Please cite this article in press as: Goldshtein et al., Transcription Factor Binding in Embryonic Stem Cells Is Constrained by DNA Sequence Repeat Symmetry, Biophysical Journal (2020), https://doi.org/10.1016/j.bpj.2020.02.009

Biophysical *Journal*
Article

Biophysical Society

Transcription Factor Binding in Embryonic Stem Cells Is Constrained by DNA Sequence Repeat Symmetry

Matan Goldshtein,[1] Meir Mellul,[2] Gai Deutch,[3] Masahiko Imashimizu,[4] Koh Takeuchi,[4] Eran Meshorer,[5,6] Oren Ram,[2,*] and David B. Lukatsky[3,*]

[1]Avram and Stella Goldstein-Goren Department of Biotechnology Engineering, Ben-Gurion University of the Negev, Beer-Sheva, Israel; [2]Department of Biological Chemistry, The Institute of Life Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel; [3]Department of Chemistry, Ben-Gurion University of the Negev, Beer-Sheva, Israel; [4]Molecular Profiling Research Center for Drug Discovery, National Institute of Advanced Industrial Science and Technology, Tokyo, Japan; [5]Department of Genetics, The Institute of Life Sciences and [6]The Edmond and Lily Safra Center for Brain Sciences, The Hebrew University of Jerusalem, Jerusalem, Israel

ABSTRACT Transcription factor (TF) recognition is dictated by the underlying DNA motif sequence specific for each TF. Here, we reveal that DNA sequence repeat symmetry plays a central role in defining TF-DNA-binding preferences. In particular, we find that different TFs bind similar symmetry patterns in the context of different developmental layers. Most TFs possess dominant preferences for similar DNA repeat symmetry types. However, in some cases, preferences of specific TFs are changed during differentiation, suggesting the importance of information encoded outside of known motif regions. Histone modifications also exhibit strong preferences for similar DNA repeat symmetry patterns unique to each type of modification. Next, using an in vivo reporter assay, we show that gene expression in embryonic stem cells can be positively modulated by the presence of genomic and computationally designed DNA oligonucleotides containing identified nonconsensus-repetitive sequence elements. This supports the hypothesis that certain nonconsensus-repetitive patterns possess a functional ability to regulate gene expression. We also performed a solution NMR experiment to probe the stability of double-stranded DNA via imino proton resonances for several double-stranded DNA sequences characterized by different repetitive patterns. We suggest that such local stability might play a key role in determining TF-DNA binding preferences. Overall, our findings show that despite the enormous sequence complexity of the TF-DNA binding landscape in differentiating embryonic stem cells, this landscape can be quantitatively characterized in simple terms using the notion of DNA sequence repeat symmetry.

SIGNIFICANCE Molecular design principles regulating developmental transitions in embryonic stem cells are still poorly understood. Development of many cancer types such as Ewing's sarcoma and gliomas is driven by cancer stem cells. Such understanding is thus essential for understanding cancer development and designing strategies for cancer treatment. Transcription factor (TF) binding to DNA constitutes a key regulatory step for establishing a genome-wide transcriptional program in embryonic stem cells, and several key master regulator TFs such as Nanog, Oct4, and Esrrb possess a unique ability to drive developmental transitions. Here, we investigate the molecular design principles responsible for TF-DNA-binding recognition specificity of such key master regulator TFs. We find that certain short repetitive DNA sequence patterns substantially influence the binding preferences of these TFs.

INTRODUCTION

Transcription factors (TF) bind DNA in a dynamic and highly combinatorial fashion. Cell types are determined by TF-specific patterns and chromatin regulators, which associate with DNA regulatory regions to activate specific transcriptional programs (1). Understanding molecular principles of protein-DNA recognition is thus essential for

Submitted December 2, 2019, and accepted for publication February 10, 2020.

*Correspondence: oren.ram@mail.huji.ac.il or lukatsky@bgu.ac.il

Matan Goldshtein and Meir Mellul contributed equally to this work.

Editor: Wilma Olson.

https://doi.org/10.1016/j.bpj.2020.02.009

Please cite this article in press as: Goldshtein et al., Transcription Factor Binding in Embryonic Stem Cells Is Constrained by DNA Sequence Repeat Symmetry, Biophysical Journal (2020), https://doi.org/10.1016/j.bpj.2020.02.009

understanding and predicting such transcriptional programs (2,3). The recently identified genome-wide TF-DNA-binding landscape for tens of TFs in differentiating human embryonic stem (ES) cells in four developmental layers demonstrates enormous complexity and plasticity of genomic DNA sequence recognition by TFs (4). In particular, these measurements identified TF-DNA binding preferences and the epigenetic landscape in the human ES cell line HUES64 (ESc) and in four additional developmental layers: mesendoderm (dMS), endoderm, mesoderm, and ectoderm (4).

In this work, we investigate DNA sequence repeat symmetries within TF binding regions to show that specific repeats are associated with TF binding specificity. We also suggest that specific repeat patterns are associated with key TF binding that then regulate early linage differentiation of human ES cells. Remarkably, we find that the majority of different TFs preferentially bind DNA sequences characterized by similar repeat symmetries. To quantitatively characterize DNA sequence repeat symmetries, we use one of the simplest possible measures for symmetry, namely, the nucleotide pair-correlation function, recently developed and tested by us (5).

Many attempts to map TFs to specific motif preferences have been already done (6); however, TF binding motifs (TFBMs) explain only a limited fraction of experimentally bound genomic sequences. For many proteins, such as POL2, THAP11, and TRIM28 analyzed in this work, specific motifs have not been identified at all (4). For example, for two key transcription regulators, EOMES and OTX2, only 6–10% and 14–22% experimentally bound DNA sequences, respectively, contain specific TFBMs for these factors (4). In recent years, it has been experimentally shown, both in vivo and in vitro, that in addition to specific TFBMs, other determinants are also responsible for efficient TF binding. DNA spatial conformation (6–8), chromatin regulators (9), and certain nonconsensus repetitive DNA sequence elements (10) can significantly affect TF-DNA binding. In general, it is now widely recognized that genomic DNA context outside of specific TFBMs affects TF-DNA binding preferences (2,11–14).

We have shown in the past using both biophysical modeling (5,10,15) and high-throughput in vitro measurements of TF-DNA binding preferences (10) that certain nonconsensus, repetitive DNA sequence elements exert an entropy-dominated, statistical interaction potential on TFs. We use the term nonconsensus TF-DNA binding free energy to describe this statistical interaction potential (10). Depending on the DNA symmetry type, the presence of certain repetitive DNA sequence elements can enhance or reduce the TF-DNA binding free energy (10). In this work, we perform a genome-wide repetitive symmetry analysis that does not utilize any biophysical model and does not have any fitting parameters. Experimentally identified genomic DNA sequences (TF binding peaks) constitute the only input to our simple computational procedure.

Here, we identify the dominant DNA repeat symmetry elements that appear to influence large clusters of TFs in different developmental layers. In particular, we show that complex rewiring of the TF-DNA binding network upon developmental transitions can be quantified by the variation of the DNA repeat symmetry type and DNA repeat symmetry strength defined below.

Although we have chosen ES cells as a model system, we expect that the mechanistic understanding of TF recognition by DNA enriched in repetitive sequence elements will be relevant for any biological or synthetic system involving TFs and DNA.

MATERIALS AND METHODS

Analysis of DNA sequence repeat symmetry from chromatin immunoprecipitation sequencing data

All chromatin immunoprecipitation sequencing (ChIP-seq) data used in our analysis are publicly available at the Gene Expression Omnibus under the accession number GSE61475 (4). We now define the measure for DNA sequence symmetry used to characterize genomic repetitive DNA sequence elements. Specifically, here we use the nucleotide pair-correlation function $\eta_{\alpha\alpha}(x)$, similar to the one used in our previous work (5). This correlation function, $\eta_{\alpha\alpha}(x)$, is proportional to the probability of finding two nucleotides of the type α separated by the relative distance x along the genome, $\eta_{\alpha\alpha}(x) = (N_{\alpha\alpha}(x) - <N_{\alpha\alpha}(x)>_{\text{rand}})/L$. For a given set of DNA sequences, $N_{\alpha\alpha}(x)$ is the total number of nucleotide pairs of the type α separated by the relative distance x, $<N_{\alpha\alpha}(x)>_{\text{rand}}$ is the corresponding average number of nucleotide pairs in the randomized sequence set, and L is the total length of DNA sequences in the set. The randomization procedure randomly reshuffles each DNA sequence in the set, keeping the GC-content of each sequence intact. The averaging, $<N_{\alpha\alpha}(x)>_{\text{rand}}$, is performed with respect to 10 random realizations of the original sequence set. Such a randomization procedure normalizes the varying genomic GC-content, allowing us to compare symmetry properties of DNA sequences from different genomic locations characterized by a variable average GC-content. We stress that in our analysis we used the simplest possible measure for characterizing DNA sequence repeats, i.e., the pair-correlation function $\eta_{\alpha\alpha}(x)$. Although this function does not account for higher-order sequence correlations, it efficiently captures the strongest effect of binary nucleotide correlations. We have demonstrated in the past (also for the case of the human genome) that taking into account higher-order correlations can be achieved using a similar method (5).

For example, 60,994 binding peaks were experimentally identified in human ES cells for serum response factor (SRF) protein (4). For each identified peak, we select a 100-basepair (bp) region in the middle of the peak. The entire set of these 60,994 sequences is used to generate the correlation functions, $\eta_{\alpha\alpha}(x)$, for this protein in this developmental layer (Fig. 1 *A*). We use this procedure to generate the correlation functions for each TF or histone modification in each developmental layer. As a result, we obtain the entire set of $\eta_{\alpha\alpha}(x)$ for 36 TFs and five histone modifications in ES cells and in four developmental layers for all reported (MNase)-based ChIP-seq data sets (4). To validate statistical significance of the results, we compute error bars for each reported correlation function. Although not all TFs and not all histone modifications were measured in each of the four developmental layers, the existing data still allow us to obtain a comprehensive, systems-level view on the TF-DNA interaction network and on rewiring of this network in the course of ES cell differentiation.

Please cite this article in press as: Goldshtein et al., Transcription Factor Binding in Embryonic Stem Cells Is Constrained by DNA Sequence Repeat Symmetry, Biophysical Journal (2020), https://doi.org/10.1016/j.bpj.2020.02.009

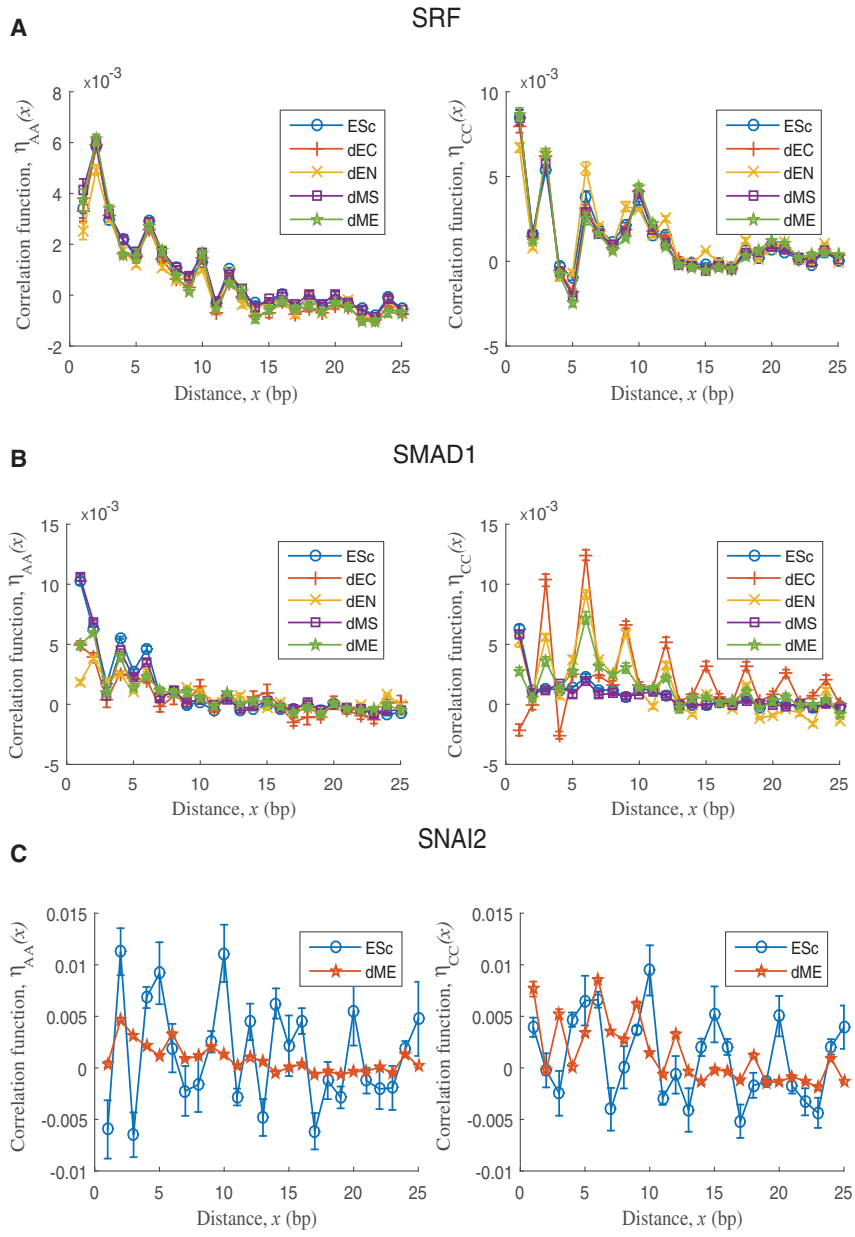

FIGURE 1 Correlation functions for the nucleotide spatial distribution identify DNA repeat symmetries selected by TFs in different developmental layers. The computed correlation function $\eta_{\alpha\alpha}(x)$ for (*A*) SRF shows nearly identical symmetry types and symmetry strengths in different developmental layers; (*B*) SMAD1 shows nearly identical symmetry types but varying symmetry strengths in different developmental layers; (*C*) SNAI2 shows varying symmetry types and varying symmetry strengths in different developmental layers. The correlation function $\eta_{TT}(x)$ behaves similar to the shown $\eta_{AA}(x)$, and $\eta_{GG}(x)$ behaves similar to the shown $\eta_{CC}(x)$ (Fig. S1). To compute error bars, we divided each set of bound DNA sequences into five randomly chosen subgroups with an equal number of sequences and calculated $\eta_{\alpha\alpha}(x)$ for each subgroup. The error bars are defined as one standard deviation of $\eta_{\alpha\alpha}(x)$ between the subgroups. To see this figure in color, go online.

In vivo reporter assay measurements in ES cells

Cloning tested plasmids

Six unique restriction sites within the multiple cloning site and enhancer sequences were inserted into PBS31 plasmids using two overhang premiers (Integrated DNA Technologies, Coralville, IA). PCR amplification of the plasmids were done using Platinum SuperFi DNA polymerase (Invitrogen, Carlsbad, CA). The minimal promoter was cloned from pMPRAdonor2 (Tarjei Mikkelsen, Addgene, Watertown, MA), using XhoI and BamHI restriction enzymes (New England Biolabs, Ipswich, MA) and Quick Ligation Kit (New England Biolabs). Transformation with heat shock was done with *Escherichia coli* XL10 with ratio of 150 ng plasmid to 200 μL bacteria. Heat shock was done at 42°C for 45 s. After cloning, plasmids were extracted with Invitrogen PureLink Quick Plasmid Midiprep Kit (Invitrogen), according to protocol.

Cell culture and transfection

Mouse R1 ES cells from A. Nagy (Toronto, Canada) were seeded on 0.1% gelatin-coated plates (G1393; Sigma-Aldrich, St. Louis, MO) and grown in an ES medium: Dulbecco's modified Eagle's medium (D5671; Sigma-Aldrich), 15% fetal bovine serum (04-007-1A; Biological Industries, Cromwell, CT), 1 mM sodium pyruvate (03042-1B; Biological Industries), 0.1 mM nonessential amino acids (01-340-1B; Biological Industries), 0.1 mM β-mercaptoethanol (M3148; Sigma-Aldrich), and 1000 U/mL leukemia inhibitory factor (ESG1107; Mercury, Rosh-Ha'ayin, Israel). For "2i" conditions, 3 μM CHIR99021 (SM-2520691-B; PeproTech, Rocky Hill, NJ) and 0.2 μM PD0325901 (SM3911091-B; PeproTech) were added to the ES medium. For transfection of plasmids, R1 cells were grown to 60–70% confluence in a 24-well plate culture dish (Corning, Durham, NC). Cells were transfected with 1 μg of DNA from each plasmid in 50 μL of Opti-MEM Reduced Serum Medium (Invitrogen) using 1.5 μL of MIRUS

transit lt-1 transfection reagent (MirusBio, Houston, TX). The transfection mixtures were removed by media exchange after 12 h.

Luciferase activities were measured 48 h after transfection. Cells were lysed with reporter lysis buffer (Promega, Madison, WI). Firefly luciferase activities were measured using GloMax 20/20 Luminometer (Promega).

Solution NMR spectroscopy: Stability of imino proton in dsDNA

DNA oligonucleotides that were purified by reverse phase cartridge were purchased from Fasmac. Double-stranded DNA (dsDNA) of 50 μM were generated in a 90% H_2O/10% D_2O solvent with 20 mM Tris-D11 (pH 7.6 at 25°C), 5 mM $MgCl_2$, and 50 mM KCl and were used for NMR experiments. The NMR experiments were performed on an AV700 spectrometer (Bruker, Billerica, MA) equipped with a 5-mm triple resonance probe at 25°C. One-dimensional spectra were recorded with a 22-ppm spectrum width, centered at 4.7 ppm, using the water gate pulse sequence for solvent suppression. Repetition delay was set to 2.0 s, and 2048 points were acquired. The experiments were repeated 192 times to ensure a sufficient sensitivity to detect the imino proton signals.

RESULTS

DNA repeat symmetry type and DNA repeat symmetry strength

We analyzed ChIP-seq data for 36 TFs and five histone modifications in four developmental layers of differentiating human ES cells (4). All the data is publicly available at the Gene Expression Omnibus under the accession number GSE61475 (4). We analyzed the DNA sequence repeat symmetry properties for each TF binding peak (4) with overall 2,595,176 peaks, ranging between 653 (OTX2) and 100,778 (HAND1) peaks across all 137 data sets.

To characterize DNA sequence repeat symmetries, we use the nucleotide pair-correlation function $\eta_{\alpha\alpha}(x)$, similar to the one introduced in our previous work (5). This correlation function, $\eta_{\alpha\alpha}(x)$, is proportional to the probability of finding two nucleotides of type α separated by the relative distance x along the genome (Materials and Methods). We normalize $\eta_{\alpha\alpha}(x)$ by the genomic GC-content, allowing us to compare symmetry properties of DNA sequences from different genomic locations characterized by a variable average GC-content (Materials and Methods). To compute $\eta_{\alpha\alpha}(x)$ for a given TF in a given developmental layer, we used the entire collection of peaks identified by ChIP-seq in the relevant data set (Materials and Methods).

The computed correlation functions, $\eta_{AA}(x)$, $\eta_{TT}(x)$, $\eta_{CC}(x)$, and $\eta_{GG}(x)$, allow us to characterize the DNA sequence recognition specificity of TFs in terms of the DNA repeat symmetry type and the DNA repeat symmetry strength defined below. We stress the fact that the coordinate x always represents the relative distance between the two nucleotides and not the absolute distance with respect to a certain specific genomic location. For example, in the case of SRF (Fig. 1 *A*), $\eta_{CC}(x)$ and $\eta_{GG}(x)$ are characterized by peaks at $x = 1$, $x = 3$, $x = 6$, and $x = 10$. The presence of such peaks means that DNA sequences bound by SRF in all developmental layers are statistically enriched in repetitive sequence patterns of the type [CC] (corresponding to $x = 1$), [CNNC] (corresponding to $x = 3$), [CNNNNNC] (corresponding to $x = 6$), etc., where N stands for any nucleotide type. On the contrary, $\eta_{AA}(x)$ and $\eta_{TT}(x)$ are characterized by the peaks at $x = 2$, $x = 6$, etc. It means that DNA sequences bound by SRF are also statistically enriched in repetitive sequence patterns of the type [ANA] (corresponding to $x = 2$), [ANNNNNA] (corresponding to $x = 6$), etc. (Fig. 1 *A*). Therefore, the relative positions of x characterized by the peaks in $\eta_{\alpha\alpha}(x)$ define the DNA repeat symmetry type. The height of the peaks (i.e., the magnitude of $\eta_{\alpha\alpha}(x)$ at a given peak position, x) defines the DNA repeat symmetry strength.

It is also interesting to note that some of the correlation functions demonstrate a nearly perfect long-range periodicity of peaks, such as $\eta_{CC}(x)$ in the case of SMAD1, in the ectoderm (Fig. 1 *B*), with peaks at $x = 3$, $x = 6$, $x = 9$, $x = 12$, etc. Such periodicity in the correlation function represents the existence of long sequence tracks of the type CNNCNNCNNC… Interestingly, in this example of SMAD1, the extent of such periodicity decreases in other developmental layers (Fig. 1 *B*).

DNA repeat symmetry selection by TFs and histone modifications in developing ES cells

We now aim to reveal a genome-wide view of DNA repeat symmetry selection by 36 TFs and five histone modifications in developing human ES cells. Fig. 1 shows examples of computed correlation functions $\eta_{\alpha\alpha}(x)$ for three TFs in different developmental lineages and in undifferentiated ES cells. Here, we selected examples from three representative groups of TFs (Fig. 1). The first group corresponds to TFs (such as SRF) that show similar $\eta_{\alpha\alpha}(x)$, regardless of developmental state (Fig. 1 *A*). The second group shows examples of TFs (SMAD1) with similar DNA repeat symmetry type (i.e., similar peak positions) but different DNA repeat symmetry strength (i.e., varying peak heights) (Fig. 1 *B*). The third group shows examples of TFs (SNAI2) for which DNA repeat symmetry type and strength across different developmental layers show clear differences (Fig. 1 *C*).

The entire set of correlation functions computed for all 36 TFs in different developmental layers is shown in Fig. S1. Although not all TFs were measured in all developmental states, the data allow a comprehensive, systems-level view of TF-DNA interaction network and dynamics during ES cell differentiation.

Interestingly, all histone modifications (H3K27ac, H3K4me1, H3K4me3, H3K9me1, and H3K27me3) show enrichment for repetitive patterns of the second group (i.e., similar DNA repeat symmetry type and varying symmetry strength across different developmental layers) (Fig. 2). In particular, for the nucleotide types A and T, for all histone

Please cite this article in press as: Goldshtein et al., Transcription Factor Binding in Embryonic Stem Cells Is Constrained by DNA Sequence Repeat Symmetry, Biophysical Journal (2020), https://doi.org/10.1016/j.bpj.2020.02.009

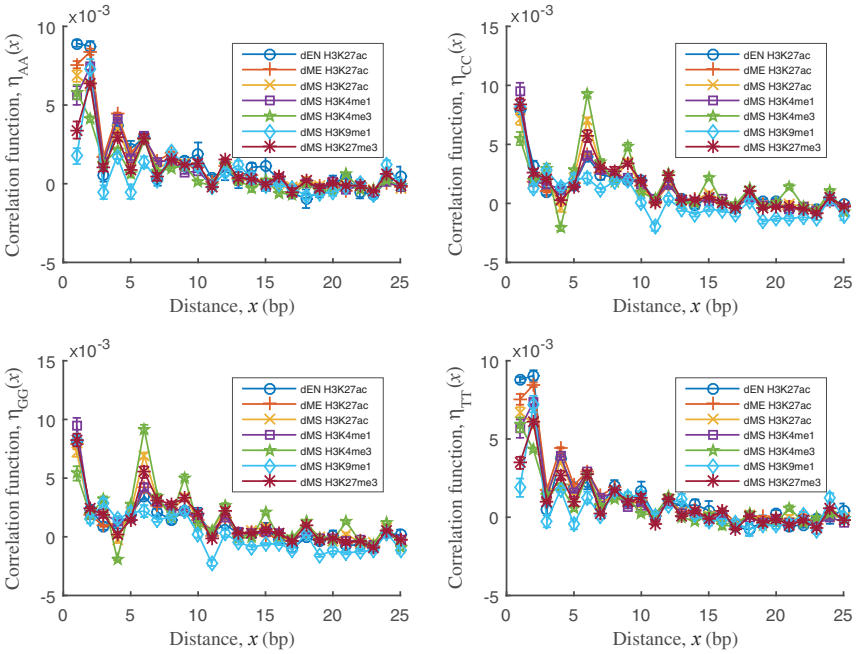

FIGURE 2 Correlation functions computed for different epigenetic histone modifications. Correlation functions in different developmental layers show similar symmetry types (i.e., similar peak positions) but varying symmetry strengths (i.e., varying peak heights). The error bars are defined similarly to Fig. 1. To see this figure in color, go online.

modifications (H3K4me3 in the dMS was an exception), the majority of correlation peaks in $\eta_{AA}(x)$ and $\eta_{TT}(x)$ are observed at $x = 2$, $x = 4$, $x = 6$, $x = 8$, $x = 10$, and $x = 12$ (Fig. 2). For H3K4me3 in dMS, there was no peak at $x = 2$ (Fig. 2). For C and G nucleotides, the peaks in $\eta_{CC}(x)$ and $\eta_{GG}(x)$ are observed at $x = 1$, $x = 6$, $x = 9$, $x = 12$, and $x = 15$ and even at larger x (Fig. 2). Therefore, based on this analysis, we suggest that histone modifications vary in the DNA repeat symmetry strength (peak heights) while keeping the DNA symmetry type (peak positions) fixed.

To compare the degree of similarity between DNA repeat symmetry preferences of the entire set of TFs and histone modifications, we show the resulting heatmaps for all TF pairs for each developmental layer (Fig. 3; Fig. S2). To construct the heatmap for each pair of TFs in a given developmental layer, we compute the Pearson correlation coefficient R between the two correlation functions $\eta_{\alpha\alpha}(x)$ characterizing the two TFs, separately for each nucleotide type α (Fig. 3, *A* and *B*). We perform this procedure for all pairs of TFs in a given developmental layer and then represent the clustered distribution of the Pearson correlation coefficients R as a heatmap (Fig. 3, *C–F*). The most notable feature of the resulting heatmap is a significant degree of clustering between different TFs in each developmental layer (Fig. 3; Fig. S2). For example, the six TFs (NANOG, OTX2, SMAD1, OCT4 (POU5F1), TCF4, and SOX2) reported as clustered in ES cells (ESc) and dMS (4) also appear almost entirely in one cluster in the heatmaps (all these TFs, except for OCT4, appear in one cluster in ESc (Fig. 3 *D*) and except for SOX2 in dMS (Fig. 3 *F*)).

Even though the cluster content undergoes transformations upon developmental transitions between different developmental layers, a high degree of clustering is apparent in all developmental layers (Fig. 3; Fig. S2). Such a high degree of clustering stems from the fact that different TFs tend to preferentially bind certain DNA repeat symmetry types (i.e., DNA sequences characterized by similar peak positions in the correlation functions, $\eta_{\alpha\alpha}(x)$) (Fig. 4; Fig. S3). Therefore, although DNA binding preferences of each TF are characterized by a unique signature represented by the entire profile of $\eta_{\alpha\alpha}(x)$, statistically, on average, there is a high degree of similarity between many TFs with respect to the selected symmetry types.

Example: DNA repeat symmetry strength statistically predicts TF-DNA binding preferences

Recently measured single-nucleotide resolution chromatin immunoprecipitation-exonuclease binding preferences for the chromatin modifier Chd2 in mouse ES cells (9) allow us to compare DNA repeat symmetry properties with the measured Chd2 occupancy profiles at each genomic location (Fig. 5). In particular, the computed DNA repeat symmetry strength (i.e., the height of the peaks in Fig. 5, *A* and *B*) constitutes a good predictor for the measured Chd2 average occupancy (Fig. 5, *C* and *D*). The correlation functions $\eta_{\alpha\alpha}(x)$ computed in a sliding window aligned with respect to transcription start sites (TSSs) show that for many peaks, the symmetry strength of DNA sequences aligned by TSSs shows an excellent correlation with the measured Chd2 occupancy (Fig. 5, *C* and *D*). In this example of Chd2, we identified the strongest predictive peaks for $\eta_{TT}(x = 2)$ (the Pearson linear correlation coefficients $R = 0.93$) and $\eta_{CC}(x = 6)$ ($R = -0.92$) (Fig. 5, *C*

Please cite this article in press as: Goldshtein et al., Transcription Factor Binding in Embryonic Stem Cells Is Constrained by DNA Sequence Repeat Symmetry, Biophysical Journal (2020), https://doi.org/10.1016/j.bpj.2020.02.009

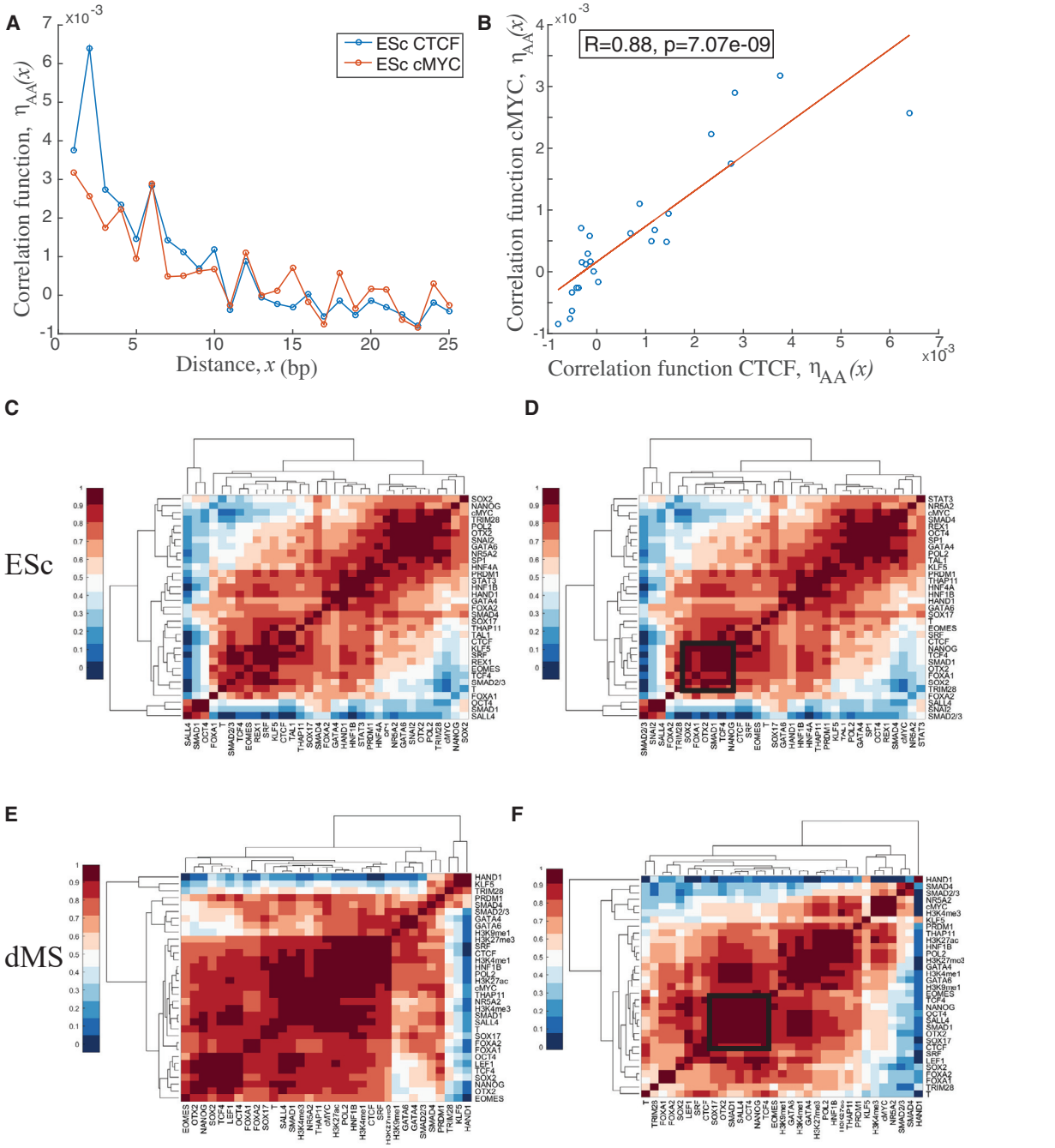

FIGURE 3 Correlation heatmaps characterizing the degree of similarity between the DNA repeat symmetry selected by TFs and histone modifications in a given developmental layer. (*A*) and (*B*) illustrate the definition of the computed Pearson correlation coefficient between each pair of TFs, using CTCF and c-MYC as examples. (*A*) The computed correlation functions $\eta_{AA}(x)$ for CTCF and c-MYC are shown. Note that in this example, out of 76,291 CTCF binding sequences, only 4235 (i.e., ~5.5%) contain the c-MYC binding motif (CACGTG). (*B*) The computed linear correlation coefficient (the Pearson correlation coefficient) between $\eta_{AA}(x)$ for CTCF (x axis) and c-MYC (y axis) is shown here. Each point in this plot represents the value of $\eta_{AA}(x)$ for CTCF (x coordinate) and c-MYC (y coordinate), respectively, for a given relative distance x. The corresponding p-value is also shown. (*C–F*) Heatmaps generated by the MATLAB (The MathWorks, Natick, MA) clustergram function represent the computed Pearson correlation coefficients for all pairs of TFs and histone modifications for the specific developmental layers, as follows: (*C*) ESc for nucleotide A, (*D*) ESc for nucleotide C, (*E*) dMS for nucleotide A, and (*F*) dMS for nucleotide C. The black frame emphasizes five out of the six TFs (NANOG, OTX2, SMAD1, OCT4 (POU5F1), TCF4, and SOX2) reported as clustered in ESc and dMS (4), which also appear in one cluster (all these TFs except OCT4 appear in one cluster in ESc, and all these TFs except SOX2 appear in one cluster in dMS) in the heatmaps (*D*) and (*F*), respectively. Only (*E*) and (*F*) (dMS) include histone modifications because ChIP-seq data for histone modifications were not reported for ESc (*C*) and (*D*). To see this figure in color, go online.

Please cite this article in press as: Goldshtein et al., Transcription Factor Binding in Embryonic Stem Cells Is Constrained by DNA Sequence Repeat Symmetry, Biophysical Journal (2020), https://doi.org/10.1016/j.bpj.2020.02.009

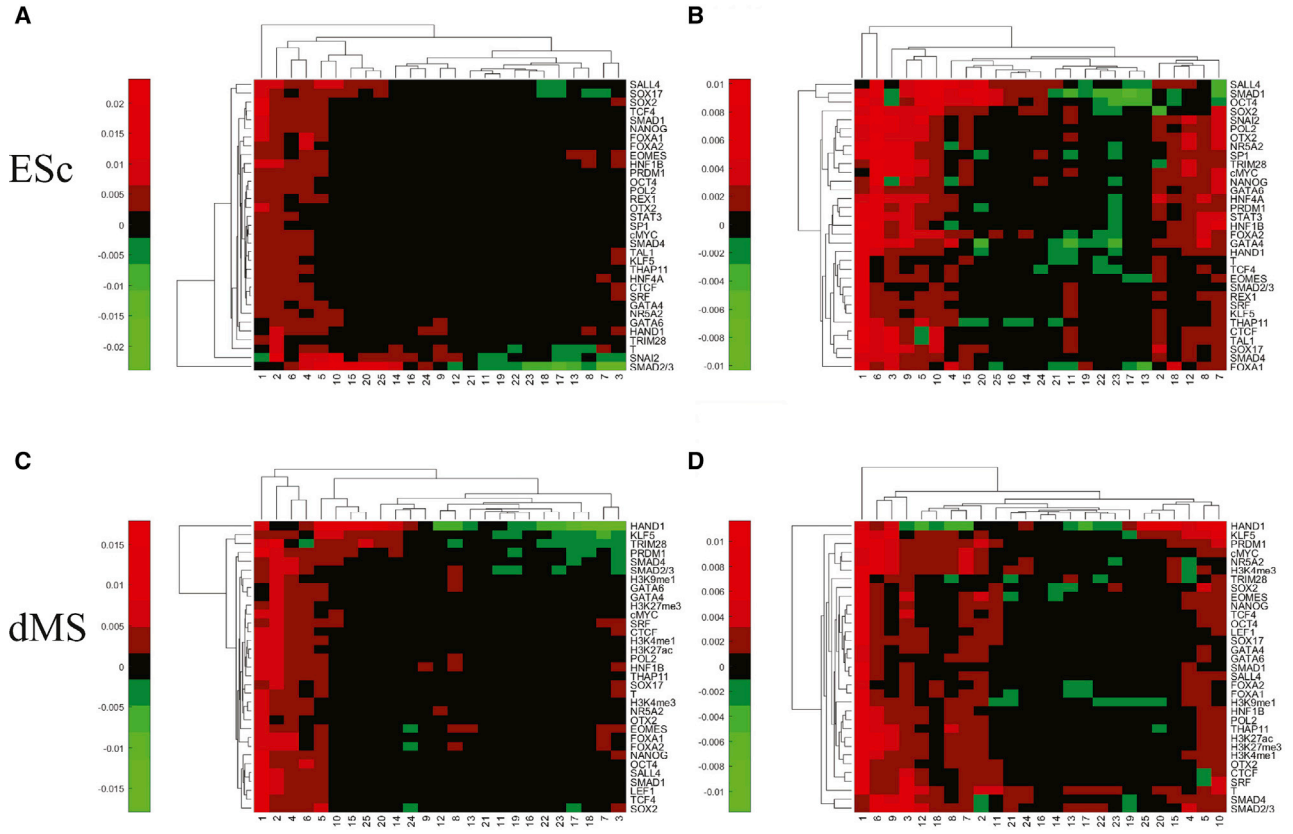

FIGURE 4 Clustered heatmap representations of the computed correlation functions $\eta_{\alpha\alpha}(x)$ for different TFs and histone modifications in a given developmental layer. TFs (y axis) are clustered by similarities of the symmetry strength (i.e., the magnitude of $\eta_{\alpha\alpha}(x)$, represented by the shown *color-code*) at different relative distances x (x axis) in a given developmental layer. The heatmaps reveal which symmetry types are more abundant in a given developmental layer for a given nucleotide type and are as follows: (*A*) ESc for nucleotide A, (*B*) ESc for nucleotide C, (*C*) dMS for nucleotide A, and (*D*) dMS for nucleotide C. Because ChIP-seq data for histone modifications were not reported for ESc (*A* and *B*), only (*C*) and (*D*) (dMS) include histone modifications. To see this figure in color, go online.

and *D*). A positive correlation means that Chd2 occupancy is higher at DNA sequences enriched in repetitive patterns of the type [ANA] and [TNT]. A negative correlation means that Chd2 occupancy is lower at DNA sequences enriched in repetitive sequence patterns of the type [CNNNNNC] and [GNNNNNG]. Other peaks of the correlation functions (in addition to $\eta_{\mathrm{TT}}(x = 2)$ and $\eta_{\mathrm{CC}}(x = 6)$) also demonstrate a statistically significant, yet weaker, predictive power.

Reporter assay measurements of gene expression induced by repetitive sequence elements

To provide a more direct validation for our findings, we performed in vivo reporter assay measurements in ES cells (Fig. 6). Briefly, using a luciferase reporter assay in mouse ES cells, we measured the ability of a few candidate DNA sequences, both genomic and computationally designed, to modulate gene expression (Fig. 6 *A*). The total length of each recognition sequence is 70 bp, and it contains the specific TFBM in the center (Fig. 6 *B*). To reduce the influence of specific TF-DNA binding in the flanking regions, we used an extended specific motif containing 10 bp (GTCACGTGAC), which is larger than the core-specific motif containing 6 bp (CACGTG). Three different flanking contexts were tested: first, the natural genomic sequences; second, the computationally designed sequences; and third, sequences with randomized flanks. All three groups of sequences have identical average GC-content (Fig. 6 *B*). Genomic sequences enriched in repetitive sequence patterns of the type [CNNC], [GNNG], and poly(A)/poly(T) were selected based on ChIP-seq binding peaks for c-MYC (Fig. 6 *B*; Fig. S1). We have also computationally designed sequences enriched in such repeats. The stochastic design procedure maximizes the number of selected repetitive sequence elements in the flanking regions around the specific TFBM (GTCACGTGAC), keeping the GC-content fixed (Fig. 6 *B*). The design procedure excludes any flanking sequences containing either the entire core-specific motif (CACGTG) or such a motif with one or two mutations. Such exclusion ensures that flanking sequences remain nonconsensus. A similar exclusion criterion was applied in the selection of genomic and randomized sequences.

Please cite this article in press as: Goldshtein et al., Transcription Factor Binding in Embryonic Stem Cells Is Constrained by DNA Sequence Repeat Symmetry, Biophysical Journal (2020), https://doi.org/10.1016/j.bpj.2020.02.009

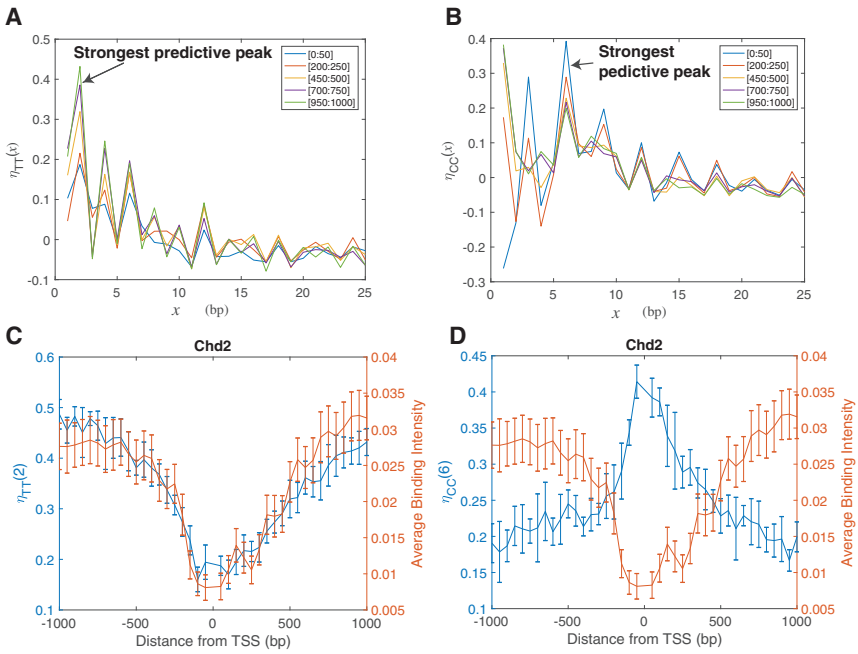

FIGURE 5 Example illustrating how the DNA symmetry strength enables us to predict TF-DNA binding preferences for the Chd2 chromatin modifier in mouse ES cells. Here, we used the chromatin immunoprecipitation-exonuclease TF-DNA binding data obtained at the single nucleotide resolution (9). (*A*) and (*B*) show the correlation functions $\eta_{TT}(x)$ (for nucleotide T) and $\eta_{CC}(x)$ (for nucleotide C), respectively, computed in a moving, sliding window with respect to the TSS for the ~14,000 aligned mouse genes. For the sliding window width, L is 50 bp. The numbers in the legend (e.g., [0:50], [200:250], etc.) represent the start and the end coordinates (with respect to the TSS) of the corresponding sliding windows, where the correlation functions were computed. The strongest predictive peaks, for $\eta_{TT}(x)$ at $x = 2$ and for $\eta_{CC}(x)$ at $x = 6$, show an excellent correlation with the measured average Chd2 binding preferences in (*C*) and (*D*), with the Pearson linear correlation coefficients $R = 0.93$ ($p = 4 \times 10^{-18}$) (*C*) and $R = -0.92$ ($p = 3 \times 10^{-17}$) (*D*), respectively. Other peaks of the correlation functions (in addition to $\eta_{TT}(x = 2)$ and $\eta_{CC}(x = 6)$) also demonstrate a statistically significant, yet weaker, predictive power. To see this figure in color, go online.

Both genomic and computationally designed sequences induced a significantly stronger transcriptional response compared to randomized sequences with exactly the same specific TFBM and identical GC-content (Fig. 6 *C*). The computed Kolmogorov-Smirnov p-values are statistically significant between genomic and random and designed and random sequence groups (the p-values are shown in the plot), yet insignificant between genomic and designed sequence groups ($p = 0.11$) (Fig. 6 *C*). These findings provide a proof of principle that both genomic and computationally designed nonconsensus repetitive sequence elements can induce a predicted functional response in ES cells, modulating the level of gene expression.

It is also important to note that randomized sequences do not induce gene expression above the control level (the Kolmogorov-Smirnov p-value between randomized and control sequence groups, $p = 0.63$). This is despite the fact that these randomized sequences contain the specific, consensus motif (Fig. 6, *B* and *C*), and this is remarkable because it demonstrates that nonconsensus repetitive elements alone are capable of significantly modulating gene expression.

Solution NMR spectroscopy: DNA repeat symmetry influences the local stability of dsDNA

In this section, we seek to establish a possible molecular mechanism responsible for the observed dependence of TF-DNA binding preferences on the presence of repetitive DNA sequence elements. The results of our statistical analysis demonstrate that many different TFs preferentially bind to DNA sequences possessing enrichment of certain well-defined repetitive patterns (Fig. 4). However, a genomic DNA sequence usually contain a combination (mixture) of several types of repeat symmetries. This fact complicates inferring a dominant molecular mechanism responsible for the effect. In our past work (10), for several TFs, we tested in vitro TF-DNA binding preferences against designed DNA sequences characterized by pure repeat symmetries, such as pure poly(A)/poly(T)/poly(C)/poly(G) or pure [αNα], where α stands for one of the four nucleotides (10). In particular, we observed that the myc-associated factor X (MAX) and c-MYC bind stronger to DNA sequences enriched in [αNα]compared to poly(α) repetitive patterns (10).

Our working hypothesis here is that DNA sequence repeat symmetry determines the local bp stability of DNA molecules. In addition, we hypothesize that in its turn, the local stability of DNA molecules influences the TF-DNA binding free energy. To provide an initial support of the first hypothesis, we performed a solution NMR experiment to compare local stability of dsDNA via imino proton resonances for several dsDNA sequences characterized by different repeat patterns with identical GC-content and containing the identical specific c-MYC/MAX binding motif (GTCACGT GAC) in the center (Fig. 7). In particular, it is generally accepted that those imino protons hydrogen-bonded in dsDNA can exchange with water protons only after the opening of the bps. Those imino protons that show substantial exchange with water cannot be observed by the NMR experiment, and therefore, the magnitude of the imino proton resonances detected by NMR experiment can reflect local DNA bp stability (16). Strikingly, we observe that DNA sequences enriched in poly(A)/poly(T)/poly(C)/poly(G) repeats (shown in *black* in Fig. 7) are less stable than DNA sequences enriched in AT/CG/TG/AG/TC/AC

Please cite this article in press as: Goldshtein et al., Transcription Factor Binding in Embryonic Stem Cells Is Constrained by DNA Sequence Repeat Symmetry, Biophysical Journal (2020), https://doi.org/10.1016/j.bpj.2020.02.009

A

B

c-Myc recognition sequence: Genomic

CATGCTCCACCCCGCCCTGCTCCGGGGCTCGTCACGTGACAGCGGCCCCGGCGGGCGAACCCTCCCCCTC
CTAAGACACAAAATGGCTGCCGGCAGCTTCGTCACGTGACCCCTTTGTGCCTTTCGGGCCCCCGAACTCC
TCCTCCTCGGAGCTTTCCGGGCTGCCCCGGGTCACGTGACTCGGACACCGCCCCTCCTCCCTTCTCCCGC
TTTTTGGGGCTAGGATTGGTTGGACCCTCCGTCACGTGACATGCAGTCGCCGGATCGCCTTTGAGTCTCG
GCTCACAGCGGGTCCTGCAGGCAGAACGCGGTCACGTGACTGATGTCCCTCCGGCTGGCAGCTAGCAATA
CTTCTGGAATTCTCCTTCCCCAGCCTCCCAGTCACGTGACTGCTTAATCTCCTCCCTCTGTGGCTCTTTC

c-Myc recognition sequence: Designed

AATTTTTCCGCCGCCGCCGCCGCCGCCGCCGTCACGTGACAAACCGCCGCCGCCGCCGCCGCCGCCGTTG
AAAAAAAAAGTTTTTGCCGCCGCCGCCGCGGTCACGTGACTTTTTTTTCCGCCGCCGCCGCCGCCGCCAA
TTTTTTTGACGCCGCCGCCGCCGCCGCCGGTCACGTGACAACCGCCGCCGCCCCCCCCCCTTTTTTC
GCGGCGGCGGCGGCGAAATTTTTTTTTTTCGTCACGTGACGCCGCCGCCGCGGCGGCGAAAATTTTTTTT
GTTGGCGGCGGCCGCCGCCGCCGAAAAAAATGTCACGTGACTTTTTTTGCGGCGCCGCCGCCGCAAAAAAG
CCGCCGCCGCCCCCCCCCTTTTTTTTAAAATGTCACGTGACAACGCCGCCGCCGCCTTTTTTTTTTTTTCC

c-Myc recognition sequence: Randomized

GCGCCCCCAGTACCGTCTTCTTCCCCAGCGGTCACGTGACACCCCGGCCAGCCGGCCCTGGCCGGGCCCC
ACTGGTACCTCCCGGAGCGATGACAGTCCTGTCACGTGACTGGTTACAACCCCCCTAGCCCGCTAGCCTT
CGCCCTGCTACCCTATCCCTGCGGCCCCCAGTCACGTGACTGGGCCTCCCTGCGCTCTCGGCCCTTTCCG
TTCTGCGCTGGTTATCCCTTTCGGTGGGGGGTCACGTGACTGTCACACTGATCCTTTGCGAATGCAGGGC
ATCCTGGTAGGGGGTACACCGCACAGCGGAGTCACGTGACCAAGGTCCGGGCCATTAACTGTCCGGCCCT
GACTCTCATATCCTGCTTGTTACTTCCCCGGTCACGTGACCATCTCTACTCGGTCCTTGTCCTCTCCTCC

C

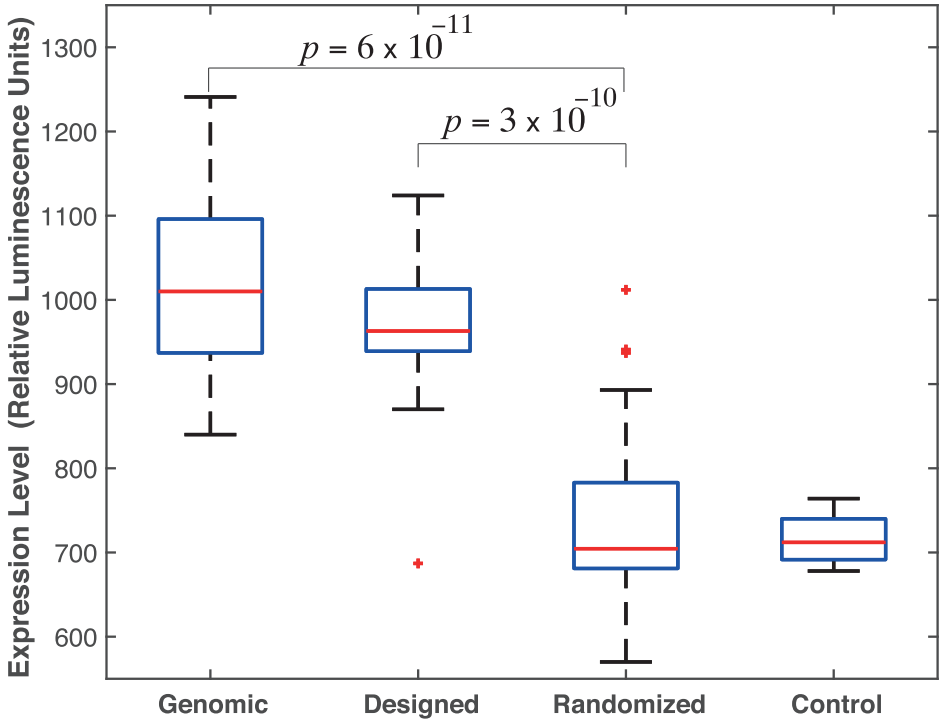

FIGURE 6 Construction of plasmids for in vivo reporter assay measurements and the results of these measurements. (*A*) The luciferase reporter assay measures the expression induced by the presence of different c-MYC recognition sequences containing the extended specific, consensus c-MYC recognition motif containing 10 nucleotides (GTCACGTGAC), surrounded by nonconsensus flanking sequence elements. The total length of each recognition sequence is 70 bp. (*B*) Three different flanking contexts were tested: first, the natural genomic sequences; second, the computationally designed sequences; and third, the sequences with randomized flanks. All three groups of sequences have identical average GC-content. Designed sequences were generated using a

(legend continued on next page)

Please cite this article in press as: Goldshtein et al., Transcription Factor Binding in Embryonic Stem Cells Is Constrained by DNA Sequence Repeat Symmetry, Biophysical Journal (2020), https://doi.org/10.1016/j.bpj.2020.02.009

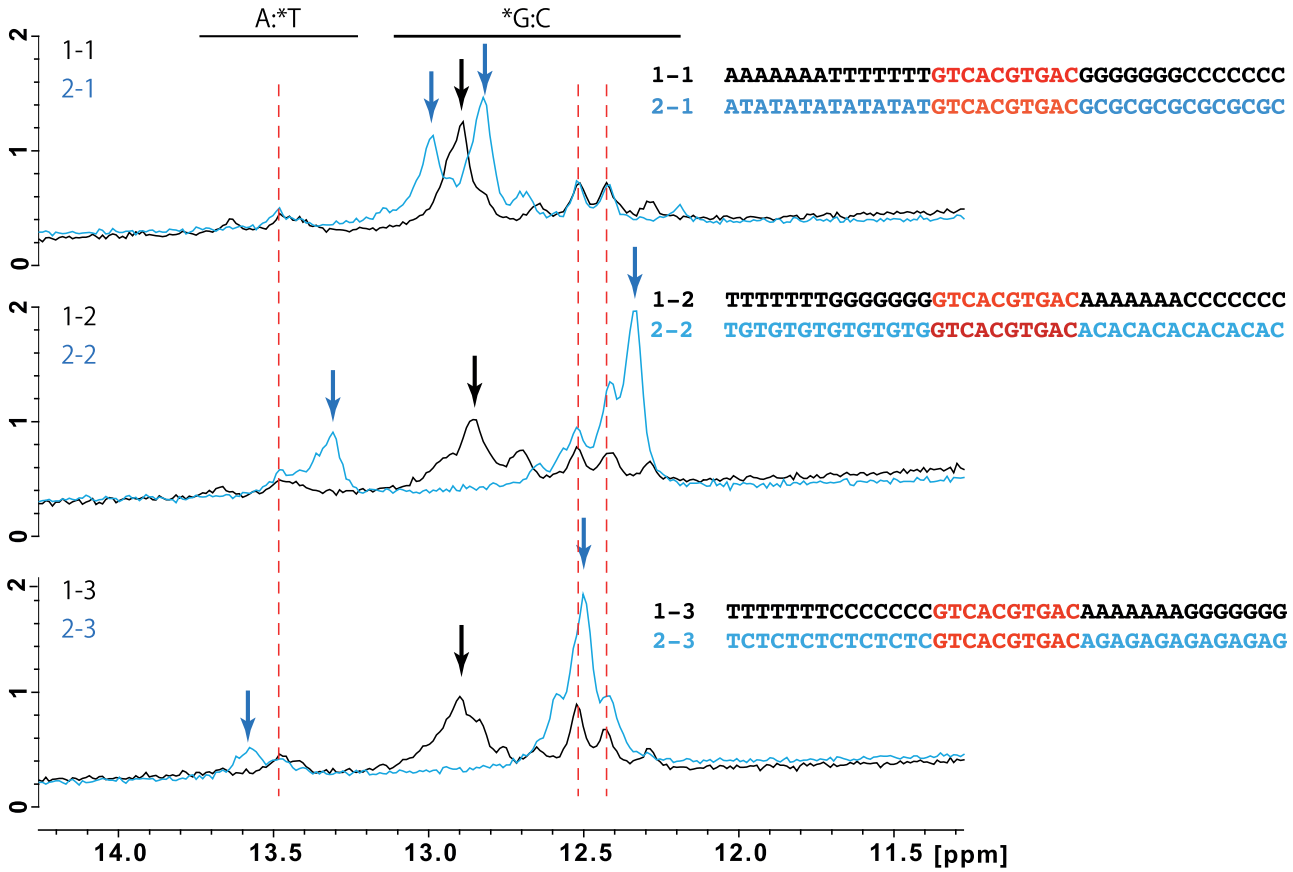

FIGURE 7 Solution NMR spectroscopy experiment measuring imino proton exchange in dsDNA with different repeat symmetry types. All shown DNA sequences are characterized by the identical GC-content and contain the identical specific c-MYC/MAX binding motif (highlighted in *red*). A higher signal indicates a stable bp (slow exchange of the imino proton with water proton), whereas a lower signal indicates an unstable bp (rapid exchange of the imino proton with water proton). The results of this experiment support our working hypothesis that DNA sequence repeat symmetry determines the local stability of dsDNA. The red dotted line represents the signal from the common specific c-MYC/MAX binding motif. The pairs of DNA sequences shown in the graph are color-coded and also marked by the numbers 1-1, 2-1; 1-2, 2-2; and 1-3, 2-3, respectively. Here, it is clearly seen that DNA sequences enriched in poly(A)/poly(T)/poly(C)/poly(G) repeats (*black*) are less stable than DNA sequences enriched in AT/CG/TG/AG/TC/AC repeats (*blue*). To see this figure in color, go online.

repeats (shown in *blue*), as indicated by the stronger imino proton resonance intensity (Fig. 7). This provides initial support for our working hypothesis.

DISCUSSION

Here, we showed that the DNA sequence repeat symmetry represents an important sequence signature providing the specificity for TF binding in developing human ES cells. In particular, binding preferences of TFs toward genomic DNA can be quantitatively characterized by two statistically defined parameters: the DNA repeat symmetry type and the DNA repeat symmetry strength, respectively. The first parameter (symmetry type) corresponds to the positions of the peaks of the correlation functions $\eta_{AA}(x)$, $\eta_{TT}(x)$, $\eta_{CC}(x)$, and $\eta_{GG}(x)$ with respect to the relative distance x between the two given nucleotides. The second parameter (symmetry strength) corresponds to the magnitude of the heights of these correlation functions at particular peak positions, x. The correlation functions, $\eta_{\alpha\alpha}(x)$, are defined

stochastic computational procedure, maximizing the number of repetitive patterns of the type [CNNC], [GNNG], poly(A), and poly(T). These patterns were identified as enriched in DNA sequences bound by c-MYC in ES cells, analyzing ChIP-seq binding peak data (4). The design procedure excludes any flanking sequences containing either the core specific motif (CACGTG) or such a motif with one or two mutations. Such exclusion ensures that flanking sequences remain nonconsensus. A similar exclusion criterion was applied in the selection of genomic and randomized sequences. (*C*) We measured luciferase luminescence for each type of c-MYC recognition flanking contexts for plasmids transfected in mouse ES cells. Five repeated measurements were performed for each of the sequences shown in (*B*). The computed Kolmogorov-Smirnov p-values are statistically significant between genomic and random and designed and random sequence groups (the p-values are shown in the plot) yet are insignificant between genomic and designed sequence groups ($p = 0.11$). The "Control" box corresponds to the case in which the entire c-MYC recognition sequences is absent from the plasmid shown in (*A*) and only the minimal promoter is retained. To see this figure in color, go online.

Please cite this article in press as: Goldshtein et al., Transcription Factor Binding in Embryonic Stem Cells Is Constrained by DNA Sequence Repeat Symmetry, Biophysical Journal (2020), https://doi.org/10.1016/j.bpj.2020.02.009

in such a way that different genomic positions can be quantitatively compared, despite the fact that the average GC-content is varying along the genome.

The most striking conclusion that emerges from the DNA repeat symmetry analysis of bound genomic regions is the fact that despite the enormous sequence complexity of the TF-DNA binding landscape, only a few DNA repeat symmetry types are selected by the majority of TFs in all developmental layers (Fig. 4; Fig. S3). For example, the correlation functions $\eta_{AA}(x)$ and $\eta_{TT}(x)$, for the majority of TFs, are peaked at $x = 1$, $x = 2$, $x = 4$, and $x = 6$ (Fig. 4; Fig. S3). The correlation functions $\eta_{CC}(x)$ and $\eta_{GG}(x)$, for the majority of TFs, show peaks at $x = 1$, $x = 3$, $x = 6$, $x = 9$, and $x = 12$ (Fig. 4; Fig. S3). In addition, yet to a lesser degree, additional repeating patterns are observed at $x = 5$ and $x = 10$, for the C and G nucleotide types (Fig. 4; Fig. S3). This means that statistically, on average, only a few dominant repetitive DNA sequence patterns significantly contribute to establishing the genome-wide TF-DNA binding profile. However, despite the observed similarity, DNA binding preferences of each TF are characterized by a unique signature represented by the entire profile of $\eta_{\alpha\alpha}(x)$, i.e., the entire profile of the DNA symmetry type (peak positions along x) and the DNA symmetry strength (peak heights).

The DNA repeat sequence symmetry selection by five different histone modifications in different developmental layers appears to be robust with respect to the DNA repeat symmetry type (i.e., peak positions) (Fig. 2; Fig. S4). The specificity for different histone modifications toward genomic DNA in human ES cells is thus tuned by tuning the DNA repeat symmetry strength (i.e., peak heights) while keeping the DNA repeat symmetry type (i.e., peak positions) nearly invariant for all histone modifications (Fig. 2). Overall, our findings suggest that DNA repeat symmetry significantly contributes toward establishing the genome-wide TF-DNA binding profile in developing human ES cells.

The central question remains open: what molecular mechanism is responsible for the observed genome-wide dependence of TF binding preferences on DNA repeat symmetry? We have suggested in the past, and validated experimentally for a number of TFs, that certain repetitive DNA sequence elements exert the entropy-dominated TF-DNA binding free energy landscape (5,10,15,17). The main difficulty here stems from the fact that in ES cells (as well as in any other type of cells), there are additional factors, besides the DNA sequence alone, that influence TF-DNA binding preferences. Such factors include, first of all, protein-protein interactions affected by protein expression levels, DNA and histone epigenetic modifications, nucleosome occupancy, and three-dimensional chromatin folding, which is affected by the number of epigenetic factors (9,18–25). In addition, the ChIP-seq data that we used in our analysis provide only a static snapshot of the dynamic TF-DNA interaction network, leaving important kinetic effects outside the scope of such measurements.

To provide a more direct validation of our central working hypothesis that certain nonconsensus, repetitive DNA sequence elements can directly influence TF-DNA binding, we performed initial in vivo reporter assay measurements for several candidate DNA sequences enriched with repeats (Fig. 6). We showed that both genomic and computationally designed nonconsensus repetitive sequence elements can induce a predicted functional response in ES cells, modulating the level of gene expression (Fig. 6 *C*). Remarkably, sequences containing specific TFBM and randomized flanking regions (i.e., randomized sequence group) do not induce gene expression above the control level (Fig. 6, *B* and *C*), demonstrating that nonconsensus repetitive elements alone are capable of significantly modulating gene expression. We also put forward a working hypothesis that DNA sequence repeat symmetry influences the local stability of dsDNA, which in turn influences the TF-DNA binding strength (Fig. 7). Further experimental development of these ideas, on a higher-throughput level, will constitute the subject of our future work.

SUPPORTING MATERIAL

Supporting Material can be found online at https://doi.org/10.1016/j.bpj.2020.02.009.

AUTHOR CONTRIBUTIONS

M.G., M.M., G.D., M.I., K.T., E.M., O.R., and D.B.L. designed the project. M.G., G.D., and D.B.L. developed theoretical models and performed the data analysis. M.M., E.M., and O.R. designed, performed, and analyzed in vivo reporter assay measurements in ES cells. M.I. and K.T. designed and performed NMR measurements. M.G., M.M., G.D., M.I., K.T., E.M., O.R., and D.B.L. wrote the manuscript.

ACKNOWLEDGMENTS

We acknowledge the generous support and professional help of Prof. Smadar Cohen and the Regenerative Medicine and Stem Cell Research Center at Ben-Gurion University of the Negev. We thank Dr. Vladimir Teif for critical reading of the manuscript.

This work was supported by the Israel Science Foundation under the grant numbers 1140/17 and 2143/19 (to E.M. and O.R.). We also would like to acknowledge the funding of the European Research Council Starting Grant number 715260.

REFERENCES

1. Ptashne, M., and A. Gann. 2002. Genes & Signals. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

2. von Hippel, P. H. 2007. From "simple" DNA-protein interactions to the macromolecular machines of gene expression. *Annu. Rev. Biophys. Biomol. Struct.* 36:79–105.

3. von Hippel, P. H. 2014. Increased subtlety of transcription factor binding increases complexity of genome regulation. *Proc. Natl. Acad. Sci. USA.* 111:17344–17345.
4. Tsankov, A. M., H. Gu, …, A. Meissner. 2015. Transcription factor binding dynamics during human ES cell differentiation. *Nature.* 518:344–349.
5. Goldshtein, M., and D. B. Lukatsky. 2017. Specificity-determining DNA triplet code for positioning of human preinitiation complex. *Biophys. J.* 112:2047–2050.
6. Slattery, M., T. Zhou, …, R. Rohs. 2014. Absence of a simple code: how transcription factors read the genome. *Trends Biochem. Sci.* 39:381–399.
7. Gordân, R., N. Shen, …, M. L. Bulyk. 2013. Genomic regions flanking E-box binding sites influence DNA binding specificity of bHLH transcription factors through DNA shape. *Cell Rep.* 3:1093–1104.
8. Rossi, M. J., W. K. M. Lai, and B. F. Pugh. 2018. Genome-wide determinants of sequence-specific DNA binding of general regulatory factors. *Genome Res.* 28:497–508.
9. de Dieuleveult, M., K. Yen, …, M. Gérard. 2016. Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells. *Nature.* 530:113–116.
10. Afek, A., J. L. Schipper, …, D. B. Lukatsky. 2014. Protein-DNA binding in the absence of specific base-pair recognition. *Proc. Natl. Acad. Sci. USA.* 111:17140–17145.
11. Xin, B., and R. Rohs. 2018. Relationship between histone modifications and transcription factor binding is protein family specific. *Genome Res.* 28:321–333.
12. Le, D. D., T. C. Shimko, …, P. M. Fordyce. 2018. Comprehensive, high-resolution binding energy landscapes reveal context dependencies of transcription factor binding. *Proc. Natl. Acad. Sci. USA.* 115:E3702–E3711.
13. Esadze, A., C. A. Kemme, …, J. Iwahara. 2014. 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 Res.* 42:7039–7046.
14. Pugh, B. F., and B. J. Venters. 2016. Genomic organization of human transcription initiation complexes. *PLoS One.* 11:e0149339.
15. Sela, I., and D. B. Lukatsky. 2011. DNA sequence correlations shape nonspecific transcription factor-DNA binding affinity. *Biophys. J.* 101:160–166.
16. Guéron, M., M. Kochoyan, and J. L. Leroy. 1987. A single mode of DNA base-pair opening drives imino proton exchange. *Nature.* 328:89–92.
17. Imashimizu, M., A. Afek, …, D. B. Lukatsky. 2016. Control of transcriptional pausing by biased thermal fluctuations on repetitive genomic sequences. *Proc. Natl. Acad. Sci. USA.* 113:E7409–E7417.
18. Eberharter, A., and P. B. Becker. 2002. Histone acetylation: a switch between repressive and permissive chromatin. Second in review series on chromatin dynamics. *EMBO Rep.* 3:224–229.
19. Javaid, N., and S. Choi. 2017. Acetylation- and methylation-related epigenetic proteins in the context of their targets. *Genes (Basel).* 8:196.
20. Quina, A. S., M. Buschbeck, and L. Di Croce. 2006. Chromatin structure and epigenetics. *Biochem. Pharmacol.* 72:1563–1569.
21. Verdone, L., E. Agricola, …, E. Di Mauro. 2006. Histone acetylation in gene regulation. *Brief. Funct. Genomics Proteomics.* 5:209–221.
22. Rotem, A., O. Ram, …, B. E. Bernstein. 2015. Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state. *Nat. Biotechnol.* 33:1165–1172.
23. Teif, V. B., and A. G. Cherstvy. 2016. Chromatin and epigenetics: current biophysical views. *AIMS Biophys.* 3:88–98.
24. Fiore, C., and B. Cohen. 2016. Interactions between pluripotency factors specify cis-regulation in embryonic stem cells. *Genome Res.* 26:778–786.
25. Teif, V. B., Y. Vainshtein, …, K. Rippe. 2012. Genome-wide nucleosome positioning during embryonic stem cell development. *Nat. Struct. Mol. Biol.* 19:1185–1192.