Introduction to Chromatin IP – sequencing (ChIP-seq) data analysis

Epigenomics Data Analysis Workshop

Stockholm, 25 October 2021

Agata Smialowska

NBIS, SciLifeLab, Stockholm University

NBIS NATIONAL BIOINFORMATICS INFRASTRUCTURE SWEDEN

SciLifeLab

UNIVERSITET STOCKHOLMS

Chromatin state and gene expression

PEV
Position effect variegation
in *Drosophila* eye
(nature.com)

First observed by
H. Muller
1930

Juxtaposition of eye colour genes with heterochromatin results in the “mottled” eye colouration (red and white).

Proteins, which bind heterochromatin, act to “spread” the silencing signal by providing a forward feedback loop.

Heterochromatin Protein 1; Histone methyltransferase Su(var)3-9; H3K9 methylation

Chromatin immunoprecipitation

- John T. Lis and David Gilmour, used UV irradiation to **cross-link** proteins bound to DNA in living bacterial cells, immuno-precipitated the complexes and analysed them on a dot blot (1984);

- **ChIP-on-chip**: use microarrays to analyse ChIP fragments genome-wide (late 1990's);

- **ChIP-seq**: use massively parallel sequencing (a.k.a. next generation sequencing, NGS) to perform genome-wide study in a non-biased fashion (first publications 2007).

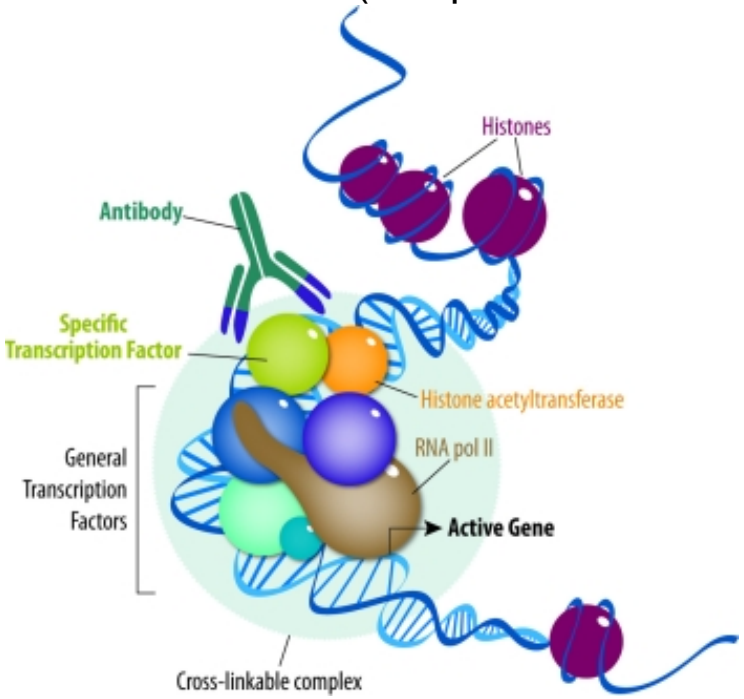
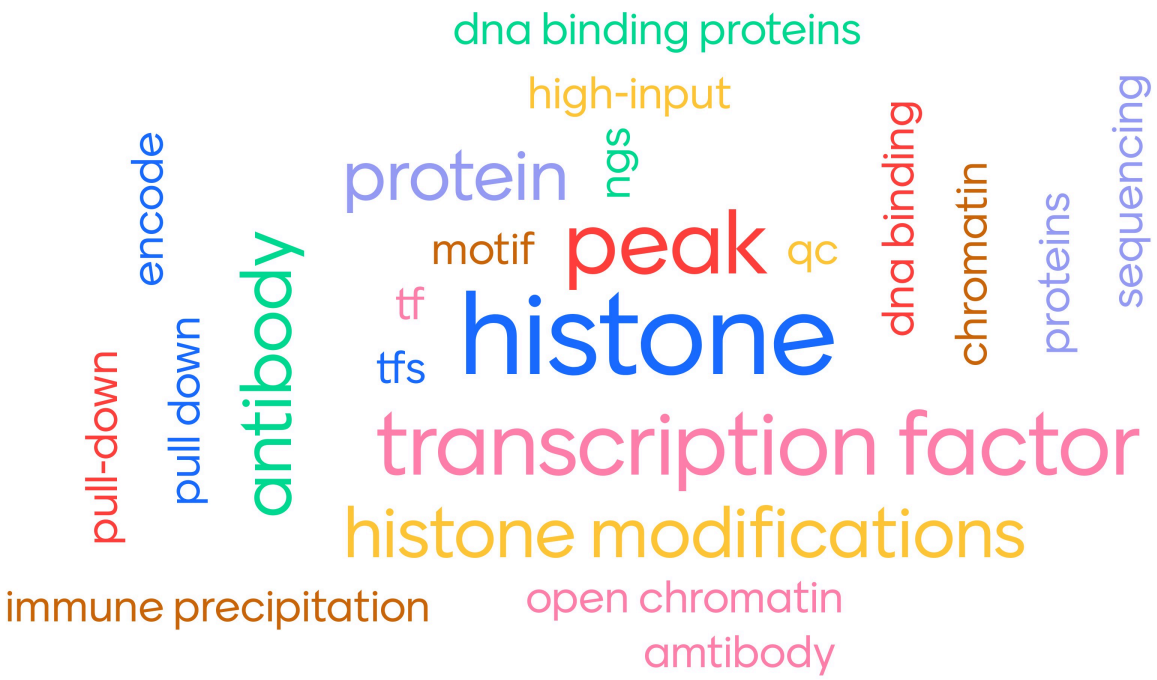


image: RnDsystems

Keywords you associate with ChIP-seq

Mentimeter



20

keywords we associate with ChIP-seq

Workflow of a ChIP-seq study

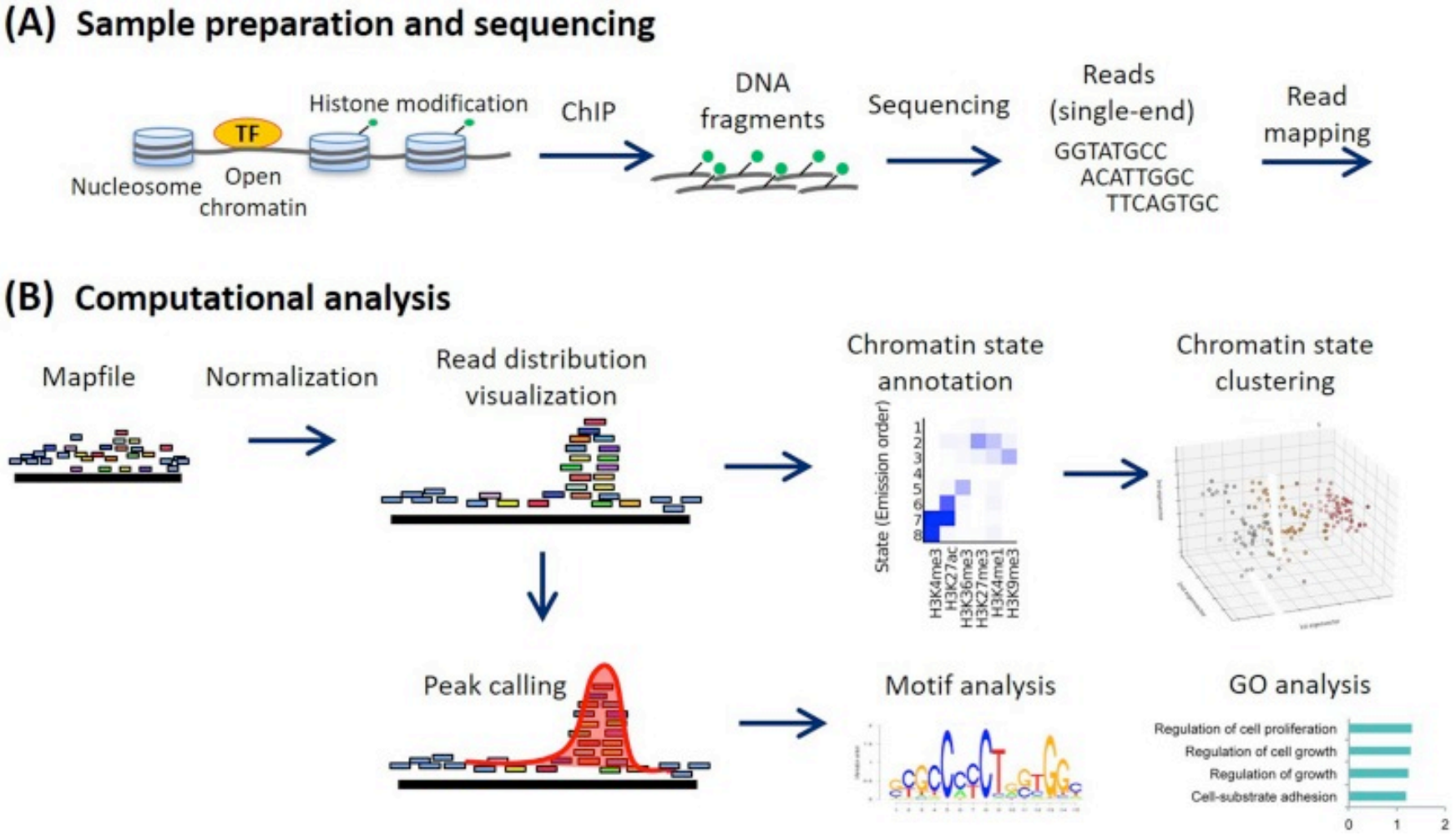


image: Nakato et al, 2021

Workflow of a ChIP-seq study

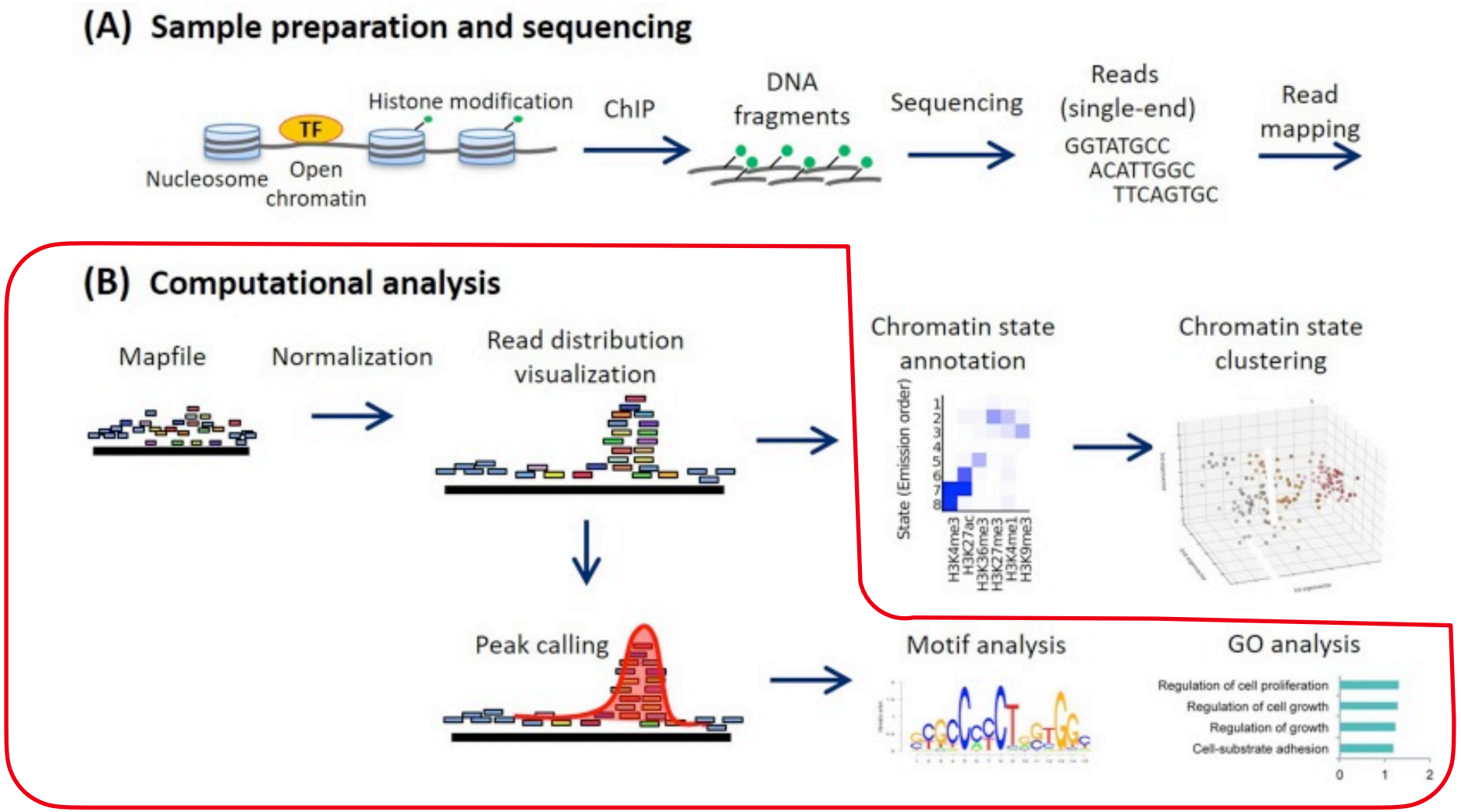
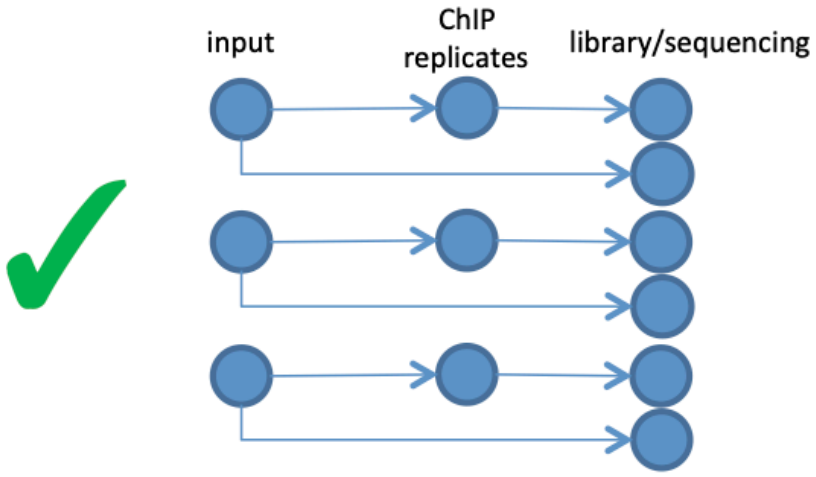
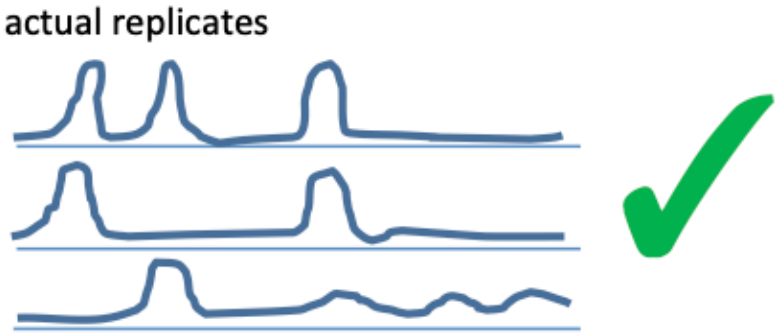


image: Nakato et al, 2021

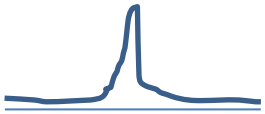
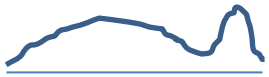
Experiment design

- Sound experimental design: replication, randomisation, control and blocking (R.A. Fisher, 1935)
- In the absence of a proper design, it is essentially impossible to partition biological variation from technical variation
- Please visit section *Experimental Design and Data Management* on the course website for more information





Sequencing depth depends on data type

	Transcription Factors	Chromatin Remodellers Histone marks	Chromatin Remodellers Histone marks RNA polymerase II
			
	point-source	mixed signal	broad signal
Human:	TF: 20 M H3K4me3: 25 M	? H3K36me3: 35 M	? H3K27me3: 40 M H3K9me3: >55 M

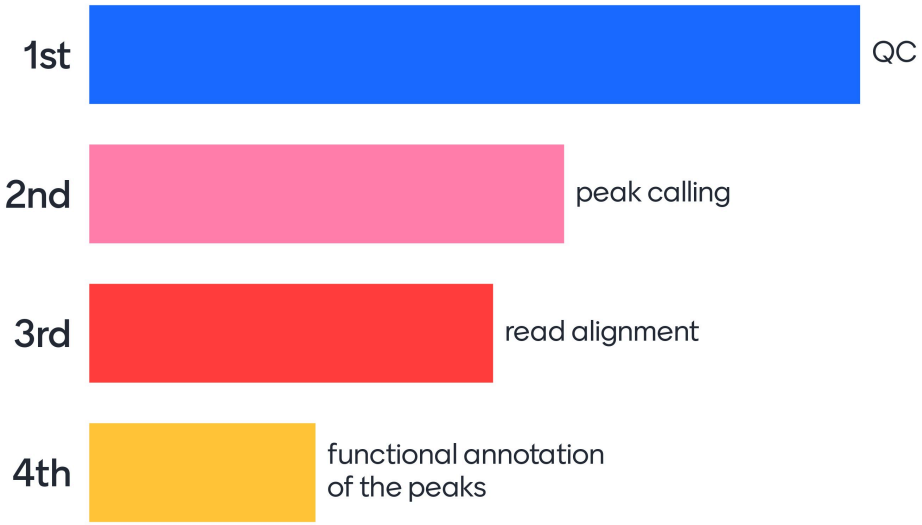
No clear guidelines for mixed and broad type of peaks

Source: The ENCODE consortium; Jung et al, NAR 2014

- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources

Rank the steps of data analysis byy importance to analysis and interpretation of ChIP-seq data

Mentimeter



17

Which part of the analysis is of key importance to analysis and interpretation of ChIP-seq experiments

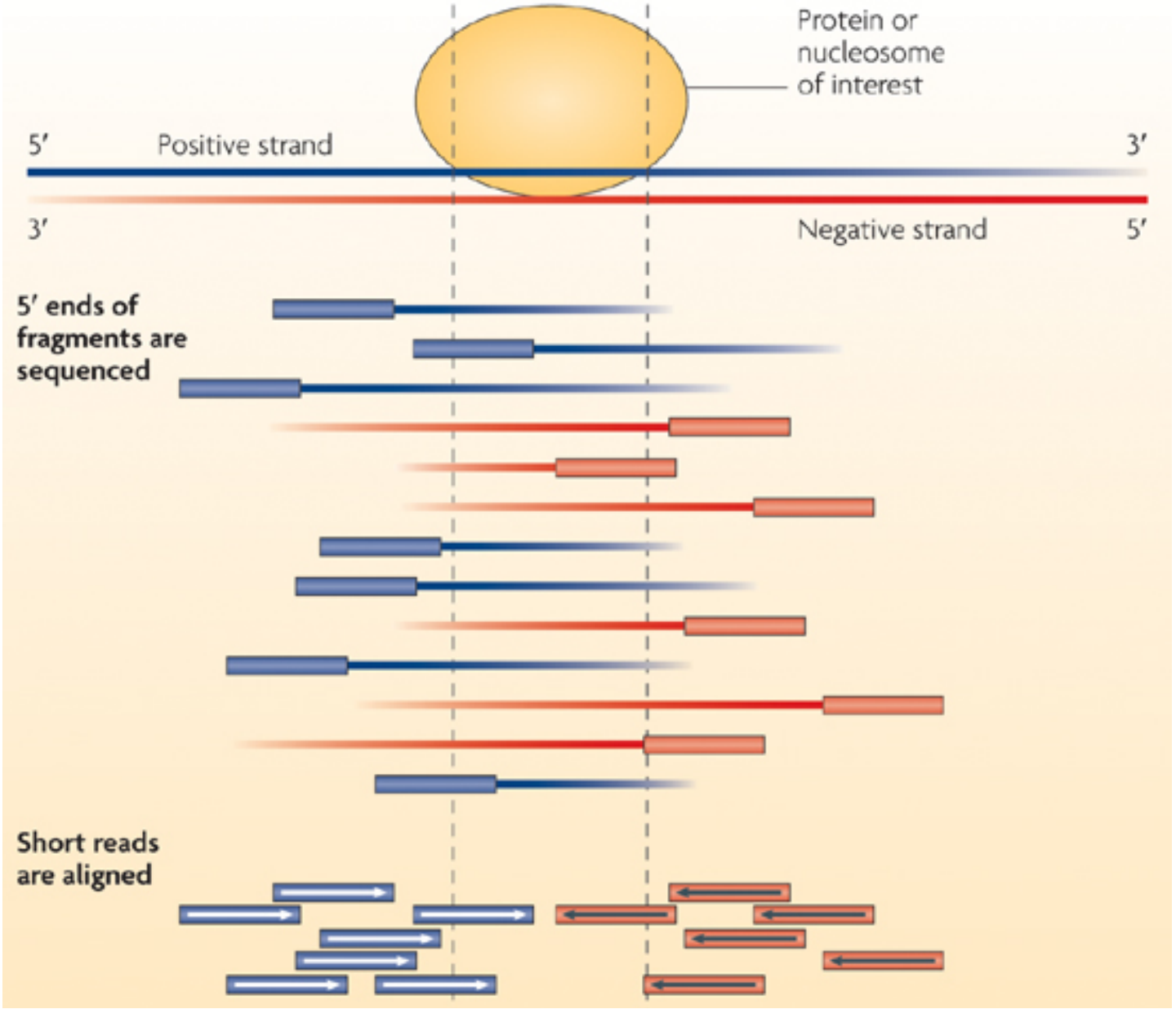
- ChIP – sequencing: introduction from a bioinformatics point of view

- Principles of analysis of ChIP-seq data

- ChIP-seq: downstream analyses

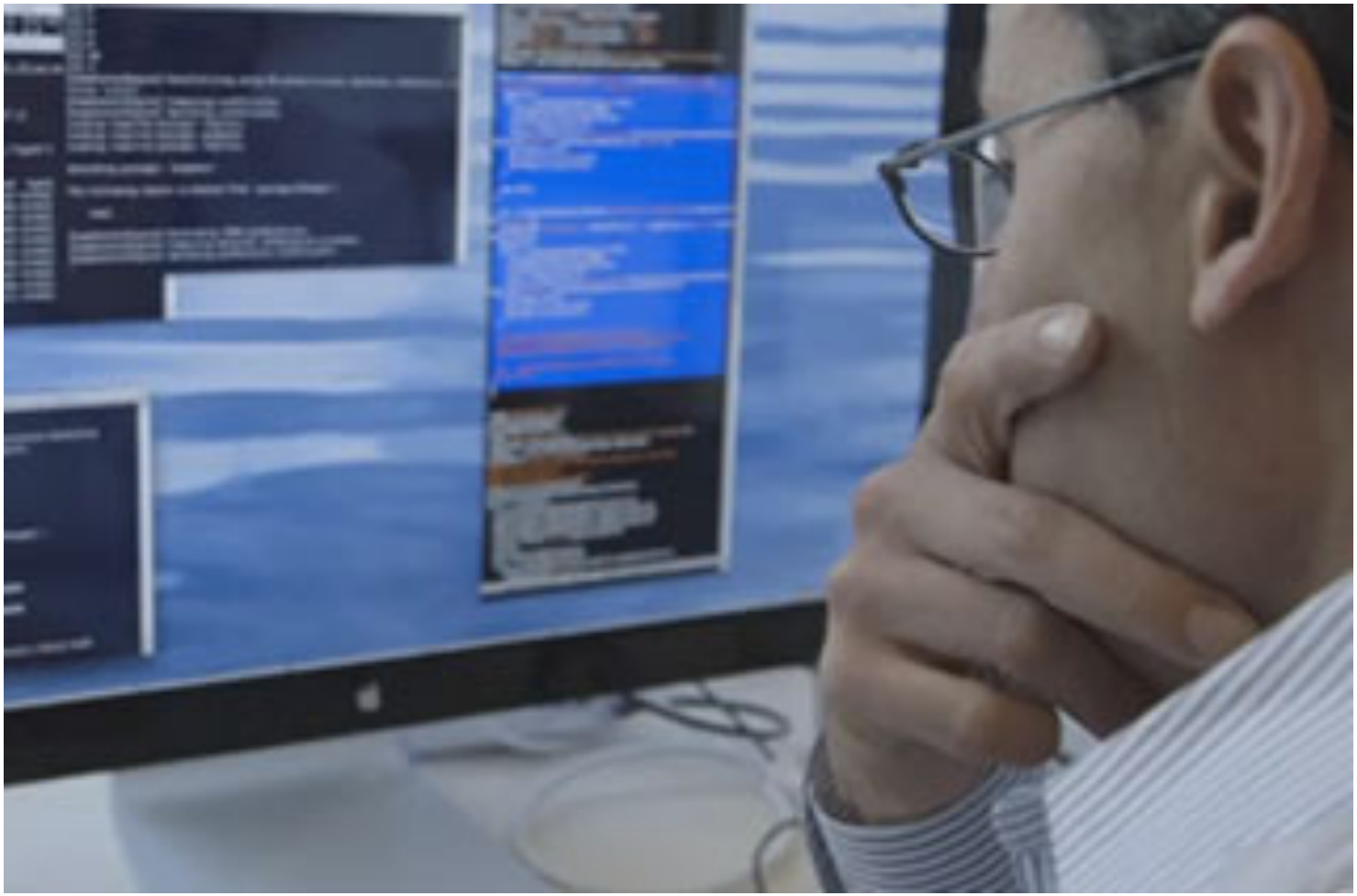
- Resources

Chromatin = DNA + proteins



Park, Nature Rev Genetics, 2009

Data analysis



Workflow of a ChIP-seq study

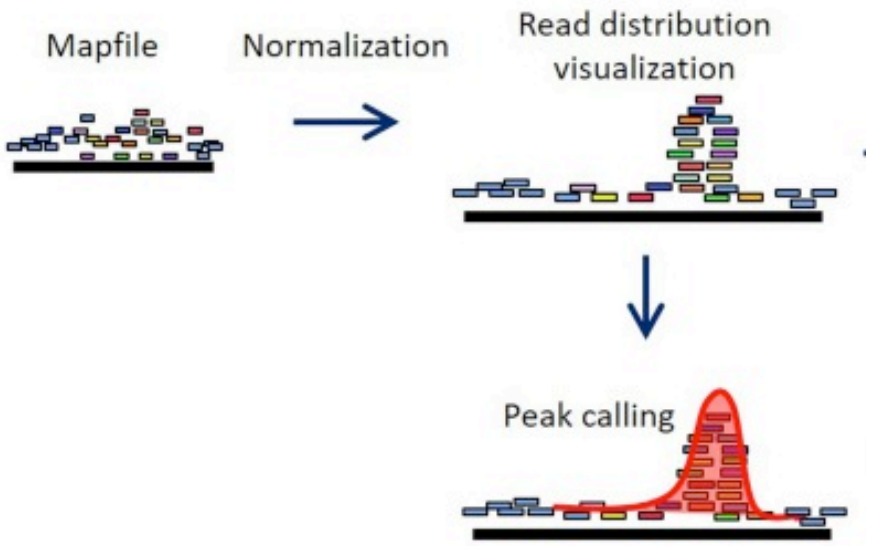


image: Nakato et al, 2021

Workflow of a ChIP-seq study

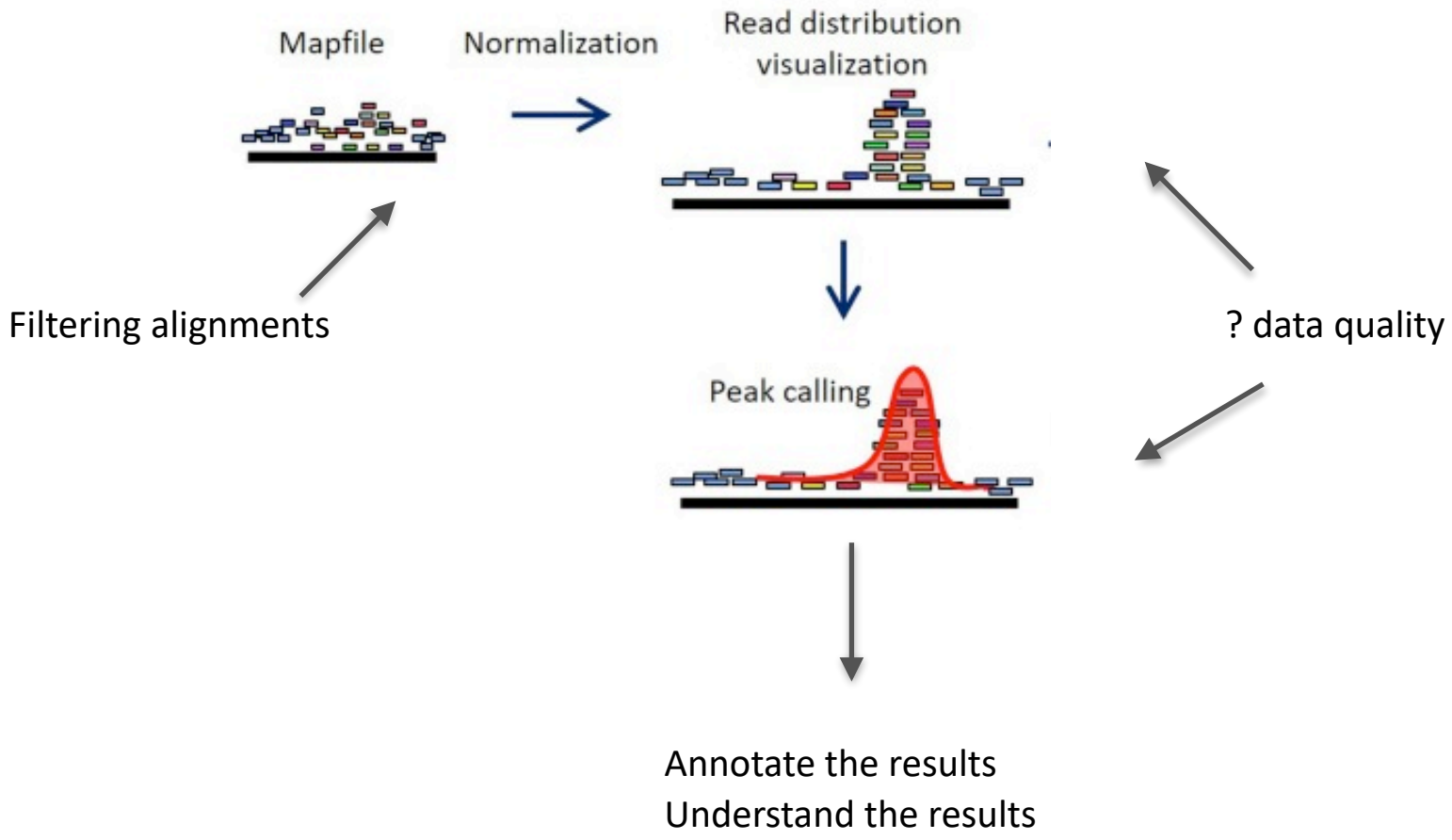
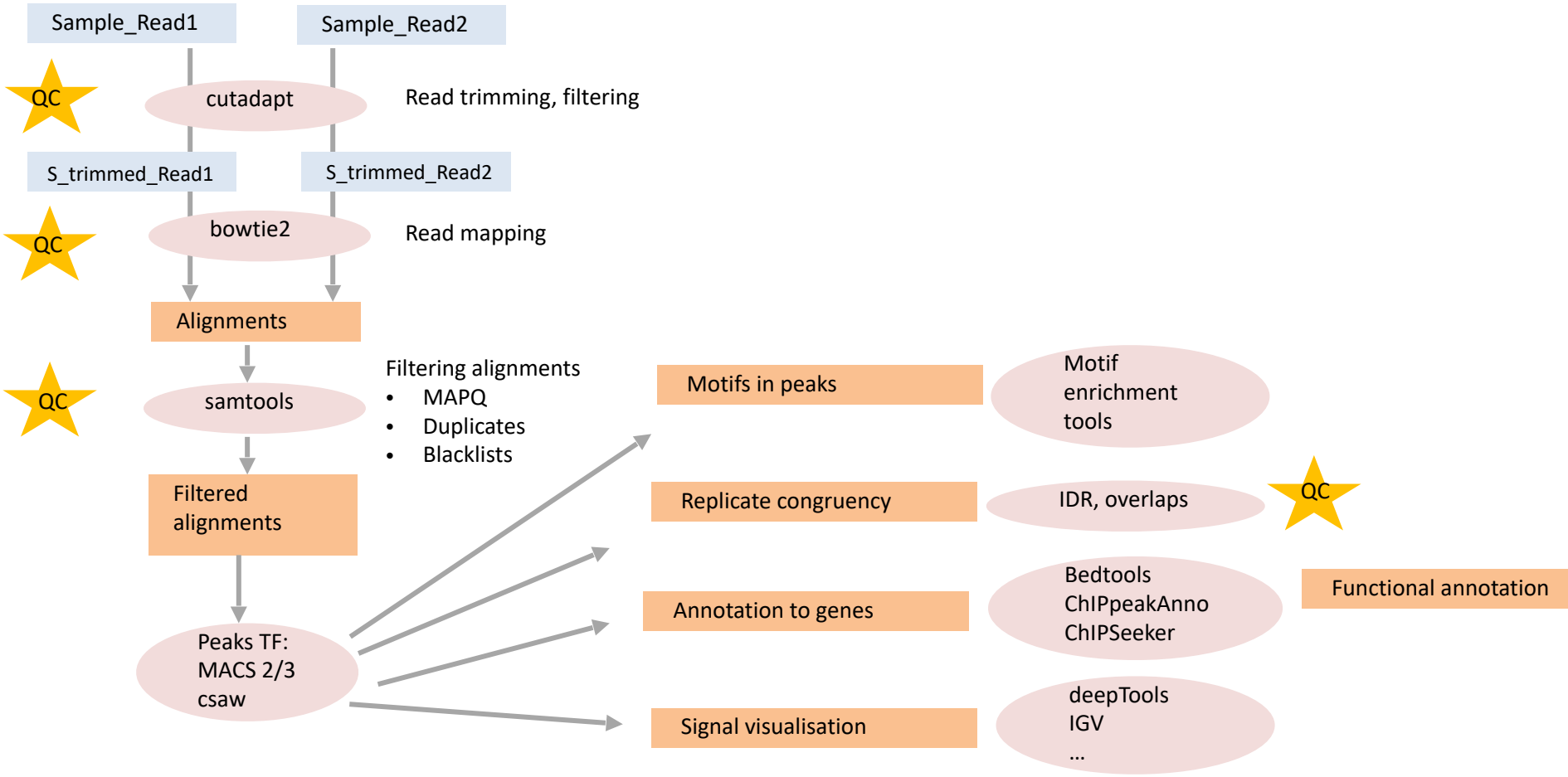


image: Nakato et al, 2021

Analysis workflow



Word of caution!

ChIP-seq experiments are more unpredictable than RNA-seq!

Inconsistency sources:

- chromatin structure
- non-specific antibody
- crosslinking
- PCR over-amplification
- Other?

- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources

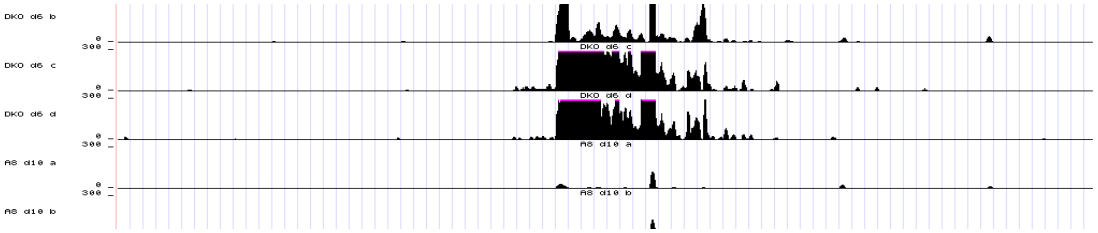
Two questions to address

- 1. Did the ChIP part of the ChIP-seq experiment work? Was the enrichment successful?

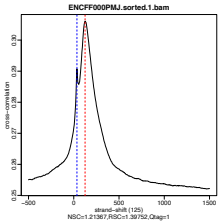
- 2. Where are the binding sites (of the protein of interest)?

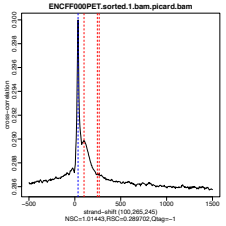
ChIP-seq QC: did the ChIP work?

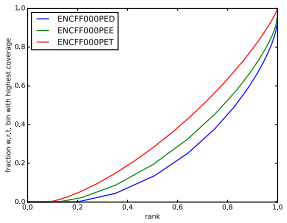
- 1. Inspect the signal (mapped reads, coverage profiles) in genome browser



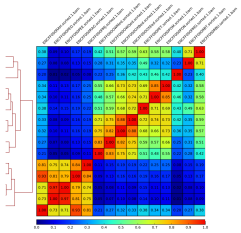
- 2. Compute peak-independent quality metrics (cross correlation, cumulative enrichment)

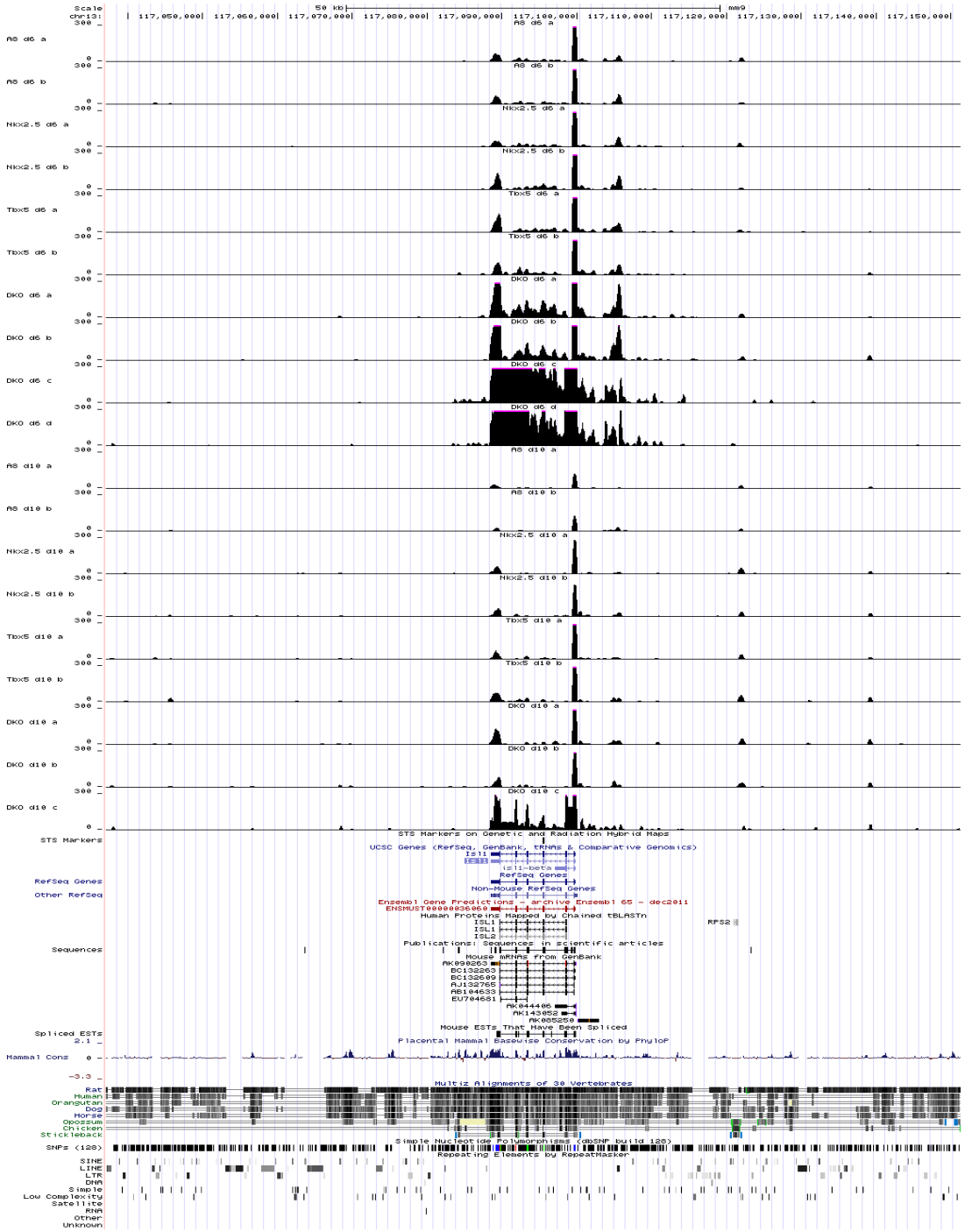






- 3. Assess replicate consistency (correlations between replicates of the same condition)





tag density distribution
reproducibility
similarity of coverage
signal at known sites
…
Spotting inconsistencies
Confounding factors
Under-sequenced libraries
…

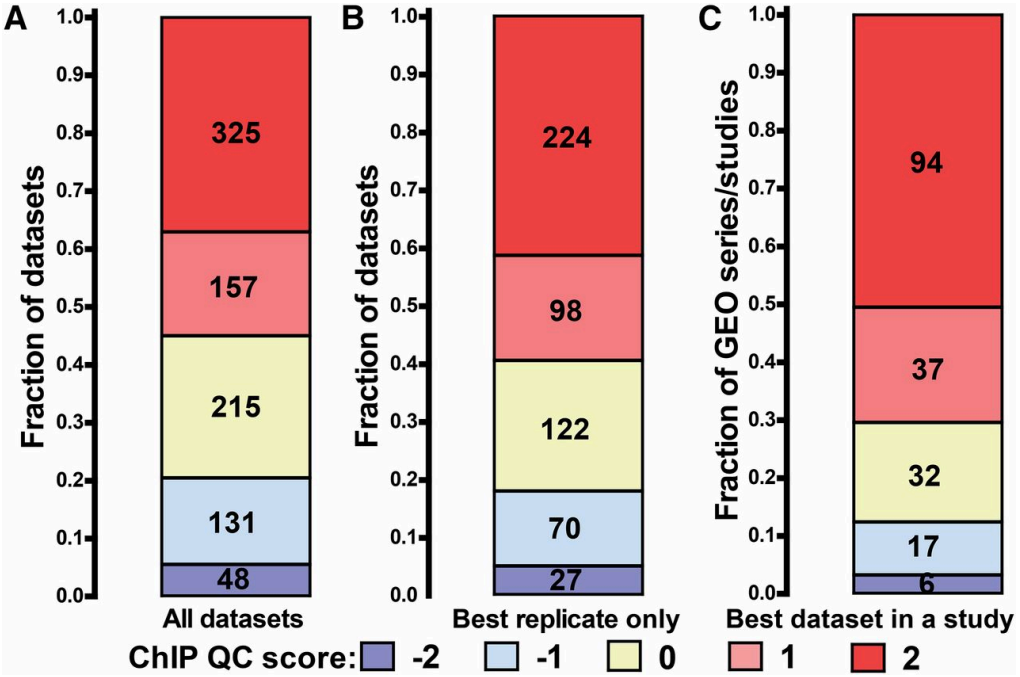
How do I know my data is of good quality?

Objective (i.e. peak independent) metrics to quantify enrichment in ChIP-seq;

for TF in mammalian systems:

Normalised Strand Correlation NSC
Relative Strand Correlation RSC

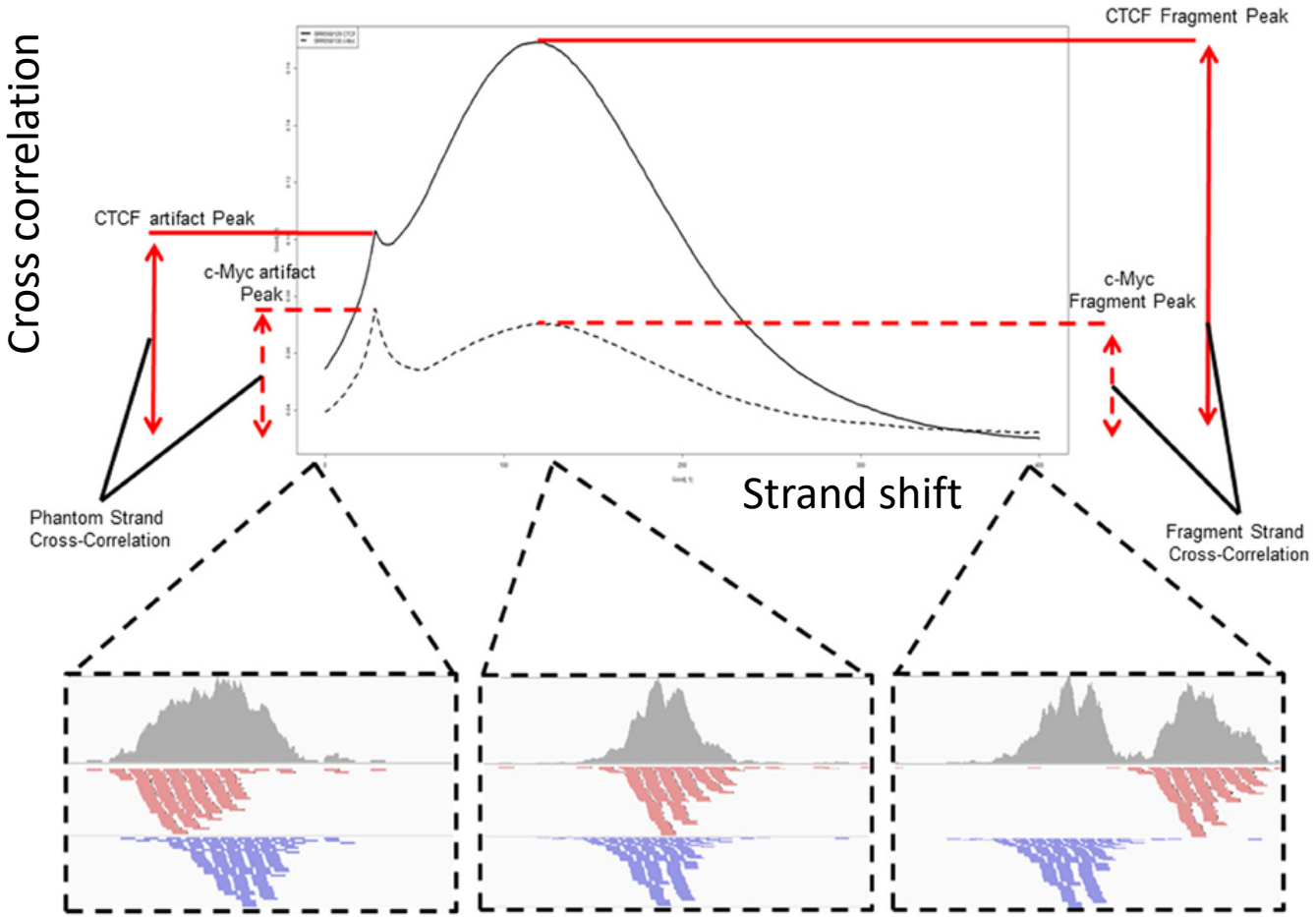
Large-scale quality analysis of published ChIP-seq data sets:
20% low quality
25% intermediate quality
30% inputs have metrics similar to IPs



Marinov et al, G3 2013

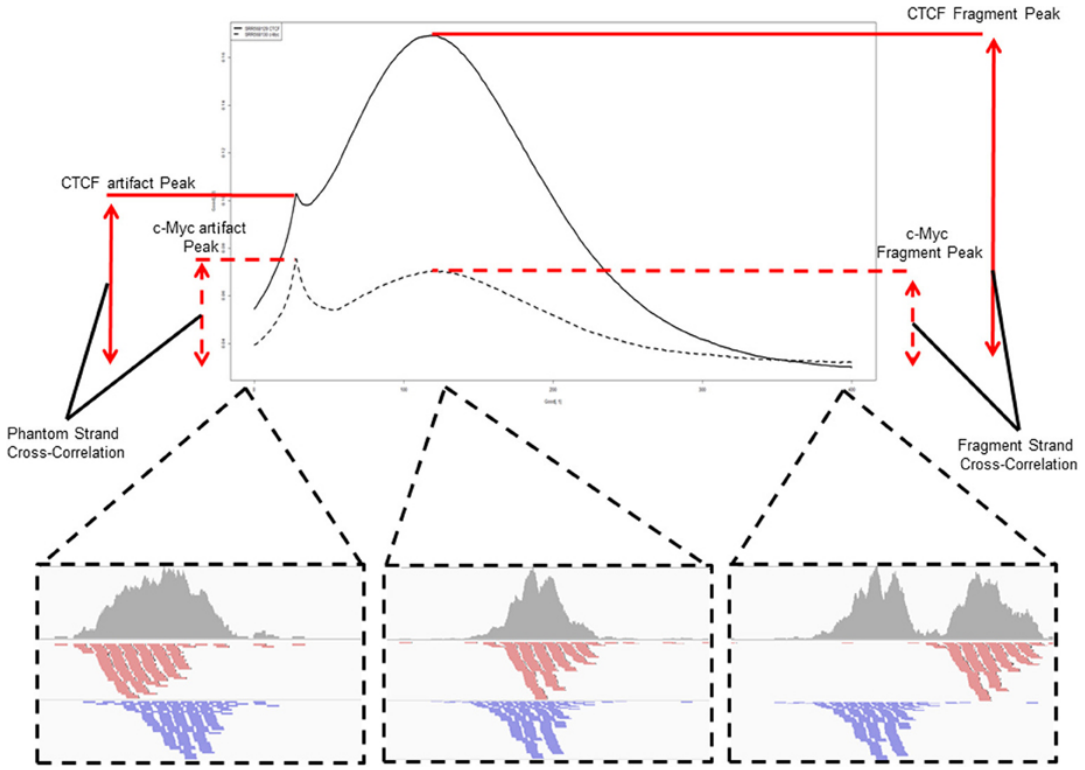
Strand cross-correlation

The correlation between signal of the 5’ end of reads on the (+) and (-) strands is assessed after successive shifts of the reads on the (+) strand and the point of maximum correlation between the two strands is used as an estimation of fragment length.



Carroll et al, Front Genet 2014

Strand cross-correlation



$$NSC = \frac{\text{Max CC value (fLen)}}{\text{Min CC}}$$

$$RSC = \frac{\text{Max CC} - \text{Min CC}}{\text{Phantom CC} - \text{Min CC}}$$

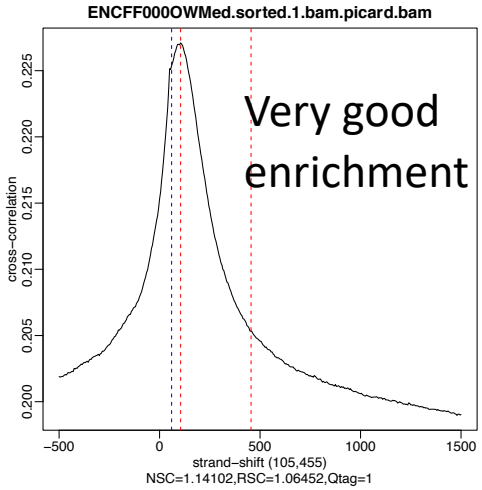
NSC > 1.05

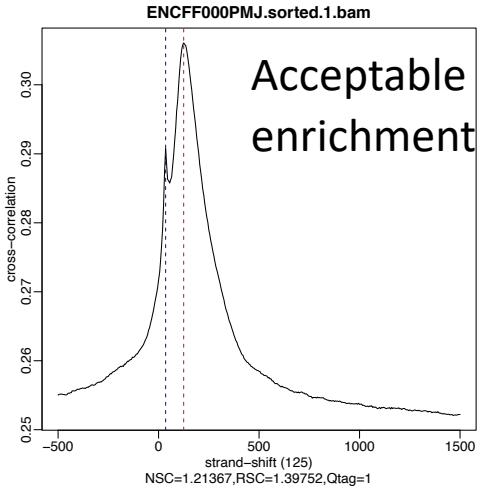
RSC > 0.8

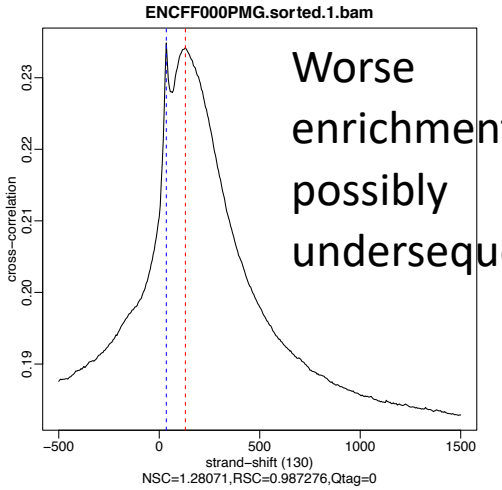
Carroll et al, Front Genet 2014

Cross-correlation plots

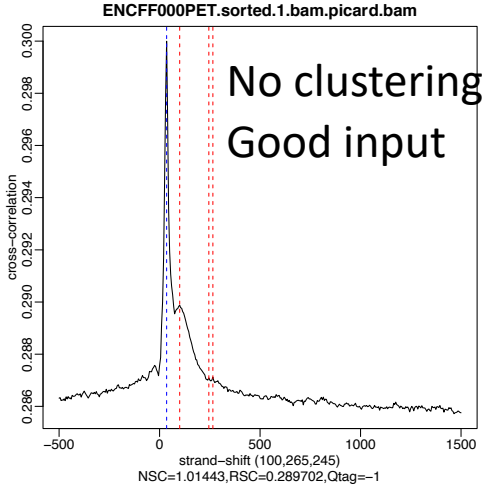
ChIP

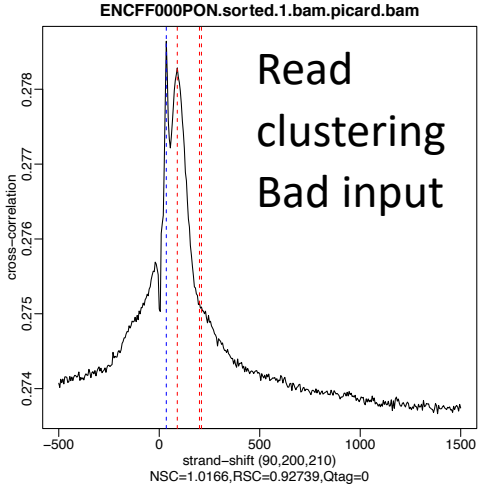




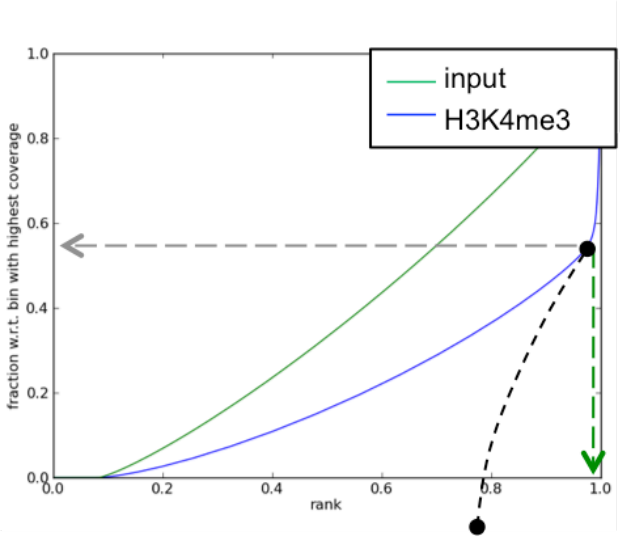


Input



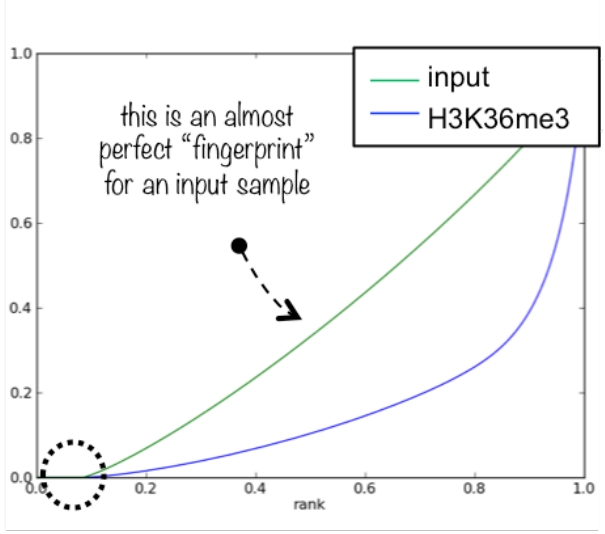


Cumulative enrichment aka “Fingerprint” is another metric for successful enrichment

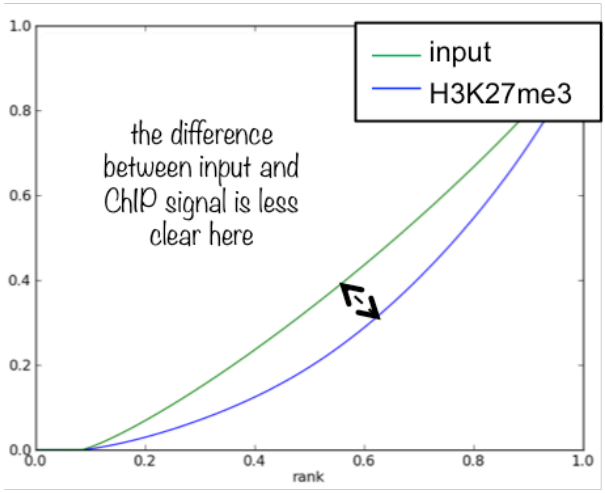


when counting the reads contained in **97%** of all genomic bins, only ca. 55% of the maximum number of reads are reached, i.e. 3% of the genome contain a very large fraction of reads!

→ this indicates very localized, very strong enrichments!
(as every biologist hopes for in a ChIP for H3K4me3)



pay attention to where the curves start to rise – this already gives you an assessment of how much of the genome you have not sequenced at all (i.e. bins containing zero reads – for this example, ca. 10% of the entire genome do not have any read)



H3K27me3 is a mark that yields broad domains instead of narrow peaks

→ it is more difficult to distinguish input and ChIP, it does not mean, however, that this particular ChIP experiment failed

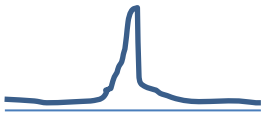
http://deeptools.readthedocs.org

Diaz et al, Genome Biol 2012

Peak calling

appropriate methodologies depend on data type

Transcription Factors



point source

MACS 2 / 3
window approaches

Chromatin Remodellers

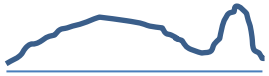
Histone marks

mixed

Chromatin Remodellers

Histone marks

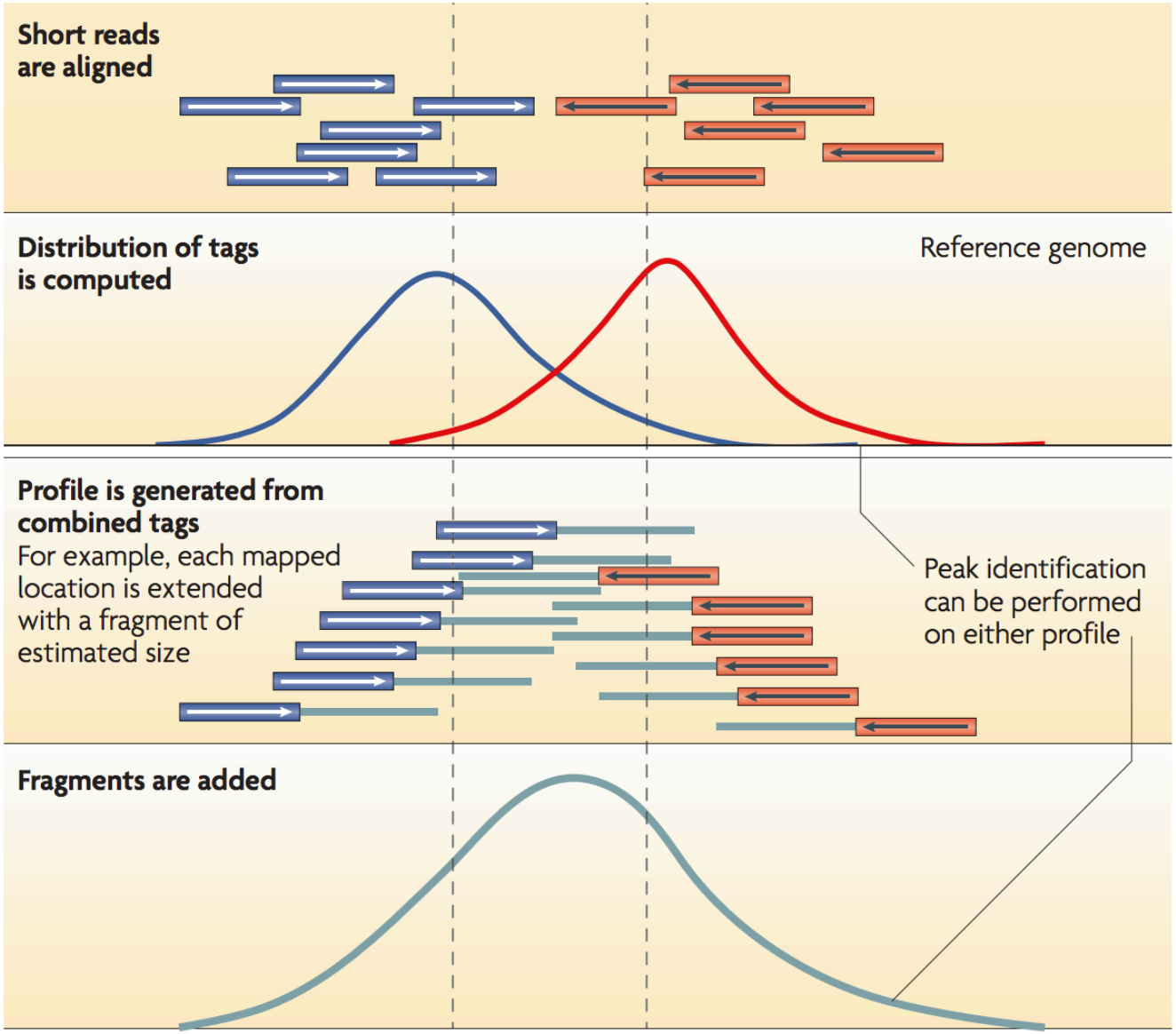
RNA polymerase II



broad

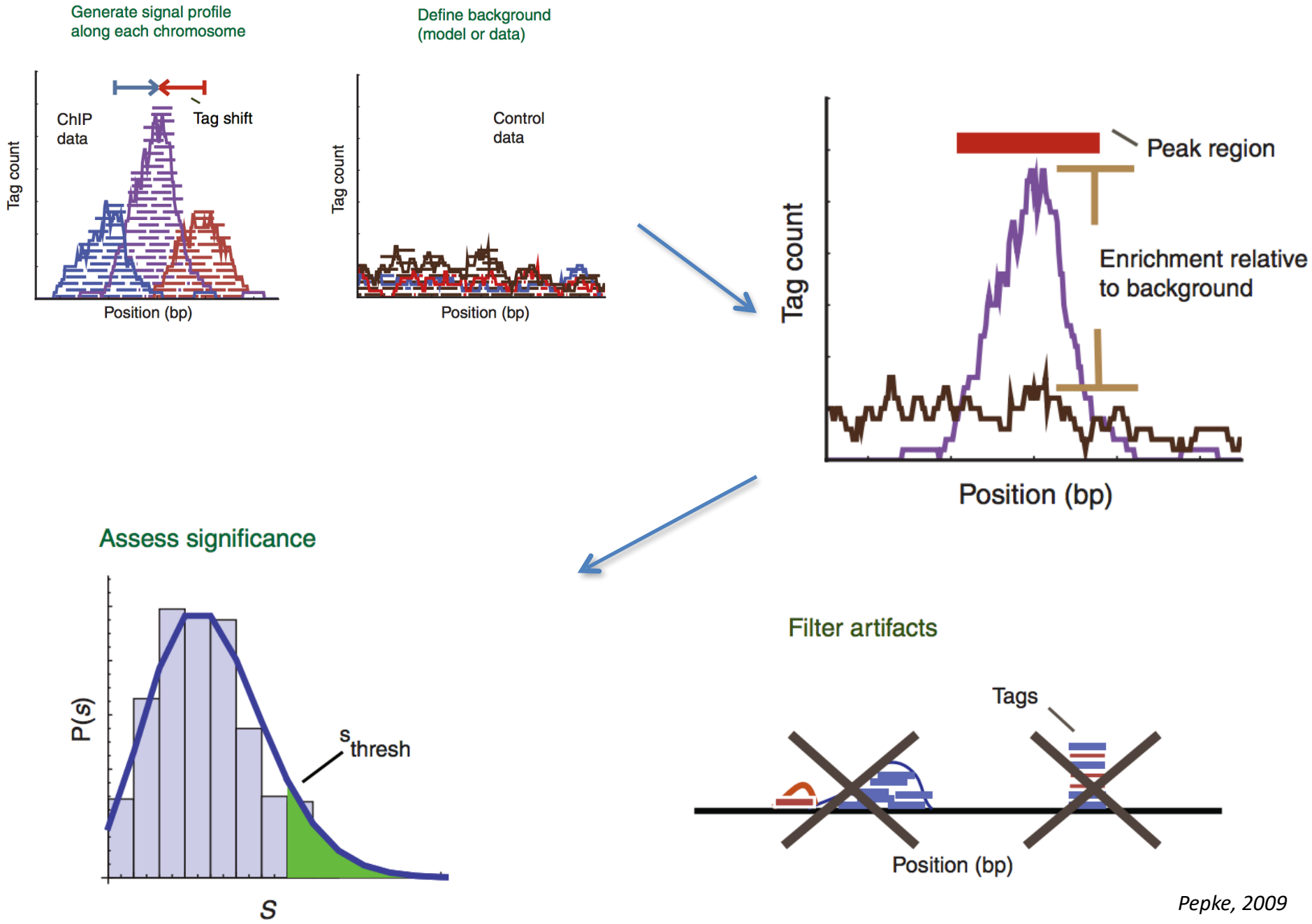
MACS 3 in broad mode
windows approaches
Epic2 (SICER)

Principle of peak detection



Symmetry in reads mapped to opposite DNA strands

Computation of enrichment model


Generate signal profile
along each chromosome
ChIP
data
Tag shift
Tag count
Position (bp)
Define background
(model or data)
Control
data
Tag count
Position (bp)
Peak region
Enrichment relative
to background
Tag count
Position (bp)
Assess significance
P(s)
s
thresh
S
Filter artifacts
Tags
Position (bp)

Pepke, 2009

Comparison of peak calling algorithms

The problem with comparisons: small number of data sets, parameter settings, choice of performance metrics

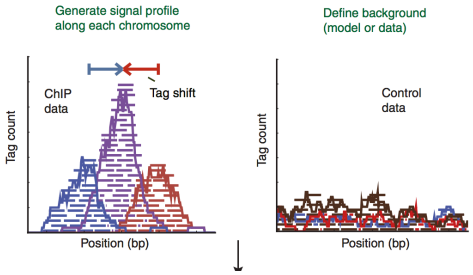
Features that define the best ChIP-seq peak calling algorithms

Reuben Thomas, Sean Thomas, Alisha K. Holloway and Katherine S. Pollard

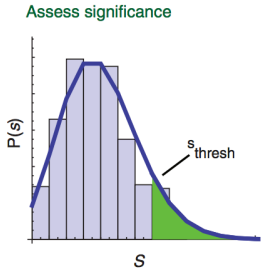
Corresponding author: Katherine S Pollard, Gladstone Institutes, San Francisco, CA 94158, USA. Tel.: 415-734-2711. Fax: 415- 355-0141; E-mail: katherine.pollard@gladstone.ucsf.edu

- Model-based Analysis for ChIP-Seq version 2 / 3 (MACS 2/3)
- MUltiScale enrIchment Calling for ChIP-Seq (MUSIC)
- Genome wide Event finding and Motif discovery (GEM)
- Zero-Inflated Negative Binomial Algorithm (ZINBA)
- Bayesian ChangePoint (BCP)
- Threshold-based method (TM)

Comparison of peak calling algorithms



Identification of enriched sites (peak candidates)



Testing candidates for significance

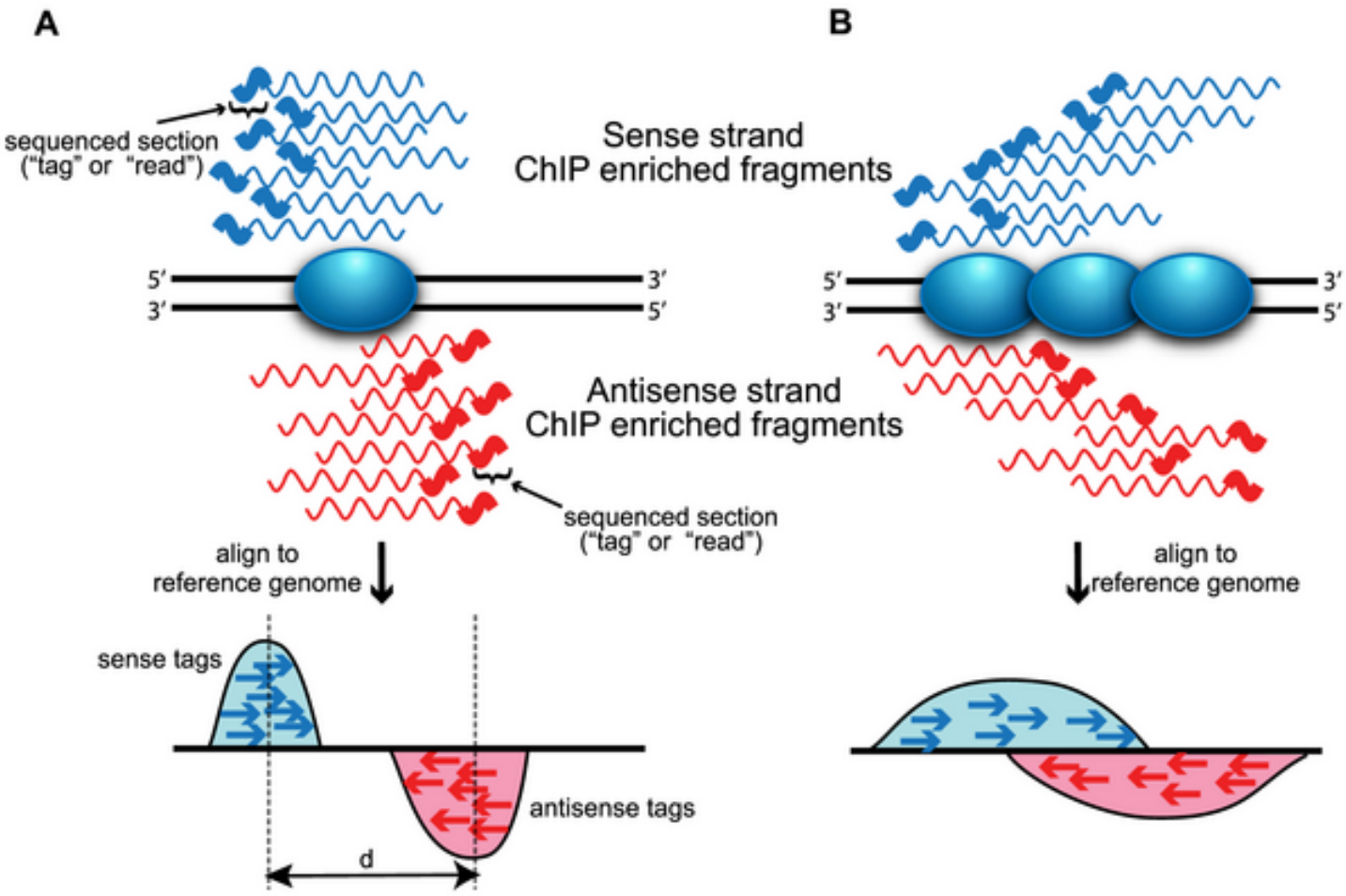
Thomas et al, 2017: surveyed 30 methods and identified 12 features of the two sub-problems that distinguish methods from each other.

- methods that explicitly combine the signals from ChIP and input samples are less powerful than methods that do not
- methods that use windows of different sizes are more powerful than the ones that do not
- for statistical testing of candidate peaks, methods that use a Poisson test to rank their candidate peaks are more powerful than those that use a binomial test

Point-source vs. broad peak detection

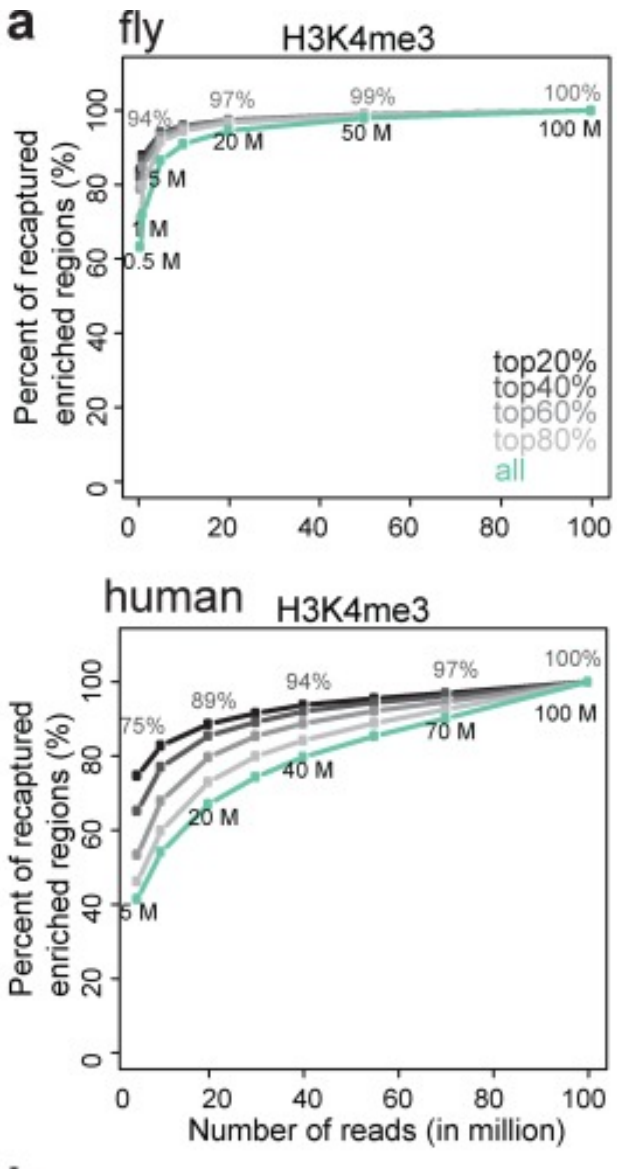
Sequence-specific binding (TFs)

Distributed binding (histones, RNApol2)



Wilbanks 2010

How to define sufficient read depth?



detection sensitivity - recovery of true enrichment region

↓

percent increase in enriched regions recaptured when an additional 1 million reads are sequenced

↓

'sufficient depth' - the sequencing depth at which the percent gain per 1 million additional sequence reads falls below 1%

Jung et al, NAR 2014

Peak calling

Chromatin Remodellers

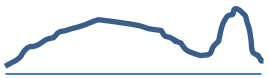
Histone marks

mixed signal

Chromatin Remodellers

Histone marks

RNA polymerase II

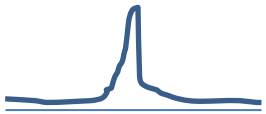


broad signal

- **MACS 3** in broad mode: window-based method using local statistics
- **csaw:** Scoring in moving windows
- **Epic2** (SICER): tendency of histone modifications to cluster to form the domains. This method identifies islands as *clusters* of enriched windows. Islands, rather than individual windows of fixed length, are the fundamental units of interest

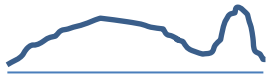
Bäst i test (2017)

- BCP and MACS2 have the best operating characteristics on simulated transcription factor binding data.



- GEM has the highest fraction of the top 500 peaks containing the binding motif of the immunoprecipitated factor, with 50% of its peaks within 10 base pairs of a motif.

- BCP and MUSIC perform best on histone data

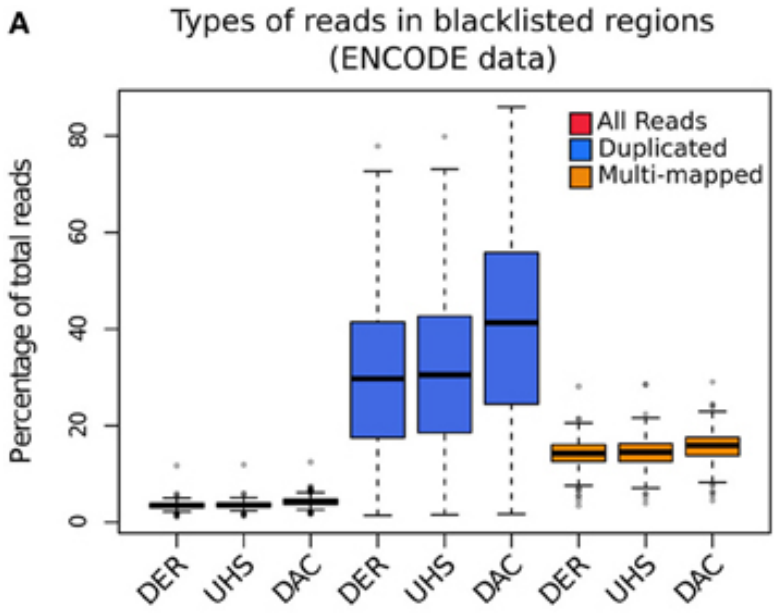


New players (2019 - 2020):

- Epic2 (SICER) (Stovner et al, 2019) – diffuse signals - (Spatial Clustering for Identification of ChIP-Enriched Regions)

- Genrich (unpubl.) – dedicated ATAC-seq mode (Tn5 cut sites), can also call ChIP-seq peaks (fragments), leverages using replicates

Blacklist: “Hyper-chippable” regions



Reads mapped to these regions should be filtered out prior to peak calling

Tracks available from UCSC for human, mouse, fly and worm

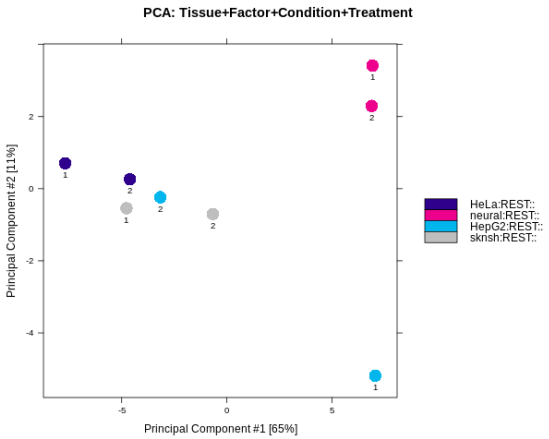
DER – Duke Excluded Regions
(11 repeat classes)
UHS – Ultra High Signal
(open chromatin)
DAC – consensus excluded regions

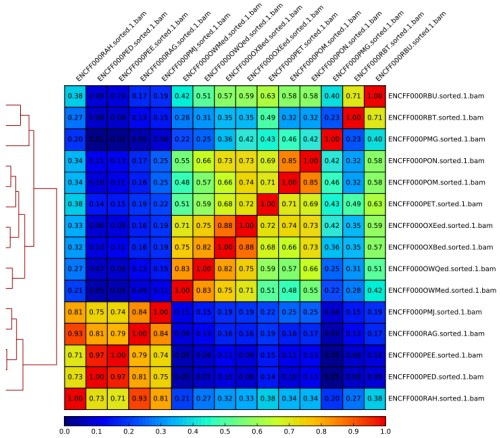
Carroll et al, Front Genet 2014

Replicate congruency

If two replicates measure the same underlying biology, the most significant peaks which are likely to be genuine signals, are expected to have high consistency between replicates. Peaks with low significance, which are more likely to be noise, are expected to have low consistency.

- PCA, sample clustering (signal in enrichment regions)



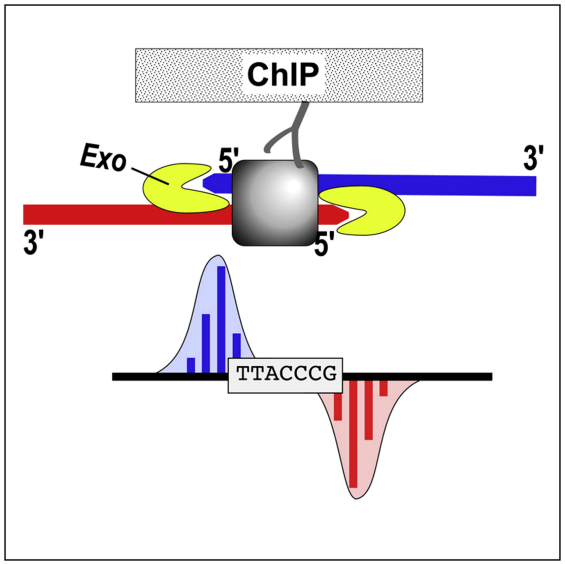


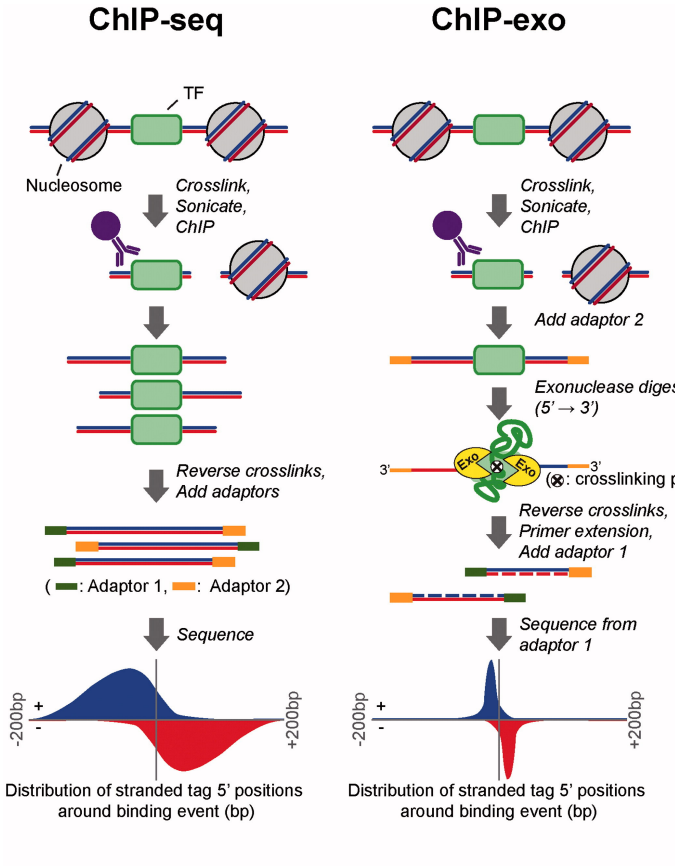
- Irreproducible discovery rate (IDR) - compare a pair of ranked lists of identifications (such as ChIP-seq peaks). These ranked lists should not be pre-thresholded, i.e they should provide identifications across the entire spectrum of high confidence/enrichment (signal) and low confidence/enrichment (noise). This method helps to set an optimal cutoff for significance.

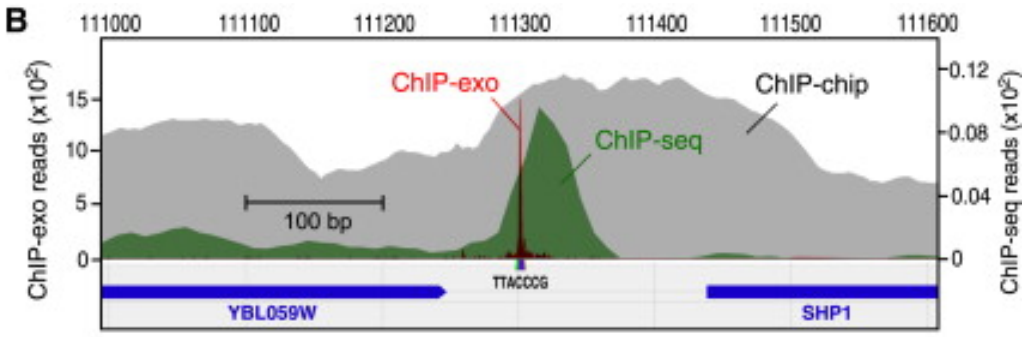
Quality considerations

- ChIP-seq quality guidelines from the ENCODE project: relative strand cross-correlation, library complexity (and irreproducible discovery rate)
- Antibody validation
- Appropriate sequencing depth (depending on genome size and peak type). For human genome and broad-source peaks, min. 40-50M reads is required.
- Experimental replication
- Fraction of reads in peaks (FRiP) > 1%
- Cross correlation (correlation of the density of sequences aligned to opposite DNA strands after shifting by the fragment size)
- Experimental verification of known binding sites (and sites not bound as negative controls)

ChIP-exo: improvement in binding site identification

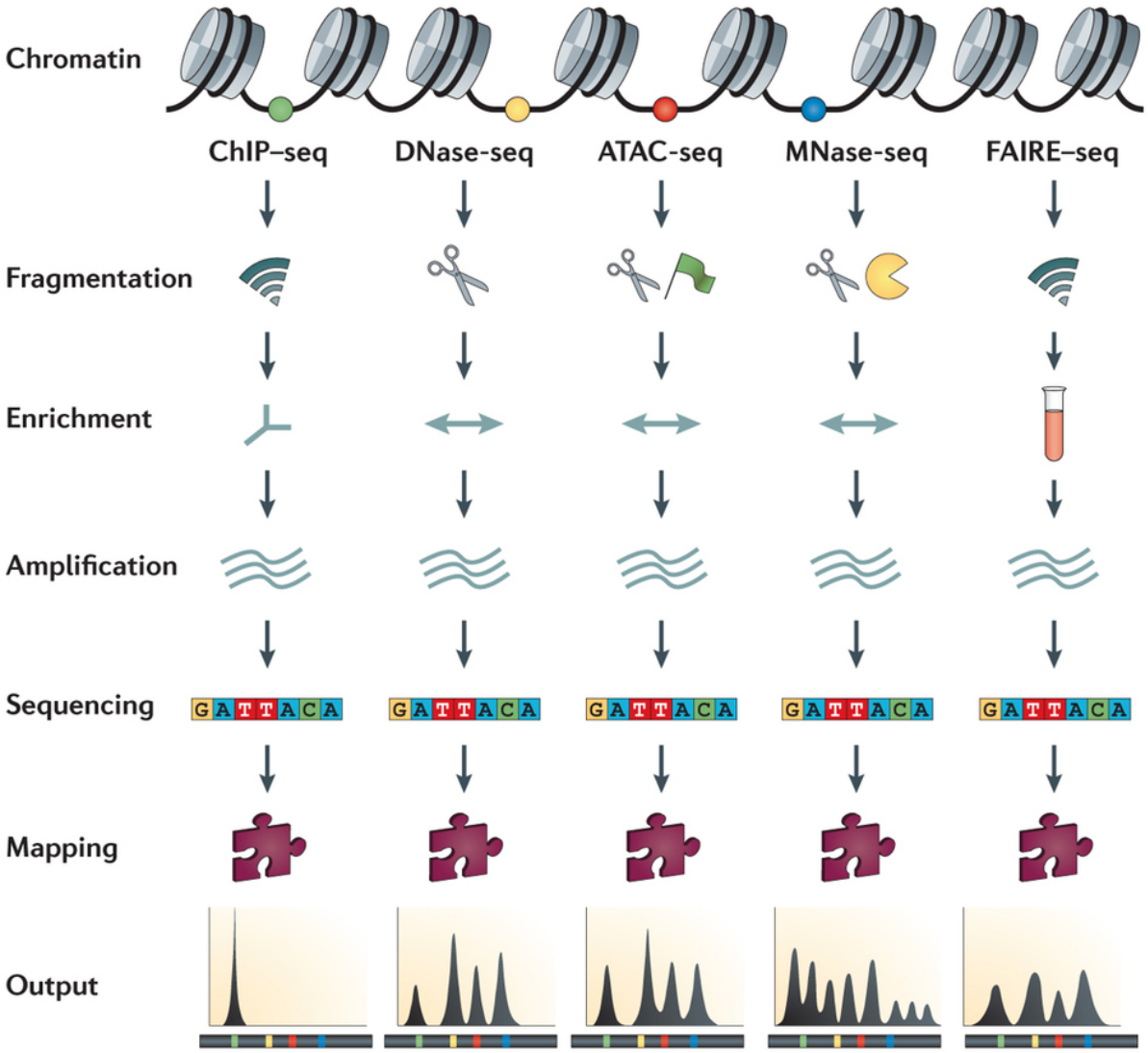






Pugh 2015
Rhee and Pugh, Cell 2011

Other functional genomics techniques

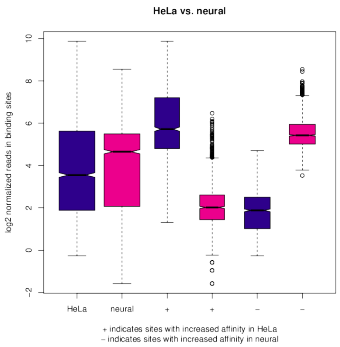


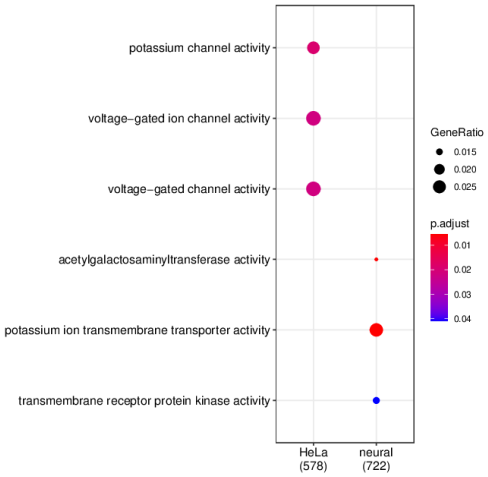
Clifford et al, Nature Rev Genet, 2014

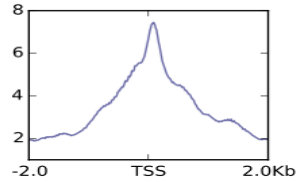
- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources

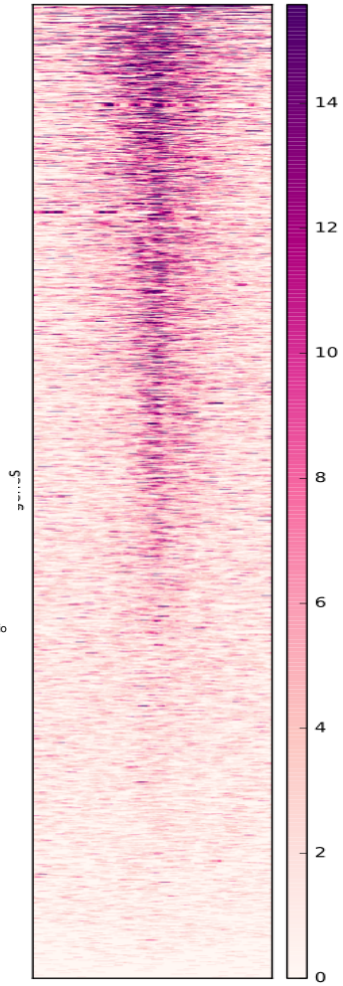
ChIPseq downstream analyses

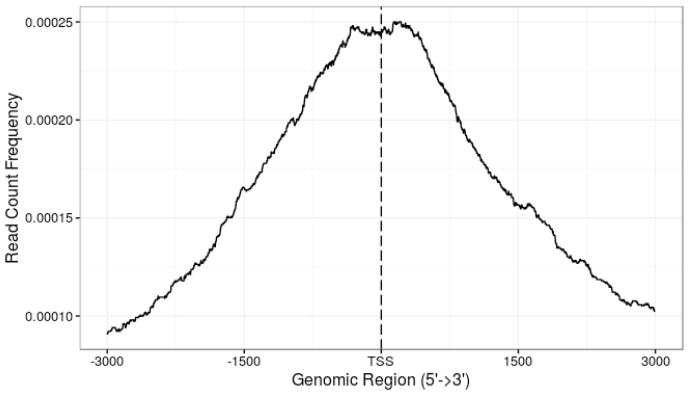
- Validation (wet lab)
- Downstream analysis
  - Motif discovery
  - Annotation
  - Integration of binding and expression data
  - Integration of various binding datasets
  - Differential binding

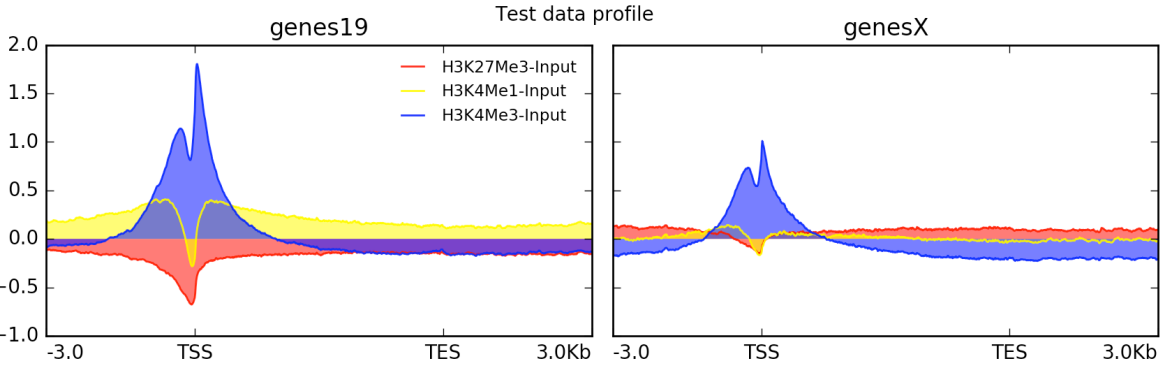


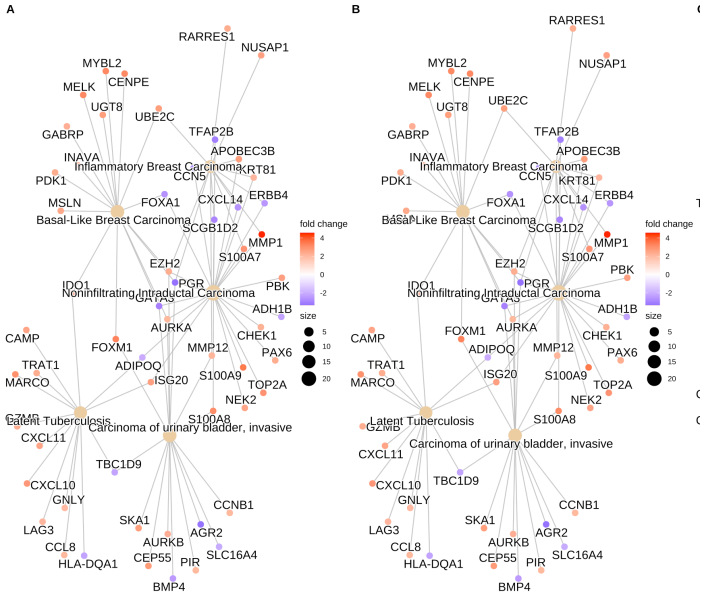


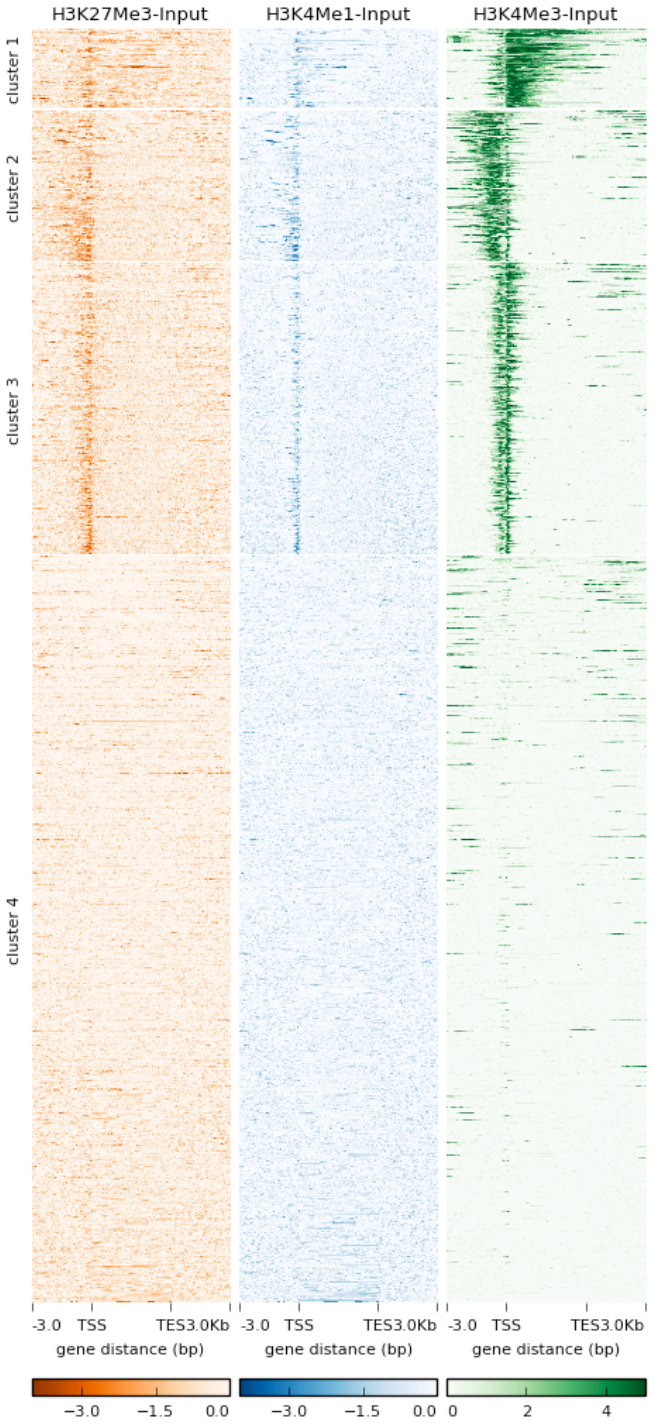





Read Count Frequency
0.00025
0.00020
0.00015
0.00010
-3000
-1500
TSS
1500
3000
Genomic Region (5'->3')


Test data profile
genes19
genesX
H3K27Me3-Input
H3K4Me1-Input
H3K4Me3-Input
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-3.0
TSS
TES
3.0Kb


A
B
RARRES1
NUSAP1
MYBL2
CENPE
MELK
UGT8
UBE2C
GABRP
TFAP2B
APOBEC3B
INAVA
Inflammatory Breast Carcinoma
CCN5
KRT81
PDK1
MSLN
FOXA1
CXCL14
ERBB4
Basal-Like Breast Carcinoma
SCGB1D2
MMP1
S100A7
EZH2
IDO1
PGR
Noninfiltrating Intraductal Carcinoma
PBK
ADH1B
AURKA
CHEK1
CAMP
FOXM1
MMP12
PAX6
TRAT1
ADIPOQ
MARCO
ISG20
S100A9
TOP2A
NEK2
GZMB
Latent Tuberculosis
S100A8
Carcinoma of urinary bladder, invasive
CXCL11
TBC1D9
CXCL10
GNLY
CCNB1
SKA1
LAG3
AGR2
CCL8
AURKB
HLA-DQA1
CEP55
PIR
SLC16A4
BMP4
fold change
size


H3K27Me3-Input
H3K4Me1-Input
H3K4Me3-Input
cluster 1
cluster 2
cluster 3
cluster 4
-3.0
TSS
TES3.0Kb
gene distance (bp)
−3.0
−1.5
0.0
0
2
4

- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources

The Epigenomics Roadmap Project

http://www.roadmapepigenomics.org/

- Reference human epigenomes
- DNA methylation, histone modifications, chromatin accessibility and small RNA transcripts
- Stem cells and primary *ex vivo* tissues
- 111 tissue and cell types
- 2,804 genome-wide datasets

Further reading

- Impact of artefact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Carrol et al, Front. Genet. 2014
- Impact of sequencing depth in ChIP-seq experiments. Jung et al, NAR 2014
- ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Landt et al, Genome Res. 2012
- http://genome.ucsc.edu/ENCODE/qualityMetrics.html#definitions
- https://www.encodeproject.org/data-standards

Resources for broad region analysis

- https://omictools.com/peak-calling-category
- https://www.encodeproject.org/chip-seq/histone/

Questions?

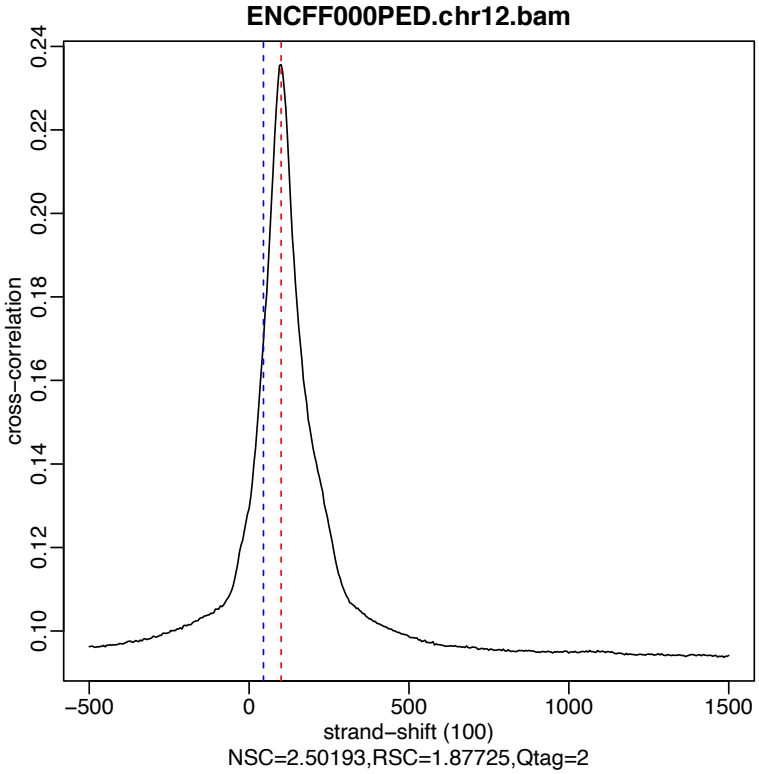
agata.smialowska@nbis.se

- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources
- Exercise overview

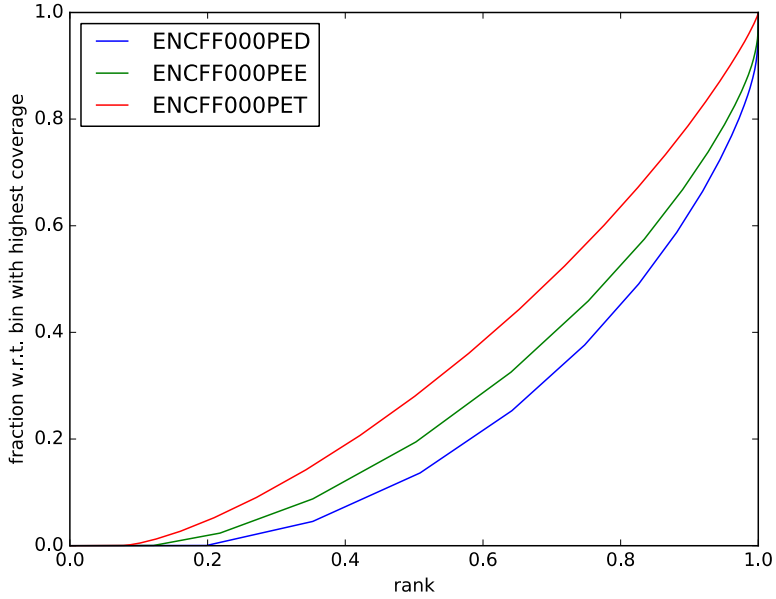
Exercise - ChIP-seq data processing

- 1. Quality control
- 2. Alignment preprocessing
- 3. Peak calling
- 4. Exploratory analysis (sample clustering)
- 5. Visualisation

Did my ChIP work?

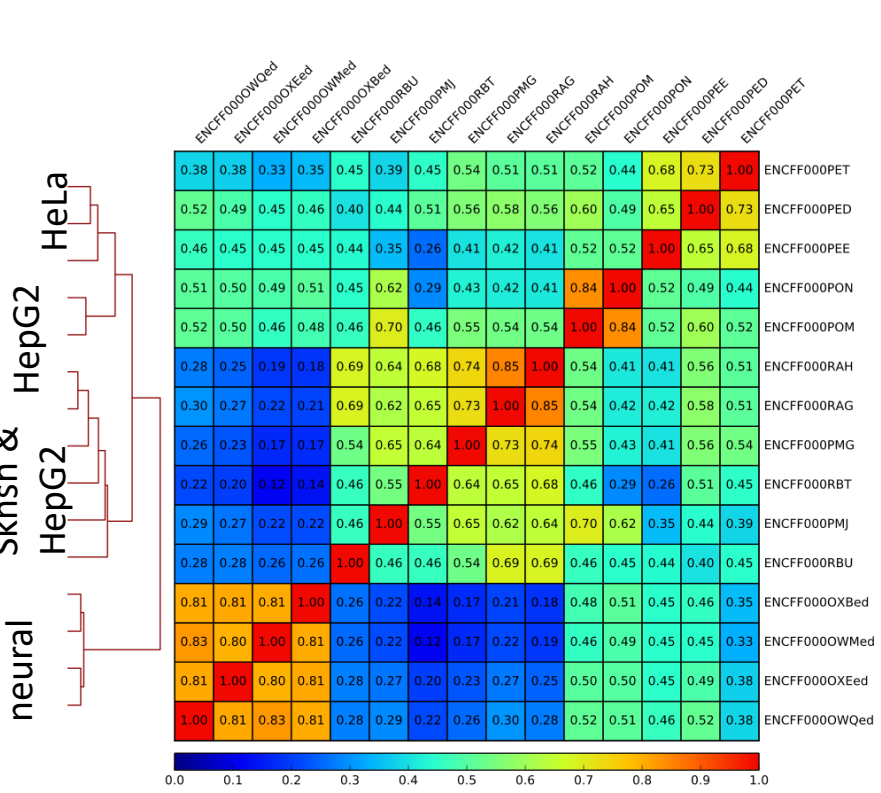


Cross-correlation

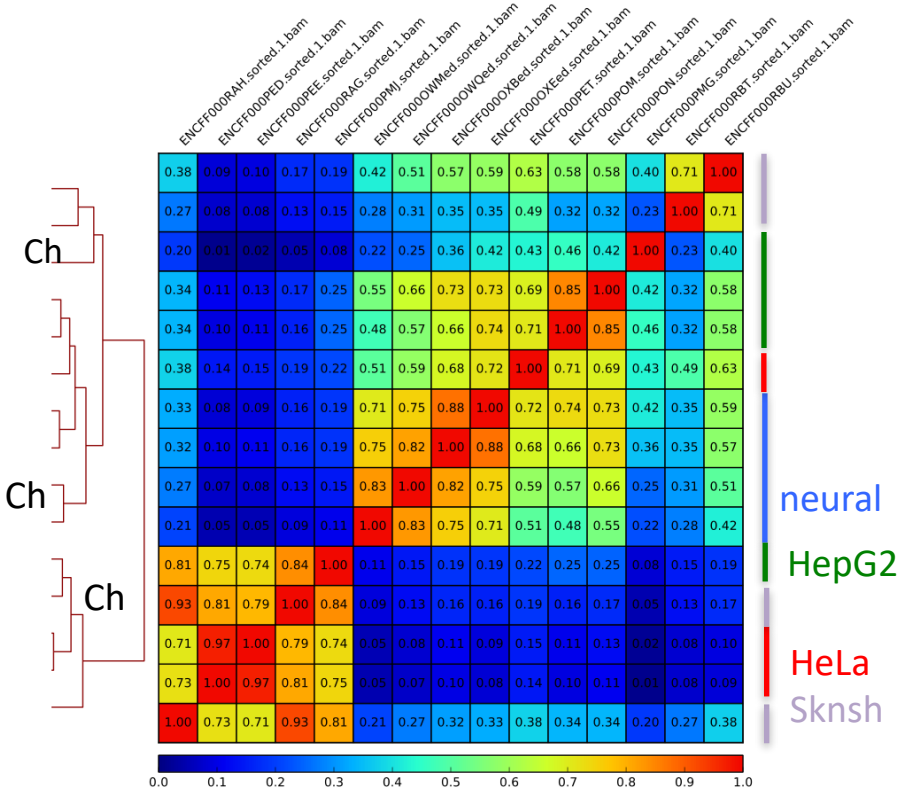


Cumulative enrichment

Exploratory analysis



Clustering of libraries
by reads mapped in bins,
genome – wide (spearman)

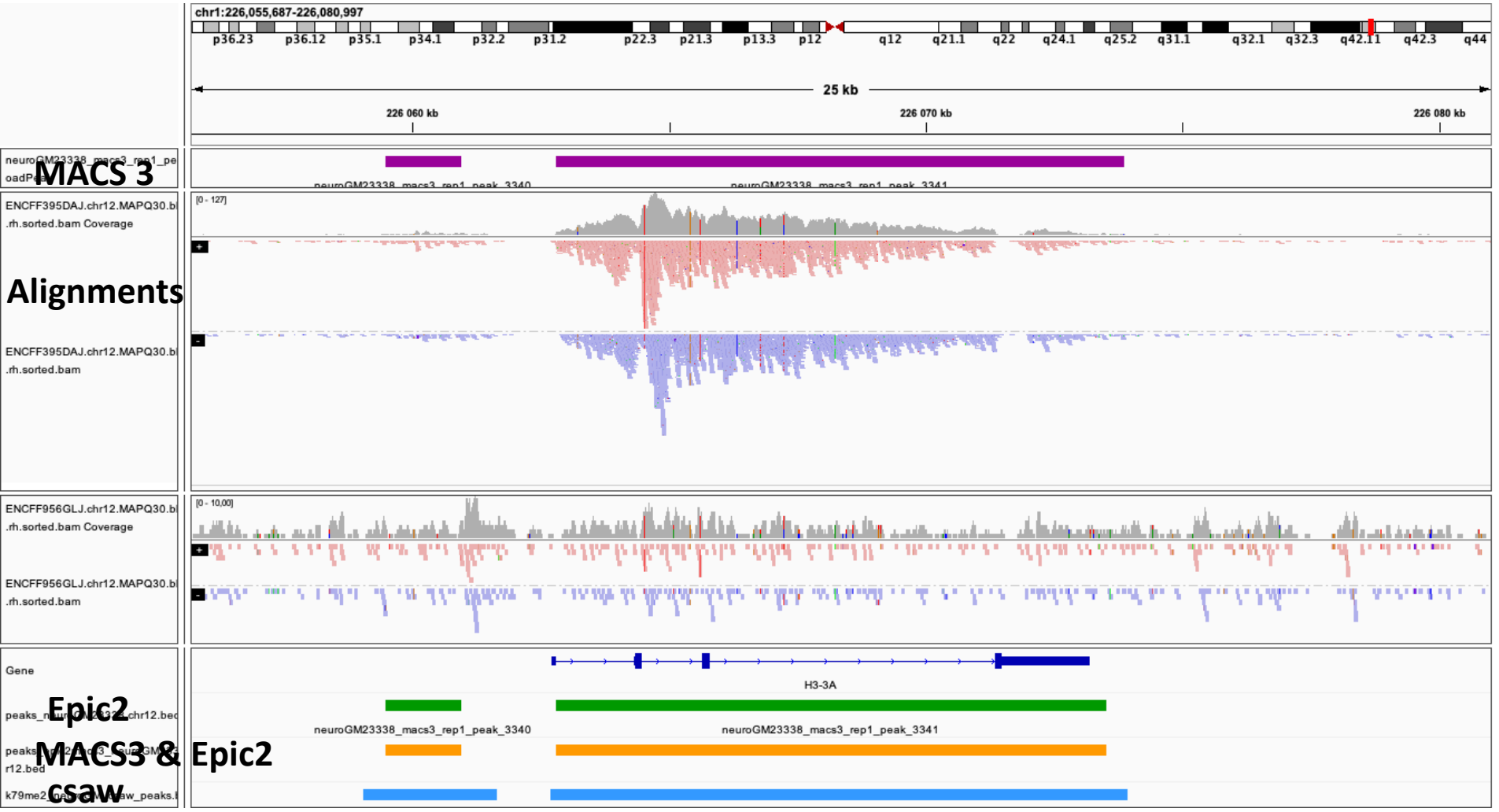


Clustering of libraries
by reads mapped in peaks
(pearson)

Exercise: broad peaks

- Model-based Analysis of ChIP-Seq (MACS) 3 in `broad peak` mode;
- csaw: Detection of differentially bound regions in ChIP-seq data with sliding windows, with methods for normalization and proper FDR control;
- epic2: focus on detection of clustered enrichment sites; ultraperformant reimplementation of SICER (speed, low memory overhead and ease of use)
- Good sequencing depth data from ENCODE; H3K79me2 (transcribed regions of active genes)

MACS 3, epic2, csaw



That’s all for now,

time to do some hands-on work

Library quality control and preprocessing

- FastQC / Prinseq
- Trim adapters if any adapter sequences are present in the reads (as determined by the QC)
- In some cases, you'll observe k-mer enrichment (especially if the data is ChIP-exo, a new variation of ChIP-seq) – it is not necessarily a bad thing, if sequence duplication levels are low; however it may indicate **low complexity of the library** – a warning sign that the enrichment in ChIP was not successful or the libraries are over-amplified (often the latter is the consequence of the former)

Mapping reads to the reference genome

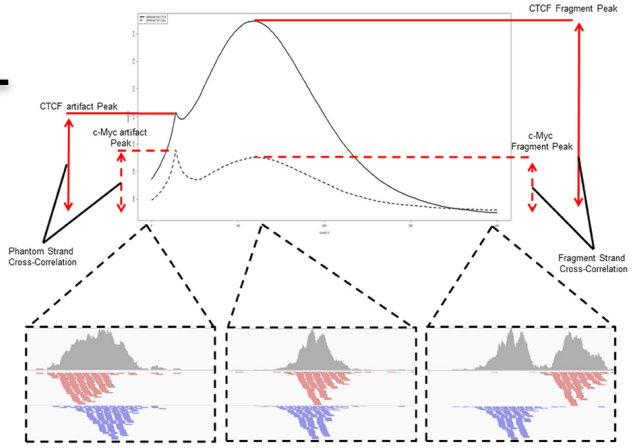
- Choose the right reference: assembly version (not always the newest is best) and type (primary assembly, or assembly from individual chromosome sequences + non-chromosomal contigs; not the top level assembly); choose the matching annotation file (GTF, GFF)
- Read mapping: **global alignment**
- Mappers (= aligners): Bowtie, BWA, BBMap, Novoalign, … (lots of tools are available)

- Visualise data in genome browser
  - BAM files or tracks (wig, bedgraph, bigWig)
  - Local (IGV) or web-based (UCSC genome browser)
  - Data quality assessment

Cross-correlation profiles, RSC and NSC

- Metrics to quantify the fragment length signal and the ratio of fragment length signal to read length signal
- Relative Cross Correlation (RSC) - ChIP to artifact signal

$$\frac{CC(\text{Fragment length}) - \min(CC)}{CC(\text{read length}) - \min(CC)}$$



- Normalised Cross Correlation (NSC)

$$\frac{CC(\text{Fragment length})}{\min(CC)}$$

- TFs: fragment lengths are often greater than the size of the DNA binding event, the distinct clustering of (+) and (-) reads around this site is very apparent
- NSC>1.1 (higher values indicate more enrichment; 1 = no enrichment)
- RSC>0.8 (0 = no signal; <1 low quality ChIP; >1 high enrichment
- Broad peaks: this clustering may be more diffuse (fragment length < peak)