NAR Genomics and Bioinformatics, 2024, **6**, lqae123
https://doi.org/10.1093/nargab/lqae123
Advance access publication date: 24 September 2024
Pipelines and Workflows for Biological Data Analysis

OXFORD

GenomicLinks: deep learning predictions of 3D chromatin interactions in the maize genome

Luca Schlegel [1], Rohan Bhardwaj [1], Yadollah Shahryary [1], Defne Demirtürk [1], Alexandre P. Marand [2], Robert J. Schmitz [3] and Frank Johannes [1,*]

[1]TUM School of Life Sciences, Plant Epigenomics, Technical University of Munich, Freising, 85354, Germany
[2]Department of Molecular, Cellular, and Developmental Biology, University of Michigan, Ann Arbor, MI 48109, USA
[3]Department of Genetics, University of Georgia, Athens, GA 30602, USA

[*]To whom correspondence should be addressed. Tel: +49 8161 71 3621; Fax: +49 8161 71 5636; Email: frank@johanneslab.org

Abstract

Gene regulation in eukaryotes is partly shaped by the 3D organization of chromatin within the cell nucleus. Distal interactions between *cis*-regulatory elements and their target genes are widespread, and many causal loci underlying heritable agricultural traits have been mapped to distal non-coding elements. The biology underlying chromatin loop formation in plants is poorly understood. Dissecting the sequence features that mediate distal interactions is an important step toward identifying putative molecular mechanisms. Here, we trained GenomicLinks, a deep learning model, to identify DNA sequence features predictive of 3D chromatin interactions in maize. We found that the presence of binding motifs of specific transcription factor classes, especially bHLH, is predictive of chromatin interaction specificities. Using an *in silico* mutagenesis approach we show the removal of these motifs from loop anchors leads to reduced interaction probabilities. We were able to validate these predictions with single-cell co-accessibility data from different maize genotypes that harbor natural substitutions in these TF binding motifs. GenomicLinks is currently implemented as an open-source web tool, which should facilitate its wider use in the plant research community.

Introduction

The spatial configuration of chromatin within the nucleus of eukaryotic cells is fundamental to genome regulation. Its structural dynamics are critical for various cellular processes, including DNA replication, repair, spatiotemporal gene expression patterning and transposable element silencing (1–3). Chromosome conformation capture (3C), coupled with next-generation sequencing (e.g. Hi-C (4), Hi-Chromatin Immunoprecipitation (ChIP) (5), ChIA-PET (6), Capture-C (7), Capture Hi-C (8), 4C (9), 5C (10) or *in situ* Hi-C (11), have emerged as powerful tools to interrogate 3D chromatin biology in a high-throughput fashion. These methods have led to systematic insights into the hierarchical organization of chromatin, including the identification of A/B compartments (4), Topologically Associated Domains (TADs) (12) and chromatin loops in the form of enhancer-promoter interactions (EPIs) (5). In animals, TAD and chromatin loop formation is largely facilitated by the CTCF-cohesin complex, which binds DNA to physically clamp down chromatin and force loop extrusion (11). Although the general principles of chromatin organization are conserved, plants are distinct, in which they lack CTCF proteins. Perhaps, as a result, TAD structures display less defined boundaries and are often indistinguishable from local A/B compartments, particularly in large and complex plant genomes (13). How plants initiate and stabilize 3D chromatin interactions remains poorly understood. Several studies point to an enrichment of particular transcription factor (TF) families at loop anchors (14), which has led to the hypothesis that TFs mediate loop formation through processes like dimerization (15,16). Hence, the presence of specific TF binding motifs may be an important determinant of 3D chromatin interactions. Testing this hypothesis in a high-throughput manner is experimentally challenging. Moreover, additional DNA sequence features, beyond known TF binding motifs, may be important contributors to loop formation but remain difficult to identify. Machine learning (ML) methods trained on 3C data in animals have emerged as powerful tools to dissect the DNA sequence grammar underlying chromatin biology. Deep Learning (DL) models, in particular, have demonstrated remarkable success in predicting 3D chromatin contacts directly from DNA sequences (17), with prediction accuracies larger than 70%. The power of these models is that they can be combined with *in silico* mutagenesis, where arbitrary DNA sequence mutations are induced and evaluated for their impact on looping probabilities. This approach provides a framework to systematically screen genetic variants underlying agronomically important complex traits that have been mapped to (distal) *cis*-regulatory elements (18). Examples of such traits in maize include *tb1* (19), *ZmCCT9*, *Vgt1*, *prol1.1*, and *UPA2* (20,21). Such knowledge can generate concrete hypotheses for experimental validation using genome editing and/or present novel breeding targets. Harnessing DL models for studying plant chromatin biology is a promising goal. However, models trained on animal data perform poorly when applied to plants, yielding prediction accuracies as low as 0.1% (see SI material). This

Received: May 24, 2024. **Revised:** July 25, 2024. **Editorial Decision:** August 28, 2024. **Accepted:** August 30, 2024

observation points to fundamental differences in the type of DNA sequence features that predict chromatin looping, and suggests that DL models need to be trained directly on plant genomic data to optimize performance. Here we implement such an approach. Using maize (*Zea mays*) as an experimental system, we developed GenomicLinks, a DL model capable of predicting 3D chromatin interactions from DNA sequences with high accuracy. Leveraging high-quality Hi-ChIP-seq dataset for training, we integrated the most successful architectural features from DL models used in the animal field. The model combines dual convolutional neural networks (CNNs) with a long short-term memory (LSTM) network, enabling the identification of spatial and sequential aspects of chromatin interactions. The application of GenomicLinks, coupled with a TF-centered *in silico* mutagenesis approach, revealed binding motifs of specific TF classes as crucial predictors of chromatin looping. We were able to validate these predictions with single-cell co-accessibility data from nine different maize genotypes that harbor natural mutations in these TF binding motifs. GenomicLinks is currently implemented as an open-source web tool at genomiclinks.com, which should facilitate its wider use among plant researchers.

Material and methods

Hi-ChIP data collection

To construct a high-quality dataset of true chromatin interactions, we obtained Hi-ChIP data for the histone modifications H3K4me3 (associated with transcriptional activation) and H3K27me3 (associated with transcriptional repression) from Ricci et al. (2019) (18). These particular histone modifications were selected because they are the only ones published and are well-known markers of active promoters and enhancers (H3K4me3) and repressive chromatin (H3K27me3), respectively. These datasets were generated through an initial Hi-C process, followed by ChIP, using antibodies against H3K4me3 (indicative of transcriptional activity) and H3K27me3 (indicative of transcriptional suppression), while excluding regions characteristic of constitutive heterochromatin (22–24). The Hi-ChIP raw data were processed using the HiC-pro pipeline (version 2.8.054) (25) and mapped to *Zea mays* B73 reference genome (version 4) (26), where alignments with a MAPQ score greater than five were retained. ChIP-seq pulldown efficiency was assessed through the analysis of dangling-end and self-ligation read pairs. Loop detection was performed using FitHiChIP (27) with valid read pairs, applying a 5 kb bin size, adjusting for coverage bias and setting a false discovery rate threshold of less than 0.01 to distinguish H3K4me3 and H3K27me3 Hi-ChIP loops. The sequenced data were also filtered to include only interactions with a minimum genomic distance of 20 kb and a maximum genomic distance of 2 Mb.

Data preprocessing

To enhance accuracy and ensure our model's comparability with similar deep learning frameworks (17,28), we refined FitHiChIP's original 5 kb anchor definitions to 2.5 kb regions centered on each anchor's midpoint, to improve computational efficiency. Nucleotide sequences for these 2.5 kb windows were extracted from the *Zea mays* B73 reference genome (version 4) (26) for deep learning analysis. These sequences, encompassing 101,172 anchor pairs, constituted our set of true positive interactions and were labeled as '1'. To create the negative training set, we started with the positive set and permuted the anchor positions, ensuring that the negative set covered the same genomic space as the positive set for biological relevance. We kept a similar distance distribution of negative and positive samples. To improve the model's ability to distinguish between positive and negative sets, we further adjusted the positions of each permuted anchor pair by randomly shifting it 5–25 base pairs left or right and labeled them as '0'. We found that permuting anchor pairs without these additional bp shifts result in no learning progress. This suggests that the model learns sequence features that determine the potential of two genomic regions to interact rather than the 3D interaction architecture itself. Some alternative methods in the literature (29) opted to define negative sets by extracting random anchor sequences from the reference genome. In heterochromatin-rich genomes, like that of maize, this latter strategy runs into the danger that the models learn DNA features that distinguish euchromatin from heterochromatin, rather than interacting from non-interaction regions, and was therefore not employed here. Next, we combined our positive and negative training sets to form the 'complete training set' (N=202 344). This complete set was shuffled to ensure that the order of the anchor pairs does not bias the model during training. Finally, one-hot encoding was applied to the 2.5 kb anchor pairs in the complete training set to convert the A, T, C and G nucleotide sequence into a binary matrix representation. These binary matrices, along with their labels, were stored in the H5 file format, making it compatible with TensorFlow (30) and Keras (31) for data handling during training.

Model architecture and training process

We developed a dual CNN architecture to independently analyze each anchor, followed by an LSTM layer (see Figure 1). This configuration has proven successful in animal systems (32–36). Our model takes two 2.5kb input anchor sequences as input, representing potential genomic interaction sites. The architecture employs a one-dimensional CNN with two convolutional layers, featuring filter sizes of 256 and 512, and kernel sizes of 64 and 32. These layers extract high-level features from the input sequences (37), which are then processed by max-pooling layers for dimensionality reduction. Subsequently, the outputs were concatenated and fed into one bidirectional LSTM layer, comprising 256 units. This LSTM layer effectively captures long-range dependencies within the input sequences (38). The final prediction was obtained through a dense layer with 512 units and a sigmoid activation function. To enhance generalization and prevent overfitting, we incorporated batch normalization and dropout layers. The optimization process utilizes the Adam optimizer with a learning rate of 1e-4, guided by the binary cross-entropy loss function. The entire architecture was implemented using the TensorFlow (30) and Keras (31) libraries (jlab github). The model summary, including detailed parameter configurations, is provided in supplementary materials (Supplementary Figure S1). During the training process, we continually monitored various statistical metrics. One crucial metric is specificity at a given sensitivity threshold for the validation set. This metric measures the model's ability to correctly classify negative samples while maintaining a specified level of sensitivity. We utilized the EarlyStopping callback with patience of 10 epochs to ensure training halts when performance plateaus or deteriorates. ModelCheckpoint was employed to save the model weights

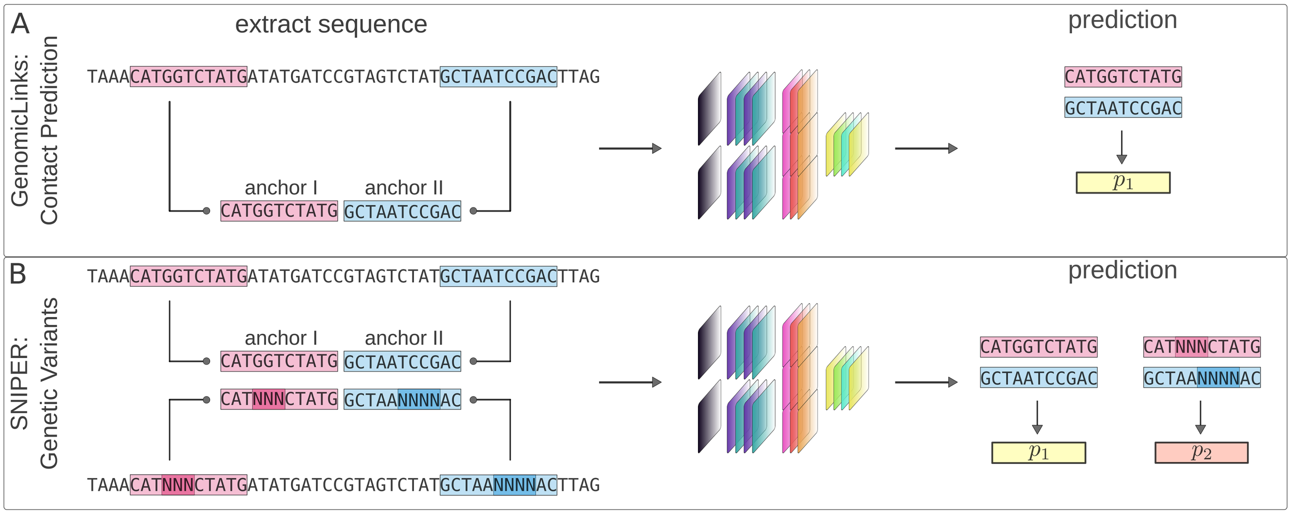


Figure 1. Workflow of GenomicLinks deep learning for predicting chromatin interactions in the maize genome. (**A**) GenomicLinks was trained on Hi-ChIP data to predict 3D chromatin interactions from DNA sequence features flanking chromatin interaction sites (anchors). The trained model accepts any 2.5 kb anchor pairs as input and outputs an interaction probability p_1. (**B**) SNIPER is an *in silico* mutagenesis software extension embedded in the GenomicLinks framework. SNIPER can introduce targeted sequence mutations in selected anchor pairs and resupplies these to GenomicLinks for evaluation. GenomicLinks uses the mutated anchor pairs to assess chromatin loop stability, yielding interaction probabilities p_1 for the original sequences and p_2 for the mutated ones. This approach provides a setup to test the impact of specific genetic variants on chromatin loop stability.

only if they outperform the previous best epoch. The training process spanned 100 epochs, with each epoch comprising a batch size of 100 samples. We used the Adam optimizer to minimize the binary cross-entropy loss, a common choice for binary classification tasks (39). Detailed training logs and access to saved model weights can be found in our GitHub repository (jlab github).

In silico mutagenesis of TFBS

To enable automated, targeted *in silico* mutagenesis of TFBS within loop anchors, we introduce SNIPER, an extension of the GenomicLinks toolkit, designed to analyze sequence features influencing genomic interactions (jlab github, Figure 1B). SNIPER provides a targeted and a genome-wide mode for integrating *in silico* mutations, deletions or motif replacements.

Web server implementation

For web server implementation, a combination of Java scripting, PERL, HTML, CSS and PHP was employed, all running on an Apache server utilizing the Hypertext Transfer Protocol (HTTP). The front end of the web interface was written in JAVA scripts, PHP and HyperText Markup Language (HTML). We utilized PHP and PERL for scripting common gateway interfaces and web–server interactions due to the platform independence and open-source nature of Apache, MySQL and PHP technologies. A step-by-step user guide for GenomicLinks and SNIPER can be found in the SI materials.

Evaluation of TFBS on chromatin looping using *in silico* mutagenesis

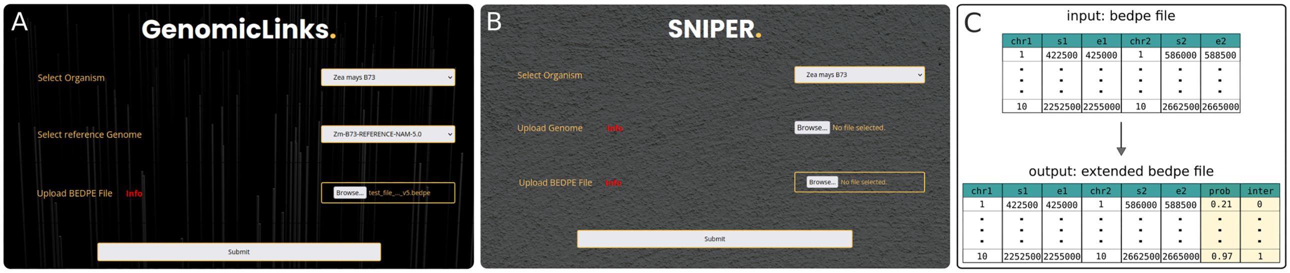
To determine the impact of transcription factor binding sites (TFBS) on Hi-ChIP loops, we used SNIPER to scan and replace nucleotides within any identified TFBS motifs, based on exact matches to the consensus sequence from JASPAR (40–46) database, with an 'N' to denote an unknown nucleotide. We assessed the impact of these modifications by comparing them with original predictions. Anchor pairs with a probability drop of more than 50% post-modification were classified as 'Relevant Anchor Pairs' (RAPs), and their associated motifs were termed 'relevant motifs'. Focusing on chromosome 1 as a test case, we identified 670 RAPs, encompassing 25 287 unique sequences with an average motif length of 8.5 base pairs. For comparison, we created a set of Control Anchor Pairs based on control motifs, which had minimal impact on loop probability, under comparable conditions of count and motif length. Notably, these control motifs were more prevalent, occurring in 2623 Control Anchor Pairs.

Validation of *in silico* predictions using maize genetic variation data

We process single nucleotide variant (SNV) data from the maize 282 diversity panel (47). The complete SNV dataset comprised approximately 2.2 million single nucleotide polymorphisms (SNPs) or indels, exclusively for chromosome 1. To prepare the dataset for GenomicLinks analysis, we updated the Hi-ChIP anchor pairs for each of the 278 maize inbred lines. We extracted DNA sequences from the reference genome and incorporated mutations as specified by the SNV matrix. Mutations were updated accordingly; unknown nucleotides ('N') were encoded as a uniform distribution vector (0.25, 0.25, 0.25 and 0.25)^T, while deletions ('D') were represented by a zero vector (0, 0, 0 and 0)^T. It is important to note that insertions ('I') were excluded from this process. We applied GenomicLinks to predict chromatin interactions within chromosome 1 for all 278 samples, creating a matrix of dimensions 278 x 16 410.

Validation of *in silico* predictions using maize single-cell co-accessibility data

To further validate the *in silico* predictions that specific SNVs affect chromatin loop probabilities, we analyzed single-cell ATAC-seq data (48) for 10 genotypes apart of the maize 282 diversity panel (B73 (reference), B97, CML103, CML52, KI3, Ky21, M162W, M37W, OH7B and Oh43), and processed the



Select Organism: Zea mays B73
Select reference Genome: Zm-B73-REFERENCE-NAM-5.0
Upload BEDPE File: Browse... test_file_..._v5.bedpe

Select Organism: Zea mays B73
Upload Genome: Browse... No file selected.
Upload BEDPE File: Browse... No file selected.

chr1	s1	e1	chr2	s2	e2
1	422500	425000	1	586000	588500
⋮	⋮	⋮	⋮	⋮	⋮
10	2252500	2255000	10	2662500	2665000

chr1	s1	e1	chr2	s2	e2	prob	inter
1	422500	425000	1	586000	588500	0.21	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	2252500	2255000	10	2662500	2665000	0.97	1

Figure 2. GenomicLinks user interface. (**A**) The GenomicLinks web portal interface for uploading BEDPE files, currently supporting *Zea mays* with options for reference genome versions v4 or v5. (**B**) Display of the input format required, and the output generated by GenomicLinks, adding predicted interaction probabilities (prob) and interaction presence (inter) for chromatin anchor pairs (s_1, e_1 and s_2, e_2). (**C**) Display of the input format required and the output generated by GenomicLinks, adding predicted interaction probabilities (prob) and interaction presence (inter) for chromatin anchor pairs (s_1, e_1 and s_2, e_2).

sequenced outputs as previously described (49). To identify regions of co-accessibility as a proxy for chromatin interactions, we first partitioned the nuclei by accessible chromatin regions (ACRs) matrix to restrict the analysis to each genotype in isolation. Nuclei from a single genotype were then aggregated in *X* pseudocells using *k* nearest neighbors, where *X* is an integer determined as the number of nuclei from the focal genotype divided by *k*, and *k* is an integer of the square root of the total number of nuclei from the focal genotype. Tn5 integrations per pseudocell across all ACRs were then scaled per million using the function *cpm* from the R package, *edgeR (50)*. Initial values of co-accessibility between ACRs were estimated as Spearman Correlation Coefficients across pseudocells, conditioning ACRs to be less than 500-kb and more than 2-kb apart. We assessed whether co-accessibility values dropped in anchor pairs, whose SNVs in RAPs were predicted to result in a loss of interaction probability. As a control, we assessed decreases in co-accessibility values in RAPs, where SNVs did not lead to predicted losses in interaction probabilities.

Results

Description of GenomicLinks

GenomicLinks is an open-source web tool designed to predict 3D chromatin interactions from DNA sequence in maize. GenomicLinks builds on a deep learning model, comprising two CNNs and LSTMs and has been trained on curated maize Hi-ChIP data (Figure 1 and Material and methods). Users can upload a pair (or multiple pairs) of genomic sequences in either FASTA or BEDPE formats. These inputs are automatically processed and submitted to the trained model for interaction predictions. GenomicLinks outputs interaction probabilities for each sequence pair from the BEDPE files or with each individual sequence pair from FASTA files (Figure 2).

Comparative benchmarking of GenomicLinks

We evaluated GenomicLinks' classification performance using standard metrics (Figure 3). The evaluation used the complete training data, including 101 172 true positive and an equal number of true negative interactions. This dataset was divided into training (80%, 161 875 samples), validation (10%, 20 234 samples) and test (10%, 20 234 samples) sets. We obtained the following classification performance: accuracy: 0.848, sensitivity (Recall): 0.853, specificity: 0.843, F1-score: 0.848, precision: 0.844, ROC AUC: 0.925, and AUPRC: 0.926). As there are currently no published DL models for predicting chromatin interactions in plant genomes, we sought to compare these classification statistics with those of other models that have been developed for similar purposes in other systems, for example, for the prediction of EPIs in human cell lines. These comparisons are summarized in Supplementary Table S1. We found that GenomicLinks' performance matches or exceeds that of similarly structured models such as SPEID, SEPT and CLNN-loop, even if the latter had been trained and tested on the same cell line, which tends to yield the best results. Applying models trained on other species or different cell lines to our maize data without retraining and adjusting the model architecture is impractical. For instance, applying the human-trained DeepMILO, designed for evaluating non-coding variant impacts on CTCF/cohesin-mediated insulator loops, correctly predicts only 0.1% of maize Hi-ChIP loops, corresponding to 144 out of 101 172 loops (see SI material).

In silico predictions of TF-mediated 3D chromatin contacts

We set out to identify specific DNA sequence features underlying our DL predictions. Several studies have pointed to an enrichment of particular classes of TFs at loop anchors (15,16). This has led to the hypothesis that TFs may mediate loop formation through processes like dimerization/protein interactions (51). To test whether the presence of TFBS in loop anchors contribute to our predictions, we applied SNIPER to perform a TF-centered *in silico* mutagenesis (see Material and Methods). Focusing on chromosome 1 as a test case, we selected anchor pairs, whose 3D interactions were supported empirically by Hi-ChIP data as well as computationally by GenomicLinks (N = 16 083 out of 16 410). Utilizing the JASPER database, we identified all high-confidence TFBS for maize (82 000) within these sequences and applied SNIPER for targeted deletions of these sites (see Material and Methods and Figure 1B). We observed that all anchor pairs contained at least one TFBS, with over 95% of these anchors having less than 20% of their base pairs overlapping TFBSs. The mutated sequences were then re-analyzed using GenomicLinks for interaction prediction. Our analysis identified a subset of approximately 4% of anchor pairs (678 out of 16 083), where TFBS deletions led to a strong probability shift of at least 50%, indicating a shift from interacting to non-interacting state. We refer to this subset as RAPs. This results highlights cases where removal of only approximately 205-bp motif sequence from 5-kb loop anchors (on average) are sufficient to compromise

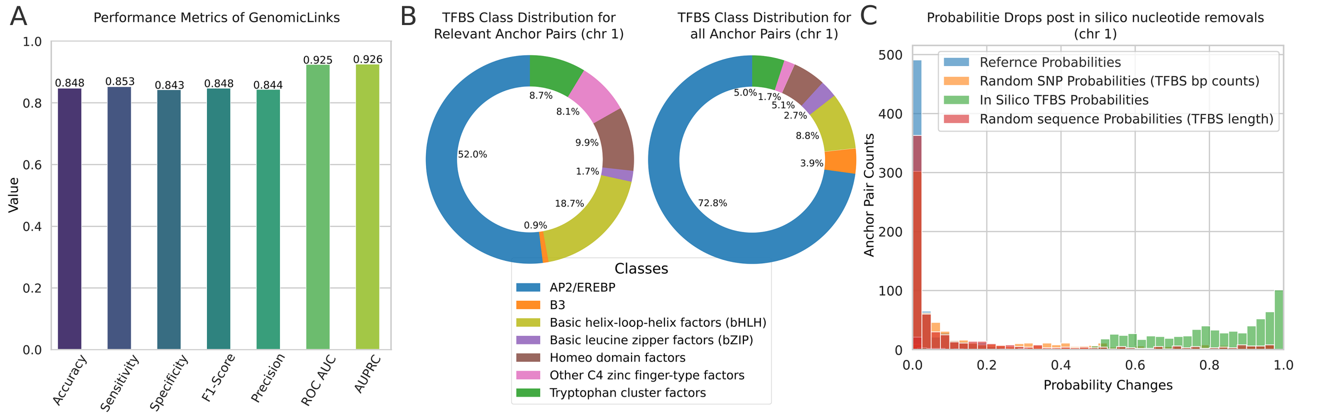


Figure 3. GenomicLinks performance statistics and impact analysis of *in silico* mutagenesis of TFBSs. (**A**) Performance metrics, including accuracy, sensitivity, specificity, F1-score, precision and AUC values. (**B**) *Left:* Distribution of TFBS motif classes within all 16 410 Hi-ChIP anchor pairs on chromosome 1 of the *B73* reference genome. *Right:* Distribution TFBS motif classes within anchor pairs predicted to lose interaction after removal of TFBS, with *bHLH* (yellow) and *Other C4 zinc finger factors* (pink) significantly affecting loop stability. (**C**) Histogram displaying the effect of *in silico* TFBS motif knockout on interaction probability: original (blue) versus collapsing interactions (green), alongside control scenarios (red/orange) verifying the specificity to the targeted TFBS motifs and their position.

looping probabilities. For comparison, we also included two benchmark scenarios into our analysis. In the first, we introduced random SNPs, matching the total base pairs of all motifs within each anchor. In the second, we randomly repositioned deleted motif segments within the anchor sequence (see Figure 3C). No loss of interaction probabilities could be detected in the two control benchmark scenarios (Figure 3C), thus providing additional evidence that these motif sequences drive the predictions. Further examination of the frequency distribution of specific TF classes in the RAPs compared with all loops revealed significant shifts in class distribution (see Figure 3B). The most significant changes were found for the *AP2/EREBP* and *Basic helix-loop-helix* (*bHLH*) classes ($\chi^2 = 87.21$, df = 6, P = 1.85e-23), with the *AP2/EREBPs* showing a substantial depletion in the RAPs (all loops: 72.8%, RAPs: 52%), and the *bHLHs* an enrichment (all loops: 8.7%; RAPs: 18.7%). Interestingly, in mammals, members of the *bHLH* class have been shown to cooperate with histones to facilitate DNA access (52) and contribute to the 3D rewiring of chromatin architecture during cell lineage transitions (53).

Natural mutations in TFBS lead to predicted losses in 3D chromatin contacts

The above analysis demonstrated that *in silico* deletions of entire TFBS can lead to a loss of predicted chromatin interactions in a subset of regions. In natural settings, complete deletions of TFBS are less common than SNVs in the form of SNPs or small indels. We asked if naturally occurring SNVs among maize inbred lines in these TFBS are sufficient to compromise predicted looping probabilities. To assess this, we screened genotype data from 9 maize NAM founder lines and retained only those RAPs that contained at least one SNPs or small indels in a TFBS (Material and methods, Figure 4A). Depending on the NAM line, this resulted in 172–551 RAPs (average of 487) on chromosome 1 for downstream analysis Figure 4B. Employing SNIPER (see Figure 1B), we replaced the B73 sequence in each TFBS with the SNP/indel genotype of the respective NAM line. On average, this translated to only 4.9 bp being substituted in a total of 205 bp of TFBS space per anchor pair (i.e. 2.39% of the TFBS space, see Supplementary Figure S2). After re-applying GenomicLinks to the SNV substituted RAPs, we found that 13% (average of 65 RAPs) exhibited a drop in predicted interaction probability of at least 10%, thus demonstrating considerable sensitivity to minor genetic variation (Figure 4C).

Single-cell co-accessibility validates predicted interaction losses

To empirically validate our predicted interactions losses in the 487 RAPs, we analyzed single-cell ATAC-seq data for 10 NAM lines (48), and quantified chromatin co-accessibility, which has been widely used as a proxy for chromatin interactions (see Material and Methods) (54). As a control, we performed the same quantification in RAPs where SNP/indels mutations had no predicted effect on looping. We found that RAPs that exhibited reduced loop stability as a result of SNP/indel mutations within TFBS also showed a corresponding decrease in co-accessibility strength across cells compared to the B73, with B97 being the exception (Pooled genotype test, $\chi^2 = 7.89$, df = 1, $P = 5.0\text{e-}03$, Figure 4E–G). A similar trend was observed when comparing the number of co-ACRs instead co-accessibility strength (Pooled genotype test, Fisher's Exact Test = 0.82, $P = 1.05\text{e-}02$), which indicated that TFBS-associated SNPs/indels also reduced the extent of open chromatin in the RAPs.

Predicting the impact of genetic variation on chromatin looping in 277 maize inbred lines

We sought to extend our genotype-based predictions of differential chromatin looping to the panel of 277 maize diverse inbred lines (47), which has become an important genomic resource in maize. To that end, we identified all APs genome-wide containing published SNVs (101 172 APs genome-wide). This analysis yielded a large matrix of individual loop predictions (277 × 101 172). We identified significant variability in loop stability across anchor pairs among genotypes (Supplementary Figure S3, Figure 5B). Interestingly, certain genotypes displayed a general trend toward loop loss (e.g. NC328, CI66 and A641), with an average loop stability of only 76% genome-wide. As a control, we examined a specific anchor pair on chromosome 1 surrounding the *tb1*

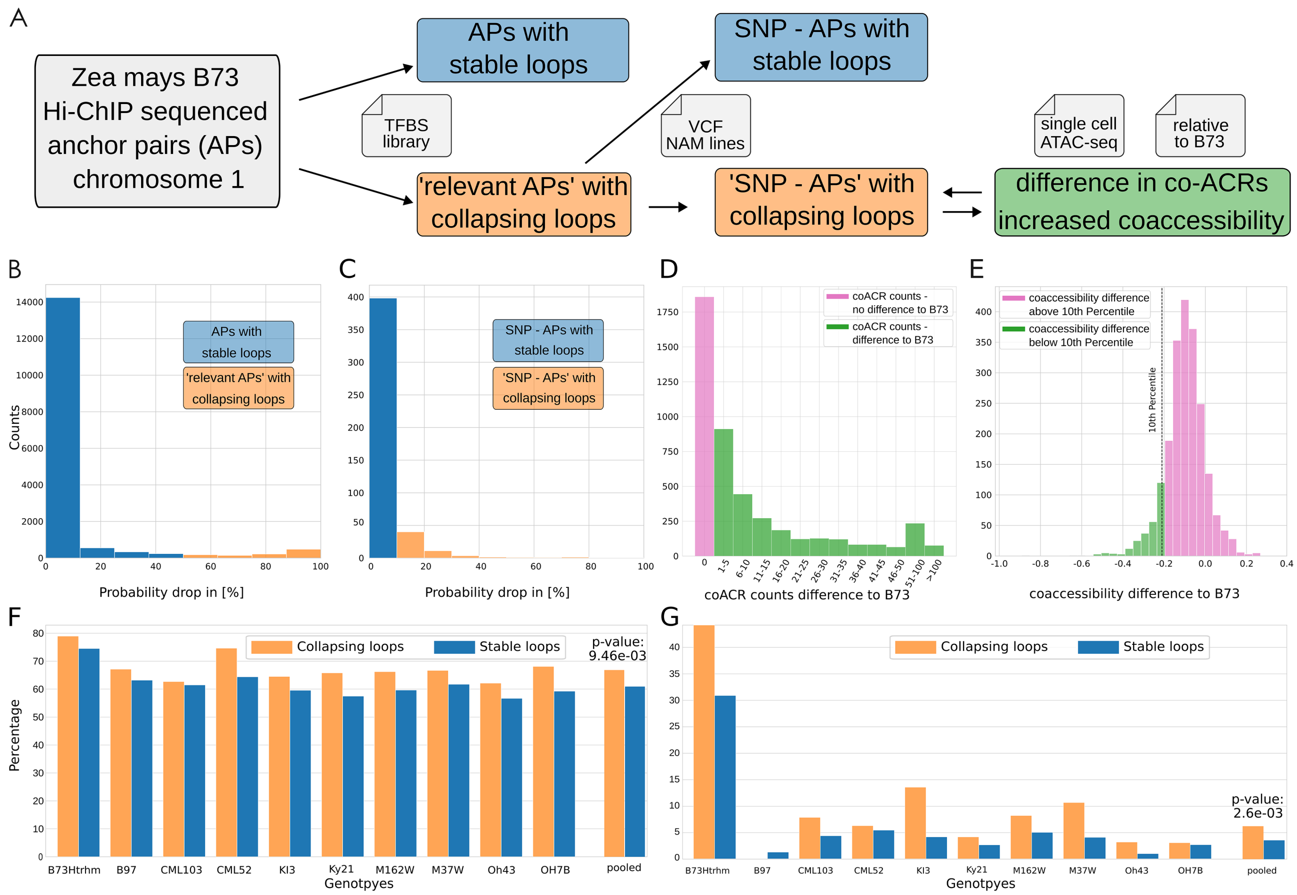


Figure 4. Validating GenomicLinks predictions using natural genetic variation and single cell ATAC-seq data. (**A**) Workflow diagram illustrating the process for extracting RAPs from Hi-ChIP sequenced data on chromosome 1 of *Zea mays B73*. APs are categorized based on their loop stability after *silico test* removal of TFBS motifs. Further analysis mutates SNVs across 10 genotypes to these APs, we call them 'SNP-APs'. Evaluation of 'SNP-APs' reveals a higher occurrence of collapsing loops in those with reduced coACR counts and increased co-accessibility differences relative to the B73 reference. (**B**) Predicted probability drops of APs post-TFBS *silico test* removal indicate a majority of stable loops, defined by a drop of less than 50% from the original loop prediction for B73. (**C**) Predicted probability drops of relevant APs post-SNV *silico test* mutation show a delineation between stable and collapsing loops, now with a tighter 10% cutoff from the original prediction for B73. (**D**) Distribution of co-ACR count differences for all SNP-APs relative to *B73*, highlighting those with a positive difference (labeled green) for further loop analysis. (**E**) Distribution of average co-accessibility differences for all SNP-APs compared to *B73*, highlighting those below the 10th percentile (labeled green) for further loop analysis. (**F**) Normalized analysis for all genotypes reveals that collapsing loops more frequently occur than stable ones in SNP-APs with positive co-ACR count differences (green subset from panel D). (**G**) Normalized analysis across all genotypes indicates that collapsing loops are more frequent than stable ones in SNP-APs with co-accessibility differences below the 10th percentile (green subset from panel E), though B97 deviates from this trend.

locus, which is a well-characterized distal *cis*-regulatory element known to form stable loops in modern maize (18,19,55). Indeed, despite the presence of genetic variation around this locus, with some genotypes showing base substitution in up to 454 bp within this anchor pair, loop probability remained consistently high (>99%) in all 277 inbred lines (Figure 5A). Our complete computational predictions could be tested in future empirical studies of the diversity panel using a combination of targeted QTL mapping and population-level single-cell co-accessibility and/or HiC data. Overall, our results demonstrate that GenomicLinks learned meaningful DNA sequence features predictive of chromatin looping in maize, and that the model can be used to test the impact of specific genomic variants on chromatin looping.

Discussion

Deep learning models like GenomicLinks have the potential to reveal complex and nested sequence motifs and genomic features responsible for the 3D structure of chromatin, which traditional sequence-based methods might miss. This computational framework enables an in-depth analysis of how genetic variants influence chromatin loop stability, a key factor in gene regulation. Additionally, it facilitates the exploration of regulatory variations associated with complex traits and investigates how non-coding loci, through physical interactions with distal genes, could modify gene expression patterns and influence phenotypic outcomes. Yet GenomicLinks has several limitations that are inherent to the 'black box' nature of deep learning models. The model's decision-making process is usually obscured by the complex interplay of millions of parameters across its node and layer architecture. The nested neural structure with its nonlinear operations between each layer makes it challenging to directly interpret how it evaluates and integrates local and distal genomic features. We already introduced a way to try to interpret DL decisions: *in silico* mutagenesis, where we strategically change sequences to observe how such changes impact model predictions. This method can be useful to identify crucial sequence features for the model's decisions, labeling significant genomic regions

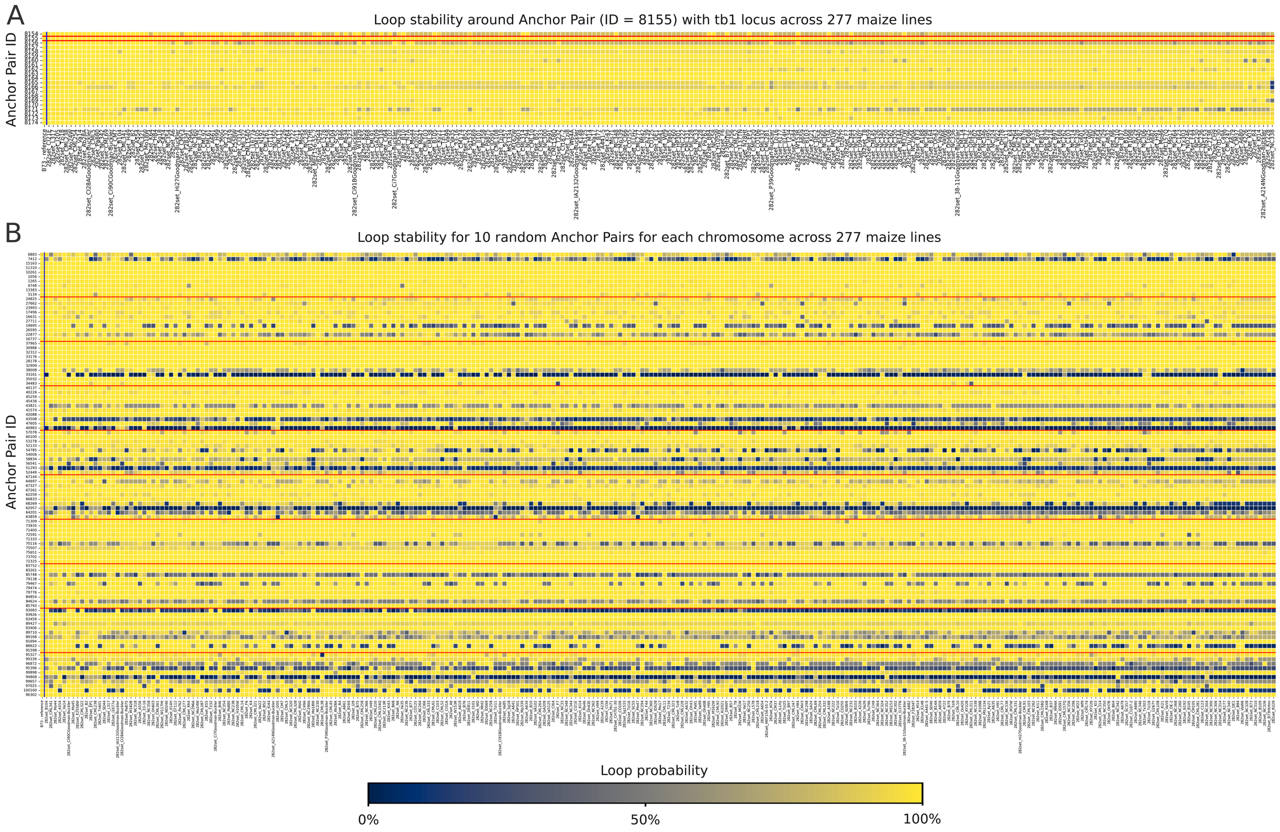


Figure 5. Impact of genetic variation on chromatin looping. (**A**) Interaction probabilities for anchor pairs in the region chr1:270276250-270488750, including the tb1 locus, across 277 maize inbred lines, demonstrating consistent loop stability for tb1. (**B**) Interaction probabilities visualized for 10 randomly selected anchor pairs across all 10 chromosomes, illustrating variability in chromatin loop stability. The analysis reveals genotypes with a higher tendency to lose loop stability (right), identifies regions that consistently exhibit stable interactions and pinpoints anchor pairs prone to collapse across almost all genotypes.

for chromatin interaction predictions. In the context of deep learning analysis, this approach is known as input perturbation, which systematically tests the model's sensitivity and reliance on specific input features. To further address the 'black box' nature of deep learning models, class activation mapping (CAM) provides a more direct method for visualizing influential input segments. CAM produces heat maps that highlight which parts of the input significantly influence the model's decisions, offering insights into the impactful regions within the input sequence for specific anchor pairs. However, CAM's utility is primarily limited to the convolutional layers, and it does not adequately address the complexities introduced by LSTM layers, which are crucial for capturing long-distance dependencies within the chromatin. Therefore, while CAMs can be valuable for in-depth analysis of particular anchor pairs, they are less effective for comprehensive analysis across large datasets. Based on work in animals, it is unlikely that GenomicLinks can be directly applied to other plant species without re-training (56,57). However, transfer learning is a promising strategy to mitigate this limitation. The concept of transfer learning involves retraining the model, initially trained on a specific species like *Zea mays*, on new datasets from different species. This process allows the model to adapt, learning both species-specific features and universal chromatin organization principles. Furthermore, it provides a method for estimating the proportion of chromatin features that are universal versus those that are genome-dependent by comparing models trained directly on various species and those adapted via transfer learning.

Data availability

The data and code supporting the findings of this study are available at the following locations: - The GenomicLinks software and all associated scripts can be accessed at our GitHub repository: https://github.com/jlab-code/GenomicLinks Alternatively: https://doi.org/10.6084/m9.figshare.25771830.v2. The following accession numbers have been used in this study: Hi-ChIP data: GSE120304, scATAC data: GSE155178 and NAM coACRs data: GSE165787.

Supplementary data

Supplementary Data are available at NARGAB Online.

Acknowledgements

This work was partly supported by the Technical University of Munich Institute for Advanced Study, funded by the German Excellence Initiative and the European Seventh Framework Programme under grant agreement no. 291763. We thank members of the Johannes lab for helpful feedback. This research was supported by the National Science Foundation (IOS-1856627) and the UGA Office of Research to RJS, and the National Institute of General Medical Sciences of the National Institutes of Health (1R00GM144742) to APM.

Funding

No external funding.

Conflict of interest statement

None declared.

References

1. Slotkin,R.K. and Martienssen,R. (2007) Transposable elements and the epigenetic regulation of the genome. *Nat. Rev. Genet.*, **8**, 272–285.
2. Deniz,Ö., Frost,J.M. and Branco,M.R. (2019) Regulation of transposable elements by DNA modifications. *Nat. Rev. Genet.*, **20**, 417–431.
3. Stewart-Morgan,K.R., Petryk,N. and Groth,A. (2020) Chromatin replication and epigenetic cell memory. *Nat. Cell Biol.*, **22**, 361–371.
4. Lieberman-Aiden,E., van Berkum,N.L., Williams,L., Imakaev,M., Ragoczy,T., Telling,A., Amit,I., Lajoie,B.R., Sabo,P.J., Dorschner,M.O., *et al.* (2009) Comprehensive mapping of long-range interactions reveals folding principles of the human genome. *Science*, **326**, 289–293.
5. Mumbach,M.R., Rubin,A.J., Flynn,R.A., Dai,C., Khavari,P.A., Greenleaf,W.J. and Chang,H.Y. (2016) HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nat. Methods*, **13**, 919–922.
6. Fullwood,M.J., Liu,M.H., Pan,Y.F., Liu,J., Xu,H., Mohamed,Y.B., Orlov,Y.L., Velkov,S., Ho,A., Mei,P.H., *et al.* (2009) An oestrogen-receptor-alpha-bound human chromatin interactome. *Nature*, **462**, 58–64.
7. Dryden,N.H., Broome,L.R., Dudbridge,F., Johnson,N., Orr,N., Schoenfelder,S., Nagano,T., Andrews,S., Wingett,S., Kozarewa,I., *et al.* (2014) Unbiased analysis of potential targets of breast cancer susceptibility loci by Capture Hi-C. *Genome Res.*, **24**, 1854–1868.
8. Hughes,J.R., Roberts,N., McGowan,S., Hay,D., Giannoulatou,E., Lynch,M., De Gobbi,M., Taylor,S., Gibbons,R. and Higgs,D.R. (2014) Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nat. Genet.*, **46**, 205–212.
9. Dekker,J., Rippe,K., Dekker,M. and Kleckner,N. (2002) Capturing chromosome conformation. *Science*, **295**, 1306–1311.
10. Dostie,J., Richmond,T.A., Arnaout,R.A., Selzer,R.R., Lee,W.L., Honan,T.A., Rubio,E.D., Krumm,A., Lamb,J., Nusbaum,C., *et al.* (2006) Chromosome Conformation Capture Carbon Copy (5C): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, **16**,1299 –1309.
11. Rao,S.S., Huntley,M.H., Durand,N.C., Stamenova,E.K., Bochkov,I.D., Robinson,J.T., Sanborn,A.L., Machol,I., Omer,A.D., Lander,E.S., *et al.* (2014) A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, **159**, 1665–1680.
12. Dixon,J.R., Selvaraj,S., Yue,F., Kim,A., Li,Y., Shen,Y., Hu,M., Liu,J.S. and Ren,B. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, **485**, 376–380.
13. Dong,P., Tu,X., Chu,P.-Y., Lü,P., Zhu,N., Grierson,D., Du,B., Li,P. and Zhong,S. (2017) 3D Chromatin Architecture of Large Plant Genomes Determined by Local A/B Compartments. *Mol. Plant*, **10**, 1497–1509.
14. Feng,S., Cokus,S.J., Schubert,V., Zhai,J., Pellegrini,M. and Jacobsen,S.E. (2014) Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis. *Mol. Cell*, **55**, 694 –707.
15. O'Malley,R.C., Huang,S.C., Song,L., Lewsey,M.G., Bartlett,A., Nery,J.R., Galli,M., Gallavotti,A. and Ecker,J.R. (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, **165**, 1280–1292.
16. Karaaslan,E.S., Wang,N., Faiß,N., Liang,Y., Montgomery,S.A., Laubinger,S., Berendzen,K.W., Berger,F., Breuninger,H. and Liu,C. (2020) Marchantia TCP transcription factor activity correlates with three-dimensional chromatin structure. *Nat Plants*, **6**, 1250–1261.
17. Piecyk,R.S., Schlegel,L. and Johannes,F. (2022) Predicting 3D chromatin interactions from DNA sequence using Deep Learning. *Comput. Struct. Biotechnol. J.*, **20**, 3439–3448.
18. Ricci,W.A., Lu,Z., Ji,L., Marand,A.P., Ethridge,C.L., Murphy,N.G., Noshay,J.M., Galli,M., Mejía-Guerra,M.K., Colomé-Tatché,M., *et al.* (2019) Widespread long-range cis-regulatory elements in the maize genome. *Nat Plants*, **5**, 1237–1249.
19. Studer,A., Zhao,Q., Ross-Ibarra,J. and Doebley,J. (2011) Identification of a functional transposon insertion in the maize domestication gene tb1. *Nat. Genet.*, **43**,1160–1163.
20. Salvi,S., Sponza,G., Morgante,M., Tomes,D., Niu,X., Fengler,K.A., Meeley,R., Ananiev,E.V., Svitashev,S., Bruggemann,E., *et al.* (2007) Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize. *Proc. Natl. Acad. Sci. U.S.A.*, **104**, 11376–11381.
21. Castelletti,S., Tuberosa,R., Pindo,M. and Salvi,S. (2014) A MITE transposon insertion is associated with differential methylation at the maize flowering time QTL Vgt1. *G3*, **4**,805–812.
22. Roudier,F., Ahmed,I., Bérard,C., Sarazin,A., Mary-Huard,T., Cortijo,S., Bouyer,D., Caillieux,E., Duvernois-Berthet,E., Al-Shikhley,L., *et al.* (2011) Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. *EMBO J.*, **30**, 1928–1938.
23. Zhang,X., Clarenz,O., Cokus,S., Bernatavichute,Y.V., Pellegrini,M., Goodrich,J. and Jacobsen,S.E. (2007) Whole-genome analysis of histone H3 lysine 27 trimethylation in Arabidopsis. *PLoS Biol.*, **5**, e129.
24. Zhang,X., Bernatavichute,Y.V., Cokus,S., Pellegrini,M. and Jacobsen,S.E. (2009) Genome-wide analysis of mono-, di- and trimethylation of histone H3 lysine 4 in Arabidopsis thaliana. *Genome Biol.*, **10**, R62.
25. Servant,N., Varoquaux,N., Lajoie,B.R., Viara,E., Chen,C.-J., Vert,J.-P., Heard,E., Dekker,J. and Barillot,E. (2015) HiC-Pro: an optimized and flexible pipeline for Hi-C data processing. *Genome Biol.*, **16**, 259.
26. Jiao,Y., Peluso,P., Shi,J., Liang,T., Stitzer,M.C., Wang,B., Campbell,M.S., Stein,J.C., Wei,X., Chin,C.-S., *et al.* (2017) Improved maize reference genome with single-molecule technologies. *Nature*, **546**, 524–527.
27. Bhattacharyya,S., Chandra,V., Vijayanand,P. and Ay,F. (2019) Identification of significant chromatin contacts from HiChIP data by FitHiChIP. *Nat. Commun.*, **10**, 4221.
28. Wall,B.P.G., Nguyen,M., Harrell,J.C. and Dozmorov,M.G. (2024) Machine and deep learning methods for predicting 3D genome organization. arXiv doi: https://arxiv.org/abs/2403.03231, 04 March 2024, preprint: not peer reviewed.
29. Zhou,J. and Troyanskaya,O.G. (2015) Predicting effects of noncoding variants with deep learning–based sequence model. *Nat. Methods*, **12**, 931–934.
30. Abadi,M., Agarwal,A., Barham,P., Brevdo,E., Chen,Z., Citro,C., Corrado,G.S., Davis,A., Dean,J., Devin,M., *et al.* (2015) TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems. Google, Mountain View, CA, USA.
31. Chollet,F. (2015) Keras: The Python Deep Learning Library. GitHub, San Francisco, CA, USA.
32. Trieu,T., Martinez-Fundichely,A. and Khurana,E. (2020) DeepMILO: a deep learning approach to predict the impact of non-coding sequence variants on 3D chromatin structure. *Genome Biol.*, **21**, 79.

33. Jing,F., Zhang,S.-W. and Zhang,S. (2020) Prediction of enhancer-promoter interactions using the cross-cell type information and domain adversarial neural network. *BMC Bioinformatics*, **21**, 507.
34. Singh,S., Yang,Y., Póczos,B. and Ma,J. (2019) Predicting enhancer-promoter interaction from genomic sequence with deep neural networks. *Quant Biol*, **7**, 122–137.
35. Hong,Z., Zeng,X., Wei,L. and Liu,X. (2020) Identifying enhancer-promoter interactions with neural network based on pre-trained DNA vectors and attention mechanism. *Bioinformatics*, **36**, 1037–1043.
36. Min,X., Ye,C., Liu,X. and Zeng,X. (2021) Predicting enhancer-promoter interactions by deep learning and matching heuristic. *Brief. Bioinform.*, **22**, 1733–1745.
37. Bello,M., Nápoles,G., Sánchez,R., Bello,R. and Vanhoof,K. (2020) Deep neural network to extract high-level features and labels in multi-label classification problems. *Neurocomputing*, **413**, 259–270.
38. Hochreiter,S. and Schmidhuber,J. (1997) Long short-term memory. *Neural Comput.*, **9**, 1735–1780.
39. Zhang,Z. (2018) Improved Adam optimizer for deep neural networks. In: *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*. IEEE, Banff, AB, Canada.
40. Fornes,O., Castro-Mondragon,J.A., Khan,A., van der Lee,R., Zhang,X., Richmond,P.A., Modi,B.P., Correard,S., Gheorghe,M., Baranašić,D., *et al.* (2020) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **48**,D1, D87–D92.
41. Yanagisawa,S. and Schmidt,R.J. (1999) Diversity and similarity among recognition sequences of Dof transcription factors. *Plant J.*, **17**, 209–214.
42. Kozaki,A., Hake,S. and Colasanti,J. (2004) The maize ID1 flowering time regulator is a zinc finger protein with novel DNA binding properties. *Nucleic Acids Res.*, **32**, 1710–1720.
43. Niu,X., Helentjaris,T. and Bate,N.J. (2002) Maize ABI4 binds coupling element1 in abscisic acid and sugar response genes. *Plant Cell*, **14**, 2565–2575.
44. Vollbrecht,E., Springer,P.S., Goh,L., Buckler,E.S. and Martienssen,R. (2005) Architecture of floral branch systems in maize and related grasses. *Nature*, **436**, 1119–1126.
45. Weirauch,M.T., Yang,A., Albu,M., Cote,A.G., Montenegro-Montero,A., Drewe,P., Najafabadi,H.S., Lambert,S.A., Mann,I., Cook,K., *et al.* (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell*, **158**, 1431–1443.
46. Boer,D.R., Freire-Rios,A., van den Berg,W.A., Saaki,T., Manfield,I.W., Kepinski,S., López-Vidrieo,I., Franco-Zorrilla,J.M., de Vries,S.C., Solano,R., *et al.* (2014) Structural basis for DNA binding specificity by the auxin-dependent ARF transcription factors. *Cell*, **156**, 577–589.
47. Bukowski,R., Guo,X., Lu,Y., Zou,C., He,B., Rong,Z., Wang,B., Xu,D., Yang,B., Xie,C., *et al.* (2017) Construction of the third-generation Zea mays haplotype map. *Gigascience*, **7**, gix134.
48. Hufford,M.B., Seetharam,A.S., Woodhouse,M.R., Chougule,K.M., Ou,S., Liu,J., Ricci,W.A., Guo,T., Olson,A., Qiu,Y., *et al.* (2021) De novo assembly, annotation, and comparative analysis of 26 diverse maize genomes. *Science*, **373**, 655–662.
49. Marand,A.P., Chen,Z., Gallavotti,A. and Schmitz,R.J. (2021) A cis-regulatory atlas in maize at single-cell resolution. *Cell*, **184**, 3041–3055.
50. Robinson McCarthy,D.J. and Smyth,G.K. (2010) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, **26**, 139–140.
51. Ji,X., Dadon,D.B., Powell,B.E., Fan,Z.P., Borges-Rivera,D., Shachar,S., Weintraub,A.S., Hnisz,D., Pegoraro,G., Lee,T.I., *et al.* (2016) 3D chromosome regulatory landscape of human pluripotent cells. *Cell Stem Cell*, **18**, 262–275.
52. Michael,A.K., Stoos,L., Crosby,P., Eggers,N., Nie,X.Y., Makasheva,K., Minnich,M., Healy,K.L., Weiss,J., Kempf,G., *et al.* (2023) Cooperation between bHLH transcription factors and histones for DNA access. *Nature*, **619**, 385–393.
53. Dall'Agnese,A., Caputo,L., Nicoletti,C., di Iulio,J., Schmitt,A., Gatto,S., Diao,Y., Ye,Z., Forcato,M., Perera,R., *et al.* (2019) Transcription factor-directed re-wiring of chromatin architecture for somatic cell nuclear reprogramming toward trans-differentiation. *Mol. Cell*, **76**, 453–472.
54. Gate,R.E., Cheng,C.S., Aiden,A.P., Siba,A., Tabaka,M., Lituiev,D., Machol,I., Gordon,M.G., Subramaniam,M., Shamim,M., *et al.* (2018) Genetic determinants of co-accessible chromatin regions in activated T cells across humans. *Nat. Genet.*, **50**, 1140–1150.
55. Igartua,E., Contreras-Moreira,B. and Casas,A.M. (2020) TB1: from domestication gene to tool for many trades. *J. Exp. Bot.*, **71**, 4621–4624.
56. Fudenberg,G., Kelley,D.R. and Pollard,K.S. (2020) Predicting 3D genome folding from DNA sequence with Akita. *Nat. Methods*, **17**, 1111–1117.
57. Schwessinger,R., Gosden,M., Downes,D., Brown,R.C., Marieke Oudelaar,A., Telenius,J., Teh,Y.W., Lunter,G. and Hughes,J.R. (2020) DeepC: predicting 3D genome folding using megabase-scale transfer learning. *Nat. Methods*, **17**, 1118.

Received: May 24, 2024. **Revised:** July 25, 2024. **Editorial Decision:** August 28, 2024. **Accepted:** August 30, 2024