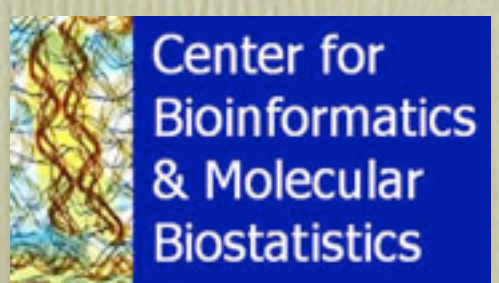
Re-Cracking the Nucleosome Positioning Code

Mark Segal
Center for Bioinformatics & Molecular Biostatistics
UCSF Division of Biostatistics
http://www.biostat.ucsf.edu/cbmb



Outline

- The Second Genetic Code
  - Nucleosome Occupancy (Positioning)
- Basis of the Code
  - Periodicities and the Spectral Envelope
- Performance of the Code
- Conclusions

A celebrated comp bio paper:

From the ISCB newsletter

Segal and colleagues published a study in *Nature* (442, 772-778, 2006) hypothesizing that the instructions for wrapping DNA around nucleosomes are contained in the DNA itself, using a statistical computational model to predict exactly how that is done, and completing the proof by verifying the predictions with experiments in yeast.

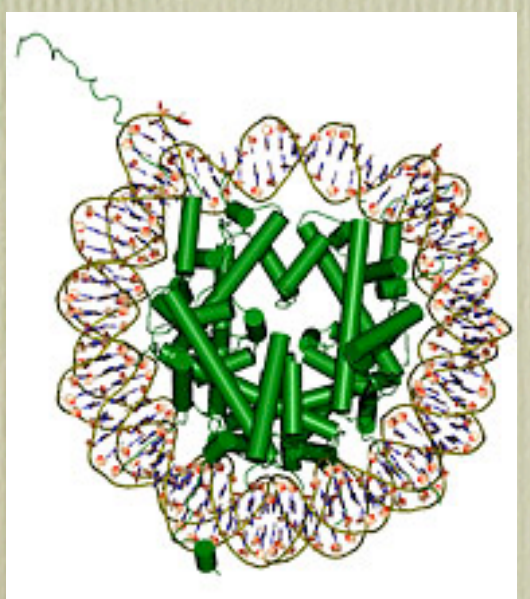
"This important paper brought Segal a lot of attention," says Lengauer. "It was featured in *Nature's* 'News and Views' section in an article by Tim Richmond, and the work was also described in a 'Making the Paper' section. And it made The New York Times on July 25, 2006.

The New York Times

Scientists Say They've Found a Code Beyond Genetics in DNA

By NICHOLAS WADE
Published: July 25, 2006

Researchers believe they have found a **second** code in DNA in addition to the genetic code. The genetic code specifies all the proteins that a cell makes. The second code, superimposed on the first, sets the placement of the nucleosomes, miniature protein spools around which the DNA is looped. The spools both protect and control access to the DNA itself.

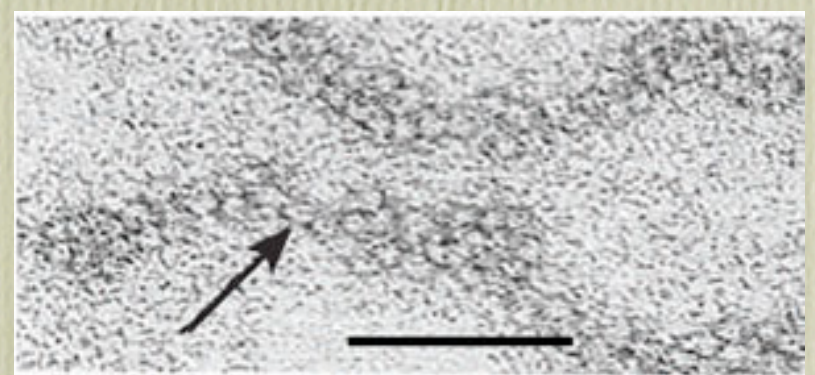


In a living cell, the DNA double helix wraps around a nucleosome, and binds to some of its proteins, known as histones.

News and Views

Nature **442**, 750-752 (17 August 2006)

Genomics: Predictable packaging



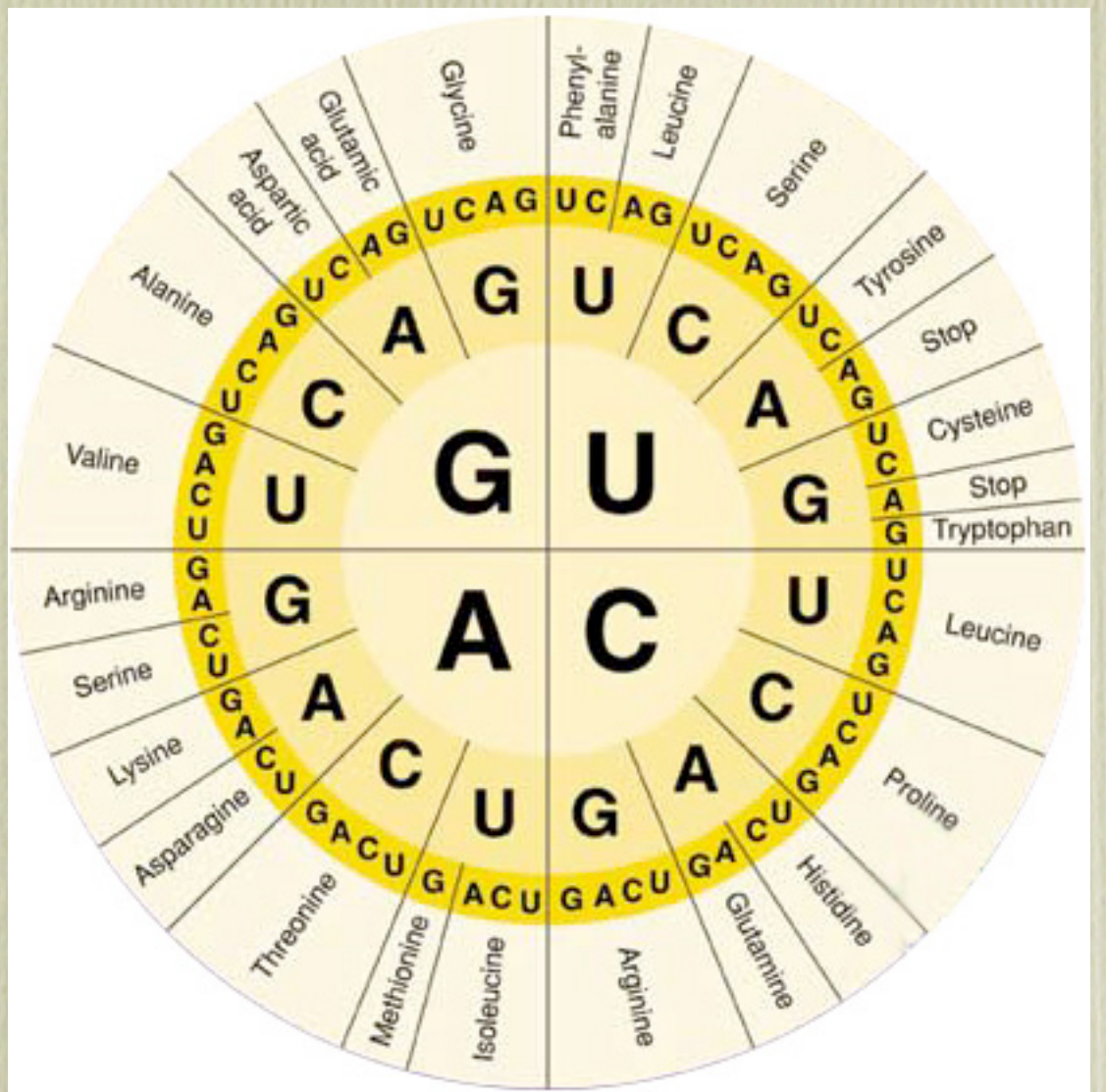
Chromatin fibres showing nucleosomes (small circles). Regions containing a single nucleosome (arrow) may represent transitions between different packing arrangements. Segal *et al*.[1] have shown that the locations of the nucleosomes may be determined directly by the DNA sequence. Scale bar, 0.1 um.

TITLE: A genomic code for nucleosome positioning

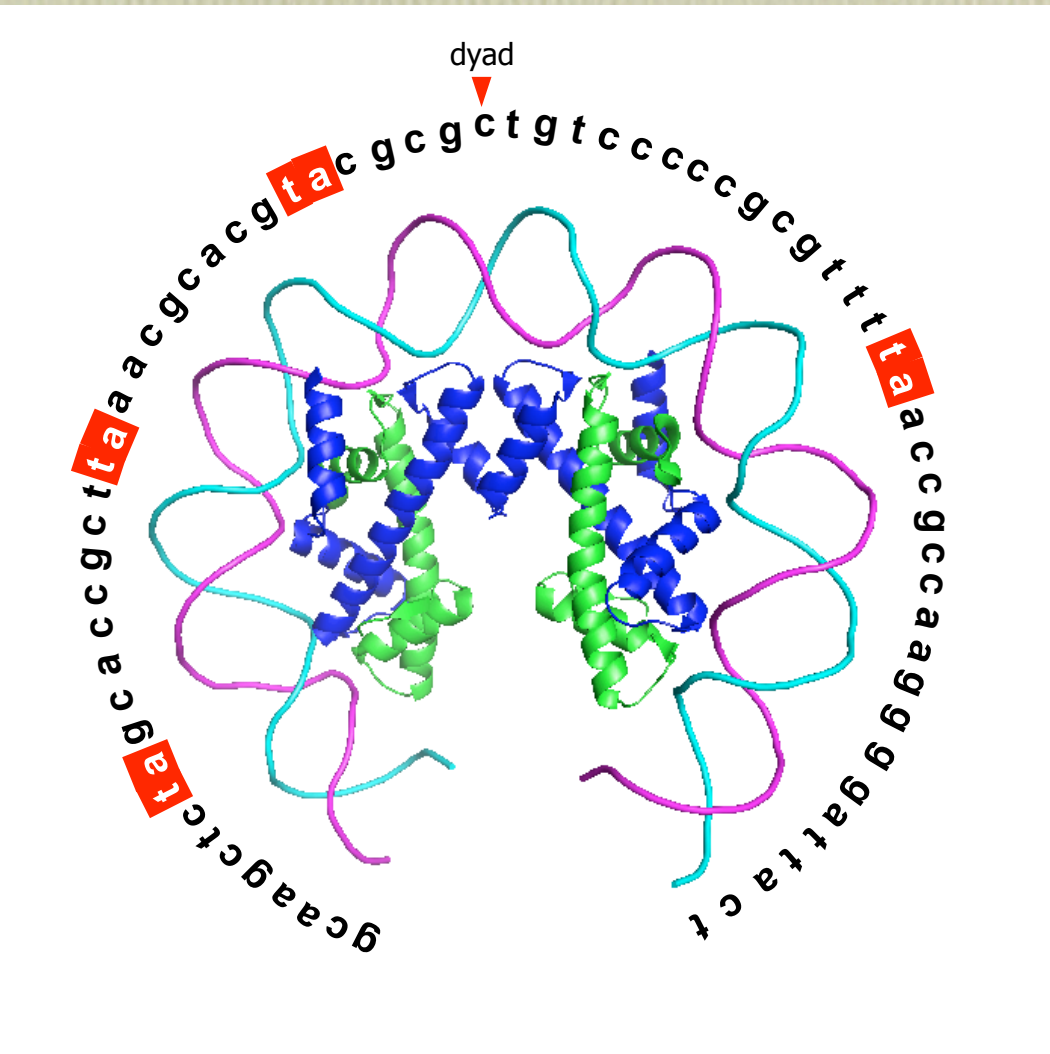
ABSTRACT: Eukaryotic genomes are packaged into nucleosome particles that occlude the DNA from interacting with most DNA binding proteins. Nucleosomes have higher affinity for particular DNA sequences, reflecting the ability of the sequence to bend sharply, as required by the nucleosome structure. However, it is not known whether these sequence preferences have a significant influence on nucleosome position in vivo, and thus regulate the access of other proteins to DNA. Here we isolated nucleosome-bound sequences at high resolution from yeast and used these sequences in a new computational approach to construct and validate experimentally a nucleosome-DNA interaction model, and to predict the genome-wide organization of nucleosomes. Our results demonstrate that genomes encode an intrinsic nucleosome organization and that this intrinsic organization can explain ~50% of the in vivo nucleosome positions. This nucleosome positioning code may facilitate specific chromosome functions including transcription factor binding, transcription initiation, and even remodelling of the nucleosomes themselves.

FIRST PARAGRAPH: Eukaryotic genomic DNA exists as highly compacted nucleosome arrays called chromatin. Each nucleosome contains a 147-base-pair (bp) stretch of DNA, which is sharply bent and tightly wrapped around a histone protein octamer. This sharp bending occurs at every DNA helical repeat (~10 bp), when the major groove of the DNA faces inwards towards the histone octamer, and again ~5 bp away, with opposite direction, when the major groove faces outward. Bends of each direction are facilitated by specific dinucleotides. Neighbouring nucleosomes are separated from each other by 10–50-bp-long stretches of unwrapped linker DNA; thus, 75–90% of genomic DNA is wrapped in nucleosomes.

The First Genetic Code



The Second Genetic Code



The Second Genetic Code

Sequence	Alignment Position			
	5	10	15	20
1-forward	C A T T A A G C C C G T A T T T C A A T T
1-reverse comp.	A A T T G T A A T A C G G G C T T A A T G
2-forward	A A A C T C T T C A T A T T
2-reverse comp.	A A T A T G A A G A G T T T
3-forward	A T G A A T T A A T G C A G T T C A A
3-reverse comp.	T T G A A C T G C A T T A A T T C A T
4-forward	G C T T A G T C T G C C T T A T
4-reverse comp.	A A A A G G C A G A C T A A G C
5-forward	T A T A A G C A A T A G C T T C T A
5-reverse comp.	T A G A A G C T A T T G C T T A T A

Add each sequence with a forward and a backward offset
(counts at each position will represent a three basepair moving average)

Sequence	Alignment Position			
	5	10	15	20
1-forward	C A T T A A G C C C G T A T T T C A A T T
1-forward (offset: +1)	C A T T A A G C C C G T A T T T C A A T
1-forward (offset: -1)	A T T A A G C C C G T A T T T C A A T T
1-reverse	A A T T G T A A T A C G G G C T T A A T G
1-reverse (offset: +1)	A A T T G T A A T A C G G G C T T A A T
1-reverse (offset: -1)	A T T G T A A T A C G G G C T T A A T G
1-forward	A A A C T C T T C A T A T T
1-forward (offset: +1)	A A A C T C T T C A T A T T
1-forward (offset: -1)	A A A C T C T T C A T A T T
1-reverse	A A T A T G A A G A G T T T
1-reverse (offset: +1)	A A T A T G A A G A G T T T
1-reverse (offset: -1)	A A T A T G A A G A G T T T
1-forward	A T G A A T T A A T G C A G T T C A A
1-forward (offset: +1)	A T G A A T T A A T G C A G T T C A A
1-forward (offset: -1)	A T G A A T T A A T G C A G T T C A A
1-reverse	T T G A A C T G C A T T A A T T C A T
1-reverse (offset: +1)	T T G A A C T G C A T T A A T T C A T
1-reverse (offset: -1)	T T G A A C T G C A T T A A T T C A T
1-forward	G C T T A G T C T G C C T T A T
1-forward (offset: +1)	G C T T A G T C T G C C T T A T
1-forward (offset: -1)	G C T T A G T C T G C C T T A T
1-reverse	A A A A G G C A G A C T A A G C
1-reverse (offset: +1)	A A A A G G C A G A C T A A G C
1-reverse (offset: -1)	A A A A G G C A G A C T A A G C
1-forward	T A T A A G C A A T A G C T T C T A
1-forward (offset: +1)	T A T A A G C A A T A G C T T C T A
1-forward (offset: -1)	T A T A A G C A A T A G C T T C T A
1-reverse	T A G A A G C T A T T G C T T A T A
1-reverse (offset: +1)	T A G A A G C T A T T G C T T A T A
1-reverse (offset: -1)	T A G A A G C T A T T G C T T A T A

Count dinucleotides

AA = 1	CA = 1	GA = 0	TA = 5
AC = 0	CC = 0	GC = 0	TC = 4
AG = 1	CG = 0	GG = 0	TG = 0
AT = 2	CT = 4	GT = 2	TT = 10

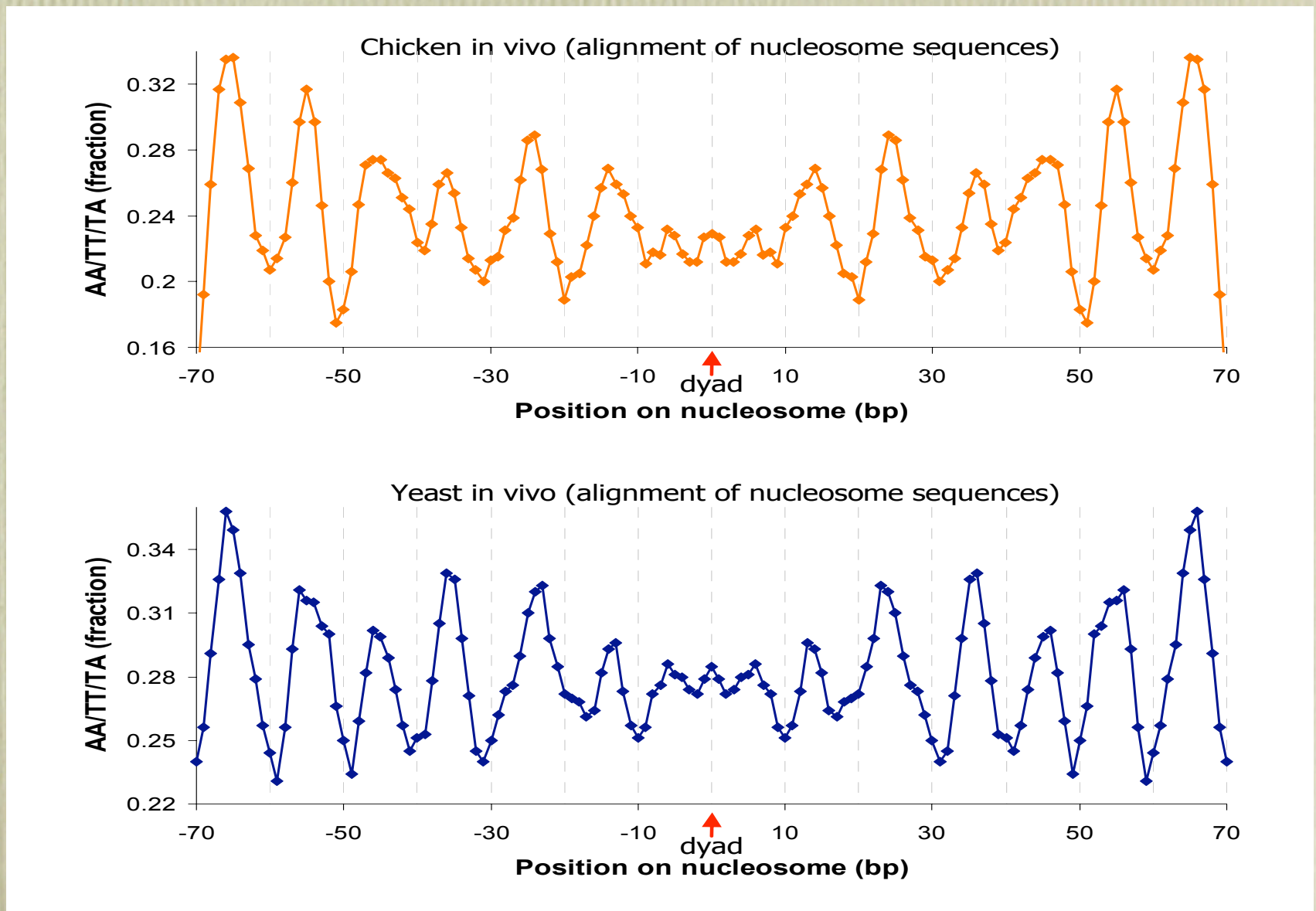
Represent as conditional counts

Nucleotide in position i	Nucleotide in position i-1			
	A	**C**	**G**	**T**
A	1	1	0	5
C	0	0	0	4
G	1	0	0	0
T	2	4	2	10

Derive conditional probabilities

Nucleotide in position i	Nucleotide in position i-1			
	A	**C**	**G**	**T**
A	0.25	0.2	0	0.263
C	0	0	0	0.211
G	0.25	0	0	0
T	0.5	0.8	1	0.526

The Second Genetic Code



The Second Genetic Code

These position specific *di*nucleotide distributions define the nucleosome-DNA model, such that probability assigned to 147bp sequence S is $P(S) = P_1(S_1) \prod_{i=2}^{147} P_i(S_i|S_{i-1})$.

A *legal configuration* specifies a *set* of 147bp nucleosomes and a start position for each such that no two nucleosomes overlap and the minimum distance between them is 10bp.

The probability of every configuration is then given by the Boltzmann distribution. While the number of configurations is vast, a forward-backward dynamic programming method enables efficient estimation of the probability of placing a nucleosome that starts at each basepair in the genome.

Rosen, O., Stoffer, D. (2007).

Biometrika, 94, 2, 335 – 345.

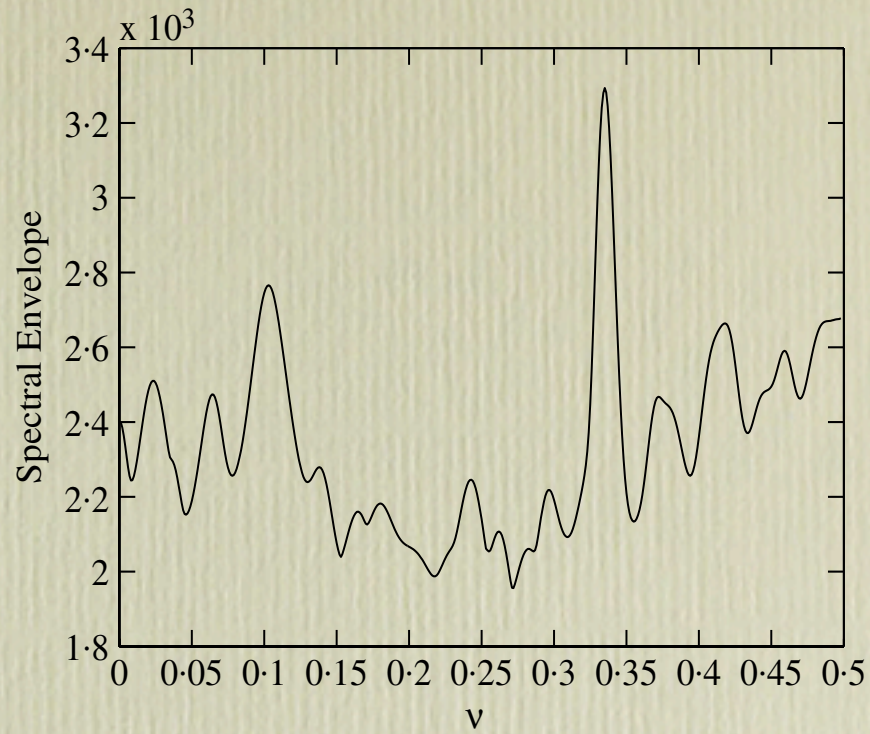


Fig. 5. Spectral envelope for part of a coding sequence in Herpesvirus saimiri.

"The spectral envelope picks up a signal at one cycle every three bps, which occurs often in coding sequences we have analyzed. There is another peak in the spectral envelope indicating a signal at one cycle every 10 bps. This signal is particularly interesting because, while the double helix makes one turn about every 10 base-pairs, the 10 bps signal is **rarely** seen and the importance of this twisting is not clear."

Contradiction: *Rare vs Everywhere*

- Not seen since not looking (sufficiently).
- Not seen due to spectral envelope limitations: power /thresholding, robustness...
- Not seen due to use of nt (*vs di*-nt) alphabet.

- Not seen since not there.

Spectral Envelope

Periodogram of a real-valued time series X_t at frequency ω :

$$I_n(\omega) = \left| n^{-1/2} \sum_{t=1}^{n} X_t \exp(-2\pi i \omega t) \right|^2$$

The spectral density, $f(\omega)$, is limit $(n \to \infty)$ of $E[I_n(\omega)]$

$$\mathrm{var}(X_t) = 2 \int_0^{1/2} f(\omega) d\omega = \sigma^2$$

Constructs generalize to k dimensional multivariate time-series, $\mathbf{Y_t}$, with spectral density $f_Y(\omega)$ a $k \times k$ complex-valued Hermitian matrix.

Categorical Time Series

$X_t = c_j$ when the time series is in state c_j at time t

Recast one dimensional categorical time series as a multivariate k dimensional time-series, $\mathbf{Y}_t = \mathbf{e}_j$ when $X_t = c_j$

The scaling process uses a vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$ to convert the multivariate time series to a univariate, real valued series, $X_t(\boldsymbol{\beta})$, by assigning category c_j value β_j : $X_t(\boldsymbol{\beta}) = \boldsymbol{\beta}' \mathbf{Y_t}$

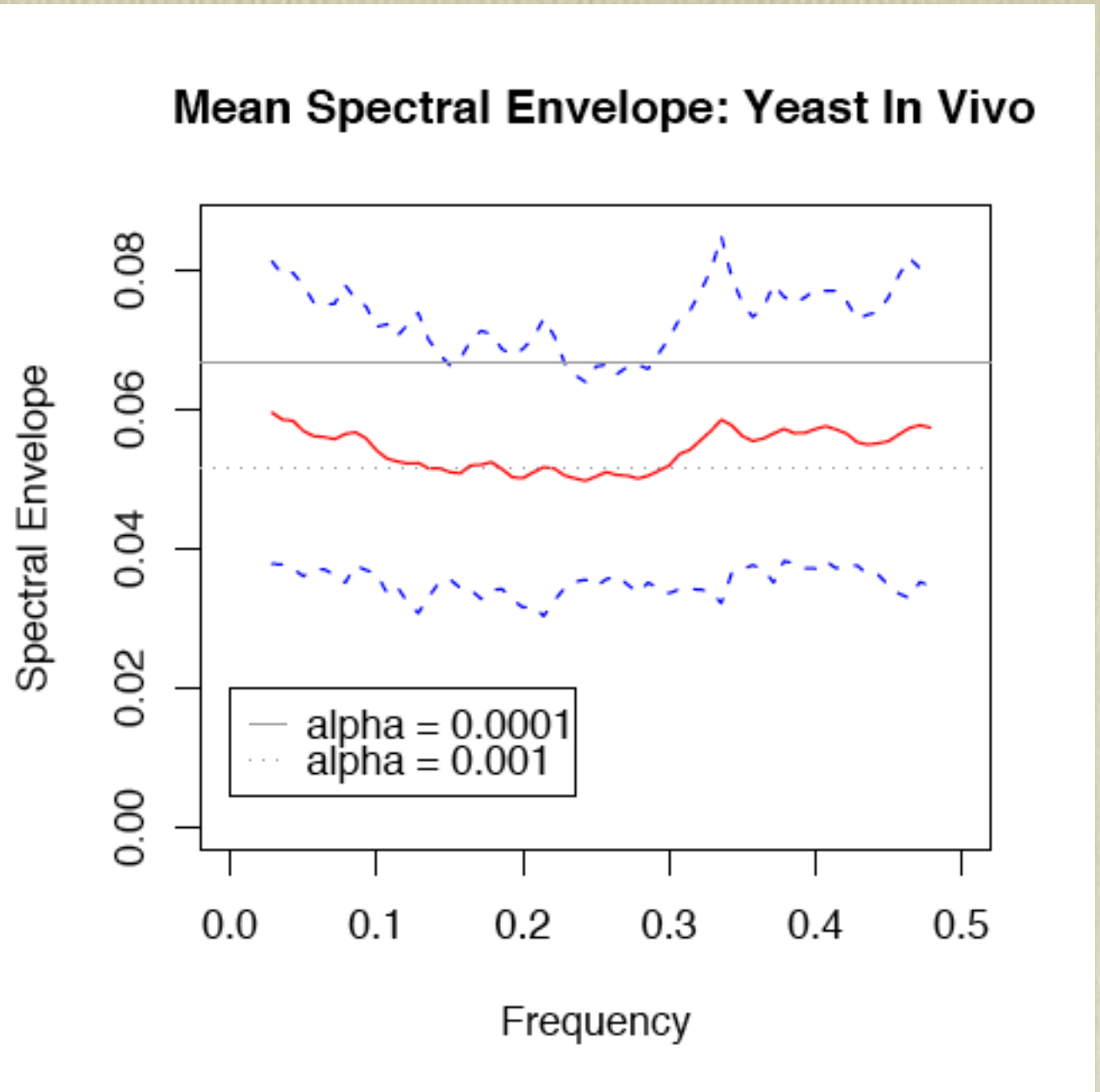
$f_Y(\omega; \boldsymbol{\beta}) = \boldsymbol{\beta}' f_Y^{re}(\omega) \boldsymbol{\beta}$ Key Question: how to choose β?

Maximize power at each ω -- leads to optimization criteria:

$$\lambda(\omega) = \sup_{\boldsymbol{\beta}} \left\{ \frac{\boldsymbol{\beta}' f_Y^{re}(\omega) \boldsymbol{\beta}}{\boldsymbol{\beta}' V \boldsymbol{\beta}} \right\}$$

Sequence Data

- Application of the spectral envelope to sequence data comes via mapping the relevant categories to a multivariate time series.
- We use a *di*-nucleotide alphabet, with and without select combination of categories (e.g. AA/TT/AT).
- Efficient R implementation using the discrete Fourier transform.

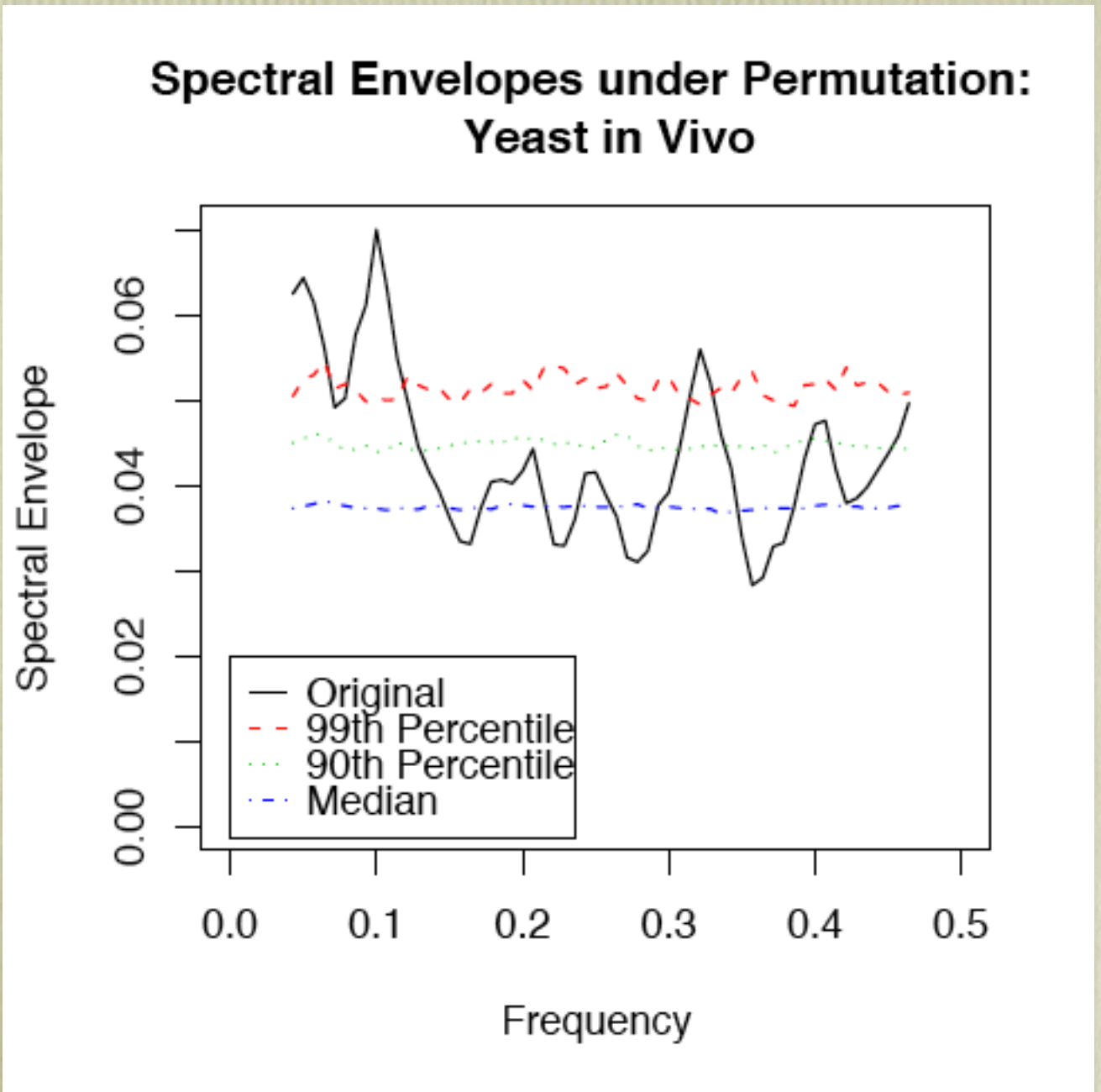


No peak at $\omega = 0.1$

So, the spectral envelope is not suitably sensitive.

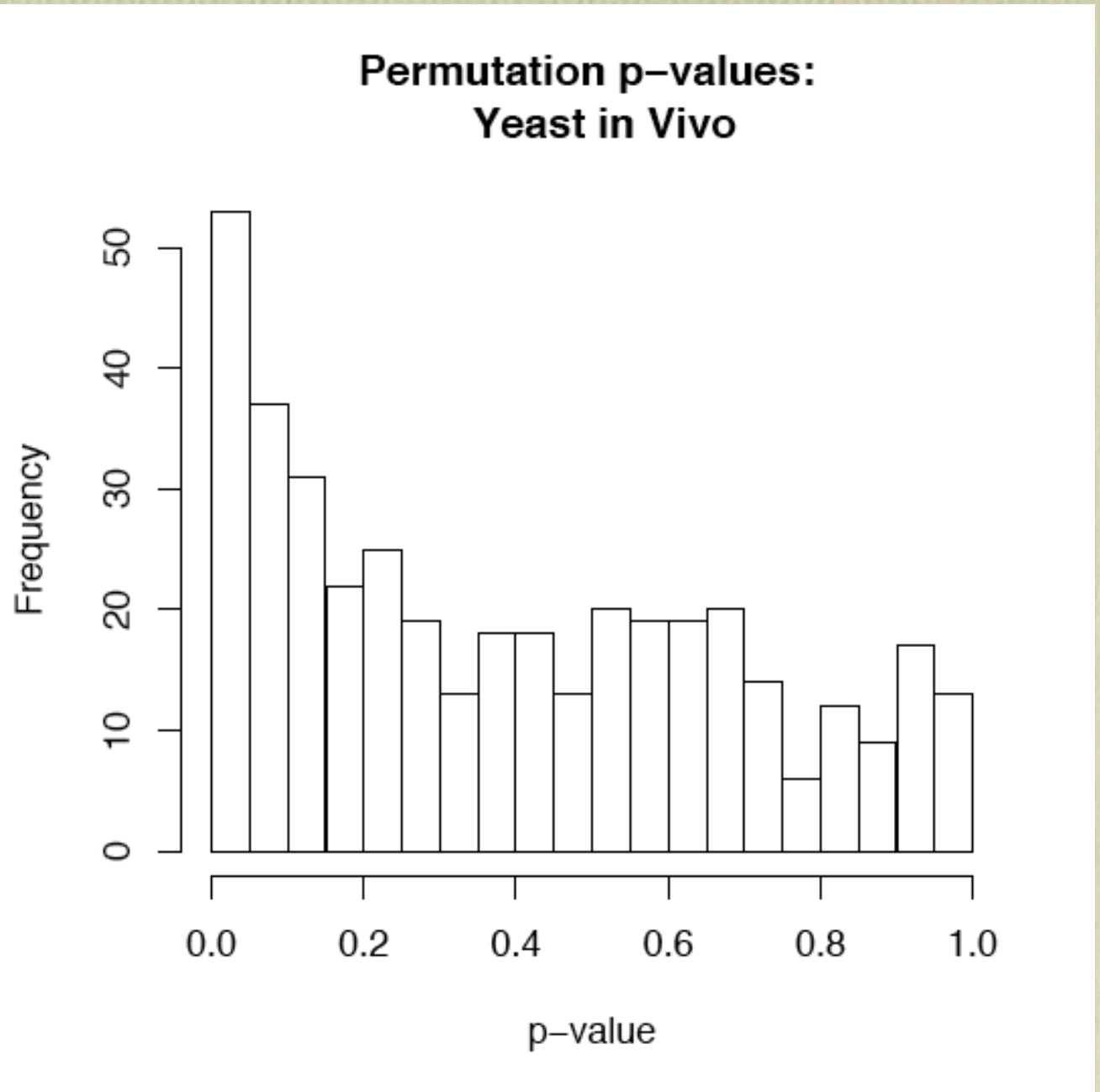
But, wait, this is the mean envelope over 398 sequences...

Combined Sequences



So, on an aggregate level, we do detect a signal at $\omega = 0.1$.

Begs question as to periodic behavior of the individual nucleosome bound sequences.



So, while about 50 sequences exhibit a significant peak at $\omega = 0.1$, the bulk do not display the characterizing 10bp periodicity.

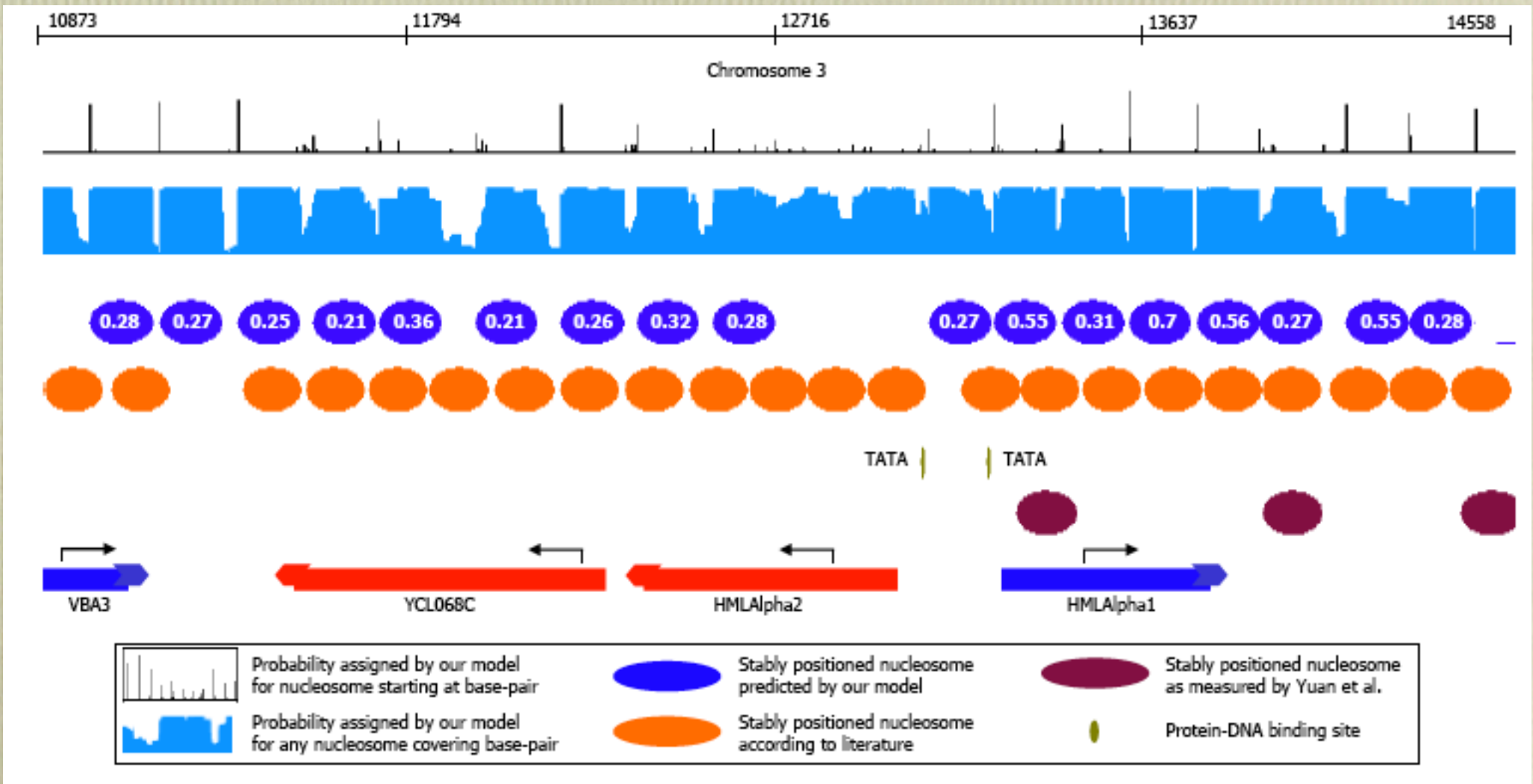
Same story for the other systems: yeast *in vitro*, chicken, synthetic.

- While this calls into question the basis for the code, does not preclude other aspects of the probability / steric model from explaining nucleosome occupancy.

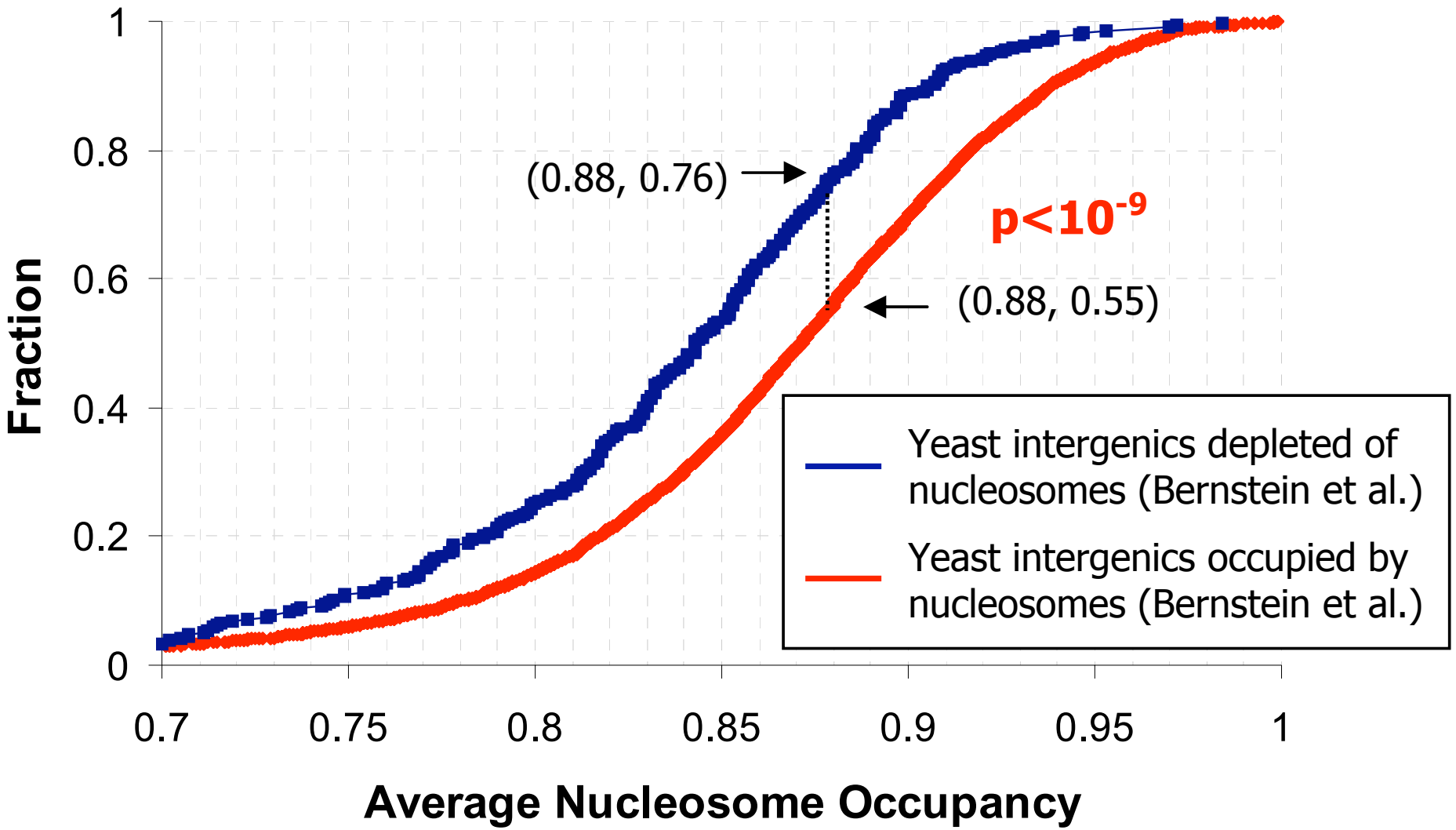
- Segal *et al* (2006) proffer a mound of evidence in order to demonstrate legitimacy of the positioning code.

- We next scrutinize the key components thereof.

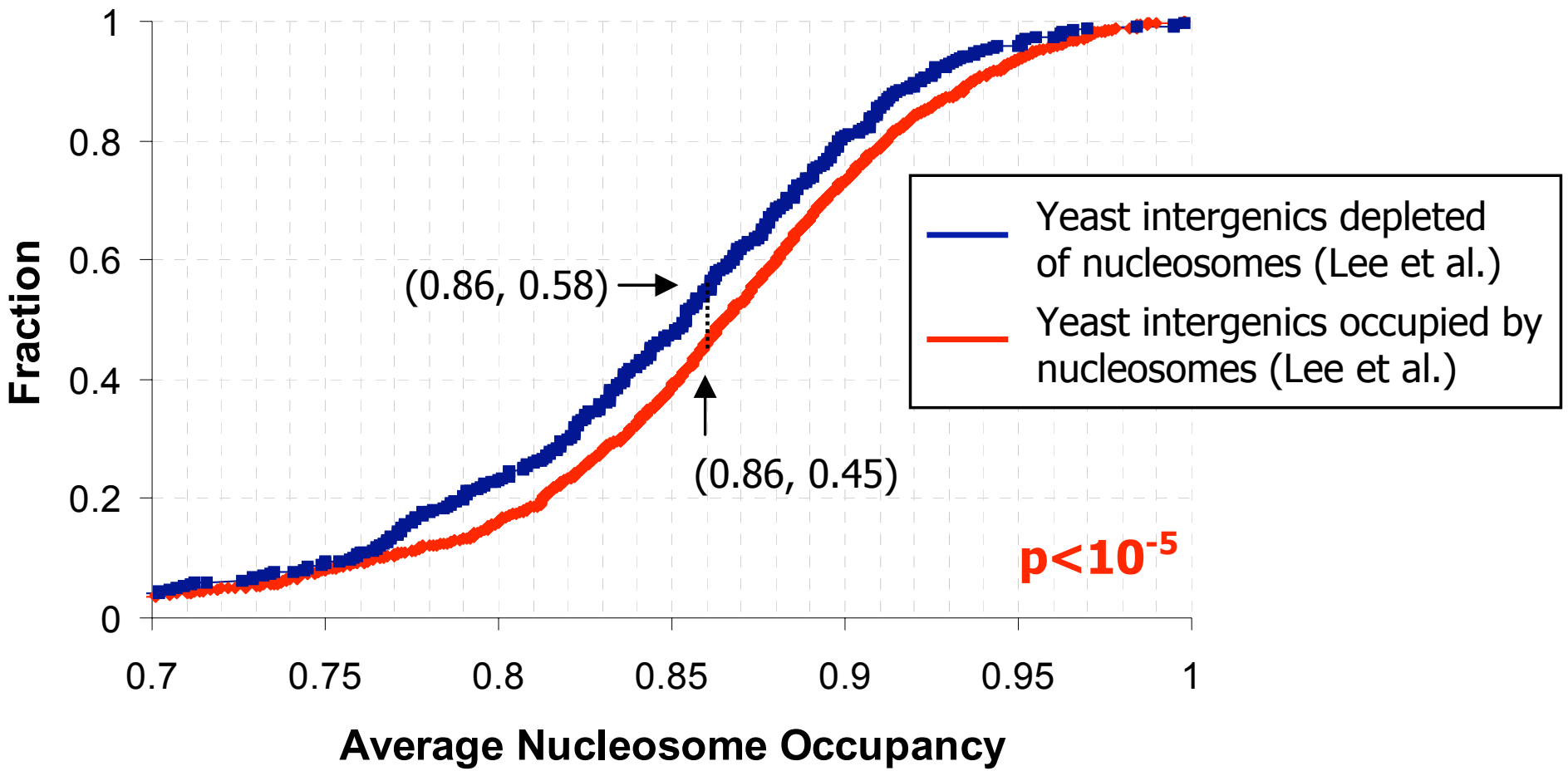
Lots of figures like...



And a few like Supp Fig 24:



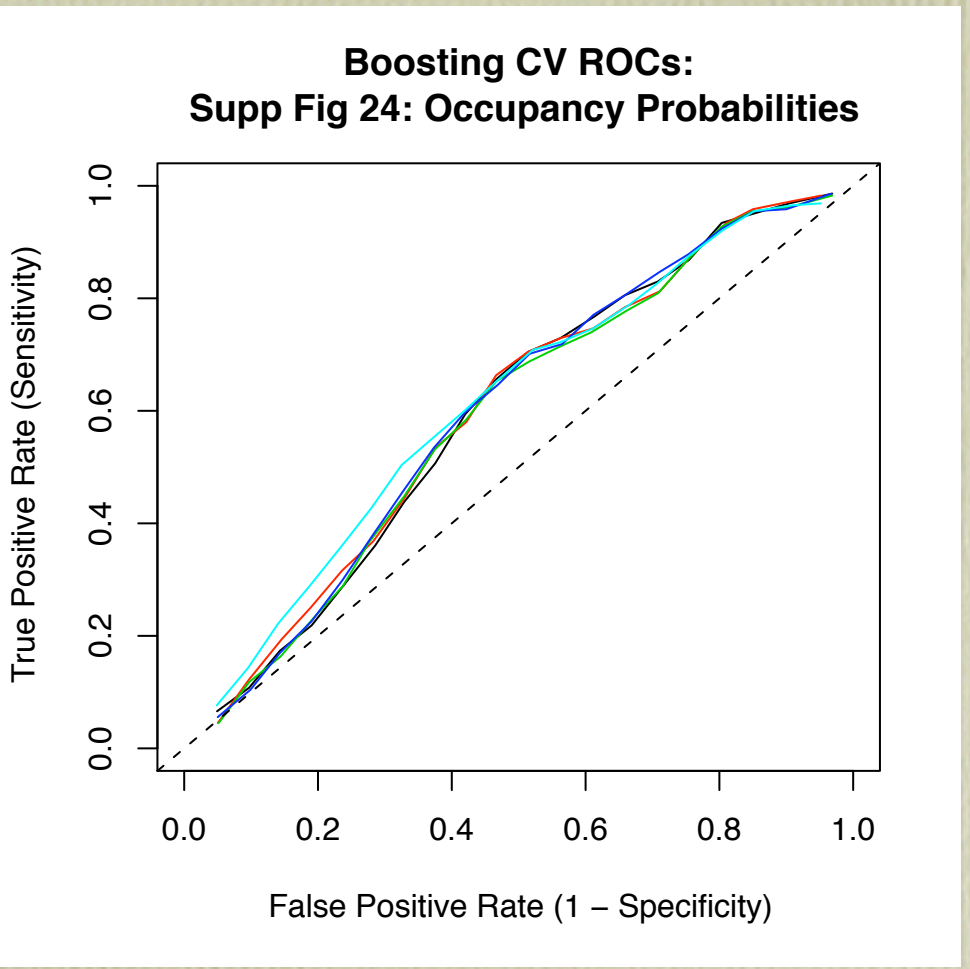
Here is Supp Fig 25:

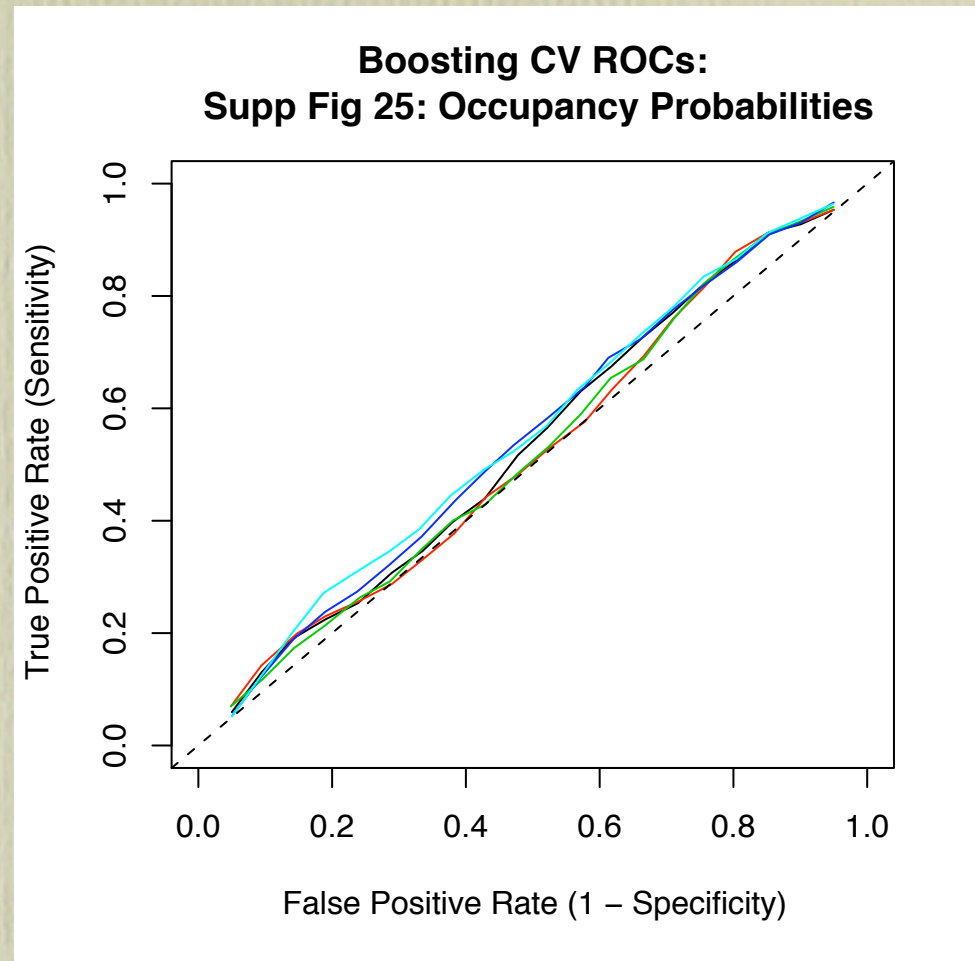


- Finally, a genome-scale, discriminatory perspective.
    - Null model won't look so good.
- How were these data obtained and analyzed?
- What about other data acquisition and analysis possibilities?

- We will leave aside issues surrounding the selection of sequences corresponding to nuclesome depleted *vs* occupied regions.
    - Attained by (arbitrary) thresholding.
- But we will revisit Average Nucleosome Occupancy probabilities (ANOPs), which embody the second genetic code, after assessing the analysis approach employed.

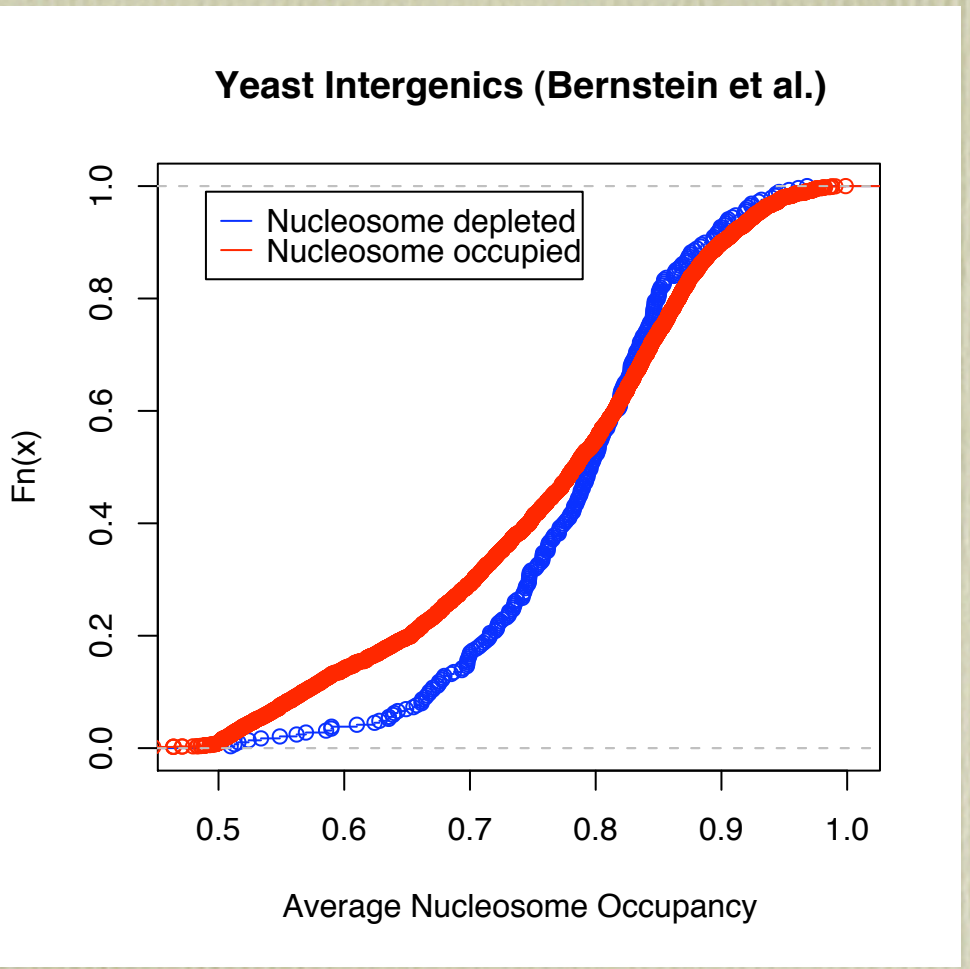
- Two sequence sets: depleted, occupied
- ANOPs for each: 294, 5387 (Supp 24)
- Differences assessed via Kolmogorov-Smirnov (KS) testing of empirical cdfs
- Easy to show that optimization inherent in KS equates to optimizing the sum: sensitivity + specificity
- Not bad; not *necessarily* good: ROC / CV

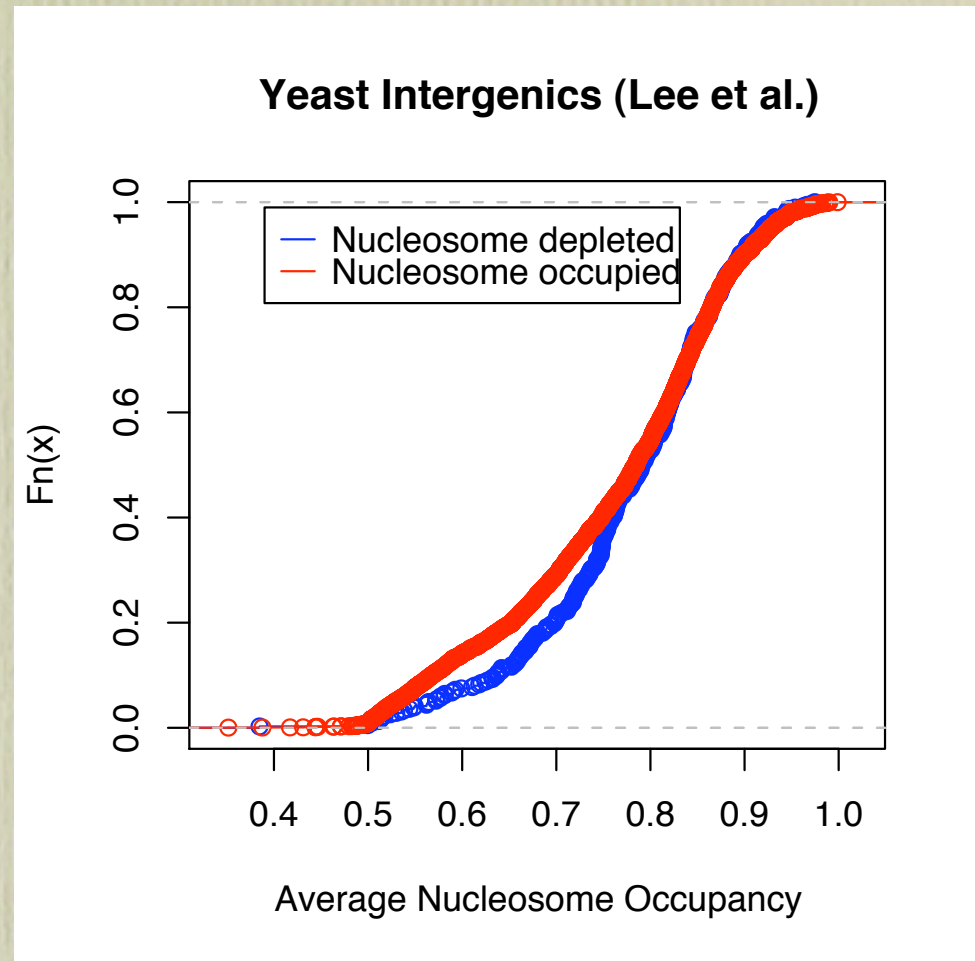

Boosting CV ROCs:
Supp Fig 24: Occupancy Probabilities
True Positive Rate (Sensitivity)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (1 – Specificity)


Boosting CV ROCs:
Supp Fig 25: Occupancy Probabilities
True Positive Rate (Sensitivity)
0.0
0.2
0.4
0.6
0.8
1.0
False Positive Rate (1 – Specificity)

- Before abandoning the second genetic code based on these ROC curves we thought it prudent to check the input ANOPs.

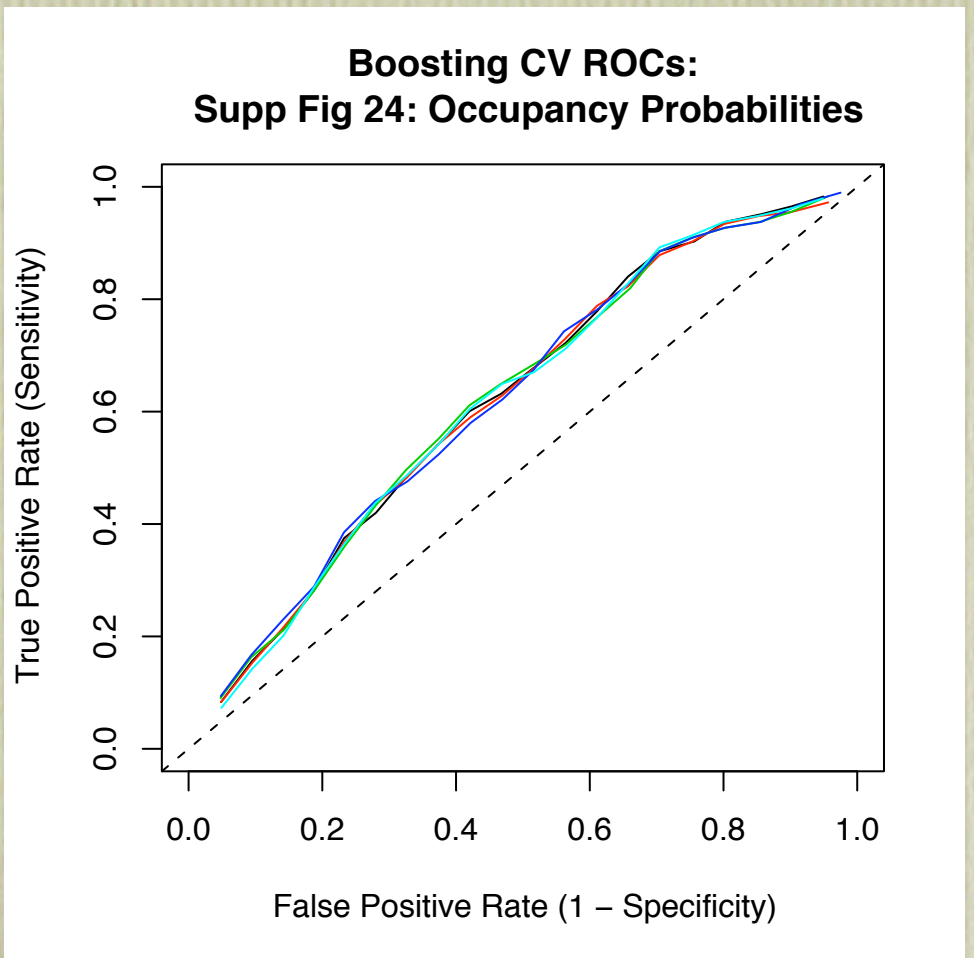
- These had been computed by submitting the depleted, occupied sequences to the script provided by Segal *et al*.

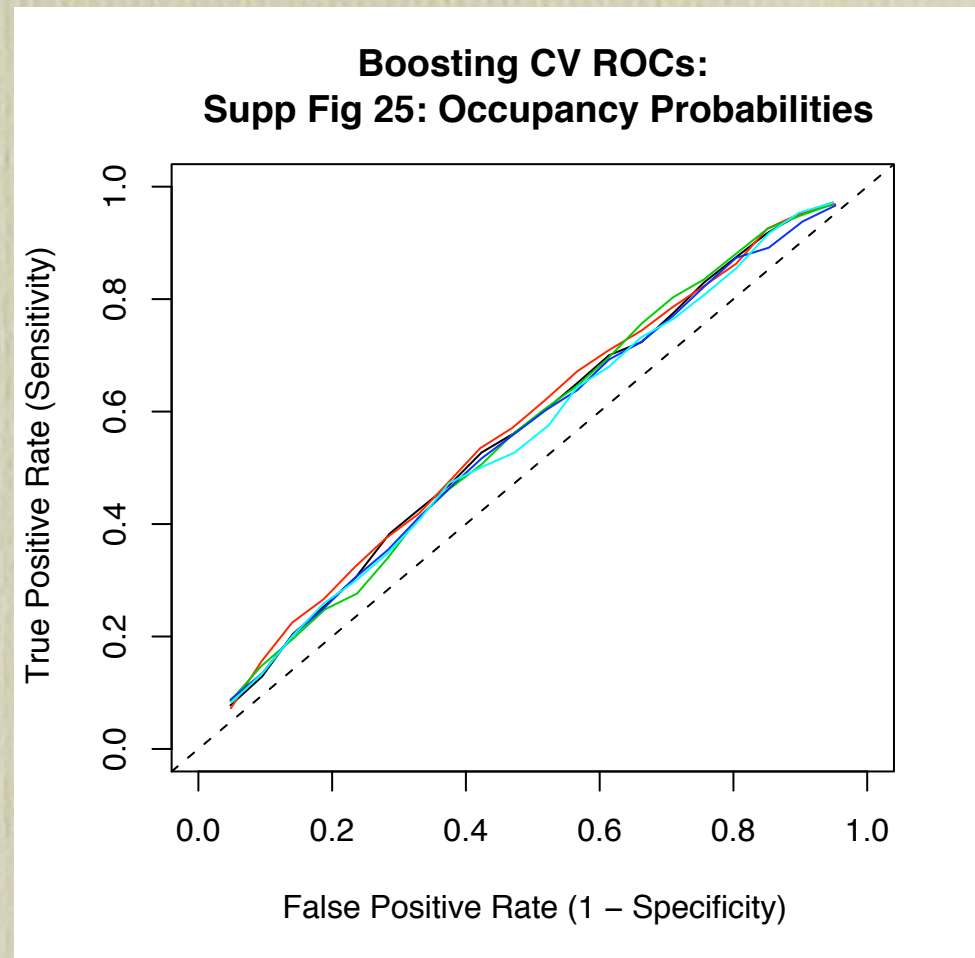

Yeast Intergenics (Bernstein et al.)
Nucleosome depleted
Nucleosome occupied
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0.5
0.6
0.7
0.8
0.9
1.0
Average Nucleosome Occupancy


Yeast Intergenics (Lee et al.)
Nucleosome depleted
Nucleosome occupied
Fn(x)
0.0
0.2
0.4
0.6
0.8
1.0
0.4
0.5
0.6
0.7
0.8
0.9
1.0
Average Nucleosome Occupancy

- Why the disparity with published ANOPs?
- To avoid boundary effects either extensive flanking sequence (5kb), or chromosome level computation with subsequent sequence specific assignment recommended.
- As seen, this has a profound impact.
- Interpretation, implications very unclear: target sequence length, location...

Reverting to the original ANOPs shows modest improvement:



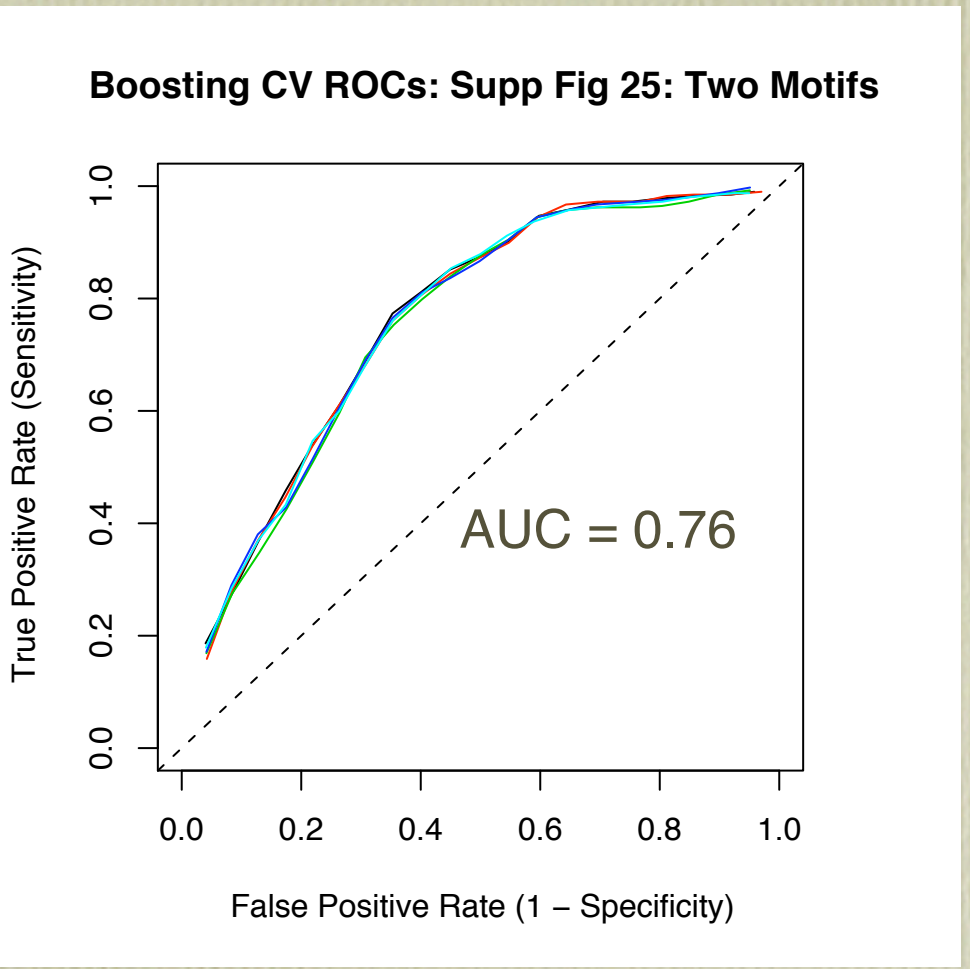


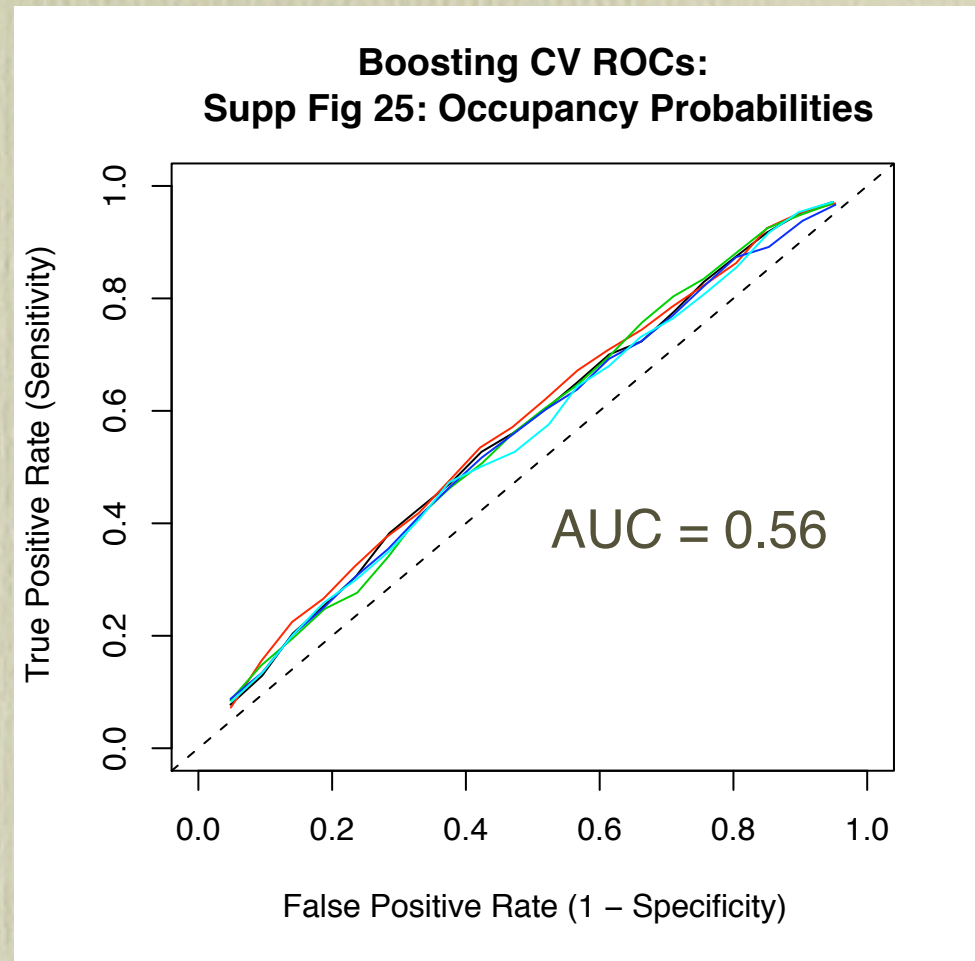
- The *classification* performance of the code is not particularly convincing.
- Yet such criteria *have* to form the basis on which the code is judged.
- Well, what else can be done?
- Have two sequence sets: depleted, occupied.
- Use *motif* (pattern) finding methods to obtain features for downstream classification.

- There is a sizeable literature on finding patterns common to sets of sequences.
- Three broad approaches: model-based, enumerative, and dictionary-based.
- Generative, discriminative modes for each.
- WordSpy: hybrid dictionary, model-based technique with steganography origins.
- Outperformed competitors in a recent benchmarking (TFBS) study.

- Applying WordSpy to the Supp Fig 24 (Bernstein) study yields leading motifs that include poly(dA.dT) and variants on the Rap1 binding motif (CACCCATACAT).

- We use the frequency of occurrence for just these 2 motifs as (classifier) features and apply them to an independent dataset (also yeast intergenics): Supp Fig 25 (Lee).


Boosting CV ROCs: Supp Fig 25: Two Motifs
True Positive Rate (Sensitivity)
False Positive Rate (1 – Specificity)
AUC = 0.76
0.0
0.2
0.4
0.6
0.8
1.0


Boosting CV ROCs:
Supp Fig 25: Occupancy Probabilities
True Positive Rate (Sensitivity)
False Positive Rate (1 – Specificity)
AUC = 0.56
0.0
0.2
0.4
0.6
0.8
1.0

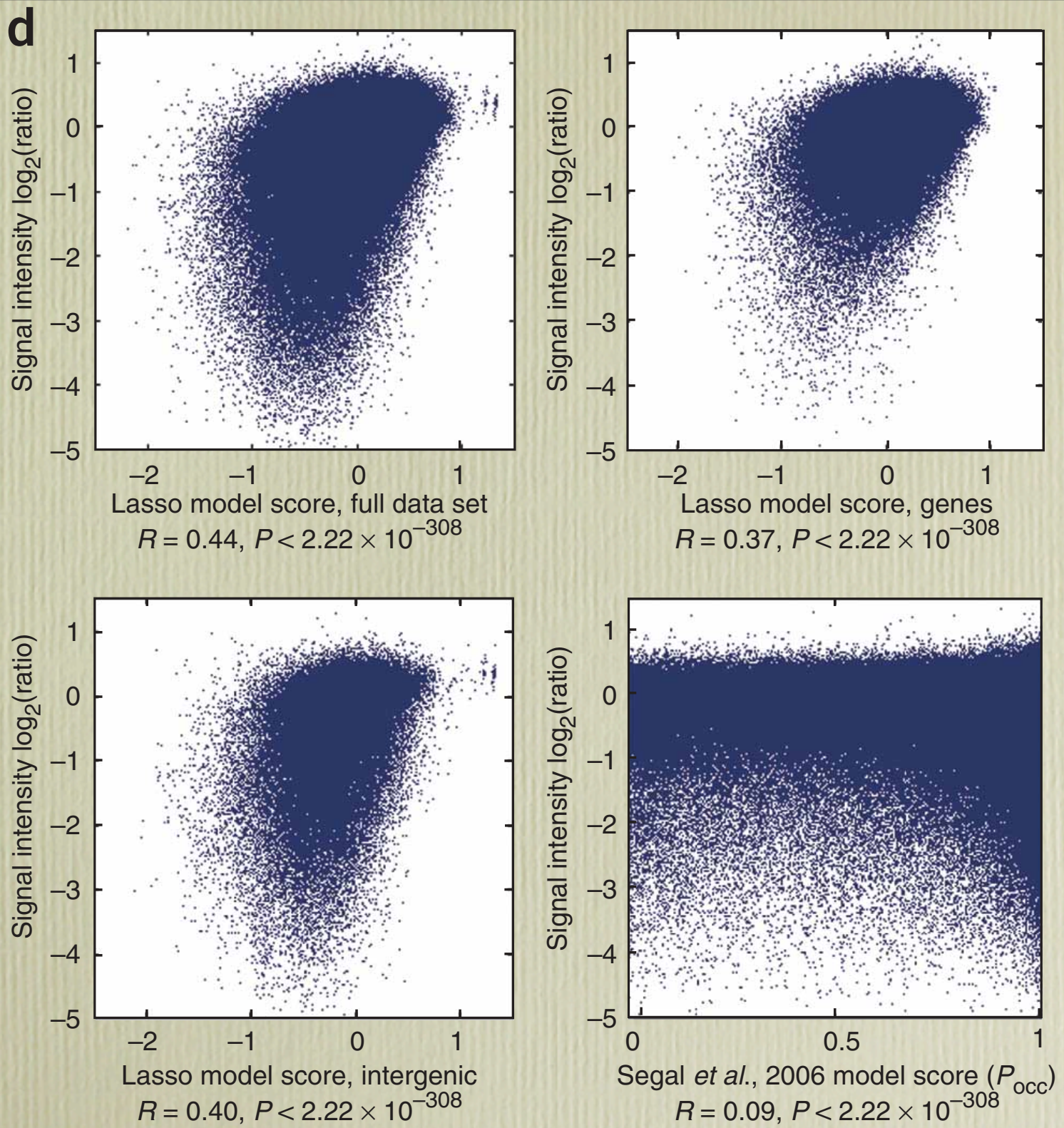
- So, the two motif model is superior to that based on Segal's proposed positioning code.
- Further, there is experimental support for a "code" based on each of these motifs: e.g., nucleosomes are depleted in the vicinity of Rap1 consensus sites and this depletion can be reversed by the small molecule rapamycin or by removing Rap1 binding sites.
- However, we are not claiming a new code.

Conclusions

- Seemingly serious questions surround both the basis for, and performance of, the proposed second genetic code.
- While a wealth of additional evidence is proffered by Segal *et al*., this does not address the necessary discriminatory aspects.

Epilog

- Many subsequent investigations: all reach similar conclusions:
  - importance of poly(dA.dT) tracts (*cf* GC content) select motifs (including rap1), and local structural properties.
  - periodicities, as embodied by Segal's code, have considerably lesser role.


d
Signal intensity log2(ratio)
1
0
–1
–2
–3
–4
–5
–2
–1
0
1
Lasso model score, full data set
R = 0.44, P < 2.22 × 10–308
Lasso model score, genes
R = 0.37, P < 2.22 × 10–308
Lasso model score, intergenic
R = 0.40, P < 2.22 × 10–308
0
0.5
1
Segal et al., 2006 model score (Pocc)
R = 0.09, P < 2.22 × 10–308

Model	Summary	Performance (Pearson R)						Correlation with %G+C (Yeast, 150 bp windows)
		Synthetic oligonucleotides (Microarray) [9]	Synthetic oligonucleotides (Sequencing) [9]	Yeast *in vitro* [9]*	Yeast *in vivo* [2]*	*C. elegans* adjusted nucleosome coverage [33]**	*C. elegans* normalized occupancy [33]**	
Lasso model (this study)	See methods.	0.44	0.41	0.86	0.38	0.49	0.66	0.85
Kaplan et al., 2009[9]	Probabilistic model based on in vitro 5-mer preferences and periodic dinucleotide signal.	0.51	0.45	0.89	0.34	0.47	0.61	0.87
Field et al., 2008[24]	Probabilistic model based on 5-mer preferences measured *in vivo* (yeast) and periodic dinucleotide signals.	0.47	0.45	0.74	0.39	0.46	0.61	0.64
Lasso model[2]	Linear regression model trained on *in vivo* nucleosome occupancy data. Uses DNA structural parameters, excluding sequences and transcription factor binding sites (ABF1, REB1, and STB2) as inputs.	0.23	0.22	0.63	0.45	0.38	0.5	0.55
Peckham et al., 2007[25]	SVM classifier trained on overrepresented k-mers (k=1-6) found in nucleosome occupied and depleted sequences determined *in vivo* yeast data.	0.43	0.39	0.48	0.22	0.29	0.33	0.57
Yuan and Liu, 2008[26]	Computes predicted nucleosome occupancy based on periodic dinucleotide signals found in nucleosomal and linker DNA sequences determined from *in vitro* and *in vivo* experiments in yeast.	0.02	0.05	0.35	0.27	0.36	0.48	0.30
Miele et al., 2008[29]	Computes free energy landscape of nucleosome formation using an estimation of dinucleotide-dependent DNA flexibility and intrinsic curvature.	0.32	0.26	0.38	0.22	0.21	0.25	0.49
%G+C, 150 bp windows	The proportion or percentage of guanine and cytosine bases in a DNA sequence.	0.53	0.49	0.78	0.25	0.42	0.47	1
Segal et al., 2006[23] (Downloaded Jan 2007)	Probabilistic model trained on yeast data, using a position specific scoring matrix derived from a collection of nucleosome-bound sequences obtained from in vitro selection experiments.	NaN	NaN	0.05	0.09	0.05	0.05	0.07
Segal et al., 2006[23] (Downloaded Aug 2009)	Probabilistic model trained on yeast data, using a position specific scoring matrix derived from a collection of nucleosome-bound sequences obtained from in vitro selection experiments.	NaN	NaN	-0.2	0.001	-0.06	-0.05	-0.21
Ioshikhes et al., 2006[22]	Computes the correlation of periodic AA/TT dinucleotide motif found in a set of 204 eukaryotic and viral nucleosomal sequences determined through in vivo and *in vitro* experiments[20].	-0.03	-0.03	0.01	0.07	-0.03	-0.01	0.01
Tolstorukov et al., 2007,2008[31,32]	Estimates the dinucleotide-dependent cost of deformation caused by threading a given sequence on a template comprising the path of DNA found on the experimentally determined structure of the nucleosome core particle.	-0.01	-0.004	0	-0.001	0.001	0.001	0.0003

Acknowledgments

- Eran Segal graciously provided data and patiently, yet rapidly, responded to queries.
- Guo-Cheng Yuan and Jun Liu provided a preprint and helpful comments.
- Jun Song and Tim Hughes provided helpful comments.