Fitness landscape for nucleosome positioning

Donate Weghorn and Michael Lässig[1]

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.

biophysics | nucleosome-depleted regions | evolution of regulation | quantitative traits | inference of 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.

[1]To 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.

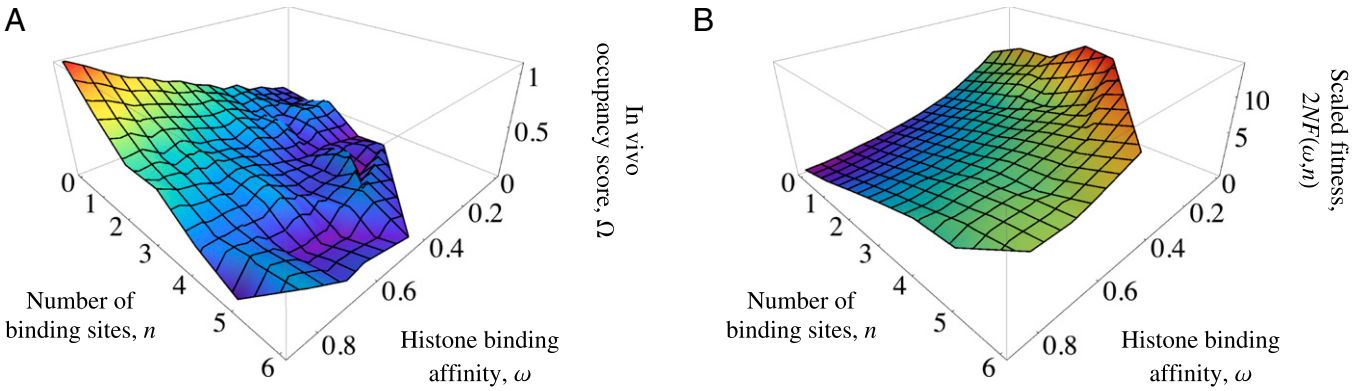

Fig. 1. In vivo nucleosome occupancy and fitness for yeast intergenic sequence segments. (*A*) The mean nucleosome occupancy, $\Omega(\omega, n)$, is plotted against two molecular phenotypes: the elasticity-mediated histone binding affinity, ω, and the number of transcription factor binding sites, *n*. Occupancy data in *S. cerevisiae* are taken from (3) and shown for nonoverlapping intergenic sequence segments of length 100 bp. Data points not shown reflect insufficient phenotype counts (ω, n). (*B*) The scaled fitness landscape $2NF(\omega, n)$ inferred from the genomic phenotype distribution (by Eq. **1**). This landscape shows that direct selection acts on both phenotypes and establishes sequence-dictated nucleosome positioning as a primary mode of the evolution of intergenic DNA.

inference method can be applied to other quantitative traits with 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 mean nucleosome occupancy ω 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 ω decreases 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 ω 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, *n*, 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. 1*A*, 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, $W(\omega, n)$, with the corresponding distribution $P_0(\omega, n)$ evaluated in a suitable null model. To obtain $W(\omega, n)$, we construct a tiling of the yeast genome into nonoverlapping segments of fixed length $\ell = 100$ 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 $W(\omega, n)$ for intergenic sequence in *S. cerevisiae* is shown in Fig. S2*A*. 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 phenotypic null distribution may be approximated as a product, $P_0(\omega, n) = P_0(\omega)P_0(n)$. We obtain the marginal distribution $P_0(\omega)$ 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 $P_0(\omega, n)$ is shown in Fig. S2*B*.

We now can infer the scaled phenotype-fitness map $2NF(\omega, n)$ as the log-likelihood score of the genomic phenotype distribution and the null distribution (29, 30):

$$2NF(\omega, n) = \log\left[\frac{W(\omega, n)}{P_0(\omega, n)}\right] + \text{const.} \qquad [1]$$

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 $F(\omega, n)$ 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, $\omega < 0.5$ (*SI Text* and Fig. S3*A*). Similarly, our results are insensitive to variations of the tiling length ℓ within the length range of functional NDRs, as shown in Fig. S3*B*.

The scaled fitness landscape $2NF(\omega, n)$ inferred for *S. cerevisiae* intergenic sequence is shown in Fig. 1*B*. It reveals substantial selection on both histone binding affinity and transcriptional regulation: We find scaled fitness differences $|2N\Delta F| \leq 10$ in our set of intergenic segments. Importantly, the selection on histone binding affinity is a primary effect; that is, the overrepresentation of NDRs

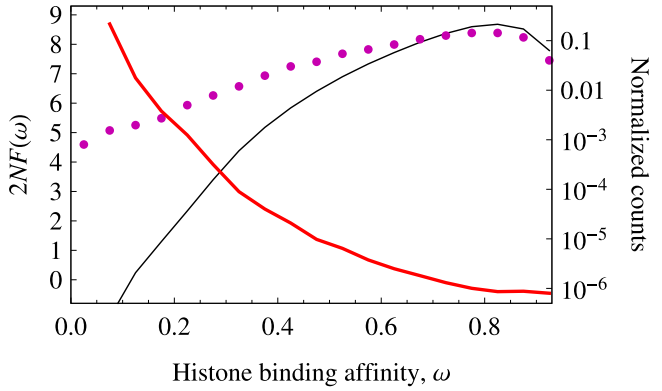

Fig. 2. Selection against nucleosome formation. Distribution of histone binding affinity for nonoverlapping intergenic segments of length 100 bp in *S. cerevisiae*, $W(\omega)$ (purple ●), compared with the analogous distribution from random sequence $P_0(\omega)$ (solid black line). Both distributions are evaluated in bins of width 0.05. The effective scaled fitness landscape for histone binding affinity, $2NF(\omega)$ (red line), is the log-likelihood of the distributions $W(\omega)$ and $P_0(\omega)$.

in the yeast genome cannot be explained by direct selection on regulatory site content alone. Our finding of substantial direct selection on ω 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-dictated nucleosome positioning within and across species.

Selection Against Nucleosome Formation. As shown in Fig. 1*B*, 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, $F(\omega)$, which is most convenient for our subsequent evolutionary analysis. This landscape is inferred from the marginal distributions $W(\omega)$ and $P_0(\omega)$ by an equilibrium relation analogous to Eq. **1**, and is shown in Fig. 2. Again, the function $F(\omega)$ is insensitive to the fraction of segments under selection and to the choice of tiling length (*SI Text* and Fig. S3).

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

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

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 $|\Delta\omega| \gtrsim 0.1$ 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 Resequencing 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 $\Delta\omega = \overline{\omega}_+ - \overline{\omega}_-$, where $\overline{\omega}_+$ denotes the larger and $\overline{\omega}_-$ 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, $\langle x_+ \rangle$, with increasing deleterious effect. Fig. 3 shows the data points $(\Delta\omega, x_+)$ and the resulting average frequencies in bins of the affinity difference. These data permit a linear fit of $\langle x_+ \rangle$ as a function of $\Delta\omega$,

$$\langle x_+ \rangle(\Delta\omega) = \frac{1}{2} - \gamma\Delta\omega, \qquad [3]$$

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 $p_{\mathrm{eq}}(x;\sigma)$ then determines the expected frequency of the deleterious allele, $\langle x_+ \rangle$ (32) (*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, $\langle x_+ \rangle(\sigma)$, 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

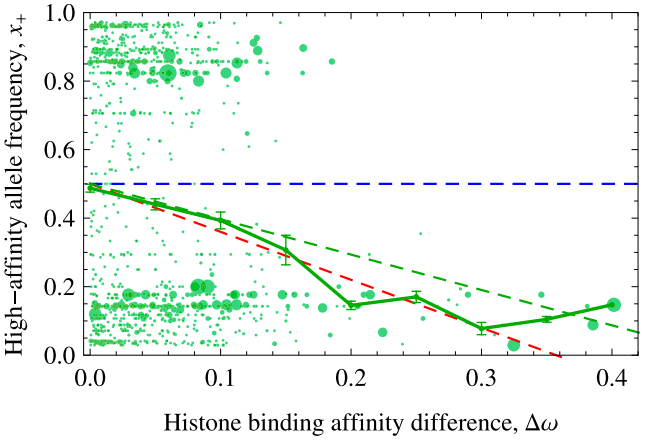

Fig. 3. Selection on SNPs. The data points show the frequency of the high-affinity allele, x_+, as a function of the phenotypic effect (i.e., the difference $\Delta\omega$ between both alleles) for SNPs in intergenic *S. paradoxus* NDRs with $\omega < 0.4$ (green dots, with size indicating the number of SNPs contributing to the data point). From these data, we evaluated the effect-dependent average frequency $\langle x_+ \rangle$ (in $\Delta\omega$-bins of size 0.05; green dots with error bars, joined by solid green line). Its approximately linear decrease follows Eq. **3** (least-squares fit, dashed green line) and shows that there is weak selection against alleles of higher affinity. The prediction from the fitness landscape $F(\omega)$ (dashed red line; see text) is in good agreement with the data. The expectation under neutrality is a constant, $\langle x_+ \rangle(\Delta\omega) = 1/2$ (dashed blue line), and is inconsistent with the data.

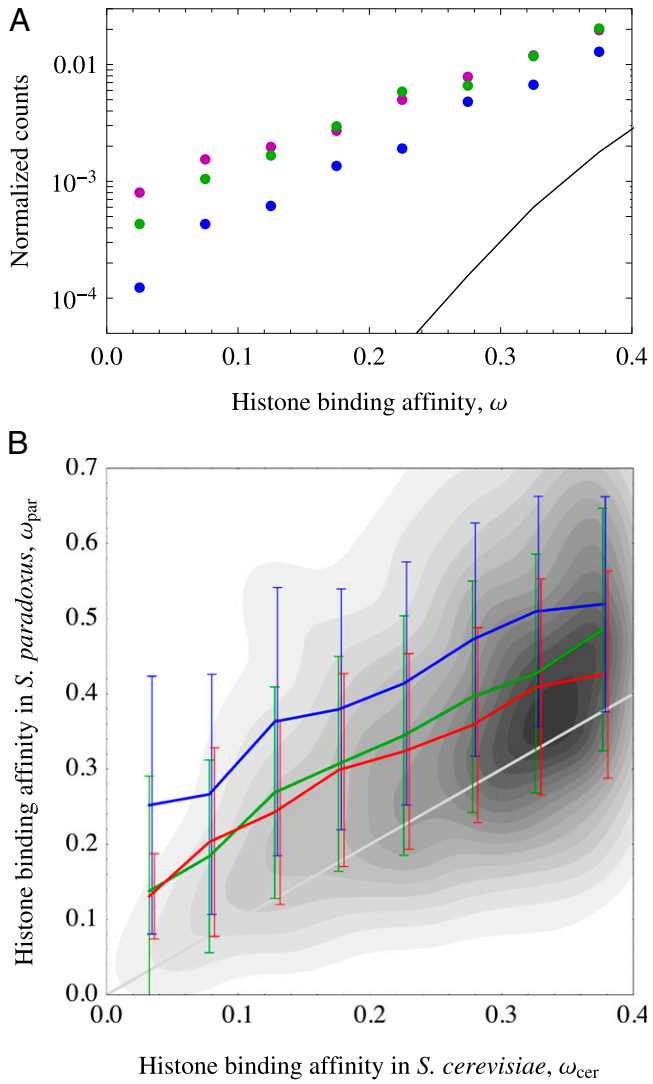

Fig. 4. Cross-species evolution of histone binding affinity. (*A*) Distribution of histone binding affinity, $W(\omega)$, for intergenic segments of length 100 bp with $\omega < 0.4$ in *S. paradoxus* (green ●) and in *S. cerevisiae* (purple ●, same as Fig. 2). These distributions are very similar, which is consistent with evolutionary equilibrium under selection given by the fitness landscape $F(\omega)$. In contrast, simulated neutral evolution (blue ●) already leads to a significant reduction of low-affinity counts over the same evolutionary distance, and would approach the neutral equilibrium distribution $P_0(\omega)$ (black line, same as Fig. 2) in the long-time limit. (*B*) Cross-species distribution of affinity pairs $(\omega_{cer}, \omega_{par})$ for NDRs in *S. cerevisiae* and their aligned sequences in *S. paradoxus* (gray contour areas). The conditional average (green line) and standard deviation (green bars) of ω_{par} is plotted as a function of ω_{cer}. We compare these data with the conditional distributions $P(\omega_{par}|\omega_{cer})$ for simulated evolution in the fitness landscape $F(\omega)$ and under neutrality (average, red and blue lines; standard deviation, red and blue bars). The cross-species data are consistent with evolution under directional selection against nucleosome formation. At the same time, the near-neutral standard deviation shows the variability of cross-species affinity evolution under this fitness model.

for cross-species evolution: The phenotype distribution $W(\omega)$ and, hence, the number of NDRs below a given affinity threshold are conserved. Fig. 4*A* compares the genomic distributions $W(\omega)$ for *S. cerevisiae* and *S. paradoxus* intergenic regions. These distributions indeed are strikingly similar between the two species. We can compare this conservation with simulated neutral evolution of an ensemble of sequence segments with the *S. cerevisiae* distribution $W(\omega)$ as the initial condition (*Methods* and *SI Text*). Already over the distance between *S. cerevisiae* and *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 $1,500 \pm 40$ non-overlapping intergenic NDRs with length 100 bp and $\omega < 0.4$ in the actual *S. cerevisiae* and *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 *S. cerevisiae* and *S. paradoxus*, and to 170 at neutral equilibrium. Similar results are obtained in a three-species comparison of *S. cerevisiae*, *S. paradoxus*, and *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 binding affinity, we mapped the set of intergenic NDR segments with $\omega < 0.4$ in *S. cerevisiae* onto their aligned segments in *S. paradoxus* (*Methods* and *SI Text*). Fig. 4*B* shows the contour lines and binned averages of the resulting scatter plot $(\omega_{cer}, \omega_{par})$. These pairs have lower mean affinity values in *S. cerevisiae* compared with *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 *S. paradoxus* NDRs).

We can compare the actual process with *in silico* evolution under selection, using a Wright–Fisher simulation of the *S. cerevisiae* NDR sequences in the fitness landscape $F(\omega)$ (for details, see *Methods* and *SI Text*). Fig 4*B* shows the binned average and standard deviation of the resulting conditional distribution $P(\omega_{par}|\omega_{cer})$ for cross-species phenotype evolution. We find both quantities to be in quantitative agreement with the observed divergence statistics between *S. cerevisiae* and *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 *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. 1*B*). 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. 1*B*, 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—at least in the densely packed genomes of prokaryotes and unicellular eukaryotes. We have shown that a joint phenotype-fitness map can disentangle selective effects on such functions, i.e., distinguish direct from apparent selection. We expect this method to be applicable to 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 5′ genomic coordinate r of a nucleosome. The free energy of a DNA sequence segment $(a_r, a_{r+1}, \ldots, a_{r+d-1})$ is given by

$$\Delta G(r) = \sum_{r'=r}^{r+d-3} \sum_{i=1}^{3} \frac{A_i}{2} \left[\phi_i^{\mathrm{nuc}}(\mathbf{a}_{r'}) - \phi_i^0(\mathbf{a}_{r'})\right]^2,$$

where $\mathbf{a}_{r'} = (a_{r'}, a_{r'+1}, a_{r'+2})$ denote trinucleotide subsegments; $\phi_i^{\mathrm{nuc}}(\mathbf{a}_{r'})$ $(i = 1, 2, 3)$ are the roll, twist, and tilt deformations in the nucleosome state, $\phi_i^0(\mathbf{a}_{r'})$ are the intrinsic deformations in the unbound state (35), A_i 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

$$Z = \sum_{N=0}^{\infty} \sum_{r_1,\ldots,r_N} \exp\left[-\beta\left(-\eta N + \sum_{k=1}^{N} \Delta G(r_k)\right)\right]$$

with the no-overlap constraint $r_{k+1} \geq r_k + d$ $(k = 1, \ldots, N-1)$. The partition function depends on the temperature via $k_B T = \beta^{-1}$ and on the chemical potential η, which are adjusted to in vivo conditions. This determines the expected single-nucleotide nucleosome occupancies (36),

$$\mathcal{O}(r) = -\beta^{-1} \sum_{r'=r-d+1}^{r} \frac{\partial \log Z}{\partial \Delta G(r')},$$

and the expected mean occupancy

$$\omega(r, \ell) = \frac{1}{\ell} \sum_{r'=r}^{r+\ell-1} \mathcal{O}(r')$$

over sequence segments of length ℓ. The dependence of ω 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 substructure (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 $\ell = 100$ bp, using a dynamic programming algorithm with an upper bound of 0.95 of the predicted mean nucleosome occupancy ω 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, $p_{\mathrm{eq}}(x; \sigma) = (x(1-x))^{\mu-1} e^{\sigma x}/Z_{\mathrm{eq}}$, where $\sigma = 2N\Delta F$ is the scaled selection coefficient, $\mu = 2N\mu_0$ is the scaled neutral mutation rate, and Z_{eq} is a normalization factor. From this distribution, we determine the allele frequency spectrum for polymorphic loci, $p_{\mathrm{eq}}(k; m, \sigma)$, in a set of m isolates by binomial sampling $(k = 1, \ldots, m-1)$. This distribution produces an average frequency of the deleterious allele, $\langle x \rangle(\sigma) = 1/2 + a\sigma + O(\sigma^2)$, 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 $F(\omega)$. 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. S6.

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).

1. Lee W, et al. (2007) A high-resolution atlas of nucleosome occupancy in yeast. *Nat Genet* 39(10):1235–1244.
2. Bai L, Morozov AV (2010) Gene regulation by nucleosome positioning. *Trends Genet* 26(11):476–483.
3. Tsankov AM, Thompson DA, Socha A, Regev A, Rando OJ (2010) The role of nucleosome positioning in the evolution of gene regulation. *PLoS Biol* 8(7):e1000414.
4. Luger K, Mäder AW, Richmond RK, Sargent DF, Richmond TJ (1997) Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature* 389(6648):251–260.

5. Field Y, et al. (2009) Gene expression divergence in yeast is coupled to evolution of DNA-encoded nucleosome organization. *Nat Genet* 41(4):438–445.
6. Shivaswamy S, et al. (2008) Dynamic remodeling of individual nucleosomes across a eukaryotic genome in response to transcriptional perturbation. *PLoS Biol* 6(3):e65.
7. Jiang C, Pugh BF (2009) Nucleosome positioning and gene regulation: Advances through genomics. *Nat Rev Genet* 10(3):161–172.
8. Radman-Livaja M, Rando OJ (2010) Nucleosome positioning: How is it established, and why does it matter? *Dev Biol* 339(2):258–266.
9. Swamy KBS, Chu W-Y, Wang C-Y, Tsai H-K, Wang D (2011) Evidence of association between nucleosome occupancy and the evolution of transcription factor binding sites in yeast. *BMC Evol Biol* 11:150.
10. Segal E, et al. (2006) A genomic code for nucleosome positioning. *Nature* 442(7104): 772–778.
11. Thåström A, et al. (1999) Sequence motifs and free energies of selected natural and non-natural nucleosome positioning DNA sequences. *J Mol Biol* 288(2):213–229.
12. Widom J (2001) Role of DNA sequence in nucleosome stability and dynamics. *Q Rev Biophys* 34(3):269–324.
13. Segal E, Widom J (2009) Poly(dA:dT) tracts: Major determinants of nucleosome organization. *Curr Opin Struct Biol* 19(1):65–71.
14. Schones DE, et al. (2008) Dynamic regulation of nucleosome positioning in the human genome. *Cell* 132(5):887–898.
15. Yuan G-C, et al. (2005) Genome-scale identification of nucleosome positions in S. cerevisiae. *Science* 309(5734):626–630.
16. Cairns BR (2009) The logic of chromatin architecture and remodelling at promoters. *Nature* 461(7261):193–198.
17. Whitehouse I, Rando OJ, Delrow J, Tsukiyama T (2007) Chromatin remodelling at promoters suppresses antisense transcription. *Nature* 450(7172):1031–1035.
18. Kornberg RD, Stryer L (1988) Statistical distributions of nucleosomes: Nonrandom locations by a stochastic mechanism. *Nucleic Acids Res* 16(14A):6677–6690.
19. Mavrich TN, et al. (2008) A barrier nucleosome model for statistical positioning of nucleosomes throughout the yeast genome. *Genome Res* 18(7):1073–1083.
20. Milani P, et al. (2009) Nucleosome positioning by genomic excluding-energy barriers. *Proc Natl Acad Sci USA* 106(52):22257–22262.
21. Möbius W, Gerland U (2010) Quantitative test of the barrier nucleosome model for statistical positioning of nucleosomes up- and downstream of transcription start sites. *PLoS Comp Biol* 6(8):e1000891.
22. van Nimwegen E (2007) Finding regulatory elements and regulatory motifs: A general probabilistic framework. *BMC Bioinformatics* 8(Suppl 6):S4.
23. Tirosh I, Sigal N, Barkai N (2010) Divergence of nucleosome positioning between two closely related yeast species: Genetic basis and functional consequences. *Mol Syst Biol* 6:365.
24. Warnecke T, Batada NN, Hurst LD (2008) The impact of the nucleosome code on protein-coding sequence evolution in yeast. *PLoS Genet* 4(11):e1000250.
25. Washietl S, Machné R, Goldman N (2008) Evolutionary footprints of nucleosome positions in yeast. *Trends Genet* 24(12):583–587.
26. Kenigsberg E, Bar A, Segal E, Tanay A (2010) Widespread compensatory evolution conserves DNA-encoded nucleosome organization in yeast. *PLOS Comput Biol* 6(12): e1001039.
27. Prendergast JG, Semple CA (2011) Widespread signatures of recent selection linked to nucleosome positioning in the human lineage. *Genome Res* 21(11):1777–1787.
28. Vaillant C, Audit B, Arneodo A (2007) Experiments confirm the influence of genome long-range correlations on nucleosome positioning. *Phys Rev Lett* 99(21):218103.
29. Berg J, Willmann S, Lässig M (2004) Adaptive evolution of transcription factor binding sites. *BMC Evol Biol* 4:42.
30. Mustonen V, Kinney J, Callan CG, Jr., Lässig M (2008) Energy-dependent fitness: A quantitative model for the evolution of yeast transcription factor binding sites. *Proc Natl Acad Sci USA* 105(34):12376–12381.
31. Liti G, et al. (2009) Population genomics of domestic and wild yeasts. *Nature* 458 (7236):337–341.
32. Wright S (1937) The distribution of gene frequencies in populations. *Proc Natl Acad Sci USA* 23(6):307–320.
33. Thurman RE, et al. (2012) The accessible chromatin landscape of the human genome. *Nature* 489(7414):75–82.
34. Stadler M, et al. (2011) DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature* 480(7378):490–495.
35. Goodsell DS, Dickerson RE (1994) Bending and curvature calculations in B-DNA. *Nucleic Acids Res* 22(24):5497–5503.
36. Percus JK (1976) Equilibrium state of a classical fluid of hard rods in an external field. *J Stat Phys* 15(6):505–511.
37. 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(6937): 241–254.