Genome-Wide Motif Statistics are Shaped by DNA Binding Proteins over Evolutionary Time Scales

Long Qian¹ and Edo Kussell^{1,2}

¹*Department of Biology and Center for Genomics and Systems Biology, New York University, New York, New York 10003, USA*

²*Department of Physics, New York University, New York, New York 10003, USA*

(Received 24 February 2016; revised manuscript received 10 August 2016; published 14 October 2016)

The composition of a genome with respect to all possible short DNA motifs impacts the ability of DNA binding proteins to locate and bind their target sites. Since nonfunctional DNA binding can be detrimental to cellular functions and ultimately to organismal fitness, organisms could benefit from reducing the number of nonfunctional DNA binding sites genome wide. Using *in vitro* measurements of binding affinities for a large collection of DNA binding proteins, in multiple species, we detect a significant global avoidance of weak binding sites in genomes. We demonstrate that the underlying evolutionary process leaves a distinct genomic hallmark in that similar words have correlated frequencies, a signal that we detect in all species across domains of life. We consider the possibility that natural selection against weak binding sites contributes to this process, and using an evolutionary model we show that the strength of selection needed to maintain global word compositions is on the order of point mutation rates. Likewise, we show that evolutionary mechanisms based on interference of protein-DNA binding with replication and mutational repair processes could yield similar results and operate with similar rates. On the basis of these modeling and bioinformatic results, we conclude that genome-wide word compositions have been molded by DNA binding proteins acting through tiny evolutionary steps over time scales spanning millions of generations.

DOI: [10.1103/PhysRevX.6.041009](https://doi.org/10.1103/PhysRevX.6.041009)

Subject Areas: Biological Physics, Statistical Physics

I. INTRODUCTION

DNA sequences encode information that is read and interpreted through molecular binding by proteins including transcription factors (TFs), nucleosomes, the RNA polymerase complex, and DNA replication machinery. A DNA binding protein must discriminate a small number of functional binding sites on DNA (~ 1 – 100) from the much larger set of possible sites genome wide ($\sim 10^6$ – 10^9), a fundamental molecular process that has been the subject of numerous biophysical studies (see, e.g., Refs. [1–4]). The size of a binding site varies depending on the DNA binding protein, with typical values in the range of 4–10 base pairs (bp) [5], while subsets of DNA nucleotide “words” or k -mers that are specifically bound by the protein are loosely referred to as its *binding motifs*. Given this challenging recognition problem, DNA binding proteins are not as specific as one would naively expect. Most DNA binding proteins recognize degenerate patterns; i.e., they can bind strongly to tens or hundreds of different possible words and weakly to thousands or more [6,7]. Indeed, genome-wide measurements consistently detect proteins bound to DNA at

nonfunctional (or ectopic) sites throughout the genome [8,9]. Given these empirical observations, a fundamental question is whether the functional demands of DNA binding proteins—which are dictated by organismal physiology and typically require a high degree of discrimination between functional and nonfunctional sites—impose evolutionary constraints on genomic DNA sequences, e.g., to enhance their global capacity for target discrimination. In general, we would like to know how strongly and over what time scales can natural selection act on such constraints, as well as whether and how such a process has shaped genomes. To address these key questions at the nexus of molecular biophysics and evolutionary biology, we develop a new approach for global genomic analysis, and use it to perform wide-ranging analyses of DNA sequence evolution.

We identify two major consequences of nonfunctional protein-DNA binding. First, there are *interference effects*, where such binding can disrupt various molecular processes [10], including transcription [11], gene regulation [12], replication [13,14], and mutational repair [15,16]. For example, ectopic binding within a protein coding region could hinder transcription of the gene, slowing down or preventing synthesis of complete mRNA molecules. Recent analyses of cancer sequencing data indicate that transcription factors bound to DNA can interfere with mutational repair machinery, which leads to elevated mutation rates at DNA binding motifs in certain cell types [15]. Second, there is a *titration effect*, where nonfunctional

Published by the American Physical Society under the terms of the [Creative Commons Attribution 3.0 License](https://creativecommons.org/licenses/by/3.0/). Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

binding of a regulatory protein titrates copies of the protein away from its functional sites, thereby reducing the efficiency of gene regulation [17]. While these deleterious effects may be small for any given protein, each genome encodes a large number of DNA binding proteins ($\sim 10^2$ – 10^3) that are each expressed at $\sim 10^1$ – 10^4 copies within each cell. The aggregate effect of nonfunctional DNA binding genome wide could have a significant impact on cellular physiology. To increase specificity, DNA binding proteins often act *cooperatively* by binding multiple, closely spaced binding motifs on DNA, which enables the formation of stabilizing protein-protein contacts [18]. Although molecular cooperativity has important regulatory functions unrelated to specificity, by lowering the free energy of binding at functional sites it also reduces the frequency of ectopic binding. In eukaryotes, DNA is packaged and organized into higher-scale structures that modulate the accessibility of different genomic regions to DNA binding proteins [18]. By rendering parts of the genome inaccessible to DNA binding, genome organization can reduce interference and titration effects. Nevertheless, accessible genomic regions constitute a large amount of DNA (4.5×10^8 bp in the human genome) and a substantial proportion of transcribed regions [19], while cooperative binding reduces but does not eliminate non-functional binding.

If nonfunctional binding constitutes a significant burden on cellular processes, then evolution could lead to a reduction of the number of ectopic binding sites genome wide. Until now, very few examples of genome-wide avoidance of binding motifs have been found. Analyses in bacteria have shown that promoter elements [20,21], transcription and translation boundary signals [22,23], and restriction sites [24–27] are often statistically avoided in genome sequences, yet these constitute only a handful of binding motifs that impact a tiny portion of DNA sequence space in any given species. A more general finding was made soon after the publication of the *E. coli* genome, when it was noticed on the basis of motif prediction algorithms that coding regions tend to avoid predicted recognition sequences for a collection of known *E. coli* DNA binding proteins [28]. In eukaryotes, it remains largely unknown whether genome sequences have been substantially shaped by DNA binding-related evolutionary constraints. Importantly, most previous studies focused on strong DNA binding sites that are genetically well characterized, but overlooked weak DNA binding motifs, which are more difficult to define experimentally. Since for any given TF there are exponentially more weak binding k -mers than strong ones [6,7], weak ectopic binding to a large number of sites genome wide is potentially much more detrimental to cellular functions than strong ectopic binding, which may be comparatively rare.

Here, we set out to determine whether signals of evolutionary constraints imposed by protein-DNA binding

of arbitrary strength can be detected in genomes, and if so, what are the magnitudes of the relevant evolutionary forces. By correlating *in vitro* measurements of DNA binding with genome-wide word statistics, we show that genomes have evolved to reduce the occurrence of weak binding motifs. We demonstrate that the distinct set of DNA binding proteins coded in each species' genome imposes a large set of global, evolutionary constraints that have ubiquitously shaped genome-wide motif statistics. We show that a hallmark signal of this process can be detected in all available genomes. We introduce a mathematical model of this process and use it to infer the time scales over which evolution under DNA binding constraints has shaped genomic motif composition across all domains of life.

II. RESULTS

A. Protein-DNA binding measurements and genome-wide motif statistics

To determine whether genome sequences exhibit a signature of global evolution to avoid nonfunctional protein-DNA binding, we use *in vitro* measurements of protein-DNA binding specificities. A number of techniques are available that yield DNA binding specificity information [29]; however, for our purposes it is particularly important to assess not only the strongest binding sequences but the full range of binding strengths to all possible DNA motifs for each DNA binding protein. Currently, the only technique that enables an unbiased measurement of this kind is a protein-binding microarray [30], which consists of $\sim 44\,000$ spots, each corresponding to a unique 60 bp DNA probe. The set of probes is designed to span the set of all possible 8-mers (total 32896), such that each nonpalindromic 8-mer is present in at least 32 different spots (palindromic 8-mers occur in at least 16 spots). Measurement of protein-DNA binding is performed by purifying a protein, which is applied to the microarray and detected by a fluorophore-conjugated antibody. A value proportional to binding strength is obtained for every possible 8-mer by averaging the fluorescence measurements over all spots that contain a given 8-mer. Such binding measurements are available for several hundred DNA binding domains from several different organisms [31]. We initially study the mouse data set [6], which is one of the most extensive, in which binding measurements are available for over 100 TFs. For each TF, the log binding intensity values are centered to the median, and normalized by their dispersion, yielding a binding score b_i for every possible 8-mer word i . Figure 1(a) shows the distribution of binding scores for a single TF (Mafk) across all 8-mers (see File S1 in Supplemental Material [32] for all TFs). Words in the positive or negative tails correspond to very strong or very weak binding, respectively. For most TFs, the majority of words lie along a continuum of binding levels without substantial gaps, consistent with previous

observations that TFs typically exhibit degeneracy of their target preferences [33,34].

To correlate binding measurements with the genomic motif composition, we analyze the intron regions of the mouse genome, which constitute ~ 750 Mb or nearly 30% of the total genomic DNA. Introns are ideal for detecting binding-related evolutionary constraints because (i) they are largely devoid of local selective pressures that would complicate the analysis (e.g., protein coding-related selection in exons) and (ii) they reside in genic regions and are thus generally accessible to binding factors. Simply correlating *in vitro* binding scores with genomic word frequencies, however, is highly misleading due to two effects. First, the nucleotide composition (G/C content) of the genome is a major predictor of word usage, and it has been argued that G/C content may be determined by both mutational biases (i.e., unequal rates of mutation between the four nucleotide bases of DNA) as well as natural selection [35,36]. TFs that bind words composed of more (less) frequent nucleotides will automatically exhibit positive (negative) correlations between binding scores and genomic k -mer frequencies, regardless of whether or not G/C content has been evolutionarily shaped by binding-related global constraints. In our analysis, we therefore use *relative* word frequencies, i.e., normalized by expectation based on genome

nucleotide composition. Second, in vertebrates and plants, the dinucleotide CpG is hypermutable (e.g., in mice its mutation rate is ~ 10 times the average point mutation rate) causing genome-wide depletion of words that contain it [37]. The distribution of k -mer frequencies in mouse introns is multimodal due to differences in words' CpG content [Fig. 1(b)]. To account for this large effect, which masks smaller differences among words, we correlate k -mer binding scores versus relative word frequencies separately for words with different CpG content.

The example in Fig. 1(c) shows pronounced negative correlations in each CpG category, indicating that the stronger binding a word, the less frequently it is used in the genome relative to expectation. Correlating binding scores of all words against their relative frequencies reveals highly significant negative correlations in each CpG category for a majority of TFs (63/115), with comparatively few TFs exhibiting positive correlations in each CpG category (9/115) [Fig. 1(d)]. For most TFs, we observe that words with below-average binding ($b_i < 0$) exhibit highly significant negative correlations of relative frequencies with binding scores (92/115) while very few have positive correlations (5/115) (see Fig. S1 in Supplemental Material [32]). Similar results are obtained in worm (21 TFs), fly (14 TFs), human (41 TFs), and yeast (89 TFs)

FIG. 1. Mouse genomic binding landscape and *in vitro* DNA binding measurements. (a) Distribution of binding scores b_i for the Mafk TF over all 8-mer words i . (b) Distribution of 8-mer word frequencies f_i in mouse introns (black curve); f_i are shown normalized with respect to expectation based on genome-wide nucleotide composition. Words are shown as separate histograms according to their CpG counts (colored bars). (c) Correlation of b_i and $\log f_i$ for the Mafk TF in separate CpG categories. Color indicates density of points. (d) Correlation coefficients (Spearman's ρ) of b_i versus f_i are shown for each mouse TF separately computed conditioned on word CpG content. Bars are shown only for statistically significant correlations, with p value $< 10^{-6}$. (e) Binding scores of words (b_i) are correlated with the average binding score of their mutational neighborhoods ($\tilde{b}_i$); results shown for Mafk, $\rho(b, \tilde{b}) = 0.87$. (f) Correlation of f_i and $\tilde{f}_i$ over all 8-mer words i for mouse introns; words are colored according to CpG content, and frequencies are normalized as in (b) and (c).

genomes (see Fig. S2, Table S1, and File S1 in Supplemental Material [32]). We conclude that statistically significant correlations exist between binding scores and genomic relative word frequencies, and that, in general, weak binding words are avoided compared to even weaker binding words.

B. Genomic hallmark of global protein-DNA binding constraints

We sought a more general method that could be applied in the absence of *in vitro* measurements to detect global sequence constraints due to protein-DNA binding. We noticed that, consistent with the biophysics of protein-DNA interaction [1], TF binding scores of words that differ at a single nucleotide are strongly correlated: for each word i , we plot its binding score b_i versus the average binding score $\tilde{b}_i$ of its “mutational neighbors”—all words that differ from i at a single nucleotide [see Fig. 1(e) herein and File S1 in Supplemental Material [32]]. For all 241 TFs that we analyze, we find a general statistical rule that similar words have similar binding strengths. Therefore, if genomes have evolved under protein-DNA binding constraints, then we should detect a strong correlation between the frequencies of similar words, since similar words would be under similar constraints. In Fig. 1(f), we show the result for the mouse genome, where for each 8-mer word i , its relative frequency f_i is plotted versus the average relative frequency $\tilde{f}_i$ of its mutational neighbors. Consistent with the hypothesis, we observe a strong correlation of f_i and $\tilde{f}_i$ ($\rho > 0.85$ within each CpG group). Below, we refer to such word frequency correlations, observed in a given genome, as *word-neighbor correlations*.

We test for word-neighbor correlations in a large collection of fully sequenced genomes spanning all domains of life. Word frequencies are measured separately in exon and intron regions, and normalized by using appropriate null models that account for mutational biases and other compositional effects. For exons, we use synonymous codon and dicodon shuffling schemes to construct partially randomized DNA sequences that preserve amino acid sequences, genomic codon biases, and nucleotide base composition. The dicodon shuffling scheme additionally preserves the frequencies of all k -mers for $k \leq 4$ [22]. The randomized sequences are scanned to determine expected word frequencies for exons. For introns, we compute expected word frequencies based on genome-wide nucleotide (1-mer), dinucleotide (2-mer), or trinucleotide (3-mer) frequencies (see Sec. IV).

All tested genomes (947 bacterial or archaeal genomes, 1304 eukaryotic chromosomes from 75 species) exhibit striking correlation between f_i and $\tilde{f}_i$ in exons [Fig. 2(a) herein and Figs. S3, S4, S7, S23 of Supplemental Material [32]] and in introns (Figs. S5, S6, S6B of Supplemental Material [32]) for each of the null models. Further analysis indicates that RNA-related constraints, such as secondary structures and RNA binding motifs, do not play a significant role in shaping word frequencies in genic regions, and that DNA binding proteins provide the major set of global constraints (see Sec. 2.2 of Supplemental Material [32]). For noncoding portions of the genomes, we analyze word statistics across intergenic regions (Fig. S25 of Ref. [32]), as well as in the DNaseI hypersensitive regions of the human genome, which are experimentally determined

FIG. 2. Word-neighbor correlations are observed in genomes across all domains of life. (a) Word frequency plots of f_i versus $\tilde{f}_i$ in exons of representative bacterial and eukaryotic genomes. Points correspond to all possible 6 bp words. Frequencies are shown relative to the null expectation from synonymous codon shuffling, where a value of 1 means the observed frequency is equal to the expected frequency. For eukaryotes, frequencies are computed across all chromosomes. (b) Word frequency plots from exons of individual human chromosomes (see Fig. S9 of Supplemental Material [32] for all data).

binding-accessible regions (Fig. S22 of Ref. [32]). In all cases, we observe similarly high levels of word-neighbor correlation, and additionally, we find that normalized word frequencies correlate strongly across all tested regions (Figs. S8 and S24 of Ref. [32]). Eukaryotic genomes are further analyzed on a chromosome-by-chromosome basis. In human chromosomes, for example, the overall shape of the f versus $\tilde{f}$ plots from exons is qualitatively similar across chromosomes [Fig. 2(b) herein and Fig. S9 of Ref. [32]]. Comparing relative word frequencies f_i between different chromosomes, we find high correlation coefficients ($\rho > 0.9$) for most chromosome pairs within each genome (Table S4 of Ref. [32]). Deviations are observed for short or Y chromosomes [Fig. 2(b) herein and Fig. S10A of Ref. [32]; see also Sec. II F].

C. Ancient phylogenetic signal of global evolutionary constraints

Evidence of a persistent global evolutionary process acting on k -mers can be found in a principal component analysis (PCA) we perform on relative word frequencies in introns (2-mer null) across eukaryotic species (Fig. 3).

Chromosomes within a single genome form tight clusters that demarcate species. Although major groups of eukaryotes can be clustered to a certain extent using genome-wide di-, tri-, and tetranucleotide frequencies [38,39] as well as codon biases [40], we achieve a significantly higher resolution using data exclusively from the intron regions of individual chromosomes. Closely related species are spatially proximate, with plants, invertebrates, and vertebrates following different directions in the PCA space (Fig. S17 of Ref. [32]).

Our analysis indicates that primates experienced a “jump” in global constraints from the rest of the mammals (Fig. S18 of Ref. [32]), while within this group, no species clusters are detected (Fig. 3), indicating that the global constraints have not significantly changed since the last common ancestor. In prokaryotes, phylogenetic signals are generally very weak or nonexistent when assessed across all genomes (Fig. S19 of Ref. [32]), but may persist at the genus and species level (Fig. S20 of Ref. [32]), indicating that over the longest evolutionary time scales global constraints can change so extensively that the most ancient parts of the signal are extremely faint and difficult to detect.

FIG. 3. Cross-species comparison of intron word frequencies using principle components analysis (PCA). Each point in the PCA plots corresponds to a single chromosome projected on the two principal axes. Principal axes are computed using the 6-mer relative word frequency vectors (2-mer null model) of all chromosomes shown in each plot. See Table S3 of Ref. [32] for further species information.

D. Evolutionary model of global selection

To analyze the role of natural selection in shaping global word frequencies, we examine the consequences of a global selective process acting differentially on words across a genome using an evolutionary model. A sequence of length L is represented by a k -mer composition vector $\mathbf{p}$ in the set $\mathcal{L} = \{\mathbf{p} | p_i \geq 0, \sum_{i=1}^n p_i = L\}$, where $n = 4^k$ is the total number of possible words and p_i is the number of occurrences of word i in the sequence. We denote by $\mathbf{s}$ the selection vector whose i th component is the selective coefficient s_i acting on each word of type i . We consider a large population of evolving sequences, each represented by a composition vector. At each generation, each sequence contributes offspring proportional to its total fitness, $\exp(\mathbf{s} \cdot \mathbf{p})$. The offspring mutate according to a mutational transfer operator $G(\mathbf{p} | \mathbf{p}')$, which gives the probability of a mutational transition $\mathbf{p}' \rightarrow \mathbf{p}$. Under these dynamics, the expected frequency of $\mathbf{p}$ within the population at generation t , denoted by $P_t(\mathbf{p})$, evolves according to the equation

$$\lambda_t P_{t+1}(\mathbf{p}) = \sum_{\mathbf{p}' \in \mathcal{L}} e^{\mathbf{s} \cdot \mathbf{p}'} G(\mathbf{p} | \mathbf{p}') P_t(\mathbf{p}'), \quad (1)$$

where λ_t is a normalizing factor, which measures the average population fitness at generation t , given by summing both sides over all values of $\mathbf{p}$. The irreducibility of the transfer operator G guarantees (by the Perron-Frobenius theorem [41]) that the linear mapping of $P_t \rightarrow P_{t+1}$ converges to a unique steady-state distribution $P(\mathbf{p})$, which dictates the population structure at mutation-selection balance.

The fitness model assumes that the contribution of each word to the global fitness is independent of all other words, which is a natural assumption given the small effect sizes one expects for individual words. Since each locus along DNA mutates independently, the transfer operator can be decomposed as a product over all loci, which implies that $P(\mathbf{p})$ is a multinomial distribution [26]. A multinomial random variable $\mathbf{x}$, denoted $\mathbf{x} \sim \text{mult}(\mathbf{f}, L)$, results when sampling L independent events from a discrete probability distribution f_i , where x_i is the number of outcomes of type i . With this notation, we have $\mathbf{p} \sim \text{mult}(\mathbf{f}, L)$, where $\mathbf{f}$ is the single-locus distribution of words at mutation-selection balance. This distribution $\mathbf{f}$ is obtained as the unique positive solution of the eigenvalue problem

$$e^{\mathbf{S}} \mathbf{A} \mathbf{f} = c \mathbf{f}, \quad (2)$$

where $\mathbf{S} = \text{diag}(\mathbf{s})$ and A_{ij} is the probability that word j mutates into word i . Once more the Perron-Frobenius theorem guarantees a unique positive eigenvector $\mathbf{f}$ with associated eigenvalue c , which determines the equilibrium average fitness, $\lambda_t \rightarrow \lambda = c^L$. For known selective coefficients and mutational transition rates, Eq. (2) can be solved numerically.

To invert the relationship, and obtain an expression for s_i given $\mathbf{f}$, we use the fact that the point mutation rate u (per bp per generation) is extremely small. We express A explicitly as $A = e^{uk(M-I)}$, where I is the identity matrix and M_{ij} is the probability that a single mutation converts word j into word i ; M_{ij} is nonzero only for mutational neighbors, i.e., words i and j that differ by exactly one mutation. If we assume unbiased mutation, $M_{ij} = 1/(3k)$ for any pair of mutational neighbors i and j , and expand A in small u to first order, i.e., $A \approx I + uk(M-I)$ neglecting contributions from double (or higher order) mutants that occur within a single word locus, then Eq. (2) yields

$$ce^{-s_i} f_i \approx f_i + uk \left(-f_i + \frac{1}{3k} \sum_{j \in N(i)} f_j \right), \quad (3)$$

where $N(i)$ is the set consisting of the $3k$ mutational neighbors of word i . The second term within the parentheses is $\tilde{f}_i$, the average frequency of mutational neighbors. Dividing both sides by f_i and taking logarithms, we find

$$s_i \approx uk(1 - \tilde{f}_i/f_i) + \log c. \quad (4)$$

Selective coefficients relative to u are thus determined up to an additive constant by the ratio of a word's frequency and the average frequency of its mutational neighbors. More generally, if words mutate with different rates ku_j , and M_{ij} is not necessarily unbiased, a generalization is easily obtained by defining the diagonal matrix $U \equiv \text{diag} \tilde{u}$ and using $A = e^{k(M-I)U} \approx I + k(M-I)U$ in Eq. (2), which yields

$$s_i \approx ku_i(1 - \tilde{J}_i/J_i) + \log c, \quad (5)$$

where $J_i \equiv ku_i f_i$ is the mutational flux out of word i and $\tilde{J}_i \equiv \sum_{j \in N(i)} ku_j f_j M_{ij}$ is the mutational flux into word i from its mutational neighbors [42].

Figures 4(a)–4(c) show examples of the equilibrium solution [Eq. (2)] when selective coefficients are randomly sampled from a normal distribution with standard deviation σ . Words experiencing similar pressures fall on the lines in the $(\tilde{f}, f)$ plane predicted by Eq. (4) [Fig. 4(a)]. This result encapsulates a basic insight of our analysis: the genome-wide frequency of a word says little if anything about the global pressure it experiences. Words can be under- or overrepresented simply because their mutational neighbors are under negative or positive pressures, respectively. Indeed, selective coefficients can be inferred only when a word is viewed relative to its mutational neighbors.

When selection is weak relative to mutation [$\sigma < u$, Fig. 4(b)], the word cloud collapses toward a point in which all words are effectively neutral, while under strong

FIG. 4. Evolutionary model of global selection. (a)–(c) Solution of the model at mutation-selection balance. Plots in the $(\tilde{f}, f)$ plane of the equilibrium word frequencies for independent, normally distributed selective coefficients $s_i \sim \mathcal{N}(0, \sigma^2)$, using the indicated values of σ . We numerically solve Eq. (2) to determine the equilibrium word frequencies. Color shows different bins of selective coefficients, and predicted equal-pressure lines [Eq. (4)] are drawn using each interval’s average pressure. Inset in (c) shows a zoomed view. Parameters are $k = 6$ and $u = 10^{-4}$, and frequencies are shown relative to the neutral expectation 4^{-k} . (d) Equilibrium for strongly correlated pressures on neighboring words. (e) Correlation is reduced by shuffling 50% of words’ pressures from (d).

selection [$\sigma > u$, Fig. 4(c)] the word with the maximum selective coefficient dominates the distribution. Only in the intermediate regime [$\sigma \approx u$, Fig. 4(a)] does the solution take the form of an extended word cloud, with frequencies varying from approximately twofold avoidance to twofold enrichment. A pronounced positive correlation between f_i and $\tilde{f}_i$ is seen, despite the fact that all words experience independent random pressures. This correlation results from selection, which modulates the frequency of neighbor words in order to alter mutational fluxes into words under selection (similar results are obtained with a skewed pressure distribution, Fig. S11 of Supplemental Material [32]). The correlation of frequencies can be increased further by introducing positive correlations in the pressures on similar words, resulting in an extended, rotated word cloud [Figs. 4(d) and 4(e)], while the equal-pressure lines remain unchanged.

E. Distribution of global selective coefficients in genomes

We apply the evolutionary model (Sec. II D) to measure the global selective coefficients in each genome. Word frequencies f_i are counted separately in introns and exons, and expected frequencies f'_i are obtained using the null models described above. We measure global selective coefficients using the *excess pressures* $\Delta s_i \equiv s_i - s'_i$, where s_i and s'_i are separately inferred from f_i and f'_i , respectively, using Eq. (4). One can think of the selective coefficients s'_i as a kind of “external field” that maintains words at their null expectation. For example, in exon regions the

requirements of protein coding impose a set of selective constraints on words, which we view as an external field that can be calculated using the word frequencies obtained from synonymous codon shuffling. By using excess pressures Δs_i rather than s_i , we measure the strength of selection needed to shift word frequencies to their observed values from their expected values based on the null models (see Supplemental Material [32]). In intron regions, mutational biases such as CpG hypermutation and other less pronounced biases can influence the observed frequencies of dinucleotides. Excess pressures relative to the 2-mer null model in introns measure only the global selective pressures that are not already accounted for by dinucleotide composition.

In the context of protein-DNA binding, Δs_i corresponds to the average fitness effect due to occupancy by DNA binding proteins at each occurrence of word i genome wide [43]. In Fig. 5, the distribution of selective coefficients is shown for several genomes (see Figs. S3–S7 of Ref. [32] for additional genomes and null models). In all cases, the bulk of the distribution has a width comparable to u . The selective coefficients Δs_i measured on different chromosomes are strongly correlated, with overall very similar distributions (Fig. S9 of Ref. [32]). Importantly, Eq. (4) allows us to determine the pressures on each word in spite of pressures acting on their mutational neighbors. We find that the selective pressures on similar words are indeed highly correlated (Fig. S13 of Ref. [32]). We conclude that if genome-wide word frequencies have been maintained by natural selection acting globally across a genome, the magnitude of the relevant selective pressures would be

FIG. 5. Distribution of global selective coefficients in different genomes. (a) Selective coefficients of 6-mers in four eukaryotic species. Intron data are used and Δs_i values are given with respect to the 2-mer null model. (b) Selective coefficients of 6-mers measured in two bacterial species. Exon data are used and Δs_i values are given with respect to the synonymous codon-shuffling null model. In *E. coli* (bottom panels), an inset shows the bulk of the distribution, since a small number of words including known restriction sites have large negative coefficients. Pressure distributions are shown using a value of $c = 1$ in Eq. (4), which corresponds to taking the average global fitness of a sequence to be one.

on the order of point mutation rates. Given that our mathematical model assumes mutation-selection balance, these correspond to the magnitude of selection necessary to maintain genome-wide motif statistics over evolutionary time scales (known as “purifying selection” in evolutionary biology [44]).

Since the effect sizes on individual words are tiny, we asked whether the global fitness differences between individuals are sufficiently large to constitute non-negligible selective differences for a finite size population. That is, these selective differences should be larger than the magnitude of demographic fluctuations, which scales as $\sim 1/N_{\text{eff}}$, where N_{eff} is the suitably defined effective population size that accounts for fluctuating population sizes [45]. If two individuals differ at l sites, or kl words, and each word contributes an average effect size $\pm \bar{s}$, their global fitness difference $\Delta S \sim \bar{s} \sqrt{kl}$. In human populations, pairs of individuals differ at 4.6×10^6 sites on average [46], and taking $k = 6$ and $\bar{s} \sim u$ (Fig. 4), with $u = 10^{-8}$ [47], we have $\Delta S \approx 5 \times 10^{-5}$. Effective global selection thus requires $N_{\text{eff}} \gtrsim 10^4$. In *E. coli*, isolates differ on average at 1%–2% of sites [48], or approximately 7×10^4 sites. Taking $\bar{s} \sim u$, and using $u = 2.2 \times 10^{-10}$ [49], we find $\Delta S \approx 1.4 \times 10^{-7}$; hence, global selection requires $N_{\text{eff}} \gtrsim 10^7$. The per-word selective coefficients we measure in these genomes are thus sufficient to maintain genome-wide statistics given the present levels of diversity in these populations and previous estimates of their effective population sizes [45]. We note that the total benefit of global adaptation, compared to nonadapted sequences, scales as Lu , and is therefore much larger than the global fitness differences that exist within any globally adapted population.

F. Effect of recombination on global selection

Our evolutionary model [Eq. (1)] propagates the expected frequencies of different word vectors $P_i(\mathbf{p})$ and, thus, carries the implicit assumption of an infinitely large population. For finite asexual populations of size N_{eff} with genomes of size $L \gg 1$, deviations from this equilibrium will occur due to Muller’s ratchet [50,51]. This general effect has been discussed in a statistical physics perspective [52], and we briefly review the same reasoning here. Because of the large genotype space, the maximally fit sequence—i.e., one in which the most favorable word is present at each locus—is rapidly lost from the population, because the rate of mutation away from this sequence, Lu , is much larger than the rate of back mutation u . Within a few generations, due to the finite population size, every sequence has accumulated at least one deleterious word, and the fittest sequence in the population is now one that carries at least one mutation. This process drives the population away from the fittest sequence, reaching an equilibrium distribution that is different from $P(\mathbf{p})$, which depends in a complex manner on N_{eff} , L , and u [52]. Recombination, however, can reverse the ratchet effect, by enabling the previously fittest sequence to be recovered. Similarly, for sexual populations, in genomic regions with low recombination rates, ratchetlike effects of various kinds (known as Hill-Robertson effects [53,54]) will oppose global adaptation.

Accordingly, simulations of a nonrecombining finite population initialized at the predicted mutation-selection balance evolve to lower fitness (Fig. S15 or Ref. [32]), while recombination enables the equilibrium genome composition to be efficiently maintained under global

selection (Fig. S14 of Ref. [32]). Similar behavior in a different simulation model has previously been shown [53]. Consistent with this predicted dependence of global selection on recombination, we find that chromosomes whose word frequencies exhibit deviations from the genome-wide average tend to be nonrecombining sex chromosomes or to have smaller genetic length than average (Fig. S10 of Ref. [32]). In the human genome, the genome-wide average number of mutations per crossover per generation is 0.87 [55], while in bacteria homologous recombination replaces small fragments of a genome with homologous fragments from other cells, and occurs with rates per site that are comparable to or greater than u [56], allowing global selection to maintain genome motif composition. Since global adaptation results in a total fitness benefit that scales as Lu , and which is only achieved for sufficiently high recombination rates, the presence of global selection would thus yield a generic selective advantage for recombination.

G. Neutral mechanisms for evolution of word-neighbor correlations

Using a separate set of analyses, we determine whether neutral mechanisms, i.e., ones that operate in the absence of selective differences between individuals, could account for word frequency distributions and word-neighbor correlations. The spontaneous mutation process can exhibit certain species-specific biases, such that mutational transition rates depend on the type of nucleotide and in some cases on its immediately proximate nucleotides (see Supplemental Material [32] for a detailed discussion). To determine whether these types of mutational biases can account for word-neighbor correlations, we run a wide range of tests, which include performing an analysis of variance on k -mer frequencies using their dinucleotide and trinucleotide composition (Table S5 of Ref. [32]), analysis of word-neighbor correlations using regression residuals (Fig. S16 of Ref. [32]), a dicodon shuffling scheme that accounts for mutational biases in bacteria (Fig. S23 of Ref. [32]), and explicit incorporation of mutational biases into evolutionary models (Fig. S12 of Ref. [32]). In each of these tests, mutational biases are unable to account for the observed word frequency distributions and word-neighbor correlations (see Sec. 2.1 of Supplemental Material [32]).

Another potentially relevant neutral mechanism involves repeat-derived sequences in genomes. These are portions of the genome that have resulted from local duplications and transposon-mediated events, which amplify certain types of sequences and could possibly skew word composition as a result. Since large eukaryotic genomes have a substantial amount of repeat-derived sequence, we test whether word-neighbor correlations might arise from a balance between amplification of specific classes of mobile and/or repeat-containing elements and mutational degradation. We analyze the repeat-masked sequence of the human genome, a procedure that removes approximately 45% of the sequence

(Fig. S21 of Ref. [32]), and find that it exhibits strong word-neighbor correlations ($r \geq 0.92$ for all null models, Fig. S21C of Ref. [32]); similar results are found across a wide range of species (Fig. S25 of Ref. [32]). Repeat expansion is, therefore, not responsible for word-neighbor correlations, and word frequencies are strongly correlated between repeat-masked and repeat-derived regions (Fig. S21B of Ref. [32]). A different possibility is that, due to ubiquitous small insertions and deletions (indels) occurring in all genomic regions, word composition could be determined by slippage-based mutational mechanisms in which DNA polymerase inserts or deletes short runs of DNA (see, e.g., Ref. [57]). While repeat masking cannot detect this finer-scale process, it is well known from comparative genomics and mutation accumulation studies in different species that indels occur between 0.03 and 0.13 times as frequently as point mutations [35,49,58–60]. Our evolutionary model shows that processes that change word frequencies at much slower rates than the point mutation rate cannot yield the observed word-neighbor clouds [Fig. 4(b)]. These results indicate that neither mutational biases nor neutral processes involving repetitive DNA can account for the ubiquitous word-neighbor correlations we observe.

Lastly, we consider interference by DNA binding proteins with DNA replication and repair machinery as a process that could in principle introduce species-specific mutational biases on k -mers of various sizes. Detecting such binding-related mutational biases in natural populations would require statistical power beyond that available in current data sets. Recent analysis of sequencing data from human cancer cell lines indicates that such interference effects exist for certain DNA binding proteins in specific cell types under UV exposure, although their molecular mechanisms are currently unknown [15]. Assuming in our model that words mutate with rates proportional to their probability of being bound q_i , we obtain from Eq. (5), in the absence of selection ($s_i = 0$), that mutational fluxes must balance and, therefore, $f_i \sim m_i/q_i$, where m_i is the right eigenvector of M_{ij} . Equilibrium word frequencies would therefore be inversely proportional to binding probability, as expected if the severity of interference increases with binding strength. This relation may implicitly involve nonlinearities, since q_i could depend on the genome composition $\mathbf{f}$ and other parameters [43].

III. DISCUSSION

Variation in genome composition across species has been extensively studied, from global G/C content differences to codon usage biases [40,61–65]. Early work by Karlin and co-workers found that dinucleotide frequencies differ between species [38,66], while further differences were later noticed at the level of longer k -mer words [39,67]. There has been little consensus on the underlying causes of

such differences, and a wide variety of mechanisms and explanations have been proposed [35,36,40,61,63,68,69]. Several works have demonstrated that mutational biases and neutral mechanisms cannot explain basic differences in bacterial genome composition [36,63,69], in codon usage biases [62], or in genome-wide patterns of variation in flies [44]. The possibility of multiple layers of “overlapping codes” that constrain coding sequences was proposed and tested in a wide range of prokaryotic and eukaryotic species in Ref. [22], which along with several other studies revealed constraints on coding regions resulting from translation initiation signals [23], promoter elements [20,21], restriction sites [24–27], repetitive sequences [70,71], RNA secondary structures [72], and microRNA target sites [73]. Viral genomes are known to experience global constraints on motif usage due to host immune mechanisms [74,75], and the effect of such constraints on the coevolution of specific host innate immune genes and viral genomes has also been noticed [76]. Here, we propose that a large number of species-specific constraints exist in all genomes due to the ubiquitous molecular process of protein-DNA binding.

We present several lines of evidence indicating that genome-wide word frequencies have been shaped by global protein-DNA binding constraints over long evolutionary time scales. We show that extensive *in vitro* binding data exhibit statistically significant correlations with genome-wide relative word frequencies. These correlations are usually negative, indicating that word avoidance scales with binding strength. We identify a general hallmark of global binding constraints—the strongly correlated frequencies of similar words—which we observe in all genomes. We argue that this correlation results from the biophysics of protein-DNA binding: since similar words have similar binding strengths, their protein-DNA binding constraints should likewise be correlated, resulting in the observed word-neighbor correlations. We analyze a wide range of null models, which demonstrate that word-neighbor correlations cannot be attributed to known mutational biases. Finally, we analyze the phylogenetic signal of global sequence evolution, which we now discuss.

The persistence of a phylogenetic signal in relative word frequencies over the large evolutionary distances seen in Fig. 3 herein and Figs. S17 and S18 of Ref. [32] cannot be explained by an unconstrained neutral divergence process. For example, *D. melanogaster* and *M. musculus* diverged over 600 million years ago, yet their relative word frequencies (2-mer null) exhibit a correlation of $\rho = 0.43$ (Table S4 of Ref. [32]). Billions of generations separate these genomes, and considering that their per bp per generation mutation rates are $\sim 10^{-8}$ [77], on average every neutral position will have mutated one or more times. Since our analysis is performed in introns, most of the ancient signal would have been destroyed, leaving essentially no correlation. We propose that this phylogenetic signal persists

because DNA binding proteins’ specificities are conserved over long evolutionary time scales, and they continue to exert similar global constraints on genomes. Consistent with this prediction, the motif specificities of homologous transcription factors from flies and mice have been shown *in vitro* to be highly conserved [78].

The ability to sensitively resolve eukaryotic species based on word usage arises from two major factors: (a) that global word usage is influenced by a large number of DNA binding protein motifs and thus provides a summary signature of each genome’s DNA binding protein repertoire, and (b) that introns are largely free of other evolutionary constraints, such as protein coding pressures, which allows these genomic regions to adapt much more fully to global constraints. Much further analysis would be necessary to determine which specific features of the global word usage signal are responsible for phylogenetic resolution. In the well-resolved portions of the bacterial phylogeny, the γ proteobacteria (Fig. S20 of Ref. [32]), we test whether the extremes of the word usage distribution are necessary for phylogenetic resolution. We remove all known restriction sites from the analysis, since these are generally the most strongly avoided words, and find no change in our ability to resolve these species (data not shown). Likewise, how changes in DNA binding proteins map to divergence of global word usage and how the total number of DNA binding proteins in a genome influences the strength of this divergence remain to be studied. To this end, an interesting case is presented by bacteria with strongly reduced genomes and tiny cell sizes, such as *Mycoplasma*, which are known to have very few transcriptional regulators per genome [79], yet which show typical word-neighbor correlations (Fig. S4 of Ref. [32]). Since individual DNA binding proteins each exhibit a continuum of binding scores (File S1 of Ref. [32]), the concentration of DNA binding proteins in the cell may be as important as the overall binding repertoire of these proteins in determining the magnitude of DNA binding constraints acting on a genome.

At the level of individual transcription factors, the correlations of binding scores with genomic word frequencies are highly significant, but the magnitudes of the correlations are not particularly high (Fig. 1 herein and Figs. S1 and S2 of Ref. [32]). This finding suggests that a typical word experiences weak global constraints from a number of distinct binding proteins, with each word’s frequency reflecting an average over a collection of different binding scores, or, equivalently, that the global word distribution is not driven by a small subset of the genome’s DNA binding proteins, but rather reflects an average set of constraints determined by a substantial fraction of them. While these constraints appear to be predominantly due to selection against weak binding motifs, a number of strong positive correlations are observed across the different data sets. Positive correlations are expected when a DNA

binding protein functions in a genome-wide process that involves localization to a large number of sites on DNA, e.g., DNA binding factors that mediate chromatin remodeling on a large scale, as occurs during developmental transitions. Intriguingly, the Zic proteins, which correspond to the three largest positive correlations in the mouse data set [Fig. 1(d), right edge], were recently shown to be involved in binding a large number of sites in the genome that become accessible to DNA binding during the differentiation of cerebellar granule neurons [80].

We consider the possibility that genomic word composition is maintained by global selection, and show that if this were the case, the magnitude of selective coefficients would be on the order of the point mutation rate u . Likewise, we show that if word composition evolves due to binding-related mutational biases, the rate of this process would be on the order of u . In either case, the relevant evolutionary rates are extremely small, since, e.g., $u \sim 10^{-10}$ per generation (*E. coli* [49]) or $\sim 10^{-8}$ (*H. sapiens* [47]). In both cases, the phylogenetic signal of word composition (Fig. 5 herein and Figs. S17 and S18 of Ref. [32]) is preserved over long evolutionary time scales due to strong conservation of DNA binding proteins' binding preferences. However, the two mechanisms make several different predictions that may become testable as larger DNA sequencing data sets become available in different species.

We show that the efficacy of global selection depends on recombination (Figs. S14 and S15 of Ref. [32]), and consistently we find that chromosomes with low recombination rates deviate in word composition from the rest of the genome (Fig. S10 of Ref. [32]). The binding-related mutational bias mechanism predicts no effect of recombination on word composition. However, low recombination rates might be correlated with low accessibility to DNA binding factors in certain cases [81]; hence, it is not straightforward to rule out binding-related biases as a possible explanation. A much finer-scale analysis of recombination rates genome wide combined with reliable DNA accessibility measurements could in principle be used to determine which mechanism is operative in any given species.

Under both mechanisms, the evolutionary rates scale with u , but the underlying cause of this scaling is potentially quite different. For binding-related biases, such scaling is automatic, since the biased mutation rates are all multiples of the point mutation rate. For global selection, the selective coefficients s_i could in principle be arbitrary, and their scaling with u in all genomes we examine would imply some universal set of constraints related to the molecular processes that protein-DNA binding interferes with. While the origin of such scaling is not obvious, neither are the evolutionary forces that have tuned the point mutation rate itself, which varies in nature across 7 orders of magnitude and appears to scale inversely with genome

length in microbes [77,82]. Indeed, one can consider the possibility that point mutation rates themselves have evolved to be on the order of global selective coefficients, which would enable genomes to gain the fitness benefit conferred by global adaptation. Another possibility is that protein-DNA binding has a tiny impact on DNA replication rates, in which case removing one binding site would decrease replication time by an average time unit scaling proportionally with the binding strength and inversely with the genome length, leading to global selective coefficients scaling with u . It is also possible that protein-DNA binding interferes with mutational repair processes, but does not lead to mutational biases. Instead, it could slow down or impair the cell division process, resulting in global selective coefficients that are proportional to both u and the binding strength. Some of these possibilities are testable by highly sensitive laboratory measurements.

Lastly, the two mechanisms differ in their response to population size. Binding-related mutational biases are not predicted to be sensitive to population size, and would operate identically in large or small populations. The efficiency of global selection in maintaining genome composition, on the other hand, will be sensitive to population size, though the effect may be small and the amount of recombination will determine precisely how much more effective global selection will be in larger populations. It may, therefore, be possible, with sufficiently high resolution data and detailed population genetic modeling, to distinguish the mechanisms using measurements from related species of substantially different population sizes.

Our extensive analysis of genomic data demonstrates that protein-DNA binding constraints constitute a universal evolutionary force that acts on genomes. We propose that this force arises from the functions of a diverse and distinct set of DNA binding processes within each genome that together generate a characteristic set of global constraints. As these protein-DNA binding constraints maintain genome-wide motif compositions over long evolutionary time scales, their hallmark can be detected in the correlated frequencies of words and their mutational neighbors. Our findings introduce a new view on genomic evolution, in which the global composition of a genome is constantly being molded by the complement of DNA binding proteins that it expresses.

IV. METHODS

A. Genomic sequence data sets

Sequences are downloaded from Ref. [83]. We manually curate a large, representative set of genomes across all bacterial and archaeal groups (947 genomes). For eukaryotes, all fully assembled and annotated genomes are included in the intron analysis (75 species, 1304 chromosomes; mitochondria and plastid sequences are not included; see Table S3 of Ref. [32]). Among eukaryotes,

exon analysis is performed for 17 species (251 chromosomes; see Table S3 of Ref. [32]). To extract the intron segments, coding sequence coordinates are obtained from the annotation files and then mapped to the full-length genome. For both exon and intron extraction, only one of the splicing isoforms is selected at random.

B. Measuring genomic word statistics

Word frequencies are collected by a sliding window of k bp ($k = 2, \dots, 8$) across the sequences. For the coding region analysis, frequencies are collected on each open reading frame (i.e., multiple exons are joined for eukaryotes) and then combined for all open reading frames on a chromosome. Start and stop codons are ignored. For the intron regions, each intron segment is read separately and then the frequencies are combined. Gaps (N tracts in the eukaryotic sequences), if encountered, are replaced randomly by four bases at equal frequencies. Word frequencies are counted on the plus strand of DNA, and the counts of reverse-complement words are combined.

C. Constructing the null models in exon and intron sequences

For the coding region analysis using codon shuffling as the null model, synonymous codons are shuffled across the entire chromosome sequence. Start and stop codons are ignored. Each shuffled sequence is then scanned as above. The expected word frequencies f'_i are calculated as the average of word counts from 1000 such shuffled sequences. For the intron analysis using nucleotide base composition as the null model (1-mer null), base compositions are calculated on the chromosomal scale. For a word w with n_A, n_T, n_C, n_G counts of A, T, C , and G bases, respectively, the expected frequencies are calculated as $f'_w(1\text{-mer}) \equiv p_A^{n_A} p_T^{n_T} p_C^{n_C} p_G^{n_G}$, where p_A, p_T, p_C, p_G are base compositions measured from the full-length sequences. For analysis using the dinucleotide (2-mer null) or trinucleotide (3-mer null) models, a Markov chain model is used to compute expectations. A k -mer word $w = (w_1 w_2 w_3 \dots w_k)$ is composed of an overlapping set of dinucleotides $(w_1 w_2), (w_2 w_3), \dots, (w_{k-1} w_k)$. The probability of observing each dinucleotide (ab) is measured across all introns conditional on a being at the first position, and denoted $p(b|a)$. Using these measured values, the expected frequency of the word w is given by

$$f'_w(2\text{-mer}) \equiv p_{w_1} p(w_2|w_1) p(w_3|w_2) \cdots p(w_k|w_{k-1}),$$

where p_{w_1} is the measured frequency of nucleotide w_1 . Similarly, for the trinucleotide model, the word is composed of an overlapping set of trinucleotides, and

$$f'_w(3\text{-mer}) \equiv p_{w_1 w_2} p(w_3|w_1 w_2) p(w_4|w_2 w_3) \cdots \times p(w_k|w_{k-2} w_{k-1}),$$

where $p(c|ab)$ is the probability of observing trinucleotide (abc) conditional on the dinucleotide (ab) occupying the first two positions, and $p_{w_1 w_2}$ is the measured frequency of dinucleotide ($w_1 w_2$).

D. Principle components analysis of genomic word frequencies

Principle component analyses are performed on the correlation matrix between relative word frequencies using the 2-mer null model. For eukaryotic intron regions (Fig. 3 herein and Figs. S17 and S18 of Ref. [32]), relative word frequencies of 1049 chromosomes from 61 species are used for the PCA (Fig. S17 legend and Table S3 of Ref. [32]). Excluded are chromosomes that are shorter than 1 Mb or having <0.8 correlation between the measured relative frequencies or excess pressures of reverse complement words on the plus strand, as well as the chromosomes of the bird *Ficedula albicollis*, because its word frequencies take on a split distribution in the $(\tilde{f}, f)$ plane. The word frequencies on these chromosomes are so extreme that they dominate the PCA analysis. The plots for subgroups shown in each panel are produced by PCA on each subgroup alone. For bacterial coding regions (Figs. S19 and S20 of Ref. [32]), relative word frequencies are computed using codon shuffling as the null model, and counts are combined on both strands of all analyzed exon sequences.

ACKNOWLEDGMENTS

We wish to thank Benjamin Greenbaum, Alexander Grosberg, Oskar Hallatschek, Eugene Koonin, Stanislas Leibler, Bruce Levin, Michael Lynch, David Pollock, Michael Puruggannan, Matthew Rockman, David Roos, Ilya Ruvinsky, Eugene Shakhnovich, Boris Shraiman, Stephen Small, Duncan Smith, and Joshua Weitz for valuable discussions. This work was supported by a Human Frontiers Science Program Young Investigator's Grant.

-
- [1] O.G. Berg and P.H. von Hippel, *On the Specificity of DNA-Protein Interactions*, *Proc. Natl. Acad. Sci. U.S.A.* **83**, 1608 (1986).
 - [2] U. Gerland and T. Hwa, *On the Selection and Evolution of Regulatory DNA Motifs*, *J. Mol. Evol.* **55**, 386 (2002).
 - [3] T. Hu, A.Y. Grosberg, and B. Shklovskii, *How Proteins Search for Their Specific Sites on DNA: The Role of DNA Conformation*, *Biophys. J.* **90**, 2731 (2006).
 - [4] V. Mustonen, J. Kinney, C.G. Callan, and M. Lässig, *Energy-Dependent Fitness: A Quantitative Model for the Evolution of Yeast Transcription Factor Binding Sites*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 12376 (2008).
 - [5] A.J. Stewart, S. Hannenhalli, and J.B. Plotkin, *Why Transcription Factor Binding Sites are Ten Nucleotides Long*, *Genetics* **192**, 973 (2012).

- [6] G. Badis, M. F. Berger, A. A. Philippakis, S. Talukder, A. R. Gehrke, S. A. Jaeger, E. T. Chan, G. Metzler, A. Vedenko, X. Chen, H. Kuznetsov, C.-F. Wang, D. Coburn, D. E. Newburger, Q. Morris, T. R. Hughes, and M. L. Bulyk, *Diversity and Complexity in DNA Recognition by Transcription Factors*, *Science* **324**, 1720 (2009).
- [7] A. Jolma, J. Yan, T. Whittington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale, *DNA-Binding Specificities of Human Transcription Factors*, *Cell* **152**, 327 (2013).
- [8] X.-Y. Li *et al.*, *Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm*, *PLoS Biol.* **6**, e27 (2008).
- [9] D. A. Cusanovich, B. Pavlovic, J. K. Pritchard, and Y. Gilad, *The Functional Consequences of Variation in Transcription Factor Binding*, *PLoS Genet.* **10**, e1004226 (2014).
- [10] I. J. Finkelstein and E. C. Greene, *Molecular Traffic Jams on DNA*, *Annu. Rev. Biophys.* **42**, 241 (2013).
- [11] V. Epshtein, F. Toulmé, A. R. Rahmouni, S. Borukhov, and E. Nudler, *Transcription through the Roadblocks: The Role of RNA Polymerase Cooperation*, *EMBO J.* **22**, 4719 (2003).
- [12] D. Villar, P. Flicek, and D. T. Odom, *Evolution of Transcription Factor Binding in Metazoans—Mechanisms and Functional Implications*, *Nat. Rev. Genet.* **15**, 221 (2014).
- [13] C. P. Guy, J. Atkinson, M. K. Gupta, A. A. Mahdi, E. J. Gwynn, C. J. Rudolph, P. B. Moon, I. C. van Knippenberg, C. J. Cadman, M. S. Dillingham, R. G. Lloyd, and P. McGlynn, *Rep Provides a Second Motor at the Replisome to Promote Duplication of Protein-Bound DNA*, *Mol. Cell* **36**, 654 (2009).
- [14] D. Smith and I. Whitehouse, *Intrinsic Coupling of Lagging-Strand Synthesis to Chromatin Assembly*, *Nature (London)* **483**, 434 (2012).
- [15] R. Sabarinathan, L. Mularoni, J. Deu-Pons, A. Gonzalez-Perez, and N. López-Bigas, *Nucleotide Excision Repair is Impaired by Binding of Transcription Factors to DNA*, *Nature (London)* **532**, 264 (2016).
- [16] M. A. M. Reijns, H. Kemp, J. Ding, S. M. de Proce, A. P. Jackson, and M. S. Taylor, *Lagging-Strand Replication Shapes the Mutational Landscape of the Genome*, *Nature (London)* **518**, 502 (2015).
- [17] R. C. Brewster, F. M. Weinert, H. G. Garcia, D. Song, M. Rydenfelt, and R. Phillips, *The Transcription Factor Titration Effect Dictates Level of Gene Expression*, *Cell* **156**, 1312 (2014).
- [18] K. M. Lelli, M. Slattery, and R. S. Mann, *Disentangling the Many Layers of Eukaryotic Transcriptional Regulation*, *Annu. Rev. Genet.* **46**, 43 (2012).
- [19] R. E. Thurman *et al.*, *The Accessible Chromatin Landscape of the Human Genome*, *Nature (London)* **489**, 75 (2012).
- [20] M. W. Hahn, J. E. Stajich, and G. A. Wray, *The Effects of Selection against Spurious Transcription Factor Binding Sites*, *Mol. Biol. Evol.* **20**, 901 (2003).
- [21] J. L. Froula and M. Pilar Francino, *Selection against Spurious Promoter Motifs Correlates with Translational Efficiency across Bacteria*, *PLoS One* **2**, e745 (2007).
- [22] S. Itzkovitz, E. Hodis, and E. Segal, *Overlapping Codes within Protein-Coding Sequences*, *Genome Res.* **20**, 1582 (2010).
- [23] W. R. Whitaker, H. Lee, A. P. Arkin, and J. E. Dueber, *Avoidance of Truncated Proteins from Unintended Ribosome Binding Sites within Heterologous Protein Coding Sequences*, *ACS Synth. Biol.* **4**, 249 (2015).
- [24] M. Gelfand and E. Koonin, *Avoidance of Palindromic Words in Bacterial and Archaeal Genomes: A Close Connection with Restriction Enzymes*, *Nucleic Acids Res.* **25**, 2430 (1997).
- [25] E. Rocha, A. Danchin, and A. Viari, *Evolutionary Role of Restriction/Modification Systems as Revealed by Comparative Genome Analysis*, *Genome Res.* **11**, 946 (2001).
- [26] L. Qian and E. Kussell, *Evolutionary Dynamics of Restriction Site Avoidance*, *Phys. Rev. Lett.* **108**, 158105 (2012).
- [27] M. Pleška, L. Qian, R. Okura, T. Bergmiller, Y. Wakamoto, E. Kussell, and C. C. Guet, *Bacterial Autoimmunity due to a Restriction-Modification System*, *Curr. Biol.* **26**, 404 (2016).
- [28] K. Robison, A. M. McGuire, and G. M. Church, *A Comprehensive Library of DNA-Binding Site Matrices for 55 Proteins Applied to the Complete Escherichia Coli K-12 Genome*, *J. Mol. Biol.* **284**, 241 (1998).
- [29] G. D. Stormo and Y. Zhao, *Determining the Specificity of Protein-DNA Interactions*, *Nat. Rev. Genet.* **11**, 751 (2010).
- [30] M. F. Berger, A. A. Philippakis, A. M. Qureshi, F. S. He, P. W. Estep III, and M. L. Bulyk, *Compact, Universal DNA Microarrays to Comprehensively Determine Transcription-Factor Binding Site Specificities*, *Nat. Biotechnol.* **24**, 1429 (2006).
- [31] M. A. Hume, L. A. Barrera, S. S. Gisselbrecht, and M. L. Bulyk, *UniPROBE, Update 2015: New Tools and Content for the Online Database of Protein-Binding Microarray Data on Protein-DNA Interactions*, *Nucleic Acids Res.* **43**, D117 (2015).
- [32] See Supplemental Material at <http://link.aps.org/supplemental/10.1103/PhysRevX.6.041009> for methods, text, Figs. S1–S27, Tables S1–S6, and the data file.
- [33] G. D. Stormo, *DNA Binding Sites: Representation and Discovery*, *Bioinformatics* **16**, 16 (2000).
- [34] A. Stewart and J. B. Plotkin, *The Evolution of Complex Gene Regulation by Low-Specificity Binding Sites*, *Proc. R. Soc. B* **280**, 20131313 (2013).
- [35] M. Lynch, *Rate, Molecular Spectrum, and Consequences of Human Mutation*, *Proc. Natl. Acad. Sci. U.S.A.* **107**, 961 (2010).
- [36] E. P. C. Rocha and E. J. Feil, *Mutational Patterns Cannot Explain Genome Composition: Are There Any Neutral Sites in the Genomes of Bacteria?*, *PLoS Genet.* **6**, e1001104 (2010).
- [37] A. Hodgkinson and A. Eyre-Walker, *Variation in the Mutation Rate across Mammalian Genomes*, *Nat. Rev. Genet.* **12**, 756 (2011).
- [38] C. Burge, A. M. Campbell, and S. Karlin, *Over- and Under-Representation of Short Oligonucleotides in DNA Sequences*, *Proc. Natl. Acad. Sci. U.S.A.* **89**, 1358 (1992).
- [39] T. Abe, H. Sugawara, S. Kanaya, M. Knouchi, and T. Ikemura, *Self-Organizing Map (SOM) Unveils and Visualizes Hidden Sequence Characteristics of a Wide Range of Eukaryote Genomes*, *Gene* **365**, 27 (2006).
- [40] S. L. Chen, W. Lee, A. K. Hottes, L. Shapiro, and H. H. McAdams, *Codon Usage between Genomes is Constrained*

- by *Genome-Wide Mutational Processes*, *Proc. Natl. Acad. Sci. U.S.A.* **101**, 3480 (2004).
- [41] R. Horn and C. Johnson, *Matrix Analysis* (Cambridge University Press, Cambridge, England, 1991).
- [42] We note that our model ignores the local, overlapping structure of words within a genome, effectively representing the genome as a “bag of words.” This approximation is fully justified because the operation of shifting a k -mer window by one position along a DNA sequence results typically in a jump to a very different word. The only exceptions occur within a very small subset of loci that contain short sequence repeats of size $\geq k$.
- [43] We denote by S^l the average fitness reduction due to interference effects of nonfunctional binding at any single locus on DNA, including interference with DNA replication and transcriptional processes, and by S_j^f the average fitness reduction due to titration of a single copy of protein j away from any of its functional sites. In these terms, the selective coefficients can be written as $\Delta s_i = \sum_j (S^l + S_j^f) q_{ij}$, where q_{ij} is the probability that a given locus with word i is bound by the j th DNA binding protein, and $q_i \equiv \sum_j q_{ij}$ is the probability that a word i is bound by any protein. The quantity q_{ij} involves partition sums that depend on the genome’s word frequency vector $\mathbf{f}$, the concentrations of the DNA binding proteins, the genome size L , and the set of binding energies E_{ij} for protein j and word i . Such complex dependencies preclude direct computation from first principles, and we present them here only to delineate the molecular and physiological underpinnings of global selective coefficients.
- [44] G. Sella, D. A. Petrov, M. Przeworski, and P. Andolfatto, *Pervasive Natural Selection in the Drosophila Genome?*, *PLoS Genet.* **5**, e1000495 (2009).
- [45] B. Charlesworth, *Effective Population Size and Patterns of Molecular Evolution and Variation*, *Nat. Rev. Genet.* **10**, 195 (2009).
- [46] The 1000 Genomes Project Consortium, *A Global Reference for Human Genetic Variation*, *Nature (London)* **526**, 68 (2015).
- [47] J. C. Roach, G. Glusman, A. F. Smit, C. D. Huff, R. Hubley, P. T. Shannon, L. Rowen, K. P. Pant, N. Goodman, M. Bamshad, J. Shendure, R. Drmanac, L. B. Jorde, L. Hood, and D. J. Galas, *Analysis of Genetic Inheritance in a Family Quartet by Whole-Genome Sequencing*, *Science* **328**, 636 (2010).
- [48] P. D. Dixit, T. Y. Pang, F. W. Studier, and S. Maslov, *Recombinant Transfer in the Basic Genome of Escherichia coli*, *Proc. Natl. Acad. Sci. U.S.A.* **112**, 9070 (2015).
- [49] H. Lee, E. Popodi, H. Tang, and P. L. Foster, *Rate and Molecular Spectrum of Spontaneous Mutations in the Bacterium Escherichia coli as Determined by Whole-Genome Sequencing*, *Proc. Natl. Acad. Sci. U.S.A.* **109**, E2774 (2012).
- [50] H. J. Muller, *The Relation of Recombination to Mutational Advance*, *Mutat. Res.* **1**, 2 (1964).
- [51] J. Felsenstein, *The Evolutionary Advantage of Recombination*, *Genetics* **78**, 737 (1974).
- [52] L. Peliti, *Introduction to the Statistical Theory of Darwinian Evolution*, [arXiv:cond-mat/9712027](https://arxiv.org/abs/cond-mat/9712027).
- [53] B. Charlesworth, A. Betancourt, V. Kaiser, and I. Gordo, *Genetic Recombination and Molecular Evolution*, *Cold Spring Harbor Symp. Quant. Biol.* **74**, 177 (2009).
- [54] N. H. Barton, *Mutation and the Evolution of Recombination*, *Phil. Trans. R. Soc. B* **365**, 1281 (2010).
- [55] M. I. Jensen-Seaman, T. S. Furey, B. A. Payseur, Y. Lu, K. M. Roskin, C.-F. Chen, M. A. Thomas, D. Haussler, and H. J. Jacob, *Comparative Recombination Rates in the Rat, Mouse, and Human Genomes*, *Genome Res.* **14**, 528 (2004).
- [56] M. Vos and X. Didelot, *A Comparison of Homologous Recombination Rates in Bacteria and Archaea*, *ISME J.* **3**, 199 (2009).
- [57] D. Tautz, M. Trick, and G. A. Dover, *Cryptic Simplicity in DNA is a Major Source of Genetic Variation*, *Nature (London)* **322**, 652 (1986).
- [58] P. D. Keightley, U. Trivedi, M. Thomson, F. Oliver, S. Kumar, and M. L. Blaxter, *Analysis of the Genome Sequences of Three Drosophila melanogaster Spontaneous Mutation Accumulation Lines*, *Genome Res.* **19**, 1195 (2009).
- [59] S. B. Montgomery *et al.*, *The Origin, Evolution, and Functional Impact of Short Insertion-Deletion Variants Identified in 179 Human Genomes*, *Genome Res.* **23**, 749 (2013).
- [60] Y. Zhu, M. L. Siegal, D. W. Hall, and D. A. Petrov, *Precise Estimates of Mutation Rate and Spectrum in Yeast*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, E2310 (2014).
- [61] D. L. Hartl, E. N. Moriyama, and S. A. Sawyer, *Selection Intensity for Codon Bias*, *Genetics* **138**, 227 (1994).
- [62] R. Hershberg and D. A. Petrov, *Selection on Codon Bias*, *Annu. Rev. Genet.* **42**, 287 (2008).
- [63] F. Hildebrand, A. Meyer, and A. Eyre-Walker, *Evidence of Selection upon Genomic GC-Content in Bacteria*, *PLoS Genet.* **6**, e1001107 (2010).
- [64] A. Tats, T. Tenson, and M. Remm, *Preferred and Avoided Codon Pairs in Three Domains of Life*, *BMC Genomics* **9**, 463 (2008).
- [65] G. R. Moura, M. Pinheiro, A. Freitas, J. L. Oliveira, J. C. Frommlet, L. Carreto, A. R. Soares, A. R. Bezerra, and M. A. S. Santos, *Species-Specific Codon Context Rules Unveil Non-neutrality Effects of Synonymous Mutations*, *PLoS One* **6**, e26817 (2011).
- [66] A. Campbell, J. Mrazek, and S. Karlin, *Genome Signature Comparisons among Prokaryote, Plasmid, and Mitochondrial DNA*, *Proc. Natl. Acad. Sci. U.S.A.* **96**, 9184 (1999).
- [67] J. Raymond and E. B. Alsop, *Resolving Prokaryotic Taxonomy with rRNA: Longer Oligonucleotide Word Lengths Improve Genome and Metagenome Taxonomic Classification*, *PLoS One* **8**, e67337 (2013).
- [68] K. Vetsigian and N. Goldenfeld, *Genome Rhetoric and the Emergence of Compositional Bias*, *Proc. Natl. Acad. Sci. U.S.A.* **106**, 215 (2009).
- [69] R. Hershberg and D. A. Petrov, *Evidence that Mutation Is Universally Biased towards AT in Bacteria*, *PLoS Genet.* **6**, e1001115 (2010).
- [70] M. Ackermann and L. Chao, *DNA Sequences Shaped by Selection for Stability*, *PLoS Genet.* **2**, e22 (2006).
- [71] W.-H. Lin and E. Kussell, *Evolutionary Pressures on Simple Sequence Repeats in Prokaryotic Coding Regions*, *Nucleic Acids Res.* **40**, 2399 (2012).

- [72] L. Katz and C. B. Burge, *Widespread Selection for Local RNA Secondary Structure in Coding Regions of Bacterial Genes*, *Genome Res.* **13**, 2042 (2003).
- [73] J. J. Forman, A. Legesse-Miller, and H. A. Collier, *A Search for Conserved Sequences in Coding Regions Reveals That the let-7 microRNA Targets Dicer within Its Coding Sequence*, *Proc. Natl. Acad. Sci. U.S.A.* **105**, 14879 (2008).
- [74] B. Greenbaum, A. J. Levine, G. Bhanot, and R. Rabadan, *Patterns of Evolution and Host Gene Mimicry in Influenza and Other RNA Viruses*, *PLoS Pathogens* **4**, e1000079 (2008).
- [75] B. Greenbaum, S. Cocco, A. J. Levine, and R. Monasson, *Quantitative Theory of Entropic Forces Acting on Constrained Nucleotide Sequences Applied to Viruses*, *Proc. Natl. Acad. Sci. U.S.A.* **111**, 5054 (2014).
- [76] B. Greenbaum, R. Rabadan, and A. Levine, *Patterns of Oligonucleotide Sequences in Viral and Host Cell RNA Identify Mediators of the Host Innate Immune System*, *PLoS One* **4**, e5969 (2009).
- [77] J. W. Drake, B. Charlesworth, D. Charlesworth, and J. F. Crow, *Rates of Spontaneous Mutation*, *Genetics* **148**, 1667 (1998).
- [78] K. R. Nitta, A. Jolma, Y. Yin, E. Morgunova, T. Kivioja, J. Akhtar, K. Hens, J. Toivonen, B. Deplancke, E. E. M. Furlong, and J. Taipale, *Conservation of Transcription Factor Binding Specificities across 600 Million Years of Bilateria Evolution*, *eLife* **4**, e04837 (2015).
- [79] C. M. Fraser *et al.*, *The Minimal Gene Complement of Mycoplasma genitalium*, *Science* **270**, 397 (1995).
- [80] C. L. Frank, F. Liu, R. Wijayatunge, L. Song, M. T. Biegler, M. G. Yang, C. M. Vockley, A. Safi, C. A. Gersbach, G. E. Crawford, and A. E. West, *Regulation of Chromatin Accessibility and Zic Binding at Enhancers in the Developing Cerebellum*, *Nat. Neurosci.* **18**, 647 (2015).
- [81] P. Termolino, G. Cremona, M. F. Consiglio, and C. Conicella, *Insights into Epigenetic Landscape of Recombination-Free Regions*, *Chromosoma* **125**, 301 (2016).
- [82] P. D. Sniegowski, P. J. Gerrish, T. Johnson, and A. Shaver, *The Evolution of Mutation Rates: Separating Causes from Consequences*, *BioEssays* **22**, 1057 (2000).
- [83] See <ftp.ncbi.nlm.nih.gov/genomes/refseq>. The list of genomes for bacteria and archaea is found in the directory /GENOME_REPORTS/prokaryotes.txt, and for eukaryotes in the directory /GENOME_REPORTS/eukaryotes.txt. Bacterial plasmid sequences are not used.