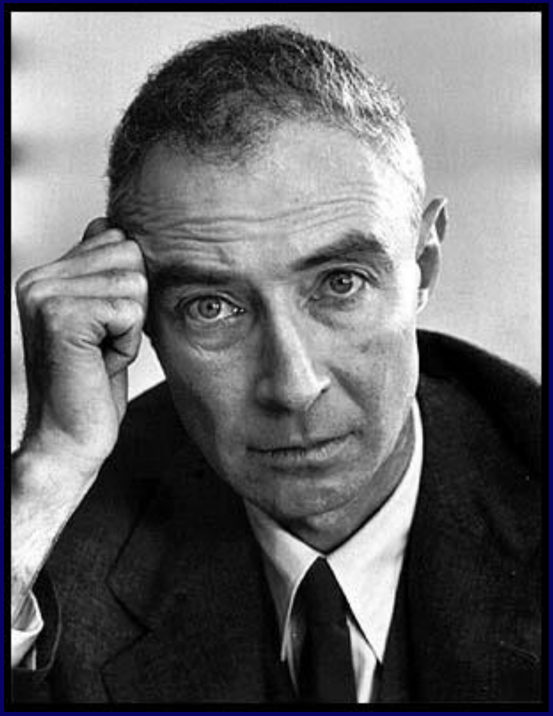
Using model-based methods to analyze NGS data

Steve Qin
Department of Biostatistics
and Bioinformatics
Rollins School of Public Health
Emory University

EMORY
UNIVERSITY

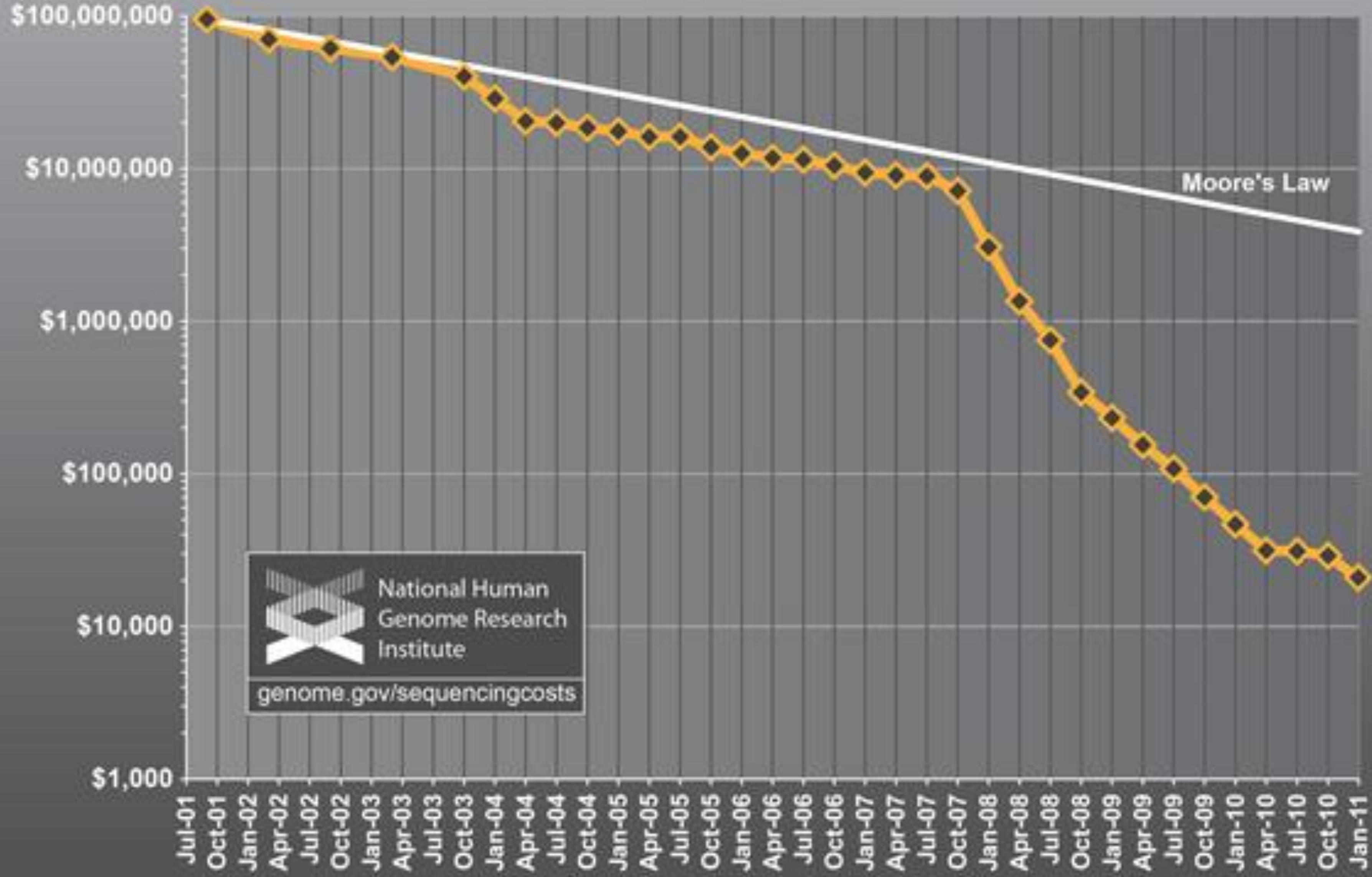


"… deep things in science are not found because they are useful; they are found because it was possible to find them"

-- Robert Oppenheimer

Next generation sequencing technologies

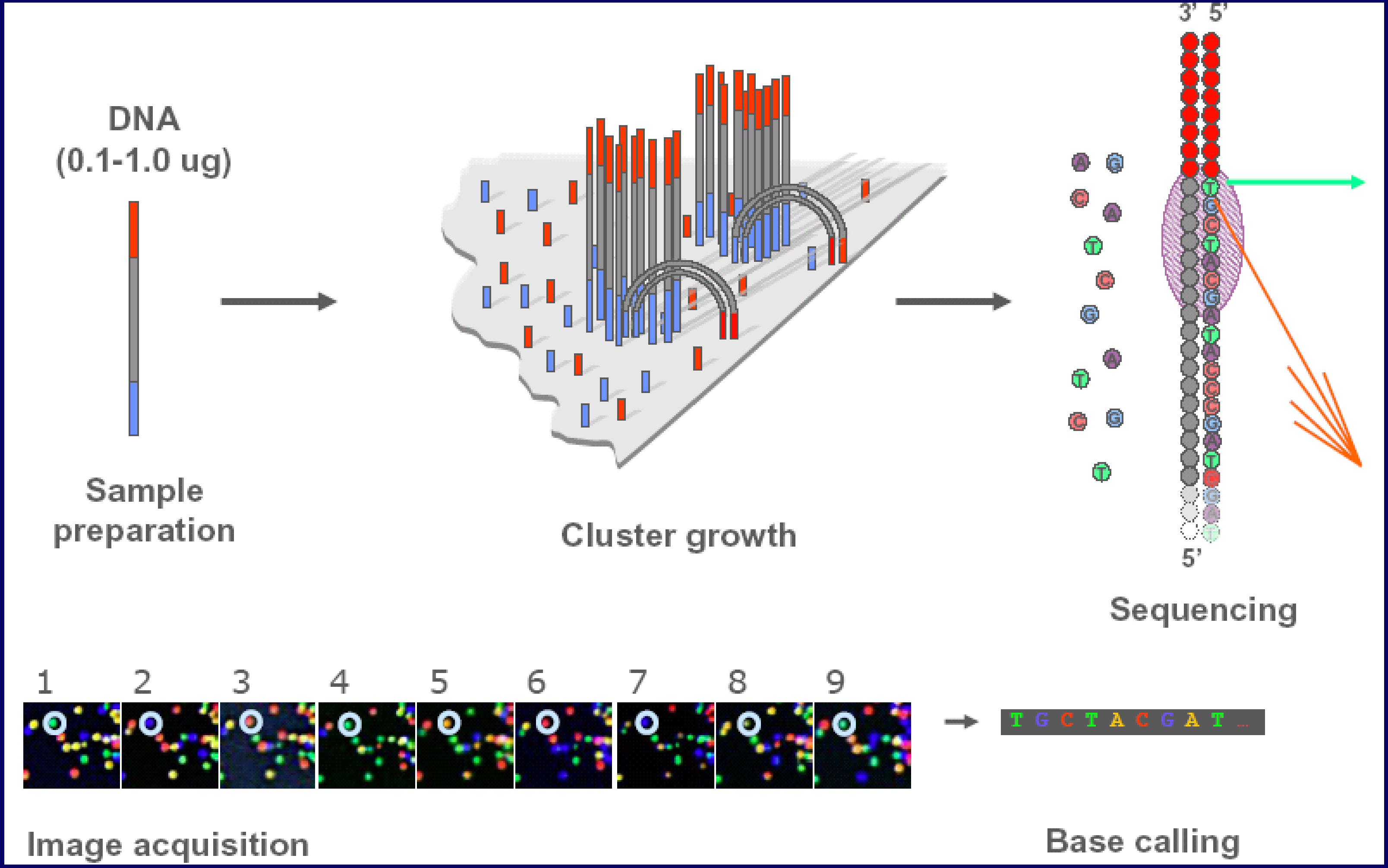
Cost per Genome



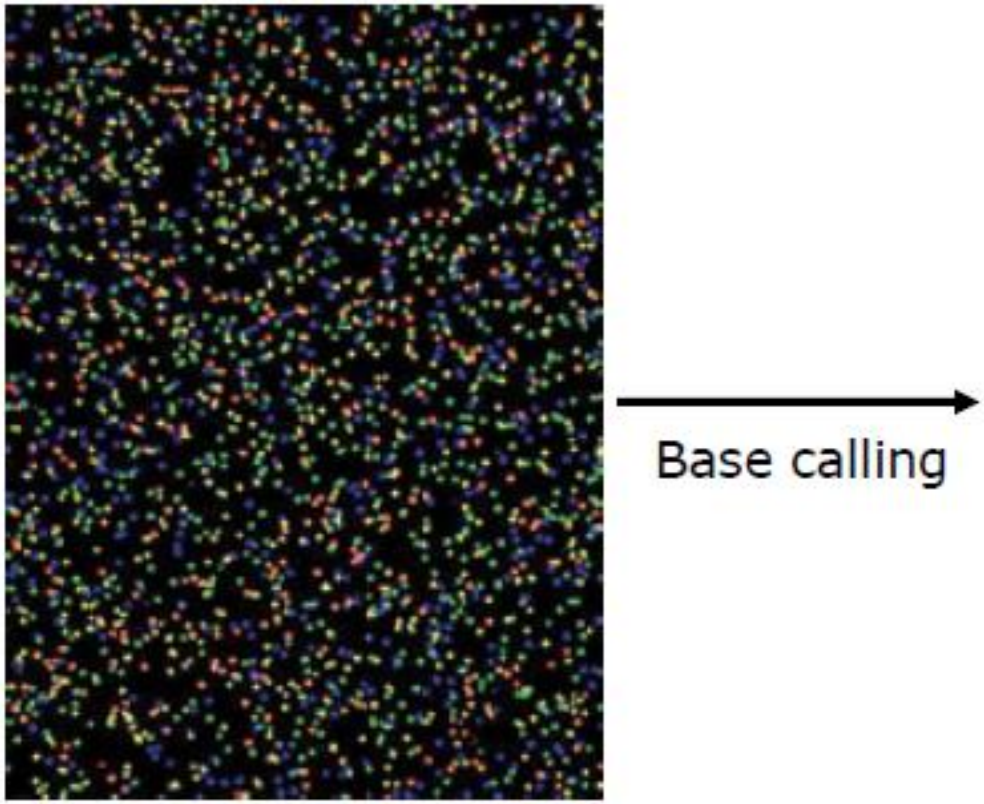
National Human Genome Research Institute

genome.gov/sequencingcosts

Illumina sequencing technology

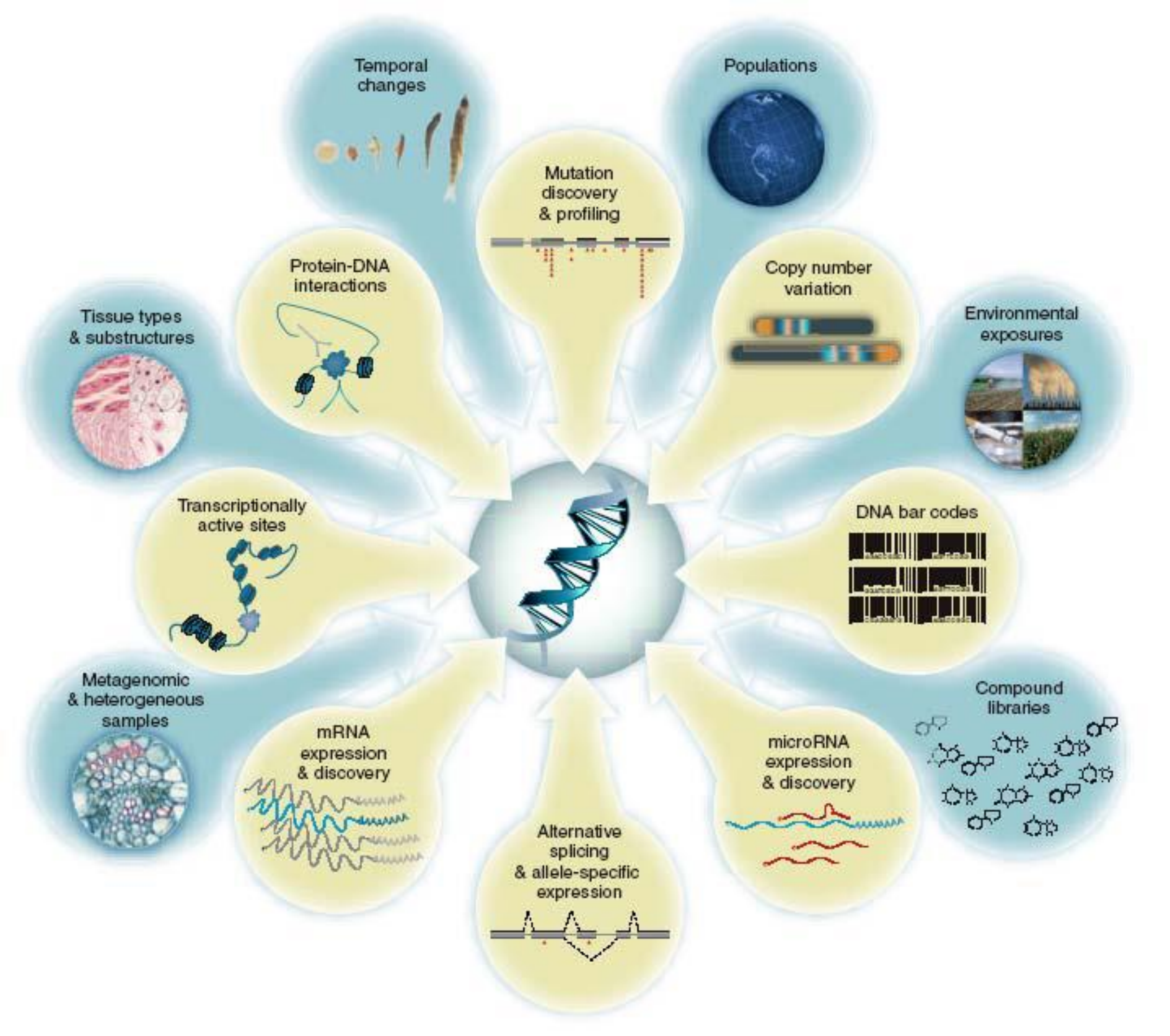


Sequence data



```
AAAAATCTCTTCCTGAACCATTCAGAAAATGC
AAAAAAAAAAAAAAAAAAAAAAAAAAAAAAAA
AACAGACCTAAAATCGCTCATTGCATATTCTT
AACCAGGCGACCTGCGACTCCTTGACGTTGAC
ATGTTAGGGTTGTACGGTAGAACTCCTATTAT
ATTGCCCAGAAAGTACCTGAGCTATCAGTGAT
ATCCCGATCCCGGTTACAGAGTCCATTGTAGA
ACCACCCAACAATGACTAATCAAACTAACCTC
ATGGGGAAATATTGCAATTATGTAAAGGTAA
ATGTTTAAAAGTCCACTTTTAAAACTATATTT
ATATAACTCTCTTCCCTCTCACTCTTTCTCTC
AGGGAACTACTCCCACCCTGGAGCCTCCGTAG
AAAAGATATATATATATATATATTCATAATTA
AGTCGACCCTGCACCTGGTCCTGCGTCTGAGA
ATTTGGTGAGTAATTAAAGAGAGTAGTAGCAT
GGTCTGTTTGTCGTATGCCGTCTTCTTCTTTT
ATTGAAAGAAGTCTTTCTAGAAATGTTAAATA
AGGGACTGAAGCTGCTGGGGCCATGTTTTTAG
AGAAAATATTAAAATCTTTGAAGAAGAAGAAG
AAGGGGATTTAGAGGGTTCTGCGGGCAAATTT
AGAACCCTCCATAAACCTGGAGTGACTATATG
AATAAGTCGGTTCAGGAGATCCAAGGAACCTT
ATTGGGTTTGGCTGTATCCCACCCCGTTACAA
CGGGGATAAGTGTGGTTTCGAAGAAGATATAA
```

Applications of NGS


Temporal changes
Populations
Mutation discovery & profiling
Protein-DNA interactions
Copy number variation
Tissue types & substructures
Environmental exposures
Transcriptionally active sites
DNA bar codes
Metagenomic & heterogeneous samples
Compound libraries
mRNA expression & discovery
microRNA expression & discovery
Alternative splicing & allele-specific expression

Kahvejian, A., *et al (2008) Nature Biotechnology, 26(10)*

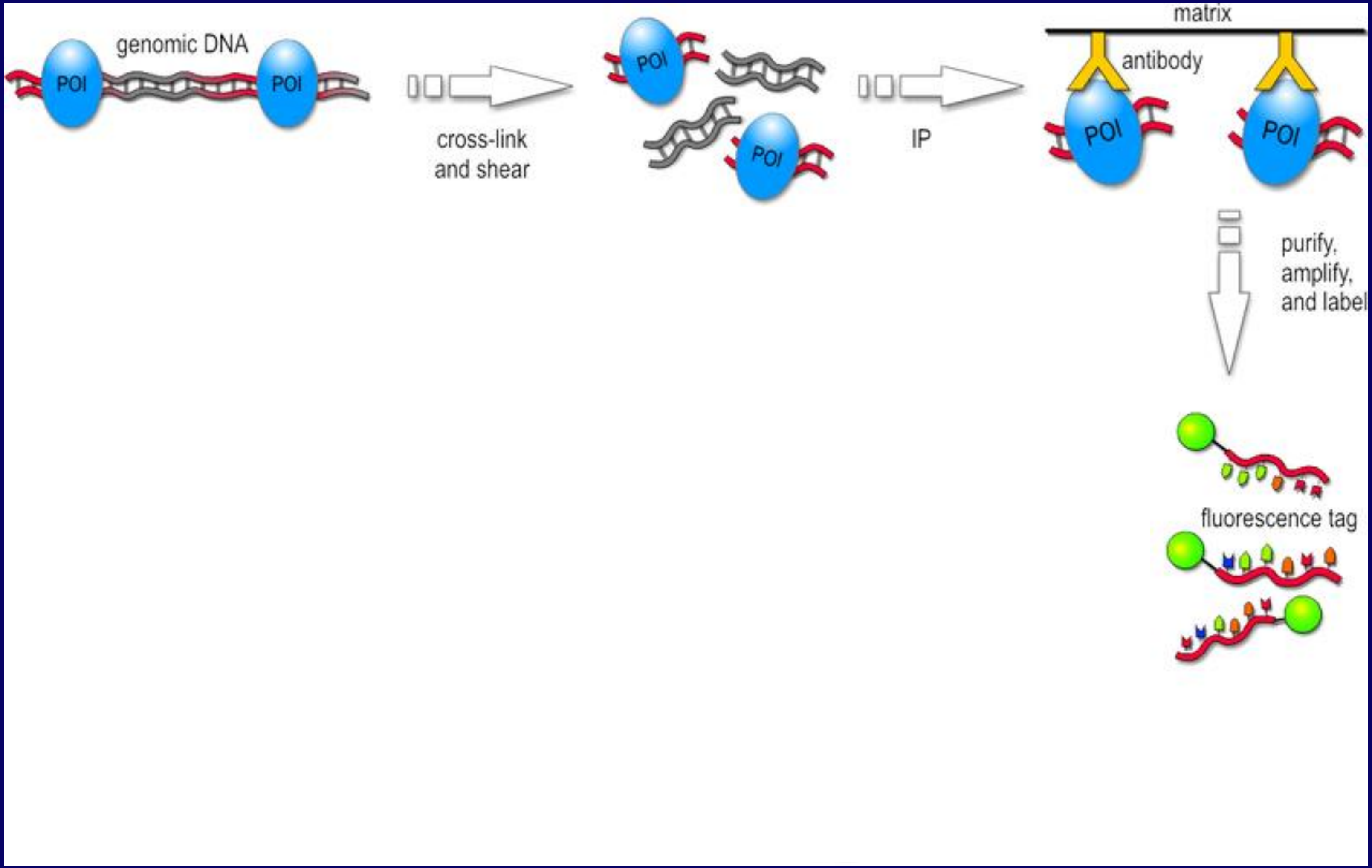
So many –seq, so little time

ALEXA-Seq, Apopto-Seq, AutoMeDip-Seq, Bind-n-Seq, Bisulfite-Seq, ChIP-Seq, ClIP-Seq, CNV-Seq, DGE-Seq, DNA-Seq, DNase-Seq, F-Seq, FRT-Seq, HITS-CLIP, indel-Seq, MBD-Seq, MeDIP-Seq, MethylCap-Seq, microRNA-Seq, mRNA-Seq, NA-Seq, NSR-Seq, PAS-Seq, Peak-Seq, ReChIP-Seq, RIP-Seq, RNA-Seq, rSW-Seq, SAGE-Seq, Sono-Seq, Tn-Seq…

So many –seq, so little time

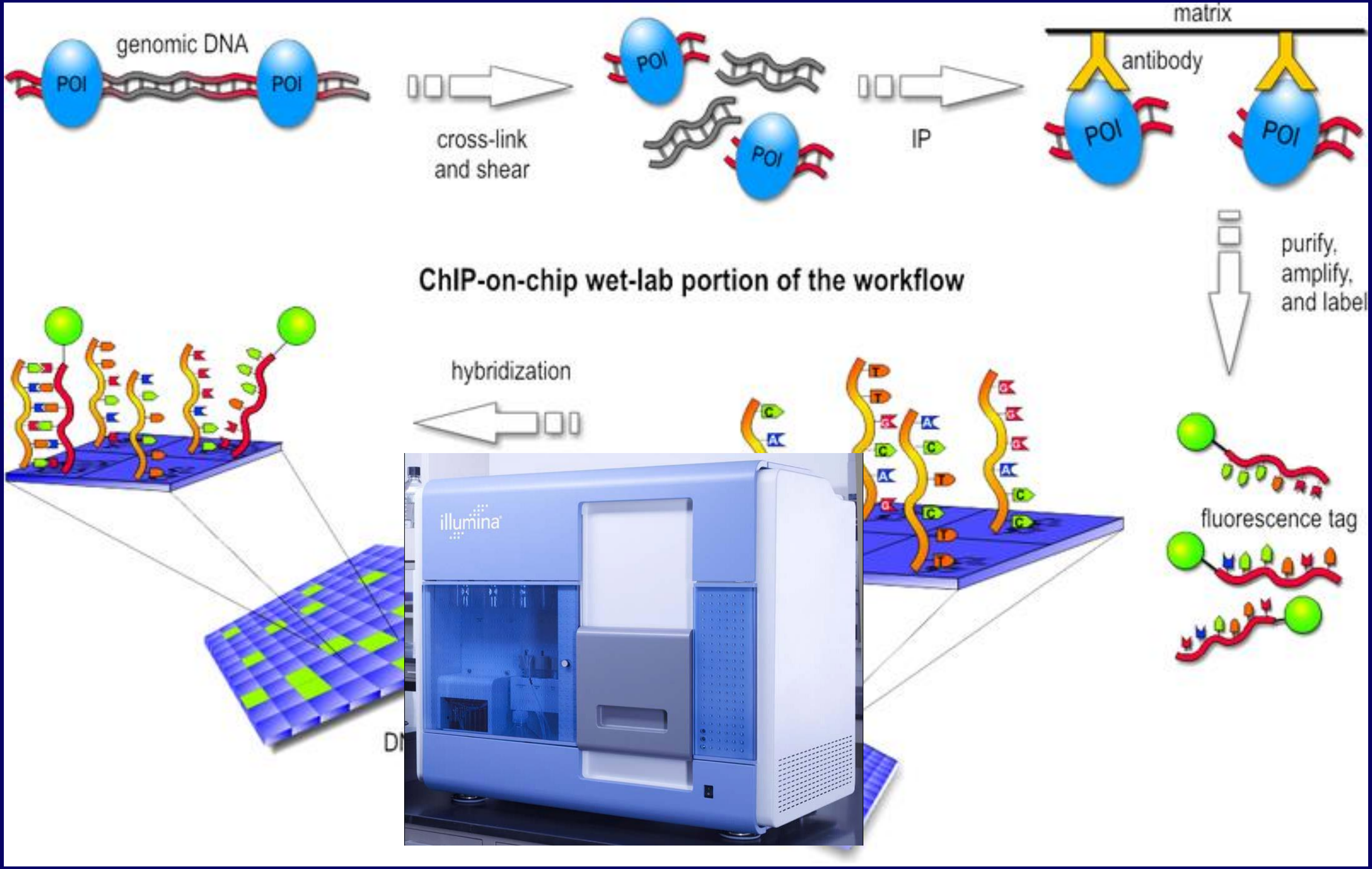
ALEXA-Seq, Apopto-Seq, AutoMeDip-Seq, Bind-n-Seq, Bisulfite-Seq, **ChIP-Seq**, ClIP-Seq, CNV-Seq, DGE-Seq, DNA-Seq, DNase-Seq, F-Seq, FRT-Seq, HITS-CLIP, indel-Seq, MBD-Seq, MeDIP-Seq, MethylCap-Seq, microRNA-Seq, mRNA-Seq, NA-Seq, NSR-Seq, PAS-Seq, Peak-Seq, ReChIP-Seq, RIP-Seq, **RNA-Seq**, rSW-Seq, SAGE-Seq, Sono-Seq, Tn-Seq…

Chromatin Immunoprecipitation



ChIP-chip on Wikipedia

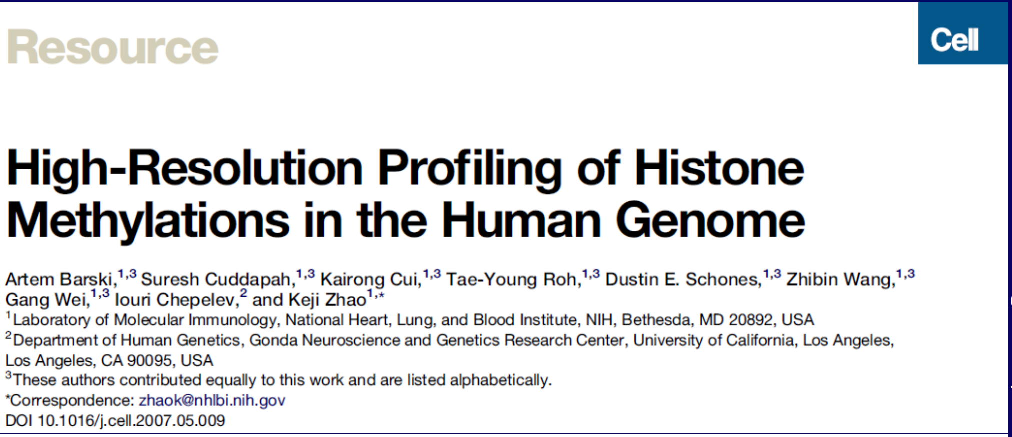
ChIP-chip and ChIP-Seq technologies



Ren et al. 1999; Iyer *et al.* 2000

ChIP-chip on

ChIP sequencing



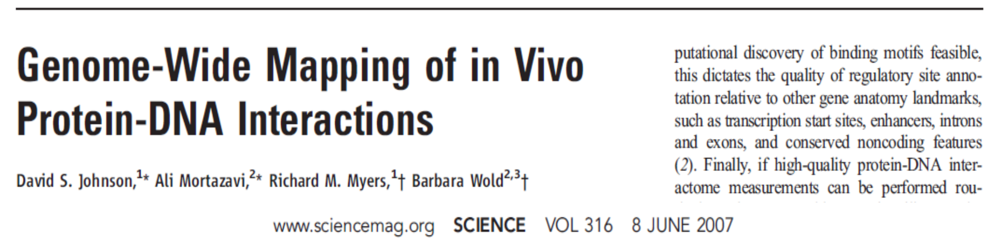
Resource

Cell

High-Resolution Profiling of Histone Methylations in the Human Genome

Artem Barski,[1,3] Suresh Cuddapah,[1,3] Kairong Cui,[1,3] Tae-Young Roh,[1,3] Dustin E. Schones,[1,3] Zhibin Wang,[1,3]
Gang Wei,[1,3] Iouri Chepelev,[2] and Keji Zhao[1,*]
[1]Laboratory of Molecular Immunology, National Heart, Lung, and Blood Institute, NIH, Bethesda, MD 20892, USA
[2]Department of Human Genetics, Gonda Neuroscience and Genetics Research Center, University of California, Los Angeles,
Los Angeles, CA 90095, USA
[3]These authors contributed equally to this work and are listed alphabetically.
*Correspondence: zhaok@nhlbi.nih.gov

DOI 10.1016/j.cell.2007.05.009

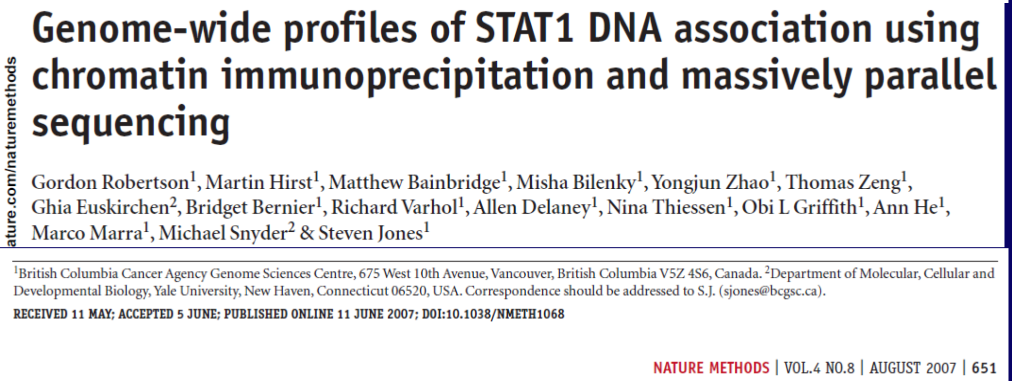


Genome-Wide Mapping of in Vivo Protein-DNA Interactions

David S. Johnson,[1*] Ali Mortazavi,[2*] Richard M. Myers,[1†] Barbara Wold[2,3†]

putational discovery of binding motifs feasible, this dictates the quality of regulatory site annotation relative to other gene anatomy landmarks, such as transcription start sites, enhancers, introns and exons, and conserved noncoding features (2). Finally, if high-quality protein-DNA interactome measurements can be performed rou-

www.sciencemag.org SCIENCE VOL 316 8 JUNE 2007



ature.com/naturemethods

Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing

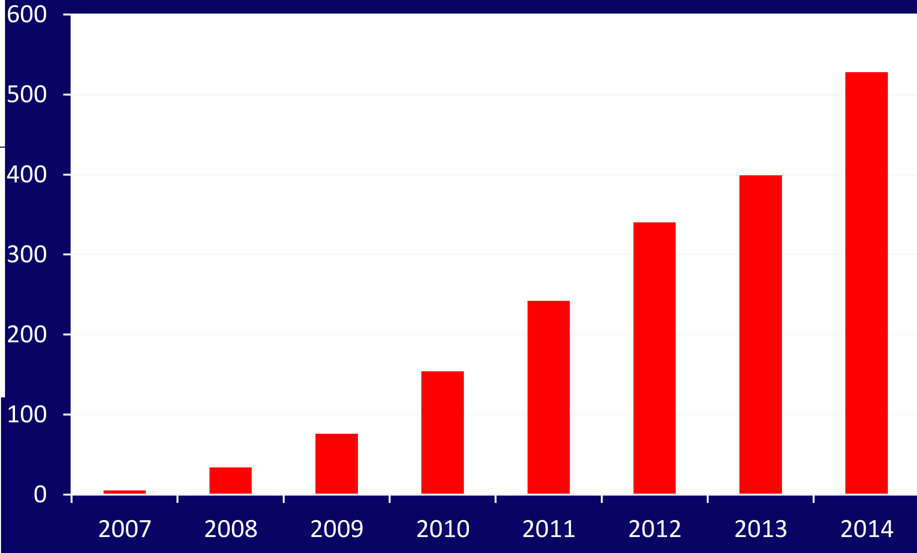
Gordon Robertson[1], Martin Hirst[1], Matthew Bainbridge[1], Misha Bilenky[1], Yongjun Zhao[1], Thomas Zeng[1], Ghia Euskirchen[2], Bridget Bernier[1], Richard Varhol[1], Allen Delaney[1], Nina Thiessen[1], Obi L Griffith[1], Ann He[1], Marco Marra[1], Michael Snyder[2] & Steven Jones[1]

[1]British Columbia Cancer Agency Genome Sciences Centre, 675 West 10th Avenue, Vancouver, British Columbia V5Z 4S6, Canada. [2]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut 06520, USA. Correspondence should be addressed to S.J. (sjones@bcgsc.ca).

RECEIVED 11 MAY; ACCEPTED 5 JUNE; PUBLISHED ONLINE 11 JUNE 2007; DOI:10.1038/NMETH1068

NATURE METHODS | VOL.4 NO.8 | AUGUST 2007 | 651

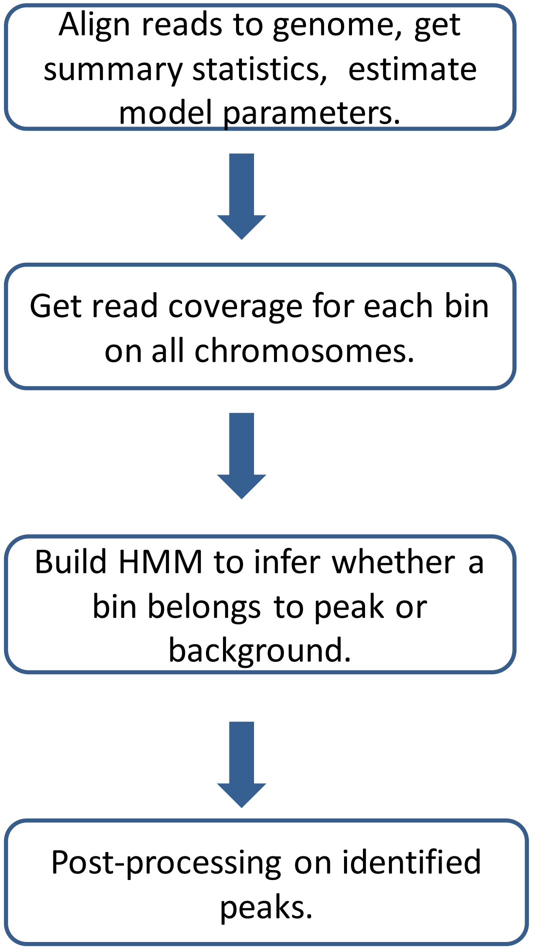
ChIP-Seq papers



Outline

- Hidden Markov model for peak detection
- Hierarchical Hidden Markov model for combining ChIP-seq and ChIP-chip data
- Hybrid Monte Carlo strategy for Motif finding

HPeak algorithm

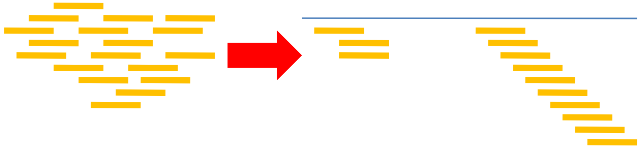


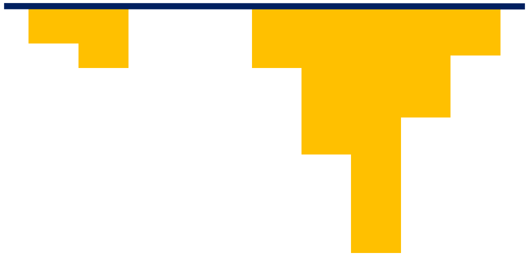
Align reads to genome, get summary statistics, estimate model parameters.

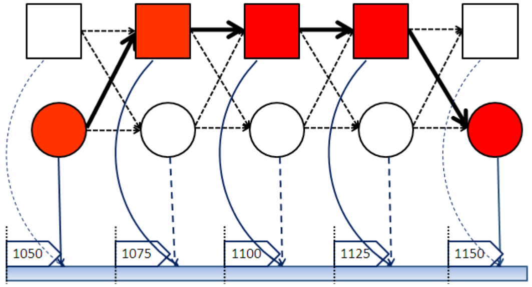
Get read coverage for each bin on all chromosomes.

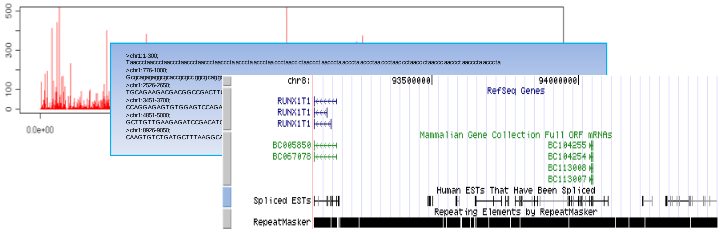
Build HMM to infer whether a bin belongs to peak or background.

Post-processing on identified peaks.



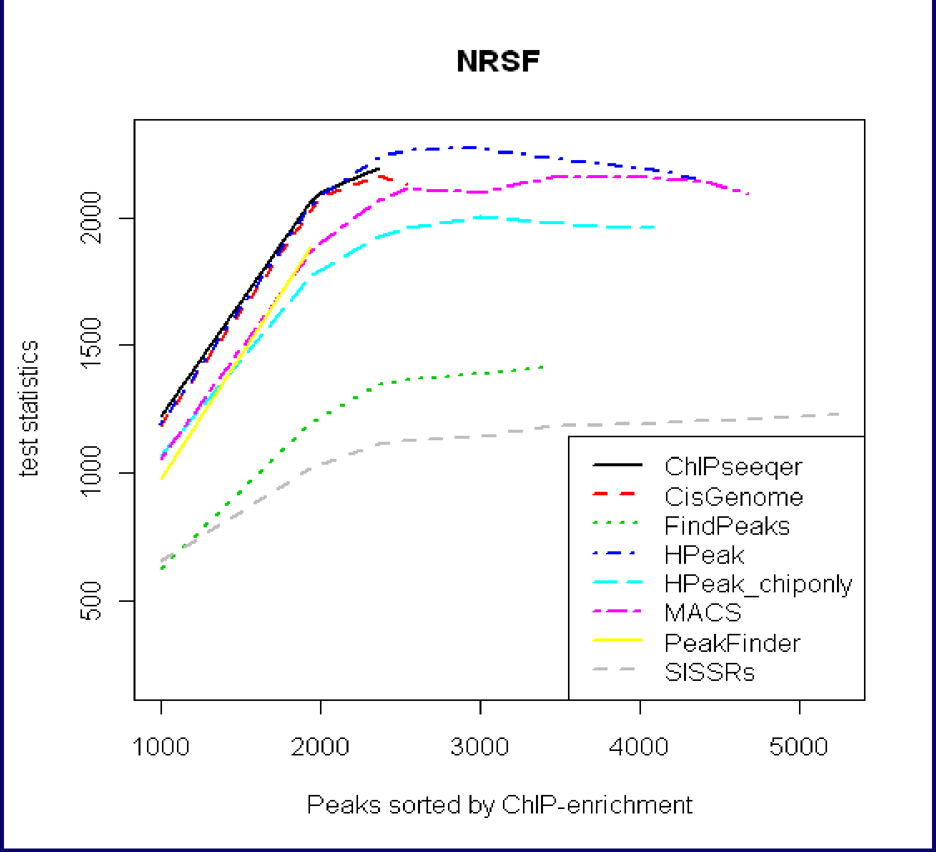


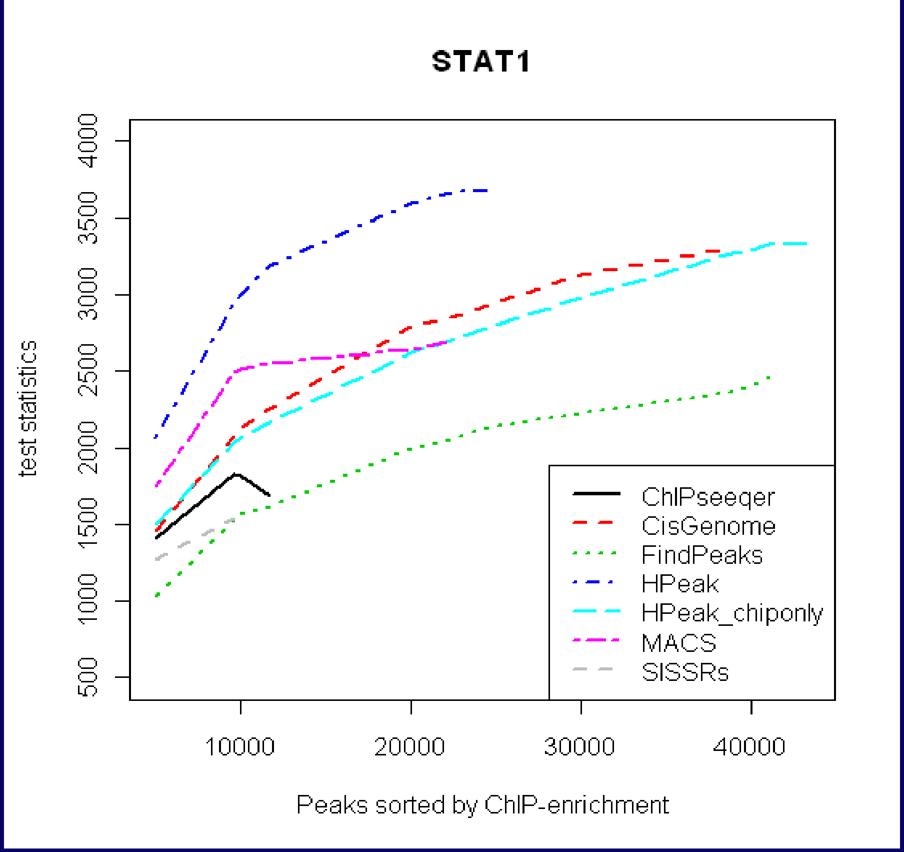




```
>chr1:1-300;
Taaccctaaccctaaccctaaccctaaccctaaccctaacccta acccta accctaa ccctaacc ctaaccct aaccctaacccta accctaa ccctaac cctaacc ctaaccc aaccct aaccctaacccta
>chr1:776-1000;
Gcgcagagaggcgcaccgcgcc ggcg cagg
>chr1:2526-2650;
TGCAGAAGACGACGGCCGACTT
>chr1:3451-3700;
CCAGGAGAGTGTGGAGTCCAGA
>chr1:4851-5000;
GCTTGTTGAAGAGATCCGACAT
>chr1:8926-9050;
CAAGTGTCTGATGCTTTAAGGC
```

Motif enrichment results for NRSF and STAT1 data





HPeak performance

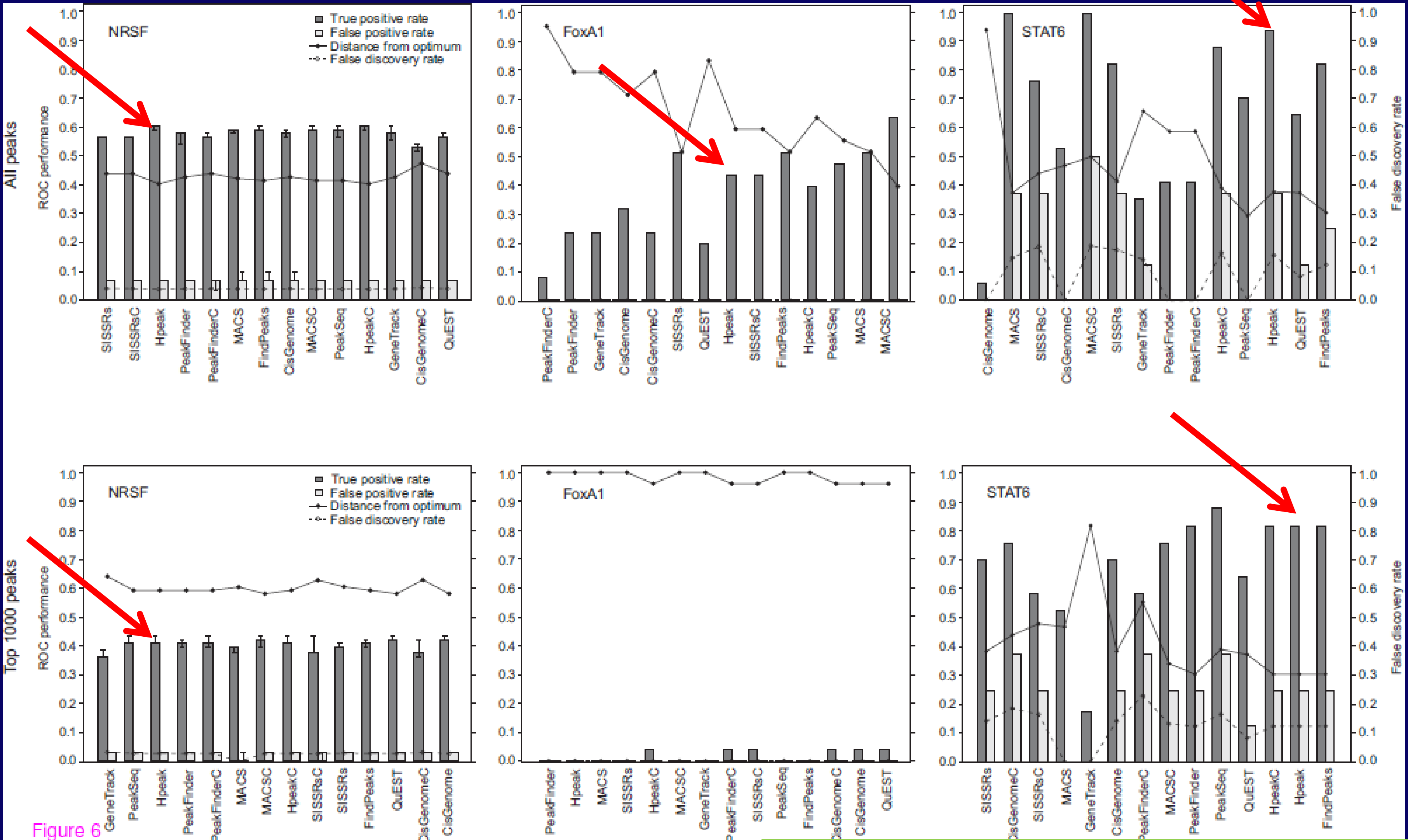


Figure 6

Laajala et al. *BMC Bioinformatics*, 2009

GP and ZIP distribution

- Do not require mean equal to variance which is useful to model over-dispersion and under-dispersion.

$$\mathrm{P}(\mathrm{Y} = \mathrm{y} \mid \lambda, \phi) = \left(\frac{\lambda}{1+\phi\lambda}\right)^{y} \frac{(1+\phi\lambda)^{y-1}}{y!} \exp\left\{\frac{-\lambda(1+\phi\lambda)}{1+\phi\lambda}\right\}$$

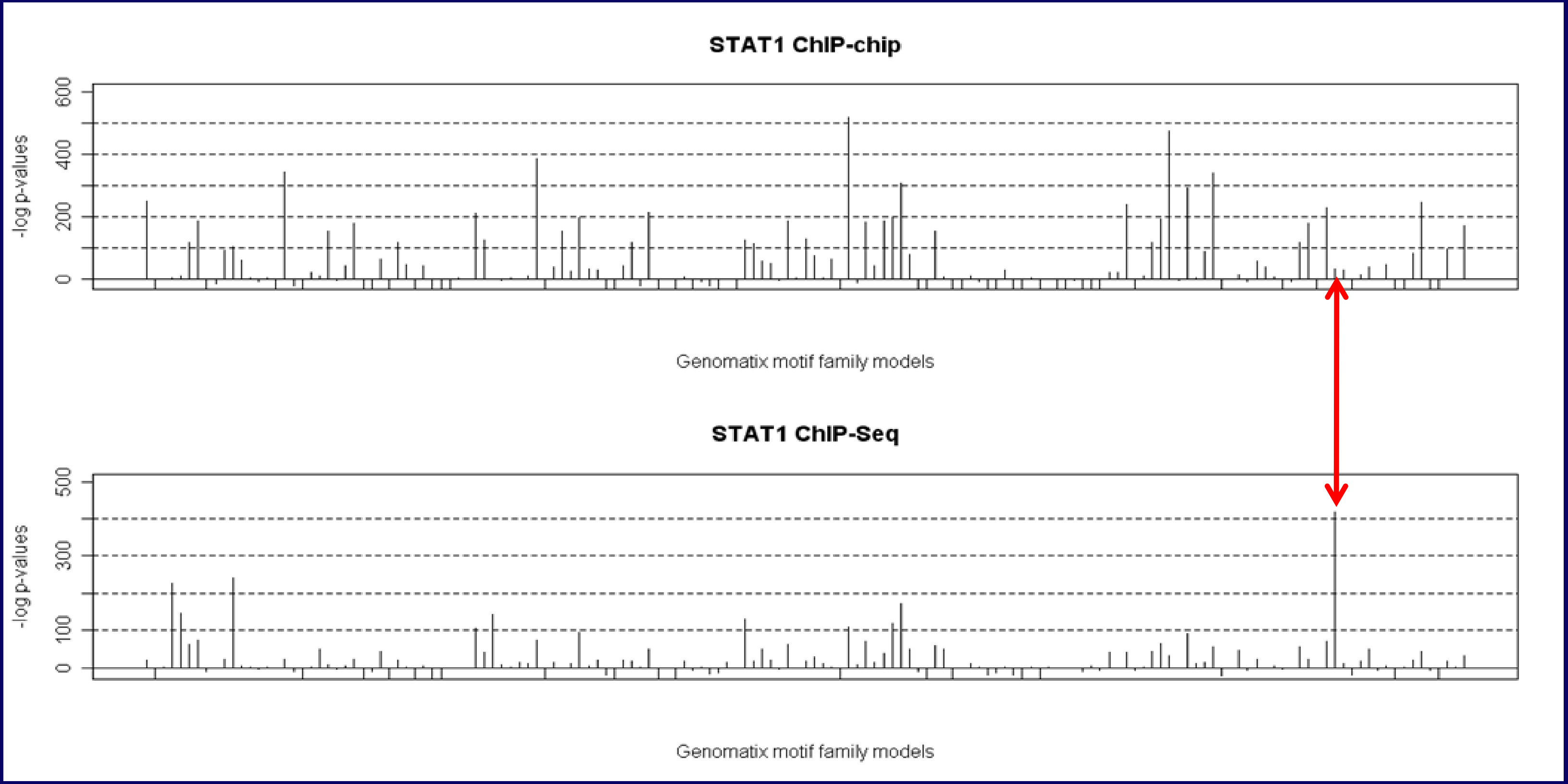
$$E(Y) = \lambda$$

$$Var(Y) = \lambda(1+\phi\lambda)^{2}$$

- Zero-inflated Poisson distribution

$$f(Y \mid \pi, \mu) = \begin{cases} (1-\pi) + \pi e^{-\mu} & \text{if } \; x = 0 \\ \dfrac{\pi e^{-\mu} \mu^{x}}{x!} & \text{if } \; x = 0 \end{cases}$$

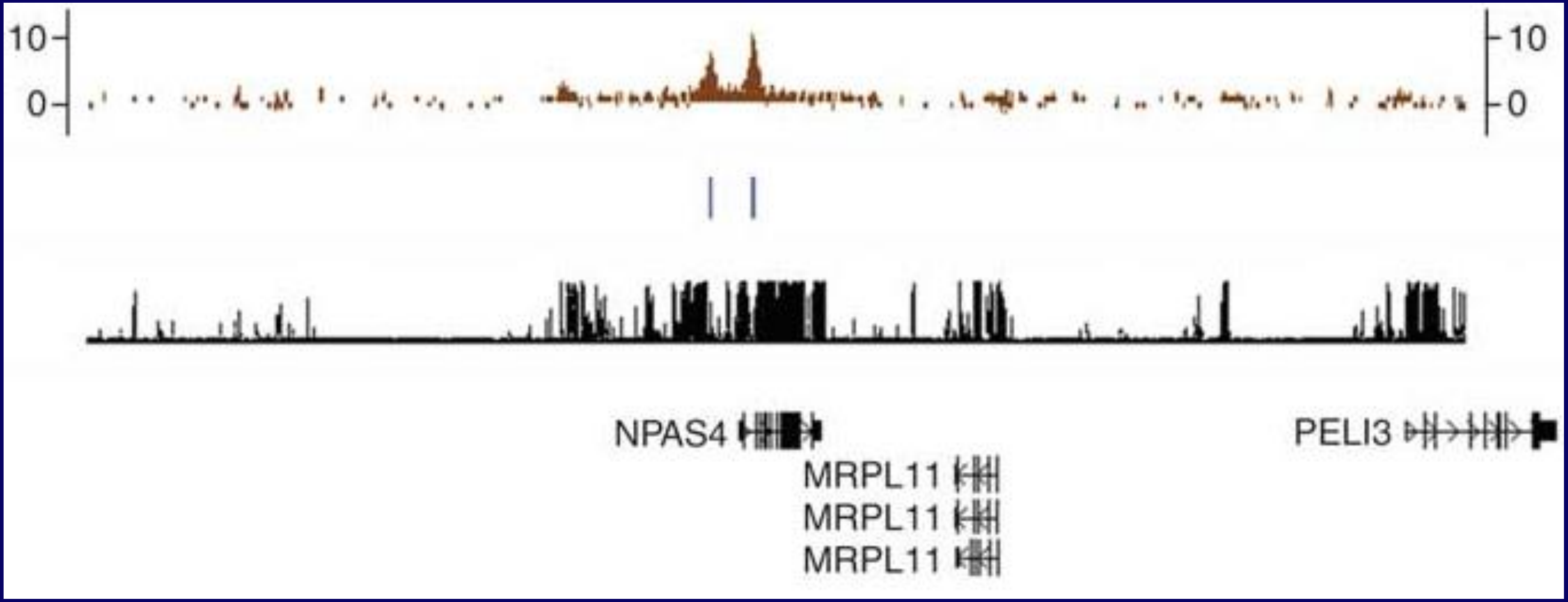
Comparison between ChIP-seq and ChIP-chip



Outline

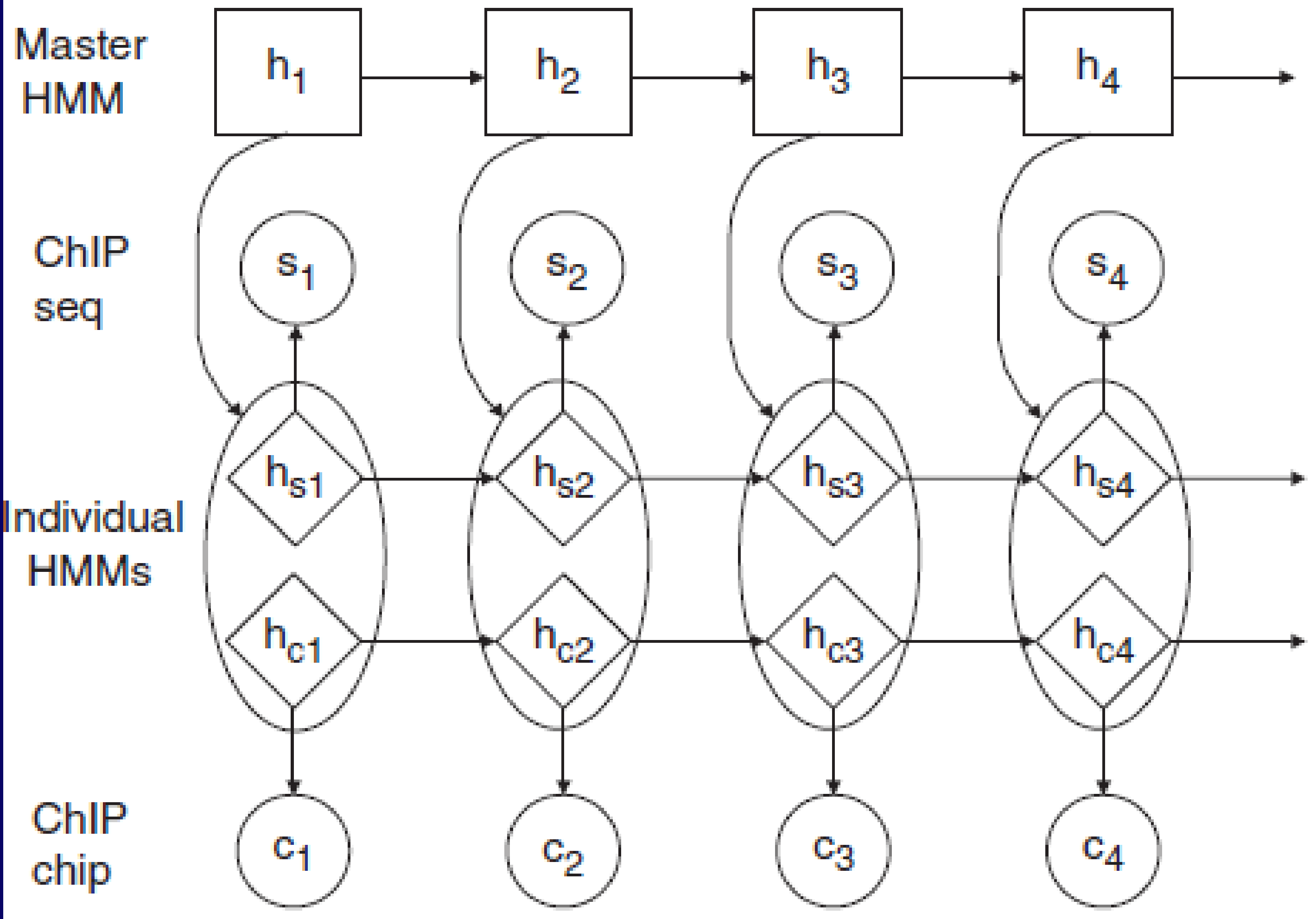
- Hidden Markov model for peak detection
- Hierarchical Hidden Markov model for combining ChIP-seq and ChIP-chip data
- Hybrid Monte Carlo strategy for Motif finding

Joint analysis of ChIP-chip and ChIP-seq

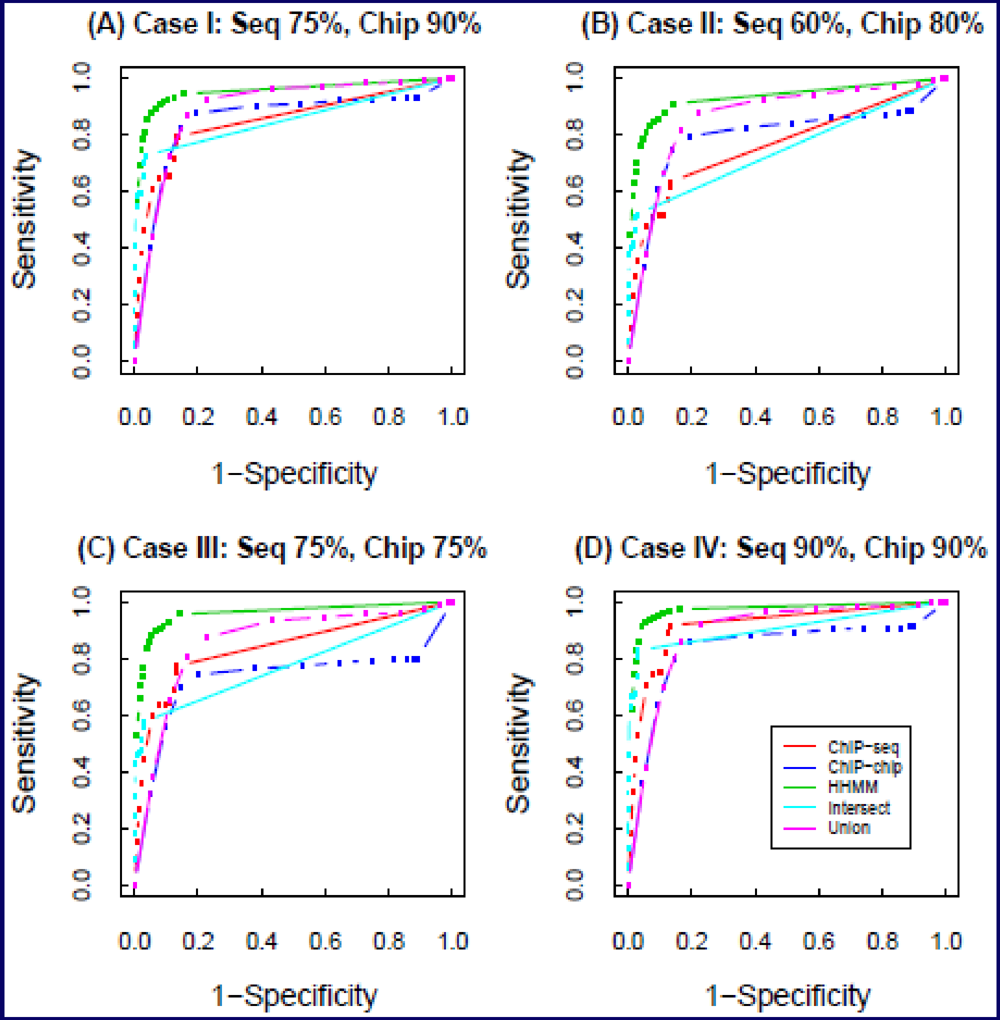


Ji et al. *Nat Biotechnology*, 2008

Hierarchical HMM

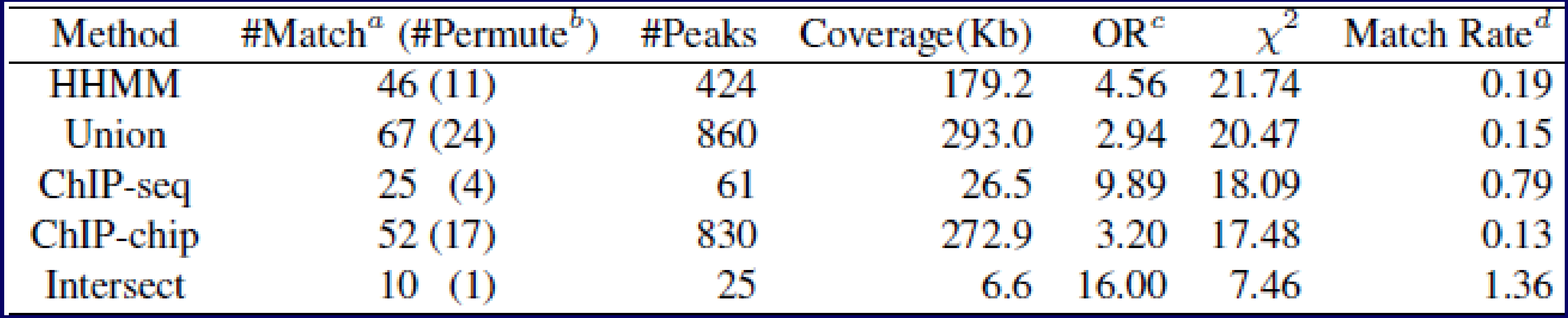


Simulated data results



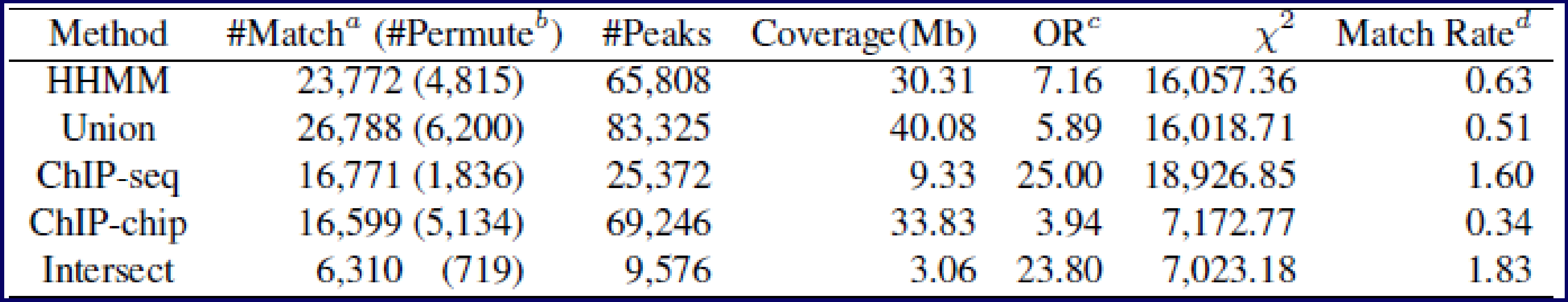
Real Data Analyses

- NRSF



Method	#Match[a] (#Permute[b])	#Peaks	Coverage(Kb)	OR[c]	χ^2	Match Rate[d]
HHMM	46 (11)	424	179.2	4.56	21.74	0.19
Union	67 (24)	860	293.0	2.94	20.47	0.15
ChIP-seq	25 (4)	61	26.5	9.89	18.09	0.79
ChIP-chip	52 (17)	830	272.9	3.20	17.48	0.13
Intersect	10 (1)	25	6.6	16.00	7.46	1.36

- CTCF

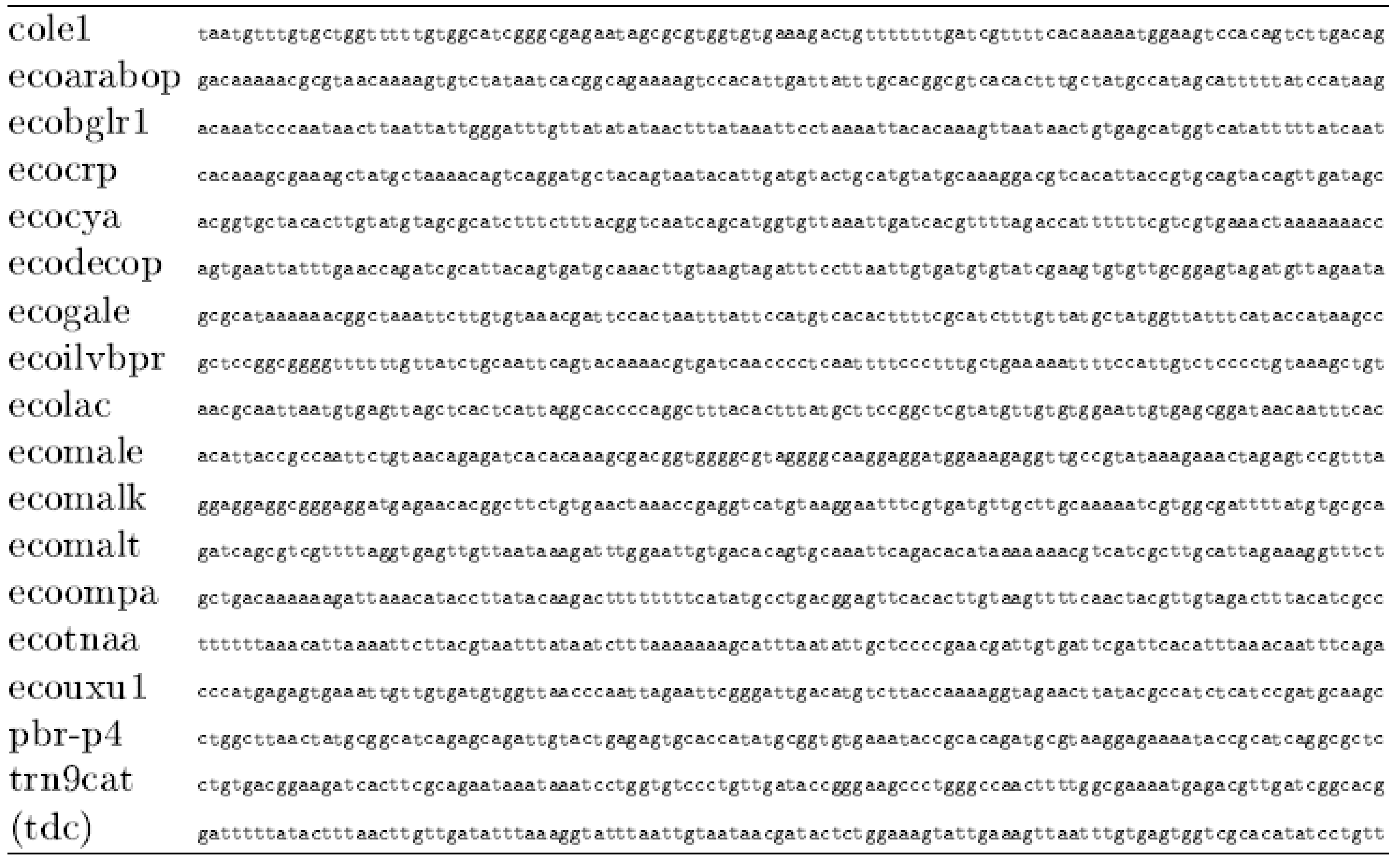


Method	#Match[a] (#Permute[b])	#Peaks	Coverage(Mb)	OR[c]	χ^2	Match Rate[d]
HHMM	23,772 (4,815)	65,808	30.31	7.16	16,057.36	0.63
Union	26,788 (6,200)	83,325	40.08	5.89	16,018.71	0.51
ChIP-seq	16,771 (1,836)	25,372	9.33	25.00	18,926.85	1.60
ChIP-chip	16,599 (5,134)	69,246	33.83	3.94	7,172.77	0.34
Intersect	6,310 (719)	9,576	3.06	23.80	7,023.18	1.83

Outline

- Hidden Markov model for peak detection
- Hierarchical Hidden Markov model for combining ChIP-seq and ChIP-chip data
- Hybrid Monte Carlo strategy for Motif finding

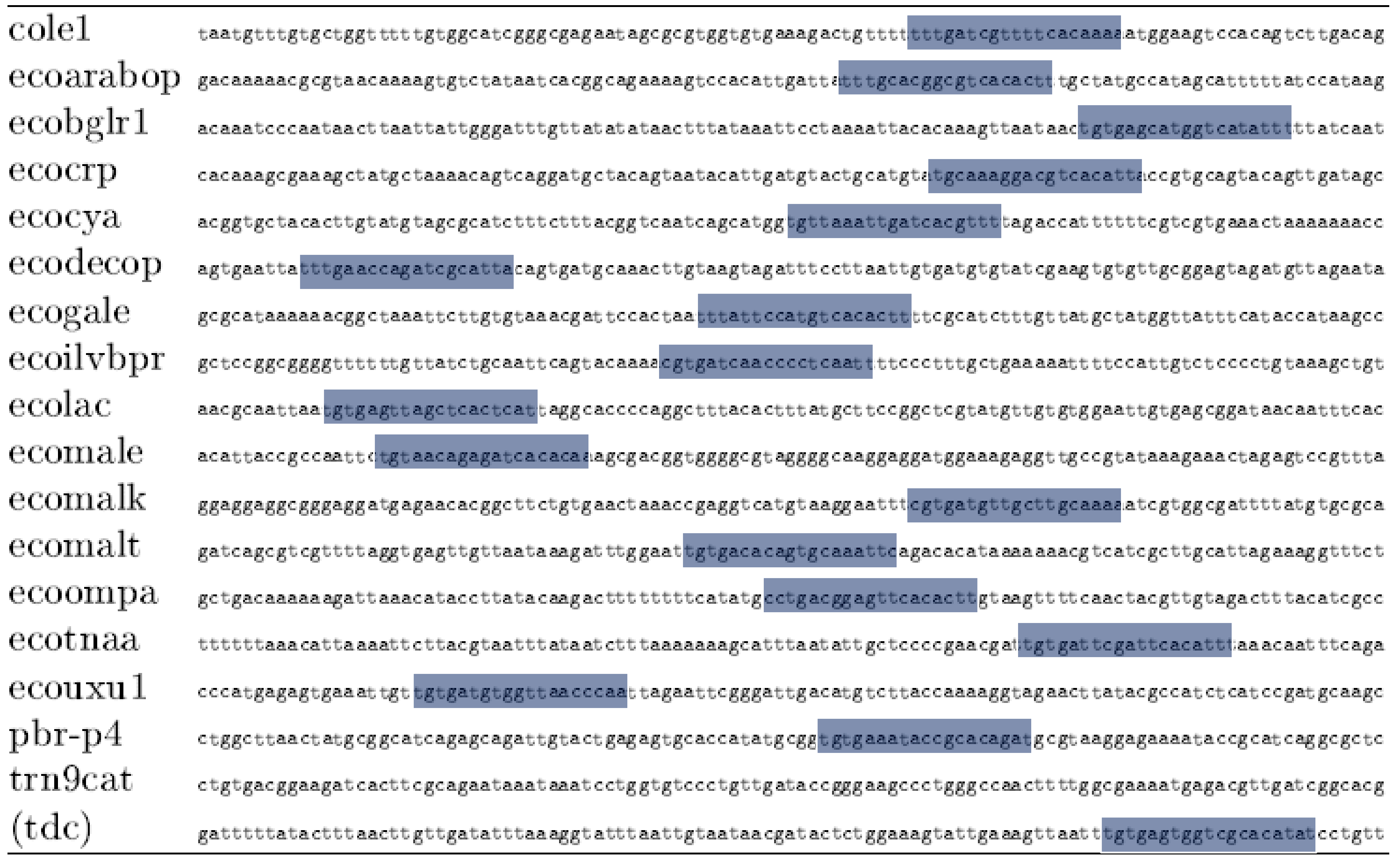
Example: cyclic receptor protein (CRP)



cole1	taatgtttgtgctggtttttgtggcatcgggcgagaatagcgcgtggtgtgaaagactgtttttttgatcgttttcacaaaaatggaagtccacagtcttgacag
ecoarabop	gacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgattatttgcacggcgtcacactttgctatgccatagcatttttatccataag
ecobglr1	acaaatcccaataacttaattattgggatttgttatatataactttataaattcctaaaattacacaaagttaataactgtgagcatggtcatatttttatcaat
ecocrp	cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatgtatgcaaaggacgtcacattaccgtgcagtacagttgatagc
ecocya	acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcatggtgttaaattgatcacgttttagaccattttttcgtcgtgaaactaaaaaaacc
ecodecop	agtgaattatttgaaccagatcgcattacagtgatgcaaacttgtaagtagatttccttaattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata
ecogale	gcgcataaaaaaacggctaaattcttgtgtaaacgattccactaatttattccatgtcacacttttcgcatctttgttatgctatggttatttcataccataagcc
ecoilvbpr	gctccggcggggttttttgttatctgcaattcagtacaaaacgtgatcaacccctcaattttccctttgctgaaaaattttccattgtctcccctgtaaagctgt
ecolac	aacgcaattaatgtgagttagctcactcattaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcac
ecomale	acattaccgccaattctgtaacagagatcacacaaagcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ecomalk	ggaggaggcgggaggatgagaacacggcttctgtgaactaaaccgaggtcatgtaaggaatttcgtgatgttgcttgcaaaaatcgtggcgattttatgtgcgca
ecomalt	gatcagcgtcgttttaggtgagttgttaataaagatttggaattgtgacacagtgcaaattcagacacataaaaaaacgtcatcgcttgcattagaaaggtttct
ecoompa	gctgacaaaaaagattaaacataccttatacaagactttttttttcatatgcctgacggagttcacacttgtaagttttcaactacgttgtagactttacatcgcc
ecotnaa	ttttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaacgattgtgattcgattcacatttaaacaatttcaga
ecouxu1	cccatgagagtgaaattgttgtgatgtggttaacccaattagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
pbr-p4	ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgcggtgtgaaataccgcacagatgcgtaaggagaaaataccgcatcaggcgctc
trn9cat	ctgtgacggaagatcacttcgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaaaatgagacgttgatcggcacg
(tdc)	gatttttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagttaatttgtgagtggtcgcacatatcctgtt

Stormo and Hartzell, 1989

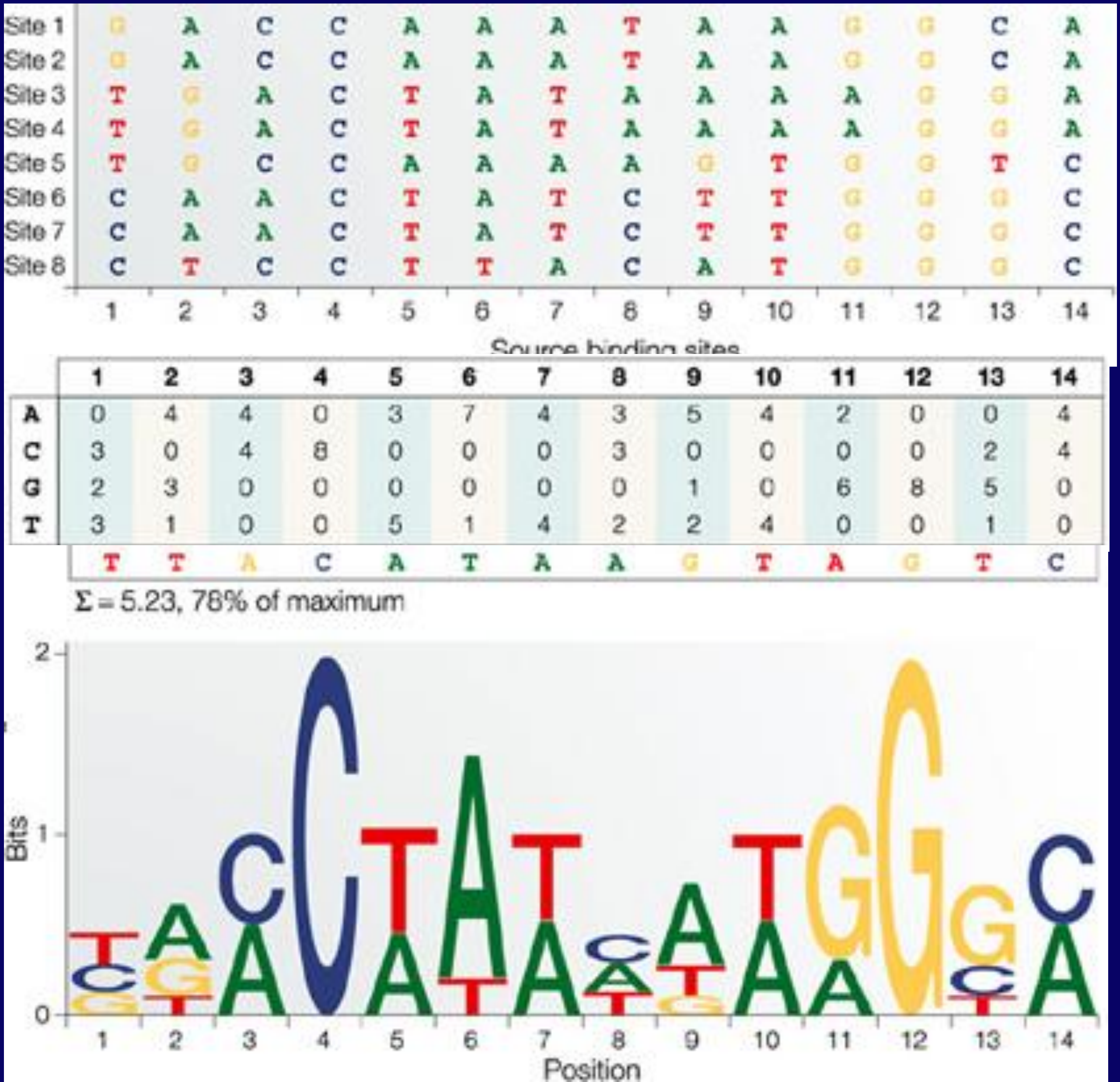
Example: cyclic receptor protein (CRP)



cole1	taatgtttgtgctggttttttgtggcatcgggcgagaatagcgcgtggtgtgaaagactgtttt**tttgatcgttttcacaaaa**atggaagtccacagtcttgacag
ecoarabop	gacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgatta**tttgcacggcgtcacactt**tgctatgccatagcatttttatccataag
ecobglr1	acaaatcccaataacttaattattgggatttgttatatataactttataaattcctaaaattacacaaagttaataac**tgtgagcatggtcatattt**ttatcaat
ecocrp	cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatgta**tgcaaaggacgtcacatta**ccgtgcagtacagttgatagc
ecocya	acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcatgg**tgttaaattgatcacgttt**tagaccattttttcgtcgtgaaactaaaaaaacc
ecodecop	agtgaatta**tttgaaccagatcgcatta**cagtgatgcaaacttgtaagtagatttccttaattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata
ecogale	gcgcataaaaaaacggctaaattcttgtgtaaacgattccactaa**tttattccatgtcacactt**ttcgcatctttgttatgctatggttatttcataccataagcc
ecoilvbpr	gctccggcggggtttttttgttatctgcaattcagtacaaaa**cgtgatcaacccctcaatt**ttccctttgctgaaaaatttttccattgtctcccctgtaaagctgt
ecolac	aacgcaattaa**tgtgagttagctcactcat**taggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcac
ecomale	acattaccgccaattc**tgtaacagagatcacaca**aagcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ecomalk	ggaggaggcgggaggatgagaacacggcttctgtgaactaaaccgaggtcatgtaaggaatttt**cgtgatgttgcttgcaaaa**atcgtggcgattttatgtgcgca
ecomalt	gatcagcgtcgttttaggtgagttgttaataaagatttggaat**tgtgacacagtgcaaattc**agacacataaaaaaacgtcatcgcttgcattagaaaggtttct
ecoompa	gctgacaaaaaagattaaacataccttatacaagactttttttttcatatg**cctgacggagttcacactt**gtaagttttcaactacgttgtagactttacatcgcc
ecotnaa	ttttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaacgat**tgtgattcgattcacattt**aaacaatttcaga
ecouxu1	cccatgagagtgaaattgt**tgtgatgtggttaacccaa**ttagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
pbr-p4	ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgcgg**tgtgaaataccgcacagat**gcgtaaggagaaaataccgcatcaggcgctc
trn9cat	ctgtgacggaagatcacttcgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaaaatgagacgttgatcggcacg
(tdc)	gatttttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagttaatt**tgtgagtggtcgcacatat**cctgtt

Stormo and Hartzell

Transcription factor binding site (TFBS)



	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C

Source binding sites

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
A	0	4	4	0	3	7	4	3	5	4	2	0	0	4
C	3	0	4	8	0	0	0	3	0	0	0	0	2	4
G	2	3	0	0	0	0	0	0	1	0	6	8	5	0
T	3	1	0	0	5	1	4	2	2	4	0	0	1	0
	T	T	A	C	A	T	A	A	G	T	A	G	T	C

Σ = 5.23, 78% of maximum

Existing *de novo* motif finding algorithms

• Consensus	Hertz *et al.* 1990
• Gibbs Motif Sampler	Lawrence *et al.* 1993
• MEME	Bailey and Elkan 1994
• AlignACE	Roth *et al.* 1998
• BioProspector	Liu *et al.* 2001
• MDScan	Liu *et al.* 2002
• Mobydick	Bussemaker *et al.* 2000
…	
Review	Tompa *et al.* 2005

Motif identification model

a_1

aaaggtcgagtagctactcgatcgatactagcaatcgttaccctagctcgatcgaaa

a_2

acgtgagatcagctatgaccgatagctactcgataaccg

a_3

gaatagctactcgatcgatactagcaatcgttaccctagctcgatcgagatggaaagactataa

...

a_J

acgtgagatcagctatcgatcgattgataactactcgtacgtat

Alignment variable $A = \{a_1, a_2 ..., a_J\}$

Posterior distributions

- The posterior conditional distribution for alignment variable ***A***

$$p(a_j = l \mid \boldsymbol{\theta_0}, \boldsymbol{\Theta}, \boldsymbol{R_j}, \boldsymbol{A_{-j}}) \propto \prod_{k=1}^{4} \theta_{0k}^{h_k(\boldsymbol{R_j})} \prod_{i=1}^{w} \prod_{k=1}^{4} \left(\frac{\theta_{ik}}{\theta_{0k}} \right)^{h_k(r_{j,l+i-1})} \propto \prod_{i=1}^{w} \prod_{k=1}^{4} \left(\frac{\theta_{ik}}{\theta_{0k}} \right)^{h_k(r_{j,l+i-1})}$$

DNA sequence data $\boldsymbol{R} = (\boldsymbol{R_1}, \ldots, \boldsymbol{R_J})$

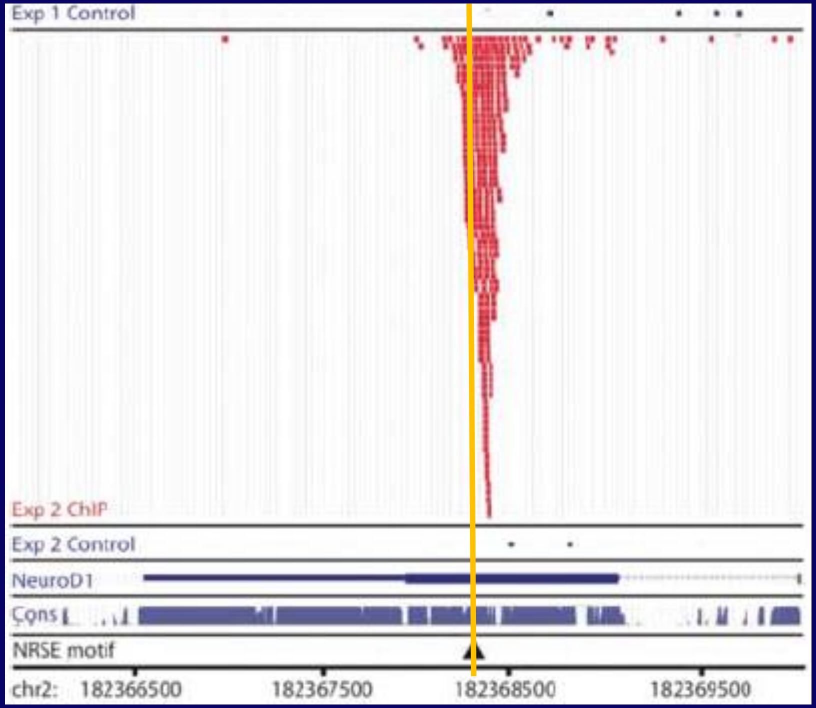
Lawrence *et al. Science* 1993, Liu *et al. JASA* 1995

Why *de novo* motif search

- The only option when the TF binding motif pattern is unknown.
- Reassuring to be able to rediscover the known TFBS motif.
- Many “known” motif patterns are biased and inaccurate.
- Multiple co-factors are often required in transcription regulation in eukaryotes.
- Binding specificity for some TFs may change under different conditions.

Challenges faced

- How to handle large number of input sequences?
- How to utilize sequencing depth information?



Johnson *et al. Science*

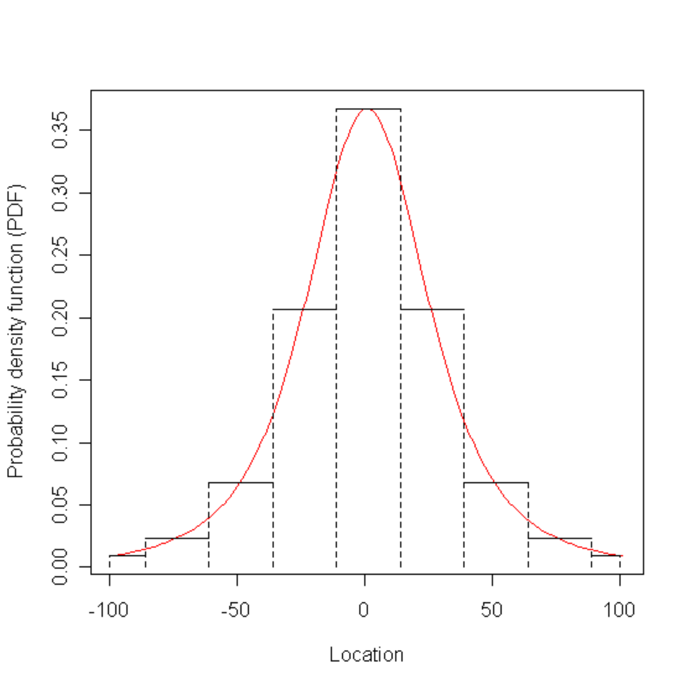
Features of our new algorithm

- Incorporate sequencing depth information in the statistical model.
- Generalize the product multinomial model to allow inter-dependent positions within the motif.
- Adopt a hybrid Monte Carlo strategy to speed up the traditional Gibbs sampler-based algorithm.

The informative prior

- The prior is symmetric and centered at the peak summit.
- The prior probabilities stem from Student’s *t*-distribution with df=3.

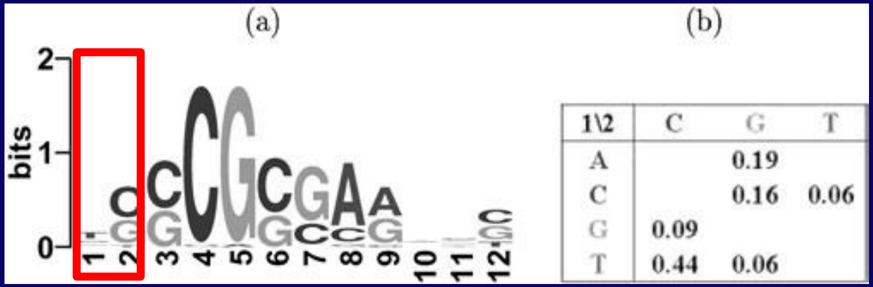
$$p(a_j = l) \propto t_3\left(\text{int}\left[\frac{| l + w/2 - s_j | + u/2}{u} \right] \right)$$



Modeling inter-dependent positions

- Zhou and Liu

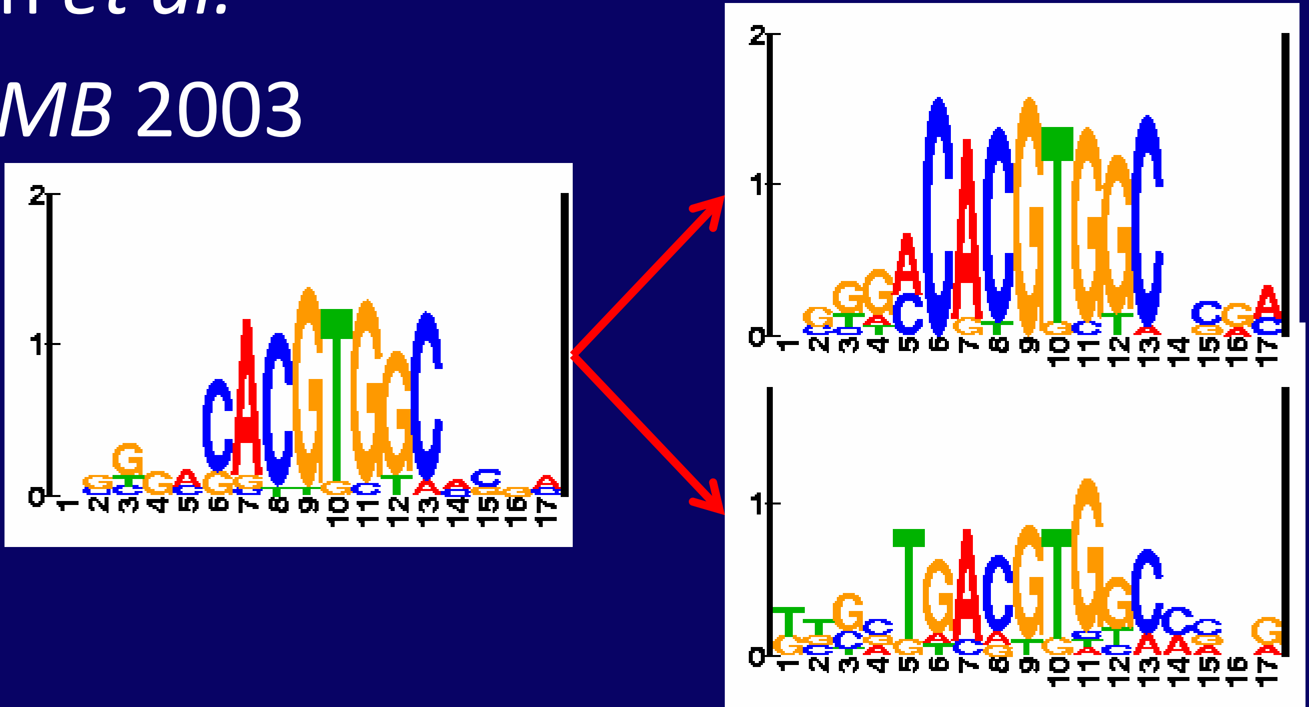
  Bioinformatics 2005



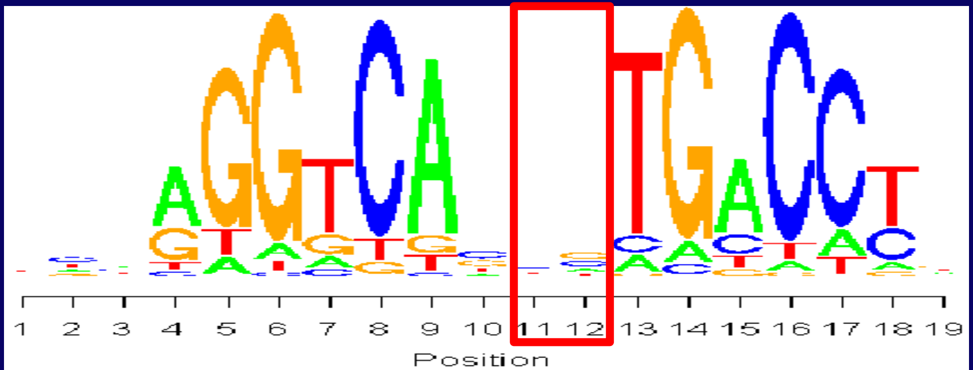
1\2	C	G	T
A		0.19	
C		0.16	0.06
G	0.09		
T	0.44	0.06	

- Barash *et al.*

  RECOMB 2003

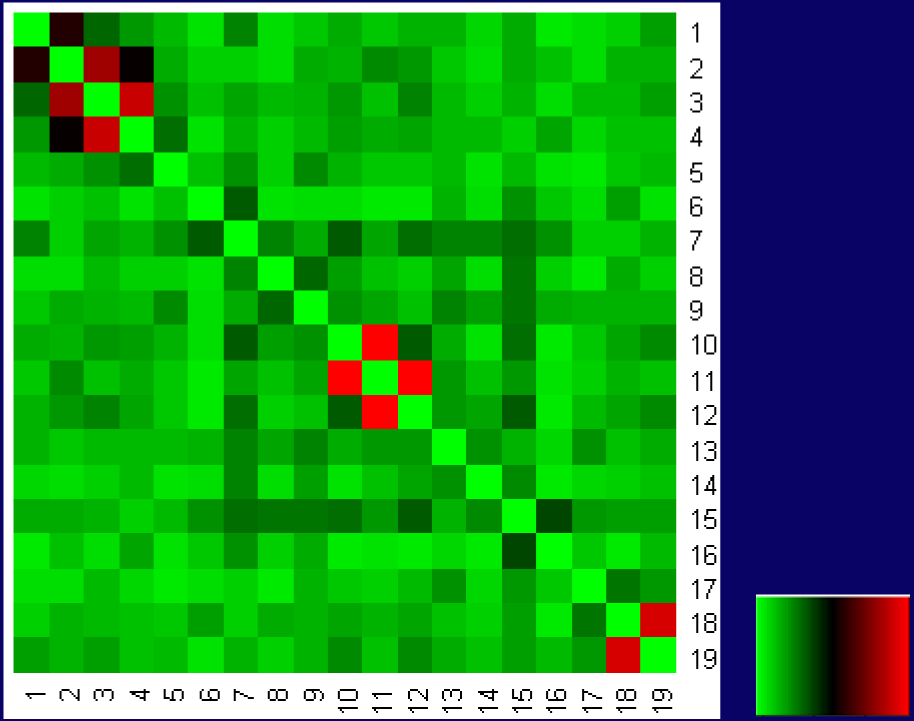


Detect intra-dependent position pairs



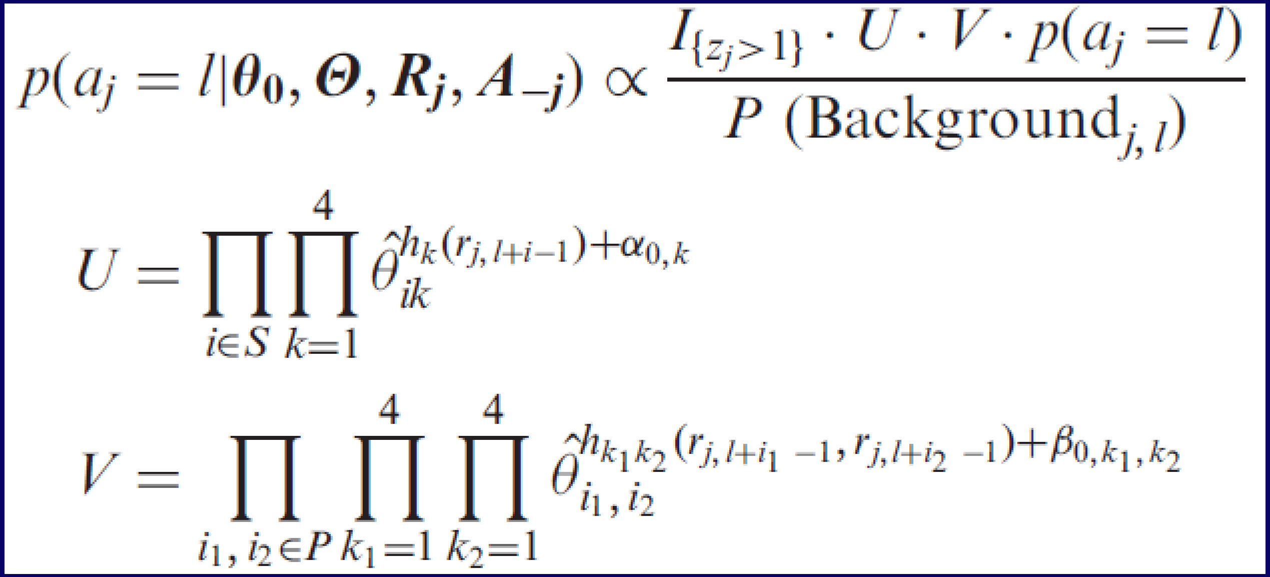
$$d_{ij} = \sum_{x=1}^{4}\sum_{y=1}^{4}\left|\hat{\eta}_{xy}(r_i,r_j) - \hat{\eta}_x(r_i)\hat{\eta}_y(r_j)\right|$$

	A	C	T	G	
A	0.03 (0.04)	0.15 (0.25)	0.28 (0.16)	0.03 (0.03)	0.49
C	0.00 (0.00)	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)	0.01
T	0.05 (0.04)	0.34 (0.24)	0.06 (0.17)	0.03 (0.03)	0.48
G	0.00 (0.00)	0.02 (0.01)	0.00 (0.01)	0.00 (0.00)	0.02
	0.08	0.52	0.34	0.06	1



New algorithm

- The posterior conditional distribution of alignment variable ***A*** under the new statistical model.



$$p(a_j = l|\boldsymbol{\theta}_0, \boldsymbol{\Theta}, \boldsymbol{R}_j, A_{-j}) \propto \frac{I_{\{z_j>1\}} \cdot U \cdot V \cdot p(a_j = l)}{P\left(\text{Background}_{j,l}\right)}$$

$$U = \prod_{i \in S} \prod_{k=1}^{4} \hat{\theta}_{ik}^{h_k(r_{j,l+i-1})+\alpha_{0,k}}$$

$$V = \prod_{i_1, i_2 \in P} \prod_{k_1=1}^{4} \prod_{k_2=1}^{4} \hat{\theta}_{i_1,i_2}^{h_{k_1k_2}(r_{j,l+i_1-1}, r_{j,l+i_2-1})+\beta_{0,k_1,k_2}}$$

Prioritized hybrid Monte Carlo

- Subject each sequence to either stochastic sampling or greedy search.
- Input sequences are not created equal.
- ChIP-enrichment is indicative of binding affinity.

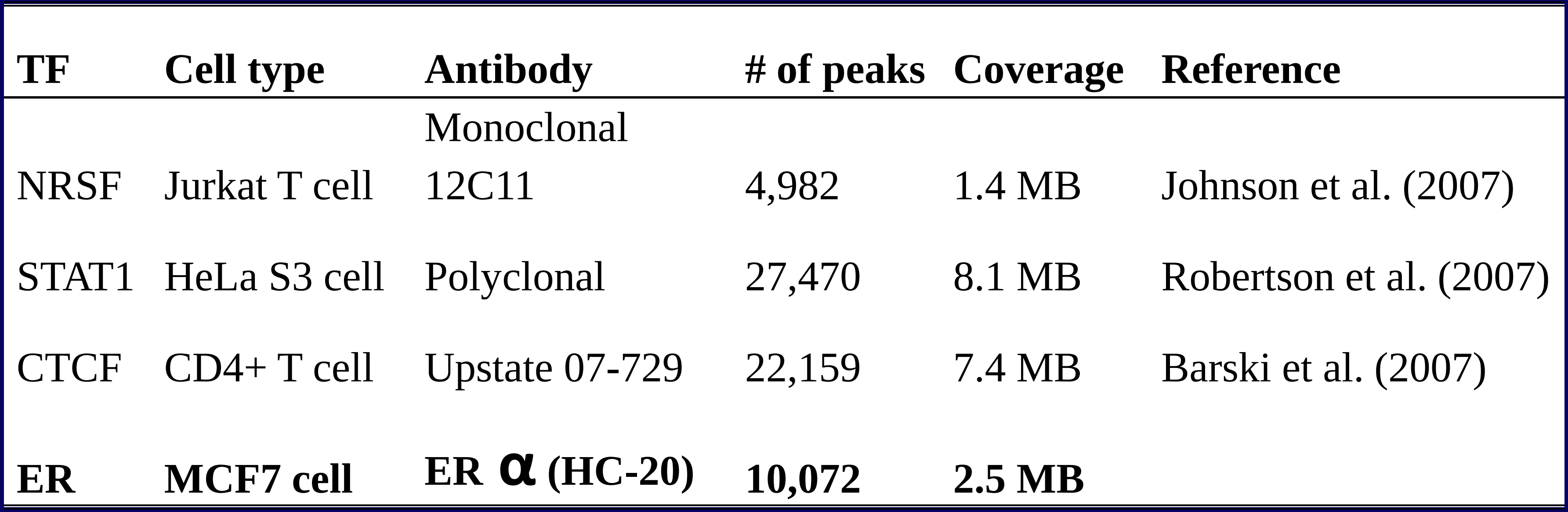
Implementation

- **H**ybrid **M**otif **S**ampler (HMS).
- Gibbs sampler type iterative procedure.
- Run multiple chains to avoid trapping in local mode.

Performance comparison

- Two established and popular motif discovery tools:
  - MEME (Bailey and Elkan 1994),
    - EM-based motif finding algorithm,
    - widely used.
  - MDscan (Liu *et al.* 2002),
    - designed to analyze ChIP-chip data,
    - combines word enumeration and probability matrix updating,
    - take into account ChIP-chip ranking,
    - very fast.

Real data analysis



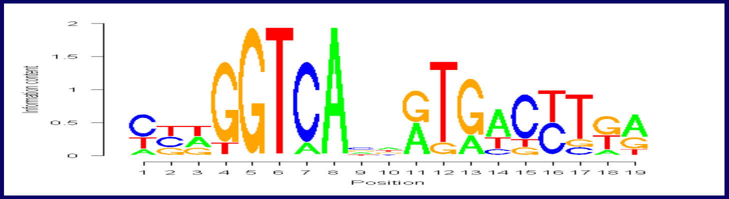
TF	Cell type	Antibody	# of peaks	Coverage	Reference
NRSF	Jurkat T cell	Monoclonal 12C11	4,982	1.4 MB	Johnson et al. (2007)
STAT1	HeLa S3 cell	Polyclonal	27,470	8.1 MB	Robertson et al. (2007)
CTCF	CD4+ T cell	Upstate 07-729	22,159	7.4 MB	Barski et al. (2007)
ER	**MCF7 cell**	**ER α (HC-20)**	**10,072**	**2.5 MB**	

Performance evaluation

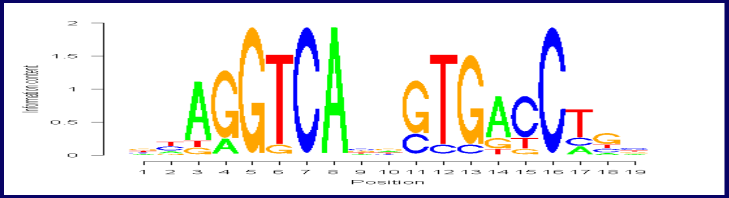
- Cross validation
  - Randomly separate all peaks into two halves: training and testing.
  - Run motif finding algorithms on the training data to predict the motif pattern.
  - Scan testing data using the identified motif pattern and compare to a set of control sequences.
- Testing
  - Using Chi-square test statistics to quantify motif enrichment .
  - Estimate FDR and plot FDR versus Chi-square test statistics.

Compare ER motif patterns

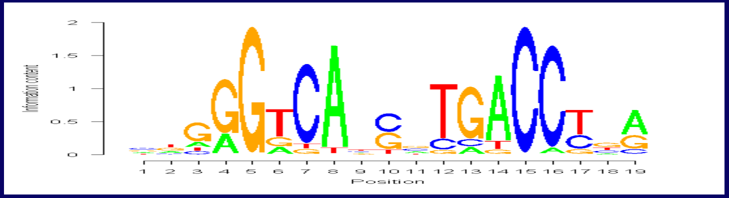
- V$ER01*



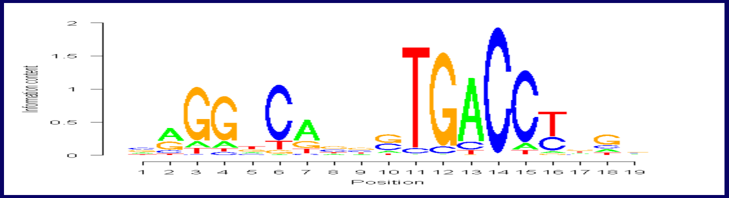
- V$ER02*



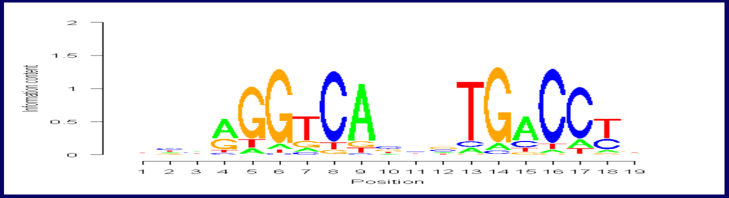
- V$ER03*



- MEME

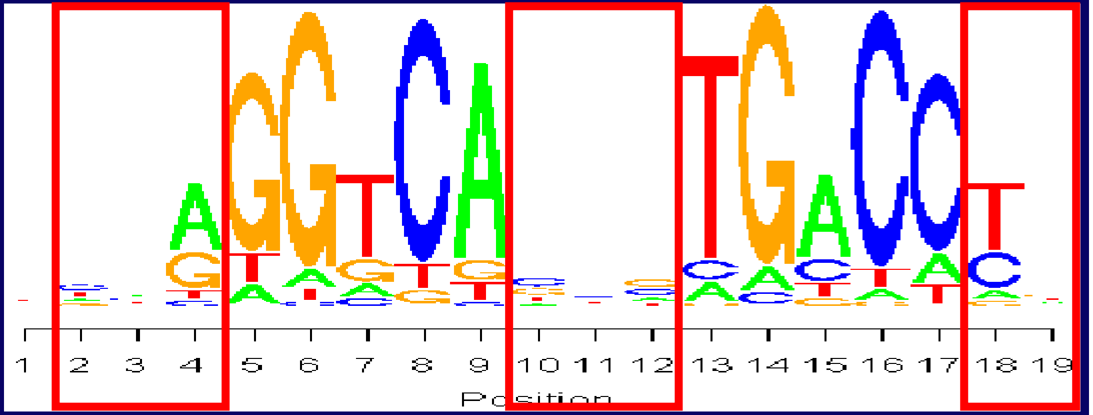


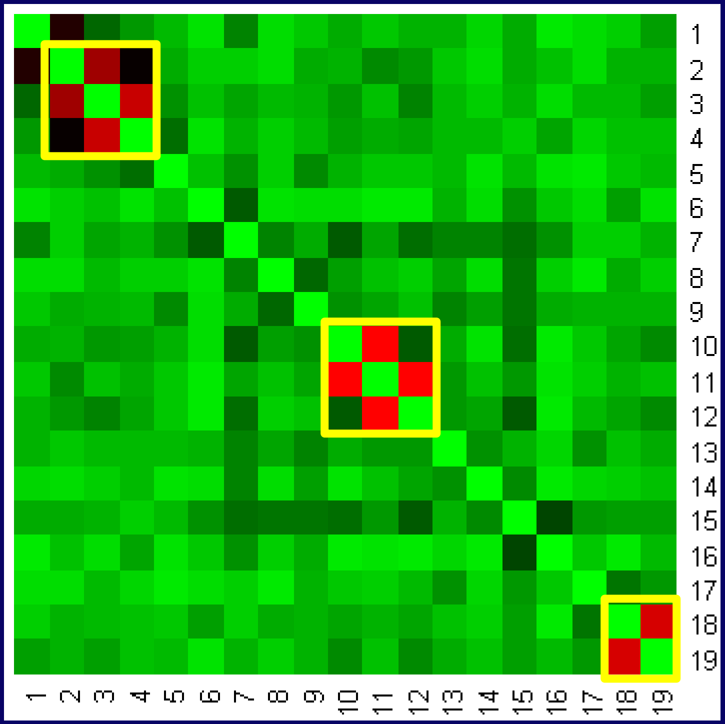
- HMS



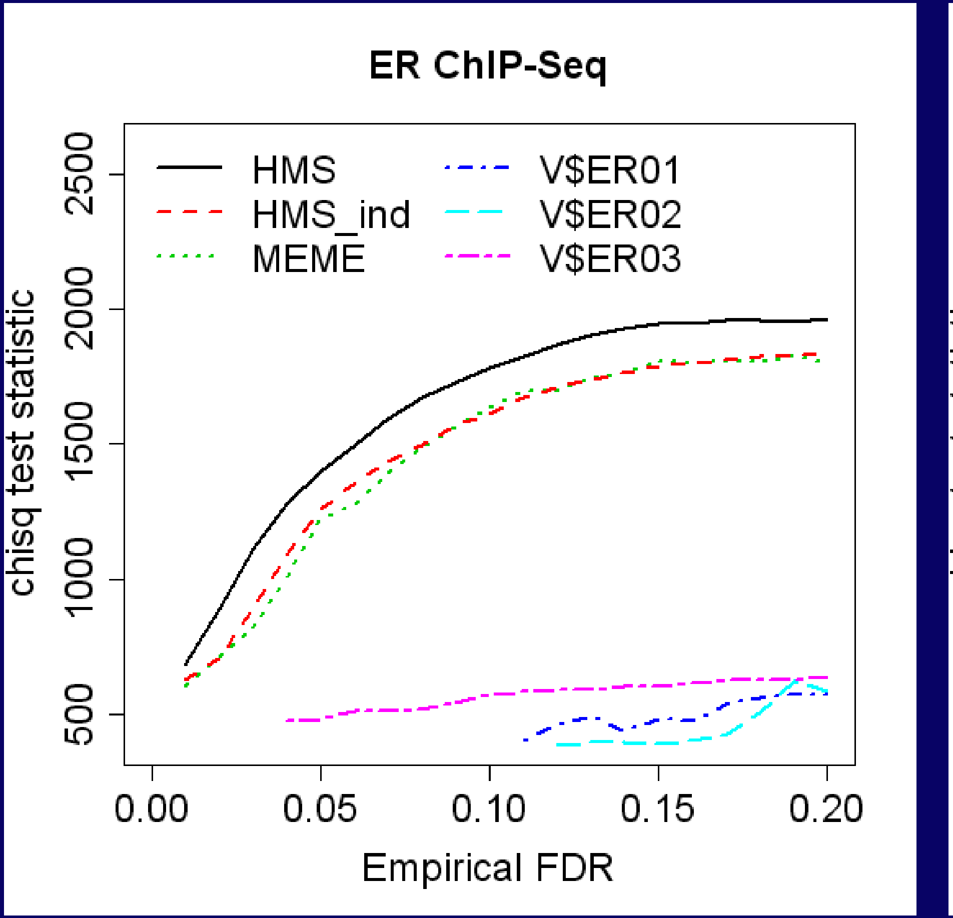
* Genomatix
understanding gene regulation

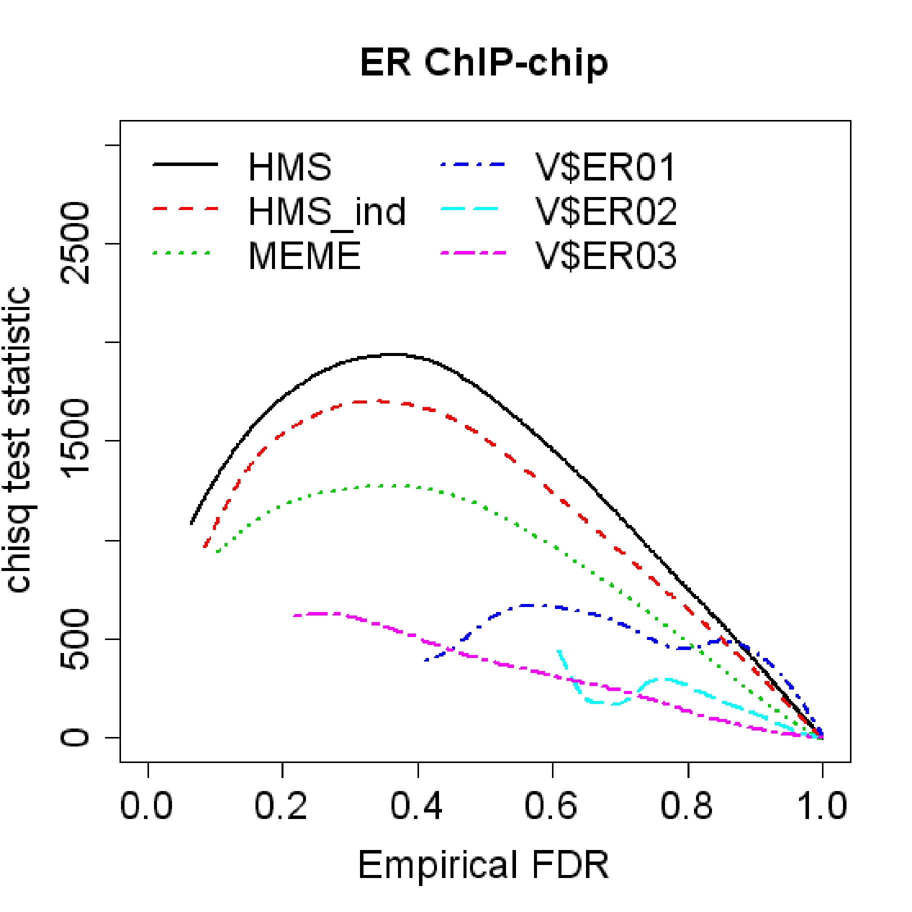
Positions show inter-dependency inside the ER motif





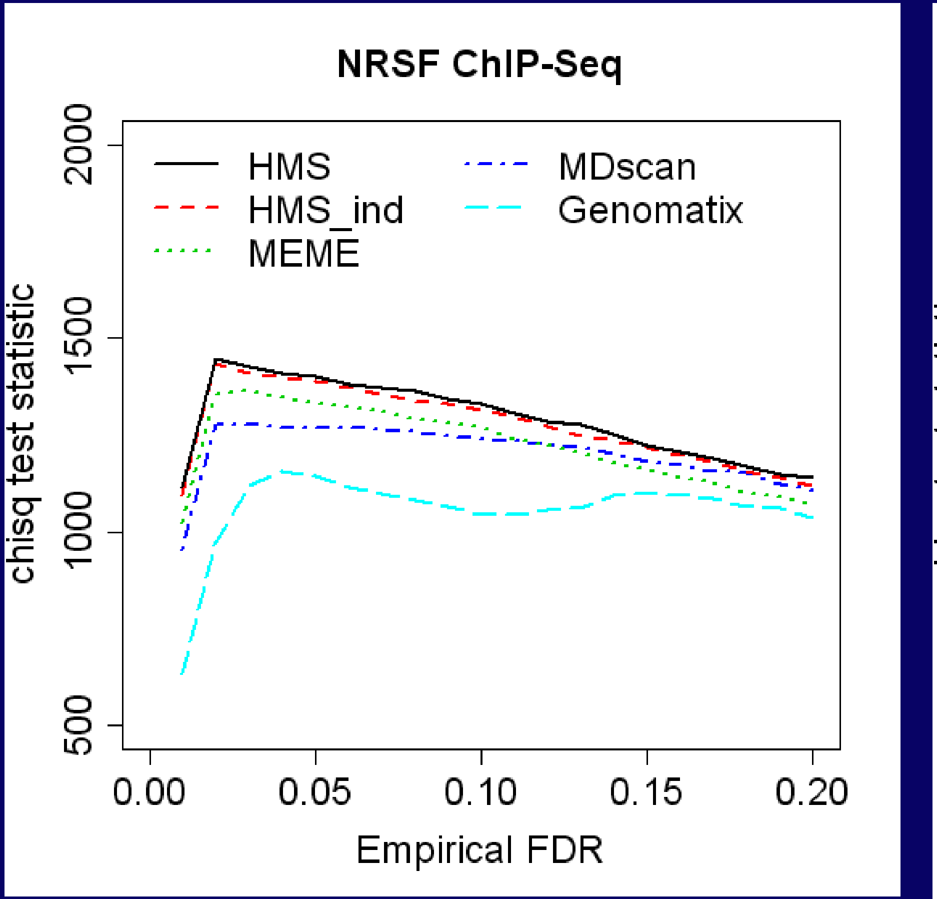
Compare ER motif enrichment

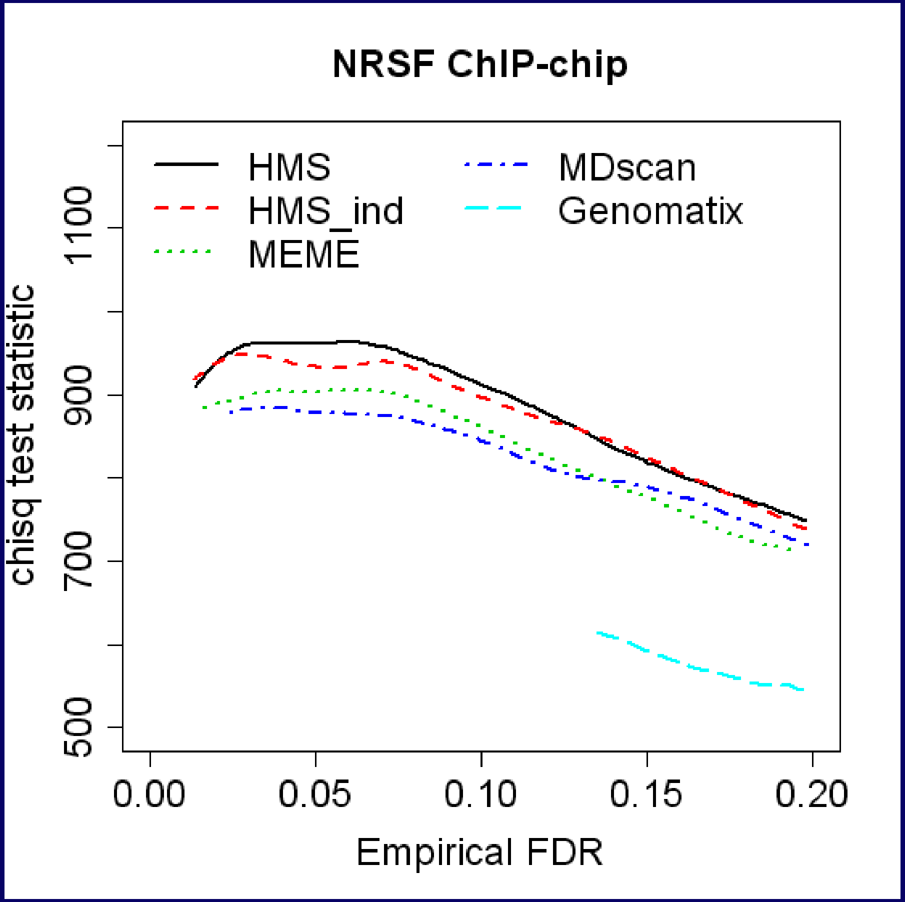




Carroll *et al. Nature Genetics* 2006

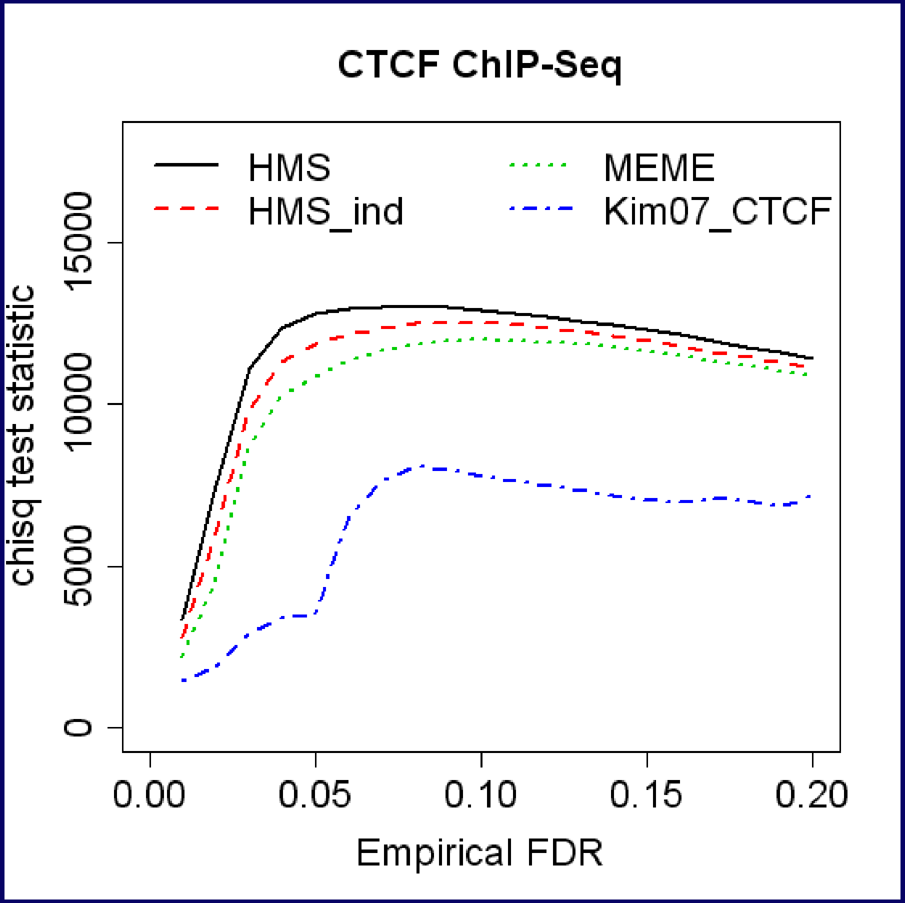
Compare NRSF motif enrichment

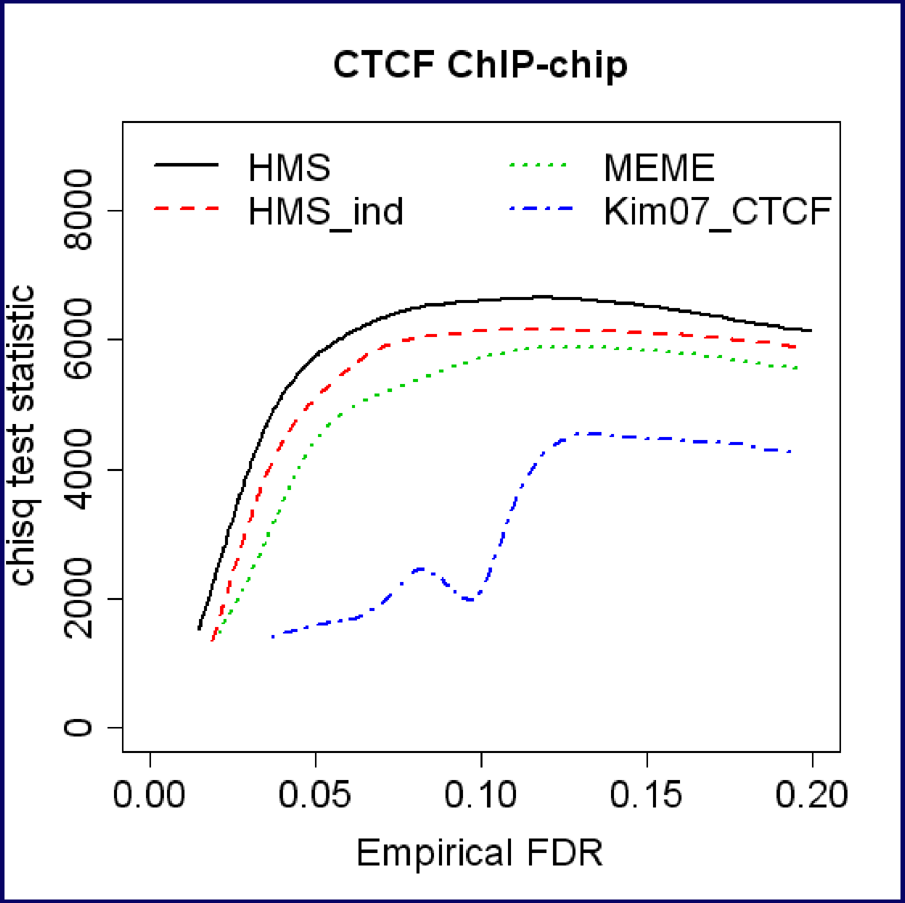




Johnson *et al. Science*

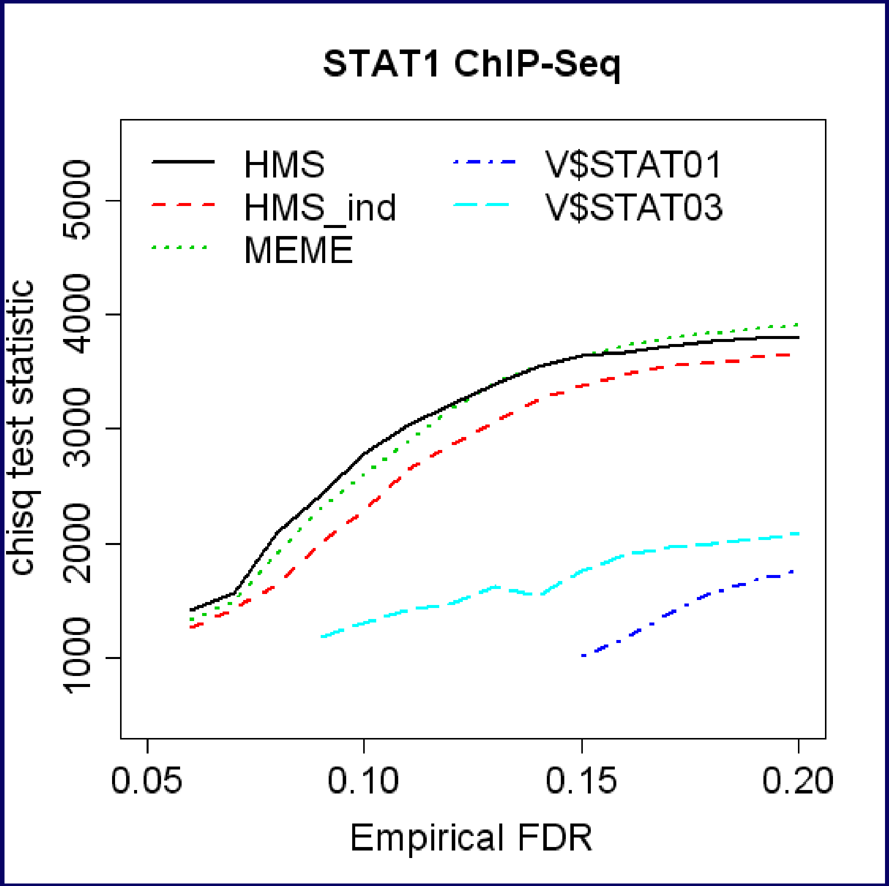
Compare CTCF motif enrichment

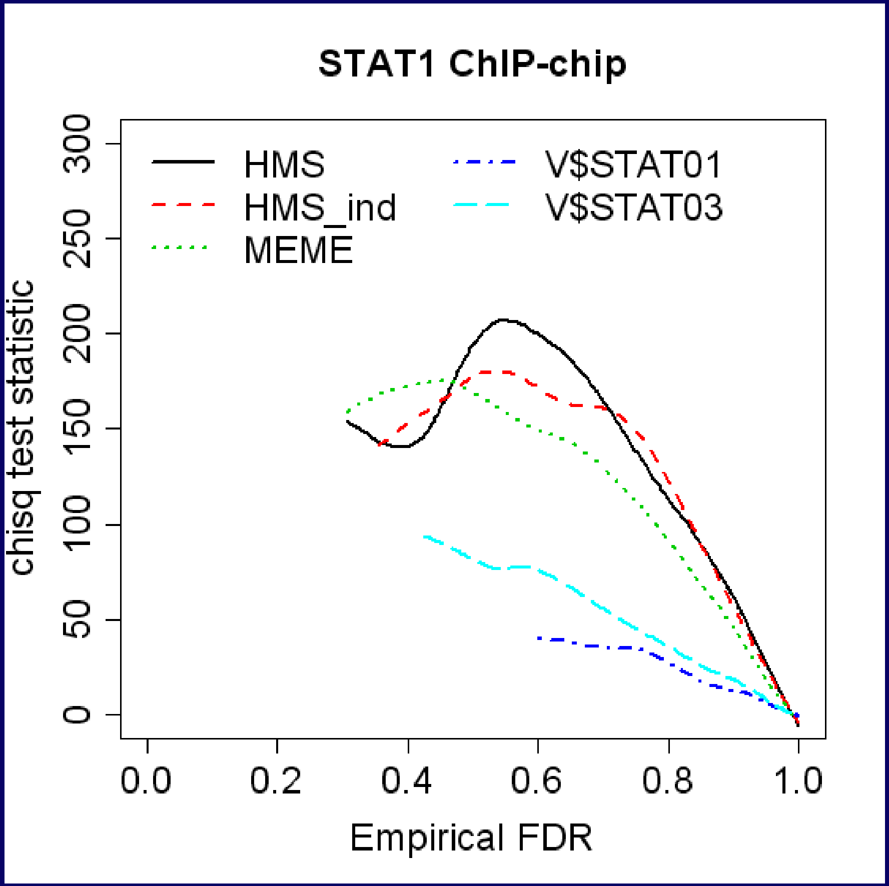




Kim *et al. Cell* 2007

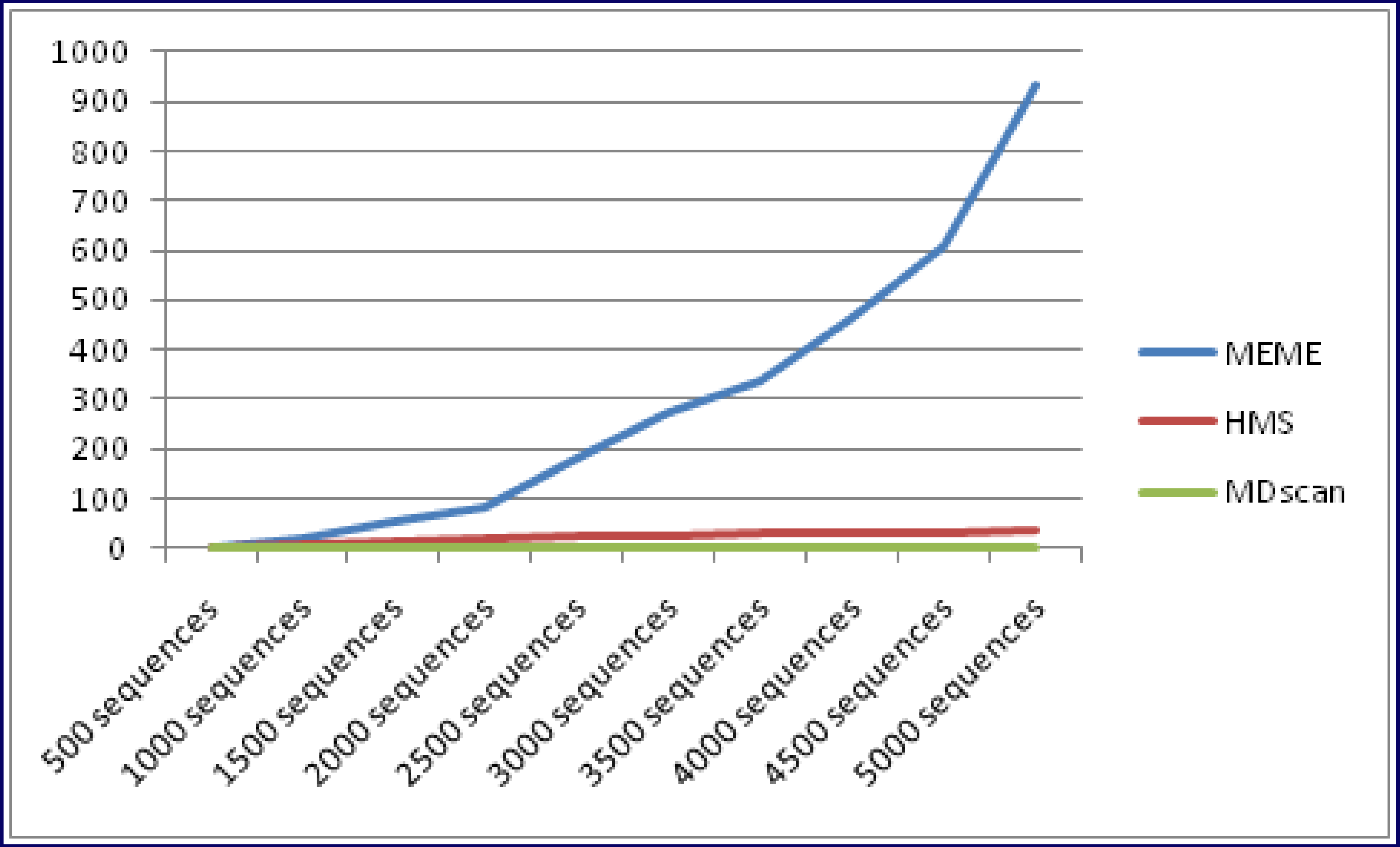
Compare STAT1 motif enrichment





Euskirchen *et al. Genome Res*

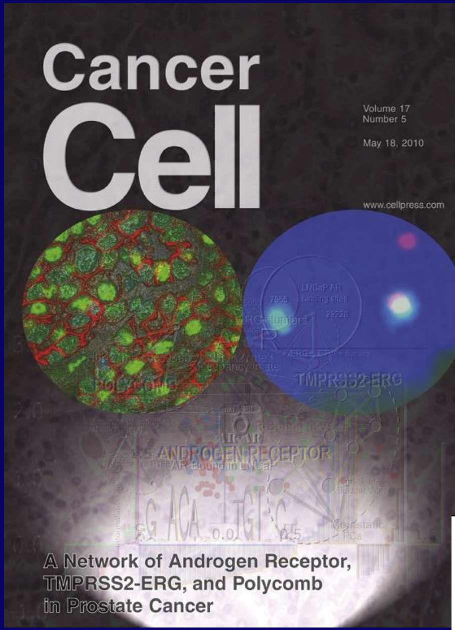
Computation time



Summary

- ChIP-Seq data offers abundant information and provides much improved opportunity for studying protein-DNA interaction.
- There are many biological and technical factors that affect the ChIP-Seq data we observe, careful modeling is critical in order to process ChIP-Seq data efficiently and thoroughly.
- New sequencing data are different from microarray, ChIP-chip data. Methods developed there do not work well for analyzing sequencing data, new models and algorithms need to be developed.

Apply to cancer genomics



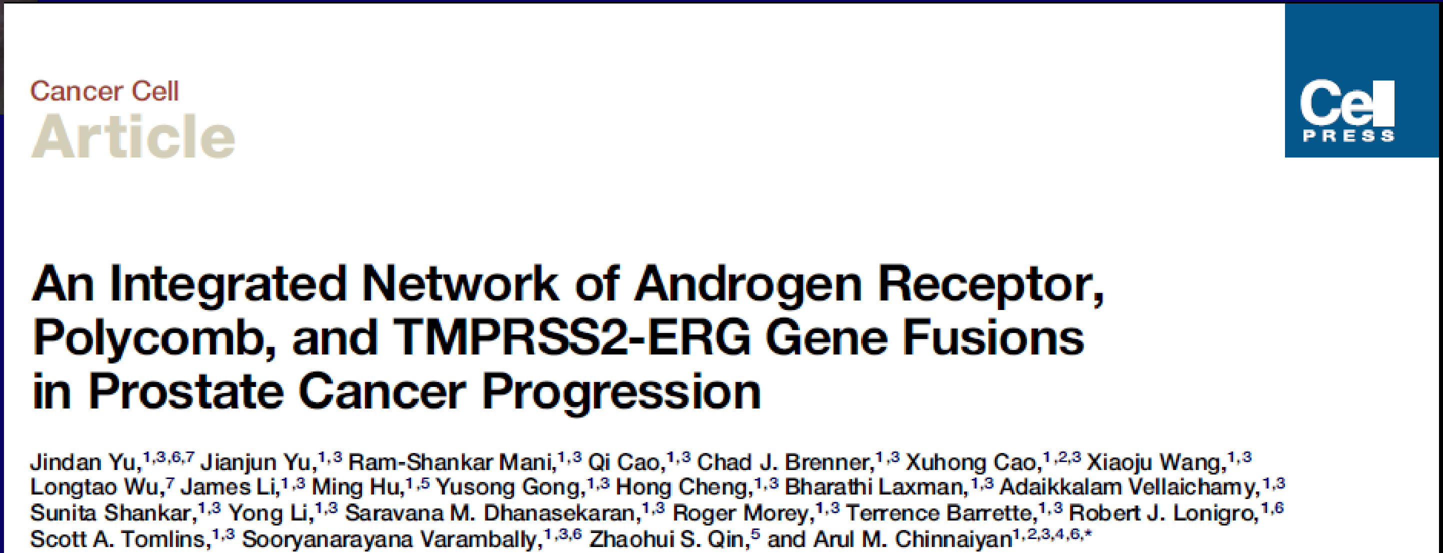
Cancer Cell

Volume 17
Number 5

May 18, 2010

www.cellpress.com

A Network of Androgen Receptor, TMPRSS2-ERG, and Polycomb in Prostate Cancer



Cancer Cell
Article

An Integrated Network of Androgen Receptor, Polycomb, and TMPRSS2-ERG Gene Fusions in Prostate Cancer Progression

Jindan Yu,[1,3,6,7] Jianjun Yu,[1,3] Ram-Shankar Mani,[1,3] Qi Cao,[1,3] Chad J. Brenner,[1,3] Xuhong Cao,[1,2,3] Xiaoju Wang,[1,3] Longtao Wu,[7] James Li,[1,3] Ming Hu,[1,5] Yusong Gong,[1,3] Hong Cheng,[1,3] Bharathi Laxman,[1,3] Adaikkalam Vellaichamy,[1,3] Sunita Shankar,[1,3] Yong Li,[1,3] Saravana M. Dhanasekaran,[1,3] Roger Morey,[1,3] Terrence Barrette,[1,3] Robert J. Lonigro,[1,6] Scott A. Tomlins,[1,3] Sooryanarayana Varambally,[1,3,6] Zhaohui S. Qin,[5] and Arul M. Chinnaiyan[1,2,3,4,6,*]

Reference

- Qin ZS, Yu J, Shen J, Maher CA, Hu M, Kalyana-Sundaram S, Yu J, Chinnaiyan AM. (2009) HPeak: An HMM-based Algorithm for Defining Read-enriched Regions in ChIP-Seq Data. *BMC Bioinformatics.* **11** 369.

  http://www.sph.umich.edu/csg/qin/HPeak/
- Choi H, Nesvizhskii A, Ghosh D, Qin ZS. (2009) Hierarchical Hidden Markov Model with Application to Joint Analysis of ChIP-chip and ChIP-seq Data. *Bioinformatics* **25** 1715-1721.

  http://sourceforge.net/projects/chipmeta/
- Hu M, Yu J, Taylor, JMG, Chinnaiyan AM, **Qin ZS.** (2010) On the Detection and Refinement of Transcription Factor Binding Sites Using ChIP-Seq Data. *Nucleic Acids Res.* **38** 2154-2167.

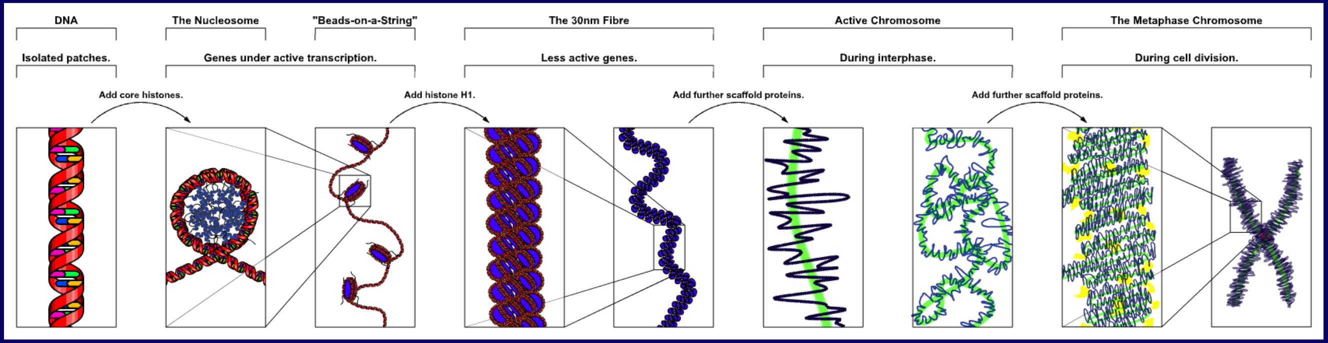
  http://www.sph.umich.edu/csg/qin/HMS/
- Hu M, Zhu Y, Taylor JMG, Liu JS, Qin ZS (2011). Using Poisson mixed-effects model to quantify exon-level gene expression in RNA-seq. *Bioinformatics.* **28** 63-68.

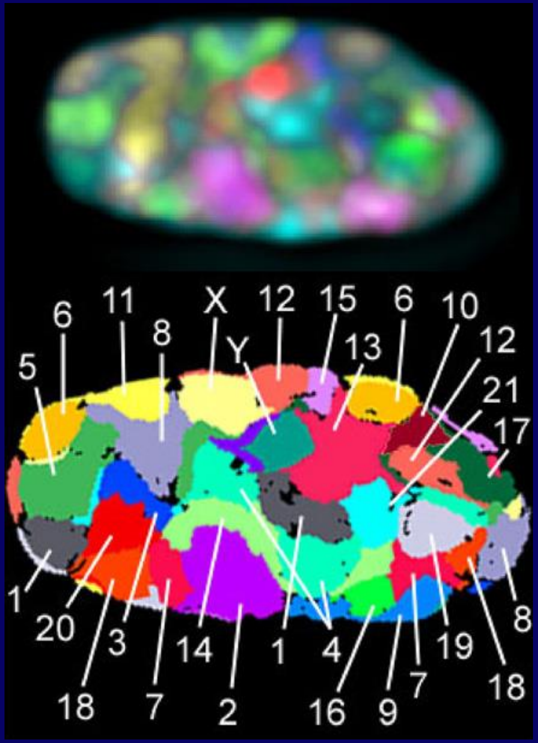
  http://www.stat.purdue.edu/~yuzhu/pome.html

Statistical model to infer chromosomal structures from Hi-C data

Chromosome folding

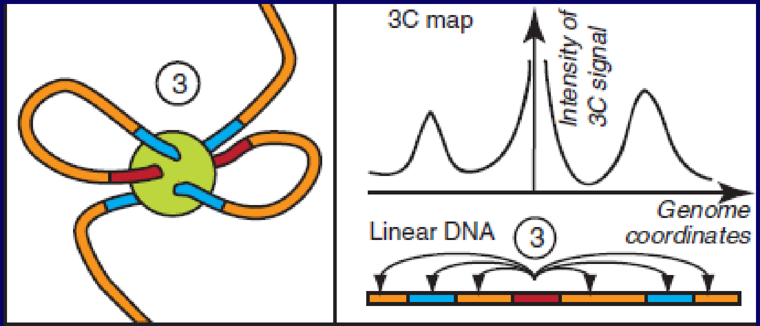
How can a two meter long polymer fit into a nucleus of ten micrometer (10^{-5} m) diameter?





http://en.wikipedia.org/wiki/Chromosome

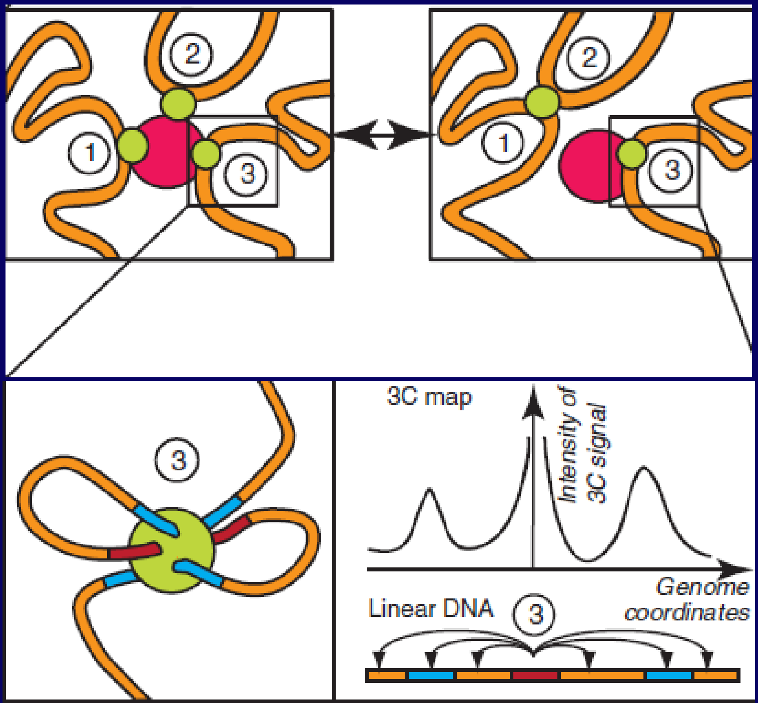
Chromosome Conformation Capture (3C)
Dekker et al. *Science* 2002



Fine scale: (0-kb)

Naumova and Dekker *J of Cell Science* 2010

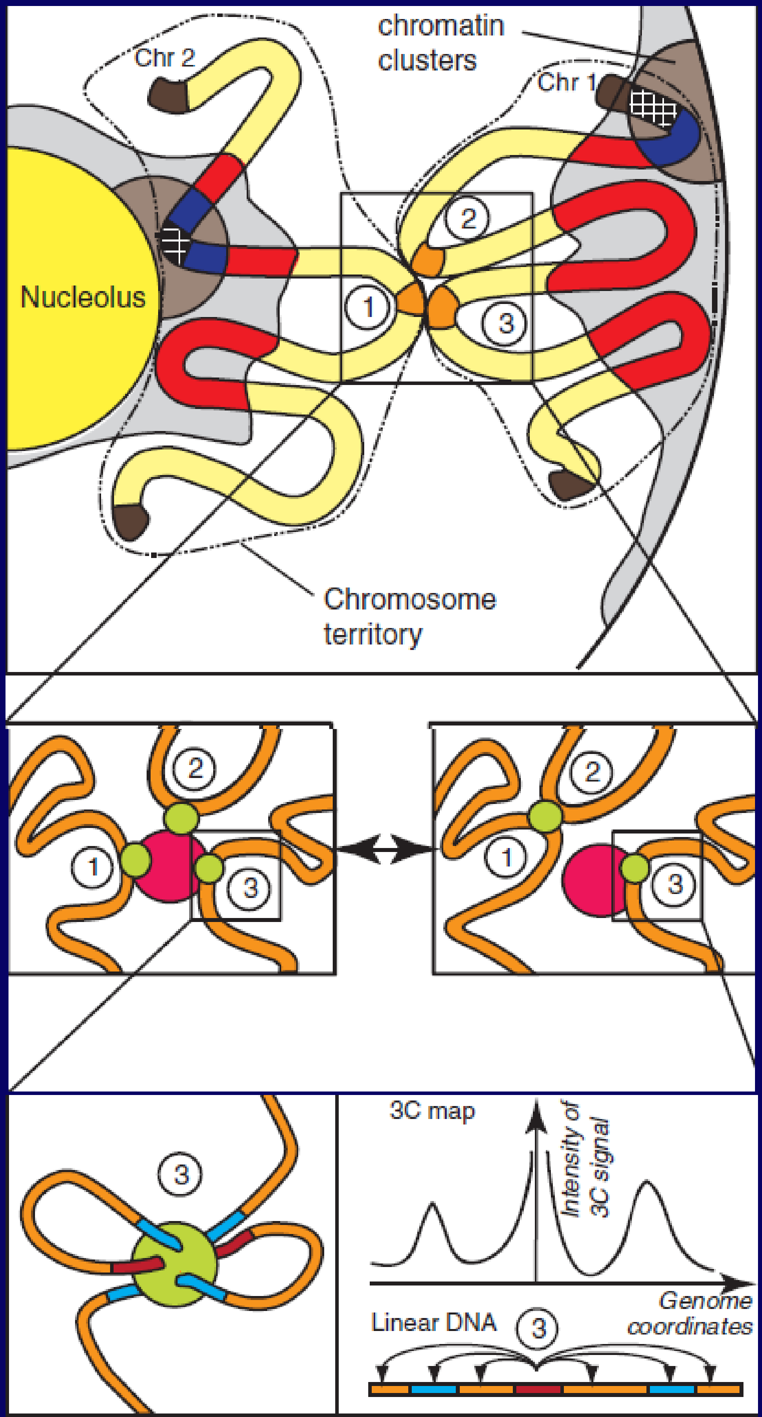
3C-on-chip/Circular 3C (4C)
5C



Intermediate: (0-Mb)

Fine scale: (0-kb)

Naumova and Dekker *J of Cell Science* 2010



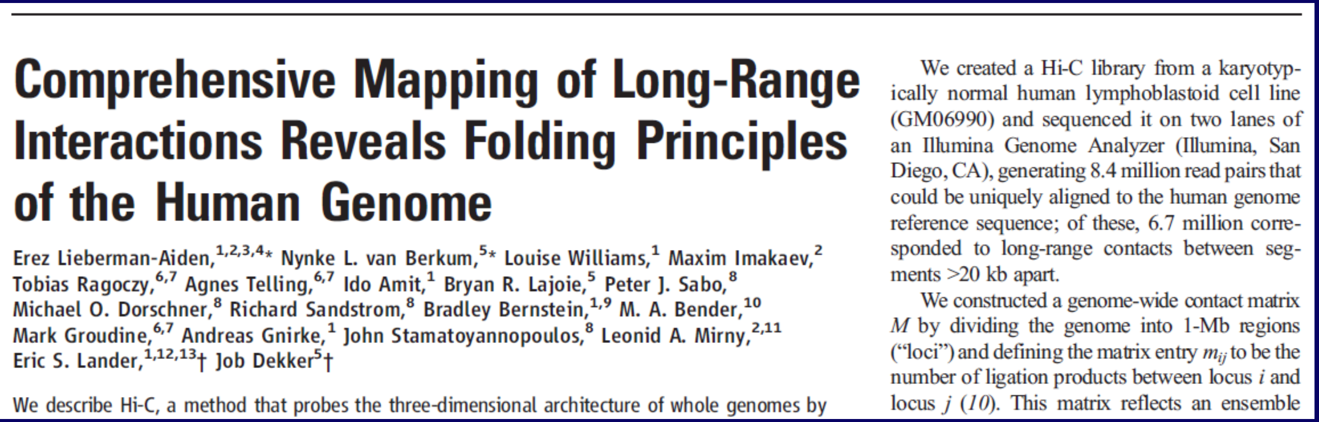
Whole genome

Intermediate: (0-Mb)

Fine scale: (0-kb)

Naumova and Dekker *J of Cell Science* 2010

9 October 2009 | $10
Science
nature biotechnology
VOLUME 30 NUMBER 1 JANUARY 2012
www.nature.com/naturebiotechnology
THE SCIENCE AND BUSINESS OF BIOTECHNOL
Arranging genomes in 3D
Pigeonpea staple sequenced
Deeper RNA sequencing
Cell
Volume 148
Number 3
February 3, 2012
www.cell.com
Genome in 3D
Hypoxia in Disease



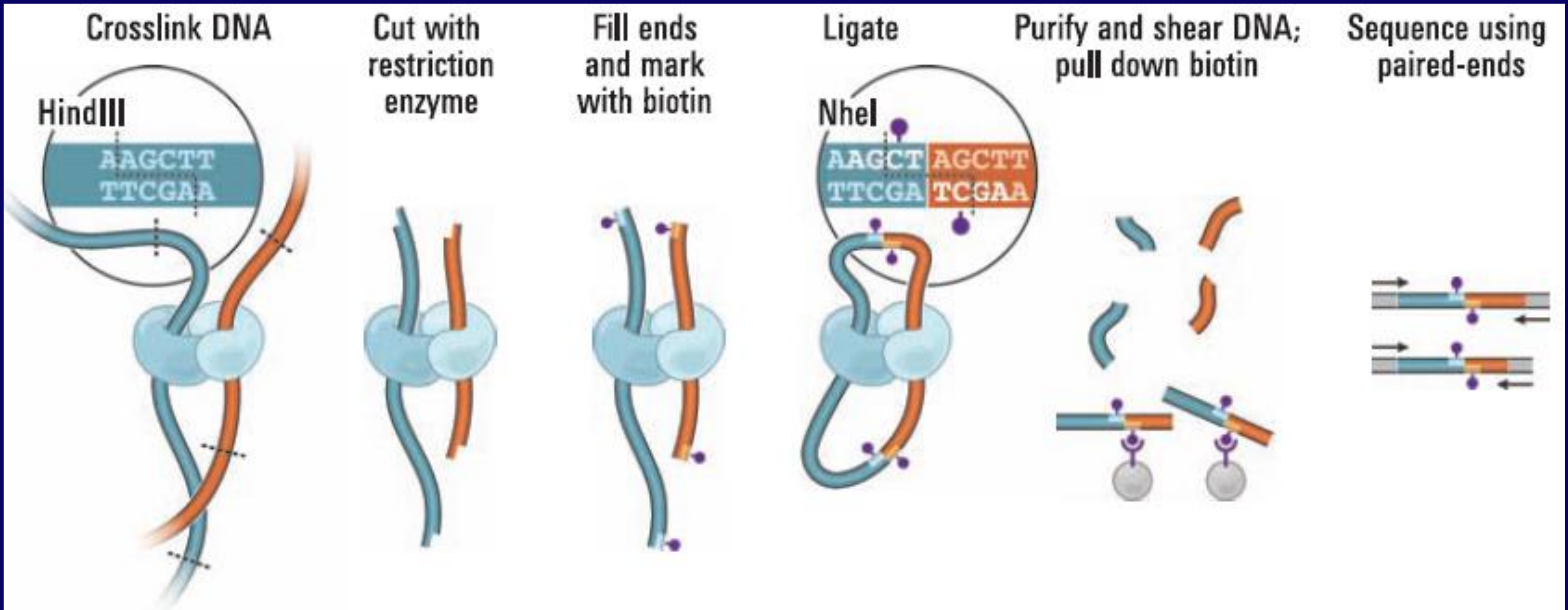
Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

Erez Lieberman-Aiden,[1,2,3,4]* Nynke L. van Berkum,[5]* Louise Williams,[1] Maxim Imakaev,[2] Tobias Ragoczy,[6,7] Agnes Telling,[6,7] Ido Amit,[1] Bryan R. Lajoie,[5] Peter J. Sabo,[8] Michael O. Dorschner,[8] Richard Sandstrom,[8] Bradley Bernstein,[1,9] M. A. Bender,[10] Mark Groudine,[6,7] Andreas Gnirke,[1] John Stamatoyannopoulos,[8] Leonid A. Mirny,[2,11] Eric S. Lander,[1,12,13]† Job Dekker[5]†

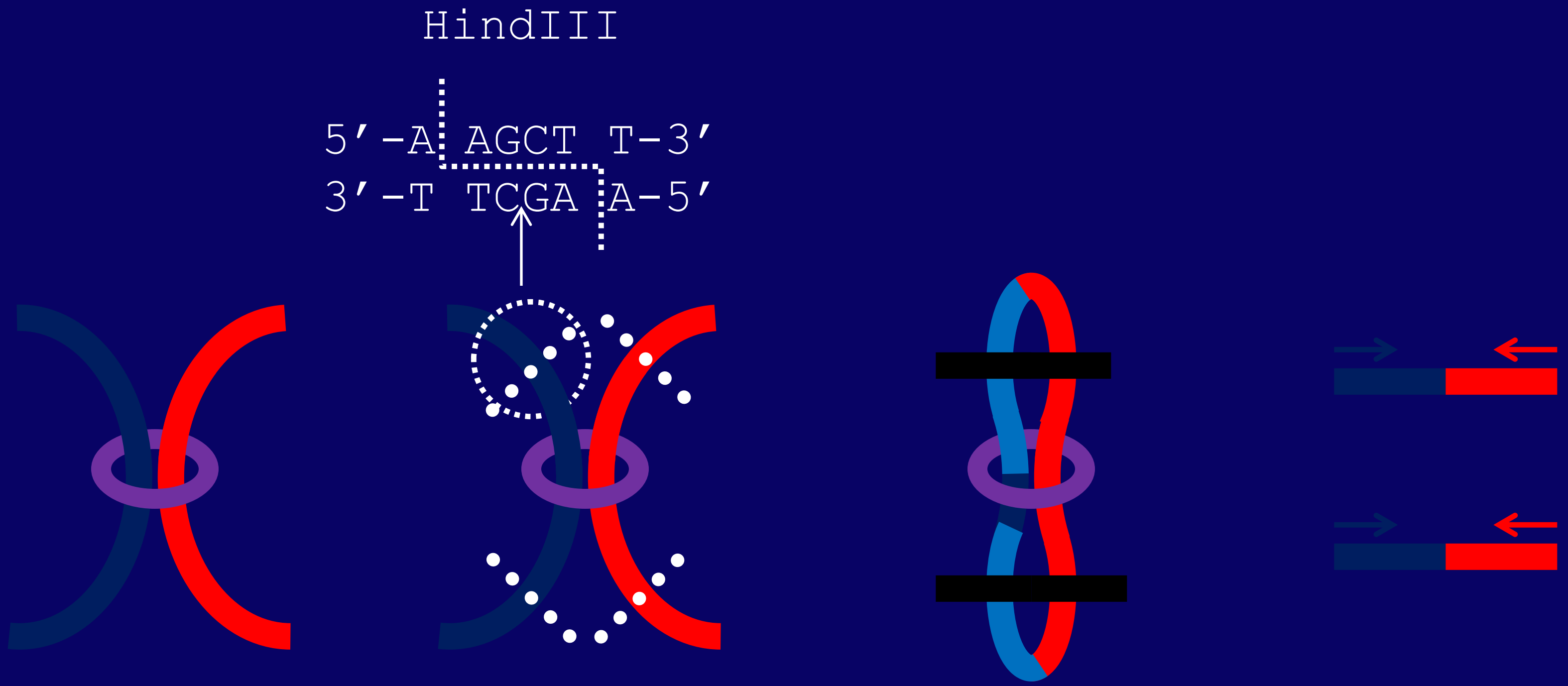
We describe Hi-C, a method that probes the three-dimensional architecture of whole genomes by

We created a Hi-C library from a karyotypically normal human lymphoblastoid cell line (GM06990) and sequenced it on two lanes of an Illumina Genome Analyzer (Illumina, San Diego, CA), generating 8.4 million read pairs that could be uniquely aligned to the human genome reference sequence; of these, 6.7 million corresponded to long-range contacts between segments >20 kb apart.

We constructed a genome-wide contact matrix M by dividing the genome into 1-Mb regions ("loci") and defining the matrix entry m_{ij} to be the number of ligation products between locus i and locus j (10). This matrix reflects an ensemble


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

Hi-C: one cell

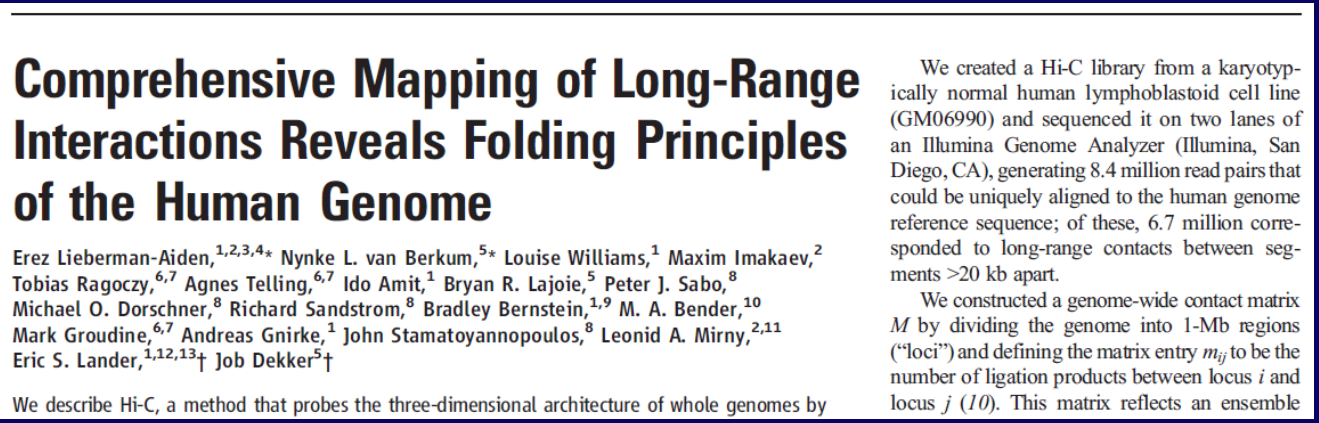


Cross-link DNA

Restriction enzyme cut

Ligation and shear

Paired-end sequencing



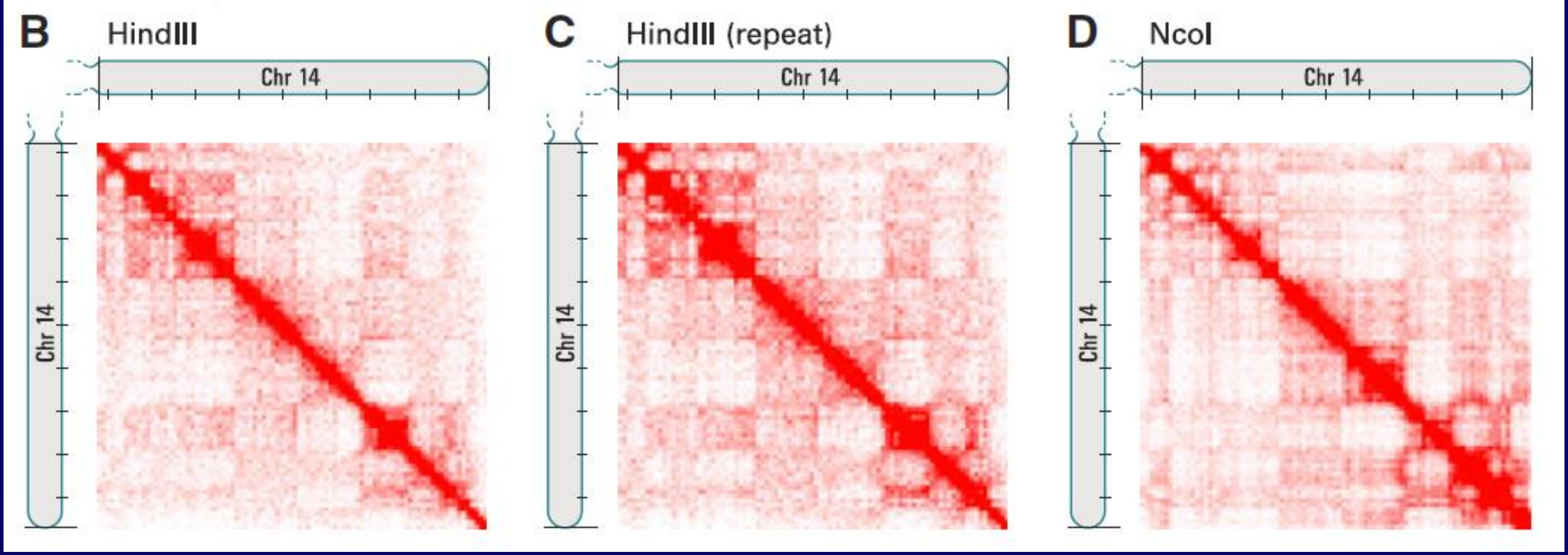
Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome

Erez Lieberman-Aiden,[1,2,3,4]* Nynke L. van Berkum,[5]* Louise Williams,[1] Maxim Imakaev,[2] Tobias Ragoczy,[6,7] Agnes Telling,[6,7] Ido Amit,[1] Bryan R. Lajoie,[5] Peter J. Sabo,[8] Michael O. Dorschner,[8] Richard Sandstrom,[8] Bradley Bernstein,[1,9] M. A. Bender,[10] Mark Groudine,[6,7] Andreas Gnirke,[1] John Stamatoyannopoulos,[8] Leonid A. Mirny,[2,11] Eric S. Lander,[1,12,13]† Job Dekker[5]†

We describe Hi-C, a method that probes the three-dimensional architecture of whole genomes by

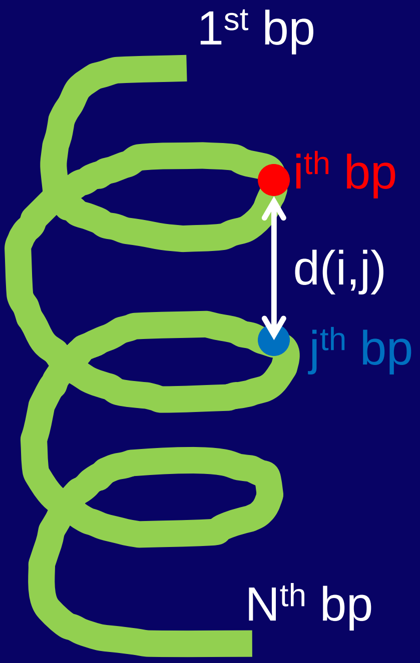
We created a Hi-C library from a karyotypically normal human lymphoblastoid cell line (GM06990) and sequenced it on two lanes of an Illumina Genome Analyzer (Illumina, San Diego, CA), generating 8.4 million read pairs that could be uniquely aligned to the human genome reference sequence; of these, 6.7 million corresponded to long-range contacts between segments >20 kb apart.

We constructed a genome-wide contact matrix *M* by dividing the genome into 1-Mb regions ("loci") and defining the matrix entry m_{ij} to be the number of ligation products between locus *i* and locus *j* (*10*). This matrix reflects an ensemble

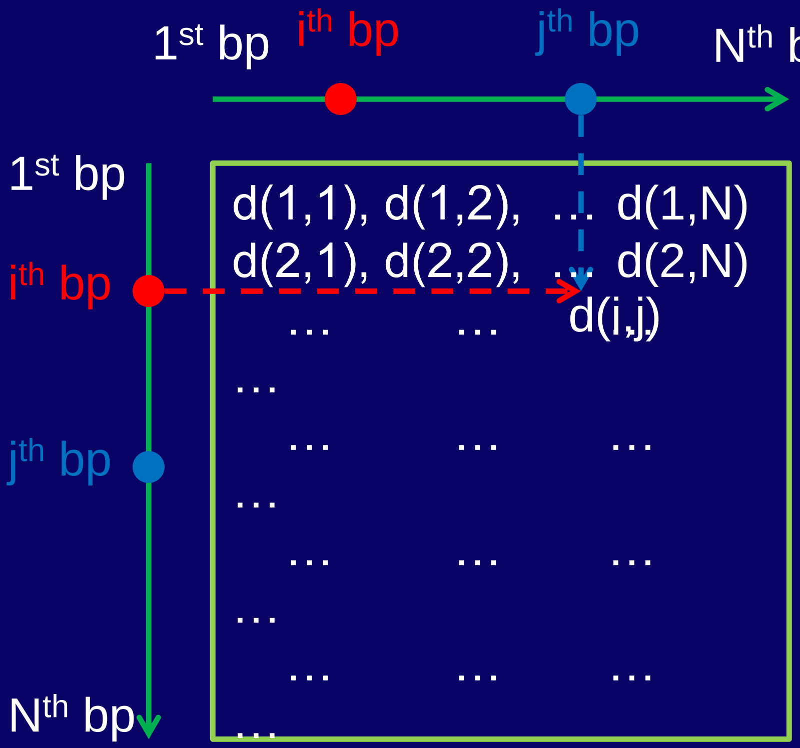


	chr1	chr2	chr3	chr4	chr5	chr6	chr7	chr8	chr9	chr10	chr11	chr12	chr13	chr14	chr15	chr16	chr17	chr18	chr19	chr20	chr21	chr22	chrX	chrY
chr1	2242	788	611	388	557	617	705	412	471	538	681	536	157	268	409	492	542	176	635	327	164	502	221	11
chr2	0	860	312	199	292	345	373	253	241	242	354	272	82	140	224	238	293	92	317	193	71	245	101	6
chr3	0	0	621	145	237	255	281	204	186	227	251	206	49	94	160	181	238	65	244	133	55	193	101	3
chr4	0	0	0	277	148	130	189	114	101	127	165	124	38	83	103	110	144	48	139	78	41	128	58	1
chr5	0	0	0	0	622	212	263	170	168	176	261	204	50	91	161	173	223	65	226	105	50	187	82	3
chr6	0	0	0	0	0	731	317	207	204	199	256	222	62	127	174	193	281	73	244	150	59	198	95	4
chr7	0	0	0	0	0	0	806	197	216	241	315	232	67	150	206	232	267	83	281	147	76	227	95	8
chr8	0	0	0	0	0	0	0	434	130	164	210	170	35	86	135	150	155	66	180	94	43	147	79	4
chr9	0	0	0	0	0	0	0	0	517	163	210	188	43	76	117	157	196	49	196	91	36	175	58	1
chr10	0	0	0	0	0	0	0	0	0	482	228	197	53	83	144	151	201	66	226	104	44	173	68	6
chr11	0	0	0	0	0	0	0	0	0	0	872	238	58	138	176	217	257	95	289	174	58	221	118	8
chr12	0	0	0	0	0	0	0	0	0	0	0	607	63	105	134	191	236	60	210	143	57	160	82	6
chr13	0	0	0	0	0	0	0	0	0	0	0	0	110	27	47	44	59	11	72	20	10	37	15	1
chr14	0	0	0	0	0	0	0	0	0	0	0	0	0	242	85	78	98	29	113	62	34	83	46	0
chr15	0	0	0	0	0	0	0	0	0	0	0	0	0	0	437	114	181	45	172	93	46	128	61	2
chr16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	538	198	52	227	95	39	169	82	9
chr17	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	781	61	243	126	68	184	75	2
chr18	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	134	64	42	23	66	34	8
chr19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	825	143	63	207	89	3
chr20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	386	42	105	64	5
chr21	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	144	45	23	1
chr22	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	521	74	4
chrX	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	170	0
chrY	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2

Hi-C Data Representation



3D chromosomal structure



d(1,1), d(1,2), … d(1,N)
d(2,1), d(2,2), … d(2,N)
… … d(i,j)
…
… … …
…
… … …
…
… … …
…
… … …
…
… … …

Pair-wise spatial distance matrix

…
d(N,1),d(N,2), …d(N,N)

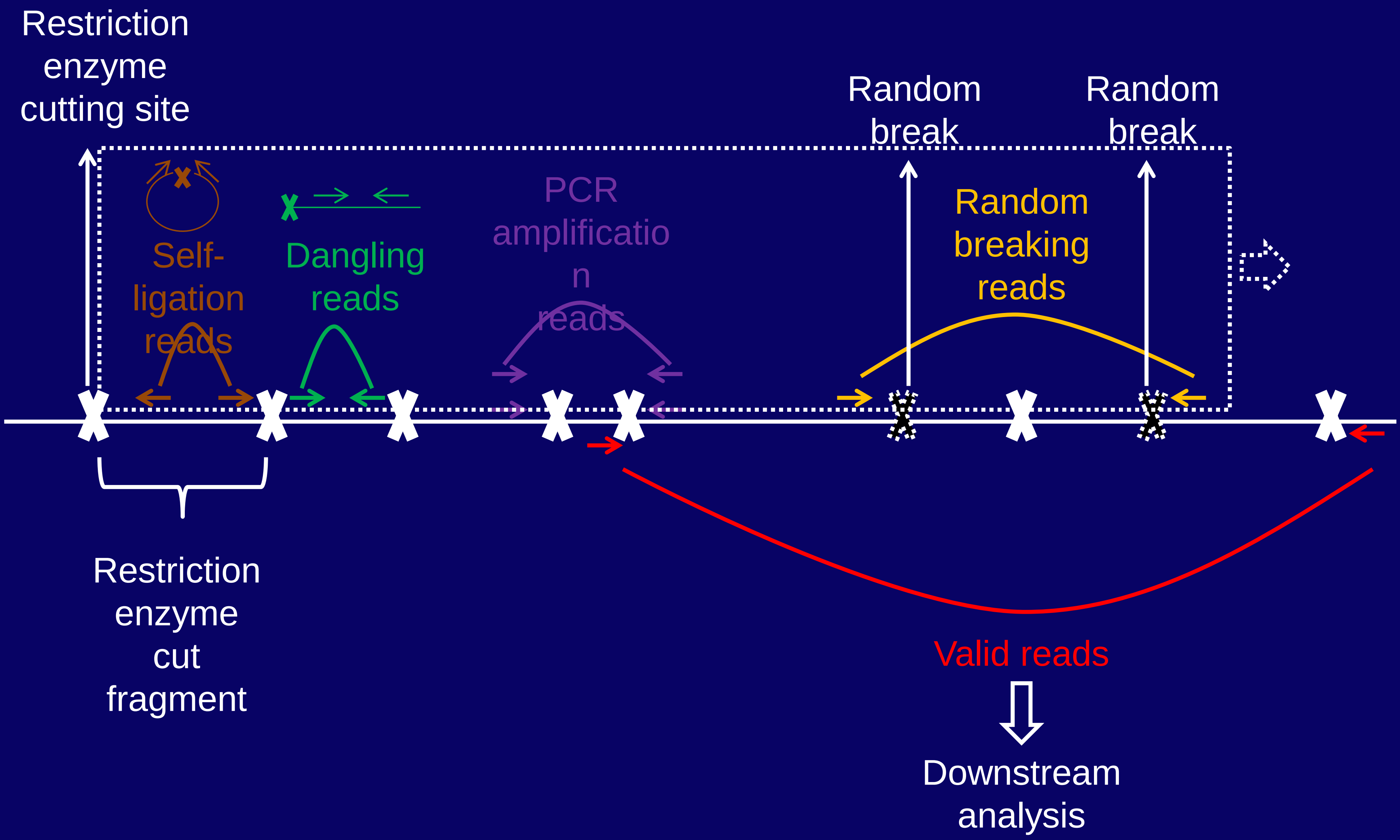
High resolution

High throughput

Challenges

- Quality control and pre-processing of the reads,
- Any bias in the data? and if so, how to normalize?
- Whether it is possible, and if so, how, to infer the 3-dimesnional chromosomal structure based on the Hi-C data?

Hi-C Data Preprocess



Restriction enzyme cutting site

Random break

Random break

Self-ligation reads

Dangling reads

PCR amplification reads

Random breaking reads

Restriction enzyme cut fragment

Valid reads

Downstream analysis

Methods for Hi-C Bias Reduction

- Normalization (equal ‘visibility’, no assumption on biases)

➢ Iterative correction and eigenvector decomposition (ICE) (Imakaev, et al, 2012)

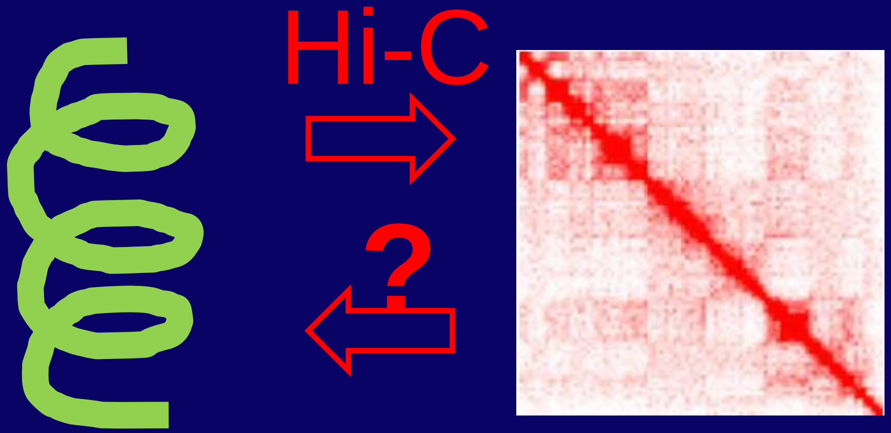
➢ Sequential component normalization (SCN)
(Cournac, et al, 2012)

- Correction (posit a statistical model on biases)

➢ Yaffe & Tanay’s method (Yaffe & Tanay, 2011)
Fragment level (4KB, 10^{12}), 420 parameters

➢ **HiCNorm (Hu et al, 2012)**
Any resolution level
1MB, 10^{6}, 3 parameters

3D structure prediction



- Challenges:
  - Sequencing uncertainties
  - Biases: enzyme, GC content, mappability

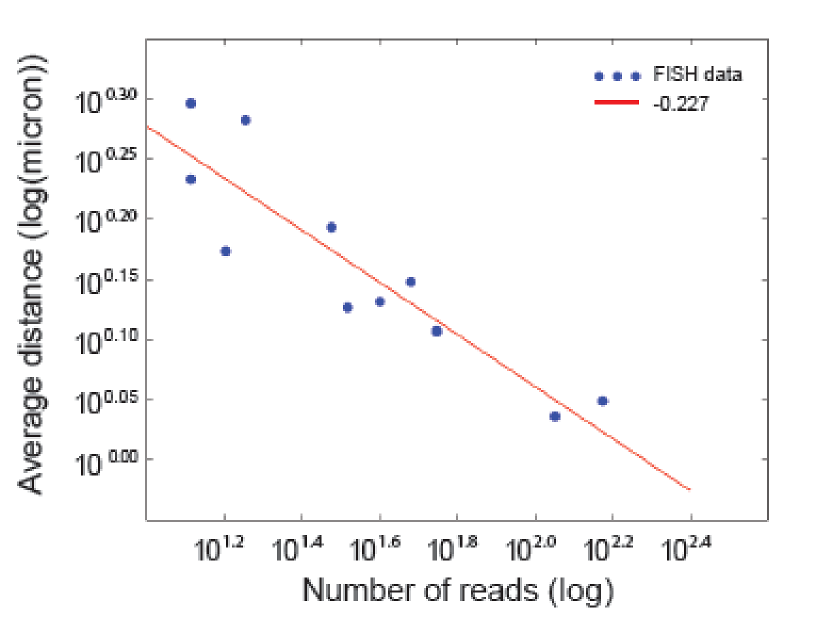
Yaffe and Tanay, 2011

What does the number mean?

- The Hi-C experiment is conducted on millions of cells,
- A captured pair-end read is from one cell,
- A number in the matrix (loci *i* and *j*) indicates the frequency of capture (link *i* and *j)* in the cell population,
- Do those numbers say anything about 3D distance?

Motivation and the key assumption

- Number of paired-end reads spanning the two loci is inversely proportional to the 3D spatial distance between them (obtained from fluorescence in situ hybridization(FISH)).



Lieberman-Aiden et al, 2009

Existing methods

- Optimizations-based method (Baù, et al, 2010, Duan, et al, 2010)
  - Biophysical properties of chromatin fiber.
  - No consideration of systematic biases.
  - No statistical inference.

- Statistical method: MCMC5C (Rousseau et al, 2011)
  - Normal model for count data.
  - No consideration of systematic biases.

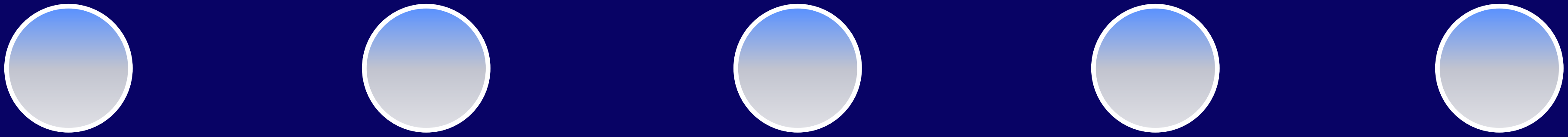
Model

ACGTAGCTAGATACTGTAGTACATCGATAGCGTAGTTTGGAACCTGAGGGTAAACCTGGAGGGGATCATG

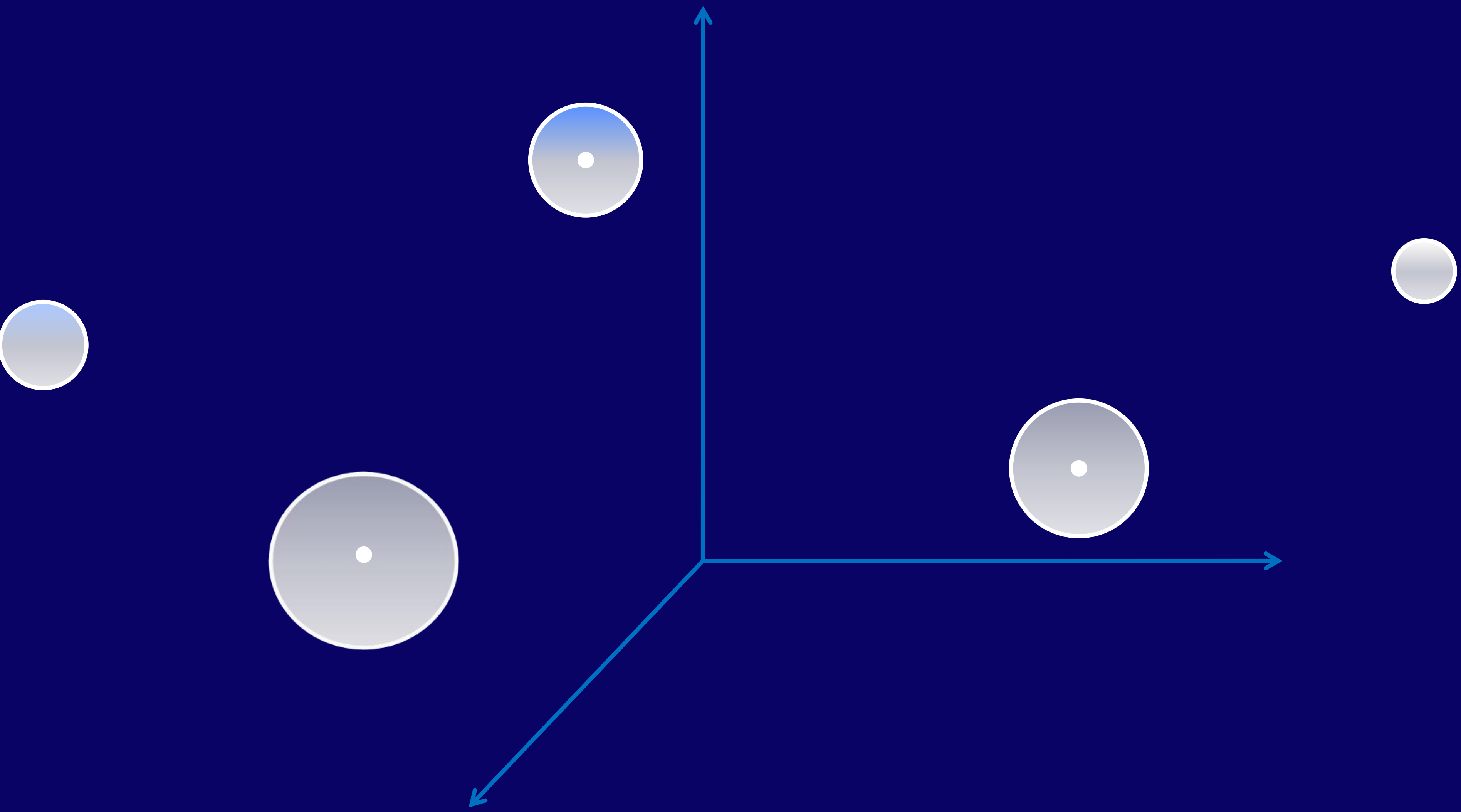
Model

ACGTAGCTAGATACT GTAGTACATCGATAG CGTAGTTTGGAACCT GAGGGTAAACCTGG AGGGGAT

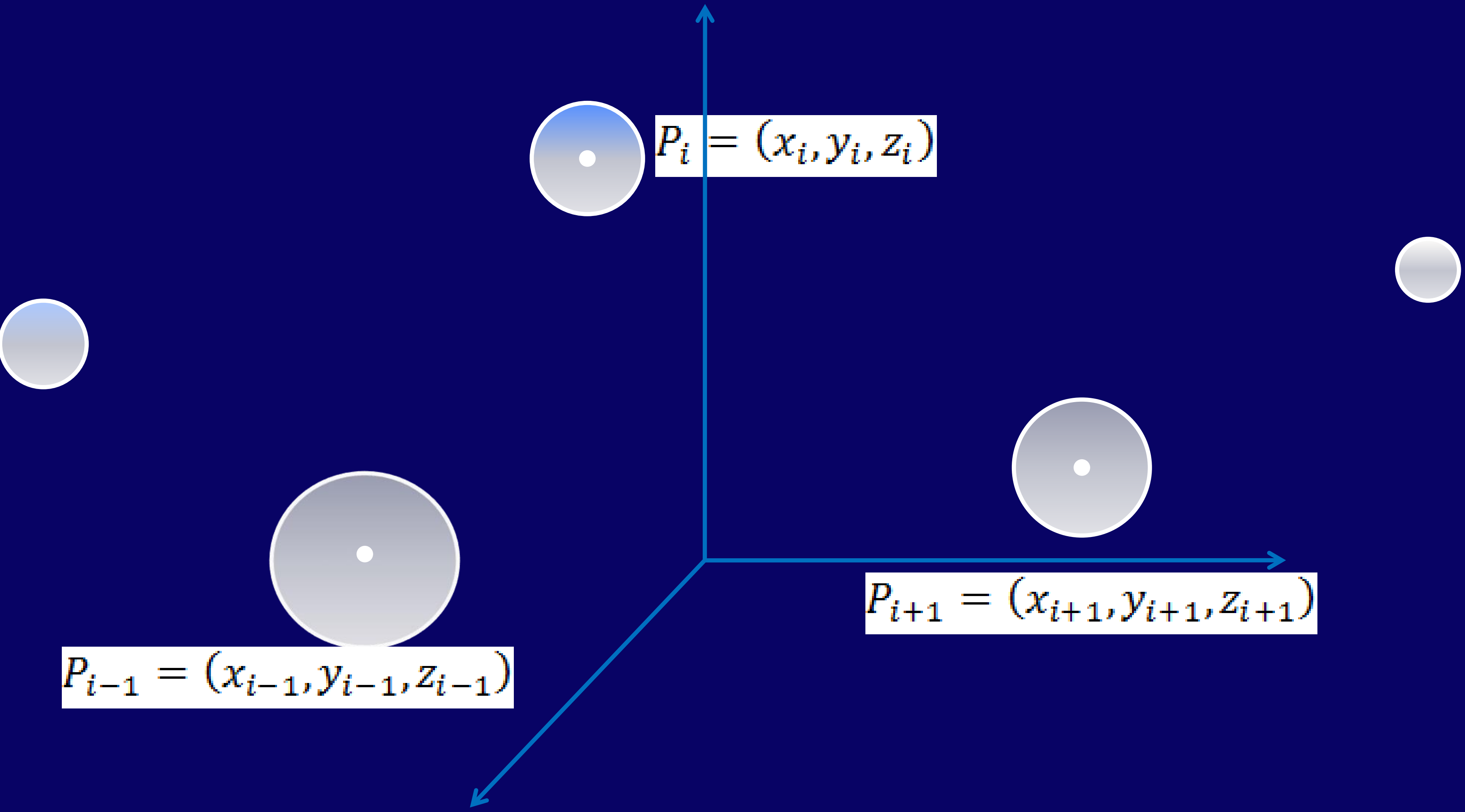
Model



Beads-on-string



Beads-on-string

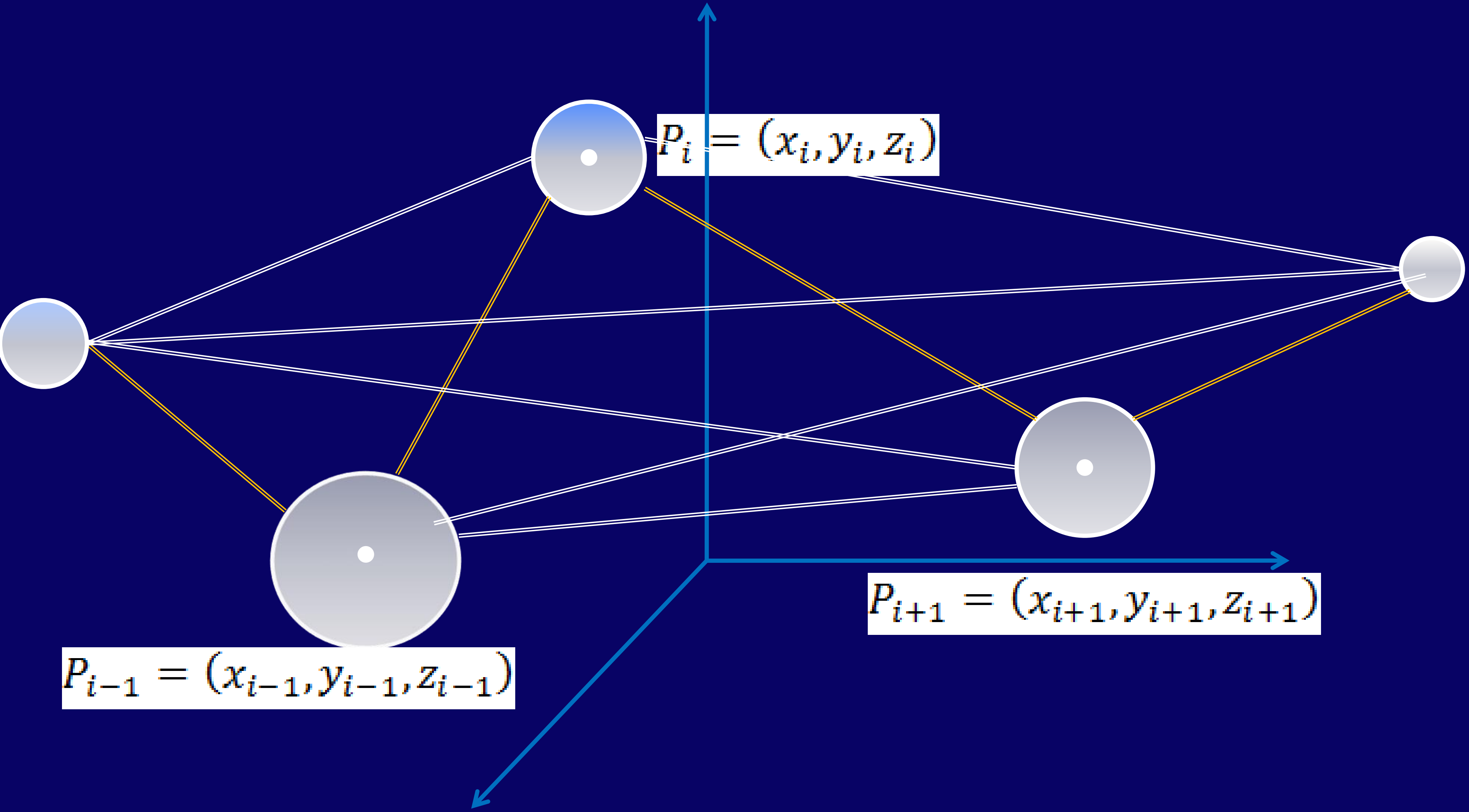


$P_i = (x_i, y_i, z_i)$

$P_{i+1} = (x_{i+1}, y_{i+1}, z_{i+1})$

$P_{i-1} = (x_{i-1}, y_{i-1}, z_{i-1})$

Beads-on-string

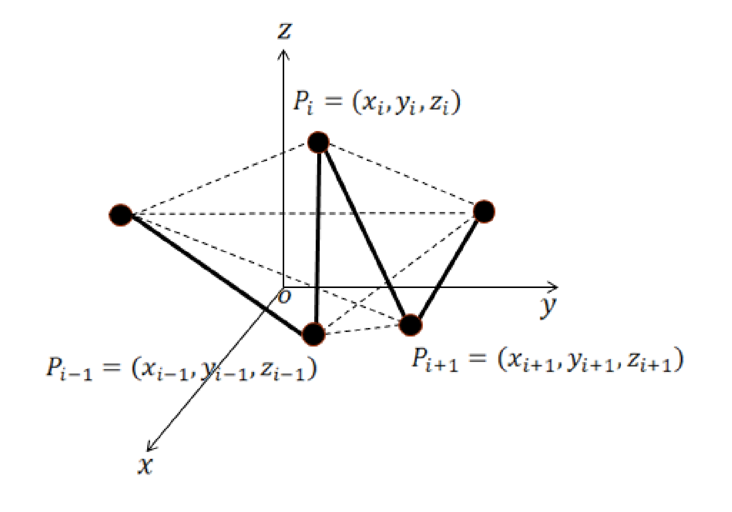


$P_i = (x_i, y_i, z_i)$

$P_{i+1} = (x_{i+1}, y_{i+1}, z_{i+1})$

$P_{i-1} = (x_{i-1}, y_{i-1}, z_{i-1})$

Bayesian statistical model



z
$P_i = (x_i, y_i, z_i)$
o
y
x
$P_{i-1} = (x_{i-1}, y_{i-1}, z_{i-1})$
$P_{i+1} = (x_{i+1}, y_{i+1}, z_{i+1})$

u_{ij} : number of reads between loci i and j.

d_{ij} : 3D Euclidian distance between loci i and j.

enz_i : number of enzyme cut site in locus i.

gcc_i : mean GC content in locus i.

map_i : mean mappability score in locus i.

$$u_{ij} \sim Poisson(\theta_{ij})$$

$$\log(\theta_{ij}) = \beta_0 + \beta_1 \log(d_{ij}) + \beta_{enz} \log(enz_i enz_j)$$

$$+ \beta_{gcc} \log(gcc_i gcc_j) + \beta_{map} \log(map_i map_j)$$

Bayesian Statistical Model

- Likelihood: $\binom{N}{2}$ data points, $3N + 5$ parameters

$$L(u_{ij}, 1 \leq i < j \leq N | x_i, y_i, z_i, 1 \leq i \leq N, \beta_0, \beta_1, \beta_e, \beta_g, \beta_m) = \prod_{1 \leq i < j \leq N} \frac{e^{-\theta_{ij}} \theta_{ij}^{u_{ij}}}{u_{ij}!}$$

$$\log(\theta_{ij}) = \beta_0 + \beta_1 \log\left(\sqrt{(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2}\right)$$
$$+ \beta_e \log(e_i e_j) + \beta_g \log(g_i g_j) + \beta_m \log(m_i m_j)$$

Statistical Inference

- Algorithm: Bayesian 3D constructor for Hi-C data (BACH)
  - Initialization 1: use Poisson regression to obtain the initial values of model parameters.
  - Initialization 2: use sequential important sampling to get the initial 3D chromosomal structure .
  - Refinement: use Gibbs sampler with hybrid Monte Carlo to refine the initial values for parameters.

SIS in BACH: Outline

- Goal: use sequential importance sampling to sequentially put *N* loci into 3D space, i.e. sample from:

$$\pi(x_i, y_i, z_i, 1 \le i \le N | u_{ij}, 1 \le i < j \le N)$$

- Bridging distributions:

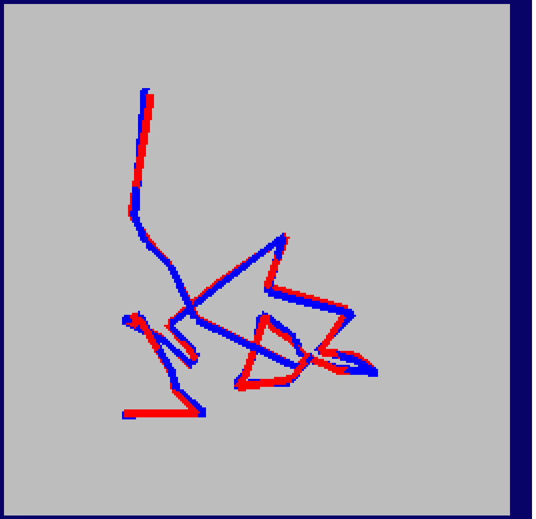
$$\pi_t(x_i, y_i, z_i, 1 \le i \le t | u_{ij}, 1 \le i < j \le t)$$

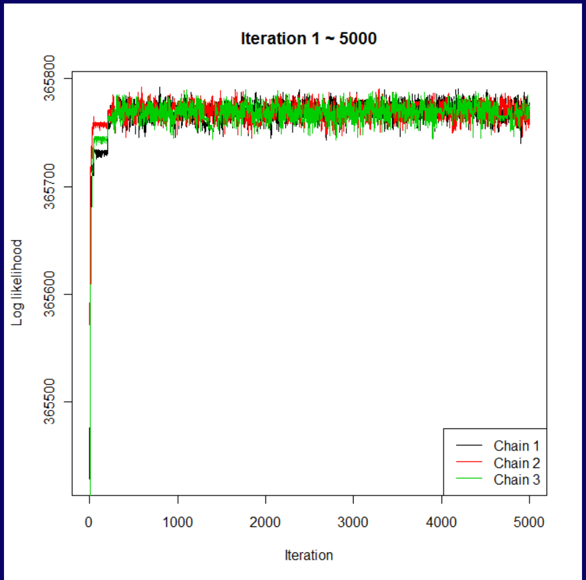
- Proposal distributions (given the previous *t*-1 loci, put the *t* th locus in to 3D space):

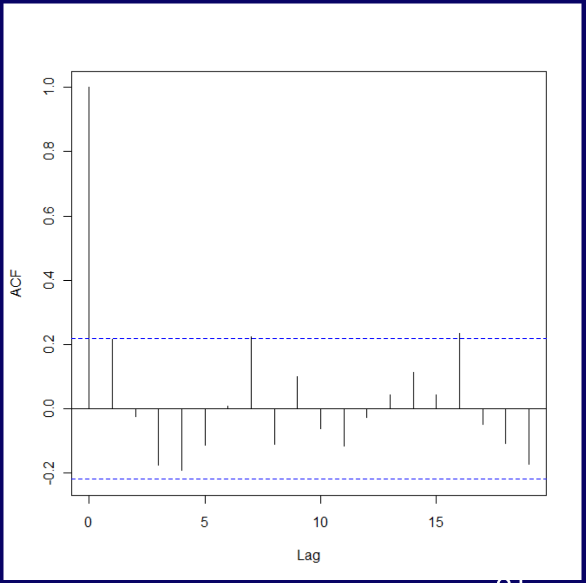
$$g_t(x_t, y_t, z_t | x_i, y_i, z_i, 1 \le i \le t - 1, u_{ij}, 1 \le i < j \le t)$$

Simulation study

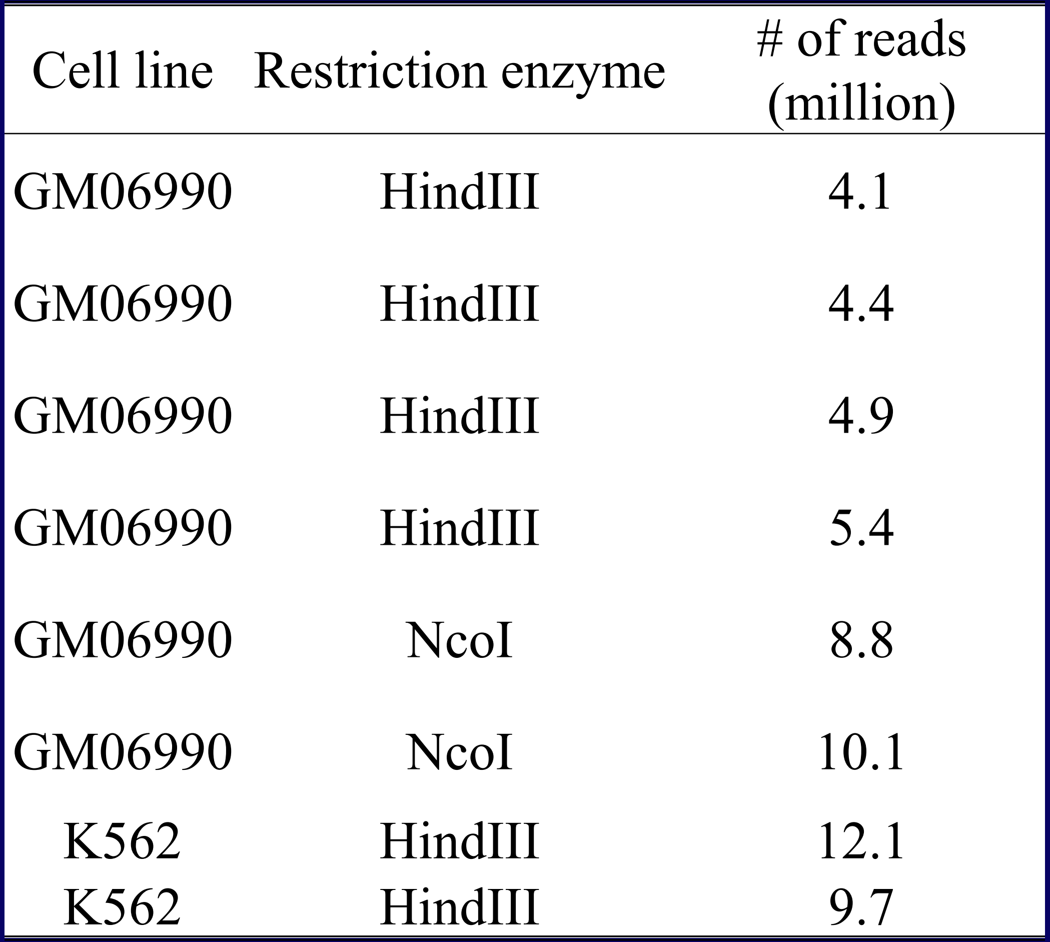
- Use random walk to simulate a 3D structure with 33 loci (red lines). Simulate Hi-C contact map from the posited model.
- Predicted 3D structure (blue lines) aligns well with true 3D structure (RMSD = 0.0091).





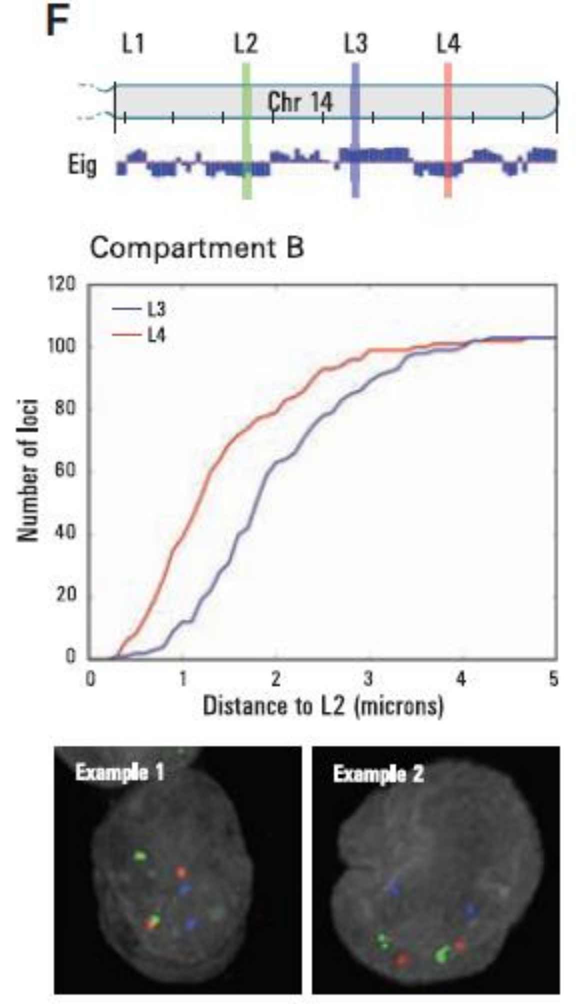


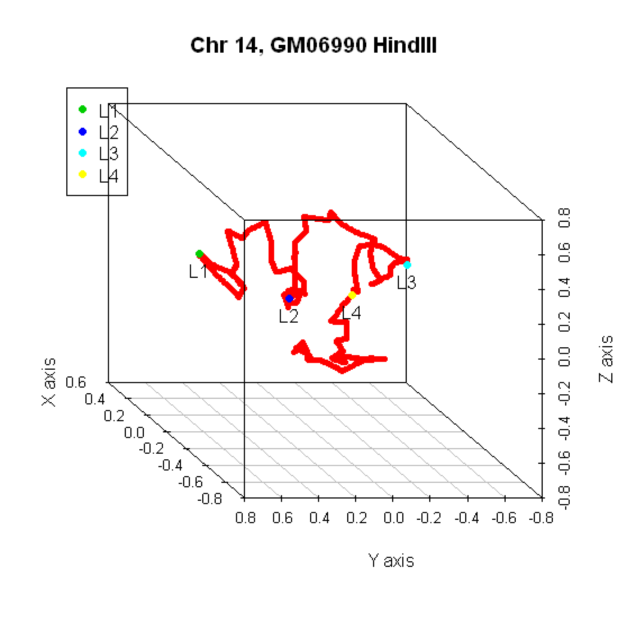
Human Hi-C data

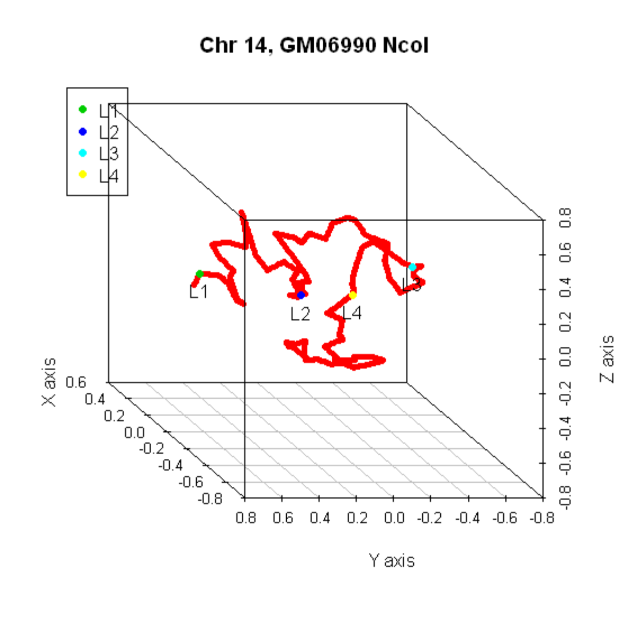


Cell line	Restriction enzyme	# of reads (million)
GM06990	HindIII	4.1
GM06990	HindIII	4.4
GM06990	HindIII	4.9
GM06990	HindIII	5.4
GM06990	NcoI	8.8
GM06990	NcoI	10.1
K562	HindIII	12.1
K562	HindIII	9.7

Real Hi-C data from Lieberman-Aiden et al. 2009

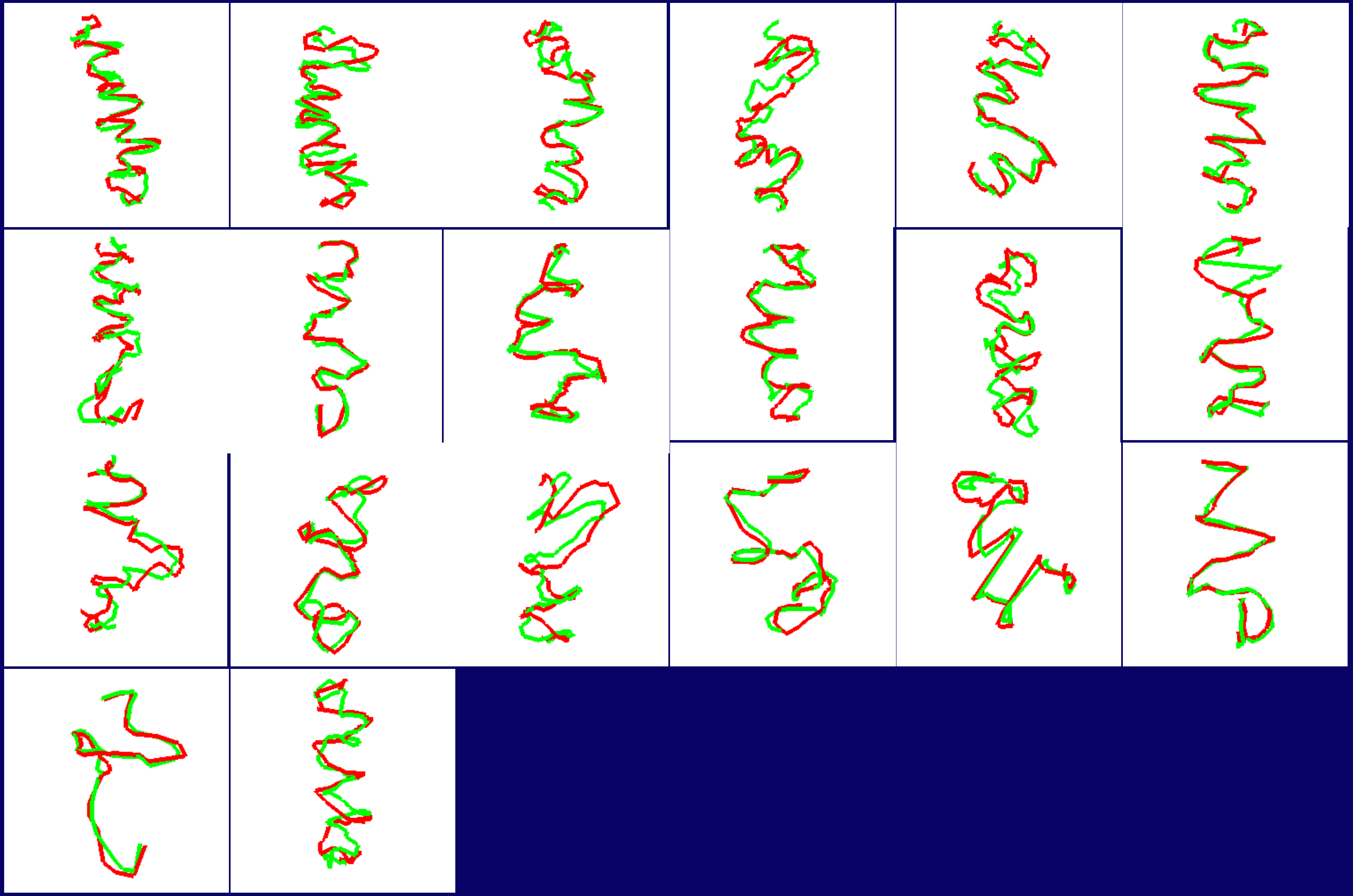






d(L2, L4) = 1.4042, d(L2, L3) = 1.9755, significant

mESC: Hind3 vs. Nco1



Whole Chromosome 3D Model

- Two compartments

➢ Compartment A: gene rich, active transcription

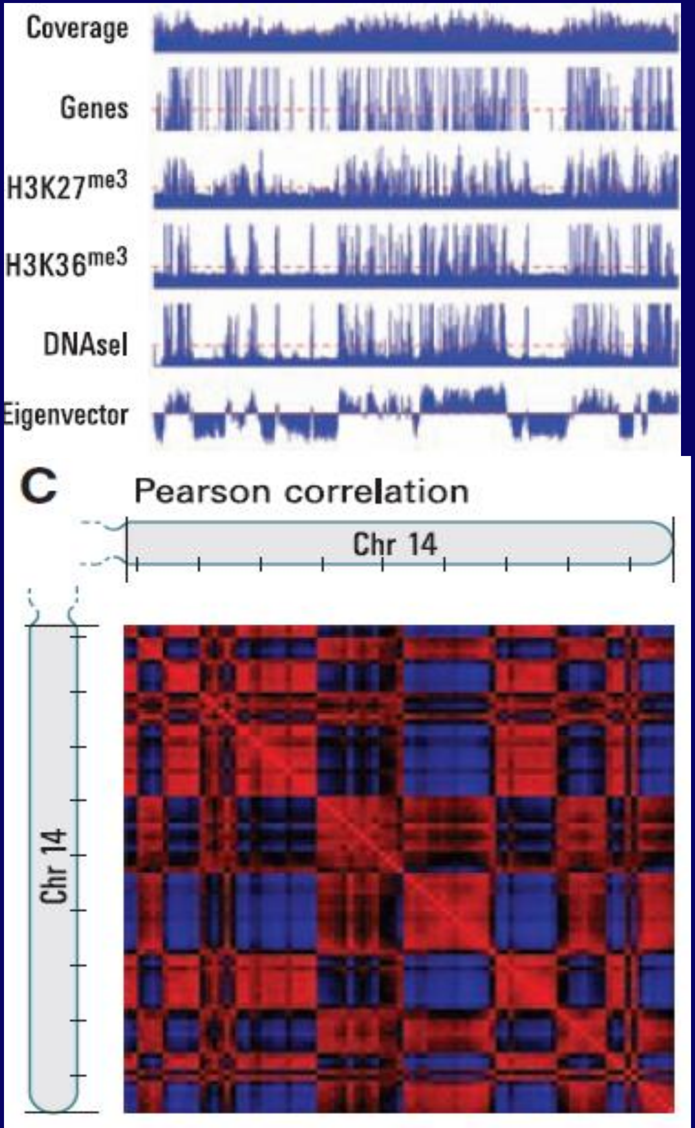
➢ Compartment B: gene poor, inactive transcription

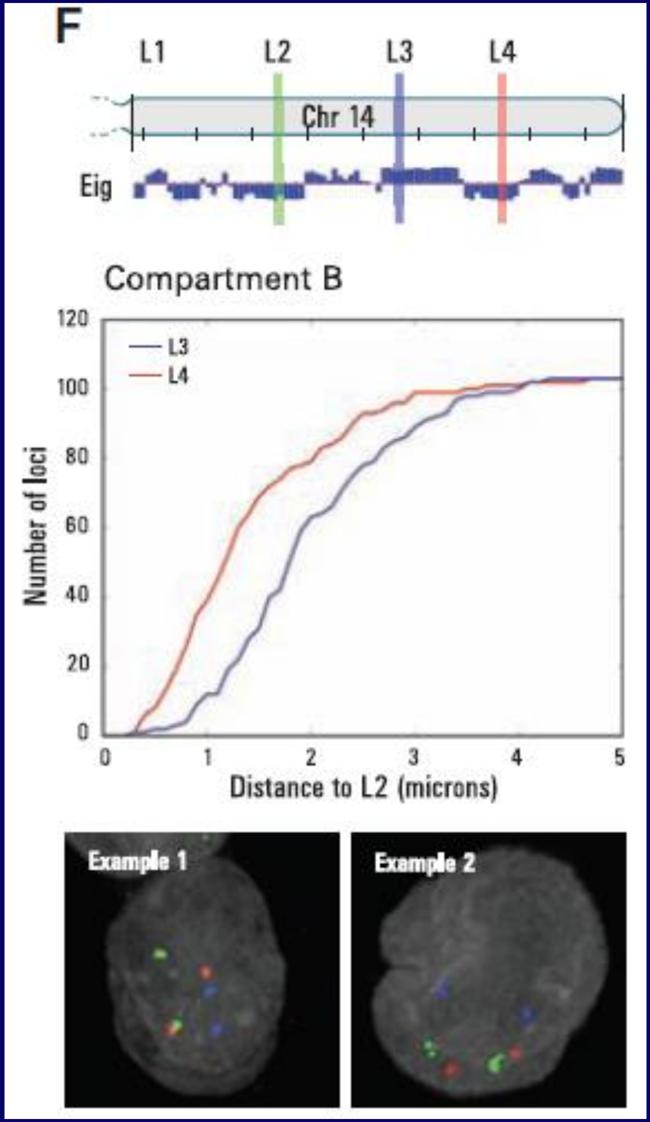
- Same compartment: strong chromatin interactions, spatially close

- Different compartments: weak chromatin interaction, spatially isolated

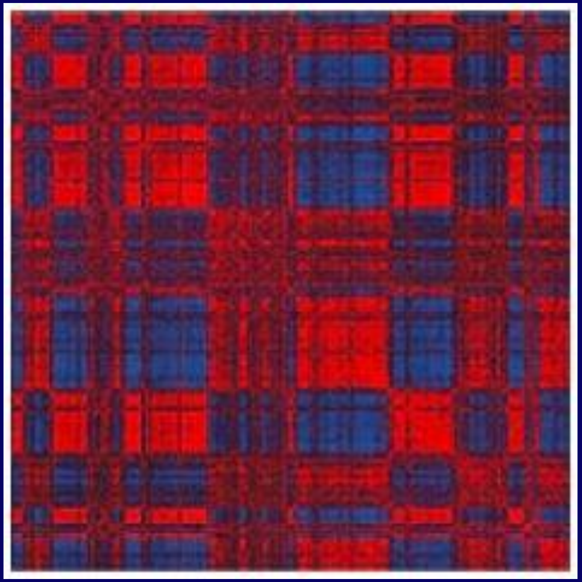
Lieberman-Aiden, et al, 2009

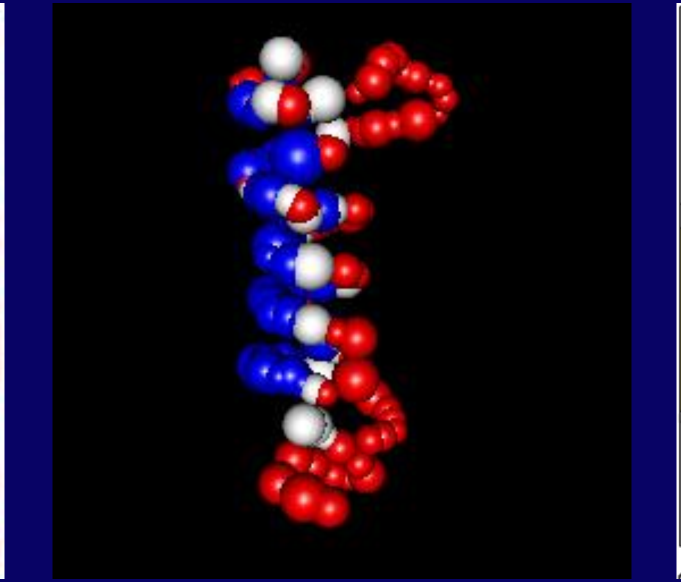
Two compartment model

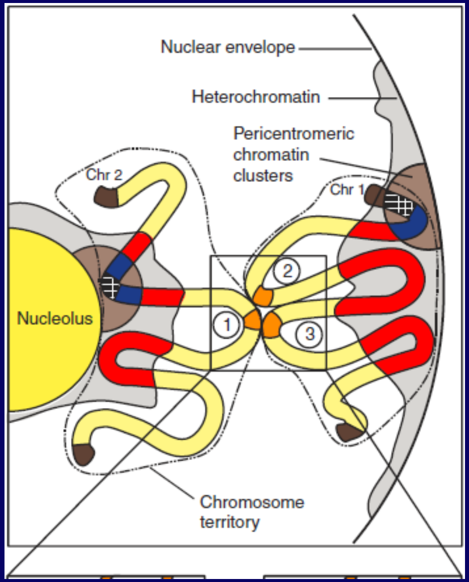




Whole Chromosome Model



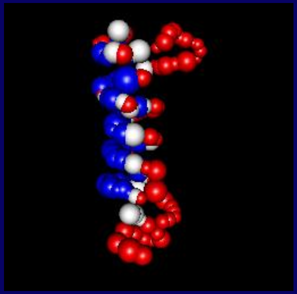




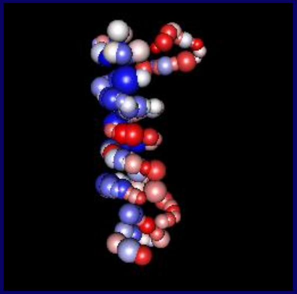
Lieberman-Aiden, et al, 2009
Naumova and Dekker, 2010

Other Features (Chromosome 2)

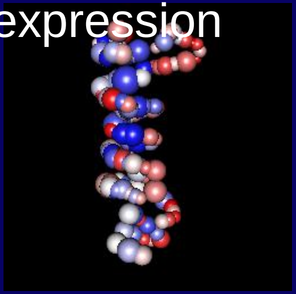
Compartment



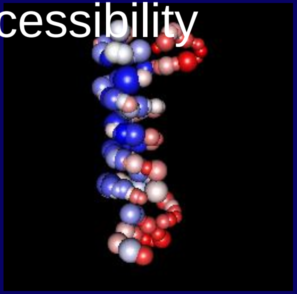
Gene density



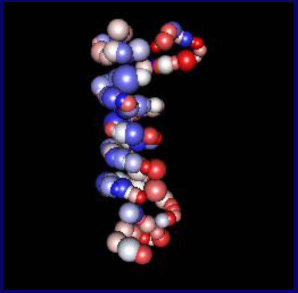
Gene expression



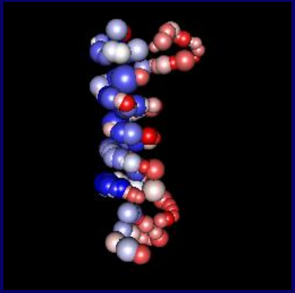
Chromatin accessibility



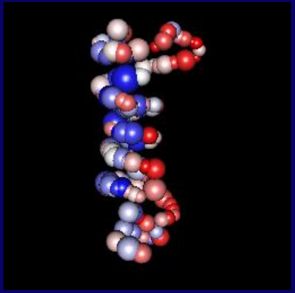
RNA polymerase II



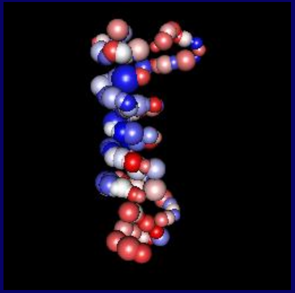
DNA replication time



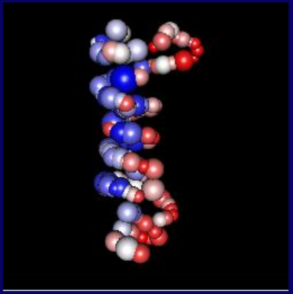
H3K36me3



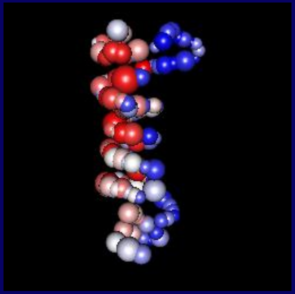
H3K27me3



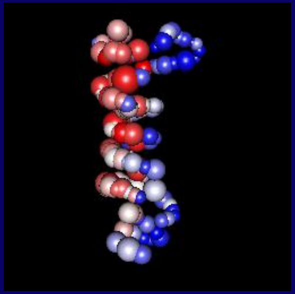
H3K4me3



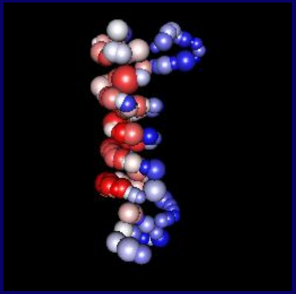
H3K9me3



H3K20me3



Lamina interaction



Conclusions

- BACH--Reconstruct chromosome 3D structures
- Remove systematic biases
- Consistent with FISH data
- Elongation of chromatin is highly associated with genetic/epigenetic features.
- Separation of compartments of A and B can be visualized.

More questions to be answered

- Is there a consensus? Or a dominant 3D chromosomal structure?
  - Completely random?
  - Mixture of distinct structures?
- Rigorous inference
  - Variance of the structure
- Computation

References

- Hu M, Deng K, Selvaraj S, Qin ZS, Ren B, Liu JS. (2012) HiCNorm: removing biases in Hi-C data via Poisson regression. *Bioinformatics*. In press.

  http://www.people.fas.harvard.edu/~junliu/HiCNorm/
- Hu M, Deng K, Qin ZS, Dixon J, Selvaraj S, Fang J, Ren B, Liu JS. (2012) Bayesian inference of three-dimensional chromosomal organization. *PLoS Computational Biology*. In press.

  http://www.people.fas.harvard.edu/~junliu/BACH/
- Hou C, Li L, Qin ZS, Corces, VG. (2012) Gene Density, Transcription and Insulators Contribute to the Partition of the Drosophila Genome into Physical Domains. *Mol Cell*. **48** 471-484 (with preview article of Xu and Felsenfeld (2012) Order from Chaos in the Nucleus. *Mol Cell 48. 327-328).*

  .
- Dixon JR, Selvaraj S, Yue F, Kim A, Li Y, Shen Y, Hu M, Liu JS and Ren B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature* , 485, 376-380.

Hu M, Deng K, Qin ZS, Liu JS (2013) Understanding spatial organizations of chromosomes via statistical analysis of Hi-C data. Quantitative Biology **1**. 156-174.

Acknowledgements

Ming Hu

Ke Deng

Jun S. Liu

Li Li

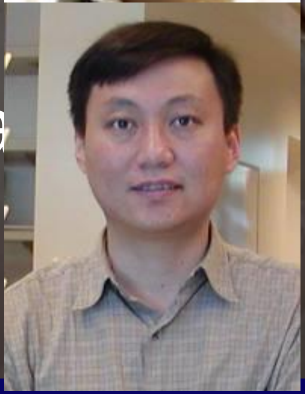
Chunhui Hou

Victor Corces

Jesse Dixon

Siddarth Selvara

Bing Ren



Thank You

Questions: zhaohui.qin@emory.edu