

# GenomicRanges HOWTOs

**Bioconductor Team**

**Edited: June 2024; Compiled: October 9, 2024**

## Contents

1	Introduction . . . . .	2
1.1	Purpose of this document. . . . .	2
1.2	Prerequisites and additional recommended reading . . . . .	2
1.3	Input data and terminology used across the HOWTOs . . . . .	2
2	HOWTOs . . . . .	3
2.1	How to read single-end reads from a BAM file . . . . .	3
2.2	How to read paired-end reads from a BAM file . . . . .	4
2.3	How to read and process a big BAM file by chunks in order to reduce memory usage . . . . .	6
2.4	How to compute read coverage . . . . .	7
2.5	How to find peaks in read coverage . . . . .	8
2.6	How to retrieve a gene model from the UCSC genome browser . . . . .	9
2.7	How to retrieve a gene model from Ensembl . . . . .	10
2.8	How to load a gene model from a GFF or GTF file . . . . .	12
2.9	How to retrieve a gene model from <i>AnnotationHub</i> . . . . .	13
2.10	How to annotate peaks in read coverage . . . . .	15
2.11	How to prepare a table of read counts for RNA-Seq differential gene expression . . . . .	15
2.12	How to summarize junctions from a BAM file containing RNA-Seq reads . . . . .	17
2.13	How to get the exon and intron sequences of a given gene. . . . .	18
2.14	How to get the CDS and UTR sequences of genes associated with colorectal cancer . . . . .	20
2.14.1	Build a gene list . . . . .	20
2.14.2	Identify genomic coordinates . . . . .	21
2.14.3	Extract sequences from BSgenome . . . . .	22
2.15	How to create DNA consensus sequences for read group ‘families’. . . . .	24
2.15.1	Sort reads into groups by start position . . . . .	24
2.15.2	Remove low frequency reads . . . . .	27
2.15.3	Create a consensus sequence for each read group family . . . . .	28
2.16	How to compute binned averages along a genome. . . . .	29

3	Session Information	30
---	---------------------	----

# 1 Introduction

---

## 1.1 Purpose of this document

This document is a collection of *HOWTO*s. Each *HOWTO* is a short section that demonstrates how to use the containers and operations implemented in the [GenomicRanges](#) and related packages ([IRanges](#), [Biostrings](#), [Rsamtools](#), [GenomicAlignments](#), [BSgenome](#), and [GenomicFeatures](#)) to perform a task typically found in the context of a high throughput sequence analysis.

Unless stated otherwise, the *HOWTO*s are self contained, independent of each other, and can be studied and reproduced in any order.

## 1.2 Prerequisites and additional recommended reading

We assume the reader has some previous experience with *R* and with basic manipulation of `GRanges`, `GRangesList`, `Rle`, `RleList`, and `DataFrame` objects. See the “An Introduction to Genomic Ranges Classes” vignette located in the [GenomicRanges](#) package (in the same folder as this document) for an introduction to these containers.

Additional recommended readings after this document are the “Software for Computing and Annotating Genomic Ranges” paper [Lawrence et al. (2013)] and the “Counting reads with `summarizeOverlaps`” vignette located in the [GenomicAlignments](#) package.

To display the list of vignettes available in the [GenomicRanges](#) package, use `browseVignettes("GenomicRanges")`.

## 1.3 Input data and terminology used across the HOWTOs

In order to avoid repetition, input data, concepts and terms used in more than one *HOWTO* are described here:

- **The [pasillaBamSubset](#) data package:** contains both a BAM file with single-end reads (`untreated1_chr4`) and a BAM file with paired-end reads (`untreated3_chr4`). Each file is a subset of chr4 from the "Pasilla" experiment.

```
> library(pasillaBamSubset)
> untreated1_chr4()

[1] "/Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/library/pasillaBamSubset/extdata/untreated1_chr4.bam"

> untreated3_chr4()

[1] "/Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/library/pasillaBamSubset/extdata/untreated3_chr4.bam"
```

See `?pasillaBamSubset` for more information.

```
> ?pasillaBamSubset
```

- **Gene models and *TxDb* objects:** A *gene model* is essentially a set of annotations that describes the genomic locations of the known genes, transcripts, exons, and CDS, for a given organism. In *Bioconductor* it is typically represented as a *TxDb* object but also sometimes as a *GRanges* or *GRangesList* object. The [txdbmaker](#) and [GenomicFeatures](#) packages contain tools for making and manipulating *TxDb* objects.

## 2 HOWTOs

### 2.1 How to read single-end reads from a BAM file

As sample data we use the [pasillaBamSubset](#) data package described in the introduction.

```
> library(pasillaBamSubset)
> un1 <- untreated1_chr4() # single-end reads
```

Several functions are available for reading BAM files into *R*:

```
readGAlignments()
readGAlignmentPairs()
readGAlignmentsList()
scanBam()
```

`scanBam` is a low-level function that returns a list of lists and is not discussed further here. See `?scanBam` in the [Rsamtools](#) package for more information.

Single-end reads can be loaded with the `readGAlignments` function from the [GenomicAlignments](#) package.

```
> library(GenomicAlignments)
> gal <- readGAlignments(un1)
```

Data subsets can be specified by genomic position, field names, or flag criteria in the `ScanBamParam`. Here we input records that overlap position 1 to 5000 on the negative strand with `flag` and `cigar` as metadata columns.

```
> what <- c("flag", "cigar")
> which <- GRanges("chr4", IRanges(1, 5000))
> flag <- scanBamFlag(isMinusStrand = TRUE)
> param <- ScanBamParam(which=which, what=what, flag=flag)
> neg <- readGAlignments(un1, param=param)
> neg
```

`GAlignments` object with 37 alignments and 2 metadata columns:

	seqnames	strand	cigar	qwidth	start	end
	<Rle>	<Rle>	<character>	<integer>	<integer>	<integer>
[1]	chr4	-	75M	75	892	966
[2]	chr4	-	75M	75	919	993
[3]	chr4	-	75M	75	967	1041
...	...	...	...	...	...	...
[35]	chr4	-	75M	75	4997	5071
[36]	chr4	-	75M	75	4998	5072
[37]	chr4	-	75M	75	4999	5073
	width	njunc	flag	cigar		

```

      <integer> <integer> | <integer> <character>
[1]          75          0 |          16          75M
[2]          75          0 |          16          75M
[3]          75          0 |          16          75M
...          ...          ... |          ...          ...
[35]         75          0 |          16          75M
[36]         75          0 |          16          75M
[37]         75          0 |          16          75M
-----
seqinfo: 8 sequences from an unspecified genome

```

Another approach to subsetting the data is to use `filterBam`. This function creates a new BAM file of records passing user-defined criteria. See `?filterBam` in the *Rsamtools* package for more information.

## 2.2 How to read paired-end reads from a BAM file

As sample data we use the *pasillaBamSubset* data package described in the introduction.

```

> library(pasillaBamSubset)
> un3 <- untreated3_chr4() # paired-end reads

```

Paired-end reads can be loaded with the `readGAlignmentPairs` or `readGAlignmentsList` function from the *GenomicAlignments* package. These functions use the same mate paring algorithm but output different objects.

Let's start with `readGAlignmentPairs`:

```

> un3 <- untreated3_chr4()
> gapairs <- readGAlignmentPairs(un3)

```

The `GAlignmentPairs` class holds only pairs; reads with no mate or with ambiguous pairing are discarded. Each list element holds exactly 2 records (a mated pair). Records can be accessed as the `first` and `last` segments in a template or as `left` and `right` alignments. See `?GAlignmentPairs` in the *GenomicAlignments* package for more information.

```

> gapairs
GAlignmentPairs object with 75409 pairs, strandMode=1, and 0 metadata columns:
      seqnames strand :      ranges --      ranges
      <Rle>  <Rle>  :    <IRanges> --    <IRanges>
[1]    chr4      +   :    169-205 --    326-362
[2]    chr4      +   :    943-979 --   1086-1122
[3]    chr4      +   :    944-980 --   1119-1155
...      ...      ... ..
[75407] chr4      +   : 1348217-1348253 -- 1348215-1348251
[75408] chr4      +   : 1349196-1349232 -- 1349326-1349362
[75409] chr4      +   : 1349708-1349744 -- 1349838-1349874
-----
seqinfo: 8 sequences from an unspecified genome

```

For `readGAlignmentsList`, mate pairing is performed when `asMates` is set to `TRUE` on the `BamFile` object, otherwise records are treated as single-end.

```
> galist <- readGAlignmentsList(BamFile(un3, asMates=TRUE))
```

`GAlignmentsList` is a more general 'list-like' structure that holds mate pairs as well as non-mates (i.e., singletons, records with unmapped mates etc.) A `mates_status` metadata column (accessed with `mcols`) indicates which records were paired.

```
> galist
```

```
GAlignmentsList object of length 96636:
```

```
[[1]]
```

```
GAlignments object with 2 alignments and 0 metadata columns:
```

	seqnames	strand	cigar	qwidth	start	end
	<Rle>	<Rle>	<character>	<integer>	<integer>	<integer>
[1]	chr4	+	37M	37	169	205
[2]	chr4	-	37M	37	326	362
	width	njunc				
	<integer>	<integer>				
[1]	37	0				
[2]	37	0				

```
-----
```

```
seqinfo: 8 sequences from an unspecified genome
```

```
[[2]]
```

```
GAlignments object with 2 alignments and 0 metadata columns:
```

	seqnames	strand	cigar	qwidth	start	end
	<Rle>	<Rle>	<character>	<integer>	<integer>	<integer>
[1]	chr4	+	37M	37	946	982
[2]	chr4	-	37M	37	986	1022
	width	njunc				
	<integer>	<integer>				
[1]	37	0				
[2]	37	0				

```
-----
```

```
seqinfo: 8 sequences from an unspecified genome
```

```
[[3]]
```

```
GAlignments object with 2 alignments and 0 metadata columns:
```

	seqnames	strand	cigar	qwidth	start	end
	<Rle>	<Rle>	<character>	<integer>	<integer>	<integer>
[1]	chr4	+	37M	37	943	979
[2]	chr4	-	37M	37	1086	1122
	width	njunc				
	<integer>	<integer>				
[1]	37	0				
[2]	37	0				

```
-----
```

```
seqinfo: 8 sequences from an unspecified genome
```

```
...
```

```
<96633 more elements>
```

Non-mated reads are returned as groups by QNAME and contain any number of records. Here the non-mate groups range in size from 1 to 9.

```
> non_mates <- galist[unlist(mcols(galist)$mate_status) == "unmated"]
> table(elementNROWS(non_mates))
```

1	2	3	4	5	6	7	8	9
18191	2888	69	60	7	8	2	1	1

## 2.3 How to read and process a big BAM file by chunks in order to reduce memory usage

A large BAM file can be iterated through in chunks by setting a `yieldSize` on the *BamFile* object. As sample data we use the [pasillaBamSubset](#) data package described in the introduction.

```
> library(pasillaBamSubset)
> un1 <- untreated1_chr4()
> bf <- BamFile(un1, yieldSize=100000)
```

Iteration through a BAM file requires that the file be opened, repeatedly queried inside a loop, then closed. Repeated calls to `readGAlignments` without opening the file first result in the same 100000 records returned each time.

```
> open(bf)
> cvg <- NULL
> repeat {
+   chunk <- readGAlignments(bf)
+   if (length(chunk) == 0L)
+     break
+   chunk_cvg <- coverage(chunk)
+   if (is.null(cvg)) {
+     cvg <- chunk_cvg
+   } else {
+     cvg <- cvg + chunk_cvg
+   }
+ }
> close(bf)
> cvg

RleList of length 8
$chr2L
integer-Rle of length 23011544 with 1 run
  Lengths: 23011544
  Values :      0

$chr2R
integer-Rle of length 21146708 with 1 run
  Lengths: 21146708
  Values :      0

$chr3L
```

```
integer-Rle of length 24543557 with 1 run
  Lengths: 24543557
  Values :      0

$chr3R
integer-Rle of length 27905053 with 1 run
  Lengths: 27905053
  Values :      0

$chr4
integer-Rle of length 1351857 with 122061 runs
  Lengths:  891   27    5   12   13   45 ... 106   75 1600   75 1659
  Values :    0    1    2    3    4    5 ...  0    1    0    1    0

...
<3 more elements>
```

## 2.4 How to compute read coverage

The “read coverage” is the number of reads that cover a given genomic position. Computing the read coverage generally consists in computing the coverage at each position in the genome. This can be done with the `coverage()` function.

As sample data we use the [pasillaBamSubset](#) data package described in the introduction.

```
> library(pasillaBamSubset)
> un1 <- untreated1_chr4() # single-end reads
> library(GenomicAlignments)
> reads1 <- readGAlignments(un1)
> cvg1 <- coverage(reads1)
> cvg1

RleList of length 8
$chr2L
integer-Rle of length 23011544 with 1 run
  Lengths: 23011544
  Values :      0

$chr2R
integer-Rle of length 21146708 with 1 run
  Lengths: 21146708
  Values :      0

$chr3L
integer-Rle of length 24543557 with 1 run
  Lengths: 24543557
  Values :      0

$chr3R
integer-Rle of length 27905053 with 1 run
  Lengths: 27905053
```

```

Values :      0

$chr4
integer-Rle of length 1351857 with 122061 runs
  Lengths: 891  27   5  12  13  45 ... 106  75 1600  75 1659
  Values :   0   1   2   3   4   5 ...   0   1   0   1   0

...
<3 more elements>

```

Coverage on chr4:

```

> cvg1$chr4

integer-Rle of length 1351857 with 122061 runs
  Lengths: 891  27   5  12  13  45 ... 106  75 1600  75 1659
  Values :   0   1   2   3   4   5 ...   0   1   0   1   0

```

Average and max coverage:

```

> mean(cvg1$chr4)
[1] 11.33746

> max(cvg1$chr4)
[1] 5627

```

Note that `coverage()` is a generic function with methods for different types of objects. See `?coverage` for more information.

## 2.5 How to find peaks in read coverage

ChIP-Seq analysis usually involves finding peaks in read coverage. This process is sometimes called “peak calling” or “peak detection”. Here we’re only showing a naive way to find peaks in the object returned by the `coverage()` function. *Bioconductor* packages [BayesPeak](#), [bumphunter](#), [Starr](#), [CexoR](#), [exomePeak](#), [RIPSeeker](#), and others, provide sophisticated peak calling tools for ChIP-Seq, RIP-Seq, and other kind of high throughput sequencing data.

Let’s assume `cvg1` is the object returned by `coverage()` (see previous *HOWTO* for how to compute it). We can use the `slice()` function to find the genomic regions where the coverage is greater or equal to a given threshold.

```

> chr4_peaks <- slice(cvg1$chr4, lower=500)
> chr4_peaks

Views on a 1351857-length Rle subject

views:
      start      end width
[1]  86849   87364   516 [ 525  538  554  580  583  585  589 ...]
[2]  87466   87810   345 [4924 4928 4941 4943 4972 5026 5039 ...]
[3] 340791  340798     8 [508 512 506 530 521 519 518 501]
[4] 340800  340885    86 [500 505 560 560 565 558 564 559 555 ...]
[5] 348477  348483     7 [503 507 501 524 515 513 512]

```



```

[6] 348488 348571 84 [554 554 559 552 558 553 549 550 559 ...]
[7] 692512 692530 19 [502 507 508 518 520 522 524 526 547 ...]
[8] 692551 692657 107 [ 530 549 555 635 645 723 725 ...]
[9] 692798 692800 3 [503 500 503]
...
[34] 1054306 1054306 1 [502]
[35] 1054349 1054349 1 [501]
[36] 1054355 1054444 90 [510 521 525 532 532 539 549 555 557 ...]
[37] 1054448 1054476 29 [502 507 516 517 508 517 525 528 532 ...]
[38] 1054479 1054482 4 [504 503 506 507]
[39] 1054509 1054509 1 [500]
[40] 1054511 1054511 1 [502]
[41] 1054521 1054623 103 [529 521 529 530 524 525 547 540 536 ...]
[42] 1054653 1054717 65 [520 519 516 528 526 585 591 589 584 ...]

> length(chr4_peaks) # nb of peaks
[1] 42

```

The weight of a given peak can be defined as the number of aligned nucleotides that belong to the peak (a.k.a. the area under the peak in mathematics). It can be obtained with `sum()`:

```

> sum(chr4_peaks)

[1] 1726347 1300700 4115 52301 3575 51233 10382 95103
[9] 1506 500 2051 500 5834 10382 92163 500
[17] 88678 1512 500 11518 14514 5915 3598 7821
[25] 511 508 503 500 1547 8961 43426 22842
[33] 503 502 501 51881 15116 2020 500 502
[41] 67010 40496

```

## 2.6 How to retrieve a gene model from the UCSC genome browser

See introduction for a quick description of what *gene models* and *TxDb* objects are. We can use the `makeTranscriptDbFromUCSC()` function from the *txdbmaker* package to import a UCSC genome browser track as a *TxDb* object.

```

> library(txdbmaker)
> ### Internet connection required! Can take several minutes...
> txdb <- makeTxDbFromUCSC(genome="sacCer2", tablename="ensGene")

```

See `?makeTxDbFromUCSC` in the *txdbmaker* package for more information.

Note that some of the most frequently used gene models are available as *TxDb* packages. A *TxDb* package consists of a pre-made *TxDb* object wrapped into an annotation data package. Go to [http://bioconductor.org/packages/release/BiocViews.html#\\_\\_\\_TxDb](http://bioconductor.org/packages/release/BiocViews.html#___TxDb) to browse the list of available *TxDb* packages.

```

> library(TxDb.Hsapiens.UCSC.hg38.knownGene)
> txdb <- TxDb.Hsapiens.UCSC.hg38.knownGene
> txdb

```

```

TxDb object:
# Db type: TxDb

```

```
# Supporting package: GenomicFeatures
# Data source: UCSC
# Genome: hg38
# Organism: Homo sapiens
# Taxonomy ID: 9606
# UCSC Table: knownGene
# UCSC Track: GENCODE V44
# Resource URL: http://genome.ucsc.edu/
# Type of Gene ID: Entrez Gene ID
# Full dataset: yes
# miRBase build ID: NA
# Nb of transcripts: 276905
# Db created by: GenomicFeatures package from Bioconductor
# Creation time: 2023-09-20 17:25:17 +0000 (Wed, 20 Sep 2023)
# GenomicFeatures version at creation time: 1.53.2
# RSQLite version at creation time: 2.3.1
# DBSCHEMAVERSION: 1.2
```

Extract the transcript coordinates from this gene model:

```
> transcripts(txdb)

GRanges object with 276905 ranges and 2 metadata columns:
      seqnames      ranges strand |      tx_id
      <Rle>       <IRanges> <Rle> | <integer>
[1]          chr1    11869-14409   + |         1
[2]          chr1    12010-13670   + |         2
[3]          chr1    29554-31097   + |         3
...
[276903] chrX_MU273397v1_alt 314193-316302   - |    276903
[276904] chrX_MU273397v1_alt 314813-315236   - |    276904
[276905] chrX_MU273397v1_alt 324527-324923   - |    276905
      tx_name
      <character>
[1] ENST00000456328.2
[2] ENST00000450305.2
[3] ENST00000473358.1
...
[276903] ENST00000710030.1
[276904] ENST00000710216.1
[276905] ENST00000710031.1
-----
seqinfo: 711 sequences (1 circular) from hg38 genome
```

## 2.7 How to retrieve a gene model from Ensembl

See introduction for a quick description of what *gene models* and *TxDb* objects are. We can use the `makeTranscriptDbFromBiomart()` function from the *txdbmaker* package to retrieve a gene model from the Ensembl Mart.

```
> library(txdbmaker)
> ### Internet connection required! Can take several minutes...
> txdb <- makeTxDbFromBiomart(biomart="ensembl",
+                             dataset="hsapiens_gene_ensembl")
```

See `?makeTxDbFromBiomart` in the `txdbmaker` package for more information.

Note that some of the most frequently used gene models are available as TxDb packages. A TxDb package consists of a pre-made `TxDb` object wrapped into an annotation data package. Go to [http://bioconductor.org/packages/release/BiocViews.html#\\_\\_\\_TxDb](http://bioconductor.org/packages/release/BiocViews.html#___TxDb) to browse the list of available TxDb packages.

```
> library(TxDb.Athaliana.BioMart.plantmart22)
> txdb <- TxDb.Athaliana.BioMart.plantmart22
> txdb

TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: BioMart
# Organism: Arabidopsis thaliana
# Resource URL: www.biomart.org:80
# BioMart database: plants_mart_22
# BioMart database version: ENSEMBL PLANTS 22 (EBI UK)
# BioMart dataset: athaliana_eg_gene
# BioMart dataset description: Arabidopsis thaliana genes (TAIR10 (2010-09-TAIR10))
# BioMart dataset version: TAIR10 (2010-09-TAIR10)
# Full dataset: yes
# miRBase build ID: NA
# transcript_nrow: 41671
# exon_nrow: 171013
# cds_nrow: 147494
# Db created by: GenomicFeatures package from Bioconductor
# Creation time: 2014-09-26 11:23:54 -0700 (Fri, 26 Sep 2014)
# GenomicFeatures version at creation time: 1.17.17
# RSQLite version at creation time: 0.11.4
# DBSCHEMAVERSION: 1.0
# TaxID: 3702
```

Extract the exon coordinates from this gene model:

```
> exons(txdb)

GRanges object with 171013 ranges and 1 metadata column:
      seqnames      ranges strand |   exon_id
      <Rle>      <IRanges> <Rle> | <integer>
[1]          1      3631-3913    + |         1
[2]          1      3996-4276    + |         2
[3]          1      4486-4605    + |         3
...          ...          ...    ... |         ...
[171011]      Pt 137869-137940    - |      171011
[171012]      Pt 144921-145154    - |      171012
[171013]      Pt 145291-152175    - |      171013
```

```
-----
seqinfo: 7 sequences (1 circular) from an unspecified genome
```

## 2.8 How to load a gene model from a GFF or GTF file

See introduction for a quick description of what *gene models* and *TxDb* objects are. We can use the `makeTranscriptDbFromGFF()` function from the *txdbmaker* package to import a GFF or GTF file as a *TxDb* object.

```
> library(txdbmaker)
> gff_file <- system.file("extdata", "GFF3_files", "a.gff3",
+                          package="txdbmaker")
> txdb <- makeTxDbFromGFF(gff_file, format="gff3")
> txdb

TxDb object:
# Db type: TxDb
# Supporting package: GenomicFeatures
# Data source: /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/library/txdbmaker/extdata/GFF3_
# Organism: NA
# Taxonomy ID: NA
# miRBase build ID: NA
# Genome: NA
# Nb of transcripts: 488
# Db created by: txdbmaker package from Bioconductor
# Creation time: 2024-10-09 19:35:44 -0400 (Wed, 09 Oct 2024)
# txdbmaker version at creation time: 1.0.1
# RSQLite version at creation time: 2.3.7
# DBSCHEMAVERSION: 1.2
```

See `?makeTxDbFromGFF` in the *txdbmaker* package for more information.

Extract the exon coordinates grouped by gene from this gene model:

```
> exonsBy(txdb, by="gene")

GRangesList object of length 488:
$Solyc00g005000.2
GRanges object with 2 ranges and 2 metadata columns:
      seqnames      ranges strand |   exon_id   exon_name
      <Rle>      <IRanges> <Rle> | <integer>   <character>
[1] SL2.40ch00 16437-17275    + |         1 Solyc00g005000.2.1.1
[2] SL2.40ch00 17336-18189    + |         2 Solyc00g005000.2.1.2
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths

$Solyc00g005020.1
GRanges object with 3 ranges and 2 metadata columns:
      seqnames      ranges strand |   exon_id   exon_name
      <Rle>      <IRanges> <Rle> | <integer>   <character>
[1] SL2.40ch00 68062-68211    + |         3 Solyc00g005020.1.1.1
[2] SL2.40ch00 68344-68568    + |         4 Solyc00g005020.1.1.2
[3] SL2.40ch00 68654-68764    + |         5 Solyc00g005020.1.1.3
```

```

-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths

$Solyc00g005040.2
GRanges object with 4 ranges and 2 metadata columns:
      seqnames      ranges strand |   exon_id   exon_name
      <Rle>      <IRanges> <Rle> | <integer> <character>
[1] SL2.40ch00 550920-550945   + |      6 Solyc00g005040.2.1.1
[2] SL2.40ch00 551034-551132   + |      7 Solyc00g005040.2.1.2
[3] SL2.40ch00 551218-551250   + |      8 Solyc00g005040.2.1.3
[4] SL2.40ch00 551343-551576   + |      9 Solyc00g005040.2.1.4
-----
seqinfo: 1 sequence from an unspecified genome; no seqlengths

...
<485 more elements>

```

## 2.9 How to retrieve a gene model from *AnnotationHub*

When a gene model is not available as a *GRanges* or *GRangesList* object or as a *Bioconductor* data package, it may be available on [AnnotationHub](#). In this *HOWTO*, will look for a gene model for *Drosophila melanogaster* on [AnnotationHub](#). Create a 'hub' and then filter on *Drosophila melanogaster*:

```

> library(AnnotationHub)
> ### Internet connection required!
> hub <- AnnotationHub()
> hub <- subset(hub, hub$species=='Drosophila melanogaster')

```

There are 87 files that match *Drosophila melanogaster*. If you look at the metadata in hub, you can see that the 7th record represents a *GRanges* object from UCSC

```

> length(hub)
[1] 467

> head(names(hub))
[1] "AH6789" "AH6790" "AH6791" "AH6792" "AH6793" "AH6794"

> head(hub$title, n=10)
[1] "Assembly"      "GDP Insertions" "BAC End Pairs" "FlyBase Genes"
[5] "RefSeq Genes"  "Ensembl Genes"  "CONTRAST"      "Human Proteins"
[9] "Spliced ESTs"  "Other mRNAs"

> ## then look at a specific slice of the hub object.
> hub[7]

AnnotationHub with 1 record
# snapshotDate(): 2024-04-30
# names(): AH6795
# $dataprovder: UCSC
# $species: Drosophila melanogaster

```

```
# $rdataclass: GRanges
# $rdatadateadded: 2013-04-04
# $title: CONTRAST
# $description: GRanges object from on UCSC track â€~CONTRASTâ€™™
# $taxonomyid: 7227
# $genome: dm3
# $sourcetype: UCSC track
# $sourceurl: rtracklayer://hgdownload.cse.ucsc.edu/goldenpath/dm3/d...
# $sourcesize: NA
# $tags: c("contrastGene", "UCSC", "track", "Gene",
#   "Transcript", "Annotation")
# retrieve record with 'object[["AH6795"]]'
```

So you can retrieve that dm3 file as a `GRanges` like this:

```
> gr <- hub[[names(hub)[7]]]
> summary(gr)

[1] "GRanges object with 13504 ranges and 5 metadata columns"
```

The metadata fields contain the details of file origin and content.

```
> metadata(gr)

$AnnotationHubName
[1] "AH6795"

$`File Name`
[1] "contrastGene"

$`Data Source`
[1] "rtracklayer://hgdownload.cse.ucsc.edu/goldenpath/dm3/database/contrastGene"

$Provider
[1] "UCSC"

$Organism
[1] "Drosophila melanogaster"

$`Taxonomy ID`
[1] 7227
```

Split the `GRanges` object by gene name to get a `GRangesList` object of transcript ranges grouped by gene.

```
> txbygn <- split(gr, gr$name)
```

You can now use `txbygn` with the `summarizeOverlaps` function to prepare a table of read counts for RNA-Seq differential gene expression.

Note that before passing `txbygn` to `summarizeOverlaps`, you should confirm that the `seqlevels` (chromosome names) in it match those in the BAM file. See `?renameSeqlevels`, `?keepSeqlevels` and `?seqlevels` for examples of renaming `seqlevels`.

## 2.10 How to annotate peaks in read coverage

[coming soon...]

## 2.11 How to prepare a table of read counts for RNA-Seq differential gene expression

Methods for RNA-Seq gene expression analysis generally require a table of counts that summarize the number of reads that overlap or 'hit' a particular gene. In this *HOWTO* we count with the `summarizeOverlaps` function from the [GenomicAlignments](#) package and create a count table from the results.

Other packages that provide read counting are [Rsubread](#) and [easyRNASeq](#). The [parathyroidSE](#) package vignette contains a workflow on counting and other common operations required for differential expression analysis.

As sample data we use the [pasillaBamSubset](#) data package described in the introduction.

```
> library(pasillaBamSubset)
> reads <- c(untrt1=untreated1_chr4(), # single-end reads
+           untrt3=untreated3_chr4()) # paired-end reads
```

`summarizeOverlaps` requires the name of a BAM file(s) and a gene model to count against. See introduction for a quick description of what a *gene models* is. The gene model must match the genome build the reads in the BAM file were aligned to. For the pasilla data this is dm3 Dmelanogaster which is available as a *Bioconductor* package. Load the package and extract the exon ranges grouped by gene:

```
> library(TxDb.Dmelanogaster.UCSC.dm3.ensGene)
> exbygene <- exonsBy(TxDb.Dmelanogaster.UCSC.dm3.ensGene, "gene")
```

`exbygene` is a *GRangesList* object with one list element per gene in the gene model.

`summarizeOverlaps` automatically sets a `yieldSize` on large BAM files and iterates over them in chunks. When reading paired-end data set the `singleEnd` argument to `FALSE`. See [?summarizeOverlaps](#) for details regarding the count modes and additional arguments.

```
> library(GenomicAlignments)
> se <- summarizeOverlaps(exbygene, reads, mode="IntersectionNotEmpty")
```

The return object is a `SummarizedExperiment` with counts accessible with the `assays` accessor:

```
> class(se)
[1] "RangedSummarizedExperiment"
attr(,"package")
[1] "SummarizedExperiment"
> head(table(assays(se)$counts))
      0      1      2      3      4      5
31188  2      6      3      4      4
```

The count vector is the same length as `exbygene`:

```
> identical(length(exbygene), length(assays(se)$counts))
[1] FALSE
```

A copy of `exbygene` is stored in the `se` object and accessible with `rowRanges` accessor:

```
> rowRanges(se)

GRangesList object of length 15682:
$FBgn00000003
GRanges object with 1 range and 2 metadata columns:
      seqnames      ranges strand |   exon_id   exon_name
      <Rle>        <IRanges> <Rle> | <integer> <character>
[1]   chr3R 2648220-2648518      + |    45123      <NA>
-----
seqinfo: 15 sequences (1 circular) from dm3 genome

$FBgn00000008
GRanges object with 13 ranges and 2 metadata columns:
      seqnames      ranges strand |   exon_id   exon_name
      <Rle>        <IRanges> <Rle> | <integer> <character>
[1]   chr2R 18024494-18024531      + |    20314      <NA>
[2]   chr2R 18024496-18024713      + |    20315      <NA>
[3]   chr2R 18024938-18025756      + |    20316      <NA>
...      ...      ...      ... |    ...      ...
[11]  chr2R 18059821-18059938      + |    20328      <NA>
[12]  chr2R 18060002-18060339      + |    20329      <NA>
[13]  chr2R 18060002-18060346      + |    20330      <NA>
-----
seqinfo: 15 sequences (1 circular) from dm3 genome

...
<15680 more elements>
```

Two popular packages for RNA-Seq differential gene expression are [DESeq2](#) and [edgeR](#). Tables of counts per gene are required for both and can be easily created with a vector of counts. Here we use the counts from our *SummarizedExperiment* object:

```
> colData(se)$trt <- factor(c("untrt", "untrt"), levels=c("trt", "untrt"))
> colData(se)

DataFrame with 2 rows and 1 column
      trt
      <factor>
untrt1  untrt
untrt3  untrt

> library(DESeq2)
> dseq <- DESeqDataSet(se, design= ~ 1)
> library(edgeR)
> edger <- DGEList(assays(se)$counts, group=rownames(colData(se)))
```



## 2.12 How to summarize junctions from a BAM file containing RNA-Seq reads

As sample data we use the *pasillaBamSubset* data package described in the introduction.

```
> library(pasillaBamSubset)
> un1 <- untreated1_chr4() # single-end reads
> library(GenomicAlignments)
> reads1 <- readGAlignments(un1)
> reads1
```

GAlignments object with 204355 alignments and 0 metadata columns:

	seqnames	strand	cigar	qwidth	start	end
	<Rle>	<Rle>	<character>	<integer>	<integer>	<integer>
[1]	chr4	-	75M	75	892	966
[2]	chr4	-	75M	75	919	993
[3]	chr4	+	75M	75	924	998
...	...	...	...	...	...	...
[204353]	chr4	+	75M	75	1348268	1348342
[204354]	chr4	-	75M	75	1348449	1348523
[204355]	chr4	-	75M	75	1350124	1350198

  

	width	njunc
	<integer>	<integer>
[1]	75	0
[2]	75	0
[3]	75	0
...	...	...
[204353]	75	0
[204354]	75	0
[204355]	75	0

-----

seqinfo: 8 sequences from an unspecified genome

For each alignment, the aligner generated a CIGAR string that describes its "geometry", that is, the locations of insertions, deletions and junctions in the alignment. See the SAM Spec available on the SAMtools website for the details (<http://samtools.sourceforge.net/>).

The `summarizeJunctions()` function from the *GenomicAlignments* package can be used to summarize the junctions in `reads1`.

```
> junc_summary <- summarizeJunctions(reads1)
> junc_summary
```

GRanges object with 910 ranges and 3 metadata columns:

	seqnames	ranges	strand	score	plus_score
	<Rle>	<IRanges>	<Rle>	<integer>	<integer>
[1]	chr4	5246-11972	*	3	1
[2]	chr4	10346-10637	*	1	1
[3]	chr4	27102-27166	*	13	11
...	...	...	...	...	...
[908]	chr4	1333752-1346734	*	1	0
[909]	chr4	1334150-1347141	*	1	1
[910]	chr4	1334557-1347539	*	1	0

```

      minus_score
      <integer>
 [1]           2
 [2]           0
 [3]           2
 ...          ...
[908]          1
[909]          0
[910]          1
-----
seqinfo: 8 sequences from an unspecified genome

```

See `?summarizeJunctions` in the [GenomicAlignments](#) package for more information.

## 2.13 How to get the exon and intron sequences of a given gene

The exon and intron sequences of a gene are essentially the DNA sequences of the introns and exons of all known transcripts of the gene. The first task is to identify all transcripts associated with the gene of interest. Our sample gene is the human TRAK2 which is involved in regulation of endosome-to-lysosome trafficking of membrane cargo. The Entrez gene id is '66008'.

```
> trak2 <- "66008"
```

The [TxDb.Hsapiens.UCSC.hg38.knownGene](#) data package contains the gene model corresponding to the UCSC 'Known Genes' track.

```
> library(TxDb.Hsapiens.UCSC.hg38.knownGene)
> txdb <- TxDb.Hsapiens.UCSC.hg38.knownGene
```

The transcript ranges for all the genes in the gene model can be extracted with the [transcriptsBy](#) function from the [GenomicFeatures](#) package. They will be returned in a named *GRangesList* object containing all the transcripts grouped by gene. In order to keep only the transcripts of the TRAK2 gene we will subset the *GRangesList* object using the `[[` operator.

```
> library(GenomicFeatures)
> trak2_txs <- transcriptsBy(txdb, by="gene")[[trak2]]
> trak2_txs

GRanges object with 5 ranges and 2 metadata columns:
      seqnames      ranges strand |      tx_id      tx_name
      <Rle>        <IRanges> <Rle> | <integer> <character>
 [1]   chr2 201377207-201451458   - |   39614 ENST00000332624.8
 [2]   chr2 201395128-201451500   - |   39615 ENST00000430254.1
 [3]   chr2 201397606-201399510   - |   39616 ENST00000486291.1
 [4]   chr2 201419197-201451453   - |   39618 ENST00000451703.5
 [5]   chr2 201420417-201433554   - |   39619 ENST00000440597.1
-----
seqinfo: 711 sequences (1 circular) from hg38 genome

```

`trak2_txs` is a *GRanges* object with one range per transcript in the TRAK2 gene. The transcript names are stored in the `tx_name` metadata column. We will need them to subset the extracted intron and exon regions:

```
> trak2_tx_names <- mcols(trak2_txs)$tx_name
> trak2_tx_names

[1] "ENST00000332624.8" "ENST00000430254.1" "ENST00000486291.1"
[4] "ENST00000451703.5" "ENST00000440597.1"
```

The exon and intron genomic ranges for all the transcripts in the gene model can be extracted with the `exonsBy` and `intronsByTranscript` functions, respectively. Both functions return a *GRangesList* object. Then we keep only the exon and intron for the transcripts of the TRAK2 gene by subsetting each *GRangesList* object by the TRAK2 transcript names.

Extract the exon regions:

```
> trak2_exbytx <- exonsBy(txdb, "tx", use.names=TRUE)[trak2_tx_names]
> elementNROWS(trak2_exbytx)

ENST00000332624.8 ENST00000430254.1 ENST00000486291.1 ENST00000451703.5
                16                8                2                3
ENST00000440597.1
                2
```

... and the intron regions:

```
> trak2_inbytx <- intronsByTranscript(txdb, use.names=TRUE)[trak2_tx_names]
> elementNROWS(trak2_inbytx)

ENST00000332624.8 ENST00000430254.1 ENST00000486291.1 ENST00000451703.5
                15                7                1                2
ENST00000440597.1
                1
```

Next we want the DNA sequences for these exons and introns. The `getSeq` function from the *Biostrings* package can be used to query a *BSgenome* object with a set of genomic ranges and retrieve the corresponding DNA sequences.

```
> library(BSgenome.Hsapiens.UCSC.hg38)
```

Extract the exon sequences:

```
> trak2_ex_seqs <- getSeq(Hsapiens, trak2_exbytx)
> trak2_ex_seqs

DNAStringSetList of length 5
[["ENST00000332624.8"]] GGCAGTGGGAGCAGCGGCAGCAGCTTCGGCTGCTGCTTTCAGGCT...
[["ENST00000430254.1"]] GTGAAGTCGCCCGCCTGTCCCTGCCACGCCGGCGGTGCTGGC...
[["ENST00000486291.1"]] ATTTGGTATTTTAACAGAGGGATCGTGATCTGGAACCTCGCTGCTC...
[["ENST00000451703.5"]] TGGGAGCAGCGGCAGCAGCTTCGGCTGCTGCTTTCAGGCTGCCGC...
[["ENST00000440597.1"]] GGAAATACTGACCCATAAACCTAAATCTCACACACGGTTTTTGA...

> trak2_ex_seqs[["uc002uyb.4"]]

NULL
```

```
> trak2_ex_seqs[["uc002uyc.2"]]
NULL
```

... and the intron sequences:

```
> trak2_in_seqs <- getSeq(Hsapiens, trak2_inbytx)
> trak2_in_seqs

DNAStringSetList of length 5
[["ENST00000332624.8"]] GTAAGAGTGCCTGGGAAATCTGGGGCCTCACTTCTTTCCTCAGCT...
[["ENST00000430254.1"]] GTGAGTATTAACATATTCTCTTTGTACCTTTTGGACAATTCTT...
[["ENST00000486291.1"]] GTAAGCCTTTGATCAAATGTCTGCAGTATGAAATAATTAGGTTTT...
[["ENST00000451703.5"]] GTAAGTCCAGTTTAATAAATATTGAAGTGCTATCGCTTATAGGAG...
[["ENST00000440597.1"]] GTAAGTAGAGCGACCTCCATGTATTTCATATTCTGACACCCTACT...

> trak2_in_seqs[["uc002uyb.4"]]
NULL

> trak2_in_seqs[["uc002uyc.2"]]
NULL
```

## 2.14 How to get the CDS and UTR sequences of genes associated with colorectal cancer

In this *HOWTO* we extract the CDS and UTR sequences of genes involved in colorectal cancer. The workflow extends the ideas presented in the previous *HOWTO* and suggests an approach for identifying disease-related genes.

### 2.14.1 Build a gene list

We start with a list of gene or transcript ids. If you do not have pre-defined list one can be created with the [KEGGREST](#) and [KEGGgraph](#) packages.

Create a table of KEGG pathways and ids and search on the term 'Colorectal cancer'.

```
> library(KEGGREST)
> li <- keggList("pathway", "hsa")
> ptag <- names(grep("Colorectal cancer", li, value=TRUE))
> ptag

[1] "hsa05210"

> tag <- gsub("path:hsa", "", ptag)
```

Use the "05210" id to query the KEGG web resource (accesses the currently maintained data).

```
> library(KEGGgraph)
> dest <- tempfile()
> retrieveKGML(tag, "hsa", dest)
```

The suffix of the KEGG id is the Entrez gene id. The [translateKEGGID2GeneID](#) simply removes the prefix leaving just the Entrez gene ids.

```
> crids <- as.character(parseKML2DataFrame(dest)[,1])
> crgenes <- unique(translateKEGGID2GeneID(crids))
> head(crgenes)

[1] "1630"      "836"      "110117499" "5290"      "5291"
[6] "5293"
```

### 2.14.2 Identify genomic coordinates

The list of gene ids is used to extract genomic positions of the regions of interest. The Known Gene table from UCSC will be the annotation and is available as a *Bioconductor* package.

```
> library(TxDb.Hsapiens.UCSC.hg38.knownGene)
> txdb <- TxDb.Hsapiens.UCSC.hg38.knownGene
```

If an annotation is not available as a *Bioconductor* annotation package it may be available in [AnnotationHub](#). Additionally, there are functions in [txdbmaker](#) which can retrieve data from UCSC and Ensembl to create a *TxDb*. See `?makeTxDbFromUCSC` for more information.

As in the previous *HOWTO* we need to identify the transcripts corresponding to each gene. The transcript id (or name) is used to isolate the UTR and coding regions of interest. This grouping of transcript by gene is also used to re-group the final sequence results.

The `transcriptsBy` function outputs both the gene and transcript identifiers which we use to create a map between the two. The `map` is a *CharacterList* with gene ids as names and transcript ids as the list elements.

```
> txbygene <- transcriptsBy(txdb, "gene")[crgenes] ## subset on colorectal genes
> map <- relist(unlist(txbygene, use.names=FALSE)$tx_id, txbygene)
> map

IntegerList of length 59
[["1630"]] 211493 211495 211496 211497 ... 211506 211507 211509 211510
[["836"]] 67353 67354 67355 67356 67357 ... 67360 67361 67362 67363
[["110117499"]] 14821
[["5290"]] 48048 48049 48050 48051 48052 48053 48054 48055 48056 48057
[["5291"]] 54973 54974 54975 54976 54977 ... 54983 54984 54985 54986
[["5293"]] 735 736 737 738 739 740 741 ... 751 752 753 754 755 756 757
[["5295"]] 69556 69557 69558 69559 69560 ... 69585 69586 69587 69588
[["5296"]] 217205 217206 217207 217208 ... 217213 217214 217215 217216
[["8503"]] 14817 14818 14819 14820 14822 14823 14824
[["842"]] 12721 12722 12723 12724 12725 ... 12728 12729 12730 12731
...
<49 more elements>
```

Extract the UTR and coding regions.

```
> cds <- cdsBy(txdb, "tx")
> threeUTR <- threeUTRsByTranscript(txdb)
> fiveUTR <- fiveUTRsByTranscript(txdb)
```

Coding and UTR regions may not be present for all transcripts specified in `map`. Consequently, the subset results will not be the same length. This length discrepancy must be taken into account when re-listing the final results by gene.

```
> txid <- unlist(map, use.names=FALSE)
> cds <- cds[names(cds) %in% txid]
> threeUTR <- threeUTR[names(threeUTR) %in% txid]
> fiveUTR <- fiveUTR[names(fiveUTR) %in% txid]
```

Note the different lengths of the subset regions.

```
> length(txid) ## all possible transcripts
[1] 921
> length(cds)
[1] 622
> length(threeUTR)
[1] 513
> length(fiveUTR)
[1] 497
```

These objects are `GRangesLists` with the transcript id as the outer list element.

```
> cds
GRangesList object of length 622:
$`736`
GRanges object with 2 ranges and 3 metadata columns:
      seqnames      ranges strand |   cds_id   cds_name exon_rank
      <Rle>        <IRanges> <Rle> | <integer> <character> <integer>
[1]   chr1 9710456-9710596      + |     798      <NA>         3
[2]   chr1 9715541-9715554      + |     799      <NA>         4
-----
seqinfo: 711 sequences (1 circular) from hg38 genome

$`737`
GRanges object with 4 ranges and 3 metadata columns:
      seqnames      ranges strand |   cds_id   cds_name exon_rank
      <Rle>        <IRanges> <Rle> | <integer> <character> <integer>
[1]   chr1 9710456-9710596      + |     798      <NA>         3
[2]   chr1 9715541-9715769      + |     800      <NA>         4
[3]   chr1 9715849-9716241      + |     802      <NA>         5
[4]   chr1 9716440-9716582      + |     803      <NA>         6
-----
seqinfo: 711 sequences (1 circular) from hg38 genome

...
<620 more elements>
```

### 2.14.3 Extract sequences from BSgenome

The `BSgenome` packages contain complete genome sequences for a given organism.

Load the `BSgenome` package for homo sapiens.

```
> library(BSgenome.Hsapiens.UCSC.hg38)
> genome <- BSgenome.Hsapiens.UCSC.hg38
```

Use `extractTranscriptSeqs` to extract the UTR and coding regions from the `BSgenome`. This function retrieves the sequences for an any `GRanges` or `GRangesList` (i.e., not just transcripts like the name implies).

```
> threeUTR_seqs <- extractTranscriptSeqs(genome, threeUTR)
> fiveUTR_seqs <- extractTranscriptSeqs(genome, fiveUTR)
> cds_seqs <- extractTranscriptSeqs(genome, cds)
```

The return values are `DNASTringSet` objects.

```
> cds_seqs

DNASTringSet object of length 622:
      width seq
[1] 155 ATGCCCCCTGGGGTGGAC...GCAGCTGCTGTGGCACCG 736
[2] 906 ATGCCCCCTGGGGTGGAC...GGTGAACGGCAGGCATGA 737
[3] 3135 ATGCCCCCTGGGGTGGAC...AAAGACAACAGGCAGTAG 738
...
[620] 570 ATGACGGAATATAAGCTG...AAGTGTGTGCTCTCCTGA 263873
[621] 44 ATGACGGAATATAAGCTG...GGGCGCCGGCGGTGTGGG 263876
[622] 1767 ATGCCCCAGCTCGGCGGC...ACCAAGTCTGCCCACTAA 275344
```

Our final step is to collect the coding and UTR regions (currently organized by transcript) into groups by gene id. The `relist` function groups the sequences of a `DNASTringSet` object into a `DNASTringSetList` object, based on the specified `skeleton` argument. The `skeleton` must be a list-like object and only its shape (i.e. its element lengths) matters (its exact content is ignored). A simple form of `skeleton` is to use a partitioning object that we make by specifying the size of each partition. The partitioning objects are different for each type of region because not all transcripts had a coding or 3' or 5' UTR region defined.

```
> lst3 <- relist(threeUTR_seqs, PartitioningByWidth(sum(map %in% names(threeUTR))))
> lst5 <- relist(fiveUTR_seqs, PartitioningByWidth(sum(map %in% names(fiveUTR))))
> lstc <- relist(cds_seqs, PartitioningByWidth(sum(map %in% names(cds))))
```

There are 239 genes in `map` each of which have 1 or more transcripts. The table of element lengths shows how many genes have each number of transcripts. For example, 47 genes have 1 transcript, 48 genes have 2 etc.

```
> length(map)

[1] 59

> table(elementNROWS(map))

 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 20 21 22 23 24 26 30
2  1  3  2  3  3  2  2  4  1  3  4  4  3  2  3  2  2  2  1  1  1  1  1
31 33 41 55 63 74
 1  1  1  1  1  1
```

The lists of DNA sequences all have the same length as `map` but one or more of the element lengths may be zero. This would indicate that data were not available for that gene. The tables below show that there was at least 1 coding region available for all genes (i.e., none of the element lengths are 0). However, both the 3' and 5' UTR results have element lengths of 0 which indicates no UTR data were available for that gene.

```
> table(elementNROWS(lstc))

 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 17 19 21 24 26 29 41 45 47
 3  7  3  5  2  2  5  2  5  4  4  3  2  2  1  1  1  1  1  1  1  1  1  1

> table(elementNROWS(lst3))

 1  2  3  4  5  6  7  8  9 10 11 12 13 16 17 21 24 30 41
 4 10  2  7  1  4  5  5  1  3  3  2  5  1  1  1  1  1  2

> names(lst3)[elementNROWS(lst3) == 0L] ## genes with no 3' UTR data
character(0)

> table(elementNROWS(lst5))

 1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 21 22 24 41 42
 4  9  3  4  4  6  5  5  1  2  2  3  3  2  1  1  1  1  1  1

> names(lst5)[elementNROWS(lst5) == 0L] ## genes with no 5' UTR data
character(0)
```

## 2.15 How to create DNA consensus sequences for read group 'families'

The motivation for this *HOWTO* comes from a study which explored the dynamics of point mutations. The mutations of interest exist with a range of frequencies in the control group (e.g., 0.1% - 50%). PCR and sequencing error rates make it difficult to identify low frequency events (e.g., < 20%).

When a library is prepared with Nextera, random fragments are generated followed by a few rounds of PCR. When the genome is large enough, reads aligning to the same start position are likely descendant from the same template fragment and should have identical sequences.

The goal is to eliminate noise by grouping the reads by common start position and discarding those that do not exceed a certain threshold within each family. A new consensus sequence will be created for each read group family.

### 2.15.1 Sort reads into groups by start position

Load the BAM file into a `GAlignments` object.

```
> library(Rsamtools)
> bamfile <- system.file("extdata", "ex1.bam", package="Rsamtools")
> param <- ScanBamParam(what=c("seq", "qual"))
> library(GenomicAlignments)
> gal <- readGAlignments(bamfile, use.names=TRUE, param=param)
```

Use the `sequenceLayer` function to lay the query sequences and quality strings on the reference.



```

> qseq <- setNames(mcols(gal)$seq, names(gal))
> qual <- setNames(mcols(gal)$qual, names(gal))
> qseq_on_ref <- sequenceLayer(qseq, cigar(gal),
+                             from="query", to="reference")
> qual_on_ref <- sequenceLayer(qual, cigar(gal),
+                             from="query", to="reference")

```

Split by chromosome.

```

> qseq_on_ref_by_chrom <- splitAsList(qseq_on_ref, seqnames(gal))
> qual_on_ref_by_chrom <- splitAsList(qual_on_ref, seqnames(gal))
> pos_by_chrom <- splitAsList(start(gal), seqnames(gal))

```

For each chromosome generate one GRanges object that contains unique alignment start positions and attach 3 metadata columns to it: the number of reads, the query sequences, and the quality strings.

```

> gr_by_chrom <- lapply(seqlevels(gal),
+   function(seqname)
+   {
+     qseq_on_ref2 <- qseq_on_ref_by_chrom[[seqname]]
+     qual_on_ref2 <- qual_on_ref_by_chrom[[seqname]]
+     pos2 <- pos_by_chrom[[seqname]]
+     qseq_on_ref_per_pos <- split(qseq_on_ref2, pos2)
+     qual_on_ref_per_pos <- split(qual_on_ref2, pos2)
+     nread <- elementNROWS(qseq_on_ref_per_pos)
+     gr_mcols <- DataFrame(nread=unname(nread),
+                           qseq_on_ref=unname(qseq_on_ref_per_pos),
+                           qual_on_ref=unname(qual_on_ref_per_pos))
+     gr <- GRanges(Rle(seqname, nrow(gr_mcols)),
+                   IRanges(as.integer(names(nread)), width=1))
+     mcols(gr) <- gr_mcols
+     seqlevels(gr) <- seqlevels(gal)
+     gr
+   })

```

Concatenate all the GRanges objects obtained in (4) together in 1 big GRanges object:

```

> gr <- do.call(c, gr_by_chrom)
> seqinfo(gr) <- seqinfo(gal)

```

'gr' is a GRanges object that contains unique alignment start positions:

```
> gr[1:6]
```

GRanges object with 6 ranges and 3 metadata columns:

	seqnames	ranges	strand	nread
	<Rle>	<IRanges>	<Rle>	<integer>
[1]	seq1	1	*	1
[2]	seq1	3	*	1
[3]	seq1	5	*	1
[4]	seq1	6	*	1
[5]	seq1	9	*	1

```
[6]      seq1          13      * |          2
                                qseq_on_ref
                                <DNAStringSetList>
[1]                               CACTAGTGGCTCATTGTAAA..
[2]                               CTAGTGGCTCATTGTAAATG..
[3]                               AGTGGCTCATTGTAAATGTG..
[4]                               GTGGCTCATTGTAAATTTT..
[5]                               GCTCATTGTAAATGTGTGGT..
[6] ATTGTAAATGTGTGGTTAA...,ATTGTAAATGTGTGGTTAA..
                                qual_on_ref
                                <BStringSetList>
[1]                               <<<<<<<<<<<<;<<<<..
[2]                               <<<<<<<<<0<<<<655<<..
[3]                               <<<<<<<<<<7;71<<;<..
[4]                               (-&- - - -, - - -) - ) , ' - ..
[5]                               <<<<<<<<<<<<<<<<;7<..
[6] <<<<<<<<;<<<8<<<<<<8...,<<<<<;<<<;<;<<<<<<<..
-----
seqinfo: 2 sequences from an unspecified genome
```

Look at `qseq_on_ref` and `qual_on_ref`.

[illegible]

2 reads align to start position 13. Let's have a close look at their sequences:

```
> mcols(gr)$qseq_on_ref[[6]]
```

DNAStrngSet object of length 2:

	width	seq	names
[1]	35	ATTGTAAATGTGTGGTTAACTCGTCCCTGGCCCA	EAS56_61:6:18:467...



## 2.15.3 Create a consensus sequence for each read group family

Compute 1 consensus matrix per chromosome:

```
> split_factor <- rep.int(seqnames(gr), elementNROWS(qseq_on_ref2))
> qseq_on_ref2 <- unlist(qseq_on_ref2, use.names=FALSE)
> qseq_on_ref2_by_chrom <- splitAsList(qseq_on_ref2, split_factor)
> qseq_pos_by_chrom <- splitAsList(start(gr), split_factor)
> cm_by_chrom <- lapply(names(qseq_pos_by_chrom),
+   function(seqname)
+     consensusMatrix(qseq_on_ref2_by_chrom[[seqname]],
+       as.prob=TRUE,
+       shift=qseq_pos_by_chrom[[seqname]]-1,
+       width=seqlengths(gr)[[seqname]]))
> names(cm_by_chrom) <- names(qseq_pos_by_chrom)
```

'cm\_by\_chrom' is a list of consensus matrices. Each matrix has 17 rows (1 per letter in the DNA alphabet) and 1 column per chromosome position.

```
> lapply(cm_by_chrom, dim)

$seq1
[1] 18 1575

$seq2
[1] 18 1584
```

Compute the consensus string from each consensus matrix. We'll put "+" in the strings wherever there is no coverage for that position, and "N" where there is coverage but no consensus.

```
> cs_by_chrom <- lapply(cm_by_chrom,
+   function(cm) {
+     ## need to "fix" 'cm' because consensusString()
+     ## doesn't like consensus matrices with columns
+     ## that contain only zeroes (e.g., chromosome
+     ## positions with no coverage)
+     idx <- colSums(cm) == 0L
+     cm["+ ", idx] <- 1
+     DNASTring(consensusString(cm, ambiguityMap="N"))
+   })
```

The new consensus strings.

```
> cs_by_chrom

$seq1
1575-letter DNASTring object
seq: NANTAGNNNCTCANTTTAAANNTTNTTTTNT...AATNATANNTTNTTNTTNTCTGNAC+++++

$seq2
1584-letter DNASTring object
seq: ++++++.....NNNANANANCTNNA+++++
```

## 2.16 How to compute binned averages along a genome

In some applications (e.g. visualization), there is the need to compute the average of a variable defined along a genome (a.k.a. genomic variable) for a set of predefined fixed-width regions (sometimes called "bins"). The genomic variable is typically represented as a named `RleList` object with one list element per chromosome. One such example is coverage. Here we create an artificial genomic variable:

```
> library(BSgenome.Scerevisiae.UCSC.sacCer2)
> set.seed(55)
> my_var <- RleList(
+   lapply(seqlengths(Scerevisiae),
+     function(seqlen) {
+       tmp <- sample(50L, seqlen, replace=TRUE) %/% 50L
+       Rle(cumsum(tmp - rev(tmp)))
+     }
+   ),
+   compress=FALSE)
> my_var

RleList of length 18
$chrI
integer-Rle of length 230208 with 9197 runs
  Lengths:  6 17 12 12 13 38 15 24 24 25 ... 24 24 15 38 13 12 12 17  7
  Values :  0  1  0  1  2  3  4  3  4  3 ...  4  3  4  3  2  1  0  1  0

$chrII
integer-Rle of length 813178 with 31826 runs
  Lengths: 35 84 50 44  7 67 18  8  7 27 ...  8 18 67  7 44 50 84 35  1
  Values : -1 -2 -1  0  1  0  1  2  1  2 ...  2  1  0  1  0 -1 -2 -1  0

$chrIII
integer-Rle of length 316617 with 12601 runs
  Lengths: 64 16  1 63 48 20 32 43 12 68 ... 12 43 32 20 48 63  1 16 65
  Values :  0  1  0  1  0  1  0  1  2  1 ...  2  1  0  1  0  1  0  1  0

$chrIV
integer-Rle of length 1531919 with 60615 runs
  Lengths:  2 19 38 14 10  8 20 ... 20  8 10 14 38 19  3
  Values :  0 -1 -2 -1 -2 -3 -2 ... -2 -3 -2 -1 -2 -1  0

$chrV
integer-Rle of length 576869 with 22235 runs
  Lengths: 10 69 31  7  3  1  5 ...  5  1  3  7 31 69 11
  Values :  0  1  2  1  2  1  0 ...  0  1  2  1  2  1  0

...
<13 more elements>
```

Use the `tileGenome` function to create a set of bins along the genome.

```
> bins <- tileGenome(seqinfo(Scerevisiae), tilewidth=100,
+   cut.last.tile.in.chrom=TRUE)
```

Compute the binned average for `my_var`:

```
> binnedAverage(bins, my_var, "binned_var")
```

GRanges object with 121639 ranges and 1 metadata column:

	seqnames	ranges	strand	binned_var
	<Rle>	<IRanges>	<Rle>	<numeric>
[1]	chrI	1-100	*	1.77
[2]	chrI	101-200	*	3.34
[3]	chrI	201-300	*	3.22
...	...	...	...	...
[121637]	2micron	6101-6200	*	-1.000000
[121638]	2micron	6201-6300	*	-0.750000
[121639]	2micron	6301-6318	*	-0.555556

-----  
seqinfo: 18 sequences (2 circular) from sacCer2 genome

The bin size can be modified with the `tilewidth` argument to `tileGenome`. See `?binnedAverage` for additional examples.

### 3 Session Information

```
R version 4.4.1 (2024-06-14)
Platform: x86_64-apple-darwin20
Running under: macOS Monterey 12.7.6

Matrix products: default
BLAS:   /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/lib/libRblas.0.dylib
LAPACK: /Library/Frameworks/R.framework/Versions/4.4-x86_64/Resources/lib/libRlapack.dylib; LAPACK version 3.11.0

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

time zone: America/New_York
tzcode source: internal

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

other attached packages:
 [1] BSgenome.Scerevisiae.UCSC.sacCer2_1.4.0
 [2] KEGGgraph_1.64.0
 [3] KEGGREST_1.44.1
 [4] BSgenome.Hsapiens.UCSC.hg38_1.4.5
 [5] BSgenome_1.72.0
 [6] rtracklayer_1.64.0
 [7] BiocIO_1.14.0
 [8] edgeR_4.2.1
 [9] limma_3.60.6
[10] DESeq2_1.44.0
```

```
[11] AnnotationHub_3.12.0
[12] BiocFileCache_2.12.0
[13] dbplyr_2.5.0
[14] txdbmaker_1.0.1
[15] TxDb.Athaliana.BioMart.plantsmart22_3.0.1
[16] TxDb.Dmelanogaster.UCSC.dm3.ensGene_3.2.2
[17] GenomicAlignments_1.40.0
[18] Rsamtools_2.20.0
[19] Biostrings_2.72.1
[20] XVector_0.44.0
[21] SummarizedExperiment_1.34.0
[22] MatrixGenerics_1.16.0
[23] matrixStats_1.4.1
[24] pasillaBamSubset_0.42.0
[25] TxDb.Hsapiens.UCSC.hg38.knownGene_3.18.0
[26] GenomicFeatures_1.56.0
[27] AnnotationDbi_1.66.0
[28] Biobase_2.64.0
[29] GenomicRanges_1.56.2
[30] GenomeInfoDb_1.40.1
[31] IRanges_2.38.1
[32] S4Vectors_0.42.1
[33] BiocGenerics_0.50.0
[34] BiocStyle_2.32.1
```

loaded via a namespace (and not attached):

```
[1] DBI_1.2.3                bitops_1.0-9
[3] httr2_1.0.5              biomaRt_2.60.1
[5] rlang_1.1.4              magrittr_2.0.3
[7] compiler_4.4.1           RSQLite_2.3.7
[9] png_0.1-8                vctrs_0.6.5
[11] stringr_1.5.1            pkgconfig_2.0.3
[13] crayon_1.5.3             fastmap_1.2.0
[15] utf8_1.2.4               rmarkdown_2.28
[17] graph_1.82.0             UCSC.utils_1.0.0
[19] purrr_1.0.2              bit_4.5.0
[21] xfun_0.48                zlibbioc_1.50.0
[23] cachem_1.1.0             jsonlite_1.8.9
[25] progress_1.2.3           blob_1.2.4
[27] DelayedArray_0.30.1      BiocParallel_1.38.0
[29] parallel_4.4.1           prettyunits_1.2.0
[31] R6_2.5.1                 VariantAnnotation_1.50.0
[33] bslib_0.8.0              stringi_1.8.4
[35] jquerylib_0.1.4          Rcpp_1.0.13
[37] bookdown_0.40            knitr_1.48
[39] Matrix_1.7-0             tidyselect_1.2.1
[41] abind_1.4-8              yaml_2.3.10
[43] codetools_0.2-20         curl_5.2.3
[45] lattice_0.22-6           tibble_3.2.1
[47] withr_3.0.1              evaluate_1.0.0
[49] xml2_1.3.6               pillar_1.9.0
```

[51] BiocManager_1.30.25	filelock_1.0.3
[53] generics_0.1.3	RCurl_1.98-1.16
[55] BiocVersion_3.19.1	hms_1.1.3
[57] ggplot2_3.5.1	munSELL_0.5.1
[59] scales_1.3.0	glue_1.8.0
[61] tools_4.4.1	locfit_1.5-9.10
[63] XML_3.99-0.17	grid_4.4.1
[65] colorspace_2.1-1	GenomeInfoDbData_1.2.12
[67] restfulr_0.0.15	cli_3.6.3
[69] rappdirs_0.3.3	fansi_1.0.6
[71] S4Arrays_1.4.1	dplyr_1.1.4
[73] Rgraphviz_2.48.0	gtable_0.3.5
[75] sass_0.4.9	digest_0.6.37
[77] SparseArray_1.4.8	rjson_0.2.23
[79] memoise_2.0.1	htmltools_0.5.8.1
[81] lifecycle_1.0.4	httr_1.4.7
[83] statmod_1.5.0	mime_0.12
[85] bit64_4.5.2	

## References

Michael Lawrence, Wolfgang Huber, Hervé Pagès, Patrick Aboyoun, Marc Carlson, Robert Gentleman, Martin T. Morgan, and Vincent J. Carey. Software for computing and annotating genomic ranges. *PLOS Computational Biology*, 4(3), 2013.