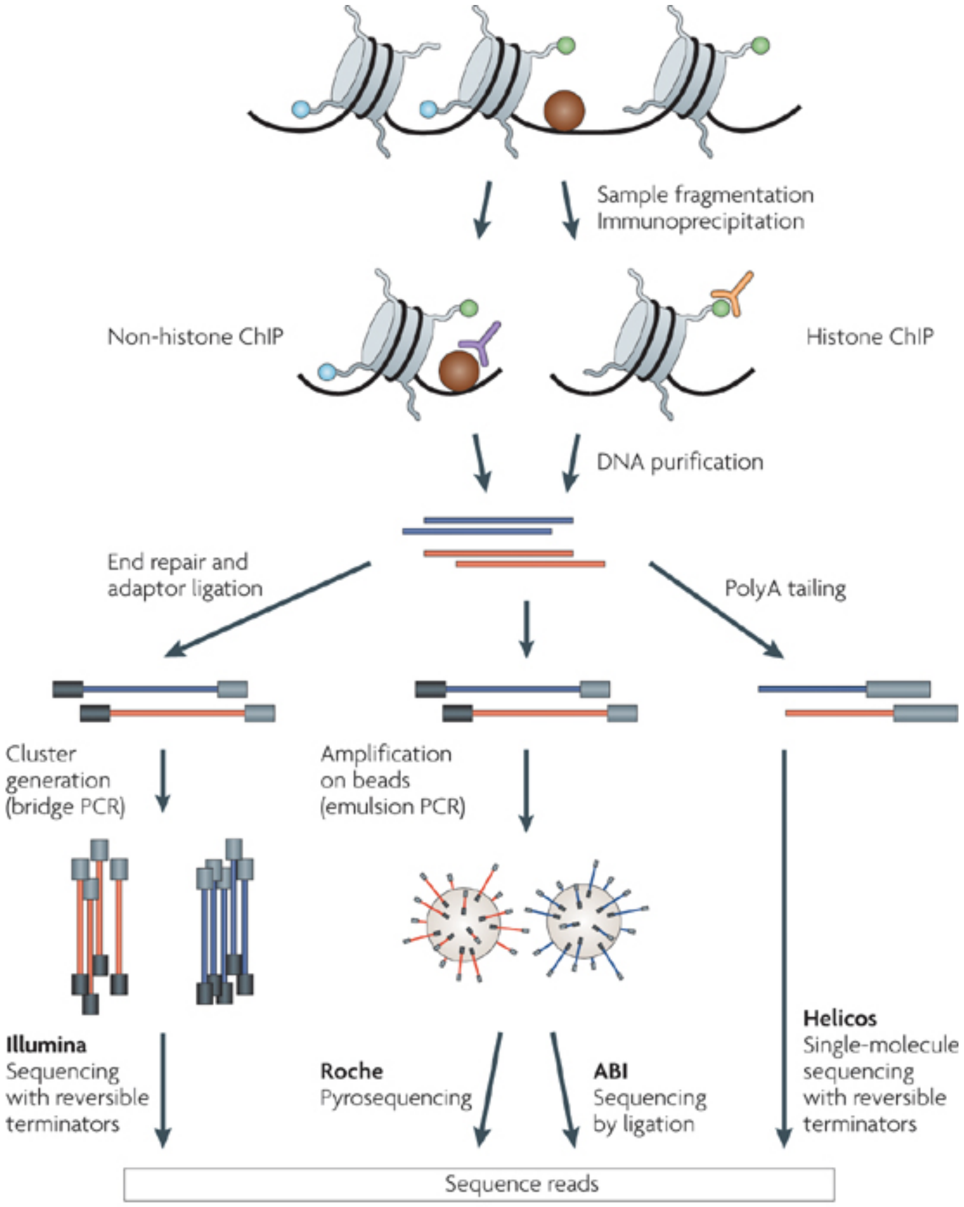
EMBL-EBI

Computational aspects of ChIP-seq

John Marioni
Research Group Leader
European Bioinformatics Institute
European Molecular Biology Laboratory

ChIP-seq

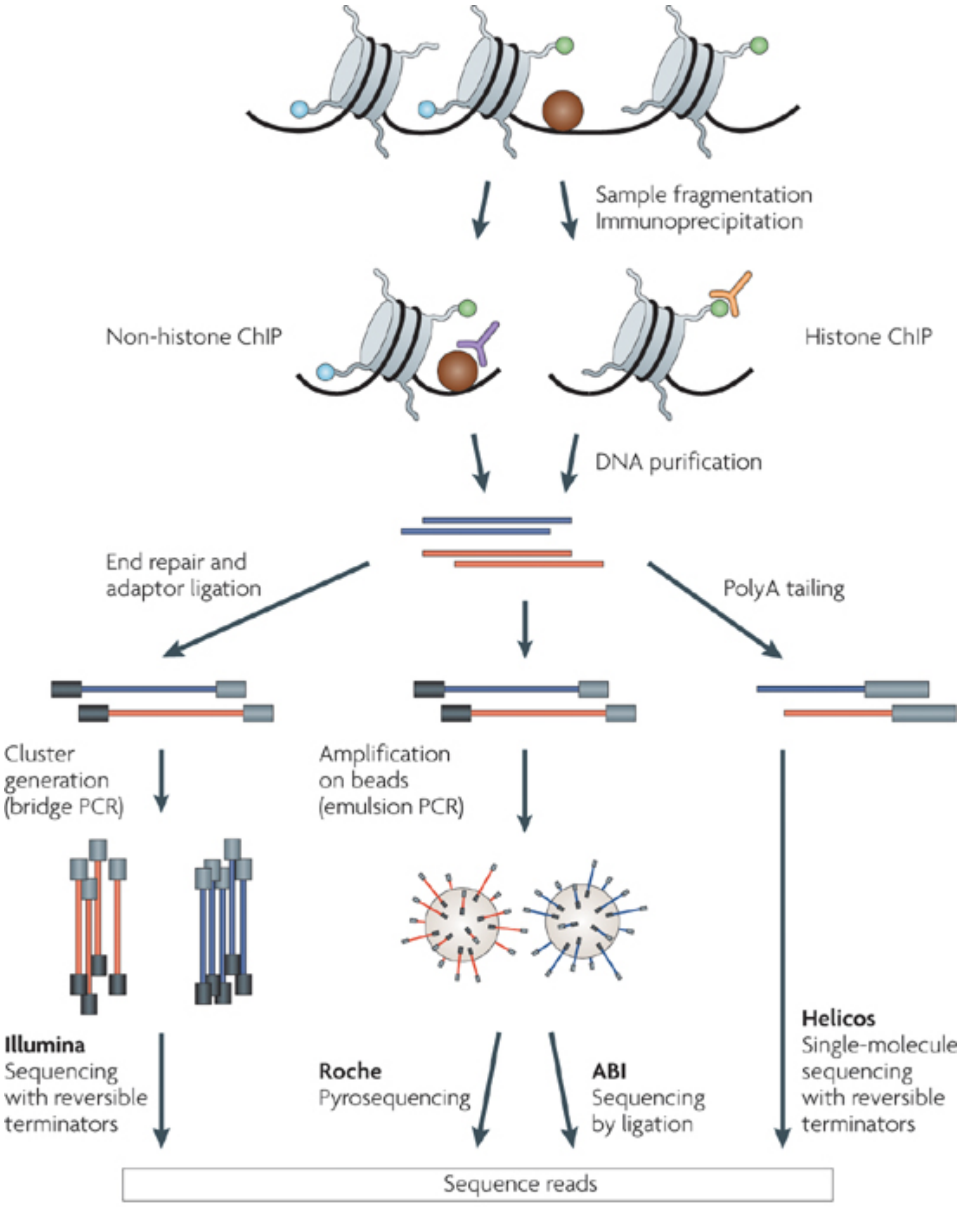
Using high-throughput sequencing to investigate DNA binding proteins or histone modifications


Sample fragmentation
Immunoprecipitation
Non-histone ChIP
Histone ChIP
DNA purification
End repair and adaptor ligation
PolyA tailing
Cluster generation (bridge PCR)
Amplification on beads (emulsion PCR)
Illumina
Sequencing with reversible terminators
Roche
Pyrosequencing
ABI
Sequencing by ligation
Helicos
Single-molecule sequencing with reversible terminators
Sequence reads

Park, Nat Rev Genet, 2009

ChIP-seq

- Other applications employ similar experimental approaches to interrogate DNaseI hypersensitivity sites and chromatin confirmation
- Will talk about these at the end of this presentation


Sample fragmentation
Immunoprecipitation
Non-histone ChIP
Histone ChIP
DNA purification
End repair and adaptor ligation
PolyA tailing
Cluster generation (bridge PCR)
Amplification on beads (emulsion PCR)
Illumina
Sequencing with reversible terminators
Roche
Pyrosequencing
ABI
Sequencing by ligation
Helicos
Single-molecule sequencing with reversible terminators
Sequence reads

ChIP-seq vs ChIP-chip

- Interrogate whole genome
- Base pair resolution
- Greater dynamic range
- Less starting material (10-50ng compared to > 2 micrograms)
- Cheaper!

ChIP-seq vs ChIP-chip

- Interrogate whole genome
- Base pair resolution
- Greater dynamic range
- Less starting material (10-50ng compared to > 2 micrograms)
- Cheaper!

One of the areas where NGS is very clearly a far superior technology to microarrays

Overview

1. Designing ChIP-seq experiments
2. Read mapping and quantifying binding
3. Applications of ChIP-seq
4. Other applications using similar techniques

Overview

1. Designing ChIP-seq experiments

2. Read mapping and quantifying binding

3. Applications of ChIP-seq

4. Other applications using similar techniques

Designing ChIP-seq experiments

Qn 1: How good is your antibody?

- ChIP-Seq data depend on antibody quality
- modENCODE project:
  - Large-scale screening for histone modifications in flies (Drosophila)
  - 20-35% of commercial 'ChIP-grade' antibodies were unusable
- Variations between antibodies
  - differences in antibody specificity can make it hard to compare data across multiple transcription factors

Celniker et al. 2009
Park 2009
Vaquerizas et al. 2008

Designing ChIP-seq experiments

Qn 2: Do you need controls?

- Controls can be generated by:
    - lysing and fragmenting (sonicating) cells but not IP-ing the sample
    - Lysing and fragmenting cells and performing a mock IP (an IP without an antibody)
    - Performing an IP with an antibody that is not know to be involved in DNA binding or chromatin modification (e.g., IGG)

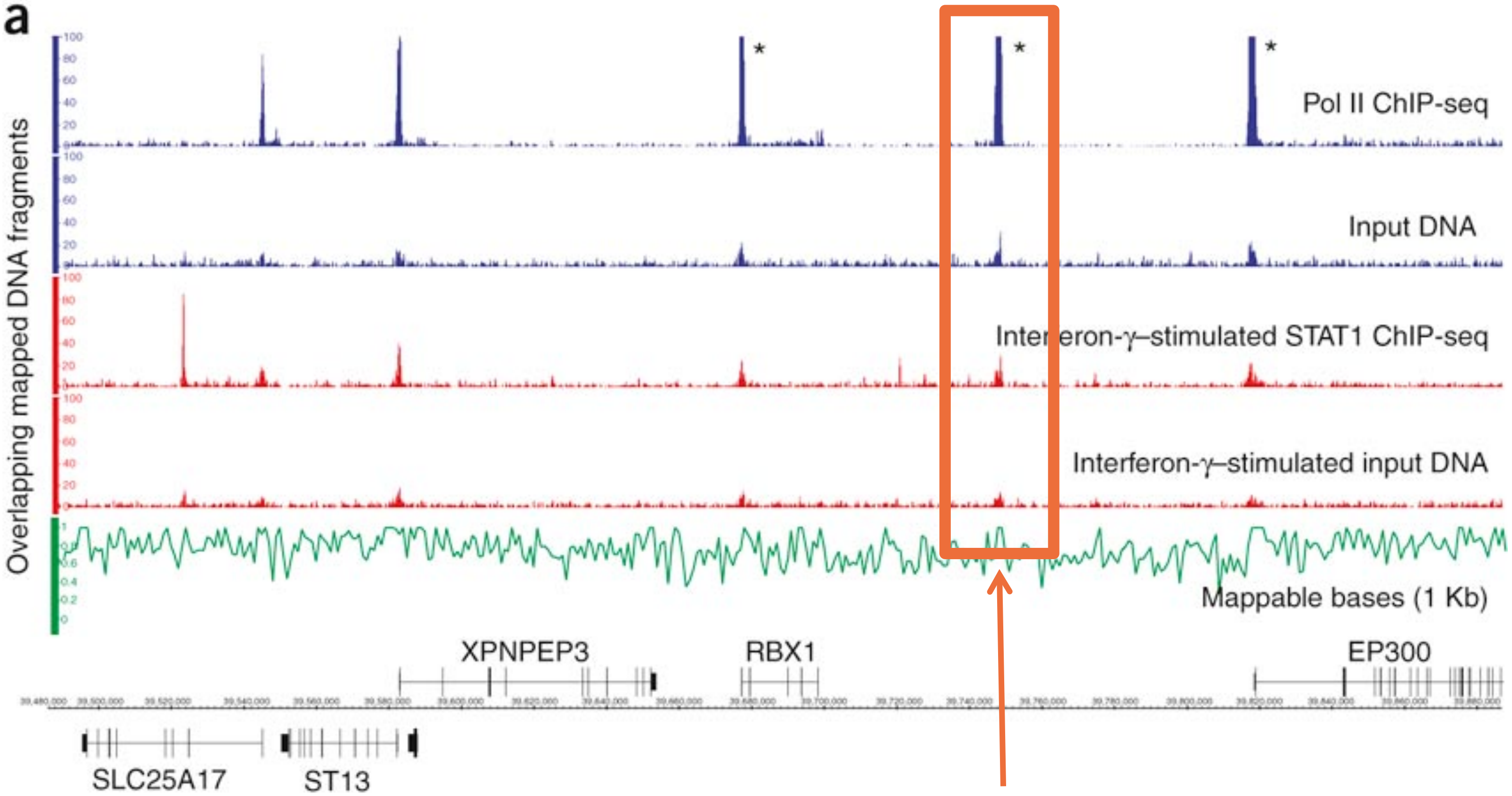
Designing ChIP-seq experiments

Qn 2: Do you need controls?

- For ChIP-Seq, lysing and fragmenting (sonicating) cells but not IP-ing the sample is the most popular way of generating a control sample
- In any event, the resulting cells can be processed into a library that is suitable for sequencing and used as a "control" or "input" sample

Designing ChIP-seq experiments

Qn 2: Do you need controls?



ChIP

Input

ChIP

Input

Peaks line up

Rozowsky et al. 2009

Designing ChIP-seq experiments

Qn 2: Do you need controls?

- Controls were skipped in early experiments:
  - Cost
  - Over-confidence in data quality
- But clearly they can control for artefacts:
  - Copy number variation
  - Incorrect mapping of repetitive genomic regions
  - Non-uniform fragmentation

If the genome of the sample being studied has been sequenced using similar technology, one can possibly use this as a control

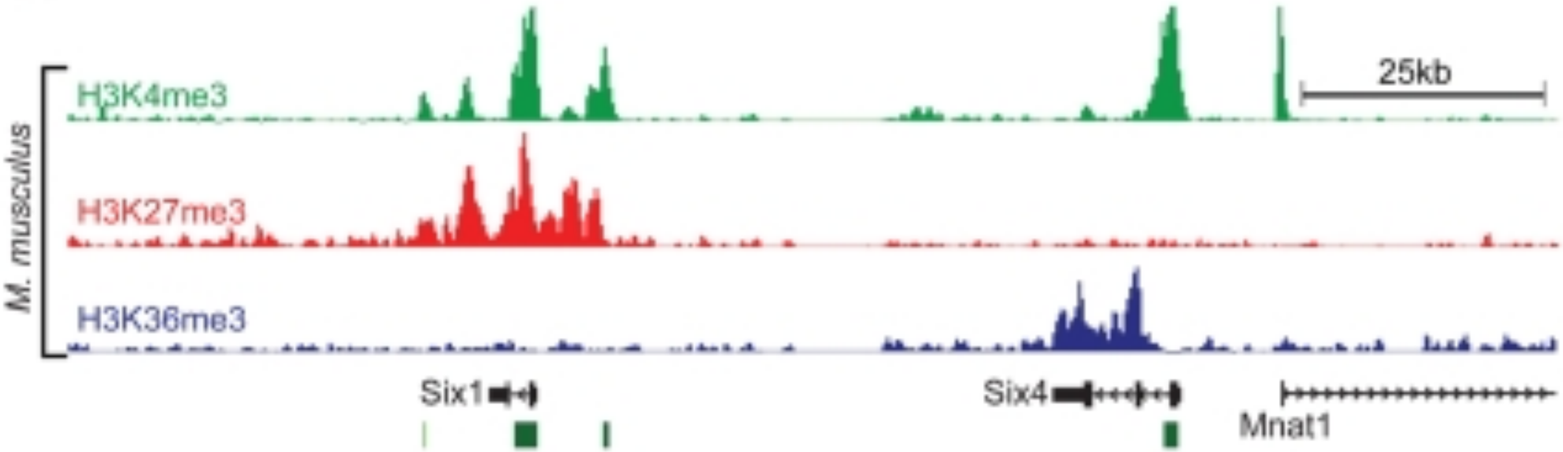
Designing ChIP-seq experiments

Qn 3: Sequencing depth

- Sequencing depth depends upon genome size, protein and the biological question
- In particular, different proteins bind to the genome in very different ways, which can effect one's ability to identify bound regions

Designing ChIP-seq experiments

Proteins bind in different ways



Activating mark (near TSS)

Peaks within body of inactive genes

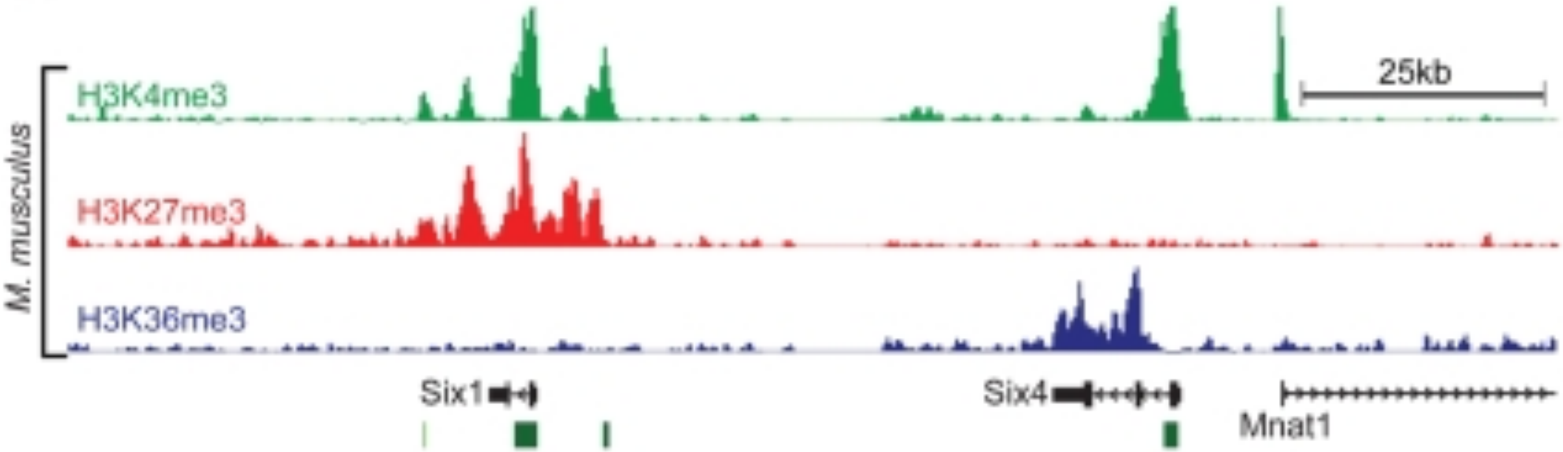
Peaks within body of active genes

Data from mouse ES cells

Ku et al. 2008

Designing ChIP-seq experiments

Proteins bind in different ways



Activating mark
(near TSS)

Peaks within body
of inactive genes

Peaks within body
of active genes

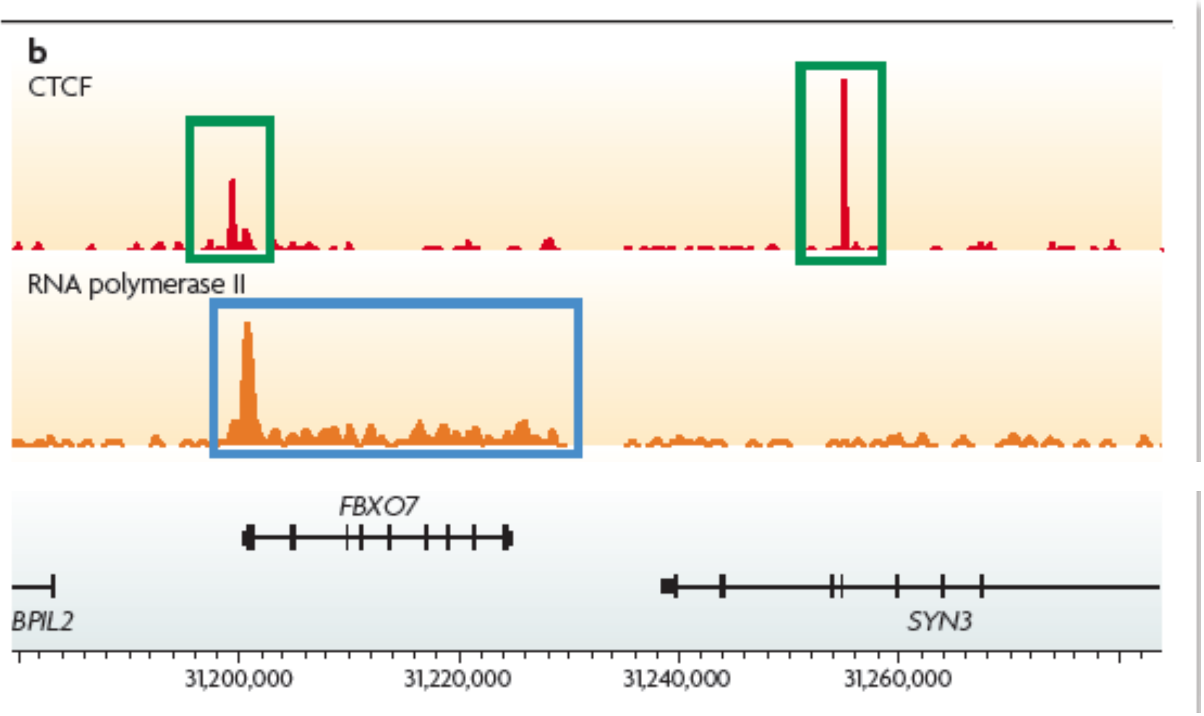
In general, the ChIP-ped regions associated with histone modifications tend to cover broad sections of the genome

Ku et al. 2008

Designing ChIP-seq experiments

Proteins bind in different ways

Transcription factor – tight, highly-peaked binding region



RNA PolII – enriched at TSS but bound throughout gene body

ChIP-Seq data from fly S2 cells

Designing ChIP-seq experiments

Proteins bind in different ways

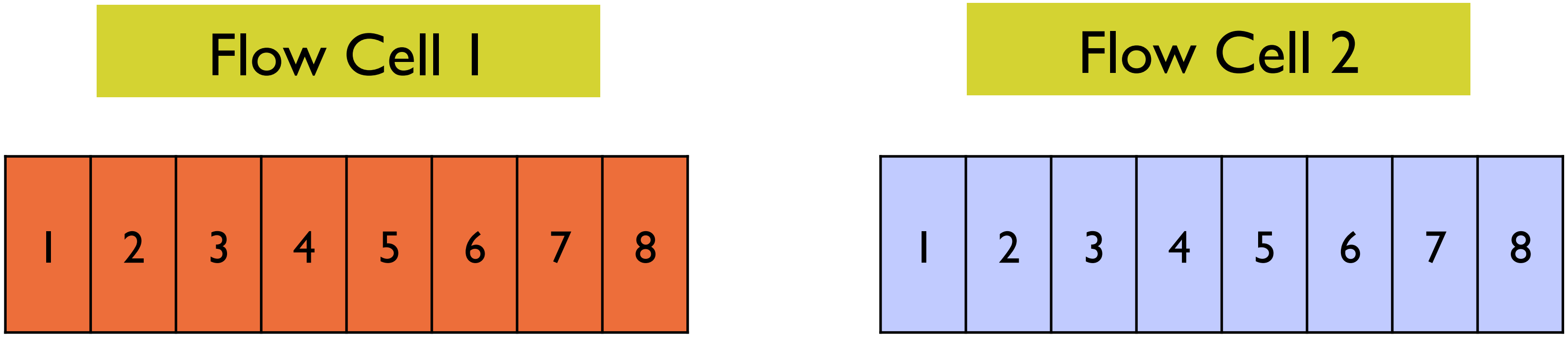
- The protein being investigated has major effects upon the binding patterns – this is important since most algorithms for calling peaks have been developed to find TF binding, where the peak is constrained and sharp
- Also, where peaks are sharper less sequencing will be required in order to accurately define its boundaries

Designing ChIP-seq experiments

Finding differentially bound regions between 2 groups

🟦 8 samples in control group

🟧 8 samples in treatment group



Flow Cell 1

1	2	3	4	5	6	7	8

Flow Cell 2

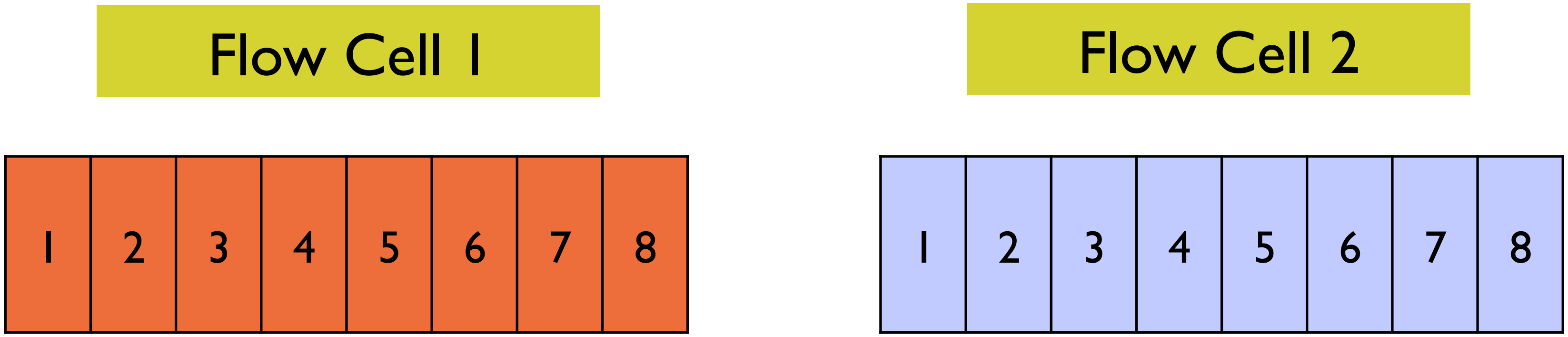
1	2	3	4	5	6	7	8

Designing ChIP-seq experiments

Finding differentially bound regions between 2 groups

🟦 8 samples in control group

🟧 8 samples in treatment group



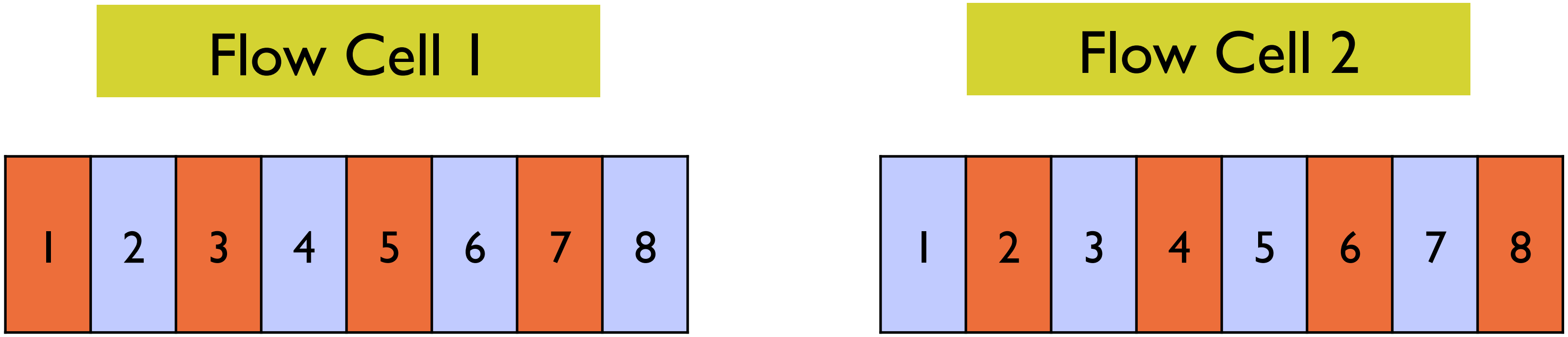
The treatment effect is completely confounded with the experimental design

Designing ChIP-seq experiments

Finding differentially bound regions between 2 groups

🟦 8 samples in control group

🟧 8 samples in treatment group



Flow Cell 1							
🟧 1	🟦 2	🟧 3	🟦 4	🟧 5	🟦 6	🟧 7	🟦 8

Flow Cell 2							
🟦 1	🟧 2	🟦 3	🟧 4	🟦 5	🟧 6	🟦 7	🟧 8

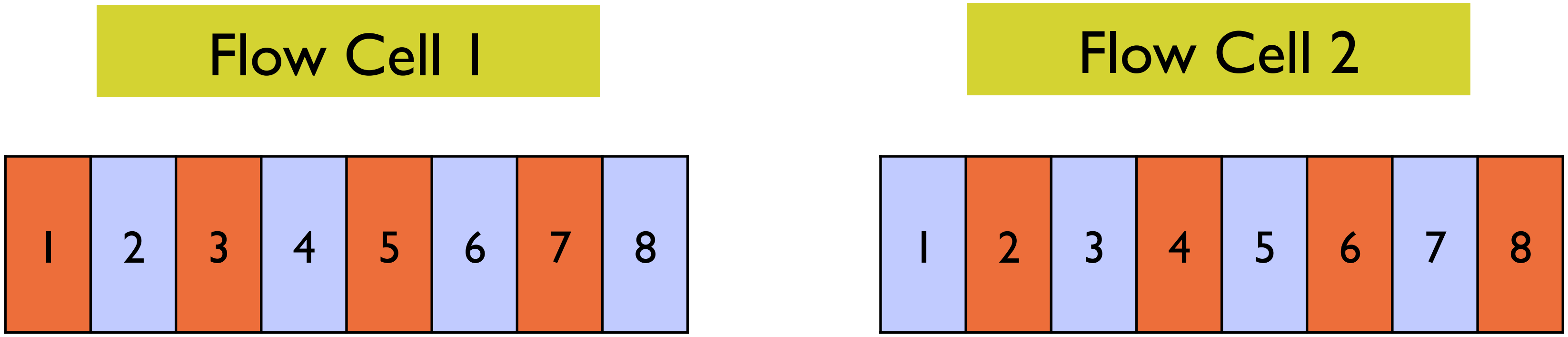
The samples have been randomized with respect to the flow cells – any flow cell effect can now be modelled

Designing ChIP-seq experiments

Finding differentially bound regions between 2 groups

🟦 8 samples in control group

🟧 8 samples in treatment group



The samples have been randomized with respect to the flow cells – any flow cell effect can now be modelled

Designing ChIP-seq experiments

Other considerations

- How many replicates?
  - The more the better!
- Do you need paired-end reads? How long should reads be?
  - Can help with mapping but not nearly as important as for identifying indels in DNA sequencing or multiple isoforms in RNA-seq

Overview

1. Designing ChIP-seq experiments
2. Read mapping and quantifying binding
3. Applications of ChIP-seq
4. Other applications using similar techniques

Read mapping and quantifying binding

- Choice of software depends upon
  - Accuracy, speed, memory, flexibility

In general alignment considerations are similar for ChIP-seq and genome sequencing – so the same considerations apply

Read mapping and quantifying binding

However, if you are interested in allele-specific binding care must be taken, since in some regions reads containing the non-reference allele might not be aligned well

Problems in mapping

reference allele reads

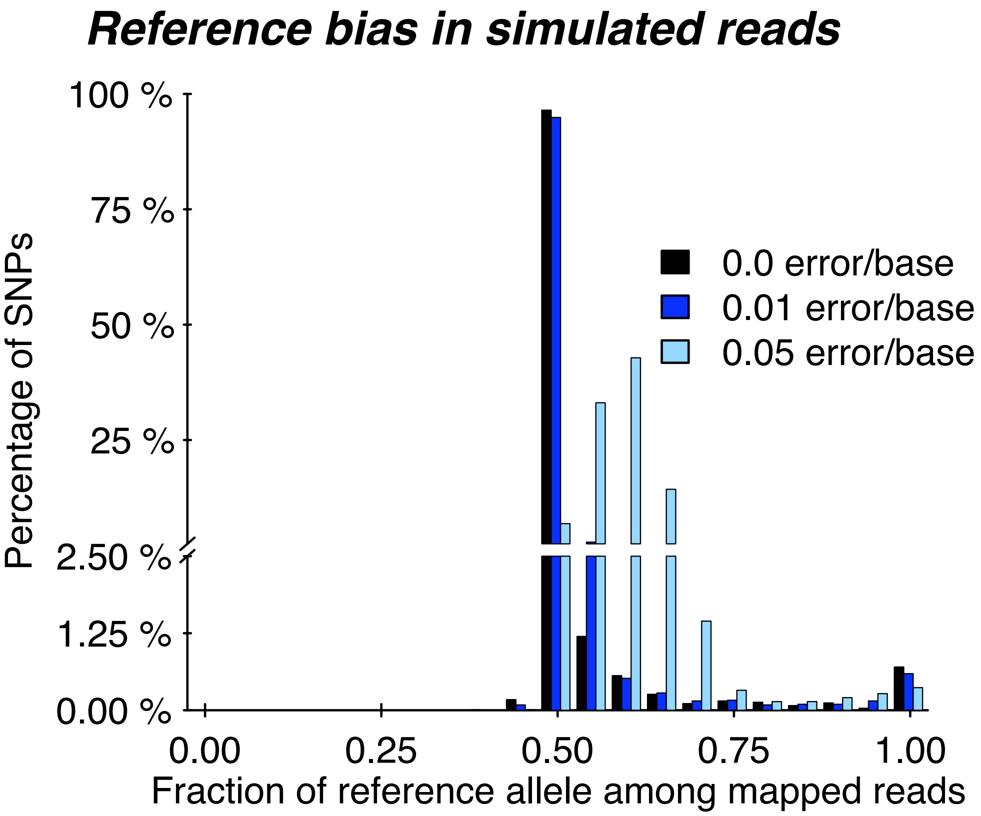
hg18 + strand

```
                                                            CCGCTGCCCTGCCCCTGGAGGGTGGCCCCACCGGCC
                                                                 .
                                                                 .
                                                                 .
                                    TGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCTG
                                   CTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCCT
                                  GCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCCC
                                 TGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGCC
                                CTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTGC
                               GCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCTG
                              TGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGCT
                             CTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCGC
                            CCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACCG
                           TCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCACC
                          CTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCAC
...ACAAGATGCCATTGTCCCCCGGCCTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCAC/GCGCTGCCCTGCCCCTGGAGGGTGGCCCCACCGGCCGAACAGCGAGCATATGCAGGAAG...
                          CTCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCAG
                           TCCTGCTGCTGCTGCTCTCCGGGGCCACGGCCAGC
                            CCTGCTGCTGCTGCTCTCCGGGGCCACGGCCAGCG
                             CTGCTGCTGCTGCTCTCCGGGGCCACGGCCAGCGC
                              TGCTGCTGCTGCTCTCCGGGGCCACGGCCAGCGCT
                               GCTGCTGCTGCTCTCCGGGGCCACGGCCAGCGCTG
                                CTGCTGCTGCTCTCCGGGGCCACGGCCAGCGCTGC
                                 TGCTGCTGCTCTCCGGGGCCACGGCCAGCGCTGCC
                                  GCTGCTGCTCTCCGGGGCCACGGCCAGCGCTGCCC
                                   CTGCTGCTCTCCGGGGCCACGGCCAGCGCTGCCCT
                                    TGCTGCTCTCCGGGGCCACGGCCAGCGCTGCCCTG
                                                                 .
                                                                 .
                                                                 .
                                                            GCGCTGCCCTGCCCCTGGAGGGTGGCCCCACCGGCC
```

non-reference allele reads

Degner et al., 2009

Some SNPs are heavily biased towards the reference allele



- For 1% of SNPs, 75% of reads (averaging across all read positions) carry the reference allele
- For 0.7% of SNPs, *all* mapped reads carry the reference allele

Degner et al., 2009

Some SNPs are heavily biased towards the reference allele

- Masking the reference allele did not solve the problem
- Instead directly accounting for mappability of different loci using simulated data is more helpful
  - Can remove loci where reads are better mapped back to the reference or non-reference allele

Degner et al., 2009

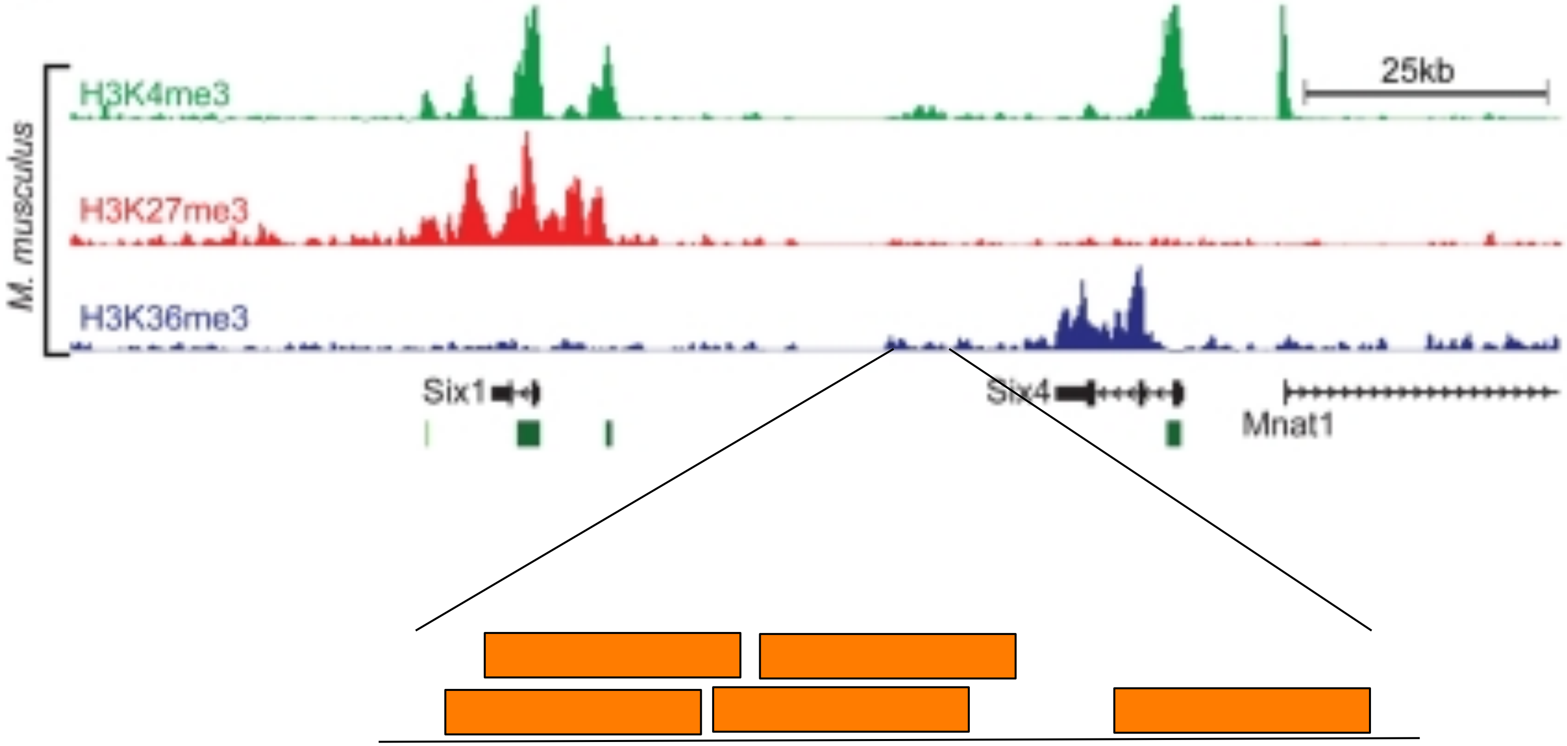
Quantifying binding

Quantifying binding - peak finding

- Good algorithms should:
  - Identify real peaks!
  - Estimate confidence (e.g., via calculation of a p-value)

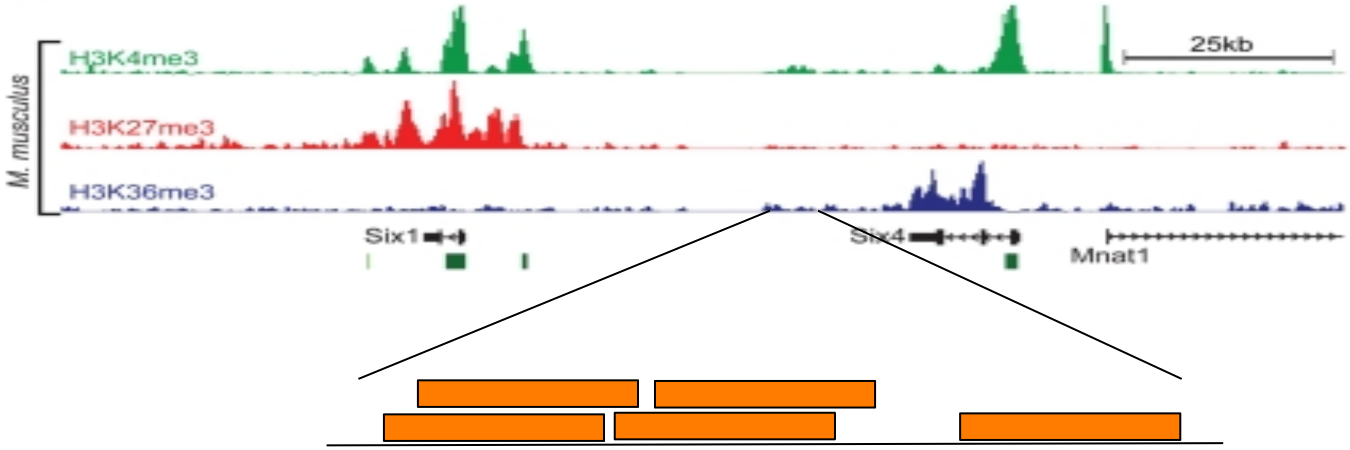
Huge number of algorithms for peak calling out there (> 60)

Quantifying binding – peak finding



Basic idea: Count the number of reads in windows and determine whether this number is above background – if so, define that region as bound

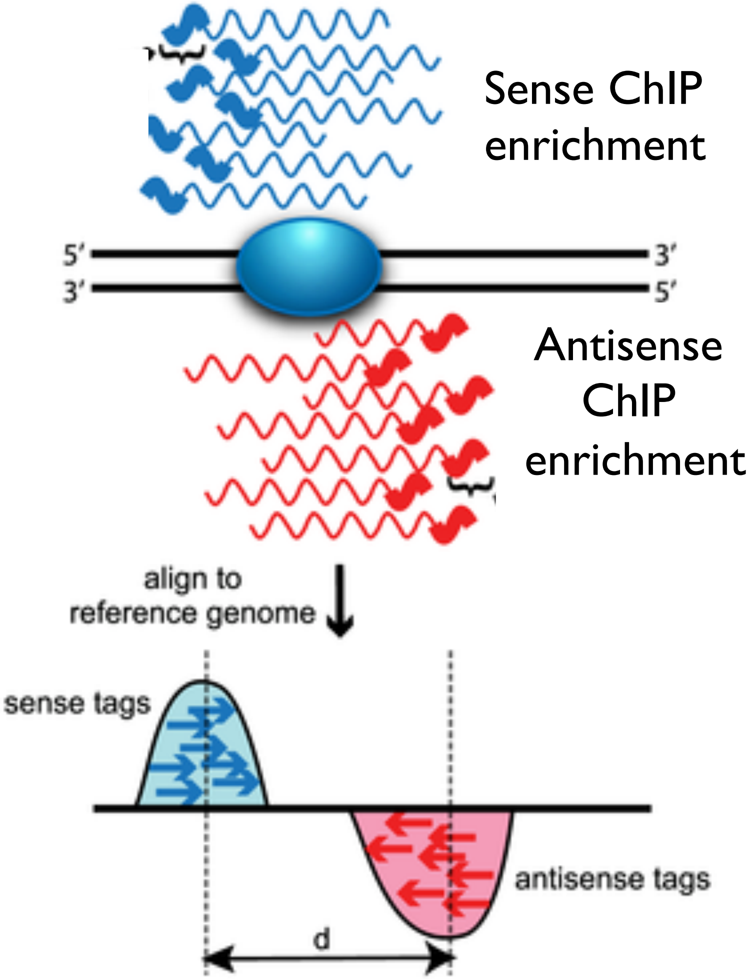
Quantifying binding – peak finding



- Calling a region as bound can be done in different ways:
  - Hard thresholds
  - HMMs
  - Compare bin counts to a background distribution determined from the input sample (assuming a Poisson or Negative Binomial distribution for example)

Quantifying binding – peak finding

- Another feature that some methods consider is that reads can be from the plus or minus strands
- In this case, for a given TF two peaks will be observed, separated by a constant distance, d
- This can be modeled either post-hoc, or by using strand specific calls

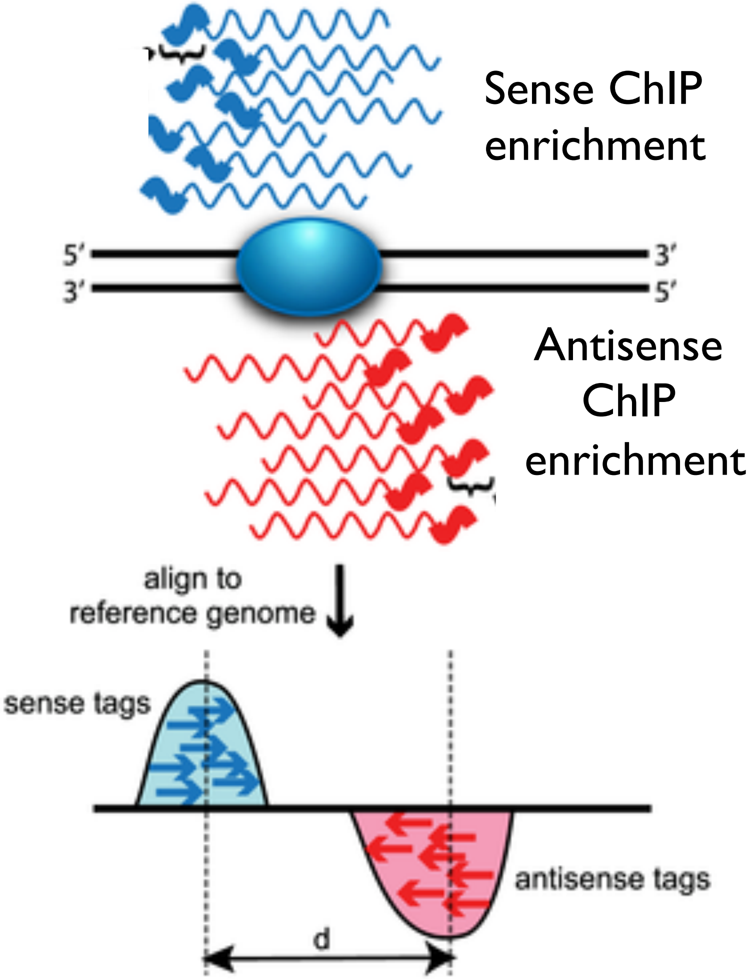

Sense ChIP enrichment

Antisense ChIP enrichment

Wilbanks et al., 2010

Quantifying binding – peak finding

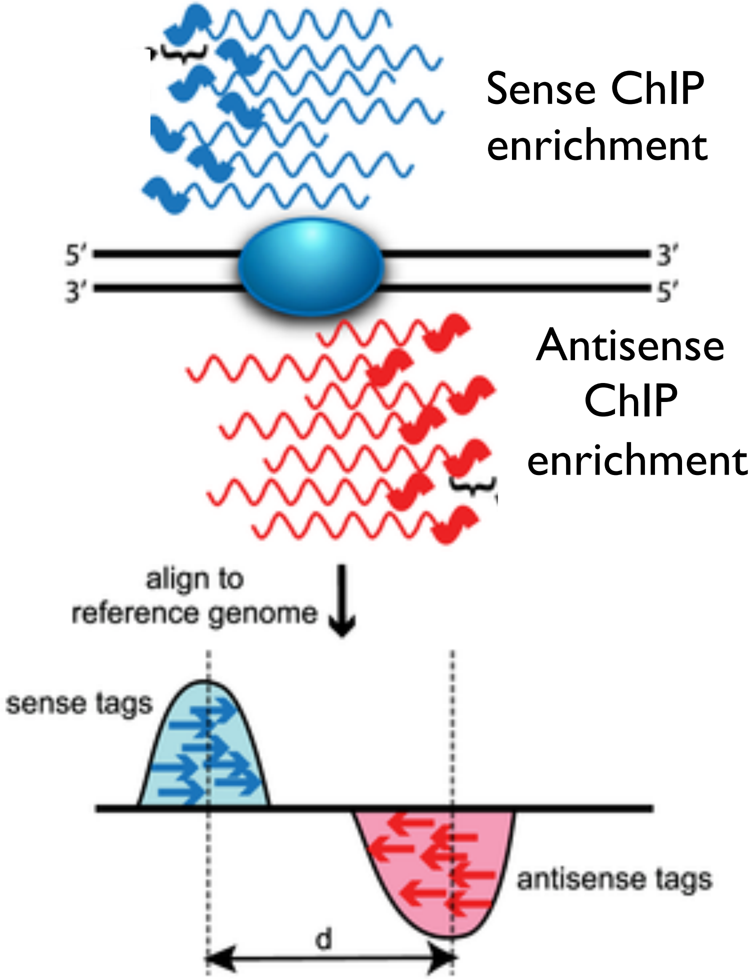
- However, this is only useful where the protein being assayed has a sharp, well defined binding site
- For histone modifications, with broad and sometimes shallow peaks, this information is less useful



Wilbanks et al., 2010

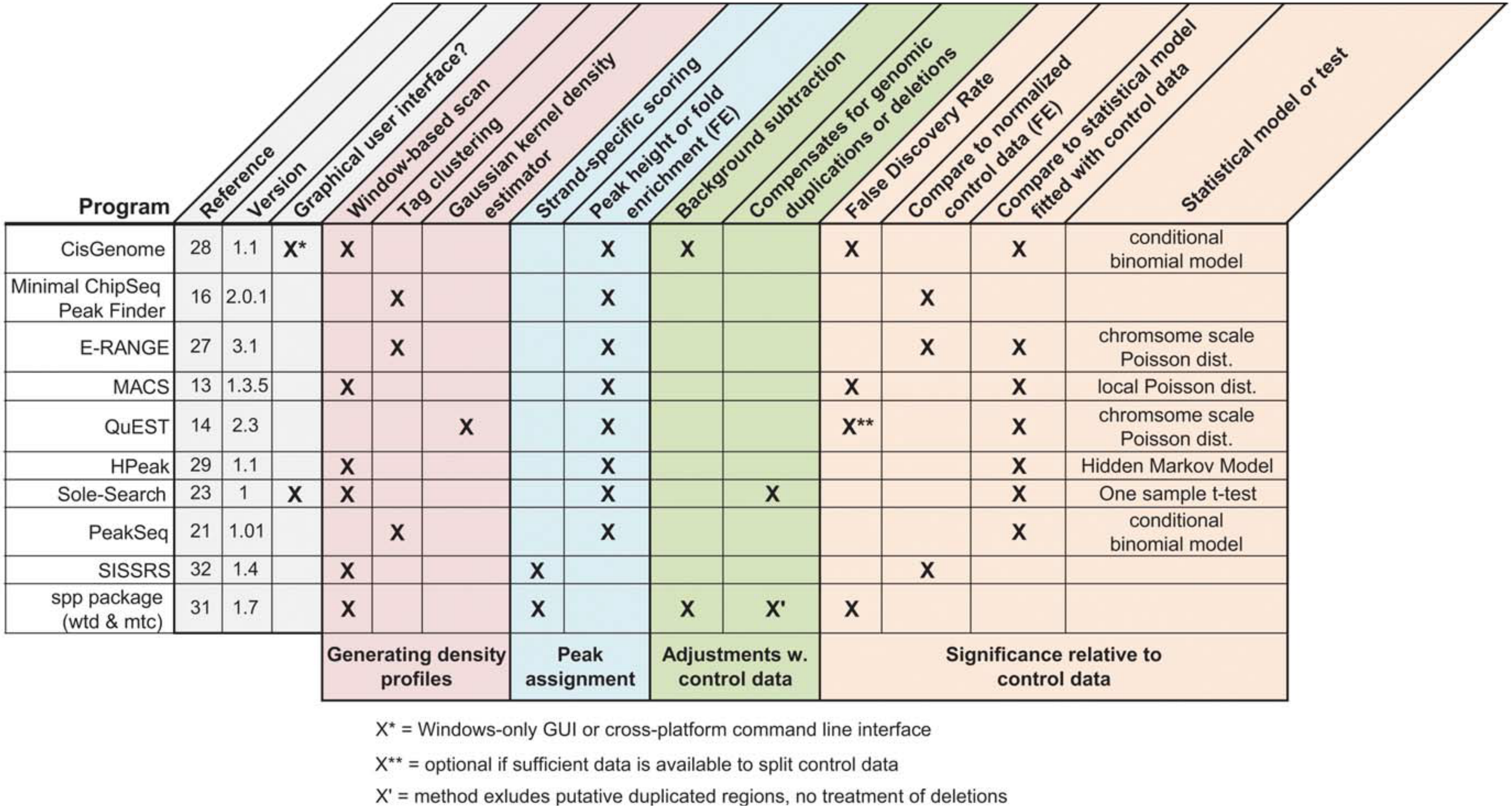
Quantifying binding – peak finding

- In general, methods have been developed for identifying regions where TFs bind – methods for identifying regions where histone modifications occur are less mature, although some approaches (e.g., those based upon HMMs) may be useful in this context[1,2]



[1] Xu, 2008
[2] http://www.ebi.ac.uk/~swilder/SWEMBL/

Summary of (some) different peak finders



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

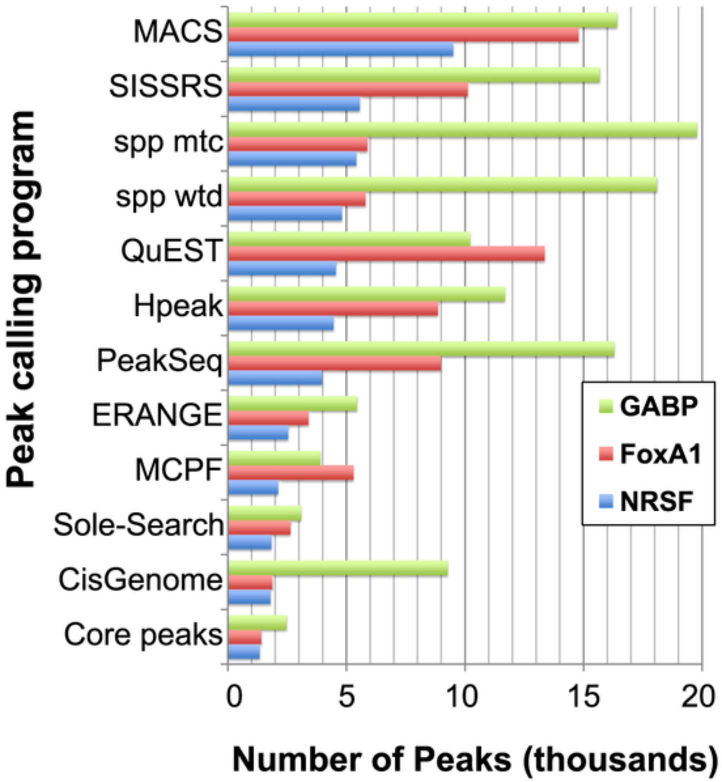
Wilbanks et al., 2010

How do methods compare?

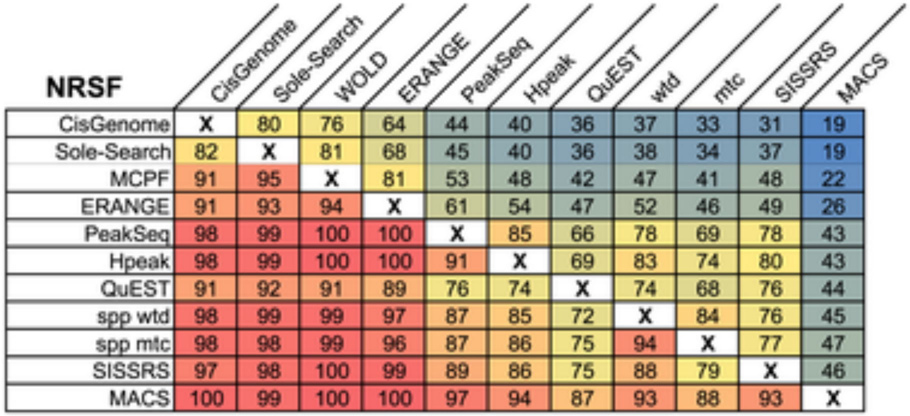
- Hard to do, since all methods rely on particular parameter values and need to be tuned accordingly to work best
- However, some groups have applied multiple methods to the same dataset using default parameters and compared results

How do methods compare?

- Wilbanks et al. compared the performance of 11 methods for calling binding sites for 3 TFs



Number of peaks called



NRSF	CisGenome	Sole-Search	WOLD	ERANGE	PeakSeq	Hpeak	QuEST	wtd	mtc	SISSRS	MACS
CisGenome	X	80	76	64	44	40	36	37	33	31	19
Sole-Search	82	X	81	68	45	40	36	38	34	37	19
MCPF	91	95	X	81	53	48	42	47	41	48	22
ERANGE	91	93	94	X	61	54	47	52	46	49	26
PeakSeq	98	99	100	100	X	85	66	78	69	78	43
Hpeak	98	99	100	100	91	X	69	83	74	80	43
QuEST	91	92	91	89	76	74	X	74	68	76	44
spp wtd	98	99	99	97	87	85	72	X	84	76	45
spp mtc	98	98	99	96	87	86	75	94	X	77	47
SISSRS	97	98	100	99	89	86	75	88	79	X	46
MACS	100	99	100	100	97	94	87	93	88	93	X

Proportion of calls in common between methods

How do methods compare?

- More encouragingly
  - Top 1,000 peaks are usually conserved (observed on previous slide)
  - Differences arise when looking for more marginal peaks
- Some common features
  - Control improves performance a lot
  - Deeper sequencing improves performance (only with control)
  - Ability to pinpoint peaks is still not very good

Wilbanks et al. 2010

What to do?

- Try several methods and take the intersection of calls?
- If biological replicates exist, only consider peaks called in multiple samples?
- Use confidence measures associated with each peak in downstream analysis?

What to do?

- Try several methods and take the intersection of calls?
- If biological replicates exist, only consider peaks called in multiple samples?
- Use confidence measures associated with each peak in downstream analysis?

In practice, many people employ some combination of the first and second points

Overview

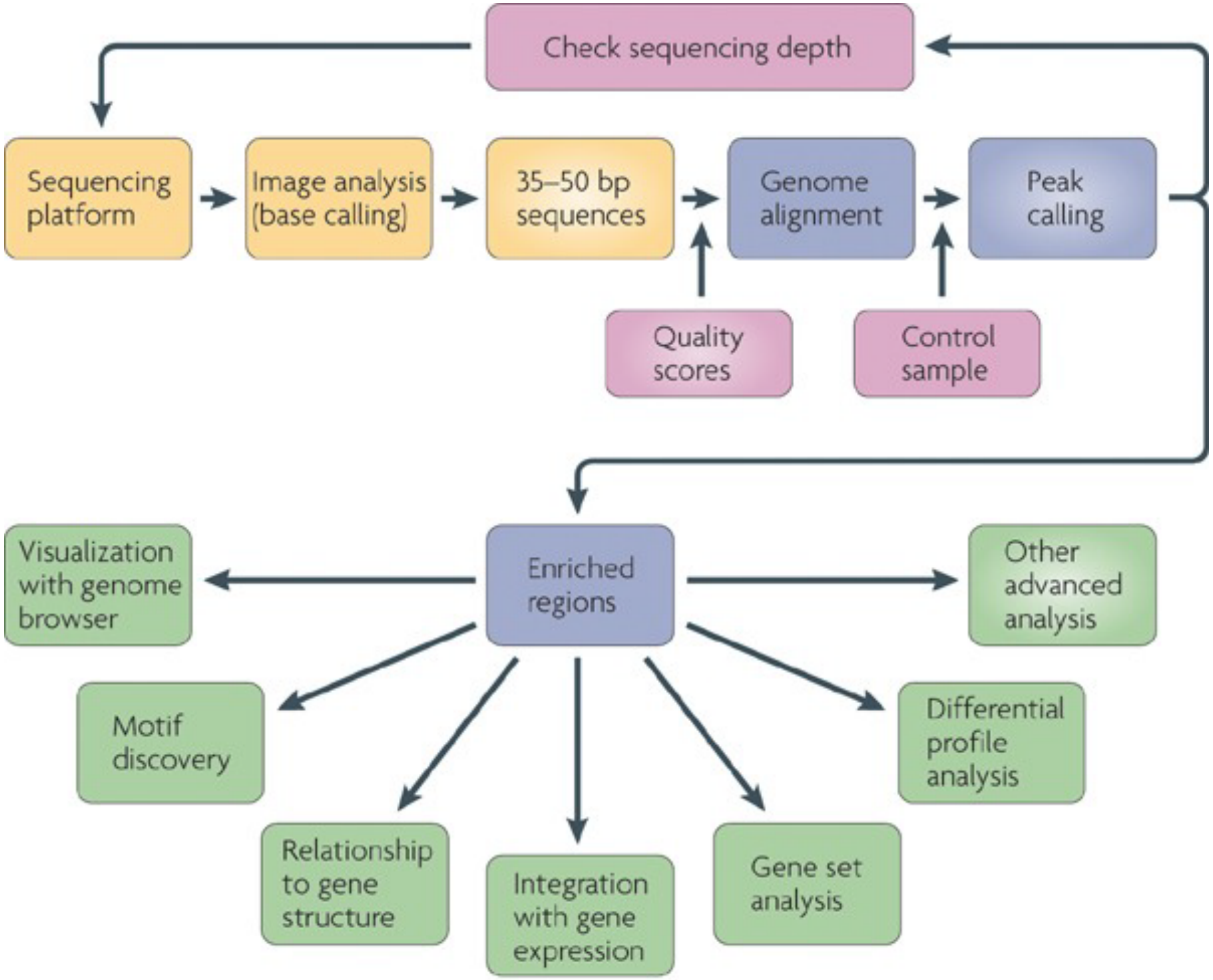
1. Designing ChIP-seq experiments

2. Read mapping and quantifying binding

3. Applications of ChIP-seq

4. Other applications using similar techniques

Downstream analysis



Park, 2009

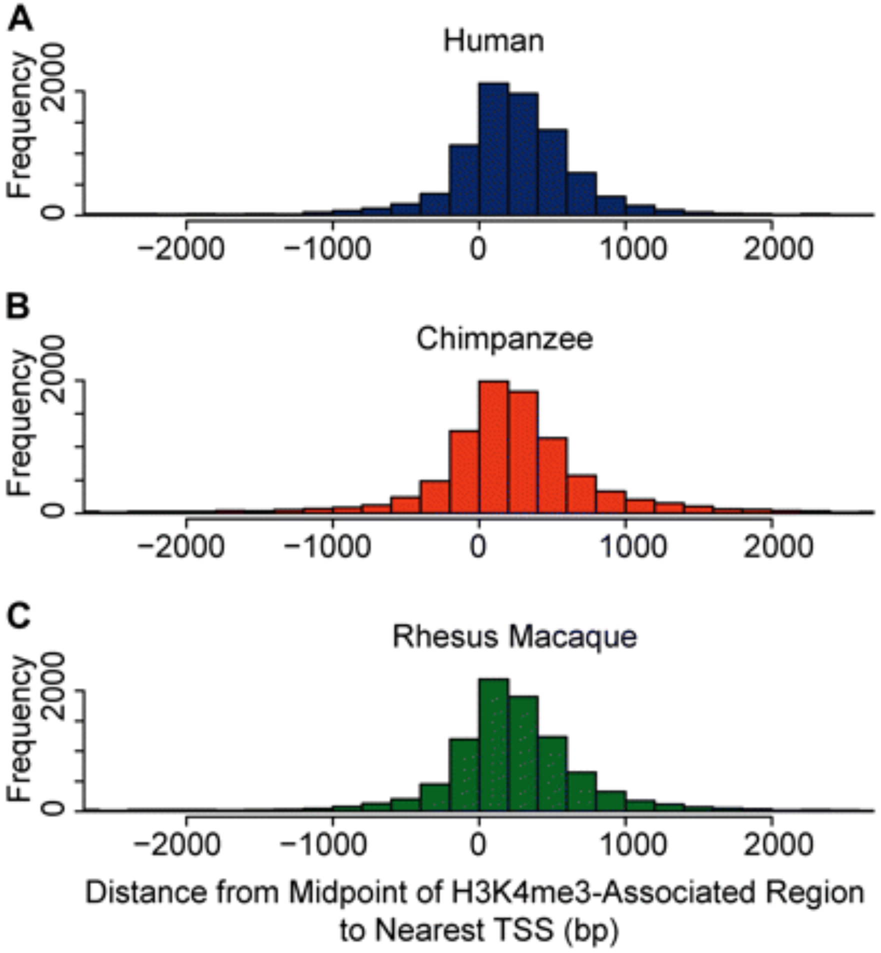
Motif discovery

- Take the set of significant bound sites and examine whether a particular motif is enriched amongst this set
  - Likely to find strong evidence of a motif for TFs
  - Less likely for histone modifications

Generally, standard motif finding algorithms (MEME, Weeder etc.) are used for this

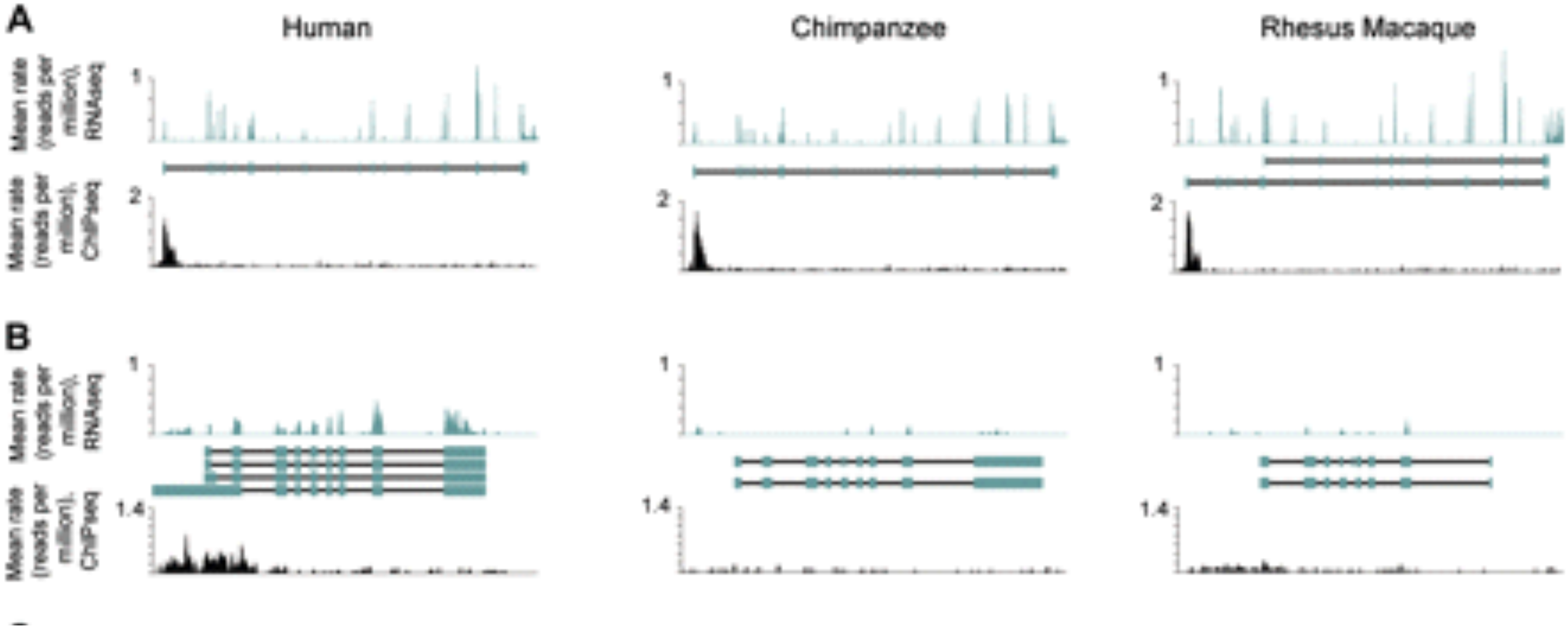
Relationship to gene structure

- Used ChIP-Seq to look for H3K4me3 regions in human, chimpanzee and rhesus macaque LCLs
- H3K4me3+ regions called using MACS and a two-step conditional cutoff; adjacent peaks were also merged
- In all three species ~61% of H3K4me3+ regions are enriched around the TSS



Cain et al., 2011

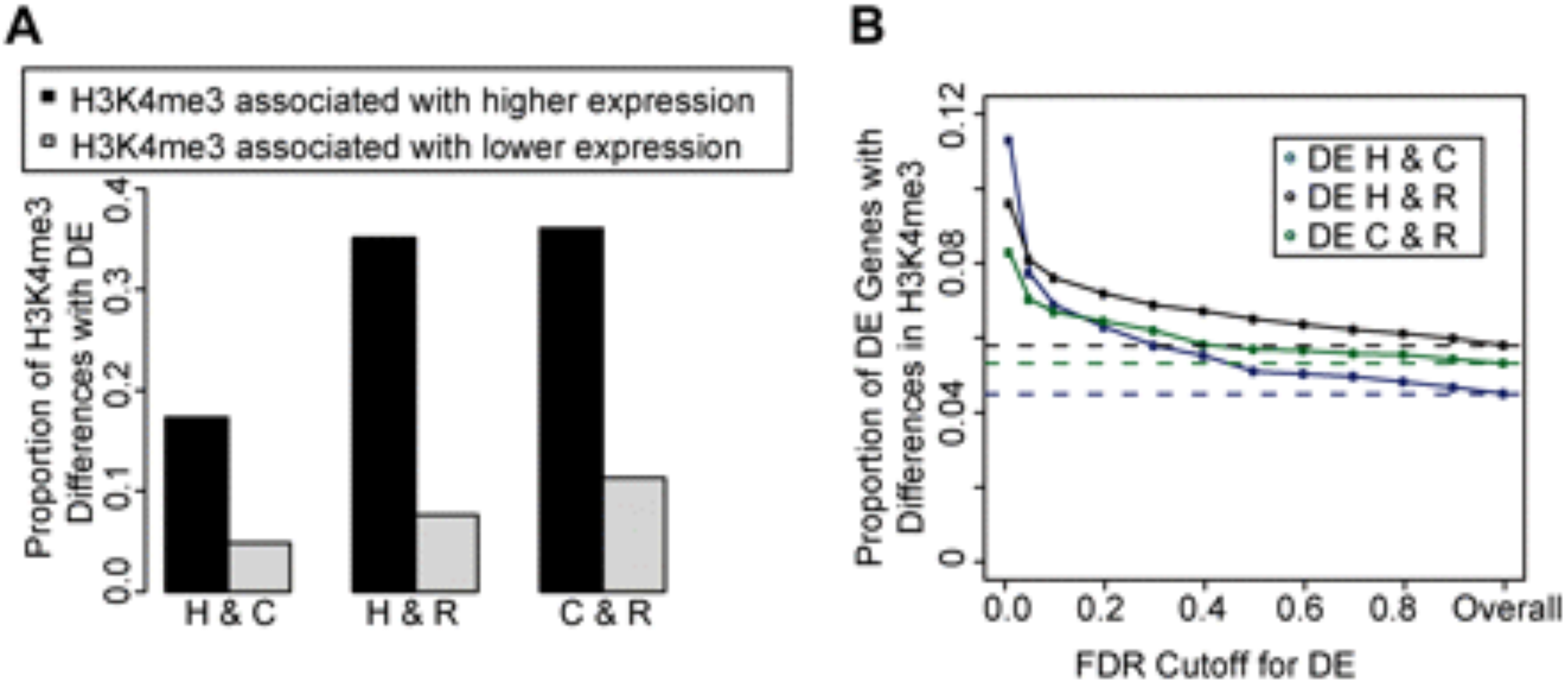
Relationship to gene expression



H3K4me3+ regions are associated with active genes

Barski et al., 2007
Cain et al., 2011

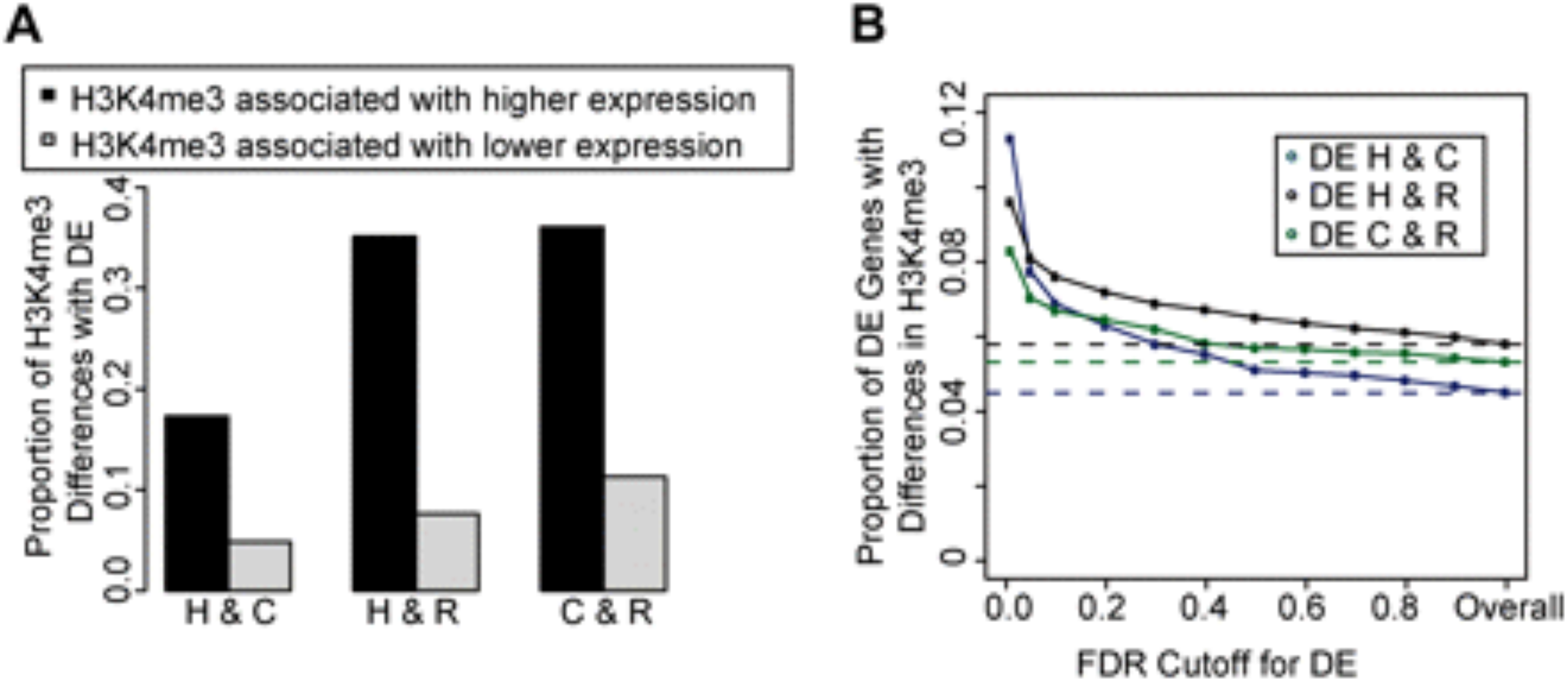
Relationship to gene expression



- Differential expression called between genes for each species using a Poisson mixed-effects model
- For each comparison, amongst the set of genes with H3K4me3 in one species but not the other, the majority of genes that were differentially expressed and overlapped with this set were more highly expressed

Cain et al., 2011

Relationship to gene expression



- The results suggest that changes in H3K4me3 status could explain between 2.5% (FDR 10%) and 6.8% (FDR 1%) of differences in gene expression levels between humans and chimpanzee, and similar proportions of differences for the other comparisons

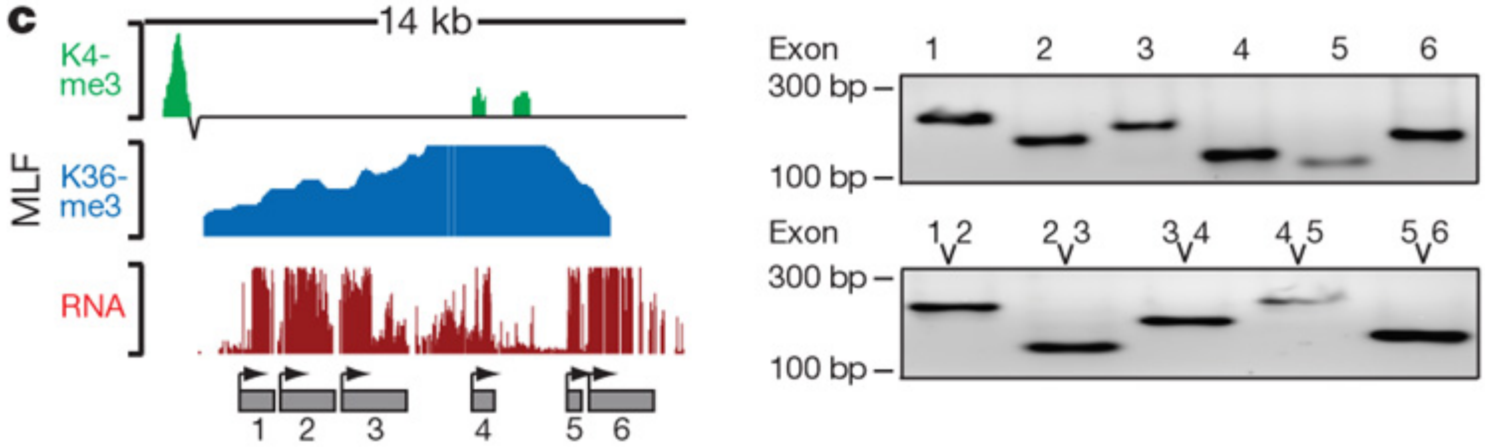
Cain et al., 2011

Combining TF binding and chromatin marks yields biological insight

- It has been shown that H3K4me3 marks active TSS sites and H3K36me3 is bound along transcribed regions[1]
- Recently, these two histone marks have been used to identify novel large intervening non-coding RNAs (lincRNAs)[2]

1. Mikkelsen et al., 2007
2. Guttman et al., 2009

Combining TF binding and chromatin marks yields biological insight



- Using conservative criteria, Guttman et al. found 1250 K4-K36 domains that did not overlap annotated genes
- Compared to other intergenic regions, these newly identified lincRNAs are more conserved; however, they are less conserved than coding sequence
- Nevertheless, they hypothesise that these lincRNAs must be functional and showed that several did have specific functions

Differential binding

- Between two sets of samples can we determine whether the same region is bound with the same intensity

What's the biological meaning?

More intense peak → Stronger binding → Stronger biological effect

Differential binding

- Between two sets of samples can we determine whether the same region is bound with the same intensity

How to measure intensity?

- Mean peak height
- Maximum peak height
- Average number of reads under peak

Differential binding

- Between two sets of samples can we determine whether the same region is bound with the same intensity

How to measure differential binding

- Can we use approaches such as DESeq?
- Will this adequately model the variance?
- Unclear at present since, frustratingly, biological replicates are not very prevalent for ChIP-Seq experiments

Overview

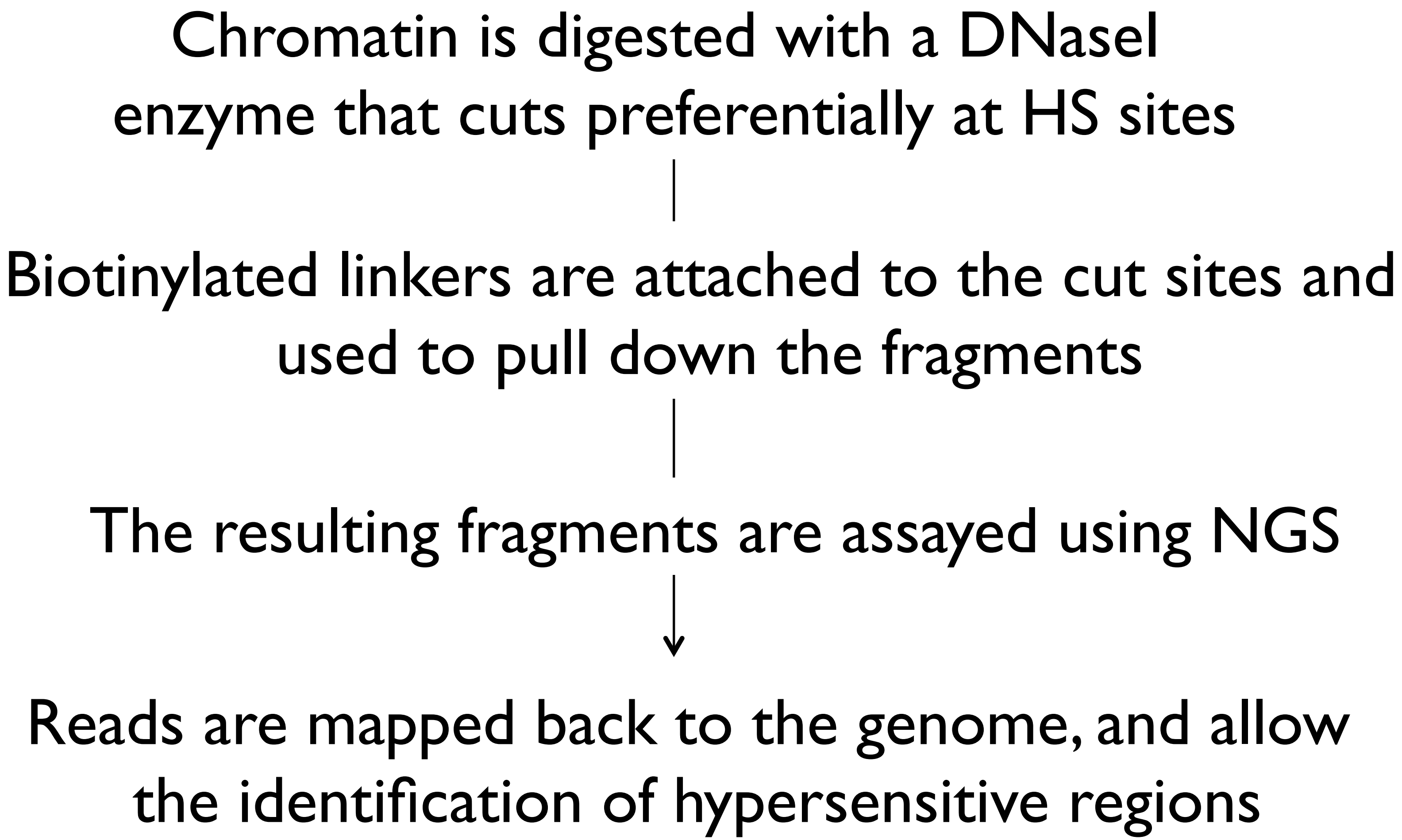
1. Designing ChIP-seq experiments
2. Read mapping and quantifying binding
3. Applications of ChIP-seq
4. Other applications using similar techniques

DNase-seq

- Regions of the genome that are hypersensitive to cleavage by DNaseI have been associated with different regulatory elements, including promoters, enhancers and silencers[1]
- More recently, they have been associated with histone modifications and TF binding[2]
- Thus, identifying these regions is of great biological interest

1. Boyle et al., 2008
2. ENCODE, 2007

DNase-seq



Chromatin is digested with a DNaseI enzyme that cuts preferentially at HS sites

Biotinylated linkers are attached to the cut sites and used to pull down the fragments

The resulting fragments are assayed using NGS

↓

Reads are mapped back to the genome, and allow the identification of hypersensitive regions

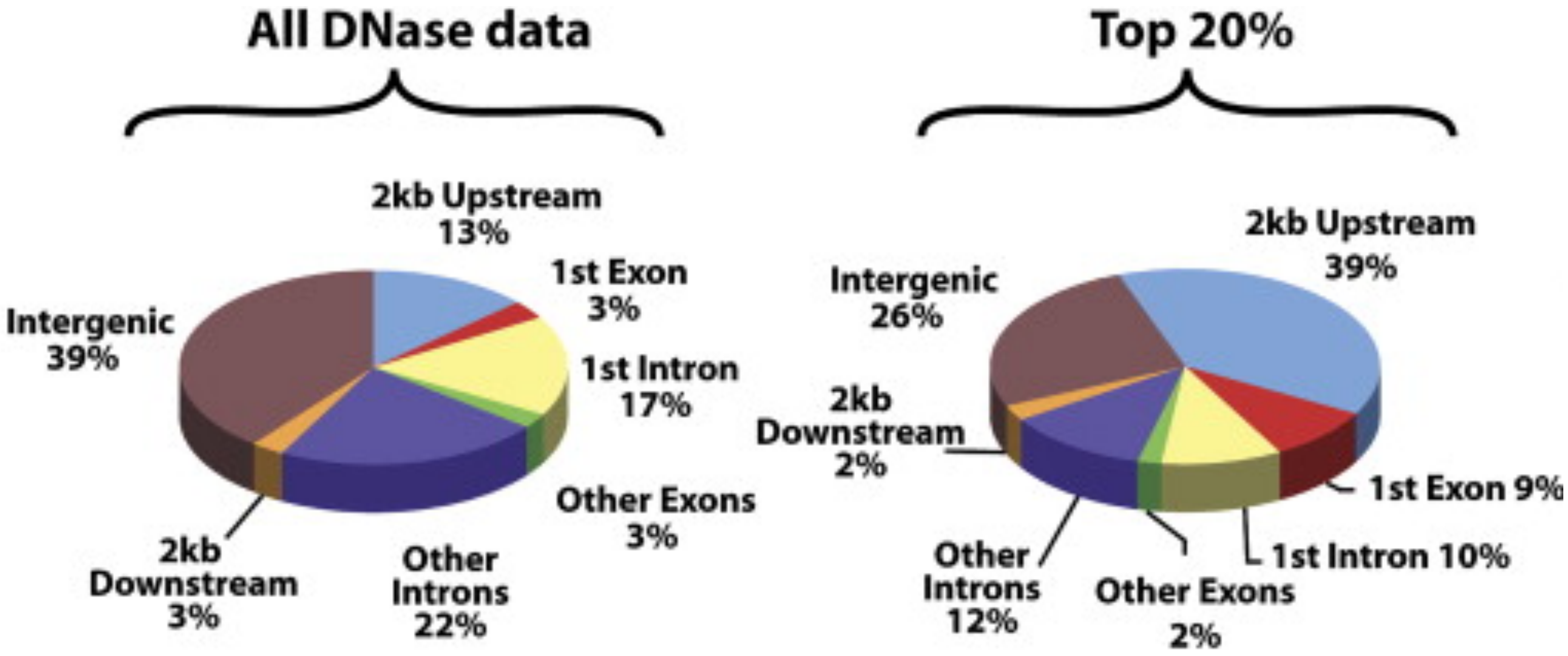
Boyle et al., 2008

DNase-seq

- Boyle et al. used DNase-seq to study primary human CD4+ cells
- They used a kernel smoothing based approach to identify ~95,000 HS sites

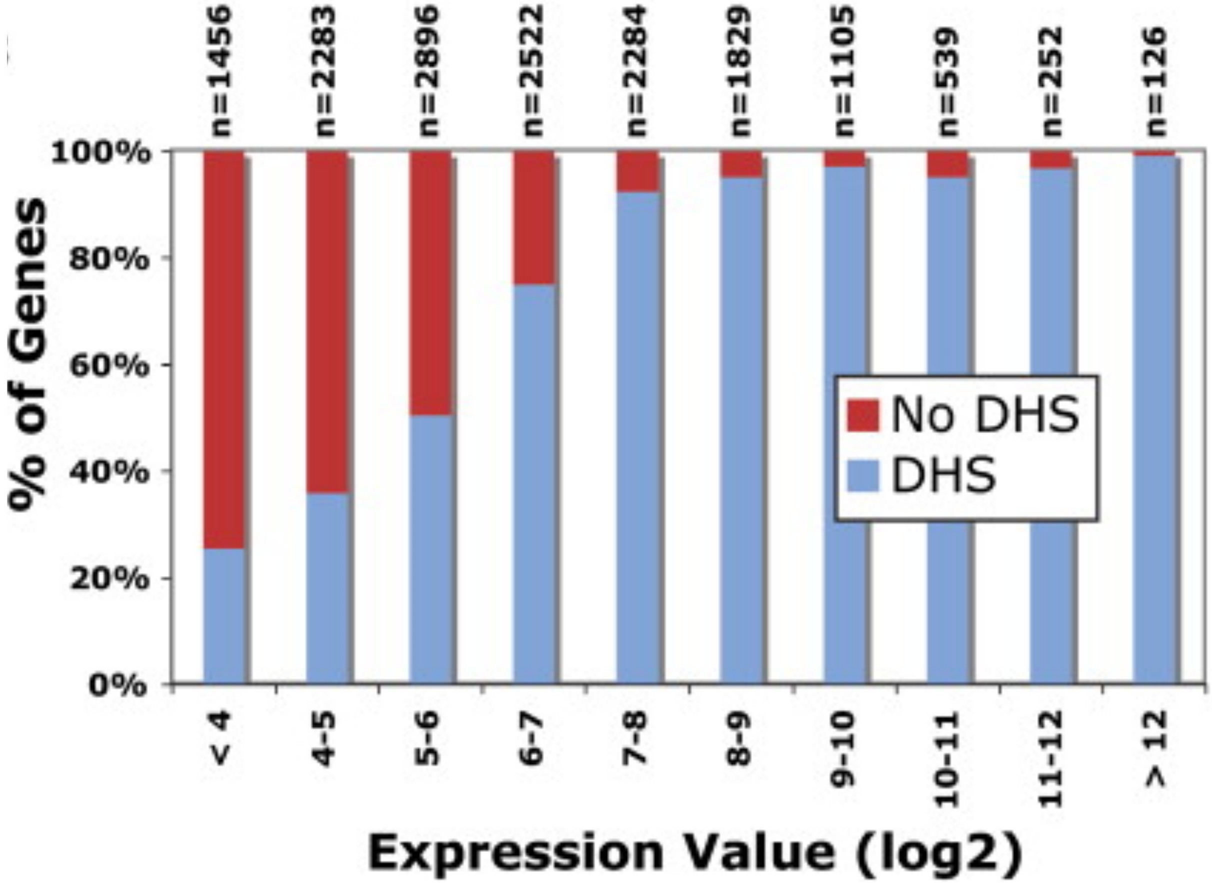
DNase-seq

- The strongest HS sites were enriched in the promoter region and the first exon of annotated transcripts



DNase-seq

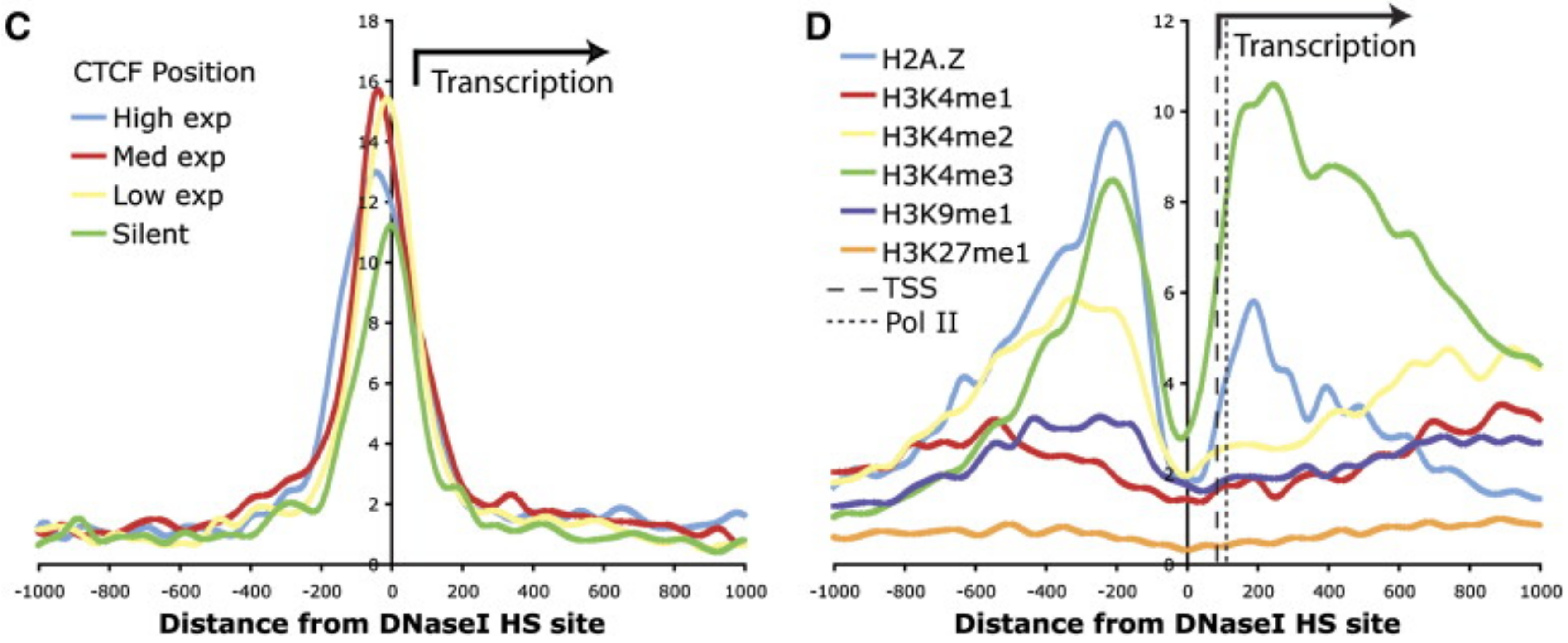
- Moreover, genes with a DNaseI HS site upstream of the 5' TSS were more highly expressed



Shows that the presence of strong DNaseI cut sites is associated with expression

DNase-seq

- Given their association with expression and location 5' of a gene's TSS, it is perhaps not surprising that DNaseI HS sites are associated with TF binding and histone modifications



These data are for highly expressed genes only

DNase-seq

- As a result, some groups have used DNaseI and histone modification data to predict whether a genomic region containing a motif is bound by a TF[1]

- This has the advantage of potentially allowing one to assay multiple TFs in one experiment – this will be especially useful for TFs with poor antibodies

[1] Pique-Regi, 2010

DNase-seq

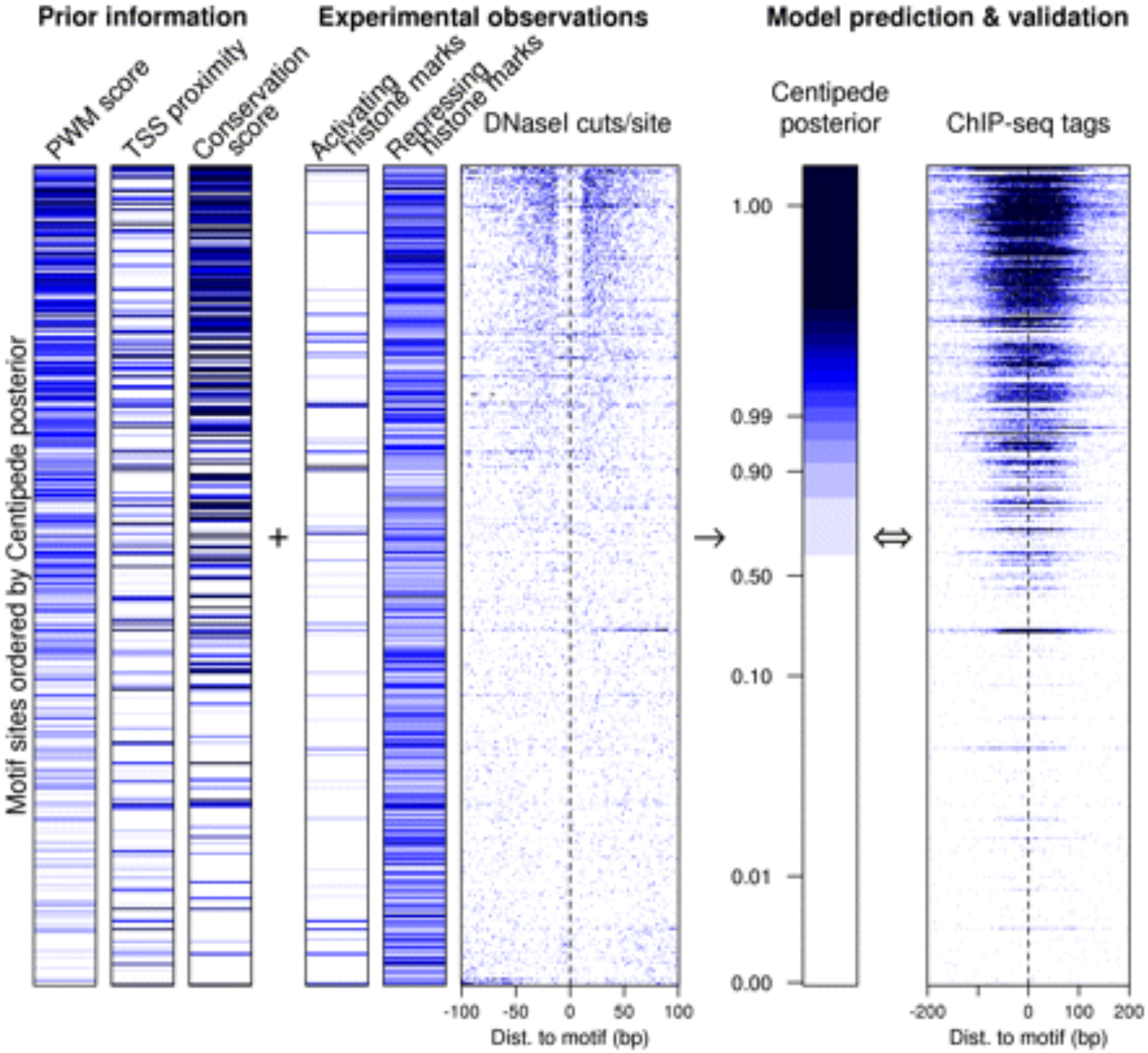
- For each site in the genome where a specific motif is present, Pique-Regi use DNaseI and histone modification data to fit the following mixture model:

$$P(\text{Data}) = \pi P(\text{Data}|\text{TF bound}) + (1-\pi)P(\text{Data}|\text{TF unbound})$$

Prior determined from conservation information, PWM score etc. Likelihood calculated by assuming the number of reads in a region around a motif follow a negative-binomial distribution and (in the bound case) the per-base pair data can be explained by a multinomial model

Can calculate the posterior probability that a region is bound

DNase-seq



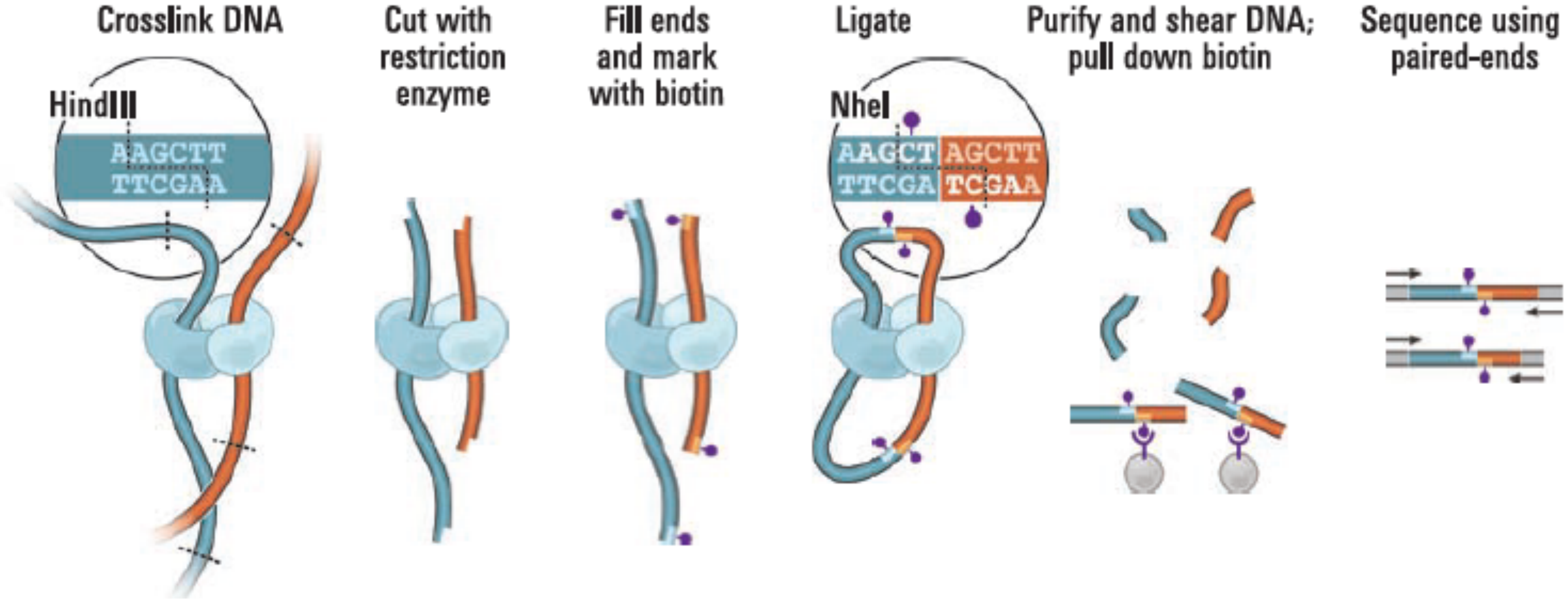
For *REST* one can see that the predictions of TF binding from the model closely follow the independently generated ChIP-Seq data

Chromatin conformation

- We have a tendency to think of a chromosome as a linear entity
- However chromatin is folded in highly complex ways, which can result in distant parts of the chromosome coming into close proximity (e.g., enhancer elements and gene promoters)

Chromatin conformation

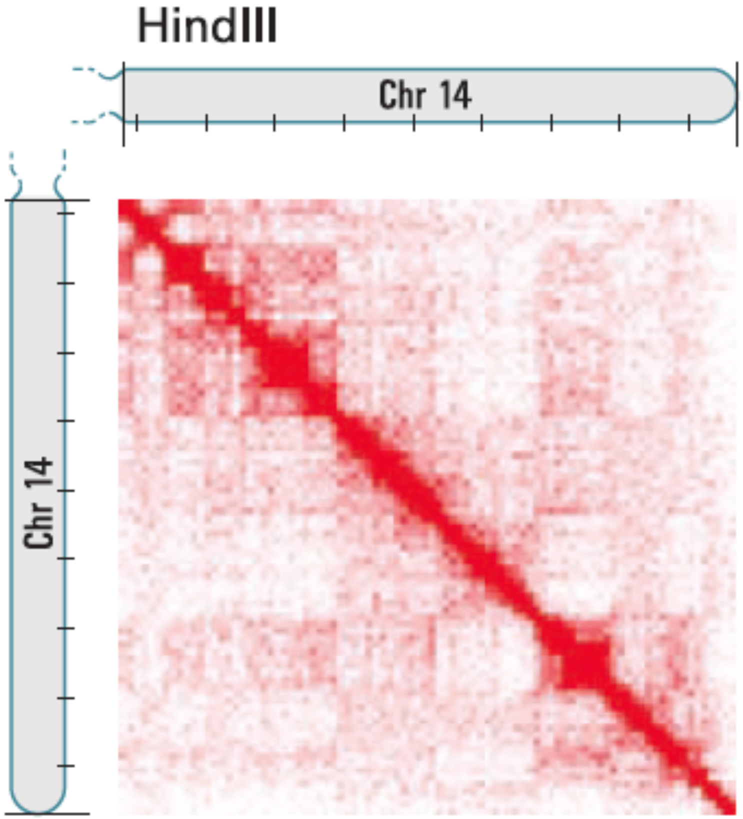
- Next-generation sequencing techniques (Hi-C) can enable us to study these interactions genome-wide


Crosslink DNA
HindIII
AAGCTT
TTCGAA
Cut with restriction enzyme
Fill ends and mark with biotin
Ligate
NheI
AAGCT AGCTT
TTCGA TCGAA
Purify and shear DNA; pull down biotin
Sequence using paired-ends

Liebermen-Aiden et al., 2009

Chromatin conformation

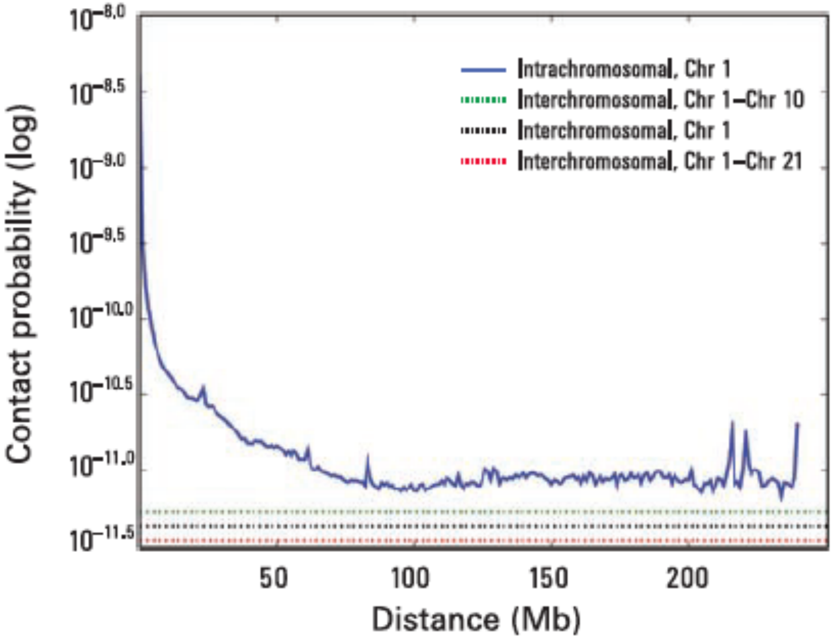
- Libermen-Aiden et al., applied this method to a CEU cell line
- They divided the genome into 1Mb windows and counted the number of reads, m_{ij} that linked window i to window j



These data can be represented as a heatmap (red = lots of links, white = no links)

Chromatin conformation

- They calculated the average contact probability within each chromosome and between chromosomes
- This showed that the probability of contact increases with reduced genomic distance
- It also shows that the probability of inter-chromosomal contact is small

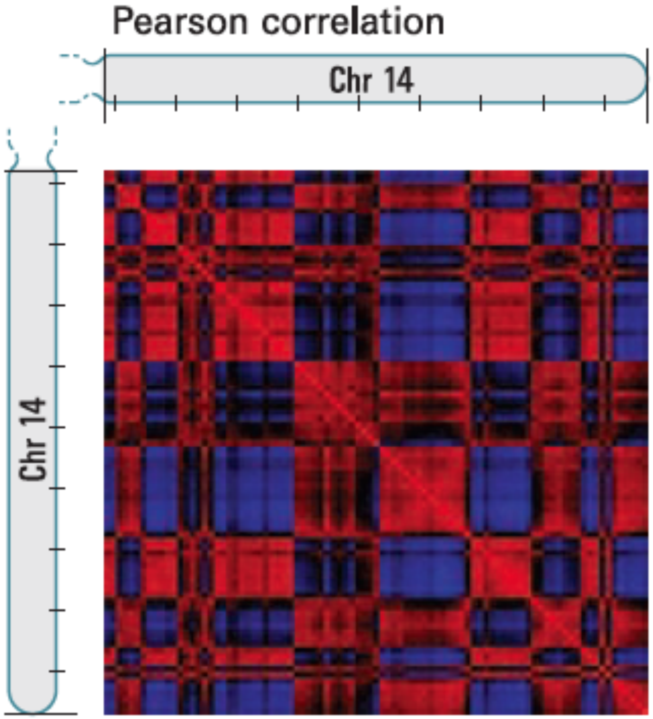


This analysis confirmed previous work suggesting there were well defined chromosomal domains

Chromatin conformation

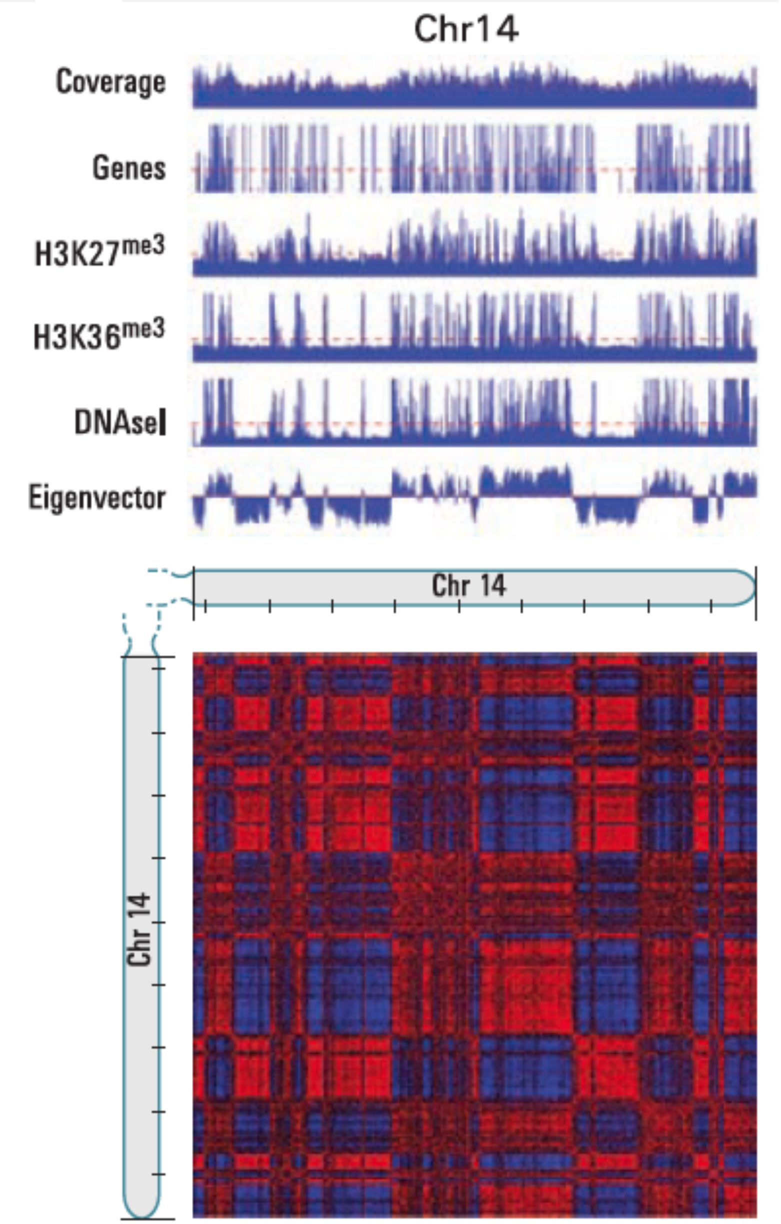
- Since the highest contacts are observed for regions that are located near one another (artefactual?), Lieberman-Aiden et al. normalized the data to account for this

Calculating the Pearson correlation matrix for the normalized data revealed that each chromosome could be broken down into two compartments (regions with lots of contacts between one another, but not to other regions)



Chromatin conformation

- By correlating the conformation data with information about histone modifications, the authors determined that one of the compartments was associated with gene dense and transcribed regions



Statistical approaches for multi-dimensional data

- Visualization techniques – heatmaps, principal components plots
- Unsupervised clustering (hierarchical approaches, linear discriminant analysis)
- Supervised clustering (using a training set to determine a rule by which other observations can be classified)

Statistical approaches for multi-dimensional data

- When you have a response variable such as gene expression and multiple explanatory variables (e.g., various histone marks, TF binding sites) how does one determine the relevant explanatory variates?
  - Stepwise regression
  - Penalized regression approaches (LASSO)
  - Bayesian regression with sparse priors

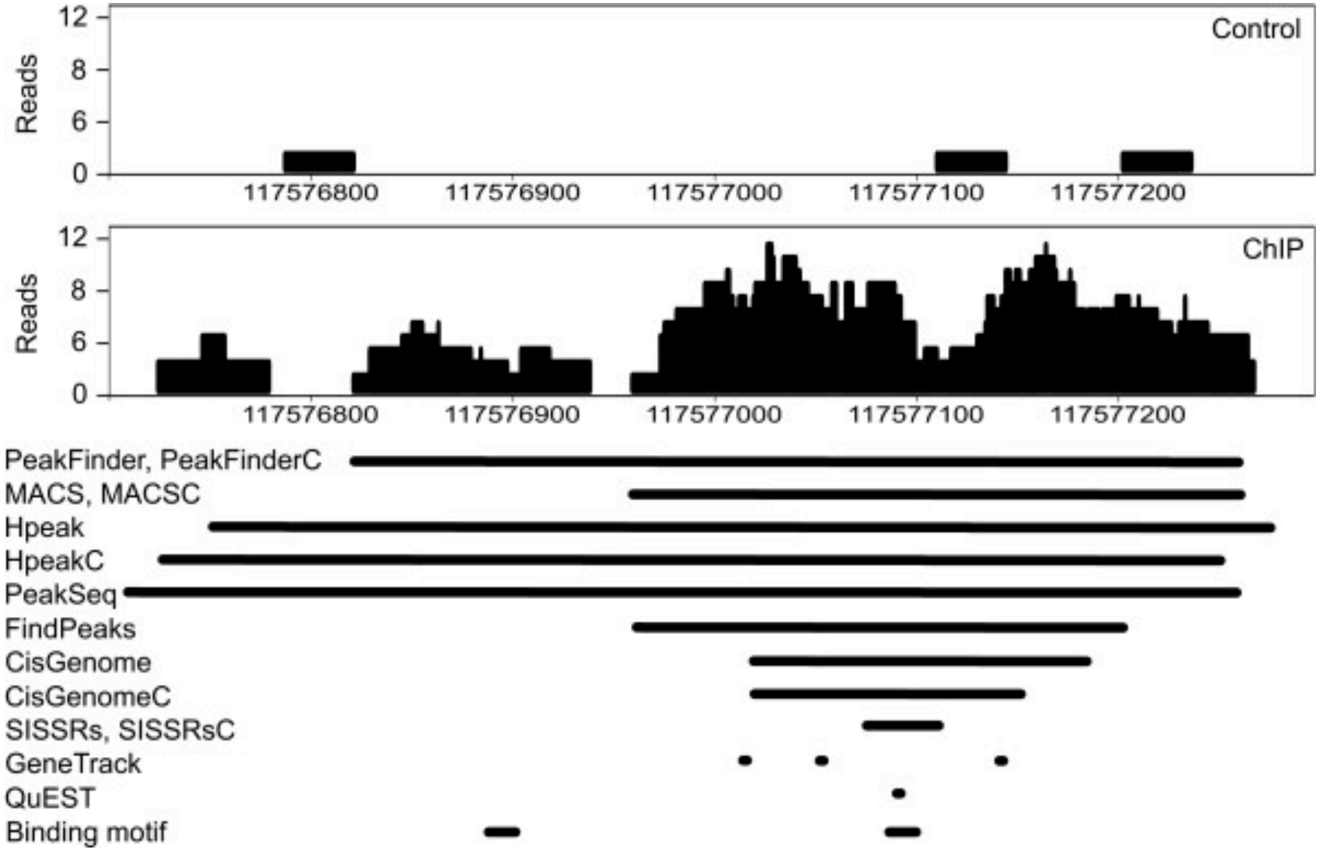
Acknowledgements

- Nick Luscombe
- Petra Schwalie
- Andre Faure
- Wolfgang Huber

How do methods compare?

- Another study by Laajala et al. also compared peak-calling methods

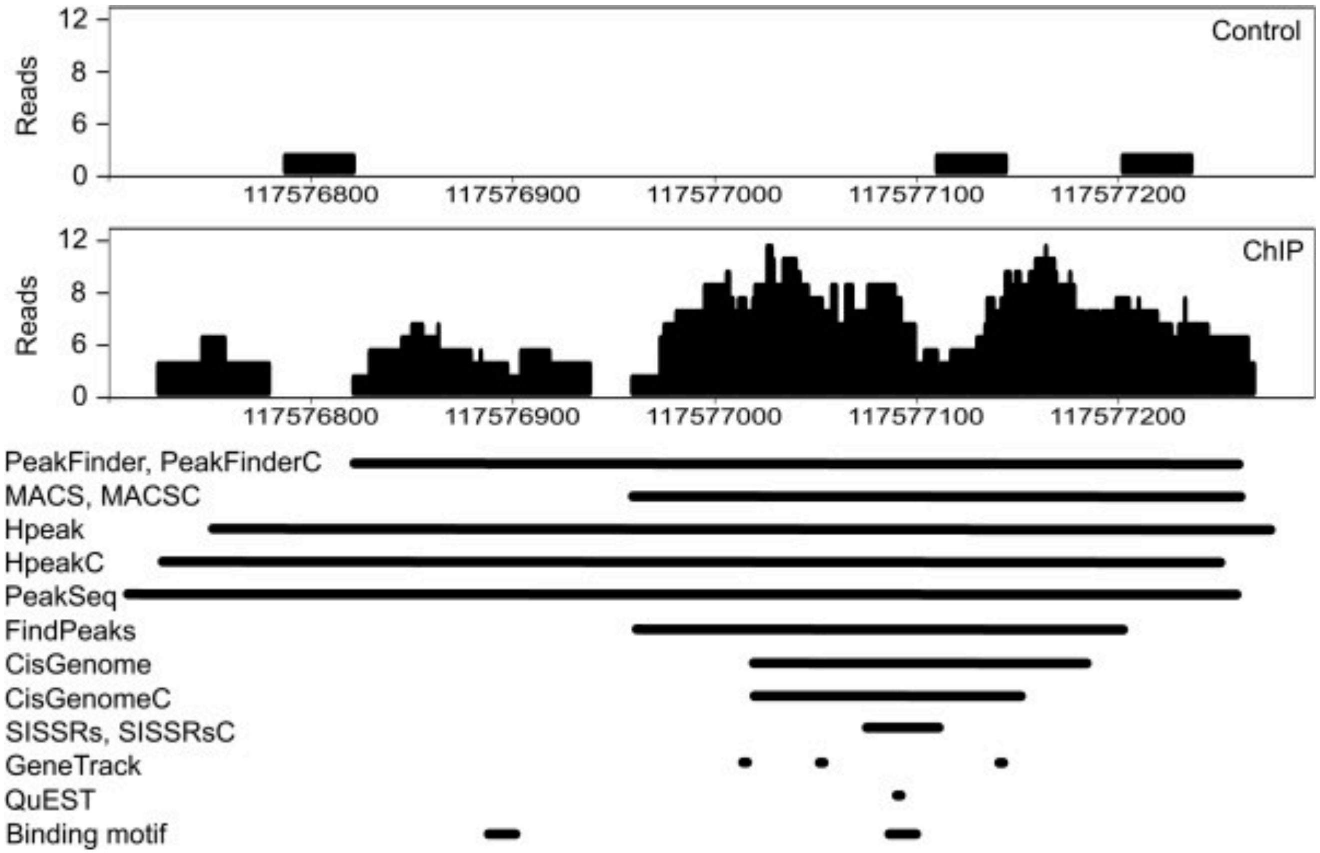
The authors applied 14 calling methods to ChIP-seq data generated for *Stat1*



How do methods compare?

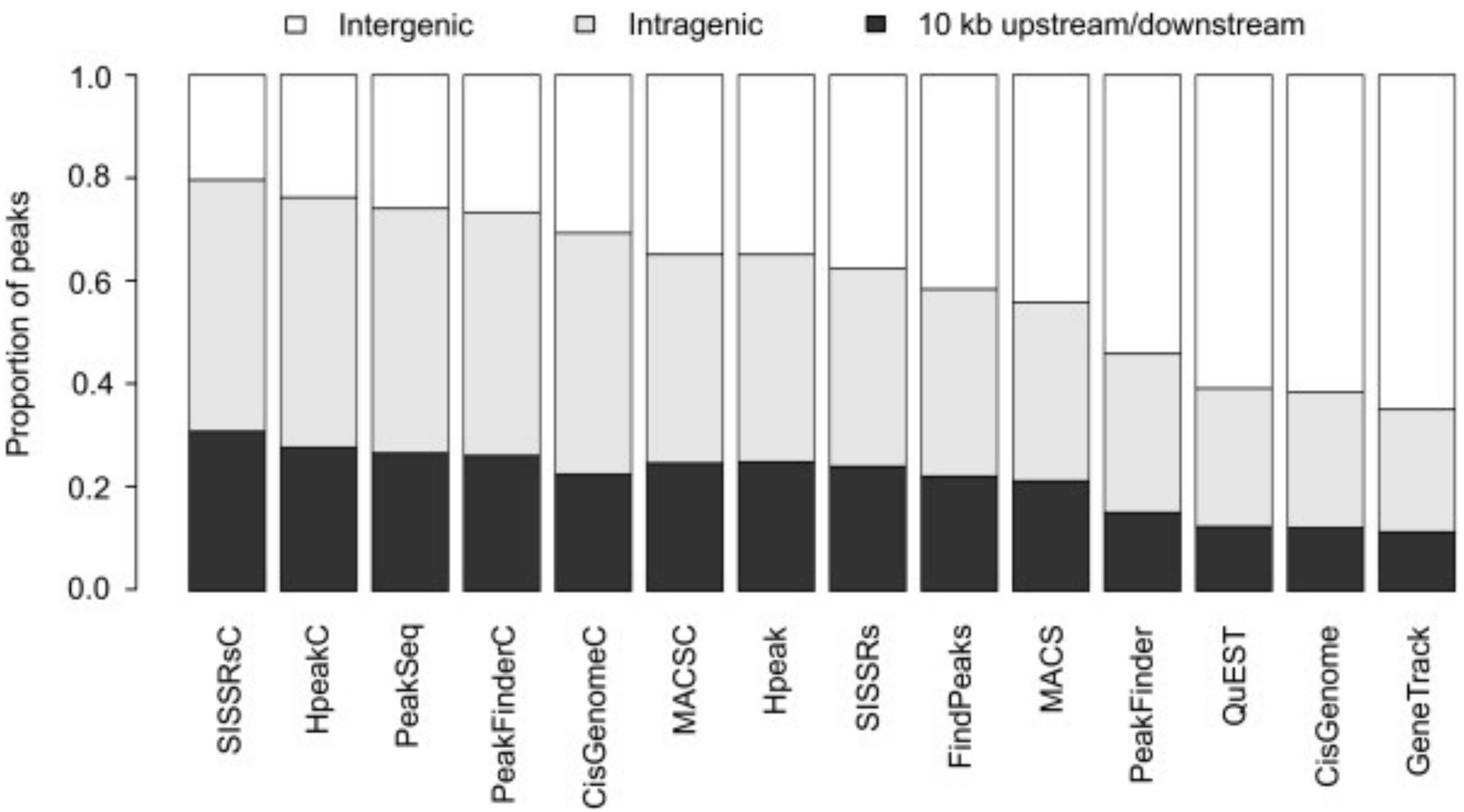
- Another study by Laajala et al. also compared peak-calling methods

The authors applied 14 calling methods to ChIP-seq data generated for *Stat1*



The length of the region identified varies hugely

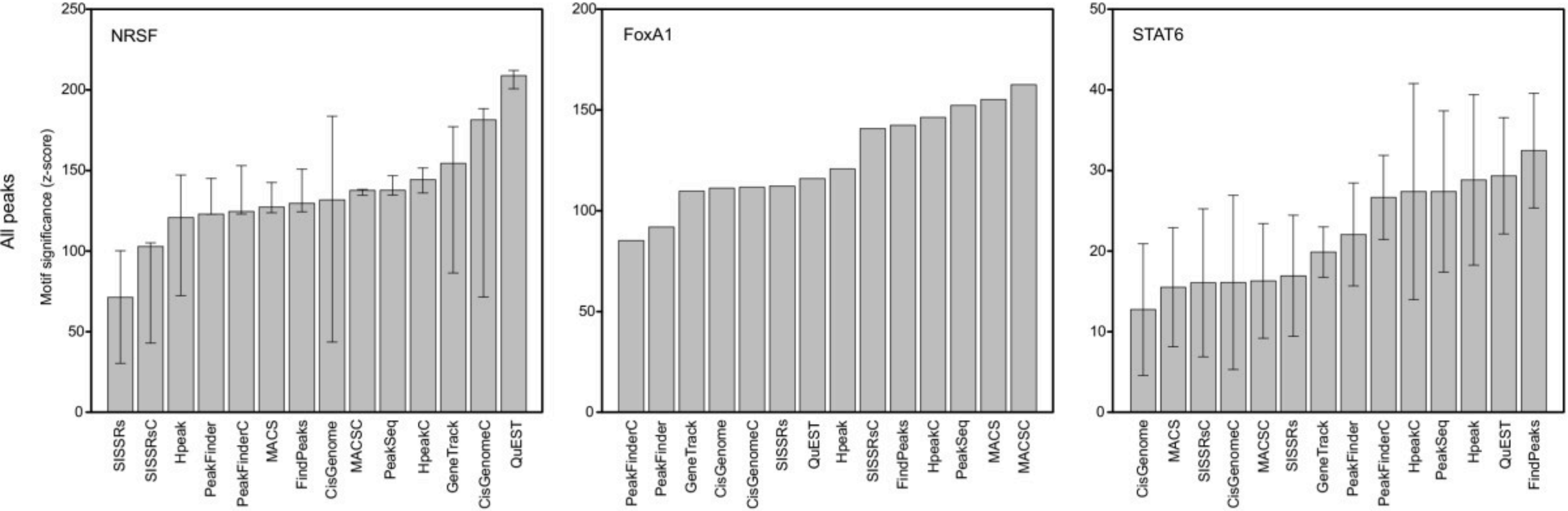
How do methods compare?



More worryingly, using different methods can result in very different biological conclusions

How do methods compare?

- Another study by Laajala et al. also compared peak-calling methods



Using known motifs as a measure of call quality (in itself quite ineffective) the authors compared different calling methods