Exercises: Using Bioconductor Annotations

Marc Carlson

29 July, 2011

These exercises will take us through various kinds of practical examples to make sure that you are comfortable using Bioconductor annotations.

Pre-requisite: The *Annotations* package successfully installed and attached.

```
> install.packages("Annotations_1.0.4.tar.gz", repos=NULL,
+                  type="source")
```

Exercise 1
Load the following toy example of a topTable from the Annotations package:

```
>   library(Annotations)
>   load(system.file("data","tt.Rda",package="Annotations"))
>   tt
```

```
          ID      logFC         t      P.Value  adj.P.Val          B
1  100127974  1.6665129  6.236165 0.0001280333 0.01280333  1.4677893
2      10013  1.6159731  4.818758 0.0008459081 0.04229540 -0.4563193
30 100130000  0.9324490  3.564729 0.0056950215 0.15242524 -2.3996197
29 100130001 -0.9453525 -3.522244 0.0060970095 0.15242524 -2.4686231
86 100130002  0.7177127  2.788740 0.0203227814 0.34371907 -3.6739124
50 100130003 -0.9524966 -2.779932 0.0206231440 0.34371907 -3.6884049
98 100130004 -0.7254350 -2.530065 0.0312897608 0.43511296 -4.0974726
9  100130006  0.4959552  2.466192 0.0348090364 0.43511296 -4.2011172
63 100130009 -1.6781056 -2.343702 0.0426907899 0.47434211 -4.3983413
88 100130011  0.6942891  2.198697 0.0543065097 0.50212044 -4.6285126
```

Now find the gene Symbol and pubmed IDs for the top gene. Then use the pubmed ID that turns up to find other genes that were associated with that publication.

Exercise 2
Modify the topTable to include the gene symbols and chromosomes that match with the gene IDs.

Exercise 3
Get the GO terms for the 2nd most relevant gene from the topTable.

Exercise 4
Load the transcriptDb package for TxDb.Hsapiens.UCSC.hg19.knownGene.db And then apply a filter on the chromosomes so that only chromosome 7 is exposed. Finally, extract the transcripts into a GRangesList object grouping by gene.

Exercise 5
The following will load a partial topTable such as you might get from the DEXseq package.

```
>   load(system.file("data",
+                    "ttDEX.Rda",
+                    package="Annotations"))
```

Notice that this table has both entrez gene IDs and exon IDs, use these to find 1) the gene symbols that correspond to the various elements, 2) the ranges for the corresponding transcripts and 3) the ranges for the corresponding exons.

Exercise 6
Read in a gapped alignment using the code below:

```
>   ga <- readGappedAlignments(system.file("extdata",
+                                          "chr7Cont1.bam",
+                                          package="Annotations"))
```

Now take that data, and the annotation data for the transcripts and use `countOverlaps` *to determine how many reads are aligned with the "CFTR" (cystic fibrosis) gene.*

Exercise 7
Begin this exercise by making a FeatureDb from the oreganno table in the oreganno track. Next, change the chromosome filtering on the TranscriptDb that we loaded earlier for hg18 so that it uses the standard set of human chromosomes (chr1:chr22,plus chrM, chrX and chrY). Then use the matchMatrix object produced by findoOverlaps() to determine which oreganno elements overlap with transcripts in the TxDb. Determine which of the genes had the most overlapping elements and then look up the gene symbol for it.