

Introduction to ATAC-seq analysis

Shamith Samarajiwa

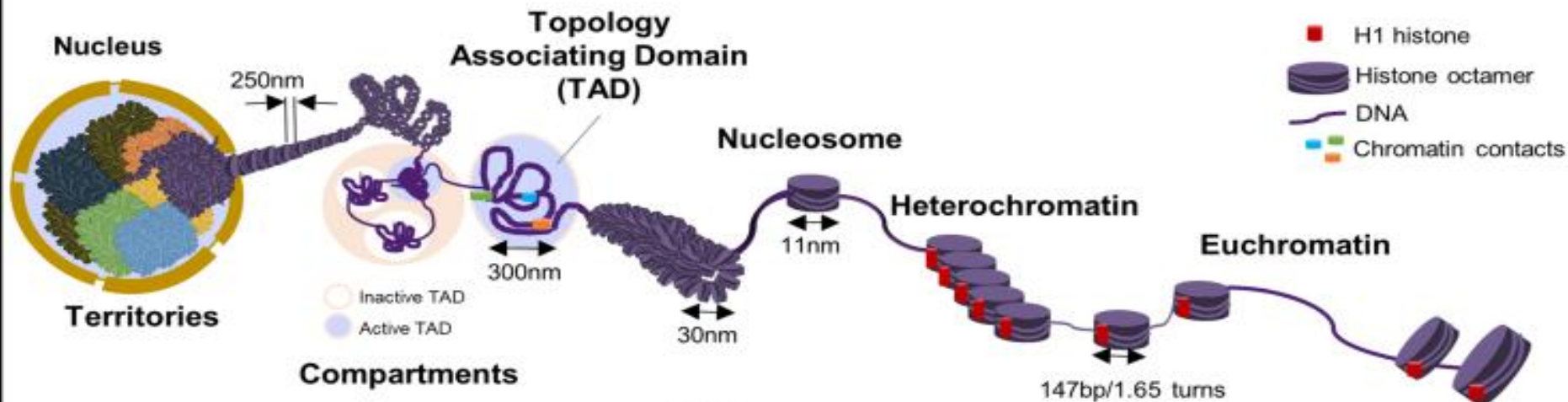
CRUK Summer School in Bioinformatics
July 2019



**UNIVERSITY OF
CAMBRIDGE**

Higher-order

Primary-order



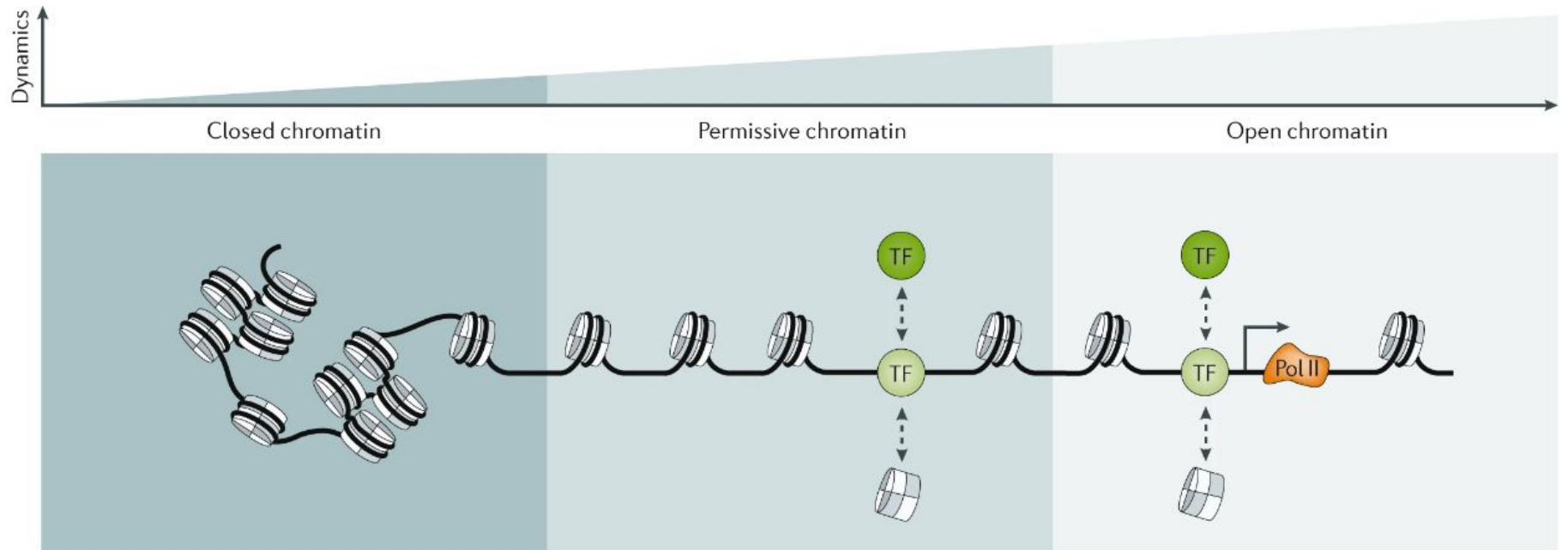
Techniques

| 4C | 5C | Hi-C | Validation | MNase-seq | DNase-seq | FAIRE-seq | ATAC-seq |
|------------------|---------------------|-------------------|--------------------|--------------------------------|-----------------------|-----------------------|------------------------------|
| <i>1-vs-Many</i> | <i>Many-vs-Many</i> | <i>All-vs-All</i> | ChIP, Imaging etc. | <i>Inferred closed regions</i> | <i>Open chromatin</i> | <i>Open chromatin</i> | <i>Open/Closed chromatin</i> |

Procedure

| Experimental | Computational | Experimental | Computational |
|------------------|-----------------------------|-------------------|-------------------|
| 1. Crosslinking | 1. Alignment | 1. Crosslinking* | 1. Size-selection |
| 2. Fragmentation | 2. Filtering | 2. Fragmentation | 2. Alignment |
| 3. Ligation | 3. Binning | 3. Size-selection | 3. Peak calling |
| 4. Detection | 4. Normalization | 4. Sequencing | 4. Normalization |
| | 5. Identifying interactions | | 5. Visualization |
| | 6. Visualization | | |

A Continuum of accessibility states



Reflects regulatory capacity of a cellular state


MENU ▾

nature | methods

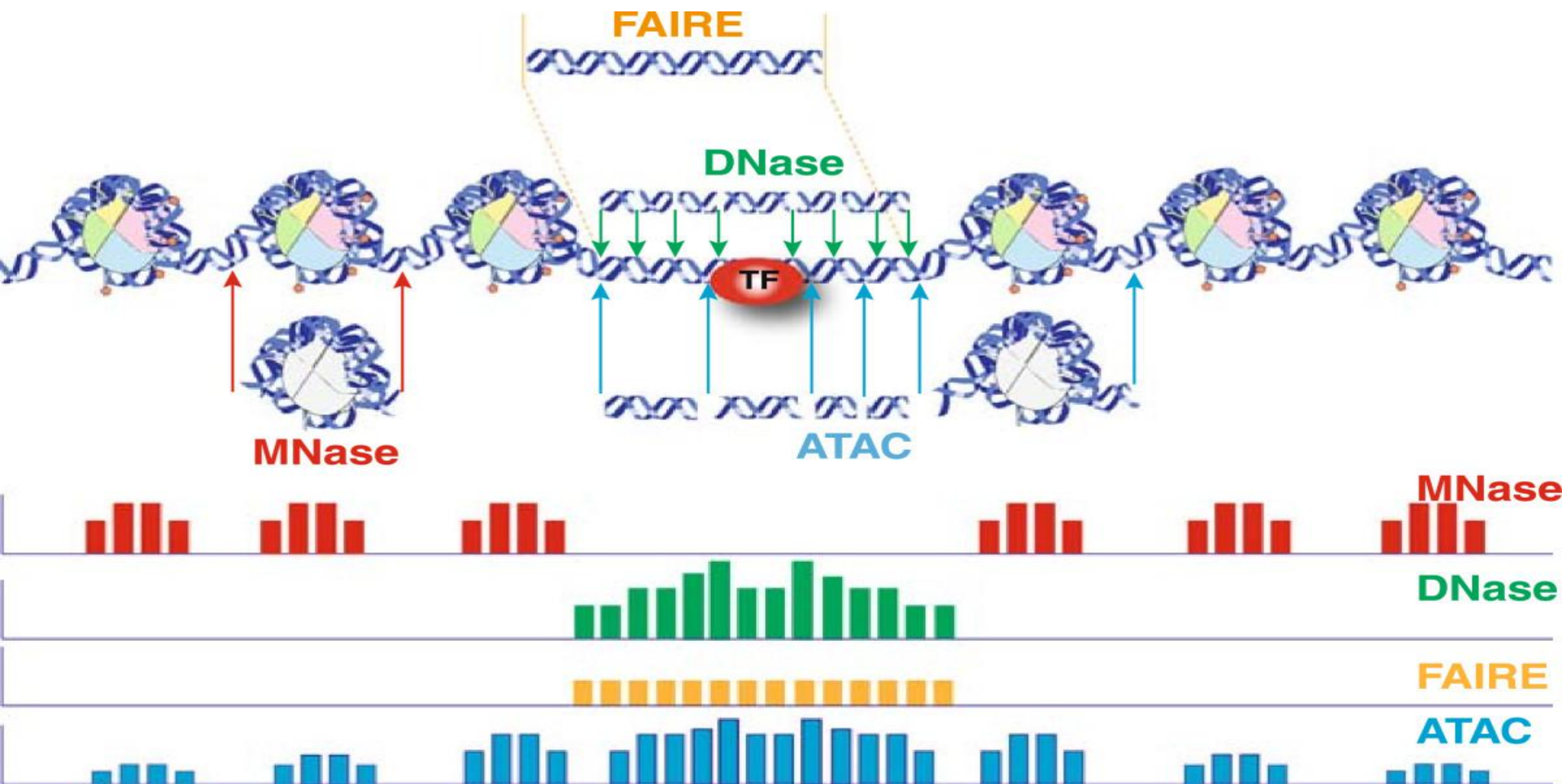
Article | Published: 06 October 2013

Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position

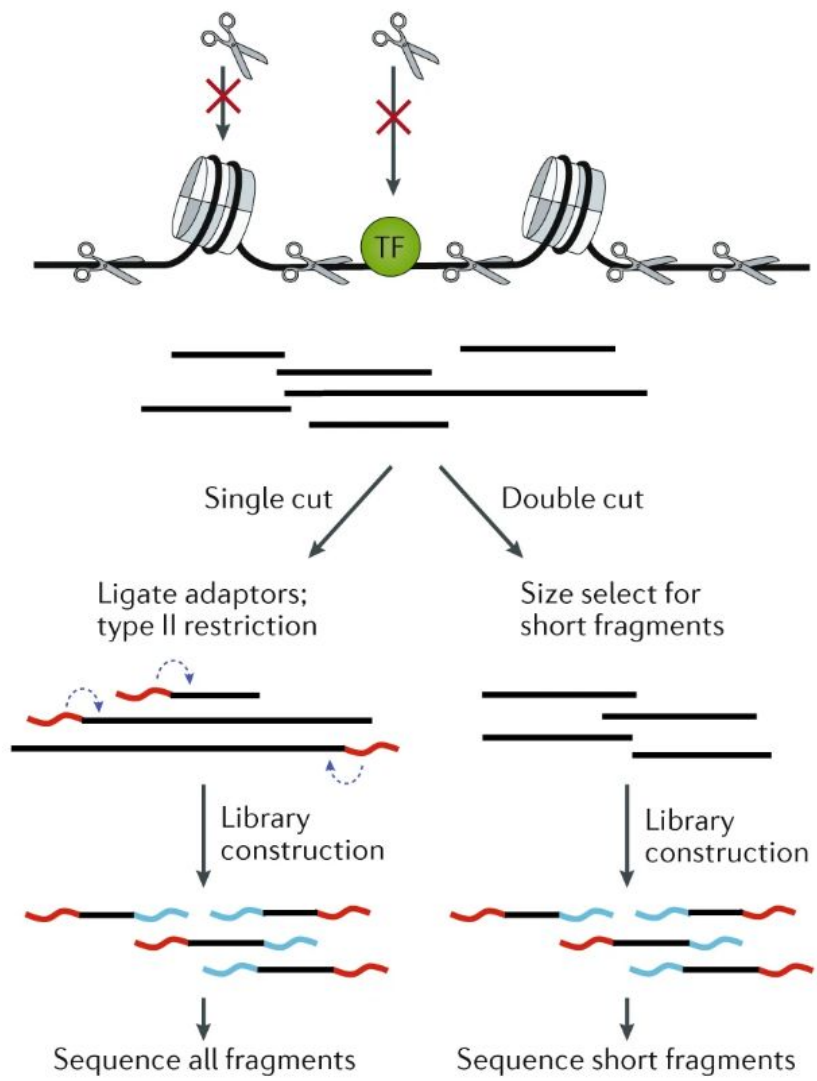
Jason D Buenrostro, Paul G Giresi, Lisa C Zaba, Howard Y Chang  & William J Greenleaf 

Nature Methods **10**, 1213–1218 (2013) | [Download Citation](#) 

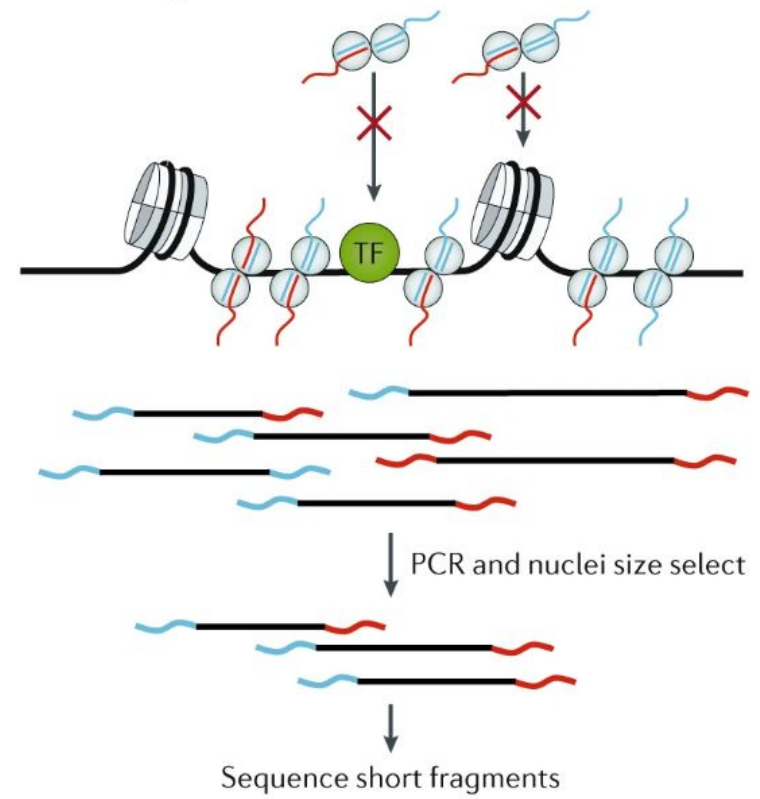
Assay for Transposase Accessible Chromatin



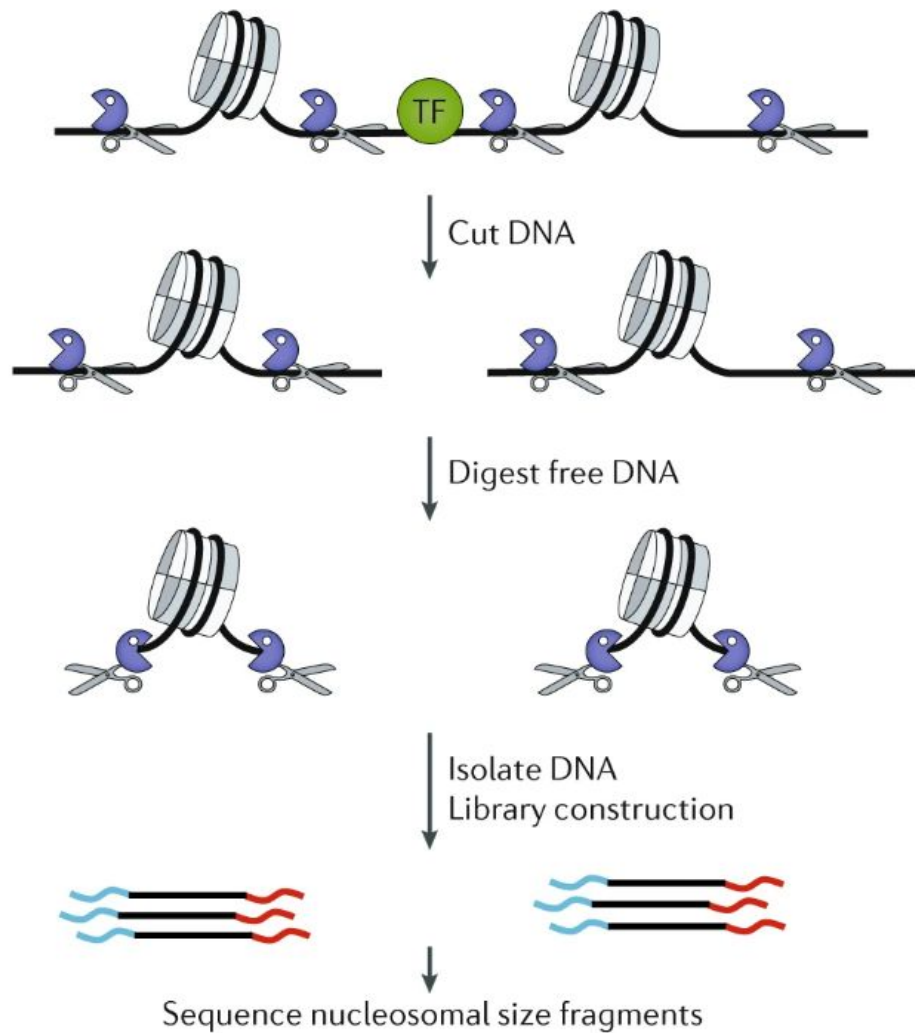
DNase-seq



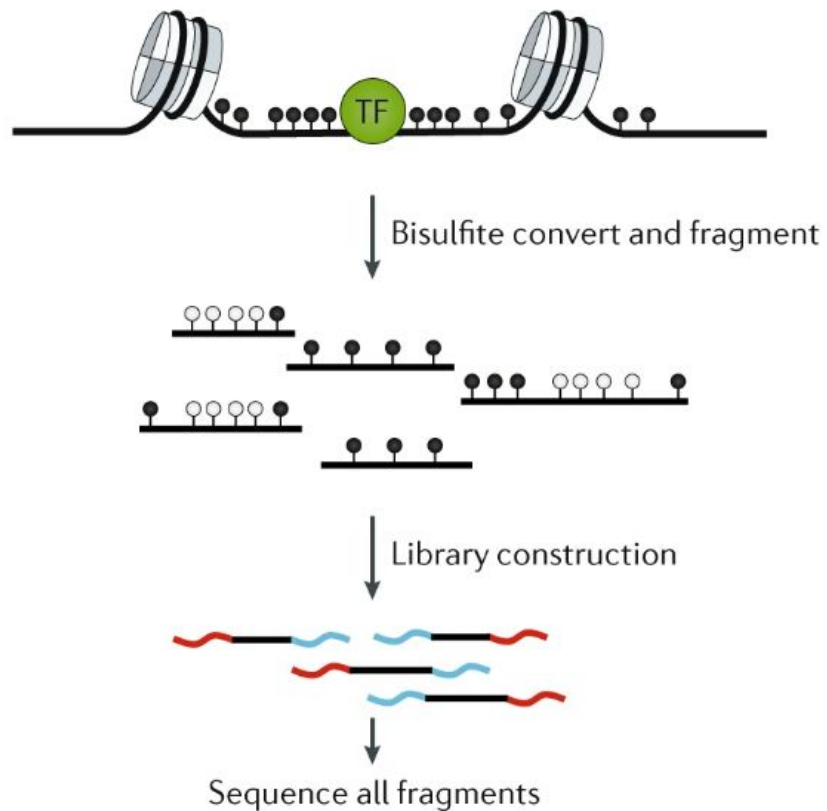
ATAC-seq



MNase-seq

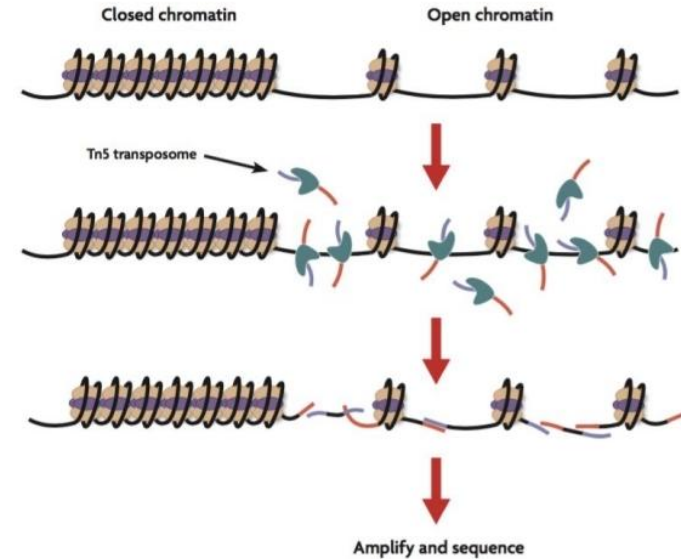


NOMe-seq

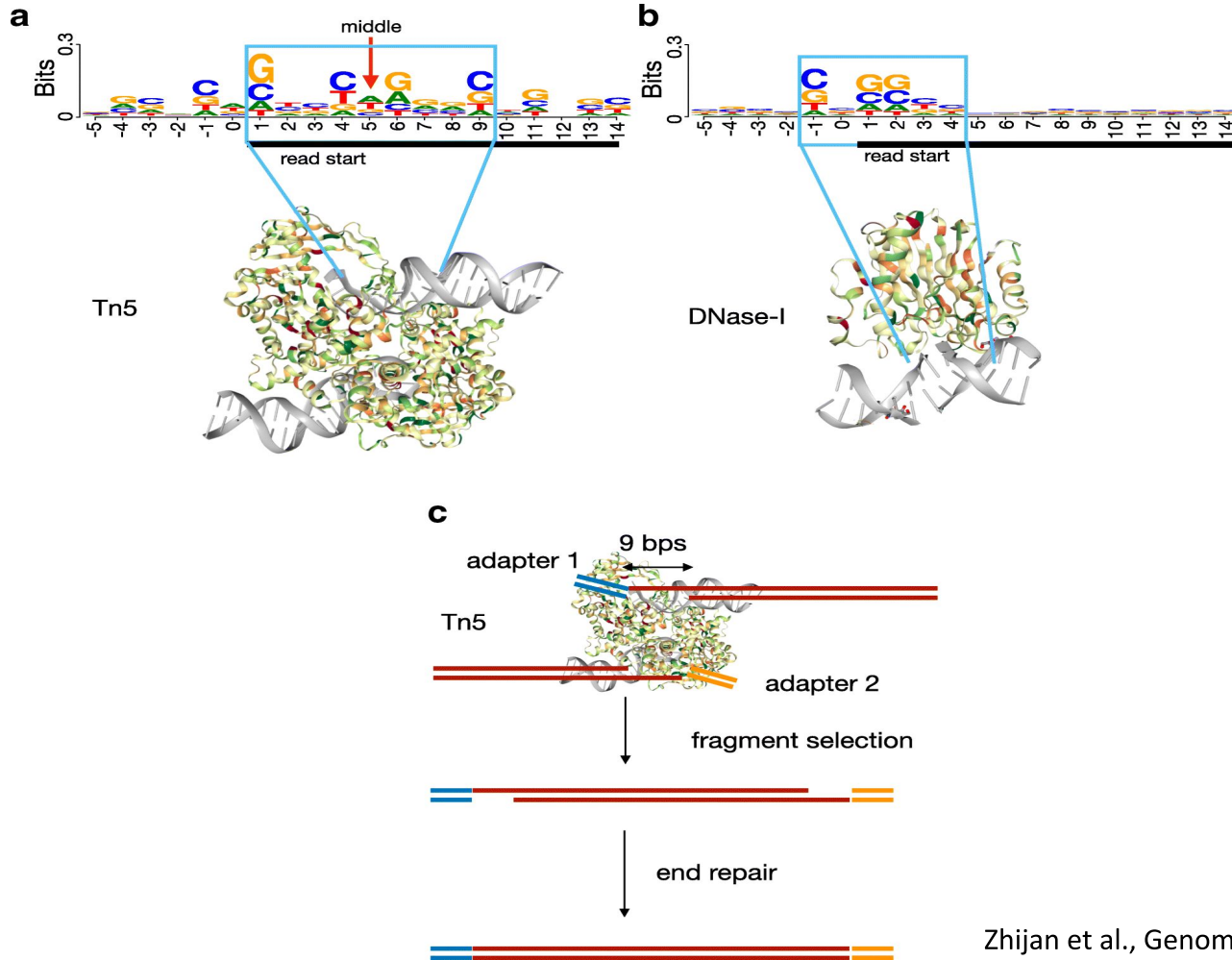


ATAC-seq

- Detects
 - Chromatin accessibility
 - Nucleosome occupancy (NFRs)
- Can also detect nucleosome packing, **positioning** and **TF footprints**.
- Does not require sonication and phenol chloroform extractions, antibodies or sensitive enzymatic digestions that can introduce potential bias.
- A hyperactive **Tn5 transposase** is used to fragment DNA and integrate into active regulatory regions.

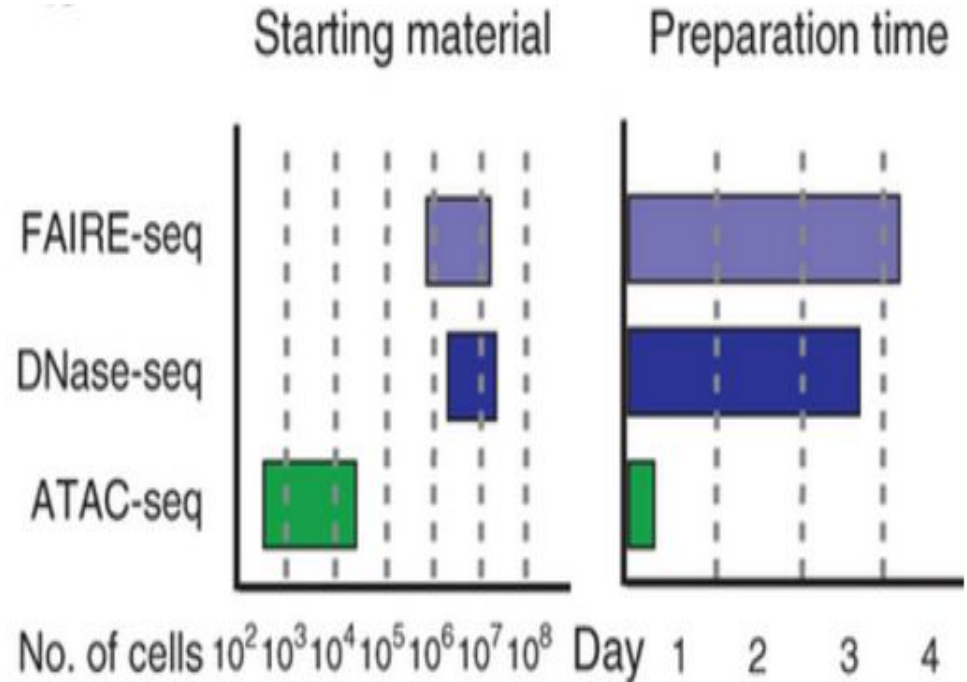


Tn5 adaptor insertion



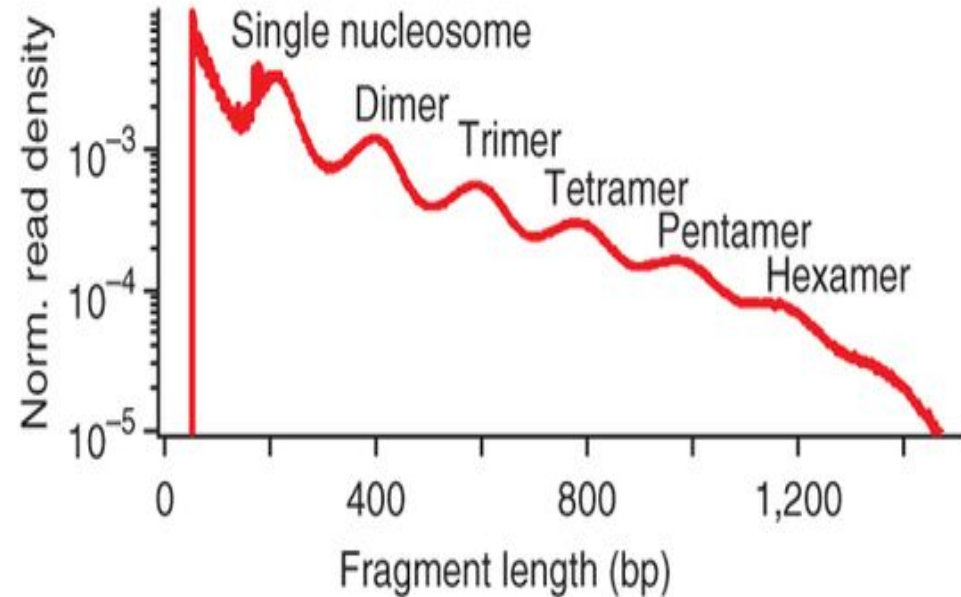
ATAC-seq

- The accessible genome comprises ~2–3% of total DNA sequence yet captures more than 90% of regions bound by TFs
- ATAC-seq is a two-step protocol
 - Insertion of Tn5 transposase with adaptors
 - PCR amplification
- Needs ~500-50,000 cells



Fragment size periodicity

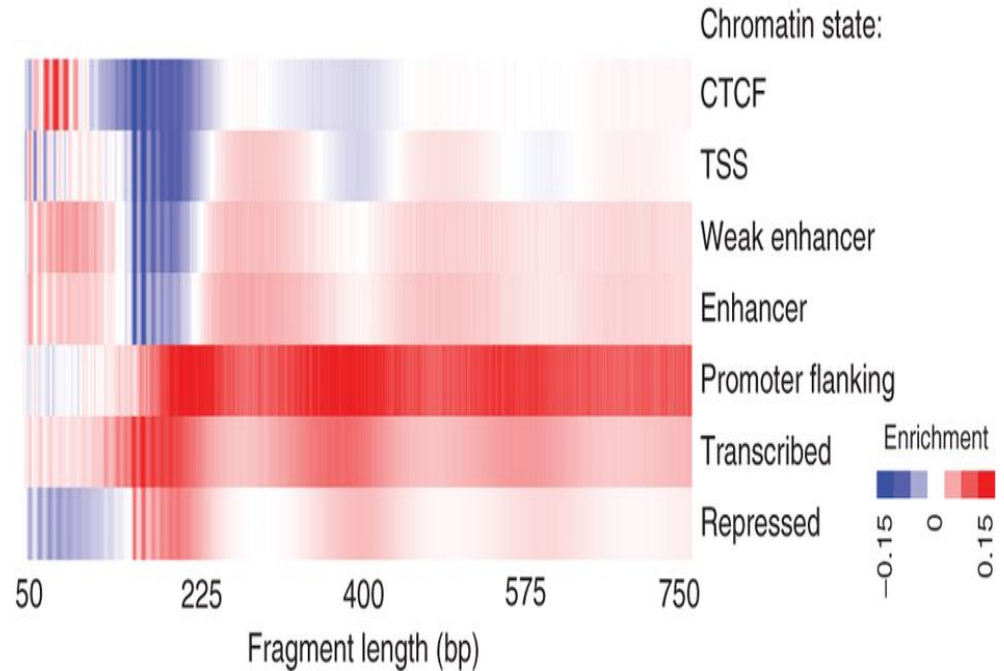
- Paired-end reads produce information about nucleosome positioning.
- Insert size distribution of fragments has a periodicity of ~ 200 bp, suggesting that fragments are protected by multiples of nucleosomes.



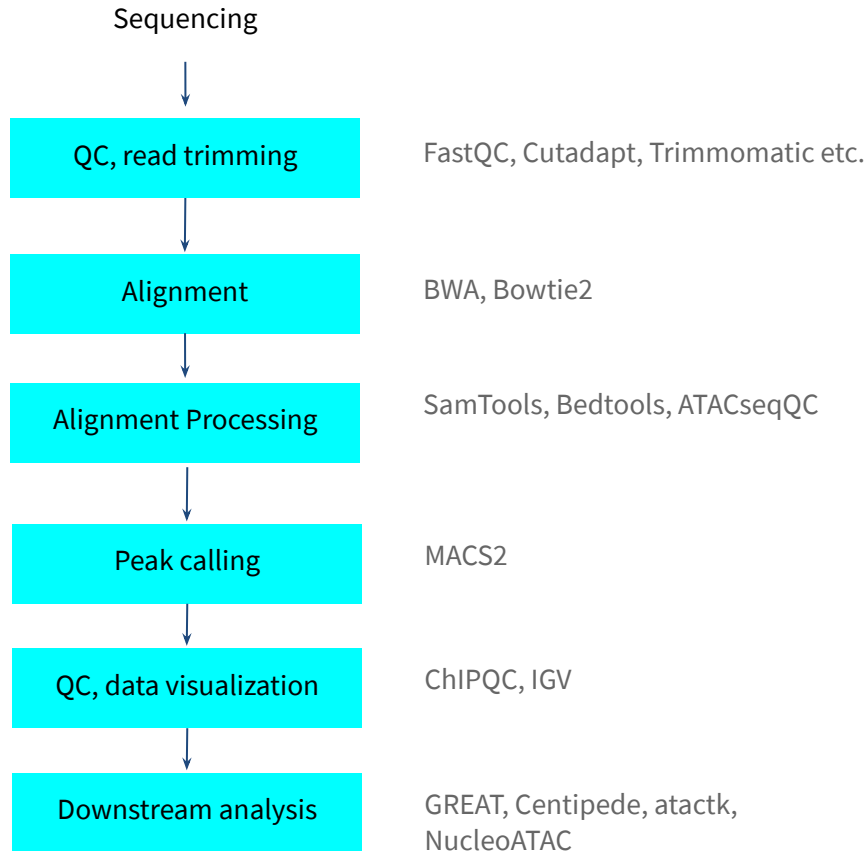
Functional States

- Different fragmentation patterns can be associated with different functional states.

eg. TSSs are more accessible than promoter flanking or transcribed regions)



Workflow of ATAC-seq data processing



Working with ATAC-seq

- Remove mitochondrial reads
 - A large fraction of ATAC-seq reads map to the mitochondrial genome (up to 40-60%)
- Remove blacklisted regions before peak calling
 - hg19 or hg38 blacklisted regions can be obtained from ENCODE
- The signal (open chromatin where the transposase was inserted) is a mixture of feature types:
 - Short fragments - signal from nucleosome free regions (**NFRs**) and open regions around DNA bound transcription factors. These are TSS rich.
 - Longer fragments - open regions from around nucleosomes. Includes +1 and -1 nucleosome positions

Differences from ChIP-seq data processing

- To identify open regions following MACS parameters are used:
`MACS2 callpeak -t bamfile --nomodel --shift -100 --extsize 200 --format BAM -g hg38`

Also try with and without the broad peaks `--broad` option.

The `--shift -100 --extsize 200` option centers a 200 bp window on the `Tn5 cut site`, which is more accurate for ATAC-seq data or single cut DNase-seq data. The 5' ends of reads represent the Tn5 cut sites; so the 5' ends of reads represent the most accessible regions.

ATAC-seq peaks are at least 200 bp long because this is about the size of a nucleosome-free region with a single nucleosome removed. Some people may use `--shift -75 --extsize 150`, with the assumption that the length of an accessible region with a single nucleosome removed is about 150 bp, which is also reasonable.

- When analysing paired-end ATAC-seq reads, many of the read pairs will span at least one nucleosome. Distribution plots of fragment length show that some fragments don't span nucleosomes (less 150-200 bp), but you will also see many fragments that do span nucleosomes (>200-400 bp). Optionally, for paired end data:

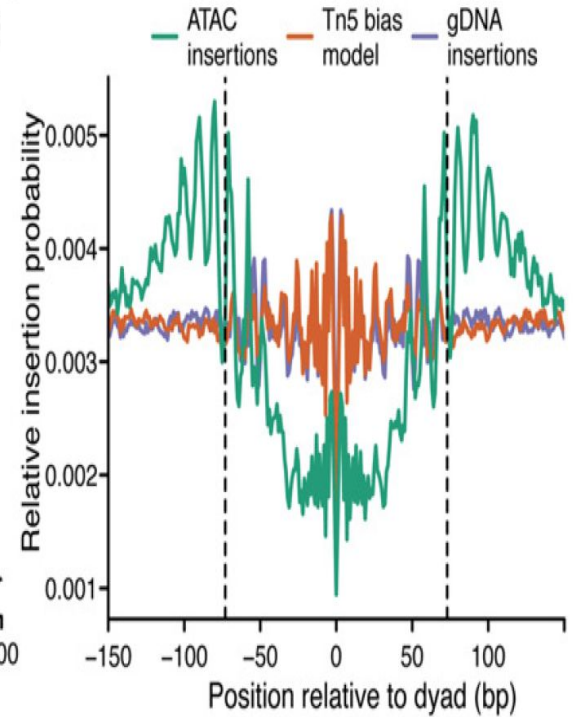
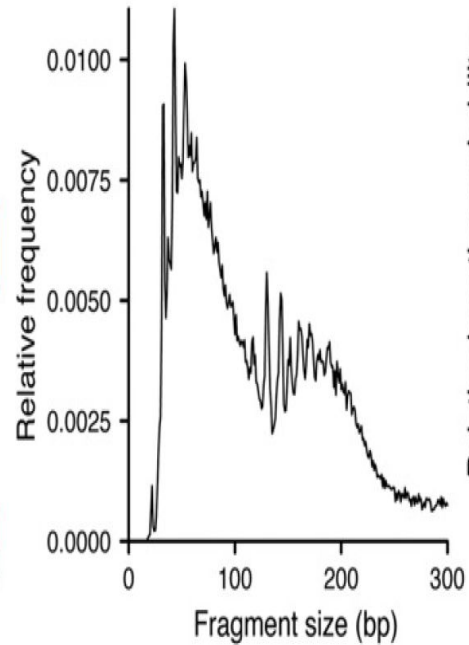
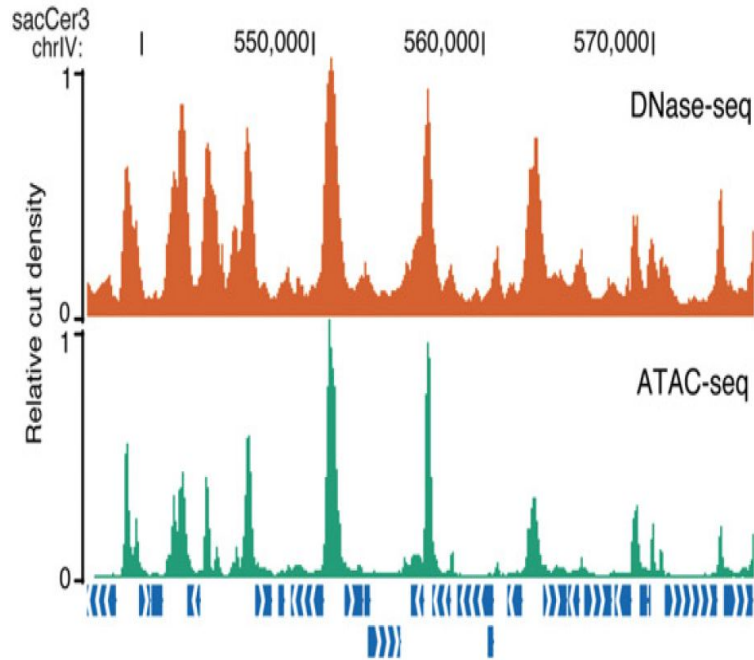
```
MACS2 callpeak -t bamfile --format BAMPE -g hg38
```

Don't use MACS2 with the default model building parameters meant for ChIP-seq. These model building parameters assume that the binding site is in the middle of the fragment, which is accurate for ChIP-seq, but not for ATAC-seq.

- For nucleosome occupancy shift and extension can centre the signal on nucleosomes (147 bp DNA is wrapped in a nucleosome)

```
MACS2 callpeak -t bamfile --nomodel --shift 37 --extsize 73 --format BAM -g hg38
```


ATAC-seq signal



Normalization and Differential Accessibility of ATAC-seq

- Normalisation across samples might be needed
 - Efficiency of the ATAC-seq protocol in assaying open regions is affected by how many transposons get into the nuclei
 - One normalisation solution is to use signal from ‘essential or housekeeping genes’. see: *Sarah K. Denny et al, Cell, 2016.*
- **THOR** is an HMM-based approach to detect and analyze differential peaks in two sets of ChIP-seq data from distinct biological conditions with replicates
 - Normalizes bam files (given a bed file of housekeeping genes)
 - Can compare two ATAC-seq datasets to perform differential openness analysis

Cutting site and TF Footprinting analysis

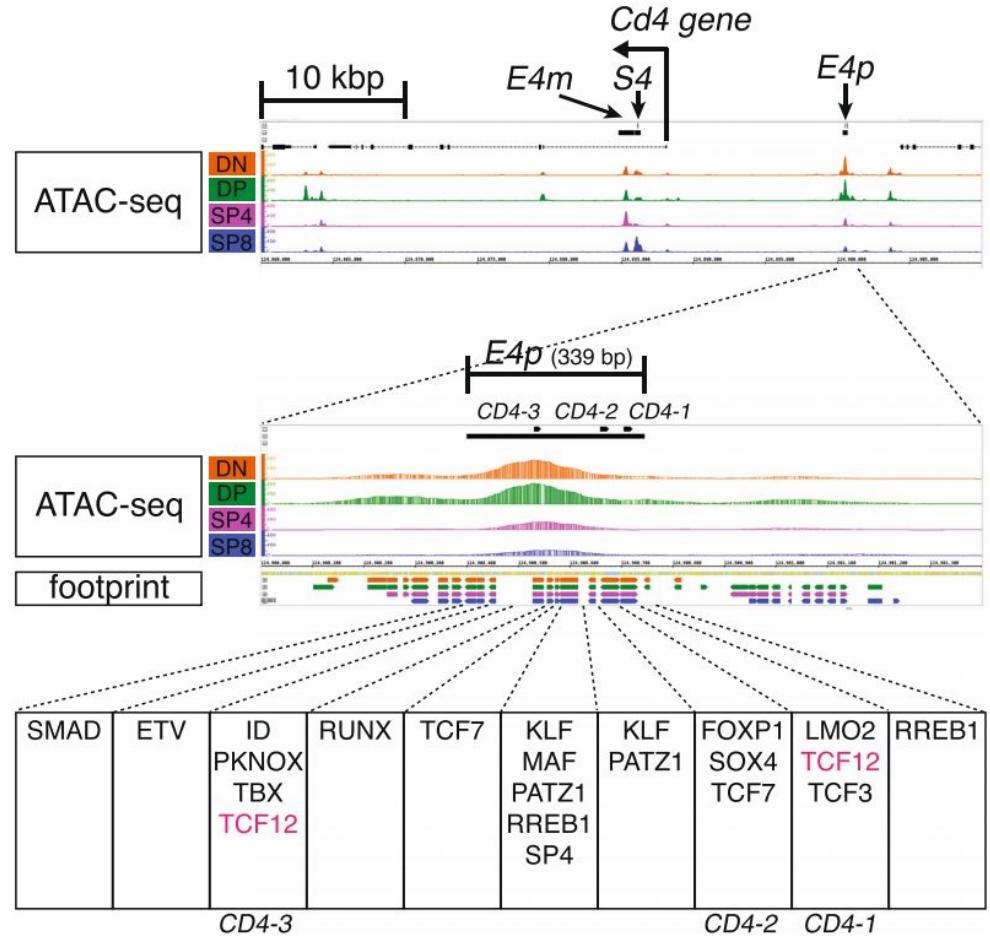
- ATAC-seq produces shorter fragments around smaller protected regions such as TF bound regions.
- Cut sites pileup signal is a good guide to TF binding sites
- Shifting reads **+4 and -5nt** depending on strand, should adjust for expected shift from **9nt** Tn5 insertion.
- TF footprinting of these regions enables the detection of enriched motifs (relative to background) for bound TFs
 - **Centipede**
 - msCentipede
 - PIQ
 - Mocap
 - **BiFET**
 - HINT-ATAC

CENTIPEDE

CENTIPEDE is a method to infer if a region of the genome is bound by a particular TF.

It uses information from a DNase-Seq / ATAC-seq experiment about the profile of reads surrounding a putative TF binding site.

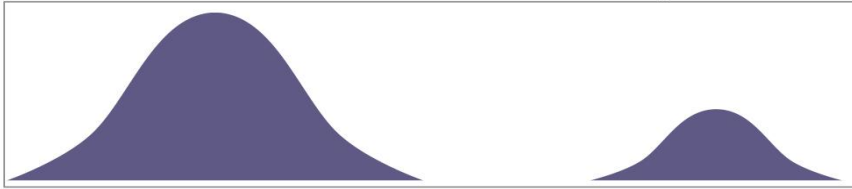
Hosoya et al., Sci Rep. 2018



BiFET

BiFET is a TF footprinting method that corrects for biases that arise from differences in chromatin accessibility levels and GC content.

Chromatin accessibility data

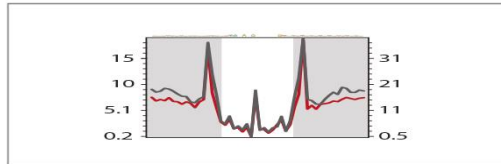


TF PWMs

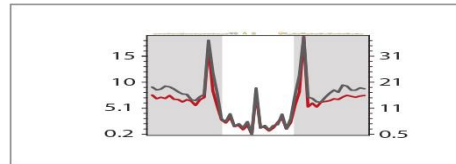


Footprinting algorithm

Footprints in
target loci



Footprints in
background loci



Bias-free Footprint Enrichment Test (BiFET)