Day 9: ChIP-seq, MACS and BEDTools

Instructors: Jessica Westfall and Lynn Sanford

Recap of the videos

1. ChIP-seq introduction
2. Evaluating ChIP-seq data
3. Peak calling with MACS
4. MEME Suite introduction
5. BEDTools introduction
6. ATAC-seq overview

Learning Objectives

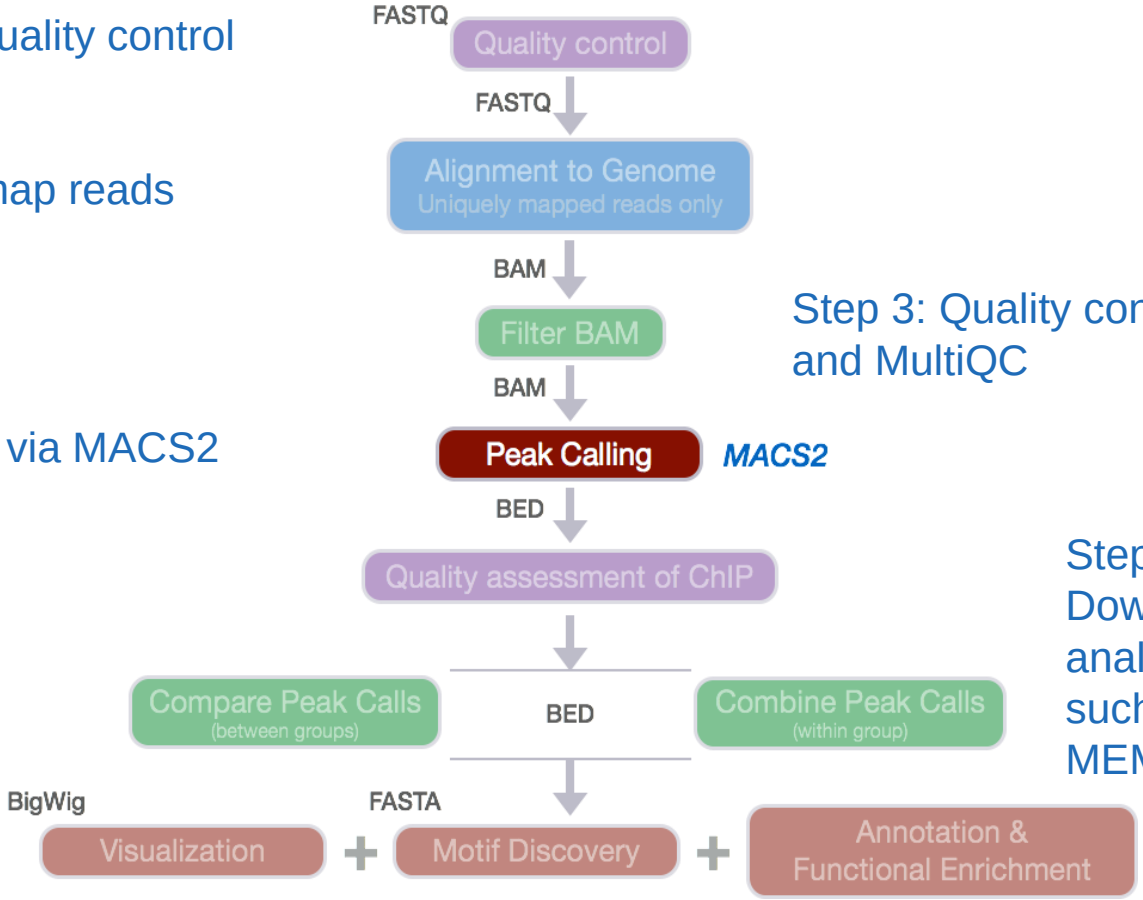
Downstream analysis of ChIP-seq and ATAC-seq data

- Demonstrate the use of a **peak calling program MACS2** to identify genomic regions with robust signal in each of these data types
  - control/input
  - ENCODE Blacklist
- **Visualize** the raw data and corresponding called peaks
- **Downstream analyses**
  - Comparing peaks to other features (e.g genes) : using BEDTools
  - Motif discovery (MEME)

Peak calling pipeline

Step 1: fastQC for quality control

Step 2: HISAT2 to map reads



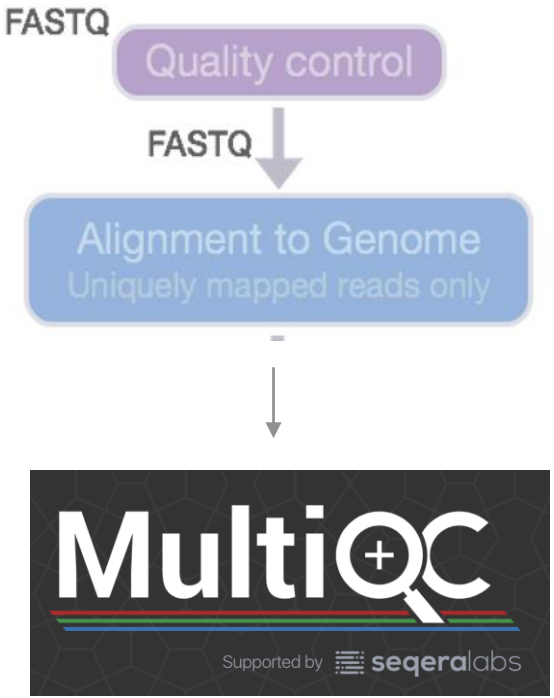
Step 3: Quality control via Preseq and MultiQC

Step 4: Peak calling via MACS2

Step 5 and beyond: Downstream analysis using tools such as BEDTools, MEME,

https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html

fastq > alignment > SAM/BAM conversion



Step 1: fastQC for quality control

Step 2: HISAT2 to map reads
**note that splicing is not relevant for mapping these DNA based methods

Step 3: Quality control via Preseq and MultiQC
multiQC will look through the files and directories that contain compatible results/reports

Quality control

fastQC

Sequence reads quality
- base sequence quality
- GC content
- sequence length and duplication
- overrepresented sequences

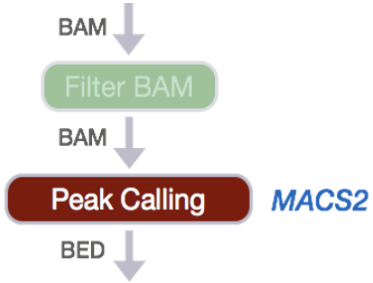
HISAT report

- Alignment rate of reads

Preseq

- Estimating complexity of sequencing library

Peak calling



Step 4: Peak calling via MACS2

ChIP-seq peak calling for enrichment

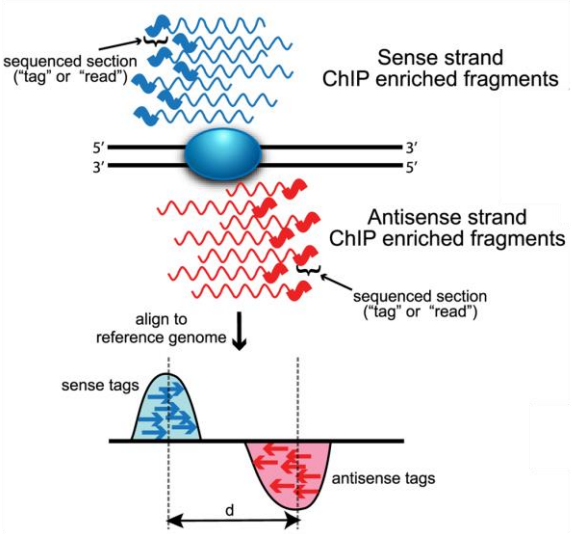
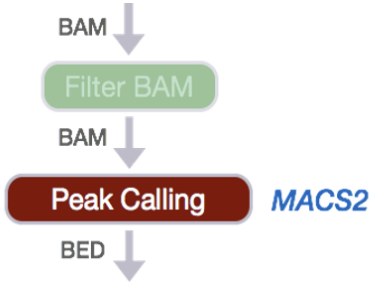


Image source: Wilbanks and Facciotti, PLoS One 2010

ChIP-seq identifies two type of enrichment

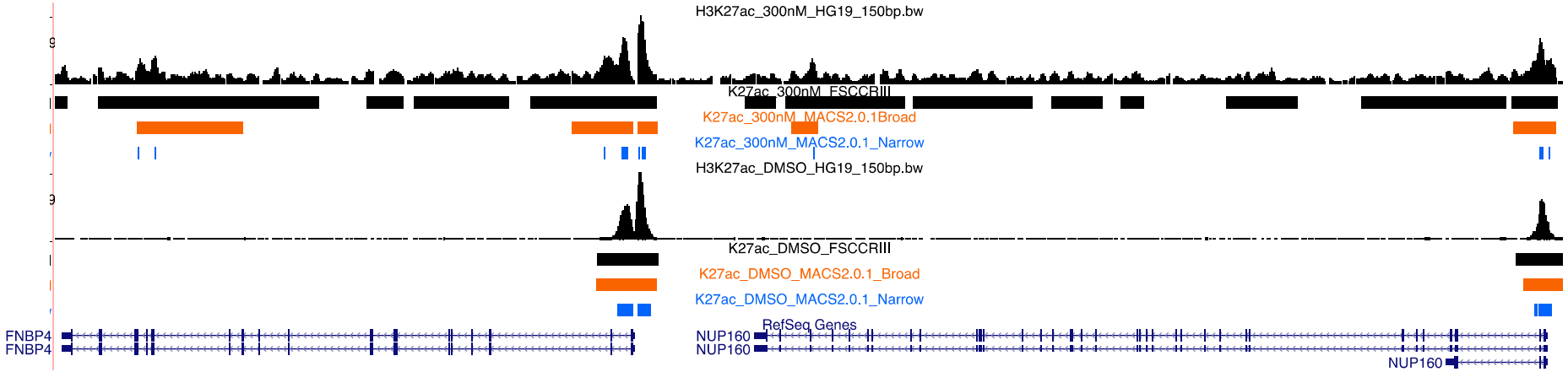
- **Broad peaks:** eg., histone modification. Here we are looking for broad peaks that cover entire gene bodies
- **Narrow peak**: eg., transcription factor binding. Here we are looking for regions of higher amplitude compared to background

Peak calling



Step 4: Peak calling via MACS2

MACS2 Broad calls
MACS2 Narrow calls
FStitch

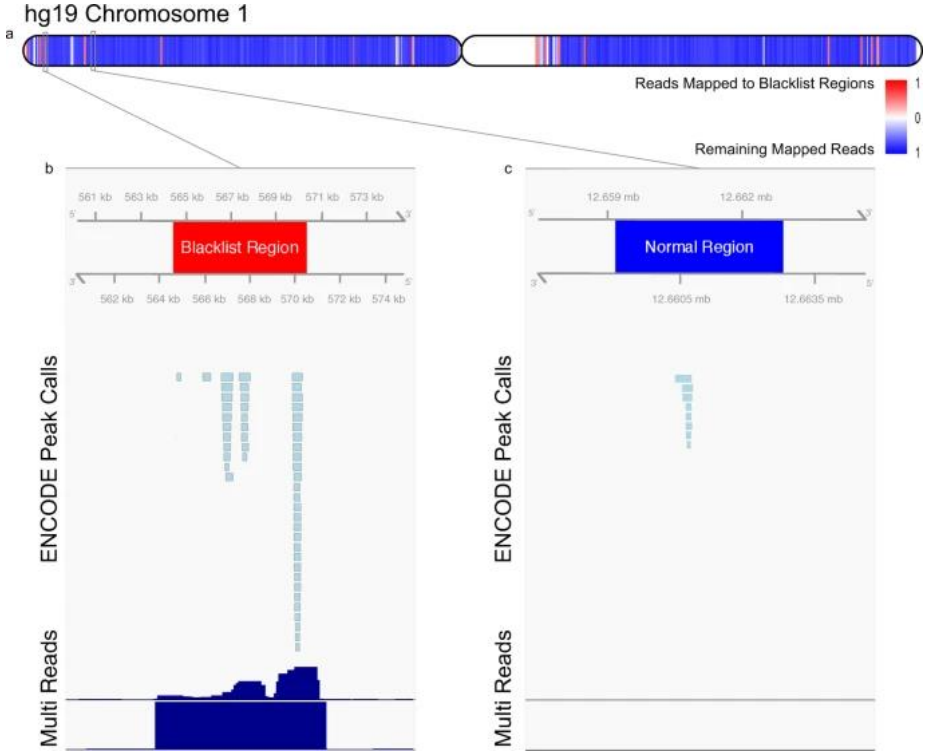


MACS genomic input/control

Controls are important!

- ChIP-seq and ATAC-seq are protocols that produce **background noise** as well as **meaningful signal**
  - Therefore, you need controls to not call background noise as peaks
- p/q value cutoffs matter and should vary based on your experiment
- Know your data type: your experiment should inform the parameters of the peak caller
- **Blacklist regions**: some genomic regions almost always show up in these protocols so remove these regions using a Blacklist

Blacklist regions should be removed



https://www.nature.com/articles/s41598-019-45839-z

These regions contain repetitive regions across the genome and almost always are enriched in ChIP-seq data.

MACS output

1. chromosome
2. start coordinate
3. end coordinate
4. name
5. score
6. strand

Standard BED file fields

7. **signalValue - Measurement of overall enrichment for the region**
8. **pValue - Statistical significance (-log10)**
9. **qValue - Statistical significance using false discovery rate (-log10)**
10. **peak - Point-source called for this peak; 0-based offset from chromStart**

narrowPeak specific fields

Image: https://hbctraining.github.io/Intro-to-ChIPseq/lessons/05_peak_calling_macs.html

Annotation files

- Files of chromosomal coordinates

Chromosome	Start coordinate	End coordinate	More metadata
chr1	20000000	20005000	…
chr5	4050000	4100000	…

- Many different file formats
  - BED, GTF, bedGraph, SAF, VCF, etc.
  - Each of these is just a text file with standardized columns
  - Coordinates can be 0- or 1-indexed, closed or open or half-open, depending on format
  - File format best friend: https://genome.ucsc.edu/FAQ/FAQformat.html

Bedtools

- Made to manipulate BED files – very useful file format
  - Usually fairly small
  - Can easily add columns if you need them

- Many different commands in one package
  - https://bedtools.readthedocs.io/en/latest/content/bedtools-suite.html
  - Some are more complicated, but many of the common commands are basic math/logic/set operations oriented toward genomic data

- After you manipulate annotation files, look at them by eye!

Additional Resources

Other Peak Callers:

- Fstitch: https://github.com/Dowell-Lab/FStitch
- SICER: https://zanglab.github.io/SICER2/
- PeakSeq: https://www.nature.com/articles/nbt.1518
- Hpeak: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-11-369
- PeakRanger: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-12-139

BEDTools Documentation https://bedtools.readthedocs.io/en/latest/
BEDTools tutorial: http://quinlanlab.org/tutorials/bedtools/bedtools.html