Best practices in the analysis of RNA-seq and ChIP-seq data
27th – 31st, July 2015
University of Cambridge, Cambridge, UK

The quality of a ChIP-seq data

Ines de Santiago
CRUK Cambridge Research Institute
Ines.desantiago@cruk.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

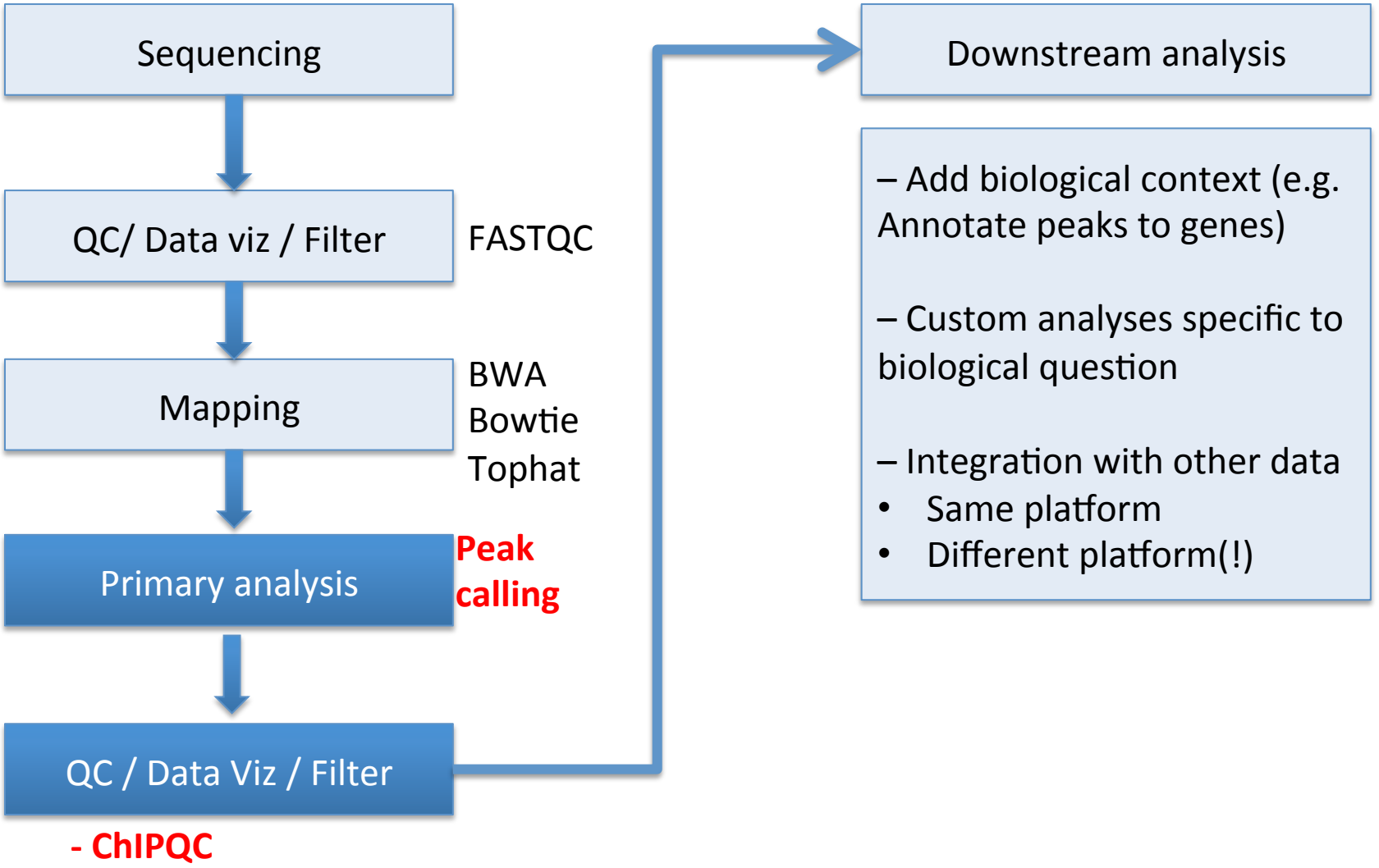
CAMBRIDGE Cancer CENTRE

CANCER RESEARCH UK

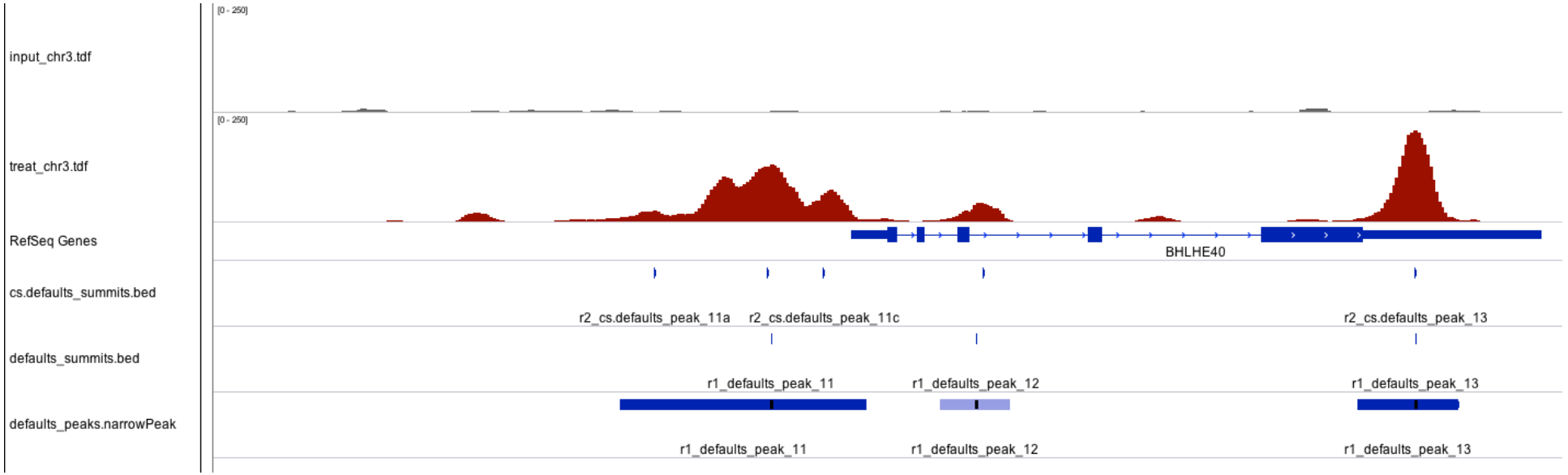
Acknowledgments

- 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
- Shamith Samarajiwa
- Suraj Menon

“Typical” ChIP-Seq Analysis workflow



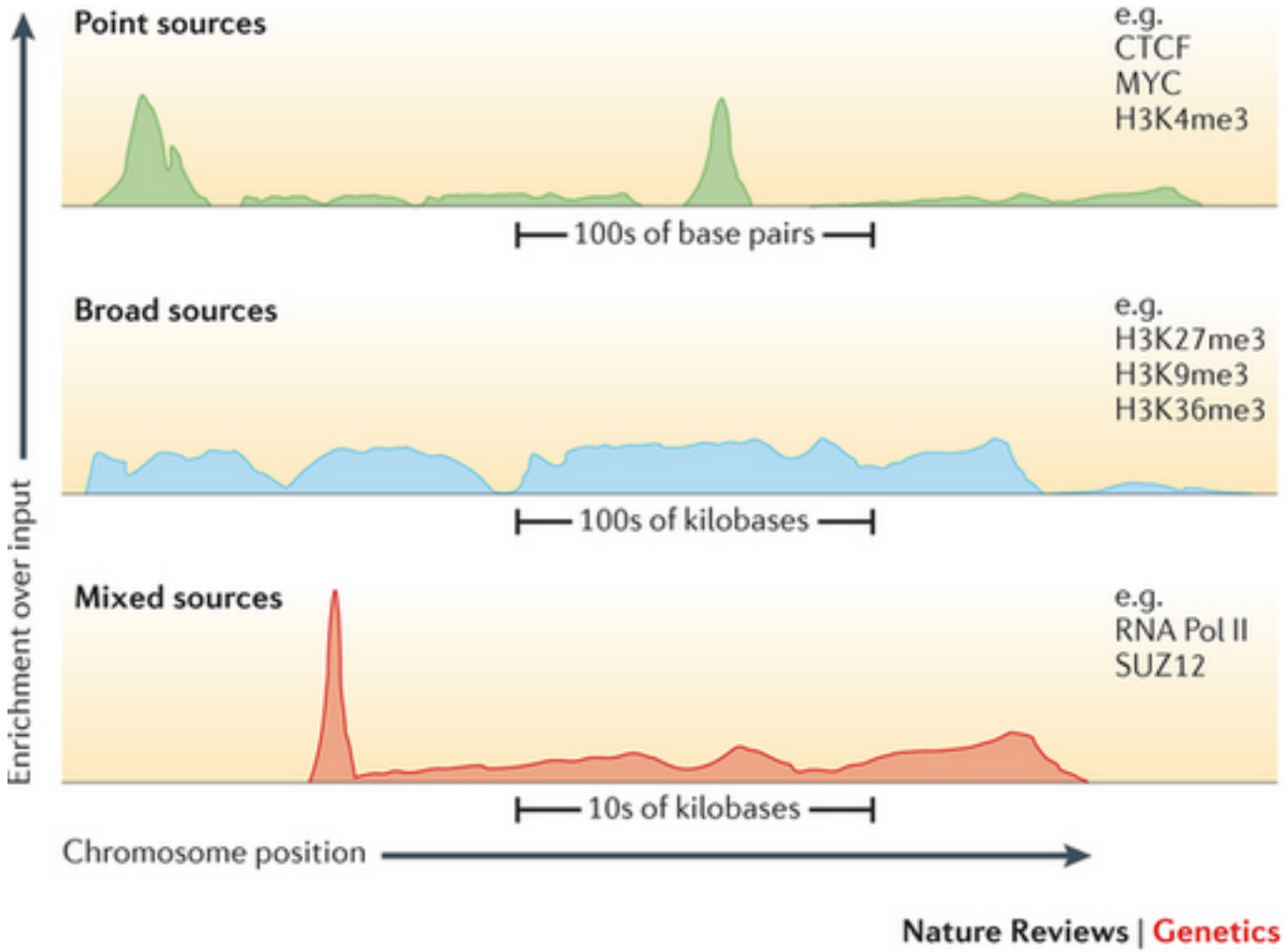
A good ChIP-seq dataset



What do we want:

- Good quality ChIP-seq enrichment over background

Narrow vs Broad peaks



http://www.nature.com/nrg/journal/v15/n2/full/nrg3642.html

What can go wrong?

- **The specificity of the antibody**
  - poor reactivity against the intended target
  - cross-reactivity with other DNA-associated proteins.

- **degree of enrichment** achieved in the affinity precipitation step.

- **Biases in the selected fragments**:
  - PCR amplification biases
  - Fragmentation biases

ENCODE guidelines for experiments: https://genome.ucsc.edu/ENCODE/experiment_guidelines.html

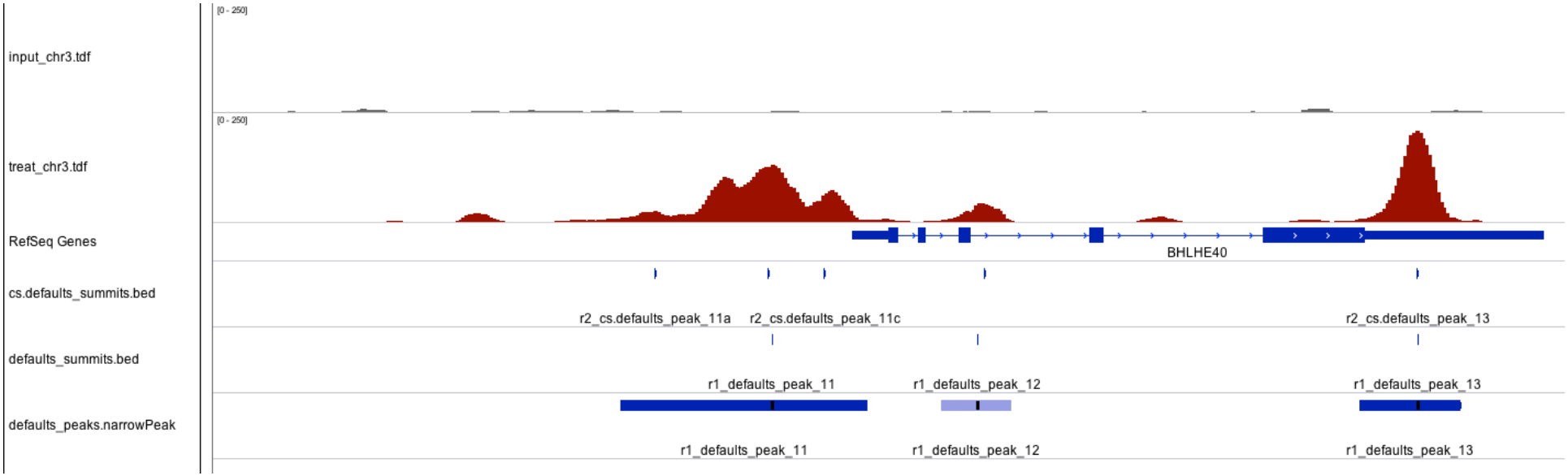
How to quantify ChIP-seq data quality?

- **ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia.**

  (Landt et al – *Genome Research 2012*)

- **ChIPQC –** Tom Carroll and Rory Stark *(Diffbind)*
- **ChIPQC** provides workflow to generate metrics per sample/experiment.
- package **SPP** (for UNIX/LINUX)

EVALUATING CHIP-SEQ DATA (QC)



What do we want:

- Good quality ChIP-seq enrichment over background

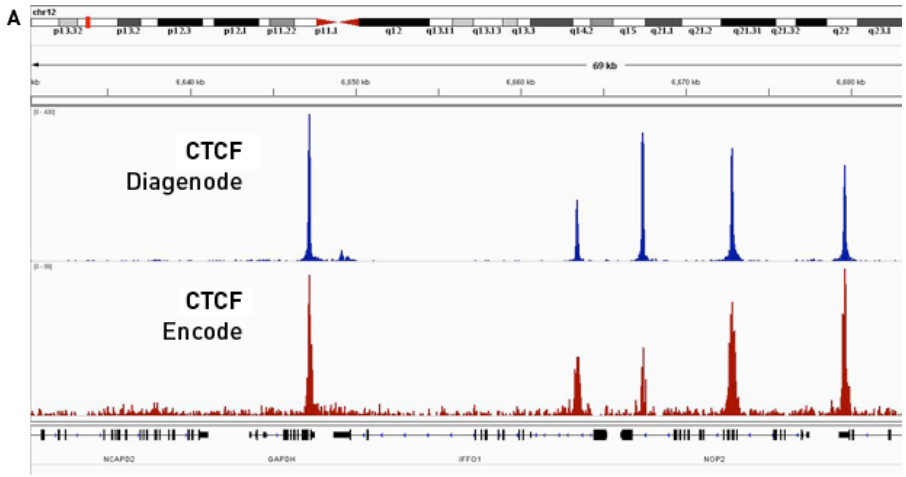
Outline

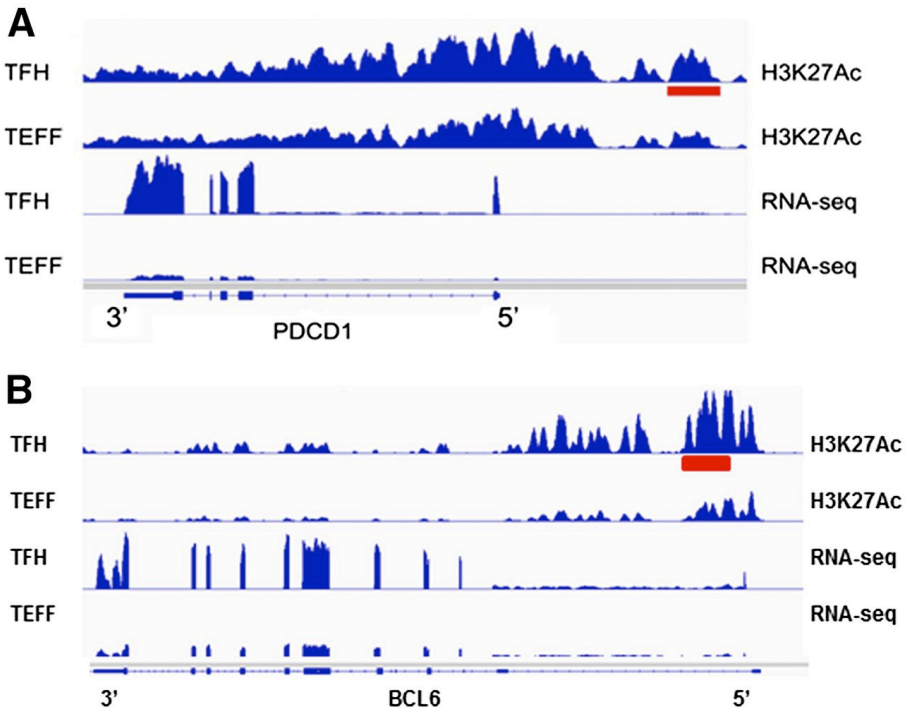
- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Fraction of reads in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads.**
- **Other factors affecting site discovery:**
  - Sequencing depth
  - Duplication rate / library complexity
  - Control sample

Outline

- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Fraction of reads in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads.**
- **Other factors affecting site discovery:**
  - Sequencing depth
  - Duplication rate / library complexity
  - Control sample

Visualisation of coverage profiles





http://www.bloodjournal.org/content/124/25/3719

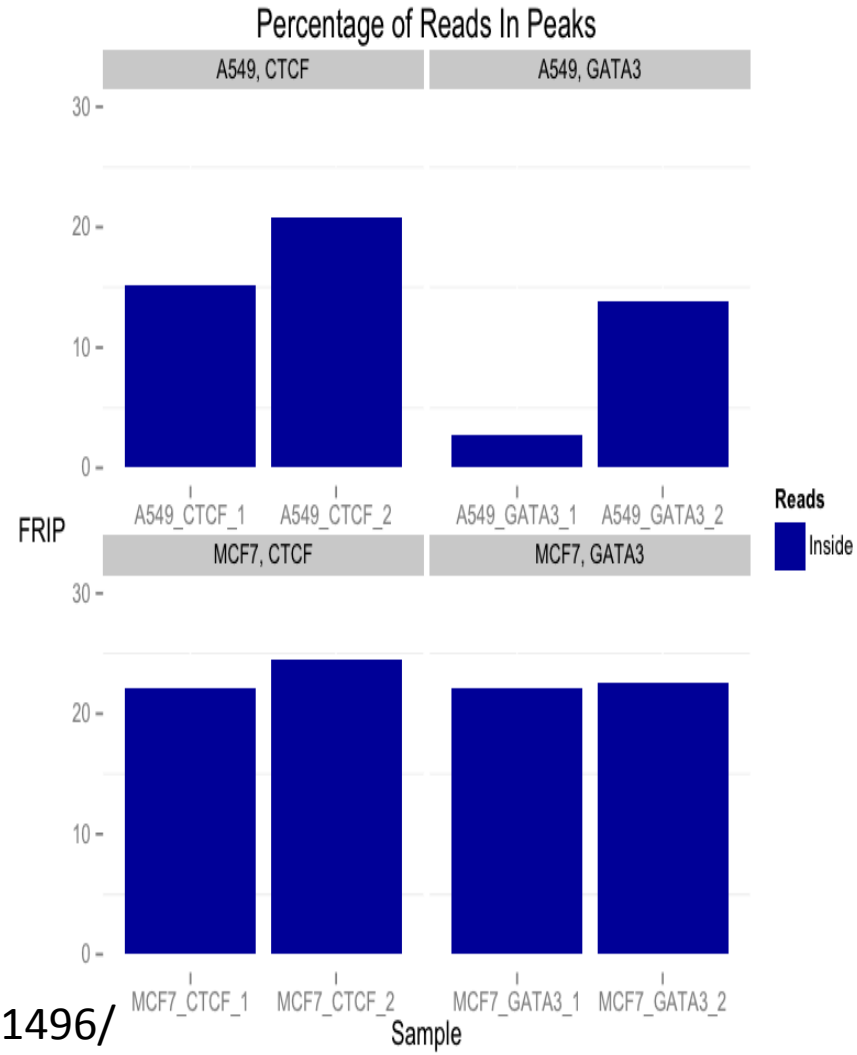
Outline

- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Fraction of reads in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads.**
- **Other factors affecting site discovery:**
  - Sequencing depth
  - Duplication rate / library complexity
  - Control sample

Measuring global ChIP enrichment (FRiP)

- useful and simple first-cut metric for the success of the immunoprecipitation
- Good quality TF > 5%

Example output from ChIPQ package:



http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431496/

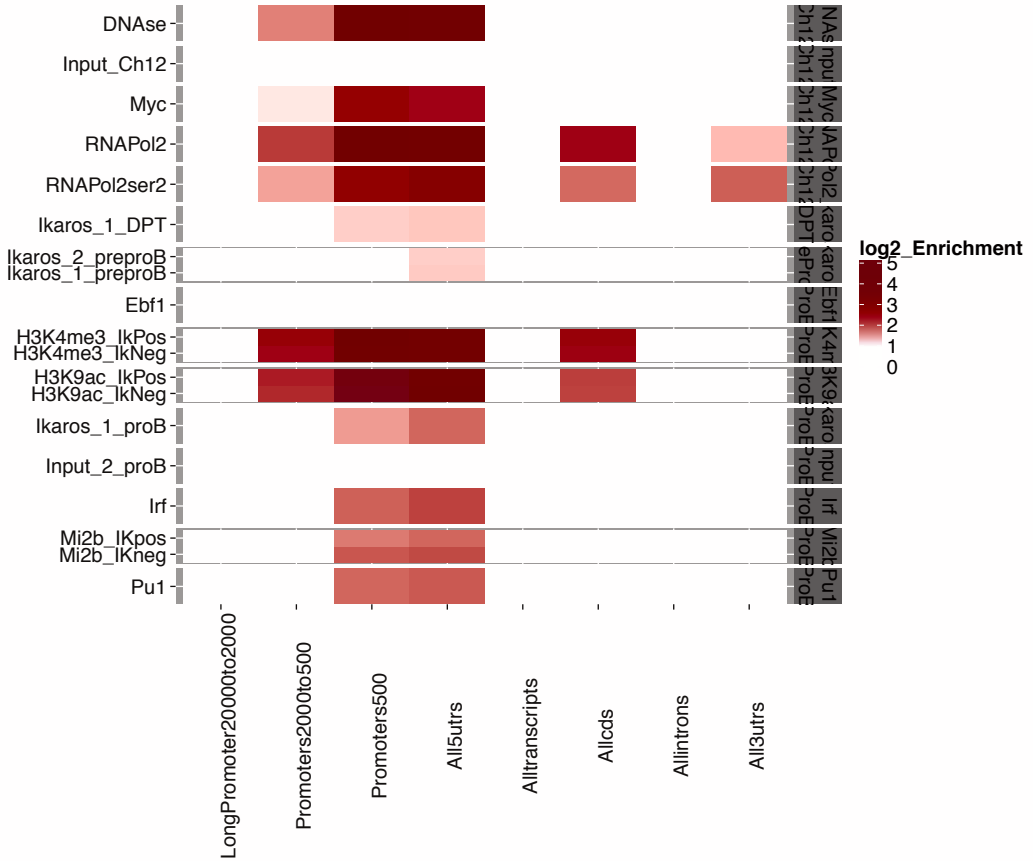
Outline

- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Fraction of reads in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads.**
- **Other factors affecting site discovery:**
  - Sequencing depth
  - Duplication rate / library complexity
  - Control sample

Enrichment in genomic intervals

- Plot relative enrichment of reads in annotated regions.

Example output from ChIPQ package:



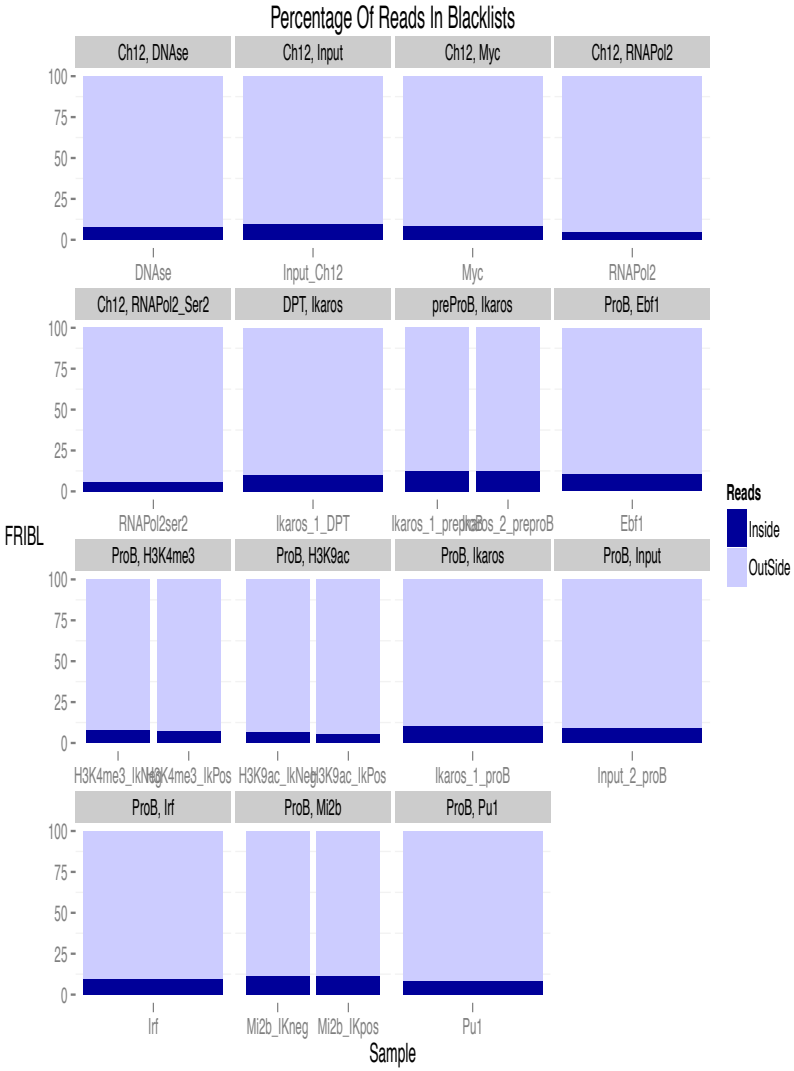
Outline

- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Fraction of reads in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads.**
- **Other factors affecting site discovery:**
  - Sequencing depth
  - Duplication rate / library complexity
  - Control sample

Signal in blacklists

- Encode empirically identified regions that showed anomalous and near-universal artefact signal
- Represent around 0.5% of genome.
  - Various reasons e.g. chromatin accessibility, repeats
- Can account for high proportion of total signal (> 10%)
  - Enriched for duplicate and multi-mapping reads
- Adversely affect fragment length calculations and in thus any analyses that require these e.g. peak calling

Example output from ChIPQ package:



Carroll et al., Front Genet. 2014 Apr 10;5:75.

Outline

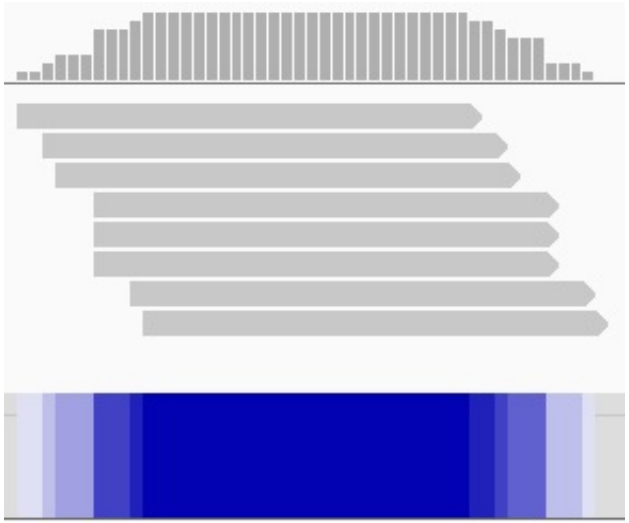
- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Fraction of reads in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads.**
- **Other factors affecting site discovery:**
  - Sequencing depth
  - Duplication rate / library complexity
  - Control sample

Dispersion of coverage

- Depth of signal: number of fragments at a genomic location.

- Expectation is that for an enriched ChIP sample, depth should show inequality in dispersion across the genome

- Build global profile of signal depth
  - Measure number of base pairs with given depth of signals.
  - Normalise to total number of reads to compare samples

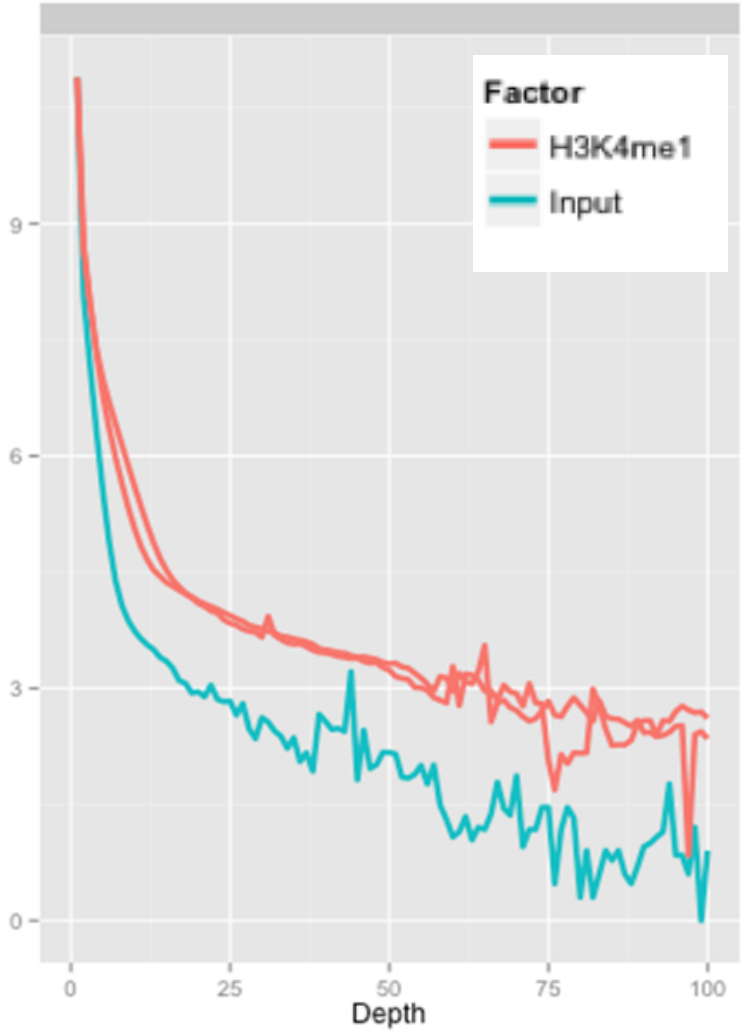


Depth	Base Pairs
1	3
2	4
3	3
5	3
6	4
7	3
8	26

Dispersion of coverage

- Global signal profile “histogram”
- Enriched (ChIP) libraries show higher number of bases at greater depths.
- Profile for inputs (no enrichment) drops off more quickly
- Gap between sample and input indicates enrichment

Example output from ChIPQ package:



Metric for dispersion of coverage: SSD

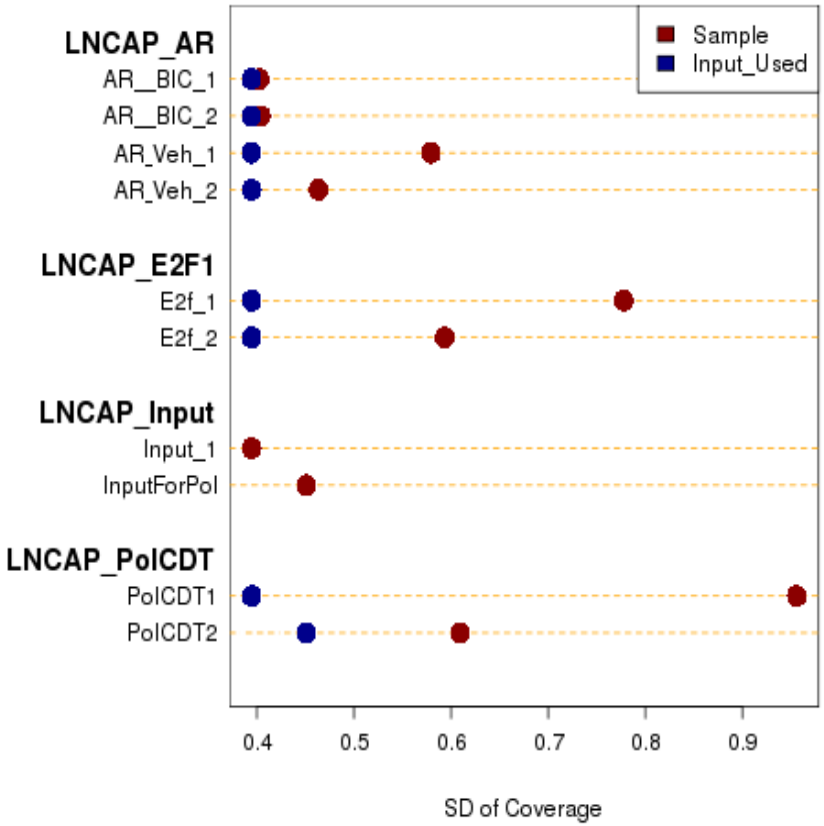
- SSD: Standardised Standard Deviation of coverage

- Metric for assessment of dispersion coverage developed in htseqtools package

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

- Provides measure of pile-up across genome
  - High for samples with enriched regions (ChIP)
  - Low for samples with uniform coverage (input)

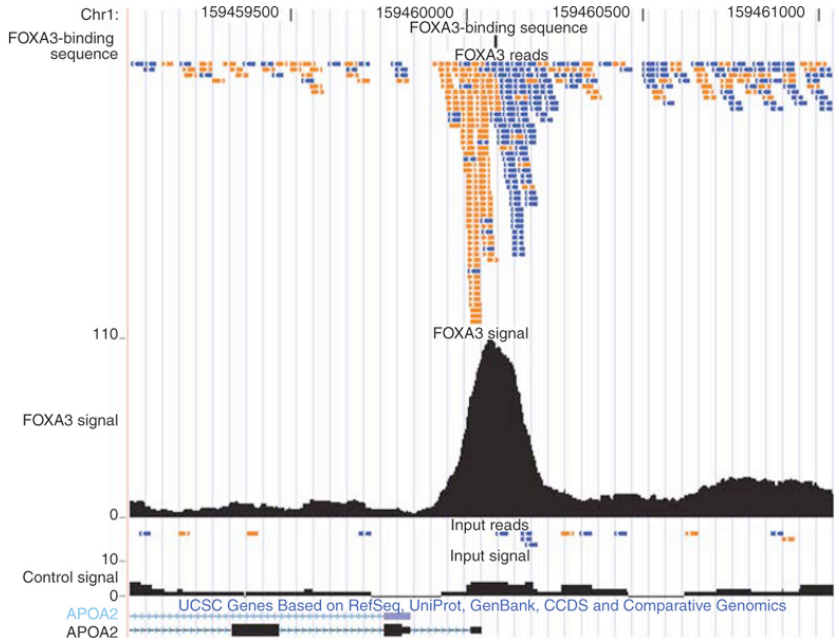
Example output from ChIPQ package:



Outline

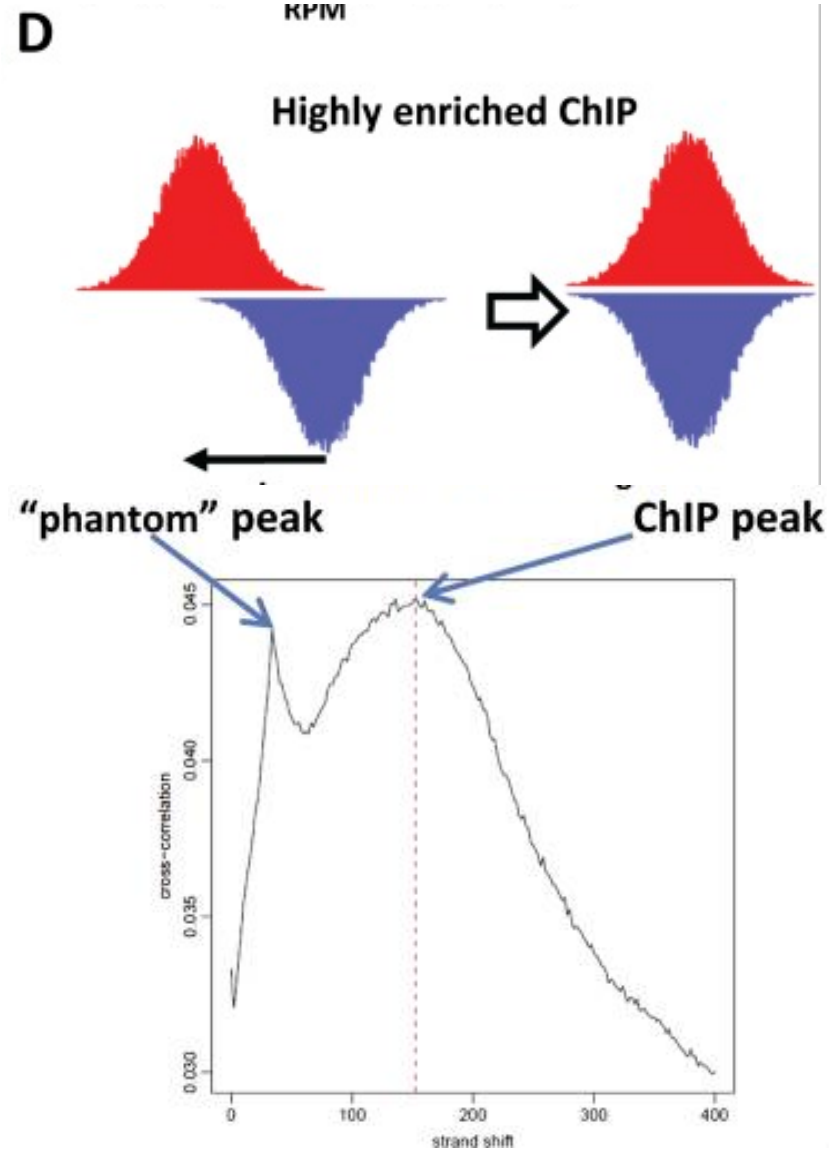
- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Fraction of reads in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads.**
- **Other factors affecting site discovery:**
  - Sequencing depth
  - Duplication rate / library complexity
  - Control sample

Clustering of Watson/Crick reads



http://www.nature.com/nmeth/journal/v6/n4/images/nmeth.f.247-F2.jpg
http://www.bloodjournal.org/content/124/25/3719

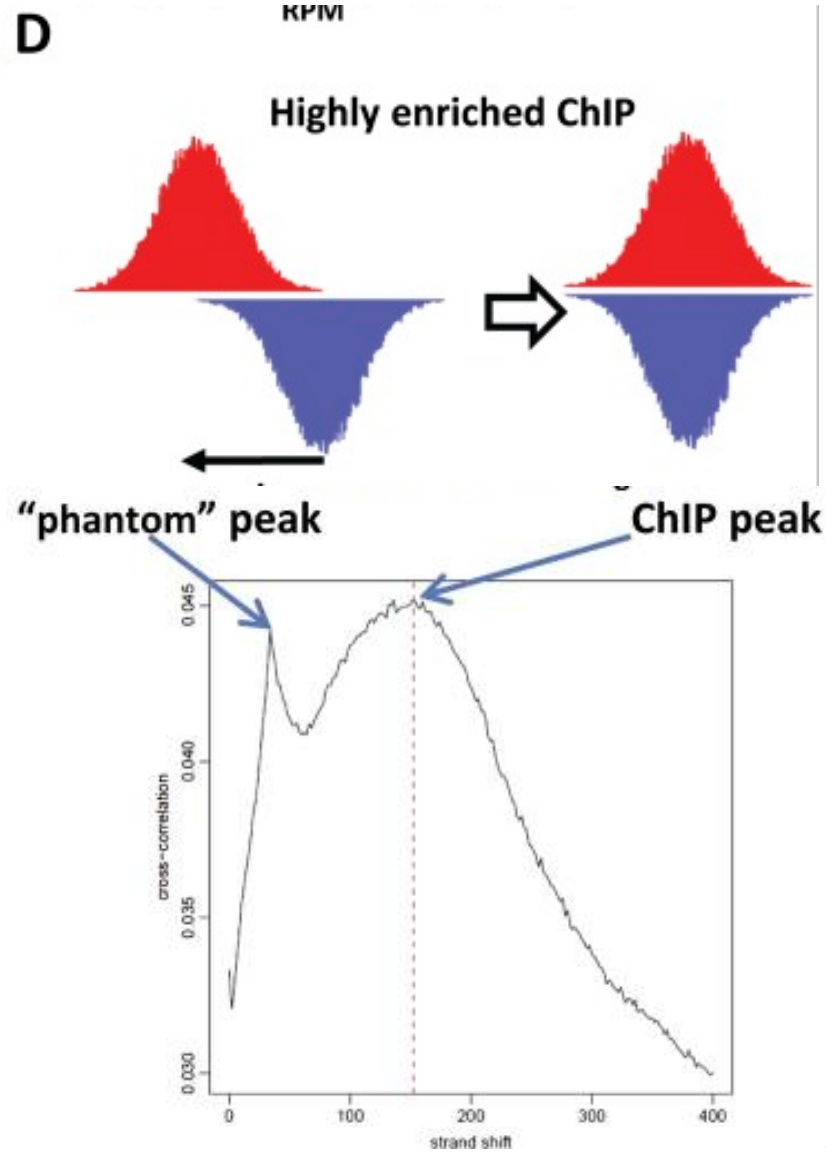
How to make a cross-correlation plot:



Clustering of Watson/Crick reads

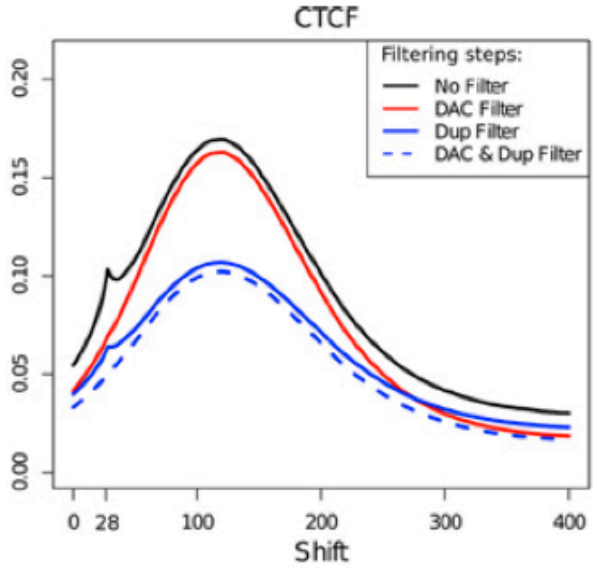
- Fragment length can be estimated from data:
  - **Cross-correlations -** Correlation of reads on positive and negative strand after successive read shifts
  - **Cross-coverage -** Coverage of reads on both strand after successive shifts of reads on one strand. Total area covered by reads will be reduced after shifting
- These provide useful QC metrics

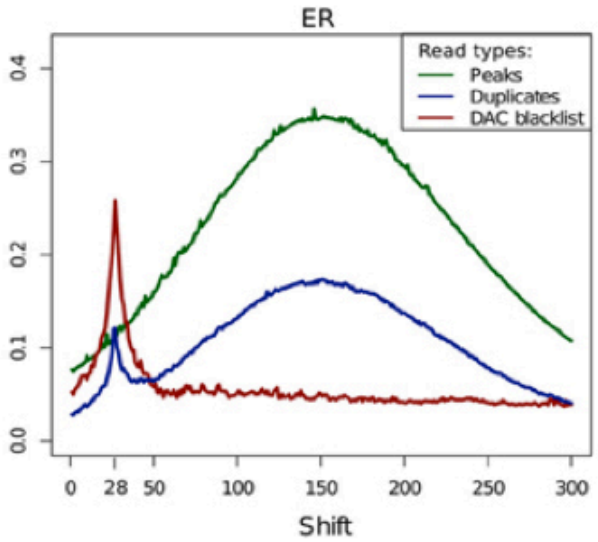
How to make a cross-correlation plot:



Clustering of Watson/Crick reads

- Cross-correlation/Cross-coverage score plots are useful for QC
- Blacklisted regions strongly contribute to read length cross-coverage peak
  - Small to non-existent peaks are seen in failed ChIPs and inputs
- ChIPQC metrics:
  - *FragCC = $CC_{fragmentlength}$*
  - *RelCC = FragCC/ $CC_{readlength}$*





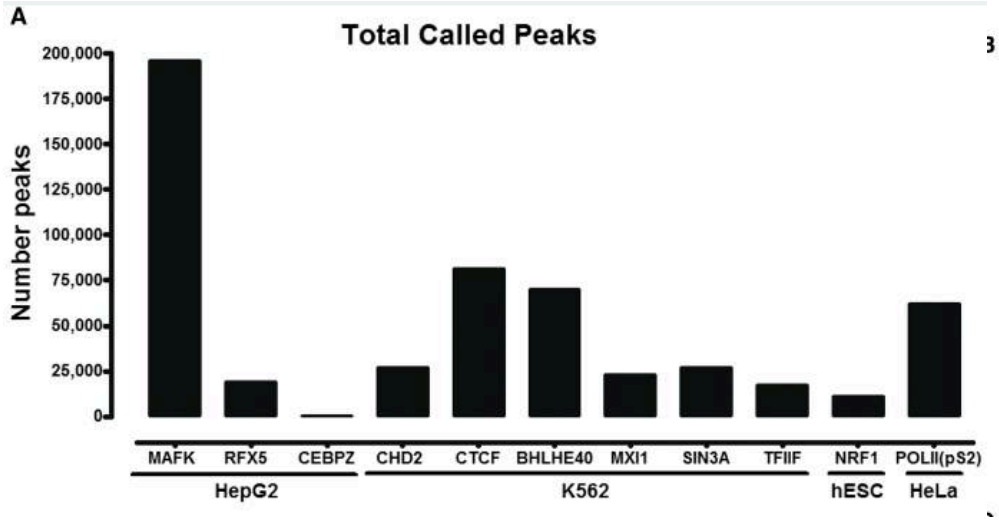
Carroll et al., Front Genet. 2014 Apr 10;5:75.

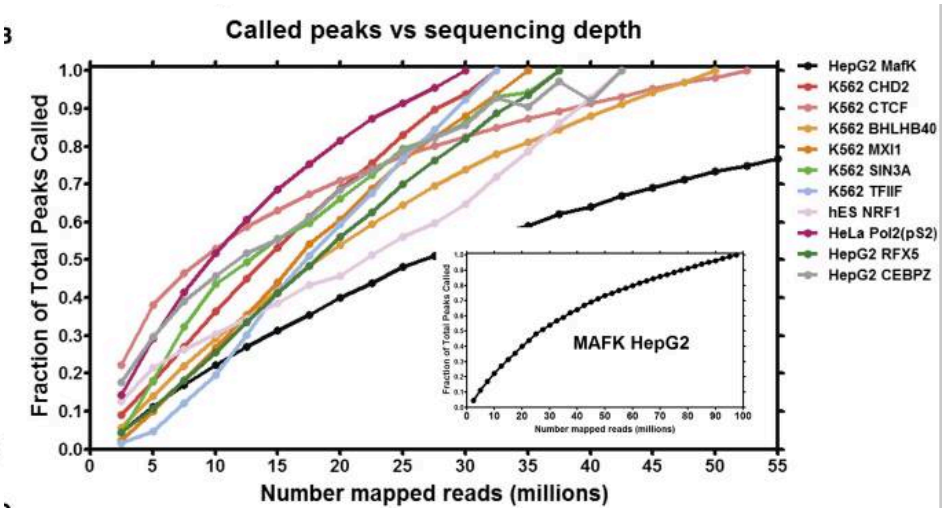
Outline

- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Fraction of reads in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads.**
- **Other factors affecting site discovery:**
  - Sequencing depth
  - Duplication rate / library complexity
  - Control sample

Sequencing Depth

Peak counts depend on sequencing depth.





http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431496/

Sequencing Depth: guidelines

Sharp peaks (TFs)

- 10M reads
- 2M worms and flies

Broad Peaks (Histones)

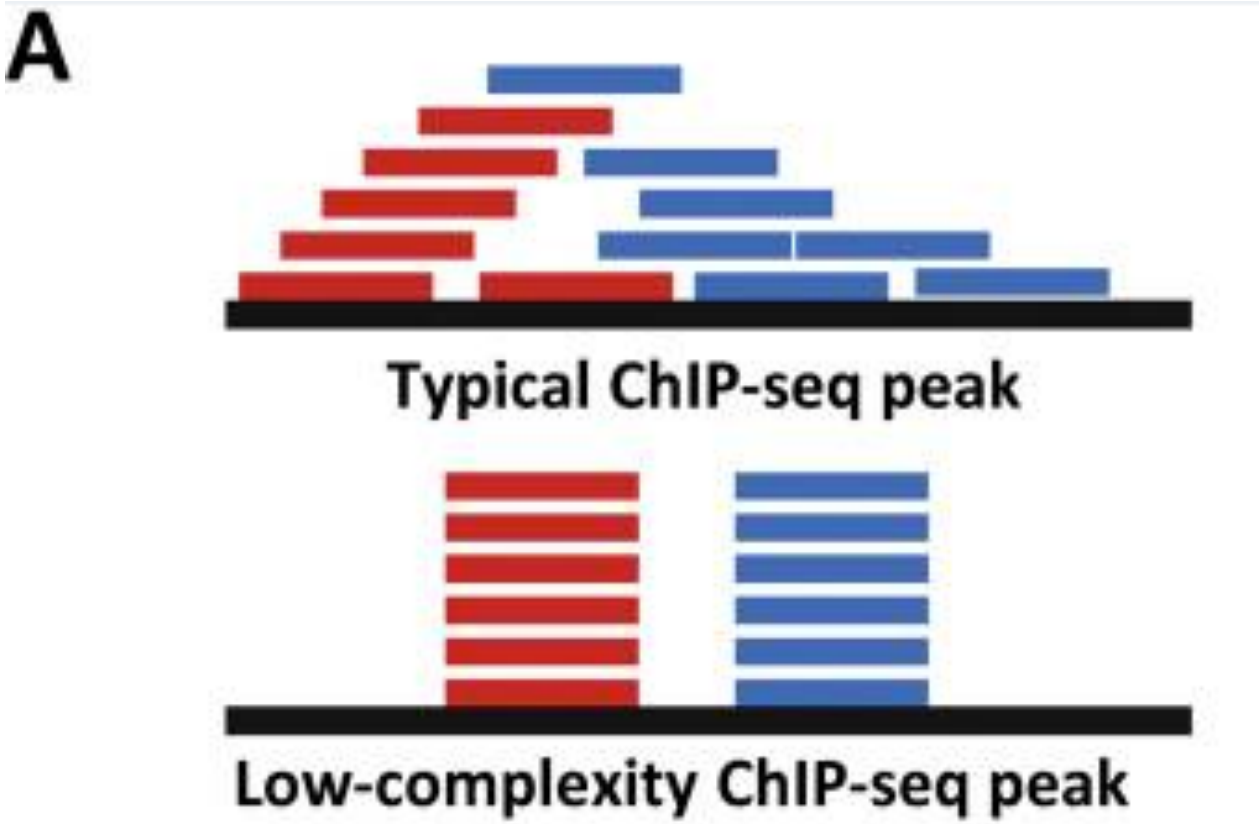
- 20M reads mammalian genomes
- 5M worms and flies

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431496/

Outline

- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Fraction of reads in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads.**
- **Other factors affecting site discovery:**
  - Sequencing depth
  - Duplication rate / library complexity
  - Control sample

Library complexity (Duplicates)



http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431496/

Library complexity (Duplicates)

- Duplicates can be artefacts
- PCR bias: certain genomic regions are preferentially amplified
- Low initial starting material
  - Overamplification -> artificially enriched regions
  - Compounded by PCR bias

- Duplicates can also be 'legitimate'
  - In highly efficient enrichments
  - In deeply sequenced ChIPs

  (Duplication rate increases with sequencing depth)

- Removing these duplicates limits the dynamic range of ChIP signal
  - Max signal for a base is (2*read length)-1

Library complexity (Duplicates)

- So what to do about duplicates?
- Keep in mind enrichment efficiency and read depth
- Thumb-rules
  - Remove duplicates prior to peak calling (some peak callers do this by default)
  - Keep duplicates for differential binding analysis
- A more objective approach:
  - htSeqTools package
  - Estimate duplicate numbers expected for sequencing depth using negative binomial model and attempt to identify significantly anomalous duplicate numbers.

Library complexity (Duplicates)

- Duplication rates are a useful QC metric
  - (Duplicate reads/Total Mapped Reads) *100
  - Expected to be low (<~ 1%) for inputs

- Non-Redundant Fraction (NRF)
  - Unique Reads/Total Mapped Reads
  - ENCODE guidelines:

    NRF >= 0.8 for 10M reads

Outline

- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Fraction of reads in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads.**
- **Other factors affecting site discovery:**
  - Sequencing depth
  - Duplication rate / library complexity
  - Control sample

Control sample

- Use of controls **highly** recommended
- **Input DNA**
  - popularly used
  - controls for CNVs, sequencing biases, fragmentation and shearing biases
- **IgG**
  - as with input but also controls for non-specific binding
  - but introduces new biases
- Controls required for
  - different types of samples (e.g. Cell lines, mice, patients)
  - different treatment groups / experimental conditions

Control sample

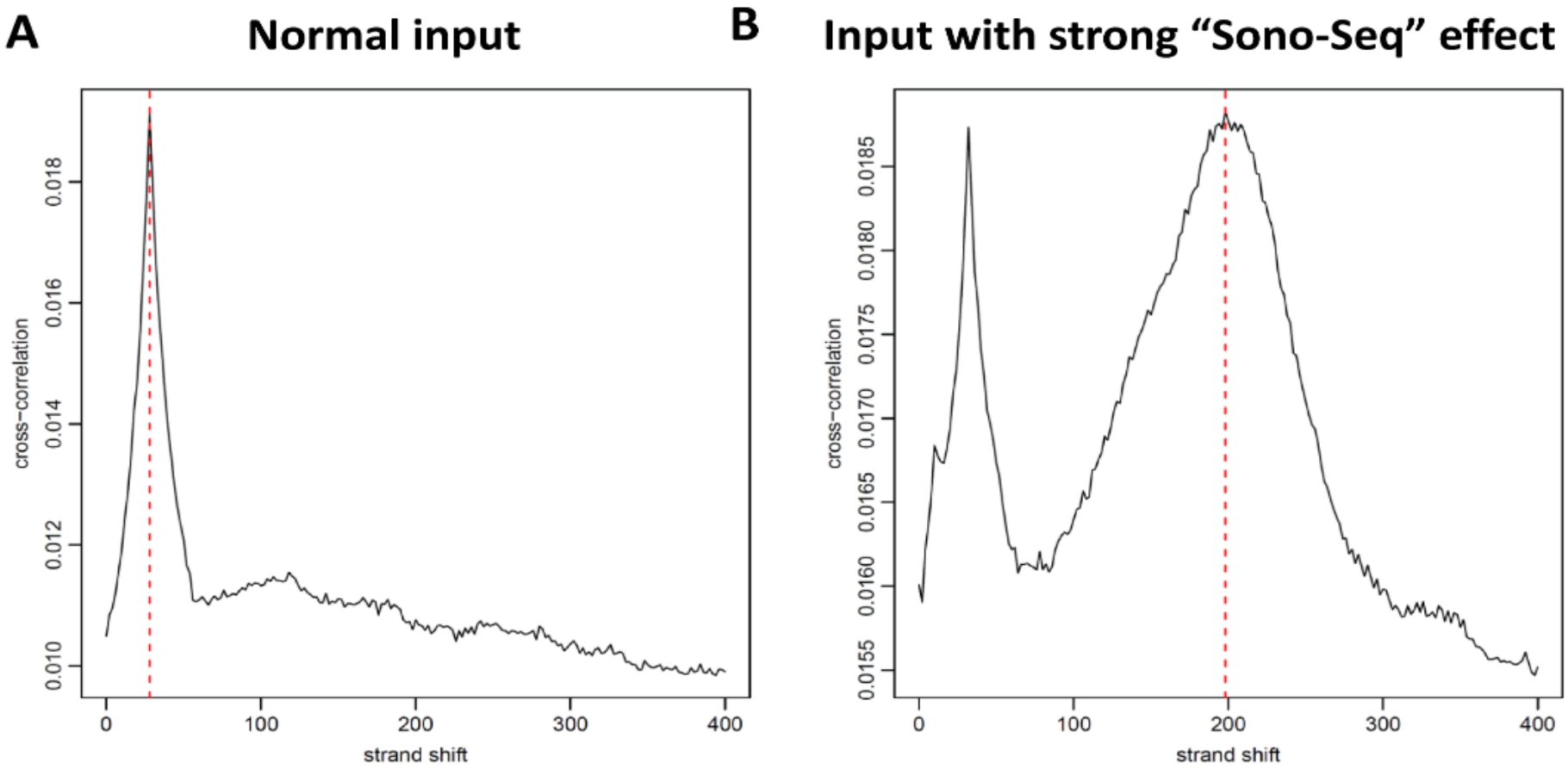


Fig. S1

http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3431496/

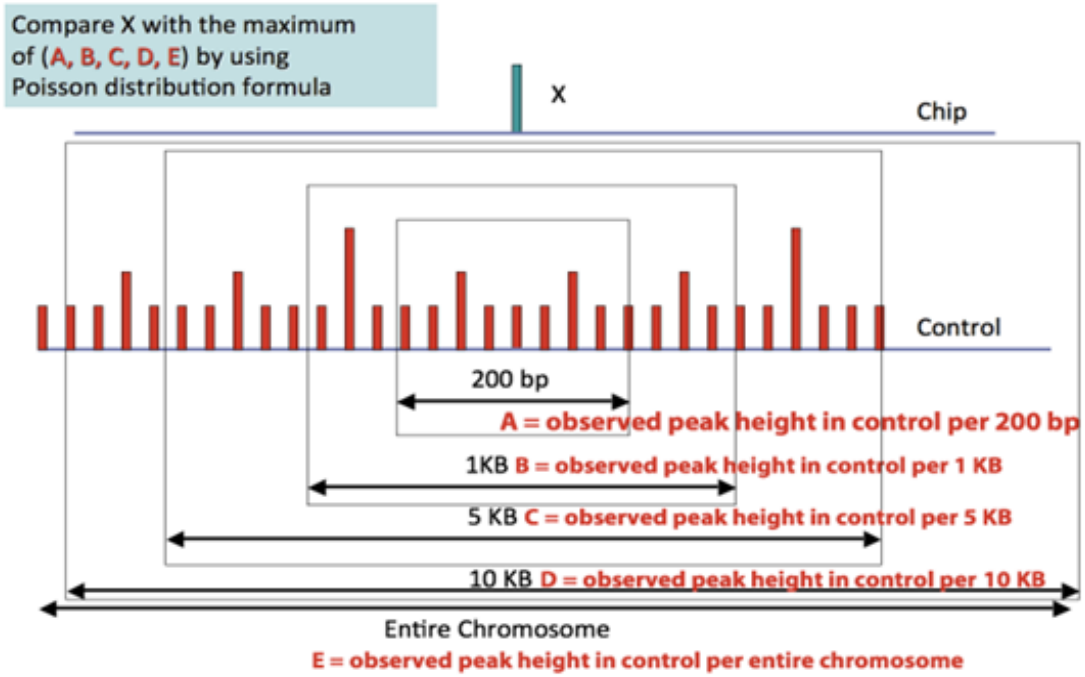
PEAK CALLING

Peak Calling:
Which Peak Caller to Use?

- Transcription factor peaks: **MACS** is very popular
- For histone marks with spanning longer regions, **Sicer** is recommended
  - MACS can be used by tweaking parameters
- Several peak callers in R/Bioconductor
  - e.g SPP, TPIC, BayesPeak
  - Not really considered gold-standard (other than SPP)
  - Often impractical: memory hungry and slow

Peak Calling: MACS

- Sliding window run across genome
- Peak height in window compared to that in windows of surrounding regions in control

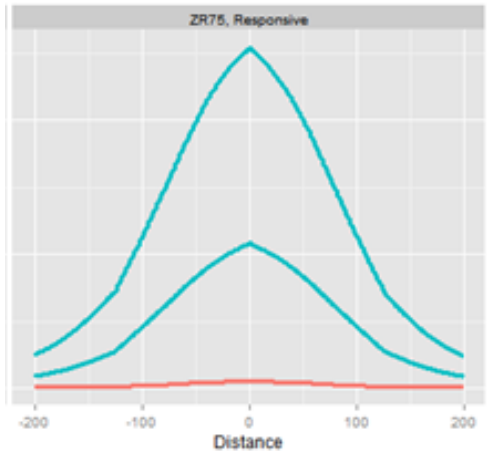


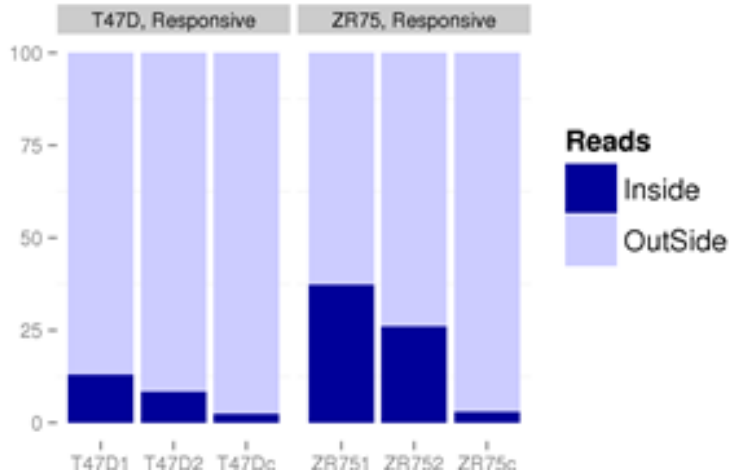
- Statistical significance of peak estimated by using Poisson distribution
  - -log10(pvalue) reported as peak score
- FDR calculated by calling peaks in control over sample

Peak Calling: Post-peak QC

- Peak profile plots
  - Mean read density at positions relative to peak summits
  - Input profiles should be flat
- Fraction of Reads in Peaks (FRIP)
  - Reads in peaks/Total mapped reads
  - Analogous to signal to noise ratio

Example output from ChIPQ package:





ChIP-Seq Practical

Working with ChIP-Seq Data in R/Bioconductor

chipqc_sweave.pdf