Best practices in the analysis of RNA-seq and ChIP-seq data

27th – 31st, July 2015

University of Cambridge, Cambridge, UK

ChIP-Seq Data Analysis:

Pre-processing, QC and Primary Analyses

Presenter: Ines de Santiago

CRUK Cambridge Research Institute

Ines.desantiago@cruk.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

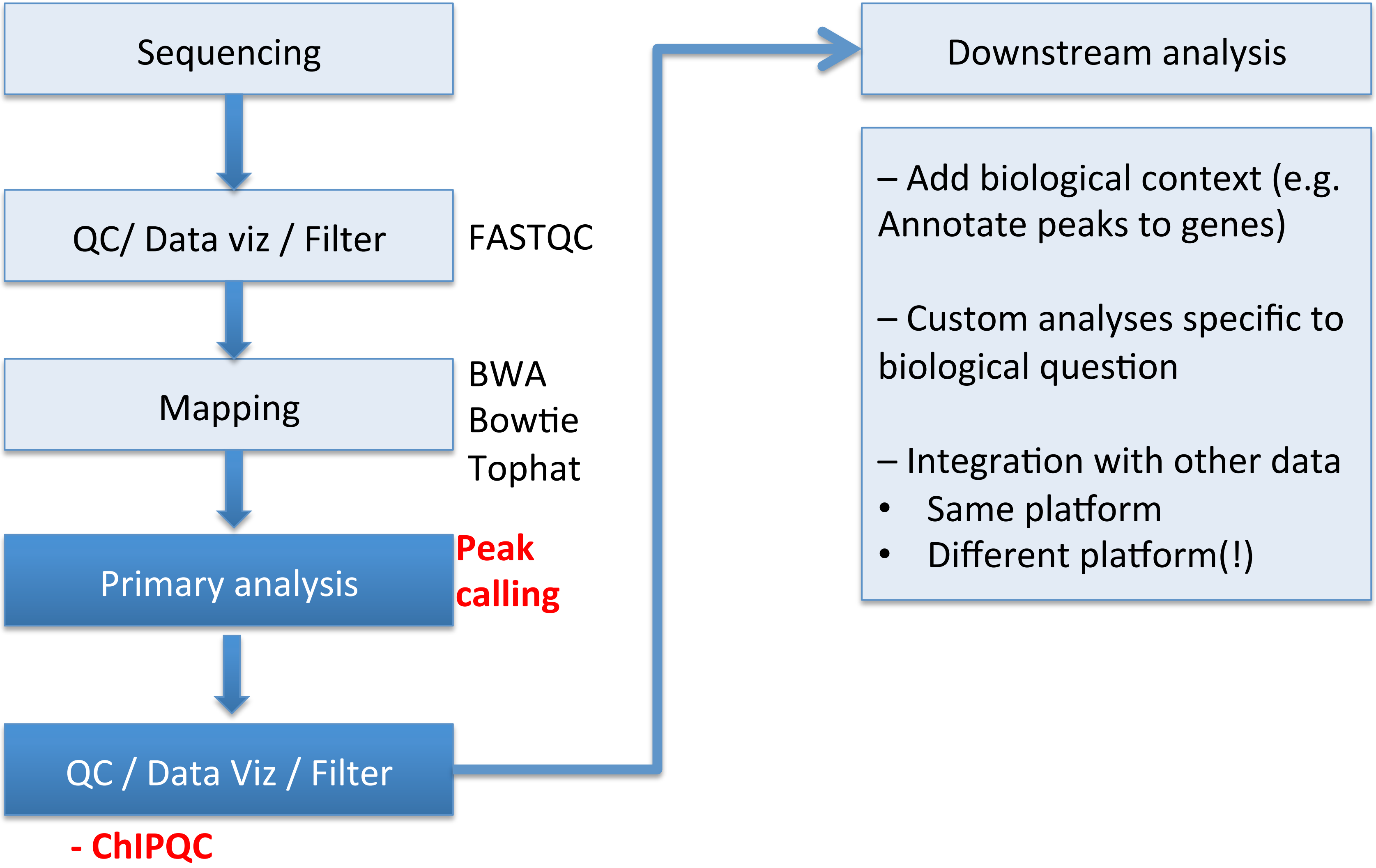
CAMBRIDGE Cancer CENTRE

CANCER RESEARCH UK

Acknowledgments

- **Tom Carroll**
  - http://bioconductor.org/help/course-materials/2014/BioC2014/Bioc2014_ChIPQC_Practical.pdf
  - http://bioconductor.org/help/course-materials/2014/BioC2014/ChIPQC_Presentation.pdf
- Shamith Samarajiwa
- Suraj Menon

“Typical” ChIP-Seq Analysis workflow



Sequencing

QC/ Data viz / Filter

FASTQC

Mapping

BWA
Bowtie
Tophat

Primary analysis

Peak calling

QC / Data Viz / Filter

- ChIPQC

Downstream analysis

– Add biological context (e.g. Annotate peaks to genes)

– Custom analyses specific to biological question

– Integration with other data

- Same platform
- Different platform(!)

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

- Highly variable patterns of enrichment between ChIPs.
  - Transcription factors show sharp/narrow peaks.
  - Histones more dispersed/broad peaks

ChIP-Seq QC resources

- **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)

Common QC/Filtering steps (ChIPQC)

- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Signal in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads.**
- **Duplication Rate**

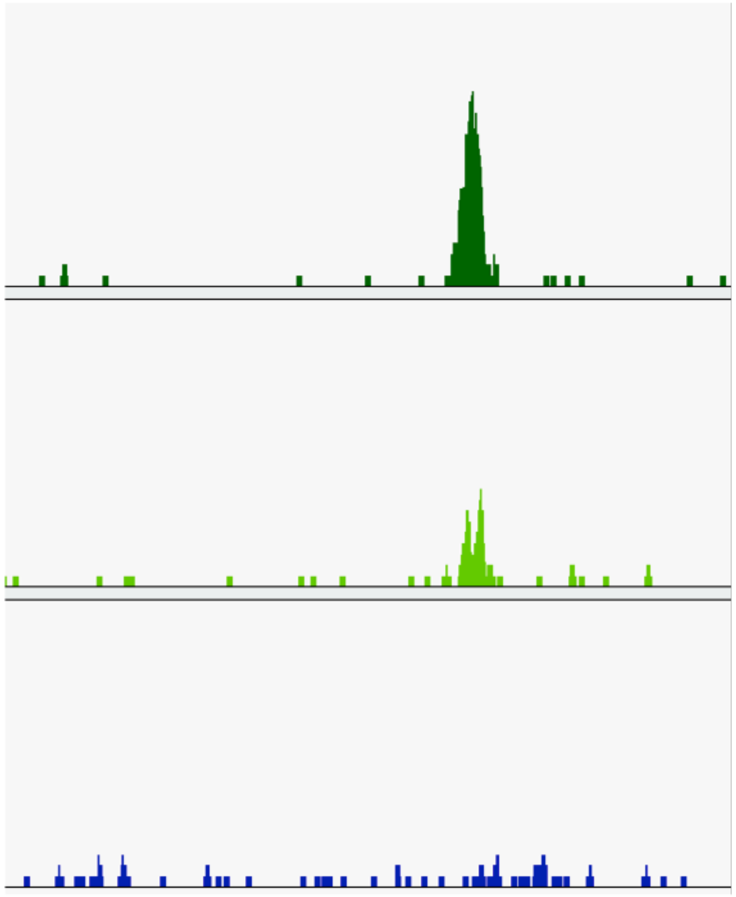
Common QC/Filtering steps (ChIPQC)

- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Signal in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads.**
- **Duplication Rate**

Distribution of Signal

Visualise coverage profiles

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

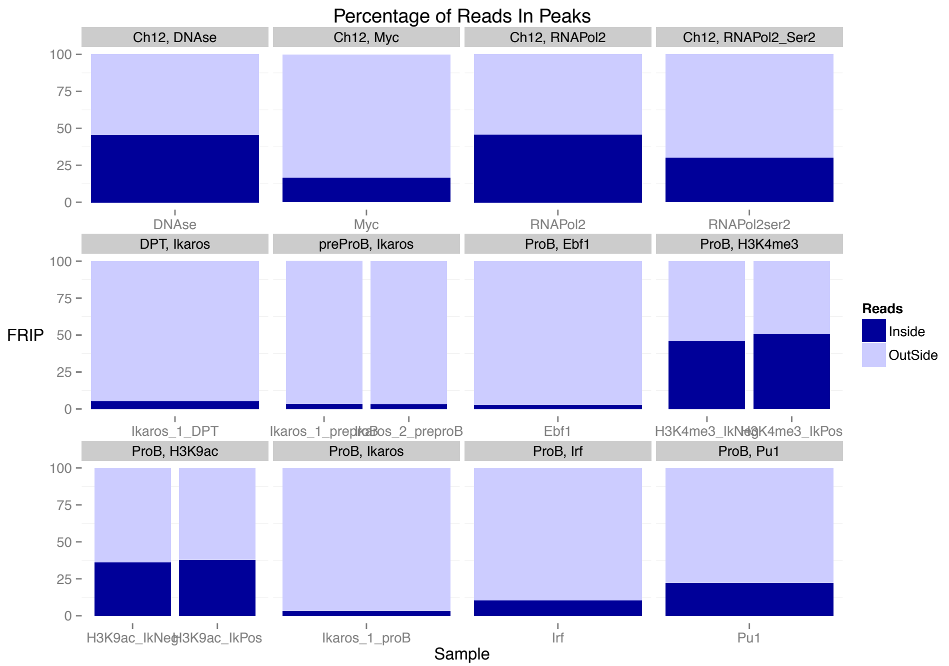


Distribution of Signal

Signal in Peaks (FRIP)

- The simplest assessment of enrichment.

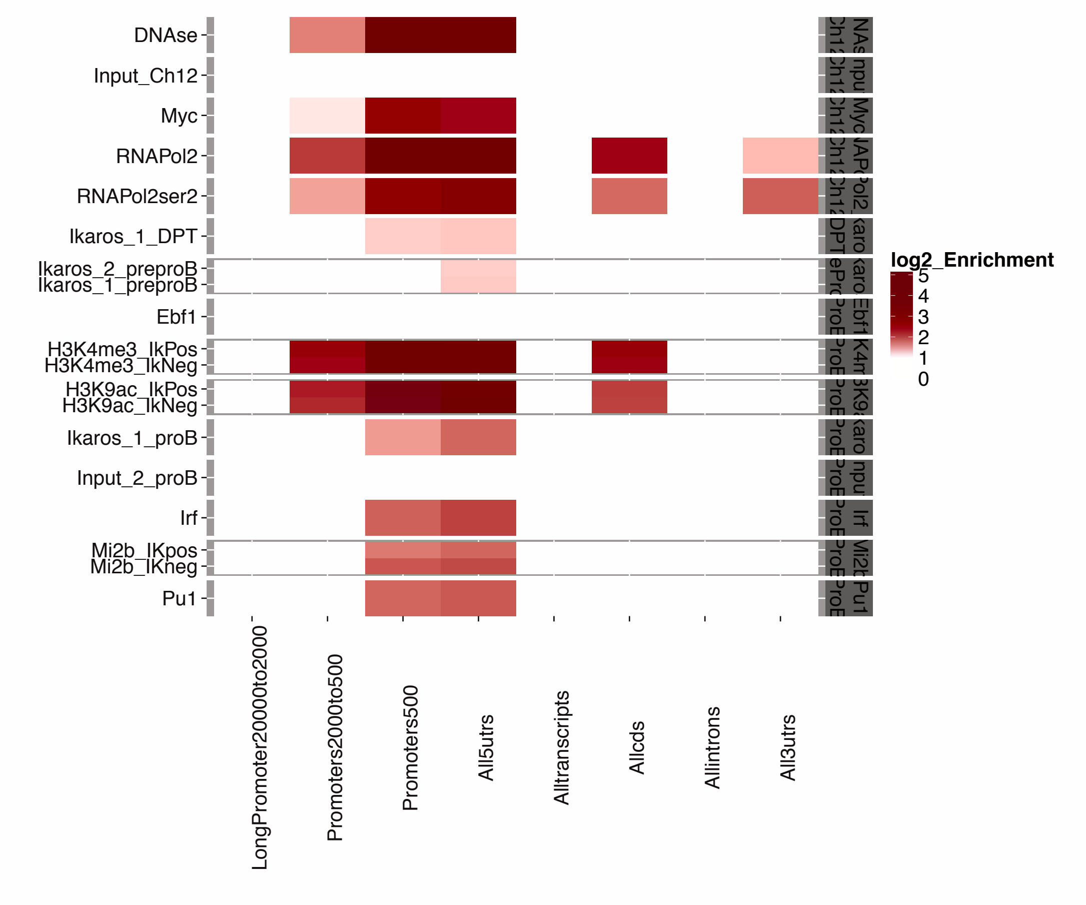
- Good quality TF > 5%
- Good quality Pol-II > 30%



Relative Enrichment in Genomic Intervals (REGI)

Signal in Peaks (FRIP)

- Plot relative enrichment of reads in annotated regions.



Distribution of Signal

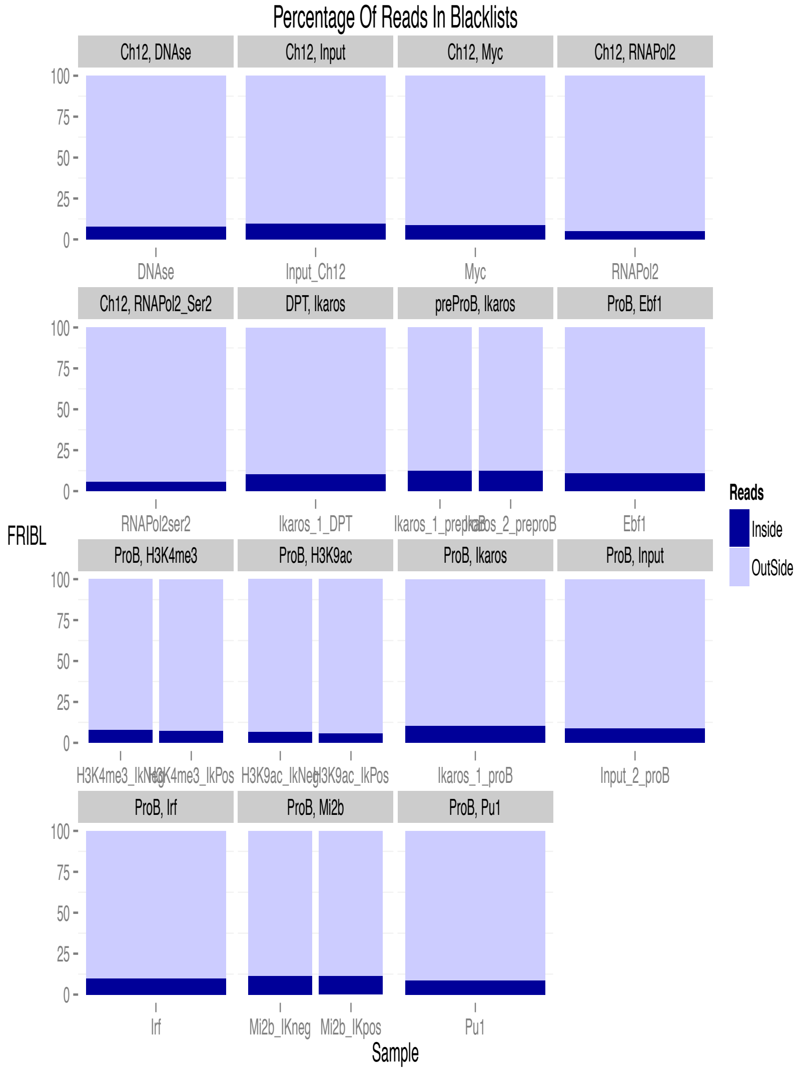
Signal in blacklists

- 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



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

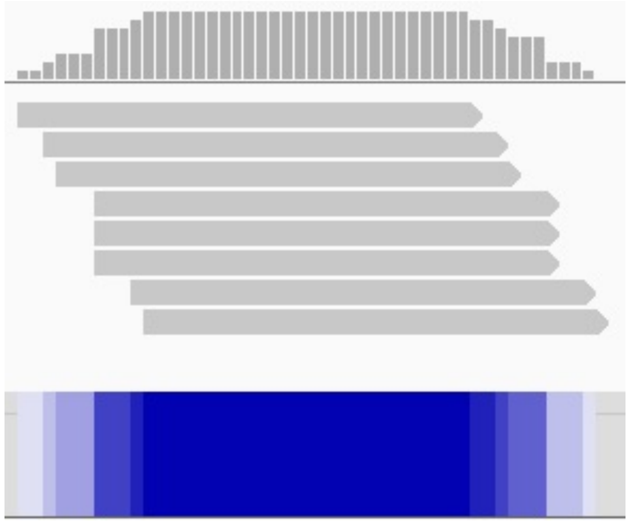
Distribution of Signal

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

Distribution of Signal

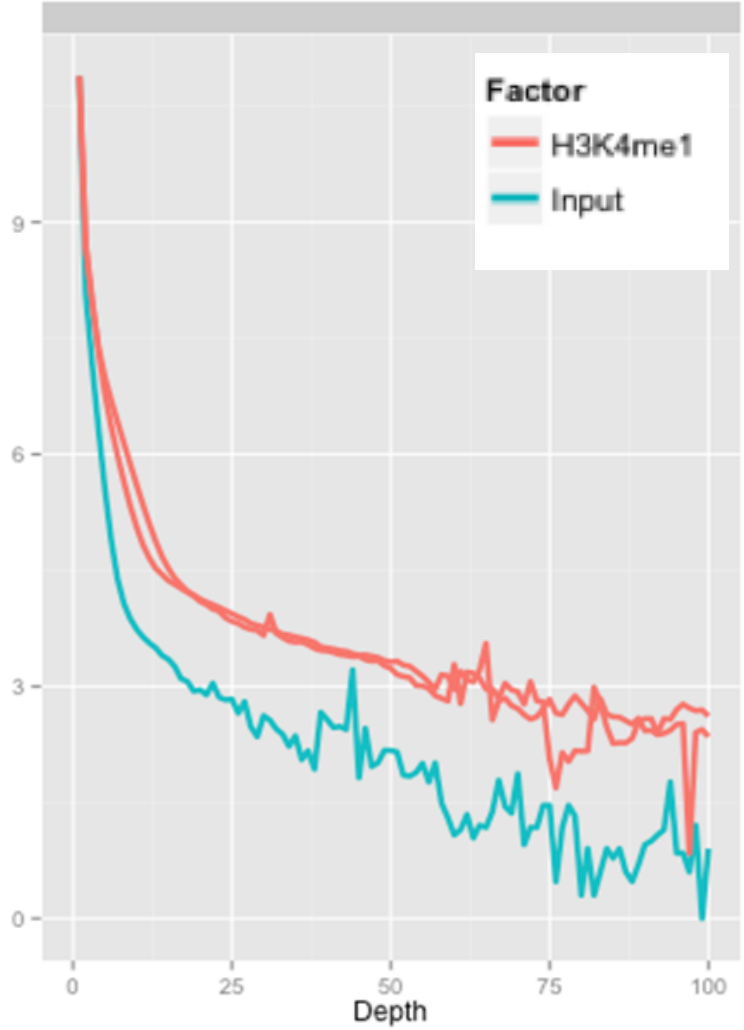
Dispersion of coverage

- 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



Distribution of Signal

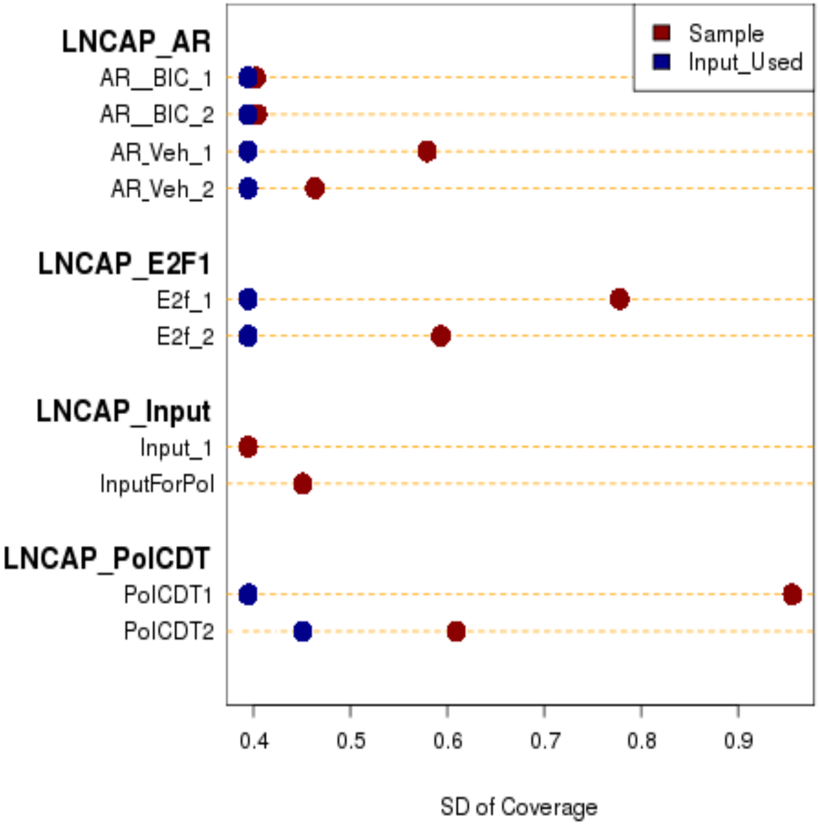
Metric for dispersion of coverage: SSD

- 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)

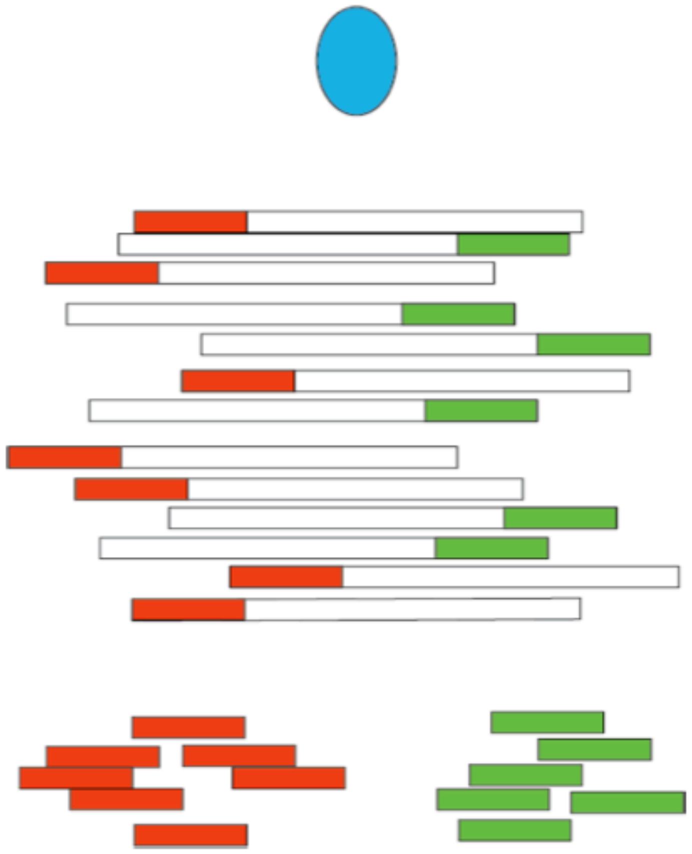


Common QC/Filtering steps (ChIPQC)

- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Signal in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads**
- **Duplication Rate**

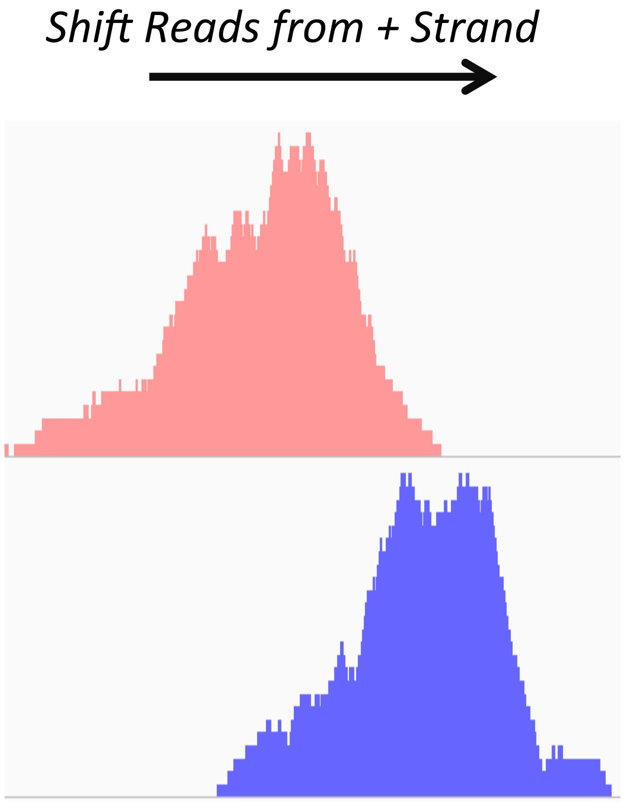
Clustering of Watson/Crick reads

- Bias in ChIP-Seq data:
  - Only ends of a fragment are sequenced
  - Shift is apparent between reads aligning to the Watson and Crick strands
  - For transcription factors the extent of this clustering related to ChIP-seq quality
- Reads need to be extended to fragment length to re-create true signal



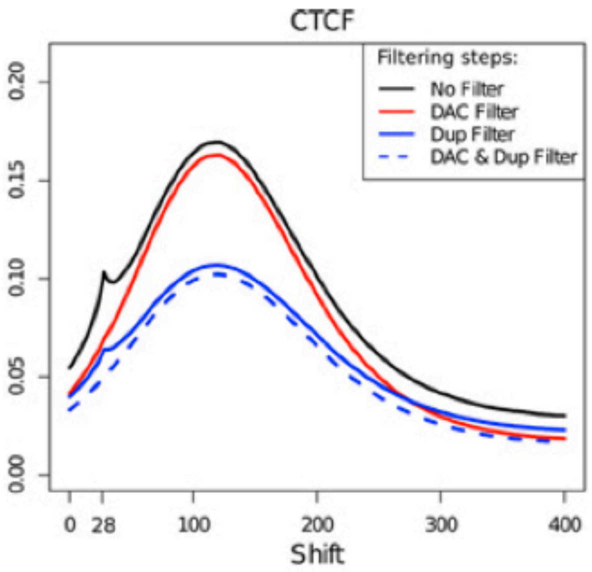
Metrics to assess W/C read clustering

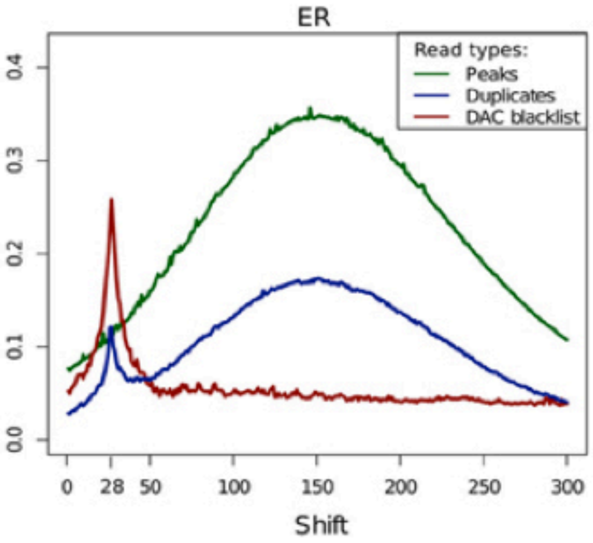
- Fragment length can be estimated from data:
  - **Cross-correlations -** Correlation of reads on positive and negative strand after successive read shifts
  - **Cross-coverage -** Coverage of reads on both strand after successive shifts of reads on one strand. Total area covered by reads will be reduced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



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:
  - $FragCC = CC_{fragmentlength}$.
  - $RelCC = FragCC/ CC_{readlength}$



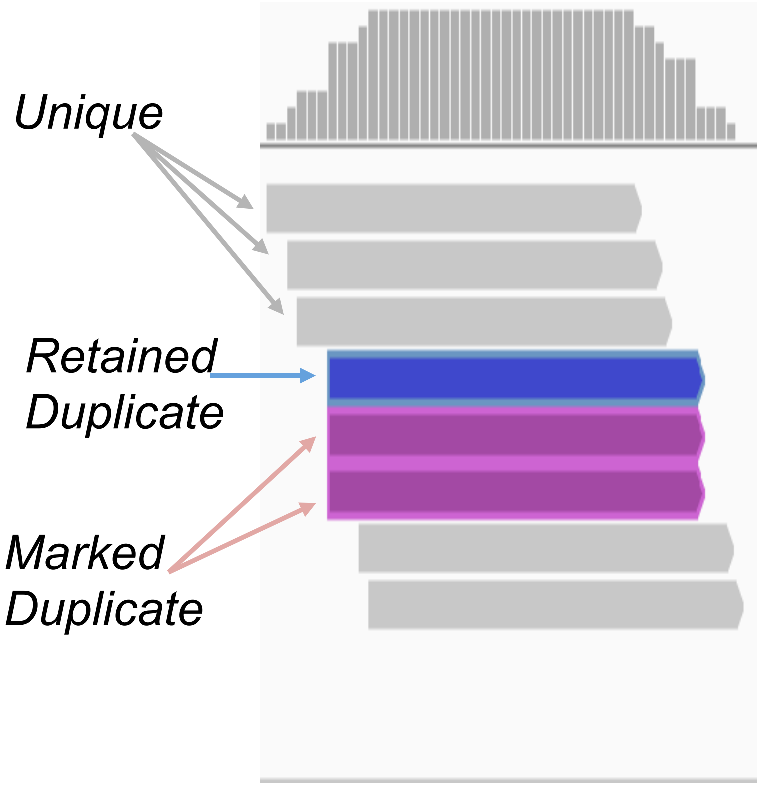


Common QC/Filtering steps (ChIPQC)

- **Distribution of Signal**
  - Visualisation of coverage profiles
  - Signal in peaks (FRIP)
  - Relative enrichment in genomic intervals (REGI)
  - Signal in blacklists (FRIBL)
  - Dispersion of coverage
- **Clustering of Watson/Crick reads**
- **Duplication Rate**

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!



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

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 significantly anomalous duplicate numbers.

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

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

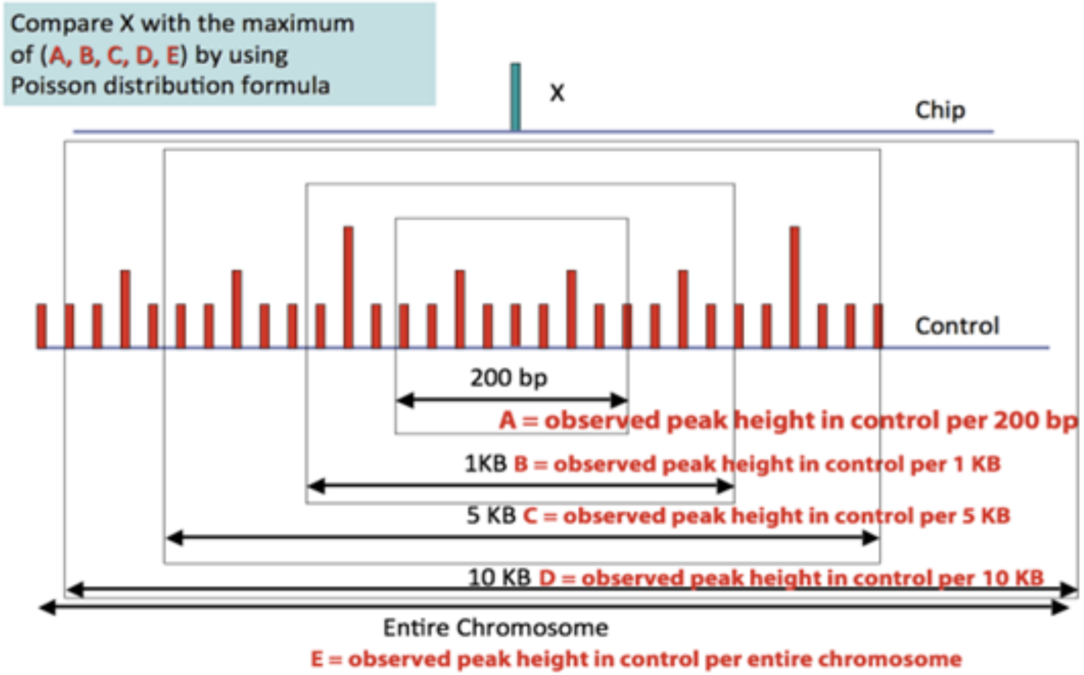
- Replicates
  - Biological (as much as possible) rather than technical
  - Different antibody for enrichment
- Check parameters of peak caller!
  - Do duplicates need to be removed?
  - Do reads need to be extended to fragment length?

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

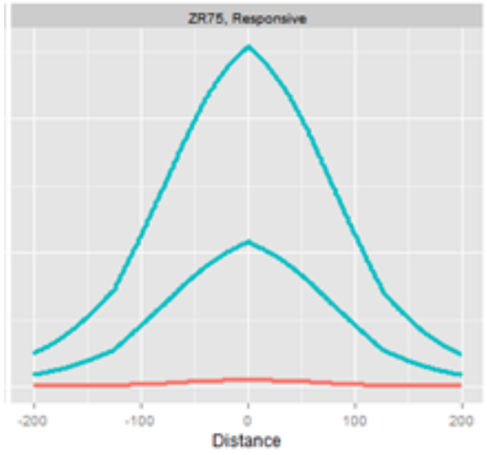
- Sliding window run across genome
- Peak height in window compared to that in windows of surrounding regions in control

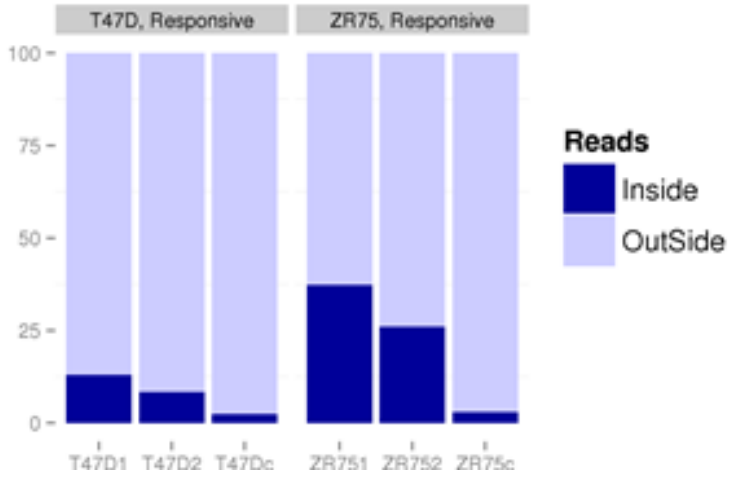

Compare X with the maximum
of (A, B, C, D, E) by using
Poisson distribution formula
X
Chip
Control
200 bp
A = observed peak height in control per 200 bp
1KB B = observed peak height in control per 1 KB
5 KB C = observed peak height in control per 5 KB
10 KB D = observed peak height in control per 10 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 ChIP-Seq Data in R/Bioconductor

chipqc_sweave.pdf