Tool-ing around with ATAC-seq

Meeta Mistry, PhD
Research Scientist
Harvard Chan Bioinformatics Core
Harvard T.H. Chan School of Public Health

The most basic repeating unit of chromatin is the nucleosome, consisting of ~147 bp of DNA wrapped around an octamer of histone proteins

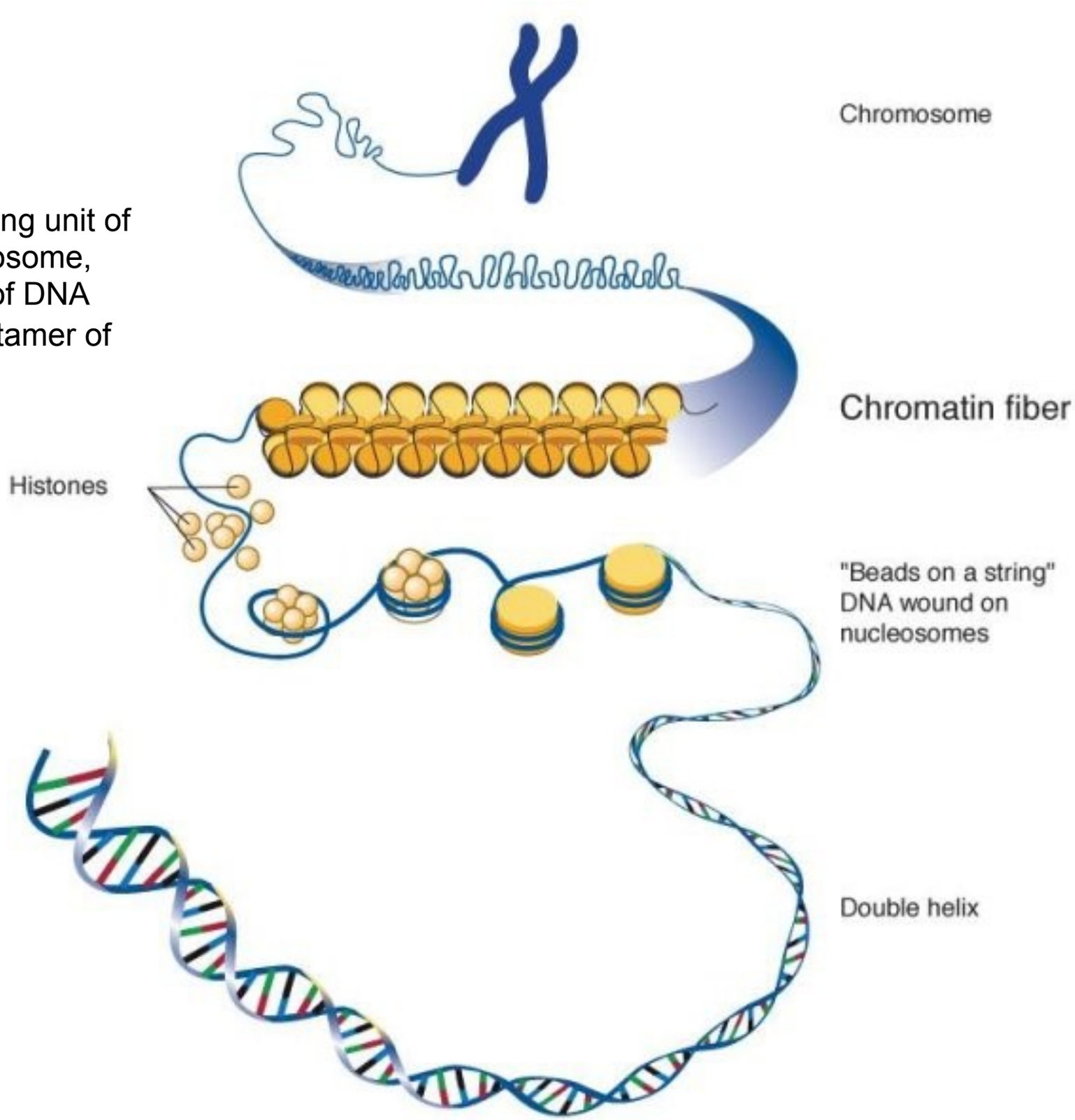
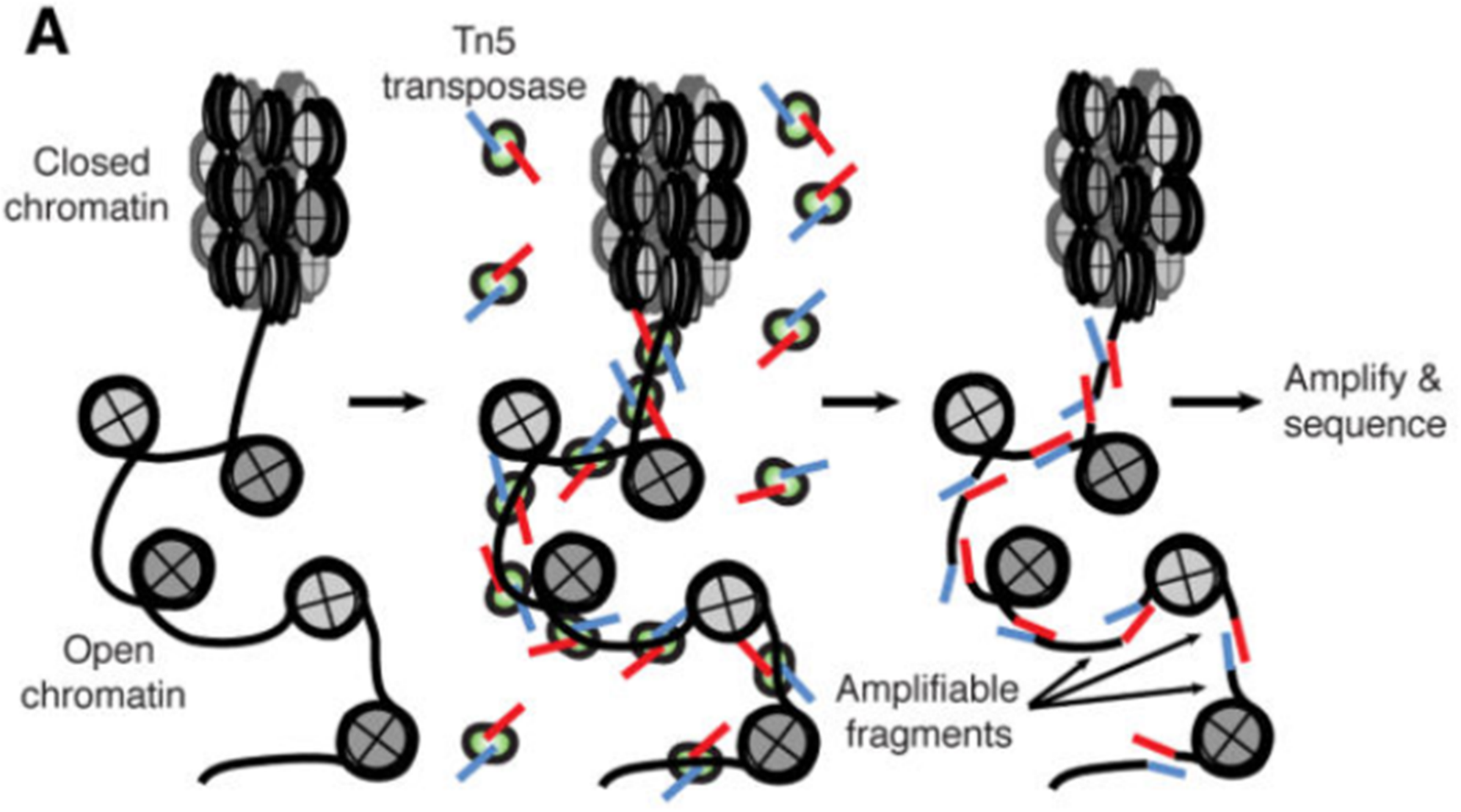


Image source: Creative Diagnostics Blog

Assay for Transposase-Accessible Chromatin (ATAC)



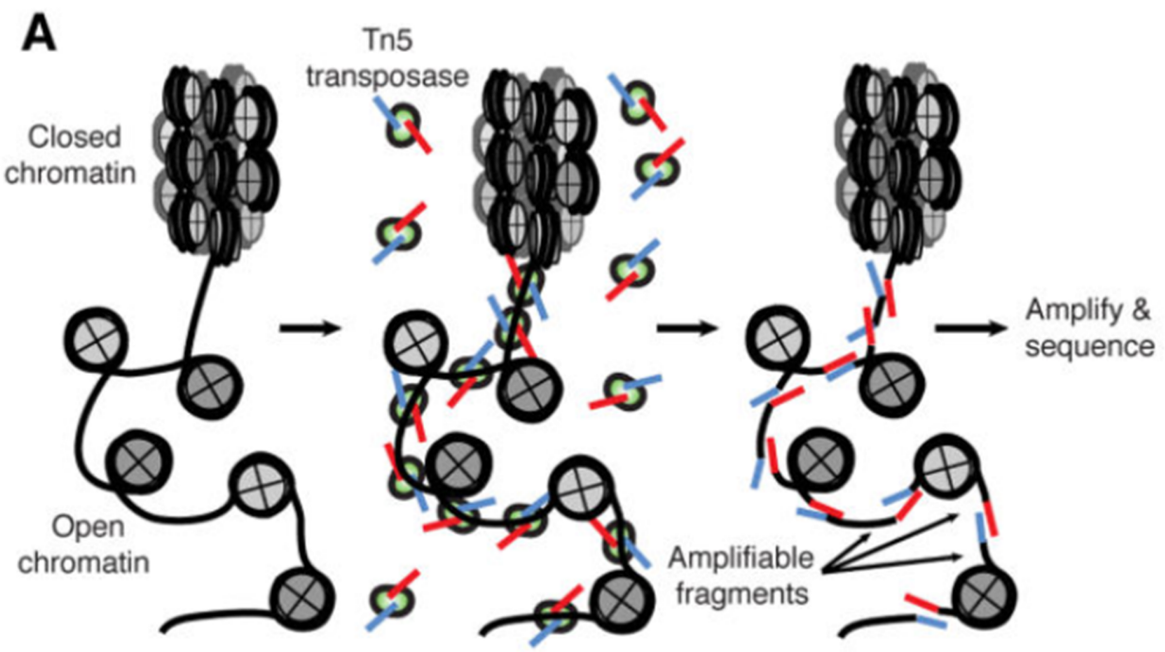
Buenrostro et al., 2015

Assay for Transposase-Accessible Chromatin (ATAC)

- Utilizes hyperactive Tn5 transposase to insert sequencing adapters into the open chromatin regions

- Tn5 tagmentation simultaneously fragments the genome and tags the resulting DNA with sequencing adapters

- Amplify and sequence



Buenrostro et al., 2015

Why ATAC-seq?

- Main advantage over existing methods is the simplicity of the library preparation protocol: Tn5 insertion followed by two rounds of PCR.
    - requires no sonication or phenol-chloroform extraction like FAIRE-seq
    - no antibodies like ChIP-seq
    - no sensitive enzymatic digestion like MNase-seq or DNase-seq
- Unlike similar methods, which can take up to four days to complete, ATAC-seq preparation can be completed in under three hours.
- Lower starting cell number than other open chromatin assays (500 to 50K cells recommended for human).

What does it give us?

- Multiple aspects of chromatin architecture simultaneously at high resolution.
    - Maps open chromatin
    - TF occupancy
    - nucleosome occupancy

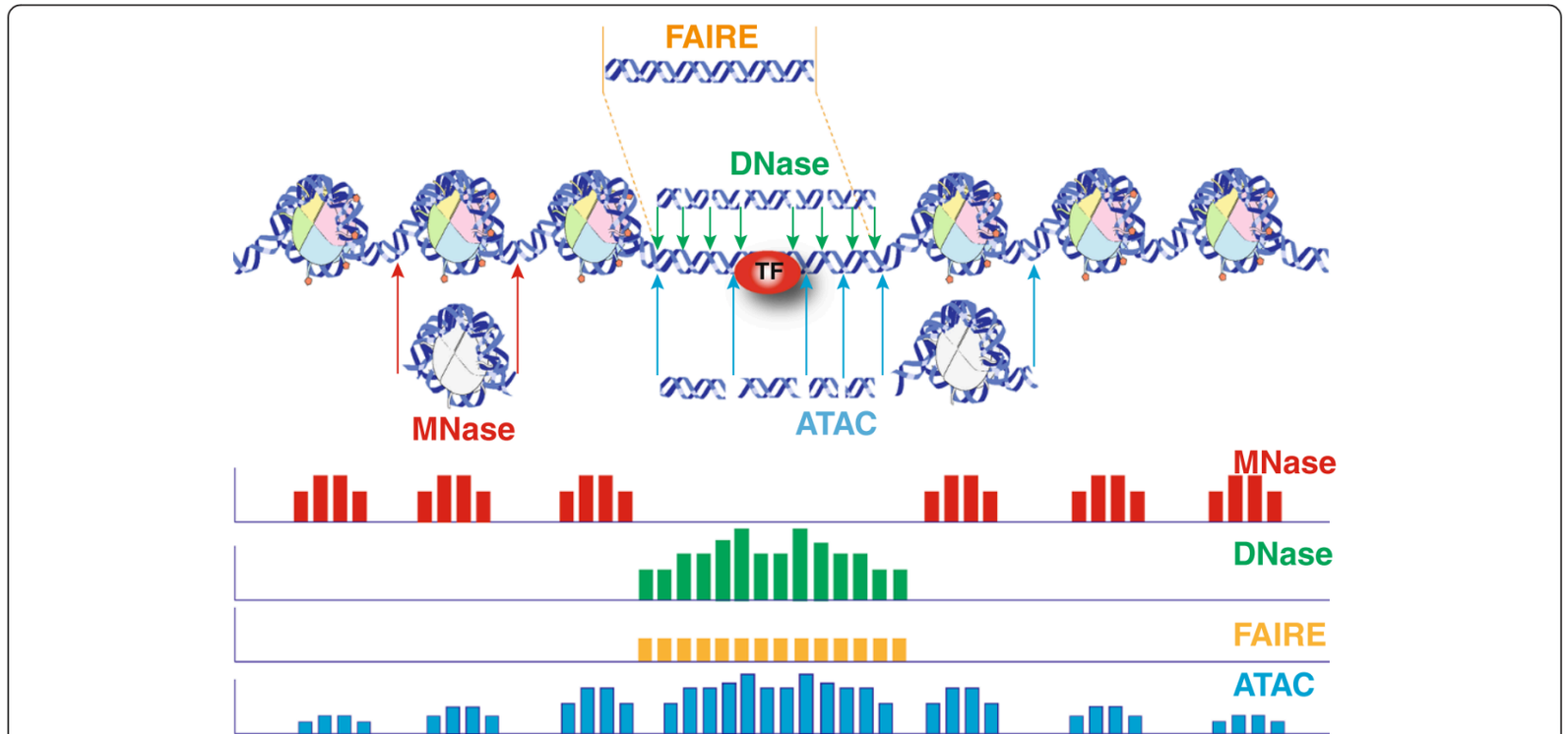


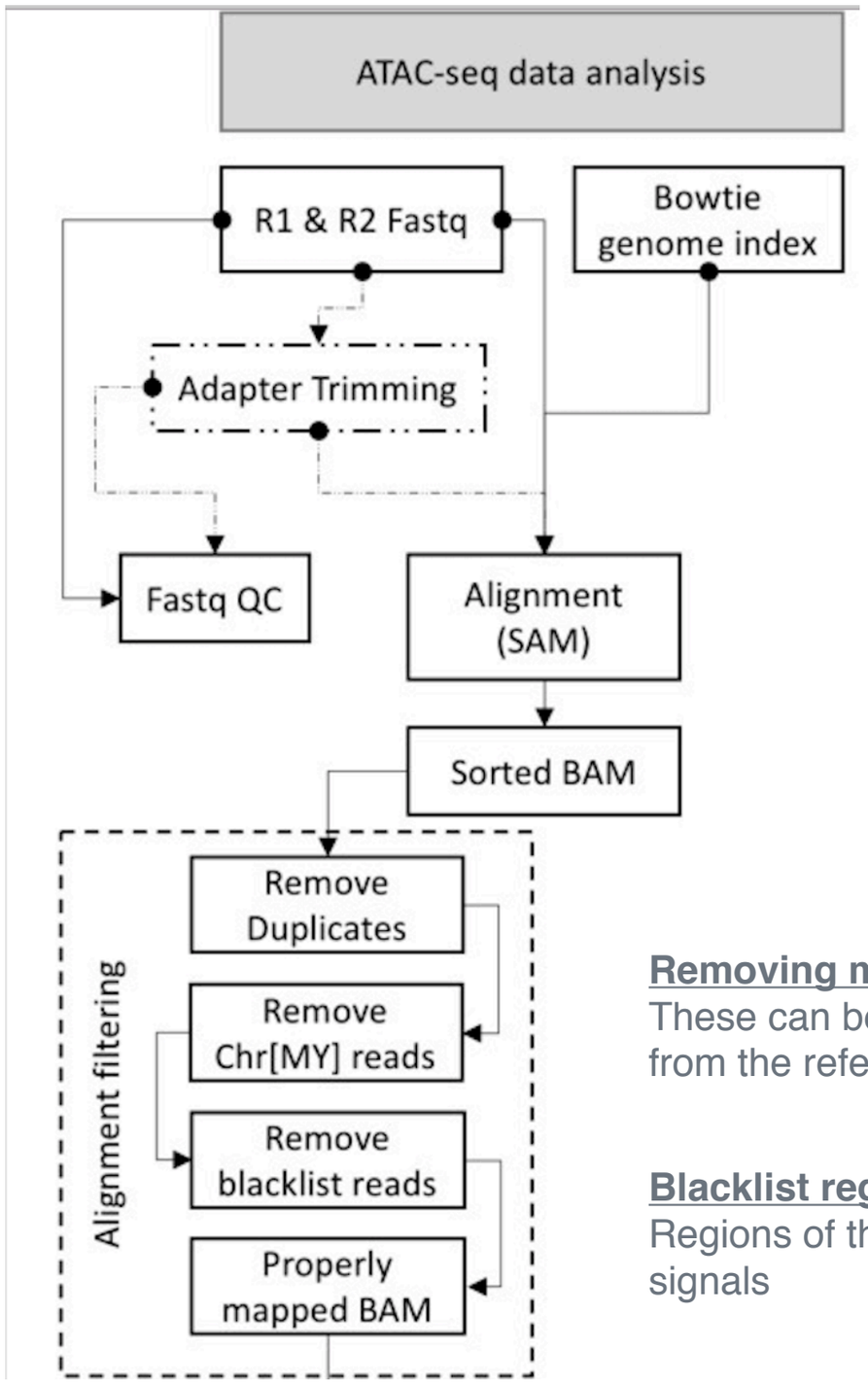
Figure 1 Schematic diagram of current chromatin accessibility assays performed with typical experimental conditions. Representative DNA fragments generated by each assay are shown, with end locations within chromatin defined by colored arrows. Bar diagrams represent data signal obtained from each assay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

Tsompana and Buck, 2014

Planning your ATAC-seq experiment

- **Replicates:** more is better
- **Controls:** not typically run, but could use genomic DNA fragmented by some other method (i.e. sonication)
- **PCR amplification:** as few cycles as possible
- **Sequencing depth:** varies based on size of reference genome and degree of open chromatin expected
- **Sequencing mode:** paired-end
- **Mitochondria:** discarded from computational analyses; option to remove during prep

Preparing the BAM



`–very-sensitive` for better alignment results

bwa as an Alternative

Mapping rates are higher (~ 2%), with an equally similar increase in the number of duplicate mappings identified. Post-filtering this translates to a significantly higher number of mapped reads and results in a much larger number of peaks being called (30% increase). When we compare the peak calls generated from the different aligners, the **bwa** peak calls are a superset of those called from the Bowtie2 aligments. Whether or not these additional peaks are true positives, is something that is yet to be determined.

Removing mitochondrial reads

These can be filtered out from the alignment files or just removed from the reference before building the index

Blacklist regions

Regions of the genome that tend to show artificially high read signals

Divate and Cheung, 2018

Peak calling with MACS2

- use the `callpeaks` command:
  - `–nomodel –nolambda;` turn off the model building and shifting and do not compute local bias lambda.
  - `–keep-dup-all;` if you have removed PCR duplicates
  - `-f BAMPE;` analyze only properly paired alignments
  - for NFR can try to do`–nomodel` with `–shift` and `–extsize` using the size of the fragments

Peak calling (and more) with Genrich

- Designed to be able to run all of the post-alignment steps through peak-calling with **one command**.
    - Removal of mitochondrial reads
    - Removal of PCR duplicates
    - Analysis of multi-mapping reads (adding fractional amount to each location)
    - Analysis of multiple replicates; collectively calls peaks. No more IDR.
    - ATAC-seq mode: intervals are centered on transposase cut sites

Open regions vs Footprinting

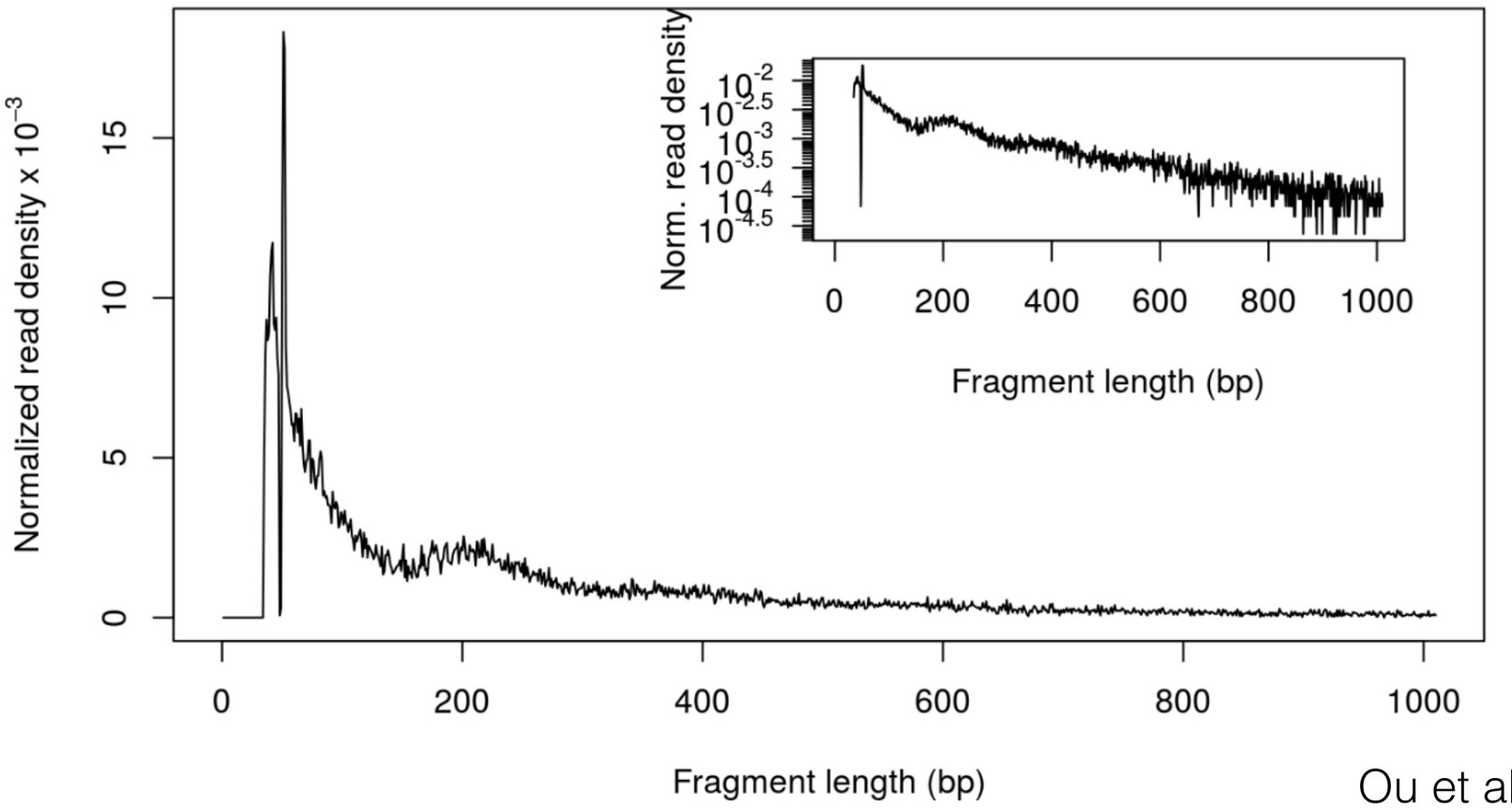
- Genome-wide maps of nucleosome positions have been generated in a number of organisms
  - coding regions have high nucleosome occupancy
  - transcriptional regulatory regions (promoters, enhancers, terminators) have low nucleosome occupancy and often contain NFR (5’ and 3’)

Alignment shifting

- when Tn5 tranposase cuts it introduces two cuts that are separated by 9bp. Therefore, reads aligning to the +/- strand need to be adjusted by +4 and -5bp to represent the center of the binding site.
- adjusted reads are written to a new BAM file

ATAC-seqQC

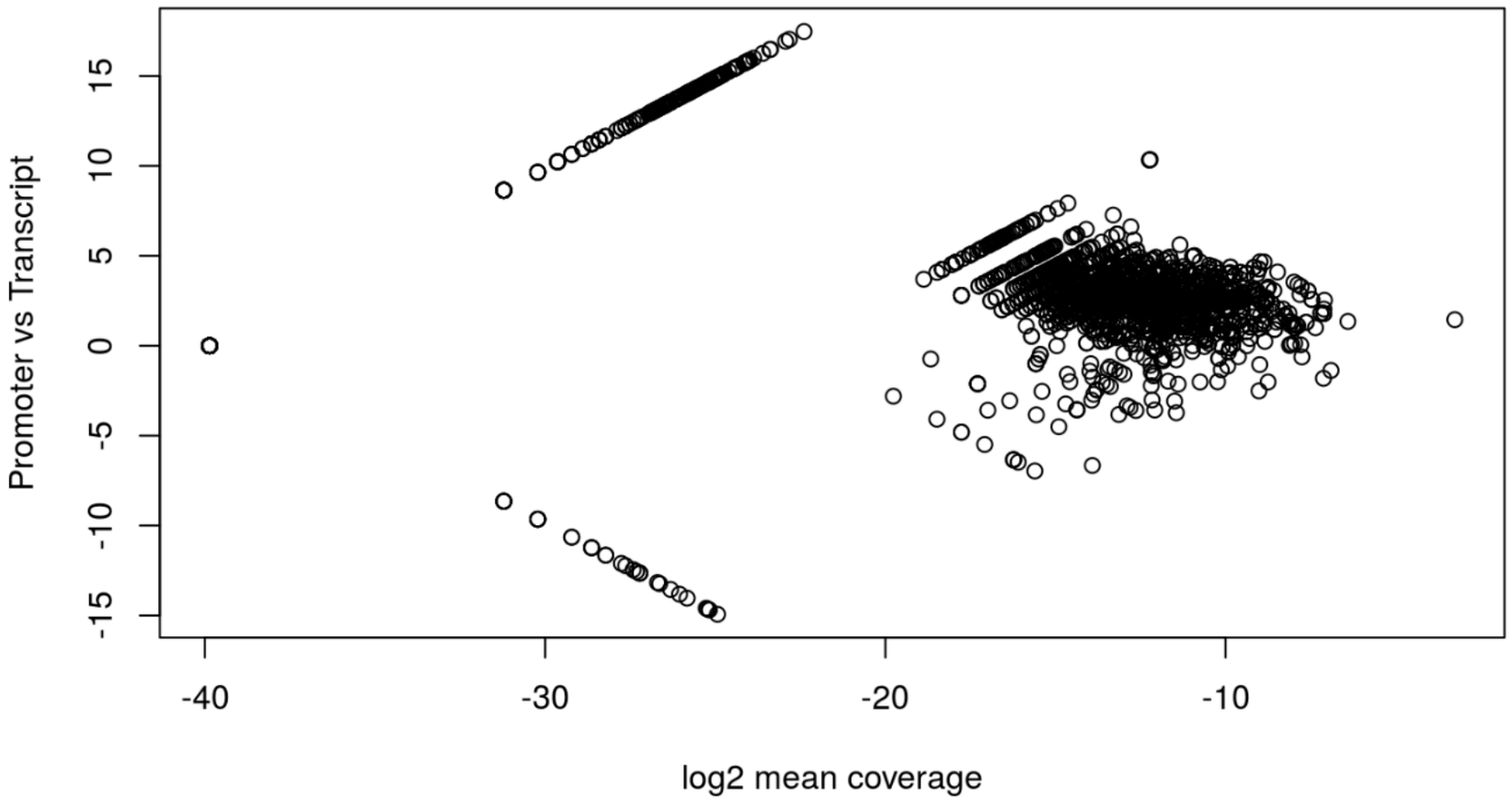
- **Fragment size distribution graph:** Tn5 transposase cuts open chromatin regions and also linker DNA. Therefore, the fragment size distribution graph of a good quality ATAC-seq library has two sharp peaks at <100 bp (open chromatin) and ~200 bp (mono-nucleosome) and smaller peaks representing di-nucleosomes and tri-nucleosomes.



Ou et al, Bioconductor 2018

ATAC-seqQC

- **Promoter/Transcript body (PT) score:** PT score is calculated as the coverage of promoter divided by the coverage of its transcript body. PT score will show if the signal is enriched in promoters.



Ou et al, Bioconductor 2018

ATAC-seqQC

- Nucleosome Free Regions (NFR) score
- Transcription Start Site (TSS) enrichment score
- Split reads: place shifted reads into bins (nucleosome free, mononucleosome, dinucleosome, trinucleosome) and use random forest to classify fragments based on fragment length, GC content and conservation scores
    - Heatmap and coverage curves

Ou et al, Bioconductor 2018

GUAVA:
GUI tool for the analysis and visualization of ATAC-seq data

- Processing from raw reads to ATAC-seq signals
- Can perform differential enrichment analysis
- Annotations and results on GO and pathway analysis
- Visualization of data tracks

Divate and Cheung, 2018

Other tools/software

- **NucleoATAC:** from Greenleaf lab but no longer actively maintained

- I-ATAC: from Jackson laboratory but requires cluster environment with all tools installed

For discussion

- Methods for integration with RNA-seq
- single-cell ATAC-seq

Thank you!

HARVARD
T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Members of the Harvard Chan Bioinformatics Core