Course Plan: Next Generation Sequencing, ChIP-seq, ATAC-seq and Epigenomics

Shamith Samarajiiwa, Dora Bihary

July 2018

Contents

Day 1: Data processing for Next Generation Sequencing

Lecture 1: Introduction to next generation sequencing (2.30-2.45pm)
Lecture 2: Brief introduction to file formats (2.45- 3.00pm)
Lecture 3: Quality control and artefact removal(3.00-3.45pm)
Practical 1: learn to use FastQC and Cutadapt (20 min) on a sample dataset
Lecture 4: Short read alignment and Quality Control (3.45-5.00pm)
Practical 2: Alignment of a ChIP-seq dataset to a reference genome using `BWA` OR `Bowtie2` and a RNA-seq dataset to `STAR` (45 min)

Day 4 ChIP-seq data analysis

Lecture 5: Introduction to ChIP-seq (9.30-10.00pm)
Lecture 6: Peak Calling (10.00-11.00pm)
Practical 3: Peak calling using MACS2 (30 min)
Lecture 7: Differential binding analysis (11.00-12.30pm)
Practical 4: THOR (and Diffbind) (20 min)
Lecture 8: Quality control methods for ChIP-seq (1 hr)
Practical 5: ChIPQC package (30 min)
Practical 6: Integrative Genome Viewer(30 min)
LUNCH (12.30-1.30pm)
Lecture 9: Downstream analysis of ChIP-seq (1.30-3.15pm)
Practical 7: Downstream analysis of ChIP-seq (30 min)
Practical 8: Identifying direct targets of transcription factors with Rcade (30 min)
Lecture 10:: Useful software utilities for the analysis of genomic data (4.30-5.00pm)

Day 5 ATAC-seq and Epigenomics

Practical 9: Useful software utilities for the analysis of genomic data (9.30-10.30am)
Lecture 11 ATAC-seq data analysis (10.30-11.30am)
Practical 10: ATAC-seq analysis (30 min)
Lecture 12 Introduction to Epigenomics and Chromatin Interactions (11.30-12.30)
LUNCH (12.30-1.30pm)

Pre-requisites:

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

Trainers:

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

Day 1: Data processing for Next Generation Sequencing

Lecture 1: Introduction to next generation sequencing (2.30-2.45pm)

- Understand differences between reference genome builds
- Introduction to Illumina sequencing technology

Lecture 2: Brief introduction to file formats (2.45- 3.00pm)

Lecture 3: Quality control and artefact removal(3.00-3.45pm)

- Use of `FastQC`, `Cutadapt` or `TrimGalore`, `Trimmomatic`, `Fastx toolkit`

Practical 1: learn to use FastQC and Cutadapt (20 min) on a sample dataset

Lecture 4: Short read alignment and Quality Control (3.45-5.00pm)

- Short read aligners
- Alignment with `BWA`, `Bowtie2` and `STAR`
- 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`
- De novo assembly, Genome graphs, Long Read technologies

Practical 2: Alignment of a ChIP-seq dataset to a reference genome using `BWA` OR `Bowtie2` and a RNA-seq dataset to `STAR` (45 min)

Day 4 ChIP-seq data analysis

Lecture 5: Introduction to ChIP-seq (9.30-10.00pm)

Lecture 6: Peak Calling (10.00-11.00pm)

- 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)

Lecture 7: Differential binding analysis (11.00-12.30pm)

Practical 4: THOR (and Diffbind) (20 min)

Lecture 8: Quality control methods for ChIP-seq (1 hr)

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

Practical 5: ChIPQC package (30 min)

Practical 6: Integrative Genome Viewer(30 min)

LUNCH (12.30-1.30pm)

Lecture 9: Downstream analysis of ChIP-seq (1.30-3.15pm)

- Normalization and Visualization
- Peal 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)
- Differential Binding (Diffbind , Thor)
- Direct target identification (Rcade) and Network Biology applications

Practical 7: Downstream analysis of ChIP-seq (30 min)

- Extract peak sequences
- Peak annotation
- Motif identification and enrichment analysis
- Ontology enrichment analysis

Practical 8: Identifying direct targets of transcription factors with Rcade (30 min)

- Integrating Gene Expression with TF binding

Lecture 10:: Useful software utilities for the analysis of genomic data (4.30-5.00pm)

- 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)
- Deeptools2

Day 5 ATAC-seq and Epigenomics

Practical 9: Useful software utilities for the analysis of genomic data (9.30-10.30am)

Lecture 11 ATAC-seq data analysis (10.30-11.30am)

Practical 10: ATAC-seq analysis (30 min)

Lecture 12 Introduction to Epigenomics and Chromatin Interactions (11.30-12.30)

LUNCH (12.30-1.30pm)