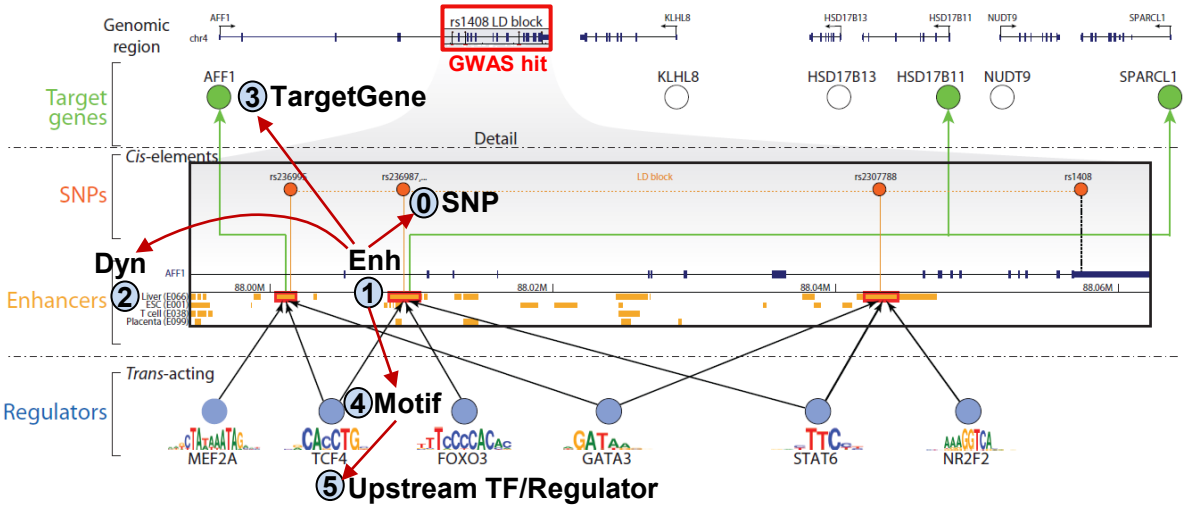
MIT 18.700/18.701/6.047/6.878/HST.507
Computational Biology: Genomes, Networks, Evolution



Lecture 6: Regulatory circuitry

Epigenome Dynamics: Joint Chromatin State Learning
Enhancer-gene linking: Correlation, Hi-C, eQTLs
TF motif discovery: Enrichment, EM, Gibbs Sampling
Deep learning convolution CNNs for motif discovery
Global motif discovery: Comparative Genomics
Motif Instance Identification: Branch Length Score
Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

MLCB - Machine Learning in Computational Biology - Fall 2024 - Profs. Manolis Kellis + Eric Alm - 6.8700/6.8701/20.s900/20.s948/HST.507

Homeworks	Project / Mentoring (Fri)	Wk	Date	Lec	Topic
			Introduction: Machine Learning, Deep Learning, Generative AI, and the Unification of Biology		
HW0 out Thu 9/5		1	Thu-Sep-05	L1	Course Overview, Machine Learning, Deep Learning, Inference, Genome, Proteins, Chemistry, Imaging
			Module 1: Genomics, Epigenomics, Single-Cell, Networks, Circuitry		
HW0 due Wed 9/11		2	Tue-Sep-10	L2	Expression Analysis, Clustering/Classification, Gaussian Mixture Models, K-means, Bayesian Inf, Gen-vs-DiscrML
HW1 out Thu 9/12	0=Self Introductions	2	Thu-Sep-12	L3	Single-cell genomics, sc-mutli-omics, non-linear embeddings, spatial transcriptomics, next-gen technologies
		3	Tue-Sep-17	L4	Sequential Data, Alignment, DynProg, Hidden Markov Models, Parsing, Posterior Decoding, HMM architectures
		3	Thu-Sep-19	L5	Epigenomics: Signal Modeling, Peak calling, Chromatin states, 3D structure, Hi-C, Genome Topology
		4	Tue-Sep-24	L6	Regulatory Genomics: Motifs, Information, ChIP, Gibbs Sampling, EM, CNNs for Genome Parsing
HW1 due Mon 9/30	1=Select previous paper(s)	4	Thu-Sep-26	L7	Regulatory Networks: Graphs, Linear Algebra, PCA, SVD, Dimentionality Reduction, TF-enhancer-gene circuitry
			Module 2: Protein Structure, Protein Language Models, Geometric Deep Learning		
HW2 out Thu 10/3		5	Tue-Oct-01	L8	Intro to structural biology
		5	Thu-Oct-03	L9	Protein structure and folding: Diffusion models, Cryo-EM, Protein design
		6	Tue-Oct-08	L10	Intro to transformers and Large Language Models LLMs
	2=Proposal+Feasibility	6	Thu-Oct-10	L11	Protein Language Models PLMs and Transfer Learning
		7	Tue-Oct-15	-	-- No Class -- Student holiday
HW2 due Mon 10/21		7	Thu-Oct-17	L12	DNA language models: Chromatin Structure
			Module 3: Chemistry, Therapeutics, Graph Neural Networks		
HW3 out Thu 10/24		8	Tue-Oct-22	L13	Overview of drug development
	3=OffHrs|Update|Feedback	8	Thu-Oct-24	L14	Intro to small molecules
		9	Tue-Oct-29	L15	Representation of small molecules: Graphs, GNNs, Transformers, RDKit
		9	Thu-Oct-31	L16	Docking: Small molecule - proteins docking
		10	Tue-Nov-05	L17	Disease Association Mapping, genetics, GWAS, linkage analysis, disease circuitry, variant-to-function
HW3 due Tue 11/12	4=OffHrs|Update|Feedback	10	Thu-Nov-07	L18	Quantitative trait mapping, molecular traits, eQTLs, mediation analysis, iMWAS, multi-modal QTLs
			Module 4: Electronic Health Records, Imaging, Evolution, Metabolism		
No HW		11	Tue-Nov-12	L19	Electronic Health Records, AllOfUs, UKBioBank, Medical Genomics, Pop-Scale Cohorts, Multi-Ancestry [not quiz'd]
		11	Thu-Nov-14	L20	**-- In-class Quiz**
		12	Tue-Nov-19	L21	Imaging methods for biological applications
	5=Midcourse report	12	Thu-Nov-21	L22	Comparative genomics, Conservation, Evolutionary signatures, PhyloCSF, RNA structure, Motif BLS2conf
		13	Tue-Nov-26	L23	Evolution, Phylogenetics, Phylogenomics, Duplication, RNA world, RNA folding, lincRNAs, RNA modifications, m6A
		13	Thu-Nov-28	-	-- No Class -- Thanksgiving Holiday
		14	Tue-Dec-03	L24	Modeling metabolism: Flux balance analysis
	6=WriteUp, Slides Due	14	Thu-Dec-05	L25	Measuring metabolism: Metabolomics and Deep Learning
			Final Projects		
	7=In Class Presentations	15	Tue-Dec-10	L26	**Project Presentations (6-8 mins/team). Report due Fri@11.59p, Slides due Mon@11.59p, Present Live Tue**

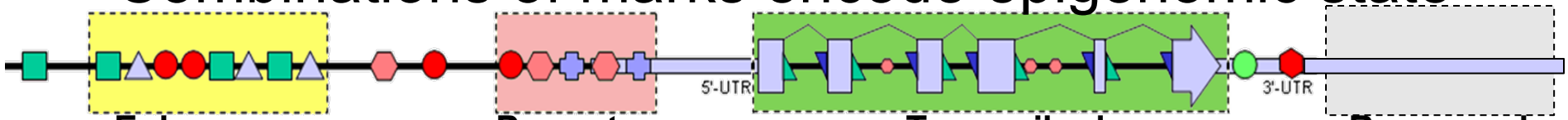
HMMs Review + Learning

Supervised + Unsupervised

Viterbi Training (max, top path)

Baum Welch Training (EM, all paths)

Combinations of marks encode epigenomic state

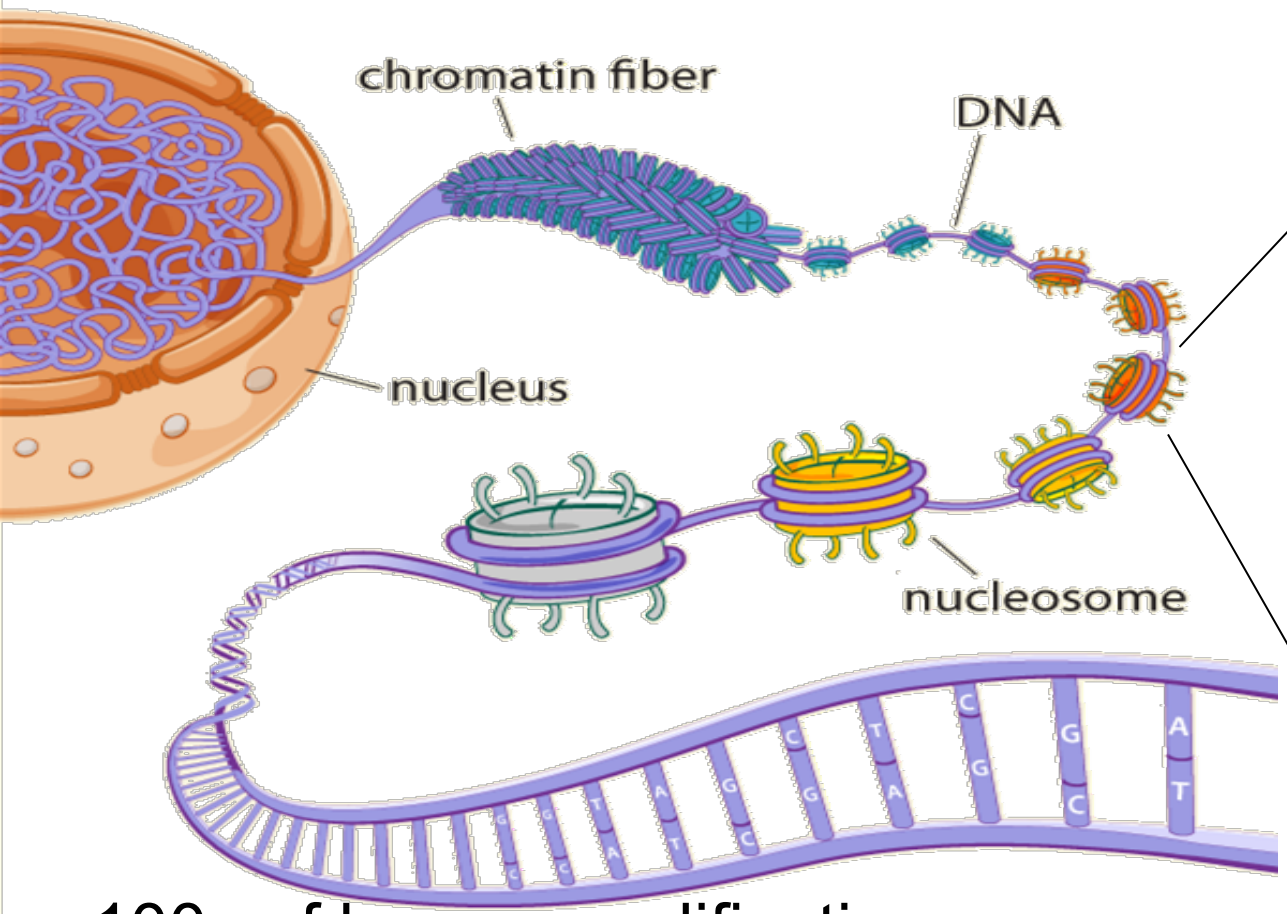


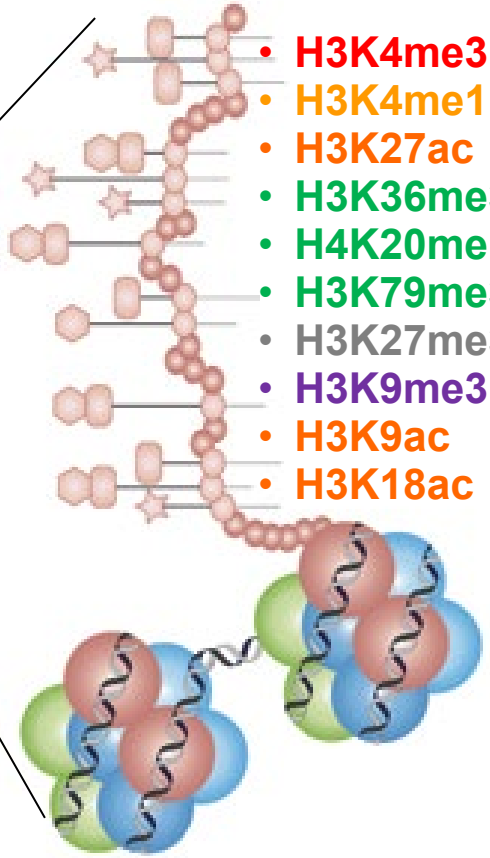
Enhancers
- H3K4me1
- H3K27ac
- DNase

Promoters
- H3K4me3
- H3K9ac
- DNase

Transcribed
- H3K36me3
- H3K79me2
- H4K20me1

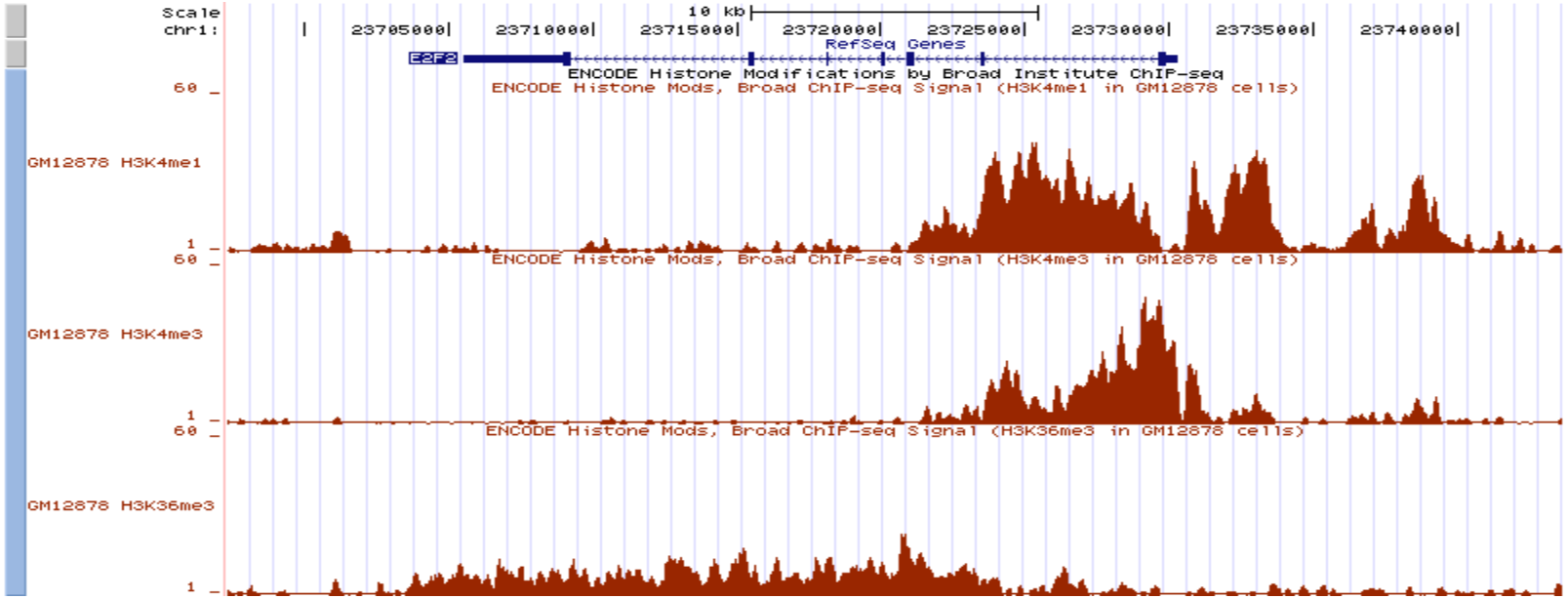
Repressed
- H3K9me3
- H3K27me3
- DNAmethyl





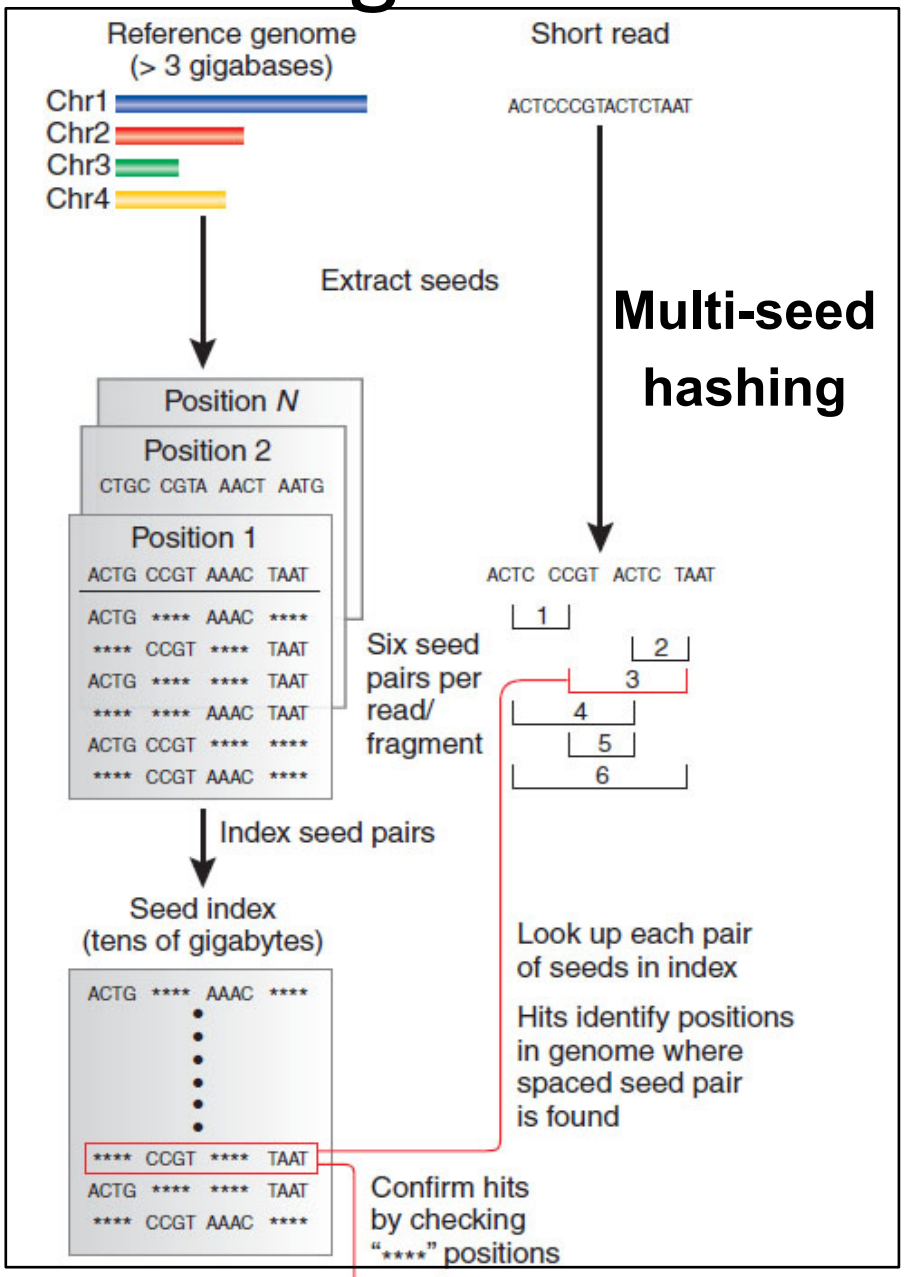
- 100s of known modifications, many new still emerging
- Systematic mapping using ChIP-, Bisulfite-, DNase-Seq

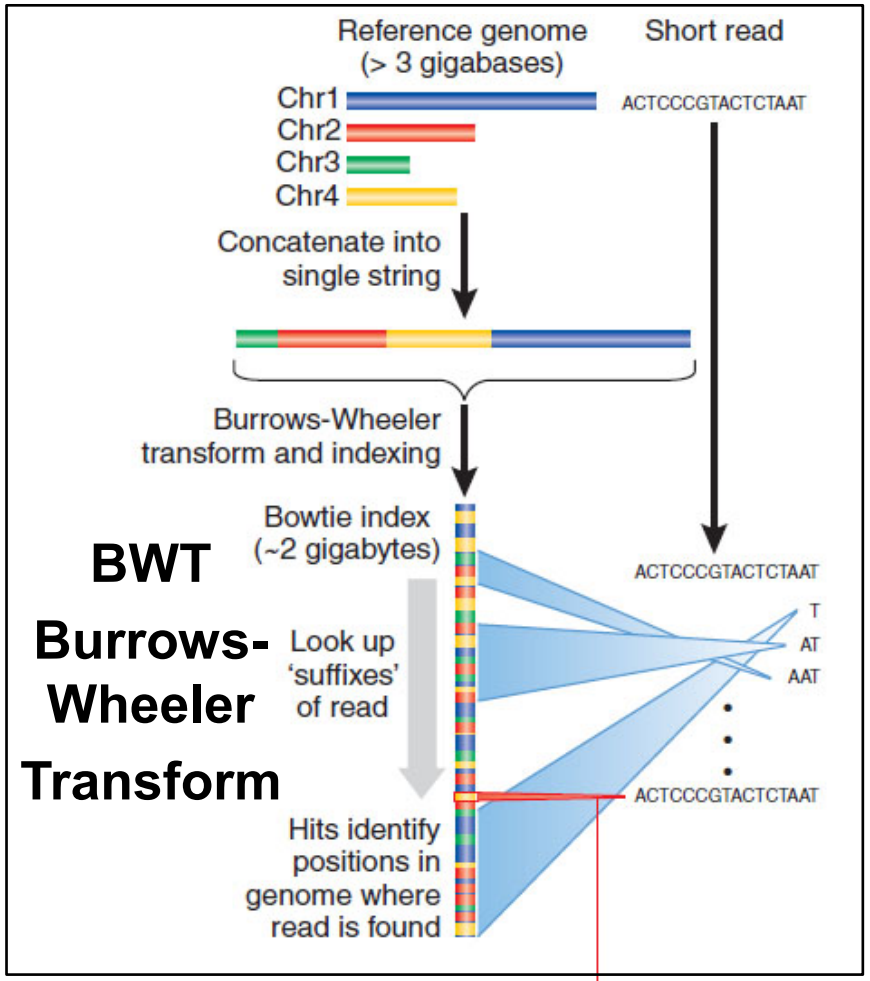
ChIP-Seq Histone Modifications: What the raw data looks like



- Each sequence tag is 30 base pairs long
- Tags are mapped to unique positions in the ~3 billion base reference genome
- Number of reads depends on sequencing depth. Typically on the order of 10 million mapped reads.

Hashing vs. Burrows Wheeler Transform



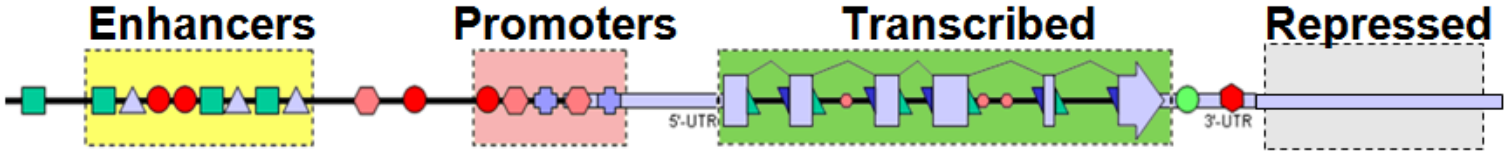


Today: How does the BW transform actually work?

Convert each hit back to genome location

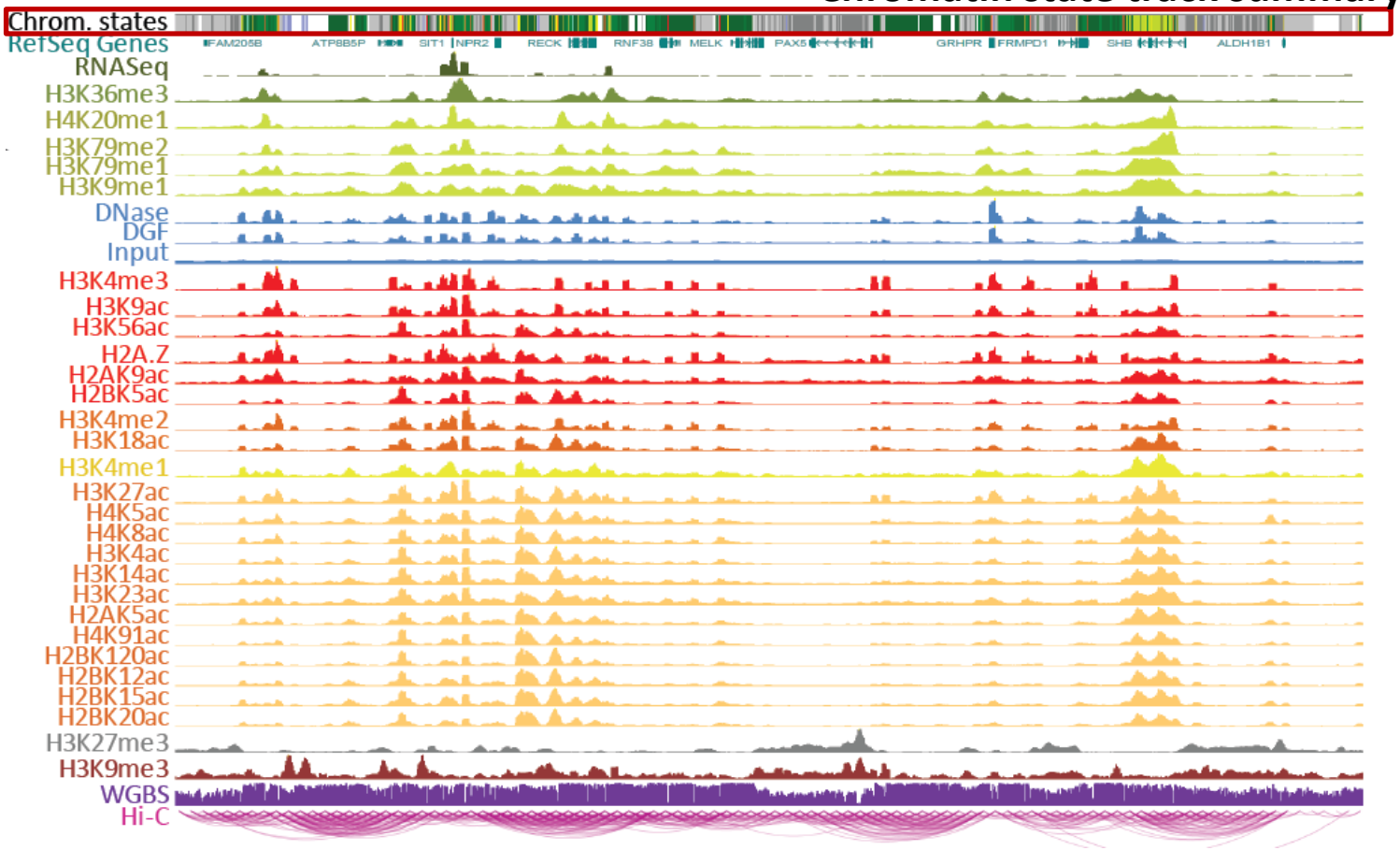
Report alignment to user

Summarize multiple marks into chromatin states



Chromatin state track summary

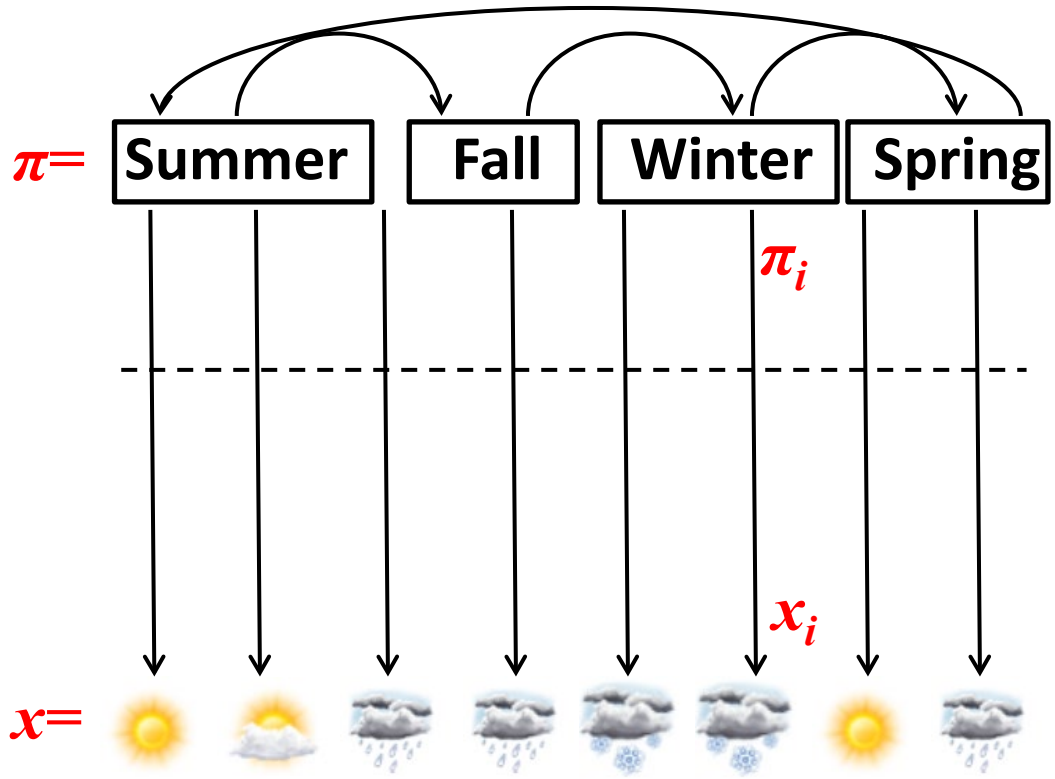
30+ epigenomics marks



WashU Epigenome Browser

ChromHMM: multi-variate hidden Markov model

HMM nomenclature for this course



Transitions: $a_{kl}=P(\pi_i=l|\pi_{i-1}=k)$

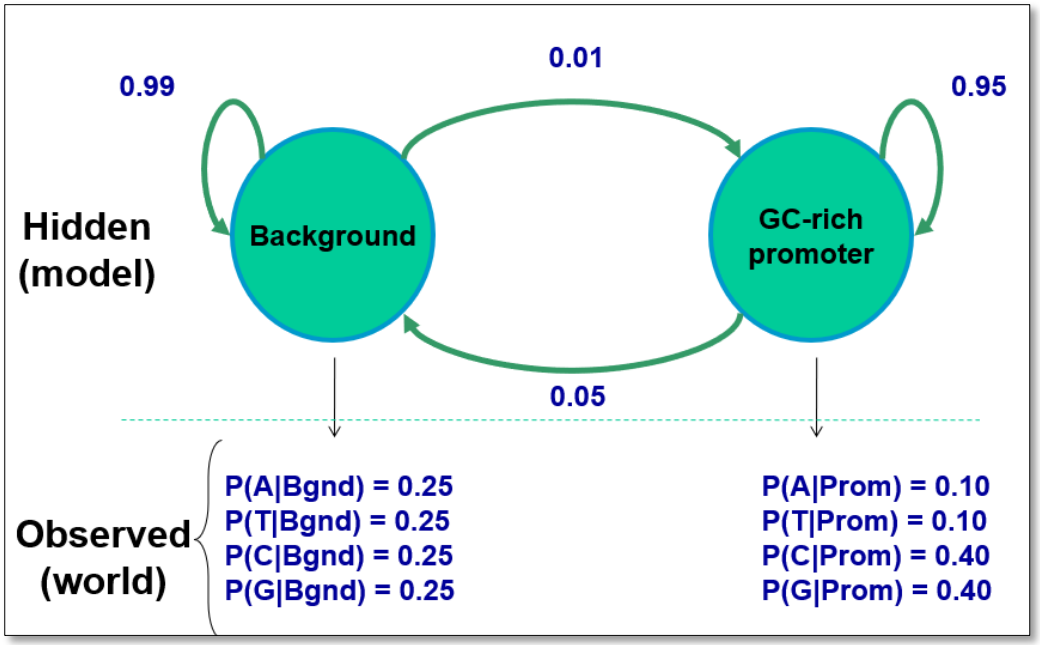
Transition probability from state k to state l

Emissions: $e_k(x_i)=P(x_i|p_i=k)$

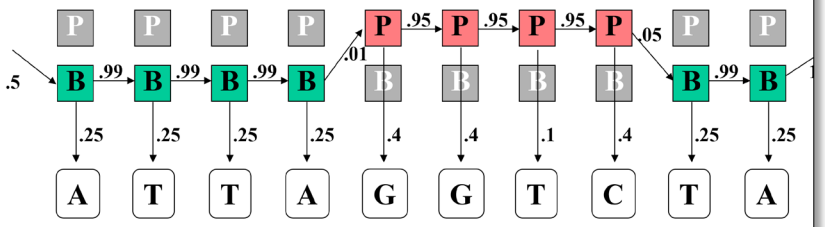
Emission probability of symbol x_i from state k

- Vector x = Sequence of observations
- Vector π = Hidden path (sequence of hidden states)
- Transition matrix A=a_{kl}=probability of $k \rightarrow l$ state transition
- Emission vector $E=e_k(x_i)$ = prob. of observing x_i from state k
- Bayes's rule: Use $P(x_i|\pi_i=k)$ to estimate $P(\pi_i=k|x_i)$

HMMs for Genome Annotation/Parsing: Viterbi Algo.



Score annotation (path π) given a seq (obs x): B-P



What is the likelihood of

π = Bgnd, Bgnd, Bgnd, Bgnd, Prom, Prom, Prom, Prom, Bgnd, Bgnd

and sequence

x = A , T, T, A, G, G, T, C, T, A

emission transition emission transition emission

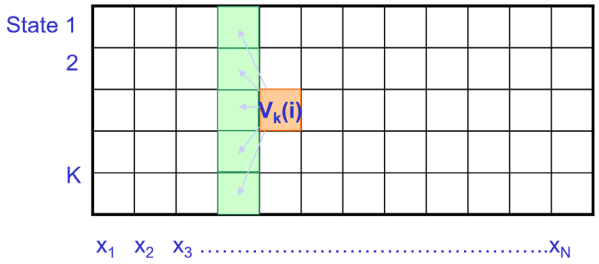
$p = ½ \times P(A \mid Bgnd)\, P(Bgnd_{i+1} \mid Bgnd_i)\, P(T \mid Bgnd)\, P(Bgnd \mid Bgnd) \ldots P(A \mid Bgnd)$

$= ½ \times (.1)^1 \times (.4)^3 \times (.25)^6 \times (0.99)^4 \times (0.01)^1 \times (0.95)^3 \times (0.05)^1$

$= 6.4 \times 10^{-10}$

Much less likely, due to high cost of transitions

The Viterbi Algorithm



Input: x = x1……xN

Initialization:

$V_0(0)=1$, $V_k(0) = 0$, for all k > 0

Iteration:

$V_k(i) = e_K(x_i) \times \max_j a_{jk}\, V_j(i\text{-}1)$

Termination:

$P(x, \pi^*) = \max_k V_k(N)$

Traceback:

Follow max pointers back

Similar to aligning states to seq

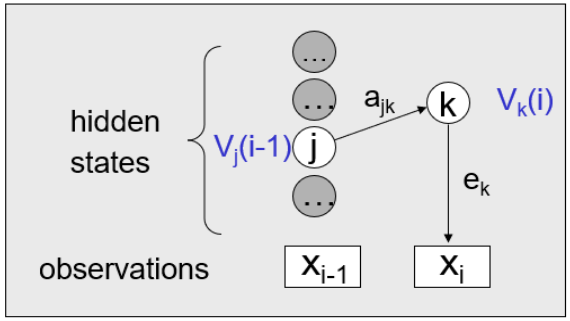
In practice:

Use log scores for computation

Running time and space:

Time: $O(K^2N)$

Space: $O(KN)$



- Assume we know V_j for the previous time step (i-1)
- Calculate $V_k(i) = e_k(x_i) * \max_j (V_j(i\text{-}1) \times a_{jk})$

$V_k(i)$: current max

$e_k(x_i)$: this emission

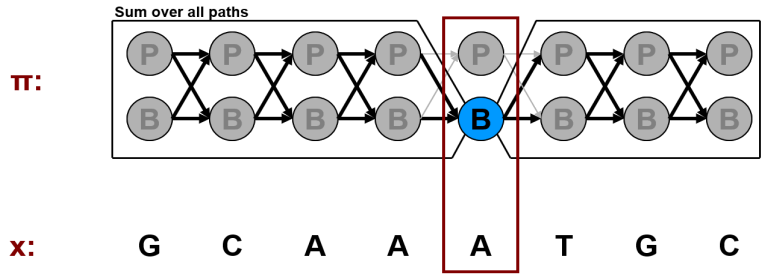
$V_j(i-1)$: max ending in state j at step i

a_{jk}: Transition from state j

all possible previous states j

HMMs over all paths: Posterior Decoding

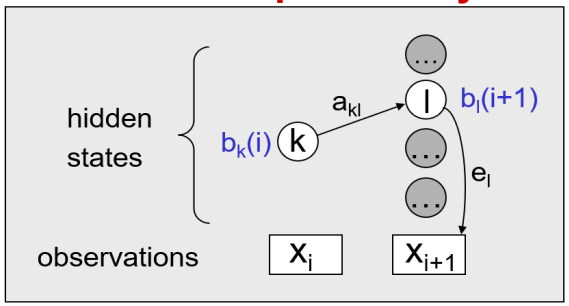
Calculate most probable label at a single position



$P(\text{Label}_i = B|x)$

- Calculate most probable label, L^*_i , at each position i
- Do this for all N positions gives us $\{L^*_1, L^*_2, L^*_3 \ldots L^*_N\}$
- How much information have we observed? Three settings:
  - **Observed nothing: Use prior information**
  - **Observed only character at position i: Prior + emission probability**
  - **Observed entire sequence: Posterior decoding**

Calculate total end probability recursively

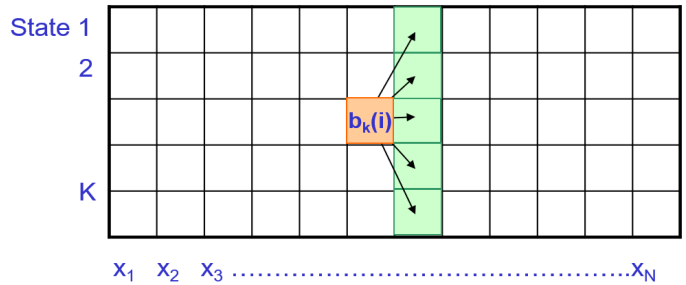


- Assume we know b_l for the next time step (i+1)
- Calculate $b_k(i)$ = sum_l ($e_l(x_{i+1})$ × a_{kl} × $b_l(i+1)$)

current max; next emission; transition to next state; prob sum from state l to end

sum over all possible next states

The Backward Algorithm



Input: x = x1……xN

Initialization:

$b_k(N) = a_{k0}$, for all k

Iteration:

$b_k(i) = \Sigma_l\ e_l(x_{i+1})\ a_{kl}\ b_l(i+1)$

Termination:

$P(x) = \Sigma_l\ a_{0l}\ e_l(x_1)\ b_l(1)$

In practice:

Sum of log scores is difficult

→ approximate exp(1+p+q)

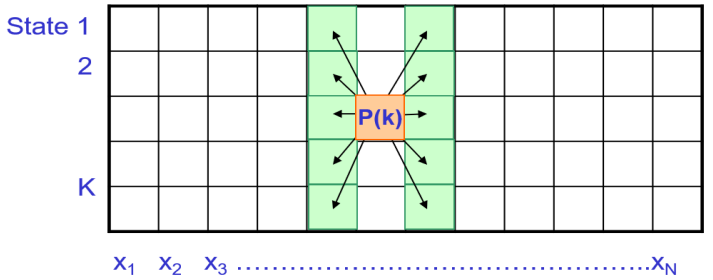
→ scaling of probabilities

Running time and space:

Time: $O(K^2N)$

Space: $O(K)$

Putting it all together: Posterior decoding



- $P(k) = P(\pi_i = k \mid x) = f_k(i) * b_k(i) / P(x)$
  - Probability that i^{th} state is k, given all emissions x
- Posterior decoding
  - Find the most likely state at position i over all possible hidden paths given the observed sequence x
  - $\hat{\pi}_i = \text{argmax}_k\ P(\pi_i = k \mid x)$
- Posterior decoding 'path' $\hat{\pi}_i$
  - For classification, more informative than Viterbi path π^*
    - More refined measure of "which hidden states" generated x
  - However, it may give an invalid sequence of states
    - Not all j→k transitions may be possible

HMM Foundations, Parsing, Decoding, Learning

1. HMM basics, evaluation, parsing, posterior decoding
   - Observations, Models, Bayes' rule, Bayesian inference
   - Markov Chains and Hidden Markov Models
   - Calculating joint probability of one (seq,parse) $P(x,\pi)$
   - Viterbi algorithm: Find best parse $\pi^* = \text{argmax}_\pi P(x,\pi)$
   - Forward algorithm: Find total $P(x)$, sum over all paths
   - Posterior Decoding: Most likely state π_i (over all paths)
2. Learning (ML training, Baum-Welch, Viterbi training)
   - Supervised: Find $e_i(.)$ and a_{ij} given labeled sequence
   - Unsupervised: given only x → annotation + params
3. Increasing the 'state' space / adding memory
   - Finding GC-rich regions vs. finding CpG islands
   - Gene structures GENSCAN, chromatin ChromHMM

The six algorithmic settings for HMMs

	One path	All paths
Scoring	1. Scoring x, one path $P(x,\pi)$ Prob of a path, emissions ✓	2. Scoring x, all paths $P(x) = \sum_{\pi} P(x,\pi)$ Prob of emissions, over all paths ✓
Decoding	3. Viterbi decoding $\pi^* = \text{argmax}_{\pi} P(x,\pi)$ Most likely path ✓	4. Posterior decoding $\pi\hat{} = \{\pi_i \mid \pi_i = \text{argmax}_k \sum_{\pi} P(\pi_i = k \mid x)\}$ Path containing the most likely state at any time point.
Learning	5. Supervised learning, given π $\Lambda^* = \text{argmax}_{\Lambda} P(x,\pi\mid\Lambda)$ 6. Unsupervised learning. $\Lambda^* = \text{argmax}_{\Lambda} \max_{\pi} P(x,\pi\mid\Lambda)$ Viterbi training, best path	6. Unsupervised learning $\Lambda^* = \text{argmax}_{\Lambda} \sum_{\pi} P(x,\pi\mid\Lambda)$ Baum-Welch training, over all paths

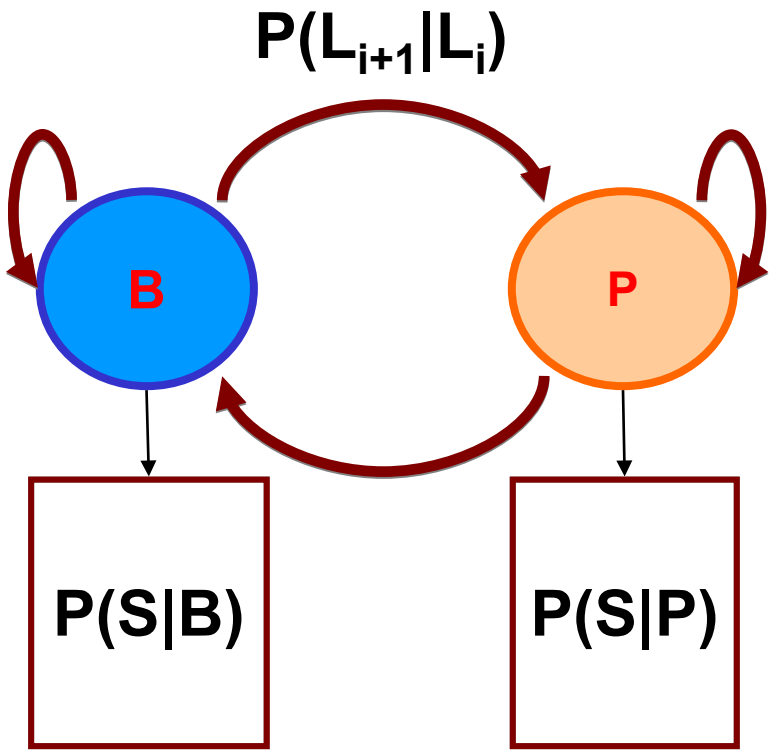
Learning: How to train an HMM

Transition probabilities

e.g. $P(P_{i+1}|B_i)$ – the probability of entering a pathogenicity island from background DNA

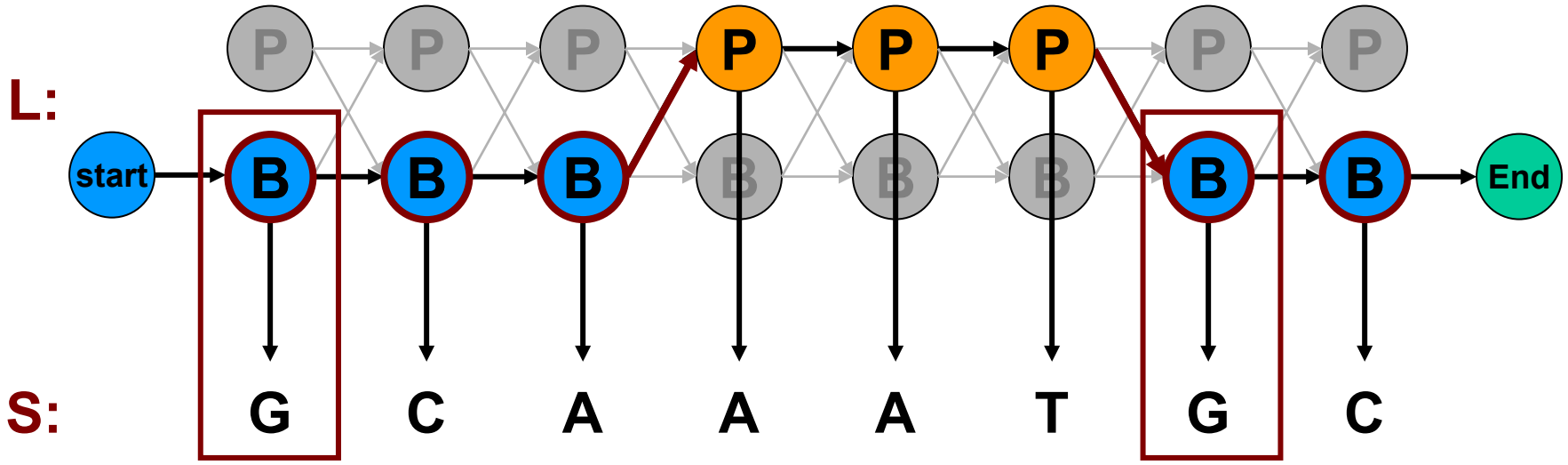
Emission probabilities

i.e. the nucleotide frequencies for background DNA and pathogenicity islands



Learning From labeled Data
→ Maximum Likelihood Estimation

If we have a sequence that has islands marked, we can simply count



$P(L_{i+1}|L_i)$

	B_{i+1}	P_{i+1}
B_i		
P_i		

P(S|B)

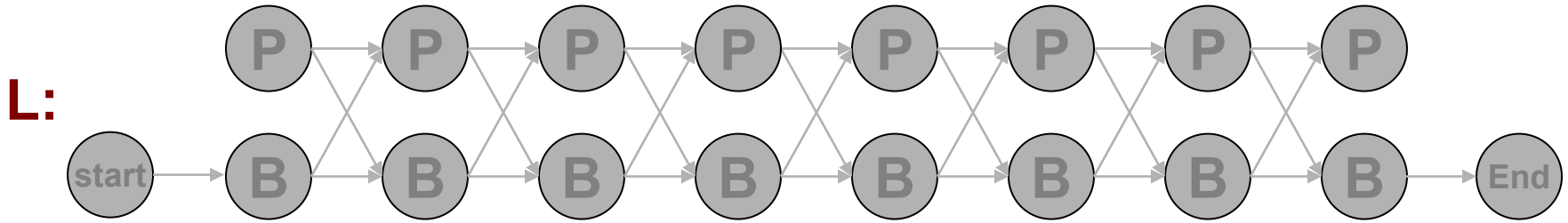
A:
T:
G:
C:
!

P(S|P)

A:
T:
G:
C:
ETC..

Unlabeled Data

How do we know how to count?



S: G C A A A T G C

?

$P(L_{i+1}|L_i)$

	B_{i+1}	P_{i+1}	End
B_i			
P_i			
Start			

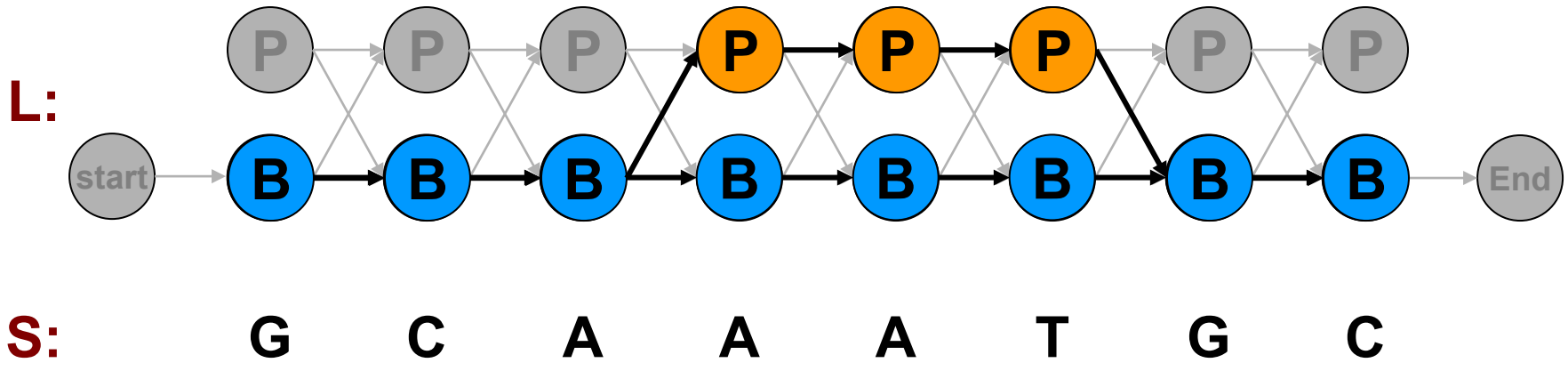
P(S|B)

A:
T:
G:
C:

P(S|P)

A:
T:
G:
C:

Unlabeled Data



An idea:

1. Imagine we start with some parameters
2. We *could* calculate the most likely path, P*, given those parameters and S
3. We *could* then use P* to update our parameters by maximum likelihood
4. And iterate (to convergence)

~~$P(L_{i+1}|L_i)^0$ $P(S|B)^0$ $P(S|P)^0$~~

~~$P(L_{i+1}|L_i)^1$ $P(S|B)^1$ $P(S|P)^1$~~

$P(L_{i+1}|L_i)^2$ $P(S|B)^2$ $P(S|P)^2$

…

$P(L_{i+1}|L_i)^K$ $P(S|B)^K$ $P(S|P)^K$

Simple case: Viterbi Training

Initialization:

Pick the best-guess for model parameters

(or arbitrary)

Iteration:

1. Perform Viterbi, to find π^*
2. Calculate A_{kl}, $E_k(b)$ according to π^* + pseudocounts
3. Calculate the new parameters a_{kl}, $e_k(b)$

Until convergence

Notes:

- Convergence to local maximum guaranteed. Why?
- Does not maximize $P(x \mid \theta)$
- In general, worse performance than Baum-Welch

	One path	All paths
Scoring	1. Scoring x, one path $P(x,\pi)$ Prob of a path, emissions ✓	2. Scoring x, all paths $P(x) = \sum_\pi P(x,\pi)$ Prob of emissions, over all paths ✓
Decoding	3. Viterbi decoding $\pi^* = \text{argmax}_\pi P(x,\pi)$ Most likely path ✓	4. Posterior decoding $\pi\text{^} = \{\pi_i \mid \pi_i = \text{argmax}_k \Sigma_\pi P(\pi_i = k \mid x)\}$ Path containing the most likely state at any time point. ✓
Learning	5. Supervised learning, given π $\Lambda^* = \text{argmax}_\Lambda P(x,\pi\mid\Lambda)$ ✓ 6. Unsupervised learning. $\Lambda^* = \text{argmax}_\Lambda \max_\pi P(x,\pi\mid\Lambda)$ ✓ Viterbi training, best path	6. Unsupervised learning $\Lambda^* = \text{argmax}_\Lambda \Sigma_\pi P(x,\pi\mid\Lambda)$ Baum-Welch training, over all paths

Expectation Maximization (EM)

Widely used in Comp Bio. This course alone:
- **HMMs: Baum Welch**
- **Expression Clustering: Fuzzy K-Means**
- **Regulatory motif discovery: MEME**
- **Systems genetics: GWAS priors/enrichments**
- **Phylogenomics: Estimate branch lengths/rates**

The basic idea is always the same:

1. Use model to estimate missing data (E step)

2. Use estimate to update model (M step)

3. Repeat until convergence

EM is a general approach for learning models (ML estimation) when there is "missing data"

1. **Initialize parameters randomly**
2. **E Step Estimate expected probability of hidden labels, Q, given current (latest) parameters and observed (unchanging) sequence**

$$Q = P(Labels | S, params^{t-1})$$

3. **M Step Choose new maximum likelihood parameters over probability distribution Q, given current probabilistic label assignments**

$$params^{t} = \arg\max_{params} E_{Q}\left[\log P(S, labels \mid params^{t-1})\right]$$

4. **Iterate**

P(S|Model) *guaranteed* to increase each iteration

EM for HMM Learning: Annotation ⇔ Parameters

Starting with our best guess of a model M, parameters θ:

Given $x = x_1 \ldots x_N$

for which the true $\pi = \pi_1 \ldots \pi_N$ is unknown,

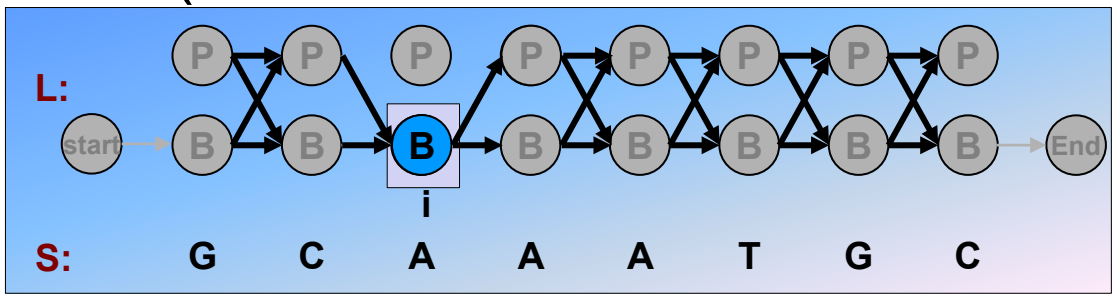
We can get to a provably more likely parameter set θ

Principle: EXPECTATION MAXIMIZATION

1. **Expected** annotations all-paths parse w/ current params (**E step**)
2. **Max**-likelihood params A_{kl}, E_k using this all-paths parse (**M step**)
3. Repeat 1 & 2, until convergence

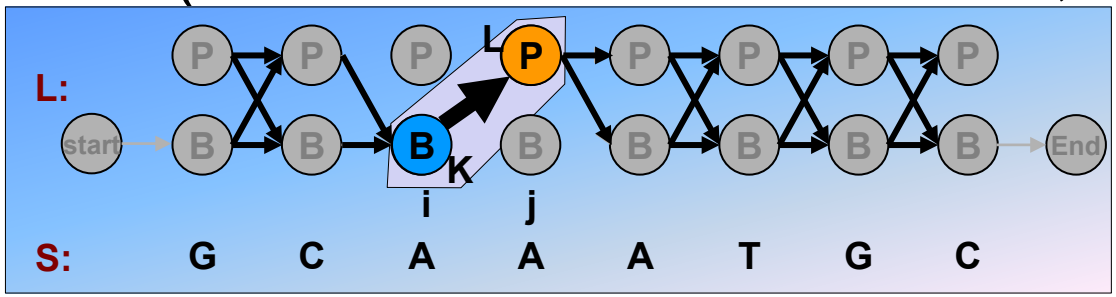
Max likelihood parameters | all-paths parse (M step)

(Sum over all emissions from k, at any time step i)



$$E_k(b) = [1/P(x)] \sum_{\{i \mid x_i = b\}} f_k(i)\, b_k(i)$$

(Sum over all k→l transitions, at any time step i)

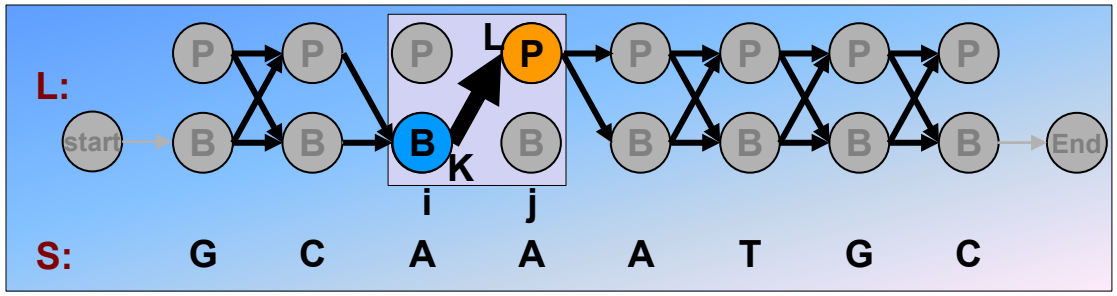


$f_k(i)$ a_{kl} $e_l(x_{i+1})$ $b_l(i+1)$

$$A_{kl} = \sum_i P(\pi_i = k, \pi_{i+1} = l \mid x, \theta) = \sum_i \frac{\quad}{P(x \mid \theta)}$$

Max likelihood parameters | all-paths parse (M step)

Derivation:
To estimate $\mathbf{A_{kl}}$:
At each position i:



Find probability transition k→l is used:

$$P(\pi_i = k, \pi_{i+1} = l \mid x) = [1/P(x)] \times P(\pi_i = k, \pi_{i+1} = l, x_1 \ldots x_N) = Q/P(x)$$

where $Q = P(x_1 \ldots x_i, \pi_i = k, \pi_{i+1} = l, x_{i+1} \ldots x_N) =$

$= P(\pi_{i+1} = l, x_{i+1} \ldots x_N \mid \pi_i = k)\, P(x_1 \ldots x_i, \pi_i = k) =$

$= P(\pi_{i+1} = l, x_{i+1} x_{i+2} \ldots x_N \mid \pi_i = k)\, f_k(i) =$

$= P(x_{i+2} \ldots x_N \mid \pi_{i+1} = l)\, P(x_{i+1} \mid \pi_{i+1} = l)\, P(\pi_{i+1} = l \mid \pi_i = k)\, f_k(i) =$

$= b_l(i+1)\, e_l(x_{i+1})\, a_{kl}\, f_k(i)$

So:
$$\mathbf{P(\pi_i = k, \pi_{i+1} = l \mid x, \theta) = \frac{f_k(i)\, a_{kl}\, e_l(x_{i+1})\, b_l(i+1)}{P(x \mid \theta)}}$$

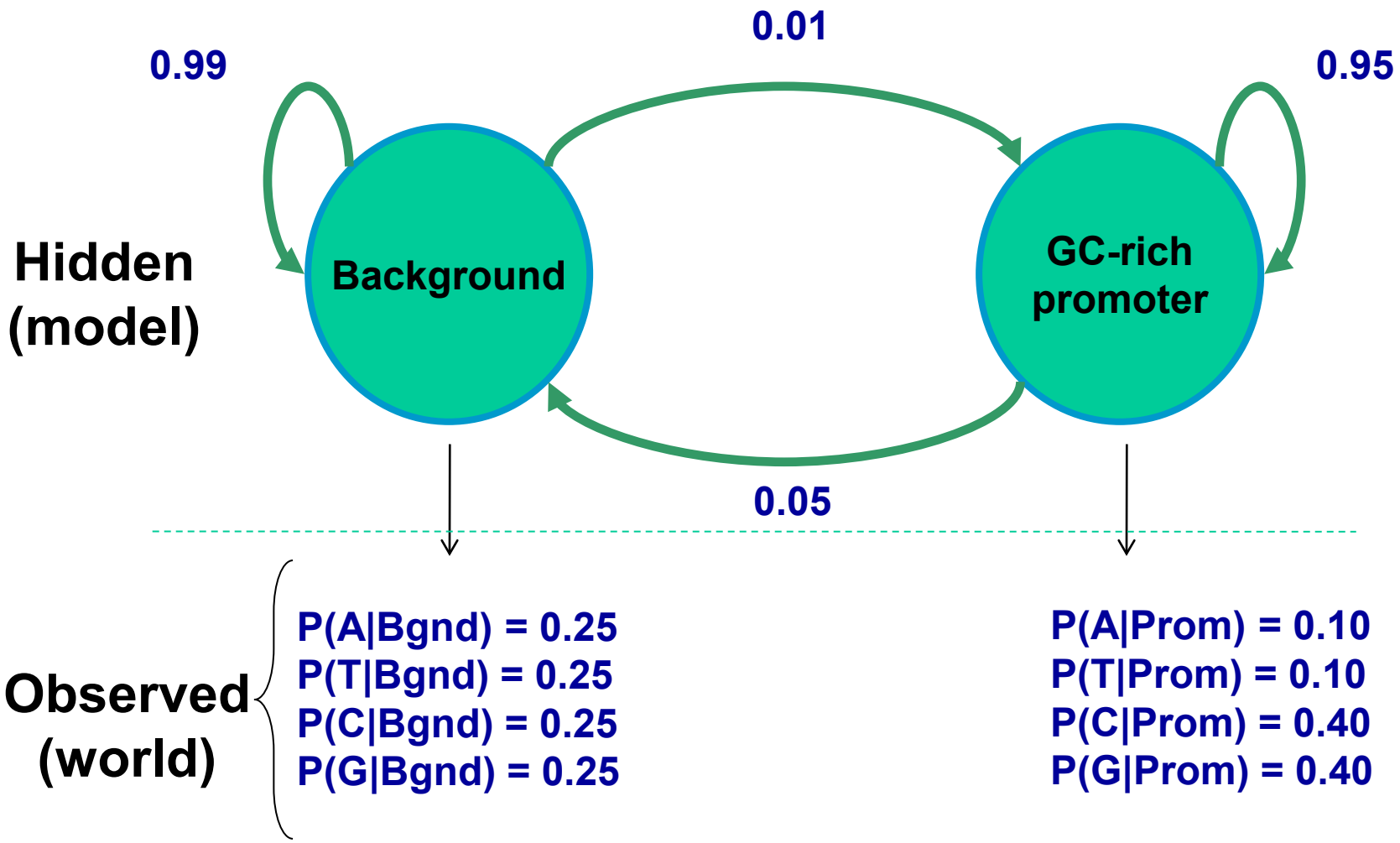
(For one such transition, at time step i→i+1)

	One path	All paths
Scoring	1. Scoring x, one path $P(x,\pi)$ ✓ Prob of a path, emissions	2. Scoring x, all paths $P(x) = \sum_{\pi} P(x,\pi)$ ✓ Prob of emissions, over all paths
Decoding	3. Viterbi decoding $\pi^* = \text{argmax}_{\pi} P(x,\pi)$ ✓ Most likely path	4. Posterior decoding $\pi\hat{} = \{\pi_i \mid \pi_i = \text{argmax}_k \sum_{\pi} P(\pi_i = k \mid x)\}$ ✓ Path containing the most likely state at any time point.
Learning	5. Supervised learning, given π $\Lambda^* = \text{argmax}_{\Lambda} P(x,\pi \mid \Lambda)$ ✓ 6. Unsupervised learning. $\Lambda^* = \text{argmax}_{\Lambda} \max_{\pi} P(x,\pi \mid \Lambda)$ ✓ Viterbi training, best path	6. Unsupervised learning $\Lambda^* = \text{argmax}_{\Lambda} \sum_{\pi} P(x,\pi \mid \Lambda)$ ✓ Baum-Welch training, over all paths

Goals for today: Computational Epigenomics

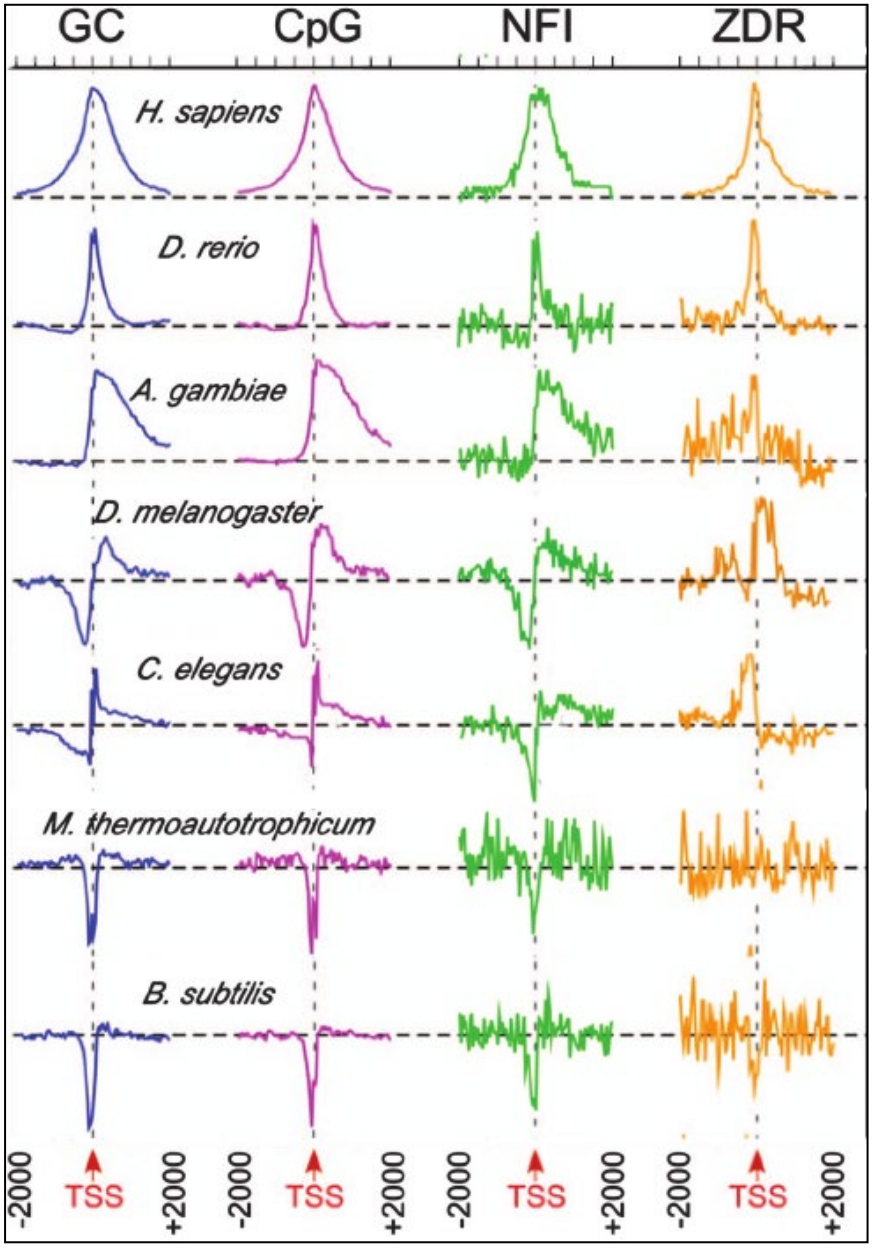
1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - HMM Foundations, Generating, Parsing, Decoding, Learning
   - Chromatin state characterization: Functional/positional enrichment
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

Detecting GC-rich regions: HMM architecture



Slide credit: Serafim Batzoglou

Example: Detecting GC-rich regions: motivation



Model genome as two states:

- P: promoter
- B: background

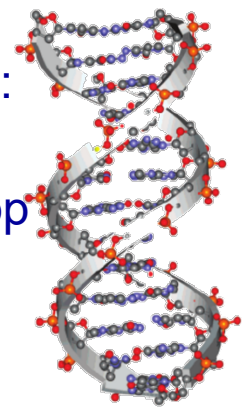
Model different nucleotide compositions

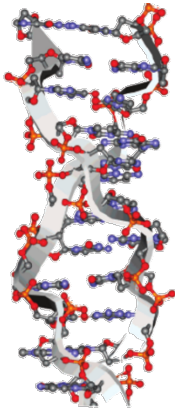
- Background:
  P(A) = P(T) = P(C) = P(G) = 25%
- Promoters
  P(A) = P(T) = 10%
  P(G) = P(C) = 40%

Note: generative model, P(A|promoter) etc
Then: reverse probabilities using Baye's rule

Model length distribution:

- Promoter: 20 bp
- Non-promoter: 100 bp


B-DNA


Z-DNA

Modified from: Khuu et al PNAS 2007; NFI: Nuclear Factor I; ZDRs: Z-DNA Forming Regions

HMMs are used broadly for genome annotation

Application	Detection of GC-rich regions	Detection of conserved regions	Detection of protein-coding exons	Detection of protein-coding conservation	Detection of protein-coding gene structures	Detection of chromatin states
Topology / Transitions	2 states, different nucleotide composition	2 states, different conservation levels	2 states, different tri-nucleotide composition	2 states, different evolutionary signatures	~20 states, different composition/ conservation, specific structure	40 states, different chromatin mark combinations
Hidden States / Annotation	GC-rich / AT-rich	Conserved / non-conserved	Coding exon / non-coding (intron or intergenic)	Coding exon / non-coding (intron or intergenic)	First/last/middle coding exon,UTRs, intron.4/3, intergenic, *(+/- strand)	Enhancer / promoter / transcribed / repressed / repetitive
Emissions / Observations	Nucleotides	Level of conservation	Triplets of nucleotides	Nucleotide triplets, conservation levels	Codons, nucleotides, splice sites, start/stop codons	Vector of chromatin mark frequencies

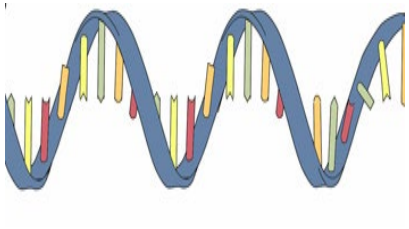
Examples of HMMs for genome annotation

Detection of conserved regions	Detection of GC-rich regions	Detection of CpG-rich regions	Detection of protein-coding exons	Detection of protein-coding gene structures	Detection of chromatin states	Detection of protein-coding conservation
2 states, different conservation levels	2 states, different nucleotide composition	8 states, 4 each +/-, different transition probabilities	2 states, different tri-nucleotide composition	~20 states, different composition/ conservation, specific structure	40 states, different chromatin mark combinations	2 states, different evolutionary signatures
Conserved / non-conserved	GC-rich / AT-rich	CpG-rich / CpG-poor	Coding exon / non-coding (intron or intergenic)	First/last/middle coding exon,UTRs, intron1/2/3, intergenic, *(+/- strand)	Enhancer / promoter / transcribed / repressed / repetitive	Coding exon / non-coding (intron or intergenic)
Level of conservation	Nucleotides	Di-Nucleotides	Triplets of nucleotides	Codons, nucleotides, splice sites, start/stop codons	Vector of chromatin mark frequencies	64x64 matrix of codon substitution frequencies
L2:alignmnt	L4:HMMs1	L5:HMMs2	L5:HMMs2	L5:HMMs2	L8:Epignmcs	L17:CompG

Genome Annotation

Genome Sequence

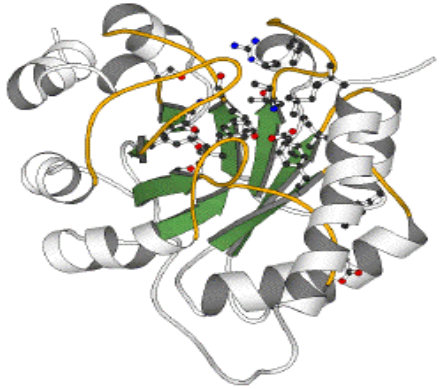
> HSCKIIBE, Human gene for casein kinase II subunit beta (EC 2.7.1.37).
ggggctgagatgtaaattagaggagctggagaggagtgcttcagagtttgggttgctttaagaaagggt
ggttccgaattctcccgtggttggagggccgaatgtgggaggagggaggataccagaggcagggaagga
gaacttgagctttactgacactgttctttttctagctgacgtgaagatgagcagctcagaggaggtgtc
ctggatttcctggttctgtgggctccgtggcaatgaattcttctgtgaagtgagttctcttcaacctcc
ctacttgccagcttcacatatcttcccaccagacgttccttcacatattccacttctacactgttctct
aaagcttttatgggagagagtgtaggtgaactagggagagacacaagtacttctgctgagttgggagtg
agaaacaagcacaacagatgcagttgtgttgatgataaggcatcacttagagcattttgcccaggtcaa
agatgaggattttgatatgggttccctcttggcttccatgtcctgacaggtggatgaagactacatcca
ggacaaatttaatcttactggactcaatgagcaggtccctcactatcgacaagctctagacatgatctt
ggacctggagcctggtgaggcaccctcagggttgttttgtgtgtgtgcgtgcactatttttctcttcaa
atctctattcacttgcctgaattttgccaaatttcctttggttctctgatttctttaaccccaaattca
tgctttattttgatcctccacctgactcttgtctagttttgtgacgtatatcacttgttctcatgtttt
tgcaagggtcagaagcccaggtttctgggtcccatgcccagatgttggatggggtaaggcccaaaagta
ggtgctaggcaaactgaatagcccgcagcccctggatatgggcagggcacctaggaaagctgaaaaaca
agtagttgcatttggccgggctgtggttcagatgaagaactggaagacaaccccaaccagagtgacctg
attgagcaggcagccgagatgctttatggattgatccacgcccgctacatcctaccaaccgtggcatc
gcccagatggtgaggcctctctgctcctacctgcctccttctgagcagtaagagacacaggttcctgca
gcaagaagtcatgtttaagccctgtttaaggaagctagctgagaagaggggaagaaccccagaacttgg
ccctgccctaatttggaagaaaggcaacacagaagtttgagagcccatctagtccagagaaggggggcct
ctggacagagttggaaggagtgccgacagagttggtatgggttgggctgcgaagggagttgcctcttct
ttacatctacctgccaacccctccattgtattcacctcagttggaaaagtaccagcaaggagactttg
gttactgtcctcgtgtgtactgtgagaaccagccaatgcttcccattGgtgagtgttgaagaagggaaa
ggaaagcaccgtgtggcagtcttatgggaaggagttggggctcaacacattggagcctgagtcctgagg
ggaggttaggtaggaatagggggatacctggcctgctgagtctggctgtctcccaggcctttcagacat
cccaggtgaagccatggtgaagctctactgccccaagtgcatggatgtgtacacacccaagtcatcaag
acaccatcacacggatggcgcctacttcggcactggtttccctcacatgctcttcatggtgcatcccga
gtaccggcccaagagacctgccaaccagtttgtgcccaggtagggagcagggagagtcattaagggtca
aaggaaaggcccaagatcccccagagagggggaggacagggcatggccctttcttgaggtctgcttctcc
cagaatcagggcatctccctgctgagtgactgtgggaaagttatttgattatctgtgcttgagttacct
tattgtagaatgttcttgagctgagaagttgggaaccacgaggctttagctctgagcaggtccatagag
gagctcaggtgggagaggtgggaatgcaggtgactggcagggcctggatggggctcatgctgctgcctct
ctgacctctgccctggcctaggctctacggtttcaagatccatccgatggcctaccagctgcagctcca
agccgccagcaacttcaagagcccagtcaagacgattcgctgattccctccccccacctgtcctgcagtc
tttgtcttttccttttcttttttgccacccctttcaggaaccctgtatggtttttagtttaaattaaagga
gtcgttatcgtggtgggaatatgaaataaagtagaagaaaaggccatgagctagtctgctggtgcttgc
tggaaggggtggagcgtggccatggaaatcgggctccacggcccagggatgg



RNA

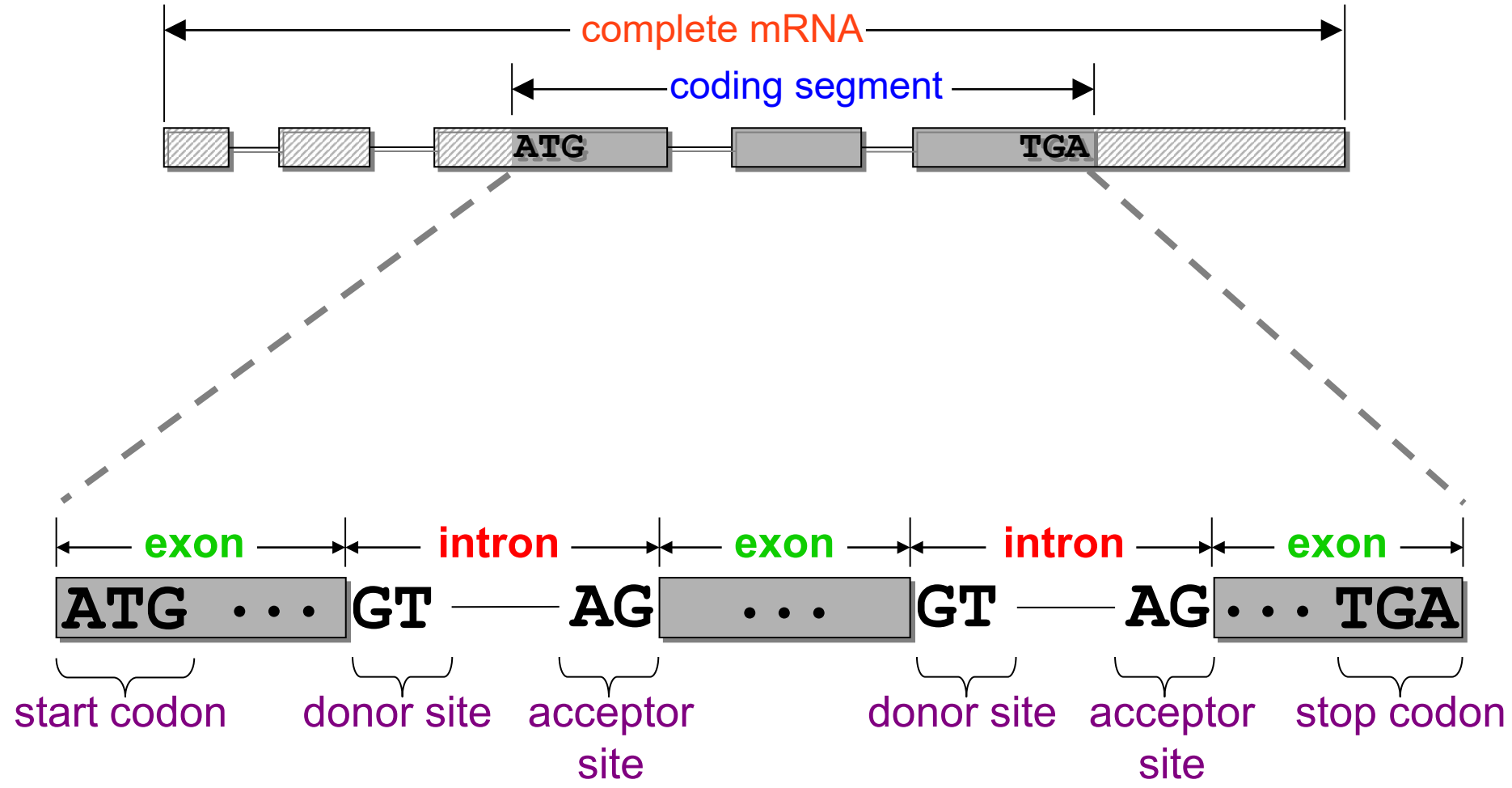
Transcription

Translation

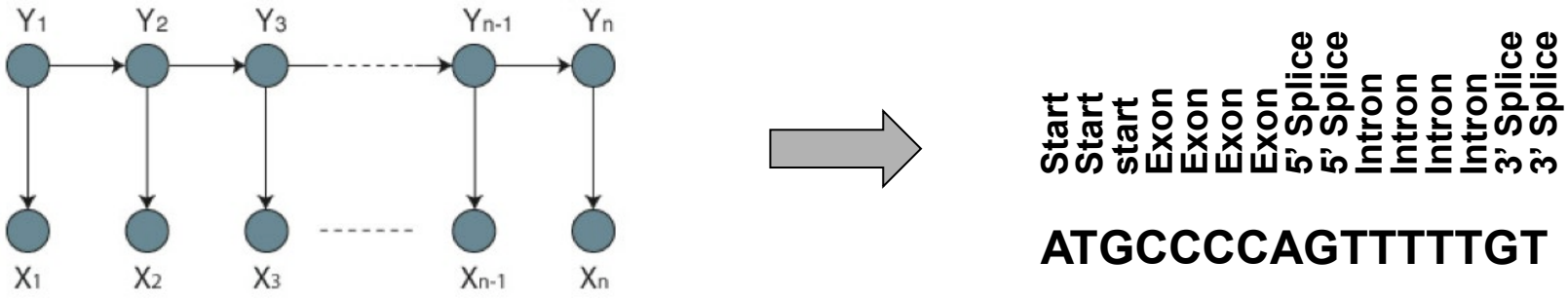


Protein

Eukaryotic Gene Structure



Gene Prediction with HMM

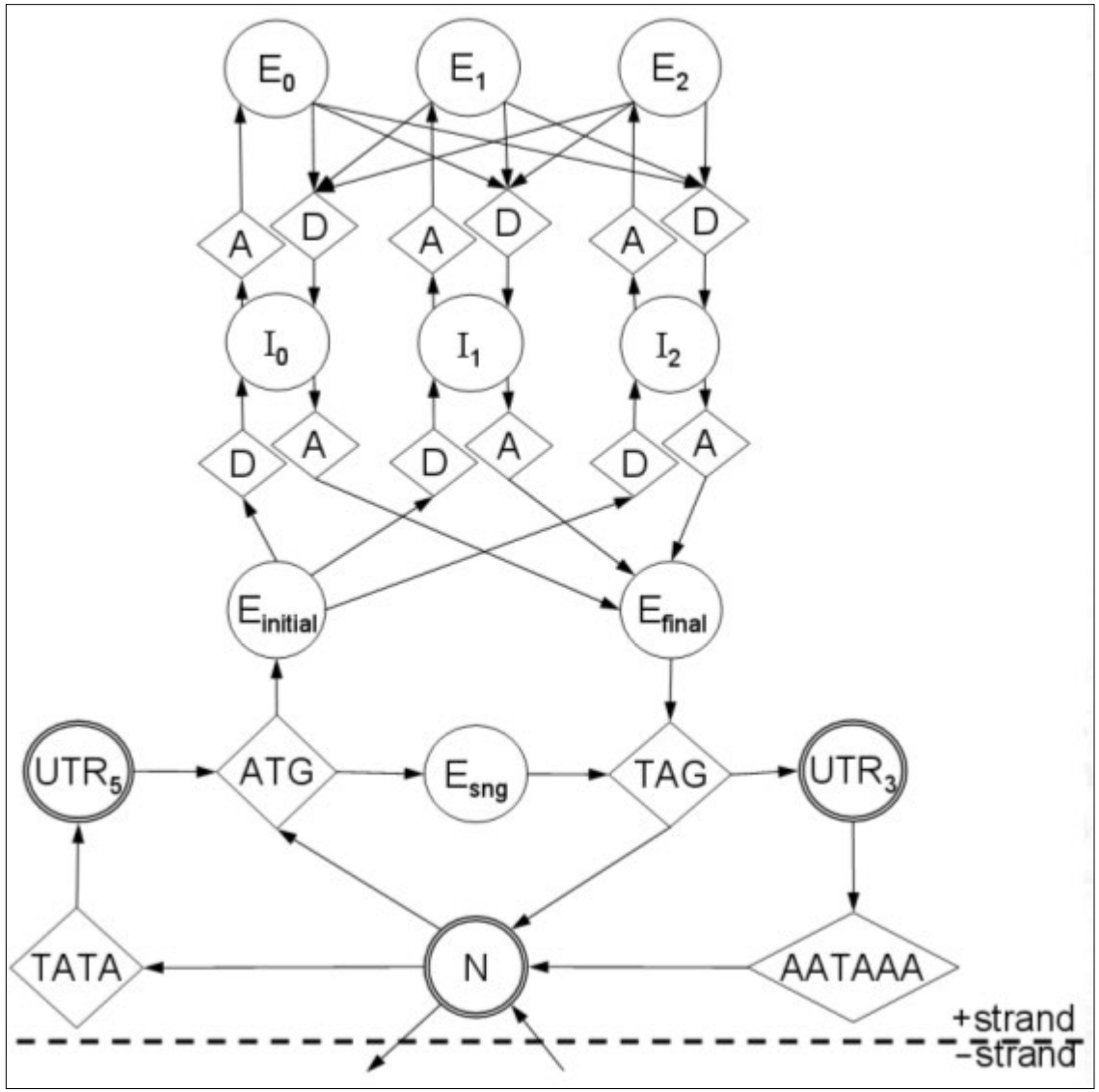


Model of joint distribution P(Y,X) = P(Labels,Seq)

For gene prediction, we are given X…

How do we select a Y efficiently?

HMM architecture matters: Protein-coding genes

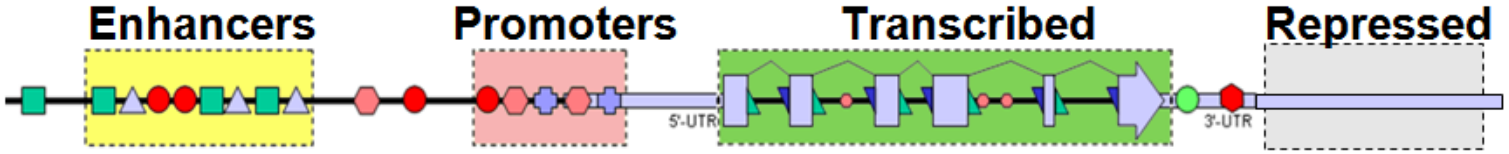


- Gene vs. Intergenic
- Start & Stop in/out
- UTR: 5' and 3' end
- Exons, Introns
- Remembering frame
  - E0,E1,E2
  - I0,I1,I2
- Sequence patterns to transition between states:
  - ATG, TAG, Acceptor/Donor, TATA, AATAA

Examples of HMMs for genome annotation

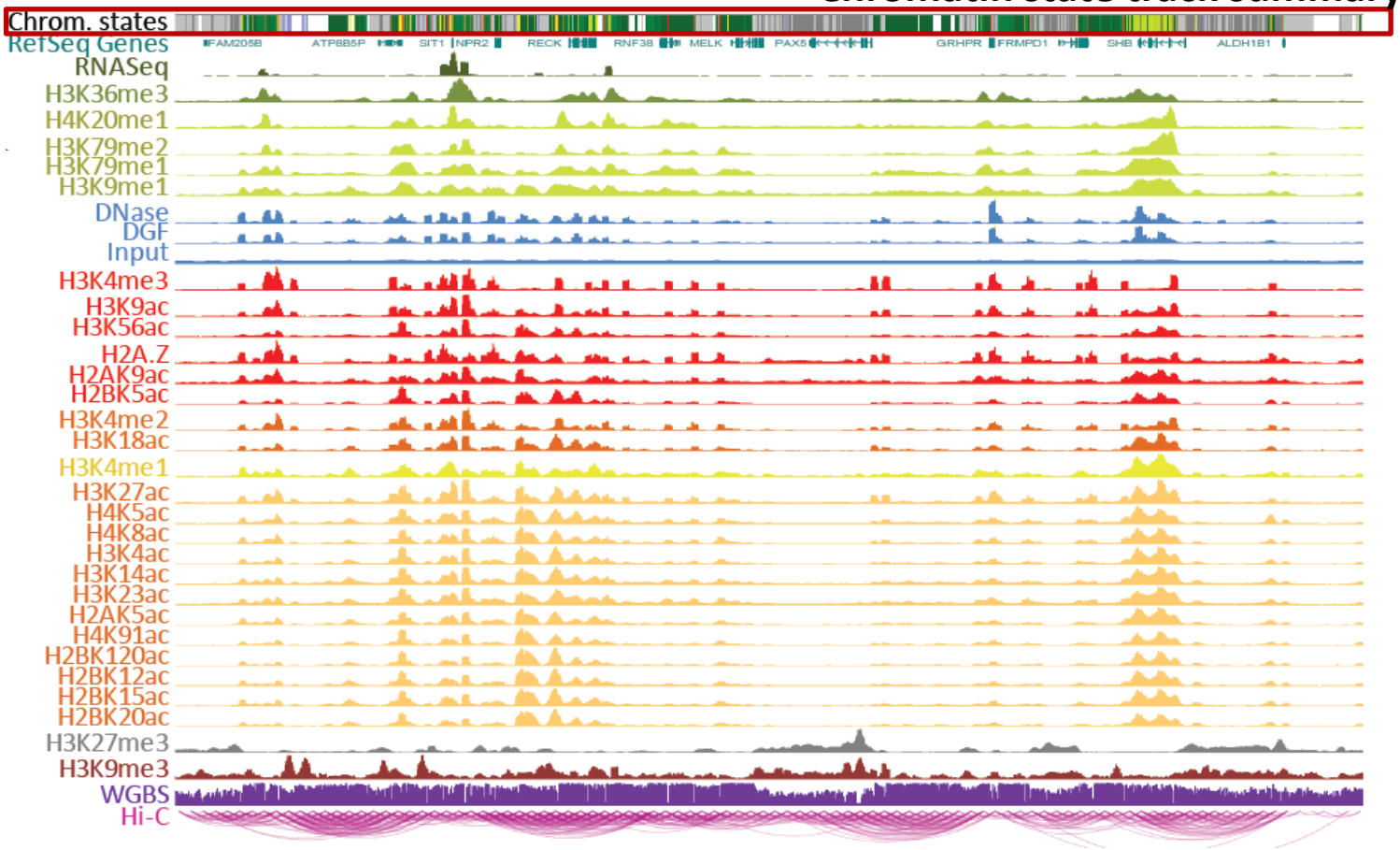
Detection of conserved regions	Detection of GC-rich regions	Detection of CpG-rich regions	Detection of protein-coding exons	Detection of protein-coding gene structures	Detection of chromatin states	Detection of protein-coding conservation
2 states, different conservation levels	2 states, different nucleotide composition	8 states, 4 each +/-, different transition probabilities	2 states, different tri-nucleotide composition	~20 states, different composition/ conservation, specific structure	40 states, different chromatin mark combinations	2 states, different evolutionary signatures
Conserved / non-conserved	GC-rich / AT-rich	CpG-rich / CpG-poor	Coding exon / non-coding (intron or intergenic)	First/last/middle coding exon,UTRs, intron1/2/3, intergenic, *(+/- strand)	Enhancer / promoter / transcribed / repressed / repetitive	Coding exon / non-coding (intron or intergenic)
Level of conservation	Nucleotides	Di-Nucleotides	Triplets of nucleotides	Codons, nucleotides, splice sites, start/stop codons	Vector of chromatin mark frequencies	64x64 matrix of codon substitution frequencies
L2:alignmnt	L4:HMMs1	L5:HMMs2	L5:HMMs2	L5:HMMs2	L8:Epignmcs	L17:CompG

Summarize multiple marks into chromatin states



Chromatin state track summary

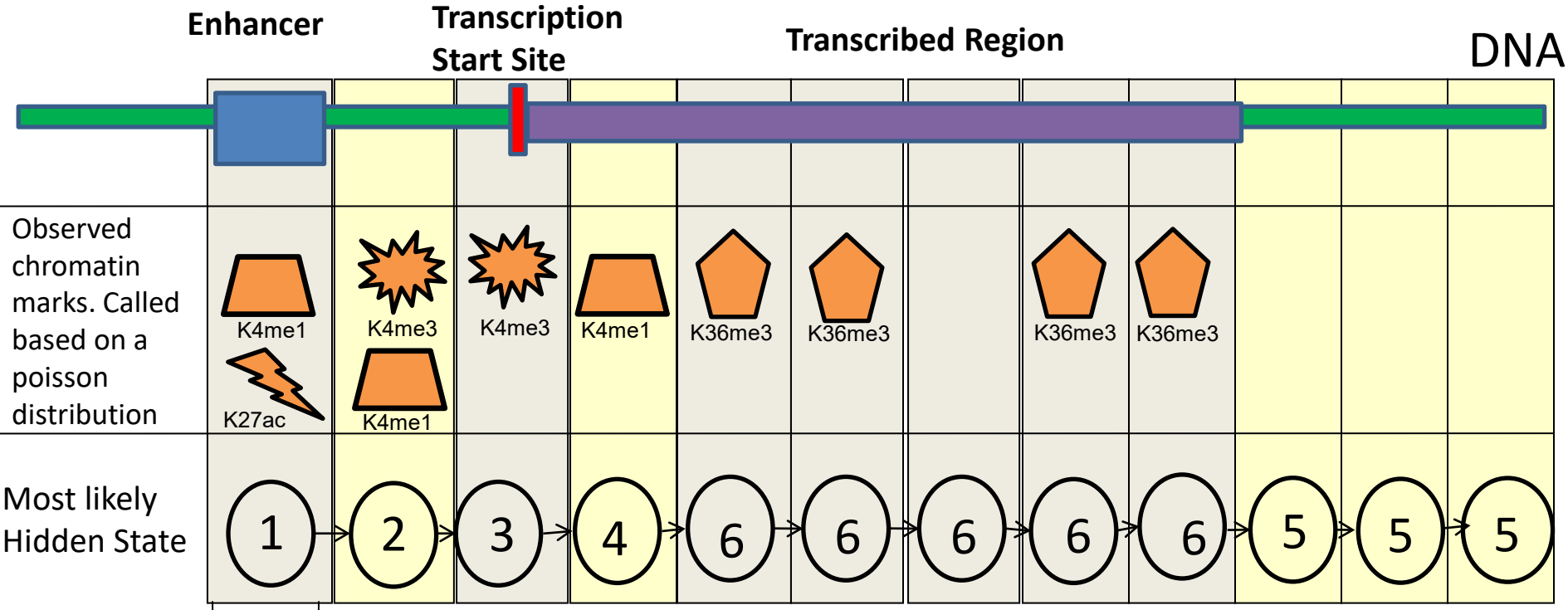
30+ epigenomics marks



WashU Epigenome Browser

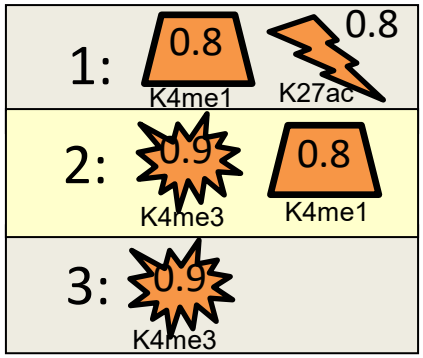
ChromHMM: multi-variate hidden Markov model

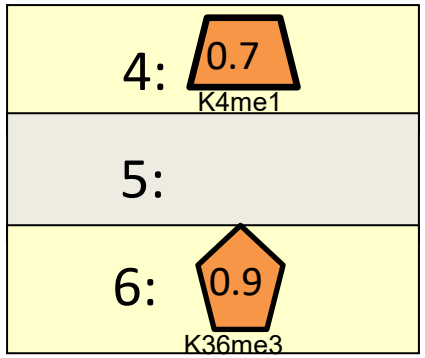
Multivariate HMM for Chromatin States



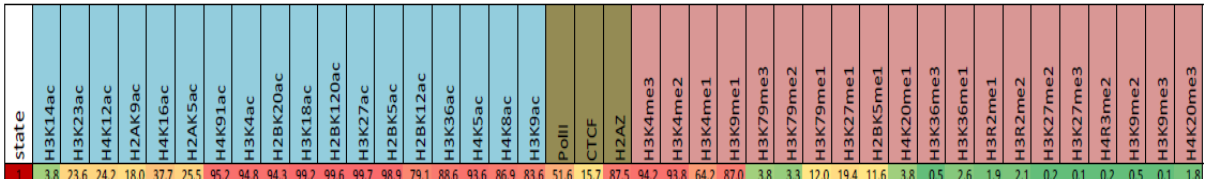
200bp intervals

High Probability Chromatin Marks in State





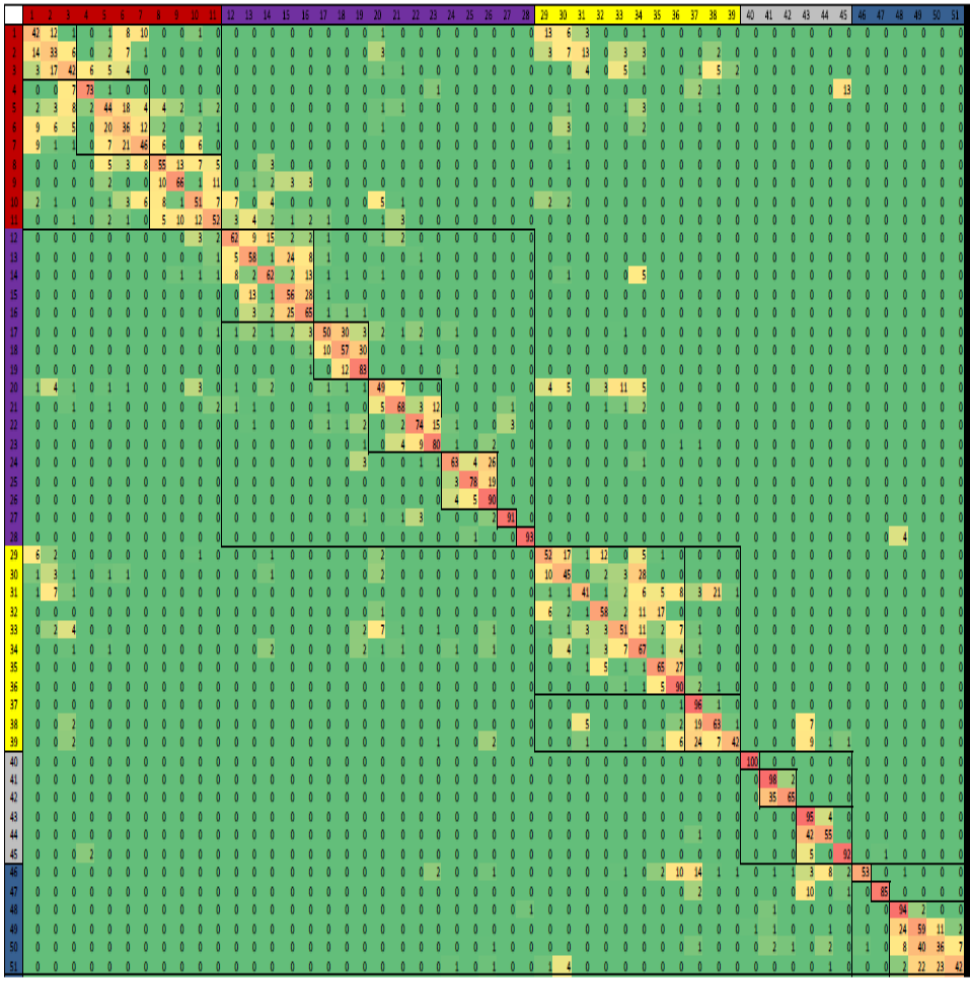
All probabilities are learned from the data



Ernst and Kellis

Nature Biotech 2010

Chromatin State: Emission & Transition Matrices



- **Emission matrix $e_k(x_i)$**
  - Multi-variate HMM
  - Emits vector of values
- **Transition matrix a_{kl}**
  - Learn spatial relationships
  - No a-priori ‘gene’ structure

Ernst and Kellis, Nature Biotech 2010, Nature 2011, Nature Methods 2012

Design Choice

- How to model the emission distribution
  - Model the signal directly
  - Locally binarize the data
- For *M* input marks each state *k* has a vector of $(p_{k1},\ldots,p_{kM})$ of parameters for independent Bernoulli random variables which determine the emission probability for an observed combination of marks

Data Binarization

- Leads to biologically interpretable models that can be robustly learned
- Let c_{ij} be the number of reads for mark *i.* mapping to bin *j.* λ_i be the average number of reads mapping to a bin for modification *i*. The input for feature *i* becomes *'1'* if

$$P(X>c_{ij})<10^{-4}$$

where *X* is a Poisson random variable with mean λ_i

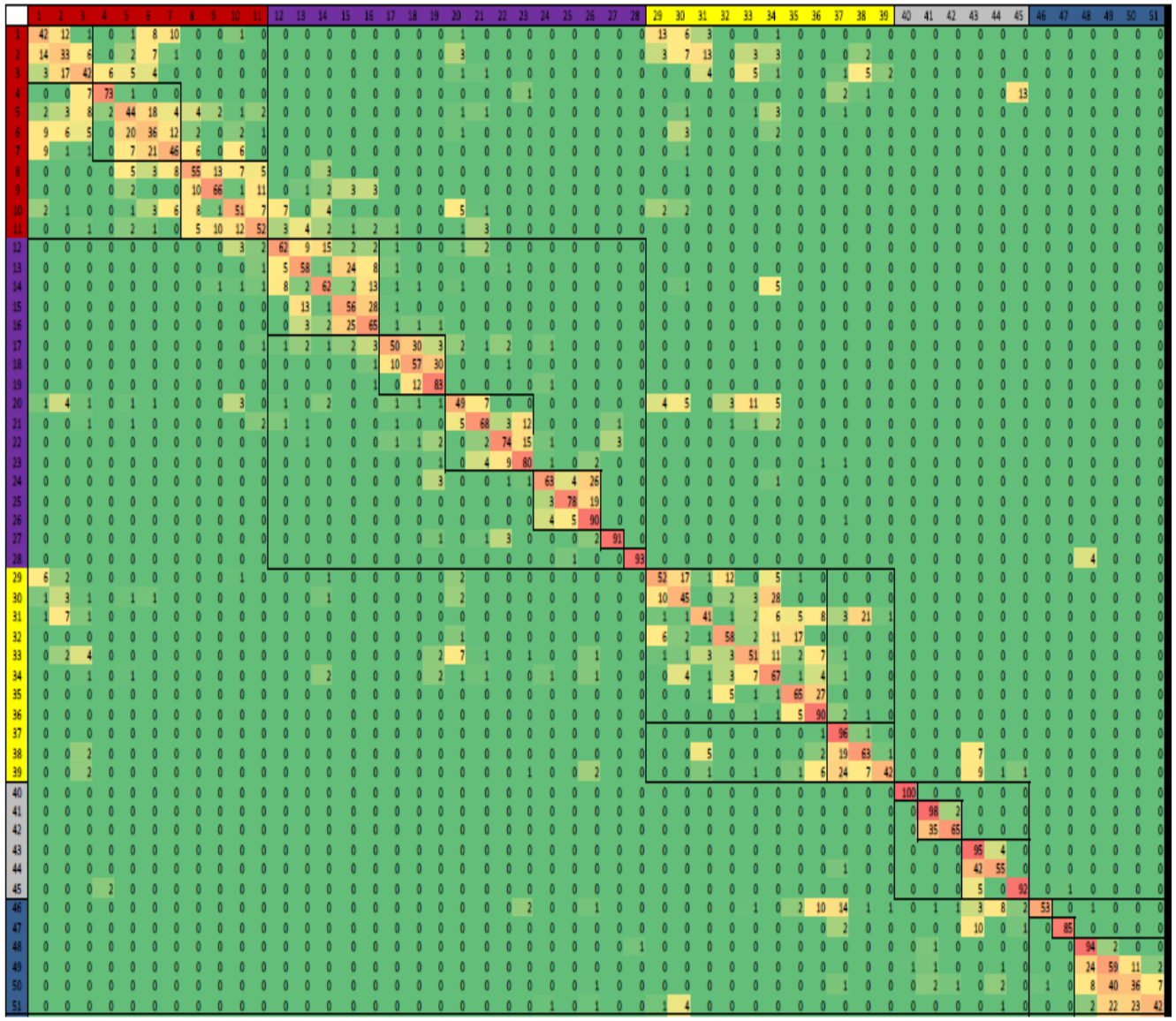
Emission Parameter Matrix $e_k(\vec{x}_i)$

state	H3K14ac	H3K23ac	H4K12ac	H2AK9ac	H4K16ac	H2AK5ac	H4K91ac	H3K4ac	H2BK20ac	H3K18ac	H2BK120ac	H3K27ac	H2BK5ac	H2BK12ac	H3K36ac	H4K5ac	H4K8ac	H3K9ac	PolII	CTCF	H2AZ	H3K4me3	H3K4me2	H3K4me1	H3K9me1	H3K79me3	H3K79me2	H3K79me1	H3K27me1	H2BK5me1	H4K20me1	H3K36me3	H3K36me1	H3R2me1	H3R2me2	H3K27me2	H3K27me3	H4R3me2	H3K9me2	H3K9me3	H4K20me3
1	3.8	23.6	24.2	18.0	37.7	25.5	95.2	94.8	94.3	99.2	99.6	99.7	98.9	79.1	88.6	93.6	86.9	83.6	51.6	15.7	87.5	94.2	93.8	64.2	87.0	3.8	3.3	12.0	19.4	11.6	3.8	0.5	2.6	1.9	2.1	0.2	0.1	0.2	0.5	0.1	1.8
2	2.5	17.5	9.2	3.2	5.9	6.3	44.6	44.4	47.0	73.2	74.1	85.9	71.2	22.1	33.5	61.9	63.3	35.4	18.1	10.9	91.2	86.7	90.4	66.9	78.3	2.4	2.2	7.9	17.6	8.7	2.3	0.6	2.2	1.7	1.5	0.4	0.5	0.2	0.4	0.1	1.4
3	0.5	5.8	1.8	1.0	0.9	1.2	12.3	9.5	8.8	22.6	21.3	22.8	12.1	2.2	4.2	8.4	12.8	7.1	11.2	16.3	77.1	93.9	80.3	45.6	74.2	1.5	1.3	4.6	4.3	7.0	8.8	0.2	1.4	2.1	1.5	0.2	2.1	0.1	0.1	0.1	1.2
4	0.1	0.8	0.1	0.4	0.3	0.2	1.5	0.9	0.7	2.1	2.1	0.5	0.2	0.1	0.1	0.1	0.3	0.3	1.9	6.2	19.0	77.9	20.8	21.1	26.4	0.3	0.0	0.1	0.0	1.3	14.9	0.1	0.2	1.4	0.9	0.0	10.2	0.1	0.1	0.4	1.3
5	0.0	0.2	0.8	1.3	2.0	0.4	26.6	12.6	6.7	15.8	23.1	26.8	24.3	1.5	4.3	0.9	1.4	7.7	53.5	20.6	21.8	87.0	11.2	2.7	15.7	6.3	3.8	2.5	0.0	5.5	14.2	0.0	0.3	0.2	0.6	0.0	0.0	0.0	0.1	0.0	1.5
6	0.1	1.8	3.6	6.9	6.1	1.9	74.5	63.5	53.0	75.7	84.3	89.4	86.7	20.8	41.5	20.5	21.6	62.7	69.2	25.5	61.2	98.3	37.4	7.1	40.3	5.3	2.7	5.6	0.6	6.0	11.6	0.0	0.5	0.4	0.9	0.0	0.0	0.0	0.0	0.1	1.6
7	1.2	8.7	20.6	43.0	53.7	9.8	98.7	98.6	95.7	99.4	99.9	99.9	99.9	76.5	93.3	81.8	76.6	99.2	88.0	26.9	77.1	99.7	38.3	2.2	37.9	32.1	24.9	14.0	2.6	6.5	16.2	0.1	1.0	0.5	1.2	0.0	0.0	0.0	0.0	0.1	1.9
8	1.2	12.7	5.2	11.9	5.6	6.8	56.9	56.1	37.5	52.4	69.8	89.1	85.5	21.3	24.8	16.7	10.3	60.8	62.1	12.0	31.4	96.7	51.7	14.9	45.3	86.3	80.1	23.8	1.7	6.7	42.2	4.5	0.4	1.1	0.9	0.0	0.0	0.0	0.0	0.0	0.5
9	0.5	7.2	3.0	1.1	0.5	2.1	4.0	7.4	2.4	2.5	11.6	35.3	28.0	2.7	2.8	2.0	1.8	8.6	34.7	4.2	4.5	79.2	41.5	23.8	36.1	86.0	82.6	12.0	1.9	6.7	43.5	7.4	0.2	0.6	0.7	0.0	0.0	0.0	0.0	0.2	0.4
10	4.1	24.8	13.1	17.5	24.4	37.0	90.4	88.6	82.0	89.8	95.5	97.0	95.1	54.0	56.4	67.2	45.7	55.7	46.6	10.2	40.8	84.6	92.3	91.4	92.8	67.1	67.2	63.4	29.2	53.8	65.2	4.5	6.8	5.7	3.4	0.1	0.0	0.3	0.1	0.0	1.0
11	1.6	21.0	3.9	3.8	3.2	8.4	28.0	26.1	14.5	22.6	37.6	56.8	47.4	6.0	6.4	13.5	8.8	18.4	30.9	6.9	20.1	92.4	93.5	94.3	94.5	73.8	74.2	55.1	24.0	57.2	79.9	8.9	6.6	4.4	2.5	0.1	0.0	0.2	0.1	0.1	0.7
12	3.6	17.0	8.9	2.2	14.1	34.9	60.3	51.0	38.8	35.6	56.6	53.3	55.3	11.9	11.5	30.0	15.4	1.7	18.0	2.7	0.7	5.4	58.2	96.0	77.8	87.4	87.0	76.3	41.6	79.8	82.2	13.4	3.1	6.4	3.6	0.3	0.0	0.7	0.2	0.1	0.4
13	1.2	10.8	3.6	0.7	2.5	7.4	9.1	6.5	2.6	2.5	6.5	7.7	5.5	0.5	1.0	5.5	3.3	0.3	10.2	1.5	0.0	2.4	56.2	83.7	82.9	92.7	92.6	64.4	38.2	80.0	89.8	12.0	2.4	3.3	2.1	0.3	0.0	0.4	0.2	0.1	0.3
14	0.7	5.3	7.9	1.0	2.4	18.0	19.8	20.7	14.6	7.9	24.5	20.1	21.8	6.7	6.6	11.3	8.6	0.3	6.9	1.7	2.1	1.3	8.0	33.0	16.3	61.8	62.1	37.9	9.7	14.3	18.3	9.7	0.2	1.7	1.2	0.2	0.0	0.1	0.3	0.4	1.0
15	0.2	1.9	2.9	0.3	0.3	1.5	0.8	1.3	0.3	0.2	1.2	1.5	1.2	0.2	0.4	0.7	1.2	0.1	5.0	0.7	0.0	0.2	11.0	17.3	29.3	84.7	82.2	33.1	8.0	26.4	56.2	5.2	0.2	0.7	0.9	0.1	0.0	0.1	0.1	0.6	0.2
16	0.0	0.3	0.4	0.1	0.1	0.5	0.4	0.6	0.2	0.1	0.5	0.4	0.3	0.1	0.2	0.2	0.3	0.0	0.6	0.3	0.1	0.1	1.2	2.8	3.9	29.0	25.2	8.5	0.7	1.3	7.9	1.2	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.4	0.1
17	1.2	9.8	2.8	0.9	2.4	7.8	6.8	6.1	2.3	4.0	3.5	8.4	3.5	0.3	1.0	9.3	6.6	0.5	3.5	1.1	0.4	1.0	52.3	68.9	83.7	22.7	23.5	61.2	48.3	64.7	57.3	21.8	2.8	4.9	2.0	1.0	0.1	0.5	0.5	0.1	0.5
18	0.3	2.6	2.1	0.4	0.5	2.4	1.4	1.6	0.5	0.6	0.9	1.6	0.7	0.2	0.5	1.9	1.9	0.1	1.6	0.7	0.1	0.1	10.4	9.7	29.1	15.0	13.5	34.0	14.9	19.9	21.4	10.6	0.5	1.2	0.9	0.5	0.1	0.1	0.3	0.2	0.2
19	0.1	0.3	0.5	0.2	0.1	0.5	0.1	0.3	0.0	0.0	0.1	0.2	0.1	0.1	0.1	0.1	0.3	0.0	0.4	0.3	0.0	0.0	0.5	0.2	2.0	4.4	2.9	7.3	1.0	1.0	2.3	1.4	0.0	0.2	0.3	0.1	0.0	0.0	0.1	0.4	0.1
20	2.5	10.7	5.4	3.1	9.9	26.2	58.2	48.8	41.7	49.3	54.8	57.1	51.5	13.0	14.1	31.6	21.7	4.0	14.5	6.7	15.6	20.9	56.8	97.1	70.5	5.4	5.6	33.8	31.1	52.6	38.4	7.4	3.4	6.9	3.8	0.5	0.1	0.7	0.3	0.0	1.0
21	0.2	0.8	2.0	1.3	7.2	11.3	32.3	15.4	11.5	5.7	18.6	8.4	12.4	2.1	1.2	3.2	2.3	0.5	15.1	6.5	0.6	4.7	17.0	68.7	33.3	8.7	6.4	37.2	9.6	65.4	87.7	9.2	1.3	7.1	5.3	0.1	0.2	1.4	0.0	0.0	0.8
22	0.1	0.1	1.1	0.6	6.2	2.4	7.8	1.8	0.6	0.1	1.5	0.8	1.5	0.1	0.1	0.7	0.7	0.1	8.5	1.0	0.0	0.0	5.3	8.4	14.6	15.5	9.1	50.0	9.6	77.5	94.1	22.9	0.5	5.6	4.6	0.1	0.0	1.4	0.0	0.1	0.7
23	0.0	0.1	0.1	0.4	2.0	1.6	4.9	1.2	0.5	0.2	0.9	0.2	0.3	0.0	0.1	0.1	0.1	0.0	1.4	1.1	0.0	0.0	0.5	2.6	1.4	0.5	0.1	5.4	1.3	19.4	36.8	2.5	0.1	1.4	1.5	0.1	0.2	0.3	0.0	0.0	0.2
24	0.3	1.8	2.1	0.9	3.2	3.8	4.0	2.3	0.9	0.6	1.1	3.6	1.9	0.1	0.4	3.7	3.9	0.3	2.2	1.0	0.1	0.1	6.0	4.5	17.2	1.3	0.2	15.6	29.8	29.3	7.1	49.5	1.3	4.7	2.2	1.0	0.0	0.6	0.3	0.1	0.3
25	0.1	0.3	0.8	0.5	0.5	0.6	0.3	0.3	0.1	0.0	0.1	0.5	0.3	0.1	0.1	0.4	0.7	0.1	0.8	0.4	0.0	0.1	0.4	0.1	1.4	0.8	0.1	2.0	8.8	2.8	0.4	60.1	0.4	1.1	0.9	0.6	0.1	0.2	0.3	1.7	0.3
26	0.1	0.2	0.6	0.2	0.2	0.2	0.1	0.2	0.0	0.0	0.1	0.3	0.2	0.0	0.1	0.2	0.4	0.0	0.6	0.3	0.0	0.0	0.3	0.1	0.8	0.2	0.0	0.8	2.3	0.9	0.2	4.2	0.1	0.2	0.4	0.1	0.0	0.0	0.1	0.1	0.0
27	0.0	0.5	4.4	0.4	1.3	1.2	1.3	0.7	0.3	0.1	0.7	2.1	2.4	0.1	0.1	1.6	2.7	0.1	21.7	1.4	0.0	0.0	1.1	1.1	3.5	4.6	1.2	9.9	3.0	7.1	31.7	34.0	0.2	0.7	1.1	0.0	0.0	0.1	0.0	0.3	0.1
28	0.0	0.0	0.2	0.3	0.0	0.4	0.1	0.2	0.2	0.0	0.3	0.1	0.1	0.1	0.0	0.1	0.1	0.0	1.3	0.3	0.0	3.3	0.2	0.5	0.5	13.8	3.4	5.7	1.3	3.0	1.5	68.8	0.1	2.7	2.7	0.3	0.2	0.8	0.4	43.0	74.9
29	4.6	8.4	11.1	6.6	20.4	54.7	88.5	88.1	89.6	86.6	95.3	86.3	86.9	68.1	60.2	67.6	42.6	10.4	13.6	3.8	24.2	7.6	24.7	84.4	25.8	4.9	5.6	14.9	17.6	21.7	5.0	2.9	0.9	4.8	3.1	0.4	0.1	0.4	0.8	0.2	4.4
30	1.2	3.6	8.4	2.4	2.6	13.5	24.9	34.5	34.4	24.6	52.1	60.1	64.6	27.4	23.8	20.7	16.7	3.2	9.1	2.9	12.0	6.9	8.8	35.8	6.5	2.8	2.9	4.4	3.7	3.0	1.0	2.8	0.1	0.9	1.0	0.0	0.1	0.1	0.4	0.6	3.4
31	1.7	7.6	4.7	1.7	2.4	6.8	17.7	18.4	21.5	37.9	31.6	35.4	20.1	9.3	13.0	48.3	57.7	3.3	1.2	5.9	69.0	10.5	16.1	41.8	11.0	0.6	0.5	1.2	10.7	2.2	0.0	1.1	1.4	2.0	1.3	1.1	0.5	0.2	0.8	0.2	1.1
32	1.4	1.6	1.0	1.9	8.1	50.1	72.4	57.2	60.9	42.5	57.6	12.1	14.8	23.6	19.5	16.0	5.4	0.3	1.6	1.7	3.6	0.1	2.1	41.4	5.0	1.8	1.7	12.2	10.9	17.1	4.1	4.1	0.9	6.5	3.0	1.3	0.3	0.6	0.5	0.3	3.8
33	1.2	4.6	0.9	0.9	1.2	9.8	12.9	10.3	7.0	12.7	10.5	9.8	5.7	1.3	2.1	6.3	4.1	0.4	2.6	4.3	8.6	1.5	16.3	77.1	23.5	0.7	0.5	5.1	10.3	10.8	3.9	1.8	1.1	2.9	1.6	0.7	0.3	0.2	0.4	0.1	0.4
34	0.2	0.7	2.0	0.5	0.7	5.4	7.2	6.2	4.9	1.2	8.9	5.7	6.2	2.9	2.2	2.3	2.3	0.1	1.5	0.9	1.0	0.2	0.6	6.3	1.0	1.9	1.7	5.2	2.6	2.2	0.6	1.4	0.0	0.5	0.6	0.0	0.0	0.0	0.1	0.2	1.0
35	0.6	1.1	0.1	1.1	1.3	32.1	29.9	20.0	19.6	16.4	11.8	1.1	0.5	3.2	5.4	4.0	1.5	0.2	0.3	1.0	2.0	0.1	0.3	2.6	0.6	0.4	0.2	1.9	2.6	2.2	0.5	1.7	0.9	4.2	2.0	2.6	0.8	0.3	0.5	0.2	0.6
36	0.2	0.3	0.0	0.5	0.2	5.7	2.5	2.1	1.0	1.6	0.8	0.1	0.1	0.3	0.6	0.3	0.1	0.1	0.1	0.3	0.3	0.0	0.1	0.3	0.1	0.2	0.0	0.2	0.4	0.4	0.1	0.4	0.3	1.3	0.8	0.8	0.5	0.1	0.3	0.3	0.1
37	0.1	0.2	0.0	0.2	0.1	0.7	0.1	0.2	0.0	0.1	0.1	0.1	0.0	0.0	0.1	0.1	0.1	0.0	0.1	0.2	0.2	0.0	0.0	0.1	0.1	0.1	0.0	0.0	0.3	0.1	0.5	0.4	0.2	1.2	0.6	0.2	0.1	0.1	0.2	0.1	0.1
38	0.4	2.5	0.6	0.8	0.2	1.9	1.0	1.9	1.4	3.8	2.2	2.4	0.9	0.8	1.4	4.8	9.2	0.4	0.1	0.7	33.0	1.3	1.5	5.9	0.9	0.2	0.1	0.2	1.7	0.3	0.0	0.3	0.8	1.0	0.9	1.0	1.1	0.2	0.6	0.2	0.3
39	0.1	0.4	0.5	0.6	0.1	1.2	0.7	0.8	1.1	0.9	1.0	0.4	0.2	0.3	0.4	1.2	2.0	0.1	4.1	86.2	12.3	1.4	1.6	5.0	1.2	0.2	0.0	0.3	0.6	0.7	0.7	0.2	0.8	2.5	1.2	0.2	0.8	0.1	0.0	0.2	0.6
40	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0
41	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.1	0.1	0.5	0.3	0.1	0.1	1.0	2.0	0.4
42	0.1	0.1	0.0	0.8	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.2	0.1	0.6	0.6	0.1	0.0	0.1	0.1	0.1	0.0	0.0	1.8	0.5	0.0	0.5	1.3	2.0	2.3	8.7	1.2	0.8	5.6	2.3	4.4
43	0.0	0.2	0.0	0.3	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.3	0.3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.3	0.4	1.0	0.1	0.5	0.5	0.1
44	0.6	1.1	0.0	1.2	0.2	0.7	0.1	0.4	0.0	0.4	0.1	0.1	0.0	0.0	0.2	0.1	0.1	0.3	0.1	0.6	2.0	0.2	0.2	0.3	0.3	0.1	0.0	0.0	0.3	1.1	0.0	1.6	3.3	5.7	3.1	9.0	9.9	0.9	2.5	0.4	1.0
45	0.1	0.7	0.0	0.5	0.0	0.4	0.1	0.1	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.1	0.6	0.6	0.4	0.1	0.4	0.1	0.1	0.0	0.0	0.0	0.1	0.2	0.3	0.5	1.0	0.9	0.4	16.4	0.2	0.3	1.1	0.4
46	1.6	0.7	0.4	1.4	4.2	6.1	5.2	3.6	0.8	4.5	2.4	0.7	0.6	0.4	0.5	0.7	0.4	0.5	0.6	2.0	1.1	2.7	2.1	28.4	5.5	0.4	0.1	1.1	8.4	63.8	16.5	5.9	17.8	48.9	25.0	9.8	7.8	11.9	3.7	1.6	17.0
47	0.0	0.1	0.0	0.2	0.0	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.0	0.6	0.0	0.0	0.0	0.5	0.0	0.0	0.0	0.0	1.9	0.1	0.4	0.7	0.3	3.1	0.1	0.7	32.3	1.8
48	0.0	0.0	0.0	0.2	0.0	0.1	0.0	0.1	0.1	0.0	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.0	0.1	0.2	0.2	0.1	0.0	0.1	0.0	0.3	0.0	0.1	0.2	0.1	0.0	1.0	0.1	0.5	0.9	0.6	0.3	0.7	0.7	11.9	37.8
49	2.3	0.5	0.6	4.3	9.5	0.2	0.3	0.3	0.2	0.2	0.3	0.1	0.2	0.2	1.4	0.4	0.7	2.4	0.9	3.0	4.3	3.0	0.3	0.5	0.7	1.2	0.1	0.1	10.4	4.4	0.3	16.1	9.0	12.5	38.0	15.3	5.6	49.1	24.5	49.3	85.1
50	21.0	4.2	7.8	16.1	42.7	1.8	1.1	2.0	0.9	0.8	1.1	0.5	0.7	1.5	6.2	2.4	2.9	12.7	14.1	25.3	16.8	36.0	8.4	17.5	14.1	8.3	1.5	1.3	60.0	40.0	9.2	80.0	68.2	70.4	84.9	64.8	28.1	87.5	76.1	79.8	97.1
51	78.3	38.3	65.7	68.2	92.6	20.6	21.4	30.6	27.7	19.7	20.9	12.5	18.3	25.3	54.7	34.7	43.4	65.5	69.6	79.8	72.7	95.1	60.9	73.4	75.8	48.8	24.9	25.3	95.2	85.3	59.0	96.0	96.5	98.4	98.6	90.5	88.5	97.2	96.4	92.5	99.6

- Multi-variate HMM emits vector of values, not just one value
- Can emit real values (SegWay) or binary presence /absence values (ChromHMM)
- Use to learn mark combinations

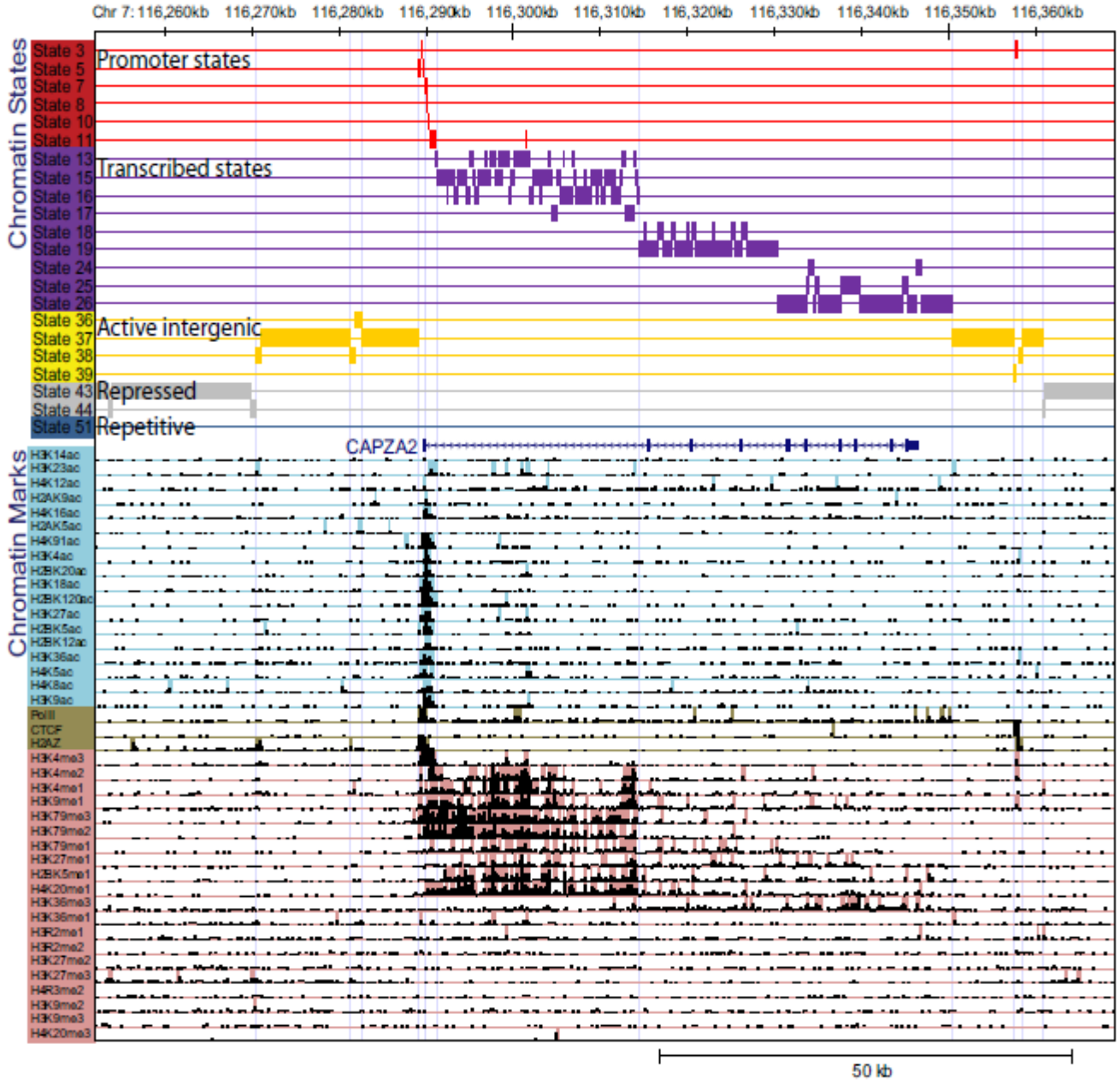
Ernst and Kellis
Nature Biotech 2010

Transition matrix a_{kl}



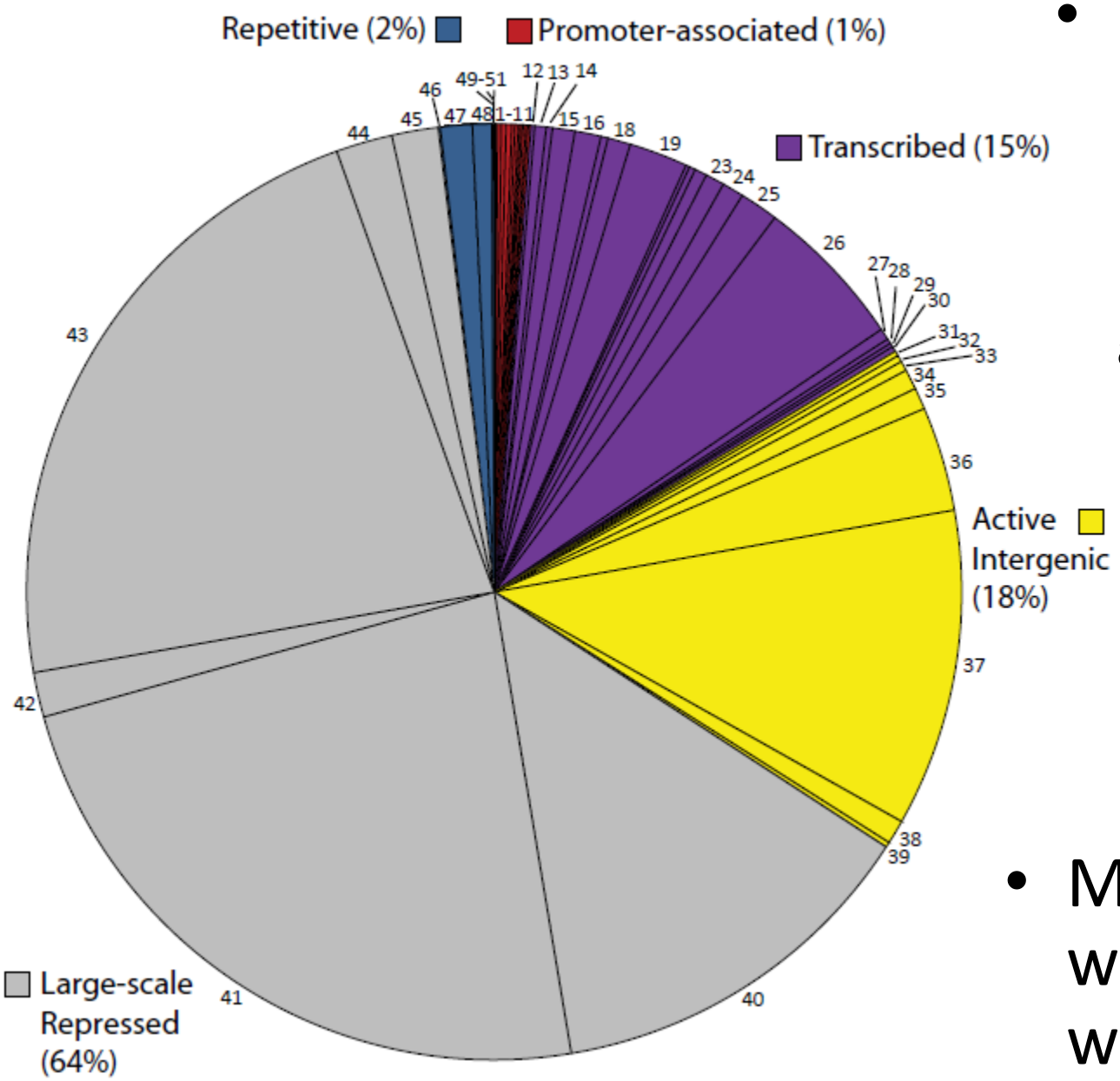
- Learns spatial relationships between neighboring states
- Reveals distinct sub-groups of states
- Reveals transitions between different groups

Example Chromatin State Annotation



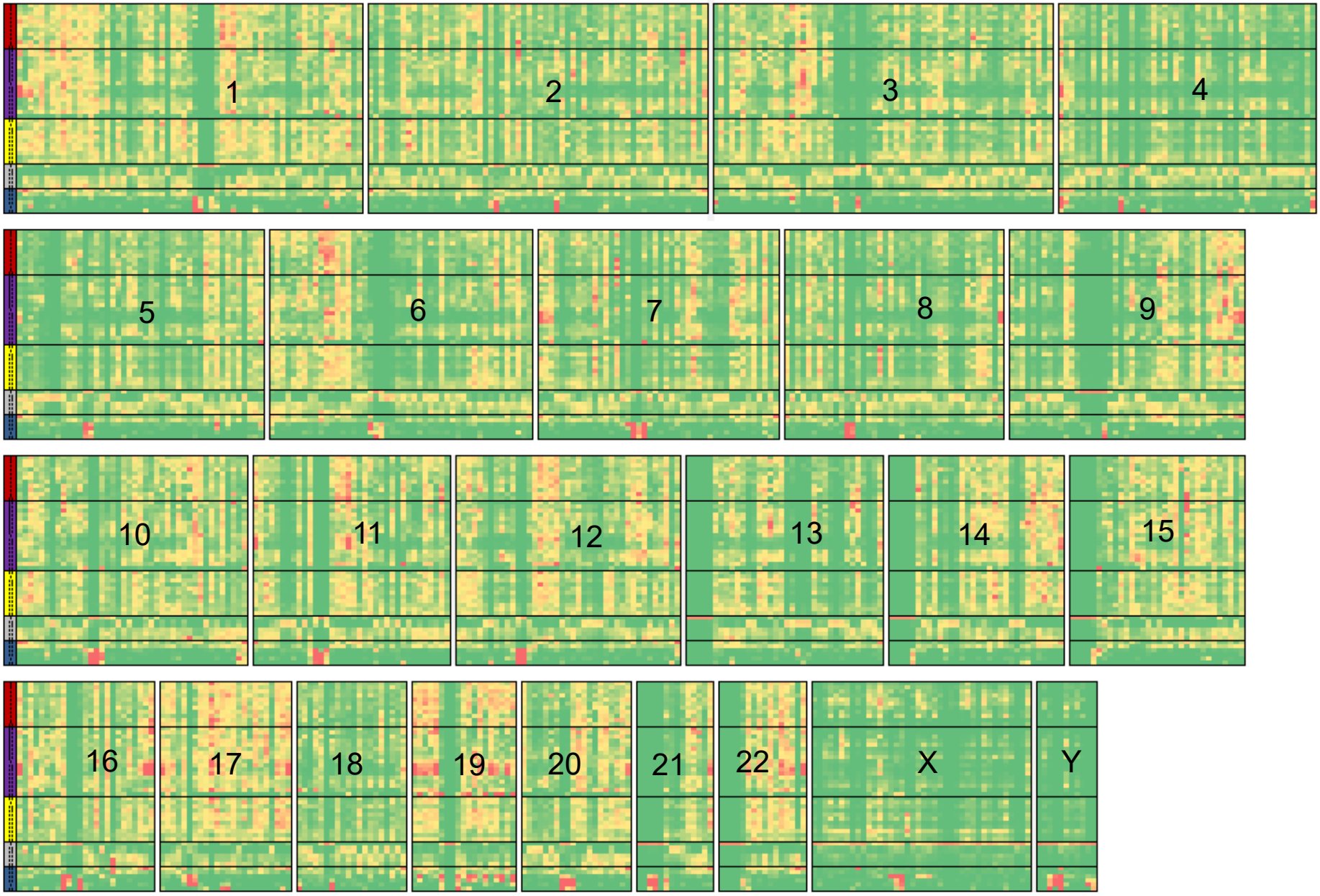
- Use Baum Welch to learn hidden states and their annotations
- Learned states correspond to known functional elements
- *De novo* discovery of major types of chromatin

Model complexity matches that of genome



- Handful of repressed states capture vast majority of genome
  - Only 1% of genome split in 14 promoter states
- Modeling power well distributed where needed

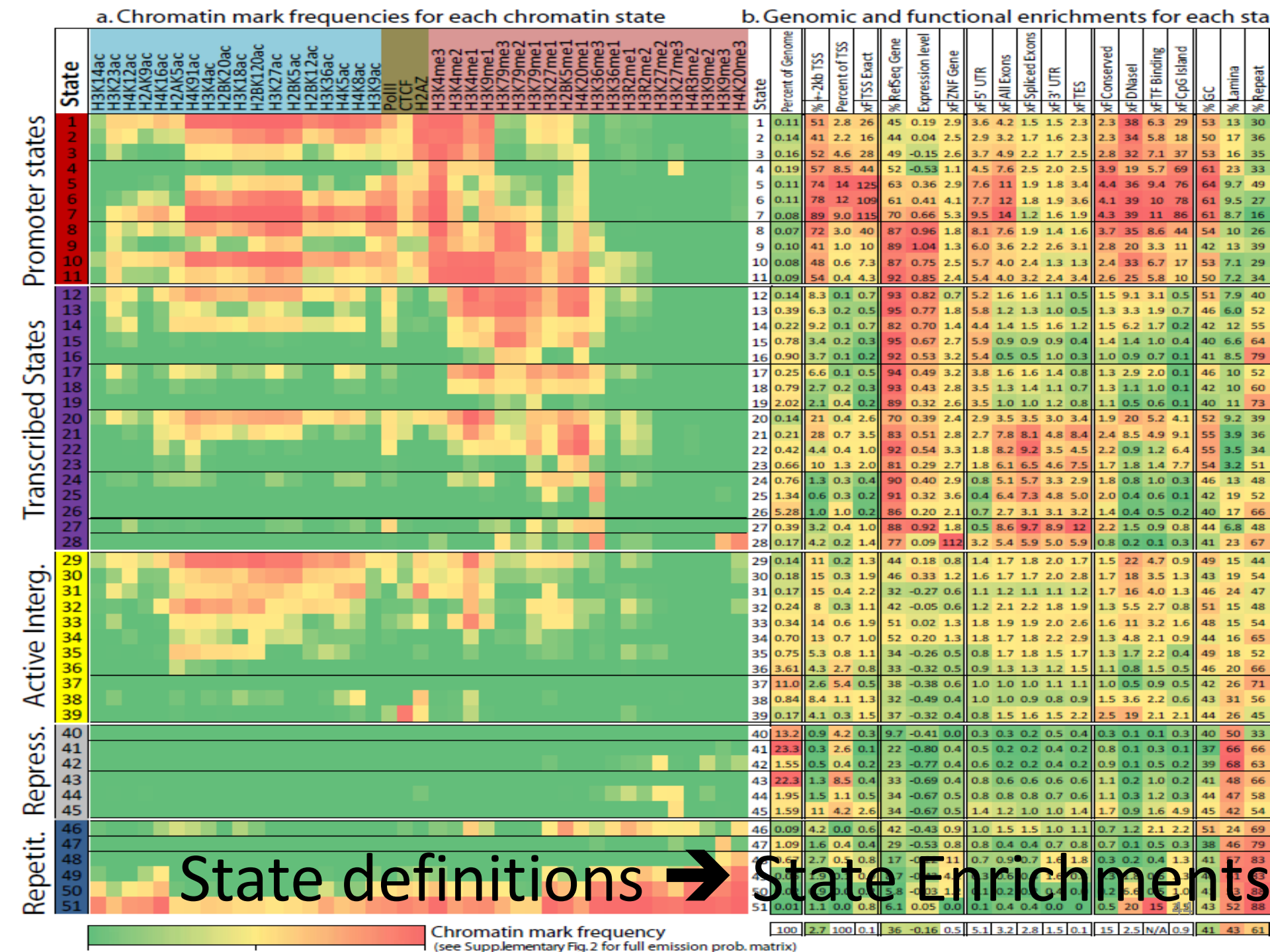
Apply genome wide to classify chromatin states *de novo*



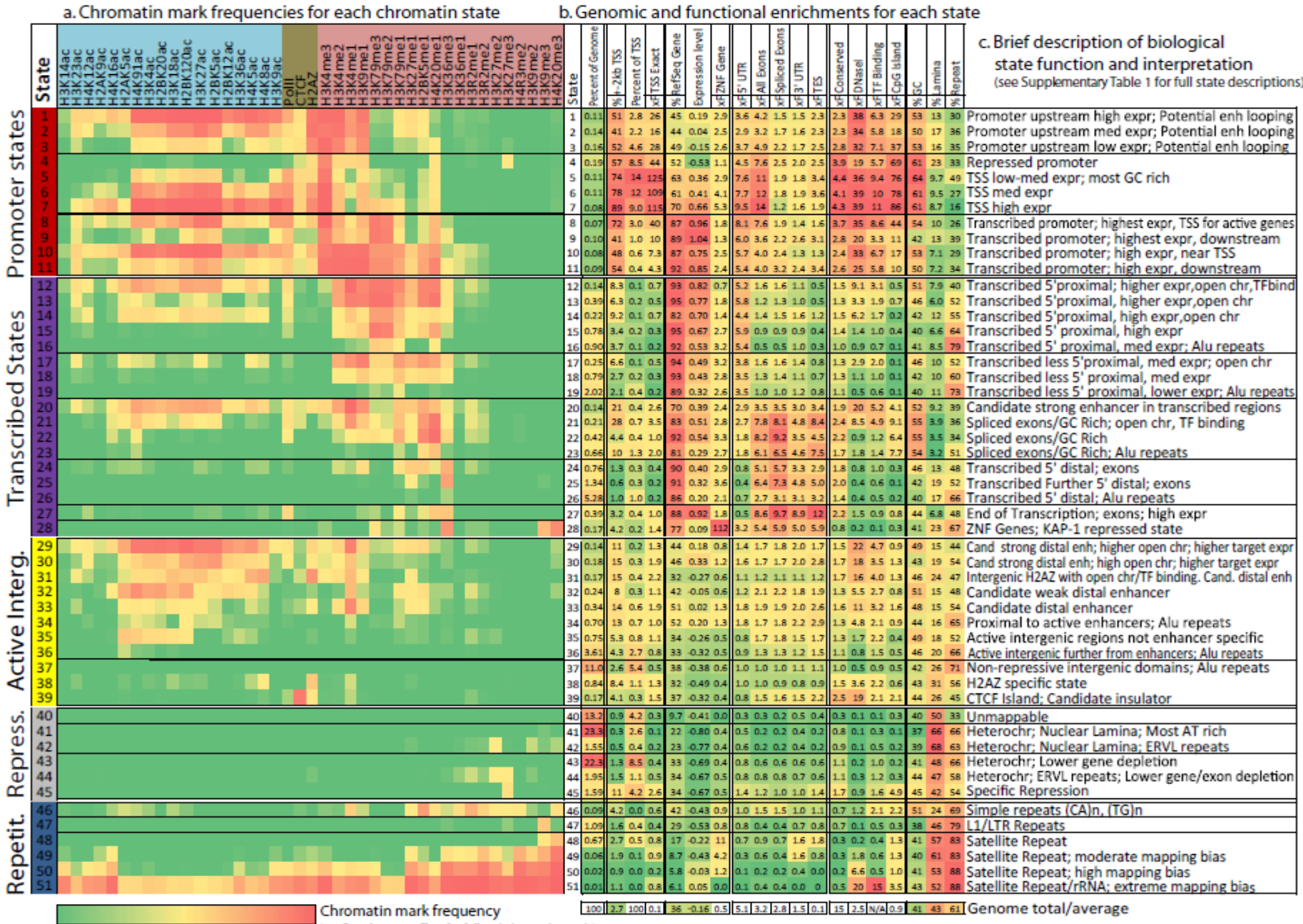
Now what? Interpret these states biologically

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - HMM Foundations, Generating, Parsing, Decoding, Learning
   - Chromatin state characterization: Functional/positional enrichment
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

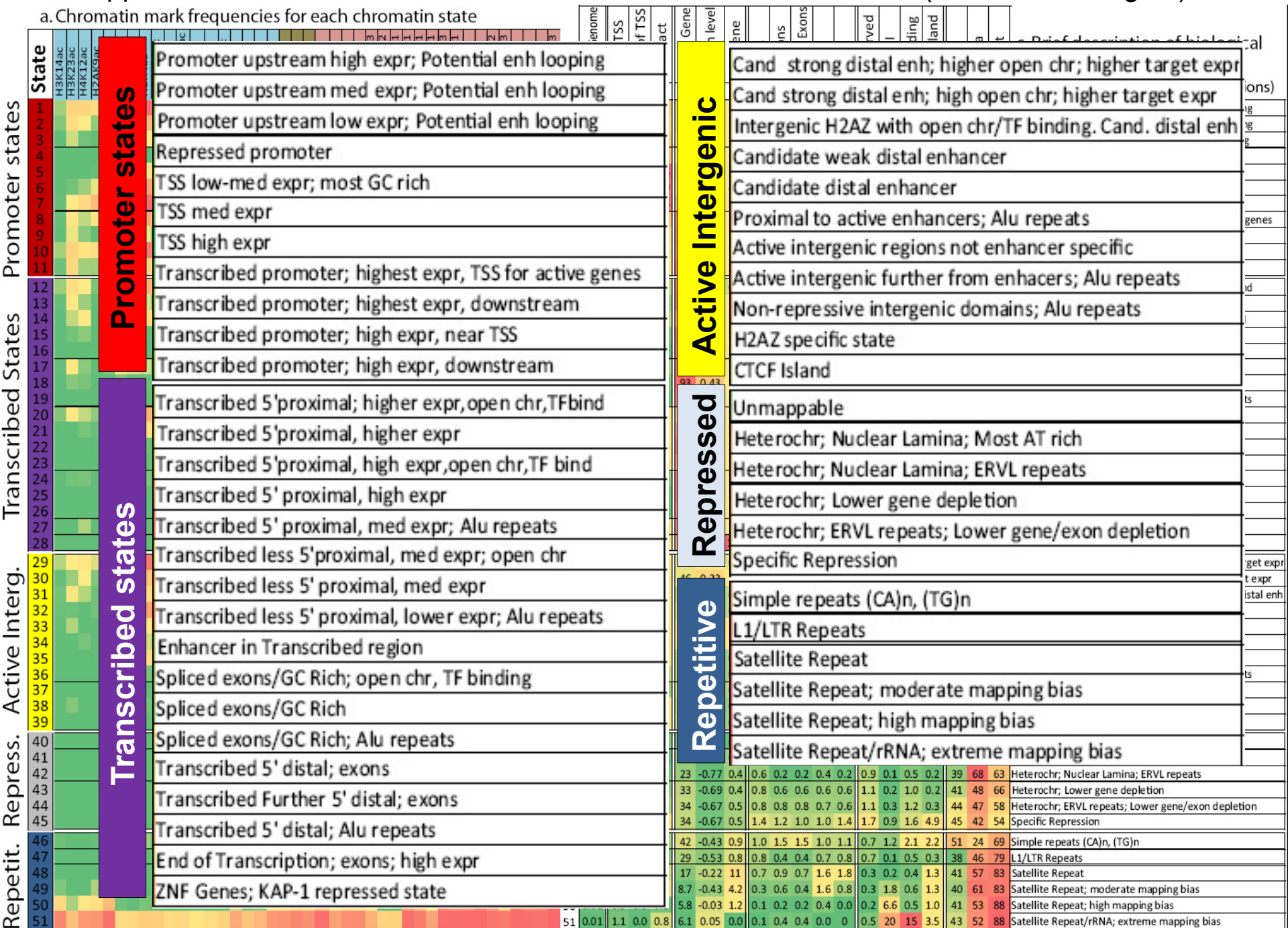

a. Chromatin mark frequencies for each chromatin state
b. Genomic and functional enrichments for each sta
State
H3K14ac
H3K23ac
H4K12ac
H2AK9ac
H4K16ac
H2AK5ac
H4K91ac
H3K4ac
H2BK20ac
H3K18ac
H2BK120ac
H3K27ac
H2BK5ac
H2BK12ac
H3K36ac
H4K5ac
H4K8ac
H3K9ac
PolII
CTCF
H2AZ
H3K4me3
H3K4me2
H3K4me1
H3K9me1
H3K79me3
H3K79me2
H3K79me1
H3K27me1
H2BK5me1
H4K20me1
H3K36me3
H3K36me1
H3R2me1
H3R2me2
H3K27me2
H3K27me3
H4R3me2
H3K9me2
H3K9me3
H4K20me3
State
Percent of Genome
% +-2kb TSS
Percent of TSS
xFTSS Exact
% RefSeq Gene
Expression level
xFZNF Gene
xF5' UTR
xFAll Exons
xFSpliced Exons
xF3' UTR
xFTES
xFConserved
xFDNaseI
xFTF Binding
xFCpG Island
% GC
% Lamina
% Repeat
Promoter states
Transcribed States
Active Interg.
Repress.
Repetit.
State definitions → State Enrichments
Chromatin mark frequency
(see Supplementary Fig. 2 for full emission prob. matrix)

Functional enrichments enable annotation of 51 distinct states



State	Percent of Genome	%/-2kb TSS	Percent of TSS	xFTSS Exact	%RefSeq Gene	Expression level	xFZNF Gene	xF5' UTR	xFAll Exons	xFSpliced Exons	xF3' UTR	xFTES	xFConserved	xFDNaseI	xFTF Binding	xFCpG Island	%GC	%Lamina	%Repeat	c. Brief description of biological state function and interpretation (see Supplementary Table 1 for full state descriptions)
Promoter states																				
1	0.11	51	2.8	26	45	0.19	2.9	3.6	4.2	1.5	1.5	2.3	2.3	38	6.3	29	53	13	30	Promoter upstream high expr; Potential enh looping
2	0.14	41	2.2	16	44	0.04	2.5	2.9	3.2	1.7	1.6	2.3	2.3	34	5.8	18	50	17	36	Promoter upstream med expr; Potential enh looping
3	0.16	52	4.6	28	49	-0.15	2.6	3.7	4.9	2.2	1.7	2.5	2.8	32	7.1	37	53	16	35	Promoter upstream low expr; Potential enh looping
4	0.19	57	8.5	44	52	-0.53	1.1	4.5	7.6	2.5	2.0	2.5	3.9	19	5.7	69	61	23	33	Repressed promoter
5	0.11	74	14	125	63	0.36	2.9	7.6	11	1.9	1.8	3.4	4.4	36	9.4	76	64	9.7	49	TSS low-med expr; most GC rich
6	0.11	78	12	109	61	0.41	4.1	7.7	12	1.8	1.9	3.6	4.1	39	10	78	61	9.5	27	TSS med expr
7	0.08	89	9.0	115	70	0.66	5.3	9.5	14	1.2	1.6	1.9	4.3	39	11	86	61	8.7	16	TSS high expr
8	0.07	72	3.0	40	87	0.96	1.8	8.1	7.6	1.9	1.4	1.6	3.7	35	8.6	44	54	10	26	Transcribed promoter; highest expr, TSS for active genes
9	0.10	41	1.0	10	89	1.04	1.3	6.0	3.6	2.2	2.6	3.1	2.8	20	3.3	11	42	13	39	Transcribed promoter; highest expr, downstream
10	0.08	48	0.6	7.3	87	0.75	2.5	5.7	4.0	2.4	1.3	1.3	2.4	33	6.7	17	53	7.1	29	Transcribed promoter; high expr, near TSS
11	0.09	54	0.4	4.3	92	0.85	2.4	5.4	4.0	3.2	2.4	3.4	2.6	25	5.8	10	50	7.2	34	Transcribed promoter; high expr, downstream
Transcribed States																				
12	0.14	8.3	0.1	0.7	93	0.82	0.7	5.2	1.6	1.6	1.1	0.5	1.5	9.1	3.1	0.5	51	7.9	40	Transcribed 5'proximal; higher expr,open chr,TFbind
13	0.39	6.3	0.2	0.5	95	0.77	1.8	5.8	1.2	1.3	1.0	0.5	1.3	3.3	1.9	0.7	46	6.0	52	Transcribed 5'proximal, higher expr,open chr
14	0.22	9.2	0.1	0.7	82	0.70	1.4	4.4	1.4	1.5	1.6	1.2	1.5	6.2	1.7	0.2	42	12	55	Transcribed 5'proximal, high expr,open chr
15	0.78	3.4	0.2	0.3	95	0.67	2.7	5.9	0.9	0.9	0.9	0.4	1.4	1.4	1.0	0.4	40	6.6	64	Transcribed 5' proximal, high expr
16	0.90	3.7	0.1	0.2	92	0.53	3.2	5.4	0.5	0.5	1.0	0.3	1.0	0.9	0.7	0.1	41	8.5	79	Transcribed 5' proximal, med expr; Alu repeats
17	0.25	6.6	0.1	0.5	94	0.49	3.2	3.8	1.6	1.6	1.4	0.8	1.3	2.9	2.0	0.1	46	10	52	Transcribed less 5'proximal, med expr; open chr
18	0.79	2.7	0.2	0.3	93	0.43	2.8	3.5	1.3	1.4	1.1	0.7	1.3	1.1	1.0	0.1	42	10	60	Transcribed less 5' proximal, med expr
19	2.02	2.1	0.4	0.7	89	0.32	2.6	3.5	1.0	1.0	1.2	0.8	1.1	0.5	0.6	0.1	40	11	73	Transcribed less 5' proximal, lower expr; Alu repeats
20	0.14	21	0.4	2.6	70	0.39	2.4	2.9	3.5	3.5	3.0	3.4	1.9	20	5.2	4.1	52	9.2	39	Candidate strong enhancer in transcribed regions
21	0.21	28	0.7	3.5	83	0.51	2.8	2.7	7.8	8.1	4.8	8.4	2.4	8.5	4.9	9.1	55	3.9	36	Spliced exons/GC Rich; open chr, TF binding
22	0.42	4.4	0.4	1.0	92	0.54	3.3	1.8	8.2	9.2	3.5	4.5	2.2	0.9	1.2	6.4	55	3.5	34	Spliced exons/GC Rich
23	0.66	10	1.3	2.0	81	0.29	2.7	1.8	6.1	6.5	4.6	7.5	1.7	1.8	1.4	7.7	54	3.2	51	Spliced exons/GC Rich; Alu repeats
24	0.76	1.3	0.3	0.4	90	0.40	2.9	0.8	5.1	5.7	3.3	2.9	1.8	0.8	1.0	0.3	46	13	48	Transcribed 5' distal; exons
25	1.34	0.6	0.3	0.2	91	0.32	3.6	0.4	6.4	7.3	4.8	5.0	2.0	0.4	0.6	0.1	42	19	52	Transcribed Further 5' distal; exons
26	5.28	1.0	1.0	0.2	86	0.20	2.1	0.7	2.7	3.1	3.1	3.2	1.4	0.4	0.5	0.2	40	17	66	Transcribed 5' distal; Alu repeats
27	0.39	3.2	0.4	1.0	88	0.92	1.8	0.5	8.6	9.7	8.9	12	2.2	1.5	0.9	0.8	44	6.8	48	End of Transcription; exons; high expr
28	0.17	4.2	0.2	1.4	77	0.09	112	3.2	5.4	5.9	5.0	5.9	0.8	0.2	0.1	0.3	41	23	67	ZNF Genes; KAP-1 repressed state
Active Interg.																				
29	0.14	11	0.2	1.3	44	0.18	0.8	1.4	1.7	1.8	2.0	1.7	1.5	22	4.7	0.9	49	15	44	Cand strong distal enh; higher open chr; higher target expr
30	0.18	15	0.3	1.9	46	0.33	1.2	1.6	1.7	1.7	2.0	2.8	1.7	18	3.5	1.3	43	19	54	Cand strong distal enh; high open chr; higher target expr
31	0.17	15	0.4	2.2	32	-0.27	0.6	1.1	1.2	1.1	1.1	1.2	1.7	16	4.0	1.3	46	24	47	Intergenic H2AZ with open chr/TF binding. Cand. distal enh
32	0.24	8	0.3	1.1	42	-0.05	0.6	1.2	2.1	2.2	1.8	1.9	1.3	5.5	2.7	0.8	51	15	48	Candidate weak distal enhancer
33	0.34	14	0.6	1.9	51	0.02	1.3	1.8	1.9	1.9	2.0	2.6	1.6	11	3.2	1.6	48	15	54	Candidate distal enhancer
34	0.70	13	0.7	1.0	52	0.20	1.3	1.8	1.7	1.8	2.2	2.9	1.3	4.8	2.1	0.9	44	16	65	Proximal to active enhancers; Alu repeats
35	0.75	5.3	0.8	1.1	34	-0.26	0.5	0.8	1.7	1.8	1.5	1.7	1.3	1.7	2.2	0.4	49	18	52	Active intergenic regions not enhancer specific
36	3.61	4.3	2.7	0.8	33	-0.32	0.5	0.9	1.3	1.3	1.2	1.5	1.1	0.8	1.5	0.5	46	20	66	Active intergenic further from enhancers; Alu repeats
37	11.0	2.6	5.4	0.5	38	-0.38	0.6	1.0	1.0	1.0	1.1	1.1	1.0	0.5	0.9	0.5	42	26	71	Non-repressive intergenic domains; Alu repeats
38	0.84	8.4	1.1	1.3	32	-0.49	0.4	1.0	1.0	0.9	0.8	0.9	1.5	3.6	2.2	0.6	43	31	56	H2AZ specific state
39	0.17	4.1	0.3	1.5	37	-0.32	0.4	0.8	1.5	1.6	1.5	2.2	2.5	19	2.1	2.1	44	26	45	CTCF Island; Candidate insulator
Repress.																				
40	13.2	0.9	4.2	0.3	9.7	-0.41	0.0	0.3	0.3	0.2	0.5	0.4	0.3	0.1	0.1	0.3	40	50	33	Unmappable
41	23.3	0.3	2.6	0.1	22	-0.80	0.4	0.5	0.2	0.2	0.4	0.2	0.8	0.1	0.3	0.1	37	66	66	Heterochr; Nuclear Lamina; Most AT rich
42	1.55	0.5	0.4	0.2	23	-0.77	0.4	0.6	0.2	0.2	0.4	0.2	0.9	0.1	0.5	0.2	39	68	63	Heterochr; Nuclear Lamina; ERVL repeats
43	22.3	1.3	8.5	0.4	33	-0.69	0.4	0.8	0.6	0.6	0.6	0.6	1.1	0.2	1.0	0.2	41	48	66	Heterochr; Lower gene depletion
44	1.95	1.5	1.1	0.5	34	-0.67	0.5	0.8	0.8	0.8	0.7	0.6	1.1	0.3	1.2	0.3	44	47	58	Heterochr; ERVL repeats; Lower gene/exon depletion
45	1.59	11	4.2	2.6	34	-0.67	0.5	1.4	1.2	1.0	1.0	1.4	1.7	0.9	1.6	4.9	45	42	54	Specific Repression
Repetit.																				
46	0.09	4.2	0.0	0.6	42	-0.43	0.9	1.0	1.5	1.5	1.0	1.1	0.7	1.2	2.1	2.2	51	24	69	Simple repeats (CA)n, (TG)n
47	1.09	1.6	0.4	0.4	29	-0.53	0.8	0.8	0.4	0.4	0.7	0.8	0.7	0.1	0.5	0.3	38	46	79	L1/LTR Repeats
48	0.67	2.7	0.5	0.8	17	-0.22	11	0.7	0.9	0.7	1.6	1.8	0.3	0.2	0.4	1.3	41	57	83	Satellite Repeat
49	0.06	1.9	0.1	0.9	8.7	-0.43	4.2	0.3	0.6	0.4	1.6	0.8	0.3	1.8	0.6	1.3	40	61	83	Satellite Repeat; moderate mapping bias
50	0.02	0.9	0.0	0.2	5.8	-0.03	1.2	0.1	0.2	0.2	0.4	0.0	0.2	6.6	0.5	1.0	41	53	88	Satellite Repeat; high mapping bias
51	0.01	1.1	0.0	0.8	6.1	0.05	0.0	0.1	0.4	0.4	0.0	0	0.5	20	15	3.5	43	52	88	Satellite Repeat/rRNA; extreme mapping bias
	100	2.7	100	0.1	36	-0.16	0.5	5.1	3.2	2.8	1.5	0.1	15	2.5	N/A	0.9	41	43	61	Genome total/average

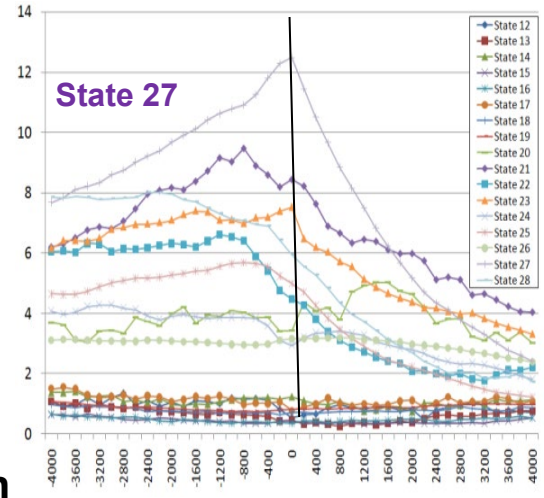
Application of ChromHMM to 41 chromatin marks in CD4+ T-cells (Barski'07, Wang'08)



Functional properties of discovered chromatin states

GO Category	State 3	State 4	State 5	State 6	State 7	State 8
Cell Cycle Phase	2.10 ($2x10^{-7}$)	0.57 (1)	1.61 (0.001)	1.45 (1)	1.15 (1)	1.51 (1)
Embryonic Development	1.24 (1)	2.82 ($9x10^{-23}$)	1.07 (1)	0.85 (1)	0.54 (1)	1.00 (1)
Chromatin	1.20 (1)	0.48 (1)	2.2 ($1.4x10^{-7}$)	1.64 (1)	0.85 (1)	0.85 (1)
Response to DNA Damage Stimulus	1.20 (1)	0.35 (1)	1.55 (0.074)	2.13 ($6.5x10^{-11}$)	1.97 ($1.0x10^{-4}$)	0.84 (1)
RNA Processing	0.49 (1)	0.26 (1)	1.31 (1)	1.91 ($4.2x10^{-11}$)	2.64 ($8.7x10^{-24}$)	2.45 ($3.0x10^{-4}$)
T cell Activation	0.77 (1)	0.88 (1)	1.27 (1)	0.70 (1)	0.79 (1)	4.72 ($2x10^{-7}$)

Promoter state ➔ gene GO function

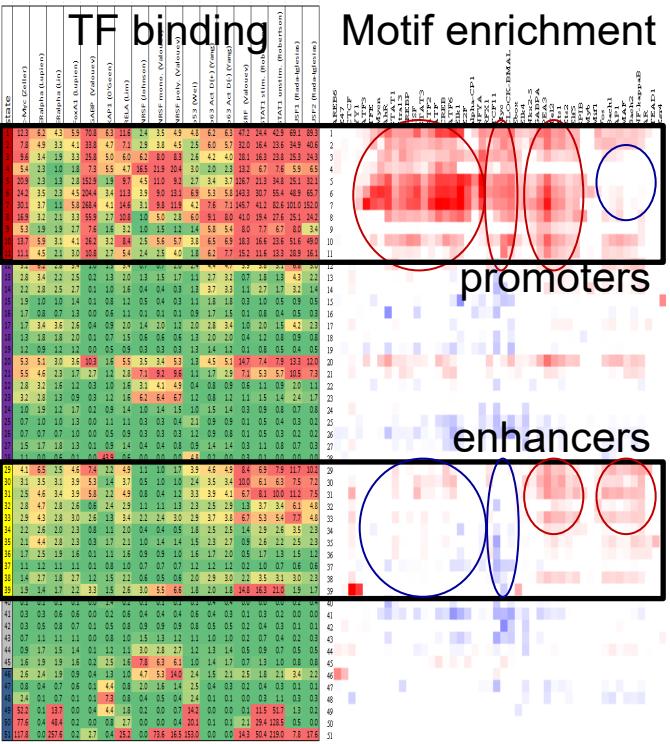


Transcription End State

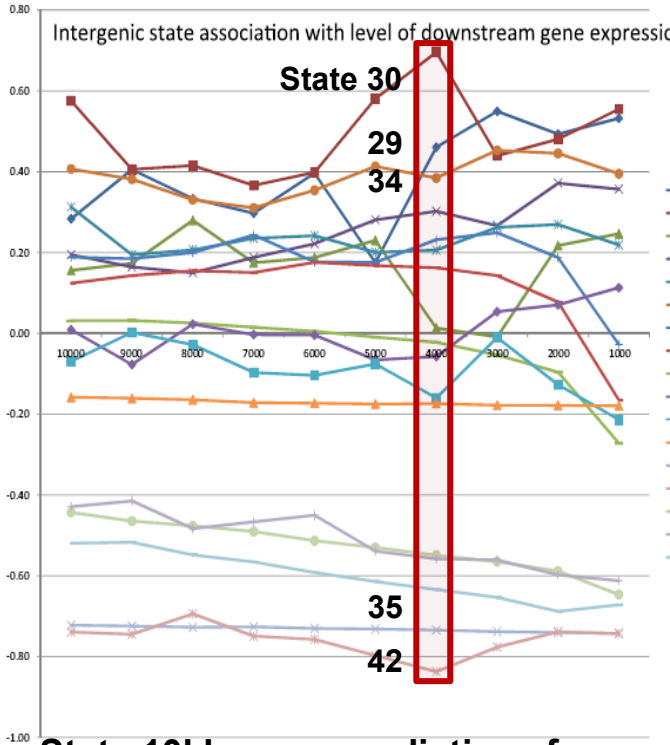
State 28: 112-fold ZNF enrich

"The achievement of the repressed state by wild-type KAP1 involves decreased recruitment of **RNA polymerase II**, reduced levels of histone **H3 K9 acteylation** and **H3K4 methylation**, an increase in **histone occupancy**, enrichment of **trimethyl histone H3K9**, **H3K36**, and **histone H4K20** …" MCB 2006.

ZNF repressed state recovery



Promoter vs. enhancer regulation

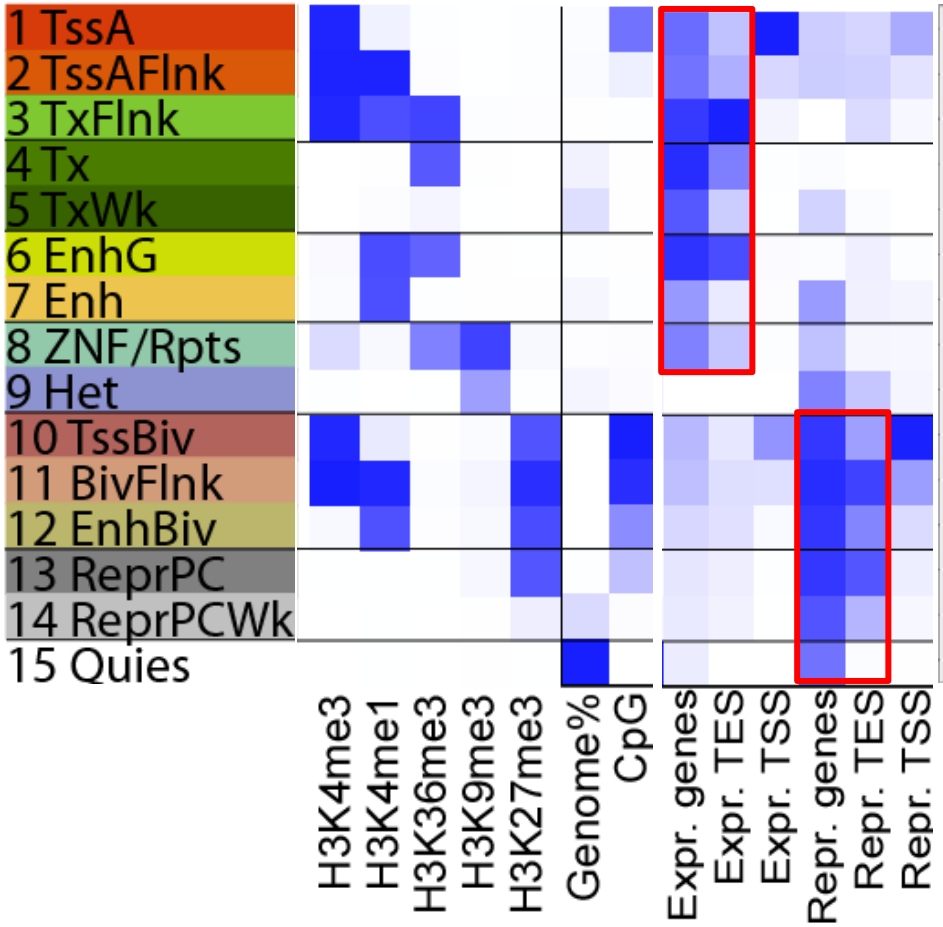


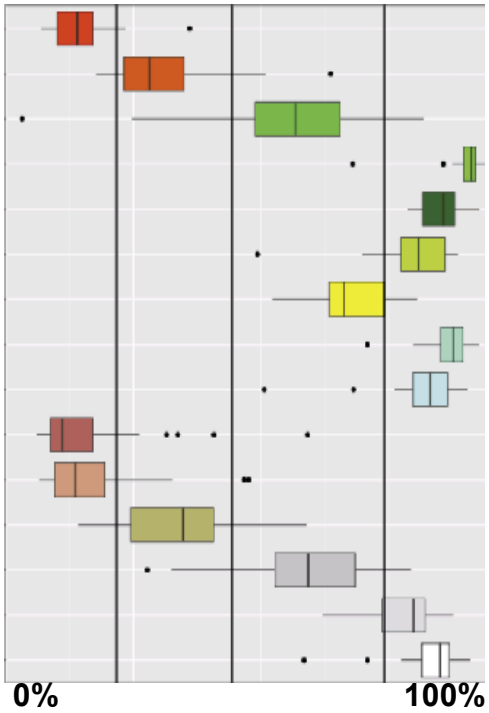
State 10kb away predictive of expr.

state	stalk	variable heterochromatic	acrocentric	gneg	gpos25	gpos50	gpos75	gpos100
40	7.6	6.6	6.1	0.6	0.4	0.5	0.3	0.5
41	0.0	0.2	0.4	0.5	0.4	0.9	1.7	2.5
42	0.0	0.2	0.3	0.5	0.5	1.1	1.7	2.2
43	0.0	0.1	0.1	1.1	1.2	1.3	1.1	0.7
44	0.0	0.1	0.1	1.2	1.3	1.3	1.0	0.6
45	0.0	0.1	0.1	1.3	1.6	1.3	0.8	0.4
46	0.0	0.1	0.2	1.7	1.7	0.7	0.4	0.2
47	0.0	0.2	0.1	1.2	1.3	1.3	0.9	0.6
48	0.0	3.2	6.2	0.8	2.2	0.2	0.5	0.4
49	0.0	3.6	11.2	0.5	1.8	0.2	0.4	0.1
50	0.0	4.7	12.0	0.6	0.6	0.2	0.2	0.1
51	0.0	4.4	12.7	0.5	1.4	0.2	0.2	0.1
% Overall	0.6	3.9	3.6	42.1	6.8	13.6	13.1	16.2

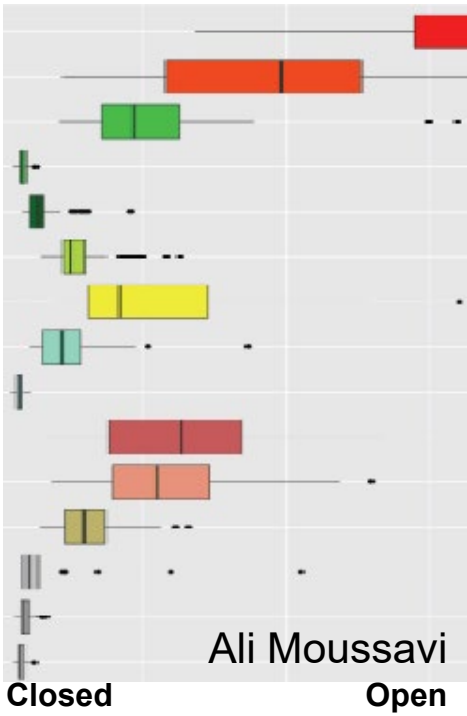
Distinct types of repression
- **Chrom bands / HDAC resp**
- **Repeat family / composition**

States show distinct mCpG, DNase, Tx, Ac profiles





DNA methylation
(WGBS, 37 epigenomes)



DNA accessibility
(DNaseI, 53 epigenomes

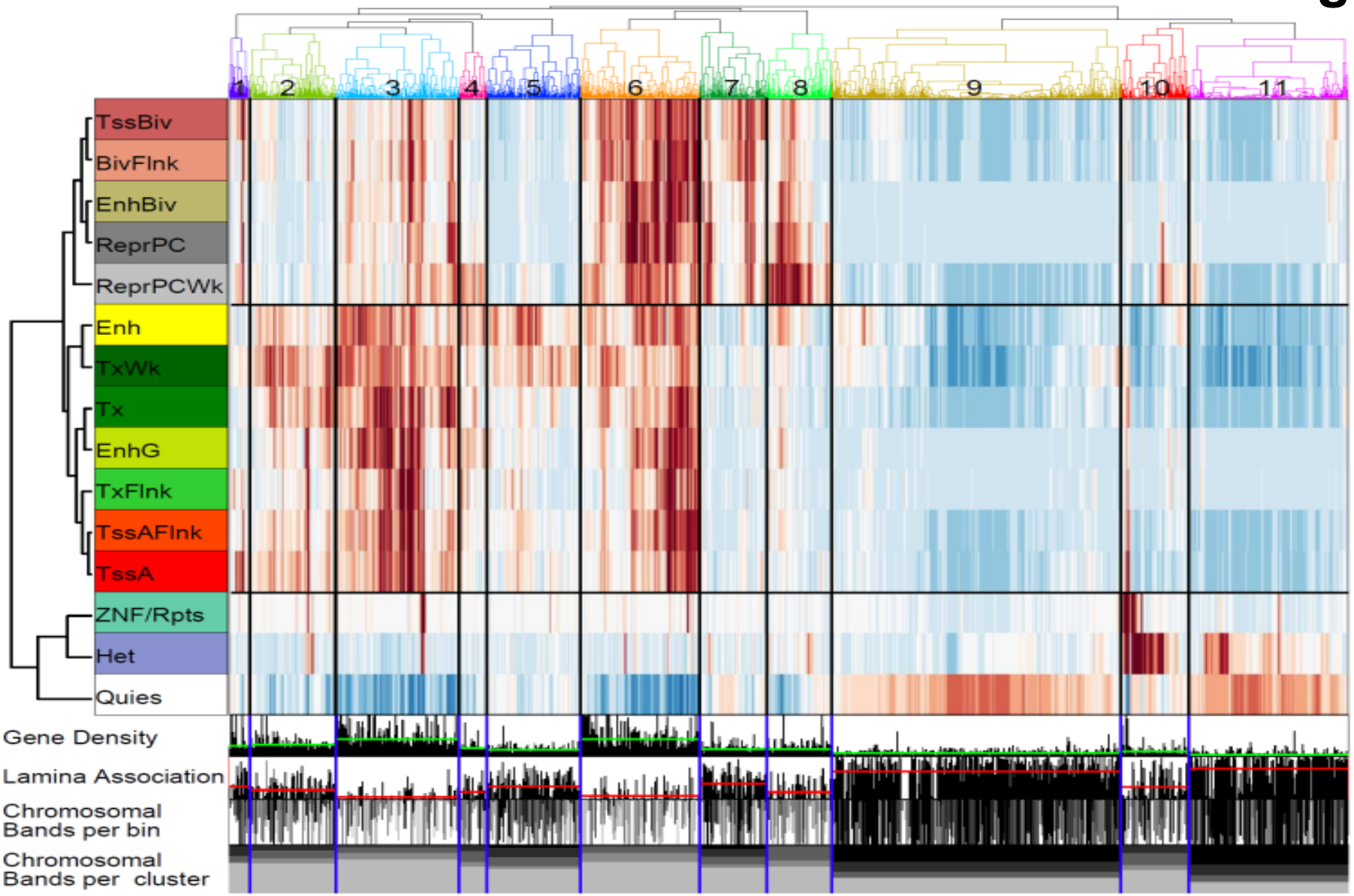
TssA vs. **TssBiv**: diff. activity, both open, both unmethylated!

Enh vs. **ReprPC**: diff. activity, both intermediate DNase/Methyl

Tx: Methylated, closed, actively transcribed

➔ Distinct modes of repression: **H3K27me3** vs. **DNAme** vs. **Het**

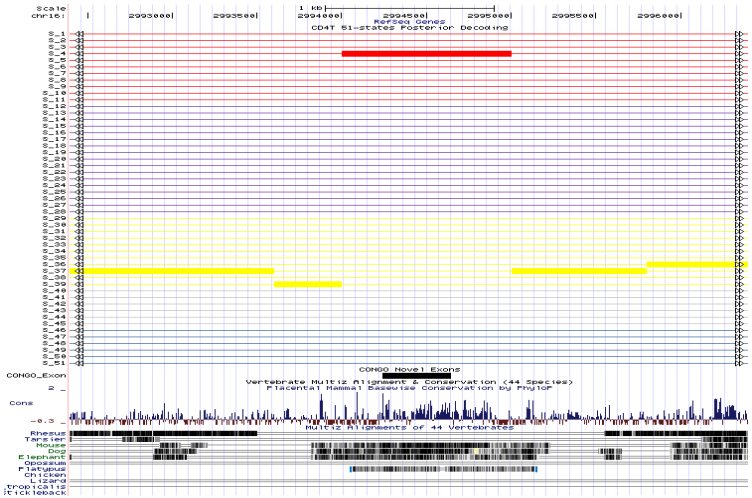
Chromosomal ‘domains’ from chromatin state usage



- State usage ➔ gene density, lamina, cytogenetic bands
- Quies/ZNF/het | gene rich/poor, each active/repressed

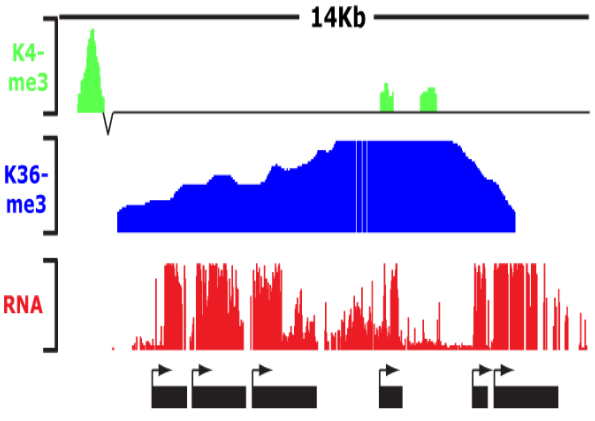
Applications to genome annotation

New protein-coding genes

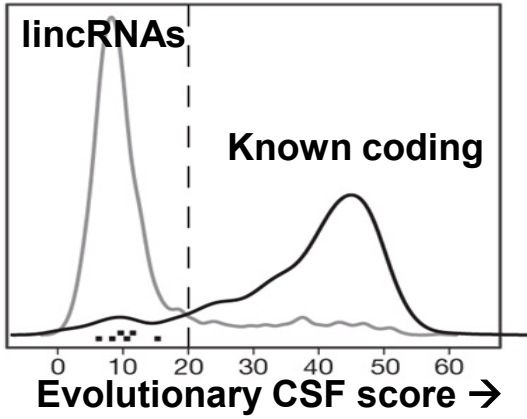


In promoter(short)/low-expr states

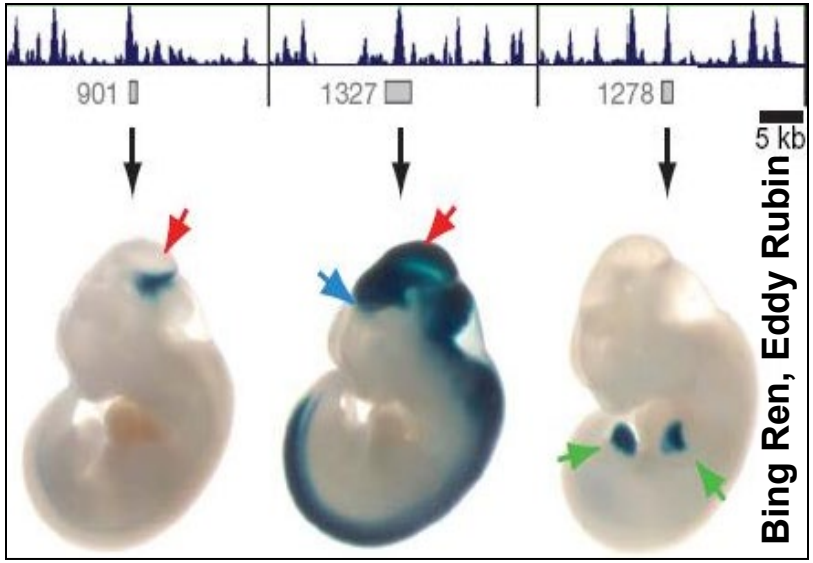
Long intergenic non-coding RNAs/lincRNAs



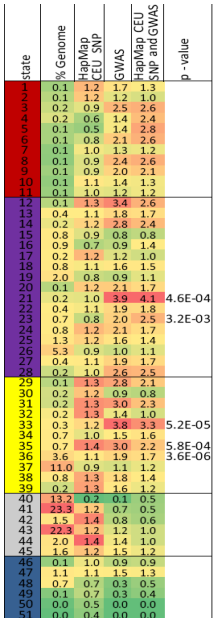
Chromatin signature: promoter / transcribed

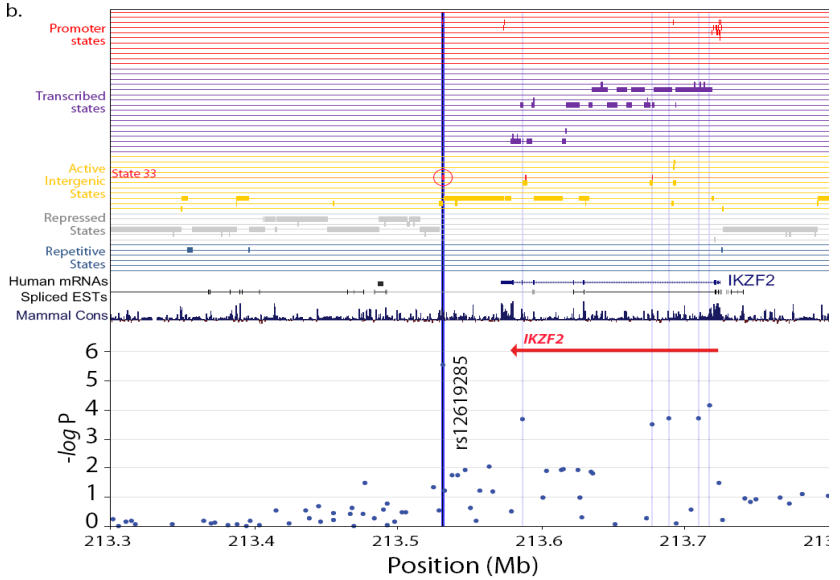


Evolutionary signature: not protein-coding



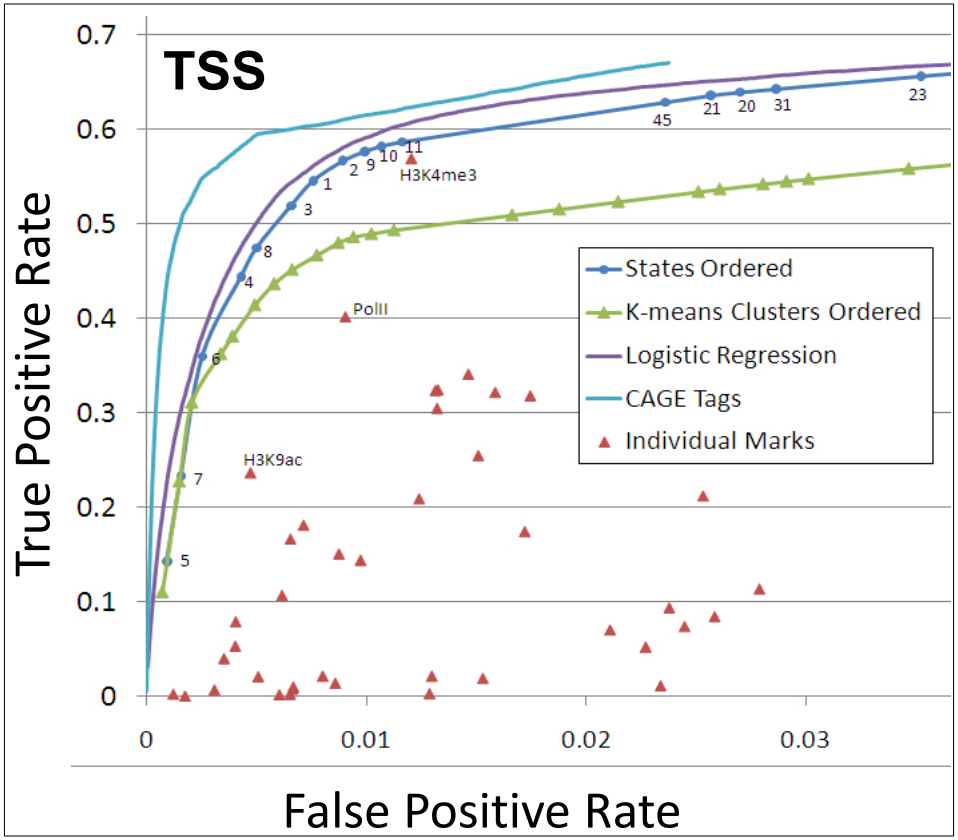
New developmental enhancer regions

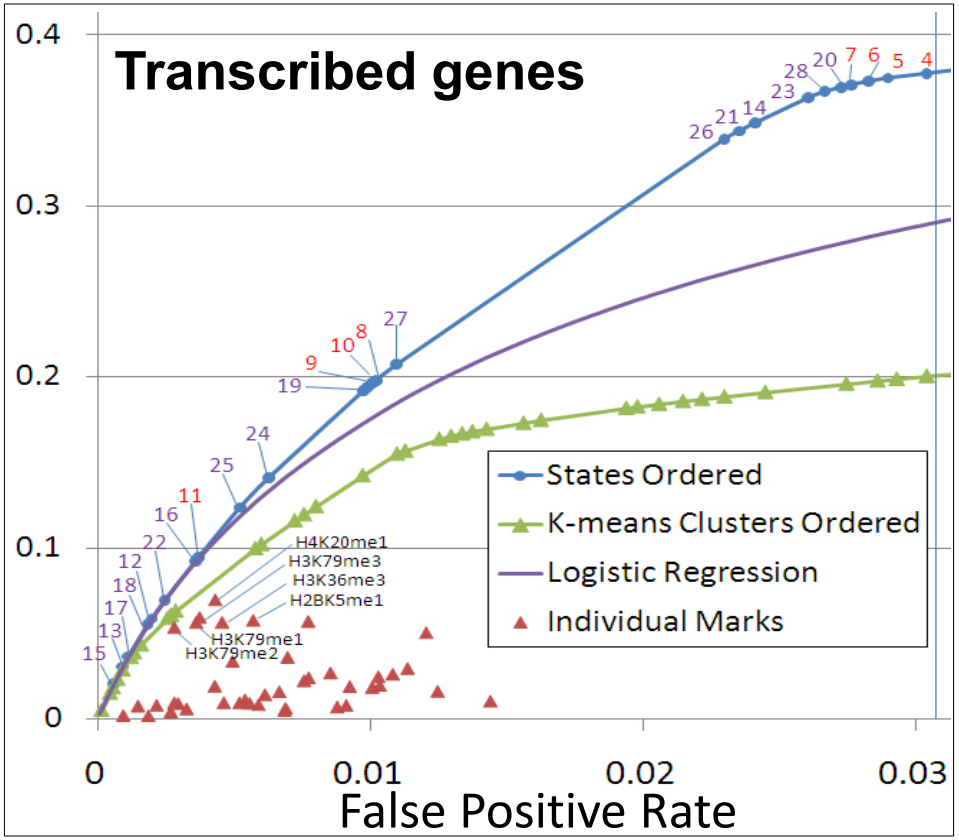




Assign candidate functions to intergenic SNPs from genome-wide association studies

Discovery power for promoters, transcripts





- Significantly outperforms single-marks
- Similar power to supervised learning approach
- CAGE experiments give possible upper bound

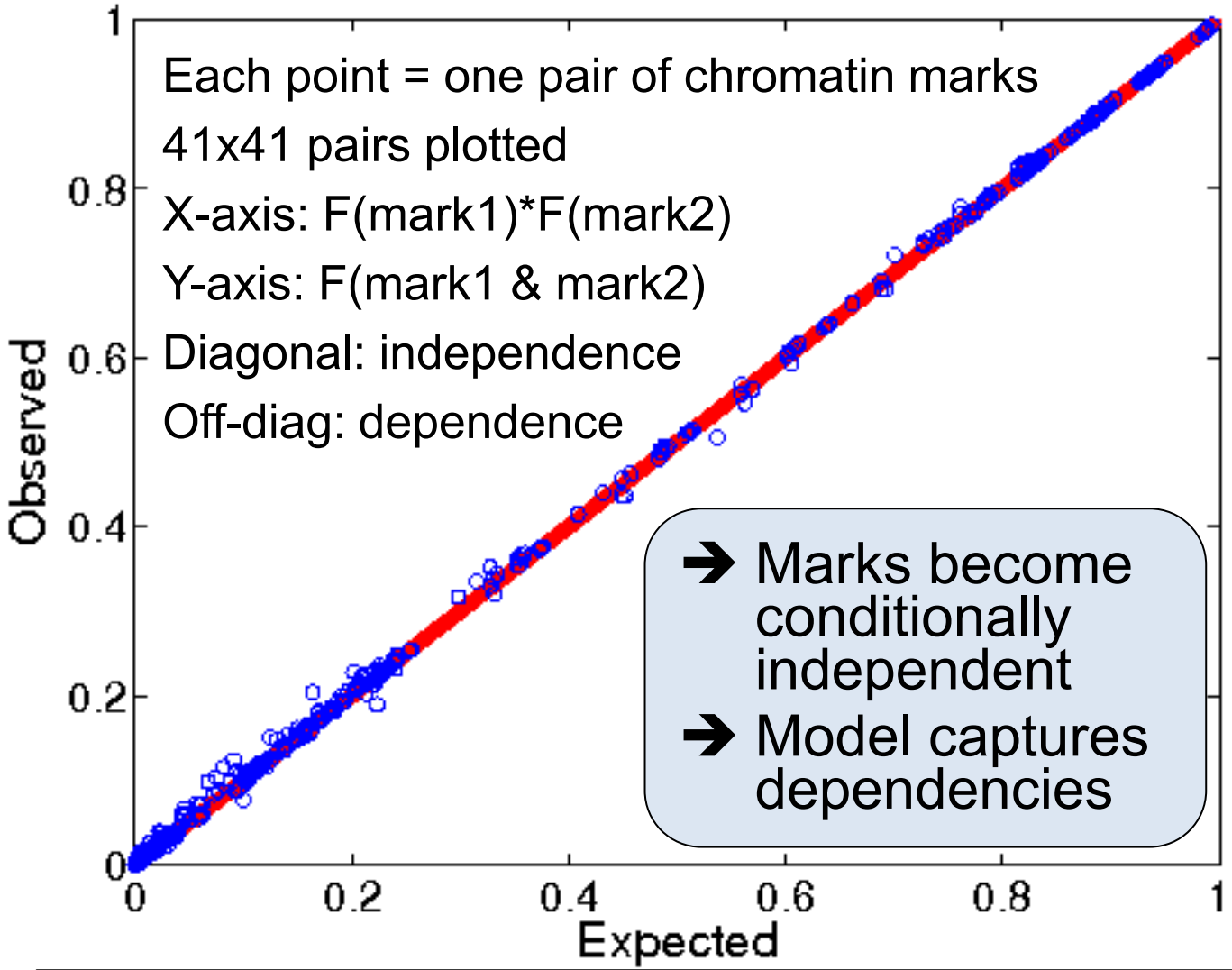
Goals for today: Computational Epigenomics

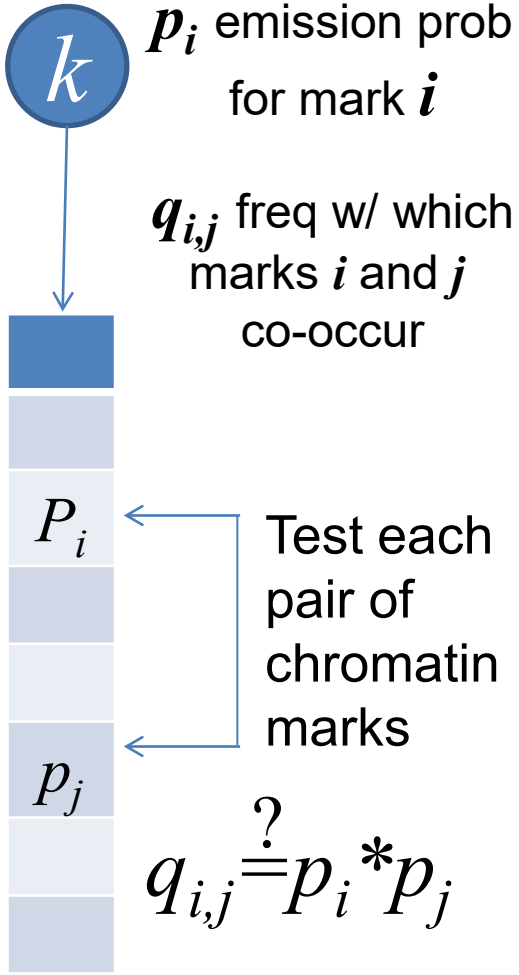
1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - HMM Foundations, Generating, Parsing, Decoding, Learning
   - ChromHMM: Multi-variate HMM for chromatin state learning
4. Model complexity: selecting the number of states/marks
   - Capturing dependencies. State-conditional mark independence
   - Selecting the number of states, selecting number of marks
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

State-conditional mark independence

Do hidden states actually capture dependencies between marks?

Pairwise Expected vs. Observed Mark Co-Occurence

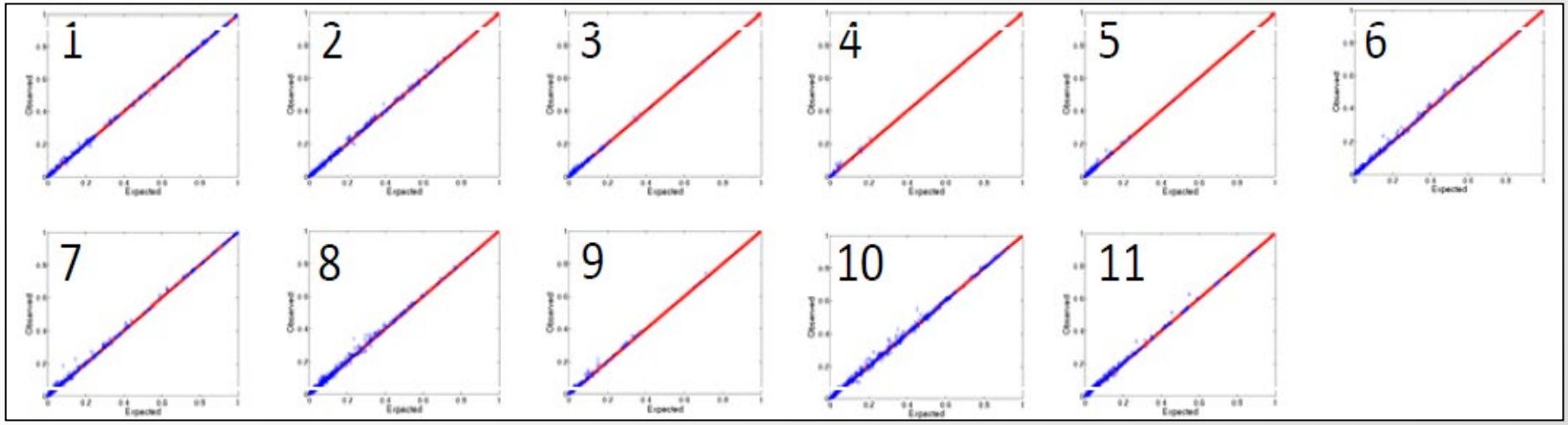




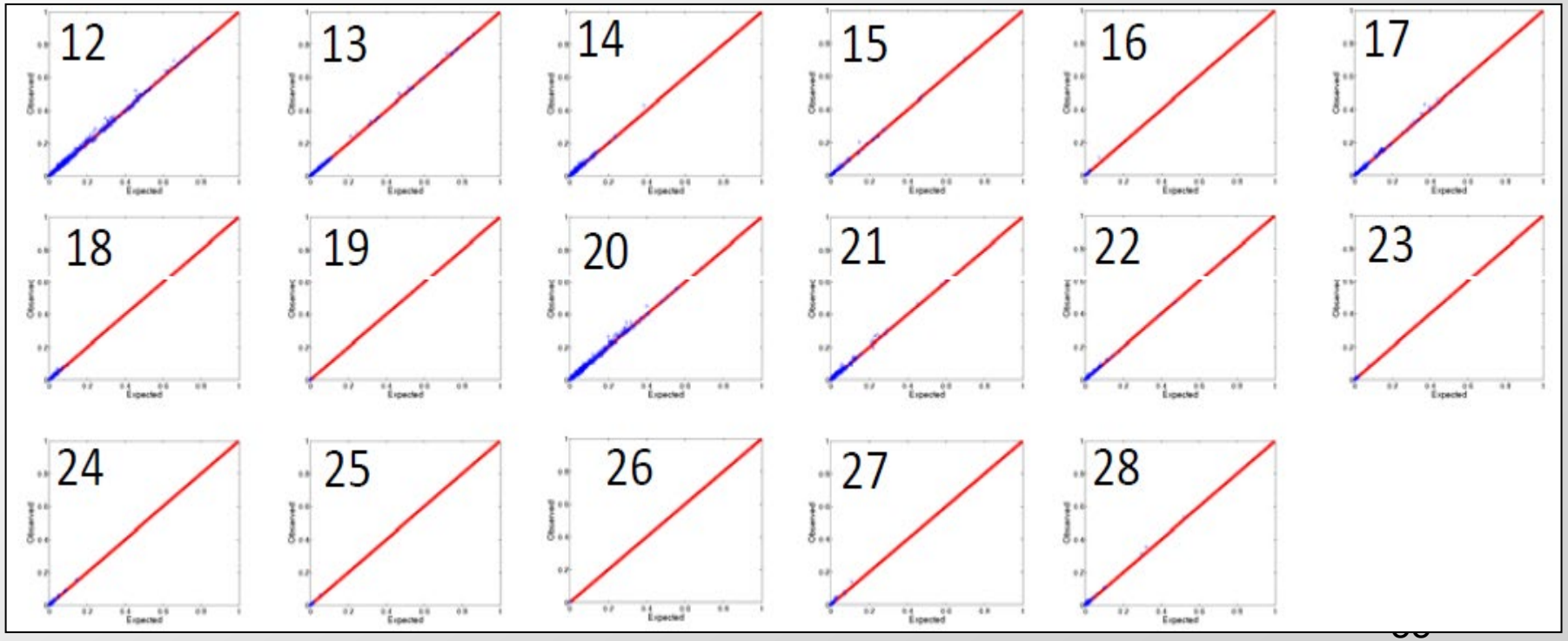
Multi-variate HMM emits entire vector of marks at a time
Model assumes mark independence **conditional** upon state
In fact, it specifically seeks to **capture** these dependencies

Test conditional independence for each state

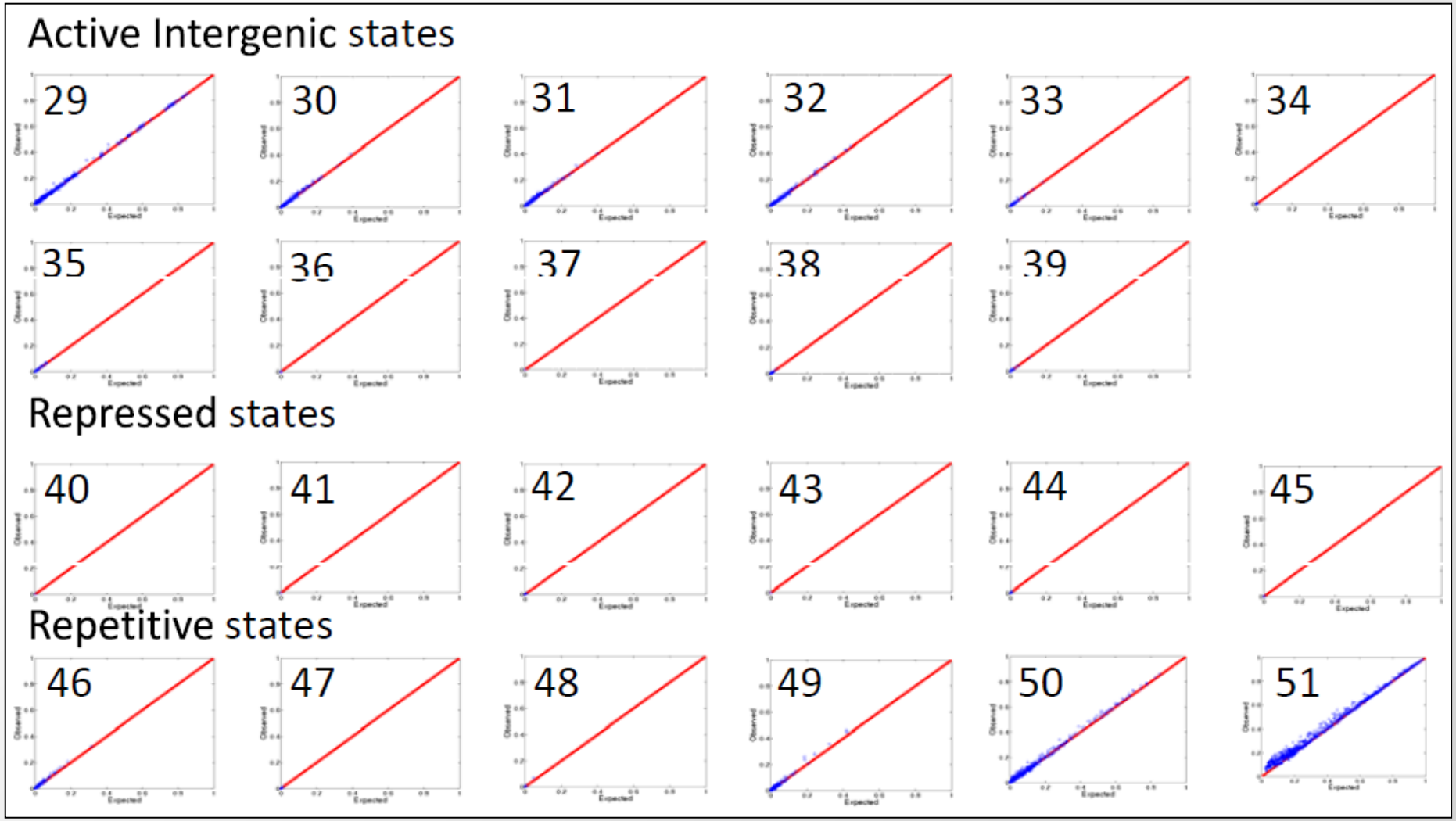
Promoter states



Transcribed states

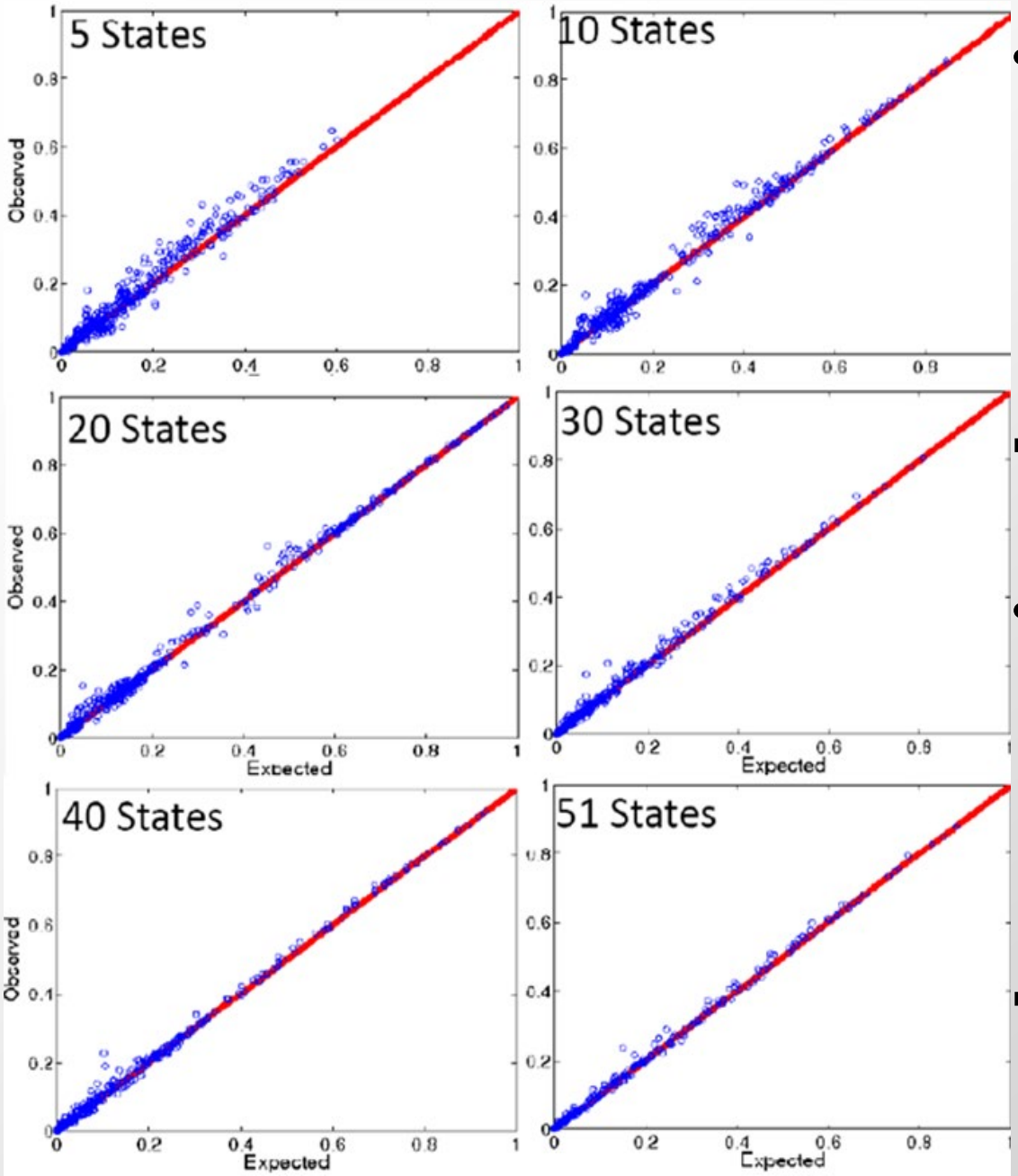


Non-independence reveals cases of model violation



- **Repetitive states show more dependencies**
- **Conditional independence does not hold**

As more states are added, dependencies captured



- With only 5 states in HMM, not enough power to distinguish different properties

➔ Dependencies remain

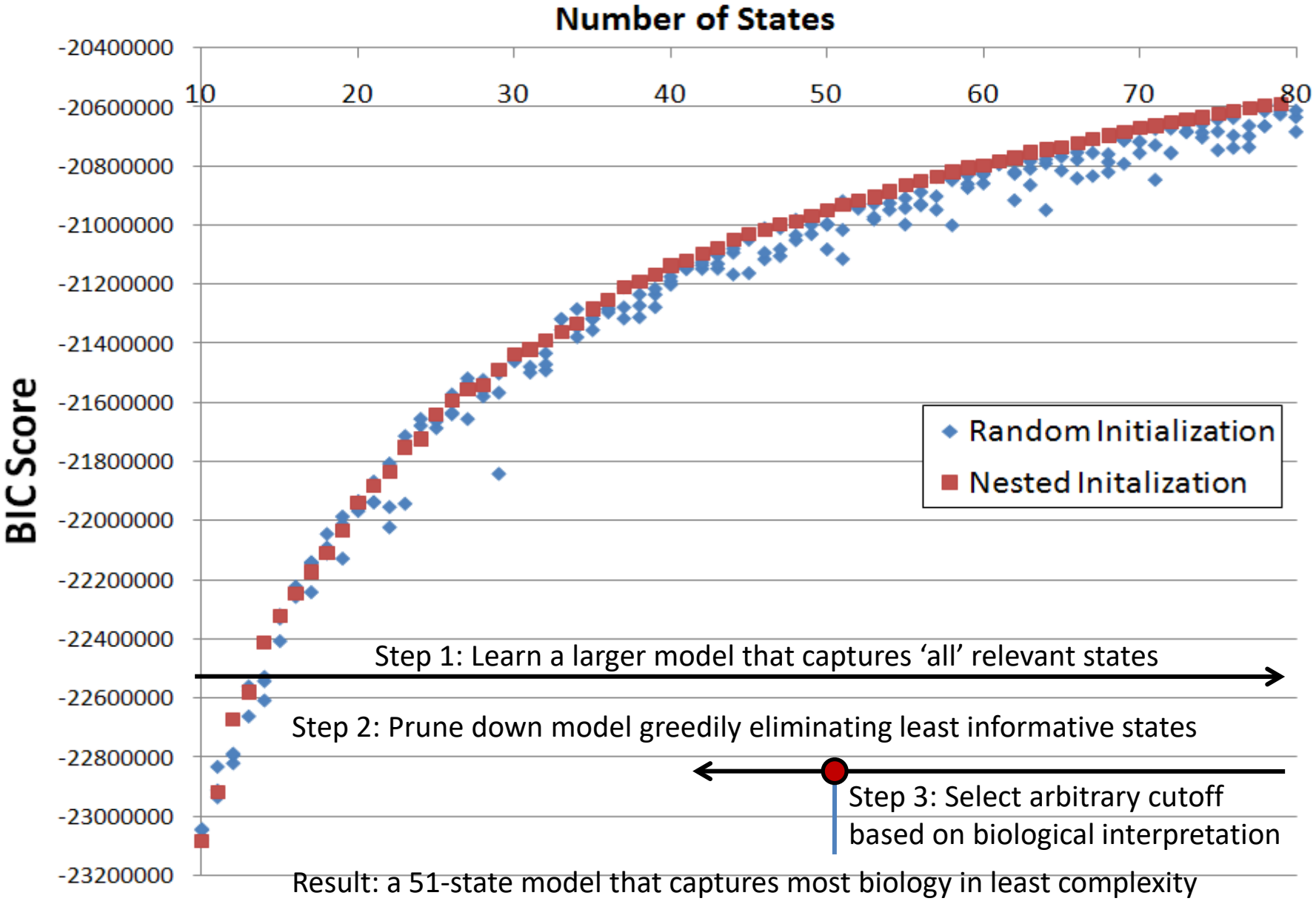
- As model complexity increases, states learned become more precise

➔ Dependencies captured

Goals for today: Computational Epigenomics

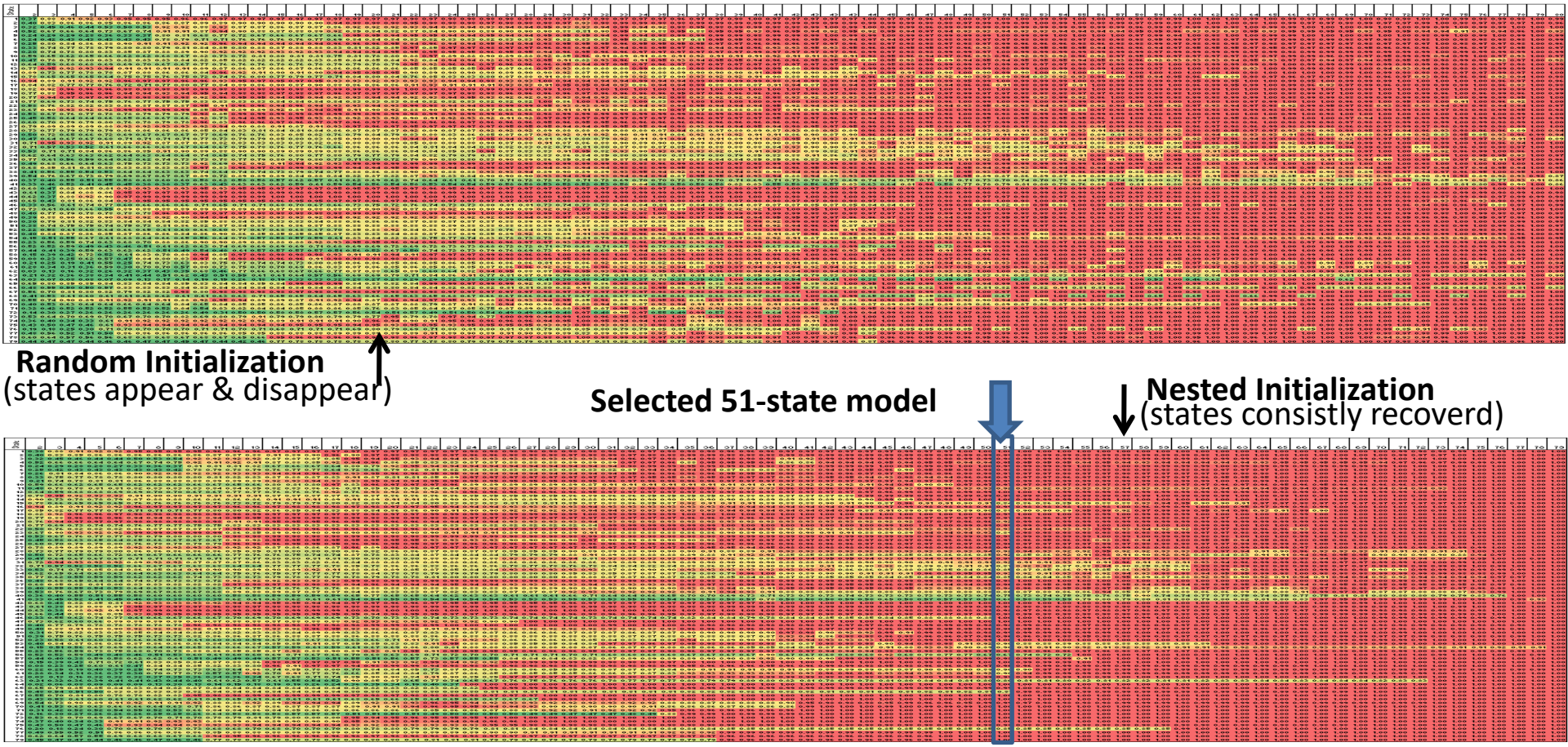
1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - HMM Foundations, Generating, Parsing, Decoding, Learning
   - ChromHMM: Multi-variate HMM for chromatin state learning
4. Model complexity: selecting the number of states/marks
   - Capturing dependencies. State-conditional mark independence
   - Selecting the number of states, selecting number of marks
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

Comparison of BIC Score vs. Number of States for Random and Nested Initialization



- Standard model selection criteria fail due to genome complexity: more states always preferred
- Instead: Start w/complex model, keep informative states, prune redundant states. Pick cutoff

Recovery of 79-state model in random vs. nested initialization



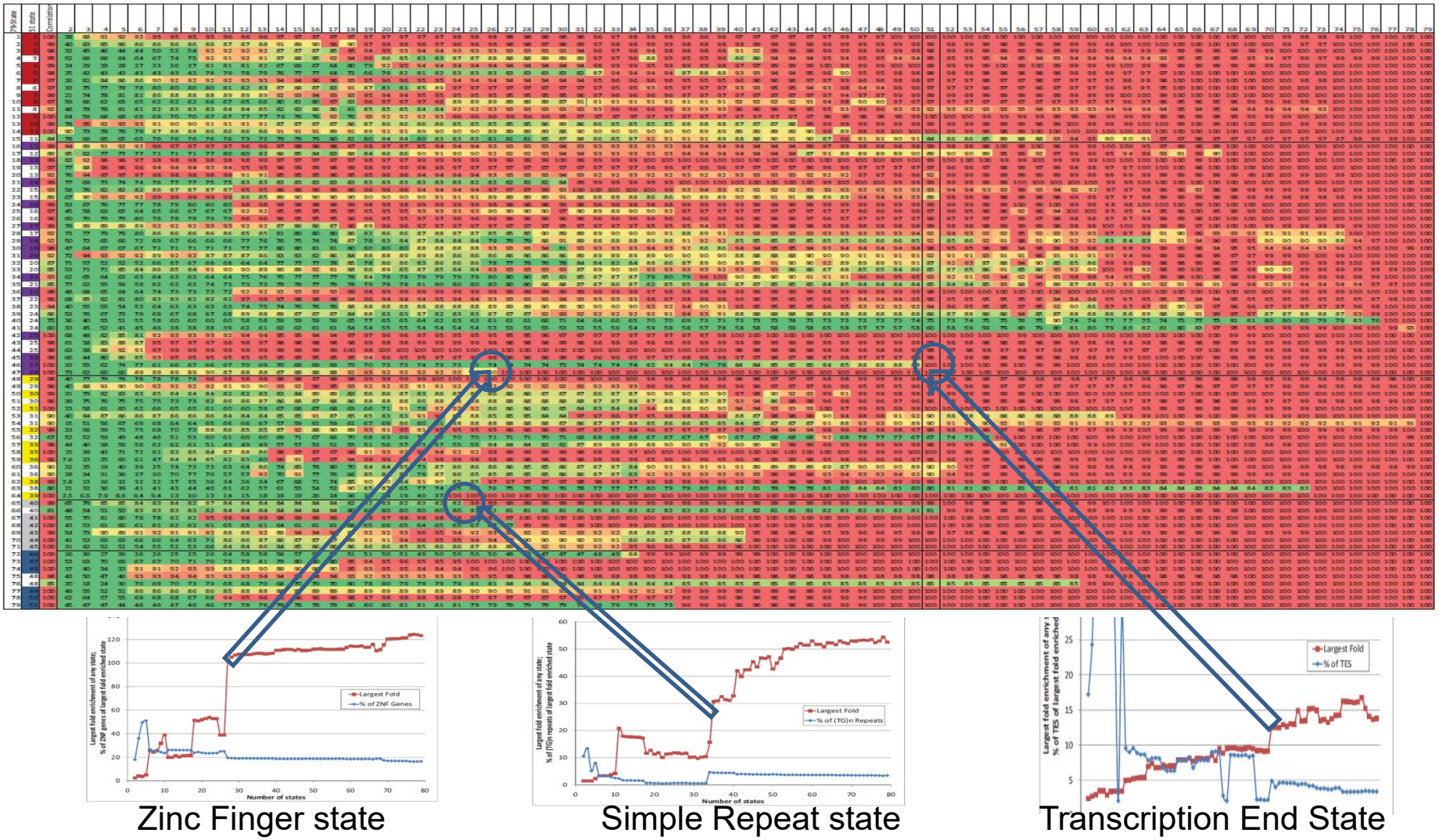
Nested initialization approach:

- **First pass:** learn models of increasing complexity
- **Second pass:** form nested set of emission parameter initializations by greedily removing states from best BIC model found

Nested models criteria:

- Maximize sum of correlation of emission vectors with nested model
- Models learned in parallel

Functional recovery with increasing numbers of states



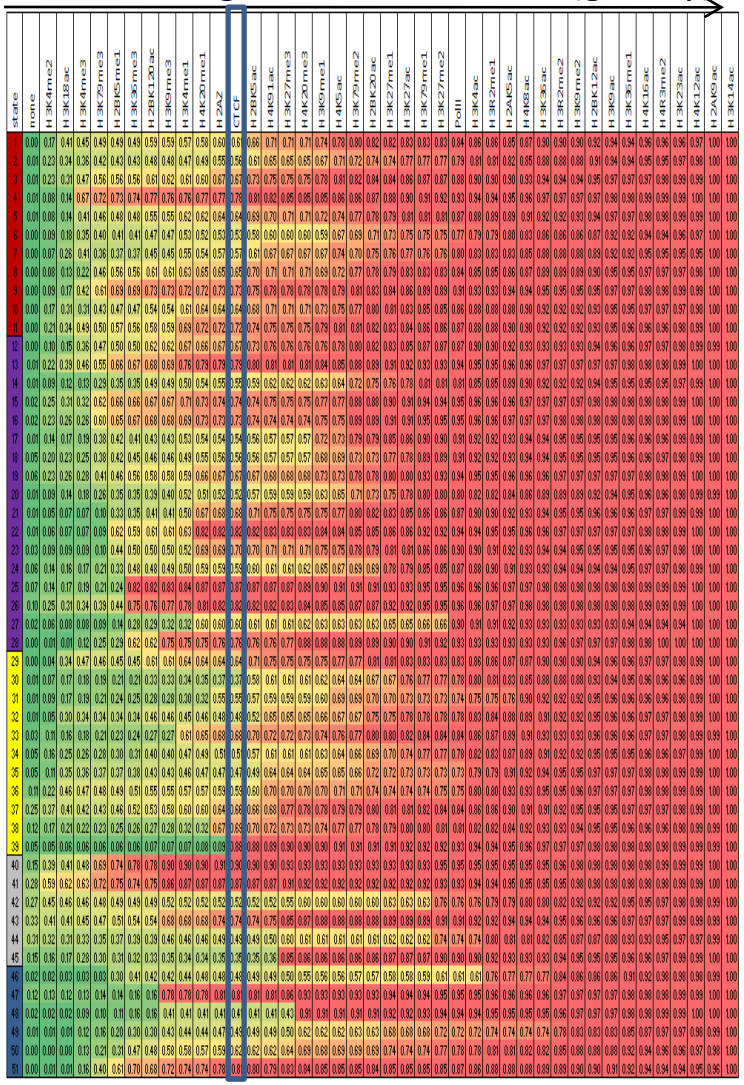
Zinc Finger state | Simple Repeat state | Transcription End State

- Red: Maximum fold functional enrichment for corresponding biological category
- Blue: Percent of that functional category that overlaps regions annotated to this state
- Top plot: Correlation of emission parameter vector for that state to closest state

Chromatin state recovery with increasing numbers of marks

Which states are well-recovered?

Increasing numbers of marks (greedy)



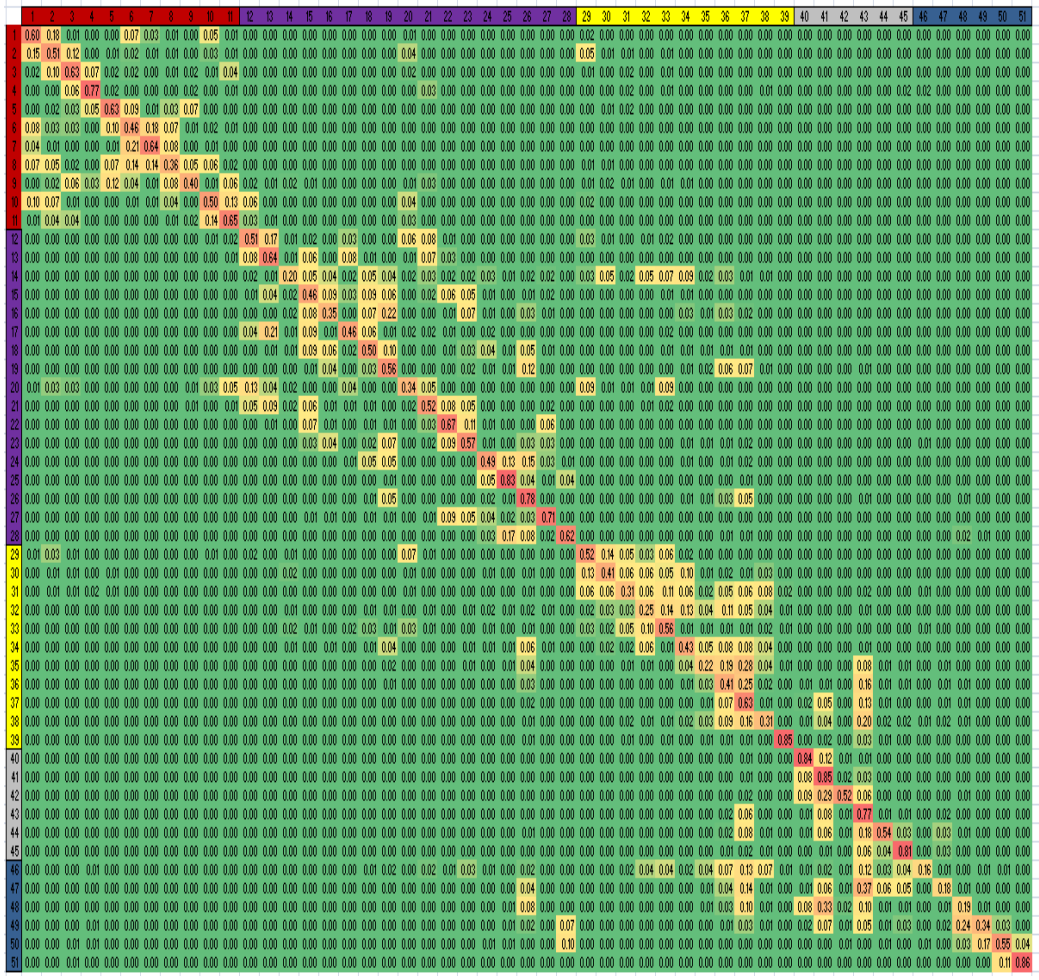
State Inferred with all 41 marks

Recovery of states with increasing number of marks

Precisely what mistakes are made?

(for a given subset of 11 ENCODE marks)

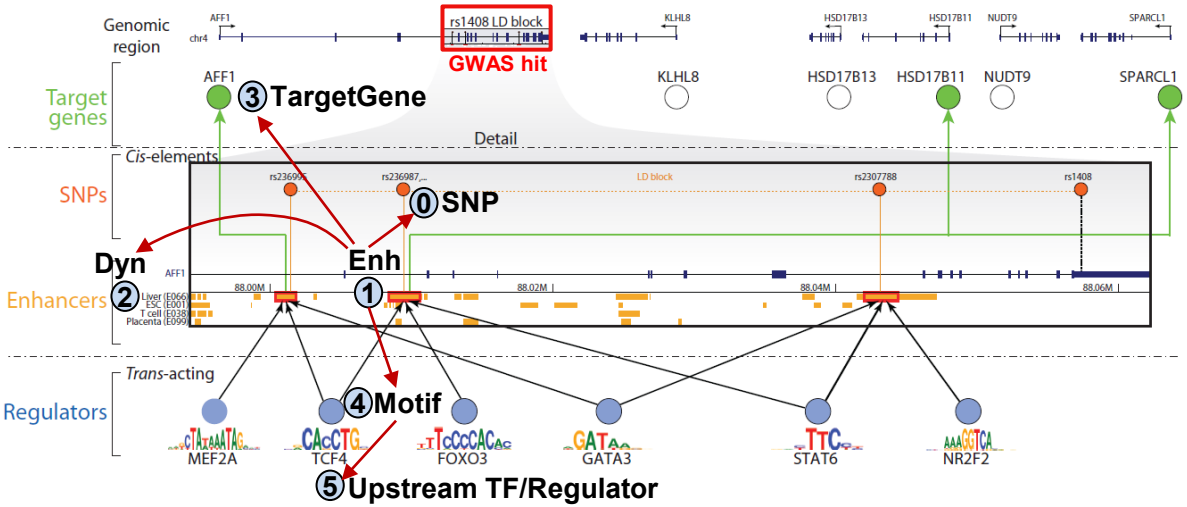
State Inferred with subset of marks



State Inferred with all 41 marks

State confusion matrix with 11 ENCODE marks

MIT 18.700/18.701/6.047/6.878/HST.507
Computational Biology: Genomes, Networks, Evolution



Lecture 5: Regulatory circuitry

Epigenome Dynamics: Joint Chromatin State Learning

Enhancer-gene linking: Correlation, Hi-C, eQTLs

TF motif discovery: Enrichment, EM, Gibbs Sampling

Deep learning convolution CNNs for motif discovery

Global motif discovery: Comparative Genomics

Motif Instance Identification: Branch Length Score

Regulatory region dissection: MPRA, HiDRA

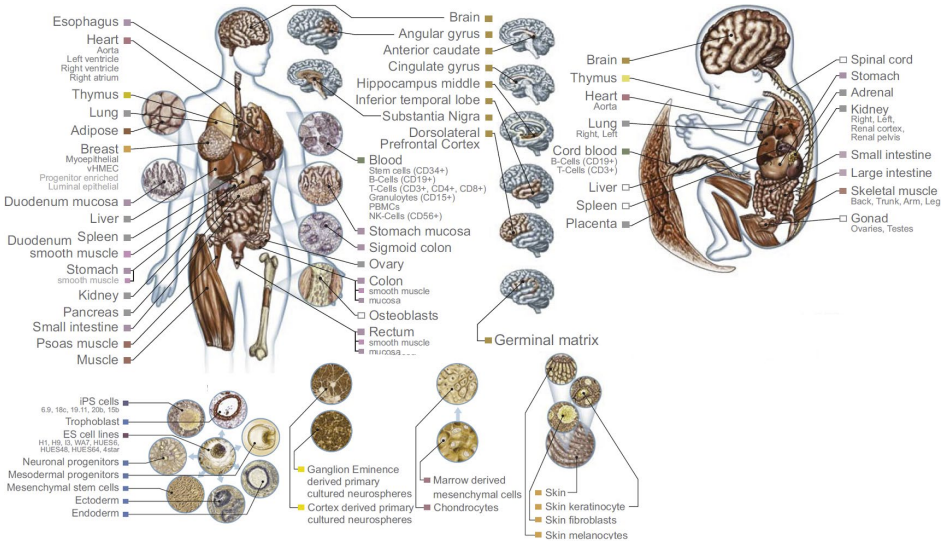
Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

Goals for today: Computational Epigenomics

1. Introduction to Epigenomics
   - Overview of epigenomics, Diversity of Chromatin modifications
   - Antibodies, ChIP-Seq, data generation projects, raw data
2. Primary data processing: Read mapping, Peak calling
   - Read mapping: Hashing, Suffix Trees, Burrows-Wheeler Transform
   - Quality Control, Cross-correlation, Peak calling, IDR (similar to FDR)
3. Discovery and characterization of chromatin states
   - HMM Foundations, Generating, Parsing, Decoding, Learning
   - ChromHMM: Multi-variate HMM for chromatin state learning
4. Model complexity: selecting the number of states/marks
   - Selecting the number of states, selecting number of marks
   - Capturing dependencies and state-conditional mark independence
5. Learning chromatin states jointly across multiple cell types
   - Stacking vs. concatenation approach for joint multi-cell type learning
   - Defining activity profiles for linking enhancer regulatory networks

Epigenomic mapping across 100+ tissues/cell types

Diverse tissues and cells



Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood…)

Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood…)

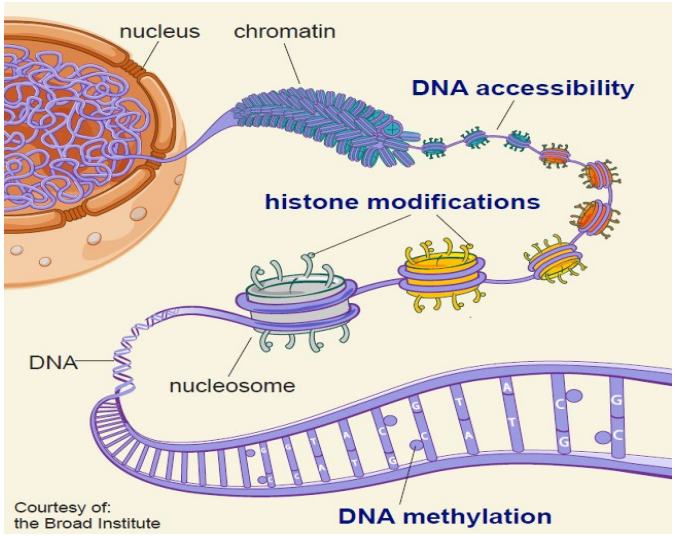
ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench…)

ROADMAP epigenomics PROJECT

ENCODE

X

Diverse epigenomic assays



Histone modifications
- **H3K4me3, H3K4me1, H3K36me3**
- **H3K27me3, H3K9me3, H3K27/9ac**
- **+20 more**

Open chromatin:
- **DNA accessibility**

DNA methylation:
- **WGBS, RRBS, MRE/MeDIP**

Gene expression
- **RNA-seq, Exon Arrays**

ENCODE: Study nine marks in nine human cell lines

9 marks

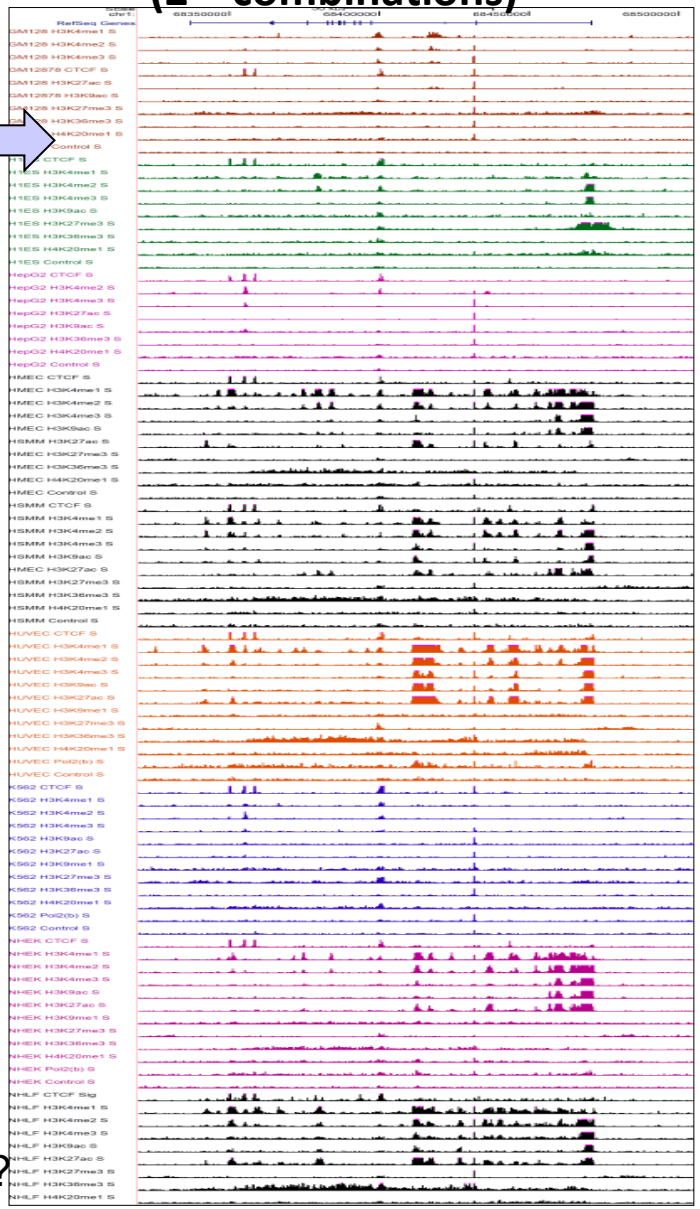
H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K9ac
H3K27me3
H4K20me1
H3K36me3
CTCF
+WCE
+RNA

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

81 Chromatin Mark Tracks
(2^{81} combinations)



Brad Bernstein ENCODE Chromatin Group

b.

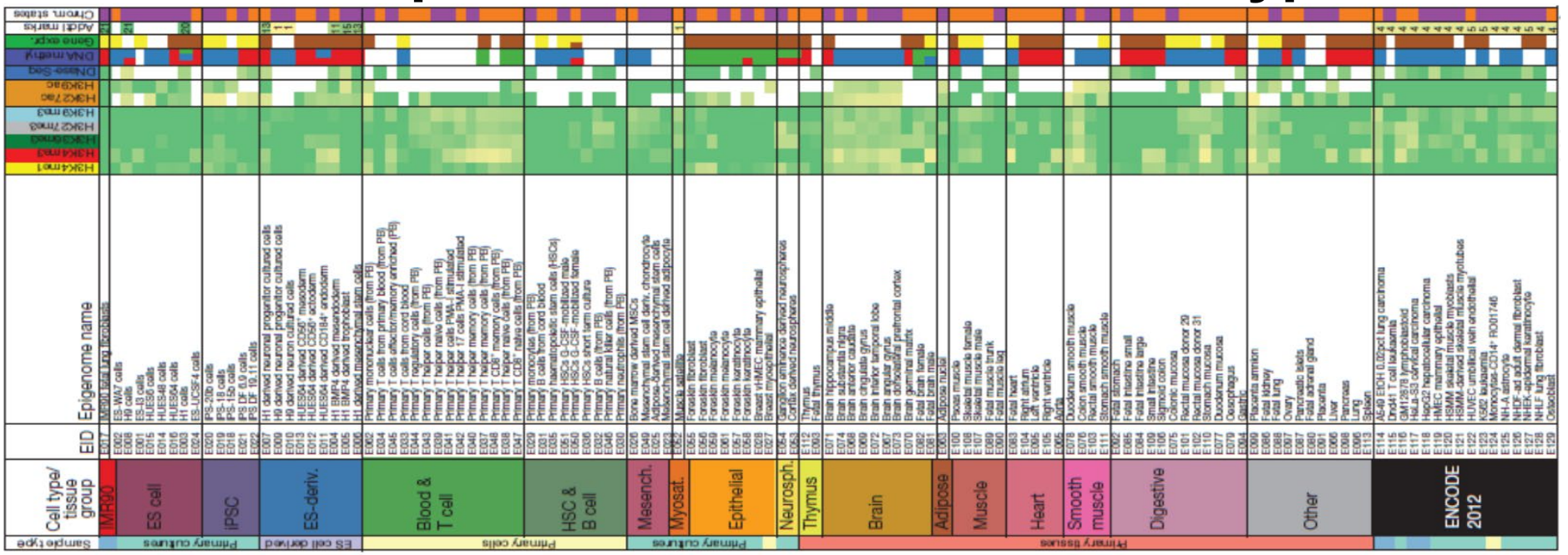
Chromatin States	State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE
	1	16	2	2	6	17	93	99	96	98	2
	2	12	2	6	9	53	94	95	14	44	1
	3	13	72	0	9	48	78	49	1	10	1
	4	11	1	15	11	96	99	75	97	86	4
	5	5	0	10	3	88	57	5	84	25	1
	6	7	1	1	3	58	75	8	6	5	1
	7	2	1	2	1	56	3	0	6	2	1
	8	92	2	1	3	6	3	0	0	1	1
	9	5	0	43	43	37	11	2	9	4	1
	10	1	0	47	3	0	0	0	0	0	1
	11	0	0	3	2	0	0	0	0	0	0
	12	1	27	0	2	0	0	0	0	0	0
	13	0	0	0	0	0	0	0	0	0	0
	14	22	28	19	41	6	5	26	5	13	37
	15	85	85	91	88	76	77	91	73	85	78

Chromatin Mark Observation Frequency (%)

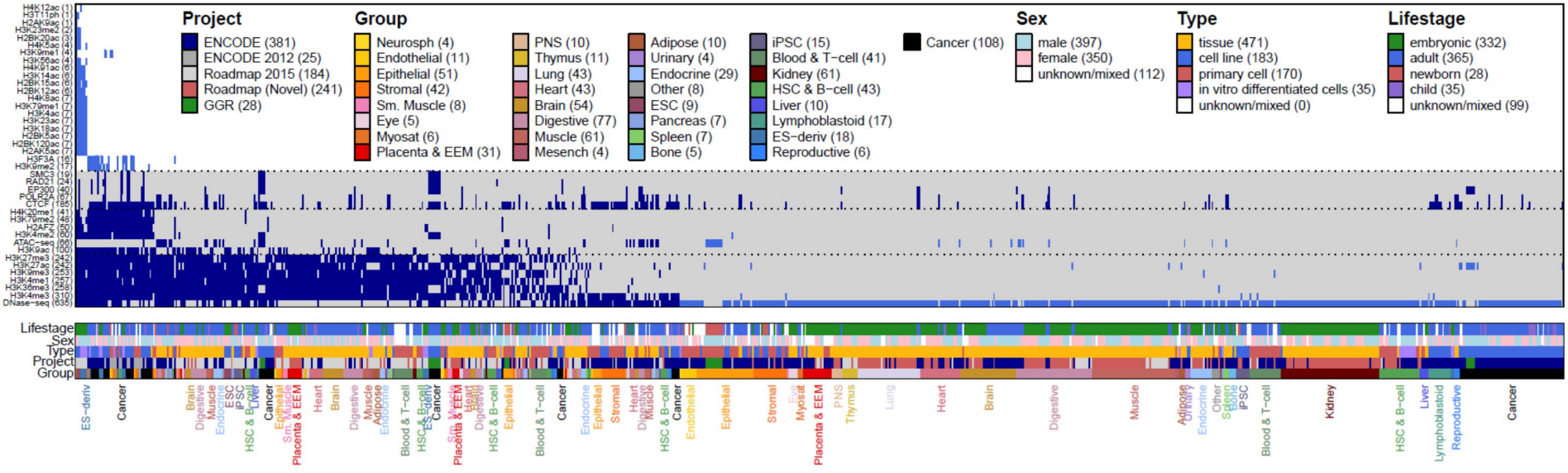
How to learn single set of chromatin states?

Ernst et al, *Nature* 2011

Roadmap 2015: 12 marks x 127 cell types



ENCODE 2019: 40 marks x 834 cell types



Solution 1: Learn independent models and cluster

state	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE
1	13.2	72.0	0.2	9.1	47.9	77.8	49.5	1.3	10.2	0.7
2	11.9	1.9	6.1	9.0	52.7	93.7	95.0	14.1	44.1	0.9
3	16.4	1.5	2.4	5.5	17.0	92.6	99.0	95.7	98.1	1.9
4	11.4	0.6	14.5	11.3	96.3	99.3	75.1	97.2	85.7	3.7
5	5.3	0.2	9.5	2.6	88.1	56.8	5.3	84.4	24.9	1.5
6	6.7	0.9	1.0	3.2	58.3	74.7	8.4	5.8	5.4	0.8
7	1.6	0.6	1.6	1.3	56.5	2.7	0.4	5.9	1.6	0.6
8	91.5	1.8	0.9	2.8	6.3	3.3	0.4	0.5	1.0	0.8
9	4.6	0.3	43.2	43.1	36.5	11.5	1.9	9.1	3.9	1.3
10	1.2	0.1	47.2	2.7	0.4	0.0	0.1	0.3	0.3	0.5
11	0.4	0.1	2.7	1.7	0.2	0.1	0.1	0.2	0.3	0.4
12	0.9	26.8	0.0	2.1	0.4	0.1	0.1	0.1	0.1	0.4
13	0.2	0.4	0.0	0.1	0.1	0.0	0.0	0.0	0.1	0.1
14	21.9	27.9	19.1	41.0	5.7	4.8	25.9	5.3	13.1	37.5
15	85.2	85.1	91.5	88.4	76.2	76.9	90.8	72.7	85.0	78.3

Promoter

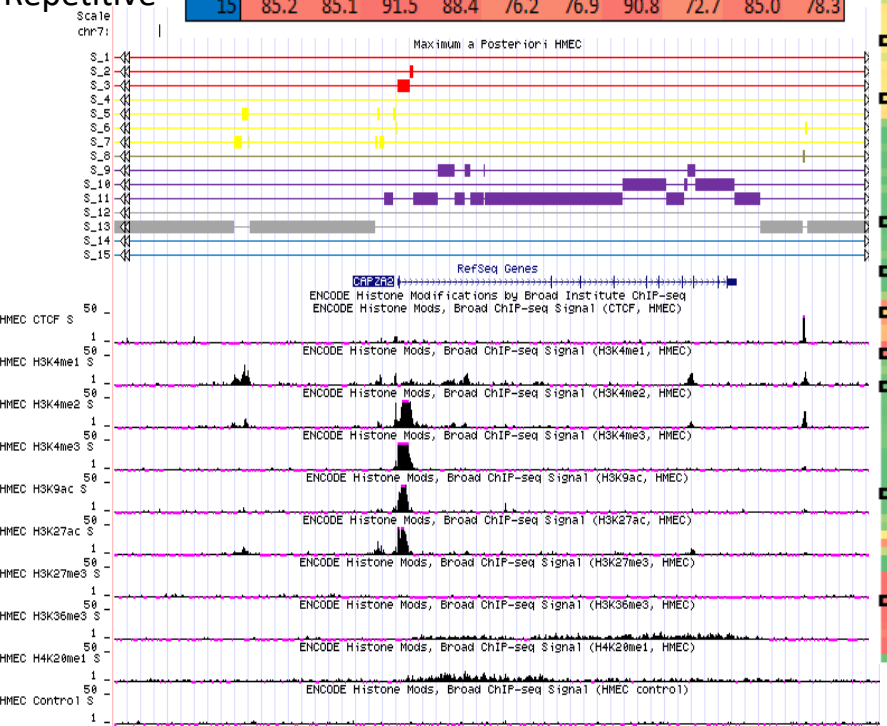
Candidate enhancer

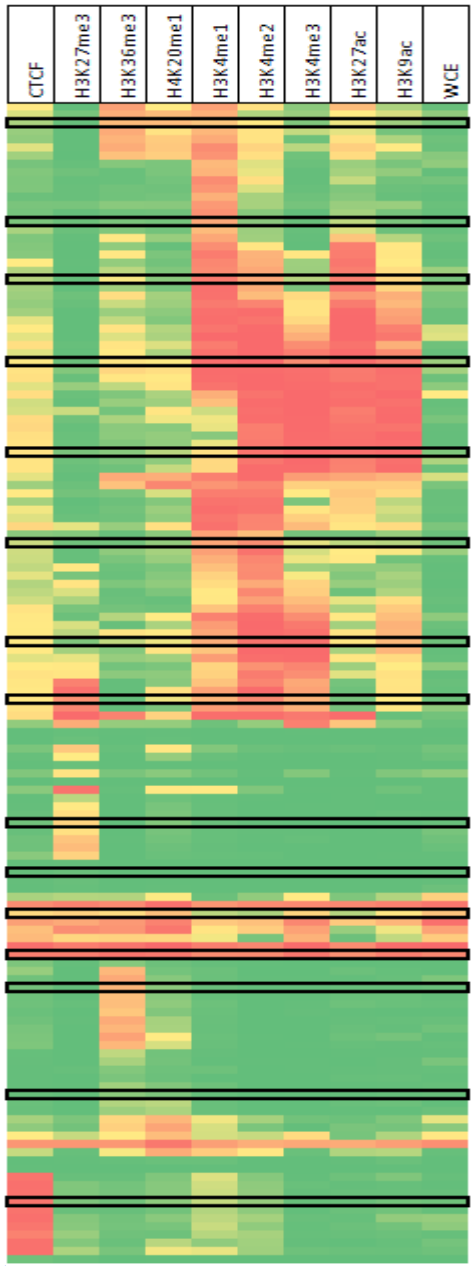
Insulator

Transcribed

Repressive

Repetitive





Basic approach:

a) Train a k-state model in each cell type independently

b) Cluster models learned independently

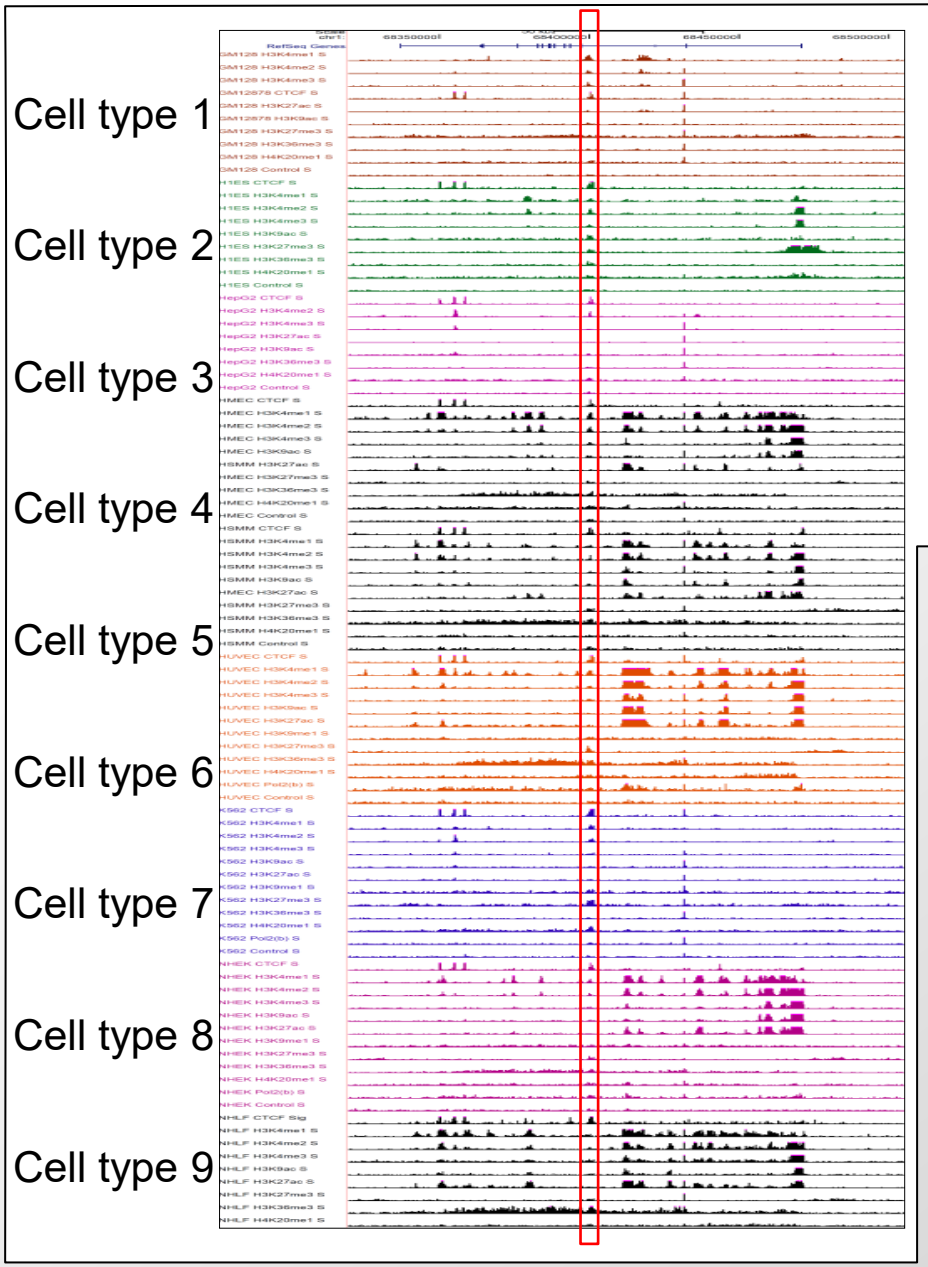
c) Merge clusters and re-apply to each cell type

How to cluster

a) Using emission probability matrix: most similar definitions

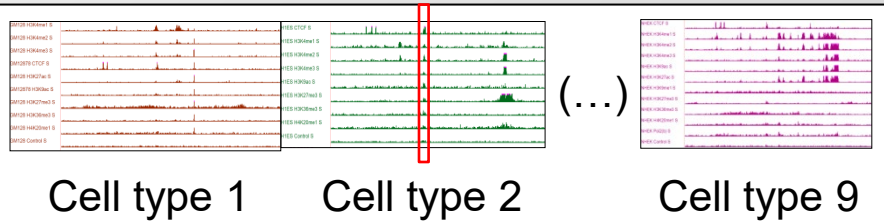
b) Using genome annotation: posterior probability decoding

Joint learning of states across multiple cell types



Solution 2: Stacking

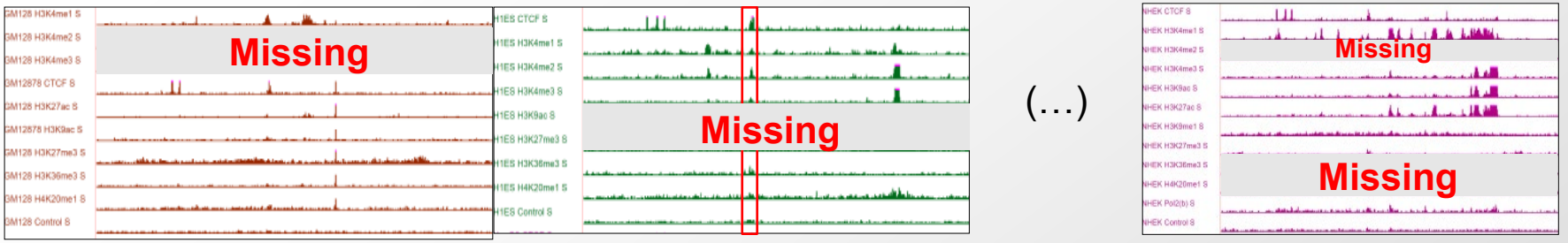
- Learns each combination of activity as a separate state
- Ex: ES-specific enhancers: enhancer marks in ES, no marks in other cell types



Solution 3: Concatenation

- Requires that profiled marks are the same (or treat as missing data)
- Ensures common state definitions across cell types

Joint learning with different subsets of marks (Solution 3)



Cell type 1 Cell type 2 Cell type 9

Option (a) Treat missing tracks as missing data

- EM framework allows for unspecified data points
- As long as pairwise relationship observed in some cell type

Option (b) Chromatin mark imputation

- Explicitly predict max-likelihood chromatin track for missing data
- Less powerful if ultimate goal is chromatin state learning

ENCODE: Study nine marks in nine human cell lines

9 marks

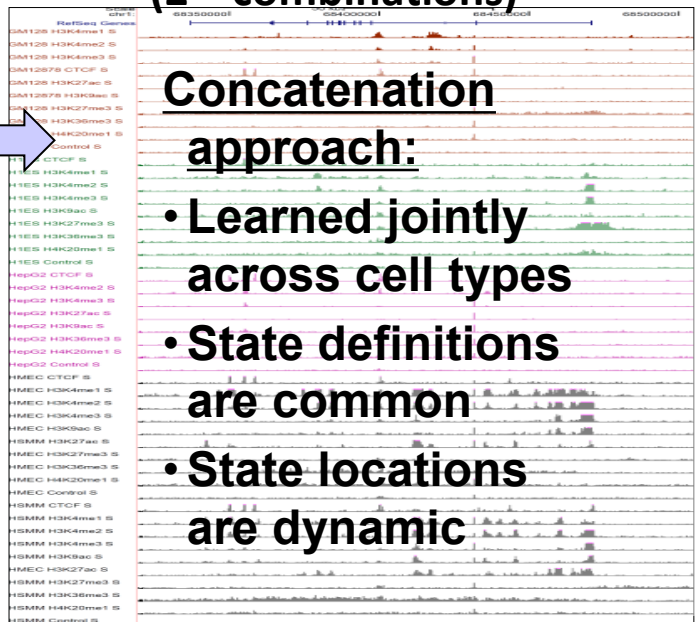
H3K4me1
H3K4me2
H3K4me3
H3K27ac
H3K9ac
H3K27me3
H4K20me1
H3K36me3
CTCF
+WCE
+RNA

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

81 Chromatin Mark Tracks (2^{81} combinations)



Concatenation approach:

- Learned jointly across cell types
- State definitions are common
- State locations are dynamic

Brad Bernstein ENCODE Chromatin Group

b.

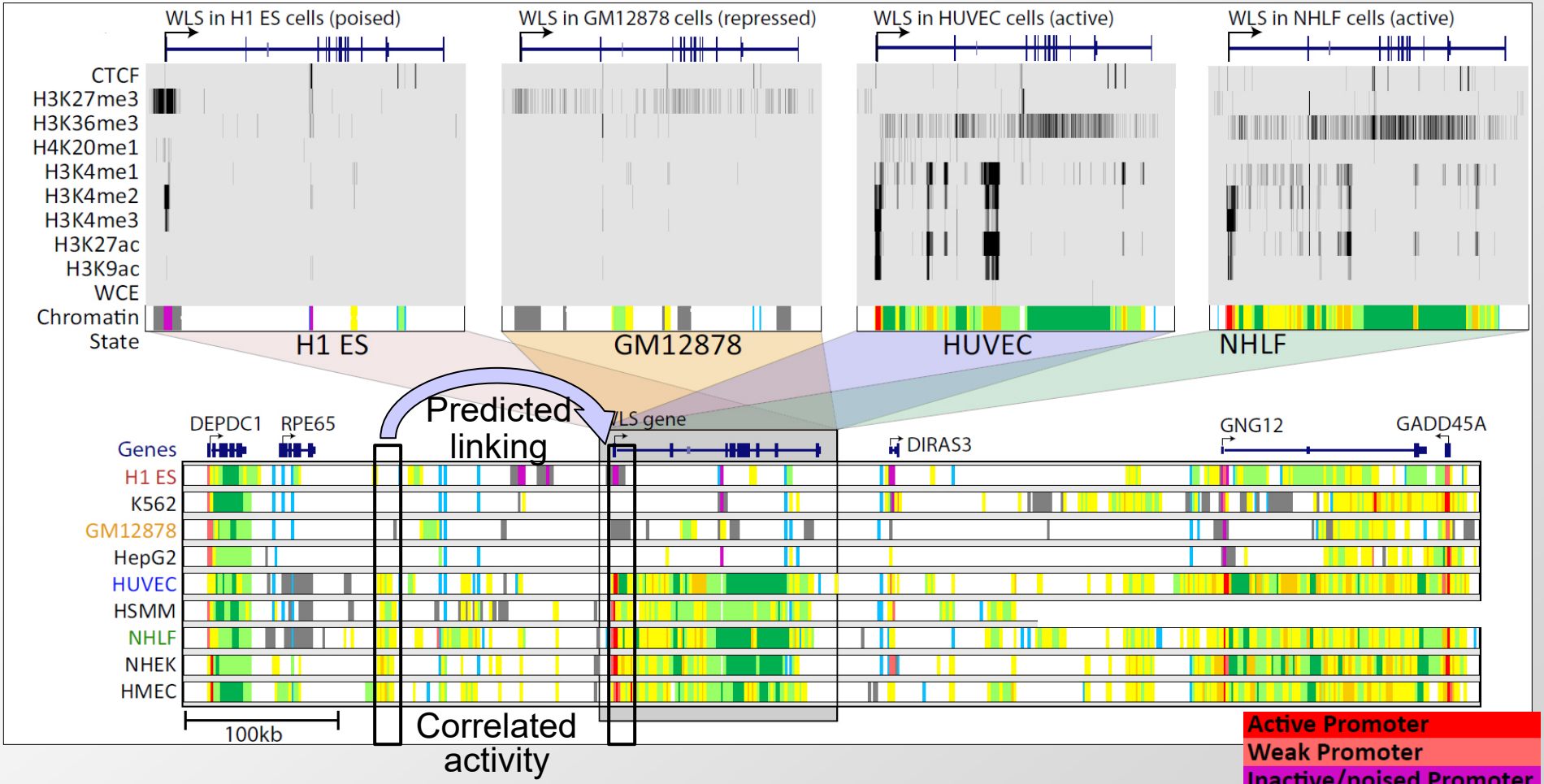
State	CTCF	H3K27me3	H3K36me3	H4K20me1	H3K4me1	H3K4me2	H3K4me3	H3K27ac	H3K9ac	WCE	Coverage: Median	Coverage: H1 ES	Coverage: GM	Median Length	+/-2kb TSS	Conserved non-exon	DNase (K562)	C-Myc (K562)	NF-κB (GM12878)	Transcript	Nuclear Lamina (NHLF)	Candidate state annotation
1	16	2	2	6	17	93	99	96	98	2	0.6	0.5	1.2	1.0	83	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	13.6	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 States

Chromatin Mark Observation Frequency (%)

(%) (fold) (kb) (%) Functional enrichments (fold)

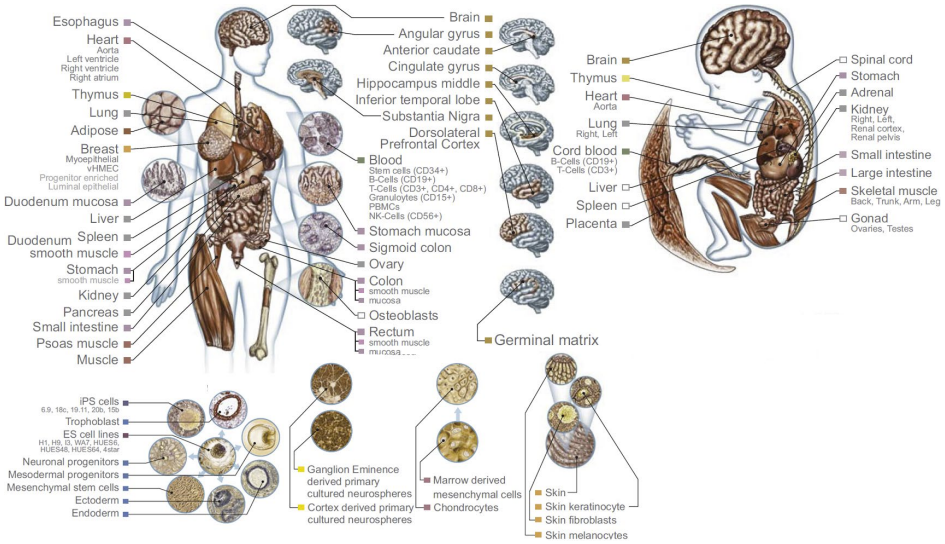
Chromatin states dynamics across nine cell types



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

Epigenomic mapping across 100+ tissues/cell types

Diverse tissues and cells



Adult tissues and cells (brain, muscle, heart, digestive, skin, adipose, lung, blood…)

Fetal tissues (brain, skeletal muscle, heart, digestive, lung, cord blood…)

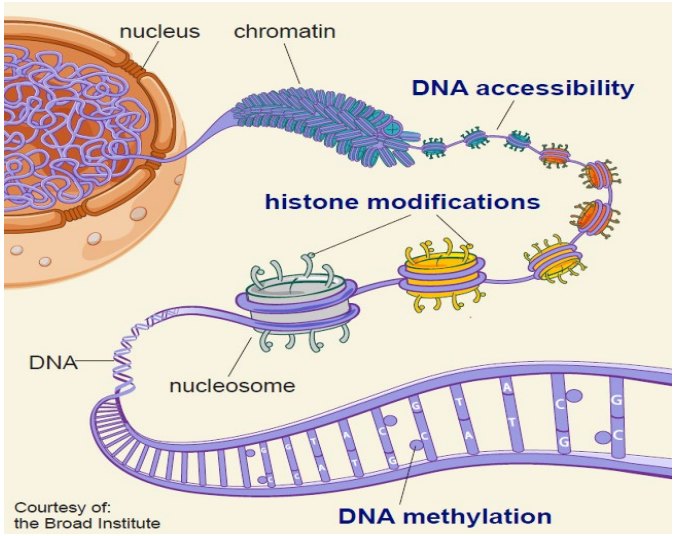
ES cells, iPS, differentiated cells (meso/endo/ectoderm, neural, mesench…)

ROADMAP epigenomics PROJECT

ENCODE

X

Diverse epigenomic assays



Histone modifications
- **H3K4me3, H3K4me1, H3K36me3**
- **H3K27me3, H3K9me3, H3K27/9ac**
- **+20 more**

Open chromatin:
- **DNA accessibility**

DNA methylation:
- **WGBS, RRBS, MRE/MeDIP**

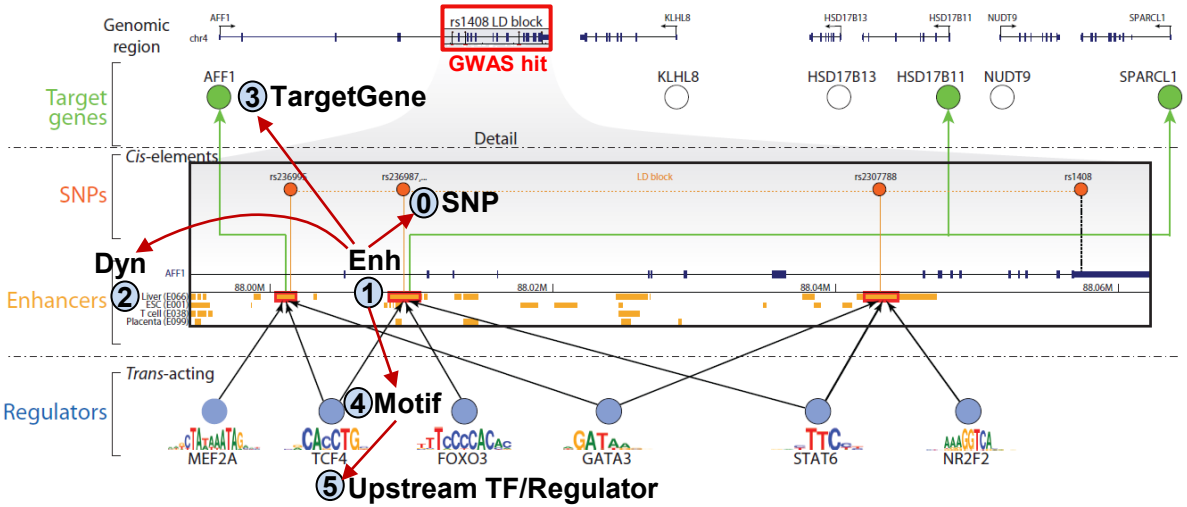
Gene expression
- **RNA-seq, Exon Arrays**

Chromatin state annotations across 127 epigenomes

Reveal epigenomic variability: enh/prom/tx/repr/het

Anshul Kundaje

MIT 18.700/18.701/6.047/6.878/HST.507
Computational Biology: Genomes, Networks, Evolution



Lecture 5: Regulatory circuitry

Epigenome Dynamics: Joint Chromatin State Learning

Enhancer-gene linking: Correlation, Hi-C, eQTLs

TF motif discovery: Enrichment, EM, Gibbs Sampling

Deep learning convolution CNNs for motif discovery

Global motif discovery: Comparative Genomics

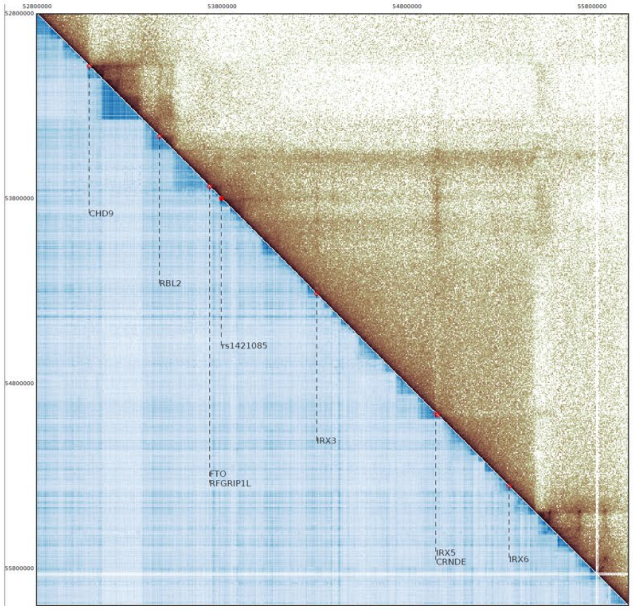
Motif Instance Identification: Branch Length Score

Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

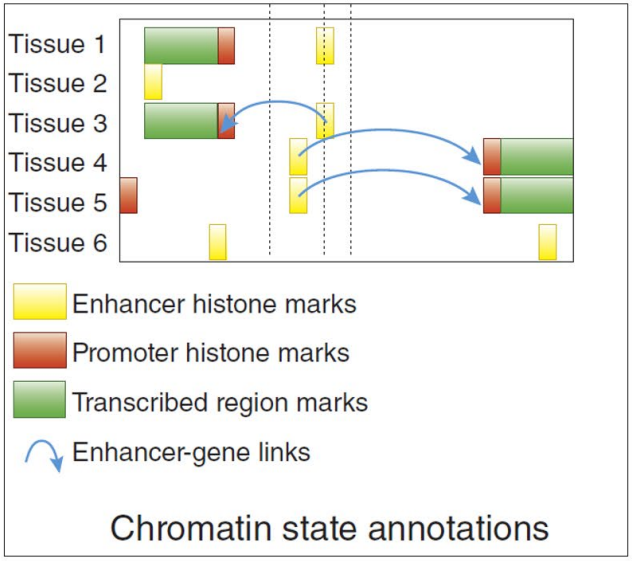
Link enhancers to their target genes 3 lines of evidence:

Physical



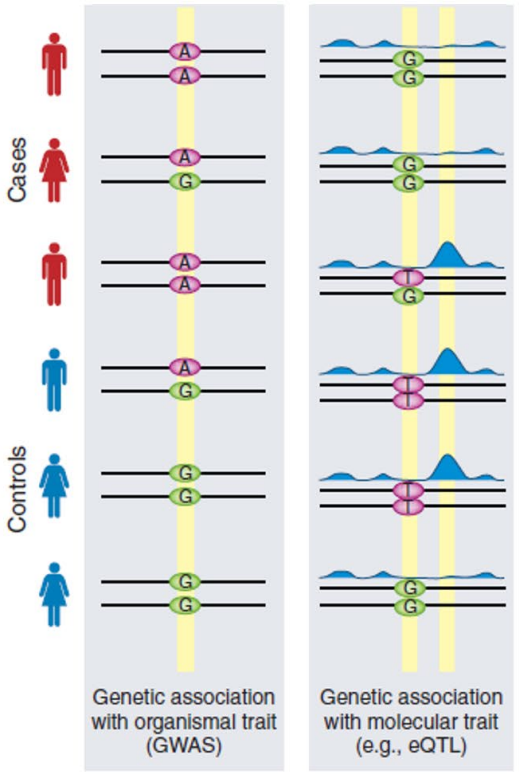
Hi-C: Physical proximity in 3D

Functional



Enhancer-gene activity correlation

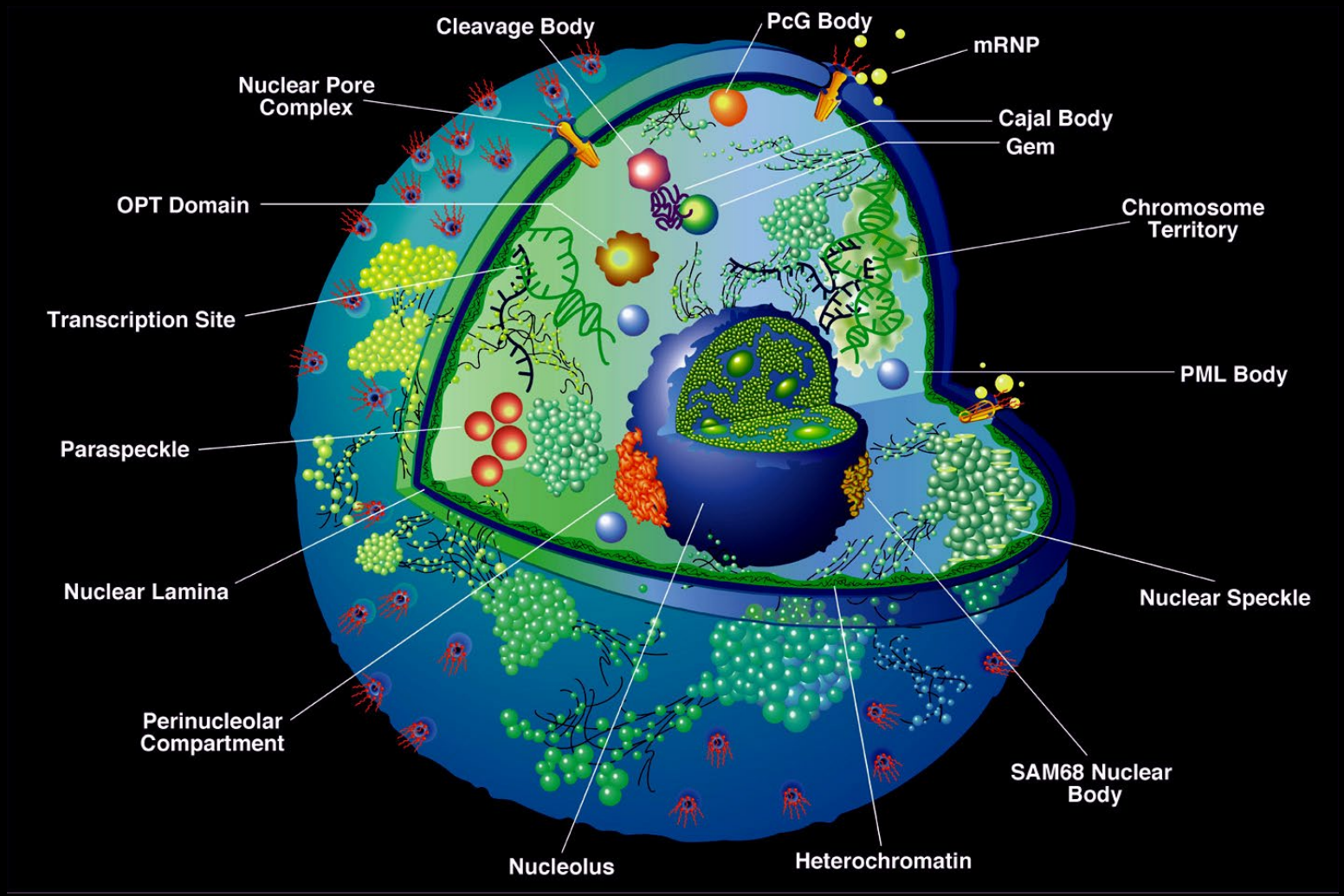
Genetic



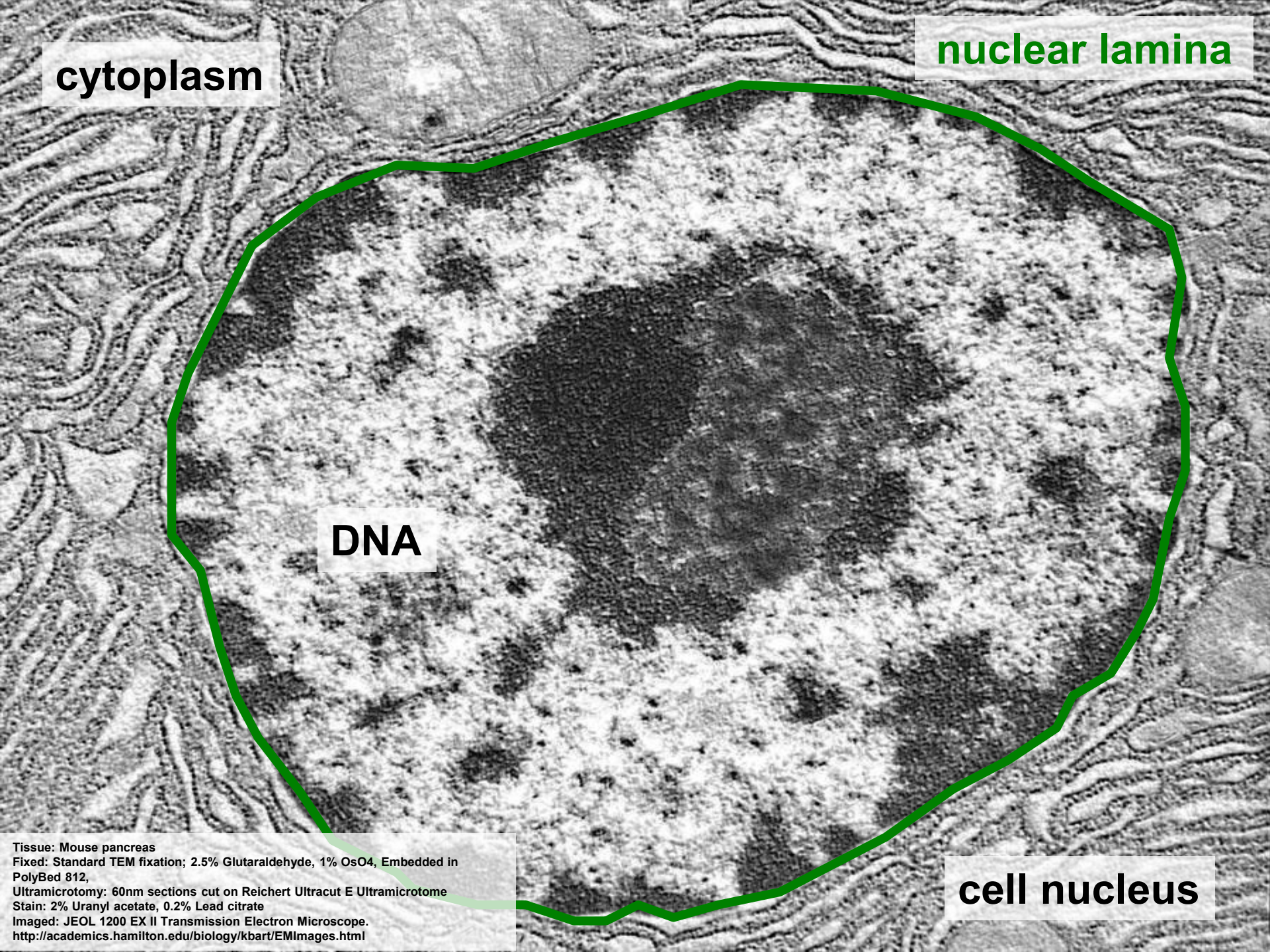
eQTL evidence: SNP effect on expression

Complementary evidence at physical, functional, genetic level

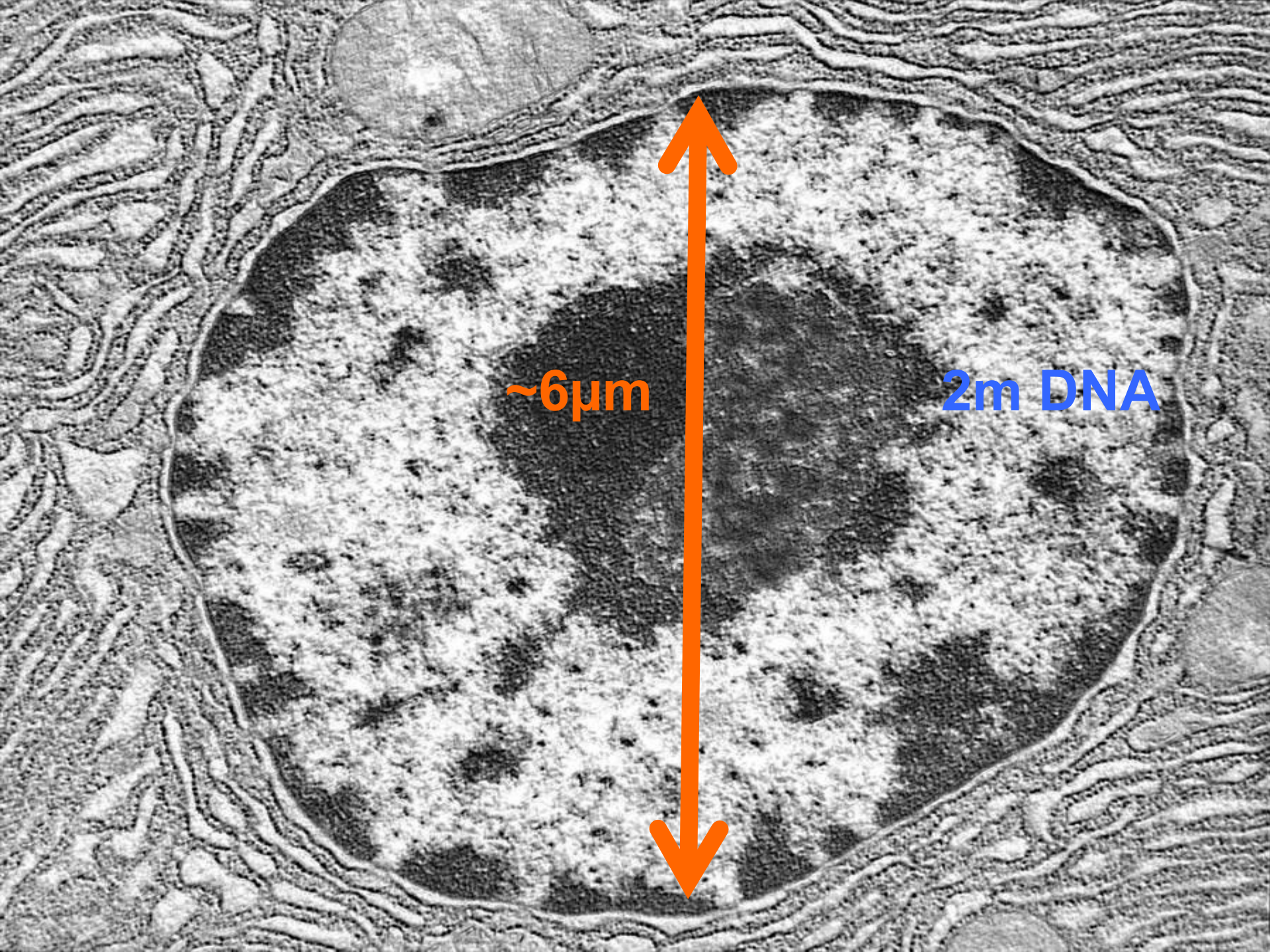
A model of the (mammalian) nucleus



Spector, D.L. 2001. Nuclear domains. J. Cell Science 114, 2891-2893.

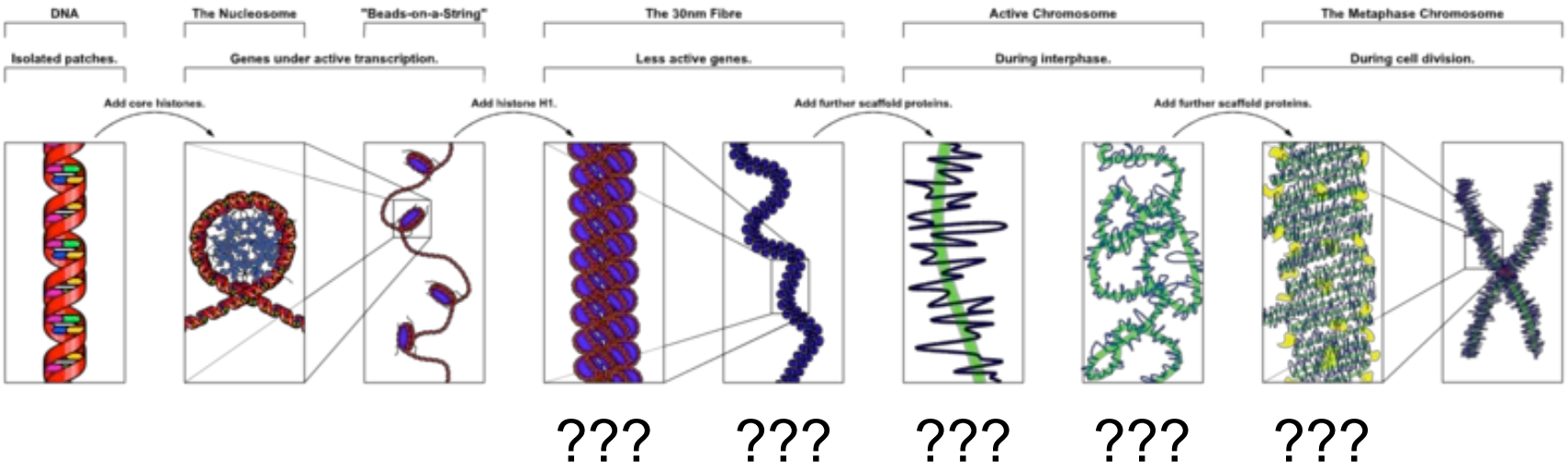


Tissue: Mouse pancreas
Fixed: Standard TEM fixation; 2.5% Glutaraldehyde, 1% OsO4, Embedded in PolyBed 812,
Ultramicrotomy: 60nm sections cut on Reichert Ultracut E Ultramicrotome
Stain: 2% Uranyl acetate, 0.2% Lead citrate
Imaged: JEOL 1200 EX II Transmission Electron Microscope.
http://academics.hamilton.edu/biology/kbart/EMImages.html

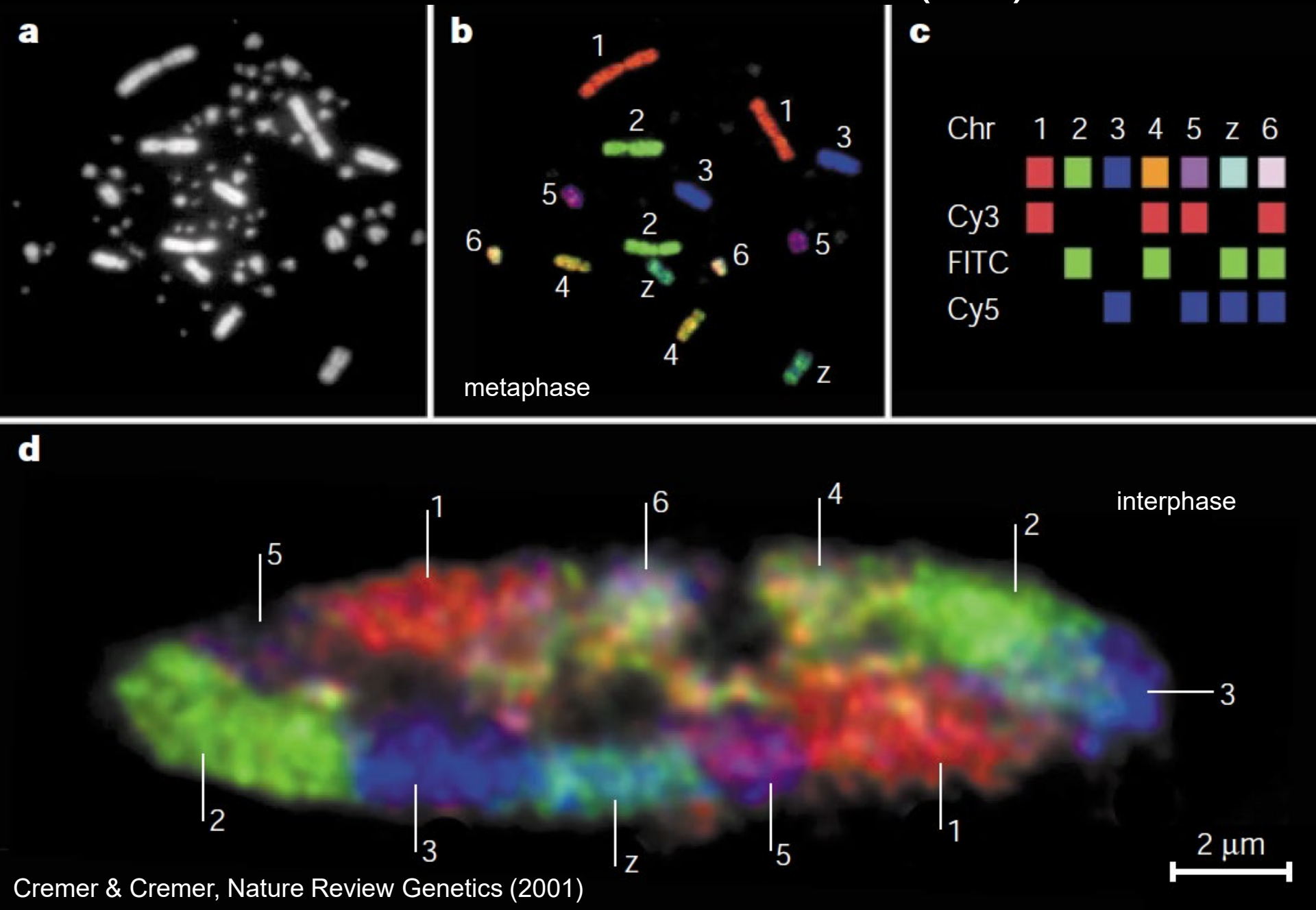

~6µm
2m DNA

DNA compaction

- DNA is **locally** compacted using *histone octamers* to form *nucleosomes*
- DNA is **globally** compacted by way of *chromosomes* (at least, during cell division / mitosis)
- Intermediate packaging mechanisms are subject of heavy speculation



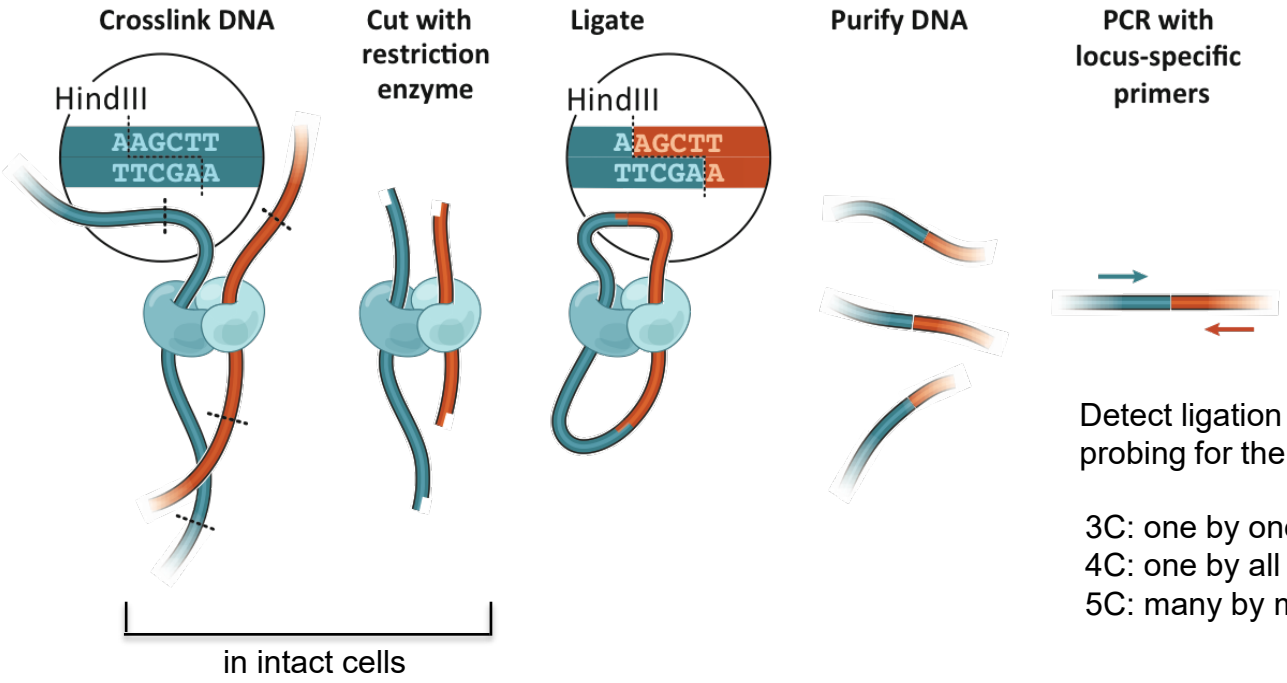
Chromosome territories (CT)



Cremer & Cremer, Nature Review Genetics (2001)

3C: Chromosome Conformation Capture

- Detects physical interactions between genomic elements
- Interacting elements are converted into ***ligation products***



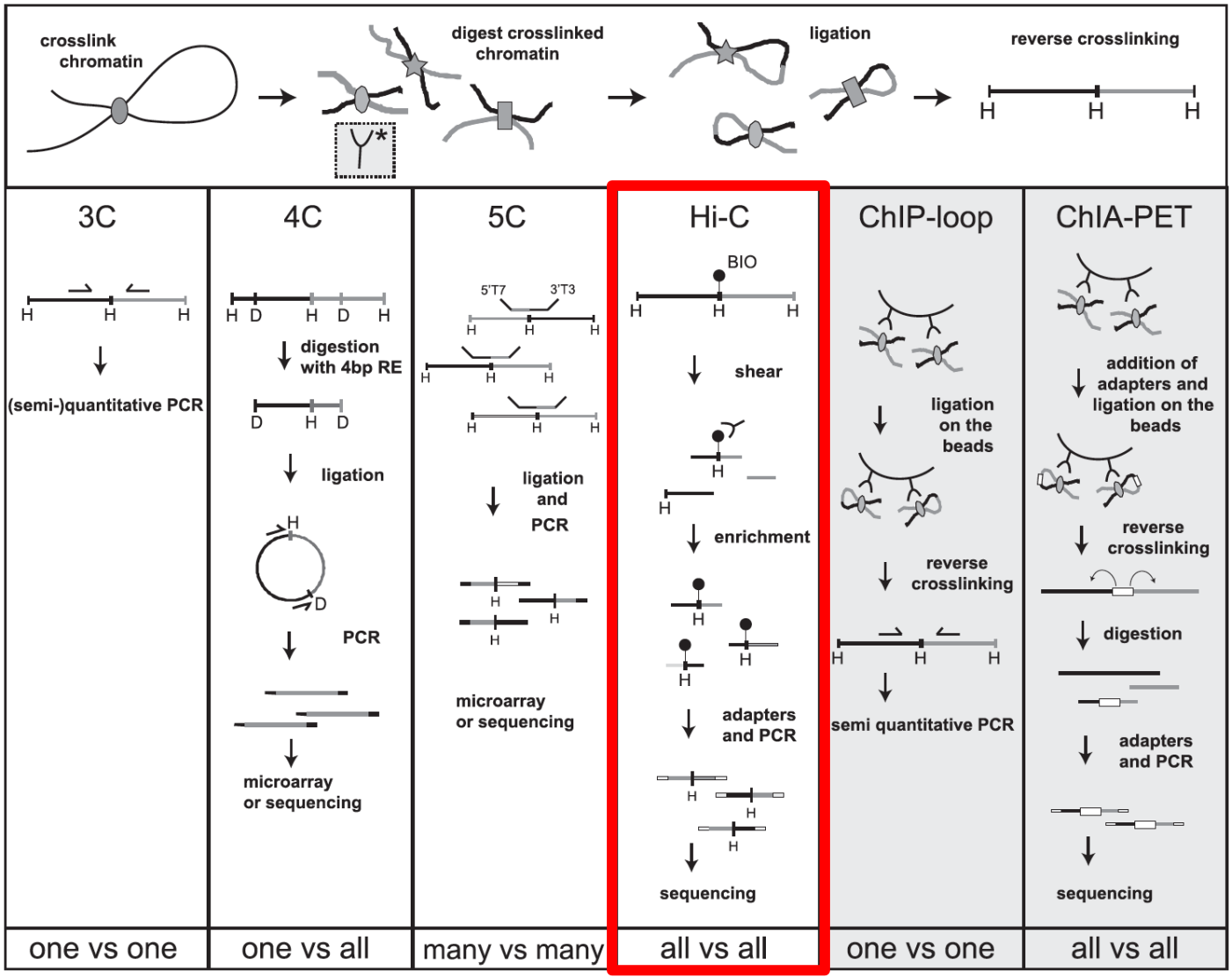
Detect ligation events by probing for them:

3C: one by one (PCR)
4C: one by all (microarray or seq)
5C: many by many (microarray or seq)

Dekker *et al.* Science 2002

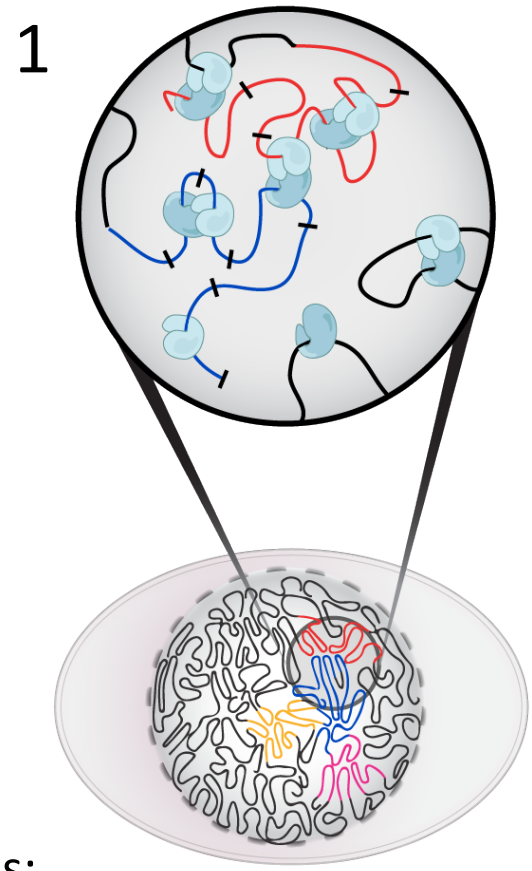
Dostie *et al*. Genome Res. 2006

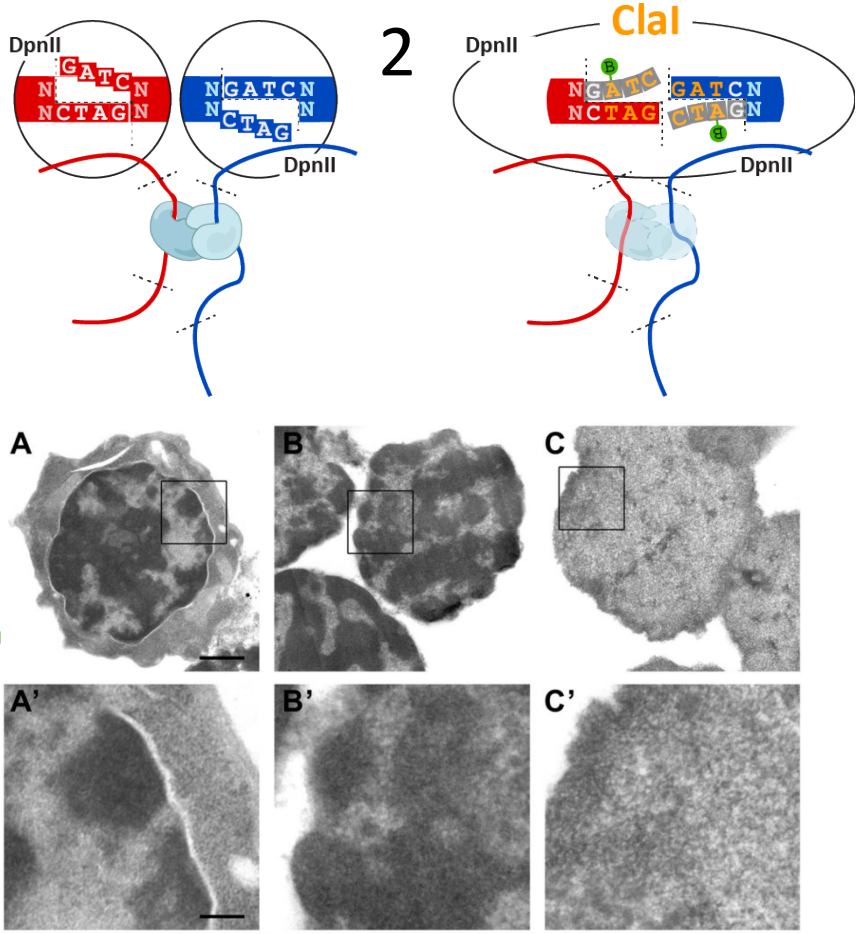
Chromosome Conformation Capturing (3C) based methods



de Wit & de Laat, Genes & Dev. (2012)

Hi-C vs. What-you-See



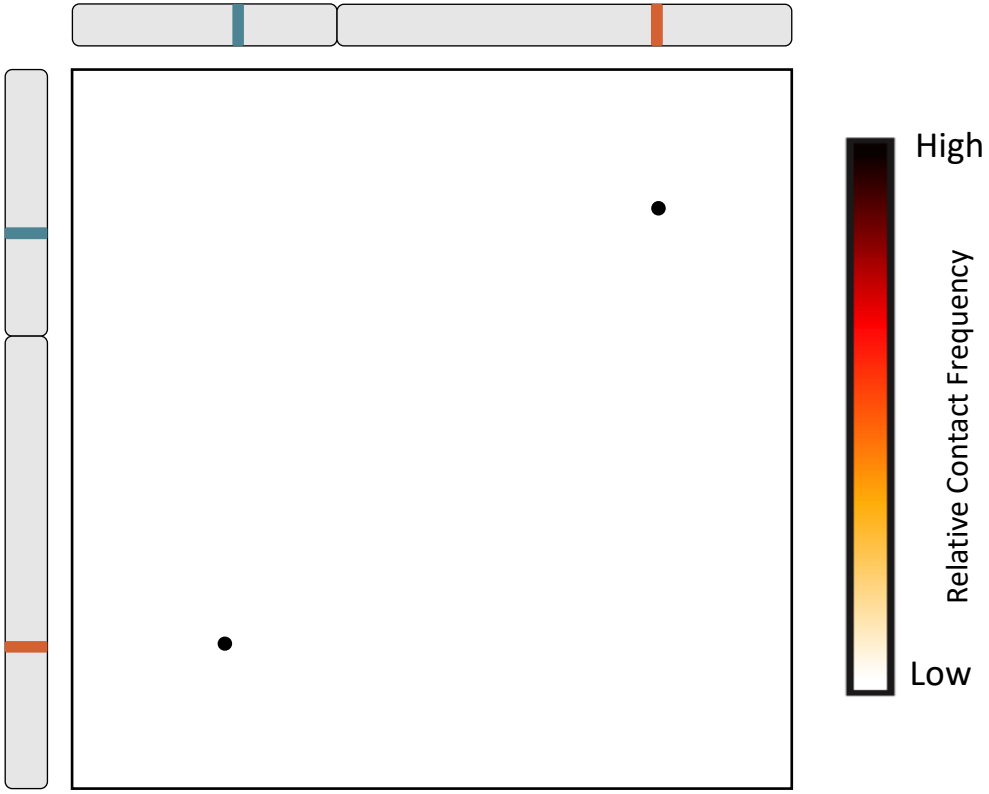


Critical steps:

1. Crosslinking to fix conformation
2. Digestion and re-ligation
3. Sequencing (biotinylated) junctions

Lieberman-Aiden et al., *Science* 2009
Belaghzal et al., *Nat Meth* 2017

Hi-C: genome-wide 3C

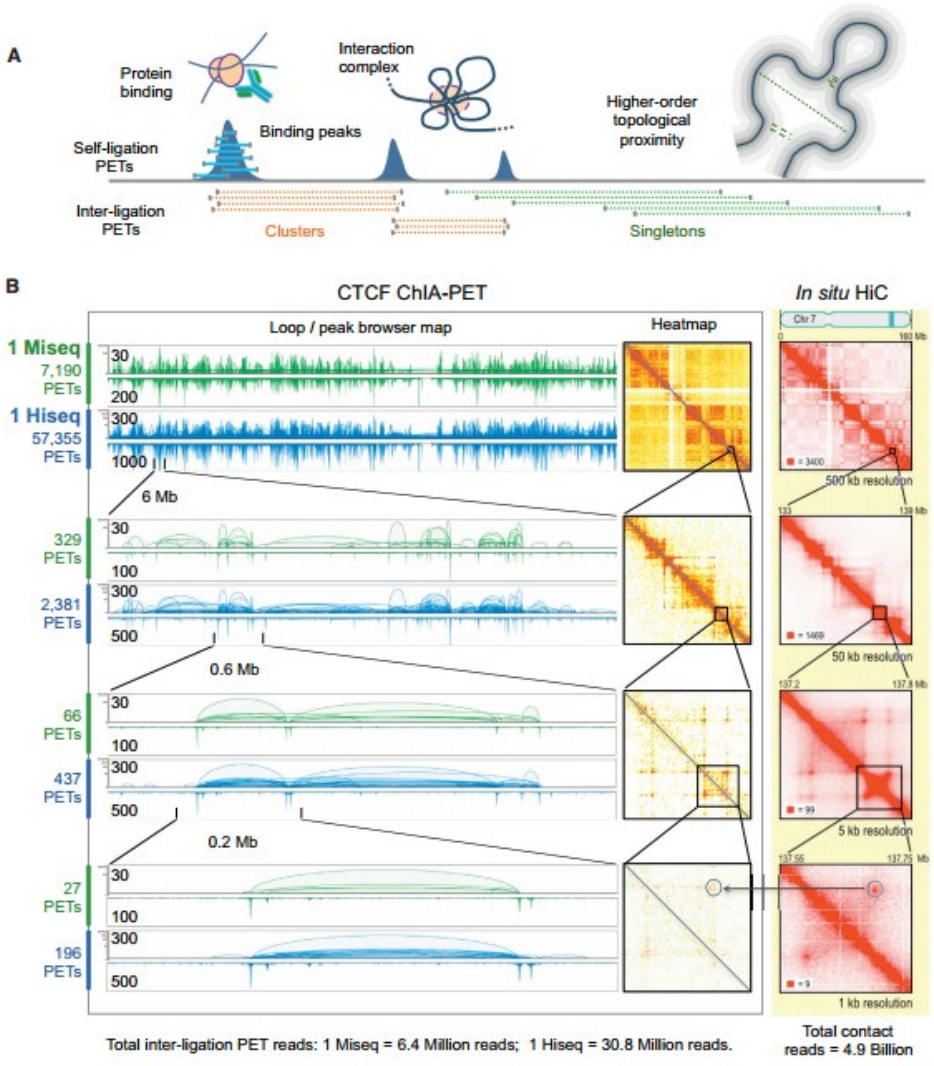


Territoriality

Kerpedjiev P, **Abdennur N**, Lekschas F, McCallum C, Dinkla K, Strobelt H, Luber JM, Ouellette SB, Azhir A, Kumar N, Hwang J, Lee S, Alver BH, Pfister H, Mirny LA, Park PJ, Gehlenborg N. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biol.* 2018 Aug 24;19(1):125.

ChIA-PET: Chromatin Interaction Analysis using Paired-End-Tag sequencing

1. self-ligation peaks: binding sites
2. Inter-ligation: long range interaction
3. Consistence between CTCF ChIA-PET and Hi-C
4. ChIA-PET has higher resolution than Hi-C



Correlation-based links of enhancer networks

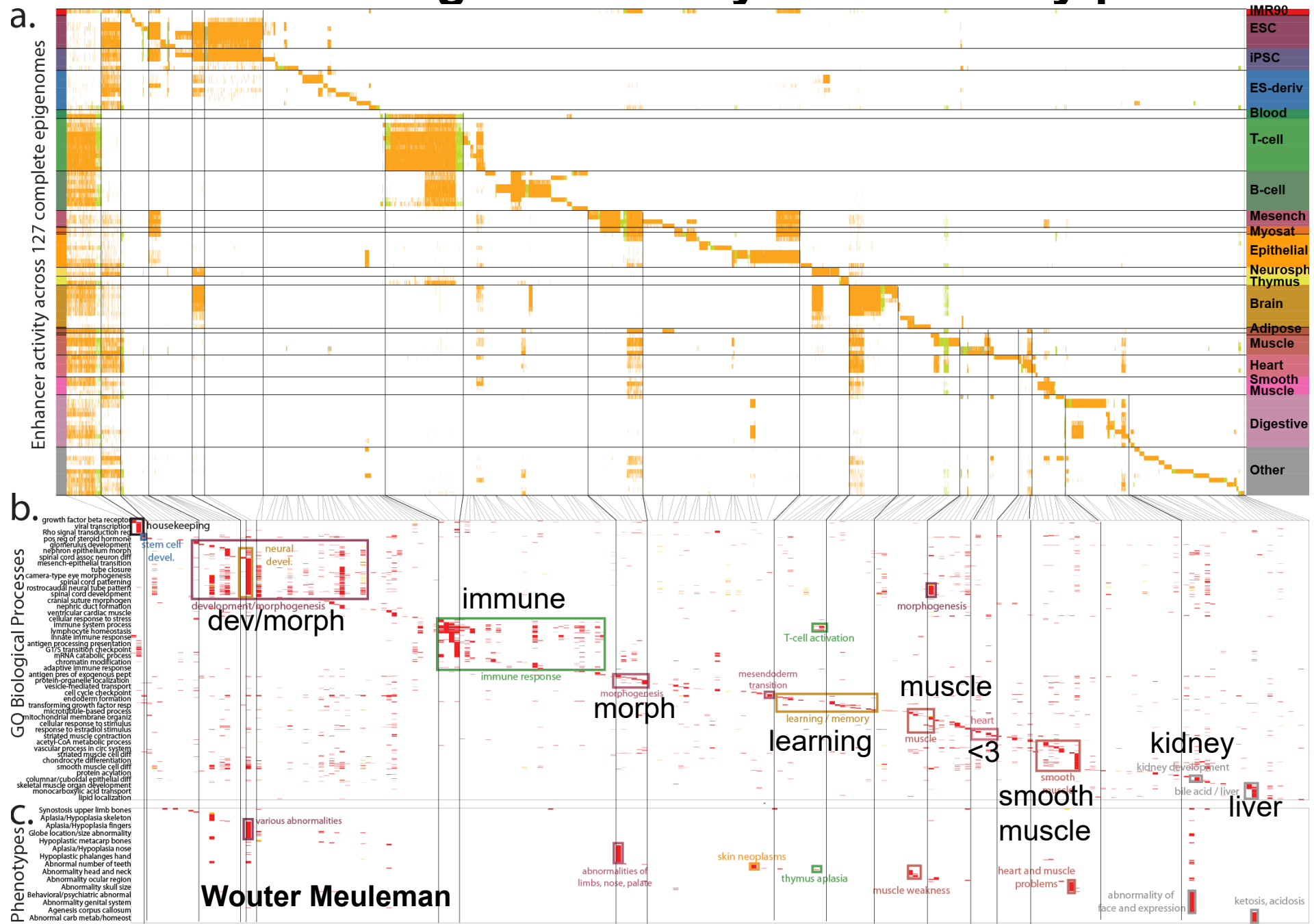
Regulators → Enhancers → Target genes

Chromatin state annotations across 127 epigenomes

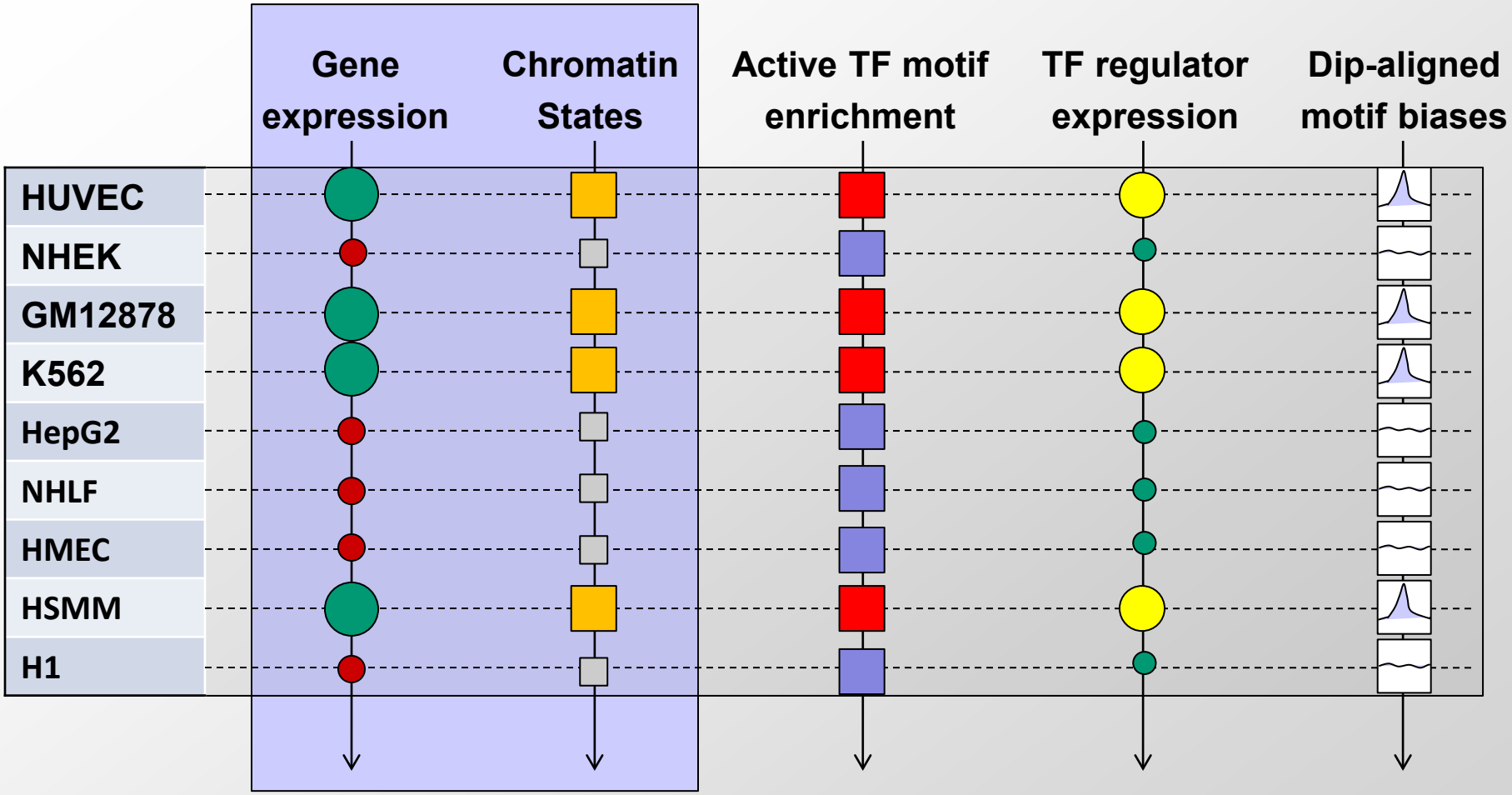
Reveal epigenomic variability: enh/prom/tx/repr/het

Anshul Kundaje

2.3M enhancer regions ⇔ only ~200 activity patterns



Introducing multi-cell activity profiles

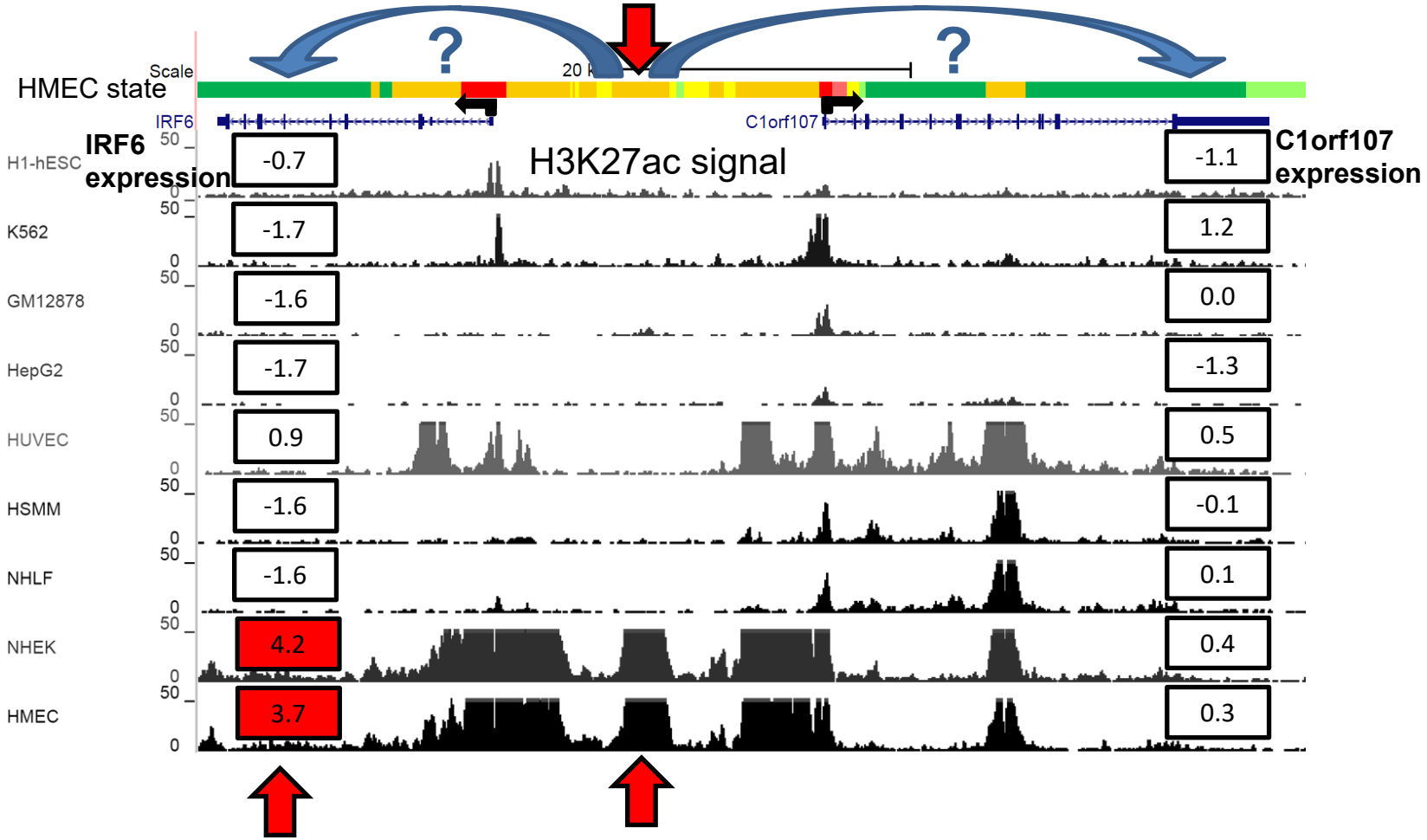


Link enhancers to target genes

ON
OFF
Active enhancer
Repressed
Motif enrichment
Motif depletion
TF On
TF Off
Motif aligned
Flat profile

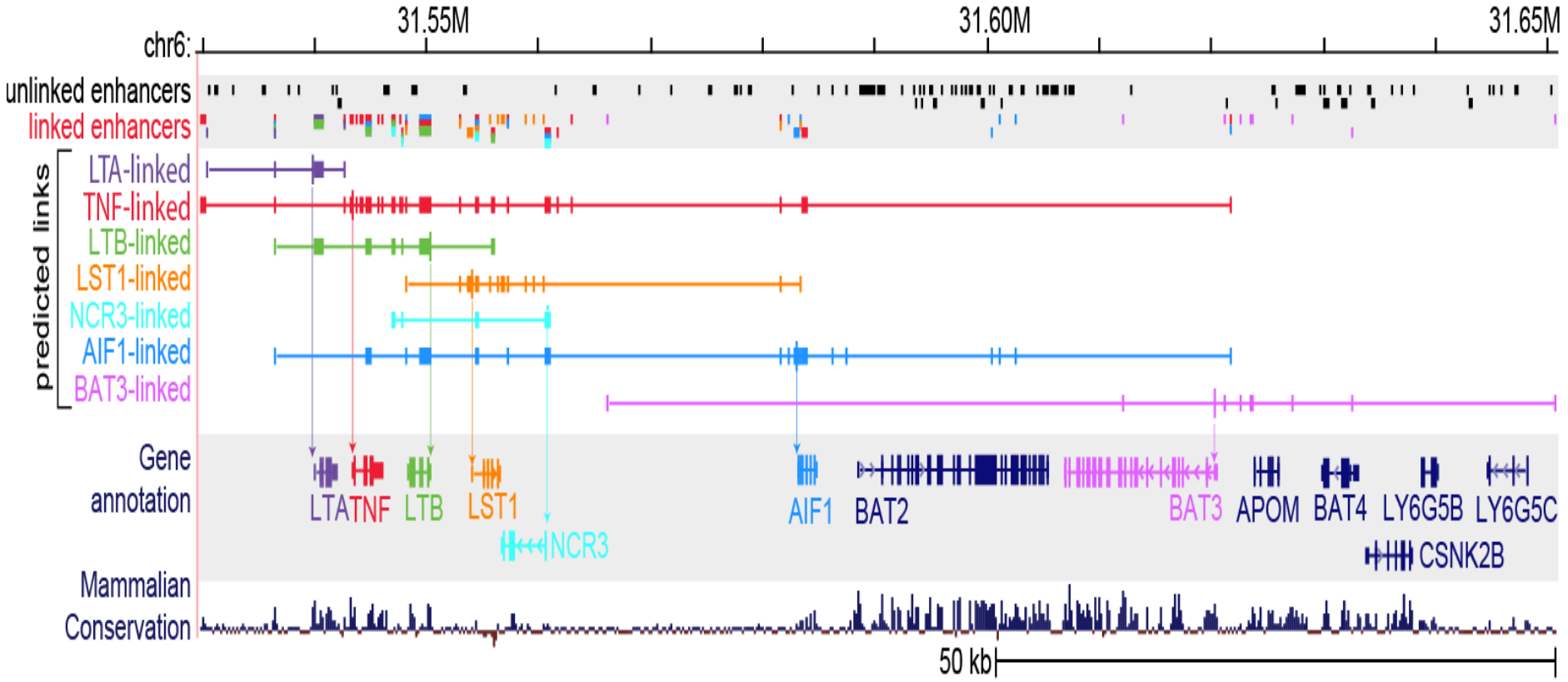
Activity-based linking of enhancers to target genes

Finding correct target of enhancer in divergently transcribed genes



Compute correlations between gene expression levels and enhancer associated histone modification signals

Visualizing 10,000s predicted enhancer-gene links

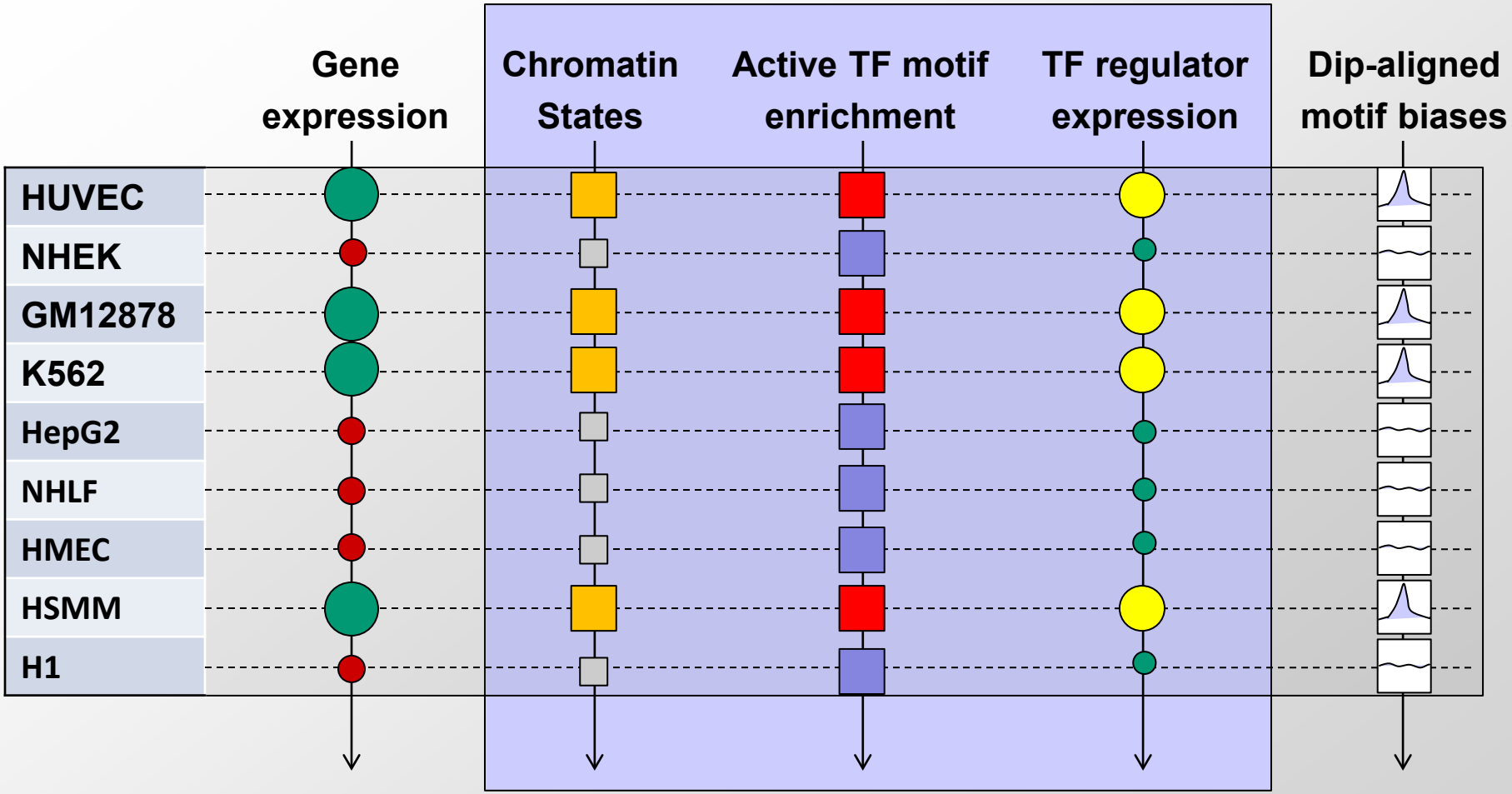


- Overlapping regulatory units, both few and many
- Both upstream and downstream elements linked
- Enhancers correlate with sequence constraint

Chromatin dynamics: linking enhancer networks

TFs → enhancers → target genes

Introducing multi-cell activity profiles



Link TFs to target enhancers

Predict activators vs. repressors

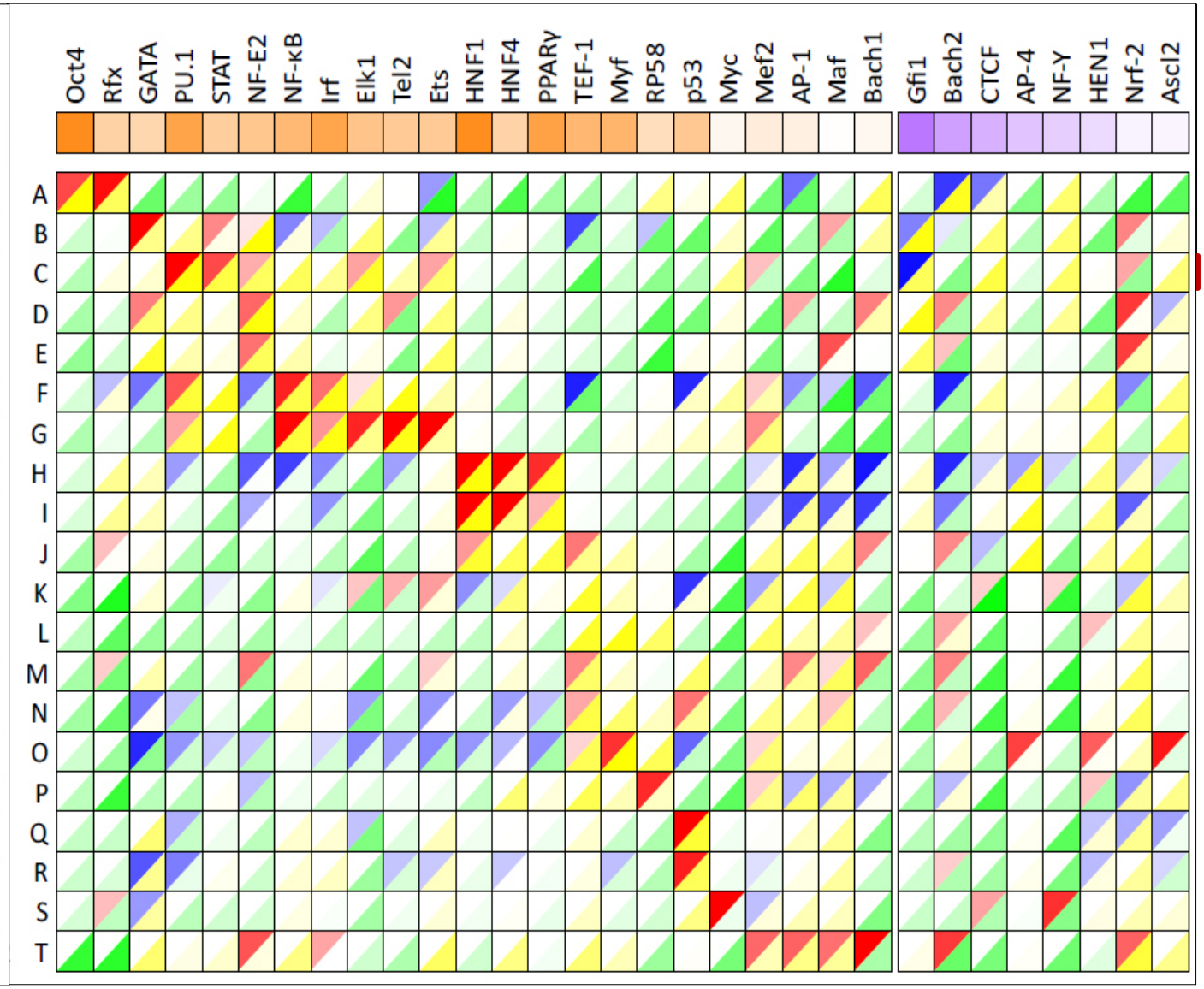
ON
OFF
Active enhancer
Repressed
Motif enrichment
Motif depletion
TF On
TF Off
Motif aligned
Flat profile

Coordinated activity reveals activators/repressors

Enhancer activity

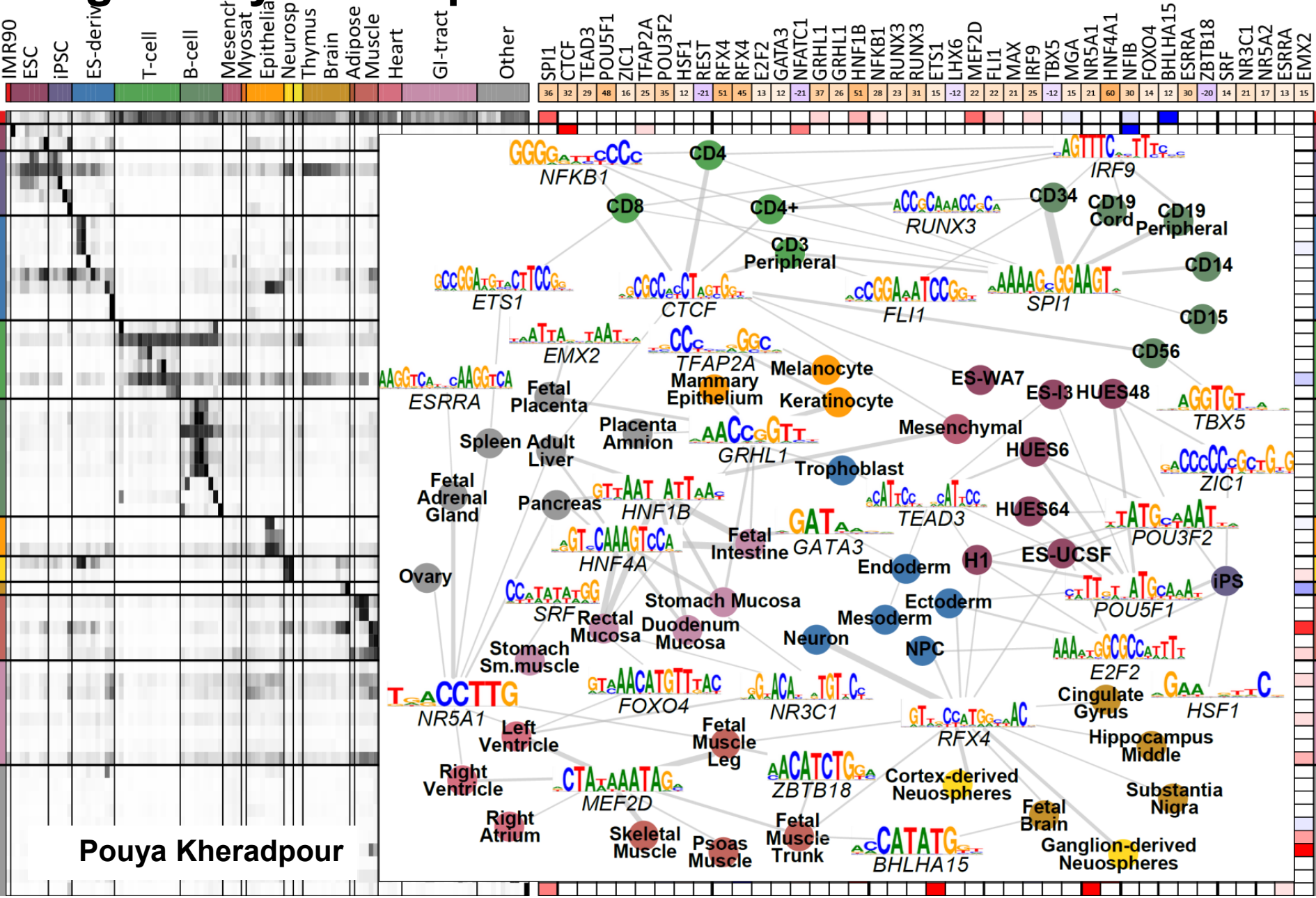
	H1 ES	K562	GM12878	HepG2	HUVEC	HSMM	NHLF	NHEK	HMEC
A	44	3	2	3	2	2	2	2	2
B	1	48	2	4	2	2	2	2	2
C	2	48	47	8	5	5	4	5	3
D	4	46	9	10	45	10	11	14	10
E	4	48	12	14	10	10	10	44	26
F	0	1	48	1	1	1	1	1	1
G	2	5	48	6	46	16	13	12	7
H	1	2	2	51	1	1	1	2	1
I	1	1	1	36	1	1	1	1	1
J	3	7	6	49	31	30	18	12	10
K	1	2	2	2	46	3	4	4	3
L	2	3	2	2	47	45	20	7	9
M	3	3	5	6	49	26	19	47	34
N	2	2	6	5	5	47	18	46	31
O	1	1	2	1	2	45	5	2	3
P	1	3	4	2	4	21	45	4	4
Q	2	1	5	5	3	3	4	49	45
R	1	1	3	2	3	3	3	45	8
S	2	4	4	4	8	6	5	8	43
T	11	40	27	25	45	41	39	45	41

Activity signatures for each TF



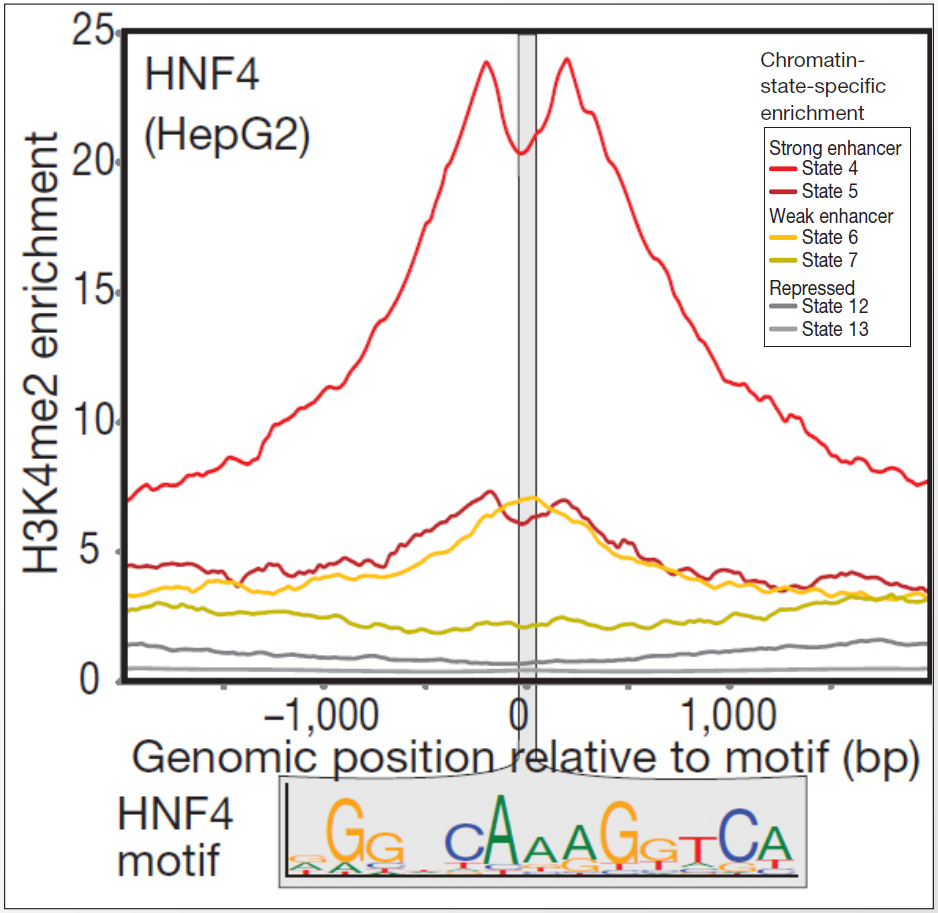
- Enhancer networks: Regulator → enhancer → target gene

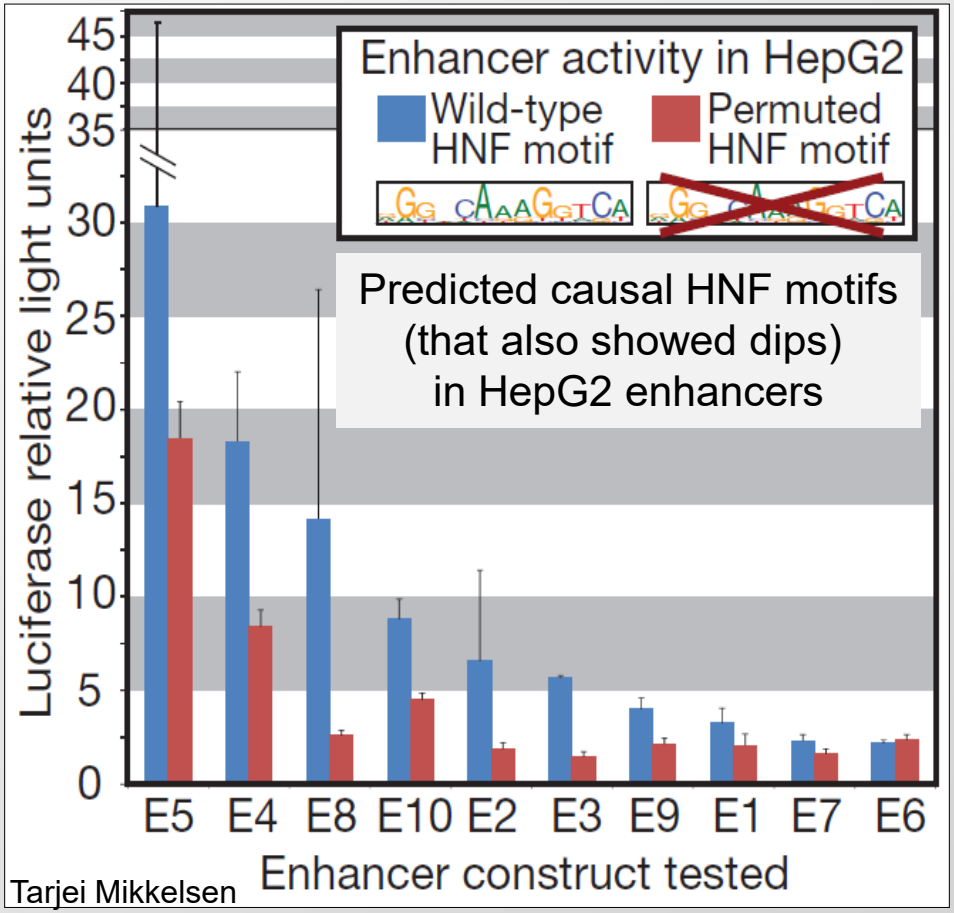
Regulatory motifs predicted to drive enhancer modules



- **Activator and repressor motifs consistent with tissues**

Causal motifs supported by dips & enhancer assays





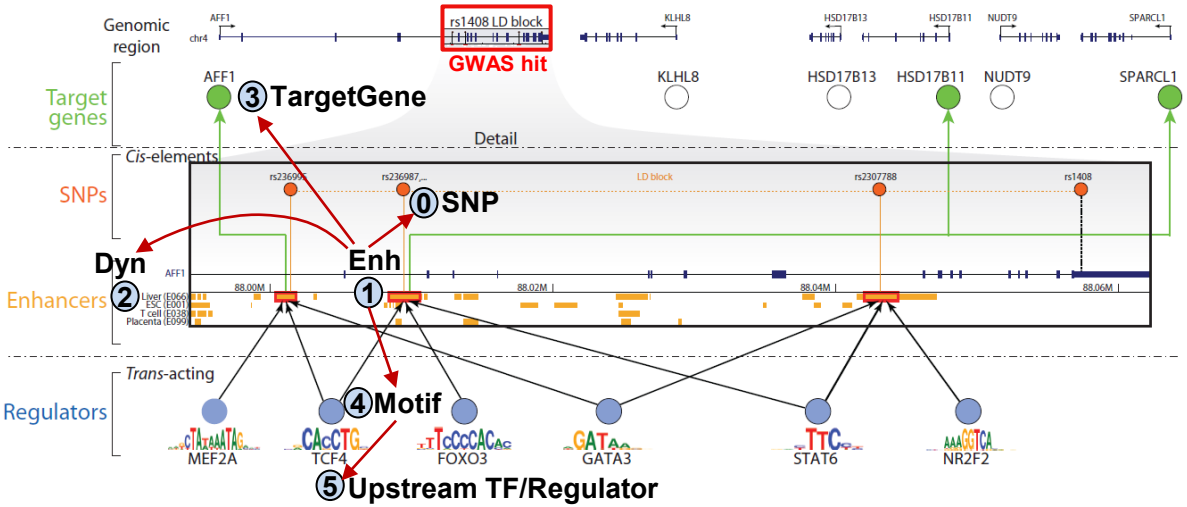
Dip evidence of TF binding (nucleosome displacement)

Enhancer activity halved by single-motif disruption

➔ Motifs bound by TF, contribute to enhancers

MIT 18.700/18.701/6.047/6.878/HST.507

Computational Biology: Genomes, Networks, Evolution



Lecture 5: Regulatory circuitry

Epigenome Dynamics: Joint Chromatin State Learning

Enhancer-gene linking: Correlation, Hi-C, eQTLs

TF motif discovery: Enrichment, EM, Gibbs Sampling

Deep learning convolution CNNs for motif discovery

Global motif discovery: Comparative Genomics

Motif Instance Identification: Branch Length Score

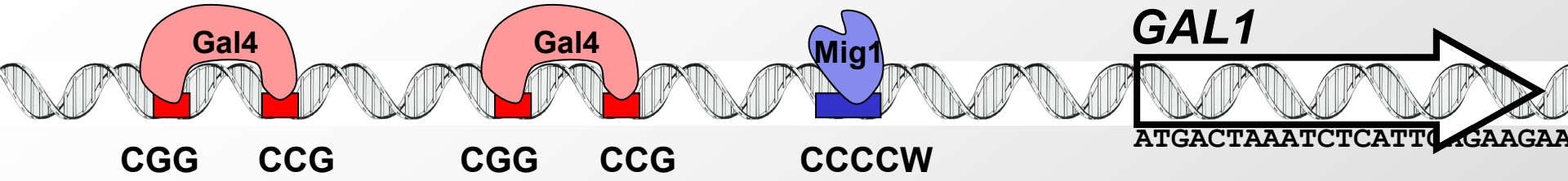
Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions Z_{ij} from motif matrix
   - M step: Find max-likelihood motif from all positions Z_{ij}
3. Gibbs Sampling: Sample from joint (M,Z_{ij}) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

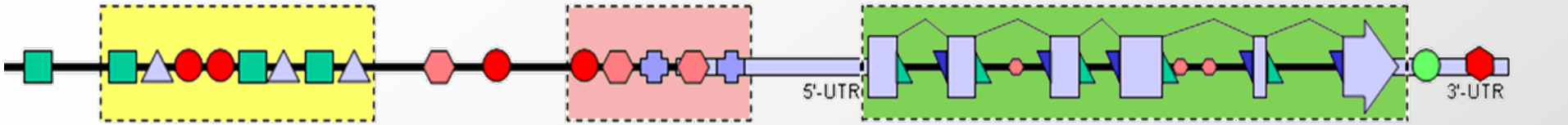
Regulatory motif discovery



- Regulatory motifs
  - Genes are turned on / off in response to changing environments
  - No direct addressing: subroutines (genes) contain sequence tags (motifs)
  - Specialized proteins (transcription factors) recognize these tags

- What makes motif discovery hard?
  - Motifs are short (6-8 bp), sometimes degenerate
  - Can contain any set of nucleotides (no ATG or other rules)
  - Act at variable distances upstream (or downstream) of target gene

The regulatory code: All about regulatory motifs



Enhancer regions
Where in the body?

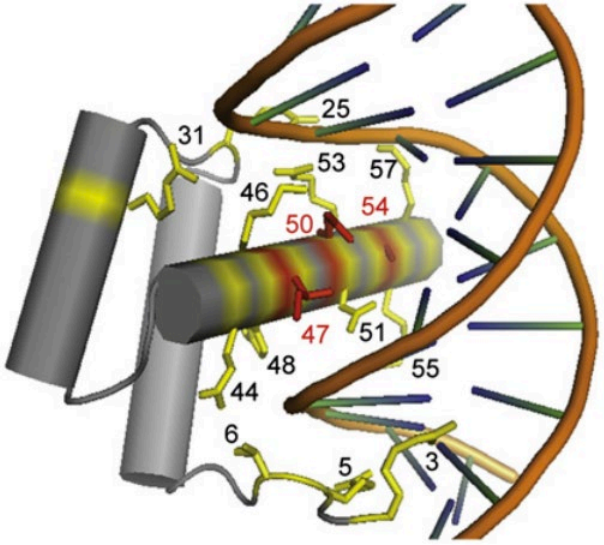
Promoter motifs
When in time?

Splicing signals
Which variants?

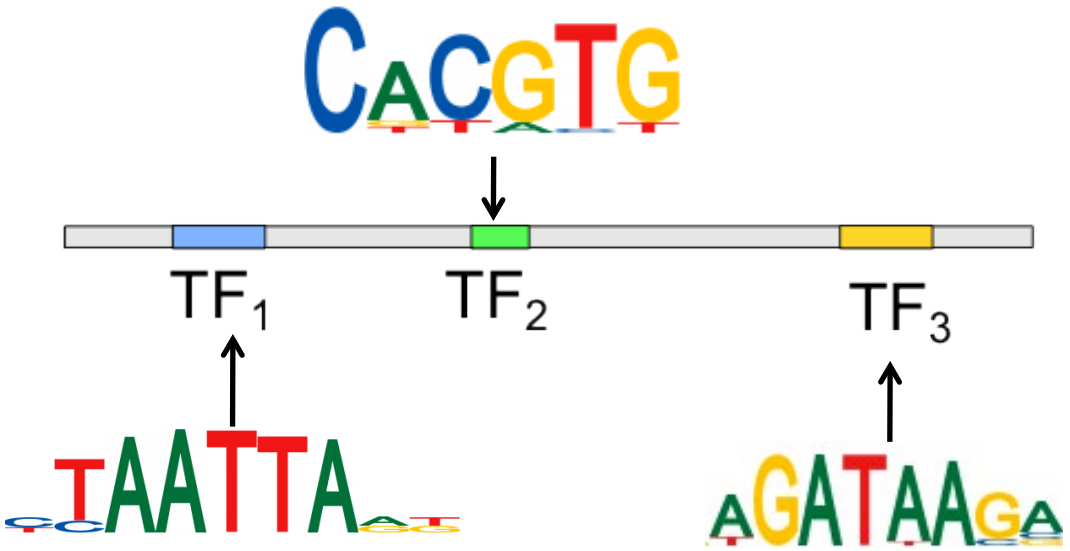
Motifs at RNA level
Which subsets?

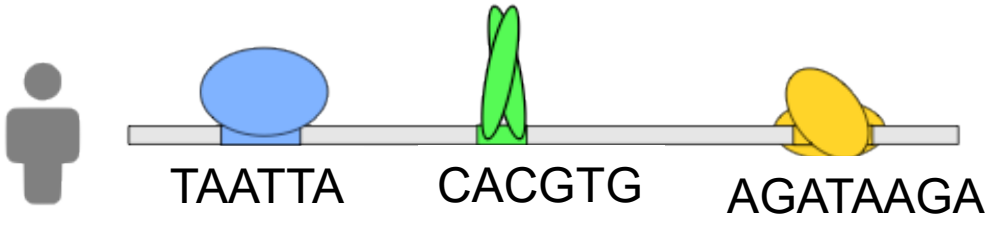
- The parts list: ~20-30k genes
  - Protein-coding genes, RNA genes (tRNA, microRNA, snRNA)
- The circuitry: constructs controlling gene usage
  - Enhancers, promoters, splicing, post-transcriptional motifs
- The regulatory code, complications:
  - Combinatorial coding of 'unique tags'
    - Data-centric encoding of addresses
  - Overlaid with 'memory' marks
    - Large-scale on/off states
  - Modulation of the large-scale coding
    - Post-transcriptional and post-translational information
- Today: discovering motifs in co-regulated promoters and *de novo* motif discovery & target identification

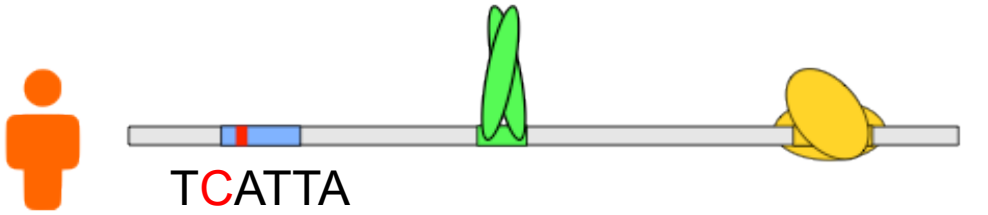
TFs use DNA-binding domains to recognize specific DNA sequences in the genome



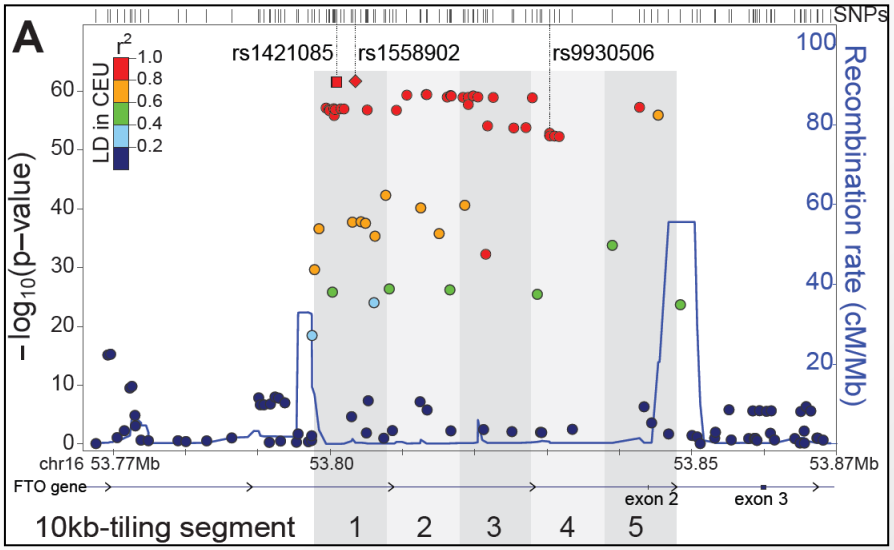
DNA-binding domain of
Engrailed



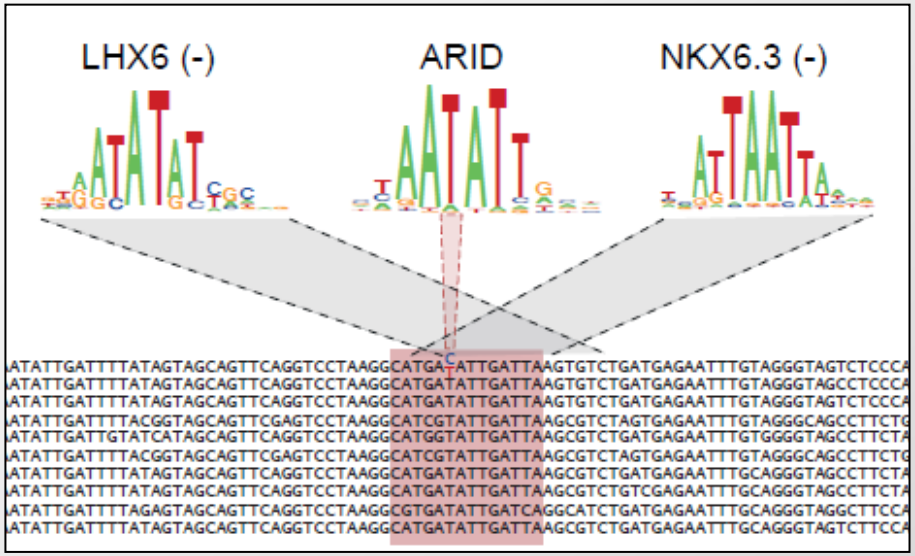




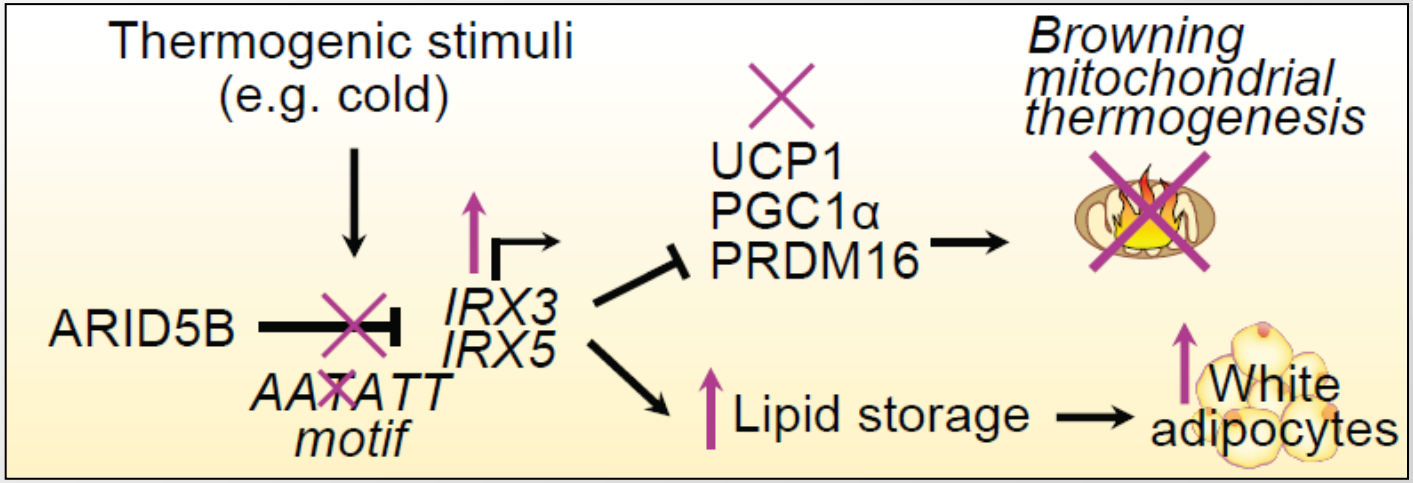
Disrupted motif at the heart of FTO obesity locus



Strongest association with obesity



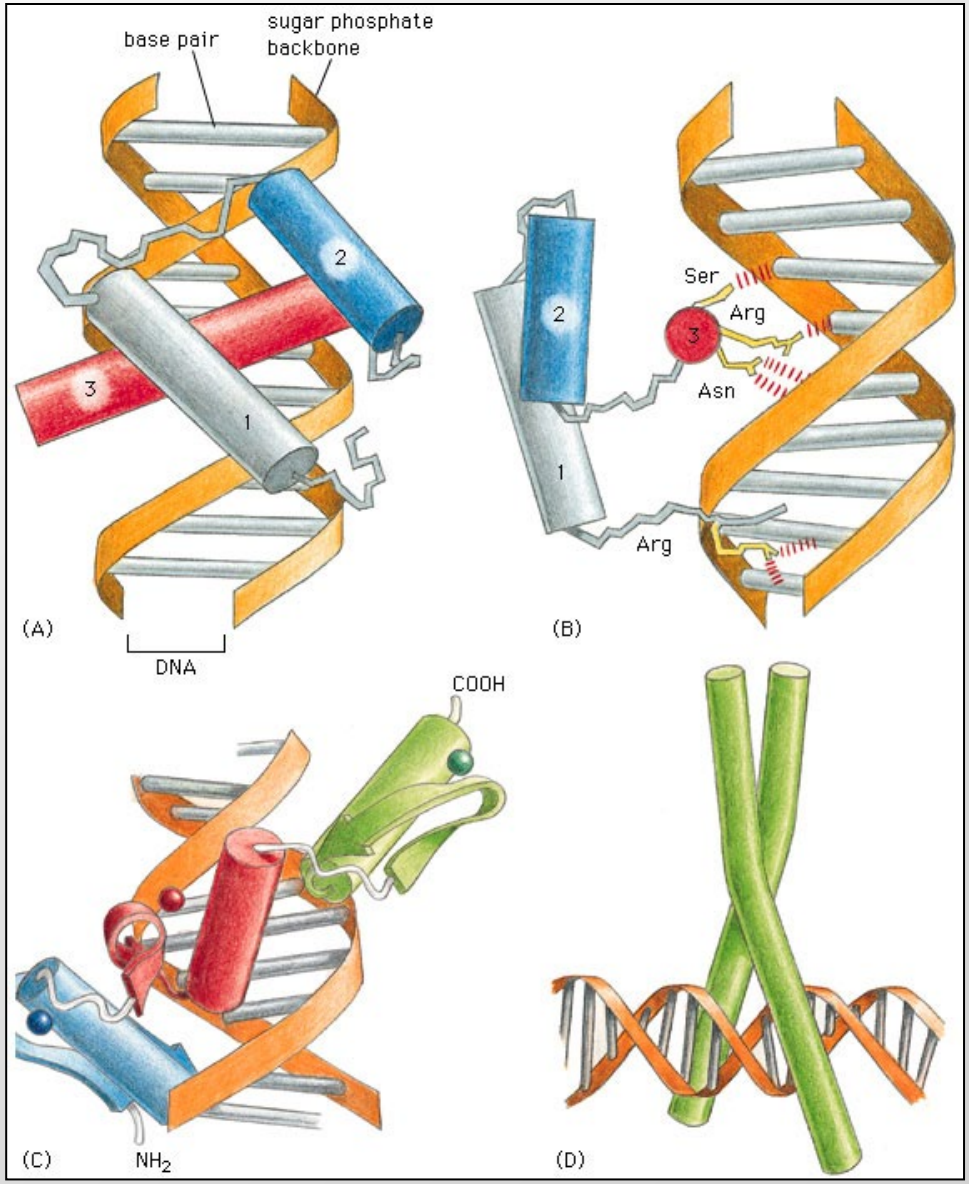
C-to-T disruption of AT-rich regulatory motif



Restoring motif restores thermogenesis

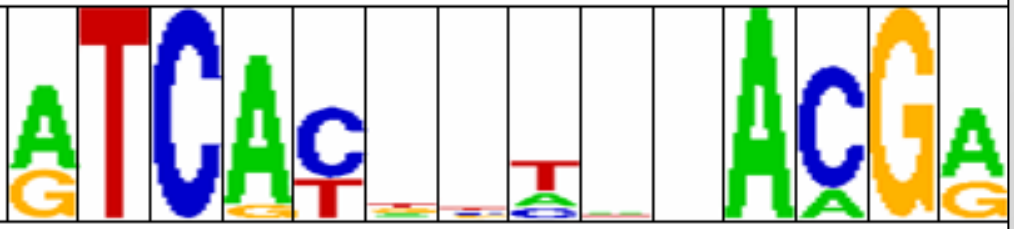
Regulator structure ⇔ recognized motifs

- Proteins ‘feel’ DNA
  - Read chemical properties of bases
  - Do NOT open DNA (no base complementarity)
- 3D Topology dictates specificity
  - Fully constrained positions: ➔ every atom matters
  - “Ambiguous / degenerate” positions ➔ loosely contacted
- Other types of recognition
  - MicroRNAs: complementarity
  - Nucleosomes: GC content
  - RNAs: structure/seqn combination



Motifs summarize TF sequence specificity

Target genes bound by ABF1 regulator		Coordinates		Genome sequence at bound site
ACS1	acetyl CoA synthetase	-491	-479	|ATCATTCTGGACG|
ACS1	acetyl CoA synthetase	-433	-421	|ATCATCTCGGACG|
ACS1	acetyl CoA synthetase	-311	-299	|ATCATTTGCCACG|
CHA1	catabolic L-serine dehydratase	-280	-254	A|ATCACCGCGAACG|GA
ENO2	Enolase	-470	-461	ggcgttat|GTCACTAACGACG|tgcacca
HMR	silencer	-256	-283	ATCAATAC|ATCATAAAATACG|AACGATC
LPD1	lipoamide dehydrogenase	-288	-300	gat|ATCAAAATTAACG|tag
LPD1	lipoamide dehydrogenase	-301	-313	gat|ATCACCGTTGACG|tca
PGK	phosphoglycerate kinase	-523	-496	CAAACAA|ATCACGAGCGACG|GTAATTTC
RPC160	RNA pol III/C 160 kDa subunit	-385	-349	|ATCACTATATACG|TGAA
RPC40	RNA pol III/C 40 kDa subunit	-137	-116	|GTCACTATAAACG|
rpL2	ribosomal protein L2	-185	-167	TAAT|aTCAcgtcACACG|AC
SPR3	CDC3/10/11/12 family homolog	-315	-303	|ATCACTAAATACG|
YPT1	TUB2	-193	-172	CCTAG|GTCACTGTACACG|TATA

Position		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Position Weight Matrix (PWM)	A	**56**	4	4	**81**	4	23	15	27	31	31	**89**	23	4	**58**
	G	**32**	4	4	12	4	31	23	4	19	23	4	4	**89**	**35**
	C	4	4	**89**	4	**58**	12	23	19	19	23	4	**69**	4	4
	T	4	**89**	4	4	**35**	35	39	50	31	23	4	4	4	4
Motif Logo															
Consensus		**R**	**T**	**C**	**A**	**Y**	N	N	H	N	N	**A**	**C**	**G**	**R**

- Summarize information
- Integrate many positions
- Measure of information
- Distinguish motif vs. motif instance
- Assumptions:
  - Independence
  - Fixed spacing

Motifs are not limited to DNA sequences

- Splicing Signals at the RNA level
  - Splice junctions
  - Exonic Splicing Enhancers (ESE)
  - Exonic Splicing Surpressors (ESS)
- Domains and epitopes at the Protein level
  - Glycosylation sites
  - Kinase targets
  - Targetting signals
  - MHC binding specificities
- Recurring patterns at the physiological level
  - Expression patterns during the cell cycle
  - Heart beat patterns predicting cardiac arrest
    - Final project in previous year, now used in Boston hospitals!
  - Any probabilistic recurring pattern

Approaches to regulatory motif discovery

Region-based motif discovery

- Expectation Maximization (e.g. MEME)
  - Iteratively refine positions / motif profile
- Gibbs Sampling (e.g. AlignACE)
  - Iteratively sample positions / motif profile
- Enumeration with wildcards (e.g. Weeder)
  - Allows global enrichment/background score
- Peak-height correlation (e.g. MatrixREDUCE)
  - Alternative to cutoff-based approach

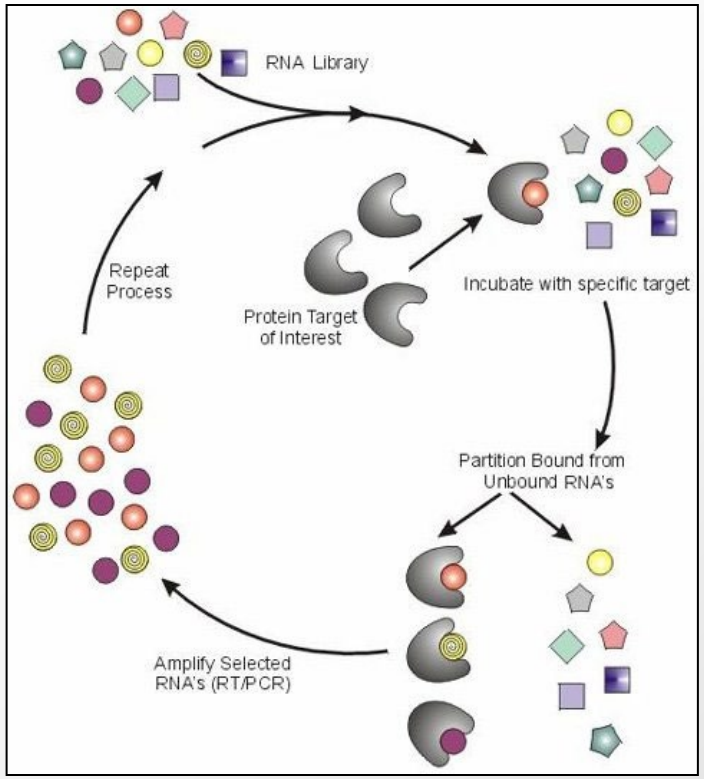
Genome-wide

- Conservation-based discovery (e.g. MCS)
  - Genome-wide score, up-/down-stream bias

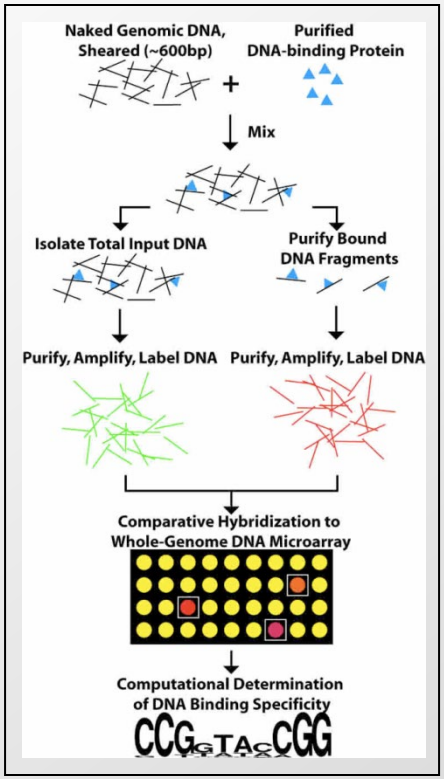
In vitro / trans

- Protein Domains (e.g. PBMs, SELEX)
  - In vitro motif identification, seq-/array-based

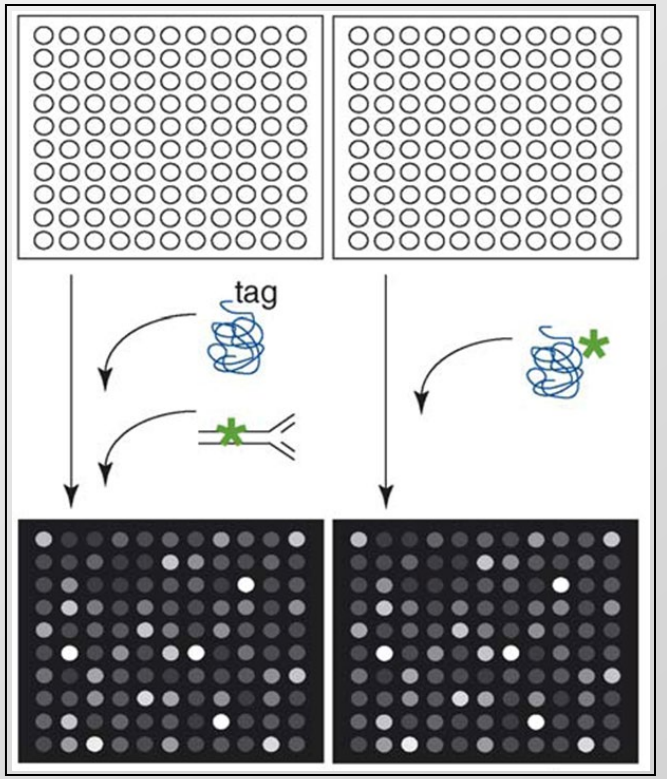
Experimental factor-centric discovery of motifs



SELEX (Systematic Evolution of Ligands by Exponential Enrichment; Klug & Famulok, 1994).

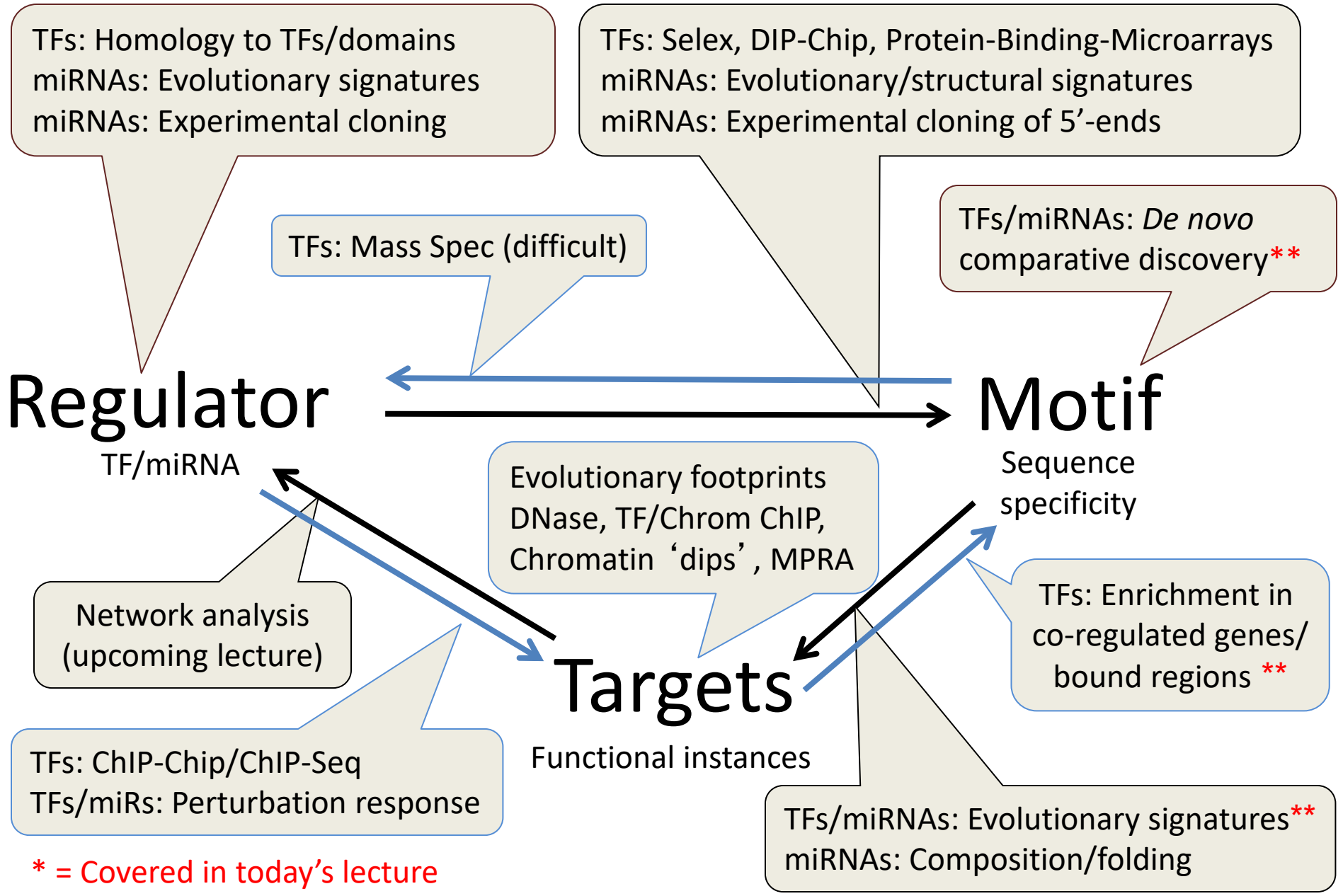


DIP-Chip (DNA-immunoprecipitation with microarray detection; Liu et al., 2005)



PBMs (Protein binding microarrays; Mukherjee, 2004)
Double stranded DNA arrays

Challenges in regulatory genomics

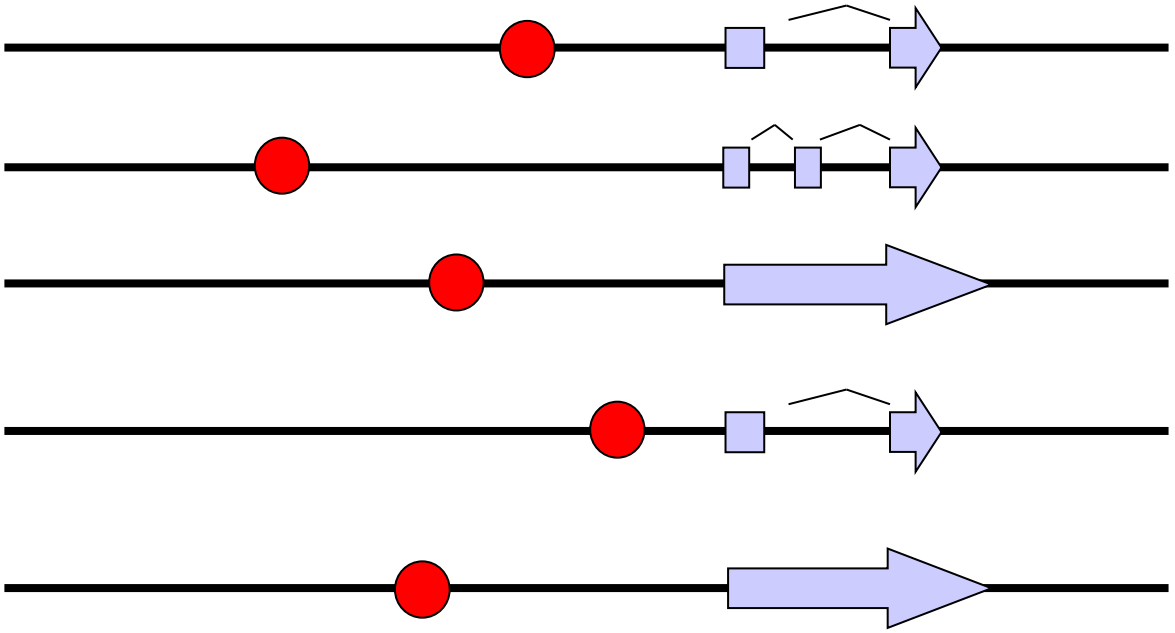


Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions Z_{ij} from motif matrix
   - M step: Find max-likelihood motif from all positions Z_{ij}
3. Gibbs Sampling: Sample from joint (M,Z_{ij}) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

Enrichment-based discovery methods

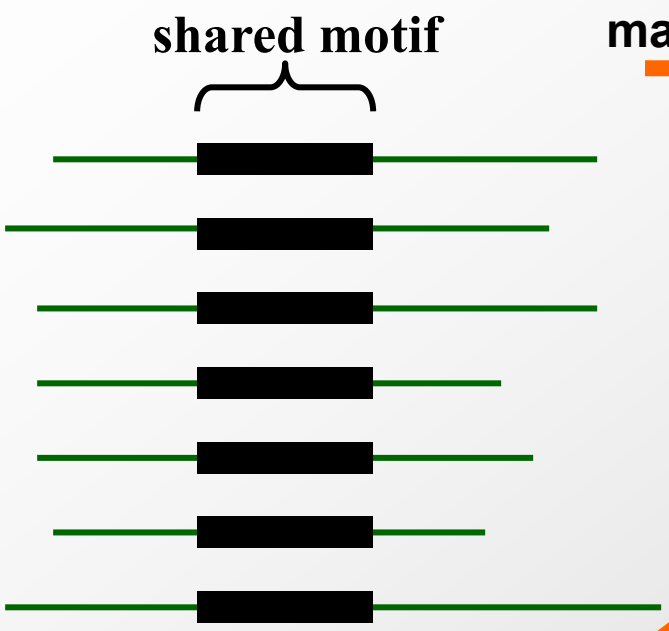
Given a set of co-regulated/functionally related genes, find common motifs in their promoter regions



- **Align the promoters to each other using local alignment**
- **Use expert knowledge for what motifs should look like**
- **Find ‘median’ string by enumeration (motif/sample driven)**
- **Start with conserved blocks in the upstream regions**

Starting positions ⇔ Motif matrix

- given aligned sequences ➔ easy to compute profile matrix



maximization

sequence positions

	1	2	3	4	5	6	7	8
A	0.1	0.3	0.1	0.2	0.2	0.4	0.3	0.1
C	0.5	0.2	0.1	0.1	0.6	0.1	0.2	0.7
G	0.2	0.2	0.6	0.5	0.1	0.2	0.2	0.1
T	0.2	0.3	0.2	0.2	0.1	0.3	0.3	0.1

expectation

given profile matrix

- easy to find starting position probabilities

Key idea: Iterative procedure for estimating both, given uncertainty (learning problem with hidden variables: the starting positions)

Three options for assigning points, and their parallels across K-means, HMMs, Motifs

Update rule	Update assignments (E step) ➔ Estimate hidden labels	Algorithm implementing E step in each of the three settings			Update model parameters (M step) ➔ max likelihood
		Expression clustering	**HMM learning**	**Motif discovery**	
The hidden label is:		Cluster labels	State path π	Motif positions	
Pick a best	Assign each point to best label	**K-means:** Assign each point to nearest cluster	**Viterbi training:** label sequence with best path	**Greedy:** Find best motif match in each sequence	Average of those points assigned to label
Average all	Assign each point to all labels, probabilistically	**Fuzzy K-means:** Assign to all clusters, weighted by proximity	**Baum-Welch training:** label sequence w all paths (posterior decoding)	**MEME:** Use all positions as a motif occurrence weighed by motif match score	Average of all points, weighted by membership
Sample one	Pick one label at random, based on their relative probability	**N/A:** Assign to a random cluster, sample by proximity	**N/A:** Sample a single label for each position, according to posterior prob.	**Gibbs sampling:** Use one position for the motif, by sampling from the match scores	Average of those points assigned to label(a sample)

Basic Iterative Approach

Given: length parameter *W,* training set of sequences

set initial values for **motif**

do

➔ re-estimate ***starting-positions*** from ***motif***

➔ re-estimate ***motif*** from ***starting-positions***

until convergence (change < ε)

return: ***motif, starting-positions***

Representing Motif M(k,c) and Background B(c)

- Assume motif has fixed width, *W*
- Motif represented by matrix of probabilities: ***M(k,c)***
  the probability of character ***c*** in column ***k***

$$M = \begin{array}{c|ccc} & 1 & 2 & 3 \\ A & 0.1 & 0.5 & 0.2 \\ C & 0.4 & 0.2 & 0.1 \\ G & 0.3 & 0.1 & 0.6 \\ T & 0.2 & 0.2 & 0.1 \end{array}$$

(~CAG)

- **Background represented by B(c), frequency of each base**

$$B = \begin{array}{c|c} A & 0.26 \\ C & 0.24 \\ G & 0.23 \\ T & 0.27 \end{array}$$

(near uniform)

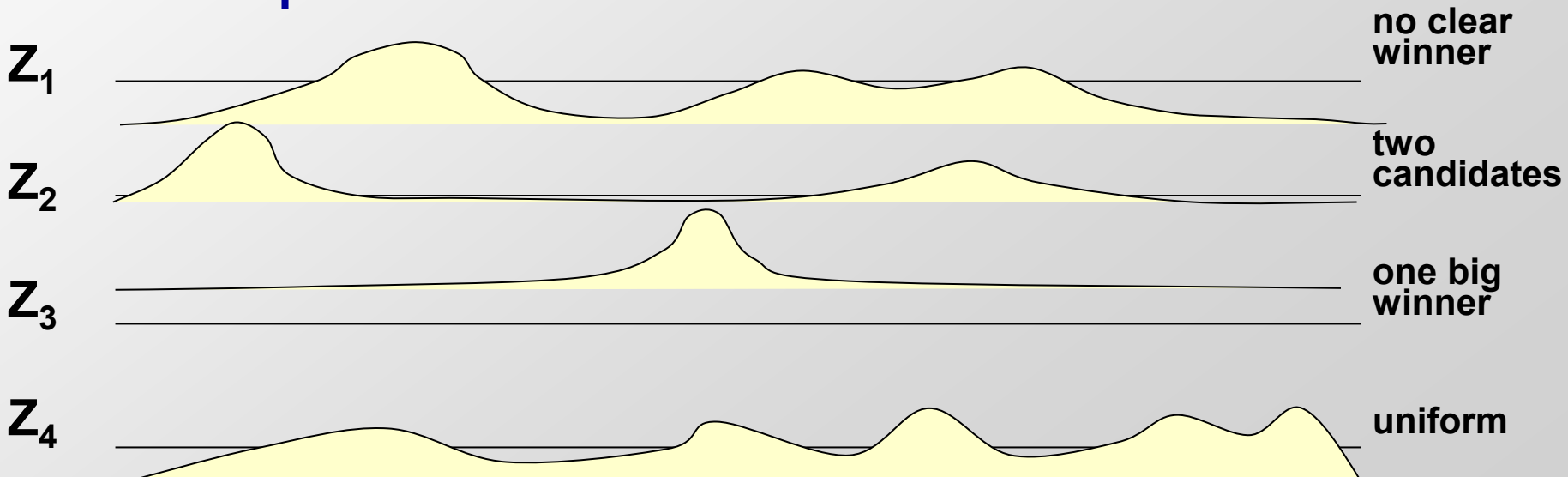
(see also: di-nucleotide etc)

Representing the starting position probabilities (Z_{ij})

- the element Z_{ij} of the matrix Z represents the probability that the motif starts in position *j* in sequence *i*

$$Z = \begin{array}{c} \\ \texttt{seq1} \\ \texttt{seq2} \\ \texttt{seq3} \\ \texttt{seq4} \end{array} \begin{array}{cccc} \texttt{1} & \texttt{2} & \texttt{3} & \texttt{4} \\ \texttt{0.1} & \texttt{0.1} & \texttt{0.2} & \texttt{0.6} \\ \texttt{0.4} & \texttt{0.2} & \texttt{0.1} & \texttt{0.3} \\ \texttt{0.3} & \texttt{0.1} & \texttt{0.5} & \texttt{0.1} \\ \texttt{0.1} & \texttt{0.5} & \texttt{0.1} & \texttt{0.3} \end{array}$$

Some examples:



Starting positions (Z_{ij}) ⇔ Motif matrix M(k,c)

M-step

E-step

	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8
c=A	0.1	0.3	0.1	0.2	0.2	0.4	0.3	0.1
c=C	0.5	0.2	0.1	0.1	0.6	0.1	0.2	0.7
c=G	0.2	0.2	0.6	0.5	0.1	0.2	0.2	0.1
c=T	0.2	0.3	0.2	0.2	0.1	0.3	0.3	0.1

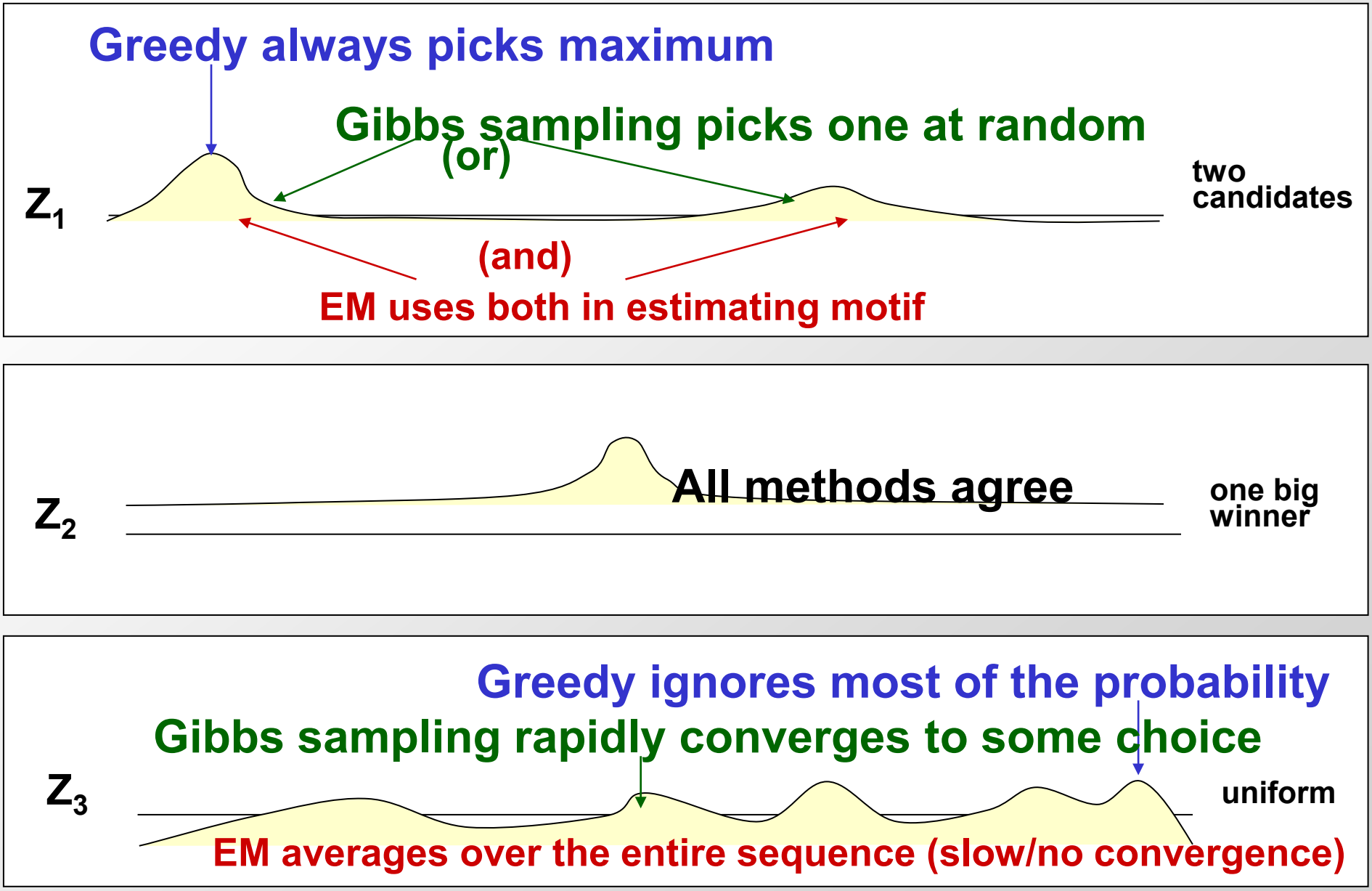
Starting positions: Z_{ij} **Motif: M(k,c)**

- Z_{ij}: Probability that on sequence i, motif start at position j
- M(k,c): Probability that k^{th} character of motif is letter c

- **Computing Z_{ij} matrix from M(k,c) is straightforward**
  - **At each position, evaluate start probability by multiplying across the matrix**

- **Three variations for re-computing motif M(k,c) from Z_{ij} matrix**
  - **Expectation maximization** ➔ **All starts weighted by Z_{ij} prob distribution**
  - **Gibbs sampling** ➔ **Single start for each seq X_i by sampling Z_{ij}**
  - **Greedy approach** ➔ **Best start for each seq X_i by maximum Z_{ij}**

Three examples for Greedy, Gibbs Sampling, EM



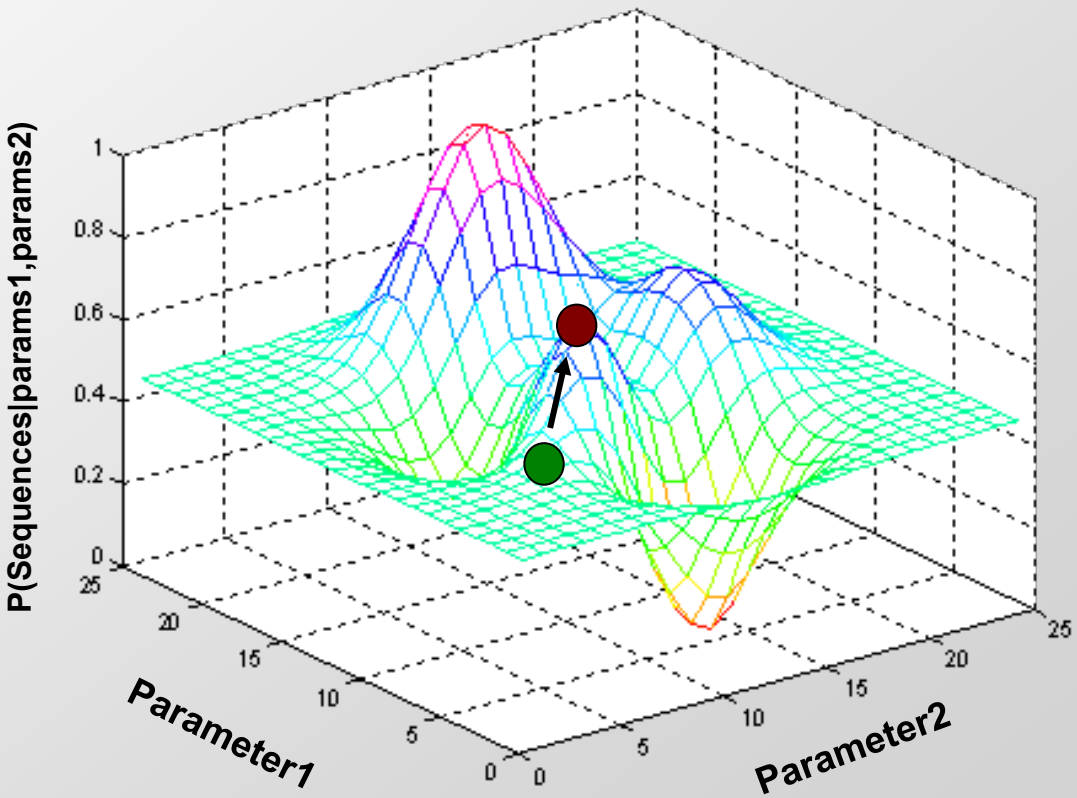
P(Seq|Model) Landscape

EM searches for parameters to increase P(seqs|parameters)

Useful to think of P(seqs|parameters) as a function of parameters

EM starts at an initial set of parameters

And then "climbs uphill" until it reaches a local maximum



Where EM starts can make a big difference

One solution: Search from Many Different Starts

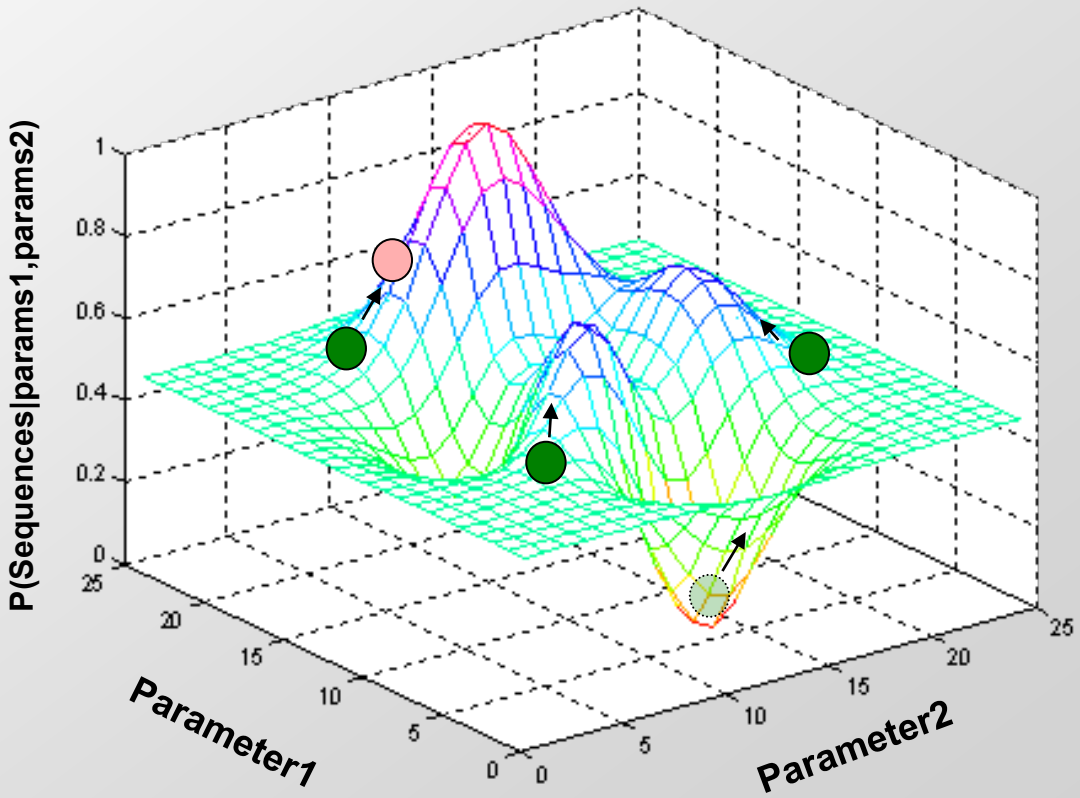
To minimize the effects of local maxima, you should search multiple times from different starting points

MEME uses this idea

Start at many points

Run for one iteration

Choose starting point that got the “highest” and continue



Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions Z_{ij} from motif matrix
   - M step: Find max-likelihood motif from all positions Z_{ij}
3. Gibbs Sampling: Sample from joint (M,Z_{ij}) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

Three options for assigning points, and their parallels across K-means, HMMs, Motifs

Update rule	Update assignments (E step) ➔ Estimate hidden labels	Algorithm implementing E step in each of the three settings			Update model parameters (M step) ➔ max likelihood
		Expression clustering	**HMM learning**	**Motif discovery**	
The hidden label is:		Cluster labels	State path π	Motif positions	
Pick a best	Assign each point to best label	**K-means:** Assign each point to nearest cluster	**Viterbi training:** label sequence with best path	**Greedy:** Find best motif match in each sequence	Average of those points assigned to label
Average all	Assign each point to all labels, probabilistically	**Fuzzy K-means:** Assign to all clusters, weighted by proximity	**Baum-Welch training:** label sequence w all paths (posterior decoding)	**MEME:** Use all positions as a motif occurrence weighed by motif match score	Average of all points, weighted by membership
Sample one	Pick one label at random, based on their relative probability	**N/A:** Assign to a random cluster, sample by proximity	**N/A:** Sample a single label for each position, according to posterior prob.	**Gibbs sampling:** Use one position for the motif, by sampling from the match scores	Average of those points assigned to label(a sample)

Gibbs Sampling

- A general procedure for sampling from the joint distribution of a set of random variables $\Pr(U_1 \ldots U_n)$ by iteratively sampling from for each j $\Pr(U_j \mid U_1 \ldots U_{j-1}, U_{j+1} \ldots U_n)$
- Useful when it's hard to explicitly express means, stdevs, covariances across the multiple dimensions
- Useful for supervised, unsupervised, semi-supervised learning
  - Specify variables that are known, sample over all other variables
- Approximate:
  - Joint distribution: the samples drawn
  - Marginal distributions: examine samples for subset of variables
  - Expected value: average over samples
- Example of Markov-Chain Monte Carlo (MCMC)
  - The sample approximates an unknown distribution
  - Stationary distribution of sample (only start counting after burn-in)
  - Assume independence of samples (only consider every 100)
- Special case of Metropolis-Hastings
  - In its basic implementation of sampling step
  - But it's a more general sampling framework

Gibbs Sampling for motif discovery

- First application to motif finding: Lawrence et al 1993
  - Can view as a stochastic analog of EM for motif discovery task
  - Less susceptible to local minima than EM
- EM maintains distribution $\boldsymbol{Z_i}$ over the starting points for each seq
- Gibbs sampling selects specific starting point $\mathbf{a_i}$ for each seq

  ➔ but keeps resampling these starting points

given: length parameter *W,* training set of sequences

choose random positions for *a*

do

pick a sequence $\boldsymbol{X_i}$

estimate ***p*** given current motif positions *a* (update step)

(using all sequences but X_i)

sample a new motif position a_i for X_i (sampling step)

until convergence

return: *p, a*

Popular implementation: AlignACE, BioProspector

AlignACE: first statistical motif finder

BioProspector: improved version of AlignACE

Both use basic Gibbs Sampling algorithm:

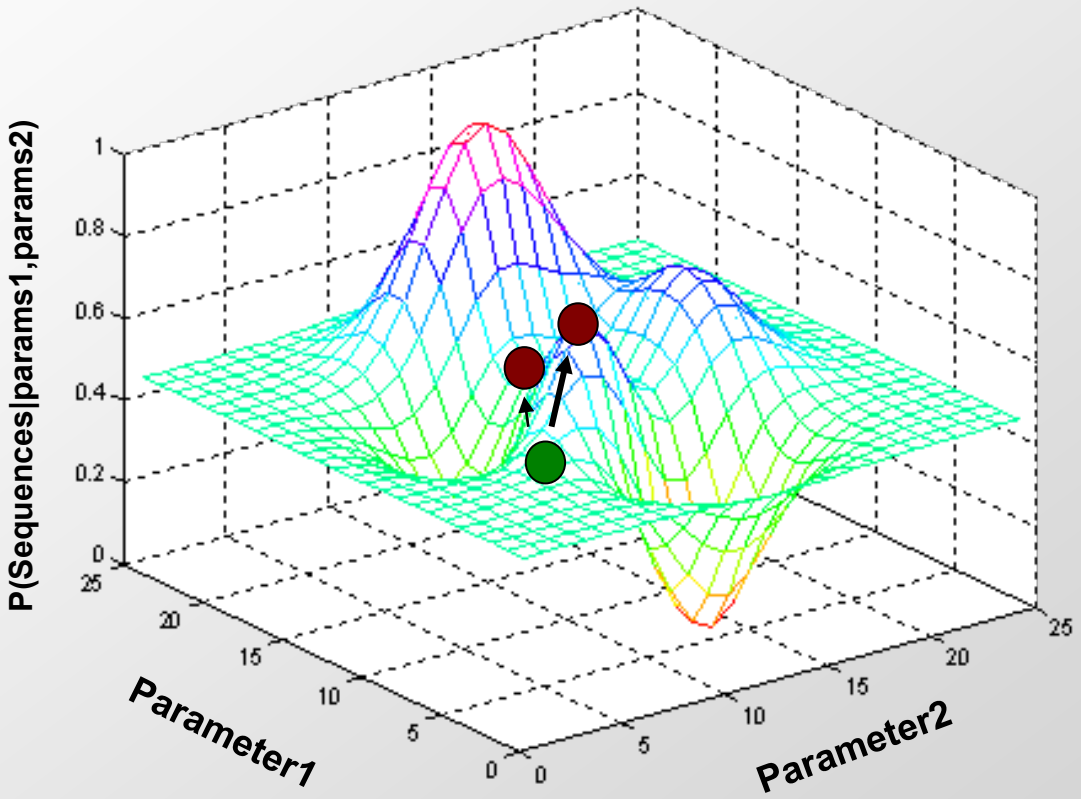
1. Initialization:
   a. Select random locations in sequences $X_1, \ldots, X_N$
   b. Compute an initial model M from these locations
2. Sampling Iterations:
   a. Remove one sequence X_i
   b. Recalculate model
   c. Pick a new location of motif in X_i according to probability the location is a motif occurrence

In practice, run algorithm from multiple random initializations:

1. Initialize
2. Run until convergence
3. Repeat 1,2 several times, report common motifs

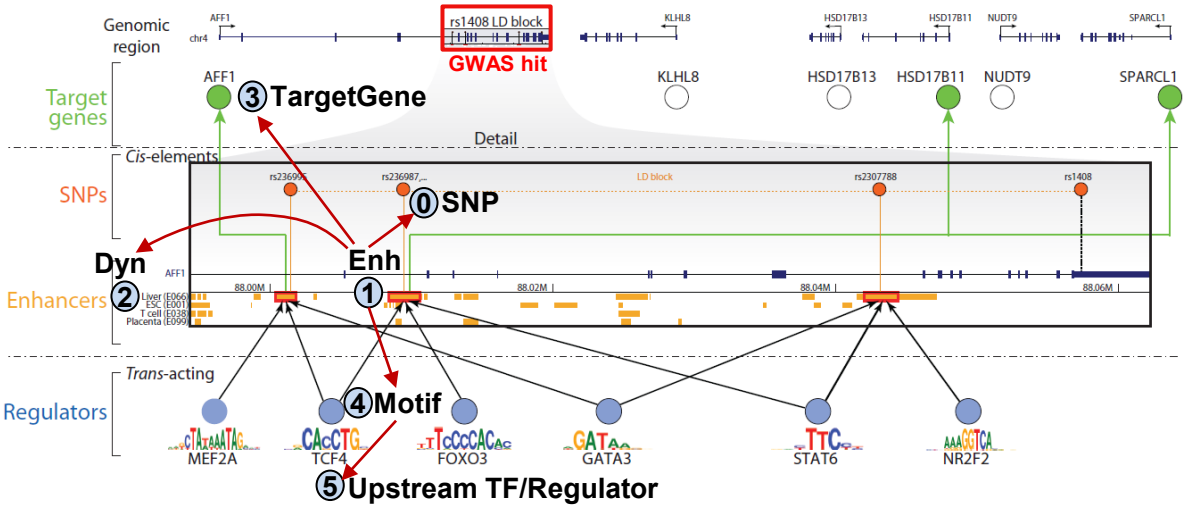
Gibbs Sampling and Climbing

Because gibbs sampling does always choose the best new location it can move to another place not directly uphill



In theory, **Gibbs Sampling less likely to get stuck a local maxima**

MIT 18.700/18.701/6.047/6.878/HST.507
Computational Biology: Genomes, Networks, Evolution

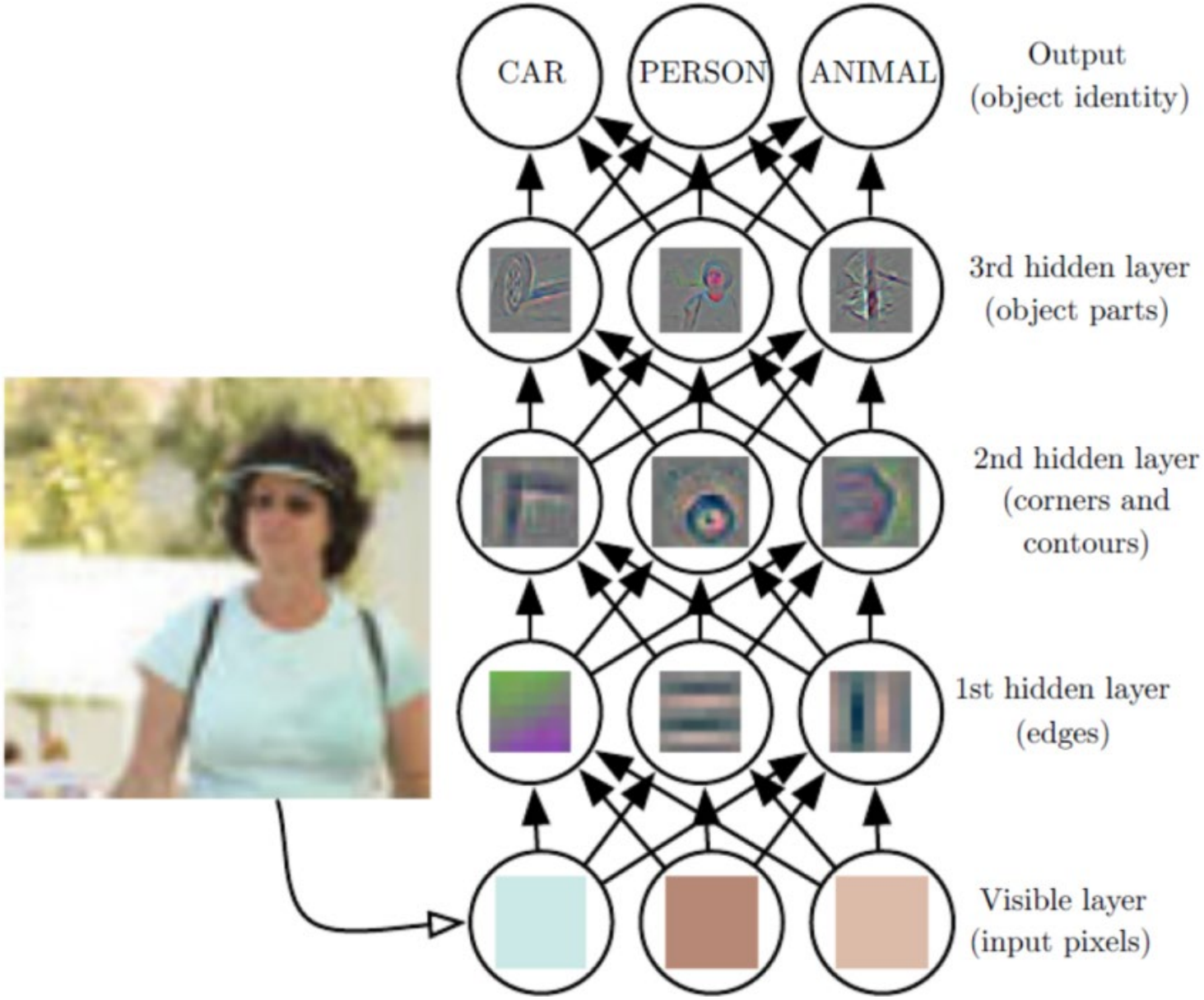


Lecture 5: Regulatory circuitry

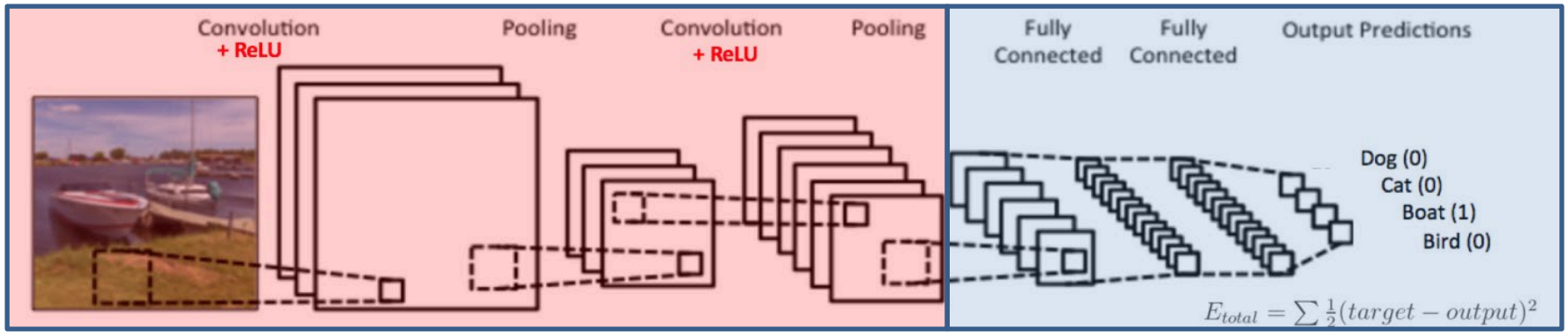
Epigenome Dynamics: Joint Chromatin State Learning
Enhancer-gene linking: Correlation, Hi-C, eQTLs
TF motif discovery: Enrichment, EM, Gibbs Sampling
Deep learning convolution CNNs for motif discovery
Global motif discovery: Comparative Genomics
Motif Instance Identification: Branch Length Score
Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

Human Vision ⇔ many **layers** of **abstraction** ⇔ **Deep** learning



Key idea: **Representation** learning



‘Modern’ Deep learning:
Hierarchical Representation Learning
Feature extraction

‘Classical’ Fully-connected
Neural Networks
Classification

In deep learning, the two tasks are **coupled**:

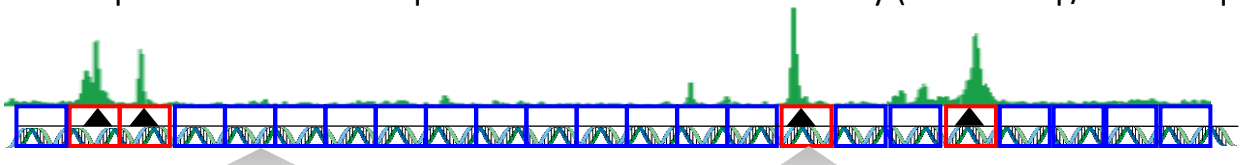
- the classification task “drives” the feature extraction
- Extremely powerful and general paradigm
  - ➔ **Be creative**! The field is still at its infancy!
  - ➔ New application domains (e.g. beyond images) can have **structure** that current architectures **do not capture/exploit**
  - ➔ Genomics/biology/neuroscience can help drive development of **new architectures**

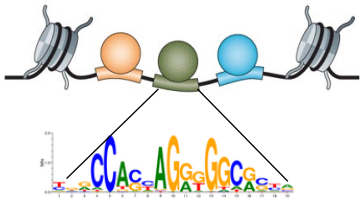
Key design principles of CNNs (+brain counterparts)

Property	Human Visual System Property	Deep Learning CNN Building Block
Locality	Low-level neurons respond to local patches (receptive field)	Local computation of convolutional filters (not a fully-connected network)
Filters	Specialized neurons carry out low-level detection operation	Low-level filters carry out the same operation throughout the network
Layers / abstraction	Layers of neurons learn increasingly abstract 'concepts'	Layers of hidden units, abstract concepts learned from simpler parts / building blocks
Threshold	Neurons fire after cross activation threshold ➔ non-linearity	Activation functions introduce non-linearities ➔ expand universe of functions
Pooling	Higher-level neurons invariant to exact position, sum/max of prev.	Max/Avg pooling layers: positional invarnce reduced # parameters, speed up compute
Multimodal	Different neurons extract different features of image	Multiple filters applied simultaneously, each captures different aspects of original image
Sampling Density	Central vision sampled densely by photoreceptors than periphery	Adjust stride of filter application to denser (slower) vs. sparser (faster) sampling
Saturation	Neurons 'tired' after activation, signal quiets down	Limiting weight of individual hidden units, dropout learning, regularization
Learn/Reinforcement	Useful connections strengthened over time	Back-propagation, adjusting weights across the hierarchy
Feed-fward edges	Neurons with long connections from lower levels to higher ones	Residual networks (ResNets) feed lower-level signal, avoid vanishing gradients

Predictive model of regulatory DNA

Transcription factor ChIP-seq data OR chromatin accessibility (DNase-seq / ATAC-seq data)





…GACTTGAAACGGCATTG
… Inactive (0) (0.3)

…GACAGATAATGCATTGA…
Active (+1) (20.2)

…GACAGATAATGCATTGA…

…ACTGTCATGGATATTCT…

…GATATTCTACTGTAAG…

DNA sequences (S_i)

…CAACCTTGAACGGCATTG…

…GACTTGAAACGGCATTG…

…CAGTATGCATACGTGAA…

Classification or Regression model $F(S_i)$

Arvey et al. 2012
Ghandi et al. 2014
Setty et al. 2015

Class = +1 (20.2)

Class = +1 (10.6)

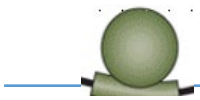
Class = +1 (15.8)

Measured Labels (Y_i)

Class = 0 (0.3)

Class = 0 (1.2)

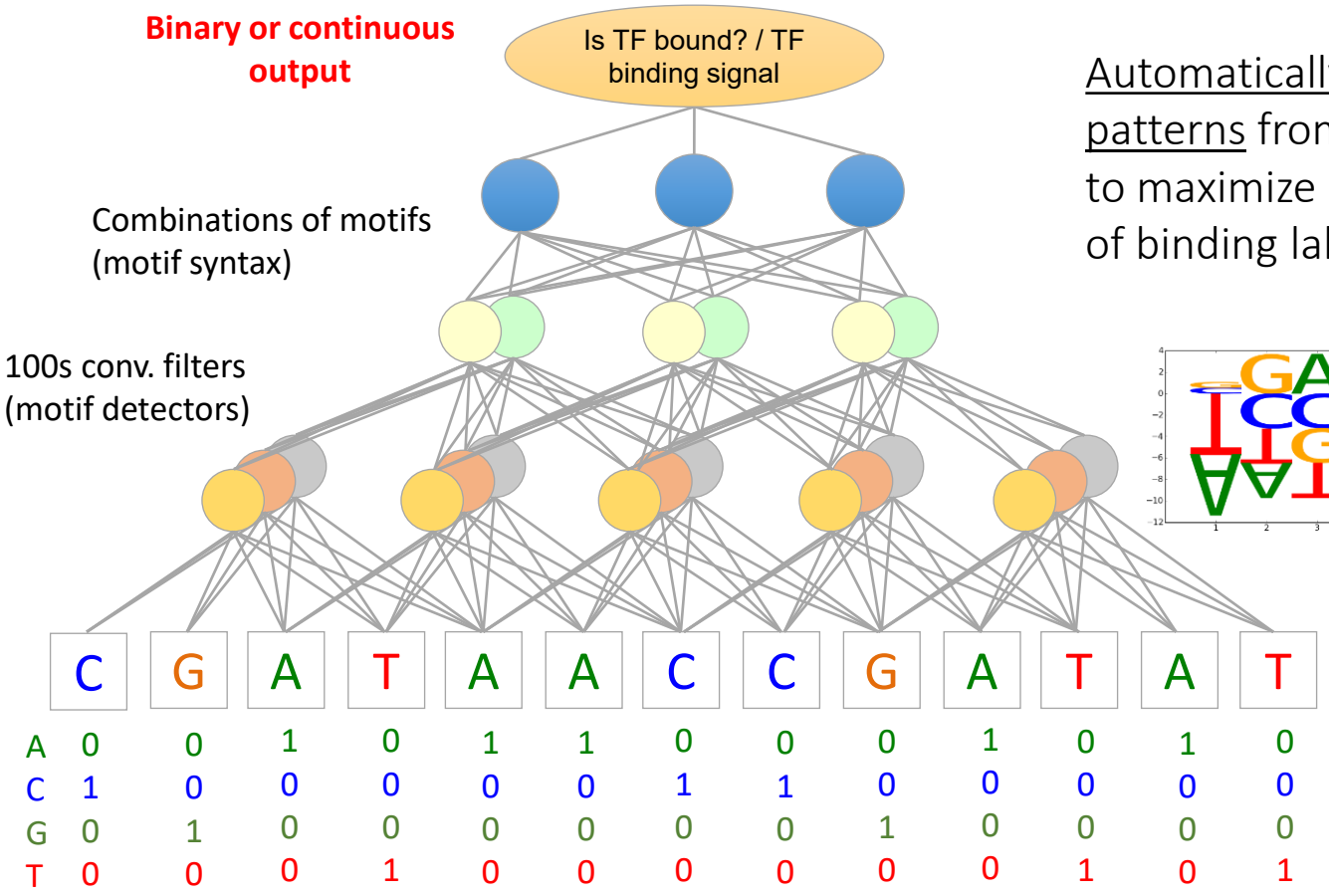
Class = 0 (3.5)



Bound

Unbound

Convolutional neural network (CNN) with DNA sequence inputs

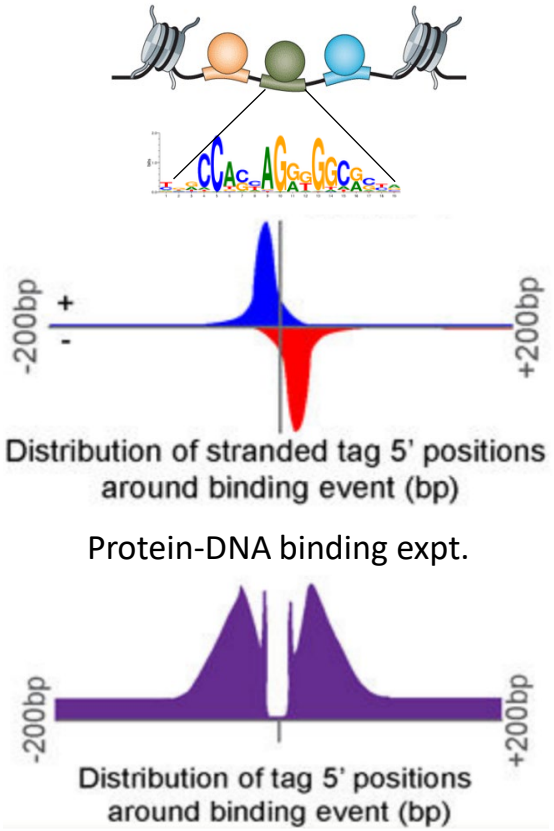


Automatically learns predictive patterns from raw DNA sequence to maximize prediction accuracy of binding labels

One-hot encoded input: DNA sequence represented as ones and zeros

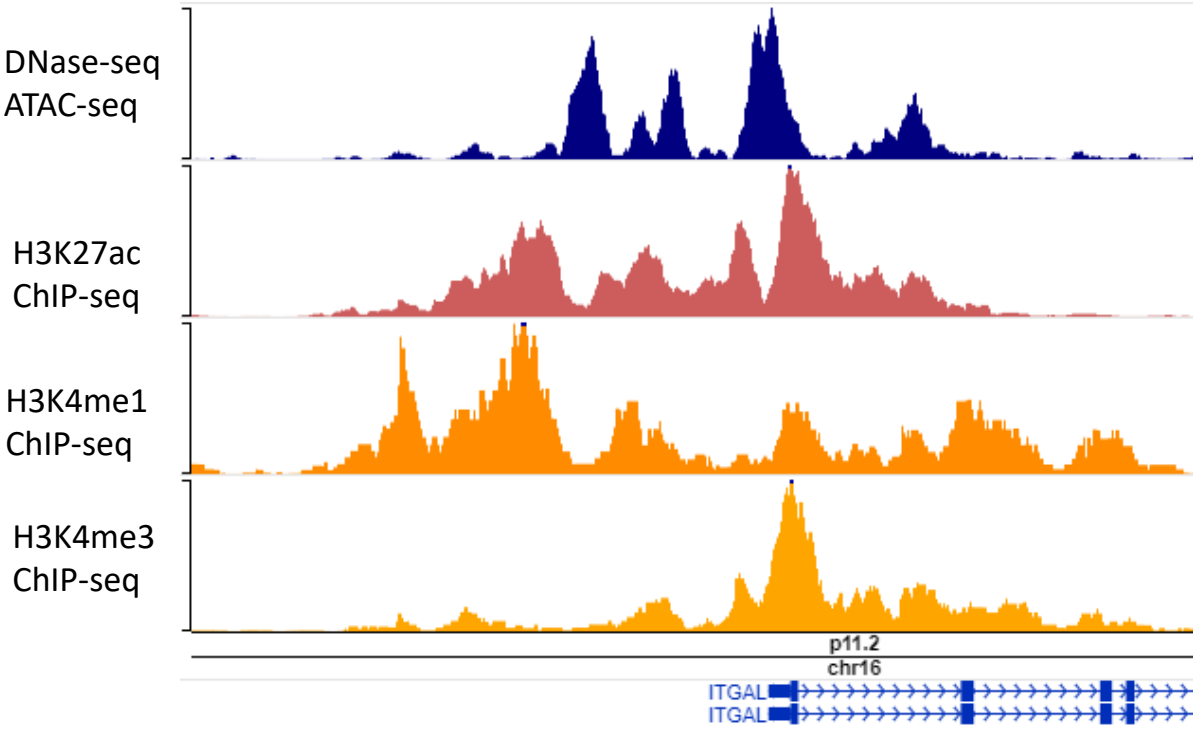
Alipanahi et al. 2015 (DeepBind)
Kelley et al. 2016 (Basset)
Zhou et al. 2015 (DeepSEA)

High-resolution ‘shapes’ and ‘spans‘ of TF and chromatin profiles capture exquisite information about protein-DNA contacts



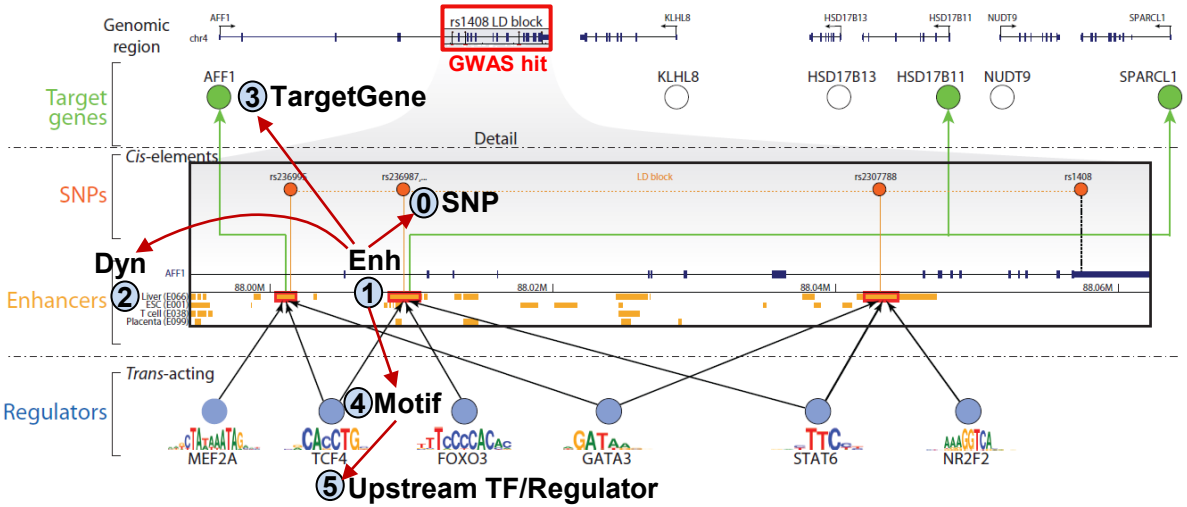
Protein-DNA binding expt.

DNA accessibility experiments



https://doi.org/10.3109/10409238.2015.1051505

MIT 18.700/18.701/6.047/6.878/HST.507
Computational Biology: Genomes, Networks, Evolution



Lecture 5: Regulatory circuitry

Epigenome Dynamics: Joint Chromatin State Learning
Enhancer-gene linking: Correlation, Hi-C, eQTLs
TF motif discovery: Enrichment, EM, Gibbs Sampling
Deep learning convolution CNNs for motif discovery
Global motif discovery: Comparative Genomics
Motif Instance Identification: Branch Length Score
Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

Regulatory genomics: motifs, instances, regions

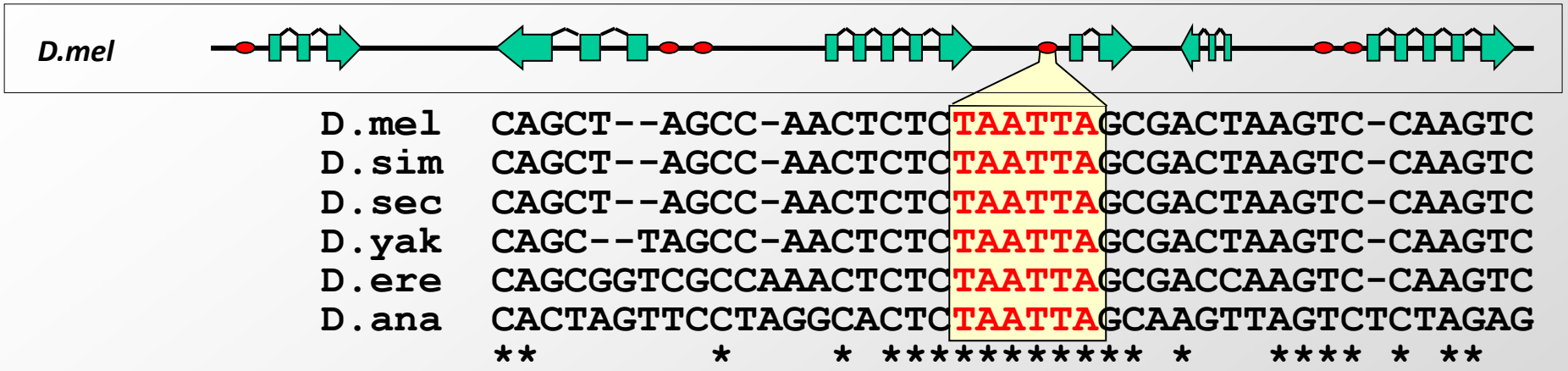
1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions Z_{ij} from motif matrix
   - M step: Find max-likelihood motif from all positions Z_{ij}
3. Gibbs Sampling: Sample from joint (M,Z_{ij}) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

Motivation for *de novo* genome-wide motif discovery

- Both TF and region centric approaches are not comprehensive and are biased
- TF centric approaches generally require transcription factor (or antibody to factor)
  - Lots of time and money
  - Also have computational challenges
- *De novo* discovery using conservation is unbiased but can't match motif to factor and require multiple genomes

Evolutionary signatures for regulatory motifs

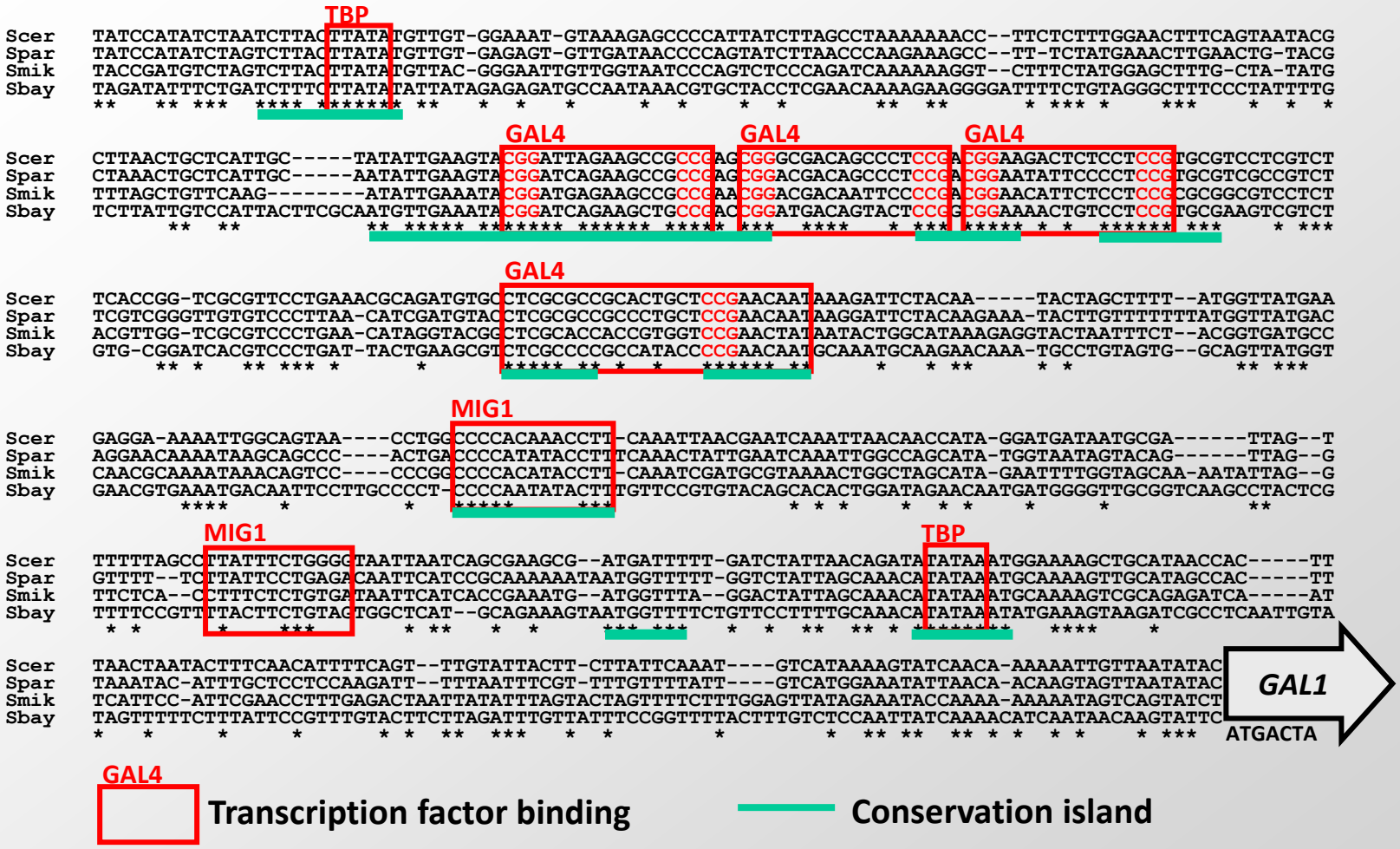
Known engrailed binding site



- Start by looking at known motif instances
- Individual motif instances are preferentially conserved
- Can we just take conservation islands and call them motifs?
  - No. Many conservation islands are due to chance or perhaps due to non-motif conservation

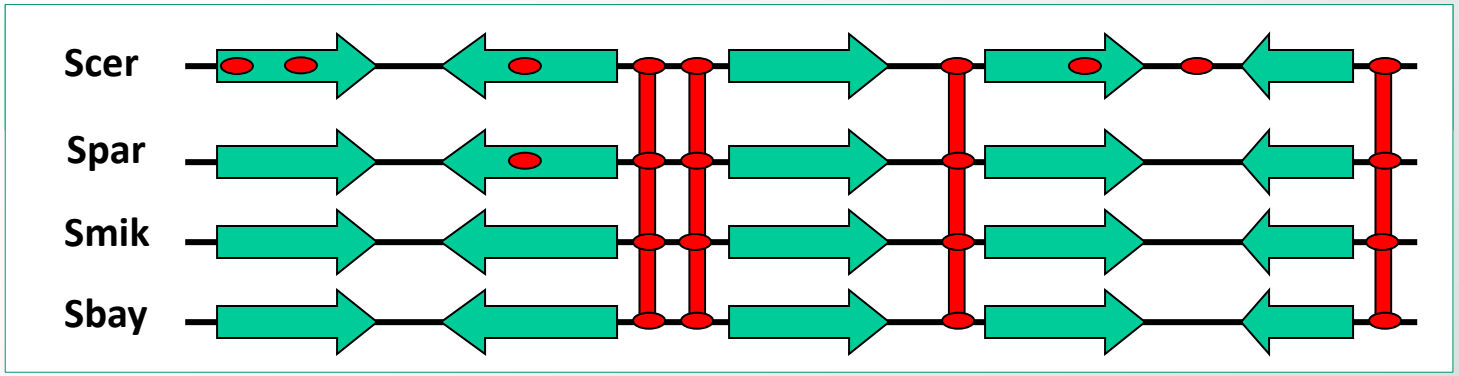
Kellis *el al*, Nature 2003
Xie *et al.* Nature 2005
Stark *et al*, Nature 2007

Conservation islands overlap known motifs



Increase power by testing conservation in many regions

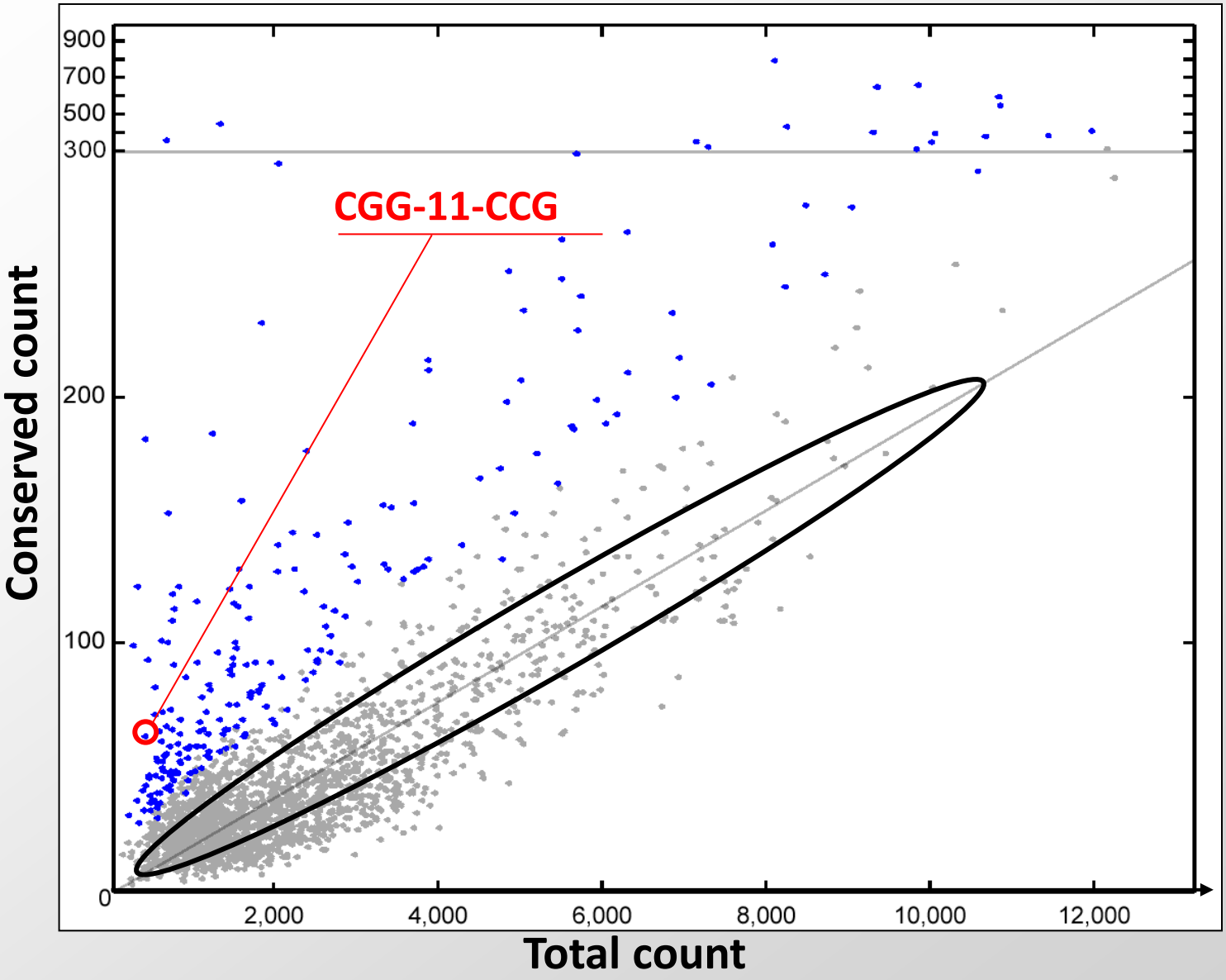
Genome-wide conservation



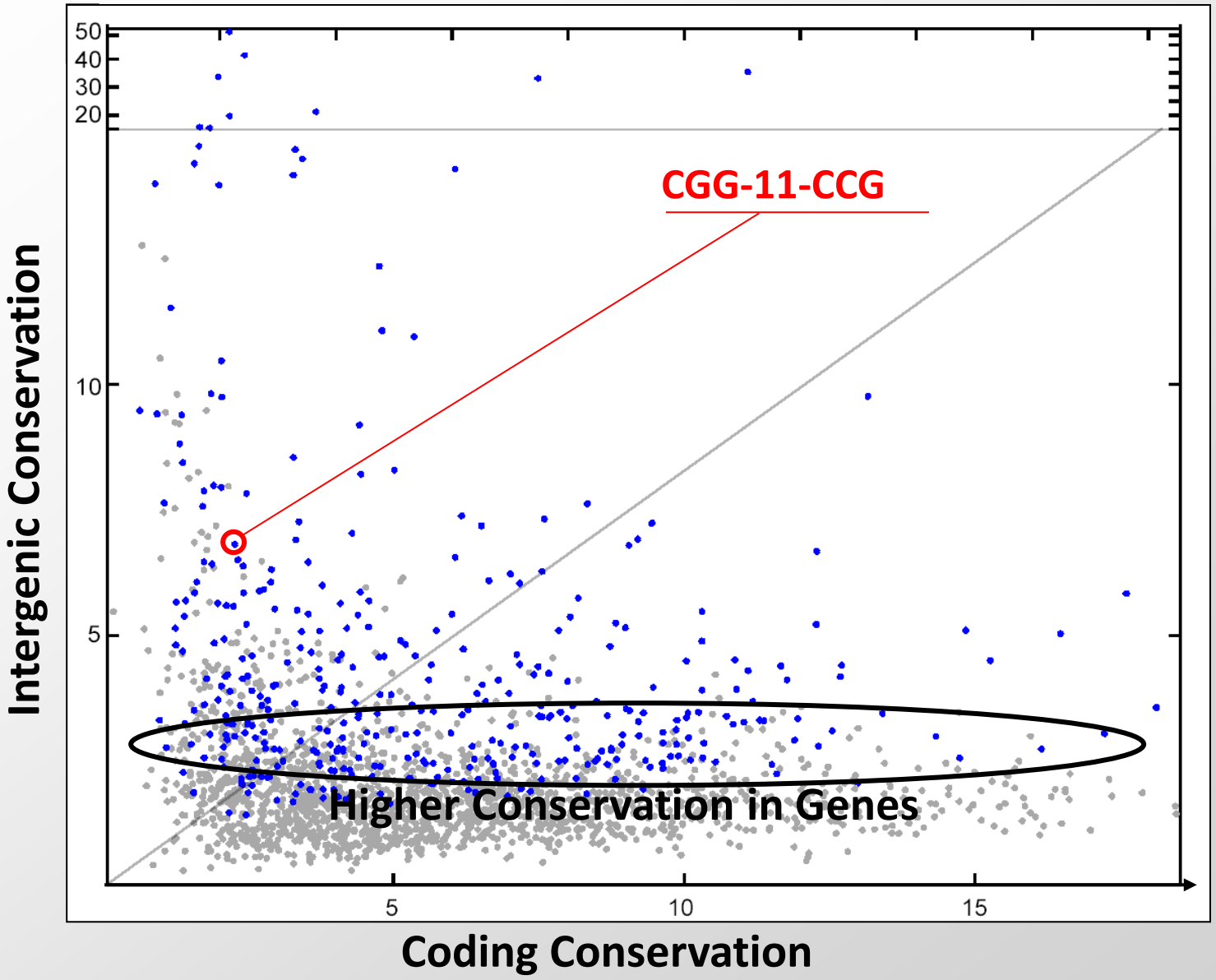
Evaluate conservation within:	Gal4	Controls
(1) All intergenic regions	13%	2%
(2) Intergenic : coding	13% : 3%	2% : 7%
(3) Upstream : downstream	12:0	1:1

A signature for regulatory motifs

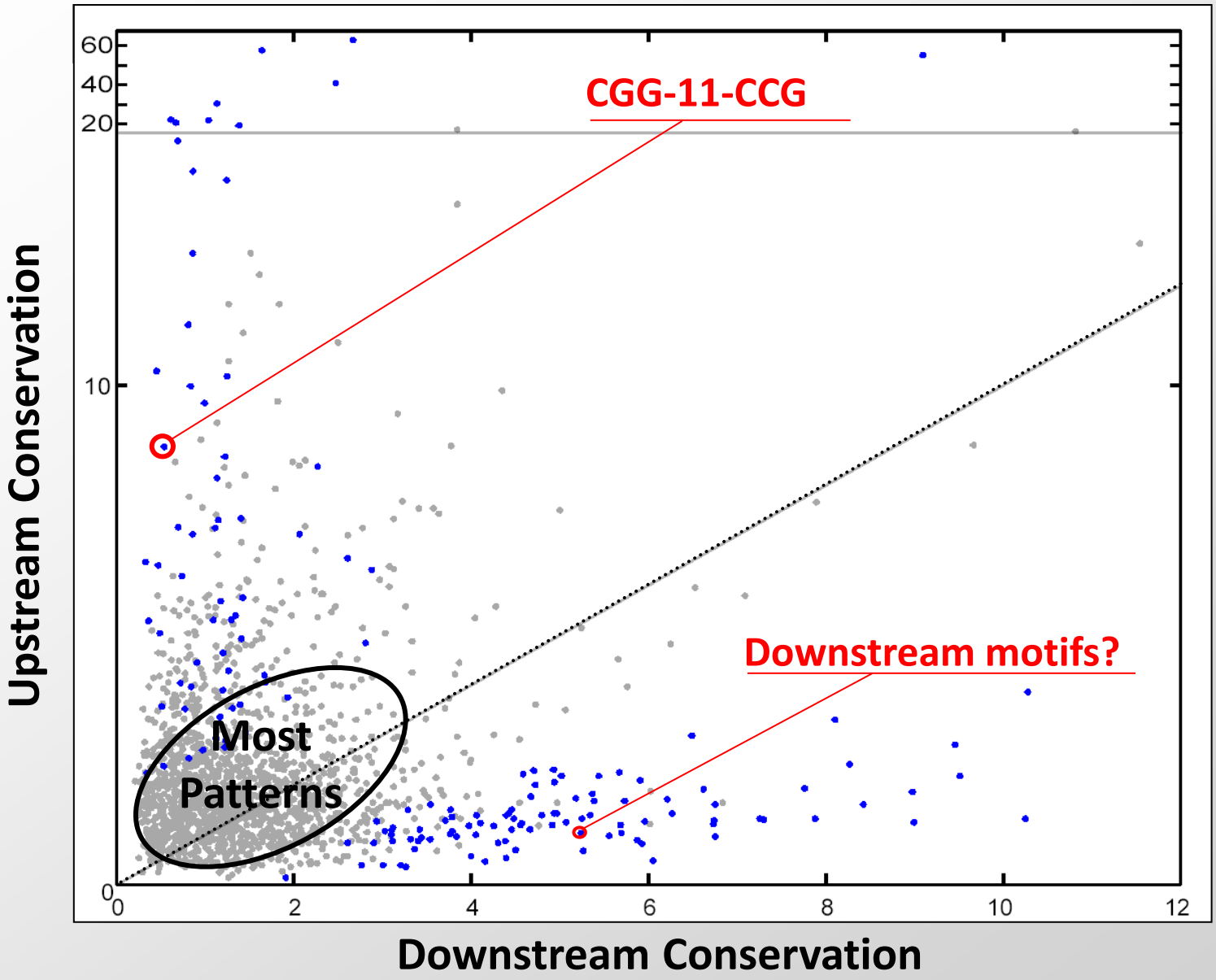
Test 1: Intergenic conservation



Test 2: Intergenic vs. Coding



Test 3: Upstream vs. Downstream



Conservation for TF motif discovery

1. **Enumerate motif seeds**

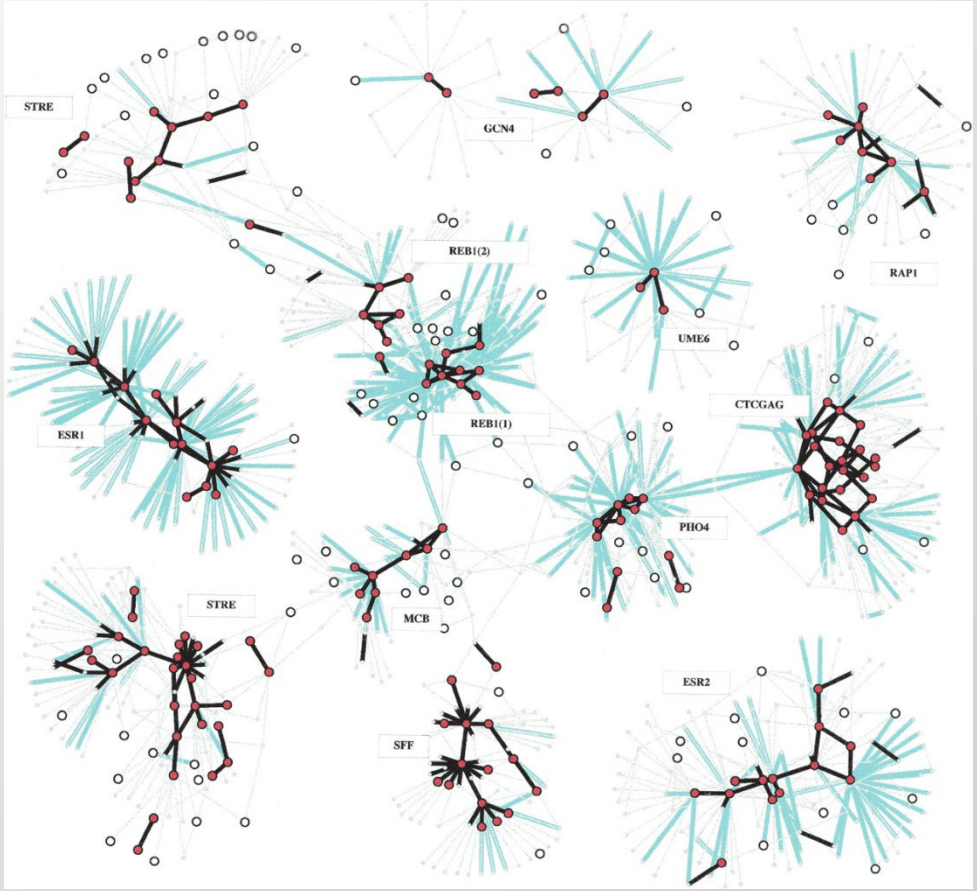
   - Six non-degenerate characters with variable size gap in the middle
2. **Score seed motifs**
   - Use a conservation ratio corrected for composition and small counts to rank seed motifs
3. **Expand seed motifs**

   - Use expanded nucleotide IUPAC alphabet to fill unspecified bases around seed using hill climbing
4. **Cluster to remove redundancy**
   - Using sequence similarity

Kellis, Nature 2003

Learning motif degeneracy using evolution

- Record frequency with which one sequence is "replaced" by another in evolution
- Use this to find clusters of k-mers that correspond to a single motif



Tanay, Genome Research 2004

Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions Z_{ij} from motif matrix
   - M step: Find max-likelihood motif from all positions Z_{ij}
3. Gibbs Sampling: Sample from joint (M,Z_{ij}) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

Validation of the discovered motifs

- Because genome-wide motif discovery is *de novo*, we can use functional datasets for validation
  - Enrichment in co-regulated genes
  - Overlap with TF binding experiments
  - Enrichment in genes from the same complex
  - Positional biases with respect to transcription start
  - Upstream vs. downstream / inter vs. intra-genic bias
  - Similarity to known transcription factor motifs
- Each of these metrics can also be used for discovery
  - In general, split metrics into discovery vs. validation
  - As long as they are *independent* !
  - Strategies that combine them all lose ability to validate
    - Directed experimental validation approaches are then needed

Similarity to known motifs

- If discovered motifs are real, we expect them to match motifs in large databases of known motifs
- We find this (significantly higher than with random motifs)
- Why not perfect agreement?
  - Many known motifs are not conserved
  - Known motifs are biased; may have missed real motifs

MCS	Discovered motif	Known Factor
46.8	**GGGCGG**R	SP-1
34.7	**GCCATnTT**g	YY1
32.7	**CACGTG**	MYC
31.2	G**ATTGGY**	NF-Y
30.8	**TGA**n**TCA**	AP-1
29.7	**GGGAGG**RR	MAZ
29.5	**TGACGTM**R	CREB
26.0	**CGGCCAT**YK	NF-MUE1
25.0	**TGACCTTG**	ERR□
22.6	**CCGGAAR**Y	ELK-1
19.8	S**CGGAAG**Y	GABP
17.9	**CA**T**TTCC**K	STAT1

Xie, Nature 2005

70/174 mammalian motifs

MCS	Discovered motif	Known Factor
65.6	CTAATTAAA	en
57.3	TTKCAATTAA	repo
54.9	WATTRATTK	ara
54.4	AAATTKATGC	prd
51	GCAATAAA	vvl
46.7	DTAATTTRYN	Ubx
45.7	TGATTAAT	ap
43.1	YMATTAAAA	abd-A
41.2	AAACNNGTT	
40	RATTKAATT	
39.5	GCACGTGT	ftz
38.8	AACASCTG	br-Z3

Stark, Nature 2007

35/145 fly motifs

Positional bias of motif matches

- Motifs are involved in initiation of transcription

→Motif matches biased versus TSS

  – 10% of fly motifs
  – 34% of mammalian motifs

→Depletion of TF motifs in coding sequence

  – 57% of fly motifs

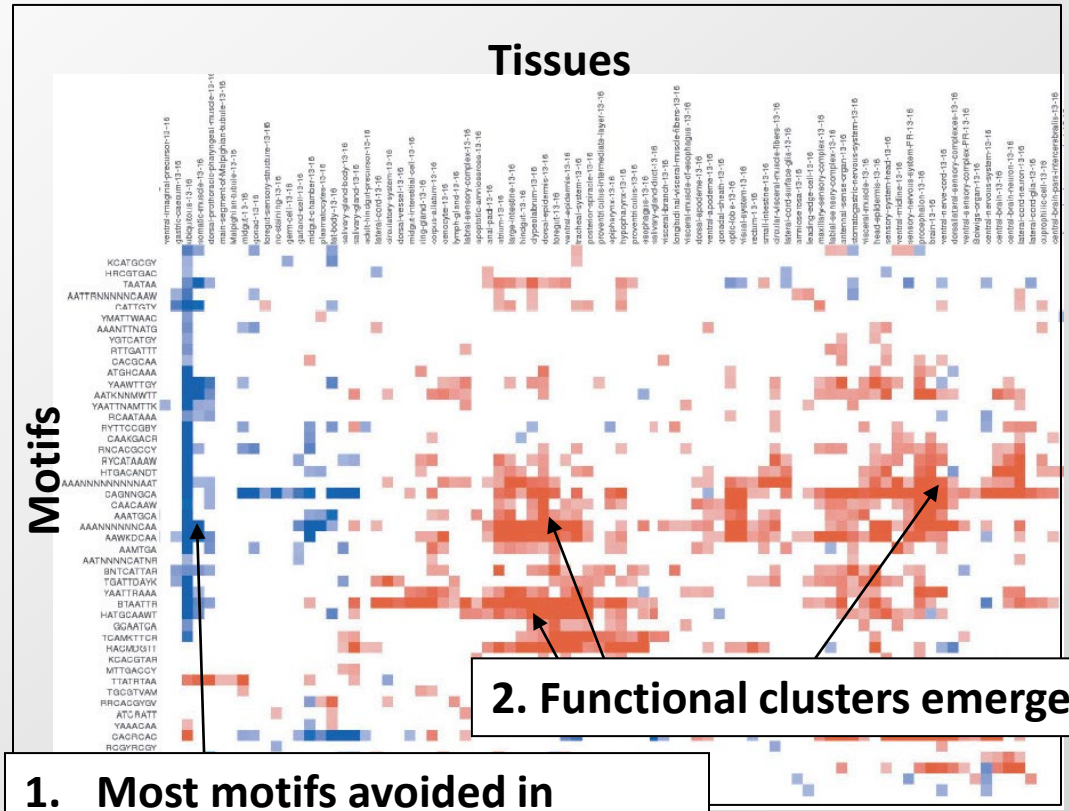
→Clustering of motif matches

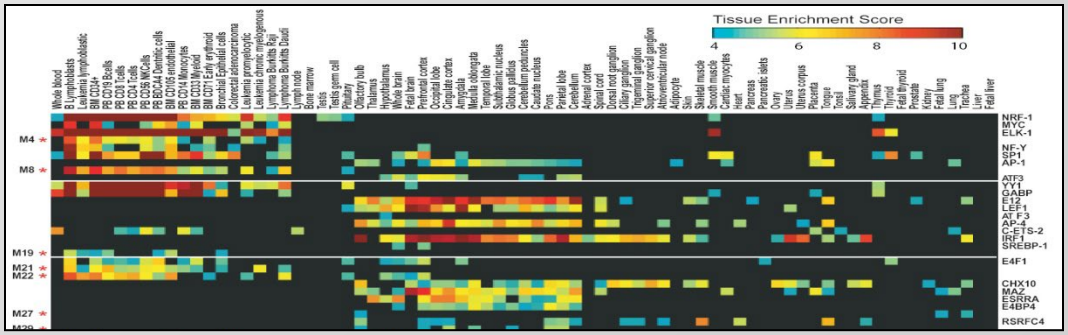
  – 19% of fly motifs

Motifs have functional enrichments

For both fly (top) and mammals (bottom), motifs are enriched in genes expressed in specific tissues

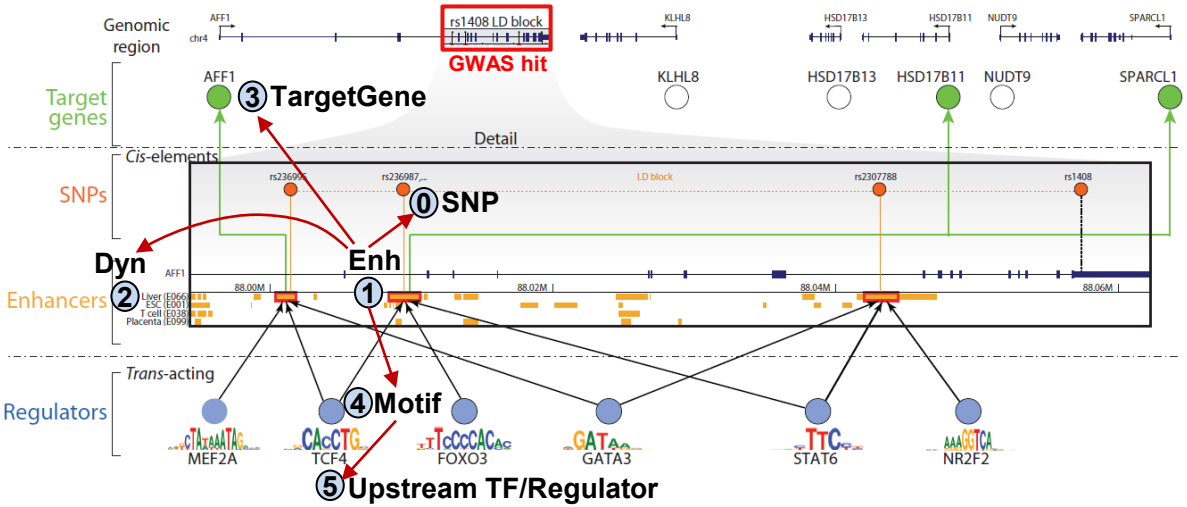
Reveals modules of cooperating motifs





MIT 18.700/18.701/6.047/6.878/HST.507

Computational Biology: Genomes, Networks, Evolution



Lecture 5: Regulatory circuitry

Epigenome Dynamics: Joint Chromatin State Learning

Enhancer-gene linking: Correlation, Hi-C, eQTLs

TF motif discovery: Enrichment, EM, Gibbs Sampling

Deep learning convolution CNNs for motif discovery

Global motif discovery: Comparative Genomics

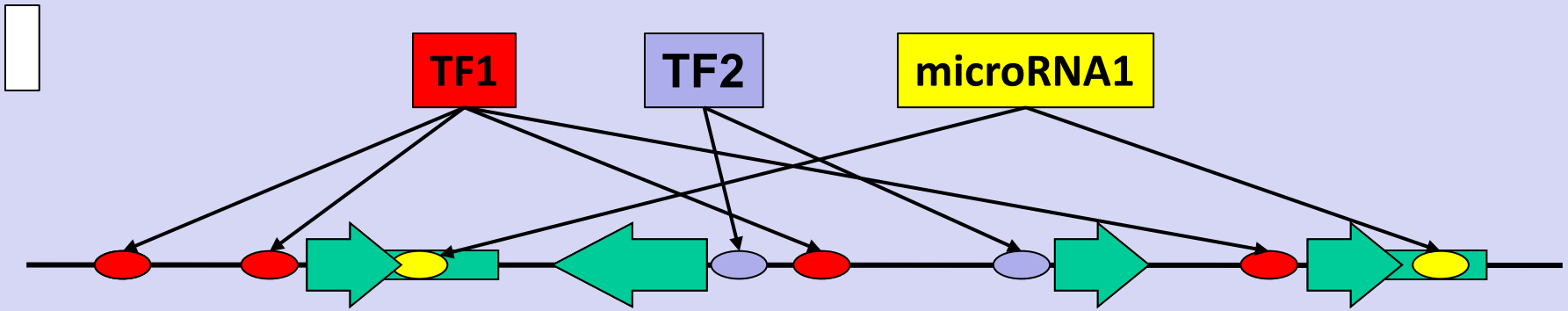
Motif Instance Identification: Branch Length Score

Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions Z_{ij} from motif matrix
   - M step: Find max-likelihood motif from all positions Z_{ij}
3. Gibbs Sampling: Sample from joint (M,Z_{ij}) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays



Motif instance identification

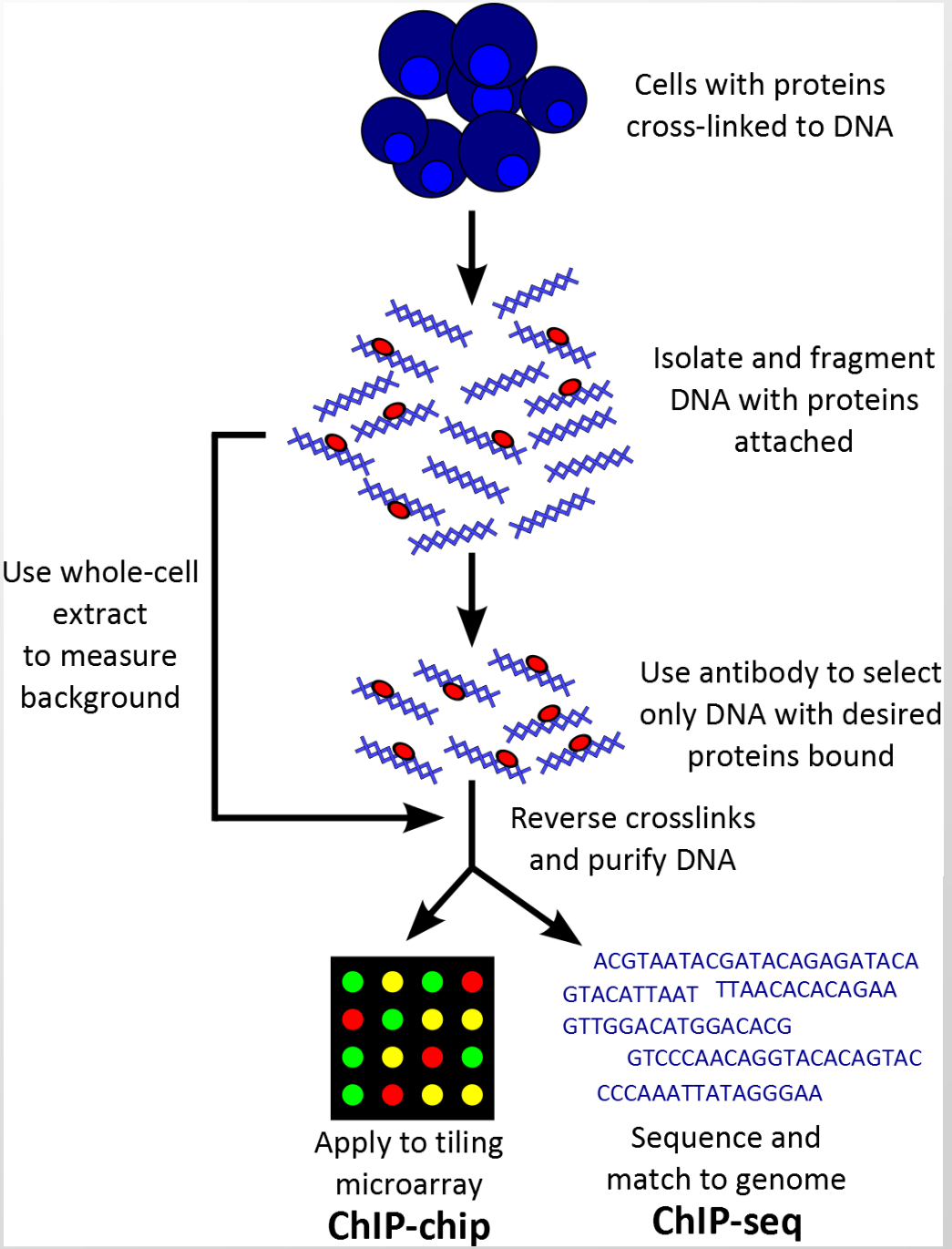
How do we determine the functional binding sites of regulators?

Kheradpour, Stark, Roy, Kellis, Genome Research 2007

Experimental target identification: ChIP-chip/seq

Limitations :

- Antibody availability
- Restricted to specific stages/tissues
- Biological functionality of most binding sites unknown
- Resolution can be limited (can't usually identify the precise base pairs)



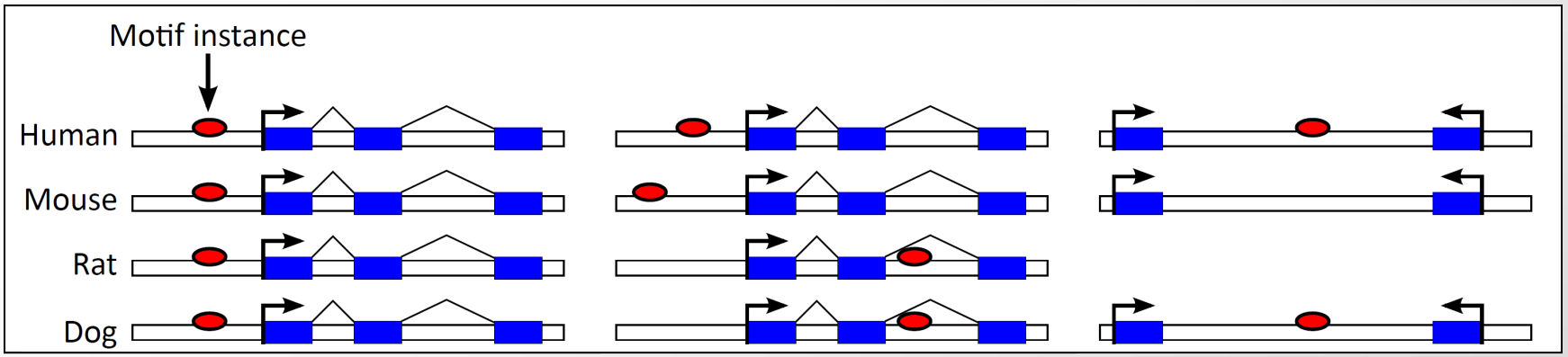
Ren et al., 2000; Iyer et al., 2001 (ChIP-chip)

Robertson et al., 2007 (ChIP-seq)

Computational target identification

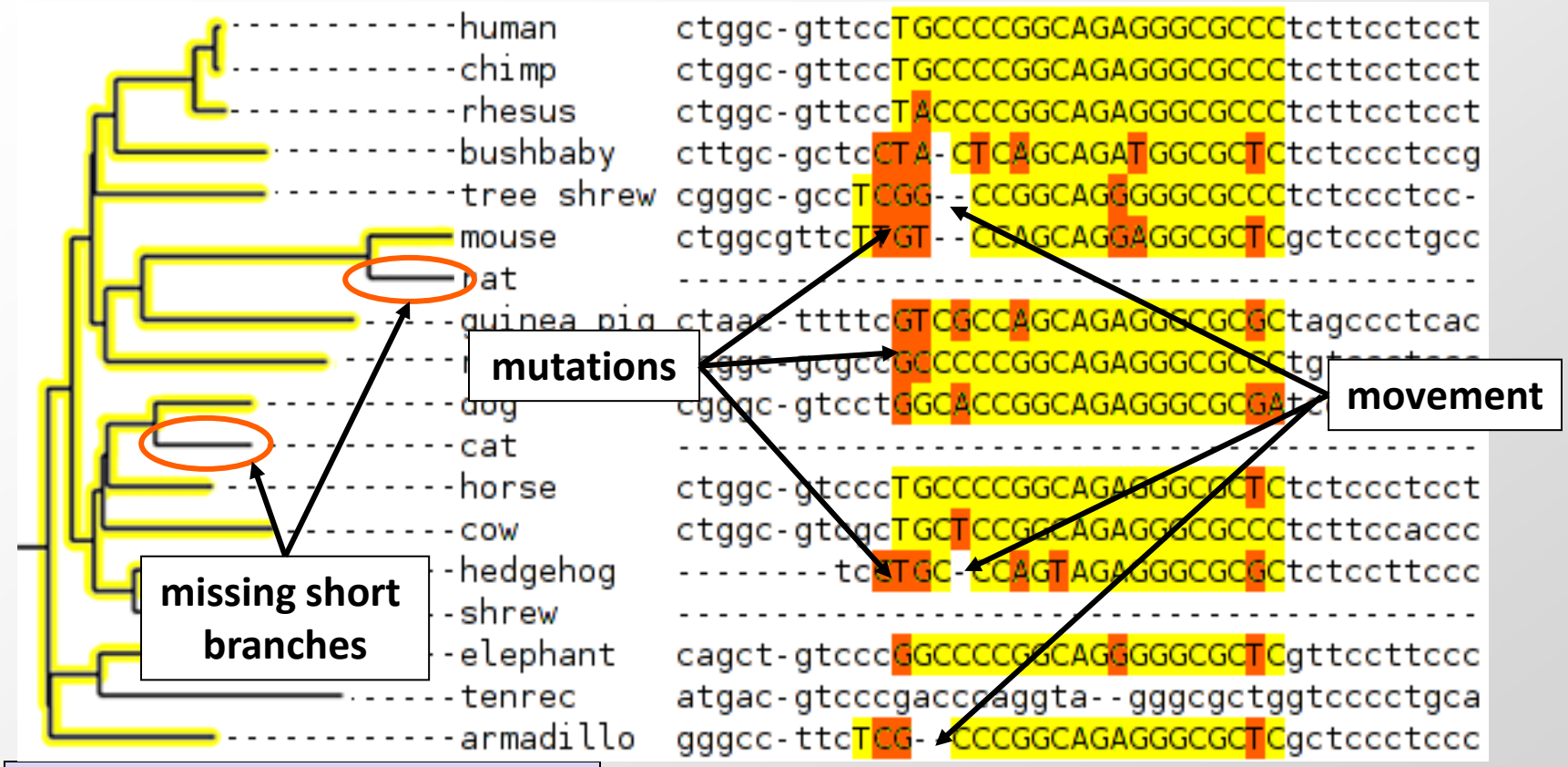
- Single genome approaches using motif clustering (e.g. Berman 2002; Schroeder 2004; Philippakis 2006)
  - Requires set of specific factors that act together
  - Miss instances of motifs that may occur alone
- Multi-genome approaches (phylogentic footprinting) (e.g. Moses 2004; Blanchette and Tompa 2002; Etwiller 2005; Lewis 2003)
  - Tend to either require absolute conservation or have a strict model of evolution

Challenges in target identification

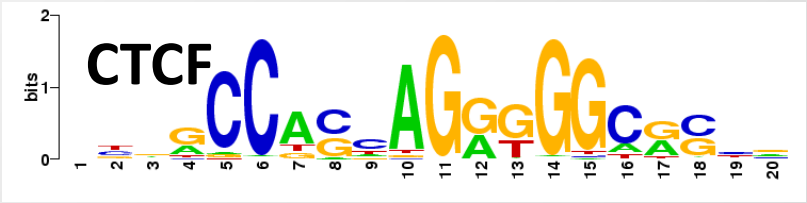


- Simple case
  - Instance fully conserved in orthologous position near genes
- Motif turn-around/movement
  - Motif instance is not found in orthologous place due to birth/death or alignment errors
- Distal/missing matches
  - Due to sequencing/assembly errors or turnover
  - Distal instances can be difficult to assign to gene

Computing Branch Length Score (BLS)



BLS = 2.23sps (78%)



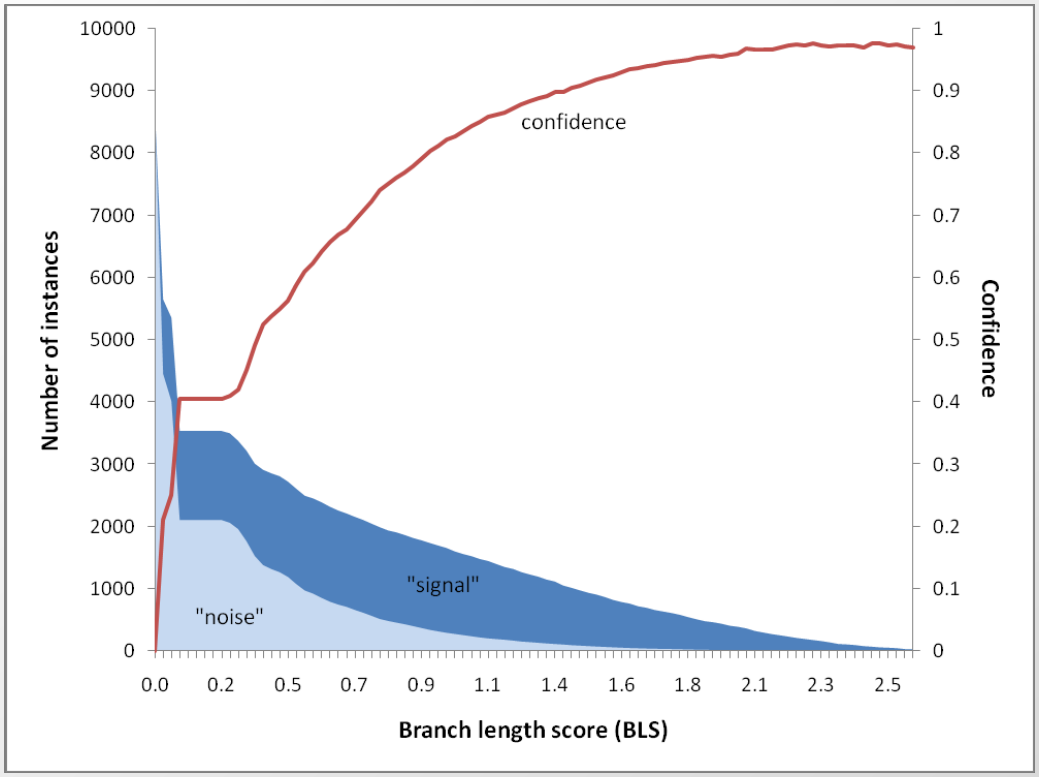
Allows for:

1. **Mutations permitted by motif degeneracy**
2. **Misalignment/movement of motifs within window (up to hundreds of nucleotides)**
3. **Missing motif in dense species tree**

Branch Length Score → Confidence

1. Evaluate chance likelihood of a given score
   - Sequence could also be conserved due to overlap with un-annotated element (e.g. non-coding RNA)
2. Account for differences in motif composition and length
   - For example, short motif more likely to be conserved by chance

Branch Length Score ➔ Confidence



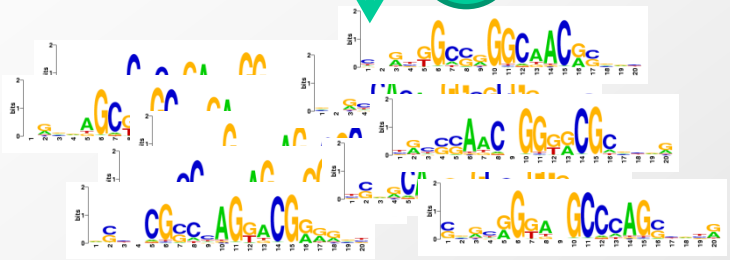
1. **Use motif-specific shuffled control motifs determine the expected number of instances at each BLS by chance alone or due to non-motif conservation**
2. **Compute Confidence Score as fraction of instances over noise at a given BLS (=1 – false discovery rate)**

Producing control motifs

When evaluating the conservation, enrichment, etc, of motifs, it is useful to have a set of "control motifs"

Original motif

1 **Produce 100 shuffles of our original**



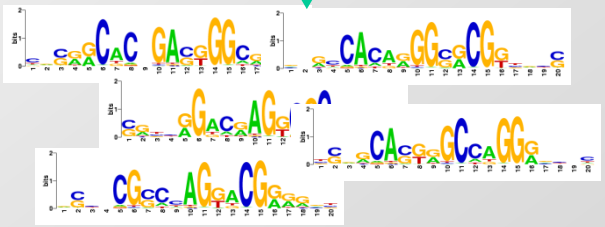
Genome sequence

Known motifs

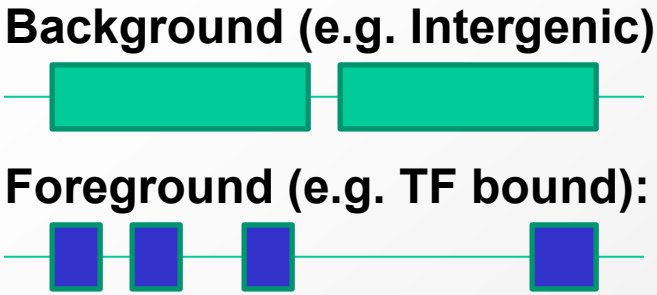
2 **Filter motifs, requiring they match the genome with about (+/- 20%) of our original motif**

3 **Sort potential control motifs based on their similarity to other known motifs**

4 **Cluster potential control motifs and take at most one from each cluster, in increasing order of similarity to known motifs**

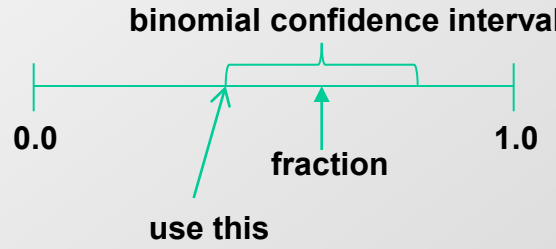


Computing enrichments: background vs. foreground



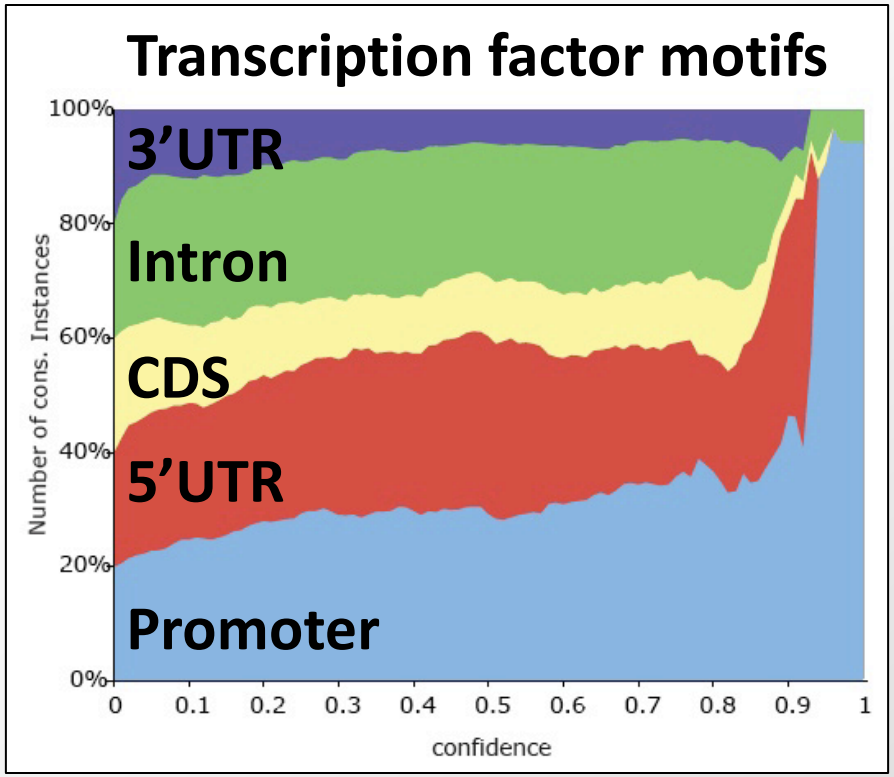
$$\frac{\#\text{ in foreground}}{\#\text{ in background}} \div \frac{\text{size of foreground}}{\text{size of background}}$$

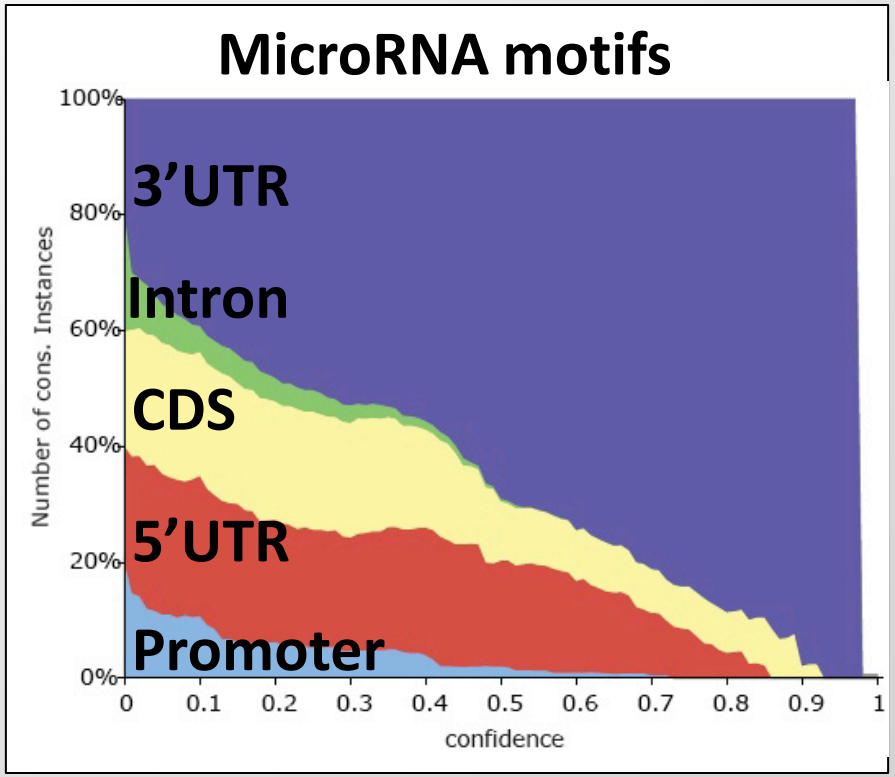
$$\frac{\#\text{ in foreground}}{\#\text{ in background}} \div \frac{\#\text{ control in foreground}}{\#\text{ control in background}}$$



- Background vs. forgeround
  - co-regulated promoters vs. all genes
  - Bound by TF vs. other intergenic regions
- Enrichment: ***fraction of motif instances in foreground*** vs. ***fraction of bases in foreground***
- Correct for composition/conservation level: compute enrichmt w/control motifs
  - Fraction of motif instances can be compared to **fraction of control motif instances in foreground**
  - A hypergeometric p-value can be computed (similar to χ^2, but better for small numbers)
- Fractions can be made more conservative using a binomial confidence interval

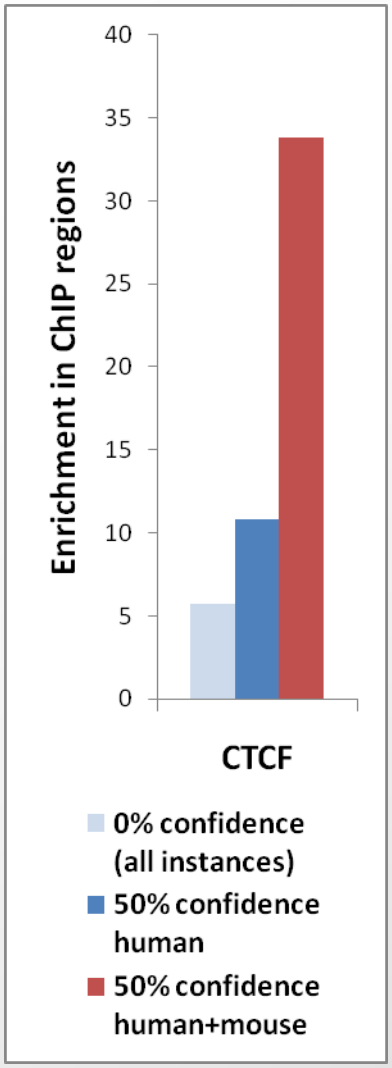
Confidence selects for functional instances



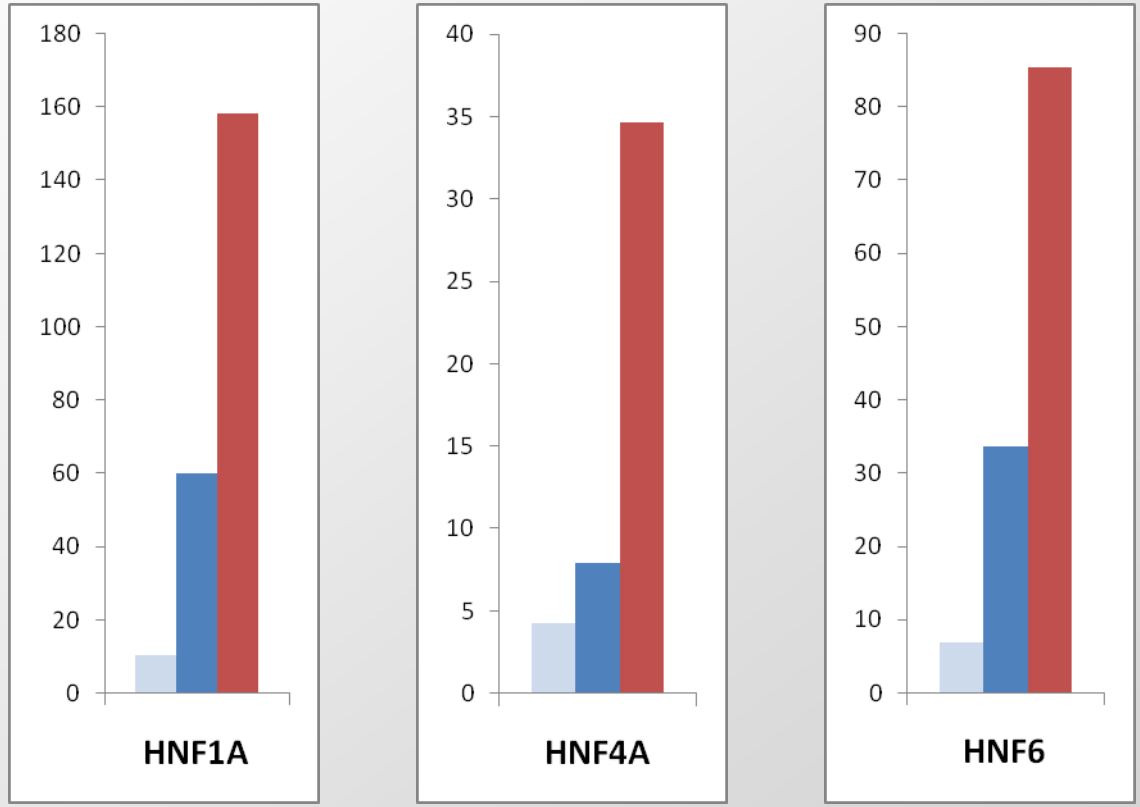


1. Confidence selects for transcription factor motif instances in promoters and miRNA motifs in 3' UTRs

More enrichment when binding conserved

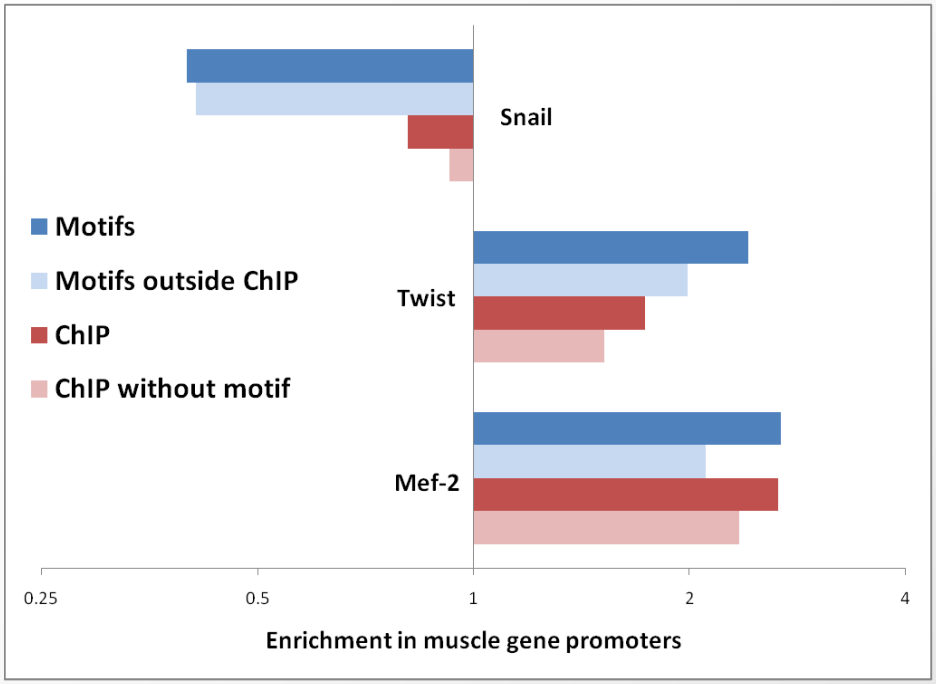


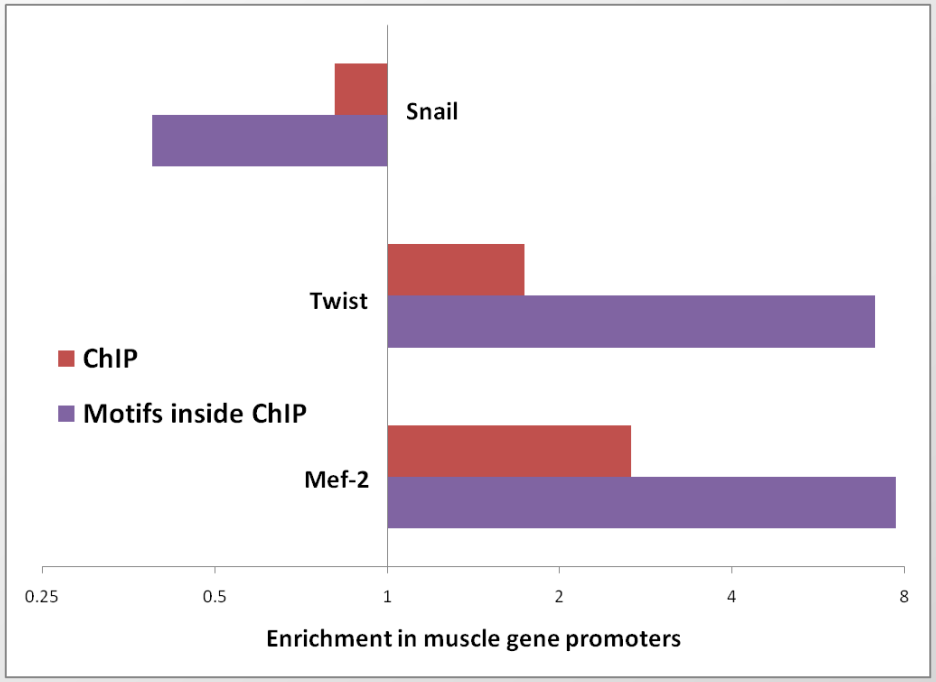
Human: Barski, *et al.*, *Cell* (2007)
Mouse: Bernstein, unpublished



1. ChIP bound regions may not be conserved
2. For CTCF we also have binding data in mouse
3. Enrichment in intersection is dramatically higher
4. Trend persists for other factors where we have multi-species ChIP data

Comparing ChIP to Conservation

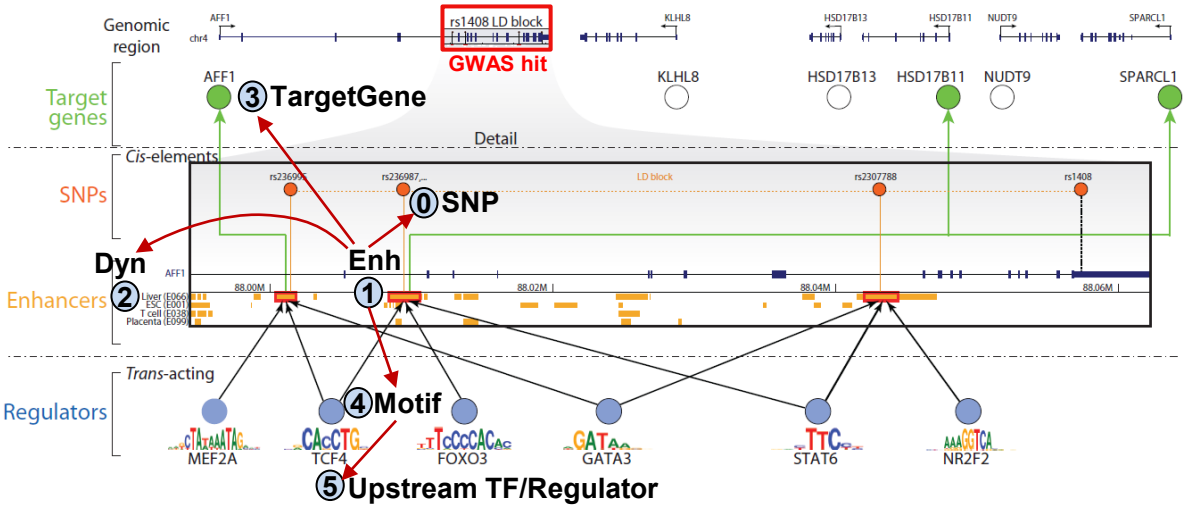




1. **Motifs at 60% confidence and ChIP have similar enrichments (depletion for the repressor Snail) in the functional promoters**
2. **Enrichments persist even when you look at non-overlapping subsets**
3. **Intersection of two regions has strongest signal**
4. **Evolutionary and experimental evidence is complementary**
   - **ChIP includes species specific regions and differentiate tissues**
   - **Conserved instances include binding sites not seen in tissues surveyed**

ChIP data from: Zeitlinger, *et al*., *G&D* (2007); Sandmann, *et al,*. *G&D* (2007); Sandmann, *et al., Dev Cell* (2006)

MIT 18.700/18.701/6.047/6.878/HST.507
Computational Biology: Genomes, Networks, Evolution



Lecture 5: Regulatory circuitry

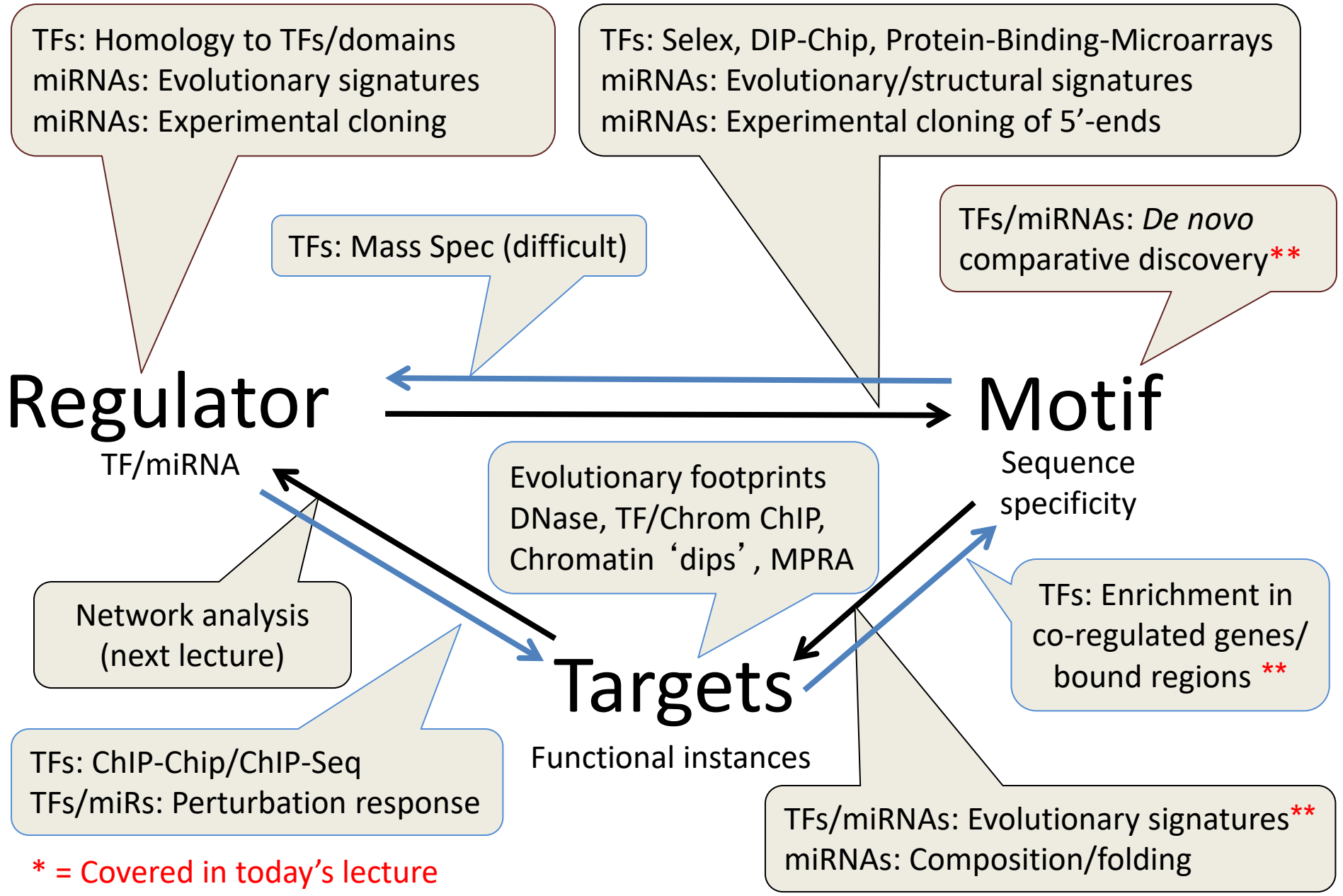
Epigenome Dynamics: Joint Chromatin State Learning
Enhancer-gene linking: Correlation, Hi-C, eQTLs
TF motif discovery: Enrichment, EM, Gibbs Sampling
Deep learning convolution CNNs for motif discovery
Global motif discovery: Comparative Genomics
Motif Instance Identification: Branch Length Score
Regulatory region dissection: MPRA, HiDRA

Prof. Manolis Kellis – manoli@mit.edu – MIT Computer Science & AI Lab

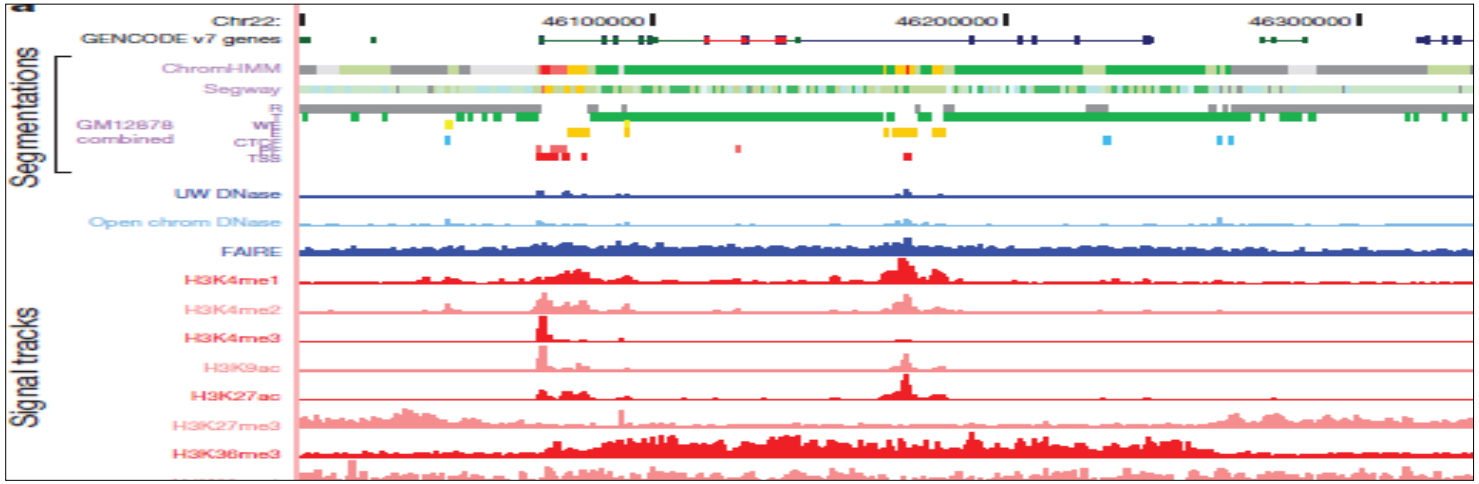
Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions Z_{ij} from motif matrix
   - M step: Find max-likelihood motif from all positions Z_{ij}
3. Gibbs Sampling: Sample from joint (M,Z_{ij}) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

Challenges in regulatory genomics



From Identification to Large-Scale Confirmation and Dissection of Candidate Regulatory Regions



ENCODE Project Consortium et al, 2012

ENCODE, Roadmap Epigenomics, *et al*: Histone marks, TF binding, DNase, FAIRE, …

➔ identification of **candidate** regulatory regions

Next challenge: confirm/dissect 10,000s of regions!

- Test **thousands** of candidate regulatory regions at once
- Identify regulatory positions at or near **nucleotide level** resolution independent of sequence motifs
- Distinguish **activating vs. repressive** nucleotides

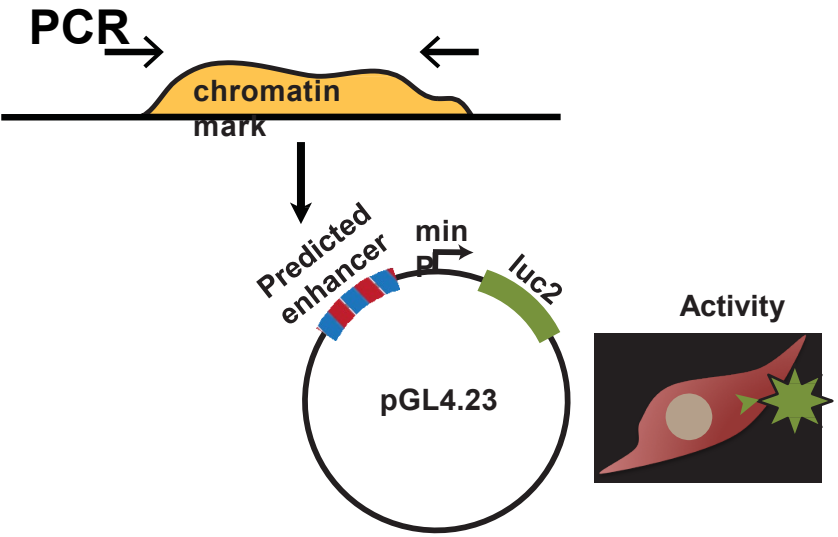
Problem: Not all annotated enhancers are real

VISTA Enhancer Browser
whole genome enhancer browser

2659 *in vivo* tested elements
1444 elements with enhancer activity

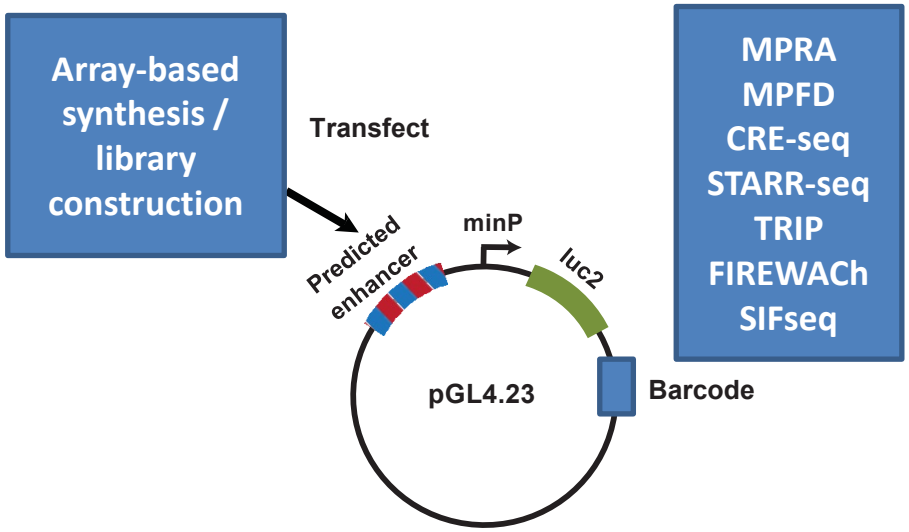
Visel *et al.* NAR 2007

Luciferase assays



Slow, tedious, time-consuming

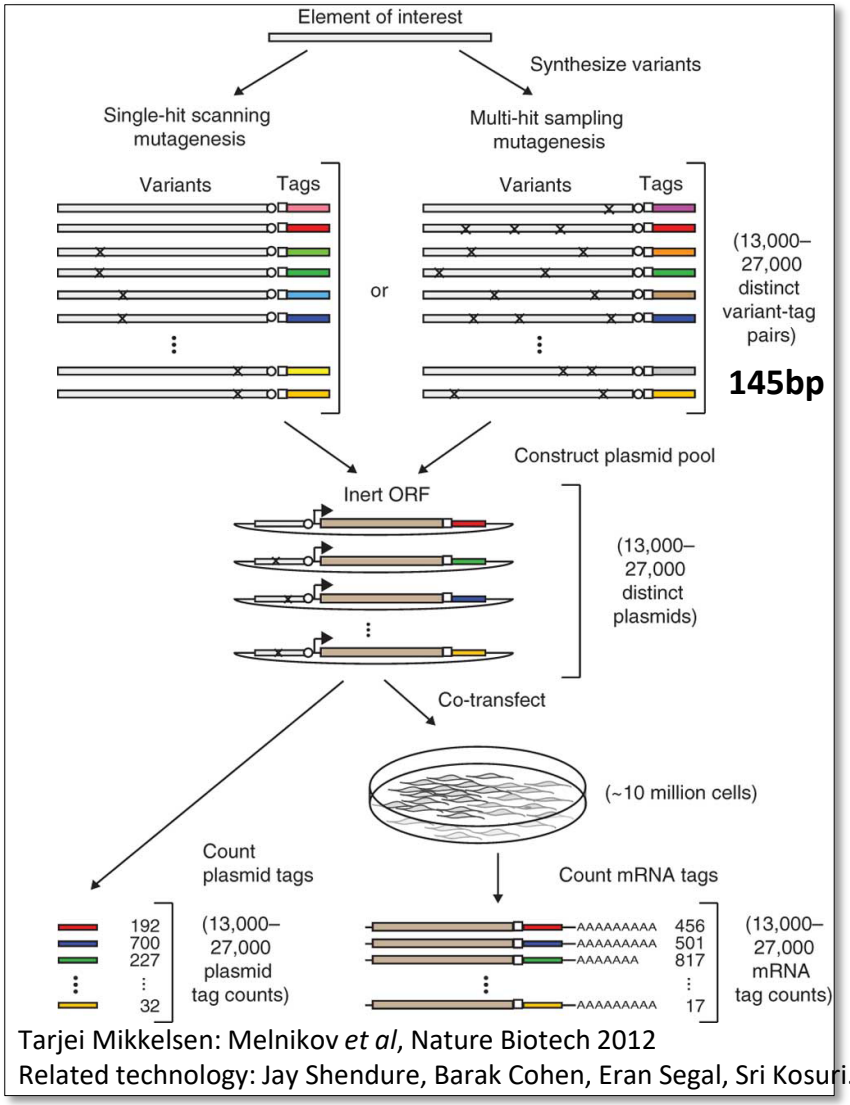
Massively-parallel assays



10,000+ elements at a time

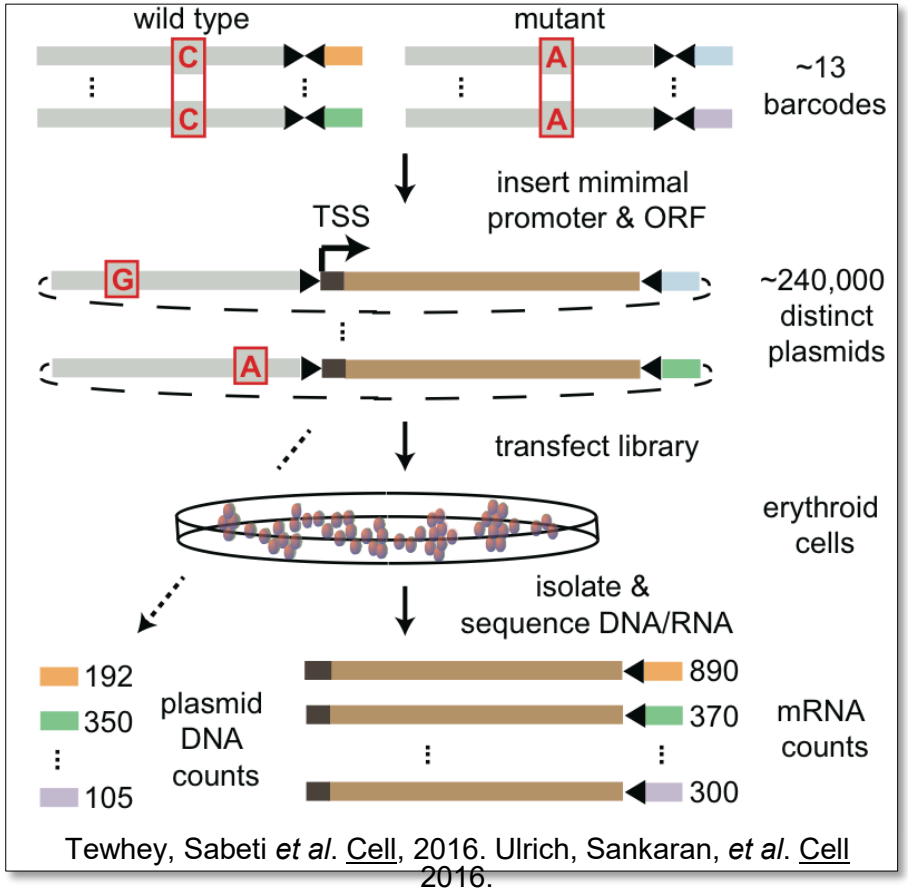
Difference between endogenous epigenomic signatures (e.g. H3K27ac)
vs. being able to actually drive expression of a reporter gene
(take DNA sequence segment out of context)

Enabling Technology: Massively Parallel Reporter Assay (MPRA)



Tarjei Mikkelsen: Melnikov *et al*, Nature Biotech 2012
Related technology: Jay Shendure, Barak Cohen, Eran Segal, Sri Kosuri…

- Synthesize many enhancer versions ➔ insert upstream
- Couple each with a barcode ➔ insert downstream
- Make 10,000s of elements ➔ plasmids, transfection
- High-throughput test in diff. cell types ➔ 10k measurements

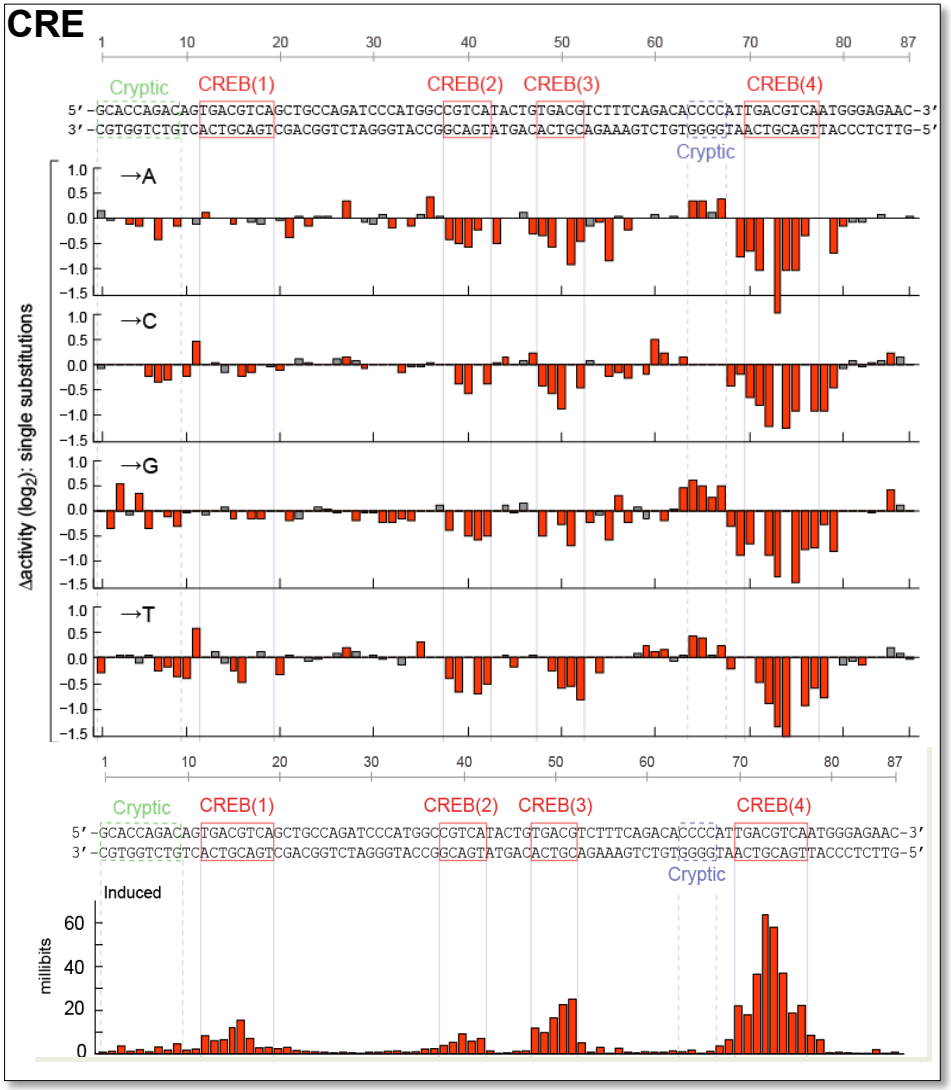


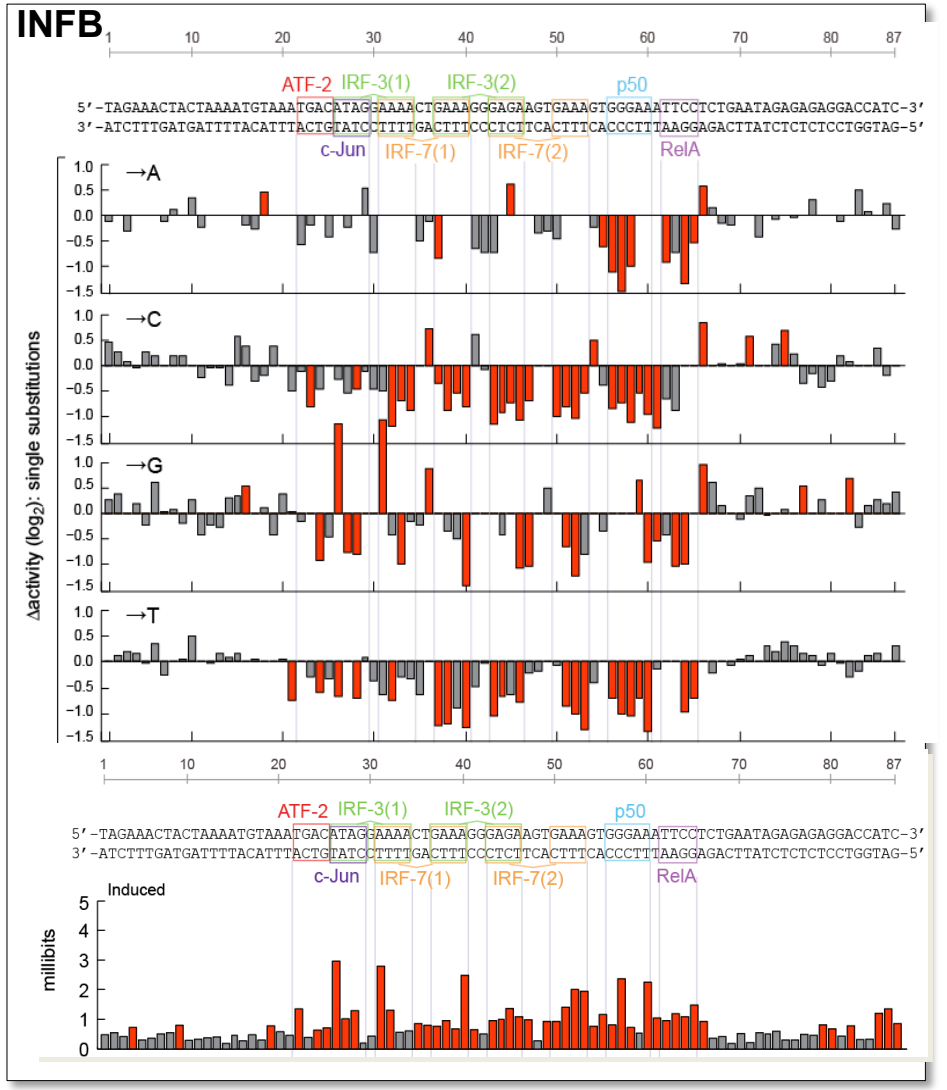
Tewhey, Sabeti *et al*. Cell, 2016. Ulrich, Sankaran, *et al*. Cell 2016.

Application: Test 10,000 variants in 1 experiment

Can we achieve (1) **large scale application** (2) **nucleotide level resolution**, and (3) **direction of effect**, all without knowing motifs or precise 145bp to test?

High-resolution tiling dissection of individual regulatory regions

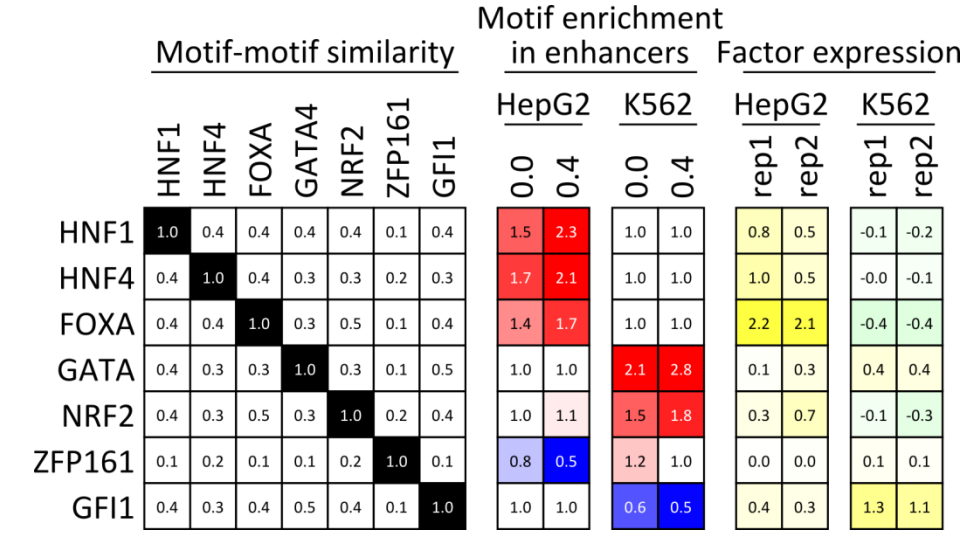


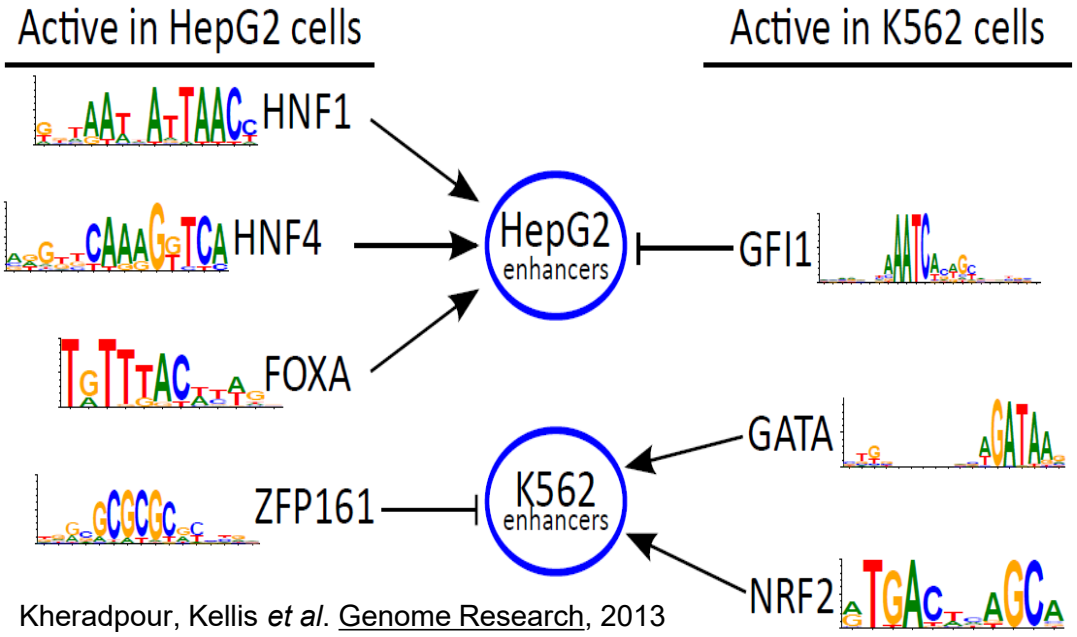


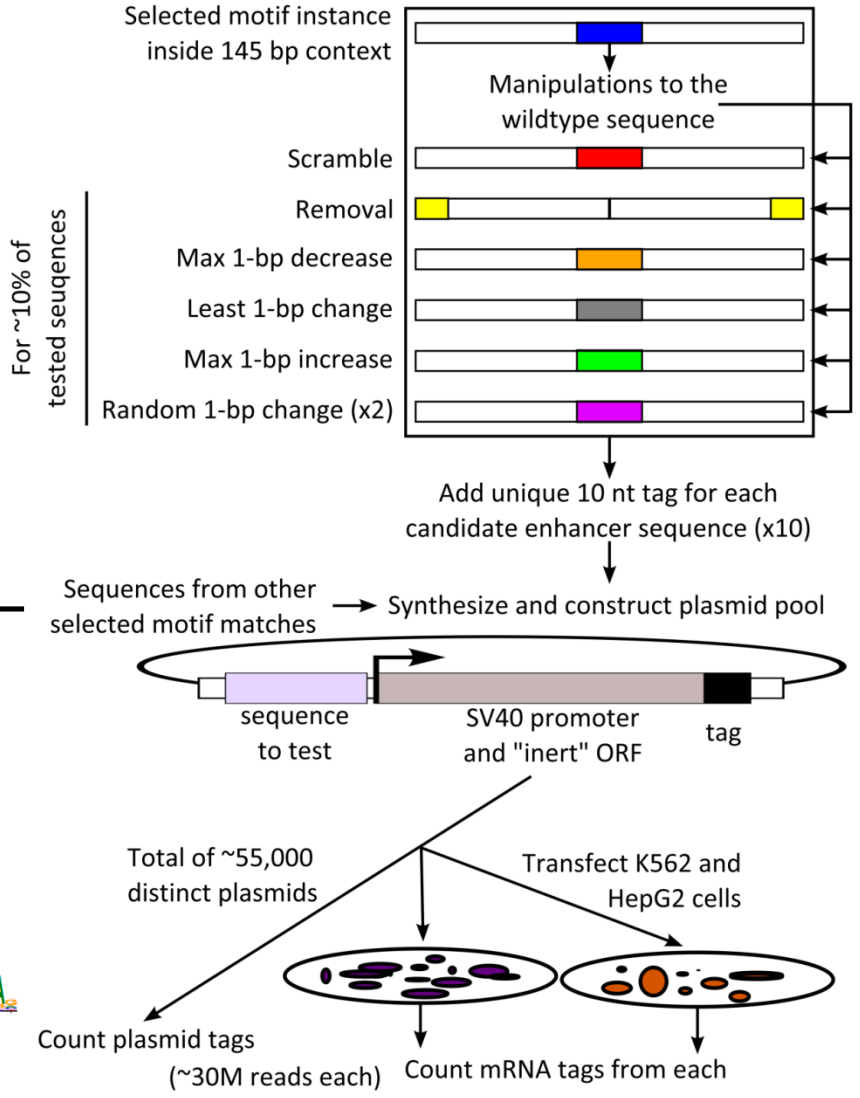
Challenge: hundreds of constructs needed for each region
Can test thousands of regions jointly?

Systematic motif disruption in 2000 regions

for 5 activators and 2 repressors in 2 human cell lines



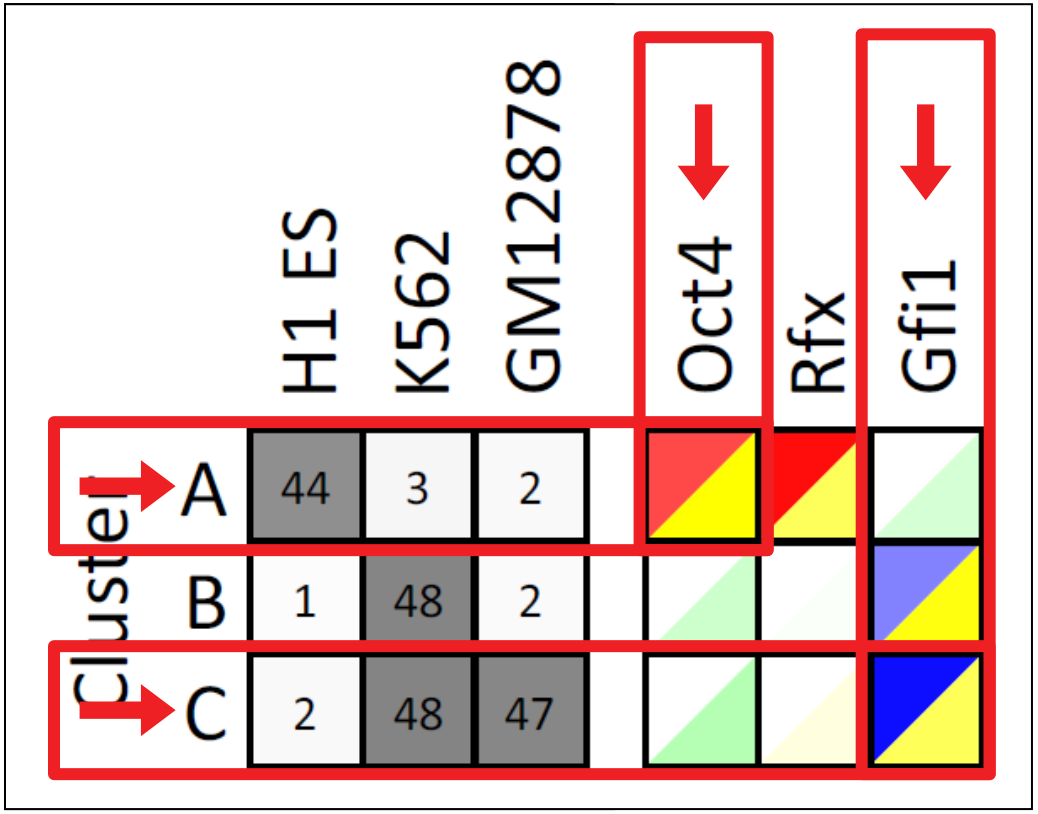


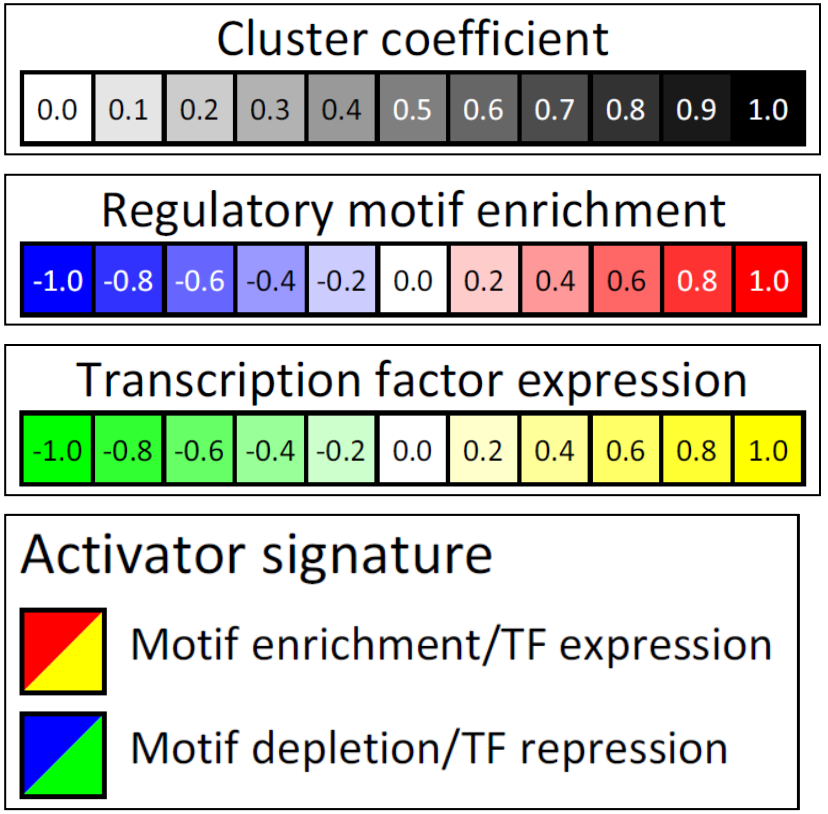


54000+ measurements (x2 cells, 2x repl)

Kheradpour, Kellis *et al*. Genome Research, 2013

What to perturb: Guided by computational predictions

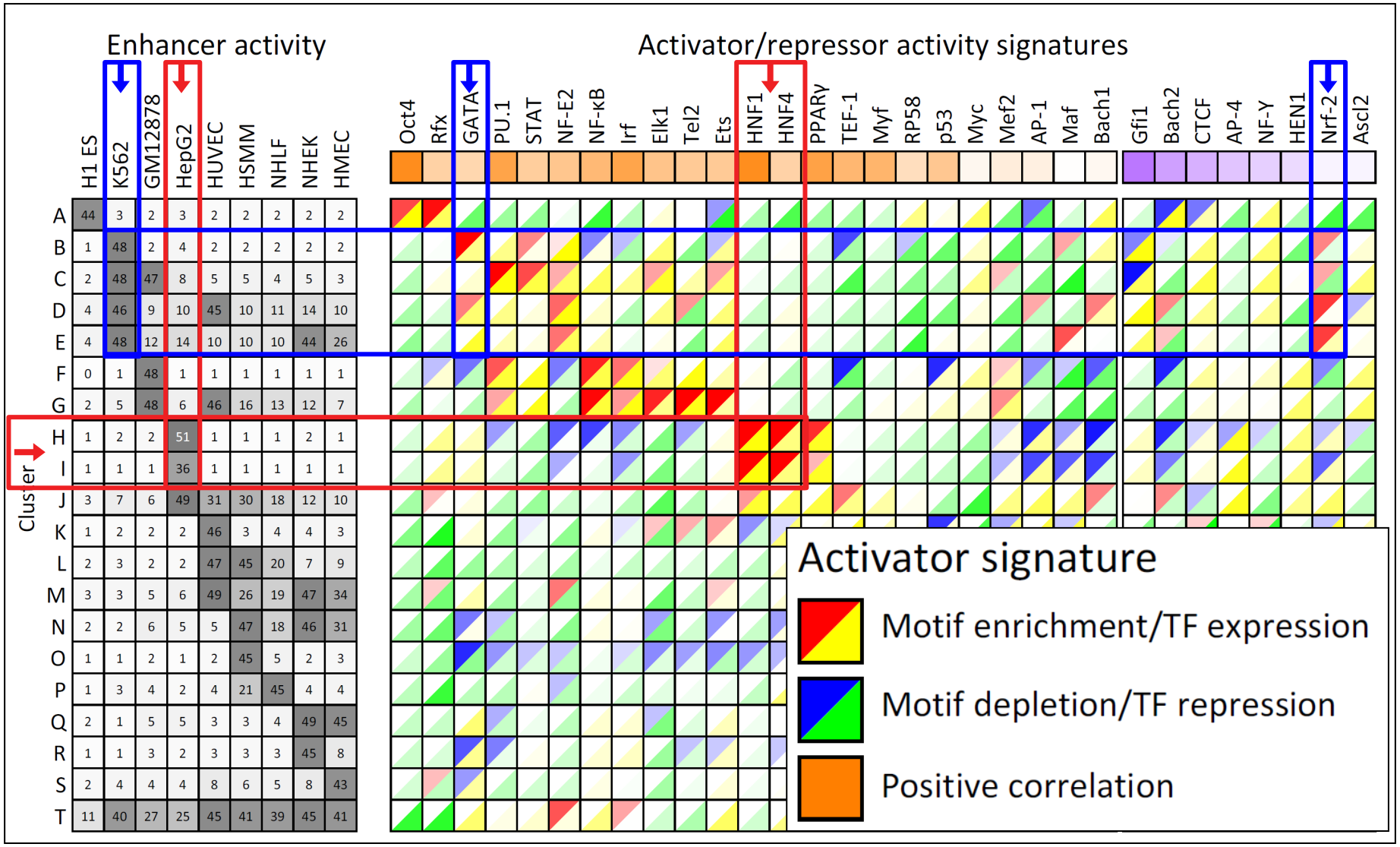




- Chromatin mark-based cell line specific enhancers
- Oct4 predicted activator of embryonic stem cells
- Gfi1 predicted repressor K562/GM12878 cells

Coordinated activity reveals activators/repressors

HNF1 and HNF4 are predicted activators of HepG2 enhancers



- Model: Disruption of the motif site would abolish enhancer state

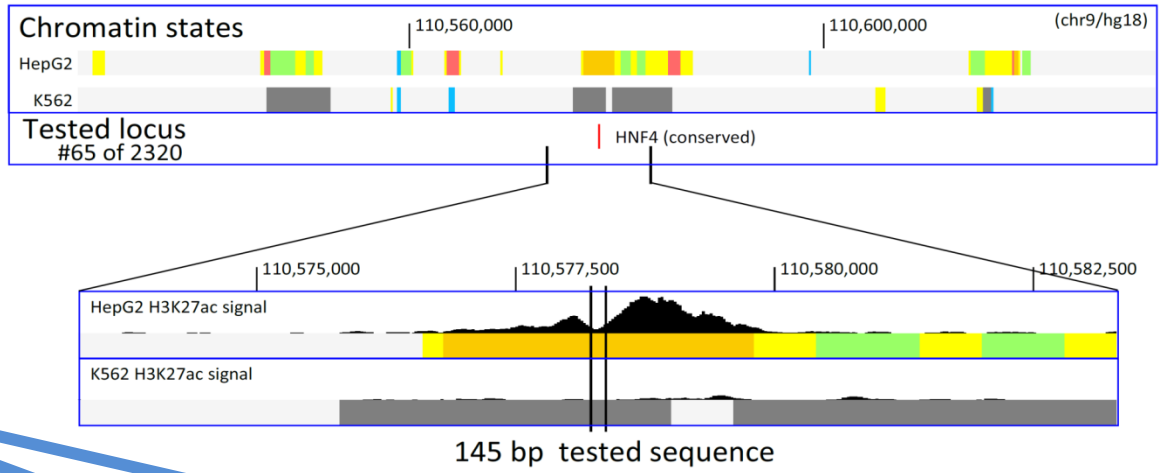
Example activator: conserved HNF4 motif match

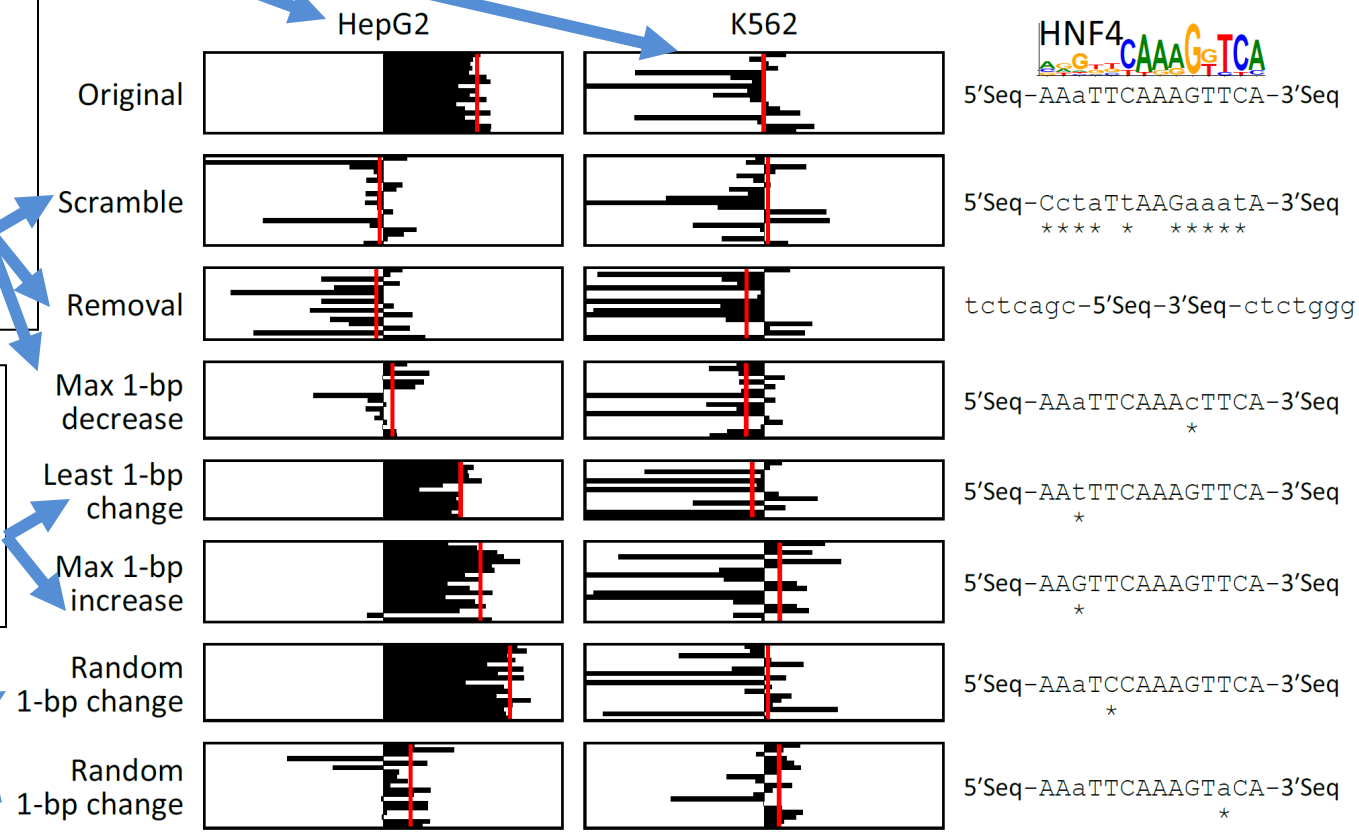
WT expression specific to HepG2

Motif match disruptions reduce expression to background

Non-disruptive changes maintain expression

Random changes depend on effect to motif match

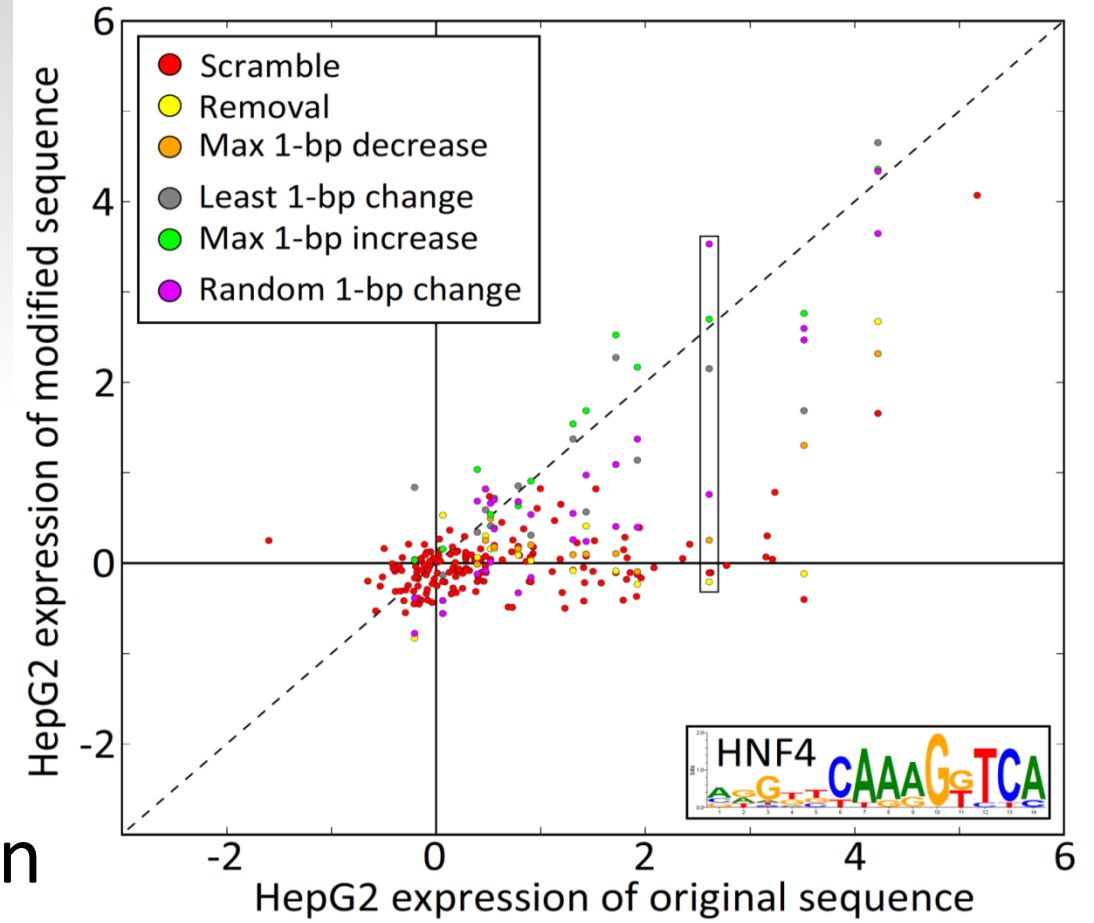


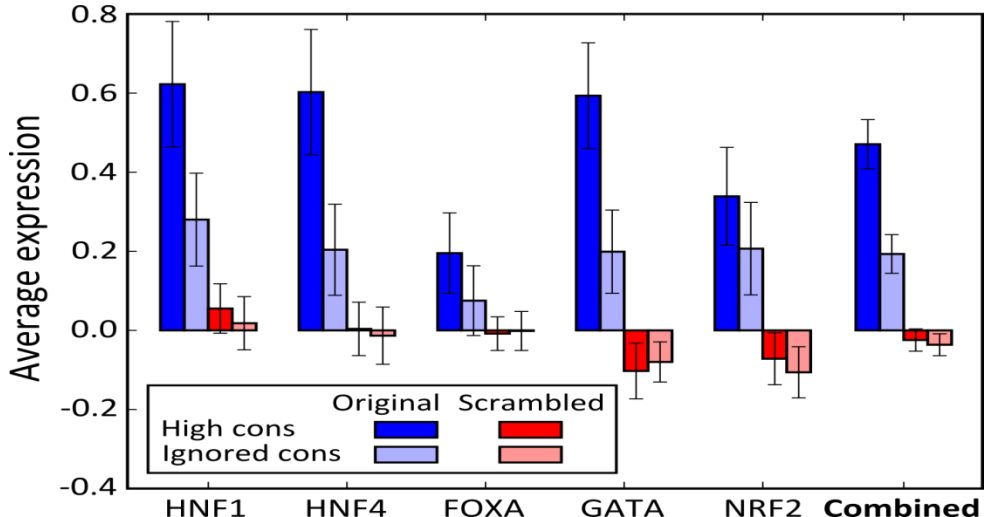


Results hold across 2000+ enhancers

- Scramble abolishes reporter expression
- Neutral mutations show no change
- Increasing mutations show more expression

- Repressor mutations ➔ expression increase
- Motif context matters

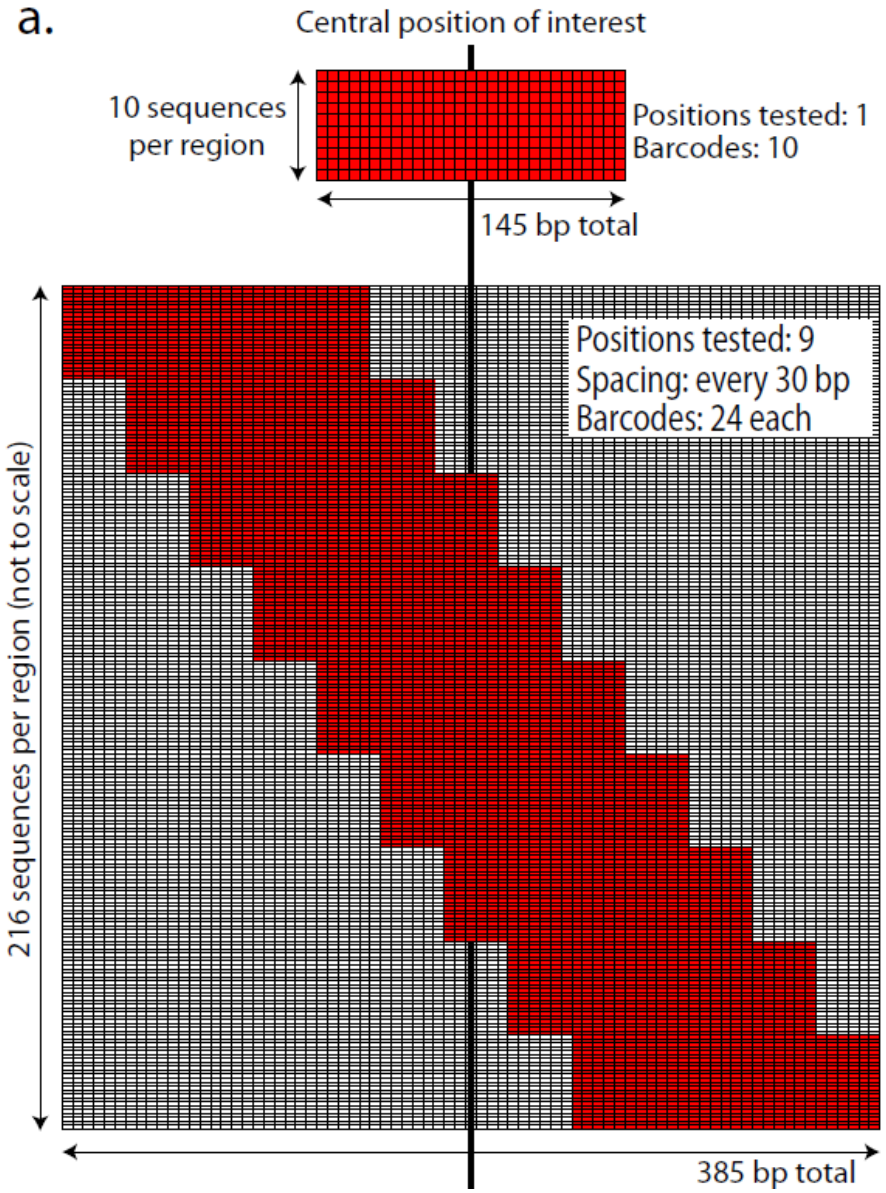




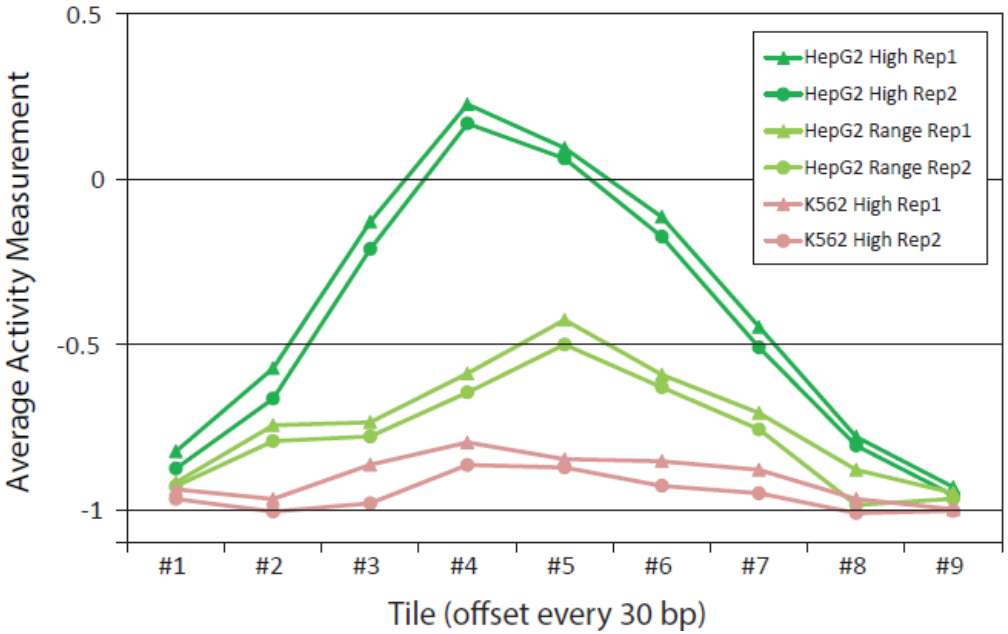
Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions Z_{ij} from motif matrix
   - M step: Find max-likelihood motif from all positions Z_{ij}
3. Gibbs Sampling: Sample from joint (M,Z_{ij}) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

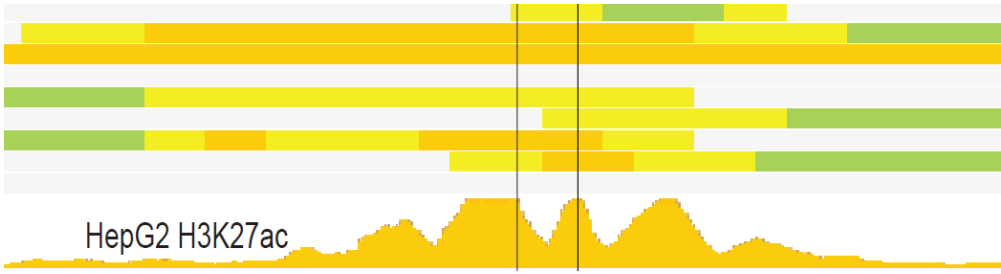
Effect of enhancer position on reporter activity



Tile regions at 30-bp offsets

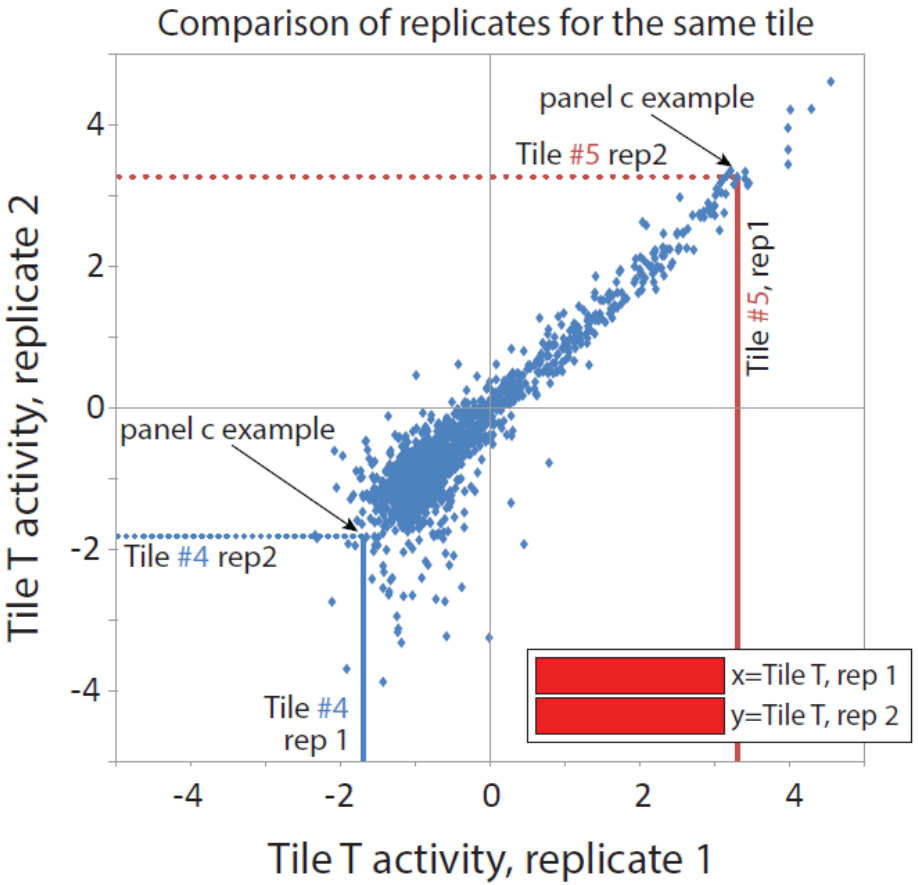


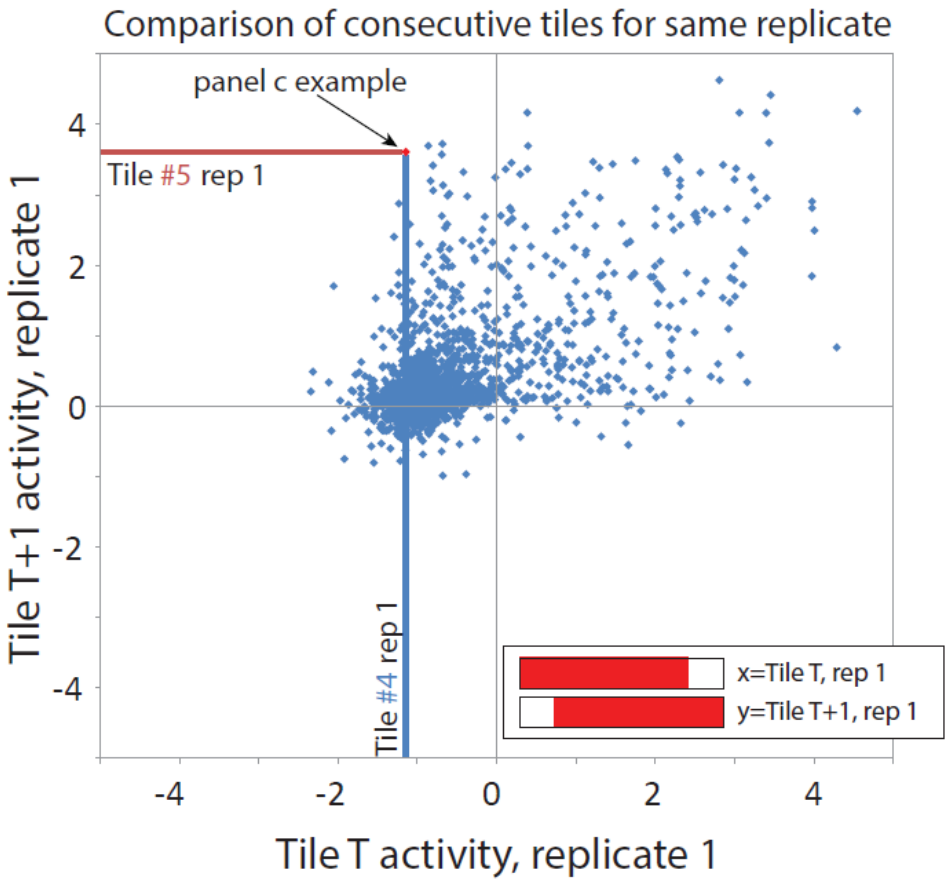
Centers of selected regions show strongest activity



Chromatin dips in matched cell show strongest activity

An offset of 30-bp can make a big difference

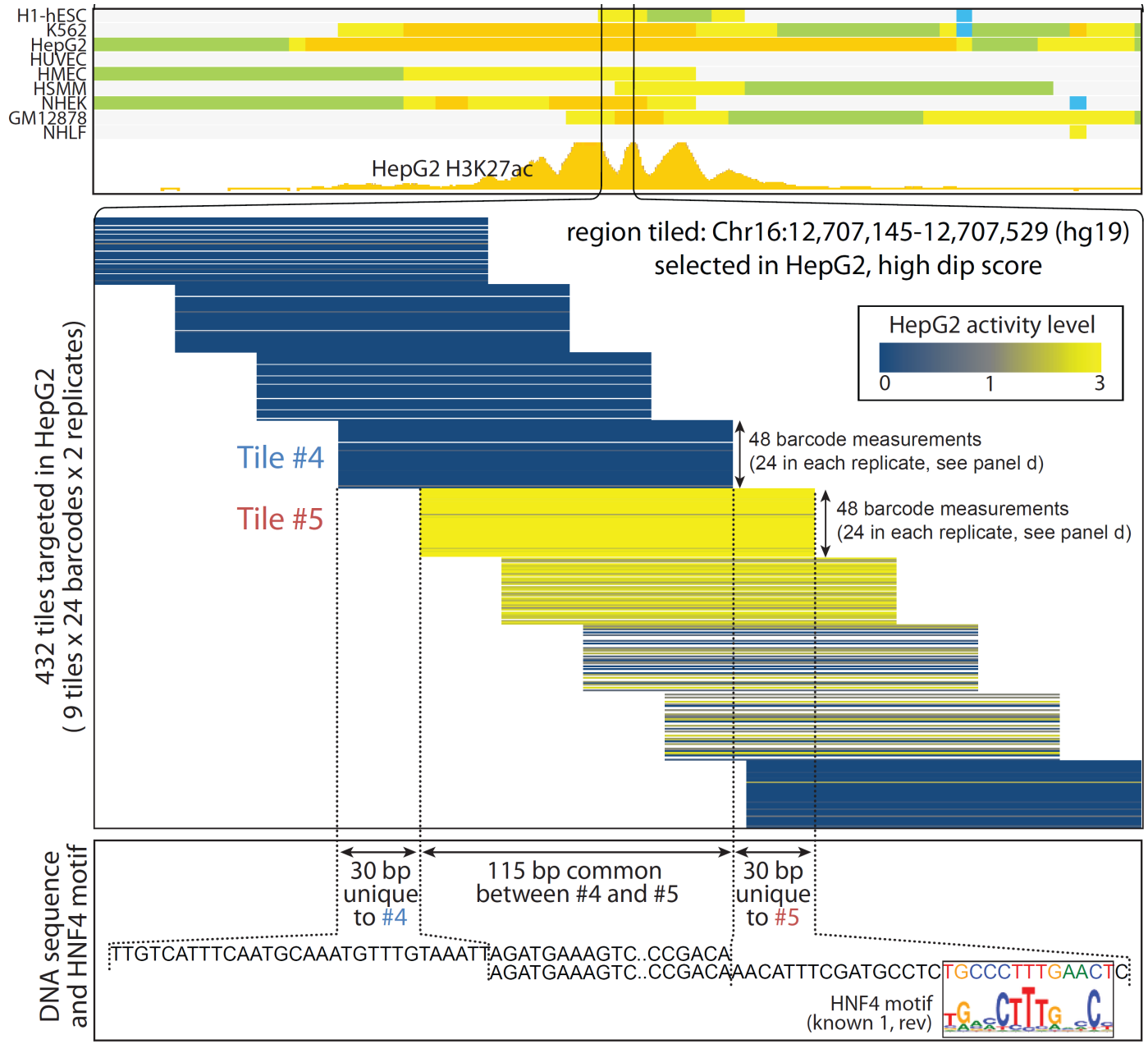




Replicates of same tile are highly consistent

Consecutive tiles can differ greatly

Consecutive tile diffs due to motif inclusion/exclusion



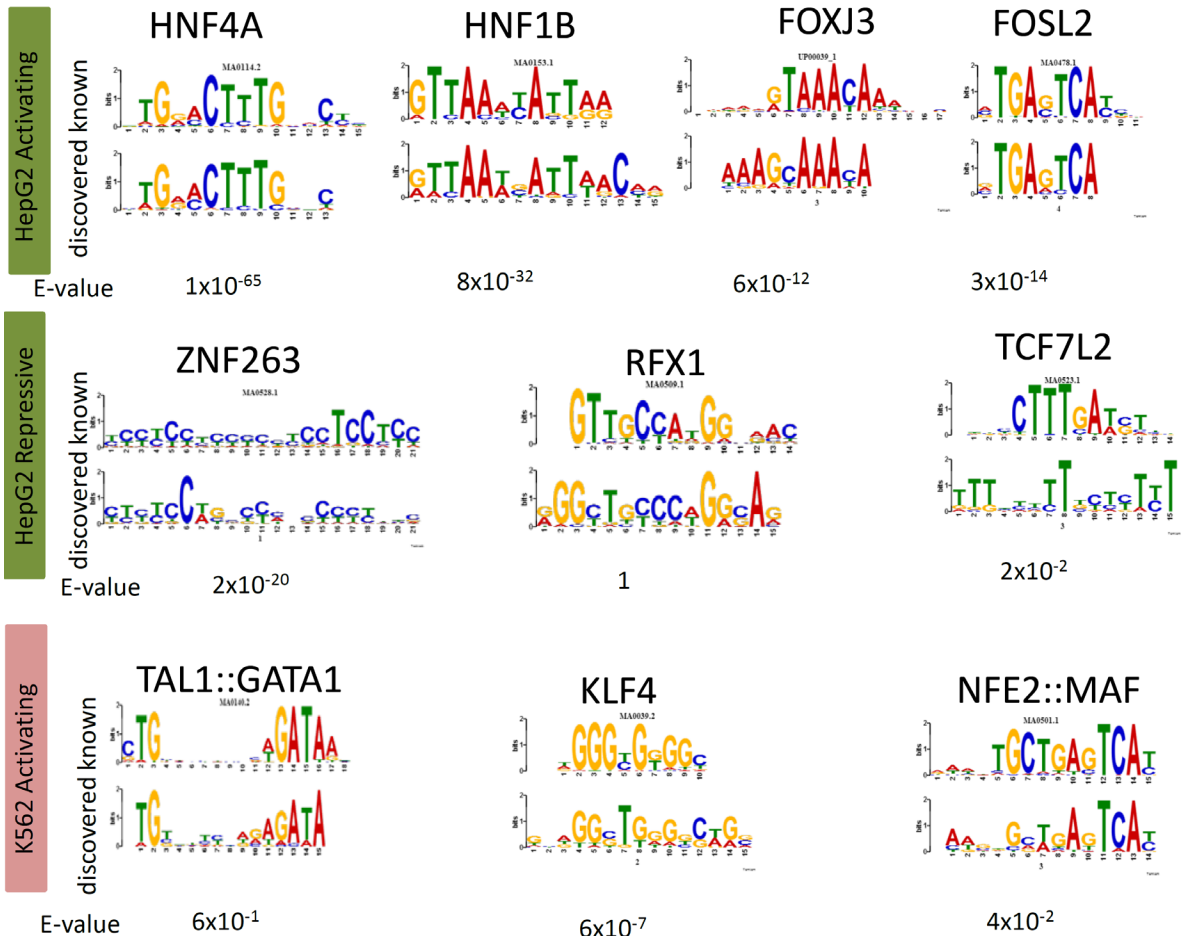
- **Inclusion/ exclusion of 30-bp intervals**
  - ➢ **Akin to systematic disruption**
  - ➢ **Increase resolution from tile (145bp) to offset (30bp)**

Applications:

- **Use to discover motifs?**
- **Further increase resolution?**

Tile differences allow motif discovery

	HepG2 Activating	HepG2 Repressing	K562 Activating	K562 Repressing
GATA_known14	1.0	1.0	6.4	1.0
LMO2_2	1.0	1.0	4.4	1.0
TAL1_disc1	1.0	1.0	3.0	1.0
CPHX_1	1.0	1.0	2.7	1.0
JDP2_2	1.0	1.0	2.4	1.0
NFE2L2_3	1.0	1.0	2.0	1.0
HNF4_known9	4.6	0.3	1.0	1.0
NR2F6_2	3.7	1.0	1.0	1.0
HNF1B_4	3.5	0.6	1.0	1.0
RXRA_known10	3.4	0.9	1.0	1.0
PPARA_4	3.1	1.0	1.0	1.0
HNF1A_4	2.8	1.0	1.0	1.0
HNF1_4	2.6	1.0	1.0	1.0
TCF7L2_disc1	2.4	1.0	1.0	1.0
AP1_known4	2.3	1.0	1.6	1.0
SMARC_disc1	2.2	1.0	1.1	1.0
CEBPB_known1	2.1	1.0	1.0	1.0
TLX2_2	2.0	1.0	1.0	1.0
FOXJ2_4	2.0	1.0	1.0	1.0
EGR1_disc2	2.0	1.0	1.0	1.0
RFX2_3	1.0	3.0	1.0	1.0
RFX5_known9	1.0	2.9	1.0	1.0

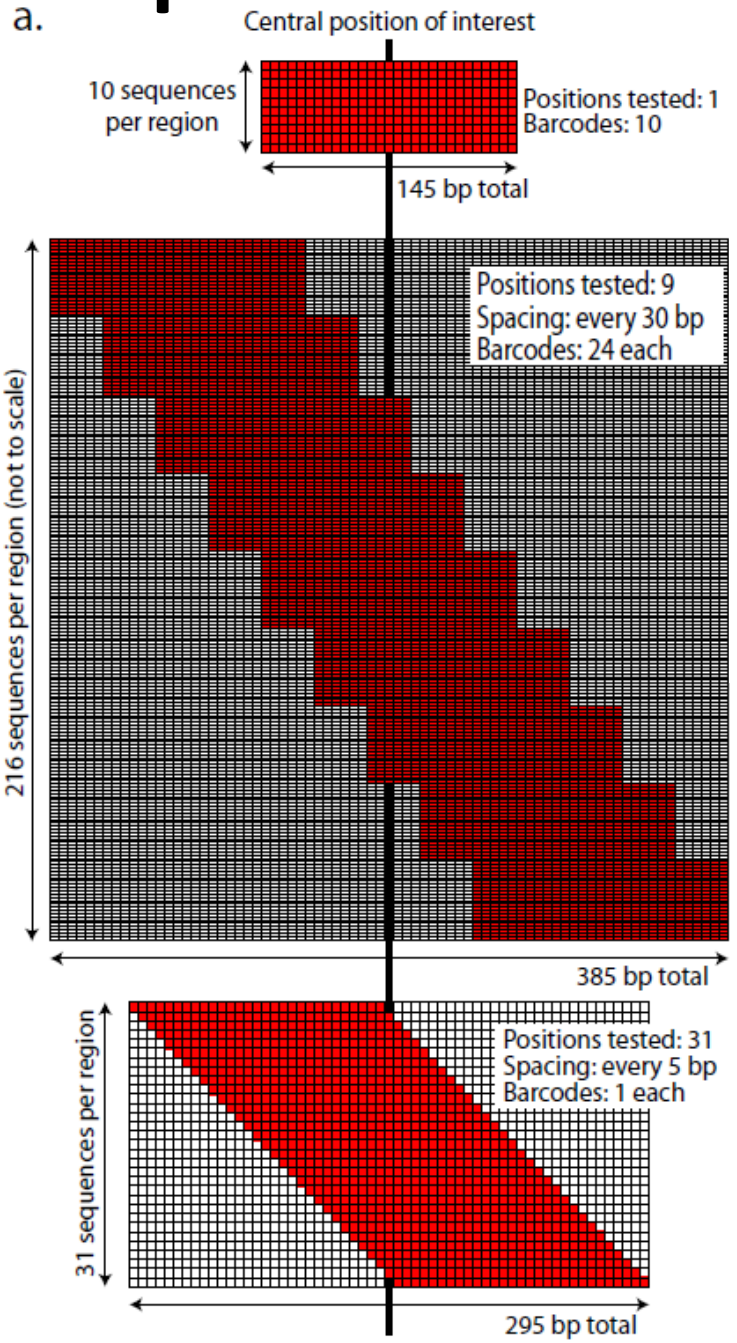


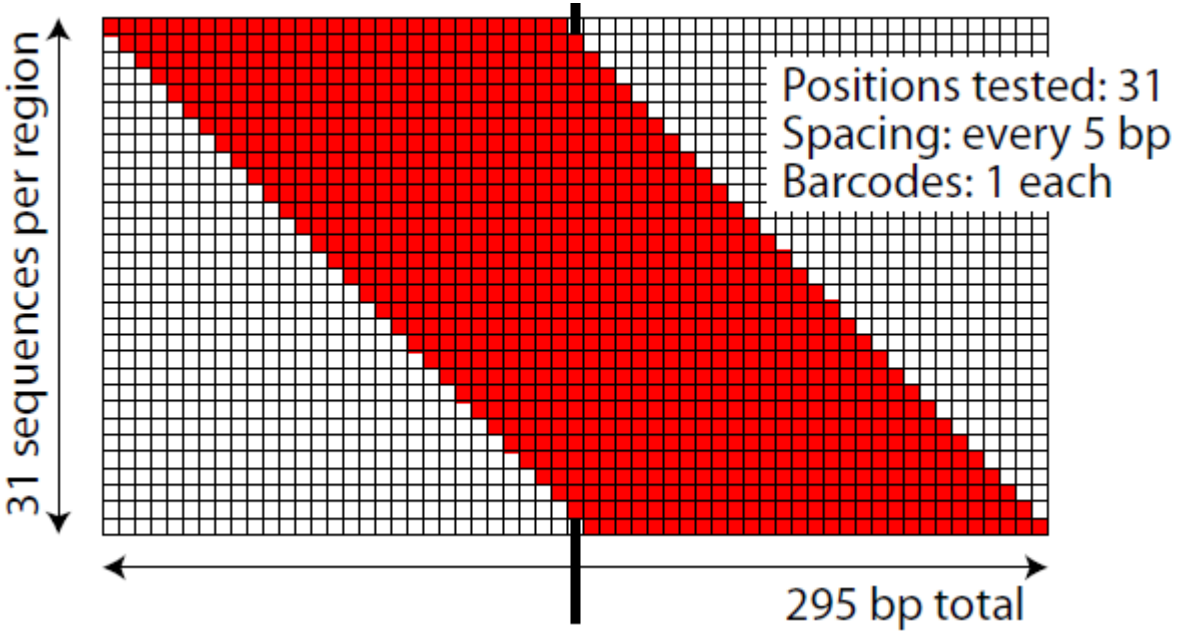
- Increased resolution allows testing of only 30-bp intervals
- *De novo* discovered motifs match known motifs
- Discovery distinguishes activating vs. repressive factors

Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions Z_{ij} from motif matrix
   - M step: Find max-likelihood motif from all positions Z_{ij}
3. Gibbs Sampling: Sample from joint (M,Z_{ij}) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

Experimental design for high-resolution tiling

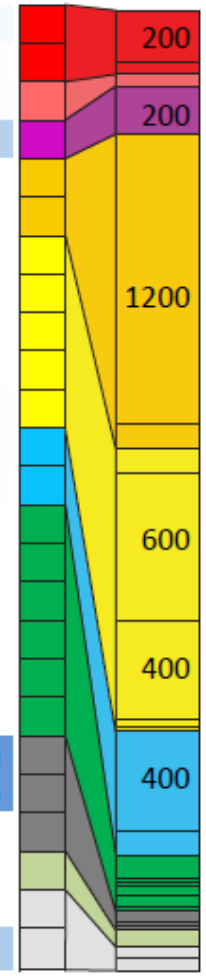




- ***Tile 295-bp regions (vs. 385)***
- ***Tile @ 5-bp offsets (vs @30)***
- ***Center on DNase peaks (vs dips)***
- ***Single barcode / offset (vs 24)***
- ***244K spot array (vs 54K)***
- ***25 chromatin states (vs. Enh)***
- ***Both minP and SV40 promoters***

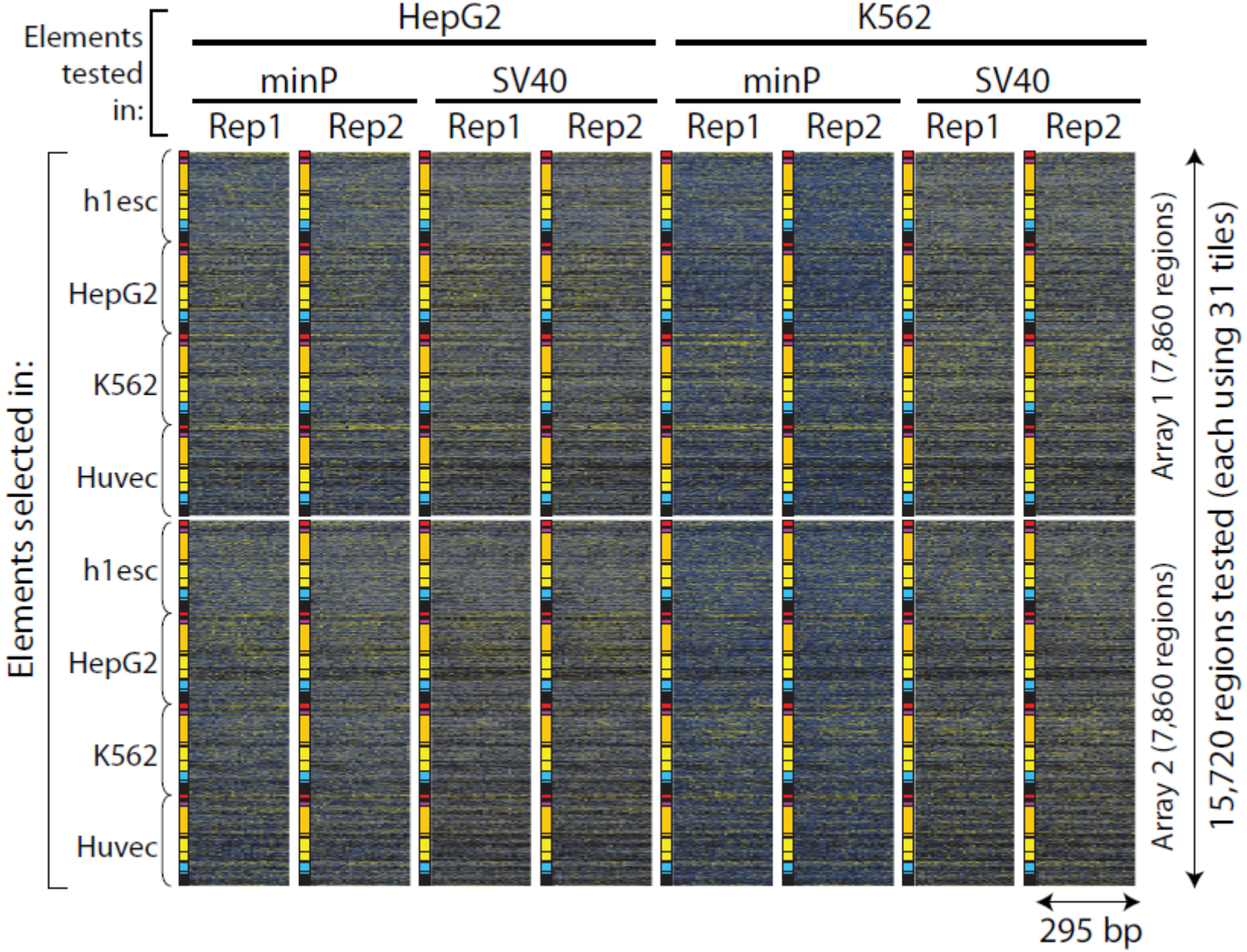
Chromatin state vs. reporter activity of DNase elmts

State	H3K36me3	H4K20me1	Control	FAIRE	POL2	H3K27ac	H3K9ac	H3K4me3	H3K4me2	H3K4me1	DNase UW	DNase DUKE	CTCF	H3K27me3
1 Tss	2	6	3	33	98	96	99	100	96	43	86	92	36	5
2 TssF	21	24	.7	4	55	72	86	98	99	73	24	15	4	2
3 PromF	3	6	.2	.6	7	2	8	32	89	44	5	4	2	2
4 PromP	.1	7	.5	8	38	2	37	88	97	61	51	87	32	42
5 Enh	11	7	4	49	69	98	73	34	97	98	89	67	17	.1
6 EnhF	5	2	.6	4	11	94	29	1	54	93	20	8	2	.1
7 EnhWF	.9	2	.2	.8	1	.6	1	0	7	88	4	4	1	.4
8 EnhW	.2	8	.7	17	19	11	6	.6	62	94	63	60	17	.7
9 DNaseU	.4	1	.1	3	3	2	.5	.1	3	9	95	14	.3	.3
10 DNaseD	1	3	.2	.3	2	.5	.7	.1	.3	3	2	69	.9	.3
11 FaireW	1	2	8	22	.8	10	6	2	.2	.3	.5	1	.6	2
12 CtcfO	6	9	1	35	19	2	1	.3	10	17	71	87	100	2
13 Ctcf	.3	2	.1	.5	.3	.1	.2	0	.7	1	6	16	99	2
14 Gen5'	70	36	.7	3	52	26	9	1	23	84	9	12	4	0
15 Elon	93	4	.3	.7	1	.3	.2	0	.1	.3	.4	1	1	0
16 ElonW		.4		.3	.8	.1	0	0		.1	.1	.2	0	0
17 Gen3'	93	18	.4	1	83	1	.6	.1	.1	.7	2	7	3	0
18 Pol2	.7	6	.1	.7	83	.9	.6	0	.1	2	3	7	1	.3
19 H4K20	3	86	.3	1	2	.2	.3	0	.2	2	.4	1	.5	.6
20 ReprD	.5	15	.5	1	2	.5	.6	1	11	21	4	48	12	90
21 Repr	0	2	.2	.1	.1	.1	0	0	0	0	0	1	.1	90
22 ReprW	0	.2	0	.1	0	0	0	0	0	0	0	.4	0	0
23 Low	.1	0	0	.2	.1	0	0	0	0	0	0	.2	0	0
24 Quies	0	0	0	0	0	0	0	0	0	0	0	0	0	0
25 Art	77	73	41	70	59	35	38	25	32	52	41	63	64	47



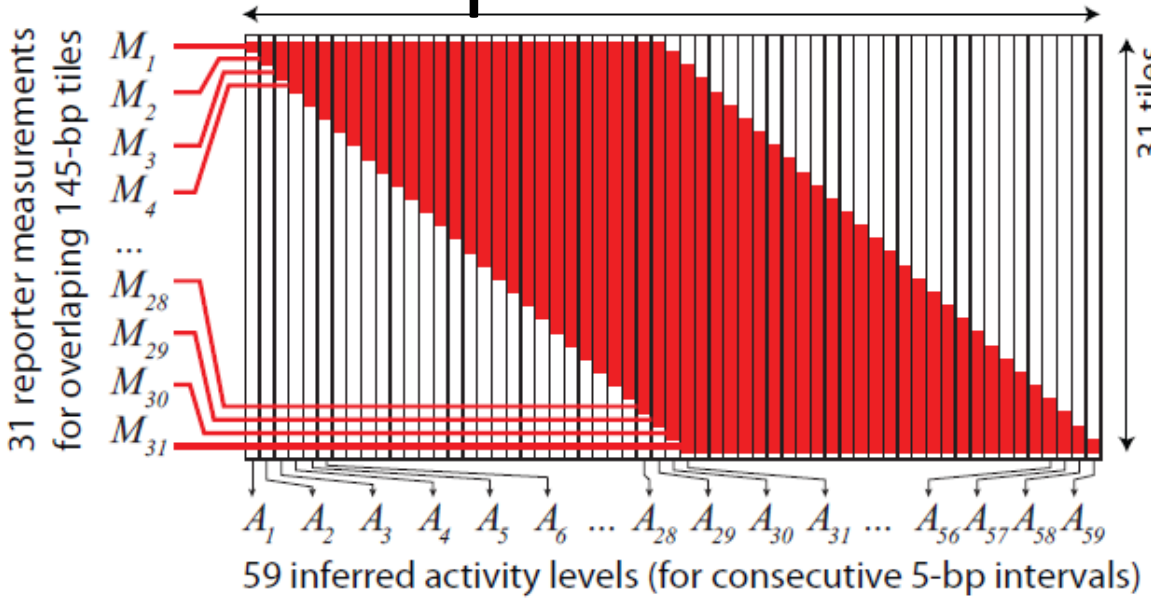
Select 15,720 DNase elements across all 25 chromatin states

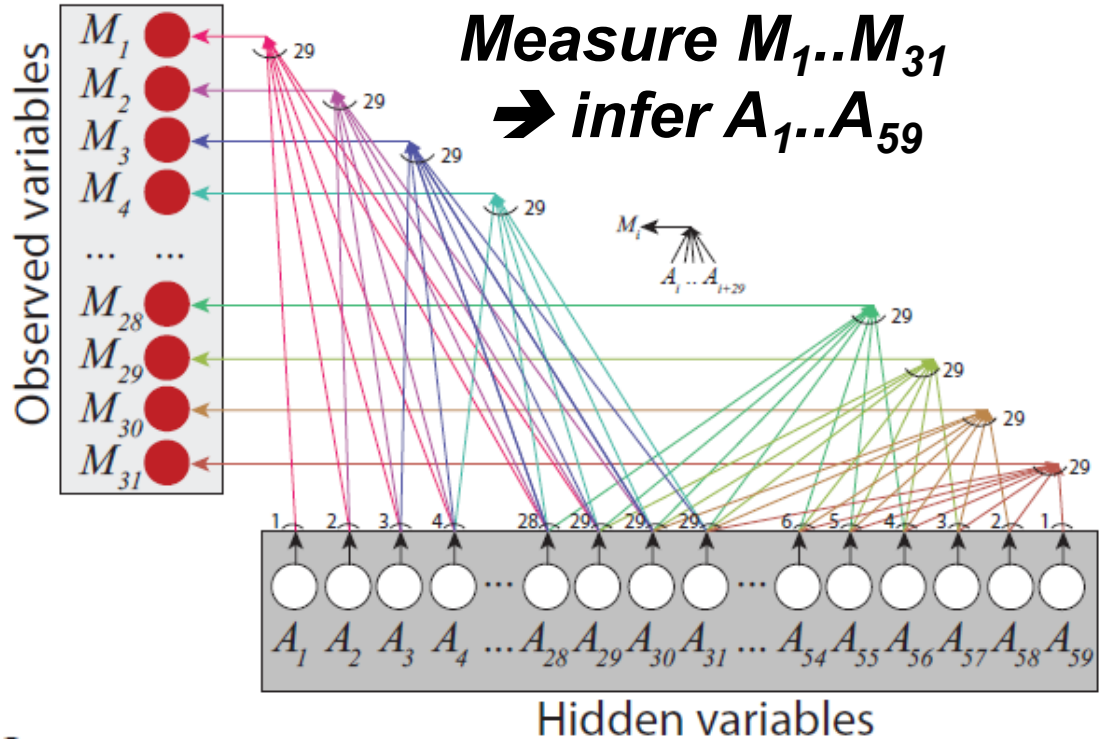
Regions selected in 4 cell types, tiled in HepG2,K562



15,720 regions x 31 offsets x 2 promoters x 2 reps x 2 cell lines

Computational inference model





Measure $M_1..M_{31}$ ➔ infer $A_1..A_{59}$

Joint samples from multivariate normal

$$\begin{bmatrix} A_{r,t} \\ M_{r,t} \end{bmatrix} \sim N(\mu_{x_{r,t}}, \Sigma_{x_{r,t}})$$

True activity normally distributed

$$A_{r,t,k} \sim N(\mu_{a_r}, \sigma_a^2) \quad k = 1, \ldots, K$$

Observation is mean of true signals

$$\mu_{a_r} = \frac{1}{|\widehat{M_r}|} \sum_{m \in \widehat{M_r}} m$$

Measurements normally distributed

$$M_{r,t,j} \sim N(\mu_{m_{r,t,j}}, \sigma_{m_r}^2)$$

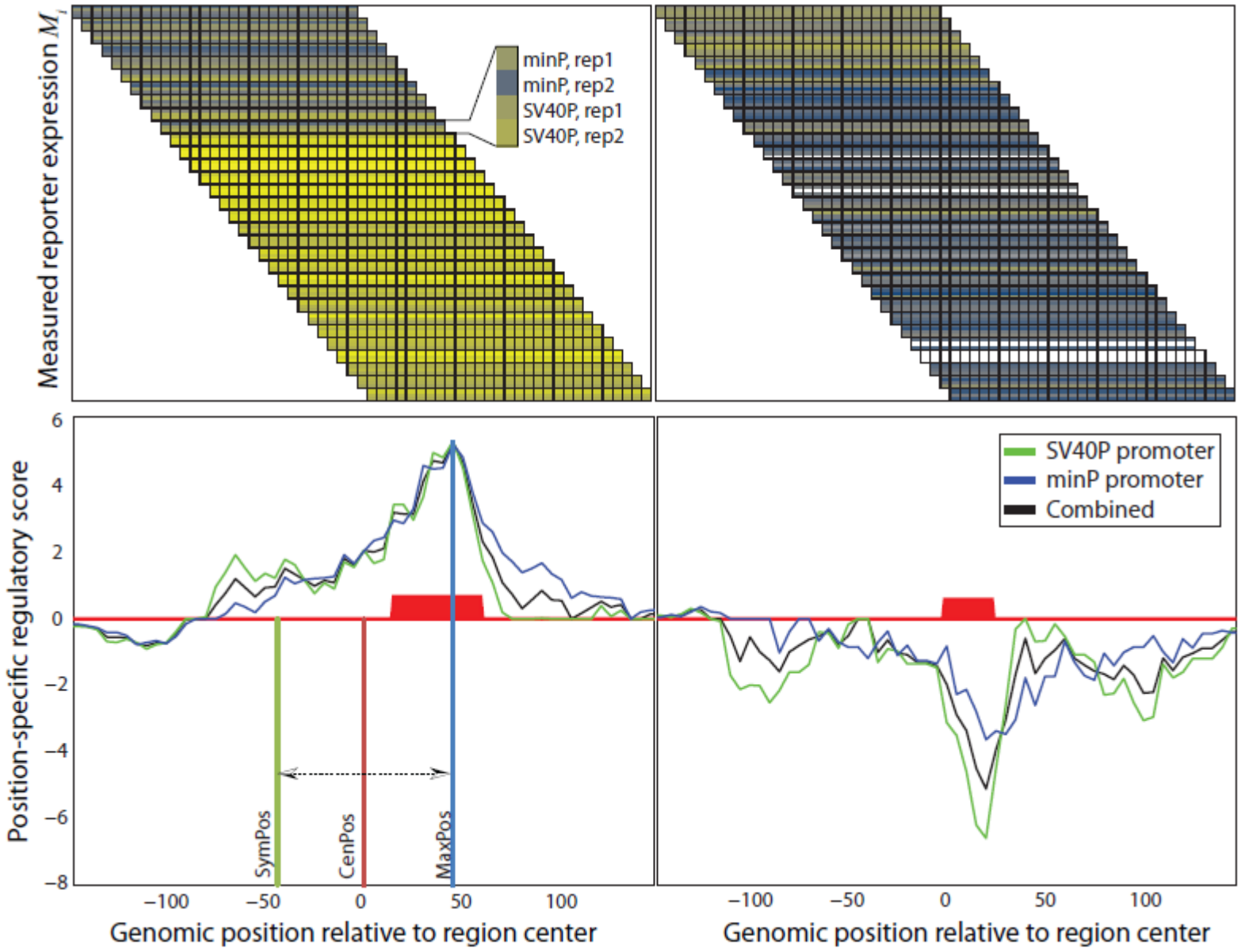
Mean set to empirical mean

$$\mu_{m_{r,t,j}} = \frac{1}{N} \sum_{i=0}^{N-1} A_{r,t,j+i}$$

Variance set to empirical variance

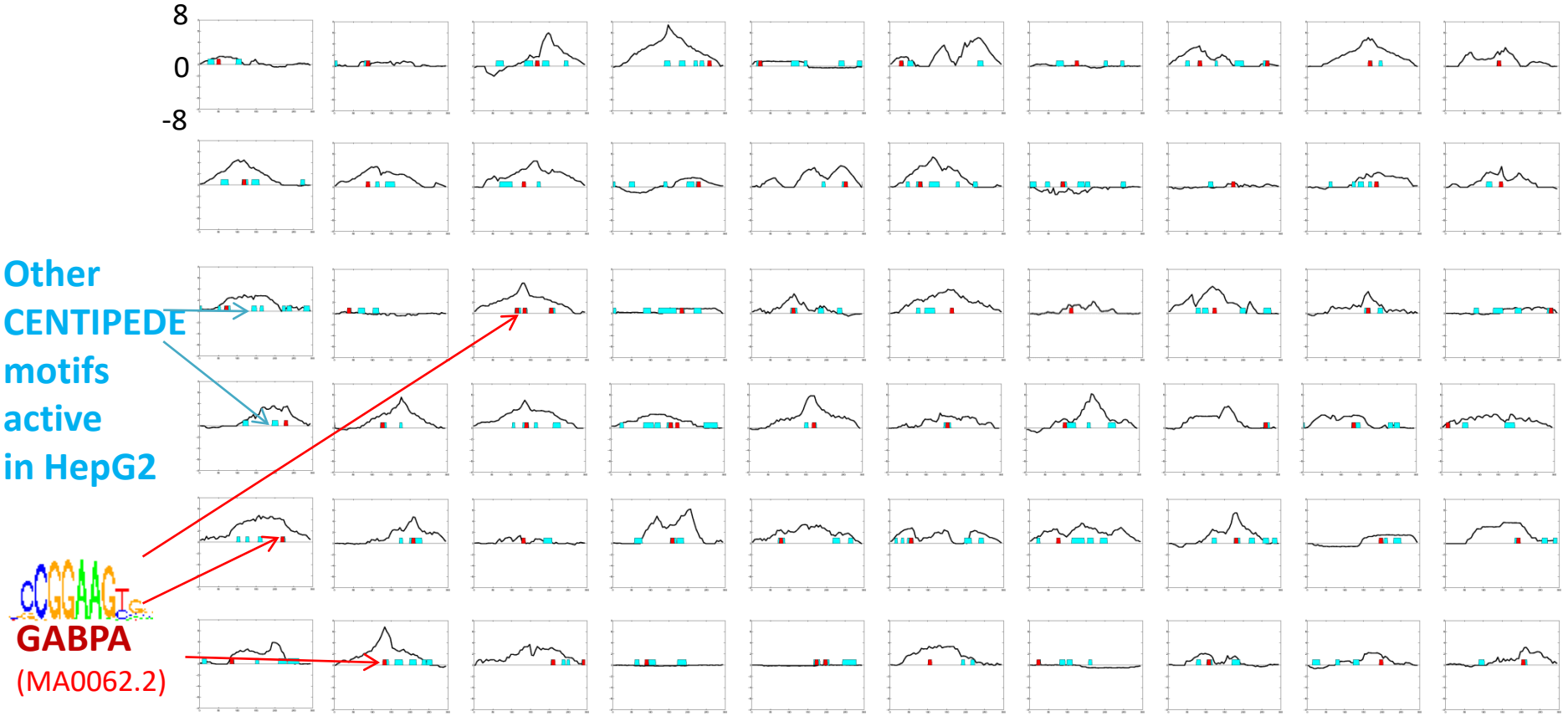
$$\sigma_{m_r}^2 = \frac{1}{|\widehat{M_r}|} \sum_{m \in \widehat{M_r}} (m - \frac{1}{|\widehat{M_r}|} \sum_{m \in \widehat{M_r}} m)^2$$

Examples of tiling data deconvolution



Detect activating/repressive elements at high resolution

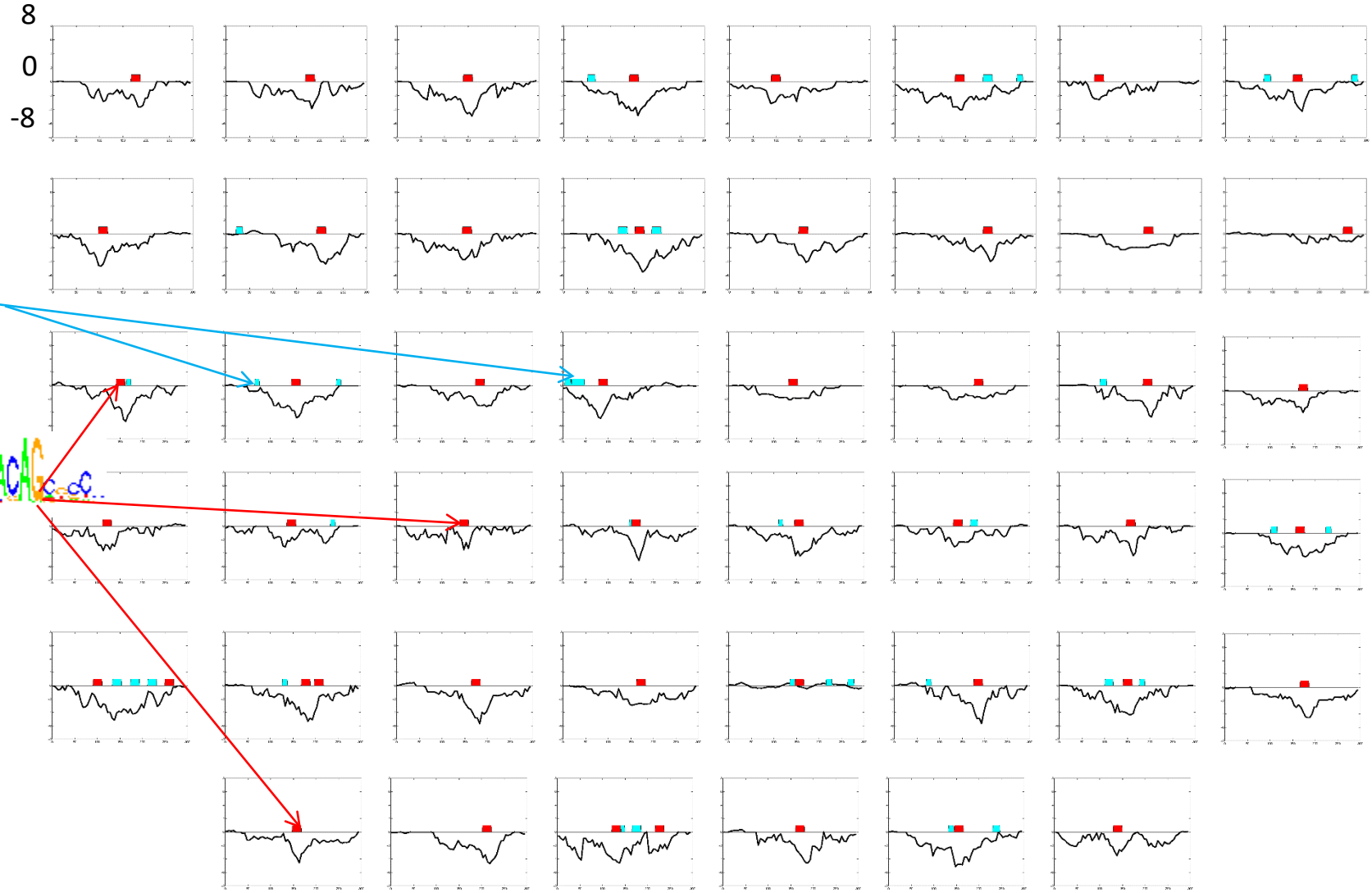
Deconvolved regulatory signal vs. activator motif



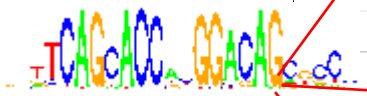
60 sites containing GABPA HepG2 motifs predicted by CENTIPEDE

CENTIPEDE predictions (Pique-Regi, et al, 2011)

Deconvolved regulatory signal vs. repressor motif



Other CENTIPEDE motifs active in HepG2

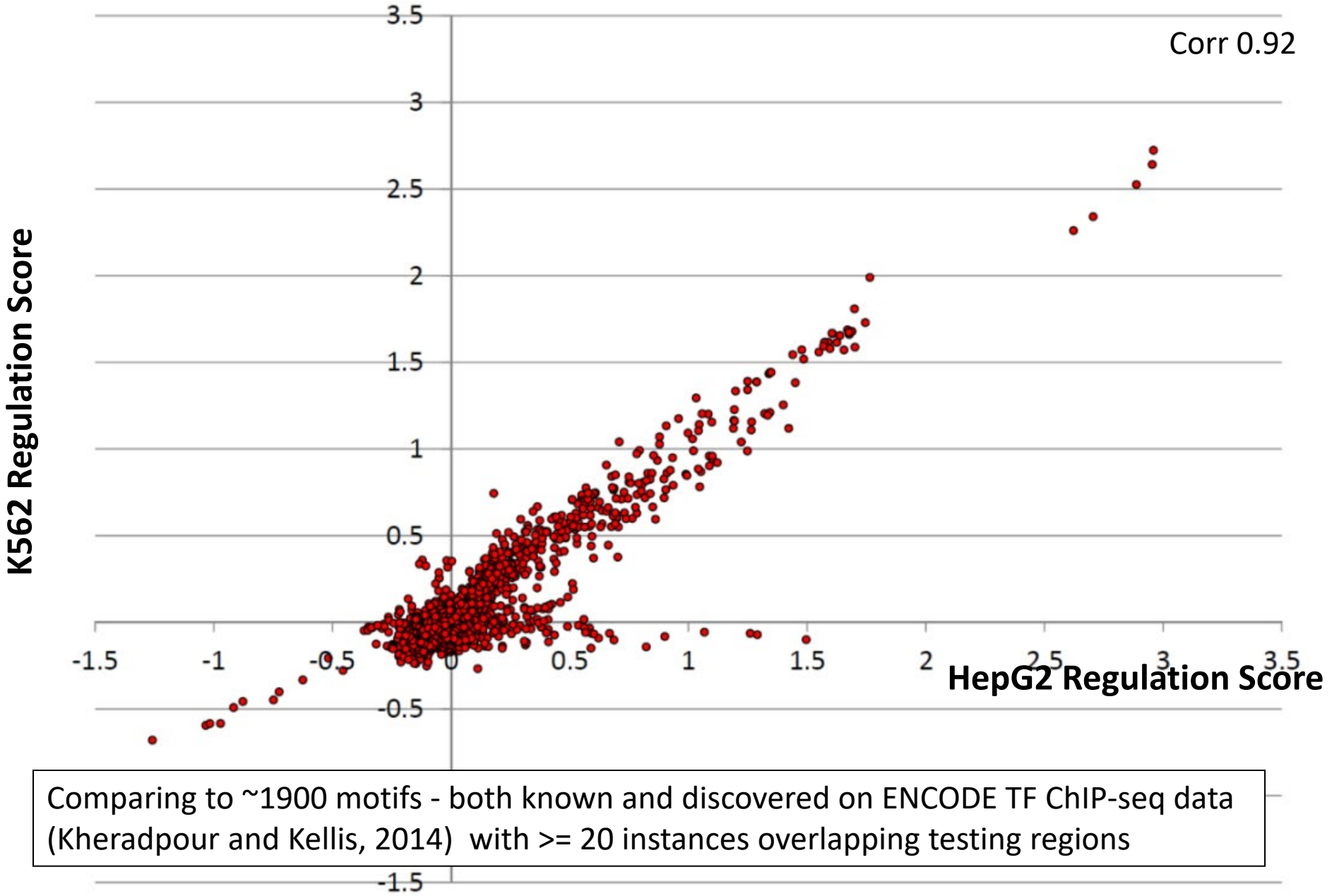


NRSF
(M00256)

46 sites containing NRSF HepG2 motifs predicted by CENTIPEDE

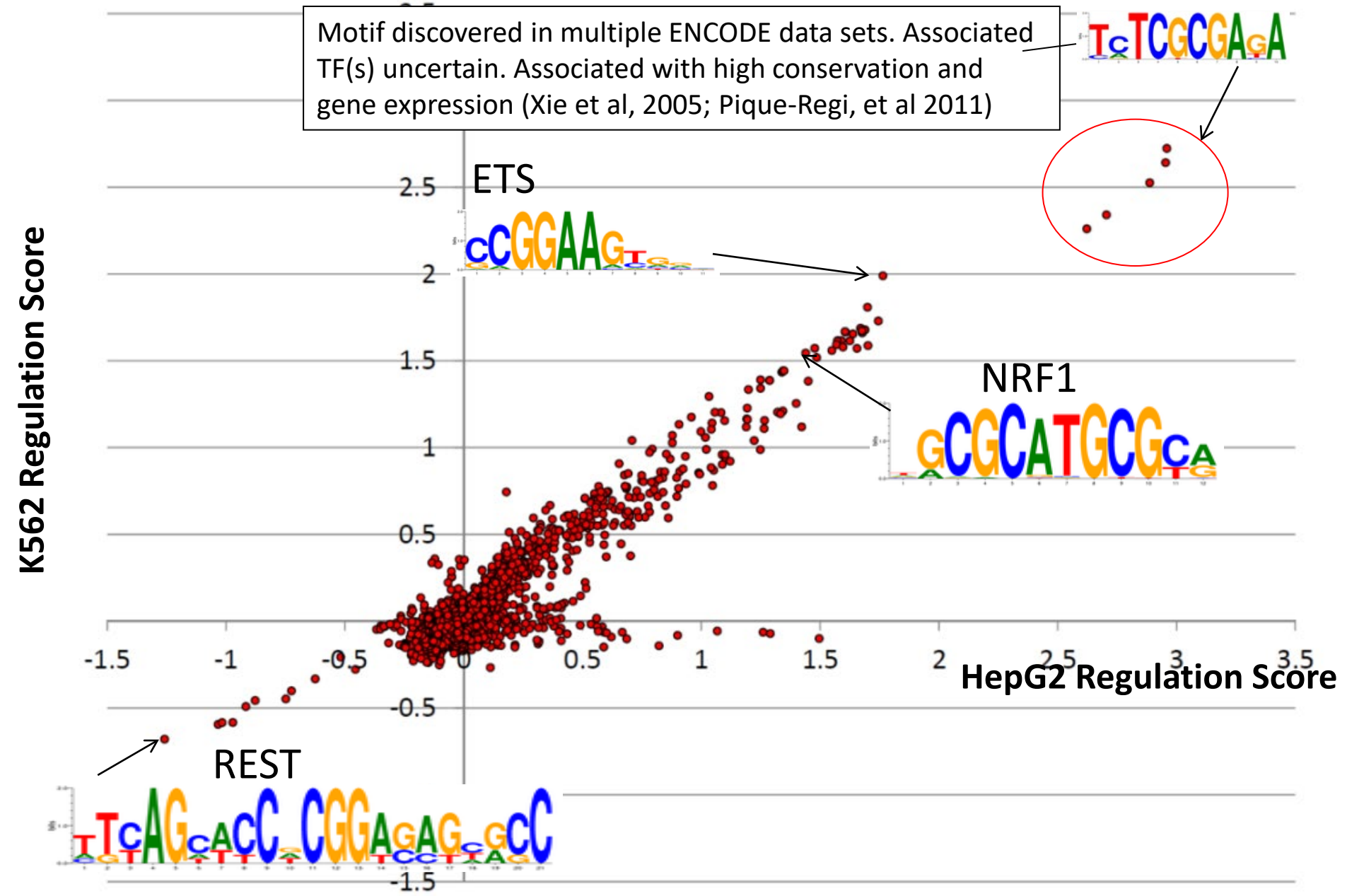
CENTIPEDE predictions (Pique-Regi, et al, 2011)

Aggregate Motif Score Highly Correlated between K562 and HepG2

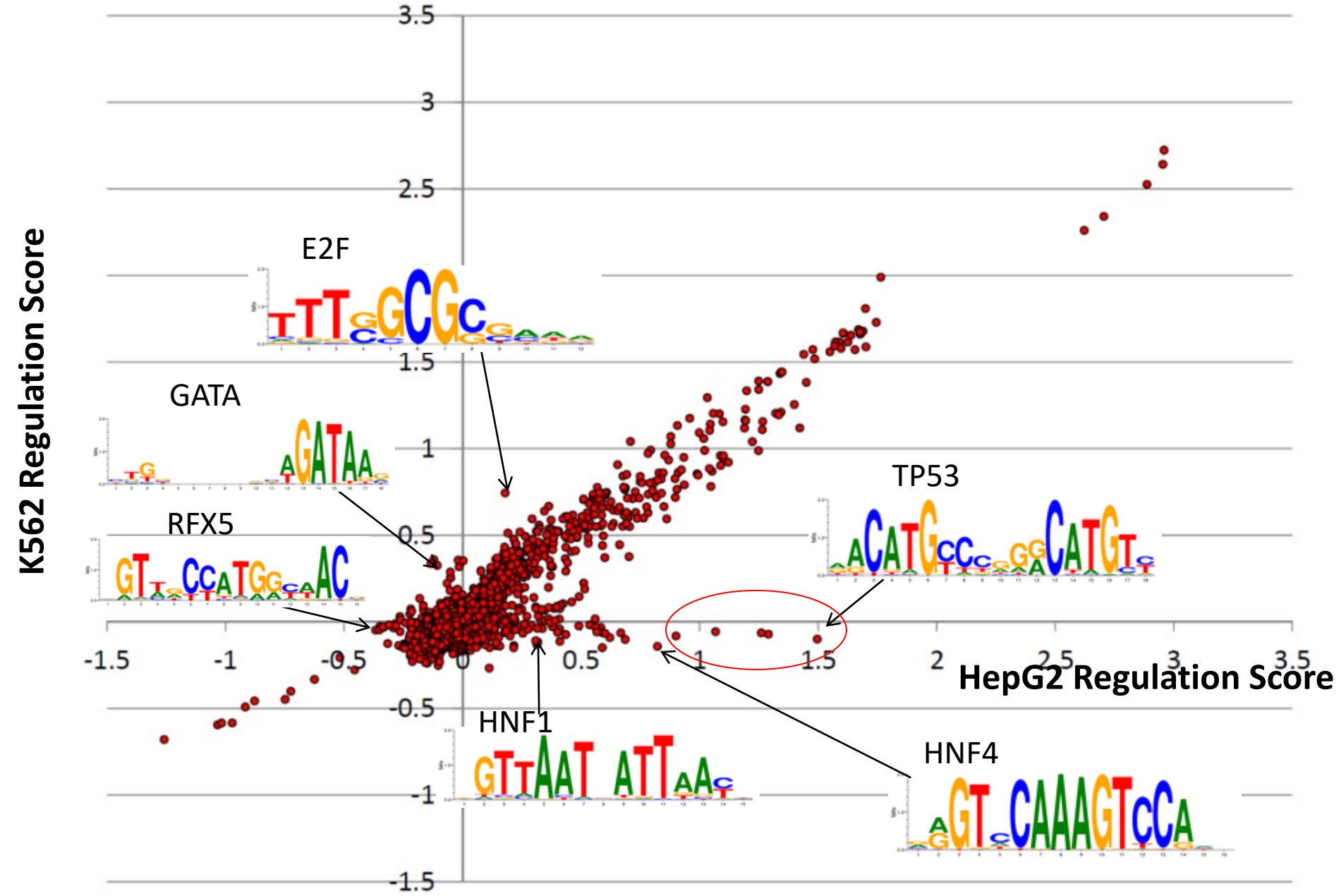


Comparing to ~1900 motifs - both known and discovered on ENCODE TF ChIP-seq data (Kheradpour and Kellis, 2014) with >= 20 instances overlapping testing regions

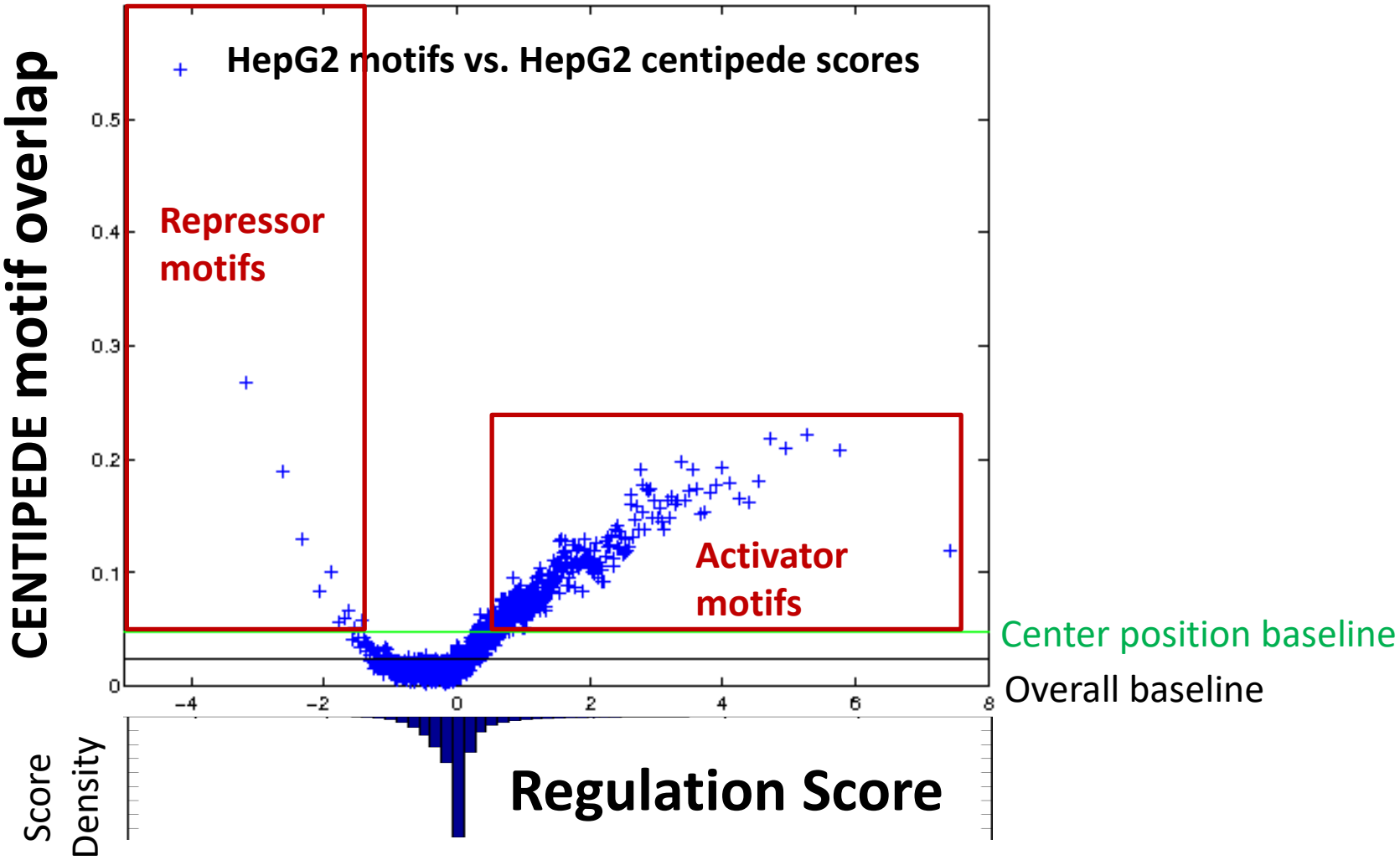
Top Activating and Repressive Motifs Revealed



Cell Type Specific Motifs Revealed

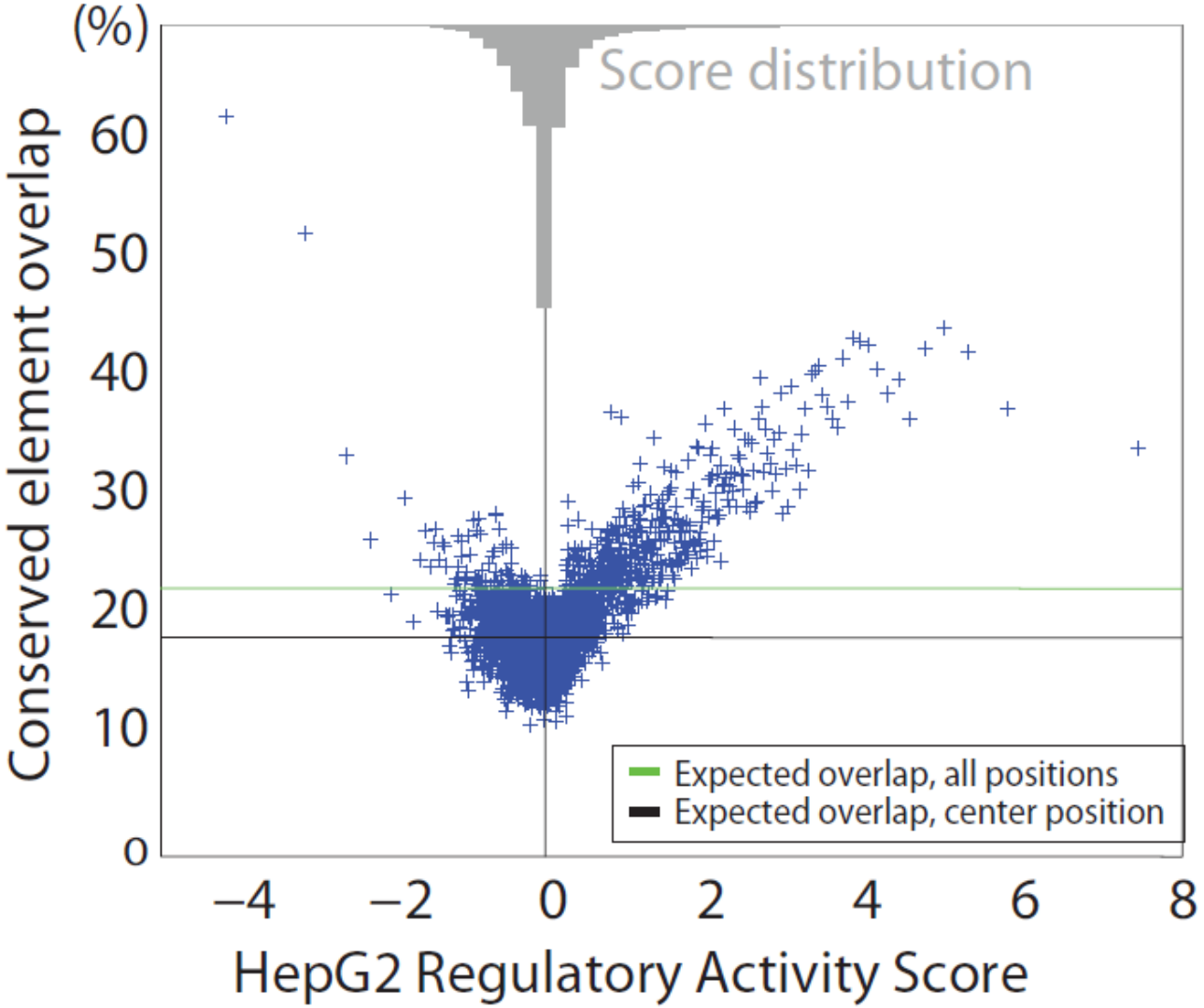


Inferred positions match regulatory motifs



Predicted Activation and Repressive Bases Strongly Enrich for Predicted Binding Sites in HepG2 + K562

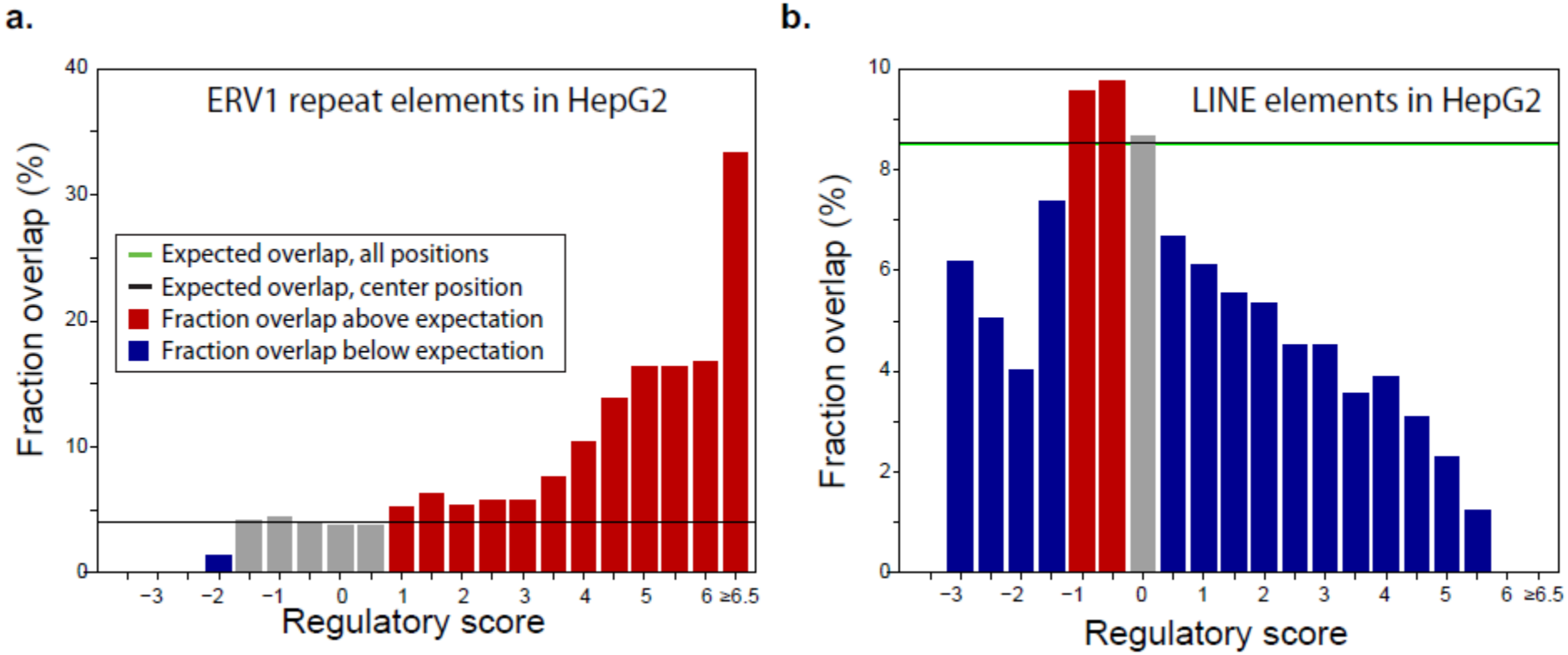
Active/repressed positions are evol. conserved



Strongest enrichment for repressive positions

Slight depletion at strongest activating positions

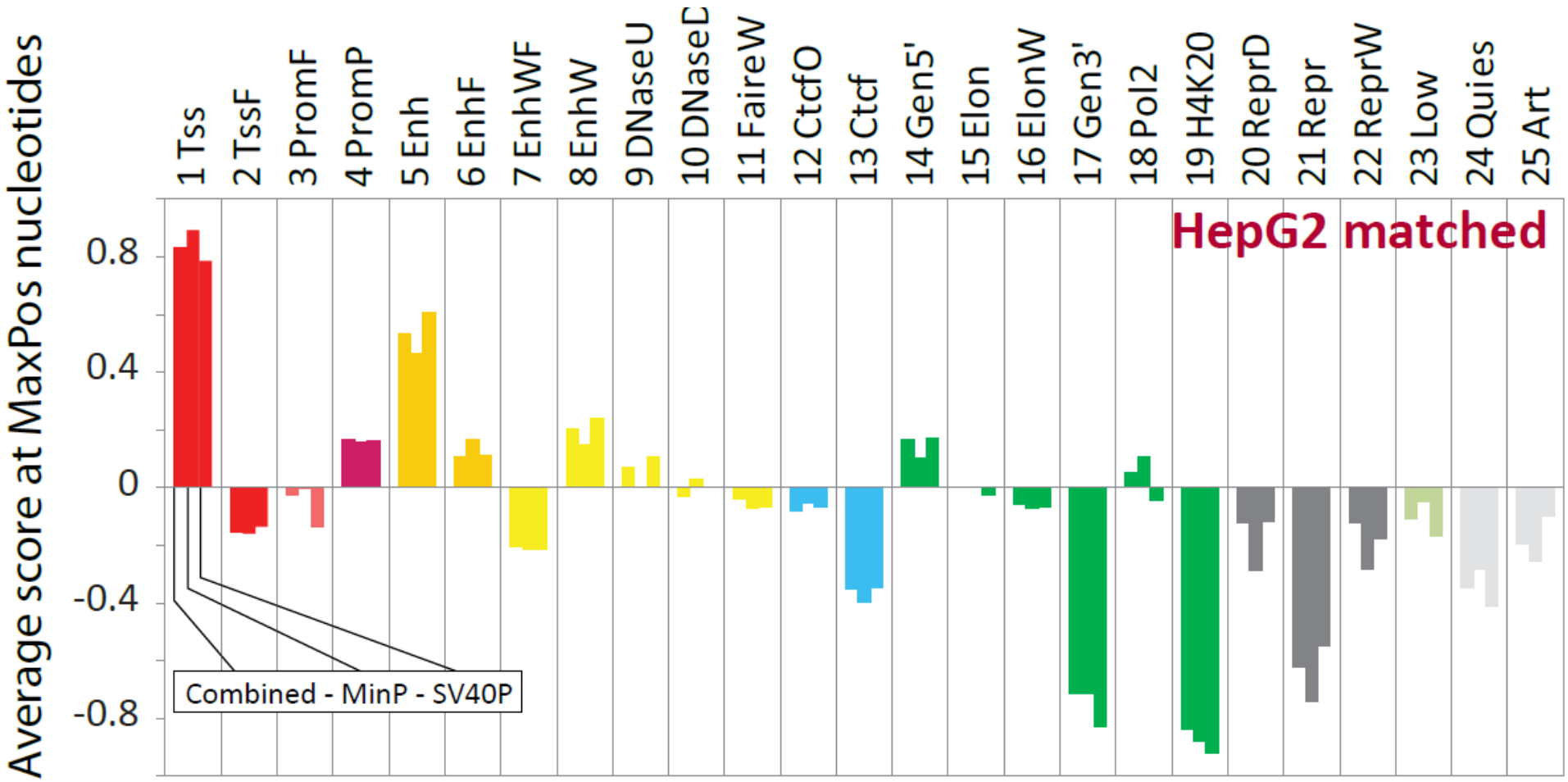
ERV1 repeat elements can drive activity



Strongest activating nucleotides match ERV1 repeats (by contrast, LINE elements strongly depleted)

Enable rapid evolution of gene-regulatory networks

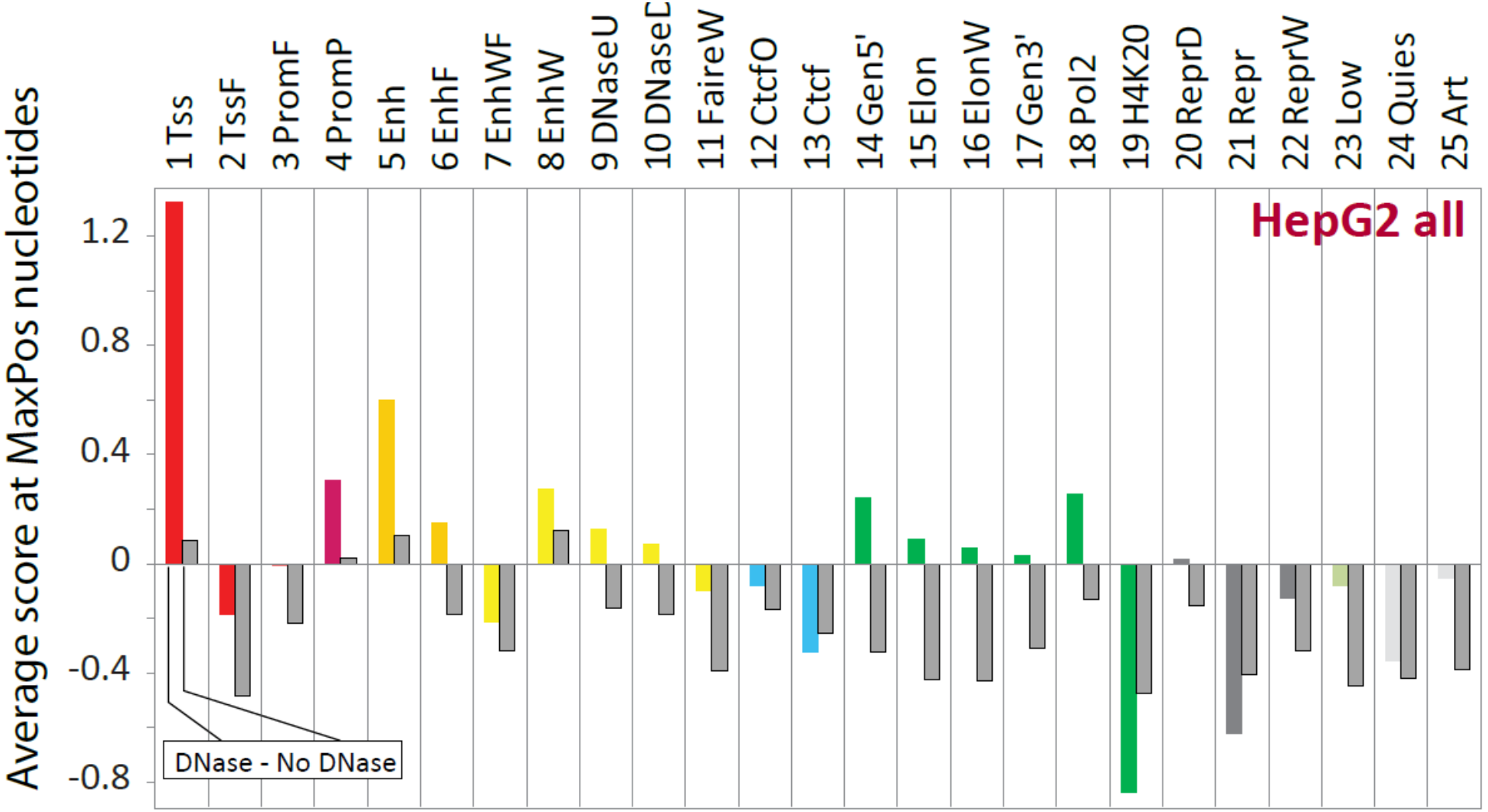
DNase elements in different chromatin states differ in their activity levels



Promoter, Enhancer regions highly activating.

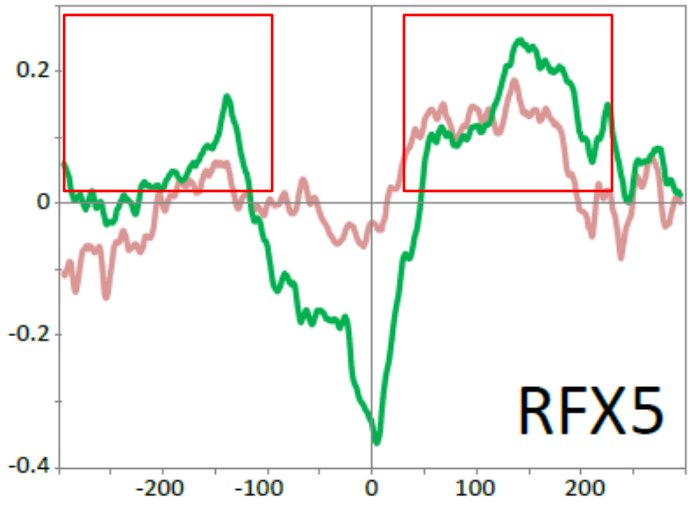
ReprPC regions highly repressive

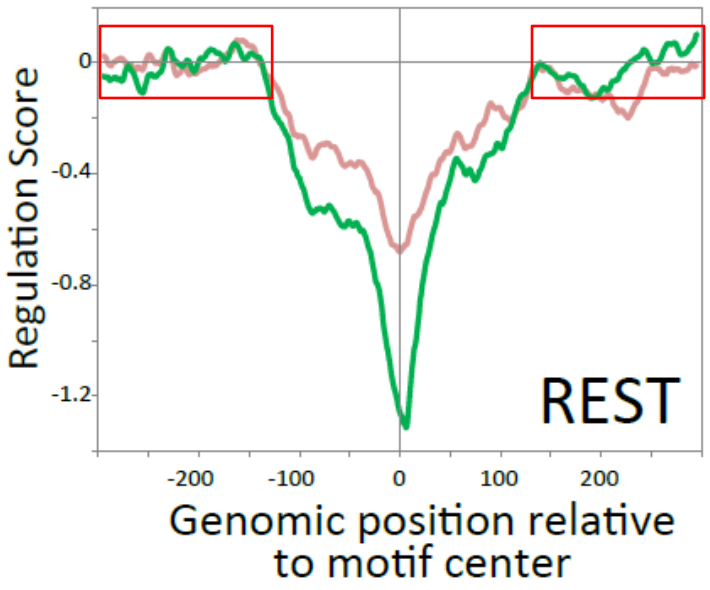
Accessible regions drive stronger activity

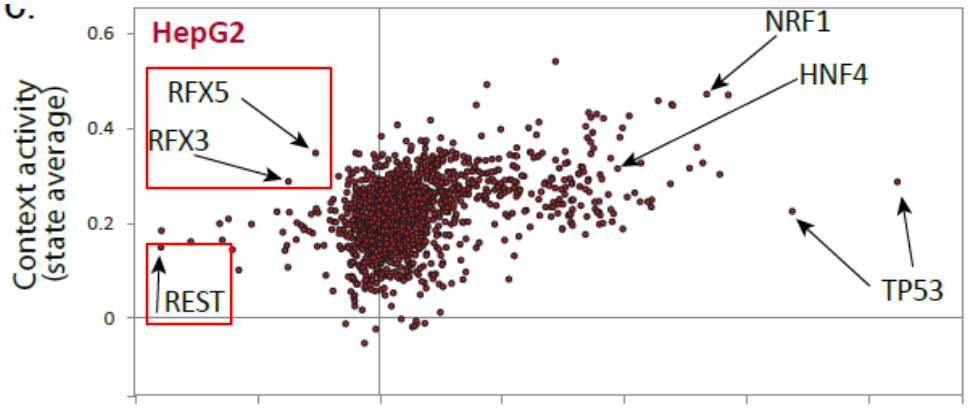


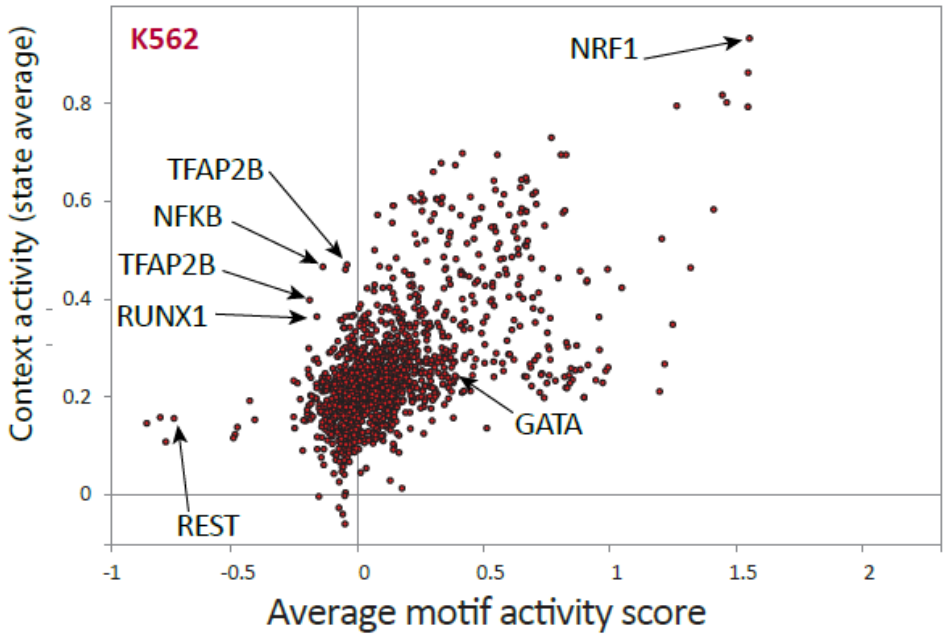
For both activating and repressive positions

Discovery of repressors that act in active regions







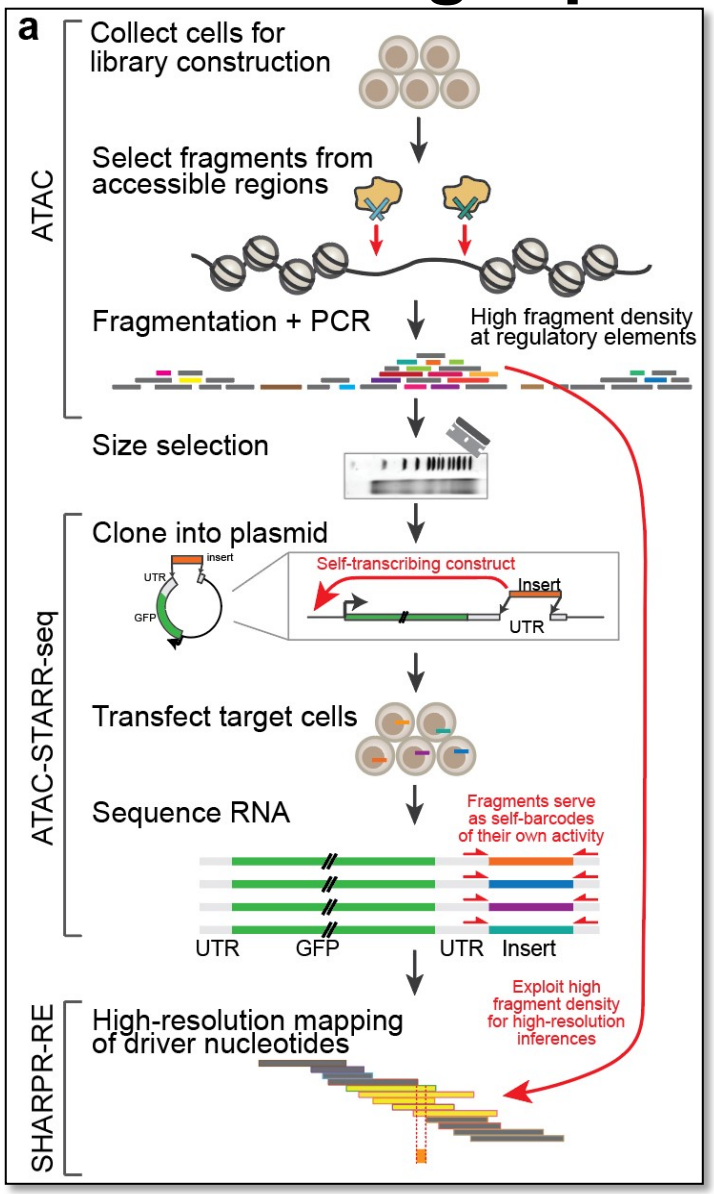


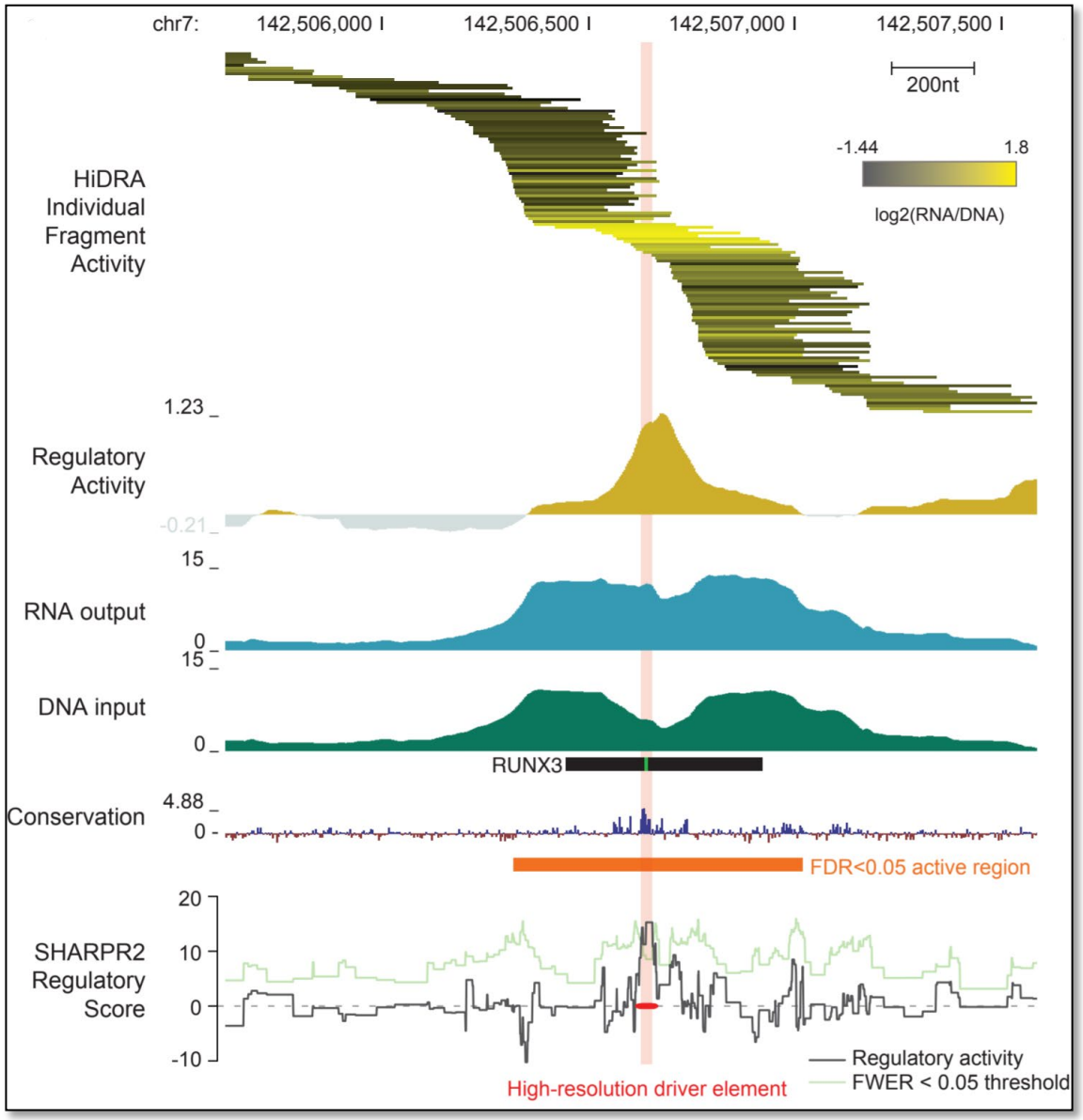
- ***REST acts as a repressor in repressive regions (as expected)***
- ***But RFX5 acts as a repressor only in active regions (modulator?)***

Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions Z_{ij} from motif matrix
   - M step: Find max-likelihood motif from all positions Z_{ij}
3. Gibbs Sampling: Sample from joint (M,Z_{ij}) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Position offset matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays

HiDRA: Longer probes + Hi-res dissection + 7M tests

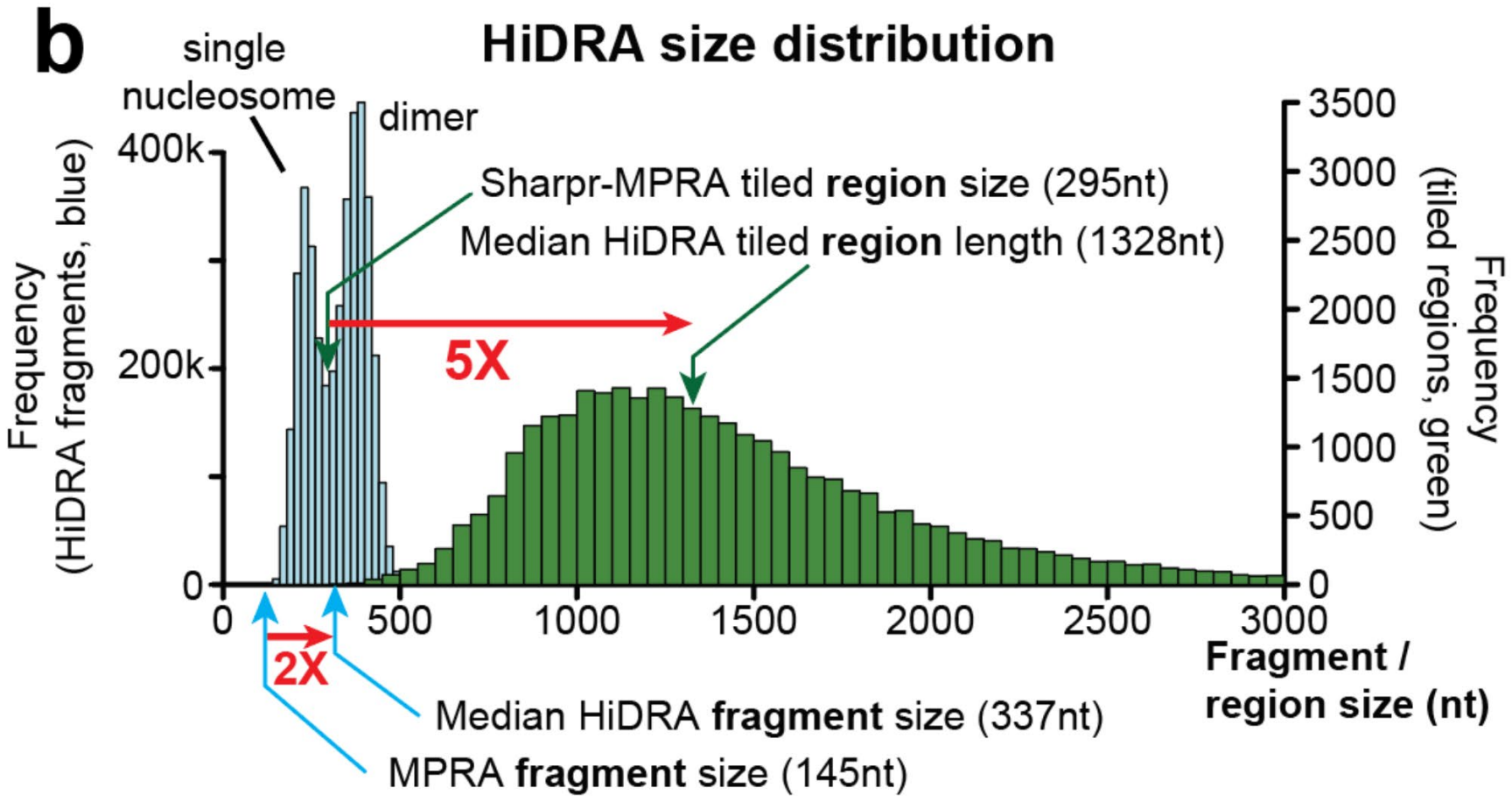




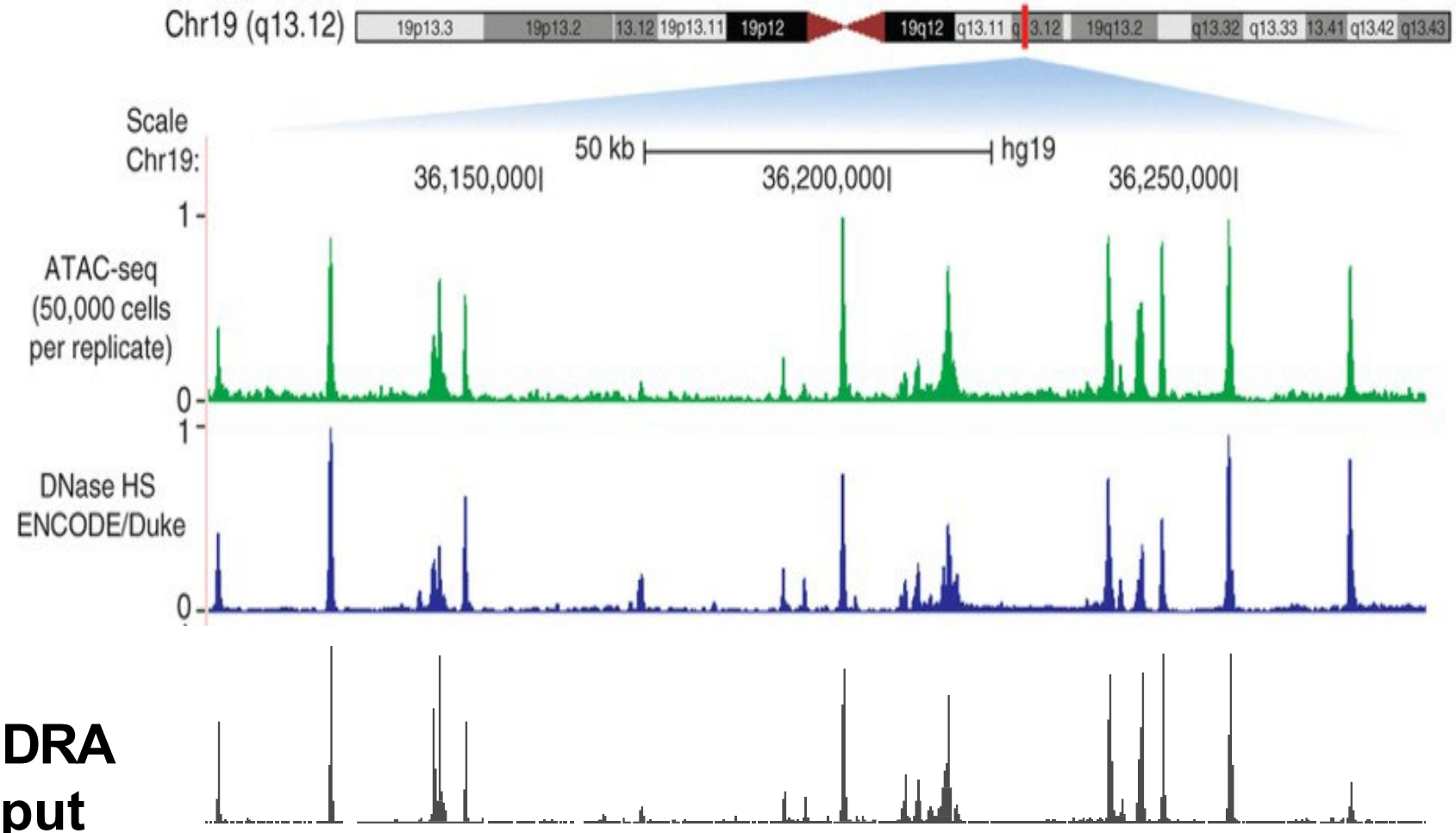
ATAC selection ➔ *No synthesis* ➔ ***7M tests***
3'UTR incorp. ➔ *Self-transcribe* ➔ ***No barcode***
Dense, random start/end ➔ ***Region tiling***

High-resolution inference of driver nucleotides
➔*Exploit differences between neighboring fragments*
➔*Driver nucleotides match motifs, evolut. conservation*

HiDRA enables testing of larger fragments



HiDRA input DNA library recapitulates DNase/ATAC-Seq

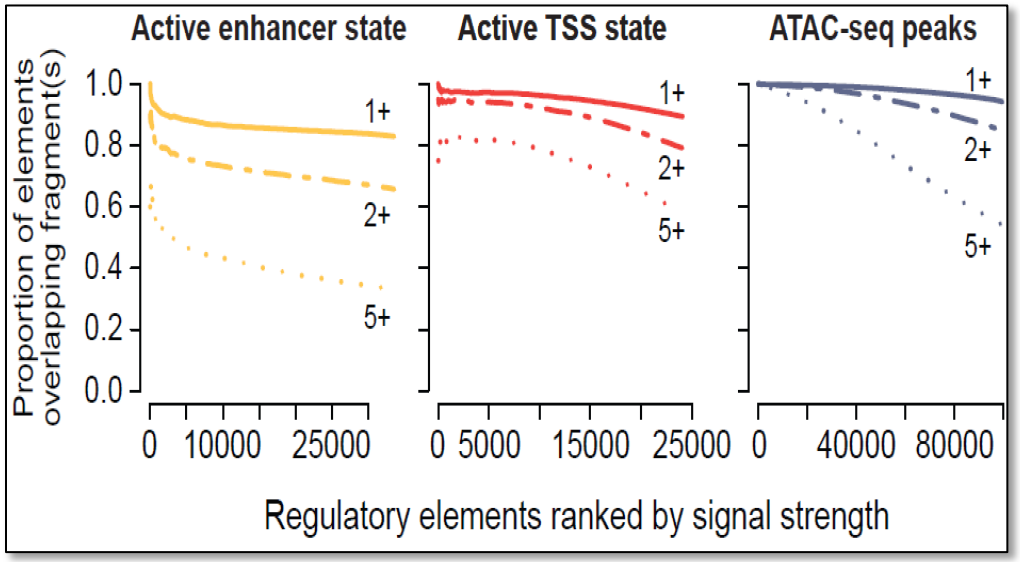


HiDRA input DNA library

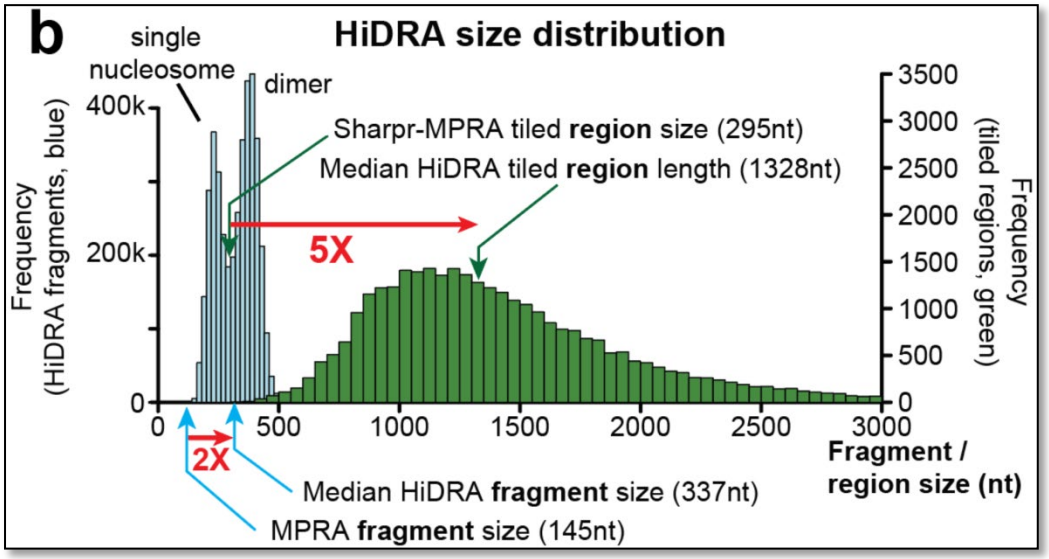
Fig 1c from Buenrostro *et al.* Nature Methods 2013

Preferential selection of putative regulatory elements

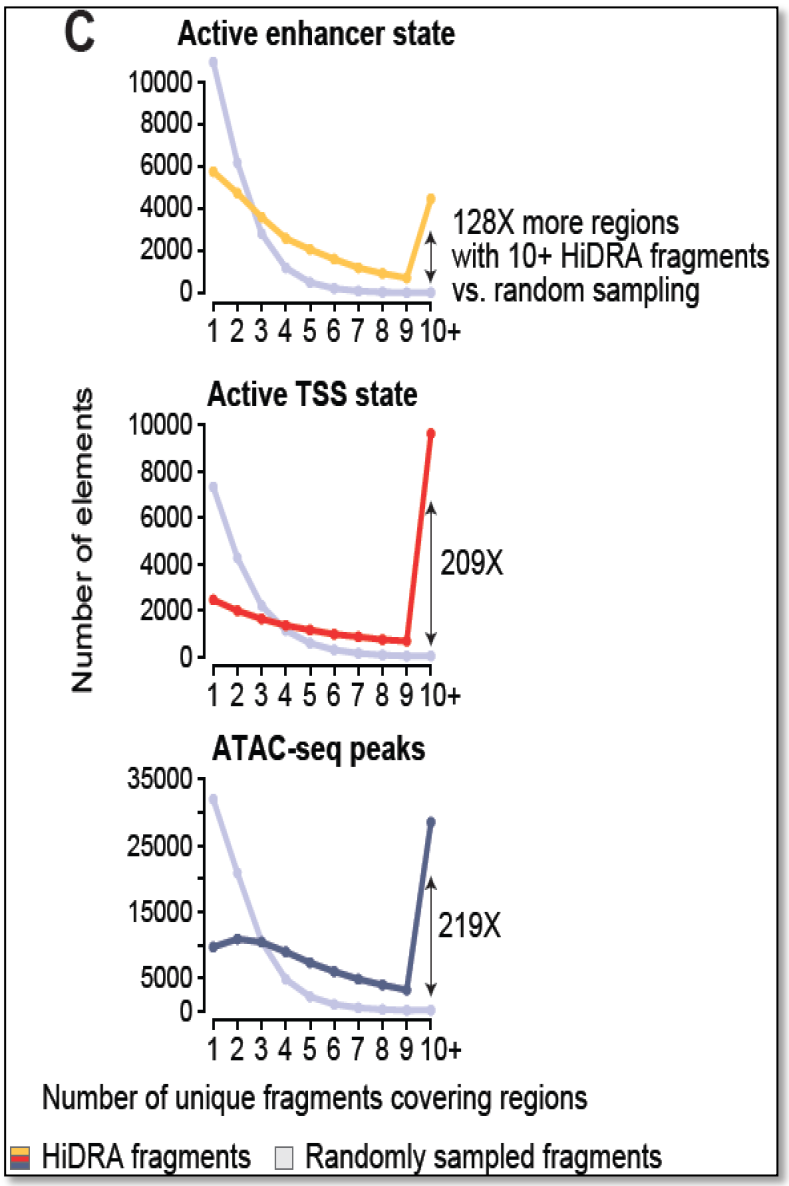
HiDRA input DNA library: long, active, densely-covered regions



HiDRA DNA library captures more active elements

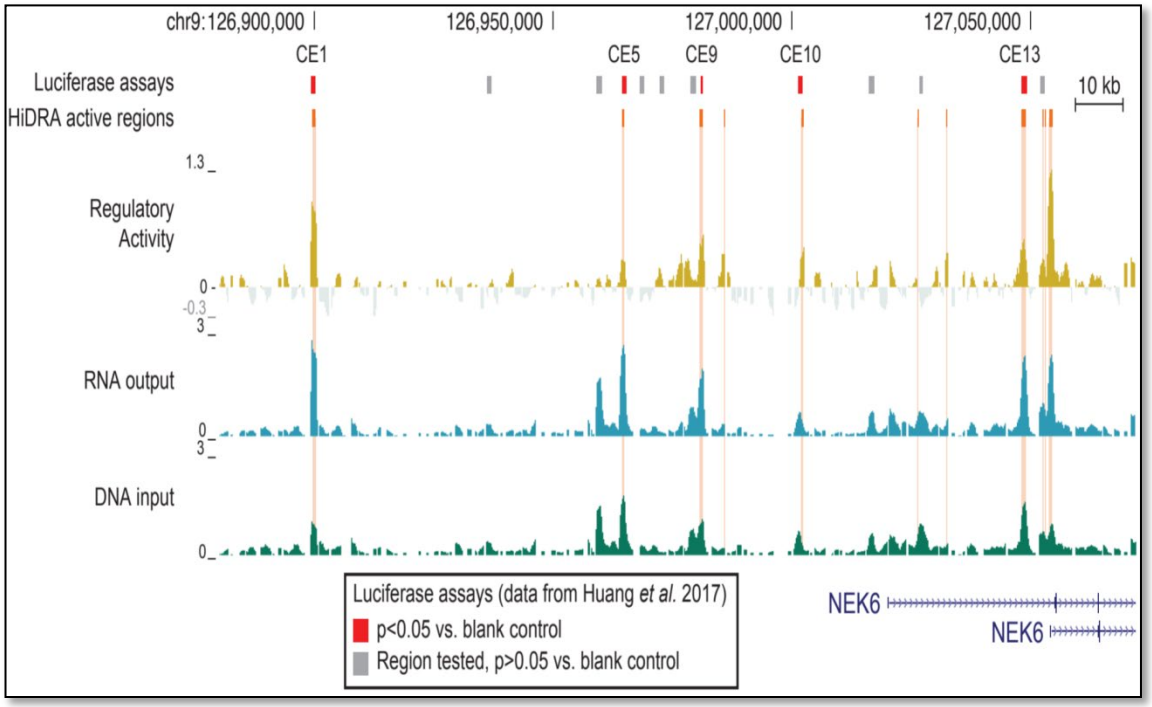


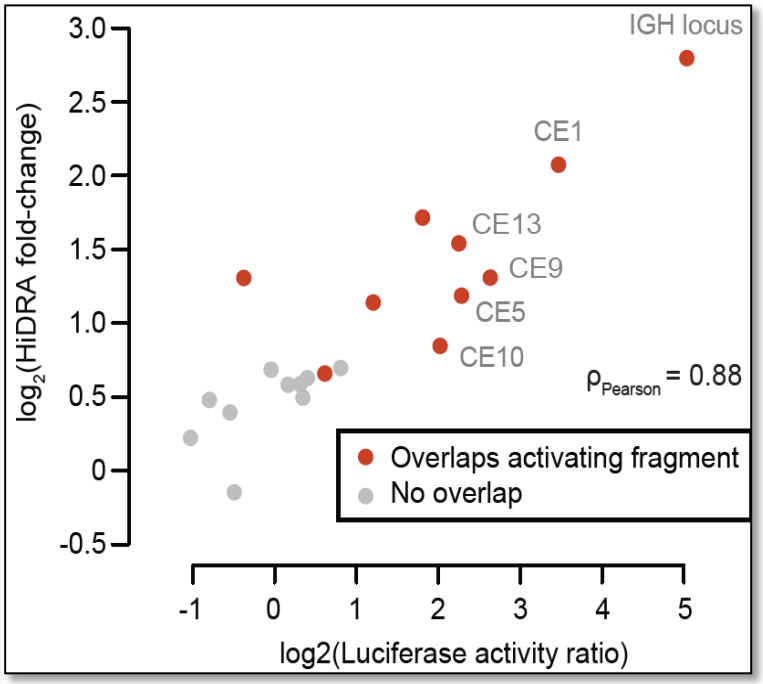
Fragments: 99% are 169-477 nt (median: 337nt)
Regions: 99% are 513-4,036 nt (median: 1,328nt)



Up to 200-fold higher coverage for putative regulatory elements

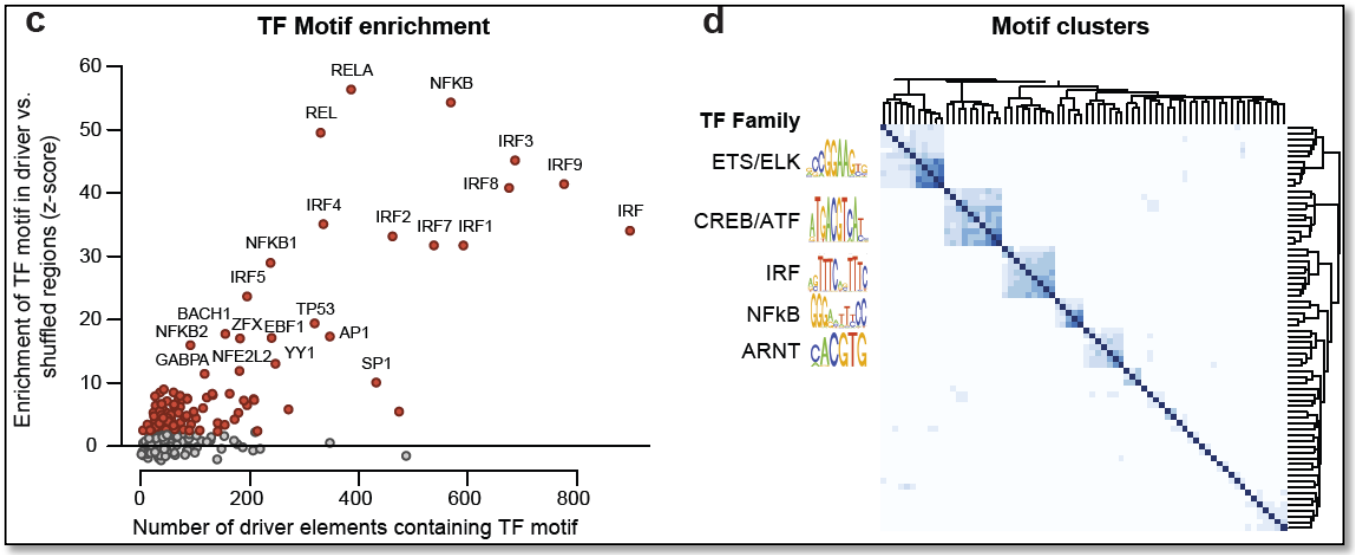
HiDRA captures known enhancers, known motifs





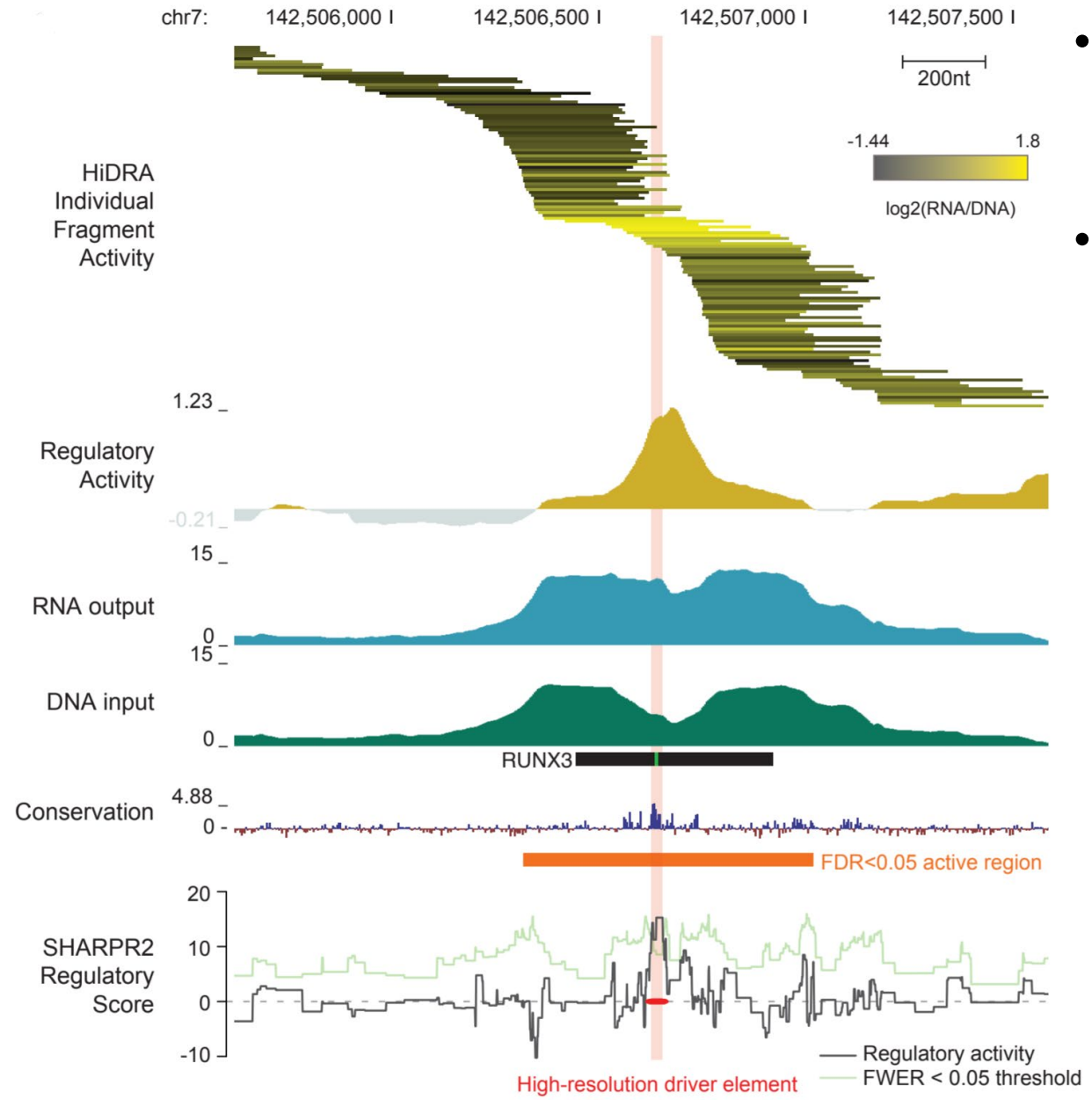
High sensitivity / high specificity vs. Luciferase assays

Quantitative read-out

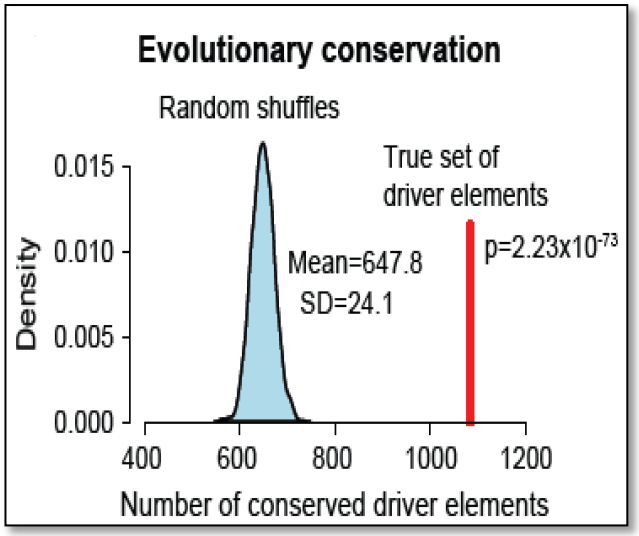


Capture known motifs

Sharpr2 algorithm infers high-resolution driver nucleotides

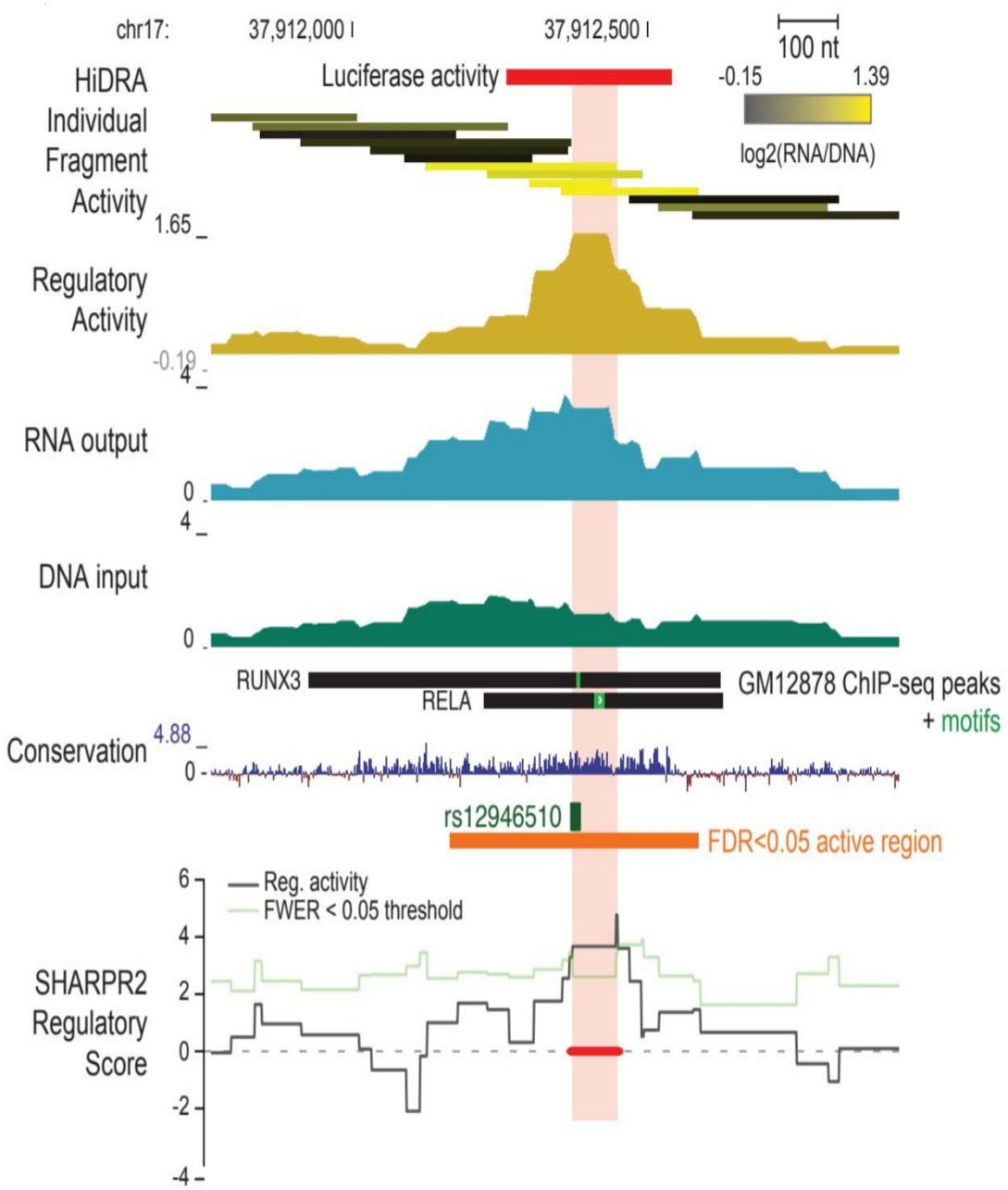


- **Exploit differences between neighboring fragments**
- **Driver nucleotides match motifs, evolutionary conservation**



- **Enrichment: $P<10^{-73}$**

HiDRA high-resolution drivers help dissect GWAS loci

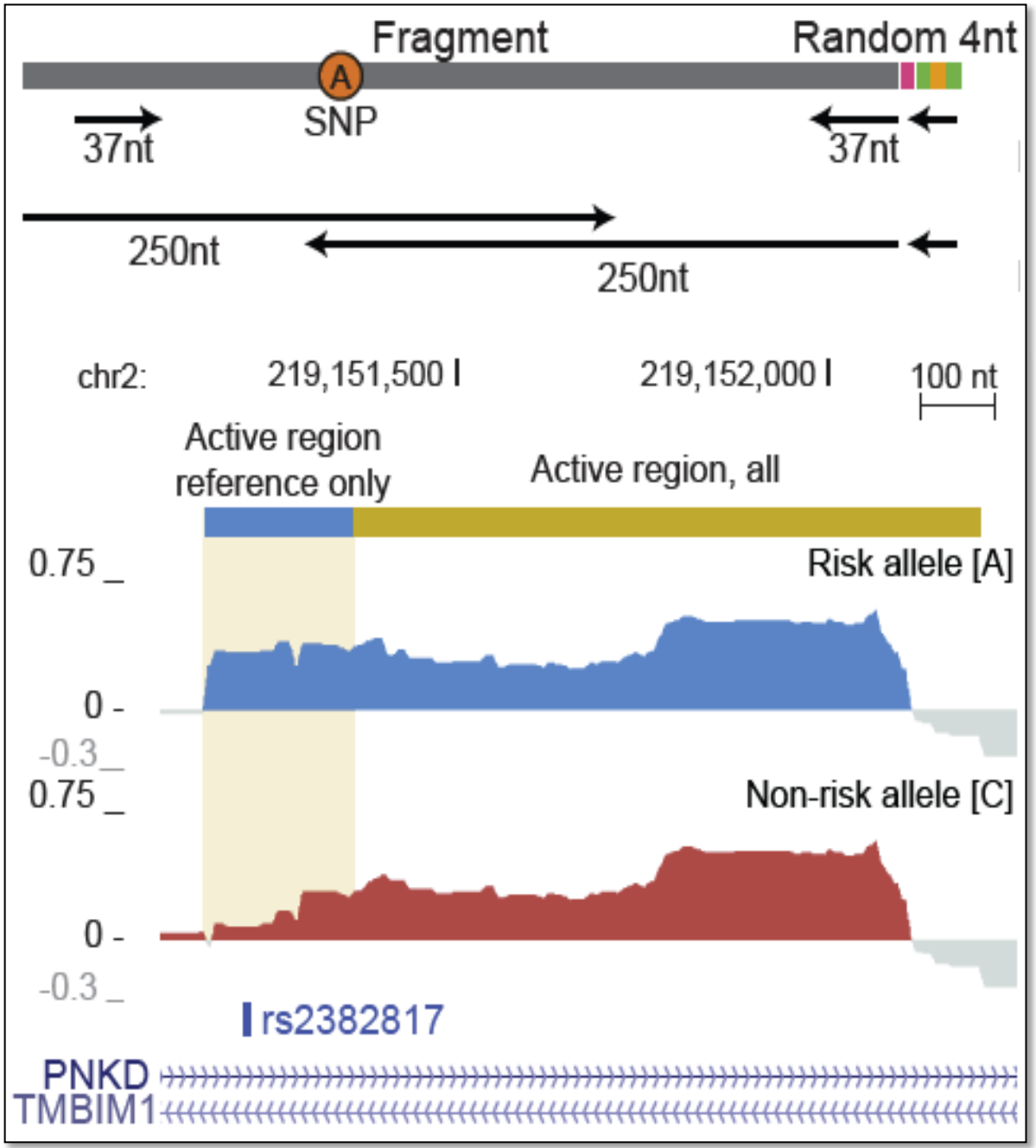


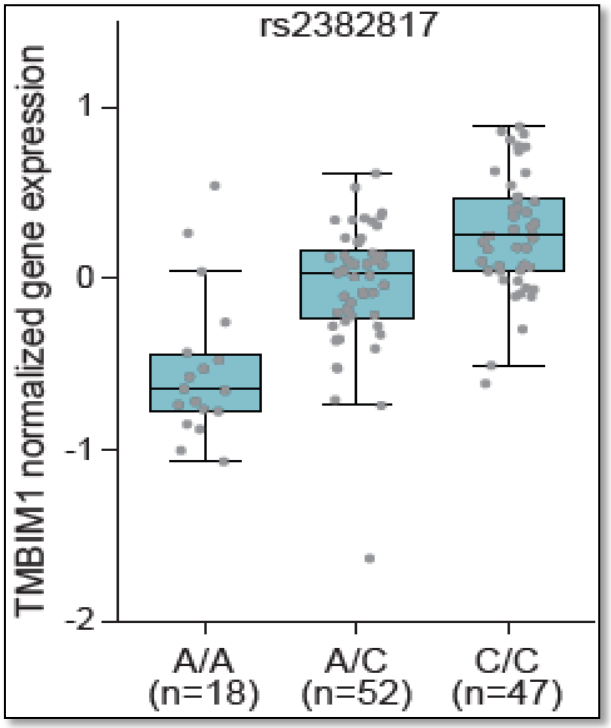
➔**General method to dissect non-coding variation**

➔**Applicable to millions of genomic regions simultaneously**

Pinpoint causal GWAS variants

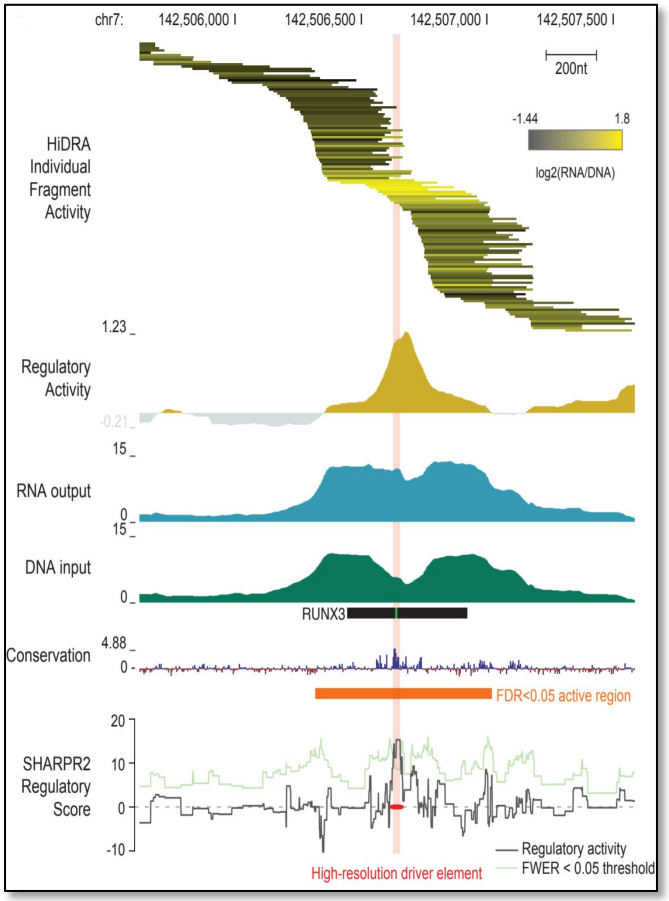
HiDRA activity differences between risk and non-risk alleles





Allele-specific activity for IBD-associated rs2382817

HiDRA summary



- 7M fragments tested in one experiment
- Longer fragments (~350nt on average)
- High reproducibility, 0.95 for higher-activity elmt
- Up to 200-fold enrichment for regulatory regions
- High-resolution dissection of driver nts
- Captures known motifs, conserved nucleotides
- Pinpoints driver SNPs in GWAS loci
- Reveals diffs between risk and non-risk alleles
- **General tool for testing regulatory regions**

bioRxiv beta

doi.org/10.1101/193136

High-resolution genome-wide functional dissection of transcriptional regulatory regions in human

Xinchen Wang[1,2,3,†], Liang He[2,3], Sarah M. Goggin[2], Alham Saadat[2], Li Wang[2], Melina Claussnitzer[2,4,*], Manolis Kellis[2,3,*]

Regulatory genomics: motifs, instances, regions

1. Introduction to regulatory motifs / gene regulation
   - Expts vs. comp. Co-regulated genes (EM, Gibbs). Conserv.
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions Z_{ij} from motif matrix
   - M step: Find max-likelihood motif from all positions Z_{ij}
3. Gibbs Sampling: Sample from joint (M,Z_{ij}) distribution
   - Sampling motif positions based on the Z vector
   - More likely to find global maximum, easy to implement
4. Evolutionary signatures for *de novo* motif discovery
   - Genome-wide conservation scores, motif extension
   - Validation of discovered motifs: functional datasets
5. Evolutionary signatures for instance identification
   - Phylogenies, Branch length score ➔ Confidence score
6. *De novo* dissection of regulatory regions in high-resolution
   - Massively-parallel reporter assays. Positioning matters.
   - 5-bp tiling for high-res dissection: Sharpr-MPRA. Insights
   - HiDRA: random ATAC fragmentation + self-reporter assays