Characterizing DNA binding sites –
high throughput approaches

Biol4230 Tues, April 24, 2018
Bill Pearson wrp@virginia.edu 4-2818 Pinn 6-057

- Reviewing sites: affinity and specificity
  - representation
  - binding and specificity
  - (equilibria and competition)
- Comprehensive site identification
  - binding, consensus, and conservation
- What does complete understanding look like?
  - have DNA sequence, identify binding affinity/occupancy
  - have protein sequence of binding domain, identify DNA target

To learn more:

1. Stormo, G. D. & Zhao, Y. Determining the specificity of protein-DNA interactions. *Nat Rev Genet* **11,** 751–760 (2010).
2. Weirauch, M. T. *et al.* Evaluation of methods for modeling transcription factor sequence specificity. *Nat Biotechnol* **31,** 126–134 (2013).
3. ENCODE Project Consortium. A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol* **9,** e1001046 (2011).
4. Noyes, M. B. *et al.* Analysis of homeodomain specificities allows the family-wide prediction of preferred recognition sites. *Cell* **133,** 1277–1289 (2008).

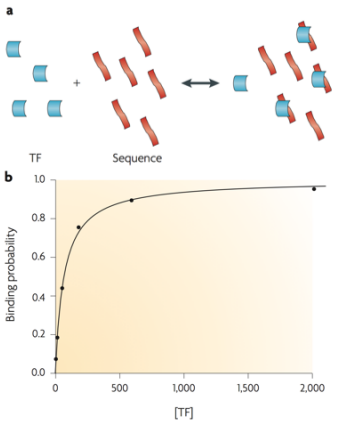
DNA-Protein interaction: binding vs specificity

Dynamic questions:

- Is DNA site *S* bound to a transcription factor *TF*
- Is the site bound frequently enough to affect transcription
- Where is most of the *TF* binding?
  - on specific DNA sites
  - on non-specific sites
  - on all sites with $K_d < 10^{-x}$
  - there are typically 10^6 more non-specific than specific sites (but are all accessible)
- what happens when the *TF* changes state?
  - higher concentration
  - more active (tighter binding) because of co-factor/modification

DNA-Protein interaction: binding vs specificity

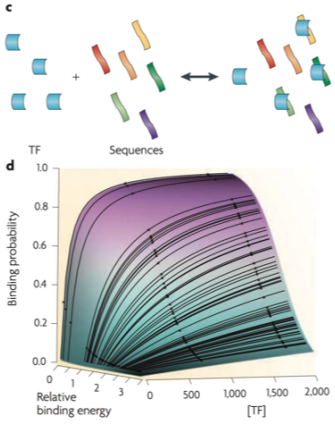
Binding



$$TF + S \underset{k_{off}}{\overset{k_{on}}{\rightleftharpoons}} TF \cdot S \quad K_d = \frac{k_{off}}{k_{on}} = \frac{[TF][S]}{[TF \cdot S]}; \Delta G^\circ = RT \ln K_d$$

$$P(\text{S bound}) = \frac{[TF \cdot S]}{[TF \cdot S] + [S]} = \frac{[TF]}{[TF] + K_d}$$

Specificity



$$\text{Spec} \equiv \sum_{S_i} \frac{K_d(S_i)}{\sum K_d(S_i)} \ln \frac{K_d(S_i)}{\langle K_d(S_i) \rangle}$$

Stormo *Nat Rev Genet* **11,** 751–760 (2010).

Terminology: Sites vs Motifs

{Sites} <-> Motif

Think restriction sites:

EcoRI: {GAATTC} <-> GAATTC
HincII {GTTAAC,GTTGAC,GTCAAC,GTCGAC} <-> GTYRAC

Transcription factor motifs should be quantitative, give different scores to different sites, reflecting differences in binding affinity.

Also: site is specific location in genome

Representations/Models of Protein-DNA binding

- Transcription factors don't bind to just one sequence
- A "Consensus sequence" is usually the preferred site, but similar sequences also bind well
- Not all variants bind equally well; some positions contribute more to the specificity than others

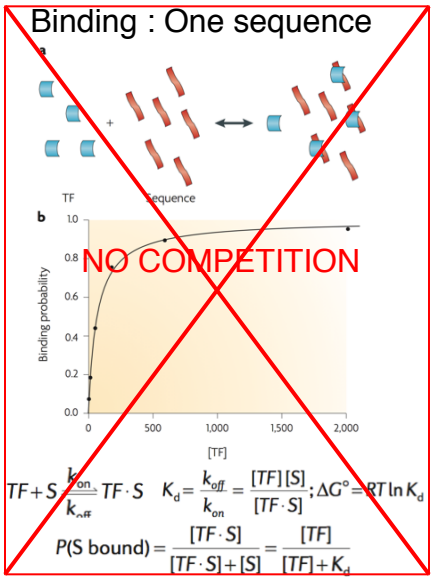
Position Weight Matrix Model (PWM, also PSSM)

log(2)-odds

A	-2.76	1.82	0.06	1.23	0.96	-2.92
C	-1.46	-3.11	-1.22	-1.00	-0.22	-2.21
G	-1.76	-5.00	-1.06	-0.67	-1.06	-3.58
T	1.67	-1.66	1.04	-1.00	-0.49	1.84

DNA-Protein interaction: binding vs specificity

Binding : One sequence



Stormo *Nat Rev Genet* **11,** 751–760 (2010).

DNA-Protein interaction: binding vs specificity

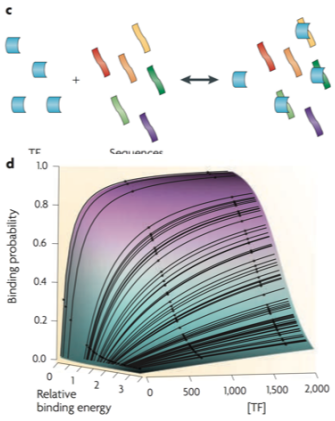
Dynamic questions:

- Is DNA site *S* bound to a transcription factor *TF*
- Is the site bound frequently enough to affect transcription
- Where is most of the *TF* binding?
  - on specific DNA sites
  - on non-specific sites
  - on all sites with $K_d < 10^{-x}$
  - there are typically 10^6 more non-specific than specific sites (but are all accessible)
- what happens when the *TF* changes state?
  - higher concentration
  - more active (tighter binding) because of co-factor/modification

DNA-Protein interaction: binding vs specificity

- Where is most of the *TF*?
  - on specific DNA sites
  - on non-specific sites
  - on all sites with $K_d < 10^{-x}$
  - there are typically 10^6 more non-specific than specific sites (but are all accessible)
- What happens when the *TF* changes state?
  - higher concentration
  - more active (tighter binding) because of co-factor/ modification

Specificity



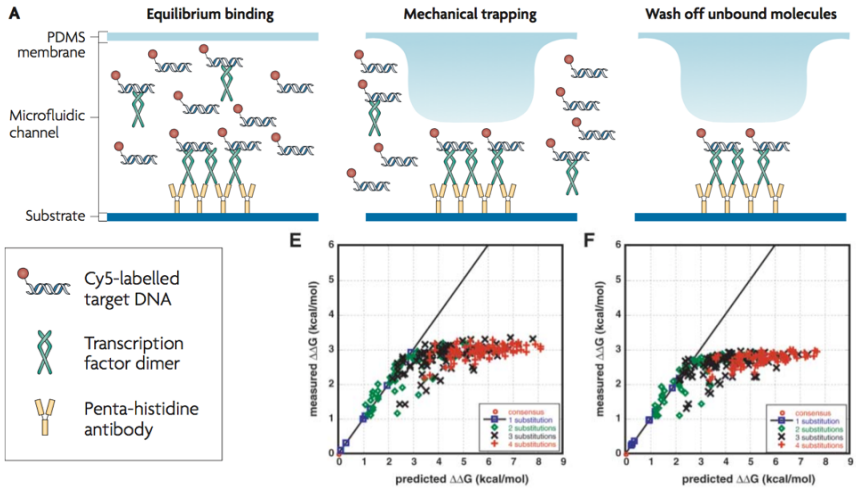
$$\mathrm{Spec} \equiv \sum_{S_i} \frac{K_d(S_i)}{\sum K_d(S_i)} \ln \frac{K_d(S_i)}{\langle K_d(S_i) \rangle}$$

Stormo *Nat Rev Genet* **11,** 751–760 (2010).

Transcription factor binding – modern approaches

- Have (functional) protein?
  - measure affinities of protein against large sets of random DNA sequences (chromatin?)
  - transform protein into cells, look at reporter genes
- Have antibody to protein?
  - ChIP-Chip/ChIP-seq – measure where the factor is on chromosomal DNA (in specific states)
  - peak width ALWAYS larger than binding sites
  - isolate surrounding DNA sequence, use consensus strategies (meme) also works with other chromatin modifications
- Have co-expressed sets of genes?
  - identify the genes, isolate sequences near promoters (enhancers?)
  - use consensus strategies (meme)

Transcription factor binding – direct measurements



Stormo *Nat Rev Genet* **11,** 751–760 (2010).

Maerkl et al. *Science* **315,** 233–237 (2007).

Transcription factor binding – direct measurements

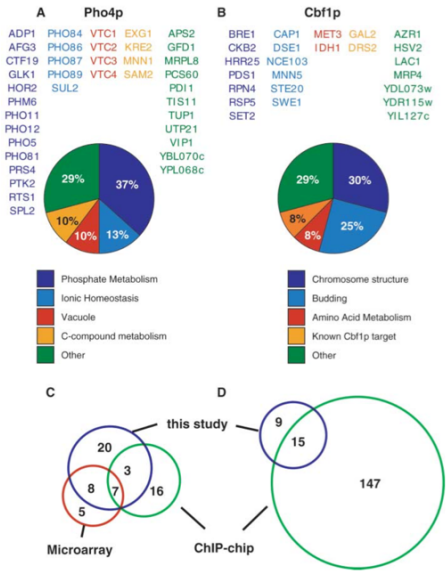
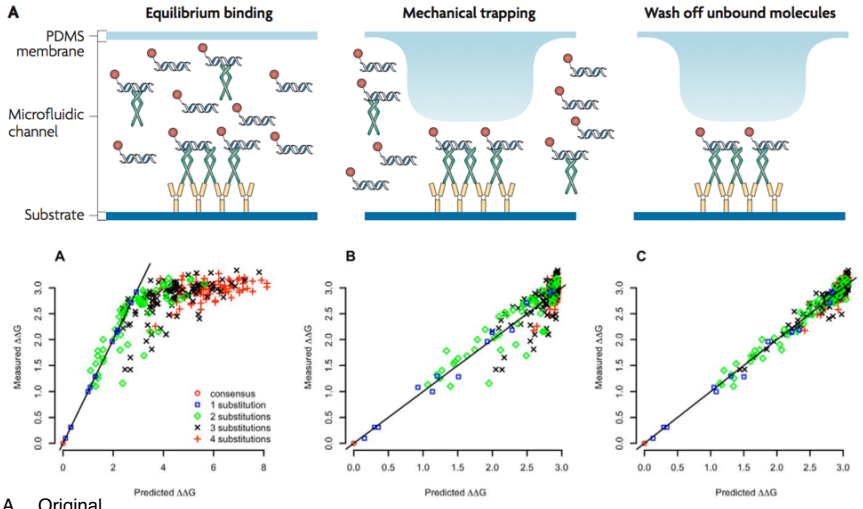


Fig. 4. In vivo function prediction for Pho4p and Cbf1p. (A and B) Genes with regulatory sequences determined to be bound by our in silico method. All genes shown here have a Pocc of above 0.2 and a sensu stricto conservation score of 25% or above. Pie charts show the functional distribution of the gene sets. (C and D) Venn diagrams comparing our predicted gene sets to gene sets determined with use of expression microarrays and ChIP-chip.

Maerkl et al. *Science* **315,** 233–237 (2007).

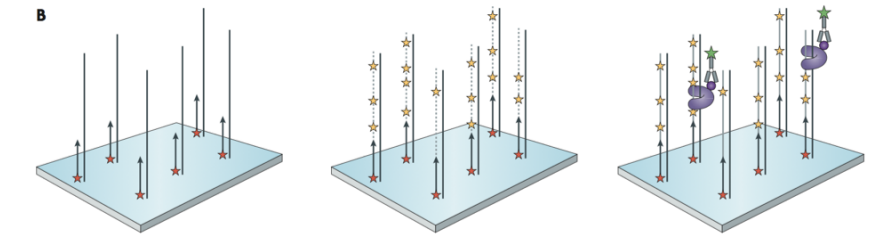
Transcription factor binding – direct measurements



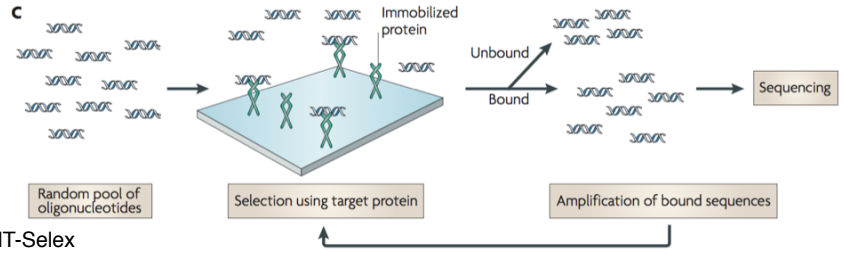
A. Original
B. BEEML (Binding Energy Est. ML) with NS energy
C. BEEML w/ NS, di-nucleotide

Zhao, Y. et al. *PLoS Comput Biol* **5,** e1000590 (2009).

Transcription factor binding – direct measurements



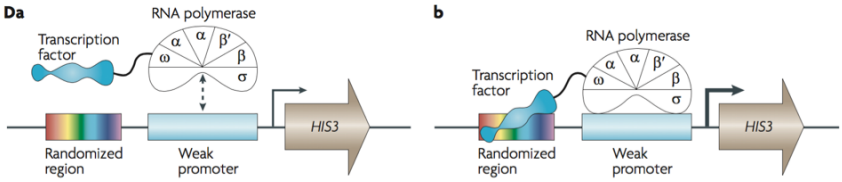
Protein Binding Microarray PBM



HT-Selex

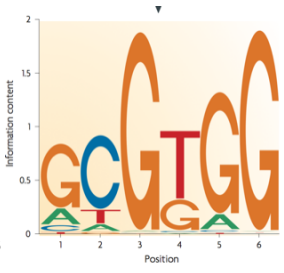
Stormo *Nat Rev Genet* **11,** 751–760 (2010).

Transcription factor binding – direct (reporter) measurements



Bacterial one-hybrid

1. Provide E. coli with essential yeast gene (HIS3) under control of a weak promoter behind a randomized binding site
2. transfect (add externally) transcription factor (sometimes linked to RNA-Pol subunit)
3. Plate out colonies expressing HIS3
4. Sequence everything that is still present
5. Most abundant (sequence seen most often) randomized region sequences grew the best, thus most binding



Stormo *Nat Rev Genet* **11,** 751–760 (2010).

High-throughput *in vitro* binding site analyses

- Can give good, quantitative models of intrinsic binding specificity
- More data alone isn't sufficient to give better models, also need good analysis methods
- Log-odds method is based on assumptions (independence) that may not be true
- Energetic models can give better descriptions
  - Non-linear relationship between binding affinity and binding probability at high TF concentration

High-throughput *in vitro* binding site analyses – does it work?

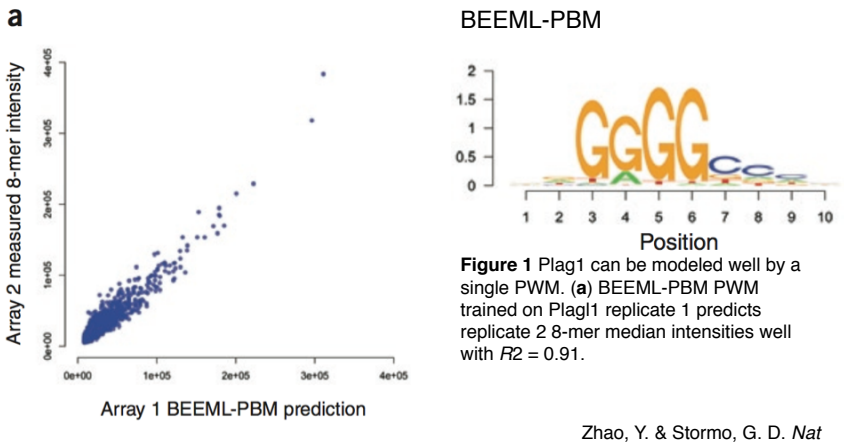
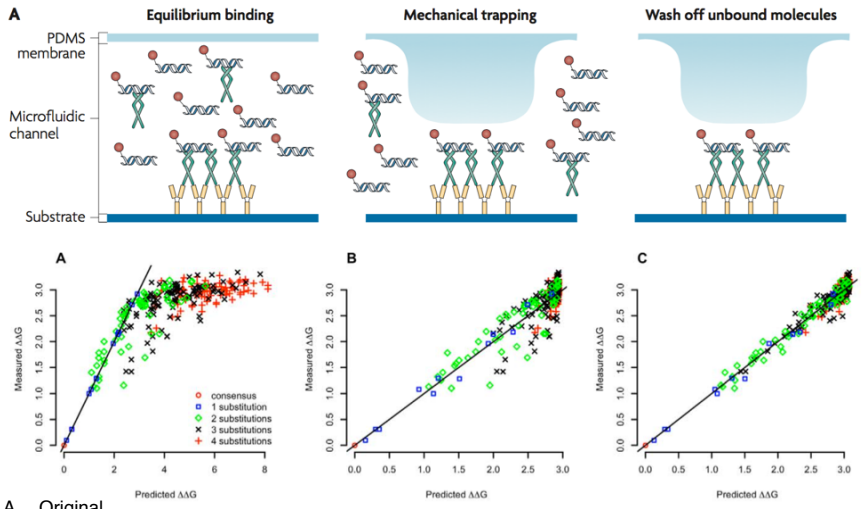


Figure 1 Plag1 can be modeled well by a single PWM. (**a**) BEEML-PBM PWM trained on Plagl1 replicate 1 predicts replicate 2 8-mer median intensities well with $R2 = 0.91$.

Zhao, Y. & Stormo, G. D. *Nat Biotechnol* **29,** 480–483 (2011).

Transcription factor binding – direct measurements



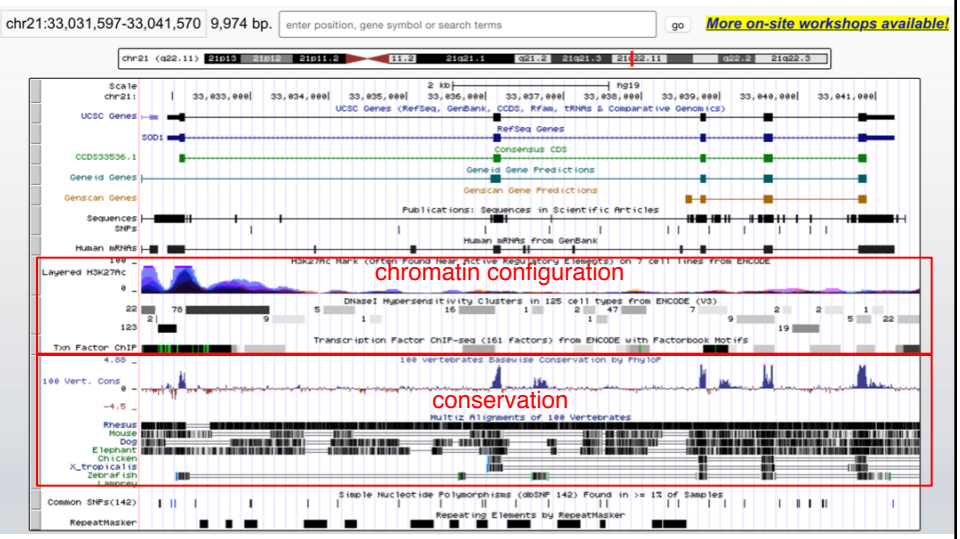
A. Original
B. BEEML with NS energy
C. BEEML w/ NS, di-nucleotide

Zhao, Y. et al. *PLoS Comput Biol* **5,** e1000590 (2009).

Transcription factor binding – modern approaches

- Have (functional) protein?
  - measure affinities of protein against large sets of random DNA sequences (chromatin?)
  - transform protein into cells, look at reporter genes
- Have antibody to protein?
  - ChIP-Chip/ChIP-seq – measure where the factor is on chromosomal DNA (in specific states)
  - peak width ALWAYS larger than binding sites
  - isolate surrounding DNA sequence, use consensus strategies (meme) also works with other chromatin modifications
- Have co-expressed sets of genes?
  - identify the genes, isolate sequences near promoters (enhancers?)
  - use consensus strategies (meme)

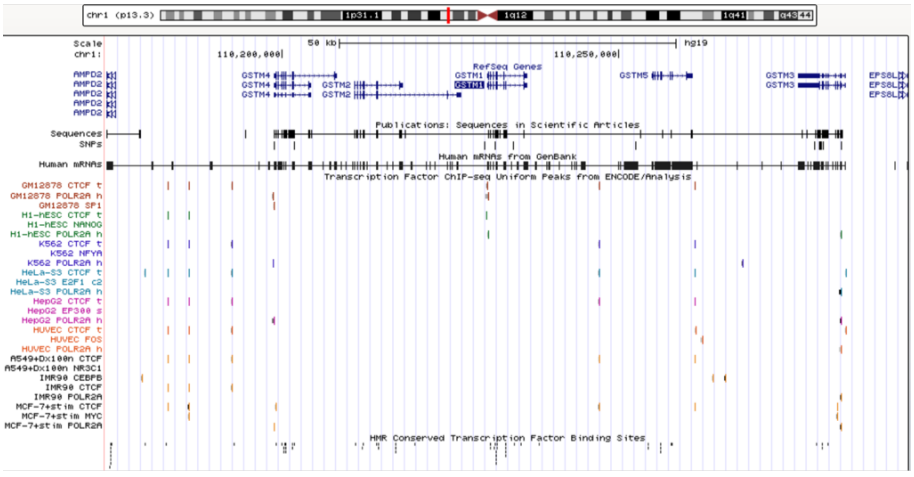
Identifying regulatory sites in chromatin



fasta.bioch.virginia.edu/biol4230

21

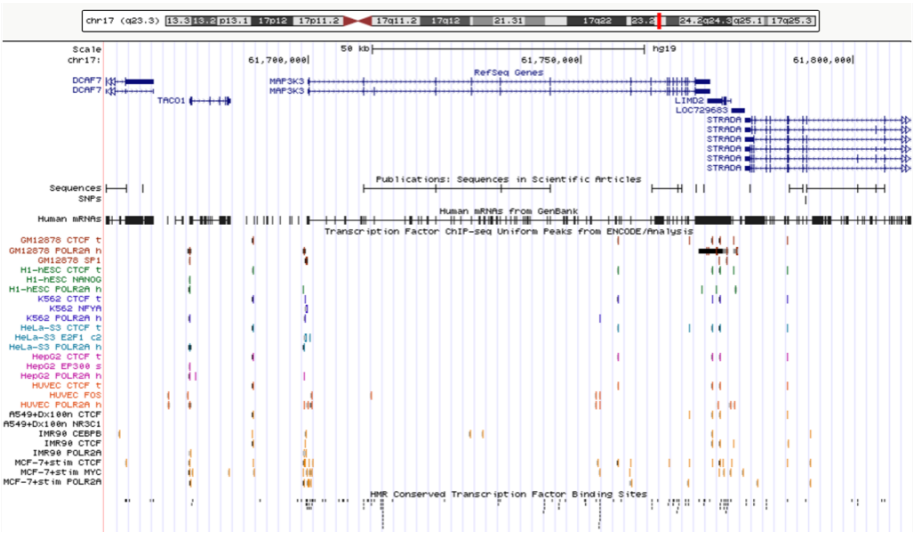
Regulatory sites in chromatin: GSTM1



fasta.bioch.virginia.edu/biol4230

22

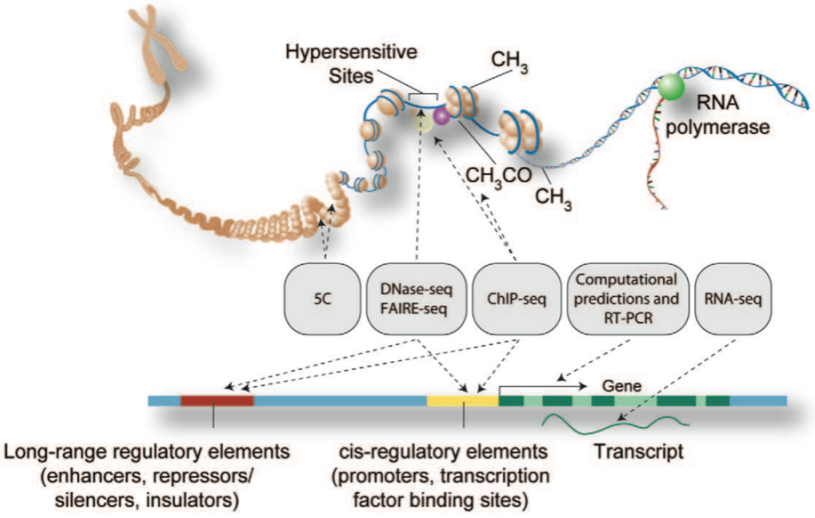
Regulatory sites in chromatin: MAP3K3



fasta.bioch.virginia.edu/biol4230

Regulatory sites in chromatin

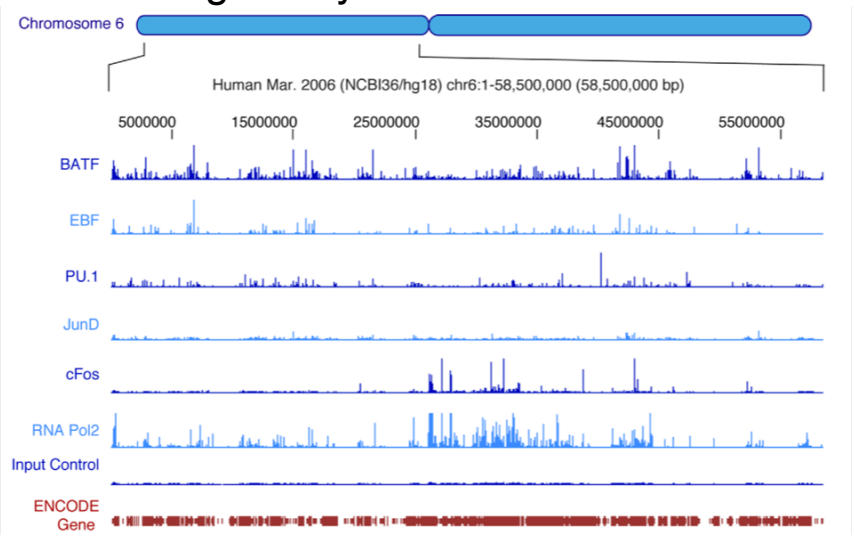
A.



ENCODE Project Consortium. *PLoS Biol* **9,** e1001046 (2011).

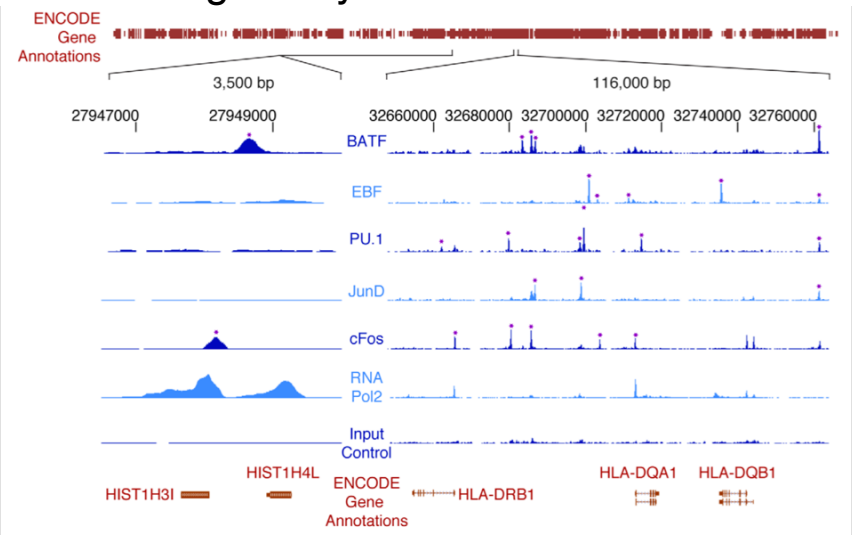
fasta.bioch.virginia.edu/biol4230

Regulatory sites in chromatin



ENCODE Project Consortium. *PLoS Biol* **9,** e1001046 (2011).

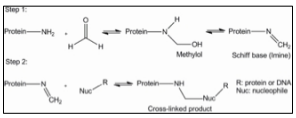
Regulatory sites in chromatin

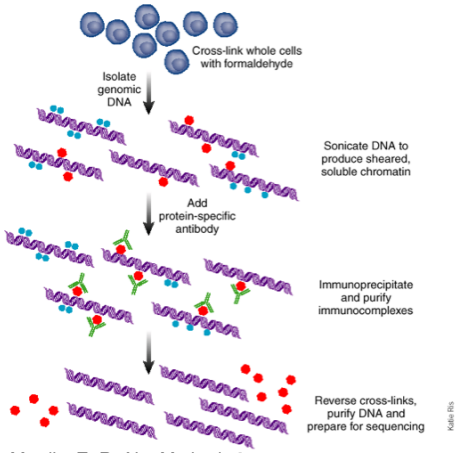


ENCODE Project Consortium. *PLoS Biol* **9,** e1001046 (2011).

Chromatin ImmunoPrecipitation - Sequencing

formaldehyde cross-linking

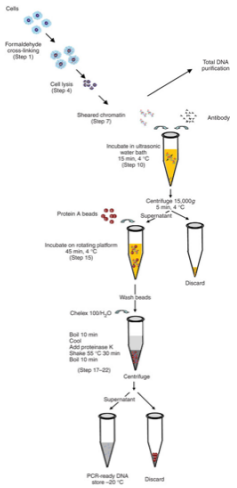




Mardis, E. R. *Nat Methods* **4,** 613–614 (2007).

fasta.bioch.virginia.edu/biol4230

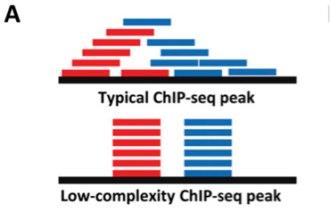
ChIP-Seq



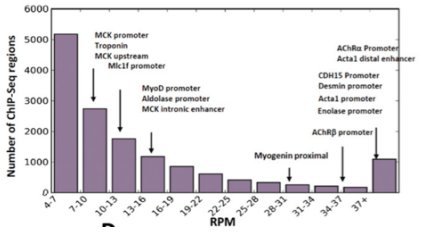
Nature Protocols **1**, 179 - 185 (2006)

27

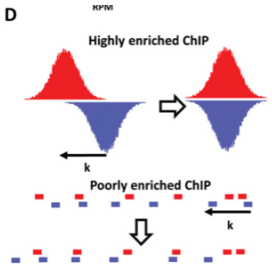
What do ChIP-Seq signals look like?]



ChIP-Seq signals should be "complex" (map across a region, with a peak)



ChIP-Seq counts on known muscle genes have a wide dynamic range



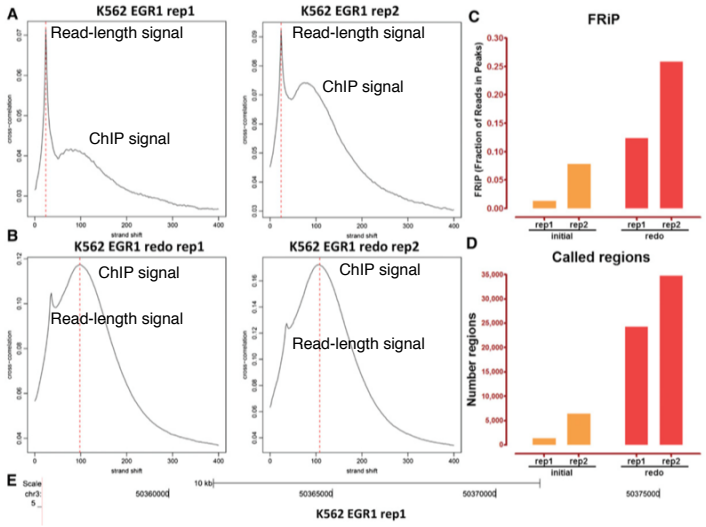
Good ChIP-seq peaks have offset reads on the two strands

Landt, S. G. *et al. Genome Res* **22,** 1813–1831 (2012).

fasta.bioch.virginia.edu/biol4230

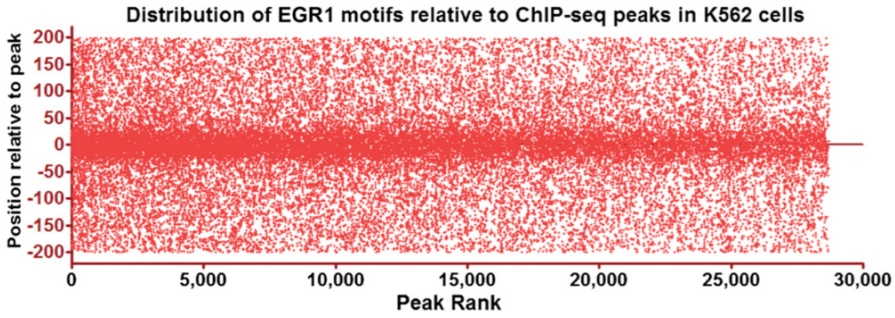
28

What do ChIP-Seq signals look like?



Landt, S. G. *et al. Genome Res* **22,** 1813–1831 (2012).

There are typically 100 – 1,000X as many motif/PWM matches as detectable binding sites



But "sites" are much more concentrated at ChIP-seq peaks
Given a set of intervals from peaks, find sites with consensus methods (meme)

Landt, S. G. *et al. Genome Res* **22,** 1813–1831 (2012).

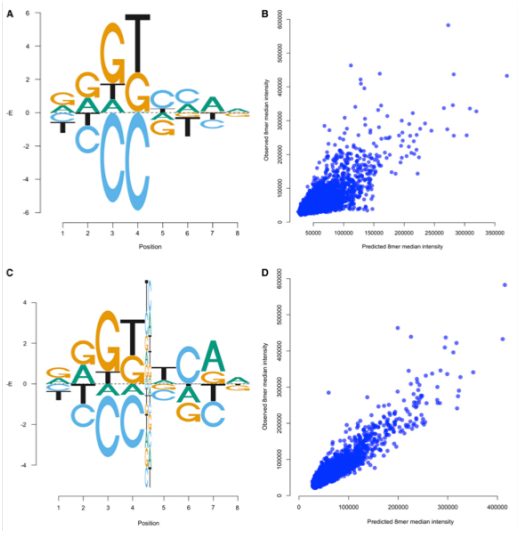
ChIP-seq summary:

- Result quality depends on antibody, immunoprecipitation, negative controls – look for reproducible peaks
- Most reads (signal) do not come from peaks
- Many more PWM sites than peaks, but sites more concentrated near peaks
- High peaks ≠ large effect
- Qualitative – enriches regions of interest

Transcription factor binding – modern approaches

- Have (functional) protein?
  - measure affinities of protein against large sets of random DNA sequences (chromatin?)
  - transform protein into cells, look at reporter genes
- Have antibody to protein?
  - ChIP-Chip/ChIP-seq – measure where the factor is on chromosomal DNA (in specific states)
  - binding sites ALWAYS larger than peak width
  - isolate surrounding DNA sequence, use consensus strategies (meme) also works with other chromatin modifications
- Have co-expressed sets of genes?
  - identify the genes, isolate sequences near promoters (enhancers?)
  - use consensus strategies (meme)

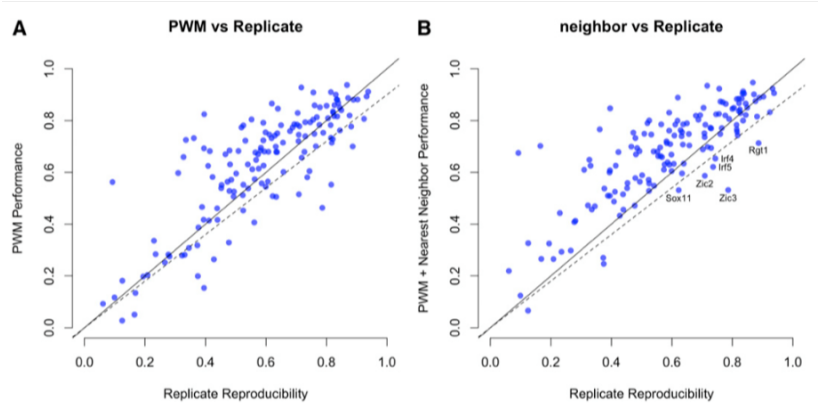
Transcription factor binding – position independence



Binding energy model including interactions makes more accurate predictions of in vitro binding specificity than the PWM for Hnf4a. (A) Graphical representation of Hnf4a binding energies estimated from PBM data under the PWM model (Supporting Information, Figure S1). Negatives of binding energy (in units of RT) are plotted on the y-axis. Energies are normalized such that the average energy at each position is 0. This energy logo is equivalent to the "affinity logo" from Foat et al. (2006). (B) Performance of model shown in A on test PBM data. (C) Binding energy model estimated from the same training data but including interaction energies between positions 4 and 5 (Figure S2). (D) Performance of the energy model including interactions on test PBM data.

Zhao, Y., et al. *Genetics* **191,** 781–790 (2012).

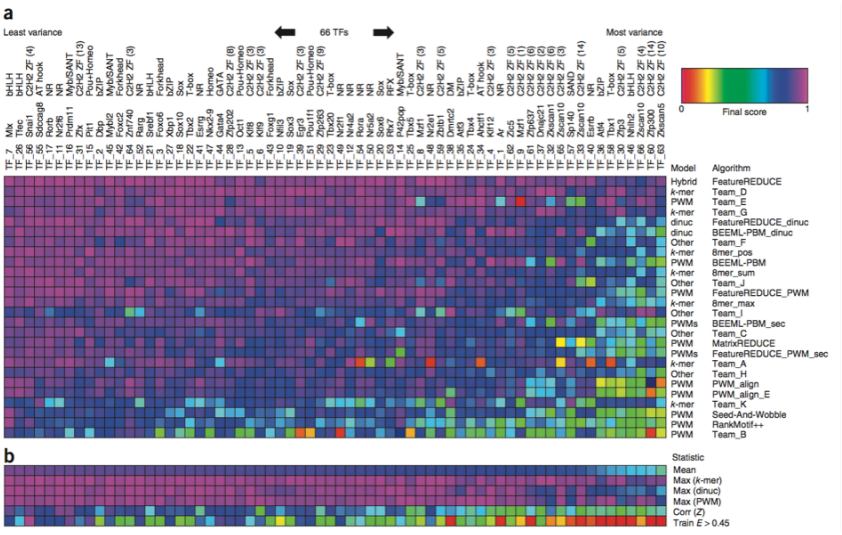
Transcription factor binding – position independence



(From abstract): We find that the specificity of most TFs is well fit with the simple PWM model, but in some cases more complex models are required. We introduce a binding energy model (BEM) that can include energy parameters for nonindependent contributions to binding affinity. We show that in most cases where a PWM is not sufficient, a BEM that includes energy parameters for adjacent dinucleotide contributions models the specificity very well.

Zhao, Y., et al. *Genetics* **191,** 781–790 (2012).

How well do methods work?



Weirauch, M. T. *et al.*. *Nat Biotechnol* **31,** 126–134 (2013).

fasta.bioch.virginia.edu/biol4230

35

How well do methods work?

Algorithm	Mean	Esrrb	Gata4	Tbx20	Tbx5	Zfx	Gal4	Phd1	Rap1	Reb1
ChIPmunk	0.741	0.718	0.655	0.809	0.776	0.780	0.523	0.792	0.841	0.780
FeatureREDUCE_PWM	0.725	0.684	0.726	0.631	0.679	0.753	0.785	0.723	0.770	0.780
FeatureREDUCE_dinuc	0.721	0.685	0.729	0.624	0.679	0.761	0.794	0.731	0.714	0.780
BEEML-PBM	0.703	0.688	0.726	0.663	0.699	0.798	0.761	0.732	0.849	0.416
PWM_align_E	0.703	0.695	0.700	0.620	0.483	0.765	0.842	0.669	0.785	0.770
PWM_align	0.695	0.698	0.702	0.618	0.473	0.763	0.769	0.680	0.788	0.770
Seed-And-Wobble	0.693	0.675	0.633	0.609	0.558	0.729	0.749	0.712	0.804	0.774
FeatureREDUCE	0.681	0.625	0.725	0.529	0.683	0.805	0.781	0.727	0.703	0.558
MEME-ChIP	0.679	0.694	0.692	0.791	0.595	0.455	0.596	0.672	0.831	0.791
BEEML-PBM_sec	0.678	0.703	0.736	0.661	0.675	0.793	0.761	0.552	0.726	0.495
Team_E	0.663	0.577	0.714	0.636	0.599	0.789	N/A	N/A	N/A	N/A
FeatureREDUCE_sec	0.653	0.699	0.637	0.627	0.582	0.704	0.733	0.720	0.611	0.564
8mer_sum_hi	0.637	0.633	0.717	0.527	0.533	0.755	0.721	0.607	0.594	0.651
RankMotif++	0.630	0.511	0.666	0.609	0.423	0.669	0.749	0.733	0.680	0.633
MatrixREDUCE	0.628	0.347	0.659	0.568	0.572	0.791	0.759	0.730	0.454	0.775
BEEML-PBM_dinuc	0.610	0.677	0.744	0.573	0.716	0.803	0.382	0.731	0.411	0.453
Team_D	0.598	0.580	0.670	0.468	0.470	0.721	0.623	0.658	0.614	0.580
8mer_sum	0.567	0.496	0.603	0.415	0.425	0.717	0.631	0.675	0.572	0.575

< 0.50 AUROC 0.85

In vitro defined PWM's accurately predict in vivo binding

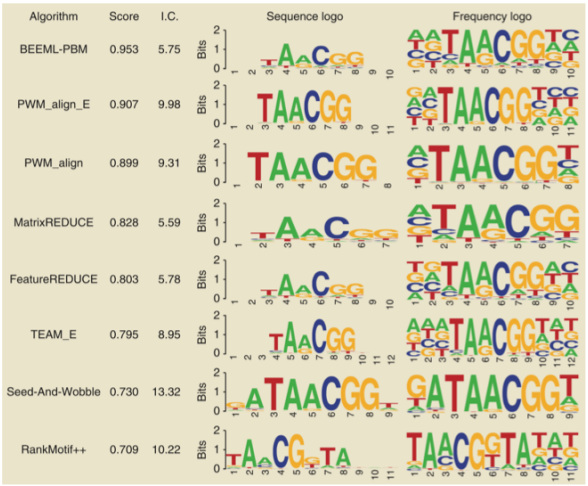
Weirauch, M. T. *et al.*. *Nat Biotechnol* **31,** 126–134 (2013).

fasta.bioch.virginia.edu/biol4230

36

Information content vs accuracy

Figure 4 Characteristics of Klf9 motifs produced by the eight PWM-based algorithms evaluated in this study. The algorithms are ranked top to bottom in order of the overall score of their PWM for this TF in our evaluation scheme. Two popular visualization methods of the PWMs produced by each algorithm are depicted. On the left are traditional sequence logos39,40, which display the information content of each nucleotide at each position; the total information content (I.C.) of the PWM is given to the left of this logo. On the right are frequency logos, in which the height of each nucleotide corresponds to its frequency of occurrence at the given position40.



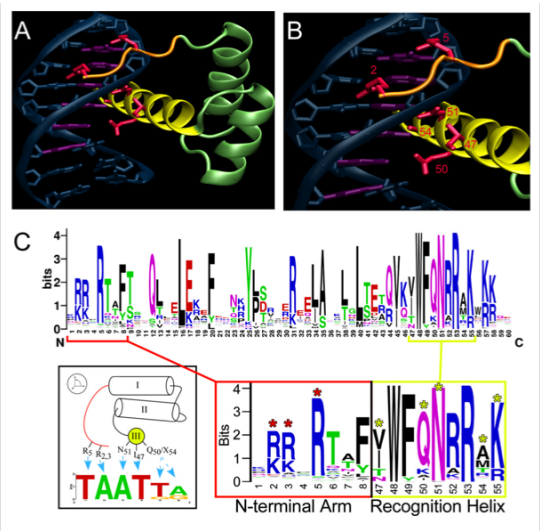
Weirauch, M. T. *et al.*. *Nat Biotechnol* **31,** 126–134 (2013).

DNA-Protein interaction: what is *complete* understanding?

1. Understand the DNA binding site
2. Identify the amino-acids that *read* the DNA sequence
3. understand how changes in the protein change the DNA binding site
4. *predict* DNA binding site preferences from protein sequence (engineering)

Noyes, M. B. *et al. Cell* **133,** 1277–1289 (2008).

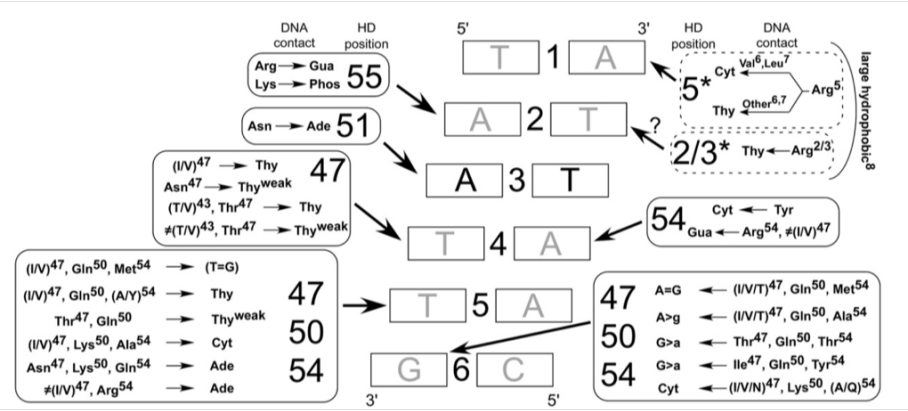
DNA-Protein interaction (homeobox): what is *complete* understanding?



Noyes, M. B. *et al. Cell* **133,** 1277–1289 (2008).

fasta.bioch.virginia.edu/biol4230

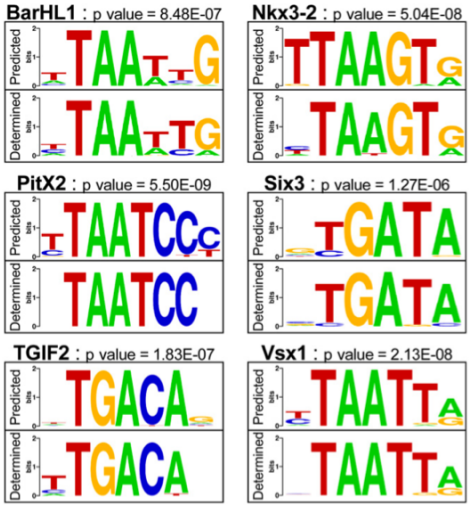
DNA-Protein interaction (homeobox): what is *complete* understanding?



Noyes, M. B. *et al. Cell* **133,** 1277–1289 (2008).

fasta.bioch.virginia.edu/biol4230

DNA-Protein interaction (homeobox): what is *complete* understanding?



Comparison of the Predicted and Determined Recognition Motifs for Six Human Homeodomains: The specificities of the human factors were determined with the B1H system. In each case, the "determined" compares favorably with the "predicted" motif generated with our algorithm.

For the homeobox family, it is possible to predict the DNA binding site from the amino-acid sequence

Noyes, M. B. *et al. Cell* **133,** 1277–1289 (2008).

Characterizing DNA binding sites – high throughput approaches

- Affinity and specificity
  - transcription factors have higher affinity for their specific binding site than non-specific sites
  - but there are 10^6 – 10^7 more non-specific sites
  - ratios of specific/non-specific binding are < 10^6
  - a large fraction of transcription factor binging is non-specific
- High-throughput in vitro methods provide accurate binding constants
  - PWM (independent positions) usually provides accurate model of binding
  - for a fraction of sites, a binding energy term that includes non-independence helps
- ChIP-Seq provides large lists of binding sites
  - but small fraction of motif matches
- For large, highly studied families (homeobox), the amino-acid recognition code is understood