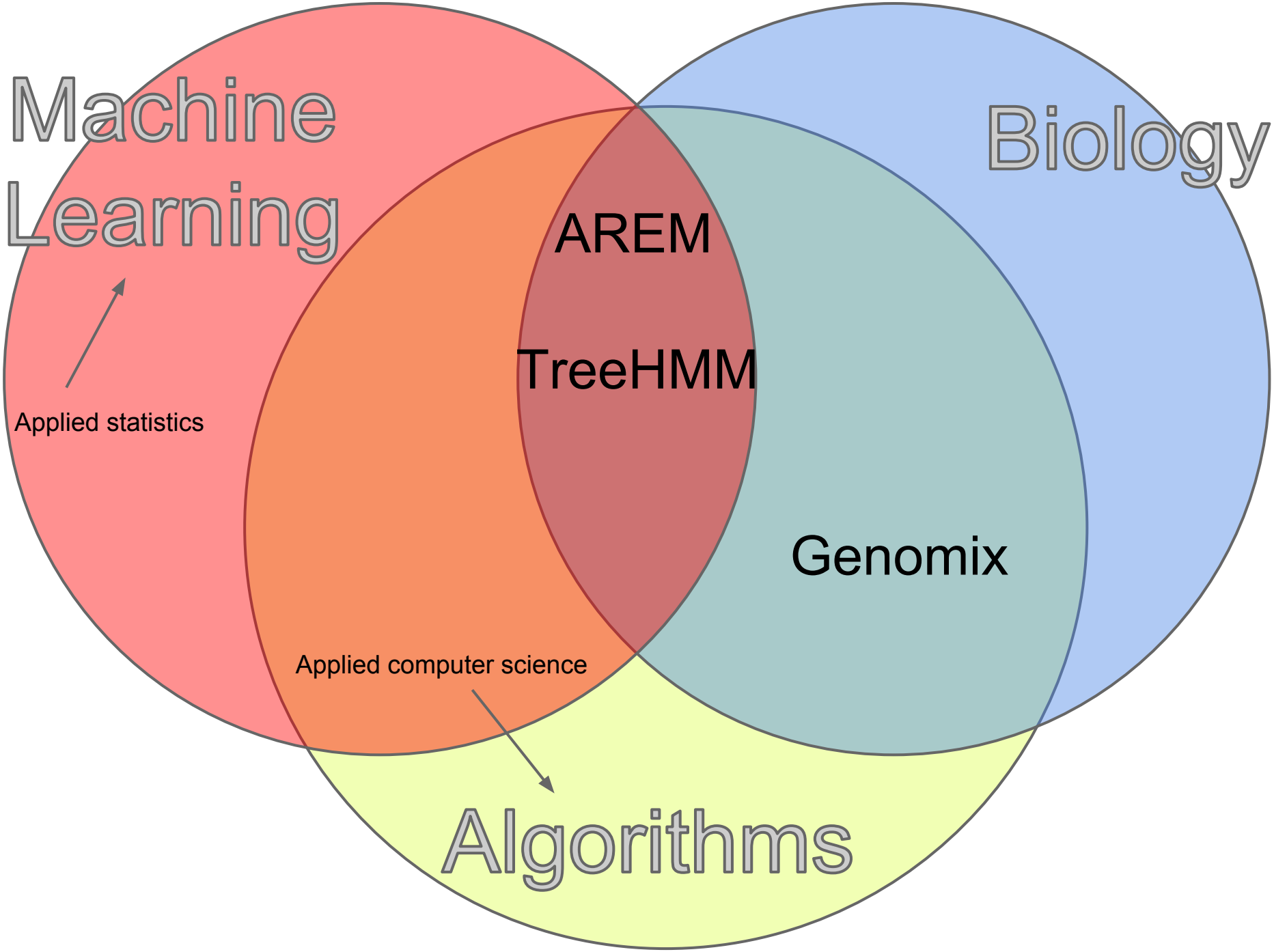
Toward a More Accurate Genome: Algorithms for the Analysis of High-Throughput Sequencing Data

Dissertation Defense
W. Jacob B. Biesinger

Tuesday, May 27th


Machine Learning
Biology
AREM
TreeHMM
Genomix
Applied statistics
Applied computer science
Algorithms

REFEREED JOURNAL PUBLICATIONS

Integrative ChIP-seq/Microarray Analysis Identifies a CTNNB1 Target Signature Enriched in Intestinal Stem Cells and Colon Cancer
PLOS One

Biallelic genome modification in F0 Xenopus tropicalis embryos using the CRISPR/Cas system
genesis

Discovering and mapping chromatin states using a tree hidden Markov model
BMC Bioinformatics

Transcriptome-wide analyses of CstF64–RNA interactions in global regulation of mRNA alternative polyadenylation
Proceedings of the National Academy of Sciences

Genome-wide analysis of hepatic LRH-1 reveals a promoter binding preference and suggests a role in regulating genes of lipid metabolism in concert with FXR
BMC Genomics

AREM: aligning short reads from ChIP-sequencing by expectation maximization
Journal of Computation Biology

Genome-wide localization of SREBP-2 in hepatic chromatin predicts a role in autophagy
Cell Metabolism

Combined biological and computational approaches to understand the role of Get1/Grhl3 in epidermal differentiation
Journal of Investigative Dermatology

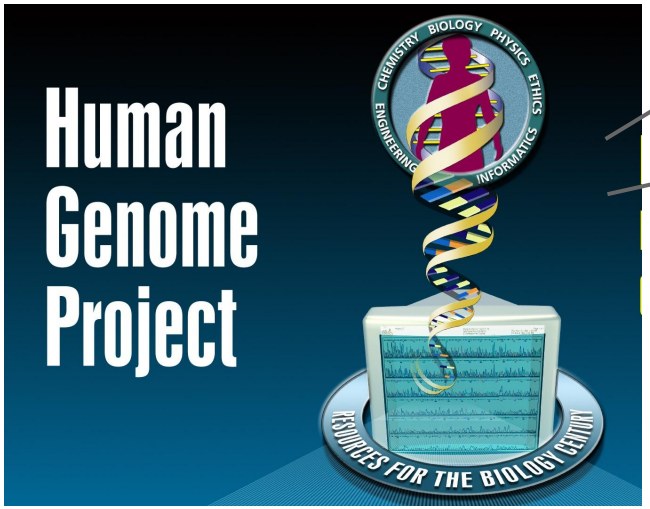
ChIP-seq

iCLIP-seq

Computational Methods

Probabilistic Models

A brief history of DNA sequencing



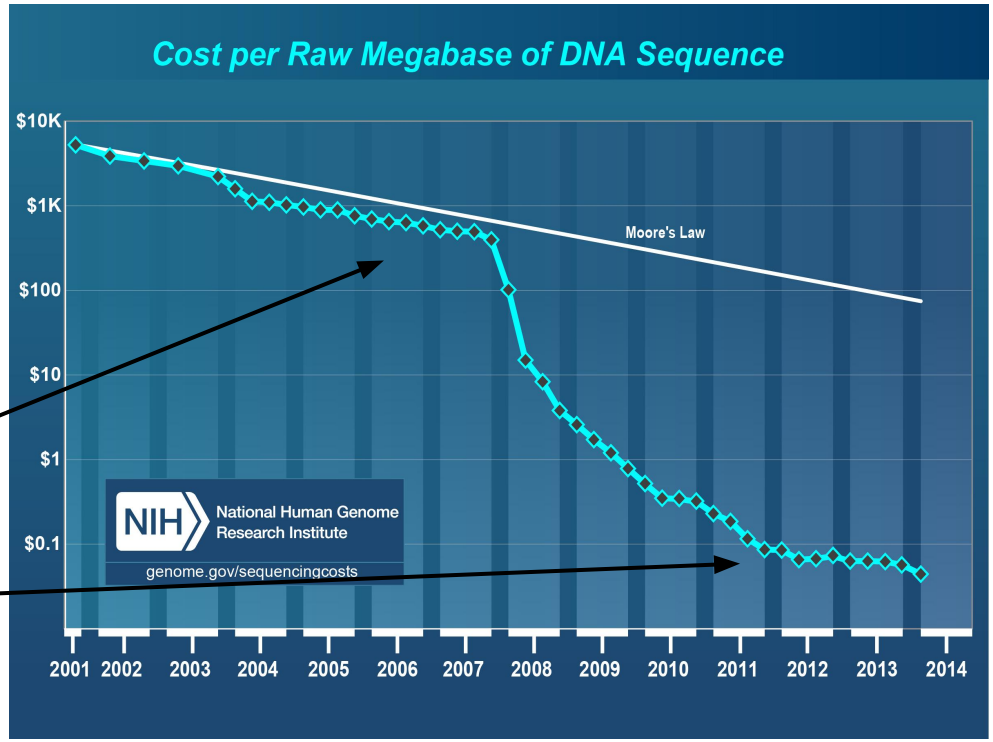
New biological insight

Seeded technological revolution: high-throughput sequencing

- Completed in 2003 at a cost of $3 billion and 10 years of labor and planning
- First time we've determined the sequence of a large genome

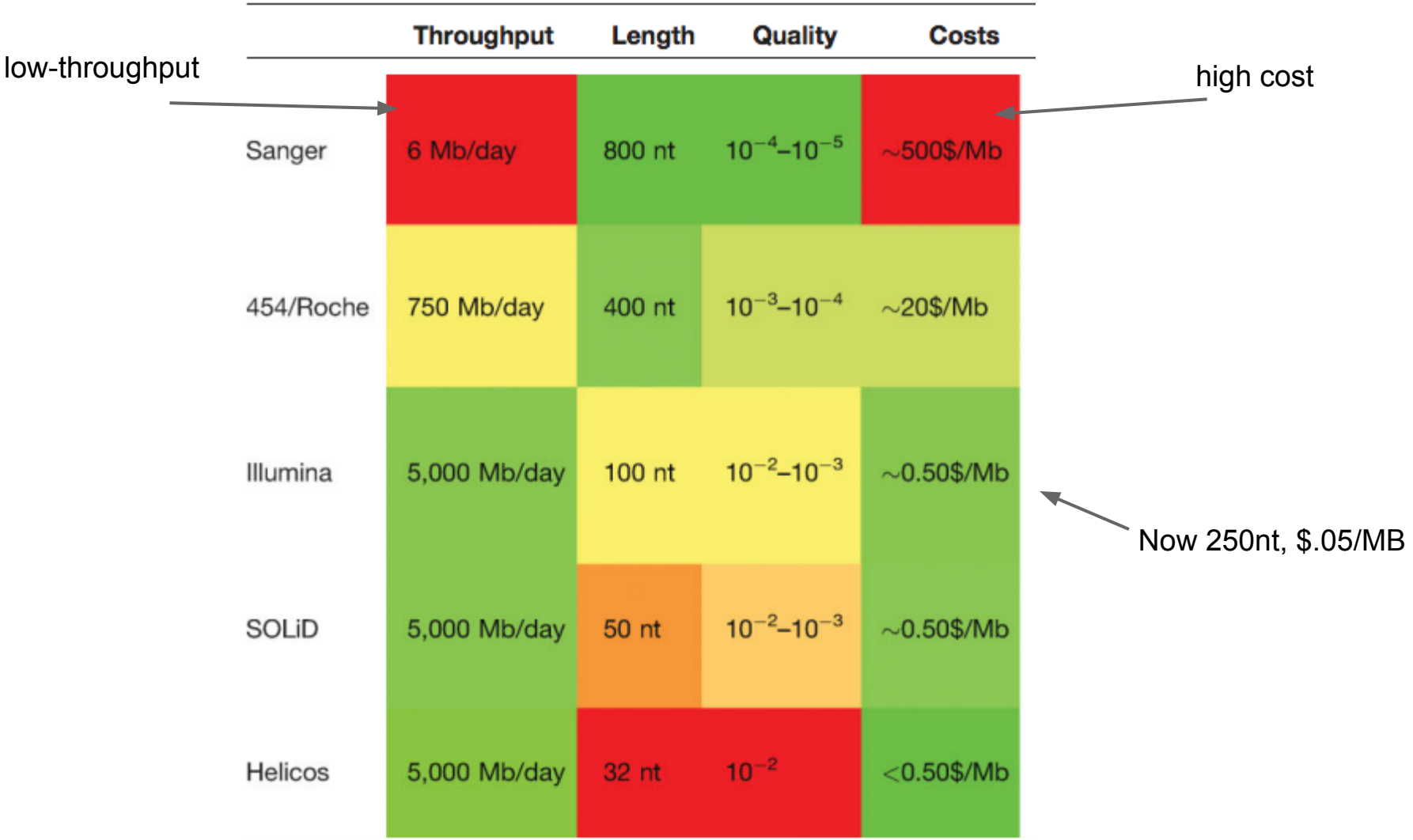
XPRIZE: $10 million for 100 genomes @ $1,000 each

XPRIZE cancelled: "Outpaced by Innovation"



http://www.genome.gov/sequencingcosts/

A brief history of DNA sequencing



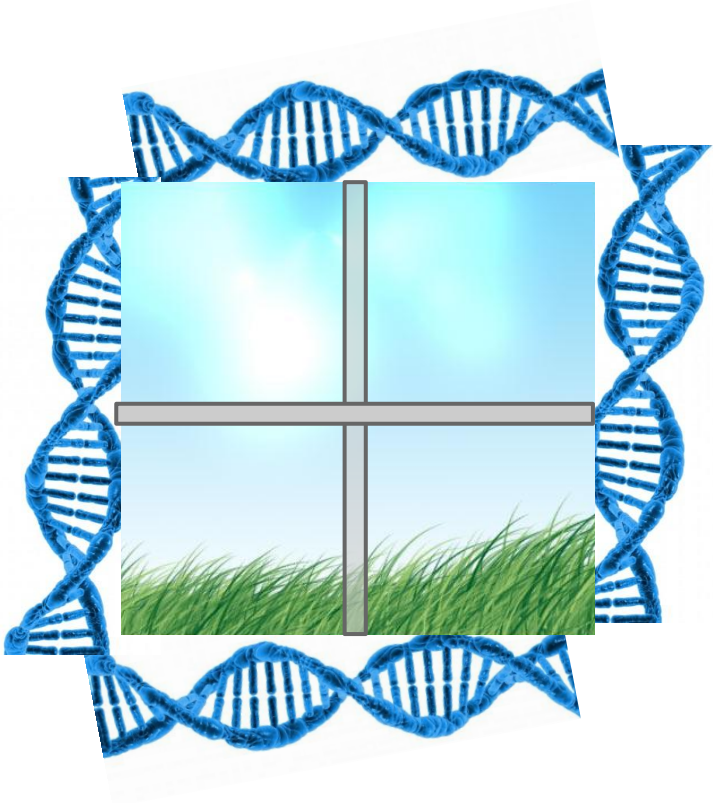
Kercher et al., Bioessays (2010).

A revolution in biology

- HTS has changed the way that much of biology is done today

New experimental methods

Targeted resequencing
Whole-genome sequencing
Exome sequencing
ChIP-seq
RNA-seq
MeDIP-seq
CLIP-seq
ChIRP-seq
Hi-C
ChIA-PET



New (or rebranded) fields of study

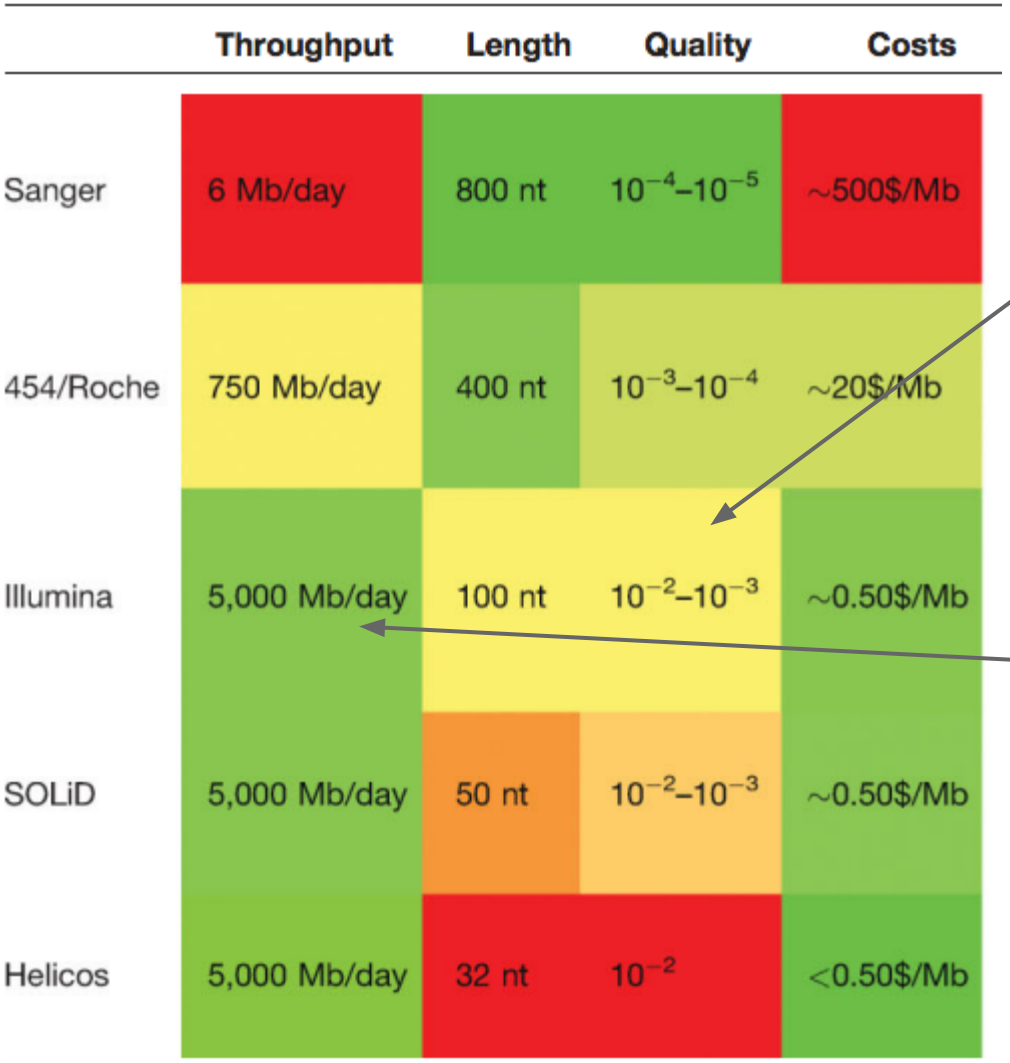
Genomics
Transcriptomics
Metabolomics
Microbiomics
Toxicogenomics
Epigenomics
Interactonomics
Circadiomics

HTS (already) has real impact

- Clinical Impact
  - Discovery of inheritable genetic disorders
  - Cancer biology (identify cancer subtypes)
  - Evolution and spread of infectious diseases
  - Prenatal diagnostics
  - Now transitioning into clinical laboratory
  - Lead to personalized therapies

- Basic Biology
  - Gene expression levels
  - Identify regulatory network structure
  - Elucidate fundamental biological processes
    - find promoter TATA binding, splicing mechanisms, the drivers of cellular state/stem cell “stemness”

Limitations of HTS methods



	Throughput	Length	Quality	Costs
Sanger	6 Mb/day	800 nt	10^{-4}–10^{-5}	~500$/Mb
454/Roche	750 Mb/day	400 nt	10^{-3}–10^{-4}	~20$/Mb
Illumina	5,000 Mb/day	100 nt	10^{-2}–10^{-3}	~0.50$/Mb
SOLiD	5,000 Mb/day	50 nt	10^{-2}–10^{-3}	~0.50$/Mb
Helicos	5,000 Mb/day	32 nt	10^{-2}	<0.50$/Mb

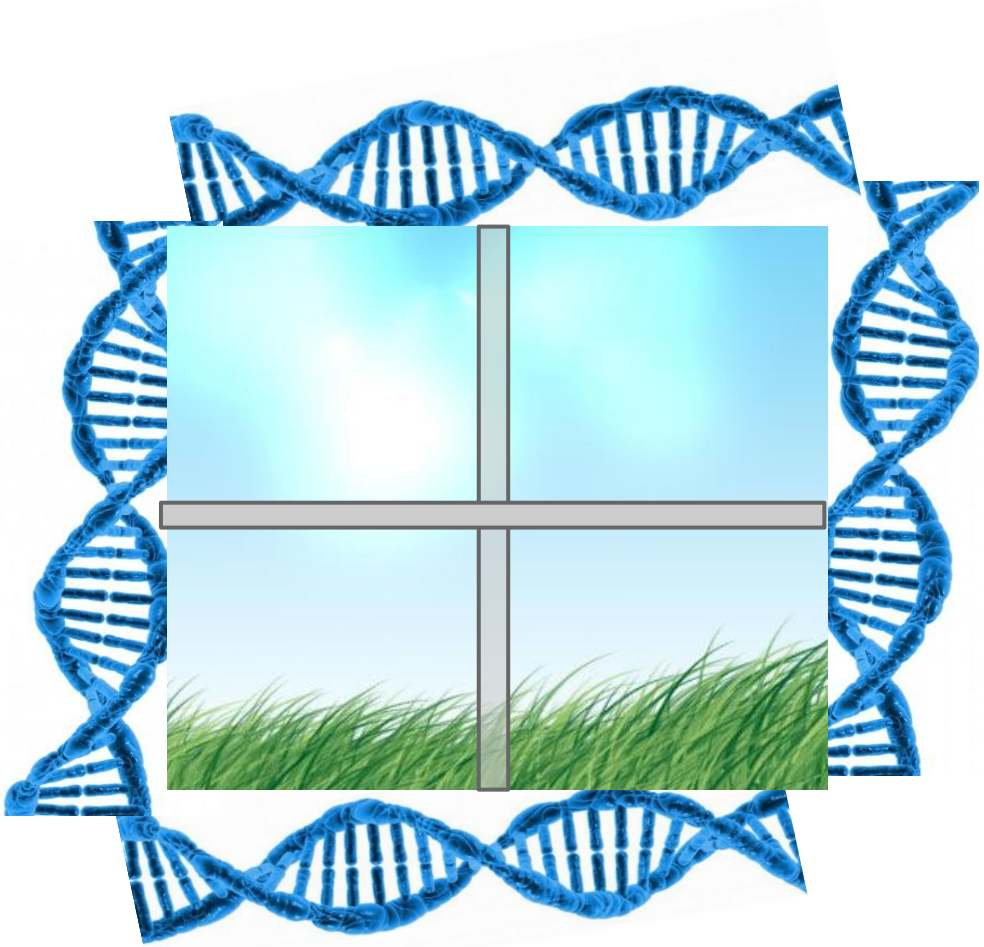
You can't trust 1/100 bases

We all **wish** the error rate were uniform

All kinds of hidden biases based on the sequence composition (GC-content, strand-bias, positional bias,

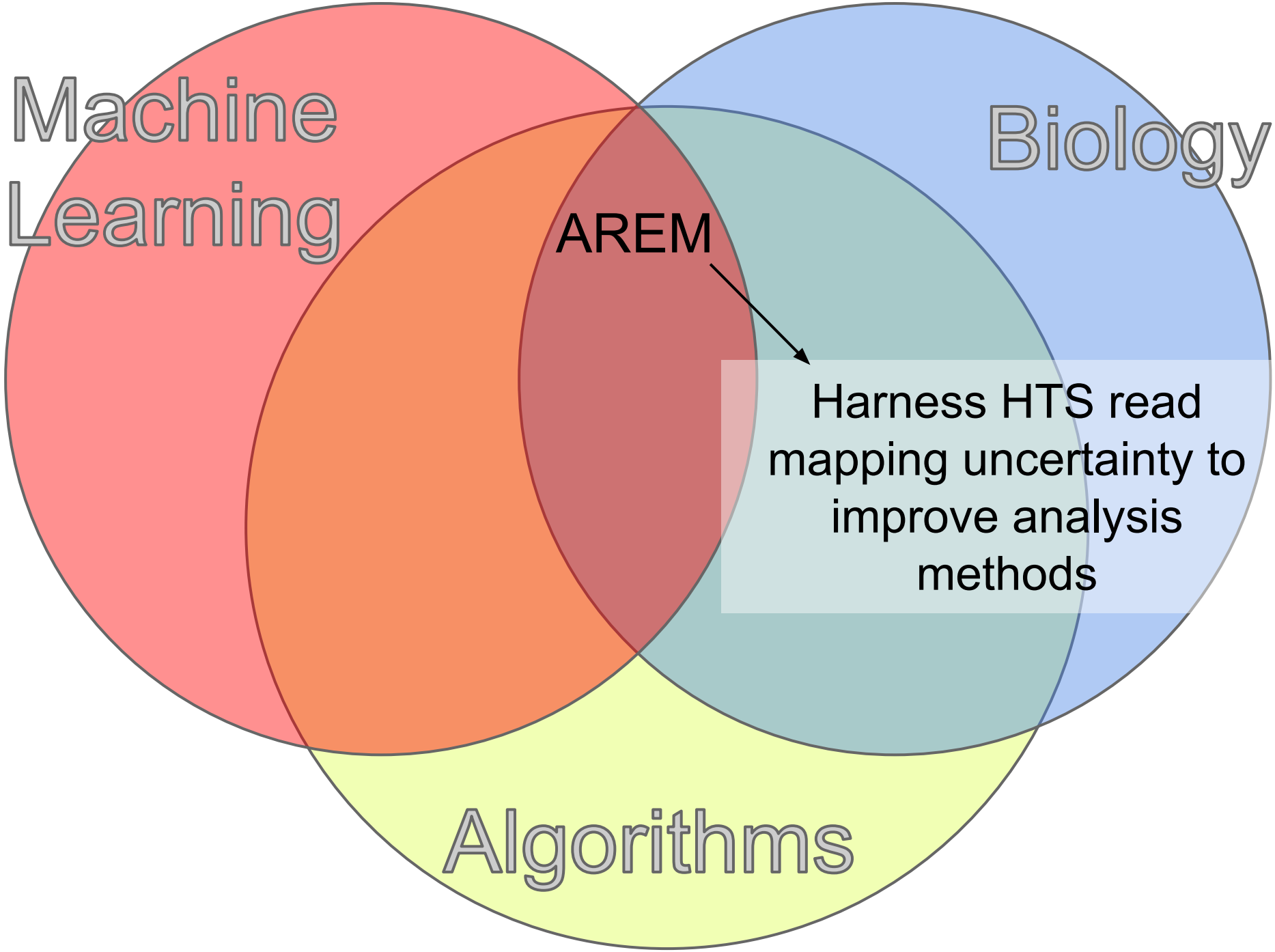
But we have much more data. How can we best use it all?

Computational biology to the rescue!



Detect and correct errors and biases

See the biology beyond the letters


Machine Learning
Biology
AREM
Harness HTS read mapping uncertainty to improve analysis methods
Algorithms

Resolve ambiguity through Machine Learning

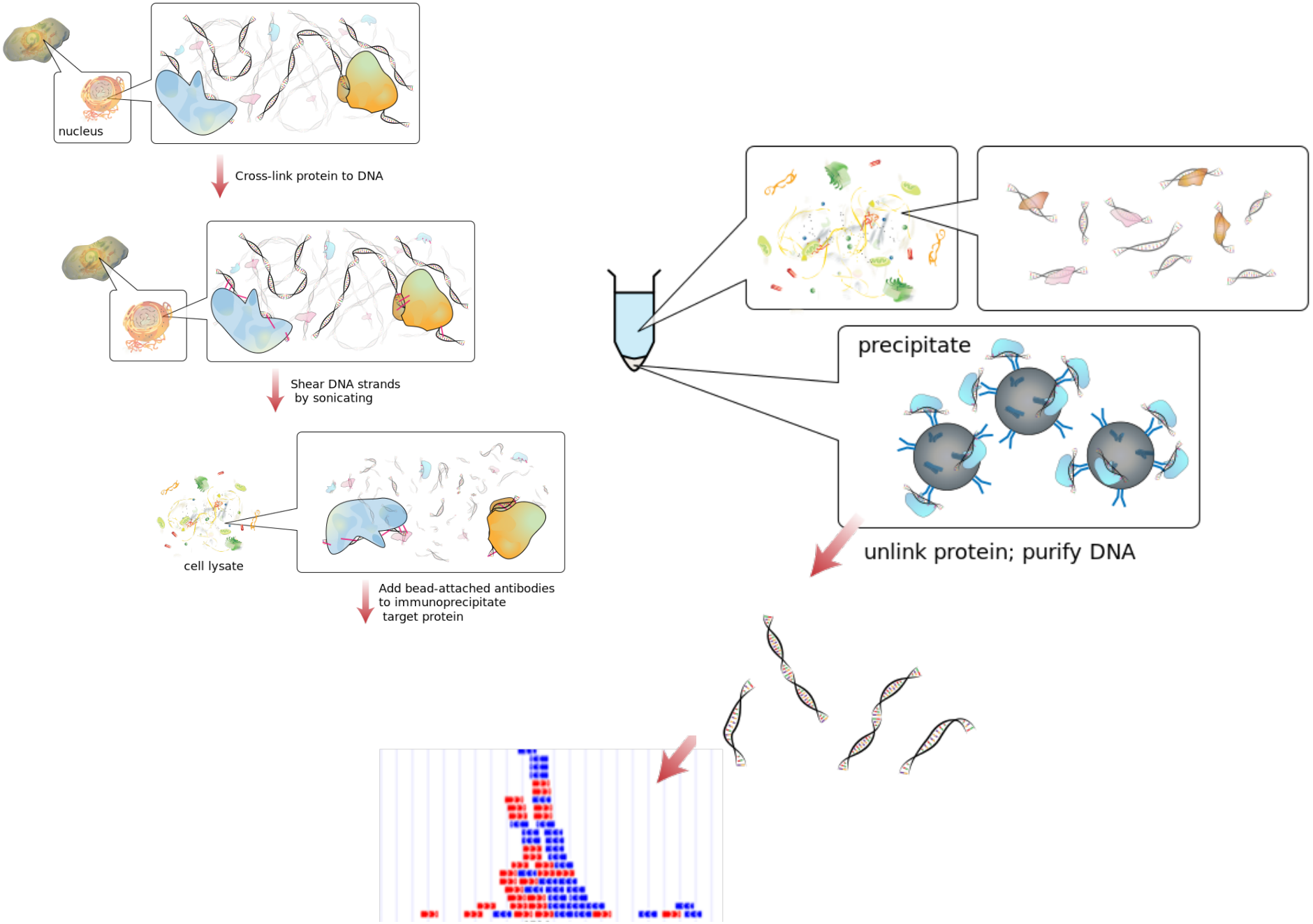
- Most genomes are riddled with *repetitive sequence*
  - Variable lengths (six to several thousand bp)
  - Up to 66% of the Human genome*
  - ~30% of reads map ambiguously**
  - Ambiguous reads often excluded completely or some subset are included at random

AREM: Aligning Reads by Expectation-Maximization

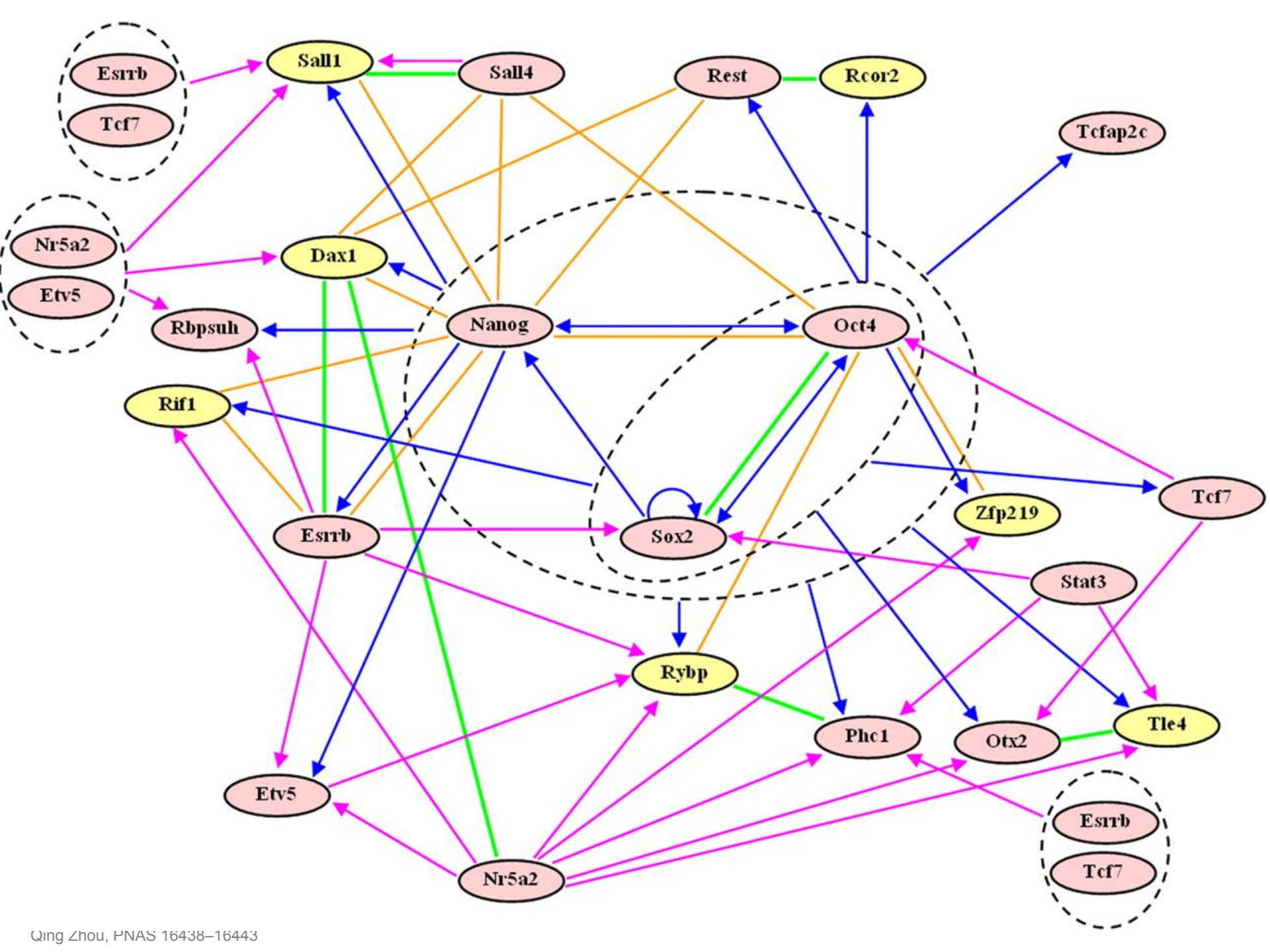
General framework for resolving repeats; we demonstrate how with ChIP-seq data

*Koning et al. PLOS Genetics 7 (12): e1002384

**Langmead et. al. Genome Biology 10 (2009) R25



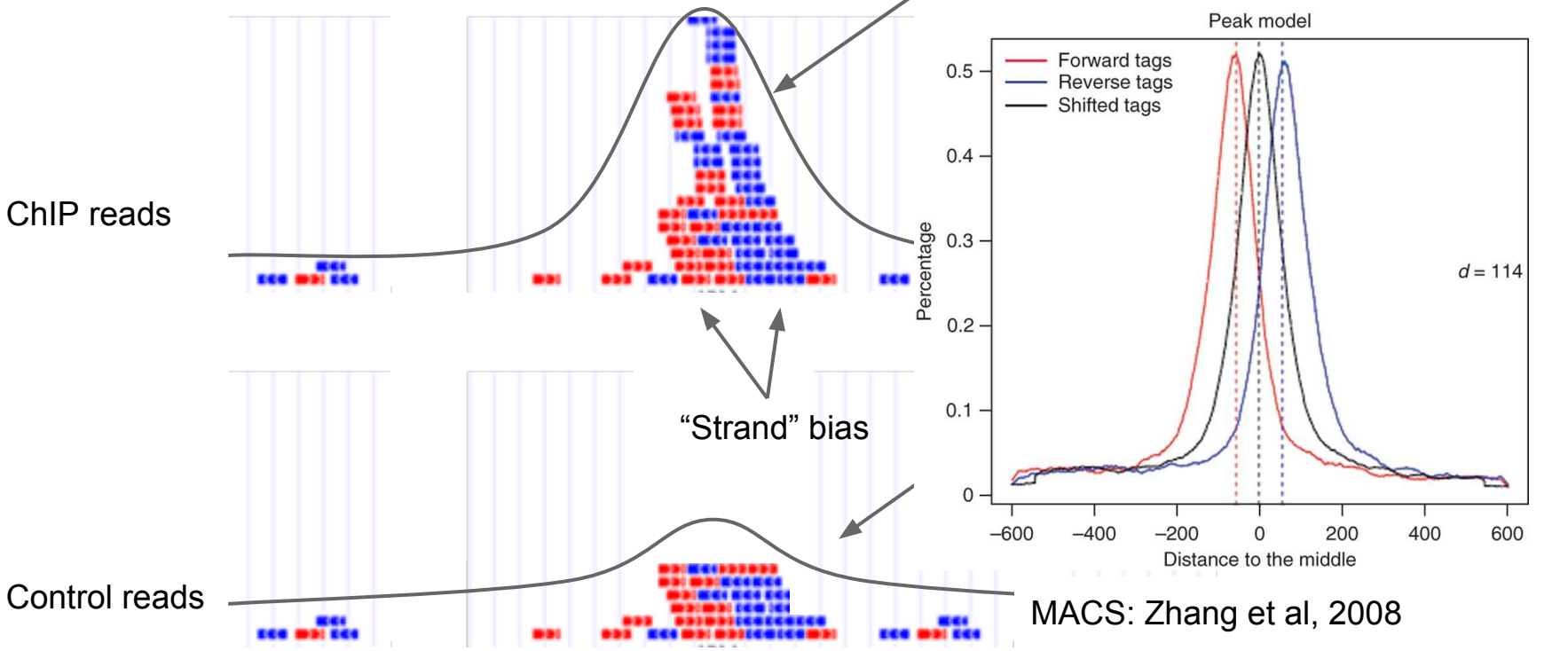
wikipedia.org/wiki/ChIP-sequencing


Esrrb
Tcf7
Sall1
Sall4
Rest
Rcor2
Tcfap2c
Nr5a2
Etv5
Dax1
Rbpsuh
Nanog
Oct4
Rif1
Tcf7
Esrrb
Sox2
Zfp219
Stat3
Rybp
Phc1
Otx2
Tle4
Etv5
Esrrb
Tcf7
Nr5a2

Identifying Peaks

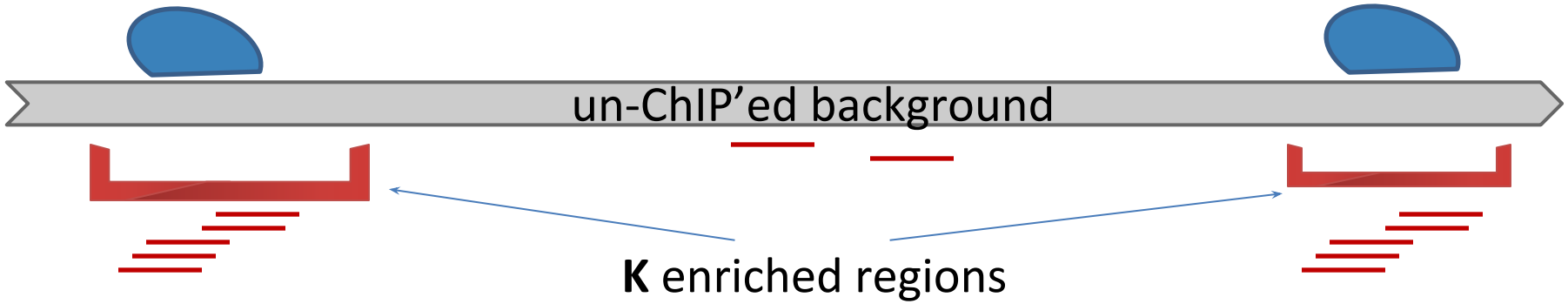
- Look for regions with many reads piled together

Treat as Nx1 dataset (N is in 10's of milions)
Smooth via kernel density



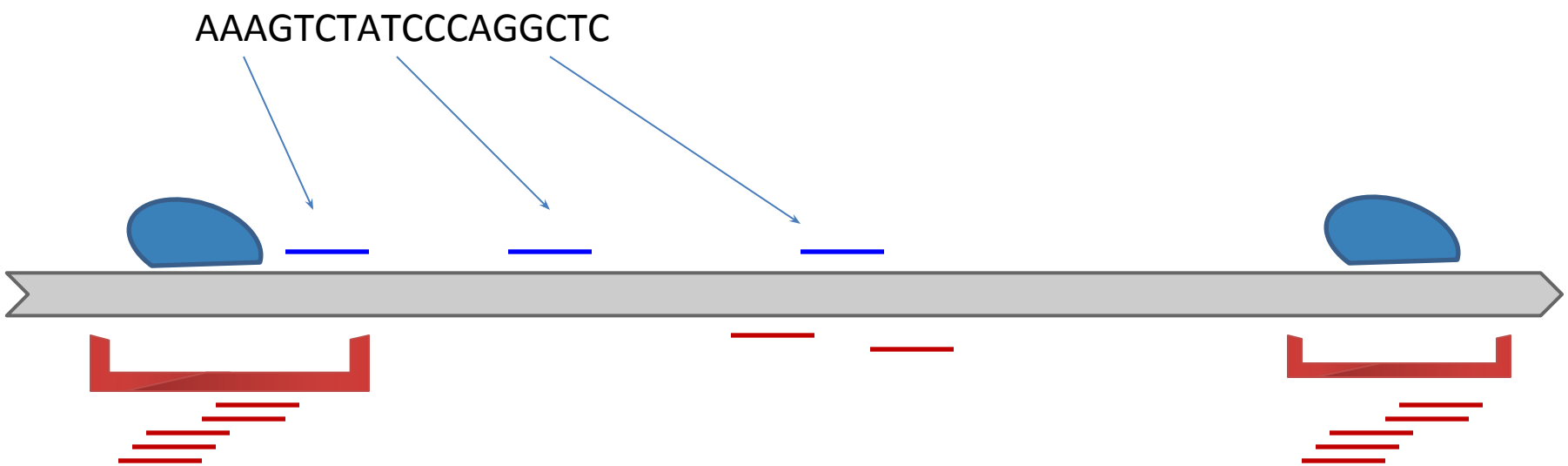
MACS: Zhang et al, 2008

A mixture model for ChIP-Seq



- Each read has some probability of belonging to each of the peak and background regions
- Identify best peak configuration by maximizing read likelihood

A mixture model for ChIP-Seq



- Which region is the most likely **source** of the ambiguous reads?
- The alignment with highest likelihood
- (Not so simple if we're unsure where the K enriched regions are located)

Maximize Likelihood via E-M

$$\hat{\Theta} = \arg\max_{\Theta} p(r_1, r_2, \cdots r_N; \Theta)$$

Overall problem: find best peak configuration

$$p(r_i) = \sum_{u_i \in K, z_i \in n_i} p(r_i | u_i = k, z_i = j) p(u_i = k) p(z_i = j)$$

Consider all possible peak sources and all possible alignments

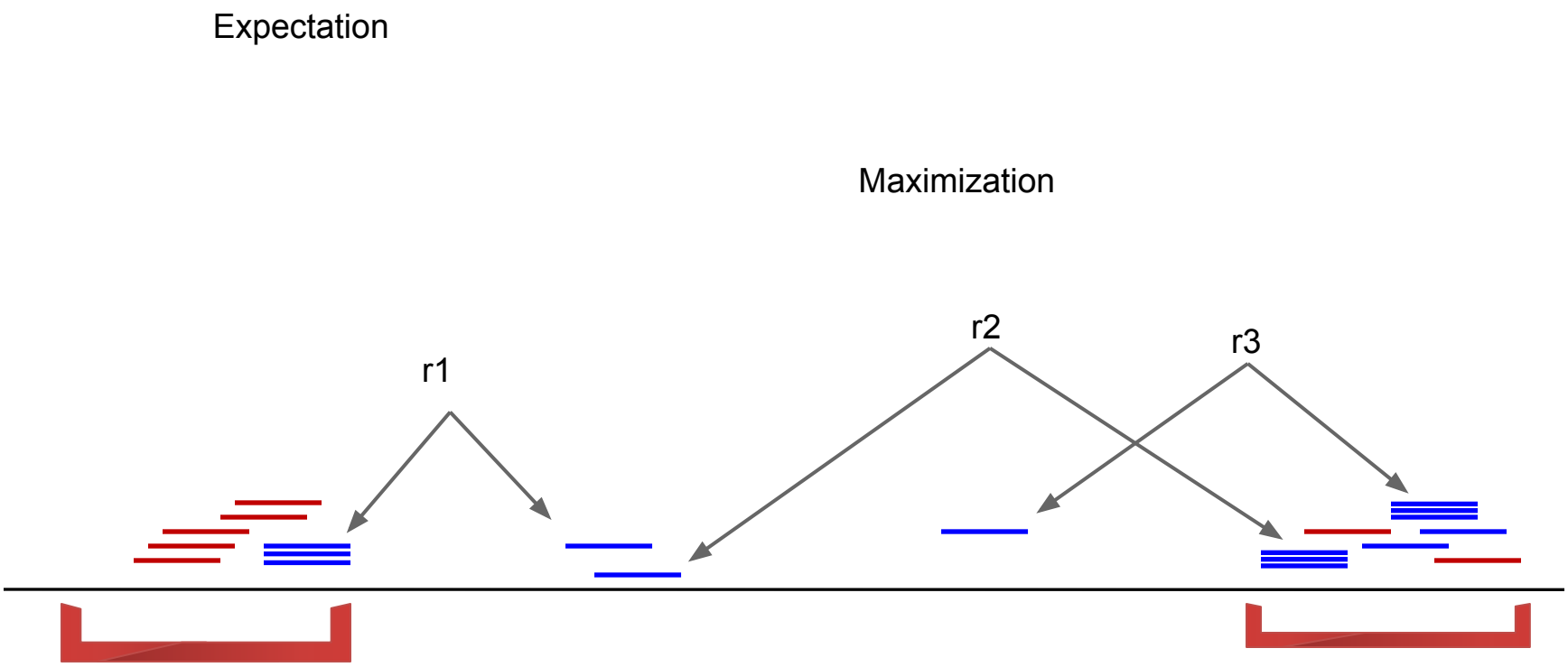
Expectation (With regions fixed, update alignments)

Maximization
(With alignments fixed, find best regions)

$$Q^{(t)}(z_i = j | R) = \frac{1}{C} P(r_i | z_i = j, \Theta^{(t)}) P(z_i = j)$$

$$\Theta^{(t+1)} = \arg\max_{\Theta} \sum_{i=1}^{N} \sum_{j=0}^{n_i} Q^{(t)}(z_i = j | R) \log P(r_i | z_i = j, \Theta)$$

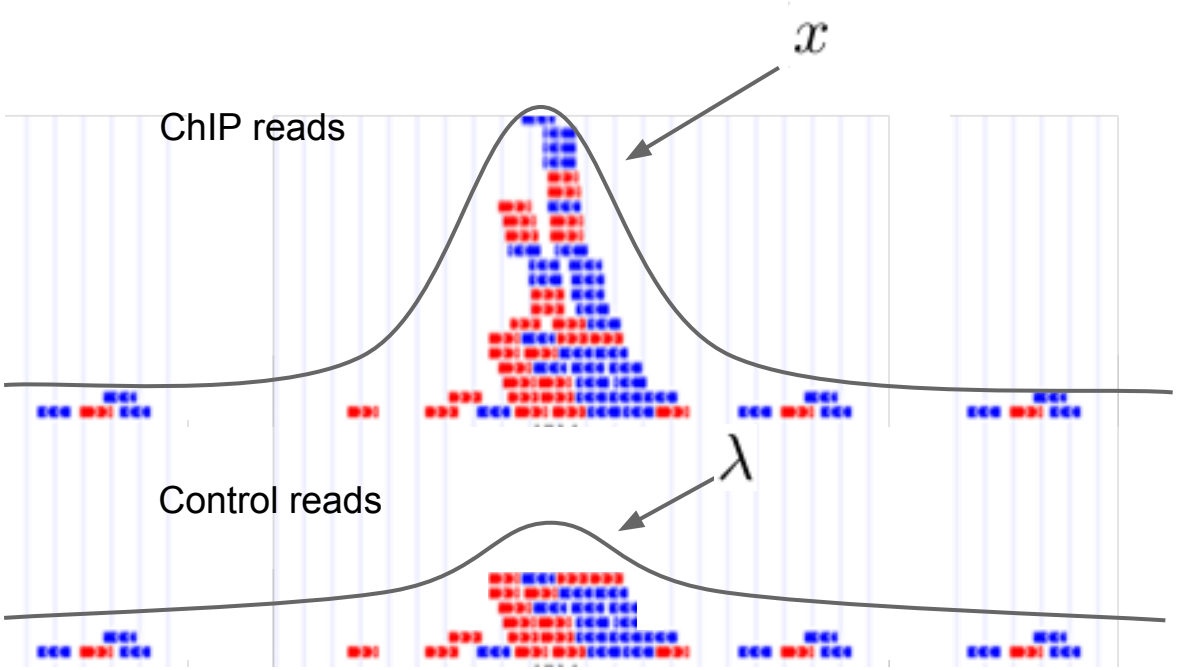
Expectation Maximization in action



E-M is a machine learning method with many applications, especially in mixture models.

Accounting for non-uniform control

- Define alignment likelihood as poisson survival of peak vs. unenriched background

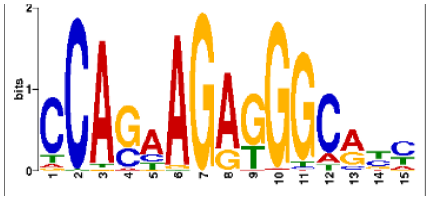


$$g(x,\lambda) = e^{-\lambda}\sum_{k=0}^{x}\frac{\lambda^k}{k!}$$

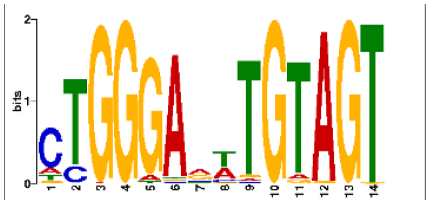
$$p(r_i|z_i = j, u_i = k) = \frac{1}{C_i}\max\left(-\log_{10}(1-g(x,\lambda)), m\right)$$

Test datasets

- We used motif presence to indicate peak quality
- **Cohesin –** structural protein, known to bind repetitive regions of the genome
  - D4Z4 sub-telomeric repeat associated with Facioscapulohumeral Disorder *
  - Cohesin often co-localizes with CTCF (motif in 80% peaks from mouse and human)
- **Srebp-1–** traditional transcription factor
  - Contains a well-characterized binding motif



CTCF binding motif



Srebp-1 binding motif

* Zeng et. al. PLoS Genetics, 5(7) 2009

AREM shows better performance in repeat regions than other peak finders

Cohesin

	Method		Alignments	Peaks	New	FDR	Motif	Repeat
1	MACS	---	2,368,229	18,556	---	2.80%	81.67%	56.55%
1	SICER	---	2,368,229	17,092	---	12.71%	82.55%	70.42%
1	AREM	1	2,368,229	19,012	---	1.90%	81.32%	55.30%
2	AREM	10	7,616,647	19,881	1,404	3.80%	81.04%	58.88%
2	AREM	20	12,312,878	19,935	1,517	3.70%	80.88%	59.66%
2	AREM	40	20,527,010	19,863	1,546	3.20%	80.93%	60.34%
2	AREM	80	34,537,311	19,820	1,538	2.90%	80.73%	60.91%

1. Allow for sequences with one alignment.
2. Allow for sequences with up to 10-80 possible alignments.

8% more peaks, similar FDR, many peaks in repeats!

AREM shows better performance in repeat regions than other peak finders

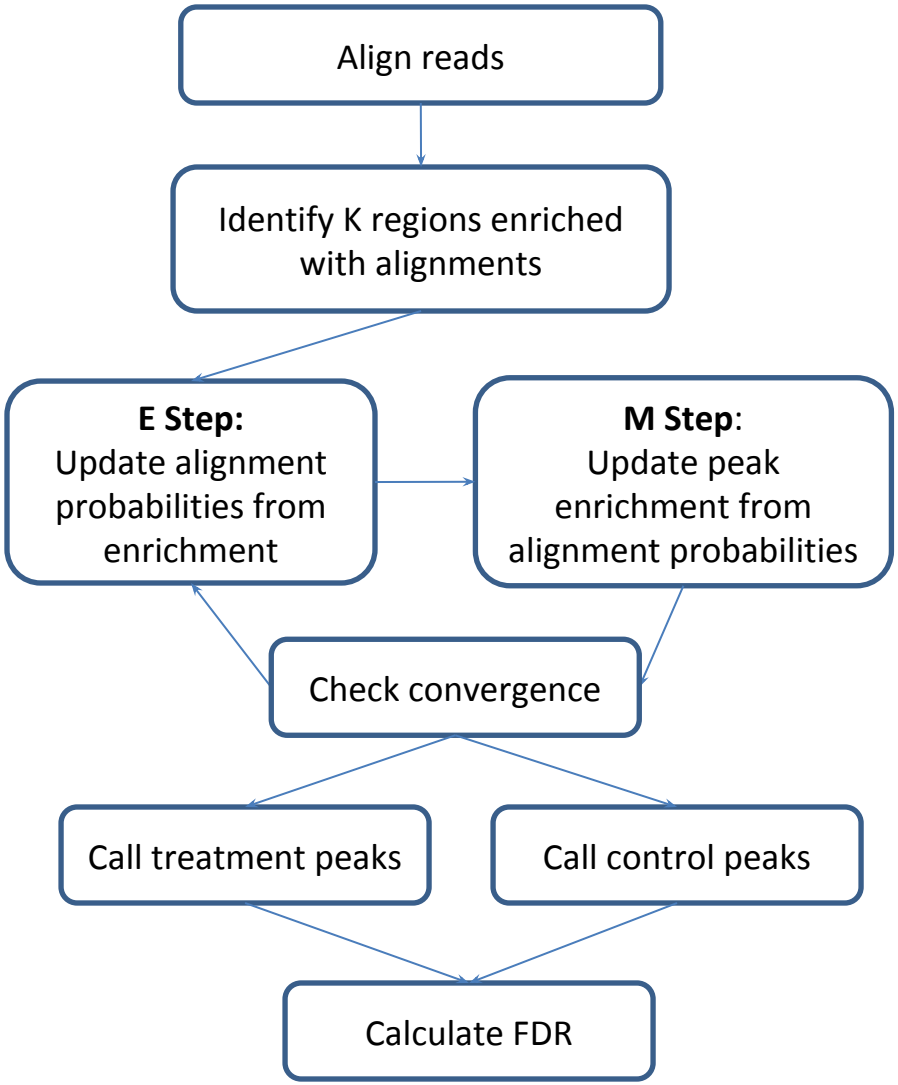
Srebp-1

	Method		Alignments	Peaks	New	FDR	Motif	Repeat
1	MACS	---	10,482,005	721	---	4.85%	46.60%	53.95%
1	SICER	---	10,482,005	622	---	9.0%	59.00%	77.33%
1	AREM	1	10,482,005	1,438	---	8.0%	39.08%	53.47%
2	AREM	10	28,347,869	1,815	262	10.5%	39.22%	56.04%
2	AREM	20	44,493,532	1,748	227	8.0%	39.95%	55.97%
2	AREM	40	72,453,642	1,685	248	8.2%	40.34%	56.46%
2	AREM	80	118,744,757	1,695	272	7.3%	40.66%	56.73%

5% more peaks called at lower FDR

1. Allow for sequences with one alignment.
2. Allow for sequences with up to 10-80 possible alignments.

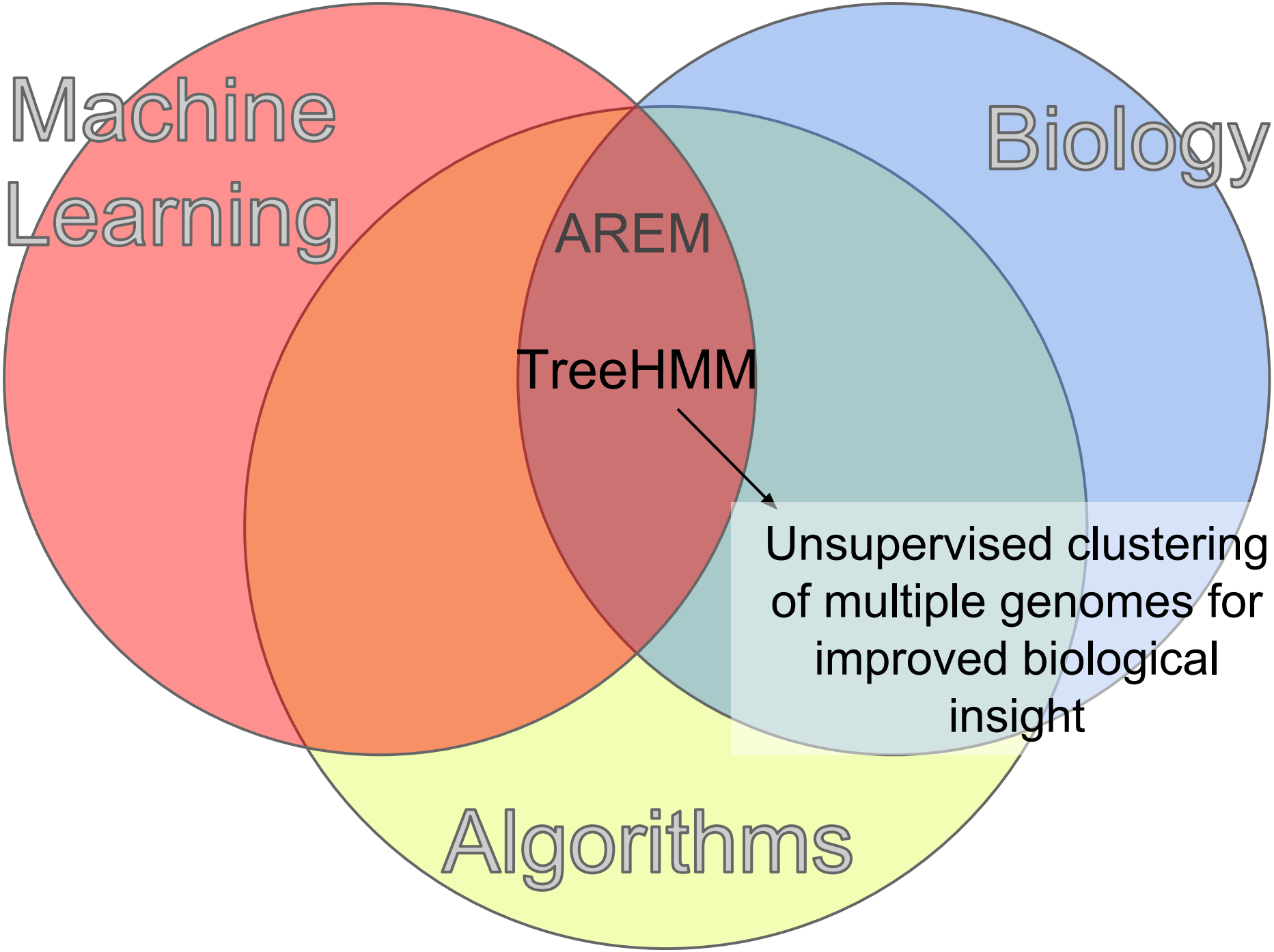
Availability



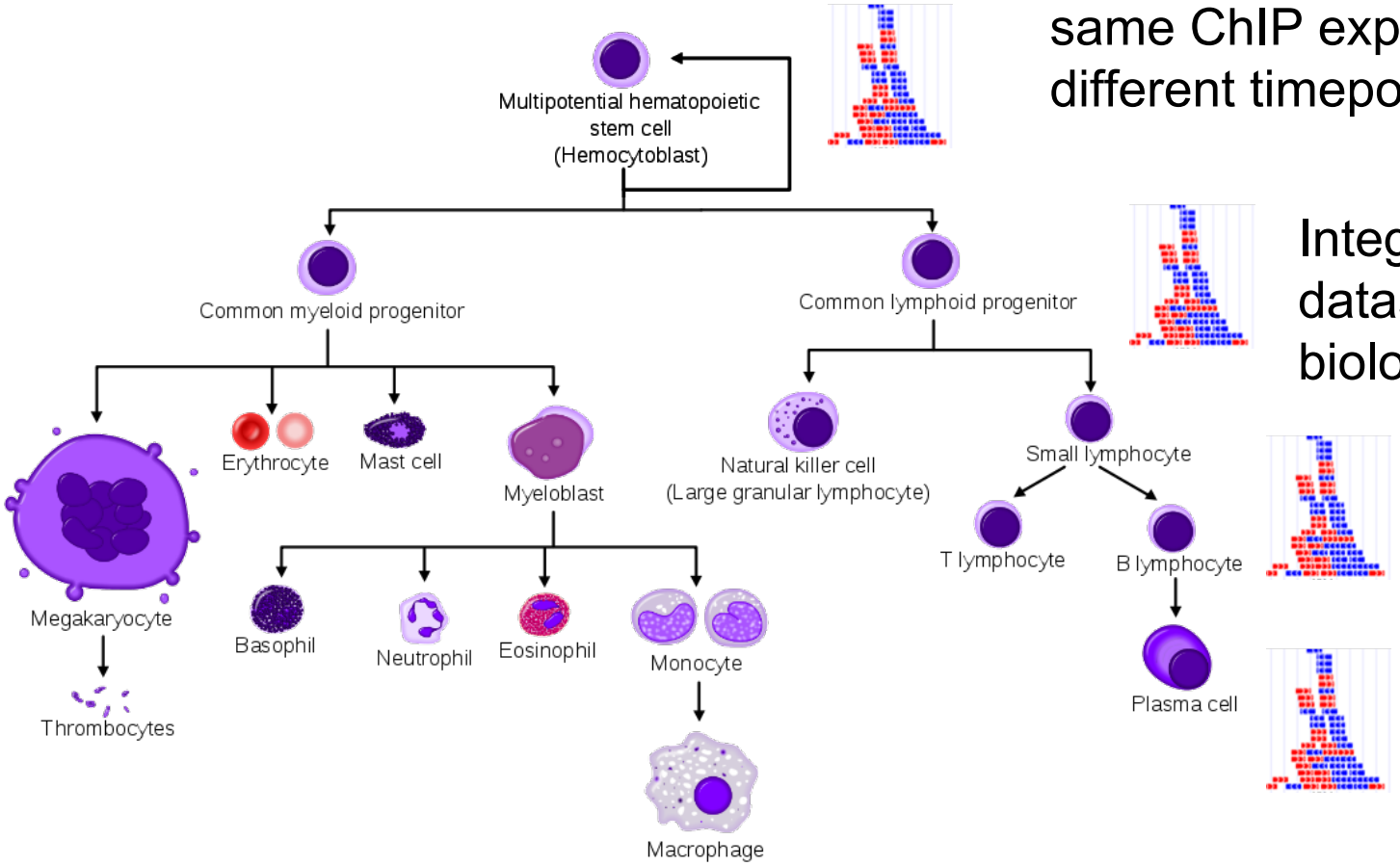
- Realigns and calls peaks:
  - 12 million alignments
    - < 20 minutes
    - < 1.6 GB RAM
  - 120 million alignments
    - < 30 minutes
    - < 6 GB RAM
- AREM is a python package
- Download from github.com/uci-cbcl/arem

AREM can be applied in other contexts

- Repeat problem plagues all of HTS analysis
- AREM framework can be applied to other analysis methods
  - RNA-seq: re-align ambiguous reads to the most abundant transcripts
  - SNP/variant calling: re-align ambiguous reads to the genotypes that the reads agree with
  - Many other possibilities


Machine Learning
Biology
AREM
TreeHMM
Unsupervised clustering of multiple genomes for improved biological insight
Algorithms

Scaling up: multiple ChIP datasets from multiple cell types



Determine binding site **dynamics** by performing the same ChIP experiment at different timepoints/cell stages

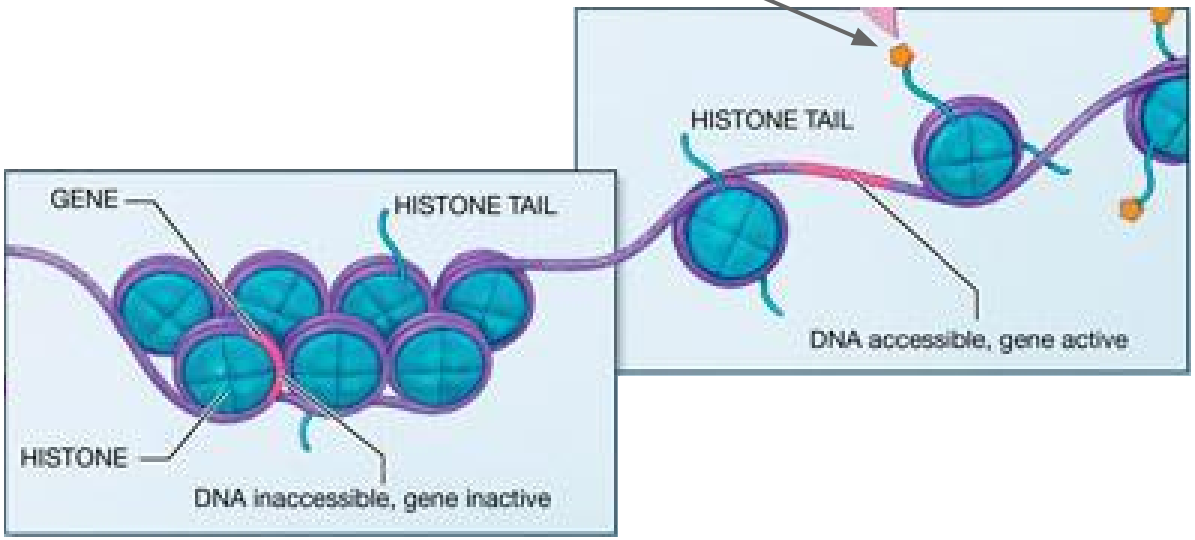
Integrate multiple datasets for biological insight

Scaling up: multiple ChIP datasets from multiple species

Nine ChIP-seq experiments

- CTCF
- H3k27me3
- H3k36me3
- H4k20me1
- H3k4me1
- H3k4me2
- H3k4me3
- H3k27ac
- H3k9ac

Histone modifications (**not** transcription factors)



wikipedia.org/wiki/Epigenetics

Scaling up: multiple ChIP datasets from multiple species

Nine ChIP-seq experiments

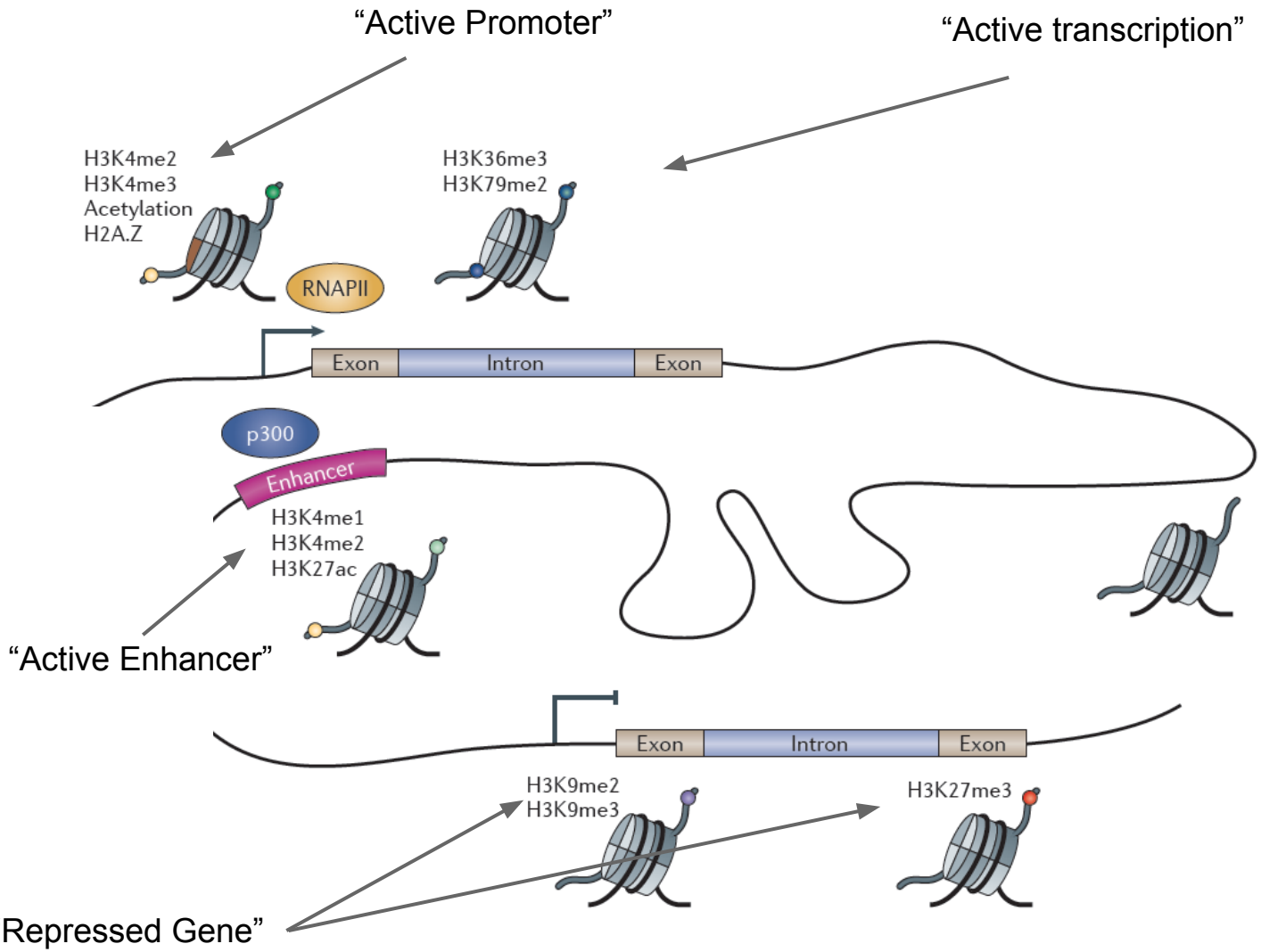
- CTCF
- H3k27me3
- H3k36me3
- H4k20me1
- H3k4me1
- H3k4me2
- H3k4me3
- H3k27ac
- H3k9ac

Nine human cell types

- embryonic stem cell (H1 ES)
- erythrocytic leukaemia cells (K562)
- B-lymphoblastoid cells(GM12878)
- hepatocellular carcinoma cells (HepG2)
- umbilical vein endothelial cells (HUVEC)
- skeletal muscle myoblasts (HSMM)
- normal lung fibroblasts (NHLF)
- normal epidermal keratinocytes (NHEK)
- mammary epithelial cells (HMEC)

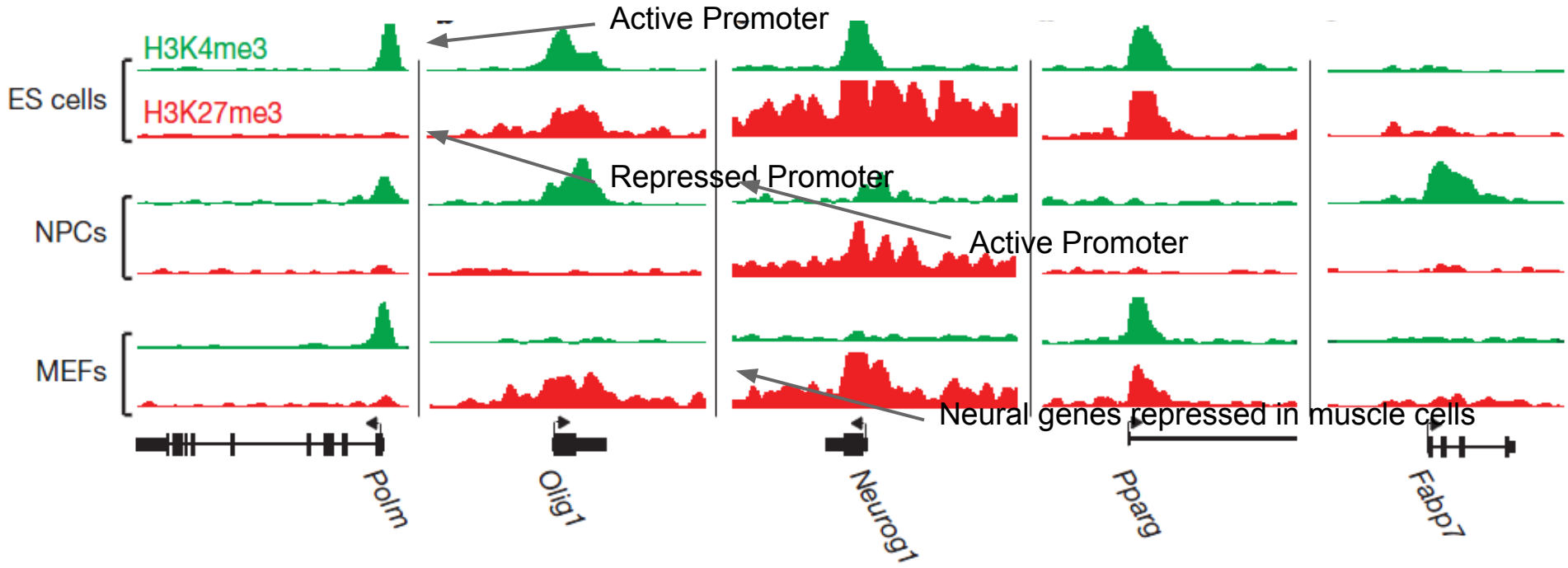
Ernst et al, Nature, 2011

Histone mark combinations indicate gene function



Zhou et al, Nature Rev. Gen., 2011

Binding dynamics across cell types



Olig1: Neural transcription factor

Polm: DNA polymerase (gene needed in all cell types)

- ES cells: Embryonic stem cells
- NPCs: Neural progenitor cells
- MEFs: Embryonic fibroblasts (muscle)

- *Neurog1*: Neurogenesis transcription factor
- *Pparg*: Adipogenesis transcription factor
- *Fabp7*: Neural progenitor marker

Mikkelsen et al., Nature 2007

Scaling up ... ets fr...

Can we automatically learn biology's histone code?

Nine ChIP-seq experiments

- CTCF
- H3k27me3
- H3k36me3
- H4k20me1
- H3k4me1
- H3...
- ...

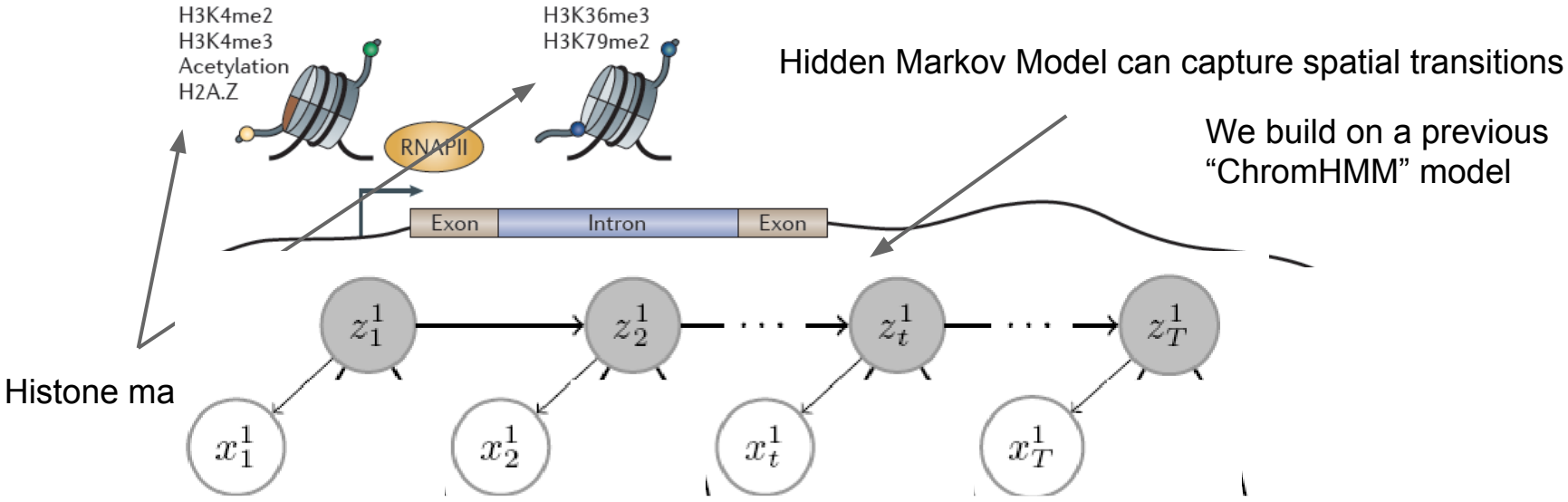
- ...n (H1 ES)
- erythrocytic leukaemia cells (K562)
- B-lymphoblastoid cells(GM12878)
- hepatocellular carcinoma cells (HepG2)
- ...bilical ...in endothelial cells (HUVEC)
- ...le myoblasts (HSMM)
- ...roblasts (NHLF)
- ...l keratinocytes (NHEK)
- ...thelial cells (HMEC)

What does the data tell us about cell differentiation?

Ernst et al, Nature, 2011

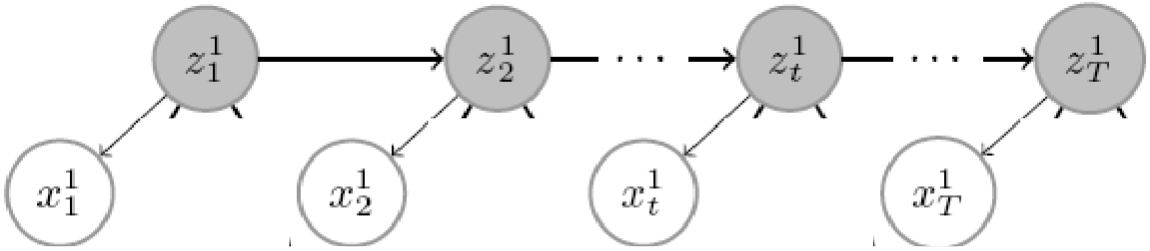
Unsupervised Learning

- Family of machine learning methods to recognize patterns in datasets
  - Includes K-means, hierarchical clustering, self-organizing maps, and many other methods

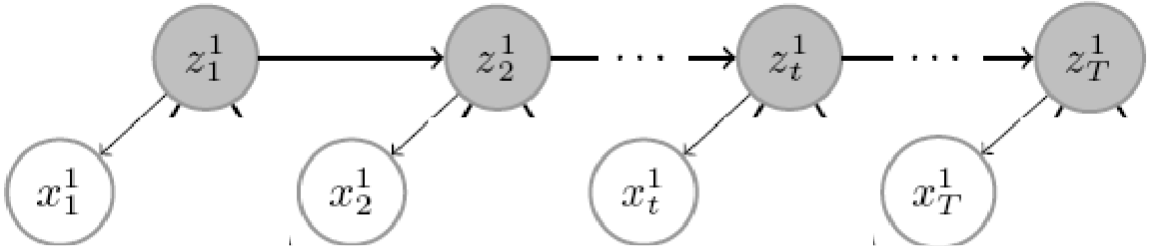


What about multiple cell types?

ES cells



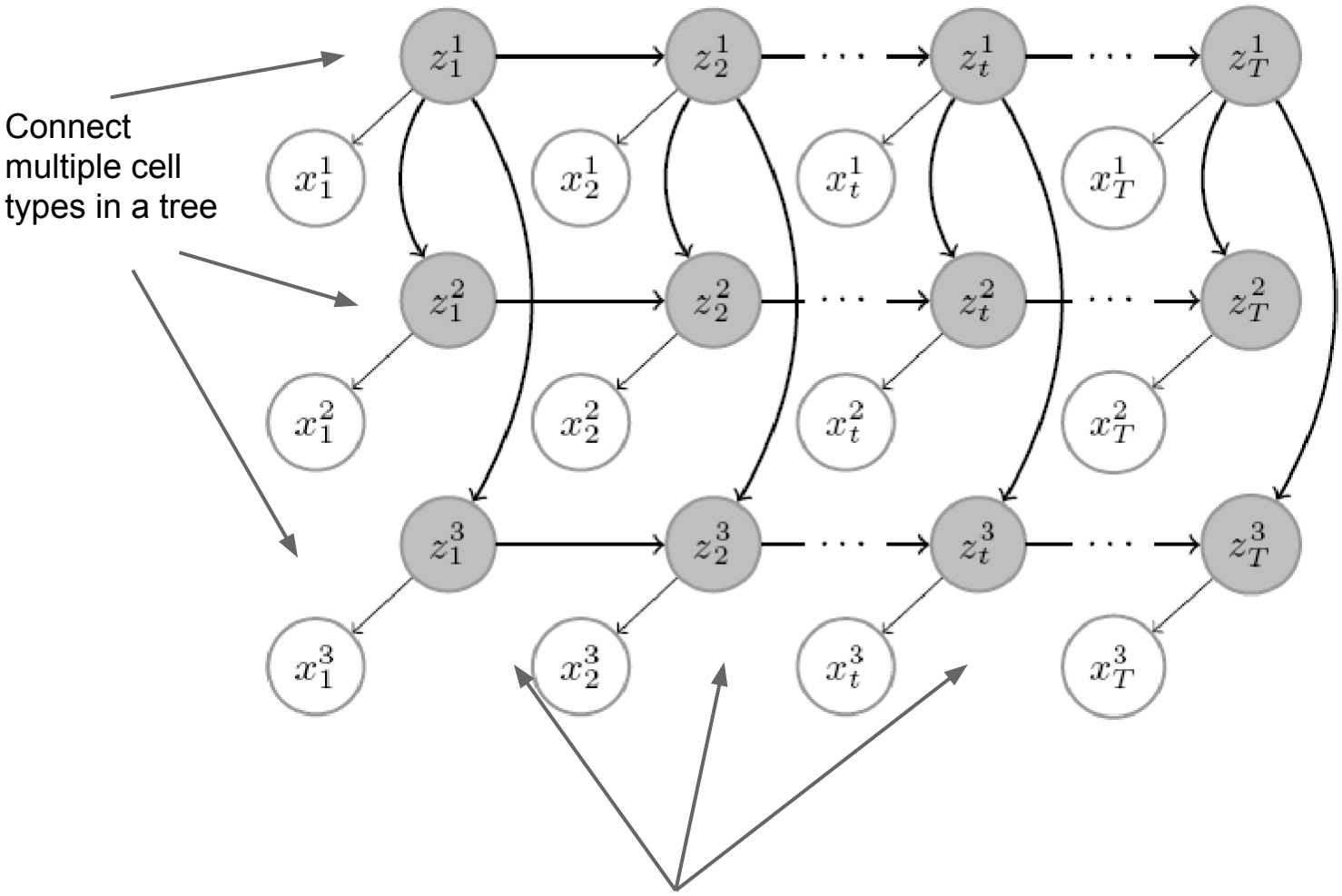
NPCs



x_i^j: **observed** histone marks (position i, species j)

z_i^j: **hidden** chromatin state (to be inferred)

A new “TreeHMM” for lineages



TreeHMM Recap

- Data: M x N x L matrix of binary histone mark presence
  - M species, related to each other by a tree
  - N contiguous genomic regions
  - L different histone marks
- Use a Tree Hidden Markov Model to do unsupervised learning
- We are given K, the number different histone states to find
  - K x L “emission” matrix
  - K x K “transition” matrix for root species
  - K x K x K “transition” matrix for other species

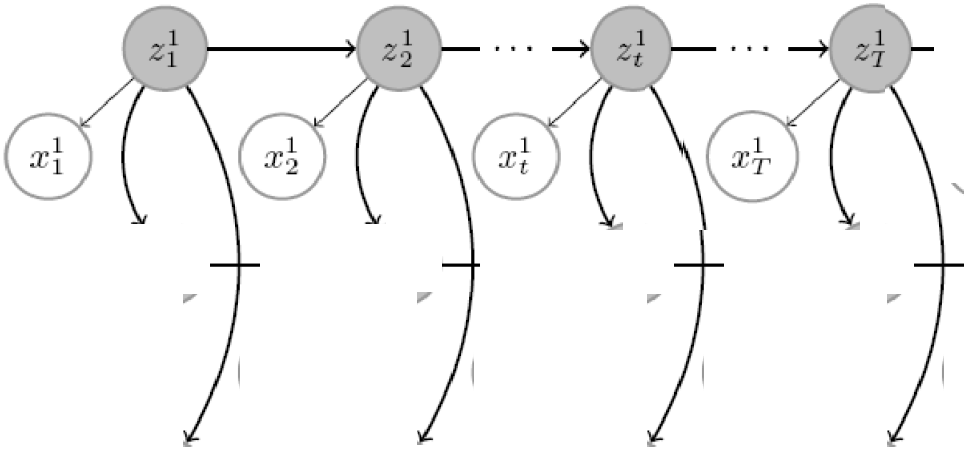
Inference in TreeHMM

- Just one problem: Inferring the hidden state in this model is intractable when K or M is large (K^M state space)
- We use **variational methods** to approximate the model
  - Choose a tractable family of surrogate models, then optimize them to look like the more complicated model

Mean field: Optimize single nodes separately

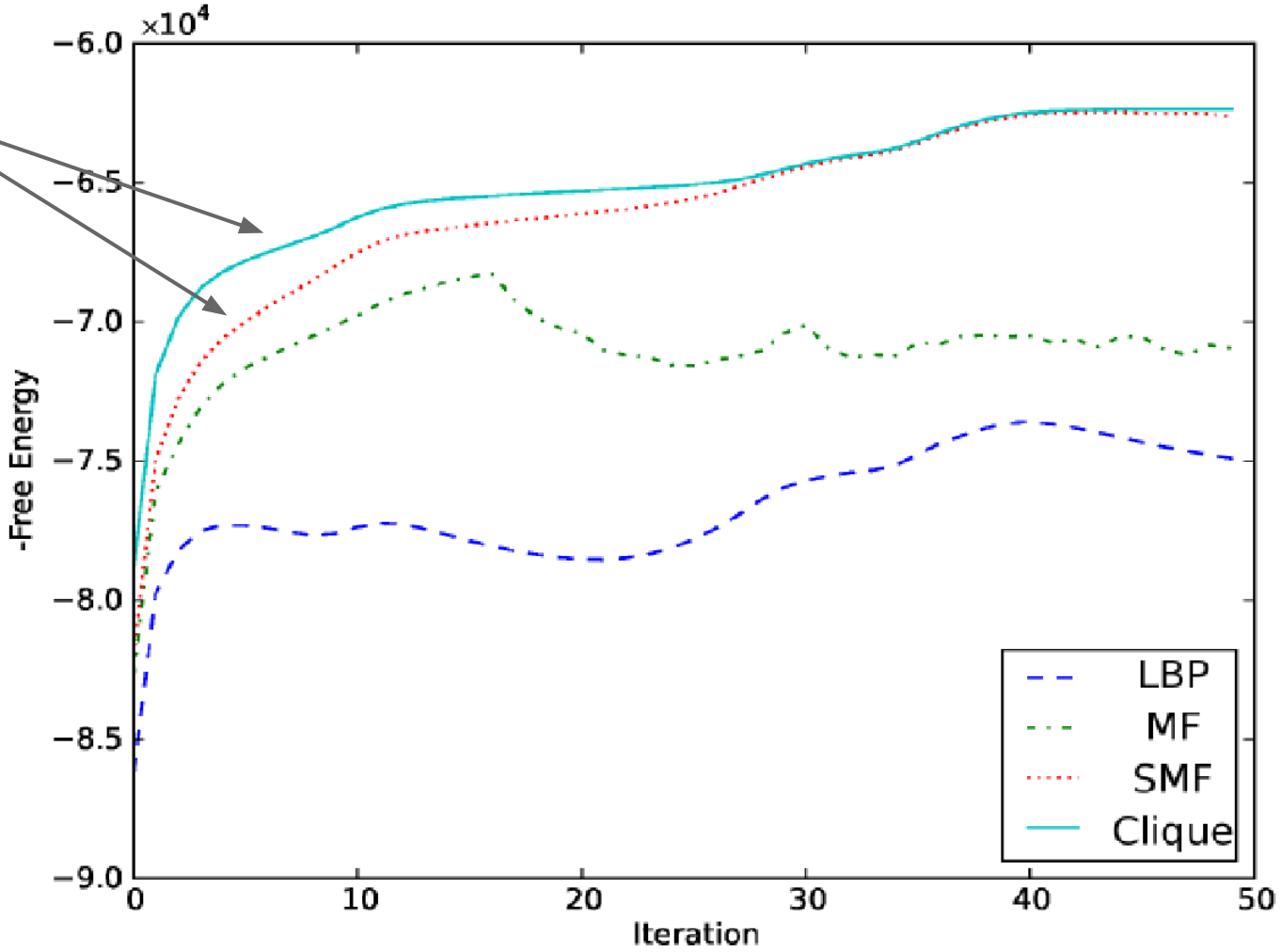
Structured mean field:
Optimize complete HMM chains separately

Loopy belief propagation

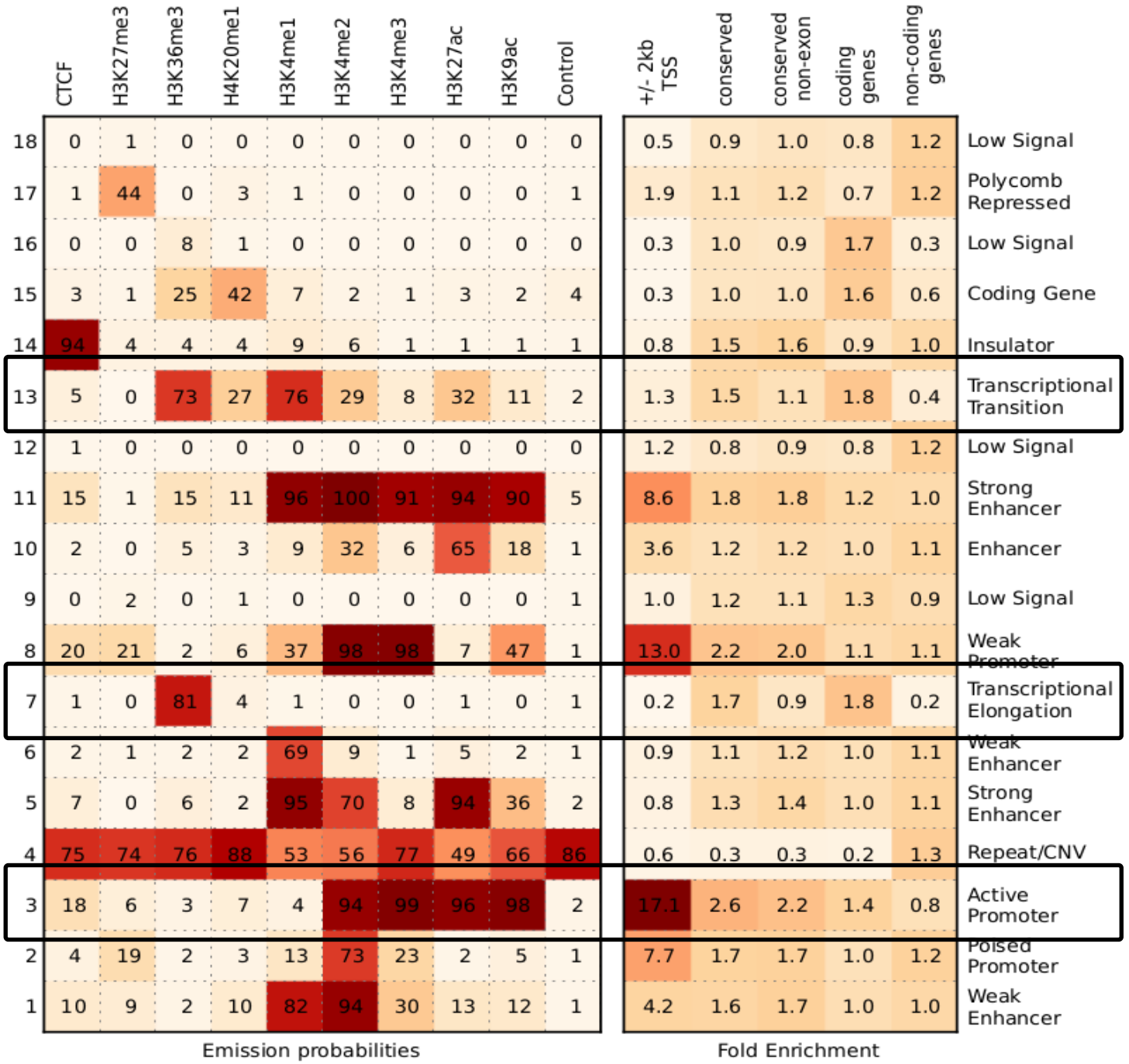


How good are the approximations?

Structured mean field approximates the exact model very well

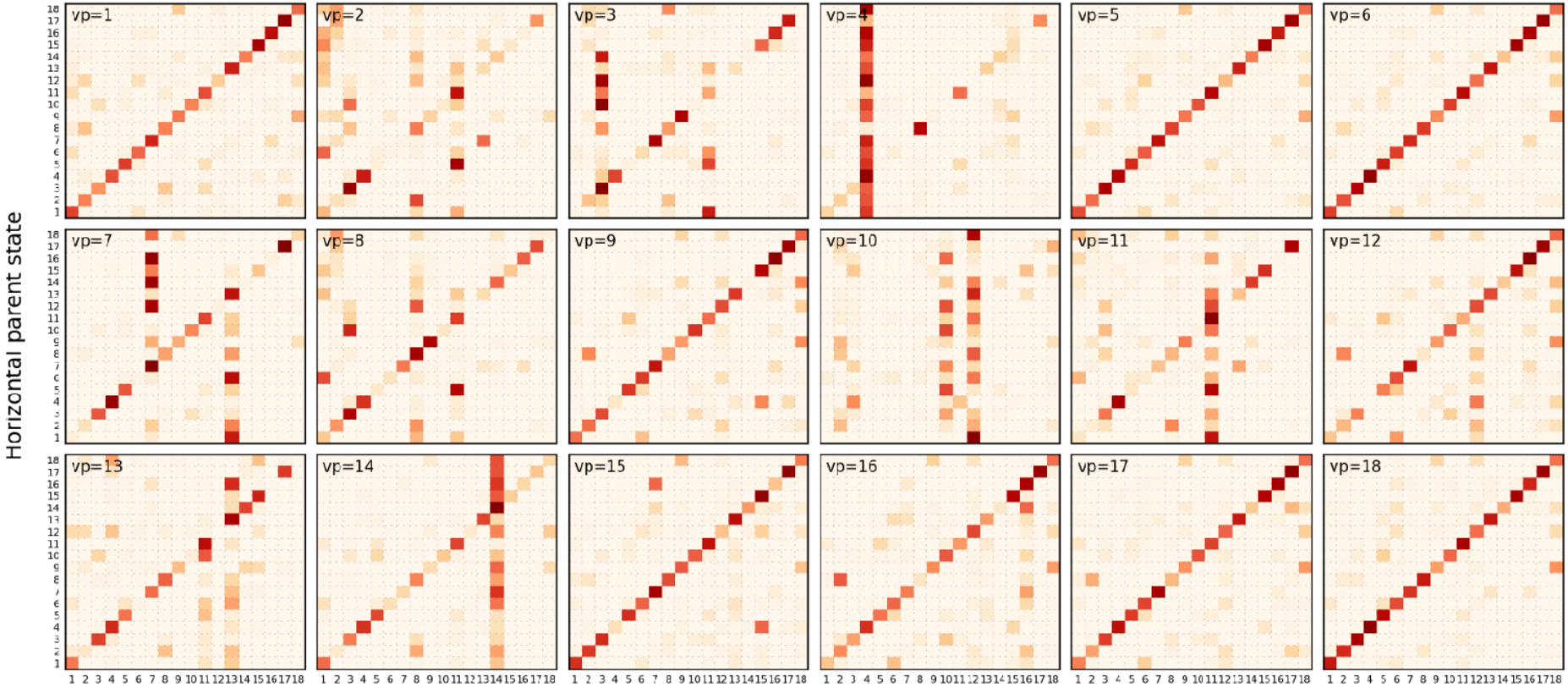


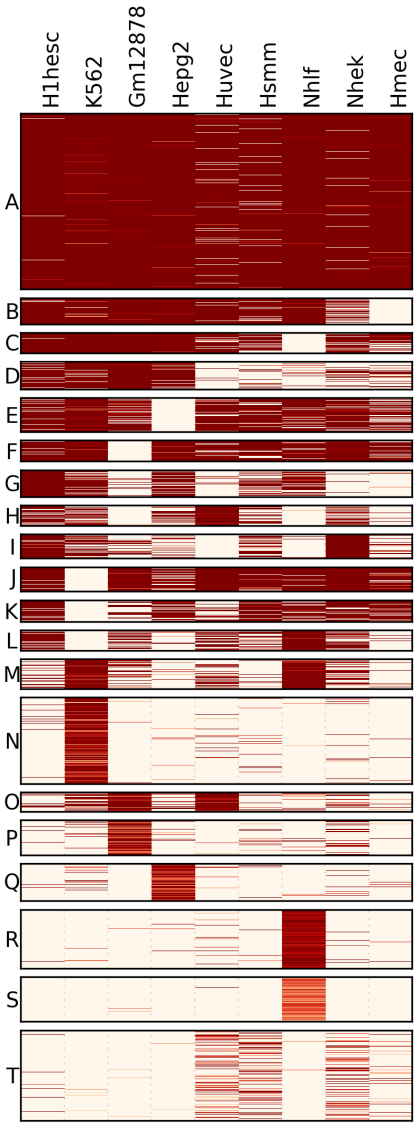
K=5, very small dataset


CTCF
H3K27me3
H3K36me3
H4K20me1
H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K9ac
Control
+/- 2kb TSS
conserved
conserved non-exon
coding genes
non-coding genes
18 0 1 0 0 0 0 0 0 0 0 0.5 0.9 1.0 0.8 1.2 Low Signal
17 1 44 0 3 1 0 0 0 0 1 1.9 1.1 1.2 0.7 1.2 Polycomb Repressed
16 0 0 8 1 0 0 0 0 0 0 0.3 1.0 0.9 1.7 0.3 Low Signal
15 3 1 25 42 7 2 1 3 2 4 0.3 1.0 1.0 1.6 0.6 Coding Gene
14 94 4 4 4 9 6 1 1 1 1 0.8 1.5 1.6 0.9 1.0 Insulator
13 5 0 73 27 76 29 8 32 11 2 1.3 1.5 1.1 1.8 0.4 Transcriptional Transition
12 1 0 0 0 0 0 0 0 0 0 1.2 0.8 0.9 0.8 1.2 Low Signal
11 15 1 15 11 96 100 91 94 90 5 8.6 1.8 1.8 1.2 1.0 Strong Enhancer
10 2 0 5 3 9 32 6 65 18 1 3.6 1.2 1.2 1.0 1.1 Enhancer
9 0 2 0 1 0 0 0 0 0 1 1.0 1.2 1.1 1.3 0.9 Low Signal
8 20 21 2 6 37 98 98 7 47 1 13.0 2.2 2.0 1.1 1.1 Weak Promoter
7 1 0 81 4 1 0 0 1 0 1 0.2 1.7 0.9 1.8 0.2 Transcriptional Elongation
6 2 1 2 2 69 9 1 5 2 1 0.9 1.1 1.2 1.0 1.1 Weak Enhancer
5 7 0 6 2 95 70 8 94 36 2 0.8 1.3 1.4 1.0 1.1 Strong Enhancer
4 75 74 76 88 53 56 77 49 66 86 0.6 0.3 0.3 0.2 1.3 Repeat/CNV
3 18 6 3 7 4 94 99 96 98 2 17.1 2.6 2.2 1.4 0.8 Active Promoter
2 4 19 2 3 13 73 23 2 5 1 7.7 1.7 1.7 1.0 1.2 Poised Promoter
1 10 9 2 10 82 94 30 13 12 1 4.2 1.6 1.7 1.0 1.0 Weak Enhancer
Emission probabilities
Fold Enrichment

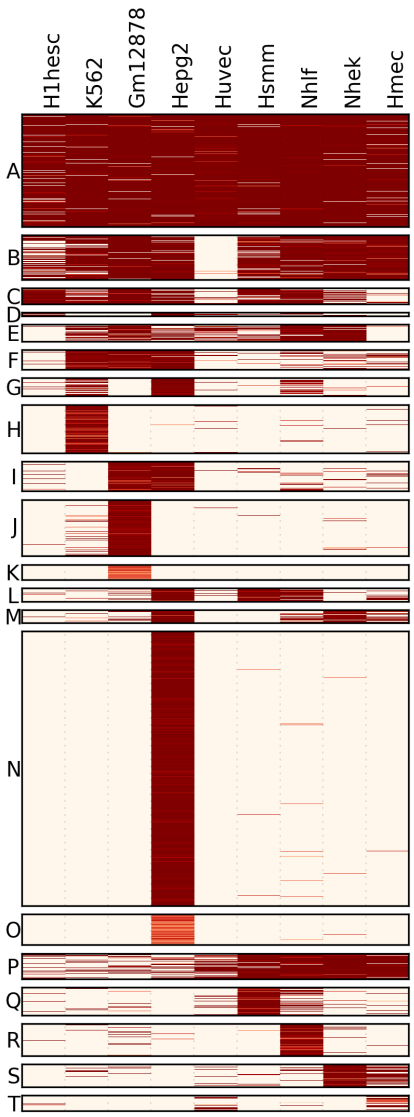
Spatial transitions between states

- Vertical parent specific transition matrix

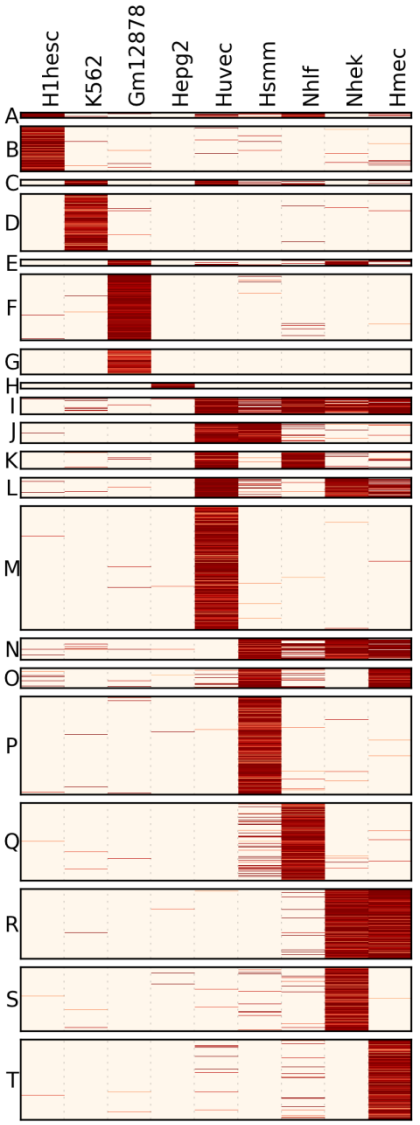




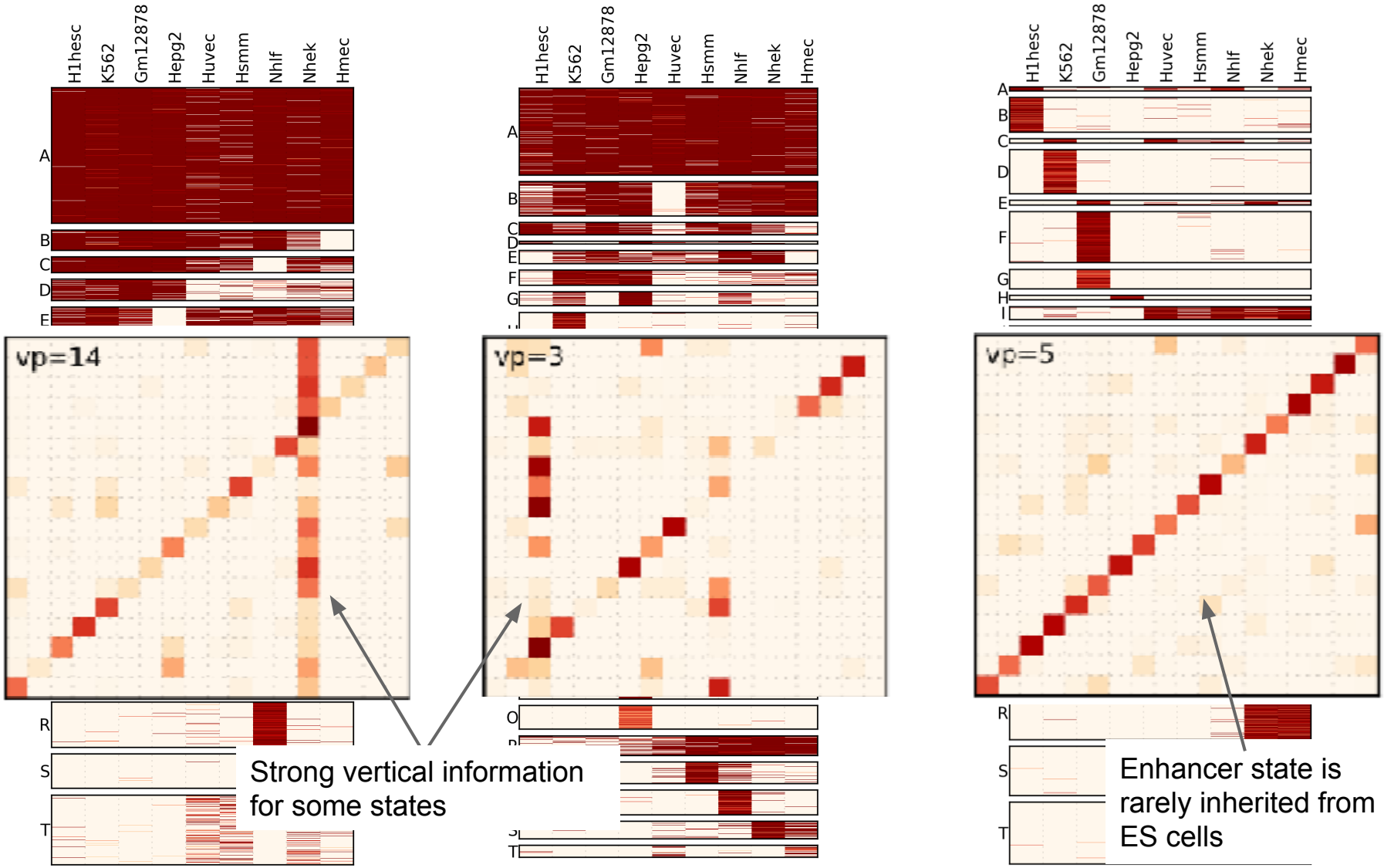
Insulator (CTCF)
(state 14)



Active Promoter
(state 3)



Strong Enhancer
(state 5)



Insulator (CTCF) (state 14)

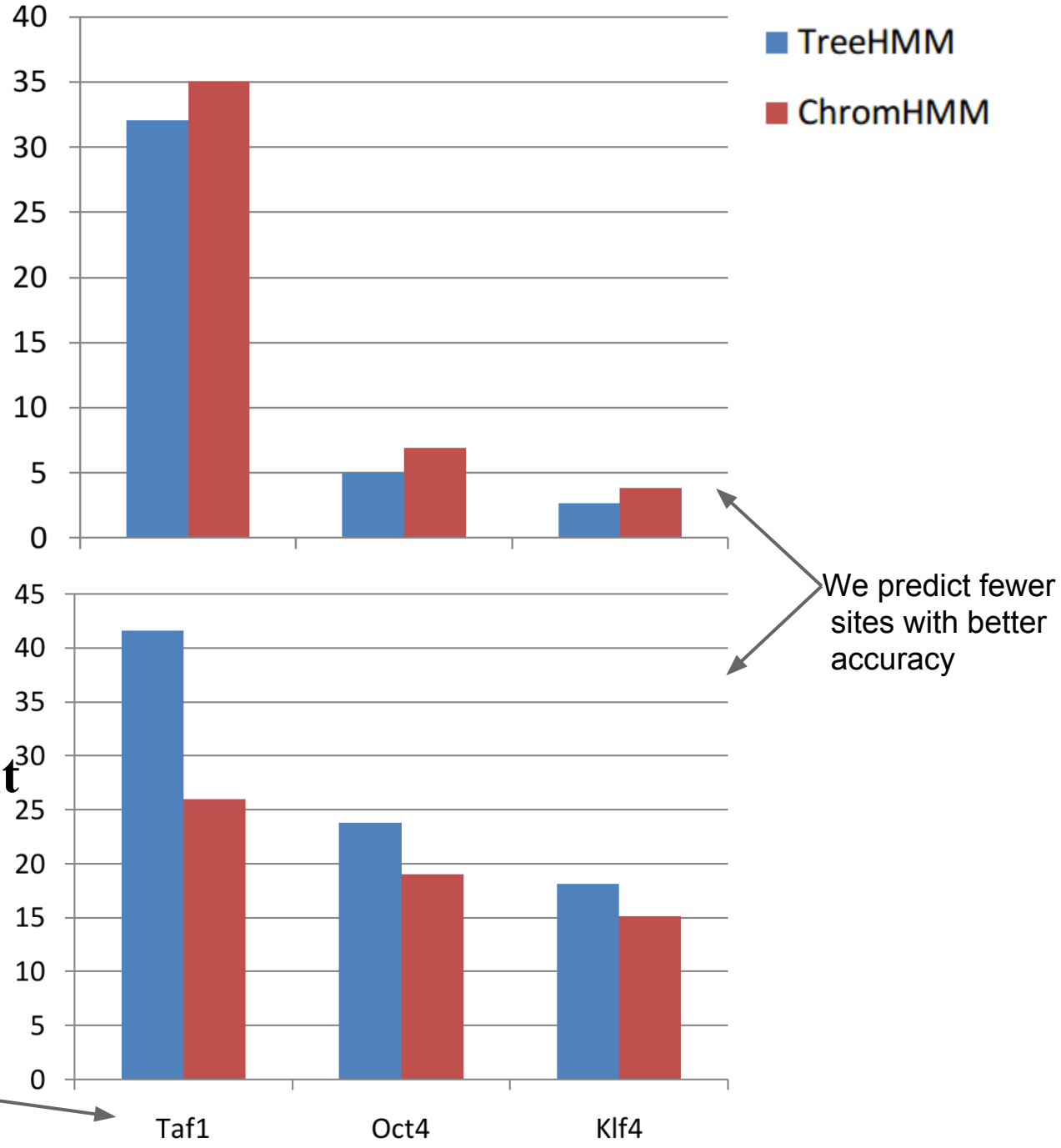
Active Promoter (state 3)

Strong Enhancer (state 5)

Validation and comparison

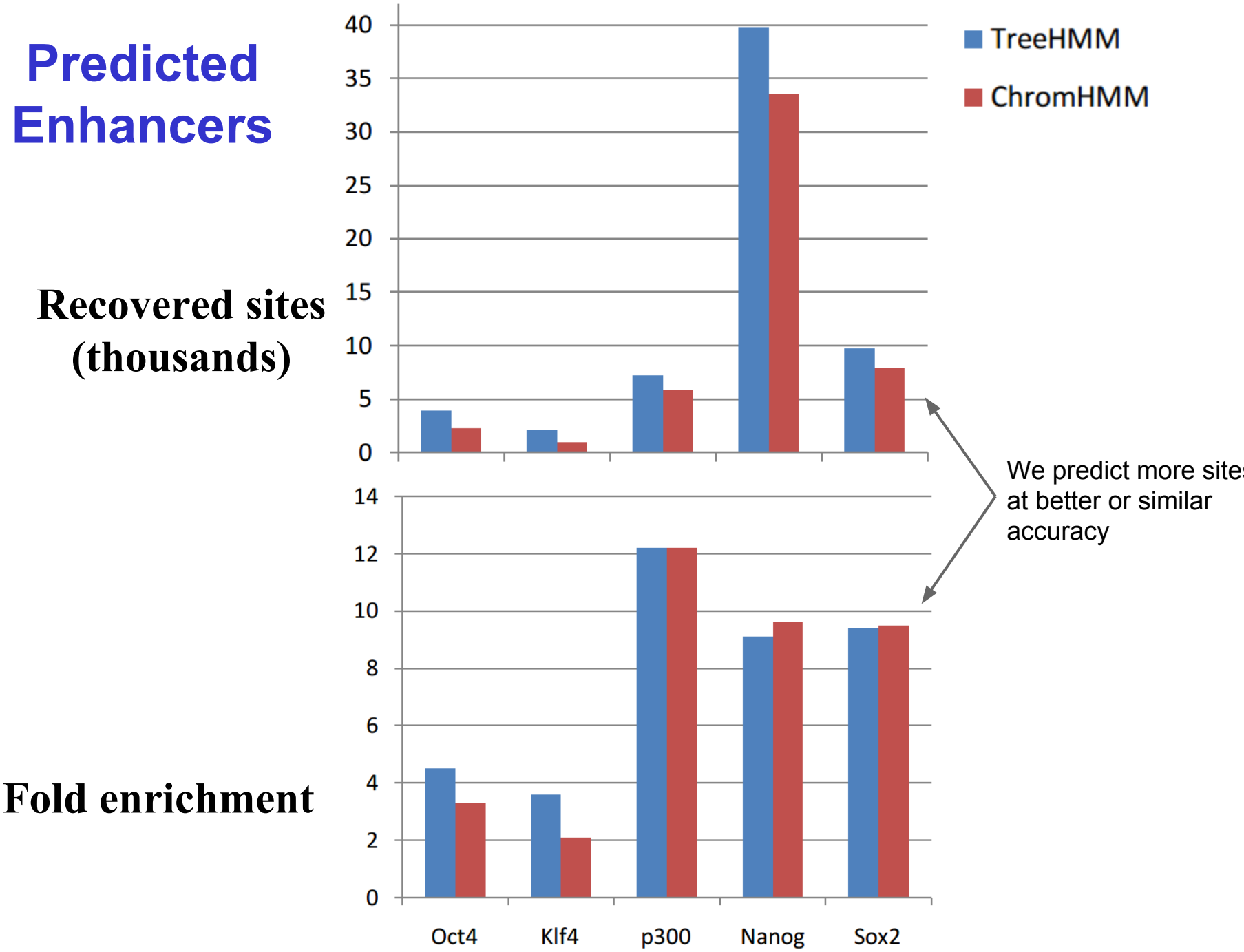
- Do our predicted states have any grounding in real biology?
- Validate them using a different dataset: transcription factor ChIP-seq
  - Record **number** of recovered TF binding sites
  - Record **fold enrichment** vs. random overlap with TF binding sites
- Compare vs. ChromHMM (like our model, but no tree component)

Predicted Promoters



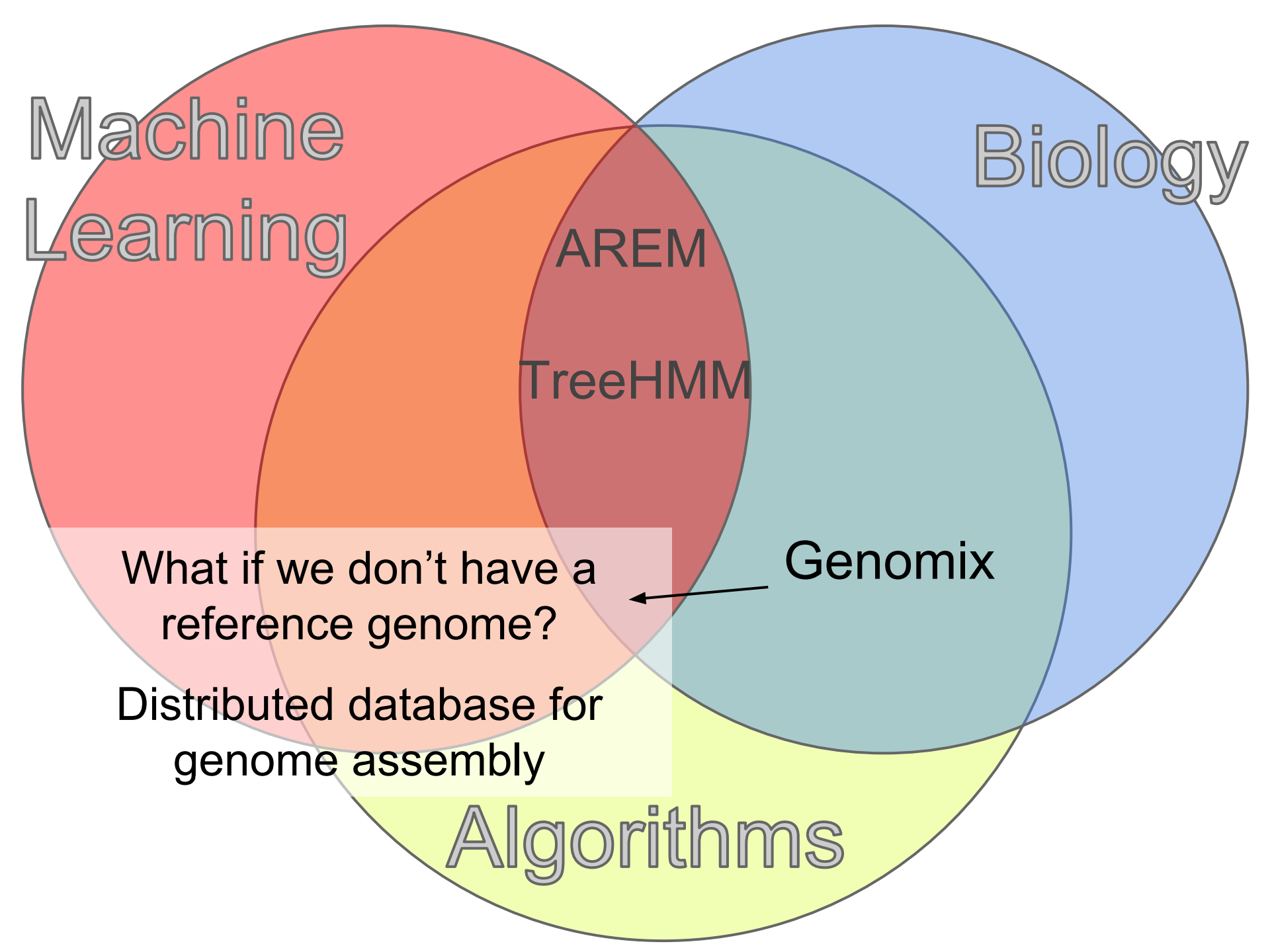
Taf1 is part of all cells' promoter machinery

Predicted Enhancers



TreeHMM take home messages

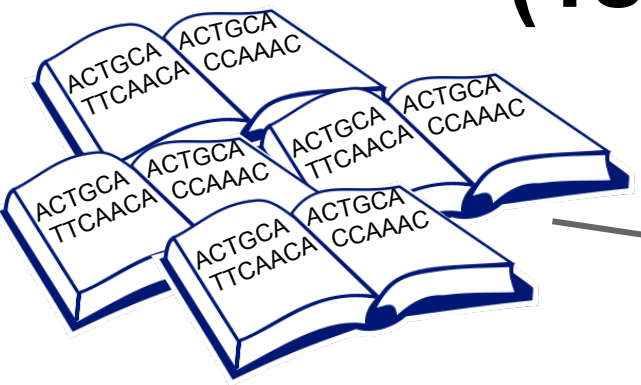
- None of this would be possible without interesting data and lots of it
  - In several organisms, there are many new datasets doing comprehensive surveys of biological function
- Extending models toward real biology is worth it
  - Can lead to improved accuracy and new insights
  - Machine Learning has tools for many more models than now employed in biology


Machine Learning
Biology
AREM
TreeHMM
What if we don't have a reference genome?
Genomix
Distributed database for genome assembly
Algorithms

No reference genome?

- For most analyses, we use the standard "reference" genome
  - Many organisms don't have a reference
  - Others have a poor quality reference
  - Some samples are too different from the reference
    - Cancer genomes are genetically unstable, subject to "genome shattering"
- Low cost of sequencing makes it feasible to do *de novo* assembly using HTS reads
  - Human genome cost ~$3 billion and 10 years, finishing in 2003
  - Done today in a few weeks for a few thousand dollars

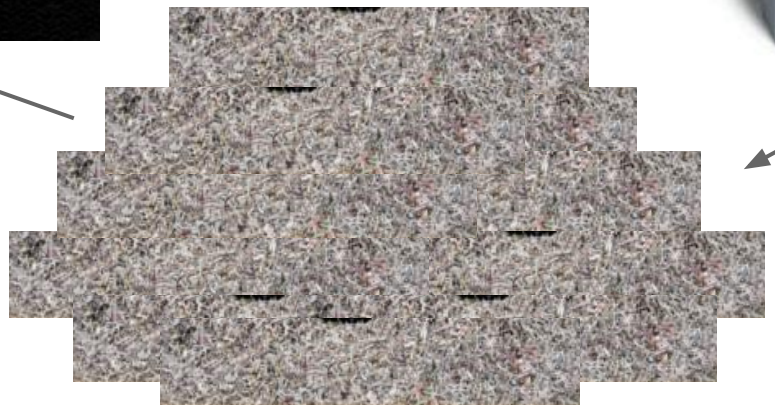
Genome Assembly (1st Approximation)



Extraction, sonication

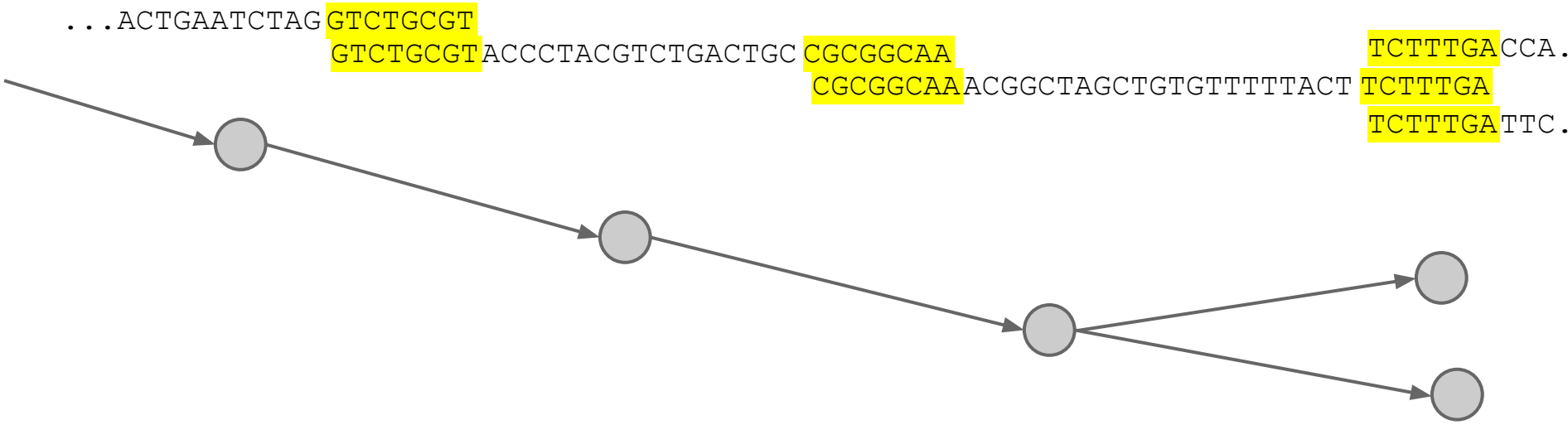
High-Throughput Sequencing

hundreds of millions of fragments



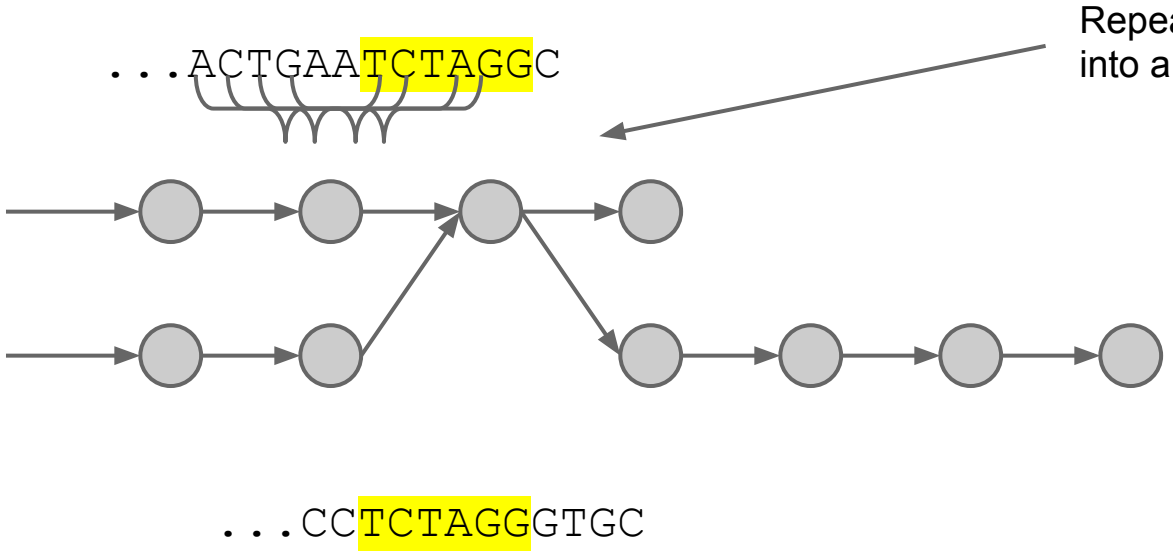
Assembly

Overlap Graphs



- Find prefix/suffix overlaps between all reads
- Form a graph where the reads are nodes and overlaps are edges
- Find a Hamiltonian path through the graph (touching all nodes once)

De Bruijn Graphs



- Form a graph where all Kmers of the reads are nodes and edges correspond to shared K-1mers
- Find an Eulerian path through the graph (touching all edges once)

Pros and Cons

Overlap Graph:

- Requires comparison between all reads
- Hamiltonian path is harder than Eulerian
- Errors detected via consensus sequences
- Can handle repeats shorter than read length
- Mostly suited for a few, long reads

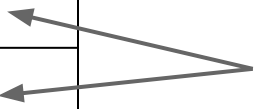
De Bruijn Graph:

- Scales with complexity of genome, not # of reads*
- Many errors show as unique graph structures
- Error identification is critical
- Without additional work, handles only repeat < K
- Well-suited for many short, low-quality reads

* except for errors...

Our goal:
scalable de Bruijn graph assembly

Algorithm	Genome Size		
	Small	Medium	Large
Velvet	✓✓	✓	✗
ABySS	✓	✓	✓
Ray	✓	✓	✓
Contrail	✗	✓	✓
Genomix	✓	✓✓	✓✓



If you have enough memory

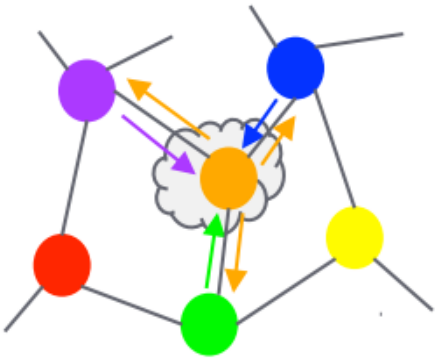
Interlude: a different revolution

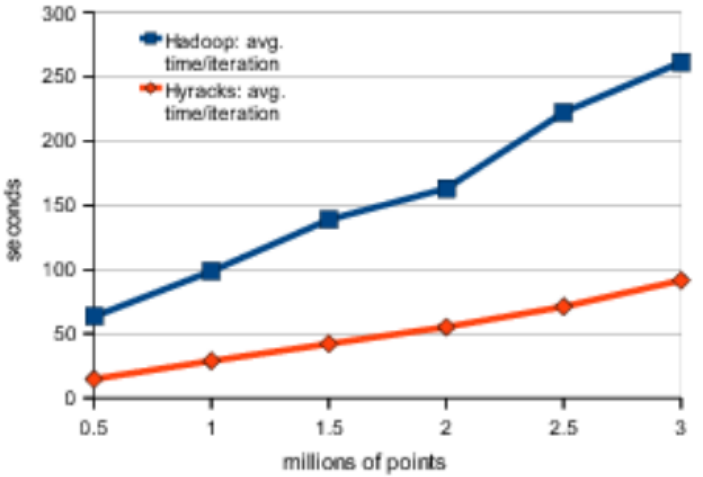
- Storage is cheap and everything is recorded
  - Social network data, browsing/shopping history, search terms, retail information
- Traditionally, all this would be kept in large databases for easy query
  - Now, the data can't fit on one machine
- Demand for **scalable** alternatives
  - Google revealed MapReduce and GFS papers **(internet scale)**
  - Hadoop (open source) soon followed

In 2013,
50% of Fortune 50

Scalable algorithms need scalable frameworks

- Hyracks: efficient and flexible alternative to Hadoop
  - Seamless use of available memory and disk space
  - Additional operators, index structures
  - Brings relational DB concepts to the cloud
- Pregelix: scalable graph algorithms
  - Hyracks-based open source Pregel implementation
  - Handles all scheduling, network, message handling, etc
  - *Think like a vertex*



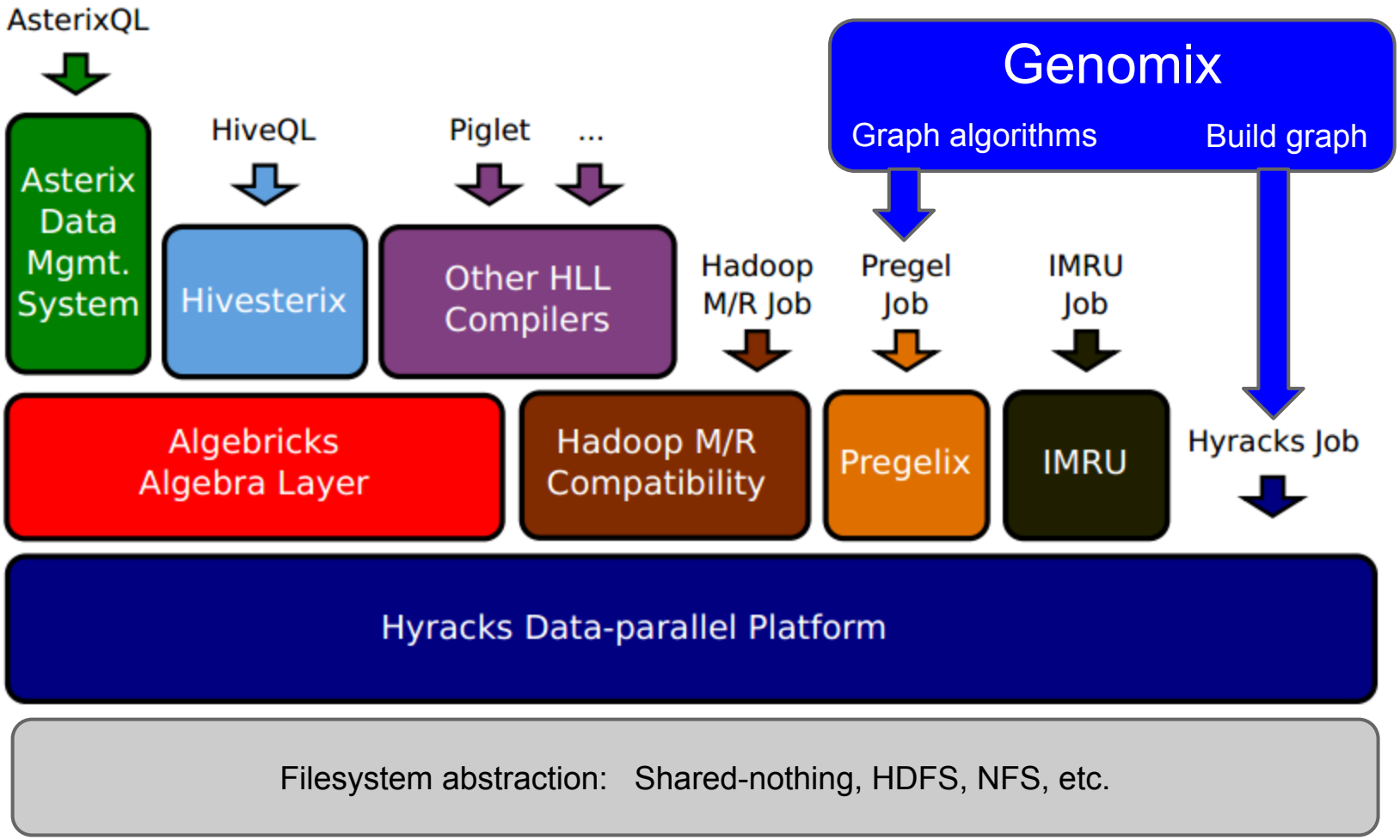


K-means clustering in the cloud

http://www.cs.washington.edu/mssi/2010/MikeCarey.pdf

Hyracks stack

Hyracks



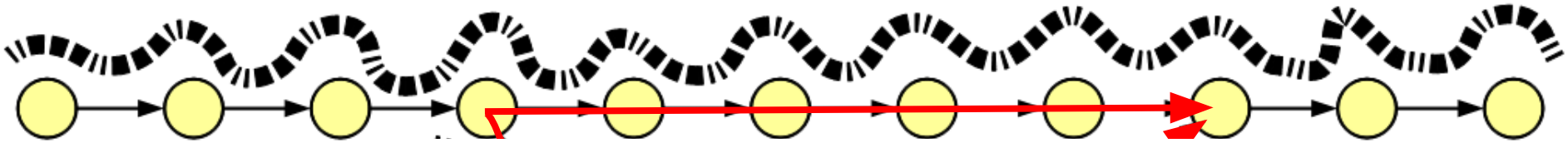
De Bruijn graph of the **small** genome of E. coli after error correction

Chaisson et al., Genome Research 2009

Sequencing Errors (Bubble Merge)

sequence

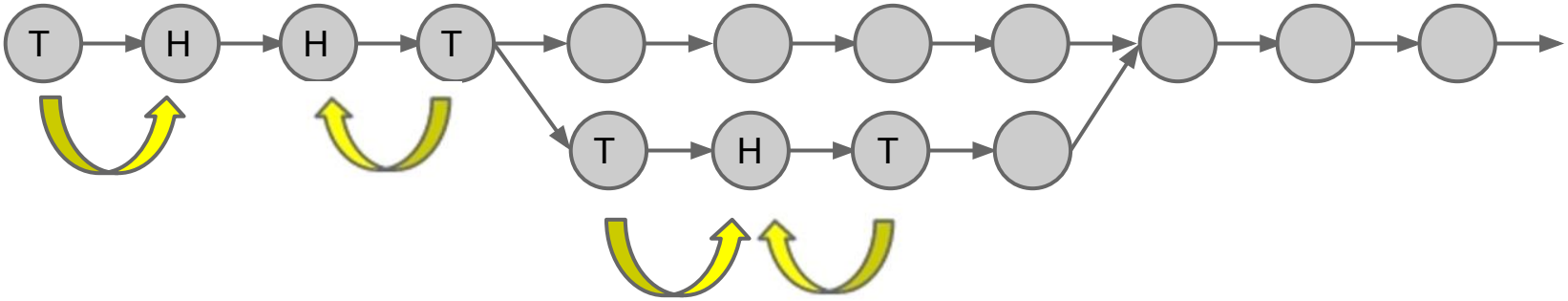
ATGGAAGTCGCGGAATC
ATGGAAGTGGCGGAATC



- Breadth-first search identifies common paths
- Extract and compare path sequences
- When paths are similar, prune “worst” one

Graph Compression

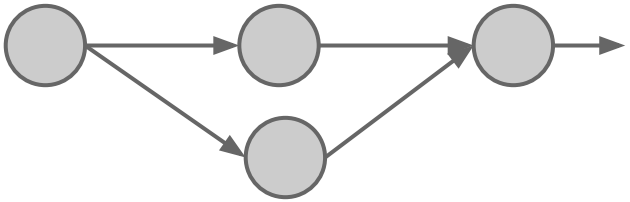
- We can collapse long chains of nodes into single nodes representing long chains
  - All later operations will take fewer iterations



- Use a randomized algorithm to coordinate nodes

Graph Compression

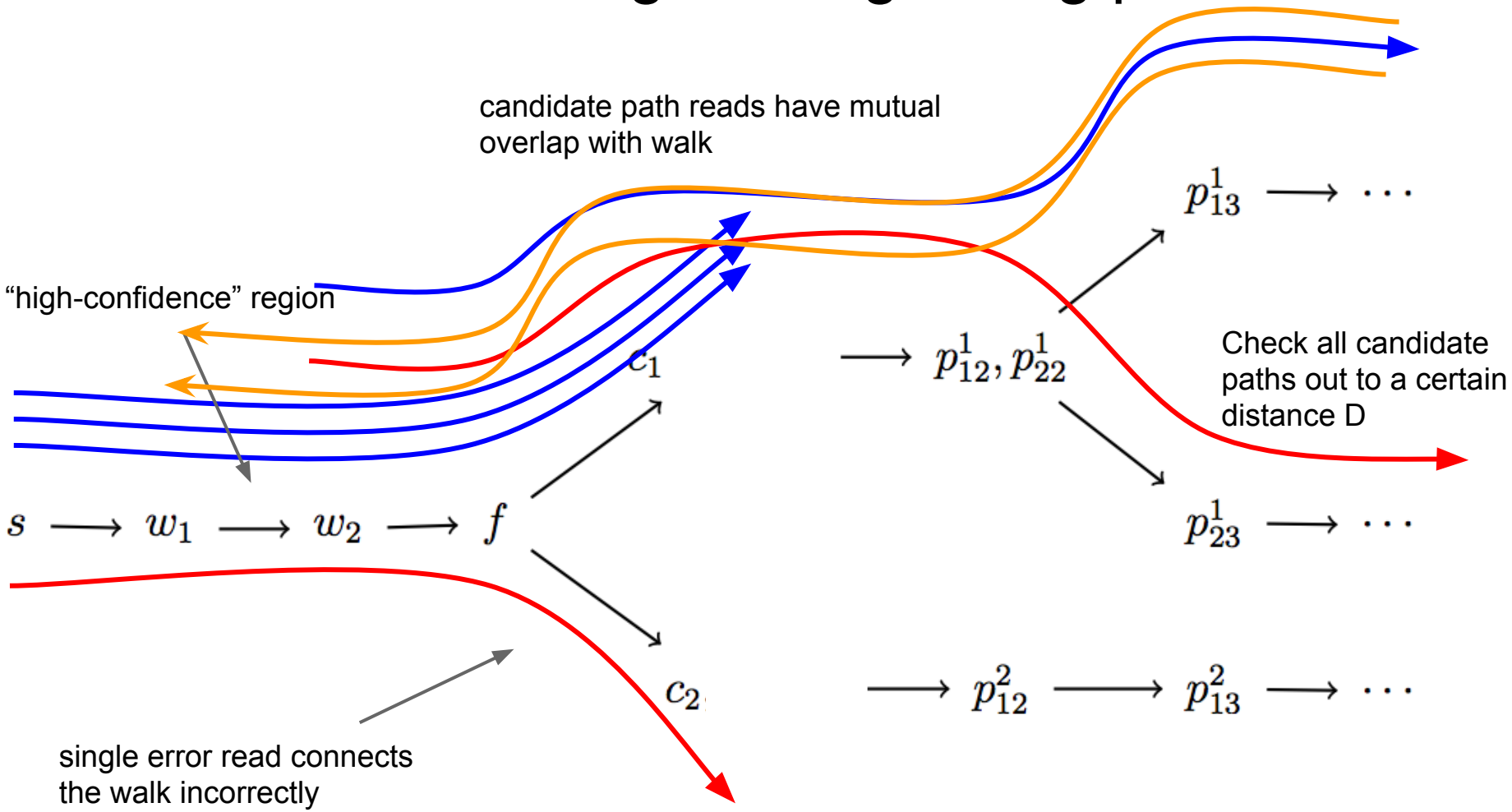
- We can collapse long chains of nodes into single nodes representing long chains
  - All later operations will take fewer iterations



- Use a randomized algorithm to coordinate nodes

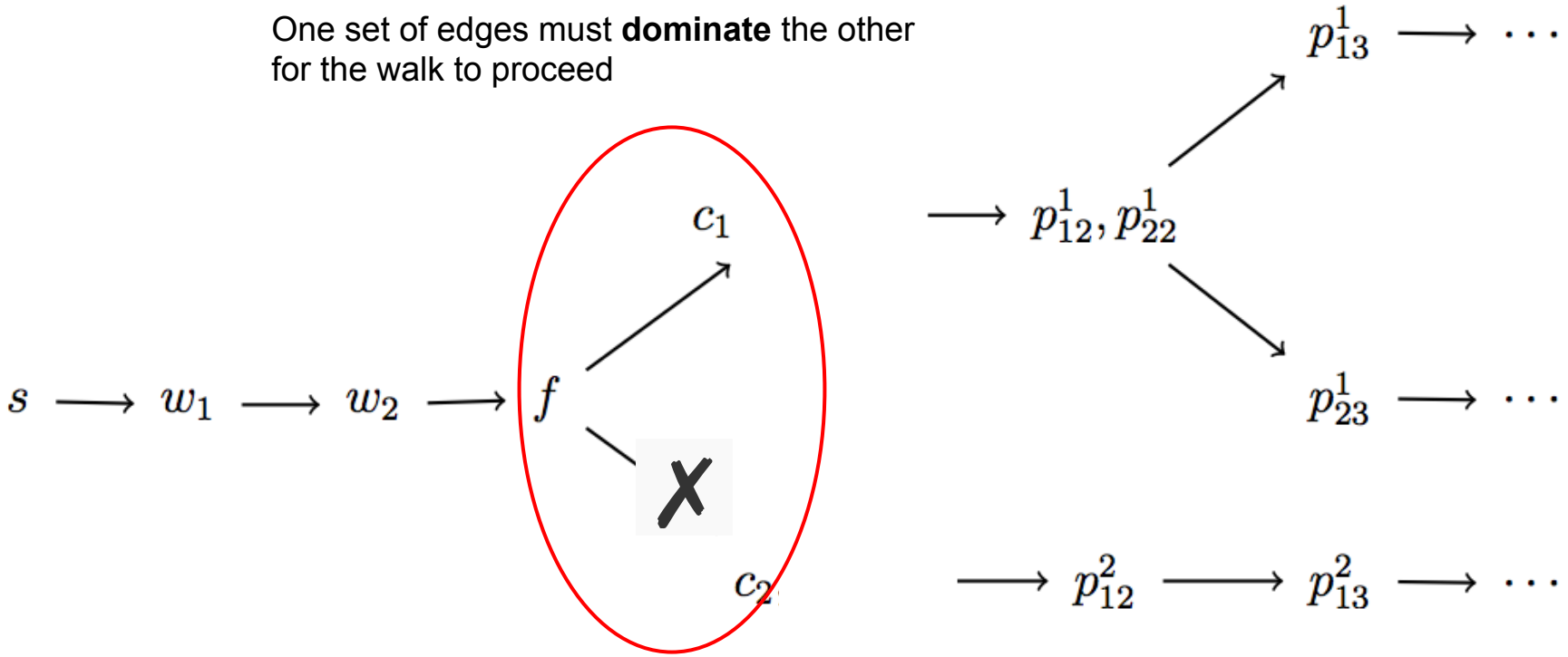
Scaffolding

- Use the reads to guide a growing path

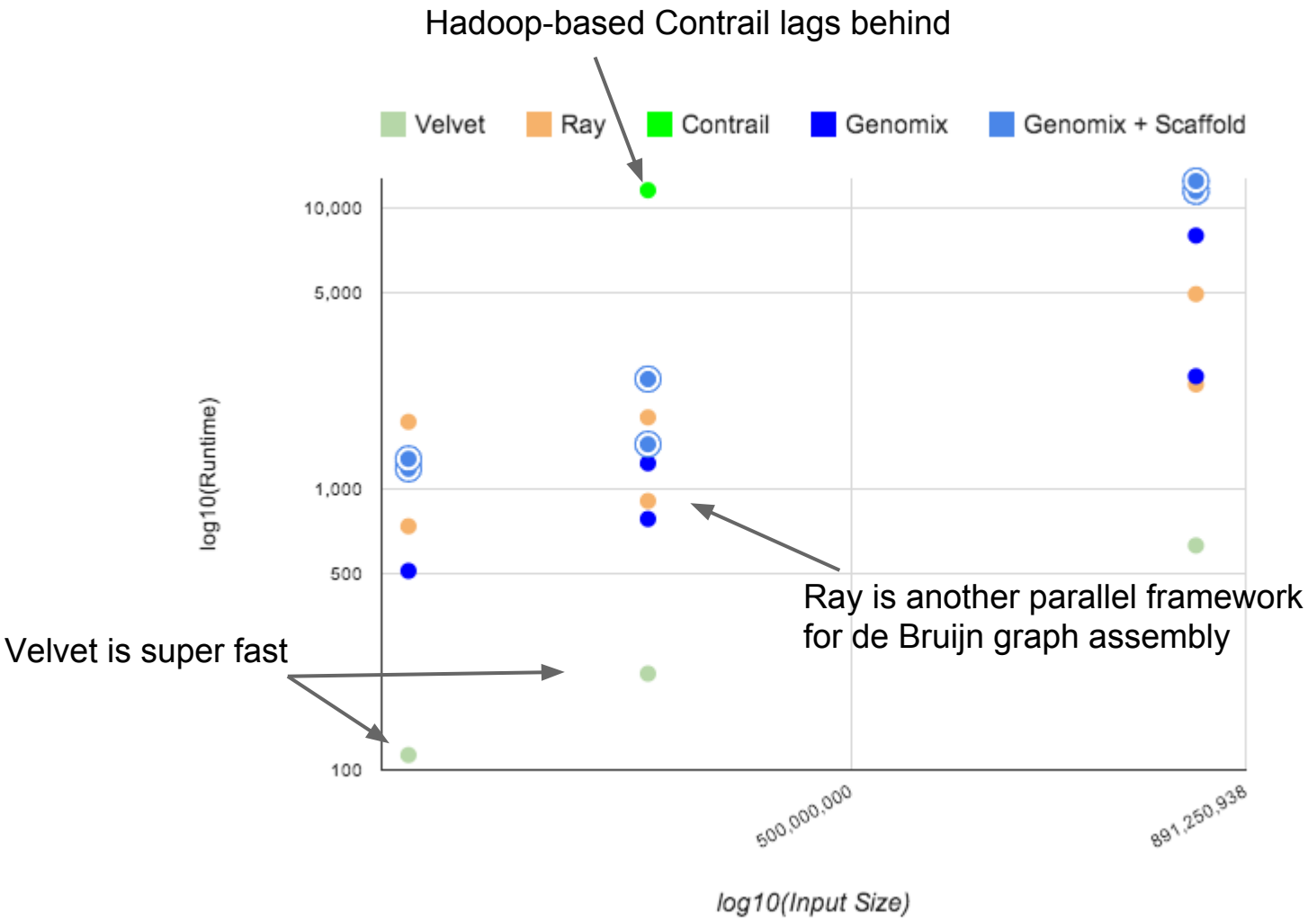


Scaffolding

- Use the reads to guide a growing path

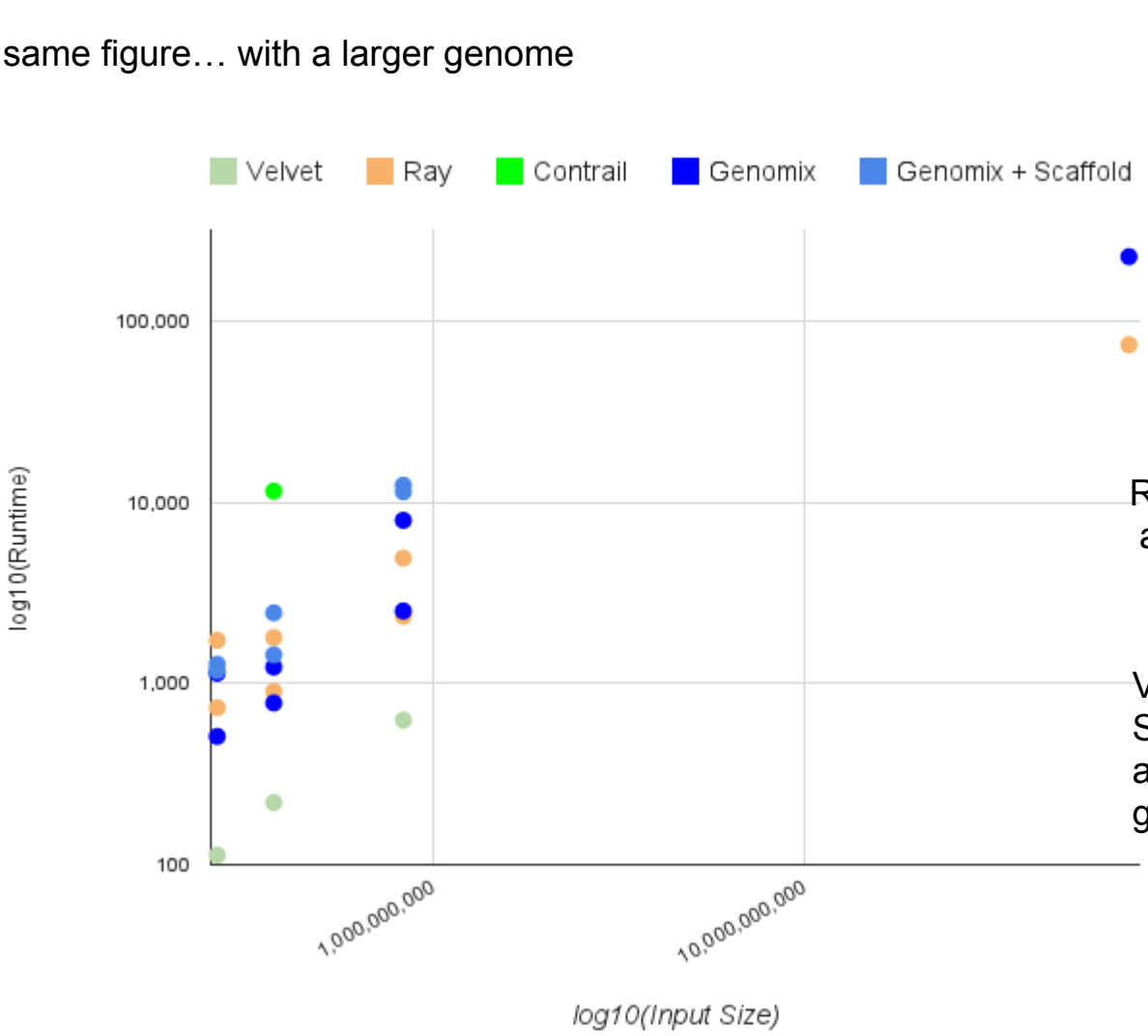


Timings: Small Genomes



With Human Genome

same figure… with a larger genome



Genomix has no memory constraints, though having it will speed up computation

Ray requires ~450GB RAM across machines

Velvet is a no-show… Some groups reported assembling a human genome with 2TB RAM...

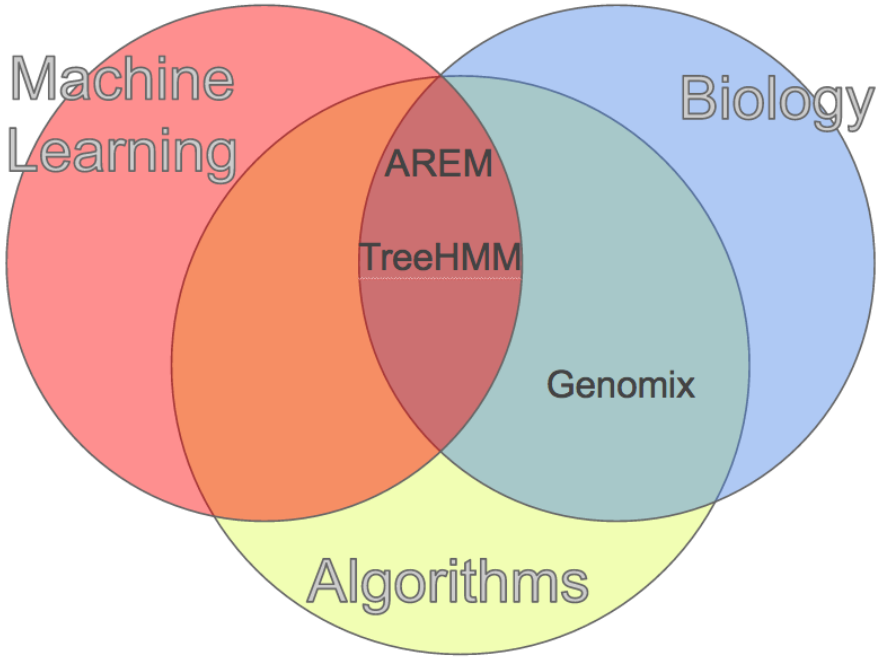
Assembly Accuracy

	N_{50}	SNPs	Indels <5bp / ≥ 5bp	Inv./Rel/Trans.	Time
Rhodo (4.6MB genome, 101bp SE reads, 180x coverage)					
Velvet	4312	899	324 / 6	1 / 2 / 2	628 (1 CPU)
Ray	3800	295	118 /4	0 / 2 / 3	2,347 (8 CPU)
Genomix	3878	573	138 / 4	2 / 3 / 9	11,402 (8 CPU)
E. coli (4.6MB genome, 36bp SE reads, 80x coverage)					
Velvet	8735	24	0 / 0	0 / 2 / 0	220 (1 CPU)
Ray	12425	37	1 / 1	0 / 1 / 0	904 (8 CPU)
Genomix	10756	9	0 / 0	8 / 6 / 0	2,452 (8 CPU)
Staph (2.9MB genome, 101bp SE reads, 90x coverage)					
Velvet	22361	100	9 / 4	0 / 2 / 0	113 (1 CPU)
Ray	3718	49	4 / 2	0 / 4 / 0	735 (8 CPU)
Genomix	7881	77	3 / 1	3 / 3 / 0	1,176 (8 CPU)

Assembly is tough, but at least you can scale up!

- Assembly was once relegated to small bacterial genomes
- Dropping costs and better tech are making assembly available to much larger genomes
  - Important for understudied organisms
  - Important for cancers and other diseases where genome structure is affected
- Don’t need huge servers w/ beefy RAM
  - Small clusters will do the job (rent them on EC2!)

Recap



- Three algorithms leveraging HTS data in different ways
  - AREM enables analysis in repetitive regions of the genome
  - TreeHMM synthesizes multiple datasets in related cell types to better annotate the genome
  - Genomix applies when the reference is inadequate or unavailable and provides a scalable solution to assembly
- HTS **requires** solid computational models and algorithms to be successful

Exciting time to be in biology!

- Costs continue to drop, quality is increasing
- New experimental methods are revealing comprehensive, in-depth biology at scales we've never seen before
- Computational methods are required to overcome errors, but also to model biological realities

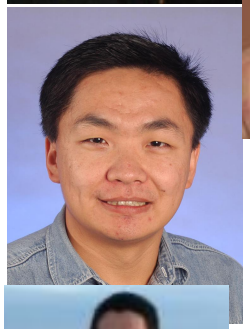
Acknowledgements

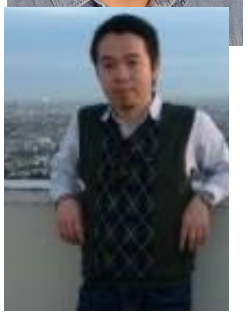
INSTITUTE FOR GENOMICS AND BIOINFORMATICS
University of California, Irvine

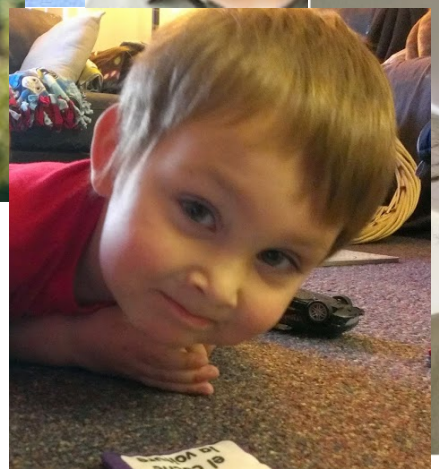
UC Irvine
Center for Complex Biological Systems

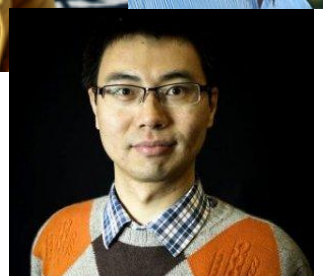
NATIONAL LIBRARY OF MEDICINE

NIH National Institutes of Health
Turning Discovery Into Health





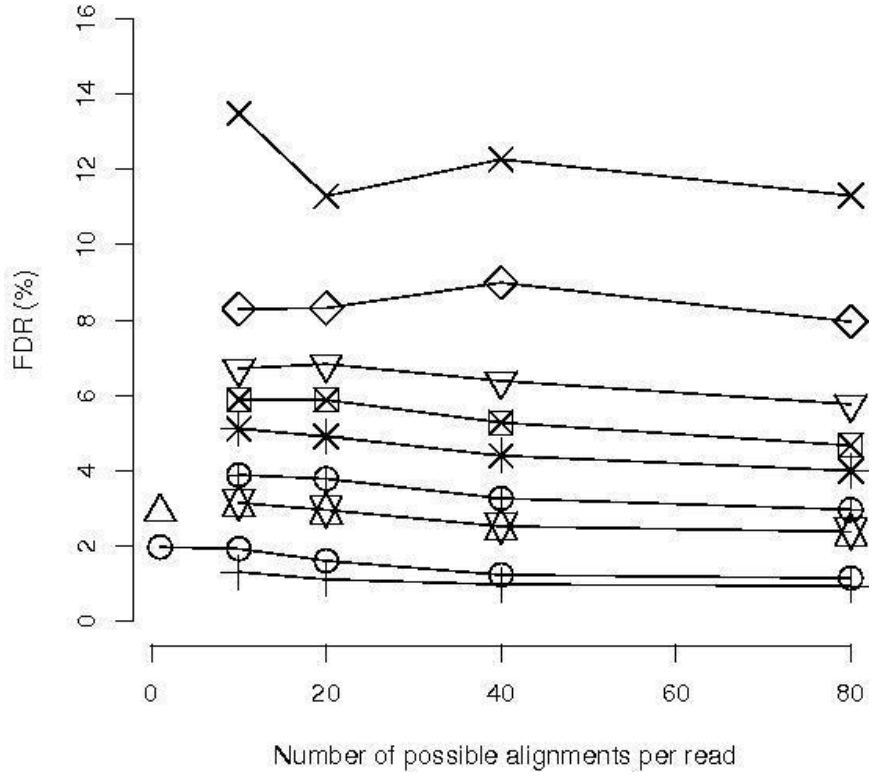


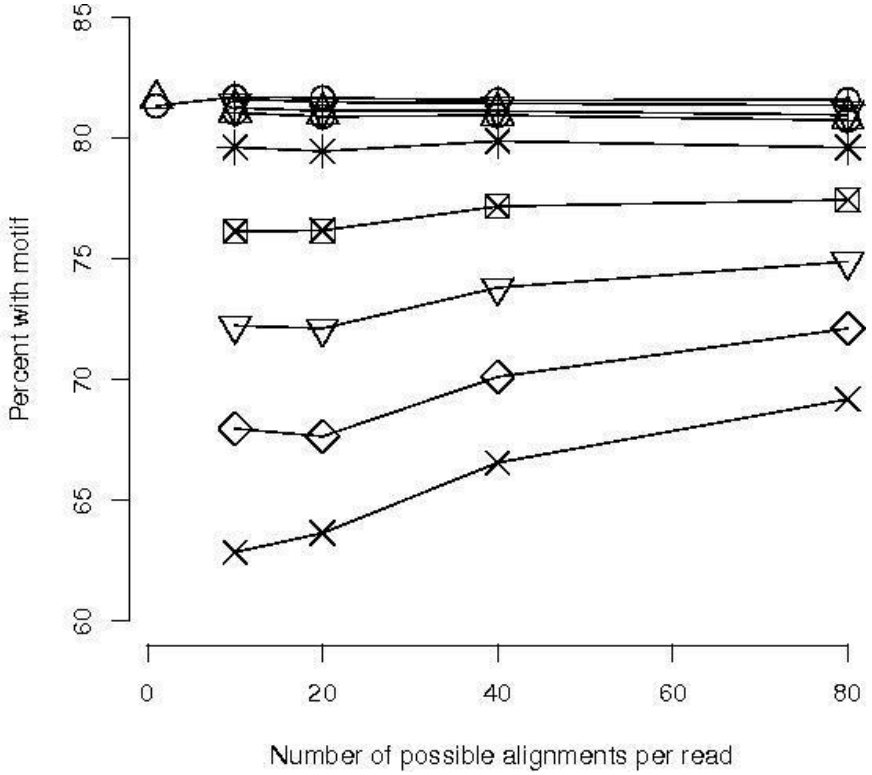


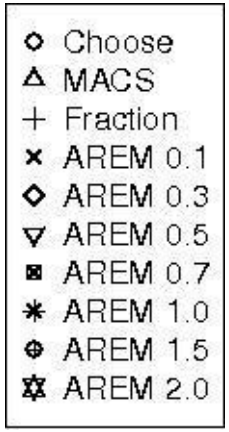
Method	# Alignments	# Peaks	Peak Bases	FDR	New Peaks	Motif	Repeat
Cohesin							
MACS	2,368,229	18,556	9,546,641	2.8%	—	81.67%	56.55%
SICER	2,368,229	17,092	17,374,108	12.71%	—	82.55%	70.42%
AREM 1	2,368,229	19,012	9,353,567	1.9%	—	81.32%	55.30%
AREM 10	7,616,647	19,881	10,225,479	3.8%	1,404	81.04%	58.88%
AREM 20	12,312,878	19,935	10,531,465	3.7%	1,517	80.88%	59.66%
AREM 40	20,527,010	19,863	10,744,836	3.2%	1,546	80.93%	60.34%
AREM 80	34,537,311	19,820	10,972,796	2.9%	1,538	80.73%	60.91%
Srebp-1							
MACS	10,482,005	721	495,968	4.85%	—	46.60%	53.95%
SICER	10,482,005	622	963,778	9.0%	—	59.00%	77.33%
AREM 1	10,482,005	1,438	880,284	8.0%	—	39.08%	53.47%
AREM 10	28,347,869	1,815	996,346	10.5%	262	39.22%	56.04%
AREM 20	44,493,532	1,748	959,646	8.0%	227	39.95%	55.97%
AREM 40	72,453,642	1,685	983,459	8.2%	248	40.34%	56.46%
AREM 80	118,744,757	1,695	987,746	7.3%	272	40.66%	56.73%

Table 2.1: **Comparison of peak-calling methods for cohesin and Srebp-1.** Three peak callers (MACS, SICER, and AREM) were run on both datasets. For AREM, the maximum number of retained alignments per read is varied (from 1 to 80). The total number of peaks and bases covered by peaks is reported as well as the FDR by swapping treatment and control. For both datasets, AREM's minimum enrichment score was fixed at 1.5 with 20 maximum alignments per read. For comparison, the motif background rate of occurence was 4.5% (CTCF) and 27% (Srebp-1) in 100,000 genomic samples, sized similarly to Rad21 MACS peaks and Srebp-1 MACS peaks, respectively.

Min Score and performance

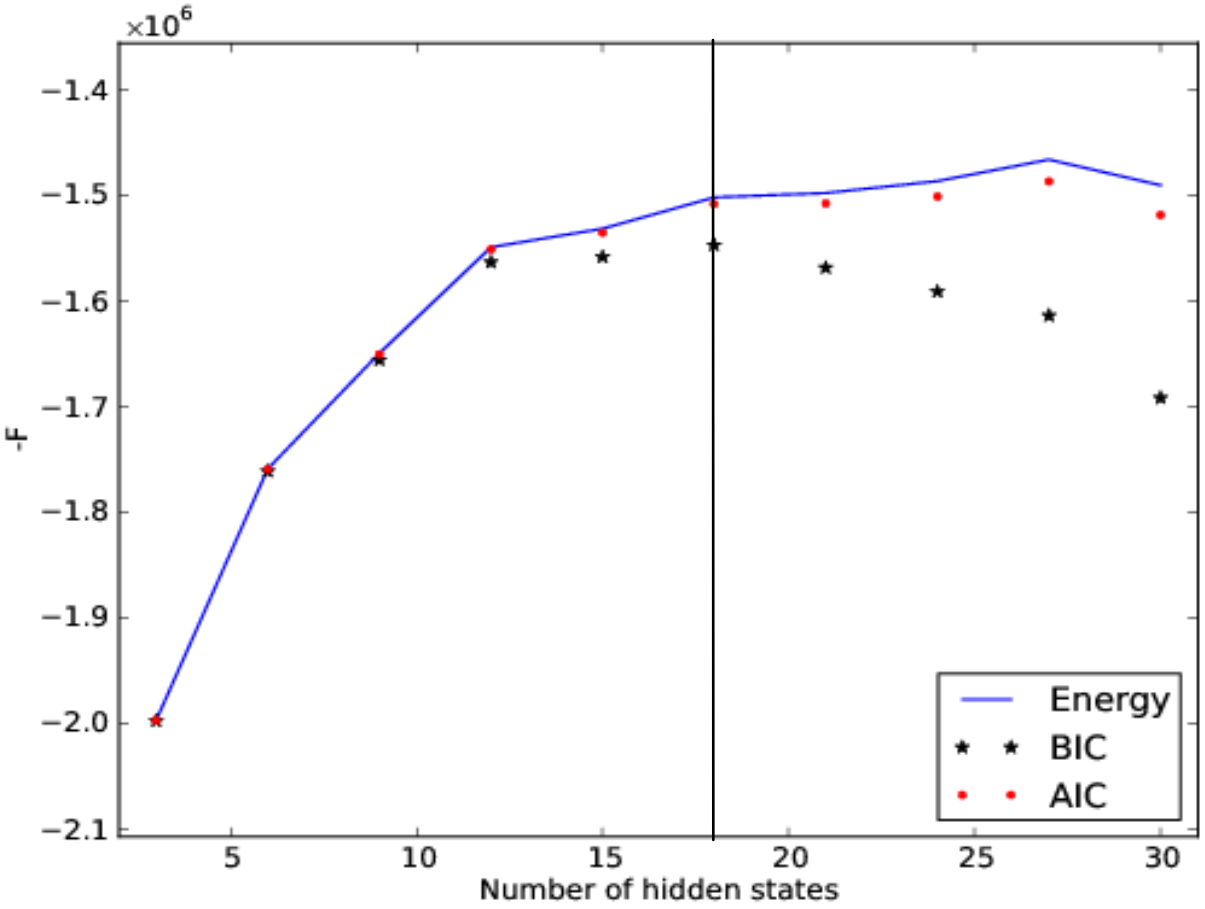






How many chromatin states?

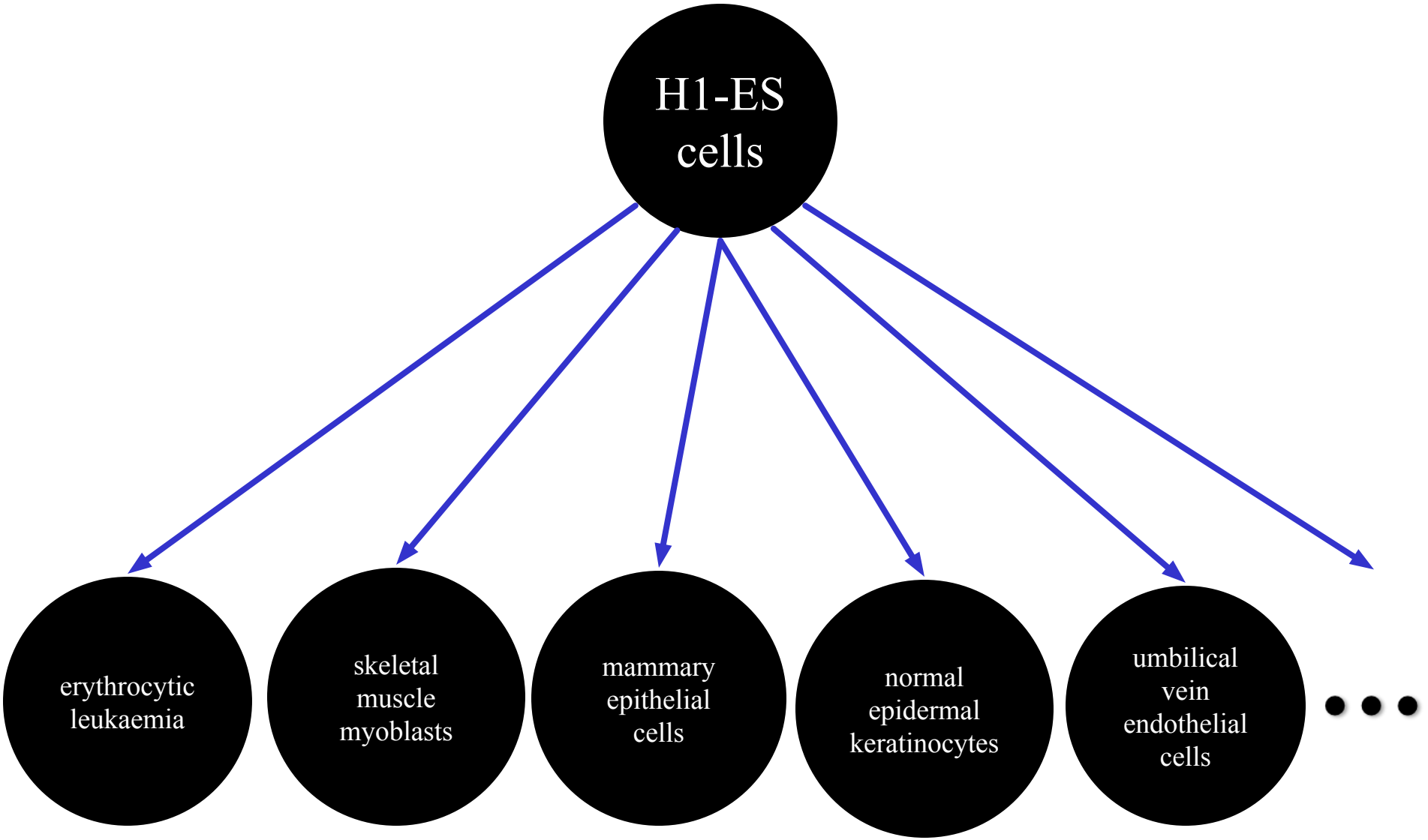
Apply tree-HMM to the Broad dataset
(9 chromatin modification in 9 cell types):

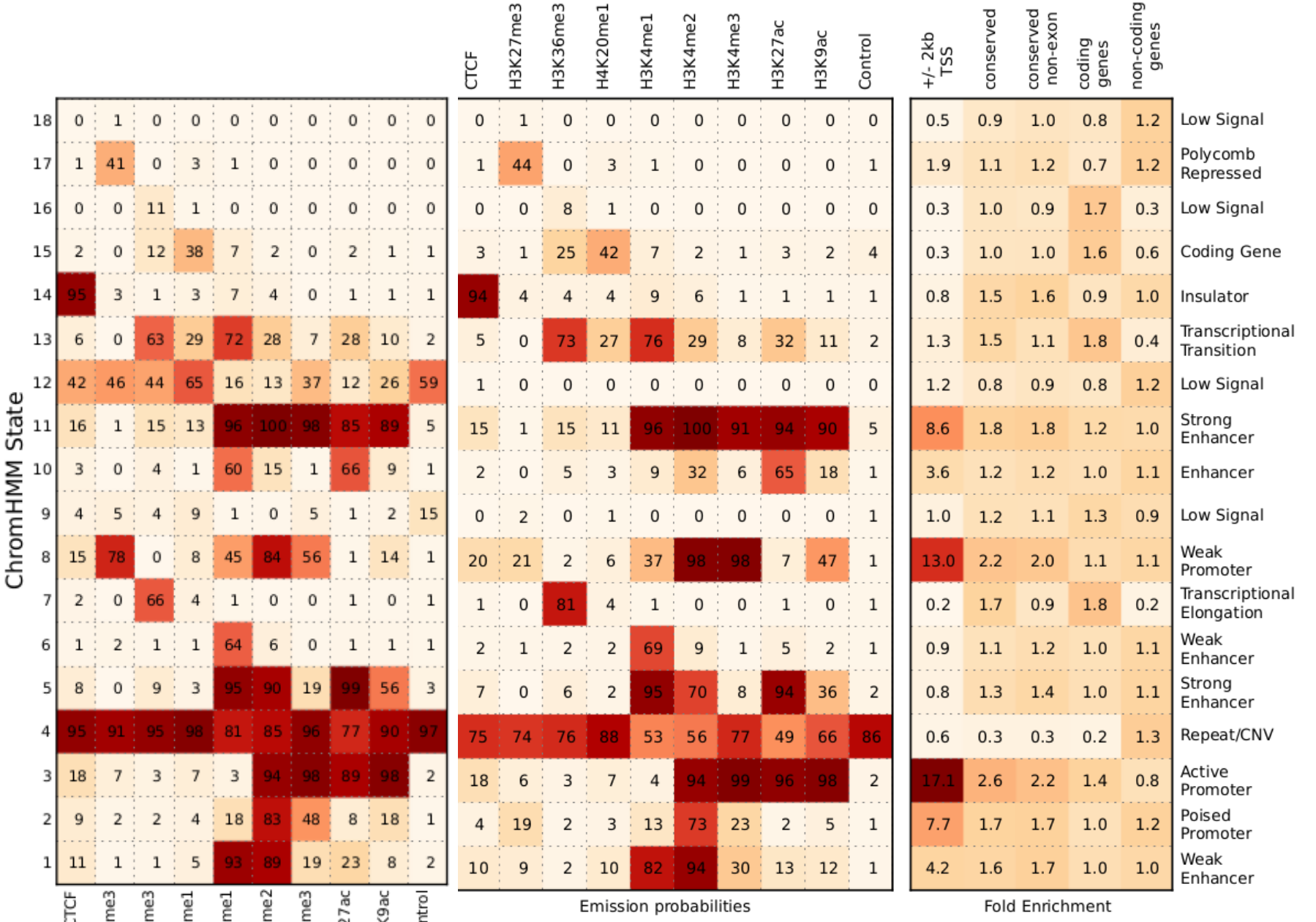


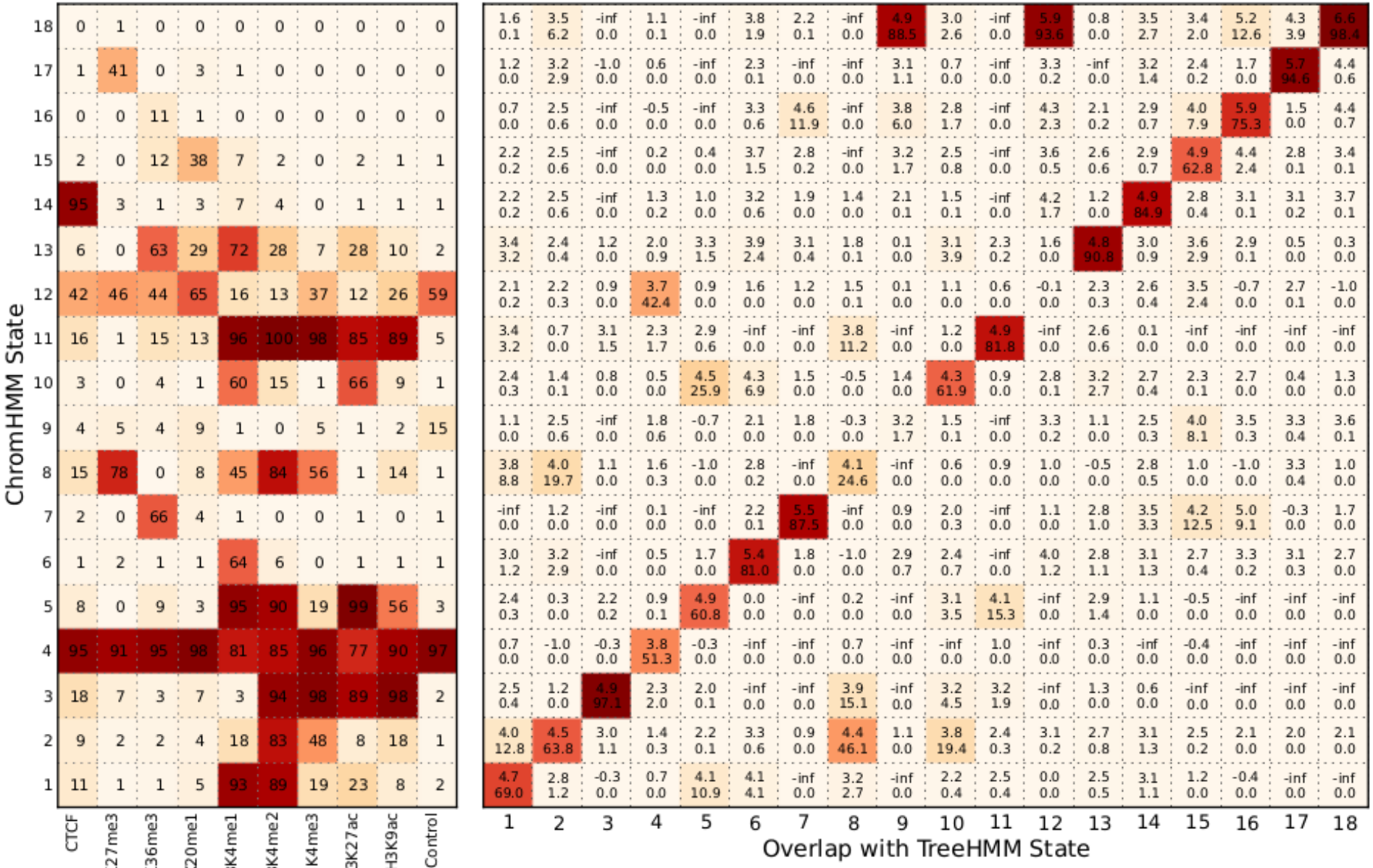
The optimal number of states is **18** by BIC.

Model selection: AIC (Akaike information criterion), BIC (Bayesian information criterion)

A Simple Lineage Tree




ChromHMM State
CTCF
H3K27me3
H3K36me3
H4K20me1
H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K9ac
Control
Emission probabilities
+/- 2kb TSS
conserved
conserved non-exon
coding genes
non-coding genes
Fold Enrichment
18 0 1 0 0 0 0 0 0 0 0 0.5 0.9 1.0 0.8 1.2 Low Signal
17 1 41 0 3 1 0 0 0 0 0 1 44 0 3 1 0 0 0 0 1 1.9 1.1 1.2 0.7 1.2 Polycomb Repressed
16 0 0 11 1 0 0 0 0 0 0 0 0 8 1 0 0 0 0 0 0 0.3 1.0 0.9 1.7 0.3 Low Signal
15 2 0 12 38 7 2 0 2 1 1 3 1 25 42 7 2 1 3 2 4 0.3 1.0 1.0 1.6 0.6 Coding Gene
14 95 3 1 3 7 4 0 1 1 1 94 4 4 4 9 6 1 1 1 1 0.8 1.5 1.6 0.9 1.0 Insulator
13 6 0 63 29 72 28 7 28 10 2 5 0 73 27 76 29 8 32 11 2 1.3 1.5 1.1 1.8 0.4 Transcriptional Transition
12 42 46 44 65 16 13 37 12 26 59 1 0 0 0 0 0 0 0 0 0 1.2 0.8 0.9 0.8 1.2 Low Signal
11 16 1 15 13 96 100 98 85 89 5 15 1 15 11 96 100 91 94 90 5 8.6 1.8 1.8 1.2 1.0 Strong Enhancer
10 3 0 4 1 60 15 1 66 9 1 2 0 5 3 9 32 6 65 18 1 3.6 1.2 1.2 1.0 1.1 Enhancer
9 4 5 4 9 1 0 5 1 2 15 0 2 0 1 0 0 0 0 0 1 1.0 1.2 1.1 1.3 0.9 Low Signal
8 15 78 0 8 45 84 56 1 14 1 20 21 2 6 37 98 98 7 47 1 13.0 2.2 2.0 1.1 1.1 Weak Promoter
7 2 0 66 4 1 0 0 1 0 1 1 0 81 4 1 0 0 1 0 1 0.2 1.7 0.9 1.8 0.2 Transcriptional Elongation
6 1 2 1 1 64 6 0 1 1 1 2 1 2 2 69 9 1 5 2 1 0.9 1.1 1.2 1.0 1.1 Weak Enhancer
5 8 0 9 3 95 90 19 99 56 3 7 0 6 2 95 70 8 94 36 2 0.8 1.3 1.4 1.0 1.1 Strong Enhancer
4 95 91 95 98 81 85 96 77 90 97 75 74 76 88 53 56 77 49 66 86 0.6 0.3 0.3 0.2 1.3 Repeat/CNV
3 18 7 3 7 3 94 98 89 98 2 18 6 3 7 4 94 99 96 98 2 17.1 2.6 2.2 1.4 0.8 Active Promoter
2 9 2 2 4 18 83 48 8 18 1 4 19 2 3 13 73 23 2 5 1 7.7 1.7 1.7 1.0 1.2 Poised Promoter
1 11 1 1 5 93 89 19 23 8 2 10 9 2 10 82 94 30 13 12 1 4.2 1.6 1.7 1.0 1.0 Weak Enhancer


ChromHMM State
18 0 1 0 0 0 0 0 0 0 0
17 1 41 0 3 1 0 0 0 0 0
16 0 0 11 1 0 0 0 0 0 0
15 2 0 12 38 7 2 0 2 1 1
14 95 3 1 3 7 4 0 1 1 1
13 6 0 63 29 72 28 7 28 10 2
12 42 46 44 65 16 13 37 12 26 59
11 16 1 15 13 96 100 98 85 89 5
10 3 0 4 1 60 15 1 66 9 1
9 4 5 4 9 1 0 5 1 2 15
8 15 78 0 8 45 84 56 1 14 1
7 2 0 66 4 1 0 0 1 0 1
6 1 2 1 1 64 6 0 1 1 1
5 8 0 9 3 95 90 19 99 56 3
4 95 91 95 98 81 85 96 77 90 97
3 18 7 3 7 3 94 98 89 98 2
2 9 2 2 4 18 83 48 8 18 1
1 11 1 1 5 93 89 19 23 8 2
CTCF
H3K27me3
H3K36me3
H4K20me1
H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K9ac
Control
1.6 0.1 3.5 6.2 -inf 0.0 1.1 0.1 -inf 0.0 3.8 1.9 2.2 0.1 -inf 0.0 4.9 88.5 3.0 2.6 -inf 0.0 5.9 93.6 0.8 0.0 3.5 2.7 3.4 2.0 5.2 12.6 4.3 3.9 6.6 98.4
1.2 0.0 3.2 2.9 -1.0 0.0 0.6 0.0 -inf 0.0 2.3 0.1 -inf 0.0 -inf 0.0 3.1 1.1 0.7 0.0 -inf 0.0 3.3 0.2 -inf 0.0 3.2 1.4 2.4 0.2 1.7 0.0 5.7 94.6 4.4 0.6
0.7 0.0 2.5 0.6 -inf 0.0 -0.5 0.0 -inf 0.0 3.3 0.6 4.6 11.9 -inf 0.0 3.8 6.0 2.8 1.7 -inf 0.0 4.3 2.3 2.1 0.2 2.9 0.7 4.0 7.9 5.9 75.3 1.5 0.0 4.4 0.7
2.2 0.2 2.5 0.6 -inf 0.0 0.2 0.0 0.4 0.0 3.7 1.5 2.8 0.2 -inf 0.0 3.2 1.7 2.5 0.8 -inf 0.0 3.6 0.5 2.6 0.6 2.9 0.7 4.9 62.8 4.4 2.4 2.8 0.1 3.4 0.1
2.2 0.2 2.5 0.6 -inf 0.0 1.3 0.2 1.0 0.0 3.2 0.6 1.9 0.0 1.4 0.0 2.1 0.1 1.5 0.1 -inf 0.0 4.2 1.7 1.2 0.0 4.9 84.9 2.8 0.4 3.1 0.1 3.1 0.2 3.7 0.1
3.4 3.2 2.4 0.4 1.2 0.0 2.0 0.9 3.3 1.5 3.9 2.4 3.1 0.4 1.8 0.1 0.1 0.0 3.1 3.9 2.3 0.2 1.6 0.0 4.8 90.8 3.0 0.9 3.6 2.9 2.9 0.1 0.5 0.0 0.3 0.0
2.1 0.2 2.2 0.3 0.9 0.0 3.7 42.4 0.9 0.0 1.6 0.0 1.2 0.0 1.5 0.1 0.1 0.0 1.1 0.0 0.6 0.0 -0.1 0.0 2.3 0.3 2.6 0.4 3.5 2.4 -0.7 0.0 2.7 0.1 -1.0 0.0
3.4 3.2 0.7 0.0 3.1 1.5 2.3 1.7 2.9 0.6 -inf 0.0 -inf 0.0 3.8 11.2 -inf 0.0 1.2 0.0 4.9 81.8 -inf 0.0 2.6 0.6 0.1 0.0 -inf 0.0 -inf 0.0 -inf 0.0 -inf 0.0
2.4 0.3 1.4 0.1 0.8 0.0 0.5 0.0 4.5 25.9 4.3 6.9 1.5 0.0 -0.5 0.0 1.4 0.0 4.3 61.9 0.9 0.0 2.8 0.1 3.2 2.7 2.7 0.4 2.3 0.1 2.7 0.0 0.4 0.0 1.3 0.0
1.1 0.0 2.5 0.6 -inf 0.0 1.8 0.6 -0.7 0.0 2.1 0.0 1.8 0.0 -0.3 0.0 3.2 1.7 1.5 0.1 -inf 0.0 3.3 0.2 1.1 0.0 2.5 0.3 4.0 8.1 3.5 0.3 3.3 0.4 3.6 0.1
3.8 8.8 4.0 19.7 1.1 0.0 1.6 0.3 -1.0 0.0 2.8 0.2 -inf 0.0 4.1 24.6 -inf 0.0 0.6 0.0 0.9 0.0 1.0 0.0 -0.5 0.0 2.8 0.5 1.0 0.0 -1.0 0.0 3.3 0.4 1.0 0.0
-inf 0.0 1.2 0.0 -inf 0.0 0.1 0.0 -inf 0.0 2.2 0.1 5.5 87.5 -inf 0.0 0.9 0.0 2.0 0.3 -inf 0.0 1.1 0.0 2.8 1.0 3.5 3.3 4.2 12.5 5.0 9.1 -0.3 0.0 1.7 0.0
3.0 1.2 3.2 2.9 -inf 0.0 0.5 0.0 1.7 0.0 5.4 81.0 1.8 0.0 -1.0 0.0 2.9 0.7 2.4 0.7 -inf 0.0 4.0 1.2 2.8 1.1 3.1 1.3 2.7 0.4 3.3 0.2 3.1 0.3 2.7 0.0
2.4 0.3 0.3 0.0 2.2 0.2 0.9 0.1 4.9 60.8 0.0 0.0 -inf 0.0 0.2 0.0 -inf 0.0 3.1 3.5 4.1 15.3 -inf 0.0 2.9 1.4 1.1 0.0 -0.5 0.0 -inf 0.0 -inf 0.0 -inf 0.0
0.7 0.0 -1.0 0.0 -0.3 0.0 3.8 51.3 -0.3 0.0 -inf 0.0 -inf 0.0 0.7 0.0 -inf 0.0 -inf 0.0 1.0 0.0 -inf 0.0 0.3 0.0 -inf 0.0 -0.4 0.0 -inf 0.0 -inf 0.0 -inf 0.0
2.5 0.4 1.2 0.0 4.9 97.1 2.3 2.0 2.0 0.1 -inf 0.0 -inf 0.0 3.9 15.1 -inf 0.0 3.2 4.5 3.2 1.9 -inf 0.0 1.3 0.0 0.6 0.0 -inf 0.0 -inf 0.0 -inf 0.0 -inf 0.0
4.0 12.8 4.5 63.8 3.0 1.1 1.4 0.3 2.2 0.1 3.3 0.6 0.9 0.0 4.4 46.1 1.1 0.0 3.8 19.4 2.4 0.3 3.1 0.2 2.7 0.8 3.1 1.3 2.5 0.2 2.1 0.0 2.0 0.0 2.1 0.0
4.7 69.0 2.8 1.2 -0.3 0.0 0.7 0.0 4.1 10.9 4.1 4.1 -inf 0.0 3.2 2.7 -inf 0.0 2.2 0.4 2.5 0.4 0.0 0.0 2.5 0.5 3.1 1.1 1.2 0.0 -0.4 0.0 -inf 0.0 -inf 0.0
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
Overlap with TreeHMM State

Promoters				
Factor	treeHMM		ChromHMM	
	All	Unique	All	Unique
Taf1	32,069 (41.6x)	1,489 (15.2x)	35,082 (26.0x)	4,502 (6.7x)
Oct4	4,980 (23.8x)	231 (8.7x)	6,932 (19x)	2,183 (12x)
Klf4	2,622 (18.1x)	105 (5.7x)	3,819 (15.1x)	1,302 (10.3x)
p300	141 (1.0x)	16 (0.9x)	1,597 (6.4x)	1,472 (11.8x)
Nanog	1,556 (1.5x)	227 (1.7x)	8,650 (4.7x)	7,321 (7.7x)
Sox2	412 (1.6x)	63 (2.0x)	2,509 (5.7x)	2,160 (9.8x)

Enhancers				
Factor	treeHMM		ChromHMM	
	All	Unique	All	Unique
Taf1	8,095 (2.5x)	4,293 (4.4x)	5,611 (2.2x)	1,809 (5.3x)
Oct4	3,914 (4.5x)	2,060 (7.8x)	2,274 (3.3x)	420 (4.5x)
Klf4	2,143 (3.6x)	1,294 (7.1x)	1,003 (2.1x)	154 (2.4x)
p300	7,253 (12.2x)	1,517 (8.4x)	5,861 (12.2x)	125 (2.0x)
Nanog	39,829 (9.1x)	7,941 (6.0x)	33,561 (9.6x)	1,673 (3.5x)
Sox2	9,786 (9.4x)	2,185 (6.9x)	7,952 (9.5x)	351 (3.1x)