Lab 4: Analyze second generation sequencing data

The purpose of this lab is to get students familiar with the typical workflow of analyzing second-generation sequencing data including sequence alignment, basic manipulation and visualization of alignment files, SNP calling, and read count summarization.

Before the lab:

Students should preinstall following software: bowtie, and Bioconductor package Rsamtools, DESeq, edgeR, DSS and R package ROCR. Optionally students can install samtools and BEDtools, which require compilation from source codes. The compilation is straightforward on Linux system, but requires additional software on Mac and Windows. Specifically one needs to install Xcode on Mac or Cygwin on Windows in order to compile C++ codes. If this turns out to be too difficult it can be skipped. The instructor can provide binary executable for Mac OSX 10.6 (Snow Leopard).

Data:

We will use following three datasets to demonstrate the usage of some useful software. The data can be obtained from the class website as one zipped file.

1. Reference genome and a sequence read file (with 10000 reads) for a type of bacteriophage. We will use this to practice sequence alignment using bowtie. The small sizes of the reference genome (only 5386 bps) and raw sequence read file make the computation fast.
2. RNA-seq and ChIP-seq for Cmyc binding data for K562 cell line. The aligned read files (in BED format) were obtained from ENCODE project. The files were processed so that only reads for chromosome 22 were kept. The files are then converted to BAM format. We will use these files to practice using Rsamtools for summarizing read counts.
3. Simulated RNA-seq count data for two group comparison with 3 replicates in each group. We will use this to practice using Bioconductor packages for differential expression analysis for RNA-seq data.

In the lab:

I. Sequence alignment using the phage data. In this practice we will align the reads to the reference genome. Following below steps:

1. Create index files based on reference genome: `bowtie-build phage-ref.fa phage`

2. Align the reads allowing 3 mismatches, and output result in a file called reads.sam in SAM format: `bowtie -v 3 -f -S phage reads.fa reads.sam`

II. Using samtools to convert the alignment file into BAM format then visualize and call SNPs. [***optional if you have samtools installed***.]

1. First convert the alignment to BAM format, then index and sort BAM file (this is necessary for viewing and SNP calling).

```
samtools view -bS reads.sam > reads.bam
samtools sort reads.bam reads.sorted
samtools index reads.sorted.bam
```

2. View the alignment in text viewer:

```
samtools tview reads.sorted.bam phage-ref.fa
```

Press space bar or left/right arrows to move forward/backward. Or one can type "/" and input a genomic location such as "phage:2750". There are two SNPs can be seen at 2793 and 3133 bps.

3. SNP calling. Samtools provide SNP calling functionalities with "bcftools". But the BAM file must be first "pileup"ed, meaning the reads need to be summarized at all basepair positions. Run following commands to call SNP:

```
samtools mpileup -uf phage-ref.fa reads.sorted.bam > reads.pileup
bcftools view reads.pileup > SNP.vcf
```

III. **Using Rsamtools to handle bam files and perform analyses**.

The first step in RNA-seq or ChIP-seq analysis is to count the number of reads in genomic windows. The windows are usually genes or exons in RNA-seq, and equal sized intervals in the whole genome in ChIP-seq.

In this lab we will practice achieving such tasks using Rsamtools in combination with other Bioconductor packages we've learned: GenomicRanges and GenomicFeatures. The data were RNA-seq and c-MYC binding ChIP-seq data for K562 cell line. For the sake of easier computation only reads for chromosome 22 were included in the data.

It was known that c-MYC binding is correlated with gene expression. We will obtain number of reads in the gene body from the RNA-seq data as gene expression levels. Number of reads in the gene promoters will be obtained from the ChIP-seq data for c-MYC binding strength. We'll show these two values are strongly correlated.

Additionally, we will practice to obtain the read counts in equal sized bins around the transcription starting sites (TSS) then look at the average counts from all genes. This is sometimes referred to as “meta-gene profile” of the binding, which shows the average shape of the “peaks”. More example of these can be seen at www.factorbooks.org, for example, http://www.factorbook.org/mediawiki/index.php/MAX.

Follow the R code (sgseq-lab.R) on the website to perform above analysis.

IV. Use Bioconductor to detect DE genes from RNA-seq data.

Follow the R code (sgseq-lab.R) on the website to do DE analysis for RNA-seq data. Compare results from DESeq, edgeR and DSS.