Introduction to ChIP-seq

Shamith Samarajiwa
MRC Cancer Unit
University of Cambridge

CRUK Bioinformatics Summer School 2021
27th July 2021

What is ChIP-seq?

- A combination of chromatin immunoprecipitation (ChIP) with ultra high-throughput massively parallel sequencing. The typical ChIP assay usually take 4–5 days, and require approx. 10^6 ~ 10^7 cells.

- Allows mapping of Protein–DNA interactions or chromatin modifications ***in vivo*** on a **genome scale**.

- Enables investigation of
  - Transcription Factor binding
  - DNA binding proteins (HP1, Lamins, HMGA etc)
  - RNA Pol-II occupancy
  - Histone modification marks

- Single cell ChIP-seq is also possible (*Rotem et al, 2015 Nat. Biotech.*)

The Chromatin Immunoprecipitation Method

DNA binding protein ChIP-seq

Croslinking

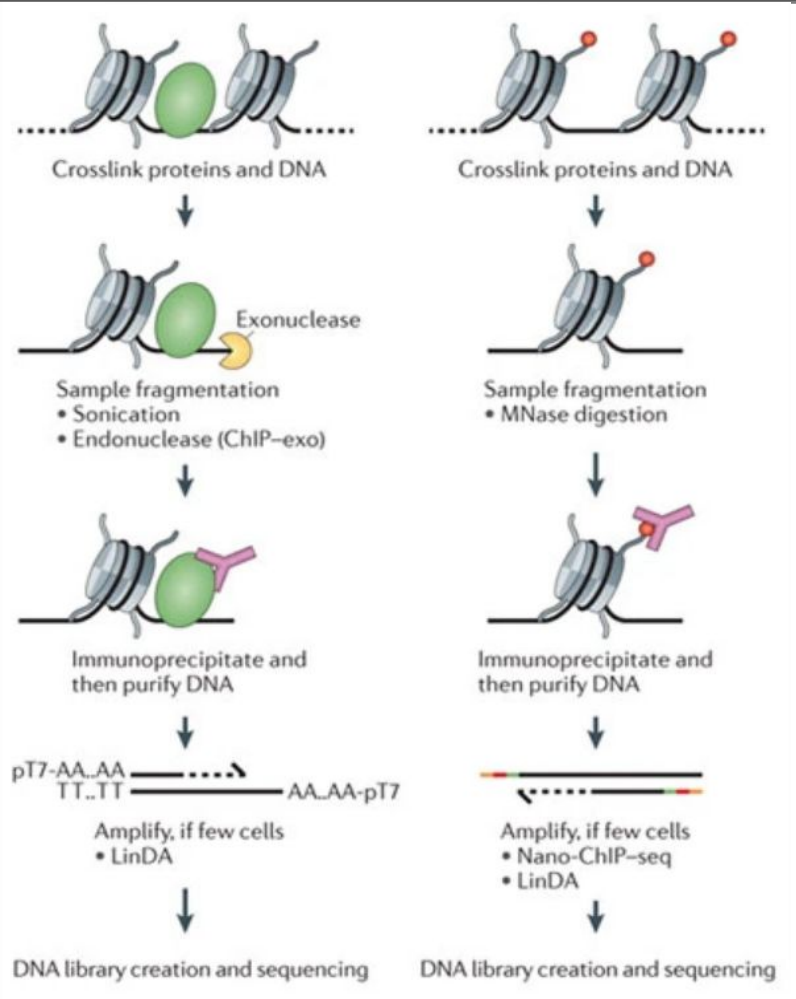
Fragmentation

Co-IP

Cross link removal and DNA fragment purification

Library Preparation

Sequencing



Histone ChIP-seq

Important Considerations

- Good Experimental Design!
- Optimize Conditions (Cells, Antibodies, Sonication etc.)
- Sufficient amount of starting material (ChIP DNA depends on cell type, abundance of the mark or protein, quality of antibody etc)
- **Biological Replicates (at least 2 or more)!!**
  - sample biological variation & improve signal to noise ratio
  - capture the desired effect size
  - statistical power to test null hypothesis
- ChIP-seq controls – **Knockouts or Knockdowns, Input Control** (Try not to use IgG)

What generates a ChIP signal?

ChIP-Seq signal depends on:

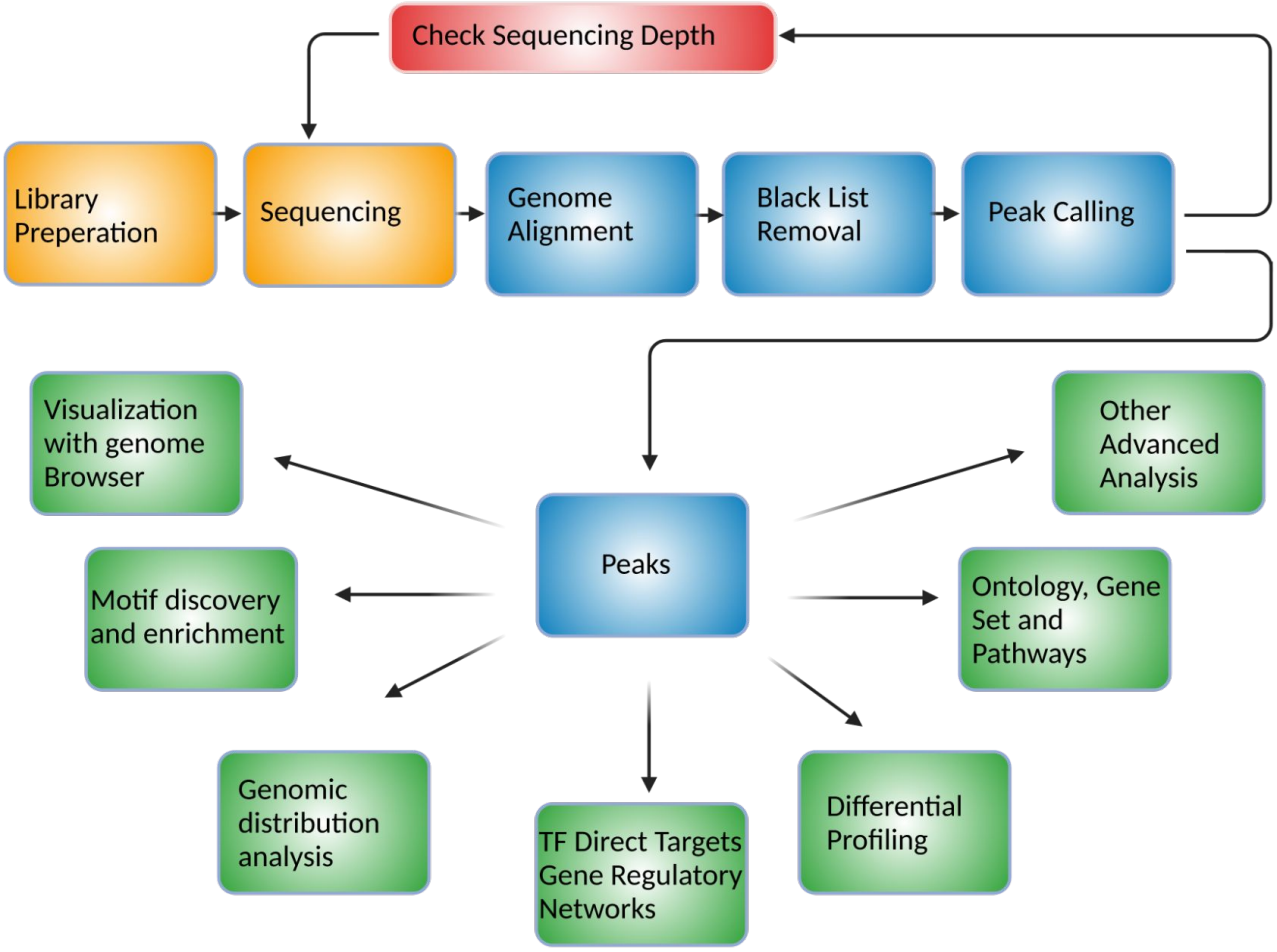
- The number of active binding sites
- The number of starting genomes (number of cells)
- IP efficiency (antibody quality, biological model used)
- GC rich content (bias in fragment selection, during amplification)

Globally

- Open chromatin regions fragment more easily than closed regions (open region will generate more reads than closed one due to non-random fragmentation)
- Differential mappability of short reads to repeat-rich genomic regions (Teytelman et al., 2009, Aird et al., 2011)
- Hyper-ChIPable regions

Locally

ChIP-seq workflow

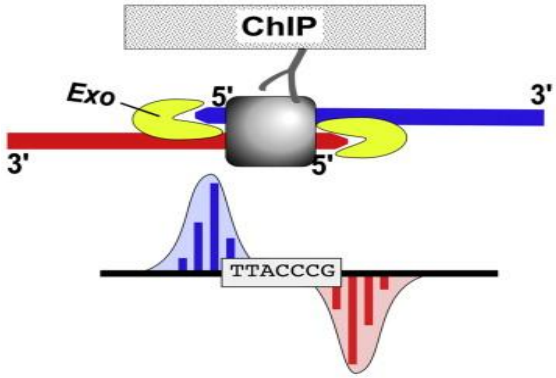


Modified from Furey, T. Nat Rev Genet 13, 840–852 (2012)

Technologies for nucleic acid-protein interaction detection

Older technologies:

- ChIP-chip: combines ChIP with microarray technology.
- ChIP-PET: ChIP with paired end tag sequencing



Newer:

- ChIP-exo: ChIP-seq with exonuclease digestion
- HiChIP: Chromatin interactions and histone modification
- CUT&Tag: Cleavage Under Targets and Tagmentation
- CLIP-seq / HITS-CLIP/ iCLIP: cross-linking immunoprecipitation high throughput sequencing for RNA-Protein binding
- Sono-seq: Sonication of cross linked chromatin sequencing.
- Hi-C: High throughput long distance chromatin interactions
- DRIP-seq: R-loop (DNA-RNA) interaction detection
- ATAC-seq: Assay for Transposon Accessible Chromatin

Best Practices

These guidelines address :

- Antibody validation (IP specificity and quality)
- Experimental replication and controls
- Biological replicates
- Sequencing depth
- Data quality assessment
- Data and metadata reporting

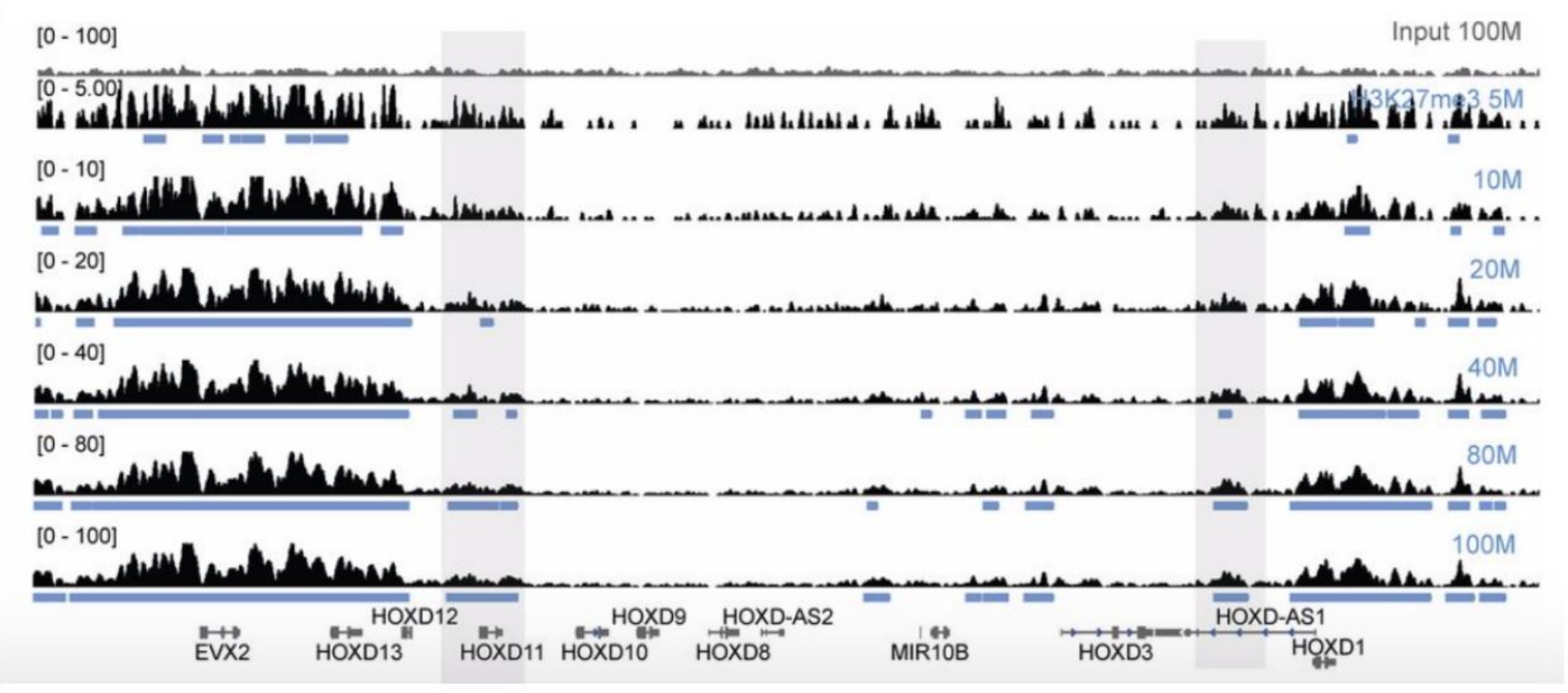
Experimental guidelines:

- 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.
- Tian et al., "Identification of factors associated with duplicate rate in ChIP-seq data" PLoS ONE 2019.

ChIP-Seq signal & sequencing depth



Rule of thumb: More prominent peaks are identified with fewer reads, compared to weaker peaks.

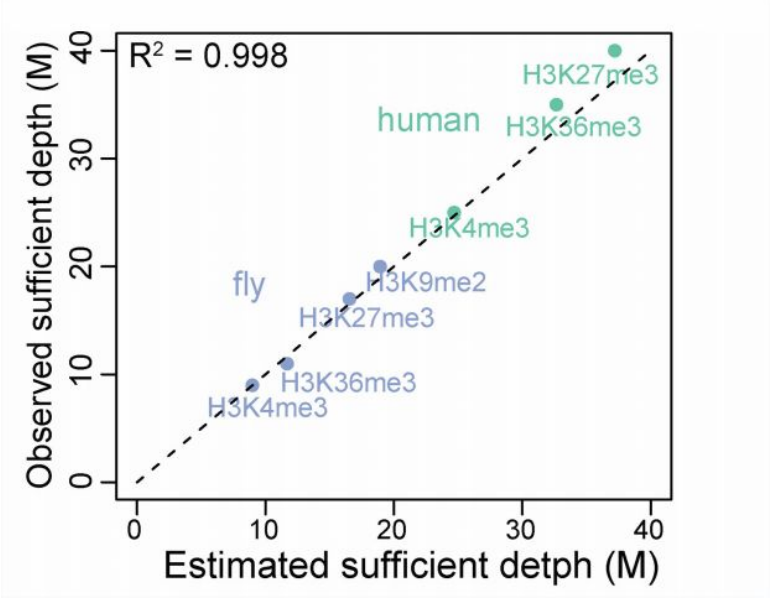
Jung et al, Nucleic Acids Res. 2014

Sequencing Depth

- Number of putative target regions continues to increase significantly as a function of sequencing depth

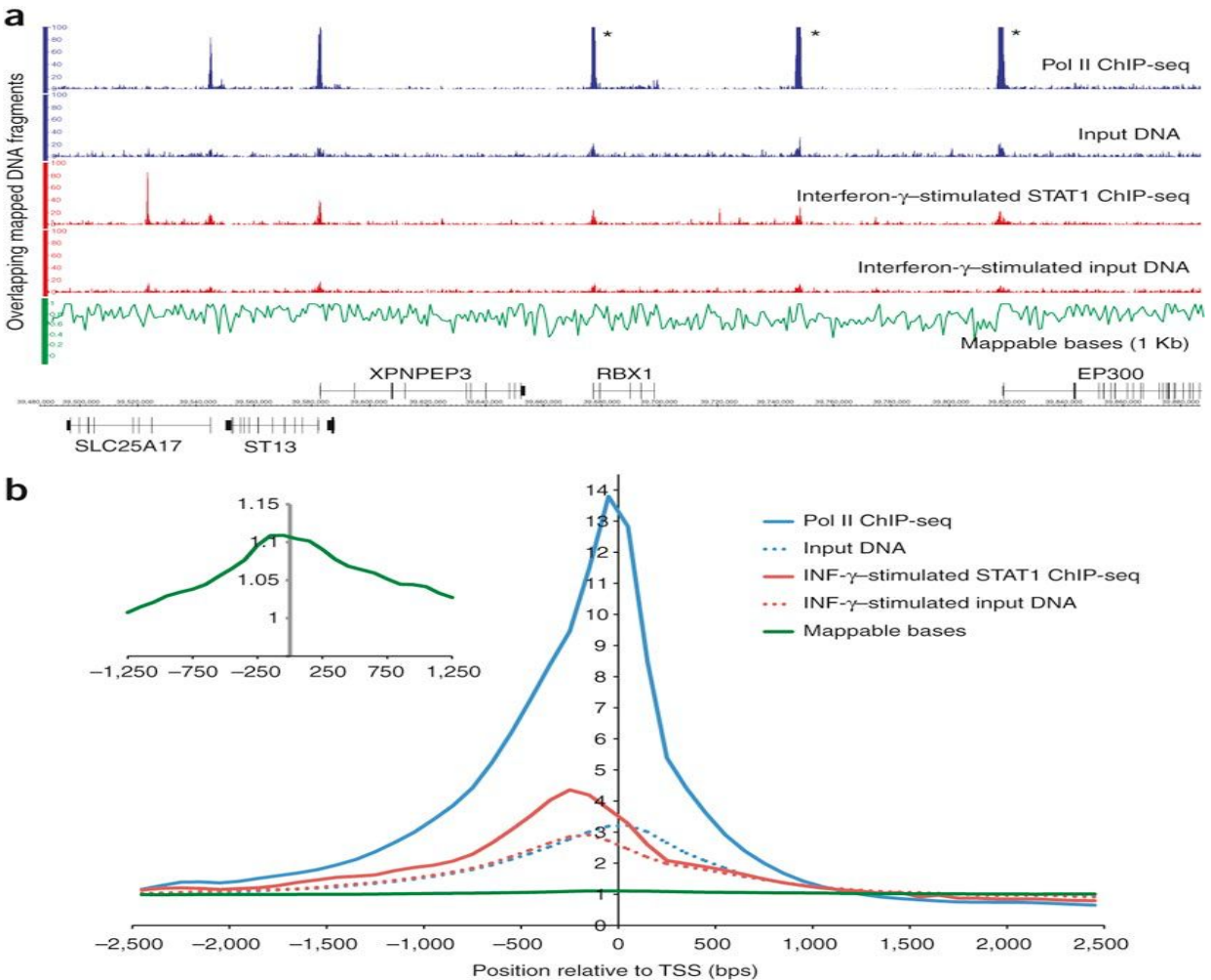
https://genohub.com/recommended-sequencing-coverage-by-application

- Narrow Peaks: 10-20 million reads
- Broad Peaks: 20-40 million reads

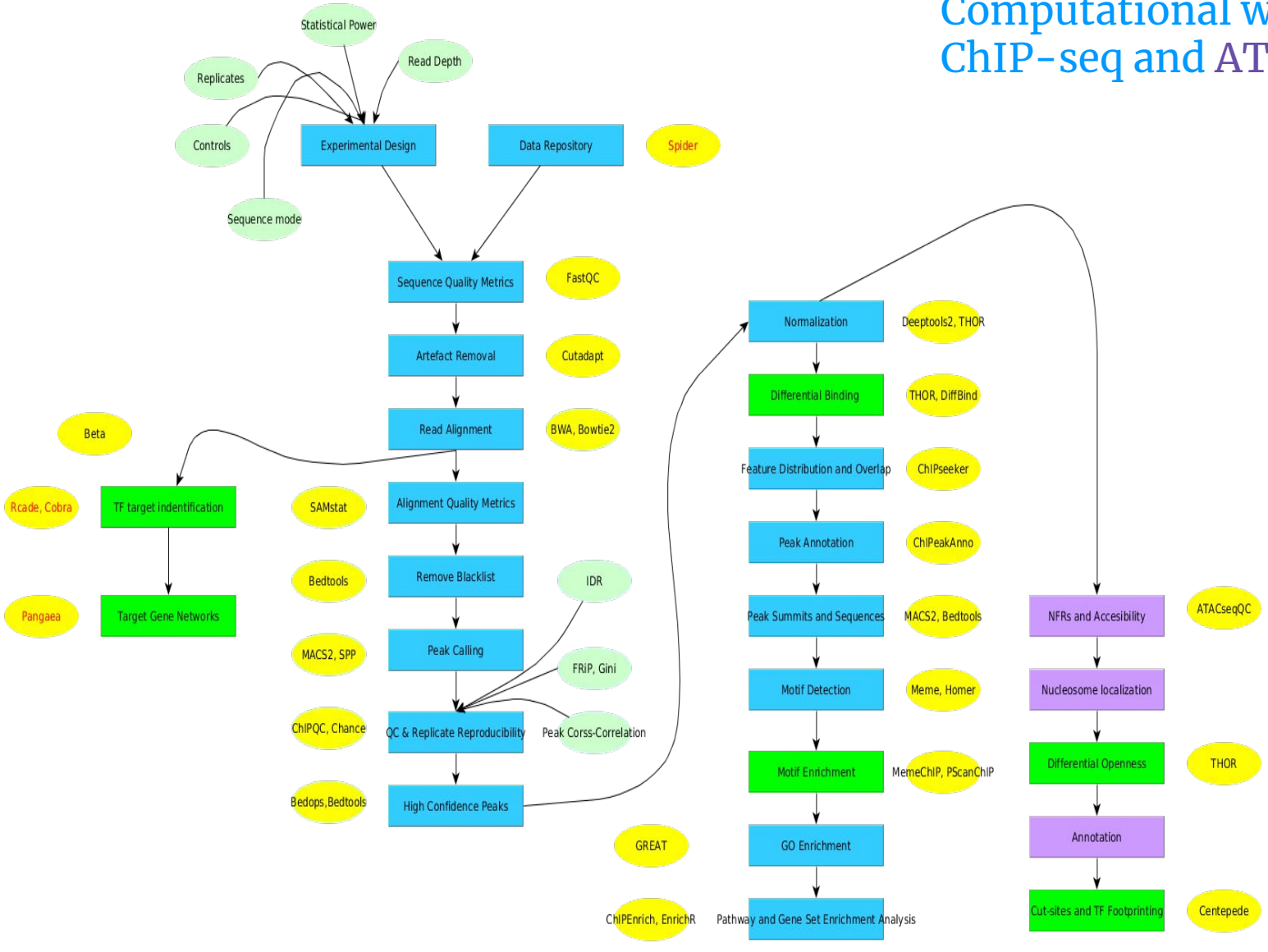


Why we use an Input Control

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



Computational workflow for ChIP-seq and ATAC-seq



Decoy and Sponge databases

- **The decoy** contains human sequences missing from the hg19 reference, mitochondrial sequences and viral sequences integrated into the human genome. blog article on decoys
- **The sponge** contains ribosomal and mitochondrial sequences, non-centromeric Huref sequences absent in GRCh38 (hg38), centromeric models etc (Miga et al., 2015).
- These mop up ambiguous sequences, resulting in more accurate and faster alignment.

Nucleic Acids Research

Nucleic Acids Res. 2015 Nov 16; 43(20): e133. PMCID: PMC4787761
Published online 2015 Jul 10. doi: 10.1093/nar/gkv671

Utilizing mapping targets of sequences underrepresented in the reference assembly to reduce false positive alignments

Karen H. Miga,* Christopher Eisenhart, and W. James Kent

Black List Removal

- **Blacklisted** regions are genomic regions with anomalous, unstructured, high signal or read counts in NGS experiments, independent of cell type or experiment. Often found at repetitive regions (Centromeres, Telomeres, Satellite repeats)

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

- Problems:
  - tend to have a very high ratio of multi-mapping reads
  - high variance of mappability
  - Difficult to remove with simple mappability filters.

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

Black Lists

- ENCODE DAC Exclusion List Regions download: https://www.encodeproject.org/annotations/ENCSR636HFF/

The ENCODE Blacklist: Identification of Problematic Regions of the Genome

Haley M. Amemiya, Anshul Kundaje ✉ & Alan P. Boyle ✉

Scientific Reports **9**, Article number: 9354 (2019) | Cite this article

Grey Lists

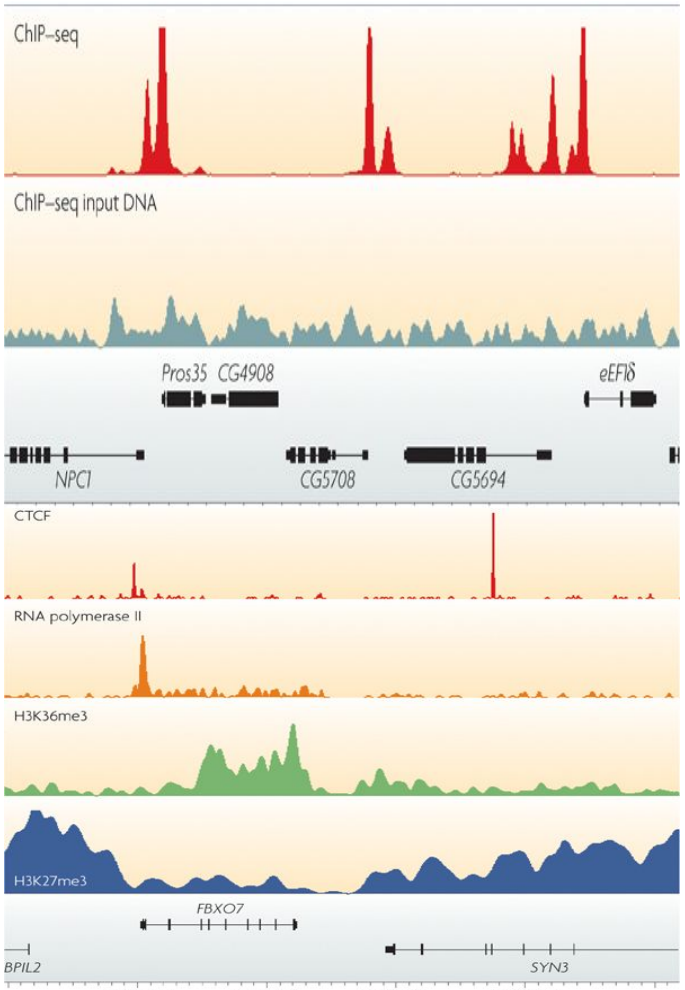
- Grey Lists represent regions of high artefact signals that are specific to cell-lines or tumour samples, and can be tuned depending on the stringency required.

- GreyListChIP bioconductor package can identify those spurious regions, so that reads in those regions can be removed prior to peak calling, allowing for more accurate insert size estimation and reducing the number of false-positive peaks.

Brown G (2021). *GreyListChIP: Grey Lists – Mask Artefact Regions Based on ChIP Inputs*.

Peak Calling

- Identifies TF binding sites or regions of histone modification.
- 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.



Sharp Peaks

input

Sharp Peaks

Broad Peaks