Fitness landscape for nucleosome positioning

Donate Weghorn and Michael Lässig

Institute for Theoretical Physics, University of Cologne, 50937 Cologne, Germany

Edited by José N. Onuchic, Rice University, Houston, TX, and approved May 21, 2013 (received for review June 27, 2012)

Histone–DNA complexes, so-called nucleosomes, are the building blocks of DNA packaging in eukaryotic cells. The histone-binding affinity of a local DNA segment depends on its elastic properties and determines its accessibility within the nucleus, which plays an important role in the regulation of gene expression. Here, we derive a fitness landscape for intergenic DNA segments in yeast as a function of two molecular phenotypes: their elasticity-dependent histone affinity and their coverage with transcription factor binding sites. This landscape reveals substantial selection against nucleosome formation over a wide range of both phenotypes. We use it as the core component of a quantitative evolutionary model for intergenic DNA segments. This model consistently predicts the observed diversity of histone affinities within wild Saccharomyces paradoxus populations, as well as the affinity divergence between neighboring Saccharomyces species. Our analysis establishes histone binding and transcription factor binding as two separable modes of sequence evolution, each of which is a direct target of natural selection.

The positional organization of nucleosomes in eukaryotic cells is of key importance for the overall chromatin structure and, thus, for the regulation of gene expression (1–3). Nucleosomes form through binding of a histone octamer to a DNA sequence segment of average length 146 base pairs (bp), which wraps around the protein complex (4). Histone-bound DNA segments are interspersed with unbound “linker” segments. Particularly prominent features of this pattern are so-called nucleosome-depleted regions (NDRs). These are extended troughs in occupancy at least ~100 bp long, primarily located in intergenic DNA. Changes in nucleosome positioning affect the accessibility of local DNA segments for binding interactions with transcription factors and lead to observable changes of gene expression in yeast (3, 5).

Explaining two correlated molecular functions—histone binding and transcriptional regulation—in the same sequence segment may be seen as a chicken-and-egg problem (6–9). Is transcription factor binding the primary function, which displaces nucleosomes to sequence segments in which transcription is neutral or deleterious? Or, conversely, does nucleosome positioning constrain transcriptional interactions? Here, we address this problem by a quantitative evolutionary analysis of yeast genomes. We infer a fitness landscape for intergenic sequence segments that measures selection on their regulatory interactions and on local nucleosome formation. We capture these functions by two molecular phenotypes, the regulatory binding site content and the histone binding affinity, which reflect distinct biophysical characteristics of a DNA segment. The fitness landscape resulting from our analysis shows substantial selection acting jointly on transcriptional interactions and on nucleosome formation. Specifically, we find broad selection against histone binding—that is, in favor of nucleosome depletion—in sequence segments ~100 bp long, although individual nucleotides within these segments are under only weak selection. Our inference of selection on nucleosome positioning is corroborated by an evolutionary analysis within and across yeast species. We model the evolution of sequence segments by mutations, genetic drift, and selection given by our fitness landscape. This model explains the observed intraspecies diversity as well as the cross-species divergence of nucleosome positioning in a quantitative way. At the end of the paper, we discuss the implications of our findings for the functional and evolutionary relationship between nucleosome positioning and transcriptional regulation and, in a broader context, for the inference of selection on correlated molecular functions.

Our evolutionary analysis is based on established biophysical models that relate the histone binding affinity and the regulatory site content of a DNA segment to its nucleotide sequence. Several mechanisms are known to influence the local probability of nucleosome formation (8). Histone-affine DNA has a specific nucleotide composition that facilitates superhelical turns around the cylinder-shaped octamer (10, 11). In contrast, histone-repelling sequence contains homopolymeric adenine segments on one strand paired with thymine segments on the other strand; these A:T tracts confer a high rigidity to the DNA double strand (12, 13).

In addition, competition with other DNA-binding proteins (3, 14, 15), as well as active rearrangement through chromatin remodelers (16, 17), may alter histone binding to DNA. All these factors contribute, to different degrees, to the positioning of nucleosomes in vivo (15). Here, we choose one particular biophysical phenotype, the elasticity-mediated histone binding affinity, to map direct selection on nucleosome formation in yeast intergenic regions. Our finding of broad selection in favor of nucleosome depletion is consistent with the known functional role of NDRs. They reflect stable barriers in the histone binding energy landscape, which constrain the positioning of nucleosomes between them (18–21).

To infer regulatory binding sites in the yeast genome, we use standard statistical models of the position-dependent binding energy profile for specific transcription factors (22).

Our findings are consistent with previous results on the evolution of nucleosome positioning. About 70% of interspecific nucleosome architecture changes in yeast are caused by cis effects as opposed to trans-acting factors (23), which supports our inference of a local histone binding phenotype. At the level of sequence evolution, it has been shown that linker regions in yeast coding sequence are more conserved than regions of higher nucleosome occupancy (24, 25), in agreement with a previous analysis of chromosome III promoters (15). More specifically, A:T-loss nucleotide changes are reduced in NDRs compared with high-occupancy regions (26), which is consistent with A:T-rich sequence disfavoring nucleosome formation. Similar signatures of selection acting on nucleotide frequencies also have been found in the human lineage (27). It is important to note, however, that observations of sequence conservation do not distinguish the evolutionary signal of direct selection acting on a specific function, in this case nucleosome formation or transcriptional interactions, from selection acting on other, potentially unrelated functions encoded in the same sequence segment. This is why we base our study on biophysically grounded models: The statistics of a biophysical trait associated with a specific function will prove to be less confounded by apparent selection than summary sequence measures. Our

Author contributions: D.W. and M.L. performed research, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

1To whom correspondence should be addressed. E-mail: mlaessig@uni-koeln.de.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1210887110/-/DCSupplemental.
Depletion in living yeast cells. This motivates our joint analysis of dependence on both phenotypes: the number of annotated transcription factor binding sites (Fig. 1) and a large sequence target, even if individual nucleotide changes are under only weak selection.

Results

Phenotypes of Histone Binding and of Transcription Factor Binding. Wrapping DNA around histones necessitates specific elastic deformations of its double strand. We evaluate the energy cost of these deformations using the model of references (20, 28). The local energy cost depends on sequence content, because different nucleotide triplets have different a priori deformations in the unbound state. Given the genomic landscape of energy costs, the resulting nucleosome occupancy of a given sequence segment is determined by equilibrium thermodynamics. We call this phenotype the histone binding affinity of the segment. Our analysis uses the thermodynamic model and algorithm of references (20, 28) (for details, see Methods). This model successfully predicts the nucleosome positioning observed under in vitro conditions, that is, without the competitive binding of transcription factors (20). As expected, the ensemble average of each increases with increasing energy cost and increases with increasing histone density (or equivalently, with the associated chemical potential) (Fig. S1). For our genomic analysis, we use a chemical potential that reproduces the genome-wide occupancy average in vivo of about 80%. With these settings, we take as the best computable phenotype to measure the elasticity-mediated histone binding affinity of a given sequence segment. By definition, this phenotype is independent of the regulatory interactions encoded in that segment. We measure these interactions by an independent phenotype, given by the number of annotated transcription factor binding sites (Methods).

We can relate these phenotypes to the in vivo nucleosome positioning in *Saccharomyces cerevisiae*, which was measured in (3). In Fig. 1A, we evaluate the mean in vivo occupancy score _Ω_ for intergenic sequence segments of length 100 bp. We find a strong dependence on both phenotypes: _Ω_ is an increasing function of _ω_ and a decreasing function of _n_. We conclude that DNA rigidity and transcription factor binding jointly contribute to nucleosome depletion in living yeast cells. This motivates our joint analysis of selection on exactly these phenotypes, to which we now turn.

Phenotype-Dependent Fitness Landscape. To infer a map between phenotype and fitness, we compare the genomic distribution of phenotype value pairs, _W_(_ω_, _n_), with the corresponding distribution _P_0(_ω_, _n_) evaluated in a suitable null model. To obtain _W_(_ω_, _n_), we construct a tiling of the yeast genome into nonoverlapping segments of fixed length _ℓ_ = 100 bp. This procedure is designed to avoid overcounting in longer NDRs and to make the phenotype data comparable between segments (for details, see Methods). The resulting distribution _W_(_ω_, _n_) for intergenic sequence in *S. cerevisiae* is shown in Fig. S2A. As a genomic null model, we use uncorrelated random sequence, which implies that nucleotide triplets conferring specific local elasticity properties are scrambled in the null model. The resulting phenotype null distribution may be approximated as a product, _P_0(_ω_, _n_) = _P_0(_ω_) _P_0(_n_). We obtain the marginal distribution _P_0(_ω_) using the same tiling procedure as in the actual yeast genome (which ensures that our results are insensitive to its bioinformatic details). This distribution is shown as a black line in Fig. 2. The marginal distribution _P_0(_n_) can even be evaluated analytically, using the information content (or relative entropy) of the binding motifs of individual transcription factors. Details on both components of the null model are given in SI Text. The resulting joint distribution _P_0(_ω_, _n_) is shown in Fig. S2B.

We now can infer the scaled phenotype-fitness map _2NF_(_ω_, _n_) as the log-likelihood score of the genomic phenotype distribution and the null distribution (29, 30):

\[
2NF(\omega, n) = \log \frac{W(\omega, n)}{P_0(\omega, n)} + \text{const.}
\]  

All fitness values on the left-hand side are measured in units of 1/(2N), where _N_ is the effective population size. This landscape is defined up to an arbitrary constant, because only fitness differences (selection coefficients) enter the evolution of phenotype frequencies. Our inference of selection involves several assumptions. First, Eq. 1 is valid if nucleosome positioning is at an evolutionary equilibrium of mutations, genetic drift, and selection. This assumption is corroborated by our cross-species analysis described below. Second, the landscape _P_0(_ω_, _n_) is inferred from all intergenic sequence segments. The underlying uniformity assumption may be relaxed: If the fraction of segments under selection against histone binding is anywhere above ~20%, our inference of selection essentially remains unchanged in the regime of reduced affinity, _ω_ < 0.5 (SI Text and Fig. S3A). Similarly, our results are insensitive to variations of the tiling length _ℓ_ within the length range of functional NDRs, as shown in Fig. S3B.

The scaled fitness landscape _2NF_(_ω_, _n_) inferred for *S. cerevisiae* intergenic sequence is shown in Fig. 1B. It reveals substantial selection on both histone binding affinity and transcriptional regulation: We find scaled fitness differences _2NF_ ≤ 10 in our set of intergenic segments. Importantly, the selection on histone binding affinity is a primary effect; that is, the overrepresentation of NDRs...
in the yeast genome cannot be explained by direct selection on regulatory site content alone. Our finding of substantial direct selection on \( \omega \) gives an a posteriori justification for our choice of this phenotype. Before we discuss the implications of the inferred fitness landscape, we test its predictions for evolution of sequence-directed nucleosome positioning within and across species.

**Selection Against Nucleosome Formation.** As shown in Fig. 1B, the selection on histone binding affinity does not depend strongly on the regulatory phenotype \( n \). Therefore, it can be evaluated in good approximation from an effective fitness landscape for histone binding affinity, \( 2Nf(\omega) \) (red line), the log-likelihood of the distributions \( W(\omega) \) and \( P_0(\omega) \).

The effective fitness landscape shows that selection in favor of nucleosome depletion acts across a broad range of affinity values, beyond what could be considered a nucleosome-free region. This implies that there is predominantly directional selection on affinity changes,

\[
2N\Delta F = -\alpha \Delta \omega,
\]

with an average proportionality constant \( \alpha = 11 \pm 1 \) obtained from a linear fit to the function \( 2Nf(\omega) \) in the range \( \omega < 0.8 \). Affinity changes of \( |\Delta \omega| > 0.1 \) are under substantial selection, i.e., they lead to fitness changes of magnitude \( |2N\Delta F| > 1 \). However, most point mutations confer smaller affinity changes and are only weakly selected. The efficacy of selection on nucleosome formation is not caused by large effects of single mutations, but by the multitude of elasticity-changing mutations in an extended sequence segment.

**Selection on Affinity Polymorphisms.** We now show that the fitness landscape of Eq. 2 correctly predicts the frequency bias of intergenic single-nucleotide polymorphisms (SNPs) that is related to selection against nucleosome formation. From the Saccharomyces Genome Resequeencing Project, we obtained the genomes of 35 Saccharomyces paradoxus isolates and their alignments (Methods). We choose this species for the analysis because it has a simpler population structure than S. cerevisiae (31). We analyze SNPs in nonoverlapping intergenic NDRs with \( \omega < 0.4 \) identified on the S. paradoxus reference genome. To determine the SNP allele frequency as a function of the associated phenotypic effect, we compute the average binding affinity in the two subpopulations carrying either allele. In this way, we obtain a polarized phenotype difference \( \Delta \omega = \bar{\omega}_a - \bar{\omega}_d \), where \( \bar{\omega}_a \) denotes the larger and \( \bar{\omega}_d \) the smaller of the two subpopulation averages. Under selection against histone binding, we expect a decrease in the average frequency of the high-affinity allele, \( \langle \chi_+ \rangle \), with increasing deleterious effect. Fig. 3 shows the data points \( (\Delta \omega, \chi_+) \) and the resulting average frequencies in bins of the affinity difference. These data permit a linear fit of \( \langle \chi_+ \rangle \) as a function of \( \Delta \omega \),

\[
\langle \chi_+ \rangle(\Delta \omega) = \frac{1}{2} - \gamma \Delta \omega,
\]

with a proportionality constant \( \gamma = 1.0 \pm 0.1 \). On the other hand, our fitness landscape predicts the scaled selection coefficient \( \sigma \equiv 2N\Delta F = -\alpha \Delta \omega \) for each of these SNPs according to Eq. 2. Assuming approximate linkage equilibrium, the classic equilibrium allele frequency distribution \( p_{eq}(x;\sigma) \) then determines the expected frequency of the deleterious allele, \( \langle \chi_+ \rangle \) (Methods). To leading order, we obtain a linear dependence as in Eq. 3 with a predicted value \( \gamma_F = 1.4 \pm 0.1 \). This is in good agreement with the observed value for S. paradoxus polymorphisms. Here, we treat the S. paradoxus isolates as a mixed population. Performing this analysis separately for the three major subpopulations in the sample (31), we find that population structure has only a minor influence on the signal of selection (Fig. S4).

Our polymorphism analysis establishes a quantitative inference of selection on NDRs on a microevolutionary timescale, despite the fact that individual mutations are under only weak to moderate selection. Importantly, apparent selection acting on sequence traits other than those relevant to nucleosome depletion is generally random with respect to the phenotype polarization. Therefore, the expectation value of the frequency of the deleterious allele as a function of the selection coefficient, \( \langle \chi_+ \rangle(\sigma) \), is affected only to a small extent by sequence conservation, say, due to the presence of transcription factor binding sites.

**Conservation of Histone Binding Affinity and Equilibrium.** Our equilibrium theory of nucleosome positioning makes a definite prediction that population structure has only a minor influence on the signal of selection (Fig. S4).
Already over the distance between \textit{S. cerevisiae} and \textit{S. paradoxus}, the neutrally evolved sequences show a significant decrease in low-affinity counts, which is inconsistent with the data. For example, we obtain a conserved number of about 1.500 ± 40 non-overlapping intergenic NDRs with length 100 bp and \( \omega < 0.4 \) in the actual \textit{S. cerevisiae} and \textit{S. paradoxus} genomes. In contrast, the count of NDRs with the same characteristics drops to about 980 for simulated neutral evolution over the evolutionary distance between \textit{S. cerevisiae} and \textit{S. paradoxus}, and to 170 at neutral equilibrium. Similar results are obtained in a three-species comparison of \textit{S. cerevisiae}, \textit{S. paradoxus}, and \textit{Saccharomyces bayanus}.

The observed cross-species conservation of affinity distribution \( W(\omega) \) and NDR number corroborates the assumption of evolutionary equilibrium underlying our analysis. The equilibrium state is characterized by detailed balance: Between two species, the number of genome segments increasing in affinity above a given threshold equals the number of segments decreasing below the same threshold. As we show below, this turnover describes the occupancy variability of individual NDRs between species.

To test the predictions of our fitness model for the divergence statistics of histone affinity, we mapped the set of intergenic NDR segments with \( \omega < 0.4 \) in \textit{S. cerevisiae} onto their aligned segments in \textit{S. paradoxus} (Methods and SI Text). Fig. 4B shows the contour lines and binned averages of the resulting scatter plot \( (\omega_{\text{par}}, \omega_{\text{cer}}) \). These pairs have lower mean affinity values in \textit{S. cerevisiae} compared with \textit{S. paradoxus}. This merely reflects our choice of base species (the opposite effect is observed if the alignment is constructed from a base set of \textit{S. paradoxus} NDRs).

We can compare the actual process with \textit{in silico} evolution under selection, using a Wright–Fisher simulation of the \textit{S. cerevisiae} NDR sequences in the fitness landscape \( F(\omega) \) (for details, see Methods and SI Text). Fig 4B shows the binned average and standard deviation of the resulting conditional distribution \( P(\omega_{\text{par}} | \omega_{\text{cer}}) \) for cross-species phenotype evolution. We find both quantities to be in quantitative agreement with the observed divergence statistics between \textit{S. cerevisiae} and \textit{S. paradoxus}. We conclude that our fitness landscape captures selection in favor of nucleosome depletion also over longer evolutionary times.

We also can compare the cross-species data to simulations of neutral evolution. Across the whole range of affinity values on \textit{S. cerevisiae} NDRs, neutral evolution leads to an average affinity gain—i.e., an average loss of NDR function—that is inconsistent with the observed process. At the same time, the standard deviation of the cross-species affinity change is similar to the neutral value; i.e., the fitness landscape does not strongly constrain phenotype variability. This is in accordance with previous findings showing a high variance across loci in the divergence of both NDR occupancy and A:T enrichment (3).

**Discussion**

We have inferred a phenotype-fitness map \( F(\omega, n) \) for yeast intergenic sequence segments, which measures selection depending on histone binding affinity and regulatory site content (Fig. 1B). This map offers a quantitative solution to the chicken-and-egg problem posed in the introduction: Can we rank nucleosome positioning and transcriptional regulation with respect to their selective effects on intergenic sequence? As shown in Fig. 1B, fitness has a genuinely two-dimensional phenotype target: there are two chickens. Histone binding and transcription factor binding are separable primary modes of the evolution of intergenic DNA, subject to direct selection of comparable strength. The selection on histone binding spans an extended set of nucleosome-depleted intergenic segments, which have affinity values up to above 50%. This result contrasts with the merely passive role of DNA methylation that has been inferred from cell-type specific variations of the methylation pattern in human and mouse (33, 34).

Direct selection on nucleosome affinity has an important biological consequence. It establishes a set of nucleosome-depleted
regions that are earmarked for interactions with transcription factors. The reduced nucleosome affinity not only increases the equilibrium coverage with transcription factors, but also may speed up the search kinetics of factor molecules toward their binding sites. Because these effects are largely independent of the actual coverage with binding sites, they facilitate binding site turnover and the adaptive formation of new sites. At the same time, the directional selection against histone binding given by our fitness landscape does not favor a specific affinity value, which is consistent with the observed cross-species variability of the affinity phenotype. This may suggest a two-tier model of selection on nucleosome-depleted intergenic regions: Elasticity-mediated directional selection broadly reduces nucleosome coverage, whereas balancing selection jointly tunes nucleosome and transcription factor coverage to gene-specific values.

The phenotypes used in this paper, histone binding affinity and regulatory site content, are distilled from the underlying cellular biophysics. A phenotype-based inference of selection is particularly relevant for histone binding, a quantitative trait that has extended (>100 bp) sequence targets with small phenotypic effects of individual mutations. Only by mapping nucleotide changes onto an affinity phenotype can we infer substantial aggregate selection against nucleosome formation. However, given the complexity of the molecular machinery of transcriptional regulation and chromatin organization, our analysis in terms of just two phenotypes is necessarily incomplete. For example, histone binding in vivo is expected to depend on additional sequence features besides our elasticity-mediated binding phenotype (10). Integrating additional phenotypes into the inference of selection leads to a higher-dimensional fitness landscape, which can be analyzed for its principal directions of selection. The projection on the two phenotypes used in this paper likely will lead to an underestimate but will not generate a spurious signal of selection. A more comprehensive analysis can also address fitness interactions or interference selection; our results suggest an avenue to infer these effects by a phenotype-based approach.

From a broader perspective, this paper is a case study analyzing quantitative traits that are encoded in overlapping sequence and represent coupled molecular functions. This scenario is at some distance from idealized models of population genetics and quantitative genetics but probably is typical of a broader class of complex molecular functions, for which we can measure or infer at least some key phenotypes.

Methods

Histone Binding Affinity. The biophysical model for histone binding underlying our analysis follows (20, 28). This model defines a histone-binding free energy landscape $\Delta G(r)$ as a function of the $r$ genomic coordinate $r$ of a nucleosome. The free energy of a DNA sequence segment $(a_i,a_{i+1},...,a_{i+d-1})$ is given by

$$
\Delta G(i) = \sum_{r=i}^{i+d-3} \sum_{\ell=1}^{3} A_\ell \left[ \phi_{\ell}^{\text{nt}}(a_{i+\ell}) - \phi_{\ell}^{\text{nt}}(a_i) \right]^2,
$$

where $a_i = (a_i,a_{i+1},a_{i+2})$ denote trinucleotide subsegments; $\phi_{\ell}^{\text{nt}}(a_i)$ are the roll, twist, and tilt deformations in the nucleosome state, $\phi_{\ell}^{\text{nt}}(a_i)$ are the intrinsic deformations in the unbound state (35), $A_\ell$ denotes the corresponding elastic constants, and we use a core binding length $d=125$ bp (28). The statistics of nucleosome positioning is then given by standard equilibrium thermodynamics. It may be derived from the grand canonical partition function

$$
Z = \sum_{\Delta G} \exp \left[ -\beta \left( \sum_{r} -\Delta G(r) \right) \right]
$$

with the no-overlap constraint $\Delta G(r) = \delta_{r,0}$ for $r = 1, \ldots, N - 1$. The partition function depends on the temperature via $k_B T = \frac{1}{\beta}$ and on the chemical potential $\mu_r$ which are adjusted to in vivo conditions. This determines the expected single-nucleotide nucleosome occupancies (36),

$$
O(r) = \beta^{-1} \sum_{r=1}^{N} \frac{\partial \ln Z}{\partial \Delta G(r)}
$$

and the expected mean occupancy

$$
\bar{O}(r) = \sum_{r=1}^{N} O(r).
$$

over sequence segments of length $i$. The dependence of $\bar{O}$ on local binding energies and on the chemical potential is shown in Fig. S1.

Data Analysis. We used genomic sequences and their alignments from University of California, Santa Cruz (UCSC) Genome Browser (sacCer3) for the interspecific analysis of $S$. cerevisiae and $S$. paradoxus. Up to a threshold, insertions and deletions were corrected to exclude alignment uncertainties. This procedure did not affect our cross-species analysis (for details, see SI Text and Fig. S5). The resulting total sequence length was 7.7 × 10^6 bp, with 1.5 × 10^6 bp in intergenic regions (37). The second dataset, obtained from the Saccharomyces Genome Resequencing Project, contains aligned genomes of 35 $S$. paradoxus strains, including SNPs. This dataset has a well-separable subsample (31). To control for demographic effects, we partitioned this dataset into three groups (European, Far Eastern, and American). We obtained annotated transcription factor binding sites on $S$. cerevisiae from the SwissRegulon Portal (Feb 2012) (22). Only nonoverlapping binding sites with a posterior probability >0.5 were used. To identify low-occupancy regions predicted by our affinity model, we constructed a tiling of the genome into nonoverlapping segments of fixed length $i=100$ bp, using a dynamic programming algorithm with an upper bound of 0.95 of the predicted mean nucleosome occupancy $\bar{O}$ in each individual segment. Experimental in vivo nucleosome occupancy scores for $S$. cerevisiae were obtained from the Gene Expression Omnibus database (accession series GSE22211) (3) and processed to reduce the effects of measurement uncertainties (SI Text).

Polymorphism Statistics. To predict the expected deleterious allele frequency given by the fitness landscape, we use the equilibrium allele frequency spectrum for a two-allele locus, $p_{\sigma|\omega}(x) = \frac{1}{2}(1 - x)^{\sigma}x^{\omega}/Z_{\text{eq}}$, where $\sigma = 2N_{\text{AF}}$ is the scaled selection coefficient, $\mu = 2N_{\text{AF}}$ is the scaled neutral mutation rate, and $Z_{\text{eq}}$ is a normalization factor. From this distribution, we determine the allele frequency spectrum for polymorphic loci, $p_{\sigma|\omega}(k; m, n)$, in a set of $m$ isolates by binomial sampling ($k=1, \ldots, m-1$). This distribution produces an average frequency of the deleterious allele, $\langle x \rangle (\sigma) = 1/2 + an + O(\sigma^2)$, with a proportionality constant $a = 0.127$ (for $\mu = 0.02$).

Modeling Sequence Evolution. We use a Wright–Fisher simulation for a population of NDR sequences evolving under mutations, genetic drift, and selection, given by the fitness landscape $F(x)$. The evolutionary time for simulation of the cross-species evolution is chosen so that the average sequence divergence in the set of predicted NDRs equals the observed real value of 13%. Simulations of neutral evolution use the same model, but without selection. More details are given in SI Text and Fig. S5.

Acknowledgments. We thank Alain Arneodo for providing the sequence-based algorithm to compute histone binding energies and nucleosome occupancy (28) and Stephan Schiffels for kindly making his Wright-Fisher evolution model algorithm available. We also are grateful for stimulating discussions with Ville Mustonen. This work was supported by Deutsche Forschungsgemeinschaft Grant SFB 680, by German Federal Ministry of Education and Research Grant 0315893-SYbacol, and in part by the National Science Foundation (NSF) under Grant NSF PHY05-51164 during a visit to the Kavli Institute for Theoretical Physics (Santa Barbara, CA).


Supporting Information

Weghorn and Lässig 10.1073/pnas.1210887110

SI Text

Null Distributions of Nucleosome Affinity and of Regulatory Site Content. Our inference of selection is based on a comparison of the genomic count distribution \( W(\omega, n) \) with a null distribution \( P_0(\omega, n) \); these distributions are shown in Fig. S2. The null distribution is approximately of product form, \( P_0(\omega, n) = P_0(\omega)P_0(n) \). Its components are obtained as follows:

1. The distribution \( P_0(\omega) \) is obtained from random sequence with *Saccharomyces cerevisiae* genome-wide average nucleotide frequencies, using the same tiling procedure as for the real sequence. This tiling identifies sets of nonoverlapping segments, which avoids overcounting. We use a fixed inference length \( \ell = 100 \) bp, which makes the average occupancy values \( \omega \) comparable between different segments. The resulting distribution \( P_0(\omega) \) is shown in Fig. 2.

2. The distribution \( P_0(n) \) is estimated from the relative entropy (Kullback–Leibler divergence) between regulatory sites and background intergenic sequence (1). Each transcription factor is associated with relative entropy

\[
D = \sum_{i=1}^{L} \sum_{a} q_i(a) \log \left( \frac{p_i(a)}{p_0(a)} \right),
\]

where \( q_i(a) \) \( (i = 1, \ldots, L; \ a \in \{A, C, G, T\}) \) denotes its position weight matrix, and we use background nucleotide frequencies \( p_0(A) = p_0(T) = 0.33, \ p_0(C) = p_0(G) = 0.17 \). A given sequence segment containing \( n \) binding sites has a probability

\[
P_0 \sim \exp \left( -\frac{1}{n} \sum_{a=1}^{n} D_a \right)
\]

under the null model. By averaging over all segments with a given value of \( n \), we estimate the null distribution

\[
P_0(n) \sim \exp \left( -\langle D \rangle_n \right).
\]

In this analysis, we use the position weight matrices of 158 *S. cerevisiae* transcription factors, as given by the SwissRegulon Portal (Feb 2012) (2).

Inference of Selection on Nucleosome Binding Affinity. Here we show that the selection on nucleosome binding resulting from our analysis is insensitive to changes of the inference procedure:

1. The inference of a phenotype–fitness map as described in the main text implicitly assumes that all intergenic sequence segments counted in the distribution \( W(\omega) \) are under selection on histone binding. Because genes are arranged in close succession on the *S. cerevisiae* genome, intergenic regions are comparatively short. This suggests that a significant fraction of them indeed is functional and, hence, under comparable selection. Specifically, if we assume that the distribution \( W(\omega) \) contains a fraction \( \lambda \) of segments evolving under the fitness landscape \( F(\omega) \) and a fraction \( (1-\lambda) \) evolving neutrally, we decompose this distribution according to a mixture model,

\[
W(\omega) = \lambda Q(\omega) + (1-\lambda)P_0(\omega),
\]

with \( 0 < \lambda \leq 1 \). For example, assuming that selection is limited to segments with binding affinities \( \omega < 0.75 \), we can estimate \( \lambda \) from the data by minimization of \( \chi^2 \) between \( (1-\lambda)P_0(\omega) \) and \( W(\omega) \) in the high-\( \omega \) regime. This yields \( \lambda = 0.26 \), in accordance with the expectation of one to two functional nucleosome-depleted regions (NDRs) per intergenic region. Our inference of selection, however, is robust over a broad range of possible \( \lambda \) values. As shown in Fig. S3A, the inferred scaled fitness landscape

\[
2NF(\omega) = \log \left( \frac{Q(\omega)}{P_0(\omega)} \right) + \text{const.}
\]

is nearly independent of \( \lambda \) in the regime \( 0.2 < \lambda \leq 1 \) and \( \omega < 0.5 \).

2. The genomic distributions \( W(\omega) \) and \( P_0(\omega) \) are obtained using a tiling procedure with a fixed length \( \ell = 100 \) bp, which is an approximate lower bound for extended linker regions (3). As shown in Fig. S2B, the inferred scaled fitness landscape \( 2N\omega(\omega) \) is insensitive to changes of \( \ell \) across the length range of NDRs in yeast (4). As expected, the selection signal is weaker for inference lengths \( \ell \) significantly below 100 bp, because it is confounded by nonfunctional short linker regions.

DNA Sequence Analysis. As null model for the selection on \( \omega \), we use random sequence with single-nucleotide frequencies corresponding to the average genome-wide *S. cerevisiae* frequencies. In particular, nucleotide triplets conferring specific local elasticity properties are scrambled in the null model. This takes into account the selection for elevated A:T (histone-repelling sequence containing homopolymeric adenine segments on one strand paired with thymine segments on the other strand) content as well as for specific nucleosome-averse sequence configurations. Using average nucleotide frequencies from intergenic sequences does not, however, change our conclusions.

To exclude alignment uncertainties in the cross-species comparison, insertions and deletions below a fraction of 2% in *Saccharomyces paradoxus* are removed with respect to the reference species *S. cerevisiae*. This means that insertions in *S. paradoxus* are cut out, whereas deletions are filled with the corresponding nucleotides of *S. cerevisiae*. The upper threshold of 2% total amount of allowed insertions and deletions is low enough to ensure that it does not have an effect on the statistical analysis carried out. Fig. S5B shows the distribution of mean occupancies on the original *S. paradoxus* sequences, and Fig. S5A shows the same distributions as Fig. 4B of the main text, but limited to NDRs that do not contain any insertions or deletions.

Analysis of in Vivo Nucleosome Data. Experimental in vivo nucleosome occupancy scores (4) are processed to reduce the effects of measurement uncertainties. We remove extended count voids (<10 counts >1,000 bp) and regions with extremely high counts (1,000-bp average greater than twice the average occupancy score). To normalize the scores, we divide by the average occupancy score.

In Silico Evolution of NDRs. In the Wright–Fisher model of evolution under mutation-selection-drift dynamics, we use the NDR sequences plus 150 bp of flanking region on both sides. Each original *S. cerevisiae* NDR sequence then is evolved as a population of \( N = 50 \) individuals with a scaled neutral mutation rate \( 2N\mu_0 = \mu = 0.02 \). Neutral evolution is simulated using

Weghorn and Lässig www.pnas.org/cgi/content/short/1210887110
the same Wright–Fisher simulation, but without selection. The *S. cerevisiae* genome is divided into pieces 500 bp long, each of which is evolved as a population of \( N = 100 \) individuals (\( \mu = 0.02 \)). In both cases, the temporal separation between the initial and the simulated sequences is set to achieve \( \sim 13\% \) average sequence divergence between the two species, corresponding to the observed real value in our set of NDR segments.

Because we use a linear fit to the fitness landscape \( F(\omega) \), the Wright–Fisher simulation of evolution under mutation-selection-drift dynamics does not depend on the initial value of histone binding affinity \( \omega \), but only on phenotypic differences. Therefore, the evolution of each *S. cerevisiae* NDR sequence (with phenotype \( \omega_{\text{par}} \)) and its flanking regions can be modeled independently of its genomic context. We thus obtain evolved genotypes, which are inserted into the *S. paradoxus* genomic background to obtain the corresponding aligned phenotypes \( \omega_{\text{par}} \).

The long-range correlations inherent in the biophysical modeling of nucleosome density along the genome are negligible in this context, which is shown in Fig. S6. The linear approximation for the fitness landscape leads to a slight overestimation of the evolutionary constraint for larger phenotype values (Fig. 4B).

**NDR Repositioning.** Comparison of aligned sequences, as in Fig. 4B, does not account for possible relocations of low-affinity sequences. In such cases, an observed increase in \( \omega \) on an NDR might disappear when it is compared with the shifted, “orthologous” NDR in the other species. Such processes are not captured by our fitness landscape or the distribution of functional sequences. To find relocations, we investigated the local environment of NDRs: Instead of considering \( \omega \) on the aligned sequence in *S. paradoxus*, we first compared it with the histone binding affinity on NDRs that overlap; i.e., we allowed for a wobble on the length scale of 100 bp. We found that all qualitative features of the cross-species divergence (Fig. 4B) stay the same. Considering NDRs that do not overlap as functionally separate, we next asked whether the relocation of NDRs with \( \omega < 0.4 \) is random or shows any additional spatial correlations. To this end, we looked at the distribution of distances between segments with \( \omega < 0.4 \) in *S. paradoxus* and the nearest such segment in *S. cerevisiae* within a symmetric window of width 2 kbp. We found that positional correlations beyond overlaps are comparatively rare (\( \approx 5\% \)) and there is no deviation from a uniform distribution. We conclude that the NDRs of our set evolved approximately independently, which validates our inference of selection.


---

**Fig. S1.** Dependence of histone binding affinity \( \omega \) on histone binding energy. We plot the average histone binding affinity phenotype \( \omega \) as a function of mean histone binding energy, \( \Delta G = \frac{1}{4} \sum_{i=1}^{4} \Delta G(r_i) \), given by the biophysical model described in Methods. Free energies and affinities are evaluated for tiled *S. cerevisiae* intergenic segments of length \( \ell = 100 \) bp. Different colors represent different total genomic nucleosome coverage, corresponding to different values of the chemical potential \( \eta \): yellow, 30\% (\( \eta = 77 k_B T \)); purple, 61\% (\( \eta = 80 k_B T \)); and blue, 80\% (\( \eta = 84 k_B T \)). The in vivo value used for the analysis of the main text. As expected, histone binding affinities correlate negatively with the associated mean histone binding energy and positively with the total genomic nucleosome coverage.
Fig. S2. Joint phenotype distributions. (A) Normalized genomic counts $W(\omega, n)$, evaluated in tiled *S. cerevisiae* intergenic segments of length $\ell = 100$ bp (Methods). (B) Null distribution $P_0(\omega, n)$, evaluated as described in SI Text. Note the logarithmic scale on the z-axis.

Fig. S3. Robustness of selection inference. (A) Selection on histone binding can be inferred using a mixture model $W(\omega) = \lambda Q(\omega) + (1 - \lambda)P_0(\omega)$, assuming different fractions $\lambda$ of segments under selection. The resulting scaled fitness landscape, $2NF(\omega) = \log[Q(\omega)/P_0(\omega)] + \text{const}$, is shown for $\lambda = 1$ (red, as in main text), $\lambda = 0.5$ (dotted gray), and $\lambda = 0.25$ (solid gray). The (arbitrary) additive normalization is chosen so that the landscapes collapse in the low-$\omega$ regime. The dependence on $\lambda$ is weak throughout the regime relevant to our analysis, $\omega < 0.5$. (B) Selection on histone binding can be inferred from affinity distributions obtained with different values of the tiling length $\ell$. The scaled fitness landscape $2NF(\omega) = \log[W(\omega)/P_0(\omega)] + \text{const}$ is shown for $\ell = 50$ and 70 bp (dashed gray, bottom to top), $\ell = 90, 110, 130$ bp (solid gray, bottom to top), and $\ell = 100$ bp (red, same as Fig. 2). The inference of selection is nearly independent of $\ell$ across the range of typical NDR sizes (solid gray). For values of $\ell$ significantly below 100 bp, the selection signal is confounded by regular, nonfunctional linker regions (dashed gray).
Controlling for *S. paradoxus* population substructure in the inference of selection on single-nucleotide polymorphisms (SNPs). The data points show the frequency of the high-affinity allele, $x_+$, as a function of the phenotypic effect (i.e., the difference $\Delta \omega$ between both alleles) for SNPs in intergenic *S. paradoxus* NDRs with $\omega < 0.4$ (green dots, with size indicating the number of SNPs contributing to the data point). These data were obtained from splitting the *S. paradoxus* population sample into three major subpopulations (European, Far Eastern, and American) (6). We evaluated the effect-dependent average frequency $\langle x_+ \rangle$ in $\Delta \omega$-bins of size 0.025 (green dots with error bars, joined by solid green line); dashed green line, linear least-squares fit, yielding $\langle x_+ \rangle(\Delta \omega) = 1/2 - (0.6 \pm 0.2) \Delta \omega$. The small-\(\sigma\) prediction from theory, averaged over the three subpopulations, is given by $\langle x_+ \rangle(\sigma) = 1/2 + 0.076 \sigma (\mu = 0.02)$, giving for the prediction of the average allele frequency from the fitness landscape $\langle x_+ \rangle(\Delta \omega) = 1/2 - (0.8 \pm 0.1) \Delta \omega$ (red line), again in good agreement with the data. The expectation in a neutral scenario is a constant $\langle x_+ \rangle(\Delta \omega) = 1/2$ (blue line) and is inconsistent with the real data.
Comparison with \textit{S. paradoxus} data without correction of insertions and deletions relative to \textit{S. cerevisiae}. (A) Distribution $W(\omega)$ of the histone binding affinity $\omega$ on nonoverlapping intergenic segments of length $l = 100$ bp in original \textit{S. paradoxus}, with no insertions and deletions altered (green●) and in \textit{S. cerevisiae} (purple●, same as Figs. 2 and 4A), compared with the analogous distribution from random sequence, $P_0(\omega)$ (solid black line, same as Figs. 2 and 4A). (B) Cross-species distribution of affinity pairs $(\omega_{\text{cer}}, \omega_{\text{par}})$ for NDRs in \textit{S. cerevisiae} and their aligned sequences in \textit{S. paradoxus} (gray contour areas). The conditional average of $\omega_{\text{par}}$ as a function of $\omega_{\text{cer}}$ (green line) is compared with simulated evolution under neutrality (blue line). Standard deviations are given by error bars. Here we use only NDRs with no sequence insertions or deletions between \textit{S. cerevisiae} and \textit{S. paradoxus}, leaving 143 data points out of 1,521. This shows that the results of the cross-species comparison reported in Fig. 4 are robust to changes in our alignment procedure (SI Text).
Effect of genomic background on histone binding affinity. Here we tested the influence of the long-range correlations in nucleosome occupancy on our evolution model. Dashed blue line: NDR sequences obtained from the neutral *in silico* evolution were inserted into the same genomic background of *S. paradoxus* used for the insertion of the NDRs evolved under selection in Fig. 4B. Solid blue line: Neutrally evolved NDRs in the neutrally evolved background. The similarity between these results shows that the influence of the genomic background is negligible, and the reinsertion of NDRs into the *S. paradoxus* background gives an essentially unbiased result for *in silico* evolution of NDRs under selection. Green line: Bin average of $\omega_{\text{par}}$ as a function of $\omega_{\text{cer}}$ of the cross-species distribution for NDRs in *S. cerevisiae* and their aligned sequences in *S. paradoxus* (same as Fig. 4B). Standard deviations are given by error bars; for illustration purposes, the dashed blue and green error bars are shifted slightly relative to the solid blue ones.