

UCSF

UC San Francisco Electronic Theses and Dissertations

Title

Chromatin Regulatory Signatures in *Saccharomyces cerevisiae*

Permalink

<https://escholarship.org/uc/item/5x90z4md>

Author

Wu, Randy

Publication Date

2008

Peer reviewed|Thesis/dissertation

Chromatin Regulatory Signatures in *Saccharomyces cerevisiae*

by

Randy Wu

DISSERTATION

Submitted in partial satisfaction of the requirements for the degree of

DOCTOR OF PHILOSOPHY

in

Biophysics

in the

GRADUATE DIVISION

of the

UNIVERSITY OF CALIFORNIA, SAN FRANCISCO

Copyright 2008
by
Randy Wu

Dedication and Acknowledgements

I dedicate this Ph.D. thesis to my mother Joy T. Yang, Ph.D. She is a loving parent in addition to a stellar scientist and I could not have completed my work without her unfailing support and guidance.

I would also like to thank Hao Li, Ph.D. for his wisdom and outstanding mentorship.

Finally, I want to thank Stephanie T. Yang for her love and inspiration.

The text of this thesis is a reprint of the material as it appears in *Genome Research*. The coauthor listed in this publication directed and supervised the research that forms the basis for the dissertation/thesis.

Wu, R.Z. and Li, H. (2008) Directed A/T-tracts: A Novel Signature for Yeast Nucleosome-Free Regions. *Genome Res* (submitted).

The text of this thesis includes a reprint of the material as it appears in BMC Bioinformatics. Randy Wu performed the primary calculations in this publication, produced most of the data, and wrote the manuscript. The work is comparable to a standard thesis. -- Hao Li, PhD.

Wu, R.Z., Chiavorapol, C., Zheng, J., Liang S. and Li, H. (2007) fREDUCE: Detection of degenerate regulatory elements using correlation with expression. *BMC Bioinformatics* **8**, 399.

Chromatin Regulatory Signatures in *Saccharomyces cerevisiae*

Randy Wu

Abstract

Eukaryotic transcriptional regulation is mediated by the organization of chromatin in promoter regions. This thesis describes three projects which examine the relationships between chromatin and transcriptional regulation in the budding yeast *S. cerevisiae*. First we describe a novel computational algorithm fREDUCE for the elicitation of regulatory motifs given sequence and expression data as inputs. fREDUCE is used to find T_nC motifs, novel repetitive sequences occurring prominently within nucleosome-free regions of promoters. The second chapter describes the relationships between T_nC and chromatin structure in fine details. We conclude that T_nC motifs constitute directional signature sequences which likely play roles in defining the locations of nucleosome-free regions in a majority of yeast promoters. Finally, we also undertake a quantitative and systematic examination of the relationship between transcription factors, their binding sites, and their corresponding chromatin environments. We find that the yeast transcriptome encompasses a diverse set of signature TF-chromatin relationships. Taken together, these three studies examine multiple facets of the intricate nature of chromatin regulation in a simple eukaryotic organism.

Table of Contents

Introduction	1
Chapter 1: fREDUCE	10
Chapter 2: Directed A/T Tracts	38
Chapter 3: TF-Chromatin Signatures	81
UCSF Library Release	94

List of Tables

Chapter 1

Table 1. fREDUCE predictions from 65 yeast ChIP-chip experiments of Harbinson	30
Table 2a: fREDUCE predictions in comparison to non-degenerate predictions made by REDUCE.	32
Table 2b: fREDUCE elicitation of the HNF-4 binding site from human hepatocyte expression data.	33
Table 3. fREDUCE predictions for regulators with poorly characterized specificities.	34

Chapter 2

Table 1: Poly(dA:dT) coverage and copy number	64
---	----

List of Figures

Introduction

- Figure 1. Views of chromatin at various levels of detail. 6
- Figure 2. The *PHO5* gene as a model system for examining the role of nucleosome context in transcription 7
- Figure 3. The determination of genome-wide nucleosome positions using a tiling-array approach 8

Chapter 1

- Figure 1. The fREDUCE algorithm. 35
- Figure 2. Comparison of fREDUCE to six other algorithms on 65 yeast ChIP-chip benchmarks. 36
- Figure 3. Scalability of fREDUCE. 37

Chapter 2

- Figure 1. Poly(dA:dT) tracts in strong- and weak-NFR promoters. 65
- Figure 2. Poly(dA:dT) track fine variations in NFR positions. 67
- Figure 3. Poly(dA:dT) enrichments occur independently of sampled transcription factor binding sites. 69
- Figure 4. NFR-specific 5' G:C capping of poly(dA:dT) tracts. 70
- Figure 5. Poly(dA:dT) capping is offset from tract enrichments toward the NFR central axis. 71
- Figure 6. Two contrasting mechanistic models of NFR definition by poly(dA:dT) signals. 72
- Figure 7. Locations of promoter elements in strong-NFR subgroups favors the Central Definition model. 73
- Figure 8. Summary of mechanistic hypotheses describing how poly(dA:dT) tracts can lead to the formation of

nucleosome-free regions.	74
Supplementary Figure 1 A) Average nucleosome profiles for divergently-transcribed vs. single-direction promoters. B) Observed vs. expected number of divergently-transcribed promoters in strong- vs. weak-NFR classes.	75
Supplementary Figure 2 Observed vs. expected number of TATA-box containing promoters in strong- vs. weak-NFR classes.	76
Supplementary Figure 3 Poly(dA:dT) enrichments in subgroups of the strong-NFR class.	77

Chapter 3

Figure 1. Deriving TF-chromatin profiles for transcription factors in the <i>S. cerevisiae</i> genome.	90
Figure 2. Distributions of the derived parameters α , β , γ , δ .	91
Figure 3. Hierarchical clustering analysis of 122 TFs according to derived parameters.	93

Introduction

Chromatin refers collectively to the intricate complexes of protein and nucleic acids that make up chromosomes inside the nuclei of eukaryotic cells. The fundamental packing unit of chromatin is the nucleosome, which consists of 147 base pairs of DNA wrapped around a core octamer of histone proteins[1,2]. Individual nucleosomes connected by intervening linker DNA comprise the “beads-on-a-string”[3] model of chromatin organization at its lowest level. The “beads-on-a-string” are further packed into a hierarchical scheme of successively more complex macromolecular structures, the largest of which is the metaphase chromosome (Figure 1).

The elaborate organization of eukaryotic DNA as chromatin is thought to serve several essential biological functions. First, chromatin condenses the immense amounts of genomic DNA (3 billion base pairs in human), allowing it to be packaged into the small volume of typical cells (microns in diameter) in an orderly way. Second, chromatin provides DNA with a structural scaffold, giving it the physical integrity required for cellular events such as mitosis and meiosis in which large-scale genomic restructuring takes place. Finally, chromatin serves multiple regulatory functions that control DNA replication and gene expression.

Chromatin can exert regulatory influences on the expression patterns of underlying genes through 1) the structural repression of genes as heterochromatin[4], 2) the alteration of expression through covalent post-translational modifications of histone C-termini tails (reviewed in [5,6]) and 3) the more subtle influence on transcription factor

binding sites (TFBS) exerted by nucleosome-positioning. Among these categories of chromatin regulation, the last is perhaps the least well-understood and is the focal topic of this thesis.

It is generally appreciated that the nucleosomal context of a TFBS is an important determinant of its function. Perhaps the most best-known system for examining the detailed relationship between nucleosome-placement and transcription is the *S. cerevisiae* *PHO5* promoter. *PHO5*, encoding an alkaline phosphatase, is under regulation from the transcription factor Pho4p. The Pho4p binding site is situated in the *PHO5* promoter in a nucleosome-free region which is flanked by four positioned nucleosomes[7]. The placement of the Pho4p site in a nucleosome-free region is, in this case, critical to correct gene function[8]. In a more recent study, it has been demonstrated that both binding sites that are nucleosome-occupied as well as those that are nucleosome-free can contribute to transcriptional regulation, albeit with different functional roles[9]. In this case, it was found that *extra*-nucleosomal binding sites are initially recognized by the transcription factor and contribute to the induction of transcription, while *intra*-nucleosomal binding sites are initially hidden from the transcription factor but contribute to the steady state levels of gene activation.

The complex interplay between nucleosome positioning and transcription is made possible by the fact that nucleosomes do not occupy DNA in an *ad hoc* way. It is now well established that nucleosomes generally occupy predictable and well-defined positions in eukaryotic genomes with only modest (but functionally significant) variations over different environmental conditions. As such, with the advent of high throughput methods (including tiling microarrays and sequencing) it has been possible to

reconstruct the genome-wide nucleosomal “atlas” for several organisms[10,11]. The availability of these data allows unprecedented opportunities for the detailed computational study of how nucleosomes relate to their underlying genes. This thesis presents a set of three studies, each of which addresses one question related to the phenomena of transcriptional regulation through nucleosome placement.

fREDUCE: detection of degenerate regulatory motifs using correlation with expression

One of the central questions this thesis attempts to address is: how do DNA sequences affect the positioning of overlying nucleosomes? The first two chapters of this thesis address this question in two successive parts. First it is necessary to identify *what* candidate sequences may contribute to the positions of nucleosomes. Because we are concerned primarily with transcriptional regulation we restrict our analysis to intergenic sequences that lie 5' to genes. Thus we desire an algorithm whose inputs consist of promoter sequences and nucleosome positioning data and which outputs DNA motifs. While many existing algorithms (reviewed in Chapter 1) have these characteristics, we were concerned that, unlike TF binding sites with highly defined motifs, nucleosome-influencing sequences are likely to be degenerate motifs occurring in high copy number. Thus, traditional algorithms may potentially miss important candidates.

To address this concern we devised the novel algorithm fREDUCE, which is an improved version of its predecessor REDUCE (Regulatory Element Detection Using Correlation with Expression). Intuitively, REDUCE works by considering input sequence and expression data as vectors; a defined set of motifs are processed into

vectors of motif counts per promoter, vectors are correlated with expression data, and those motifs with the most significant correlations are chosen. fREDUCE allows the consideration of motifs expressed in the form of IUPAC symbols for multiple bases, allowing the systematic treatment of degenerate sites in the REDUCE scheme. Key to the fREDUCE algorithm are a number of computational shortcuts which dramatically improve its efficiency.

Directed A/T-tracts: a novel signature for nucleosome-free regions in yeast

Once potential motifs are identified, they must be analyzed for how they relate to the positions of nucleosomes. Many of the top-scoring sequence candidates identified by fREDUCE take the form of poly-A and poly-T repeats. These A/T-tracts generally appear to negatively correlate with positions of nucleosomes, indicating early on that they may function to repel nucleosomes. Another curious observation from the fREDUCE analysis is that many of the tracts appear as GAn or TnC; that is, there appears to be a G/C cap that is placed onto the ends of A/T tracts in a directionally specific manner. In the second chapter, we describe a set of computations which characterize the specific distributions of A/T tracts as well as the manners of their capping in yeast intergenic regions.

Our results were twofold. First, we observed that A/T tracts followed a characteristic distribution within nucleosome-free regions: A-tracts appear 3' relative to T-tracts and the two are symmetric relative to the central coordinate or the NFR. Furthermore, the width of the NFR appears to correspond with the locations of the A/T-tracts in a strongly quantitative way. Second, G/C capping of A/T-tracts, while a

phenomenon occurring in intergenic regions in general, is especially prominent within NFRs. The highly characteristic placement of capped A/T-tracts led us to hypothesize that they play formative functions in defining the bounds of promoter NFRs, and that their directionality may be important in this capacity.

Surveying TF-chromatin profiles in the yeast genome

Having analyzed in extensive detail the functions of a putative set of NFR-directing sequences, in Chapter 3 we step back and look at how nucleosome positions affect transcriptional regulation. We try to address a number of questions such as 1) Do TF binding sites on the whole prefer to lie in nucleosome-free regions, where are are presumably more free from steric hinderance? 2) How do nucleosomes affect the ability of TFs to bind to their cognate sites? 3) How do nucleosomes affect the ability of bound transcription factors to carry out their intended function? All of these questions can be asked independently for each of the 122 TFs surveyed.

We do this by considering the nucleosome-occupation, TF-binding, and sequence conservation data collectively and creating a set of derived parameters each of which answers one of the stated questions. The TF-chromatin relationships of each transcription factor can thus be characterized by a set of four derived parameters. By clustering TFs according to the derived parameters, we gain a global view of trends in TF-chromatin relations in the transcriptome.

Figures

Figure 1. Views of chromatin at various levels of detail.

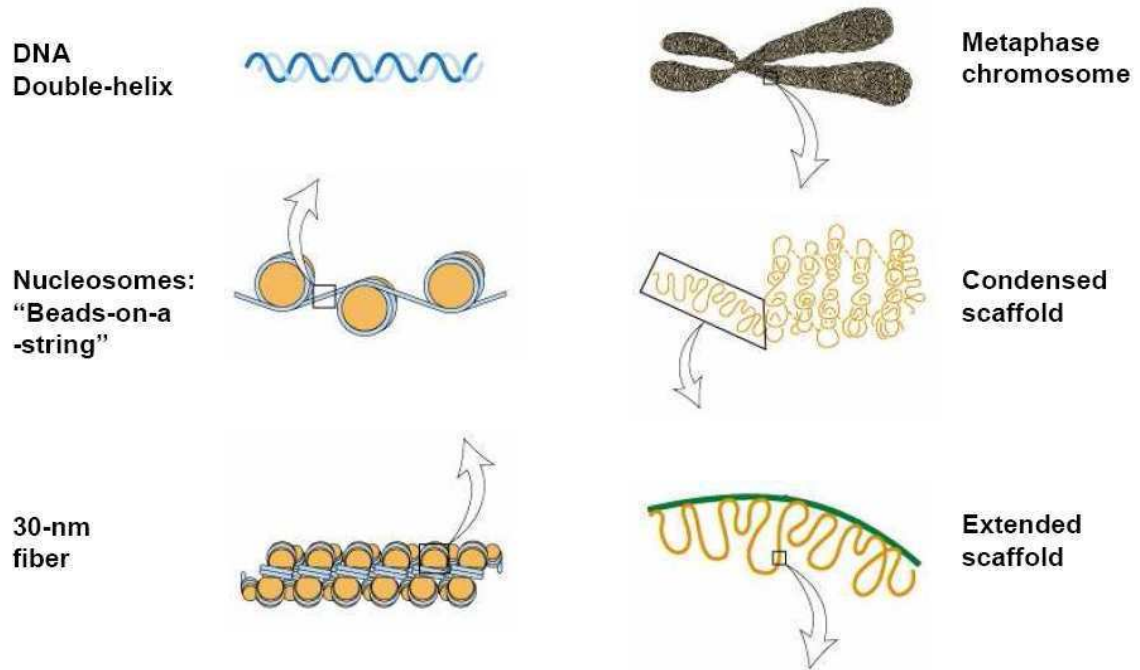
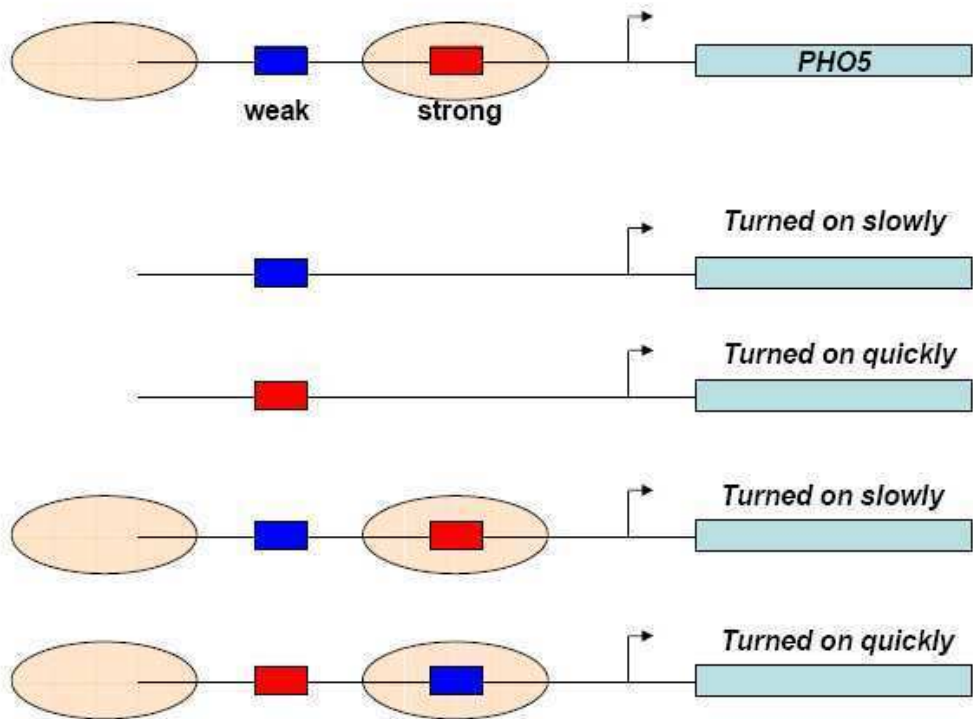
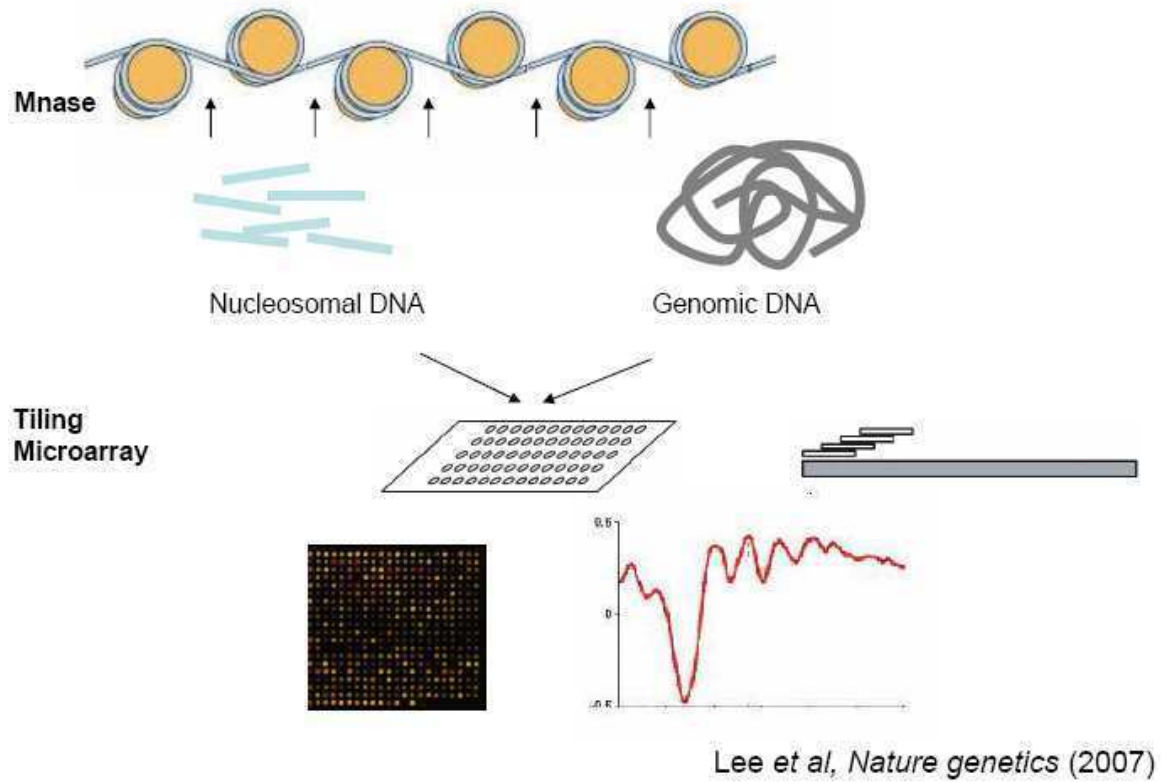


Figure 2. The *PHO5* gene as a model system for examining the role of nucleosome context in transcription (Lam et al. 2008). *PHO5* contains two binding sites for the transcription factor Pho4p: a strong site located in a nucleosome and a weak site located in a nucleosome-free region. It was observed that the induction profiles of genes containing mutant promoters behaved as if the nucleosomal Pho4p site was masked.



Lam et al, Nature (2008)

Figure 3. The determination of genome-wide nucleosome positions using a tiling-array approach (Lee et al. 2007).



References

1. Richmond TJ, Davey CA (2003) The structure of DNA in the nucleosome core. *Nature* 423: 145-150.
2. Richmond TJ, Finch JT, Rushton B, Rhodes D, Klug A (1984) Structure of the nucleosome core particle at 7 Å resolution. *Nature* 311: 532-537.
3. Li HJ (1975) A model for chromatin structure. *Nucleic Acids Res* 2: 1275-1289.
4. Grewal SI, Jia S (2007) Heterochromatin revisited. *Nat Rev Genet* 8: 35-46.
5. Narlikar GJ, Phelan ML, Kingston RE (2001) Generation and interconversion of multiple distinct nucleosomal states as a mechanism for catalyzing chromatin fluidity. *Mol Cell* 8: 1219-1230.
6. Grunstein M (1997) Histone acetylation in chromatin structure and transcription. *Nature* 389: 349-352.
7. Almer A, Horz W (1986) Nuclease hypersensitive regions with adjacent positioned nucleosomes mark the gene boundaries of the PHO5/PHO3 locus in yeast. *EMBO J* 5: 2681-2687.
8. Straka C, Horz W (1991) A functional role for nucleosomes in the repression of a yeast promoter. *EMBO J* 10: 361-368.
9. Lam FH, Steger DJ, O'Shea EK (2008) Chromatin decouples promoter threshold from dynamic range. *Nature* 453: 246-250.
10. Lee W, Tillo D, Bray N, Morse RH, Davis RW, et al. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* 39: 1235-1244.
11. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, et al. (2008) Nucleosome organization in the *Drosophila* genome. *Nature* 453: 358-362.

Chapter 1

fREDUCE: Detection of Degenerate Regulatory Elements

Using Correlation with Expression

Abstract

The precision of transcriptional regulation is made possible by the specificity of physical interactions between transcription factors and their cognate binding sites on DNA. A major challenge is to decipher transcription factor binding sites from sequence and functional genomic data using computational means. While current methods can detect strong binding sites, they are less sensitive to degenerate motifs. We present fREDUCE, a computational method specialized for the detection of weak or degenerate binding motifs from gene expression or ChIP-chip data. fREDUCE is built upon the widely applied program REDUCE, which elicits motifs by global statistical correlation of motif counts with expression data. fREDUCE introduces several algorithmic refinements that allow efficient exhaustive searches of oligonucleotides with a specified number of degenerate IUPAC symbols. On yeast ChIP-chip benchmarks, fREDUCE correctly identified motifs and their degeneracies with accuracies greater than its predecessor REDUCE as well as other known motif-finding programs. We have also used fREDUCE to make novel motif predictions for transcription factors with poorly characterized binding sites. We demonstrate that fREDUCE is a valuable tool for the prediction of

degenerate transcription factor binding sites, especially from array datasets with weak signals that may elude other motif detection methods.

Introduction

Transcriptional regulation is modulated by a complex network of interactions between regulatory proteins and their binding targets on DNA. To comprehensively understand gene regulation at a systems level, a primary goal is to decipher the “regulatory code” that consists of knowledge of all transcriptional regulators, their DNA binding profiles, and their regulatory targets [1]. Regulatory information can be inferred from the combined analysis of genomic sequence with an abundance of microarray based methods such as ChIP-chip (chromatin immunoprecipitation on microarray)[2-3] and transcription factor perturbation experiments [4-5]. However, highly reliable regulator specificities have been unattainable for many regulators probed by such genomic-scale methods [1] since weak signals from regulators are often very difficult to isolate from experimental noise.

Thus, from a computational standpoint, a major challenge is to develop techniques that can extract maximal regulator specificity information from imperfect data. A common strategy among computational tools developed for this purpose is to first obtain a small group of genes in which a given motif may be statistically over-represented, from which the motif can then be elicited using methods such as position weight matrix updating and word enumeration [5-10]. While highly effective in some cases, a potential drawback of this approach is that the process of isolating a subgroup of sequences, typically done using clustering, cutoffs, or functional categorization, can be

arbitrary. The delineation of signal from background may be poor for noisy experimental data, where cutoffs can lead to significant loss of information. Other algorithms, such as dictionary- [11] or steganalysis-based [12] methods, do not rely on clustering but can benefit from subgroup selection.

A technique used by many motif-finding algorithms is to integrate expression data into the search process [12-14]. For example, the algorithm REDUCE (Regulatory Element Detection Using Correlation with Expression) avoids subgroup selection in a natural way by genome-wide fitting of motif counts to expression data [15]. REDUCE is a deterministic method that first enumerates oligonucleotides and then identifies words whose occurrence in promoter sequences correlate most strongly with expression data. This procedure is applied iteratively to produce a set of oligonucleotides that produce the best simultaneous fit to the data. REDUCE requires only a single expression dataset and makes use of the entire genomic dataset (both signal and background) to assess the significance of individual motifs. This method, which has already been widely applied [16-21], allows greater sensitivity to weak transcriptional signals and facilitates the discovery of combinatorial effects between regulators.

One weakness of REDUCE is that it can miss weak but biologically significant variants of the regulator site. Highly degenerate motifs whose individual variants fall below the detection threshold will be missed altogether. This is particularly the case for regulators in higher mammalian genomes, which can exhibit strong site to site variation in specificity. Thus, we have generalized the REDUCE approach to examine words containing degenerate IUPAC symbols representing multiple bases (i.e. S=C or G). However, a straightforward extension of REDUCE using exhaustive enumeration of

degenerate motifs becomes impractical when the motif length or number of degenerate positions increase. Specifically, by including m IUPAC symbols in a word of length l the motif search space increases by a factor of $\frac{l!}{m!(l-m)!} \left(\frac{11}{4}\right)^m$ where 11 is the number of IUPAC symbols (excluding A,C,G,T). For example, the computational cost is increased by 340-fold for $l=10$ and $m=2$, and by 3500-fold for $m=3$. Therefore, we have developed fast-REDUCE (fREDUCE), a significant re-implementation of the REDUCE algorithm that allows efficient searches of the extended space of degenerate motifs. We have applied fREDUCE to detect multiple motifs for transcription factor binding sites in yeast as well as human.

Results

Algorithm. The original version of REDUCE identifies motifs by exhaustively correlating all oligonucleotides up to length l in promoter sequences with expression data. However, the direct computation of the Pearson correlation coefficient is computationally laborious and is not well suited for analyzing large spaces of degenerate oligonucleotides. fREDUCE uses the following strategy to efficiently compute the Pearson coefficients of the most significant degenerate motifs (Figure 1): 1) A list of degenerate motifs that can be derived from the sequence data is generated. 2) For each degenerate motif, we can quickly compute a “pseudo-Pearson” coefficient, an estimate of the actual Pearson coefficient. The pseudo-Pearson coefficient is guaranteed to be an upper-bound on the actual Pearson coefficient and is used as a filter to eliminate most (typically >99.9%) of the motif list. 3) Actual Pearson coefficients are computed and the top motif is found and

4) The contribution from the top motif is subtracted from the expression data to form a residual, which is used for subsequent rounds of motif searching.

Performance Assessment with Yeast ChIP-chip To assess the performance of fREDUCE, we applied the algorithm to 352 ChIP-chip experiments from Harbinson *et. al.* [1] involving 203 known and putative transcription factors in the budding yeast *S. cerevisiae*. For each ChIP-chip experiment, we correlated the normalized array data to the corresponding yeast intergenic sequences, eliciting motifs of up to length 8 and containing up to 2 IUPAC degenerate symbols. In order to verify the correctness of our predictions, we compared these results to a benchmarking set consisting of 65 high confidence motif logos assembled from the predictions of six separate motif finding algorithms [1]. For 47 of 65 benchmarks fREDUCE produced an IUPAC motif that was identical to the annotated motif, including correct degeneracies (Table 1). In comparison, we ran AlignACE [22-23] on the same 65 ChIP-chip experiments. Using the same filtering and comparison criteria, we found that AlignACE detected the annotated motif for only 36 of 65 regulators. We also compared the performance of fREDUCE with those of the other 5 motif finding algorithms used to assemble the benchmark motifs (Figure 2). Even though the benchmark motifs are likely to be biased toward the six programs from which they were originally found, fREDUCE still stood out as having the best individual performance.

We also examined the performance of fREDUCE on 38 regulators for which Harbinson *et. al.* detected motifs with lower confidence. Noting that many of these 38

predicted motifs could contain inaccuracies, fREDUCE matched 7 of these predictions while alignACE matched 3.

Comparison to the original REDUCE and to MatrixREDUCE To assess the ability of fREDUCE to correctly capture motif degeneracies, we systematically compared the predictions made by fREDUCE to those made by its predecessor REDUCE on the subset of benchmark motifs containing significant degeneracy. Of 15 degenerate benchmark motifs, fREDUCE assigned IUPAC degenerate symbols identically to the benchmark in 11 cases (Table 2a). In the 4 remaining cases (HAP1, MSN2, STB5 and SUM1) fREDUCE made a prediction which is consistent with the benchmark motif while having a different degeneracy (e.g. CGGkGwTA vs. CGGwsTTA for STB5). In all of these cases, fREDUCE assigns the degenerate motif a more significant p-value than the corresponding non-degenerate motif. We note that in some cases motif degeneracies can be detected by the original REDUCE as separate motif predictions. This is especially true for regulators with strong signal (AFT2, CIN5, FHL1, GCN4, SFP1 and YAP7). However, in 5 cases degeneracies successfully predicted by fREDUCE were not detectable at all by REDUCE (CAD1, PHO4, SNT2, TEC1 and YAP1). This is typically characteristic of regulators with weaker signal.

We also compared the performance of fREDUCE to MatrixREDUCE, a recently introduced REDUCE-variant that refines motifs elicited by REDUCE into Position Specific Affinity Matrices (PSAM) [24-25]. MatrixREDUCE matched 43 of the 65 benchmarks as well as 6 of 38 motifs in the lower confidence set. In the high confidence set, six predictions were specific to fREDUCE (HAP4, HSF1, INO4, LEU3, NFG1 and

THI2) while two were specific to MatrixREDUCE (MCM1, SIP4). Specific predictions from the lower confidence set included ROX1, SWI5, UME1 for fREDUCE and PUT3, RLM1 for MatrixREDUCE. Overall, fREDUCE has a slightly stronger joint performance with 9 uniquely correct predictions from the two sets versus MatrixREDUCE's 4. In the former cases, MatrixREDUCE did not seem to begin with the correct seed, suggesting that an enumeration strategy is beneficial for some regulators. In the latter cases, fREDUCE does not find the correct motif because the long and fuzzy nature of these motifs makes them too costly for enumeration. We note that some of these differences are dependent on run parameters; with the parameters we have used MatrixREDUCE took an order of magnitude longer to run on average than fREDUCE (data not shown).

Prediction of novel motifs from yeast ChIP-chip Next we looked to see whether fREDUCE was capable of detecting novel motifs for transcription factors with uncharacterized specificities. Of the remaining transcription factors in the ChIP-chip study with no benchmark logo, we found 24 cases where fREDUCE made nontrivial (not repetitive poly-dA/dT sequences) motif predictions with p-values under 10^{-3} (Table 3). In all of these cases, there has been little to no experimental information available regarding the specificity, and existing computation methods have yielded little additional insight. Nevertheless, in a few cases we found evidence in the literature which supports the novel motif predictions we have made with fREDUCE. For example, the binding site of ARO80, a regulator of the aromatic amino acid structural genes, has been characterized in two genes as being tandem repeats of the sequences TAACCG and

TTGCCG [26]. From the ChIP-chip data, fREDUCE elicited the motif GATAACCG with high significance ($p=10^{-41}$) as well as the degenerate motif T(A/G)CCG(A/C) ($p = 10^{-5.6}$), which is similar to both of the characterized repeat elements and reflects their degeneracies. We also considered the regulator MTH1, which negatively regulates the glucose sensing signal transduction pathway by interacting with the transcriptional repressor Rgt1p [27]. Although it is unknown whether Mth1p has intrinsic DNA sequence specificity, Rgt1p has been shown to have the specificity CGGANNA [28]. fREDUCE found the matching motif GGAGRA ($p=10^{-3.57}$), which is compatible with the notion that Mth1p binds to DNA in association with Rgt1p.

Motif Elicitation in Human Hepatocytes In higher eukaryotes, motifs tend to be more degenerate and dispersed among longer intergenic regions. A common benchmark set used to evaluate the performance of computational algorithms in higher eukaryotes is the liver specific dataset [29]. Krivan et. al compiled a set of experimentally defined regulatory elements upstream of genes that were expressed exclusively in liver or in a small number of tissues including liver. From this set of genes, they found that hepatocyte-specific gene expression is mainly regulated by a small set of transcription factors (TFs), including HNF-1, HNF-3, HNF-4, and C/EBP. HNF-1, HNF-4, and C/EBP are known to be transcriptional activators based on TRANSFAC [30] annotation.

We ran fREDUCE on human adult hepatocyte expression data to capture binding sites of liver-specific transcription factors. fREDUCE captured both the forward and reverse complement of the HNF-4 binding site as well as two key degeneracies in the motif core as published in Krivan *et. al.* (Table 2b). HNF-4 is known to be linked to gene

expression in mature liver [29], which is consistent with the expression data set used in our analysis. In contrast, REDUCE was not able to capture the known binding sites, which is most likely due to the degeneracy involved in the known consensus. These results show the potential of using fREDUCE to identify regulatory elements in higher eukaryotes, including human.

Discussion

Despite the availability of powerful techniques such as ChIP-chip, the binding specificities of many transcription factors remain uncharacterized. This can be due to several reasons, including 1) regulators that have few genomic targets 2) regulators which interact weakly or indirectly with their targets and 3) regulators which bind to their maximal set of targets only under very specific environmental cues, which may be hard to find experimentally. fREDUCE offers increased sensitivity in these cases because it 1) uses the entire array data set for correlation and 2) searches all possible degeneracies. While fREDUCE is in some respects similar to motif regressor [14] and matrixREDUCE, a key distinction is that fREDUCE detects degenerate motifs *de novo* by exhaustive enumeration. In contrast, matrixREDUCE refines degeneracies from non-degenerate seeds and motif regressor selects among candidate matrices using correlation with expression. Thus, fREDUCE may be advantageous when motifs are difficult to detect in a non-degenerate form or are missed in the candidate set.

By comparison to 65 benchmark logos in yeast, we see that fREDUCE is comparable to or greater in detection power versus algorithms like AlignACE for strong motifs that are relatively easy to detect. Even in these cases, fREDUCE outperforms the

original REDUCE algorithm by accurately predicting known degeneracies. The most advantageous use of fREDUCE, however, is for the detection of weak motifs which may lie at the border of detection. It is difficult to verify the correctness of many of the motifs elicited in these cases because of their poor characterization. Nevertheless, we have found two cases where fREDUCE was sensitive to subtle signals: ARO80, for which sites are highly degenerate, and MTH1, which may have a weak signal due an indirect interaction with DNA. We have also shown that fREDUCE is capable of capturing the HNF-4 binding site in hepatocytes, demonstrating that this algorithm is generally applicable to the detection of degenerate motifs in mammalian cells.

Conclusions

We have presented the motif prediction algorithm fREDUCE, a refined variation of REDUCE specialized for the detection of degenerate motifs. The two primary strengths of fREDUCE are 1) it maximizes data utilization by fitting all expression data and 2) it searches motif degeneracies in a comprehensive and unbiased way. We have shown that fREDUCE is an improvement upon the existing REDUCE algorithm for degenerate binding profiles and that it can outperform existing motif finding methods on yeast CHIP-chip benchmarks. Furthermore, fREDUCE is able to detect degenerate signals in yeast and human. Thus, fREDUCE should be a valuable computation tool for the detection of subtle motifs.

Methods

Algorithm.

The pearson correlation between expression values and counts of a possibly degenerate motif D is given by:

$$P(D) = \frac{\sum_{i=1}^G (E_i - \bar{E})(n_i^D - \bar{n}^D)}{\sqrt{\sum_{i=1}^G (E_i^2 - \bar{E}^2)} \cdot \sqrt{\sum_{i=1}^G (n_i^D n_i^D - \bar{n}^2)}}$$

Where i is an index over genes, E_i is the expression of gene i , n_i^D is the number of motif counts matching D in sequence i , \bar{n} is the average of n_i^D over all genes and G is the total

number of genes. Let g_i be the normalized gene expression: $g_i = \frac{E_i - \bar{E}}{\sqrt{\sum_{i=1}^G (E_i^2 - \bar{E}^2)}}$, so

that $\sum_{i=1}^G g_i = 0$ and $\sum_{i=1}^G g_i^2 = 1$. Then the Pearson coefficient reduces to:

$$P(D) = \frac{\sum_{i=1}^G g_i n_i^D}{\sqrt{\sum_{i=1}^G (n_i^D n_i^D - \bar{n}^2)}}$$

Since $n_i^D = \sum_S n_i^S$, where the sum is over all non-degenerate nucleotide motifs S that match

D , we can pre-compute and store a table of $\sum_{i=1}^G g_i n_i^S$ for all S and readily construct the

numerator of $P(D)$ for any D . However, the denominator is not linear in n_i^D and cannot

be expressed as a sum over S . Nevertheless we can compute a pseudo-Pearson

coefficient:

$$\tilde{P}(D) = \frac{\sum_{i=1}^G g_i n_i^D}{\sqrt{\tilde{n}^2 - G\bar{n}^2}}$$

where $\tilde{n}^2 = \sum_S \sum_{i=1}^G n_i^S n_i^S$ can be constructed as a sum over S .

Since $\sum_{i=1}^G n_i^D n_i^D = \sum_{i=1}^G \left(\sum_{S_1} n_i^{S_1} \right) \left(\sum_{S_2} n_i^{S_2} \right) = \sum_{S_1} \sum_{S_2} \sum_{i=1}^G n_i^{S_1} n_i^{S_2} \geq \tilde{n}^2$, we have $|P(D)| \leq |\tilde{P}(D)|$.

Hence the magnitude of pseudo-Pearson coefficient is an upper bound for the magnitude of the actual Pearson coefficient, allowing rapid screening of all degenerate motifs.

Actual Pearson values can then be computed for a small subset of motifs with pseudo-Pearson values above a given threshold. This scheme is effective except for motifs where $\tilde{n}^2 < G\bar{n}^2$, in which case the Pearson coefficient must be computed directly. Thus, fREDUCE will give a computational advantage as long as the average motif count \bar{n} is less than one.

Specifically, fREDUCE uses the following procedure:

- (1) For each oligonucleotide string S of length L that appears in the sequence, we pre-

compute the quantities $p_d^S = \sum_{i=1}^G g_i n_i^S$, $\bar{n}^S = \frac{1}{G} \sum_{i=1}^G n_i^S$, and $\overline{n^2}^S = \sum_{i=1}^G n_i^S n_i^S$

- (2) We generate a list of all possible nucleotides containing up to l degeneracies matching the set of S .
- (3) We rapidly compute corresponding quantities for all degenerate strings D

matching S : $p_d = \sum_{i=1}^G g_i n_i^D = \sum_S p_d^S$, $\bar{n} = \frac{1}{G} \sum_{i=1}^G n_i^D = \sum_S \bar{n}^S$, and

$\tilde{n}^2 = \sum_S \sum_{i=1}^G n_i^S n_i^S = \sum_S \overline{n^2}^S$ and use them to construct the pseudo-Pearson

coefficient $p_D / \sqrt{\tilde{n}^2 - G\bar{n}^2}$. We save only those motifs whose pseudo-Pearson coefficients exceed a threshold corresponding to the p-value cutoff for its motif

class. For the motifs whose pseudo-Pearson coefficients cannot be calculated directly (because $\tilde{n}^2 \leq G\bar{n}^2$), we compute the true Pearson.

- (4) We sort the remaining motifs in decreasing order of the magnitudes of their pseudo-Pearson and compute true Pearson coefficients in this order. We stop computing when the magnitude of the pseudo-Pearson value of the current motif in the list falls below the magnitude of the true Pearson coefficient of the top motif.
- (5) Finally, we compute the residual gene expression $\tilde{g}_i = g_i - P(D)n_i^D$, that is, the expression data after the effect of motif D has been taken into account. After a renormalization, the residual is used to carry out subsequent rounds of motif finding.

To estimate the statistical significance of motifs, we note that since $|P(D)| \ll 1$, its distribution is well approximated by a Normal distribution. We convert $P(D)$ into a z-score:

$$Z(D) = P(D) \sqrt{\frac{G-2}{1-P(D)^2}}$$

This z-score is used to derive the p-value [15]:

$$pvalue = \frac{2}{\sqrt{2\pi}} \int_{Z(D)}^{\infty} e^{-\frac{t^2}{2}} dt$$

To correct for multiple testing, we first apply a Bonferroni correction factor of $\binom{L}{m} D^m 4^{L-m}$ to each motif of length L containing m IUPAC symbols. This factor corresponds to the total number of motifs in the class of L and m , where $D=11$ or 15 depending on whether 3-fold IUPAC symbols are included. We then apply a second

correction factor for the total number of motif classes examined for a particular run. For example, with the settings ($L=7$ and $m=1$) we would examine all motifs in the classes (6,0), (6,1), (7,0) and (7,1) giving a second correction factor of 4 for each motif (we require a minimum motif length of 6). This weighted method of correction has the advantage of accounting for the fact that motif classes with larger values of L and m tend to give higher numbers of false positives.

fREDUCE performance testing. We ran fREDUCE on the REB1_YPD ChIP-chip data from Harbison et. al. [1] for varying L and m on an 2.40 GHz Intel Xeon processor. In all runs, the known Reb1p binding site CGGGTAA or close variants appeared as the top motif (data not shown).

Motif Detection from Yeast ChIP-chip. We applied fREDUCE to 354 yeast ChIP-chip experiments involving 203 known and putative transcription factors [1]. Each experiment was analyzed with fREDUCE using the corresponding set of yeast intergenic sequences, searching all motifs up to length 8 containing up to 2 two-fold IUPAC degenerate symbols. We filtered the set of motifs found for each fREDUCE run by three criteria. First, since yeast intergenic sequences have relatively low G/C content, we eliminated motifs which only contained the letters A/T/W as such motifs tend to have inflated correlation coefficients. From the remaining list of motifs, we chose the top three most significant motifs for further comparison. Accounting for the fact that we are eliciting motifs from several hundred experiments, we also discarded motifs with corrected p-values less significant than 10^{-2} . If the given transcription factor was

associated with CHIP-chip data under multiple environmental conditions, then filtered motifs from all conditions were combined and the top three chosen. The final motifs for each transcription factor were compared to reference motifs predicted by Harbinson *et. al.* based on a composite of several motif finding algorithms [1]. We extracted IUPAC representations of reference motifs from [31], which contained 102 specificities of which 65 were considered high confidence. Each reference motif was compared to their corresponding fREDUCE predictions using a sliding window string comparison. Predicted motifs are considered a match if there is at least one window where all IUPAC characters are consistent between both strings. Motif predictions made for transcription factors with no reference motifs were compared to literature.

Comparison to non-degenerate REDUCE. From the 65 high confidence benchmarks, we selected cases where the annotated motif had at least one IUPAC character. In 15 of these cases, both fREDUCE and REDUCE made correct, if not correctly degenerate predictions. In 11 of these 15 cases fREDUCE made the correct IUPAC assignments. For each of these 11 cases, we considered whether the degeneracy can be assembled from non-degenerate motifs with $p < 0.01$ predicted by REDUCE.

Comparison to other motif-finding algorithms. We obtained the alignACE package and ran all CHIP-chip data with the default parameters using probes with p-values below 0.001. The output alignment was converted into an IUPAC string using the method described by Cavener *et. al.* [32] and the resulting motifs were compared to reference motifs in the same way as the fREDUCE motif predictions. Details of alignACE motifs

found and comparisons to alignACE motifs from Harbison *et al.* are available in Supp. Table 1. We also obtained MatrixREDUCE [33] and ran all ChIP-chip data against the provided yeast sequence file Y5_600_Bst.fa. Default parameters were used except that we set max_motif=10 for consistency with our fREDUCE runs. For the other five algorithms, we tallied the total number of references to each algorithm from the list of matrices on Harbison *et al.* supporting website [34].

Motif Detection from Human Liver Tissue. 158 custom made Affymetrix gene expression arrays for 79 different human tissues (2 replicates each) were obtained from Novartis in a publicly available database [35-36]. The arrays were normalized using gcrma [37-38] and the probes were annotated using Ensembl gene annotation [39] for human build 35. To study adult liver specific gene expression, we first normalized expression values for each liver tissue replicate against the average expression of all other tissues (excluding the 2 liver tissue experiments) The expression value of each gene in liver tissue experiments is represented as the following z-score:

$$z^{n,g} = \frac{E^{n,g}_{liver} - \mu^{g}_{other}}{\sigma^{g}_{other}}$$

Where n is the liver tissue experiment replicate number, g is the index over genes, E^{ng}_{liver} is the expression value of gene g in replicate n , μ_{other}^g is the mean expression value of gene g in non-liver tissue experiments, and σ_{other}^g is the standard deviation of gene g in non-liver tissue experiments.

Human genomic sequences (build 35) were extracted 1000bp upstream from the transcriptional start site (TSS) if known, or from the initiation codon, based on Ensembl v35 [40]. The repeat masked promoter sequences were mapped to corresponding z-scores, which represent gene expression. This resulted in a final set of 11,710 paired z-scores and promoter sequences for input into fREDUCE. We then ran fREDUCE on the z-scores for each replicate of the liver tissue on the basis that a higher z-score translates to higher expression in liver tissues compared to the other tissues. Two different sets of parameters were run on each replicate as follows: length 8 with 0 IUPAC symbols and length 8 with 2 IUPAC symbols.

Software Availability and Requirements

- **Project Name:** fREDUCE
- **Project Home Page:** <http://genome3.ucsf.edu:8080/freduce>
- **Operating system:** Linux
- **Programming languages:** C++

Source code and example usage are included in the release file fREDUCE-1.0.tar.gz.

References

1. Harbison CT, Gordon DB, Lee TI, Rinaldi NJ, Macisaac KD, Danford TW, Nannett NM, Tagne JB, Reynolds DB, Yoo J, Jennings EG, Zeitlinger J, Pokholok DK, Kellis M, Rolfe PA, Takusagawa KT, Lander ES, Gifford DK, Fraenkel E, Young RA: **Transcriptional regulatory code of a eukaryotic genome.** *Nature* 2004, **431**: 99-104.
2. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **290**: 2306–2309.
3. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, Hannett NM, Harbison CT, Thompson CM, Simon I, Zeitlinger J, Jennings EG, Murray HL, Gordon DB, Ren B, Wyrick JJ, Tagne JB, Volkert TL, Fraenkel E, Gifford DK, Young RA: **Transcriptional regulatory networks in *Saccharomyces cerevisiae*.** *Science* 2002, **298**: 799–804.
4. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, Kidd MJ, Meyer MR, Slade D, Lum PY, Stepaniants SB, Shoemaker DD, Gachotte D, Chakraburttty K, Simon J, Bard M, Friend SH: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102**: 109-26.
5. Wang W, Cherry M, Botstein D, Li H: **A systematic approach to reconstructing transcription networks in *Saccharomyces cerevisiae*.** *Proc Natl Acad Sci U S A* 2002, **26**: 16893-98.
6. Bailey TL, Elkan C: **Fitting a mixture model by expectation maximization to discover motifs in biopolymers.** *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 1994, **2**: 28-36.
7. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **262**: 208-14.
8. Wang T, Stormo GD: **Combining phylogenetic data with co-regulated genes to identify regulatory motifs.** *Bioinformatics* 2003, **19**: 2369–2380.
9. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J. Mol. Biol.* 1997, **281**: 827-42.
10. Liu X, Brutlag DL, Liu JS: **BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.** *Proc Pac Symp Biocomput.* 2001, 127-38.
11. Bussemaker HJ, Li H, Siggia ED: **Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis.** *Proc Natl Acad Sci U S A* 2000, **97**: 10096-100.
12. Wang G, Zhang W: **A steganalysis-based approach to comprehensive identification and characterization of functional regulatory elements.** *Genome Biol.* 2006, **7**: R49.
13. Beer MA, Tavazoie S: **Predicting gene expression from sequence.** *Cell* 2004, **177**: 185-98.

14. Conlon EM, Liu XS, Lieb JD, Liu JS: **Integrating regulatory motif discovery and genome-wide expression analysis.** *Proc. Natl. Acad. Sci. USA* 2003, **100**: 3339-44.
15. Bussemaker HJ, Li H, Siggia ED: **Regulatory element detection using correlation with expression.** *Nat Genet.* 2001, **27**:167-71.
16. Kim K, Duncan K, Li H, Guthrie C: **Functional specificity of shuttling hnRNPs revealed by genome-wide analysis of their RNA binding profiles.** *RNA* 2005, **11**: 383-93.
17. Wang W, Cherry JM, Nochomovitz Y, Jolly E, Botstein D, Li H: **Inference of combinatorial regulation in yeast transcriptional networks: A case study of sporulation.** *Proc Natl Acad Sci U S A* 2005, **102**: 1998-2003.
18. Klebes A, Sustar A, Kechris K, Li H, Schubiger G, Kornberg TB: **Regulation of cellular plasticity in Drosophila imaginal disc cells by the Polycomb group, trithorax group and lama genes.** *Development* 2005, **132**: 3753-65.
19. Koerkamp MG, Rep M, Bussemaker HJ, Hardy GP, Mul A, Piekarska K, Szigyarto CA, De Mattos JM, Tabak HF: **Dissection of transient oxidative stress response in Saccharomyces cerevisiae by using DNA microarrays.** *Mol. Biol. Cell* 2002, **13**: 2783-2794.
20. Orian A, van Steensel B, Delrow J, Bussemaker HJ, Li L, Sawado T, Williams E, Loo LW, Cowley SM, Yost C, Pierce S, Edgar BA, Parkhurst SM, Eisenman RN: **Genomic binding by the Drosophila Myc, Max, Mad/Mnt transcription factor network.** *Genes Dev.* 2003, **17**: 1101-14.
21. van Steensel BD, Bussemaker HJ: **Genome-wide analysis of Drosophila GAGA factor target genes reveals context-dependent DNA binding.** *Proc. Natl. Acad. Sci. USA* 2003, **100**: 2580-85.
22. Roth FR, Hughes JD, Estep PE, Church GM: **Finding DNA regulatory motifs within unaligned non-coding sequences clustered by whole-genome mRNA quantitation.** *Nature Biotech.* 1998, **16**: 939-45.
23. **AlignACE Homepage** [<http://atlas.med.harvard.edu>]
24. Foat BC, Morozov AV, Bussemaker HJ: **Statistical mechanical modeling of genome-wide transcription factor occupancy data by MatrixREDUCE.** *Bioinformatics* 2006, **22**:141-9.
25. Foat BC, Houshmandi SS, Olivas WM, Bussemaker HJ: **Profiling condition-specific, genome-wide regulation of mRNA stability in yeast.** *Proc. Natl. Acad. Sci. USA* 2005, **102**: 17675-80.
26. De Rijcke M, Seneca S, Punyammalee B, Glansdorff N, Crabeel M: **Characterization of the DNA target site for the yeast ARGR regulatory complex, a sequence able to mediate repression of induction by arginine.** *Mol. Cell Biol.* 1992, **12**: 68-81.
27. Lakshmanan J, Mosley AL, Ozcan S: **Repression of transcription by Rgt1 in the absence of glucose requires Std1 and Mth1.** *Curr. Genet.* 2003, **44**: 19-25.
28. Kim JH, Polish J, Johnston M: **Specificity and regulation of DNA binding by the yeast glucose transporter gene repressor Rgt1.** *Mol. Cell Bio.* 2003, **23**: 5208-16.
29. Krivan W, Wasserman WW: **A predictive model for regulatory sequences directing liver-specific transcription.** *Genome Res.* 2001, **11**: 1559-1566.

30. Matys V, Fricke E, Geffers R, Gossling E, Haubrock M, Hehl R, Hornischer K, Karas D, Kel AE, Kel-Margoulis OV, Kloos DU, Land S, Lewicki-Potapov B, Michael H, Munch R, Reuter I, Rotert S, Saxel H, Scheer M, Thiele S, Wingender E: **TRANSFAC: transcriptional regulation, from patterns to profiles**. *Nucleic Acids Res*. 2003, **31**: 374-378.
31. **Final Motifs**
[http://fraenkel.mit.edu/Harbison/release_v24/final_set/Final_InTableS2_v24.motifs]
32. Cavener DR: **Comparison of the consensus sequence flanking translational start sites in Drosophila and vertebrates**. *Nucleic Acids Res*. 1987, **15**: 1353-61.
33. **MatrixREDUCE Homepage**.
[<http://bussemaker.bio.columbia.edu/software/MatrixREDUCE/>]
34. **The Fraenkel Lab – Harbison et al. Final Motif Logos**
[http://fraenkel.mit.edu/Harbison/release_v24/final_set/Final_motifs/]
35. Su AI, Wiltshire T, Batalov S, Lapp H, Ching KA, Block D, Zhang J, Soden R, Hayakawa M, Kreiman G, Cooke MP, Walker JR, Hogenesch JB: **A gene atlas of the mouse and human protein-encoding transcriptomes**. *Proc Natl Acad Sci U S A* 2004, **101**: 6062-67.
36. **GNF SymAtlas** [<http://symatlas.gnf.org/SymAtlas/>]
37. Wu Z, Irizarry RA: **Preprocessing of oligonucleotide array data**. *Nat Biotechnol* 2004, **22**: 656-658; author reply 658.
38. **A Model Based Background Adjustment for Oligonucleotide Expression Arrays** [<http://www.bepress.com/jhubiostat/paper1/>]
39. Hubbard, T. *et al.*: **Ensembl 2005**. *Nucleic Acids Res* 2005, **33**: D447-453.
40. Curwen V, Eyraas E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M: **The Ensembl automatic gene annotation system**. *Genome Res* 2004, **14**: 942-950.

Tables

Table 1. fREDUCE predictions from 65 yeast ChIP-chip experiments of Harbinson *et al.* Check marks (√) indicate that fREDUCE matched the IUPAC string corresponding to the benchmark logo. The results of a similar analysis for AlignACE is given in the right column.

Factor	Known Site	Condition	Motif	p-value	fREDUCE match?	AlignACE Match?
ABF1	rTCAYt....Acg	YPD	rTGATm	22.4	√	√
ACE2	tGCTGGT	YPD	kGCTGGy	6.2	√	
AFT2	GGGTGy	H2O2Lo	rGGTGy	91.5	√	√
AZF1	YwTTkcKkTyycgykky	YPD	mTTTTw	14.8		
BAS1	TGACTC	YPD	TGACTCCG	37.2	√	√
CAD1	mTTAsTmAkC	YPD	GmTTAsTA	4.2	√	√
CBF1	tCACGTG	YPD	CACGTG	90.7	√	√
CIN5	TTAygTAA	YPD	TTAyrTAA	59.4	√	√
DAL82	GATAAGa	RAPA	GATAAG	9.4	√	
DIG1	TgAAAca	YPD	TGAAACA	18	√	
FHL1	rTGTayGGrtg	YPD	GTayGGrT	141.2	√	√
FKH1	tTgTTTAc	YPD	yTGTTkAC	28.8	√	
FKH2	aaa.GTAAACaAa	YPD	GTAAACA	23.7	√	√
GAL4	CGG.....cCg	YPD	TTCGGAGC	4.9		√
GAT1	aGATAAG	RAPA	GATAAG	13.3	√	
GCN4	TGAsTCa	YPD	rTGAsTCA	166.7	√	√
GLN3	GATAAGa.a	RAPA	GATAAG	38.2	√	
HAP1	GGmraTA.CGs	YPD	kTTATCGG	60.3	√	√
HAP4	g.CcAAAtcA	YPD	CCAATsAr	21.7	√	√
HSF1	TTCya.....TTC	H2O2Hi	TTCyrGAA	109.5	√	√
IME1		H2O2Hi				
INO2	CAcaTGc	YPD	kCACATGC	12.8	√	
INO4	CATGTGaaaa	YPD	CAyrTG	89.2	√	√
LEU3	cCGgtacCGG	YPD	CGGkACCG	10.8	√	√
MBP1	rACGCGt	YPD	ACGCGT	126.9	√	√
MCM1	tttCC.rAt..gg	Alpha	yTTCCTAA	5.7		√
MET4	RMmAwsTGKSgyGsc	SM	CrCGyG	14.8		
MSN2	mAGGGGsgg	H2O2Hi	rGGGGy	20.8	√	
NDD1	tt.CC.rAw..GG	YPD	CTCGAGGC	12.3		√
NRG1	GGaCCCT	YPD	AGGGTCs	11.3	√	√
PDR1	ccGCCgRAWra	YPD	CCrwACAT	11.4		
PHD1	sc.GC.gg	YPD	mTGCAk	21.1		√
PHO2	SGTGCGsygyG	Pi-				
PHO4	CACGTGs	Pi-	sCACGTGs	14.1	√	
RAP1	tGyayGGrtg	SM	GyrTGGGT	57.1	√	√
RCS1	ggGTGca.t	H2O2Lo	GGGTGCA	43.6	√	√
RDS1	kCGGCCGa	H2O2Hi	TCCGCGG	35.6	√	
REB1	CGGGTAA	YPD	CGGGTAAy	136.7	√	√
RFX1	TTgccATggCAAC	YPD	GTCGTCCG	3.2		√
RLR1	ATTTTCttCwTt	YPD				
RPN4	TTTGCCACC	H2O2Lo	TyGCCACC	109.8	√	√

SFP1	ayCcrTACay	SM	yCCrTACA	31.6	√	√
SIG1	ArGmAwCrAmAA	H2O2Hi				
SIP4	CGG.y.AATGGrr	SM	CTCGGCC	58.4		
SKN7	G.C..GsCs	H2O2Lo	GsCyGGCC	37.7	√	
SNT2	yGGCGCTAyca	YPD	GrTAGCGC	96.1	√	√
SOK2	tGCAG..a	BUT14	GGTrCAGA	5.6		
SPT2	ymtGTmTytAw	YPD	TkyATA	6.2		
SPT23	rAAATsaA	YPD	wTkAAA	25.1		
STB1	rracGCsAaa	YPD	wCGCGT	4	√	
STB4	TCGg..CGA	YPD	CGGryCGA	7.1	√	√
STB5	CGGwstTAta	YPD	CGGkGwTA	24	√	
STE12	tgAAACa	YPD	TGAAACA	38.9	√	√
SUM1	gyGwCAswaaw	YPD	GyGTCAs	25.0	√	√
SUT1	gcsGsg..sG	YPD	wCkCCG	49.8		
SWI4	raCgCsAAA	YPD	CGCsAAAA	12.6	√	√
SWI6	tttcGCGt	YPD	TTTCsk	11.6	√	
TEC1	rrGAATG	YPD	rrGAATGT	22.4	√	
THI2	gmAAcy.twAgA	Thi-	GGAAACyS	4.5	√	
TYE7	tCACGTGAy	YPD	TCACGTGr	70.8	√	√
UME6	taGCCGCCsa	YPD	GCsGCy	154.3	√	√
YAP1	TTaGTmAGc	YPD	mTkACTAA	13.6	√	√
YAP7	mTkAsTmAk	H2O2Hi	mTTAsTAA	121.9	√	√
YDR026c	ttTACCCGm	YPD	CCGGGTAA	23.2	√	√
ZAP1	ACCCTmAAGGTyrT	YPD	wAyATT	16.5		

Table 2a: fREDUCE predictions in comparison to non-degenerate predictions made by REDUCE. Benchmark logos and their corresponding motifs are shown for reference. P-values are shown as $-\log_{10}$ values.

TF	REDUCE (p-value)	fREDUCE (p-value)	Benchmark Logo	Benchmark Motif
AFT2	GGGTGC (61.8) GGGTGT (31.6)	GGGTGy (91.5)		GGGTGy
CAD1	ATTAGTA (2.9) -	GmTTAsTA (4.2)		mTTAsTmAkC
CIN5	TATGTAA (17.8) TACGTAA (15.6)	TTAyrTAA (59.4)		TTAyGTAA
FHL1	TGTACGG (59.4) GTATGGG (30.5)	GTAYGGrT (159.7)		rTGTayGGrt
GCN4	TGACTCA (103.3) GAGTCAT (36.4)	rTGAsTCA (166.7)		TGAsTCA
HAP1	TATCGG (38.8) -	kTTATCGG (60.3)		GGmraTA.CGs
MSN2	AAGGGG (8.6) -	rGGGGy (20.8)		mAGGGGsgg
PHO4	CACGTGC (6.4) -	sCACGTGs (14.1)		CACGTGS
SFP1	CCGTACA (12.2) CCCATAC (10.4)	yCCrTACA (31.6)		ayCcrTAcay
SNT2	GGCGCTA (49.7) CGCTATC (7.0)	GCGCTAyC (96.1)		yGGCGCTAyca
STB5	CGGTGTT (7.0) -	CGGkGwTA (24.0)		CGGwstTAta
SUM1	TGTCAC (11.4) TGACAC (8.9)	GwCAGTAA (25.0)		gyGwCAswaa
TEC1	AGAATG (13.0) -	rrGAATGT (22.4)		rrGAATG
YAP1	ATTAGT (10.9) -	TTAGTmAk (13.6)		TTaGTmAkC
YAP7	TTACTAA (50.1) TTAGTAA (41.7) TGACTAA (15.9)	TTAsTAAk (118.6)		mTkAsTmAk

Table 2b: fREDUCE elicitation of the HNF-4 binding site from human hepatocyte expression data.


TF	REDUCE (p-value)	fREDUCE (p-value)	Benchmark Logo	Benchmark Motif
HNF-4	-	GRMCTTTG (7.4)		TGrmCTTTG

Table 3. fREDUCE predictions for regulators with poorly characterized specificities. We searched the literature for evidence supporting our motif predictions and the matching examples are highlighted. *The annotated motifs for Rgt1p.

<u>Regulator</u>	<u>Predicted Site</u>	<u>P-value</u>	<u>Motif from Literature Search</u>
ARG80	TTYTCY	34.3	CYNYAAANKRMAR
ARO80	TRCCGM	5.6	TWRCCG
ASK10	AYTTKA	9.1	
CST6	TYAAWA	7.0	
DAT1	WTTSAA	16.7	
ECM22	GCRSCC	16.2	TCGTATA
EDS1	TWTTSA	8.4	
FAP7	WTRAAG	11.3	
GAT3	CCTSGGC	15.2	
GCR2	TTCAWW	5.0	CTTCC
HAL9	WTTRAA	14.7	
HIR3	WTTRAA	22.0	ACGCTAAA
IME4	YACACAC	17.8	
MAL13	CCASSG	11.6	
MAL33	GRCAS	13.8	
MET18	WTTCAA	8.2	
MGA1	TTTRAY	5.9	
MSN1	MMCCCA	3.8	
MTH1	GGAGRA	3.4	CGGANNA *
OAF1	CGCASY	4.9	CGGNNNTNAN ₉₋₁₂ CCG
RGM1	CSGSCC	27.1	
RTG1	ATYTRA	10.3	
SIP3	WTCAAW	7.6	
SMK1	WTGWAG	3.9	
STB2	CAAGGYC	3.1	
STB6	TATSAW	5.6	
STP4	AARMTT	24.1	
TOS8	RCACMC	20.7	
UPC2	MATSAA	4.5	
WAR1	TYAAGW	6.6	
YBR239c	WATAYT	16.8	
YDR049W	AWTGAW	3.5	
YER051w	AKYACT	3.9	
YER130C	CAARTW	3.1	
YFL052w	WTCAAK	3.6	
YGR067C	TTYAAW	4.6	
YKR064W	WGTTTRA	6.3	
YLR278C	KTTMAA	7.2	
YML081W	WCAAMT	3.7	
YNR063W	TCAARTA	2.4	
YPR196W	WTCAAW	10.3	

Figures

Figure 1. The fREDUCE algorithm. A set of possible IUPAC strings are generated from the input sequence. For each IUPAC string, we compute a pseudo-Pearson coefficient, which is an estimate and upper bound on the true Pearson coefficient. After the vast majority of motifs are filtered out using the pseudo-Pearson value, we then compute true Pearson coefficients for the remaining motifs and select the top motif. The residual expression value is then used to iteratively derive subsequent motifs.

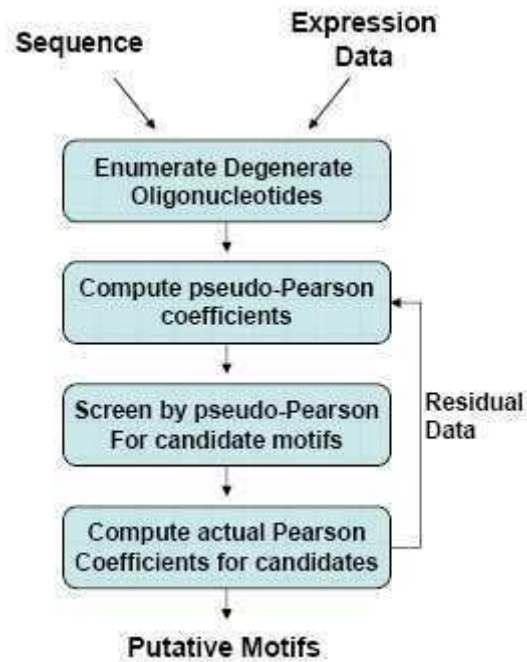


Figure 2. Comparison of fREDUCE to six other algorithms on 65 yeast ChIP-chip benchmarks. AlignACE* indicates results of running AlignACE from scratch, while the performance of other methods were compiled from the Harbison *et. al* supporting website.

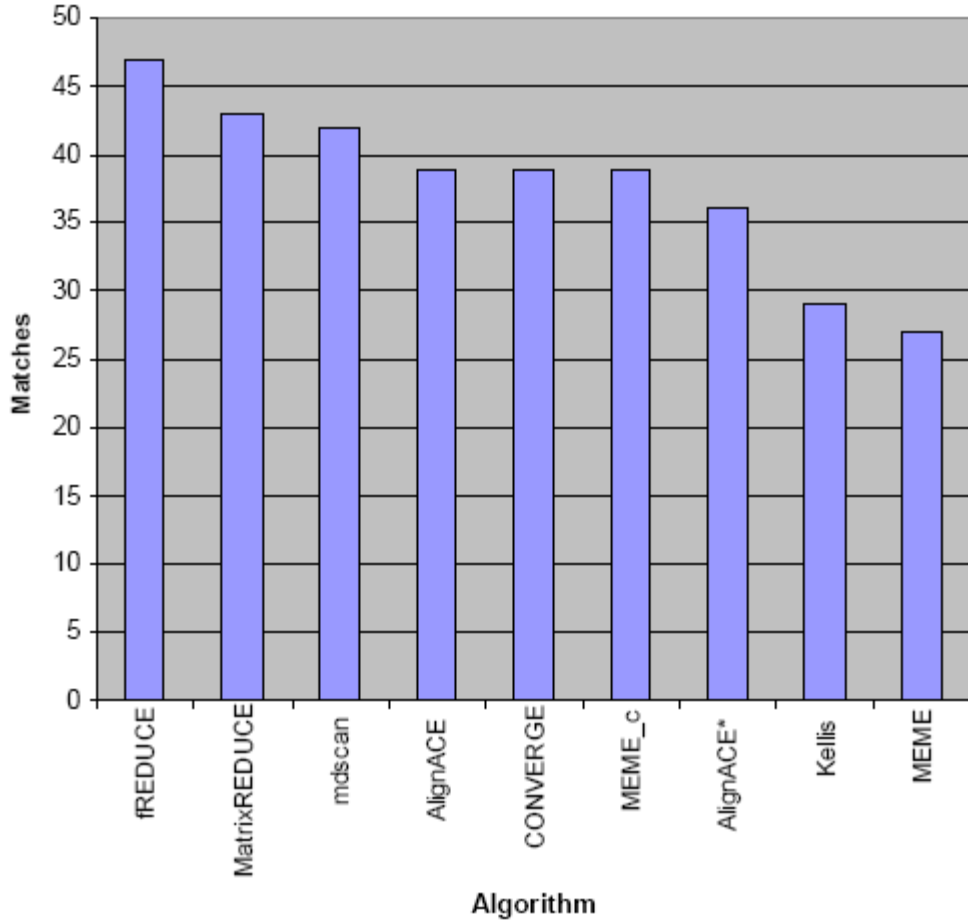
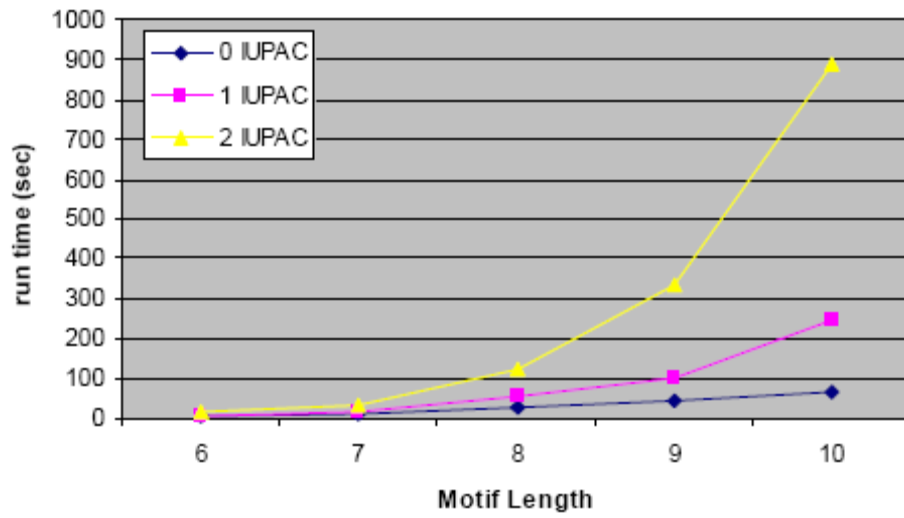


Figure 3. Scalability of fREDUCE. The performance of fREDUCE on yeast ChIP-chip experiment REB1_YPD for various motif lengths and numbers of degeneracies.



Chapter 2

Directed A/T-tracts:

A Novel Signature for Yeast Nucleosome Free Regions

Abstract

Eukaryotic transcriptional regulation is mediated by the organization of nucleosomes in promoter regions. A highly stereotyped chromatin organization is seen in most *S. cerevisiae* promoters, where nucleosome-free regions (NFR) are flanked by well-ordered nucleosomes. By analyzing groups of promoters with varying nucleosome occupancy patterns, we found that yeast promoters with well-defined NFRs are characterized by positioned patterns of poly(dA:dT) tracts with two signature features. First, poly(dA:dT) tracts are highly localized in a strand-dependent fashion where poly(dA) tracts lie proximal to transcriptional start sites and poly(dT) tracts are distal. Collectively the inverted tracts define an axis of symmetry coinciding with NFR centers. Second, poly(dA:dT) tracts exhibit a novel “capping” effect where tracts preferentially terminate in G:C residues in a direction-dependent manner. In NFRs, capping is greatly increased and is localized to the poly(dA:dT) symmetric axis. Both signature features quantitatively co-vary with fine positional variations between NFRs, establishing a closely-knit relationship between poly(dA:dT) tracts, their capping patterns, and the central coordinates of NFRs. Based on our data, we hypothesize that localized stretches of short poly(dA:dT) tracts constitute directional signals in yeast promoters which

facilitate NFR placement in a manner independent of specific transcription factors. We present a model of NFR origination in yeast in which directed poly(dA:dT) tracts contribute to the definition of a central NFR nucleation site, and provide data which distinguishes this model from an alternative model where tracts act as boundary elements that anchor flanking nucleosomes.

Introduction

Eukaryotic DNA is packaged as chromatin: highly organized arrays of nucleosomes which profoundly affect the functions of underlying sequence[1,2,3]. Because chromatin structure plays critical regulatory roles[4], promoter sequences must not only dictate their own regulatory logic but also coordinate the patterns of nucleosomes that are superimposed upon them. Current hypotheses describing how genomic sequence is mapped to nucleosome positioning are encompassed by two paradigms. Because DNA affinities to the histone core can differ over a 1000-fold range depending on sequence[5], one view is that *in vivo* nucleosome positions can largely be specified by the thermodynamic preferences of nucleosomes for genomic DNA[6]. Recent efforts have attempted to deduce nucleosome positioning from periodic dinucleotide patterns that confer physical properties favorable for the sharp DNA bending required for incorporation into nucleosomes[6,7], but their predictive power over random guessing is modest [8,9,10]. An contrasting view is that a small number of strategically positioned nucleosomes can serve as boundaries against which other nucleosomes fall into place through statistical packing[11,12,13,14]. The positions of these key “barrier” nucleosomes must be highly regulated, and is likely to involve a combination of *cis*

acting sequences that work through intrinsic DNA-histone interactions as well as *trans* acting sequences that signal transcription factors and chromatin remodeling complexes[2,15]. The balance between these two paradigms is not clear and is likely to depend on genomic context.

The barrier nucleosome paradigm is likely to be especially relevant in promoters where nucleosomes are organized around nucleosome-free regions (NFR), spans of nucleosome-deficient sequence emanating in the 5' direction from transcriptional start sites (TSS)[16]. Prevalent in yeast[10,16,17], fly[18] and human[19], NFRs appear to be a conserved mode of promoter nucleosomal organization in most eukaryotes. In *S. cerevisiae*, NFRs appear in up to 95% of promoters[11], have a typical span of ~140bp[10,17] and incorporate the histone variant H2A.Z into flanking nucleosomes[20,21]. It has been suggested that the highly defined nucleosomes flanking NFRs anchor a large part of nucleosome organization in the *Saccharomyces* genome using the barrier nucleosome principle[11]. The key question remains, however, of how the positions of the “keystone” boundary nucleosomes are specified through sequence.

High-throughput nucleosome mapping studies in yeast have universally linked nucleosome-free regions with the enrichment of poly(dA:dT) tracts[10,16], contiguous stretches of homopolymeric dA or dT that are over-represented in the intergenic regions of many eukaryotes[22,23]. In addition, two recent computational approaches to nucleosome prediction have found that poly(dA:dT) tracts as short as length 3 have significant discriminative power in distinguishing nucleosomal vs. non-nucleosomal sequence[9,24]. This association between NFRs and poly(dA:dT) tracts is typically attributed to the latter's physical rigidity[25], which is thought to destabilize nucleosomes.

However, the impact of poly(dA:dT) rigidity on nucleosome positioning has not been clearly demonstrated: long tracts ($l \geq 10$) only modestly destabilize nucleosomes *in vitro*[26], and the *in vivo* impact of shorter tracts is not known.

In this study we explore in detail two novel characteristics of poly(dA:dT) tracts in the *S. cerevisiae* genome and their relationships with patterns of nucleosome-free regions. First, we show that positional distributions of poly(dA) tracts and their inverse poly(dT) tracts are related by symmetry across the central NFR axis. Second, we demonstrate that poly(dA:dT) tracts in NFRs exhibit an oriented, terminal specific “capping” by G:C. We demonstrate not only that both features are specific to NFR-containing promoters, but also that they co-vary with fine variations among NFRs of different sizes and localizations. The highly organized placement of poly(dA:dT) tracts in promoters, their orientation-specific terminal characteristics and their intricate correlations with NFRs suggest that directed poly(dA:dT) tracts may constitute signature sequences which influence NFR placement. Models of how poly(dA:dT) tracts may guide NFR formation and mechanistic implications are discussed.

Results

Promoters classification into “strong” and “weak” NFR classes

We re-examined the genome-wide nucleosome positioning map presented by Lee et al[10], aligning promoters according to mapped transcriptional start sites (TSS). To avoid possible convoluting effects of divergently transcribed regions, we exclude these except in cases where the divergent TSSs are sufficiently far from each other (>1000bp). We used a Self-organizing Map (SOM)[27] to arrange the final, filtered set of 2118

promoters in a visually coherent manner (Fig1A). Two qualitatively distinct nucleosome occupancy patterns were evident from the SOM (Fig1B). The majority of promoters (84%) comprise the “strong-NFR” class, which are characterized by a single well-defined nucleosome-free region (NFR) emanating from the TSS that occupies the core promoter region (-150 to 0). The remaining promoters comprise the “weak-NFR” class, which have non-stereotyped nucleosome occupancy patterns characterized by diffuse nucleosome deficiency in the entire promoter region up to -400. The weak-NFR promoters encompasses a broad range of atypical nucleosome architectures, featuring delocalized nucleosomes and promoters with multiple localized NFRs of varying lengths. Divergently transcribed promoters are overrepresented in the weak class, but their inclusion did not qualitatively affect the overall nucleosome-occupancy patterns of either class (FigS1). The biological significance of this grouping has been investigated by Tirosh et al., who made a similar classification and saw differences in many features including histone turnover, binding site locations, H2A.Z occupancy, expression noise, and expression diversity[28]. We observed similar trends in our grouping: for example, the weak-NFR class was overrepresented in TATA-containing promoters (FigS2).

First poly(dA:dT) signature: localized, strand-dependent tracts symmetrically distributed about NFR centers

We examined the frequencies of poly(dA:dT) tracts of varying lengths as a function of distance from the TSS, considering poly(dA) and poly(dT) tracts separately (Fig1C). To facilitate comparison between tracts of varying lengths, we expressed each frequency as a percent enrichment relative to background intergenic tract frequencies. In

the strong-NFR promoter class, this analysis revealed a striking poly(dA:dT) localization pattern. For lengths greater than $l=2$, poly(dA) enrichment is strongly peaked near -60 and fall off sharply to background at -90 and -30. Surprisingly, poly(dT) tracts have a 5' offset from poly(dA): they are enriched between -120 and -60 with a shallower peak near -100. Within peak regions, location-specific enrichments of both poly(dA) and poly(dT) increase monotonically with repeat length. Finally, poly(dA) enrichment is usually greater than that of poly(dT) for similar-length tracts; for the longest tracts ($l \geq 6$), maximum enrichments exceed +200% for poly(dA) and +175% for poly(dT). In contrast, these characteristics are not observed in the weak-NFR class, where poly(dA:dT) enrichments are generally much smaller in magnitude than in the strong-NFR class. We observe a slight enrichment of poly(dT) downstream of -60 and a slight deficiency of poly(dA) in the same region, but these trends do not significantly increase for longer tract lengths.

It is telling to take the difference between the poly(dA) and poly(dT) enrichments (Fig1D). The resulting enrichment difference curves track the asymmetry between the inverted tracts; their x-intercepts give the coordinates at which poly(dA) tracts become more abundant than poly(dT). For the strong-NFR class, the enrichment differences of all six length-classes intercept the x-axis at -80, meeting there with a common point of inflection. This curve has an approximate C_2 symmetry: rotation of the curve about the coordinate -80 by 180 degrees will result in a similar curve. A similar C_2 symmetry about -80 is manifest in the underlying sequence: from the point of view of an observer situated at -80, tracts in both directions appear identical until the symmetry is broken by a gene on one side (Fig1E). From here on we will refer to -80 simply as the “symmetric

axis”. Note that the coordinate of each poly(dA:dT) tract was assigned according to its 5'-most base, introducing a slight 5' shift. Therefore the symmetric axis is closer to -75 when the discrete lengths of tracts are considered, placing it directly at the center of the core promoter region.

Poly(dA:dT) positions correlate with fine NFR variations

We have shown that strong NFRs in yeast are associated with a sequence pattern consisting of poly(dA:dT) tracts whose positional distributions are symmetric about NFR centers. This result suggests that poly(dA:dT) tracts may directly influence NFR placement but leaves open the alternative that both features are independently associated with the biological specialization of genes in the strong-NFR class. To resolve this issue, we analyzed how poly(dA:dT) tract localization relates to fine NFR positional variations within subgroups of the strong-NFR class. Strong-NFR promoters ordered by the self-organizing map gave a graded arrangement where 5' NFR boundaries are progressively shifted toward the TSS (Fig2A). By comparison, 3' NFR boundaries shifted little, and some of this variation may be attributable to experimental error in TSS determination. Taken together, this amounts to a gradual narrowing of the NFR width from 202bp to 98bp (inferred from peak-to-peak distances of boundary nucleosomes) over 1781 promoters.

We segmented the ordered strong-NFR promoters into 6 equal subgroups (I-VI). Group I promoters are the most similar to the weak-NFR class: they have the longest NFRs and have the most NFR positional variability, or “fuzziness”. Group VI, which contains the shortest NFRs, also had a high degree of variability; the remaining

subgroups were more homogeneous. The peak-to-peak NFR center coordinate shifted toward the TSS by approximately 15bp per subgroup (fitted with linear regression), spanning a range from -130 to -50. We then plotted the poly(dA:dT) enrichment difference as before for each subcluster (Fig2B; see FigS3 for raw enrichment values). As expected, the overall magnitude of each subgroup's tract enrichments is dependent on its degree of fuzziness. Group I enrichments peak near +100% whereas Group IV, the most localized subgroup, reached peak enrichments of +300%.

The overall qualities of poly(dA:dT) enrichment are intact for every subgroup, with 5' poly(dT) bias and a symmetric 3' poly(dA) bias. However, the symmetric axis for each group is shifted in a way that directly corresponds to its fine NFR position: wide NFRs have axes shifted away from the TSS, whereas narrow NFRs have axes shifted toward the TSS. To quantify this relationship, we plotted symmetric axis coordinates versus their corresponding NFR center coordinates for each subgroup (Fig2C). Linear regression gave an excellent fit ($r^2 = 0.90$) with a slope very close to 1: a 1bp shift in the poly(dA:dT) symmetric axis will produce a corresponding 1bp shift in the NFR center position. Their concordance with fine positional variations between NFRs reinforces the notion that poly(dA:dT) tracts distributed symmetrically across a central axis may have a direct influence on NFR placement.

Poly(dA:dT) tracts show independence from transcription factor binding sites

Poly(dA:dT) have long been implicated as regulatory elements in *S. cerevisiae*[29,30], and in particular they have been seen to influence the regulatory behavior of transcription factors with adjacent binding sites[31]. For example, the

insertion of a Reb1 binding site juxtaposed with a 3' T₇ sequence was sufficient to induce NFR formation even in the context of a coding region[21]. Furthermore, Lee et al. recently showed that binding sites for Reb1 and Abf1 are localized specifically to a narrow region centered at -100bp upstream of the TSS[10], which is slightly upstream of the symmetric axis. These observations not only suggest a synergistic relationship between poly(dA:dT) tracts and these particular transcription factors, but also raises the question of whether localized tract enrichments are seen solely as a consequence of juxtaposition with localized transcription factor binding sites (TFBS).

To address this question, we first used the set of all bound and functionally conserved transcription factor binding sites[31] to survey which factors may be spatially coupled to tract enrichment. Because tract localization is seen only in the strong-NFR class, we first screened factors using their relative TFBS abundances in strong- vs. weak-NFR promoters (Fig3A). Interestingly, the TFBSs of the majority of factors are vastly overrepresented in the weak-NFR class; for example, Skn7 favors weak-NFR promoters by more than 10:1 (normalized by class size). Only 4 of 45 factors examined (Reb1, Hsf1, Abf1 and Rpn4) were overrepresented in the strong-NFR class, and of these only Reb1 and Abf1 have bound and functionally conserved sites in a significant number of promoters genome-wide (226 and 209, respectively).

We reasoned that if localized poly(dA:dT) tracts are manifest mostly in the context of abundant and localized sites such as those of Reb1/Abf1, then by restricting our analysis to promoters containing high-confidence Reb1 or Abf1 sites poly(dA:dT) enrichment signals should be greatly enhanced over those of background promoters. However, this is not the case: tract enrichments are similar in magnitude between

Reb1/Abf1-site containing promoters and background (Fig3B). To see whether poly(dA:dT) localization could arise from general juxtaposition with other transcription factors, we also assessed tract enrichment in a set of TFBS-depleted promoters: strong-NFR promoters from which promoters containing annotated TFBS have been excluded. To maximize the stringency of this exclusion step, the TFBS used for filtering include non-conserved sites as well as sites with less stringent binding thresholds ($p < 0.005$) for 118 transcription factors. The final set of TFBS-depleted promoters ($n=674$), however, show no decrease in localized poly(dA:dT) tract enrichments compared to the unfiltered set (Fig3B). In fact, TFBS-depleted promoters actually show slightly increased poly(dA:dT) enrichment. Thus, while it is possible that poly(dA:dT) enrichments can be explained by juxtaposition to other abundant, localized and NFR-specific transcription factors not covered by the MacIssac et al. study, the most likely interpretation is that poly(dA:dT) positioning patterns arise from TF-independent tract function.

Curiously, both Abf1- and Reb1-specific tract enrichment profiles have slight variations from the overall strong-NFR profile. Both Abf1 and Reb1 have sharper poly(dT) enrichments than background, and the Reb1 poly(dT) peak is shifted by ~30bp to the opposite side of the symmetric axis. Reb1 also lacks a prominent poly(dA) peak. These factor-specific deviations from average poly(dA:dT) localizations likely reflect individualized factor-tract relationships.

Independent assortment of poly(dA) and poly(dT) tracts suggests that individual promoters do not require dual symmetric tracts

By analyzing the averaged localization profiles of poly(dA:dT) tracts over large numbers of promoters, we have seen a characteristic symmetry of opposing tracts with respect to the NFR center. There are two ways in which this collective symmetry could arise: either promoters are individually symmetric or the symmetry is a consequence of overlaying individually asymmetric promoters. We address this question by considering whether the number of promoters containing both poly(dA) and poly(dT) tracts in enriched regions exceed the expected number given independent assortment of poly(dA) and poly(dT). For this purpose we analyze the 297 well-aligned promoter sequences of strong-NFR subgroup IV: because these promoters are highly uniform, we can estimate the number of functional poly(dA:dT) tracts in each promoter by counting them in fixed windows centered at tract-enriched positions. This counting analysis is presented in Table 1, which shows the fraction of promoters containing at least a single copy of poly(dA) or poly(dT) in their respectively enriched regions in comparison to the fraction of promoters containing both. For all length cutoffs, the number of promoters with co-occurring poly(dA) and poly(dT) is comparable to or lower than the overlap expected from independent assortment. This data is consistent with a model where poly(dA) tracts proximal to the TSS are positioned independently with respect to poly(dT) tracts distal to the TSS. Therefore, this counting analysis disfavors models where poly(dA:dT) tracts are mechanistically constrained to act as inverted pairs in individual promoters.

Second poly(dA:dT) signature: NFR-specific and terminal-specific G:C capping of tracts

Our analysis of poly(dA:dT) tracts has suggested that their placement in promoters is not constrained by functional coupling to either transcription factors or to symmetric counterpart tracts. Nonetheless, it is difficult to imagine that poly(dA:dT) tracts can function entirely autonomously given their low information content. Therefore, we investigated the possibility that there is additional information content in sequences flanking poly(dA:dT) by looking at the terminal base pair composition at both ends of poly(dA:dT) tracts of various lengths. First, we examined the background set of all yeast intergenic sequences (Fig4A), where poly(dA) and poly(dT) were pooled due to the lack of reference directions. Surprisingly, even in this background set we saw that, relative to poly(dA), tracts have a significant preference for the incorporation of G nucleotides at both terminal positions (accordingly, poly(dT) have terminal bias for C). We refer to this phenomenon as “G:C capping” of poly(dA:dT) and we define the G:C capping rate as the proportion of poly(dA:dT) tracts which terminate in this manner. In the yeast intergenic background, the G:C capping rate is a steadily increasing function of tract length; for very long tracts the capping rate at both termini is greater than 40% (Fig4A), a significant increase from the expected rate of 25.7% (computed by renormalizing single base frequencies; see Methods).

An interesting feature of “G:C capping” is that there is an asymmetry in capping rates between the two tract termini. By convention, we designate the G:C capping terminus relative to the poly(dA) strand: tracts manifest as GA_n and T_nC will be referred to as “5’ G:C capped” whereas tracts of the form A_nG and CT_n will be referred to as “3’ G:C capped” (Fig4A). In the yeast intergenic background, it is apparent that 5’ capping

rates are consistently greater than 3' capping rates, a difference which tends to increase for longer tract lengths.

To see whether oriented G:C capping could play a role in poly(dA:dT) specification of NFRs, we considered the location dependency of G:C capping in the context of promoter regions (strong and weak NFR classes considered together) (Fig4B). In this context, we now consider poly(dA) and poly(dT) separately due to the symmetry breaking TSS. Intriguingly, 5' G:C capping has many features similar to poly(dA:dT) enrichment. For poly(dA), 5' capping rates are far above background at the center of the core promoter region (-150 to 0). Near its peak, 5' poly(dA) capping is an increasing function of tract length, with capping rates of >64% for tract lengths 6 and above. 5' capping rates fall back to background levels at the edges of the core promoter region as well as in distal promoter regions (-300 to -150). An identical effect is seen for the 5' G:C capping of poly(dT) tracts. However, in contrast to the strongly context dependent 5' G:C capping rates, 3' G:C capping rates are uniform in both core and distal promoter regions, where their modest dependencies on tract length are consistent with background capping rates.

Finally, we de-convoluted 5' G:C capping into contributions from the strong-NFR vs. weak-NFR promoters (Fig 4C, $l=5$ shown). Whereas the 5' G:C capping rate for strong NFR promoters is strongly peaked in core promoter regions, the 5' G:C capping rate is much more delocalized in weak-NFR promoters. In weak-NFR promoters, poly(dT) tracts show a relatively uniform capping rate in both core and distal promoters regions while poly(dA) capping rates actually decline in core promoter regions. In both cases, local capping rates in weak-NFR promoters differ substantially from the typical

pattern seen in strong-NFR promoters. These observations are consistent with the notion that NFRs are arranged in delocalized and atypical ways in the weak class.

Thus, our data suggests that there are two separate effects which contribute to the observed G:C capping of poly(dA:dT) tracts. First, there is a background G:C capping effect that is prevalent in the bulk of intergenic yeast sequences. This background G:C capping is non-specific: it is present in both core and distal promoter regions, increases at a moderate rate as tract lengths increase and is largely similar between the 5' and 3' termini. Superimposed on top of this background effect is a second G:C capping effect which is both 5' terminal-specific (relative to poly(dA)) and NFRs-specific and which most strongly affects poly(dA:dT) tracts near the symmetric axis. It is this second G:C capping effect which constitutes an additional poly(dA:dT)-based sequence signature for NFRs.

Tract capping is skewed toward the symmetric axis

Although trends in 5' G:C capping have many similarities with poly(dA:dT) enrichment, there is one important distinction: 5' G:C capping for poly(dA) and poly(dT) both occur directly at the symmetric axis, whereas enrichment for poly(dA) is distinct from poly(dT) and occur at regions flanking the symmetric axis. To highlight this distinction, we chose a particular length class (≥ 5), renormalized their capping and enrichment curves relative to their range of values, and co-plotted them (Fig5A). Relative to enrichment peaks, there is a ~20bp shift in 5' G:C capping toward the NFR central axis for both poly(dA) and poly(dT). Thus, when capping biases are

superimposed on top of tract enrichments, the result is that each tract population contains a 5' capped subpopulation that is skewed toward the symmetric.

We offer two interrelated interpretations of this skewed capping effect. First, capped tracts are directional signals, both in the sense that capping is specific to the 5' terminus and in the sense that capping occurs on tracts skewed toward the NFR central axis. Thus, capped tracts may highlight a sense of local orientation in promoter regions; either they “point” toward the central axis of the NFR or away from the NFR's boundaries. Second, the fact that capping is localized to the symmetric axis suggests that capping reinforces some aspect of poly(dA:dT) tract function that is especially significant near NFR centers. Most promoters contain multiple tracts in poly(dA:dT) enriched regions (Table 1), and capping may give additional specificity to particular tracts as a way of “highlighting” them in the context of a group of tracts. Based on the size of the shift between enrichment and capping curves (~20bp), we estimate the typical length of a poly(dA:dT) enriched region to be ~40bp in a single promoter. In summary, both the position and directionality of poly(dA:dT) tract capping may important for facilitating the manner in which tracts relate to NFR positions (Fig5B).

Discriminating between “Central” and “Boundary” NFR definition models

Assuming that poly(dA:dT) tracts act as directional signals with roles in NFR specification, we consider two general NFR definition models which describe how poly(dA:dT) tracts can potentially influence NFR formation (Fig6). In the “Central” definition model, poly(dA:dT) tracts mark a specific location within the promoter as the center of a nascent NFR. Once the central site is defined, a set of downstream events

allows nucleosomes to be spaced equidistantly in opposite directions to create the nucleosome-free region. This model is supported by the strong 1-to-1 concordance between tract enrichments and positions of NFR centers. Furthermore, NFR-specific 5' capping localizes specifically to the central axis and could play a role in the definition of the hypothetical center. An additional advantage of this model is that, in principle, only a single tract or set of unidirectional tracts is required to specify each NFR. However, one drawback of the Central model is that it does not explain how the extent of each NFR is specified.

An alternative to the Central model is the "Boundary" model, which posits that poly(dA:dT) tracts act as directional boundary elements which anchor the NFR's flanking nucleosomes. In this model, nucleosomes are directed away from the capped ends of poly(dA:dT) tracts. This model requires poly(dA:dT) tracts to be present in correct orientations at both nucleosome boundaries in each NFR. This requirement is an obvious drawback, as we demonstrated earlier that such a scenario is unlikely.

In order to differentiate between these two models, we now consider the detailed relationships between poly(dA:dT) locations and NFR positioning. We use the centroids of poly(dA:dT) enrichment as single numerical indicators of tract location. Intuitively, the centroid corresponds to the expected position of the poly(dA:dT) tract after taking a weighted average across the enrichment peak. We schematically represent the positions of five promoter elements: 5' nucleosome boundary, 3' nucleosome boundary, NFR center, poly(dT) centroid and poly(dA) centroid across the six strong-NFR subgroups (Fig7). We then assign a best-fit slope to each of the five promoter elements using linear regression. The slope of each element, with units of bp/subgroup, represents the average

number of base pairs the element shifts toward the TSS per each group of 297 promoters from the SOM.

If the Boundary model is correct, we would expect poly(dA:dT) tract positions to vary in lockstep with the positions of boundary nucleosomes. This is not the case: the slope of 5' nucleosomes is 24 bp/group whereas the slope of the adjacent poly(dT) tracts is only 10bp/group. The lack of coordination between 5' nucleosomes and poly(dT) is clear by contrasting groups I and VI: in the former case the two features are separated by over 80bp, whereas in the latter case the separation is less than 10bp. A lack of coordination in the opposite sense is manifest between 3' nucleosomes and poly(dA): 3' nucleosomes barely shift (slope = 5 bp/group) whereas poly(dA) tracts have much greater shift (slope =13 bp/group). Globally, it is apparent that distances between poly(dA) and poly(dT) do not narrow as NFRs do. This lack of coordination between shifts in poly(dA:dT) positions and their respective nucleosome boundaries argues against the Boundary model. On the other hand, the slope of the NFR central coordinate (15 bp/group) is similar to slopes of poly(dA) and poly(dT) centroids. Poly(dA) and poly(dT) tracts tend to remain at a relatively fixed distance from each other (~40-50bp) and maintain the NFR center between them regardless of NFR position or width. The alignment preference of poly(dA:dT) tracts toward the central axis, rather than toward boundaries, favors the Central model of relating tracts with NFRs.

Discussion

We have reported the computational characterization of two previously unknown features of poly(dA:dT) tracts in yeast promoters: 1) a strand-specific localization that is

approximately C_2 symmetric about the center of the core promoter region and 2) a terminal-specific capping by G:C residues. The association between these two poly(dA:dT) tract characteristics and nucleosome-free regions in promoters is supported by both qualitative and quantitative lines of evidence. First, both strand-dependent tract enrichment and 5' G:C tract capping were found to have characteristic, localized distributions that are specific to the strong-NFR class of promoters. Second, shifts in NFR positions across different sets of promoters were mirrored by corresponding shifts in tract positions. Collectively, they argue that a direct mechanistic relationship between poly(dA:dT) tracts at defined promoter positions and the specification of NFR placement at these promoters is highly plausible.

Our work also suggests several mechanistic guidelines for how poly(dA:dT) tracts may translate into NFRs. First, we believe that the directionality of tracts is important for providing a sense of local orientation in promoters, perhaps for binding of a factor that recognizes the direction of DNA. This directionality is manifest in multiple aspects: in the C_2 symmetry of tracts about NFR centers, in the terminal-specific tract capping preferences and in the way that capping is spatially skewed toward the symmetric axis. Second, individual promoters are unlikely to require complementary tracts on both sides of the symmetric axis (although in many cases these may be present) even though this is seen after tract profiles are averaged over many promoters. Third, our work suggests that poly(dA:dT) tracts have better spatial correlation with NFR centers than with boundary nucleosomes. Thus, we believe that positioned poly(dA:dT) tracts are not likely to act as boundary elements but instead play a role in defining the NFR center coordinate. Finally, tract specification of NFRs does not generally seem depend on transcription factors.

However, tracts may have specialized relationships with particular transcription factors which occur prominently in strong-NFR promoters such as Reb1 and Abf1.

Shortcomings of our current model are that it does not explain 1) how the extent of each NFR is defined and 2) how the -1 and +1 nucleosomes are spaced equidistantly from the central region.

NFRs appear in most promoter regions and have well-defined localizations; thus the sequences that collectively define the spatial patterning of NFRs must be highly abundant as well as highly specific. An ongoing challenge is to understand how low information content sequences such as poly(dA:dT) tracts can contribute to both criteria. There is a tradeoff of coverage for specificity as tract lengths increase: longer tracts are more specific but will occur in fewer promoters. How long does a poly(dA:dT) tract have to be in order to exert a functionally significant effect? Assuming that poly(dA:dT) tracts underlie a general NFR-specification mechanism, we can estimate an upper bound on the minimal functional tract length. In order for strong-NFR promoters to be covered by at least one functionally relevant poly(dA:dT) tract, tracts as short as length 4 must be invoked (Table 1; 96% of promoters are covered in this case). Indeed, our analysis has shown that even poly(dA:dT) tracts shorter than length 4 have significant enrichment peaks (Fig1C). Thus it is very likely that short tracts on the order of length 4 make non-negligible contributions to NFR specification in many promoters.

Short tracts alone, however, seem to lack the necessary specificity. Therefore, we reasoned that short poly(dA:dT) tracts must work within context: there must be additional sequence signals specific to NFRs which allow functionally significant poly(dA:dT) tracts to be distinguished from decoy tracts. Because the typical promoter

contains multiple co-localized tracts (Table 1), one possibility is that tracts benefit from cooperativity. G:C tract capping, which increases tract information content, is likely to be another important source of specificity. Because tract capping is centered at the poly(dA:dT) symmetric axis, capped tracts seem to be a more location-specific indicator of the central NFR coordinate than tracts in general. Thus, both capped and uncapped tracts appear to be important: uncapped tracts provide a context in which capped tracts are emphasized. Together, they may act as a directed signal that helps to “point out” where NFRs should be centered.

Our work represents a departure from the view that poly(dA:dT) tracts are haphazardly positioned promoter elements which generically displace nucleosomes by virtue of their physical rigidity. Here it is important to emphasize that much of the work which suggest that poly(dA:dT) tend to exclude nucleosomes by rigidity pertain only to very long ($l > 20$) tracts[33,34,35,36]. Effects are modest for shorter tracts: 10bp poly(dA:dT) tracts only destabilize nucleosomes by ~ 0.2 - 0.3 kcal/mol[32], while incorporating an A_{16} tract into the middle of nucleosomal DNA only resulted in a 1.7 fold (0.35 kcal/mol) destabilization[26]. In this shorter length range ($10 \leq l \leq 20$) poly(dA:dT) tracts can in many cases be incorporated into positioned nucleosomes *in vitro*[33,34,35] and *in vivo*[36]. Even if tracts at these lengths can perturb nucleosome positions, they are likely to affect only a minority of promoters genome-wide ($\sim 14\%$ of strong-NFR promoters by Table 1). By contrast, the lengths of poly(dA:dT) tracts for which we have shown to be functionally correlated to NFRs are much shorter. While tracts as short as length 4 have been seen to adopt straight structures *in vivo*[25], it is unlikely that their rigidity alone can exert thermodynamically significant nucleosome exclusion effects.

Finally, *in vitro* selection experiments have found that even the the most *disfavorable* sequences destabilize nucleosomes by only modest amounts (< 0.7 kcal/mol relative to bulk genomic DNA)[5]. Even if core promoter sequences destabilize nucleosomes by comparable amounts, these free energy differences are on the order of thermal fluctuations and would be insufficient to keep NFR boundary nucleosomes in their fixed positions.

We conjecture that poly(dA:dT) tracts, rather than acting as static nucleosome-repelling elements, play an integral role in a series of well-orchestrated chromatin-remodeling events that transform a random distribution of promoter nucleosomes into the characteristic open NFR architecture (Fig8). One hypothesis is that poly(dA:dT) tracts can facilitate interaction with chromatin remodeling enzymes which remove nucleosomes and/or slide them away from the NFR center; the presence of multiple tracts may facilitate remodeling processivity. Yeast contains numerous candidate chromatin remodeling complexes including Swi2/Snf2, Ino80, Isw1, Isw2, and RSC[37]. For example, RSC is known to mediate nucleosome sliding at Pol II promoters [38,39] and its ATPase subunit, Sth1, can track along one strand of duplex DNA with 3' to 5' polarity[40]. It will also be interesting to see whether H2A.Z, which is deposited at both NFR boundary nucleosomes, plays a role in NFR-specification that is coordinated with poly(dA:dT) tracts.

Despite unresolved mechanistic details, we have provided insight into a class of promoter-specific sequences which correspond to the positions of key nucleosomes. By postulating that poly(dA:dT) motifs form the basis of a dynamic mechanism of NFR formation, we bring a fresh perspective on this ubiquitous class of sequences. Finally,

the essential features of NFRs are seen in the promoters of multicellular eukaryotes with small variations. It is not yet known whether characteristic poly(dA:dT) sequences patterns are seen in NFR-containing promoters of higher organisms, but even if sequence patterns have diverged the mechanistic principles of NFR formation suggested in yeast may still be preserved.

Methods

Promoter classification.

S. cerevisiae nucleosome positioning data mapped at 4bp resolution (Lee et al.) was aligned relative to 4799 mapped transcriptional start sites (David et al.) from -400 to +400. For all single direction promoters we considered the distance between the TSS and the corresponding start codon, filtering out promoters where the TSS is downstream or >500bp upstream of ATG. Furthermore, we excluded short promoters of less than 150bp, yielding a total of 1842 filtered single direction promoters. This analysis was repeated in divergently transcribed intergenic regions where the TSS to TSS distance was greater than 1kb. In these cases the opposing termini were extracted separately to give 276 additional promoters. All promoters were oriented relative to the direction of transcription. Self-organizing Map analysis was performed using Cluster with default parameters (Xdim=49 and Ydim=1) and visualized with Treeview; both are available from the Eisen Lab (<http://rana.lbl.gov/EisenSoftware.htm>).

Poly(dA:dT) enrichment analysis.

We computed background frequencies of all poly(dA:dT) sequences within all yeast intergenic sequences (3.5Mb) as motif count per total sequence length. Motif counts were tabulated only according to the longest observed tracts; e.g. GAAAAT registers as AAAA but not AAA or AA. For the background set, poly(dA) and poly(dT) were pooled due to lack of reference direction. Position-dependent frequencies in promoter sequences were computed in 21bp sliding windows centered at the reference position. Any poly(dA:dT) tract whose 5' end was located in this window was tabulated to occur in this window. For each window, frequencies were computed as total motif number over all sequences per window size. For both background and location-specific analyses tracts of $l \geq 6$ were counted as a single class. Enrichment scores were computed as:

$$Enrichment = \frac{freq_{obs} - freq_{background}}{freq_{background}}$$

Transcription factor binding site analysis

We used TFBS annotation from MacIssac et al. with ChIP-chip $p < 0.001$ and conservation in 2 additional yeasts species (http://fraenkel.mit.edu/improved_map/). For each factor we counted the number of promoters containing bound and conserved binding sites (“orfs_by_factor_p0.001_cons2.txt”) for factor in strong- and weak-NFR classes. Only factors with sites in more than 30 promoters genome-wide (45 factors total meet this criterion) were considered. For factors overrepresented in the weak-NFR cluster the p-value is $\log_{10}(P(X \leq n))$ whereas for factors overrepresented in the strong-NFR cluster the p-value is $-\log_{10}(P(X \geq n))$, where n is the number of promoters occurring

in the strong-NFR class. For the TFBS exclusion analysis we excluded from the 1781 strong-NFR promoters all promoters in the file “orfs_by_factor_p0.005_cons0.txt”.

Tract counting analysis

Non-overlapping tracts were counted in fixed windows of the 297 aligned promoters of strong-NFR subgroup IV. Windows were 80bp wide and centered at -60 for poly(dA) and at -100 for poly(dT). The expected number of promoters with both poly(dA) and poly(dT) coverage given independent assortment of poly(dA) and poly(dT) is given by:

$$E = \frac{n_A n_T}{n_{total}}$$

where n_A is the number of promoters containing poly(dA) tracts, n_T is the number of promoters containing poly(dT) tracts, and n_{total} is the total number of promoters.

G:C capping analysis

Background capping rates were found by pooling all instances of poly(dA) and poly(dT) in all yeast intergenic sequences and computing 5' and 3' terminal base compositions separately for each length class. To compute the expected background capping rate r , we renormalize background single base frequencies f_σ given that the capping base X is different than the tract base Y :

$$r(X, Y) = \frac{f_X}{\sum_{\sigma \neq Y} f_\sigma}$$

thus:

$$r(G, A) = \frac{f_G}{f_C + f_G + f_T} = \frac{.1738}{.1779 + .1738 + .3245} = .257 \approx r(C, T)$$

In promoter sequences, we define capping rates at upstream and downstream termini c_L^{up} and c_L^{down} by base X for Y -tracts of length L at coordinate i as:

$$c_L^{up}(X, Y, i) = p(\sigma_{i-1} = X \mid \sigma_{i-1} \neq Y, \sigma_{i+L} \neq Y, \sigma_k = Y, i \leq k < i + L)$$

and

$$c_L^{down}(X, Y, i) = p(\sigma_{i+L} = X \mid \sigma_{i-1} \neq Y, \sigma_{i+L} \neq Y, \sigma_k = Y, i \leq k < i + L)$$

that is: the probability that the upstream or downstream capping base is X given contiguous bases Y at positions i through $i+L-1$ flanked by non Y bases at $i-1$ and $i+L$. Capping rates shown were smoothed over 41bp windows centered at i .

Nucleosome boundary calculations

For each subgroup, we computed the average nucleosome occupancy profile using aligned data (Lee et al.). The maxima of peaks corresponding to +1 and -1 nucleosomes were taken to be their average central coordinates within each subgroup. Boundary coordinates were inferred by adjusting central coordinates by 73bp.

Poly(dA:dT) centroid calculations

For each subgroup, we first computed an extended motif enrichment profile from -300 to +50, which was then smoothed using a Gaussian ($\sigma=5$ bp). We then selected a promoter region that was inclusive of the major enrichment peaks for length classes 4, 5, and 6+. This region is defined according to the largest region spanned by x-intercepts of

the smoothed enrichment curves of these three length classes. The centroids for each length class was calculated within this region as follows:

$$Centroid_k = \frac{\sum_i iE_{ik}}{\sum_i E_{ik}}$$

Where i is the promoter position relative to TSS, $k=4, 5$ or $6+$ is an index over motif length classes, and E is the enrichment. To filter out background signals only locations with enrichment values above 0% were tabulated. The final centroid position for the subgroup was taken as the median value among the three length classes.

Tables

Table 1: Poly(dA:dT) coverage (fraction of promoters containing tract) and copy number (average number of tracts per promoter) in strong-NFR subgroup IV core promoters (n=297)

Tract Length Cutoff	Poly(dA) coverage	Poly(dT) coverage	dA or dT coverage	dA + dT coverage (expected)	dA avg. copies	dT avg. copies
3	96%	96%	100%	92%(92%)	3.02	3
4	80%	80%	96%	65%(64%)	1.71	1.44
5	59%	57%	83%	32%(33%)	0.89	0.83
6	34%	31%	58%	7%(11%)	0.42	0.37
7	23%	21%	41%	3%(5%)	0.27	0.23
8	15%	13%	27%	1%(2%)	0.16	0.14
9	10%	8%	18%	0%(1%)	0.11	0.084
10	9%	6%	14%	0%(1%)	0.088	0.064

Figures

Figure 1. Poly(dA:dT) tracts in strong- and weak-NFR promoters. **A)** Nucleosome occupancy data of 2118 *S. cerevisiae* promoters were clustered using a Self-Organized Map and partitioned into strong- and weak-NFR classes on a visual basis. **B)** Averaged nucleosome occupancy for strong- and weak-NFR promoters. Strong-NFR promoters are characterized by a single defined nucleosome-free region adjoining the TSS whereas weak-NFR promoters exhibit a variety of diffuse nucleosome patterns across the entire promoter. **C)** Location-specific enrichment (vs. intergenic background) of poly(dA) and poly(dT) tracts of varying lengths, taken over 21bp windows. Colors represent differing tract lengths. Tracts of lengths 6 and greater were considered collectively for statistical accuracy. Coordinates are relative to transcription start sites (TSS). Note the lack of significant tract enrichments in weak-NFR promoters. **D)** Difference between poly(dA) and poly(dT) enrichment; values above the x-axis indicate greater poly(dA) enrichment. Dashed line: symmetric axis at -80. **E)** Illustrating the C_2 symmetry of poly(dA:dT) tracts with respect to the symmetric axis.

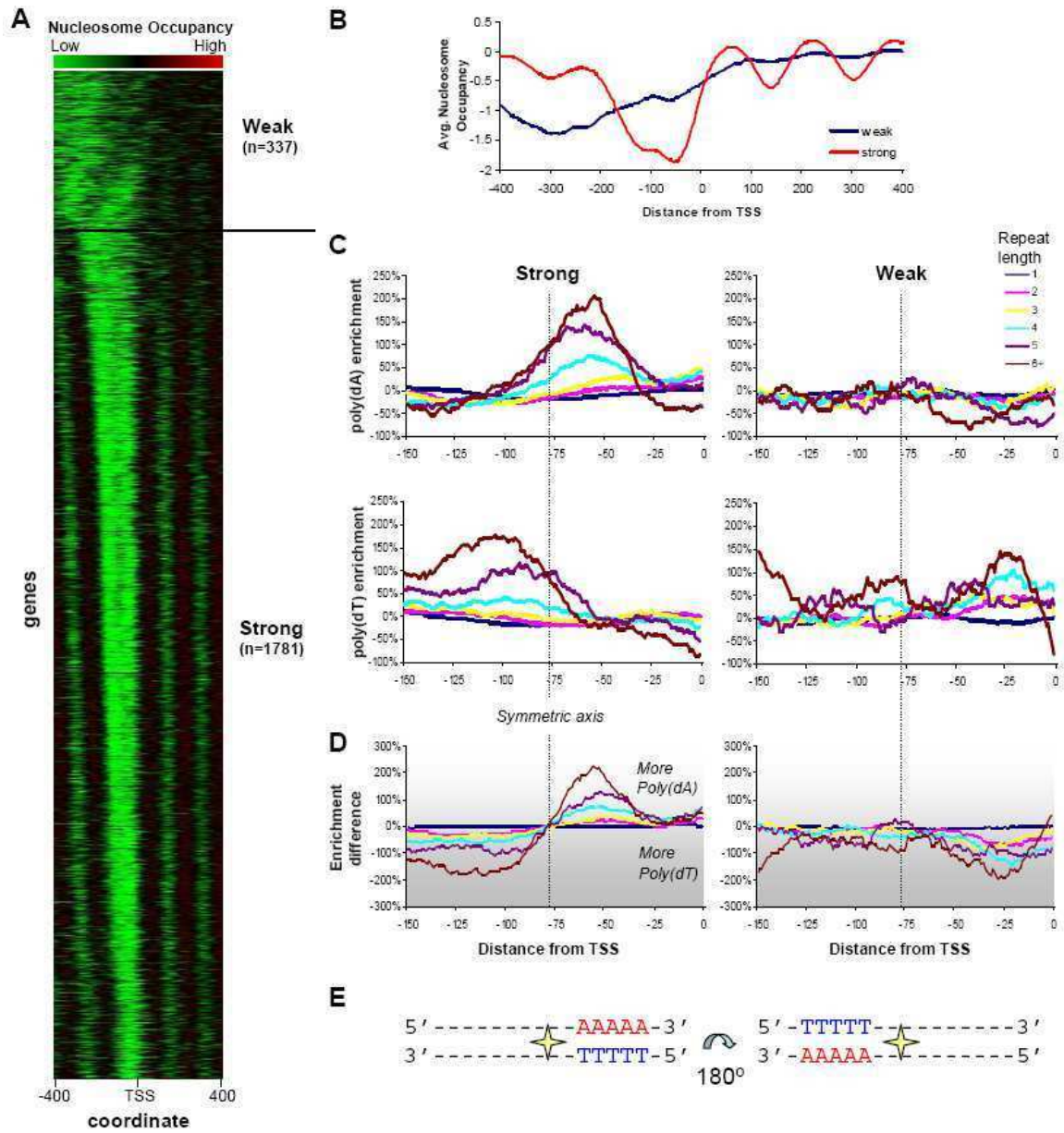


Figure 2. Poly(dA:dT) track fine variations in NFR positions. **A)** The 1781 strong-NFR promoters, which show progressive decrease in NFR length, are divided into 6 equal subgroups I-VI. **B)** Poly(dA:dT) enrichment differences in each subgroup. Arrows denote locations of symmetric axes for individual subgroups. **C)** Linear regression plot of subgroup NFR center positions vs. symmetric axis coordinates, showing 1-to-1 tracking. The multiple points per subgroup are enrichment difference x-intercepts for poly(dA:dT) lengths from 2 to 6+.

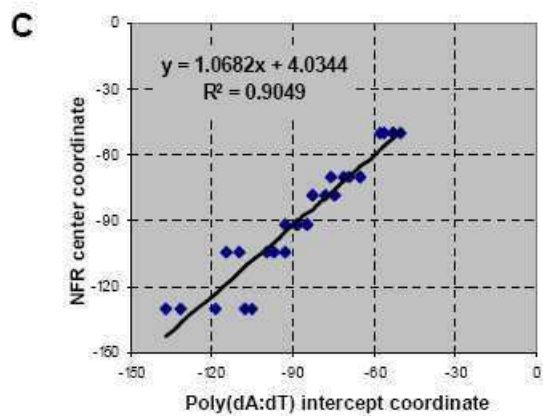
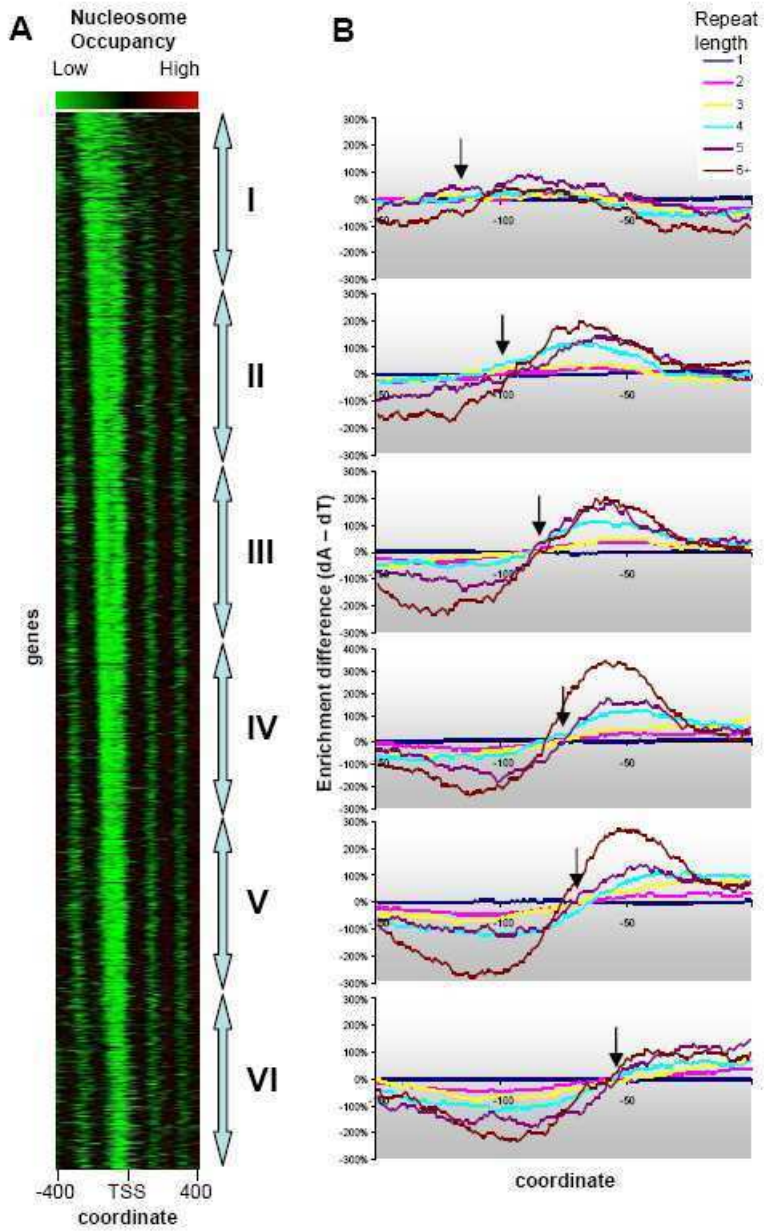


Figure 3. Poly(dA:dT) enrichments occur independently of sampled transcription factor binding sites. **A)** Ranking of transcription factors by overrepresentation of bound and functionally conserved sites (MacIssac et al.) in strong-NFR promoters. Most transcription factors have large numbers of sites in weak-NFR promoters and few in strong-NFR promoters. Prominent exceptions include general transcription factors Reb1 and Abf1. **B)** Poly(dA:dT) tract enrichments (length \geq 4) for subsets of strong-NFR promoters: Abf1-containing, Reb1-containing, or TFBS-depleted. TFBS-depleted promoters were selected by excluding promoters containing moderately bound ($p < 0.005$) binding sites (no conservation requirement) for any of 118 TFs.

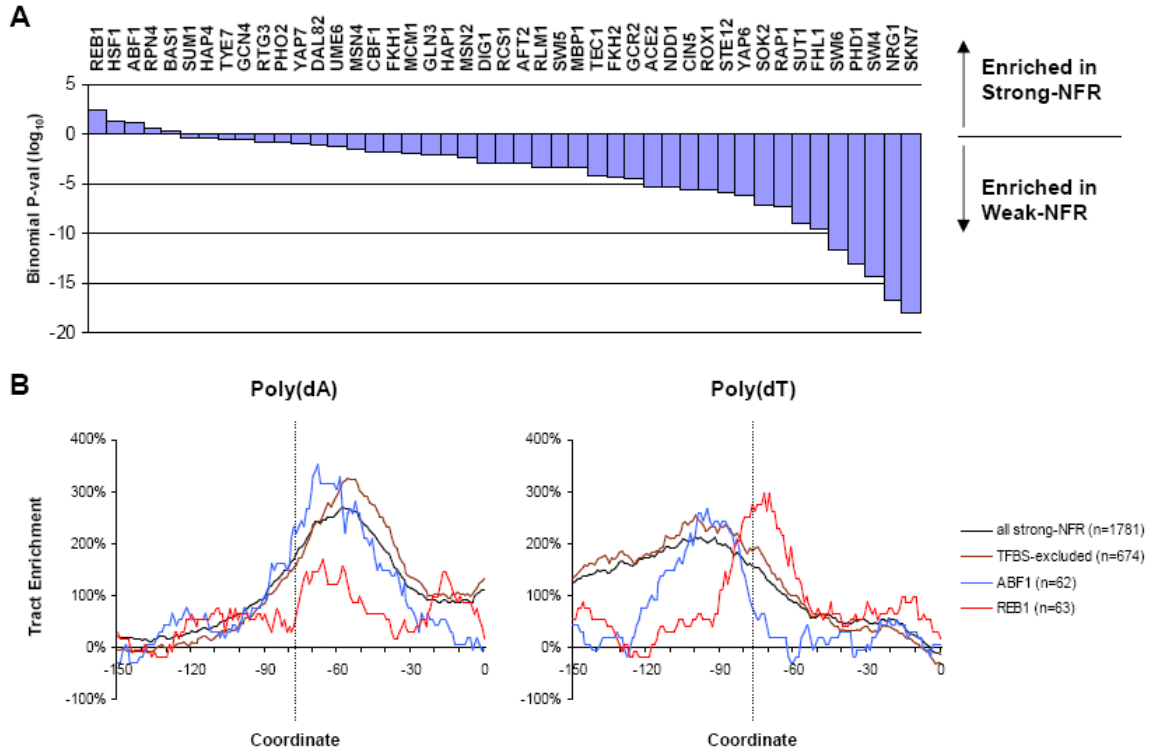


Figure 4. NFR-specific 5' G:C capping of poly(dA:dT) tracts. **A)** Background (all yeast intergenic regions) G:C capping rates of poly(dA:dT) tracts for 5' and 3' termini over different tract lengths. Examples of both 5' and 3' G:C capping are illustrated: with respect to poly(dA), tracts tend to incorporate G residues at terminal positions at higher than expected frequencies. **B)** Promoter-specific G:C capping rates (strong- and weak-NFR classes combined). 5' capping for poly(dA) and poly(dT) show prominent increases in core promoter regions (-150 to 0, shaded) but not in distal promoter regions; 3' capping remains at background levels in core and distal promoter regions. Dashed line: symmetric axis. Thus motifs of the form GA_n and T_nC occur prominently in NFR central regions. **C)** 5' capping rates for strong- vs. weak-NFR promoters.

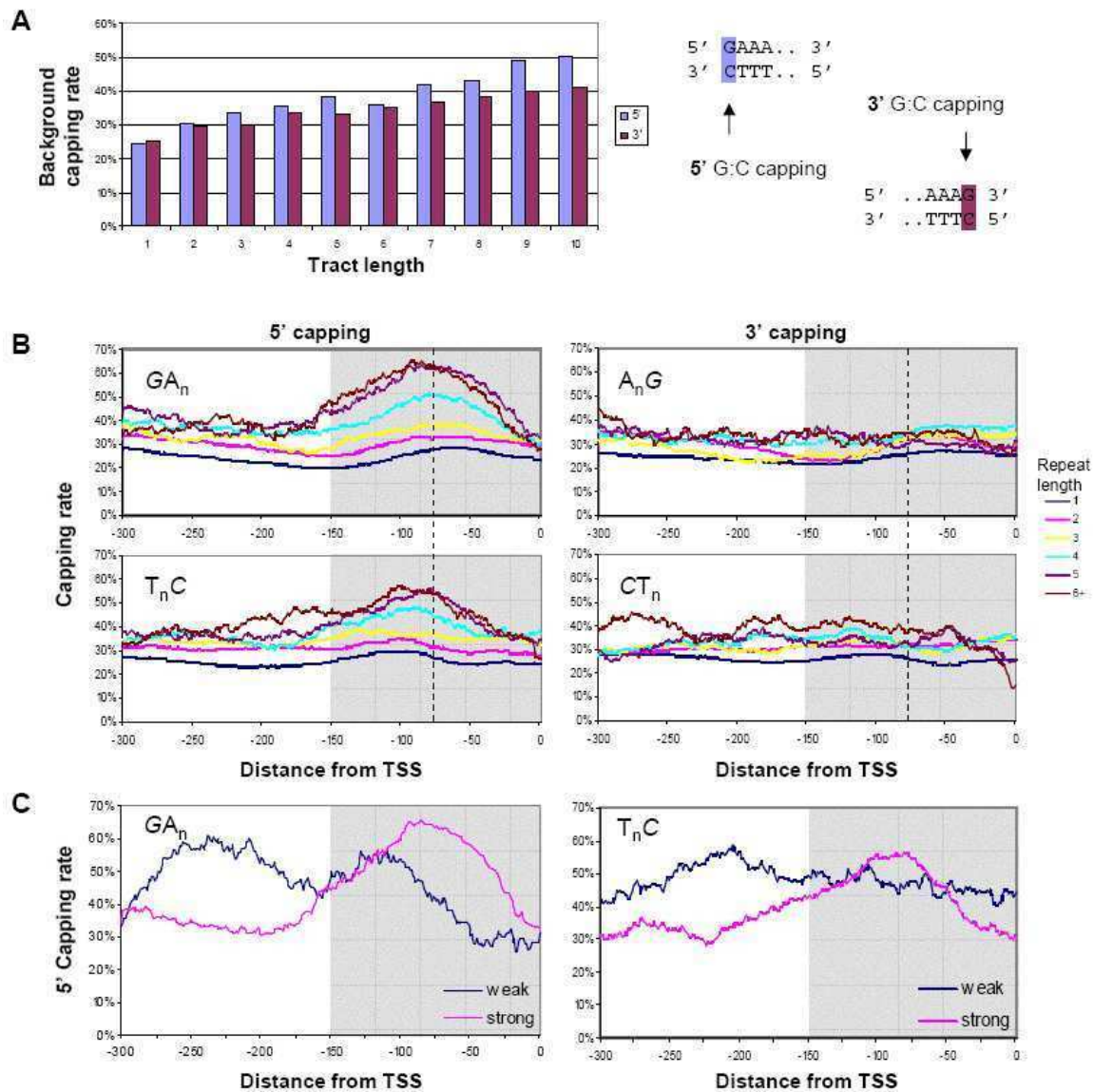


Figure 5. Poly(dA:dT) capping is offset from tract enrichments toward the NFR central axis. **A)** Poly(dA:dT) enrichments (blue) and capping rates (red) are each renormalized to their respective range of values and co-plotted. A ~20bp shift capping shift toward the central NFR axis is manifest for both poly(dA) and poly(dT). **B)** Illustrating a hypothetical set of poly(dA:dT) tracts in promoters, where capped poly(dA:dT) tracts exist as a subpopulation of “leader” sequences that directionally orient spans of tracts. Blue regions denote poly(dA:dT) tract enriched regions and red arrows denote capped tracts that point toward the 5' capped terminus.

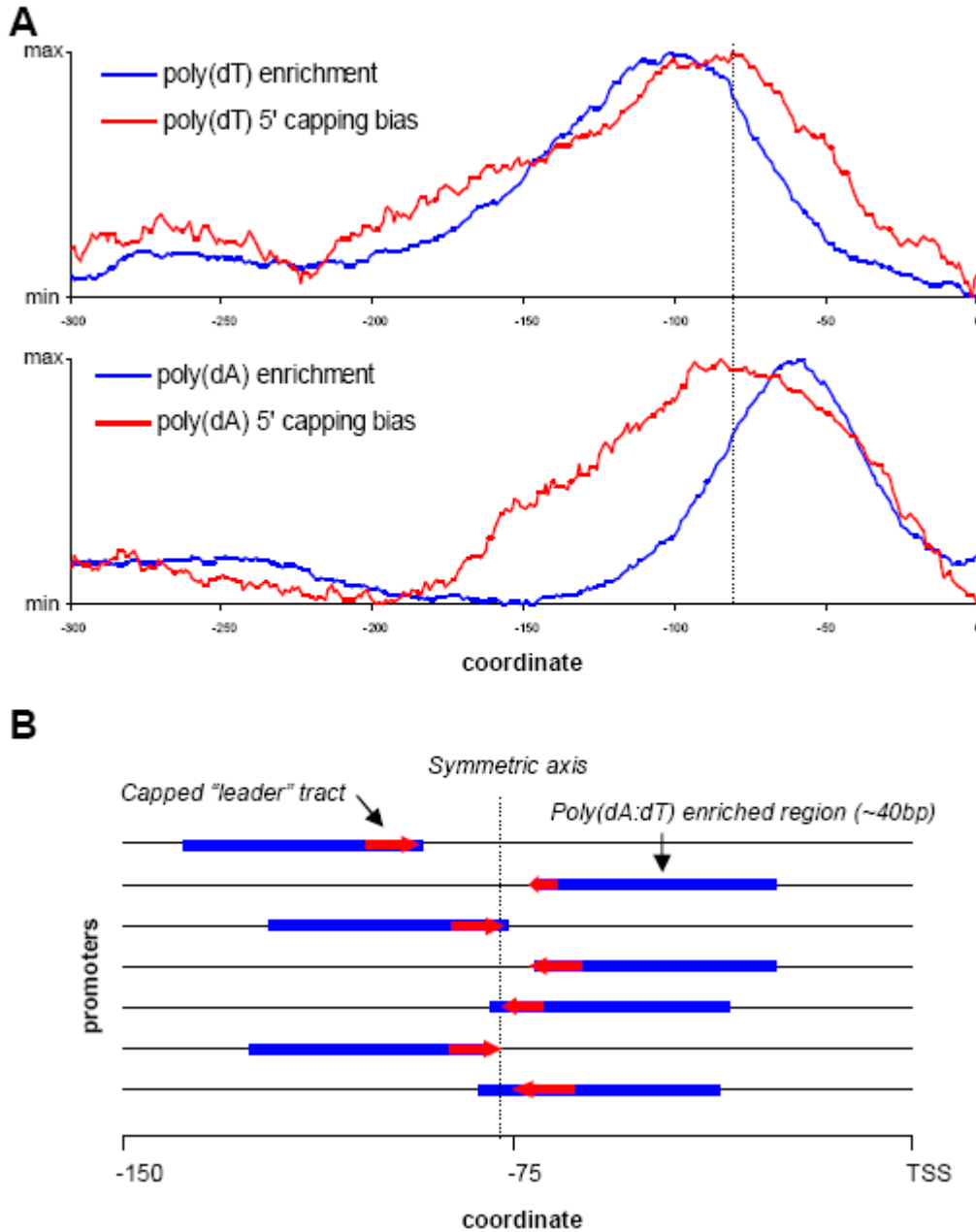


Figure 6. Two contrasting mechanistic models of NFR definition by poly(dA:dT) signals. In the “Central” NFR definition model, tracts define a single location within the promoter as the NFR center, and a separate mechanism spaces nucleosomes equidistantly from the center. In the “Boundary” NFR definition model, separate tracts at each end of the NFR act as directional boundary elements that prevent nucleosome incursion into the NFR.

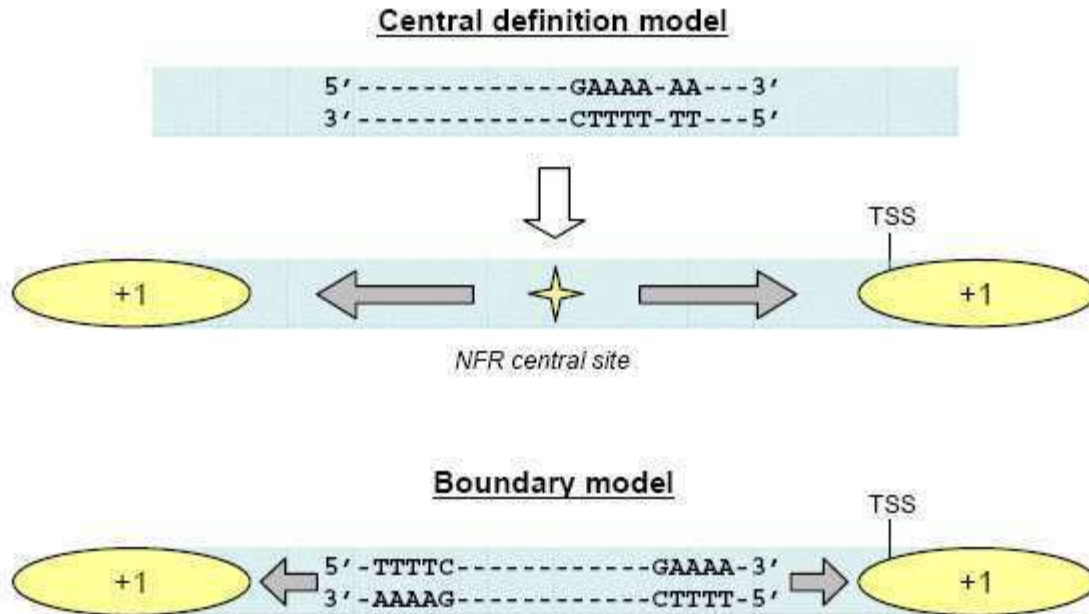


Figure 7. Locations of promoter elements in strong-NFR subgroups favors the Central Definition model. Five promoter elements (5' and 3' nucleosome boundaries, NFR center, and poly(dA)/(dT) centroid locations) are represented for each strong-NFR subgroup. The slopes of each element, which represent shifts of each element per subgroup, were derived using linear regression. Poly(dA:dT) centroids track not with NFR boundaries but with NFR centers, thus supporting the Central NFR definition model.

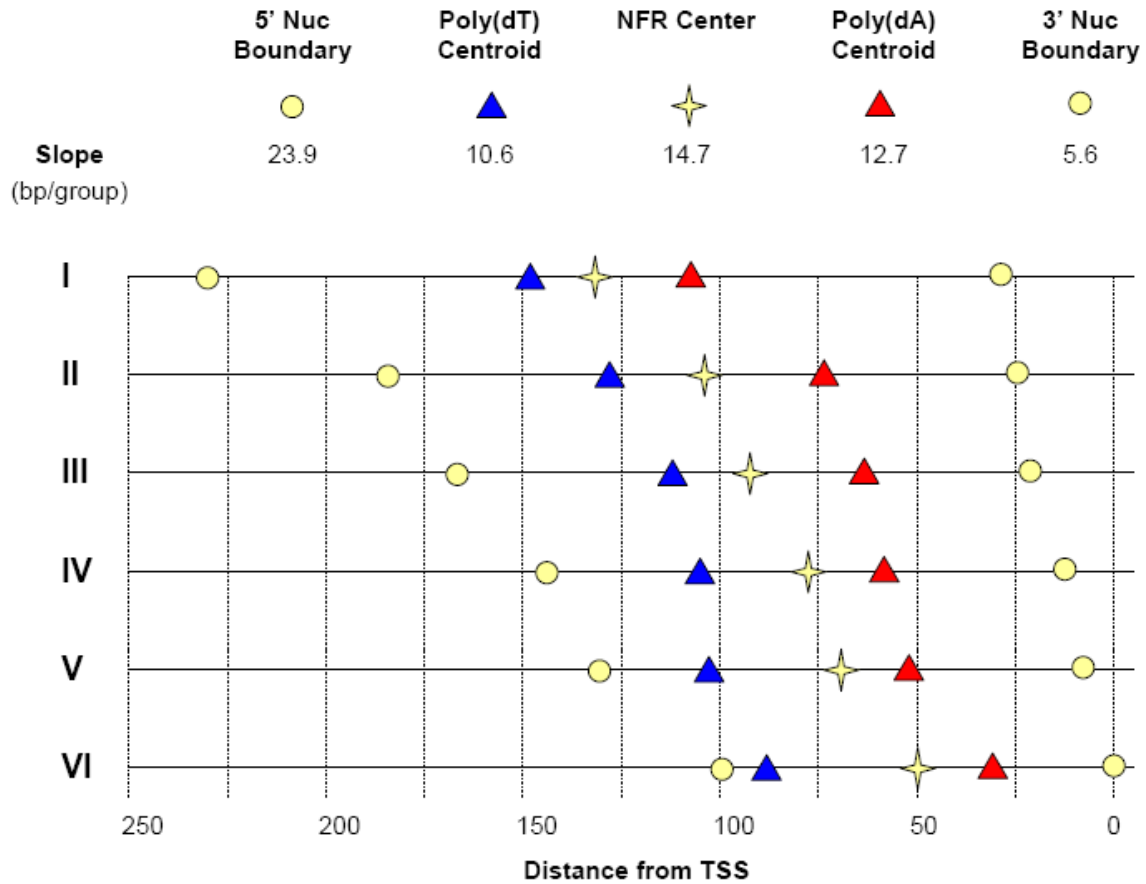
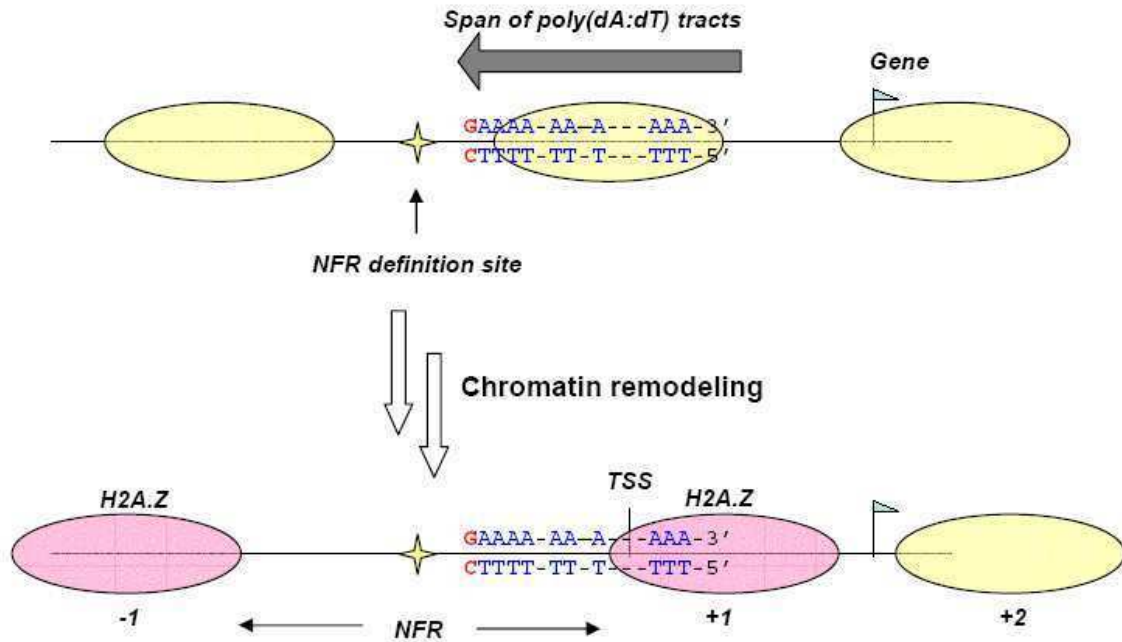
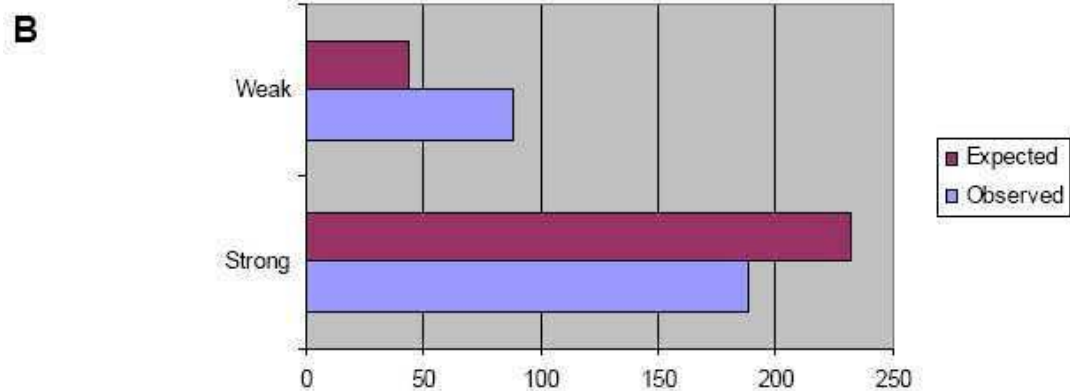
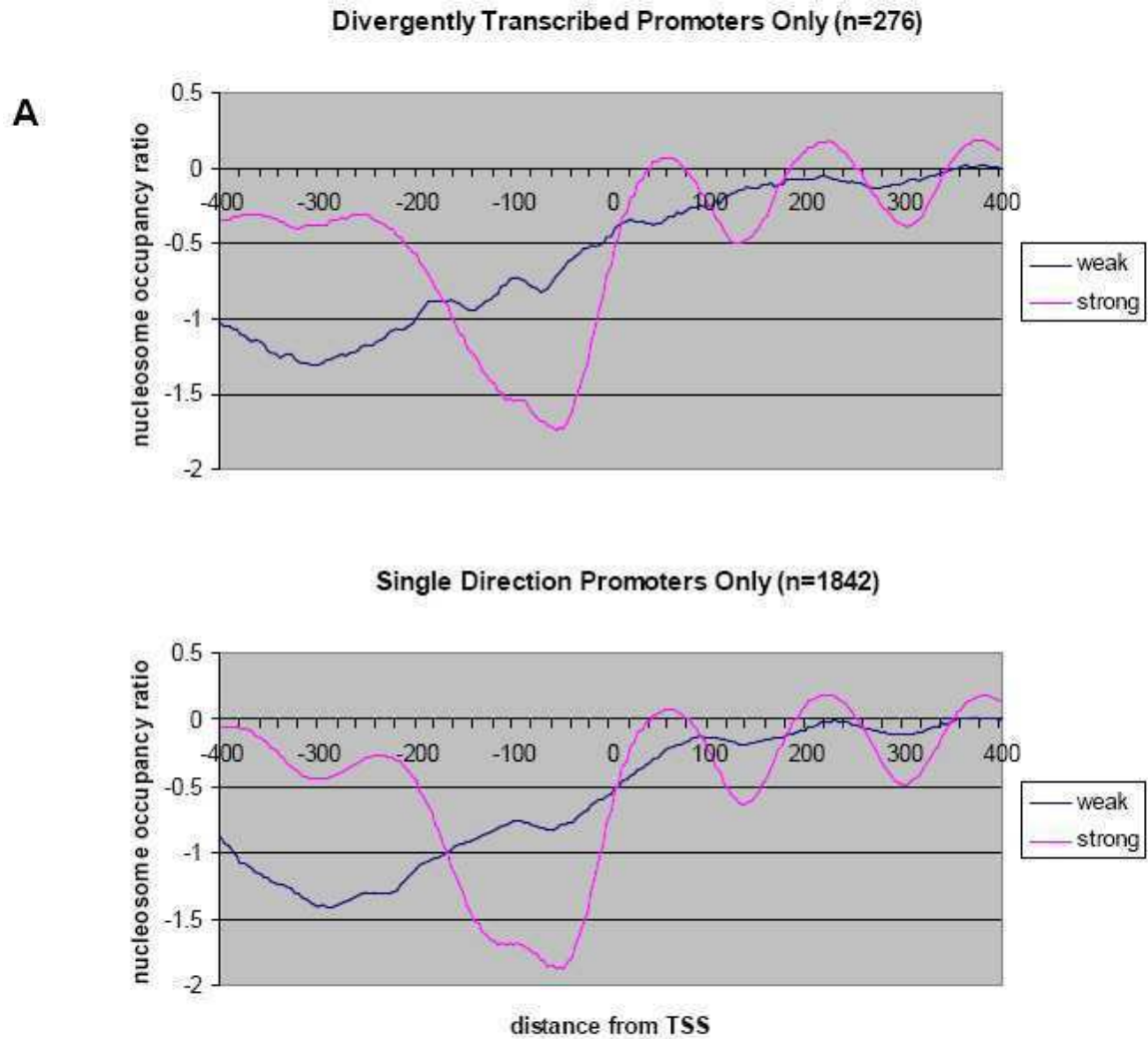


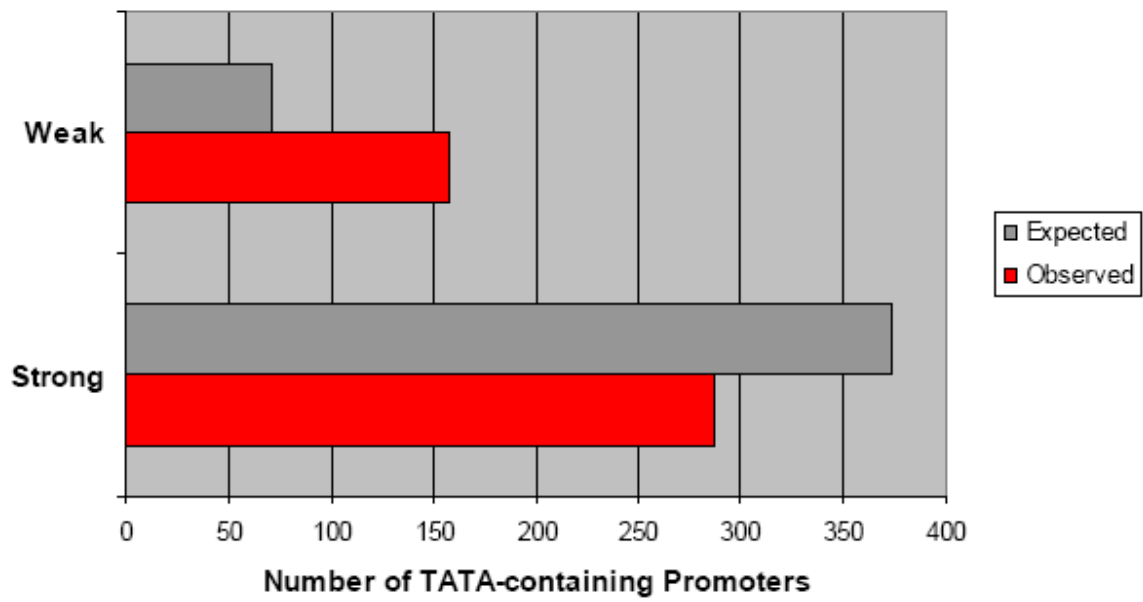
Figure 8. Summary of mechanistic hypotheses describing how poly(dA:dT) tracts can lead to the formation of nucleosome-free regions. 5' G:C capped tracts in conjunction with uncapped tracts mediate the definition of the NFR central coordinate. Tracts may facilitate the action of a chromatin-remodeling complex such as RSC, which gives rise to the nucleosome-free pattern. H2A.Z deposition into NFR-flanking nucleosomes may also be functionally relevant in this process.



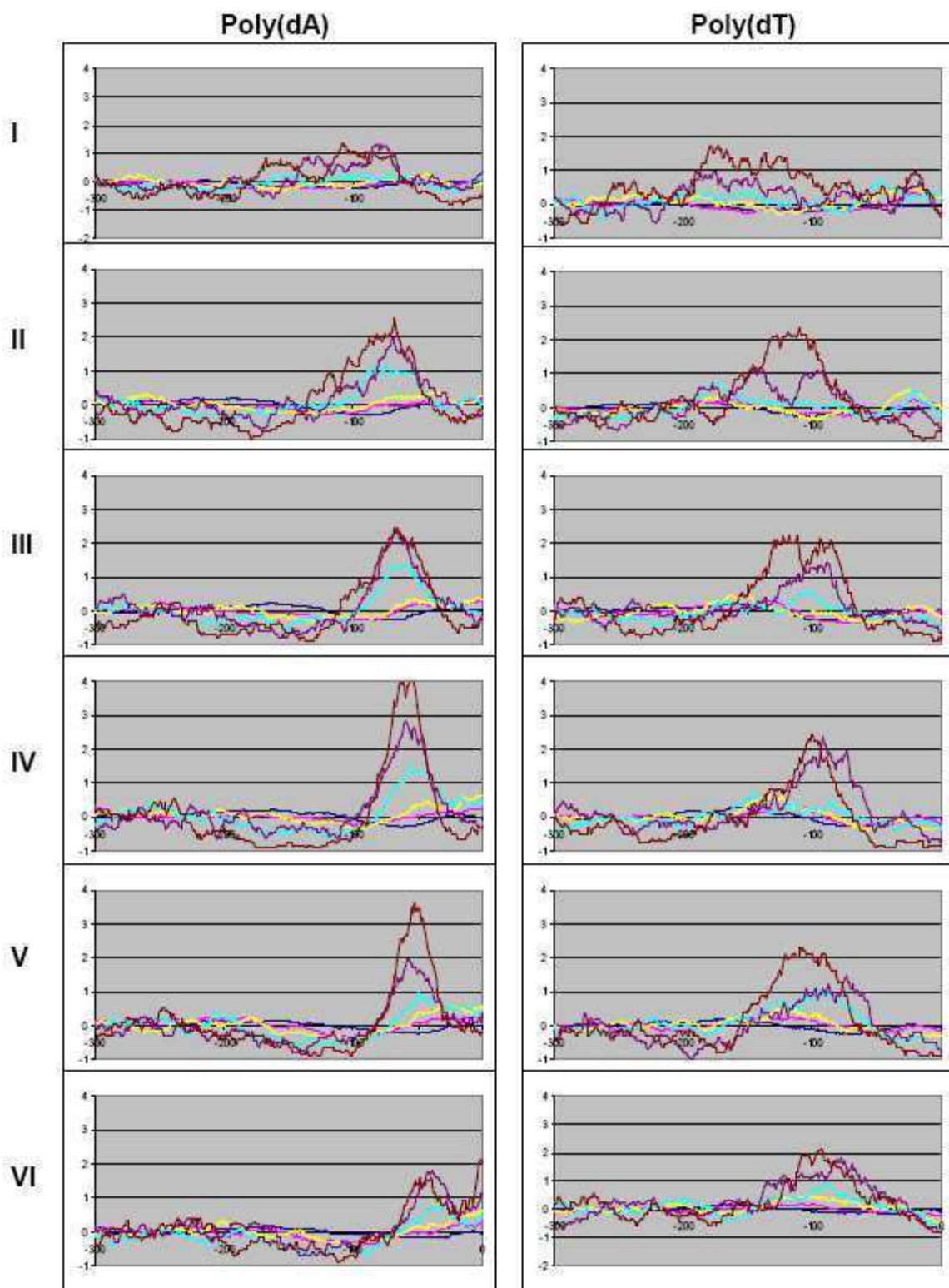
Supplementary Figure 1 **A)** Average nucleosome profiles for divergently-transcribed vs. single-direction promoters. **B)** Observed vs. expected number of divergently-transcribed promoters in strong- vs. weak-NFR classes.



Supplementary Figure 2 Observed vs. expected number of TATA-box containing promoters in strong- vs. weak-NFR classes.



Supplementary Figure 3 Poly(dA:dT) enrichments in subgroups of the strong-NFR class.



References

1. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389: 251-260.
2. Li B, Carey M, Workman JL (2007) The role of chromatin during transcription. *Cell* 128: 707-719.
3. Kornberg RD, Lorch Y (1999) Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell* 98: 285-294.
4. Lam FH, Steger DJ, O'Shea EK (2008) Chromatin decouples promoter threshold from dynamic range. *Nature* 453: 246-250.
5. Thastrom A, Lowary PT, Widlund HR, Cao H, Kubista M, et al. (1999) Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J Mol Biol* 288: 213-229.
6. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, et al. (2006) A genomic code for nucleosome positioning. *Nature* 442: 772-778.
7. Ioshikhes IP, Albert I, Zanton SJ, Pugh BF (2006) Nucleosome positions predicted through comparative genomics. *Nat Genet* 38: 1210-1215.
8. Segal MR (2008) Re-cracking the nucleosome positioning code. *Stat Appl Genet Mol Biol* 7: Article14.
9. Peckham HE, Thurman RE, Fu Y, Stamatoyannopoulos JA, Noble WS, et al. (2007) Nucleosome positioning signals in genomic DNA. *Genome Res* 17: 1170-1177.
10. Lee W, Tillo D, Bray N, Morse RH, Davis RW, et al. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* 39: 1235-1244.
11. Mavrigh TN, Ioshikhes IP, Venters BJ, Jiang C, Tomsho LP, et al. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* 18: 1073-1083.
12. Kornberg R (1981) The location of nucleosomes in chromatin: specific or statistical. *Nature* 292: 579-580.
13. Kornberg RD, Stryer L (1988) Statistical distributions of nucleosomes: nonrandom locations by a stochastic mechanism. *Nucleic Acids Res* 16: 6677-6690.
14. Gupta S, Dennis J, Thurman RE, Kingston R, Stamatoyannopoulos JA, et al. (2008) Predicting human nucleosome occupancy from primary sequence. *PLoS Comput Biol* 4: e1000134.
15. Rando OJ, Ahmad K (2007) Rules and regulation in the primary structure of chromatin. *Curr Opin Cell Biol* 19: 250-256.
16. Yuan GC, Liu YJ, Dion MF, Slack MD, Wu LF, et al. (2005) Genome-scale identification of nucleosome positions in *S. cerevisiae*. *Science* 309: 626-630.
17. Shivaswamy S, Bhinge A, Zhao Y, Jones S, Hirst M, et al. (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol* 6: e65.
18. Mavrigh TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, et al. (2008) Nucleosome organization in the *Drosophila* genome. *Nature* 453: 358-362.
19. Ozsolak F, Song JS, Liu XS, Fisher DE (2007) High-throughput mapping of the chromatin structure of human promoters. *Nat Biotechnol* 25: 244-248.

20. Albert I, Mavrich TN, Tomsho LP, Qi J, Zanton SJ, et al. (2007) Translational and rotational settings of H2A.Z nucleosomes across the *Saccharomyces cerevisiae* genome. *Nature* 446: 572-576.
21. Raisner RM, Hartley PD, Meneghini MD, Bao MZ, Liu CL, et al. (2005) Histone variant H2A.Z marks the 5' ends of both active and inactive genes in euchromatin. *Cell* 123: 233-248.
22. Behe MJ (1995) An overabundance of long oligopurine tracts occurs in the genome of simple and complex eukaryotes. *Nucleic Acids Res* 23: 689-695.
23. Karlin S, Blaisdell BE, Sapolsky RJ, Cardon L, Burge C (1993) Assessments of DNA inhomogeneities in yeast chromosome III. *Nucleic Acids Res* 21: 703-711.
24. Yuan GC, Liu JS (2008) Genomic sequence is highly predictive of local nucleosome depletion. *PLoS Comput Biol* 4: e13.
25. Suter B, Schnappauf G, Thoma F (2000) Poly(dA.dT) sequences exist as rigid DNA structures in nucleosome-free yeast promoters in vivo. *Nucleic Acids Res* 28: 4083-4089.
26. Anderson JD, Widom J (2001) Poly(dA-dT) promoter elements increase the equilibrium accessibility of nucleosomal DNA target sites. *Mol Cell Biol* 21: 3830-3839.
27. Holdaway RM, White MW (1990) Computational neural networks: enhancing supervised learning algorithms via self-organization. *Int J Biomed Comput* 25: 151-167.
28. Tirosh I, Barkai N (2008) Two strategies for gene regulation by promoter nucleosomes. *Genome Res* 18: 1084-1091.
29. Struhl K (1985) Naturally occurring poly(dA-dT) sequences are upstream promoter elements for constitutive transcription in yeast. *Proc Natl Acad Sci U S A* 82: 8419-8423.
30. Schlapp T, Rodel G (1990) Transcription of two divergently transcribed yeast genes initiates at a common oligo(dA-dT) tract. *Mol Gen Genet* 223: 438-442.
31. Macisaac KD, Gordon DB, Nekludova L, Odom DT, Schreiber J, et al. (2006) A hypothesis-based approach for identifying the binding specificity of regulatory proteins from chromatin immunoprecipitation data. *Bioinformatics* 22: 423-429.
32. Getts RC, Behe MJ (1992) Isolated oligopurine tracts do not significantly affect the binding of DNA to nucleosomes. *Biochemistry* 31: 5380-5385.
33. Bao Y, White CL, Luger K (2006) Nucleosome core particles containing a poly(dA.dT) sequence element exhibit a locally distorted DNA structure. *J Mol Biol* 361: 617-624.
34. Losa R, Omari S, Thoma F (1990) Poly(dA).poly(dT) rich sequences are not sufficient to exclude nucleosome formation in a constitutive yeast promoter. *Nucleic Acids Res* 18: 3495-3502.
35. Puhl HL, Behe MJ (1995) Poly(dA).poly(dT) forms very stable nucleosomes at higher temperatures. *J Mol Biol* 245: 559-567.
36. Verdone L, Camilloni G, Di Mauro E, Caserta M (1996) Chromatin remodeling during *Saccharomyces cerevisiae* ADH2 gene activation. *Mol Cell Biol* 16: 1978-1988.
37. Narlikar GJ, Fan HY, Kingston RE (2002) Cooperation between complexes that regulate chromatin structure and transcription. *Cell* 108: 475-487.

38. Cairns BR, Lorch Y, Li Y, Zhang M, Lacomis L, et al. (1996) RSC, an essential, abundant chromatin-remodeling complex. *Cell* 87: 1249-1260.
39. Parnell TJ, Huff JT, Cairns BR (2008) RSC regulates nucleosome positioning at Pol II genes and density at Pol III genes. *EMBO J* 27: 100-110.
40. Saha A, Wittmeyer J, Cairns BR (2005) Chromatin remodeling through directional DNA translocation from an internal nucleosomal site. *Nat Struct Mol Biol* 12: 747-755.

Chapter 3

Transcription Factor-Chromatin Profiles in the Yeast Genome

Abstract

When transcription factors interact with their cognate binding sites on DNA, this interaction must take place in the context of the chromatin environment in which the DNA is situated. Transcription factor binding sites which are located in the context of nucleosomal DNA may have different binding and functional properties than sites which are located in nucleosome-free DNA. In this study we systematically survey 122 transcription factors in the *S. cerevisiae* genome to determine how each is affected by chromatin context. High throughput data for TF binding, conservation, and nucleosome occupation is integrated into several numerical parameters for every TF, each of which indicates individual relationships between the transcriptional regulator and its nucleosomal setting. While the spectrum of TF-chromatin profiles in the yeast genome is quite diverse, we find that certain signatures, such as high preference for binding and conservation in nucleosome-free regions, predominate. Thus our survey presents a valuable first step in the systematic assessment of TF-chromatin relationships in a whole organism.

Introduction

The physical accessibility of regions of eukaryotic genomic DNA to its interacting protein machinery is strongly influenced by local chromatin structure. When Transcription Factors (TFs) recognize their cognate DNA binding sequences, they must do so in the context of the chromatin's physical space. Intuitively, because contact with nucleosomes sterically occludes a substantial fraction of the DNA's binding surfaces, it seems that nucleosomes should tend to hinder the binding of most proteins. Indeed, the binding of certain TFs (e.g. Pho4p) have been shown to exhibit a strong dependence on nucleosomal context, where TF binding depends on the removal of local nucleosomes through ATP-dependent remodeling[1,2]. This notion is also supported by the general correspondence between transcriptional activation and promoter nucleosome deficiency[3]. It may not be the case, however, that chromatin must hinder the binding of all DNA-interacting proteins. Hypothetically, nucleosomes may actually enhance the binding of DNA-interacting proteins by providing cooperative interactions. Currently, this question has not yet been systematically addressed.

A separate but related question to consider is how chromatin affects the functionality of TFs once they are actually bound to DNA. It may be that some TFs behave identically regardless of underlying nucleosomes while other TFs exhibit behaviors that strongly depend on chromatin context. To address these questions, we must consider the functional role of TFs in chromatin contexts in a manner which is independent from TF binding.

In this study, we use a novel computation method to consider the interrelationships between three different and possibly independent binary parameters for

all known TF binding sequences: TF binding, TF functionality and nucleosome occupation. This is made possible by the recent availability of such data on a genome-wide scale for *S. cerevisiae*. These include 1) a comprehensive catalog of motifs and binding sites for 122 TFs[4,5], 2) the mapping of conserved regulatory sequences through comparative genomic analyses using additional *sensu stricto* yeasts[6] and 3) the high-resolution (4bp) mapping of nucleosome positions in the yeast genome[7] (Fig1). These data offer an unprecedented opportunity to systematically dissect the relationships between TF binding, nucleosome positions, and TF function on a genome-wide scale. In our approach, for every TF we compile several parameters, each of which quantifies a particular TF-chromatin relationship. These parameters, in aggregate, form a nuanced and complex portrait of the relationship between regulatory factors and chromatin in a simple eukaryotic organism.

Results

Tabulation of binary parameters for each TFBS

To comprehensively identify Transcription Factor Binding Sites (TFBS) in the *S. cerevisiae* genome, we used annotated position-specific scoring matrices (PSSM) to select the top 1000 sites for each of 122 TFs. For each TFBS, we tabulated three independent binary parameters:

a: true if TFBS is in a nucleosome-free region, false otherwise

b: true if TFBS is bound to factor, false otherwise

c: true if TFBS is functionally conserved, false otherwise

Where nucleosome-free regions are demarked by troughs in the smoothed nucleosome-occupancy tiling array data, binding is assigned according to ChIP-chip, and functional conservation is assigned according to sequence conservation among *sensu stricto* yeasts (see Methods for details). Thus, the set of parameters (a, b, c) describe each TFBS in the yeast genome according to the independent criteria of nucleosome-occupancy, regulator protein binding, and functional conservation. Finally, for each TF we tabulate the number of TFBS in each of the eight categories defined over the space of all possible values of (a, b, c) .

Derived parameters capture TF-chromatin relationships for each TF

After tabulating these parameters for all TFs across the entire genome, we compute for each TF a set of three derived parameters, each of which measures a particular TF-chromatin relationship. The first derived parameter, α , measures the intrinsic tendency for a TF to have sites located in NFRs:

$$\alpha = \frac{NFR/NOR}{NFR_{bg}/NOR_{bg}}$$

Where NFR is the number of TFBS in nucleosome-free regions, NOR is the number of TFBS in nucleosome-occupied regions, and NFR_{bg} and NOR_{bg} are the total lengths of NFR and NOR sequences in the genome. This parameter gives the ratio of binding sites for a given TF in NFRs vs. NORs and is normalized by the expected ratio given by the relative lengths of such sequences. Intuitively, if the placement of sites for a given TF has no preference or aversion for nucleosomes, then α is unity.

We also introduce another derived parameter β , which quantifies the relative likelihood that a given TF will be bound to its cognate site in nucleosome-free versus nucleosome-occupied regions:

$$\beta = \frac{NFR_{bound} / NOR_{bound}}{NFR_{unbound} / NOR_{unbound}}$$

Where the subscripts indicate that we only consider those sites where $b=true$ for *bound* and $b=false$ for *unbound*. Here the ratio in the numerator gives the intrinsic odds that a TF will be bound to a given site in a nucleosome-free region. The denominator gives a similar ratio for nucleosome-occupied regions. The ratio of these two ratios, then, gives the relative odds that a TF will bind with the presence or absence of a nucleosome and addresses the question of how the nucleosomal environment affects TF binding. TFs with $\beta > 1$ have a built-in preference for preferring to bind to nucleosome-free regions whereas those TFs with $\beta < 1$ prefer to bind to nucleosome-occupied regions.

We also introduce a derived parameter γ , which is similar to β except that it addresses functionally conserved sites rather than bound sites:

$$\gamma = \frac{NFR_{conserved} / NOR_{conserved}}{NFR_{nonconserved} / NOR_{nonconserved}}$$

Thus, similarly to the function of β , γ addresses the question of how nucleosomal context affects the likelihood that a TFBS is functionally conserved. TFs with $\gamma > 1$ are more likely to be functionally conserved in nucleosome-free regions whereas TFs with $\gamma < 1$ are more likely to be functionally conserved in nucleosome-occupied regions.

Finally, we introduce the derived parameter δ , which measures a quantity similar to α except that only bound and conserved sites are taken into account. This metric can be seen as a more accurate version of α when a TF has a large number of bound and conserved sites, but may be less accurate otherwise.

$$\delta = \frac{NFR_{bound,conserved} / NOR_{bound,conserved}}{NFR_{bg} / NOR_{bg}}$$

Distributions of derived parameters for the yeast transcriptome

The distributions of all four derived parameters are shown in Figure 2 as logarithmic values. The histogram for α shows a relative normal distribution with a slight skew toward negative values. This indicates that, on the whole, unfiltered transcription factor binding sites are largely unbiased in terms of preference for nucleosome-free or nucleosome-occupied positions. In contrast, TFBS which are both bound and conserved have a fairly strong bias for nucleosome-free regions, as is evinced by the distribution of δ . Despite this bias, a large peak is present for the distribution of δ at 0, indicating that a significant portion of TFs in the yeast genome have little NFR bias even when binding and conservation are taken into account. The binding and conservation preferences of TFs also show a fairly broad spectrum. The distribution of β has a slight positive bias where the distribution of γ , after correction for overall conservation bias, is fairly centered at 0.

Hierarchical clustering of TFs based on derived parameters

In order to order TFs into coherent groups with similar TF-chromatin properties, we hierarchically clustered all 122 TFs according to the values of the four derived parameters (Figure 3). Several prominent patterns emerged. One of the most striking groupings consists of 11 TFs which have positive values for all four derived parameters (cluster 3). These TFs, which include the general transcription factors Reb1p and Abf1p

(see Chapter 2 for the significant thereof), are characterized by some of the strongest values of α among all TFs, indicating the strongest intrinsic preference to reside in NFRs. TFs in this group also have consistently positive values of β and γ : sites are more likely to be bound and conserved in nucleosome-free regions.

Related to cluster 3 is cluster 2, a large group of 34 TFs which include, among others, the general factors Rap1p and Cbf1p, the adaptor protein Mcm1p, and the Gcn4 regulatory protein. The true distinction between clusters 3 and 2 is unclear: cluster 2 generally has smaller values of α , but remains similar in terms of δ ; the differences between these two groups are subtle.

In direct contrast to clusters 2 and 3 are clusters 5 and 6, which are characterized by predominantly negative values of α , β , and γ . Our interpretation of these groups is that they represent TFs which preferentially bind to sites within nucleosome-occupies stretches of DNA. Similarly, those sites within nucleosome-occupies regions are more likely to be functionally relevant. While the reasons for such preference are not currently understood, some of this preference may be based on structural reasons. Intriguingly, cluster5 contains a large number of transcription factors of the bZIP family such as Arr1p, Cst6p and Yap6p (Saccharomyces Genome Database, www.yeastgenome.org). Thus, the preference of these TFs for nucleosomal sites may be tied to their structures.

Finally, in two of the observed groups the values of β have the opposite signs as the values of γ . In group 1 β tends to be positive and γ negative, while in group 3 the exact opposites are true. This demonstrates that the binding preferences and functional preferences of a given TF need not be in accord with respect to their chromatin environment.

Conclusion

We have described a preliminary but systematic computational study of the relationships between TFs and their chromatin environments for 122 TFs in the yeast genome. Our survey reveals substantial diversity in such TF-chromatin relationships. While a majority of TFs, as might be expected, prefer to bind and function within nucleosome-free environments, a significant minority of TFs have exactly the opposite behavior and instead prefer nucleosomal environments. Another result is that the binding

preferences and functional preferences of TFs appear to be independent of one another, as many TFs have opposing values. These initial data should provide a good foundation for continuing work in elucidating how chromatin context affects transcription factor function.

Methods

Regulator binding site annotation. We obtained position-specific scoring matrices for 122 yeast regulators from the Fraenkel lab website (http://fraenkel.mit.edu/improved_map/) and used them to score the ~3 megabases of yeast intergenic sequence. The top 1000 scoring positions were selected to represent the most likely binding sites for each regulator. We also imposed a minimum score of 7.8, which effectively filters out hits from very short matrices (typically 5-6 bases in length) with low information content. We designate the lowest score for each regulator as its minimum scoring criteria. We consider a site to be conserved if the aligned positions of at least two *sensu stricto* orthologs meet the minimum scoring criteria for that regulator. We consider a site to be bound by its regulator if the p-value of its corresponding probe for the YPD ChIP-chip experiment is below 0.01.

Nucleosome-free region annotation. To locate nucleosome-free regions, we applied a heuristic peak-selection strategy to the whole-genome nucleosome occupancy data by Lee et al (<http://chemogenomics.stanford.edu/supplements/03nuc/index.html>). First, we smoothed out local irregularities in the array data by replacing each probe value with the median value of an 11-probe window centered on the probe. Then, for each non-3' only intergenic region, we took the mean and standard deviation of all probes between 500 bp upstream and 500 bp downstream of the IGR. Troughs in the array data which fell below one standard deviation were considered as NFR candidates. Troughs with center-to-center distances below 200 bp were iteratively merged. The final set of troughs which were within 1000 bp of a start codon and which were at least 100 bp in length were selected as nucleosome-free regions.

Normalization of γ Because nucleosome-free regions in yeast typically have higher conservation rates and are more easily alignable, the values of γ_{raw} collectively show a large skew toward ratios below 1. To normalize these statistics, we permute the PSSM for each regulator 10 times and annotate sites for each permuted matrix in the same way as the original ones. The corrected conservation preference is then:

$$\gamma_{\text{corrected}} = \gamma_{\text{raw}} - \text{median}(\gamma_{\text{permuted}})$$

References

1. Venter U, Svaren J, Schmitz J, Schmid A, Horz W (1994) A nucleosome precludes binding of the transcription factor Pho4 in vivo to a critical target site in the PHO5 promoter. *EMBO J* 13: 4848-4855.
2. Lam FH, Steger DJ, O'Shea EK (2008) Chromatin decouples promoter threshold from dynamic range. *Nature* 453: 246-250.
3. Lee CK, Shibata Y, Rao B, Strahl BD, Lieb JD (2004) Evidence for nucleosome depletion at active regulatory regions genome-wide. *Nat Genet* 36: 900-905.
4. MacIsaac KD, Wang T, Gordon DB, Gifford DK, Stormo GD, et al. (2006) An improved map of conserved regulatory sites for *Saccharomyces cerevisiae*. *BMC Bioinformatics* 7: 113.
5. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, et al. (2002) Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* 298: 799-804.
6. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423: 241-254.
7. Lee W, Tillo D, Bray N, Morse RH, Davis RW, et al. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* 39: 1235-1244.

Figures

Figure 1. Deriving TF-chromatin profiles for transcription factors in the *S. cerevisiae* genome. A preliminary set of TFBS are derived by scanning the intergenic sequence with 122 PSSMs. Sites are further filtered using ChIP-chip binding data, *sensu stricto* sequence conservation data, and nucleosomal occupation data.

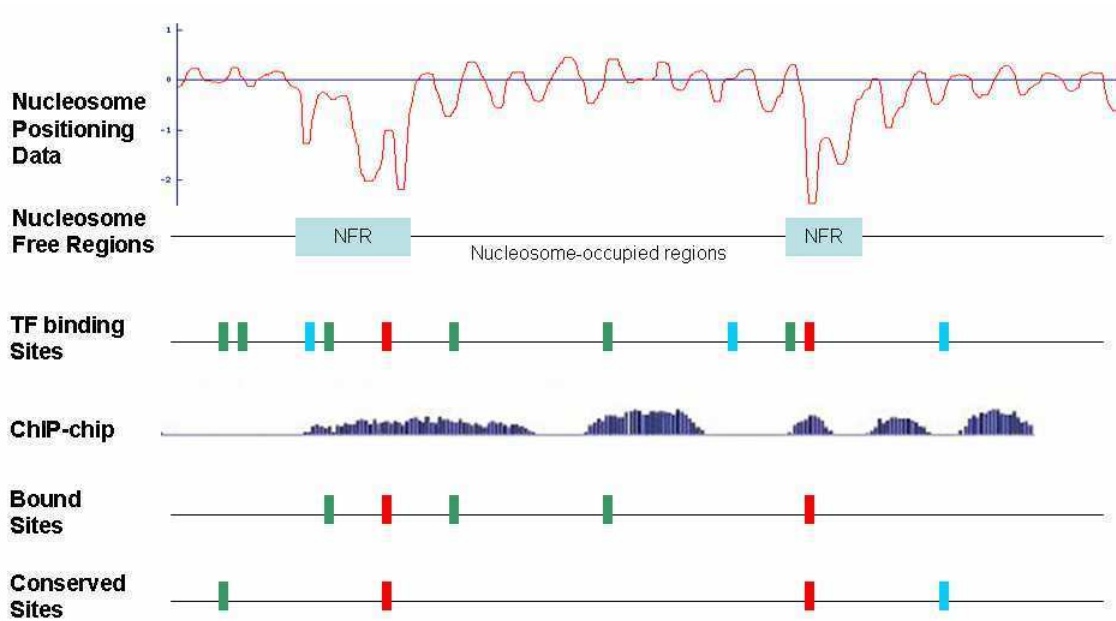
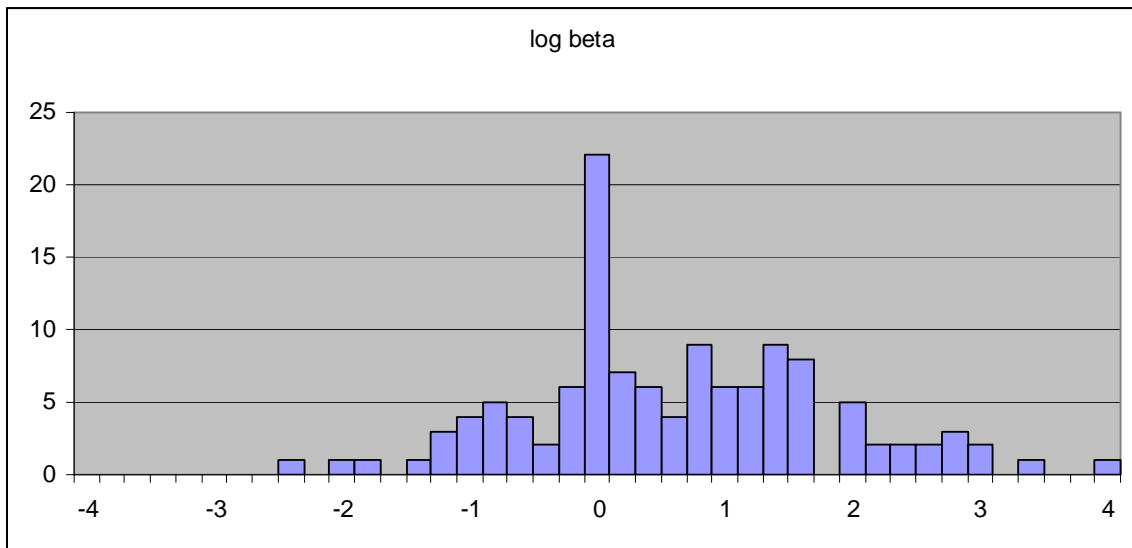
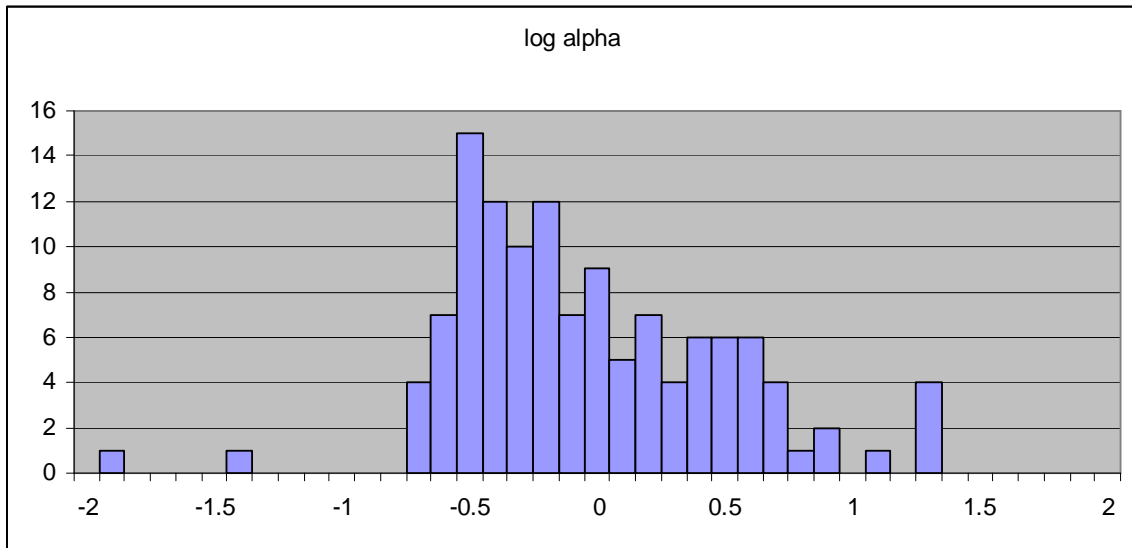


Figure 2. Distributions of the derived parameters α , β , γ , δ



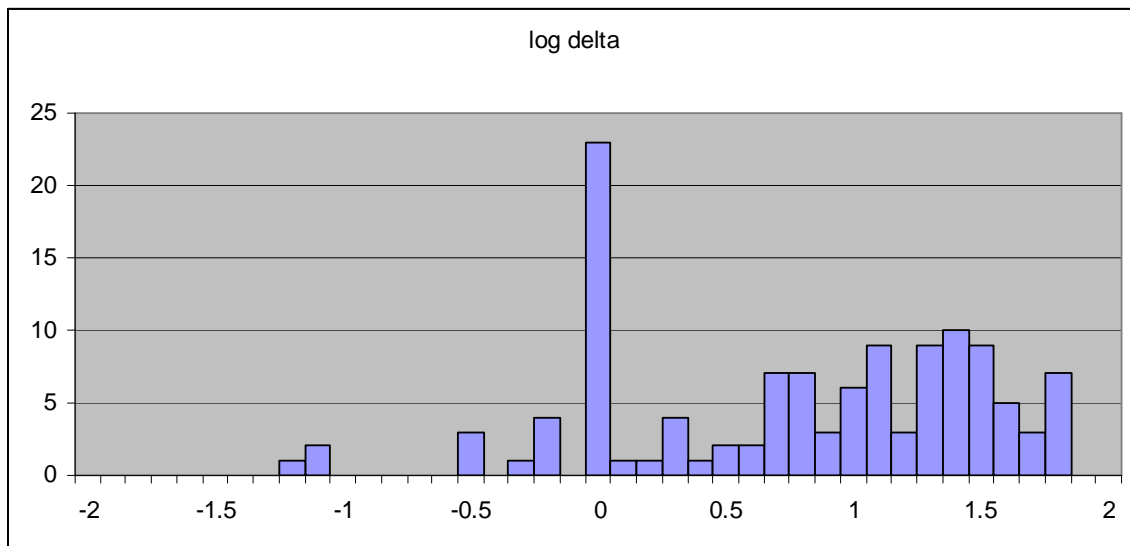
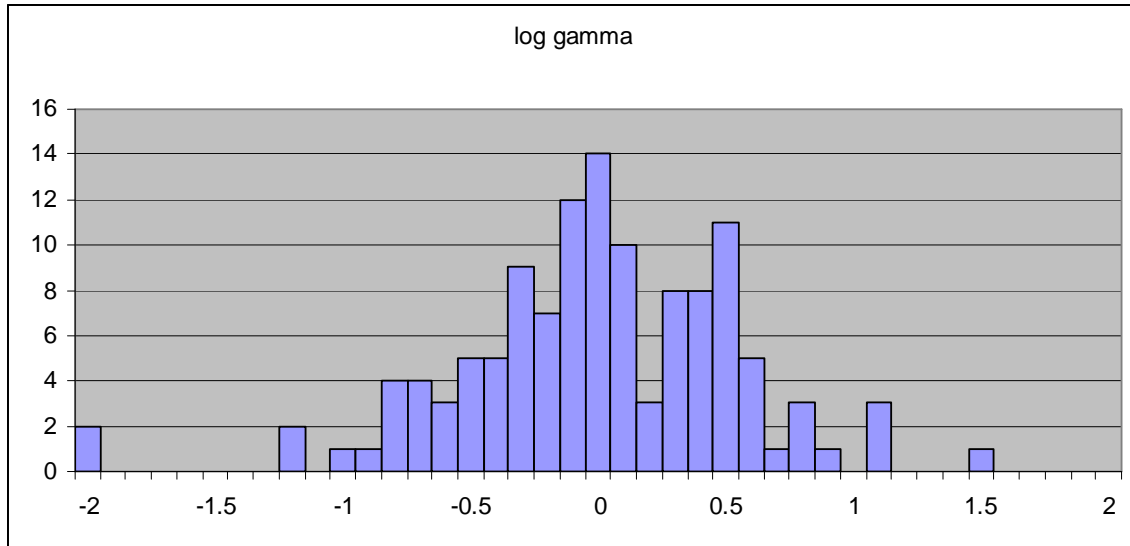
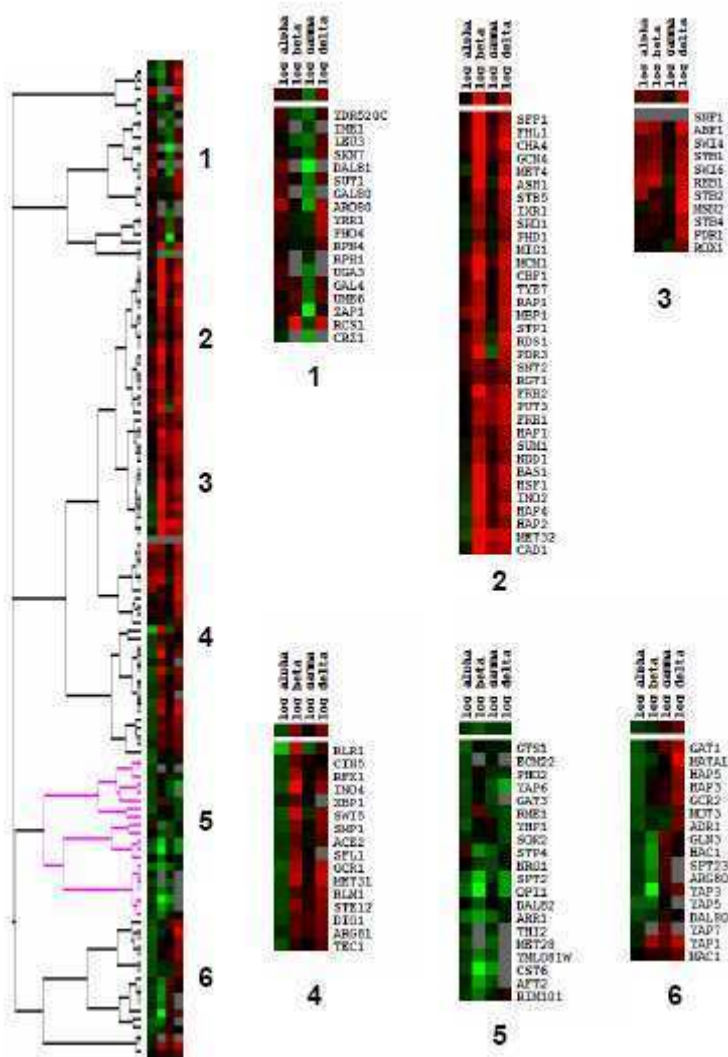


Figure 3. Hierarchical clustering analysis of 122 TFs according to derived parameters.



UCSF Library Release

Publishing Agreement

It is the policy of the University to encourage the distribution of all theses and dissertations. Copies of all UCSF theses and dissertations will be routed to the library via the Graduate Division. The library will make all theses and dissertations accessible to the public and will preserve these to the best of their abilities, in perpetuity.

Please sign the following statement:

I hereby grant permission to the Graduate Division of the University of California, San Francisco to release copies of my thesis or dissertation to the Campus Library to provide access and preservation, in whole or in part, in perpetuity.

Randy Wu

1/09/09

A handwritten signature in black ink, appearing to read "Randy Wu", with a long, sweeping flourish extending downwards and to the right.