**ENAR 2014 NGS short course computer lab**

The purpose of this lab is to get students familiar with typical workflow of analyzing next-generation sequencing data, in particular ChIP-seq and BS-seq. We will cover basics of

* Sequence alignment.
* Manipulation of alignment result files.
* ChIP-seq: peak calling and peak comparison.
* BS-seq: smoothing and DMR detection.
* Several useful Bioconductor packages for genomic data analyses.

**Before lab:**

Students should preinstall following software. T**his is very important, because we won’t have time to install these in class!!!**.

* bowtie (http://bowtie-bio.sourceforge.net/index.shtml).
* Bioconductor package Rsamtools, bsseq, DSS, GenomicRanges, GenomicFeatures. Run following commands in R to install.

source("http://bioconductor.org/biocLite.R")

biocLite()

biocLite("Rsamtools")

biocLite("bsseq")

biocLite("GenomicRanges")

biocLite("GenomicFeatures")

* MACS (<http://liulab.dfci.harvard.edu/MACS/>).
* samtools (<http://samtools.sourceforge.net>).
* Integrated genome viewer (<http://www.broadinstitute.org/igv/home>).

Note that many of the software run in command line. Under Windows system, you might encounter difficulties, and need to install additional software (shell, perl, python, gcc, etc.). Please look for help from your local system administrator.

**It is fine if you have trouble install some of the software, we will provide all lab exercise results. So if you cannot perform one exercise, the next one won’t be affected.**

**Data**:

We will use following datasets in the lab.

1. Reference genome and a sequence read file (with 10000 reads) for a type of *bacteriophage*. This will be used for practicing sequence alignment using bowtie. The small sizes of the reference genome (only 5386 bps) and raw sequence read file make the computation fast.
2. ChIP-seq data for Cmyc and Pol2 binding, and control for K562 cell line. The aligned read files (in BED format) were obtained from ENCODE project. The files were processed to keep only reads aligned to chromosome 22, and then converted to BAM format. We will use these files to practice peak calling.
3. A set of BS-seq data. The data are for caner vs. normal comparison. There are three replicates in each group. Data from a small part of genome on chromosome 21 is extracted for computational efficiency.

All data can be downloaded from xxx as a zip file.

**Lab practice:**

**I. Sequence alignment using the phage data**.

In this practice we will align the reads to the reference genome. The necessary data are stored in “alignment” folder.

***1. Create index files based on reference genome:***

bowtie-build phage-ref.fa phage

After this step, 6 .ebwt files will be created. These are the index files for the phage genome. Note that the index files for many species have been pre-built, and available to download from bowtie’s website.

***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

After this step, bowtie print alignment result statistics on the screen. A file called reads.sam is created. Open reads.sam to take a brief look.

***3. Convert the alignment to BAM format, then index and sort BAM file.***

samtools view -bS reads.sam > reads.bam

samtools sort reads.bam reads.sorted

samtools index reads.sorted.bam

Note that it is desirable to save the alignment results in BAM format, because the bam files are smaller and faster to process.

***4. View the alignment in samtools text viewer:***

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

Press space bar or arrow keys to move forward/backward. Or one can type “/” and input a genomic location such as “phage:2750”. Two SNPs can be visually detected at 2793 and 3133 bps.

**II. ChIP-seq peak calling.**

The necessary data are stored in “ChIP-seq” folder.

***1. Use MACS to call peaks from Cmy and Pol2 ChIP-seq data.***

Go to the correct directory, and run following in command window:

macs14 -t K562Cmyc\_chr22.bam -c K562\_chr22\_Input.bam -n K562\_Cmyc

macs14 -t K562Pol2\_chr22.bam -c K562\_chr22\_Input.bam -n K562\_Pol2

This will generate some .xls and .bed files. K562\_Cmyc\_peaks.bed and K562\_Pol2\_peaks.bed are peak lists in bed format.

***2. Visualize the data using IGV.***

IGV takes .tdf files for reads counts. First covert the bam files into tdf with following steps:

1. Select reference genome as “Human hg18”.
2. Select “Run igvtools” form “Tools” menu.
3. In command, select “Count”. Then select Input file. Keep other options as default, and click “Run”.

After conversion, you’ll get two files “K562Cmyc\_chr22.bam.tdf” and “K562Pol2\_chr22.bam.tdf”. Now click “File”-> “Load from File”, and select the two tdf files, then switch to chr22. The coverage will be shown as bar plots. Try zooming and moving the viewer window to see some peaks. Genomic region can be input to the text box at top. For example, try put in “chr22:18,216,605-18,229,012” to see a peak.

***3. Comparison of peaks.***

The overlaps of different lists of peaks can be easily computed using R/Bioconductor packages. The ***GenomicRanges*** package is designed to operate on lists of genomic intervals.

To look at the overlaps of Pol2 and Cmyc binding, run following codes in R.

library(GenomicRanges)

## read in peak calling results

peaks.pol2 = read.table("K562\_Pol2\_peaks.bed")

peaks.cmyc = read.table("K562\_Cmyc\_peaks.bed")

## look at the first a few peaks

head(peaks.pol2)

head(peaks.cmyc)

## create "GRanges" objects for the peaks

peaks.pol2.gr = GRanges(seqnames=Rle(peaks.pol2[,1]),

ranges=IRanges(peaks.pol2[,2], peaks.pol2[,3]))

peaks.cmyc.gr = GRanges(seqnames=Rle(peaks.cmyc[,1]),

ranges=IRanges(peaks.cmyc[,2], peaks.cmyc[,3]))

## look at the GRanges objects

head(peaks.pol2.gr)

head(peaks.cmyc.gr)

## look at the overlaps - overlap percentage is high

mean(peaks.pol2.gr %over% peaks.cmyc.gr)

mean(peaks.cmyc.gr %over% peaks.pol2.gr)

***4. Explore overlaps of peaks and gene transcriptional start sites***

Sometimes it is of interest to explore whether binding sites significantly overlap certain genomic features, such as ***transcriptional start sites (TSS)*** of genes. The ***GenomicFeatures*** package provide functionalities to access gene annotations. With that, the overlaps can be calculated in similar way.

For the lab practice, the refseq gene annotation for human hg18 genome has been obtained to save students time. Following R codes below to create the TSS for human refseq genes on chromosome 22, and then explore their overlaps with Pol2 and Cmyc binding.

library(GenomicFeatures)

## load in saved hg18 database

refGene.hg18 = loadDb("refGene.hg18.sqlite")

tx = transcripts(refGene.hg18)

tx.chr22 = tx[seqnames(tx)=="chr22"] ## genes on chr 22

tx.chr22 ## 925 transcripts on chr22

## construct regions around TSS (+/- 1000 bp).

## This is a little tricky, be careful in strands directions.

strands=strand(tx.chr22)

ix.gene.minus=which(strands=="-")

TSS=start(tx.chr22)

TSS[ix.gene.minus]=end(tx.chr22)[ix.gene.minus]

ext=1000

TSS.GRanges=GRanges(seqnames=Rle("chr22", length(tx.chr22)),

ranges=IRanges(start=TSS-ext, end=TSS+ext))

## follow lines compute percent of TSS covered by binding peaks.

mean(TSS.GRanges %over% peaks.pol2.gr)

mean(TSS.GRanges %over% peaks.cmyc.gr)

**III. Analysis of BS-seq data**

Follow the R script below to perform basic smoothing and DMR detection using bsseq Bioconductor package.

library(bsseq)

dat = read.table("BSseq.txt", header=TRUE)

head(dat)

dim(dat)

## make BSseq data object

dat.bsseq = BSseq(chr=dat$chr, pos=dat$pos, Cov=as.matrix(dat[,3:8]),

M=as.matrix(dat[,9:14]), sampleNames=c("C1","C2","C3","N1","N2","N3"))

dat.bsseq

## smoothing - this will take a couple minutes

dat.bsseq.smooth = BSmooth(dat.bsseq)

## DMR detection

tests = BSmooth.tstat(dat.bsseq.smooth, c("C1","C2", "C3"),

c("N1","N2", "N3"), estimate.var="same")

dmrs = dmrFinder(tests, cutoff=c(-5, 5))

dmrs

## visualize a DMR

range = GRanges(seqnames=dmrs$chr[1],

IRanges(dmrs$start[1], dmrs$end[1]))

plotRegion(dat.bsseq.smooth,region=range, extend=5000)