Visualisation and assessment of ChIP-seq quality

Thomas Carroll

Head of Bioinformatics,
MRC Clinical Sciences Centre,
Imperial College London

BioC 2014

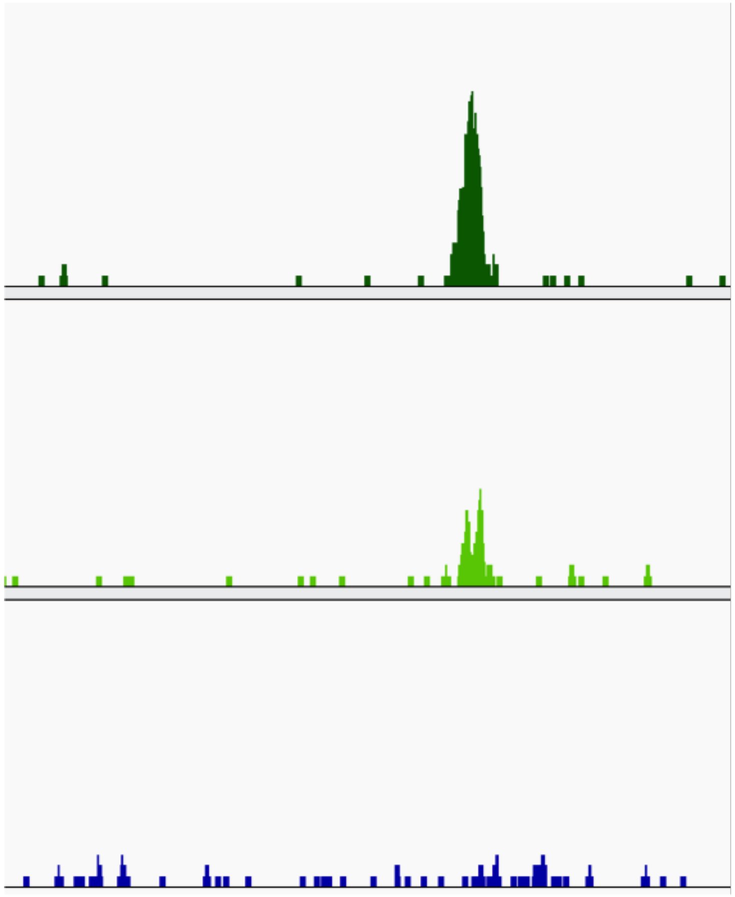
ChIP-seq is noisy

- ChIP-seq/ChiP-exo/DNA-seq/MNase-seq is noisy.
- Experimental biases:
    - Fragmentation/digestion.
    - IP strength/efficiency and specificity.
    - PCR Bias (Overamplification from low starting material)
- Highly variable patterns of enrichment between ChIPs.
    - Transcription factors may show sharp/narrow peaks.
    - Polymerase II will show mix of sharp/narrow and dispersed/broad peaks

Always visualise your data

- Coverage graphs.
  - Wigs (Okay)
  - bedGraphs (Okay)
  - BigWigs (Great)
- Allows for quick assessment of data...

  ..but dependent on user's interpretation/experience.



High-thoughput ChIP-seq quality control with **ChIPQC**

- Need methods to quantify informative characteristics about your ChIP-seq data.

- **ChIPQC –** Tom Carroll and Rory Stark *(Diffbind)*.

- **ChIPQC** provides workflow to generate metrics per sample/experiment.

ChIP-seq metrics

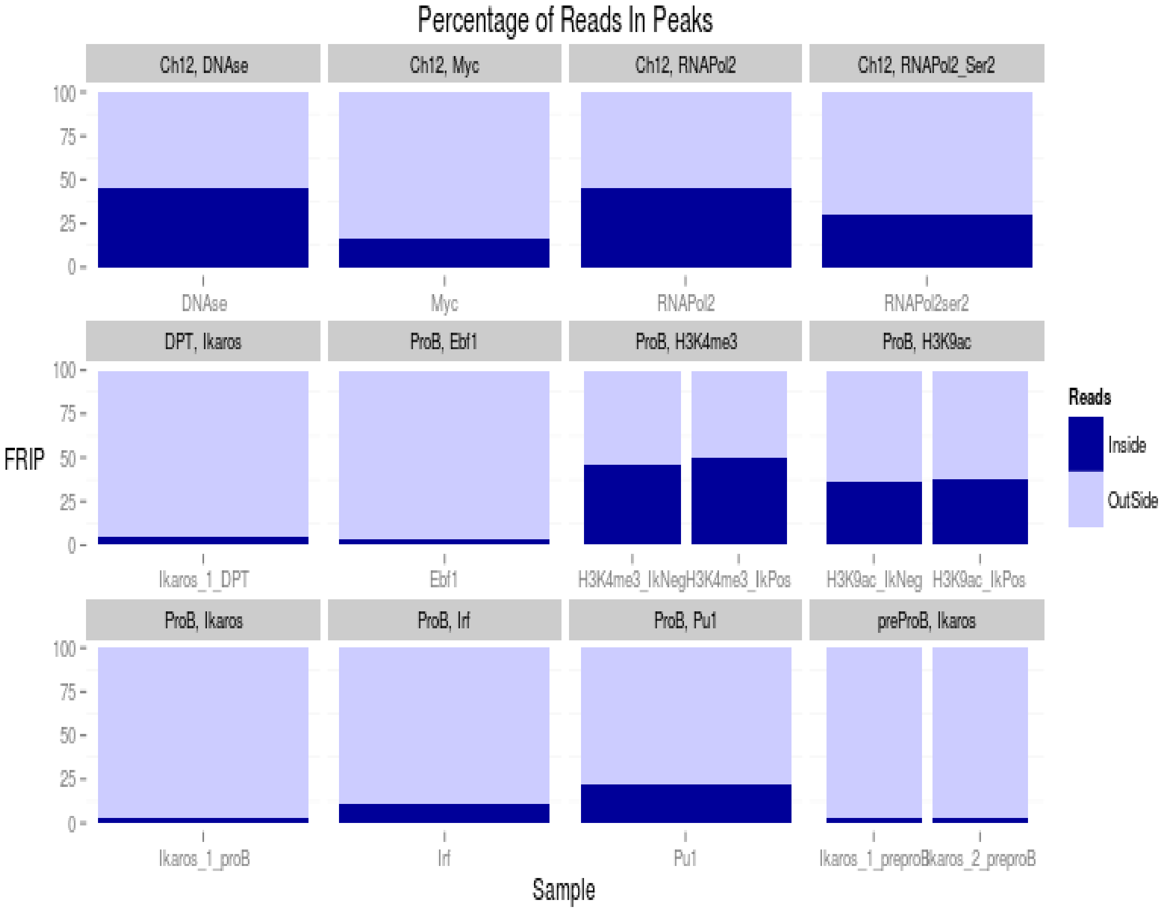
- **Distribution of Signal**
- **Clustering of Watson/Crick reads.**
- **Duplication Rate.**

Distribution of Signal

- Within enriched regions
- Within/across expected annotation
- Across the genome
- Within known artefact regions

Signal in Peaks (FRIP)

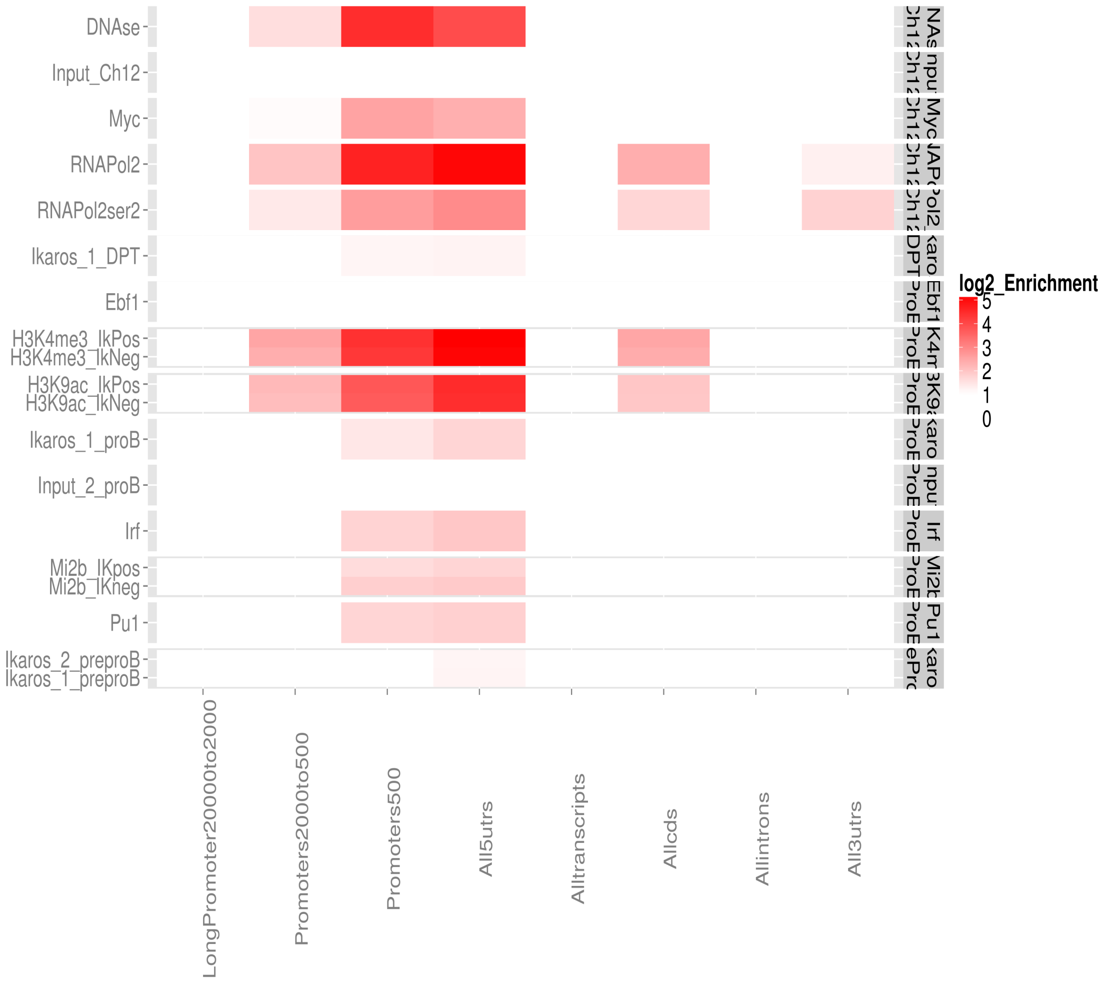
- The simplest assessment of enrichment.
  - Call enriched regions over input
  - Measure fraction of reads in peaks (FRIP)
  - Good quality TF > 5%
  - Good quality Pol-II > 30%



Relative Enrichment in Genomic Intervals (REGI).

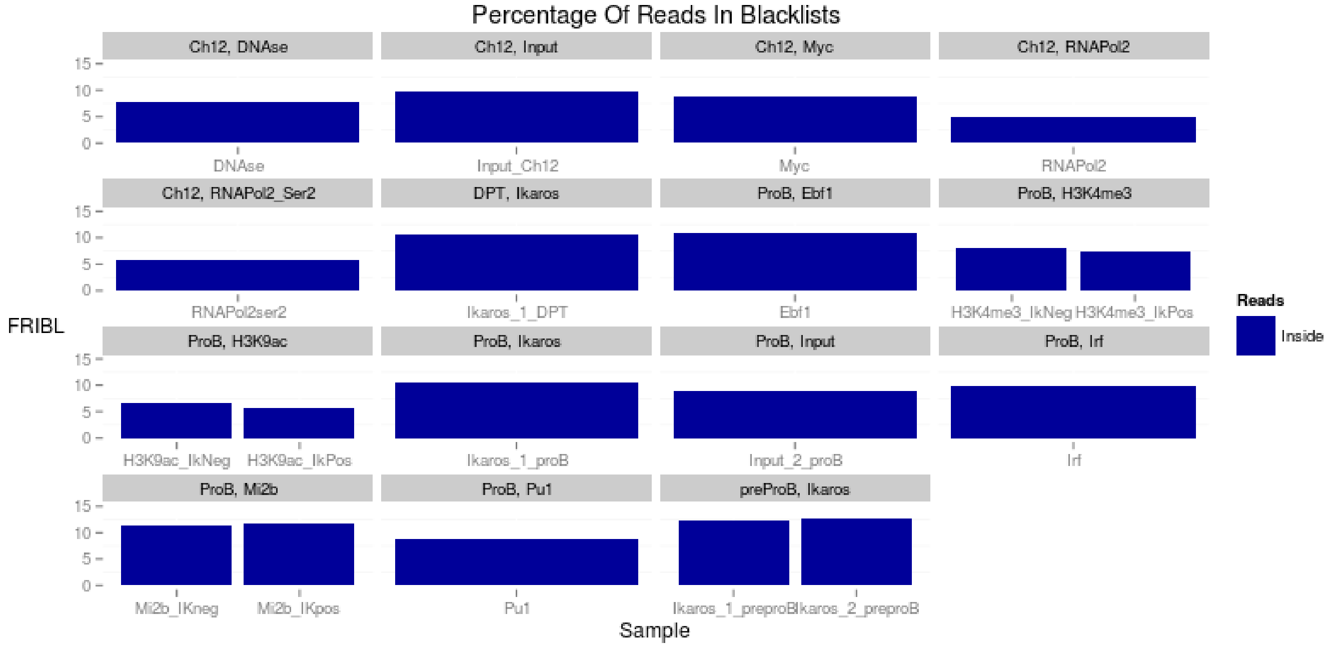
- Expected enrichment in genomic regions

- Plot relative enrichment of reads in annotated regions.



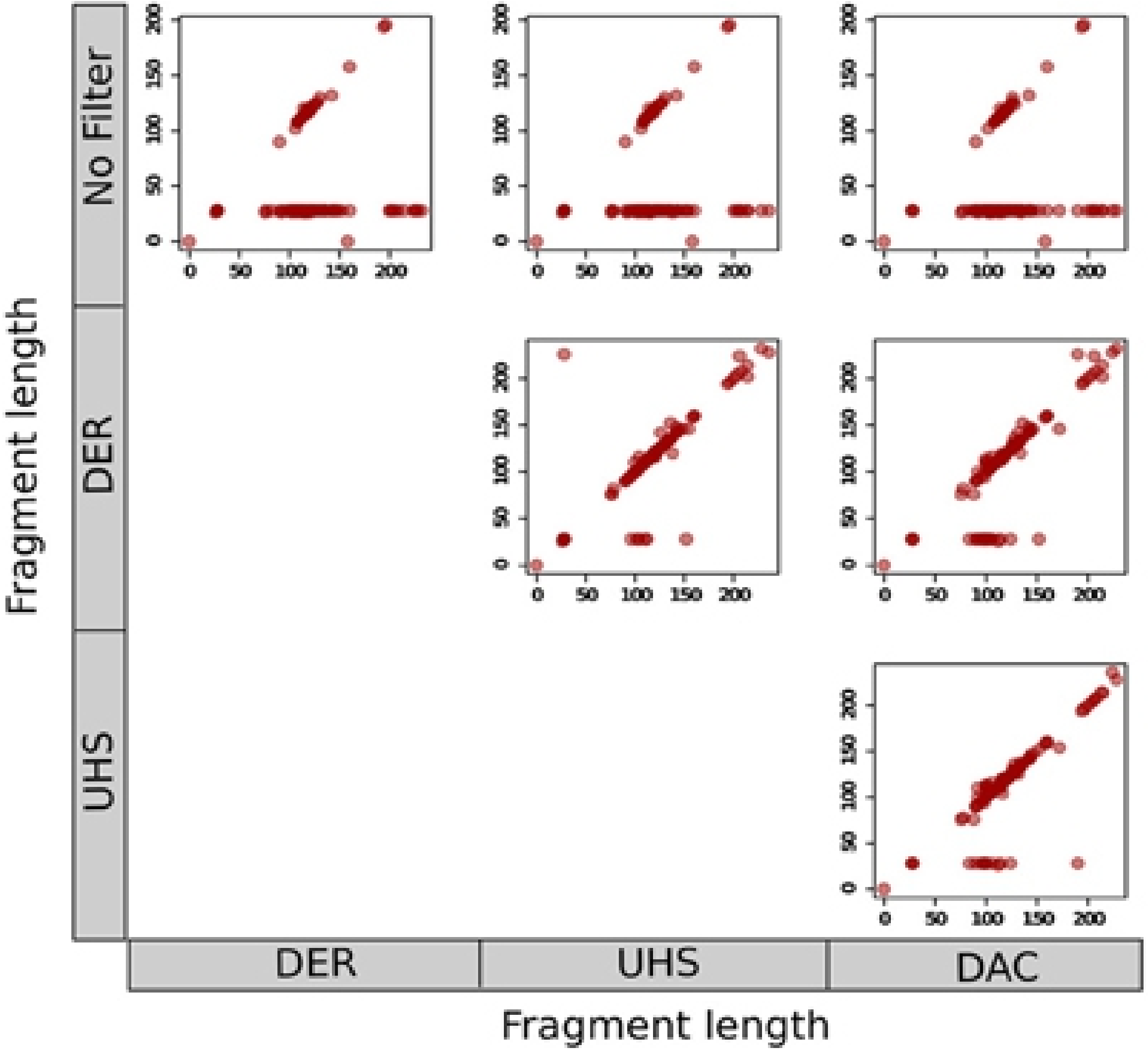
Signal in Blacklists (FRIBL)

- Work from Encode (Kudaje A) has produced curated list of conserved high signal artefact regions.
- Available for many species including human, mouse and drosophila genomes.
- Represent around 0.5% of genome.
- Can account for high proportion of total signal (> 10%).



Why worry about blacklists?

- Can affect -
  - Normalisation between samples.
  - Fragment length estimation.
  - Quality metrics for ChIP-seq.

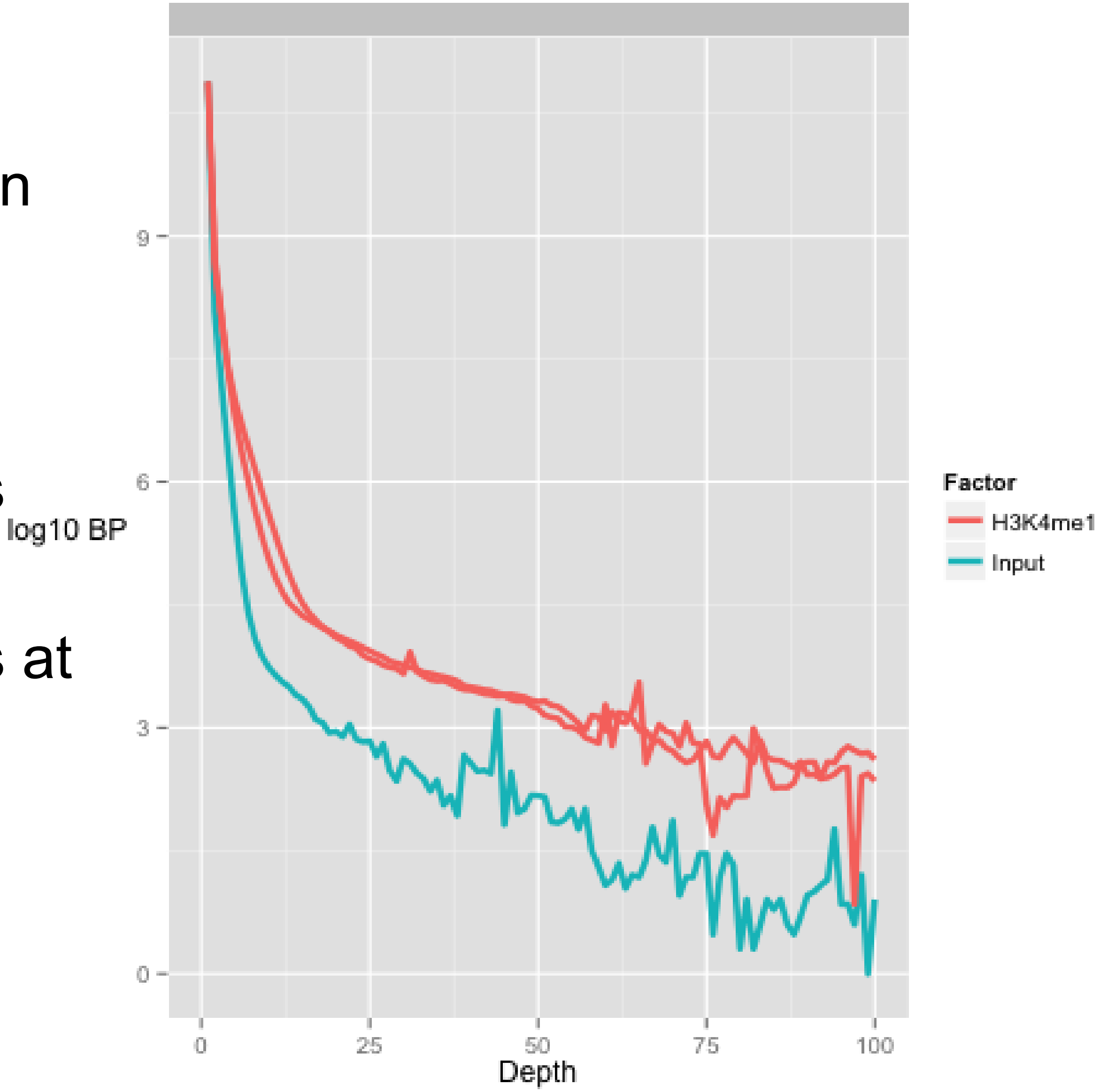


Predicted fragment length before and after blacklisting

Carroll et al 2014

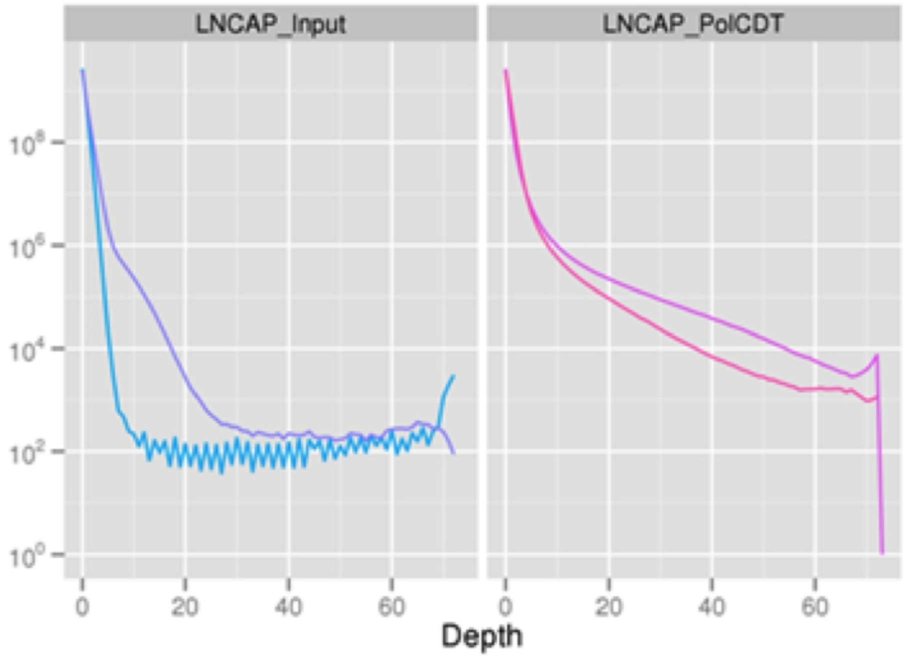
Global signal profile

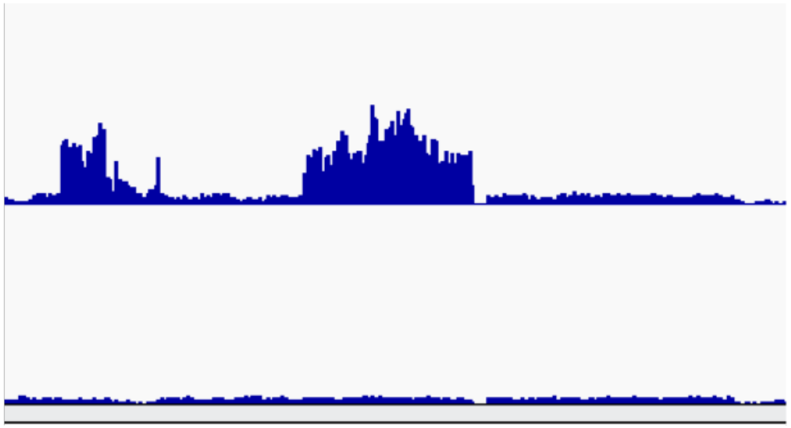
- A simple method to review global distribution is as histograms.
- More enriched samples show higher number of bases at greater depths
- Input samples show higher number of bases at low depths



Global Signal Profile

- Presence of stretch of high signal depth
- Identify anomalous signal region as candidate for blacklisting.





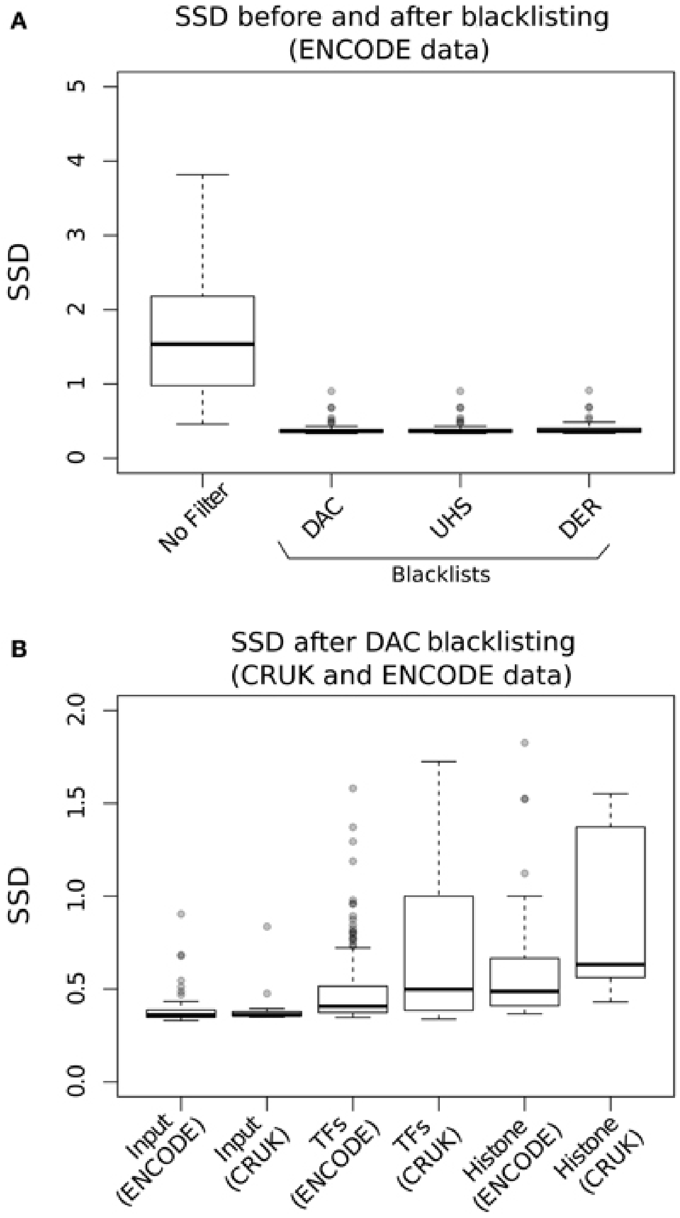
Metric of Global Signal Profile - SSD

- SSD developed in htseqtools package.
- Normalised standard deviation of coverage.
- Provides measure of pile-up across genome
  - Sample with regions of high signal (High SSD score)
  - Sample with low signal across genome (Low SSD score)

- Provides no measure of signal structure.

SSD and Blacklists

- SSD is very sensitive high signal artefact regions.
- Input SSD scores reduced after Blacklisting
- Sample SSD scores remain higher.



SSD before and after blacklisting (ENCODE data)

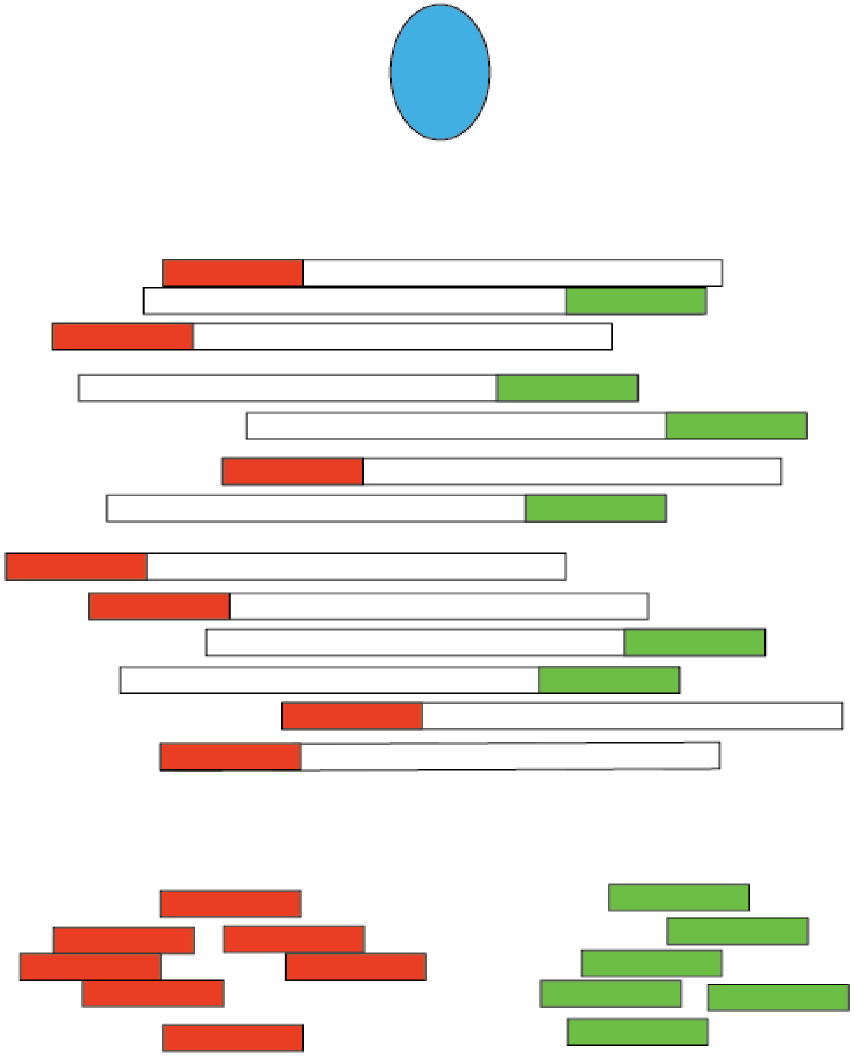
SSD after DAC blacklisting (CRUK and ENCODE data)

Carroll et al 2014

Clustering of Watson/Crick reads.

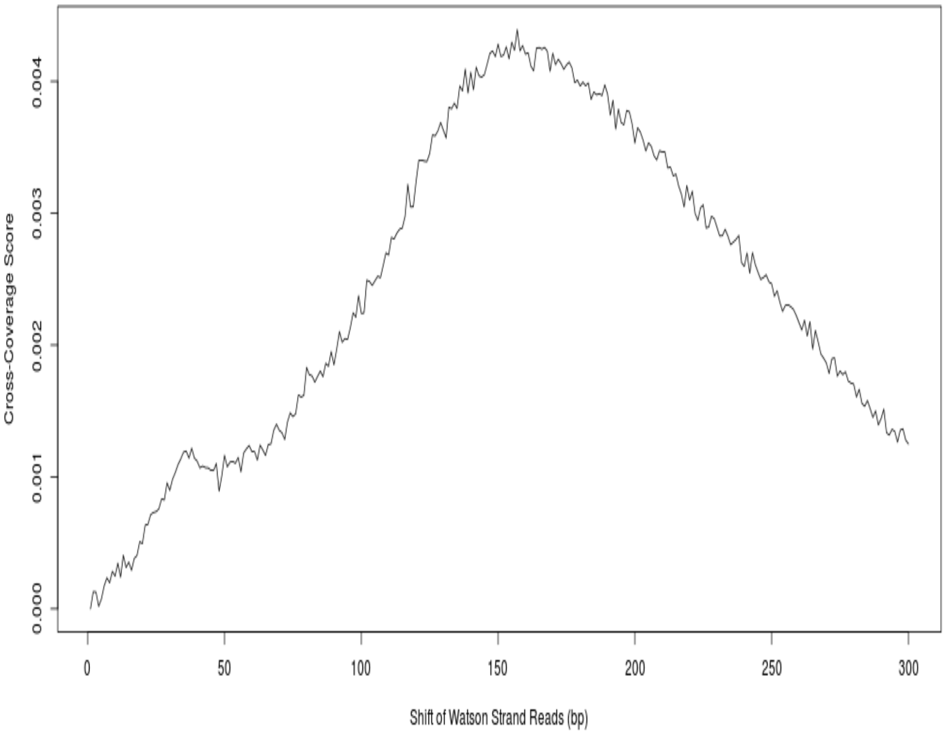
Watson and Crick reads cluster around epigenetic marks

- ChIP-seq is typically single ended.
- ChIP-seq watson and crick reads cluster around binding events.
- For transcription factors the extent of this clustering related to ChIP-seq quality.



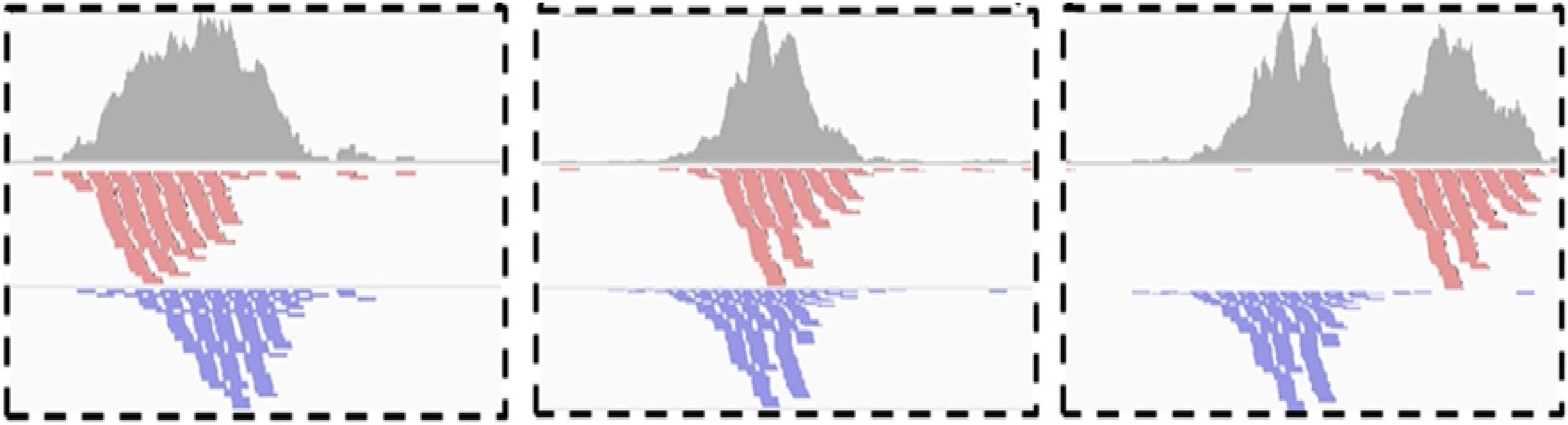
Assessing W/C read clustering

- Convert total coverage to cross-coverage scores to allow for comparison between samples (and regions)
  - *Cross-Coverage Score =(Coverage*$_0$ *– Coverage*$_n$*)/Coverage*$_0$



- *Frag_CC = Cross-coverage score at fragment length.*

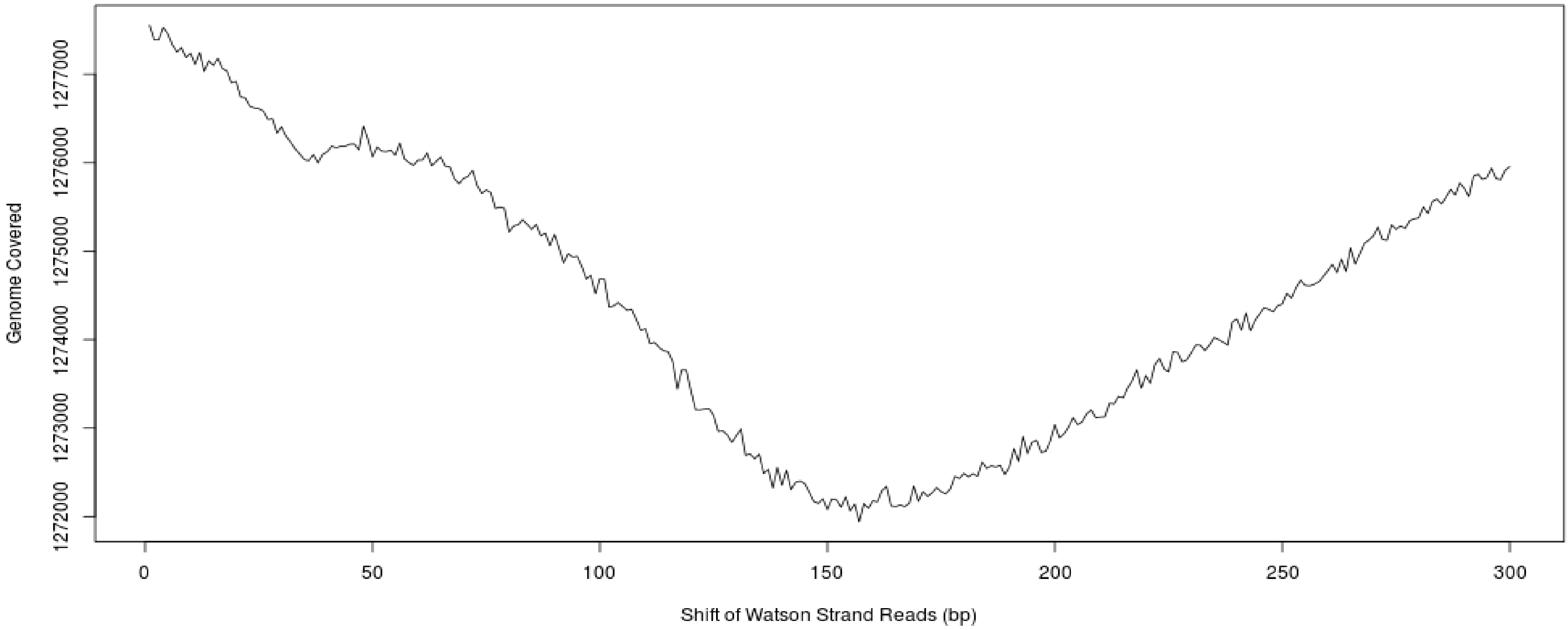
Assessing W/C read clustering



- Slide Watson reads along binding site (5' to 3').
- Total area covered by signal will reduce after shifting Watson reads by fragment length

Assessing W/C read clustering

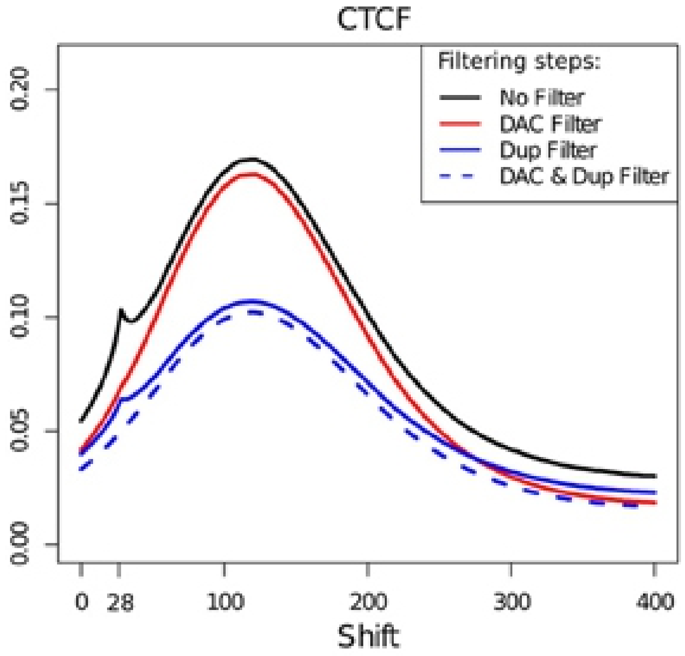
- Applied across genome.
- Expect reduction at fragment length.

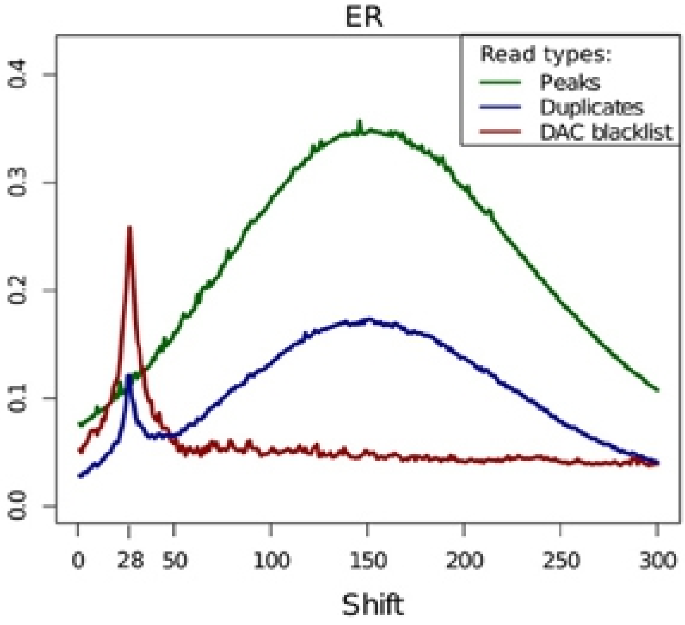


Read-length cross-coverage peak

- Blacklisted regions strongly contribute to read length cross-coverage peak

- *Rel_CC = Frag_CC/ read length cross-coverage score.*



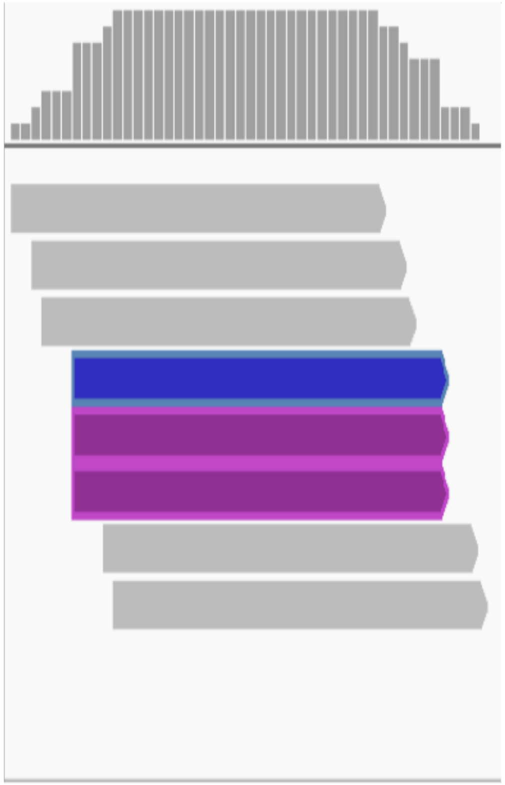


Carroll et al 2014

Duplication Rate

Duplicate FAQ

- Typically ChIP-seq is single end sequenced
  - Reads with same start position considered duplicates
- Removing duplicates saturates dynamic range of signal.
  - Maximum signal at base is 2*read length



Why worry about duplicates

- “Read duplicates arise from experimental artefacts”
  - Is true
- “All read duplicates arise from experimental artefacts”
  - Is false.

- So we need to consider that duplicates may be enriched for artefacts..
- ..but contribute to genuine ChIP-signal

Duplicates (the bad kind)

- Low starting material.
  - If initial starting material is low this can lead to overamplification of this material.
  - Biases in PCR will compound this problem.
  - Can lead to artificially enriched regions.

Duplicates (bad kind 2)

- Blacklists with ultra high signal are high in duplicates.
- Masking blacklisted regions prior to analysis removes this problem

Duplicates (The Good and Misunderstood)

- Duplicates will also exist within highly efficient (or even inefficient ChIP) when deeply sequenced ChIP.
- **Removal of duplicates can lead to a saturation and so underestimation of ChIP-signal!**

Duplicates

- Consider enrichment efficiency and sequencing depth.
- Remove duplicates prior to peak calling.
- Retain duplicates for differential binding analysis.

Practical.

- All data is /data/ChIPQC/
- Handout and R code in /data/ChIPQC/ or on Bioc2014 materials page.
- We will work through first examples.
- Few questions using what we learnt.