Regulatory Genomics: ChIP-seq course plan

Shamith Samarajiiwa, Dora Bihary

September 2017

Contents

1 Gene Regulatory Interactomics: Transcription Factor & Epigenomic ChIP-seq, ATAC-seq data analysis

Pre-requisites:

- Completed a R programming and Unix course.
- Experience with R/Bioconductor packages such as tidyr, dplyr, ggplot2, biomaRt and GenomicRanges

Lecturers: Dr. Shamith Samarajiwa (ss861@mrc-cu.cam.ac.uk) Dr. Dora Bihary (db679@mrc-cu.cam.ac.uk)

2 Day 1: Data processing for Next Generation Sequencing

2.1 L1: Brief introduction to file formats (2.30-3.00pm)

2.2 L2: Quality control and artefact removal(3.00-3.30pm)

- Use of `FastQC` and `Cutadapt`

```
Practical 1: learn to use FastQC and Cutadapt (30 min) on a sample dataset
```

2.3 L3: Short read alignment to a reference genome (3.30-5.00pm)

- Understand differences between reference genome builds
- Introduction to Illumina sequencing
- Short read alignment with `BWA`, `Bowtie2` and `STAR`
- Sequencing Coverage and Depth
- Mappability
- Use of decoy and sponge databases
- Alignment quality
- Use of `Samtools`, `Picard tools`, `Samstat` and `Qualimap`
- Visualization using `IGV` and `Tablet`
- Other technologies

```
Practical 2: Alignment of a ChIP-seq dataset to a reference genome using BWA, Bowtie2
and a RNA-seq dataset to STAR (30 min)
```

3 Day 3 ChIP-seq and ATAC-seq Analysis

3.1 LUNCH (12.30-1.30pm)

3.2 L4: Introduction to ChIP-seq (1.30-2.00pm)

- Brief intorduction to ChIP-seq
- Statistical aspects
- Sequencing depth
- Why we need controls
- Artefact removal
- Peak calling

3.3 L5: Peak Calling (2.00-3.30pm)

- Narrow vs. Broad peaks
- IDR and Dealing with replicates (generating high confidence peak sets)
- Statistical and Practical aspects of peak calling (`MACS2`)
- Understanding the differences between Transcription Factor ChIP-seq and Epigenomic ChIP-seq
- Identifying broad and narrow peaks.

```
Practical 3: peak calling using MACS2 (30 min)
Practical 4: View BAM and Bed files on IGV, use IGV tools (15 min)
```

3.4 L6: Quality control methods for ChIP-seq (3.30-5.00pm)

- Blacklists and Graylists
- Use of `ChIPQC` to understand and interpret different QC methods and metrics

```
Practical 5: ChIPQC package (30 min)
```

4 Day 4

4.1 L7: Useful software for analysis of genomic data (9.30-11.00pm)

- Genomic coordinate systems
- `SRAtoolkit`
- `bedtools`
- `UCSC utilities`
- `GenomicRanges`
- `USCS table browser`
- Visualizing ChIPseq data with `ChIPseeker` and `rtracklayer`
- `IGV` genome browser
- Working with and manipulating peaks (`Bedtools`, `ChIPseeker`)

```
Practical 6: Using Bedtools, GenomicRanges and ChIPseeker for peak analysis (1 hr)
```

4.2 L8: Downstream analysis of ChIP-seq and ATAC-seq data (11.00-12.30 noon)

- Normalization and Visualization
- Peak annotation (`ChIPpeakAnno`)
- Feature distribution of peaks
- Functional Enrichment - ontology (`GREAT` and `rGREAT`) and gene-set enrichment (`ChIPEnrich`)
- Use peak summits to get Fasta sequence (`Bedtools`)
- Motif detection and motif enrichment analysis (`MEME Suite` and `PscanChIP`)
- Using Postion Weight Matrix databases (`Jaspar` and others)

4.3 LUNCH (12.00-1.00pm)

4.4 L8: Downstream analysis of ChIP-seq data contd. (1.00-1.30pm)

```
Practical 7: ChIPpeakAnno, GREAT, ChIPEnrich, MEMEChIP

Learn to use:
ChIPpeakAnno and biomaRt for annotation and sequence extraction
MEME Suite and MEMEChIP for motif identification and sequence extraction
GREAT for ontology enrichment
ChIPEnrich for gene set enrichment
```

4.5 L9: Identifying direct targets of Transcription Factors (1.30-2.30pm)

- Identifying direct targets of TFs by integrating ChIP-seq and RNA-seq using Rcade or Beta
- How to use TF direct targets to reverse engineer regulatory networks
- Network topology analysis

```
Practical 9: Rcade (30 min)
```

4.6 L10: Differential binding analysis (2.30-3.30pm)

- Identifying differential peaks / binding (`DiffBind` and `Thor`)

```
Practical 10: DiffBind (30 min)
```

4.7 L11: Introduction to Epigenomics and Chromatin Interactions (3.30-4.30pm)

- Introduction to Histone modifications
- Chromatin Segmentation Analysis
- Introduction to Hi-C analysis
- Hi-C data processing workflow
- Generating normalized contact matrices
- A/B compartments, TADs
- `DiffHiC`