Quality control in ChIP-seq data

Dóra Bihary

MRC Cancer Unit, University of Cambridge

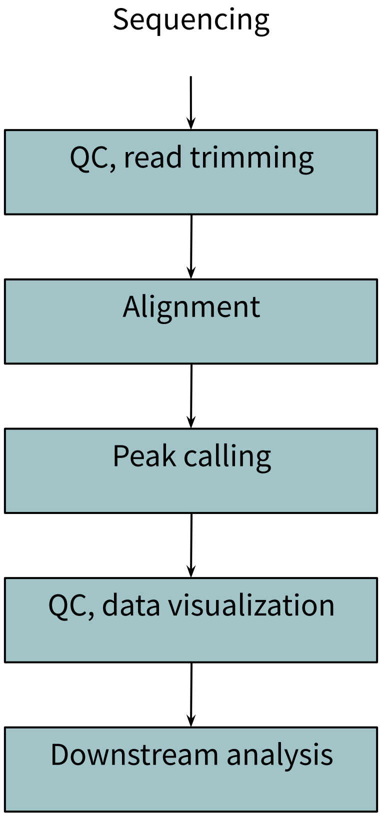
CRUK CI Bioinformatics Summer School

July 2019

Overview

- Introduction
- Distribution of signal
  - Coverage profiles
  - FRiP: fraction of reads in peaks
  - REGI: relative enrichment in genomic intervals
  - FRiBl: fraction of reads in blacklisted regions
- Clustering of Watson/Crick reads
- Other factors affecting site discovery
  - Sequencing depth
  - Duplication rate, library complexity
  - Controls

Workflow of ChIP-seq data processing



Sequencing

QC, read trimming

Alignment

Peak calling

QC, data visualization

Downstream analysis

FastQC, Cutadapt, Trimmomatic etc.

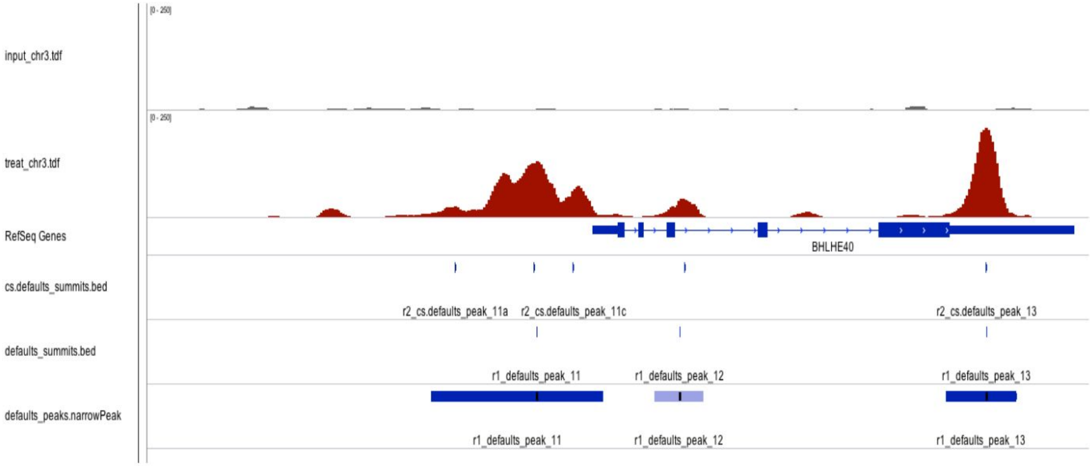
BWA, Bowtie

MACS2 etc.

ChIPQC, SPP, IGV

Looking at ChIP-seq data

- A good quality ChIP-seq experiment will have high enrichment over background
- Ways to quantify the quality:
  - Number of reads in peaks
  - High peaks, low background
  - Sequencing depth
  - Diverse library (duplications)
  - Low enrichment in control
  - Similarity of replicates
  - Genes closeby



- Tools to quantify quality:
  - ChIPQC (T Carroll, *Front Genet*, 2014.)
  - SPP package - Unix/Linux (PV Karchenko, *Nature Biotechnol,* 2008.)
  - ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia (Landt et al, *Genome Research*, 2012.)

Things that can go wrong

- The specificity of the antibody
  - Poor reactivity against the target of the experiment
  - High cross-reactivity with other proteins
- Degree of enrichment
- Biases during library preparation
  - PCR amplification bias
  - Fragmentation bias

- These can all affect the quality of the data and the number of sites detected
- Identification and removal of technical noise from the data is important

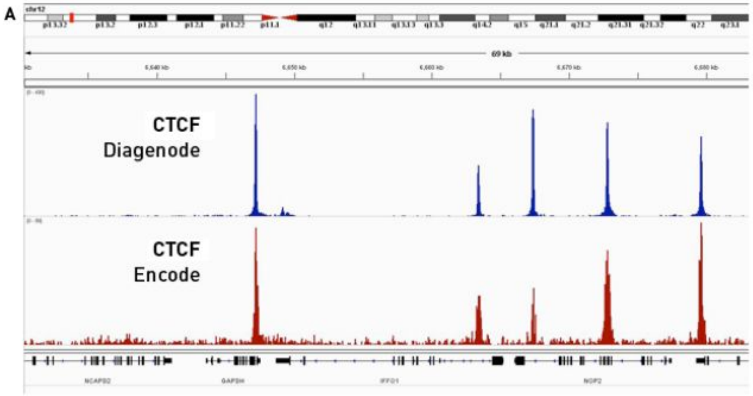
[1] Landt et al, *Genome Research*, 2012.

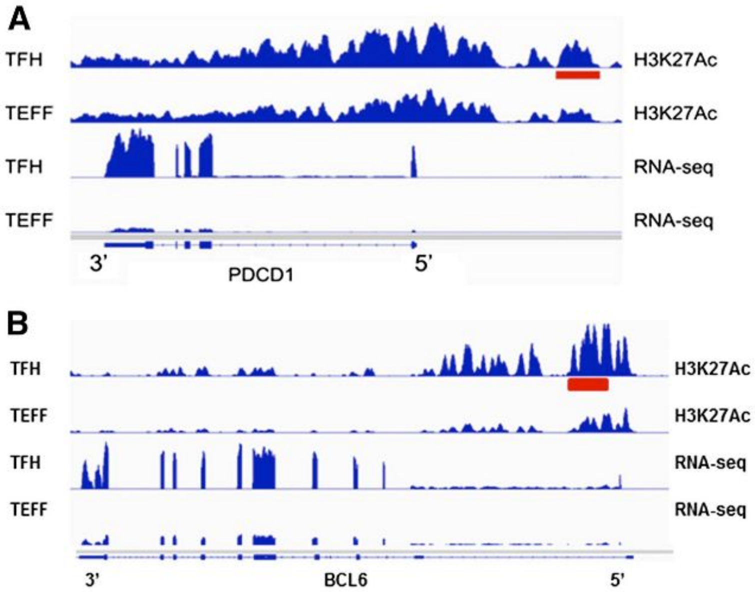
Overview

- Introduction
- Distribution of signal
  - Coverage profiles
  - FRiP: fraction of reads in peaks
  - REGI: relative enrichment in genomic intervals
  - FRiBl: fraction of reads in blacklisted regions
- Clustering of Watson/Crick reads
- Other factors affecting site discovery
  - Sequencing depth
  - Duplication rate, library complexity
  - Controls

Visualisation of coverage profiles

- Using IGV or USCS genome browser

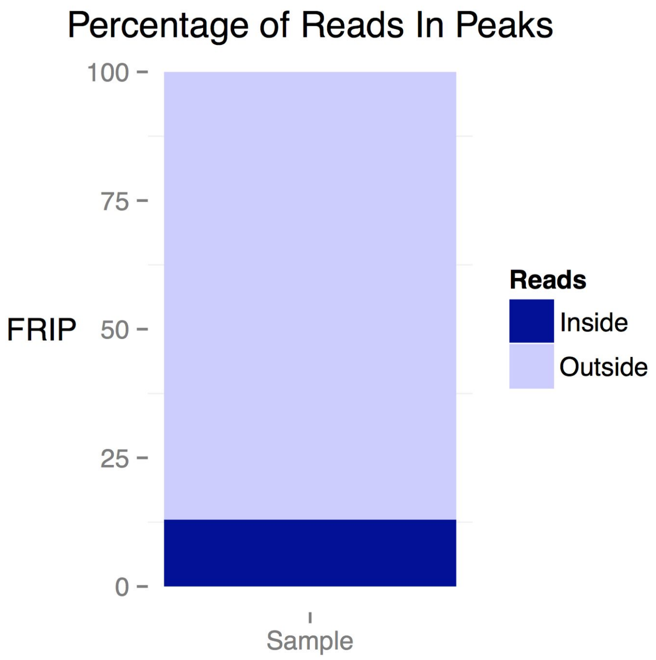




[1] Weinstein et al, *Blood*, 2014.

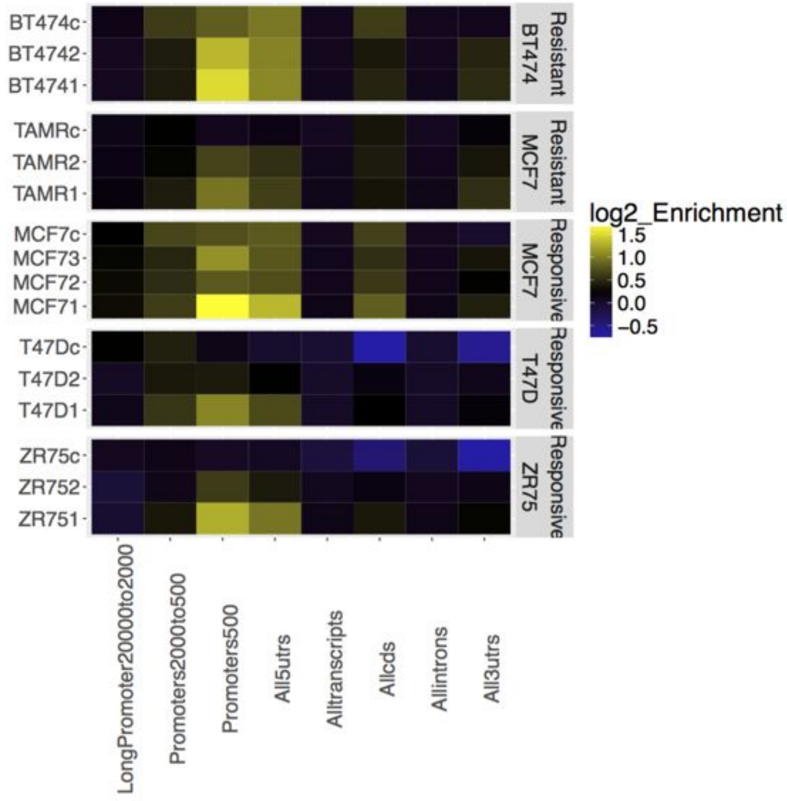
FRiP - fragment of reads in peaks

- A useful metric to measure global ChIP enrichment
- Gives a quick understanding of the success of immunoprecipitation
- Guideline: in case of good quality FRiP is > 5%
  - But there are known examples of good quality data with FRiP < 1%



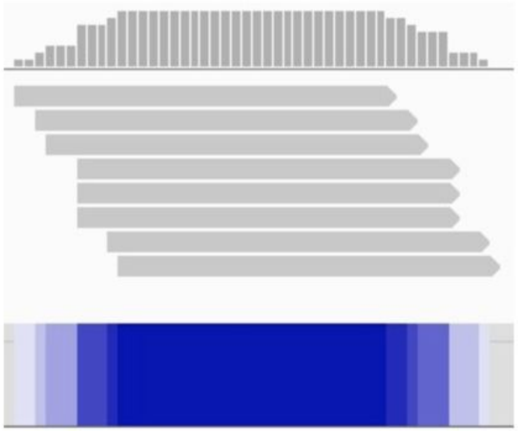
REGI - relative enrichment in genomic intervals

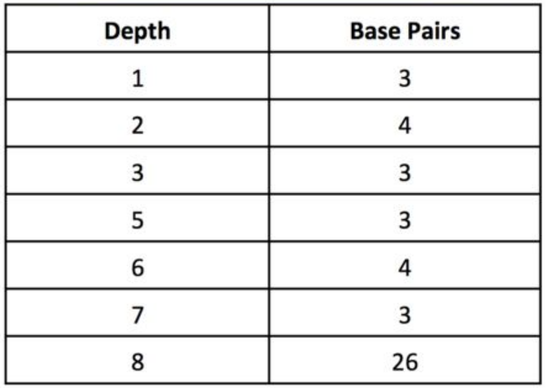
- Proteins might have a high expected enrichment in certain genomic regions, like promoters, UTRs, introns, etc.
- This plot helps to identify whether our experiment worked as expected and/or to reveal interesting behaviour



Dispersion of coverage

- The depth of coverage is the number of fragments at a specific genomic region
- To build a coverage profile
  - Measure the number of base pairs with a given depth of coverage
  - Normalise to the number of reads to compare samples
- We expect the depth to have large diversity in an enriched ChIP dataset

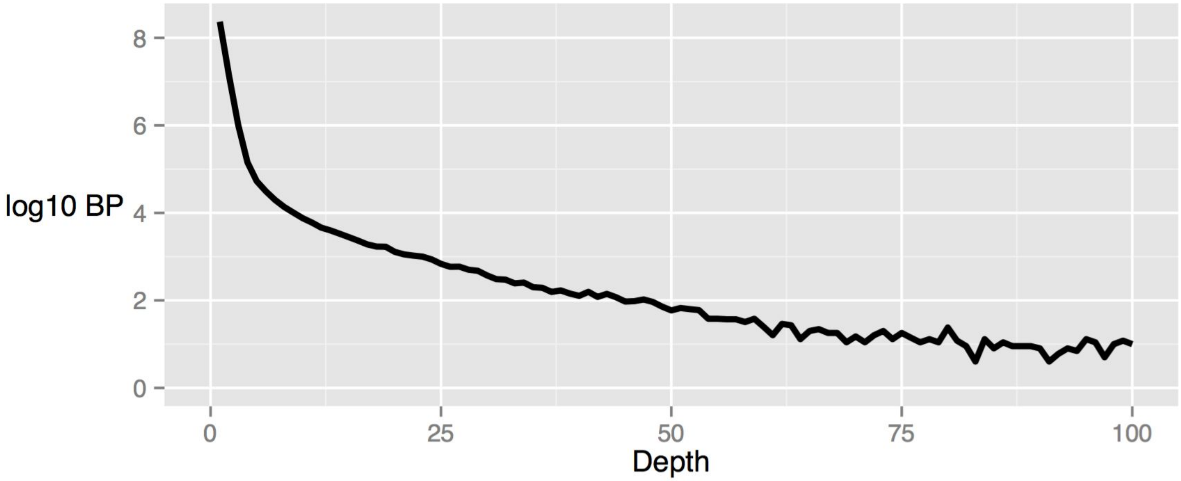




Depth	Base Pairs
1	3
2	4
3	3
5	3
6	4
7	3
8	26

Dispersion of coverage

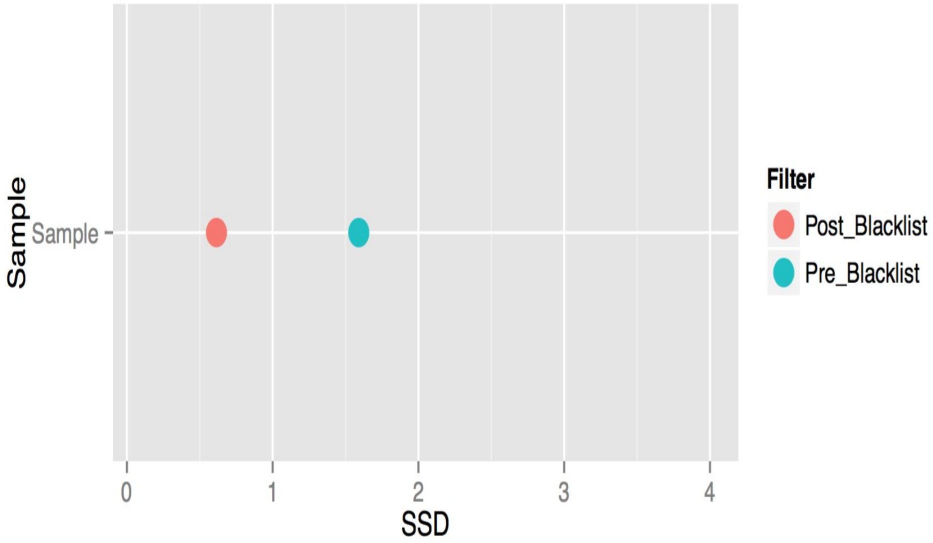
- Dispersion coverage profile plotted with ChIPQC
- More enriched libraries have higher number of bases at greater depths
- Profile of control samples usually drops more quickly
- The gap between samples and controls indicates enrichment



Dispersion of coverage

- **SSD**: standardised standard deviation
- Metric to assess dispersion coverage developed in htSeqTools package
- Provides measure of pile-up across the genome, it is expected to be:
  - High for samples with enriched regions
  - Low for controls with uniform coverage
- This measure is highly influenced by regions, where the coverage is high because of some mapping error, like blacklisted regions

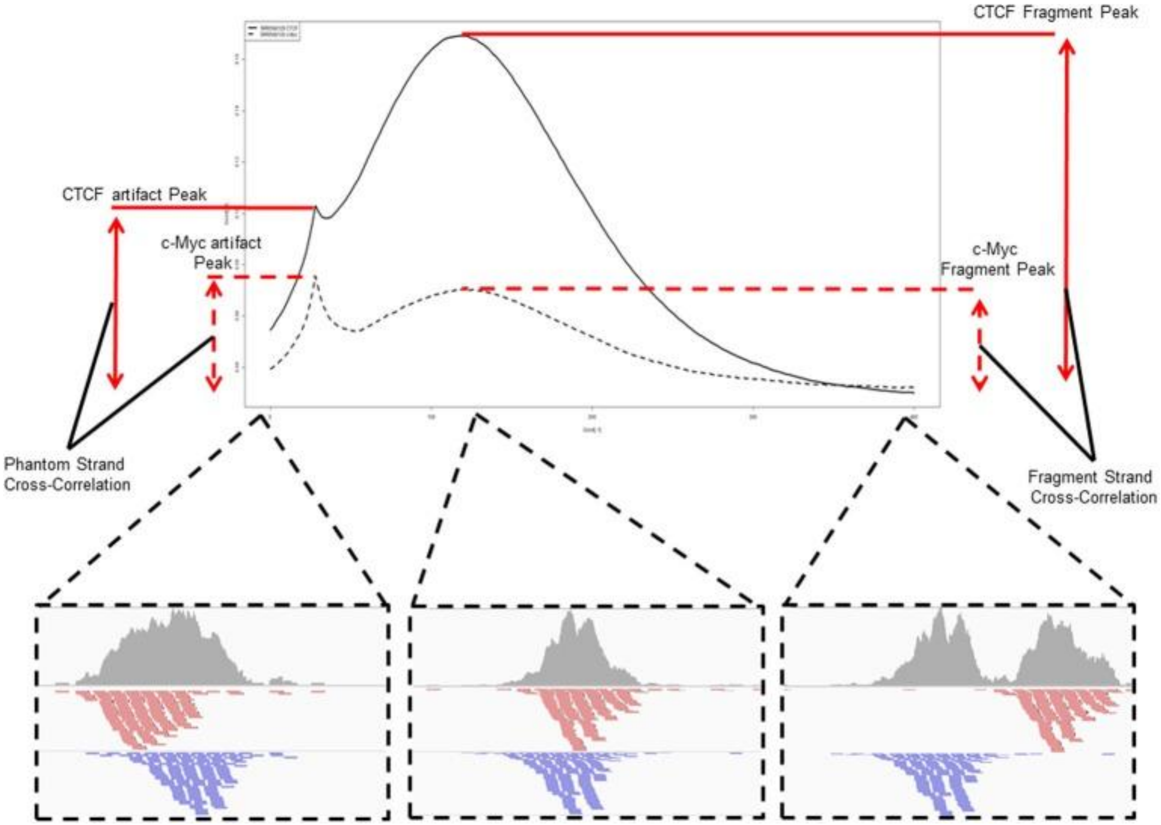
$$SSD = \frac{SD}{\sqrt{n}}$$



Overview

- Introduction
- Distribution of signal
  - Coverage profiles
  - FRiP: fraction of reads in peaks
  - REGI: relative enrichment in genomic intervals
  - FRiBl: fraction of reads in blacklisted regions
- Clustering of Watson/Crick reads
- Other factors affecting site discovery
  - Sequencing depth
  - Duplication rate, library complexity
  - Controls

Clustering of Watson/Crick reads



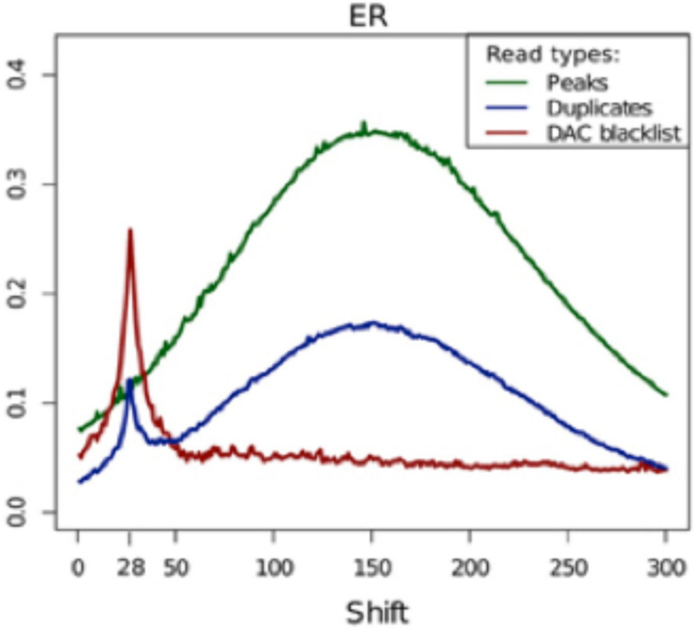
Clustering of Watson/Crick reads

- Fragment length can be estimated from the data
  - Cross-correlation: correlation of reads on positive and negative strand after successive read shifts
  - Cross-coverage: coverage of reads on both strands after successive shifts of reads on one strand; the area covered by reads will be reduced after the shifting
- These metrics are computed in ChIPQC:

$$FragCC = CC_{fragmentLength}$$

$$RelCC = \frac{FragCC}{CC_{readLength}}$$

- Blacklisted regions have a large contribution to read-length cross-coverage peaks

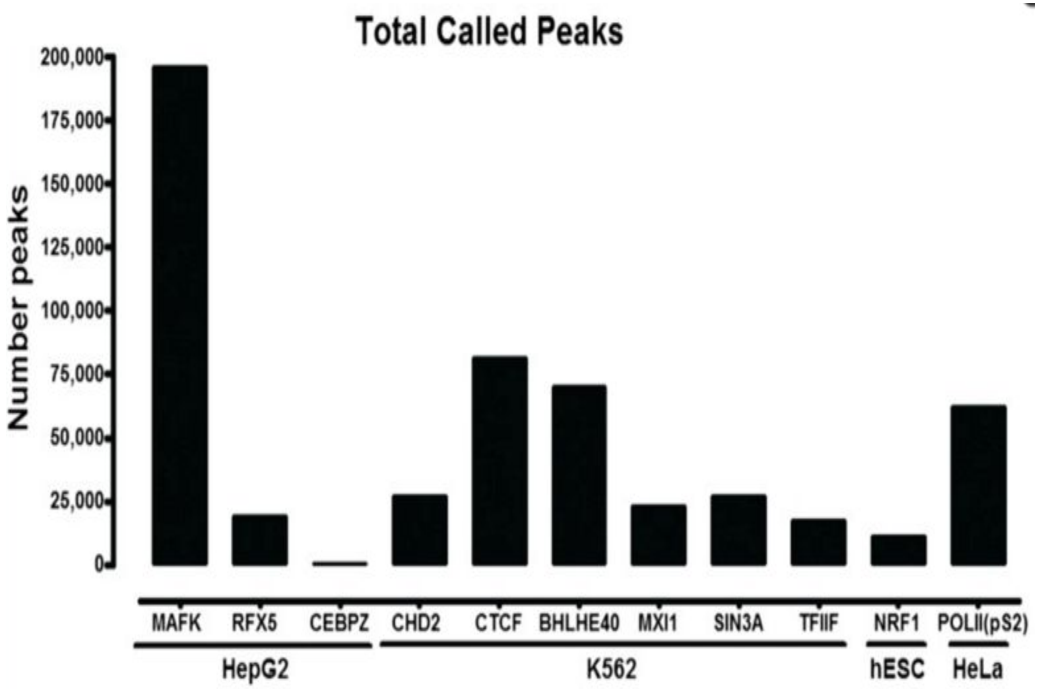


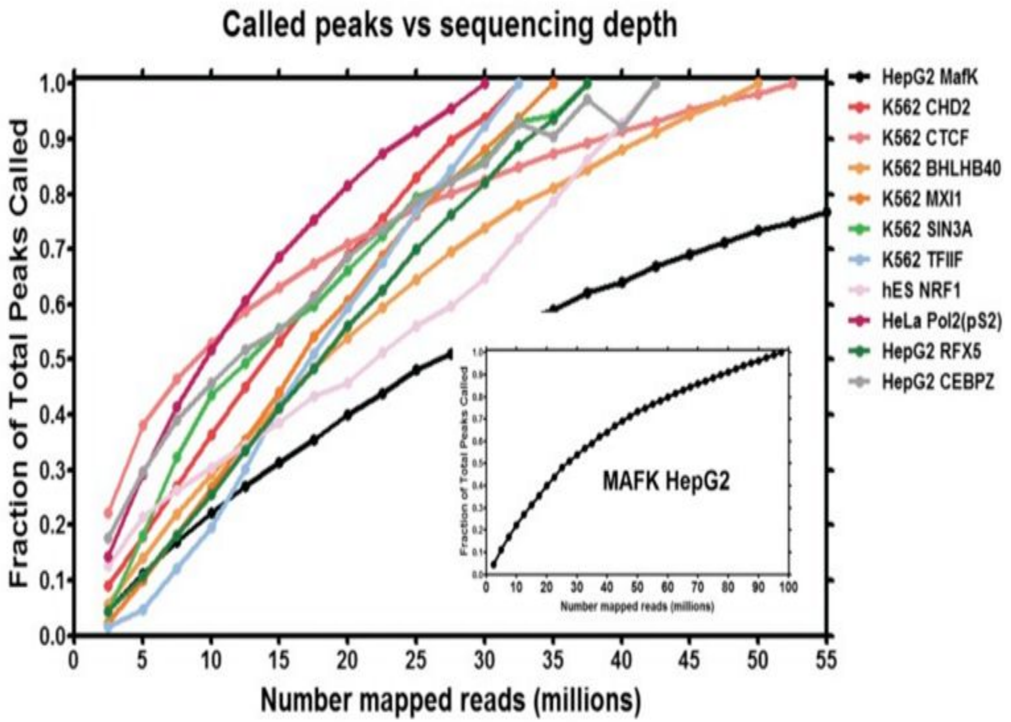
Overview

- Introduction
- Distribution of signal
  - Coverage profiles
  - FRiP: fraction of reads in peaks
  - REGI: relative enrichment in genomic intervals
  - FRiBl: fraction of reads in blacklisted regions
- Clustering of Watson/Crick reads
- Other factors affecting site discovery
  - Sequencing depth
  - Duplication rate, library complexity
  - Controls

Sequencing depth

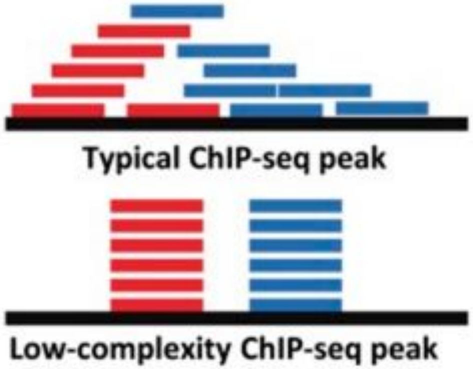
- The number of peaks depends on the depth of sequencing
- Some ENCODE guidelines:
  - Sharp peaks (like transcription factors):
    - Mammalian: 10M reads
    - Worms and flies: 2M reads
  - Broad peaks (some histone marks):
    - Mammalian: 20M reads
    - Worms and flies: 5M reads





[1] Landt et al, *Genome Research*, 2012.

Duplication rate, library complexity

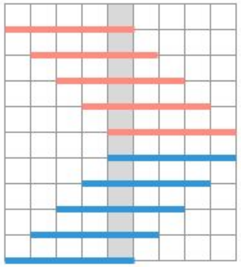


- Duplication rate is also a QC metric:
  - Expected to be low (<1%) for inputs

$$\frac{DuplicateReads}{TotalMappedReads} \times 100$$

- Duplicates can be artefacts:
  - PCR bias: certain genomic regions are preferentially amplified
  - Low initial starting material can introduce artificially enriched regions with overamplification
- Duplicates can also be "legitimate":
  - It is unavoidable in highly enriched experiments and deeply sequenced ChIPs since it is naturally increasing with the sequencing depth
- Removing duplicates limits the dynamic range of ChIP signal:
  - Maximum signal/base: one fragment on each strand in each possible position of the read

$Signal_{max} = 2*readLength$



[1] Landt et al, *Genome Research*, 2012.

Duplication rate, library complexity

- What to do with duplicates?
- Always keep in mind enrichment efficiency and read depth
- Some approaches:
    - Remove all duplicates
    - Don’t remove duplicates as long as it has a reasonable rate
    - Remove duplicates for some analysis:
        - Remove duplicates before peak-calling
        - Keep duplicates for differential binding analysis
    - htSeqTools:
        - Estimate duplicate numbers expected taking into account the sequencing depth and using negative binomial model
        - Attempt to identify significantly outstanding duplicate numbers

Control/input samples

- The use of some kind of a control is always recommended
- You need different controls for:
  - Different cell lines, cell types
  - Different organisms
  - Different treatments/conditions
- Types of controls:
  - Input DNA:
    - Most popularly used
    - Controls for CNVs, sequencing-, fragmentation- and shearing biases
  - IgG:
    - Also controls for non-specific binding
    - Introduces other biases

Acknowledgement

- Ines de Santiago

  https://github.com/bioinformatics-core-shared-training/ngs-in-bioc/blob/master/Day3/Lect7.ChIP_QC_presentation.pdf

- Tom Carroll

  http://bioconductor.org/help/course-materials/2014/BioC2014/ChIPQC_Presentation.pdf

  https://github.com/bioinformatics-core-shared-training/ngs-in-bioc/blob/master/Lectures/Lect6b_ChIP--Seq%20Data%20Analysis.pdf