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

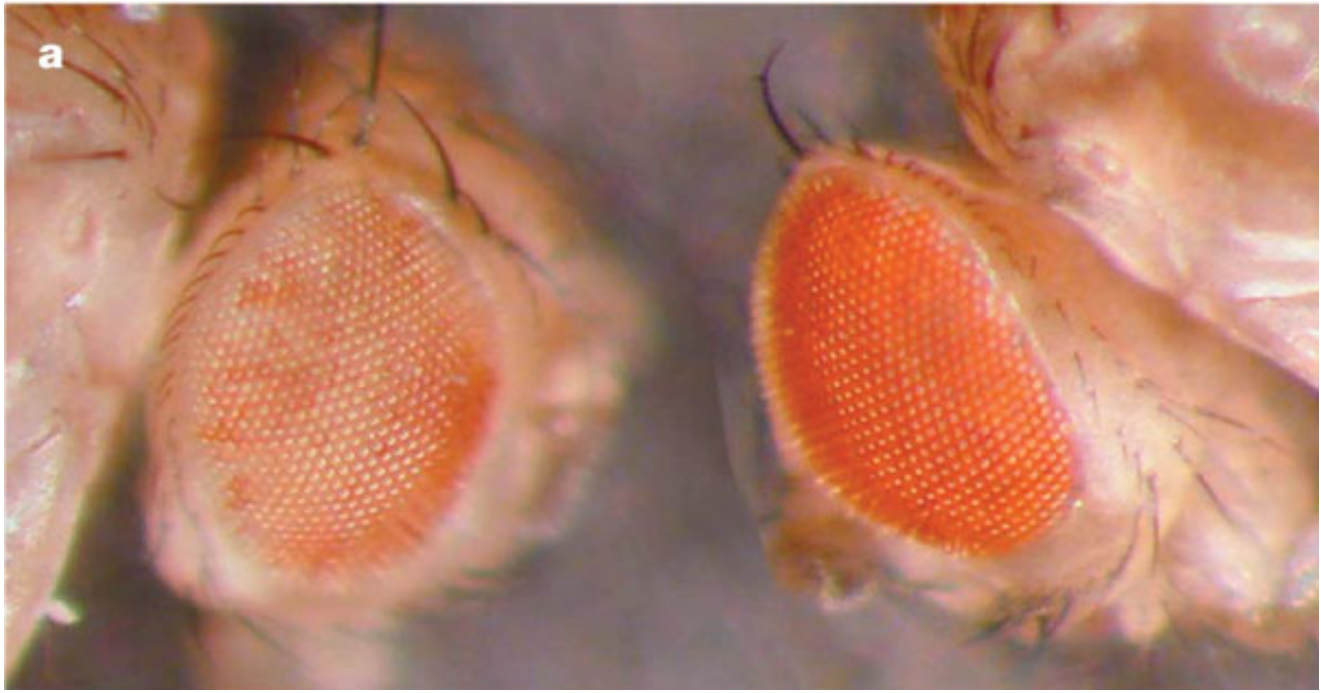
Introduction to Bioinformatics Using NGS Data

27 January 2016

Agata Smialowska

BILS, SciLife Lab, Stockholm University

Chromatin state and gene expression



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

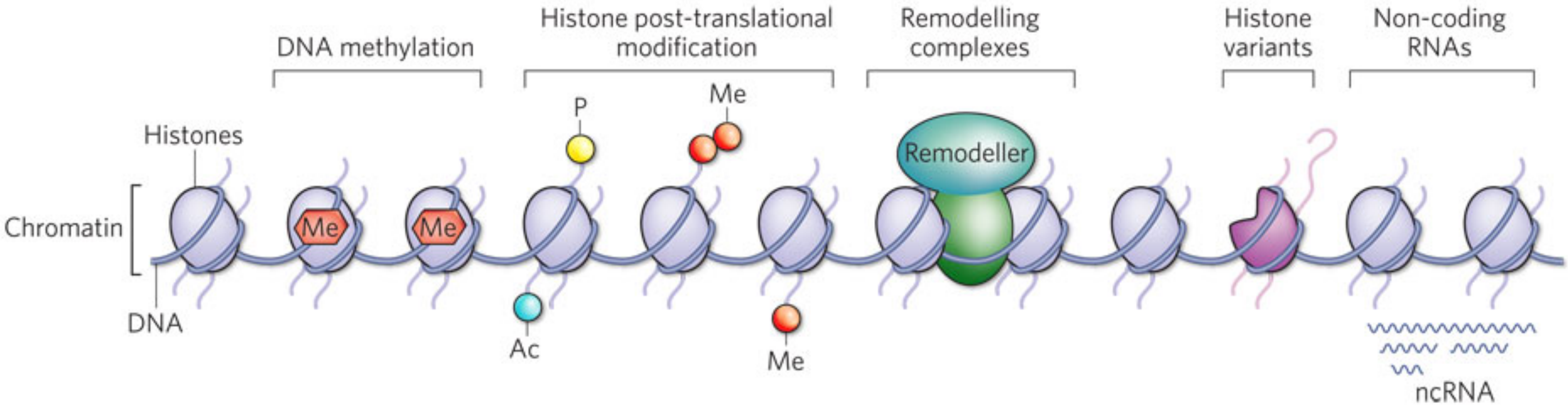
First observed by
H. Muller
1930

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

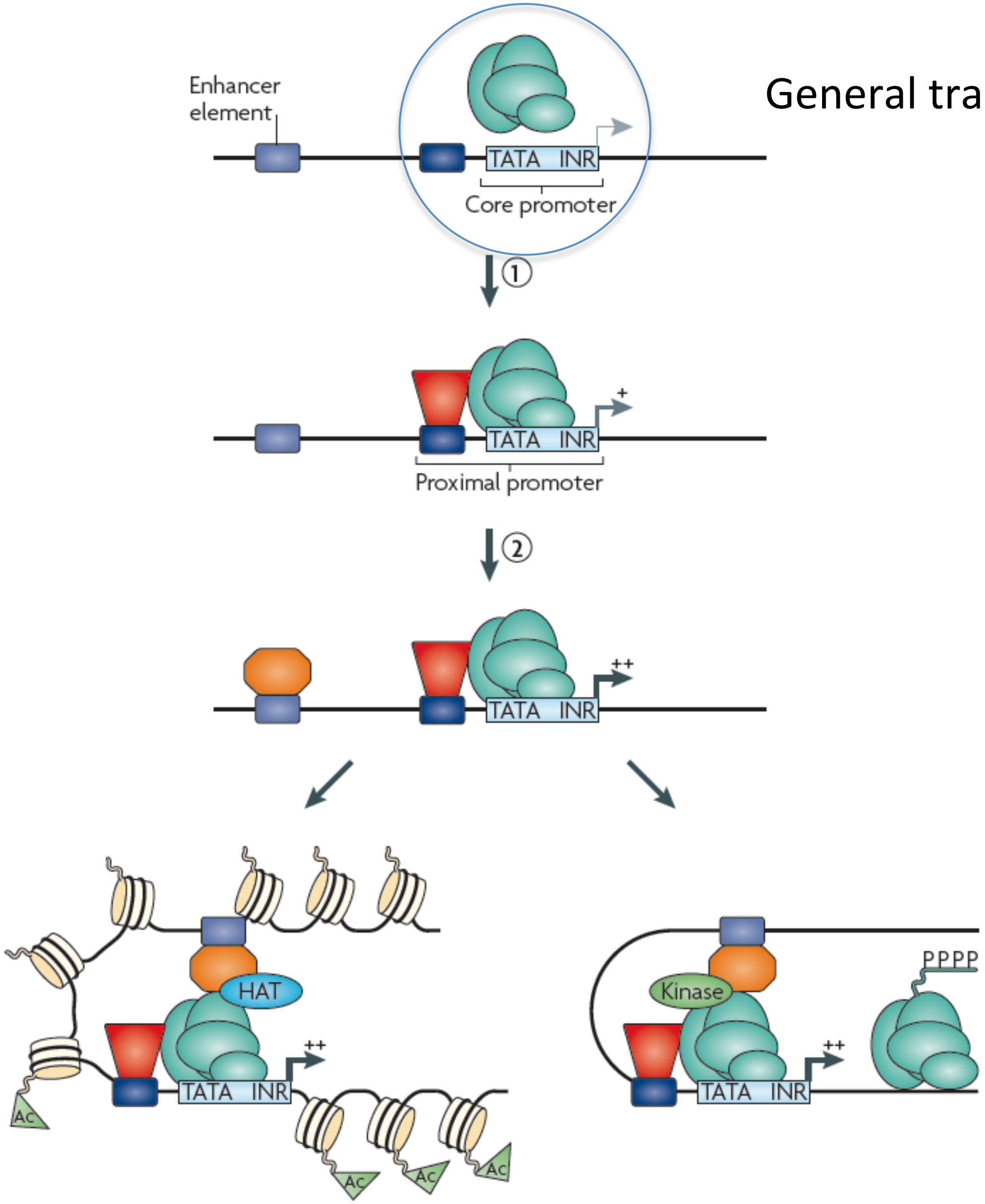
Chromatin / epigenetic signatures



Histone methylation
H3 K4 me3 – active gene promoters
H3 K36 me3 – bodies of active genes
H3 K27 me3 – facultatively repressed genes
H3 K9 me3 – silent chromatin (heterochromatin)

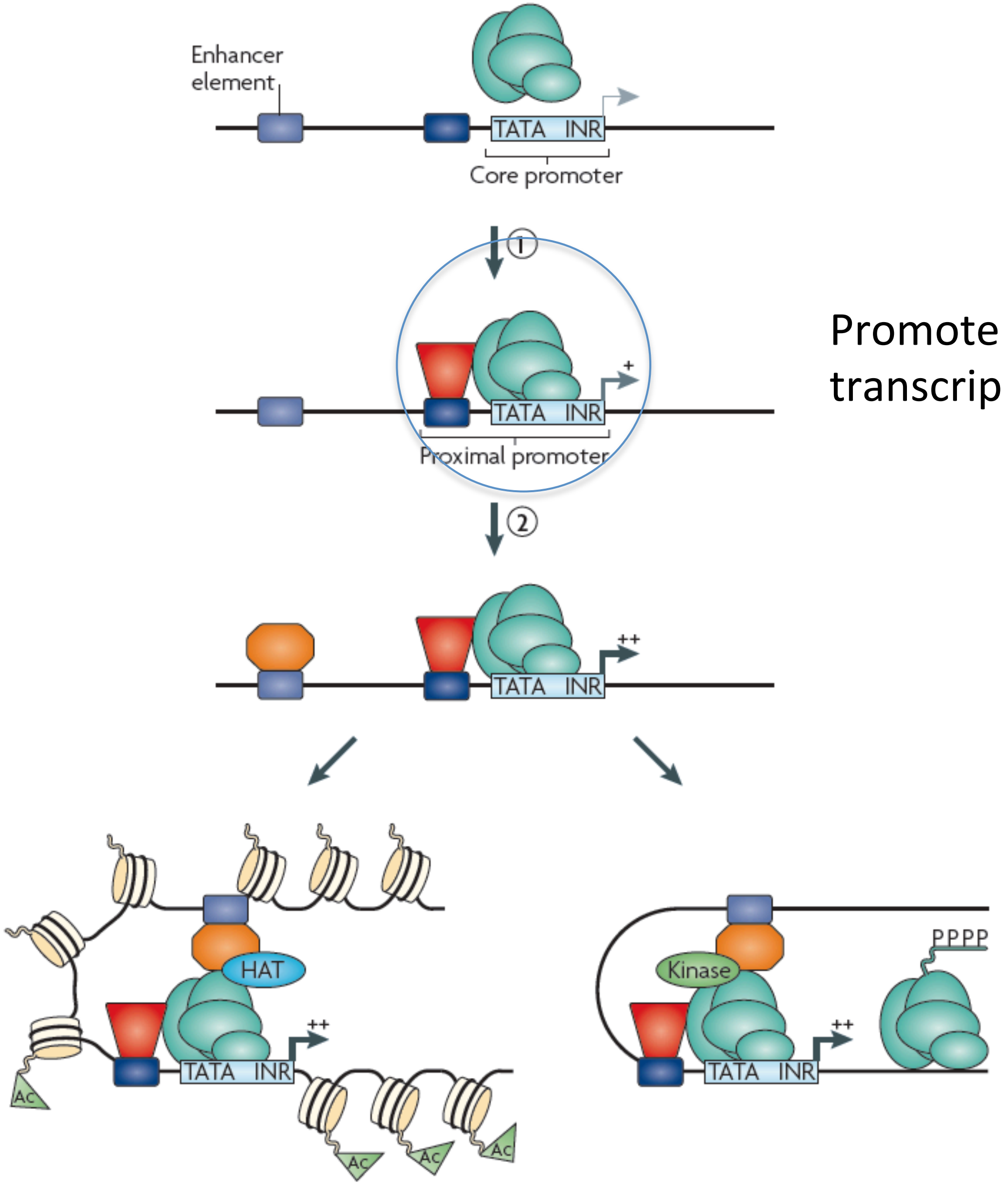
Dulac, Nature 2010

Applications



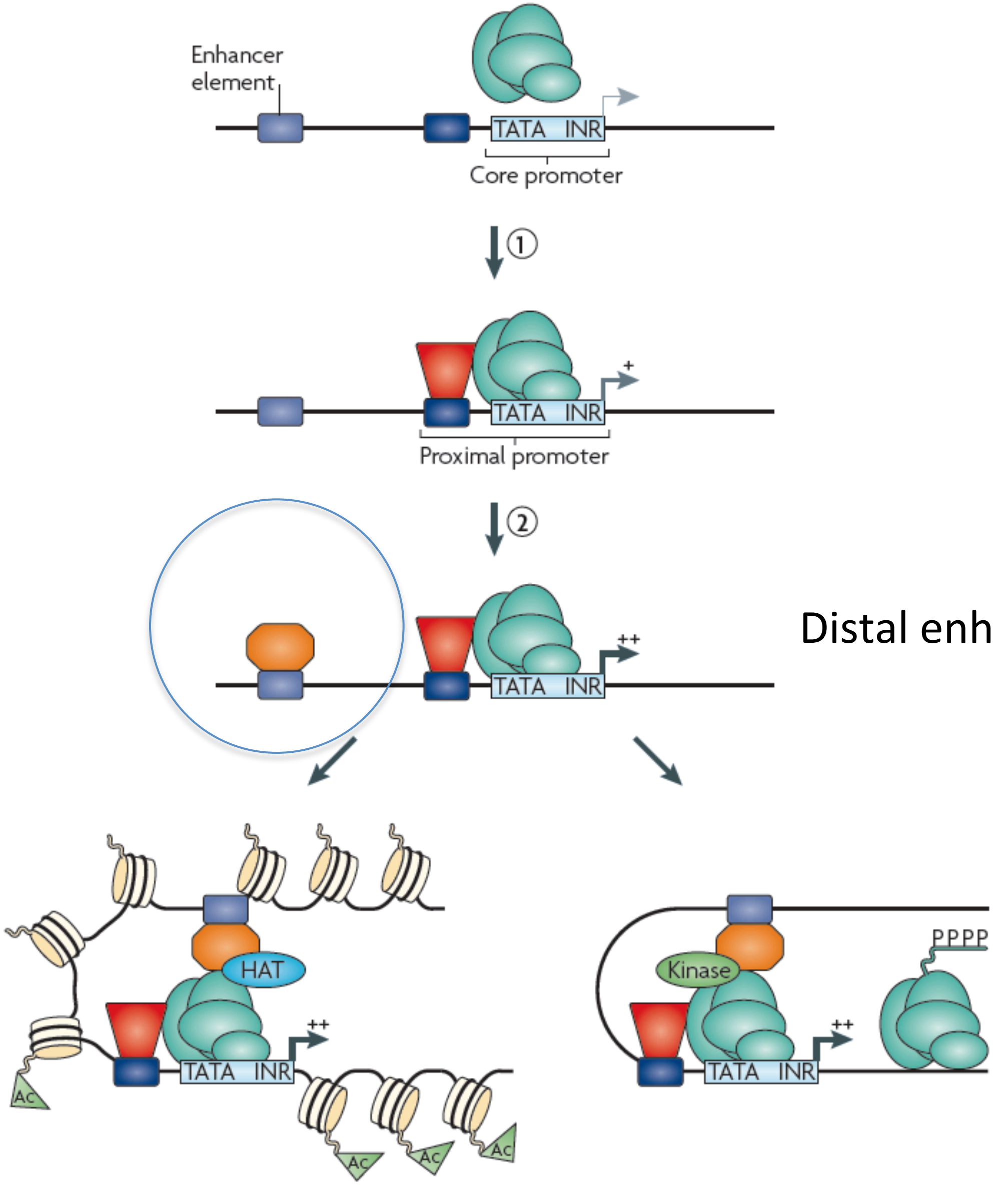
General transcription machinery

Applications



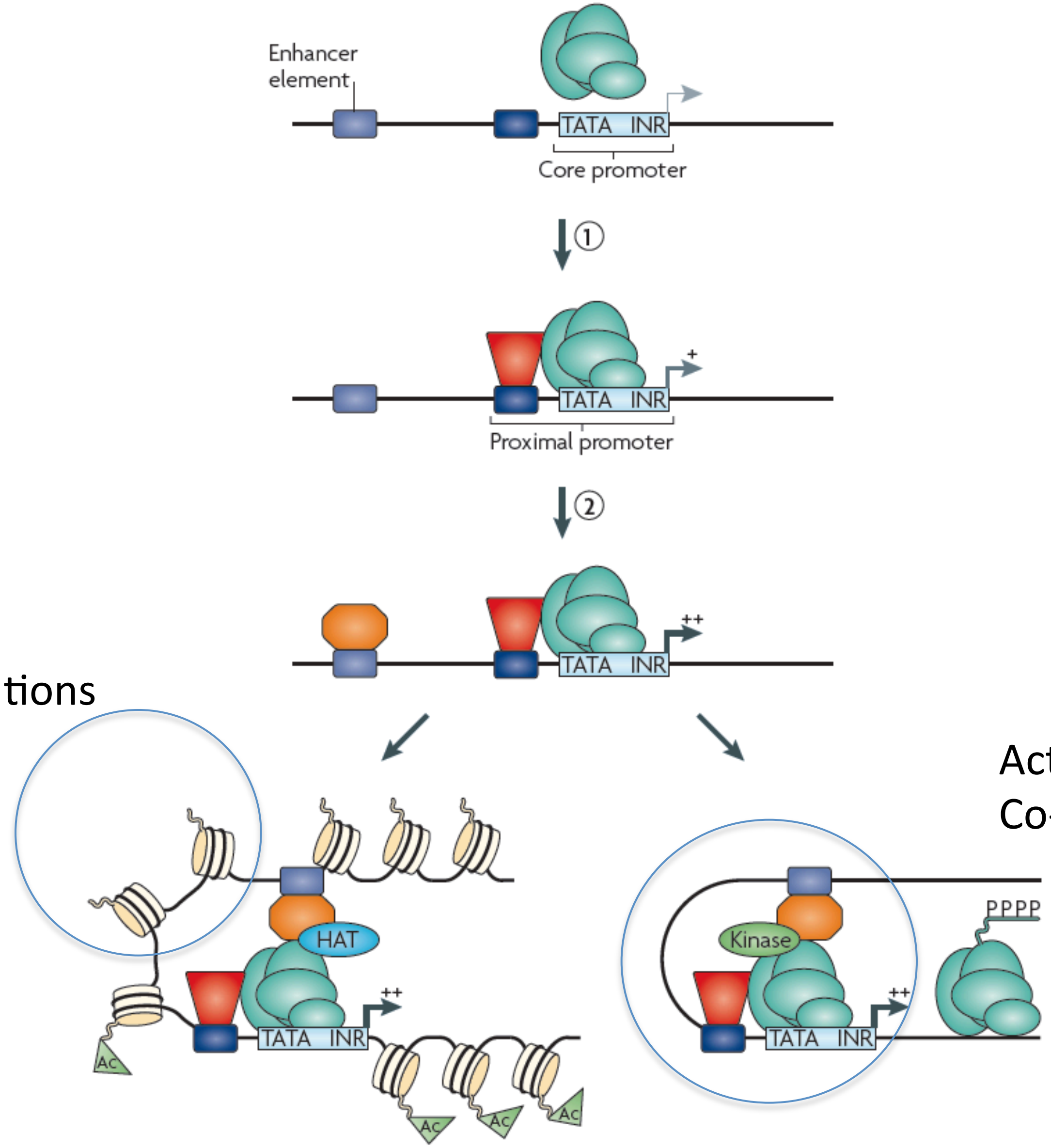
Promoter-associated transcription factors

Applications



Distal enhancers

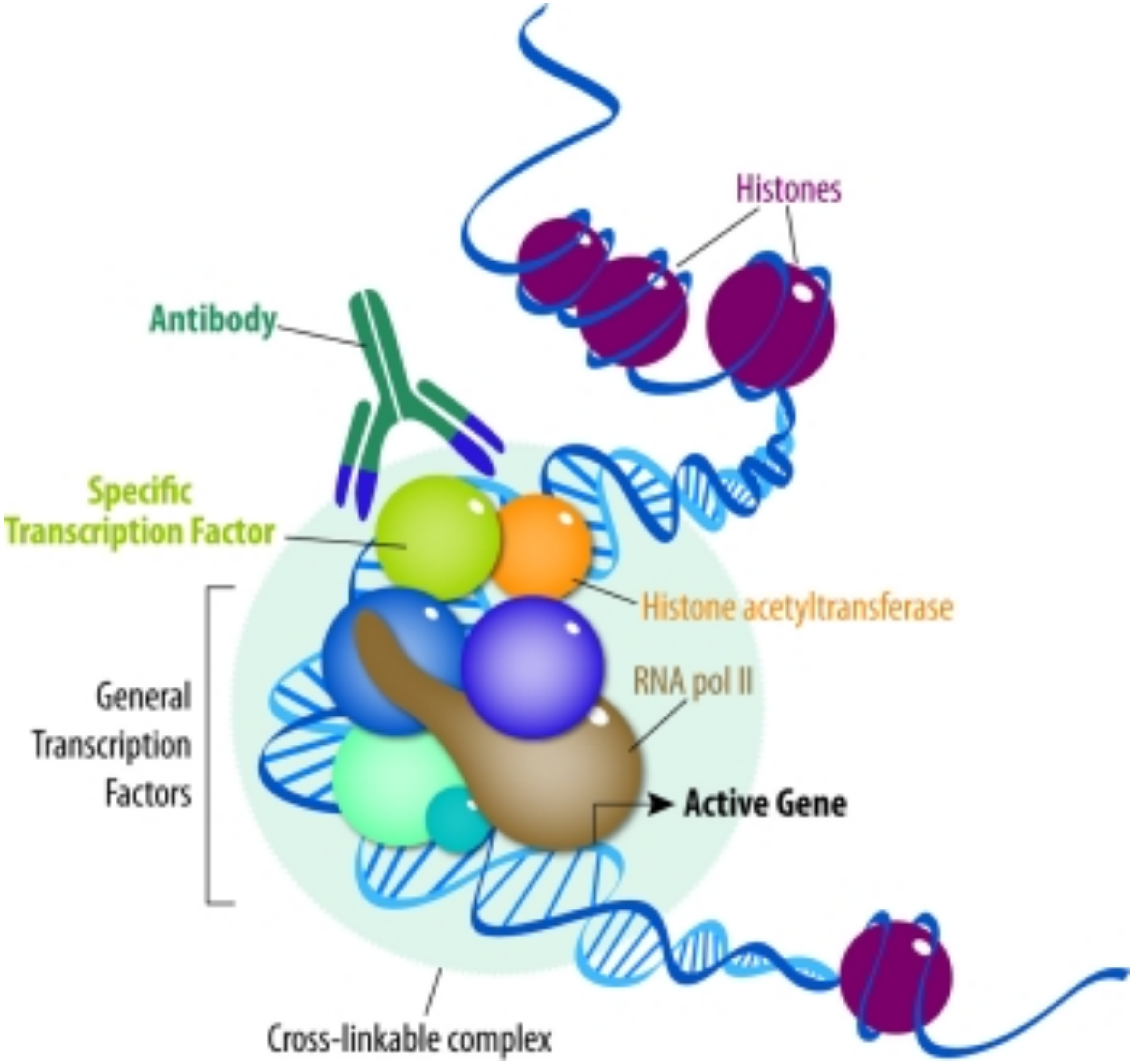
Applications



Histone modifications and variants

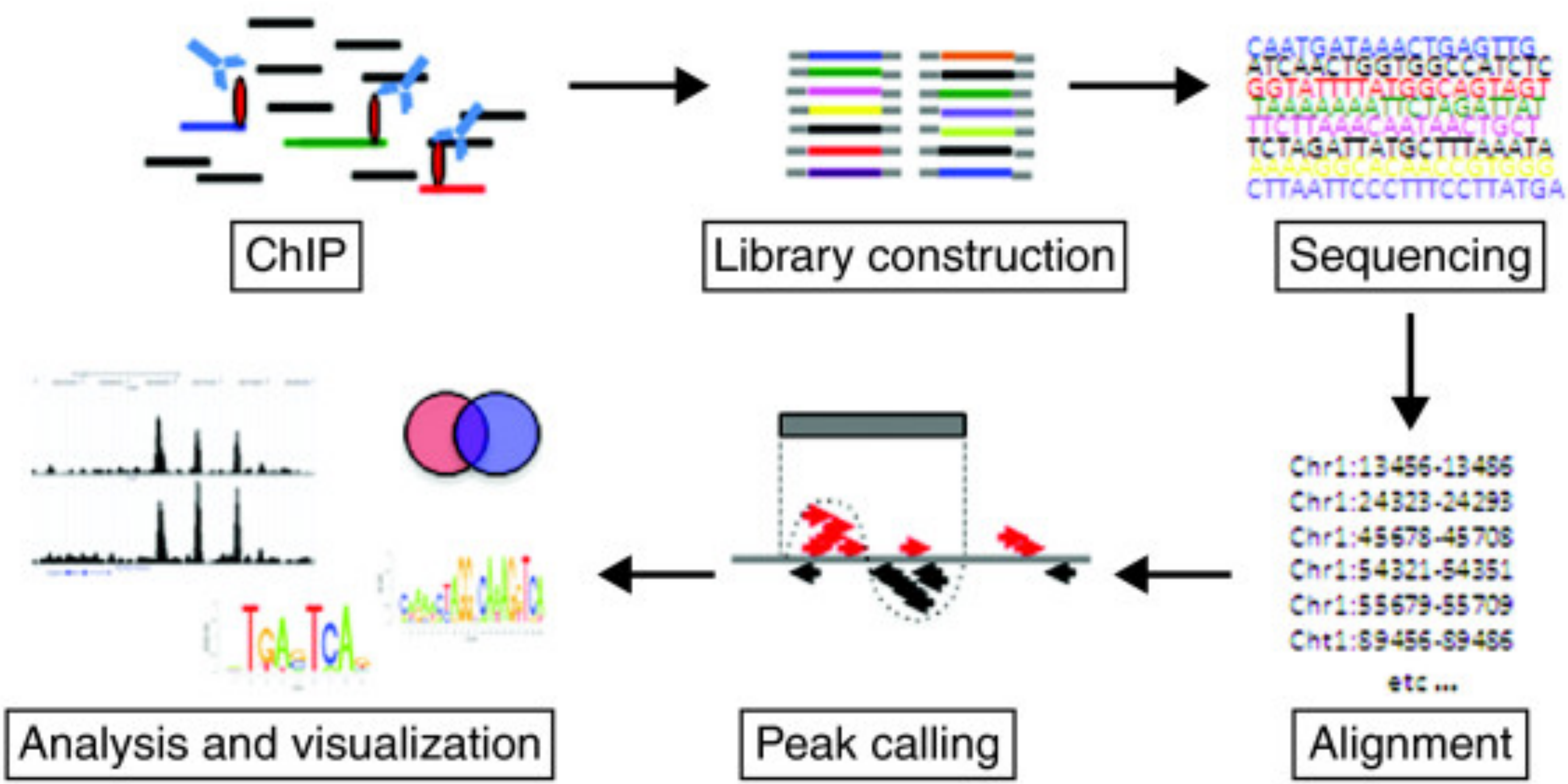
Activation states
Co-factors

Chromatin immunoprecipitation


Histones
Antibody
Specific Transcription Factor
Histone acetyltransferase
RNA pol II
General Transcription Factors
Active Gene
Cross-linkable complex

RnDsystems

ChIP-seq workflow



Liu, Pott and Huss, BMC Biology 2010

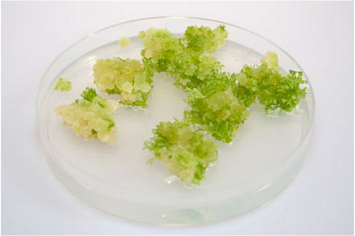
Critical factors

- Antibody selection
- Library cloning and sequencing
- Algorithm for peak detection
- Proper control sample (input chromatin or mock IP)

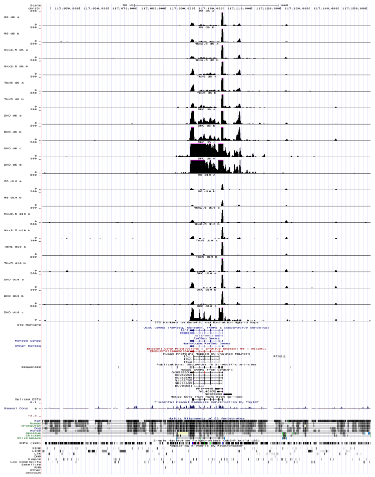
- Reproducibility in chromatin fragmentation
- Cross-linker choice
- Enough material and biological replicates

What you need

to get to the point of doing sequence tag alignments? (wet lab)



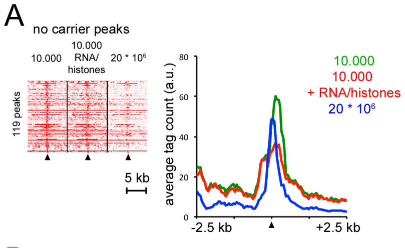
- reproducible experimental system
- molecular biology lab/reagents/expertise
- well conceived study design
- reliable library construction and sequencing lab/reagents/expertise
- modern computer running bowtie and fastqc



to build and view tracks in the genome browser,
call ChIP peaks, perform QC

- Linux / Mac OS machine / access to a server or an HTC cluster (SNIC / Uppmax)
- beginner bioinformatics expertise

to perform solid downstream analyses



- combination of advanced genomics, bioinformatics and biology experience (either one individual or a team working together).

Experiment design

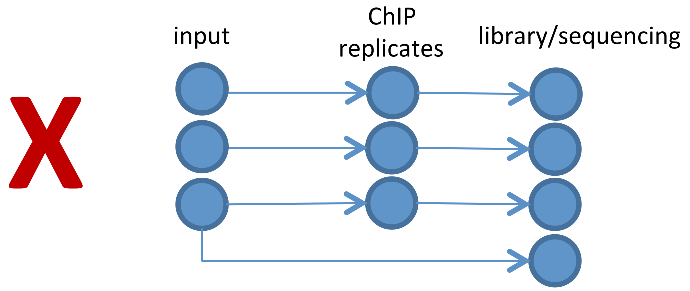
- Sound experimental design: replication, randomisation and blocking (R.A. Fisher, 1935)
- In the absence of a proper design, it is essentially impossible to partition biological variation from technical variation
- Sequencing depth: depends on the structure of the signal; cannot be easily scaled to 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

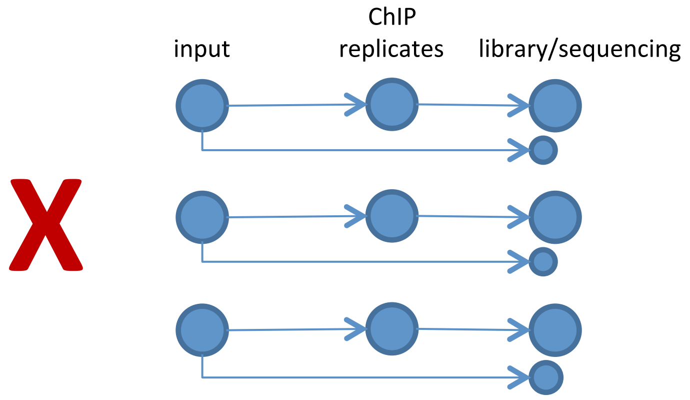
Experiment design

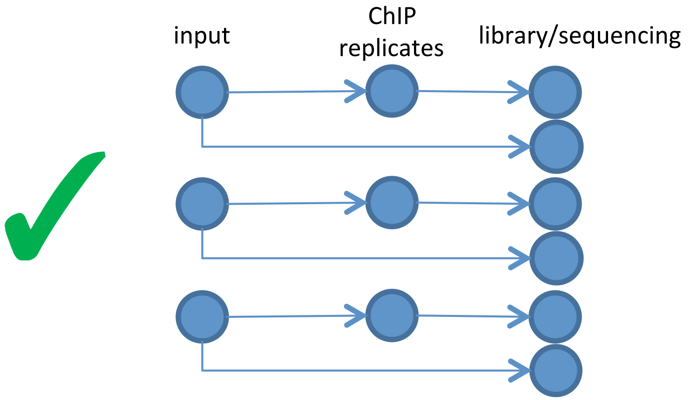
Ideal design:

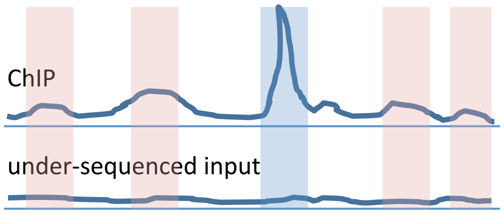
Each sample has a matched input
Input sequenced to a comparable depth as IP sample

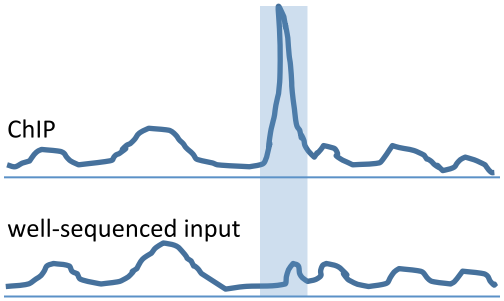
≥2 biological replicates for site identification
≥3 biological replicates for differential binding





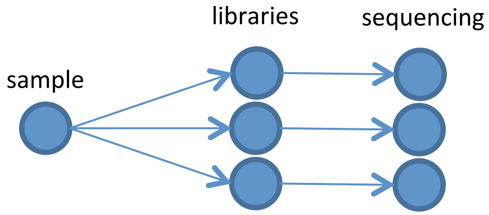






Importance of biological replicates

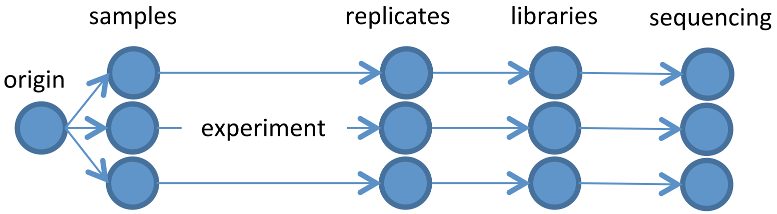
✗



sample
libraries
sequencing

technical replicates are generally a waste of time and money

✗

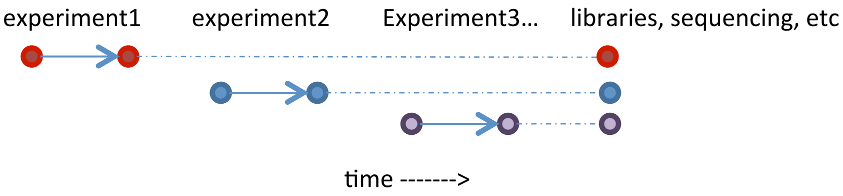


origin
samples
experiment
replicates
libraries
sequencing

many studies do not account for batch effects

i. time
ii. origin

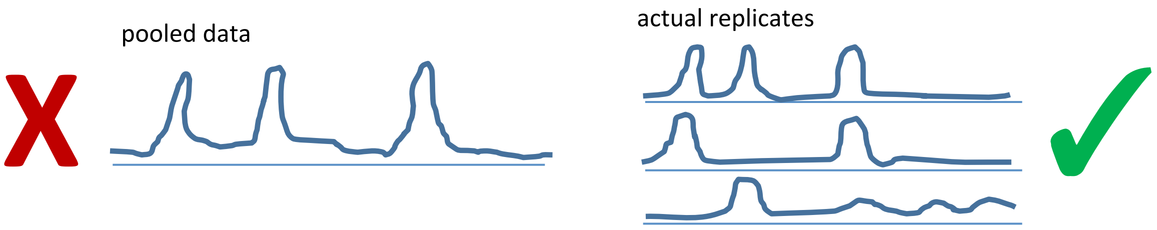
✓



experiment1
experiment2
Experiment3…
libraries, sequencing, etc
time ------->

so if you care about reproducibility

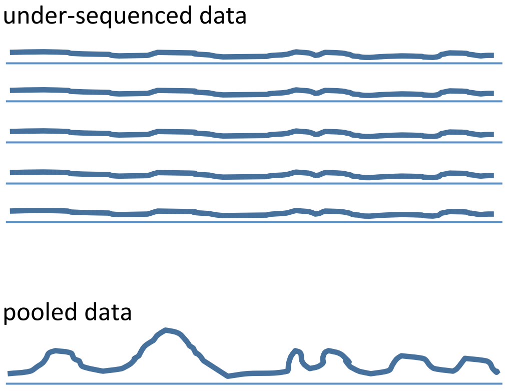
Importance of sequencing depth



pooled data

actual replicates

if you need to pool your data, then it is under-sequenced

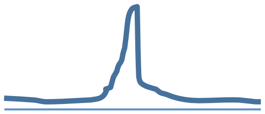


under-sequenced data

pooled data

Sequencing depth depends on data type

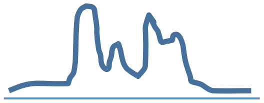
Transcription Factors



point-source

Chromatin Remodellers

Histone marks

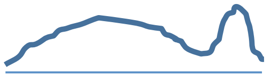


mixed signal

Chromatin Remodellers

Histone marks

RNA polymerase II



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

- Exercise overview

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

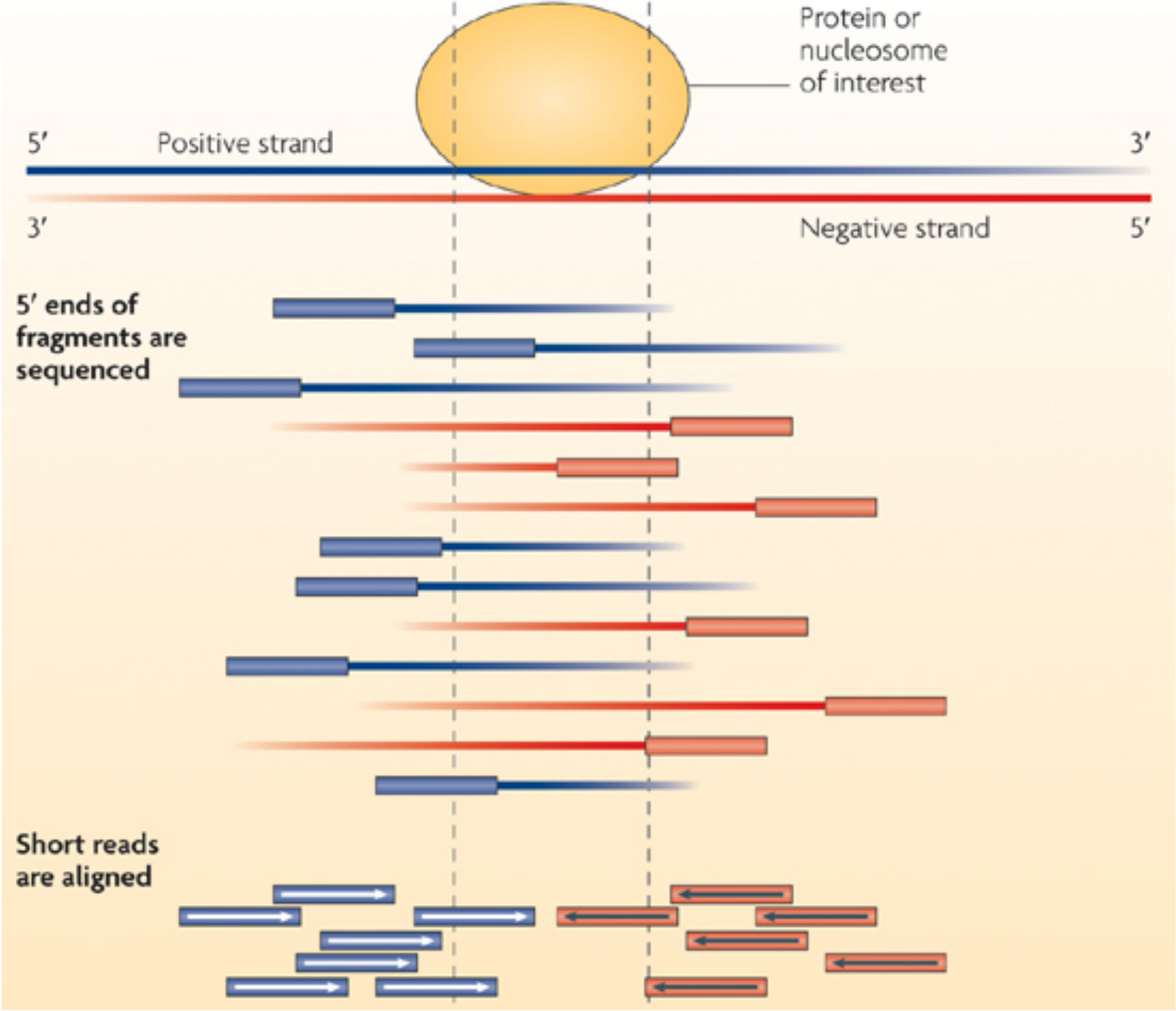
- Principles of analysis of ChIP-seq data

- ChIP-seq: downstream analyses

- Resources

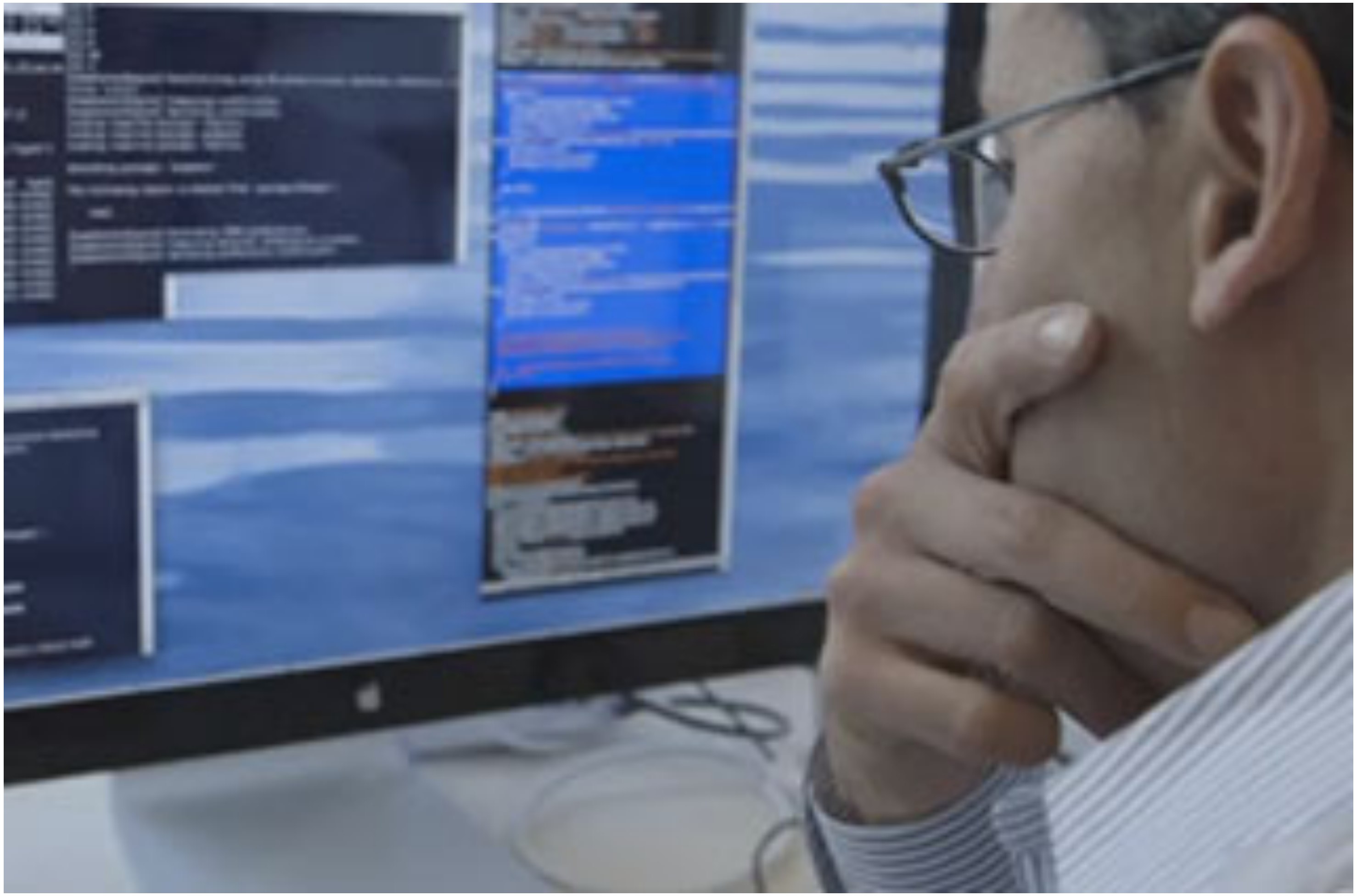
- Exercise overview

Chromatin = DNA + proteins

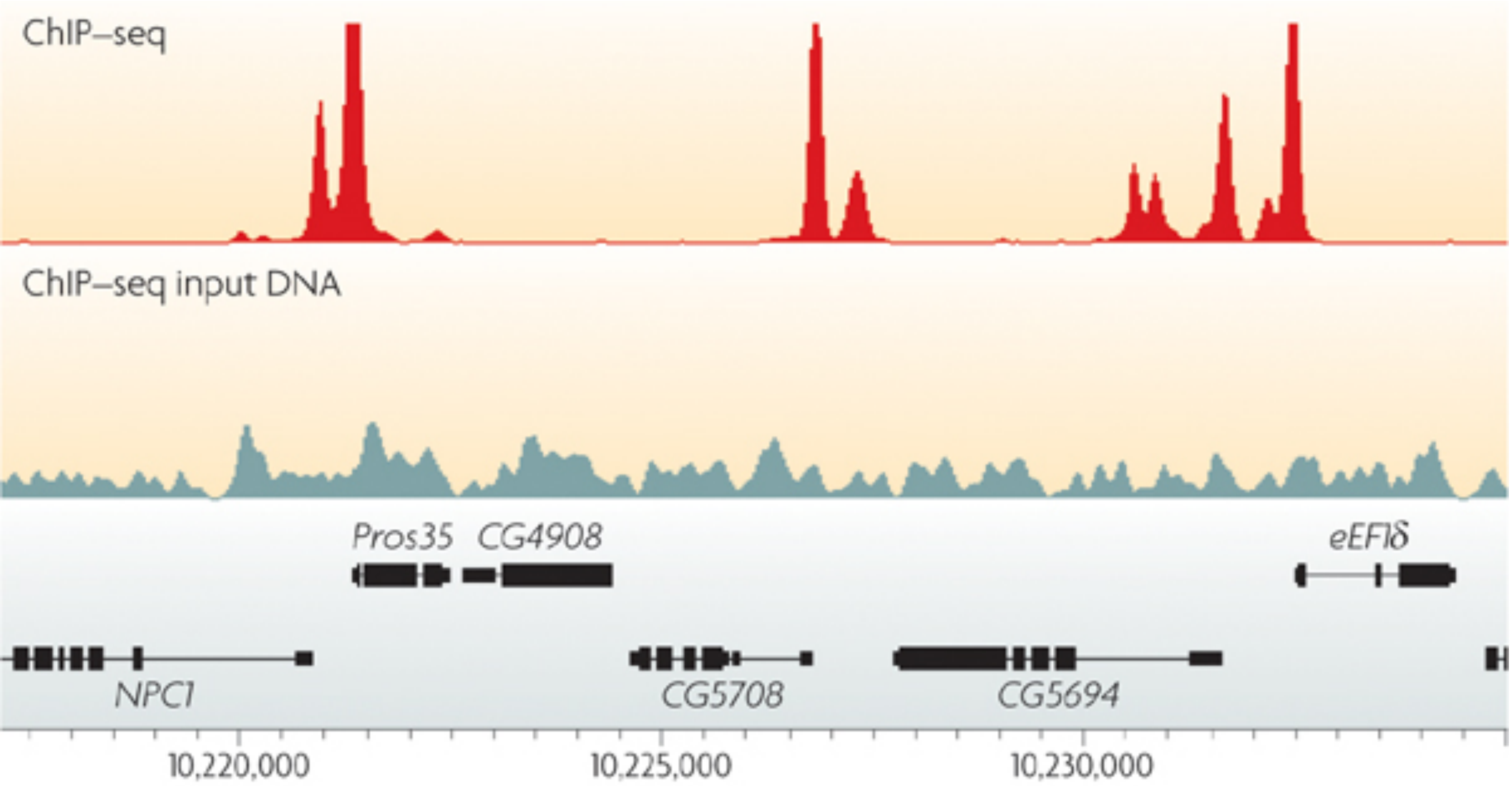

Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned

Park, Nature Rev Genetics, 2009

Data analysis



Profile of protein binding sites vs input



Chromator (*Drosophila*) – protein binding methylated histones

Park, Nature Rev Genetics, 2009

Workflow of a ChIPseq study

design study

obtain input chromatin

perform precipitation

construct library

sequence library

filter sequences

align sequences

identify peaks / regions of enrichment

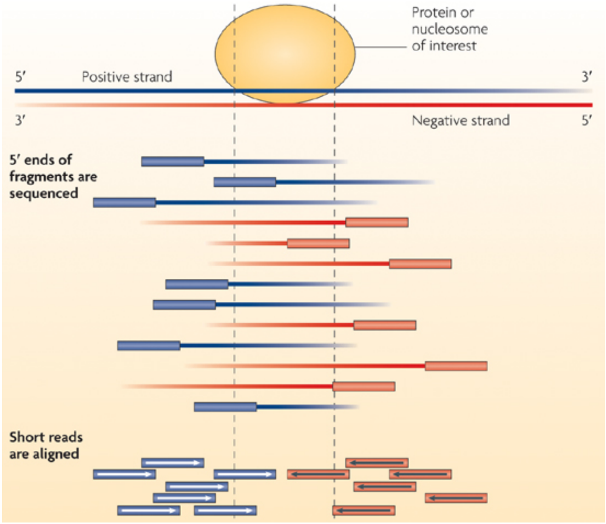
assess data quality

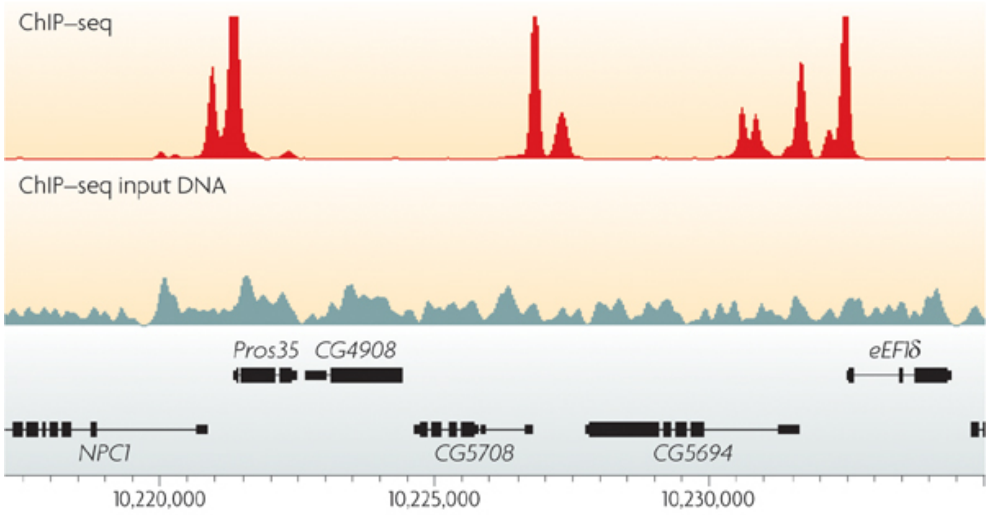
understand the data

downstream analyses

Iterative process

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


Protein or nucleosome of interest
5′
Positive strand
3′
3′
Negative strand
5′
5′ ends of fragments are sequenced
Short reads are aligned



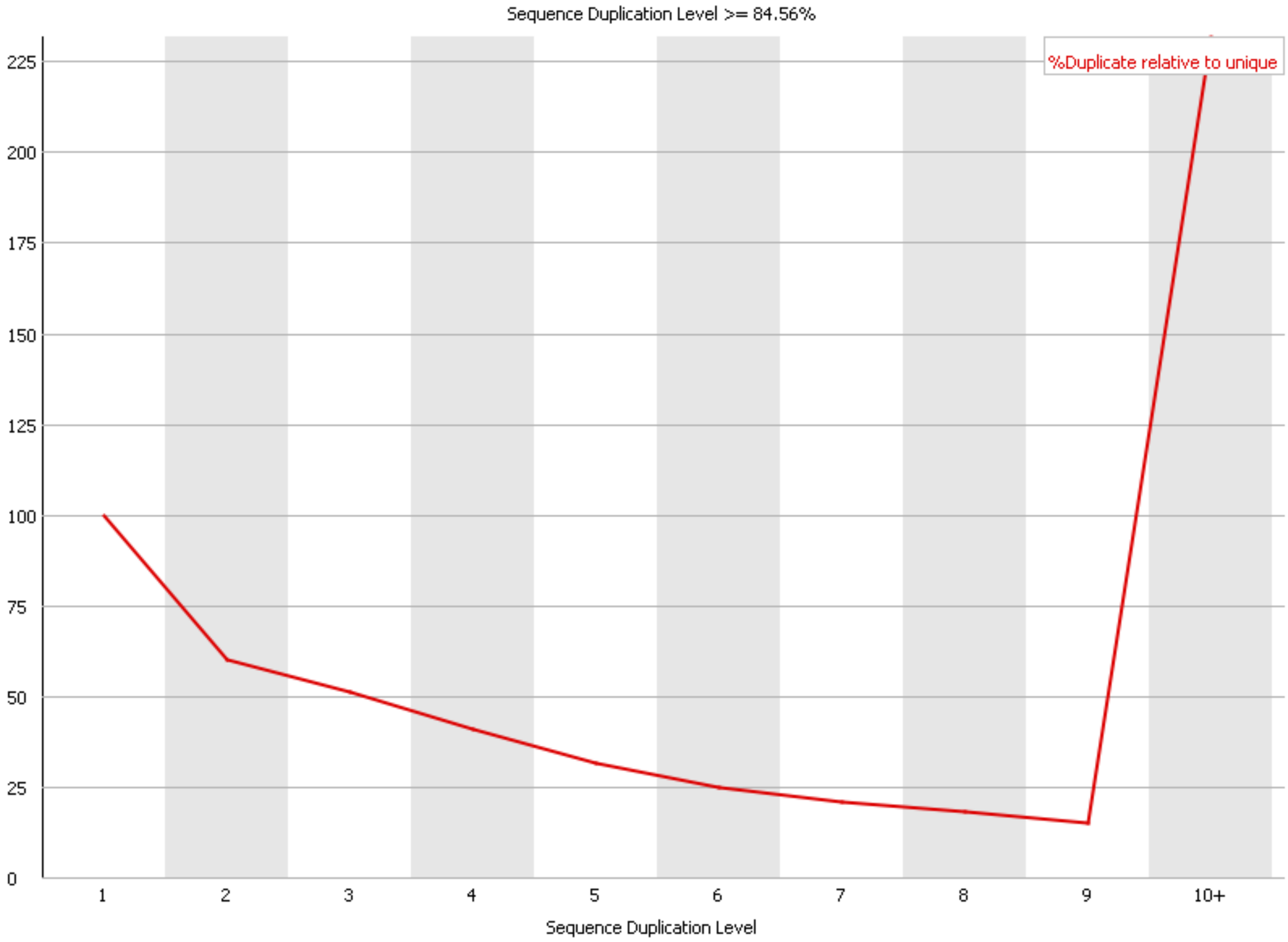
Park, Nature Rev Genetics, 2009

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 succesfull
- Filter out redundant (duplicated) reads; some peak callers (MACS) do that automatically

Quality control: tag uniqueness – library complexity metric

Sequence duplication level > 70% (low complexity library)



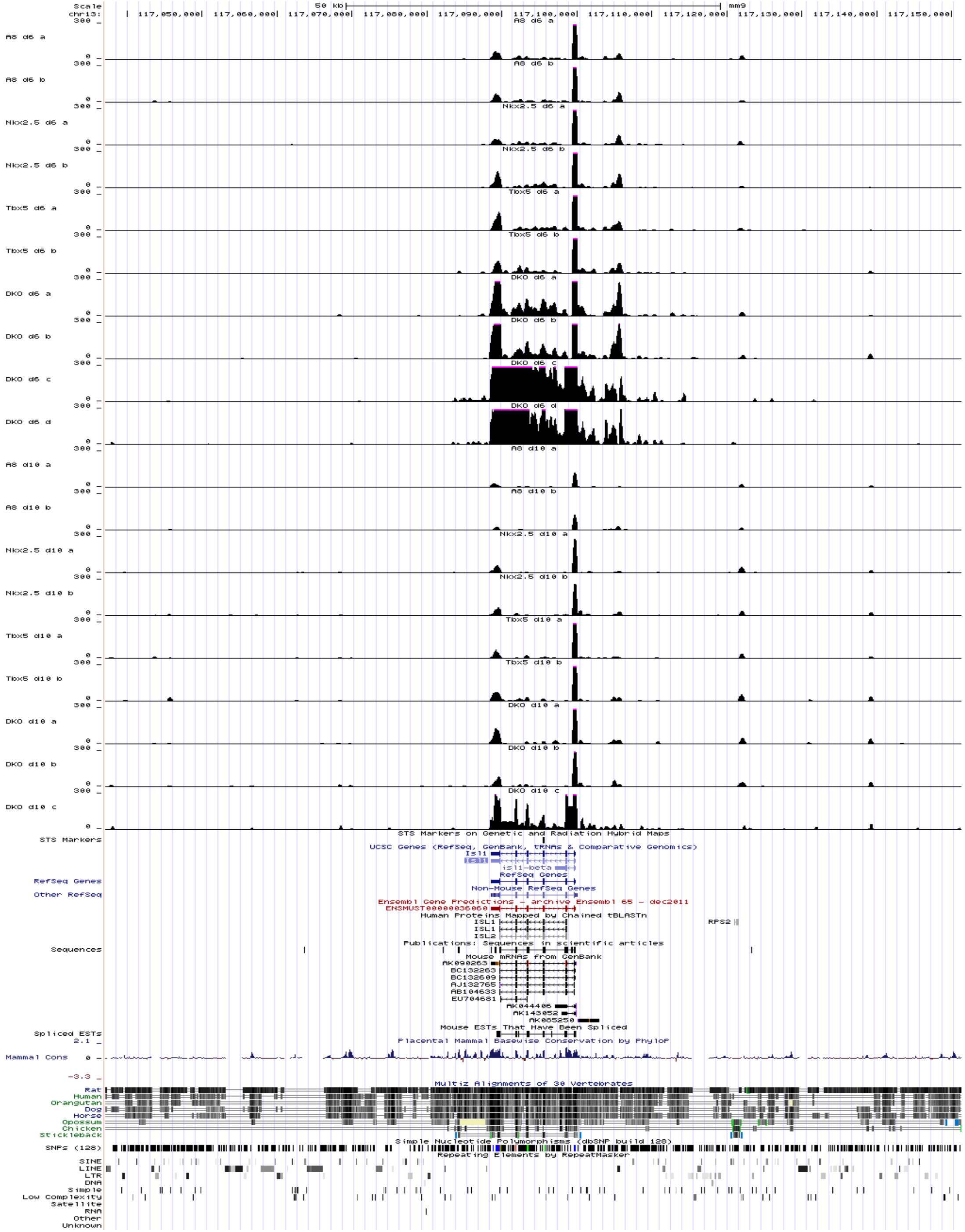
NRF = Non-redundant fraction (of reads)

- the proportion of duplicates within a data set compared to the total number of sequence reads has been used as a measure of ChIP-quality
- recently formalized by the ENCODE consortium as the Non-Redundant Fraction (NRF)
- guidelines for NRF suggest that less than 20% of reads should be duplicates for 10 million reads sequenced

Mapping reads to the reference genome

- Choose the right reference: assembly version (not always the newest is best) and type (primary assembly, or assemble 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



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

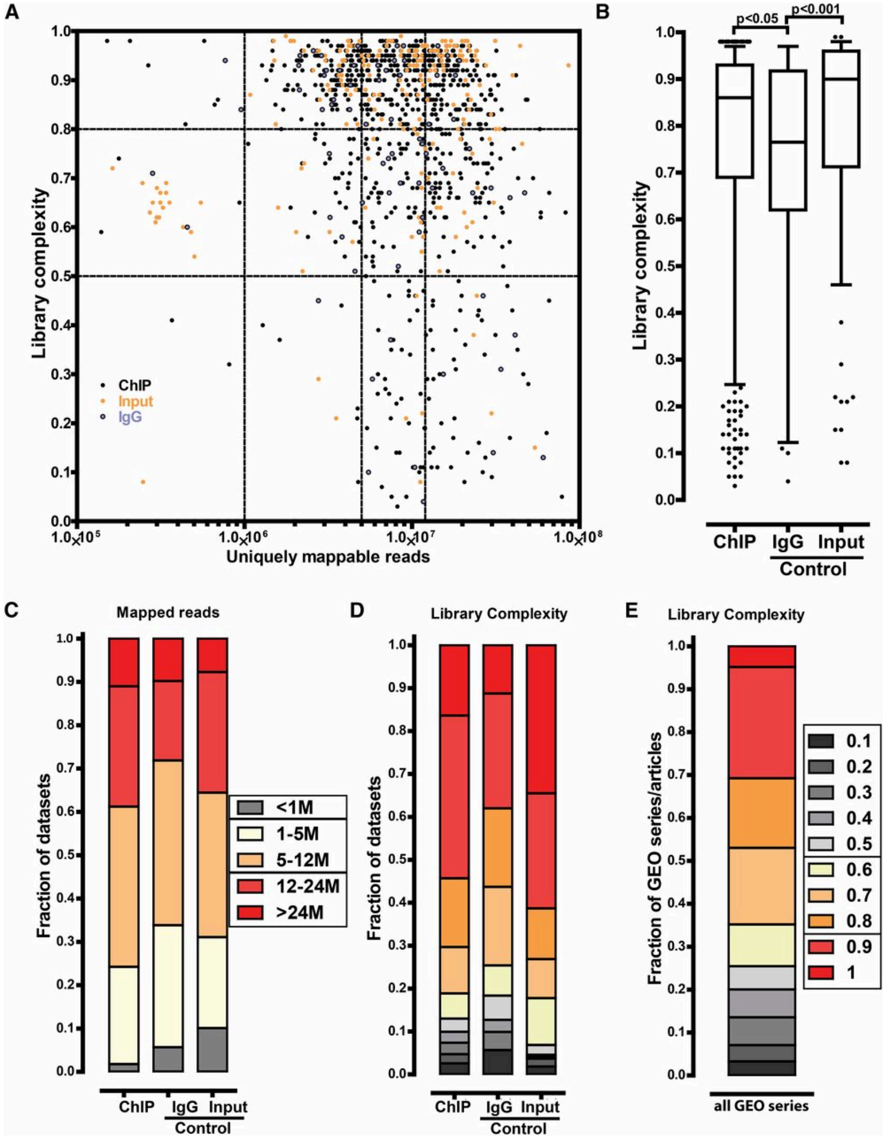
Word of caution!

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

Error sources:

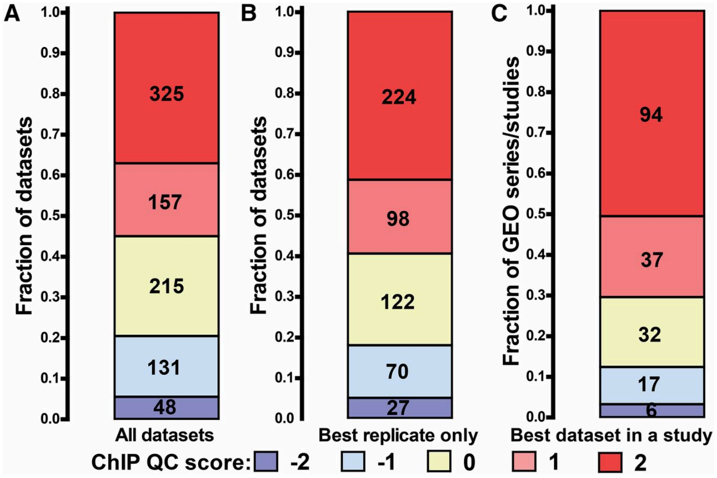
- chromatin structure
- PCR overamplification
- non-specific antibody
- other things?

How do I know my data is of good quality?



Objective metrics to quantify enrichment in ChIP-seq; for TF in mammalian systems:
NSC, RSC

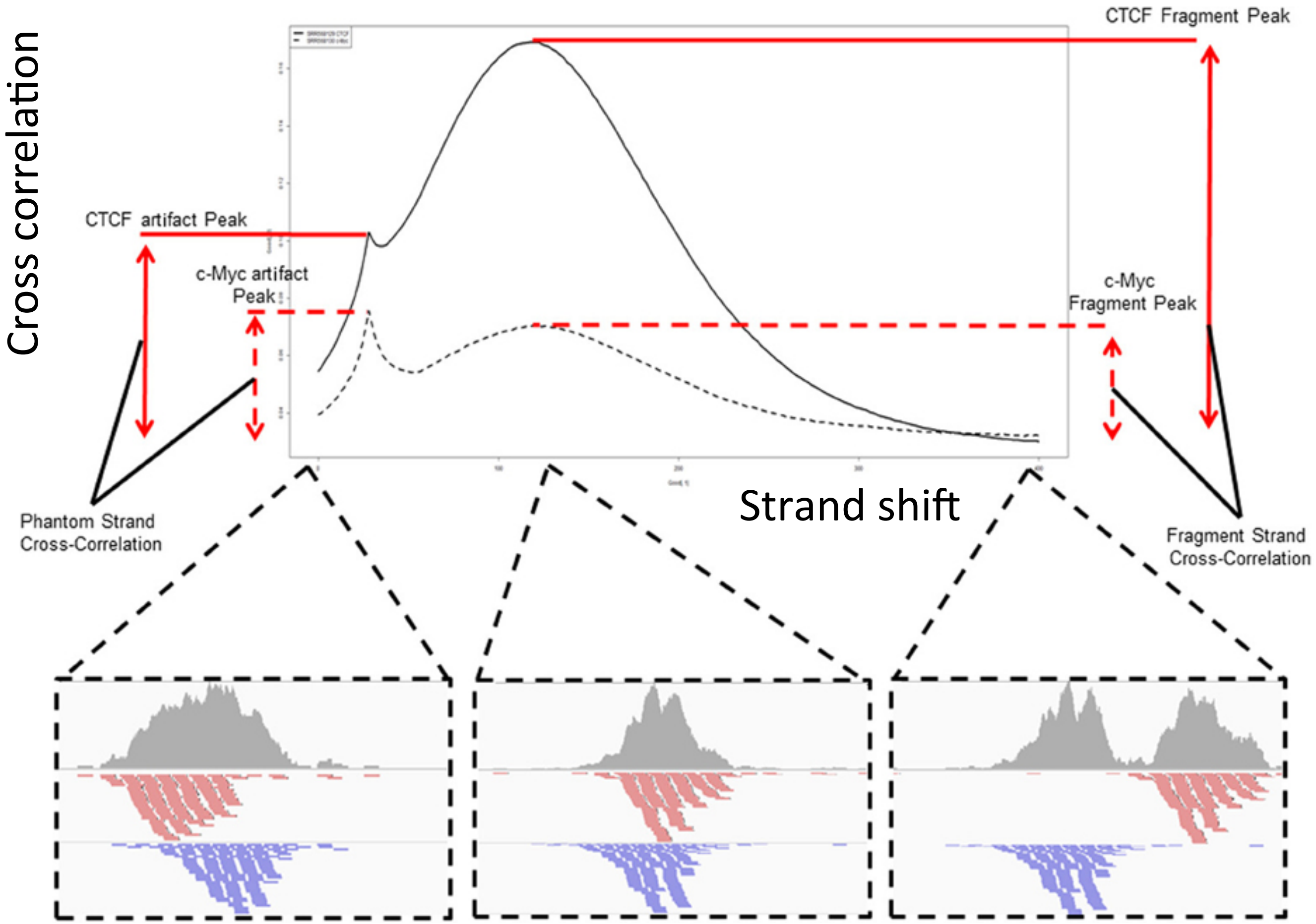
Large-scale quality analysis of published ChIP-seq data sets:
20% low quality
25% intermediate quality
30% inputs have metrics similar to IPs



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.

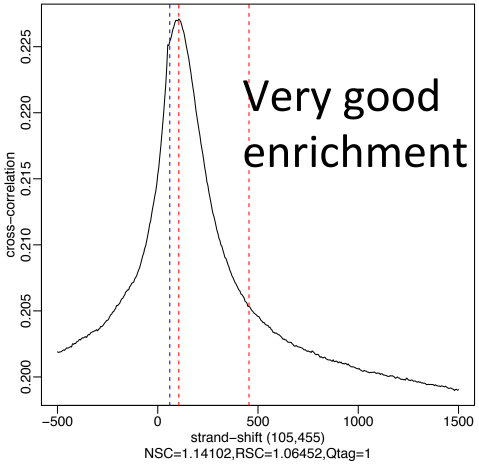


Carroll et al, Front Genet 2014

Cross correlation plots

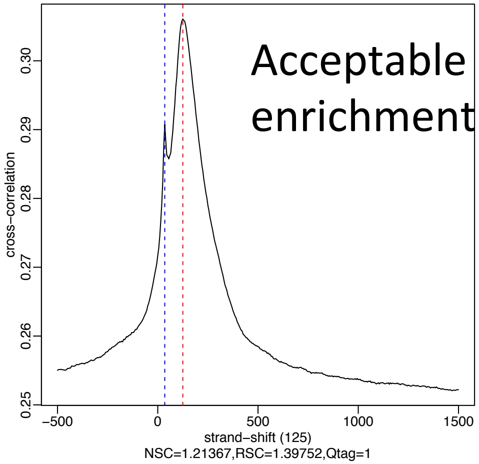
ChIP

ENCFF000OWMed.sorted.1.bam.picard.bam



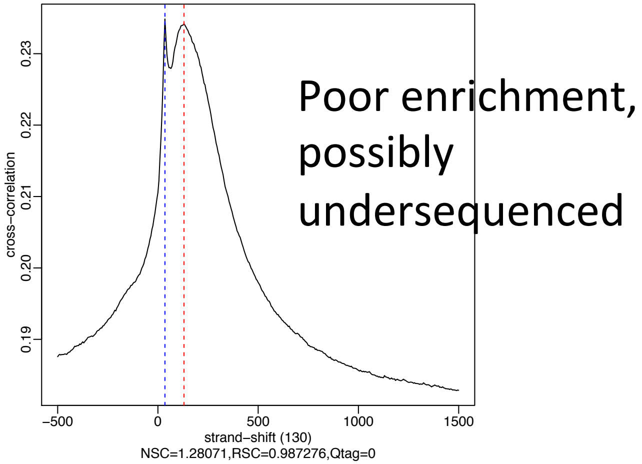
Very good enrichment

ENCFF000PMJ.sorted.1.bam



Acceptable enrichment

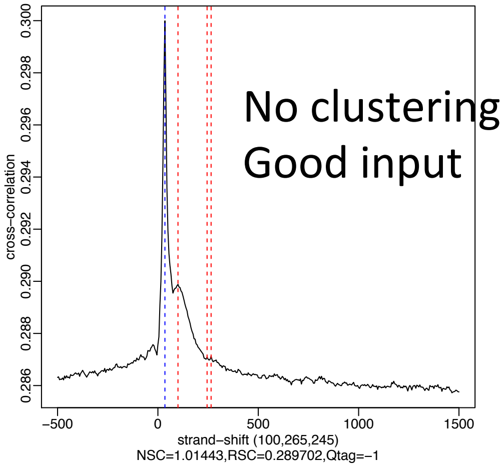
ENCFF000PMG.sorted.1.bam



Poor enrichment, possibly undersequenced

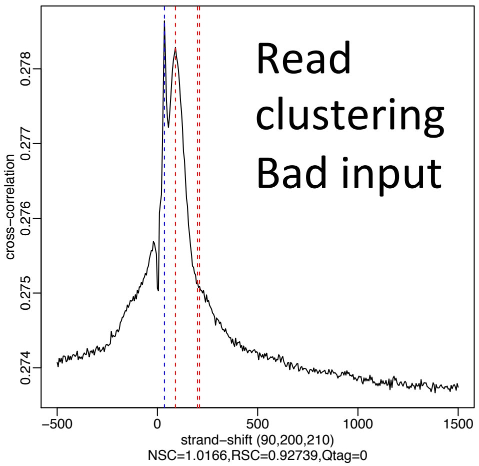
Input

ENCFF000PET.sorted.1.bam.picard.bam



No clustering
Good input

ENCFF000PON.sorted.1.bam.picard.bam



Read clustering
Bad input

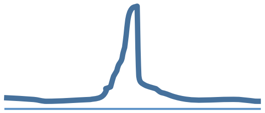
Quality considerations

- 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)

Peak calling

appropriate methodologies depend on data type

Transcription
Factors

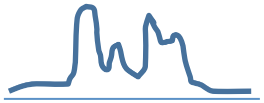


punctate

SPP
MACS

Chromatin
Remodellers

Histone marks



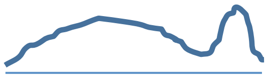
mixed signal

-

Chromatin
Remodellers

Histone marks

RNA polymerase II

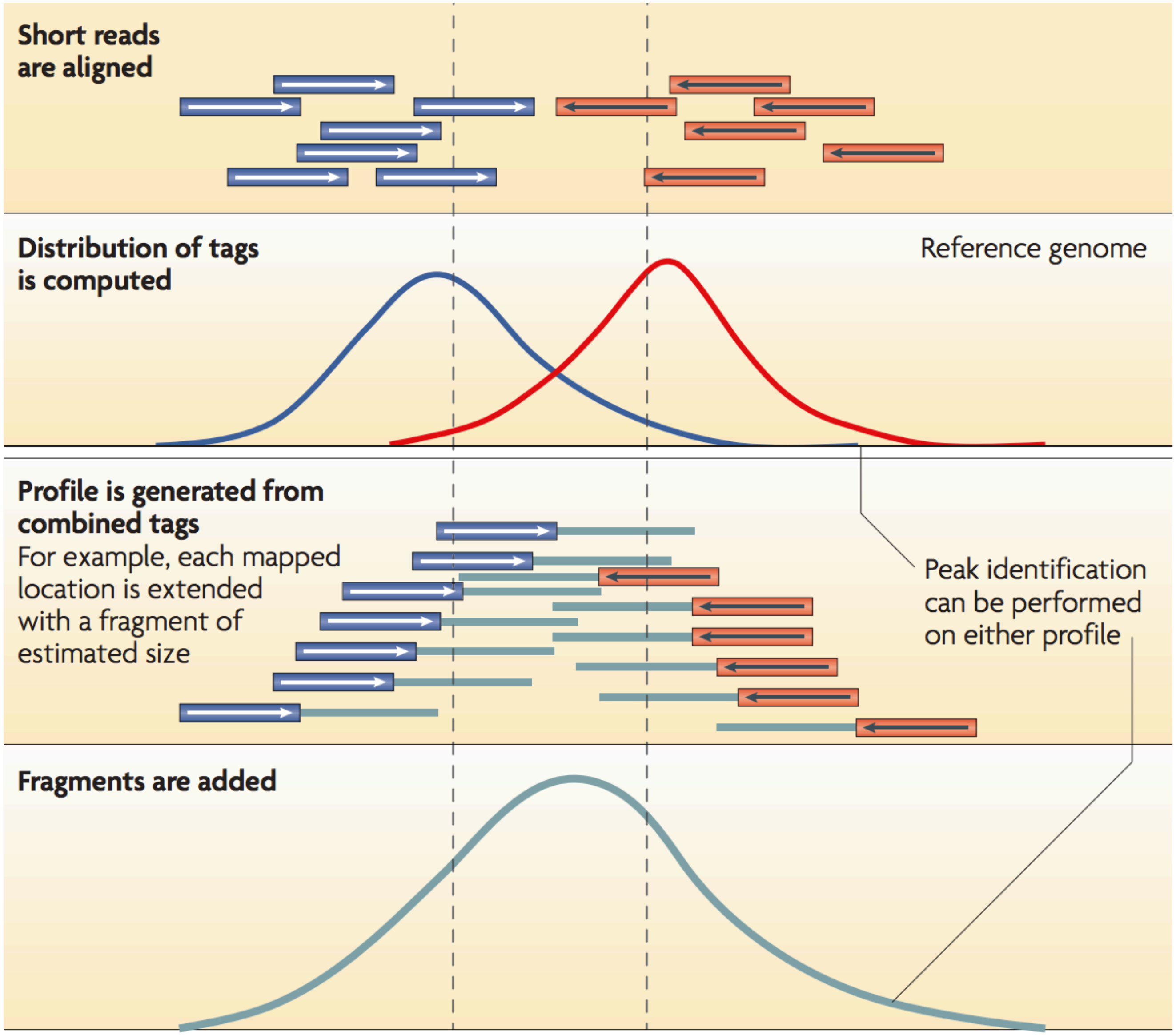


broad signal

-

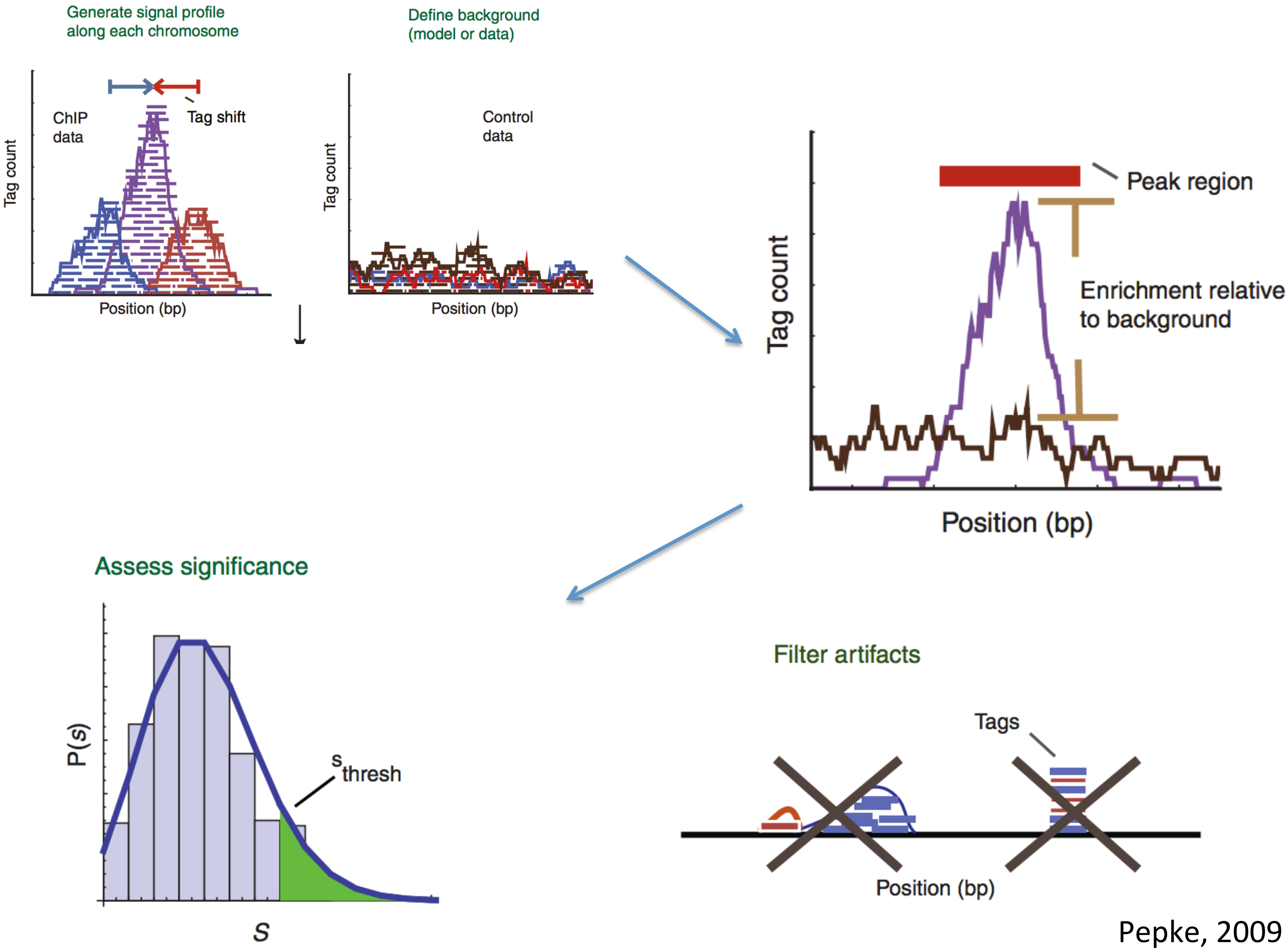
This is an active area of algorithm development

Principle of peak detection


Short reads are aligned
Distribution of tags is computed
Reference genome
Profile is generated from combined tags
For example, each mapped location is extended with a fragment of estimated size
Peak identification can be performed on either profile
Fragments are added

Asymmetry in reads mapped to opposite DNA strands

Computation of enrichment model



Generate signal profile along each chromosome

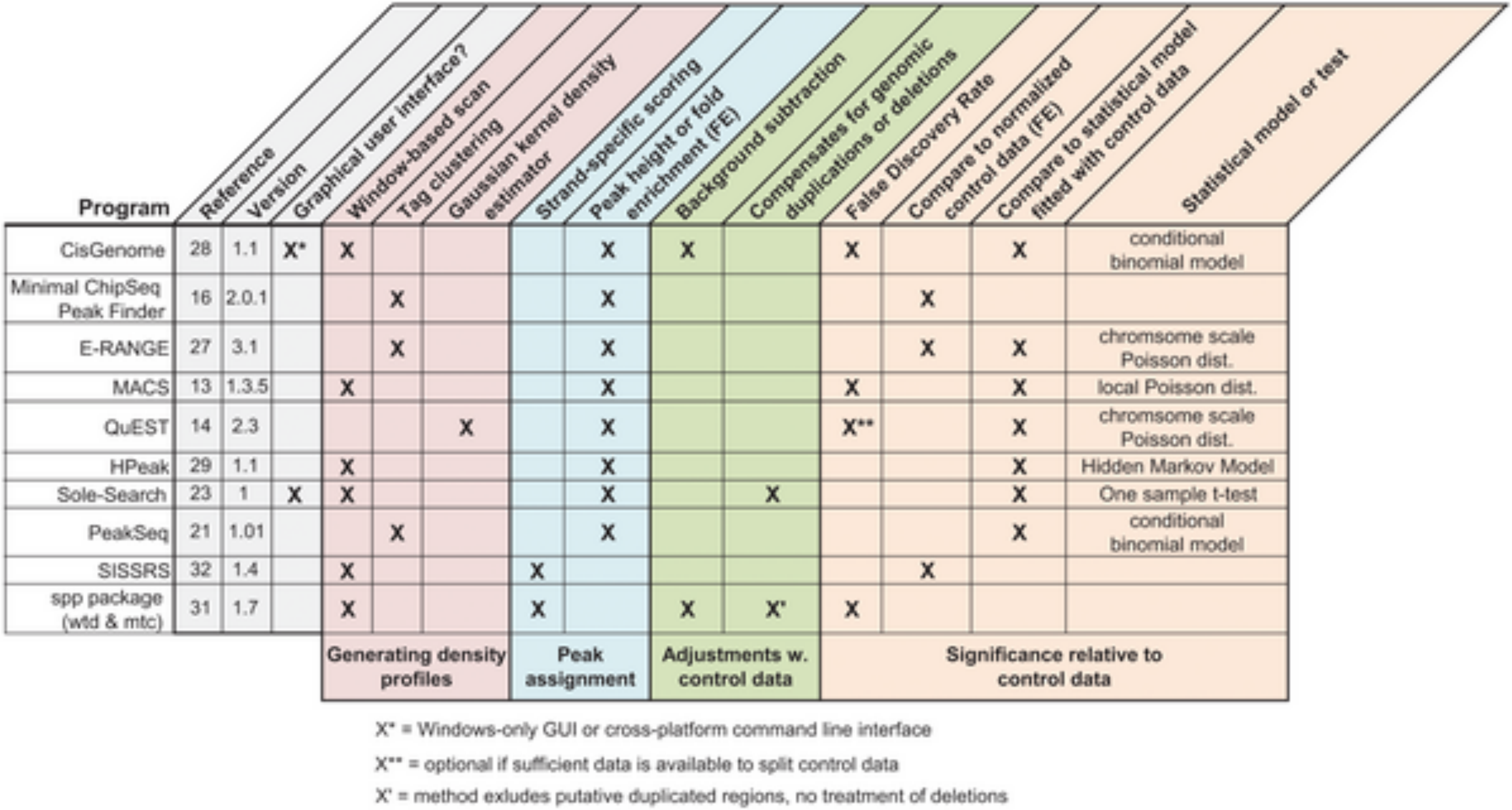
Define background (model or data)

Assess significance

Filter artifacts

Pepke, 2009

Comparison of peak calling algorithms



Program	Reference	Version	Graphical user interface?	Window-based scan	Tag clustering	Gaussian kernel density estimator	Strand-specific scoring	Peak height or fold enrichment (FE)	Background subtraction	Compensates for genomic duplications or deletions	False Discovery Rate	Compare to normalized control data (FE)	Compare to statistical model fitted with control data	Statistical model or test
CisGenome	28	1.1	X*	X				X	X		X		X	conditional binomial model
Minimal ChipSeq Peak Finder	16	2.0.1			X			X				X		
E-RANGE	27	3.1			X			X				X	X	chromsome scale Poisson dist.
MACS	13	1.3.5		X				X			X		X	local Poisson dist.
QuEST	14	2.3				X		X			X**		X	chromsome scale Poisson dist.
HPeak	29	1.1		X				X					X	Hidden Markov Model
Sole-Search	23	1	X	X				X		X			X	One sample t-test
PeakSeq	21	1.01			X			X					X	conditional binomial model
SISSRS	32	1.4		X			X					X		
spp package (wtd & mtc)	31	1.7		X			X		X	X'	X			
				Generating density profiles			Peak assignment		Adjustments w. control data		Significance relative to control data			

X* = Windows-only GUI or cross-platform command line interface

X** = optional if sufficient data is available to split control data

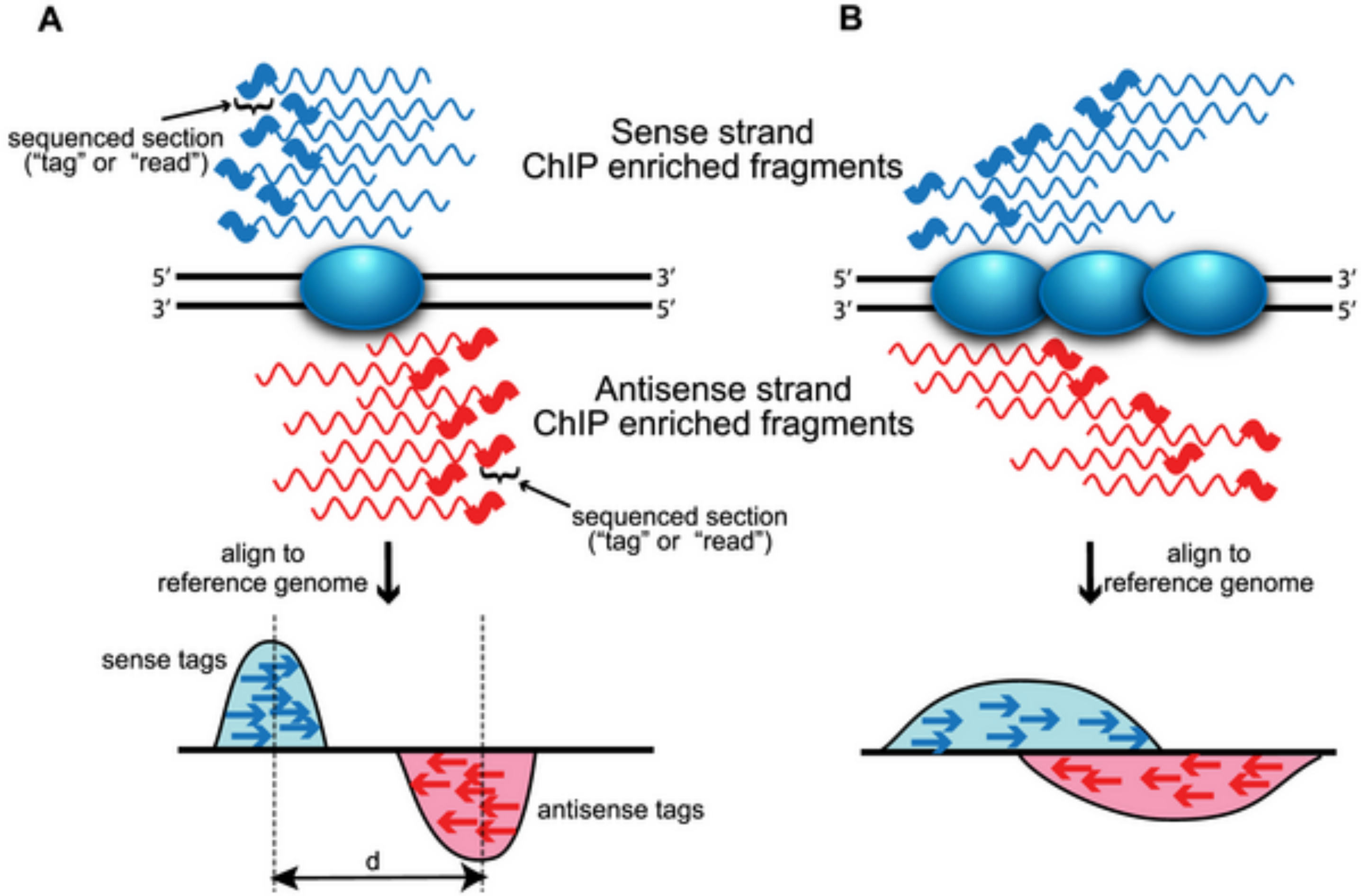
X' = method exludes putative duplicated regions, no treatment of deletions

Wilbanks 2010

Point-source vs. broad peak detection

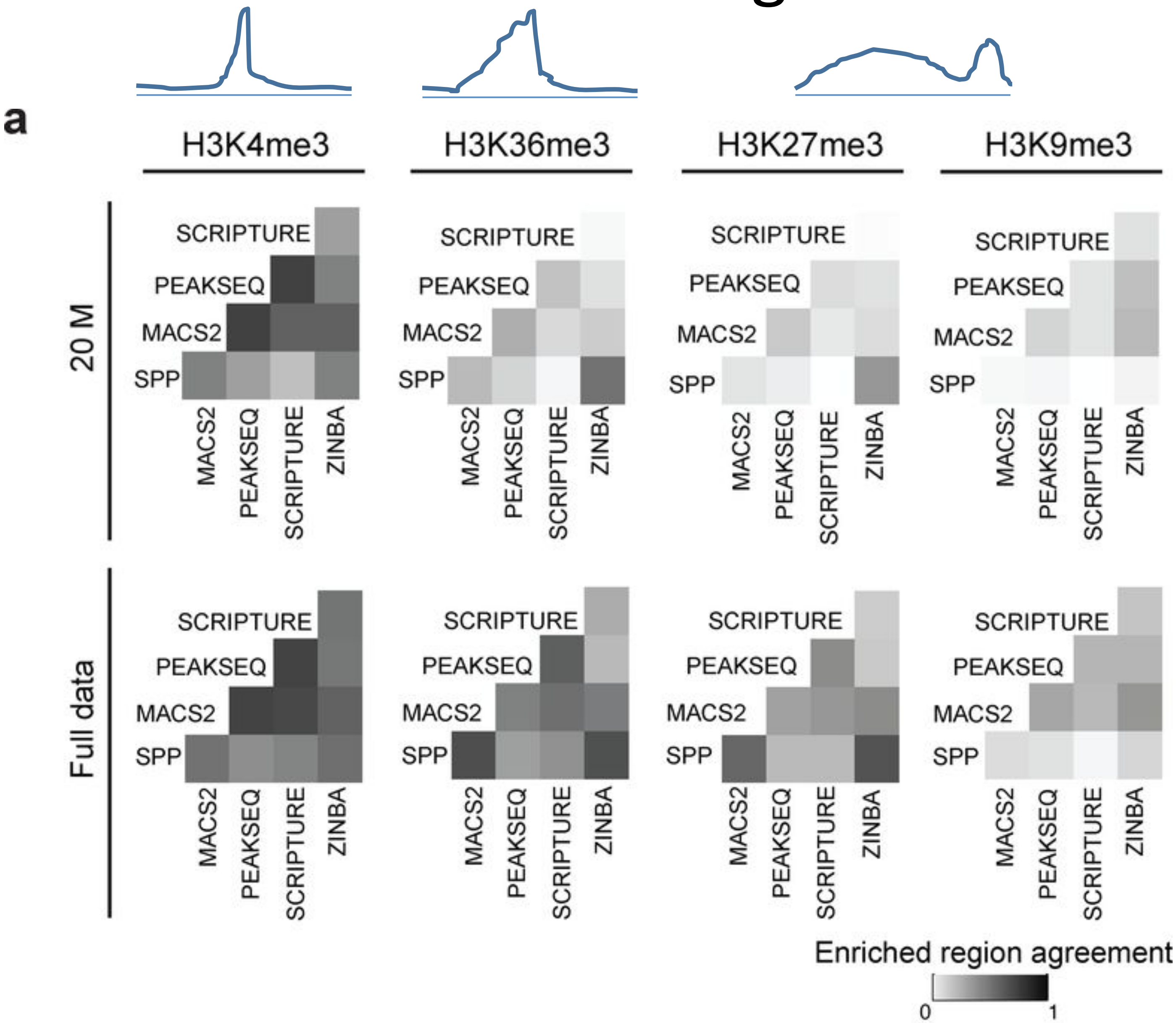
Sequence-specific binding (TFs)

Distributed binding (histones, RNApol)



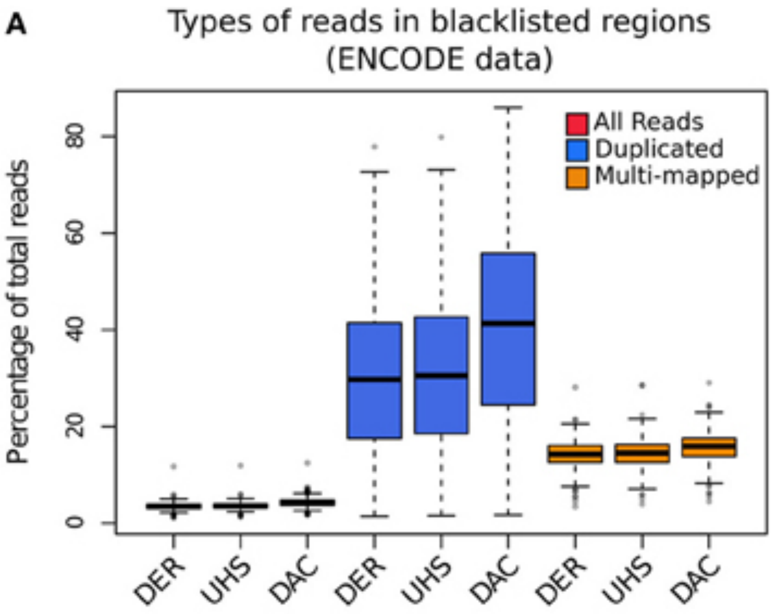
Wilbanks 2010

Comparison of enriched regions detected by various algorithms



Jung 2014

“Hyper-chippable” regions



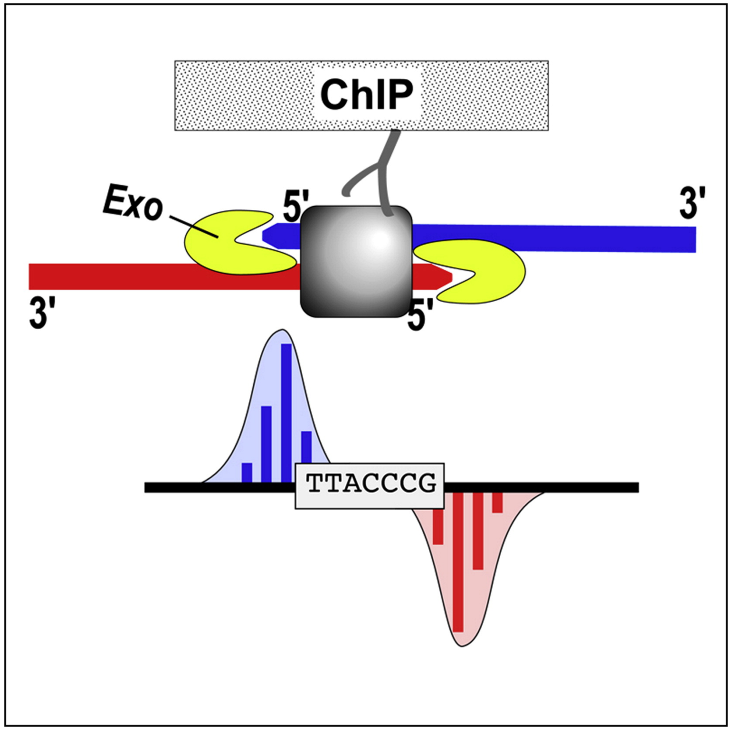
Reads mapped to these regions should be filtered out prior to peak calling

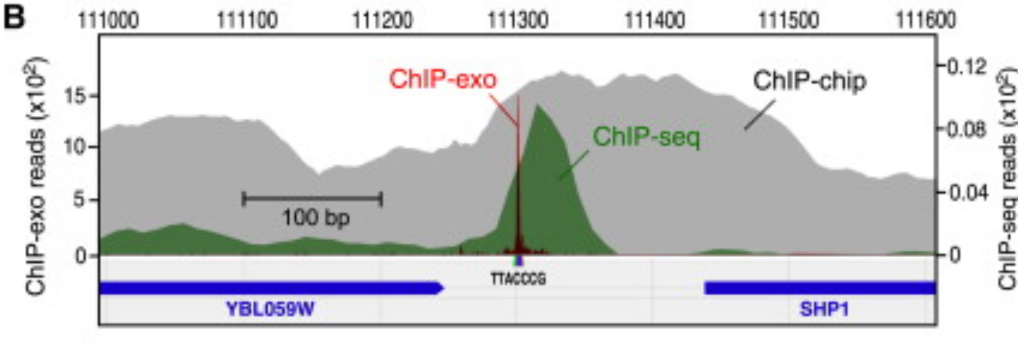
Tracks available from UCSC for human, mouse, fly and worm

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

Carroll et al, Front Genet 2014

ChIP-exo: improvement in binding site identification





Rhee and Pugh, Cell 2011

IDR = Irreproducible Discovery Rate

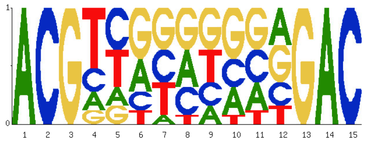
- Measure of consistency between replicates
- IDR describes the expected probability that the selected signals come from the “error” group for a given threshold; the “error” group for IDR refers to the irreproducible (inconsistent between replicates) group
- The selection made by IDR criterion is a combined results of ranking of the significance scores on individual replicates and consistency between replicates.
- For example, signals that have consistent rankings on both replicates but moderately ranked may be selected before the signals that have a very high score on one replicate but low on the other.

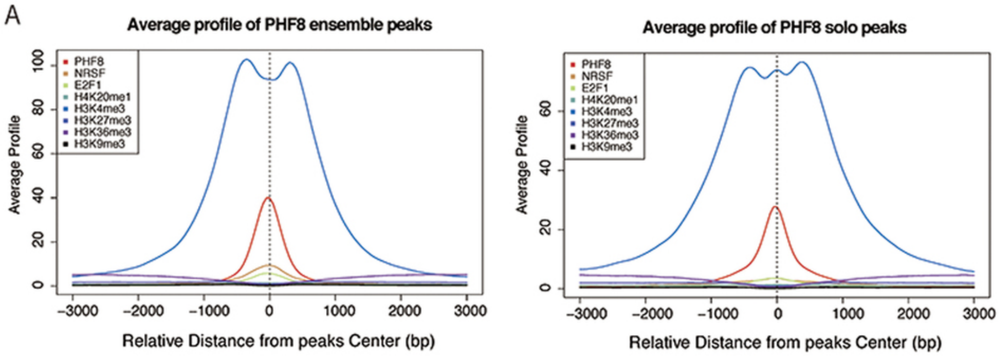
- ChIP – sequencing: introduction from a bioinformatics point of view
- Principles of analysis of ChIP-seq data
- ChIP-seq: downstream analyses
- Resources
- Exercise overview

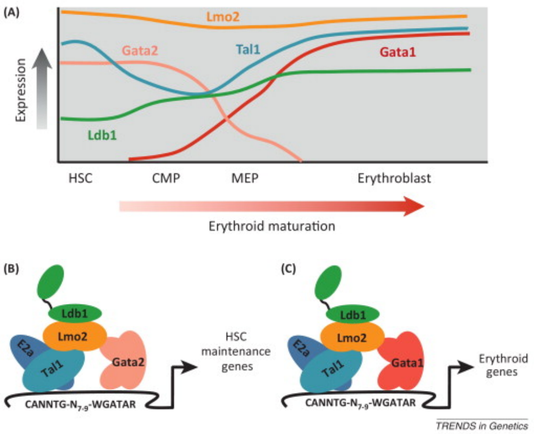
ChIPseq downstream analyses

- Validation

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







Peak annotation

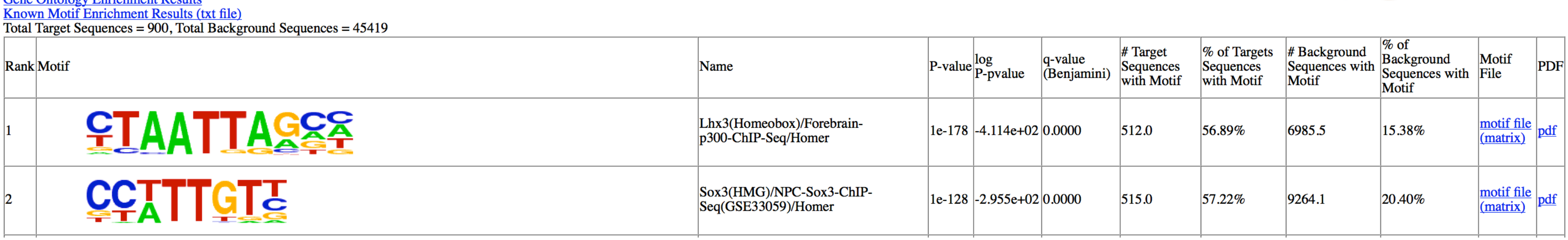
- Identification of nearest genomic features
- BEDTools,
- BEDops,
- PeakAnnotator,
- CisGenome,
- In R: ChIPPeakAnno

Motif detection

- Enrichment of known sequence motifs (CEAS, Transfac Match, HOMER)
- *De novo* motif detection (MEME, CisFinder, HMS, DREME, ChIPMunk, HOMER)

Enrichment of known motifs (Homer):

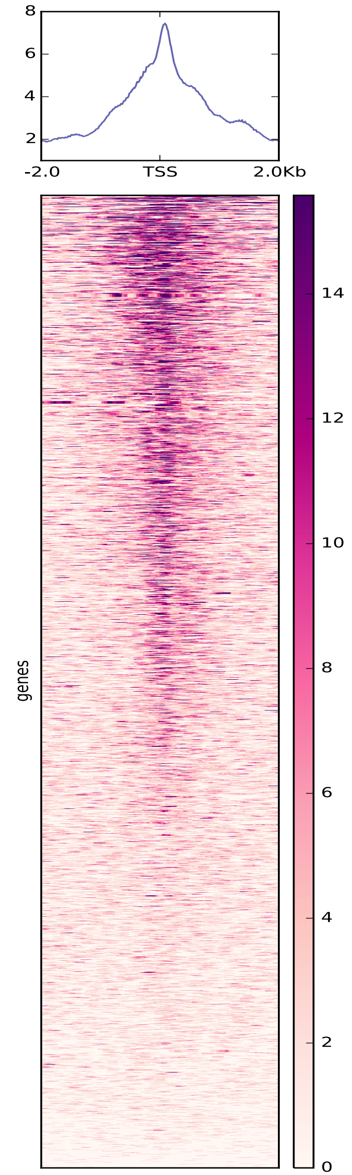
Homer Known Motif Enrichment Results



Homer *de novo* Motif Results
Gene Ontology Enrichment Results
Known Motif Enrichment Results (txt file)
Total Target Sequences = 900, Total Background Sequences = 45419

Rank	Motif	Name	P-value	log P-pvalue	q-value (Benjamini)	# Target Sequences with Motif	% of Targets Sequences with Motif	# Background Sequences with Motif	% of Background Sequences with Motif	Motif File	PDF
1		Lhx3(Homeobox)/Forebrain-p300-ChIP-Seq/Homer	1e-178	-4.114e+02	0.0000	512.0	56.89%	6985.5	15.38%	motif file (matrix)	pdf
2		Sox3(HMG)/NPC-Sox3-ChIP-Seq(GSE33059)/Homer	1e-128	-2.955e+02	0.0000	515.0	57.22%	9264.1	20.40%	motif file (matrix)	pdf

Signal visualisation and interpretation



deepTools
Ngsplots
seqMiner

- Clustering
- Heatmaps
- Profiles
- Comparison of different datasets

Mapping of a TF in relation to the transcription start site

Differential occupancy

- Use algorithms developed for differential expression and summarise reads found in peaks; normalisation; statistical testing; R environment
  - edgeR
  - DESeq(2)
  - DiffBind (implements several normalisation methods)

- Calculate enrichment in sliding windows
  - DROMPA
  - Diffreps

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

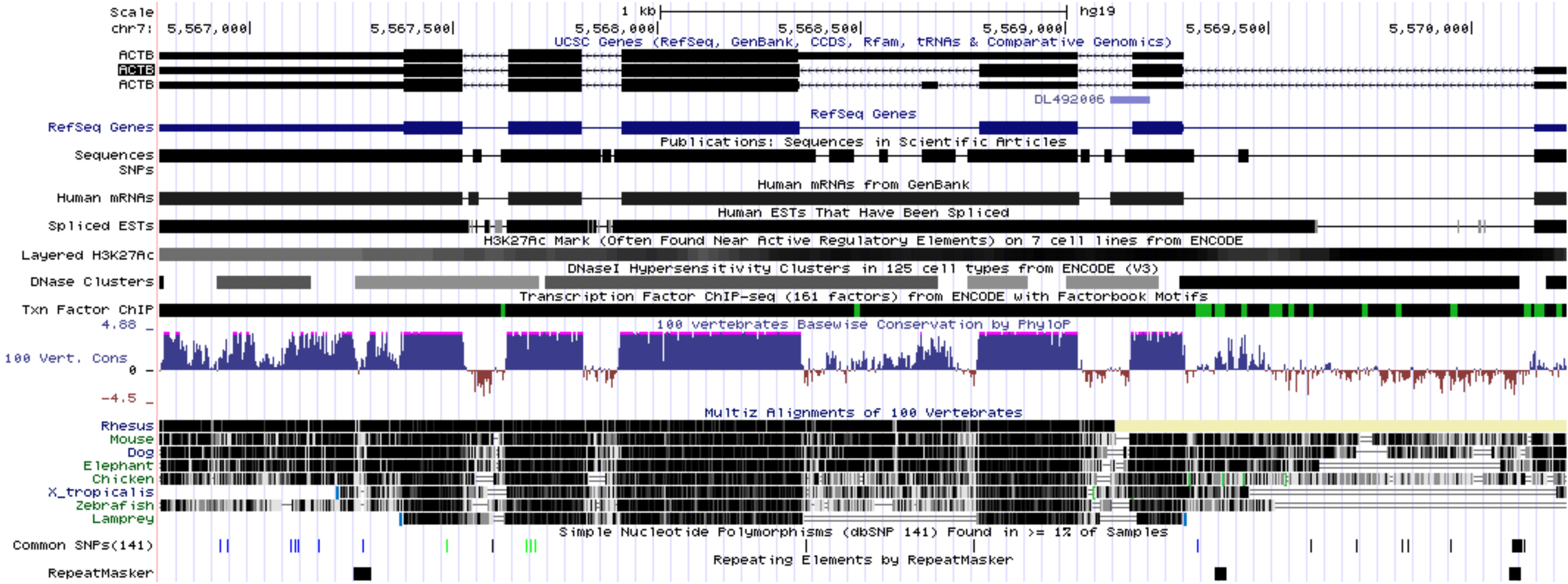
Where to obtain data?

The ENCODE project

www.encodeproject.org

- Encyclopedia of DNA elements
- Identification of regulatory DNA elements in human (and mouse) genome
- www.encodeproject.org
- 240 human and 55 mouse DNA binding proteins
- 1464 human and 432 mouse samples
- RNA profiling, protein-DNA interaction, chromatin condensation, DNA methylation, …
- 2009 - ongoing

Human ACTB locus as seen in the UCSC Genome Browser


Scale
1 kb
hg19
chr7:
5,567,000
5,567,500
5,568,000
5,568,500
5,569,000
5,569,500
5,570,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ACTB
ACTB
ACTB
DL492006
RefSeq Genes
RefSeq Genes
Publications: Sequences in Scientific Articles
Sequences
SNPs
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
4.88 _
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0 _
-4.5 _
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 141) Found in >= 1% of Samples
Common SNPs(141)
Repeating Elements by RepeatMasker
RepeatMasker

Gene model

Alternative transcripts

Histone modifications

Chromatin structure

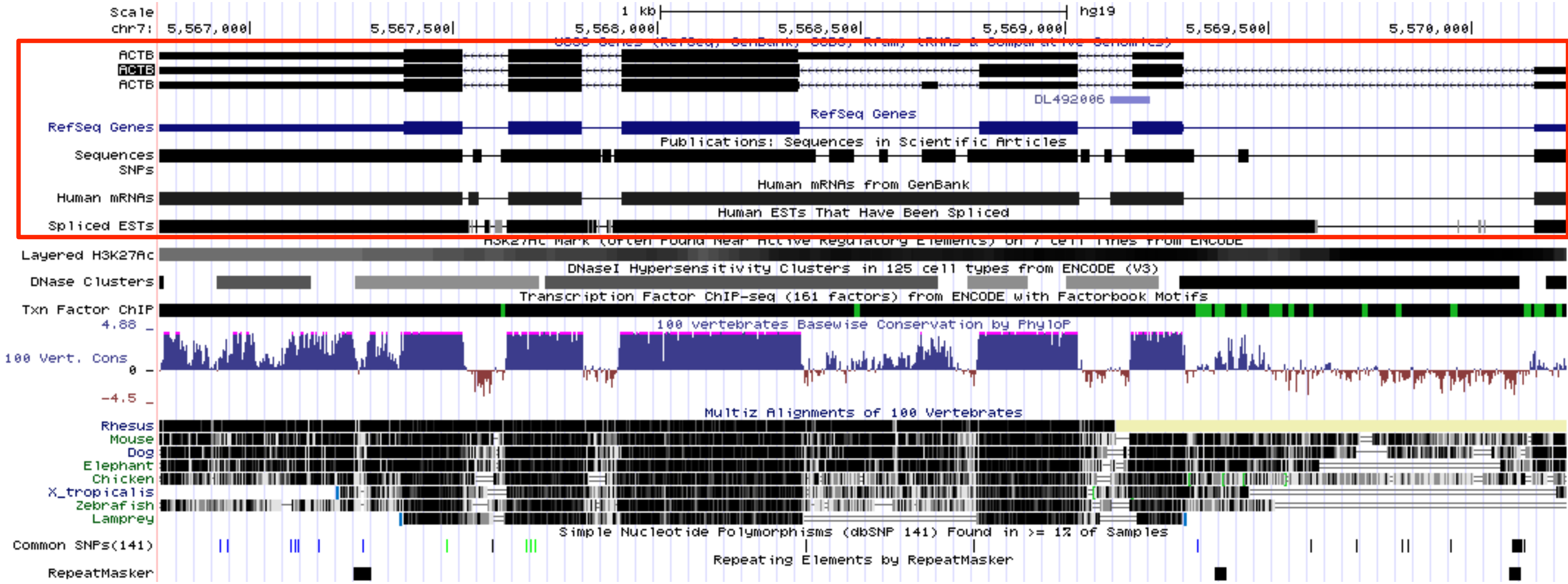
Transcription factor binding sites

DNA conservation

Single nucleotide polymorphisms (SNP)

Repeats

Human ACTB locus as seen in the UCSC Genome Browser


Scale
chr7:
5,567,000
5,567,500
5,568,000
5,568,500
5,569,000
5,569,500
5,570,000
1 kb
hg19
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ACTB
ACTB
ACTB
DL492006
RefSeq Genes
RefSeq Genes
Publications: Sequences in Scientific Articles
Sequences
SNPs
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 141) Found in >= 1% of Samples
Common SNPs(141)
Repeating Elements by RepeatMasker
RepeatMasker

Gene model

Alternative transcripts

Histone modifications

Chromatin structure

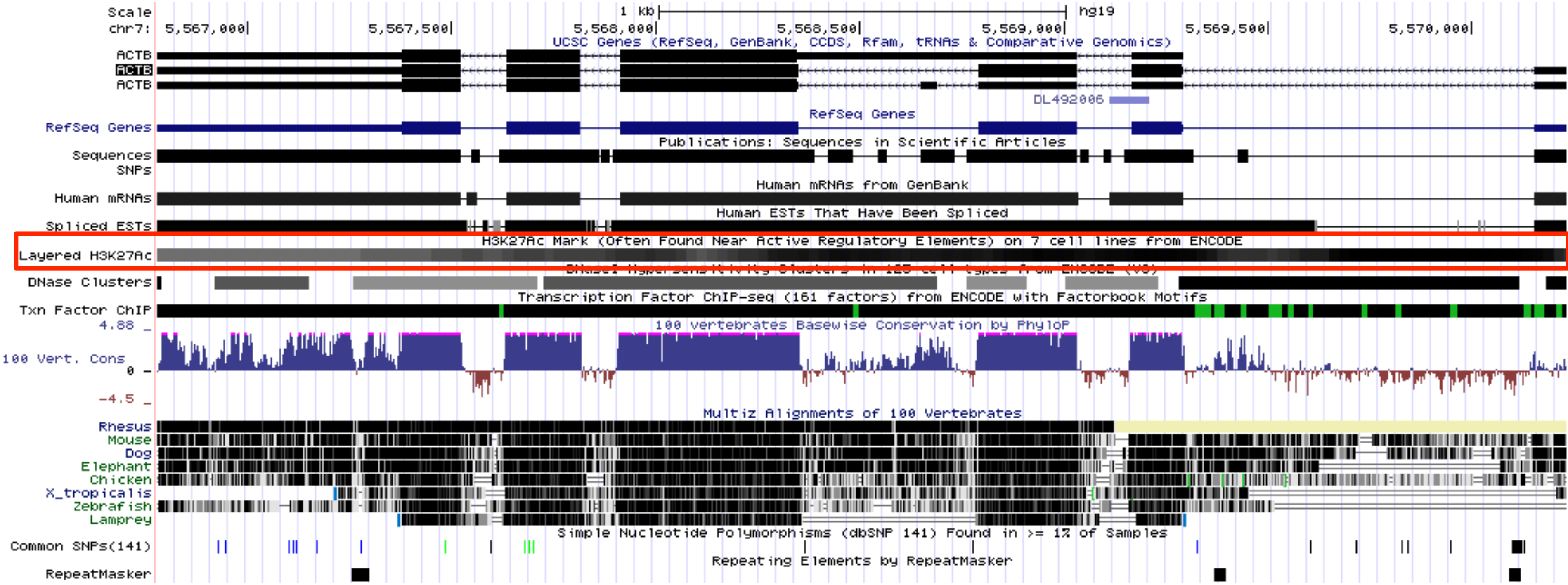
Transcription factor binding sites

DNA conservation

Single nucleotide polymorphisms (SNP)

Repeats

Human ACTB locus as seen in the UCSC Genome Browser


Scale
chr7:
1 kb
hg19
5,567,000
5,567,500
5,568,000
5,568,500
5,569,000
5,569,500
5,570,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ACTB
ACTB
ACTB
DL492006
RefSeq Genes
RefSeq Genes
Publications: Sequences in Scientific Articles
Sequences
SNPs
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V2)
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
4.88
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0
-4.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 141) Found in >= 1% of Samples
Common SNPs(141)
Repeating Elements by RepeatMasker
RepeatMasker

Gene model

Alternative transcripts

Histone modifications

Chromatin structure

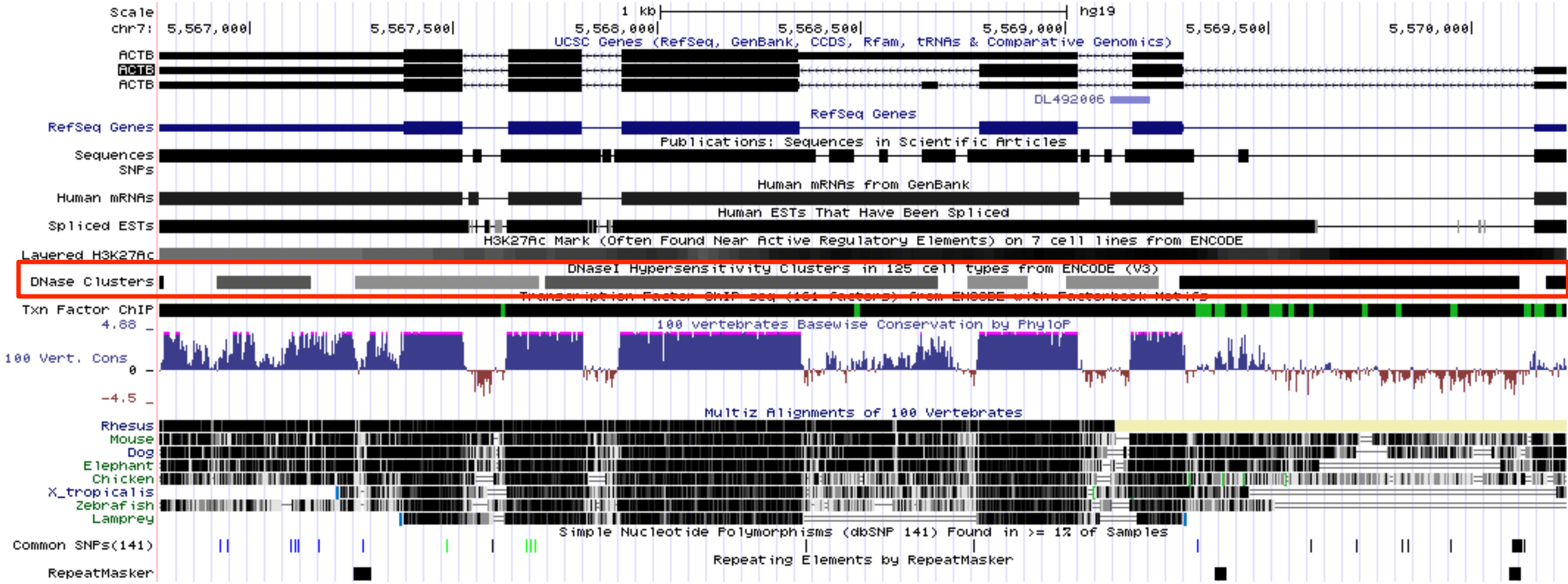
Transcription factor binding sites

DNA conservation

Single nucleotide polymorphisms (SNP)

Repeats

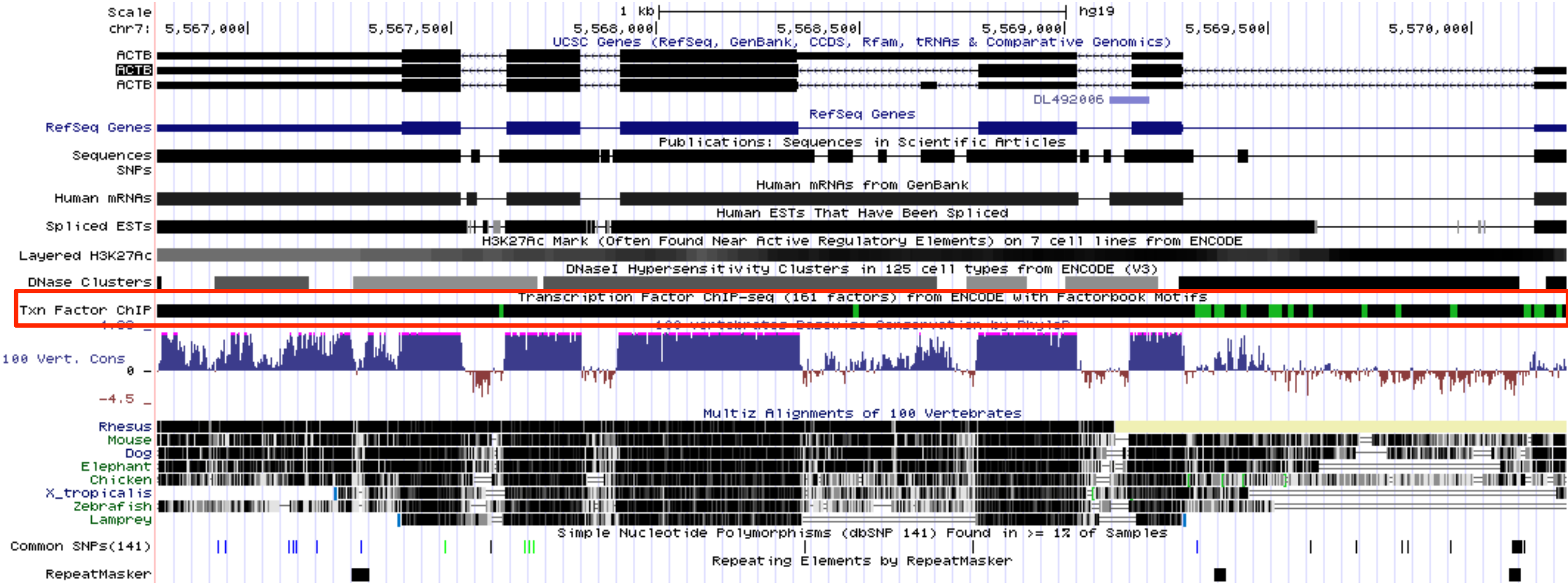
Human ACTB locus as seen in the UCSC Genome Browser


Scale
1 kb
hg19
chr7: 5,567,000 5,567,500 5,568,000 5,568,500 5,569,000 5,569,500 5,570,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ACTB
ACTB
ACTB
DL492006
RefSeq Genes
RefSeq Genes
Publications: Sequences in Scientific Articles
Sequences
SNPs
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
100 vertebrates Basewise Conservation by PhyloP
4.88
100 Vert. Cons
0
-4.5
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 141) Found in >= 1% of Samples
Common SNPs(141)
Repeating Elements by RepeatMasker
RepeatMasker

Gene model
Alternative transcripts
Histone modifications
Chromatin structure

Transcription factor binding sites
DNA conservation
Single nucleotide polymorphisms (SNP)
Repeats

Human ACTB locus as seen in the UCSC Genome Browser


Scale
chr7:
1 kb
hg19
5,567,000
5,567,500
5,568,000
5,568,500
5,569,000
5,569,500
5,570,000
UCSC Genes (RefSeq, GenBank, CCDS, Rfam, tRNAs & Comparative Genomics)
ACTB
ACTB
ACTB
DL492006
RefSeq Genes
RefSeq Genes
Publications: Sequences in Scientific Articles
Sequences
SNPs
Human mRNAs from GenBank
Human mRNAs
Human ESTs That Have Been Spliced
Spliced ESTs
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Layered H3K27Ac
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
DNase Clusters
Transcription Factor ChIP-seq (161 factors) from ENCODE with Factorbook Motifs
Txn Factor ChIP
1.00 _
100 vertebrates Basewise Conservation by PhyloP
100 Vert. Cons
0 -
-4.5 _
Multiz Alignments of 100 Vertebrates
Rhesus
Mouse
Dog
Elephant
Chicken
X_tropicalis
Zebrafish
Lamprey
Simple Nucleotide Polymorphisms (dbSNP 141) Found in >= 1% of Samples
Common SNPs(141)
Repeating Elements by RepeatMasker
RepeatMasker

Gene model
Alternative transcripts
Histone modifications
Chromatin structure

Transcription factor binding sites
DNA conservation
Single nucleotide polymorphisms (SNP)
Repeats

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 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 ChIPseq 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 ChIPseq 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)
- Quality control
  - ChIPQC
- Differential expression
  - edgeR
  - DESeq, DESeq2
  - DiffBind (compatible with objects used for ChIPQC, wrapper for DESeq and edgeR DE functions)
- Peak Annotation
  - ChIPpeakAnno (annotating peaks with genome context information)

- 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. Visualisation
- 5. Statistical analysis of differential occupancy

This afternoon & tomorrow morning

Questions?

agata.smialowska@bils.se

That’s all for now,

time to do some hands-on work

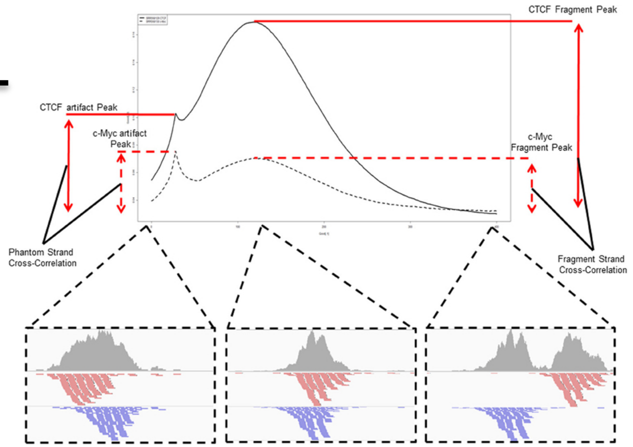
Cross-correlation profiles, RSC and NSC

- 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

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

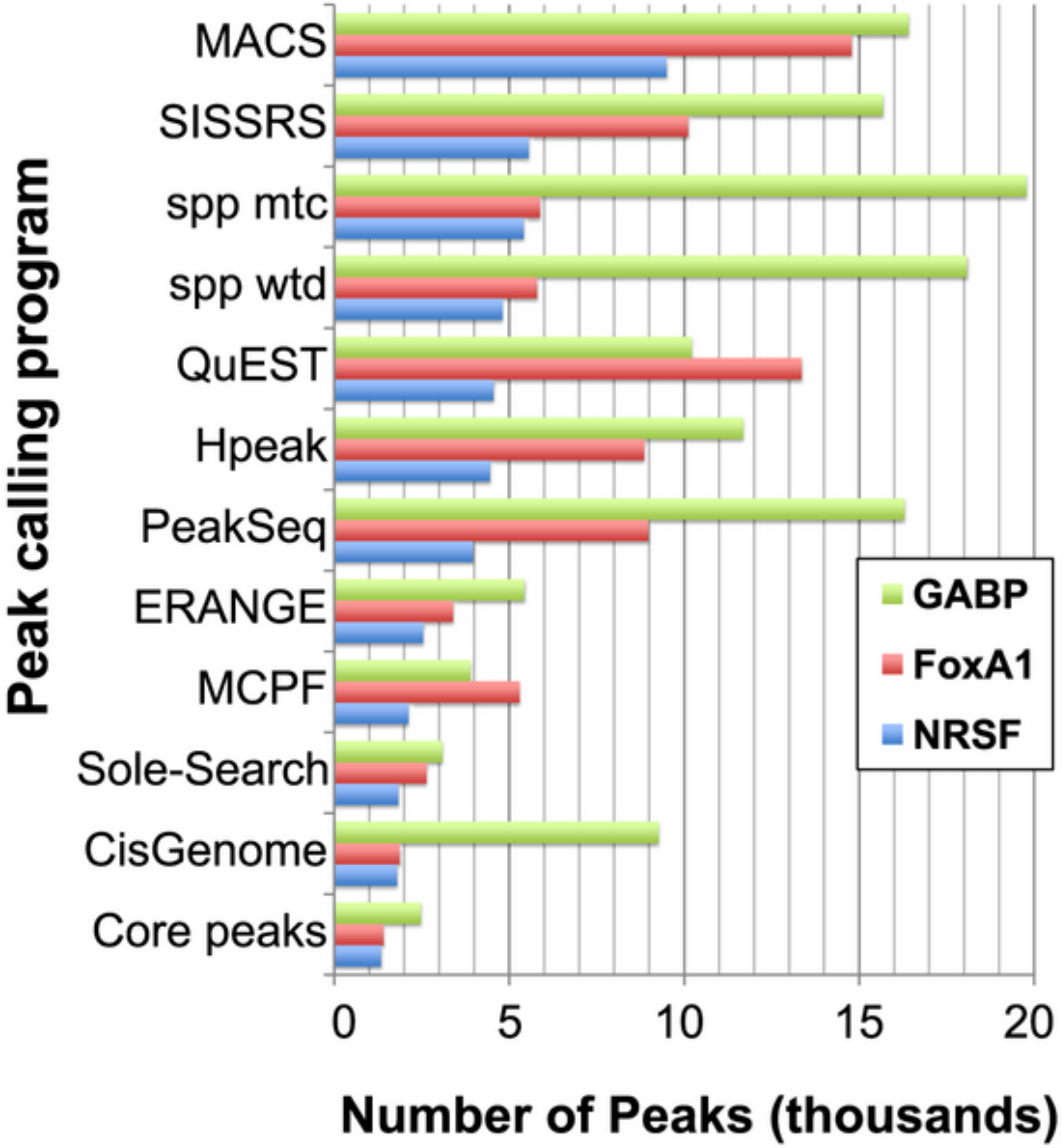
- Normalised Cross Correlation (NSC)

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

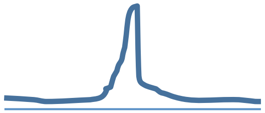


- TFs: fragment lengths are often greater than the size of the DNA binding event, the distinct clustering of (+) and (-) reads around this site is very apparent
- NSC>1.1 (higher values indicate more enrichment; 1 = no enrichment)
- RSC>0.8 (0 = no signal; <1 low quality ChIP; >1 high enrichment
- Broad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

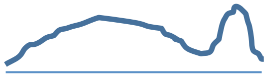
Comparison of peak calling algorithms



Peak overlap (Ho et al, 2012)



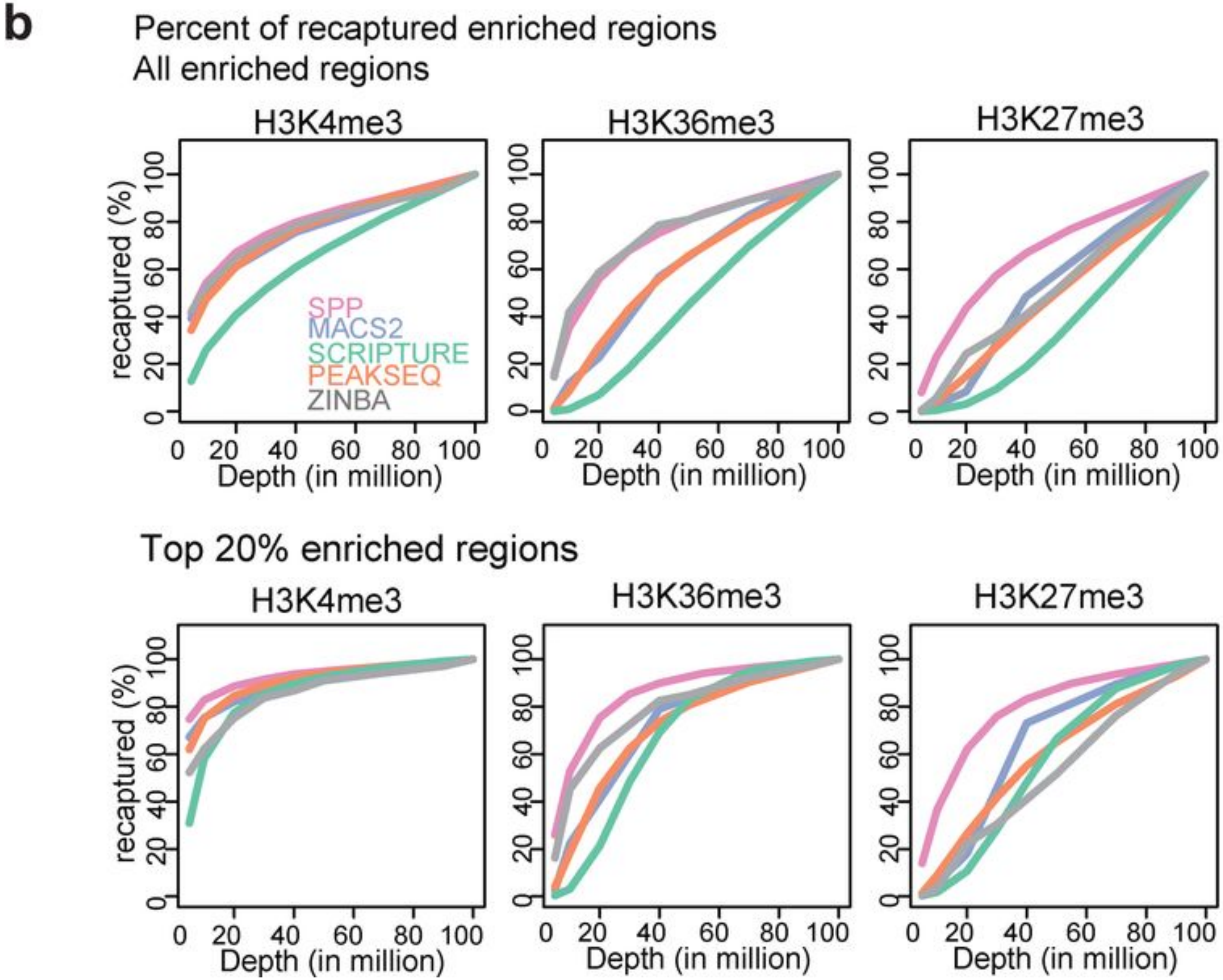
> 50 %



20 %

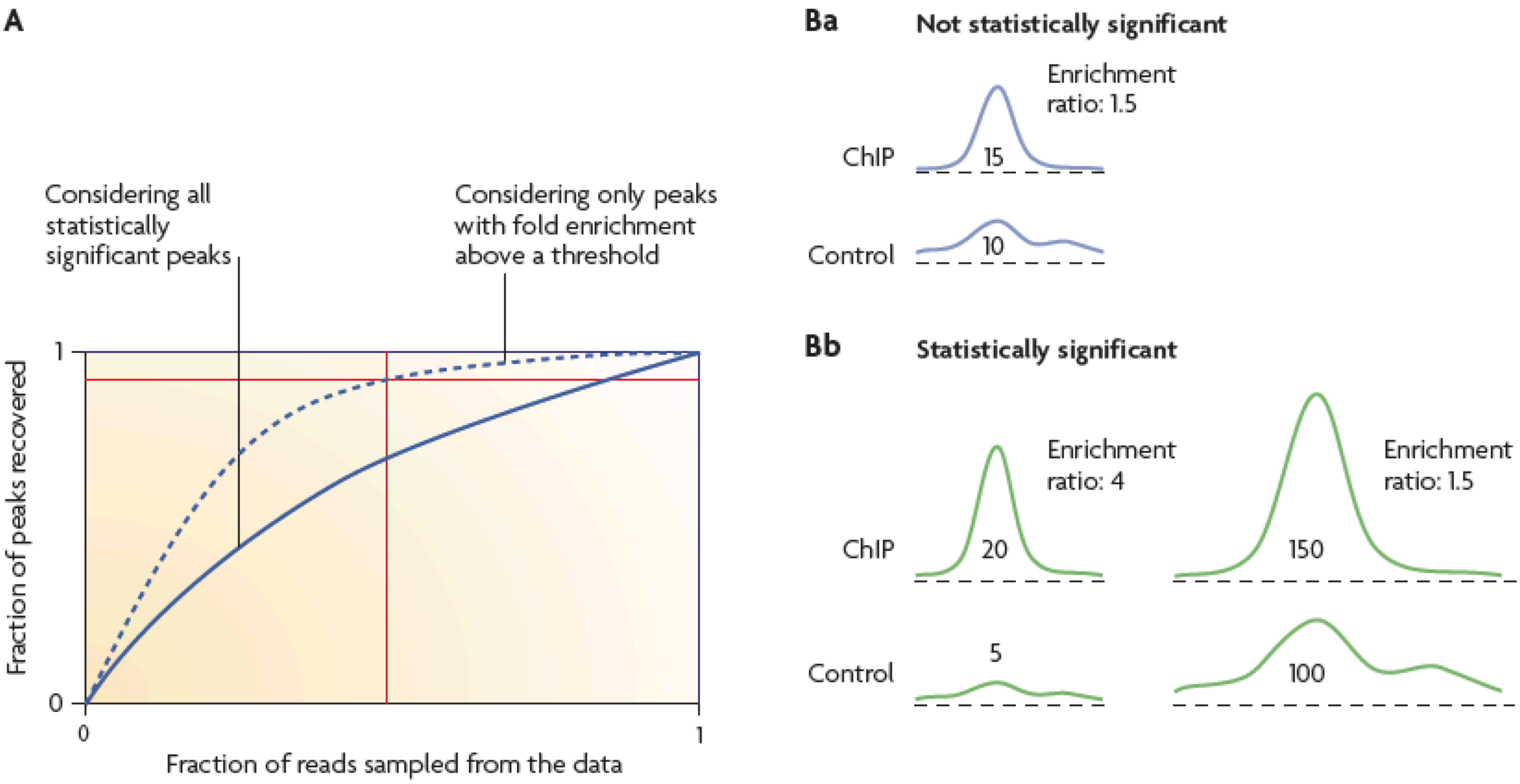
Wilbanks 2010

Effect of sequencing depth on regions detected by various algorithms



Jung 2014

Fold enrichment = signal / background


A
Considering all statistically significant peaks
Considering only peaks with fold enrichment above a threshold
Fraction of peaks recovered
Fraction of reads sampled from the data
Ba
Not statistically significant
Enrichment ratio: 1.5
ChIP
Control
Bb
Statistically significant
Enrichment ratio: 4
Enrichment ratio: 1.5
ChIP
Control