Assessing base-resolution DNA mechanics on the genome scale

Jiang, Wen-Jie; Hu, Congcong; Lai, Futing; Pang, Weixiong; Yi, Xinyao; Xu, Qianyi; Wang, Haojie; Zhou, Jialu; Zhu, Hanwen; Zhong, Chunge; Kuang, Zeyu; Fan, Ruiqi; Shen, Jing; Zhou, Xiaorui; Wang, Yu-Juan; Wong, Catherine C.L.; Zheng, Xiaoqi; Wu, Hua-Jun

Published in:
Nucleic Acids Research

Published: 13/10/2023

Document Version:
Final Published version, also known as Publisher's PDF, Publisher's Final version or Version of Record

Published version (DOI):
10.1093/nar/gkad720

Publication details:
Jiang, W.-J., Hu, C., Lai, F., Pang, W., Yi, X., Xu, Q., Wang, H., Zhou, J., Zhu, H., Zhong, C., Kuang, Z., Fan, R., Shen, J., Zhou, X., Wang, Y.-J., Wong, C. C. L., Zheng, X., & Wu, H.-J. (2023). Assessing base-resolution DNA mechanics on the genome scale. *Nucleic Acids Research*, *51*(18), 9552-9566. https://doi.org/10.1093/nar/gkad720

9552–9566 *Nucleic Acids Research, 2023, Vol. 51, No. 18* *Published online 11 September 2023*
https://doi.org/10.1093/nar/gkad720

Assessing base-resolution DNA mechanics on the genome scale

Wen-Jie Jiang[1,2], Congcong Hu[3], Futing Lai[2], Weixiong Pang[4], Xinyao Yi[3], Qianyi Xu[5], Haojie Wang[6], Jialu Zhou[7], Hanwen Zhu[2], Chunge Zhong[8], Zeyu Kuang[2], Ruiqi Fan[9], Jing Shen[9], Xiaorui Zhou[1], Yu-Juan Wang[1], Catherine C.L. Wong[10,*], Xiaoqi Zheng[11,*] and Hua-Jun Wu[1,2,*]

[1]Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Peking University Cancer Hospital and Institute, 100142 Beijing, China, [2]School of Basic Medical Sciences, Center for Precision Medicine Multi-Omics Research, Peking University Health Science Center, 102206 Beijing, China, [3]Department of Mathematics, Shanghai Normal University, 200234 Shanghai, China, [4]Department of Mathematics, Shanghai Ocean University, 201306 Shanghai, China, [5]University of California, San Diego, CA 92103, USA, [6]Institute of Genetics and Developmental Biology, Chinese Academy of Sciences, 100101 Beijing, China, [7]Department of Gynecology and Obstetrics, Chinese PLA General Hospital, 100853 Beijing, China, [8]College of Life and Health Sciences, Northeastern University, 110819 Shenyang, China, [9]Central Laboratory, Key Laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Peking University Cancer Hospital and Institute, 100142 Beijing, China, [10]Department of Medical Research Center, State Key Laboratory of Complex Severe and Rare Diseases, Peking Union Medical College Hospital, Chinese Academy of Medical Science & Peking Union Medical College, 100730 Beijing, China and [11]Center for Single-Cell Omics, School of Public Health, Shanghai Jiao Tong University School of Medicine, 200025 Shanghai, China

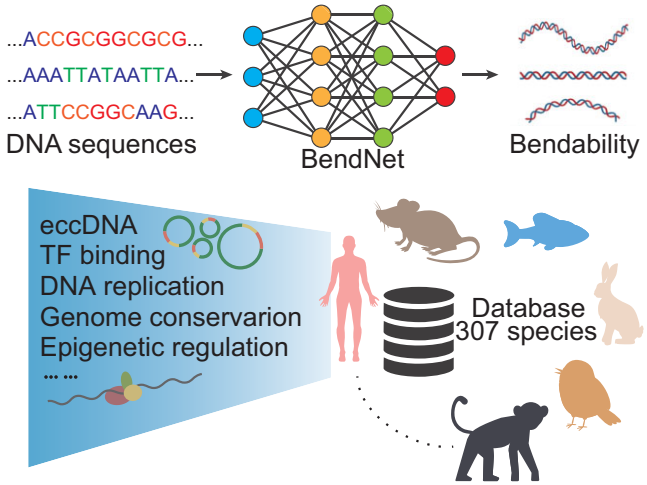
Received February 13, 2023; Revised August 09, 2023; Editorial Decision August 13, 2023; Accepted August 18, 2023

ABSTRACT

Intrinsic DNA properties including bending play a crucial role in diverse biological systems. A recent advance in a high-throughput technology called loop-seq makes it possible to determine the bendability of hundred thousand 50-bp DNA duplexes in one experiment. However, it's still challenging to assess base-resolution sequence bendability in large genomes such as human, which requires thousands of such experiments. Here, we introduce 'BendNet'—a deep neural network to predict the intrinsic DNA bending at base-resolution by using loop-seq results in yeast as training data. BendNet can predict the DNA bendability of any given sequence from different species with high accuracy. To explore the utility of BendNet, we applied it to the human genome and observed DNA bendability is associated with chromatin features and disease risk regions involving transcription/enhancer regulation, DNA replication, transcription factor binding and extrachromosomal circular DNA generation. These findings expand our understanding on DNA mechanics and its association with transcription regulation in mammals. Lastly, we built a comprehensive resource of genomic DNA bendability profiles for 307 species by applying BendNet, and provided an online tool to assess the bendability of user-specified DNA sequences (http://www.dnabendnet.com/).

*To whom correspondence should be addressed. Tel : +86 010 82805976; Email: hjwu@pku.edu.cn
Correspondence may also be addressed to Xiaoqi Zheng. Tel: +86 21 63846590; Email: xqzheng@shsmu.edu.cn
Correspondence may also be addressed to Catherine C.L. Wong. Tel: +86 010 69154952; Email: catclw321@126.com

GRAPHICAL ABSTRACT



INTRODUCTION

DNA-protein interactions are essential for many key cellular processes, including DNA replication (1), chromatin formation (2) and transcriptional regulation (3). These interactions require DNA bending to embrace proteins, which involves intrinsic properties of DNA fragments likely below 100 base-pair (bp) (4). Assays, such as electrophoretic mobility (5) and single-molecule fluorescence resonance energy transfer (smFRET) (6), have been developed to determine the DNA bending ability (termed as DNA bendability). Specifically, they measure the looping rate of a single DNA fragment of approximately 100 bp in length at a time, therefore is limited by its low throughput. Recent advance in a sequencing-based approach called loop-seq (7,8) has led to the vast increase in throughput, which has scaled-up looping rate detection from dozens to hundred thousand DNA duplexes in one study through combining smFRET and systematic enrichment of ligands by exponential enrichment (SELEX) (9) selection methods. In the same study, they applied loop-seq to demonstrate the contribution of DNA bendability to nucleosome organization in yeast by measuring the DNA looping rate that tiles the regions of interest at 7 bp resolution (8). However, it remains challenging to tile larger genomes such as human at base resolution, because library preparation of this method requires synthesizing equivalent numbers of DNA duplexes to the genome size that relies on constructing thousands of such libraries, and this does not consider the sequence variations between individuals.

To extend bendability assessment to DNA duplexes in large sequence space, we developed a deep learning-based method termed BendNet, which extracts and learns sequence features encoding DNA bendability using capsule networks (10) without routing (11). BendNet can predict DNA bendability of any given sequence with a high agreement to that measured by loop-seq and other low throughput approaches. We also demonstrated that BendNet can accurately provide a finer map of DNA bendability in yeast whole genome at base-resolution as compared to the original loop-seq study (8), resulting in the same biological findings. To explore the utility of BendNet, we applied it to the human genome and obtained base-resolution map of DNA bendability. We observed a rigid region (showing relatively low bendability) located at transcriptional start/end sites and enhancer centers. DNA bendability is associated with chromatin states, epigenetic markers, G4 structures, replication timing and the frequency of SNPs in GWAS catalog. The GC content is an intrinsic determinant of DNA bendability, particularly in enhancers. Most transcription factors (TFs) bind rigid DNA sequences, while some specific TFs, such as EBF1 and CTCF, show an unusually high bendability within their binding elements. Analyses of *in silico* mutagenesis on 840 TF motifs revealed important nucleotide-level features at the motif center that impact DNA bending. Our work provides a tool to assess DNA bendability for large-scale DNA sequences and expands our understanding of DNA mechanics in chromatin regulation. To make a comprehensive resource of DNA bendability, we applied BendNet to whole genomes of 307 species and implemented an easy-to-use interface through a webserver.

MATERIALS AND METHODS

Training and evaluation data

Our model was trained and evaluated based on a set of DNA fragments whose bendability is measured by loop-seq. This dataset consists of 270 806 DNA duplexes from *S. cerevisiae*, measured by five independent experiments including random, reference and codon-altered sequences in different regions of interest. After removing outliers showing extremely high frequency in each individual experiment, and averaging the bendability of DNA duplexes with multiple measurements, we obtained in total 264 860 valid DNA duplexes and their bendability measurements as our processed data for the following analyses.

We trained two separated models to comprehensively evaluate our BendNet method. For training the primary model, 264 860 DNA duplexes were randomly split into 70% training (185 402), 20% validation (52 972) and 10% hold-out test (26 486) sets. For training the second model, we hold out one of the five independent experimental data which contains bendability of DNA duplexes tiling the chromosome V of yeast genome in 7-bp resolution as test set (82 404), and randomly split the left data into training (146 010) and validation sets (36 502) with the splitting ratio of 8:2.

Model architecture

The BendNet architecture consists of a convolutional module and a capsule module. The convolutional module includes multiple sequential convolutional blocks, each of which has three stacked convolutional layers with increased numbers of kernels followed by dropout and batch normalization. The output of each convolutional block is fed into a capsule block, which contains a capsule, dropout and batch normalization layers. The results of all capsule blocks are stacked, and subsequently output to a fully connected layer to produce a regression score as bendability prediction. More specifically,

(1) **Input.** BendNet takes one-hot encoding of DNA sequences as input, i.e. N*50-bp DNA sequences were transformed into N*50*4 matrices, to the model. To ensure the generalization of our model, we randomly reverse complemented 50% of the input DNA sequences in training.
(2) **Convolutional module.** This module contains multiple convolutional blocks, each of which has three convolutional layers with kernel sizes 2*1, 3*1 and 4*1, respectively. To capture the global structure of the input sequences, we used increased numbers of kernels for each convolutional block. In detail, three layers of the first convolutional block have 16 kernels of size 2*1, 32 kernels of size 3*1 and 64 kernels of size 4*1, while the numbers are 32, 64 and 128 for the second convolutional block, and so on. The number of convolutional blocks is a hyperparameter to be determined by the validation set.
(3) **Capsule module.** The capsule module consists of three capsule blocks, and each capsule block contains a capsule, a dropout and a batch normalization layer.
(4) **Output.** The results of three capsule blocks are stacked and output to a dense layer to get a regression score.

BendNet with three convolutional blocks has in total 287 442 parameters, including 285 484 trainable parameters and 1958 non-trainable parameters. The mean squared error (MSE) is used as the loss function, and the Adaptive Moment Estimation algorithm is used to update the parameters. The model with the minimum MSE on the validation set in 200 epochs was used as the final model. BendNet was written in Python using the TensorFlow and Keras frameworks.

Hyperparameter optimization approach

Our model has two types of hyperparameters, i.e. structural-related hyperparameters including the number of convolutional/capsule modules, the number of capsule classes and the number of dimensions, and nonstructural-related hyperparameters including learning rate, dropout rate in convolutional and capsule blocks. We adopted different strategies to optimize two types of hyperparameters. For three structure-related hyperparameters which affect the model complexity, we used control variates to assign different values to the three hyperparameters and recorded their minimum validation loss in 100 epochs. The hyperparameters were retained when validation loss is no longer reduced significantly. In other words, if two models with different hyperparameters are comparable in accuracy, we will adopt the simpler one with smaller hyperparameters.

The genetic algorithm is adopted to tune the nonstructural-related hyperparameters. Basically, the validation mean square error is used to measure the fitness of a set of hyperparameters. By randomly generating parameters, and exchanging parameters as mutation and cross operations. We generated 50 generations in this genetic algorithm with each of which was composed of 20 individuals. The best two individuals that have the minimum validation loss in 100 epochs in each generation were selected to cross and mutate to produce the next generation. The best performance was achieved with a learning rate of 0.02585, a dropout rate of 0.17632 in the convolutional block, and a dropout rate of 0.14818 in the capsule block.

Model comparison

We compared BendNet with two widely used machine learning models – Random Forest and SVR, and six state-of-the-art deep learning models – AlexNet, VGG16, GoogleNet22, ResNet34, DNAcycP and DeepBend with default parameters on the same dataset. To train these models, we used an unbiased checkpoint strategy—we retained the model with the smallest loss in the validation set - and trained them for 200 epochs on the same device (CentOS, GTX3090Ti). To make an unbiased comparison, we included both the model prediction results of BendNet before and after hyperparameter optimization.

Independent validation datasets

(1) ***Gallus gallus***. This dataset includes the relative electrophoretic mobilities of seven DNA fragments in CTCF binding regions of Gallus gallus. The relative electrophoretic mobility of a DNA fragment measures whether the corresponding region is bendable. In detail, DNA fragments are loaded into pBEND2 plasmid, and digested with restriction enzymes into a set of probes of equal length.
(2) ***E. coli***. The dataset includes the relative length (RL) of 56 DNA fragments of length 57 bp provided by Wang *et al.* (12). According to their study, mutations of 56 non-redundant *E. coli* DNA fragments were induced *in vitro* within a 57 bp region which is located at the ilvIH operon TSS –83 bp to –140 bp. RL is defined as the ratio of the apparent length to the actual length, which is proportional to the bendability of the DNA fragment.
(3) ***Human***. The dataset contains Rbound/Rfree ratios of the 35 fragments which were composed of five types of 20 bp p53 response elements (p53 consensus sequence, p21/waf1/cip1, symmetric sequence, ribosomal gene cluster sequence (RGC), SV40 promoter sequence) and five recombinant plasmids (pCon30, pWaf30, pSS30, pRGC30 and pSV30) in human. These plasmids contain the above five types of p53 binding sites flanked by tandemly duplicated DNA sequences. The plasmids were cleaved at the seven restriction sites and 35 DNA fragments were obtained. The Rbound/Rfree ratio was calculated by the p53DBD-DNA complex's electrophoretic mobility compared to free DNA. A fragment with a large Rbound/Rfree ratio is supposed to be more bendable.
(4) ***Mouse***. We obtained the bendability of over 90 000 mouse genomic DNA fragments from a recent loop-seq experiment (13).

Genome annotations and features in human

The annotation of the protein-coding gene (PCG), long non-coding RNA (lncRNA), non-coding RNA (ncRNA) and pseudogene were obtained from the Gencode GRCh38.p13 version of human genome. Housekeeping genes and tissue-specific genes were obtained from

a previous study (14). Sequence conservation was obtained from the UCSC 'phastCons46wayPlacental' track.

DNA bendability groups of protein-coding genes

We categorized all protein-coding genes (PCGs) into three equally sized groups based on their average bendability within the regions spanning from TSS –20 to +70 bp. These groups were labeled as rigid, intermediate and bendable, corresponding to the low, intermediate and high levels of bendability observed in the respective regions respectively. Meanwhile, we also performed the same analysis on transcription end sites (TESs).

Enhancer and super-enhancer annotation

We downloaded enhancers of the human genome from Fantom5 (15) and used bedtools (16) to remove enhancers that overlap with promoters. Then the remaining enhancers were used for further analyses. Super-Enhancers were downloaded from CircleBase (17).

DNA bendability analysis in TSSs and TESs

We obtained ATAC-seq, DNase-seq, MNase-seq, H3K4me3, H3K27ac, H3K79me2 and POLR2A signals in GM12878, K562 and HepG2 cell lines from ENCODE. The sequencing signals in ±500 bp of rigid, intermediate and bendable TSSs were obtained and compared through the Wilcoxon rank sum test.

Association between DNA bendability and replication timing

We obtained replication timing data in three cell lines from Massey *et al.* (18). We divided genomic regions into 100 groups with equal numbers of DNA fragments based on replication timing, ranging from low (late) to high (early). We then calculated the average replication timing signal and DNA bendability for each group, and visualized their association using scatter plots.

DNA bendability at G4 structure

We obtained G4 regions from a previous G4-seq study (19) and analyzed DNA bendability distribution in those regions.

DNA bendability analysis in chromatin states and GWAS SNPs

We obtained 15 core chromatin states in four cell lines (GM12878, A549, HepG2 and K562) from chromHMM (20). We calculated the average bendability of each 200 bp region, and reported the mean bendability for each chromatin state. To examine the correlation between DNA bendability and GC content, we computed the average DNA bendability and GC content within 200 bp genomic regions. For the GWAS SNP analysis, we fragmented the human genome into 200 bp windows, and calculated the frequency of GWAS SNPs (21) in each window. We then grouped all windows based on the frequency of GWAS SNPs, and calculated the average bendability in each group. In order to eliminate the influence of total SNPs (22) to the analysis. We calculated the corrected GWAS SNP counts for each window by dividing the number of GWAS SNP counts by the total SNP counts. These corrected counts were then used to investigate the relationship between the relative number of disease-associated variances and DNA bendability. Furthermore, we also considered the significance of GWAS SNPs in the analysis by comparing DNA bendability to mean $-\log_{10}$ significant scores of GWAS SNPs in each window.

Transcription factor binding analysis

We obtained the BigWig files and BED (peak) files of TFs in GM12878 from ENCODE. We used bwtool to extract the ChIP-seq signal and bendability in the peak regions of each TF. We employed two methods to depict the DNA bendability in the binding sites of different TFs: (i) we computed the mean profile of bendability and each TF ChIP-seq signal across all peaks in ±500 bp regions of peak centers, and calculated the Pearson correlation coefficient between them; (ii) we subtracted the average bendability of surrounding background regions (–500 bp to –100 bp and 100–500 bp) from the average bendability in peak centers (±100 bp) to obtain the relative bendability height in each of the 152 TFs. To analyze the association between bendability and TF ChIP-seq signals, we calculated the average ChIP-seq signal per peak, and identified the top and bottom 25% peak groups for each TF. The average bendability per peak in the two groups was then calculated and compared by using the Wilcoxon test. For TF binding and co-binding analysis, the ChIP-seq peaks of each TF were overlapped with rigid and bendable TSSs (±500 bp region), and the number of overlapping peaks was calculated and compared.

Protein interaction network and enrichment analysis

The protein-protein interactions were obtained from string database (https://string-db.org/). The enrichment analysis was performed under the GO term 'cell composition'.

eccDNA analysis

We obtained eccDNAs of five categories (ncRNA, PC: Protein-coding, pseudogene, snoRNA and snRNA) from circlebase (17) and predicted their corresponding bendability scores. We generated random sequences with the same length by bedtools (16) as the control group for the comparison.

Nucleotide effects at TF motif sites via in silico mutagenesis

We obtained the genomic loci of 840 transcription factor (TF) binding motifs from the JASPAR database (http://expdata.cmmt.ubc.ca/JASPAR/downloads/UCSC_tracks/2022/hg38/). As TFs can have variable number of motif loci in the genome, we randomly selected 10000 positions per motif. We also randomly picked the same number of genomic loci from the human genome 1000 times to generate a background sequence set. For

above two sets, the corresponding flanking sequences of 50 bp in length are extracted, with the motif sequence located at the center. We next randomly mutated each position in the sequence to the other three nucleotides, resulting in a total of 150 mutated sequences for each motif locus. We then applied BendNet to predict bendability of all sequences and investigated the nucleotide effects in the motif and flanking regions on DNA bendability.

Genome-wide DNA bendability of 307 species

We obtained the genome sequences of 307 species from the Ensembl database and applied BendNet to predict their DNA bendability. All predictions are available at (https://zenodo.org/deposit/7699690 and https://zenodo.org/deposit/7663786).

Data visualization

Heatmaps and average profile plots are generated by deeptools (23). The rest plots are drawn in the R environment using basic plot functions, ggplot2 and pheatmap packages. The Wilcoxon test is used for calculating all the *P*-values.

RESULTS

Overview of BendNet

We developed BendNet (Figure 1A), inspired by capsule network (10,11) and homogeneous vector capsules (24), to predict bendability of DNA duplex from sequence alone. The BendNet architecture consists of a convolutional module and a capsule module (Figure 1A). The convolutional module includes multiple consecutive convolutional blocks, each of which has three stacked convolutional layers with increased kernel and channel sizes followed by dropout and batch normalization layers. The output of each convolutional block is fed into a capsule block, which contains a capsule, a dropout and a batch normalization layer. The results from all the capsule blocks are stacked, and subsequently supplied to a fully connected layer to produce a regression score as bendability prediction.

We trained BendNet on the dataset of intrinsic DNA bendability measured by loop-seq (7). The dataset contains the bendability of 270 806 DNA duplexes from five independent experiments including random, reference and codon-altered sequences in different regions of interest in yeast. We first examined the data distribution, and found an abnormally high frequency at the bendability score of 0.02096348 (over 9000 DNA duplexes). We removed these data points, as these outliers can introduce bias to the prediction model in statistical learning (25), and obtained bendability of 264 860 DNA duplexes as DNABend dataset. We then split the dataset into 70% training, 20% validation and 10% hold-out test sets to train our model. BendNet learns the optimal architecture by using a genetic algorithm (26) proposed by *Mitchell* in training and validation sets (Supplementary Figure 1A–C), and achieves Pearson correlation between our predictions and experimental measurements of 0.793 in the test set (Figure 1B, C). We compared BendNet to other state-of-the-art machine learning and deep learning models including Random Forest (RF) (27), Support Vector Regression (SVR) (28), AlexNet (29), ResNet (30), GoogleNet (31), VGG (32), DNAcycP (33) and DeepBend (34). BendNet outperforms all other models in terms of prediction accuracy (Figure 1D), in the meanwhile, it is the fastest deep learning model in both training and inference processes, which saves 38% training time and 25% inference time than the second fastest model (Supplementary Figure 1D). Additionally, we conducted an unbiased comparison between BendNet (before hyperparameter optimization) and six other deep learning models. BendNet still outperformed the other models on both the test set and external dataset (13) (Supplementary Figure 1E, F).

To demonstrate the potential of BendNet in uncovering biological discoveries, we held out one of the five independent experimental datasets which contain bendability of DNA duplexes tiling the chromosome V of yeast genome in 7-bp resolution as a test set, and trained another BendNet model. The model achieved a correlation of 0.757 between the predicted values and the actual values (Supplementary Figure 1G, H). A visualization on chromosome V depicts an overall high agreement between BendNet predictions and loop-seq measurements (Figure 2A). BendNet predictions are as sensitive as loop-seq measurements in detecting DNA bendability in individual regions (Figure 2A), and successfully uncovered the previous observation of low DNA bendability within nucleosome-depleted regions (NDRs) (8).

To demonstrate the generalization of BendNet, we collected the intrinsic cyclizability of DNA duplexes measured by three distinct low-throughput experimental assays and loop-seq. The measurements include relative ionization mobilities of 9 data points of a CTCF binding region in *Gallus gallus* (35), relative lengths (RL) of 56 DNA duplexes in *E. coli* (12), Rbound/Rfree ratios of 35 DNA duplexes in human (36) and bendability of over 90 000 DNA duplexes of the mouse genome measured by loop-seq (13). Although based on distinct protocols and performed on sequences from different species, predictions by BendNet show an overall high consistency with the above experimental measurements (Figure 2B–D, Supplementary Figure 1I). In particular, BendNet achieves Pearson correlations of 0.773 for the *E. coli* dataset, 0.654 for the human dataset and 0.762 for the mouse dataset. There results demonstrated that BendNet is an accurate, computationally efficient and generalizable model suitable for large-scale DNA bendability prediction tasks.

BendNet predicts base-resolution DNA bendability in human

To explore the role of DNA bendability in mammals, we applied BendNet to the human genome and obtained base-resolution bendability predictions in the genome scale. The bendability follows an approximately normal distribution with negative skewness (Supplementary Figure 2A), which reveals an over-representation of high bendability regions. This is probably due to a higher frequency of nucleosome-occupied regions compared to NDRs in the human genome, as nucleosome-occupied regions are generally more bendable than NDRs. To further clarify the pattern of DNA bendability in different gene types, we examined it at protein-coding genes (PCGs), long non-coding RNAs (lncRNAs), non-coding RNAs (ncRNAs) and pseu-

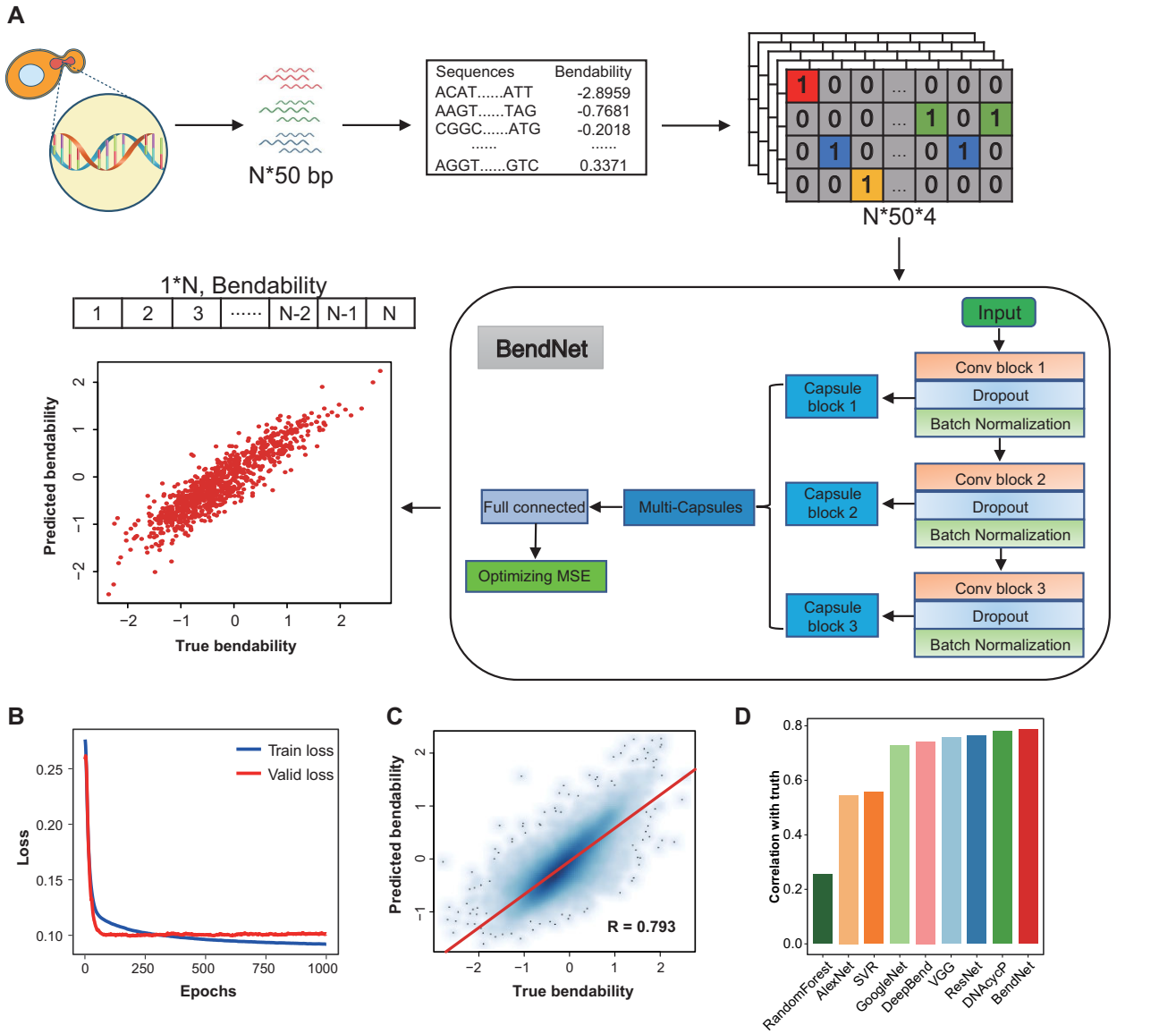


Figure 1. Overview of BendNet, a multi-capsule network for predicting bendability from DNA sequences. (**A**) Workflow of BendNet. DNA fragments of 50 bp and their bendability measured by 'loop-seq' are used to train our model. All input DNA fragments are transformed into binary matrices by one-hot encoding scheme. BendNet starts with a convolutional module, which includes three consecutive convolutional blocks with different sizes and numbers of convolution kernels. After each convolutional block, there is a capsule block consisting of three layers (capsule, dropout and batch normalization). Three capsule blocks are concatenated together and followed by a fully connected layer to get output. (**B**) Training loss and validation loss of BendNet. (**C**) Consistency between true and predicted bendability on the hold-out test set. (**D**) Accuracy comparison of different models.

dogenes. Generally, we observed a clearly defined region of rigid DNA (with low bendability) at both transcription start sites (TSSs) (Figure 3A) and transcription end sites (TESs) (Figure 3B). More specifically, PCGs exhibit the strongest bendability drop that matches the nucleosome depletion at both TSSs and TESs (Supplementary Figure 2B, C); lncRNAs show a weaker bendability decrease than PCGs, and display a more obvious change in TESs than in TSSs which coincides with the nucleosome depletion at TESs but not TSSs; ncRNAs demonstrates an oscillatory bendability pattern, with a notably low bendability at TESs which is associated with a decreased nucleosome occupancy; pseudogene displays a minimal bendability change at both TSSs and TESs. These results validate the previous finding of the association between rigid DNA and nucleosome depletion at TSSs, and demonstrate the same relationship at TESs by analyzing distinct bendability patterns of four major gene types.

Effect of DNA bendability in TSS and TES regions

We next sought to explore the effect of DNA bendability on chromatin regulation at different gene promoters in GM12878. We divided all PCGs into three equal groups (rigid, intermediate and bendable) according to their average bendability at TSS –20 to +70 bp regions (Supplementary Figure 2D). Here 'rigid' represents DNA sequences with low bendability, 'bendable' presents DNA sequences

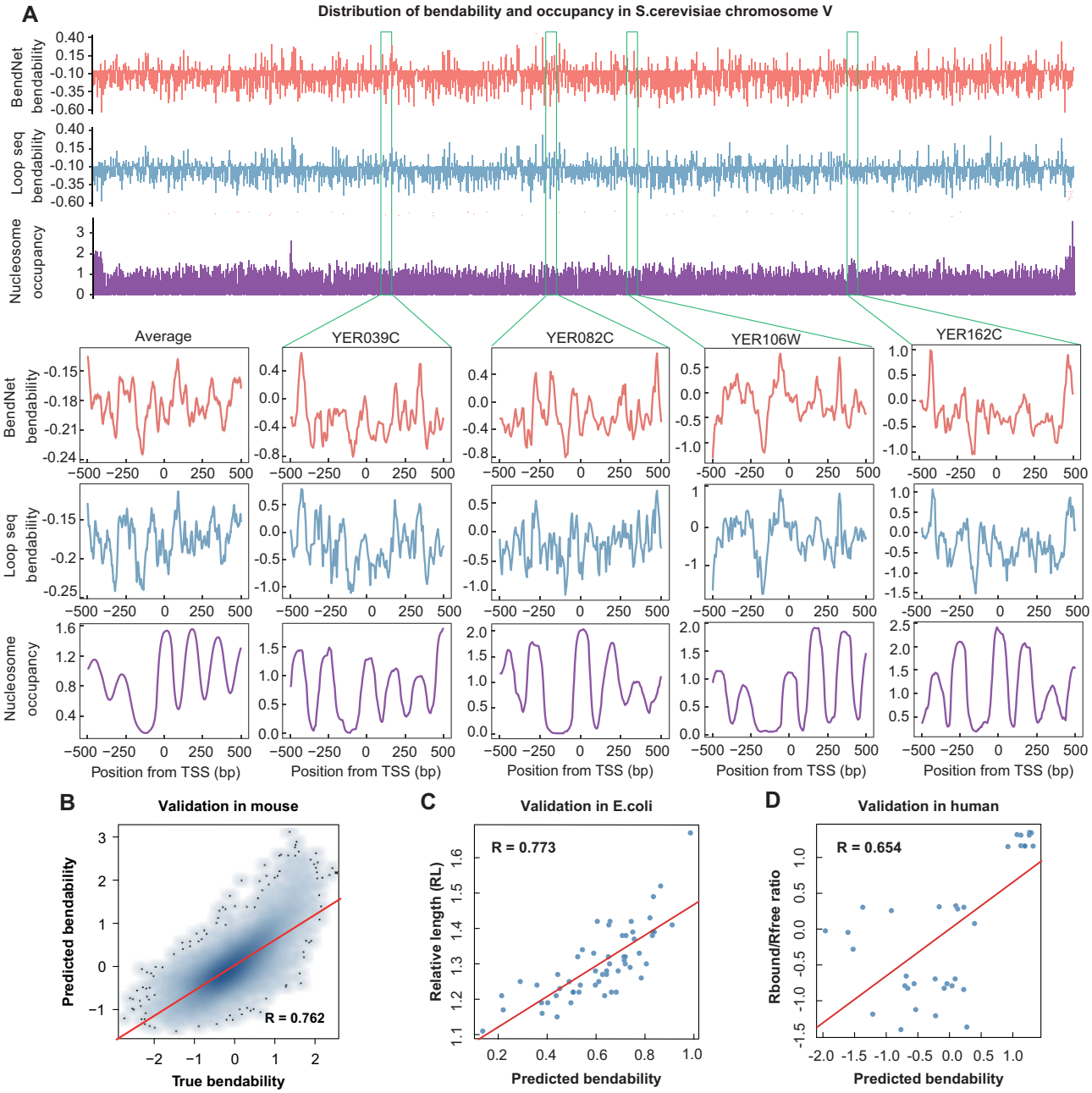


Figure 2. Evaluation of BendNet results in independent datasets. (**A**) Top panel: Alignment of DNA bendability by loop-seq, predicted bendability by BendNet, and nucleosome occupancy of *S.cerevisiae* chromosome V. Bottom panel: Average BendNet prediction, loop-seq measurement and nucleosome occupancy signal over all genes in *S.cerevisiae* chromosome V (left). BendNet prediction, loop-seq measurements and nucleosome occupancy signals of four example genes (YER039C, YER082C, YER106W and YER162C) (right). (**B**) Correlation between loop-seq bendability and predicted bendability in the mouse genome. (**C**) Correlation between relative lengths (RL) (a surrogate of DNA bendability) and predicted bendability in *E. coli.* (**D**) Correlation between Rbound/Rfree ratios (a surrogate of DNA bendability) and predicted bendability in human.

with high bendability and 'intermediate' presents DNA sequences with bendability in-between. For each group, a representative gene promoter is illustrated (Supplementary Figure 2E). As expected, DNA bendability is positively associated with nucleosome occupancy measured by MNase (Figure 3C). By analyzing ATAC-seq and DNase-seq, we found rigid TSSs (with low bendability) are more accessible (Supplementary Figure 2F, G) than bendable TSSs (with high bendability). Moreover, the rigid TSSs are found to be enriched with the H3K79me2 ChIP-seq and RNA polymerase II (Pol II) signals, which are associated with transcriptional initiation and elongation, respectively (Supplementary Figure 2H, I). Consistent with this observation, the histone modifications of transcriptional activation such as H3K4me3 and H3K27ac are enriched in rigid TSSs (Supplementary Figure 2J, K), and DNA methylation is relatively low in rigid TSSs (Supplementary Figure 2L). These observations are consistent across different

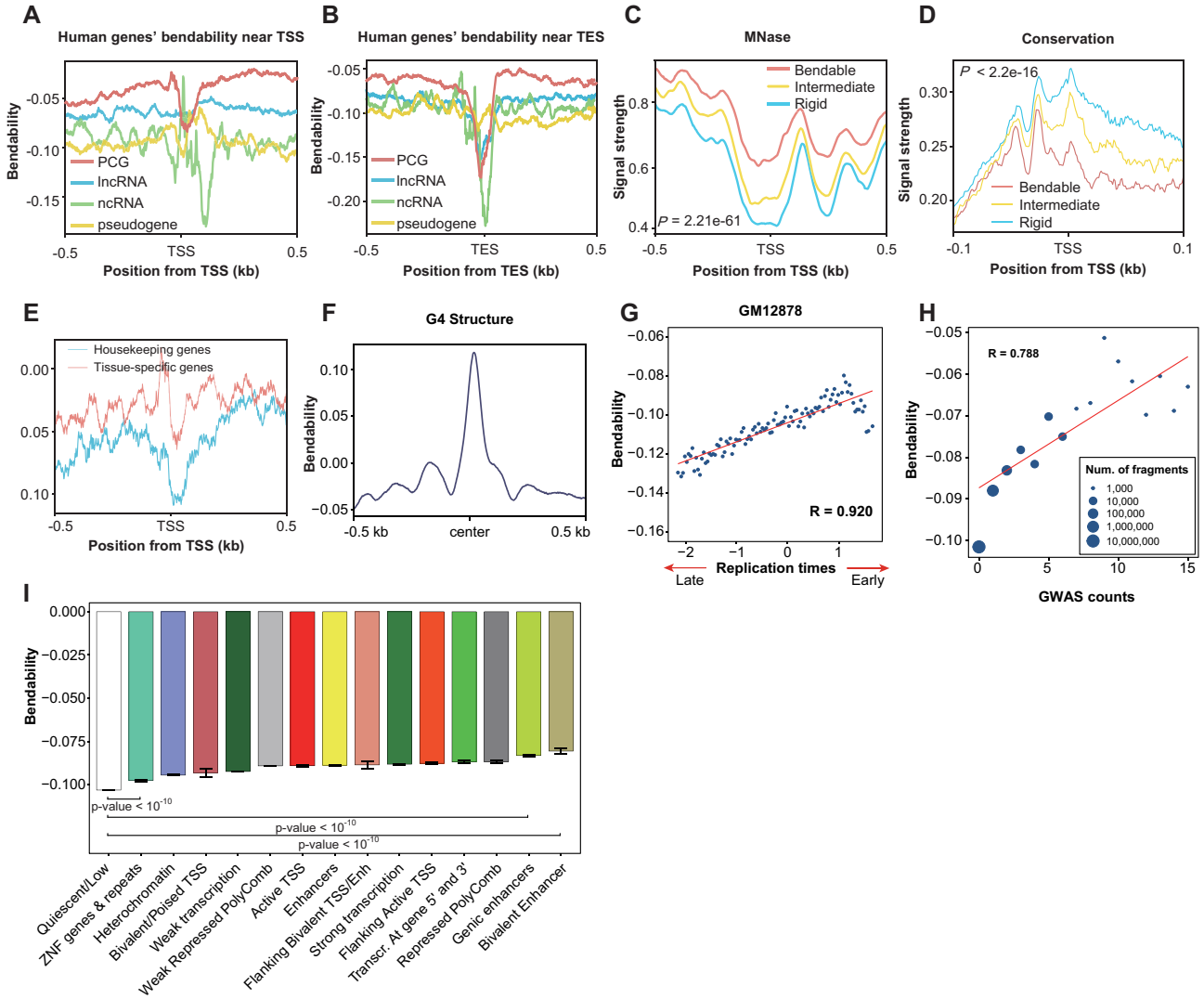


Figure 3. Association between DNA bendability and functional elements in human genome. (A, B) The average bendability of protein-coding genes (PCGs), lncRNAs, ncRNAs and pseudogenes in TSS (**A**) and TES (**B**) regions. TSS, transcription start site. TES, transcription end site. (C, D) Signals of MNase (**C**) and conservation (**D**) in TSSs of three PCG groups. *P*-values are calculated by the *Wilcoxon test* between signals in the rigid group and bendable group. (**E**) The average bendability of housekeeping and tissue-specific genes around TSS. (**F**) Bendability near G4 structures. (**G**) Correlation between DNA bendability and replication timing in GM12878. (**H**) Correlation between average bendability and GWAS SNP counts. (**I**) The average bendability of 15 chromatin states in GM12878 cell line, including Heterochromatin, ZNF Genes & Repeats, Quiescent/Low, Bivalent/Poised TSS, Weak Transcription, Active TSS, Flanking Bivalent TSS/Enh, Strong Transcription, Enhancers, Flanking Active TSS, Transcr. At Gene 5 and 3, Weak Repressed PolyComb, Repressed PolyComb, Genic Enhancers and Bivalent Enhancers. *P*-values are calculated by the Wilcoxon test.

cell types (Supplementary Figure 3A-T) indicating an intrinsic role of DNA mechanics in transcriptional control through epigenetic regulations. In addition, we observed that rigid TSSs are more conserved across species than bendable TSSs (Figure 3D) and are enriched for the housekeeping genes (Figure 3E), which suggests that rigid TSSs are functionally essential during both evolution and cellular development.

It is worth noting that DNA bendability at TESs is not associated with that at the same TSSs (Supplementary Figure 4A). Therefore, we performed the above analysis on TESs without considering their bendability patterns at corresponding TSSs (Supplementary Figure 4B). As expected, low nucleosome occupancy and DNA methylation are associated with rigid DNA at TESs (Supplementary Figure 4C, D). However, chromatin accessibility and histone activation marks are associated with bendable DNA at TESs which is opposite to the observation at TSSs (Supplementary Figure 4E–J), and its variation in bendability is not due to the difference of the sequence conservation (Supplementary Figure 4K). The greater reduction in bendability at TES compared to that at TSS may be due to the AT-rich sequences of polyAs (Supplementary Figure 4L). These results may suggest a different effect of DNA mechanics on epigenetic regulation at TESs.

Effect of DNA bendability in G4 structure

G4 (G-quadruplex) structure is a DNA secondary structure frequently found in the human genome, and plays an impor-

tant role in genomic instability (37). To investigate whether G4 structures are associated with DNA bendability, we analyzed the DNA bendability of G4 sequences in the human genome. We found an enrichment of bendable DNA at the center of G4 sequences (Figure 3F). G4 sequences are considered to form knot-like folding structures through Hoogsteen hydrogen bonding of four guanines (38) both *in vivo* and *in vitro*, and as a result are more bendable than their surrounding regions.

Effect of DNA bendability on replication timing

Eukaryotic genomes undergo replication in a specific order and timing. Various origins along the chromosomes start to replicate at specific times during cell division. This process is known as the replication timing program. However, it is still unclear to what extent replication initiation sites are determined by local sequences (39). To investigate whether DNA bendability plays a role in replication origins, we analyzed DNA bendability at various replication timing regions on a genome-wide scale. Our analysis showed that the DNA bendability is highly correlated with replication timing across multiple cell lines (Figure 3G, Supplementary Figure 4M, N). These results suggest a potential role of intrinsic DNA properties, such as DNA bendability, in regulating the replication timing program.

DNA bendability in disease associated regulatory regions

Large scale epigenomic study reveals that regulatory regions are enriched in disease-associated traits (40). Therefore, we investigated if the bendability of a DNA fragment (200 bp) is associated with the frequency of disease-associated variants. We found that DNA bendability is positively associated with the frequency of the single nucleotide polymorphisms (SNPs) in the genome-wide association studies (GWAS) catalog (21) ($R = 0.788$, Figure 3H). This finding still holds after correcting for the total number of SNPs in the relevant regions ($R = 0.666$, Supplementary Figure 4O) or taking significant levels of GWAS SNPs into consideration ($R = 0.731$, Supplementary Figure 4P). There analyses suggest DNA intrinsic property may play an important role in establishing functional genomic regions in human. To gain further insights into the effect of DNA bendability on broad regulatory regions, we calculated the average bendability of 15 chromatin states defined by chromHMM in GM12878 (20). Functional regulatory regions such as enhancer regions are generally more bendable than heterochromatin and Quiescent/Low regions (Figure 3I). The same pattern was also observed in three other cell lines (Supplementary Figure 4Q).

Effect of DNA bendability in enhancer regions

Enhancers are key regulatory elements that play an important role in tissue development and diseases. Enhancers demonstrate a clear decrease of DNA bendability at the central region (Figure 4A). Interestingly, we observed a stronger correlation between DNA bendability and GC content in enhancers (Figure 4B, $R = 0.431$) than in random genomic regions (Figure 4C, $R = 0.309$). We performed motif analysis on enhancer sequences with the highest (top 0.5%) and lowest (bottom 0.5%) bendability by using HOMER (41). Rigid enhancers are enriched with AT-rich transcription factor (TF) binding motifs, such as REM19, RLR1 and Fra1 (Figure 4D); while bendable enhancers are enriched with GC-rich TF binding motifs, such as ZNF341, Etv2 and ERG (Figure 4E). The above results prompt an intrinsic role of GC content in determining DNA bendability in regulatory regions. Therefore, we extended our analysis to 15 chromatin states defined by ChromHMM in GM12878 (42), and observed the highest correlation ($R = 0.351$) between GC content and bendability in Flanking Bivalent TSS/Enh, and lowest correlation ($R = 0.231$) in Bivalent Enhancer (Figure 4F). These findings are consistent across cell lines (Supplementary Figure 4R).

Effect of DNA bendability on TF binding

Next, we sought to investigate the effect of DNA bendability on transcription factor (TF) binding. By analyzing the ChIP-seq data of 152 TFs in GM12878 from ENCODE (43), we found distinct bendability patterns in TF binding sites, quantified by the Pearson correlation between the average profile of bendability and the ChIP-seq signal at TF binding regions (Figure 5A, top). At the TF binding sites compared to the surrounding regions, over two thirds of TFs such as CREB1 exhibit depressed bendability, about one third of TFs such as CTCF display elevated bendability, and some TFs in between such as STAT5A show no clear bendability pattern (Figure 5B, Supplementary Figure 5A). We also tried another measurement of TF binding strength by measuring the bendability height in the TF binding regions, and obtained similar results (Figure 5A bottom).

We then performed GO enrichment analysis (44,45) and protein-protein interaction (PPI) (46) network analysis on TFs with rigid ($R \leq -0.3$ in Figure 5A) and bendable ($R > 0.3$ in Figure 5A) binding sites. We found that TFs with rigid binding sites are enriched in transcription regulator complex, nuclear chromosome and transferase complex (Supplementary Figure 5B), while TFs with bendable binding sites are enriched for cohesion complex, condensed chromosome, transcriptional repressor complex and SWI/SNF superfamily-type complexes (Supplementary Figure 5C).

To further explore the effect of bendability on the binding strength of individual TFs, we obtained binding peaks with the top and bottom quarters of ChIP-seq signals, and investigated the bendability difference between them for each of the 152 TFs. We found that strong binding peaks exhibited lower bendability than weak peaks for TFs with overall rigid binding sites such as CREB1 and ZNF24, and reversed patterns for TFs with overall bendable binding sites such as SMC3 and CTCF (Figure 5C). These findings indicate the complex role of DNA mechanics in determining the TF binding strength.

Next, we sought to investigate the functional role of DNA bendability for TFs with overall rigid binding sites, which are mostly located in the promoter regions (Figure 5A). Most TFs regulate gene expression through binding and co-binding to *cis*-regulatory elements. It is well accepted that

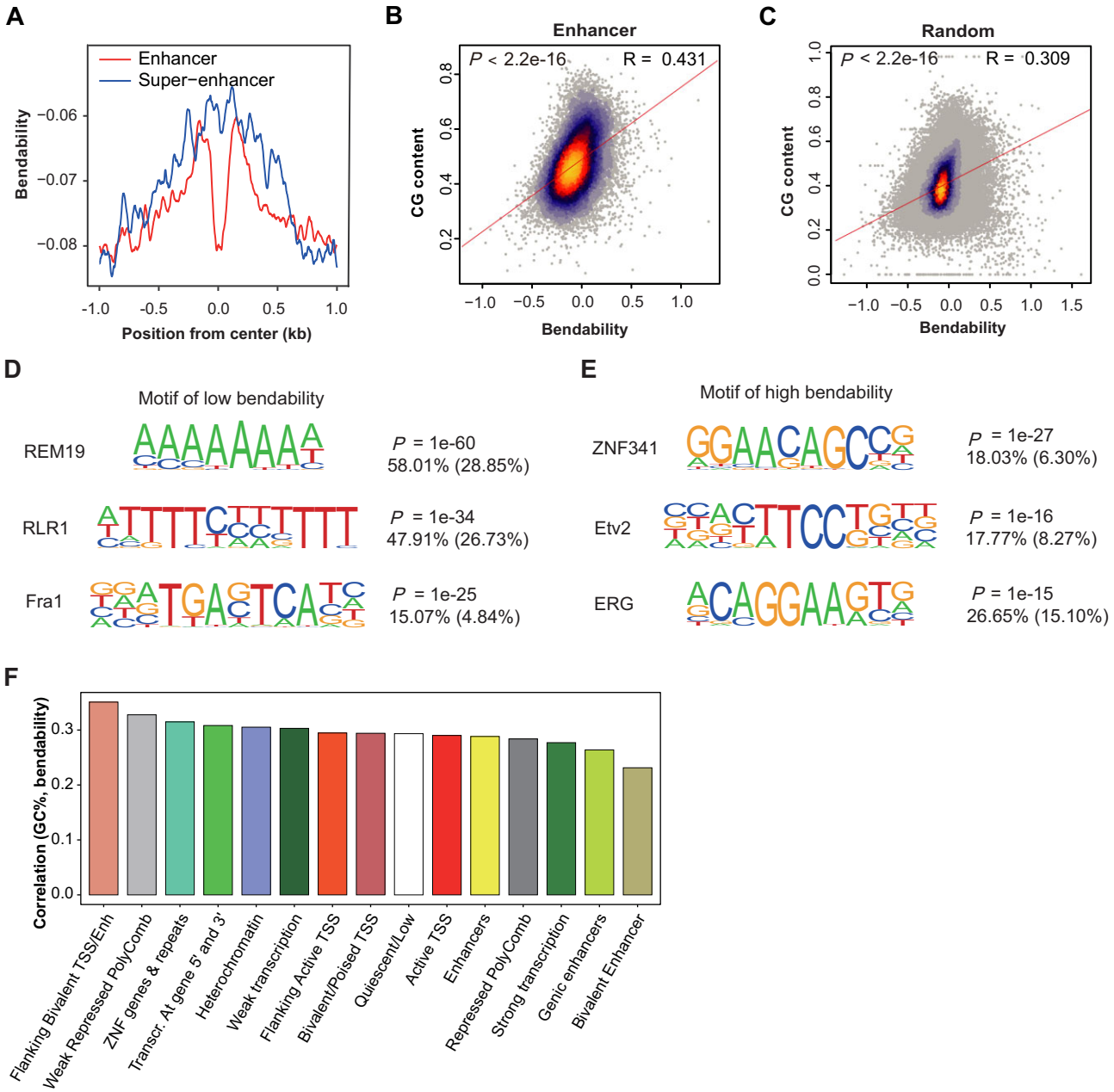


Figure 4. DNA bendability in enhancer regions and its association with GC content. (**A**) The average bendability of enhancer and super-enhancer regions. (B, C) Associations between GC content and average bendability of enhancer (**B**) and random sequences (**C**) in the human genome. (D, E) Top three motifs of rigid (**D**) and bendable enhancers (**E**) by using HOMER. (**F**) Correlation between GC content and bendability of 15 chromatin states in GM12878.

sequence motif is critical but not exclusive for TF binding. What other DNA properties contribute to the TF binding remains elusive. Here, we showed that the TF binding and co-binding are influenced by the intrinsic bendability of DNA. Specifically, we observed more binding and co-binding sites for all 152 tested TFs at rigid TSSs than bendable TSSs (Supplementary Figure 6A, B). These findings suggest the critical role of DNA bendability in both TF binding and co-binding, and may serve as an intrinsic regulator of gene transcription other than sequence motifs.

Competition between nucleosome and TFs at bendable DNAs

We then sought to investigate the functional implication of DNA bendability for TFs with overall bendable binding sites (Figure 5A and Supplementary Figure 6C–K). These TFs are mostly involved in 3D genome folding and co-bind with CTCF, including CTCF itself, cohesin subunits RAD21 and SMC3, and regulators of promoter-enhancer loops—YY1 and ZNF143. Further analyses demonstrate CTCF binding sites (CBSs) exhibit the strongest bendability peaks (Figure 5D, Supplementary Figure 6L–N), while bendability peaks of the other four factors are due to their

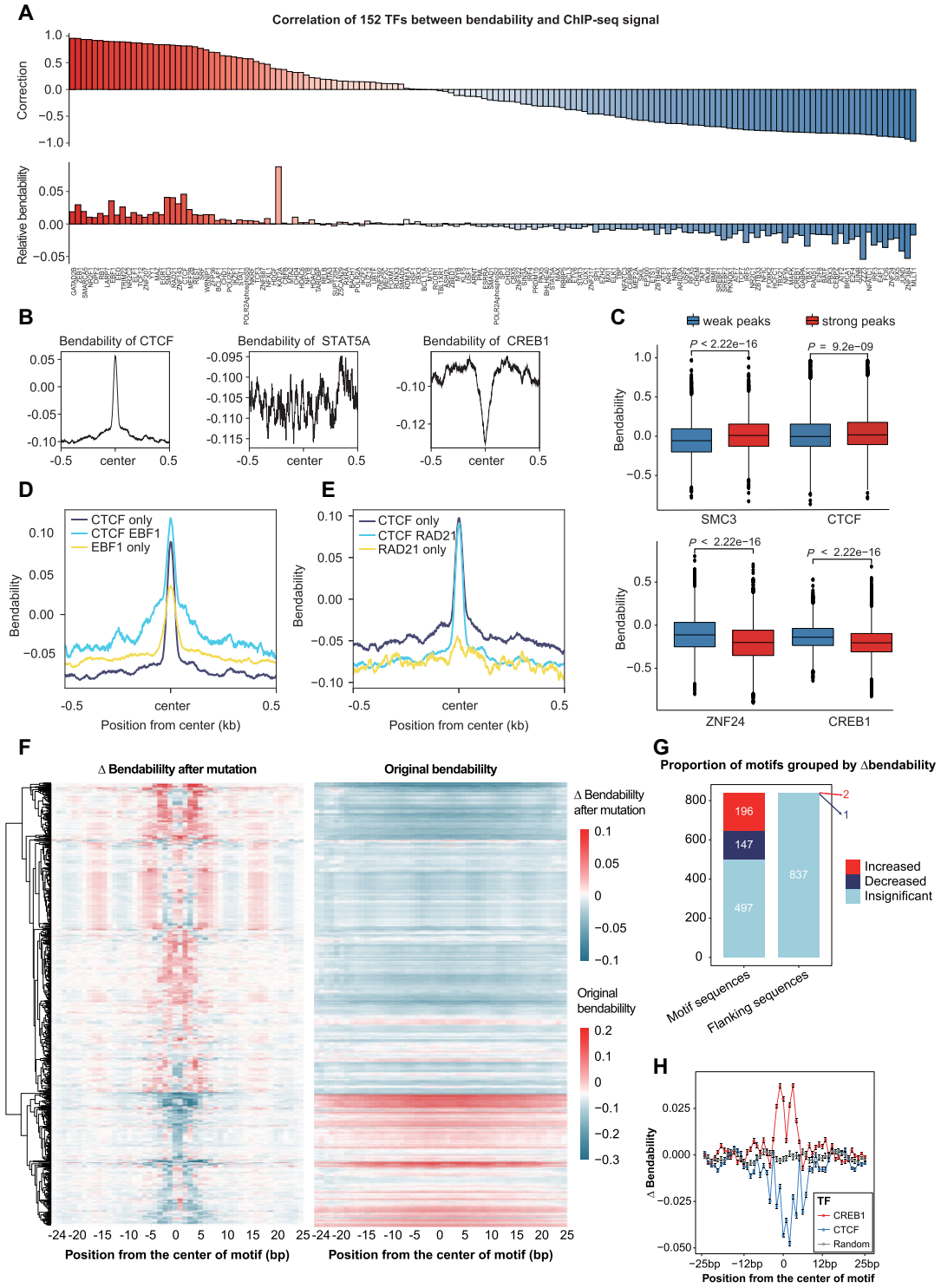


Figure 5. Sequence effect of DNA bendability on TF binding. (**A**) Correlations between bendability and ChIP-seq signals of 152 TFs in GM12878 cell line (top). Relative bendability is calculated as the difference between average bendability in the peak central ±100 bp region and average bendability in the surrounding regions (–500 bp to –100 bp and +100 bp to +500 bp) (bottom). KAT2A has only 11 ChIP-seq peaks in the GM12878 cell line, resulting in significant data fluctuations. (**B**) The average bendability of three representative TFs (CTCF, STAT5A and CREB1) showing positive, weak and negative correlations to their ChIP-seq signals at the binding sites. (**C**) The bendability of weak and strong peaks in four TFs (SMC3, CTCF, ZNF24 and CREB1). Weak and strong peaks refer to ChIP-seq peaks with 25% lowest and 25% highest ChIP-seq signals, respectively. The lower and upper hinges of the boxes represent the first and third quantiles, the whiskers extend 1.5 times the interquartile range from the hinges and the line represents the median. (D, E) Average bendability in binding sites of CTCF (dark blue), EBF1 (**D**) and RAD21 (**E**) (yellow), and their co-binding sites (sky blue). (**F**) Bendability changes after in silico mutagenesis of 840 TF motifs (left) and their original values (right). (**G**) Three groups of motifs, indicated increased bendability change (Δbendability > 0.01 & *P*-value < 0.001), decreased bendability change (Δbendability < -0.01 & *P*-value < 0.001) and no significant change. (**H**) Bendability changes (predicted disruption) after in silico mutagenesis of CTCF/CREB1 motifs and random sequences.

overlapping with CBSs (Figure 5E, Supplementary Figure 6O–Q), which underlines the central role of CTCF in regulating 3D genome organization through DNA bendability. Consistent with this observation, there is a well-defined NDR region at the CBS with up- and down-stream nucleosomes aligned accordingly, but much weaker patterns for the binding sites of the other factors (Supplementary Figure 7A–D). Moreover, we observed that CBSs without CTCF binding in specific cell lines are occupied by nucleosomes. More specifically, we investigated the MNase-seq signal in CBSs with or without actual CTCF binding in GM12878 cells. The latter exhibits a clear nucleosome occupancy signal, while the former does not (Supplementary Figure 7E–G). This mutually exclusive pattern is also observed in the other 3 tested cell lines (Supplementary Figure 7H-J), indicating the role of CTCF to compete with nucleosome proteins in binding to the specific bendable regions. This phenomenon suggests a mechanism that the CTCF and its bendable binding elements cooperate to form the anchor points in the loop formation process, because the sharply defined bendable regions in the CBSs can facilitate attracting and capturing CTCF to form accurate and stable boundaries.

Except for the above genome folding factors, four other TFs, i.e. EBF1, EGR1, MAZ and MEF2B also show overall bendable binding sites (Supplementary Figure 6D, E, H, I), which are not due to the overlap with CBSs (Figure 5D, Supplementary Figure 6L–N). By exploring the MNase-seq data, we found CREB1, MAZ and EGR1 (Supplementary Figure 7K–M), similar to CTCF, show a well-defined NDR at their binding sites, while EBF1 (Supplementary Figure 7N) and MEF2B (Supplementary Figure 7O) binding sites exhibit a near random nucleosome-occupied pattern suggesting their bindings do not reply on nucleosome depletion. Both genes are B cell-specific regulators, and EBF1 is a pioneer factor that can directly bind DNA with nucleosomes, therefore demonstrating an association between bendable DNA and nucleosome independent TF binding. These findings indicate that genome folding factors and some cell-specific regulators bind to bendable regions that have to compete with nucleosome proteins, and these factors are functionally distinct compared to the typical TFs with rigid DNA binding sites leading to intrinsically block of nucleosome occupancy.

BendNet predicts nucleotide-level DNA bending features of TF motifs

To screen for single nucleotide variations at TF binding sites that lead to local DNA bendability remodeling, we performed in silico saturation mutagenesis of 50 bp regions centered at 840 TF motifs and random sequences, and quantified the predicted DNA bendability disruptions for mutations to individual nucleotides. Predicted disruptions of motifs show a negative correlation with their original bendability. After mutagenesis, the bendability of the motifs tends to regress to the genome background (Figure 5F). In particular, predicted disruptions are larger for nucleotides around the motifs than in the flanking regions, with specific vital nucleotides for different motifs (Figure 5F, G). For instance, predicted disruptions of the CTCF motifs show a strong bendability decrease at the motif centers, specifically at positions -1 and + 1, which contrasts with the elevation of bendability at the original motif sequences (Figure 5H). Predicted disruptions of the CREB1 motifs, on the other hand, exhibit a strong bendability increase at the motif centers, specifically at positions –1 and + 3, which contrasts with the dip of bendability at the original motif sequences (Figure 5H). These analyses indicate that BendNet predicts base-resolution DNA bending features of TF motifs, and that key motif nucleotides impacting DNA bending remain uncharacterized.

DNA bendability of human eccDNA

Extrachromosomal circular DNA (eccDNA) is a type of double-stranded loop-shaped DNA derived from genomic DNA. It has been observed in diverse cell types across different species (17), but its biogenesis is largely unknown. Here, we collected 423 018 human eccDNA fragments with lengths of <1000 bp and observed their sequences are significantly bendable compared to the genomic background (Figure 6A). There is no clear bendability difference among eccDNAs generated from different genomic regions (Figure 6A). These findings suggest that DNA mechanical properties may contribute to the generation of eccDNA, which is independent of sequence locations.

Resource of DNA bendability in 307 species

Taking advantage of BendNet in its potential for cross-species prediction, we computed base-resolution genomic bendability profiles of 307 species (Supplementary Table 1). To investigate the relationship between DNA bendability and species evolution, we calculated the average bendability of each species in four classes of vertebrates (Actinopteri: $n = 83$, Aves: $n = 48$, Lepidosauria: $n = 10$, Mammalia: $n = 120$). Aves (birds) demonstrates the highest DNA bendability (–0.0726), Actinopteri (fish) the lowest (–0.0972) and Mammalia in between (–0.0837) (Figure 6B). In model species, *Mus musculus* exhibits the highest average bendability of –0.0776, while *Caenorhabditis elegans* displays the lowest average bendability of –0.1995 (Figure 6C).

Webserver of BendNet

To facilitate the easy access of the large resource of DNA bendability in model species, we developed a webserver (http://www.dnabendnet.com/) (Figure 6D). Users can query one or multiple genomic region(s) in the database to obtain the average profile and heatmap of individual regions of DNA bendability (Figure 6E). This function can assist users to obtain the DNA bendability pattern of the genomic regions of interest obtained from ChIP-seq, ATAC-seq, etc. Users can also query one or multiple DNA sequences to obtain their bendability in newly sequenced species or sequence variations, such as SNPs and somatic mutations (Figure 6F). Furthermore, our website provides predicted bendability disruptions resulting from in silico mutagenesis on 840 TF motifs. Users can query individual motif IDs or gene symbols to access this information.

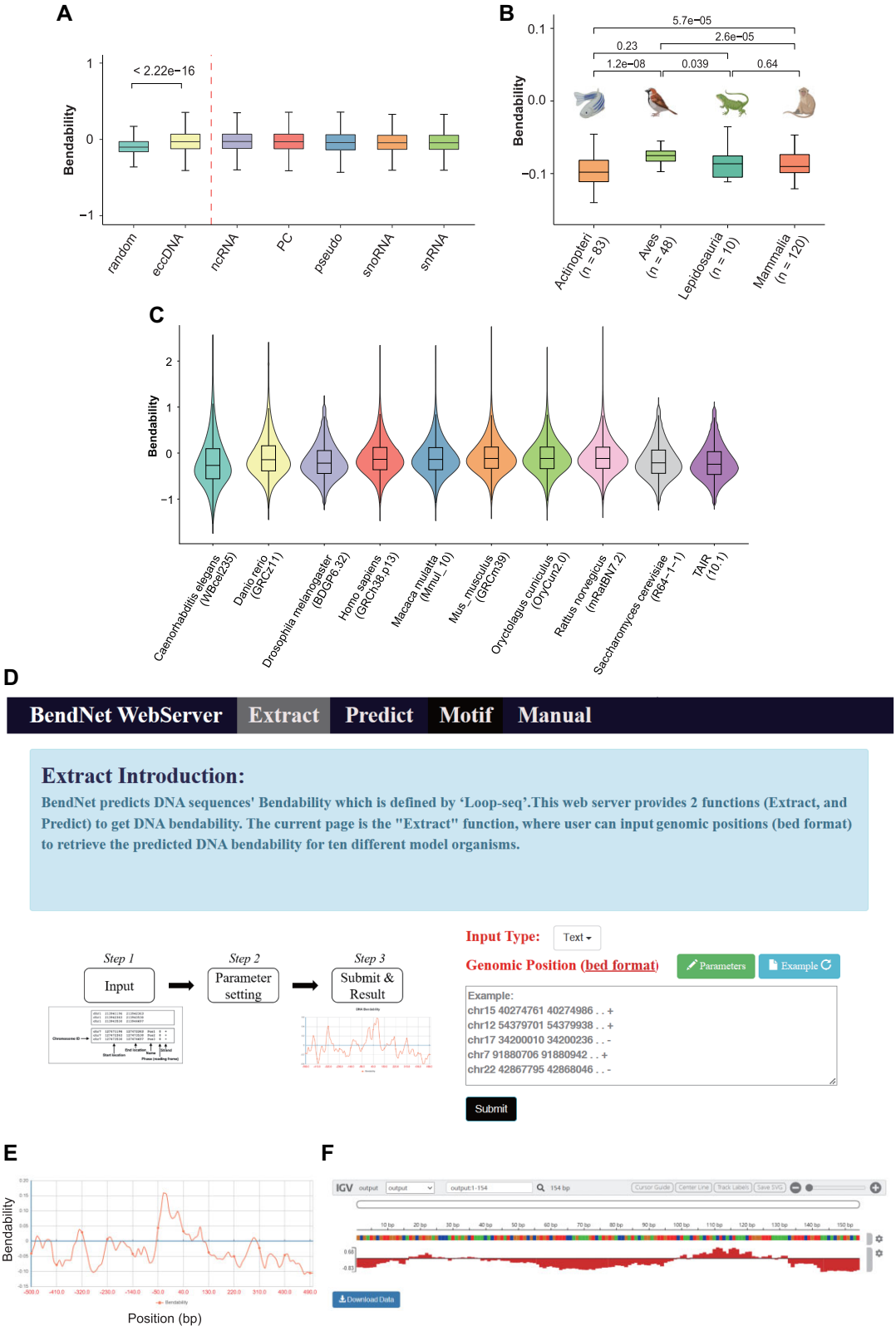


Figure 6. DNA bendability distributions of eccDNA, whole genomes of 307 species and online webserver. (**A**) The bendability of random genomic segments, eccDNAs and five types of eccDNA with different origins. (**B**) The average bendability of four classes of vertebrates (Actinopteri: $n = 83$, Aves: $n = 48$, Lepidosauria: $n = 10$, Mammalia: $n = 120$) in the genome scale. (**C**) Genome-wide DNA bendability of 10 species. (**D**) The interface of the BendNet webserver. Users could either input genomic positions of existing species in bed/bed format, or any new sequences in fasta format or raw read. (E, F) Outputs of BendNet server when inputting genomic regions in bed format (**E**) or raw DNA sequences (**F**).

DISCUSSION

Decoding DNA mechanics and its effect on chromatin regulation is one of the fundamental questions in genomics. Our work provides a tool to assess bendability, one of the mechanical properties of DNA, for massive-scale DNA sequences to study their biological consequences. Our data provide a base-resolution map of DNA bendability predictions in human, which expands our knowledge of how DNA mechanics influence chromatin regulation through DNA-macromolecular interactions. We also provide genome scale bendability predictions for 307 species as a comprehensive resource, along with a webserver to query some of the model species. By using these, researchers could study the variation of DNA bendability during evolution and its effect on gene regulation and function. The webserver could be used to investigate the effect of DNA bendability on diverse biological systems, such as nucleosome assembly during DNA replication, genome stability maintenance, epigenetic inheritance, DNA damage repair, and V(D)J recombination in B- or T-cell development. Furthermore, our framework can be easily applied to assess other DNA mechanics, such as DNA twisting, supercoiling and torsional rigidity, when enough experimental measurement data are available, and a multi-task learning strategy could further improve the prediction accuracy. The current BendNet model is built upon *in vitro* data that lacks conditional information, such as cell type and tissue specificity. However, developing experimental strategies to measure *in vivo* DNA mechanics could address this limitation in the future.

DATA AVAILABILITY

ChIP-seq of 152 TFs, ATAC-seq, DNase-seq, MNase-seq, H3K4me3, H3K27ac, H3K79me2 and POLR2A in GM12878 were obtained from Encode website (https://www.encodeproject.org/). The bendability resource for 307 species at a resolution of 10 can be found at (https://zenodo.org/deposit/7699690 and https://zenodo.org/deposit/7663786).

CODE AVAILABILITY

BendNet is freely available at (https://github.com/JiangWenJie-stack/DNABendNet) and (https://zenodo.org/record/8218189). The webserver is at http://www.dnabendnet.com/.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We gratefully acknowledge the High-performance Computing Platform of Peking University for conducting the model training and prediction.

FUNDING

Fundamental Research Funds for the Central Universities [PKU2022LCXQ027 – Clinical Medicine Plus X – Young Scholars Project, Peking University and BMU2021YJ064 to H.J.W.]; National Natural Science Foundation of China [32270683 to H.J.W., 61572327 and 61972257 to X.Z.]; Natural Science Foundation of Shanghai [20JC1413800 to X.Z.]; National Key R&D Program of China [2021YFC1712805 to H.J.W., 2018YFA0900600 to X.Z.]. Funding for open access charge: Fundamental Research Funds for the Central Universities [PKU2022LCXQ027 – Clinical Medicine Plus X – Young Scholars Project, Peking University and BMU2021YJ064 to H.J.W.]; National Natural Science Foundation of China [32270683 to H.J.W., 61572327 and 61972257 to X.Z.]; Natural Science Foundation of Shanghai [20JC1413800 to X.Z.]; National Key R&D Program of China [2021YFC1712805 to H.J.W., 2018YFA0900600 to X.Z.].

Conflict of interest statement. None declared.

REFERENCES

1. Dalrymple,B.P., Kongsuwan,K., Wijffels,G., Dixon,N.E. and Jennings,P.A. (2001) A universal protein-protein interaction motif in the eubacterial DNA replication and repair systems. *Proc. Natl. Acad. Sci. U.S.A.*, **98**, 11627–11632.
2. Korolev,N., Vorontsova,O.V. and Nordenskiöld,L. (2007) Physicochemical analysis of electrostatic foundation for DNA–protein interactions in chromatin transformations. *Prog. Biophys. Mol. Biol.*, **95**, 23–49.
3. Wang,D., Yamamoto,S., Hijiya,N., Benveniste,E.N. and Gladson,C.L. (2000) Transcriptional regulation of the human osteopontin promoter: functional analysis and DNA-protein interactions. *Oncogene*, **19**, 5801–5809.
4. Vafabakhsh,R. and Ha,T. (2012) Extreme bendability of DNA less than 100 base pairs long revealed by single-molecule cyclization. *Science*, **337**, 1097–1101.
5. Calladine,C., Collis,C.M., Drew,H.R. and Mott,M.R. (1991) A study of electrophoretic mobility of DNA in agarose and polyacrylamide gels. *J. Mol. Biol.*, **221**, 981–1005.
6. Roy,R., Hohng,S. and Ha,T. (2008) A practical guide to single-molecule FRET. *Nat. Methods*, **5**, 507–516.
7. Basu,A., Bobrovnikov,D.G., Qureshi,Z., Kayikcioglu,T., Ngo,T., Ranjan,A., Eustermann,S., Cieza,B., Morgan,M.T. and Hejna,M. (2021) Measuring DNA mechanics on the genome scale. *Nature*, **589**, 462–467.
8. Tang,L. (2021) Sequencing DNA bendability. *Nat. Methods*, **18**, 121–121.
9. Chai,C., Xie,Z. and Grotewold,E. (2011) *Methods Mol. Biol.*, Springer, pp. 249–258.
10. Sabour,S., Frosst,N. and Hinton,G.E. (2017) Dynamic routing between capsules. *Adv. Neural Inf. Process.*, **30**, 3859–3869.
11. Byerly,A., Kalganova,T. and Dear,I. (2021) No routing needed between capsules. *Neurocomputing*, **463**, 545–553.
12. Wang,Q., Albert,F.G., Fitzgerald,D.J., Calvo,J.M. and Anderson,J.N. (1994) Sequence determinants of DNA bending in the ilvlH promoter and regulatory region of Escherichia coli. *Nucleic Acids Res.*, **22**, 5753–5860.
13. Basu,A., Bobrovnikov,D.G., Cieza,B., Arcon,J.P., Qureshi,Z., Orozco,M. and Ha,T. (2022) Deciphering the mechanical code of the genome and epigenome. *Nat. Struct. Mol. Biol.*, **29**, 1178–1187.
14. Chang,C.W., Cheng,W.C., Chen,C.R., Shu,W.Y., Tsai,M.L., Huang,C.L. and Hsu,I.C. (2011) Identification of human housekeeping genes and tissue-selective genes by microarray meta-analysis. *PLoS One*, **6**, e22859.
15. Lizio,M., Harshbarger,J., Shimoji,H., Severin,J., Kasukawa,T., Sahin,S., Abugessaisa,I., Fukuda,S., Hori,F., Ishikawa-Kato,S. *et al.* (2015) Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol.*, **16**, 22.
16. Quinlan,A.R. (2014) BEDTools: the Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinformatics*, **47**, 11.12.1–11.12.34.

17. Zhao,X., Shi,L., Ruan,S., Bi,W., Chen,Y., Chen,L., Liu,Y., Li,M., Qiao,J. and Mao,F. (2022) CircleBase: an integrated resource and analysis platform for human eccDNAs. *Nucleic Acids Res.*, **50**, D72–D82.
18. Massey,D.J., Kim,D., Brooks,K.E., Smolka,M.B. and Koren,A. (2019) Next-Generation Sequencing Enables Spatiotemporal Resolution of Human Centromere Replication Timing. *Genes (Basel)*, **10**, 269.
19. Chambers,V.S., Marsico,G., Boutell,J.M., Di Antonio,M., Smith,G.P. and Balasubramanian,S. (2015) High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat. Biotechnol.*, **33**, 877–881.
20. Ernst,J. and Kellis,M. (2017) Chromatin-state discovery and genome annotation with ChromHMM. *Nat. Protoc.*, **12**, 2478–2492.
21. Buniello,A., MacArthur,J.A.L., Cerezo,M., Harris,L.W., Hayhurst,J., Malangone,C., McMahon,A., Morales,J., Mountjoy,E., Sollis,E. *et al.* (2019) The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.*, **47**, D1005–D1012.
22. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
23. Ramirez,F., Dundar,F., Diehl,S., Gruning,B.A. and Manke,T. (2014) deepTools: a flexible platform for exploring deep-sequencing data. *Nucleic Acids Res.*, **42**, W187–W191.
24. Byerly,A. and Kalganova,T. (2021) Homogeneous vector capsules enable adaptive gradient descent in convolutional neural networks. *IEEE Access*, **9**, 48519–48530.
25. Kwak,S.K. and Kim,J.H. (2017) Statistical data preparation: management of missing values and outliers. *Korean J. Anesthesiol.*, **70**, 407–411.
26. Mitchell,M. (1998) In: *An Introduction to Genetic Algorithms*. MIT Press.
27. Breiman,L. (2001) Random forests. *Mach. Learn.*, **45**, 5–32.
28. Awad,M. and Khanna,R. (2015) In: *Efficient Learning Machines*. Springer, pp. 67–80.
29. Krizhevsky,A., Sutskever,I. and Hinton,G.E. (2017) Imagenet classification with deep convolutional neural networks. *Commun. ACM*, **60**, 84–90.
30. He,K., Zhang,X., Ren,S. and Sun,J. (2016) In: *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*. pp. 770–778.
31. Szegedy,C., Liu,W., Jia,Y., Sermanet,P., Reed,S., Anguelov,D., Erhan,D., Vanhoucke,V. and Rabinovich,A. (2015) In: *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit*. pp. 1–9.
32. Simonyan,K. and Zisserman,A. (2014) Very deep convolutional networks for large-scale image recognition. arXiv doi: https://arxiv.org/abs/1409.1556, 10 April 2015, preprint: not peer reviewed.
33. Li,K., Carroll,M., Vafabakhsh,R., Wang,X.A. and Wang,J.P. (2022) DNAcycP: a deep learning tool for DNA cyclizability prediction. *Nucleic Acids Res.*, **50**, 3142–3154.
34. Khan,S.R., Sakib,S., Rahman,M.S. and Samee,M.A.H. (2023) DeepBend: an interpretable model of DNA bendability. *Iscience*, **26**, 105945.
35. MacPherson,M.J. and Sadowski,P.D. (2010) The CTCF insulator protein forms an unusual DNA structure. *BMC Mol. Biol.*, **11**, 101.
36. Nagaich,A.K., Appella,E. and Harrington,R.E. (1997) DNA bending is essential for the site-specific recognition of DNA response elements by the DNA binding domain of the tumor suppressor protein p53. *J. Biol. Chem.*, **272**, 14842–14849.
37. De,S. and Michor,F. (2011) DNA secondary structures and epigenetic determinants of cancer genome evolution. *Nat. Struct. Mol. Biol.*, **18**, 950–955.
38. Gellert,M., Lipsett,M.N. and Davies,D.R. (1962) Helix formation by guanylic acid. *Proc. Natl. Acad. Sci. U.S.A.*, **48**, 2013–2018.
39. Aladjem,M.I. and Redon,C.E. (2017) Order from clutter: selective interactions at mammalian replication origins. *Nat. Rev. Genet.*, **18**, 101–116.
40. Roadmap Epigenomics,C., Kundaje,A., Meuleman,W., Ernst,J., Bilenky,M., Yen,A., Heravi-Moussavi,A., Kheradpour,P., Zhang,Z., Wang,J. *et al.* (2015) Integrative analysis of 111 reference human epigenomes. *Nature*, **518**, 317–330.
41. Heinz,S., Benner,C., Spann,N., Bertolino,E., Lin,Y.C., Laslo,P., Cheng,J.X., Murre,C., Singh,H. and Glass,C.K. (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell*, **38**, 576–589.
42. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
43. Luo,Y., Hitz,B.C., Gabdank,I., Hilton,J.A., Kagda,M.S., Lam,B., Myers,Z., Sud,P., Jou,J., Lin,K. *et al.* (2020) New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res.*, **48**, D882–D889.
44. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S. and Eppig,J.T. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.*, **25**, 25–29.
45. Gene Ontology Consortium (2021) The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.*, **49**, D325–D334.
46. Szklarczyk,D., Gable,A.L., Lyon,D., Junge,A., Wyder,S., Huerta-Cepas,J., Simonovic,M., Doncheva,N.T., Morris,J.H., Bork,P. *et al.* (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res.*, **47**, D607–D613.