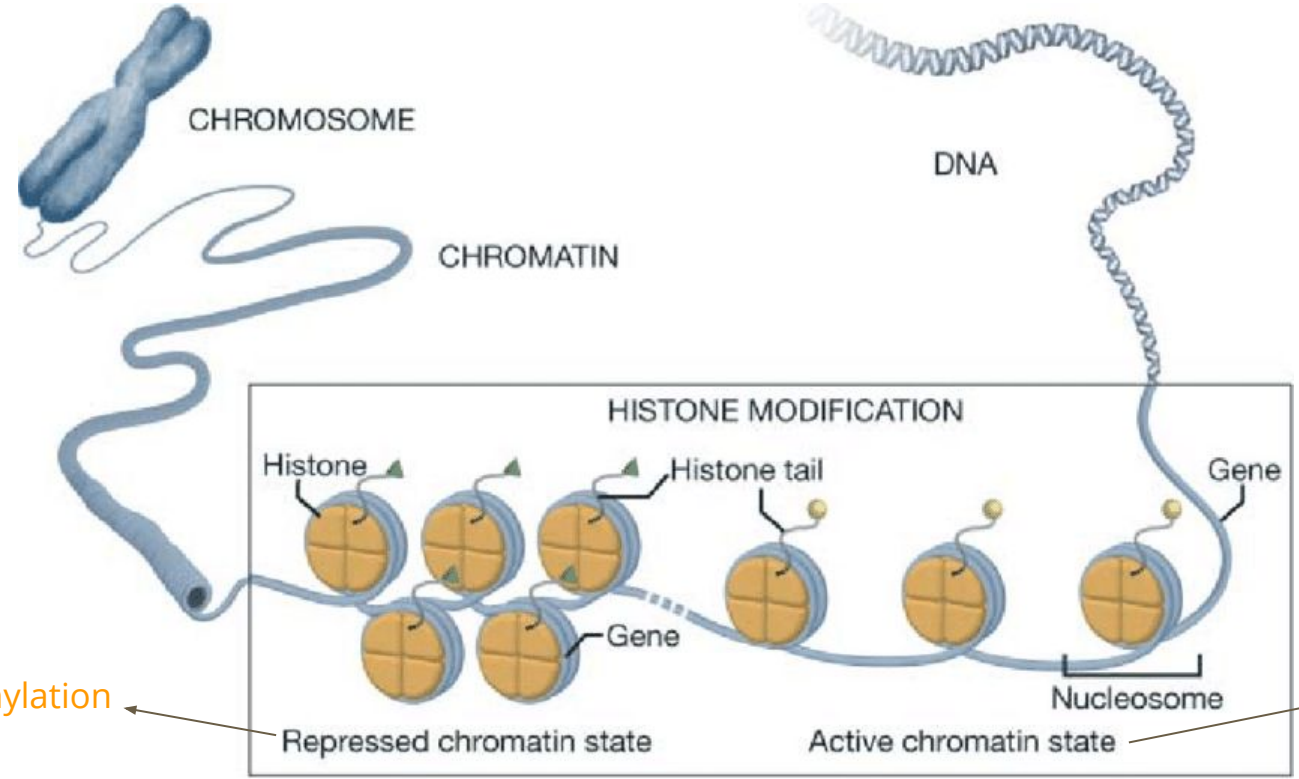
ATAC-Sequencing

SG3 Genomics
14/04/2020
Kathleen Imbach

Yan, F., Powell, D.R., Curtis, D.J. & Wong, N.C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* **21**, 22 (2020).

First, a bit about chromatin...



Pieterman, C.R.C. *et al.* Thoracic and duodenopancreatic neuroendocrine tumors in multiple endocrine neoplasia type 1: natural history and function of menin in tumorigenesis. *Endocrine-Related Cancer* **21**, R121-R142 (2014).

ATAC-seq

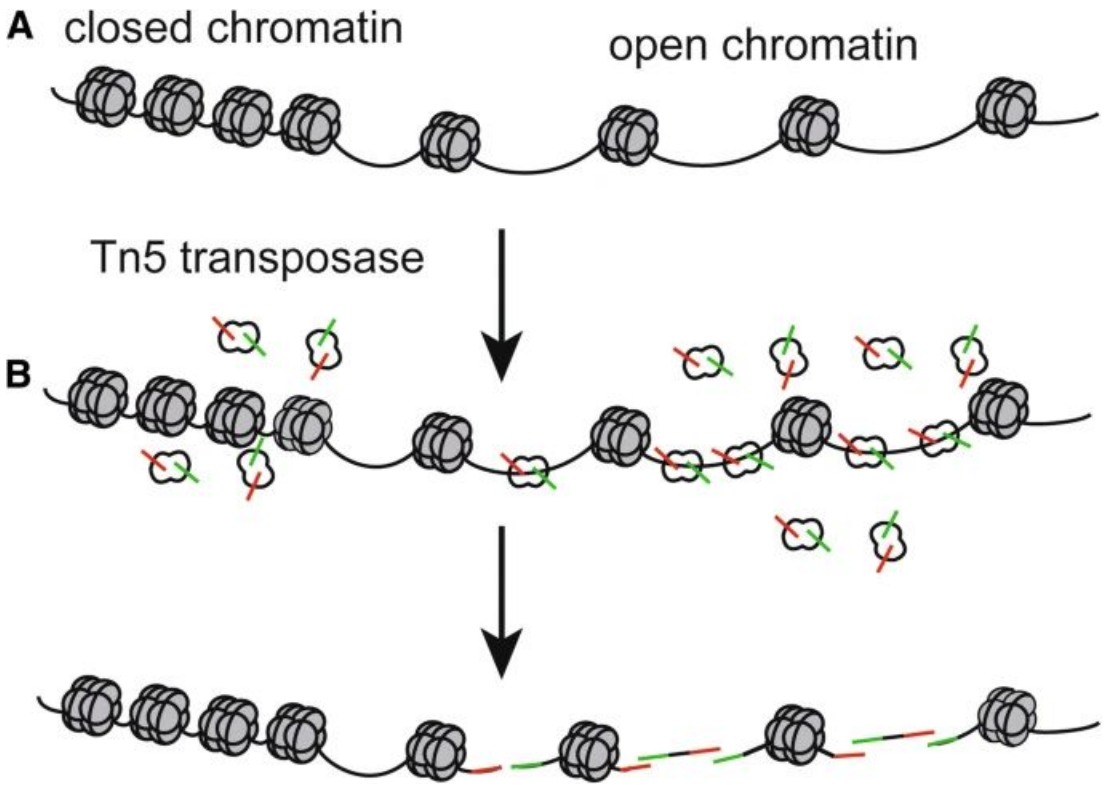
Assay of **T**ransposase **A**ccessible **C**hromatin **seq**uencing

Nature Methods by Jason Buenrostro and colleagues in 2013[1]

Genome-wide map of chromatin accessibility

Hyperactive Tn5 transposase

Low number of cells needed (500-50,000), simple protocol



1. Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y. & Greenleaf, W.J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature methods* **10**, 1213-1218 (2013).

Sun, Y., Miao, N. & Sun, T. Detect accessible chromatin using ATAC-sequencing, from principle to applications. *Hereditas* **156**, 29 (2019).

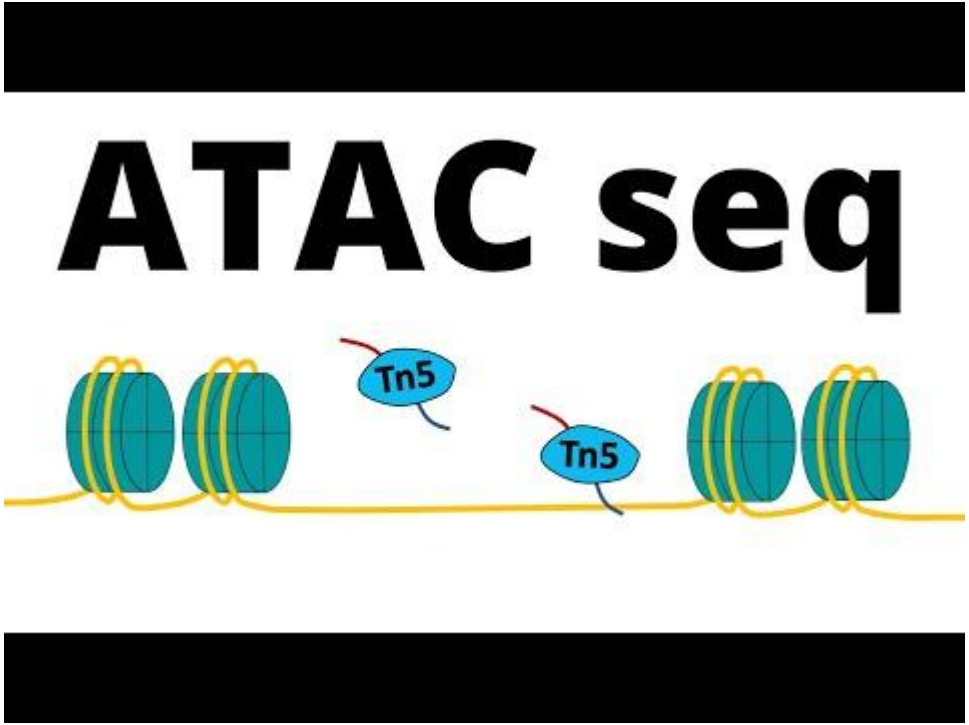
How does ATAC-seq differ from other methods?

ATAC-seq	DNase-seq	FAIRE-seq	ChIP-seq	MNase-seq	Hi-C
Reveals areas of open chromatin Uses Tn5 transposase for cleavage and tagmentation Fragments are used to identify peaks of accessible regions	Also reveals open chromatin Uses DNase I endonuclease to digest chromatin Size selection to enrich fragments in regions sensitive to DNase	Identifies regulatory regions of DNA Uses formaldehyde to crosslink chromatin Sonication and extraction used to obtain regions free of nucleosomes	Reveals binding sites of specific proteins (TFs, etc) Uses immuno-precipitation to extract fragments bound to target proteins	Identifies positions of nucleosomes Uses micrococcal nuclease to digest DNA until obstruction DNA bound to histone/protein remain for sequencing	Reveals spatial organization of genome DNA cross-linked with formaldehyde Interacting DNA regions are fragmented, ligated and sequenced

Meyer, C.A. & Liu, X.S. Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nature reviews. Genetics* **15**, 709-721 (2014).
Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R. & Lieb, J.D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome research* **17**, 877-885 (2007).

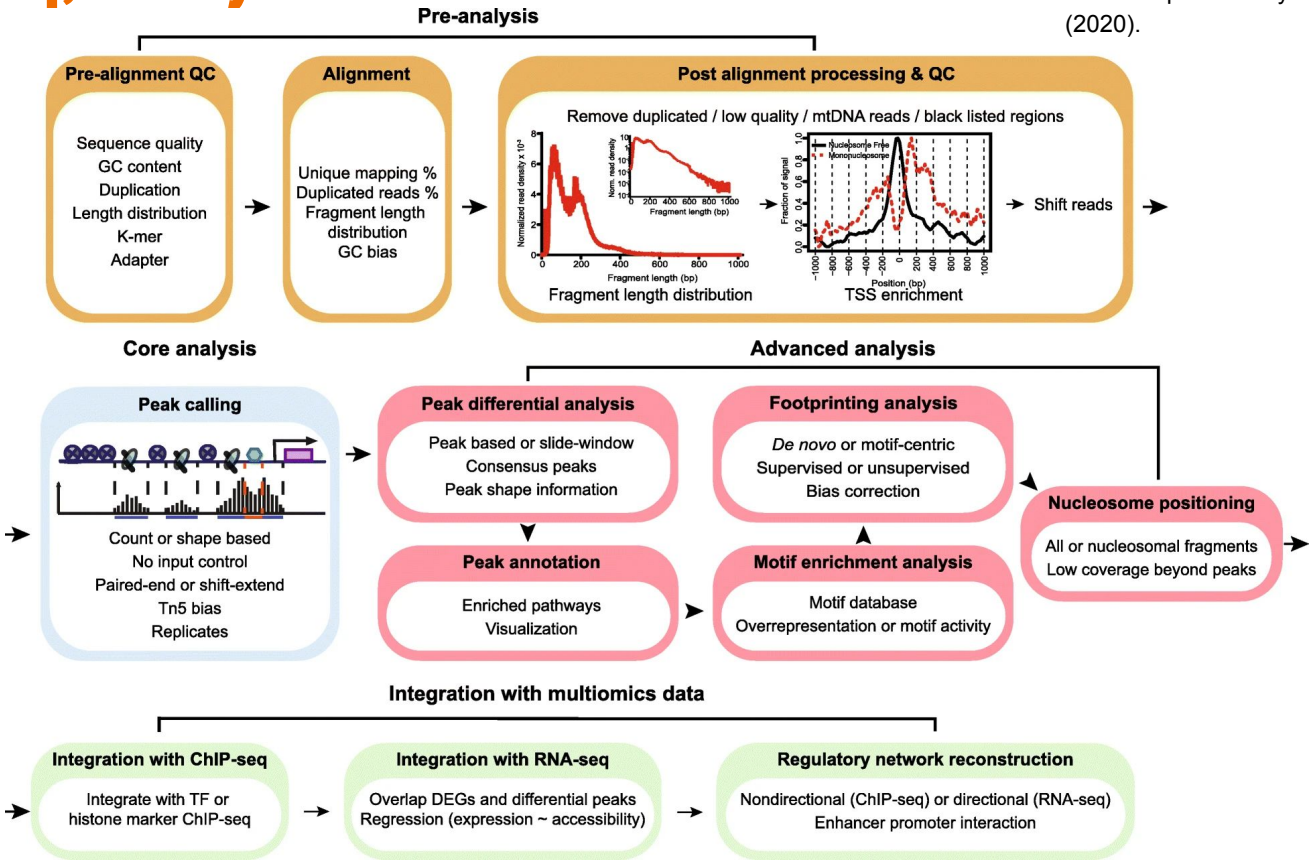
ATAC-seq, experiment overview

1. Prep cells
2. Tagmentation with Tn5 transposase
3. PCR amplification
4. Next-generation sequencing
5. Map nucleosomes & accessible regions



ATAC-seq, analysis overview

Yan, F., Powell, D.R., Curtis, D.J. & Wong, N.C. From reads to insight: a hitchhiker's guide to ATAC-seq data analysis. *Genome Biology* **21**, 22 (2020).



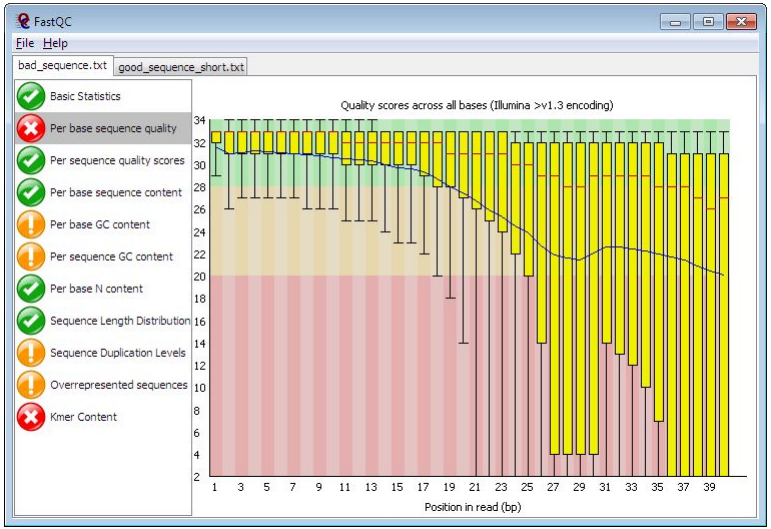
Pre-alignment QC

Standard step for most sequencing data

- Trim reads
- Evaluate base quality scores
- GC content
- Distribution of fragment lengths
- K-mer scores
- Primer/adapter contamination

Remove Nextera adapters → **cutadapt, AdapterRemoval v2, Skewer,** and **trimmomatic**

FastQC



Alignment

Map trimmed reads to genome

Tools: Bowtie2, BWA-MEM

Over 80% unique mapping rate = successful experiment

50 mil. mapped reads for open chromatin detection, 200 mil. for footprinting analyses

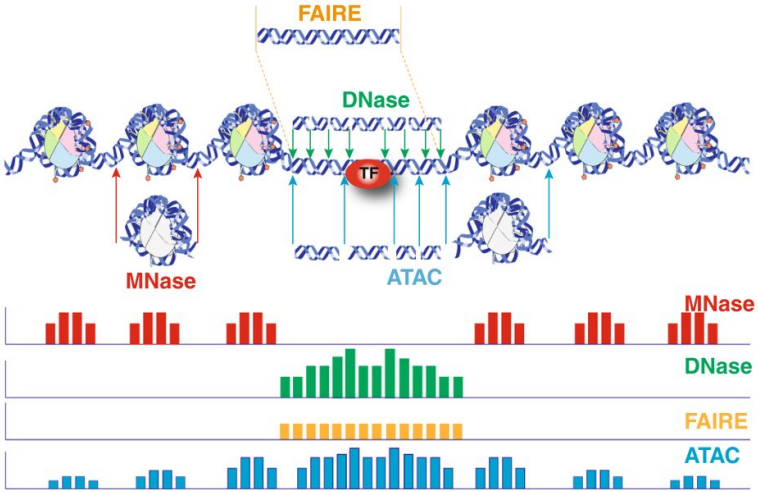


Figure 1 Schematic diagram of current chromatin accessibility assays performed with typical experimental conditions. Representative DNA fragments generated by each assay are shown, with end locations within chromatin defined by colored arrows. Bar diagrams represent data signal obtained from each assay across the entire region. The footprint created by a transcription factor (TF) is shown for ATAC-seq and DNase-seq experiments.

Tsompana, M. & Buck, M.J. Chromatin accessibility: a window into the genome. *Epigenetics & Chromatin* **7**, 33 (2014).

Post-alignment QC

Evaluate metrics of BAM file

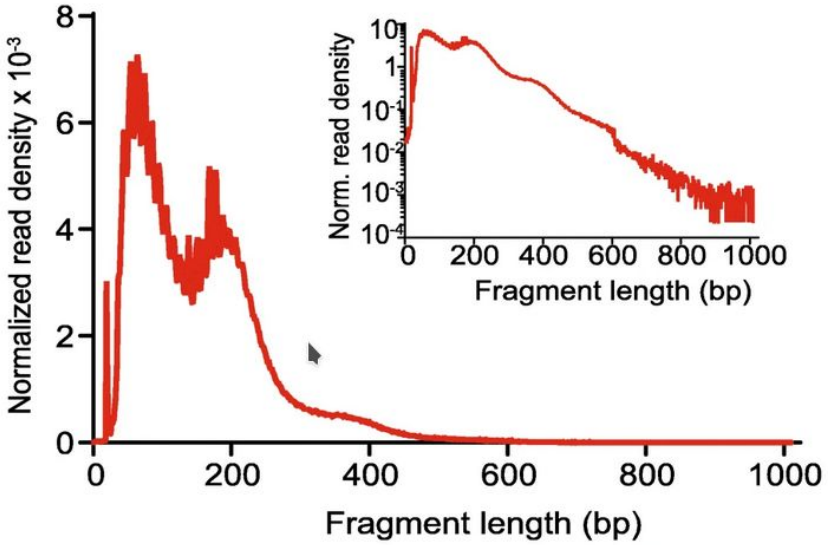
- Percentage of duplicated reads
- Unique mappings
- Fragment size distribution

Tools: Picard, SAMtools

Discard MT genome, ENCODE blacklisted regions, duplicated reads

Evaluate fragment size distribution plot

- nucleosome-free regions (NFR) < 100 bp
- mono-, di-, and tri-nucleosomes ~ 200, 400, 600 bp, respectively
- **Tool: ATACseqQC**



Peak calling

Identify accessible regions (peaks) in the genome

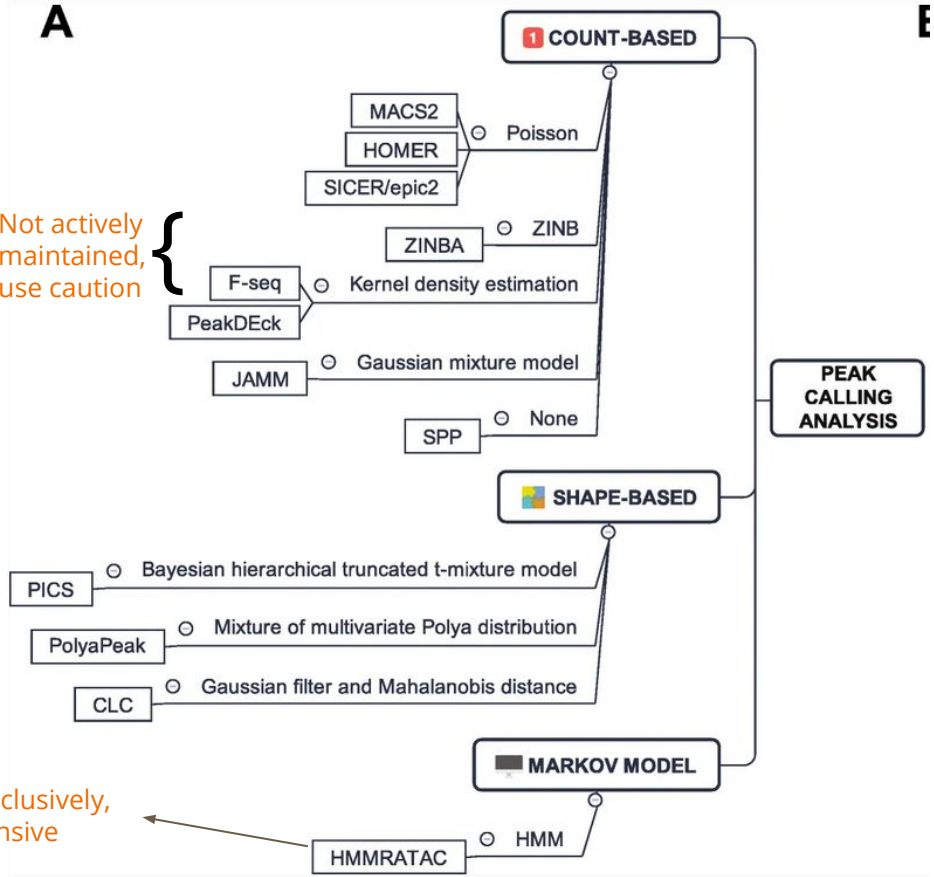
Count-based or shape-based

Should account for cleavage bias of Tn5 at GC regions

Often adapted from ChIP-seq or DNase-seq tools- only one developed specifically for ATAC-seq

Input controls, like used in ChIP-seq, not practical for ATAC-seq: tools should not require

Tools: MACS2, HOMER, HMMRATAC



Only caller for ATAC-seq exclusively, computationally more intensive

Peak differential analysis

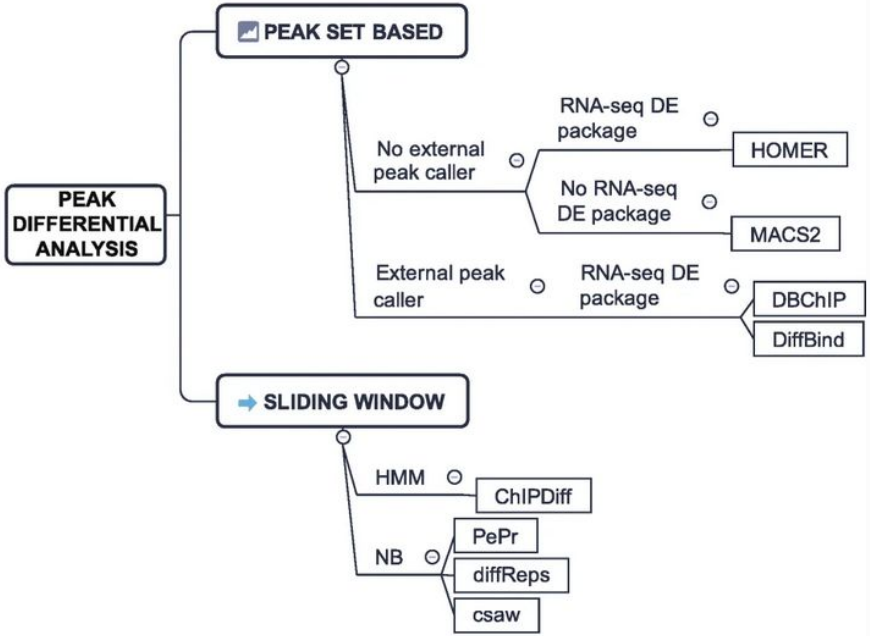
No differential peak analysis specifically for ATAC-seq

Consensus peak-based tools (often use RNA-seq DE) vs sliding window methods

- most assume reads in peak regions follow a NB distribution (like RNA-seq data)
- No shape-based tools

Peak annotation

- Nearby genes, regulatory elements
- **Tools: HOMER, ChIPseeker, ChIPpeakAnno**



Motifs

Motif: sequence of DNA to which TFs bind

- Binding position: TFBS

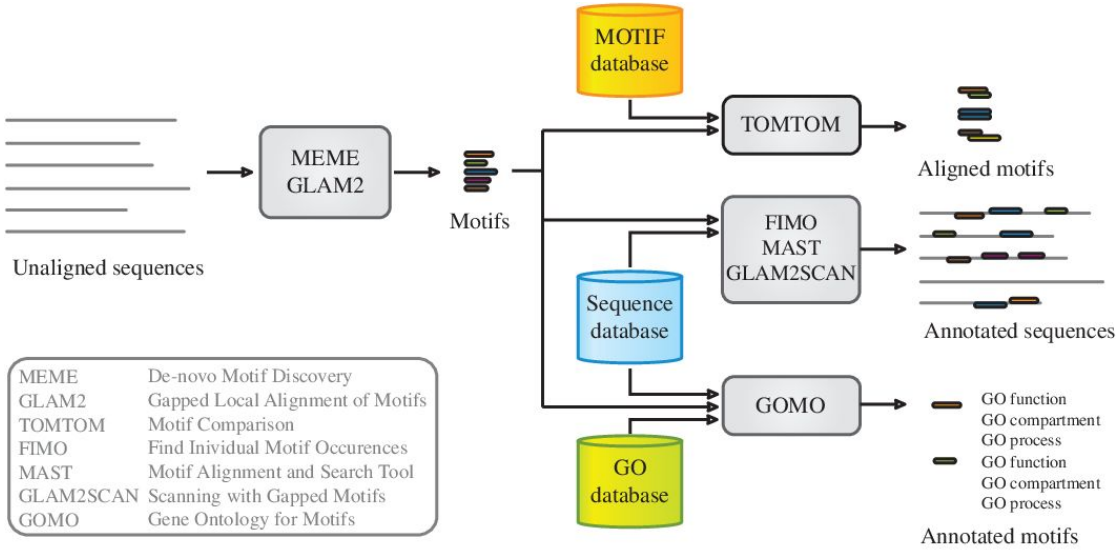
Helps to understand underlying mechanisms of chromatin accessibility

Databases available with motif information:

- JASPAR, CIS-BP,TRANSFAC, HOCOMOCO

Tools to generate likely TFBS from given peak sequences:

- **HOMER, TFBSTools, motifmatchr, PWMScan, MEME**



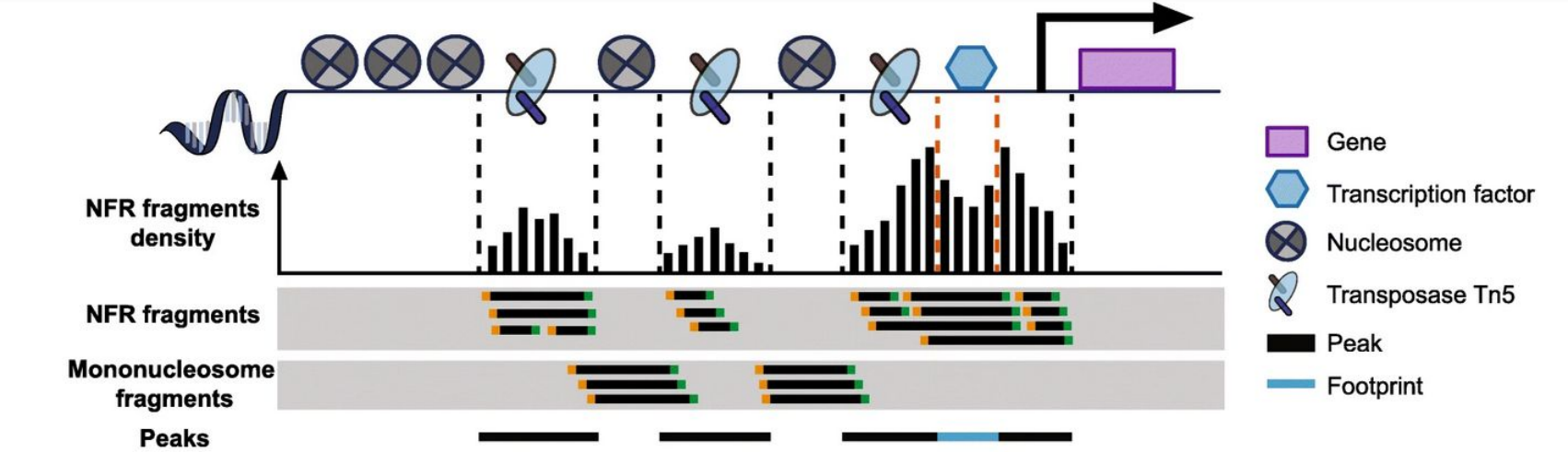
Bailey, T. *et al.* MEME Suite: tools for motif discovery and searching. *Nucleic Acids Research* **37**, W202 - W208 (2009).

Footprinting

Footprint: pattern that occurs when active TF is bound to DNA and prevents Tn5 cleavage

Must shift raw reads in *pre-processing* to account for 9 bp duplicates

Tools: de novo (**HINT-ATAC**) vs motif-centric (supervised vs unsupervised)

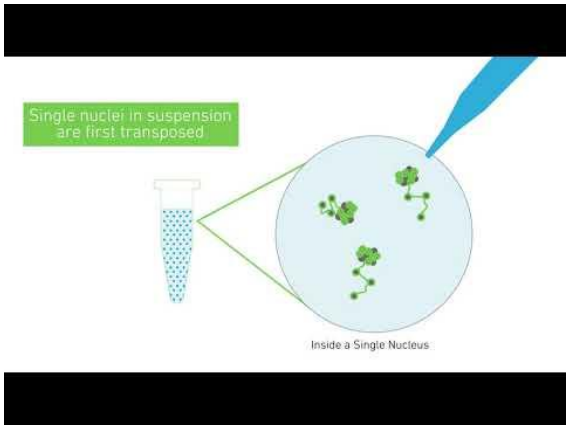


Single-cell ATAC-seq & multi-omics approaches

ATAC-seq can be done in single-cells using barcoding procedures

Clustering scATAC-seq: **SnapATAC**

Sparseness, as occurs in scRNA-seq, often results in scATAC-seq data



ATAC-seq can be combined with other technologies, such as RNA-seq and CiTE-seq

- CiTE-seq: validate results via overlapping peaks, indicate TFs that can bind to closed chromatin, reduce TFBS false positives

- RNA-seq: associate chromatin accessibility changes with gene expression changes, associate DE with TFs/motifs, better population detection (single-cell)