1. Extended Data

Figure #	Figure title One sentence only	Filename This should be the name the file is saved as when it is uploaded to our system. Please include the file extension. i.e.: *Smith_ED_Fig1.jpg*	Figure Legend If you are citing a reference for the first time in these legends, please include all new references in the main text Methods References section, and carry on the numbering from the main References section of the paper. If your paper does not have a Methods section, include all new references at the end of the main Reference list.

2. Supplementary Information:

A. Flat Files

Item	Present?	Filename This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. The extension must be .pdf	A brief, numerical description of file contents. i.e.: *Supplementary Figures 1-4, Supplementary Discussion, and Supplementary Tables 1-4.*
Supplementary Information	Yes	Supplementary_Information.docx	Supplementary Notes 1 - 21
Reporting Summary	Yes	ReportingSummary_Basu	
Peer Review Information	Yes	*TPRFile_Basu*	

B. Additional Supplementary Files

Type	Number If there are multiple files of the same type this should be the numerical indicator. i.e. "1" for Video 1, "2" for Video 2, etc.	Filename This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. i.e.: *Smith_Supplementary_Video_1.mov*	Legend or Descriptive Caption Describe the contents of the file
Supplementary Table	1	Supplementary_Table.xlsx	Supplementary Tables 1 - 21

3. Source Data

Parent Figure or Table	Filename This should be the name the file is saved as when it is uploaded to our system, and should include the file extension. i.e.: *Smith_SourceData_Fig1.xls,* or *Smith_ Unmodified_Gels_Fig1.pdf*	Data description i.e.: Unprocessed Western Blots and/or gels, Statistical Source Data, etc.
Source Data Fig. 1	Fig1_source_data.xlsx	Source data for Fig 1
Source Data Fig. 2	Fig2_source_data.xlsx	Source data for Fig 2
Source Data Fig. 3	Fig3_source_data.xlsx	Source data for Fig 3
Source Data Fig. 4	Fig4_source_data.xlsx	Source data for Fig 4

Deciphering the mechanical code of the genome and epigenome

Aakash Basu[1,2,a], Dmitriy G. Bobrovnikov[2], Basilio Cieza[3], Juan Pablo Arcon[4], Zan Qureshi[3], Modesto Orozco[4,5] and Taekjip Ha[2,3,6,7,a].

[1]Department of Biosciences, Durham University, Durham, DH7 3LE, UK.

[2]Department of Biophysics and Biophysical Chemistry, Johns Hopkins University School of Medicine, Baltimore, MD 21205, USA.

[3]Department of Biophysics, Johns Hopkins University, Baltimore, MD 21218, USA.

[4]Institute for Research in Biomedicine (IRB Barcelona), The Barcelona Institute of Science and Technology, 08028, Barcelona, Spain.

[5]Department of Biochemistry and Biomedicine. Universitat de Barcelona. 08028, Barcelona, Spain.

[6]Department of Biomedical Engineering, Johns Hopkins University, Baltimore, MD 21205.

[7]Howard Hughes Medical Institute, Baltimore, MD 21205, USA.

[a]To whom correspondence should be addressed (e-mail: aakash.basu@durham.ac.uk, tjha@jhu.edu)

Abstract

Diverse DNA-deforming processes are impacted by the local mechanical and structural properties of DNA, which in turn depend on local sequence and epigenetic modifications. Deciphering this 'mechanical code', i.e., this dependence, has been challenging due to the lack of high-throughput experimental methods. Here we present a comprehensive characterization of the mechanical code. Utilizing high-throughput measurements of DNA bendability via loop-seq, we quantitatively established how the occurrence and spatial distribution of dinucleotide, tetranucleotides, and methylated CpG, impact DNA bendability. We use our measurements to develop a physical model for the sequence and methylation dependence of DNA bendability. We validate the model by performing loop-seq on mouse genomic sequences around transcription start sites and CTCF binding sites. We apply our model to test the predictions of all atom molecular dynamics simulations, and to demonstrate that sequence and epigenetic modifications can mechanically encode regulatory information in diverse contexts.

Introduction

The local structural, shape, and mechanical properties of DNA can regulate diverse phenomena that require DNA deformation. Various techniques have revealed that these properties depend on local sequence and epigenetic modifications. For example, structural data[6] have been used to predict various shape parameters (twist/ roll/ tilt) associated with all dinucleotides. These models predict plectoneme density along genomic DNA[1], and to some extent, nucleosome organization from DNA mechanics, although they do not capture key features of nucleosome occupancy data[2]. All atom Monte Carlo (MC) simulations have predicted the sequence-dependence of local DNA shape[3,4], while molecular dynamics (MD) simulations have observed surprising polymorphic behaviors associated with the shape parameters describing B-DNA[5-8]. However, experimental validation of such predictions is extremely limited[6].

Direct experimental methods to measure how DNA sequence impacts its mechanical properties have mainly included low-throughput techniques such as measuring cyclization rates of short DNA

molecules[9], DNAse I accessibility of DNA minicircles[10,11] and DNA stretching[12] via AFM or hydrodynamic drag. Low-throughput has restricted their use to primarily studying the mechanics of isolated examples of promoter sequences[13,14], nucleosomal sequences[15-17], or sequences involved in transcription activation[18]. High-throughput SELEX-based methods have identified sequence features that are more enriched or de-enriched in highly loopable molecules by performing several rounds of selection from an initial random pool of sequences[19,20]. However, quantitative bendability values have not been assigned to any sequence, and as a result, SELEX data have not led to quantitative measures of DNA mechanics, or quantitative models.

Cytosine methylation in the CpG contexts is a form of epigenetic modification in mammalian genomes, and is capable of altering gene expression during development and in disease through preventing transcription factor binding and recruiting specialized proteins[21,22]. CpG methylation has also been suggested to impact transcription by altering the mechanical and structural properties of DNA and chromatin: CpG methylation increases the rupture force necessary for DNA strand separation[23], reduces DNA bendability[17], alters the stability of nucleosomes[17,24,25], and enhances polyamine-mediated DNA-DNA association[26]. However, methylation-induced changes in DNA physics have not been characterized in diverse sequence contexts.

We recently reported a high-throughput experimental method called loop-seq to determine the 'intrinsic cyclizability' - a measure of cyclization propensity or bendability - of as many as ~90,000 different 50 bp DNA sequences at a time[27]. Here we apply loop-seq to a large library of randomly specified sequences, correlate intrinsic cyclizability with sequence features and CpG methylation state, and develop a comprehensive characterization of the mechanical code. Application of our predictive model in various contexts suggests the functional relevance of DNA mechanics. Loop-seq provides a high-throughput experimental platform to test the predictions of MD simulations regarding distortion energy penalty and structural polymorphism.

Results

We previously reported[27] the intrinsic cyclizabilities of 12,472 50 bp DNA fragments (this set of randomly chosen sequences was termed the ′random library (supplementary note 1)). Here, we measured the intrinsic cyclizability of all sequences in the random library after we enzymatically methylated all CpG dinucleotides for a side-by-side comparison. Overall, DNA becomes more rigid upon CpG methylation, thereby suggesting that in addition to sequence, DNA mechanics depends on epigenetic modifications (Fig. 1a, Supplementary Note 2). We therefore use these two datasets as a starting point for understanding how DNA sequence and epigenetic modifications influence DNA mechanics.

Sequence features that influence DNA cyclizability

The overall G/C content is uncorrelated with intrinsic cyclizability except for its increase at very high G/C content above 65% (Fig. 1b). To test if certain dinucleotides influence intrinsic cyclizability, we sorted the random library members according to intrinsic cyclizability and divided them into 12 bins. The number of times each dinucleotide occurs in individual bins after appropriate normalization (Supplementary Note 3) (Fig. 1c) varied monotonically with bin number (there are some exceptions we will discuss later). We used a linear fit and defined its slope as the bendability quotient of each dinucleotide (Fig. 1d, Supplementary Note 3). For example, the TA dinucleotide (TpA) has the most positive bendability quotient, meaning that having more TpAs makes DNA more bendable on average. Conversely, the CG dinucleotide (CpG) has the most negative bendability quotient. Previous studies have suggested that TpA is particularly flexible[9,28-31] and is associated with highly bent DNA in nucleosomes[32,33]. Less is known about the contribution of CpG to DNA mechanics. A SELEX-based study found that CpGs are depleted among the highly loopable sequences, but they did not find specific enrichment of TA dinucleotides over other T/A dinucleotides[19].

TT and AA dinucleotides are represented highly in both the highest and the lowest intrinsic cyclizability bins (Fig. 1c), suggesting that their positional distribution could play a greater role in determining intrinsic cyclizability than their overall content. Short stretches of dA:dTs bend DNA towards the minor groove. These bend directions add in phase or cancel out if repeated at the helical or half-helical

period, resulting in globally curved or straight molecules that would have high or low intrinsic cyclizability, respectively[34-36].

The availability of high-throughput data now permits us to quantitively evaluate how phased repeats of any sequence motifs (not just short dA:dT stretches) influence DNA mechanics. There are 136 NN-NN dinucleotide pairs (such as AA-TT, AC-GG, etc.). For a given pair in a sequence, we define the pairwise distance distribution function, $\rho(i)$, as the number of times that the two dinucleotides appear separated by a distance i, appropriately normalized (Supplementary Note 4). For example, the AA-TT dinucleotide pair shows striking oscillations in $\rho(i)$ when averaged over thousands of sequences: these dinucleotides tend to more likely be separated by 10, 20 or 30 bp (i.e. multiples of the helical repeat) and less likely be separated by 5, 15 or 25 bp (i.e. odd multiples of half the helical repeat) among the 1,000 most cyclizable sequences (Fig. 1e, red curve) and vice versa among the 1,000 least cyclizable sequences (blue curve). These observations are consistent with the known fact that short dA:dT stretches repeated every 10 bp result in curved molecules[37,38].

We discovered that other dinucleotide pairs can also influence intrinsic cyclizability depending on their mutual distance (Fig. 1e). The CG – GC pair shows a similar behavior to AA-TT pair. The TT – GC pair shows the reverse effects: sequences with high cyclizability have more instances of TT – GC separation at 5, 15, or 25 bp, and vice versa. Several other pairs, such as CA – CA, show no oscillatory patterns. To quantify these effects, we define a metric called the helical repetition index (HRI) of a given dinucleotide pair in a given sequence as follows:

$$HRI = \sum_{i=10,20,30} max(\rho(i-1), \rho(i), \rho(i+1)) - \sum_{i=5,15,25} max(\rho(i-1), \rho(i), \rho(i+1))$$

A high HRI of a dinucleotide in a sequence implies that, in that sequence, the dinucleotide pair in question is more often separated by multiples of the helical repeat than by odd multiples of half the helical repeat. HRI of each dinucleotide pair showed a linear dependence on intrinsic cyclizability, and we defined the slope as a measure of the pair′s contribution towards intrinsic cyclizability (Fig. 1f). Dinucleotide pairs

with all As or Ts make a positive contribution to intrinsic cyclizability (Fig. 1f). A similar effect was observed for some of the dinucleotide pairs made of Gs or Cs only (Fig. 1f) and this is not a result of exclusion of A/Ts (supplementary note 7), but is in fact consistent with an earlier suggestion that G/C dinucleotides tend to bend DNA towards the major groove[9]. These curvatures might similarly add in phase or cancel out in sequences where G/C rich dinucleotides occur every helical or half helical repeat, respectively. In the case of hybrid pairs, where one dinucleotide is purely A/T containing and the other purely G/C containing, DNA sequence with high HRI is less cyclizable. Consecutive bend directions towards the minor and major groove every half-helical repeat or helical repeat would add in phase and cancel out respectively, presumably contributing to this effect. These findings are consistent with SELEX experiments which noted that A/T rich stretches separated by 5 bp from G/C rich stretches, on average, are more represented in highly loopable molecules[19]. In addition, there are several hybrid pairs, such as AG-GG, or GT-AG, which make significant positive or negative contributions to intrinsic cyclizability, which could not have been predicted from prior literature.

Both dynamic flexibility as well as static bends may contribute to intrinsic cyclizability as noted previously[27]. Speculatively, the helical repetition index of dinucleotides determines intrinsic cyclizability mainly by altering the static bent structure of DNA and determining whether alternate static bend directions add in phase or cancel out. On the other hand, bendability quotients, by virtue of being independent of nucleotide distribution, may reflect how dinucleotides contribute to dynamic flexibility.

We next asked whether the sequence features linked to high or low intrinsic cyclizability are found in genomic landmarks. Genes are characterized by a well-ordered array of downstream nucleosomes (labeled +1, +2, etc, depending on their distance from the TSS) and an upstream Nucleosome Depleted Region (NDR)[39,40]. Via loop-seq on *S. cerevisiae,* we discovered a narrow well-defined region of unusually low intrinsic cyclizability co-centric with the NDR and regions of high local intrinsic cyclizabilities at the locations of the downstream nucleosomes[27].

Averaged over 4,912 genes in *S. cerevisiae* that have previously been annotated as ORFs and both ends mapped with high confidence[41], G/C content shows no special features around the TSSs, other than a

general dip near the NDR (Fig. 1g), likely because NDRs are A/T rich. However, CpG content peaks at the NDR and dips at nucleosomal dyads (Fig. 1g, supplementary note 8), contrary to the overall G/C content distribution, but consistent with CpG having the most negative bendability quotient (Fig. 1d). Likewise, TpA content peaks at +/-1 nucleosomal dyads and dips near the NDR (Fig. 1g, supplementary note 8), consistent with TpA having the highest bendability quotient (Fig. 1d). The sum of the helical repetition index of the 10 dinucleotide pairs that make the most negative contribution to intrinsic cyclizability peaks near the NDRs and dips near the dyads of the +1 nucleosomes (Fig. 1g). In other words, such pairs tend to more often be separated at the helical repeat (or less often at half the helical repeat) near the NDRs of genes, and vice versa near the dyads of +1 nucleosomes. Likewise, the sum of the helical repetition index of the 10 dinucleotide pairs that make the most positive contribution to intrinsic cyclizability (such as AA-TT, GC-GC, etc, Fig. 1f) averaged over all genes, peaks at the dyads of +1 nucleosomes and dips near the NDRs (Fig. 1g). Therefore, the sequence features for high or low intrinsic cyclizability, identified from the random library, are indeed found in genomic regions where we have independently measured[27] intrinsic cycliability to be high or low, respectively. The low intrinsic cyclizability at the NDRs, and other shape properties of DNA in this region, may contribute to nucleosome depletion by impacting the nucleosome sliding activities of chromatin remodelers, as suggested earlier[27,42,43]. The high rigidity of DNA may also reduce nucleosome occupancy by energetically disfavoring extensive DNA bending[32,44-48], although its effect may be overshadowed by the action of chromatin remodelers.

Intrinsic cyclizability is related to DNA shape

DNA shape plays a fundamental role in determining the extent of protein-DNA interactions[49] and the binding specificities of transcription factors[50] and nucleosome formation[51]. It is described by parameters such as twist, roll, propeller twist, etc[52]. Propeller twist is the angle between the planes of the two bases in a base pair (Supplementary Note 9). High propeller twist has been suggested to be an attribute of rigid DNA[53]. Indeed, we found that sequences with low measured intrinsic cyclizability have high predicted propeller twist values (Fig. 2a).

Twist is the angle by which one base pair rotates relative to the adjacent base pair around the DNA duplex axis. The CpG step transitioned between high and low twist conformations in MD simulation with their relative populations depending on its tetranucleotides context[5] and even within a defined tetranucleotide context their relative weights depended on the identities of the flanking nucleotides[5,6]. To provide an experimental test of these predictions[54] we first estimated the bendability quotients of all 256 NNNN tetranucleotides (Fig. 2b, supplementary note 10). NCGN and NTAN tetranucleotides tend to have low and high bendability quotients, respectively, consistent with CpG and TpA having the lowest and highest bendability quotients, respectively (Fig. 2b). NCGNs with stronger weight toward the low twist conformation in MD simulations also showed low measured bendability quotients and vice versa (Fig. 2c), lending experimental support to the polymorphic twist behavior of CpGs[6] and suggesting that low bendability of CpGs arises from their low twist conformation.

Roll measures the rotation of a base pair plane around its long axis and may influence DNA bending[55]. All-atom MD simulations have reported the equilibrium roll angles associated with all NNNN tetranucleotides[7,8]. However, when a tetranucleotide occurs as part of a cyclized molecule, its roll angle has to deviate from the unstressed value (supplementary note 11). We tested if the large variation in bendability even for those sharing the same central dinucleotides is caused by variations in roll energy penalty incurred upon cyclization. Indeed, for 16 different NCGN tetranucleotides, the energy penalty calculated from MD simulations was higher if the experimentally determined bendability quotient is lower ($r_{Pearson}$ = -0.83) (Supplementary Note 11) (Fig. 2d). Expanding the analysis to all 16 dinucleotides (NNs) showed several other NNs where the 16 bendability quotients of tetranucleotides that contain the NN in the middle are significantly anti-correlated with the corresponding roll energy values (Supplementary Note 12). Therefore, roll energy penalty is a major determinant of DNA bendability in a broad sequence context.

Overall, MD simulations provided atomic level insights into the sequence-dependence of DNA bendability derived from loop-seq analysis.

CpG methylations stiffens DNA in all sequence contexts

CpG methylation has been proposed to affect transcription by altering the relevant physical properties of chromatin[17,23-25]. However, there has been no high-throughput characterization of how DNA mechanics is impacted by CpG methylation in a broad sequence context. Recently, DNA cyclization measurements and MD simulations suggested that CpG methylation can make DNA more rigid[17,56]. These experiments, however, were performed on a single DNA sequence, and require reevaluation because they did not account for the surface tethering position that sets the phase of preferred bending[27]. Intrinsic cyclizability, which is determined from loop-seq data with three different tethering positions, showed that CpG methylation, averaged over all sequence contexts, makes DNA more rigid (Fig. 1a). CpG methylation decreases the bendability quotients of all tetranucleotides that contain at least one CG (Fig. 2e), indicating that CpG methylation stiffens DNA in all sequence contexts.

We next examine the effect of CpG methylation on how the spatial distribution of dinucleotides influences intrinsic cyclizability. Higher helical repetition index of unmethylated CG-NN pairs make DNA more cyclizable if N = G or C, and more rigid if N = A or T (Fig. 1f). CpG methylation de-emphasize the importance of phased dinucleotide pairs: CG-NN pairs with G/C rich and A/T rich NNs decrease and increase, respectively, their contribution to intrinsic cyclizability as a result of CpG methylation (Fig. 2f). Because contributions of dinucleotide pairs towards intrinsic cyclizability likely arise from the directions of consecutive static bending every helical repeat either adding in phase or canceling out, we propose that CpG methylation suppresses DNA bending towards the major groove at CpG steps. It has indeed been suggested, based on crystal structures, that methylation increases steric hindrance at the major groove, thereby widening the major groove and suppressing bending[57]. Overall, loop-seq provides evidence that CpG methylation decreases dynamic flexibility and buffers sequence-dependent static curving of DNA.

Models to predict intrinsic cyclizability

After identifying sequence features that impact intrinsic cyclizability, we aimed to build a unified physical model that predicts DNA mechanics from sequence by incorporating the effects of NN content and NN-NN helical repetition indices (HRIs) (Fig. 1). A 50 bp DNA sequence can be described by a set of 16 NN contents (the number of times each of the 16 dinucleotides occurs in the sequence) and 136 NN-NN HRIs. We built linear models that treat intrinsic cyclizability as a linear combination of firstly just the NN contents, then just the NN-NN HRIs, and finally all the 152 parameters (16 NN contents and 136 NN-NN HRIs (supplementary note 13, Figs. 3a-b). The best fit coefficients are obtained through multivariant linear regression, using the Tiling Library as the training dataset. We tested the models by predicting the intrinsic cyclizabilities of sequences in the Random, ChrV and Cerevisiae Nucleosomal Libraries (Fig. 3c). Correlation coefficients between measured and predicted values obtained by the combined linear model are in the range 0.52 – 0.54 (Fig. 3c), which are significantly higher than those obtained using NN contents alone or NN-NN HRIs alone and are close to what we obtained using neural nets (Supplementary Note 14).

Predicted intrinsic cyclizability around 576 genes in *S. cerevisiae* captures all essential features that were identified using the measured intrinsic cyclizability values – a well-defined region of unusually low intrinsic cyclizability co-centric with the NDR, and local peaks in intrinsic cyclizability around the dyads of downstream nucleosomes (Fig. 3d). The absolute values do not match between prediction and measurement because the measured intrinsic cyclizability is defined only up to an additive constant which depends on other sequences present in the library[27].

The impact of DNA mechanics in promoters and CTCF sites

To explore relations between DNA mechanics and nucleosome organization, we compared predicted intrinsic cyclizability with measured nucleosome occupancy around the TSSs in four different organisms (Fig. 4a). For *S. cerevisiae*, predicted intrinsic cyclizability patterns broadly agree with what was measured from a subset of genes (576 genes)[27] (Fig. 3d) – there is a well-defined region of rigid DNA upstream of the +1 dyad, which coincides with the NDR (Fig. 4a). Additionally, flexibility peaks are visible at the location of the +1 and -1 nucleosomes, and to a lesser extent, the +2 nucleosome. For *S. pombe*, we

observed two marked differences[58] (Fig. 4a). First, there is no special region of rigid DNA upstream of the +1 nucleosome, and second, intrinsic cyclizability shows much stronger oscillations in phase with the nucleosome occupancy oscillation, at least up to +7 nucleosome. It is possible that in the absence of a well-defined region of rigid DNA at the NDR, paused polymerases[59], transcription factors[60] or other bound factors may serve as a similar barrier to nucleosome sliding in *S. pombe*. The more prominent oscillations in intrinsic cyclizability are indicative of a greater importance of DNA mechanics in evenly spacing gene body nucleosomes in *S. pombe* than in *S. cerevisiae*, although the overall contribution of such an effect in either species could be dwarfed by the action of chromatin remodelers. Future experiments involving altering DNA flexibility via genome editing would be necessary to understand differences in how either species translates genomic information into nucleosome positioning. In drosophila and mouse, our predictive models suggest that, as in *S. cerevisiae*, there exists a sharply-defined region of rigid DNA near TSSs. Further, in drosophila, small peaks in intrinsic cyclizability are also visible at the locations of the +1 and +2 nucleosomes.

We next predicted how CpG methylation would change the mechanical landscape around TSSs. In *S. cerevisiae*, *S. pombe*, and *D. melanogaster*, CpG methylation does not cause any significant change in the overall predicted landscape of intrinsic cyclizability around TSSs (Fig. 4a). CpG methylation also does not occur in these organisms. In contrast, our model predicts that methylation significantly alters the landscape around TSSs in mouse (Fig. 4a): the characteristic sharply-defined region of rigid DNA near TSSs becomes broadened and less well-defined. To test the prediction, we measured the intrinsic cyclizabilities near the TSS of 649 randomly selected genes in mouse with or without CpG methylation. The measured mechanical landscape indeed confirmed that CpG methylation results in the broadening and dilution of the rigid DNA region (Fig. 4b, Supplementary Note 17). It is therefore possible that methylation-induced changes in DNA mechanics near mouse promoters contributes to transcriptional silencing by altering promoter-proximal nucleosome positioning and the action of chromatin remodelers which we have shown to depend on DNA mechanics[27].

Nucleosomes are also well organized around other genomic landmarks, one notable example being the binding sites of the transcription factor CTCF[61]. CTCF sites are surrounded by well-ordered nucleosomes and the sites themselves are either free of nucleosomes or bound by fragile nucleocomes[62,63]. Our model predicts that CTCF binding sites in mouse embryonic stem cells[62,64] are characterized by a prominent peak in intrinsic cyclizability (Fig. 4c). The width of the peak extends significantly beyond the 20 bp CTCF consensus motif, and this is not caused by averaging (supplementary note 18), consistent with the report that a structured DNA region exceeds the CTCF consensus motif[65]. The sharp peak in intrinsic cyclizable may aid in the known positioning of nucleosomes at CTCF sites. Indeed, our model predicts that this peak cyclizability value, as well as the contrast in cyclizability between the peak and the surrounding DNA is higher for sites with higher nucleosome occupancy (Fig. 4d). G/C content alone cannot explain these observations (Supplementary Note 20).

We tested these model predictions on CTCT sites using loop-seq. The measured mechanical landscape shows a prominent peak in cyclizability at CTCF sites which extends beyond the 20 bp consensus motif (Fig. 4e). Further, we observed two flanking peaks in cyclizability which were not captured by the predictive model and align with nucleosome occupancy[65], suggesting that they might aid in the positioning of the flanking nucleosomes (Fig. 4e). The 5 upstream flanking peak in cyclizability is somewhat further from the CTCF motif than the corresponding peak in nucleosome occupancy, consistent with the requirement of the chromatin remodeler SNF2H for reduced nucleosome repeat length[65]. We further experimentally verified the prediction that CTCF sites with higher nucleosome occupancy have a higher contrast in intrinsic cyclizability (Fig. 4f).

Impact of DNA mechanics in non-nucleosomal contexts

Next, we examined DNA mechanics in several non-nucleosomal contexts.

Groups of A_n/T_n nucleotides present in a periodic fashion, extending for hundreds of nucleotides, has been reported in eukaryotes genomes[66,67]. Such hyperperiodic DNA is particularly pronounced in the

306 *C. elegans* genome at ~10 bp periodicity, and recent structural studies have revealed that it is curved[68]. The
307 intrinsic cyclability predicted across the hyperperiodic segment using the NN-NN phased repeat model
308 indeed shows much higher values than when the nucleotide order in the ~900 bp region is scrambled,
309 suggesting that hyperperiodic phase repeats are responsible for the curved DNA observed (Fig. 4g). The
310 functional consequences of hyperperiodic DNA extending for hundreds of nucleotides are unclear[68] and
311 require future experiments.

312 TATA binding protein (TBP) binds to promoter DNA, about 30 bp upstream of the TSS in *S.*
313 *cerevisiae*[69], during transcription initiation[70]. The TATA-box – a consensus sequence to which TBP binds,
314 is present in only 15% of promoters, although TBP binding is universally required for transcription[71]. TBP
315 binding to the TATA-box induces DNA bending by ~90° [72]. Consistent with this observation, sequence-
316 dependent curvature of DNA is linked to transcription efficiency[73], and TBP binds less efficiently to poly-
317 A tracts. Our predictive model, averaging across all categories of TSSs[74], showed a local peak about 25 bp
318 upstream of the TSSs (Fig. 4h). This coincides with the location of the TBP[69] and of the TATA box in
319 TATA-containing promoters. Our findings suggest that genome-wide, DNA mechanics might influence
320 DNA bending by TBP regardless of the presence of the TATA-box.

321 DNA bending is also ubiquitous in bacteria. The *E. coli* RNA polymerase σ70 (RNAP) open
322 promoter complex has 300° wrapping of about ~90 bp of DNA around the RNAP[75]. Our predicted intrinsic
323 cyclizability averaged around *E. coli* TSSs[76] show a sharp peak from -35 to 0 position relative to the TSS,
324 and an overall region of elevated intrinsic cyclizability extending from about position -100 to about +150
325 (Fig. 4i). It is possible that such an extended region of flexible DNA might aid in extensive DNA bending
326 required to form the open promoter complex and for DNA wrapping around RNAP although verification
327 requires future experiments.

328 Negative supercoiling of DNA by DNA gyrase is critical for bacterial life. Gyrase bends DNA at
329 its cleavage site and extensively wraps DNA up to 80 bp away[77]. Strong Gyrase Sites (SGSs) in some
330 genomes have been proposed to allow extensive DNA bending[78]. FluMu is a prophage whose 35kb genome
331 is found inserted within the *Haemophilus influenza* genome[79,80], and contains one SGS[81]. Predicted intrinsic

cyclizability along the entire genome has three most prominent peaks all within 200 bp of the cleavage site of the SGS (Fig. 3j). One of the peaks coincides with the cleavage site, where gyrase must bend DNA. The second peak lies within 80 bp of the cleavage site – this region is where the wrapping of DNA occurs, in turn requiring extensive bending. Thus the mechanical properties of DNA around SGSs may have evolved to facilitate gyrase activity.

Looping of 100 – 140 bp of DNA is required for efficient communication between enhancers and σ^{54}-dependent promoters in bacteria[82]. The bacterial protein IHF extensively bends DNA[83]. Low-throughput DNA cyclization experiments have revealed that in IHF-dependent σ^{54}-promoters, DNA cannot intrinsically bend to support this loop formation, although it can in IHF-independent promoters. Applying our predictive model, we found that indeed among the eight σ^{54}-dependent promoters on which cyclization experiments were performed earlier[82], the two that did not contain an IHF-binding site were more cyclizable than the rest in the region between the enhancer and the promoter (Fig. 4k). This is consistent with the proposal[82] that a function of IHF might be to overcome a communication block set up by the rigid mechanical properties of the IHF binding site. In certain promoters, intrinsically flexible DNA may take over the role of IHF in facilitating DNA bending.

Discussion

We obtained a comprehensive characterization of the mechanical code of the genome and epigenome using high-throughput experimental measurements. Using high-throughput loop-seq measurements, we quantified the contributions of short sequence features (such as all dinucleotide contents and HRIs) to intrinsic cyclizability and measured, in a broad sequence context, how these contributions are altered by CpG methylation. We demonstrate loop-seq to be a powerful experimental platform for testing the predictions of structural studies and MD simulations in a broad sequence context, and making new testable predictions about the structural dynamics of nucleic acids. We integrated our data into regression and machine learning based models for predicting how local sequence and epigenetic modifications impact

DNA mechanics. Applications of our predictive model, and its confirmation via new loop-seq measurements, broadly suggests that DNA mechanics may have evolved to facilitate processes that require DNA bending in both eukaryotes and bacteria.

Physical deformations of DNA are not restricted to bending, but include twisting, supercoiling, or unzipping, during, for example, transcription initiation[84] and elongation[85], the binding of various transcription factors[86], or 3D folding of the genome at local[87] and global scales[88]. Future high-throughput experimental and computational methods to understand how DNA sequence impacts such diverse modes of DNA deformability will expand our understanding of how regulatory information is encoded via the mechanical code, and how this information can be evolved, adapted, or controlled by processes that chemically alter DNA via sequence mutations, epigenetic modifications, or DNA damage.

Acknowledgements:

A.B. and T.H. would like to thank Jun S. Song for suggestions and insights pertaining to developing the linear predictive models. This work was supported by the Royal Society URF\R21\211659 (to A.B.), the Royal Society RF\ERE\210288 (to A.B.), funding from Durham University (to A.B.), the National Science Foundation Grants PHY-1430124, EFMA 1933303 (to T.H.), the National Institutes of Health Grant GM122569 (to T.H.), the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 754510 (to J.P.A), the Spanish Ministry of Science RTI2018-096704-B-100 and AGAUR, Generalitat de Catalunya, Grups de Reserca Consolidats 2017 SGR 1110 (to M.O.). T.H. is an Investigator with the Howard Hughes Medical Institute. A.B. is a Royal Society University Research Fellow.

Author contributions:

AB and TH designed research. AB performed research, analyzed data, and built the predictive models. AB and TH wrote the paper. DGB compiled nucleosome occupancy data from various organisms. BC

investigated pairwise correlation among NN-NN dinucleotide pairs in highly loopable and rigid sequences. ZQ assisted with library preparation loop-seq experiments pertaining to the random library. JPA and MO related cyclizabilities to DNA shape parameters.

Competing Interests:

The authors declare no competing interests.

Figure Legends:

Figure 1: (a) The ordinate depicts mean intrinsic cyclizabilities of sequences in the random library (blue) and in the methylated random library (red) that have at least the number of CpGs as specified by the abscissa. Intrinsic cyclizability values of the methylated random library have been adjusted to allow for comparison with those of the random library (Supplementary Note 2). For each color, from left to right, N = 12,472, 12,036, 10,451, 7,620, 4,516, 2,147, 864, 282. Error bars are s.e.m. (b) Mean intrinsic cyclizability of sequences in the random library as a function of G/C content. Mean is calculated only over those sequences in the random library whose G/C content is as specified by the abscissa. Error bars are s.e.m. From left to right, N = 30, 77, 110, 214, 324, 500, 746, 984, 1239, 1353, 1403, 1316, 1210, 982, 749, 487, 307, 201, 120, 53, 27. (c) The 12,472 sequences in the random library were sorted according to increasing intrinsic cyclizability and grouped into 12 bins with 1,039 sequences each. 4 remaining sequences were ignored. Within each bin, the normalized number of times each of the 16 dinucleotides occur is color coded and depicted (see supplementary note 3). (d) Bendability quotient for a dinucleotide is defined as the slope of the linear fit to the plot of the normalized number of times it occurs among sequences in each of the 12 bins in panel c, vs the mean intrinsic cyclizability of sequences in the 12 bins (supplementary note 3) (e) Pairwise distance distribution functions vs separation distance, averaged over the 1,000 sequences in the random library that have the most (red) or least (blue) values of intrinsic cyclizability, for four different NN-NN pairs. See Supplementary Note 5 for plotting details. (f) For a given

NN-NN dinucleotide pair, the best fit linear relationship between its helical separation extent in a sequence and the intrinsic cyclizability of that sequence, for all sequences in the random library, was obtained. The heatmap here depicts the slopes of these linear relationships for all 136 NN-NN pairs. See Supplementary Note 6 for details. (g) Nucleosome occupancy and various sequence parameters as functions of position from the dyad of the +1 nucleosome, averaged over all identified 4,904 genes in *S. cerevisiae* (see supplementary note 8 for plotting details). To calculate $\sum_{rigid} \sigma_{NN-NN}$, the ten NN-NN pairs that make the most negative contribution to intrinsic cyclizability were identified (supplementary note 8). The sum of the helical separation extents of these pairs over a 50 bp DNA fragment centered around the ordinate value was calculated for each gene. The values were averaged to obtain the abscissa value. $\sum_{flexible} \sigma_{NN-NN}$ was similarly calculated. See supplementary note 8 for heatmaps of TpA and CpG contents.

Figure 2: (a) Mean propeller twist as a function of position averaged over N = 1,000 50 bp DNA sequences in the random library that had the most (red) and least (blue) values of intrinsic cyclizability. Sequences were tiled as a series of pentamers and the associated propeller twist of the central base in the pentamer was assigned on the basis of earlier reports[4]. Error bars are s.e.m. (b) The 12,472 sequences in the random library were sorted according to increasing intrinsic cyclizability and grouped into 12 bins with 1,039 sequences each. 4 remaining sequences were ignored. Within each bin, the normalized number of times each of the 256 tetranucleotides occur is color coded and depicted (Supplementary Note 10). Tetranucleotides containing a CG in the middle (ie of the form NCGN) are indicated by a magenta circle, while those containing a TA in the middle are indicated by a green circle. (c) Bendability quotients of all 16 NCGN tetranucleotides as obtained from loop-seq of the random library (see Supplementary Note 10) vs the weighted average of the high-twist and low-twist conformations that the central CG step samples, as obtained from MD simulations[6]. Pearson's r value is shown. 95% confidence interval (CI) = 0.33, 0.89. P value determined by t-test (two-sided). (d) Bendability quotients of all 16 NCGN tetranucleotides vs roll energy. Roll energy is the energy penalty associated with the tetranucleotide having to deviate from its equilibrium roll angle when adopting a constrained conformation after looping, i.e., the tetranucleotide

being part of a circle of circumference 110 bp (Supplementary Note 11). Pearson's r value is shown. 95% confidence interval (CI) = -0.57, -0.94. P value determined by t-test (two-sided). (e) Bendability quotients for all 256 NNNNs obtained from the measured values of intrinsic cyclizability of the 12,472 sequences in the random library vs those obtained from the measured values of intrinsic cyclizability of the sequences in the methylated random library (which contain the identical set of 12,472 sequences, except all occurring CpG are cytosine methylated). Dashed line represents x=y. NNNNs are marked in red if at least one CG occurs in it (such as ACGA, CGCG, CGAC, etc). Other NNNNs (such as AAGC, GGGC, etc) are marked in blue. (f) Heatmap representing the contributions of all NN-CG dinucleotide pairs towards intrinsic cyclizability, obtained by considering the intrinsic cyclizability values of sequences in the random library (first column, identical to the 15^{th} column in figure 1f except for the color scale), and obtained from measurements on the methylated random library (second column). Contribution towards intrinsic cyclizability of a NN-CG pair is calculated as done in the case of figure 1f.

Figure 3: (a) 2-D histogram of the scatter plot between measured intrinsic cyclizabilities of sequences in the random library and the associated predicted intrinsic cyclizability. Here, prediction was made via a model where intrinsic cyclizability of a 50 bp sequence is a linear combination of the number of times each of the 16 dinucleotides occur in the sequence and a constant term (supplementary note 13). Best fit coefficients of the linear model were derived by training the model using the measured intrinsic cyclizability values of sequences in the tiling library (Supplementary Note 1). Pearson's r value is shown. 95% confidence interval (CI) = 0.34, 0.37. P = 0.0000 (determined by two-sided t-test). (b) Same as panel a, except that prediction was made using a linear model where intrinsic cyclizability of a 50 bp sequence is a linear combination of the 136 helical separation extent values in the sequence of the 136 NN-NN pairs, and a constant term (supplementary note 13). Pearson's r value is shown. 95% confidence interval (CI) = 0.36, 0.39. P = 0.0000 (determined by two-sided t-test). (c) 2-D histogram of the scatter plots between measured and predicted intrinsic cyclizability values of sequences in the random, chrV, and cerevisiae nucleosomal libraries (Supplementary Note 1). Here, prediction was made using a model were intrinsic cyclizability of

a 50 bp sequence is a linear combination of a constant term and the 16 dinucleotide contents (subject to the constraint that their sum = 49) and 136 helical separation extents that describe the sequence[89] (supplementary note 13). Coefficients were derived by training the model against the tiling library. Pearson's r values are shown. 95% confidence intervals (CIs) are (0.51, 0.53), (0.52, 0.55), and (0.52, 0.55). P = 0.0000 (determined by two-sided t-test) in all cases. (d) Measured intrinsic cyclizability, predicted intrinsic cyclizability, and nucleosome occupancy as functions of position from the dyad of the +1 nucleosome, averaged over 576 genes in *S. cerevisiae*. Mean occupancy values[90] and positions[33] were as reported earlier. Prediction was performed using the linear physical model trained using the measured intrinsic cyclizability values of the random library. See supplementary note 15 for details.

Figure 4: (a) Nucleosome occupancy, predicted intrinsic cyclizability in absence of CpG methylation, predicted intrinsic cyclizability in presence of CpG methylation, and G/C content, as a function of distance from the dyad of the +1 nucleosome (in the case of *S. cerevisiae* and *S. pombe*) or from the TSS (in the case of drosophila and mouse), averaged over a large number of genes in these four organisms. Nucleosome occupancy metrics were obtained from previous publications. See supplementary note 16 for details. (b) Measured intrinsic cyclizability as a function of position from the TSS, averaged over 629 mouse genes, in absence and presence of CpG methylation. See supplementary note 17 for details. (c) Top panel: Predicted intrinsic cyclizability (predicted by using the linear physical model trained against the random library) as a function of position from the start of the 20 bp CTCF consensus motif, averaged over 19,900 reported CTCF binding sites in mouse embryonic stem cells[64]. Bottom panel: same as the top panel, except DNA outside the 20 bp concensus sequence motif was chosen at random and not obtained from the mouse genome. See Supplementary Note 18 for details. (d) Predicted intrinsic cyclizability as a function of position (where 0 is the start of the CTCF consensus motif), averaged over the two groups of 1,000 sites that have the highest (blue) and lowest (red) nucleosome Center Weighted Occupancy (CWO[62]) at the CTCF motif. See supplementary note 19 for details. (e) Measured intrinsic cyclizability and nucleosome occupancy vs position from the edge of the CTCF motif, averaged over 433 randomly selected CTCF binding sites. See

supplementary note 21 for details. Nucleosome occupancy is the nucleosome CWO[62]. (f) Measured mean intrinsic cyclizability vs position from the edge of the CTCF motif, averaged over two sets of 433 CTCF binding sites that have the least and greatest mean nucleosome CWO[62] at the CTCF motif. See supplementary note 21 for details. (g) Predicted intrinsic cyclizability as a function of position along the 923 bp Ω4 region of the C. elegans genome[68] (blue), and along three more 923 bp DNA sequences obtained by randomizing the order of nucleotides that occur along the native Ω4 sequence (black, green, red). (h) Predicted intrinsic cyclizability as a function of distance from TSSs in *S. cerevisiae*. Data is averaged over all 11,102 annotated TSSs[74] that were more than 500 bp away from chromosome edges. The red arrow points to a local peak in intrinsic cyclizability coinciding with the known location of the TATA box. The black arrow indicates the flexibility peak associated with the +1 nucleosome dyad, as reported earlier[27] (i) Predicted intrinsic cyclizability as a function of distance from TSSs in *E. coli*. Data is averaged over 14,860 annotated TSSs[76]. (j) Predicted intrinsic cyclizability along the 35 kb genome of the MU-like prophage FluMu found inserted within the *H. influenza* genome. Arrow points to the region of the gyrase cleavage site. Right panel: zoomed view of a 1 kb region around the strong gyrase binding site in the FluMu genome. The red dashed line corresponds to the gyrase cleavage site. The dashed grey lines demarcate 80 bp from the cleavage site. (k) Predicted intrinsic cyclizability as a function of position along the 220 bp DNA fragment encompassing the enhancer and promoter of the eight σ54-promoters on which earlier DNA cyclization experiments have been reported. In bold are the two promoters (*K. pneumonia nifLA* promoter and *E. coli glnAp2* promoter) among the eight which lack an IHF binding site. Plotted on the x-axis is the position along the 220 bp fragment from the enhancer to the promoter.

References

1 Kim, S. H. *et al.* DNA sequence encodes the position of DNA supercoils. *Elife* **7**, e36557 (2018).

2 Morozov, A. V. *et al.* Using DNA mechanics to predict in vitro nucleosome positions and formation energies. *Nucleic acids research* **37**, 4707-4722 (2009).

3 Rohs, R., Sklenar, H. & Shakked, Z. Structural and energetic origins of sequence-specific DNA bending: Monte Carlo simulations of papillomavirus E2-DNA binding sites. *Structure* **13**, 1499-1509 (2005).

4 Chiu, T.-P. *et al.* GBshape: a genome browser database for DNA shape annotations. *Nucleic acids research* **43**, D103-D109 (2015).

5 Pasi, M. *et al.* μABC: a systematic microsecond molecular dynamics study of tetranucleotide sequence effects in B-DNA. *Nucleic acids research* **42**, 12272-12283 (2014).

6 Dans, P. D. *et al.* Unraveling the sequence-dependent polymorphic behavior of d (CpG) steps in B-DNA. *Nucleic acids research* **42**, 11304-11320 (2014).

7 Dans, P. D. *et al.* The static and dynamic structural heterogeneities of B-DNA: extending Calladine–Dickerson rules. *Nucleic acids research* **47**, 11090-11102 (2019).

8 Walther, J. *et al.* A multi-modal coarse grained model of DNA flexibility mappable to the atomistic level. *Nucleic acids research* **48**, e29-e29 (2020).

9 Geggier, S. & Vologodskii, A. Sequence dependence of DNA bending rigidity. *Proceedings of the National Academy of Sciences* **107**, 15421-15426 (2010).

10 Brukner, I., Jurukovski, V. & Savic, A. Sequence-dependent structural variations of DNA revealed by DNase I. *Nucleic acids research* **18**, 891-894 (1990).

11 Brukner, I., Sanchez, R., Suck, D. & Pongor, S. Sequence‐dependent bending propensity of DNA as revealed by DNase I: parameters for trinucleotides. *The EMBO journal* **14**, 1812-1818 (1995).

12 Rief, M., Clausen-Schaumann, H. & Gaub, H. E. Sequence-dependent mechanics of single DNA molecules. *Nature structural biology* **6**, 346-349 (1999).

13 Davis, N. A., Majee, S. S. & Kahn, J. D. TATA box DNA deformation with and without the TATA box-binding protein. *Journal of molecular biology* **291**, 249-265 (1999).

14 Parvin, J. D., McCormick, R. J., Sharp, P. A. & Fisher, D. E. Pre-bending of a promoter sequence enhances affinity for the TATA-binding factor. *Nature* **373**, 724-727 (1995).

15 Satchwell, S. C., Drew, H. R. & Travers, A. A. Sequence periodicities in chicken nucleosome core DNA. *Journal of molecular biology* **191**, 659-675 (1986).

16 Ngo, T. T., Zhang, Q., Zhou, R., Yodh, J. G. & Ha, T. Asymmetric unwrapping of nucleosomes under tension directed by DNA local flexibility. *Cell* **160**, 1135-1144 (2015).

17 Ngo, T. *et al.* Effects of cytosine modifications on DNA flexibility and nucleosome mechanical stability. *Nature communications* **7**, 1-9 (2016).

18 Bracco, L., Kotlarz, D., Kolb, A., Diekmann, S. & Buc, H. Synthetic curved DNA sequences can act as transcriptional activators in Escherichia coli. *The EMBO journal* **8**, 4289-4296 (1989).

19 Rosanio, G., Widom, J. & Uhlenbeck, O. C. In vitro selection of DNA s with an increased propensity to form small circles. *Biopolymers* **103**, 303-320 (2015).

20 Beutel, B. A. & Gold, L. In vitro evolution of intrinsically bent DNA. *Journal of molecular biology* **228**, 803-812 (1992).

21 Greenberg, M. V. & Bourc'his, D. The diverse roles of DNA methylation in mammalian development and disease. *Nature reviews Molecular cell biology* **20**, 590-607 (2019).

22 Jones, P. A. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nature Reviews Genetics* **13**, 484-492 (2012).

23 Severin, P. M., Zou, X., Gaub, H. E. & Schulten, K. Cytosine methylation alters DNA mechanical properties. *Nucleic acids research* **39**, 8740-8751 (2011).

24 Lee, J. Y. & Lee, T.-H. Effects of DNA methylation on the structure of nucleosomes. *Journal of the American Chemical Society* **134**, 173-175 (2012).

25 Keshet, I., Lieman-Hurwitz, J. & Cedar, H. DNA methylation affects the formation of active chromatin. *Cell* **44**, 535-543 (1986).

26 Yoo, J., Kim, H., Aksimentiev, A. & Ha, T. Direct evidence for sequence-dependent attraction between double-stranded DNA controlled by methylation. *Nature communications* **7**, 1-7 (2016).

27 Basu, A. *et al.* Measuring DNA mechanics on the genome scale. *Nature* **589**, 462-467 (2021).

28 Protozanova, E., Yakovchuk, P. & Frank-Kamenetskii, M. D. Stacked–unstacked equilibrium at the nick site of DNA. *Journal of molecular biology* **342**, 775-785 (2004).

29 Okonogi, T., Alley, S., Reese, A., Hopkins, P. & Robinson, B. Sequence-dependent dynamics of duplex DNA: the applicability of a dinucleotide model. *Biophysical Journal* **83**, 3446-3459 (2002).

30 Olson, W. K., Gorin, A. A., Lu, X.-J., Hock, L. M. & Zhurkin, V. B. DNA sequence-dependent deformability deduced from protein–DNA crystal complexes. *Proceedings of the National Academy of Sciences* **95**, 11163-11168 (1998).

31 El Hassan, M. & Calladine, C. Conformational characteristics of DNA: empirical classifications and a hypothesis for the conformational behaviour of dinucleotide steps. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences* **355**, 43-100 (1997).

32 Lowary, P. & Widom, J. New DNA sequence rules for high affinity binding to histone octamer and sequence-directed nucleosome positioning. *Journal of molecular biology* **276**, 19-42 (1998).

33 Brogaard, K., Xi, L., Wang, J.-P. & Widom, J. A map of nucleosome positions in yeast at base-pair resolution. *Nature* **486**, 496-501 (2012).

34 Crothers, D. M., Haran, T. E. & Nadeau, J. G. Intrinsically bent DNA. *J. Biol. Chem* **265**, 7093-7096 (1990).

35 Koo, H.-S., Wu, H.-M. & Crothers, D. M. DNA bending at adenine· thymine tracts. *Nature* **320**, 501-506 (1986).

36 Hagerman, P. J. Sequence dependence of the curvature of DNA: a test of the phasing hypothesis. *Biochemistry* **24**, 7033-7037 (1985).

37 Wu, H.-M. & Crothers, D. M. The locus of sequence-directed and protein-induced DNA bending. *Nature* **308**, 509-513 (1984).

38 Stefl, R., Wu, H., Ravindranathan, S., Sklenář, V. & Feigon, J. DNA A-tract bending in three dimensions: solving the dA4T4 vs. dT4A4 conundrum. *Proceedings of the National Academy of Sciences* **101**, 1177-1182 (2004).

39 Klemm, S. L., Shipony, Z. & Greenleaf, W. J. Chromatin accessibility and the regulatory epigenome. *Nature Reviews Genetics* **20**, 207-220 (2019).

40 Jiang, C. & Pugh, B. F. Nucleosome positioning and gene regulation: advances through genomics. *Nature Reviews Genetics* **10**, 161-172 (2009).

41 Xu, Z. *et al.* Bidirectional promoters generate pervasive transcription in yeast. *Nature* **457**, 1033-1037 (2009).

42 Krietenstein, N. *et al.* Genomic nucleosome organization reconstituted with pure proteins. *Cell* **167**, 709-721. e712 (2016).
43 Oberbeckmann, E. *et al.* Genome information processing by the INO80 chromatin remodeler positions nucleosomes. *Nature communications* **12**, 1-19 (2021).
44 Cloutier, T. E. & Widom, J. Spontaneous sharp bending of double-stranded DNA. *Molecular cell* **14**, 355-362 (2004).
45 Drew, H. R. & Travers, A. A. DNA bending and its relation to nucleosome positioning. *Journal of molecular biology* **186**, 773-790 (1985).
46 Hayes, J. J., Tullius, T. D. & Wolffe, A. P. The structure of DNA in a nucleosome. *Proceedings of the National Academy of Sciences* **87**, 7405-7409 (1990).
47 Widlund, H. R. *et al.* Nucleosome structural features and intrinsic properties of the TATAAACGCC repeat sequence. *Journal of Biological Chemistry* **274**, 31847-31852 (1999).
48 Shrader, T. E. & Crothers, D. M. Artificial nucleosome positioning sequences. *Proceedings of the National Academy of Sciences* **86**, 7418-7422 (1989).
49 Rohs, R. *et al.* The role of DNA shape in protein–DNA recognition. *Nature* **461**, 1248-1253 (2009).
50 Zhou, T. *et al.* Quantitative modeling of transcription factor binding specificities using DNA shape. *Proceedings of the National Academy of Sciences* **112**, 4654-4659 (2015).
51 Barozzi, I. *et al.* Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Molecular cell* **54**, 844-857 (2014).
52 Li, J. *et al.* Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. *Nucleic acids research* **45**, 12877-12887 (2017).
53 El Hassan, M. & Calladine, C. Propeller-twisting of base-pairs and the conformational mobility of dinucleotide steps in DNA. *Journal of molecular biology* **259**, 95-103 (1996).
54 Dans, P. D., Perez, A., Faustino, I., Lavery, R. & Orozco, M. Exploring polymorphisms in B-DNA helical conformations. *Nucleic acids research* **40**, 10668-10678 (2012).
55 Czapla, L., Swigon, D. & Olson, W. K. Sequence-dependent effects in the cyclization of short DNA. *Journal of chemical theory and computation* **2**, 685-695 (2006).
56 Pérez, A. *et al.* Impact of methylation on the physical properties of DNA. *Biophysical journal* **102**, 2140-2148 (2012).
57 Tippin, D. & Sundaralingam, M. Nine polymorphic crystal structures of d (CCGGGCCCGG), d (CCGGGCCm5CGG), d (Cm5CGGGCCm5CGG) and d (CCGGGCC (Br) 5CGG) in three different conformations: effects of spermine binding and methylation on the bending and condensation of A-DNA. *Journal of molecular biology* **267**, 1171-1185 (1997).
58 Moyle-Heyrman, G. *et al.* Chemical map of Schizosaccharomyces pombe reveals species-specific features in nucleosome positioning. *Proceedings of the National Academy of Sciences* **110**, 20158-20163 (2013).
59 Gilchrist, D. A. *et al.* Pausing of RNA polymerase II disrupts DNA-specified nucleosome organization to enable precise gene regulation. *Cell* **143**, 540-551 (2010).
60 Garcia, H. G. *et al.* Biological consequences of tightly bent DNA: the other life of a macromolecular celebrity. *Biopolymers: Original Research on Biomolecules* **85**, 115-130 (2007).
61 Braccioli, L. & de Wit, E. CTCF: a Swiss-army knife for genome organization and transcription regulation. *Essays in biochemistry* **63**, 157-165 (2019).

62 Voong, L. N. *et al.* Insights into nucleosome organization in mouse embryonic stem cells through chemical mapping. *Cell* **167**, 1555-1570. e1515 (2016).

63 Wiechens, N. *et al.* The chromatin remodelling enzymes SNF2H and SNF2L position nucleosomes adjacent to CTCF and other transcription factors. *PLoS genetics* **12**, e1005940 (2016).

64 Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. *Cell* **133**, 1106-1117 (2008).

65 Clarkson, C. T. *et al.* CTCF-dependent chromatin boundaries formed by asymmetric nucleosome arrays with decreased linker length. *Nucleic acids research* **47**, 11181-11196 (2019).

66 Wang, J.-P. Z. & Widom, J. Improved alignment of nucleosome DNA sequences using a mixture model. *Nucleic acids research* **33**, 6743-6755 (2005).

67 Fire, A., Alcazar, R. & Tan, F. Unusual DNA structures associated with germline genetic activity in Caenorhabditis elegans. *Genetics* **173**, 1259-1273 (2006).

68 Moreno-Herrero, F., Seidel, R., Johnson, S. M., Fire, A. & Dekker, N. H. Structural analysis of hyperperiodic DNA from Caenorhabditis elegans. *Nucleic Acids Research* **34**, 3057-3066 (2006).

69 Pugh, B. F. & Venters, B. J. Genomic organization of human transcription initiation complexes. *PloS one* **11**, e0149339 (2016).

70 Kornberg, R. D. The molecular basis of eukaryotic transcription. *Proceedings of the National Academy of Sciences* **104**, 12955-12961 (2007).

71 Cormack, B. P. & Struhl, K. The TATA-binding protein is required for transcription by all three nuclear RNA polymerases in yeast cells. *Cell* **69**, 685-696 (1992).

72 Kim, Y., Geiger, J., Hahn, S. & Sigler, P. B. Crystal structure of a yeast TBP/TATA-box complex. *Nature* **365**, 512-520 (1993).

73 Wu, J., Parkhurst, K. M., Powell, R. M., Brenowitz, M. & Parkhurst, L. J. DNA bends in TATA-binding protein· TATA complexes in solution are DNA sequence-dependent. *Journal of Biological Chemistry* **276**, 14614-14622 (2001).

74 Rossi, M. J. *et al.* A high-resolution protein architecture of the budding yeast genome. *Nature* **592**, 309-314 (2021).

75 Rivetti, C., Guthold, M. & Bustamante, C. Wrapping of DNA around the E. coli RNA polymerase open promoter complex. *The EMBO journal* **18**, 4464-4475 (1999).

76 Thomason, M. K. *et al.* Global transcriptional start site mapping using differential RNA sequencing reveals novel antisense RNAs in Escherichia coli. *Journal of bacteriology* **197**, 18-28 (2015).

77 Basu, A. *et al.* Dynamic coupling between conformations and nucleotide states in DNA gyrase. *Nature chemical biology* **14**, 565-574 (2018).

78 Oram, M., Travers, A. A., Howells, A. J., Maxwell, A. & Pato, M. L. Dissection of the bacteriophage Mu strong gyrase site (SGS): significance of the SGS right arm in Mu biology and DNA gyrase mechanism. *Journal of bacteriology* **188**, 619-632 (2006).

79 Oram, M. & Pato, M. L. Mu-like prophage strong gyrase site sequences: analysis of properties required for promoting efficient Mu DNA replication. *Journal of bacteriology* **186**, 4575-4584 (2004).

80 Fleischmann, R. D. *et al.* Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**, 496-512 (1995).

81 Morgan, G. J., Hatfull, G. F., Casjens, S. & Hendrix, R. W. Bacteriophage Mu genome sequence: analysis and comparison with Mu-like prophages in Haemophilus, Neisseria and Deinococcus. *Journal of molecular biology* **317**, 337-359 (2002).
82 Huo, Y.-X. *et al.* IHF-binding sites inhibit DNA loop formation and transcription initiation. *Nucleic acids research* **37**, 3878-3886 (2009).
83 Travers, A. DNA–protein interactions: IHF-the master bender. *Current Biology* **7**, R252-R254 (1997).
84 Revyakin, A., Liu, C., Ebright, R. H. & Strick, T. R. Abortive initiation and productive initiation by RNA polymerase involve DNA scrunching. *Science* **314**, 1139-1143 (2006).
85 Ma, J., Bai, L. & Wang, M. D. Transcription under torsion. *Science* **340**, 1580-1583 (2013).
86 Rohs, R. *et al.* Origins of specificity in protein-DNA recognition. *Annual review of biochemistry* **79**, 233 (2010).
87 Ohno, M. *et al.* Sub-nucleosomal genome structure reveals distinct nucleosome folding motifs. *Cell* **176**, 520-534. e525 (2019).
88 Lieberman-Aiden, E. *et al.* Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *science* **326**, 289-293 (2009).

89 Basu, A. (2022, Aug 29) aakashbasu2/Intrinsic-Cyclizability-Prediction-Codes: v1.0.0 [Computer Software]. Zenodo. https://doi.org/10.5281/zenodo.7031125

90 Chereji, R. V., Ramachandran, S., Bryson, T. D. & Henikoff, S. Precise genome-wide mapping of single nucleosomes and linkers in vivo. *Genome biology* **19**, 1-20 (2018).

Methods:

Primer and adapter sequences for loop-seq

25 nt left adapter: 5' - TTTCTTCACTTATCTCCCACCGTCC - 3'

25 nt right adapter: 5' - GGCAGAAGACAAGGGAACGAAATAG - 3'

Example sequence is a loop-seq library: 5’ – Left adapter – 50 nt variable region – Right adapter – 3’

Primer P1: 5’ – CAGAATCCGTCGAAGAGCCTTATCTCCCACCGTCC – 3’

Primer P2: 5’ – ACGGATTCTGCGAAGAGCTTCCCTTG/iBiodT/CTTCTGCC – 3’

Primer P2-Biotin26: 5’ – ACGGATTCTGCGAAGAGCTTCCCTTG/iBiodT/CTTCTGCC – 3’

Primer P2-Biotin29: 5’ – ACGGATTCTGCGAAGAGCTTCCCTTGTCT/iBiodT/CTGCC – 3’

Primer P2-Biotin31: 5’ – ACGGATTCTGCGAAGAGCTTCCCTTGTCTTC/iBiodT/GCC – 3’

Primer SP1: 5’ – TCGTCGGCAGCGTCAGATGTGTATAAGAGACAGCAGAATCCGTC GAAGAGCCTTATCTCCCACCGTCC – 3’

Primer SP2: 5’ – GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAGACGGATTCTGC GAAGAGCTTCCCTTGTCTTCTGCC – 3’

Nucleotide sequences

All DNA sequences that were part of various libraries where loop-seq was performed in this study, or were sequences around genomic loci where intrinsic cyclizability values were predicted as part of this study, are reported in Supplementary Tables 1 – 21. Please refer to the Key for a description of each table’s content.

Loop-seq

Loop-seq was performed according to protocol published earlier[91]. Libraries of ssDNA, containing ~90,000 different sequences, with the central 50 nt variable and flanked by 25 nt left and right adapters were obtained (Genscript), amplified via emulsion PCR (Micellula DNA Emulsion and Purification Kit (CHIMERx)) using primers P1 and P2-Biotin26, using KAPA polymerase (Roche). These primers introduce a biotin molecule, as well as the binding site for the site specific nicking enzyme Nt.BspQ1 (New England Biolabs (NEB)). The amplified library DNA fragments were bound to streptavidin-coated magnetic beads (Dynabeads MyOne Streptavidin T1, Thermo Fisher Scientific). Whenever buffer exchange was required, the magnetic beads were pulled down with rare earth magnets. The library molecules immobilized on the beads were nicked near the ends using Nt.BspQ1 to general 10 nt single-stranded complementary overhangs. The nicking solution was replaced with buffer T2.5BSA (2.375 mM Sodium Chloride, 9.5 mM tris HCl pH 8, 1 mg/ml BSA). Looping was initiated by replacing T2.5BSA with looping buffer (1M Sodium Chloride, 20 mM tris pH 8, 1 mg/ml BSA). After 1 minute incubation in looping buffer, the beads were incubated for 20 minutes in digestion buffer (1x NEB Buffer 4 (NEB), 1 mM ATP, 0.333 units/ml

RecBCD (NEB)), which digests those molecules that failed to loop in the prior looping step. Digestion buffer was then replaced with looping buffer and prepared for sequencing. The above protocol was also repeated for an identically prepared control where the digestion enzyme RecBCD was omitted.

The libraries (sample and control) were then prepared for sequencing. DNA was PCR amplified off the beads using primers P1 and P2, followed by PCR amplification using primers SQ1 and SQ2, followed by index primers available in the Nexters XT Index Kit (24 Indices 96 Samples) (Illumina). Sequencing was performed on Illumina MiSeq (for libraries with less than 20,000 sequneces) or HiSeq platforms. Sequencing results were mapped to the known sequences in the library using Bowtie 1 and SAMtools version 1.12. Cyclizability of a sequence was defined as the ratio of the relative population of the sequence in the amplified library after loop-seq, to that in the control sample, where all steps were identical except that unlooped molecules were not digested. These calculations were performed using simple MATLAB (Matworks) versions 9.0, 9.2, 9.4, and 9.6 scripts that were developed previously[91].

For any given library, loop-seq was repeated three times for three different biotin locations[27] (where primers P2-Biotin29 and P2-Biotin31 were used in the initial amplification step instead of P2-Biotin26). For each sequence, cyclizability was modeled to have a sinusoidal dependence on the position of the biotin, with periodicity equal to that of the helical repeat. The constant term of the sinusoid was defined as the intrinsic cyclizability[27].

CpG Methylation of libraries

Libraires to be methylated were first amplified via emulsion PCR as done as part of the regular loop-seq protocol[91]. Prior to immobilization on streptavidin-coated magnetic beads, CpG methylation of the amplified library was carried out using CpG Methyltransferase (New England Biolabs, product number M0226S), by following the manufacturer's protocol. The methylation mixture involved amplified DNA library (~120 ng), 10x methyltransferase reaction buffer (New England Biolabs) (5 µl), 1600 µM SAM (5 µl), Methyltransferase (4 New England Biolabs units), and water (up to 50 µl). The reaction mixture was

incubated at 37 ºC for 1 hour. Heat inactivation was not performed in order to prevent denaturation and incorrect annealing of the library.

Statistics and reproducibility

All presented loop-seq data in figures were compiled from a single run of loop-seq on the library in question. The reproducibility of the loop-seq technique has been established earlier[53].

All Pearson's r, and associated P values, were calculated using the 'corrcoef' function in MATLAB (Matworks), as described earlier[53].

Data Availability:

All new sequencing data reported as part of this study are deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under accession number PRJNA746342. All measurements of intrinsic cyclizabilities obtained from our earlier study[27] were based on sequencing data that is deposited in the National Center for Biotechnology Information (NCBI) Sequence Read Archive (SRA) under accession number PRJNA667271.

The following datasets usen in this study were downloaded from the NCBI Gene Expression omnibus (GEO) using the following accession numbers: GSE97290 (nucleosome occupancy data around +1 nucleosome dyads in *S. cerevisiae*); GSE46957 (nucleosome occupancy data around +1 nucleosome dyads in *S. pombe*); GSE69336 (nucleosome occupancy data in *D. melanogaster* around TSSs); GSE82127 (nucleosome occupancy in *M. musculus* around TSSs and CTCF binding sites); GSE11431 (location of CTCF binding sites in mouse embryonic stem cells); GSE147927 (location of TSSs in *S. cerevisiae*); GSE55199 (TSS locations in *E. coli*);

S. cerevisiae (sacCer3), *S. pombe* (spo2), *D. melanogaster* (BDGP5/ dm3) and *M. musculus* (mm9) genome sequences were downloaded from the UCSC Genome Browser (https://genome.ucsc.edu/cgi-bin/hgGateway). The *E. coli* MG1655 genome was downloaded from NCBI

(accession number NC_000913.2). The *H. influenza* genome was downloaded from NCBI GeneBank (L42023.1). The sequence of the Ω4 region of the *C. elegans* genome was downloaded from the supplementary material of this publication: Moreno-Herrero, F., Seidel, R., Johnson, S. M., Fire, A. & Dekker, N. H. Structural analysis of hyperperiodic DNA from Caenorhabditis elegans. *Nucleic Acids Research* **34**, 3057-3066 (2006)[68].

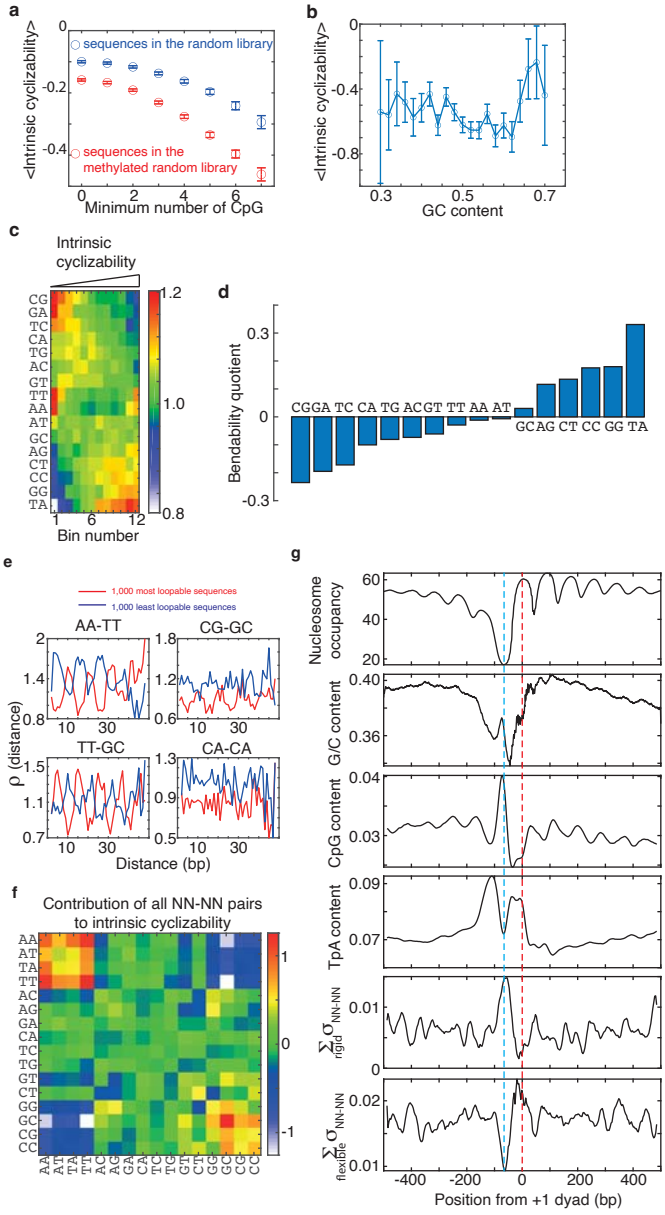
Supplementary Tables 1 – 21 provide the following data: Sequences and intrinsic cyclizability values in all libraries on which loop-seq was performed either in this study or earlier[27]; All sequences and predicted intrinsic cyclizability values around all genomic loci where we applied our predictive models to predict intrinsic cyclizability; The values of all parameters that quantify the contribution of short sequence features and their distributions to intrinsic cyclizability, as obtained in this study.

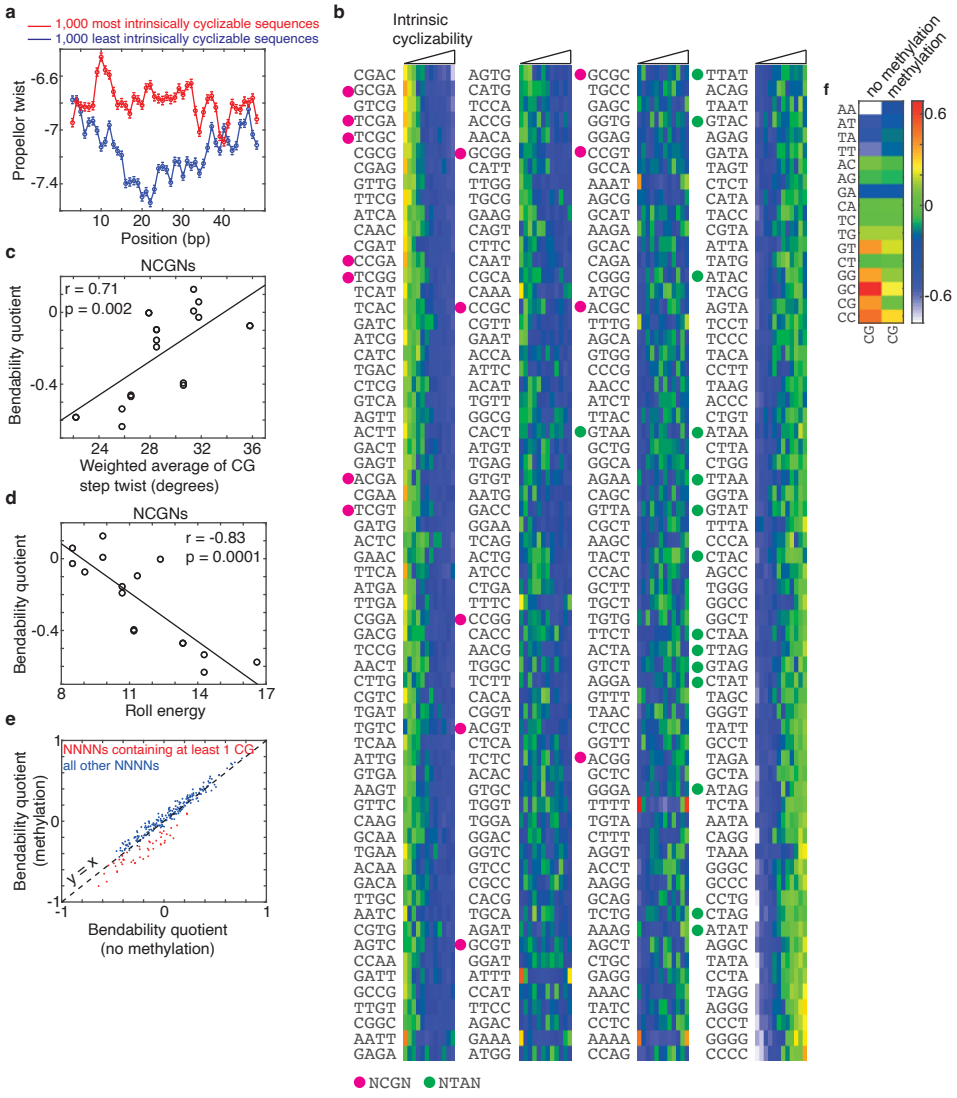
Code availability:

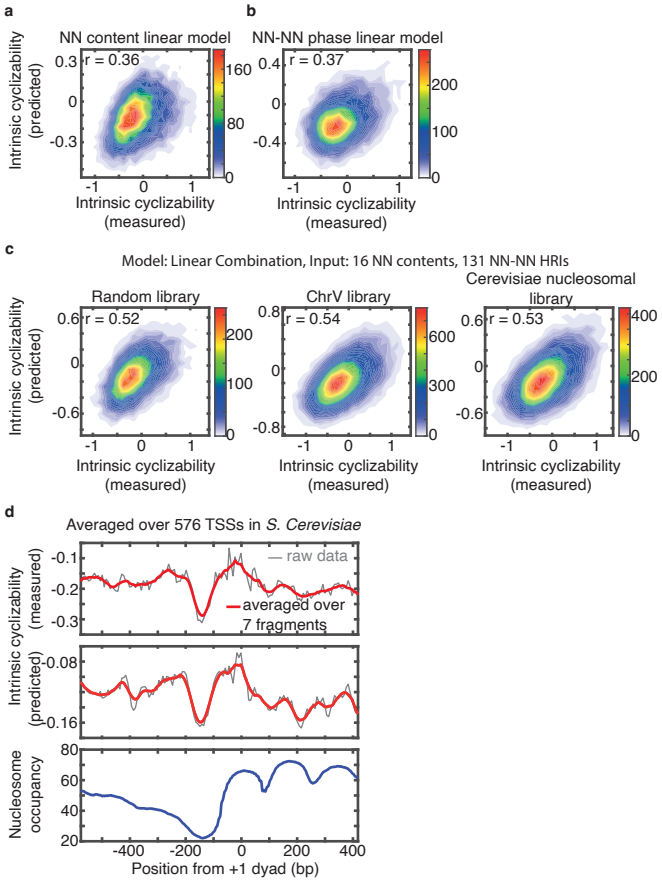
Custom MATLAB codes developed as part of this study for predicting intrinsic cyclizability based on linear regression models or neural nets have been deposited in Zenodo[89]: Basu, A. Intrinsic-Cyclizability-Prediction-Codes: v1.0.0. Zenodo. https://doi.org/10.5281/zenodo.7031125. (2022, Aug 29).

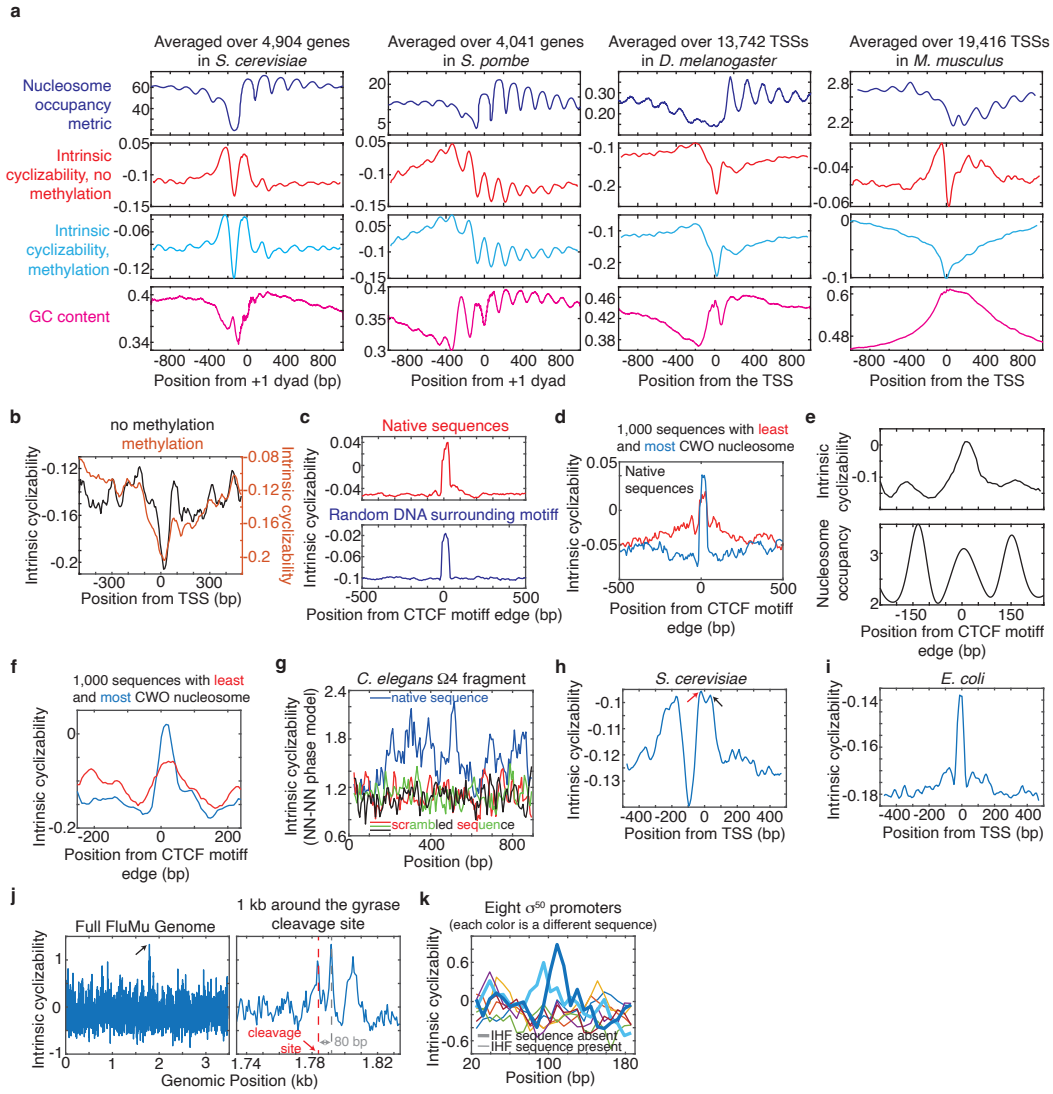
Methods only references:

91 Basu, A. in *Methods in Enzymology* Vol. 661 305-326 (Elsevier, 2021).


a
<Intrinsic cyclizability>
sequences in the random library
sequences in the methylated random library
Minimum number of CpG
b
<Intrinsic cyclizability>
GC content
c
Intrinsic cyclizability
Bin number
d
Bendability quotient
CGGA TC CA TG ACGT TT AA AT
GC AG CT CC GG TA
e
1,000 most loopable sequences
1,000 least loopable sequences
AA-TT
CG-GC
TT-GC
CA-CA
ρ (distance)
Distance (bp)
f
Contribution of all NN-NN pairs to intrinsic cyclizability
g
Nucleosome occupancy
G/C content
CpG content
TpA content
Position from +1 dyad (bp)


a
1,000 most intrinsically cyclizable sequences
1,000 least intrinsically cyclizable sequences
Propellor twist
-6.6
-7
-7.4
10
20
30
40
Position (bp)
b
Intrinsic cyclizability
NCGN
NTAN
f
no methylation
methylation
0.6
0
-0.6
c
NCGNs
r = 0.71
p = 0.002
Bendability quotient
0
-0.4
24
28
32
36
Weighted average of CG step twist (degrees)
d
NCGNs
r = -0.83
p = 0.0001
Bendability quotient
0
-0.4
8
11
14
17
Roll energy
e
NNNNs containing at least 1 CG
all other NNNNs
y = x
Bendability quotient (methylation)
1
0
-1
-1
0
1
Bendability quotient (no methylation)


a
b
NN content linear model
NN-NN phase linear model
Intrinsic cyclizability (predicted)
r = 0.36
r = 0.37
Intrinsic cyclizability (measured)
c
Model: Linear Combination, Input: 16 NN contents, 131 NN-NN HRIs
Random library
ChrV library
Cerevisiae nucleosomal library
r = 0.52
r = 0.54
r = 0.53
d
Averaged over 576 TSSs in S. Cerevisiae
raw data
averaged over 7 fragments
Intrinsic cyclizability (measured)
Intrinsic cyclizability (predicted)
Nucleosome occupancy
Position from +1 dyad (bp)


a
Averaged over 4,904 genes in S. cerevisiae
Averaged over 4,041 genes in S. pombe
Averaged over 13,742 TSSs in D. melanogaster
Averaged over 19,416 TSSs in M. musculus
Nucleosome occupancy metric
Intrinsic cyclizability, no methylation
Intrinsic cyclizability, methylation
GC content
Position from +1 dyad (bp)
Position from +1 dyad
Position from the TSS
Position from the TSS
b
no methylation
methylation
Intrinsic cyclizability
Position from TSS (bp)
c
Native sequences
Random DNA surrounding motiff
Intrinsic cyclizability
Position from CTCF motiff edge (bp)
d
1,000 sequences with least and most CWO nucleosome
Native sequences
Intrinsic cyclizability
Position from CTCF motiff edge (bp)
e
Intrinsic cyclizability
Nucleosome occupancy
Position from CTCF motiff edge (bp)
f
1,000 sequences with least and most CWO nucleosome
Intrinsic cyclizability
Position from CTCF motiff edge (bp)
g
C. elegans Ω4 fragment
native sequence
scrambled sequence
Intrinsic cyclizability (NN-NN phase model)
Position (bp)
h
S. cerevisiae
Intrinsic cyclizability
Position from TSS (bp)
i
E. coli
Intrinsic cyclizability
Position from TSS (bp)
j
Full FluMu Genome
1 kb around the gyrase cleavage site
Intrinsic cyclizability
cleavage site
80 bp
Genomic Position (kb)
k
Eight σ50 promoters
(each color is a different sequence)
Intrinsic cyclizability
IHF seqdence absent
IHF sequence present
Position (bp)