CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Design and analysis of ChIP-seq experiments

RORY STARK
UNIVERSITY OF CAMBRIDGE – CRUK

4 NOVEMBER 2016

Chapter 10
Characterization of DNA-Protein Interactions: Design and Analysis of ChIP-Seq Experiments

Rory Stark and James Hadfield

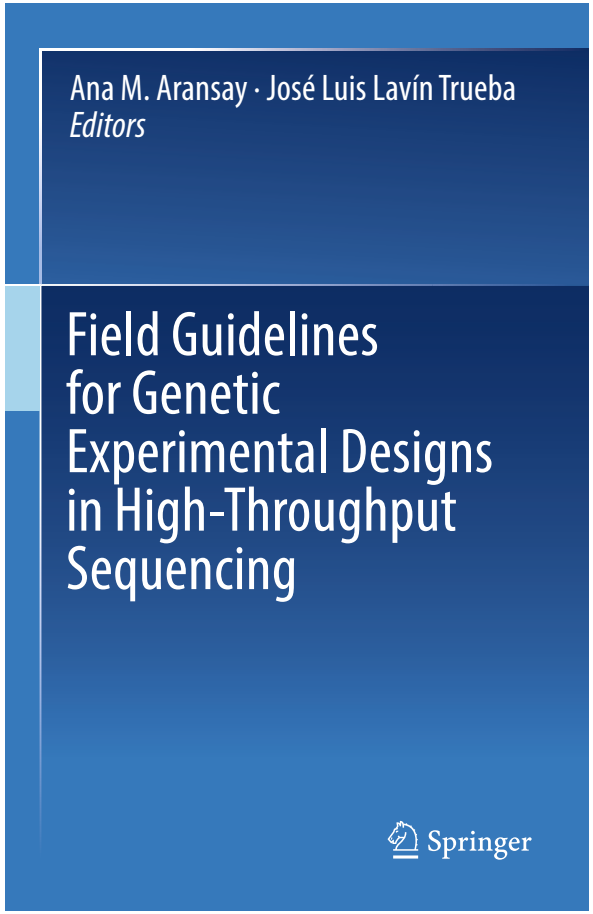
10.1 Introduction to Genome-Wide Analysis of DNA-Protein Interactions Using ChIP-seq

Within the last decade, advances in high-throughput sequencing have enabled extensive research into protein-DNA interactions on a genomic scale. These interactions include the binding of transcription factor proteins to localized positions on DNA, as well as proteins involved in other aspects of transcriptional regulation (e.g., methylases, acetylases) and in transcription itself (polymerases, etc.). The same methods can further be used to ascertain relevant aspects of chromatin state involved in transcriptional regulation, most notably key histone "marks" (including methylation and acetylation).

The primary experimental method used is chromatin immunoprecipitation followed by sequencing, or ChIP-seq. While ChIP assays have been utilized for some time, modern high-throughput sequencing has enabled the entire genome (rather than just a small number of genes or genomic loci) to be interrogated in a single experiment. Figure 10.1, generated by the ENCODE project (ENCODE Project Consortium 2011), shows the high-level picture of regulatory elements in the genome, including the aspects that may be examined using ChIP-seq. This chapter describes how to design, implement, and analyze ChIP-seq experiments to successfully address a range of biological questions involving DNA-protein interactions and transcriptional regulation.

R. Stark, B.A., M.Sc., M.Phil., D.Phil. (✉) • J. Hadfield, B.Sc., Ph.D.
Cancer Research UK Cambridge Institute, University of Cambridge, Robinson Way, Cambridge CB2 0RE, UK
e-mail: rory.stark@cruk.cam.ac.uk; James.Hadfield@cruk.cam.ac.uk

A.M. Aransay, J.L. Lavín Trueba (eds.), *Field Guidelines for Genetic Experimental Designs in High-Throughput Sequencing*,
DOI 10.1007/978-3-319-31350-4_10



rory.stark@cruk.cam.ac.uk

UNIVERSITY OF CAMBRIDGE

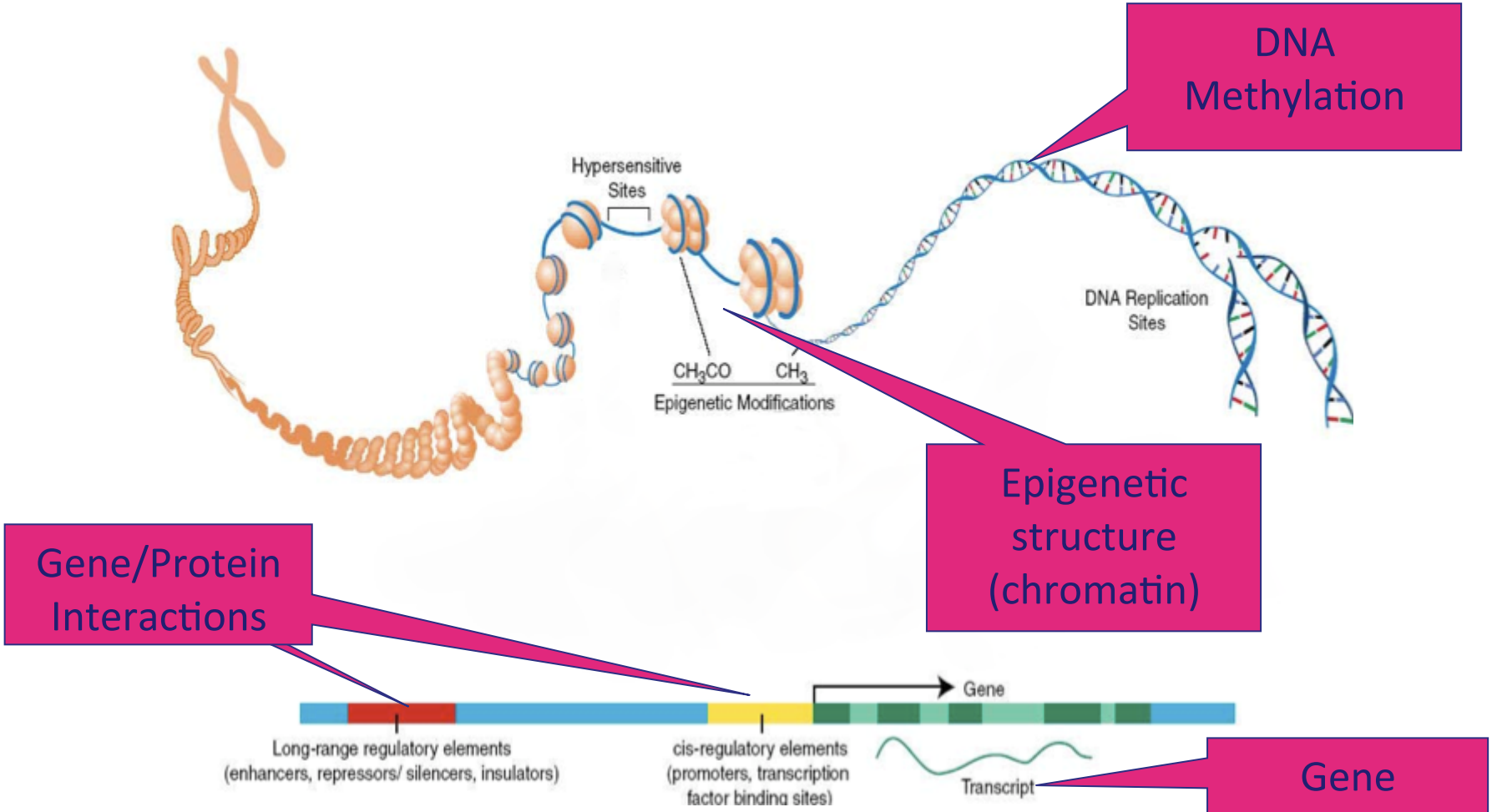
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Epigenomics and genome regulation

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Mapping regulatory elements



The ENCODE Project Consortium (2011) A User's Guide to the Encyclopedia of DNA Elements (ENCODE). PLoS Biol 9(4): e1001046. doi:10.1371/journal.pbio.1001046
http://127.0.0.1:8081/plosbiology/article?id=info:doi/10.1371/journal.pbio.1001046

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Regulatory elements of interest include…

TRANSCRIPTION FACTORS

- ChIP

HISTONE MARKS

- ChIP

DNA METHYLATION

- MeDIP etc.

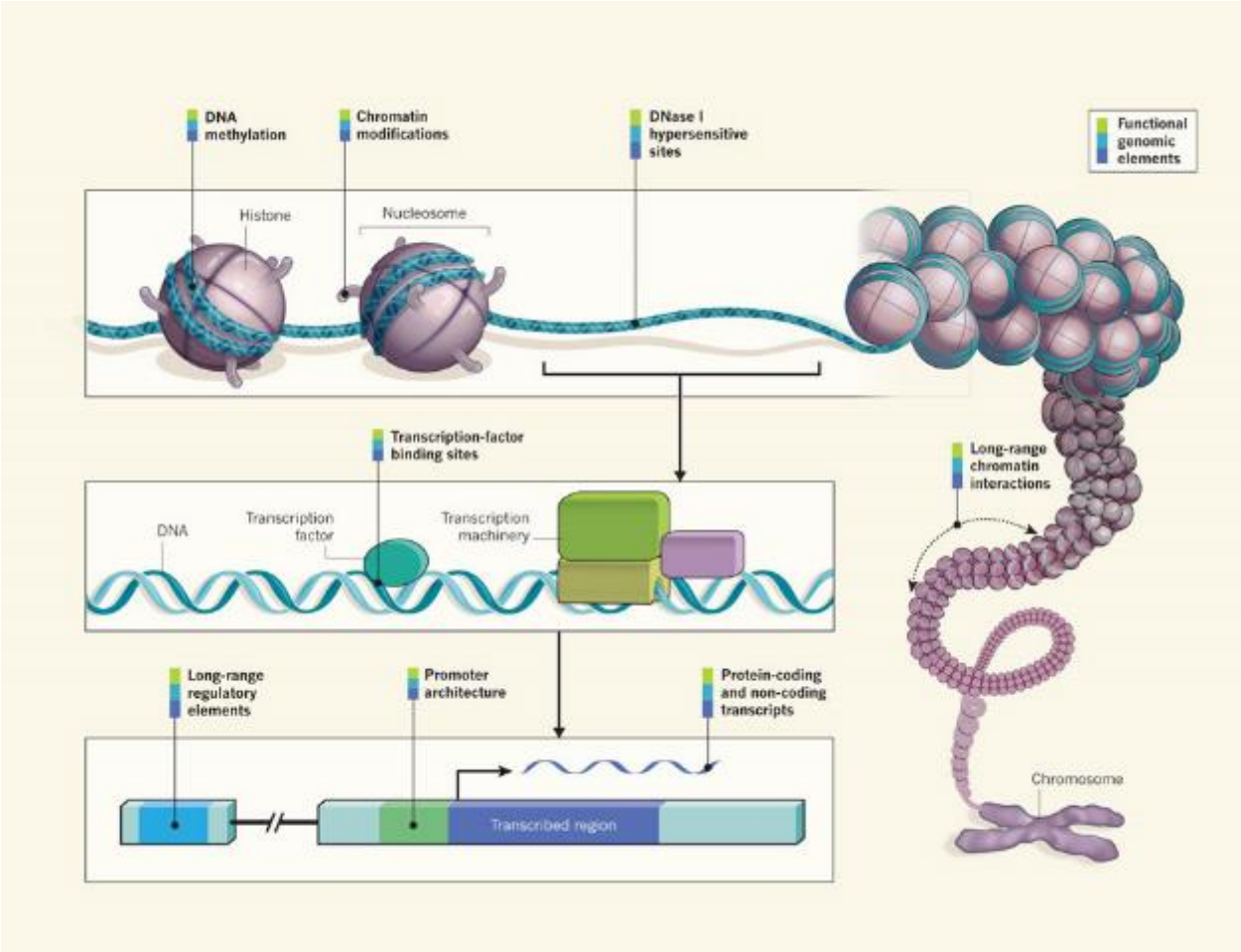
NUCLEOSOMES

RNA POLYMERASE

- Pol II ChIP

OPEN CHROMATIN

- DNase Hypersensitivity



UNIVERSITY OF CAMBRIDGE

Functional genomics and epigenomic analysis

Most **functional** studies to date have focused on RNA levels

- Well established design/analysis
- Unable to directly distinguish driver/upstream from passenger/ downstream changes
- Regulatory schema **inferred** (knockouts, modelling)

Most epigenomic studies to date have focused on **mapping**, not **function** (cf ENCODE)

- Comparisons limited to peak overlaps (co-occupancy)
- Limited quantitative analysis

Can we use ChIP-Seq to more directly **observe** regulatory events?

ChIP basics

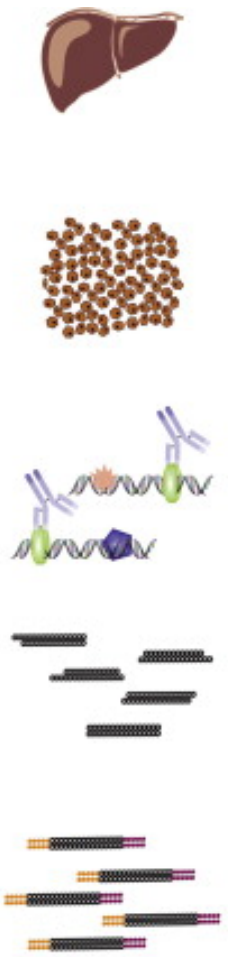
UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Mapping Protein/DNA interactions: Chromatin Immuno-Precipitation (ChIP)

Material

- Tissue
- Cross-linked cells
- Lysed chromatin fragments
- ChIPed DNA
- Sequencing library

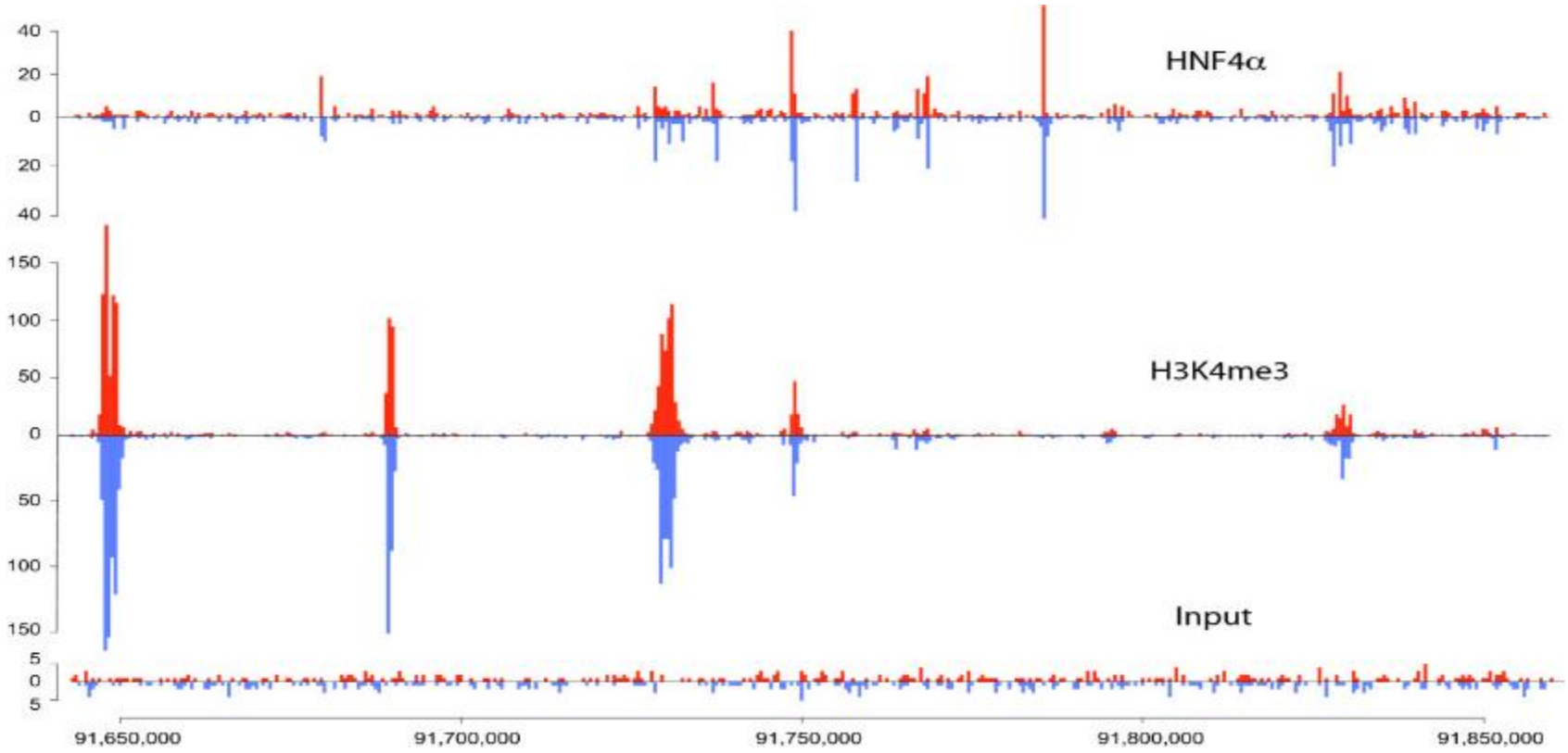


Process

- Isolate chromatin
- Cross-link
- Fragmentation
- Introduce antibody
- Precipitate
- Reverse cross-links
- Purify DNA
- Ligate adaptors
- Sequencing

Adapted from: *Schmidt, D., Wilson, M. D., Spyrou, C., Brown, G. D., Hadfield, J., & Odom, D. T. (2009). ChIP-seq: Using high-throughput sequencing to discover protein–DNA interactions. Methods, 48(3), 240-248.*

Precipitation assay efficiency



CAN YOU SEE THE "PEAKS"?

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

ChIP-seq computational challenges

DID THE CHIP WORK? (QA)

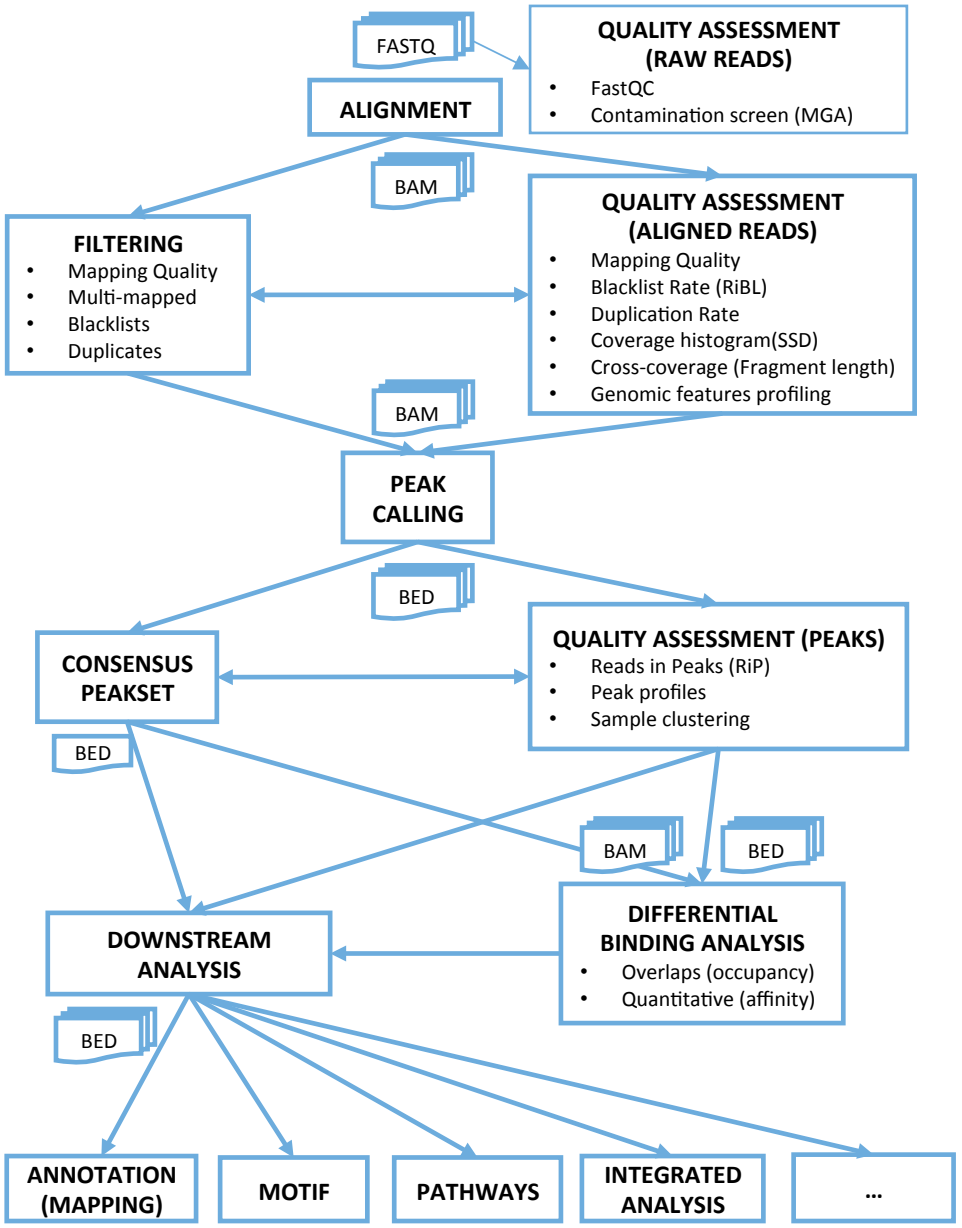
WHERE IS THE PROTEIN BOUND? (PEAK CALLING)

- In reality, only a small proportion of fragments are bound. Hopefully this enrichment can be detected computationally
- If bound fragments comprise 0.01% of sample, and ChIP enriches by 1000x, 90% of sequenced fragment will be background!

WHERE ARE PROTEINS BOUND DIFFERENTLY? (DIFFERENTIAL BINDING ANALYSIS)

WHAT DO THE BINDING SITES SIGNIFY BIOLOGICALLY? (DOWNSTREAM ANALYSIS)

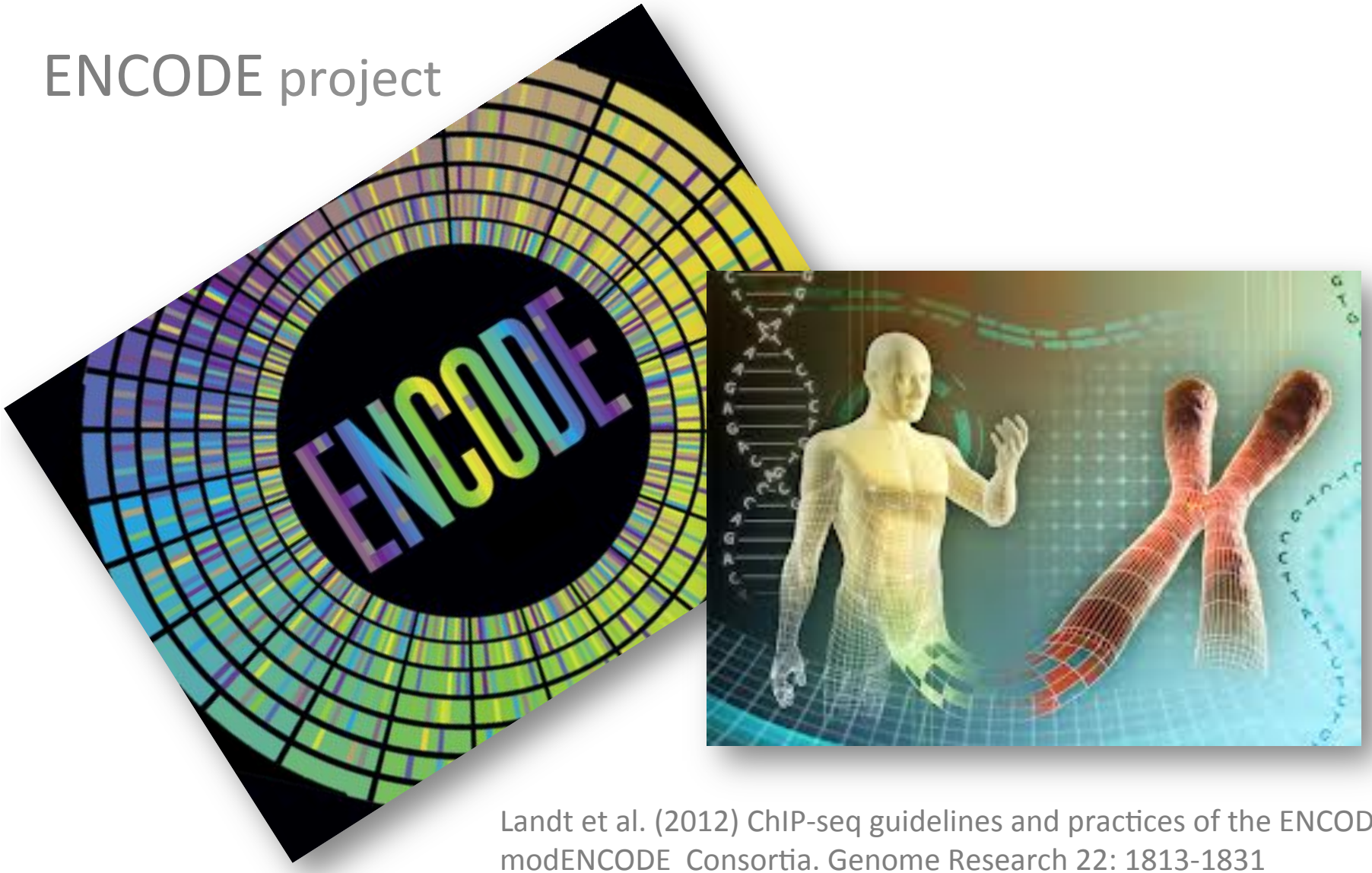
ChIP-seq analysis workflow



UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

ENCODE project



Landt et al. (2012) ChIP-seq guidelines and practices of the ENCODE and modENCODE Consortia. Genome Research 22: 1813-1831

guidelines!

Chen et al. (2012) Systematic evaluation of factors influencing ChIP-seq fidelity. Nat Methods 9: 609

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Experimental design

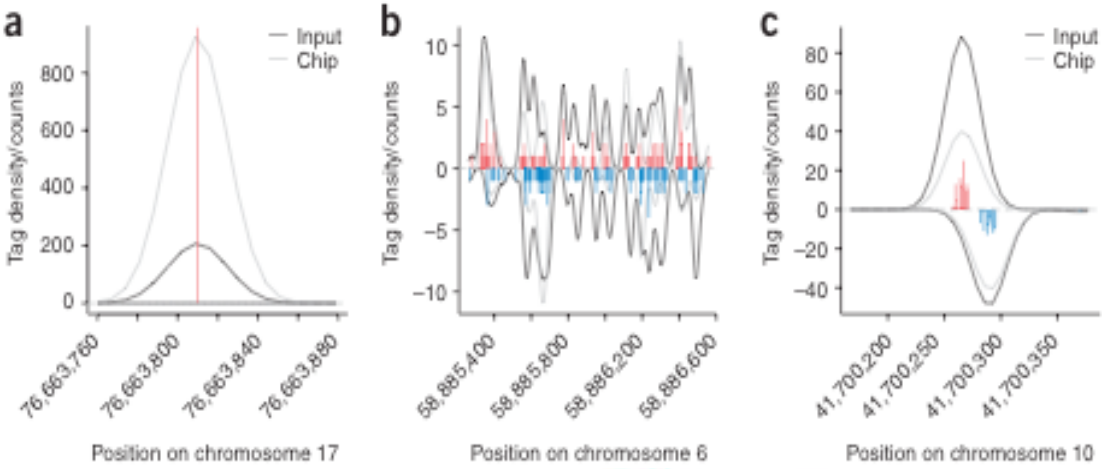
UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Experimental design: Controls

- Why use a control track?
  - Enrichment relative to “background”
  - Tissue anomalies (CNV)
  - Open chromatin
  - Experimental, technical, computational biases
  - Background distribution irregular; difficult to model accurately (not Poisson)

- Types of Controls
  - Input
  - Vehicle
  - Non-specific antibody (IgG)
- Other control issues
  - Depth
  - Freshness



UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Experimental design: Replicates

- Technical
  - Multiple lanes on a flowcell
  - Different flowcells/instruments (e.g. GA vs. HiSeq)
- Biological
  - Patient samples
  - Model organisms
- "Experimental"
  - Repeat experiment using same procedures
    - Same cell population (passages)
    - Re-grow/acquire cell population
  - Related experiments (e.g. different antibody)

- How many replicates?
  - ENCODE uses exactly 2, but only does mapping
  - Differential analysis requires 3+
  - Don't yet have statistical techniques to determine how much **power** is gained from adding replicates
  - New control for each replicate?

Experimental design: Sequencing parameters

- Single vs Paired end
  - SE generally used
- Read length
  - Long enough to map effectively – 50bp
- Read depth
  - 20-30M reads*
- Batches and randomization
- Multiplexing
  - One pool is optimal (up to 96 samples)

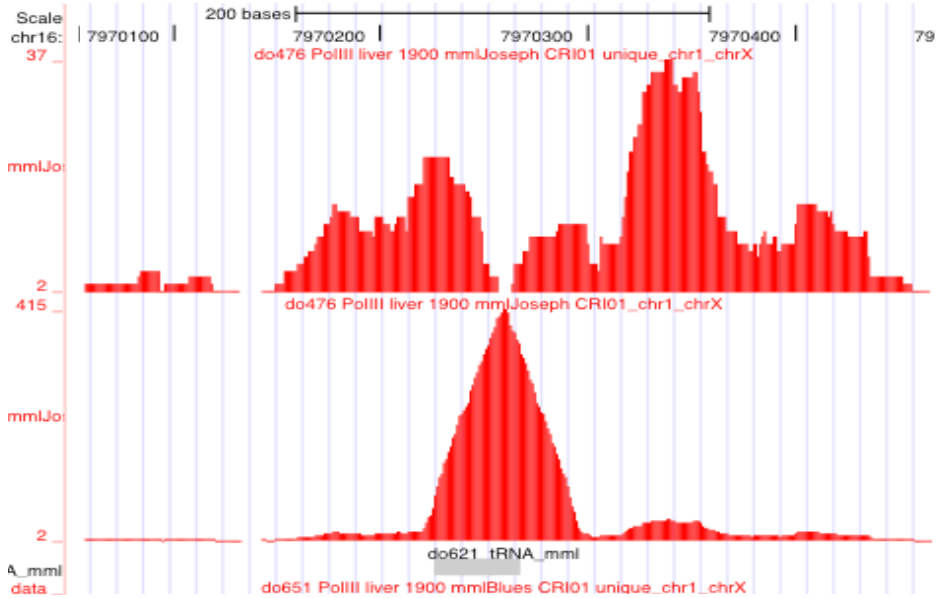
Read processing

Read filtering – Multimapped reads

ALIGNMENT QUALITY (CONFIDENCE OF CORRECT ALIGNMENT)

NON-UNIQUE ALIGNMENTS/"MULTIREADS" (REPEATS ETC.)

- One best position, but other positions within e.g. one base
- Multiple equally probable positions
- Most researchers discard non-unique alignments
- Random assignment
- Modelling alignment placement



UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Read filtering: Duplicates

TARGET IN CONTROL: <5%

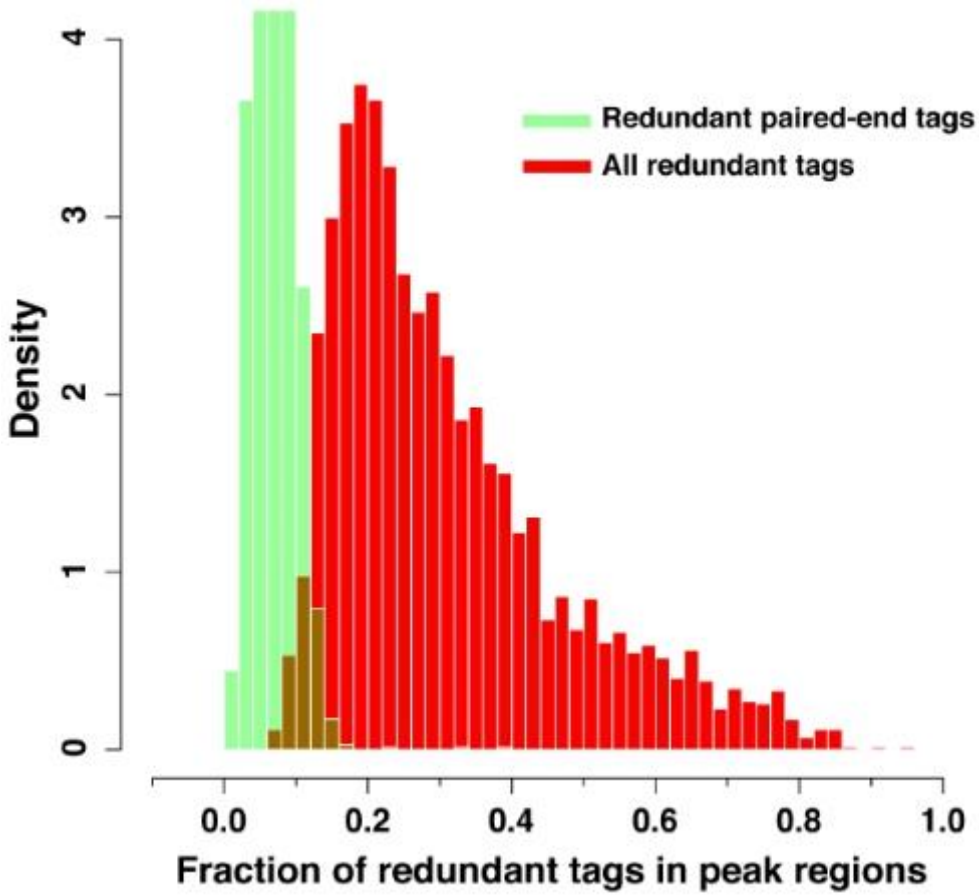
AMPLIFICATION ARTEFACTS

DUP RATE INCREASES WITH DEPTH

CALCULATING EXPECTED DUPLICATION RATES

DYNAMIC RANGE LIMITS

REQUIRED FOR DB ANALYSIS?

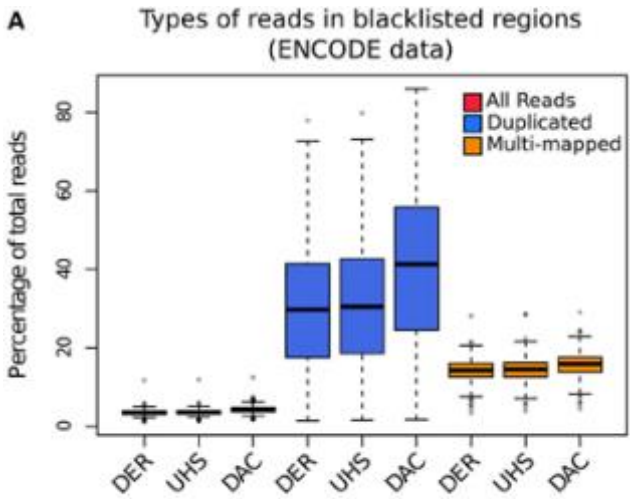


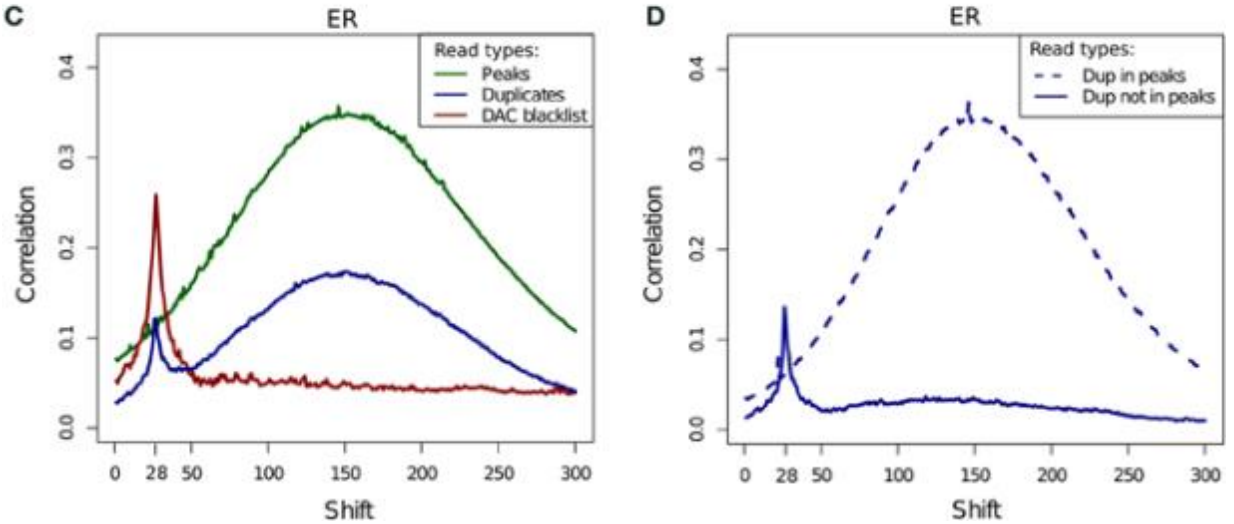
UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Read filtering: Blacklists

- Blacklisting recommended by ENCODE
- Duplicate and multi-mapped reads enriched in blacklisted regions
- Reads in blacklists bias strand shift calculations





"Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data", Carroll, Liang, Salama, Stark, and de Santiago, Frontiers in Genetics, 2014

UNIVERSITY OF CAMBRIDGE

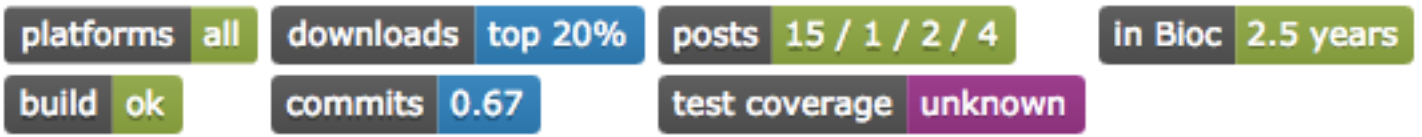
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Quality Assessment: Reads

Home Install

Home » Bioconductor 3.3 » Software Packages » ChIPQC

ChIPQC



Quality metrics for ChIPseq data

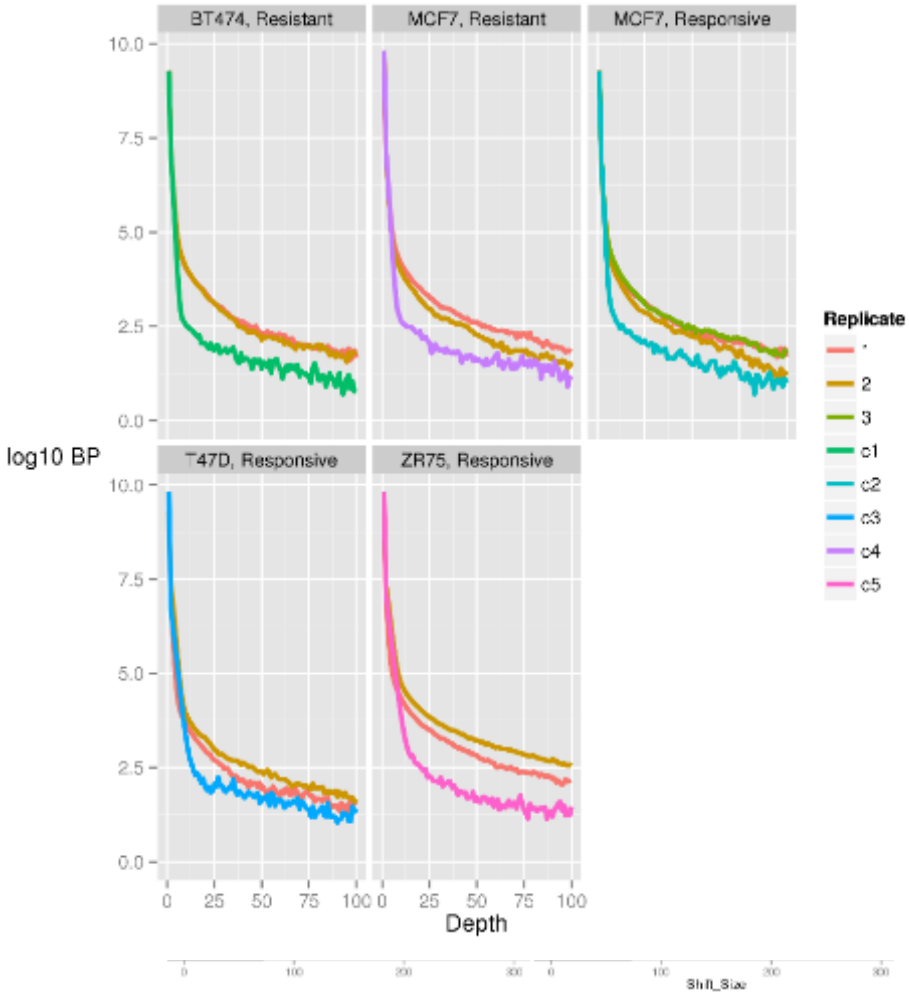
UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Coverage histogram

- Background reads (input) drop off quickly
- Enriched libraries (ChIP) drop off more slowly
- Gap between input and ChIP shows enrichment
- Marks have coverage “signature”
- Normalized standard deviation of distribution:

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

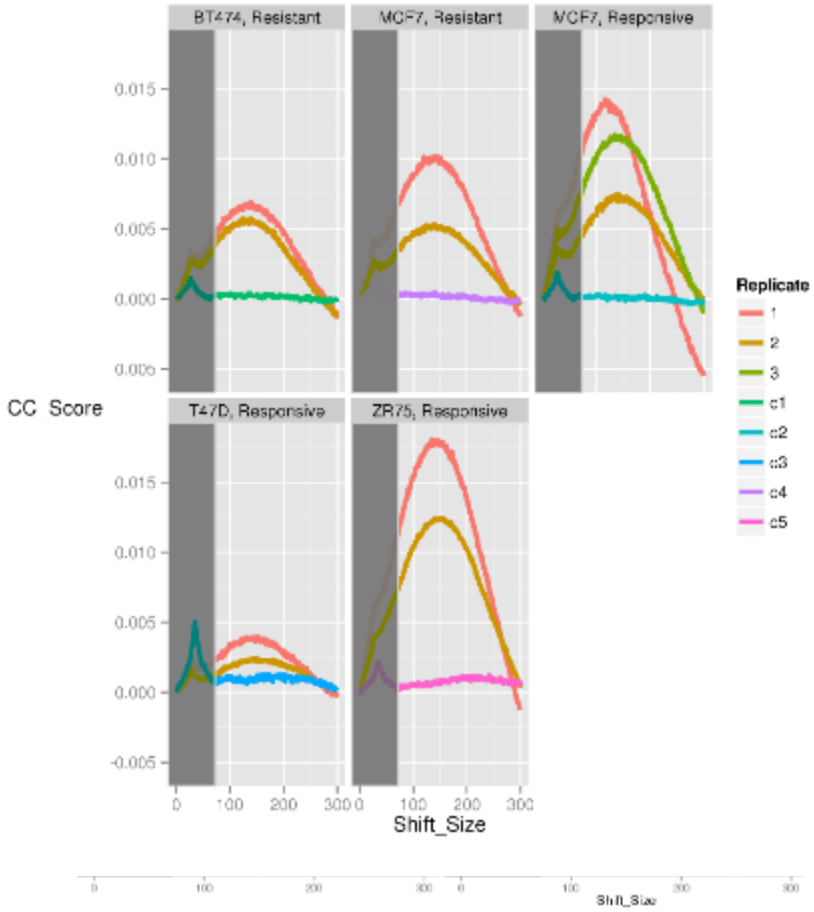


UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Fragment length estimation

- Multiple methods to estimate fragment length.
- **Cross-correlations -** Correlation of reads on positive and negative strand after successive read shifts.
- **Cross-coverage -** Coverage of reads on both strand after successive shifts of reads on one strand
- **Normalized score** - Length at max / read length

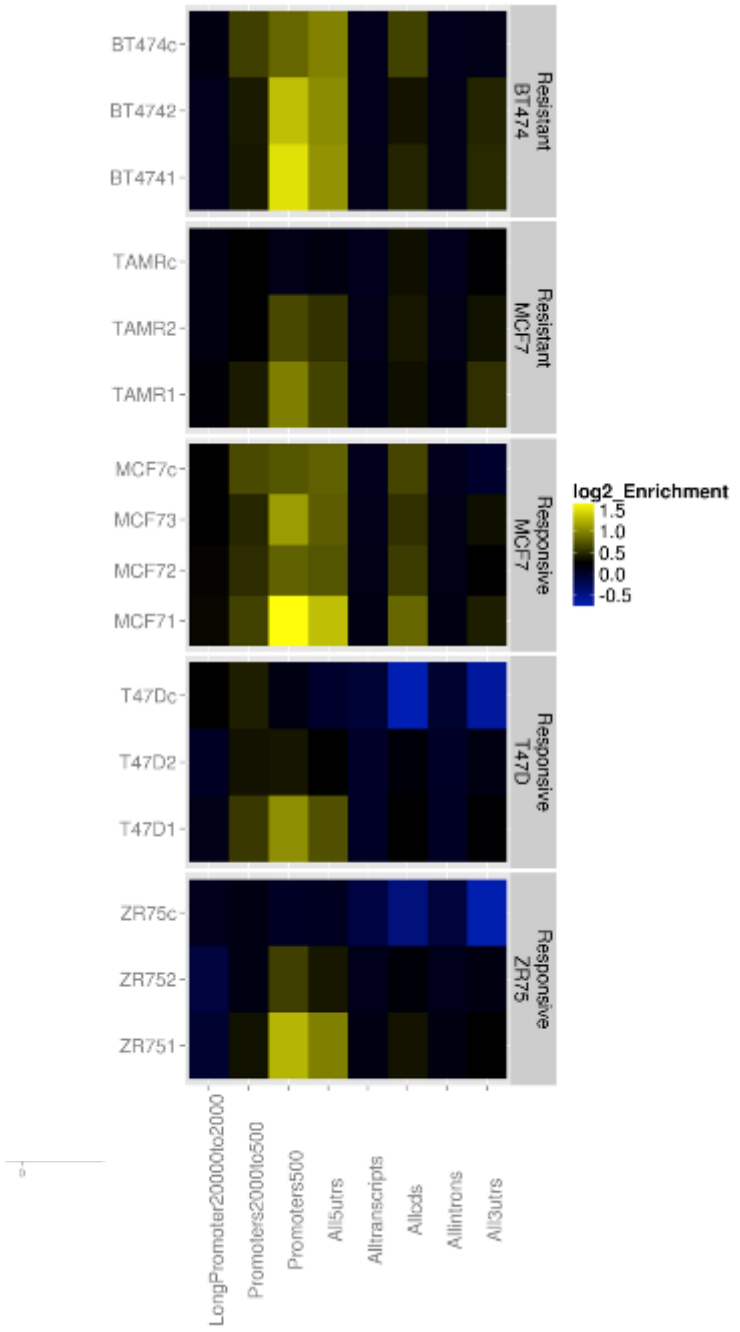


UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Genome profiling: Reads in features

- Enrichment of reads in specific genomic features
- Eg, reads in promoters vs. introns
- Relative to background (expected)
- Can provide custom annotations (eg enhancers)



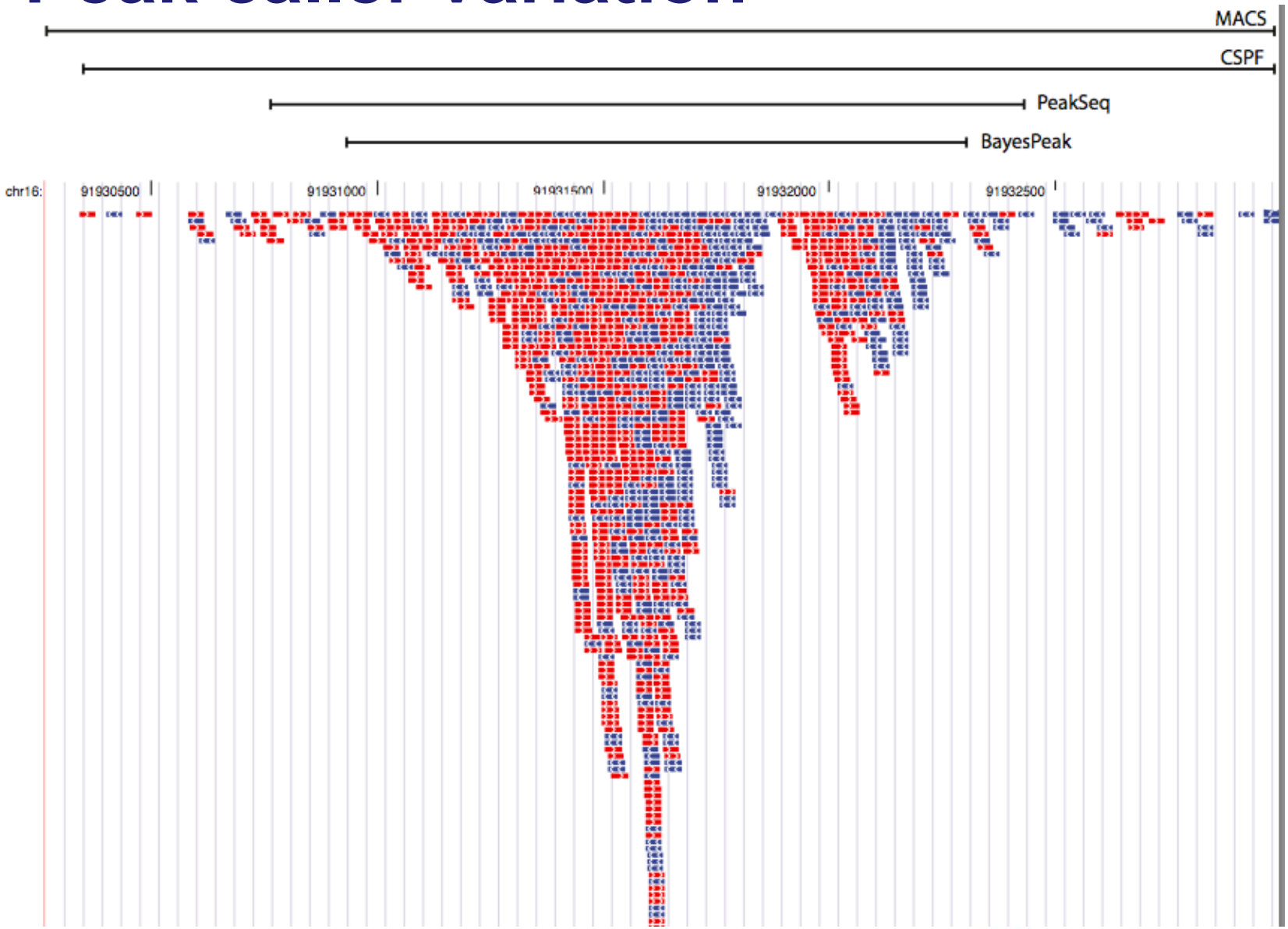
UNIVERSITY OF CAMBRIDGE

Peak Calling

Peak calling packages

	IP only, & control, either	read features	data from different strands	masks genomic repeats	scoring criteria	confidence in results, FDR estimates sensitivity / specificity	for both TF & HM
CSPF	& / or	read length no orientation	merges strands	N	simple height criteria	empirically: ROC curve	both
XSET	& / or	mean fragment length orientation	merges strands	N	simple height criteria	FDR based on randomised sample and Poisson probabilities	both
Mikkelsen et al.	IP only	no orientation	no merge / shift	Y	p-values produced by randomising the datasets	no official FDR	both
MACS	& / or	mean fragment length orientation ignores duplicated reads	shifts reads merges strands	N	Poisson p-values	FDR = no. peaks in control : IP	both
QuEST	&	orientation	shifts reads merges strands	N	kernel density estimation	FDR based on calling peaks in 1/2 the control sample	TF
FindPeaks	IP only	mean fragment length orientation	no merge / shift	N	simple height criteria	Monte-Carlo based FDR (ie. from randomised sample)	both
SISSR	& / or	mean fragment length orientation	no merge / shift	N	compares read density on different strands	FDR comparing simulated background peaks to real data	better for TF
Kharchenko et al.	&	orientation	no merge / shift	N	Poisson probabilities	FDR based on different randomised versions of the input sample	better for TF
PeakSeq	&	mean fragment length orientation	merges strands	Y	pre-processing: normalisation Binomial p-values	FDR: q-values after multiple correction adjustment	both
BayesPeak	& / or	mean fragment length orientation	no merge / shift	N	Negative Binomial distribution Bayesian posterior probabilities	posterior probabilities of enrichment presence	both

Peak caller variation



UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

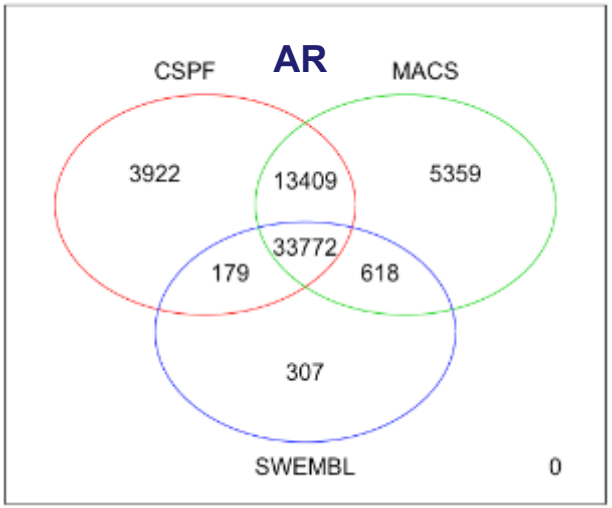
Agreement amongst peak callers

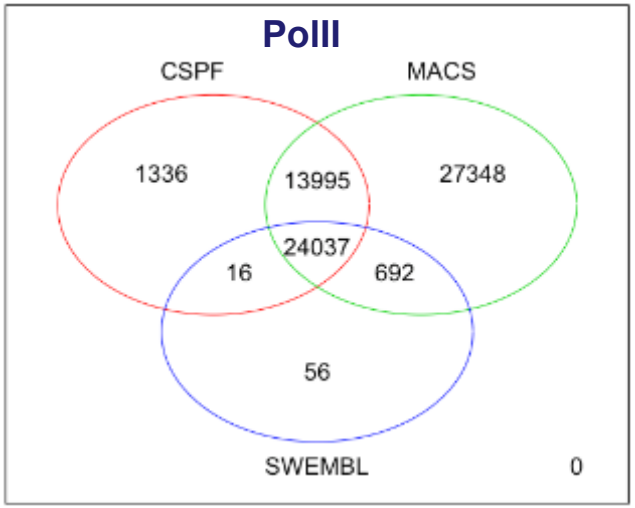
AGREEMENT ON A CORE SET OF PEAKS

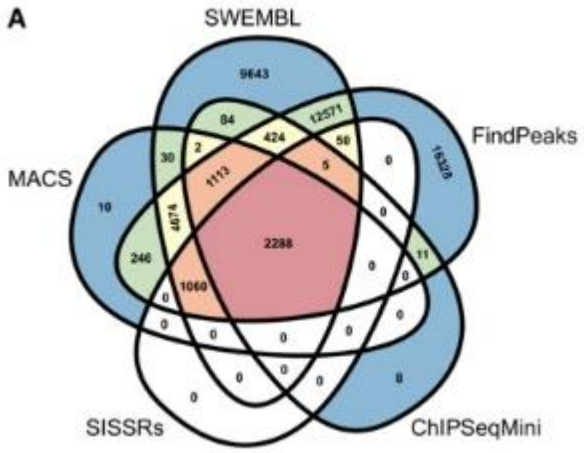
"PERMISSIVE" VS. "STRINGENT" PEAK CALLING

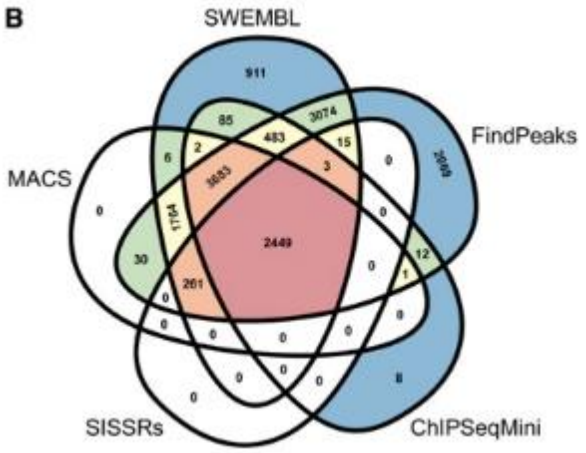
THERE IS NO ONE TRUE PEAK CALLER (TO RULE THEM ALL).

CAN WE AVOID PEAK CALLERS ALTOGETHER?









UNIVERSITY OF CAMBRIDGE

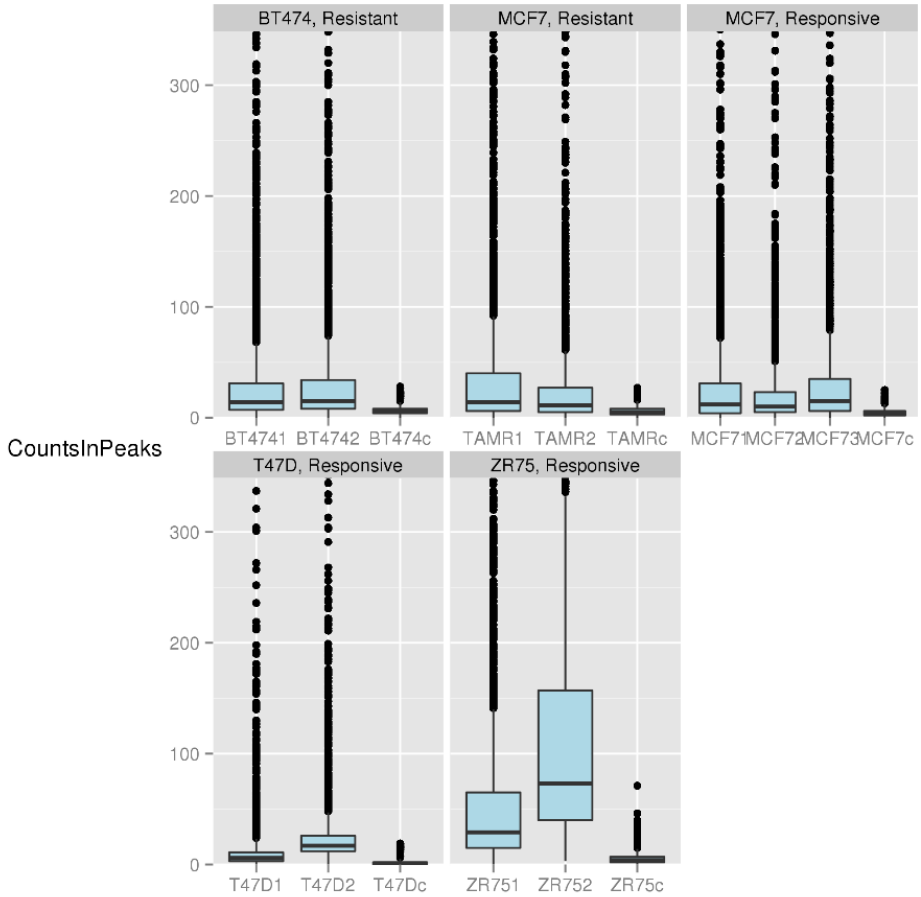
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Quality Assessment: Peaks

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Peak-based metrics I: Reads in Peaks

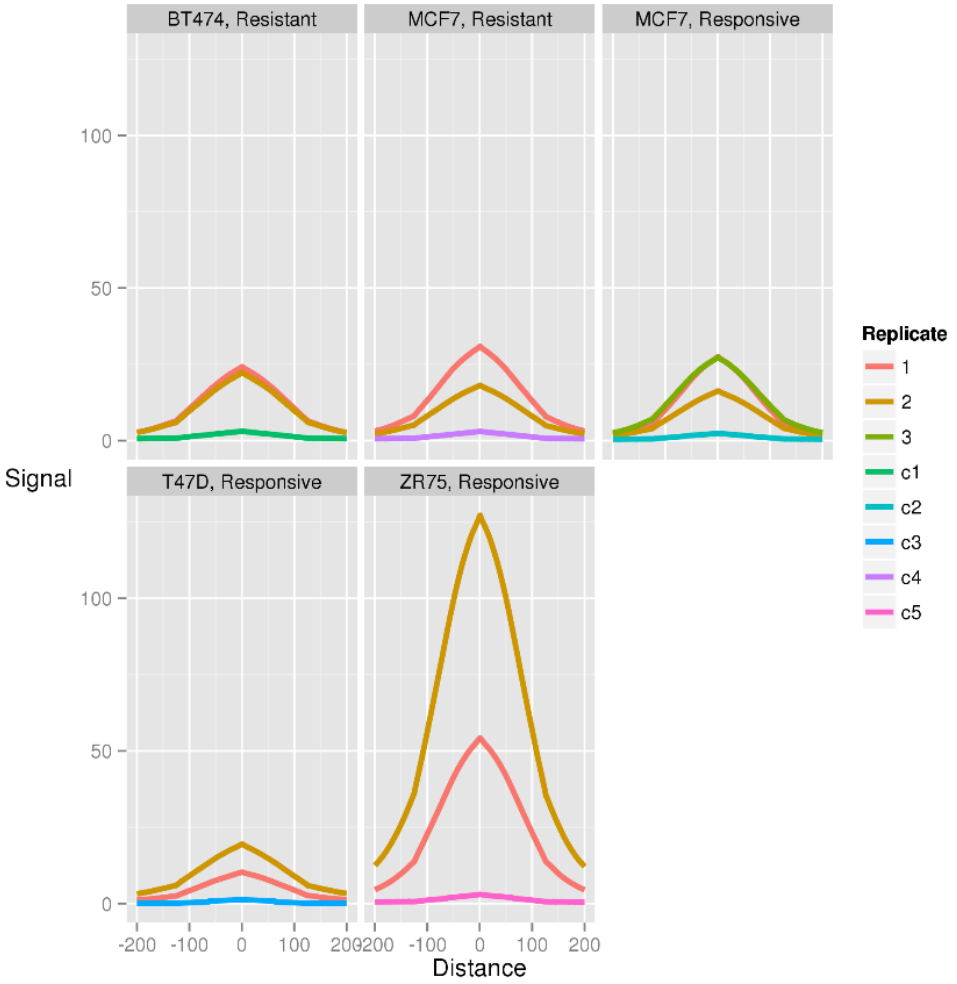


- Overall proportion of reads that overlap called peaks
- Distribution of read density across peaks

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Peak-based metrics II: Peak profiles

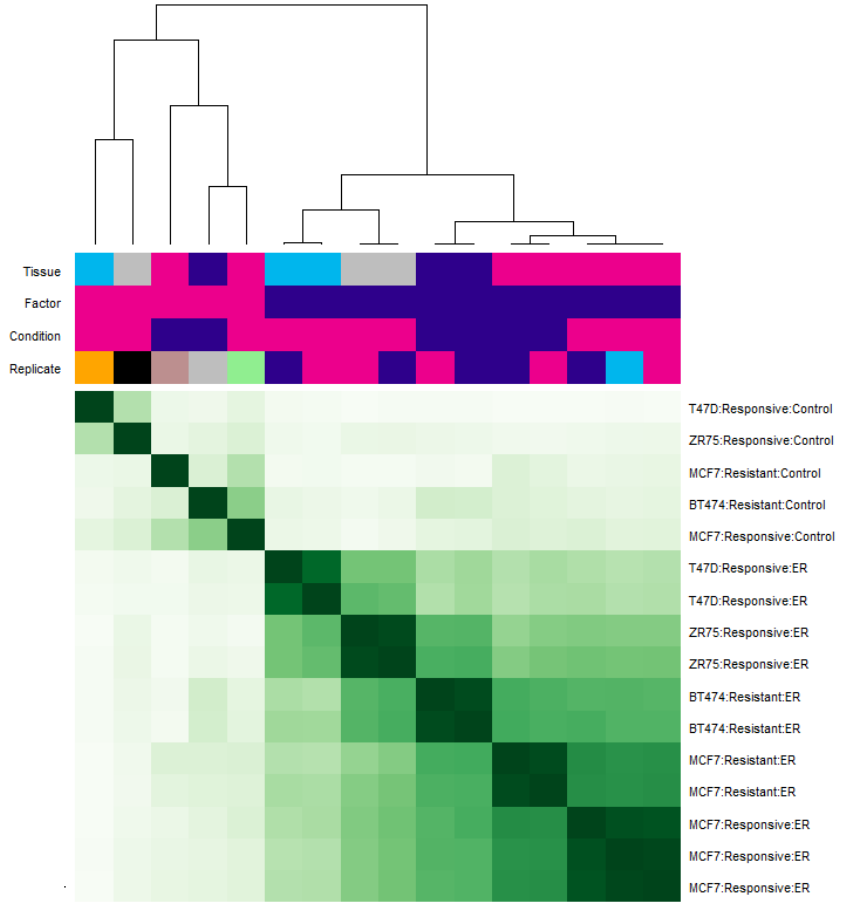


- Each peak centered at summit
- Mean density of reads at each position relative to summit
- Input controls should be flat

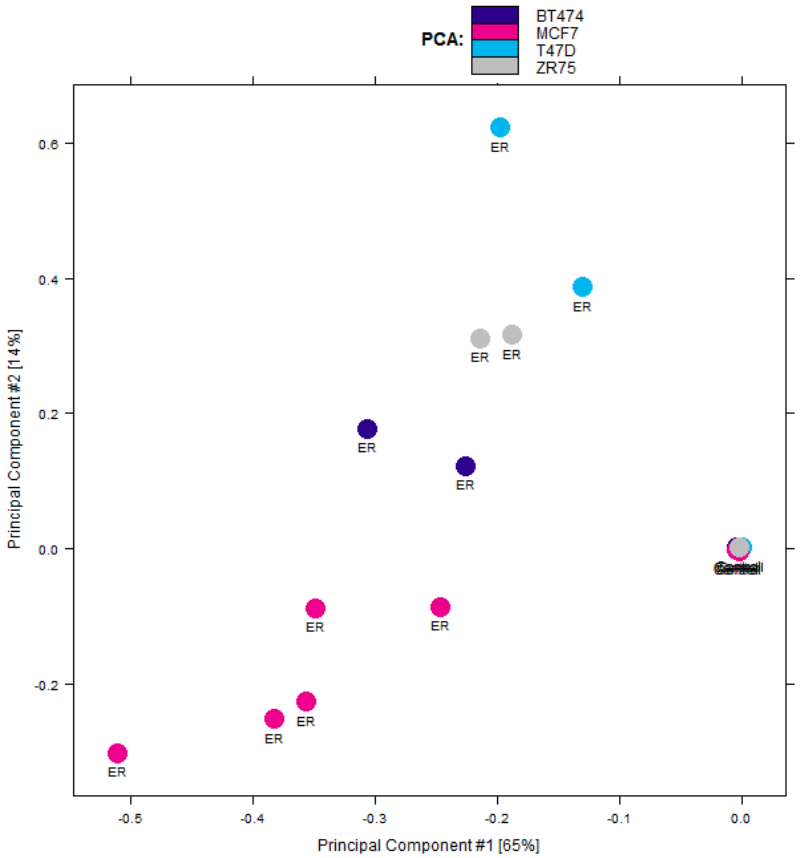
UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Peak-based metrics III: Clustering and PCA



- Clustered correlation heatmap



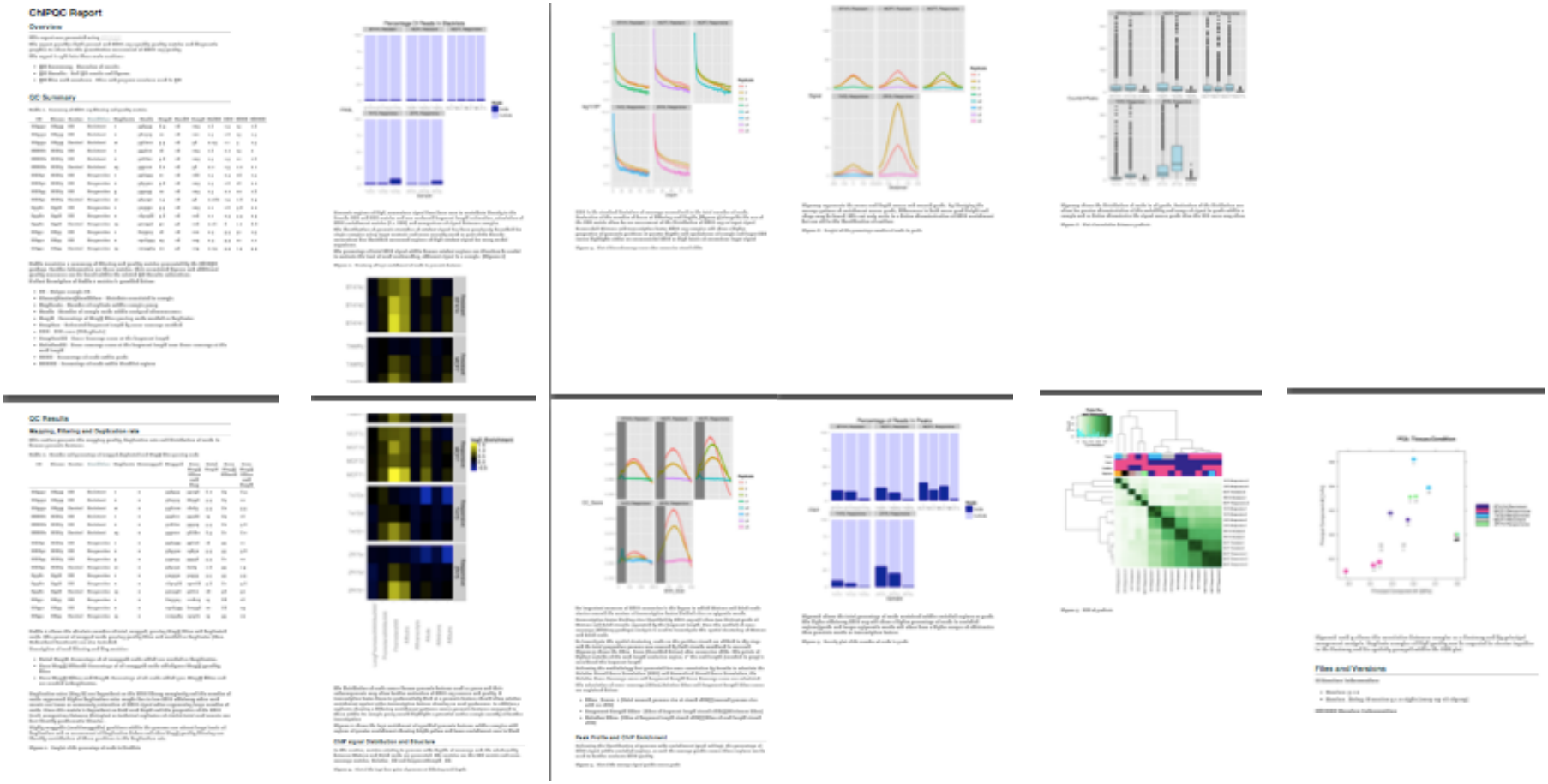
- Principal component analysis

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

ChIPQC report for example dataset

http://starkhome.com/ChIPQC/Reports/tamoxifen/ChIPQC.html



UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Acknowledgements

- **CRUK-CI Bioinformatics Core**
  - Matthew Eldridge
  - Suraj Menon
  - Thomas Carroll (**ChIPQC**, now MRC Clinical Sciences Centre)
- **Jason Carroll lab**
  - Caryn Ross-Innes
  - Vasiliki Therodorou
  - **Gordon Brown (DiffBind, GreyList)**

Backup Slides

Analysis of ChIP-seq data

EXPERIMENTAL DESIGN

- Replicates
- Controls
- Sequencing parameters

READ PROCESSING

- Alignment
- Filtering
  - Multimapped reads
  - Duplicates
  - Blacklists
- Quality assessment

PEAK CALLING

- Peak callers
- Alternatives
- Quality assessment

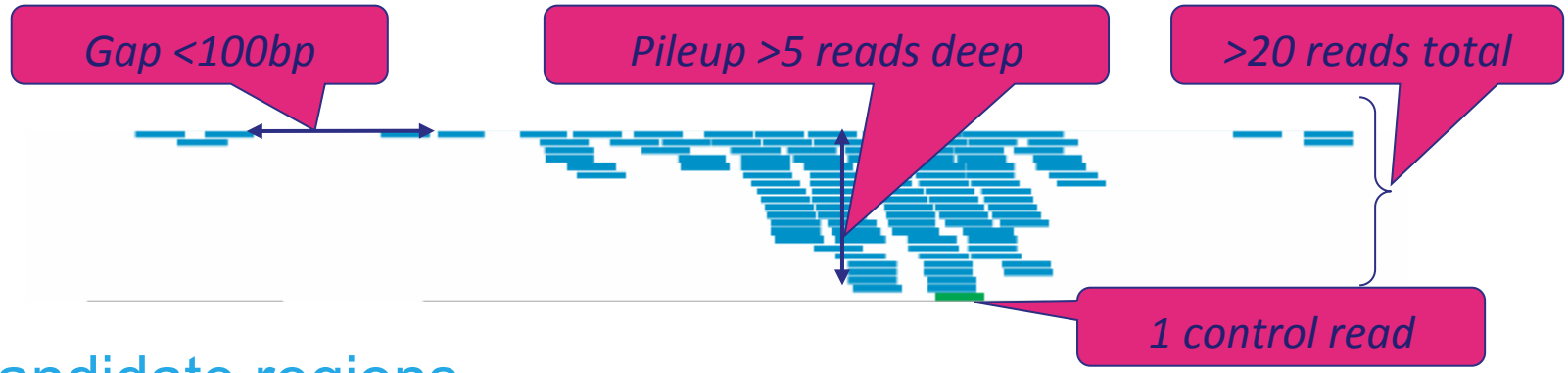
DIFFERENTIAL BINDING ANALYSIS

- Occupancy-based analysis
- Affinity-based analysis

VALIDATION AND DOWNSTREAM ANALYSIS

- Annotation
- Motif analysis

Peak calling



Find candidate regions

- Maximum distance *d* between reads
- Minimum of *n* sequence reads
- Minimum *p* peak pileup height

Determine enrichment:

$$E = \frac{n_{treatment}}{n_{input}}$$

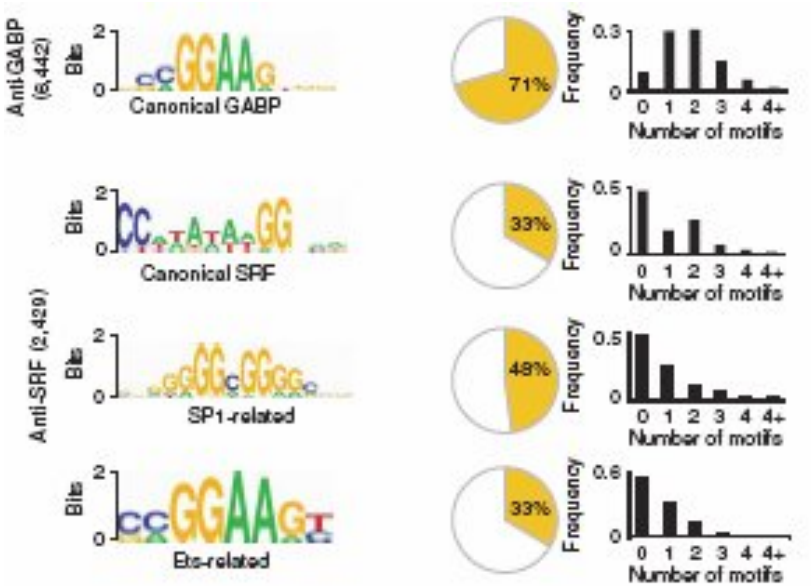
- If *E* > threshold *t*, it is an “enriched region”

Peak calling confidence statistics

RANKING BY ENRICHMENT VS. STANDARD STATISTICAL MEASURES

- p-values/FDR
- Poor FDR agreement!
- Peak callers are highly parametric, enabling fiddling to get what you want

- Validation: what is a false positive/negative?
  - Biological knowledge (e.g. literature)
  - Presence of TFBS motif
  - Proximity to genomic features (e.g. genes, promoters)
  - Experimental validation (qPCR)
  - Agreement with other marks (Pol II, open chromatin, histone marks, co-factors)

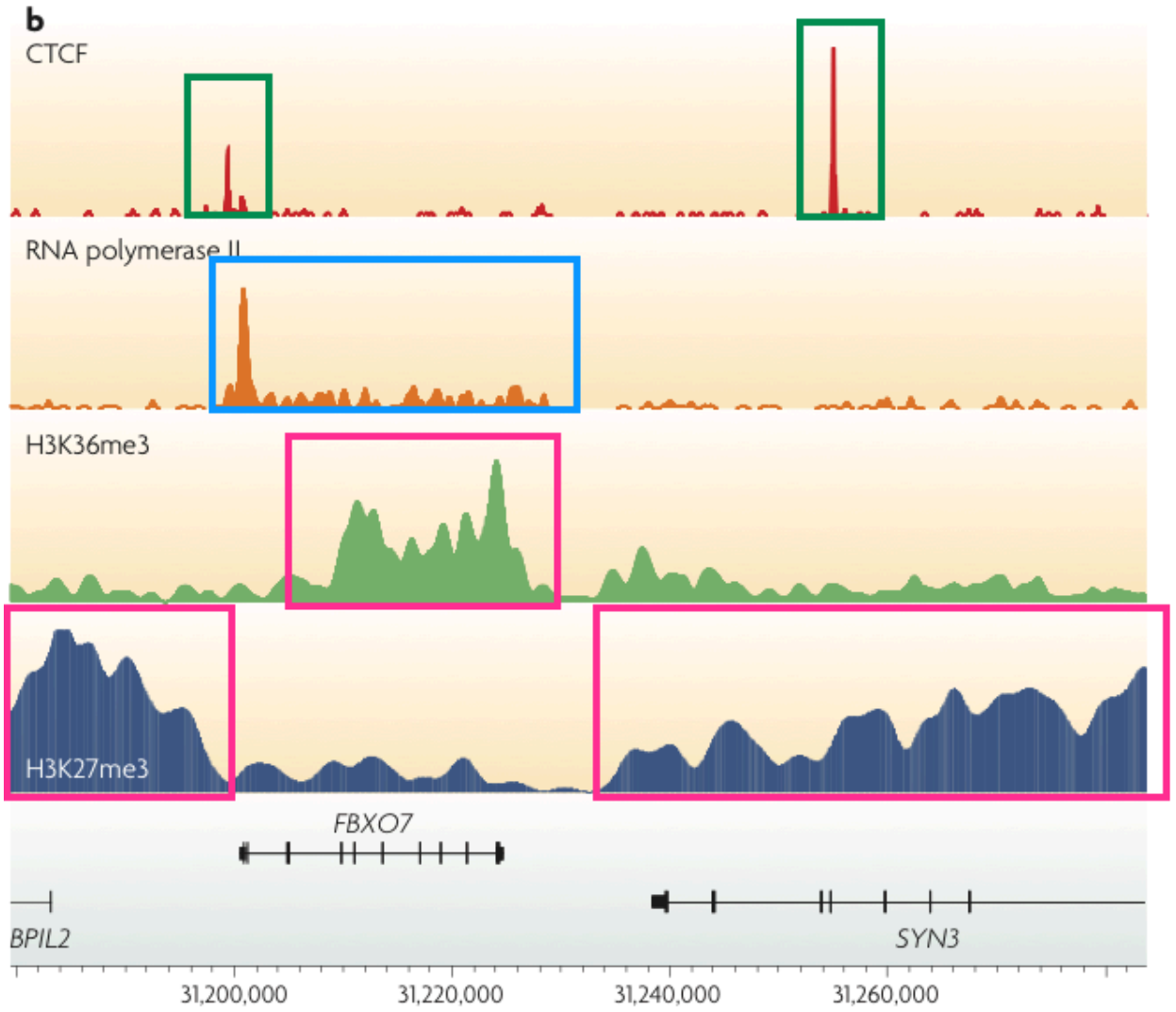


Valouev *et. al.* Nature Methods 2008

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Wide enriched regions (histone marks)



sequence-specific TFs

RNA Pol II

histone modifications

ChIP-Seq data
for fly S2 cells

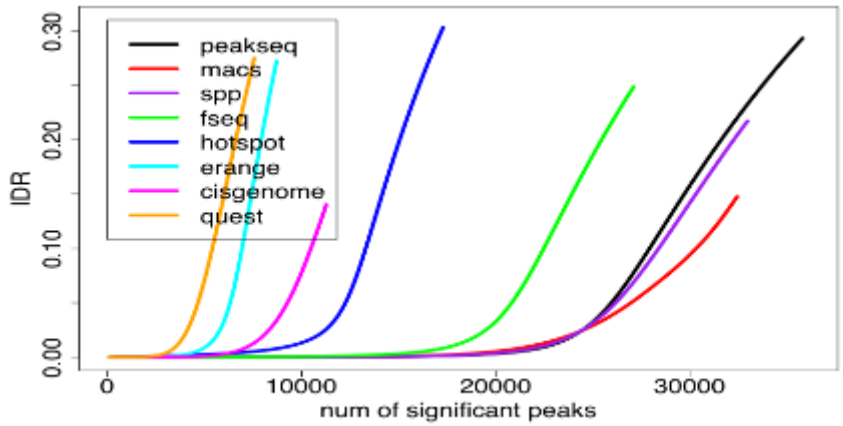
UNIVERSITY OF CAMBRIDGE

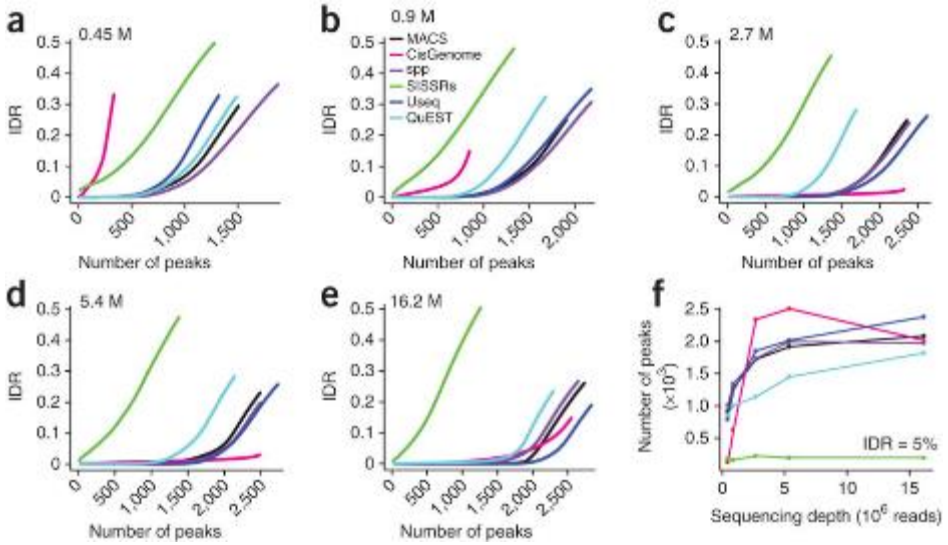
CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Consensus peaks: Irreproducible Discovery Rate (IDR)

Compare two sets of peak calls

- Two replicates
- Two peak callers (technical test)





Qunhua Li et al (2011) "*Measuring reproducibility of high-throughput experiments*"

UNIVERSITY OF CAMBRIDGE

CANCER RESEARCH UK | CAMBRIDGE INSTITUTE

Sample dataset

Sample	Tissue	Factor	Status	Rep#
MCF71	MCF7	ERα	Responsive	1
MCF72	MCF7	ERα	Responsive	2
MCF73	MCF7	ERα	Responsive	3
T47D1	T47D	ERα	Responsive	1
T47D1	T47D	ERα	Responsive	2
ZR751	ZR75	ERα	Responsive	1
ZR752	ZR75	ERα	Responsive	2
MCF7r1	MCF7	ERα	Resistant	1
MCF7r2	MCF7	ERα	Resistant	2
BT4741	BT474	ERα	Resistant	1
BT4742	BT474	ERα	Resistant	2