ChIP-seq and ATAC-seq analysis

Inma Barrasa and Bingbing Yuan

BaRC Hot Topics –May 19th 2022
Bioinformatics and Research Computing
Whitehead Institute

http://barc.wi.mit.edu/hot_topics/

OUTLINE

- Overview of ChIP-seq and ATAC-seq
  - Analysis workflow: common and specific steps
- Detailed ChIP-seq analysis
  - Quality control/adapter removal
  - Mapping
  - Peak calling, differential binding, FRiP score
  - Remove peaks overlapping with black list regions
  - Downstream analysis
- Detailed ATAC-seq analysis
  - Pre-alignment quality control
  - Mapping, post-alignment filtering and quality control
  - Peak calling, assessing peak calls with FRiP score
  - Remove peaks overlapping with black list regions
  - Downstream analysis

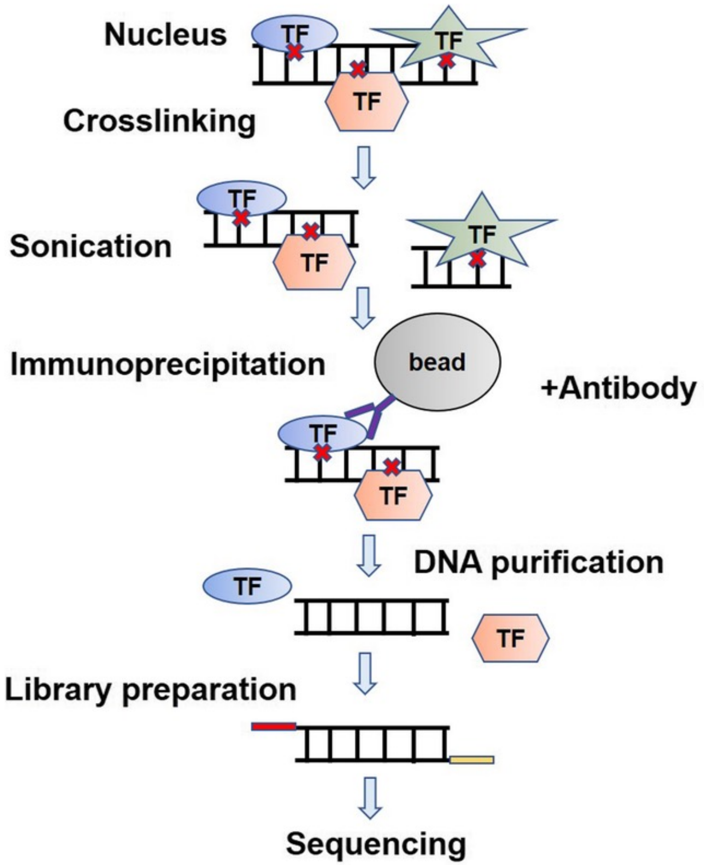
Goals of ChIP-seq and ATAC-seq

- **ChIP-seq: Ch**romatin **I**mmuno**p**recipitation **seq**uencing
  - Identify the regions of chromatin bound by a specific protein, *e.g.* transcription factor; or that are part of a nucleosome containing specific histone modifications, *e.g.* H3K27ac (correlates with enhancers), H3K4me3 (active transcription), H3K9me3 (represive mark, associated with condensed and constitutive heterochromatin)

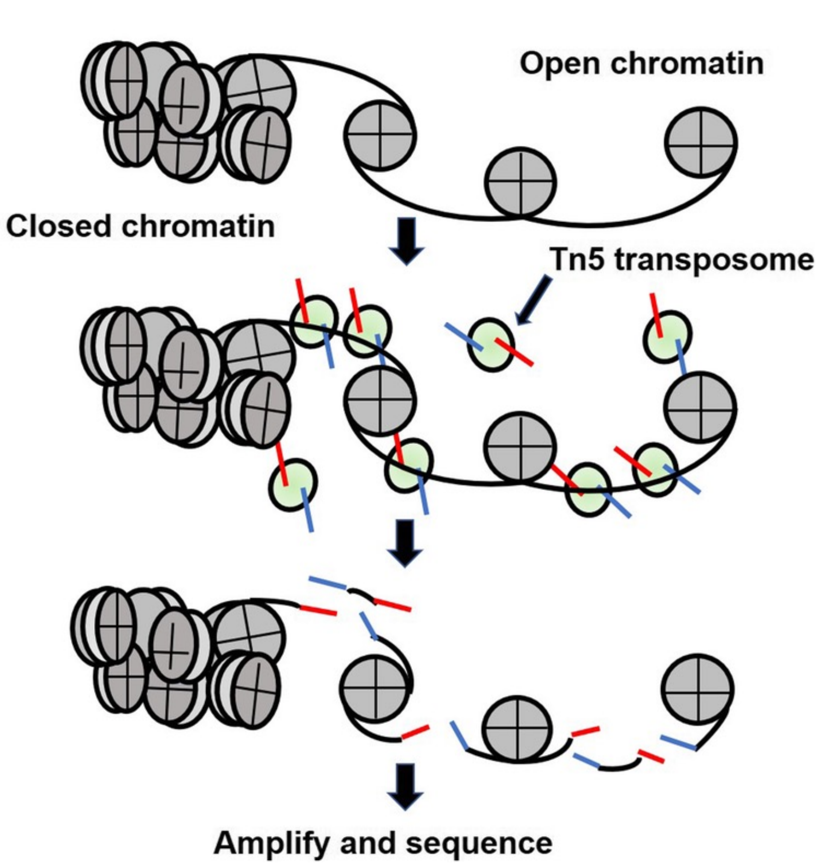
- **ATAC-seq: A**ssay for **T**ransposase-**A**ccessible **C**hromatin with high-throughput **seq**uencing
  - Identification of open chromatin regions, nucleosome positioning, and regulatory motifs

ChIP-seq and ATAC-seq overview

ChIP-seq

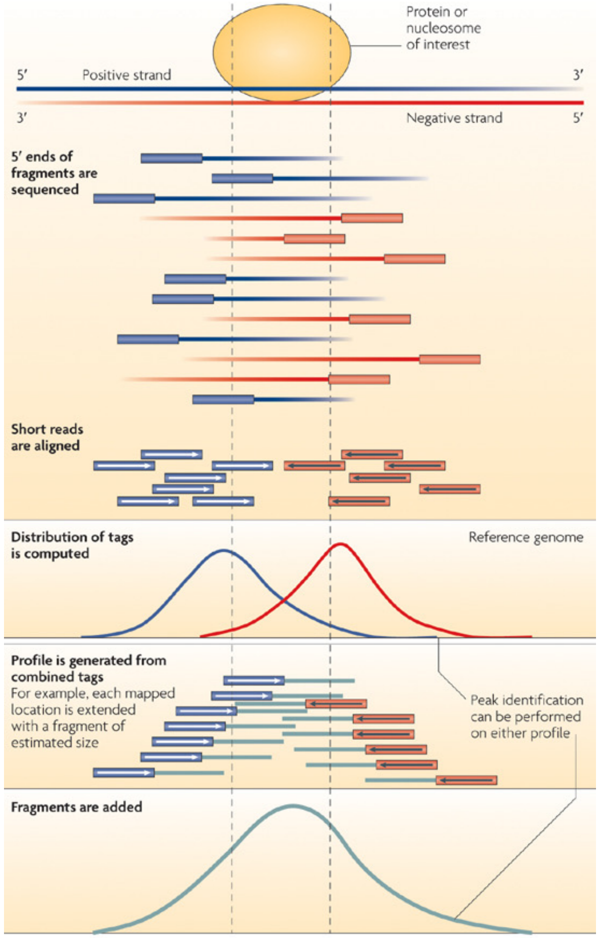


ATAC-seq



Ma, S., Zhang, Y. Profiling chromatin regulatory landscape: insights into the development of ChIP-seq and ATAC-seq. *Mol Biomed* **1,** 9 (2020).

ChIP-seq analysis workflow


Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned
Distribution of tags is computed
Reference genome
Profile is generated from combined tags
For example, each mapped location is extended with a fragment of estimated size
Peak identification can be performed on either profile
Fragments are added

Nature Reviews | Genetics

Steps in ChIP-seq data analysis

Goal: Find the regions bound by the protein assayed

Peak = region bound by the assayed protein or modified histone

1. Quality control

   Optional: remove adapters

2. Mapping

   Treat IP and control the same way (preprocessing and mapping)

3. Peak calling and differential analysis

   i) Read extension (--extsize fragment_length) and signal profile generation, based on fragment length.

   ii) Peak assignment

   iii) Differential peak binding

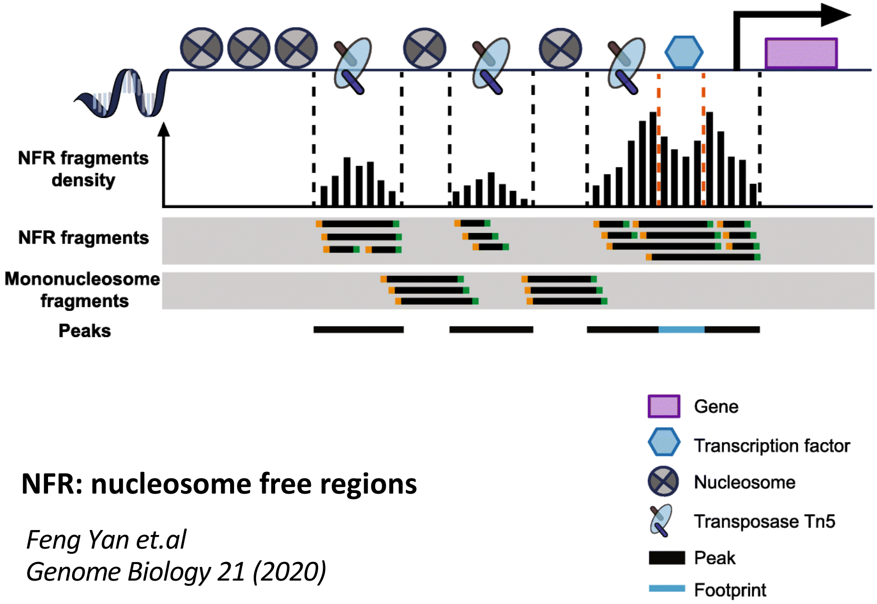
4. Peak analysis and interpretation

   i) Assign peaks to genes the factors may be regulating

   ii) Find motifs within peaks

Park, P. J., *ChIP-seq: advantages and challenges of a maturing technology, Nat Rev Genet. Oct;10(10):669-80 (2009)*

ATAC-seq analysis workflow



NFR: nucleosome free regions

Feng Yan et.al
Genome Biology 21 (2020)

Steps in ATAC-seq data analysis

Goal: Find the open chromatin regions

Peak = open chromatin region

1. Quality control

   Remove adapters if necessary
2. Mapping

   Tailored to paired end reads
3. Remove reads mapped to mitochondrial genes
4. Peak calling and differential analysis

   i) Read shift and extension and signal profile generation.

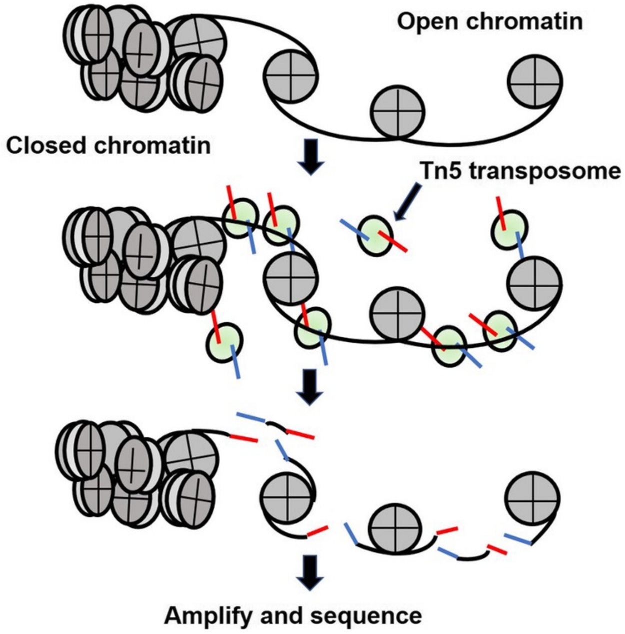
   ii) Peak assignment

5. Peak interpretation

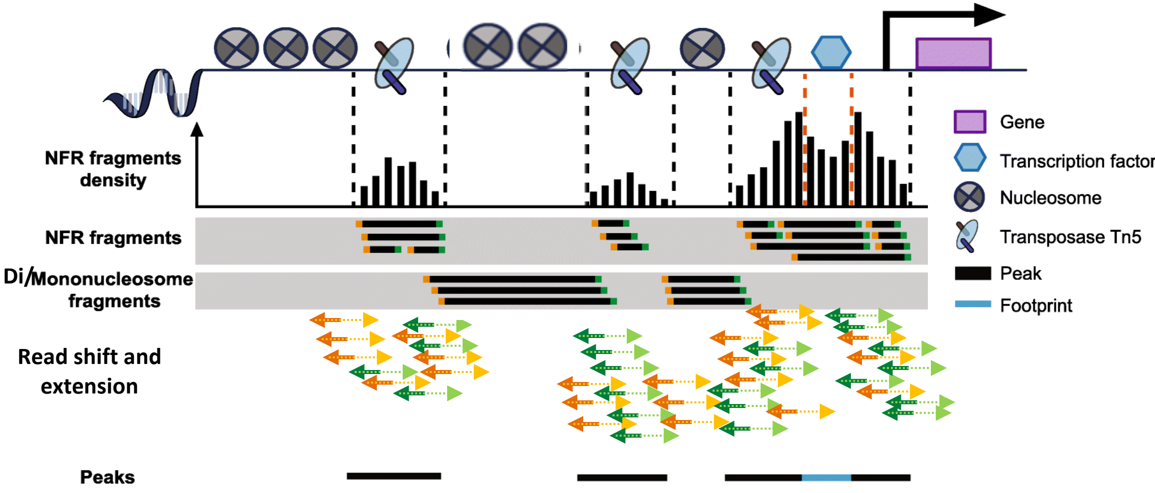
   i) Find genes next the open chromatin regions

   ii) Find motifs within peaks
6. Foot printing analysis nucleosome positioning analysis (if enough sequencing)

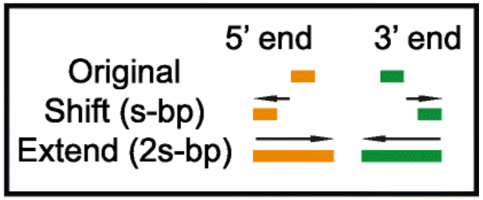
Profile generation in ATAC-seq



Ma, S., Zhang, Y. Profiling chromatin regulatory landscape: insights into the development of ChIP-seq and ATAC-seq. *Mol Biomed* **1,** 9 (2020).



Modified from Feng Yan et.al Genome Biology 21 (2020)



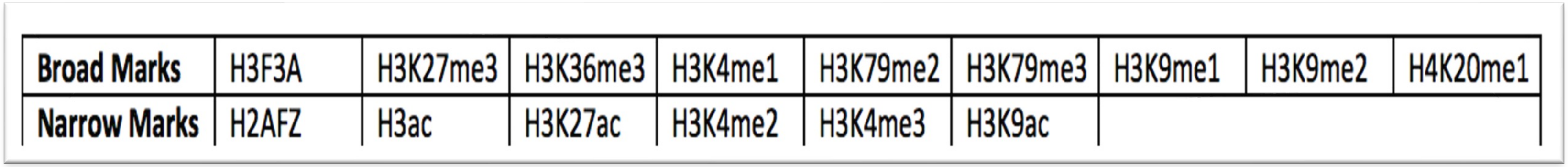
To create a signal profile centered around the cutting sites we will shift the reads 75 bp toward 3'->5' direction using *--shift -75*, and the extend the reads extended 150 bp towards 5'->3‘ using --*extsize 150*

Experimental design for ChIP-seq

- Read length:
  - Short reads is enough : 35bp-50bp
    - If the factor assayed binds repeat regions, using pair-end reads and longer reads will improve the mapping
- Control samples (Input, IgG):
  - Under the same conditions as the test sample
  - At least the same depth as test sample
- Replicates:
  - Technical replicates: not necessary
  - Biological replicates: at least 2

ENCODE Guideline for sequencing depth

- Transcription Factors:
  - Each replicate should have >=20 mil fragments
- Histone:
  - Narrow peak: >= 20 mil
  - Broad peak: >= 45 mil

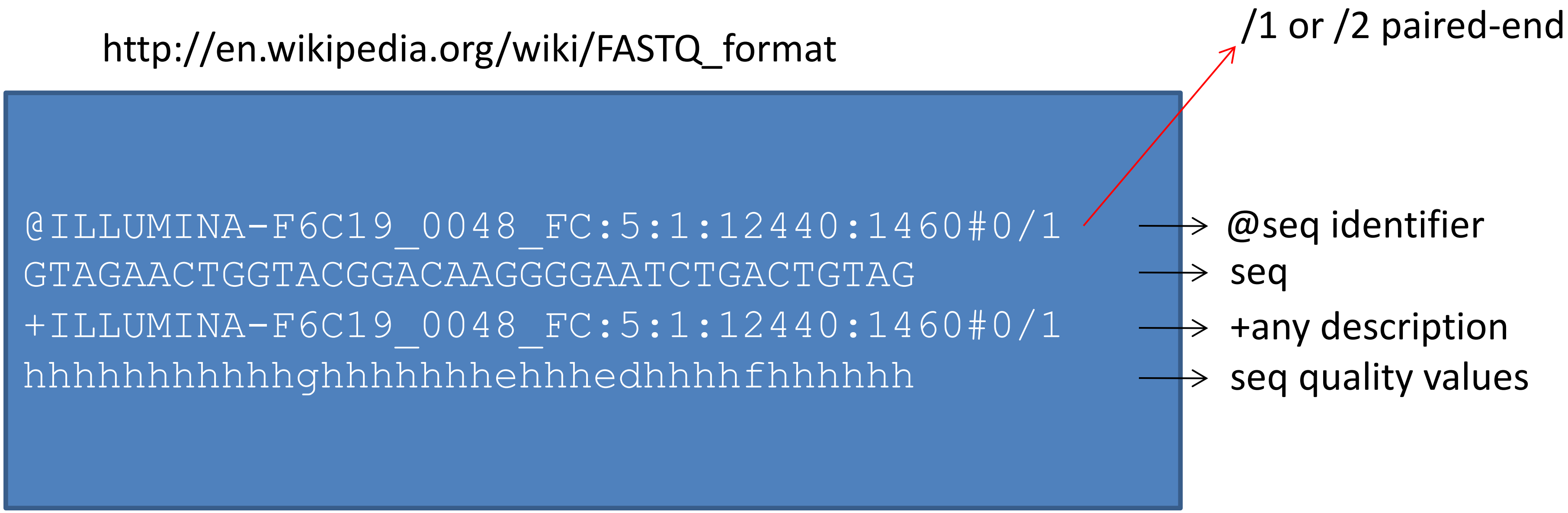


Broad Marks	H3F3A	H3K27me3	H3K36me3	H3K4me1	H3K79me2	H3K79me3	H3K9me1	H3K9me2	H4K20me1
Narrow Marks	H2AFZ	H3ac	H3K27ac	H3K4me2	H3K4me3	H3K9ac			

Illumina data format

- Fastq format:

http://en.wikipedia.org/wiki/FASTQ_format



```
@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhghhhhhhhehhhedhhhhfhhhhhh
```

/1 or /2 paired-end

@seq identifier

seq

+any description

seq quality values

Input qualities	Illumina versions
--solexa-quals	<= 1.2
--phred64	1.3-1.7
--phred33	>= 1.8

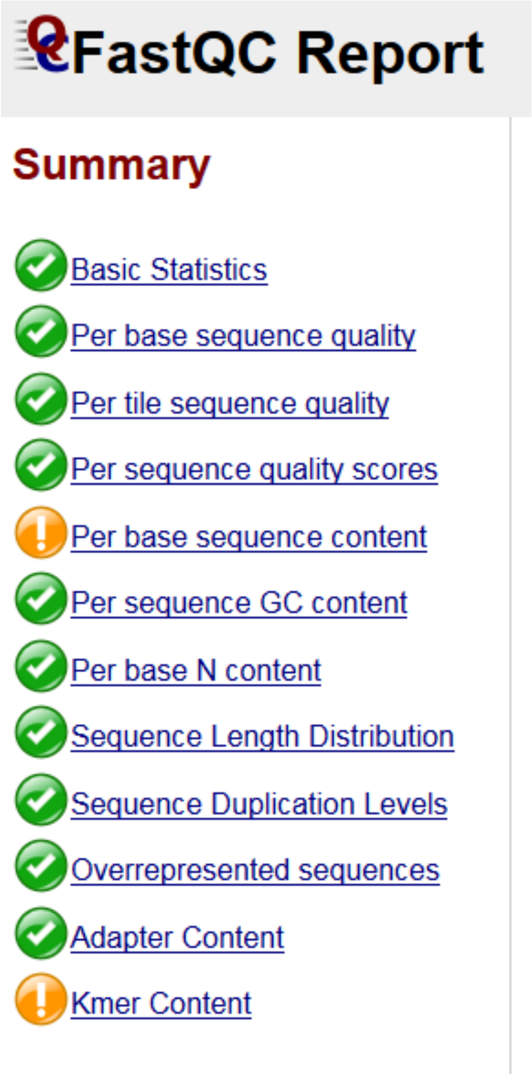
Check read quality with FastQC

(http://www.bioinformatics.babraham.ac.uk/projects/fastqc/)

1. Run FastQC to check read quality

   bsub fastqc sample.fastq

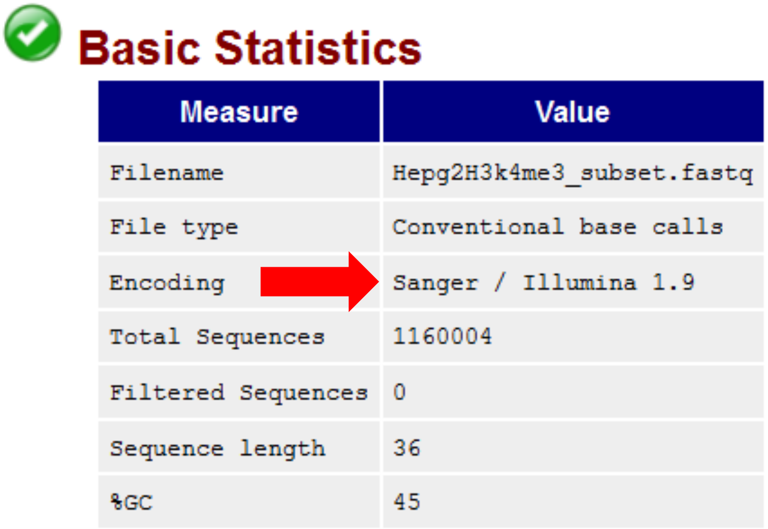
2. Open output file: "fastqc_report.html"



FastQC Report

Summary

- ✓ Basic Statistics
- ✓ Per base sequence quality
- ✓ Per tile sequence quality
- ✓ Per sequence quality scores
- ! Per base sequence content
- ✓ Per sequence GC content
- ✓ Per base N content
- ✓ Sequence Length Distribution
- ✓ Sequence Duplication Levels
- ✓ Overrepresented sequences
- ✓ Adapter Content
- ! Kmer Content



✓ Basic Statistics

Measure	Value
Filename	Hepg2H3k4me3_subset.fastq
File type	Conventional base calls
Encoding	Sanger / Illumina 1.9
Total Sequences	1160004
Filtered Sequences	0
Sequence length	36
%GC	45

Preprocessing tools

- **Fastx Toolkit** (http://hannonlab.cshl.edu/fastx_toolkit/)
  - FASTQ/A Trimmer: Shortening reads in a FASTQ or FASTQ files (removing barcodes or noise).
  - FASTQ Quality Trimmer: Trims (cuts) sequences based on quality
  - FASTQ Masker: Masks nucleotides with 'N' (or other character) based on quality

  (for a complete list go to the link above)
- **cutadapt** to remove adapters

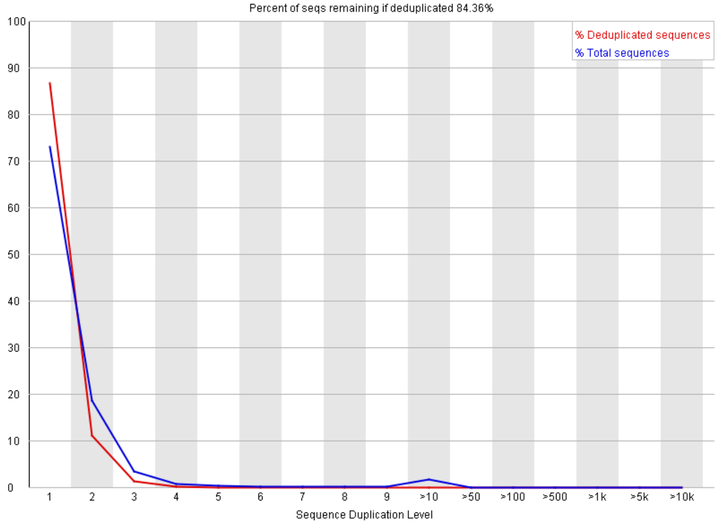
  (https://code.google.com/p/cutadapt/)

Recommendation for preprocessing

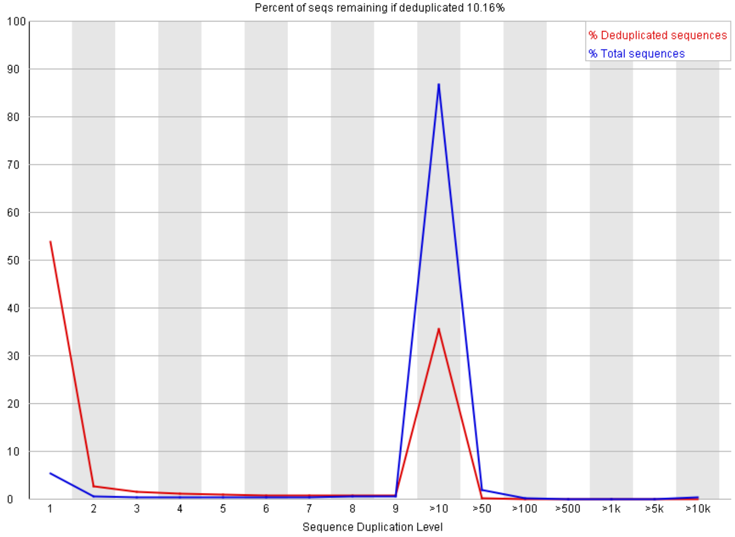
- Treat IP and control samples the same way during preprocessing and mapping.
- Watch out for preprocessing that may result in very different read length in the different samples as that can affect mapping.
- If you have paired-end reads, make sure you still have both reads of the pair after the processing is done.
- Run FastQC on the processed samples to see if the problem has been removed.

Library Complexity

- Refers to the fraction of unique fragments present in a given library.
- One proxy for library complexity is to look at the sequence duplication levels on the FastQC report:



% Complexity calculated by the genome core
85.6



% Complexity calculated by the genome core
4.95

Library Complexity II

- Low library complexity may be an indicator that:
  - A new sample and a new library should be prepared.
  - We have to find a better Ab to perform the IP.
  - We can not sequence the same sample anymore because we will not find new sequences.
- In certain experimental settings we may expect a low library complexity. *i.e.* We are profiling a protein that binds to a small subset of the genome.

Mapping
Non-spliced alignment software

- Bowtie2:
  - Supports gapped alignment.
  - Use “end-to-end“ (default) alignment mode
  - Previous Hot Topics on how to run the programs
- BWA:
  - refer to the BaRC Best practices for detailed information

    http://barcwiki.wi.mit.edu/wiki/SOPs
- Filter reads mapped by quality mapping score

```
samtools view -bq 10 file.bam > filtered.bam
```

Local genomic files needed for mapping

tak: /nfs/genomes/

- Human, mouse, zebrafish, *C.elegans*, fly, yeast, etc.
- Different genome builds
  - mm9: mouse_gp_jul_07
  - mm10: mouse_mm10_dec_11
- human_hg38_dec13 vs human_hg38_dec13_no_random
  - human_hg38_dec13 includes *_random.fa, *hap*.fa, etc.
- Sub directories:
  - bowtie
    - Bowtie1: *.ebwt
    - Bowtie2: *.bt2
  - fasta: one file per chromosome
  - fasta_whole_genome: all sequences in one file
  - gtf: gene models from Refseq, Ensembl, etc.

Example commands: Mapping the reads

```
bsub bowtie2 --phred33-quals -N 1 -x
/nfs/genomes/human_gp_feb_09_no_random/bowtie/hg19   -U
Hepg2Control_subset.fastq | samtools view -b - >
Hepg2Control_subset_hg19.N1.bam
```

-N <int>   max # mismatches in seed alignment; can be 0 or 1 (0)

-x <bt2-idx>

Optional: filter reads mapped by quality mapping score

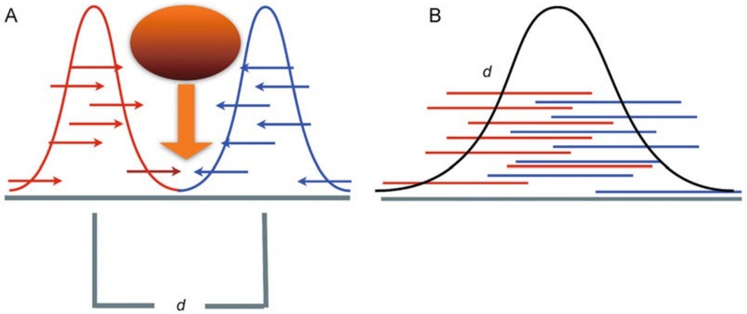
```
samtools view -bq 10 file.bam > filtered.bam
```

Estimation of the fragment length
Strand cross-correlation analysis

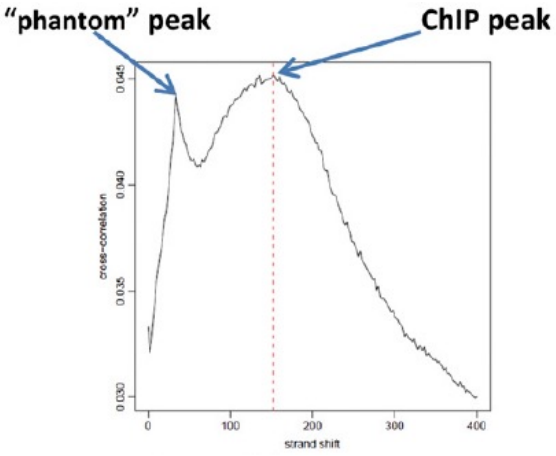
Example command:

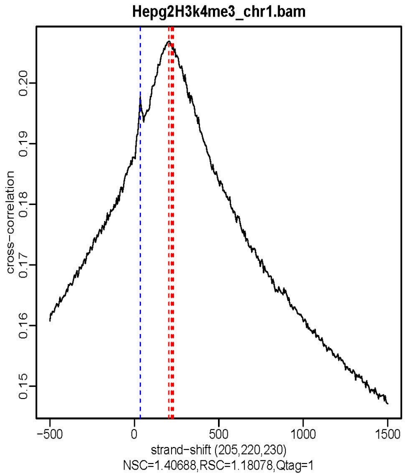
```
/nfs/BaRC_Public/phantompeakqualtools/run_spp.R  -c=H3k4me3_chr1.bam  -savp
-out=H3k4me3_chr1.run_spp.out
```

Cross-correlation: compute the Pearson's correlation between Watson and Crick strand after shifting Watson by n bases



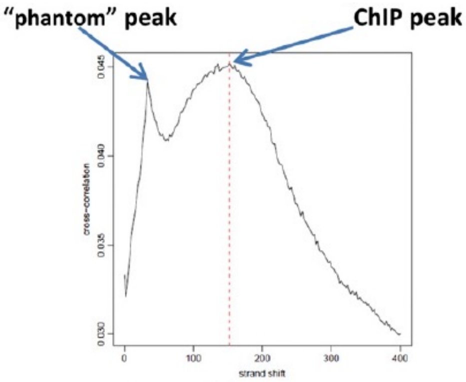
Shin et. Al. *Quantitative Biology*
March 2013, Volume 1, Issue 1, pp 54–70

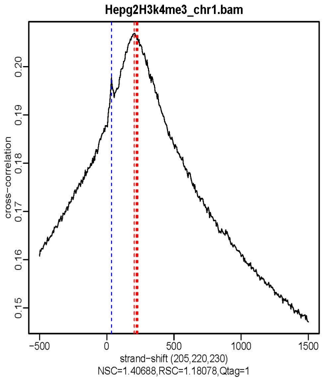


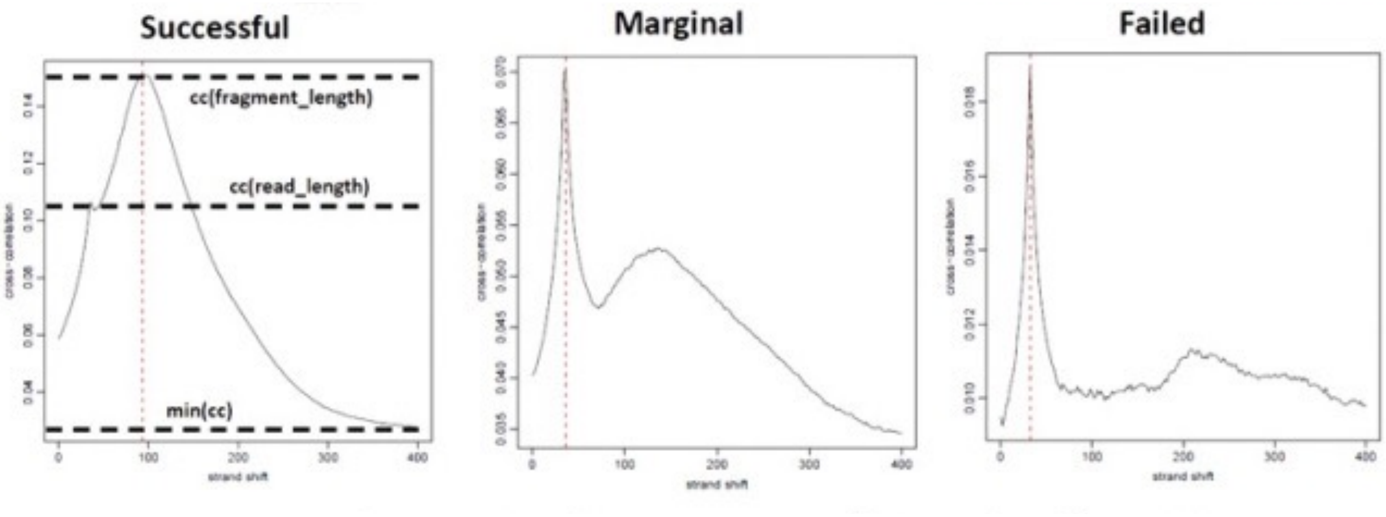


Genome Res. 2012. 22:1813

Estimation of the fragment length
Strand cross-correlation analysis







Genome Res. 2012. 22:1813

Peak calling

i) Read extension and signal profile generation

- strand cross-correlation can be used to calculate fragment length

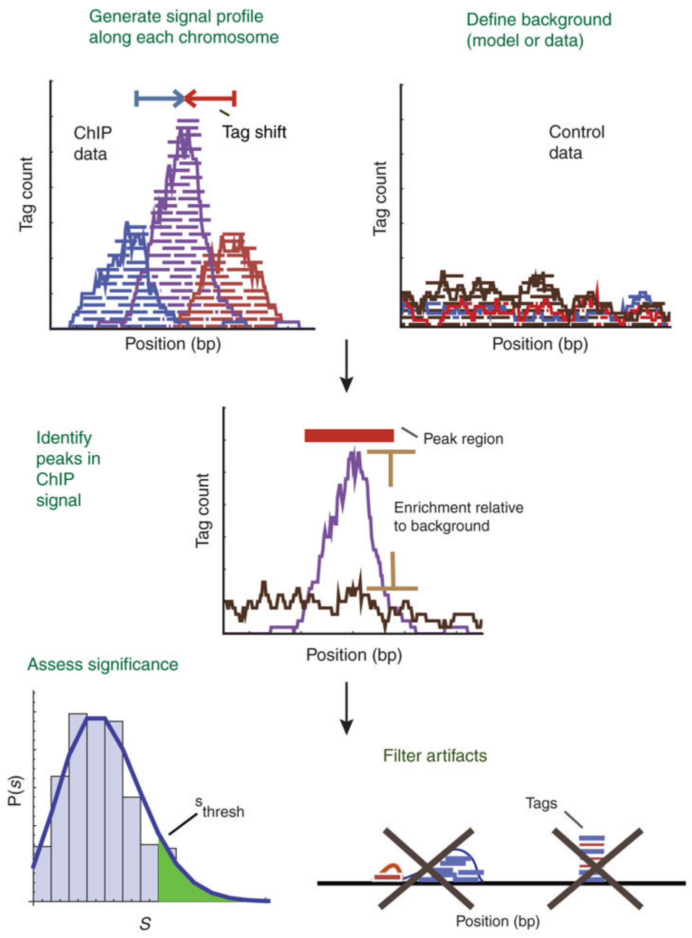
ii) Peak evaluation

- Look for fold enrichment of the sample over input or expected background
- Estimate the significance of the fold enrichment using:
  - Poisson distribution
  - background distribution from input DNA
  - model background data to adjust for local variation (MACS)

iii) Broad peaks:

two-level peak calling (MACS):

- embed stronger/narrower calls in weaker/broader calls



Pepke, S. *et al*. Computation for ChIP-seq and RNA-seq studies, *Nat Methods. Nov. 2009*

Peak calling: MACS

- MACS can calculate the fragment length but we will use a different program and give MACS the fragment length as an input parameter.
- It uses a **Poisson distribution** to assign p-values to peaks. But the distribution has a dynamic parameter, local lambda, to capture the influence of **local biases**.
- MACS default is to **filter out redundant tags** at the same location and with the same strand by allowing at most 1 tag. This works well.
- -g: You need to set up this parameter accordingly:

  Effective genome size. It can be 1.0e+9 or 1000000000, or shortcuts: 'hs' for human (2.7e9), 'mm' for mouse (1.87e9), 'ce' for *C. elegans* (9e7) and 'dm' for fruit fly (1.2e8), Default:hs
- For broad peaks like some histone modifications it is recommended to use `--nomodel` and if there is not input sample to use `--nolambda`.

Example of MACS command

MACS command

```
bsub  macs2  callpeak  -t H3k4me3_chr1.bam -c  Control_chr1.bam --name H3k4me3_chr1
-f BAM -g hs --nomodel -B --extsize "size calculated on the strand crosscorrelation analysis“
```

PARAMETERS

- -t TFILE Treatment file
- -c CFILE Control file
- --name NAME Experiment name, which will be used to generate output file names. DEFAULT: “NA”
- -f FORMAT Format of tag file, “BED” or “SAM” or “BAM” or “BOWTIE”. DEFAULT: “BED”
- --nomodel skips the step of calculating the fragment size.
- -B create a bedgraph
- --extsize EXTSIZE The arbitrary extension size in bp. When nomodel is true, MACS will use this value as fragment size to extend each read towards 3' end, then pile them up. **You can use the value from the strand cross-correlation analysis**

MACS output

Output files:

1. Excel peaks file ("_peaks.xls") contains the following columns

```
Chr, start, end, length, abs_summit, pileup,
-LOG10(pvalue), fold_enrichment, -
  LOG10(qvalue), name
```

2. "_summits.bed": contains the peak summits locations for every peaks. The 5th column in this file is -log10qvalue
3. "_peaks.narrowPeak" is BED6+4 format file. Contains the peak locations together with peak summit, fold-change, pvalue and qvalue.

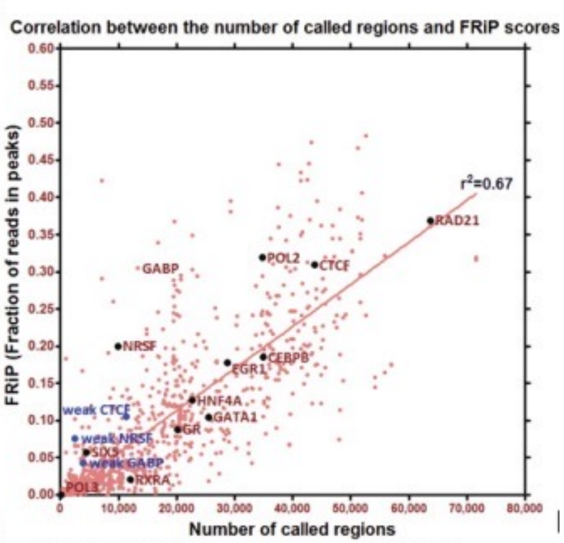
To look at the peaks on a genome browser you can upload one of the output bed files or you can also make a bedgraph file with columns (step 3 of hands on):

```
chr, start, end, fold_enrichment
```

How good is your ChIP-seq?

- Library Complexity, low duplication level
- Cross-correlation analysis
- FRiP: Fraction of Reads In Peaks
- RiBL: Reads Intercept with Black List
- Visualize peaks in genome browser
- Irreproducible Discovery Rate (IDR)
- Enrichment of reads in genomic features

FRiP: Fraction of Reads In Peaks



Genome Res. 2012. 22:1813

- To check success of the immunoprecipitation
- Often correlates with the number of called regions.
- Different proteins have different values
- Uses:
  - To compare results obtained with the same antibody across cell lines.

Can be calculated with QCmetrics from ChIPQC
Or

```
/nfs/BaRC_Public/BaRC_code/shell/calculate_FRiP_score/calculate_FRiP_score.sh
  SampleA.bam SampleA_peaks.narrowPeak > SampleA_FRiP_score.txt
```

Reads Intercept with Black List (RiBL)

- Artificially high signal :
  - genomic regions with anomalous, unstructured, high signal/read counts in next gen sequencing experiments independent of cell line and type of experiment
- Often in repeats (centromeres, telomeres, satellite repeats)
  - 0.5% of genome, but could account for >10% total signal
- Could influence peak caller, fragment length estimation
- Solution: masking regions prior to analysis
- BaRC_datasets -> ENCODE_blacklists
  - Human, mouse, fly, *C. elegans*

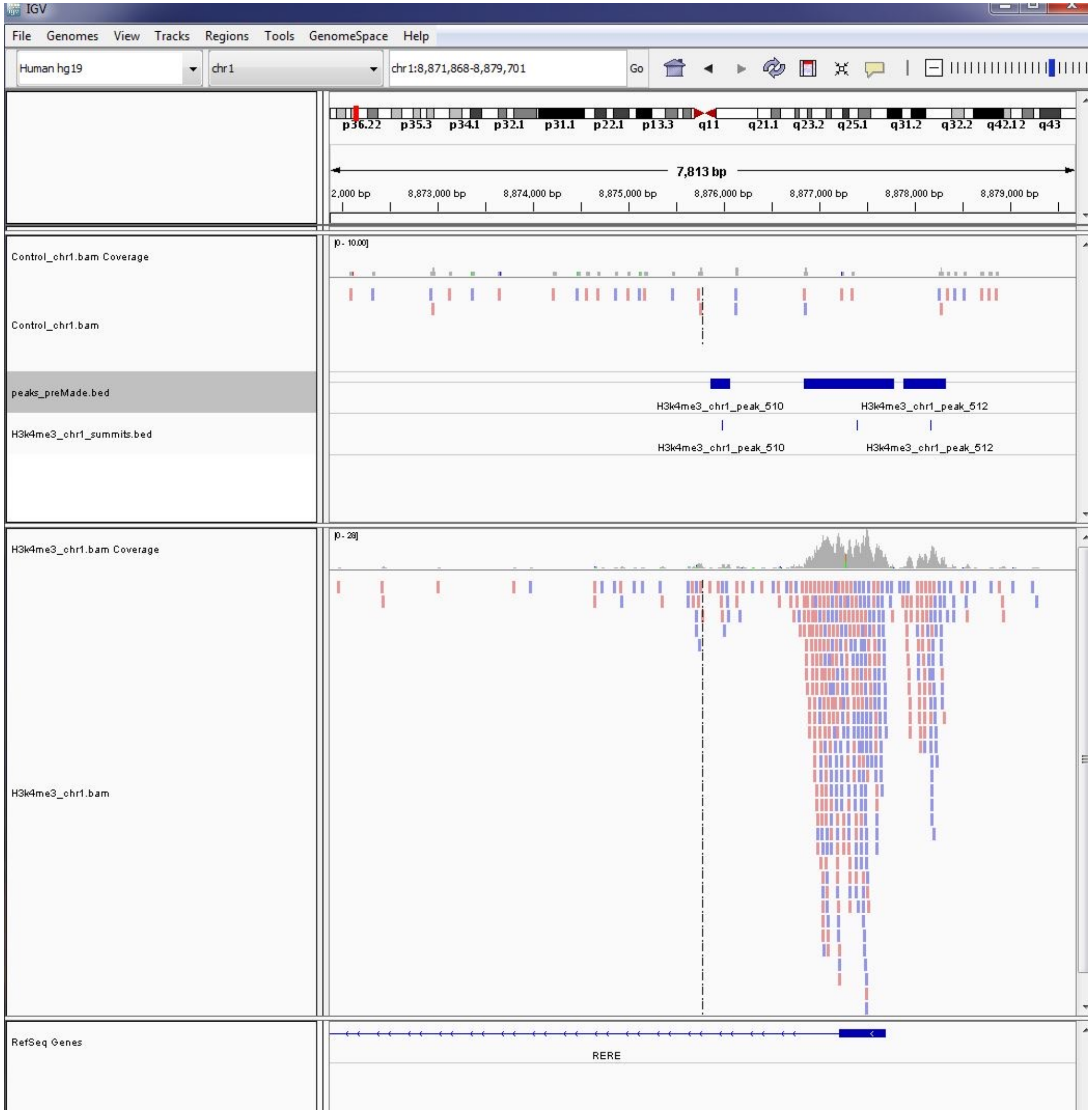
Visualize peaks in IGV

control

peaks

summit

H3K4me3


IGV
File Genomes View Tracks Regions Tools GenomeSpace Help
Human hg19
chr1
chr1:8,871,868-8,879,701
Go
p36.22 p35.3 p34.1 p32.1 p31.1 p22.1 p13.3 q11 q21.1 q23.2 q25.1 q31.2 q32.2 q42.12 q43
7,813 bp
2,000 bp 8,873,000 bp 8,874,000 bp 8,875,000 bp 8,876,000 bp 8,877,000 bp 8,878,000 bp 8,879,000 bp
Control_chr1.bam Coverage
[0 - 10.00]
Control_chr1.bam
peaks_preMade.bed
H3k4me3_chr1_peak_510
H3k4me3_chr1_peak_512
H3k4me3_chr1_summits.bed
H3k4me3_chr1_peak_510
H3k4me3_chr1_peak_512
H3k4me3_chr1.bam Coverage
[0 - 28]
H3k4me3_chr1.bam
RefSeq Genes
RERE

Comparing ChIP-seq across samples

i.e. Co-localization or differential binding

To compare two samples you can use :

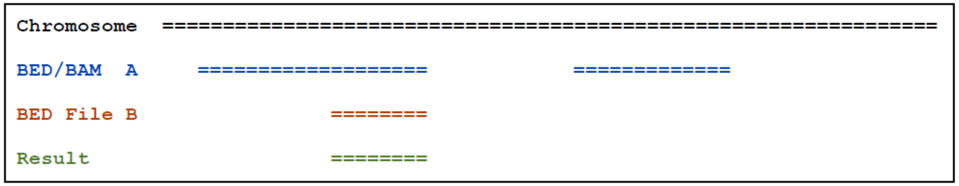
1. intersectBed (finds the subset of peaks **common** in 2 samples **or unique** to one them)
2. macs2 bdgdiff (find peaks present only in one of the samples)

If more than 2 samples follow:

/nfs/BaRC_Public/BaRC_code/Perl/compare_bed_overlaps

Linking peaks to genes: Bed tools

intersectBed



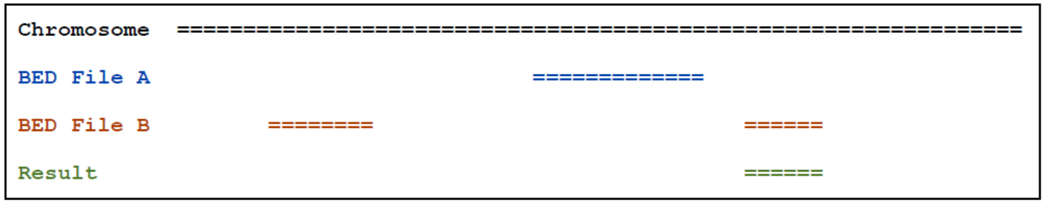
```
Chromosome  ================================================================

BED/BAM  A     ===================            =============

BED File B                ========

Result                    ========
```

closestBed



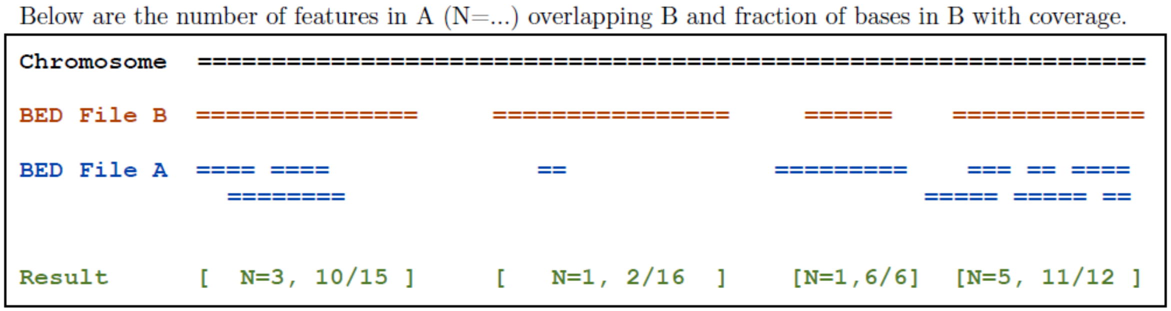
```
Chromosome  =================================================================

BED File A                             =============

BED File B         ========                                ======

Result                                                     ======
```

coverageBed



Below are the number of features in A (N=...) overlapping B and fraction of bases in B with coverage.

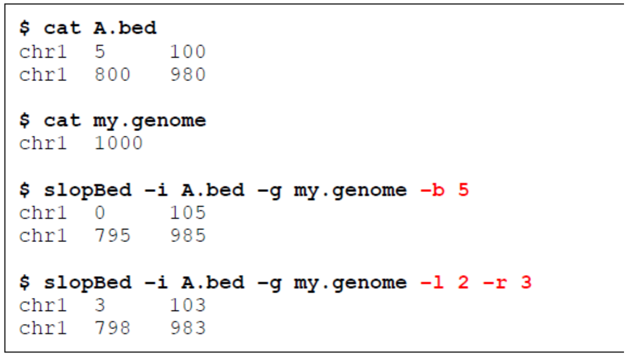
```
Chromosome  ================================================================

BED File B  ===============     ================    ======    =============

BED File A  ==== ====              ==               =========    === == ====
              ========                                         ===== ====== ==


Result      [  N=3, 10/15 ]     [   N=1, 2/16   ]   [N=1,6/6]  [N=5, 11/12 ]
```

slopBed



```
$ cat A.bed
chr1   5      100
chr1   800    980

$ cat my.genome
chr1   1000

$ slopBed -i A.bed -g my.genome -b 5
chr1   0      105
chr1   795    985

$ slopBed -i A.bed -g my.genome -l 2 -r 3
chr1   3      103
chr1   798    983
```

groupBy

It groups rows based on the value of a given column/s and it summarizes the other columns

Linking peaks to nearby genes:

Get all the genes at a certain distance (*i.e.* 3Kb) of the peak.
The distance we use depends on the area where we want to find regulatory interactions.

1. Take all genes and add 3Kb up and down with slopBed

   1.
      ```
      slopBed -b 3000 -i GRCh37.p13.HumanENSEMBLgenes.bed -g /nfs/genomes/human_gp_feb_09_no_random/anno/chromInfo.txt > HumanGenesPlusMinus3kb.bed
      ```
2. Intersect the slopped genes with peaks and get the list of unique genes overlapping

   ```
   intersectBed -wa  -a HumanGenesPlusMinus3kb.bed  -b peaks.bed | awk '{print $4}' | sort -u > Genesat3KborlessfromPeaks.txt
   ```

```
intersectBed -wa  -a HumanGenesPlusMinus3kb.bed  -b peaks.bed | head -3
chr1	45956538	45968751	ENSG00000236624_CCDC163P
chr1	45956538	45968751	ENSG00000236624_CCDC163P
chr1	51522509	51528577	ENSG00000265538_MIR4421
```

Link peaks to closest gene

For each region find the closest gene and filter based on the distance to the gene

The commands below are an example where we are looking for interactions at 3Kb or less.

```
closestBed -d -a  peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed |head
chr1	20870	21204	H3k4me3_chr1_peak_1	5.77592	chr1	14363	29806	ENSG00000227232_WASH7P	0
chr1	28482	30214	H3k4me3_chr1_peak_2	374.48264	chr1	29554	31109	ENSG00000243485_MIR1302-10	0
chr1	28482	30214	H3k4me3_chr1_peak_2	374.48264	chr1	14363	29806	ENSG00000227232_WASH7P	0
#the next two steps can also be done on excel
```

```
closestBed -d -a  peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed | groupBy -g 9,10  -c
    6,7,8, -o distinct,distinct,distinct | head -3
ENSG00000227232_WASH7P	0	chr1	14363	29806
ENSG00000243485_MIR1302-10	0	chr1	29554	31109
ENSG00000227232_WASH7P	0	chr1	14363	29806
```

```
closestBed -d -a  peaks.bed -b GRCh37.p13.HumanENSEMBLgenes.bed | groupBy -g 9,10  -c 6,7,8, -o
    distinct,distinct,distinct | awk  'BEGIN {OFS="\t"}{ if ($2<3000)  {print $3,$4,$5,$1,$2} } ' | head -5
chr1	14363	29806	ENSG00000227232_WASH7P	0
chr1	29554	31109	ENSG00000243485_MIR1302-10	0
chr1	14363	29806	ENSG00000227232_WASH7P	0
chr1	134901	139379	ENSG00000237683_AL627309.1	0
chr1	135141	135895	ENSG00000268903_RP11-34P13.15	0
```

Link peaks to closest gene (1 command)

For each region find the closest gene and filter based on the distance to the gene

The command below is an example where we are looking for interactions at 3Kb or less.

```
closestBed -d -a  peaks.bed -b
GRCh37.p13.HumanENSEMBLgenes.bed | groupBy -g 9,10  -c 6,7,8,
-o distinct,distinct,distinct | awk  'BEGIN {OFS="\t"}{ if ($2<3000)
{print $3,$4,$5,$1,$2} }'  > closestGeneAt3KborLess.bed
```

closestBed

-d print the distance to the feature in -b

groupBy

-g columns to group on

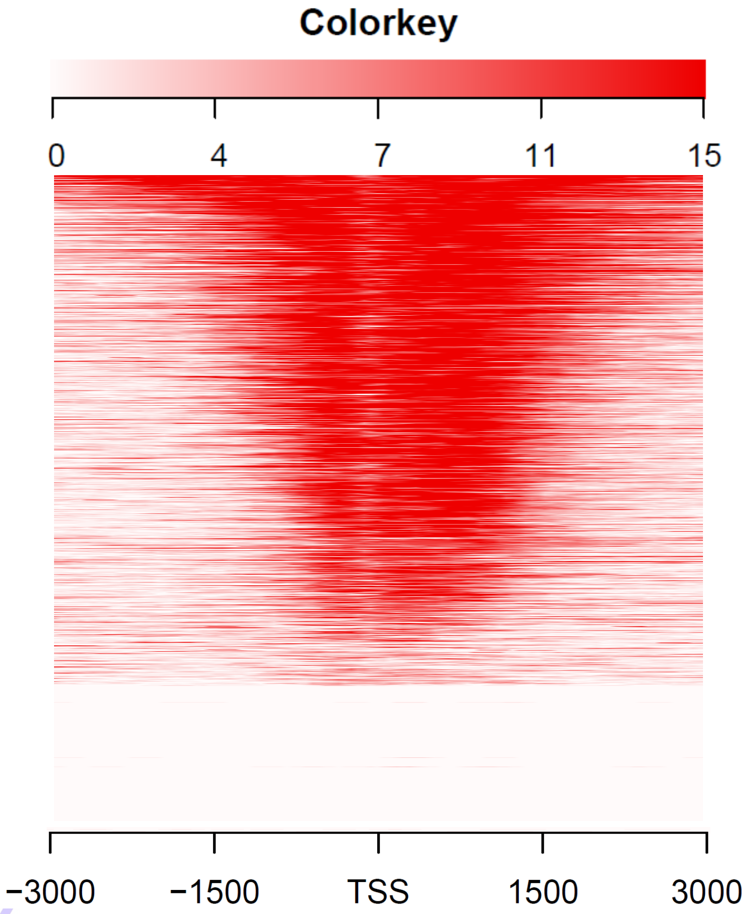
-c columns to summarize

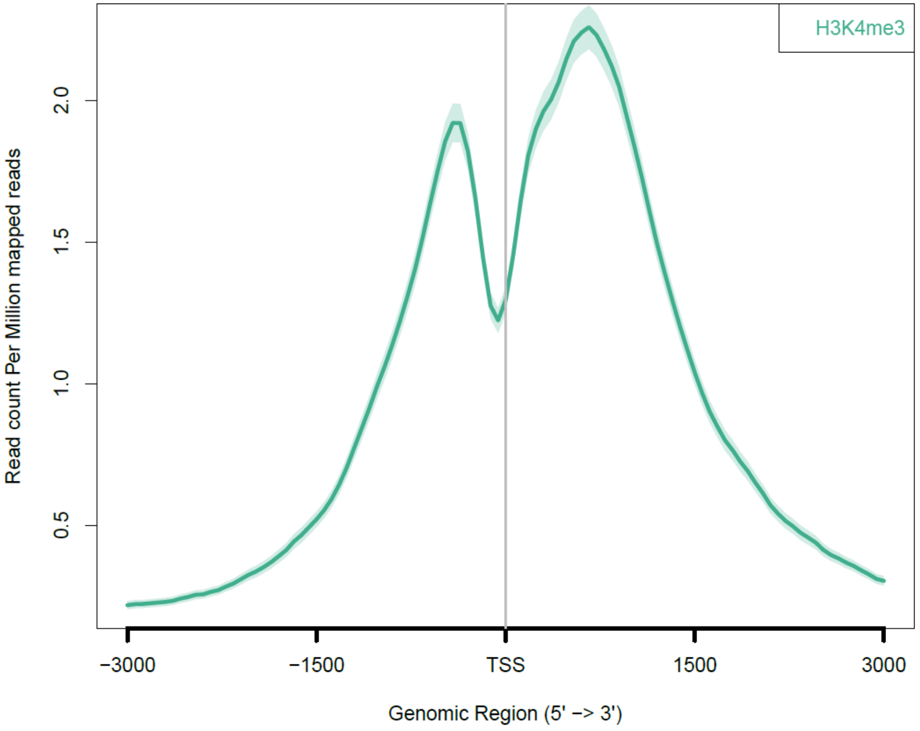
-o operation to use to summarize

Visualizing ChIP-seq reads with ngsplot

See Hot Topics: ngsplot

```
bsub ngs.plot.r -G hg19  -R tss -C H3k4me3_chr1.bam  -O H3k4me3_chr1.tss -T H3K4me3 -L 3000 -FL 300
```

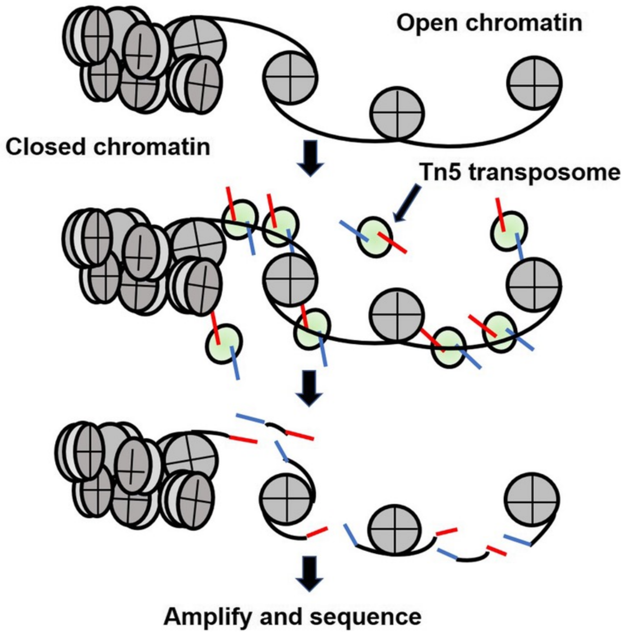




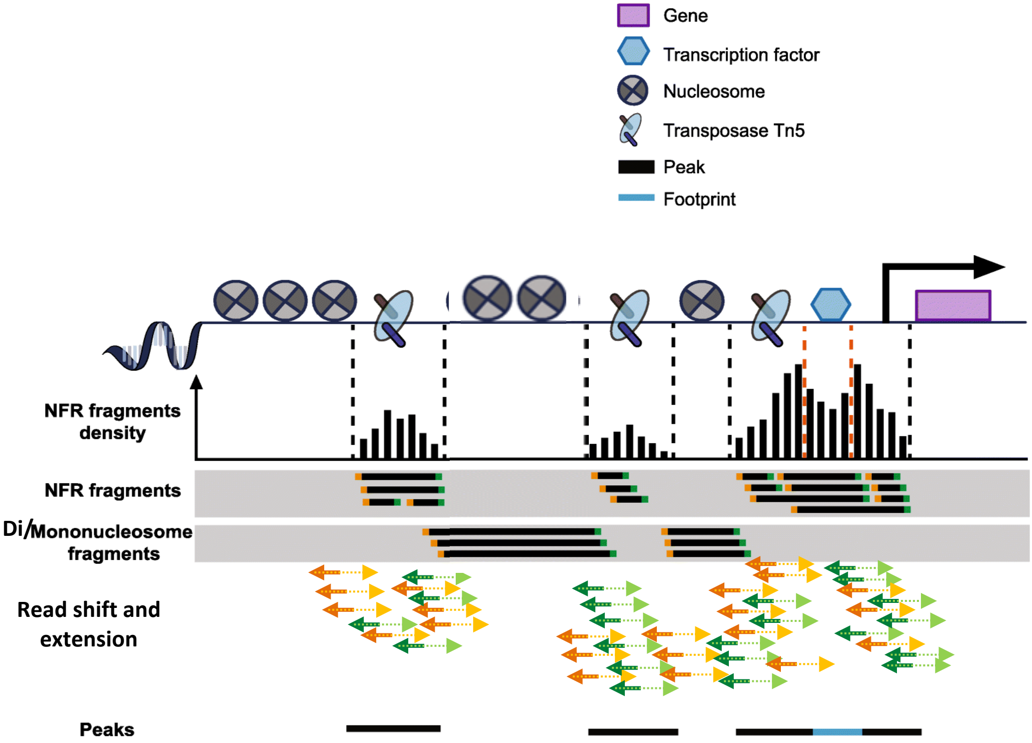
Exercises

- \\wi-files1\BaRC_Public\Hot_Topics\ChIPseq_2022\ChIPseq_2022.commands.txt

Profile generation in ATAC-seq



Ma, S., Zhang, Y. *Mol Biomed* **1,** 9 (2020)



Modified from Feng Yan et.al Genome Biology 21 (2020)

- Open chromatin
- Motif enrichment
- TF footprint
- Nucleosome positions

ATAC-seq pipeline goal

- Using short-read sequencing, identify genome-wide regions of open chromatin

ATAC-seq pipeline

1. Pre-alignment quality control
2. Aligning reads to genome
3. Post-alignment filtering
4. Post-alignment quality control
5. Peak (accessible regions) calls
6. Assessing Peak Calls with FRiP score (same as ChIP-seq)
7. Blacklist filtering for peaks (same as ChIP-seq)

ENCODE ATAC-seq pipeline:

- Includes all the steps in a single run

  Instructions to run pipeline using Whitehead server
- Human and mouse with biological replicates
- QC report (a sample report)

What sequencing works best for ATAC-seq?

- Read depth recommended by ENCODE:

  25 million non-duplicate, non-mitochondrial aligned-reads for single-end sequencing

  50 million for paired-ended sequencing (25 million fragments)

- No input control sample
- Shorter reads lengths (50x50 or 75x75) better than longer reads (100x100 or longer)
- Pair-end reads in recommended over single reads

Pre-alignment quality control

1. Check reads quality with fastqc

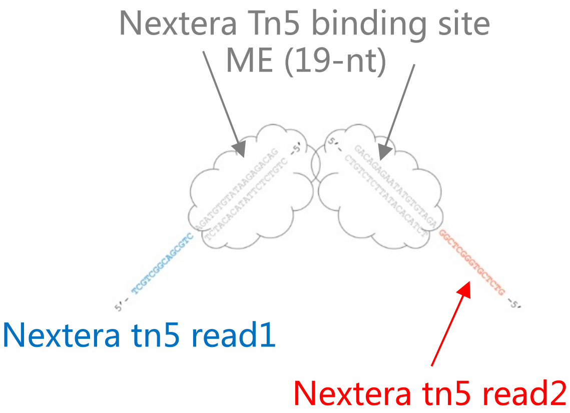
```
fastqc read1.fq read2.fq
```

2. Remove adapter when necessary:

```
trim_galore --fastqc -nextera --paired --length 30 read1.fq read2.fq
```

- -nextera: ATAC-seq experiments use the Nextera DNA Library Prep Kit
- --paired: both reads need to pass or they are both removed.
- --length: discard trimmed reads shorter than this length.

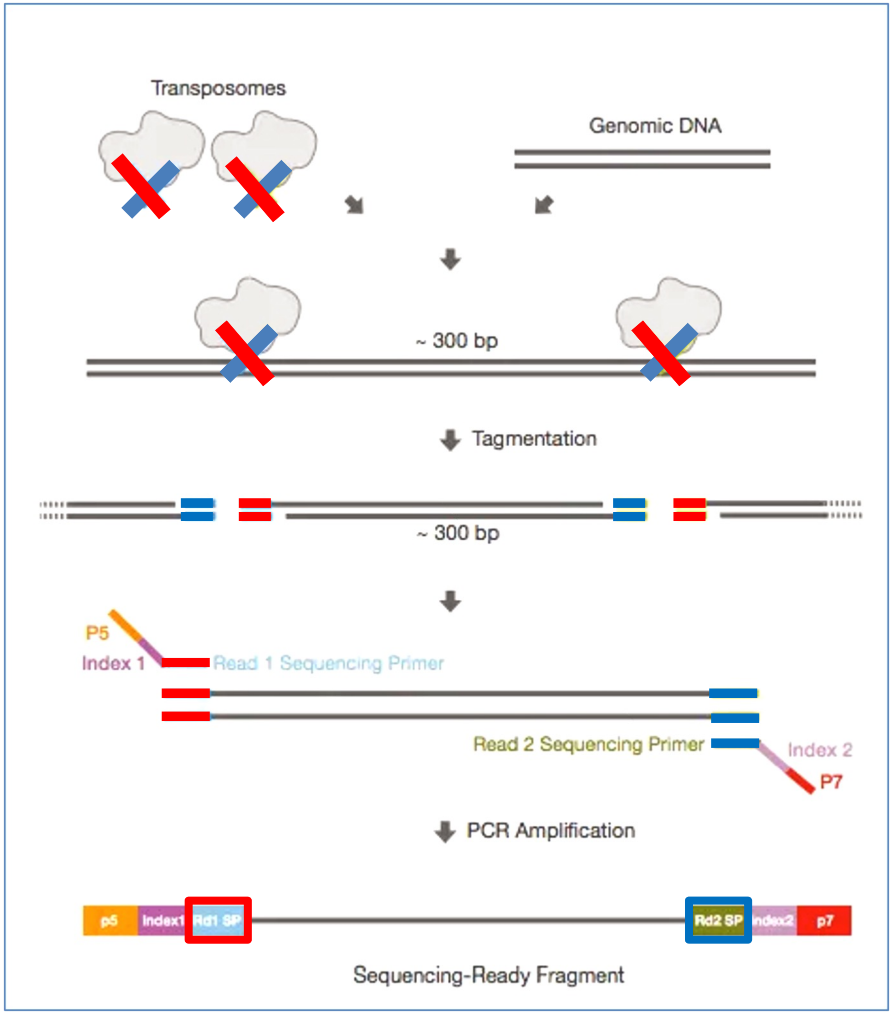
Nextera adapter in ATAC-seq


Nextera Tn5 binding site ME (19-nt)

Nextera tn5 read1

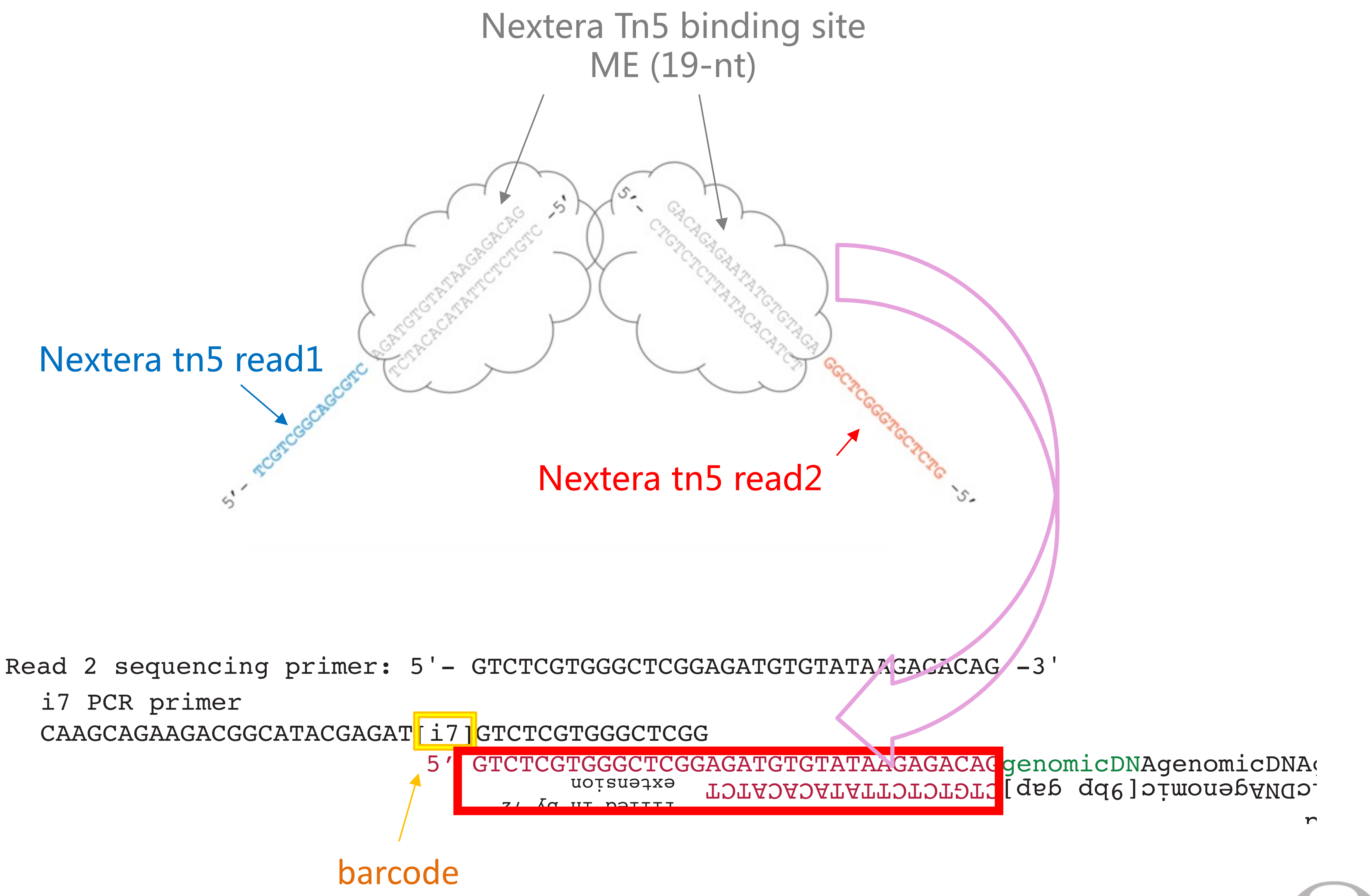
Nextera tn5 read2

5'- TCGTCGGCAGCGTC AGATGTGTATAAGAGACAG -5'
TCTACACATATTCTCTGTC
5'- GACAGAGAATATGTGTAGA
CTGTCTCTTATACACATCT GGCTCGGGTGCTCTG -5'



Modified from https://www.abpbio.com/product/tn5-transposase/

Nextera adapter in ATAC-seq



```
Read 2 sequencing primer: 5'- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG -3'
  i7 PCR primer
  CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG
                          5' GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGgenomicDNAgenomicDNA
```

Modified from https://www.plob.org/article/11443.html

Map reads to reference genome

Map reads with a non-spliced mapping tools: bowtie2 or BWA

```
bowtie2 --very-sensitive --no-discordant -p 2 -X 2000 -x hg38 -1
read1.fq -2 read2.fq | samtools view -ub - | samtools sort - >|
bowtie_out.bam
```

- --no-discordant:
  - Suppress discordant alignments for paired reads

- -X 2000:
  - Increase maximum fragment length to 2k to include nucleosome distribution
  - Used for plotting fragment size distribution

Post alignment filtering

- Remove reads with low quality score: MAPQ < 30

```
alignmentSieve -b file.bam --minMappingQuality 30 --
samFlagInclude 2 -o MAPQ30.bam"
```

- Remove duplicates with Picard's 'MarkDuplicates’

```
java -jar picard.jar MarkDuplicates I=foo.bam O=noDups.bam
M=foo.marked_dup_metrics.txt REMOVE_DUPLICATES=true
```

- Remove reads mapped to mitochondria

```
samtools view -h file.bam | grep -v chrM | samtools view -b -
h -f 0x2 - | samtools sort - > file.sorted.bam
```

-f 0x2
samFlagInclude 2
} keep only properly paired reads

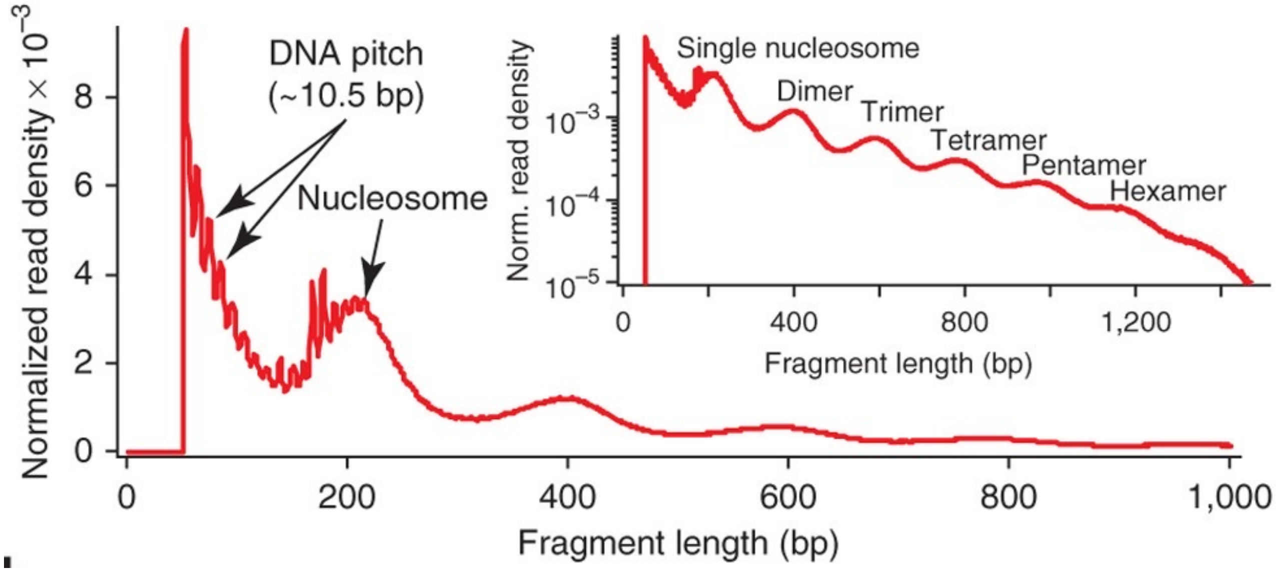
Post-alignment quality control

- Fragment size distribution
- TSS enrichment score

Fragment size distribution

calculate_ATACseq_fragment_size_distribution.R

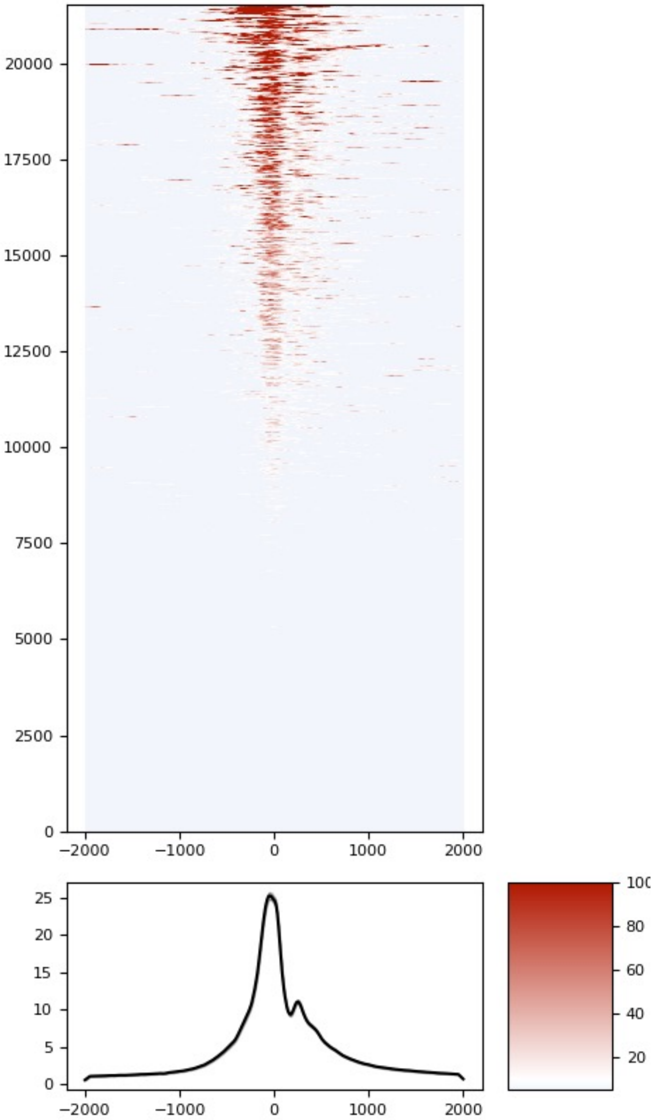
```
library("ATACseqQC")
pdf("sample.fragment_sizes.pdf", w=11, h=8.5)
fragSizeDist(”sample.bam", "sample")
dev.off()
```



Jason D Buenrostro, et.al. Nature Methods 10 (2013)

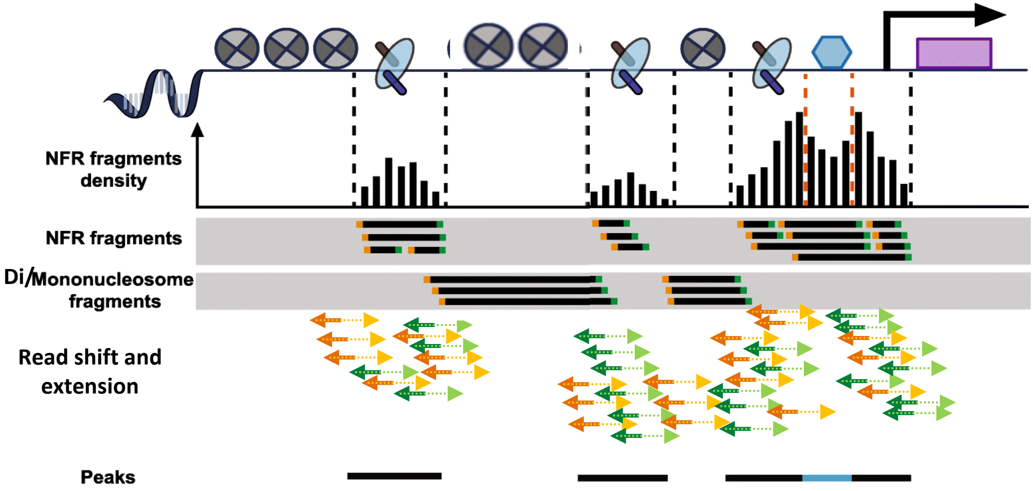
TSS enrichment score

```
calculate_TSS_enrichment_score.py
--outdir      ./tss
--outprefix   sample_tss
--fastq1       sample1_1.fq.gz
–tss               TSS.bed
--chromsizes chromInfo.txt
--bam            sample.bam  >
sample_TSS_enrichment_score.txt
```



TSS enrichment values depends on TSS annotation.
Score with TSS per transcript is smaller than score with one TSS per gene
ENCODE standard with one TSS per gene:
[*https://www.encodeproject.org/atac-seq/#standards*](https://www.encodeproject.org/atac-seq/#standards)

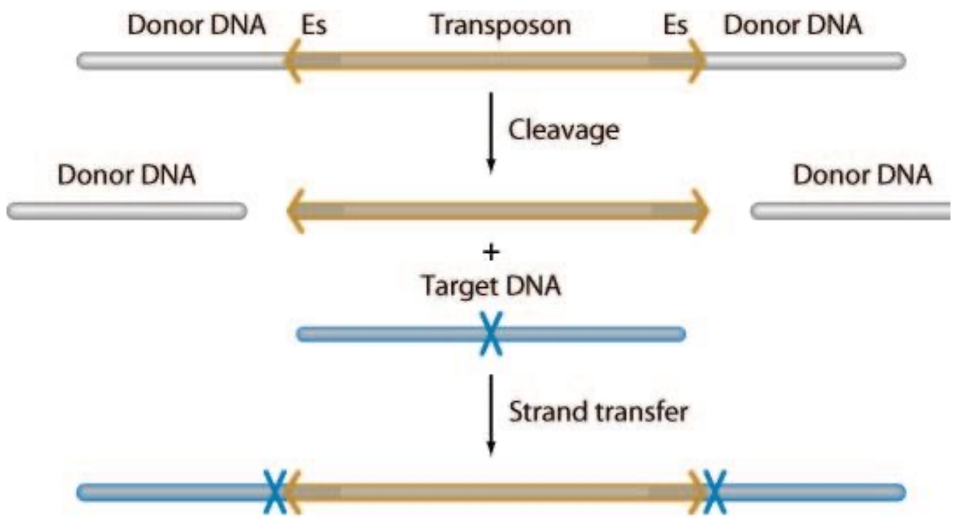
Call Peaks



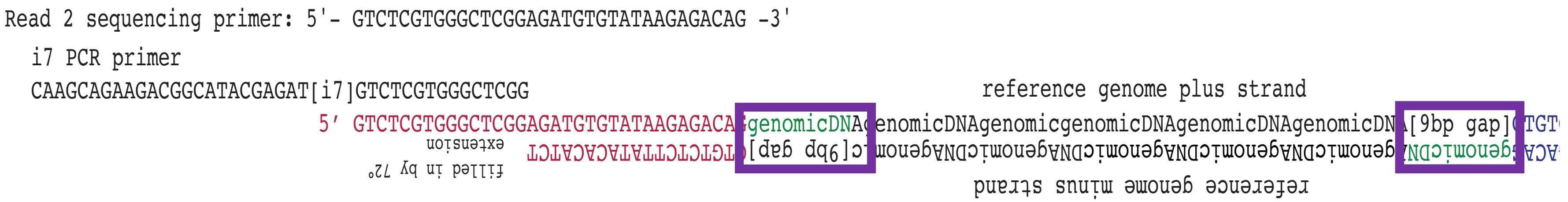
Modified from Feng Yan et.al Genome Biology 21 (2020)

Create a signal profile centered around the cutting sites

9-nt duplication



Reznikoff WS. 2008.
Annu. Rev. Genet. 42:269–86



```
Read 2 sequencing primer: 5'- GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG -3'
  i7 PCR primer
  CAAGCAGAAGACGGCATACGAGAT[i7]GTCTCGTGGGCTCGG
                                                                      reference genome plus strand
                        5' GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGgenomicDNAgenomicDNAgenomicgenomicDNAgenomicDNAgenomicDNA[9bp gap]CTGT
```

Upside-down lines:

```
ACAgenomicDNAgenomicDNAgenomicDNAgenomicDNAgenomicDNAgenomic[9bp gap]GTCTCTTATACACATCT
extension
filled in by 72°
reference genome minus strand
```

Solution: shift read + 4 bp and – 5 bp for positive and negative strand respectively

Convert bam to bed

- Consider both reads of a pair:

```
bedtools bamtobed -i foo.bam > foo.bed
```

- Convert an input paired-end alignment:

```
K00168:88:HFF7YBBXX:1:1203:28371:20304 163 chr1 3199165 42 51M = 3199703 589
ACTAAAAACAGACAAATGCTCAACATTTACATGAAATGTAAGACTAAATAT AA-FFFJA-FAJJJ<--FA-
7AJ7-AJJJF-<77AJ-A<F<FJA7FJJJJA MD:Z:51 PG:Z:MarkDuplicates XG:i:0 NM:i:0 XM:i:0
XN:i:0 XO:i:0 AS:i:0 YS:i:0 YT:Z:CP
```

```
K00168:88:HFF7YBBXX:1:1203:28371:20304 83 chr1 3199703 42 51M = 3199165 -589
ACTTTGAAAAAAATGAGTAACAGACTTCTGTTAAAATGACCACAGTGTACT
JJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJJFFFAA MD:Z:51 PG:Z:MarkDuplicates XG:i:0 NM:i:0
XM:i:0 XN:i:0 XO:i:0 AS:i:0 YS:i:0 YT:Z:CP
```

- To bed output file:

```
chr1 3199164 3199215 K00168:88:HFF7YBBXX:1:1203:28371:20304/2 42 +
chr1 3199702 3199753 K00168:88:HFF7YBBXX:1:1203:28371:20304/1 42 -
```

Shift reads

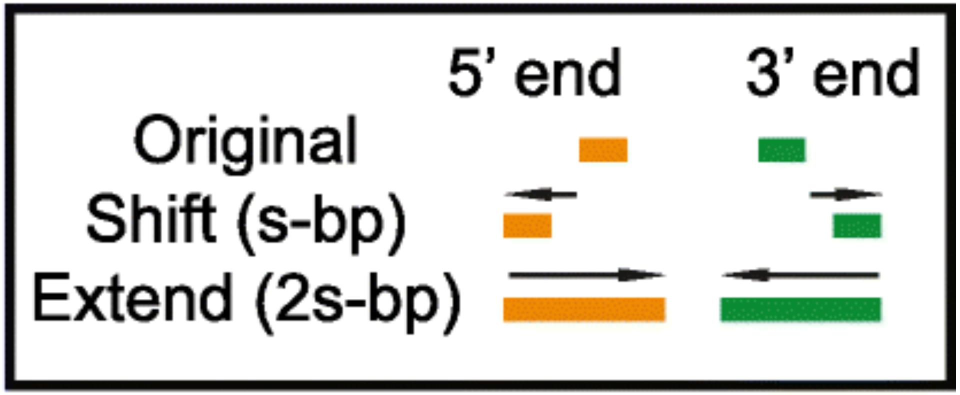
- Reads should be shifted + 4 bp and – 5 bp for positive and negative strand respectively, to account for the 9-bp duplication created by DNA repair of the nick by Tn5 transposase

```
cat foo.bed | awk -F $'\t' 'BEGIN {OFS = FS}{ if ($6 == "+") {$2 = $2 + 4} else if ($6 == "-") {$3 = $3 - 5} print $0}' >| foo_tn5.bed
```

Call peaks

```
macs2 callpeak -t foo_tn5.bed -n foo -f BED -g mm -q 0.01 --nomodel --shift -75 --extsize 150 --call-summits --keep-dup all
```

Shift the reads to create a signal profile centered around the cutting sites



Feng Yan et.al Genome Biology 21 (2020)

FRiP (Fraction of reads in peaks) score

- Fraction of all mapped reads that fall into the called peak regions
- The higher the score, the better

  According to ENCODE, score is preferably over 0.3, values greater than 0.2 are acceptable.

```
calculate_FRiP_score.py Sample.bam Sample_peaks.narrowPeak
```

Blacklist filtering for peaks

- Anomalous, unstructured, or high signal in next-generation sequencing experiments independent of cell line or experiment
- Samples from human, mouse, fly, or C. elegans
- Can be downloaded from ENCODE

```
bedtools intersect -v -a foo_peaks.narrowPeak -b
blacklist.bed > bfiltered_peaks.narrowPeak
 #  -v option: only report those "A" peaks with no overlaps in "B"
```

Identify footprints

- High depth of coverage:

  At least ~200 million of reads

- Shift reads to account for the 9-bp duplication

- Tools:

(Feng Yan et.al Genome Biology 21, 2020)

HINT-ATAC:

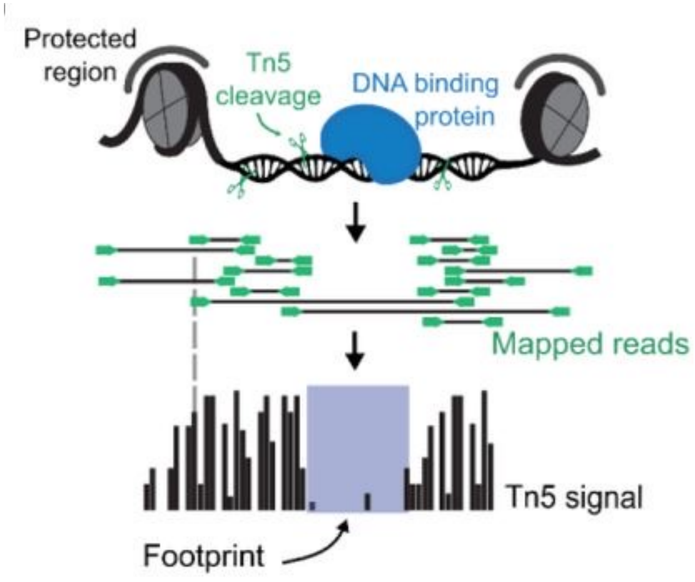
De novo

CENTIPEDE:

Motif-centric tools

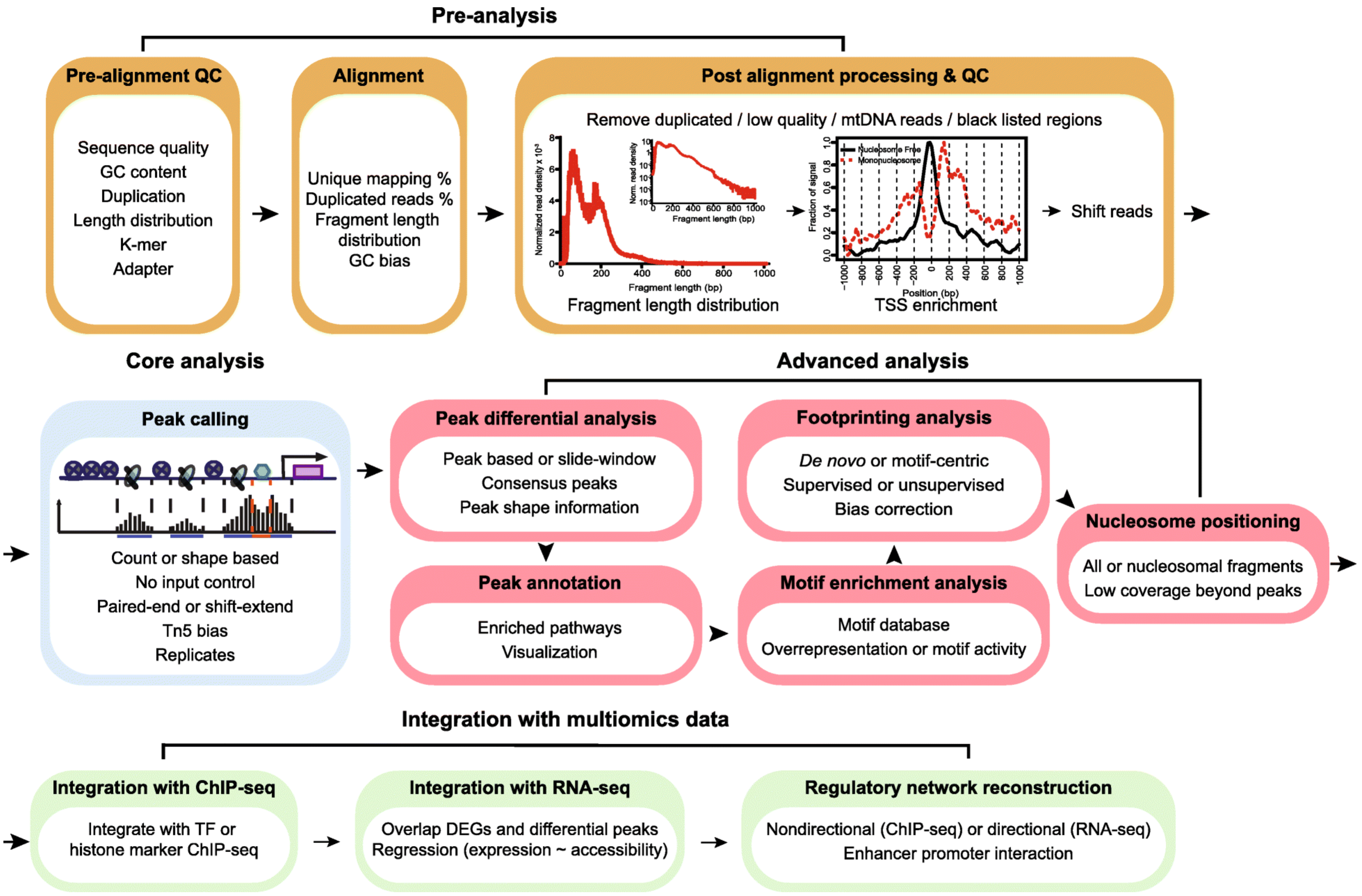
BaGFoot:

global TF footprint changes



Bentsen, M., Goymann, P., Schultheis, H. *et al. Nat Commun* 11 (2020)

ATAC-seq analysis



Pre-analysis

Pre-alignment QC

Sequence quality
GC content
Duplication
Length distribution
K-mer
Adapter

Alignment

Unique mapping %
Duplicated reads %
Fragment length distribution
GC bias

Post alignment processing & QC

Remove duplicated / low quality / mtDNA reads / black listed regions

Fragment length distribution

TSS enrichment

Shift reads

Core analysis

Peak calling

Count or shape based
No input control
Paired-end or shift-extend
Tn5 bias
Replicates

Advanced analysis

Peak differential analysis

Peak based or slide-window
Consensus peaks
Peak shape information

Peak annotation

Enriched pathways
Visualization

Footprinting analysis

De novo or motif-centric
Supervised or unsupervised
Bias correction

Motif enrichment analysis

Motif database
Overrepresentation or motif activity

Nucleosome positioning

All or nucleosomal fragments
Low coverage beyond peaks

Integration with multiomics data

Integration with ChIP-seq

Integrate with TF or histone marker ChIP-seq

Integration with RNA-seq

Overlap DEGs and differential peaks
Regression (expression ~ accessibility)

Regulatory network reconstruction

Nondirectional (ChIP-seq) or directional (RNA-seq)
Enhancer promoter interaction

Feng Yan et.al
Genome Biology 21 (2020)

Exercises:

- Mouse lung tissue postnatal (0 days) from ENCODE

  Randomly chosen 100,000 pair-end reads from the 1st replicate

- \\wi-files1\BaRC_Public\Hot_Topics\ ATACseq_2022\ ATAC-seq_2022_commands.txt

References

- Reviews and benchmark papers:

  ChIP-seq: advantages and challenges of a maturing technology *Nat Rev Genet 10, 2009*

  Computation for ChIP-seq and RNA-seq studies *Nat Methods* 6, 2009

  Practical Guidelines for the Comprehensive Analysis of ChIP-seq Data. *PLoS Comput. Biol*. 9(11), 2013

  A computational pipeline for comparative ChIP-seq analyses. *Nat Protoc* **7,** 2012

  ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res. 22(9), 2012*

  Identifying and mitigating bias in next-generation sequencing methods for chromatin biology

  Nature Reviews Genetics 15, 2014

  From reads to insight: a hitchhiker’s guide to ATAC-seq data analysis. *Genome Biol* 21, 2020

  Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods* 10, 2013
- ENCODE ATAC-seq guideline: https://github.com/ENCODE-DCC/atac-seq-pipeline
- Quality control and strand cross-correlation:

  http://code.google.com/p/phantompeakqualtools/
- MACS:

  Model-based Analysis of ChIP-Seq (MACS). *Genome Biol* 9, 2008 https://liulab-dfci.github.io/software/

  Using MACS to identify peaks from ChIP-Seq data. *Curr Protoc Bioinformatics*. 2011
- Picard Tools: https://broadinstitute.github.io/picard/
- Bedtools: https://code.google.com/p/bedtools/

  BEDTools: a flexible suite of utilities for comparing genomic features, *Bioinformatics 15, 2010*
- ngsplot: https://code.google.com/p/ngsplot/

  ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases *BMC Genomics* 15, 2014

Other resources

- Previous Hot Topics

  Quality Control

  http://barc.wi.mit.edu/education/hot_topics/NGS_QC_2017/slides4perPage.pdf

- SOPs

  http://barcwiki.wi.mit.edu/wiki/SOPs/chip_seq_peaks

  http://barcwiki.wi.mit.edu/wiki/SOPs/atac_Seq

Upcoming Hot Topics

- Enrichment Analysis (May)
- Clustering and Heatmaps (June)

Suggestions on topics of Hot Topics are welcome

Previous Hot Topics: Slides and Exercises

http://barc.wi.mit.edu/education/hot_topics/