Introduction to ChIP-seq

Shamith Samarajiwa

CRUK Summer School in Bioinformatics

July 2019

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. 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 possible (*Rotem et al, 2015 Nat. Biotech.*)

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

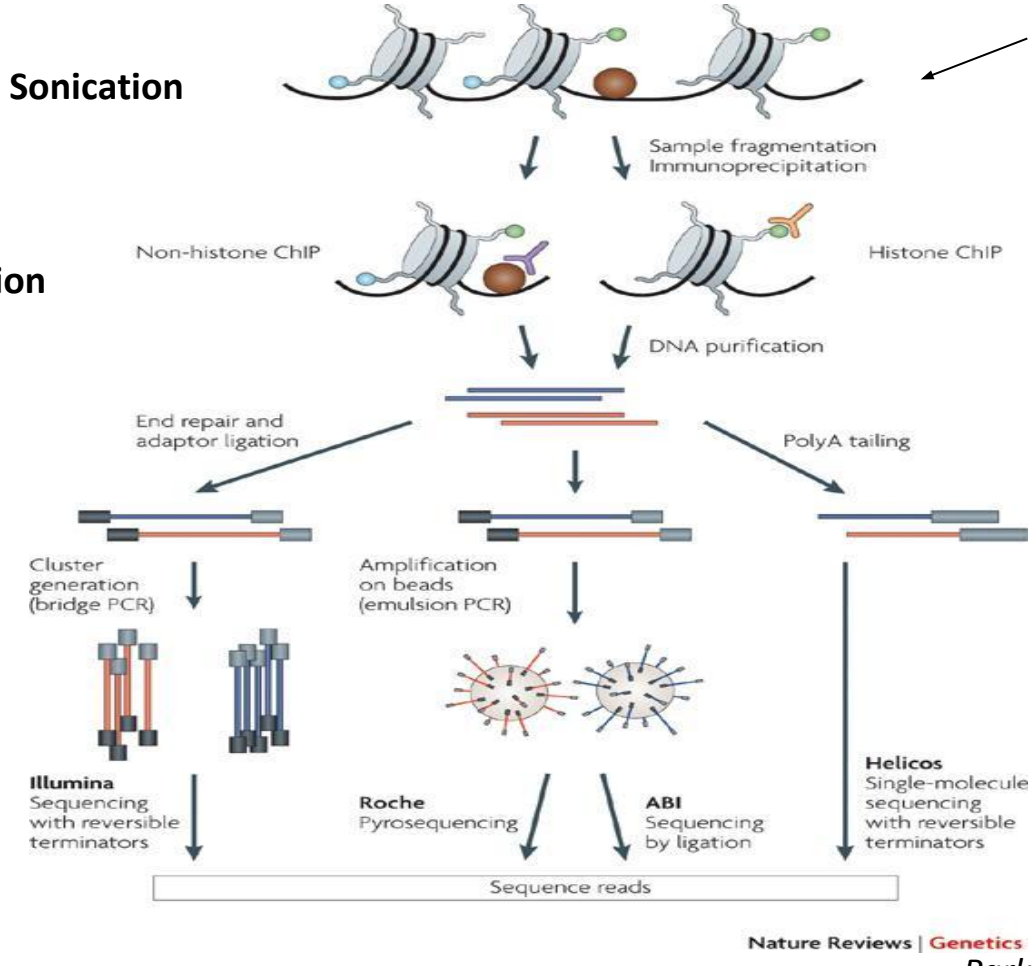
ChIP-seq methodology

Cross-link cells

Isolate genomic DNA

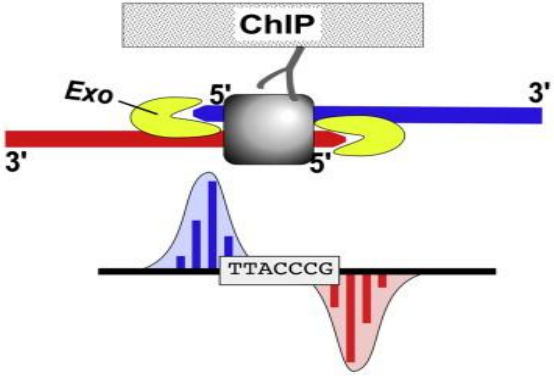
Sonication

Immuno-precipitation



Park 2009 Nat. Rev Genet.

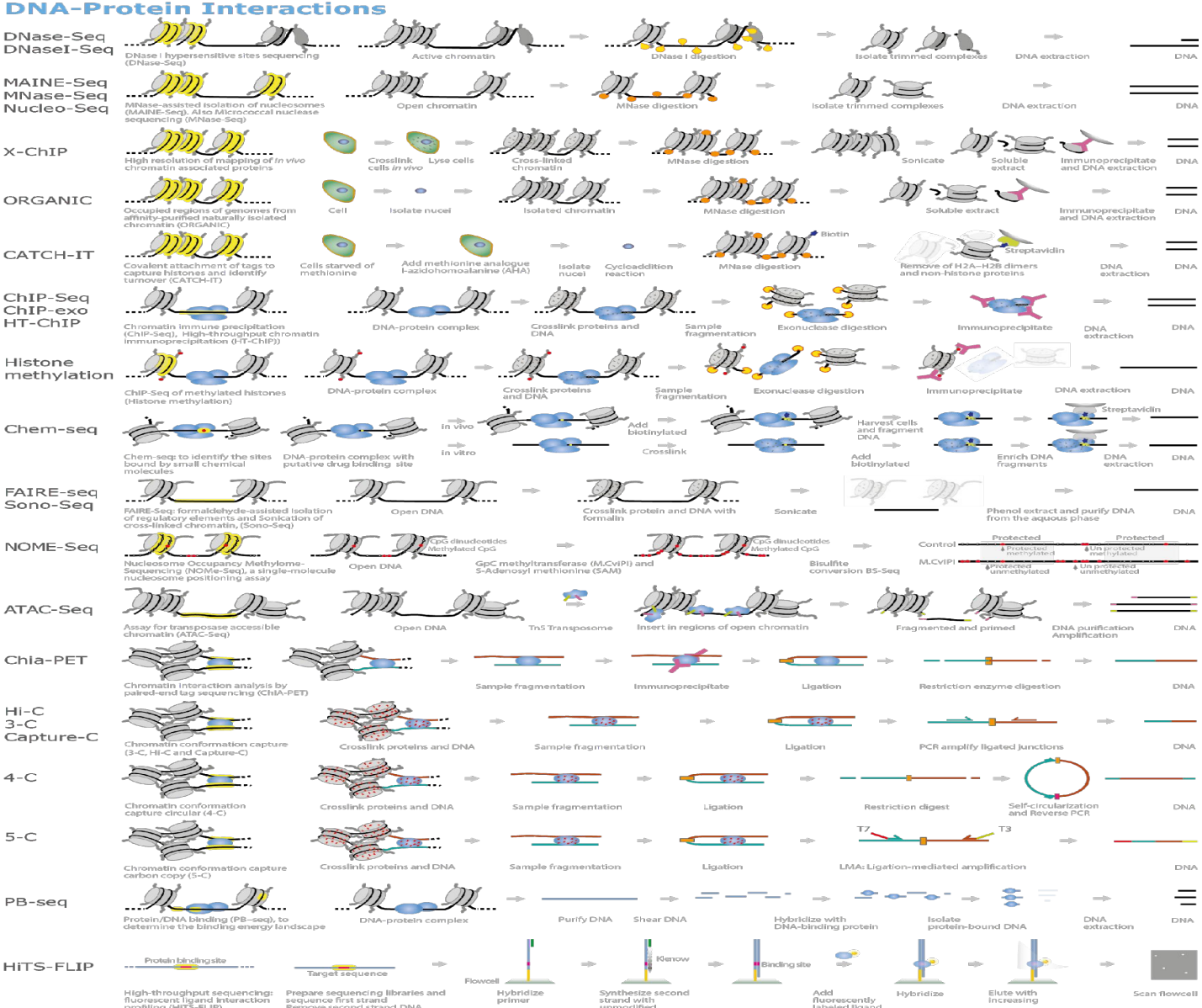
Advances in technologies for nucleic acid-protein interaction detection



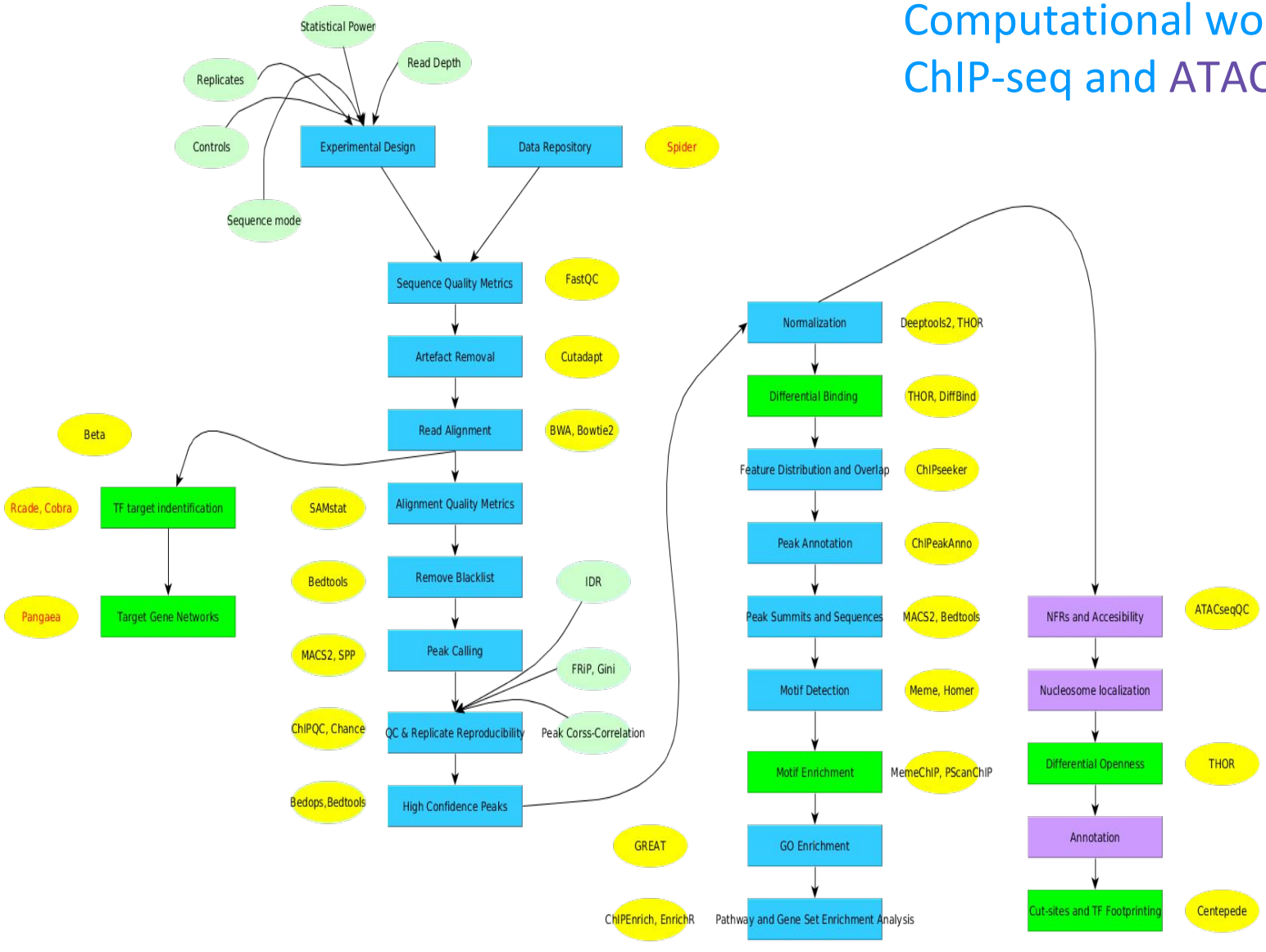
- ChIP-chip : combines ChIP with microarray technology.
- ChIP-PET : ChIP with paired end tag sequencing

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

DNA-Protein Interactions



Computational workflow for ChIP-seq and ATAC-seq



Statistical aspects and 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:

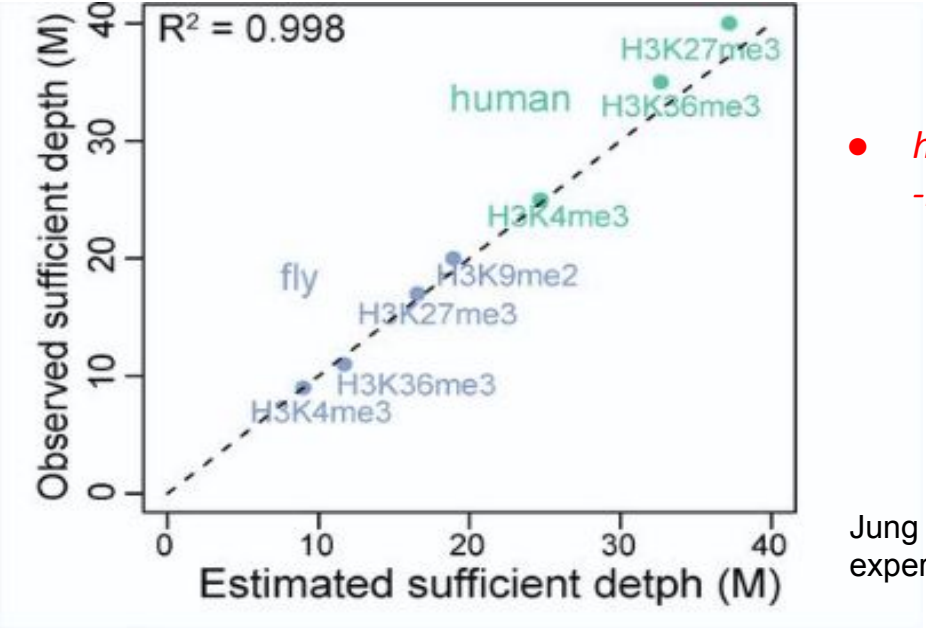
- Landt *et al.*, “ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.” *Genome Res. 2012.*
- Marinov *et al.,* “Large-scale quality analysis of published ChIP-seq data.” 2014 *G3*
- Rozowsky *et al.,* "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls"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

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

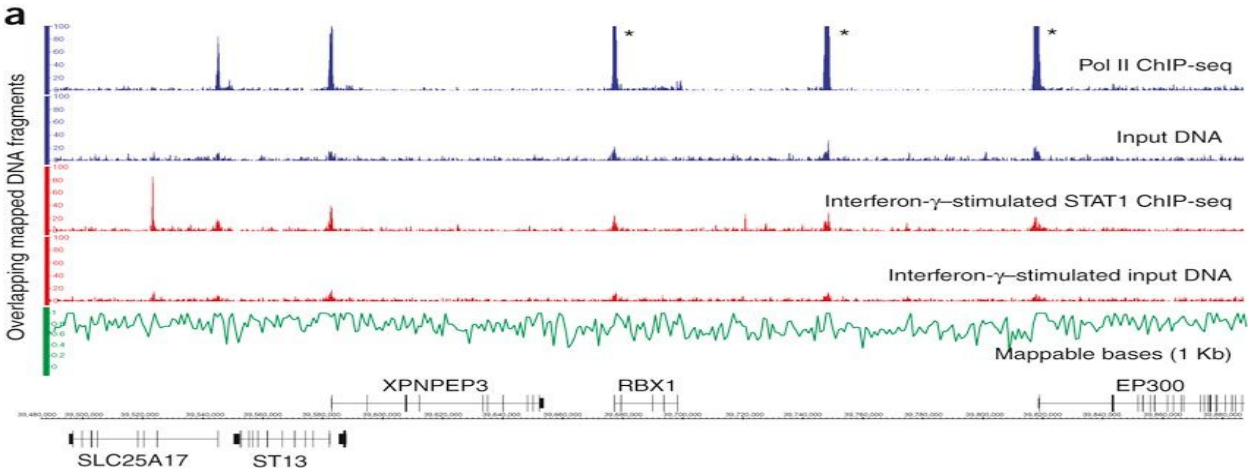


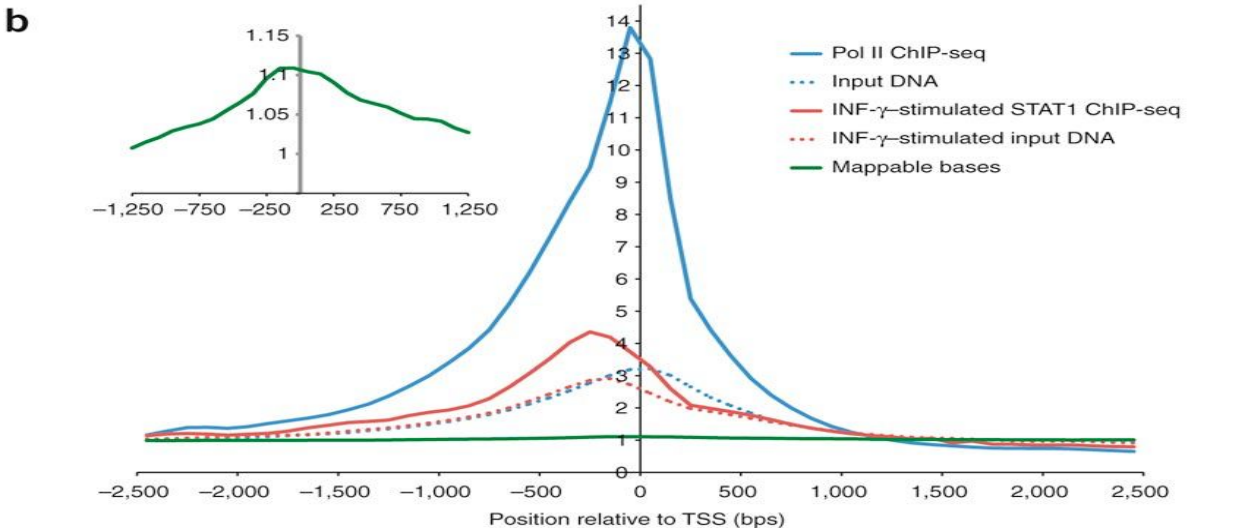
- *https://genohub.com/recommended-sequencing-coverage-by-application/*

Jung et al., 2014 Impact of sequencing depth in ChIP-seq experiments. *Nucleic Acids Res.*

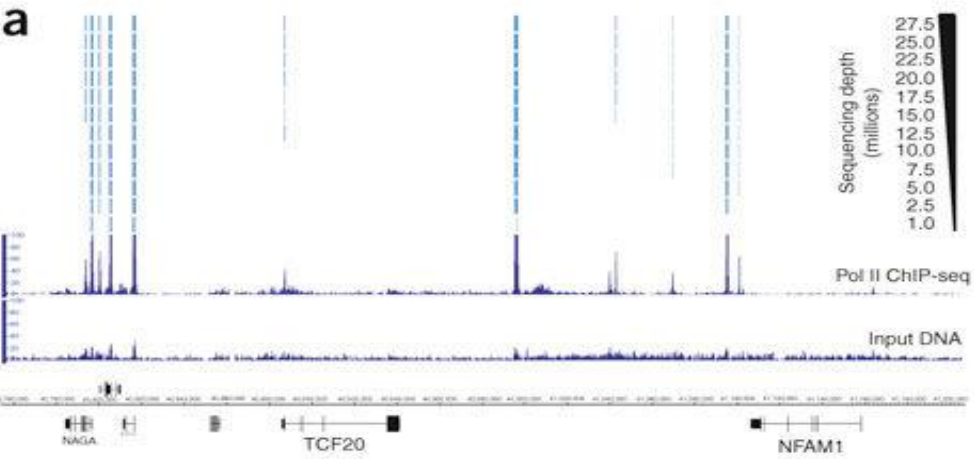
Why we need input controls

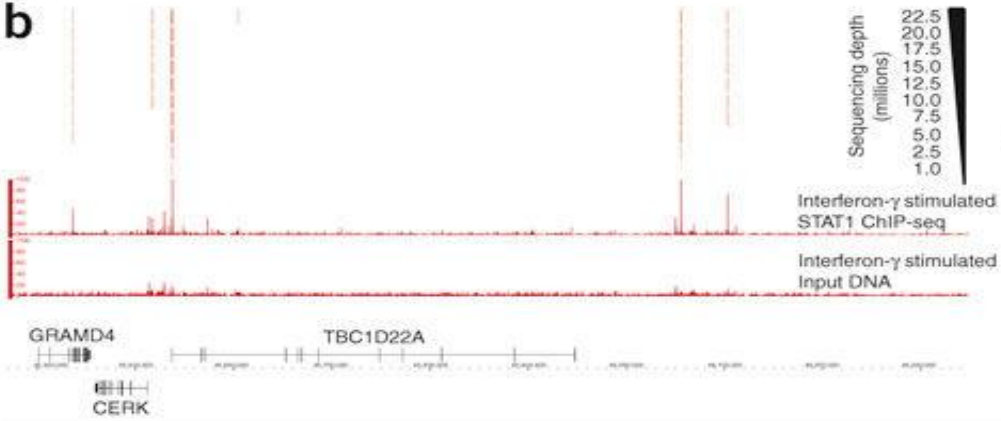
- Open chromatin regions are more easily fragmented than closed regions
- Uneven read distribution
- Repetitive sequences may appear to be enriched
- Compare ChIP-seq peak with same region in Input control

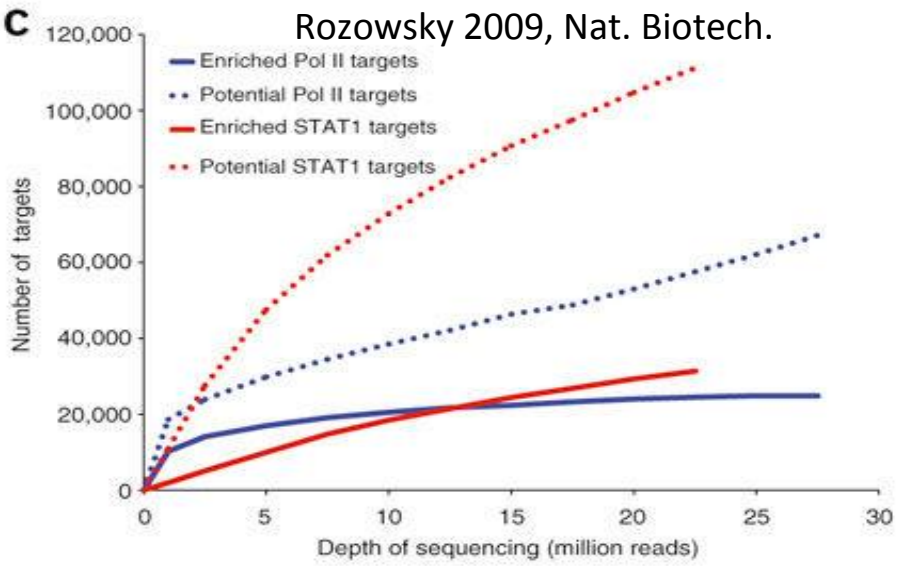




Rozowsky 2009, Nat. Biotech.







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: 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

Artefact removal 2: 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 3

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

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. The *DAC Blacklisted Regions* track was generated for the ENCODE project.

- 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

Artefact removal 4: 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 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.

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.