

SPLINTER

SPLice INTERpreter: Alternative splicing analysis toolkit

Diana Low

24 April 2017

Package version: SPLINTER 1.2.0

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 2 |
| 2 | Loading the package | 2 |
| 3 | Initializing the genome for transcript selection | 2 |
| 4 | Reading in the splicing analysis file | 2 |
| 4.1 | Additional annotation | 3 |
| 5 | Analyzing a specific gene | 4 |
| 5.1 | Inspecting a single gene in more detail (single record) | 4 |
| 5.2 | Finding relevant transcripts from the ENSEMBL database | 4 |
| 5.3 | Constructing the region of interest (ROI) | 4 |
| 5.4 | Finding transcripts that contain the ROI | 6 |
| 6 | Simulating alternatively spliced products | 7 |
| 6.1 | Simulating the outcome of exon skipping by removing an exonic region | 7 |
| 6.2 | Simulating the outcome of intron retention by inserting an intronic region | 7 |
| 6.3 | Comparing sequences before and after removal/insertion of a region | 7 |
| 7 | Designing primers to inspect splicing regions | 8 |
| 7.1 | Getting the DNA of the region of interest | 8 |
| 7.2 | Using Primer3 to design primers for alternative splicing identification | 8 |
| 7.3 | Checking primers coverage | 8 |
| 7.4 | Predicting PCR results using the primers | 9 |
| 8 | Plotting results | 9 |
| 9 | Session info | 11 |

1 Introduction

SPLINTER provides tools to analyze alternative splicing sites, interpret outcomes based on sequence information, select and design primers for site validation and give visual representation of the event to guide downstream experiments.

2 Loading the package

To load the *SPLINTER* package:

```
library(SPLINTER)
```

3 Initializing the genome for transcript selection

In this example, we will be utilizing the mm9 genome for mouse. You will need to install the appropriate package (eg. *BSgenome.Mmusculus.UCSC.mm9*) for the genome that you will be using.

```
library(BSgenome.Mmusculus.UCSC.mm9)
library(GenomicFeatures)
bsgenome <- BSgenome.Mmusculus.UCSC.mm9
```

We begin with full set of available transcripts to screen from, and read it into a *TxDb* object. One source of this (best option to ensure compatibility) would be the GTF file that you have used for alternative splicing analysis. For other sources of data, please refer to *GenomicFeatures*).

We then extract the coding sequences (CDS), and transcripts in general (coding and non-coding) (exons) from this object.

```
data_path<-system.file("extdata",package="SPLINTER")
gtf_file<-paste(data_path,"/Mus_musculus.Ensembl.NCBIM37.65.partial.gtf",sep="")
txdb <- makeTxDbFromGFF(file=gtf_file,chrominfo = Seqinfo(genome="mm9"))
## Warning: closing unused connection 6 (C:/Users/biocbuild/
## bbs-3.5-bioc/tmpdir/Rtmp27NXdk/Rinst17c424727c1/SPLINTER/extdata/
## Mus_musculus.Ensembl.NCBIM37.65.partial.gtf)
## Warning: closing unused connection 5 (C:/Users/biocbuild/
## bbs-3.5-bioc/tmpdir/Rtmp27NXdk/Rinst17c424727c1/SPLINTER/extdata/
## Mus_musculus.Ensembl.NCBIM37.65.partial.gtf)

# txdb generation can take quite long, you can save the object and load it the next time
# saveDb(txdb,file="txdb_hg19.sqlite")
# txdb<-loadDb(file="txdb_hg19.sqlite")

# extract CDS and exon information from TxDb object
thecds<-cdsBy(txdb,by="tx",use.names=TRUE)
theexons<-exonsBy(txdb,by="tx",use.names=TRUE)
```

4 Reading in the splicing analysis file

The output file from *MATS* is used here, but essentially all that is needed are coordinates of the exons (target and flanking) involved in the splicing process to be studied. For the case of exon skipping, this will include the upstream, target and downstream exons. More output types will be supported in the future.

The following types of alternative splicing events are accepted:

| Type of alternative splicing event | Definition |
|------------------------------------|----------------------------|
| SE | Skipped exon |
| RI | Retained intron |
| MXE | Mutually exclusive exon |
| A5SS | Alternative 5' splice site |
| A3SS | Alternative 3' splice site |

```
typeofAS="SE"
mats_file<-paste(data_path,"/skipped_exons.txt",sep="")
splice_data <-extractSpliceEvents(data=mats_file, filetype='mats', splicetype=typeofAS)
splice_data$data[,c(1:10)]
##           ID           Symbol   chr strand exonStart
## 4825  ENSMUSG000000052337 ENSMUSG000000052337 chr6      + 71816720
## 17227 ENSMUSG000000023110 ENSMUSG000000023110 chr14     - 55132128
## 22583 ENSMUSG000000079477 ENSMUSG000000079477 chr6      - 87965626
## 13693 ENSMUSG000000024911 ENSMUSG000000024911 chr19     + 5464334
## 18826 ENSMUSG000000027940 ENSMUSG000000027940 chr3      + 89894935
##           exonEnd upstreamStart upstreamEnd downstreamStart downstreamEnd
## 4825  71816734      71813135      71813264      71818574      71818797
## 17227 55132291      55130830      55130991      55133434      55133483
## 22583 87965712      87963632      87963692      87995067      87995239
## 13693 5464431       5464129       5464215       5464925       5465051
## 18826 89895013      89893932      89894001      89903450      89904487
```

4.1 Additional annotation

SPLINTER assumes that the main identifier is ENSEMBL, however gene symbols can be added.

```
splice_data<-addEnsemblAnnotation(data=splice_data,species="mmusculus")

# (Optional) Sorting the dataframe, if you have supporting statistical information
splice_data$data<-splice_data$data[with(splice_data$data,order(FDR,-IncLevelDifference)),]
head(splice_data$data[,c(1:10)])
##           ID Symbol   chr strand exonStart  exonEnd upstreamStart
## 1 ENSMUSG000000052337  Immt chr6      + 71816720 71816734      71813135
## 2 ENSMUSG000000023110  Prmt5 chr14     - 55132128 55132291      55130830
## 3 ENSMUSG000000079477  Rab7 chr6      - 87965626 87965712      87963632
## 4 ENSMUSG000000024911  Fibp chr19     + 5464334 5464431      5464129
## 5 ENSMUSG000000027940  Tpm3 chr3      + 89894935 89895013      89893932
##           upstreamEnd downstreamStart downstreamEnd
## 1      71813264      71818574      71818797
## 2      55130991      55133434      55133483
## 3      87963692      87995067      87995239
## 4      5464215       5464925       5465051
## 5      89894001      89903450      89904487
```

5 Analyzing a specific gene

5.1 Inspecting a single gene in more detail (single record)

Once we have defined the events, we will pick 1 event to analyze.

```
single_id='Prmt5'
pp<-which(grepl(single_id,splice_data$data$Symbol)) # Prmt5 has 1 record

splice_data$data[pp,c(1:6)] # show all records
##           ID Symbol   chr strand exonStart  exonEnd
## 2 ENSMUSG00000023110 Prmt5 chr14      -   55132128 55132291

single_record<-splice_data$data[pp[1],]
```

5.2 Finding relevant transcripts from the ENSEMBL database

To reduce search complexity, we define the valid transcripts and coding sequences with regards to our gene of interest. We find that Prmt5 has 7 transcripts, 2 of which are coding sequences.

```
valid_tx <- findTX(id=single_record$ID,tx=theexons,db=txdb)
## ENSMUST00000023873
## ENSMUST00000139964
## ENSMUST00000132227
## ENSMUST00000147214
## ENSMUST00000133552
## ENSMUST00000138367
## ENSMUST00000132801
## 7 valid transcripts found.

valid_cds<- findTX(id=single_record$ID,tx=thecds,db=txdb)
## ENSMUST00000023873
## ENSMUST00000139964
## 2 valid transcripts found.
```

5.3 Constructing the region of interest (ROI)

The makeROI function will create a list containing GRanges objects for the splicing event. This will help identify and construct relevant outputs later.

This list contains the following information:

- type: type of alternative splicing event
- name: name of gene
- roi: GRanges object of the exon
- flank: GRanges object of the flanking exons
- roi_range: GRanges list containing
 - GRanges object of Type 1
 - GRanges object of Type 2

| Type of alternative splicing | Type 1 representation | Type 2 representation (annotated only) |
|------------------------------|---|--|
| SE | isoform with event exon included | isoform with the exon skipped |
| RI | isoform with normal exon boundaries | isoform with the intron retained |
| MXE | isoform defined 1st (leftmost) in input | isoform defined 2nd in input |
| A5SS | isoform with longer exon | isoform with shorter exon |
| A3SS | isoform with longer exon | isoform with shorter exon |

```

roi <- makeROI(single_record,type="SE")
roi
## $type
## [1] "SE"
##
## $name
## [1] "ENSMUSG00000023110"
##
## $roi
## GRanges object with 1 range and 1 metadata column:
##      seqnames      ranges strand | exon_rank
##      <Rle>         <IRanges> <Rle> | <integer>
## [1] chr14 [55132128, 55132291] - | 1
## -----
##      seqinfo: 1 sequence from an unspecified genome; no seqlengths
##
## $flank
## GRanges object with 2 ranges and 1 metadata column:
##      seqnames      ranges strand | exon_rank
##      <Rle>         <IRanges> <Rle> | <integer>
## [1] chr14 [55133434, 55133483] - | 2
## [2] chr14 [55130830, 55130991] - | 1
## -----
##      seqinfo: 1 sequence from an unspecified genome; no seqlengths
##
## $roi_range
## GRangesList object of length 2:
## [[1]]
## GRanges object with 3 ranges and 0 metadata columns:
##      seqnames      ranges strand
##      <Rle>         <IRanges> <Rle>
## [1] chr14 [55133434, 55133483] -
## [2] chr14 [55132128, 55132291] -
## [3] chr14 [55130830, 55130991] -
##
## [[2]]
## GRanges object with 2 ranges and 0 metadata columns:
##      seqnames      ranges strand
## [1] chr14 [55133434, 55133483] -
## [2] chr14 [55130830, 55130991] -
##
## -----
##      seqinfo: 1 sequence from an unspecified genome; no seqlengths

```

5.4 Finding transcripts that contain the ROI

At this juncture, we look for transcripts are compatible with the ROI. Compatibility is defined as having the exact cassette (matching upstream, target, downstream) exons. In the case of intron retention, this would just be the 2 exons flanking the intron.

We notice here that Prmt5 only has 1 compatible transcript involved in the event ROI, out of 7 transcripts (or 2 coding transcripts). There are no Type 2 transcripts, which means there are no annotated transcripts of Prmt5 containing the alternative event.

```
compatible_tx<- findCompatibleEvents(valid_tx,roi=roi,verbose=TRUE)
## Checking Type 1.....
## ENSMUST00000023873
## Checking Type 2.....
## No transcripts found!
## 1 match(es) from original 7 transcripts.

compatible_cds<- findCompatibleEvents(valid_cds,valid_tx,roi=roi,verbose=TRUE)
## Checking Type 1.....
## ENSMUST00000023873
## Checking Type 2.....
## No transcripts found!
## 1 match(es) from original 2 transcripts.
```

6 Simulating alternatively spliced products

6.1 Simulating the outcome of exon skipping by removing an exonic region

```
region_minus_exon <-removeRegion(compatible_cds$hits[[1]],roi)
```

6.2 Simulating the outcome of intron retention by inserting an intronic region

```
# Not relevant for this Prmt5 skipped exon example
region_plus_exon <-insertRegion(region_minus_exon,roi)
```

6.3 Comparing sequences before and after removal/insertion of a region

```
event<-eventOutcomeCompare(seq1=compatible_cds$hits[[1]],seq2=region_minus_exon,
                           genome=bsgenome,direction=TRUE,fullseq=FALSE)

##
## ### ENSMUST00000023873 ###
## - early termination
## middle insertion GILKPK (247-252)
## middle deletion LEIGADLP (206-213)
## middle deletion EPIK (224-227)
## middle deletion KAAIL (227-231)
## multiple mismatch sites
## 3' end mismatch : ALEIGADLP SNHVIDRWLG EPIK AAILPTSIFLTNKKGFPVLSKVQQRLLIFRLLKLEVQFIITGTNHHSEKEFC SYLQYLEYLSQ
## length : 637 AA to 242 AA

event
## $alignment
## Global-Local PairwiseAlignmentsSingleSubject (1 of 1)
## pattern: [1] MAAMAVGGAGGSRVSSGRDLNCVPEIADTLGA...GVL----FLP-----PVLGILKPKSPSTQCL*
## subject: [1] MAAMAVGGAGGSRVSSGRDLNCVPEIADTLGA...AILPTSIFLTNKKGFPVL-----SKVQQRLLI
## score: 1085.5
##
## $eventtypes
## [1] "(NMD)"
```

7 Designing primers to inspect splicing regions

7.1 Getting the DNA of the region of interest

This function will return the DNA of the ROI, with exons separated by “[]” (Primer3 notation) and the junction marked by jstart.

```
aa<-getRegionDNA(roi,bsgenome)
aa
## $seq
## [1] "GTGCATAACTTTTCGGACTCTGTGTGACTATAGCAAGAGAATTGCAGTAG [] TTGGAAGTGCAGTTTATCATCACGGGAACCAACCACCTCAGAGA
##
## $jstart
## [1] 51
```

7.2 Using Primer3 to design primers for alternative splicing identification

We have included a helper function to run Primer3 from within R. You will need to define the path to your Primer3 installation. Refer to ?callPrimer3 for more details.

```
primers<-callPrimer3(seq=aa$seq,sequence_target = aa$jstart,size_range='100-500')
```

```
primers[,c(1:4)]
##   i    PRIMER_LEFT_SEQUENCE PRIMER_RIGHT_SEQUENCE PRIMER_LEFT_TM
## 1 0    ACTTTCGGACTCTGTGTGACT TCATAGGCATTGGGTGGAGG          58.967
## 2 1    TGGCATAACTTTTCGGACTCTG  GGAGTGGGGACTGCAGATAG          58.027
## 3 2    GCATAACTTTCGGACTCTGTGT  CTCCTTCTCTGAGTGGTGGT          58.673
## 4 3    TCGGACTCTGTGTGACTATAGC  ATTGGGTGGAGGGCGATTTT          59.056
## 5 4    GGACTCTGTGTGACTATAGCAAG  GCTCATAGGCATTGGGTGG          58.322
```

Alternatively, primers can be entered manually with the appropriate headers.

```
primers <- data.frame(PRIMER_LEFT_SEQUENCE="ACTTTCGGACTCTGTGTGACT",
                      PRIMER_RIGHT_SEQUENCE="TCATAGGCATTGGGTGGAGG",
                      stringsAsFactors=FALSE)
```

7.3 Checking primers coverage

As a confirmation, we can run the primers against the ROI to give the genomic location of the primer coverage.

```
cp<-checkPrimer(primers[1,],bsgenome,roi)
cp
## $total_span
## GRanges object with 1 range and 0 metadata columns:
##       seqnames          ranges strand
##       <Rle>             <IRanges> <Rle>
## [1] chr14 [55130876, 55133475]      *
## -----
## seqinfo: 1 sequence from an unspecified genome; no seqlengths
##
## $primer_left_span
## GRanges object with 1 range and 0 metadata columns:
##       seqnames          ranges strand
```



```
##          <Rle>          <IRanges> <Rle>
## [1]    chr14 [55133455, 55133475]      *
## -----
## seqinfo: 1 sequence from an unspecified genome; no seqlengths
##
## $primer_right_span
## GRanges object with 1 range and 0 metadata columns:
##      seqnames      ranges strand
##      <Rle>        <IRanges> <Rle>
## [1]    chr14 [55130876, 55130895]      *
## -----
## seqinfo: 1 sequence from an unspecified genome; no seqlengths
```

7.4 Predicting PCR results using the primers

getPCRsizes will give you the length of the PCR product produced by the set of primers.

```
pcr_result1<-getPCRsizes(cp,theexons)
pcr_result1
##              ID  bp
## 1 ENSMUST00000023873 322

tx_minus_exon <-removeRegion(compatible_tx$hits[[1]],roi)
pcr_result2<-getPCRsizes(cp,tx_minus_exon)
pcr_result2
##              ID  bp
## 1 ENSMUST00000023873 158
```

7.4.1 Selecting sizes relevant to splicing event (subset of getPCRsizes)

While getPCRsizes will return all possible PCR products for a given set of annotation, splitPCRhit will return PCR product sizes that are relevant to the splicing event in question.

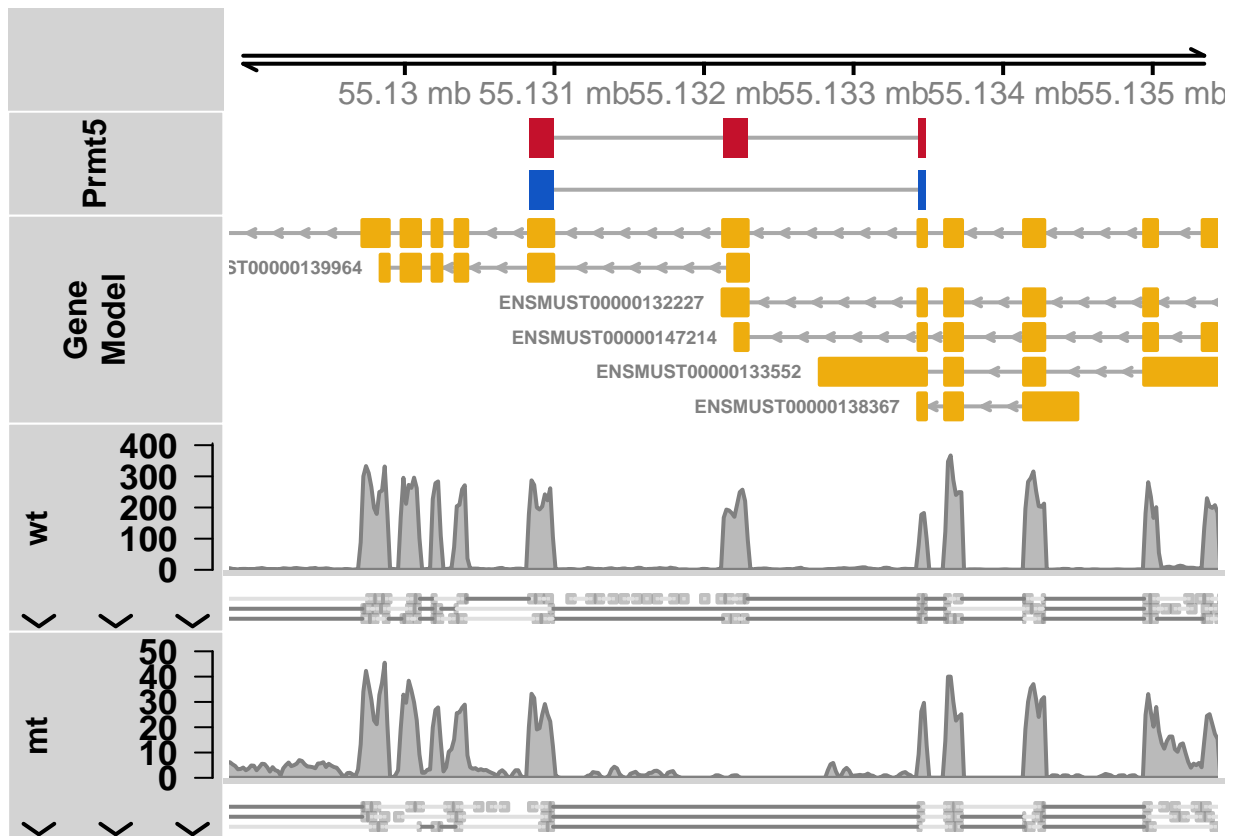
```
relevant_pcr_bands<-splitPCRhit(pcr_result1,compatible_tx)

relevant_pcr_bands
## $Type1
##              ID  bp
## 1 ENSMUST00000023873 322
##
## $Type2
## [1] ID bp
## <0 rows> (or 0-length row.names)
```

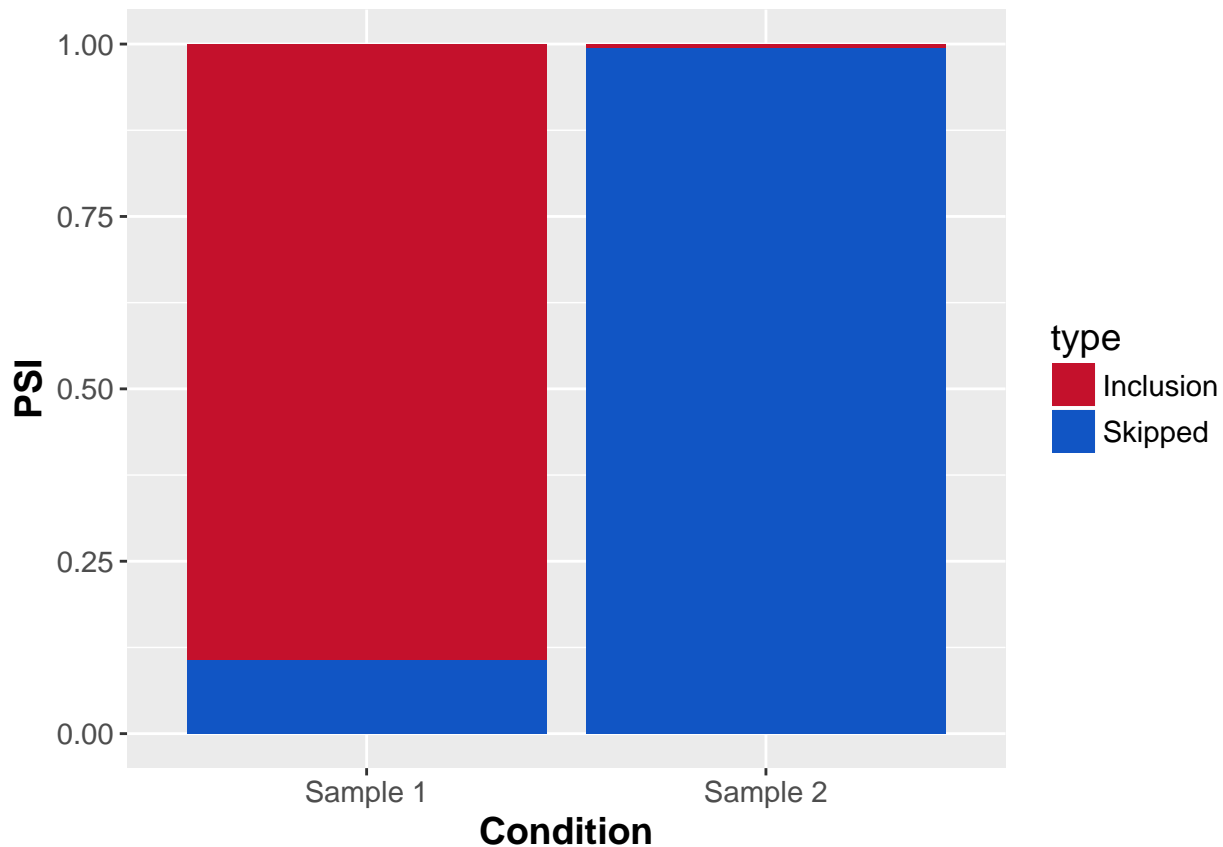
8 Plotting results

```
# reading in BAM files
mt<-paste(data_path,"/mt_chr14.bam",sep="")
wt<-paste(data_path,"/wt_chr14.bam",sep="")
```

```
# Plotting genomic range, read density and splice changes
eventPlot(transcripts=theexons,roi_plot=roi,bams=c(wt,mt),names=c('wt','mt'),
          annoLabel=single_id,rspan=2000)
```



```
# Barplot of PSI values if provided
psiPlot(single_record)
```



9 Session info

```
## R version 3.4.0 (2017-04-21)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows Server 2012 R2 x64 (build 9600)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=C
## [2] LC_CTYPE=English_United States.1252
## [3] LC_MONETARY=English_United States.1252
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.1252
##
## attached base packages:
## [1] stats4      parallel  stats      graphics  grDevices  utils      datasets
## [8] methods     base
##
## other attached packages:
## [1] GenomicFeatures_1.28.0      AnnotationDbi_1.38.0
## [3] Biobase_2.36.0             BSgenome.Mmusculus.UCSC.mm9_1.4.0
## [5] BSgenome_1.44.0            rtracklayer_1.36.0
## [7] Biostrings_2.44.0           XVector_0.16.0
```

```

## [9] GenomicRanges_1.28.0          GenomeInfoDb_1.12.0
## [11] IRanges_2.10.0                S4Vectors_0.14.0
## [13] BiocGenerics_0.22.0           SPLINTER_1.2.0
## [15] BiocStyle_2.4.0
##
## loaded via a namespace (and not attached):
## [1] httr_1.2.1                    AnnotationHub_2.8.0
## [3] splines_3.4.0                 shiny_1.0.2
## [5] Formula_1.2-1                 interactiveDisplayBase_1.14.0
## [7] latticeExtra_0.6-28           GenomeInfoDbData_0.99.0
## [9] Rsamtools_1.28.0              yaml_2.1.14
## [11] RSQLite_1.1-2                 backports_1.0.5
## [13] lattice_0.20-35               biovizBase_1.24.0
## [15] digest_0.6.12                 RColorBrewer_1.1-2
## [17] checkmate_1.8.2               colorspace_1.3-2
## [19] httpuv_1.3.3                  htmltools_0.3.5
## [21] Matrix_1.2-9                  plyr_1.8.4
## [23] XML_3.98-1.6                  biomaRt_2.32.0
## [25] zlibbioc_1.22.0               xtable_1.8-2
## [27] scales_0.4.1                  BiocParallel_1.10.0
## [29] htmlTable_1.9                 tibble_1.3.0
## [31] AnnotationFilter_1.0.0         ggplot2_2.2.1
## [33] SummarizedExperiment_1.6.0     nnet_7.3-12
## [35] Gviz_1.20.0                   lazyeval_0.2.0
## [37] mime_0.5                      survival_2.41-3
## [39] magrittr_1.5                  memoise_1.1.0
## [41] evaluate_0.10                 foreign_0.8-67
## [43] BiocInstaller_1.26.0           tools_3.4.0
## [45] data.table_1.10.4             matrixStats_0.52.2
## [47] stringr_1.2.0                 munsell_0.4.3
## [49] cluster_2.0.6                 DelayedArray_0.2.0
## [51] ensemblDb_2.0.0               compiler_3.4.0
## [53] grid_3.4.0                    RCurl_1.95-4.8
## [55] dichromat_2.0-0               VariantAnnotation_1.22.0
## [57] htmlwidgets_0.8               labeling_0.3
## [59] bitops_1.0-6                  base64enc_0.1-3
## [61] rmarkdown_1.4                 gtable_0.2.0
## [63] DBI_0.6-1                     R6_2.2.0
## [65] GenomicAlignments_1.12.0       gridExtra_2.2.1
## [67] seqLogo_1.42.0                knitr_1.15.1
## [69] Hmisc_4.0-2                   rprojroot_1.2
## [71] ProtGenerics_1.8.0             stringi_1.1.5
## [73] Rcpp_0.12.10                  rpart_4.1-11
## [75] acepack_1.4.1

```