In-Person

Genomics Compute Cluster

Peak Calling with MACS2

Northwestern INFORMATION TECHNOLOGY

RESEARCH COMPUTING AND DATA SERVICES

COMPUTING AND SOFTWARE

Access to high performance computing, research software, and global networks for conducting computationally intense research.

DATA MANAGEMENT AND SHARING

Learn about data management planning and options for storing, securing, transferring, and sharing data.

DATA SCIENCE, STATISTICS, AND VISUALIZATION

Support for collecting, analyzing, visualizing, and programming with research data.

TRAINING AND CONSULTATION

Identify events, resources, and people to help you learn computational and data skills for your research.

Research Computing and Data Services

We're here to help after the workshop!

quest-help@northwestern.edu

bit.ly/rcdsconsult

https://sites.northwestern.edu/researchcomputing/

Set up

1. log onto Quest

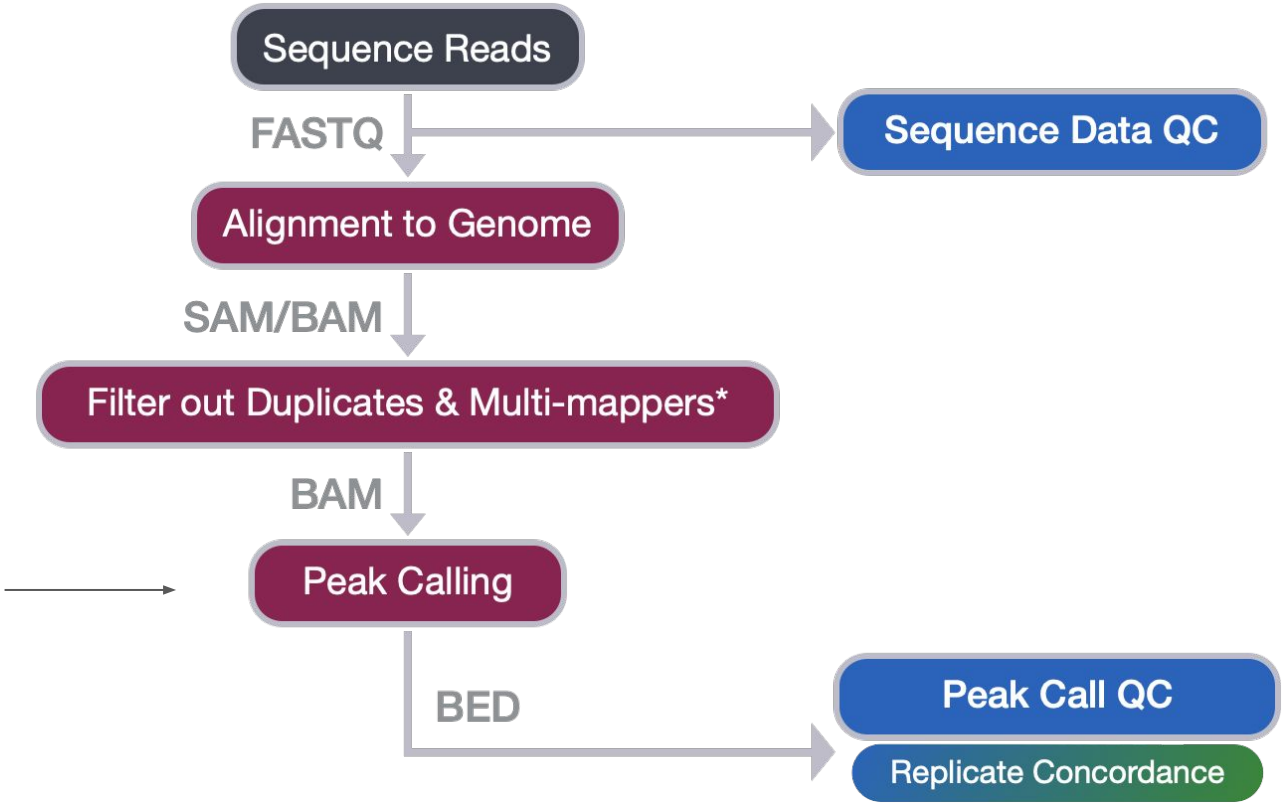
   ssh <netid>@login.quest.northwestern.edu # enter your netid password
2. Move to our classroom folder

   cd /projects/e32680
3. Make your own subfolder if you don’t have one

   mkdir <folder name>

   cd ./<folder name> #navigate to your folder

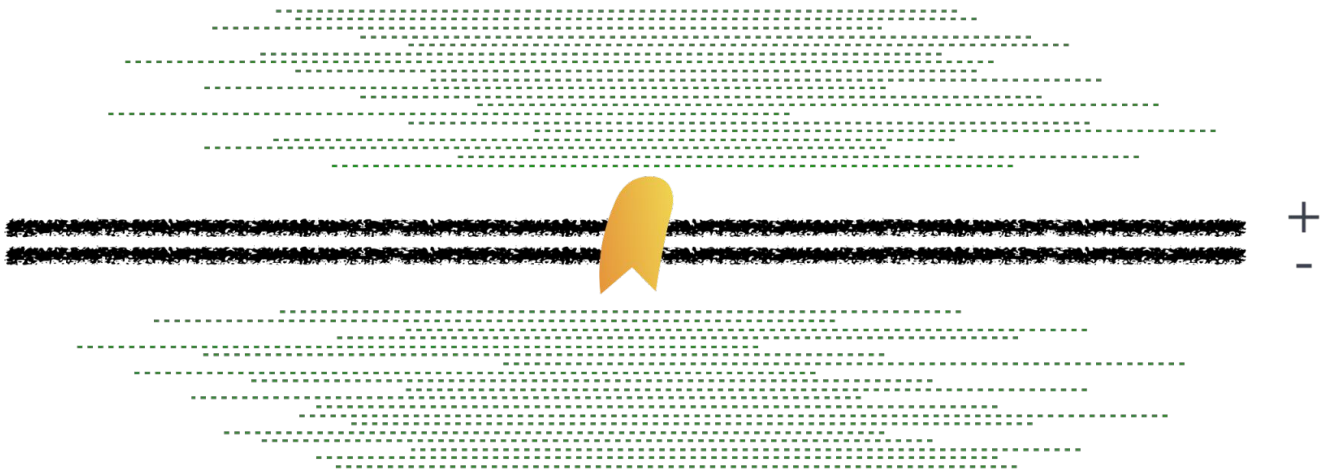
General ChIP-seq workflow



Bimodal nature of ChIP-seq data

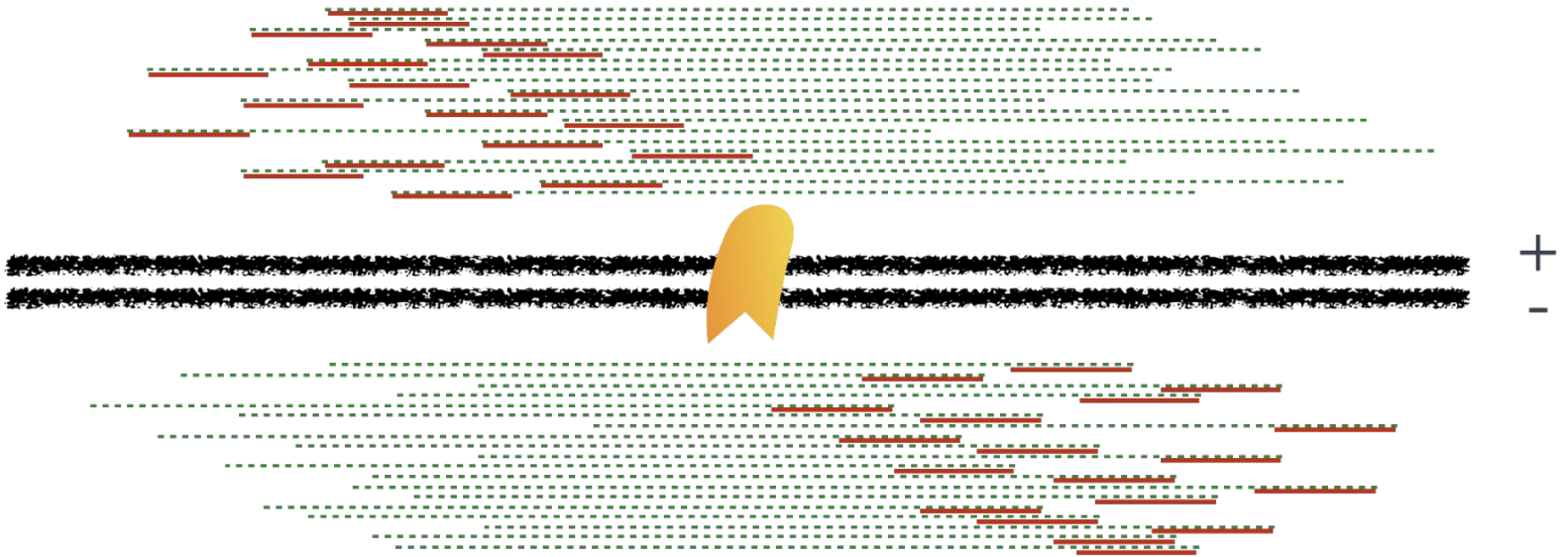
= binding site

---- = size selected DNA fragment



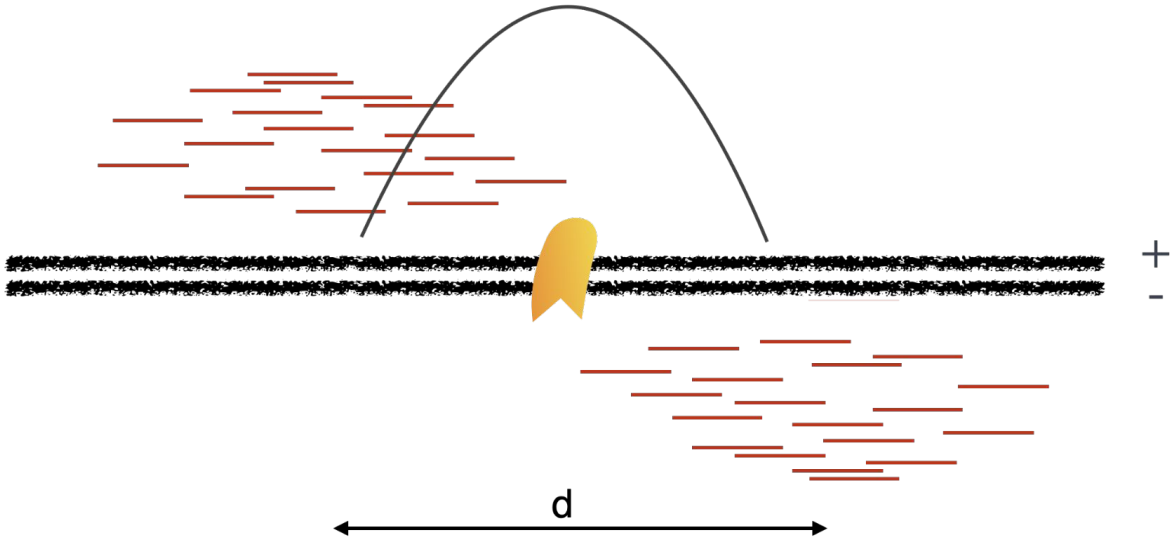
Bimodal nature of ChIP-seq data

ChIP-seq fragments are sequenced from the 5’ end



Bimodal nature of ChIP-seq data

Alignment generates a **bimodal pattern** on the plus and minus strands around binding sites

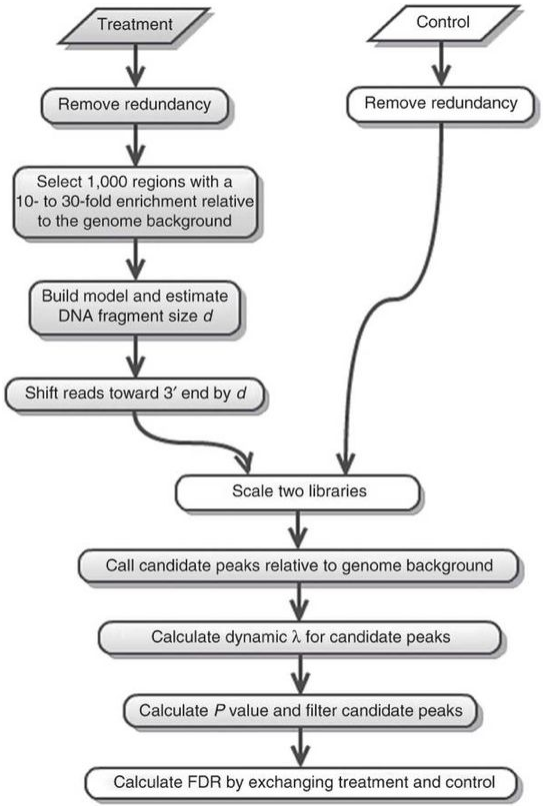


Peak calling algorithms use this pattern to estimate the relative strand shift

MACS workflow

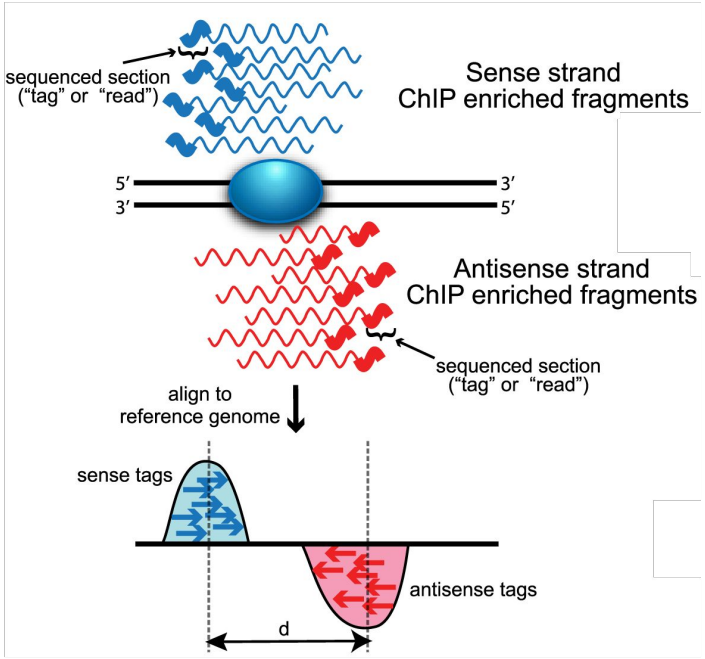
1. Estimate the shift size

2. Calling peaks



Modeling the shift size (d)

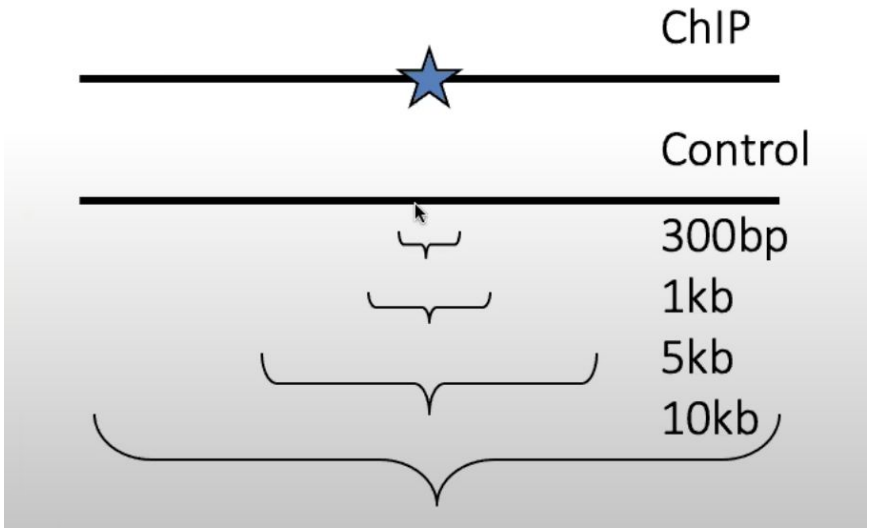
- MACS searches for highly enriched regions across the genome (600bp window with 50-fold enrichment)
- Randomly samples 1000 peaks
- Estimates d (distance between the modes of the two peaks)



Wilbanks and Faccioti, PLoS One 2010

Peak calling

- MACS models peak distribution using the Poisson distribution.
- MACS computes λ(expected number of reads in a given window) for each candidate peaks.
- A Poisson distribution *p* is computed to identify significant enrichment (*p* < 1e-5).
- Estimate FDR using the Benjamini-Hochberg correction

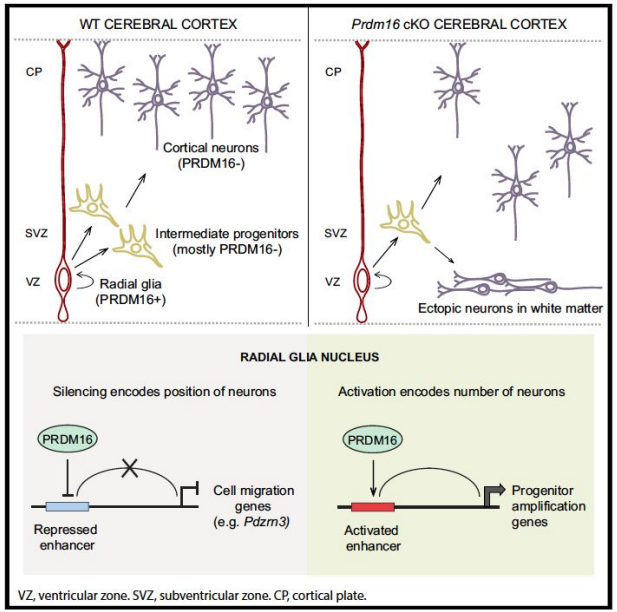


Dynamic λlocal: control local bias

Experiment setup: identify and validate the targets and activities of PRDM16 using PRDM16 conditional knockout mice

- Groups: WT and PRDM16 KO
- 2 replicates per group

Graphical Abstract



Baizabal
Neuron 2018

Course folder setup

cd /projects/e32680/03_macspeakcalling_reference

#course folder

./00_fastq

#raw data (fastq) folder

./01_bam

#bam files folder

./scripts

#scripts folder

Sbatch commands

```
#!/bin/bash
#SBATCH -A e32680
#SBATCH -p short
#SBATCH -t 1:00:00
#SBATCH --mail-type=BEGIN,END,FAIL,REQUEUE
#SBATCH --mail-user=qianliliu2020@u.northwestern.edu
#SBATCH --output=./logs/%x_%j.out
#SBATCH -N 1
#SBATCH -n 3
#SBATCH --mem-per-cpu=1gb
#SBATCH --export=NONE
#SBATCH -J MAC2_peakcalling
```

Running MACS2

```
macs2 callpeak -t ./wt_sample1_chip.bam \
 -c ./wt_sample1_input.bam \
-f BAM -g mm \
-n wt_sample1 \
--outdir
```

- -t: data file
- -c: input control
- -f: format -g ;mappable genome size
- -n prefix string for output files
- --outdir output directory

Running MACS2

```
cd ./<YourFolder>
#Navigate to your folder
cp /projects/e32680/03_macspeakcalling_reference/scripts/MACS2_1.sh .
#copy script to run MACS
nano MACS2_1.sh
#edit script
sbatch MACS2_1.sh
#run script
```

Check on jobs

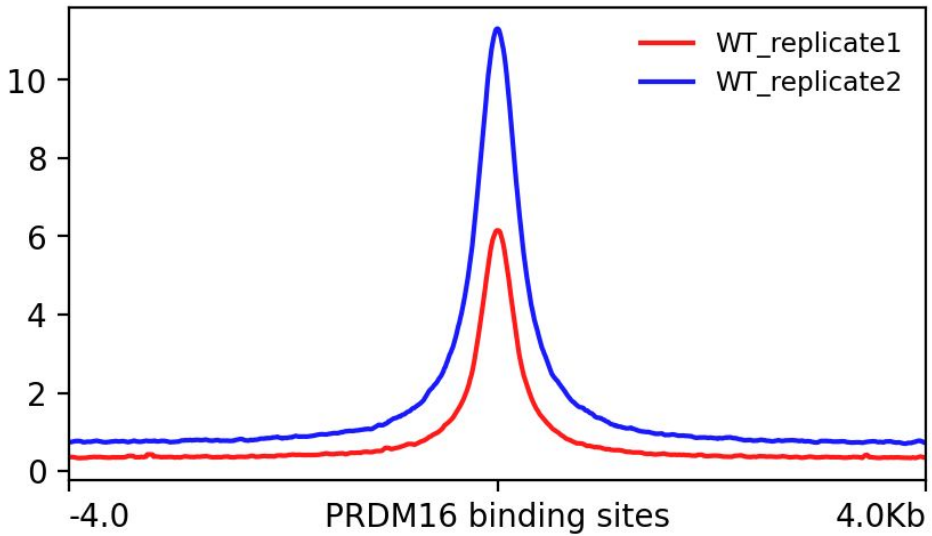
```
squeue --me
# shows all running and pending jobs
sacct -X
# shows all jobs from today
sacct -X -S 040125
# shows all jobs from this month
```

MACS2 Output

- _peaks.narrowPeak: BED6+4 format file which contains the peak locations together with peak summit, pvalue and qvalue
- _peaks.xls: a tabular file which contains information about called peaks.
- _summits.bed: The location in the peak with the highest fragment pileup. These are the predicted precise binding location and recommended to use for motif finding.
- _model.R: an R script which you can use to produce a PDF image about the model based on your data and cross-correlation plot

After peak calling

- Finding overlapping peaks between replicates.
    - Bedtools intersect
- Visualizing peaks using deepTools.



Finding overlapping peaks between replicates

```
bedtools intersect \
    -wo -f 0.3 -r \  #
    -a wt_sample1_peaks.narrowPeak \
    -b wt_sample2_peaks.narrowPeak \
    >  <YourFolder>/wt_peaks_final.bed
```

- -wo: Write the original A (file 1) and B (file 2) entries plus the number of base pairs of overlap between the two features.
- -f: Minimum overlap required as a fraction of A.
- -r: Require that the fraction overlap be reciprocal for A and B. (we require the overlap region being at least 30% in A and B)

Creating bigwig files for visualization

Module load deeptools/3.5.6

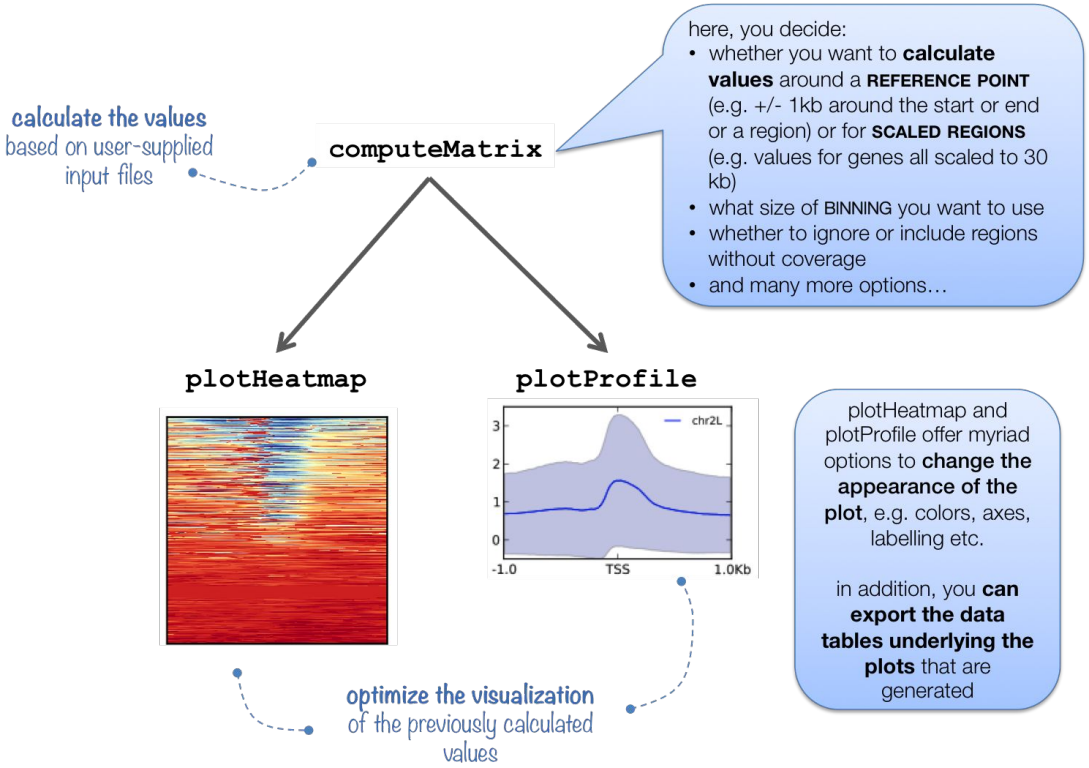
bamCoverage -b /projects/e32680/03_macspeakcalling_reference/01_bam

/WT_REP1.mLb.clN.sorted.bam \

-o <YourFolder>/wt_sample1_chip.bw \

--binSize 20

Evaluating signal in PRDM16 binding sites



https://deeptools.readthedocs.io/en/latest/

Compute matrix

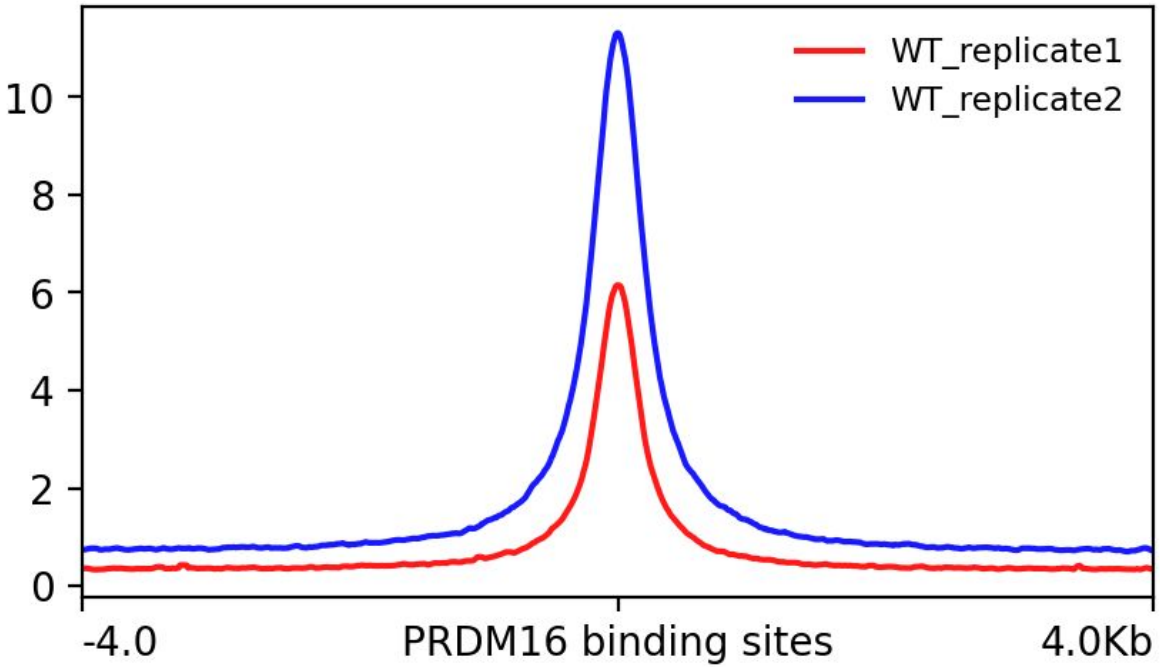
```
computeMatrix reference-point --referencePoint center \

-b 4000 -a 4000 \

-R <YourFolder>/wt_peaks_final.bed \

-S <YourFolder>/wt_sample1_chip.bw <YourFolder>/wt_sample2_chip.bw \

--skipZeros -o <YourFolder>/wt_matrix.gz
```

- `reference-point`: The reference point for plotting.
- `-b, a`: Specify a window around the reference point
- `-R`: The region file (we will use the WT replicate overlap BED file).
- `-S`: The list of bigWig files
- `--skipZeros`: Do not include regions with only scores of zero
- `-o`: output file name

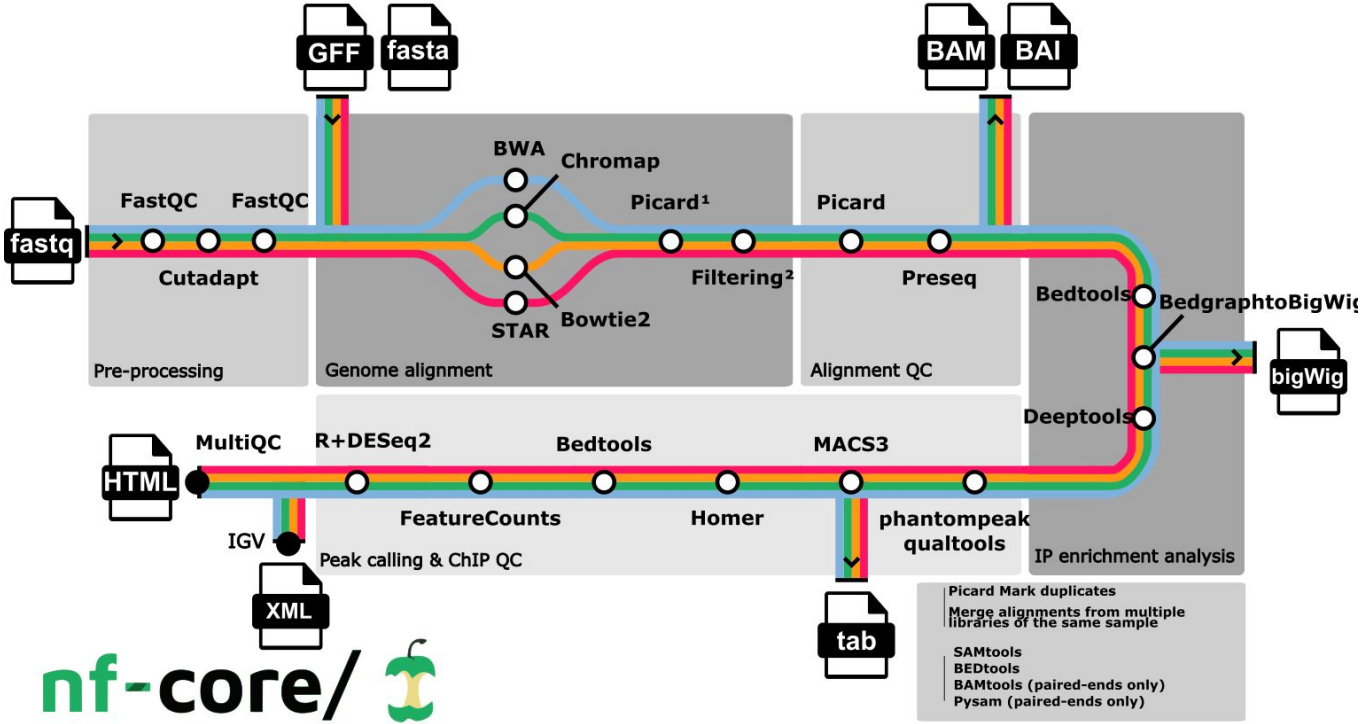
Drawing the profile plot

```
plotProfile -m ./wt_matrix.gz \

-out ./plot1_wt_replicates.png \

--regionsLabel "" \

--perGroup \

--colors red blue \

--samplesLabel "WT_replicate1" "WT_replicate2" \

--refPointLabel "PRDM16 binding sites"
```

- `-out`: output file name
- `--regionsLabel`: Labels for the regions plotted in the heatmap
- `--perGroup`: The default is to plot all groups of regions by sample
- `--colors`: List of colors to use for the plotted lines
- `--samplesLabel`: Labels for the samples plotted
- `--refPointLabel:` Label shown in the plot for the reference-point.


WT_replicate1
WT_replicate2
10
8
6
4
2
0
-4.0
PRDM16 binding sites
4.0Kb

Running MACS with Nextflow



nf-core/chipseq

https://nf-co.re/chipseq/2.0.0/