ChIP-seq Experiments

R. Gentleman

(lots of slides thanks to John Marioni)

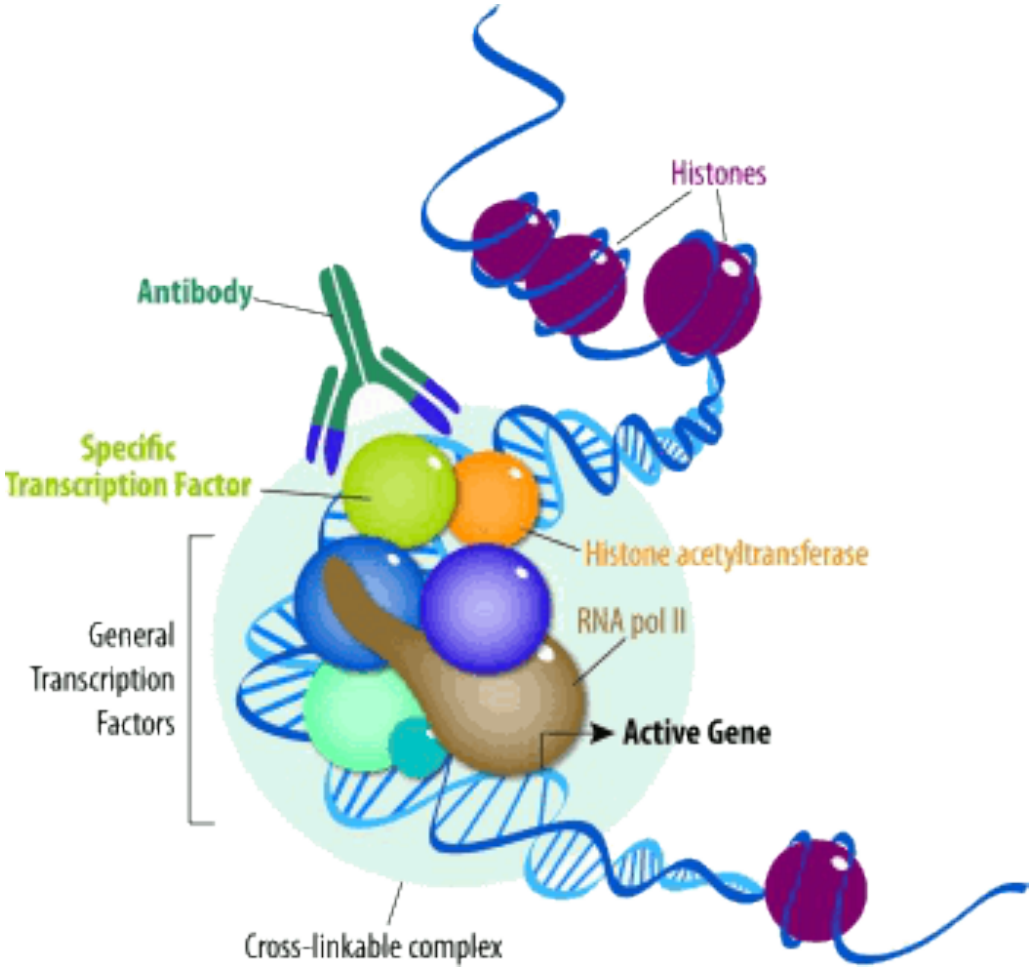
Biological Motivation

- Chromatin-immunopreciptation followed by sequencing (ChIP-seq) is a powerful tool
- epigenetics
    - histone modifications
    - methylation
- locating transcription factor (TF) DNA interactions
- detecting what nucleic acid sequences any protein is interacting with
  – ribosomal profiling

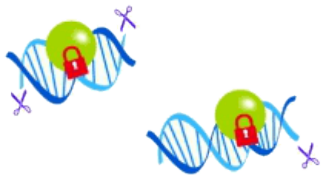
Assay Basics

- cross-link proteins to DNA or RNA
  - usually using formalin
- introduce tagged antibody that targets the protein or entity of interest
- enrich the output by selecting for the tagged protein (immuno-precipitation)
- undo cross-linking
- purify for either RNA or DNA
- sequence – and then process

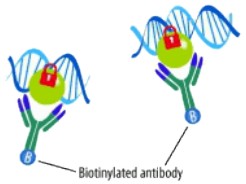
ChIP-seq Protocol



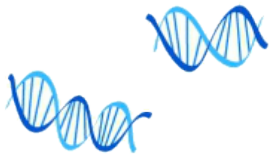
Cross-linked proteins and DNA fragments



Enrichment with antibody pull-down



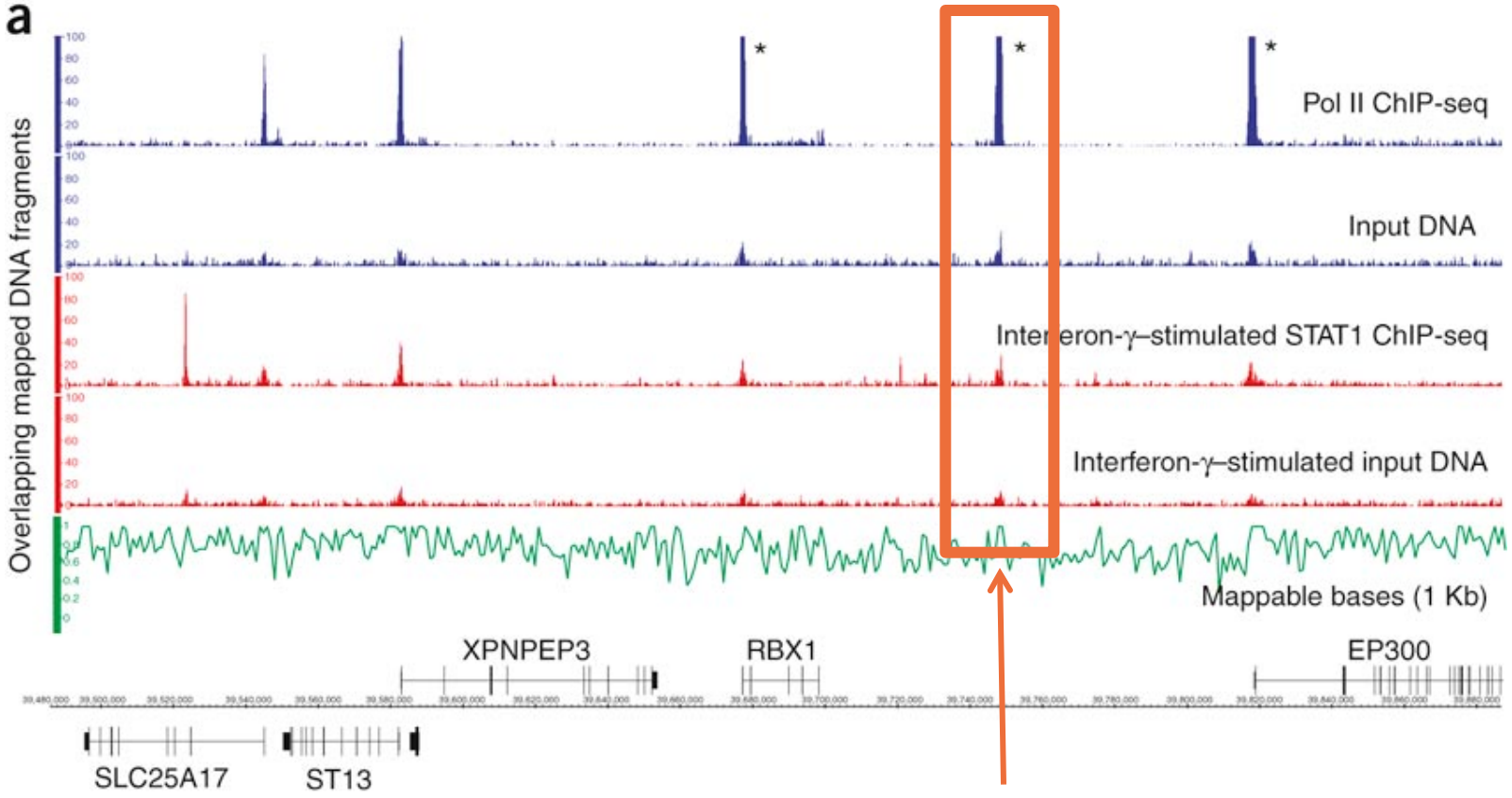
Purified DNA for sequencing



Controls

- as for all experiments it is important that relevant controls be used
  - it is not so clear what those are, and at what level they are useful
- commonly used controls:
  - input (randomly sheared DNA)
  - IgG or GFP
- used to identify anomalies in the genome or artifacts that might be due to reagents, not biology
- argues for a fairly limited set of controls

Do you need controls



Rozowsky et al. 2009

Potential Antibody issues

- there are often multiple antibodies for a particular entity
  - for TP53 there are two widely used ones
- the antibody might not be specific
- you detect direct and indirect interactions with DNA
- cross-linking may occur for spatially proximal proteins that are bound to DNA very far apart in the sequence

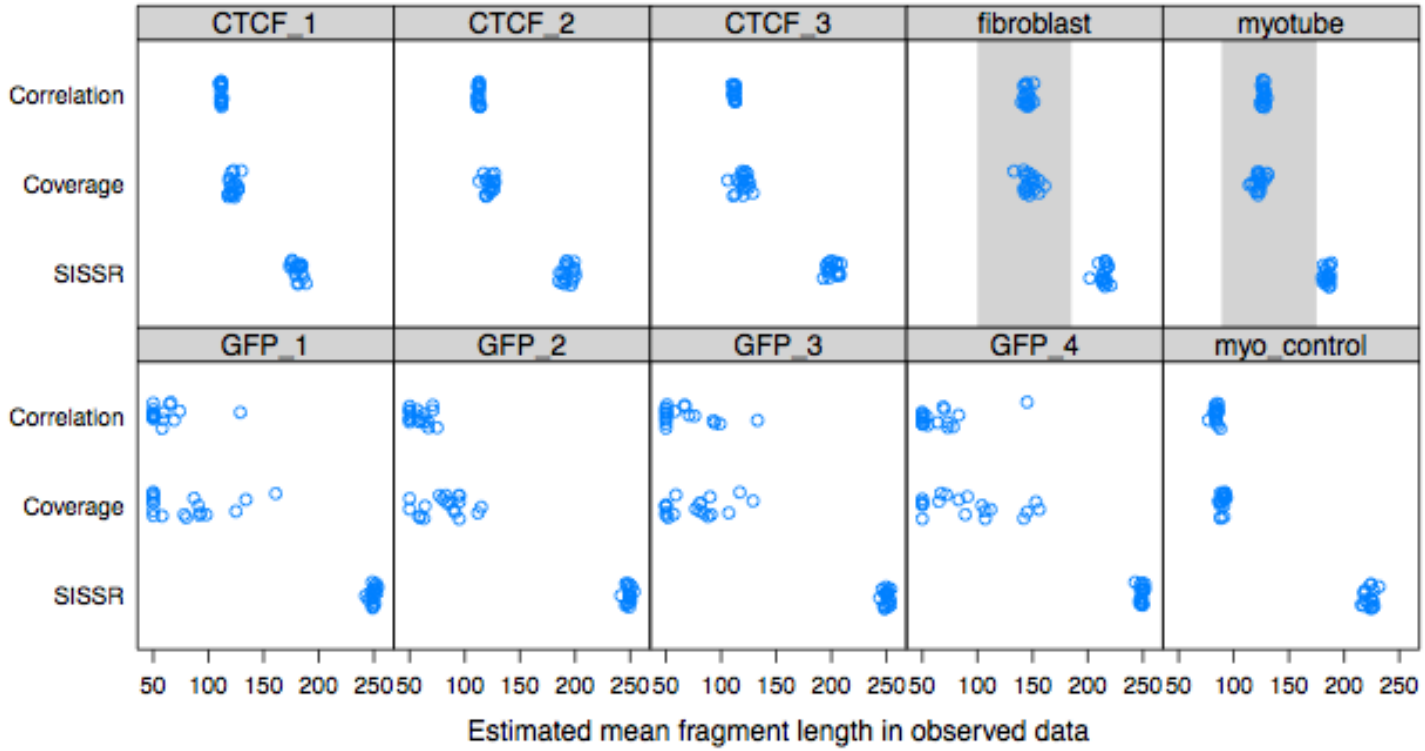
Analysis Pipeline

- QA
- map to genome
  - is our TF associated with a TE?
  - does it like repetitive DNA?
- determine fragment length
- determine foreground/background
- deal with control lane (if present)
- decide if we are looking for peaks or sausages
  - if a TF do we know the PWM (RE)?

Estimating Fragment length

- there are several methods in the literature
  - Kharchenko et al is quite good
  - Jothi et al is quite bad
- our method:
  - choose a lower bound, *w*, for the mean fragment length
  - shift each negative strand read by an amount *u*
  - compute the total number of bases covered by any read
  - find the value u_{min} of *u* for which the number of bases covered is a minimum
  - estimate the mean length by $w + u_{min}$

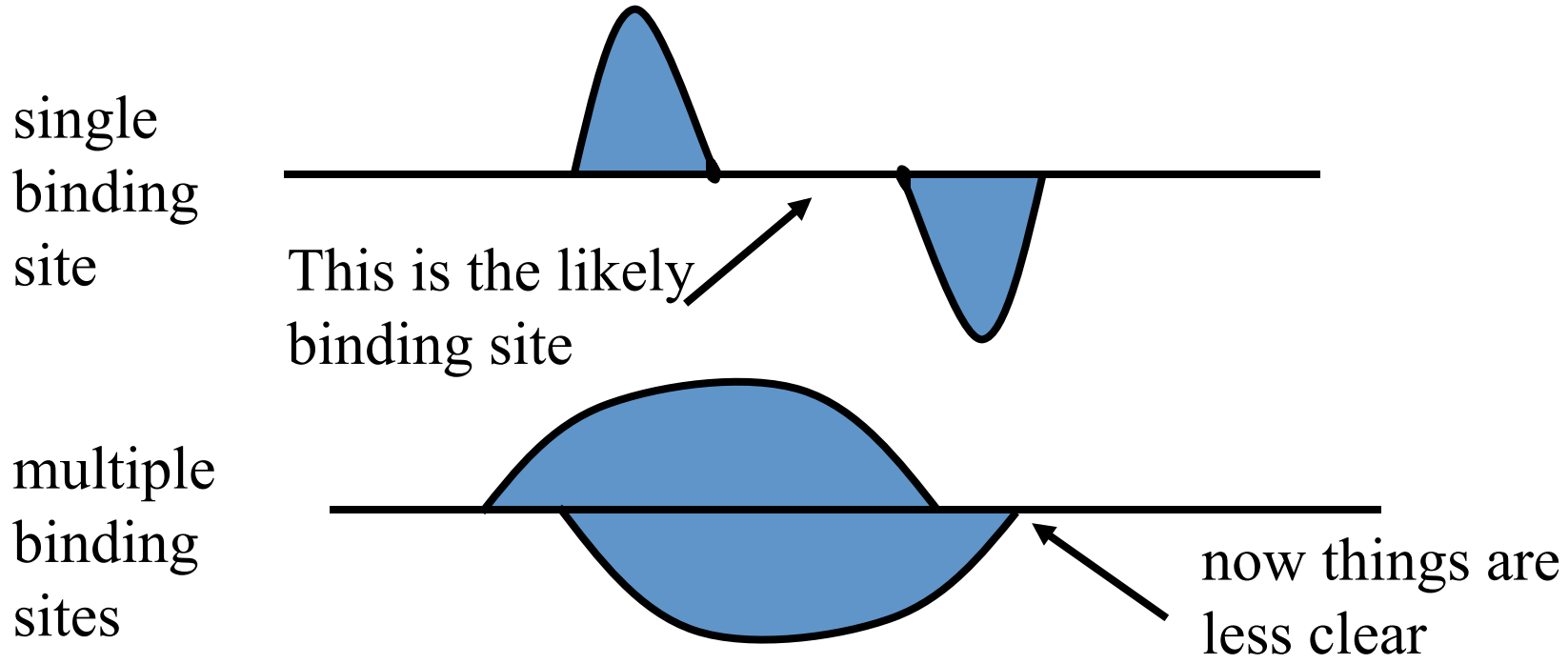
Comparison of methods



- comparison of three methods to estimate mean fragment length

Where did the TF bind?

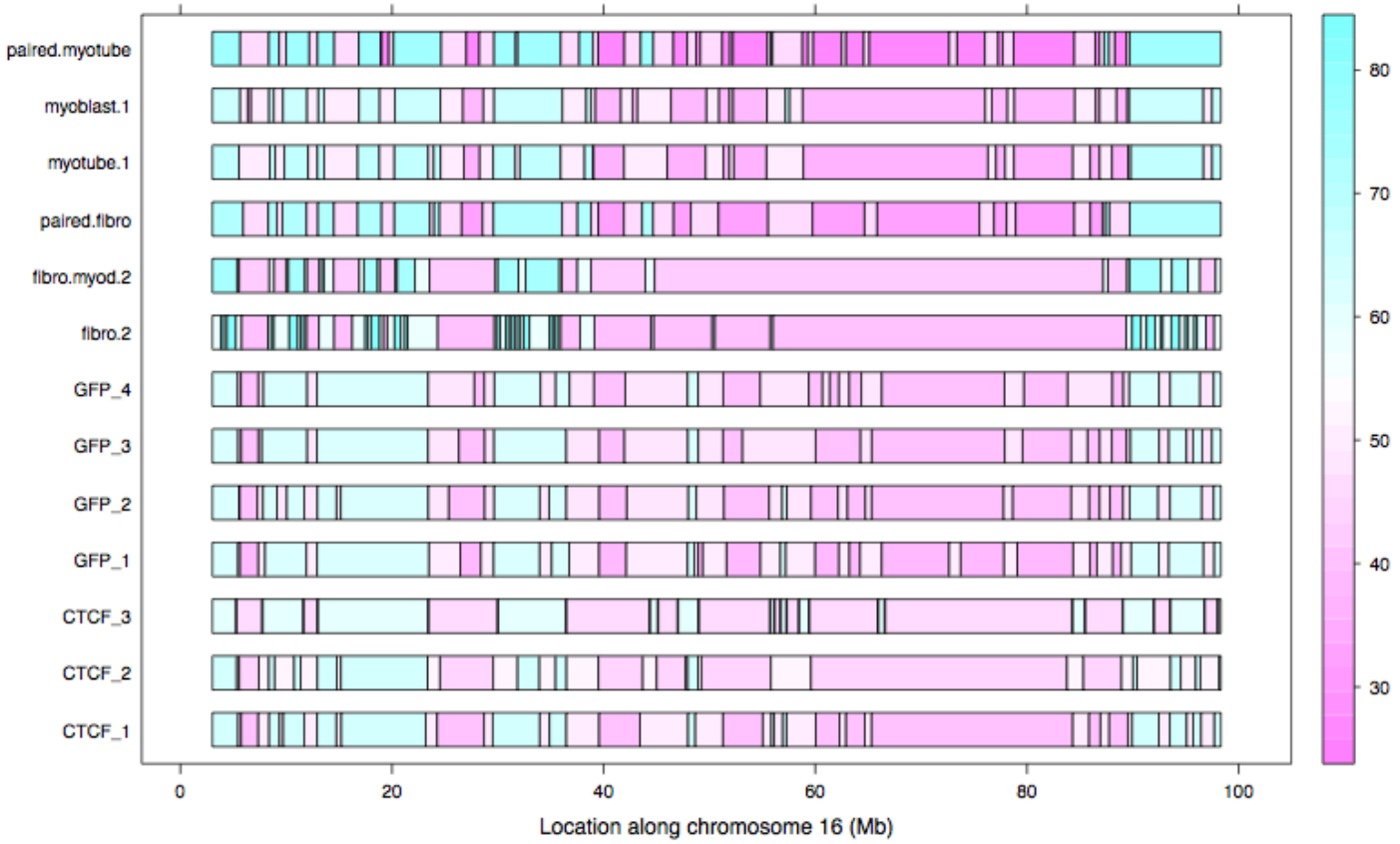
- we should get reads from both the + and - strand
- the reads on the - strand should be upstream of the binding site
- those on the + strand should be downstream



Foreground vs Background

- we observe both reads that correspond to
  - foreground: they represent the binding we are interested in
  - background:low density reads from throughout the genome
- we want to separate these two types of signal
  - the background varies within a genome and between individuals

Background variation



Observed Data

- we exclude (but ultimately won't) reads that map to more than one location
- we exclude reads that map to the same start location and orientation (since in our setting we believe that these are likely due to PCR bias)
  - depth of sequencing is important
- this forces us to think a bit about the *mappable genome:* that part of the genome we could have mapped to
  - so for 36nt reads we want to know how much of the genome is unique

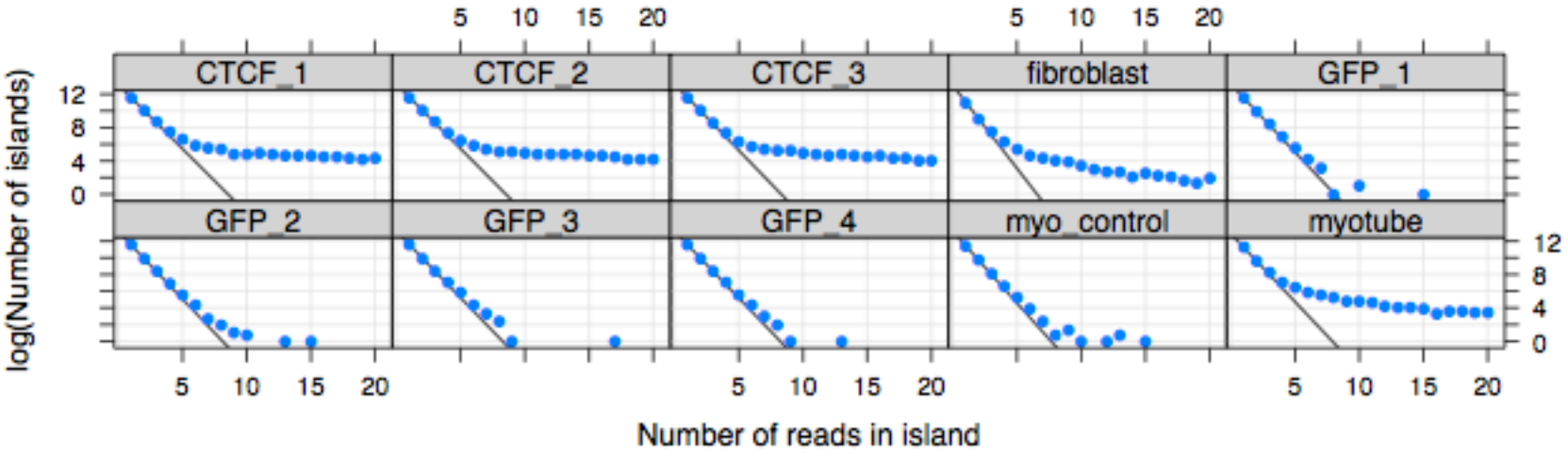
Peak Calling

- null model assumes that reads are distributed uniformly on the genome (Lander and Waterman)
- assumes fragments are length L and let α denote the probability of a new fragment starting at any base
- then the number of reads per island follows a Geometric distribution $P(N=k) = p^{k-1}(1-p)$ where $p=1-(1-\alpha)^L$
- we should only use background reads for this
- we proposed using islands of size 1 or 2 to estimate α

Peak Discovery

- given a Poisson model for the background and an estimate of α we can develop an algorithm that a peak of height k is unlikely given the background
- our original data at the Hutch (small-ish sample sizes) this worked well
- at GNE we found that larger amounts of sequence caused problems and we are developing a Negative Binomial model, with double truncation
  - large values are probably foreground
  - we see some zero inflation

Estimation of the background



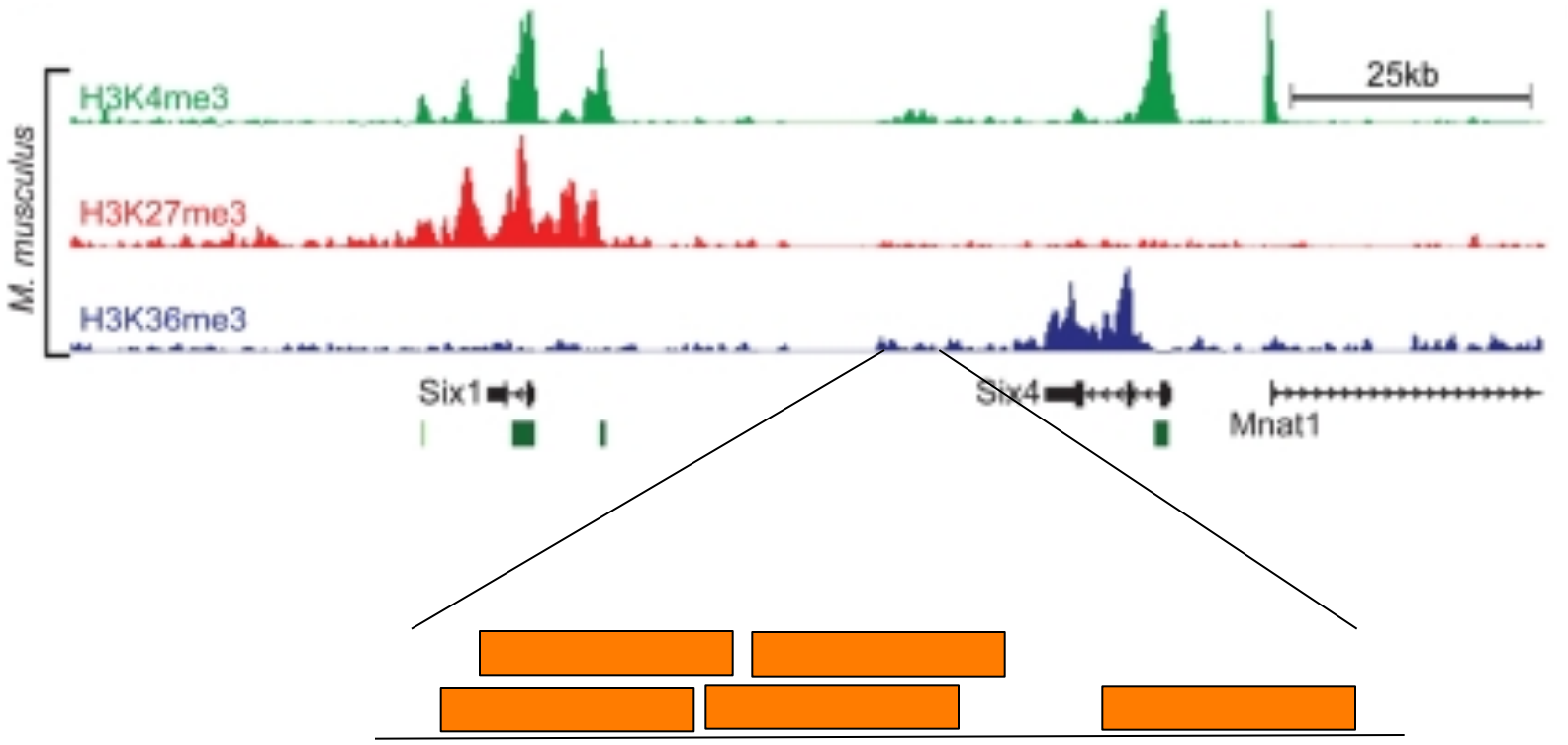
- number of reads per island for Chromosome 1 (mouse)
- black line is an estimate of p, using islands with only one or two reads

Quantifying binding - peak finding

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

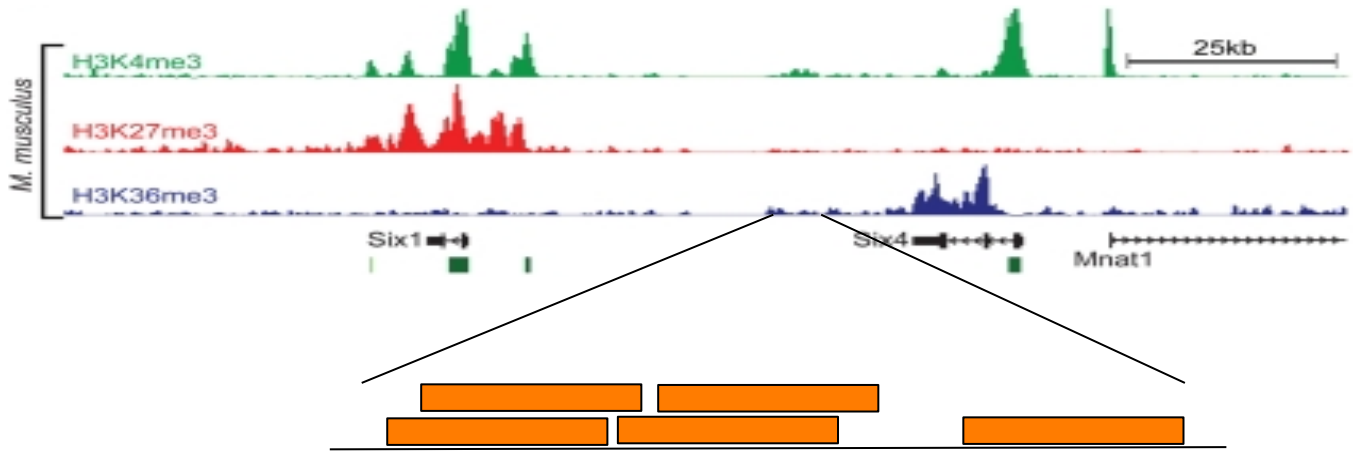
Huge number of algorithms for peak calling out there (> 60)

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

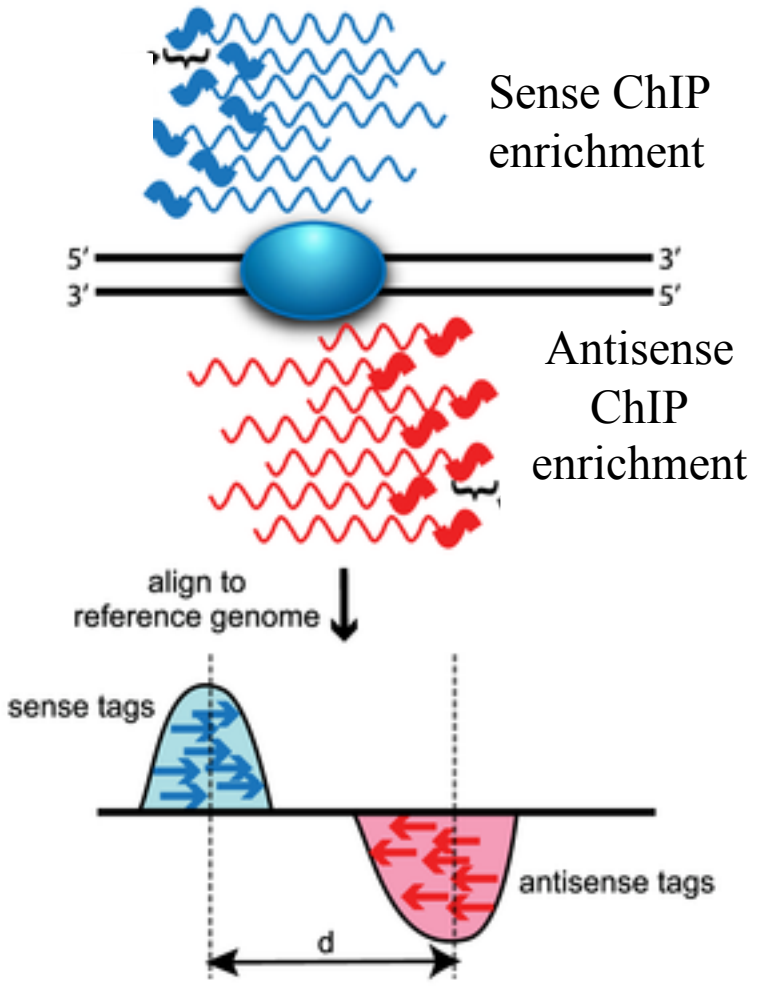
Quantifying binding – peak finding



- 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

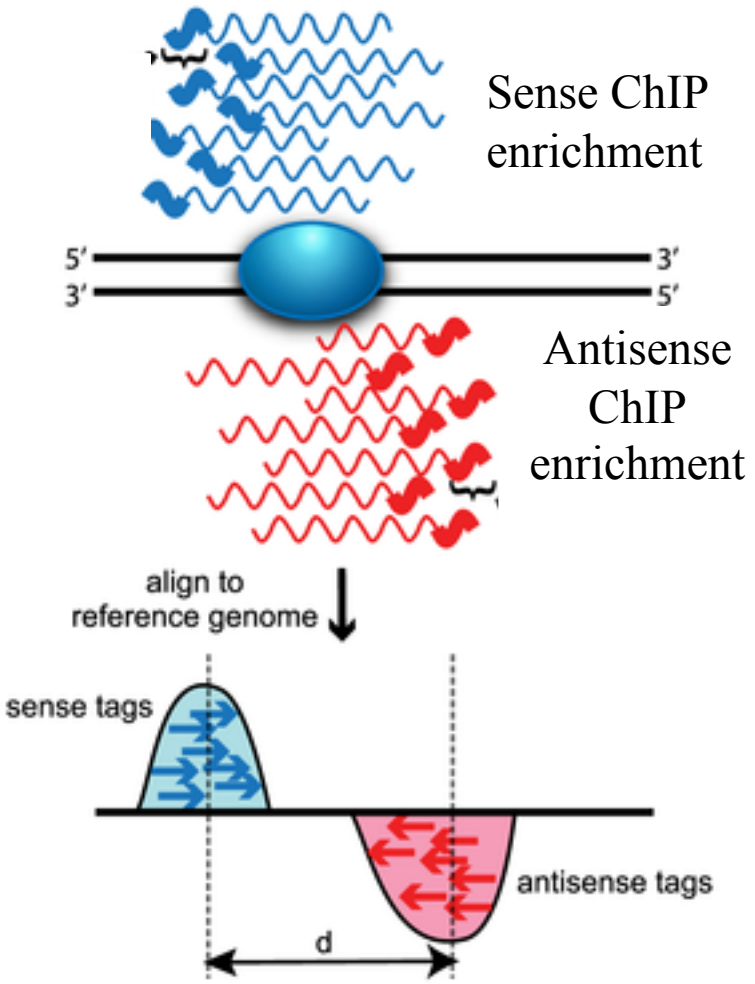
- Another feature that some methods consider is that reads can be from the plus or minus strands
- In this case, for a given TF two peaks will be observed, separated by a constant distance, d
- This can be modeled either post-hoc, or by using strand specific calls



Wilbanks et al., 2010

Quantifying binding – peak finding

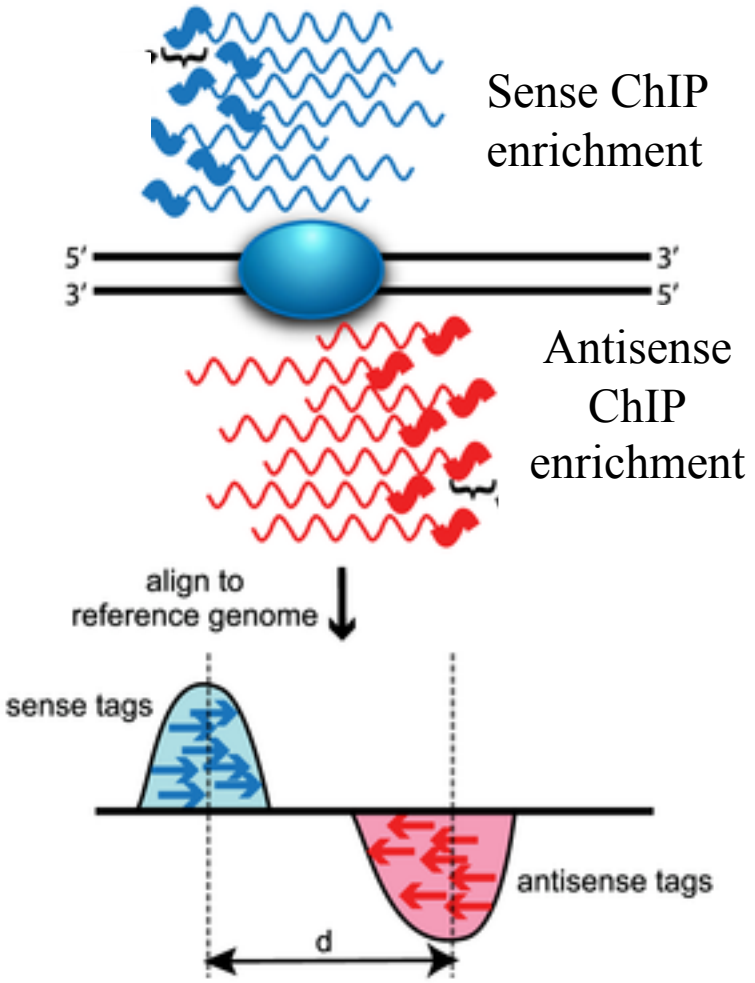
- However, this is only useful where the protein being assayed has a sharp, well defined binding site
- For histone modifications, with broad and sometimes shallow peaks, this information is less useful



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

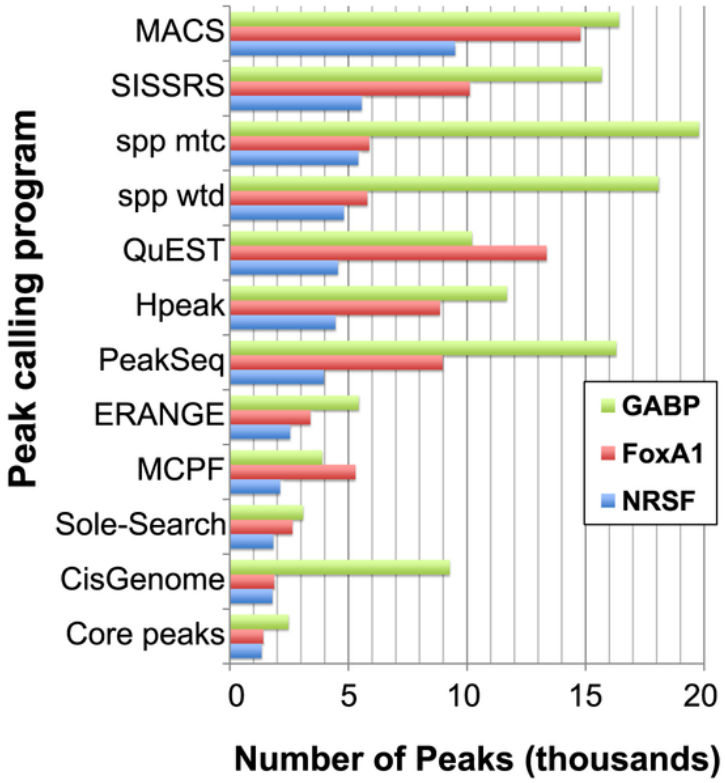
Wilbanks et al., 2010

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

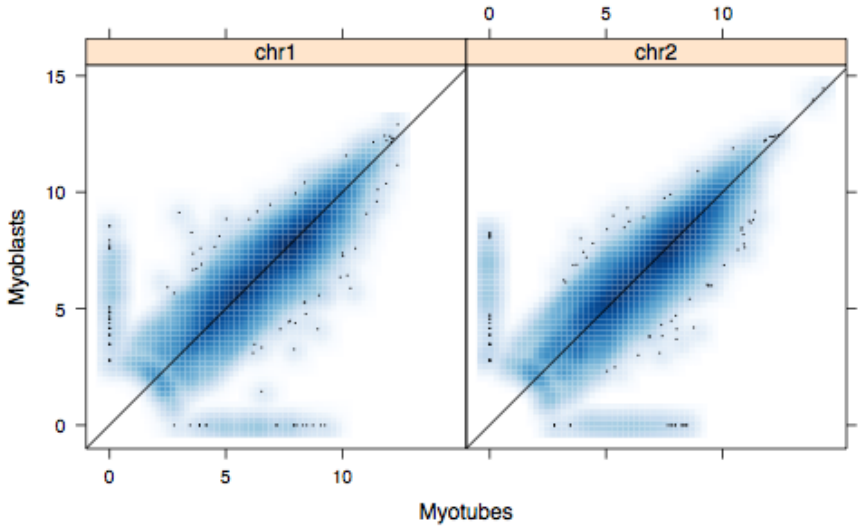
DE peaks

- suppose we have two conditions: Myoblasts and Myotubes
  - potential differences in the genome should be addressed
- we want to know which peaks are differentially expressed (low in one condition and high in the other)
- one could use some cut-off in one condition, and then look for peaks in the other
- instead we combine the data into one collection, choose a fairly relaxed cut-off to define intervals of interest
- use of PWM or RE seems like it should work, but results to date are not strongly positive

DE peaks

- we can then find DE peaks by a number of methods
- a regression approach using DESeq or edgeR seems like it should work
- normalization is an important problem
  - how to deal with different numbers of reads in the different samples

- the vertical and horizontal bars are peaks that were found in only one condition
- otherwise points far from the line are candidates for DE



Found our peaks - what next

- once we have decided what things are peaks we next need to try and interpret them
- typically that involves putting them in some form of genomic context
- IRanges/rtracklayer etc can help

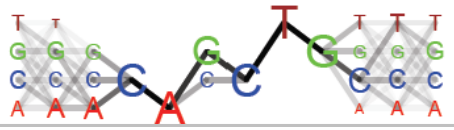
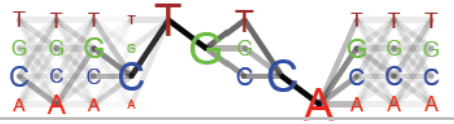
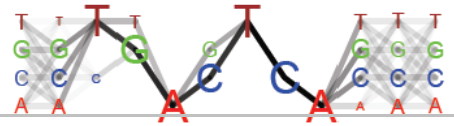
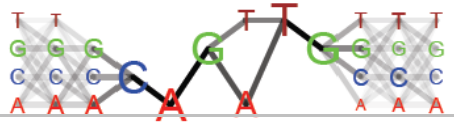
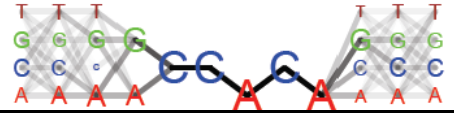
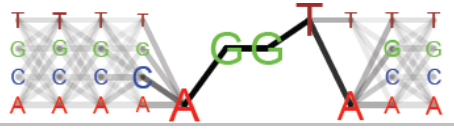
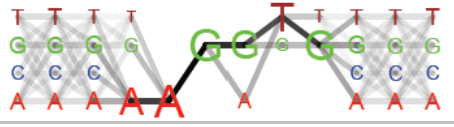
Peak Summary (cut-off 8)

	All	Up	Down	Ratio
Total	145970	8779	8861	1.009
Promoter	24887	1185	559	0.472
3’	4000	225	274	1.218
Upstream	30983	1836	1663	0.906
Downstream	30073	1795	1833	1.021
Gene	78689	4570	4738	1.037

Motif finding

- can we detect and understand motifs under the binding sites and do they differ in different contexts?
- are co-factors or other variables involved?
- MEME – now DREME are possible choices
  - MEME is low throughput
  - DREME is new (I have not tried it)
- an essential part of motif finding is to select an appropriate set of sequences to compare against
  - we try to find something similar, nearby

Myotube peaks vs. Control

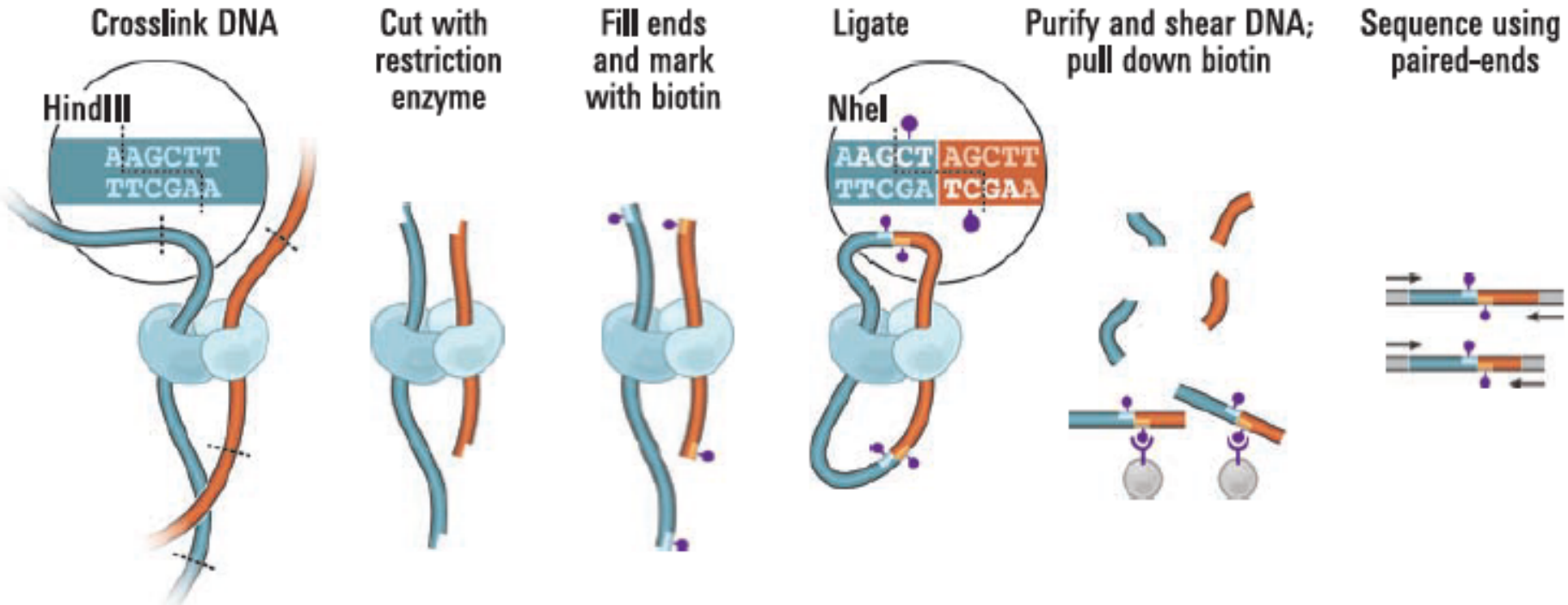
	Score	S/N	%fg	%bg	Logo
NNVCAGCTGNNN (Ebox)	-108.1	11.09	0.86	0.13	
NNNCTGBCANNN (Meis)	-40.6	1.76	0.53	0.36	
NNTGACTCANNN (c-Fos)	-32.5	2.70	0.13	0.06	
NNVCAGATGNNN (Ebox)	-27.3	1.90	0.19	0.11	
NNDRCCACANNN (Runx1)	-27.0	1.87	0.24	0.14	
NNNNAGGTANN	-40.2	0.40	0.16	0.35	
NNNAAGGTGNNN	-38.7	0.66	0.32	0.44	

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

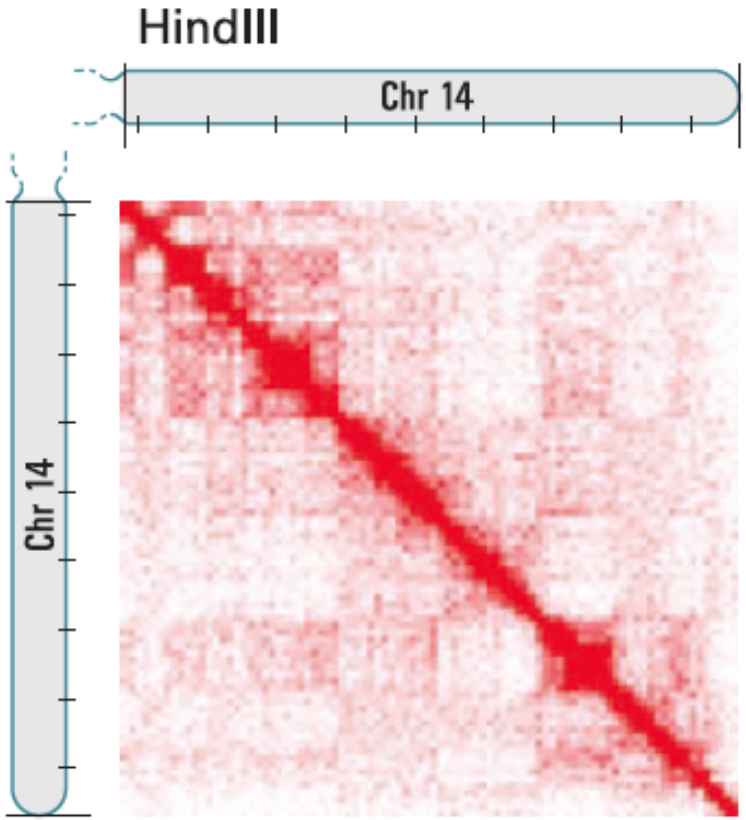
- Next-generation sequencing techniques (Hi-C) can enable us to study these interactions genome-wide



Lieberman-Aiden et al., 2009

Chromatin conformation

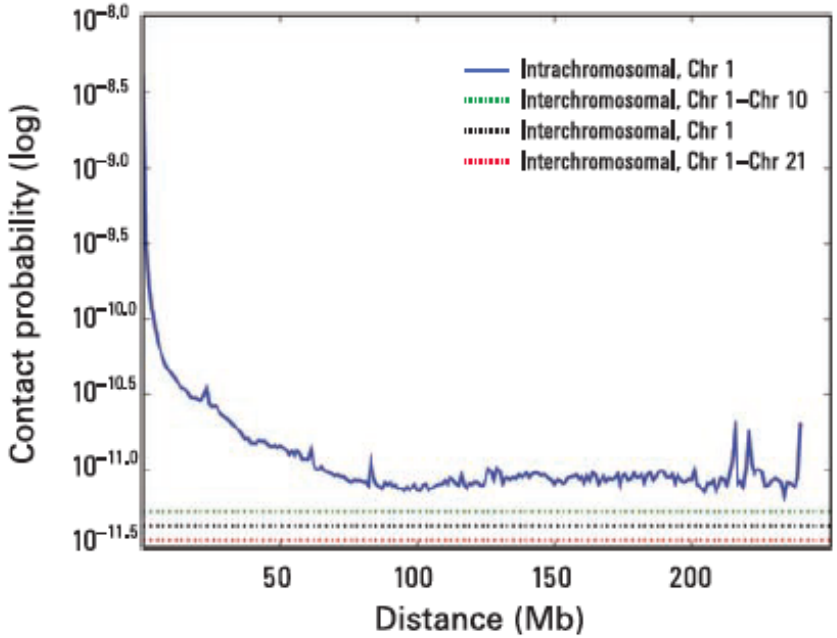
- Liberman-Aiden et al., applied this method to a CEU cell line
- They divided the genome into 1Mb windows and counted the number of reads, m_{ij} that linked window i to window j



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

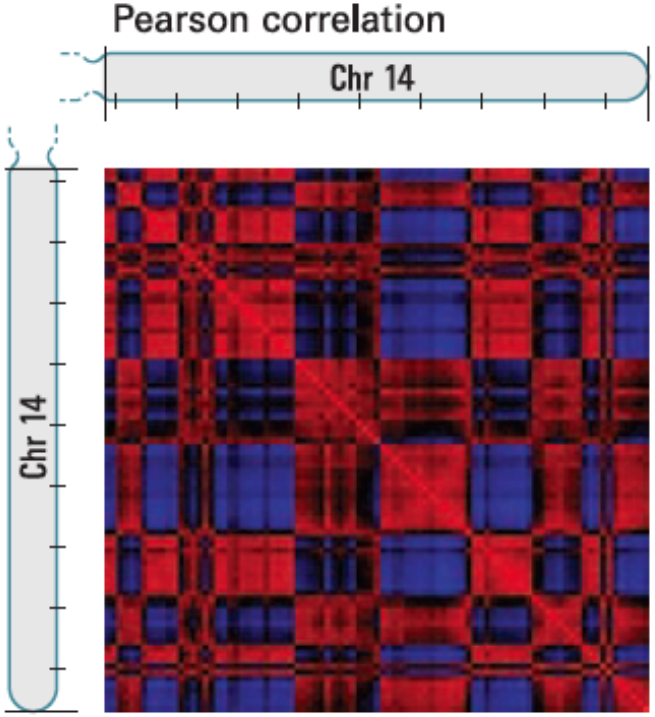


This analysis confirmed previous work suggesting there were well defined chromosomal domains

Chromatin conformation

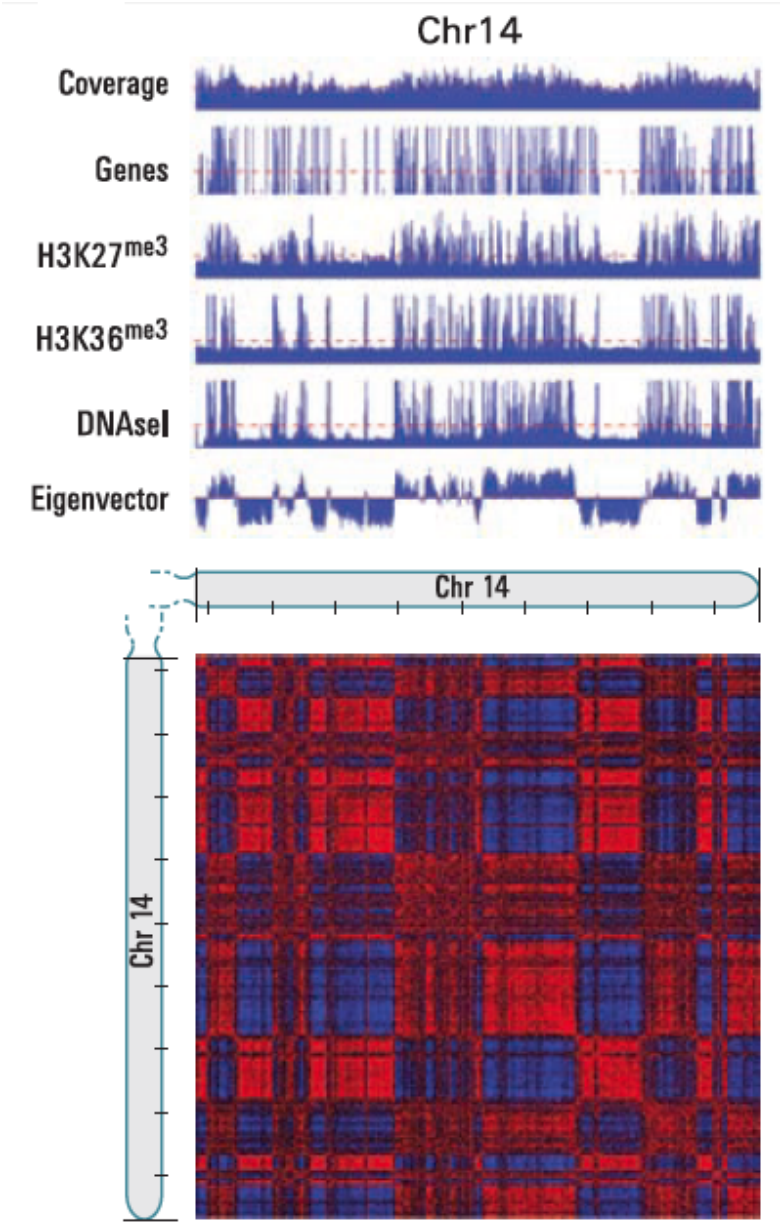
- Since the highest contacts are observed for regions that are located near one another (artefactual?), Lieberman-Aiden et al. normalized the data to account for this

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



Acknowledgements

- D. Sarkar
- S. Tapscott
- Y. Cao
- Z. Yao
- M. Lawrence
- L. Ruzzo
- W. Huber
- M. Morgan
- R. Bourgon
- J. Degenhardt