData"
1452 rows and 4 columns
Slot "replicates"
AGO6 AGO6 AGO4 AGO4
Slot "groups":
list()

Slot "data":
SL9 SL10 SL26 SL32

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

Slot "locLikelihoods" (stored on log scale):
Matrix with 0 rows.
<0 x 0 matrix>

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

**Segmentation by heuristic methods**

A fast method of segmentation can be achieved by exploiting the bimodality of the densities of small RNAs in the potential segments. In this approach, we assign each potential segment to one of two clusters for each replicate group, either as a segment or a null based on the density of sequence tags within that segment. We then combine these clusterings for each replicate group to gain a consensus segmentation map.

```r
> hS <- heuristicSeg(sD = sD, aD = aD, RKPM = 1000, largeness = 1e8, getLikes = TRUE, cl = cl)
```

```
> cS <- classifySeg(sD = sD, aD = aD, cD = hS, cl = cl)
```

**Segmentation by empirical Bayesian methods**

A more refined approach to the problem uses an existing segment map (or, if not provided, a segment map defined by the `hS` function) to acquire empirical distributions on the density of sequence tags within a segment. We can then estimate posterior likelihoods for each potential segment as being either a true segment or a null. We then identify all potential segments in the with a posterior likelihood of being a segment greater than some value ‘lociCutoff’ and containing no subregion with a posterior likelihood of being a null greater than ‘nullCutoff’. We then greedily select the longest segments satisfying these criteria that do not overlap with any other such segments in defining our segmentation map.

```r
> cS <- classifySeg(sD = sD, aD = aD, cD = hS, cl = cl)
```

```
> cS
GRanges object with 65 ranges and 0 metadata columns:

  seqnames ranges strand
  <Rle> <IRanges> <Rle>
[1] >Chr1  1-364  *
[2] >Chr1  265-393  *
[3] >Chr1  840-967  *
```
An object of class "lociData"
65 rows and 4 columns

Slot "replicates"
AGO6 AGO6 AGO4 AGO4

Slot "groups":
list()

Slot "data"
AGO6.1 AGO6.2 AGO4.1 AGO4.2
[1,] 0 0 0 0
[2,] 31 29 51 85
[3,] 24 18 14 0
[4,] 2 3 0 0
[5,] 121 120 147 560
60 more rows...

Slot "annotation":
data frame with 0 columns and 65 rows

Slot "locLikelihoods" (stored on log scale):
Matrix with 65 rows.
AGO4  AGO6
1 0.031548 0.034761
2 0.8338 0.9883
3 0.61923 0.85384
4 0.014835 0.019089
5 0.95559 0.84837
... ... ...
61 0.66692 0.93945
62 0.0076827 0.024958
63 0.98159 0.86769
64 0.12483 0.018825
65 0.32536 0.99319

Expected number of loci in each replicate group
AGO4  AGO6
29.48546 36.18992

By one of these methods, we finally acquire an annotated lociData object, with the annotations describing the co-ordinates of each segment.
We can use this `lociData` object, in combination with the `alignmentData` object, to plot the segmented genome.

```r
> par(mfrow = c(2,1), mar = c(2,6,2,2))
> plotGenome(aD, hS, chr = ">Chr1", limits = c(1, 1e5),
+ showNumber = FALSE, cap = 50)
> plotGenome(aD, cS, chr = ">Chr1", limits = c(1, 1e5),
+ showNumber = FALSE, cap = 50)
```

**Figure 1:** The segmented genome (first $10^5$ bases of chromosome 1.

Given the calculated likelihoods, we can filter the segmented genome by controlling on likelihood, false discovery rate, or familywise error rate.

```r
> loci <- selectLoci(cS, FDR = 0.05)
> loci

GRanges object with 34 ranges and 0 metadata columns:  

<table>
<thead>
<tr>
<th>seqnames</th>
<th>ranges</th>
<th>strand</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>&lt;IRanges&gt;</td>
<td></td>
</tr>
<tr>
<td>&lt;Rle&gt;</td>
<td>&lt;IRanges&gt;</td>
<td>&lt;Rle&gt;</td>
</tr>
<tr>
<td>[1] &gt;Chr1</td>
<td>265-839</td>
<td>*</td>
</tr>
<tr>
<td>[2] &gt;Chr1</td>
<td>840-967</td>
<td>*</td>
</tr>
<tr>
<td>[3] &gt;Chr1</td>
<td>17739-17909</td>
<td>*</td>
</tr>
</tbody>
</table>
```
An object of class "lociData"
34 rows and 4 columns

Slot "replicates"
AGO6 AGO6 AGO4 AGO4

Slot "groups":
list()

Slot "data":
  AGO6.1 AGO6.2 AGO4.1 AGO4.2
[1,]  31  29  51  85
[2,]  24  18  14  90
[3,]  16  21  50  17
[4,] 239 173 162 177
[5,] 301 302 1096 349
29 more rows...

Slot "annotation":
data frame with 0 columns and 34 rows

Slot "locLikelihoods" (stored on log scale):
Matrix with 34 rows.
  AGO4   AGO6
1  0.8338  0.9883
2  0.6192  0.8538
3  0.6483  0.8819
4  0.9854  0.9973
5  0.9846  0.9984
...  ...  ...
30  0.9803  0.9698
31  0.5610  0.9879
32  0.6690  0.9394
33  0.9816  0.8676
34  0.3253  0.9932

Expected number of loci in each replicate group
  AGO4   AGO6
25.3339 32.7570

The lociData objects can now be examined for differential expression with the baySeq package.
First we define the possible models of differential expression on the data. In this case, the models are of non-differential expression and pairwise differential expression.

```r
> groups(cS) <- list(NDE = c(1,1,1,1), DE = c("AGO6", "AGO6", "AGO4", "AGO4"))
```

Then we get empirical distributions on the parameter space of the data.

```r
> cS <- getPriors(cS, cl = cl)
```

Then we get the posterior likelihoods of the data conforming to each model. Since the ‘cS’ object contains null regions as well as true loci, we will use the ‘nullData = TRUE’ option to distinguish between non-differentially expressed loci and non-expressed regions. By default, the loci likelihoods calculated earlier will be used to weight the initial parameter fit in order to detect null data.

```r
> cS <- getLikelihoods(cS, nullData = TRUE, cl = cl)
```

We can examine the highest likelihood non-expressed (‘null’) regions

```r
> topCounts(cS, NULL, number = 3)
```

<table>
<thead>
<tr>
<th>seqnames</th>
<th>start</th>
<th>end</th>
<th>width</th>
<th>strand</th>
<th>AGO6.1</th>
<th>AGO6.2</th>
<th>AGO4.1</th>
<th>AGO4.2</th>
<th>likes</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA.61</td>
<td>789549</td>
<td>944194</td>
<td>154646</td>
<td>*</td>
<td>13</td>
<td>37</td>
<td>29</td>
<td>29</td>
<td></td>
</tr>
<tr>
<td>NA.28</td>
<td>372735</td>
<td>423138</td>
<td>50404</td>
<td>*</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>NA.45</td>
<td>552693</td>
<td>587497</td>
<td>34805</td>
<td>*</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
</tbody>
</table>

FDR. FWER.

| NA.61    | 0.9781566 | 0.02184336 | 0.02184336 |
| NA.28    | 0.9754303 | 0.02329653 | 0.04587638 |
| NA.45    | 0.9753956 | 0.02367247 | 0.06935198 |

The highest likelihood expressed but non-differentially expressed regions

```r
> topCounts(cS, "NDE", number = 3)
```

<table>
<thead>
<tr>
<th>seqnames</th>
<th>start</th>
<th>end</th>
<th>width</th>
<th>strand</th>
<th>AGO6.1</th>
<th>AGO6.2</th>
<th>AGO4.1</th>
<th>AGO4.2</th>
<th>likes</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA.14</td>
<td>77151</td>
<td>77519</td>
<td>369</td>
<td>*</td>
<td>176</td>
<td>147</td>
<td>275</td>
<td>131</td>
<td>0.9307905</td>
</tr>
<tr>
<td>NA.56</td>
<td>758302</td>
<td>758848</td>
<td>547</td>
<td>*</td>
<td>167</td>
<td>166</td>
<td>269</td>
<td>134</td>
<td>0.8751151</td>
</tr>
<tr>
<td>NA.42</td>
<td>550077</td>
<td>550655</td>
<td>579</td>
<td>*</td>
<td>5</td>
<td>11</td>
<td>0</td>
<td>19</td>
<td>0.8663408</td>
</tr>
</tbody>
</table>

FDR. NDE  FWER. NDE

| NA.14    | 0.06920947 | 0.06920947 |
| NA.56    | 0.09749717  | 0.18545113  |
| NA.42    | 0.10925117  | 0.29432304  |

And the highest likelihood differentially expressed regions

```r
> topCounts(cS, "DE", number = 3)
```

<table>
<thead>
<tr>
<th>seqnames</th>
<th>start</th>
<th>end</th>
<th>width</th>
<th>strand</th>
<th>AGO6.1</th>
<th>AGO6.2</th>
<th>AGO4.1</th>
<th>AGO4.2</th>
<th>likes</th>
</tr>
</thead>
<tbody>
<tr>
<td>NA.48</td>
<td>634297</td>
<td>634350</td>
<td>54</td>
<td>*</td>
<td>65</td>
<td>90</td>
<td>12</td>
<td>17</td>
<td>0.9950597</td>
</tr>
<tr>
<td>NA.16</td>
<td>238359</td>
<td>238417</td>
<td>59</td>
<td>*</td>
<td>9</td>
<td>9</td>
<td>0</td>
<td>0</td>
<td>0.9914359</td>
</tr>
<tr>
<td>NA.13</td>
<td>76799</td>
<td>77150</td>
<td>352</td>
<td>*</td>
<td>6</td>
<td>21</td>
<td>0</td>
<td>0</td>
<td>0.9784148</td>
</tr>
</tbody>
</table>

DE  FDR.DE  FWER.DE

| NA.48    | AGO6>AGO4 | 0.004940330 | 0.00494033 |
| NA.16    | AGO6>AGO4 | 0.006752299 | 0.01346211 |
Finally, to be a good citizen, we stop the cluster we started earlier:

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

Session Info

```R
> sessionInfo()

R version 4.4.0 beta (2024-04-15 r86425)
Platform: x86_64-pc-linux-gnu
Running under: Ubuntu 22.04.4 LTS

Matrix products: default
BLAS: /home/biocbuild/bbs-3.19-bioc/R/lib/libRblas.so
LAPACK: /usr/lib/x86_64-linux-gnu/lapack/liblapack.so.3.10.0

locale:
[1] LC_CTYPE=en_US.UTF-8 LC_NUMERIC=C
[3] LC_TIME=en_US.UTF-8 LC_COLLATE=en_US.UTF-8
[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

time zone: America/New_York
tzcode source: system (glibc)

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

other attached packages:
[1] segmentSeq_2.38.0 ShortRead_1.62.0
[3] GenomicAlignments_1.40.0 SummarizedExperiment_1.34.0
[5] Biobase_2.64.0 MatrixGenerics_1.16.0
[7] matrixStats_1.3.0 Rsamtools_2.20.0
[9] Biostrings_2.72.0 XVector_0.44.0
[11] BiocParallel_1.38.0 GenomicRanges_1.56.0
[13] GenomeInfoDb_1.40.0 IRanges_2.38.0
[15] S4Vectors_0.42.0 BiocGenerics_0.50.0
[17] baySeq_2.38.0

loaded via a namespace (and not attached):
[1] SparseArray_1.4.0 bitops_1.0-7 jpeg_0.1-10
[4] lattice_0.22-6 digest_0.6.35 evaluate_0.23
[7] grid_4.4.0 RColorBrewer_1.1-3 fastmap_1.1.1
References
