PLOS | COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Nonconsensus Protein Binding to Repetitive DNA Sequence Elements Significantly Affects Eukaryotic Genomes

Ariel Afek[1☯], Hila Cohen[1☯], Shiran Barber-Zucker[1], Raluca Gordân[2], David B. Lukatsky[1]*

1 Department of Chemistry, Ben-Gurion University of the Negev, Beer-Sheva, Israel, **2** Center for Genomic and Computational Biology, Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, United States of America

☯ These authors contributed equally to this work.
* lukatsky@bgu.ac.il

Citation: Afek A, Cohen H, Barber-Zucker S, Gordân R, Lukatsky DB (2015) Nonconsensus Protein Binding to Repetitive DNA Sequence Elements Significantly Affects Eukaryotic Genomes. PLoS Comput Biol 11(8): e1004429. doi:10.1371/journal.pcbi.1004429

Editor: Ilya Ioshikhes, Ottawa University, CANADA

Received: September 29, 2014

Accepted: June 30, 2015

Published: August 18, 2015

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: DBL acknowledges the financial support from the Israel Science Foundation (ISF) grant 1014/09. RG acknowledges financial support form the Alfred P. Sloan Foundation. AA is supported by the Adams Fellowship program of the Israel National Academy of Science. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

Recent genome-wide experiments in different eukaryotic genomes provide an unprecedented view of transcription factor (TF) binding locations and of nucleosome occupancy. These experiments revealed that a large fraction of TF binding events occur in regions where only a small number of specific TF binding sites (TFBSs) have been detected. Furthermore, *in vitro* protein-DNA binding measurements performed for hundreds of TFs indicate that TFs are bound with wide range of affinities to different DNA sequences that lack known consensus motifs. These observations have thus challenged the classical picture of specific protein-DNA binding and strongly suggest the existence of additional recognition mechanisms that affect protein-DNA binding preferences. We have previously demonstrated that repetitive DNA sequence elements characterized by certain symmetries statistically affect protein-DNA binding preferences. We call this binding mechanism *nonconsensus protein-DNA binding* in order to emphasize the point that specific consensus TFBSs do not contribute to this effect. In this paper, using the simple statistical mechanics model developed previously, we calculate the nonconsensus protein-DNA binding free energy for the entire *C*. *elegans* and *D*. *melanogaster* genomes. Using the available chromatin immunoprecipitation followed by sequencing (ChIP-seq) results on TF-DNA binding preferences for ~100 TFs, we show that DNA sequences characterized by low predicted free energy of nonconsensus binding have statistically higher experimental TF occupancy and lower nucleosome occupancy than sequences characterized by high free energy of nonconsensus binding. This is in agreement with our previous analysis performed for the yeast genome. We suggest therefore that nonconsensus protein-DNA binding assists the formation of nucleosome-free regions, as TFs outcompete nucleosomes at genomic locations with enhanced nonconsensus binding. In addition, here we perform a new, large-scale analysis using *in vitro* TF-DNA preferences obtained from the universal protein binding microarrays (PBM) for ~90 eukaryotic TFs belonging to 22 different DNA-binding domain types. As a result of this new analysis, we conclude that nonconsensus protein-DNA binding is a widespread phenomenon that significantly affects protein-DNA binding preferences and need not require the presence of consensus (specific) TFBSs in order to achieve genome-wide TF-DNA binding specificity.

Author Summary

Interactions between proteins and DNA trigger many important biological processes. Therefore, to fully understand how the information encoded on the DNA transcribes into RNA, which in turn translates into proteins in the cell, we need to unravel the molecular design principles of protein-DNA interactions. It is known that many interactions occur when a protein is attracted to a specific short segment on the DNA called a specific protein-DNA binding motif. Strikingly, recent experiments revealed that many regulatory proteins reproducibly bind to different regions on the DNA lacking such specific motifs. This suggests that fundamental molecular mechanisms responsible for protein-DNA recognition specificity are not fully understood. Here, using high-throughput protein-DNA binding data obtained by two entirely different methods for ~100 TFs in each case, we show that DNA regions possessing certain repetitive sequence elements exert the statistical attractive potential on DNA-binding proteins, and as a result, such DNA regions are enriched in bound proteins. This is in agreement with our previous analysis performed for the yeast genome. We use the term *nonconsensus protein-DNA binding* in order to describe protein-DNA interactions that occur in the absence of specific protein-DNA binding motifs. Here we demonstrate that the identified nonconsensus effect is highly significant for a variety of organismal genomes and it affects protein-DNA binding preferences and nucleosome occupancy at the genome-wide level.

Introduction

Binding of TFs to their target sites on the DNA is a key step during gene activation and repression. An existing paradigm assumes that the main mechanism responsible for specific TF-DNA recognition is TF binding to short (typically 6–20 bp long) DNA sequences called *specific consensus motifs*, or *specific TF binding sites* (TFBSs). It has been known for a long time, since the seminal studies of Iyer and Struhl [1], that genomic context surrounding specific TFBSs significantly influences TF-DNA binding preferences. However, general rules describing the mechanisms responsible for such influences remain unknown.

Recently, the model organism ENCODE (modENCODE) project has revealed genome-wide comprehensive maps of TF-DNA binding and nucleosome occupancy in *C. elegans* [2–7] and in *D. melanogaster* [8–10]. Remarkably, these studies have challenged the existing paradigm and revealed that a large fraction of TF-DNA binding events occurs in genomic regions depleted of specific consensus motifs. Such genomic regions with enhanced overall TF-DNA binding but depleted in consensus motifs are oftentimes of low sequence complexity, which means that they are enriched in repeated DNA sequences.

We have recently proposed that repetitive DNA sequences characterized by certain symmetries and length scales of repetitive sequence patterns (see below) exert a statistical potential on DNA-binding proteins, affecting their binding preferences [11–15]. This effect of protein binding to repetitive DNA sequences in the absence of specific base-pair recognition is different from the concept of nonspecific protein-DNA binding introduced and explored in seminal studies of von Hippel, Berg, et al. [16–21]. In particular, von Hippel and Berg defined two related mechanisms for nonspecific protein-DNA binding [19]. The first mechanism is DNA sequence-independent, and it assumes that DNA exerts an electrostatic attraction upon DNA-binding proteins, modulated by the overall DNA geometry [19]. It has been proposed that DNA-binding proteins use different conformations in specific and nonspecific binding modes [16–20, 22]. The second mechanism assumes that mutated specific DNA consensus motifs retain a reduced binding affinity for sequence-specific TFs [19]. Nonspecific protein-DNA binding might become significant since the statistical probability to find such imperfect motifs in many genomic locations by random chance is high for eukaryotic genomes [19, 23]. The importance of nonspecific protein-DNA binding has been experimentally demonstrated for a number of systems both *in vivo* [24, 25] and *in vitro* [26–31].

We demonstrated recently that repetitive DNA sequence patterns characterized by certain symmetries lead to *nonconsensus protein-DNA binding* that can be enhanced or reduced depending on the symmetry type [11]. We use the term *nonconsensus protein-DNA binding* in order to emphasize the point that the nonconsensus protein-DNA binding free energy is computed without using any experimental information on specific protein-DNA binding preferences (see below). For example, we showed that repetitive homo-oligonucleotide sequence patterns, such as repeated poly(A)/poly(T)/poly(C)/poly(G) tracts lead to statistically enhanced nonconsensus protein-DNA binding affinity [11]. Our results indicated that such nonconsensus binding significantly influences nucleosome occupancy [12], TF-DNA binding preferences [13], and transcription pre-initiation complex binding preferences [14] in yeast.

In addition, using the protein binding microarray (PBM) method, we have recently directly measured the nonconsensus protein-DNA binding free energy for several human TFs [15]. We have demonstrated that, remarkably, the magnitude of the identified nonconsensus effect reaches as much as 66% of consensus (specific) binding [15].

In this study we explore the extent and significance of the nonconsensus protein-DNA binding mechanism for a large number of proteins belonging to different structural families. First, we investigate the nonconsensus effect in more complex, multicellular organisms, using the available ChIP-seq data obtained for ~100 TFs in *C. elegans* [2, 3] and *D. melanogaster* [10, 32]. Next, we perform the analysis of high-resolution *in vitro* universal protein-DNA binding microarray (PBM) data obtained for ~90 eukaryotic TFs belonging to 22 different DNA-binding domain types [33–35]. In addition, we identify protein sequence features that statistically distinguish between proteins with stronger and weaker response to nonconsensus repetitive DNA sequence elements, respectively.

We stress the point that *in vitro* analysis is free of confounding factors present in a cell, such as nucleosomes and indirect TF-DNA binding. Our previous experimental *in vitro* study of nonconsensus protein-DNA binding was performed for only 6 TFs [15]. The present analysis of the vast amount of *in vitro* TF-DNA binding data extends this number to more than an order of magnitude, suggesting that the nonconsensus mechanism most likely represents the statistical law rather than the exception. Therefore, the results reported here strongly support our conclusion that nonconsensus protein-DNA binding is a widespread phenomenon that significantly affects protein-DNA binding preferences in eukaryotic genomes, and need not require the presence of consensus (specific) TFBSs in order to achieve genome-wide TF-DNA binding specificity.

Results

Nonconsensus free energy correlates with *C. elegans* and *D. melanogaster* TF-DNA binding preferences

We compared the predicted landscape of *nonconsensus protein-DNA binding free energy* with the genomic binding profiles of 69 transcriptional regulators in *C. elegans* [2, 3] and 30 transcriptional regulators in *D. melanogaster* [10, 32], as determined by ChIP-seq in the modENCODE project [2, 3, 8, 10]. We computed the nonconsensus binding free energy landscape using a simple approach that we developed previously [11]. Briefly, we used a set of random protein-DNA binders as a proxy for nonspecific protein-DNA interactions in a crowded cellular environment (Fig 1). Next, to each location along the *C. elegans* and *D. melanogaster* genomes, we assigned an average *free energy of nonconsensus protein-DNA binding*, $\langle F\rangle_{TF}$, where the averaging is performed over an ensemble of random binders (see Methods for further details). The free energy value at each sequence location is entropy-dominated, and it is influenced exclusively by the presence of repetitive DNA sequence patterns [11] surrounding that location. We use the term *DNA sequence correlations* to describe the repetitive DNA patterns, and the term *correlation scale* to describe the length of the patterns (Methods). The larger the correlation scale, the larger the number of repetitive sequence patterns, and thus the stronger the nonconsensus protein-DNA binding effect [11]. Importantly, the genomic DNA sequence constitutes the only input for the nonconsensus binding model, i.e. the model does not have any fitting parameters (Methods).

We found that the nonconsensus protein-DNA binding free energy correlates negatively with the combined TF occupancy in both the *C. elegans* and the *D. melanogaster* genomes, i.e. the lower the nonconsensus binding free energy, the higher the combined TF occupancy (Fig 2). Fig 2a and 2c illustrate this correlation for free energy profiles, $\langle\langle F\rangle_{TF}\rangle_{seq}$, averaged over genomic sequences aligned with respect to the TSS. A statistically significant correlation at the single gene level is also observed, on average, without sequence alignment with respect to the TSS (Fig 2b and 2d). In these analyses both genomes show statistically significant negative correlations, with the correlation being more pronounced in *C. elegans*.

We verified that the predicted free energy landscape is qualitatively robust with respect to variations in the model parameters (i.e. the sliding window width, L, and the TFBS size, M) (S1 Fig). In addition, we validated that the predicted free energy landscape is determined by the presence of repetitive sequence patterns, and *not* by the average genomic nucleotide content. To show this, we shuffled the DNA sequence in each sliding window along the genome to obtain random DNA sequences with a fixed nucleotide content, and we computed the normalized free energy, $\delta F = F - F_{\mathrm{rand}}$, where F_{rand} is the free energy of the random, shuffled sequences, averaged over different random realizations (Methods). As shown in S2 Fig, the normalized free energy δF is robust with respect to global variations in the genomic nucleotide content.

The predicted reduction in the nonconsensus free energy upstream of TSSs (Fig 2a and 2c) stems from the enhanced level of homo-oligonucleotide sequence correlations (i.e. repetitive homo-oligonucleotide sequence patterns, such as repeated poly(dA:dT) tracts). This effect can be intuitively understood in the following way. As shown in our previous work, the presence of enhanced homo-oligonucleotide sequence correlations within a DNA region generally leads to the widening of the protein-DNA binding energy spectrum in this region [11]. For example, in the statistical ensemble of random binders interacting with DNA sequence that contains long homo-oligonucleotide tracts with two alternating types of nucleotides (such as alternating poly (dA:dT) and poly(dT:dA) tracts), the width (i.e. the standard deviation) of the binding energy spectrum, σ_U^{homo}, will be universally larger than the corresponding width for the case of entirely random DNA sequence, $\sigma_U^{\mathrm{homo}} \simeq \sqrt{2}\cdot\sigma_U^{\mathrm{random}}$ [11]. This result is independent of the microscopic details of the protein-DNA interaction potential, U, and it is simply the consequence of the central limit theorem [36, 37]. The wider energy spectrum, $\sigma_U^{\mathrm{homo}} > \sigma_U^{\mathrm{random}}$, universally leads to the statistically lower free energy, $F^{\mathrm{homo}} < F^{\mathrm{random}}$ [38], and therefore to a higher nonconsensus protein-DNA binding affinity. The computed probability distributions of the nonconsensus protein-DNA binding energy and the free energy in the *C. elegans* genome, further

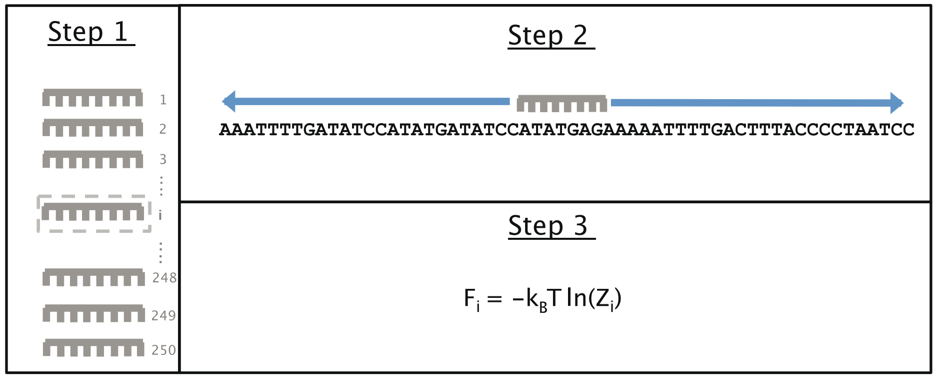

Fig 1. Cartoon illustrating our model for computing the free energy of nonconsensus protein-DNA binding. Schematic representation of the procedure for computing the nonconsensus free energy. Step 1: In order to model nonspecific TF-DNA binding, we generate an ensemble of 250 *random* TFs. Step 2: Each TF moves within a sliding window of width L bp. The TF-DNA binding energy is computed at each location of TF along the sliding window using the random potential. Step 3: For each TF we calculate the TF-DNA binding free energy. We repeat this process for all random TFs and compute the *average* nonconsensus binding free energy with respect to this ensemble of random TFs. Moving the sliding window along the genome, we assign the nonconsensus TF-DNA binding free energy at each genomic location. We assume that each random binder makes contacts with M bps upon DNA binding. For each model TF (random binder), we define the partition function of protein-DNA binding within the chosen sliding window of width L bp. We used $L = 50$ bp (i.e. the sliding window size) in our calculations of the genome-wide nonconsensus TF-DNA binding free energy profiles, and $L = 36$ bp for calculations of the nonconsensus TF-DNA free energies for *in vitro* protein binding microarray (PBM) experiments. In the latter case, we do not move the sliding window since each DNA sequence in the PBM library is 36-bp long.

doi:10.1371/journal.pcbi.1004429.g001

illustrates this mechanism (S3 Fig). Thus, the nonconsensus protein-DNA binding mechanism can significantly influence TF-DNA binding preferences in the *C. elegans* and *D. melanogaster* genomes, complementing the conventional, specific protein-DNA recognition mode.

We stress the fact that the minimum of the *average* nonconsensus protein-DNA binding free energy landscape does not align precisely with the maximum of the average TF occupancy profile in both *C. Elegans* and *D. melanogaster* genomes (Fig 2a and 2c). Such mismatch is also observed between the average nonconsensus protein-DNA binding free energy landscape and the average nucleosome profile (see below, Fig 3a and 3c), similar to the case as we previously observed for the yeast genome [12]. Combination of additional factors not taken into account in our model but present *in vivo* might explain a possible origin of such a mismatch. These factors include, first, steric constrains imposed by the presence of nucleosome particles [39]; second, steric constrains imposed by the transcription pre-initiation complex (PIC) [40]; and third, the presence of specific TFBSs [41].

Nonconsensus protein-DNA binding influences nucleosome preferences

We also assessed the effect of nonconsensus protein-DNA binding on nucleosome binding preferences in the *C. elegans* and *D. melanogaster* genomes. Genome-wide measurements of nucleosome occupancy show a typical nucleosome depleted region upstream of the TSSs, and a well-positioned +1 nucleosome [2, 4, 42]. In *D. melanogaster*, an oscillating nucleosome occupancy pattern was observed, similar to the one in yeast [43], while the *C. elegans* genome-wide nucleosome occupancy profile does not demonstrate such strong oscillations [4, 42].

The computed nonconsensus free energy landscapes show a statistically high, positive correlation with the nucleosome occupancy profile in both genomes (Fig 3). In particular, the

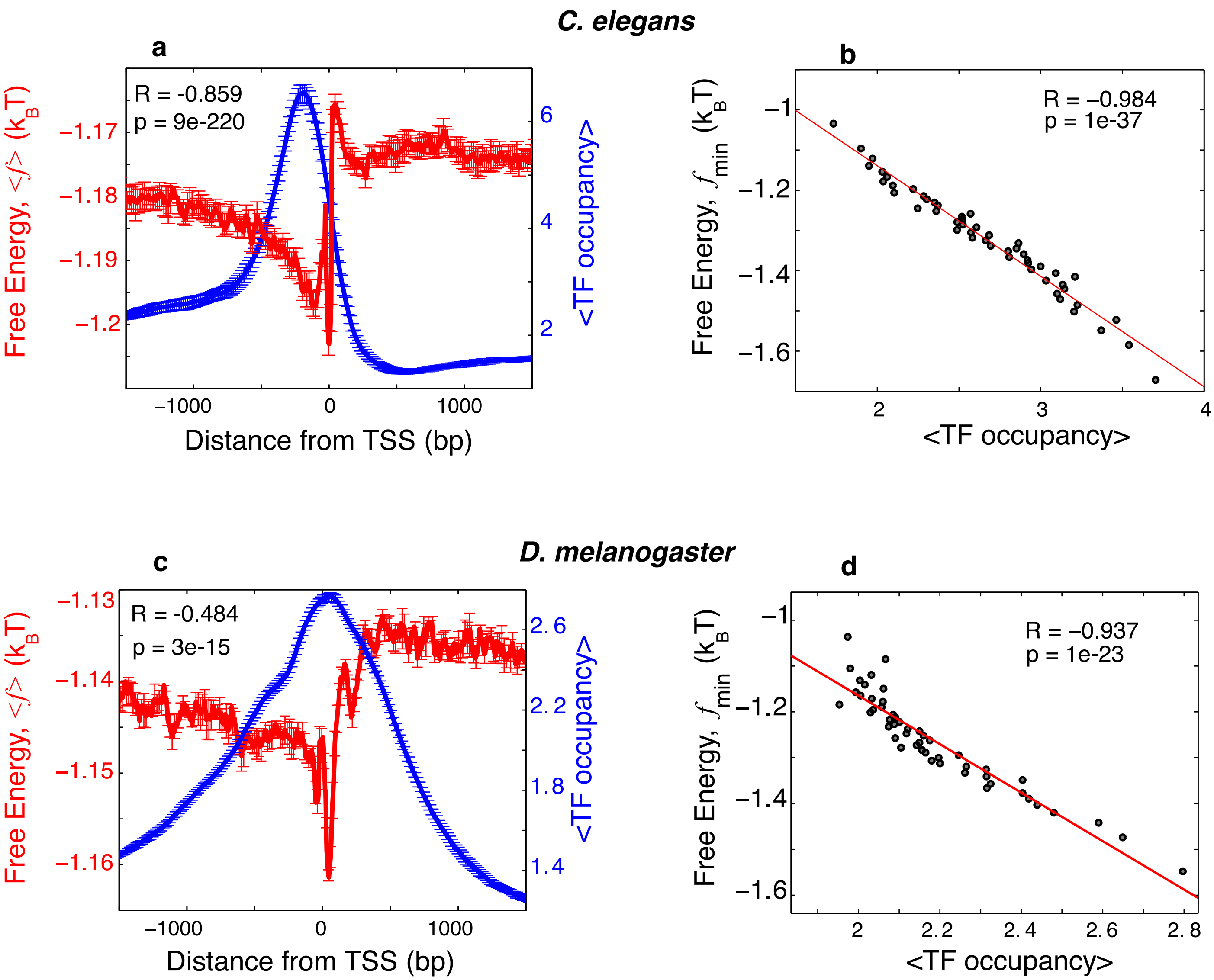

Fig 2. The free energy of nonconsensus TF-DNA binding negatively correlates with the combined TF occupancy for both *C. elegans* and *D. melanogaster* genomes. (**a**) The average *free energy of nonconsensus TF-DNA binding* per bp, $\langle f \rangle = \langle\langle F \rangle_{TF} \rangle_{seq}/M$ (red), and the average, combined occupancy profile of 69 *C. elegans* TFs (blue), plotted around the TSSs of 17,207 *C. elegans* coding genes. The notation <TF occupancy> describes the average, combined occupancy profile of all 69 TFs. The linear correlation coefficient is computed for a linear fit of $\langle f \rangle$ versus <TF occupancy> at individual genomic locations, computed every 4 bp, within the interval (-1500,1500) around the TSS. The sequences are aligned with respect to the TSS. In order to compute error bars, we divided genes into ten randomly chosen subgroups, and computed $\langle f \rangle$ for each subgroup. The error bars are defined as one standard deviation of $\langle f \rangle$ between the subgroups. The error bars for the combined TF occupancy are computed analogously. (**b**) Correlation between the minimum value of the free energy of nonconsensus TF-DNA binding, $f_{min} = \min(f)$, and the combined occupancy of all TFs, computed for individual genes in non-overlapping windows of 100 bp, within the entire interval (-1000,1000). The data was grouped into 50 bins. (**c**) Similar to (a) but showing the average free energy of nonconsensus TF-DNA binding per bp, $\langle f \rangle$ (red), and the average transcription factor occupancy (blue), around the TSSs of 12,188 *D. melanogaster* genes. (**d**) Similar to (b) but for 12,188 *D. melanogaster* genes.

doi:10.1371/journal.pcbi.1004429.g002

average nonconsensus free energy shows a pronounced minimum in the upstream nucleosome depleted region (Fig 3a and 3c), similar to the one observed in yeast [12]. In Fig 3b and 3d we also observed, at the single gene level, statistically significant correlation between the average nucleosome occupancy and the average free energy of nonconsensus binding (Methods). Sequences with lower nonconsensus protein-DNA binding free energy have, on average, lower nucleosome occupancy.

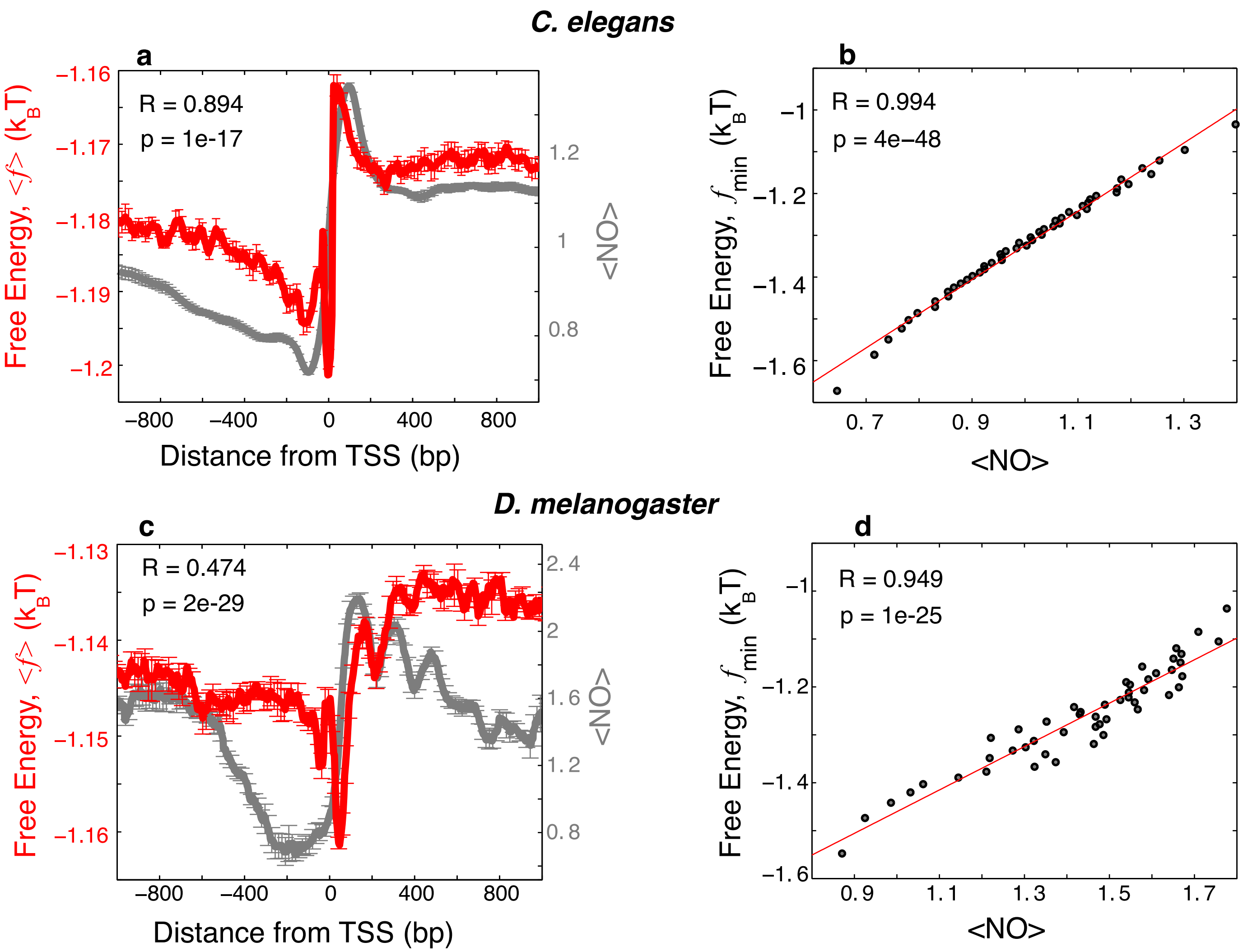

Fig 3. The free energy of nonconsensus TF-DNA binding positively correlates with the nucleosome occupancy. (**a**) The average free energy of nonconsensus TF-DNA binding per bp, $\langle f \rangle = \langle \langle F \rangle_{TF} \rangle_{seq} / M$ (red), and the average nucleosome occupancy from [4] (gray), around the TSSs of 23,287 mRNA coding and non-coding *C. elegans* genes. The linear correlation coefficient is computed for a linear fit of $\langle f \rangle$ versus the average nucleosome occupancy at individual genomic locations, computed every 4 bp, within the interval (-1000,1000). In order to compute error bars, we divided genes into five randomly chosen subgroups, and computed $\langle f \rangle$ for each subgroup. The error bars are defined as one standard deviation of $\langle f \rangle$ between the subgroups. (**b**) Correlation between the minimal value of the free energy of nonconsensus TF-DNA binding, $f_{min} = \min(f)$, and the nucleosome occupancy, computed for individual genes in non-overlapping windows of 100 bp within the interval (-1000,1000) around the TSS for each of the 23,287 genes. The data was grouped into 50 bins. (**c**) Similar to (a) but showing the average free energy of nonconsensus TF-DNA binding per bp, $\langle f \rangle$ (red), and the average H2A.Z nucleosome occupancy (grey) around the TSSs of 12,188 *D. melanogaster* genes [32]. (**d**) Similar to (b) but for 12,188 *D. melanogaster* genes.

doi:10.1371/journal.pcbi.1004429.g003

We suggest that the observed effect stems from the competition between TFs that experience enhanced nonspecific attraction towards upstream promoter regions (i.e., reduced level of the nonconsensus free energy) and nucleosome-forming histones. It is important to stress that the presence of repetitive DNA sequence elements in promoter regions might also affect histone-DNA binding due to the nonconsensus mechanism, and as a result of it, the nucleosome formation. How exactly individual histones and histone complexes respond to different repetitive DNA sequence patterns remains an open question. This issue is further complicated by the fact that several additional mechanisms influence histone-DNA binding in promoter regions. Namely, genome-wide, *in vitro* nucleosome reconstruction experiments demonstrate that nucleosome-free regions (NFR) can be formed to some extend even in the mixture of purified genomic DNA with histones [44, 45]. However, intrinsic DNA sequence preferences of nucleosomes still remain an open issue [46]. In particular, it has been recently demonstrated that AT-rich sequences present in many NFRs have little effect on the stability of nucleosomes [46]. Rather it appears that ATP-dependent chromatin modifiers constitute a major factor regulating nucleosome-binding preferences *in vivo* [43, 46].

In vitro protein-DNA binding measurements for ~90 TFs to ~45,000 short, non-genomic DNA sequences validate the nonconsensus binding mechanism

Here we provide an additional, highly significant validation for the proposed mechanism of nonconsensus protein-DNA binding by the analysis of the available *in vitro* TF-DNA binding data obtained using the protein-binding microarray (PBM) technology [35, 47–49]. The PBM technology allows to simultaneously measure binding of a TF to tens of thousands of 36-bp long DNA sequences in a single experiment [35]. The PBM method is free from the confounding factors, such as the effect of competing TFs and nucleosomes on TF-DNA binding preferences. Here, we used the currently available 'universal PBM' data for 91 TFs (belonging to 22 distinct DNA-binding domains) from *C. elegans*, *D. melanogaster*, and *mus musculus* [33, 34, 50] (Fig 4 and S1 Table). The DNA libraries used in these 'universal PBM' experiments were designed in such a way that they cover all possible 8-mer DNA sequences [35], giving an unbiased view of TF-DNA binding specificity. Overall, there are ~45,000 distinct DNA sequences in this library, and thus the TF-DNA binding strength was measured for each TF to all these sequences [33, 34, 50].

We computed the nonconsensus TF-DNA binding free energy, $\langle f \rangle_{TF}$, for each 36-bp long DNA sequence in the library using the procedure described above (Fig 1). Contrary to the case of genomic sequences, here we do not move the sliding window along the DNA sequence since each sequence is short, $L = 36$ bp, and therefore a single value of $\langle f \rangle_{TF}$ is assigned to each DNA sequence.

Remarkably, for 69 out of 91 analyzed TFs (i.e. 76%) we detected a statistically significant, negative correlation between the nonconsensus protein-DNA binding free energy and the measured *in vitro* TF-DNA binding intensity. This is in agreement with the results obtained for the *in vivo* TF-DNA binding data (Fig 2b and 2d). Twelve TFs (i.e. 13%) did not show a statistically significant correlation, and interestingly, ten TFs (i.e. 11%) showed an *opposite*, positive correlation (S1 Table). The latter observation is remarkable, since it demonstrates that a non-negligible fraction of TFs can respond to DNA symmetries (represented by our free energy model) in an opposite way compared to the majority of other TFs. However, statistically, the *average* TF-DNA binding preferences show highly significant, negative correlation with the computed free energy of nonconsensus protein-DNA binding (Fig 4) in agreement with the *in vivo* results (Fig 2b and 2d).

In order to identify what structural and sequence features are responsible for the anomalous behavior of these 11% of TFs, we classified all TFs according to the DNA-binding domain (DBD) families they belong to. However, we have not identified any particular DBD families that are unique to those 11% of TFs (S1 Table and S4 Fig). We have also not identified any preference of these TFs with respect to any particular biological function, according to the gene ontology (GO) classification. Therefore, the question what sequence and structural features of proteins are responsible for the positive correlation between the free energy and the experimentally measured *in vitro* TF occupancy remains open.

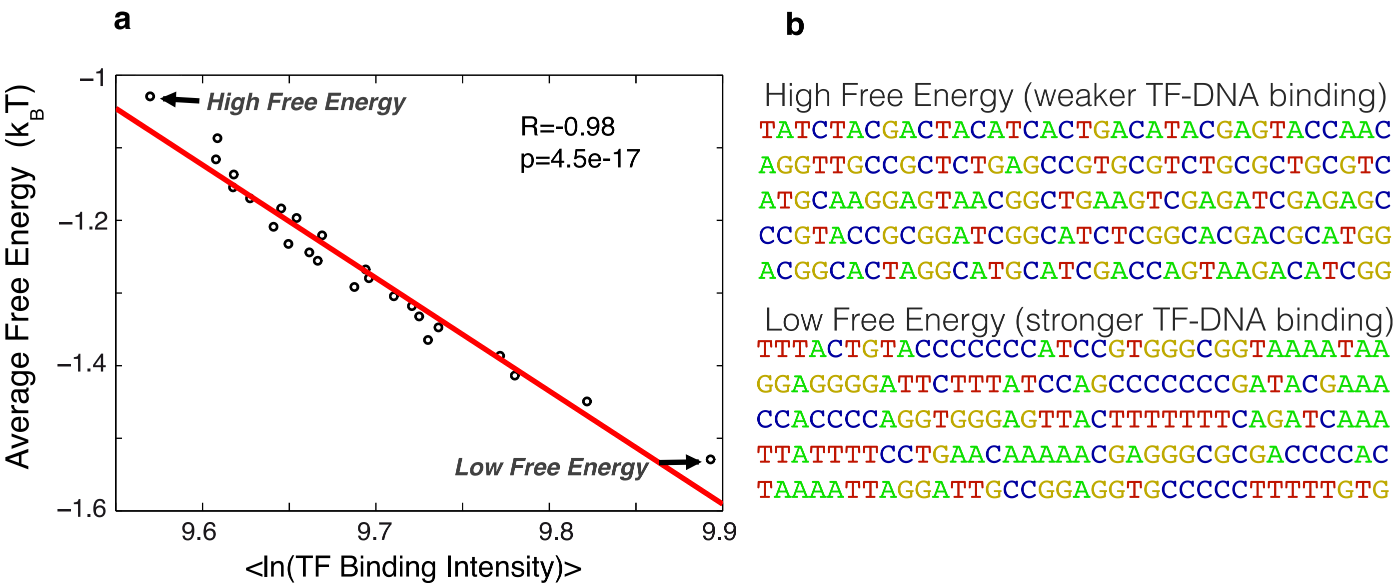

Fig 4. Random-binder model for nonconsensus protein-DNA binding provides a good statistical description of the TF-DNA binding strength measured *in vitro*. (a) Plots show the correlation between the free energy of nonconsensus TF-DNA binding, $\langle f \rangle$, and the measured average TF binding intensity of 82 mouse TFs [34]. We used the data from the PBM experiments performed on universal arrays, which provide measurements of TF binding to all possible 8-bp sequences (8-mers). The data for average TF intensity and free energy was grouped into 25 bins. **(b)** Typical examples of sequences with high and low values of the nonconsensus free energy, respectively, as determined by the *in vitro* PBM measurements.

doi:10.1371/journal.pcbi.1004429.g004

Next, in order to identify protein sequence features that might be responsible for enhanced nonconsensus TF-DNA binding, we separated TFs (we used 82 mouse TFs for this analysis) into two groups. The first group contained 41 TFs with the strongest negative correlation between the free energy and the measured TF occupancy. The second group contained the remaining 41 TFs. We have analyzed the amino acid correlation properties in these two groups of TFs. Our working hypothesis here is that enhanced amino acid sequence correlations in TF sequences are responsible for enhanced nonconsensus TF-DNA binding. We use the term "sequence correlations" in order to describe repetitive sequence patterns. We have previously used a similar analysis in order to investigate protein sequence features responsible for enhanced level of protein structural disorder and protein-protein interaction promiscuity [36]. In particular, we have analyzed the frequency of occurrence of the following repetitive amino acid sequence patterns in each TF group: [*aa*], [*aXa*], [*aXXa*], and [*aXXXa*], where *a* represents each amino acid type and *X* represents an arbitrary amino acid (S2 Table). For example, when we compute the frequency of [Lys-*X*-Lys] pattern, we count the total number of the occurrence of this pattern in each protein sequence, irrespectively to the identity of *X*. As a result of this analysis, we have identified three patterns that demonstrated a statistically significant difference of frequencies between the two TF groups: [Lys-*XX*-Lys] (enriched in the first TF group; Kolmogorov-Smirnov p-value, $p_{ks} \simeq 0.01$), [Arg-Arg] (enriched in the second TF group; $p_{ks} \simeq 0.02$), and [Leu-*X*-Leu] (enriched in the first TF group;$p_{ks} \simeq 0.05$) (S2 Table). In addition the overall compositional fraction of Lys was enriched in the first TF group ($p_{ks} \simeq$ 0.01) (S2 Table). The fact that the most statistically significant enrichment (distinguishing the two TF groups) is observed for the [Lys-*XX*-Lys] and [Arg-Arg] patterns is encouraging since positively charged Lys and Arg are obviously the key amino acids responsible for TF binding to the negatively charged DNA molecule.

Two conclusions can be drawn from our results. First, that the intrinsic propensity for nonconsensus protein-DNA binding is imprinted both into the DNA and the protein. Since our simple nonconsensus binding model treats proteins as random binders, it captures general trends in the binding profiles of most, but not all, TFs. Second, nonconsensus and specific (consensus) protein-DNA binding mechanisms are tightly interlinked, and both of these mechanisms cooperate in determining the overall protein-DNA binding preferences in eukaryotic genomes. The fact that our simple random-binder model (without any fitting parameters and without any protein-DNA binding specificity built in) provides such a good statistical description of the measured DNA binding strength for the majority of TFs strongly suggests that the nonconsensus mechanism is quite general and it represents the statistical law rather than the exception. However, more accurate, atomistic models describing nonconsensus protein-DNA binding interactions are necessary in order to improve the accuracy of our predictions for different proteins.

Discussion

Our analyses of the effect of nonconsensus protein-DNA binding demonstrate that the combined genome-wide binding preferences of 69 TFs in *C. elegans* and 30 TFs in *D. melanogaster* are significantly, negatively correlated with the predicted nonconsensus free energy landscape (Fig 2). Our analyses also show that the experimentally derived nucleosome occupancy in *C. elegans* and in *D. melanogaster* is significantly, positively correlated with the predicted nonconsensus protein-DNA binding free energy (Fig 3). This trend is qualitatively similar to the one that we previously observed in yeast [12]. The results shown in Figs 2 and 3 strongly suggest that TFs compete with nucleosomes for nonconsensus binding to DNA. Such a competition between TFs and nucleosomes could lead to the enhanced TF binding cooperativity previously predicted by Mirny [51] and Teif et al. [52]. We suggest that nonconsensus protein-DNA binding greatly enhances such nucleosome-induced cooperativity between TFs, and most importantly, in order to achieve this enhancement, promoters do not require the presence of specific, consensus TF binding sites. We stress the important point that the predicted effect of nonconsensus TF-DNA binding most likely affects many but not all TFs. We expect for example, that stress response TFs, such as for example Msn2 in yeast [53], might be insignificantly influenced by the nonconsensus mechanism.

Our model predicts that genomic loci enriched with repetitive sequences, such as in heterochromatin, should also be enriched with TF binding. However, the ChIP-seq analysis in such regions is impeded by the fact that multi-mapping reads from long repetitive region will be filtered out by most peak-calling algorithms, therefore identifying interactions in these regions remains a challenging problem [54]. Interestingly, there are evidences that regions of heterochromatin are not actually transcriptionally inert and non-coding RNA molecules are transcribed from repeated DNA sequences in pericentromeric heterochromatin in different eukaryotic genomes [55]. A recent study even demonstrated [56] that some TFs bind directly to the major satellite repeat DNA sequences that are present in pericentromeric heterochromatin regions and might play a significant role in the mouse heterochromatin formation. Further experiments and analysis of TF binding to the heterochromatin would reveal whether nonconsensus binding play an important role in these regions as well.

Our analysis of available *in vitro* TF-DNA binding data from protein-binding microarray (PBM) experiments (Fig 4 and S1 Table) demonstrates that statistically, on average, *in vitro* TF-DNA binding preferences negatively correlate with the computed nonconsensus free energy landscape, and showed qualitatively similar behavior to the one observed *in vivo* (compare Fig 2b and 2d with Fig 4). This additional analysis is important for several reasons. First, the *in vitro* TF-DNA binding preferences are not affected by the presence of other proteins and histones, which can compete with the protein or cause an indirect binding to the DNA. Second, the TF binding intensity is measured in PBM experiments at significantly higher accuracy compared to ChIP-seq experiments. Third, the usage of non-genomic sequences that cover all possible 8-mer DNA sequences, eliminates possible sequence bias that might exist in the genomic sequences, and thus PBM measurements provide an entirely independent validation of the nonconsensus protein-DNA binding effect. Finally, the present analysis performed for ~90 TFs extends our previous analysis performed for only 6 TFs [15] by more than an order of magnitude, thus strongly suggesting the generality of the nonconsensus protein-DNA binding effect in eukaryotic genomes.

Interestingly, ten TFs (i.e. 11%) showed an *opposite*, positive correlation between the free energy and the measured TF-DNA occupancy (S1 Table). The latter observation is remarkable, since it demonstrates that a non-negligible fraction of TFs can respond to DNA symmetries (represented by our free energy model) in an opposite way compared to the majority of other TFs. However, we failed to identify any particular structural, sequence, or functional features unique to this set of TFs. This failure might stem from the small number of proteins that exhibited such behavior. Yet, we were able to identify repetitive amino acid sequence patterns that are responsible for enhanced nonconsensus TF-DNA binding (S2 Table). In particular, for the group of TFs characterized by the strongest nonconsensus TF-DNA binding preferences, the most statistically significant enrichment is observed for the [Lys-*XX*-Lys] pattern, while the frequency of [Arg-Arg] pattern is reduced in this group (S2 Table). The latter result is intuitively sound since both Lys and Arg are the key amino acids responsible for TF binding to the negatively charged DNA molecule.

Importantly, in this study, our random-binder statistical mechanics model for protein-DNA interactions does not use any experimentally pre-determined information on either low-affinity or high-affinity TF-DNA binding sites. The genomic DNA sequence constitutes the only experimental parameter of the model. In addition, our model does not have any fitting parameters. Contrary to the case of specific protein-DNA binding that requires the presence of a 6 to 20-bp long specific DNA motif (unique for each individual TF), the nonconsensus protein-DNA binding effect stems from multiple nonspecific interactions between the TF and a relatively long (few tens of bp) DNA fragments enriched with repetitive sequence patterns. The fact that different TFs are affected in a *statistically similar* way by *entirely different DNA sequences* containing similar repetitive patterns constitutes the key difference between the nonconsensus and specific protein-DNA recognition modes.

What exactly is the interplay between nonconsensus DNA repetitive sequence elements and consensus (specific) sequences and how their combination influences the overall binding of proteins to the DNA and the expression levels of genes are important questions yet to be explored. We suggest that repetitive nonconsensus sequence elements might have similar influence on TF-DNA binding and on gene expression as repeats of consensus (specific) DNA sequence elements (i.e. homotypic clusters) [57]. However, an important difference between these two types of repeated sequence elements is that nonconsensus repeats can affect many different TFs in a similar way, while homotypic clusters are more specific to a limited set of TFs.

Repetitive sequence elements located near the consensus (specific) motif, could increase the TF association rate, by inducing the one-dimension "sliding" of the TF, and improving its search for the specific binding site [20, 58]. The presence of many weaker sites flanking a strong binding site could lead to a funnel effect [59–62], where the molecules are directed to the strong binding site as depicted in S5a Fig. It could also stabilize binding sites that are not strong enough individually [63, 64] and increase the ability of binding sites to "withstand mutations" [65]. We use the *C. elegans* Hlh-1 protein as an example demonstrating that nonconsensus DNA sequence elements might stabilize the binding to specific consensus elements *in vivo* (S5b

Fig). The analysis of Hlh-1 binding sites (based on the genome-wide ChIP-seq measurements [2, 3] in *C. elegans*) demonstrates that only 5% of the total number of Hlh-1 specific motifs in the genome is bound by Hlh-1 (S5b Fig). We sorted the genomic sequences containing the Hlh-1 motif (consensus motifs were reported in [3]) into two groups: the first group contains DNA sequences that were experimentally determined as being bound by Hlh-1, while the second group contains unbound DNA sequences. S5c Fig represents the average nonconsensus protein-DNA binding free energy computed for each of these two sequence groups. We observed that the nonconsensus free energy is reduced for the group that contains bound sequences as compared with the group that contains unbound sequences. The computed *p*-values show that this result is statistically significant (S5c Fig). This example supports the hypothesis that nonconsensus sequence elements might provide the funnel effect *in vivo*. Additional analysis and experimental measurements of the kinetics of TF-DNA binding to consensus (specific) sequence elements embedded in different nonconsensus DNA backgrounds, should shed more light on this hypothesis.

Future *in vitro* measurements of binding preferences for additional TFs [66], combined with high-resolution *in vivo* ChIP-seq and ChIP-exo analysis, will help to complete the molecular picture of design principles for nonconsensus protein-DNA binding and its functional significance.

Methods

Gene sets

We used the set of 23,287 *C. elegans* genes based on Wormbase annotation, WS228 [2, 67], and 12,188 *D. melanogaster* genes annotated in [10].

Experimental *in vivo* TF occupancy

We used experimentally measured binding preferences of 69 *C. elegans* TFs (S3 Table), as determined by the Gerstein and Snyder labs [2, 3]; for computing the *D. melanogaster* TF occupancy we used binding preferences of 30 TFs (S4 Table) determined by the White lab [8]. TF-DNA binding preferences for both genomes were measured using ChIP-seq assays (modENCODE project). We defined TF occupancy for each genomic location as the total number of bound TFs at each location along the genome.

Experimental nucleosome occupancy

We used experimentally measured, genome-wide, normalized nucleosome occupancy determined by the paired-end Ilumina sequencing in *C. elegans* [4, 5]; we also used the genome-wide map of H2A.Z nucleosome occupancy in *D. melanogaster* embryos (0–12 hr) (determined in [32]).

Experimental *in vitro* TF-DNA binding strength measured using PBM

We used experimentally measured *in vitro* binding intensity for the *C. elegans*, *D. melanogaster*, and *mus musculus* TFs (S1 Table), determined using the protein-binding microarray (PBM) technology [33, 35, 47–49].

Calculation of the free energy of nonconsensus protein-DNA binding

In order to compute the nonconsensus protein-DNA binding free energy landscape, we generate an ensemble of random DNA binders as a proxy for the phenomenon of nonconsensus protein-DNA binding in a crowded cellular environment [11]. Our model does not use any experimentally pre-determined protein-DNA binding preferences in order to model protein-DNA binding. The actual DNA sequences of the *C. elegans* and *D. melanogaster* genomes constitute the only input parameter for our model. In order to compute the free energy of nonconsensus protein-DNA binding at any given location along a DNA sequence, we position the center of the sliding window of width L = 50 bp at that location. The 50 bp length is a typical sliding event distance of a protein along the DNA under physiological conditions [68, 69] (Fig 1).

We assume that a model protein (random binder) makes M bp contacts with the DNA (Fig 1b) and that the model protein-DNA interaction energy at each genomic position i is simply a sum of M interaction energies:

$$U(i) = -\sum_{j=i}^{M+i-1} \sum_{\alpha=\{A,T,C,G\}} K_\alpha s_\alpha(j) \tag{1}$$

where $s_\alpha(j)$ represents the elements of a four-component vector of the type ($\delta_{\alpha A}$, $\delta_{\alpha T,}$ $\delta_{\alpha C,}$ $\delta_{\alpha G}$), and $\delta_{\alpha\beta} = 1$ if $\alpha = \beta$, or $\delta_{\alpha\beta} = 0$ if $\alpha \neq \beta$. For example, if the A nucleotide is positioned at the coordinate j along the DNA, then this vector takes the form: (1,0,0,0). If, for example, the DNA sequence contains entirely poly(A) at a given genomic location, then a random binder makes all M contacts with the A nucleotide, and hence at this location the resulting energy, Eq (1), will be simply, MK_A. In order to generate each model protein, we draw the values of K_A, K_T, K_C, and K_G from Gaussian probability distributions, $P(K_\alpha)$, with zero mean, and standard deviation $\sigma_\alpha = 2k_BT$, where T is the temperature and k_B is the Boltzmann constant. We have shown previously that the resulting free energy is qualitatively robust with respect to the choice of model parameters [11]. The energy scale, $2k_BT \simeq 1.2$ kcal/mol, is chosen to represent a typical strength of a hydrogen bond, or an electrostatic bond that a protein makes with one DNA bp [16, 19].

For each model random binder, we define the partition function of protein-DNA binding within the chosen sliding window of width L bp:

$$Z = \sum_{i=1}^{L} \exp(-U(i)/k_BT) \tag{2}$$

and the corresponding free energy of *nonconsensus* protein-DNA binding in this sliding window:

$$F = -k_BT\ln Z \tag{3}$$

We then assign the computed F to the sequence coordinate in the middle of the sliding window. Next, we move the sliding window along the DNA sequence and we compute F at each sequence location. This procedure allows us to assign the *free energy of nonconsensus protein-DNA* binding to each DNA bp within the genome.

Next, we repeat the described procedure for an ensemble of 250 model random binders (Fig 1) and compute the average free energy, $\langle F_{TF}\rangle$, over this ensemble, at each sequence location. We stress that the resulting free energy is qualitatively robust with respect to the choice of the sliding window size, L, within a wide range of values (S1 Fig). In addition, the free energy profiles are statistically robust with respect to a moderate variation of the value of M, within a typical range of the TF binding site size (S1 Fig). We verified that the predicted free energy landscape is dominated by DNA sequence correlations, and *not* by the average nucleotide composition (S2 Fig). In particular, for each random binder, in each sliding window we computed the normalized free energy, $\delta F = F - F_{rand}$, where F_{rand} is the free energy computed for a randomized sequence (in the same sliding window as F) and averaged over 25 random realizations.

p-value calculations

In order to compute the *p*-value for S5c Fig, we first selected all the 800 bp-long sequences containing the exact binding motifs for each TF. For example, genome-wide, we have overall 9258 sequences containing the consensus Hlh-1motif. Among those 9258 sequences, 442 sequences were experimentally determined as bound by Hlh-1, while the rest of 8816 sequences were unbound. In order to compute the *p*-value, we compiled 10^5 pairs of groups containing 442 and 8816 sequences, respectively, randomly chosen from the original 9258 sequences. These 10^5 pairs of groups represent randomized analogs for the original groups of bound and unbound Hlh-1motifs. Second, for each of these pairs of random groups we computed the average free energies, $\langle f \rangle$, of nonconsensus binding separately for the randomized bound and unbound groups, as described above. Third, for each pair of randomized groups we computed the difference of the integrated free energy within the interval (-400,400) between the two randomized groups. Finally, we computed the probability that this difference is equal or larger than the actual value of the difference. The latter probability was taken as the *p*-value.

Supporting Information

S1 Fig. Robustness of the computed free energy of nonconsensus protein-DNA binding with respect to (A) the TFBS size, *M*, and (B) the width of the sliding window, *L*. Plots show the normalized, average free energy per bp, $\langle \delta f \rangle = \langle \langle \delta F \rangle_{TF} \rangle_{seq} / M$, where δF is computed in the interval (-400,400) around the TSSs of 18,150 genes in *C. elegans*. The free energy *F* is computed as described in the main text, using an ensemble of 125 random DNA binders. F_{rand} is the free energy computed for a randomized sequence (in the same sliding window as *F*), and averaged over 25 random realizations.
(EPS)

S2 Fig. Robustness of the computed free energy of nonconsensus protein-DNA binding with respect to the global variability of the nucleotide content along the genome. Plot shows the average free energy per bp, $\langle f \rangle$ (blue curve) compared to the corresponding *normalized* average free energy $\langle \delta f \rangle = \langle \langle \delta F \rangle_{TF} \rangle_{seq} / M$ (red curve), where $\delta F = F - F_{rand}$. Both the unnormalized and normalized energies are plotted in the interval (-400,400) around the TSSs of 18,150 genes in *C. elegans*. The free energy *F* is computed as described in the main text, using an ensemble of 125 random DNA binders. F_{rand} is the free energy computed for a randomized sequence (in the same sliding window as *F*), and averaged over 25 random realizations. We used $M = 8$ and $L = 50$ in our calculations. The described procedure removes a possible bias in the free energy stemming from the global variability of the nucleotide content.
(EPS)

S3 Fig. Illustration of the entropy-dominated mechanism for nonconsensus TF binding to repetitive DNA sequence elements. (a) Probability distribution, $P(U)$, for two different groups of DNA sequences: the first group is composed of 1000 genomic (*C. elegans*) DNA sequences containing repetitive elements (black), and the second group of sequences is composed of randomly permuted DNA sequences from the first group (gray). The repetitive, genomic sequences are characterized by a wider standard deviation of $P(U)$ than the randomly permuted, non-repetitive sequences. The length of each sequence is 58-bp; the sequences were selected from the TSS region of the *C. elegans* genes. (**b**) The repetitive DNA sequences are characterized by the lower (statistically, on average) free energy of nonconsensus TF-DNA binding, $<F>_{TF}$ than the randomly permuted, non-repetitive DNA sequences.
(EPS)

S4 Fig. Classification of TFs with respect to their DNA-binding domain (DBD) families and the correlation between TF occupancy and free energy of nonconsensus TF-DNA binding. Each box-plot represents **(a)** DBD family, where each TF belonging to this family is represented by a single bar. The *y*-axis displays the correlation *R*-value between the measured *in vitro* TF occupancy and the free energy of nonconsensus DNA-binding. Most of the DBD families only contain TFs with negative *R*-values. One family contains two proteins with positives *R*-values (SAND). Three DBD families contain TFs that are both negatively and positively correlated with the free energy (BRLZ, GATA, HLH). Fig shows only the DBD families that contain at least two TFs (15 out of 22 DBD families), and only mouse TFs (74 out of 91 PBM-tested TFs).
(EPS)

S5 Fig. Example of how nonconsensus sequence elements can influence consensus (specific) TF-DNA binding to the specific TFBS. (a) Schematic representation of the nonconsensus funnel effect. Repetitive, nonconsensus sequence elements can increase the TF binding to the DNA near a strong, specific binding site, and to induce a one-dimension "sliding" of the TF towards the specific TFBS. **(b)** Two examples of sequences, both containing exactly the same consensus-binding motif ACAGCTG for the *C. elegans* transcription factor Hlh-1, surrounded by different nonconsensus sequence elements (Hlh-1was detected as bound [2, 3] to the specific binding motif shown in the top sequence, but it remained unbound to the *identical* specific motif shown the bottom sequence). **(c)** The computed average free energy per bp, $\langle \delta f \rangle = \langle \langle \delta F \rangle_{TF} \rangle_{seq} / M$, in the interval (-400,400) around Hlh-1specific motifs that was detected as being bound (red). Blue line corresponds to $\langle f \rangle$ computed for DNA sequences surrounding unbound Hlh-1motifs [2, 3]. The specific motifs, which were detected as being bound, are surrounded by DNA sequences with significantly lower average free energy compared to unbound motifs (computed p-value $< 10^{-5}$; see Methods).
(EPS)

S1 Table. The table shows the correlation between the computed free energy of nonconsensus TF-DNA binding, *f*, and the measured TF binding intensity for 91 TFs from [33, 34,47,50]. The linear correlation coefficient, *R*, and the *p*-value were calculated for each TF after binning the data into 50 bins (the binning is performed in a way similar to the binning performed in the plots presented in the main text). 69 out of 91 proteins statistically behave according to our model ($p < 0.05$), ten proteins exhibit the opposite behavior ($p < 0.05$), while the remaining 12 proteins show no statistically significant correlation ($p > 0.05$). For each TF in the table, the protein name and its DNA-binding domain type are specified.
(XLSX)

S2 Table. Statistical analysis of repetitive amino acid patterns and the amino acid content of TF sequences belonging to the two groups of the mouse TFs (82 TFs overall). The first group (1st average) contained 41 TFs with the strongest negative correlation between the free energy and the measured TF occupancy. The second group (2nd average) contained the remaining 41 TFs. We have analyzed the frequency of occurrence of the following repetitive amino acid sequence patterns in each TF group: [*aa*], [*aXa*], [*aXXa*], and [*aXXXa*], where *a* represents each amino acid type and *X* represents an arbitrary amino acid. The second table contains the average amino acid content in each group of TFs. The presented p-values represent the Kolmogorov-Smirnov p-values.
(XLSX)

S3 Table. List of modEncode *C. elegans* TFs.
(XLSX)

S4 Table. List of modEncode *D. melanogaster* TFs.
(XLSX)

Acknowledgments

We thank Yifat Miller and the staff of the BGU High Performance Cluster computational center.

Author Contributions

Conceived and designed the experiments: AA HC RG DBL. Performed the experiments: AA HC SBZ. Analyzed the data: AA HC SBZ RG DBL. Contributed reagents/materials/analysis tools: SBZ. Wrote the paper: AA HC RG DBL.

References

1. Iyer V, Struhl K. Poly(dA:dT), a ubiquitous promoter element that stimulates transcription via its intrinsic DNA structure. EMBO J. 1995; 14(11):2570–9. Epub 1995/06/01. PMID: 7781610
2. Gerstein MB, Lu ZJ, Van Nostrand EL, Cheng C, Arshinoff BI, Liu T, et al. Integrative analysis of the Caenorhabditis elegans genome by the modENCODE project. Science. 2010; 330(6012):1775–87. Epub 2010/12/24. doi: 10.1126/science.1196914 PMID: 21177976
3. Niu W, Lu ZJ, Zhong M, Sarov M, Murray JI, Brdlik CM, et al. Diverse transcription factor binding features revealed by genome-wide ChIP-seq in C. elegans. Genome Research. 2011; 21(2):245–54. doi: 10.1101/gr.114587.110 PMID: 21177963
4. Ercan S, Lubling Y, Segal E, Lieb JD. High nucleosome occupancy is encoded at X-linked gene promoters in C. elegans. Genome Res. 2011; 21(2):237–44. Epub 2010/12/24. doi: 10.1101/gr.115931.110 PMID: 21177966
5. Liu T, Rechtsteiner A, Egelhofer TA, Vielle A, Latorre I, Cheung MS, et al. Broad chromosomal domains of histone modification patterns in C. elegans. Genome Res. 2011; 21(2):227–36. Epub 2010/12/24. doi: 10.1101/gr.115519.110 PMID: 21177964
6. Harris TW, Antoshechkin I, Bieri T, Blasiar D, Chan J, Chen WJ, et al. WormBase: a comprehensive resource for nematode research. Nucleic Acids Res. 2010; 38(Database issue):D463–7. Epub 2009/11/17. doi: 10.1093/nar/gkp952 PMID: 19910365
7. Spencer WC, Zeller G, Watson JD, Henz SR, Watkins KL, McWhirter RD, et al. A spatial and temporal map of C. elegans gene expression. Genome Research. 2011; 21(2):325–41. doi: 10.1101/gr.114595.110 PMID: 21177967
8. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010; 330(6012):1787–97. doi: 10.1126/science.1198374 PMID: 21177974
9. Roy S, Ernst J, Kharchenko PV, Kheradpour P, Negre N, Eaton ML, et al. Identification of functional elements and regulatory circuits by Drosophila modENCODE. Science. 2010; 330(6012):1787–97. Epub 2010/12/24. doi: 10.1126/science.1198374 PMID: 21177974
10. Negre N, Brown CD, Ma L, Bristow CA, Miller SW, Wagner U, et al. A cis-regulatory map of the Drosophila genome. Nature. 2011; 471(7339):527–31. Epub 2011/03/25. doi: 10.1038/nature09990 PMID: 21430782
11. Sela I, Lukatsky DB. DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. Biophysical Journal. 2011; 101(1):160–6. Epub 2011/07/05. doi: 10.1016/j.bpj.2011.04.037 PMID: 21723826
12. Afek A, Sela I, Musa-Lempel N, Lukatsky DB. Nonspecific transcription-factor-DNA binding influences nucleosome occupancy in yeast. Biophysical Journal. 2011; 101(10):2465–75. Epub 2011/11/22. doi: 10.1016/j.bpj.2011.10.012 PMID: 22098745
13. Afek A, Lukatsky DB. Nonspecific protein-DNA binding is widespread in the yeast genome. Biophysical Journal. 2012; 102(8):1881–8. Epub 2012/07/10. doi: 10.1016/j.bpj.2012.03.044 PMID: 22768944
14. Afek A, Lukatsky DB. Genome-Wide Organization of Eukaryotic Preinitiation Complex Is Influenced by Nonconsensus Protein-DNA Binding. Biophysical Journal. 2013; 104(5):1107–15. Epub 2013/03/12. doi: 10.1016/j.bpj.2013.01.038 PMID: 23473494

15. Afek A, Schipper JL, Horton J, Gordan R, Lukatsky DB. Protein-DNA binding in the absence of specific base-pair recognition. Proceedings of the National Academy of Sciences of the United States of America. 2014; 111(48):17140–5. doi: 10.1073/pnas.1410569111 PMID: 25313048
16. von Hippel PH, Revzin A, Gross CA, Wang AC. Non-specific DNA binding of genome regulating proteins as a biological control mechanism: I. The lac operon: equilibrium aspects. Proc Natl Acad Sci U S A. 1974; 71(12):4808–12. Epub 1974/12/01. PMID: 4612528
17. Berg OG, Winter RB, von Hippel PH. Diffusion-driven mechanisms of protein translocation on nucleic acids. 1. Models and theory. Biochemistry. 1981; 20(24):6929–48. Epub 1981/11/24. PMID: 7317363
18. Berg OG, von Hippel PH. Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. J Mol Biol. 1987; 193(4):723–50. PMID: 3612791
19. von Hippel PH, Berg OG. On the specificity of DNA-protein interactions. Proc Natl Acad Sci U S A. 1986; 83(6):1608–12. Epub 1986/03/01. PMID: 3456604
20. von Hippel PH, Berg OG. Facilitated target location in biological systems. J Biol Chem. 1989; 264 (2):675–8. Epub 1989/01/15. PMID: 2642903
21. von Hippel PH. From "simple" DNA-protein interactions to the macromolecular machines of gene expression. Annu Rev Biophys Biomol Struct. 2007; 36:79–105. Epub 2007/05/05. PMID: 17477836
22. Slutsky M, Mirny LA. Kinetics of protein-DNA interaction: facilitated target location in sequence-dependent potential. Biophys J. 2004; 87(6):4021–35. Epub 2004/10/07. PMID: 15465864
23. Wunderlich Z, Mirny LA. Different gene regulation strategies revealed by analysis of binding motifs. Trends Genet. 2009; 25(10):434–40. Epub 2009/10/10. doi: 10.1016/j.tig.2009.08.003 PMID: 19815308
24. Elf J, Li GW, Xie XS. Probing transcription factor dynamics at the single-molecule level in a living cell. Science. 2007; 316(5828):1191–4. Epub 2007/05/26. PMID: 17525339
25. Hammar P, Leroy P, Mahmutovic A, Marklund EG, Berg OG, Elf J. The lac repressor displays facilitated diffusion in living cells. Science. 2012; 336(6088):1595–8. Epub 2012/06/23. doi: 10.1126/science.1221648 PMID: 22723426
26. Liebesny P, Goyal S, Dunlap D, Family F, Finzi L. Determination of the number of proteins bound non-specifically to DNA. J Phys Condens Matter. 2010; 22(41):414104. Epub 2011/03/10. doi: 10.1088/0953-8984/22/41/414104 PMID: 21386587
27. Manzo C, Zurla C, Dunlap DD, Finzi L. The effect of nonspecific binding of lambda repressor on DNA looping dynamics. Biophys J. 2012; 103(8):1753–61. Epub 2012/10/23. doi: 10.1016/j.bpj.2012.09.006 PMID: 23083719
28. Zurla C, Manzo C, Dunlap D, Lewis DE, Adhya S, Finzi L. Direct demonstration and quantification of long-range DNA looping by the lambda bacteriophage repressor. Nucleic Acids Res. 2009; 37 (9):2789–95. Epub 2009/03/12. doi: 10.1093/nar/gkp134 PMID: 19276206
29. Wang YM, Austin RH, Cox EC. Single molecule measurements of repressor protein 1D diffusion on DNA. Phys Rev Lett. 2006; 97(4):048302. Epub 2006/08/16. PMID: 16907618
30. Blainey PC, Luo G, Kou SC, Mangel WF, Verdine GL, Bagchi B, et al. Nonspecifically bound proteins spin while diffusing along DNA. Nat Struct Mol Biol. 2009; 16(12):1224–9. Epub 2009/11/10. doi: 10.1038/nsmb.1716 PMID: 19898474
31. Tafvizi A, Huang F, Leith JS, Fersht AR, Mirny LA, van Oijen AM. Tumor suppressor p53 slides on DNA with low friction and high stability. Biophys J. 2008; 95(1):L01–3. Epub 2008/04/22. doi: 10.1529/biophysj.108.134122 PMID: 18424488
32. Mavrich TN, Jiang C, Ioshikhes IP, Li X, Venters BJ, Zanton SJ, et al. Nucleosome organization in the Drosophila genome. Nature. 2008; 453(7193):358–62. Epub 2008/04/15. doi: 10.1038/nature06929 PMID: 18408708
33. Grove CA, De Masi F, Barrasa MI, Newburger DE, Alkema MJ, Bulyk ML, et al. A multiparameter network reveals extensive divergence between C. elegans bHLH transcription factors. Cell. 2009; 138 (2):314–27. Epub 2009/07/28. doi: 10.1016/j.cell.2009.04.058 PMID: 19632181
34. Badis G, Berger MF, Philippakis AA, Talukder S, Gehrke AR, Jaeger SA, et al. Diversity and complexity in DNA recognition by transcription factors. Science. 2009; 324(5935):1720–3. Epub 2009/05/16. doi: 10.1126/science.1162327 PMID: 19443739
35. Berger MF, Bulyk ML. Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. Nat Protoc. 2009; 4(3):393–411. Epub 2009/03/07. doi: 10.1038/nprot.2008.195 PMID: 19265799
36. Afek A, Shakhnovich EI, Lukatsky DB. Multi-scale sequence correlations increase proteome structural disorder and promiscuity. J Mol Biol. 2011; 409(3):439–49. Epub 2011/04/06. doi: 10.1016/j.jmb.2011.03.056 PMID: 21463640

37. Lukatsky DB, Afek A, Shakhnovich EI. Sequence correlations shape protein promiscuity. J Chem Phys. 2011; 135(6):065104. Epub 2011/08/17. doi: 10.1063/1.3624332 PMID: 21842953

38. Elkin M, Andre I, Lukatsky DB. Energy Fluctuations Shape Free Energy of Nonspecific Biomolecular Interactions. Journal of Statistical Physics. 2012; 146(4):870–7.

39. Luger K, Mader AW, Richmond RK, Sargent DF, Richmond TJ. Crystal structure of the nucleosome core particle at 2.8 angstrom resolution. Nature. 1997; 389(6648):251–60. PMID: 9305837

40. Liu X, Bushnell DA, Wang D, Calero G, Kornberg RD. Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism. Science. 2010; 327(5962):206–9. Epub 2009/12/08. doi: 10.1126/science.1182015 PMID: 19965383

41. Kasinathan S, Orsi GA, Zentner GE, Ahmad K, Henikoff S. High-resolution mapping of transcription factor binding sites on native chromatin. Nature methods. 2014; 11(2):203–9. doi: 10.1038/nmeth.2766 PMID: 24336359

42. Valouev A, Ichikawa J, Tonthat T, Stuart J, Ranade S, Peckham H, et al. A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. Genome Res. 2008; 18(7):1051–63. Epub 2008/05/15. doi: 10.1101/gr.076463.108 PMID: 18477713

43. Zhang Z, Wippo CJ, Wal M, Ward E, Korber P, Pugh BF. A packing mechanism for nucleosome organization reconstituted across a eukaryotic genome. Science. 2011; 332(6032):977–80. Epub 2011/05/21. doi: 10.1126/science.1200508 PMID: 21596991

44. Zhang Y, Moqtaderi Z, Rattner BP, Euskirchen G, Snyder M, Kadonaga JT, et al. Intrinsic histone-DNA interactions are not the major determinant of nucleosome positions in vivo. Nat Struct Mol Biol. 2009; 16(8):847–52. Epub 2009/07/22. doi: 10.1038/nsmb.1636 PMID: 19620965

45. Kaplan N, Moore IK, Fondufe-Mittendorf Y, Gossett AJ, Tillo D, Field Y, et al. The DNA-encoded nucleosome organization of a eukaryotic genome. Nature. 2009; 458(7236):362–6. Epub 2008/12/19. doi: 10.1038/nature07667 PMID: 19092803

46. Lorch Y, Maier-Davis B, Kornberg RD. Role of DNA sequence in chromatin remodeling and the formation of nucleosome-free regions. Genes Dev. 2014; 28(22):2492–7. Epub 2014/11/19. doi: 10.1101/gad.250704.114 PMID: 25403179

47. Berger MF, Philippakis AA, Qureshi AM, He FS, Estep PW 3rd, Bulyk ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nat Biotechnol. 2006; 24(11):1429–35. Epub 2006/09/26. PMID: 16998473

48. Mukherjee S, Berger MF, Jona G, Wang XS, Muzzey D, Snyder M, et al. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. Nat Genet. 2004; 36(12):1331–9. Epub 2004/11/16. PMID: 15543148

49. Gordan R, Shen N, Dror I, Zhou T, Horton J, Rohs R, et al. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. Cell Rep. 2013; 3(4):1093–104. Epub 2013/04/09. doi: 10.1016/j.celrep.2013.03.014 PMID: 23562153

50. Berger MF, Badis G, Gehrke AR, Talukder S, Philippakis AA, Pena-Castillo L, et al. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. Cell. 2008; 133(7):1266–76. Epub 2008/07/01. doi: 10.1016/j.cell.2008.05.024 PMID: 18585359

51. Mirny LA. Nucleosome-mediated cooperativity between transcription factors. Proc Natl Acad Sci U S A. 2010; 107(52):22534–9. Epub 2010/12/15. doi: 10.1073/pnas.0913805107 PMID: 21149679

52. Teif VB, Rippe K. Nucleosome mediated crosstalk between transcription factors at eukaryotic enhancers. Phys Biol. 2011; 8(4):044001. doi: 10.1088/1478-3975/8/4/044001 PMID: 21666293

53. Elfving N, Chereji RV, Bharatula V, Bjorklund S, Morozov AV, Broach JR. A dynamic interplay of nucleosome and Msn2 binding regulates kinetics of gene activation and repression following stress. Nucleic Acids Res. 2014; 42(9):5468–82. doi: 10.1093/nar/gku176 PMID: 24598258

54. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol. 2013; 9(11):e1003326. Epub 2013/11/19. doi: 10.1371/journal.pcbi.1003326 PMID: 24244136

55. Grewal SI, Elgin SC. Transcription and RNA interference in the formation of heterochromatin. Nature. 2007; 447(7143):399–406. PMID: 17522672

56. Bulut-Karslioglu A, Perrera V, Scaranaro M, de la Rosa-Velazquez IA, van de Nobelen S, Shukeir N, et al. A transcription factor-based mechanism for mouse heterochromatin formation. Nat Struct Mol Biol. 2012; 19(10):1023–30. Epub 2012/09/18. doi: 10.1038/nsmb.2382 PMID: 22983563

57. Ezer D, Zabet NR, Adryan B. Homotypic clusters of transcription factor binding sites: A model system for understanding the physical mechanics of gene expression. Computational and structural biotechnology journal. 2014; 10(17):63–9. doi: 10.1016/j.csbj.2014.07.005 PMID: 25349675

58. Kolomeisky AB. Physics of protein—DNA interactions: mechanisms of facilitated target search. Physical Chemistry Chemical Physics. 2011; 13(6):2088–95. doi: 10.1039/c0cp01966f PMID: 21113556

59. Shimamoto N. One-dimensional diffusion of proteins along DNA Its biological and chemical significance revealed by single-molecule measurements. Journal of Biological Chemistry. 1999; 274 (22):15293–6. PMID: 10336412

60. Esadze A, Kemme CA, Kolomeisky AB, Iwahara J. Positive and negative impacts of nonspecific sites during target location by a sequence-specific DNA-binding protein: origin of the optimal search at physiological ionic strength. Nucleic acids research. 2014:gku418.

61. Weindl J, Dawy Z, Hanus P, Zech J, Mueller JC. Modeling promoter search by E. coli RNA polymerase: One-dimensional diffusion in a sequence-dependent energy landscape. Journal of theoretical biology. 2009; 259(3):628–34. doi: 10.1016/j.jtbi.2009.05.006 PMID: 19463831

62. Mirny L, Slutsky M, Wunderlich Z, Tafvizi A, Leith J, Kosmrlj A. How a protein searches for its site on DNA: the mechanism of facilitated diffusion. Journal of Physics A: Mathematical and Theoretical. 2009; 42(43):434013.

63. Zhang C, Xuan Z, Otto S, Hover JR, McCorkle SR, Mandel G, et al. A clustering property of highly-degenerate transcription factor binding sites in the mammalian genome. Nucleic acids research. 2006; 34(8):2238–46. PMID: 16670430

64. Afek A, Schipper JL, Horton J, Gordân R, Lukatsky DB. Protein– DNA binding in the absence of specific base-pair recognition. Proceedings of the National Academy of Sciences. 2014; 111(48):17140–5.

65. Smith T, Husbands P, Layzell P, O'Shea M. Fitness landscapes and evolvability. Evolutionary computation. 2002; 10(1):1–34. PMID: 11911781

66. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, et al. Determination and inference of eukaryotic transcription factor sequence specificity. Cell. 2014; 158(6):1431–43. Epub 2014/09/13. doi: 10.1016/j.cell.2014.08.009 PMID: 25215497

67. Yook K, Harris TW, Bieri T, Cabunoc A, Chan J, Chen WJ, et al. WormBase 2012: more genomes, more data, new website. Nucleic Acids Res. 2012; 40(Database issue):D735–41. Epub 2011/11/10. doi: 10.1093/nar/gkr954 PMID: 22067452

68. Halford S. An end to 40 years of mistakes in DNA-protein association kinetics? Biochemical Society Transactions. 2009; 37(2):343.

69. Bonnet I, Biebricher A, Porté P-L, Loverdo C, Bénichou O, Voituriez R, et al. Sliding and jumping of single EcoRV restriction enzymes on non-cognate DNA. Nucleic acids research. 2008; 36(12):4118–27. doi: 10.1093/nar/gkn376 PMID: 18544605