UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

ChIP-Seq Data Analysis: Pre-processing, QC and Primary Analyses

Suraj Menon

Bioinformatics Core

CRUK Cambridge Institute

Acknowledgements: Rory Stark, Thomas Carroll

Bias Alert

- Especially in the practicals!
- Tech bias: Illumina short read sequencing
- Experimental/biology bias: transcription factor binding
- Not everything on the course may be universally applicable to all ChIP-Seq analyses

Limitations of R/Bioconductor
.. and thus, this course

- Some processing steps/analyses are not (yet) possible or feasible in R/BioC
- Some processing/analyses are possible in R/BioC ... BUT
    - the “best” methodology may not be in R/BioC
    - it may just be easier and/or faster to do something outside of R
- Samtools, bedtools, Picard suite etc
- Galaxy: Cistrome (for ChIP-Seq)

“Typical” ChIP-Seq Analysis workflow

- Raw reads
- QC/Data viz/Filter
- Alignment
- QC/Data viz/Filter
- Primary analysis
  - Peak calling
- QC/Data viz /Filter

- “Downstream” analyses
  - Add biological context (e.g. Annotate peaks to genes)
  - Custom analyses specific to biological question
  - Integration with other data
    - Same platform
    - Different platform(!)

- Differential Binding Analysis

PRE-PROCESSING AND DATA QC

QC very important for ChIP-Seq data!

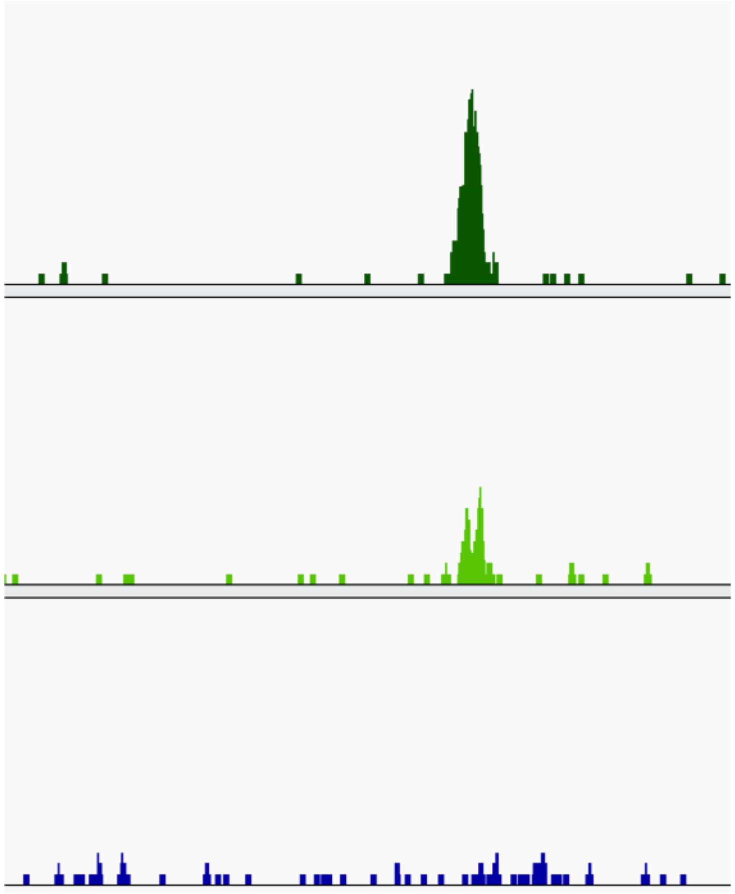
- ChIP Seq data is noisy
  - only a small proportion of reads actually represent protein-bound sequences. Mostly ‘background’
- Many sources of experimental bias
  - Antibody binding efficiency and specificity
  - Fragmentation biases
  - PCR amplification biases

Common QC/Filtering steps

- Visualisation of coverage profiles
- Dispersion of coverage
- Strand shift/ fragment length metrics
- Assessment/filtering of duplicate reads

QC: Visualise coverage profiles

- Simplest QC
  - Qualitative and subjective
- Various data formats
  - Wigs, Bams, bigWigs, bedGraphs
- Various browsers
  - UCSC, Ensembl, IGV
- Recommendation:
  - bigWigs on IGV

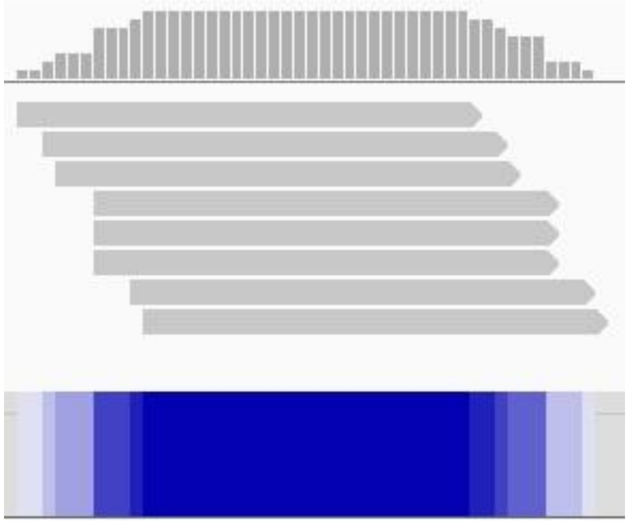


QC: Dispersion of coverage

- 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

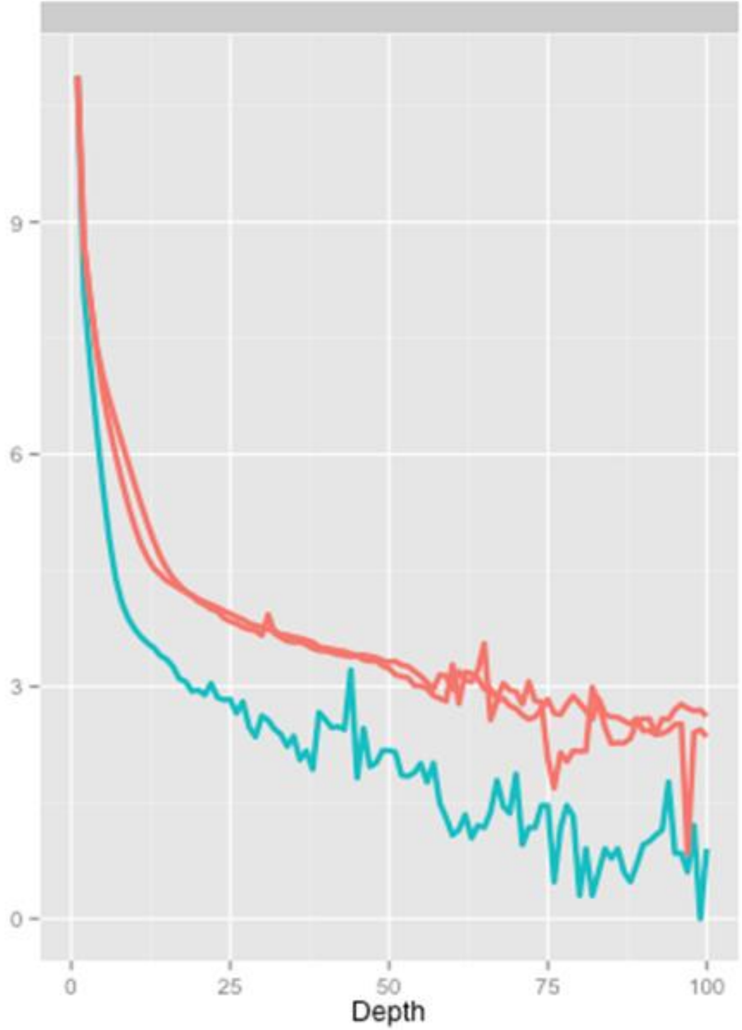
QC: Dispersion of coverage

- 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



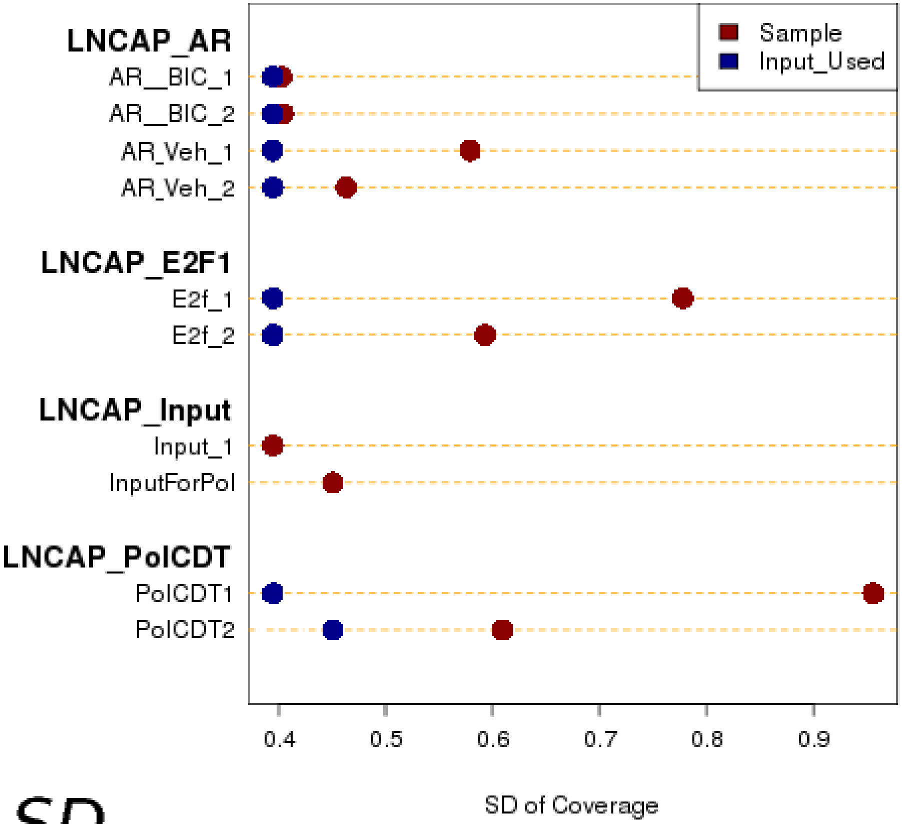
QC: Dispersion of coverage

- SSD: Standardised Standard Deviation of coverage

- Metric for assessment of dispersion of coverage

- High for samples with enriched regions (ChIP)

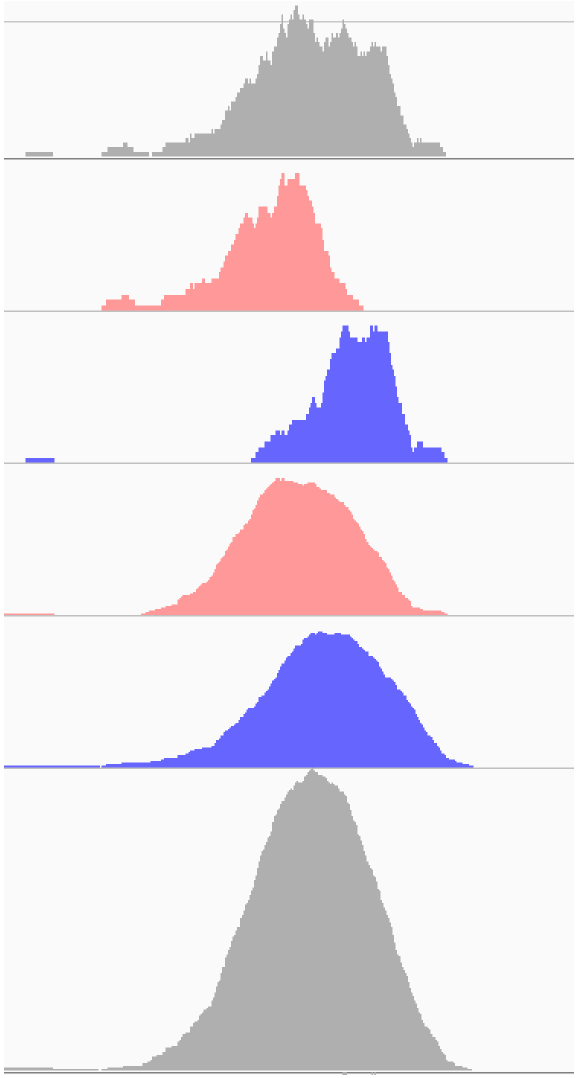
- Low for samples with uniform coverage (input)



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

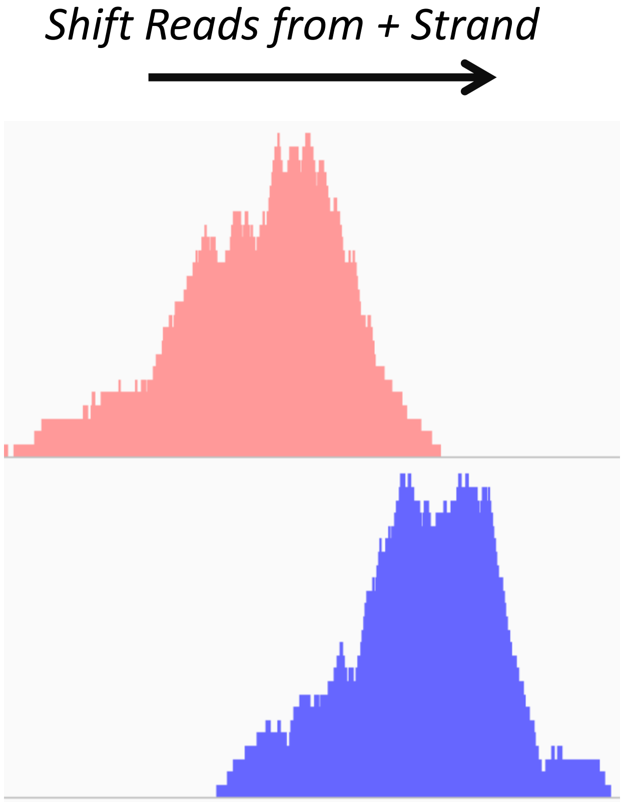
QC: Strand shift/ fragment length

- Bias in ChIP-Seq data:
  - Only ends of a fragment are sequenced
  - Shift is apparent between reads aligning to the Watson and Crick strands
  - Two distributions of peaks around centre of true enrichment
- Reads need to be extended to fragment length to re-create true signal



QC: Strand shift/ fragment length

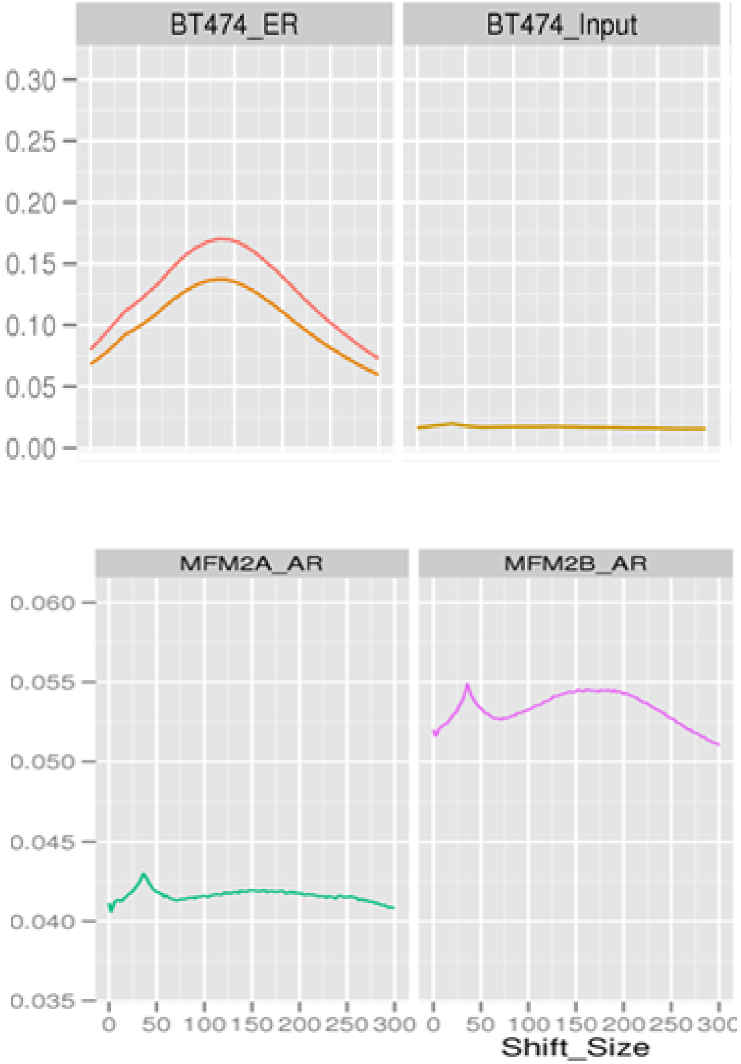
- Fragment length can be estimated from data:
  - **Cross-correlations -** Correlation of reads on positive and negative strand after successive read shifts
  - **Cross-coverage -** Coverage of reads on both strand after successive shifts of reads on one strand



- These provide useful QC metrics

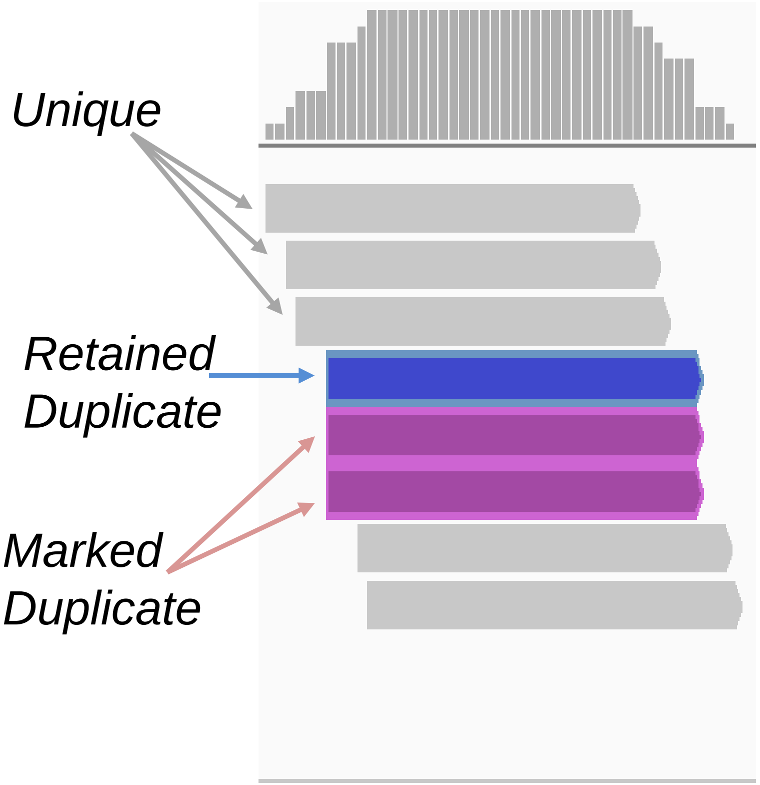
QC: Strand shift/ fragment length

- Cross-correlation/Cross-coverage score plots are useful for QC
- Peaks should be seen at the fragment length for enriched ChIP samples
- Small to non-existent peaks are seen in failed ChIPs and inputs



QC: Assessing/Filtering duplicates

- **Single-end Duplicate** is **read** with **same start** position.
- First read at duplicated position is **retained** and remaining are **marked**.
- Duplicates can represent experimental artefacts, but not all the time!



QC: Assessing/Filtering 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

QC: Assessing/Filtering 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 signficantly anomalous duplicate numbers.

QC: Assessing/Filtering 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

Further ChIP-Seq QC considerations

- Proportions of reads in biologically relevant regions
  - e.g. windows around promoters, intergenic regions

- Filtering out reads aligning to ‘blacklist’ regions
  - Encode empirically identified regions that showed anomalous and near-universal artefact signal
  - Various reasons e.g. chromatin accessibility, repeats
  - Enriched for duplicate and multi-mapping reads
  - Adversely affect fragment length calculations and in thus any analyses that require these e.g. peak calling

ChIP-Seq QC resources

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

  (Landt et al – *Genome Research 2012*)

- Bioconductor package **ChIPQC**
- R package **SPP** (for UNIX/LINUX)

PEAK CALLING

Peak Calling: Experimental Considerations

- 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

Peak Calling: Experimental Considerations

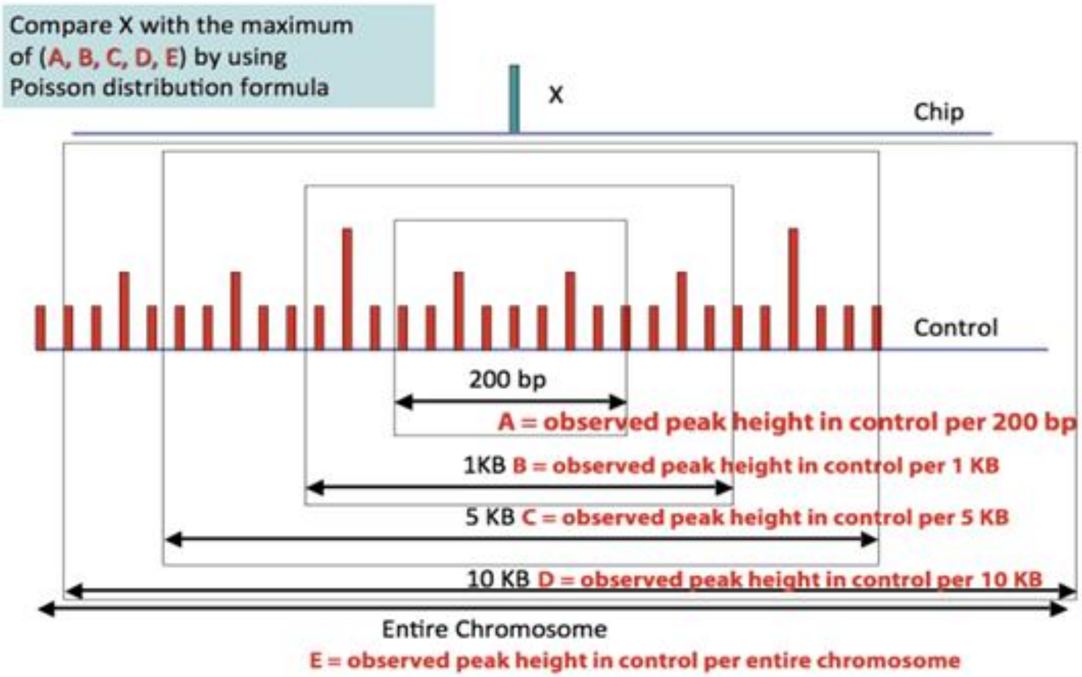
- Replicates
  - Biological (as much as possible) rather than technical
  - Different antibody for enrichment
- Check paramaters of peak caller!
  - Do duplicates need to be removed?
  - Do reads need to be extended to fragment length?

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

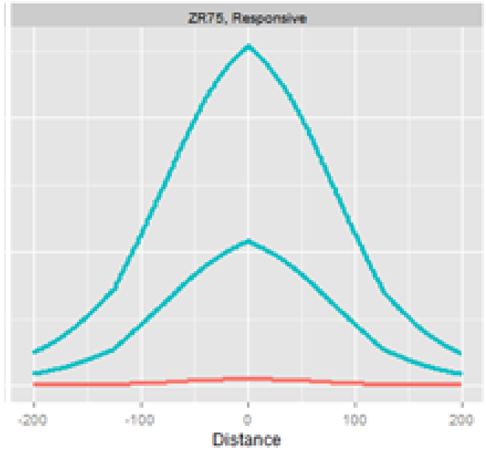


Compare X with the maximum of (A, B, C, D, E) by using Poisson distribution formula
X
Chip
Control
200 bp
A = observed peak height in control per 200 bp
1KB B = observed peak height in control per 1 KB
5 KB C = observed peak height in control per 5 KB
10 KB D = observed peak height in control per 10 KB
Entire Chromosome
E = observed peak height in control per entire chromosome

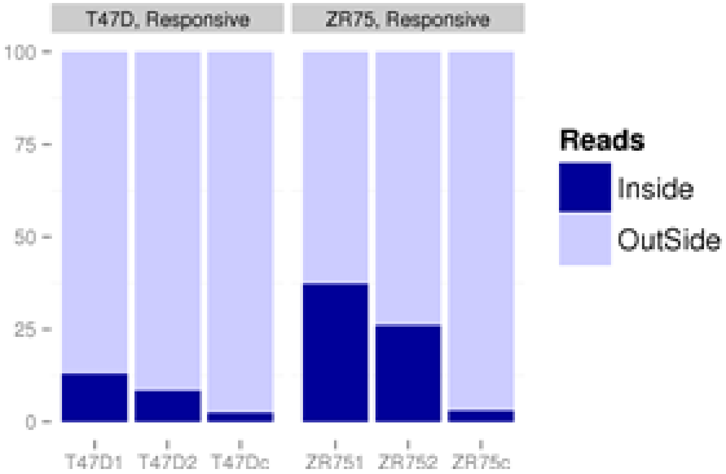
- Statistical significance of peak estimated by using Poisson distribution
  - -log10(pvalue) reported as peak score
- FDR calculated by calling peaks in control over sample

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



ChIP-Seq Practical

Working with aligned read data and peaks in R/Bioconductor