6.047/6.878
Computational Biology: Genomes, Networks, Evolution

Lecture 10
Regulatory motif discovery and target identification

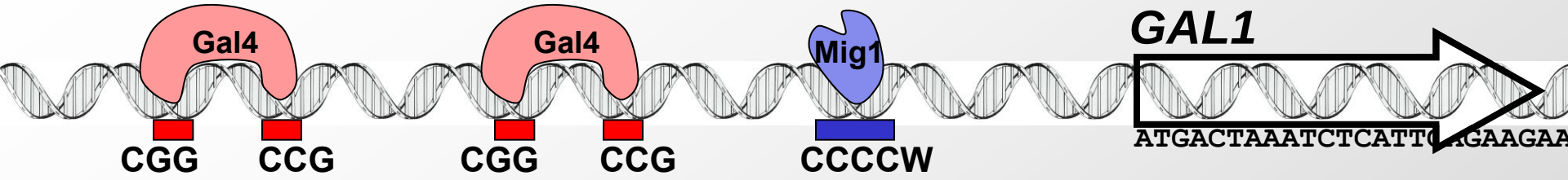
Module III: Epigenomics and gene regulation

- Computational Foundations
  - L10: Gibbs Sampling: between EM and Viterbi training
  - L11: Rapid linear-time sub-string matching
  - L11: Multivariate HMMs
  - L12: Post-transcriptional regulation
- Biological frontiers:
  - L10: Regulatory motif discovery, TF binding
  - L11: Epigenomics, chromatin states, differentiation
  - L12: Post-transcriptional regulation

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo
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
   - Foreground vs. background. Real vs. control motifs.

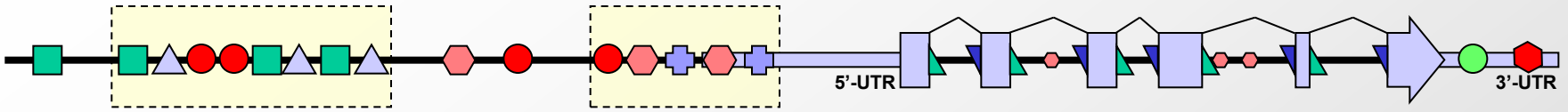
Regulatory motif discovery



- 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?

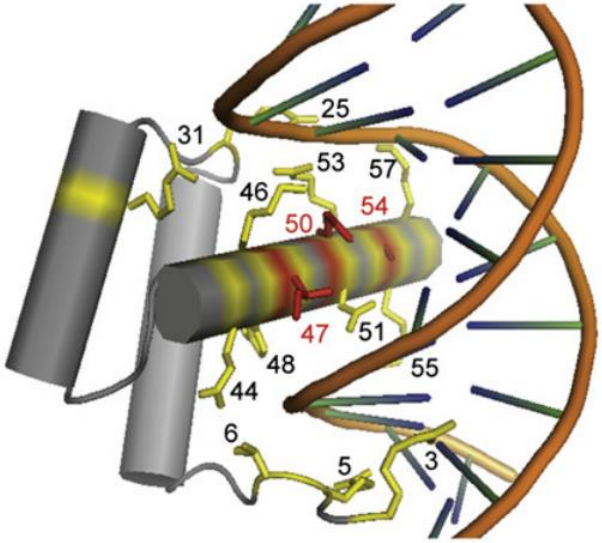
Promoter motifs
When in time?

Splicing signals
Which variants?

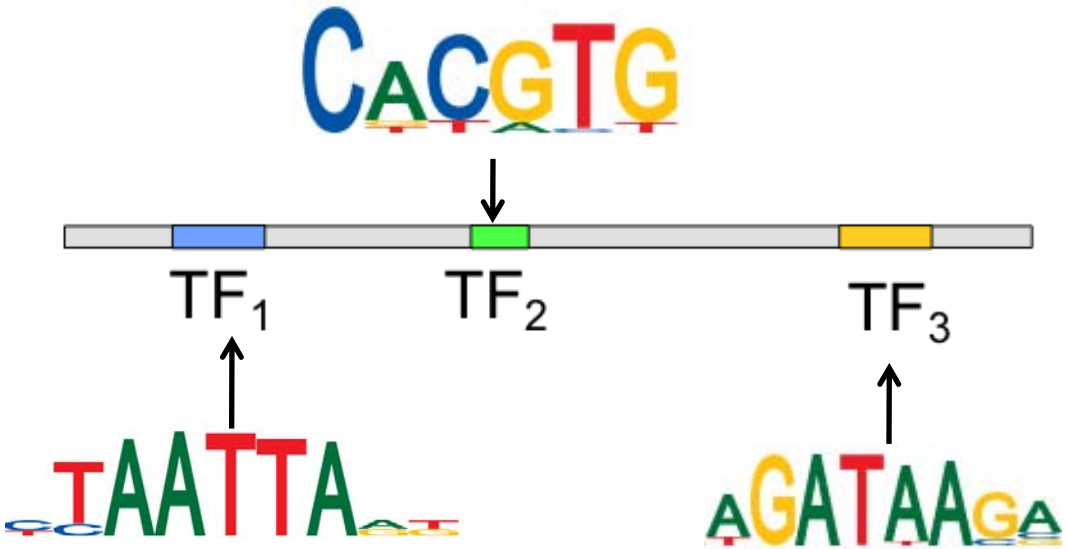
Motifs at RNA level
Which subsets?

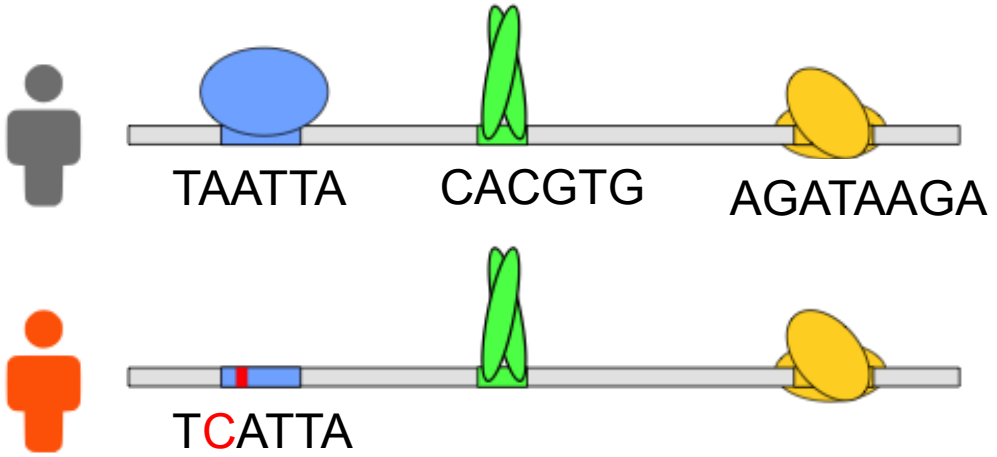
- The parts list: ~20-30k genes
  - Protein-coding genes, RNA genes (tRNA, microRNA, snRNA)
- The circuitry: constructs controlling gene usage
  - Enhancers, promoters, splicing, post-transcriptional motifs
- The regulatory code, complications:
  - Combinatorial coding of 'unique tags'
    - Data-centric 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

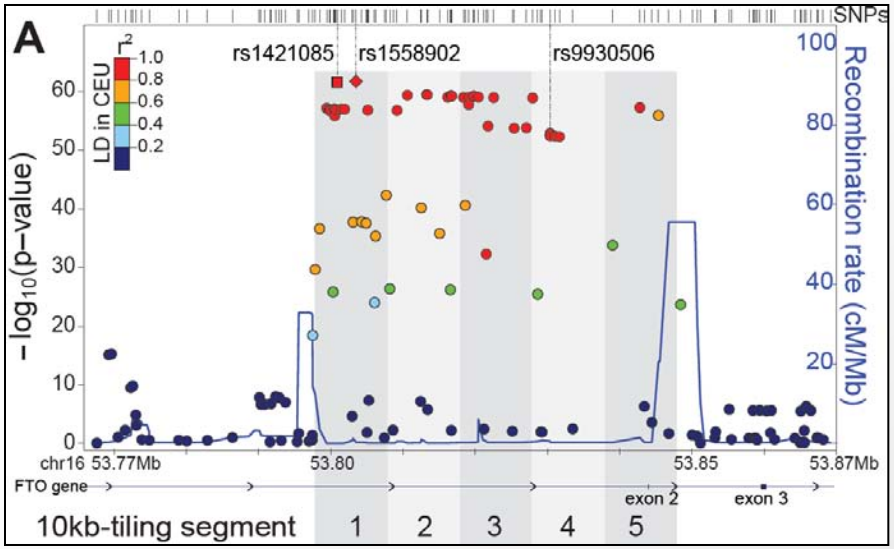


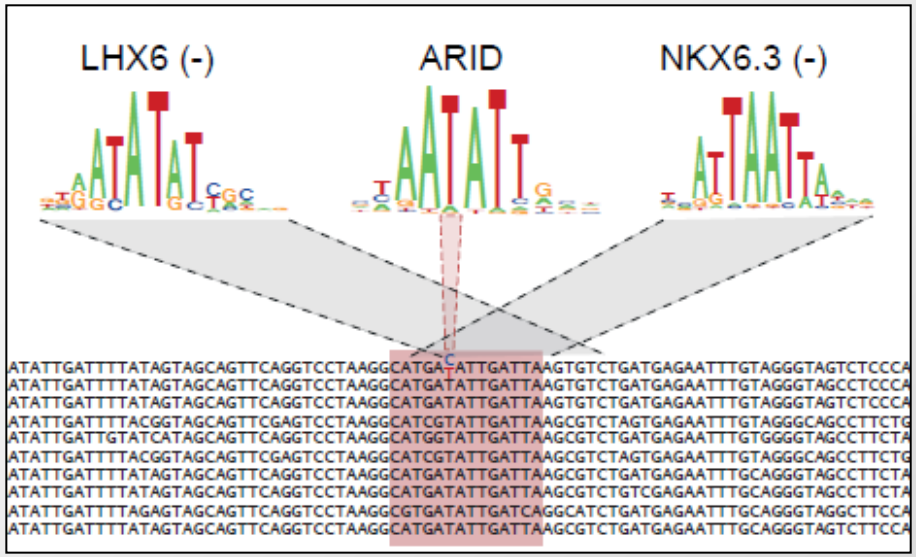
DNA-binding domain of *Engrailed*





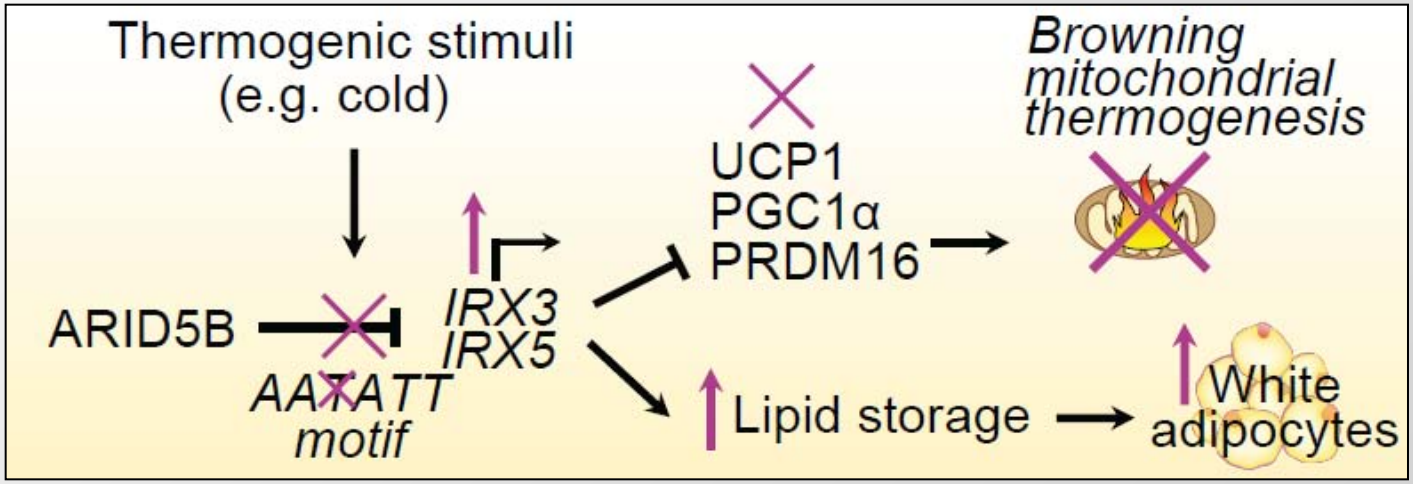
Disrupted motif at the heart of FTO obesity locus





Strongest association with obesity

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

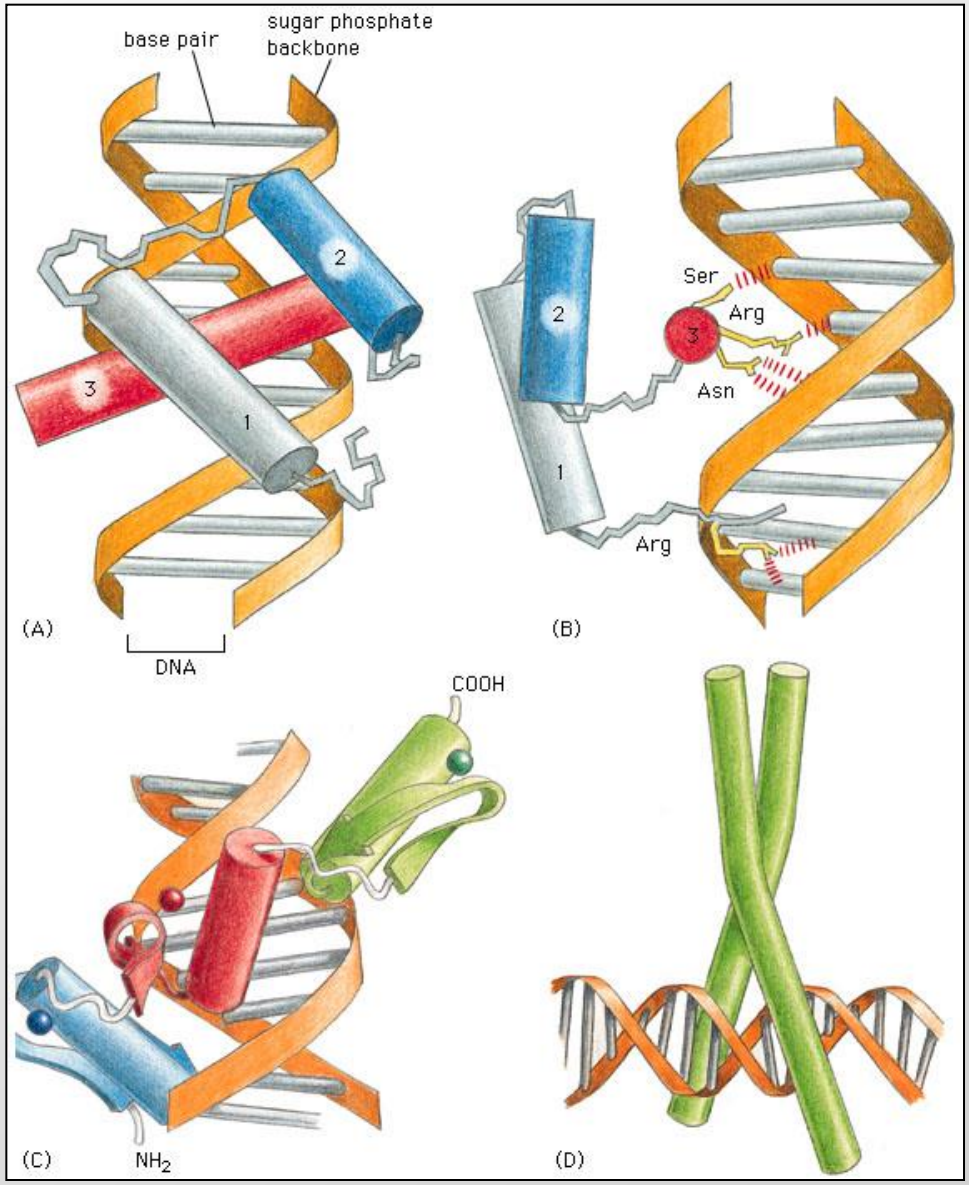


Restoring motif restores thermogenesis

Courtesy of Manolis Kellis. Used with permission.

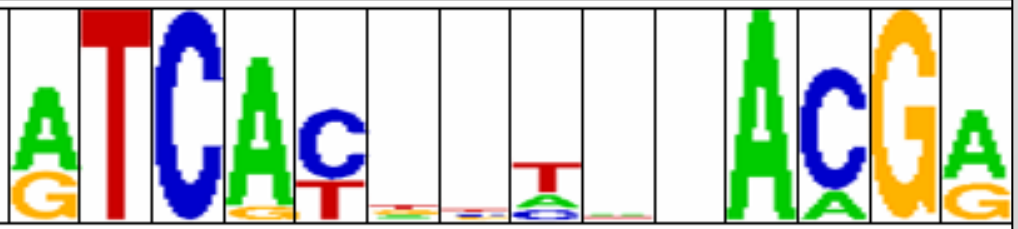
Regulator structure ⇔ recognized motifs

- 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



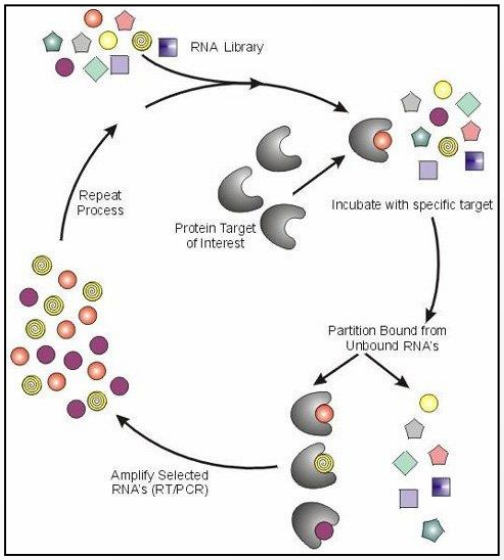
Motifs summarize TF sequence specificity

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

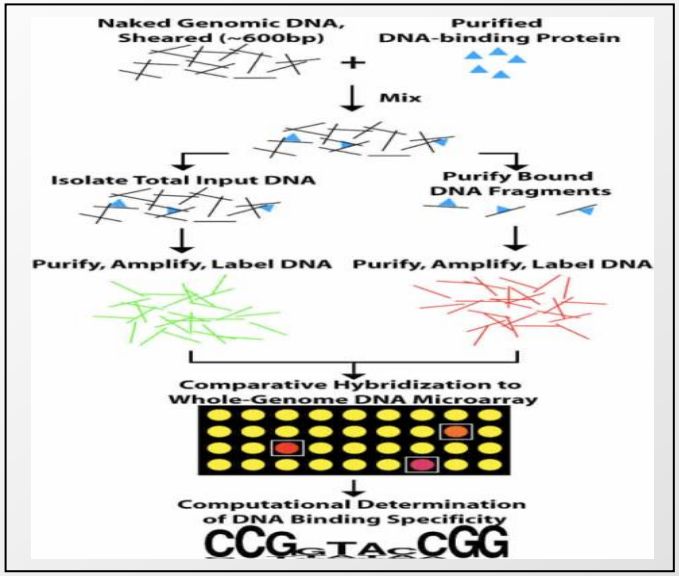
Position		1	2	3	4	5	6	7	8	9	10	11	12	13	14
Position Weight Matrix (PWM)	A	**56**	4	4	**81**	4	23	15	27	31	31	**89**	23	4	**58**
	G	**32**	4	4	12	4	31	23	4	19	23	4	4	**89**	**35**
	C	4	4	**89**	4	**58**	12	23	19	19	23	4	**69**	4	4
	T	4	**89**	4	4	**35**	35	39	50	31	23	4	4	4	4
Motif Logo														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

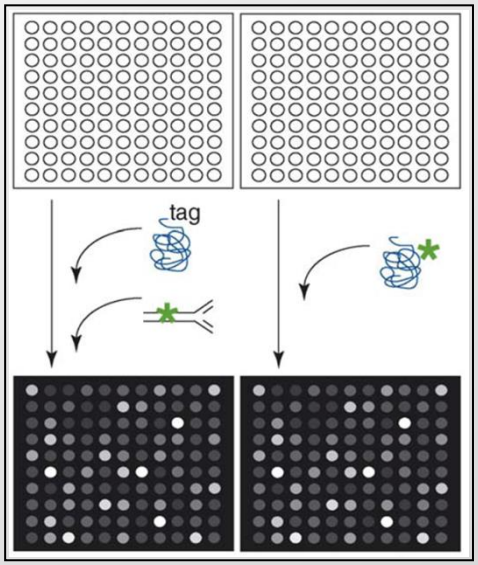
Experimental factor-centric discovery of motifs



Courtesy of the authors. Used with permission.
Source: Ray, Partha, and Rebekah R. White. "Aptamers For targeted drug delivery." Pharmaceuticals 3, no. 6 (2010): 1761-1778.



© Cold Spring Harbor Laboratory Press. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.
Source: Liu, Xiao et al. "DIP-chip: rapid and accurate determination of DNA-binding specificity." Genome Research 15, no. 3 (2005): 421-427.



© source unknown. All rights reserved. This content is excluded from our Creative Commons license. For more information, see http://ocw.mit.edu/help/faq-fair-use/.

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

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

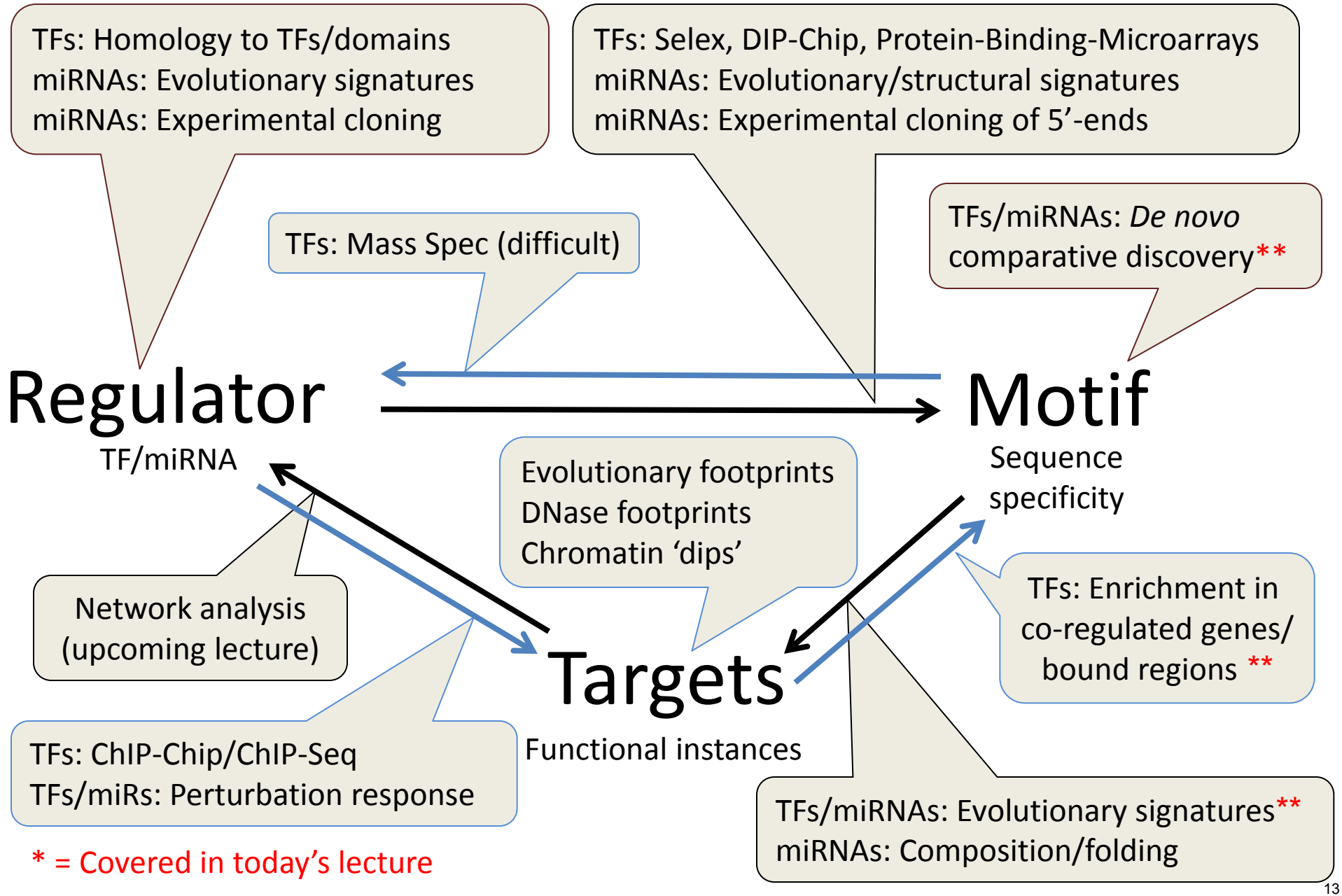
In vitro / trans

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

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

Challenges in regulatory genomics

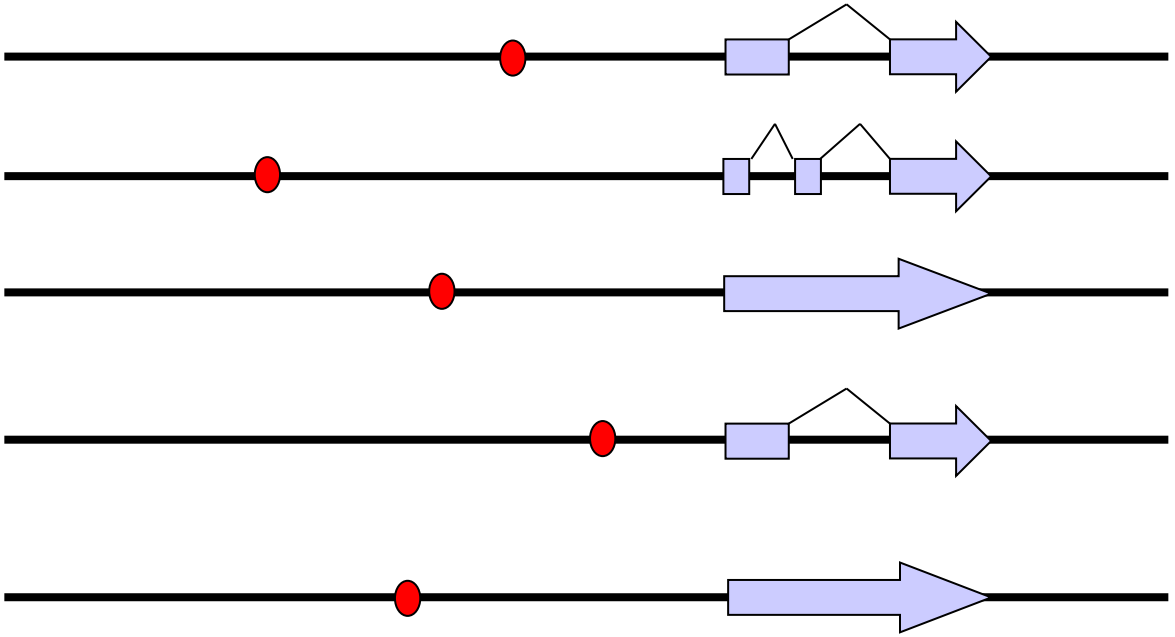


Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo
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
   - Foreground vs. background. Real vs. control motifs.

Enrichment-based discovery methods

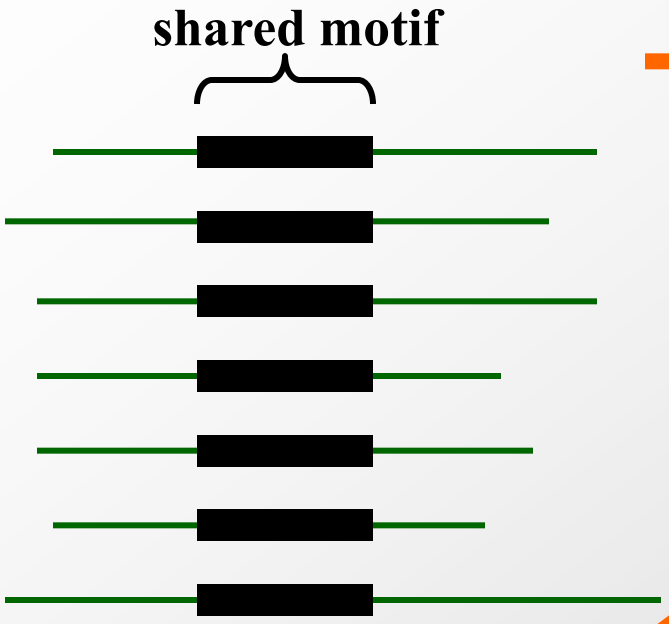
Given a set of co-regulated/functionally related genes, find common motifs in their promoter regions



- **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



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

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)

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)

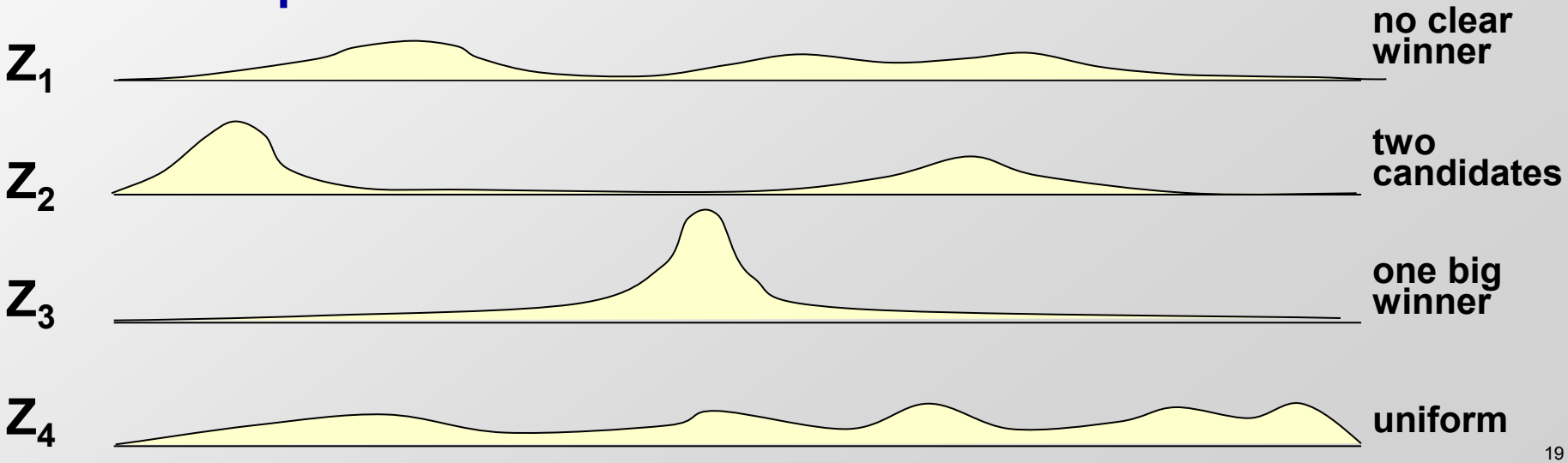
(see also: di-nucleotide etc)

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*

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

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

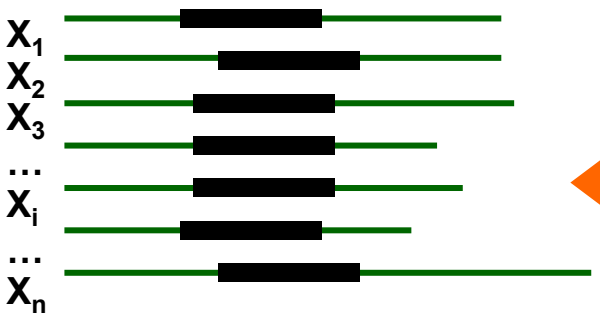
- **Computing Z_{ij} matrix from M(k,c) is straightforward**
  - **At each position, evaluate start probability by multiplying across the matrix**

- **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}**

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo
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
   - Foreground vs. background. Real vs. control motifs.

E-step:
Estimate Z_{ij} positions from matrix

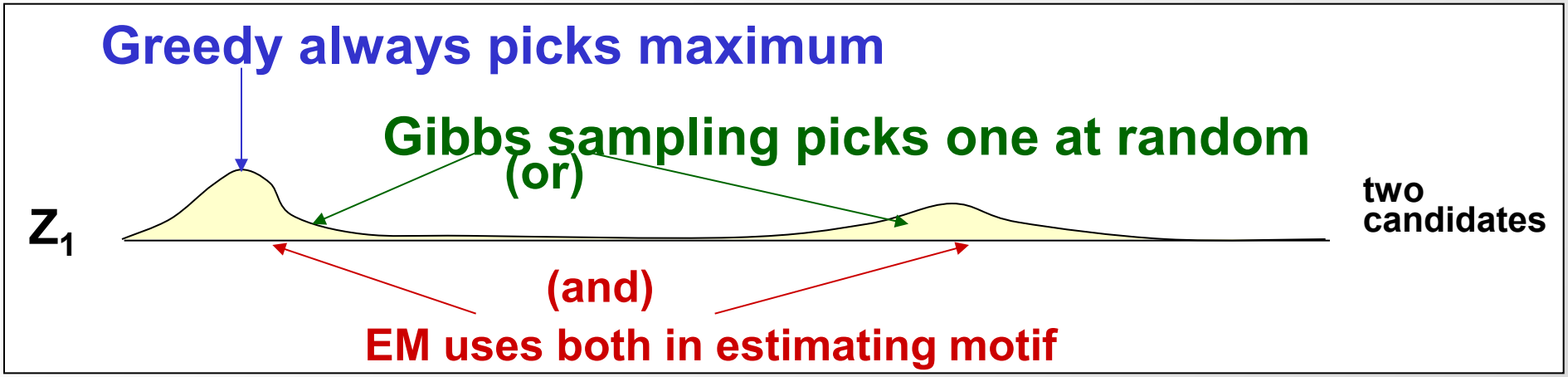


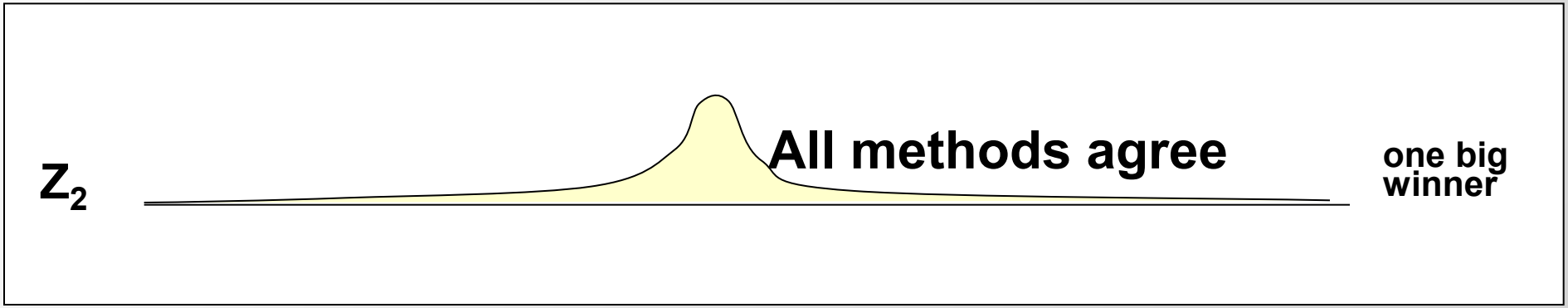
E-step

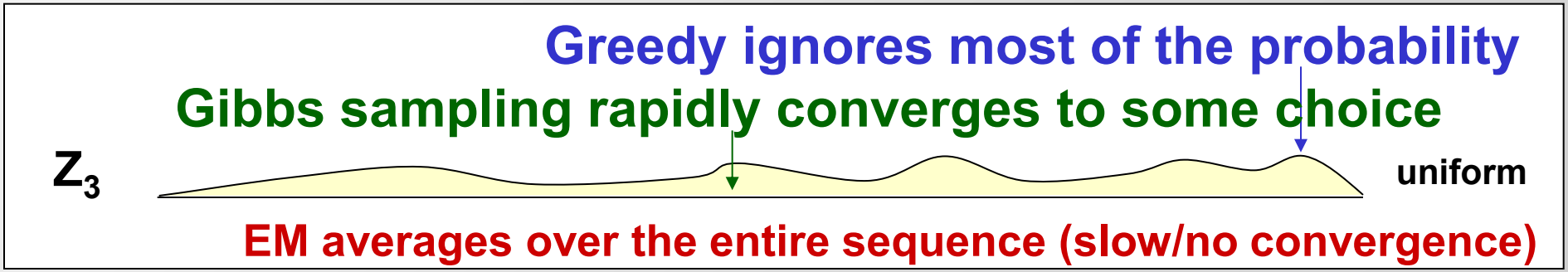
	k=1	k=2	k=3	k=4	k=5	k=6	k=7	k=8
c=A	0.1	0.3	0.1	0.2	0.2	0.4	0.3	0.1
c=C	0.5	0.2	0.1	0.1	0.6	0.1	0.2	0.7
c=G	0.2	0.2	0.6	0.5	0.1	0.2	0.2	0.1
c=T	0.2	0.3	0.2	0.2	0.1	0.3	0.3	0.1

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

Three examples for Greedy, Gibbs Sampling, EM







Calculating $P(X_i)$ when motif position is known

- Probability of training sequence X_i, given hypothesized start position j

$$\Pr(X_i \mid Z_{ij} = 1, M, B) = \underbrace{\prod_{k=1}^{j-1} B(X_{i,k})}_{\text{before motif}} \underbrace{\prod_{k=j}^{j+W-1} M(k-j+1, X_{i,k})}_{\text{motif}} \underbrace{\prod_{k=j+W}^{L} B(X_{i,k})}_{\text{after motif}}$$

- **Example:**

$X_i =$ G C **T G T** A G

$B =$

A	0.25
C	0.25
G	0.25
T	0.25

$M =$

	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

$$\Pr(X_i \mid Z_{i3} = 1, M, B) =$$

$$B(\mathrm{G}) \times B(C) \times M(1,T) \times M(2,G) \times M(3,T) \times B(A) \times B(G) =$$

$$0.25 \times 0.25 \times 0.2 \times 0.1 \times 0.1 \times 0.25 \times 0.25$$

Calculating the Z vector (using M)

- **To estimate the starting positions in *Z* at step *t***

$$Z_{ij}^{(t)} = \Pr(Z_{ij} = 1 \mid X_i, M^{(t)}) = \frac{\Pr(X_i \mid Z_{ij} = 1, M^{(t)}) \Pr(Z_{ij} = 1)}{\Pr(X_i)}$$

likelihood (numerator) · prior (numerator) · posterior (left) · evidence (denominator) · **(Bayes' rule)**

- At iteration t, calculate $Z_{ij}^{(t)}$ based on $M^{(t)}$
  - We just saw how to calculate $\Pr(X_i \mid Z_{ij}=1, M^{(t)})$
  - To obtain total probability $\Pr(X_i)$, sum over all starting positions

$$Z_{ij}^{(t)} = \frac{\Pr(X_i \mid Z_{ij} = 1, M^{(t)}) \cancel{\Pr(Z_{ij} = 1)}}{\sum_{k=1}^{L-W+1} \Pr(X_i \mid Z_{ik} = 1, M^{(t)}) \cancel{\Pr(Z_{ik} = 1)}}$$

- **Assume uniform priors (motif eq likely to start at any position)**

Calculating the Z vector: Example

$X_i =$ G C T G T A G

$p =$

	0	1	2	3
A	0.25	0.1	0.5	0.2
C	0.25	0.4	0.2	0.1
G	0.25	0.3	0.1	0.6
T	0.25	0.2	0.2	0.1

$$Z_{i1} = 0.3 \times 0.2 \times 0.1 \times 0.25 \times 0.25 \times 0.25 \times 0.25$$

$$Z_{i2} = 0.25 \times 0.4 \times 0.2 \times 0.6 \times 0.25 \times 0.25 \times 0.25$$

$\vdots$

- then normalize so that

$$\sum_{j=1}^{L-W+1} Z_{ij} = 1$$

Aside: Simplifying P(X_i)

- Probability of training sequence X_i, given hypothesized start position j

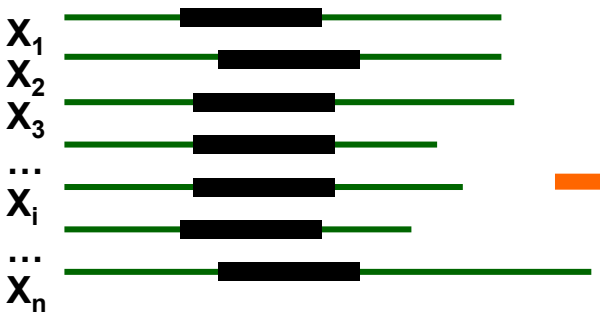
$$\Pr(X_i \mid Z_{ij} = 1, M, B) = \underbrace{\prod_{k=1}^{j-1} B(X_{i,k})}_{\text{before motif}} \underbrace{\prod_{k=j}^{j+W-1} M(k-j+1, X_{i,k})}_{\text{motif}} \underbrace{\prod_{k=j+W}^{L} B(X_{i,k})}_{\text{after motif}}$$

$$= \prod_{k=j}^{j+W-1} \underbrace{\frac{M(k-j+1, X_{i,k})}{B(X_{i,k})}}_{\text{can be stored in a matrix}} \underbrace{\prod_{k=1}^{L} B(X_{i,k})}_{\text{constant for each sequence}}$$

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo
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
   - Foreground vs. background. Real vs. control motifs.

M-step:
Max-likelih motif from Z_{ij} positions



M-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)

The M-step: Estimating the motif *M*

- recall $M(k,c)$ represents the probability of character *c* in position *k* ; $B(c)$ stores values for the background

$$M^{(t+1)}(k,c) = \frac{n_{k,c} + d}{\sum_{c}(n_{k,c} + d)}$$

pseudo-counts

where $$n_{c,k} = \sum_{i}{}' \sum_{\{j | X_{i,j+k-1} = c\}}{}' Z_{ij}$$

$$B^{(t+1)}(c) = \frac{n_{0,c} + d}{\sum_{c}(n_{0,c} + d)}$$

where $$n_{0,c} = n_c - \sum_{j=1}^{W} n_{j,c}$$

total # of c's in data set

M-step example: Estimating M(k,c) from Z_{ij}

X_1 = A C A G C A

Z_1 = 0.1 0.7 0.1 0.1

X_2 = A G G C A G

Z_2 = 0.4 0.1 0.1 0.4

X_3 = T C A G T C

Z_3 = 0.2 0.6 0.1 0.1

$$M(1,A) = \frac{Z_{1,1} + Z_{1,3} + Z_{2,1} + Z_{3,3} + 1}{Z_{1,1} + Z_{1,2} \ldots + Z_{3,3} + Z_{3,4} + 4}$$

Em approach: Avg'em all

Gibbs sampling: Sample one

Greedy: Select max

- EM: sum over full probability
  - $n_{1,A}$= 0.1+0.1+0.4+0.1 = 0.7
  - $n_{1,C}$= 0.7+0.4+0.6 = 1.7
  - $n_{1,G}$= 0.1+0.1+0.1+0.1= 0.4
  - $n_{1,T}$= 0.2 = 0.2
  - Total: T=0.7+1.7+0.4+0.2 = 3.0
- Normalize and add pseudo-counts
  - M(1,A) = (0.7+1)/(T+4) = 1.7/7=0.24
  - M(1,C) = (1.7+1)/(T+4) = 2.7/7=0.39
  - M(1,G) = (0.4+1)/(T+4) = 1.4/7=0.2
  - M(1,T) = (0.2+1)/(T+4) = 1.2/7=0.17
- M(k,c) =

	1	2	3
A	0.24	0.39	0.21
C	0.39	0.21	0.18
G	0.2	0.24	0.44
T	0.17	0.16	0.16

The EM Algorithm

- EM converges to a local maximum in the likelihood of the data given the model:

$$\prod_i \Pr(X_i \mid M, B)$$

- **Deterministic iterations max direction of ascent**
- **Usually converges in a small number of iterations**
- **Sensitive to initial starting point (i.e. values in *M*)**

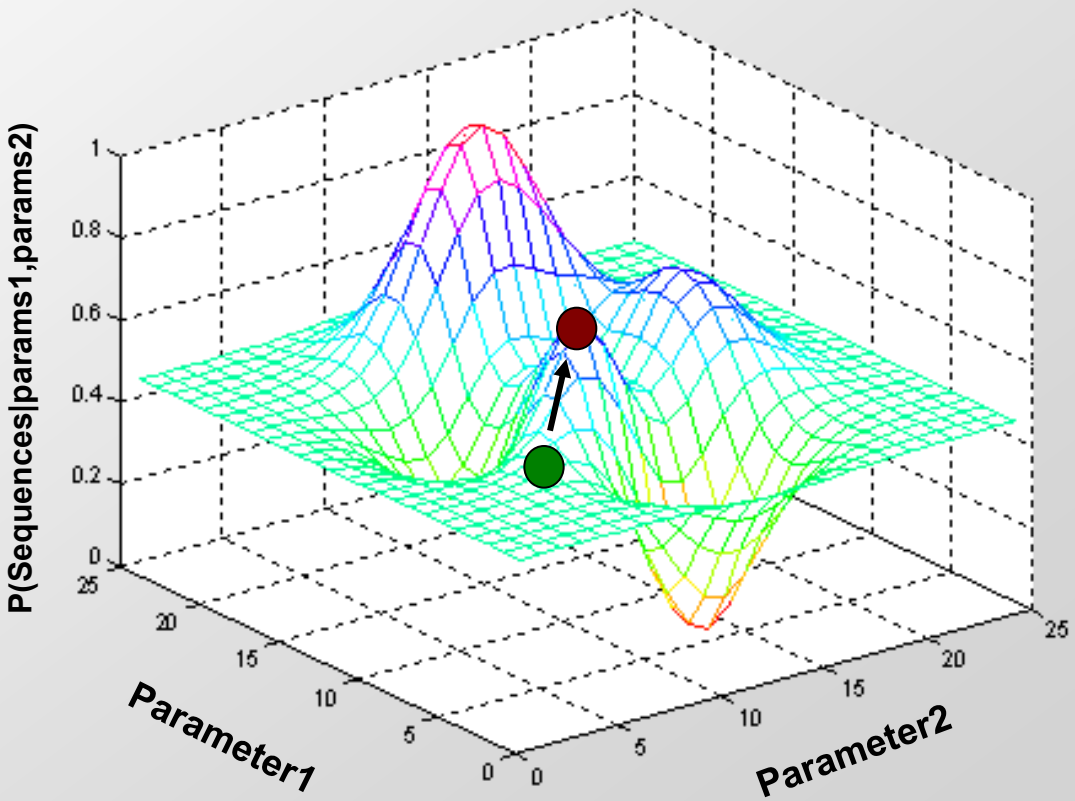
P(Seq|Model) Landscape

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

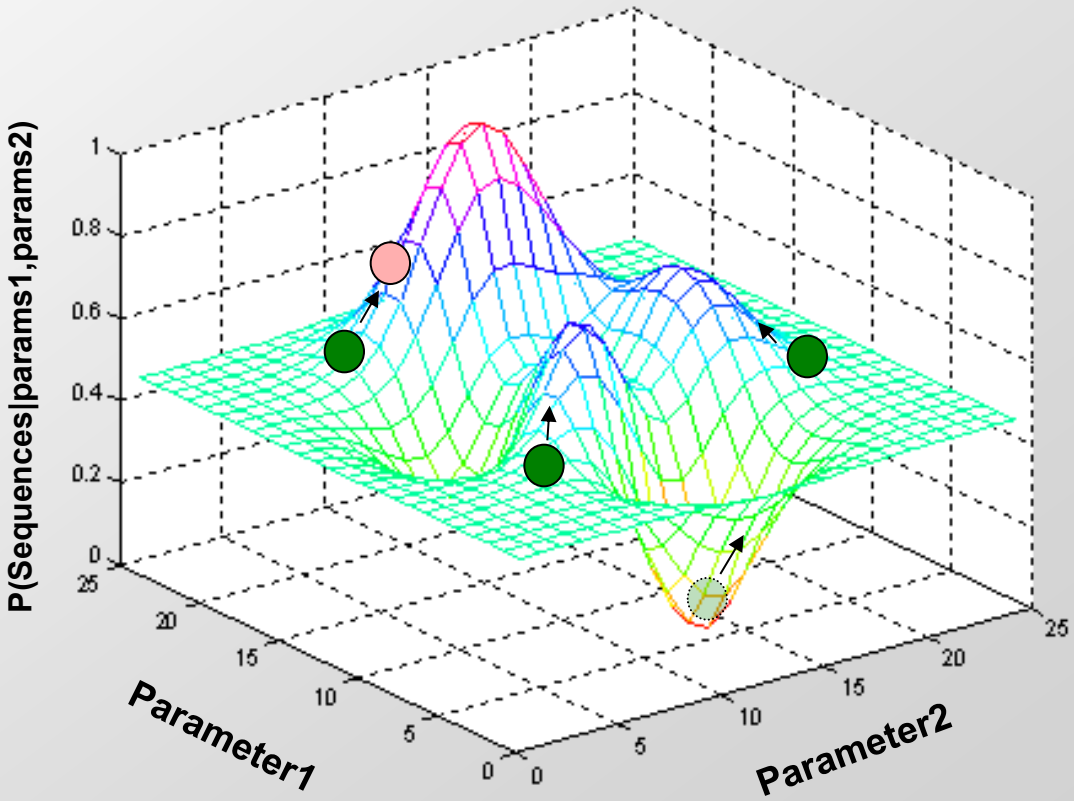
To minimize the effects of local maxima, you should search multiple times from different starting points

MEME uses this idea

Start at many points

Run for one iteration

Choose starting point that got the “highest” and continue



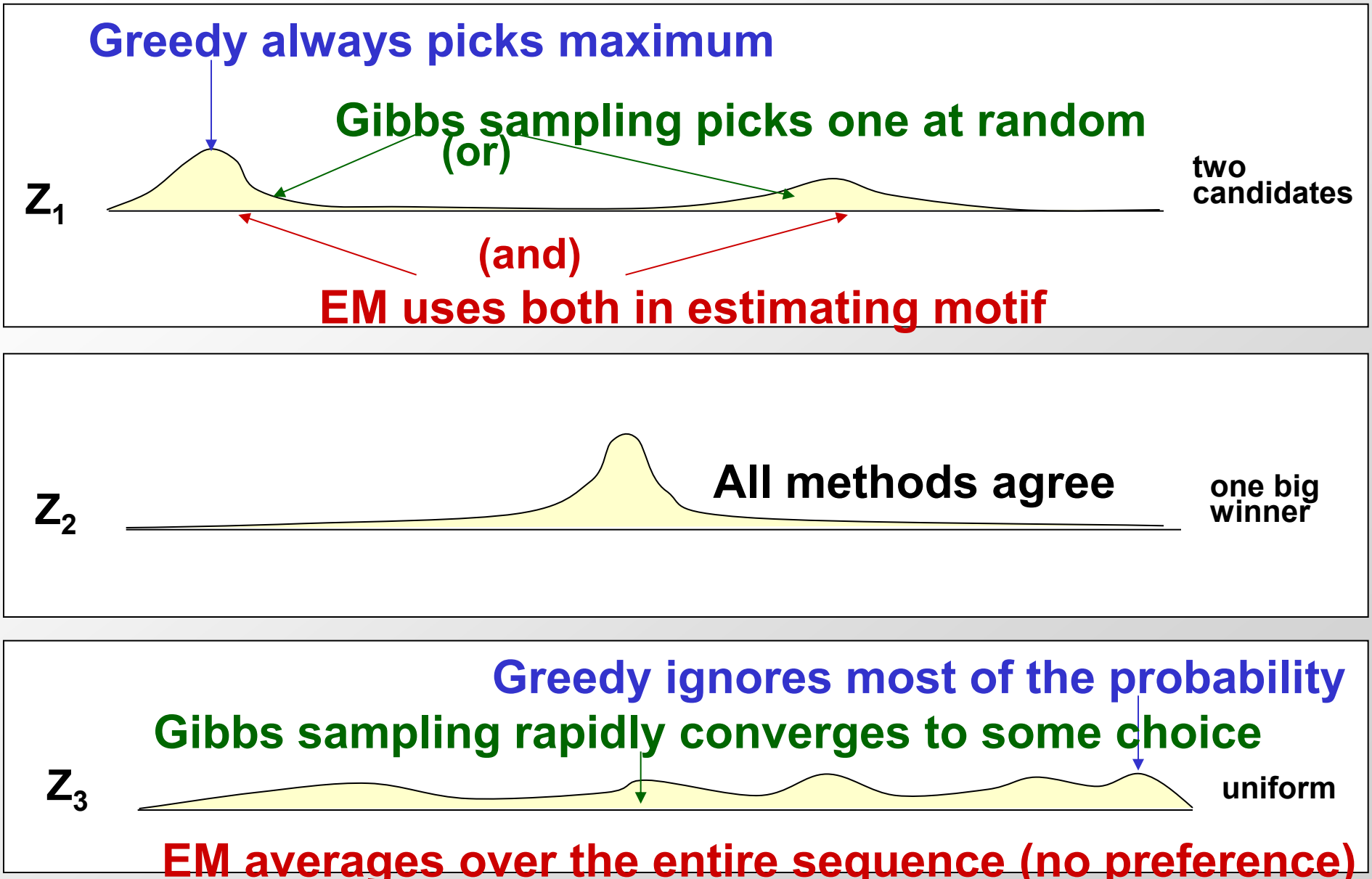
Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo
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
   - Foreground vs. background. Real vs. control motifs.

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)

Three examples of Greedy, Gibbs Sampling, EM



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 $\boldsymbol{p}$ 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 (AlignACE)

- Given:
  - $X_1, \ldots, X_N$,
  - motif length W,
  - background B,

$$\sum_{i=1}^{N}\sum_{k=1}^{W}\log\frac{M(k, X_{i,a_i+k})}{B(X_{i,a_i+k})}$$

- Find:
  - Model M
  - Locations $a_1,\ldots, a_N$ in $X_1, \ldots, X_N$

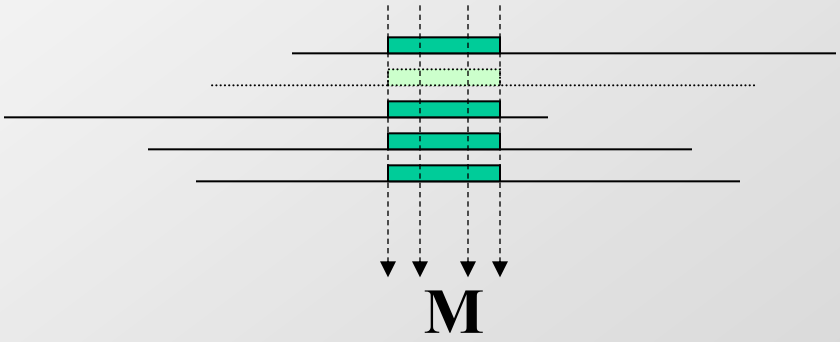
Maximizing log-odds likelihood ratio

This is the same as the EM objective (notice log and notation change)

Gibbs Sampling (AlignACE)

Predictive Update:

- Select a sequence x_i
- Remove x_i, recompute model:



$$M(k,c) = \frac{d + \sum_{s \neq i} (X_{s,a_s+k} = c)}{(N-1) + 4d}$$

where d is a pseudocount to avoid 0s

Sampling New Motif Positions

- for each possible starting position, $a_i=j$, compute a weight

$$A_j = \prod_{k=j}^{j+W-1} \frac{M(k-j+1, X_{i,k})}{B(X_{i,k})}$$

- randomly select a new starting position $\boldsymbol{a_i}$ according to these weights (normalizing across the sequence, again like with MEME)

- Note, this is equivalent to using the likelihood from MEME because:

$$A_j \propto \Pr(X_i \mid Z_{ij} = 1, p)$$

Advantages / Disadvantages

- Very similar to EM

Advantages:

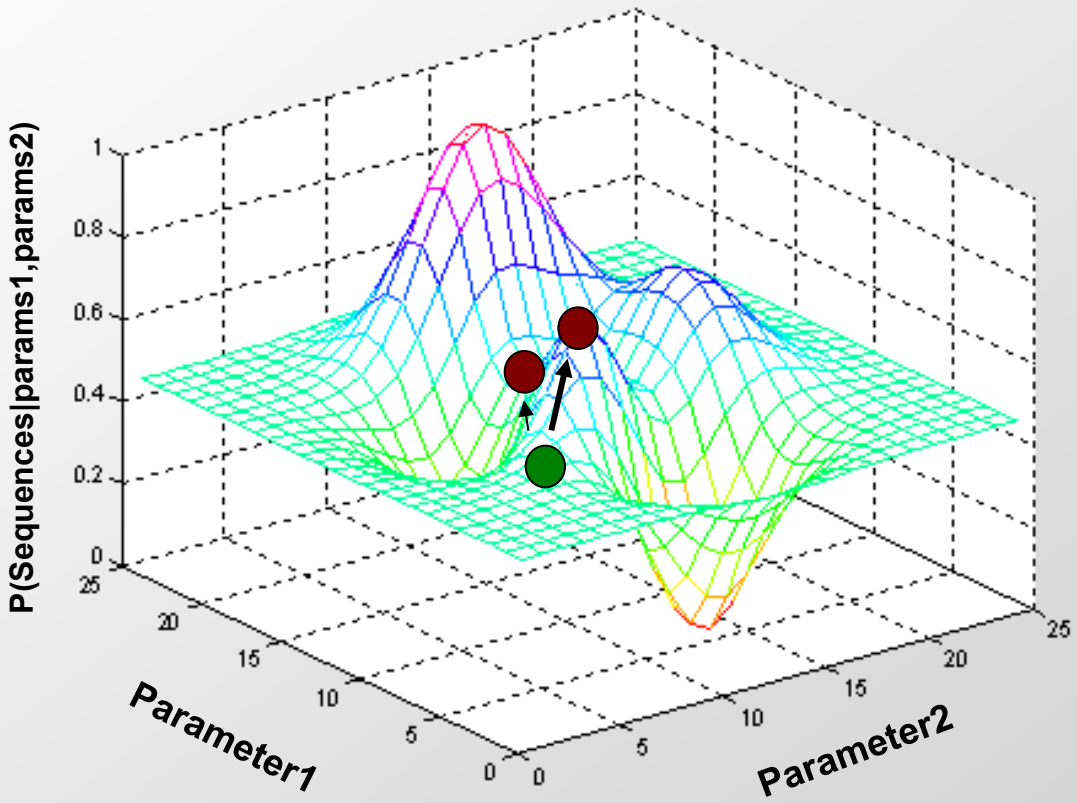
- Easier to implement
- Less dependent on initial parameters
- More versatile, easier to enhance with heuristics

Disadvantages:

- More dependent on all sequences to exhibit the motif
- Less systematic search of initial parameter space

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**

Motif discovery overview

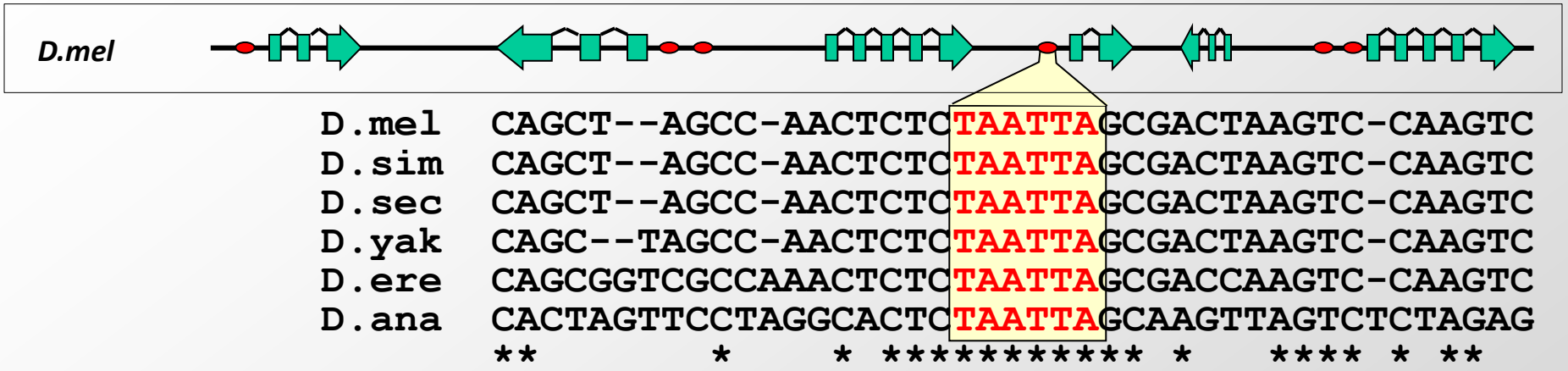
1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions Z_{ij} from motif matrix
   - M step: Find max-likelihood motif from all positions Z_{ij}
3. Gibbs 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
   - Foreground vs. background. Real vs. control motifs.

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

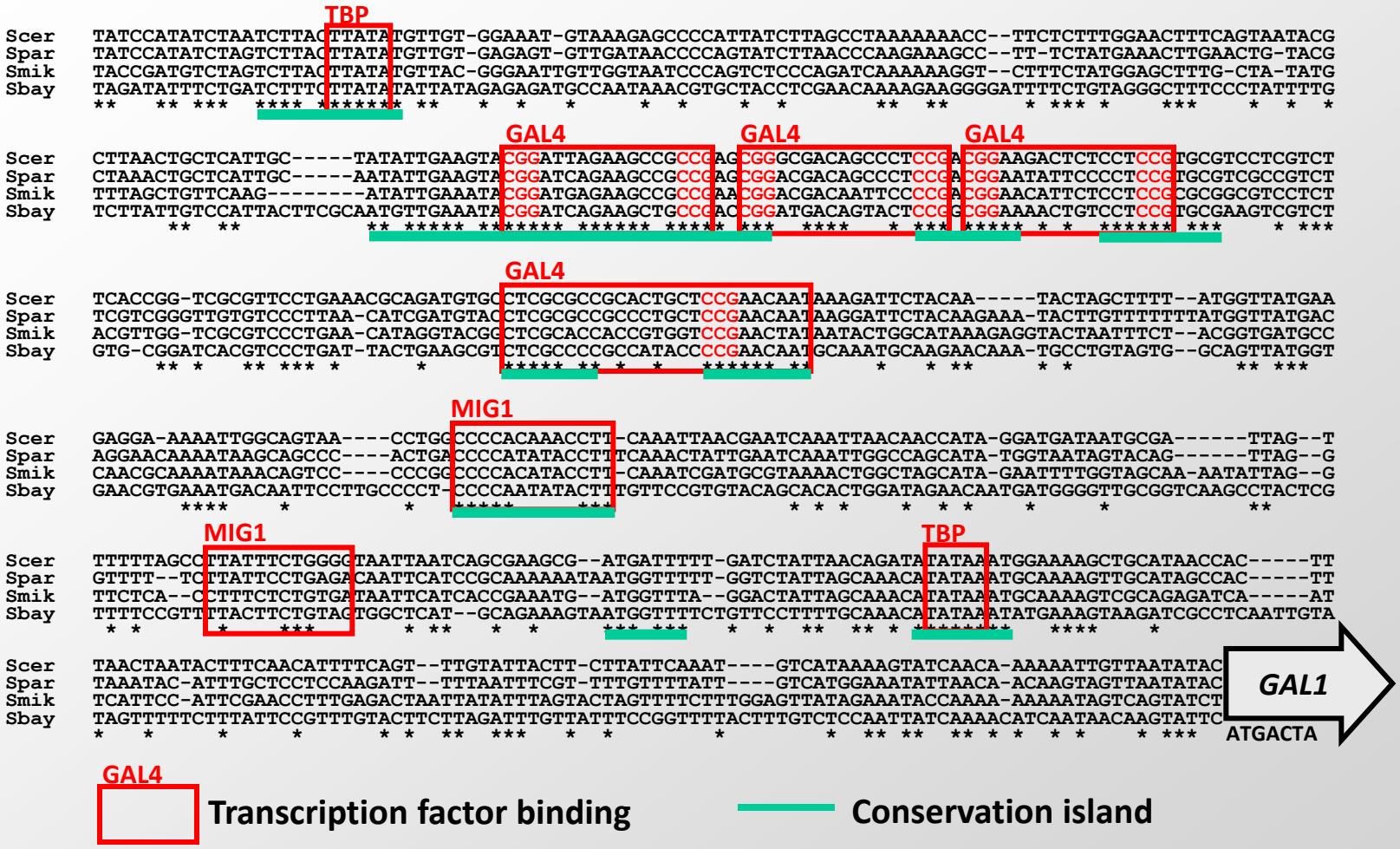
Known engrailed binding site



- 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

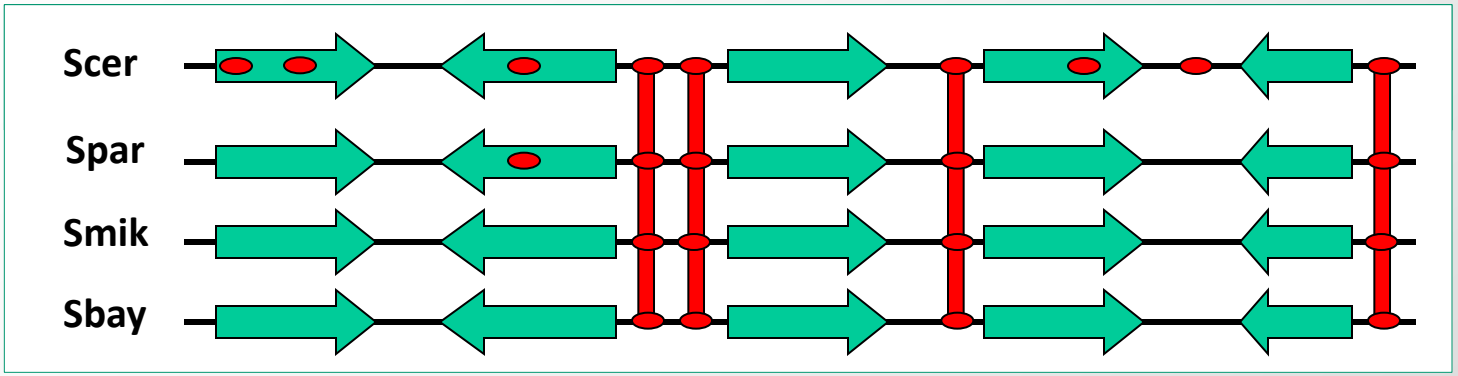
Kellis *el al,* Nature 2003
Xie *et al.* Nature 2005
Stark *et al*, Nature 2007

Conservation islands overlap known motifs



Increase power by testing conservation in many regions

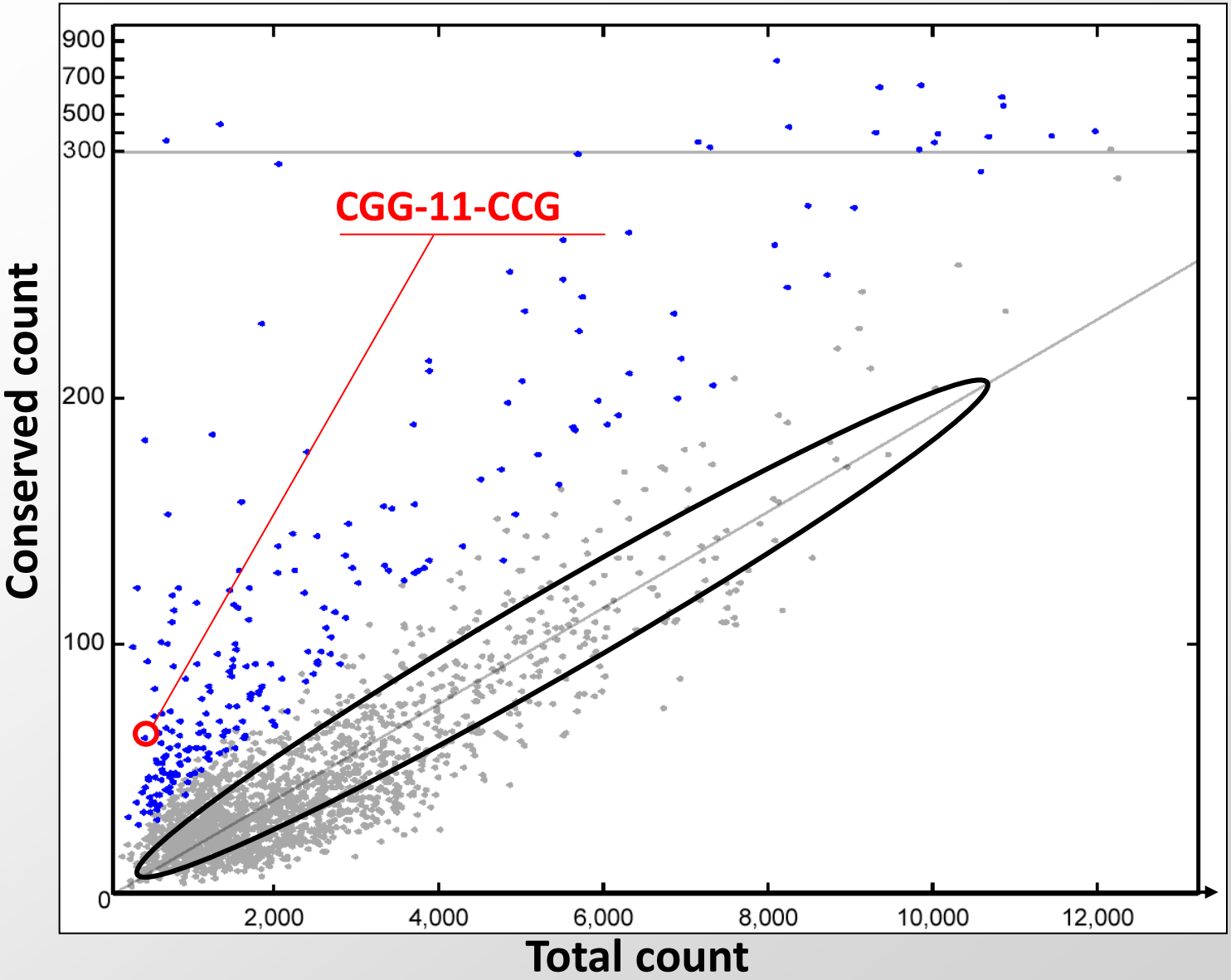
Genome-wide conservation



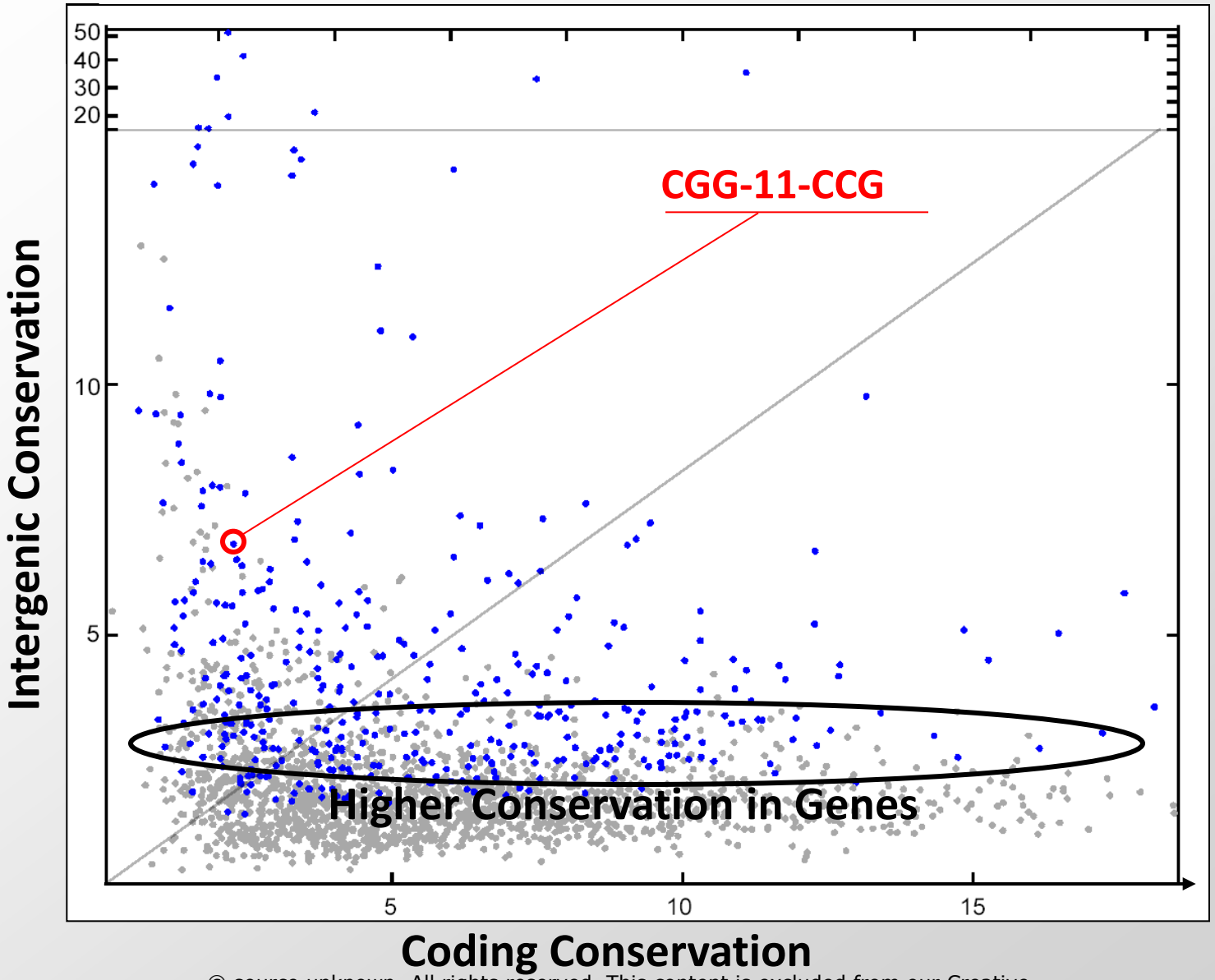
Evaluate conservation within:	Gal4	Controls
(1) All intergenic regions	13%	2%
(2) Intergenic : coding	13% : 3%	2% : 7%
(3) Upstream : downstream	12:0	1:1

A signature for regulatory motifs

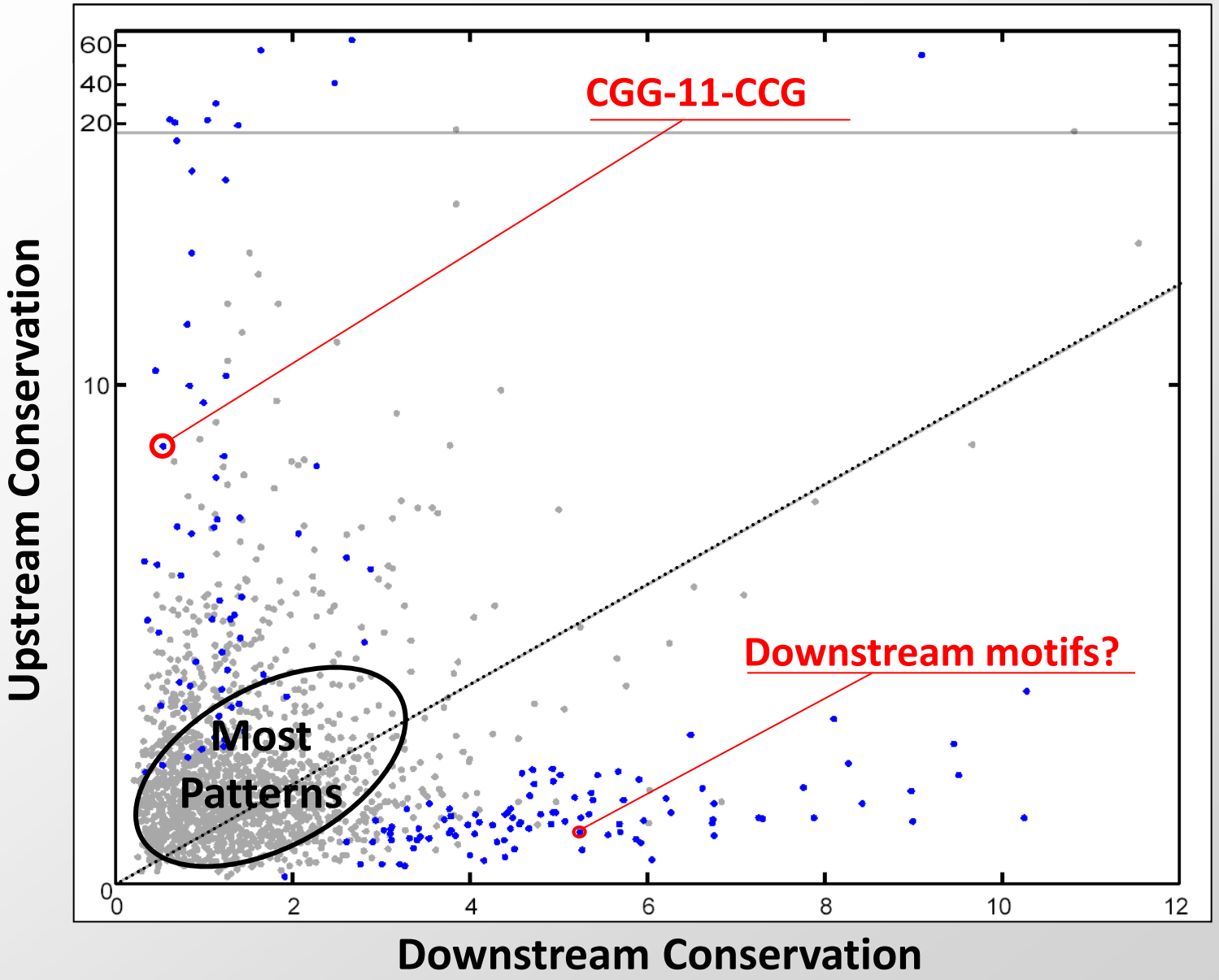
Test 1: Intergenic conservation



Test 2: Intergenic vs. Coding



Test 3: Upstream vs. Downstream



Conservation for TF motif discovery

1. **Enumerate motif seeds**

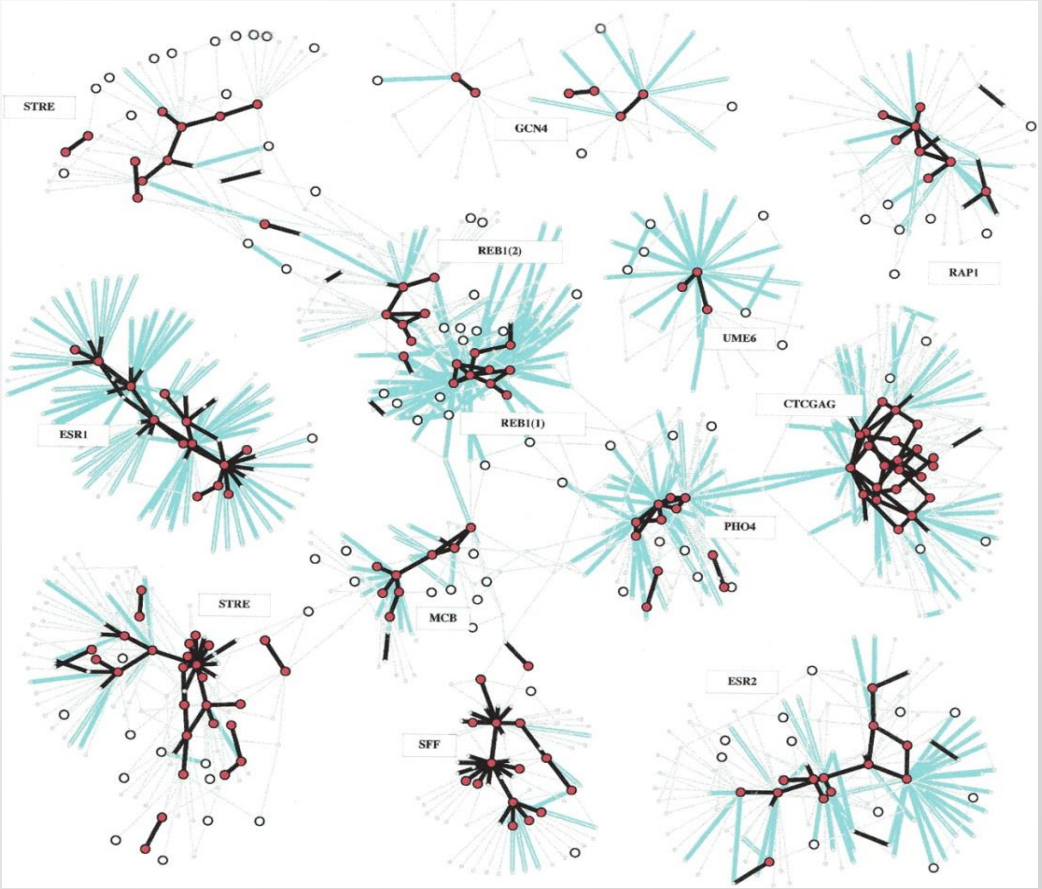
   - Six non-degenerate characters with variable size gap in the middle
2. **Score seed motifs**
   - Use a conservation ratio corrected for composition and small counts to rank seed motifs
3. **Expand seed motifs**

   - 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



Source: Tanay, Amos et al. "A global view of the selection forces in the evolution of yeast cis-regulation." Genome Research 14, no. 5 (2004): 829-834.

Tanay, Genome Research 2004

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo
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
   - Foreground vs. background. Real vs. control motifs.

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

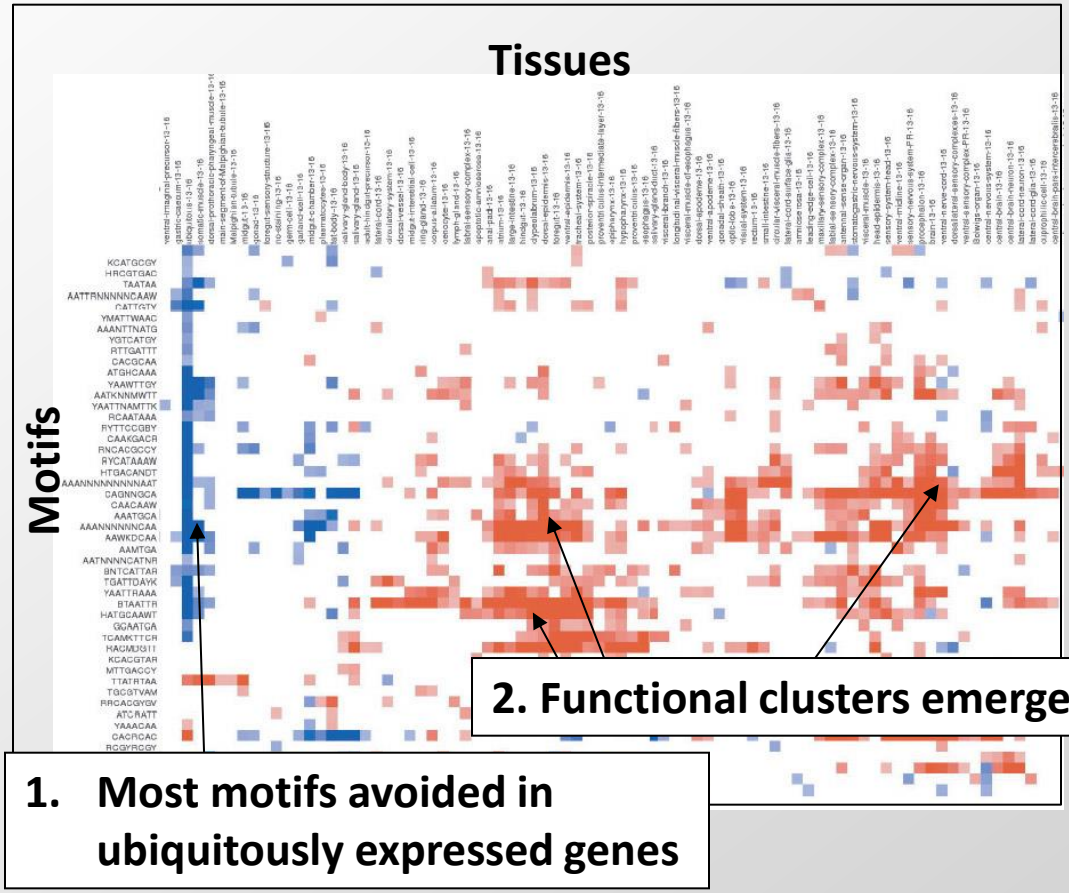
→Clustering of motif matches

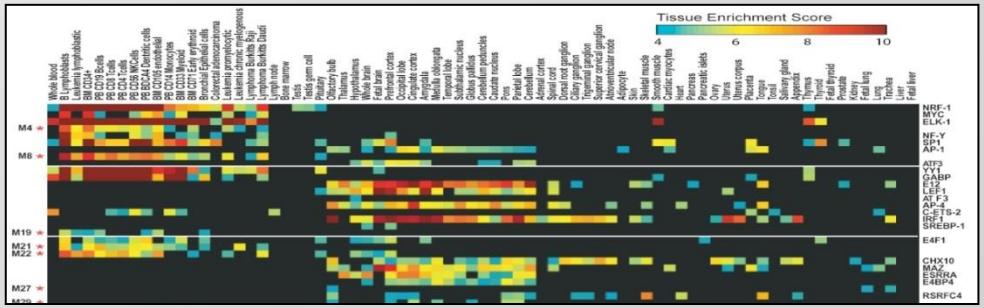
  – 19% of fly motifs

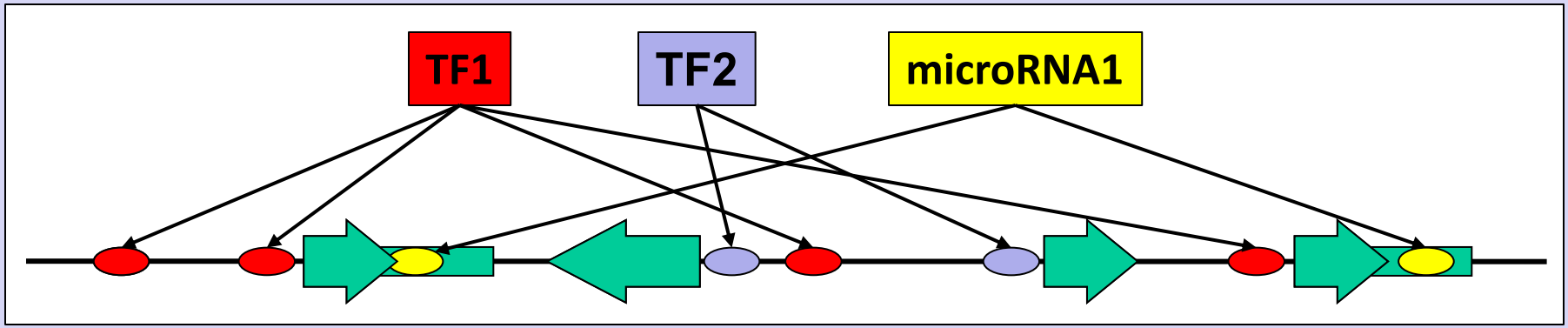
Motifs have functional enrichments

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

Reveals modules of cooperating motifs







Motif instance identification

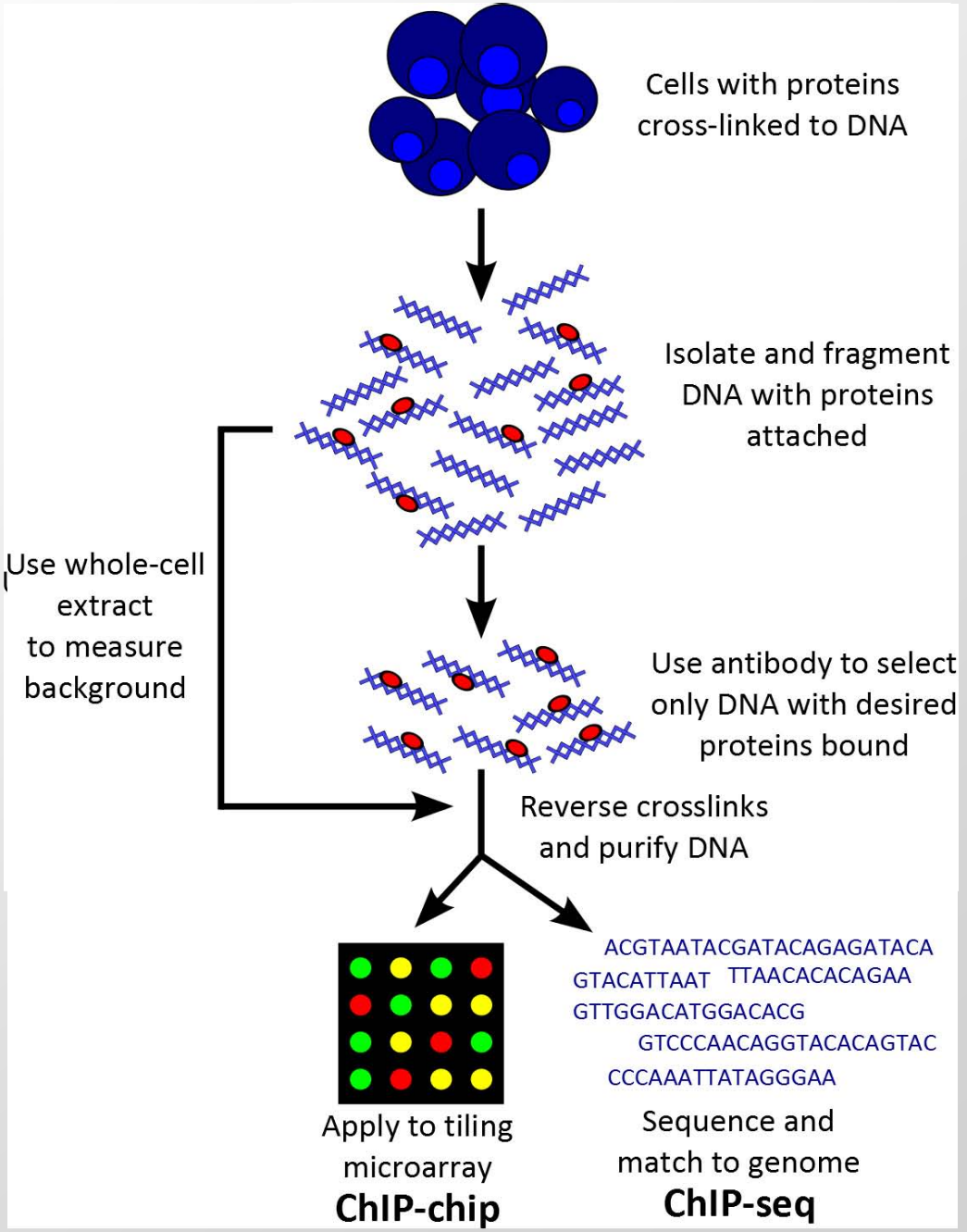
How do we determine the functional binding sites of regulators?

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)



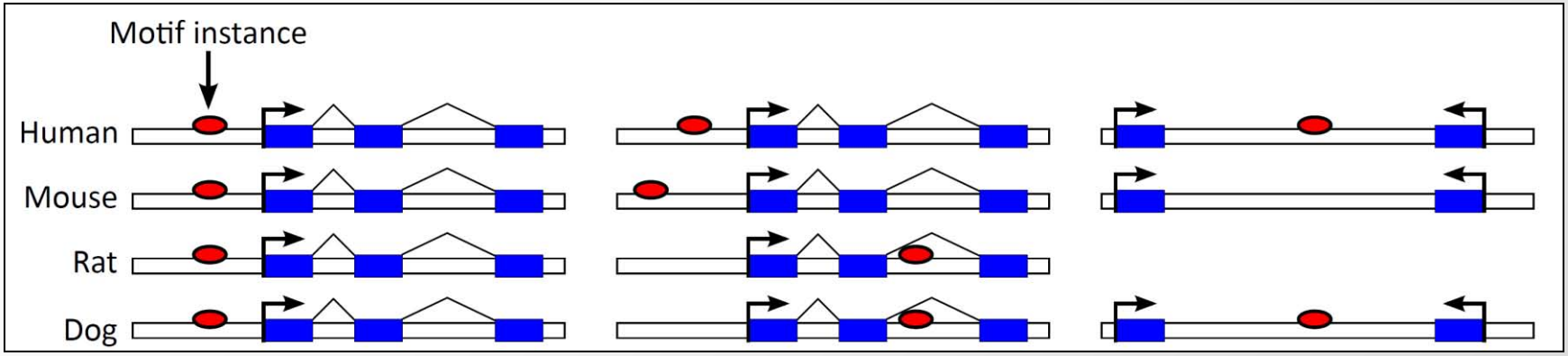
Ren et al., 2000; Iyer et al., 2001 (ChIP-chip)

Robertson et al., 2007 (ChIP-seq)

Computational target identification

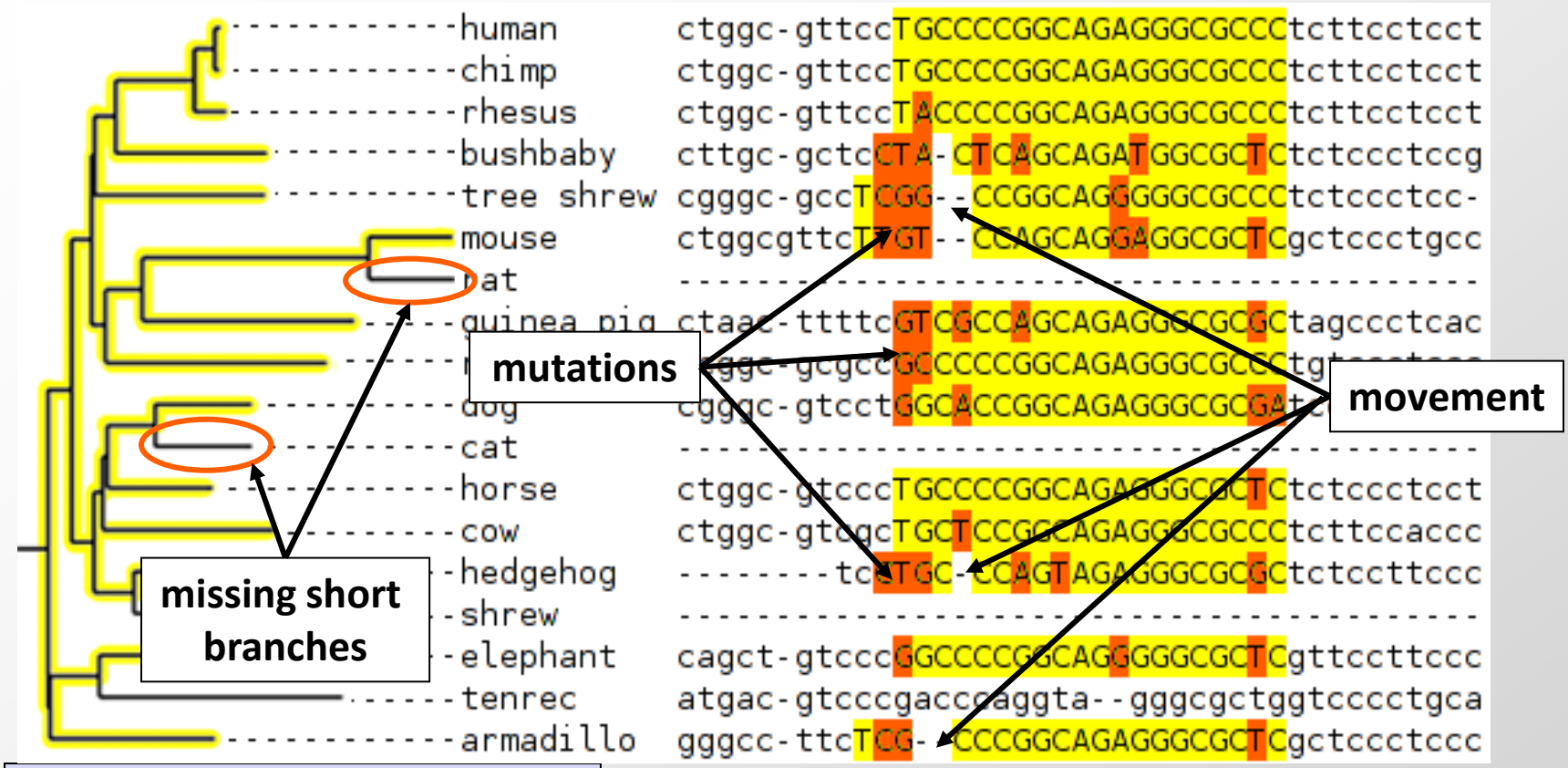
- Single genome approaches using motif clustering (e.g. Berman 2002; Schroeder 2004; Philippakis 2006)
  - Requires set of specific factors that act together
  - Miss instances of motifs that may occur alone
- Multi-genome approaches (phylogentic footprinting) (e.g. Moses 2004; Blanchette and Tompa 2002; 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%)

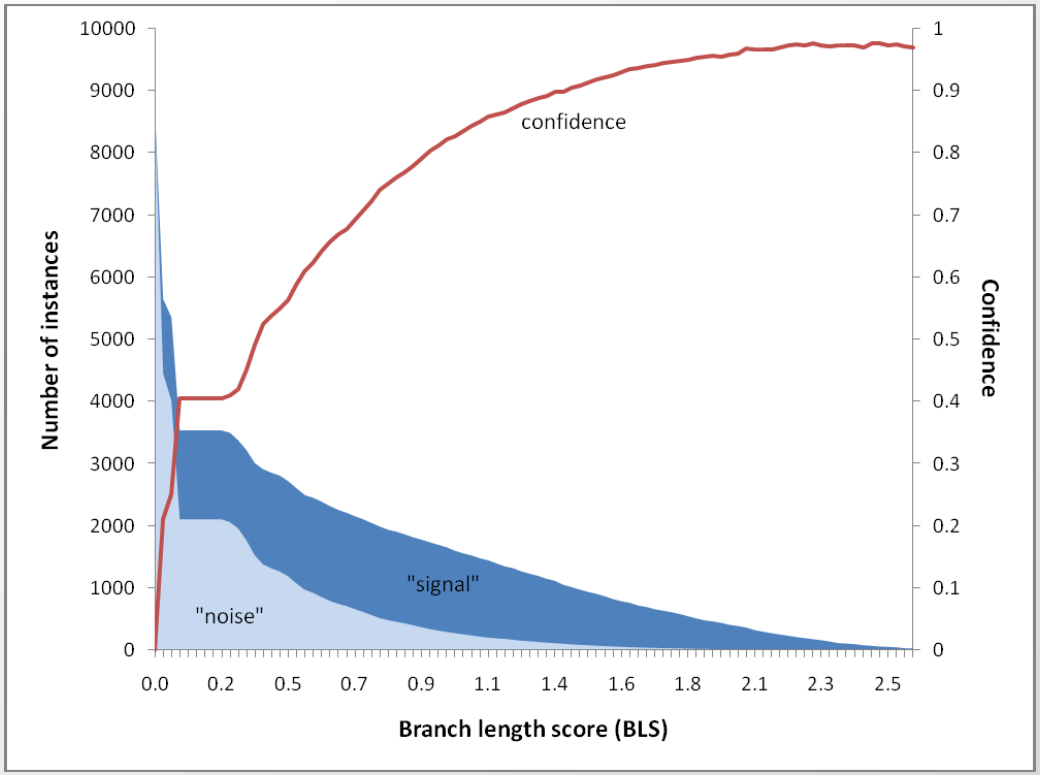
Allows for:

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

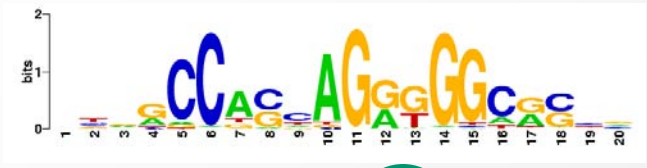
Branch Length Score ➔ Confidence



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)**

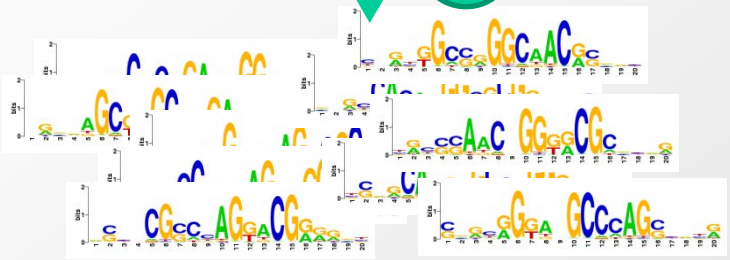
Producing control motifs

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 motif



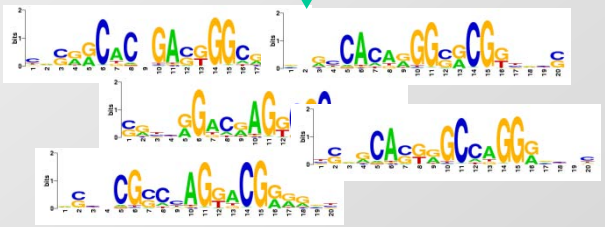
Genome sequence

Known motifs

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

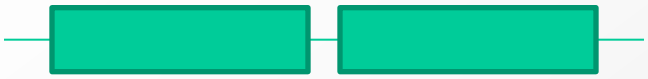
3 Sort potential control motifs based on their similarity to other known motifs

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

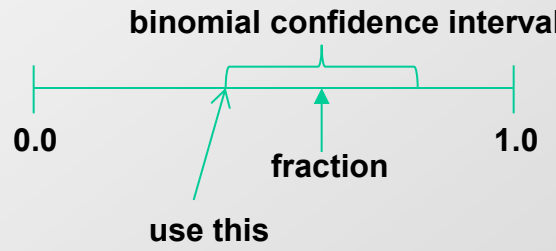
Background (e.g. Intergenic):



Foreground (e.g. TF bound):

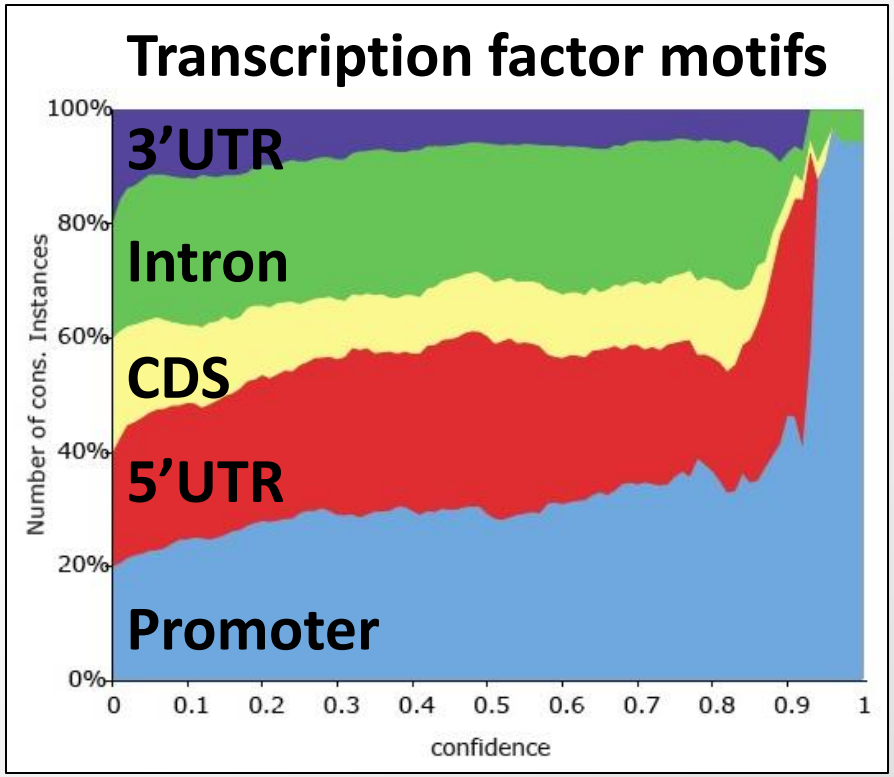
$$\frac{\#\text{ in foreground}}{\#\text{ in background}} \div \frac{\text{size of foreground}}{\text{size of background}}$$

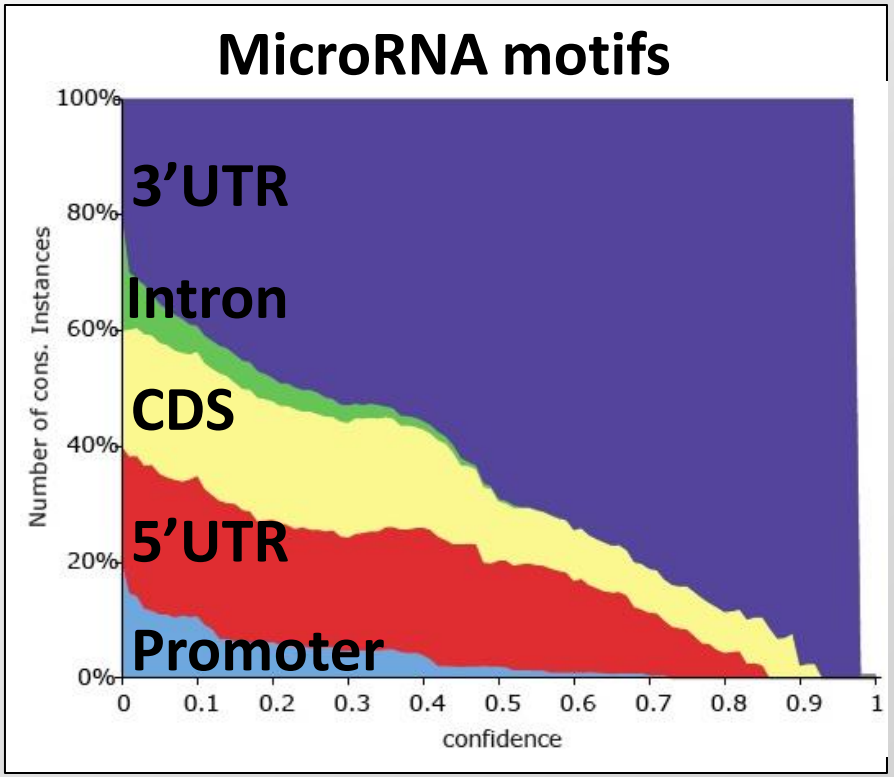
$$\frac{\#\text{ in foreground}}{\#\text{ in background}} \div \frac{\#\text{ control in foreground}}{\#\text{ control in background}}$$



- 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





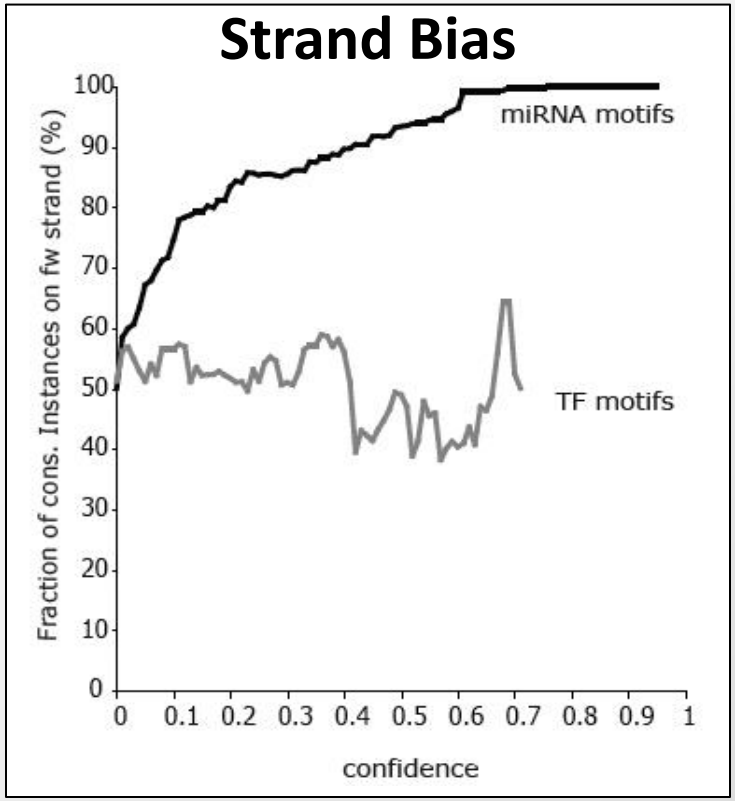
1. Confidence selects for transcription factor motif instances in promoters and miRNA motifs in 3' UTRs

Validation of discovered motif instances

Use independent experimental evidence

Look for functional biases / enrichments

Confidence selects for functional instances

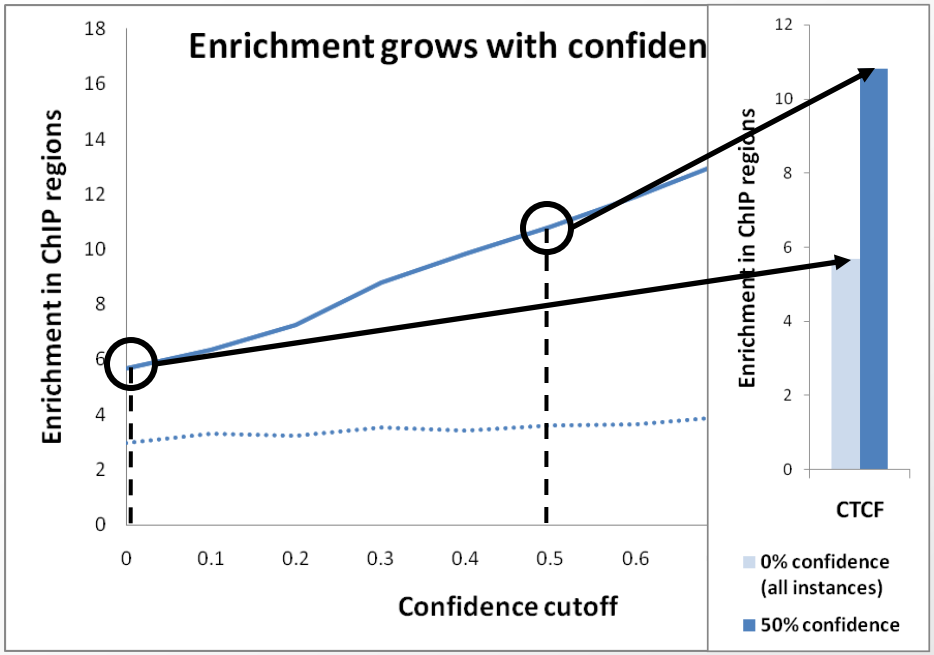


1. **Confidence selects for transcription factor motif instances in promoters and miRNA motifs in 3' UTRs**
2. **miRNA motifs are found preferentially on the plus strand, whereas no such preference is found for TF motifs**

Increased sensitivity using BLS

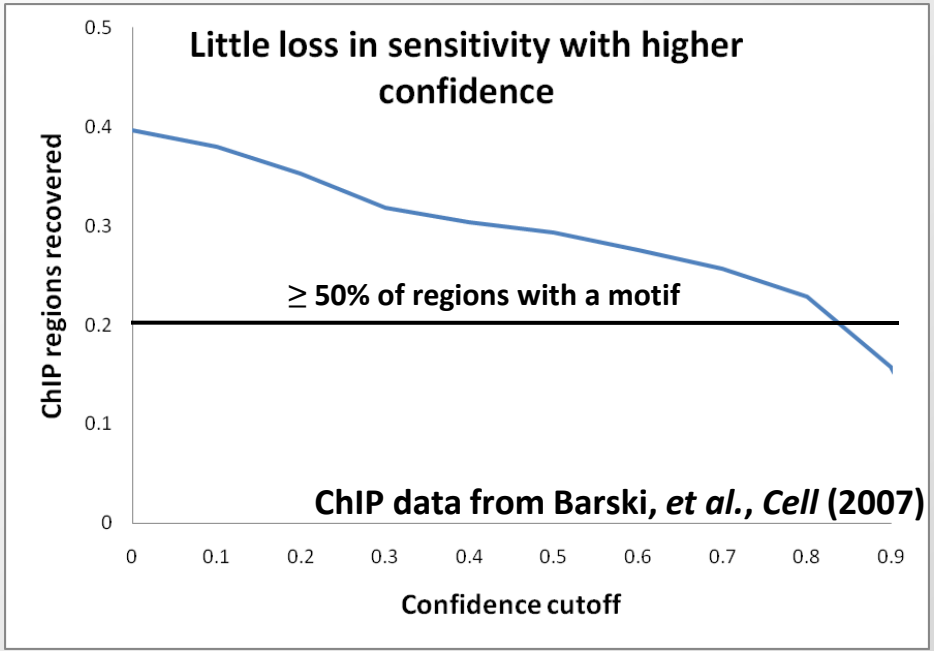
Figure 3 B removed due to copyright restrictions.
Source: Kheradpour, Pouya et al. "Reliable prediction of regulator targets using 12 Drosophila genomes." Genome Research 17, no. 12 (2007): 1919-1931.

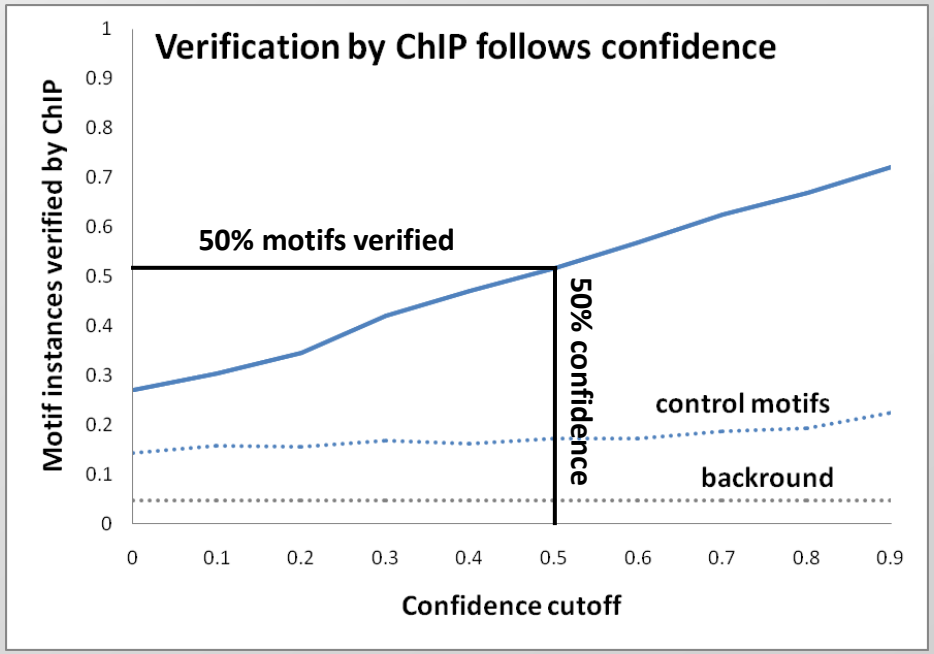
Intersection with CTCF ChIP-Seq regions



ChIP-Seq and ChIP-Chip technologies allow for identifying binding sites of a motif experimentally

- Conserved CTCF motif instances highly enriched in ChIP-Seq sites
- High enrichment does not require low sensitivity
- Many motif instances are verified

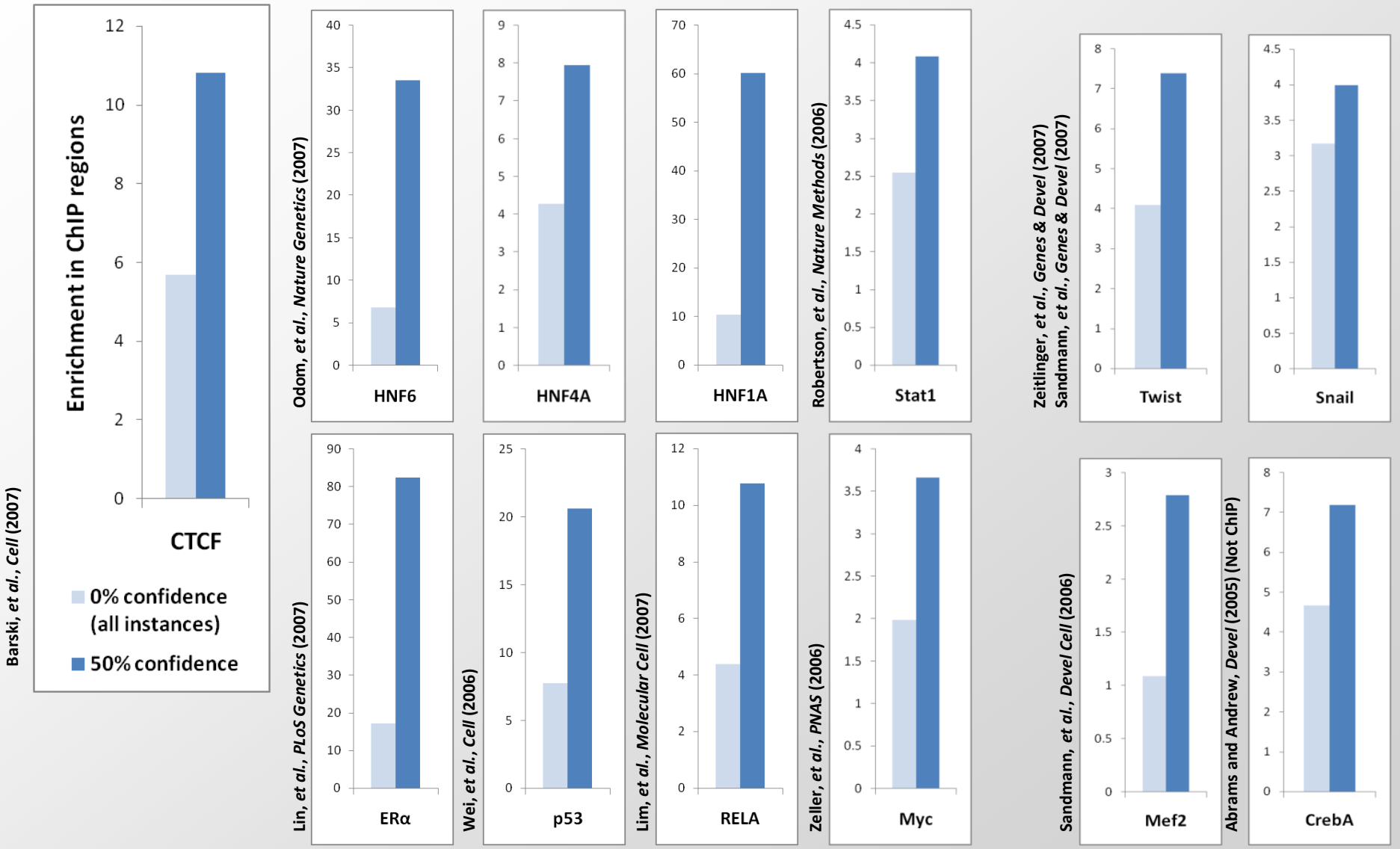




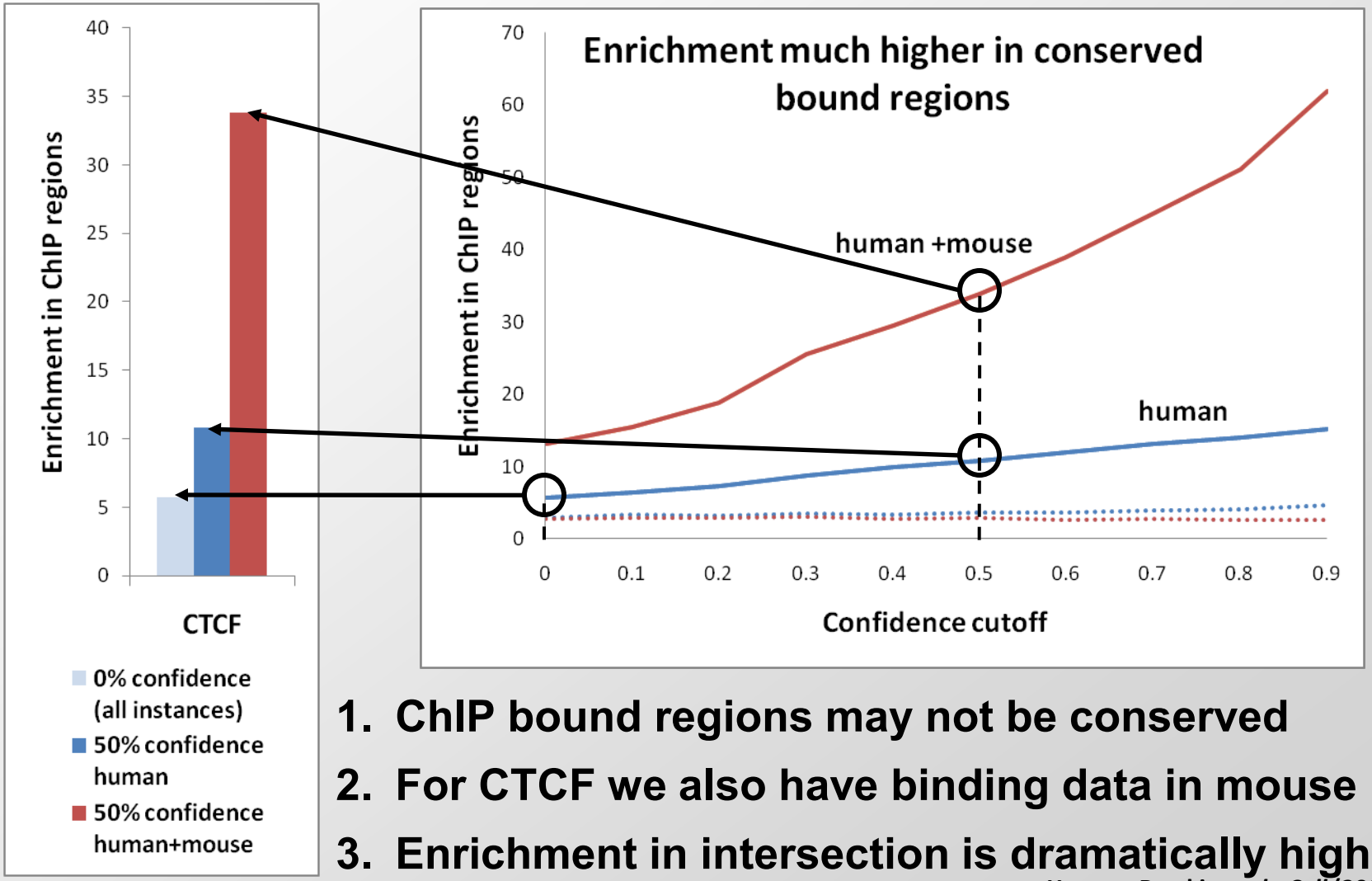
Enrichment found for many factors

Mammals

Flies



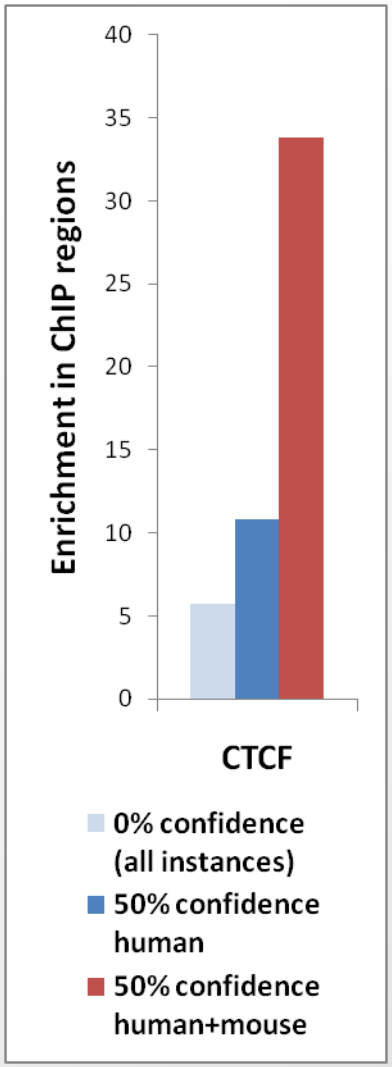
Enrichment increases in conserved bound regions



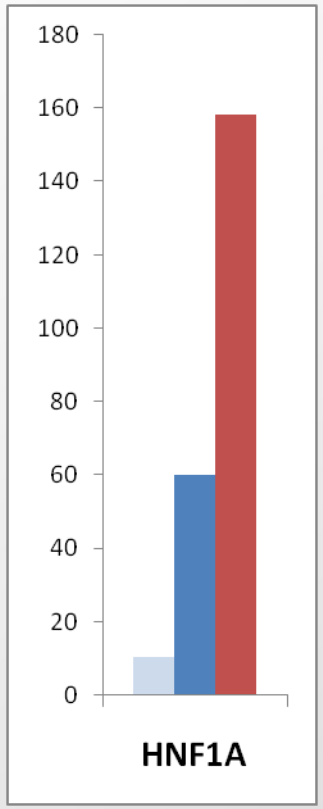
1. ChIP bound regions may not be conserved
2. For CTCF we also have binding data in mouse
3. Enrichment in intersection is dramatically higher

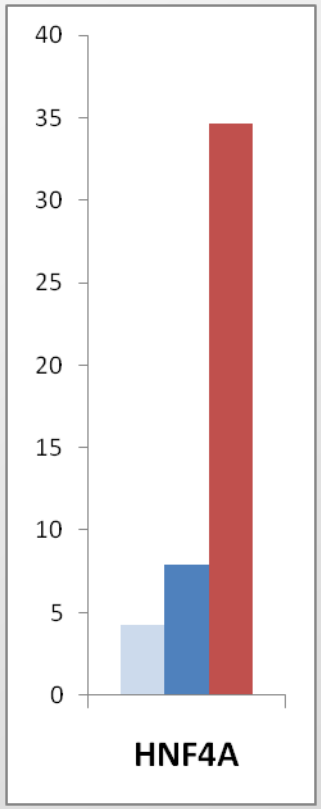
Human: Barski, *et al.*, *Cell* (2007)
Mouse: Bernstein, unpublished

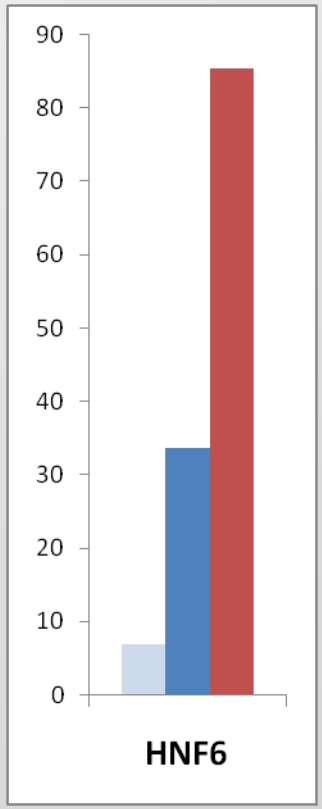
More enrichment when binding conserved



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



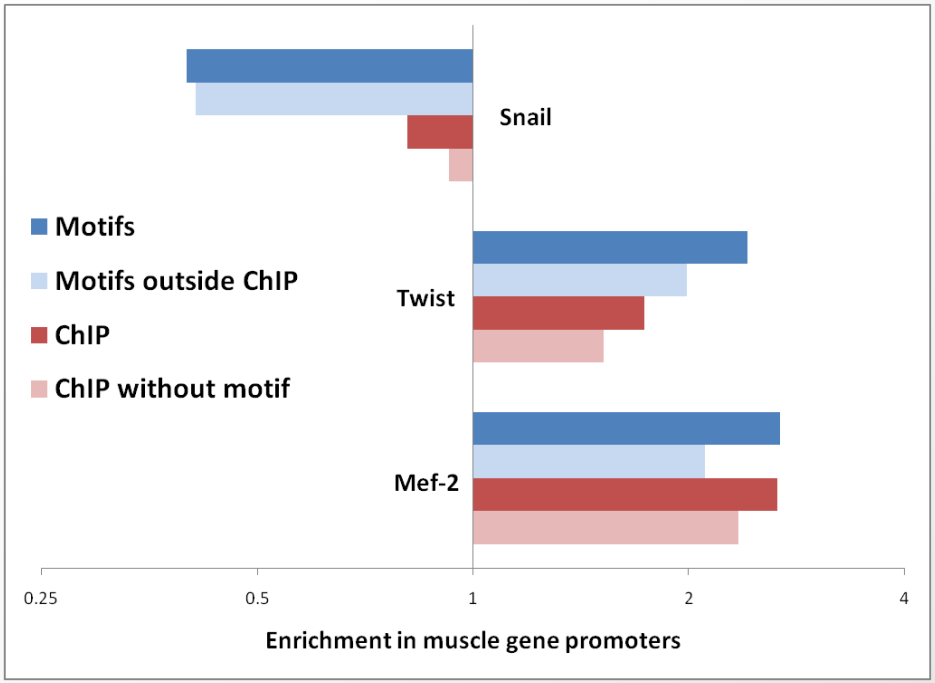


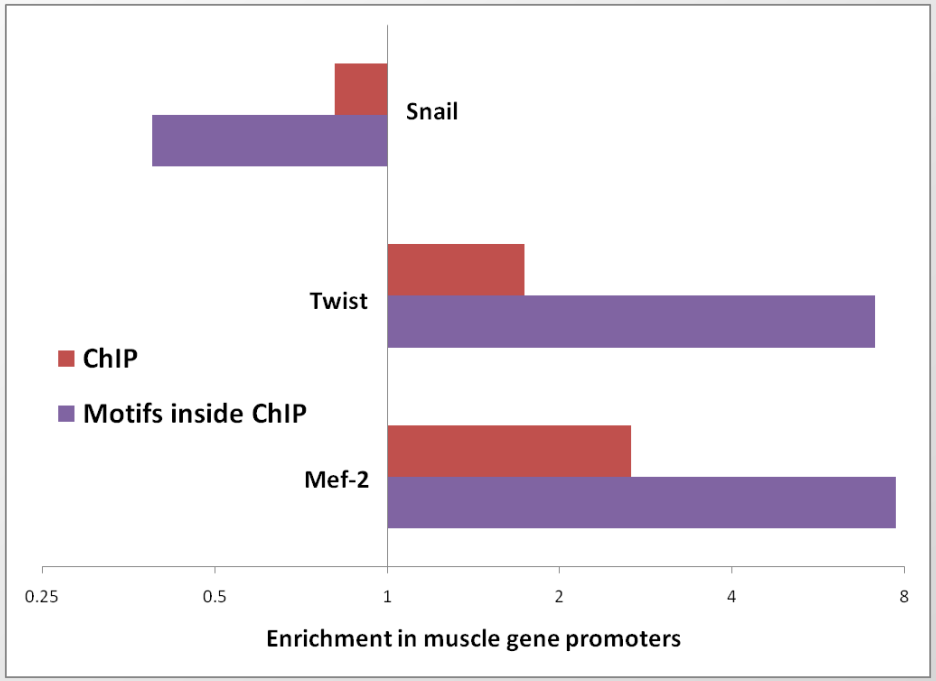


Odom, *et al.*, *Nature Genetics* (2007)

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

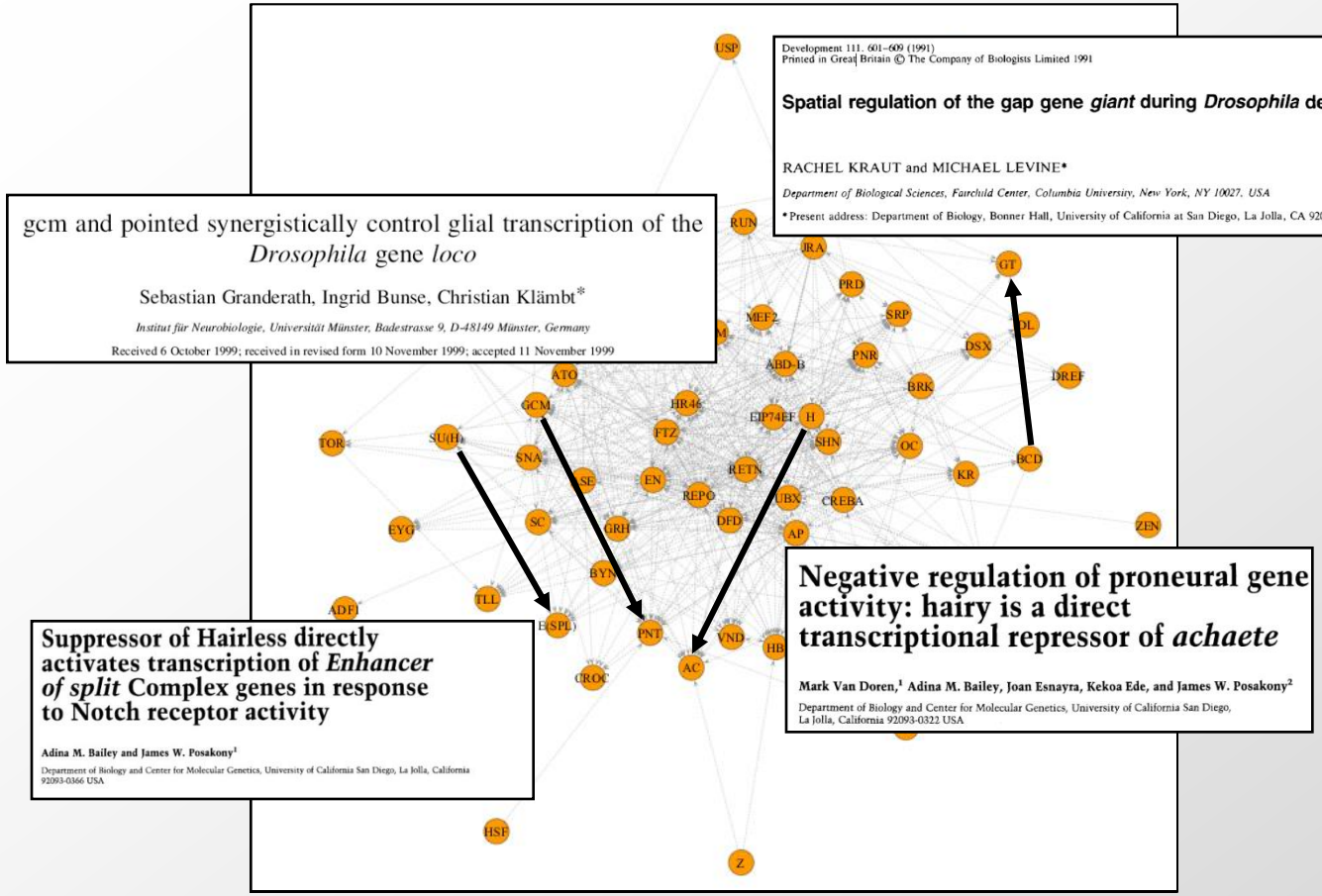




1. **Motifs at 60% confidence and ChIP have similar enrichments (depletion for the repressor Snail) in the functional promoters**
2. **Enrichments persist even when you look at non-overlapping subsets**
3. **Intersection of two regions has strongest signal**
4. **Evolutionary and experimental evidence 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)

Fly regulatory network at 60% confidence



TFs:
67 of 83 (81%)
46k instances

miRNAs:
49 of 67 (86%)
4k instances

Several connections confirmed by literature (directly or indirectly)

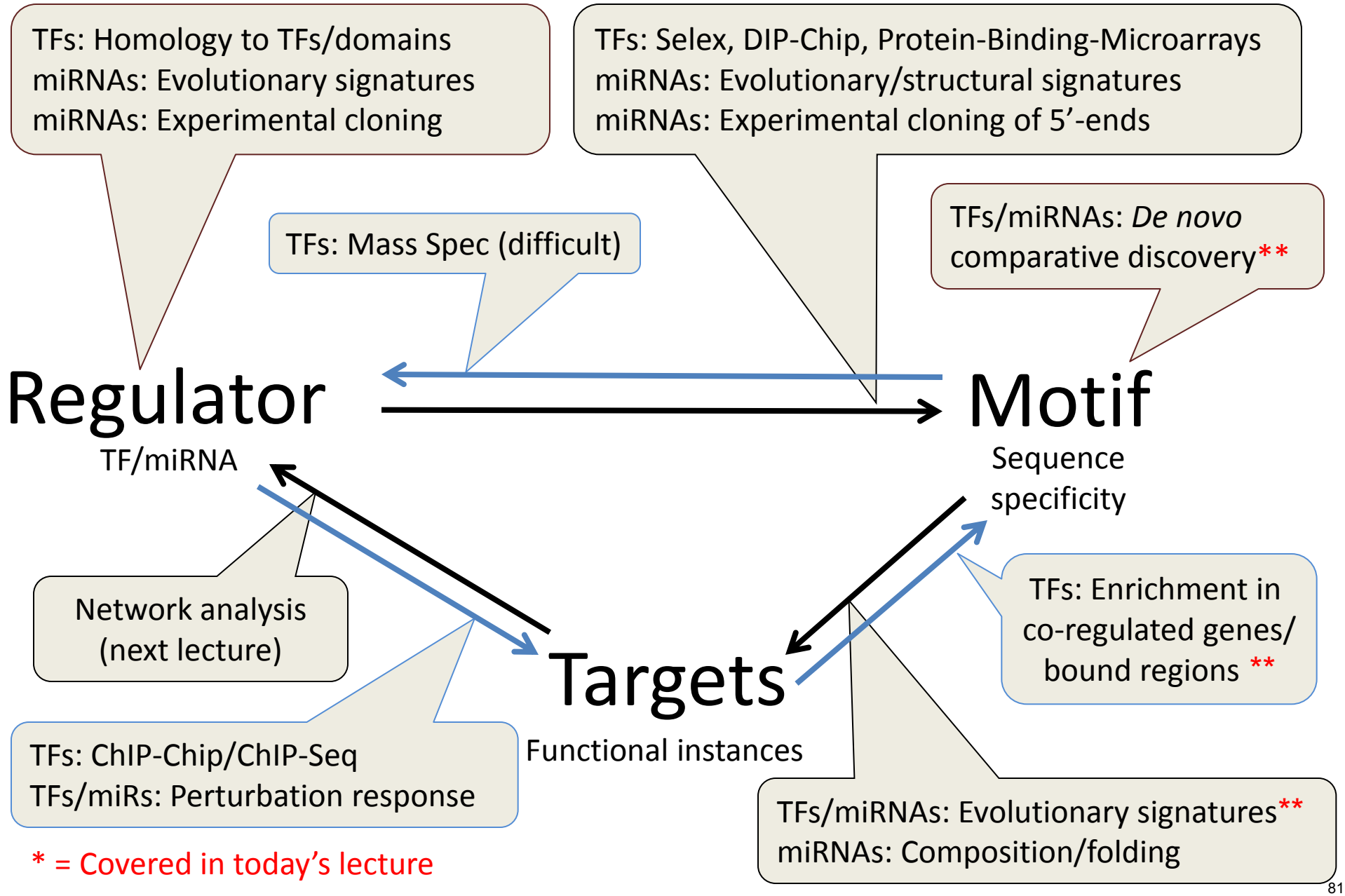
Global view of instances allows us to make network level observations:

- **46% of targets were co-expressed with their factor in at least one tissue ($P < 2 \times 10^{-3}$)**
- **TFs were more targeted by TFs ($P < 10^{-20}$) and by miRNAs ($P < 5 \times 10^{-5}$)**
- **TF in-degree associated with miRNA in-degree (high-high: $P < 10^{-4}$; low-low $P < 10^{-6}$)**

Motif discovery overview

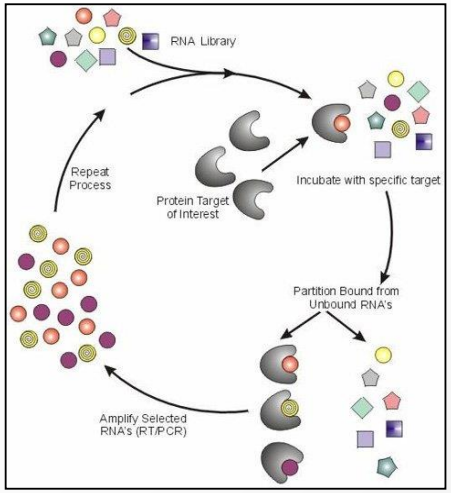
1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo
2. Expectation maximization: Motif matrix⇔positions
   - E step: Estimate motif positions Z_{ij} from motif matrix
   - M step: Find max-likelihood motif from all positions Z_{ij}
3. Gibbs 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
   - Foreground vs. background. Real vs. control motifs.

Challenges in regulatory genomics



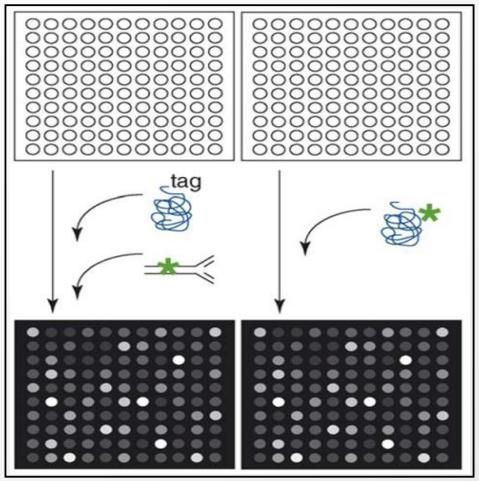
* = Covered in today's lecture

Recitation tomorrow: *in vitro* motif identification



Courtesy of the authors. Used with permission. Source: Ray, Partha, and Rebekah R. White. "Aptamers for targeted drug delivery." Pharmaceuticals 3, no. 6 (2010): 1761-1778.

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



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

- **PBMs: Protein binding microarrays**
- **SELEX: Selection-based motif identiifcation**
- **De Bruijn graphs to generate PBM probes**
- **From k-mers to motifs**
- **Gapped motifs**
- **Degenerate motifs and DNA bending (DNA shape)**
- **Relaxing independence assumptions in PWMs**

Motif discovery overview

1. Introduction to regulatory motifs / gene regulation
   - Two settings: co-regulated genes (EM,Gibbs), de novo
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
   - Foreground vs. background. Real vs. control motifs.

6.047 / 6.878 / HST.507 Computational Biology
Fall 2015