ATAC-seq Data Processing

Library Prep → Sequencing → Reads Separated by Barcode

size limitations can chop reads of same barcode

Combine split reads → Raw Read File .fastq.gz file → Check quality with **FastQC**

This is also the step where you can remove specific types of reads

Align reads to *genome* Using **Bowtie2**

Genome is basically a text file (.fasta) that has a ">" listed chromosomes and the sequence found in each one. Masked files have repeats as NNNN. Important things to note are if there are any scaffolds (genome building could not place these reads in a chromosome correctly) and how the chromosomes are designated (can be Chr1 chr1 or just 1, or some weird way like SLchrPenn1…)

Aligned .SAM file

Convert to binary format using **Samtools**

A BAM file is a binary version of a SAM file. Information is the same but in a different format. It is a MUCH smaller file size. You can delete the SAM file after converted to a BAM file. The type of information in a BAM (Binary Alignment Map) file includes:
ReadID, where it aligned, the sequence, quality of read, and more

Aligned .BAM file

Alignment Processing using **Samtools**

Processing includes sorting the reads by chromosome (1,2,3,…), indexing (creates an extra file that has info about chromosome sizes), and quality filtering (removes reads found more than twice in the genome (basically)

Sorted, Index, Quality Filtered .BAM file

Organellar Read Removal
also **Samtools**

Remove Organellar reads (Chloroplast/Mitochondria)
Done by specifying which reads to keep (chromosome ones)

Sorted, Index, Quality Filtered Organellar Read-Free .BAM file

→ PCA
Peak Calling
Visualization
Other Comparisons

Peak Calling

Sorted, Index, Quality Filtered
Organellar Read-Free
.BAM file

Create Tag Directory using **HOMER**

Not much here, it fills a folder with tagged information from your BAM file that will make it possible for HOMER to call peaks. If lots of scaffolds are in the BAM file you may have use the "-single" option

Tag Directory

Call peaks using **HOMER**

There are many options for calling peaks. For ATAC-seq we include the "-minDist 150 -regionRes1 -region" options

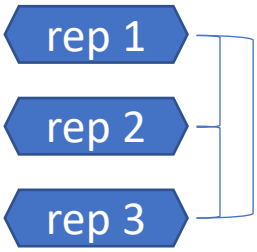
Strange file

Convert to bed file

This will pull out the coordinates, sort, and merge them

.bed File

Find Replicates using **bedtools**



Set criteria to where they have to replicate by at least 50%, which is around 150 bp

Replicated Peaks

Annotation
Comparisons to Other Peaks
Sequence Enrichment
Visualization coordinates
Locations for Read Enrichment

Visualization

Sorted, Index, Quality Filtered
Organellar Read-Free
.BAM file

Convert to bigwig file
using **bamCoverage**

bigwig files are even smaller than BAM files. You lose information about reads, but you keep the important info except it is present as read amount at each single base pair (assuming that is the resolution you set)

.bw file

Compute covergae
using **DEEPTOOLS**

Create a matrix file containing information of read amounts at specific coordinates you provide in the bed file that is analyzed

Matrix .gz file

Will either be plotting peaks (reference-point) or anchored coordinates (scale-regions). Bed file determines and computematrix specifies which one

Plot profile or heatmap plots

Alternatively
Do the same visualization
but with an APP-like interface
by using **SeqPlots**

Plot profile or heatmap plots