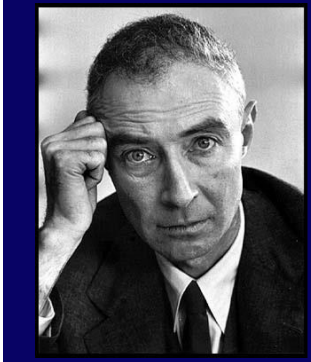
ChIP-seq and its analysis

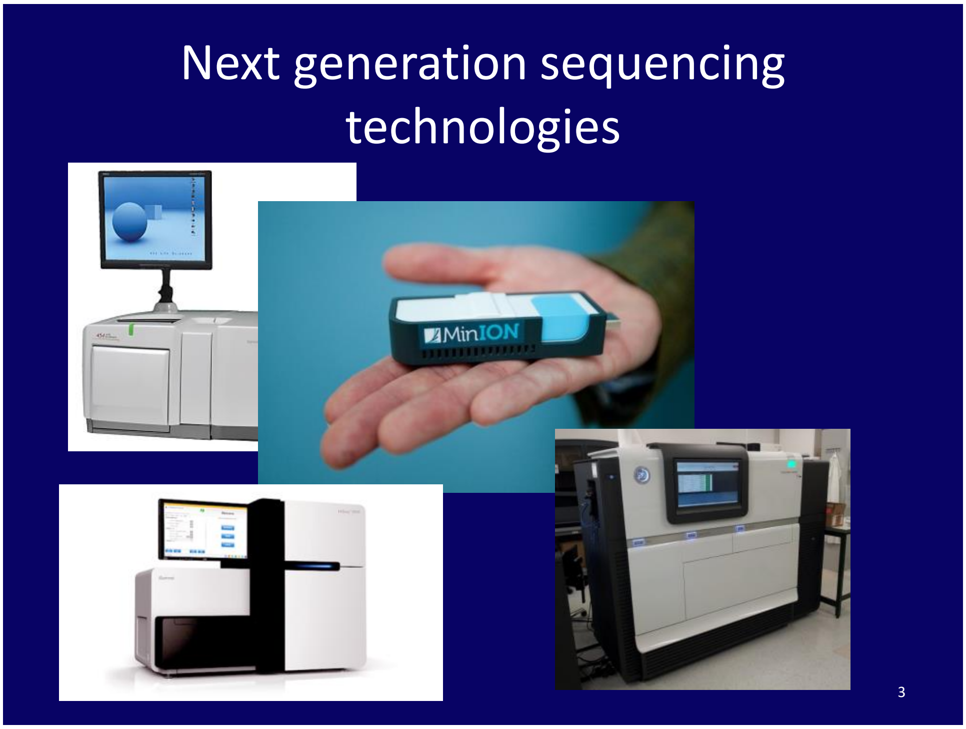
Steve Qin
Department of Biostatistics
and Bioinformatics
Rollins School of Public Health
Emory University

EMORY
UNIVERSITY



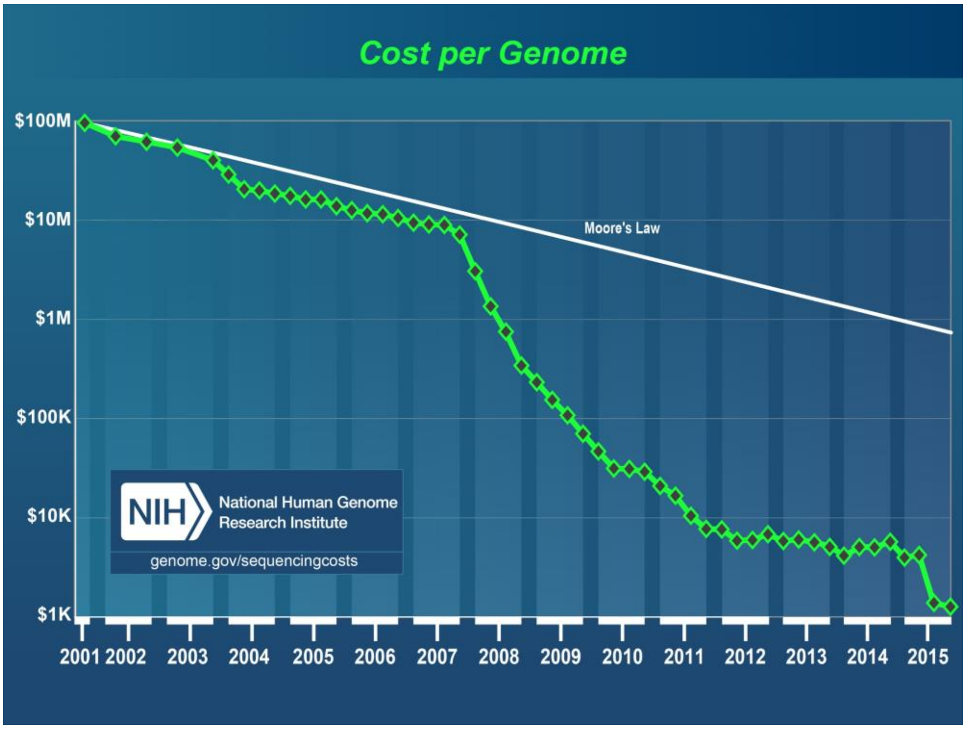
"… deep things in science are not found because they are useful; they are found because it was possible to find them"

-- Robert Oppenheimer



Next generation sequencing technologies

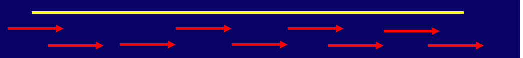
3



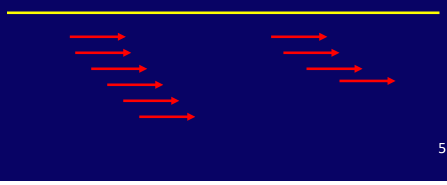
Cost per Genome

Different strategies of using sequencing technologies

- DNA-seq:
  - Whole genome sequencing
  - Uniform coverage
  - Flat

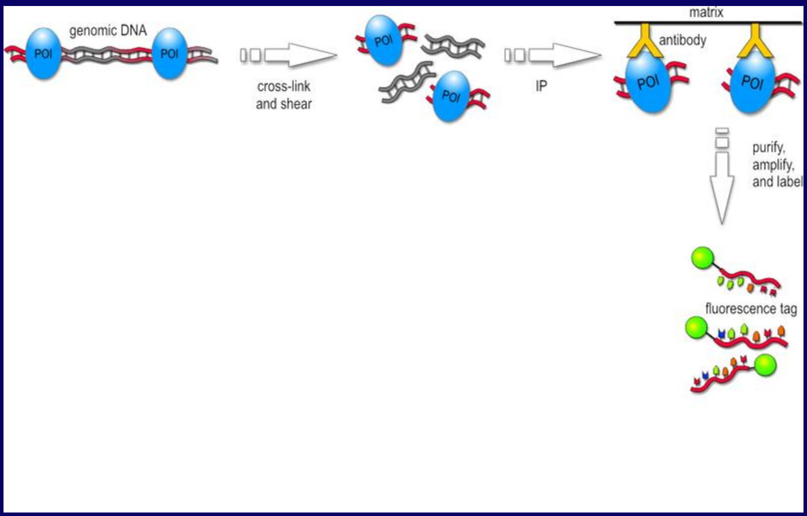


- ChIP-seq, Dnase-seq, ATAC-seq…
  - Capture-based
  - Genome-wide select sequencing
  - Subset
  - Peaky



5

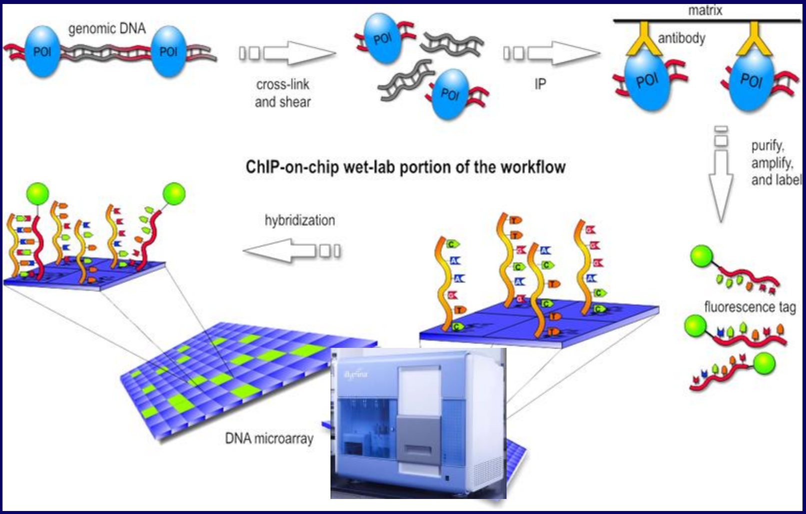
Chromatin Immunoprecipitation



ChIP-chip on Wikipedia

6

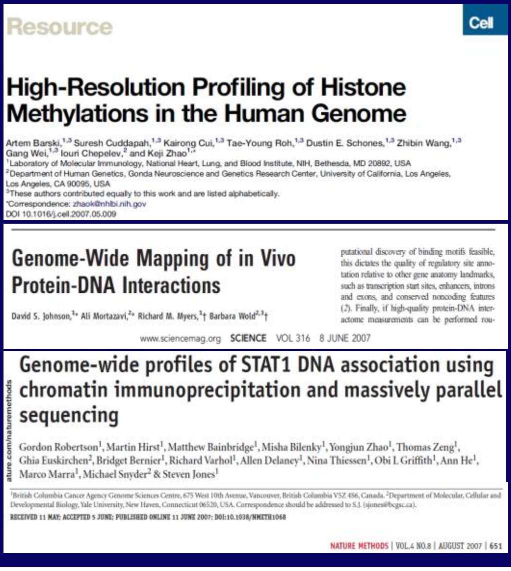
ChIP-chip and ChIP-Seq technologies



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

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]

ature.com/naturemethods

[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

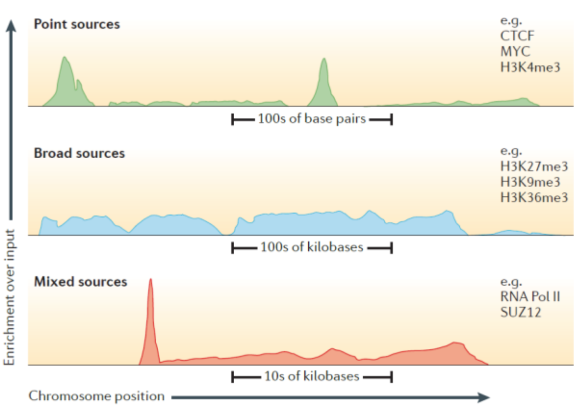
Using model-based methods to analyze ChIP-seq data

Outline

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

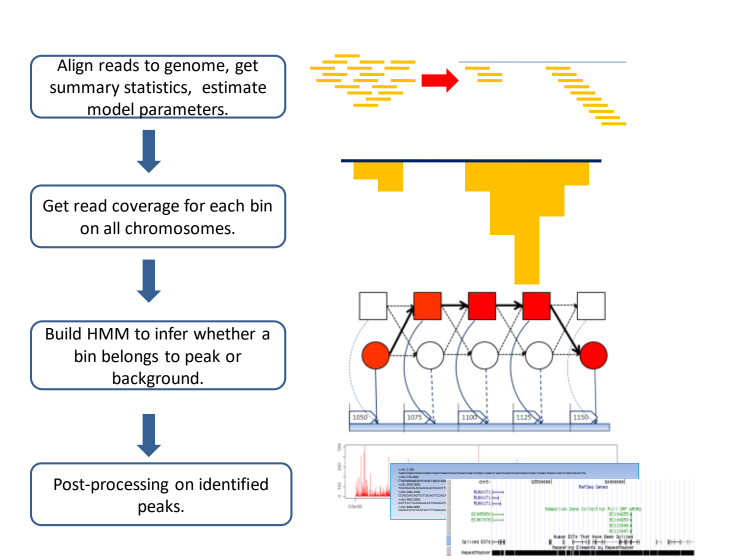
Peak calling tools

- MACS
- HOMER
- cisGenome
- PeakSeq
- Hpeak



11

HPeak algorithm



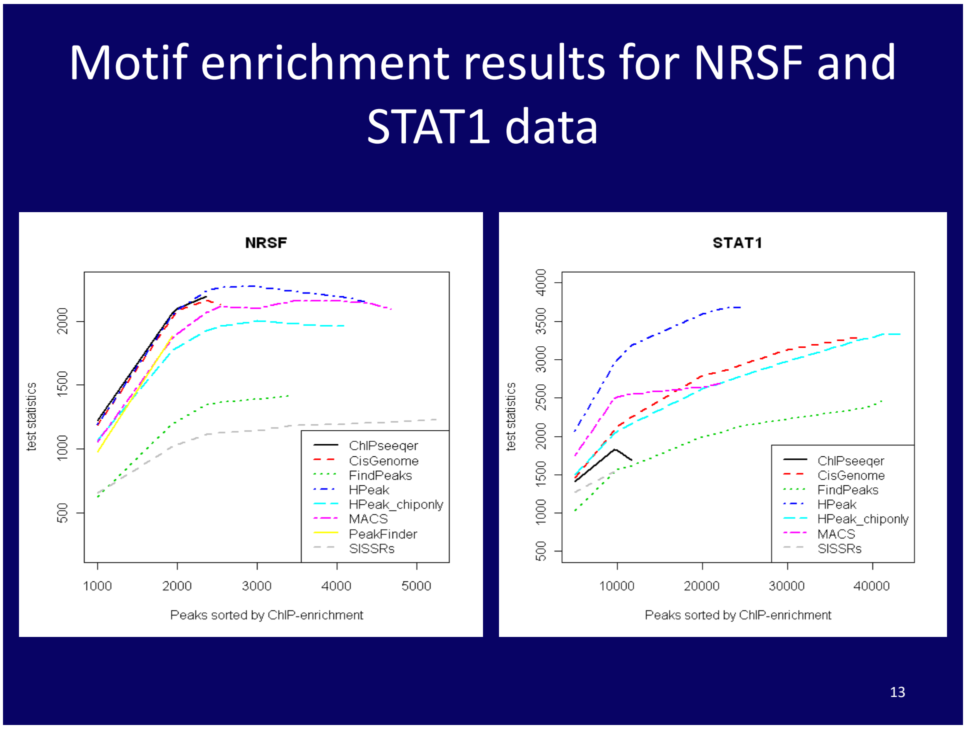
Align reads to genome, get summary statistics, estimate model parameters.

Get read coverage for each bin on all chromosomes.

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

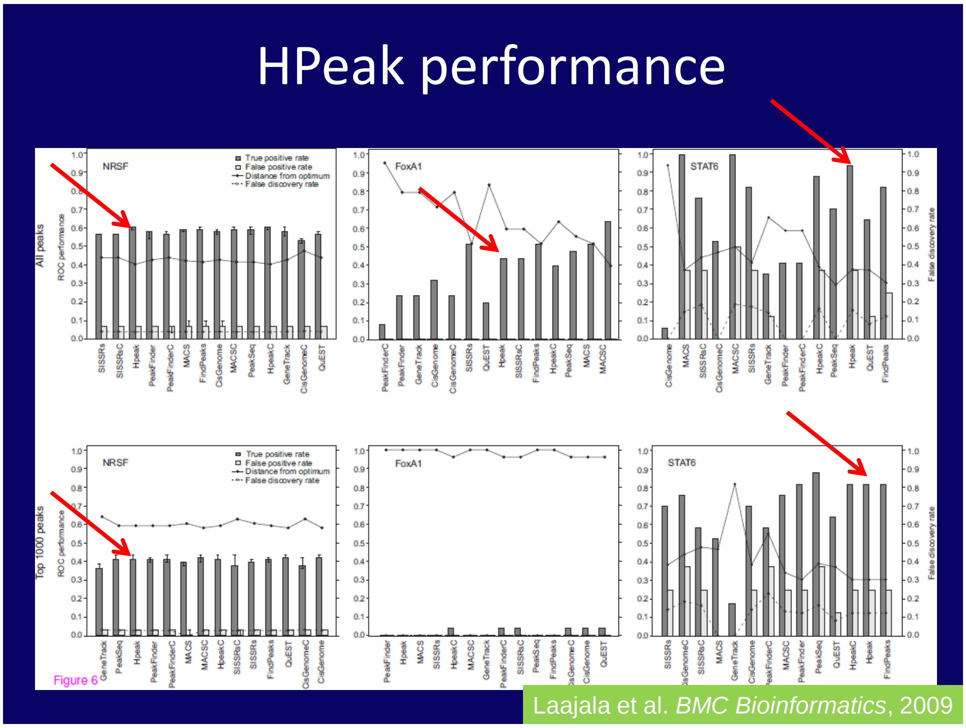
Post-processing on identified peaks.

12



Motif enrichment results for NRSF and STAT1 data

13



HPeak performance

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.

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

$$E(Y)=\lambda$$

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

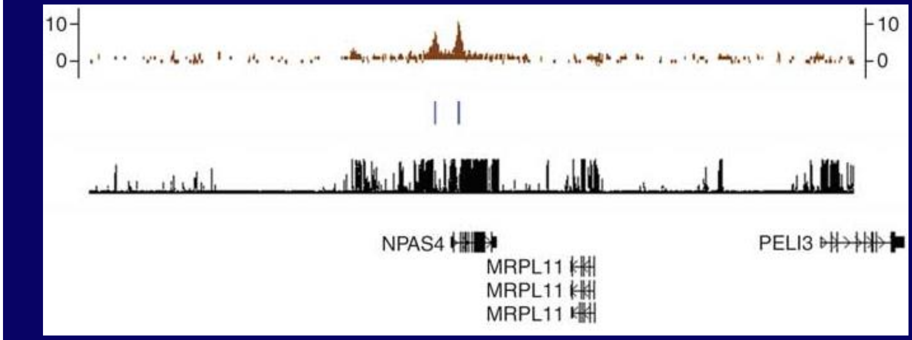
- Zero-inflated Poisson distribution

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

Outline

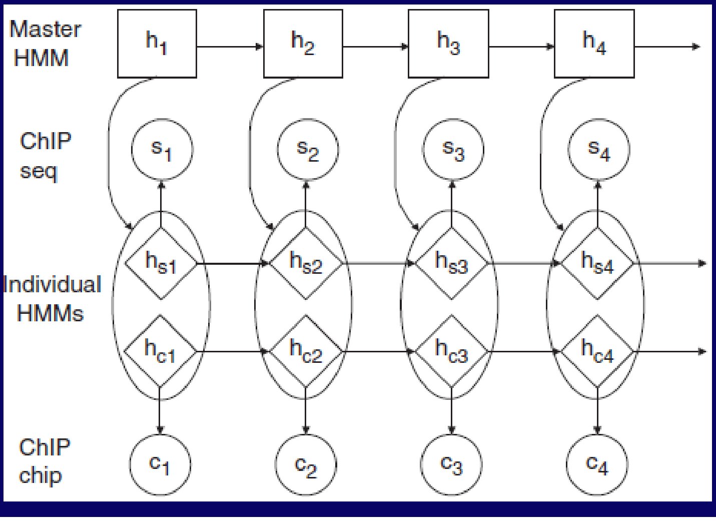
- Hidden Markov model for peak detection
- Hierarchical Hidden Markov model for combining ChIP-seq and ChIP-chip data
- Hybrid Monte Carlo strategy for Motif finding

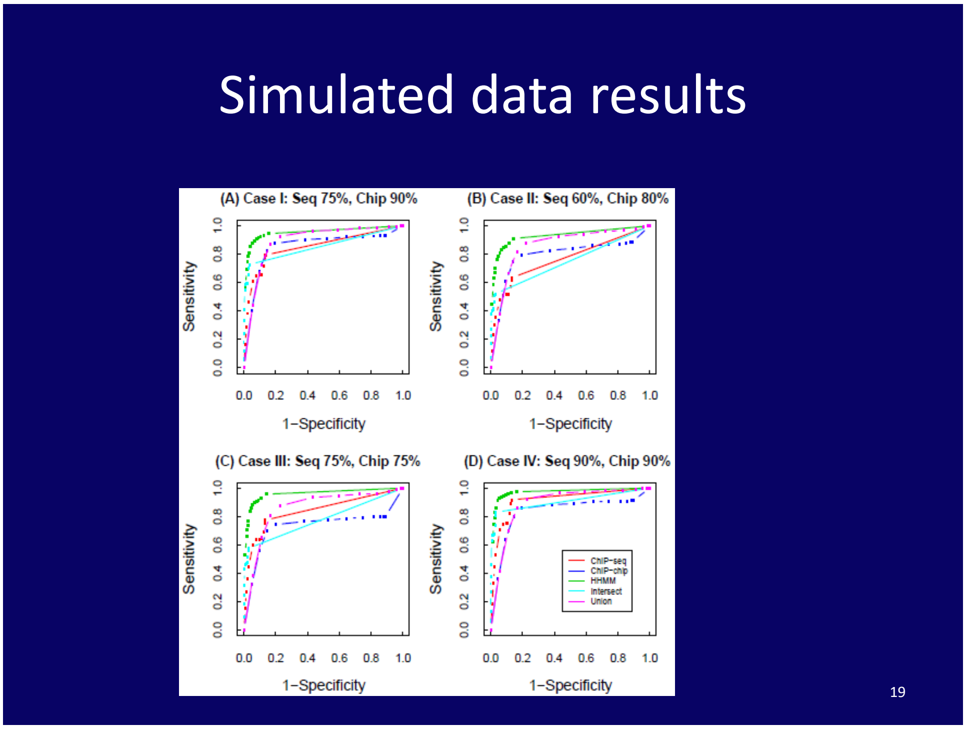
Joint analysis of ChIP-chip and ChIP-seq



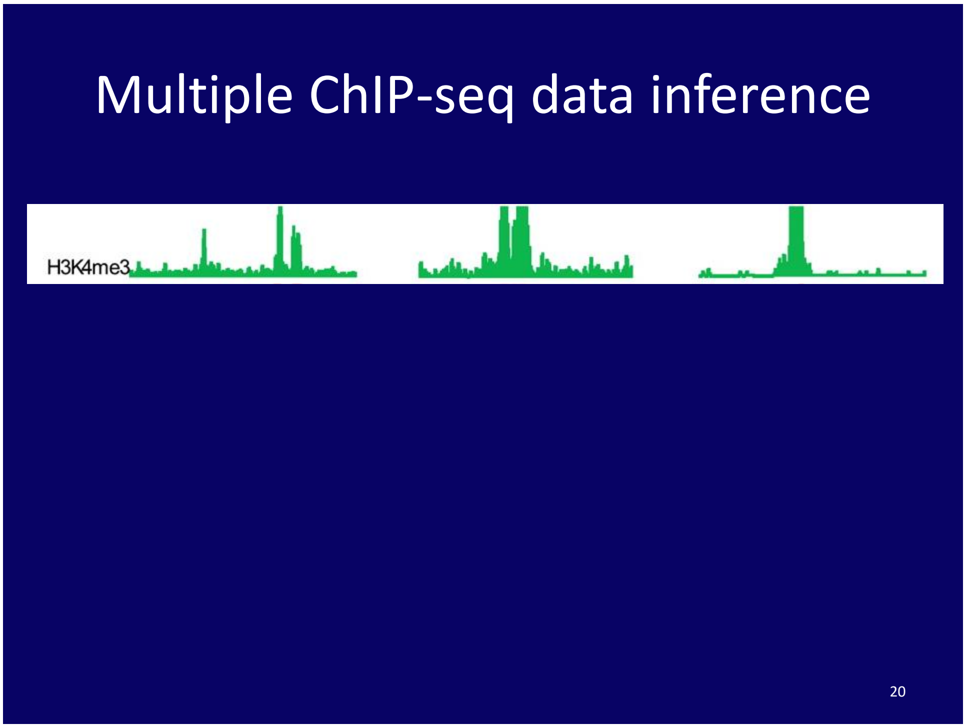
Ji et al. *Nat Biotechnology*, 2008

Hierarchical HMM



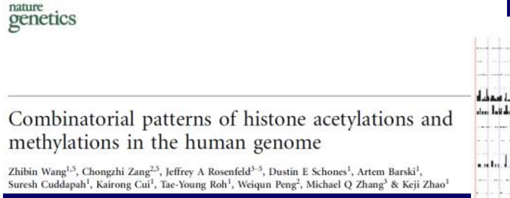


Simulated data results



Multiple ChIP-seq data inference

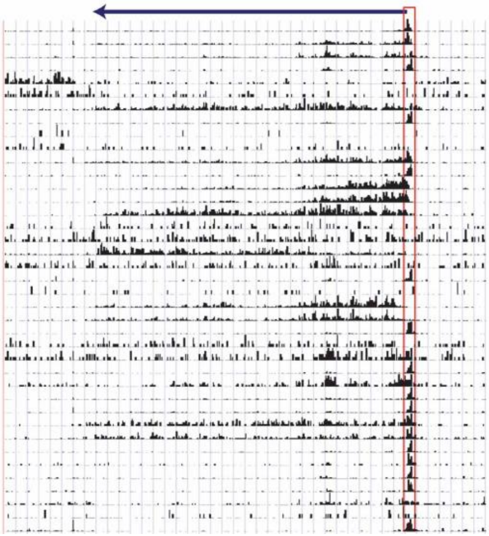
ChIP-Seq compendium



nature genetics

Combinatorial patterns of histone acetylations and methylations in the human genome

Zhibin Wang[1,5], Chongzhi Zang[2,5], Jeffrey A Rosenfeld[3–5], Dustin E Schones[1], Artem Barski[1], Suresh Cuddapah[1], Kairong Cui[1], Tae-Young Roh[1], Weiqun Peng[2], Michael Q Zhang[3] & Keji Zhao[1]



The problem

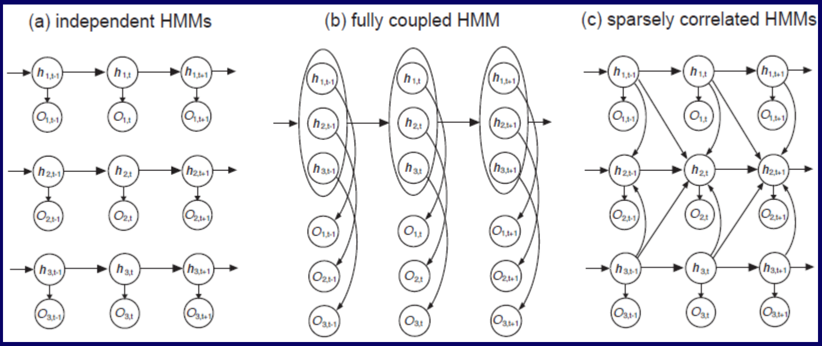
- *N* series of data, each can be modeled by an HMM,
- The goal is to infer the hidden states for all series,
- Suppose there are *k* states for each chain, then the total number of possible states for the whole datasets is k^N, the size of the transition matrix is k^{2N} ,
  - Independent: ignore correlation among the data series,
  - A single HMM to model all data together: intractable for large *N*.

22

Our goal

- Allow coupling among the chains,
- The goal is to borrow information across different experiments/datasets,
- Limit the amount of coupling allowed to reduce computation cost

Our scheme



Our learning plan

- Perform inference one series a time,
- Incorporate knowledge of hidden states in other series into the learning process,
- Assume sparsity in the correlation matrix.

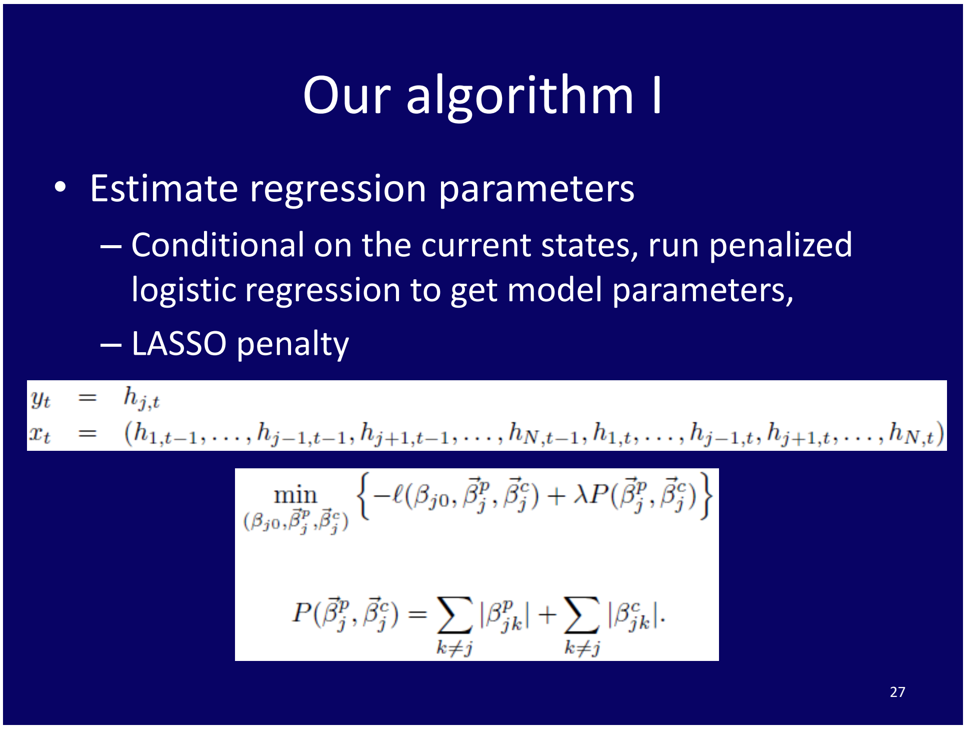
Our model

- Use an inhomogeneous HMM to incorporate correlation,
- Define the transition kernel for series *j* and time *t* as:

$$K_j(t) = \begin{pmatrix} 1 - p_{jt} & p_{jt} \\ 1 - q_{jt} & q_{jt} \end{pmatrix}$$

$p_{jt} = Pr(h_{j,t} = 1 | h_{j,t-1} = 0)$ and $q_{jt} = Pr(h_{j,t} = 1 | h_{j,t-1} = 1)$.

$$\begin{aligned} \log\left(\frac{p_{jt}}{1 - p_{jt}}\right) &= \beta^p_{j0} + \sum_{k \neq j}\left(\beta^p_{jk} h_{k,t-1} + \beta^c_{jk} h_{k,t}\right) \\ \log\left(\frac{q_{jt}}{1 - q_{jt}}\right) &= \gamma^p_{j0} + \sum_{k \neq j}\left(\gamma^p_{jk} h_{k,t-1} + \gamma^c_{jk} h_{k,t}\right) \end{aligned}$$



Our algorithm I

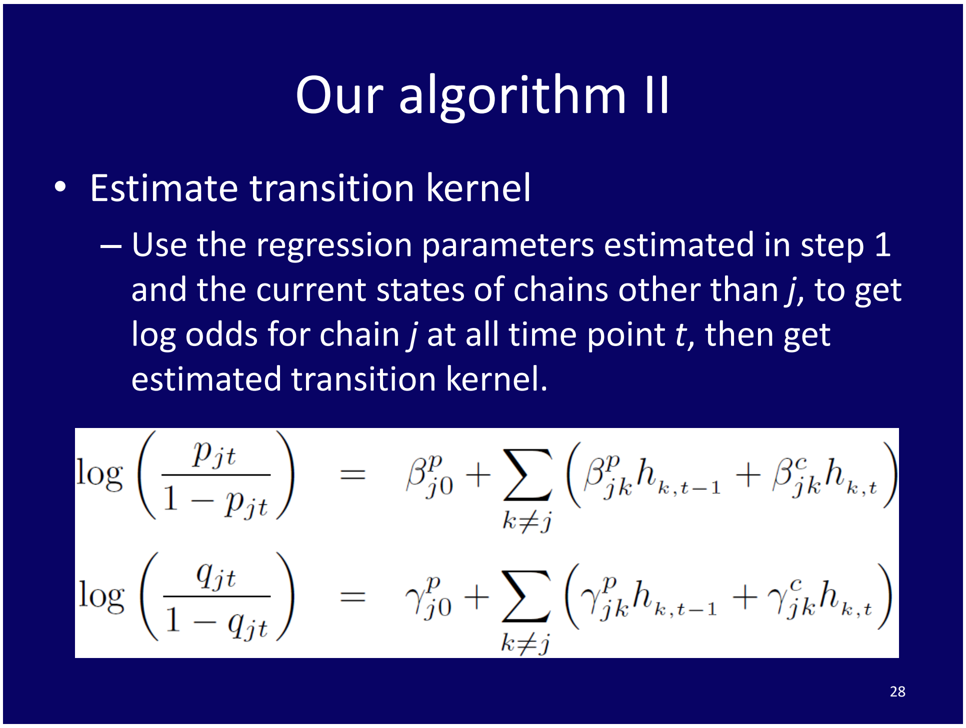
- Estimate regression parameters
  - Conditional on the current states, run penalized logistic regression to get model parameters,
  - LASSO penalty

$$\begin{aligned} y_t &= h_{j,t} \\ x_t &= (h_{1,t-1}, \ldots, h_{j-1,t-1}, h_{j+1,t-1}, \ldots, h_{N,t-1}, h_{1,t}, \ldots, h_{j-1,t}, h_{j+1,t}, \ldots, h_{N,t}) \end{aligned}$$

$$\min_{(\beta_{j0}, \vec{\beta}_j^p, \vec{\beta}_j^c)} \left\{ -\ell(\beta_{j0}, \vec{\beta}_j^p, \vec{\beta}_j^c) + \lambda P(\vec{\beta}_j^p, \vec{\beta}_j^c) \right\}$$

$$P(\vec{\beta}_j^p, \vec{\beta}_j^c) = \sum_{k \neq j} |\beta_{jk}^p| + \sum_{k \neq j} |\beta_{jk}^c|.$$

27



Our algorithm II

- Estimate transition kernel
  - Use the regression parameters estimated in step 1 and the current states of chains other than *j*, to get log odds for chain *j* at all time point *t*, then get estimated transition kernel.

$$\begin{aligned} \log\left(\frac{p_{jt}}{1-p_{jt}}\right) &= \beta_{j0}^p + \sum_{k \neq j} \left(\beta_{jk}^p h_{k,t-1} + \beta_{jk}^c h_{k,t} \right) \\ \log\left(\frac{q_{jt}}{1-q_{jt}}\right) &= \gamma_{j0}^p + \sum_{k \neq j} \left(\gamma_{jk}^p h_{k,t-1} + \gamma_{jk}^c h_{k,t} \right) \end{aligned}$$

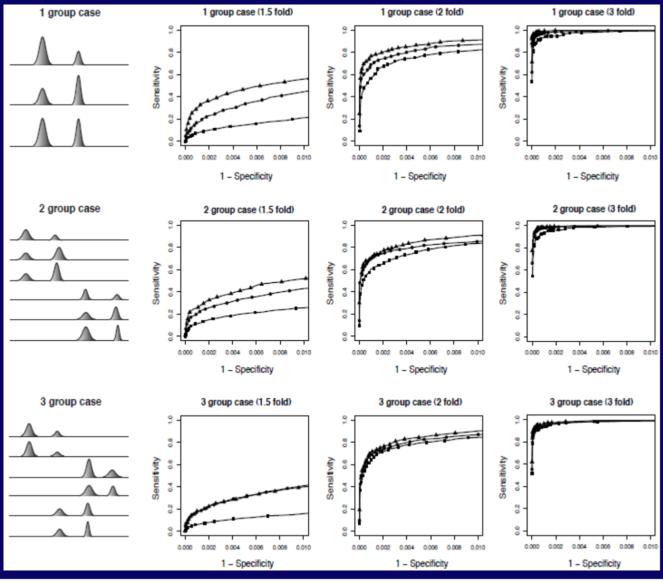
28

Our algorithm III

- Infer hidden states
  - Use the transition kernel estimated in step 2, current emission probabilities and observed data to run regular HMM (forward-backward algorithm) to get updated hidden states,
- Estimate the emission probabilities
  - Use the hidden states estimated in step 3 and observed data to update emission probabilities.

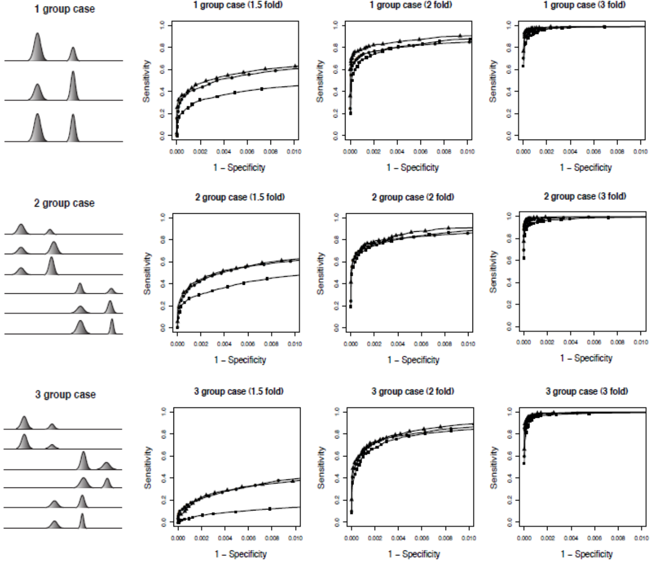
29

Simulation studies



30

Simulation studies



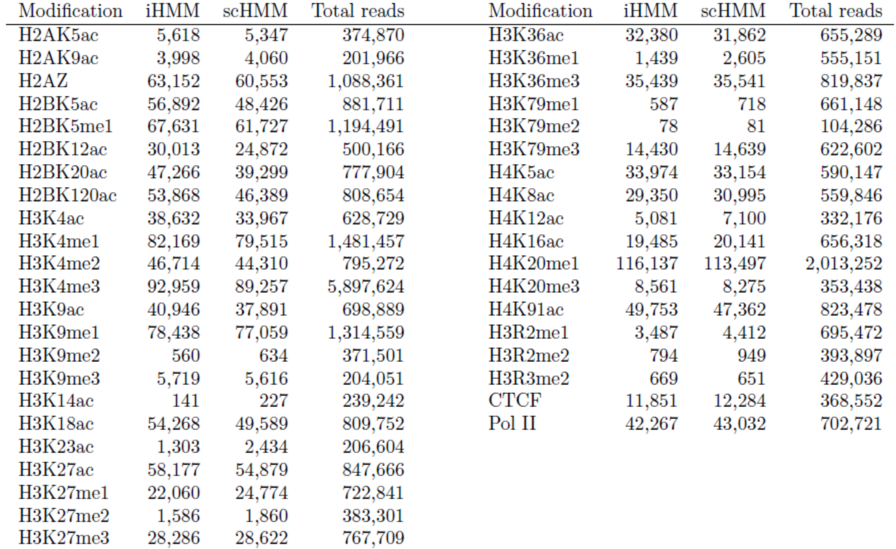
31

Real data

- In human CD4+ T cells,
- 39 histone acetylations and methylations marks + RNA polII + CTCF,
- 200 bp bin,
- 5kb up/downstream of TSS,
- Barski *et al. Cell* 2007, Wang *et al. Nature Genetics,* 2008.

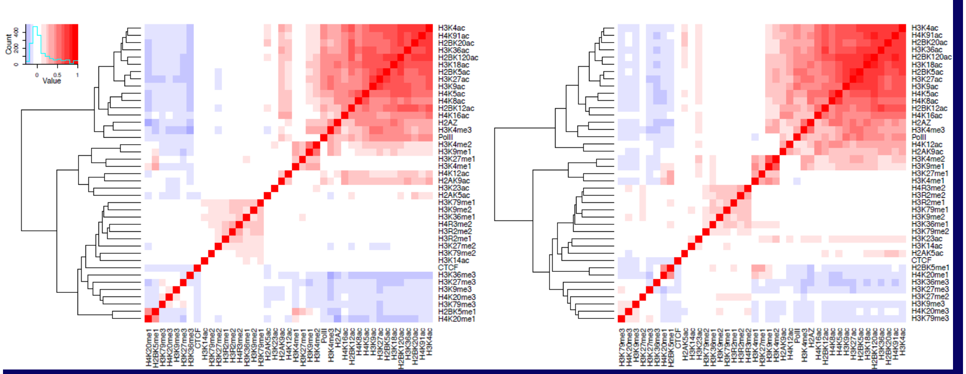
32

Real data description



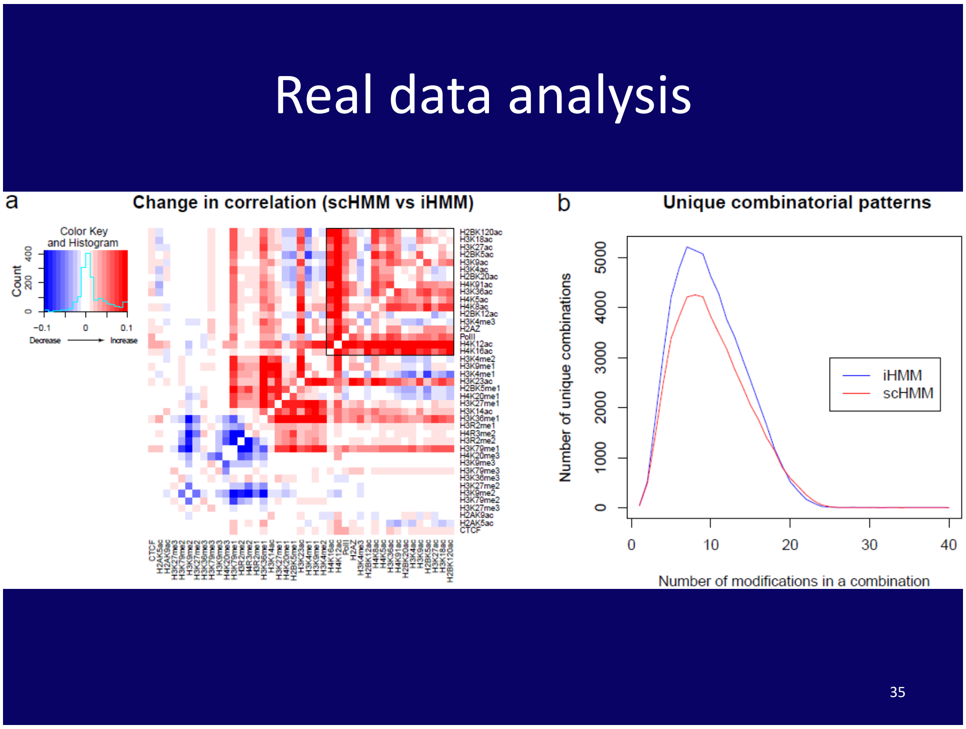
Modification	iHMM	scHMM	Total reads
H2AK5ac	5,618	5,347	374,870
H2AK9ac	3,998	4,060	201,966
H2AZ	63,152	60,553	1,088,361
H2BK5ac	56,892	48,426	881,711
H2BK5me1	67,631	61,727	1,194,491
H2BK12ac	30,013	24,872	500,166
H2BK20ac	47,266	39,299	777,904
H2BK120ac	53,868	46,389	808,654
H3K4ac	38,632	33,967	628,729
H3K4me1	82,169	79,515	1,481,457
H3K4me2	46,714	44,310	795,272
H3K4me3	92,959	89,257	5,897,624
H3K9ac	40,946	37,891	698,889
H3K9me1	78,438	77,059	1,314,559
H3K9me2	560	634	371,501
H3K9me3	5,719	5,616	204,051
H3K14ac	141	227	239,242
H3K18ac	54,268	49,589	809,752
H3K23ac	1,303	2,434	206,604
H3K27ac	58,177	54,879	847,666
H3K27me1	22,060	24,774	722,841
H3K27me2	1,586	1,860	383,301
H3K27me3	28,286	28,622	767,709
H3K36ac	32,380	31,862	655,289
H3K36me1	1,439	2,605	555,151
H3K36me3	35,439	35,541	819,837
H3K79me1	587	718	661,148
H3K79me2	78	81	104,286
H3K79me3	14,430	14,639	622,602
H4K5ac	33,974	33,154	590,147
H4K8ac	29,350	30,995	559,846
H4K12ac	5,081	7,100	332,176
H4K16ac	19,485	20,141	656,318
H4K20me1	116,137	113,497	2,013,252
H4K20me3	8,561	8,275	353,438
H4K91ac	49,753	47,362	823,478
H3R2me1	3,487	4,412	695,472
H3R2me2	794	949	393,897
H3R3me2	669	651	429,036
CTCF	11,851	12,284	368,552
Pol II	42,267	43,032	702,721

Real data analysis



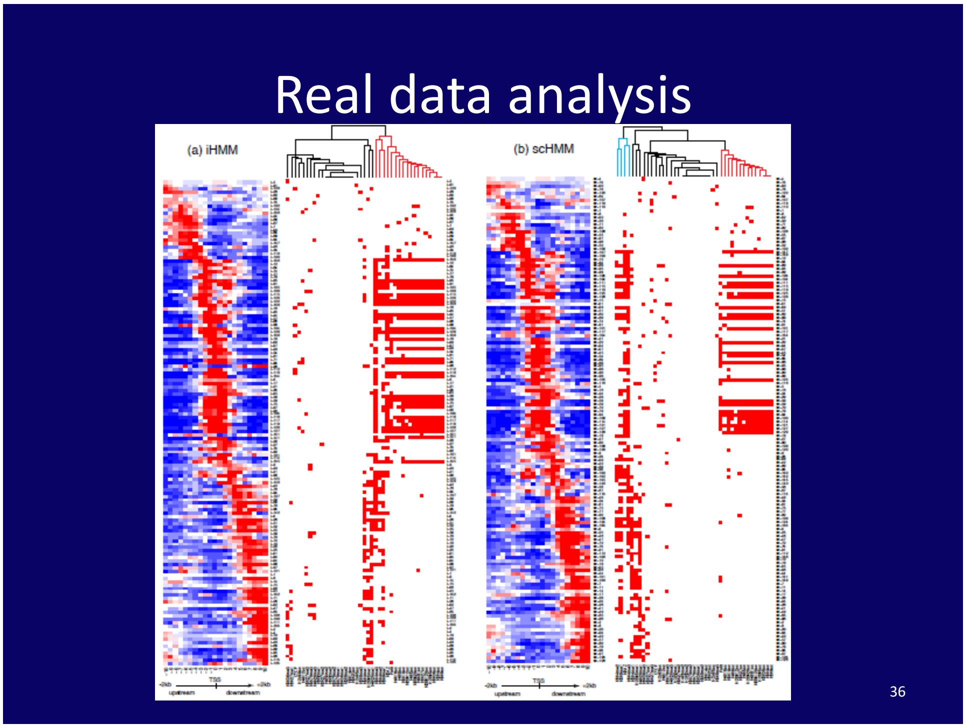
(a) Correlation between modifications in iHMM

(b) Correlation between modifications in scHMM



Real data analysis

35



Real data analysis

36

Joint inference of multiple ChIP-seq data

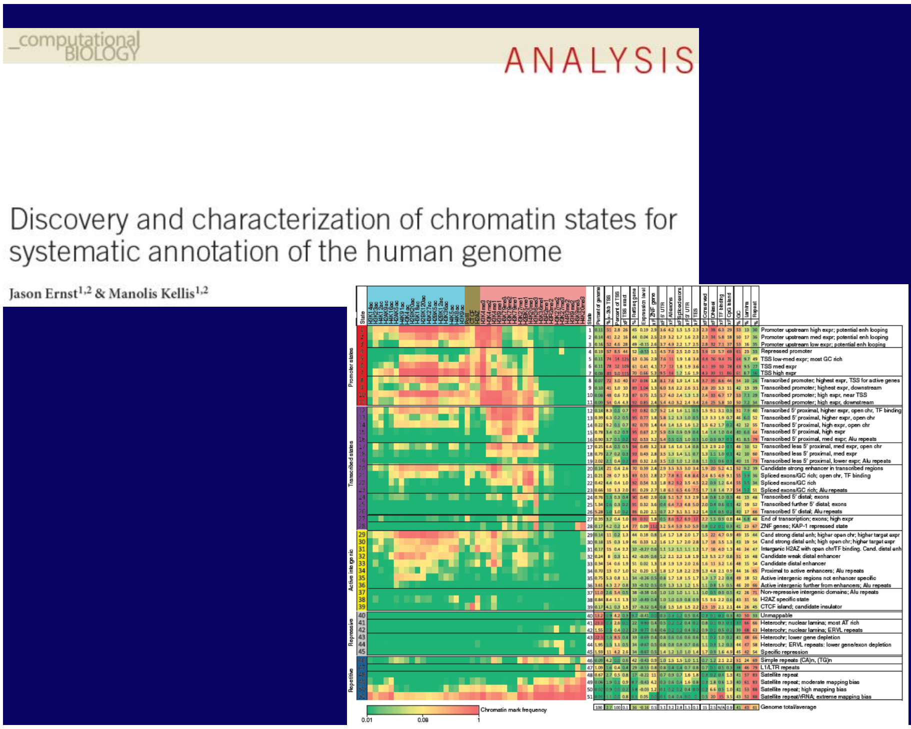
- JAMIE
  - Joint analysis of multiple ChIP-chip data
  - Wu, Ji Bioinformatics 2010
- HHMM
  - Joint analysis of ChIP-seq and ChIP-chip data
  - Choi et al. Bioinformatics 2009
- scHMM
  - Joint analysis of multiple ChIP-seq data
  - Choi et al. bioinformatics 2013

Acknowledgement

Hyung Won Choi
National University of Singapore

Debashis Ghosh
Colorado School of Public Health

Alexey Nesvizhskii
Damian Fermin
University of Michigan



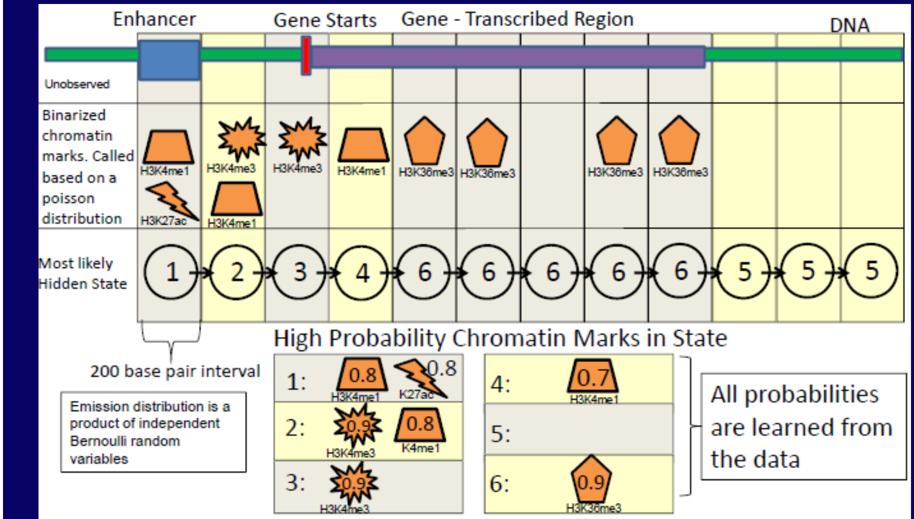
computational BIOLOGY

ANALYSIS

Discovery and characterization of chromatin states for systematic annotation of the human genome

Jason Ernst[1,2] & Manolis Kellis[1,2]

Method: Multivariate Hidden Markov Model



Emission distribution is a product of independent Bernoulli random variables

All probabilities are learned from the data

Binarization leads to explicit modeling of mark combinations and interpretable parameters

ENCODE: Study nine marks in nine human cell lines

9 marks

- H3K4me1
- H3K4me2
- H3K4me3
- H3K27ac
- H3K9ac
- H3K27me3
- H4K20me1
- H3K36me3
- CTCF
- +WCE
- +RNA

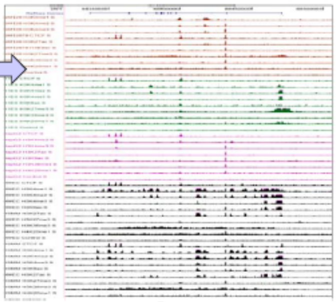
X

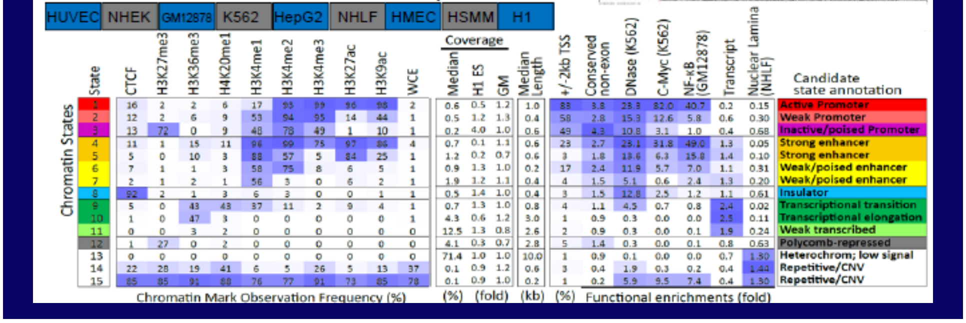
9 human cell types

HUVEC	Umbilical vein endothelial
NHEK	Keratinocytes
GM12878	Lymphoblastoid
K562	Myelogenous leukemia
HepG2	Liver carcinoma
NHLF	Normal human lung fibroblast
HMEC	Mammary epithelial cell
HSMM	Skeletal muscle myoblasts
H1	Embryonic

Brad Bernstein ENCODE Group

81 Chromatin Mark Tracks





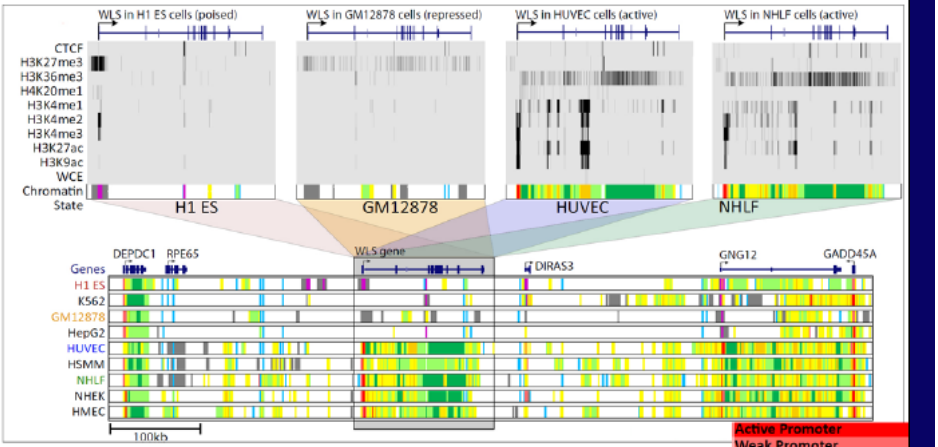
HUVEC NHEK GM12878 K562 HepG2 NHLF HMEC HSMM H1

Chromatin States

State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Coverage Median	Coverage H1 ES	Coverage GM	Median Length	+/-2kb TSS	Conserved non-exon	DNase (K562)	C-Myc (K562)	NF-κB (GM12878)	Transcript	Nuclear Lamina (NHLF)	Candidate state annotation
1	16	2	2	6	17	93	99	96	98	2	0.6	0.5	1.2	1.0	88	3.8	23.3	82.0	40.7	0.2	0.15	Active Promoter
2	12	2	6	9	53	94	95	14	44	1	0.5	1.2	1.3	0.4	58	2.8	15.3	12.6	5.8	0.6	0.30	Weak Promoter
3	13	72	0	9	48	78	49	1	10	1	0.2	4.0	1.0	0.6	49	4.3	10.8	3.1	1.0	0.4	0.68	Inactive/poised Promoter
4	11	1	15	11	96	99	75	97	86	4	0.7	0.1	1.1	0.6	23	2.7	23.1	31.8	49.0	1.3	0.05	Strong enhancer
5	5	0	10	3	88	57	5	84	25	1	1.2	0.2	0.7	0.6	3	1.8	18.6	6.3	15.8	1.4	0.10	Strong enhancer
6	7	1	1	3	58	75	8	6	5	1	0.9	1.3	1.0	0.2	17	2.4	11.9	5.7	7.0	1.1	0.31	Weak/poised enhancer
7	2	1	2	1	56	3	0	6	2	1	1.9	1.2	1.1	0.4	4	1.5	5.1	0.6	2.4	1.3	0.20	Weak/poised enhancer
8	92	2	1	3	6	3	0	0	1	1	0.5	1.4	1.0	0.4	3	1.5	12.8	2.5	1.2	1.1	0.61	Insulator
9	5	0	43	43	37	11	2	9	4	1	0.7	1.3	1.0	0.8	4	1.1	4.5	0.7	0.8	2.4	0.02	Transcriptional transition
10	1	0	47	3	0	0	0	0	0	1	4.3	0.6	1.2	3.0	1	0.9	0.3	0.0	0.0	2.5	0.11	Transcriptional elongation
11	0	0	3	2	0	0	0	0	0	0	12.5	1.3	0.8	2.6	2	0.9	0.3	0.0	0.1	1.9	0.24	Weak transcribed
12	1	27	0	2	0	0	0	0	0	0	4.1	0.3	0.7	2.8	5	1.4	0.3	0.0	0.1	0.8	0.63	Polycomb-repressed
13	0	0	0	0	0	0	0	0	0	0	71.4	1.0	1.0	10.0	1	0.9	0.1	0.0	0.0	0.7	1.30	Heterochrom; low signal
14	22	28	19	41	6	5	26	5	13	37	0.1	0.9	1.2	0.6	3	0.4	1.9	0.3	0.2	0.4	1.44	Repetitive/CNV
15	85	85	91	88	76	77	91	73	85	78	0.1	0.9	1.0	0.2	1	0.2	5.9	9.5	7.4	0.4	1.30	Repetitive/CNV

Chromatin Mark Observation Frequency (%) (%) (fold) (kb) (%) Functional enrichments (fold)

Chromatin states dynamics across nine ENCODE cell types



- Active Promoter
- Weak Promoter
- Inactive/poised Promoter
- Strong enhancer
- Strong enhancer
- Weak/poised enhancer
- Weak/poised enhancer
- Insulator
- Transcriptional transition
- Transcriptional elongation
- Weak transcribed
- Polycomb-repressed
- Heterochrom; low signal

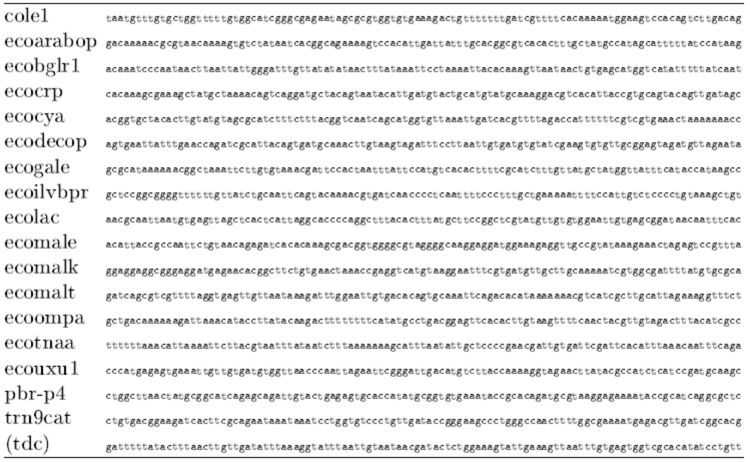
- **Single annotation track for each cell type**
- **Summarize cell-type activity at a glance**
- **Can study 9-cell activity pattern across ↓**

Ernst et al, *Nature* 2011

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

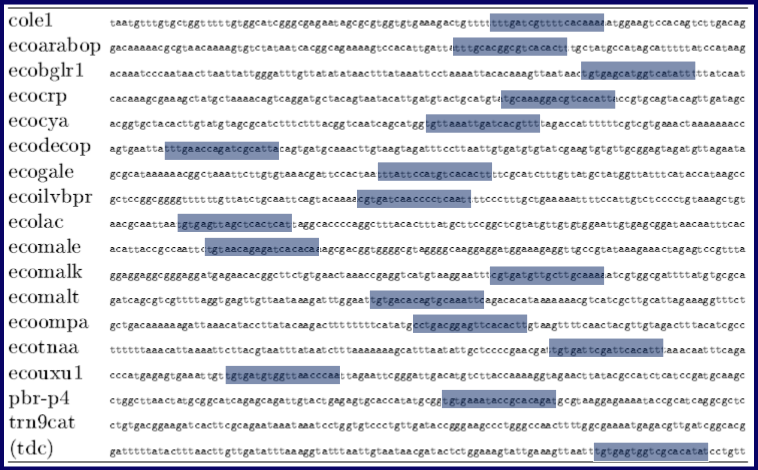
Example: cyclic receptor protein (CRP)



cole1	taatgtttgtgctggttttgtggcatcgggcgagaatagcgcgtggtgtgaaagactgtttttttgatcgttttcacaaaaatggaagtccacagtcttgacag
ecoarabop	gacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgattatttgcacggcgtcacactttgctatgccatagcatttttatccataag
ecobglr1	acaaatcccaataacttaattattgggatttgttatatataactttataaattcctaaaattacacaaagttaataactgtgagcatggtcatatttttatcaat
ecocrp	cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatgtatgcaaaggacgtcacattaccgtgcagtacagttgatagc
ecocya	acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcatggtgttaaattgatcacgttttagaccatttttcgtcgtgaaactaaaaaaacc
ecodecop	agtgaattatttgaaccagatcgcattacagtgatgcaaacttgtaagtagatttccttaattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata
ecogale	gcgcataaaaaacggctaaattcttgtgtaaacgattccactaatttattccatgtcacacttttcgcatctttgttatgctatggttatttcataccataagcc
ecoilvbpr	gctccggcggggtttttgttatctgcaattcagtacaaaacgtgatcaacccctcaatttttccctttgctgaaaaatttccattgtctcccctgtaaagctgt
ecolac	aacgcaattaatgtgagttagctcactcattaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcac
ecomale	acattaccgccaattctgtaacagagatcacacaaagcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ecomalk	ggaggaggcgggaggatgagaacacggcttctgtgaactaaaccgaggtcatgtaaggaatttcgtgatgttgcttgcaaaaatcgtggcgattttatgtgcgca
ecomalt	gatcagcgtcgttttaggtgagttgttaataaagatttggaattgtgacacagtgcaaattcagacacataaaaaaacgtcatcgcttgcattagaaaggtttct
ecoompa	gctgacaaaaaagattaaacataccttatacaagacttttttttcatatgcctgacggagttcacacttgtaagttttcaactacgttgtagactttacatcgcc
ecotnaa	ttttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaacgattgtgattcgattcacatttaaacaatttcaga
ecouxu1	cccatgagagtgaaattgttgtgatgtggttaacccaattagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
pbr-p4	ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgcggtgtgaaataccgcacagatgcgtaaggagaaaataccgcatcaggcgctc
trn9cat	ctgtgacggaagatcacttcgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaaaatgagacgttgatcggcacg
(tdc)	gatttttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagttaatttgtgagtggtcgcacatatcctgtt

Stormo and Hartzell, 1989

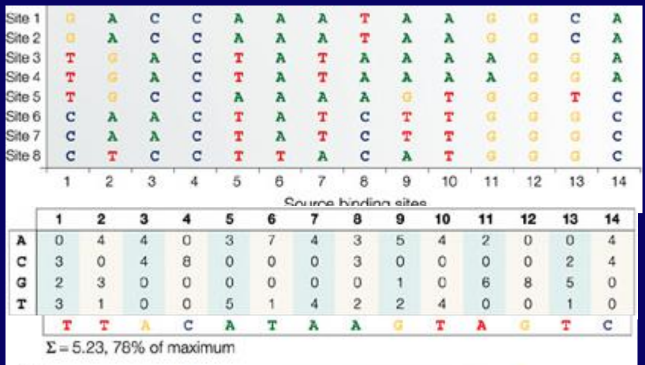
Example: cyclic receptor protein (CRP)



cole1	taatgtttgtgctggttttgtggcatcgggcgagaatagcgcgtggtgtgaaagactgttttttttgatcgttttcacaaaaatggaagtccacagtcttgacag
ecoarabop	gacaaaaacgcgtaacaaaagtgtctataatcacggcagaaaagtccacattgattatttgcacggcgtcacactttgctatgccatagcatttttatccataag
ecobglr1	acaaatcccaataacttaattattgggatttgttatatataactttataaattcctaaaattacacaaagttaataactgtgagcatggtcatatttttatcaat
ecocrp	cacaaagcgaaagctatgctaaaacagtcaggatgctacagtaatacattgatgtactgcatgtatgcaaaggacgtcacattaccgtgcagtacagttgatagc
ecocya	acggtgctacacttgtatgtagcgcatctttctttacggtcaatcagcatggtgttaaattgatcacgttttagaccatttttcgtcgtgaaactaaaaaaacc
ecodecop	agtgaattatttgaaccagatcgcattacagtgatgcaaacttgtaagtagatttccttaattgtgatgtgtatcgaagtgtgttgcggagtagatgttagaata
ecogale	gcgcataaaaaacggctaaattcttgtgtaaacgattccactaatttattccatgtcacacttttcgcatctttgttatgctatggttatttcataccataagcc
ecoilvbpr	gctccggcggggtttttgttatctgcaattcagtacaaaacgtgatcaacccctcaatttttccctttgctgaaaaatttccattgtctcccctgtaaagctgt
ecolac	aacgcaattaatgtgagttagctcactcattaggcaccccaggctttacactttatgcttccggctcgtatgttgtgtggaattgtgagcggataacaatttcac
ecomale	acattaccgccaattctgtaacagagatcacacaaagcgacggtggggcgtaggggcaaggaggatggaaagaggttgccgtataaagaaactagagtccgttta
ecomalk	ggaggaggcgggaggatgagaacacggcttctgtgaactaaaccgaggtcatgtaaggaatttcgtgatgttgcttgcaaaaatcgtggcgattttatgtgcgca
ecomalt	gatcagcgtcgttttaggtgagttgttaataaagatttggaattgtgacacagtgcaaattcagacacataaaaaaacgtcatcgcttgcattagaaaggtttct
ecoompa	gctgacaaaaaagattaaacataccttatacaagactttttttcatatgcctgacggagttcacacttgtaagttttcaactacgttgtagactttacatcgcc
ecotnaa	ttttttaaacattaaaattcttacgtaatttataatctttaaaaaaagcatttaatattgctccccgaacgattgtgattcgattcacatttaaacaatttcaga
ecouxu1	cccatgagagtgaaattgttgtgatgtggttaacccaattagaattcgggattgacatgtcttaccaaaaggtagaacttatacgccatctcatccgatgcaagc
pbr-p4	ctggcttaactatgcggcatcagagcagattgtactgagagtgcaccatatgcggtgtgaaataccgcacagatgcgtaaggagaaaataccgcatcaggcgctc
trn9cat	ctgtgacggaagatcacttcgcagaataaataaatcctggtgtccctgttgataccgggaagccctgggccaacttttggcgaaaatgagacgttgatcggcacg
(tdc)	gatttttatactttaacttgttgatatttaaaggtatttaattgtaataacgatactctggaaagtattgaaagttaatttgtgagtggtcgcacatatcctgtt

Stormo and Hartzell

Transcription factor binding site (TFBS)

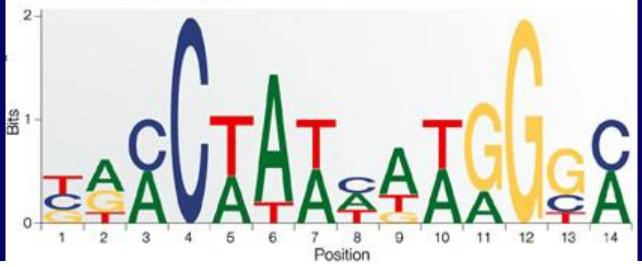


Site 1	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 2	G	A	C	C	A	A	A	T	A	A	G	G	C	A
Site 3	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 4	T	G	A	C	T	A	T	A	A	A	A	G	G	A
Site 5	T	G	C	C	A	A	A	A	G	T	G	G	T	C
Site 6	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 7	C	A	A	C	T	A	T	C	T	T	G	G	G	C
Site 8	C	T	C	C	T	T	A	C	A	T	G	G	G	C
	1	2	3	4	5	6	7	8	9	10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

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

$\Sigma = 5.23$, 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

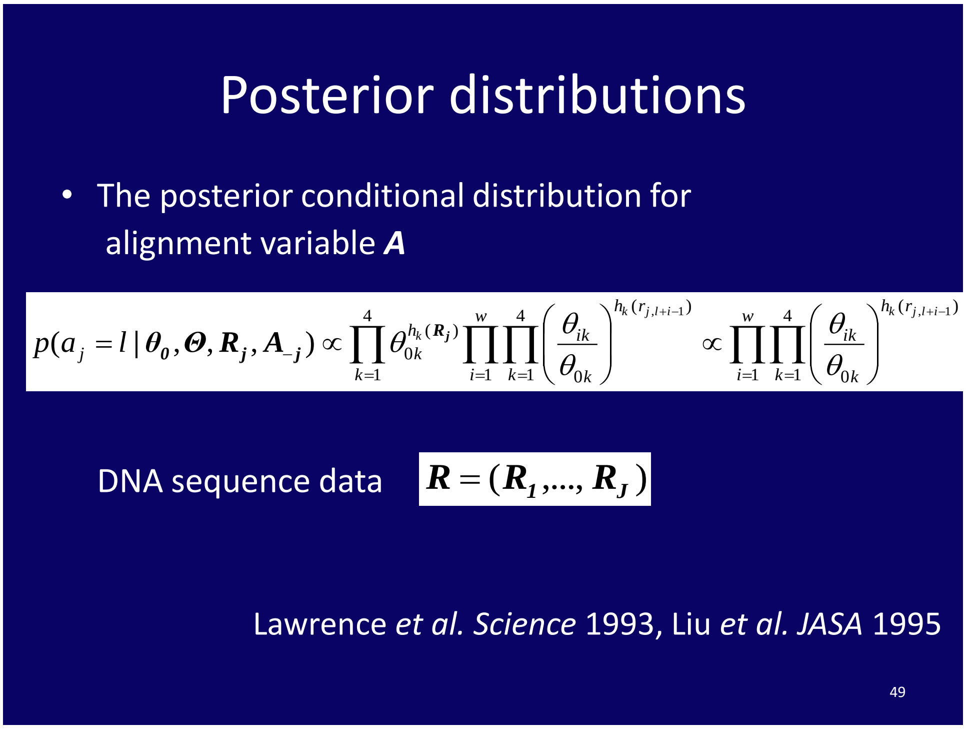
gaatagctactcgatcgatactagcaatcgttaccctagctcgatcgagatggaaagactataa

…

a_J

acgtgagatcagctatcgatcgattgataactactcgtacgtat

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



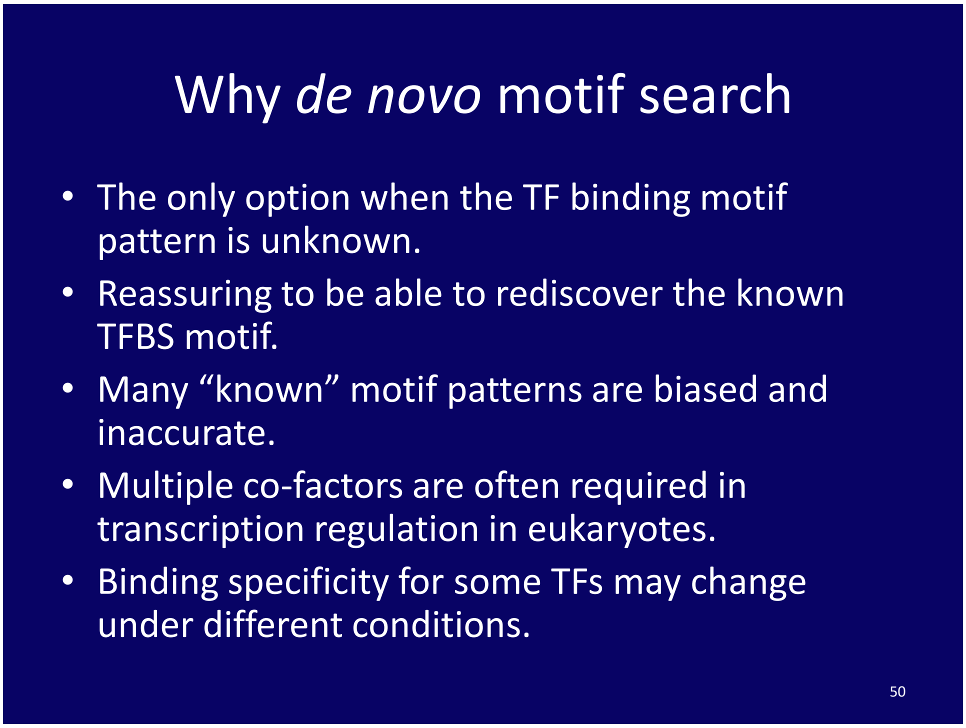
Posterior distributions

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

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

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

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

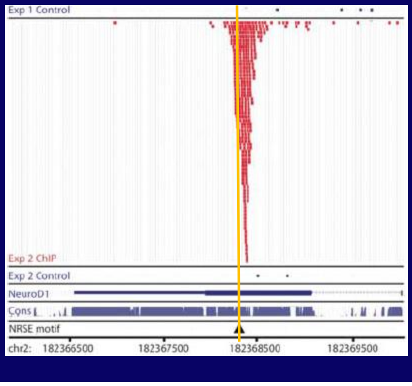


Why *de novo* motif search

- 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*

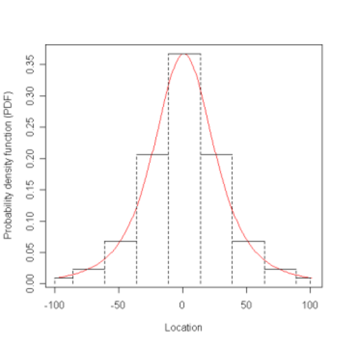
Features of our new algorithm

- 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.

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

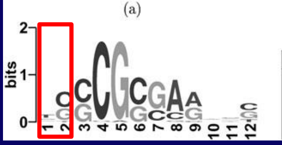


53

Modeling inter-dependent positions

- Zhou and Liu

  Bioinformatics 2005

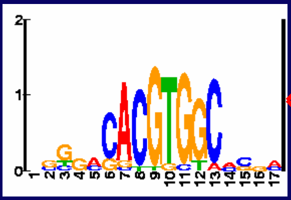


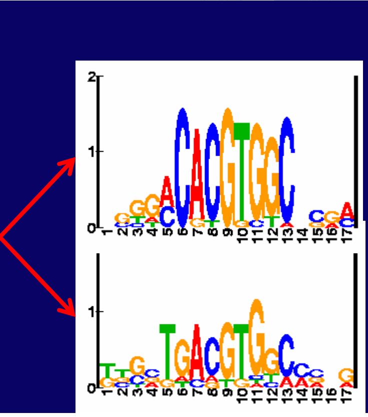
(b)

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

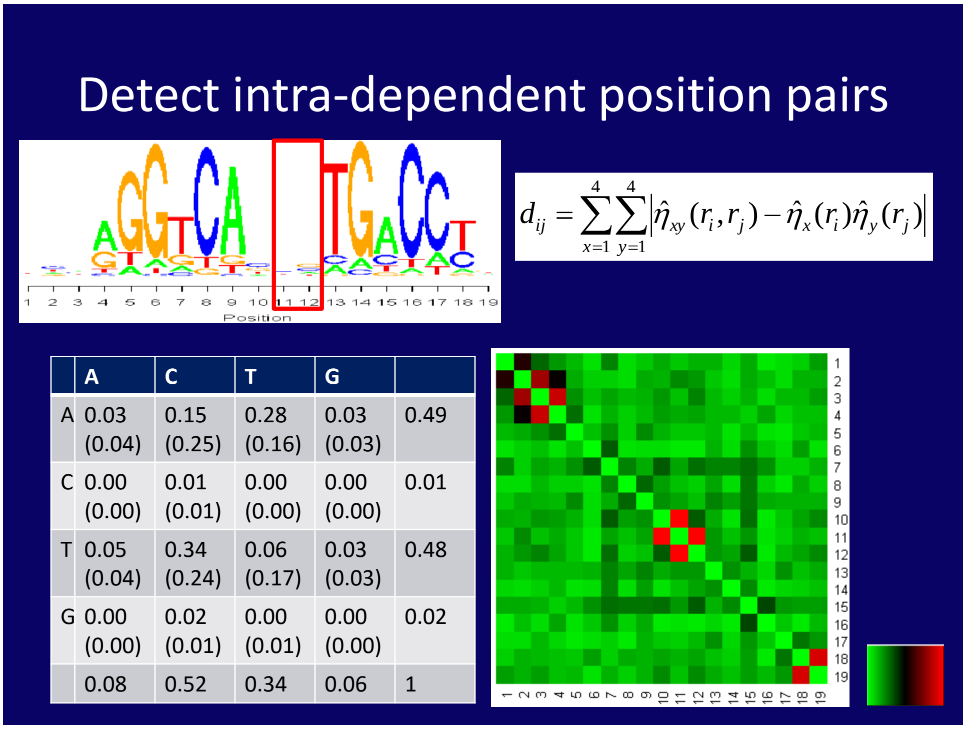
- Barash *et al.*

  RECOMB 2003





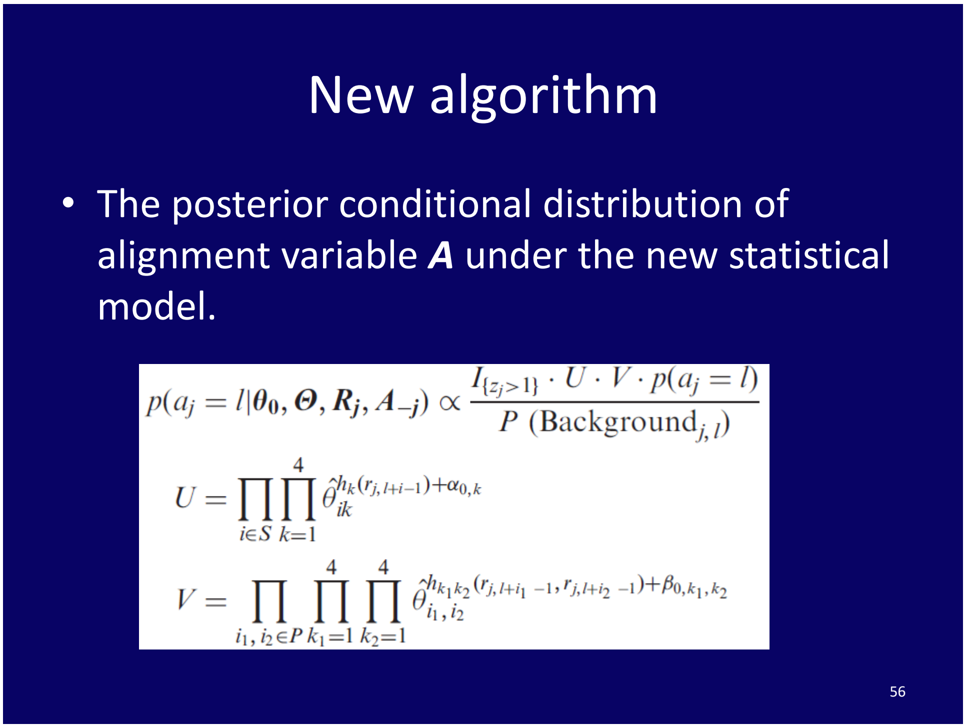
54



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

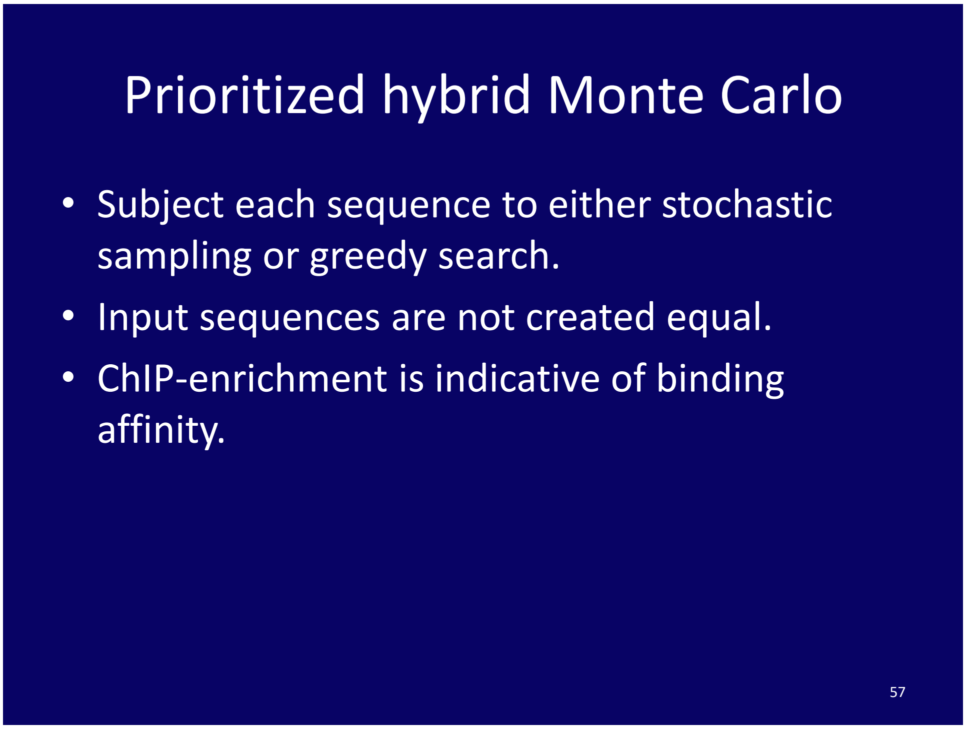
- The posterior conditional distribution of alignment variable ***A*** under the new statistical model.

$$p(a_j = l|\boldsymbol{\theta_0}, \boldsymbol{\Theta}, \boldsymbol{R_j}, \boldsymbol{A_{-j}}) \propto \frac{I_{\{z_j>1\}} \cdot U \cdot V \cdot p(a_j = l)}{P\ (\text{Background}_{j,\,l})}$$

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

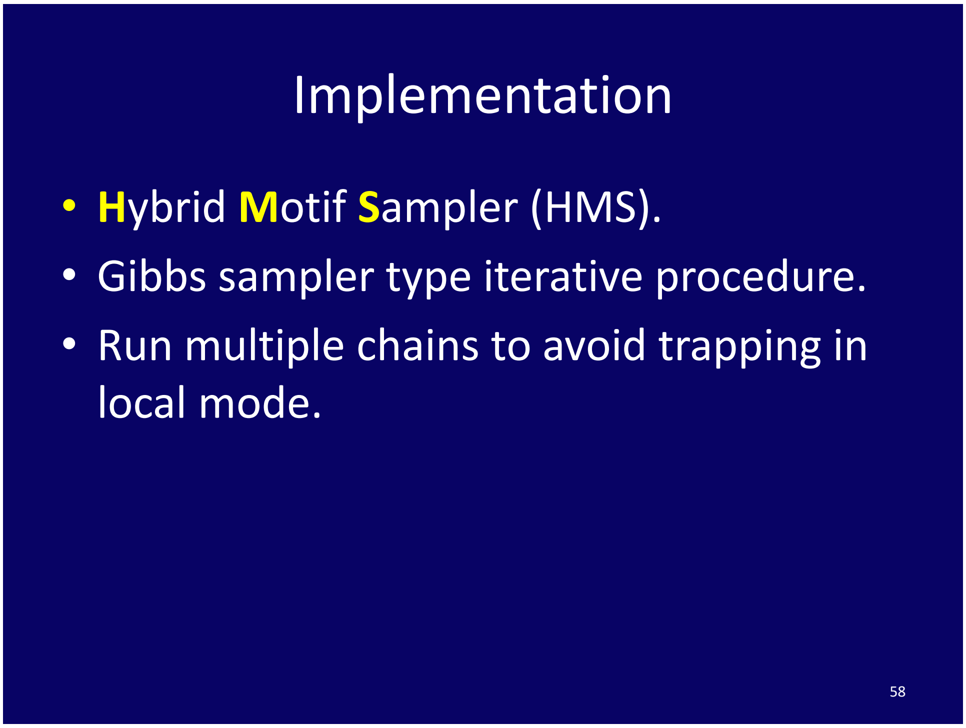
$$V = \prod_{i_1,i_2\in P}\prod_{k_1=1}^{4}\prod_{k_2=1}^{4}\hat{\theta}_{i_1,i_2}^{h_{k_1k_2}(r_{j,l+i_1-1},\, r_{j,l+i_2-1})+\beta_{0,k_1,k_2}}$$

56



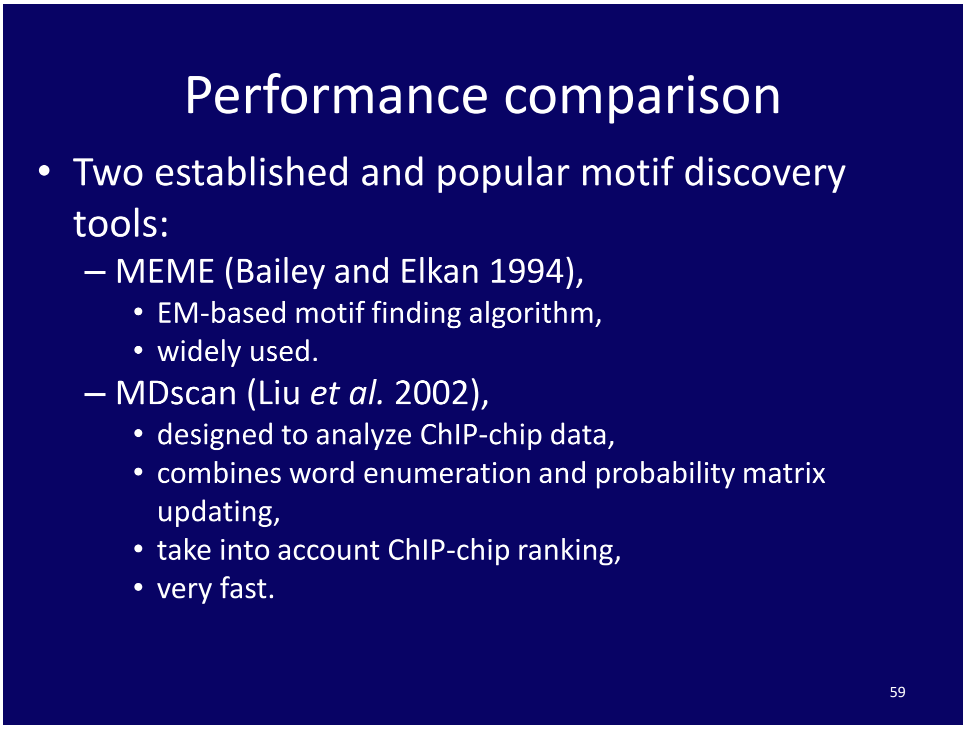
Prioritized hybrid Monte Carlo

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



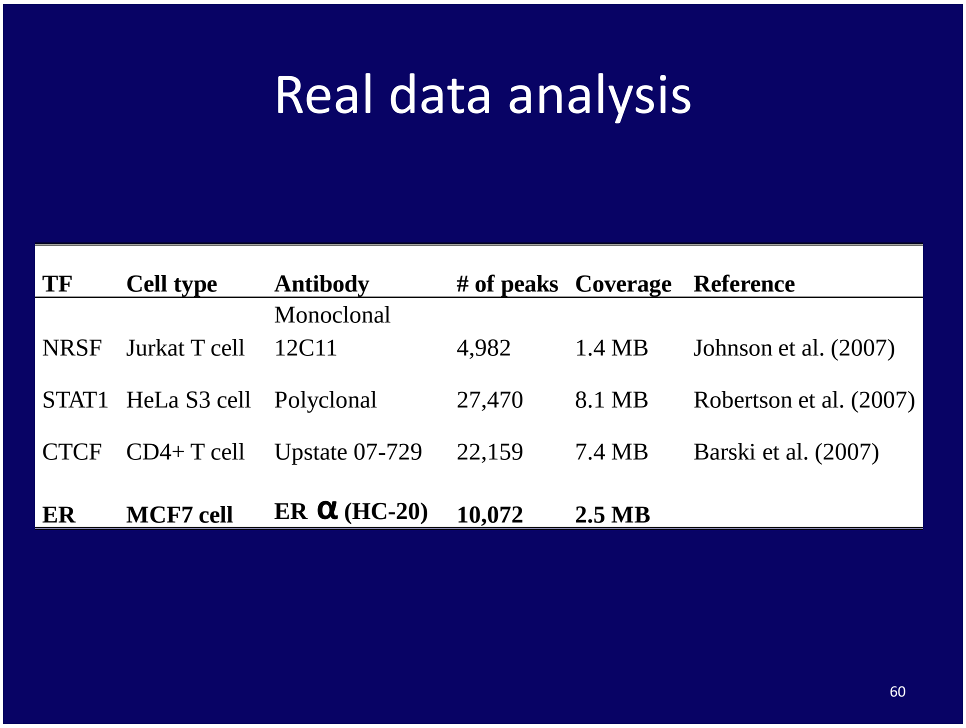
Implementation

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



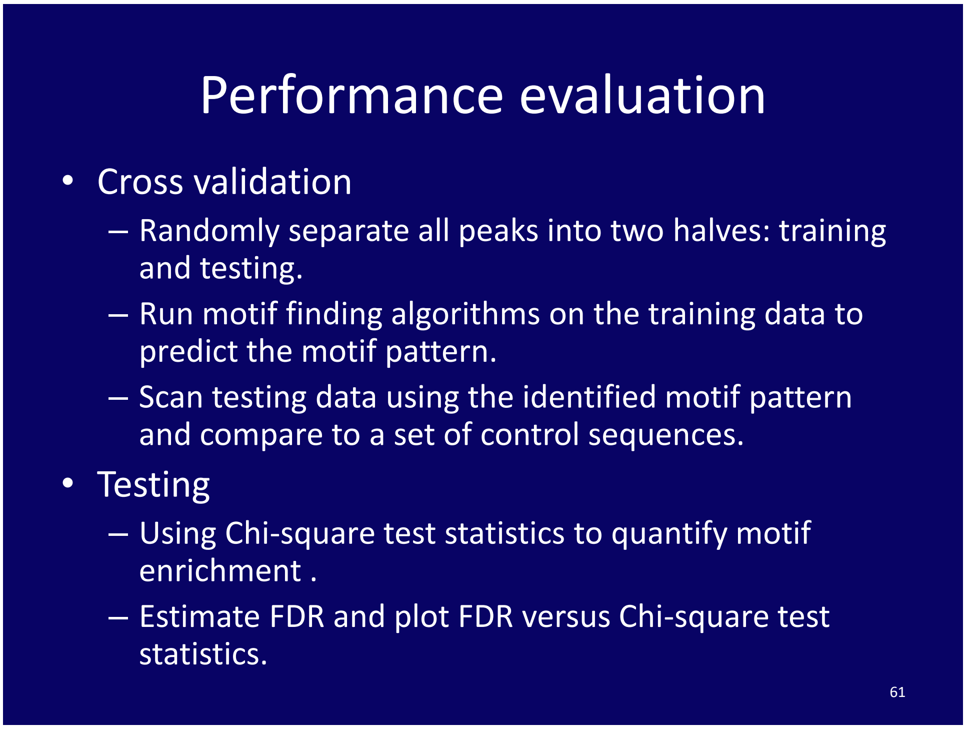
Performance comparison

- 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.



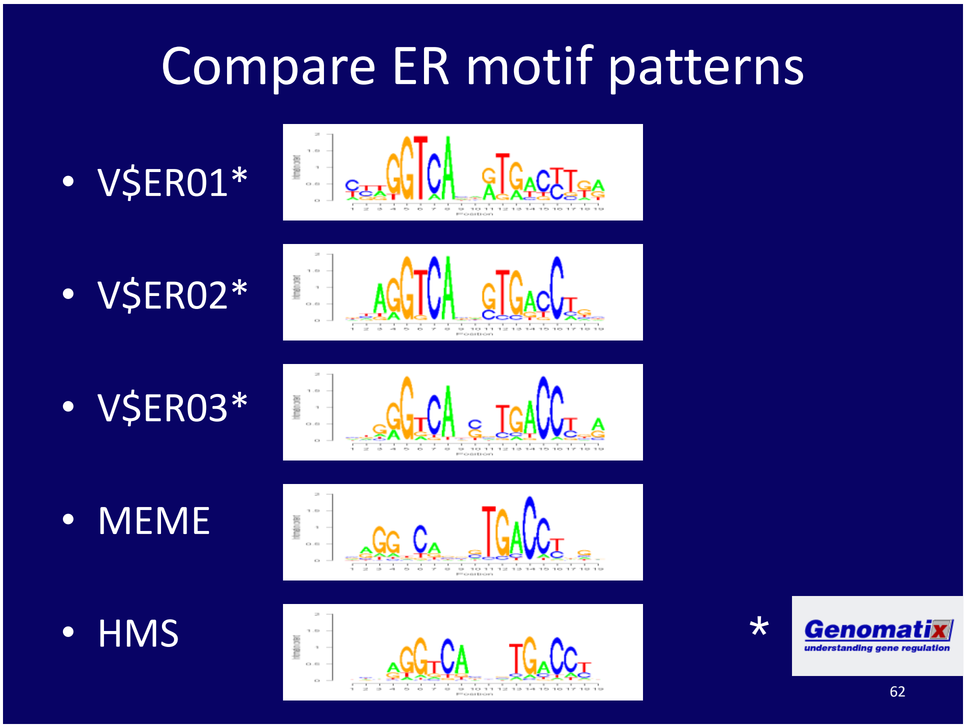
Real data analysis

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.

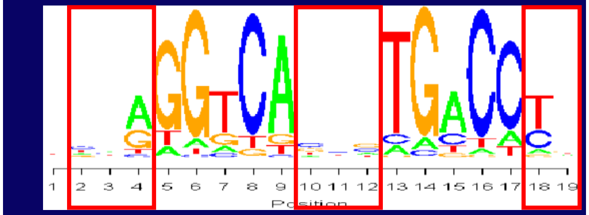


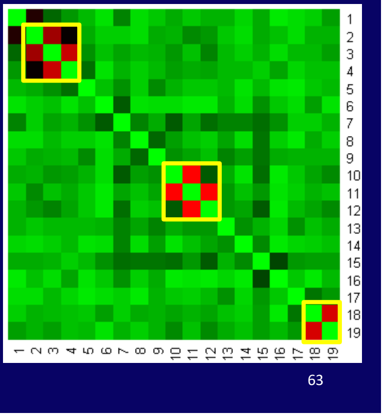
Compare ER motif patterns

- V$ER01*
- V$ER02*
- V$ER03*
- MEME
- HMS

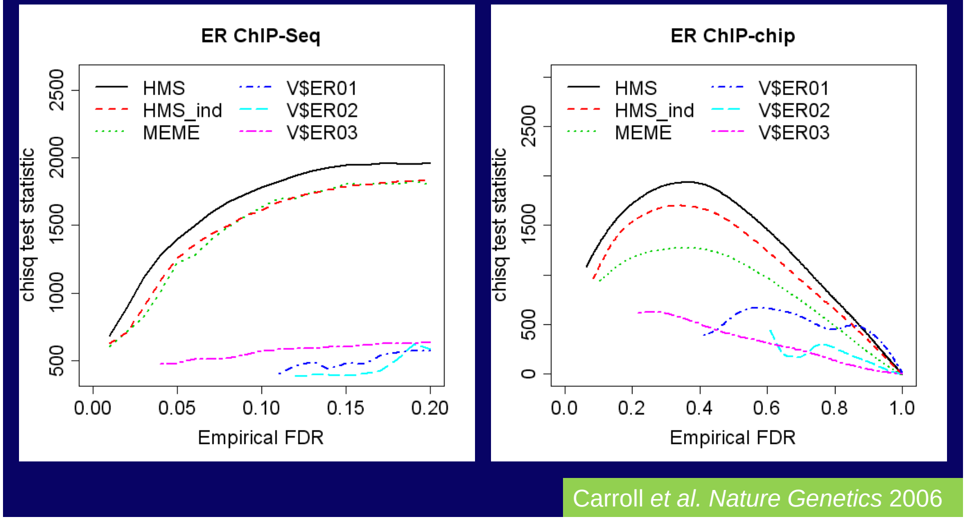
* Genomatix understanding gene regulation

Positions show inter-dependency inside the ER motif

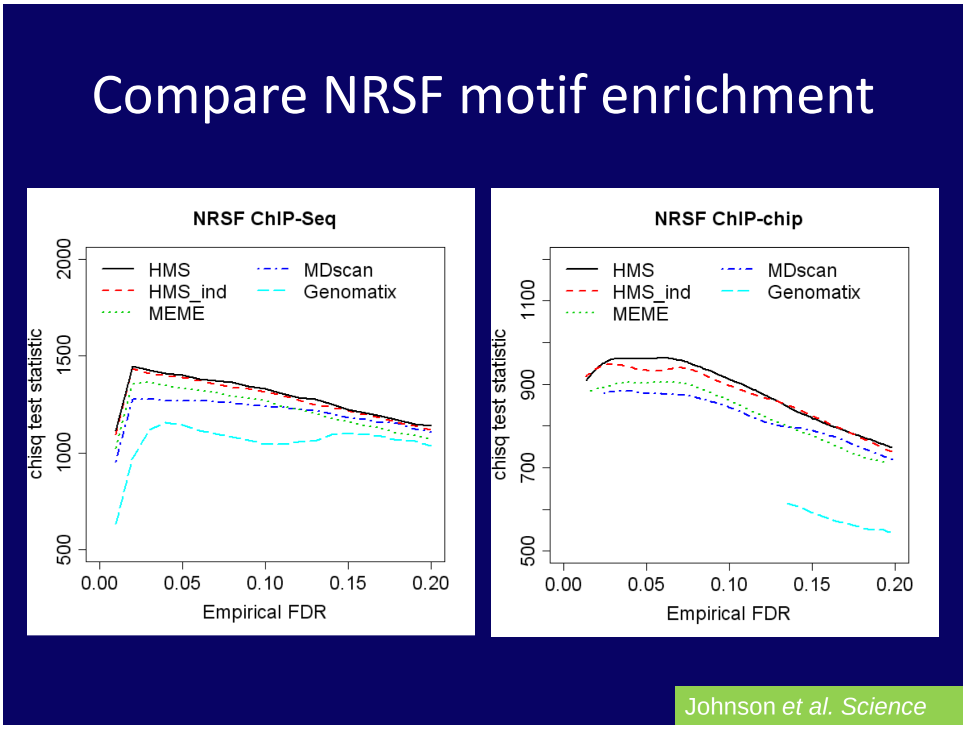




Compare ER motif enrichment

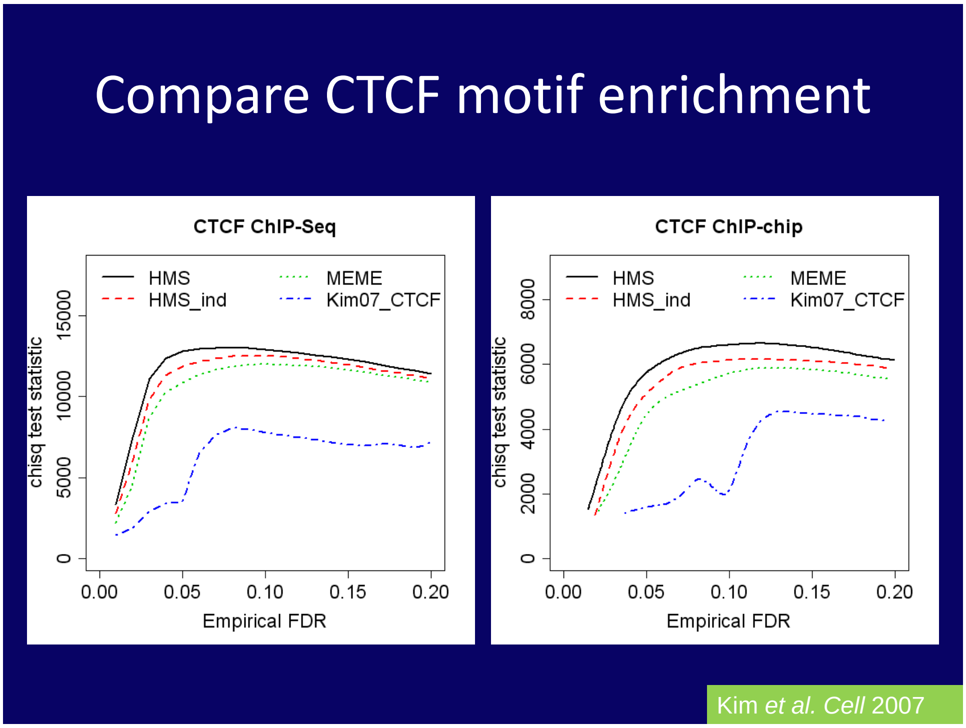


Carroll *et al. Nature Genetics* 2006



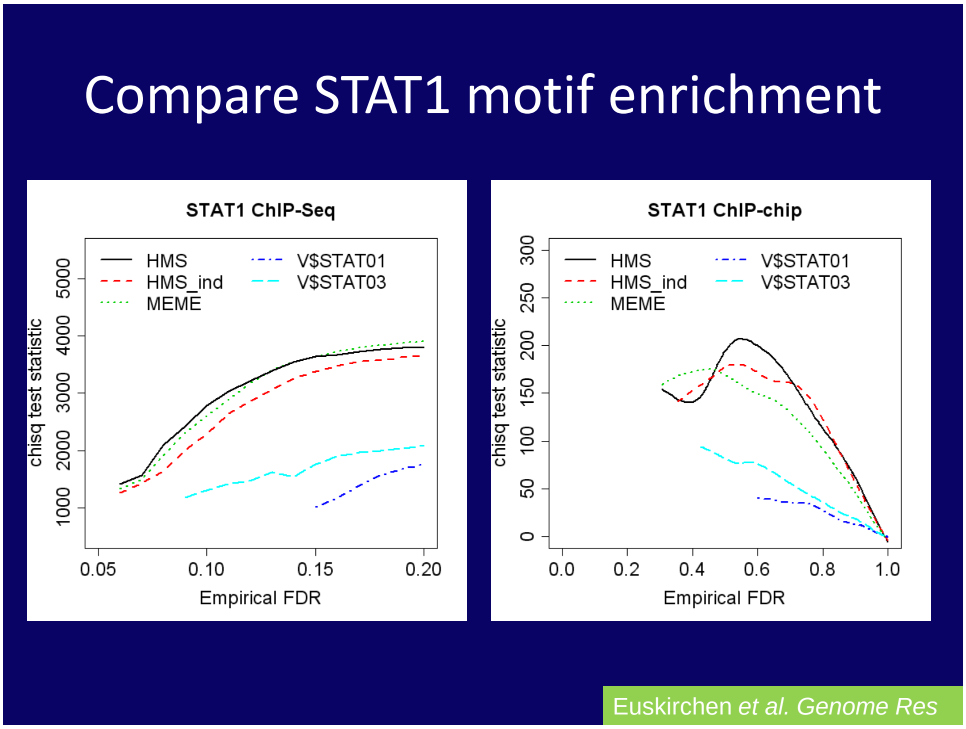
Compare NRSF motif enrichment

Johnson *et al. Science*



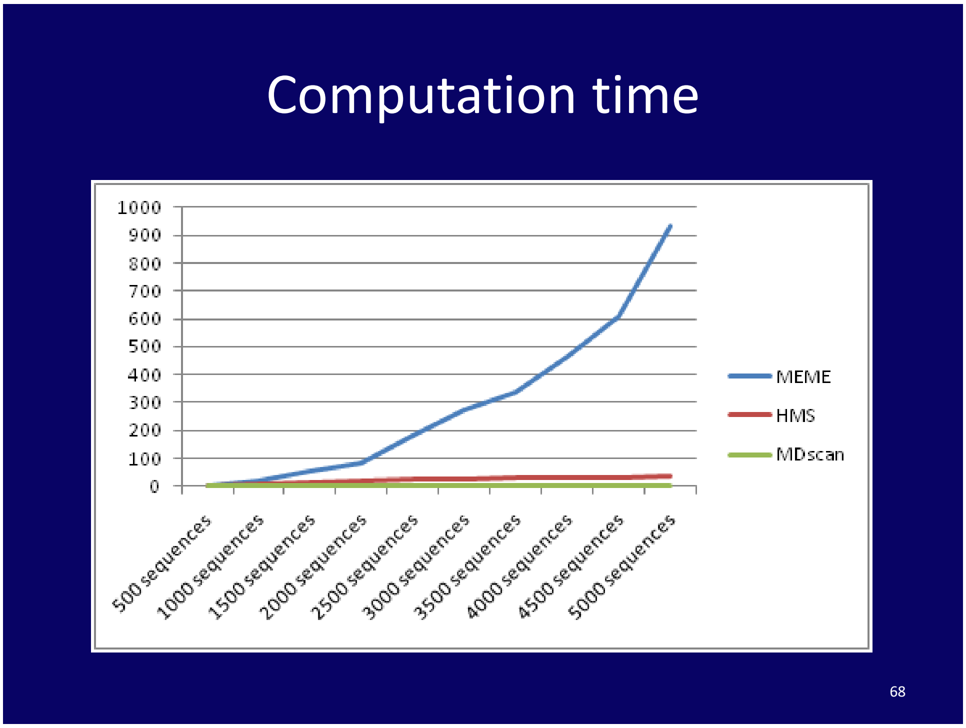
Compare CTCF motif enrichment

Kim *et al. Cell* 2007



Compare STAT1 motif enrichment

Euskirchen *et al. Genome Res*



Computation time

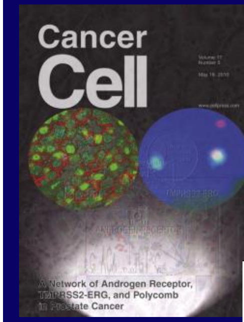
68

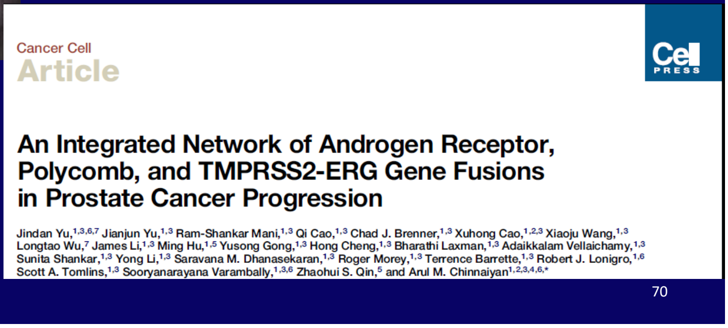
Summary

- 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.

69

Apply to cancer genomics


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



Cancer Cell

Article

Cell PRESS

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,*]

70

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.

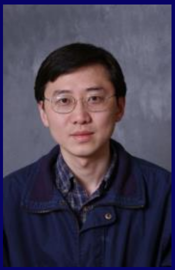
  http://www.sph.umich.edu/csg/qin/HMS/
- Hu M, Zhu Y, Taylor JMG, Liu JS, Qin ZS (2011). Using Poisson mixed-effects model to quantify exon-level gene expression in RNA-seq. *Bioinformatics.* **28** 63-68.

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

Acknowledgement

Ming Hu

Cleveland Clinics



Michael Yu Zhu

Purdue University