10116–10123 *Nucleic Acids Research, 2012, Vol. 40, No. 20* *Published online 16 August 2012*
doi:10.1093/nar/gks758

DNA breathing dynamics distinguish binding from nonbinding consensus sites for transcription factor YY1 in cells

Boian S. Alexandrov[1,2], Yayoi Fukuyo[1], Martin Lange[1], Nobuo Horikoshi[1,3], Vladimir Gelev[1], Kim Ø. Rasmussen[2], Alan R. Bishop[2] and Anny Usheva[1,*]

[1]Harvard Medical School, Department of Medicine, Endocrinology, Beth Israel Deaconess Medical Center, Boston, MA 02215, [2]Los Alamos National Laboratory, Los Alamos, NM 87545 and [3]Department of Radiation Oncology, University of Texas Southwestern Medical School, Dallas, TX, 75390 USA

Received February 28, 2012; Revised July 13, 2012; Accepted July 17, 2012

ABSTRACT

The genome-wide mapping of the major gene expression regulators, the transcription factors (TFs) and their DNA binding sites, is of great importance for describing cellular behavior and phenotypic diversity. Presently, the methods for prediction of genomic TF binding produce a large number of false positives, most likely due to insufficient description of the physiochemical mechanisms of protein–DNA binding. Growing evidence suggests that, in the cell, the double-stranded DNA (dsDNA) is subject to local transient strands separations (breathing) that contribute to genomic functions. By using site-specific chromatin immunopecipitations, gel shifts, BIOBASE data, and our model that accurately describes the melting behavior and breathing dynamics of dsDNA we report a specific DNA breathing profile found at YY1 binding sites in cells. We find that the genomic flanking sequence variations and SNPs, may exert long-range effects on DNA dynamics and predetermine YY1 binding. The ubiquitous TF YY1 has a fundamental role in essential biological processes by activating, initiating or repressing transcription depending upon the sequence context it binds. We anticipate that consensus binding sequences together with the related DNA dynamics profile may significantly improve the accuracy of genomic TF binding sites and TF binding-related functional SNPs.

INTRODUCTION

The variations in gene expression regulation are fundamental features that predetermine the tissue specificity and phenotypic diversity (1,2). A significant part of this regulation is due to the variety of the TFs and their binding sites. Recently, significant progress has been made in developing tools for their recognition and genomic annotation. However, the insufficient understanding of the genomic DNA-relevant binding criteria limits the accuracy of the description of TF binding sites. Explaining the effects of the direct recognition sequence (DRS) on the binding, and the role of the flanking regions, as well as the importance of small variations in these regions, including single-nucleotide polymorphisms (SNPs), still remains challenging (3).

The strength and specificity of TF–DNA interactions are achieved through a complex interplay of the direct points-of-contact (i.e. the DRS) and induced fit (indirect recognition) events (4,5). The indirect recognition events are sensitive to the flanking sequence and variations in the local ion-environment (4). The analysis of the indirect recognition patterns is complicated by the discrepancies between the TF binding sites sets derived from high-throughput enrichments with relatively short oligonucleotides, and the genomic sequences derived from chromatin immunoprecipitation (ChIP). Importantly, the binding data derived from short oligonucleotides lack responsiveness to SNPs in the flanking sequences (3).

Our extensive experimental and computational efforts point to sequence-specific DNA breathing dynamics, that is composed of local transient openings of the double helix (aka DNA bubbles), as an essential factor for

*To whom correspondence should be addressed. Tel: +1 617 632 0522; Fax: +1 617 632 2927; Email: ausheva@bidmc.harvard.edu

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

transcriptional activity in mammalian promoters (6–8). However, the links between TF binding and local DNA breathing, are underappreciated, and the role of the flanking sequences and the SNPs variations in the cellular TF–DNA binding is not clear.

To prove the importance of DNA breathing and the role of the flanking sequences in genomic TF binding, we have compared dynamical characteristics of various genomic binding sites for the ubiquitous Zn-finger type (His2Cys2) TF YY1. Here we report binding predictions, based on our extended Peyrard–Bishop–Dauxois model (EPBD) related Langevin molecular dynamics (LMDs) simulations (6,9) of the local DNA breathing dynamics of YY1 binding sites, together with genomic site-specific chromatin immunoprecipitation (ssChIP) and gel shift experimental tests of these predictions. We find that YY1 binding in cells depends not only on the availability of direct points-of-contact with DNA, as previously reported (10), but also on a particular DNA breathing profile, that depends on the flanking sequence and can be altered by single SNPs outside of the DRS. We report a significant correlation between DNA propensity for breathing and YY1 binding in cells. We demonstrate that the DNA breathing profile characteristics distinguish the sequences that bind to YY1 in cells from the nonbinding sequences. We demonstrate experimentally that the SNPs at the flanks that have an effect on the DRS local breathing also cause variations in YY1 binding. Further, the data indicate that the flanking DNA breathing profile is important for YY1 binding in cells. Our simulations of binding and nonbinding sites and the ssChiP data strongly suggest that YY1 binding is regulated by DNA breathing dynamics that in turn are regulated by the flanks and SNPs outside of the DRS.

EXPERIMENTAL PROCEDURES

Site-specific chromatin immunoprecippitation

ChIP is conducted following the protocol of the ChIP kit supplier (Upstate Laboratory, NY) with two different YY1 antibodies: chicken polyclonal (BIDMC, Usheva laboratory) and mouse monoclonal anti-YY1antibodies (Santa Cruz, CA) together with preimmune negative control chicken IgY and mouse IgG1 isotype control that does not react with any antigen (ab27479 Abcam). After sonication chromatin is subjected to restriction enzyme cleavage with HaeII and RsaI or RsaI and MseI before immunoprecipitation to separate the binding sites. Protein AG sepharose (Pierce) is used to capture the mouse and the IgY-anti IgY antibody-bound fragments. The immunoprecipitated PLG promoter (NT_025741.15), Homo sapiens chromosome 6 fragments are identified by Polymerase chain reaction (PCR) with the following primers: 5′-GGCAGAGGGTCTCGTC-3′ and 5′-TATT TCCTCCCCTCTTC-3′ for the HaeII–RsaI fragment, respectively, 5′-AGACTAATTGCGAGAG-3′ and 5′-CA GCAGTGCCCAGAAAG-3′ for the RsaI–MseI fragment of the human PLG promoter fragment that is located between positions −255 bp upstream and +70 bp downstream of the TSS. The 1986 bp genomic region (chr6:5200425-5202412, NT_034880), spanning the predicted *FARS2/LYRM4* promoters was PCR amplified from DNA samples of individuals homozygous for the major or minor alleles of promoter SNPs (verified by sequencing) (11). The fragment is cloned into the pGL3 basic Firefly luciferase reporter vector in two versions (the major versus the minor SNP allele haplotypes). pGL3 vectors with inserts are transiently transfected in HeLa cells. ChIP is conducted following the protocol of the ChIP kit. After sonication chromatin and the pGL3 inserts are subjected to restriction enzyme cleavage with EcoN1, Eco 01091 and HindIII or Bsp E1 before immunoprecipitation to separate the LYRM4/SNPs (5′ . . . GCCA TTTTGG . . .) and LYRM/-1424 (5′ . . . TCCATCTTTCT CCG. . .) YY1 binding sites that are finally PCR amplified with site-specific primers. The Homo sapiens chromosome 22 ssChIP fragments are located in NT_011520.12. They are identified by PCR with the following primers: A22 5′- GGGCTACAAGTGCGCATCA CCA-3′, 5′-GC ACTTTGGGAGGTCGAGGTGG-3′; B22 5′-CCCTGA AGTCCTAGGT GGGC-3′, 5′-TGTCTTAGGTCCCCG TGTGCCAA-3′; C22 5′-CTCAGCCCCACATGGC AA GGG-3′ and 5′-GCTCCAAATGGATGTGGCAGGGA -3′; D22 5′-GAGCCTCCCAAG GCTGTCCCT-3′, 5′-C ACCAGCCTCGATGGGCCAC-3′. After sonication chromatin is subjected to restriction enzyme cleavage with BbvI, FokI for A22; BamHI, FokI for B22; HinfI, Fok1 for C22; BglI, AvaI for D22.

Computer simulations

We used the EPBD DNA model, which is an extension of the Peyrard–Bishop–Dauxois nonlinear model (9) that includes inhomogeneous stacking potentials (6).

The EPBD model is a quasi-two-dimensional nonlinear model that describes transverse opening motions of the complementary strands of DNA, while distinguishing the two sides (v_n—left and u_n—right) of the double helix. The Hamiltonian of the EPBD model is:

$$H_{EPBD} = \sum_{n=1}^{N} \Big\{ D_n(e^{-a_n(u_n-v_n)} - 1)^2 + \frac{K^u_{n;n-1}}{2}(u_n - u_{n-1})^2 + \frac{K^v_{n;n-1}}{2}(v_n - v_{n-1})^2 + \frac{\rho}{4} e^{-\beta[(u_n-v_n)+(u_{n-1}-v_{n-1})]} \Big(\sqrt{K^u_{n;n-1}}(u_n - u_{n-1}) - \sqrt{K^v_{n;n-1}}(v_n - v_{n-1}) \Big)^2 \Big\}$$

In H_{EPBD}, as in the PBD model, the independent degree of freedom, u_n (v_n), is the relative displacement from the equilibrium position of the n-th nucleotide, situated on the right (left) strand of DNA, and therefore $y_n = u_n - v_n/\sqrt{2}$ quantifies the transverse stretching of the hydrogen bonds between the complementary bases. The first term: $\sum_{n=1}^{N} D_n(e^{-a_n(u_n-v_n)} - 1)^2$ of H_{EPBD} is a sum of Morse potentials, each representing the combined effects of the hydrogen bonds between the complementary bases and the electrostatic repulsion of the backbone phosphates (9). The sum is over the N base pairs of the sequence. The parameters D_n and a_n depend on the nature of the

n-th base pair: A–T versus G–C, (i.e. two hydrogen bonds versus three). The second term of the Hamiltonian represents the part of the stacking interactions between two consecutive nucleotides that influence the transverse stretching motion. The phenomenological parameters β and ρ, were introduced for the first time in (12) to fulfill the requirement for a sharp DNA melting transition. Thus, $\rho = 0$ corresponds to a purely harmonic stacking potential, while $\rho > 0$ corresponds to a decreasing stacking interaction when a neighboring base pair is stretched out of equilibrium. This nonlinear stacking dependence leads to a cooperative to the melting process. The concrete values in our Hamiltonian: $\beta = 0.35$ and $\rho = 2$ were determined based on DNA melting experiments (13).

The exponential term effectively decreases stacking interaction when one of the nucleotides is displaced away from its equilibrium position in the double helix (14,15). The stacking force constants $K^{u}_{n;n-1}(K^{v}_{n;n-1})$ depend on the nature of the base pair; the $u(v)$ indices denote the right (left) DNA strand. The inhomogeneous dinucleotide stacking force constants were determined in (6) by fitting simulation results to UV-melting curves of DNA oligos. The stacking force-constant dependence allows treatment of DNA with single base-pair resolution. The LMD computer simulations are based on the EPBD model as previously described (6,7,16). The simulations are used to generate equilibrium quantities. LMD simulations generates a number of trajectories that provide information related to the averaged bubble duration [ps], and the probability of the existence of bubbles of a certain length L [bp], beginning at base pair n, and with amplitude of the opening larger a given threshold A [Å]. The numerical data are graphically presented in MATLAB.

We emphasize that the EPBD Langevin dynamics is not necessarily a phenomenological representation of DNA's full complexity. For example, large-scale effects and the resulting dynamical time scales, caused by three-dimensional conformations and torsions of DNA molecule, are absent because of the quasi-two-dimensional character of the EPBD model. Nevertheless, the model does give a qualitative description of the dynamical issues involved in the relative propensity to bubble openings (including the time scale), as suggested by the excellent comparisons with various experiments.

The accounting of the relaxed helical DNA states (underwound), which are interrelated with the discussed here dynamically (vibrational) hot spots, will lead to twist-induced long-lived bubbles with considerably enhanced lifetimes (17,18).

Gel shift reactions

Gel shift reactions are assembled with recombinant wild-type YY1 protein isolated from *Escherichia coli* as previously described (19). In these experiments the bacterially produced YY1 protein was purified from bacterial endotoxins (BIDMC, Usheva Laboratory). The sequence of the YY1 biding oligonucleotides: −243YY1 inverted site—5′-ATCTCACATTTGCTGGAA, 5′-TTCCAGCAAATGTGAGAT; −18YY1 site -5′-CTCAAACATTTTGTAACG, 5′-TCCTACAAAATGTTTGAG; AAVP5—5′-CGAGCAGGATCTCCATTTTGACCGCG, 5′-CGCGGTCAAAATGGAGATCCTGCTCG; mAAVP5—5′-CGAGCAGGATCTCCATT TTGACCGCG, 5′-CGCGGTCAAAATGGACTAGGATGCTCG. *All fragments contain the flanking sequence (CCT) on both ends to minimize end wobbling*. The gel shift results are consistent between three independent experiments. The gel shift reactions are conducted at 37°C.

RESULTS

LMD simulations distinguish true YY1 binding from nonbinding sites in the human PLG promoter

YY1 knockdown in HeLa cells coincides with the accumulation of plasminogen (PLG) mRNA (not shown). The gene product regulates a wide variety of biologic responses directly related to the development of cardiovascular disease including atherosclerosis and restenosis (20). A large (800 bp) PLG promoter fragment has been isolated by YY1 ChIP and the sequence is available in the BIOBASE database (http://www.biobase-international.com). The exact YY1 binding position within the fragment, however, is not known.

By searching for YY1 consensus sequences in the BIOBASE fragment (Figure 1, panel a) TRANSFAC identifies two probable YY1 consensus binding sites that are centered at positions 243 and 18 upstream of the transcriptional start site (TSS). We assembled gel shift reactions with recombinant YY1 protein and the ^{33}P-labeled 18 bp long 243YY1 and 18YY1 PLG oligonucleotides comprising the YY1 consensus binding site to verify binding (panel b). Cold oligonucleotide with the strong P5 AAV YY1 binding site sequence (21) was used as a specific competitor, and a non-related oligonucleotide was used as a non-specific competitor in the gel shift reactions. We observed specific gel shifts with YY1 and both the 243YY1 and 18YY1 oligonucleotides (panel b, lanes 2–4 and 6–8).

We next asked if YY1 occupy these sites in HeLa cells by applying ssChIP with YY1 antibodies and amplification of the fragments with the individual binding sites (panel c). To separate and identify the YY1-binding fragments we applied genomic DNA sonication followed by restriction enzyme digestions. The fragments with bound YY1 were immunoprecipitated with YY1 specific antibody and PCR amplified with sequence specific primers. The ssChIP assay revealed YY1 complexes at the 18YY1 sequence by producing a 209 bp long PCR fragment (panel c, lane 3). Less than 2% PCR amplification was observed with the non-specific antibody (lane 2). The 243YY1 fragment, however, has not been PCR amplified (lane 3) suggesting that this position may not serve as YY1 binding site in HeLa cells although the gel shift reactions results in a strong and specifically YY1-shifted oligonucleotide. ChIP with another preparation of polyclonal anti-YY1 antibody also failed to recover the 243 sequence (not shown) further supporting the absence of YY1 binding at this position. Although the

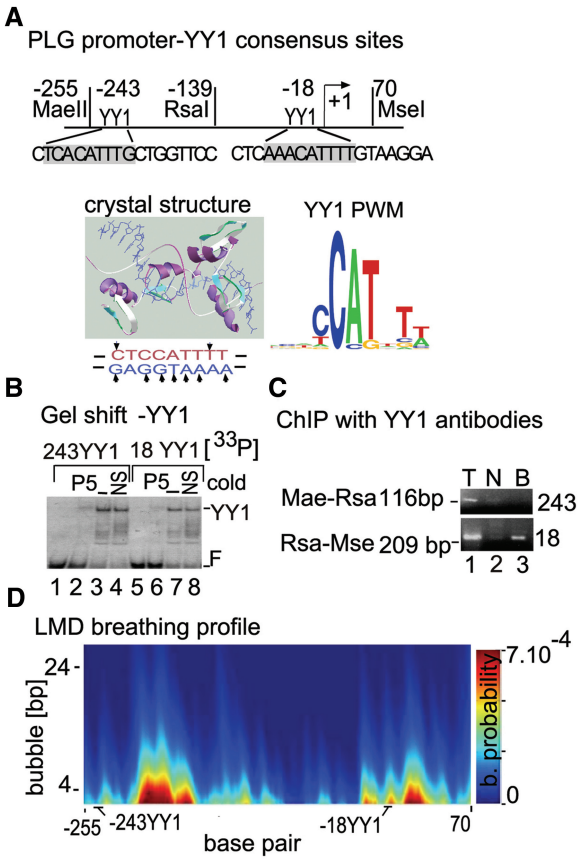

Figure 1. LMD simulations distinguish true YY1 binding in cells from nonbinding PLG promoter sites. (**A**) YY1 binding sites location relative to the TSS (+1) of the human PLG promoter (http://www.biobase-international.com). The YY1 DRS is highlighted in gray and the position of the restriction enzymes cleavage sites in the ssChIP assay are indicated at the top. The crystal structure of the YY1-P5 DNA complex (21) and the base-specific YY1 contacts on both strands (arrows) are shown below the diagram. The YY1 PWM is shown on the right. (**B**) Gel shift assays demonstrate recombinant YY1 binding to 24 bp long synthetic oligonucleotides containing the sequence 243YY1, respectively, 18YY1 (0.1 nM), as indicated above the plots. The reactions in lanes 4 and 8 received 20 nM of unrelated cold oligonucleotide competitor (NS oligo); lanes 2 and 6 received 20 nM homologous competitor cold oligonucleotide. The absence (−) of competitor oligo DNA in the reactions is indicated above the lanes. Lines 1 and 5 did not receive YY1 protein. The positions of the gel shift start (S), the free DNA (F) and YY1 complexes are indicated on the right. (**C**) ssChIP assay is used to verify genomic YY1 binding at the identified (TRANSFAC) consensus sequences. After sonication and restriction enzymes digestion the YY1 antibody-captured promoter fragments are amplified by PCR with fragment-specific primers: line 1—total DNA before antibody selection (T); line 2—pulled-down DNA with control antibody (N); line 3—pulled-down with YY1 antibody (B). The identity of the PCR-amplified fragments is shown at the left: Hae-Rsa contains the −243YY1; the 18YY1 site is located in the Rsa-Mse restriction fragment. (**D**) LMDs simulations demonstrating local DNA breathing dynamics in 305 bp long human PLG promoter fragment (http://www.biobase-international.com). The probability is determined from the lifetimes of all open states with a given length [bp] and amplitude above 3.5 Å, normalized over the time of the simulation. Probability for opening (color axis) is shown starting at specific nucleotide positions (horizontal axis), as a function of bubble length [bp] (vertical axis). Nucleotide positions are labeled relative to the TSS (TSS, +1). YY1 PWM sites are shown below the plot.

243YY1 ssChIP result is negative, it serves as a negative internal ssChIP control together with the non-immune mouse IgG antibody (line 2).

Seeking an explanation for the discrepancy between the observed bindings in cells and bindings in gel shifts (i.e. binding to short oligonucleotides) we simulated the breathing profile of the 300 bp long PLG promoter fragment (positions −255 to +50 relative to the TSS) that contain both binding sites (panel d). The EPBD-based LMD simulation data reveal a probability for bubble formation $P_{18YY1}\sim4.5\times10^{-4}$ at the −18YY1 that is similar to the previously observed breathing probability for the strong P5 AAV YY1 binding site (16). The breathing profile of the 243YY1 site, however, displays breathing probability ($P_{243YY1}\sim1\times10^{-4}$) that is 4.5 times lower than for the 18YY1 site. Importantly, the lower breathing probability at the 243YY1 ($P_{243YY1}\sim1\times10^{-4}$) coincides with the ssChIP results suggesting for absence of binding *in cells*.

Thus, only the presence of a consensus sequence that binds to the TF *in vitro* (i.e. in gel shift reactions with short oligonucleotides) is not a sufficient criterion for YY1 binding in cells. The DRS breathing profile characteristics, however, distinguish sequences that bind YY1 in ssChIP from the nonbinding PLG genomic sites. Low DNA bubble formation probability at the DRS coincides with the absence of YY1 binding at the 243 YY1 site. Further, the (ssChIP validated) active 18 YY1 site shows significantly higher probability for breathing versus the non-active 243 YY1 site.

The DRS dynamic DNA profile and YY1 binding are predetermined by the flanking sequence

To further test our understanding of how genomic YY1 binding is influenced by the flanking sequences we randomly selected YY1 PWM containing sequences on the human chromosome 22 (Figure 2). The 200 bp long A22, B22, C22 and D22 sequences were selected from promoter and intergenic regions. In each case our intention was to select for strong and nearly identical DRSs that differ only in their flanking regions. We examined these sequences with the goal to identify the effect of flanking on LMD derived DNA dynamics and YY1 binding. The identical A22 and C22 DRSs that are embedded in their specific flanking environment clearly differ in DNA dynamics ($P_{A22YY1}\sim2.2\times10^{-4}$, $P_{C22YY1}\sim4.9\times10^{-4}$, average bubble length over ~10 bp) and cellular YY1 binding in ssChIP. Similarly, B22 and D22 are with YY1-specific PWMs, but their site-specific flanking sequences, DNA dynamics profiles and YY1 binding in cells is very different. The flanking region of D22 shapes the DNA dynamics profile ($P_{D22YY1}\sim4.5\times10^{-4}$, average bubble length over ~10 bp) at the DRS that supports YY1 binding. The average DRS opening profile at B22 ($P_{B22YY1}\sim1.6\times10^{-4}$), however, is not sufficient for YY1 binding as seen by ssChIP in HeLa cells with YY1-specific antibody.

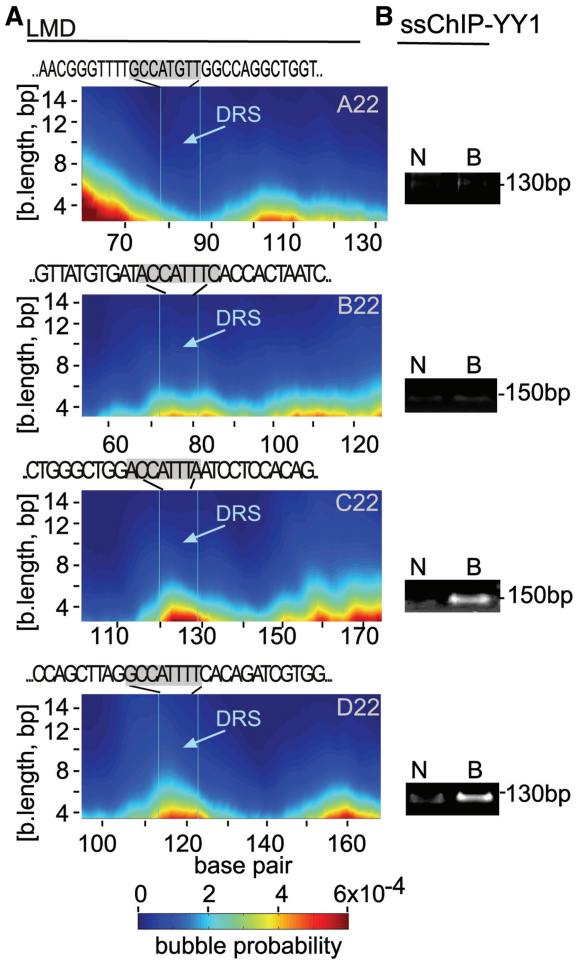

Figure 2. Fanking sequence influence the DRS breathing dynamics and genomic YY1 binding. (**A**) LMDs simulations demonstrating local DNA breathing dynamics in 150 bp long fragments A22, B22, C22 and D22 from human chromosome 22. Probability for opening (P) (color axis) at specific nucleotide positions (horizontal axis) is shown as a function of bubble length [bp] (vertical axis). Part of the flanking sequence is shown above the plots and the DRS is highlighted in gray. (**B**) ssChIP assay is used to verify genomic YY1 binding at the A22, B22, C22 and D22 DRSs. Representative PCR-amplified fragments are shown at the right of the LMD simulation plots: line 2—PCR with pulled-down DNA fragments with control antibody (C); line 3—pulled-down with YY1 antibody (B). The size of the YY1 site-specific PCR-amplified fragments is shown at the right. ssChIP results are consistent between two independent experiments.

From the results it seems that the genomic flanking regions have the strength to control DNA dynamics at the DRS and the ability of YY1 to bind.

A single SNP in the flanking can change breathing dynamics at the DRS and YY1 binding in cells

The PLG sites have nearly identical DRSs and the difference in their breathing characteristics (and hence in binding) could be attributed to the difference in their flanking sequence.

To test how the sequence variations in the flanks change DNA breathing and hence the binding, we simulated DNA breathing of various genomic fragments with

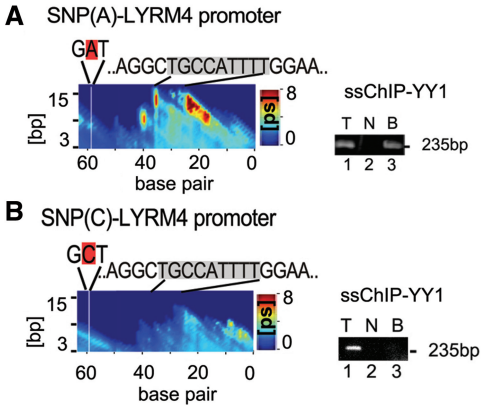

Figure 3. Long-range effect of flanking single SNPs on the DRS breathing dynamics and YY1 binding in cells. (**A**) LMDs simulations demonstrating local DNA breathing dynamics in 50 bp long human LYRM4 promoter fragment, SNP(A) haplotype (11). (**B**) Local breathing dynamics in response to the SNP(C) haplotype in the same 50 bp long LYRM4 fragment. Duration for opening (ps) (color axis) at specific nucleotide positions (horizontal axis) is shown as a function of bubble length [bp] (vertical axis). The identity of the sequence is shown above the plots and the DRS is highlighted in gray. Both sequences are identical with exception of the SNP(A)/(C) nucleotide that is highlighted in red above the plots. ssChIP assay is used to verify genomic YY1 binding at the SNP(A) and SNP(C) haplotype sequences. Representative PCR-amplified fragments are shown at the right of the LMD simulation plots: line 1—PCR with total DNA before antibody selection (T); line 2—PCR with pulled-down DNA fragments with control antibody (N); line 3—pulled-down with YY1 antibody (B). The size of the YY1 site-specific PCR-amplified fragments is shown at the right. ssChIP results are consistent between two independent experiments.

known functional SNPs in the vicinity of the YY1 consensus sites. As a main test case we selected the SNP(A) and the SNP(C) haplotype sequence of the 2 kb long schizophrenia-related fragment from human chr6 (11). TRANSFAC search identifies a YY1 consensus binding site in the vicinity of a functional SNP at the LYRM4 promoter in the considered fragment (11). The difference between the two sequences is that 30 bp upstream of the YY1 DRS the SNP(A) haplotype has adenine(A) and the SNP(C) variant has cytosine(C). The SNP(A) variant supports active LYRM4 transcription, whereas in the SNP(C) variant the gene is less active (11). To assay for YY1 binding in cells, by ssChIP, both variants were inserted into a vector plasmid for transfection into HeLa cells (Figure 3). Although a specific YY1 ssChIP fragment is produced with the SNP(A) variant, the amplification of the SNP(C) fragment does not result in a product (panel b). We next compared the breathing dynamics with the ssChIP assay data of both these sequences. Replacement of SNP(A) with the SNP(C) decreases the average bubble lifetime at the DRS more than 4-fold (panel b). The average breathing activity at the YY1 binding site in the SNP(C) variant is low with a bubble average lifetime ~2 psec, whereas in the SNP(A) variant DNA breathing results in a large bubble that extends over ~10 bp and has average lifetime ~8 psec. Both SNPs reveal long-range dynamics effects. Although there is no significant DNA dynamics in their close proximity, SNP(A) dramatically activates breathing at the DRS

that coincides with specific cellular YY1 binding in ssChIP. SNP(C) has an opposite (silencing) effect on the DRS that correspond to the absence of YY1 binding in the ssChIP.

YY1 binding in cells requires both specific breathing pattern at the DRS and at the flanking sequences

Our simulations and the ssChIP data show correlation between YY1 binding and breathing dynamic at the DRS. We sought to determine whether breathing characteristics of the DRS flanks play a role in the binding as well. We compared DNA breathing characteristics (for openings with amplitude >3.5 Å) at physiological temperature ($T = 310°K$), of the PLG and LYRM4 sites with centered YY1 DRS (Figure 4). The two ssChIP positive binding sequences (panel a) exhibit similar probability ($P = \sim 4.5 \times 10^{-4}$) for bubble formation with average bubble length over ~10 bp. Although the flanking sequences of these binding sites in cells are different, they exhibit common dynamic features: a low breathing probability ($P < 2 \times 10^{-4}$) upstream of the first Zn-finger contacts with the DNA (left side of the DRS) and a high breathing probability ($P > 4.5 \times 10^{-4}$) immediately after the last Zn-finger contacts (at the right side, immediately after the DRS). We previously reported that the strong YY1 binding site at the adeno associated P5 (P5 AAV) promoter (21) has a similar breathing pattern (16).

The DRSs in the nonbinding PLG (−243) and the LYRM/SNP(C) sequences display low breathing, which correlates with the absence of genomic binding (panel b). We identified another YY1 consensus sequence (YY1/-1424) in the 2 kb schizophrenia fragment (11) that has strong binding activity in gel shift (not shown) and no binding in ssChIP assay (panel b). Notably, the LMD simulations of the LYRM4/-1424 site demonstrate DRS breathing probability similar to the breathing at the active PLG (−18) site. This time, however, the left flank of the LYRM4/-1424 site upstream of the first Zn-finger contacts, shows a high ($P > 5 \times 10^{-4}$) bubble probability, which is in contrast to the low bubble probability ($P < 1.5 \times 10^{-4}$) at the left flank of the DRS of the active YY1 PLG (−18) site. Importantly, this opposing breathing pattern of the LYRM4/-1424 site coincides with the absence of YY1 binding in the ssChIP. To further verify the negative effect of high DNA breathing upstream of the first YY1 Zn-finger contacts (at the left side of the DRS) on the binding, we produced 80 bp long AAVP5 DNA variant with 4 mismatches at the left side—immediately before the DRS (producing in this way a constitutive DNA bubble at this location). In this case, the artificial bubble situated upstream of the first Zn-finger contacts, caused by the presence of the mismatches, yields a high breathing probability at the left, resulting in the absence of YY1 binding in gel shift (panel b, lanes 1–6). Importantly, breathing dynamics at the DRS of the mismatches-mutant and wild type remains nearly identical.

These data suggest that a specific breathing probability profile at the flanking sequences is another requirement for binding in cells. In particular (in cells) the YY1 binding coincides with low bubble probability ($P < 2 \times 10^{-4}$), upstream of the first Zn-finger contacts (at the left of the DRS), and with a high bubble probability ($P > 5 \times 10^{-4}$), downstream of the last Zn-finger (at the right side of DRS), while the breathing at the DRS should be with $P > 4 \times 10^{-4}$. Hence, the EPBD-based LMD simulations provide a specific set of breathing parameters important for the description of YY1 binding sites in cells.

DISCUSSION

The results reported here suggest a strong correlation between DNA local breathing and cellular YY1 binding. The EPBD-derived breathing profile at the DRS gives information on whether local DNA breathing seeds or inhibits the formation of particular bubbles needed for the YY1 binding. The probabilities for local DNA opening are equilibrium properties of the underlying free energy landscape, and detail information could be obtained also from the thermodynamical Poland–Scheraga model and its applications (22,23). The EPBD model is a dynamical mesoscopic model that is strongly nonlinear and with breathers (aka bubbles) (24), which constitute transient openings of the double helix. In this sense our results, although indeed thermodynamical equilibrium properties are richer because they give a qualitatively accurate idea of the dynamical information for the relative bubbles lifetimes. The last type information cannot be obtained by purely thermodynamical calculations.

The correlation between our simulations and ssChIP data clearly emphasize the role of local DRS breathing in protein recognition. YY1 is a Zn-finger protein that binds to the major grove without conformational changes in the DNA and protein itself (21). The four YY1 fingers make multiple contacts with the major groove edges of bases and most of the specific hydrogen-based contacts are restricted to the template strand. Therefore, it seems natural that the EPBD-predicted requirements for enhanced transient DNA openings at the points-of-contact region would facilitate a specific YY1 binding. In support of this suggestion is our previously published observation that a template with five mismatches creating an artificial bubble at the DRS of the P5 AAV initiator is still competent for specific YY1 binding and Initiator functionality in transcription (10,21). It is also likely that DNA breathing together with the nucleosomes instability support YY1 binding to the chromatin-organized Initiator sequences.

Such arguments could hold for other members of the large group of Zn-finger TFs and Initiators as well. Proteins that make several specific contacts with both DNA strands, however, would benefit from suppressed DRS breathing.

How exactly TFs and histones compete or cooperate in DNA binding is not precisely clear. It has been shown that the nucleosomal DNA remains fully wrapped around the histone core for only 250 ms before spontaneously unwrapping and sliding to new locations along the

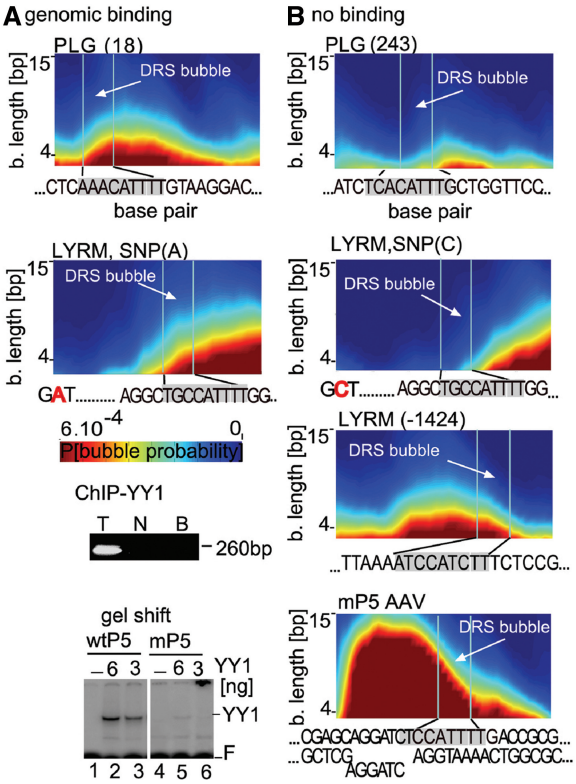

Figure 4. Breathing profile of the flanking sequences at YY1 binding and nonbinding consensus sites. LMD simulations of 50 bp long genomic fragments with centered YY1 DRS. (**A**) Fragments that bind. (**B**) Fragments that do not bind YY1 in cells. The length of the transient bubbles (in base pairs [bp]) is shown along the vertical axis. The color map represents the probability for bubble openings where the red color denotes high probability and blue color denotes low probability. The identity of the sequence for each variant is shown above the panel. The DRS is highlighted in gray. The arrow shows the bubble formation region at the DRS. ssChIP assay is used to verify YY1 binding in cells at the LYRM4-1424 consensus sequence (at the right of the panel with the breathing profile). After sonication and restriction enzyme digestion the YY1 antibody-captured LYRM4-1424 (11) fragments are amplified by PCR with fragment-specific primers: line T—total DNA before antibody selection; line N—pulled-down DNA with control antibody; line B—pulled-down with YY1 antibody as indicated above the plot. Gel shift with 80 bp long wild-type (wtP5) AAVP5 oligo and mutant (mP5) that contains 5 mismatched base pairs on the none-template strand at the left of the DRS is used to verify binding (at the left of the AAVP5 breathing probability panel). Reactions are assembled with recombinant YY1 protein (6 and 3 ng) as indicated above the plots. The reactions in lanes 1 and 4 did not receive YY1 protein (–). The positions of the free DNA (F), and YY1 complexes are indicated on the right.

DNA (25). Such spontaneous conformational fluctuations, exposures, could be envisioned as unwrapped nucleosome free state of the DNA that enable TF binding. Importantly, by treating DNA as unwrapped linear template our EPBD model accurately discriminates YY1 nonbinding from binding genomic locations.

The observation that variations in the flanking genomic sequence can exert an effect on the DRS breathing characteristics and predetermine YY1 cellular binding is not coincidental but actually plays a role in determining the position of TF binding in cells. Moreover, genomic flanking sequences seed the formation of the DRS bubbles needed for TF binding, or they may represent another yet unknown but important conformational feature of specific TF-DNA recognition (3). Our results help to better understand the mechanisms that underlie DNA–protein interactions and clearly point to the role of local DNA breathing in protein recognition.

Our simulation of PLG and the LYRM4 breathing dynamics and the ssChIP data advocate that flanking sequences can exert long-range effects on the DRS propensity for transient openings and hence on the regulation of YY1 binding in cells. We tested the idea that single SNP, further upstream of the DRS, can suppress or activate DRS breathing and YY1 binding. The striking difference in the dynamical profiles of the YY1 binding sites at the cognitive schizophrenia-related SNP(A) and SNP(C) haplotypes (11) clearly supports such a notion. Our ssChIP assay confirms that YY1 binds only to the SNP(A) template variant, strongly suggesting that a single SNP in the flanks could predetermine YY1 binding in cells. EPBD dynamic simulations revealed that, in addition to the reported absence of YY1 binding, the SNP(C) haplotype has significantly reduced dynamic activity at the DRS, consistent with our view that intact DRS and TF-specific breathing dynamics patterns are required for binding in cells.

Additionally, the data indicate that the collective openings at the flanks are also important (as the breathing profile at the DRS) for YY1 binding in cells. Specifically, our LMD simulations show that YY1 binding in cell requires a low breathing activity at the left side of the DRS, and a high breathing activity at the DRS right flank. The suppressed breathing upstream of the first Zn-finger binding contacts is likely to be needed to facilitate the initial step of the recognition, when the protein slides along non-specific DNA base pairs without losing contact. Such requirements present additional fine-tunings in TF positioning on the template defining the landing DNA location of YY1.

Together our results argue against a purely 'letter-code representations of DNA consensus sequence' view of TF binding. The reported experimental differences between presence of YY1 binding in cells, intact DRS, flanking sequence variations and SNPs can be readily explained from considerations of local sequence-specific DNA breathing dynamics. Study of other YY1 binding sites and TFs are needed to determine if there is a general dependence of TF binding in cells on the flanking and the local DRS breathing. At this stage, the available data do not allow determination of whether the breathing characteristics have only supportive or truly deterministic role as a DNA structure-related '*epigenetic*' determinant in TF binding. However, our study strongly implies that the combination of the effects of flanking sequences, functional SNPs, use of consensus sequence together with local dynamic conformational properties of DNA could revolutionize predictions of TF–DNA binding in cells. Furthermore, a combination of local breathing profile characteristics together with experimentally determined chromatin accessibility data (26), as well as accounting for the strong cooperativity among non-interacting TFs caused by their competition with nucleosomes (27,28)

could significantly simplify the cell and tissue-specific binding landscape of any TF.

FUNDING

Funding for open access charge: DOE [LANL, LDRD 20110516ECR to B.A.]; National Institutes of Health [ARRA supplement GM073911-04S to A.U.]; contract from the National Nuclear Security Administration of the US Department of Energy at Los Alamos National Laboratory [DE-AC52-06NA25396]; William F. Milton Award (to A.U.).

Conflict of interest statement. None declared.

REFERENCES

1. Skelly,D.A., Ronald,J. and Akey,J.M. (2009) Inherited variation in gene expression. *Annu. Rev. Genom. Hum. G*, **10**, 313–332.
2. Kasowski,M., Grubert,F., Heffelfinger,C., Hariharan,M., Asabere,A., Waszak,S.M., Habegger,L., Rozowsky,J., Shi,M., Urban,A.E. *et al.* (2010) Variation in transcription factor binding among humans. *Science*, **328**, 232–235.
3. Tompa,M., Li,N., Bailey,T.L., Church,G.M., De Moor,B., Eskin,E., Favorov,A.V., Frith,M.C., Fu,Y., Kent,W.J. *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.*, **23**, 137–144.
4. Sarai,A. and Kono,H. (2005) Protein-DNA recognition patterns and predictions. *Annu. Rev. Biophys. Biomol. Struct.*, **34**, 379–398.
5. Steffen,N.R., Murphy,S.D., Tolleri,L., Hatfield,G.W. and Lathrop,R.H. (2002) DNA sequence and structure: direct and indirect recognition in protein-DNA binding. *Bioinformatics*, **18(Suppl. 1)**, S22–S30.
6. Alexandrov,B.S., Gelev,V., Monisova,Y., Alexandrov,L.B., Bishop,A.R., Rasmussen,K.O. and Usheva,A. (2009) A nonlinear dynamic model of DNA with a sequence-dependent stacking term. *Nucleic Acids Res.*, **37**, 2405–2410.
7. Alexandrov,B.S., Gelev,V., Yoo,S.W., Alexandrov,L.B., Fukuyo,Y., Bishop,A.R., Rasmussen,K.O. and Usheva,A. (2010) DNA dynamics play a role as a basal transcription factor in the positioning and regulation of gene transcription initiation. *Nucleic Acids Res.*, **38**, 1790–1795.
8. Choi,C.H., Rapti,Z., Gelev,V., Hacker,M.R., Alexandrov,B., Park,E.J., Park,J.S., Horikoshi,N., Smerzi,A., Rasmussen,K.O. *et al.* (2008) Profiling the thermodynamic softness of adenoviral promoters. *Biophys. J.*, **95**, 597–608.
9. Peyrard,M. and Bishop,A.R. (1989) Statistical mechanics of a nonlinear model for DNA denaturation. *Phys. Rev. Lett.*, **62**, 2755–2758.
10. Usheva,A. and Shenk,T. (1996) YY1 transcriptional initiator: protein interactions and association with a DNA site containing unpaired strands. *Proc. Natl Acad. Sci. USA*, **93**, 13571–13576.
11. Jablensky,A., Angelicheva,D., Donohoe,G.J., Cruickshank,M., Azmanov,D.N., Morris,D.W., McRae,A., Weickert,C.S., Carter,K.W., Chandler,D. *et al.* (2011) Promoter polymorphisms in two overlapping 6p25 genes implicate mitochondrial proteins in cognitive deficit in schizophrenia. *Mol. Psychiatr.*, October 4 (doi: 10.1038/mp.2011.129; epub. ahead of print).
12. Dauxois,T., Peyrard,M. and Bishop,A.R. (1993) Dynamics and thermodynamics of a nonlinear model for DNA denaturation. *Phys. Rev. E*, **47**, 684.
13. Campa,A. and Giansanti,A. (1998) Experimental tests of the Peyrard-Bishop model applied to the melting of very short DNA chains. *Phys. Rev. E*, **58**, 3585.
14. Dauxois,T., Peyrard,M. and Bishop,A.R. (1993) Entropy-driven DNA denaturation. *Phys. Rev. E Stat. Phys. Plasmas. Fluids Relat. Interdiscip. Topics*, **47**, R44–R47.
15. Alexandrov,B.S., Wille,L.T., Rasmussen,K.O., Bishop,A.R. and Blagoev,K.B. (2006) Bubble statistics and dynamics in double-stranded DNA. *Phys. Rev. E Stat. Nonlin. Soft Matter Phys.*, **74**, 050901.
16. Alexandrov,B.S., Gelev,V., Yoo,S.W., Bishop,A.R., Rasmussen,K.O. and Usheva,A. (2009) Toward a detailed description of the thermally induced dynamics of the core promoter. *PLoS Comput. Biol.*, **5**, e1000313.
17. Jeon,J.-H., Adamcik,J., Dietler,G. and Metzler,R. (2010) Supercoiling induces denaturation bubbles in circular DNA. *Phys. Rev. Lett.*, **105**, 208101.
18. Jost,D., Zubair,A. and Everaers,R. (2011) Bubble statistics and positioning in superhelically stressed DNA. *Phys. Rev. E*, **84**, 031912.
19. Usheva,A. and Shenk,T. (1994) TATA-binding protein-independent initiation: YY1, TFIIB, and RNA polymerase II direct basal transcription on supercoiled template DNA. *Cell*, **76**, 1115–1121.
20. Plow,E.F. and Hoover-Plow,J. (2004) The functions of plasminogen in cardiovascular disease. *Trends Cardiovasc. Med.*, **14**, 180–186.
21. Houbaviy,H.B., Usheva,A., Shenk,T. and Burley,S.K. (1996) Cocrystal structure of YY1 bound to the adeno-associated virus P5 initiator. *Proc. Natl Acad. Sci. USA*, **93**, 13577–13582.
22. Poland,D. and Scheraga,H.A. (1966) Phase transitions in one dimension and the helix-coil transition in polyamino acids. *J. Chem. Phys.*, **47**, 1456.
23. Ambjörnsson,T., Banik,S.K., Krichevsky,O. and Metzler,R. (2006) Sequence sensitivity of breathing dynamics in heteropolymer DNA. *Phys. Rev. Lett.*, **97**, 128105.
24. Forinash,K., Bishop,A.R. and Lomdahl,P.S. (1991) Nonlinear dynamics in a double-chain model of DNA. *Phys. Rev. B Condens. Matter*, **43**, 10743.
25. Levitus,M., Bustamante,C. and Widom,J. (2005) Rapid spontaneous accessibility of nucleosomal DNA. *Nat. Struct. Mol. Biol.*, **12**, 46–53.
26. Kaplan,T., Li,X.-Y., Sabo,P.J., Thomas,S., Stamatoyannopoulos,J.A., Biggin,M.D. and Eisen,M.B. (2011) Quantitative models of the mechanisms that control genome-wide patterns of transcription factor binding during early Drosophila development. *PLoS Genet.*, **7**, e1001290.
27. Mirny,L.A. (2010) Nucleosome-mediated cooperativity between transcription factors. *PNAS*, **107**, 22534–22539 .
28. Miller,J.A. and Widom,J. (2003) Collaborative competition mechanism for gene activation in vivo. *Mol. Cell Biol.*, **23**, 1623–1632.