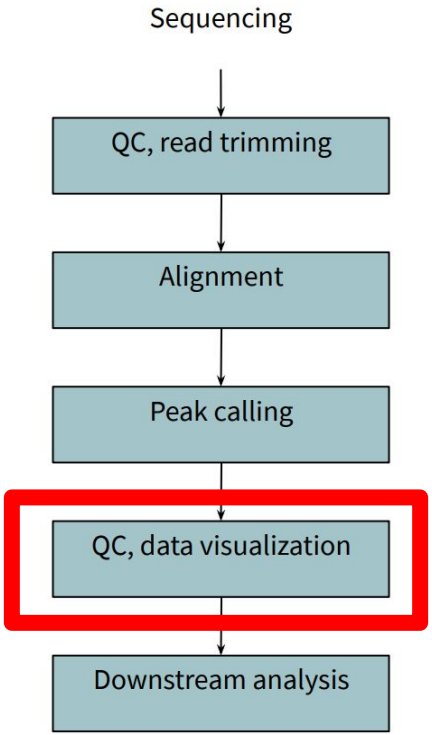
Evaluating ChIPseq Data

Shoko Hirosue
MRC Cancer Unit, University of Cambridge

CRUK CI Bioinformatics Summer School July 2020

Quality control of ChIP data



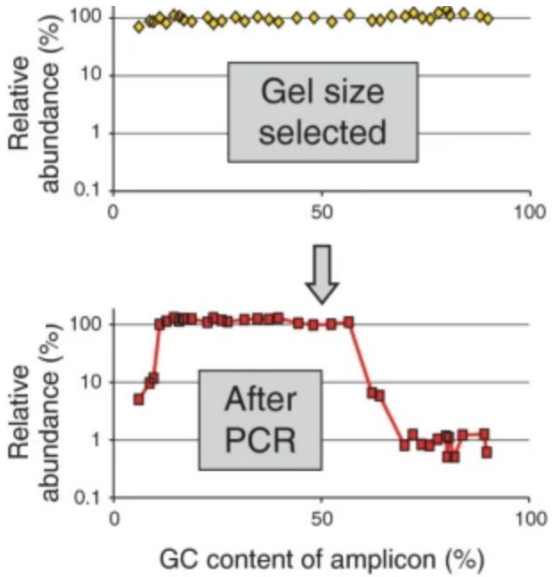
Adapted from Dora Bihary's slides

Things that could go wrong in ChIP seq experiment

- The specificity of the antibody
  - Poor reactivity against the target of the experiment
  - High cross-reactivity with other proteins

- Biases during library preparation
  - PCR amplification bias
  - Fragmentation bias

PCR Amplification bias



Aird et al. 2011, Genome Biol

Quality Control

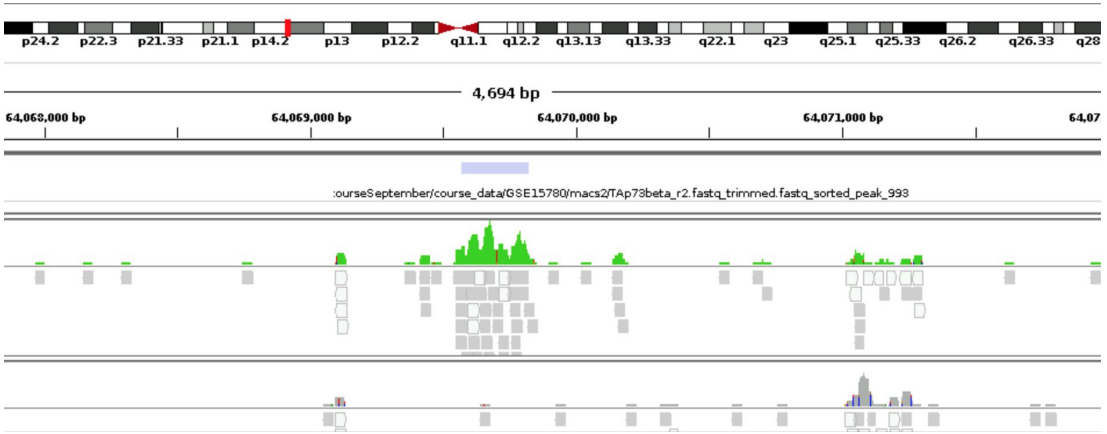
1. Browser Inspection
2. Fraction of Reads in Peaks (FRiP)
3. Uniformity of Coverage
4. Reads overlapping in Blacklisted regions (RiBL)
5. Cross-correlation analysis
6. Consistency of Replicates

1. Browser Inspection

1. Browser inspection

Using IGV or USCS genome browser

- Previously known sites
- Consistency across replicates
- Signal strength compared to input
- Accuracy of peak calls

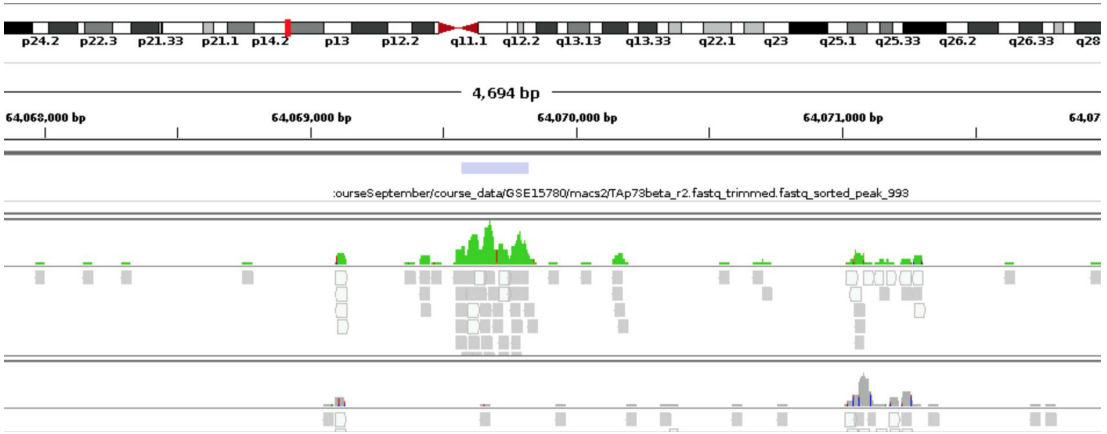


1. Browser inspection

Using IGV or USCS genome browser

- Previously known sites
- Consistency across replicates
- Signal strength compared to input
- Accuracy of peak calls

Exercise later!



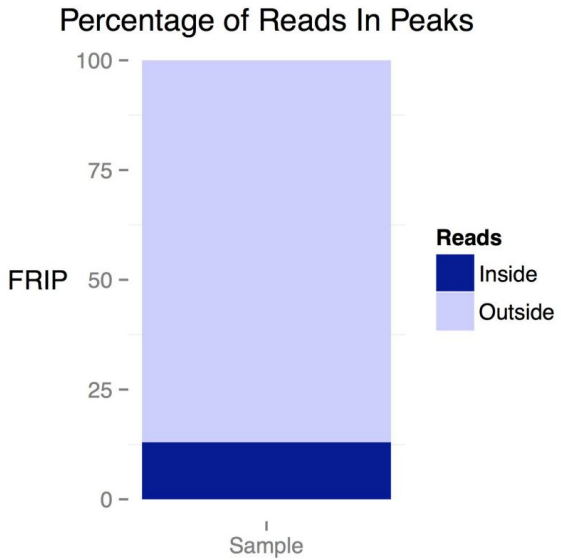
2. Fraction of Reads in Peaks

2. Measuring global ChIP enrichment (FRiP)

FRiP: **F**raction of all mapped **R**eads that fall **i**nto **P**eak regions identified by a peak-calling algorithm

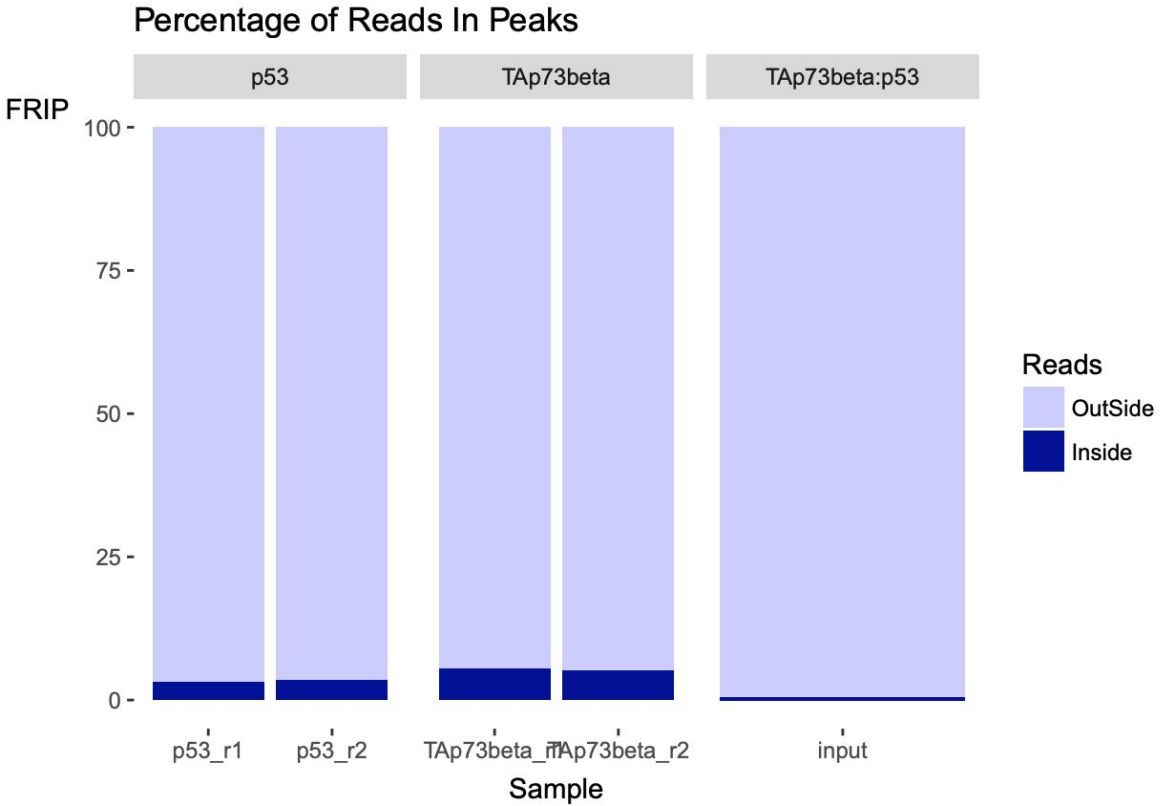
- Gives a quick understanding of the success of immunoprecipitation
- Guideline: in case of good quality FRiP is > 5%

N.B. FRiP is sensitive to the specifics of peak calling method, antibody & target factor pair, so FRiP < 1% does not automatically mean failure



Adapted from Dora Bihary's slides

What do you see in here?



Adapted from Dora Bihary's slides

3. Uniformity of Coverage

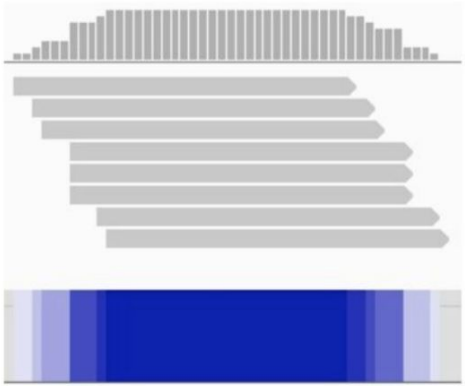
3. Uniformity of Coverage

“SSD (Standardized Standard Deviation)” : A **metric** to assess the uniformity of coverage of reads across genome

Computed by looking at the standard deviation of signal pile-up along the genome normalized to the total number of reads

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

An enriched sample typically has regions of significant pile-up so a higher SSD is more indicative of better enrichment.



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

Adapted from Dora Bihary’s slides

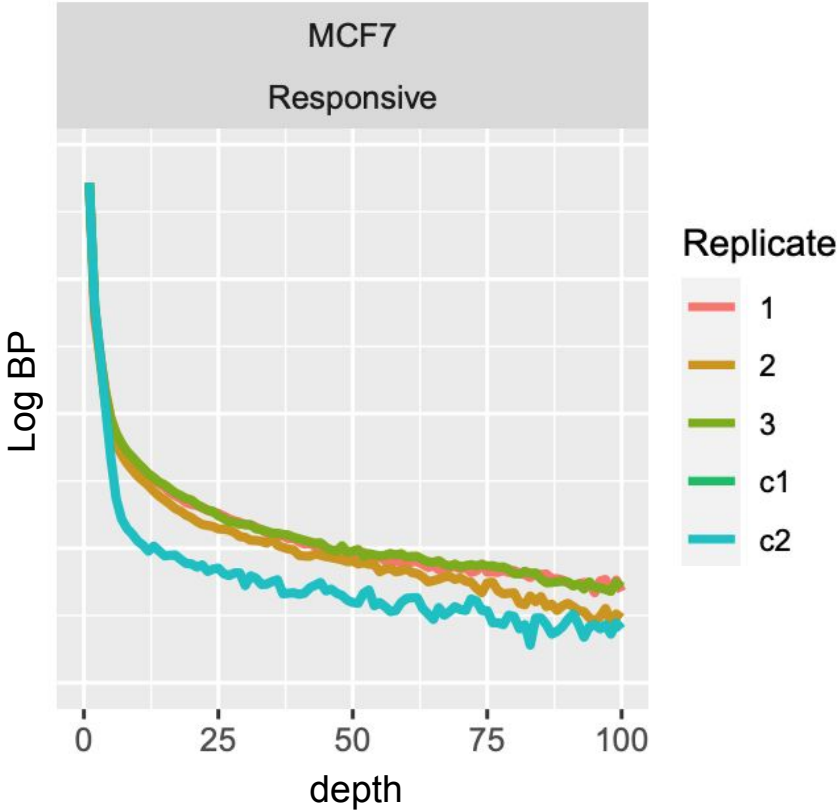
3. Uniformity of Coverage

“Coverage histogram”: visualization of coverage uniformity

X-axis (Depth): the read pileup height at a base pair position

Y-axis (log BP): Number of positions that have this pileup height in log scale

- Good enrichment: more positions (higher values on the y axis) with higher depth
- Input: Most positions in the low pile up (low x)



Documentation from bioconductor ChIPQC (https://bioconductor.riken.jp/packages/3.4/bioc/html/ChIPQC.html)
Carroll and Stark

4. Reads Overlapping in Blacklisted regions

4. Reads overlapping in Blacklisted regions (RiBL)

- BL regions: Set of regions in the genome often found at specific types of repeats such as centromeres, telomeres and satellite repeats
- BL regions show enriched signal in ChIP seq experiments regardless of what's IPed

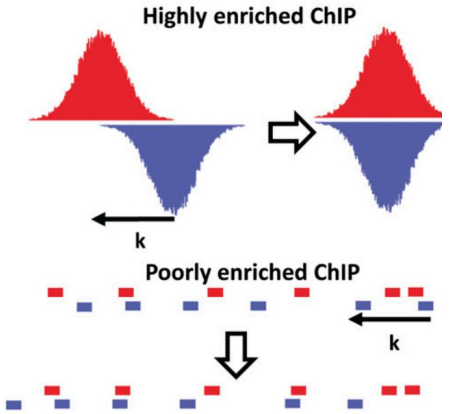
-> Leads to false positive peaks, throw off between-sample normalization!

The RiBL score acts as a guide for the level of background signal in a ChIP or input. (Lower RiBL is better)

(More about BL regions: Amemiya et al. 2019, Scientific Reports)

5. Cross-Correlation analysis

5. Cross-correlation analysis



Question: Is there a bimodal enrichment of reads?

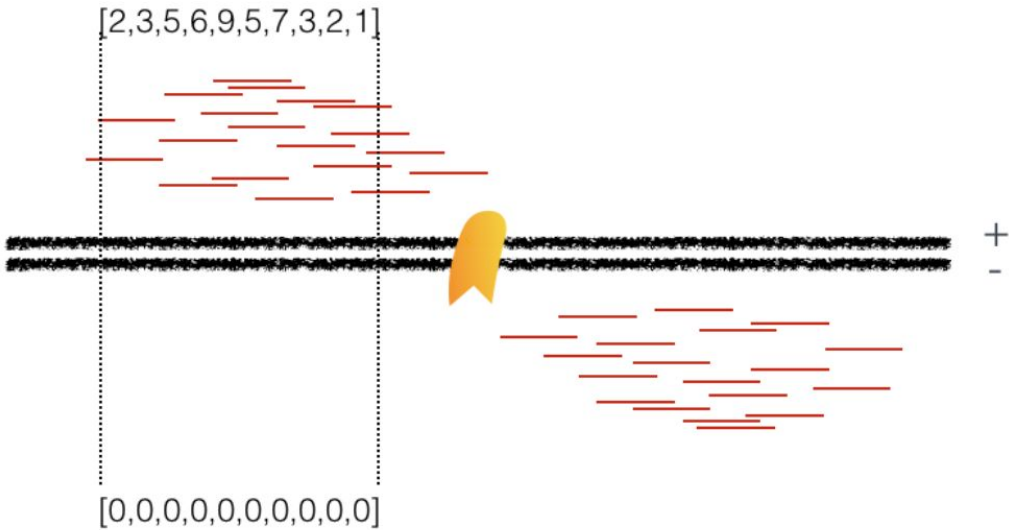
The cross-correlation metric:

- Computed as the Pearson linear correlation between the Crick strand and the Watson strand, after shifting Watson by k base pairs
- Reads are shifted in the direction of the strand they map to by an increasing number of base pairs and the Pearson correlation between the per-position read count vectors for each strand is calculated
- These Pearson correlation values are computed for every peak for each chromosome and values are multiplied by a scaling factor and then summed across all chromosomes

Landt et al., 2012, *Genome Res*

5. Cross-correlation analysis

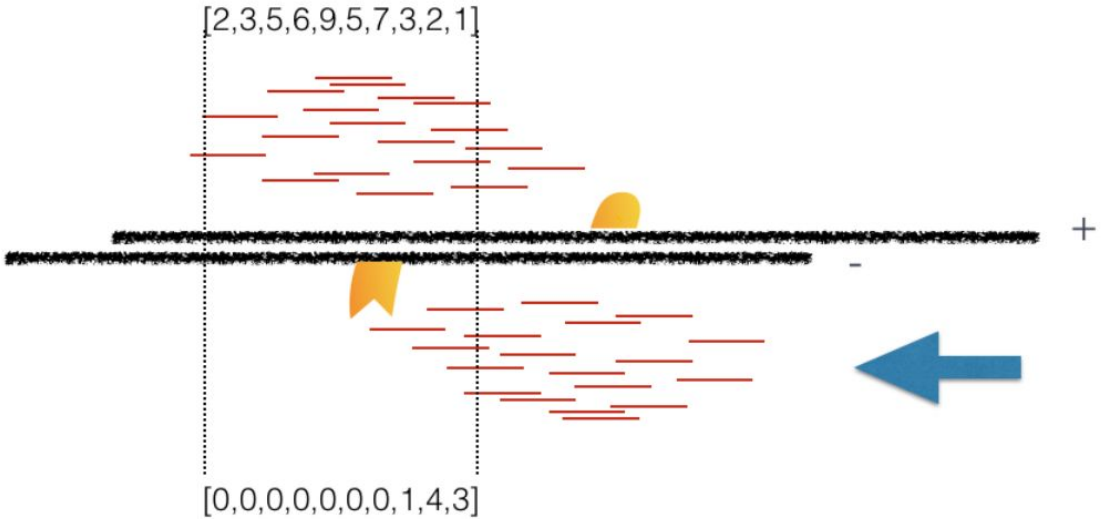
Plot 1: *At strand shift of zero, the Pearson correlation between the two vectors is 0.*



Intro to ChIPseq using HPC
Mary Piper, Meeta Mistry and Radhika Khetani
https://hbctraining.github.io/Intro-to-ChIPseq/lessons/06_combine_chipQC_and_metrics.html

5. Cross-correlation analysis

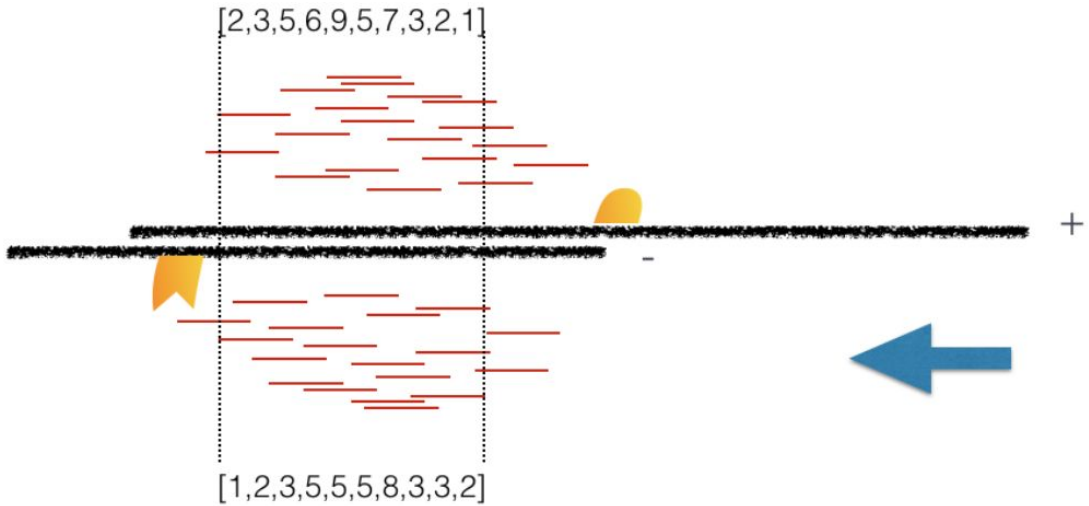
Plot 2: *At strand shift of 100bp, the Pearson correlation between the two vectors is 0.389.*



Intro to ChIPseq using HPC
Mary Piper, Meeta Mistry and Radhika Khetani
https://hbctraining.github.io/Intro-to-ChIPseq/lessons/06_combine_chipQC_and_metrics.html

5. Cross-correlation analysis

Plot 3: *At strand shift of 175bp, the Pearson correlation between the two vectors is 0.831.*



Intro to ChIPseq using HPC
Mary Piper, Meeta Mistry and Radhika Khetani
https://hbctraining.github.io/Intro-to-ChIPseq/lessons/06_combine_chipQC_and_metrics.html

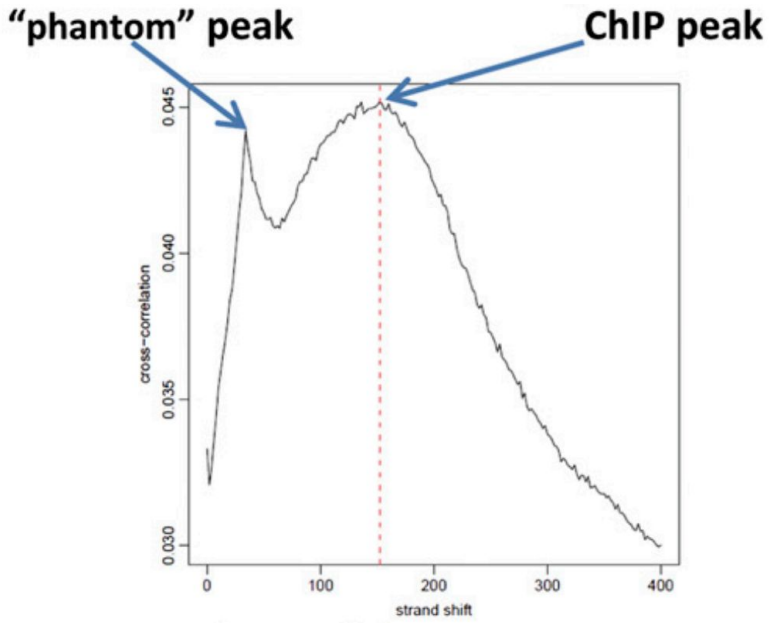
5. Cross-correlation analysis

CC (Cross-correlation): y axis. correlation of reads on positive and negative strand after successive read shifts

Once the final cross-correlation values have been calculated, they can be plotted (Y-axis) against the shift value (X-axis) to generate a cross-correlation plot

The cross-correlation plot typically produces two peaks:

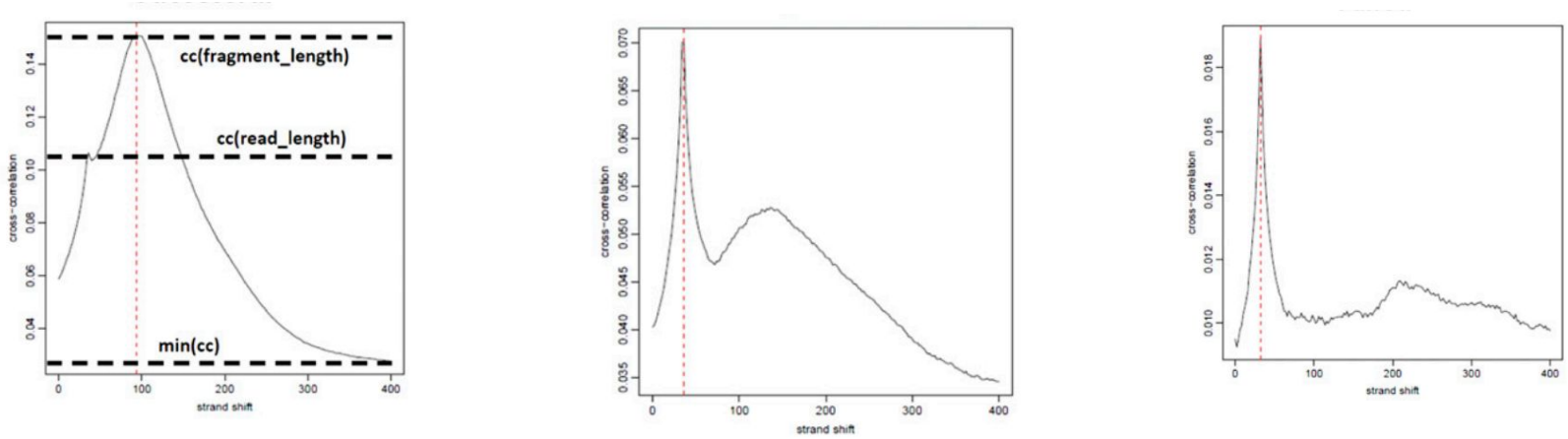
- a peak of enrichment corresponding to the predominant fragment length (high correlation value)
- peak corresponding to the read length (“phantom” peak)



Landt et al., 2012, *Genome Res*

5. Cross-correlation analysis

CC (Cross-correlation): y axis. correlation of reads on positive and negative strand after successive read shifts



Strong signal ⟷ No signal

Metrics computed in ChIPQC

- RelCC, RSC (Relative strand cross-correlation coefficient) (>1 for all samples: good signal to noise)

$$RSC = \frac{cc(fragment\ length) - min(cc)}{cc(read\ length) - min(cc)}$$

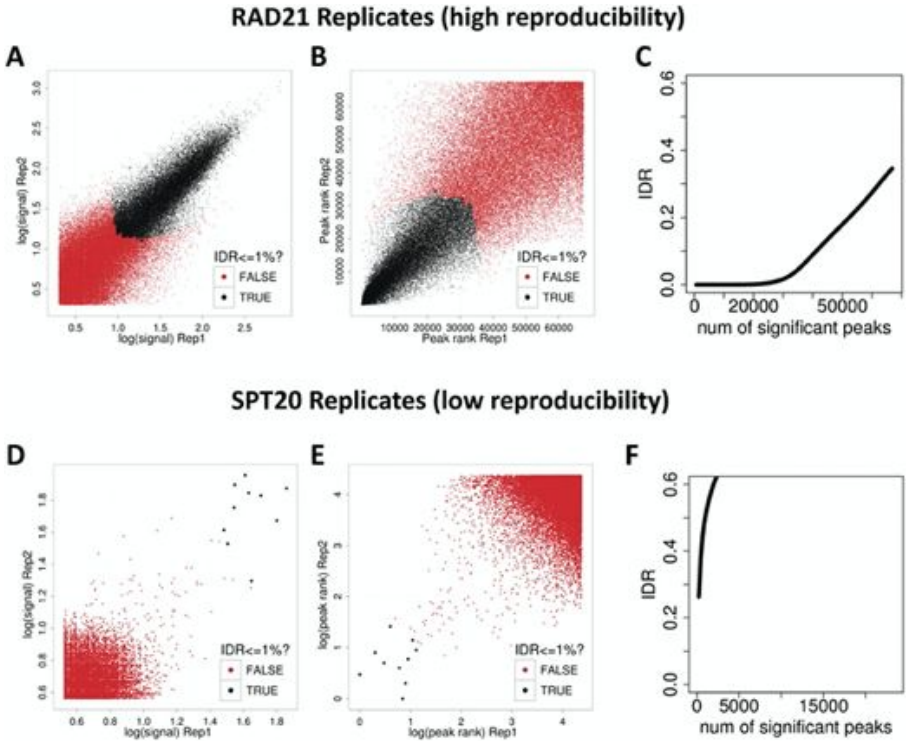
Landt et al., 2012, *Genome Res*

6. Consistency across Replicates

6. Consistency across Replicates

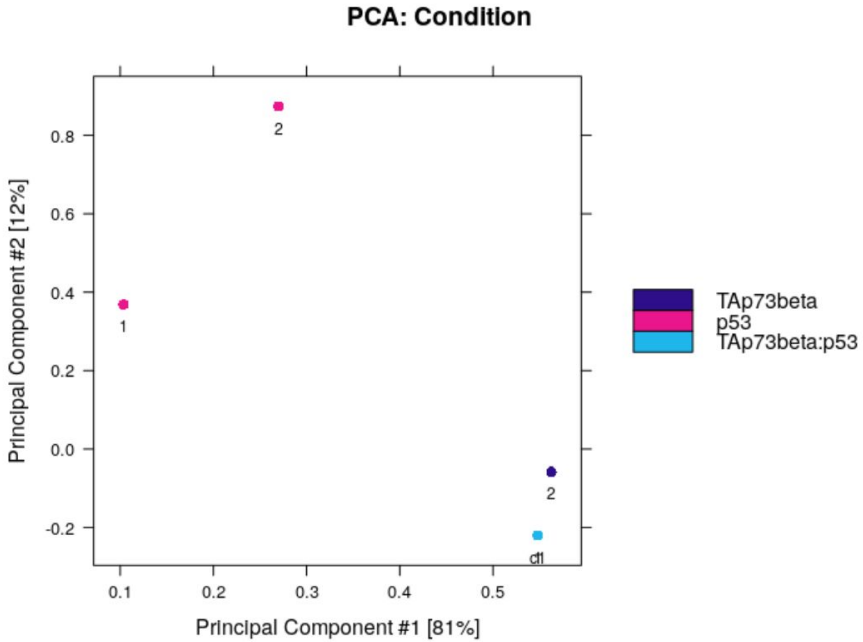
IDR (Irreproducible Discovery Rate)

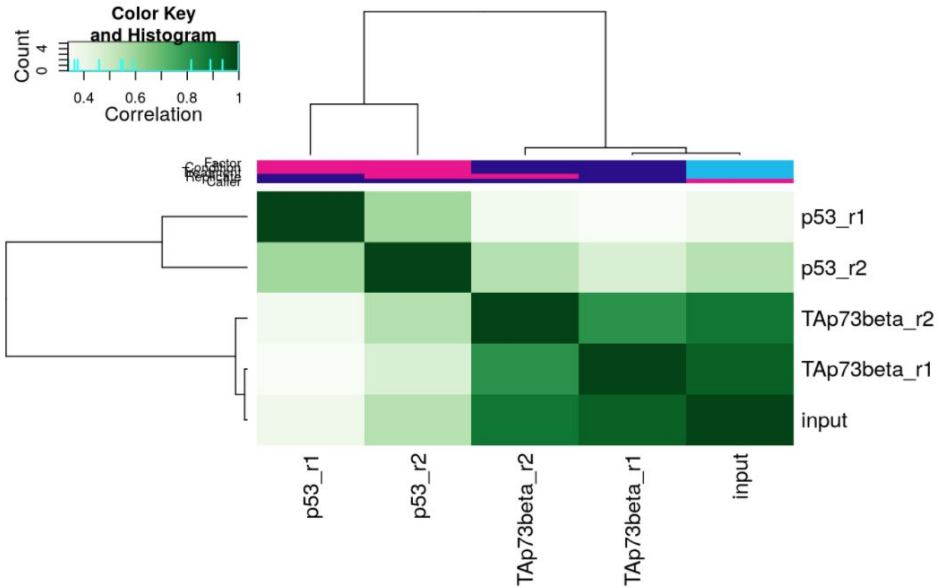
- Rank peaks from a pair of replicate datasets (eg. qvalue, FC)



Landt et al., 2012, *Genome Res.*
Li et al. 2011. *Ann Appl Stat*

6. Consistency across Replicates



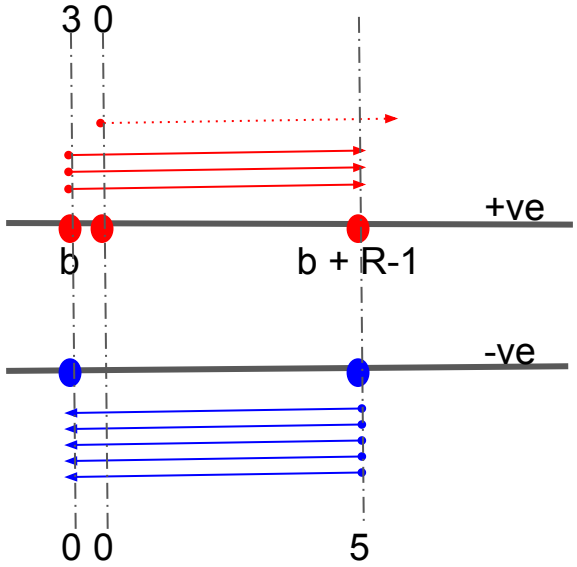


Adapted from Dora Bihary's slides

Questions?

Supplementary

5. Cross-correlation analysis



Why is there a phantom peak?

Phantom peaks: unavoidable artefact caused by “mappability”

If the sequence of R nucleotides beginning at position b occurs nowhere else in the genome: position b is mappable

the R-mer beginning at position $b+1$ matches exactly the R-mer beginning at one or more other positions in the genome: position $b+1$ is unmappable

Ramachandran et al. 2013, Bioinformatics

References

- CRUK summer school 2018 materials (https://bioinformatics-core-shared-training.github.io/cruk-summer-school-2018/)
- CRUK summer school 2019 materials (https://bioinformatics-core-shared-training.github.io/cruk-summer-school-2019/)
- Carroll et al. 2014.,Frontiers in Genetics. “Impact of artefact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data.”
- Landt et al., 2012, *Genome Res.* “ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia”
- Intro to ChIPseq using HPC, Mary Piper, Meeta Mistry and Radhika Khetani (https://hbctraining.github.io/Intro-to-ChIPseq/lessons/06_combine_chipQC_and_metrics.html)
- Li Q, Brown J, Huang H, Bickel P. 2011.” Measuring reproducibility of high-throughput experiments.” *Ann Appl Stat*
- Ramachandran et al. 2013, Bioinformatics. “MaSC: mappability-sensitive cross-correlation for estimating mean fragment length of single-end short-read sequencing data”

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.