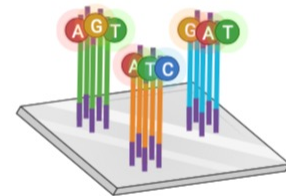


# Quality control – ChIP–Seq Data

Junfan Huang

MRC Cancer Unit  
University of Cambridge

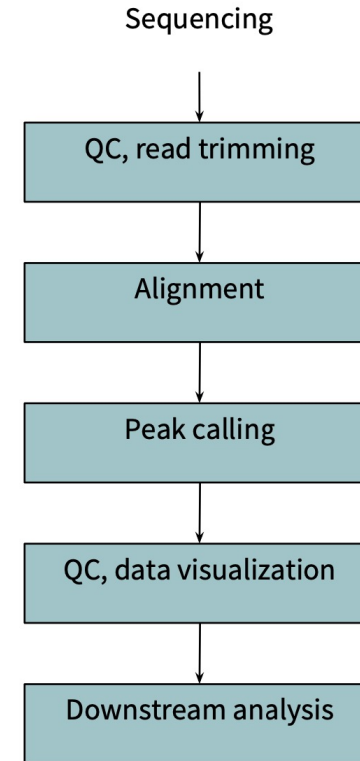
CRUK Bioinformatics Summer School 2021  
27th July 2021



# Quality control – ChIP-Seq data

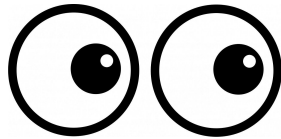
Visually

Computationally

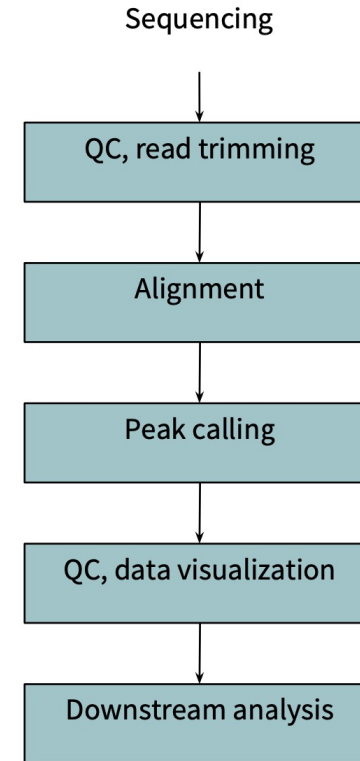


# Quality control – ChIP-Seq data

Visually

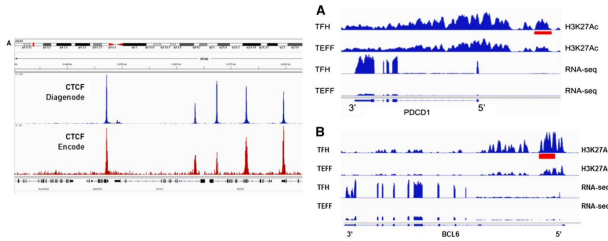


Computationally

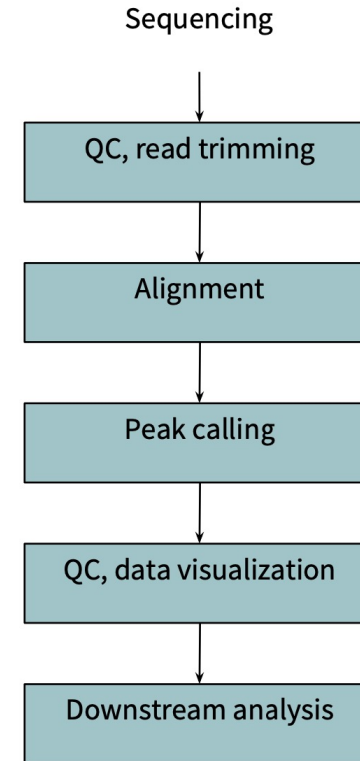


# Quality control – ChIP–Seq data

Visually (IGV or USCS genome browser )

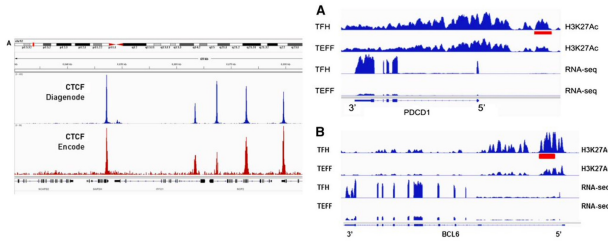


Computationally

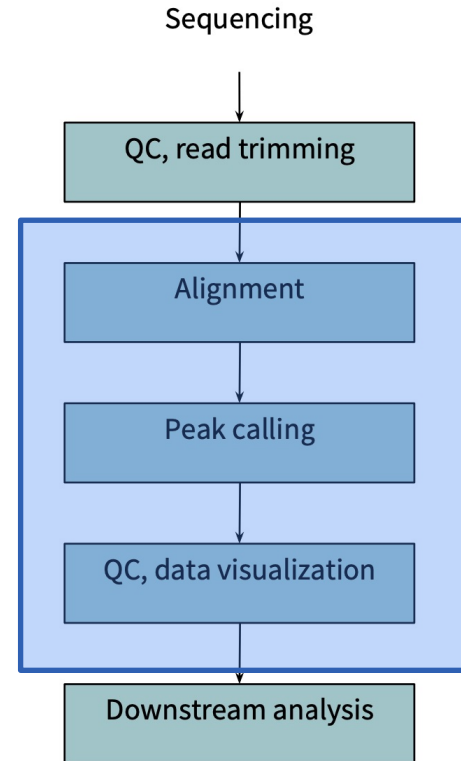


# Quality control – ChIP-Seq data

Visually



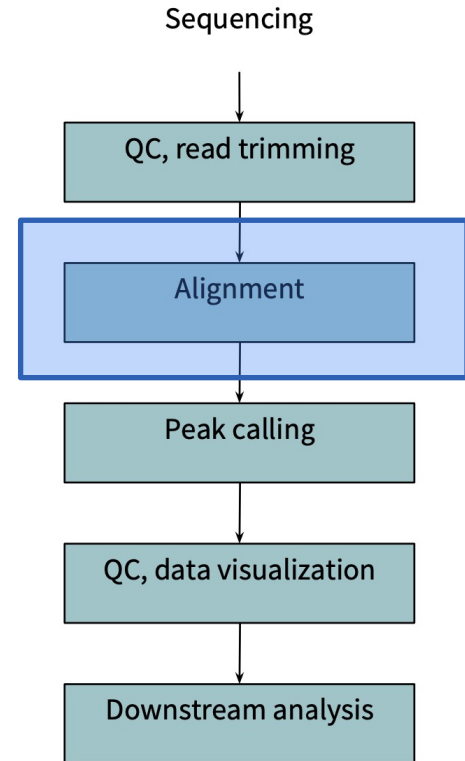
Computationally



# Quality control – ChIP-seq data

## Visually (Alignment)

- Relative Enrichment in genomic intervals (REGI)

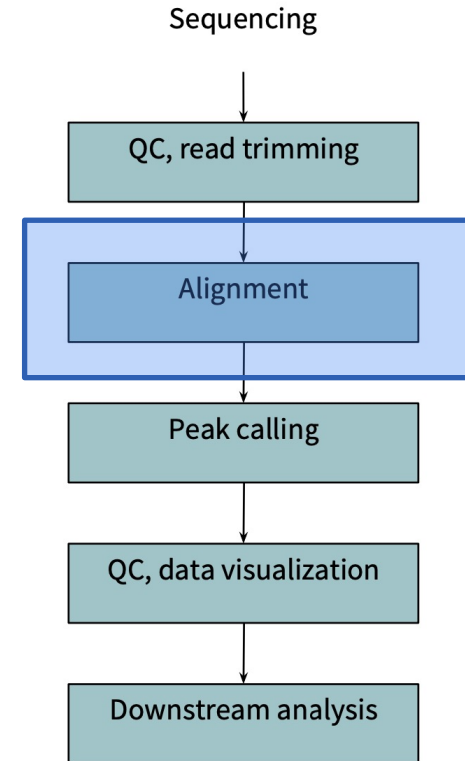


Computationally

# Quality control – ChIP-seq data

## Visually (Alignment)

- Relative Enrichment in genomic intervals (REGI)
  - Proteins might have a high expected enrichment in certain genomic regions



# Quality control – ChIP-seq data



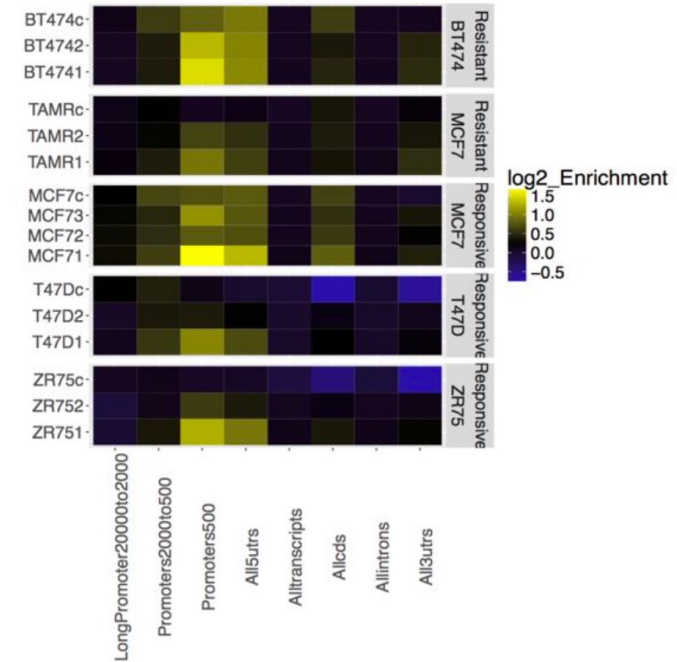
## Visually (Alignment)

- Relative Enrichment in genomic intervals (REGI)
  - Proteins might have a high expected enrichment in certain genomic regions

Promoter region

UTRs

introns



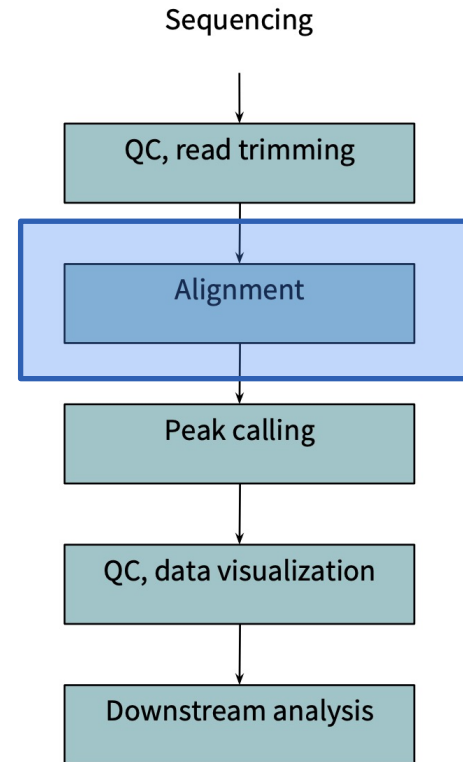


# Quality control – ALL NGS data

Visually

Computationally

- Read Mapping% (**Higher the better**)

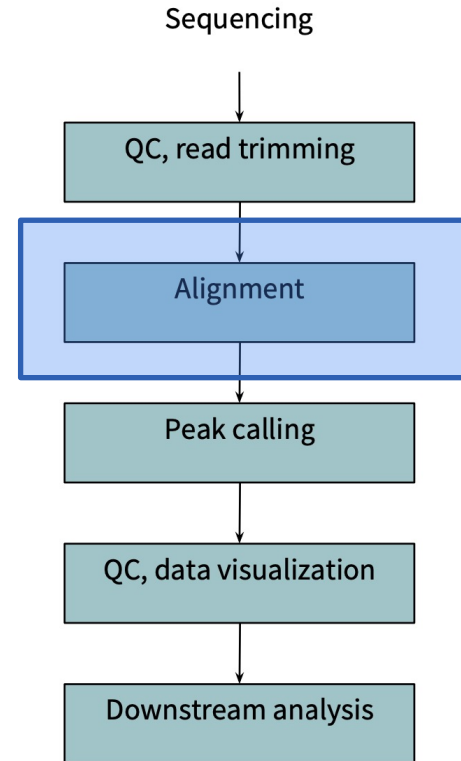


# Quality control – ChIP-seq data

Visually

Computationally (Alignment)

- Remove Blacklisted regions
- Strand cross-correlation
- PCR Bottleneck coefficient (PBC)

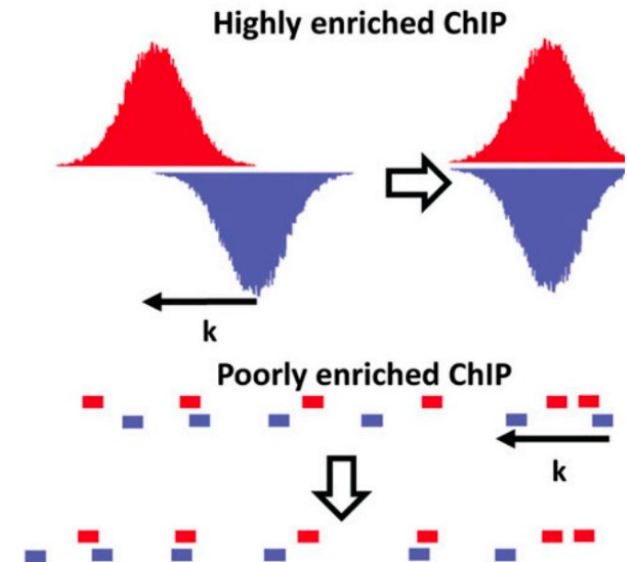


# Quality control – ChIP-seq data

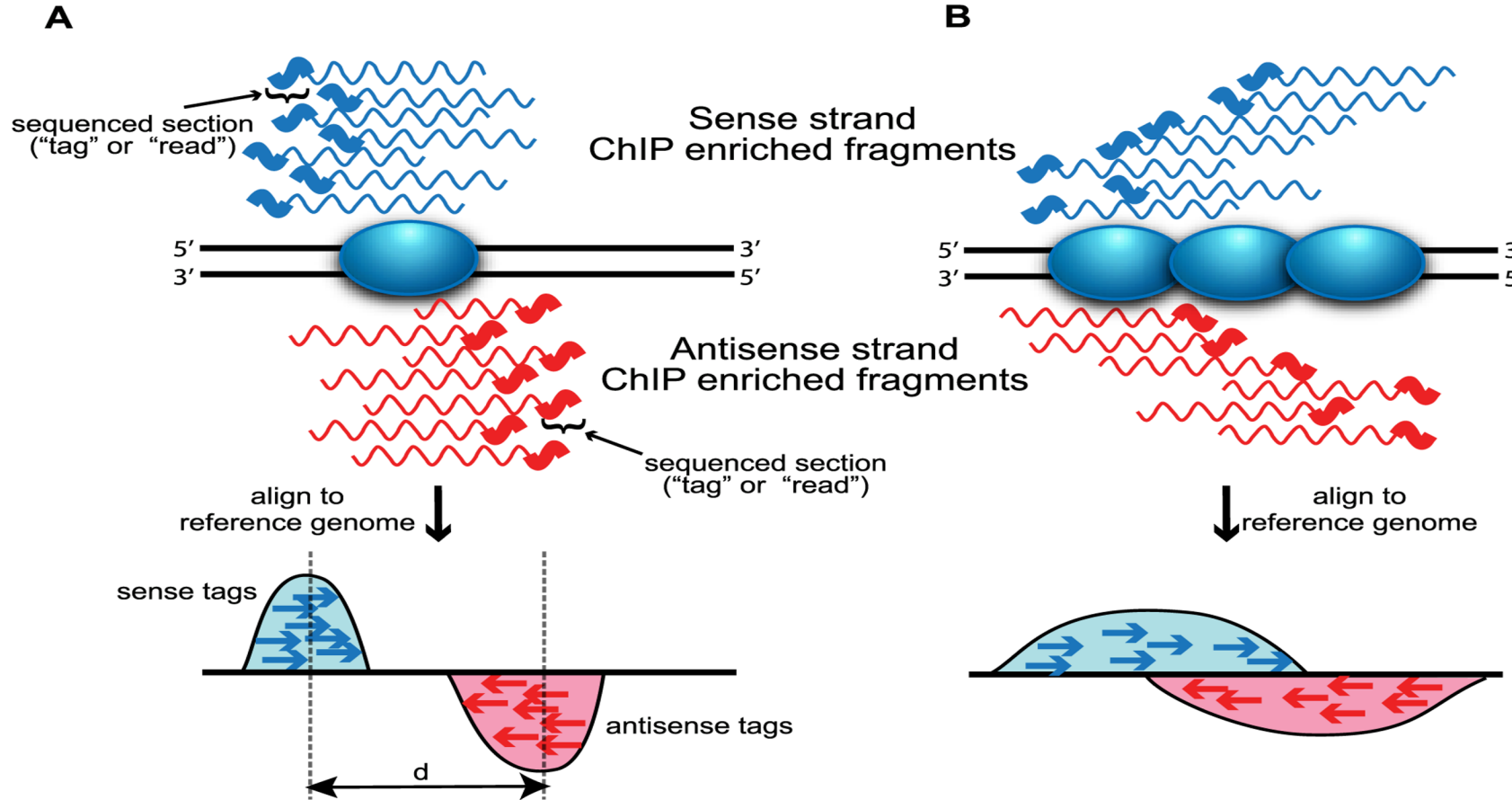
Visually

Computationally (Alignment)

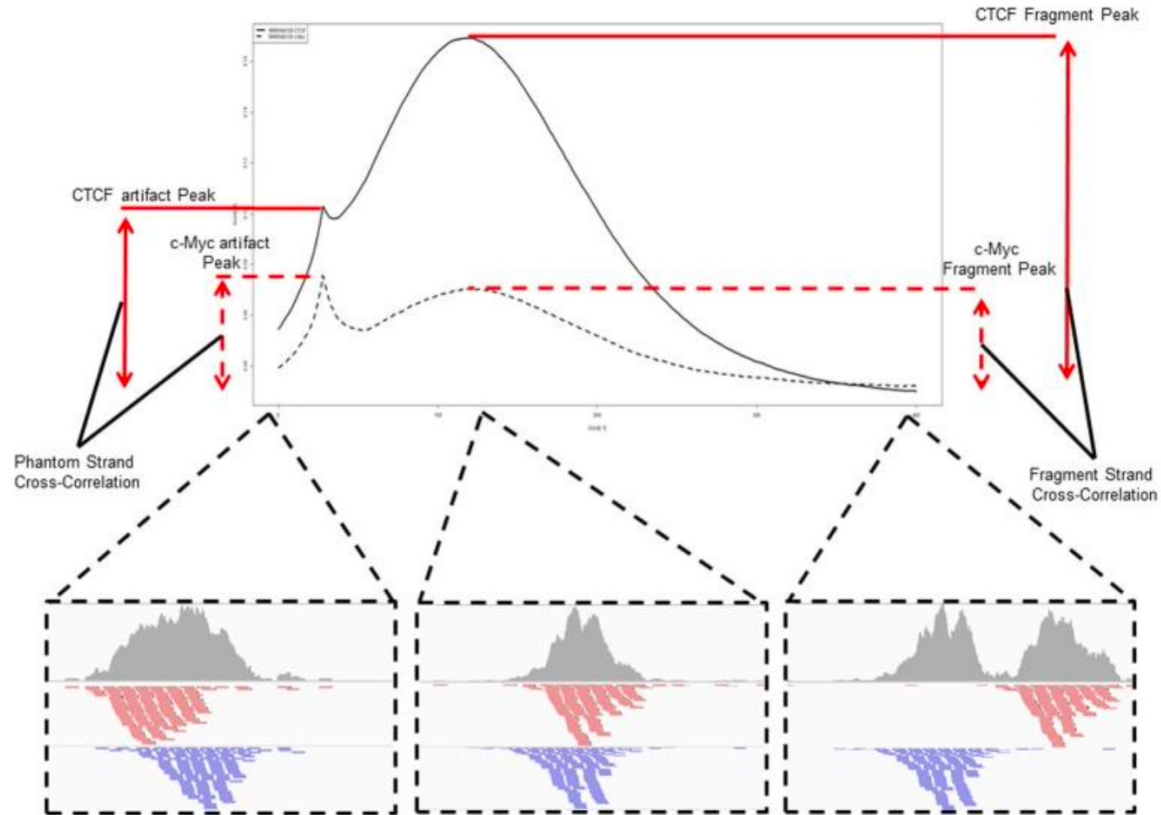
- Remove Blacklisted regions
- **Strand cross-correlation**
- PCR Bottleneck coefficient (PBC)



# Strand dependent bimodality



# Strand cross-correlation

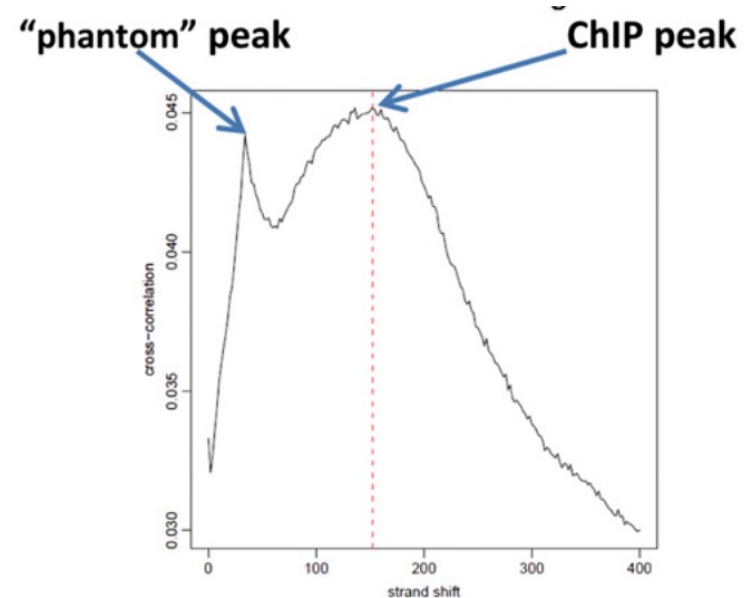


# Strand cross-correlation plot



The cross-correlation plot typically produces two peaks:

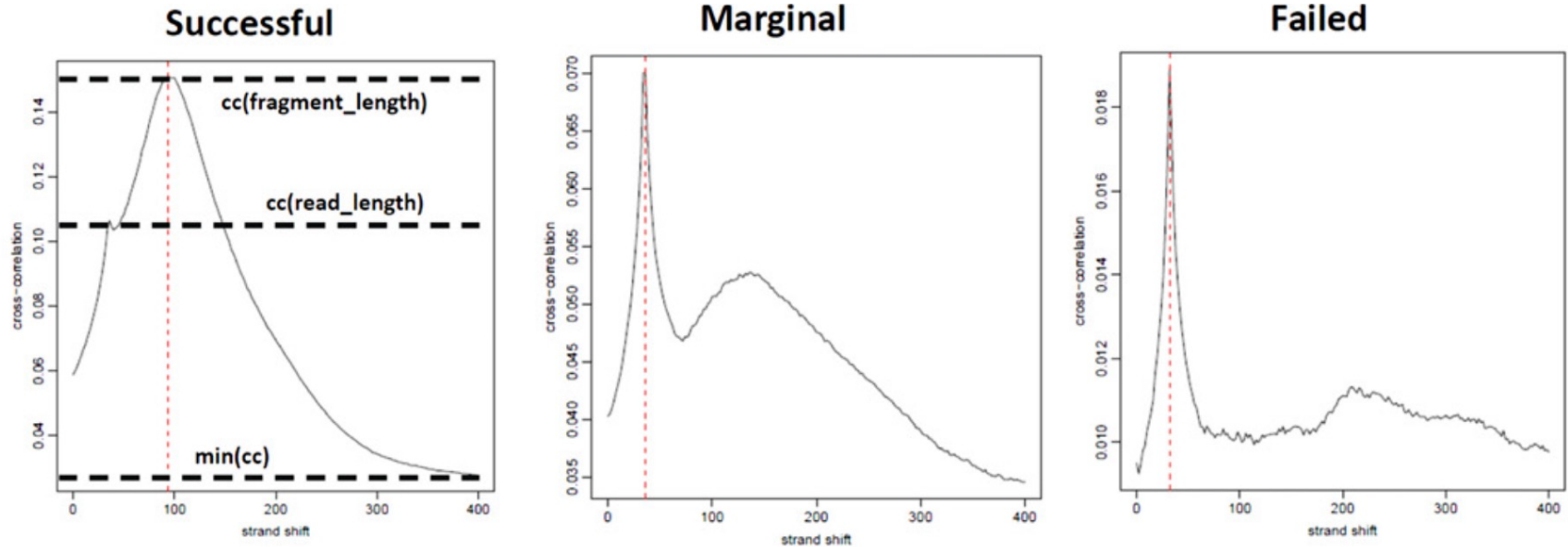
- ChIP peak - fragment length
- Phantom peak - read length



# Strand cross-correlation plot



G

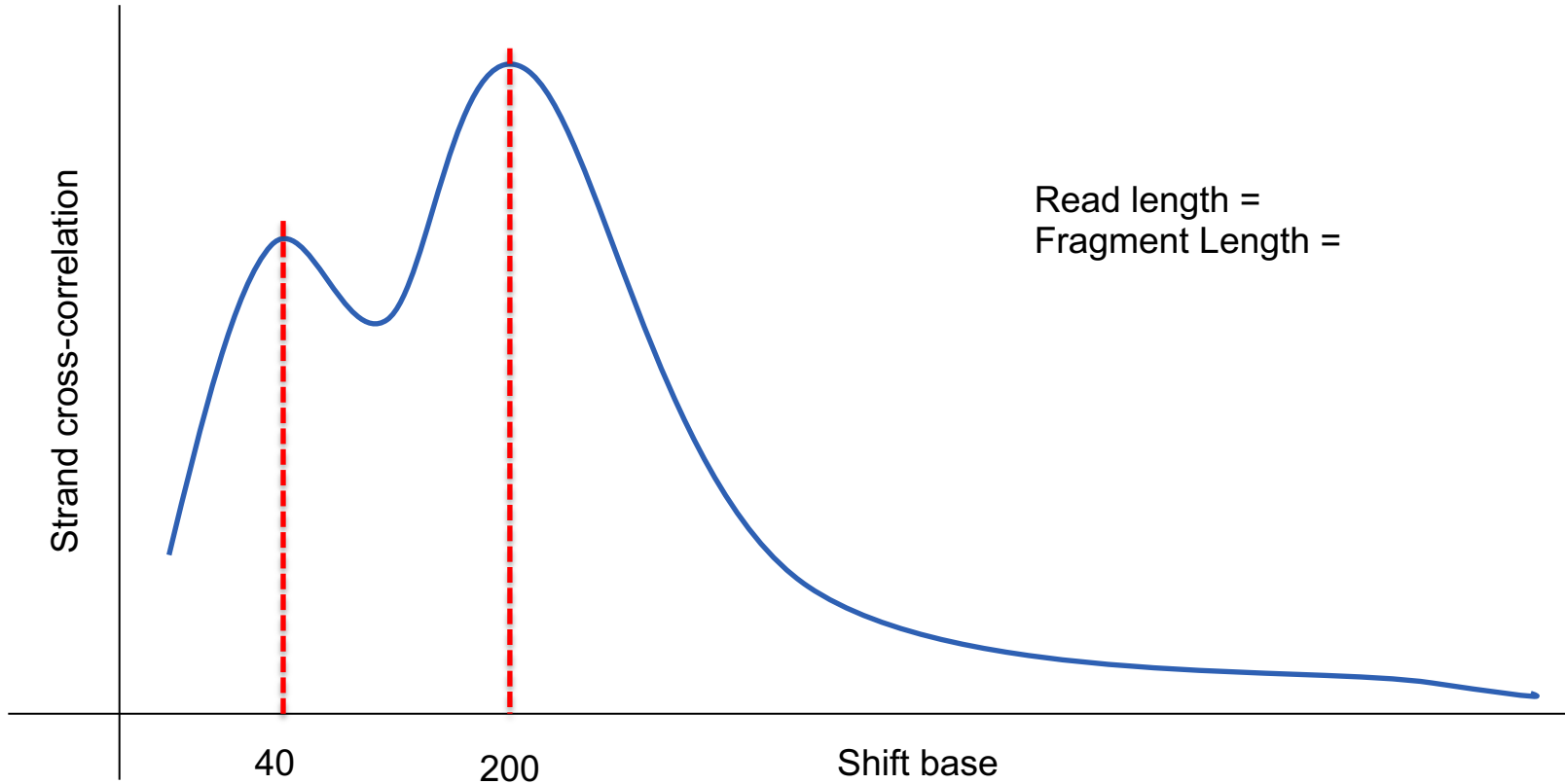


$$NSC = \frac{cc(\text{ fragment length })}{\min(cc)}$$

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

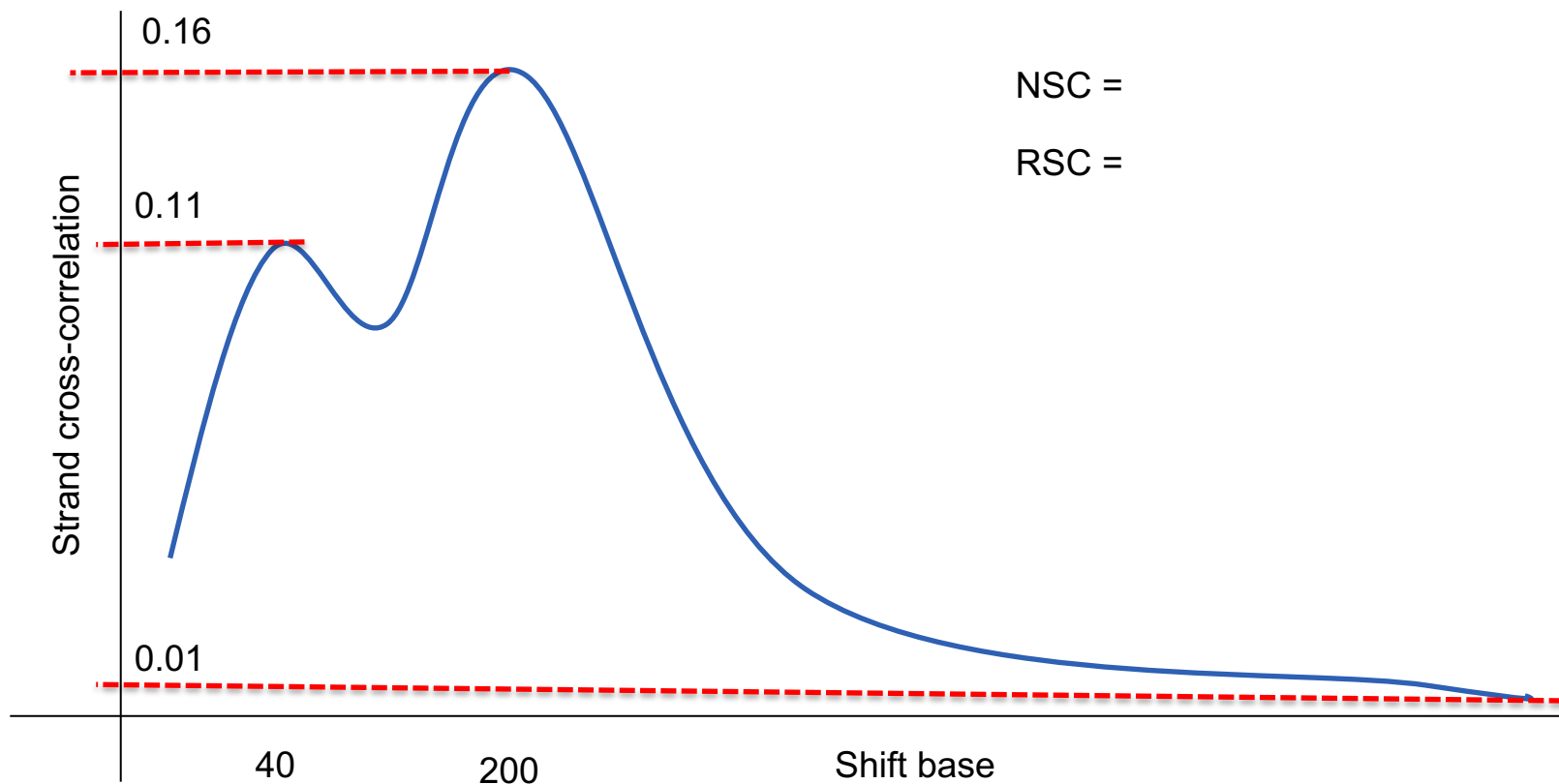
Very successful ChIP experiments generally have  $NSC > 1.05$  and  $RSC > 0.8$

# Strand cross-correlation plot



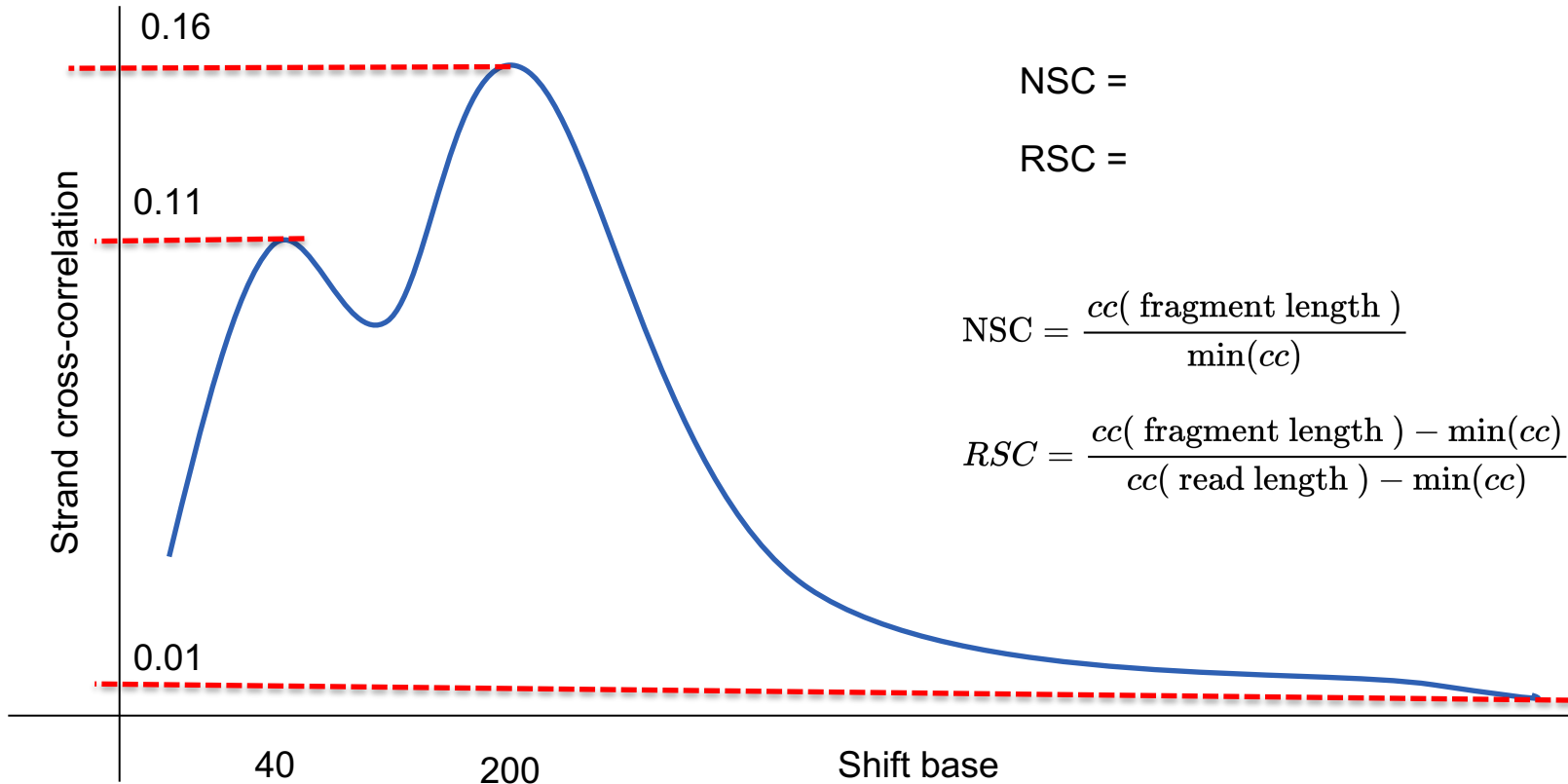


# Strand cross-correlation plot



Very successful ChIP experiments generally have  $NSC > 1.05$  and  $RSC > 0.8$

# Strand cross-correlation plot



Very successful ChIP experiments generally have  $NSC > 1.05$  and  $RSC > 0.8$

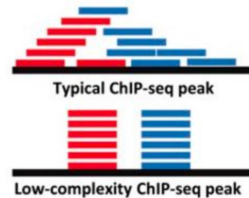
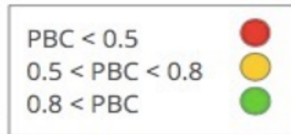
# Quality control – ChIP-seq data

Visually

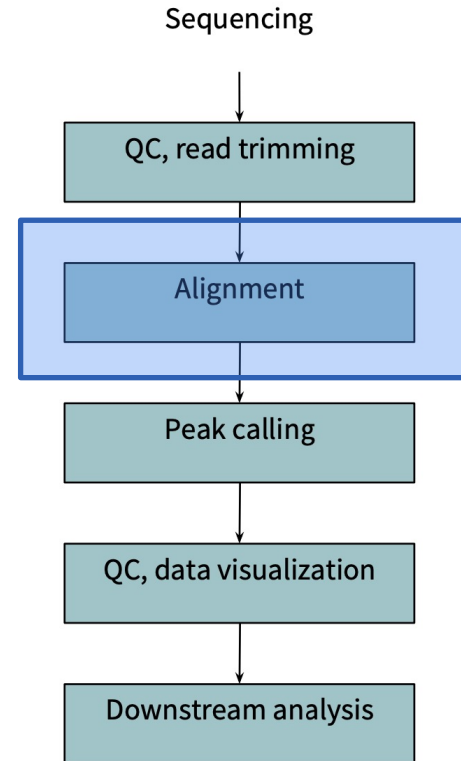
Computationally (Alignment)

- Remove Blacklisted regions
- Strand cross-correlation
- PCR Bottleneck coefficient (PBC)

$$PBC = \frac{N_1}{N_2}$$



N1: # genomic positions with one read aligned (**higher**)  
N2: # genomic positions with one or more reads

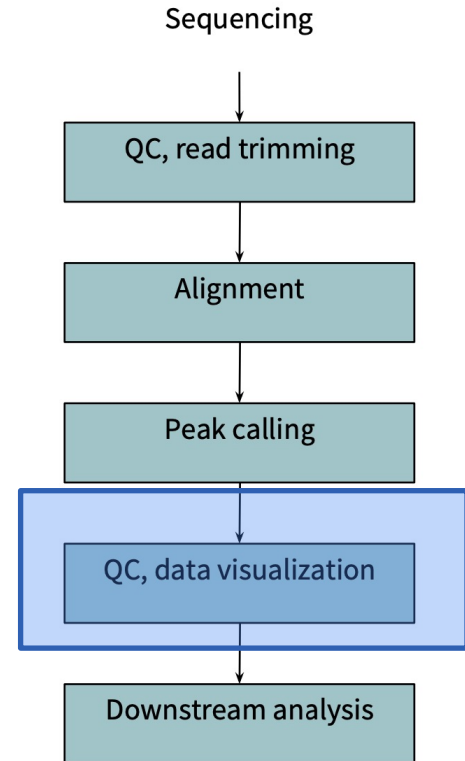


# Quality control – ChIP-seq data

Visually

Computationally (peak)

- Num of peaks with good FDR( $<0.05$ )

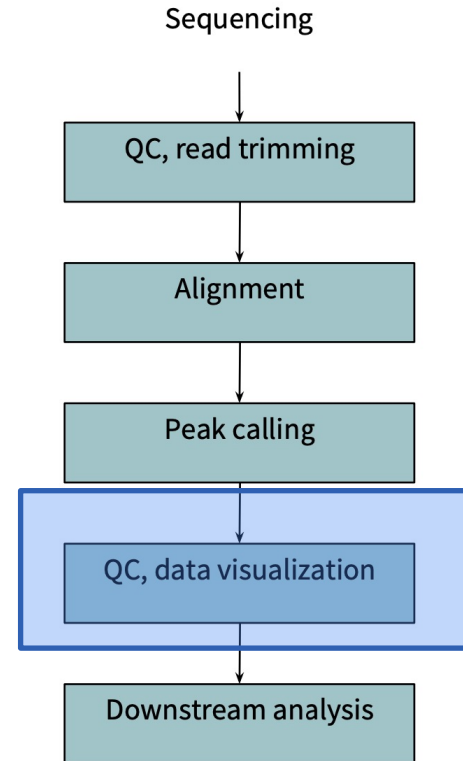


# Quality control – ChIP-seq data

Visually

Computationally (peak)

- Num of peaks with good FDR( $<0.05$ )
- **Fold Change**

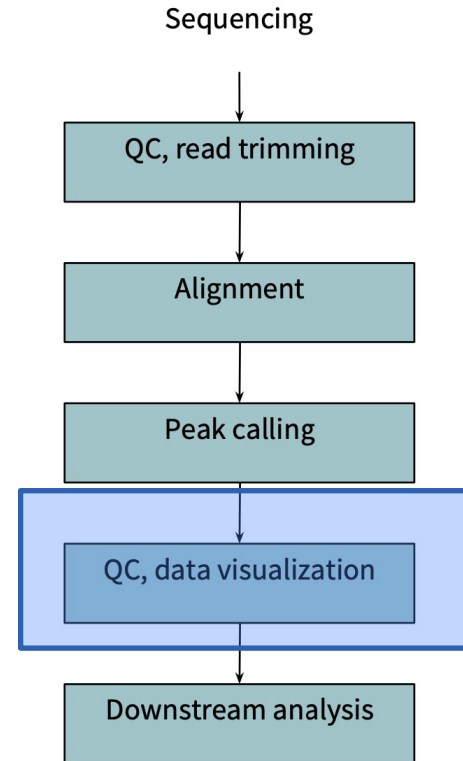
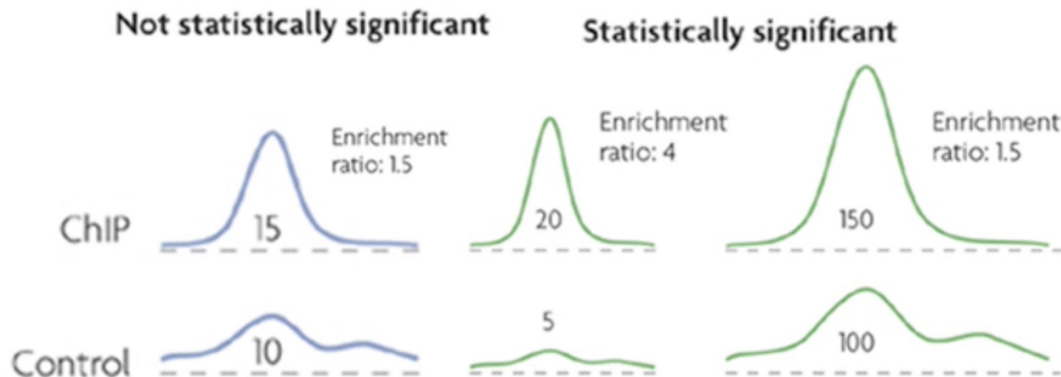


# Quality control – ChIP-seq data

Visually

Computationally (peak)

- Num of peaks with good FDR( $<0.05$ )
- **Fold Change (enrichment ratio)**



# Quality control – ChIP-seq data

Visually

Computationally (peak)

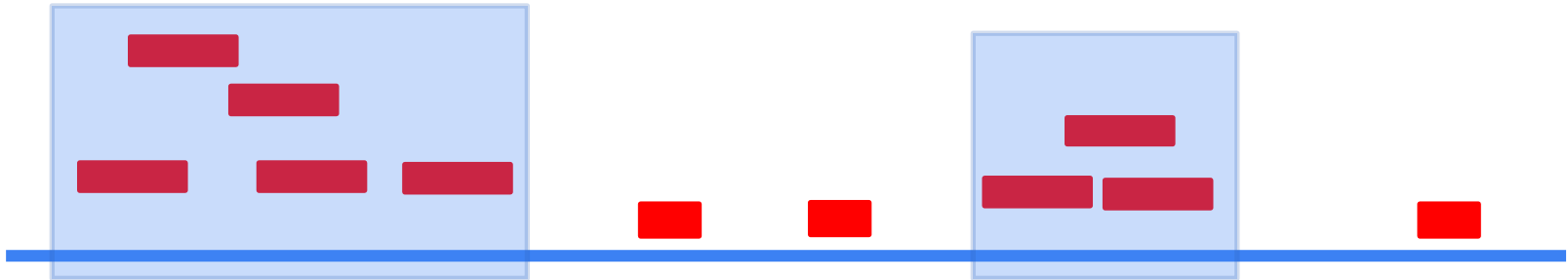
- Num of peaks with good FDR (0.05)
- Fold Change
- **Fraction of Reads in Peaks (FRiP>5%)**

# Quality control – ChIP-seq data

Visually

Computationally (peak)

- Num of peaks with good FDR (0.05)
- Fold Change
- Fraction of Reads in Peaks (FRiP>5%)



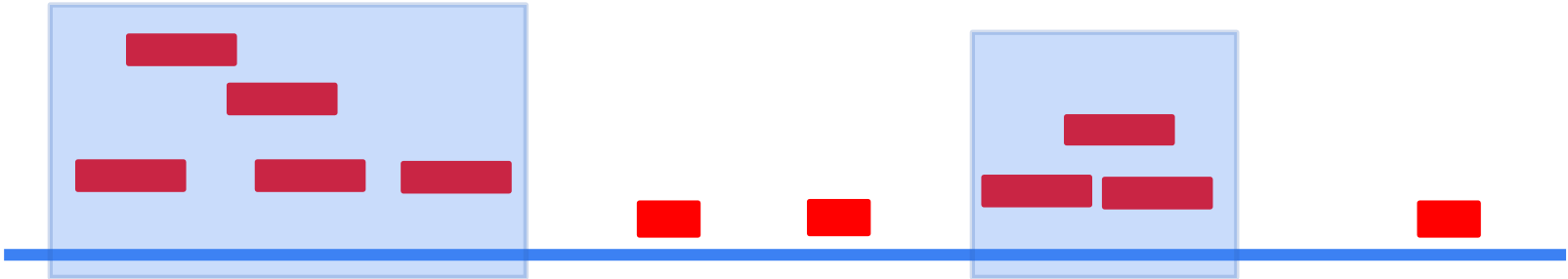


# Quality control – ChIP-seq data

Visually

Computationally (peak)

- Num of peaks with good FDR (0.05)
- Fold Change
- Fraction of Reads in Peaks (FRiP>5%) 
$$\text{FRiP} = \frac{\text{reads} \in \text{peaks}}{\text{total reads}}$$

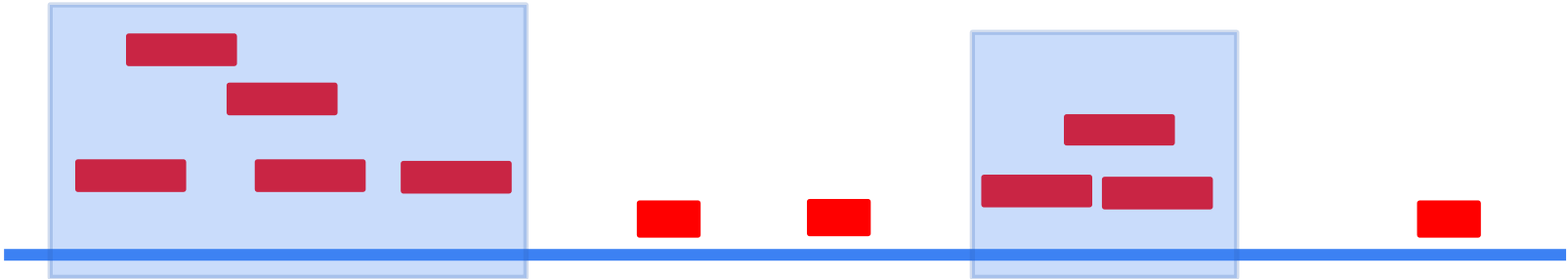


# Quality control – ChIP-seq data

Visually

Computationally (peak)

- Num of peaks with good FDR (0.05)
- Fold Change
- Fraction of Reads in Peaks (FRiP > 5%) 
$$\text{FRiP} = \frac{\text{reads} \in \text{peaks}}{\text{total reads}}$$



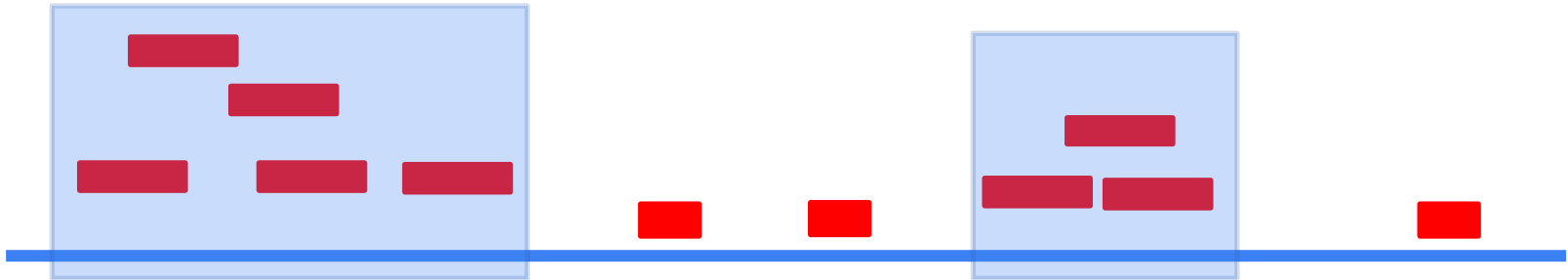
N.B. FRiP is sensitive to the specifics of peak calling method, antibody & target factor pair, so FRiP < 1% does not automatically mean failure

# Quality control – ChIP-seq data

Visually

Computationally (peak)

- Num of peaks with good FDR (0.05)
- Fold Change
- **Fraction of Reads in Peaks (FRiP>5%)** 
$$\text{FRiP} = \frac{\text{reads} \in \text{peaks}}{\text{total reads}}$$

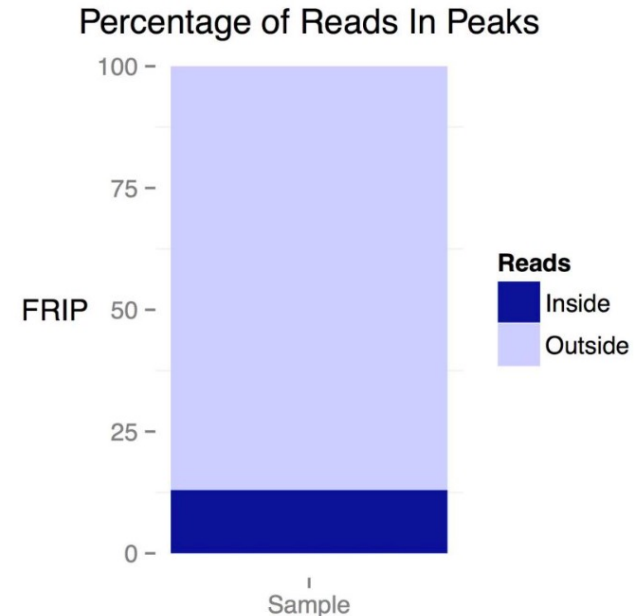


# Quality control – ChIP-seq data

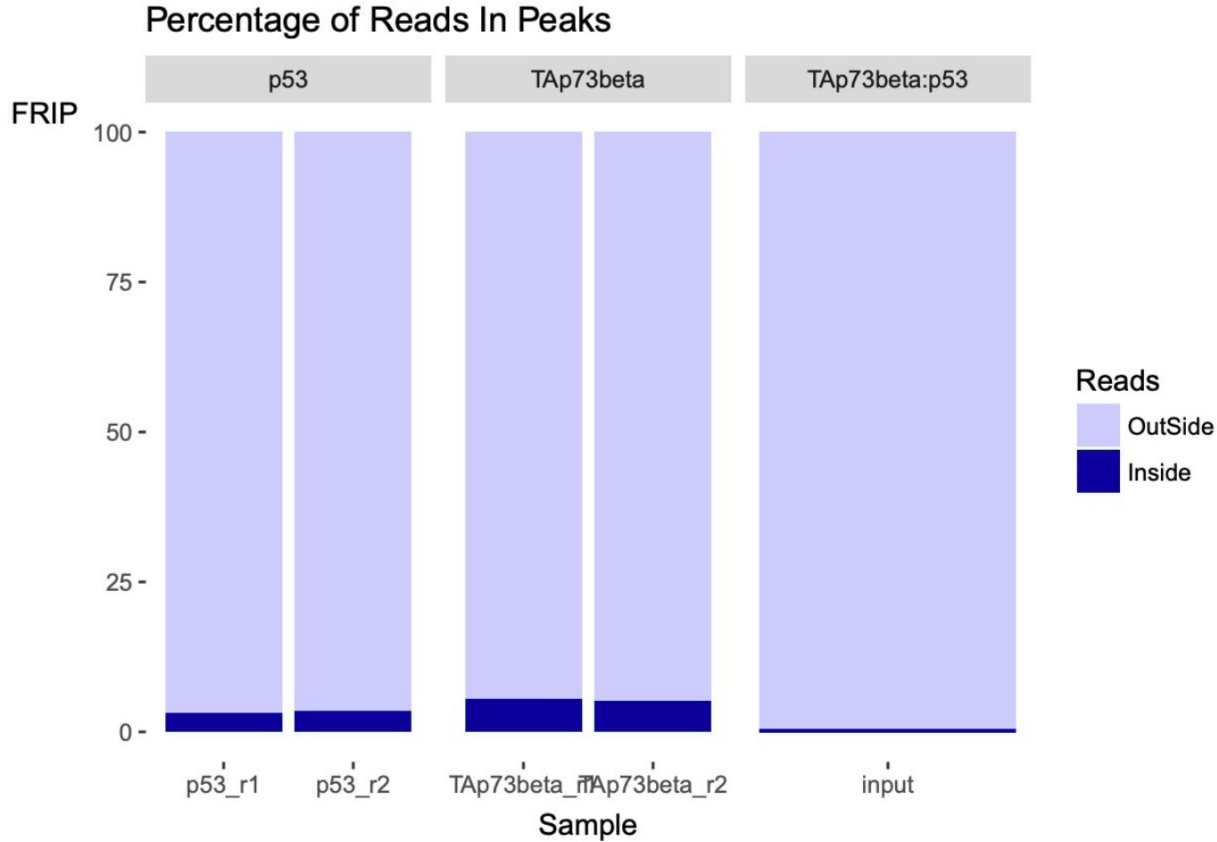
Visually

Computationally (peak)

- Num of peaks with good FDR (0.05)
- Fold Change
- **Fraction of Reads in Peaks (FRiP > 5%)**



# What do you see here?



Adapted from Dora Bihary's slides

# Quality control – ChIP-seq data

Visually

Computationally (peak)

- Num of peaks with good FDR (0.05)
- Fold Change
- Fraction of Reads in Peaks (FRiP>5%)
- **Irreproducible discovery rate (IDR - replicates)**

# Irreproducible discovery rate (IDR)

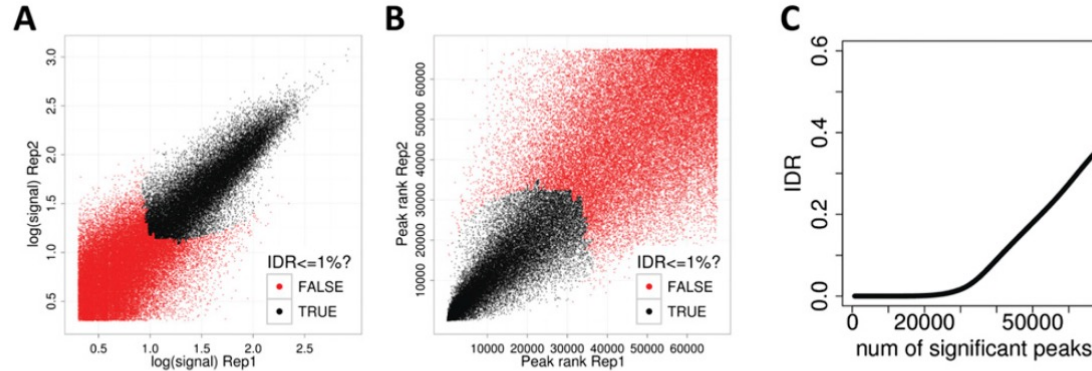


**Useful for determining reproducibility of peaks in replicates**

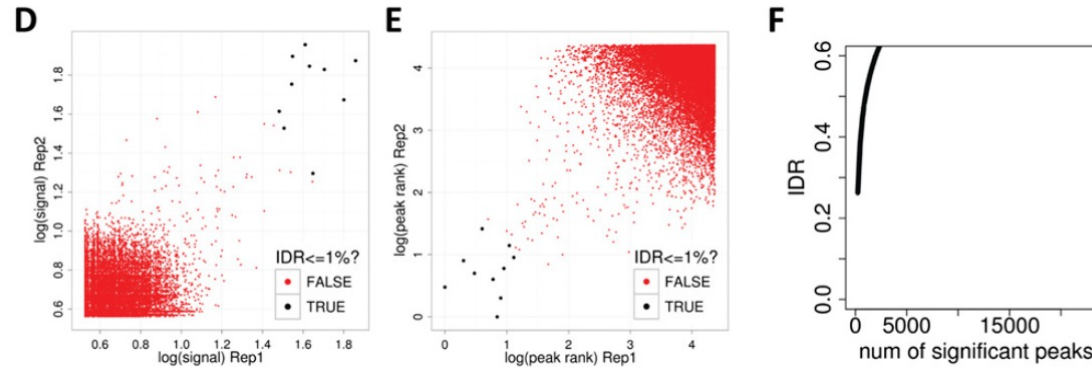
Tools: <https://github.com/kundajelab/idr>

# IDR (lower the better)

## RAD21 Replicates (high reproducibility)



## SPT20 Replicates (low reproducibility)





# Quality control – ChIP-seq data

Visually

Computationally (peak)

- Num of peaks with good FDR (0.05)
- Fold Change
- Fraction of Reads in Peaks (FRiP>5%)
- Irreproducible discovery rate (IDR – replicates)
- **Standardised standard deviation (SSD)**

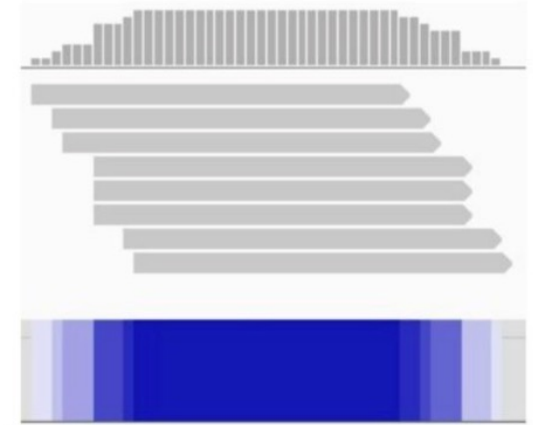
# Dispersion of coverage

The depth of coverage:

- The number of fragments at a specific genomic region

Expectation:

The depth to have large diversity in an enriched ChIP dataset !



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

# Measure the dispersion of coverage

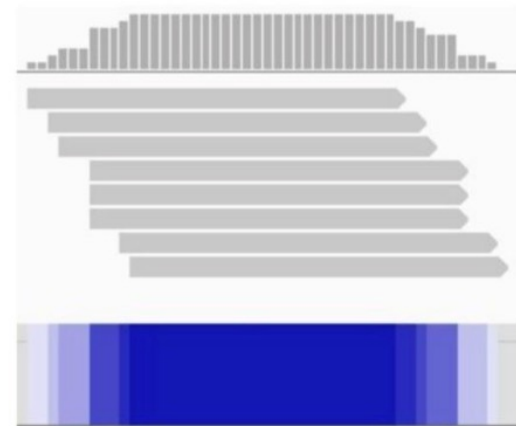
- Based on whole genome pile-up signal

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

An enriched sample: significant pile-up

SSD (**higher the better**)

- High for samples with enriched regions
- Low for controls with uniform coverage



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

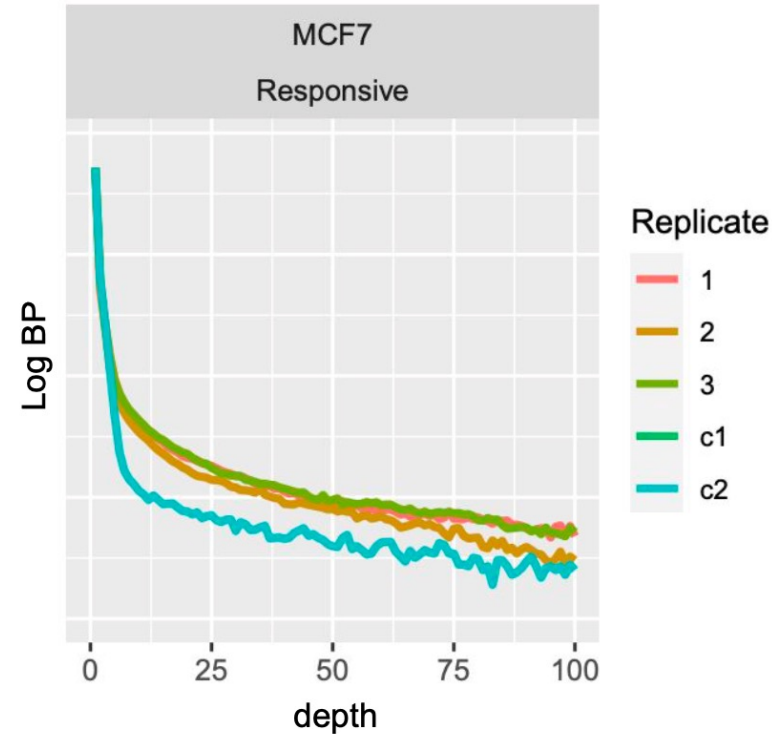
# Visualisation

## Coverage histogram

X: depth – read pileup height at a base pair position

Y: logBP – the number of positions that have this pileup height in log scale

- Good enrichment (1,2,3)
- Input (c1, c2)



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

# Database

- <http://cistrome.org/db>

| Batch                    | Species      | Biological Source                     | Factor   | Publication                                  | Quality Control<br>Library Complexity:<br>PCR bottleneck coefficient (PBC) |
|--------------------------|--------------|---------------------------------------|----------|--|--|
| <input type="checkbox"/> | Homo sapiens | HeLa; Epithelium; Cervix              | BTAF1    | Johannes F, et al. Bioinformatics 2010       |  |
| <input type="checkbox"/> | Homo sapiens | HeLa; Epithelium; Cervix              | GAPDH    | Johannes F, et al. Bioinformatics 2010       |  |
| <input type="checkbox"/> | Homo sapiens | K562; Erythroblast; Bone Marrow       | EGR1     | Tang C, et al. Electrophoresis 2010          |  |
| <input type="checkbox"/> | Homo sapiens | LS174T; Epithelium; Colon             | TCF4     | Mokry M, et al. PLoS ONE 2010                |  |
| <input type="checkbox"/> | Homo sapiens | LS174T; Epithelium; Colon             | TCF4     | Mokry M, et al. PLoS ONE 2010                |  |
| <input type="checkbox"/> | Homo sapiens | LS174T; Epithelium; Colon             | TCF4     | Mokry M, et al. PLoS ONE 2010                |  |
| <input type="checkbox"/> | Homo sapiens | LS174T; Epithelium; Colon             | TCF4     | Mokry M, et al. PLoS ONE 2010                |  |
| <input type="checkbox"/> | Homo sapiens | LS174T; Epithelium; Colon             | TCF4     | Mokry M, et al. PLoS ONE 2010                |  |
| <input type="checkbox"/> | Homo sapiens | BJ; Fibroblast; Skin                  | TERF1    | Simonet T, et al. Cell Res. 2011             |  |
| <input type="checkbox"/> | Homo sapiens | BJ; Fibroblast; Skin                  | TERF2    | Simonet T, et al. Cell Res. 2011             |  |
| <input type="checkbox"/> | Homo sapiens | 22RV1; Epithelium; Prostate           | AR       | Yu J, et al. Cancer Cell 2010                |  |
| <input type="checkbox"/> | Homo sapiens | HEK293T; Epithelium; Embryonic Kidney | PHF8     | Fortschegger K, et al. Mol. Cell. Biol. 2010 |  |
| <input type="checkbox"/> | Homo sapiens | aTconv; T Lymphocyte; Blood           | H3K4me1  | Tian Y, et al. PLoS ONE 2011                 |  |
| <input type="checkbox"/> | Homo sapiens | aTconv; T Lymphocyte; Blood           | H3K4me3  | Tian Y, et al. PLoS ONE 2011                 |  |
| <input type="checkbox"/> | Homo sapiens | BG01; Embryonic Stem Cell; Embryo     | H3K27me3 | Guenther MG, et al. Cell Stem Cell 2010      |  |

Lunch break before practical

Back at 13.30