RNA-seq Data

Bernard Pereira

November 26, 2015

Contents

1 Introduction

This is a simple practical that is designed to help you become familiar with RNA-seq data and the ways it can be handled in Bioconductor.

2 The Data

There are six BAM files in the directory Day2/bam. These BAM files were aligned using TopHat2 [2].

```
#Not in R!
ls -lh Day2/bam #What are .bai files?
```

The data used here comes from a study on esophageal squamous cell carcinoma[1], in which three patients (sample ids: 16, 18 and 19) had their tumours sequenced along with matched normal tissue[1]. Once the RNA-seq data was available for the samples, the authors performed differential expression analysis to look for genes whose expression was deregulated in the tumours. The data was downloaded from GEO (GSE29968).

We can use samtools to explore the BAM files from the command line. You can find details about the SAM format at https://samtools.github.io/hts-specs/SAMv1.pdf.

```
#Still not R!
samtools flagstat Day2/bam/16N_aligned.bam
samtools flagstat Day2/bam/16T_aligned.bam


samtools view -H Day2/bam/16N_aligned.bam
#Can you interpret the SAM file header using the SAM format specification?

samtools view  Day2/bam/16N_aligned.bam | head -1 #This the first read in the SAM file
```

We can also have a browse through some of the files in the genome browser IGV. Becoming familiar with the data and what it looks like is an essential step before any analyses are performed. The authors of the original publication identified

PTK6 as a tumour suppressor; can we see downregulation (lower expression = fewer reads) of the gene in the tumours compared to the normal samples?

Although we will not be using this functionality, we can also access the reads in a BAM file using Bioconductor.

```
library(Rsamtools)

#Regions of interest -
#We want to look at reads in these regions
chr1 <- IRanges(0,50000)
ofInterest <- RangesList(chr1=chr1)

#What we want to see
output <- c("rname", "strand", "pos", "qwidth", "seq")

#Get the data!
bam <- scanBam("Day2/bam/16N_aligned.bam", param=ScanBamParam(which=ofInterest, what=output))
bam #How many reads in this region?
```

As you might imagine, we could use features in *GenomicRanges* to find and count overlaps between a set of reads and a set of features. However, there are packages that directly provide functions for counting.

3 Counting reads

We will use the package *Rsubread*[3] to count the reads mapping to genes in the human genome.

```
library(Rsubread)
filesToCount <- dir("Day2/bam", pattern=".bam$", full.names=T)
filesToCount
```

Rsubread has a number of inbuild annotations that we can make use of. In the interests of time, we will explore the featureCounts function using the reads in one BAM files for now (16N).

```
filesToCount[1]
tmp1 <- featureCounts(filesToCount[1], annot.inbuilt="hg19", ignoreDup=F) #Approx. 1 minute
names(tmp1)
tmp1$stat
tmp1$targets
head(tmp1$annotation)
head(tmp1$counts)
dim(tmp1$counts)

?featureCounts  #Lots of parameters to play with!
tmp2 <- featureCounts(filesToCount[1], annot.inbuilt="hg19", ignoreDup=F, minReadOverlap=50)
tmp2$stat

tmp3 <- featureCounts(filesToCount[1], annot.inbuilt="hg19", ignoreDup=F,
countMultiMappingReads=T)
tmp3$stat
```

We will now count the reads in all BAM files using *Rsubread*, and use this output for our downstream analyses. As you have seen, there are a number of options to play around with; for now, however, we will keep things simple and generate a count matrix to use in the practical on differential expression.

```
#Default parameters
tmp <- featureCounts(filesToCount, annot.inbuilt="hg19", ignoreDup=F,
countMultiMappingReads=F, minReadOverlap=1)
save(tmp, file="../Day2/countMatrix.RData")
```

References

[1] Ma, S. et al. (2012) *Identification of PTK6, via RNA Sequencing Analysis, as a Suppressor of Esophageal Squamous Cell Carcinoma*, Gastroenterology, 143 (3) 675-686.

[2] Kim, D. et al. (2013) *TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions*, Genome Biology, 14:R36.

[3] Liao, Y., Smyth, GK. & Shi, W. (2013) *The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote*, Nucleic Acids Res. 41, e108.

[4] Love, MI.et al. (2014) *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*, Genome Biology, 15(12) 550.