Day 9: ChIP-seq, MACS and BEDTools

Rutendo Sigauke and Ariel Eraso

Recap of the videos

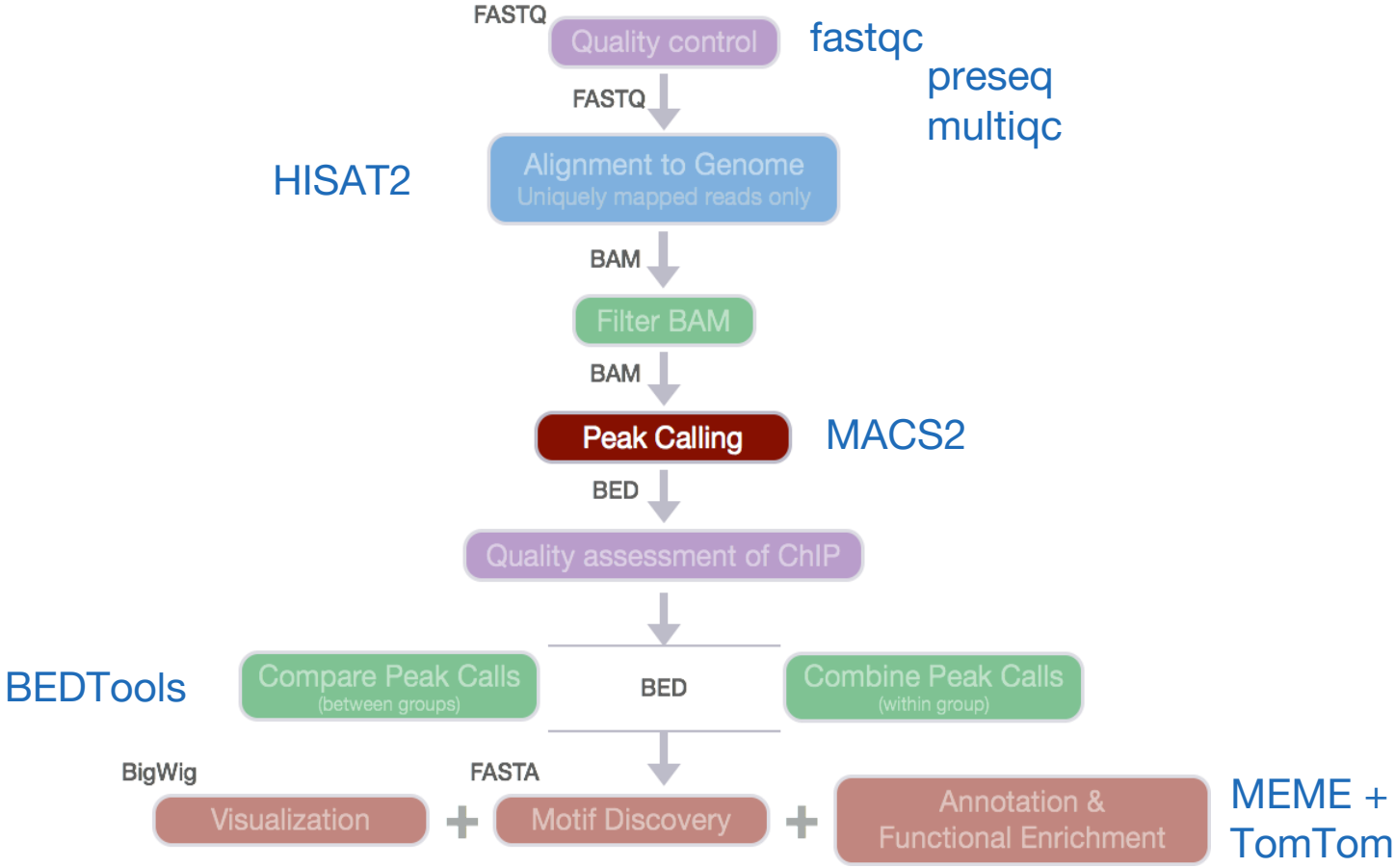
1. ChIP-seq introduction
2. Evaluating ChIP-seq data
3. Peak calling with MACS
4. MEME Suite introduction
5. BEDTools introduction
6. ATAC-seq overview

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



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

ChIP-seq peak calling for enrichment

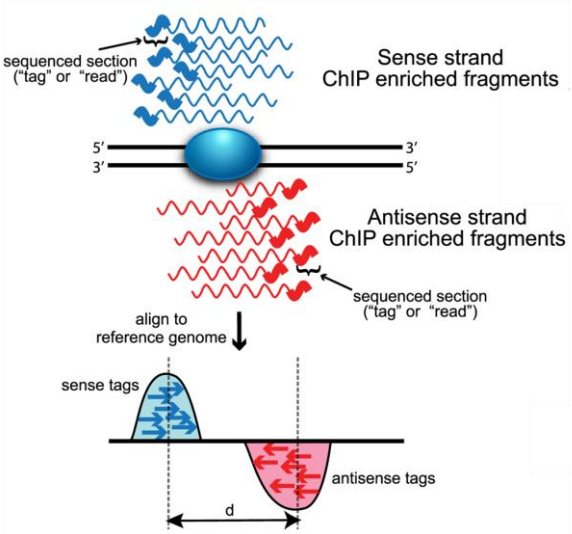


Image source: Wilbanks and Faccioti, PLoS One 2010

ChIP-seq identifies two type of enrichment

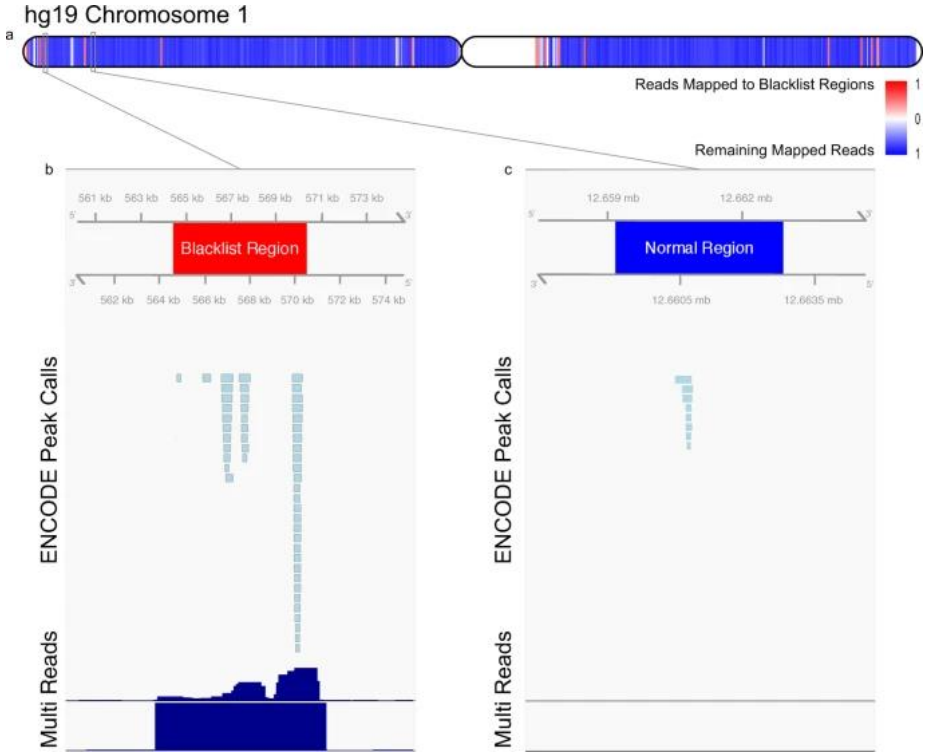
- **Broad peaks:** eg., histone modification. Here we are looking for broad peaks that cover entire gene bodies
- **Narrow peak**: eg., transcription factor binding. Here we are looking for regions of higher amplitude compared to background

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

MACS2 peak calling recommendations

Data type	**q value**	**--broad and --control flags**	**Reasoning**
ChIP-seq for TF	<0.01	--control <INPUT>	TF ChIP-seq often has very abrupt, small peaks that are well defined, so narrow peaks is necessary, and a less stringent adjusted p value is likely needed than for other data types
ChIP-seq for histone marks	<0.0001	--broad --control <INPUT>	Histone marks are often broadly dispersed without very well defined edges so a broad peak tag is useful but a very low p value helps differentiate between background and data
ATAC-seq	<0.0001	--control <INPUT>	ATAC-seq should show peaks at open chromatin across the genome similarly to histone ChIP-seq data, but with more abrupt peaks, so no broad peak tag is needed

Processing ChIP-seq Data

1. Let us pre-process ChIP-seq data and perform QC
   SR2022_worksheet_day9_preprocessing_ChIP-seq
   a. Map reads
   b. Perform QC
2. Let's run MACS to call peaks on our bam files with a genomic input.
   * *Note there is an input and anti-BACH1/GABPA ChIP-seq bam files.*
   a. Call peaks with MACS
      SR2022_worksheet_day9_MACS

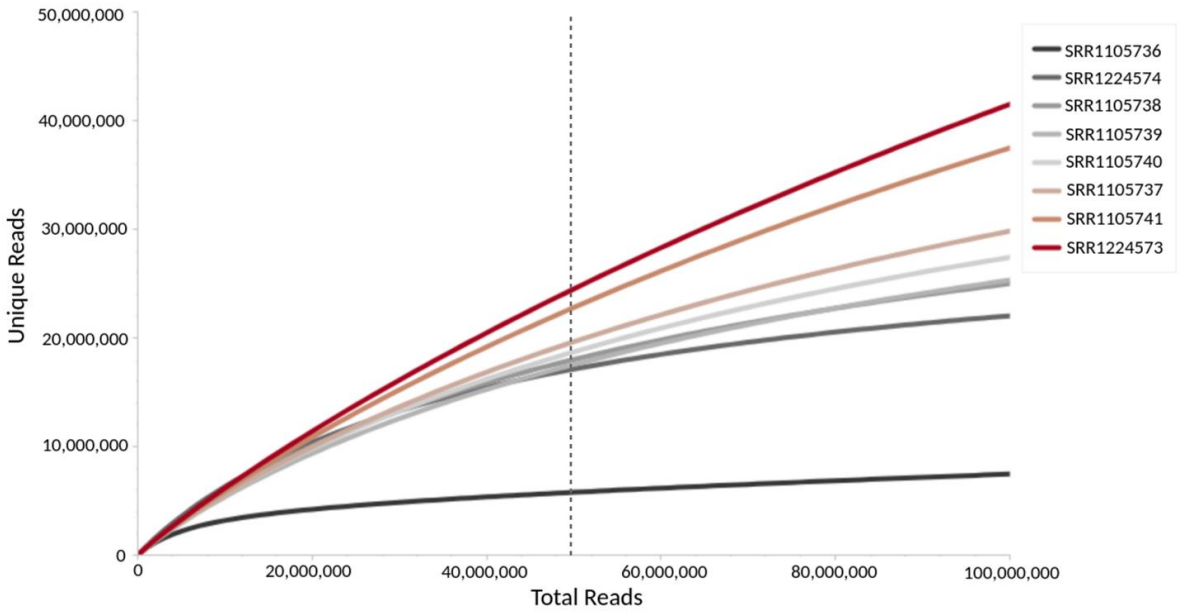
Usage

```
macs2 [-h] [--version]
    {callpeak,bdgpeakcall,bdgbroadcall,bdgcmp,bdgopt,cmbreps,bdgdiff,filterdup,predictd,pileup
```

Example for regular peak calling: `macs2 callpeak -t ChIP.bam -c Control.bam -f BAM -g hs -n test -B -q 0.01`

Example for broad peak calling: `macs2 callpeak -t ChIP.bam -c Control.bam --broad -g hs --broad-cutoff 0.1`

Preseq : Measures library complexity



Complexity = # of unique reads / total reads sequenced

```
$ preseq c_curve -o complexity_output.txt input.sorted.bam
```

Using Bedtools

BEDTools Documentation
https://bedtools.readthedocs.io/en/latest/

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 tutorial: http://quinlanlab.org/tutorials/bedtools/bedtools.html