An overview of the *girafe* package

J. Toedling, C. Ciaudo, O. Voinnet, E. Heard, E. Barillot

October 30, 2018

1 Introduction

The intent of package *girafe* is to facilitate the functional exploration of the alignments of multiple reads\(^1\) from next-generation sequencing (NGS) data to a genome.

It extends the functionality of the Bioconductor (Gentleman et al., 2004) packages *ShortRead* (Morgan et al., 2009) and *genomeIntervals*.

```r
> library("girafe")
> library("RColorBrewer")
```

If you use *girafe* for analysing your data, please cite:


Getting help

If possible, please send questions about *girafe* to the Bioconductor mailing list. See [http://www.bioconductor.org/docs/mailList.html](http://www.bioconductor.org/docs/mailList.html) Their archive of questions and responses may prove helpful, too.

2 Workflow

We present the functionality of the package *girafe* using example data that was downloaded from the Gene Expression Omnibus (GEO) database (Edgar et al., 2002, GSE10364). The example data are Solexa reads of 26 nt length derived from small RNA profiling of mouse oocytes. The data has previously been described in *Tam et al.* (2008).

\(^1\) The package has been developed for analysing single-end reads (fragment libraries) and does not support mate-pair reads yet.
2.1 Adapter trimming

We load reads that include parts of the adapter sequence.

```r
> ra23.wa <- readFastq(dirPath=exDir, pattern=
+ "aravinSRNA_23_plus_adapter_excerpt.fastq")
> show(ra23.wa)
```

```text
class: ShortReadQ
length: 1000 reads; width: 26 cycles
```

To removing adapter sequences, we use the function `trimAdapter`, which relies on the `pairwiseAlignment` function from the `Biostrings` package. The adapter sequence was obtained from the GEO page of the data.

```r
> adapter <- "CTGTAGGCACCATCAAT"
> ra23.na <- trimAdapter(ra23.wa, adapter)
> show(ra23.na)
```

```text
class: ShortReadQ
length: 1000 reads; width: 23 cycles
```

2.2 Importing aligned reads

The reads have been mapped to the mouse genome (assembly mm9) using the alignment tool *Bowtie* alignment tool (Langmead et al., 2009).

The resulting tab-delimited map file can be read into an object of class *AlignedRead* using the function `readAligned`. Both, this class and this function, are defined in the Bioconductor package *ShortRead*.

```r
> exA <- readAligned(dirPath=exDir, type="Bowtie",
+ pattern="aravinSRNA_23_no_adapter_excerpt_mm9_unmasked.bwtmap")
> show(exA)
```
The object of class `AlignedRead` can be converted into an object of class `AlignedGenomeIntervals`, which is the main class of the `girafe` package.

```r
> exAI <- as(exA, "AlignedGenomeIntervals")
> organism(exAI) <- "Mm"
```

For alignments in BAM format (Li et al., 2009), there is an alternative way of importing the data. The function `agiFromBam` can be used to directly create `AlignedGenomeIntervals` objects from indexed and sorted BAM files, making use of functionalities in the `Rsamtools` package.

### 2.3 Exploring the aligned reads

```r
> show(exAI)
```

1,286 genome intervals with 1,689 aligned reads on 22 chromosome(s).

Organism: Mm

Which chromosomes are the intervals located on?

```r
> table(seqnames(exAI))
```

<table>
<thead>
<tr>
<th></th>
<th>chr1</th>
<th>chr10</th>
<th>chr11</th>
<th>chr12</th>
<th>chr13</th>
<th>chr14</th>
<th>chr15</th>
<th>chr16</th>
<th>chr17</th>
<th>chr18</th>
<th>chr19</th>
<th>chr2</th>
<th>chr3</th>
</tr>
</thead>
<tbody>
<tr>
<td>count</td>
<td>112</td>
<td>50</td>
<td>60</td>
<td>53</td>
<td>65</td>
<td>74</td>
<td>59</td>
<td>48</td>
<td>47</td>
<td>43</td>
<td>19</td>
<td>87</td>
<td>62</td>
</tr>
</tbody>
</table>

A subset of the intervals on a specific chromosome can be obtained using subsetting via `['`. 

```r
> detail(exAI[seqnames(exAI) == "chrMT"])
```

<table>
<thead>
<tr>
<th>start</th>
<th>end</th>
<th>seq_name</th>
<th>strand</th>
<th>reads</th>
<th>matches</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>964</td>
<td>chrMT</td>
<td>+</td>
<td>1</td>
<td></td>
<td>GTTTATGAGAGGAGATAAGTTGT</td>
</tr>
<tr>
<td>2</td>
<td>11613</td>
<td>chrMT</td>
<td>+</td>
<td>10</td>
<td></td>
<td>AAGAAAGATTGCAAGAAGACTGCTA</td>
</tr>
<tr>
<td>3</td>
<td>11613</td>
<td>chrMT</td>
<td>+</td>
<td>1</td>
<td></td>
<td>AAGCAAGATTGCAAGAAGACTGCTA</td>
</tr>
<tr>
<td>4</td>
<td>11613</td>
<td>chrMT</td>
<td>+</td>
<td>1</td>
<td></td>
<td>AAGCAAGATTGCAAGAAGACTGCTA</td>
</tr>
<tr>
<td>5</td>
<td>11613</td>
<td>chrMT</td>
<td>+</td>
<td>1</td>
<td></td>
<td>AAGAAAGATTGCAAGAAGACTGTTA</td>
</tr>
</tbody>
</table>
Finally, what is the number of aligned reads per chromosome?

```r
> summary(exAI)
```

Number of aligned reads per chromosome:

<table>
<thead>
<tr>
<th>chromosome</th>
<th>chr1</th>
<th>chr10</th>
<th>chr11</th>
<th>chr12</th>
<th>chr13</th>
<th>chr14</th>
<th>chr15</th>
<th>chr16</th>
<th>chr17</th>
<th>chr18</th>
<th>chr19</th>
<th>chr2</th>
<th>chr3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>139</td>
<td>51</td>
<td>105</td>
<td>56</td>
<td>83</td>
<td>131</td>
<td>68</td>
<td>61</td>
<td>48</td>
<td>44</td>
<td>19</td>
<td>106</td>
<td>107</td>
</tr>
<tr>
<td>chr4</td>
<td>61</td>
<td>52</td>
<td>107</td>
<td>57</td>
<td>70</td>
<td>102</td>
<td>14</td>
<td>206</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2.4 Processing the aligned intervals

Sometimes, users may wish to combine certain aligned intervals. One intention could be to combine aligned reads at exactly the same position, which only differ in their sequence due to sequencing errors. Another objective could be to combine overlapping short reads that may be (degradation) products of the same primary transcript. The function `reduce` combines a set of aligned intervals into a single aligned interval, if the intervals:

- are on the same strand,
- are overlapping (or contained in each other) or directly adjacent to each other AND
- have the same `read match specificity` (see below)

**Read match specificity** By the `read match specificity r(I_i)` of an interval $I_i$, we refer to the total number of valid alignments of reads that have been aligned to $I_i$, i.e. the total numbers of intervals with the same reads aligned in the whole genome (or other set of reference sequences). If $r(I_i) = 1$, the reads that were aligned to the interval $I_i$ have no other valid alignments in the whole genome, i.e. the interval $I_i$ is the unique match position of these reads. If $r(I_i) = 2$, the reads that were aligned to the interval $I_i$ have exactly one other valid alignment to another interval $I_j, j \neq i$. The match specificity is stored in the `matches` slot of objects of the class `AlignedGenomeIntervals`.

We first demonstrate the `reduce` using a toy example data object.

```r
> D <- AlignedGenomeIntervals(robot)
```

```r
> D <- AlignedGenomeIntervals(  
+   start=c(1,3,4,5,8,10,11), end=c(5,5,6,8,9,11,13),
+   chromosome=rep(c("chr1","chr2","chr3"), c(2,2,3)),
+   strand=c("-","-","+","+","+","+","+") ,
+   sequence=c("ACATT","ACA","CGT","GTAA","AG","CT","TTT") ,
+   reads=c(1,7), matches=c(rep(1,6),3))
```
Calling the `reduce` method on these example data results in the following:

```r
> detail(reduce(D))
```

<table>
<thead>
<tr>
<th>start</th>
<th>end</th>
<th>seq_name</th>
<th>strand</th>
<th>reads</th>
<th>matches</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>chr1</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>ACATT</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>chr2</td>
<td>+</td>
<td>2</td>
<td>1</td>
<td>CGTAA</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>chr3</td>
<td>+</td>
<td>2</td>
<td>1</td>
<td>AGCT</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>chr3</td>
<td>+</td>
<td>1</td>
<td>3</td>
<td>TTT</td>
</tr>
</tbody>
</table>

Calling the `reduce` method on these data leads to the following result:

```r
> detail(reduce(S))
```

<table>
<thead>
<tr>
<th>start</th>
<th>end</th>
<th>seq_name</th>
<th>strand</th>
<th>reads</th>
<th>matches</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100768450</td>
<td>chrX</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>ATATAATACACCTGCTAAGCTGTTG</td>
</tr>
<tr>
<td>2</td>
<td>101311567</td>
<td>chrX</td>
<td>-</td>
<td>18</td>
<td>1</td>
<td>TGAGGTGCTGCTGCTGCTGCTGCTG</td>
</tr>
<tr>
<td>3</td>
<td>101311567</td>
<td>chrX</td>
<td>-</td>
<td>12</td>
<td>1</td>
<td>TGAGGTGCTGCTGCTGCTGCTGCTG</td>
</tr>
<tr>
<td>4</td>
<td>101311567</td>
<td>chrX</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>TGAGGTGCTGCTGCTGCTGCTGCTG</td>
</tr>
<tr>
<td>5</td>
<td>101311567</td>
<td>chrX</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>TGAGGTGCTGCTGCTGCTGCTGCTG</td>
</tr>
<tr>
<td>6</td>
<td>101311567</td>
<td>chrX</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>TGAGGTGCTGCTGCTGCTGCTGCTG</td>
</tr>
<tr>
<td>7</td>
<td>148346896</td>
<td>chrX</td>
<td>+</td>
<td>4</td>
<td>1</td>
<td>TGAGGTGCTGCTGCTGCTGCTGCTG</td>
</tr>
</tbody>
</table>

Calling the `reduce` method on these example data results in the following:

```r
> detail(reduce(D))
```

<table>
<thead>
<tr>
<th>start</th>
<th>end</th>
<th>seq_name</th>
<th>strand</th>
<th>reads</th>
<th>matches</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5</td>
<td>chr1</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>ACATT</td>
</tr>
<tr>
<td>2</td>
<td>8</td>
<td>chr2</td>
<td>+</td>
<td>2</td>
<td>1</td>
<td>CGTAA</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>chr3</td>
<td>+</td>
<td>2</td>
<td>1</td>
<td>AGCT</td>
</tr>
<tr>
<td>4</td>
<td>13</td>
<td>chr3</td>
<td>+</td>
<td>1</td>
<td>3</td>
<td>TTT</td>
</tr>
</tbody>
</table>

Note that the two last intervals still show overlap. However, the last interval is a non-unique match position of the respective reads (`matches`= 3), in contrast to the other intervals.

Here is another example using the data introduced above.

```r
> S <- exAI[seqnames(exAI) == "chrX" & matches(exAI) == 1L & exAI[,1] > 1e8]
> detail(S)
```

<table>
<thead>
<tr>
<th>start</th>
<th>end</th>
<th>seq_name</th>
<th>strand</th>
<th>reads</th>
<th>matches</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100768450</td>
<td>chrX</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>ATATAATACACCTGCTAAGCTGTTG</td>
</tr>
<tr>
<td>2</td>
<td>101311567</td>
<td>chrX</td>
<td>-</td>
<td>34</td>
<td>1</td>
<td>TGAGGTGCTGCTGCTGCTGCTGCTG</td>
</tr>
<tr>
<td>3</td>
<td>148346896</td>
<td>chrX</td>
<td>+</td>
<td>4</td>
<td>1</td>
<td>TGAGGTGCTGCTGCTGCTGCTGCTG</td>
</tr>
</tbody>
</table>

Calling the `reduce` method on these data leads to the following result:

```r
> detail(reduce(S))
```

<table>
<thead>
<tr>
<th>start</th>
<th>end</th>
<th>seq_name</th>
<th>strand</th>
<th>reads</th>
<th>matches</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>100768450</td>
<td>chrX</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>ATATAATACACCTGCTAAGCTGTTG</td>
</tr>
<tr>
<td>2</td>
<td>101311567</td>
<td>chrX</td>
<td>-</td>
<td>34</td>
<td>1</td>
<td>TGAGGTGCTGCTGCTGCTGCTGCTG</td>
</tr>
<tr>
<td>3</td>
<td>148346896</td>
<td>chrX</td>
<td>+</td>
<td>4</td>
<td>1</td>
<td>TGAGGTGCTGCTGCTGCTGCTGCTG</td>
</tr>
</tbody>
</table>
Notice that the reads that match the same segment of the X chromosome differ in their last base. However, since most of the reads have a 'G' as final letter, the combined aligned interval als has a 'G' as the last letter.

The additional argument method="exact" can be specified if you want to solely combine intervals that have exactly the same start and stop position (but may have reads of slightly different sequence aligned to them). Consider the following example:

```r
> S2 <- exAI[seqnames(exAI)="chr11" & matches(exAI)==1L & exAI[,1]>8e7]
> detail(S2)
```

<table>
<thead>
<tr>
<th>start</th>
<th>end</th>
<th>seq_name</th>
<th>strand</th>
<th>reads</th>
<th>matches</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>86397621</td>
<td>86397643</td>
<td>chr11</td>
<td>-</td>
<td>20</td>
<td>1</td>
<td>TAGCTTATCAGACTGATGTTTAC</td>
</tr>
<tr>
<td>86397621</td>
<td>86397643</td>
<td>chr11</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>TAGATTATCAGACTGATGTTTAC</td>
</tr>
<tr>
<td>86397621</td>
<td>86397643</td>
<td>chr11</td>
<td>-</td>
<td>2</td>
<td>1</td>
<td>TAGCTTATCAGACTGATGTTTAC</td>
</tr>
<tr>
<td>88515338</td>
<td>88515360</td>
<td>chr11</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>GGTGCAGGGAGCGCCAGTGTTTAC</td>
</tr>
<tr>
<td>96178500</td>
<td>96178522</td>
<td>chr11</td>
<td>+</td>
<td>2</td>
<td>1</td>
<td>TACCCCTGATACCCGAGTTTATTG</td>
</tr>
<tr>
<td>96178501</td>
<td>96178523</td>
<td>chr11</td>
<td>+</td>
<td>1</td>
<td>1</td>
<td>ACCCTGATGTCGCGAATTGTTGA</td>
</tr>
<tr>
<td>108873196</td>
<td>108873218</td>
<td>chr11</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>AGTCGCGTGACCGAGCGCTTACC</td>
</tr>
</tbody>
</table>

```r
> detail(reduce(S2, method="exact"))
```

<table>
<thead>
<tr>
<th>start</th>
<th>end</th>
<th>seq_name</th>
<th>strand</th>
<th>reads</th>
<th>matches</th>
<th>sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>86397621</td>
<td>86397643</td>
<td>chr11</td>
<td>-</td>
<td>23</td>
<td>1</td>
<td>TAGCTTATCAGACTGATGTTTAC</td>
</tr>
<tr>
<td>88515338</td>
<td>88515360</td>
<td>chr11</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>GGTGCAGGGAGCGCCAGTGTTTAC</td>
</tr>
<tr>
<td>96178500</td>
<td>96178522</td>
<td>chr11</td>
<td>+</td>
<td>2</td>
<td>1</td>
<td>TACCCCTGATACCCGAGTTTATTG</td>
</tr>
<tr>
<td>96178501</td>
<td>96178523</td>
<td>chr11</td>
<td>+</td>
<td>1</td>
<td>1</td>
<td>ACCCTGATGTCGCGAATTGTTGA</td>
</tr>
<tr>
<td>108873196</td>
<td>108873218</td>
<td>chr11</td>
<td>-</td>
<td>1</td>
<td>1</td>
<td>AGTCGCGTGACCGAGCGCTTACC</td>
</tr>
</tbody>
</table>

Notice that the 6th aligned interval in S2 is only shifted by 1 nt from the 5th one. By default, the function reduce would merge them into one aligned genome interval. However, when method="exact" is specified, these two intervals are not merged since they are not at exactly the same position. There are additional methods for restricting the merging to intervals with the same 5’- and 3’-ends (specify method="same5" and method="same3", respectively).

### 2.5 Visualising the aligned genome intervals

The package girafe contains functions for visualising genomic regions with aligned reads.

```r
> plot(exAI, mgi.gi, chr="chrX", start=50400000, + show="minus")
```

See the result in Figure 1.
Figure 1: A 10-kb region on the mouse X chromosome. Reads aligned to the Watson strand in this region would be shown above the chromosome coordinate axis with the annotation of genome elements in this region, while reads aligned to the Crick strand are shown below. In the region shown, there are only intervals with aligned reads on the Crick strand, and these four intervals overlap with annotated microRNA positions.

Note that the annotation of genome elements (as shown in Figure 1) has to be supplied to the function. Here the object \texttt{mgi.gi} contains most annotated genes and ncRNAs for the mouse genome (assembly: \texttt{mm9}). This object has been created beforehand\footnote{See the script \texttt{prepareAnnotation.R} in the package scripts directory for an example of how to create such an object.} and it is of class \texttt{Genome\_intervals\_stranded}, a class defined in package \texttt{genomeIntervals}.

### 2.6 Summarising the data using sliding windows

The data can be searched for regions of defined interest using a sliding-window approach implemented in the function \texttt{perWindow}. For each window, the number of intervals with aligned reads, the total number of reads aligned, the number of unique reads aligned, the fraction of intervals on the Plus-strand, and the higher number of aligned reads at a single interval within the window are reported.

```r
> exPX <- perWindow(exAI, chr="chrX", winsize=1e5, step=0.5e5)
> head(exPX[order(exPX$n.overlap, decreasing=TRUE),])
```

<table>
<thead>
<tr>
<th>chr</th>
<th>start</th>
<th>end</th>
<th>n.overlap</th>
<th>n.reads</th>
<th>n.unique</th>
<th>frac.plus</th>
<th>max.reads</th>
</tr>
</thead>
<tbody>
<tr>
<td>chrX</td>
<td>50341103</td>
<td>50441102</td>
<td>18</td>
<td>55</td>
<td>18</td>
<td>0</td>
<td>28</td>
</tr>
</tbody>
</table>
2.7 Exporting the data

The package girafe also contains methods for exporting the data into tab-delimited text files, which can be uploaded to the UCSC genome browser\(^3\) as 'custom tracks'.

Currently, there are methods for exporting the data in 'bed' format and 'bedGraph' format, either writing intervals from both strands into one file or into two separate files. Details about these track formats can be found at the UCSC genome browser web pages.

\[
> \text{export(exAI, con="export.bed",} \\
+ \quad \text{format="bed", name="example\_reads",} \\
+ \quad \text{description="Example reads",} \\
+ \quad \text{color="100,100,255", visibility="pack"})
\]

Additional arguments to the export function, besides object, con, and format, are treated as attributes for the track definition line, which specifies details concerning how the data should be visualised in the genome browser.

Users may also wish to consult the Bioconductor package rtracklayer for data transfer and direct interaction between R and the UCSC genome browser.

2.8 Overlap with annotated genome features

Next, we determine the degree of overlap of the aligned reads with annotated genomic elements. In this example, the annotated genome elements are provided as an object of class Genome\_intervals\_stranded\(^4\).

This objects needs to be created beforehand. See the script prepareAnnotation.R in the package scripts directory\(^5\) for an example of how to create such an object.

\(^3\)http://genome.ucsc.edu
\(^4\)Objects of class Genome\_intervals and AlignedGenome\_Intervals are also allowed.
\(^5\)system.file("scripts", package="girafe")
> exOv <- interval_overlap(exAI, mgi.gi)

How many elements do read match positions generally overlap?

> table(listLen(exOv))

<table>
<thead>
<tr>
<th></th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>12</th>
</tr>
</thead>
<tbody>
<tr>
<td>815</td>
<td>340</td>
<td>130</td>
<td>1</td>
<td></td>
</tr>
</tbody>
</table>

What are the 12 elements overlapped by a single match position?

> mgi.gi$ID[exOv[[which.max(listLen(exOv))]]]

[1] "Pcdha1"  "Pcdha2"  "Pcdha3"  "Pcdha4"  "Pcdha5"  "Pcdha6"  "Pcdha7"
[8] "Pcdha8"  "Pcdha9"  "Pcdha10" "Pcdha11" "Pcdha12"

And in general, what kinds of annotated genome elements are overlapped by reads?

> (tabOv <- table(as.character(mgi.gi$type)[unlist(exOv)]))

<table>
<thead>
<tr>
<th></th>
<th>gene</th>
<th>lincRNA</th>
<th>miRNA</th>
<th>ncRNA</th>
<th>pseudogene</th>
<th>rRNA</th>
<th>snoRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>238</td>
<td>15</td>
<td></td>
<td>297</td>
<td>19</td>
<td>28</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>tRNA</td>
<td>4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

We display these overlap classes using a pie chart.

> my.cols <- brewer.pal(length(tabOv), "Set3")
> pie(tabOv, col=my.cols, radius=0.88)

See the result in Figure 2.

Note that function `interval_overlap` only determines whether two intervals are overlapping by at least one base. For restricting the result to intervals overlapping by at least a certain number of bases or by a fraction of the interval’s length, see the function `fracOverlap`.

9
3 Memory usage

At the moment, girafe and the packages that it depends on, retain all the information concerning the read alignments in memory. This allows quick access to and swift operations on the data, but may limit the package’s usability on machines with low amounts of RAM.

The step with the highest RAM requirements is importing the alignments and saving them as objects of the AlignedRead class using the functionality in package ShortRead. Usually, objects of the AlignedGenomeIntervals class are created starting from AlignedRead objects and the AlignedRead objects can safely be discarded after this step. Since the data is summarised in that process, AlignedGenomeIntervals objects require about 10–100 times less memory than the original AlignedRead object\(^6\). We recommend that the import of the alignments and the generation of the AlignedGenomeIntervals are performed using a separate script which only needs to be called once on a machine with sufficient RAM.

\(^6\)e.g., an AlignedRead object for holding \(10^6\) reads of length 36 bp aligned to the mouse reference genome occupies about 1.4 Gb in RAM but is processed into an AlignedGenomeIntervals object of size 66.7 Mb
A suggestion for limiting memory usage is to perform the read alignments and import of the results in batches of a few million reads each. The batch-wise result `AlignedGenomeIntervals` objects can later be combined using the basic R function "c", the standard way of combining objects, optionally followed by calls of the `reduce` function.

For alignments in SAM/BAM format, the `Samtools` software suite (Li et al., 2009) as well as the Bioconductor package `Rsamtools` allow the user to access and import only selected subsets of the data, which also leads to a lower memory footprint. For details, please refer to the documentation of these packages.

Finally, while the processing of `AlignedRead` objects is the principal way of generating `AlignedGenomeIntervals` objects, there is also a convenience function called `AlignedGenomeIntervals`, which can be used to create these objects from simpler objects in the work space, such as data read in using basic R functions such as `scan`. This convenience function may be easier to use for importing and processing the data in manageable chunks.

When following these suggestions, most operations with the `girafe` package should be possible on a machine with 4 Gb of RAM, and we have not so far encountered a situation that requires more than 12 Gb (state as of the end of 2009). However, increased throughput of sequencing machines and longer reads will lead to increased memory requirements. Future developments of this and other NGS-related Bioconductor packages will therefore likely concern ways to reduce the memory footprint. One idea is to make use of packages like `ff`, which provide ways of swapping data from RAM to flat files on the hard disk, while still allowing fast and direct access to the data.

### 4 A word about speed

For improving the run time on machines with multiple processors, some of the functions in the `girafe` package have been implemented to make use of the functionality in the `parallel` package. If `parallel` has been attached and initialised before calling these functions, the functions will make use of `mclapply` instead of the normal `lapply` function. The number of cores to be used in parallel is determined by the `mc.cores` option (see the example below).

For example, if `parallel` is functional on a given system\(^7\), there should be an obvious speed improvement in the following code example.

```
> library("parallel")
> options("mc.cores"=4) # adjust to your machine
> exAI.R <- reduce(exAI, mem.friendly=TRUE)
```

### 5 Links to other Bioconductor packages

The `girafe` package is mostly built upon the interval notation and implementation provided by the packages `intervals` and `genomeIntervals`. Functions from the `ShortRead` package (Morgan et al.,

\(^7\)The `mclapply` function currently does not support Windows operating systems.
Biostrings provides help for working with the read nucleotide sequences. girafe also makes limited use of the Rle and IRanges classes defined in the IRanges package. Furthermore, the data can be converted between object classes defined in girafe and IRanges.

We note that many of the interval operations in girafe can also be performed using classes and functions defined in the IRanges package. However, the scope of the packages is slightly different. While IRanges is meant to be a generic infrastructure package of the Bioconductor project, the aim of girafe is to provide a single, comparatively lightweight, object class for working with reads aligned to the genome, the AlignedGenomeIntervals. This class and its methods allow easy access to such data and facilitate standard operations and workflows.

There is some overlap in functionality between girafe, IRanges, GenomicRanges and tracklayer. The range of interactions between these packages and new Bioconductor packages related to next-generation sequencing is likely to increase over the releases. Our aim is to provide users with a broad range of alternatives for selecting the classes and functions that are most suited for their workflows.

Package versions

This vignette was generated using the following package versions:

- R version 3.5.1 Patched (2018-07-12 r74967), x86_64-pc-linux-gnu
- Running under: Ubuntu 16.04.5 LTS
- Matrix products: default
- BLAS: /home/biocbuild/bbs-3.8-bioc/R/lib/libRblas.so
- LAPACK: /home/biocbuild/bbs-3.8-bioc/R/lib/libRlapack.so
- Base packages: base, datasets, grDevices, graphics, grid, methods, parallel, stats, stats4, utils
- Other packages: AnnotationDbi 1.44.0, Biobase 2.42.0, BiocGenerics 0.28.0, BiocParallel 1.16.0, Biostrings 2.50.0, DelayedArray 0.8.0, GenomeInfoDb 1.18.0, GenomicAlignments 1.18.0, GenomicRanges 1.34.0, IRanges 2.16.0, RColorBrewer 1.1-2, Rsamtools 1.34.0, S4Vectors 0.20.0, ShortRead 1.40.0, SummarizedExperiment 1.12.0, XVector 0.22.0, genomeIntervals 1.38.0, girafe 1.34.0, intervals 0.15.1, matrixStats 0.54.0, org.Mm.eg.db 3.7.0
- Loaded via a namespace (and not attached): DBI 1.0.0, GenomeInfoDbData 1.2.0, Matrix 1.2-14, RCurl 1.95-4.11, RSQLite 2.1.1, Rcpp 0.12.19, bit 1.1-14, bit64 0.9-7, bitops 1.0-6, blob 0.1.1, compiler 3.5.1, digest 0.6.18, hwriter 1.3.2, lattice 0.20-35, latticeExtra 0.6-28, memoise 1.1.0, pkgconfig 2.0.2, tools 3.5.1, zlibbioc 1.28.0

Acknowledgements

Many thanks to Nicolas Servant, Valérie Cognat, Nicolas Delhomme, and especially Patrick Aboyoun for suggestions and feedback on the package. Special thanks to Julien Gagneur and Richard Bourgon for writing genomeIntervals and for rapidly answering all my questions regarding the package. The plotting functions in package girafe are largely based on the function plotAlongChrom and its auxiliary functions from package tilingArray, most of which were written by Wolfgang Huber.

Funding: This work was supported by the Institut Curie, INCa "GepiG".
References


