Finding RNA-seq Hot Spots

Patrick Aboyoun

Fred Hutchinson Cancer Research Center

January 29, 2010

Outline

Motivation

Programming Steps

Function List

Data Classes You'll Encounter

Outline

Loading Packages

The *ShortRead* contains alignment I/O capabilities. It will also load the *IRanges* and *BSgenome* packages that we will use.

```
> library(ShortRead)

> head(search())

[1] ".GlobalEnv"
[2] "package:day3"
[3] "package:BSgenome.Scerevisiae.UCSC.sacCer2"
[4] "package:biomaRt"
[5] "package:org.Sc.sgd.db"
[6] "package:RSQLite"
```

Importing the Alignments (redux)

We will reload the Bowtie alignments of RNA-seq data to Saccharomyces cerevisiae genome, UCSC (sacCer2, June 2008).

```
> bowtieFile <- system.file("extdata", "BYe9.head.map",
+                          package="day3")
> aln <- readAligned(bowtieFile, type = "Bowtie")
> aln
```

```
class: AlignedRead
length: 1000000 reads; width: 32 cycles
chromosome: chrmt_S288C chrmt_S288C ... chr12_S288C chr12_S288C
position: 7021 12161 ... 446999 461957
strand: - - ... + -
alignQuality: NumericQuality
alignData varLabels: similar mismatch
```

Renaming the Chromosomes

```
> head(levels(chromosome(aln)), 4)

[1] "chr01_S288C" "chr02_S288C" "chr03_S288C" "chr04_S288C"

> newChrom <- chromosome(aln)
> levels(newChrom) <- sub("^chr0", "chr", levels(newChrom))
> levels(newChrom) <- sub("\\_S288C$", "", levels(newChrom))
> aln <- initialize(aln, chromosome = newChrom)
> head(levels(chromosome(aln)), 4)

[1] "chr1" "chr2" "chr3" "chr4"
```

Creating the Alignment Intervals

```
> alnRanges <- IRanges(position(aln), width=width(aln))
> head(alnRanges)

IRanges of length 6
     start    end width
[1]   7021   7052    32
[2]  12161  12192    32
[3]  94785  94816    32
[4]  59338  59369    32
[5]   6611   6642    32
[6] 129953 129984    32
```

Extending the Alignment Intervals

```
> extAlnStarts <-
+   ifelse(strand(aln) == "+", start(alnRanges), end(alnRanges)
> extAlnRanges <- IRanges(start = extAlnStarts, width = 150)
> head(extAlnRanges)

IRanges of length 6
     start    end width
[1]   6903   7052   150
[2]  12043  12192   150
[3]  94667  94816   150
[4]  59338  59487   150
[5]   6611   6760   150
[6] 129953 130102   150
```

Splitting Alignment Intervals into Groups

```
> chromStrand <- paste(chromosome(aln), strand(aln), sep="")
> extAlnList <- split(extAlnRanges, chromStrand)
> head(names(extAlnList))
```

```
[1] "chr1-"  "chr1+"  "chr10-" "chr10+" "chr11-" "chr11+"
```

Constructing the Intervals for Genes

```
> library(org.Sc.sgd.db)
> geneTable <- cbind(toTable(org.Sc.sgdCHRLOC),
+                    end = toTable(org.Sc.sgdCHRLOCEND)[,"stop"]
> geneTable <- geneTable[abs(geneTable[,"start"]) < abs(geneTabl
> chroms <- paste("chr", geneTable[,"Chromosome"], sep="")
> strand <- ifelse(geneTable[,"start"] > 0, "+", "-")
> geneRanges <- IRanges(start = abs(geneTable[,"start"]),
+                       end   = abs(geneTable[,"end"]))
> geneAnnList <- split(geneRanges, paste(chroms, strand, sep = "
> head(names(geneAnnList))

[1] "chr1-"  "chr1+"  "chr10-" "chr10+" "chr11-" "chr11+"
```

Outline

Session Information

- R version 2.10.1 Patched (2010-01-28 r51060), `x86_64-unknown-linux-gnu`
- Locale: `LC_CTYPE=en_US.UTF-8`, `LC_NUMERIC=C`, `LC_TIME=en_US.UTF-8`, `LC_COLLATE=en_US.UTF-8`, `LC_MONETARY=C`, `LC_MESSAGES=en_US.UTF-8`, `LC_PAPER=en_US.UTF-8`, `LC_NAME=C`, `LC_ADDRESS=C`, `LC_TELEPHONE=C`, `LC_MEASUREMENT=en_US.UTF-8`, `LC_IDENTIFICATION=C`
- Base packages: base, datasets, graphics, grDevices, methods, stats, tools, utils
- Other packages: AnnotationDbi 1.8.1, Biobase 2.6.1, biomaRt 2.2.0, Biostrings 2.14.8, bitops 1.0-4.1, BSgenome 1.14.2, BSgenome.Scerevisiae.UCSC.sacCer2 1.3.16, day3 0.0.2, DBI 0.2-4, IRanges 1.4.9, lattice 0.17-26, org.Sc.sgd.db 2.3.5, RCurl 1.3-0, RSQLite 0.7-3, rtracklayer 1.6.0, ShortRead 1.4.0
- Loaded via a namespace (and not attached): grid 2.10.1, [illegible]