Annotation

Marc Carlson

Fred Hutchinson Cancer Research Center

January 29, 2010

Outline

Annotations for Sequencing Technologies

Annotations for Sequencing projects
Other packages:

- rtracklayer – export to UCSC web browsers.
- GenomicFeatures – coming soon for transcript annotations (will release in spring)

biomaRt:

- Query web-based 'biomart' resource for genes, sequence, and SNPs etc.

AnnotationDbi packages:

- Organism and chip packages – contain chromosome start and stop sites for most genes.

Outline

rtracklayer basics

What rtraklayer offers: rtracklayer

- Web accessible annotations
- Source: The data is from UCSC Genome tracks

finding resources with rtracklayer

How to find data from the UCSC Genome browser in R

- creates a browserSession: `browserSession`.
- list available genomes from UCSC: `ucscGenomes`.
- set up a genome object: `genome`.
- list available tracks: `trackNames`.

```
> library(rtracklayer)
> session <- browserSession()
> head(ucscGenomes())
> genome(session) <- "hg18"
> head(trackNames(session))
```

obtaining resources with rtracklayer

Downloading the UCSC Genome browser data into R

- generate a query for UCSC: `ucscTableQuery`.
- retrieves a UCSC track: `getTable`.

```
> ##can generate a query
> query <- ucscTableQuery(session, "refGene")
> ##which in turn can be used to get the data
> track <- getTable(query)
> head(track)
> colnames(track)
```

packaging chromosome data into a RangedData object

Next we can package this data into a RangedData object

```
> library(IRanges)
> library(BSgenome)
> rdAnn <- RangedData(IRanges(start = track[,"txStart"],
+                             end   = track[,"txEnd"]),
+                     space  = track[,"chrom"],
+                     strand = track[,"strand"],
+                     id     = track[,"name"])
> rdAnn
```

Outline

BiomaRt basics

What biomaRt offers: biomaRt

- Web accessible annotations
- The data is from ensembl

finding resources at biomaRt

BiomaRt has several methods for discovery or resources.

- list available databases: `listMarts`.
- list available datasets: `listDatasets`.
- sets up a DB to be used: `useMart`.

```
> library(biomaRt)
> head(listMarts())
> mart <- useMart("ensembl")
> head(listDatasets(mart))
> ens <- useMart("ensembl", dataset="scerevisiae_gene_ensembl")
> ens
```

extracting data from biomaRt

To call `getBM` you need to to apply appropriate filters and attributes to a list of values that you supply. Attributes are what you want from the query, and filters describe the values you supply.

- list filters from the DB/Dataset: `listFilters`.
- list attributes from that DB/Dataset: `listAttributes`.
- get selected data: `getBM`.

```
> head(listFilters(ens))
> head(listAttributes(ens))
> ## example query
> getBM(attributes=c("ensembl_gene_id","chromosome_name",
+         "strand","start_position","end_position"),
+       filters="entrezgene",
+       values=c(1466398,1466399,1466400), mart=ens)
```

extracting data from biomaRt

Lets now call getBM to get ALL of the data on these fields.

```
> BMres <- getBM(attributes=c("ensembl_gene_id",
+                     "chromosome_name","strand",
+                     "start_position","end_position"), mart=ens)
```

biomaRt exercise

Using what you just learned about biomaRt, try to construct a RangedData Annotation object similar to what we did with rtracklayer.

packaging biomaRt data into a RangedData object

```
> library(IRanges)
> library(BSgenome)
> strand <- strand(ifelse(BMres[,"strand"] > 0, "+", "-"))
> rdAnno <- RangedData(IRanges(
+                          start = abs(BMres[,"start_position"]),
+                          end   = abs(BMres[,"end_position"])),
+                        space   = BMres[,"chromosome_name"],
+                        strand  = strand,
+                        gene_id = BMres[,"ensembl_gene_id"] )
> rdAnno
```

Outline

Using Annotation packages

What Annotation packages offer:

- Pre-built and versioned annotation packages
- The data is from NCBI

extracting chromosome data from Annot packages

First let't just get the data from the package.

```
> library(org.Sc.sgd.db)
> start <- toTable(org.Sc.sgdCHRLOC)
> end <- toTable(org.Sc.sgdCHRLOCEND)
> ##must check that these are the SAME!
> table(start[,1]==end[,1])
> ##If that checks out ok, then we can cbind() them together:
> end <- end[,"stop"]
> res <- cbind(start,end)
> ##filter out autonomously replicating sequences...
> res <- res[abs(res[,"start"]) < abs(res[,"end"]),]
> head(res)
```

Annotation package exercise

Using what you just learned about the annotation packages, try to construct a RangedData Annotation object similar to what we did with biomaRt and rtracklayer.

packaging annotation package data into a RangedData object

```
> library(IRanges)
> library(BSgenome)
> chroms <- paste("chr", res[,"Chromosome"], sep="")
> strand <- strand(ifelse(res[,"start"] > 0, "+", "-"))
> rdAnnot <- RangedData(IRanges(start = abs(res[,"start"]),
+                               end   = abs(res[,"end"])),
+                       space   = chroms,
+                       strand  = strand,
+                       id = res[,"systematic_name"])
> rdAnnot
```

This is the same as the contents of `extractYeastGenesAsRangedData`.