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

Epigenomics Data Analysis Workshop

Stockholm, 24 November 2020

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)

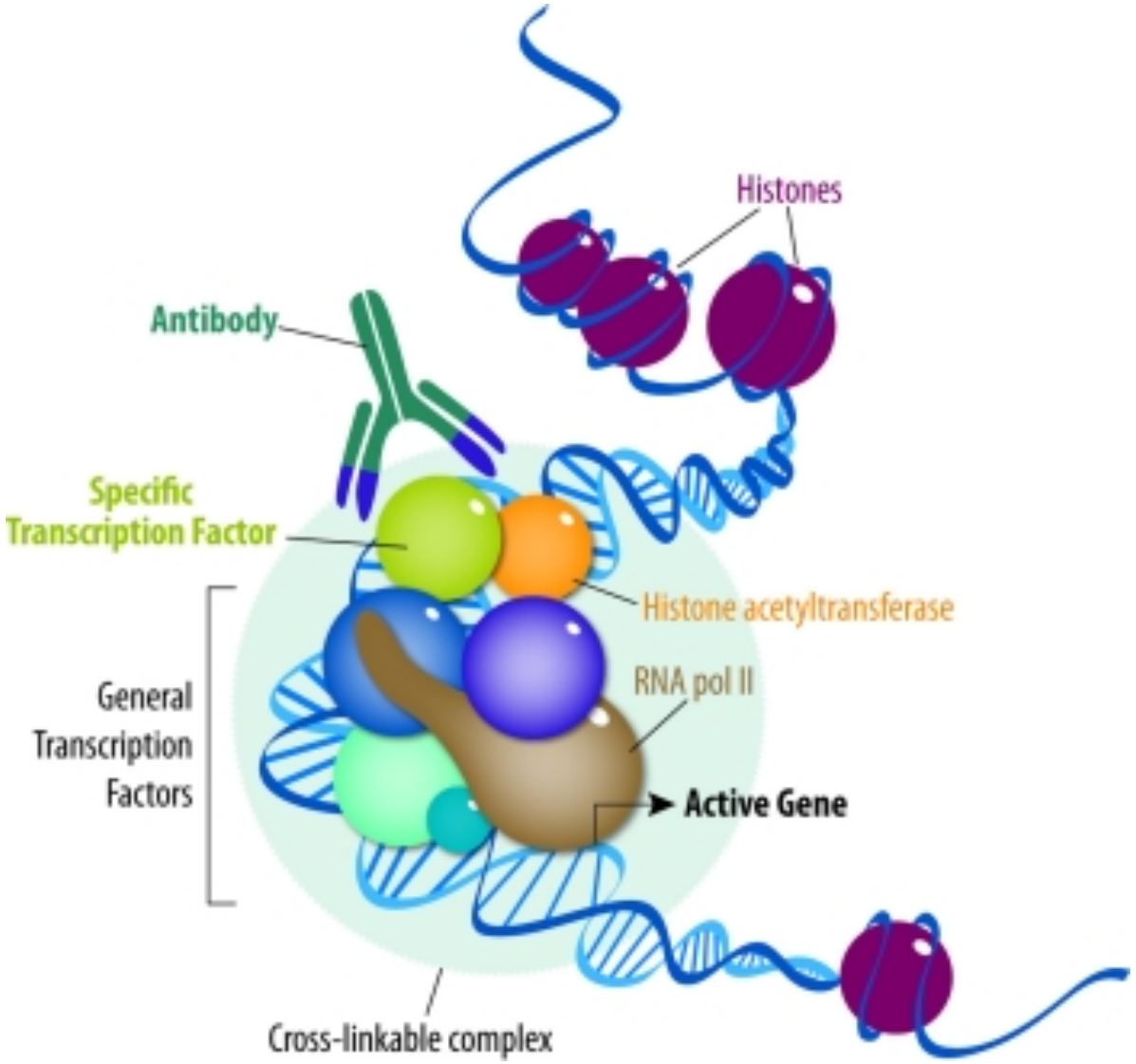
First observed by
H. Muller
1930

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



RnDsystems

www.pollev.com/AGATASMIALOW506

Give TWO keywords which you associate with ChIP-seq

Workflow of a ChIP-seq study

design study

obtain input chromatin

perform precipitation

construct library

sequence library

bioinformatic analysis

Wet lab

Experiment design

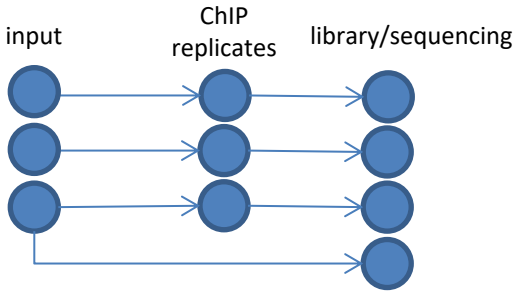
- Sound experimental design: replication, randomisation and blocking (R.A. Fisher, 1935)
- In the absence of a proper design, it is essentially impossible to partition biological variation from technical variation
- Sequencing depth: depends on the structure of the signal; cannot be linearly scaled to genome size
- Single- vs. paired-end reads: PE improves read mapping confidence and gives a direct measure of fragment size, which otherwise has to be modelled or estimated
- Other factors: Cross-linker choice, chromatin fragmentation method, antibody, …

Experiment design

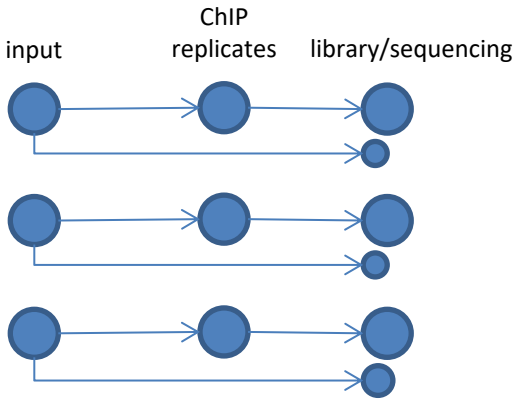
Ideal design:

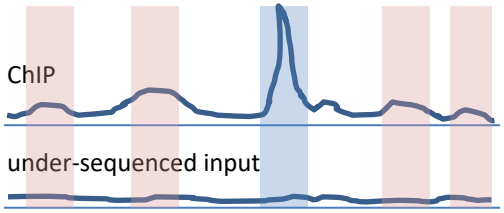
Each sample has a matched input
Input sequenced to a comparable depth as IP sample

X

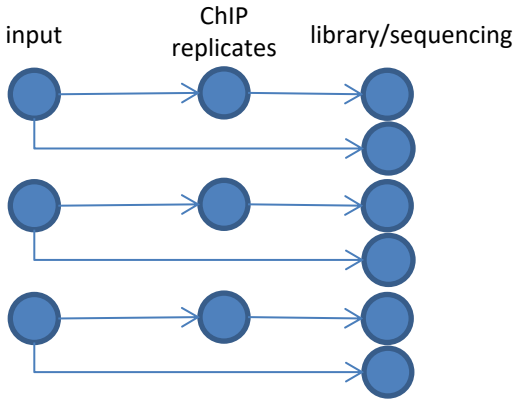


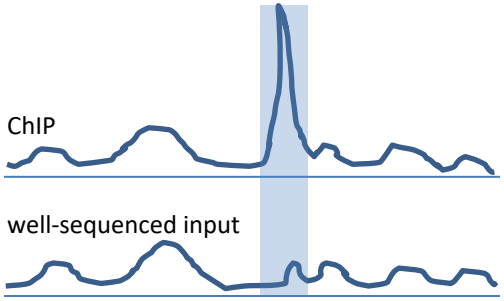
X



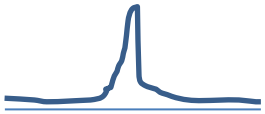
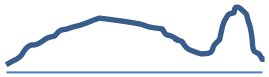


✓





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

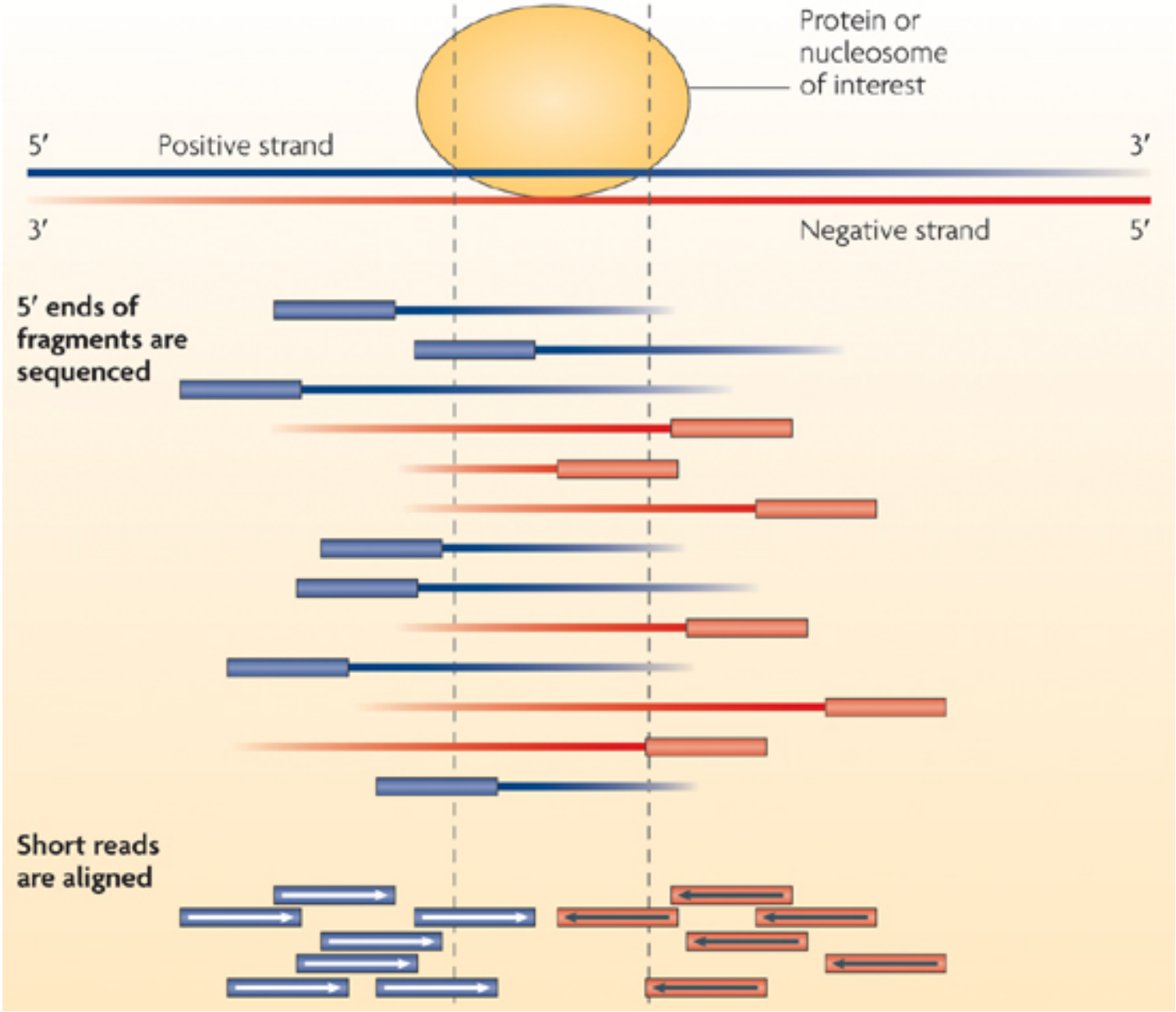
- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources

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

Which part of the analysis, in your opinion, is of key importance to analysis and interpretation of ChIP-seq data?

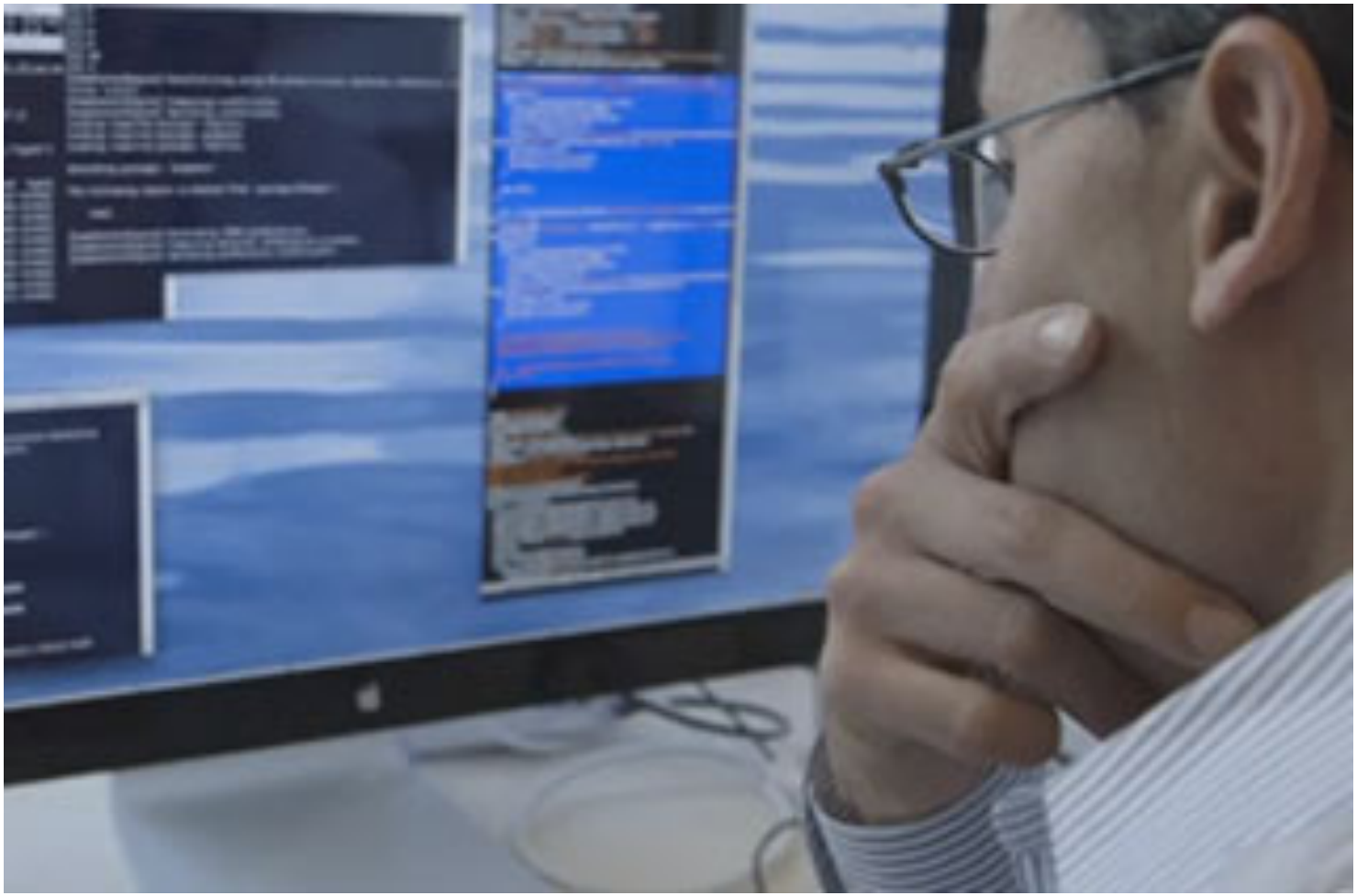
QC	**A**
peak calling	**B**
read alignment	**C**
functional annotation of peaks	**D**

Chromatin = DNA + proteins



Park, Nature Rev Genetics, 2009

Data analysis



Workflow of a ChIP-seq study

design study

obtain input chromatin

perform precipitation

construct library

sequence library

Wet lab

library quality control
filter sequences

align sequences

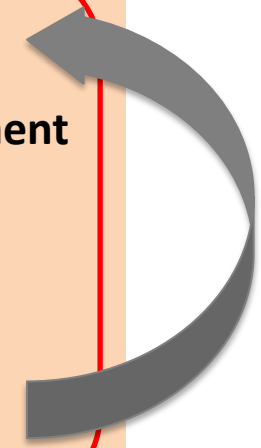
filter alignments

identify peaks / regions of enrichment

assess data quality

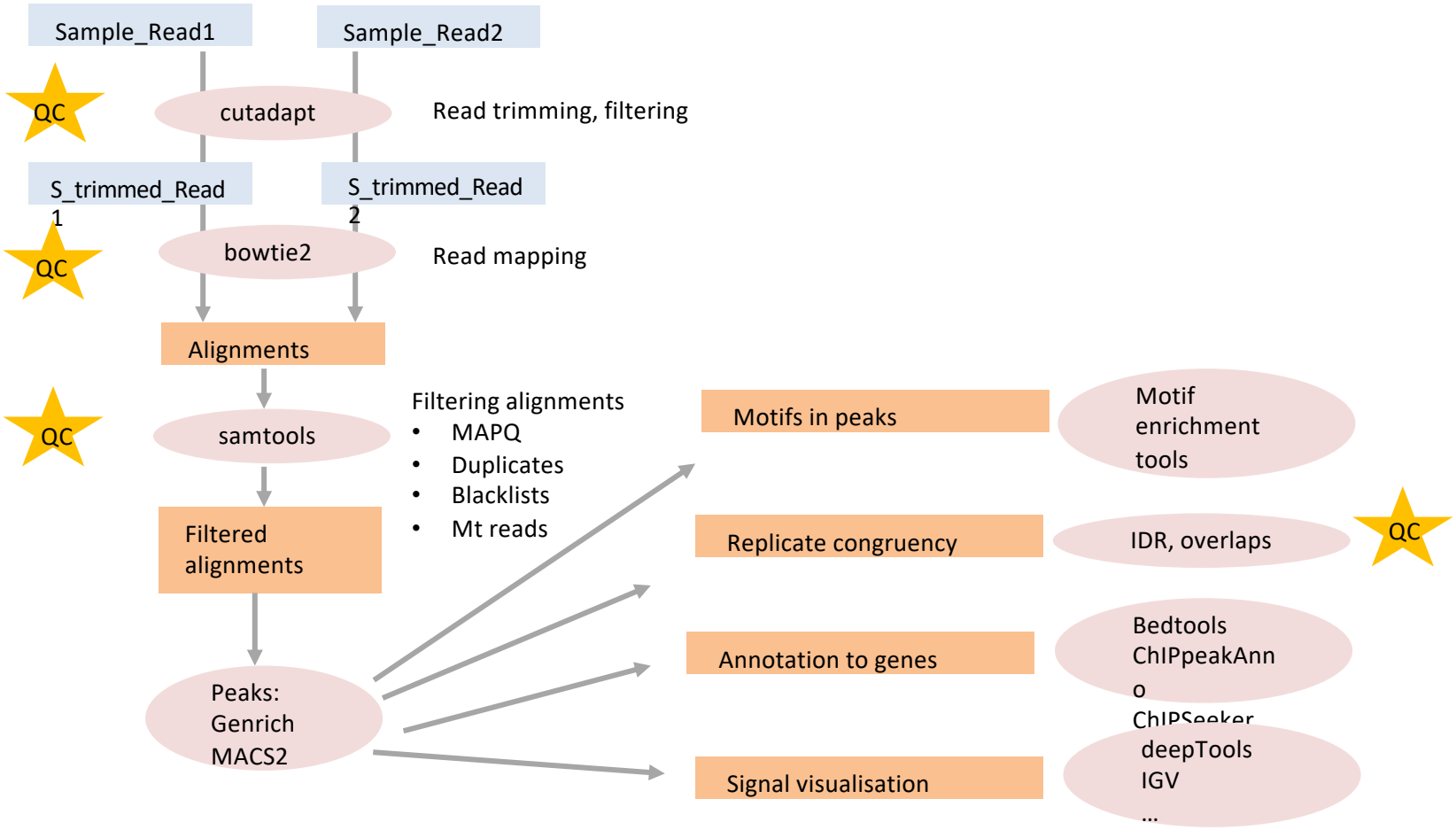
understand the data / results

downstream analyses



Iterative process

Analysis workflow



- 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)?

Word of caution!

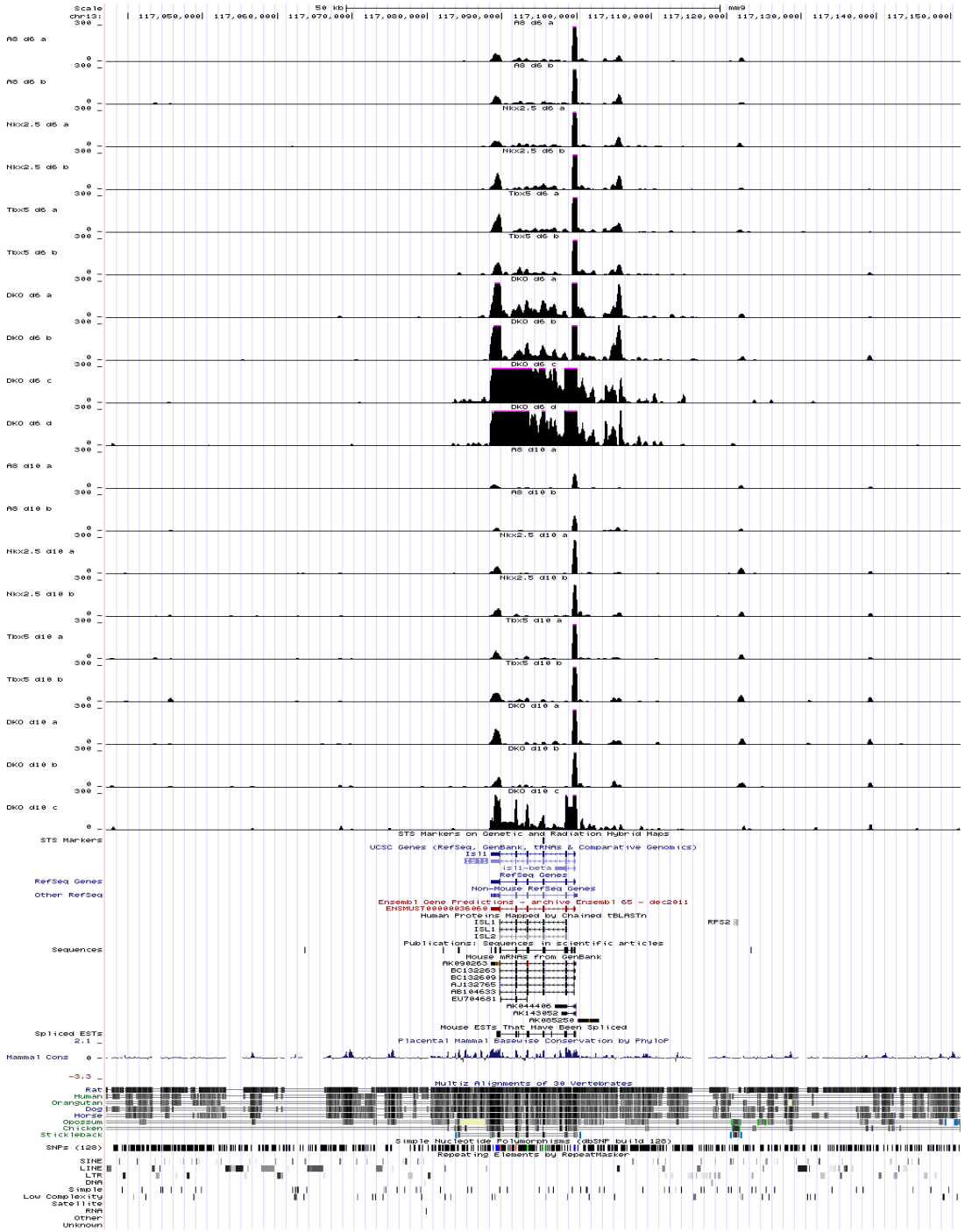
ChIP-seq experiments are more unpredictable than RNA-seq!

Inconsistency sources:

- chromatin structure
- PCR over-amplification
- non-specific antibody
- other things?

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

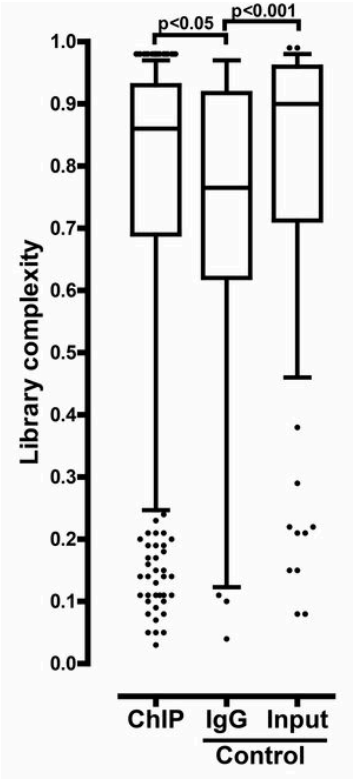
Confounding factors

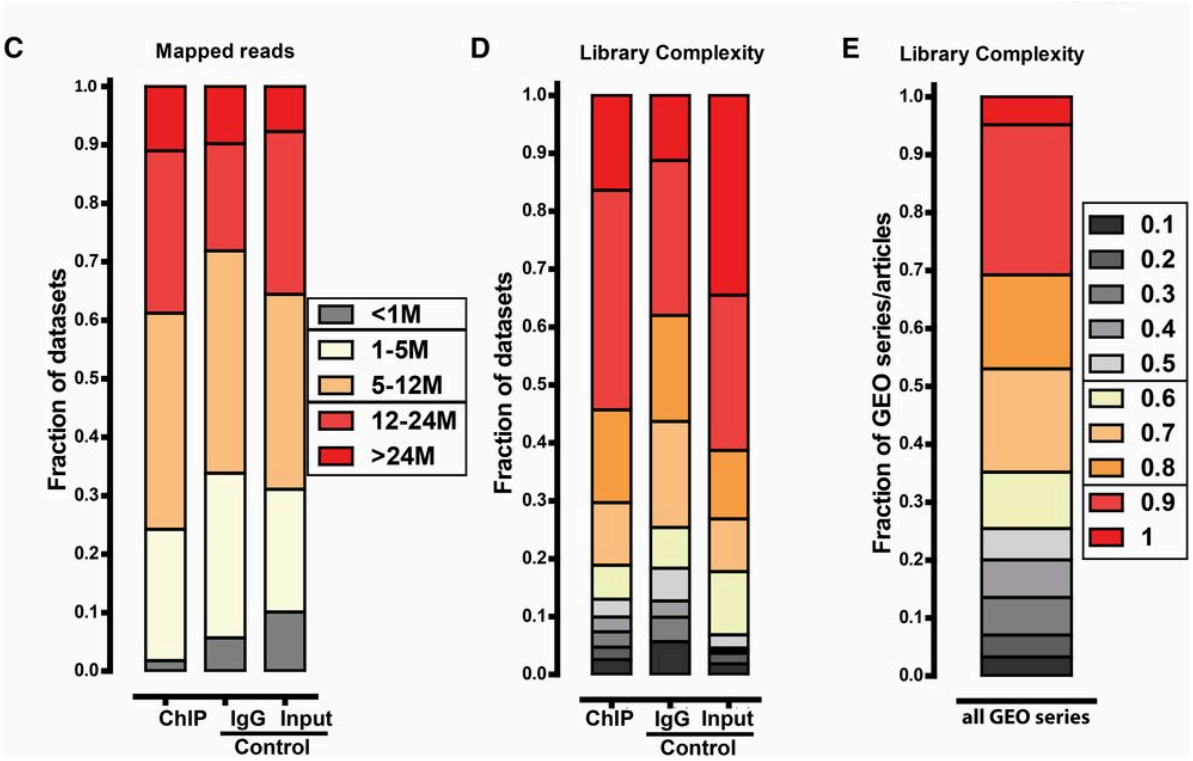
Under-sequenced libraries

…

How do I know my data is of good quality?

Library complexity

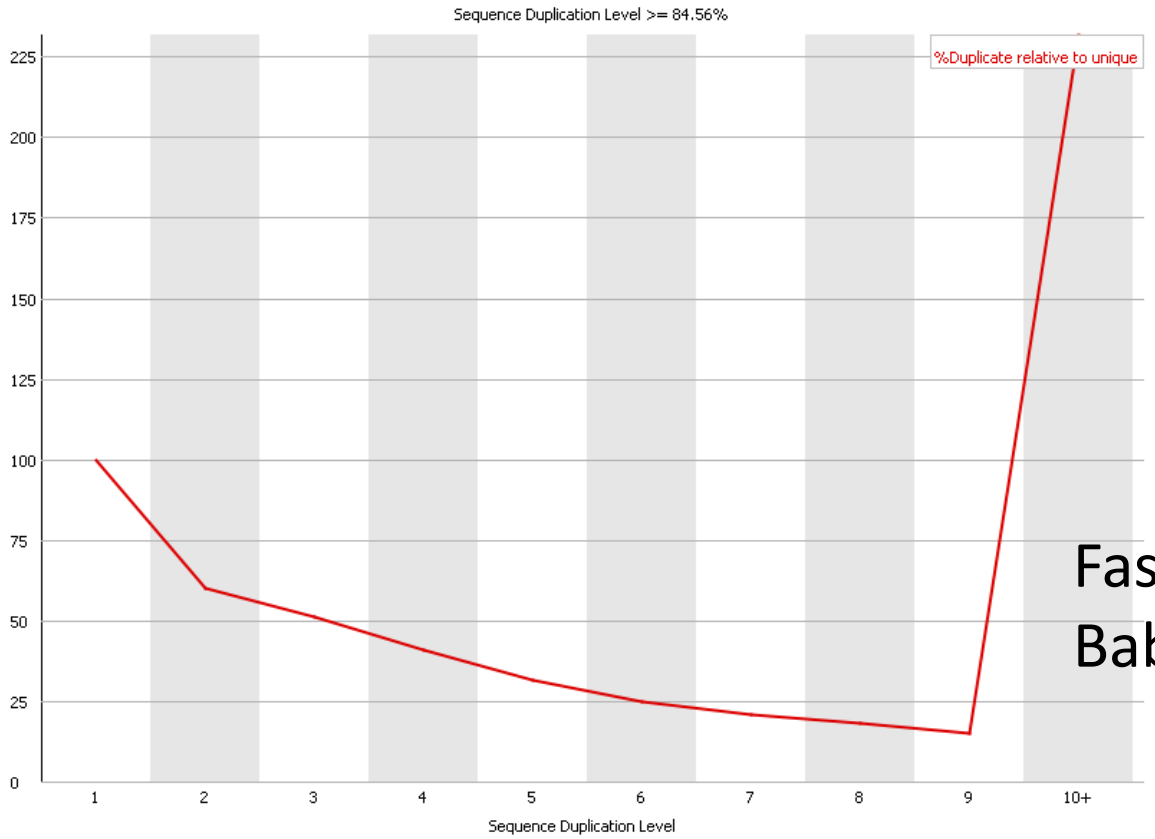




Marinov et al, G3 2013

Quality control: tag uniqueness – library complexity metric

Sequence duplication level > 80% (low complexity library)



FastQC
Babraham Institute

NRF: Non-redundant fraction (of reads): proportion of unique tags / total

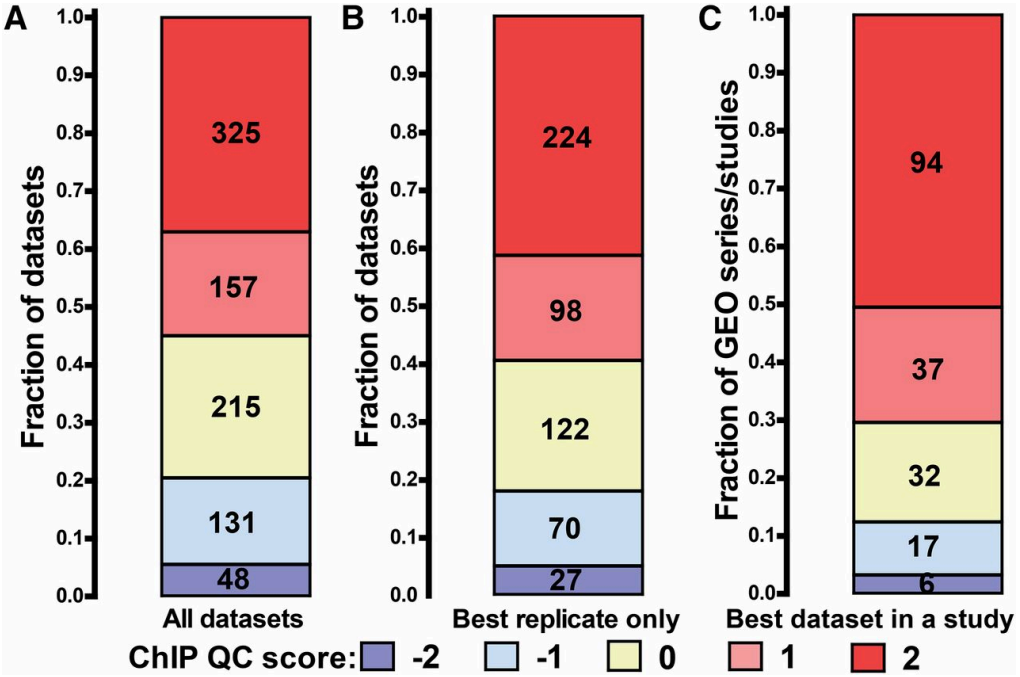
How do I know my data is of good quality?

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

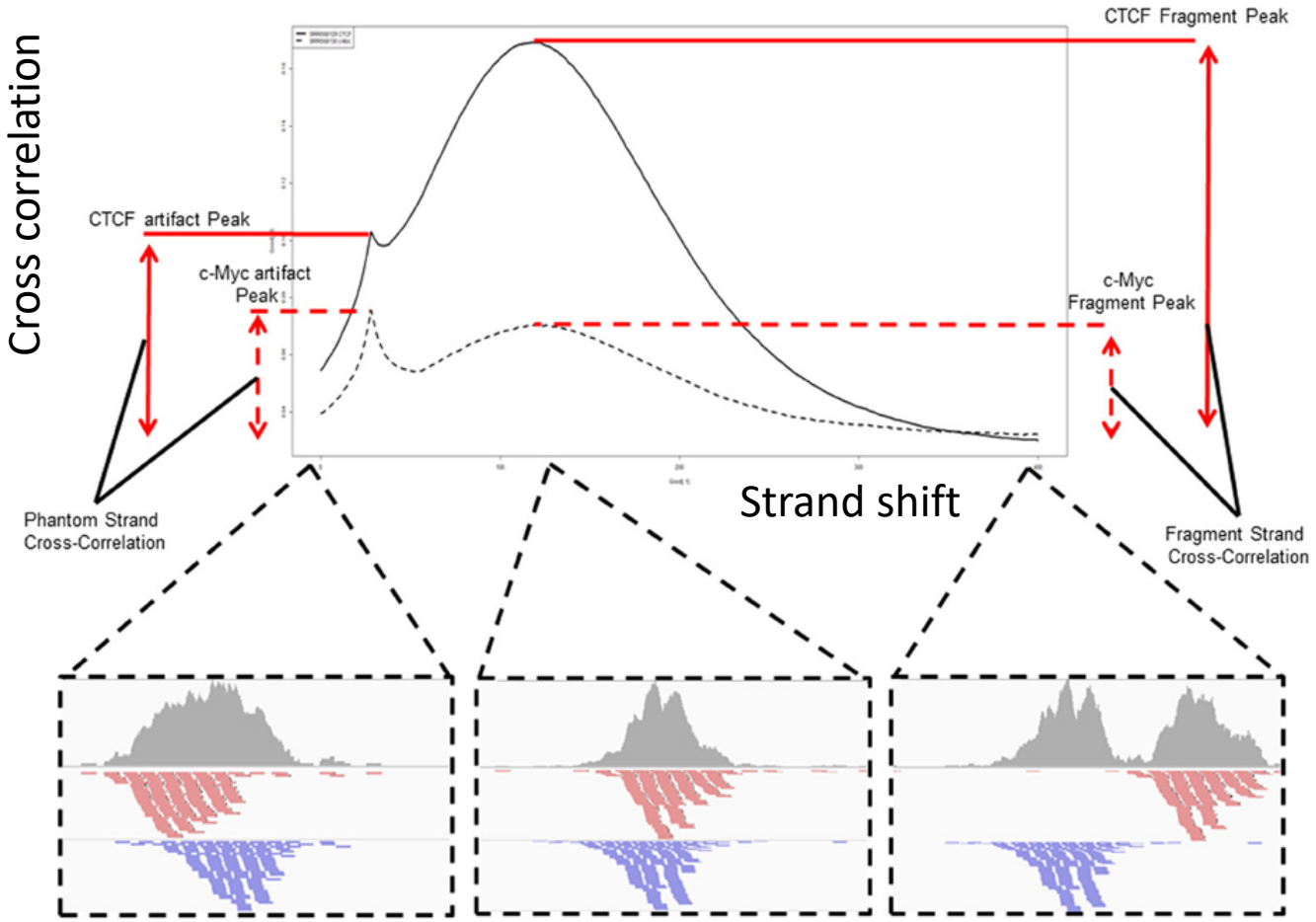
Large-scale quality analysis of published ChIP-seq data sets:
20% low quality
25% intermediate quality
30% inputs have metrics similar to IPs



Marinov et al, G3 2013

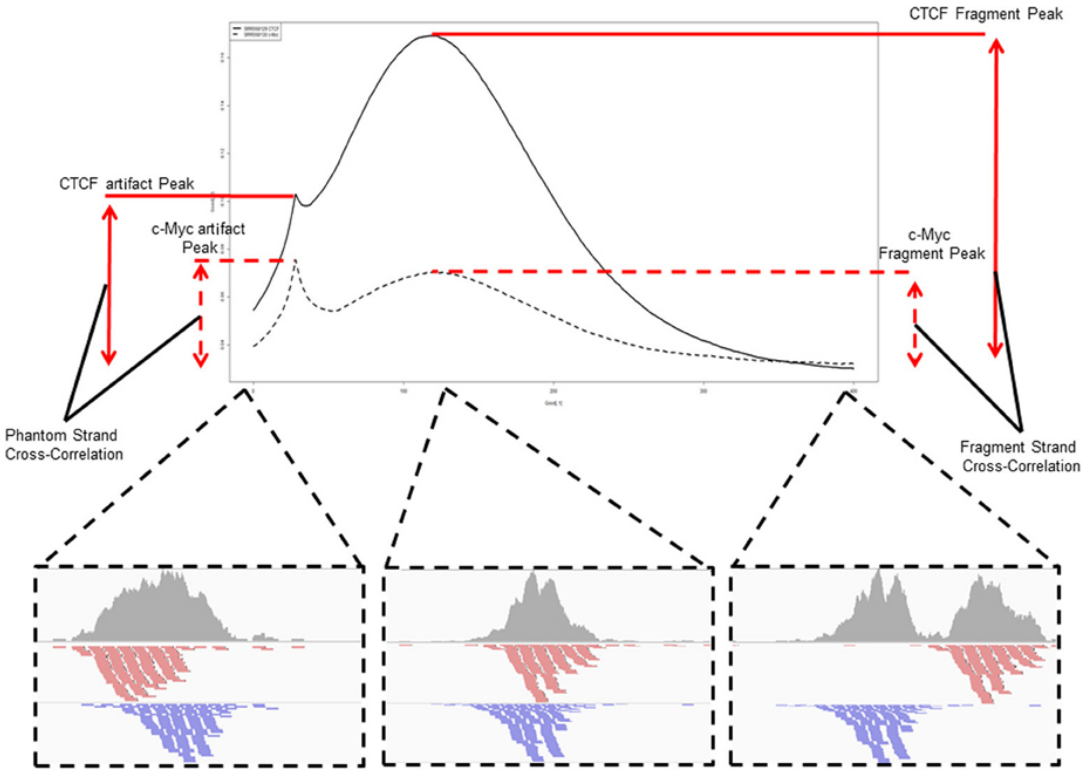
Strand cross-correlation

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



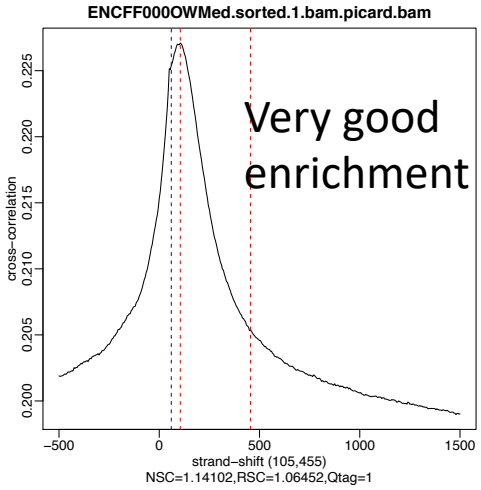
$$NSC = \frac{\text{Max CC value (fLen)}}{\text{Min CC}}$$

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

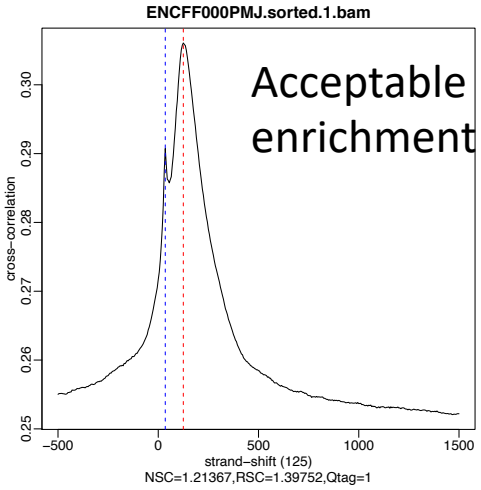
Carroll et al, Front Genet 2014

Cross-correlation plots

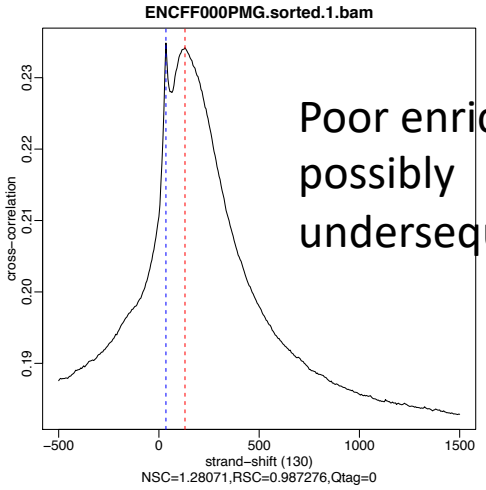
ChIP



Very good enrichment

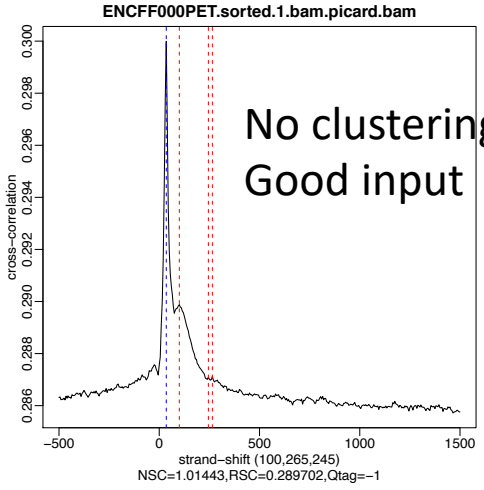


Acceptable enrichment

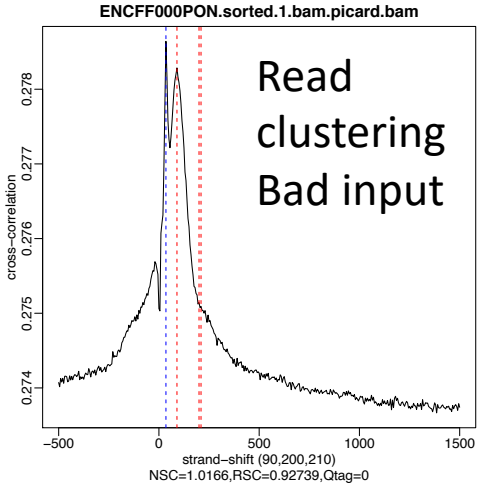


Poor enrichment, possibly undersequenced

Input

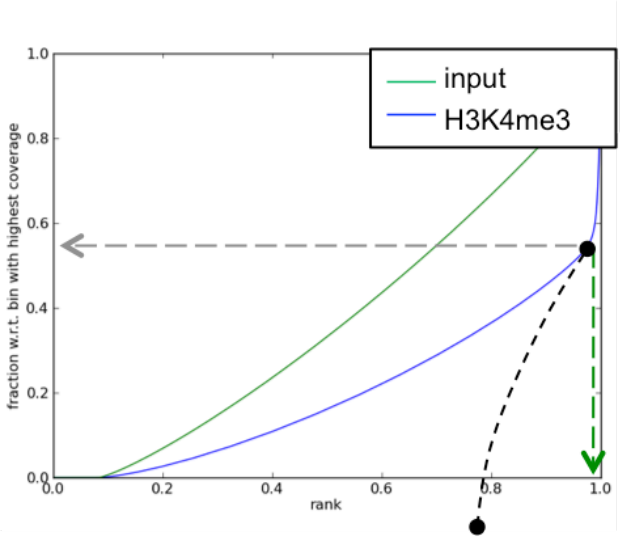


No clustering
Good input



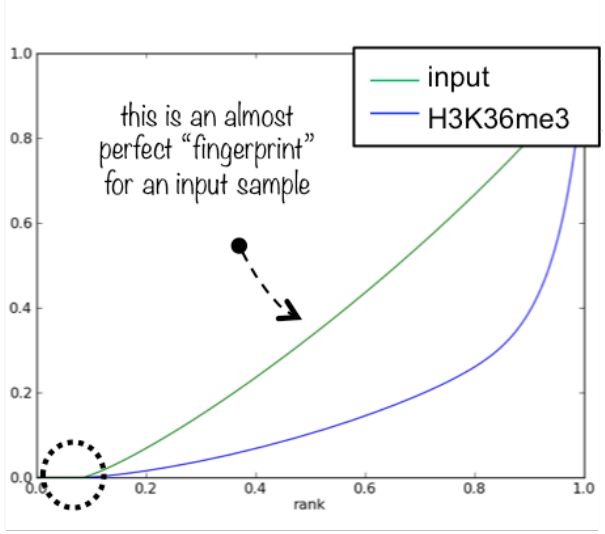
Read clustering
Bad input

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

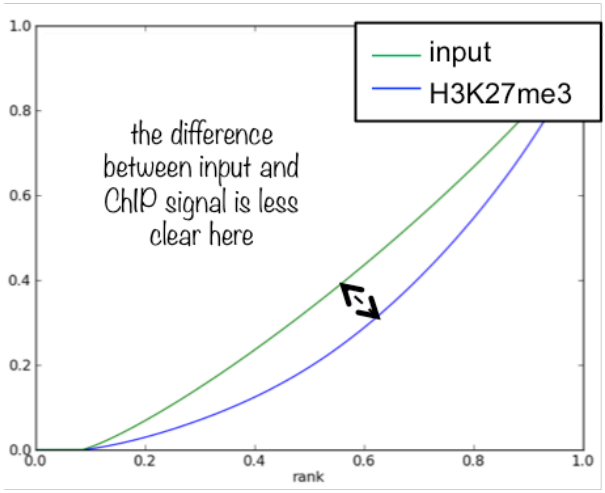


when counting the reads contained in **97%** of all genomic bins, only ca. **55%** of the maximum number of reads are reached, i.e. 3% of the genome contain a very large fraction of reads!

→ 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

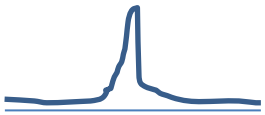
→ it is more difficult to distinguish input and ChIP, it does not mean, however, that this particular ChIP experiment failed

http://deeptools.readthedocs.org
Diaz et al, Genome Biol 2012

Peak calling

appropriate methodologies depend on data type

Transcription Factors



point source

MACS2

Chromatin Remodellers

Histone marks

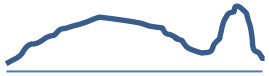
mixed

MACS2 in broad mode
windows approaches
Epic2 (SICER)

Chromatin Remodellers

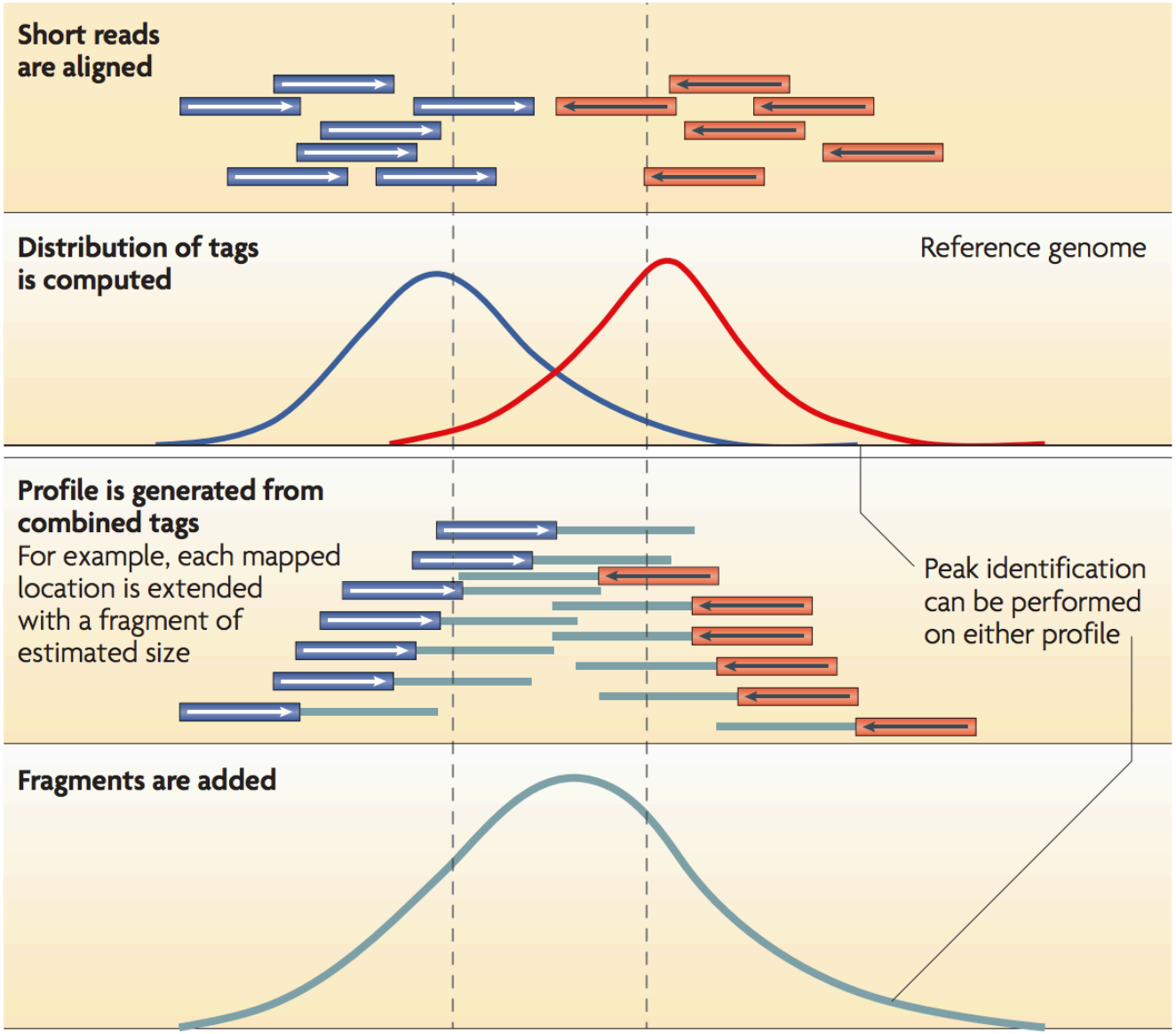
Histone marks

RNA polymerase II



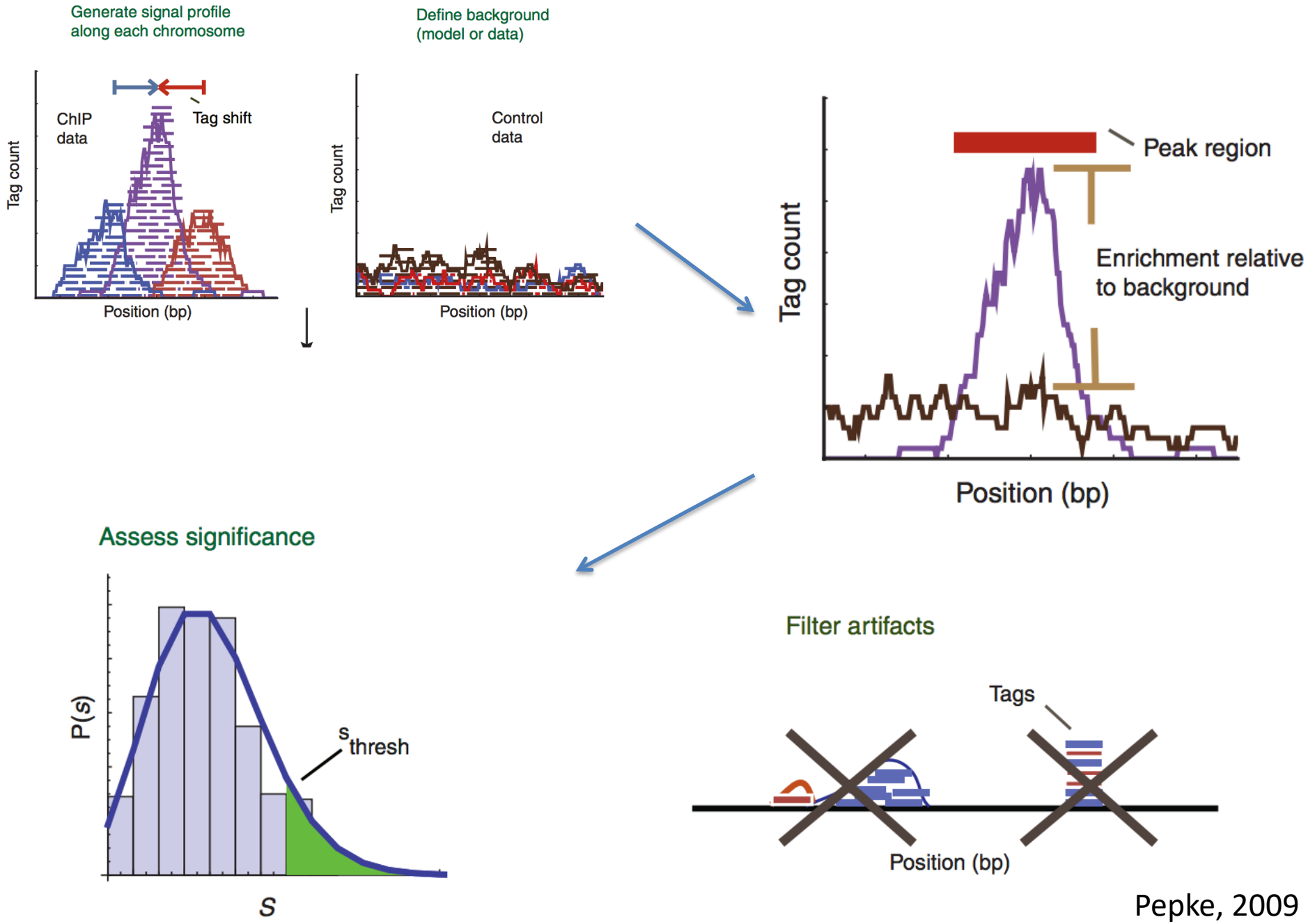
broad

Principle of peak detection



Symmetry in reads mapped to opposite DNA strands

Computation of enrichment model

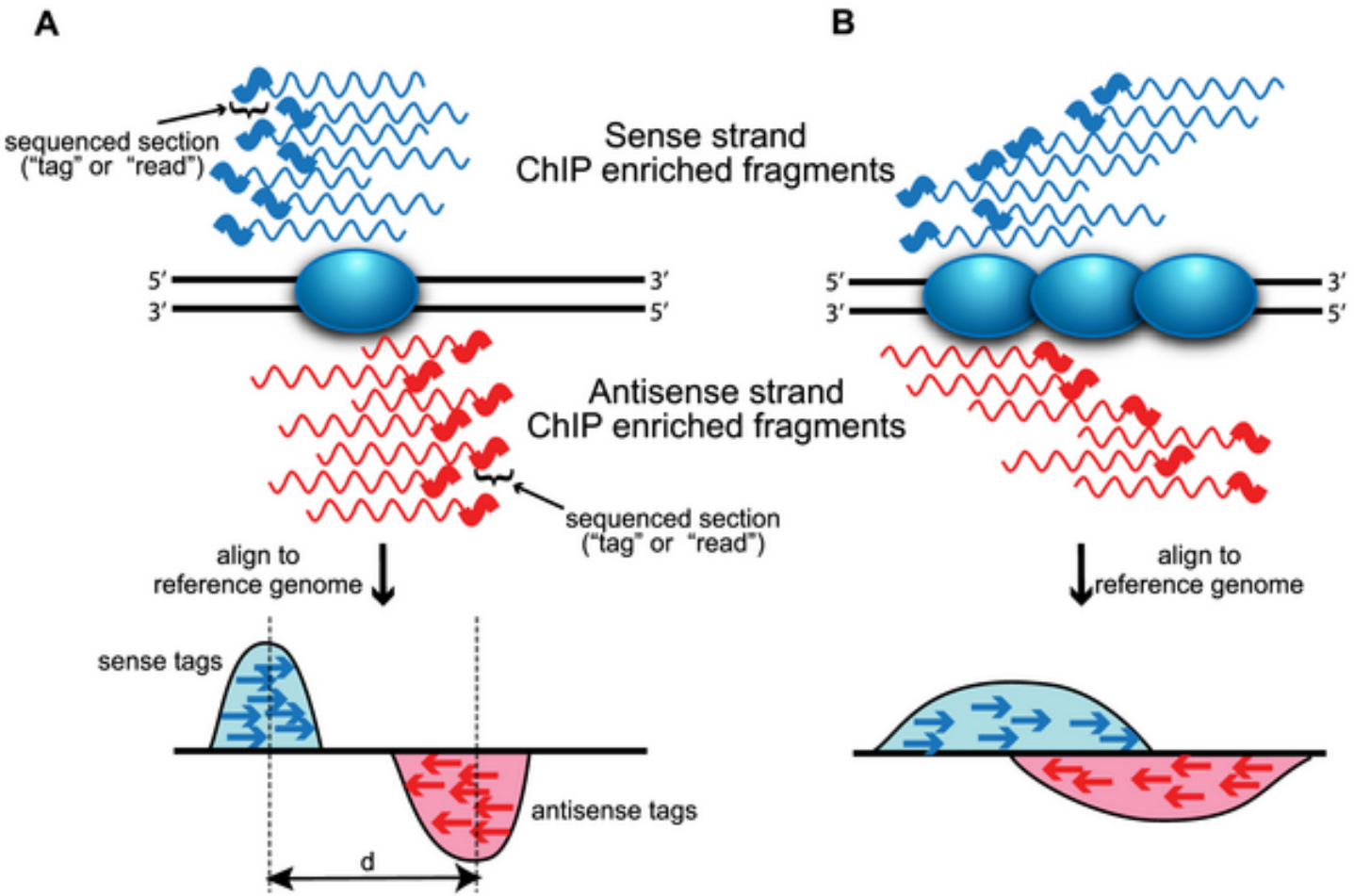

Generate signal profile
along each chromosome
ChIP
data
Tag shift
Tag count
Position (bp)
Define background
(model or data)
Control
data
Tag count
Position (bp)
Peak region
Enrichment relative
to background
Tag count
Position (bp)
Assess significance
P(s)
s
thresh
S
Filter artifacts
Tags
Position (bp)

Pepke, 2009

Point-source vs. broad peak detection

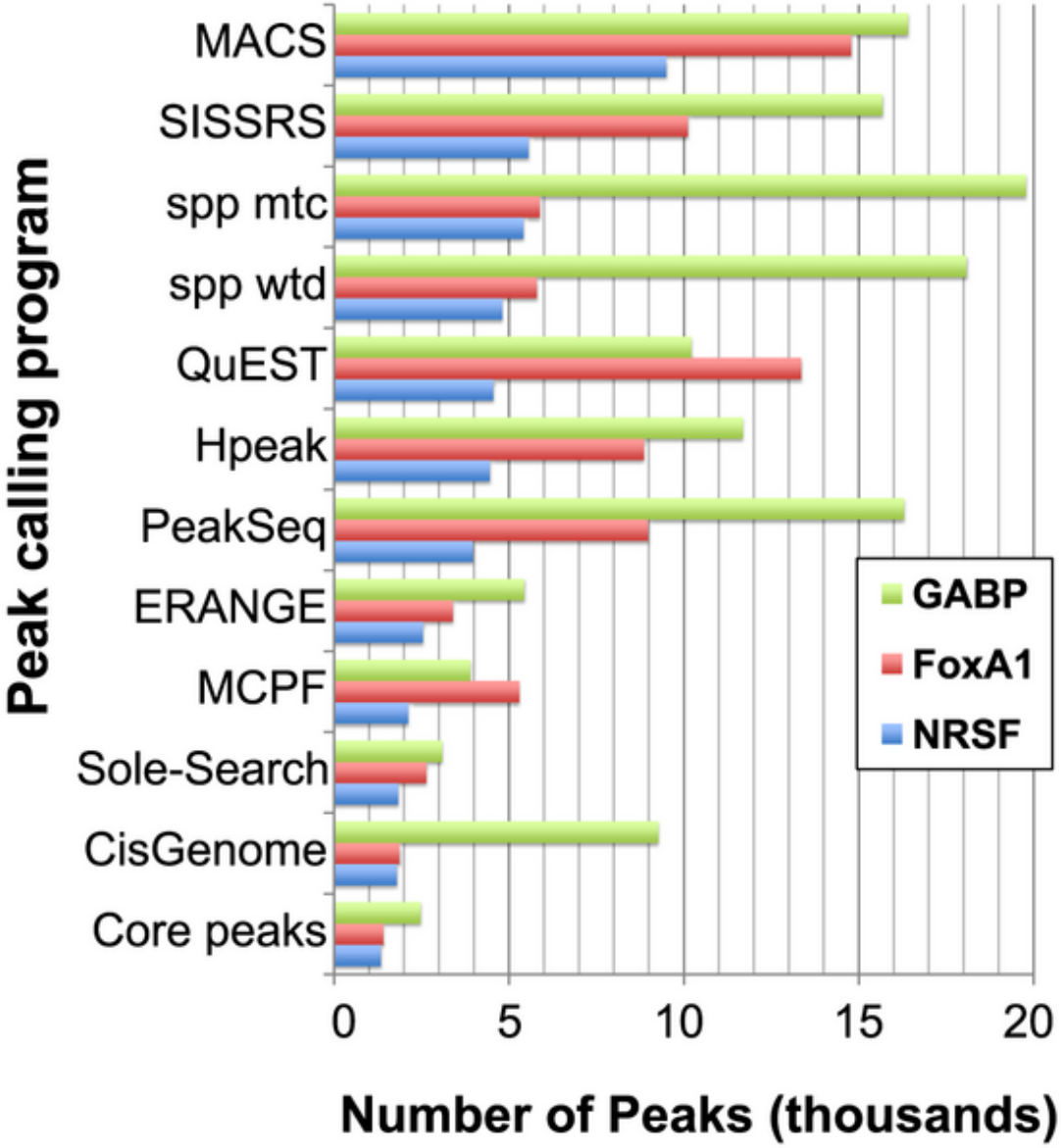
Sequence-specific binding (TFs)

Distributed binding (histones, RNApol2)

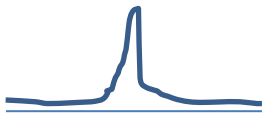


Wilbanks 2010

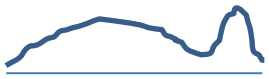
Comparison of peak calling algorithms



Peak overlap (Ho et al, 2012)



> 50 %



20 %

Wilbanks 2010

Comparison of peak calling algorithms

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

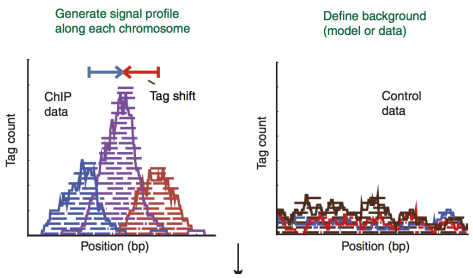
Features that define the best ChIP-seq peak calling algorithms

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

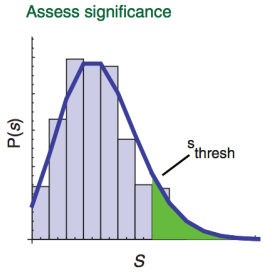
Corresponding author: Katherine S Pollard, Gladstone Institutes, San Francisco, CA 94158, USA. Tel.: 415-734-2711. Fax: 415- 355-0141; E-mail: katherine.pollard@gladstone.ucsf.edu

- Model-based Analysis for ChIP-Seq version 2 (MACS2)
- 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)



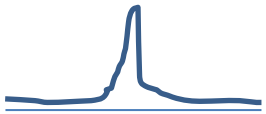
Testing candidates for significance

Thomas et al, 2017: serveyed 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

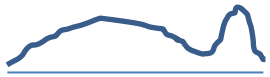
Bäst i test (2017)

- 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

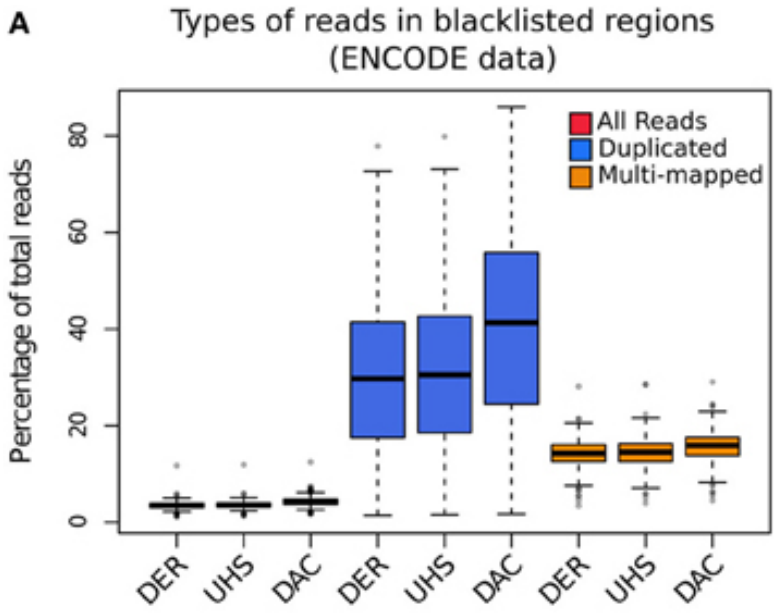


New players (2019 - 2020):

- 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

“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

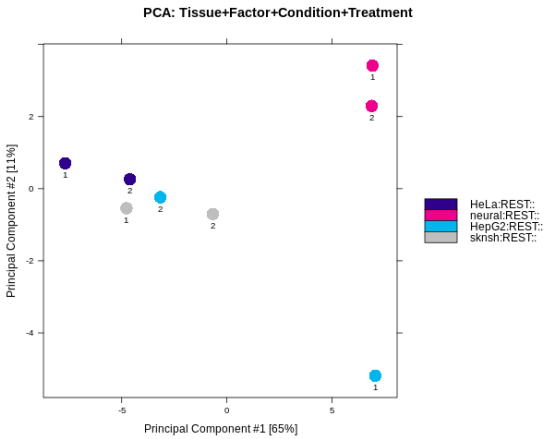
DER – Duke Excluded Regions
(11 repeat classes)
UHS – Ultra High Signal
(open chromatin)
DAC – consensus excluded regions

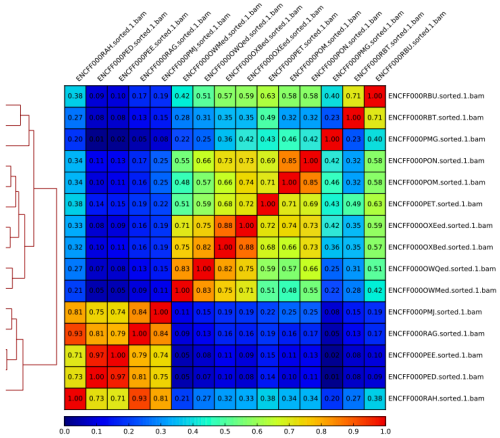
Carroll et al, Front Genet 2014

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)



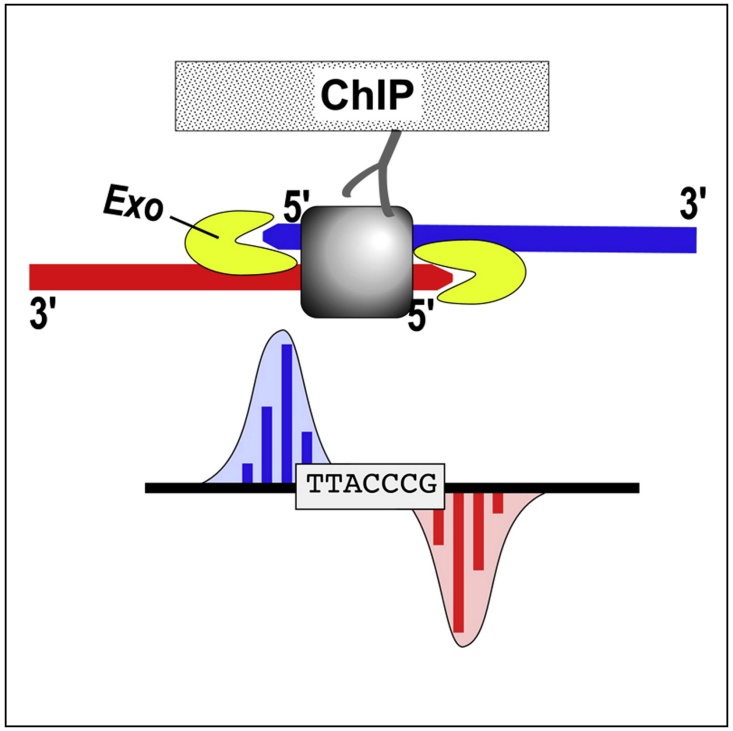


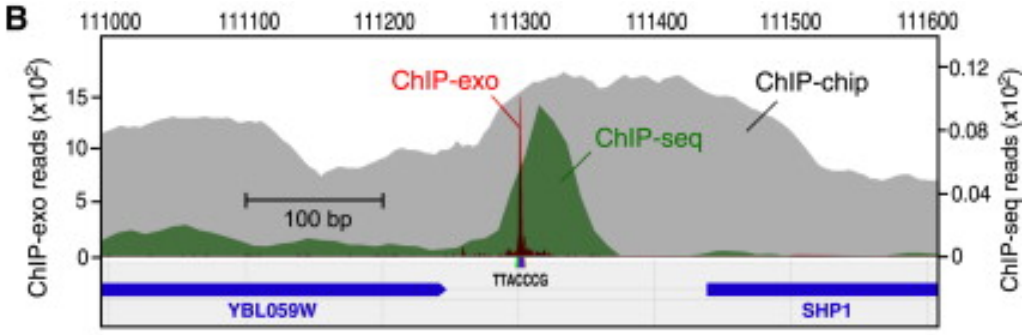
- Irreproducible discovery rate (IDR) - compare a pair of ranked lists of identifications (such as ChIP-seq peaks). These ranked lists should not be pre-thresholded, i.e they should provide identifications across the entire spectrum of high confidence/enrichment (signal) and low confidence/enrichment (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, 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)

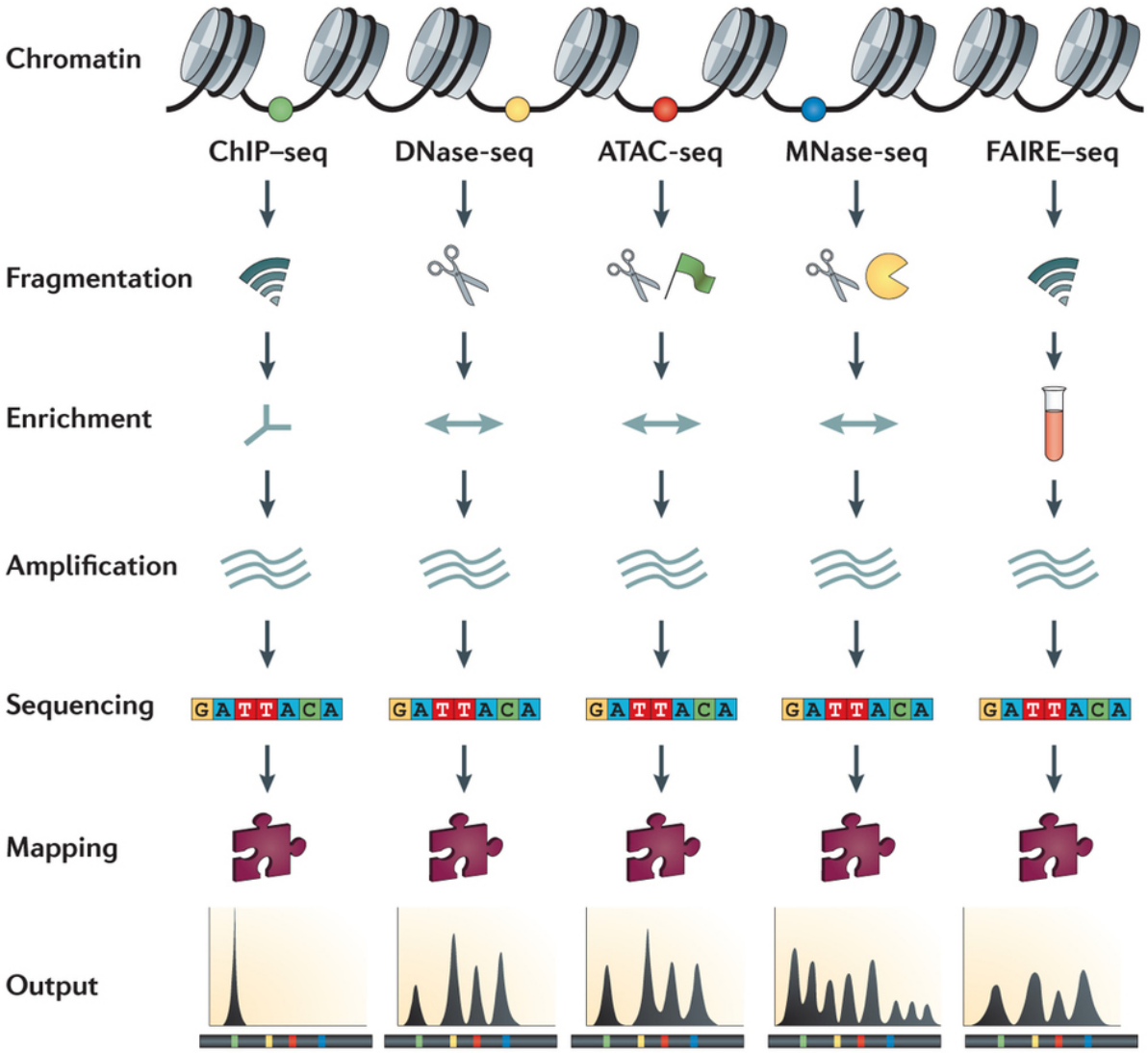
ChIP-exo: improvement in binding site identification





Rhee and Pugh, Cell 2011

Other functional genomics techniques



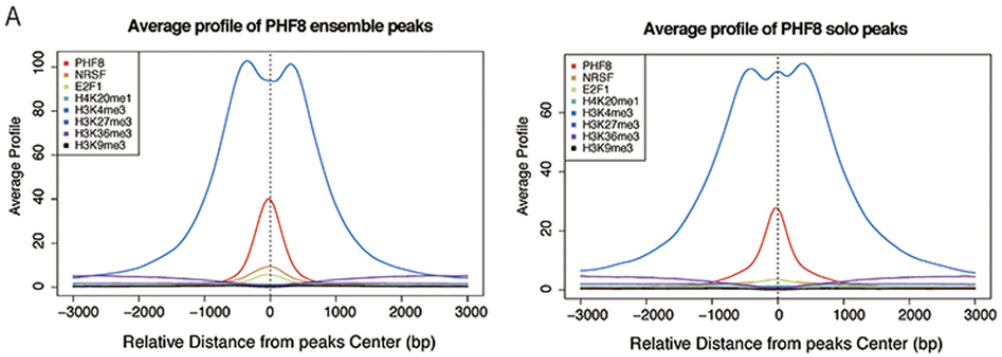
Clifford et al, Nature Rev Genet, 2014

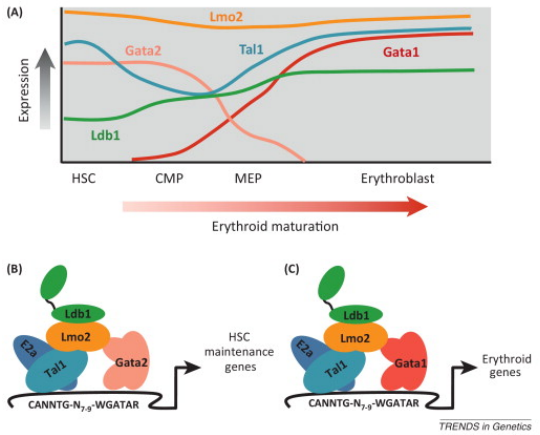
- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources

ChIPseq downstream analyses

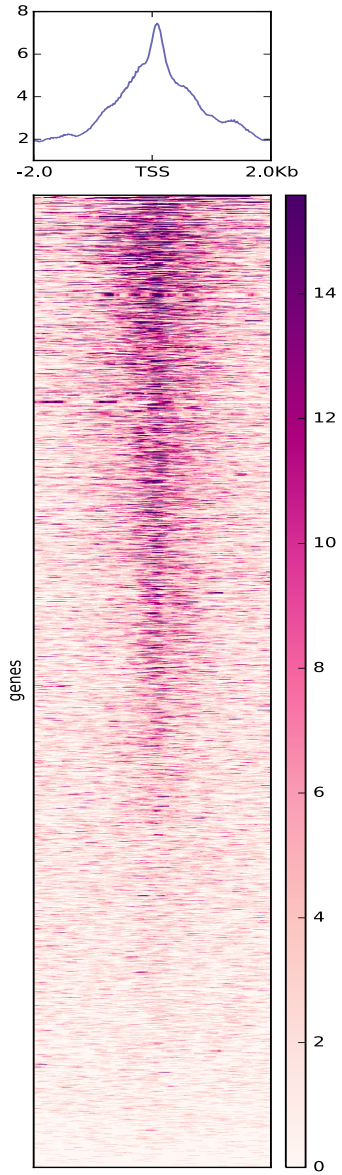
- Validation (wet lab)

- Downstream analysis
  - Motif discovery
  - Annotation
  - Integration of binding and expression data
  - Integration of various binding datasets
  - Differential binding





Signal visualisation and interpretation



deepTools
ngsplots
seqMiner

- Clustering
- Heatmaps
- Profiles
- Comparison of different datasets

Binding profile of a TF in relation to the transcription start site

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

Further reading

- Impact of artifact 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

Bioconductor ChIP-seq resources

- General purpose tools:
  - Rsubread (read mapping; not ideal for global alignment)
  - Rbowtie (global alignment)
  - GenomicRanges (tools for manipulating range data)
  - Rsamtools (SAM / BAM support)
  - htSeqTools (tools for NGS data; post-alignment QC)
  - chipseq (utilities for ChIP-seq analysis)
- Peak calling
  - SPP
  - BayesPeak (HMM and Bayesian statistics)
  - MOSAiCS (model-based one and two Sample Analysis and Inference for ChIP-Seq)
  - iSeq (Hidden Ising models)
  - ChIPseqR (developed to analyse nucleosome positioning data)
  - Csaw (a pipeline for ChIP-seq analysis, including statistical analysis of differential occupancy)
- Quality control
  - ChIPQC
- Differential occupancy
  - edgeR
  - DESeq2
  - DiffBind (compatible with objects used for ChIPQC, wrapper for DESeq and edgeR DE functions)
- Peak Annotation
  - ChIPpeakAnno (annotating peaks with genome context information)
  - ChIPSeeker (functional annotation of peaks)

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

Questions?

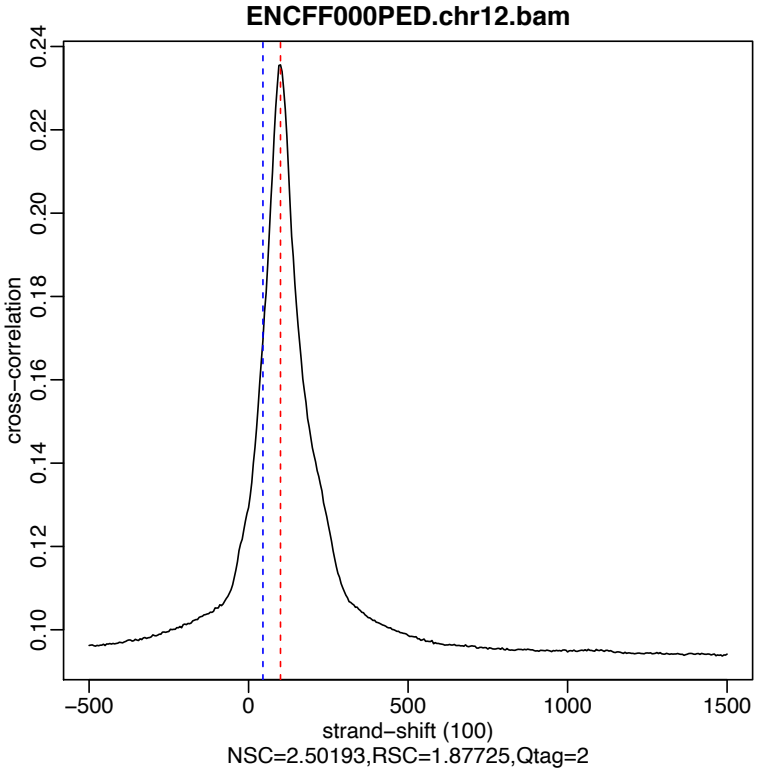
agata.smialowska@nbis.se

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

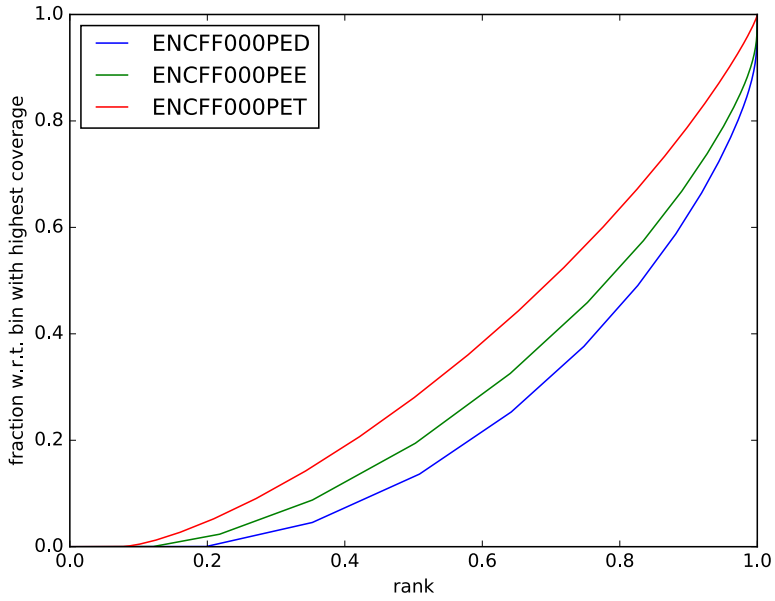
Exercise

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

Did my ChIP work?

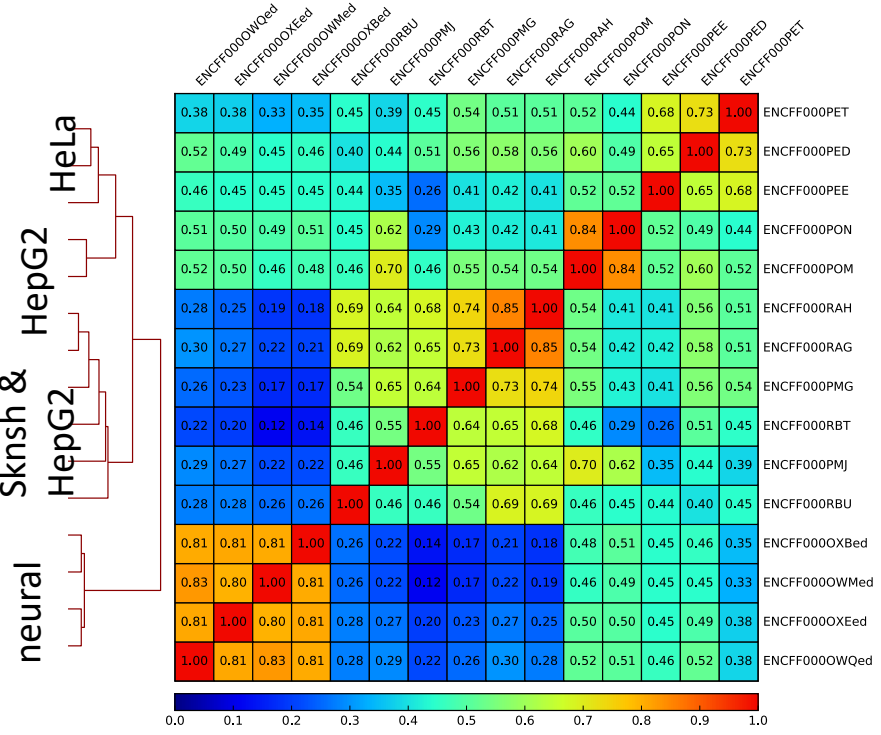


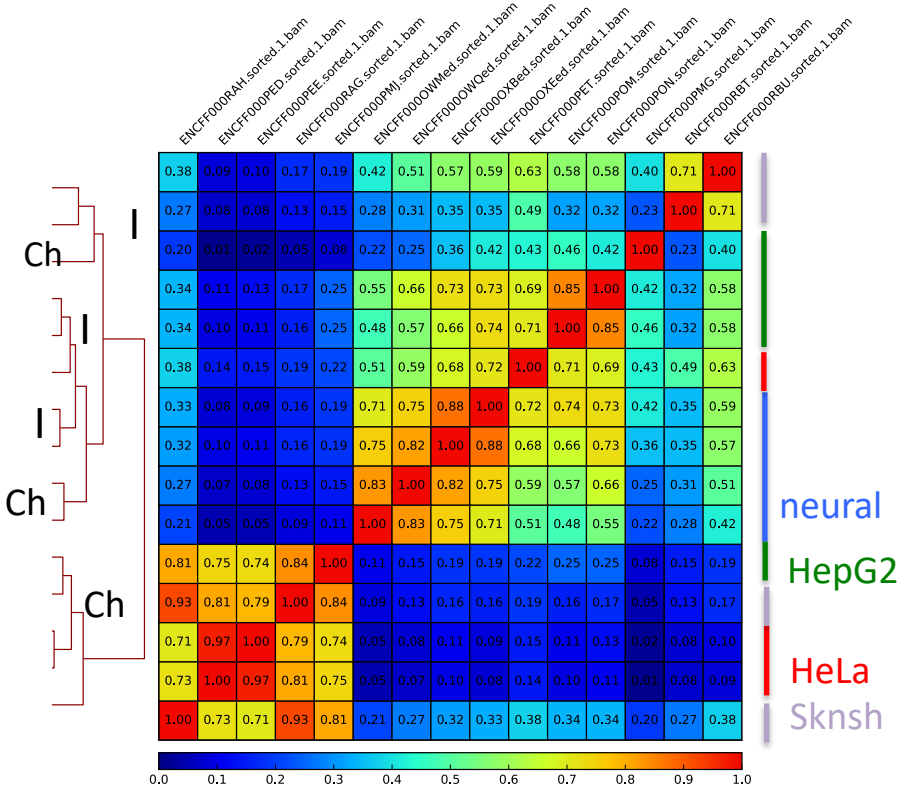
Cross-correlation



Cumulative enrichment

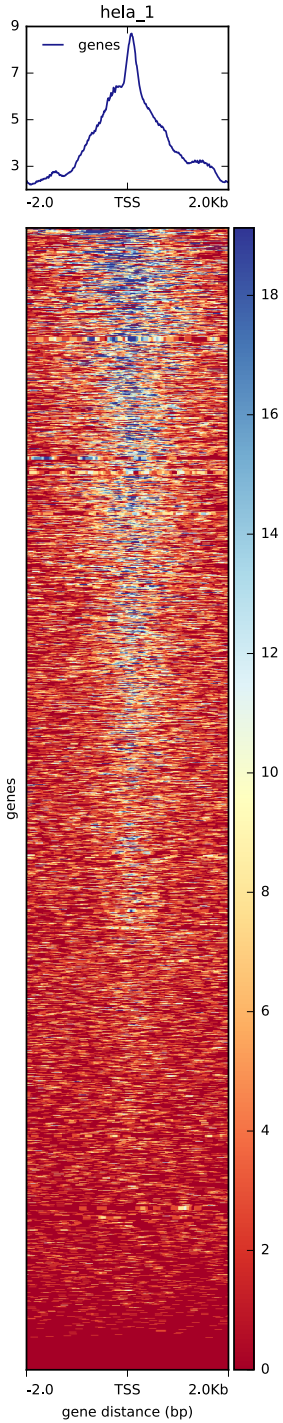
Exploratory analysis





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

Clustering of libraries
by reads mapped in peaks
(pearson)



Binding profile around TSS

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

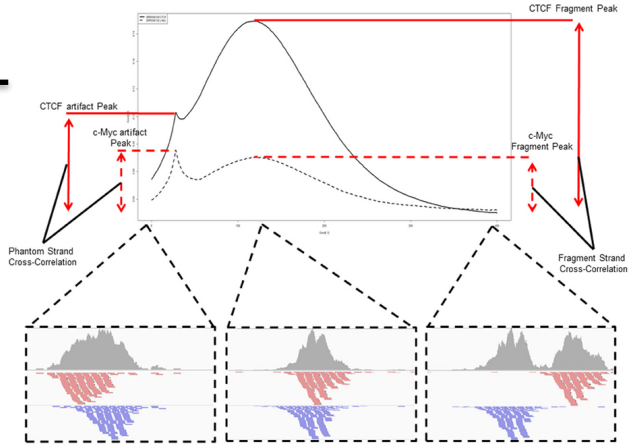
- Choose the right reference: assembly version (not always the newest is best) and type (primary assembly, or assembly from individual chromosome sequences + non-chromosomal contigs; not the top level assembly); choose the matching annotation file (GTF, GFF)
- Read mapping: **global alignment**
- Mappers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