High-level S4 containers for HTS data (solutions to exercises)

Hervé Pagès

27-28 February 2012

Exercise 1

a. *Find the SAM Spec online and investigate the meaning of predefined tags NH and NM.*

b. *Load BAM file untreated3_chr4.bam into a GappedAlignments object and subset this object to keep only the alignments satisfying the 2 following conditions:*

- *The alignment corresponds to a query with a* unique alignment *(aka* unique match *or* unique hit*).*
- *The alignment is a* perfect match *(i.e. no insertion, no deletion, no mismatch).*

c. *Do those alignments have gaps?*

Solution: The SAM Spec are available on the SAMtools website http://samtools.sourceforge.net/. According to the table of predefined tags (section 1.5 of the Spec: *The alignment section: optional fields*):

- NH: Number of reported alignments that contains the query in the current record
- NM: Edit distance to the reference, including ambiguous bases but excluding clipping

Therefore alignments that correspond to a query with a *unique hit* are those for which NH is 1. And alignments that are *perfect matches* are those for which NM is 0.

We start by loading BAM file untreated3_chr4.bam (located in the *SeattleAdvancedWorkshop2012Data* package). We need to pass a *ScanBamParam* object to `readGappedAlignments()` in order to load the NH and NM tags:

```
> library(GenomicRanges)
> library(Rsamtools)
> library(SeattleAdvancedWorkshop2012Data)
```

```
> param <- ScanBamParam(tag=c("NH", "NM"))
> gal4 <- readGappedAlignments(pathto_untreated3_chr4(),
+                              use.names=TRUE, param=param)
> gal4

GappedAlignments with 175346 alignments and 2 elementMetadata cols:
                    seqnames strand       cigar    qwidth     start       end
                       <Rle>  <Rle> <character> <integer> <integer> <integer>
  SRR031715.1138209     chr4      +         37M        37       169       205
   SRR031714.776678     chr4      -         37M        37       184       220
  SRR031715.3258011     chr4      -         37M        37       187       223
  SRR031715.4791418     chr4      +         37M        37       193       229
  SRR031715.1138209     chr4      -         37M        37       326       362
   SRR031714.756385     chr4      +         37M        37       943       979
  SRR031714.2355189     chr4      +         37M        37       944       980
  SRR031714.5054563     chr4      +         37M        37       946       982
  SRR031715.4533153     chr4      -         37M        37       946       982
                ...      ...    ...         ...       ...       ...       ...
  SRR031715.3832729     chr4      +         37M        37   1348349   1348385
  SRR031715.4873052     chr4      -         37M        37   1348350   1348386
  SRR031714.1650928     chr4      +         37M        37   1349196   1349232
  SRR031714.1650928     chr4      -         37M        37   1349326   1349362
  SRR031714.1650928     chr4      +         37M        37   1349708   1349744
  SRR031714.1650928     chr4      -         37M        37   1349838   1349874
  SRR031714.5192891     chr4      +         37M        37   1351640   1351676
  SRR031715.2351056     chr4      +         37M        37   1351640   1351676
   SRR031714.864195     chr4      +         37M        37   1351760   1351796
                       width      ngap   |        NH        NM
                   <integer> <integer>   | <integer> <integer>
  SRR031715.1138209        37         0   |         1         0
   SRR031714.776678        37         0   |         1         2
  SRR031715.3258011        37         0   |         1         1
  SRR031715.4791418        37         0   |         1         1
  SRR031715.1138209        37         0   |         1         0
   SRR031714.756385        37         0   |         1         0
  SRR031714.2355189        37         0   |         1         0
  SRR031714.5054563        37         0   |         1         0
  SRR031715.4533153        37         0   |         8         1
                ...       ...       ... ...       ...       ...
  SRR031715.3832729        37         0   |         2         2
  SRR031715.4873052        37         0   |         2         2
  SRR031714.1650928        37         0   |         5         0
  SRR031714.1650928        37         0   |         5         0
  SRR031714.1650928        37         0   |         5         0
  SRR031714.1650928        37         0   |         5         0
  SRR031714.5192891        37         0   |         4         2
```

```
  SRR031715.2351056        37         0   |         4         2
   SRR031714.864195        37         0   |         3         2
  ---
  seqlengths:
      chr2L    chr2R    chr3L    chr3R     chr4     chrM     chrX  chrYHet
   23011544 21146708 24543557 27905053  1351857    19517 22422827   347038
```

Then we subset `gal4`:

```
> has_unique_hit <- elementMetadata(gal4)$NH == 1L
> is_perfect_match <- elementMetadata(gal4)$NM == 0L
> keep <- has_unique_hit & is_perfect_match
> table(keep)

keep
 FALSE   TRUE
 74773 100573

> gal4[keep]

GappedAlignments with 100573 alignments and 2 elementMetadata cols:
                    seqnames strand       cigar    qwidth     start       end
                       <Rle>  <Rle> <character> <integer> <integer> <integer>
  SRR031715.1138209     chr4      +         37M        37       169       205
  SRR031715.1138209     chr4      -         37M        37       326       362
   SRR031714.756385     chr4      +         37M        37       943       979
  SRR031714.2355189     chr4      +         37M        37       944       980
  SRR031714.5054563     chr4      +         37M        37       946       982
  SRR031714.5054563     chr4      -         37M        37       986      1022
  SRR031715.1722593     chr4      -         37M        37      1108      1144
  SRR031715.2202469     chr4      -         37M        37      1114      1150
  SRR031714.2355189     chr4      -         37M        37      1119      1155
                ...      ...    ...         ...       ...       ...       ...
  SRR031715.1467928     chr4      -         37M        37   1322595   1322631
  SRR031715.4365847     chr4      +         37M        37   1322760   1322796
  SRR031715.4365847     chr4      -         37M        37   1322908   1322944
  SRR031715.3400894     chr4      +         37M        37   1322994   1323030
  SRR031715.3400894     chr4      -         37M        37   1323141   1323177
  SRR031714.5105755     chr4      -         37M        37   1323236   1323272
  SRR031715.3618930     chr4      -         37M        37   1323418   1323454
  SRR031714.2538775     chr4      +         37M        37   1335326   1335362
  SRR031714.5201273     chr4      +         37M        37   1335329   1335365
                        width      ngap   |        NH        NM
                    <integer> <integer>   | <integer> <integer>
  SRR031715.1138209        37         0   |         1         0
  SRR031715.1138209        37         0   |         1         0
   SRR031714.756385        37         0   |         1         0
```

```
  SRR031714.2355189        37         0   |         1         0
  SRR031714.5054563        37         0   |         1         0
  SRR031714.5054563        37         0   |         1         0
  SRR031715.1722593        37         0   |         1         0
  SRR031715.2202469        37         0   |         1         0
  SRR031714.2355189        37         0   |         1         0
               ...       ...       ... ...       ...       ...
  SRR031715.1467928        37         0   |         1         0
  SRR031715.4365847        37         0   |         1         0
  SRR031715.4365847        37         0   |         1         0
  SRR031715.3400894        37         0   |         1         0
  SRR031715.3400894        37         0   |         1         0
  SRR031714.5105755        37         0   |         1         0
  SRR031715.3618930        37         0   |         1         0
  SRR031714.2538775        37         0   |         1         0
  SRR031714.5201273        37         0   |         1         0
  ---
  seqlengths:
      chr2L    chr2R    chr3L    chr3R     chr4     chrM     chrX  chrYHet
   23011544 21146708 24543557 27905053  1351857    19517 22422827   347038
```

And yes:

```
> table(grepl("N", cigar(gal4)[keep], fixed=TRUE))

FALSE  TRUE
98350  2223
```

... some of those alignments have gaps!

Exercise 2
Use the TxDb.Dmelanogaster.UCSC.dm3.ensGene package and the result of Exercise 1 to count the number of unique hits *per transcript, that is, the number of hits from reads with a* unique alignment*.*

Solution: First we subset `gal4` to keep only the alignments corresponding to a query with a *unique hit*:

```
> gal4uh <- gal4[has_unique_hit]
```

Then we turn `gal4uh` into a *GRangesList* object:

```
> grl4uh <- as(gal4uh, "GRangesList")
> grl4uh

GRangesList of length 130399:
$SRR031715.1138209
GRanges with 1 range and 0 elementMetadata cols:
```

```
      seqnames     ranges strand
         <Rle>  <IRanges>  <Rle>
  [1]     chr4 [169, 205]      +

$SRR031714.776678
GRanges with 1 range and 0 elementMetadata cols:
      seqnames     ranges strand
  [1]     chr4 [184, 220]      -

$SRR031715.3258011
GRanges with 1 range and 0 elementMetadata cols:
      seqnames     ranges strand
  [1]     chr4 [187, 223]      -

...
<130396 more elements>
---
seqlengths:
    chr2L    chr2R    chr3L    chr3R     chr4     chrM     chrX  chrYHet
 23011544 21146708 24543557 27905053  1351857    19517 22422827   347038
```

Then we load the transcript annotations corresponding to the reference genome that was used to align the reads:

```
> library(TxDb.Dmelanogaster.UCSC.dm3.ensGene)
> TxDb.Dmelanogaster.UCSC.dm3.ensGene

TranscriptDb object:
| Db type: TranscriptDb
| Data source: UCSC
| Genome: dm3
| Genus and Species: Drosophila melanogaster
| UCSC Table: ensGene
| Resource URL: http://genome.ucsc.edu/
| Type of Gene ID: Ensembl gene ID
| Full dataset: yes
| transcript_nrow: 23017
| exon_nrow: 69155
| cds_nrow: 59573
| Db created by: GenomicFeatures package from Bioconductor
| Creation time: 2011-09-29 18:24:41 -0700 (Thu, 29 Sep 2011)
| GenomicFeatures version at creation time: 1.5.27
| RSQLite version at creation time: 0.9-4
| DBSCHEMAVERSION: 1.0
| package: GenomicFeatures
```

Then we extract the exons grouped by transcript from this *TranscriptDb* object:

```
> exbytx <- exonsBy(TxDb.Dmelanogaster.UCSC.dm3.ensGene, by="tx", use.names=TRUE)
> exbytx

GRangesList of length 23017:
$FBtr0089116
GRanges with 11 ranges and 3 elementMetadata cols:
       seqnames           ranges strand |   exon_id   exon_name exon_rank
          <Rle>        <IRanges>  <Rle> | <integer> <character> <integer>
   [1]     chr4 [251356, 251521]      + |         1        <NA>         1
   [2]     chr4 [252561, 252603]      + |         2        <NA>         2
   [3]     chr4 [252905, 253474]      + |         3        <NA>         3
   [4]     chr4 [254891, 254971]      + |         4        <NA>         4
   [5]     chr4 [255490, 255570]      + |         5        <NA>         5
   [6]     chr4 [257021, 257101]      + |         6        <NA>         6
   [7]     chr4 [257895, 258185]      + |         7        <NA>         7
   [8]     chr4 [260940, 261024]      + |         8        <NA>         8
   [9]     chr4 [263892, 264211]      + |         9        <NA>         9
  [10]     chr4 [264260, 264374]      + |        10        <NA>        10
  [11]     chr4 [265806, 266500]      + |        11        <NA>        11

...
<23016 more elements>
---
seqlengths:
     chr2L  chr2LHet     chr2R  chr2RHet ...   chrXHet   chrYHet      chrM
  23011544    368872  21146708   3288761 ...    204112    347038     19517
```

Finally we use `countOverlaps()` to count the number of hits per transcript:

```
> txhits <- countOverlaps(exbytx, grl4uh)
> length(txhits)

[1] 23017

> head(txhits)

FBtr0089116 FBtr0300800 FBtr0300796 FBtr0300799 FBtr0300798 FBtr0300797
        365         406         410         370         410         407

> sum(txhits)  # total nb of hits

[1] 194659

> head(sort(txhits, decreasing=TRUE))

FBtr0089175 FBtr0089176 FBtr0089177 FBtr0112904 FBtr0289951 FBtr0089243
      14376       14051       13811        5433        5411        5410
```

Note that this counting is still a very rough one because:

- The fact that the reads are actually *paired-end* is ignored.
- A hit is counted even if it's not *compatible* with the splicing of the transcript.

Some tools are currently under active development in the *GenomicRanges* and *Rsamtools* packages to address this. They should be available in BioC 2.10 (to be released on April 2nd, 2012).