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 ω 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

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

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

$$P_0 \sim \exp\left(-\sum_{\alpha=1}^{n} D_\alpha\right)$$

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

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

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 λ 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 λ from the data by minimization of χ^2 between $(1 - \lambda)P_0(\omega)$ and $W(\omega)$ in the high-ω 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 λ values. As shown in Fig. S3*A*, the inferred scaled fitness landscape

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

is nearly independent of λ 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. S3*B*, the inferred scaled fitness landscape $2NF(\omega)$ is insensitive to changes of ℓ across the length range of NDRs in yeast (4). As expected, the selection signal is weaker for inference lengths ℓ significantly below 100 bp, because it is confounded by nonfunctional short linker regions.

DNA Sequence Analysis. As null model for the selection on ω, 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. S5*A* shows the distribution of mean occupancies on the original *S. paradoxus* sequences, and Fig. S5*B* shows the same distributions as Fig. 4*B* 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 $2N\mu_0 \equiv \mu = 0.02$ (5, 6). Neutral evolution is simulated using

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 ~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 ω, but only on phenotypic differences. Therefore, the evolution of each *S. cerevisiae* NDR sequence (with phenotype ω_{cer}) 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 ω_{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. 4*B*).

NDR Repositioning. Comparison of aligned sequences, as in Fig. 4*B*, does not account for possible relocations of low-affinity sequences. In such cases, an observed increase in ω 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 ω 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. 4*B*) 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.

1. Cover TM, Thomas JA (2006) *Elements of Information Theory*, Wiley Series in Telecommunications and Signal Processing (Wiley Interscience, New York).
2. van Nimwegen E (2007) Finding regulatory elements and regulatory motifs: A general probabilistic framework. *BMC Bioinformatics* 8(Suppl 6):S4.
3. Jiang C, Pugh BF (2009) Nucleosome positioning and gene regulation: Advances through genomics. *Nat Rev Genet* 10(3):161–172.
4. 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.
5. Tsai IJ, Bensasson D, Burt A, Koufopanou V (2008) Population genomics of the wild yeast Saccharomyces paradoxus: Quantifying the life cycle. *Proc Natl Acad Sci USA* 105(12):4957–4962.
6. Liti G, et al. (2009) Population genomics of domestic and wild yeasts. *Nature* 458(7236): 337–341.

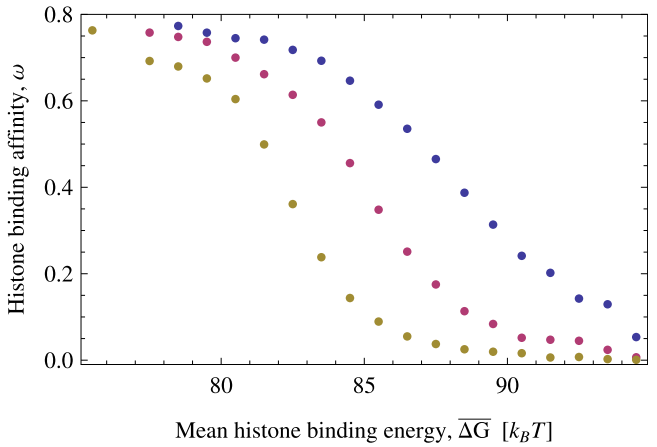

Fig. S1. Dependence of histone binding affinity ω on histone binding energy. We plot the average histone binding affinity phenotype ω as a function of mean histone binding energy, $\overline{\Delta G} = \frac{1}{\ell}\sum_{r'=r}^{r+\ell-1} \Delta G(r')$, given by the biophysical model described in *Methods*. 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 η: yellow, 30% ($\eta = 77\ k_BT$); purple, 61% ($\eta = 80\ k_BT$); and blue, 80% ($\eta = 84\ k_BT$), 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.

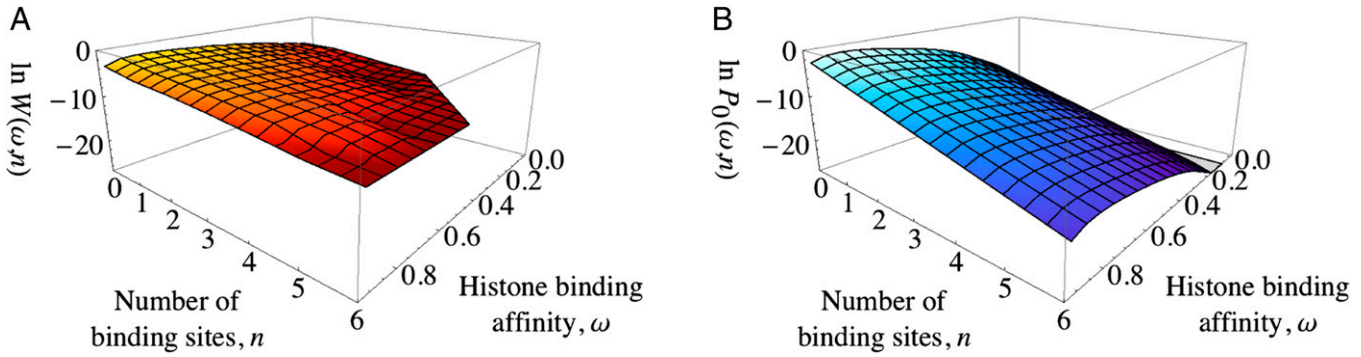

Fig. S2. Joint phenotype distributions. (*A*) Normalized genomic counts $W(\omega, n)$, evaluated in tiled *S. cerevisiae* intergenic segments of length $\ell = 100$ bp (*Methods*). (*B*) Null distribution $P_0(\omega, n)$, evaluated as described in *SI Text*. Note the logarithmic scale on the *z*-axis.

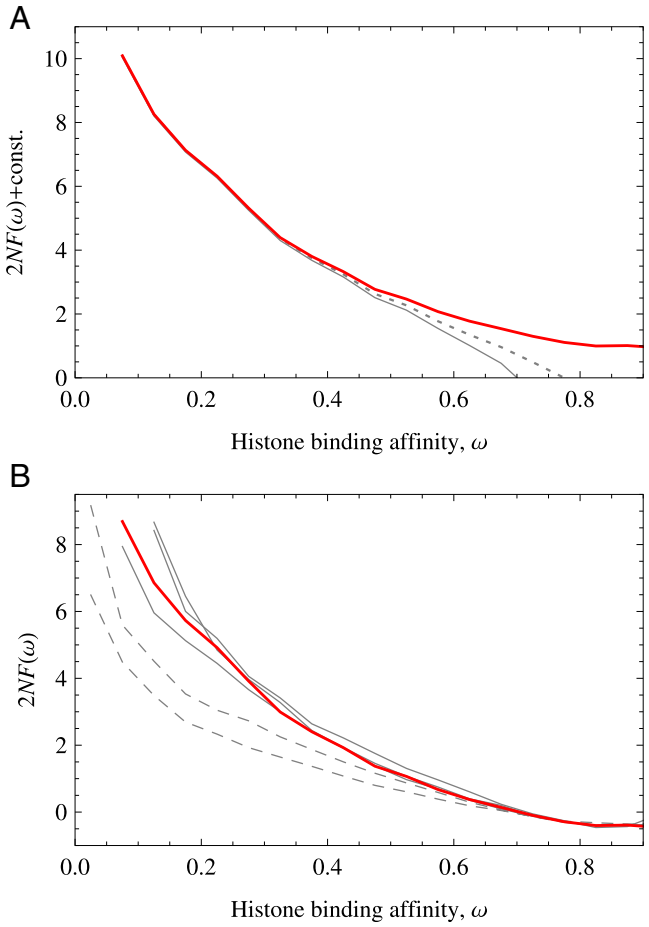

Fig. S3. Robustness of selection inference. (*A*) Selection on histone binding can be inferred using a mixture model $W(\omega) = \lambda Q(\omega) + (1-\lambda)P_0(\omega)$, assuming different fractions λ of segments under selection. The resulting scaled fitness landscape, $2NF(\omega) = \log[Q(\omega)/P_0(\omega)] + \text{const.}$, is shown for $\lambda = 1$ (red, as in main text), $\lambda = 0.5$ (dotted gray), and $\lambda = 0.26$ (solid gray). The (arbitrary) additive normalization is chosen so that the landscapes collapse in the low-ω regime. The dependence on λ 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 ℓ. 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 ℓ across the range of typical NDR sizes (solid gray). For values of ℓ significantly below 100 bp, the selection signal is confounded by regular, nonfunctional linker regions (dashed gray).

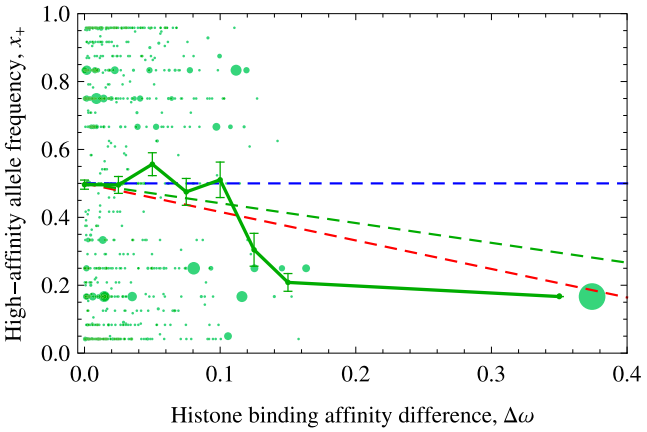

Fig. S4. Controlling for *S. paradoxus* population substructure in the inference of selection on single-nucleotide polymorphisms (SNPs). The data points show the frequency of the high-affinity allele, x_+, as a function of the phenotypic effect (i.e., the difference $\Delta\omega$ between both alleles) 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 $\langle x_+ \rangle_F(\Delta\omega) = 1/2 - (0.8 \pm 0.1)\Delta\omega$ (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.

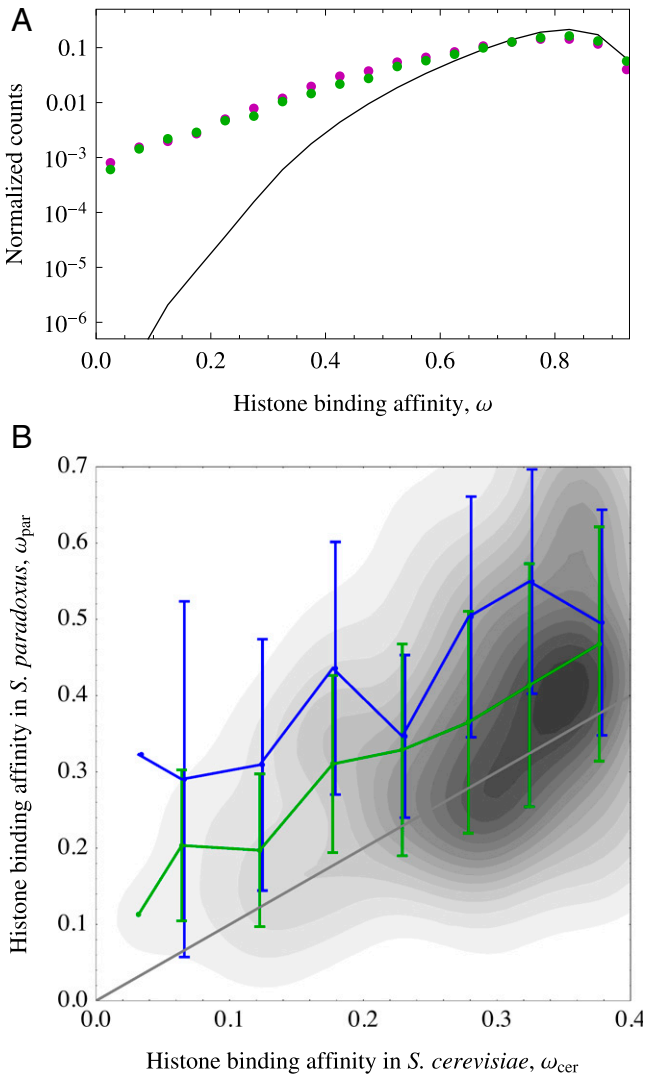

Fig. S5. Comparison with *S. paradoxus* data without correction of insertions and deletions relative to *S. cerevisiae*. (*A*) Distribution $W(\omega)$ of the histone binding affinity ω on nonoverlapping intergenic segments of length $\ell = 100$ bp in original *S. paradoxus*, with no insertions and deletions altered (green ●) and in *S. cerevisiae* (purple ●, same as Figs. 2 and 4*A*), compared with the analogous distribution from random sequence, $P_0(\omega)$ (solid black line, same as Figs. 2 and 4*A*). (*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 of ω_{par} as a function of ω_{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 *S. cerevisiae* and *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*).

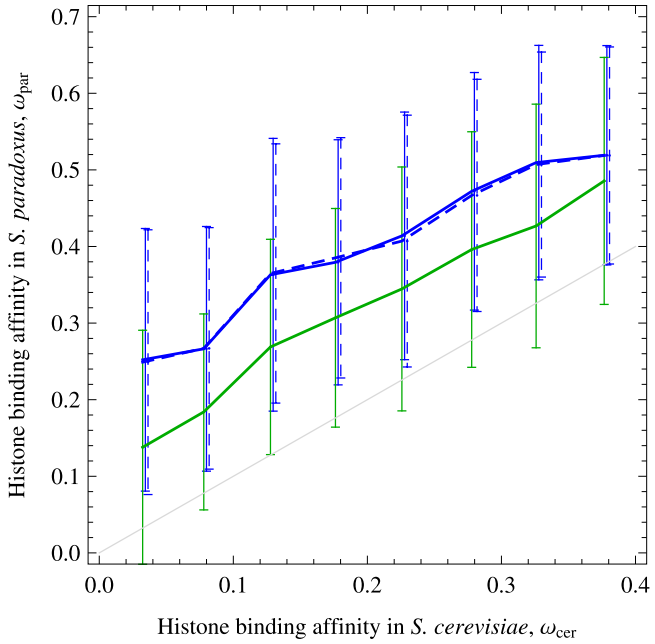

Fig. S6. Effect of genomic background on histone binding affinity. Here we tested the influence of the long-range correlations in nucleosome occupancy on our evolution model. Dashed blue line: NDR sequences obtained from the neutral *in silico* evolution were inserted into the same genomic background of *S. paradoxus* used for the insertion of the NDRs evolved under selection in Fig. 4*B*. 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 ω_{par} as a function of ω_{cer} of the cross-species distribution for NDRs in *S. cerevisiae* and their aligned sequences in *S. paradoxus* (same as Fig. 4*B*). 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.