Genomic_Ranges_Assignment

piotyama

April 21, 2017

Genomic Ranges Assignment, Lecture slides 132-139.

setwd for all chuncks

Install Bioconductor's Primary Packages

```
source("http://bioconductor.org/biocLite.R")
```

```
## Bioconductor version 3.4 (BiocInstaller 1.24.0), ?biocLite for help
```

```
## A new version of Bioconductor is available after installing the most
##   recent version of R; see http://bioconductor.org/install
```

```
biocLite("GenomicRanges")
```

```
## BioC_mirror: https://bioconductor.org
```

```
## Using Bioconductor 3.4 (BiocInstaller 1.24.0), R 3.3.2 (2016-10-31).
```

```
## Installing package(s) 'GenomicRanges'
```

```
## package 'GenomicRanges' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Paulotyama\AppData\Local\Temp\RtmpqKFOEV\downloaded_packages
```

```
## installation path not writeable, unable to update packages: boot, cluster,
##   foreign, lattice, MASS, rpart, survival
```

```
biocLite("GenomicFeatures")
```

```
## BioC_mirror: https://bioconductor.org
```

```
## Using Bioconductor 3.4 (BiocInstaller 1.24.0), R 3.3.2 (2016-10-31).
```

```
## Installing package(s) 'GenomicFeatures'
```

```
## package 'GenomicFeatures' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Paulotyama\AppData\Local\Temp\RtmpqKFOEV\downloaded_packages
```

```
## installation path not writeable, unable to update packages: boot, cluster,
##   foreign, lattice, MASS, rpart, survival
```

```
biocLite("rtracklayer")
```

```
## BioC_mirror: https://bioconductor.org
```

```
## Using Bioconductor 3.4 (BiocInstaller 1.24.0), R 3.3.2 (2016-10-31).
```

```
## Installing package(s) 'rtracklayer'
```

```
## package 'rtracklayer' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\Paulotyama\AppData\Local\Temp\RtmpqKFOEV\downloaded_packages
## installation path not writeable, unable to update packages: boot, cluster,
##   foreign, lattice, MASS, rpart, survival
library(IRanges)
## Warning: package 'IRanges' was built under R version 3.3.3
## Loading required package: BiocGenerics
## Loading required package: parallel
##
## Attaching package: 'BiocGenerics'
## The following objects are masked from 'package:parallel':
##
##     clusterApply, clusterApplyLB, clusterCall, clusterEvalQ,
##     clusterExport, clusterMap, parApply, parCapply, parLapply,
##     parLapplyLB, parRapply, parSapply, parSapplyLB
## The following objects are masked from 'package:stats':
##
##     IQR, mad, xtabs
## The following objects are masked from 'package:base':
##
##     anyDuplicated, append, as.data.frame, cbind, colnames,
##     do.call, duplicated, eval, evalq, Filter, Find, get, grep,
##     grepl, intersect, is.unsorted, lapply, lengths, Map, mapply,
##     match, mget, order, paste, pmax, pmax.int, pmin, pmin.int,
##     Position, rank, rbind, Reduce, rownames, sapply, setdiff,
##     sort, table, tapply, union, unique, unsplit, which, which.max,
##     which.min
## Loading required package: S4Vectors
## Warning: package 'S4Vectors' was built under R version 3.3.3
## Loading required package: stats4
##
## Attaching package: 'S4Vectors'
## The following objects are masked from 'package:base':
##
##     colMeans, colSums, expand.grid, rowMeans, rowSums
library(GenomicRanges)
## Warning: package 'GenomicRanges' was built under R version 3.3.3
## Loading required package: GenomeInfoDb
library(rtracklayer)
```

import a file with variants (SNPs, indels, etc...) from chr1 of Mus musculus File found in the course repository: BCB546X-Spring2016/bds-files/chapter-09-working-with-range-data

```
dbsnp137 <- import("mm10_snp137_chr1_trunc.bed.gz")
```

create a mouse transcript db using GenomicFeatures

```
biocLite("TxDb.Mmusculus.UCSC.mm10.ensGene")
```

```
## BioC_mirror: https://bioconductor.org
## Using Bioconductor 3.4 (BiocInstaller 1.24.0), R 3.3.2 (2016-10-31).
## Installing package(s) 'TxDb.Mmusculus.UCSC.mm10.ensGene'
## installing the source package 'TxDb.Mmusculus.UCSC.mm10.ensGene'
## installation path not writeable, unable to update packages: boot, cluster,
##   foreign, lattice, MASS, rpart, survival
```

```
library(TxDb.Mmusculus.UCSC.mm10.ensGene)
```

```
## Loading required package: GenomicFeatures
## Warning: package 'GenomicFeatures' was built under R version 3.3.3
## Loading required package: AnnotationDbi
## Loading required package: Biobase
## Welcome to Bioconductor
##
##     Vignettes contain introductory material; view with
##     'browseVignettes()'. To cite Bioconductor, see
##     'citation("Biobase")', and for packages 'citation("pkgname")'.
```

```
txdb <- TxDb.Mmusculus.UCSC.mm10.ensGene
```

checking the contents of txdb

```
genes(txdb)
```

```
## GRanges object with 39017 ranges and 1 metadata column:
##                       seqnames                ranges strand |
##                          <Rle>             <IRanges>  <Rle> |
##   ENSMUSG00000000001     chr3 [108107280, 108146146]      - |
##   ENSMUSG00000000003     chrX [ 77837901,  77853623]      - |
##   ENSMUSG00000000028    chr16 [ 18780447,  18811987]      - |
##   ENSMUSG00000000031     chr7 [142575529, 142578143]      - |
##   ENSMUSG00000000037     chrX [161117193, 161258213]      + |
##                  ...      ...                   ...    ... .
##   ENSMUSG00000099329    chr17   [29549799, 29674187]      + |
##   ENSMUSG00000099330     chr7   [44538106, 44538178]      - |
##   ENSMUSG00000099331    chr11   [46170087, 46170143]      + |
##   ENSMUSG00000099332    chr17   [85618067, 85618265]      + |
##   ENSMUSG00000099334     chr4   [74635214, 74635351]      + |
##                              gene_id
##                          <character>
##   ENSMUSG00000000001 ENSMUSG00000000001
##   ENSMUSG00000000003 ENSMUSG00000000003
##   ENSMUSG00000000028 ENSMUSG00000000028
##   ENSMUSG00000000031 ENSMUSG00000000031
##   ENSMUSG00000000037 ENSMUSG00000000037
##                  ...                ...
```

```
##   ENSMUSG00000099329 ENSMUSG00000099329
##   ENSMUSG00000099330 ENSMUSG00000099330
##   ENSMUSG00000099331 ENSMUSG00000099331
##   ENSMUSG00000099332 ENSMUSG00000099332
##   ENSMUSG00000099334 ENSMUSG00000099334
##   -------
##   seqinfo: 66 sequences (1 circular) from mm10 genome
```

```
transcripts(txdb)
```

```
## GRanges object with 94647 ranges and 2 metadata columns:
##                 seqnames             ranges strand |     tx_id
##                    <Rle>          <IRanges>  <Rle> | <integer>
##       [1]           chr1 [3054233, 3054733]      + |         1
##       [2]           chr1 [3102016, 3102125]      + |         2
##       [3]           chr1 [3466587, 3513553]      + |         3
##       [4]           chr1 [4529017, 4529123]      + |         4
##       [5]           chr1 [4807788, 4848410]      + |         5
##       ...            ...                ...    ... .       ...
##   [94643] chrUn_GL456381     [16623, 16721]      - |     94643
##   [94644] chrUn_GL456385     [31243, 31343]      + |     94644
##   [94645] chrUn_GL456385     [32719, 32818]      + |     94645
##   [94646] chrUn_JH584304     [52190, 59667]      - |     94646
##   [94647] chrUn_JH584304     [52691, 59690]      - |     94647
##                     tx_name
##                 <character>
##       [1] ENSMUST00000160944
##       [2] ENSMUST00000082908
##       [3] ENSMUST00000161581
##       [4] ENSMUST00000180019
##       [5] ENSMUST00000134384
##       ...                ...
##   [94643] ENSMUST00000184505
##   [94644] ENSMUST00000178705
##   [94645] ENSMUST00000180206
##   [94646] ENSMUST00000179505
##   [94647] ENSMUST00000178343
##   -------
##   seqinfo: 66 sequences (1 circular) from mm10 genome
```

```
exons(txdb)
```

```
## GRanges object with 348801 ranges and 1 metadata column:
##                  seqnames             ranges strand |   exon_id
##                     <Rle>          <IRanges>  <Rle> | <integer>
##        [1]           chr1 [3054233, 3054733]      + |         1
##        [2]           chr1 [3102016, 3102125]      + |         2
##        [3]           chr1 [3466587, 3466687]      + |         3
##        [4]           chr1 [3513405, 3513553]      + |         4
##        [5]           chr1 [4529017, 4529123]      + |         5
##        ...            ...                ...    ... .       ...
##   [348797] chrUn_JH584304     [55112, 55701]      - |    348797
##   [348798] chrUn_JH584304     [56986, 57151]      - |    348798
##   [348799] chrUn_JH584304     [58564, 58835]      - |    348799
##   [348800] chrUn_JH584304     [58564, 59690]      - |    348800
```

```
##   [348801] chrUn_JH584304     [59592, 59667]      - |    348801
##   -------
##   seqinfo: 66 sequences (1 circular) from mm10 genome
```

```
promoters(txdb)
```

```
## Warning in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE): GRanges object contains 3 out-of-bou
##   chr4_JH584293_random, chr4_JH584295_random, and
##   chr5_JH584296_random. Note that only ranges located on a
##   non-circular sequence whose length is not NA can be considered
##   out-of-bound (use seqlengths() and isCircular() to get the lengths
##   and circularity flags of the underlying sequences). You can use
##   trim() to trim these ranges. See ?`trim,GenomicRanges-method` for
##   more information.

## Warning in valid.GenomicRanges.seqinfo(x, suggest.trim = TRUE): GRanges object contains 3 out-of-bou
##   chr4_JH584293_random, chr4_JH584295_random, and
##   chr5_JH584296_random. Note that only ranges located on a
##   non-circular sequence whose length is not NA can be considered
##   out-of-bound (use seqlengths() and isCircular() to get the lengths
##   and circularity flags of the underlying sequences). You can use
##   trim() to trim these ranges. See ?`trim,GenomicRanges-method` for
##   more information.
```

```
## GRanges object with 94647 ranges and 2 metadata columns:
##                 seqnames             ranges strand |     tx_id
##                    <Rle>          <IRanges>  <Rle> | <integer>
##       [1]           chr1 [3052233, 3054432]      + |         1
##       [2]           chr1 [3100016, 3102215]      + |         2
##       [3]           chr1 [3464587, 3466786]      + |         3
##       [4]           chr1 [4527017, 4529216]      + |         4
##       [5]           chr1 [4805788, 4807987]      + |         5
##       ...            ...                ...    ... .       ...
##   [94643] chrUn_GL456381     [16522, 18721]      - |     94643
##   [94644] chrUn_GL456385     [29243, 31442]      + |     94644
##   [94645] chrUn_GL456385     [30719, 32918]      + |     94645
##   [94646] chrUn_JH584304     [59468, 61667]      - |     94646
##   [94647] chrUn_JH584304     [59491, 61690]      - |     94647
##                     tx_name
##                 <character>
##       [1] ENSMUST00000160944
##       [2] ENSMUST00000082908
##       [3] ENSMUST00000161581
##       [4] ENSMUST00000180019
##       [5] ENSMUST00000134384
##       ...                ...
##   [94643] ENSMUST00000184505
##   [94644] ENSMUST00000178705
##   [94645] ENSMUST00000180206
##   [94646] ENSMUST00000179505
##   [94647] ENSMUST00000178343
##   -------
##   seqinfo: 66 sequences (1 circular) from mm10 genome
```

collapse all overlapping exons in the mouse TranscriptDb

```
collapsed_exons <- reduce(exons(txdb), ignore.strand=TRUE)
```

create an object with only exons from chr1

```
chr1_collapsed_exons <- collapsed_exons[seqnames(collapsed_exons) == "chr1"]
```

inspect our variant file

```
summary(width(dbsnp137))
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.000   1.000   1.000   1.142   1.000 732.000
```

adjusting variant width so as to find overlap with exon ranges

```
dbsnp137_resized <- dbsnp137
zw_i <- width(dbsnp137_resized) == 0
dbsnp137_resized[zw_i] <- resize(dbsnp137_resized[zw_i], width=1)
```

pull out those variants that overlap exons on chromosome 1

```
hits <- findOverlaps(dbsnp137_resized, chr1_collapsed_exons,ignore.strand=TRUE)
```

determine the number of variants that are exonic

```
length(unique(queryHits(hits)))
```

```
## [1] 57623
```

determine the proportion of variants that are exonic

```
length(unique(queryHits(hits)))/length(dbsnp137_resized)
```

```
## [1] 0.02134185
```

determine the number of variants per exon

```
var_counts <- countOverlaps(chr1_collapsed_exons, dbsnp137_resized, ignore.strand=TRUE)
```

append this to our GRanges object that includes exons

```
chr1_collapsed_exons$num_vars <- var_counts
```

write this into a csv file

```
write.table(chr1_collapsed_exons, file="chr1_collapsed_exons.csv", row.names=FALSE,sep=",")
```