Introduction to ChIP-seq and ATAC-seq

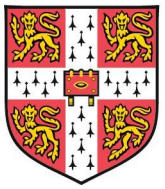
Shamith Samarajiwa

University of Cambridge

CRUK Bioinformatics Summer School

September 2017

MRC Medical Research Council



UNIVERSITY OF CAMBRIDGE

Important!!!

- Good Experimental Design
- Optimize Conditions (Cells, Antibodies, Sonication etc.)
- **Biological Replicates (at least 3)**!!
  - sample biological variation & improve signal to noise ratio
  - capture the desired effect size
  - statistical power to test null hypothesis
- ChIP-seq controls – **Knockout, Input** (Try not to use IgG)

What is ChIP Sequencing?

- Combination of chromatin immunoprecipitation (ChIP) with ultra high-throughput massively parallel sequencing.

- Allows mapping of protein–DNA interactions ***in vivo*** on a genome scale.
- Enables mapping of transcription factors binding, DNA binidng proteins (HP1, Lamins, HMGA etc), RNA Pol II occupancy or Histone modification marks at genome scale.

- The typical ChIP assay usually take 4–5 days, and require approx. 10^6~ 10^7 cells.

Origins of ChIP-seq technology

- Barski, A., Cuddapah, S., Cui, K., Roh, T. Y., Schones, D. E., Wang, Z., et al. “High-resolution profiling of histone methylations in the human genome.” *Cell 2007*

- Johnson, D. S., Mortazavi, A., Myers, R. M., and Wold, B. “Genome-wide mapping of *in vivo* protein-DNA interactions.” *Science* 316, 2007

- Mikkelsen, T. S., Ku, M., Jaffe, D. B., Issac, B., Lieberman, E., Giannoukos, G., et al. “Genome-wide maps of chromatin state in pluripotent and lineage-committed cells.” *Nature* 2007

- Robertson et al., "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing." *Nat Methods.* 2007

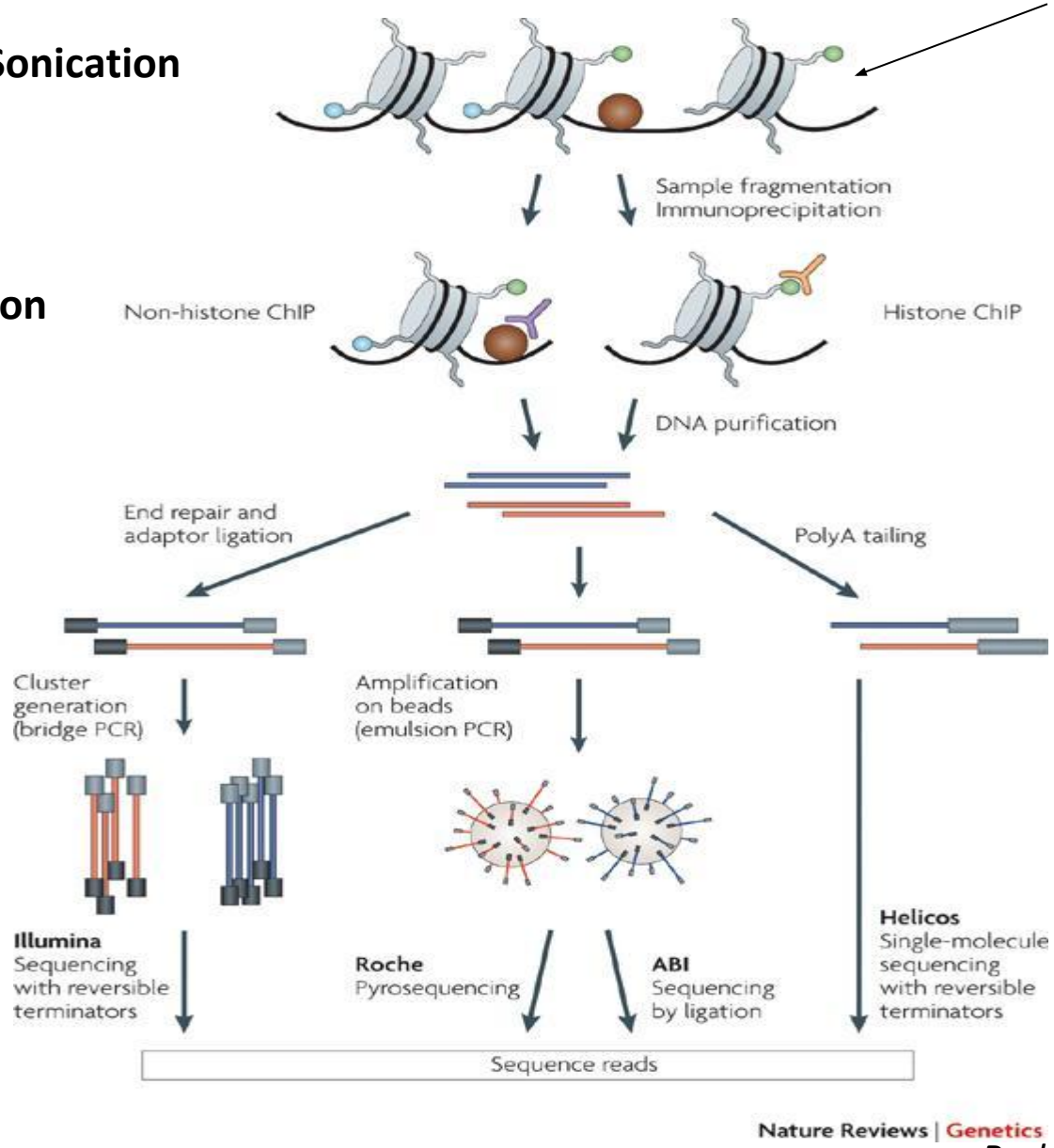
ChIP-seq methodology

Cross-link cells

Isolate genomic DNA

Sonication

Immuno-precipitation



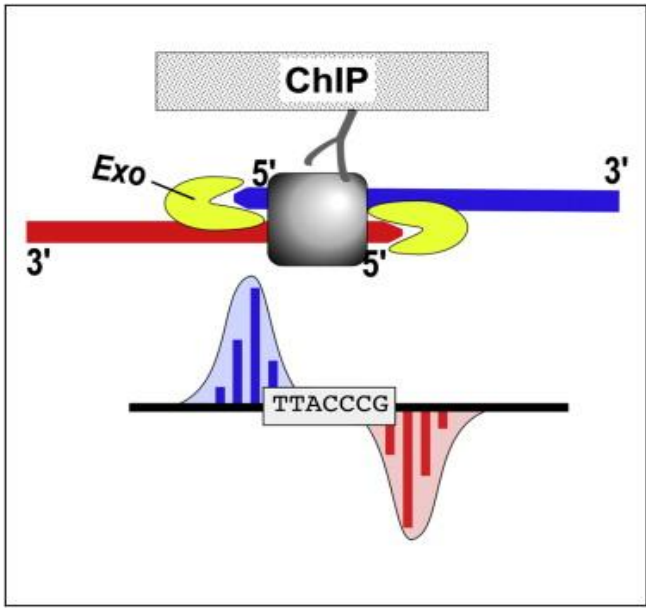
Park 2009 Nat. Rev Genet.

Advances in technologies for nucleic acid-protein interaction detection

- ChIP-chip : combines ChIP with microarray technology.

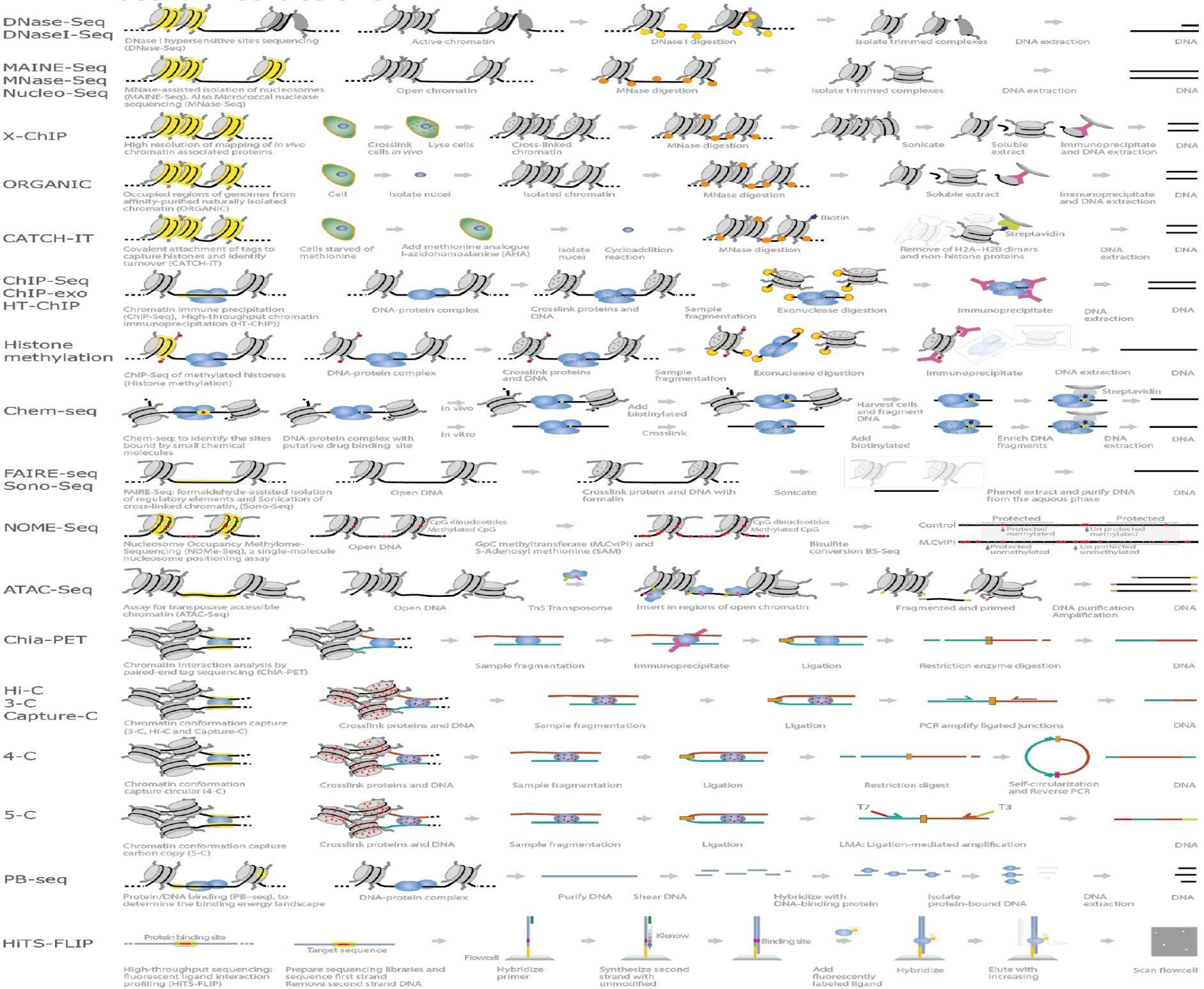
- ChIP-PET : ChIP with paired end tag sequencing

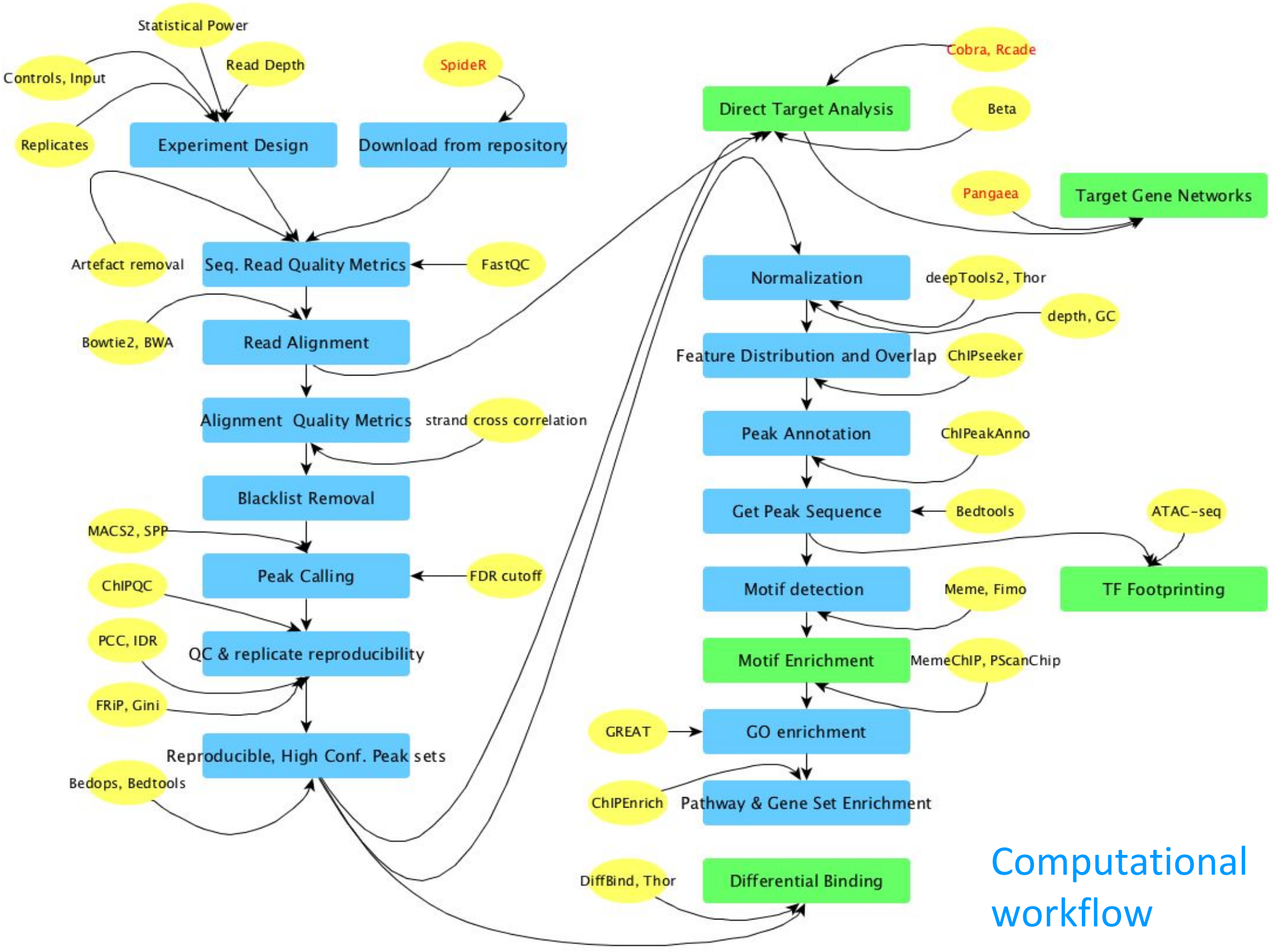
- ChIP-exo : ChIP-seq with exonuclease digestion



- CLIP-seq / HITS-CLIP : cross-linking immunoprecipitation high throughput sequencing
- ATAC-seq : Assay for Transposon Accessible Chromatin
- Sono-seq : Sonication of cross linked chromatin sequencing.
- Hi-C: High throughput long distance chromatin interactions

DNA-Protein Interactions





Computational workflow

Statistical aspects and best practices

These guidelines address :

- Antibody validation
- Experimental replication
- Sequencing depth
- Data and metadata reporting
- Data quality assessment
- Replicates

Experimental guidelines:

- Landt *et al.*, “ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.” *Genome Res. 2012.*
- Marinov *et al.,* “Large-scale quality analysis of published ChIP-seq data.” 2014 *G3*
- Rozowsky *et al.,* "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls" *Nat Biotechnol.* 2009

Statistical aspects:

- Cairns *et al.*, “Statistical Aspects of ChIP-Seq Analysis.” *Adv. in Stat Bioinf.*, 2013.
- Carroll TS *et al.*, “Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data.” *Front Genet.* 2014
- Bailey *et al.*, "Practical guidelines for the comprehensive analysis of ChIP-seq data.” *PLoS Comput Biol.* 2013.
- Sims et al., “Sequencing depth and coverage: key considerations in genomic analyses.” *Nat. Rev. Genet*. 2014.

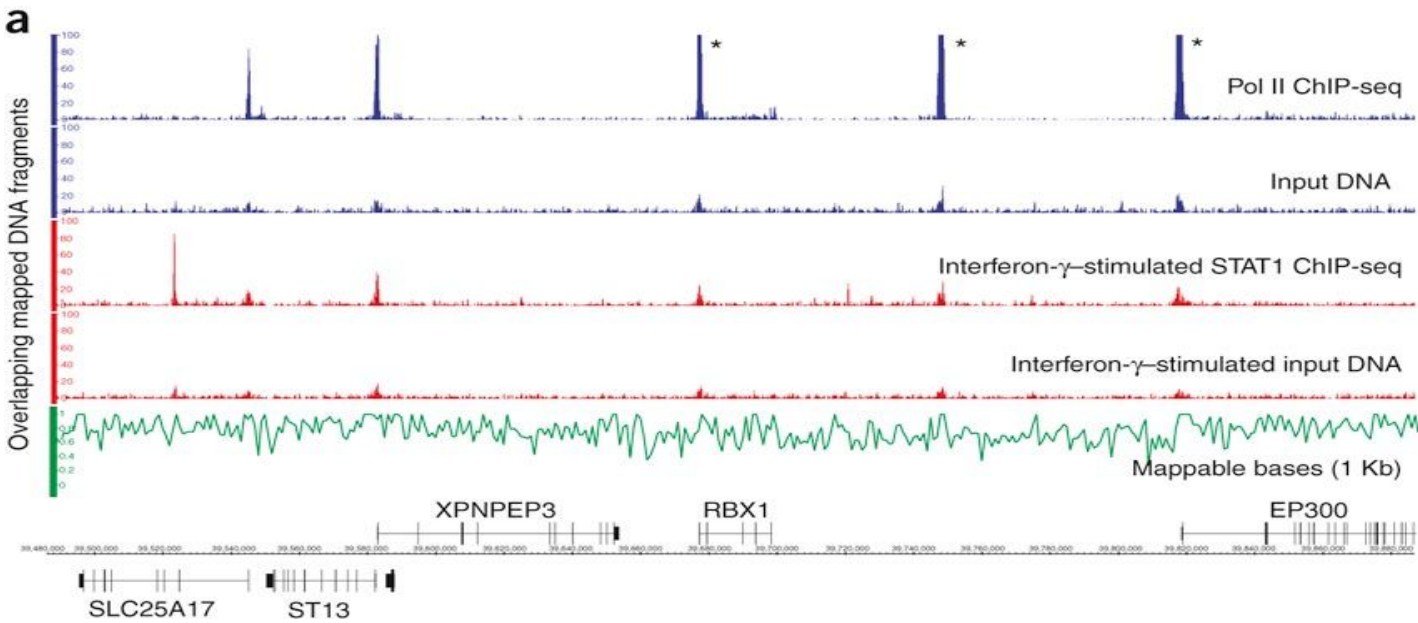
Sequencing depth for ChIP-seq

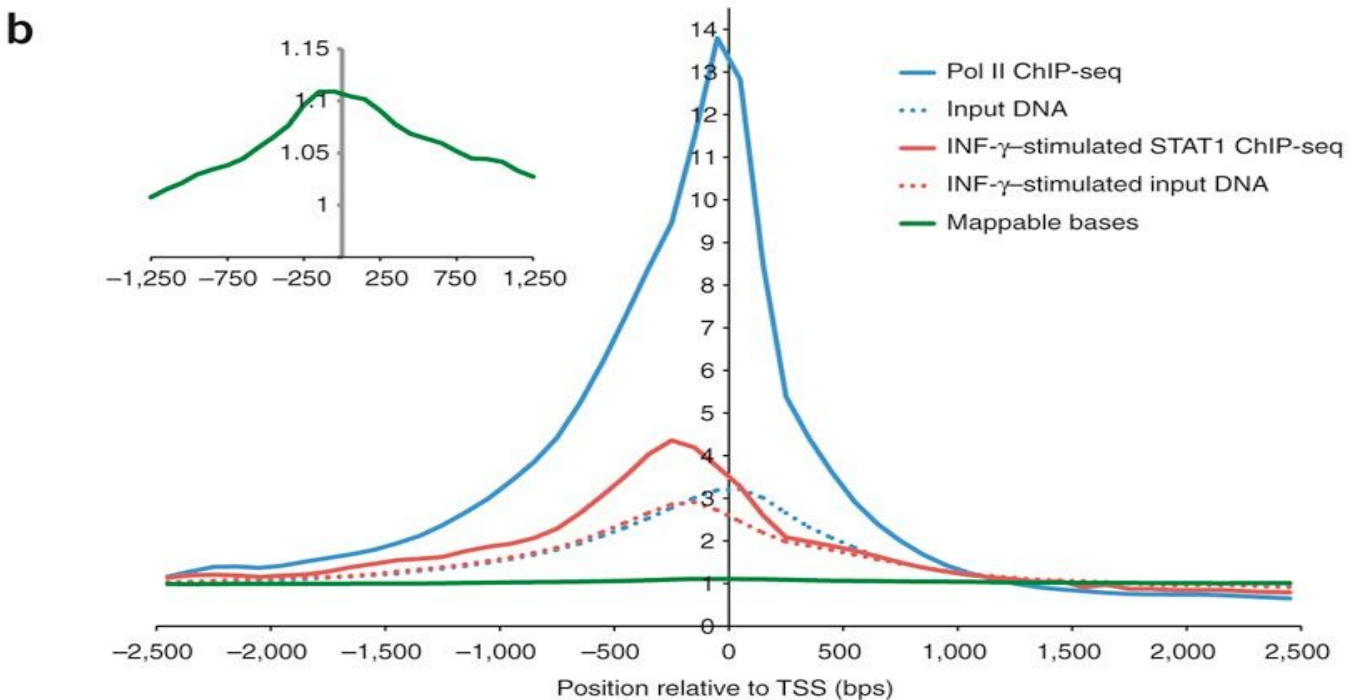
- More prominent peaks are identified with fewer reads, versus weaker peaks that require greater depth
- Number of putative target regions continues to increase significantly as a function of sequencing depth
- Narrow Peaks: 15-20 million reads, Broad Peaks: 20-60 million reads

- https://genohub.com/recommended-sequencing-coverage-by-application/

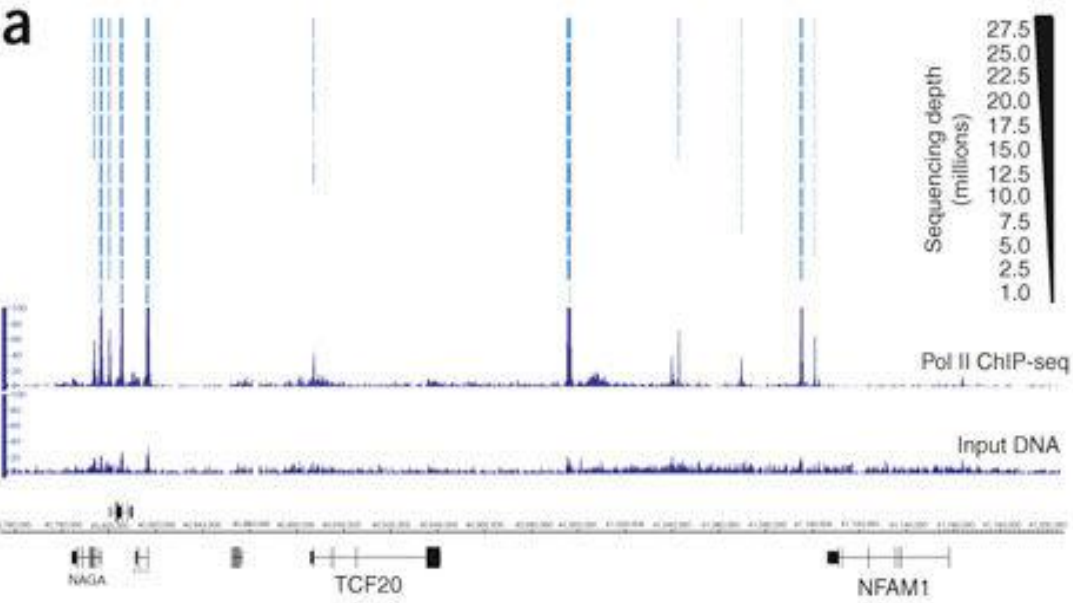
Why we need input controls

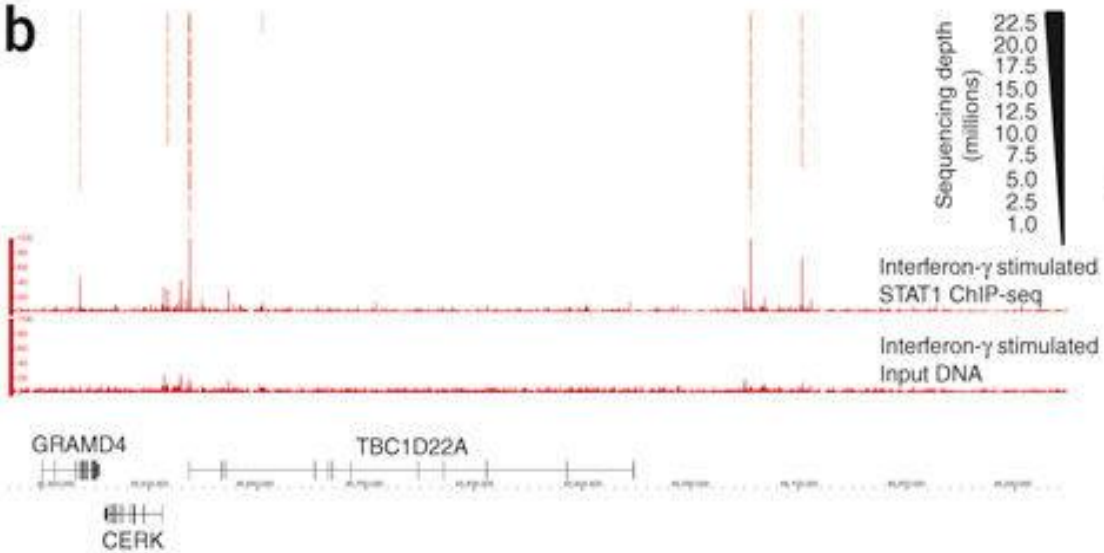
- Open chromatin regions are more easily fragmented than closed regions.
- Uneven read distribution
- Repetitive sequences may appear to be enriched.
- Compare ChIPseq peak with same region in Input control.

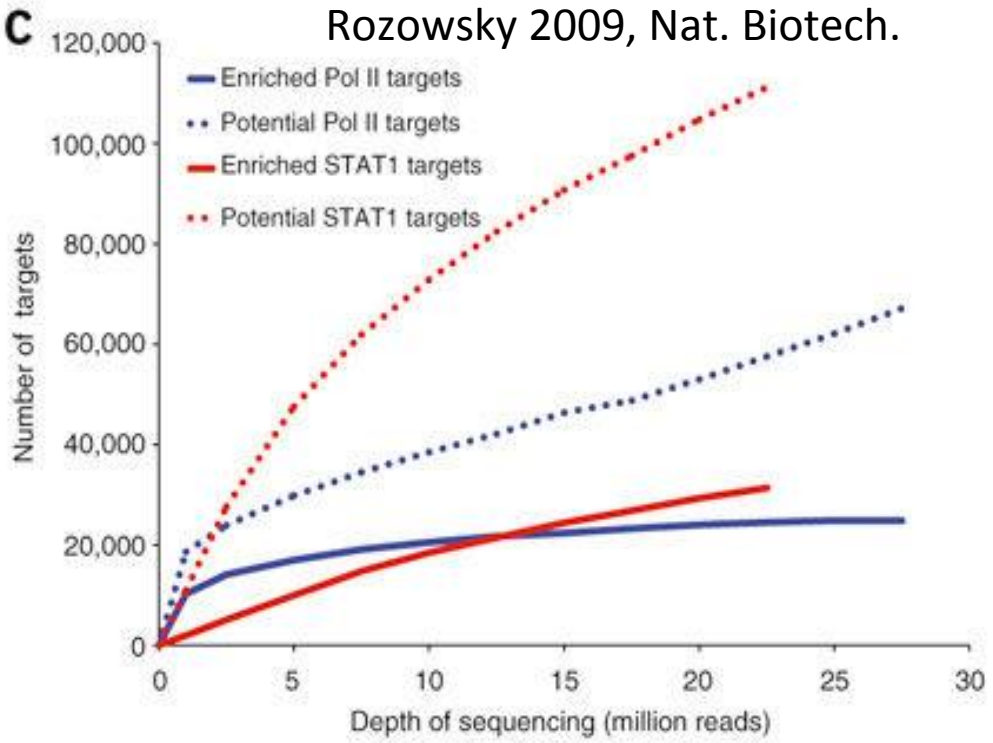




Rozowsky 2009, Nat. Biotech.







d

Number of replicas	Number of targets	True positives	False positives	False negatives	Sensitivity	Positive predictive value
1	20,902	18,833	2,069	5,906	0.761	0.901
2	20,733	19,257	1,476	5,482	0.778	0.929
3	20,328	19,126	1,202	5,613	0.773	0.941

(**a**) Fragment density signal tracks for Pol II and the input-DNA control as well as the target regions that are identified (significantly enriched) as a function of the number of mapped sequence reads. The same numbers of sequence reads are used for both sample and control. More prominent peaks are identified with fewer reads, whereas weaker peaks require greater depth. (**b**) Similar plot with STAT1 and matching interferon-γ–stimulated HeLa input-DNA control. (**c**) The number of putative Pol II (blue line) and STAT1 (red line) targets identified and the fraction for each of these that are enriched relative to input DNA as a function of the number of mapped sequence reads. Although the number of putative targets continues to increase for both Pol II and STAT1, the number of enriched targets begins to plateau. The number of Pol II targets appears to saturate faster than for STAT1 targets. (**d**) Summarized results of analyzing 9 million mapped Pol II ChIP-seq sequence reads using one, two or three biological replicates. We calculate sensitivity and positive predictive values using the targets identified with all the available sequence reads (~29 million uniquely mapped reads) as gold standard positives and the remainder as negatives. Only a marginal gain in positive predictive value at the cost of sensitivity is gained by using three biological replicates instead of two biological replicates.

Artefact removal 1: Blacklisted regions

- Once reads have been aligned to the reference genome, “blacklisted regions” are removed from **BAM** files before peak calling.

- **Blacklisted** regions are genomic regions with anomalous, unstructured, high signal or read counts in NGS experiments, independent of cell type or experiment.

- These regions tend to have a very high ratio of multi-mapping to unique mapping reads and a high variance of mappability and simple mappability filters do not account for them.

- These regions are often found at repetitive regions (Centromeres, Telomeres, Satellite repeats) and are troublesome for high throughput sequencing aligners and when computing genome wide correlations.

- These regions also confuse peak callers and result in spurious signal.

Artefact removal 2

• The *DAC Blacklisted Regions* aim to identify a comprehensive set of regions in the human genome that have anomalous, unstructured, high signal/read counts in NGS experiments, independent of cell line and type of experiment.

80 open chromatin tracks (DNase and FAIRE data-sets) and 20 ChIP-seq input/control tracks spanning ~60 human tissue types/cell lines in total used to identify these regions with signal artefacts. These regions tend to have a very high ratio of multi-mapping to uniquely mapping reads and high variance in mappability. The *DAC Blacklisted Regions* track was generated for the ENCODE project.

• The *Duke Excluded Regions* contains problematic regions for short sequence tag signal detection (such as satellites and rRNA genes).

• *Grey Lists* represent regions of high artefact signals that are specific to your cell-type or sample, and can be tuned depending on the stringency required.

Artefact removal 3

Resources:

Where to get Blacklist BED file:

- https://sites.google.com/site/anshulkundaje/projects/blacklists

How they were generated:

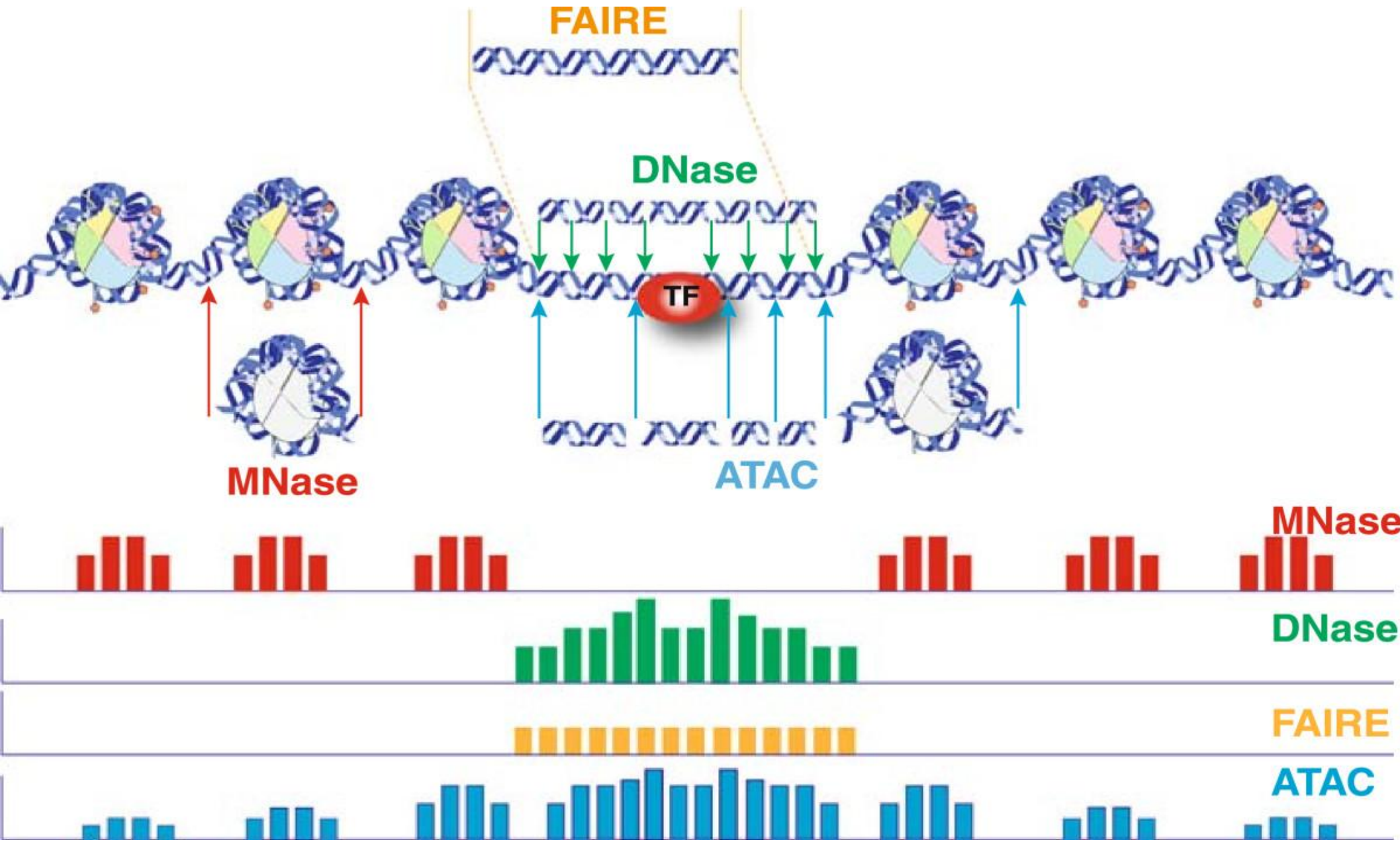
- https://docs.google.com/file/d/0B26FxqAtrFDwWGFCdXE1SlFYRmM/edit

ChIPseq Quality control :

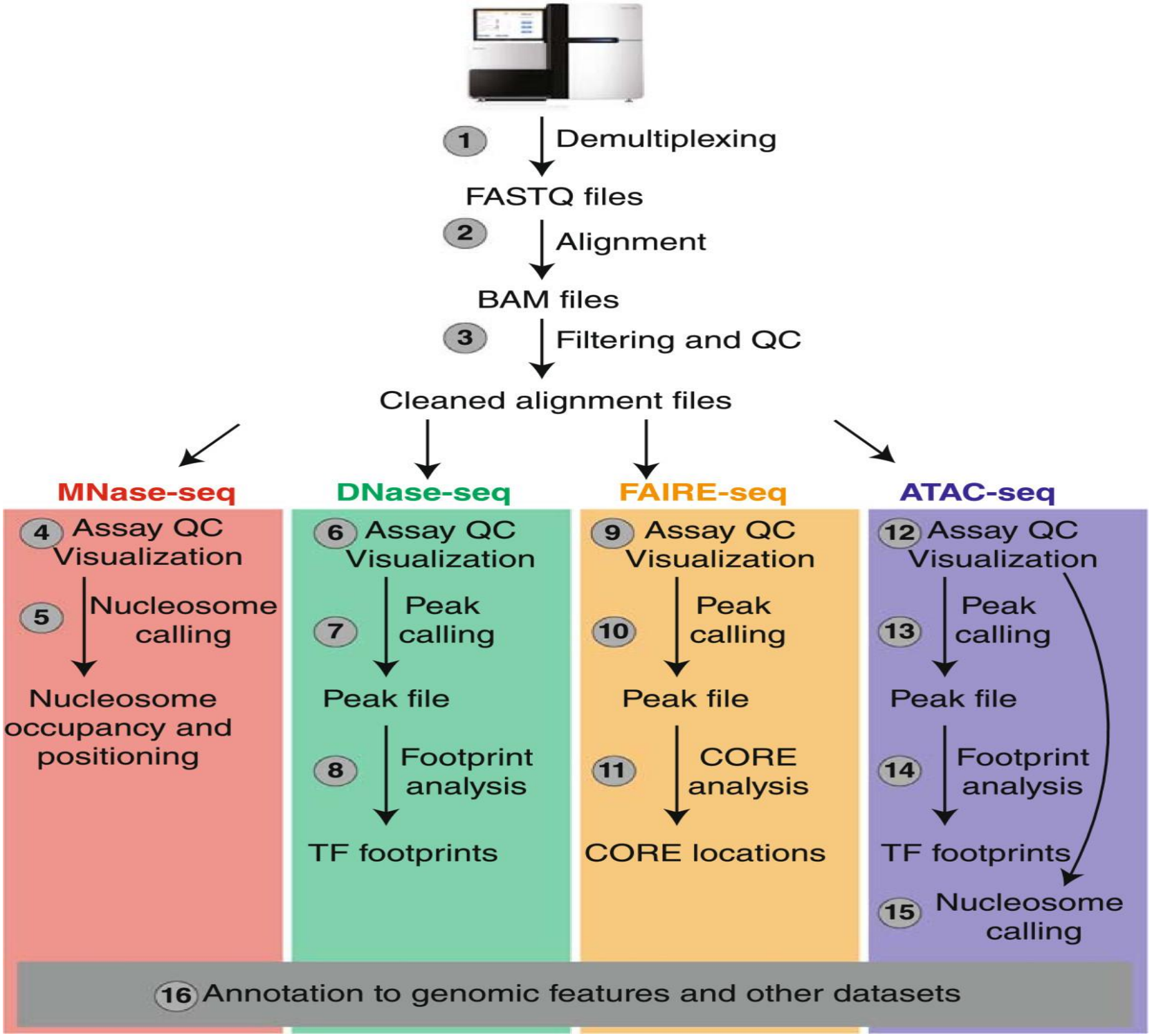
- Carroll *et al.,* "Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data." *Front Genet.* 2014

- **GreyListChIP**
- **ChIPQC**

Detecting Chromatin Accessibility

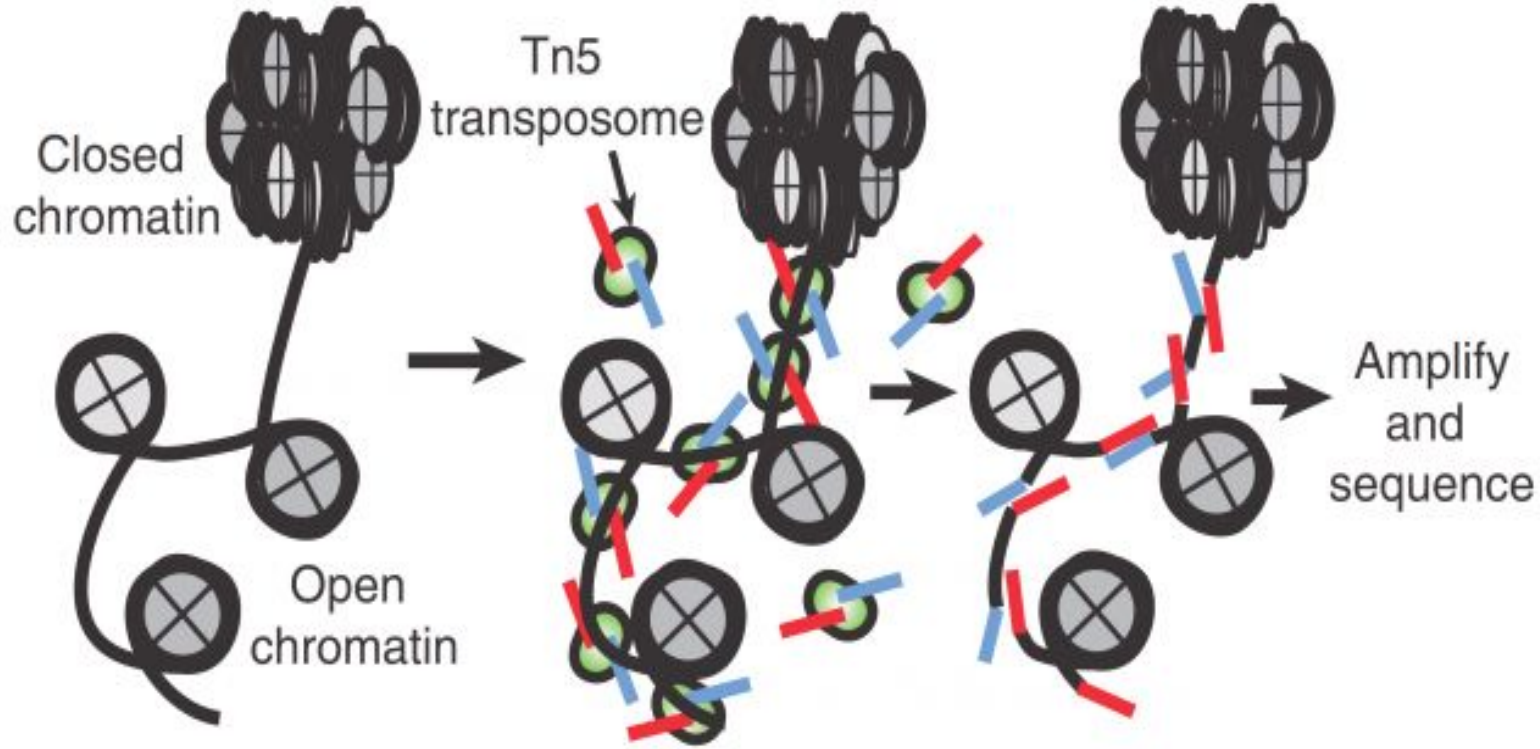


Tsompana and Buck, *Epigenetics & Chromatin* 2014



Tsompana and Buck, *Epigenetics & Chromatin* 2014

ATAC-seq

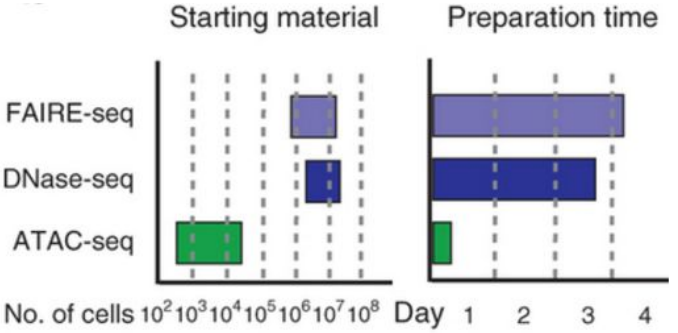


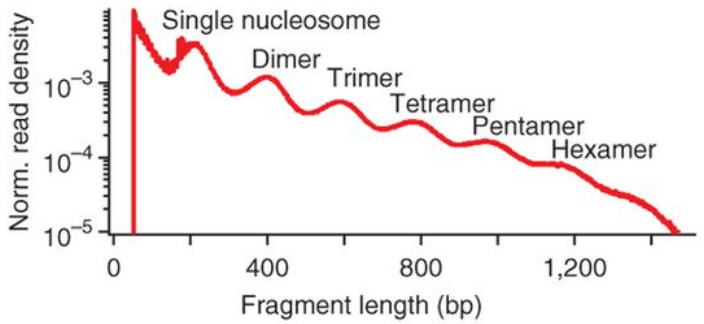
- Enables measurement of chromatin structure modifications (nucleosome free regions) on gene regulation.
- Does not require antibodies or tags that can introduce potential bias.
- Hyperactive Tn5 transposase is used to fragment DNA and integrate into active regulatory regions.
- During ATAC-seq, 500–50,000 unfixed nuclei are tagged *in vitro* with sequencing adapters by purified Tn5 transposase.
- Can also detect nucleosome packing, positioning and TF footprints.

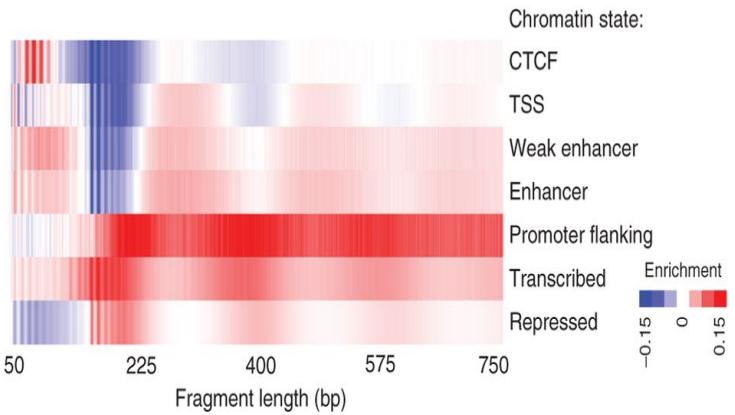
JD Buenrostro et al, Nature Methods, 2013.

ATAC-seq

- Two-step protocol
  - Insertion of Tn5 transposase with adaptors
  - PCR amplification
- Needs ~500-50,000 cells
- Paired-end reads produce information about nucleosome positioning.
- Insert size distribution of fragments has a periodicity of ~200bp, suggesting that fragments are protected by multiplies of nucleosomes
- Different fragmentation patterns can be associated with different functional states (eg. TSSs are more accessible than promoter flanking or transcribed regions)

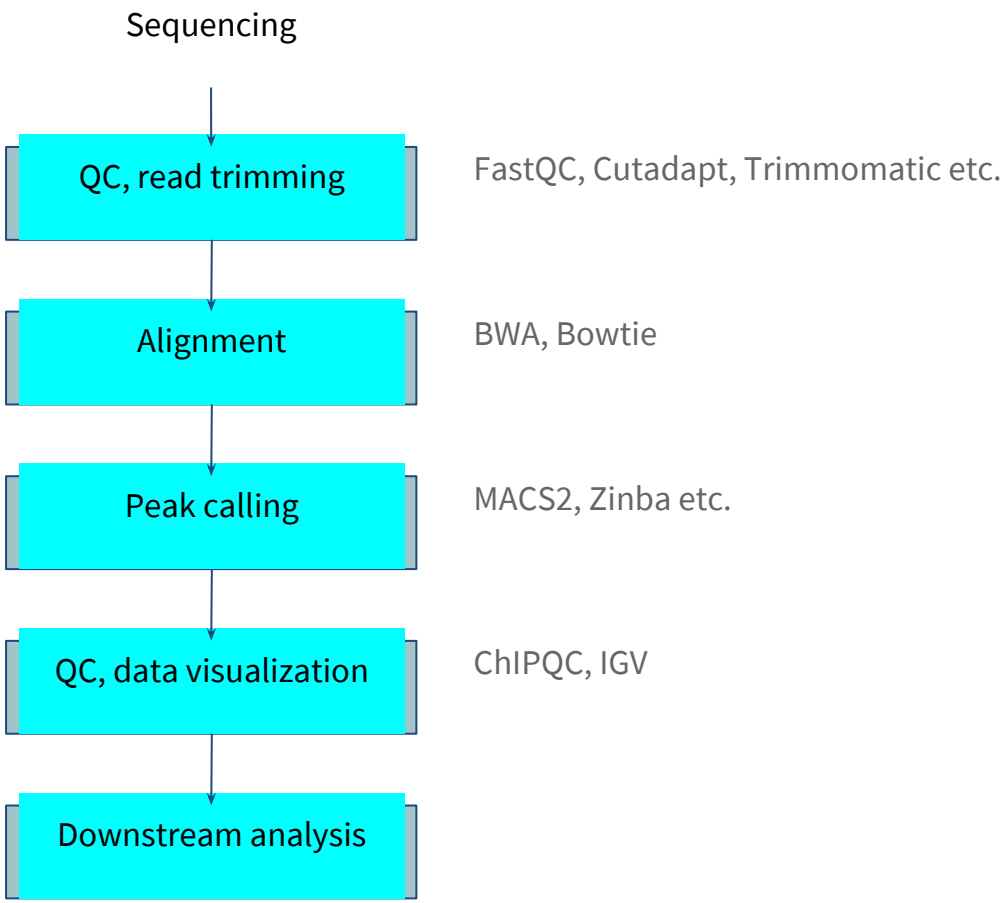






[1] JD Buenrostro et al, *Nature Methods*, 2013.

Workflow of ATAC-seq data processing



Differences from ChIP-seq data processing

- Use the fragment length for smoothing when calling peaks with MACS2
  - MACS2 documentation says when using DNAse-seq type data:
    - "... all 5' ends of sequenced reads should be extended in both direction to smooth the pileup signals. If the wanted smoothing window is 200bps, then use '--nomodel --shift -100 --extsize 200"
  - --nomodel: don't build shifting model
  - --shift: when this value is negative, ends will be moved toward 3'->5' direction
  - --extsize: extend reads in 5'->3' direction to fix-sized fragments
  - Use the fragment size for smoothing - you can calculate it with ChIPQC
- Remove mitocondrial reads
  - A large fraction of ATAC-seq reads map to mitocondrial genome (up to 40-60%) that you will want to remove
  - Blacklisted regions contain the mitocondrial genome
- Normalisation across samples might be needed
  - Efficiency of the ATAC-seq protocol in assaying open regions might be different based on how much transposome gets into nuclei
  - For a solution of normalisation see: *Sarah K. Denny et al, Cell, 2016.*

Peak Calling

- Identifies TF binding sites
- Count based - Define regions. Count the number of reads falling into each region. When a region contains a statistically significant number of reads, call that region a peak.
- Shape based - Consider individual candidate binding sites. Model the spatial distribution of reads in surrounding regions, and call a peak when the read distribution conforms to the expected distribution near a binding site.