Received: 8 December 2021 | Accepted: 20 June 2022

DOI: 10.1002/tpg2.20249

The Plant Genome

ORIGINAL RESEARCH

Modeling chromatin state from sequence across angiosperms using recurrent convolutional neural networks

Travis Wrightsman[1] | **Alexandre P. Marand**[2] | **Peter A. Crisp**[3] | **Nathan M. Springer**[4] | **Edward S. Buckler**[1,5,6]

[1]Section of Plant Breeding and Genetics, Cornell Univ., Ithaca, NY 14853, USA

[2]Dep. of Genetics, Univ. of Georgia, Athens, GA 30602, USA

[3]School of Agriculture and Food Sciences, Univ. of Queensland, Brisbane, QLD 4072, Australia

[4]Dep. of Plant and Microbial Biology, Univ. of Minnesota, Saint Paul, MN 55108, USA

[5]Institute for Genomic Diversity, Cornell Univ., Ithaca, NY 14853, USA

[6]USDA-ARS, Ithaca, NY 14853, USA

Correspondence
Travis Wrightsman, Section of Plant Breeding and Genetics, Cornell Univ., Ithaca, NY, USA 14853.
Email: tw493@cornell.edu

Assigned to Associate Editor Thomas Jacobs.

Funding information
NSF Graduate Research Fellowship, Grant/Award Number: DGE-1650441; NSF Postdoctoral Fellowship in Biology, Grant/Award Number: DBI-1905869; Australian Research Council (ARC) Discovery Early Career Award, Grant/Award Number: DE200101748; USDA-ARS, Grant/Award Number: NSF : IOS-1934384

Abstract

Accessible chromatin regions are critical components of gene regulation but modeling them directly from sequence remains challenging, especially within plants, whose mechanisms of chromatin remodeling are less understood than in animals. We trained an existing deep-learning architecture, DanQ, on data from 12 angiosperm species to predict the chromatin accessibility in leaf of sequence windows within and across species. We also trained DanQ on DNA methylation data from 10 angiosperms because unmethylated regions have been shown to overlap significantly with ACRs in some plants. The across-species models have comparable or even superior performance to a model trained within species, suggesting strong conservation of chromatin mechanisms across angiosperms. Testing a maize (*Zea mays* L.) held-out model on a multi-tissue chromatin accessibility panel revealed our models are best at predicting constitutively accessible chromatin regions, with diminishing performance as cell-type specificity increases. Using a combination of interpretation methods, we ranked JASPAR motifs by their importance to each model and saw that the TCP and AP2/ERF transcription factor (TF) families consistently ranked highly. We embedded the top three JASPAR motifs for each model at all possible positions on both strands in our sequence window and observed position- and strand-specific patterns in their importance to the model. With our publicly available across-species 'a2z' model it is now feasible to predict the chromatin accessibility and methylation landscape of any angiosperm genome.

Abbreviations: ACR, accessible chromatin region; ATAC-seq, assay for transposase-accessible chromatin with sequencing; auPR, area under the precision-recall curve; CNN, convolutional neural network; GIA, global importance analysis; scATAC, single-cell ATAC; TF, transcription factor.

1 | INTRODUCTION

Accessible chromatin regions (ACRs) are known to play a critical role in eukaryotic gene regulation but their comprehensive identification in plants remains a challenge (Marand et al., 2017; Weber et al., 2016). Current methods to assay chromatin accessibility are highly environment specific and relatively expensive compared with DNA sequencing, limiting the number of species or conditions that can be investigated. Assaying chromatin accessibility in plants comes with additional unique challenges: the cell wall makes plant nuclei hard to isolate and many active transposon families shuffle, create, and destroy regulatory regions over time (Hirsch & Springer, 2017). Regions that lack DNA methylation are known to be stable over developmental time and overlap significantly with ACRs in plants with larger genomes (Crisp et al., 2020), suggesting they may contain a superset of ACRs across cell types. Computational models capable of predicting chromatin accessibility and methylation state directly from DNA sequence would enable a wide range of previously intractable studies on gene regulation across evolutionary time as well as estimation of noncoding variant effects for use in contexts such as breeding. Plants also provide an excellent system to study the genetic basis of adaptation (Anderson et al., 2011). Now that it is feasible to assemble genomes of thousands of species, regulatory regions that control adaptation can be identified, providing valuable insight on how to breed crops resilient to climate change. Recent advances in machine learning, particularly deep learning, have catalyzed a vast number of applications to biological prediction including RNA abundance (Agarwal & Shendure, 2020; Avsec, Agarwal, et al., 2021; Washburn et al., 2019), chromatin state (Kelley, 2020; Quang & Xie, 2016; Zhou & Troyanskaya, 2015), and transcription factor (TF) binding (Tu et al., 2020) directly from DNA sequence. Many of these models have so far only been trained within a single species to predict within the same species, usually using held-out chromosomes as a test set to control for sequence relatedness.

At a high level, plant chromatin has characteristics similar to animal chromatin: chromatin is organized into hierarchical compartments, distal regulatory regions are colocalized to genes through chromatin looping, and various histone modifications signal a wide variety of local chromatin states. However, the exact mechanisms driving chromatin accessibility are known to be quite different in terms of specific histone modifications (Lu et al., 2019), pioneer factors (Yamaguchi, 2021), and chromatin looping mechanisms (Doğan & Liu, 2018). Because of these differences, plant-specific chromatin accessibility models are likely to be necessary.

We know that TF binding sites are strongly conserved across evolutionary time (Chen et al., 2018; Tu et al., 2020) and highly enriched in ACRs (Shlyueva et al., 2014). Certain deep-learning model architectures, such as convolutional neural networks (CNNs), have already been shown effective for predicting chromatin accessibility within species by recognizing important motifs (Quang & Xie, 2016; Zhou & Troyanskaya, 2015) and their spatial relationships (Avsec, Weilert, et al., 2021). These CNN-based architectures can accurately predict chromatin accessibility in humans (Liu et al., 2017; Quang & Xie, 2016; Zhou & Troyanskaya, 2015) as well as in plants (Shen et al., 2021; Zhao et al., 2021). However, the vast majority of previous work has focused on improving performance within species and across cell types with little focus on across-species prediction (Kelley, 2020) or prediction within unobserved species. Previous work (Chen et al., 2018; Krützfeldt et al., 2020) has observed that CNNs require much larger training data sets than earlier model architectures to achieve equivalent or better performance. By incorporating multiple species into the training data, we not only increase the number of observations but also the total evolutionary time between observations, which reduces confounding neutral variation within conserved sequences. For the purposes of predicting regulatory regions in unobserved plant species, training a model across species will be critical to learn important motifs and syntax that are conserved across longer evolutionary time periods. Therefore, we predicted that previously published deep-learning architectures could work well across species and make accurate chromatin accessibility and methylation predictions in related unobserved species.

Here, we train DanQ (Quang & Xie, 2016) to predict chromatin accessibility in leaf using assay for transposase-accessible chromatin with sequencing (ATAC-seq) data from 12 angiosperm species (Lu et al., 2019), comparing the performance of within-species models to across-species models. We also train DanQ to predict unmethylated regions using methylation data from 10 angiosperm species including five previously published grasses (Crisp et al., 2020). Using a maize single-cell ATAC (scATAC) accessibility atlas (Marand et al., 2021), we see that the accessibility model has similar performance across cell types but is highly variable across regions with different levels of cell-type specificity. Using various interpretation methods designed for CNNs, we compare and contrast which motifs were

Core Ideas

- Cross-species models of chromatin state from sequence are comparable or superior to within-species models.
- Model performance is highest on accessible regions open in many tissues.
- Transcription factor motifs can be ranked by importance to each species and chromatin state.

important across angiosperms for predicting chromatin accessibility in leaves or methylation state. Our publicly available pan-angiosperm chromatin state models are an important stepping stone toward a better understanding of gene regulation and adaptation.

2 | MATERIALS AND METHODS

2.1 | Software environment

The software environment for the experiments was managed by conda (v4.10.3). Packages were downloaded from the conda-forge (conda-forge Community, 2015) and bioconda (Grüning et al., 2018) channels. Software versions not explicitly mentioned in the methods are defined in the conda environment files in the companion code repository on Zenodo (Wrightsman et al., 2021). All of the code and data files required to reproduce this manuscript, including all experiments and figures, are available in the associated Zenodo repository. The trained parameters for all models are also available in the associated Zenodo repository.

2.2 | Raw data

The angiosperm ATAC-seq peaks (Lu et al., 2019) were downloaded from NCBI GEO accession GSE128434. Genomes and annotations for *Arabidopsis thaliana* (L.) Heynh. (TAIR10) (Lamesch et al., 2011), *Eutrema salsugineum* (Pall.) Al-Shehbaz & Warwick (v1.0) (Yang et al., 2013), common bean (*Phaseolus vulgaris* L.) (v1.0) (Schmutz et al., 2014), soybean [*Glycine max* (L.) Merr.] (Wm82.a2.v1) (Schmutz et al., 2010), *Brachypodium distachyon* (L.) Beauv. (v3.0) (The International Brachypodium Initiative, 2010), rice (*Oryza sativa* L.) (v7.0) (Ouyang et al., 2007), green foxtail [*Setaria viridis* (L.) P. Beauv.] (v1.0) (Mamidi et al., 2020), poplar (*Populus trichocarpa* Torr. & A. Gray) (v3.0) (Tuskan et al., 2006), and sorghum [*Sorghum bicolor* (L.) Moench] (v3.1 and v3.1.1) (McCormick et al., 2017) were downloaded from Phytozome. Reference genomes and annotations for maize (*Zea mays* L.) (AGPv4.38) (Jiao et al., 2017) and barley (*Hordeum vulgare* L.) (IBSC_v2) (Mascher et al., 2017) were downloaded from Ensembl Plants. The genome and annotation for asparagus (*Asparagus officinalis* L.) (v1.1) (Harkess et al., 2017) was downloaded from the Asparagus Genome Project website. Unmethylated regions for the grasses were downloaded from the supplemental information of Crisp et al. (2020). For the unmethylated regions, the maize AGPv4 genome and annotation was downloaded from MaizeGDB. The grapevine (*Vitis vinifera* L.) genome and annotation (Genoscope.12X) (The French–Italian Public Consortium for Grapevine Genome Characterization, 2007) were downloaded from the Genoscope website.

JASPAR 2020 Core *Plantae* (Fornes et al., 2019) motifs and clusters were downloaded from the JASPAR website. Maize AGPv4 RepeatMasker annotations were downloaded from NCBI. Yeast and human cell-line GM12878 ATAC-seq peaks (Schep et al., 2015) were downloaded from NCBI GEO accession GSE66386. The yeast (sacCer3 April 2011) (Mewes et al., 1997) and human (hg19) (Church et al., 2011) genomes were downloaded from NCBI. Maize scATAC-seq peaks (Marand et al., 2021) were downloaded from NCBI GEO accession GSE155178. Genome files were indexed using SAMtools (Danecek et al., 2021).

2.3 | Unmethylated region calling on non-grass species

Unmethylated region analysis on the nongrass species was performed as per Crisp et al. (2020) using the data summarized in Supplemental Table S14. Briefly, sequencing reads were trimmed and quality checked using Trim galore! (0.6.4_dev), powered by cutadapt (v1.18) (Martin, 2011) and fastqc (v0.11.4). For all libraries, 20 bp were trimmed from the 5′ ends of both R1 and R2 reads and aligned with bsmap (v2.74) (Xi & Li, 2009) to the respective genomes with the following parameters: -v 5 to allow five mismatches, -r 0 to report only unique mapping pairs, and -p 1 and -q 20 to allow quality trimming to Q20. Output SAM files were parsed with SAMtools (Li et al., 2009) fixsam, sorted, and then indexed. Picard MarkDuplicates (Broad Institute, 2019) was used to remove duplicates, BamTools filter to remove improperly paired reads, and bamUtil clipOverlap (Jun et al., 2015) to trim overlapping reads so as to only count cytosines once per sequenced molecule in a pair for paired-end reads. The methylratio.py script from bsmap was used to extract per-site methylation data summaries for each context (CH/CHG/CHH) and reads were summarized into nonoverlapping 100-bp windows tiling the genome. Whole-genome bisulfite sequencing pipelines are available on GitHub. To identify unmethylated regions, each 100-bp tile of the genome was classified into one of six domains or types—'missing data' (including 'no data' and 'no sites'), 'high CHH/RdDM', 'Heterochromatin', 'CG only', 'Unmethylated' or 'intermediate'—in preferential order as per Crisp et al. (2020).

2.4 | Training data preprocessing

Interval manipulation was done using a combination of the GNU coreutils, gawk, and bedtools (Quinlan & Hall, 2010). We created our positive observations by symmetrically extending each accessible or unmethylated region from the midpoint by half of the window size (300, 600, or 1000 bp).

Our negative observations are randomly sampled from the rest of the genome not covered by the union of the resized positive observations and the original peaks. Observations were labeled as 'genic' if more than half of the range overlapped with a gene annotation, as 'proximal' if not genic and more than half of the range was within the proximal cutoff (2 kb), and as 'distal' if neither genic nor proximal. Previous work (Krützfeldt et al., 2020; Wei & Dunbrack, 2013) has shown that classifiers train best on balanced sets with an equal number of positive and negative examples but should be tested on the true class distribution to get an accurate performance estimate. Therefore, for the across-species models, we randomly sampled 6% of the observations and divided them equally between a validation and test set. For the within-species models, we randomly chose a hold-out chromosome to follow best practice for reducing contamination of related sequences between the training and test sets. As a heuristic to select held-out chromosomes across genome assemblies of varying contiguity, we randomly select within chromosomes that are at least a million base pairs long and have more than five positive observations. We then down sampled the remaining observations to obtain a training set for the across-species models with a balanced representation of species and target class. The Ns were encoded as vectors with equal probability assigned to each base as opposed to all zeros, which is another common practice. Sequences were extracted using BioPython (Cock et al., 2009) and pyfaidx (Shirley et al., 2015)

2.5 | Training and evaluating models

The DanQ, Basset, CharPlant, and DeeperDeepSEA architectures were implemented and trained using Keras (Chollet, 2015) and TensorFlow (TensorFlow Developers, 2022). The across-species models were tested on a given species and trained on the remainder. Within-species models were tested on a held-out chromosome and trained on the other chromosomes. Because our ratio of accessible to inaccessible chromatin observations is heavily unbalanced, we focus more on the area under the precision-recall curve (auPR) to measure model performance as opposed to the more commonly reported area under the receiver operating characteristic curve. Performance metrics were measured using scikit-learn (Pedregosa et al., 2011) and curves were plotted using matplotlib (Hunter, 2007). Each model was trained three times to obtain an estimate of variability in performance because of the stochastic nature of the model variable initialization. For comparison between models, we used the first of the three trained models.

The bag-of-*k*-mers model was trained and tested independently on the within-species maize accessibility and methylation training data using code adapted from Tu et al. (2020) and compared with the within-species maize accessibility and methylation models. The Basset, CharPlant, and DeeperDeepSEA architectures were trained in the same manner as the across-species DanQ configuration with the same window size of 600 bp. For the two-step masked model comparison, we masked the maize-held-out accessibility and methylation model predictions to zero if more than half of a region overlapped with an annotated repeat from RepeatMasker. We used pybedtools (Dale et al., 2011) to compute overlaps between the test set and the repeats. We preprocessed the yeast and human cell line ATAC-seq peaks in the same manner as the angiosperm ATAC-seq peaks and used the maize-held-out model to make predictions on the yeast and human peaks.

The grasses accessibility model was trained and evaluated in the same manner as the across-species angiosperm accessibility model but restricted to only grass species. The 'balDist' accessibility model extended the training data balancing to distance class in addition to chromatin state, meaning the training data had equal representation for each species, distance class (genic, proximal, distal), and target class (accessible or inaccessible or unmethylated or methylated). The 'exp' accessibility model changed the activation function on the convolutional layer from rectified linear unit to exponential. The 'all_v_AtZm' accessibility model was tested on *Arabidopsis* and maize and trained on the rest of the angiosperm species. All trained model weights are available on Zenodo (Wrightsman et al., 2021).

The dendrogram in Figure 1 was plotted using the Phylo package of Biopython (Talevich et al., 2012).

2.6 | Analysis of maize scATAC-seq data

The scATAC-seq peaks were preprocessed in the same manner as the other peaks to generate uniform 600-bp regions. Peaks were classified as open in a cell type if their counts per million (a normalized depth measurement) value was greater than $\log_2 5$ in that cell type, which would represent no reads observed in that peak in that cell type, based on the methods reported in Marand et al. (2021). Accessibility was predicted using the maize-held-out model.

2.7 | TF-MoDISco and *k*-mer occlusion

We ran TF-MoDISco (Shrikumar et al., 2020) with a sliding window size of 15 bp, a flank size of 5 bp, and a target seqlet false discovery rate of 0.15. For converting seqlets to patterns, we set 'trim_to_window_size' to 15 bp, 'initial_flank_to_add' to 5 bp and specified a final minimum cluster size of 60.

The *k*-mer-occlusion method involves masking (replacing with Ns) a sliding *k*-mer across each sequence in a given model's test set. The difference between the model's masked and unmasked prediction is the *k*-mer's 'effect size'. We ran the *k*-mer-occlusion method with a *k*-mer size of 10 bp on all

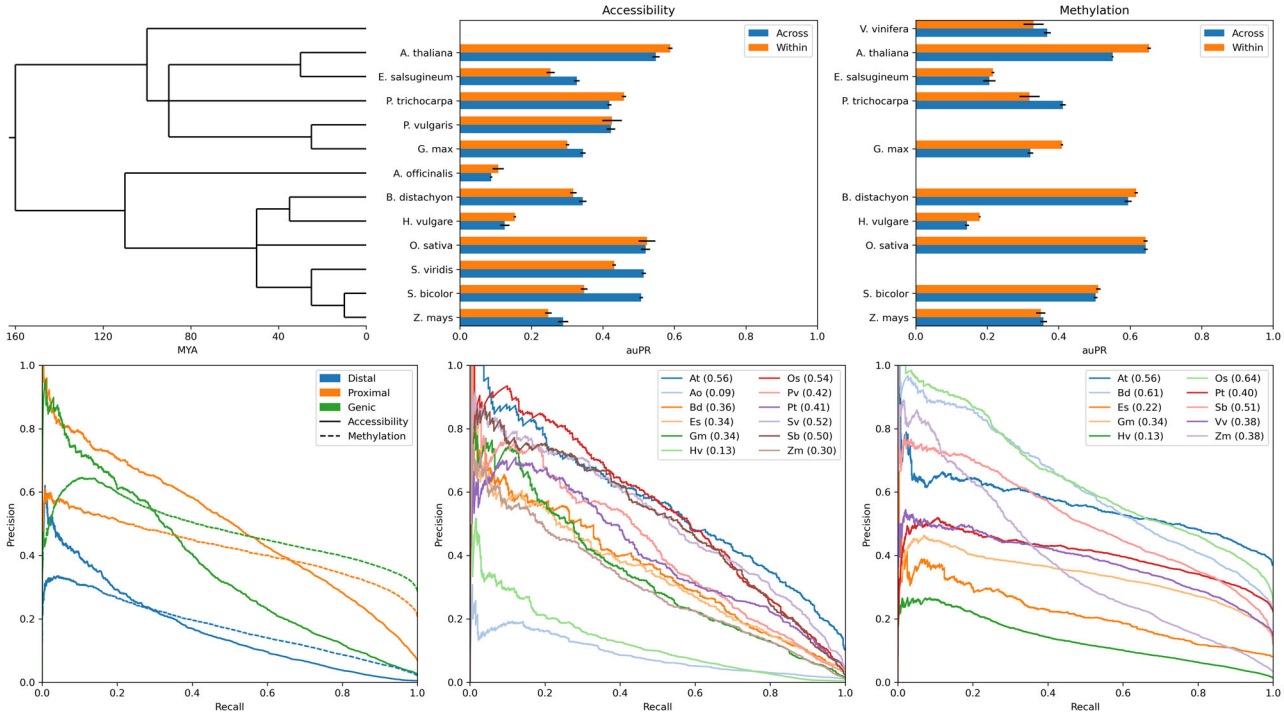

FIGURE 1 Performance of the across-species chromatin state classifiers. The top middle and top right show the mean and standard error (due to variability in the stochastic model training process) of the area under the precision-recall curve (auPR) for the accessibility and methylation models, respectively, per species for both the within- and across-species training configurations. The bottom left is the precision-recall curve across all hold-out species for the across-species models split by distance class and chromatin feature. The bottom middle and bottom right are the precision-recall curves for the across-species accessibility and methylation models, respectively, split by species. The auPR is shown in parentheses within the figure legends

species and chromatin feature pairs. The top 5% accessibility- or methylation-reducing *k*-mers per species and chromatin feature were classified as 'high-effect' *k*-mers. We performed an all-by-all global alignment of the high-effect *k*-mers per species and chromatin feature using Biopython's pairwise aligner (Cock et al., 2009). Using the alignment distance matrix, we clustered these high-effect *k*-mers into 100 representative *k*-mers using *k*-medoids (Bauckhage, 2015). We took the 100 medoid *k*-mers for each species and chromatin feature pair and did another all-by-all global alignment to create another distance matrix. The embedded *k*-mers coordinates were created using the MDS function in scikit-learn's manifold package. High-effect *k*-mers were matched to JASPAR 2020 CORE *plantae* motifs using FIMO (Grant et al., 2011) and a *q*-value threshold of 0.05.

2.8 | Positional global importance analysis

Global importance analysis (GIA) (Koo et al., 2021) measures the average difference in model predictions from a sampled background set of sequences to the same set with the sequence embedded within them. We ran a positional GIA (pGIA) analysis for each species and chromatin feature pair by embedding the consensus motifs of the 530 JASPAR 2020 CORE *plantae* TFs in both orientations at each possible position within 1,000 generated 600-bp sequences. The 600-bp sequences were generated using a profile model, where bases were sampled at each position according to their relative frequency in the model's test set at that position. The GNU parallel (Tange, 2018) was used to speed up the pGIA analysis.

JASPAR motifs were ranked by their maximum global importance across all positions. The TF families and classes were obtained from the JASPAR API (v1).

2.9 | Manuscript

This manuscript was formatted with Manubot (Himmelstein et al., 2019).

3 | RESULTS

3.1 | Recurrent CNNs accurately model chromatin state across species

To train a successful chromatin state classifier, we needed to choose a window size that balanced genomic context with

resolution. We tested a few different model configurations and decided upon 600-bp windows because higher window sizes showed diminishing returns on performance on our validation set while decreasing our effective resolution (Supplemental Figure S1). We preprocessed the ATAC-seq and unmethylated peaks by taking the midpoint and symmetrically extending to half the window size in both directions to obtain our positive observations. Negatives were sampled from the rest of the genome. After preprocessing, we had 26,280 training regions per species (315,360 total) for the across-species accessibility models and 35,652 training regions per species (356,520 total) for the methylation models split evenly between classes.

As a baseline for comparison to previous, within-species, chromatin state CNN models as well as our across-species models, we trained within-species DanQ model configurations for each of the angiosperm species in our data. We also trained across-species model configurations each using a different species as a test set. Generally, we observed that a given across-species model has a comparable, if not superior, auPR to the within-species model (Figure 1, top middle and top right). Although the across-species accessibility model auPR and areas under the receiver-operating characteristic curve vary substantially (Figure 1, bottom middle and bottom right; Supplemental Figure S2), they are also within the range of those observed in the original DanQ and DeepSEA human models and superior to the bag-of-*k*-mers model within maize (Supplemental Figure S3). We also see that both within-species and across-species performance decreases as genome size increases (Supplemental Figure S4). When comparing the accessibility and methylation models, we see the same trends in performance for each species. To assess whether recurrent CNNs were a better architecture choice for across-species accessibility models over standard CNNs, we trained across-species accessibility configurations of the Basset, CharPlant, and DeeperDeepSEA architectures. We observed that DanQ was a superior architecture for across-species accessibility modeling for almost all hold-out species (Supplemental Figure S5).

To see if the models were more accurate in predicting accessible or unmethylated regions near or within genes, where these regions are known to be enriched, we looked at the precision-recall curves across different distance classes (genic, proximal, or distal). Observations were labeled as genic if more than half of the range overlapped with a gene annotation, as proximal if not genic and more than half of the range was within the proximal cutoff (2 kb), and as distal if neither genic nor proximal. We see that the across-species models for both chromatin features perform the worst on distal regions but show contrasting results on the genic and proximal regions (Figure 1, bottom left). This could be driven by the imbalanced distribution of regions between the distance classes, with accessible regions biased toward the proximal class and unmethylated regions toward the genic class (Supplemental Figure S6). In particular, barley has proportionally many more distal accessible and unmethylated regions, which could explain the lower overall performance. The across-species accessibility models are very precise when calling inaccessible chromatin, with most of the errors being false positives, particularly in distal regions (Supplemental Figure S7). We see a much different result in the methylation model, which shows only a slight bias toward false positives.

To control for potential *trans*-driven transposon silencing, we tested a two-step model that takes the predictions of the a2z model and then masks them with zeros if they overlap annotated transposons in maize. We see that these two-step repeat-masked models do much better (ΔauPR 0.15 for accessibility and 0.07 for methylation) than the naive models (Supplemental Figure S8), suggesting a relatively straightforward way to reduce false positives in larger plant genomes with more transposon-derived sequence.

Finally, we wanted to assess how far out in evolutionary time the angiosperm model could work. We ran the model against ATAC-seq data from yeast and a human GM12878 cell line (Schep et al., 2015). We see the plant-trained model has some ability (Supplemental Figure S9) to predict chromatin accessibility in yeast (auPR 0.21), if not human cell lines (auPR 0.02).

3.2 | Leaf-trained models struggle to predict cell-type-specific ACRs

Knowing the a2z models are capable of working across species, we then asked how well the leaf-trained accessibility models could work across cell types. We used scATAC-seq data from six maize organs (Marand et al., 2021) as a multiple cell type test set for our single-tissue model. Using a model trained on every species with ATAC-seq data except maize, we predicted the accessibility of each scATAC peak as well as negatives sampled from the rest of the genome. Looking at the area under the threshold-recall curve, we see that the model does better on peaks that are accessible across many cell types, with a sharp decrease in peaks only accessible in five or fewer cell types, which are likely to be a mix of false positives and highly cell-type-specific peaks (Figure 2, left). The model does best on peaks that are generally open across many cell types, which comprise the largest portion of the training data (Supplemental Figure S10). This is clearly shown when looking at the overall precision-recall curves in the best (guard cell) and worst (trichoblast) cell types as well as a union of all cell types. There is not a substantial difference between the three (Figure 2, right).

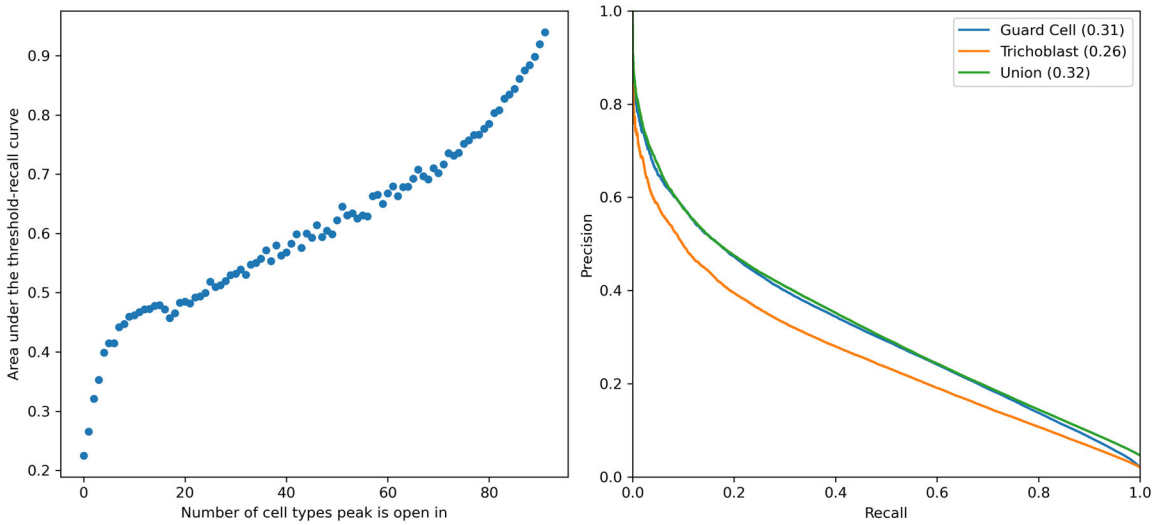

FIGURE 2 Across-cell type performance of the maize accessibility model. The left plot shows the area under the threshold-recall curve for each set of peaks grouped by the number of cell types they are accessible in. The right plot shows the precision-recall curves for peaks accessible in the guard cell (best) and trichoblast (worst) cell types as well as peaks open in any cell type (union). The area under each curve is shown in parentheses in the figure legend

3.3 | Interpretation methods reveal important conserved and species-specific motifs

Although chromatin state models that work across angiosperms are a useful tool, we may be able to gain new insights into chromatin biology by dissecting what motifs and higher-order motif patterns the model is learning to use to separate accessible from inaccessible chromatin or unmethylated from methylated regions. We started with the attribution tool TF-MoDISco to identify important motifs in the maize and *Arabidopsis* test sets using their respective held-out models. Although TF-MoDISco qualitatively identified many important motifs (Supplemental Figure S11), most of them ranked similarly by attribution score and therefore could not be quantitatively compared in terms of effect size or importance relative to each other.

To obtain better estimates of sequence-effect size, we developed a method that masks sliding windows across a set of sequences and evaluates the change in the model prediction, which we refer to as the *k*-mer occlusion method. Using a *k*-mer size of 10 bp, representing a common estimate of core binding site length, we ran a *k*-mer occlusion to get effect sizes for each *k*-mer in the test set, binned *k*-mers into 'high-effect' and 'null-effect', and then scanned them for matches to JASPAR 2020 CORE *plantae* (Fornes et al., 2019) binding motifs. For our accessibility models, we see that approximately 20–40% of high-effect *k*-mers match with JASPAR motifs, whereas our methylation models generally seem to have poor matching between JASPAR motifs and high-effect *k*-mers (Supplemental Figure S12). To look at how similar the high-effect *k*-mers were between chromatin features and species, we used *k*-medoids to get a subset of representative *k*-mers and then visualized the distances between them using multidimensional scaling. Surprisingly, the high-effect *k*-mers across species and chromatin features cluster together, with slight separation between methylation and accessibility (Figure 3, left). However, there is no separation between species (Supplemental Figure S13) nor monocots and dicots (Figure 3, middle, right) for either chromatin feature.

To understand which known biological motifs were being recognized as important to the model, we used a recently developed model interpretation method known as GIA (Koo et al., 2021). First, we ranked JASPAR motifs by their maximum global importance across all positions for each model (Table 1) and saw both species-specific and common TFs across the models. One of the most remarkable observations is that the top 10 motifs in the *Arabidopsis* model are all from the TCP family. The maize accessibility model also ranked TCP motifs in the top 10 but behind Dof-type motifs. The *Arabidopsis* and maize methylation models rank the same two motifs at the top and share mostly the same families between the rest. Next, we looked at the positional effects of the top three TFs across *Arabidopsis* accessibility (Figure 4, top left) and methylation (bottom left) as well as maize accessibility (top right) and methylation (bottom right). The most striking feature is the sawtooth pattern seen across both species and chromatin feature models; however, the cause of this pattern is unclear. The *Arabidopsis* accessibility model shows a clear bias toward the center of the accessible regions for the top three TFs, whereas the other models are not as consistent.

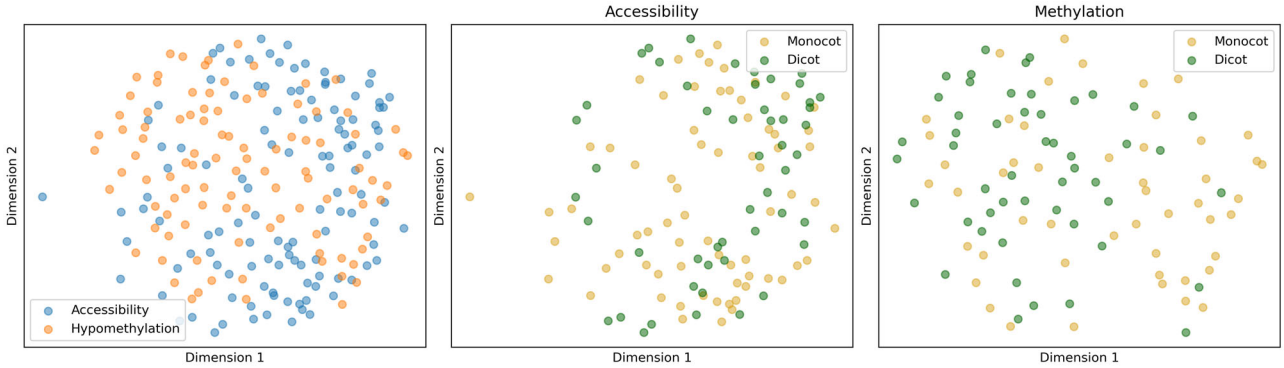

FIGURE 3 Multidimensional scaling of the high-effect medoid *k*-mer distance matrix across all species and chromatin feature model combinations. Each point is a high-effect *k*-mer in a given species and chromatin feature combination

TABLE 1 Top 10 JASPAR motifs for four pan-angiosperm models ranked by max global importance across all possible embedding positions. Transcription factor (TF) family or class (if family was not available) according to JASPAR is shown in parentheses under each TF

	Accessibility		Methylation	
Rank	***Arabidopsis thaliana***	***Zea mays***	***Arabidopsis thaliana***	***Zea mays***
1	TCP1 (TCP)	AT5G66940 (Dof-type)	ERF104 (AP2/ERF)	ERF104 (AP2/ERF)
2	TCP14 (TCP)	OBP3 (Dof-type)	AT4G18450 (AP2/ERF)	AT4G18450 (AP2/ERF)
3	At1g72010 (TCP)	AT1G69570 (Dof-type)	ERF9 (AP2/ERF)	RAP211 (AP2/ERF)
4	TCP21 (TCP)	OBP1 (Dof-type)	BPC5 (BBR-BPC)	BPC5 (BBR-BPC)
5	TCP19 (TCP)	AT2G28810 (Dof-type)	ERF2 (AP2/ERF)	ERF9 (AP2/ERF)
6	TCP7 (TCP)	AT5G02460 (Dof-type)	LEP (AP2/ERF)	ESE1 (AP2/ERF)
7	At2g45680 (TCP)	TCP1 (TCP)	BPC1 (BBR-BPC)	AT5G66940 (Dof-type)
8	TCP20 (TCP)	At1g72010 (TCP)	ESE1 (AP2/ERF)	BPC1 (BBR-BPC)
9	OJ1581_H09.2 (TCP)	TCP21 (TCP)	ERF10 (AP2/ERF)	ERF2 (AP2/ERF)
10	TCP2 (TCP)	BPC5 (BBR-BPC)	BPC6 (BBR-BPC)	LEP (AP2/ERF)

4 | DISCUSSION

We have shown that recurrent CNNs, DanQ in particular, are an effective architecture on which to base across-species sequence to chromatin state models. By incorporating sequence data from multiple species, we not only increase the size of our training data set, a critical factor for deep-learning models, but also reduce the amount of confounding neutral variation around functional motifs. Being able to predict chromatin state across species also opens the door for studies of regulatory regions in additional angiosperm species with only genomic sequence data. Beyond angiosperms, the a2z model's predictive ability in yeast suggests it is capable of working effectively across wide evolutionary timescales. Unsurprisingly, we noticed that the performance across different peak classes relates to their relative abundance in the training set. Future work looking at ways to balance or weight observations in rarer peak classes would likely improve the generalizability of the models. This is particularly important for working toward better across-tissue chromatin state models, where the tissue-specific peaks are usually the minority in any given data set, as well as with larger genomes, where distal peaks are more prevalent.

Further, most sequence-based model architectures, including DanQ, only take in *cis* sequence, which is known (Xiao & Wagner, 2015) to account for only a portion of the variation in local chromatin state. Model architectures that can effectively incorporate *trans* factors, such as chromatin-remodeling TFs on neighboring regulatory elements (Taberl et al., 2011) or small RNA silencing (Ito, 2011), will likely surpass current methods but their across-species applicability remains an open question. By far, the most prevalent error of the accessibility models in particular is calling false positives, which may be due to lack of *trans* information. A portion of these false positives may also be under called ATAC-seq peaks that are open in very specific cell types, because the peaks from Lu et al. (2019) were called with relatively conservative thresholds.

Interpreting deep-learning models remains a challenge but is an especially critical one to overcome. Here we use occlusion- and perturbation-based methods instead of gradient-based approaches like TF-MoDISco and saliency

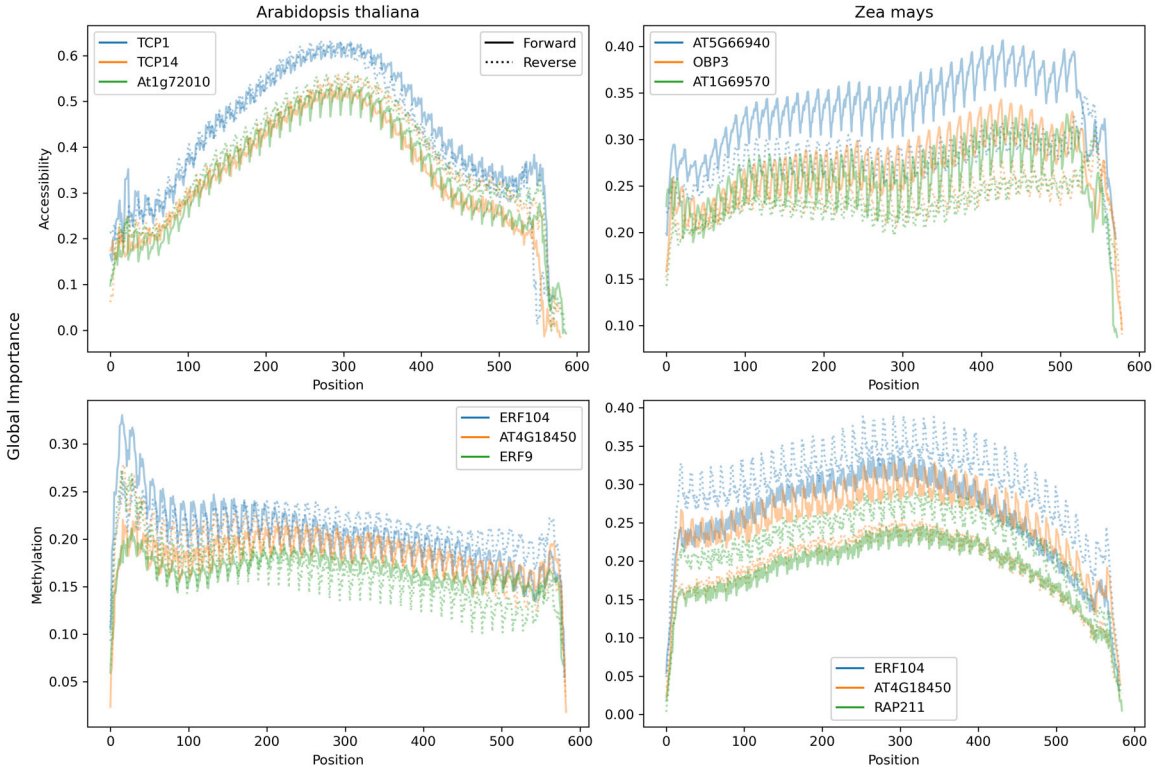

FIGURE 4 Positional global importance analysis plots for *Arabidopsis* (left) and maize (right) accessibility (top) and methylation (bottom). The solid and dotted lines represent the importance scores for the positive and negative strand, respectively. Only the top three JASPAR motifs ranked by the maximum global importance across the sequence were plotted

maps to trade longer computational times for reduced noise (Kim et al., 2019) in effect estimates. Particularly because eukaryotic TF binding sites are known to be degenerate (Stewart et al., 2012), point mutation effect sizes in regulatory sequences are likely to be small and harder to estimate accurately with our limited data. The lack of separation between clades and species in the multidimensional scaling plots for each chromatin feature is not too surprising. The across-species models must learn to prioritize motifs that are generalizable across species and so potential species- or clade-specific motifs are ignored. The sawtooth pattern, which is stronger in some TFs than in others, could be a manifestation of the model learning a helical face bias for specific TF binding. Further controls will be necessary to investigate that hypothesis, as the pattern may also be an artifact of the maximum pooling or long short-term memory layers. Not all of the pGIA results agree with current theory. For example, some of the motifs have a noticeable strand bias, but enhancers are known to operate in an orientation-independent (Arnold et al., 2013) manner. Given some of them are relatively simple motifs, it is possible that these matches are surrogates for important nonbinding motifs. We chose to rank JASPAR motifs by maximum global importance across the sequence as a rough estimate for importance to regulating the given chromatin feature state, though other methods of ranking could be preferable depending on the use case. Because positive observations are created by extending from the midpoint, the effect of TFs that bind to the center of accessible or unmethylated regions will be easier to estimate because they are more aligned across the test set sequences. In contrast, TFs that bind to the edges of accessible or unmethylated regions are not aligned because the lengths of the true, unextended ATAC-seq peaks are not equal.

The top 10 JASPAR motifs are very different between the features but remarkably similar between the species within each feature. Of the two known (Jin et al., 2021; Lai et al., 2021; Tao et al., 2017) plant pioneer TFs (LEC1 and LEAFY), only LEAFY is present in JASPAR but does not show up in the top 10 motifs for any of the models. This is not unexpected, as it is a floral TF, and our models are trained on leaf accessible regions. The strong presence of the TCP family in the highly ranked accessibility TFs is promising, because they are known (Yang et al., 2020) to be involved in chromatin remodeling. What role the Dof-type TFs play in accessibility is still unclear because of the wide variety of roles they play (Noguero et al., 2013). The shared top two motifs between the methylation models have evidence that they are involved in plant pathogen response (Bethke et al., 2009; Ou et al., 2011). Knowing that plant immunity genes are among the most variable (Van de Weyer et al., 2019), it would be interesting to see

if these unmethylated regions are harboring a large library of rapidly inducible resistance genes that remain mostly inaccessible until needed. With the high similarity in binding motifs by definition within families, it is quite possible that some highly ranked TFs are false positives because of association with the few causal TFs in the same family. Although it is useful to use JASPAR motifs as specific testable hypotheses, there are only 530 motifs in the database, and with the lowest estimates of angiosperm TF gene count starting at ∼1,500 (Lang et al., 2010), critical TFs may still be missing.

Moving forward, more focus is necessary on collecting high-quality accessible regions across a variety of cell types to train models that are capable of simultaneously generalizing across both tissues and species. Lessons learned from successful across-species and across-tissue chromatin state models could be applied to improve more task-specific sequence models such as enhancer prediction (Min et al., 2017) or promoter-enhancer contact prediction (Li et al., 2019). With the release of highly accurate protein-folding models, such as AlphaFold2 (Jumper et al., 2021), the missing species-specific TF binding motifs in any genome may finally be feasible to estimate using simulated DNA docking approaches. Now that many deep-learning-based approaches borrowed from other fields (Avsec, Agarwal, et al., 2021; Tu et al., 2020) have been shown to be successful in mapping genomic sequence to a variety of cellular phenotypes, better interpretation methods to assess what these black-box models are learning will be important to optimize toward more biologically relevant architectures.

ACKNOWLEDGMENTS

This work was funded by an NSF Graduate Research Fellowship (DGE-1650441) and the USDA–ARS to T.W., an NSF Postdoctoral Fellowship in Biology (DBI-1905869) to A.P.M., an Australian Research Council (ARC) Discovery Early Career Award (DE200101748) to P.A.C., the NSF IOS-1934384 to N.M.S., and the USDA–ARS to E.S.B. The Texas Advanced Computing Center supported a portion of the compute time for the analyses with their Frontera system. Peter Koo contributed helpful comments during the analyses.

AUTHOR CONTRIBUTIONS

Travis Wrightsman: Conceptualization; Data curation; Formal analysis; Funding acquisition; Investigation; Methodology; Project administration; Resources; Software; Validation; Visualization; Writing – original draft; Writing – review & editing. Alexandre P. Marand: Formal analysis; Methodology; Resources; Supervision; Writing – review & editing. Peter A. Crisp: Formal analysis; Resources; Writing – review & editing. Nathan M. Springer: Methodology, Resources, Writing – review & editing. Edward S. Buckler: Conceptualization, Funding acquisition, Methodology, Supervision, Writing – review & editing.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Travis Wrightsman https://orcid.org/0000-0002-0904-6473
Alexandre P. Marand https://orcid.org/0000-0001-9100-8320
Peter A. Crisp https://orcid.org/0000-0002-3655-0130
Nathan M. Springer https://orcid.org/0000-0002-7301-4759
Edward S. Buckler https://orcid.org/0000-0002-3100-371X

REFERENCES

Agarwal, V., & Shendure, J. (2020). Predicting mRNA abundance directly from genomic sequence using deep convolutional neural networks. *Cell Reports*, *31*, 107663. https://doi.org/10.1016/j.celrep.2020.107663

Anderson, J. T., Willis, J. H., & Mitchell-Olds, T. (2011). Evolutionary genetics of plant adaptation. *Trends in Genetics*, *27*, 258–266. https://doi.org/10.1016/j.tig.2011.04.001

Arnold, C. D., Gerlach, D., Stelzer, C., Boryń, Ł. M., Rath, M., & Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, *339*, 1074–1077. https://doi.org/10.1126/science.1232542

Avsec, Ž., Agarwal, V., Visentin, D., Ledsam, J. R., Grabska-Barwinska, A., Taylor, K. R., Assael, Y., Jumper, J., Kohli, P., & Kelley, D. R. (2021). Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, *18*, 1196–1203. https://doi.org/10.1038/s41592-021-01252-x

Avsec, Ž., Weilert, M., Shrikumar, A., Krueger, S., Alexandari, A., Dalal, K., Fropf, R., McAnany, C., Gagneur, J., Kundaje, A., & Zeitlinger, J. (2021). Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, *53*, 354–366. https://doi.org/10.1038/s41588-021-00782-6

Bauckhage, C. (2015). *NumPy/SciPy recipes for data science: k-Medoids clustering*. https://doi.org/10.13140/2.1.4453.2009

Bethke, G., Unthan, T., Uhrig, J. F., Pöschl, Y., Gust, A. A., Scheel, D., & Lee, J. (2009). Flg22 regulates the release of an ethylene response factor substrate from MAP kinase 6 in *Arabidopsis thaliana*via ethylene signaling. *Proceedings of the National Academy of Sciences*, *106*, 8067–8072. https://doi.org/10.1073/pnas.0810206106

Broad Institute. (2019). Picard toolkit. GitHub. https://github.com/broadinstitute/picard

Chen, L., Fish, A. E., & Capra, J. A. (2018). Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties. *PLoS Computational Biology*, *14*, e1006484. https://doi.org/10.1371/journal.pcbi.1006484

Chollet, F. (2015). Keras. GitHub. https://github.com/keras-team/keras

Church, D. M., Schneider, V. A., Graves, T., Auger, K., Cunningham, F., Bouk, N., Chen, H.-C., Agarwala, R., McLaren, W. M., Ritchie, G. R. S., Albracht, D., Kremitzki, M., Rock, S., Kotkiewicz, H., Kremitzki, C., Wollam, A., Trani, L., Fulton, L., Fulton, R., … Hubbard, T. (2011). Modernizing reference genome assemblies. *PLoS Biology*, *9*, e1001091. https://doi.org/10.1371/journal.pbio.1001091

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B.,

& de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*, 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

conda-forge Community. (2015). *The conda-forge Project: Community-based software distribution built on the conda package format and ecosystem*. Zenodo. https://doi.org/10.5281/zenodo.4774216

Crisp, P. A., Marand, A. P., Noshay, J. M., Zhou, P., Lu, Z., Schmitz, R. J., & Springer, N. M. (2020). Stable unmethylated DNA demarcates expressed genes and their *cis*-regulatory space in plant genomes. *Proceedings of the National Academy of Sciences*, *117*, 23991–24000. https://doi.org/10.1073/pnas.2010250117

Dale, R. K., Pedersen, B. S., & Quinlan, A. R. (2011). Pybedtools: A flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics*, *27*, 3423–3424. https://doi.org/10.1093/bioinformatics/btr539

Danecek, P., Bonfield, J. K., Liddle, J., Marshall, J., Ohan, V., Pollard, M. O., Whitwham, A., Keane, T., McCarthy, S. A., Davies, R. M., & Li, H. (2021). Twelve years of SAMtools and BCFtools. *GigaScience*, *10*, giab008. https://doi.org/10.1093/gigascience/giab008

Doğan, E. S., & Liu, C. (2018). Three-dimensional chromatin packing and positioning of plant genomes. *Nature Plants*, *4*, 521–529. https://doi.org/10.1038/s41477-018-0199-5

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., Santana-Garcia, W., Tan, G., Chèneby, J., Ballester, B., Parcy, F., Sandelin, A., Lenhard, B., Wasserman, W. W., & Mathelier, A. (2019). JASPAR 2020: Update of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, *48*, D87–D92. https://doi.org/10.1093/nar/gkz1001

The French–Italian Public Consortium for Grapevine Genome Characterization. (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*, *449*, 463–467. https://doi.org/10.1038/nature06148

Grant, C. E., Bailey, T. L., & Noble, W. S. (2011). FIMO: Scanning for occurrences of a given motif. *Bioinformatics*, *27*, 1017–1018. https://doi.org/10.1093/bioinformatics/btr064

Grüning, B., Dale, R., Sjödin, A., Chapman, B. A., Rowe, J., Tomkins-Tinch, C. H., Valieris, R., & Köster, J. (2018). Bioconda: Sustainable and comprehensive software distribution for the life sciences. *Nature Methods*, *15*, 475–476. https://doi.org/10.1038/s41592-018-0046-7

Harkess, A., Zhou, J., Xu, C., Bowers, J. E., Van der Hulst, R., Ayyampalayam, S., Mercati, F., Riccardi, P., McKain, M. R., Kakrana, A., Tang, H., Ray, J., Groenendijk, J., Arikit, S., Mathioni, S. M., Nakano, M., Shan, H., Telgmann-Rauber, A., Kanno, A., … Chen, G. (2017). The asparagus genome sheds light on the origin and evolution of a young Y chromosome. *Nature Communications*, *8*, 1279. https://doi.org/10.1038/s41467-017-01064-8

Himmelstein, D. S., Rubinetti, V., Slochower, D. R., Hu, D., Malladi, V. S., Greene, C. S., & Gitter, A. (2019). Open collaborative writing with Manubot. *PLoS Computational Biology*, *15*, e1007128. https://doi.org/10.1371/journal.pcbi.1007128

Hirsch, C. D., & Springer, N. M. (2017). Transposable element influences on gene expression in plants. *Biochimica Et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1860*, 157–165. https://doi.org/10.1016/j.bbagrm.2016.05.010

Hunter, J. D. (2007). Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering*, *9*, 90–95. https://doi.org/10.1109/mcse.2007.55

The International Brachypodium Initiative. (2010). Genome sequencing and analysis of the model grass *Brachypodium distachyon*. *Nature*, *463*, 763–768. https://doi.org/10.1038/nature08747

Ito, H. (2011). Small RNAs and transposon silencing in plants. *Development, Growth & Differentiation*, *54*, 100–107. https://doi.org/10.1111/j.1440-169x.2011.01309.x

Jiao, Y., Peluso, P., Shi, J., Liang, T., Stitzer, M. C., Wang, B., Campbell, M. S., Stein, J. C., Wei, X., Chin, C.-S., Guill, K., Regulski, M., Kumari, S., Olson, A., Gent, J., Schneider, K. L., Wolfgruber, T. K., May, M. R., Springer, N. M., … Ware, D. (2017). Improved maize reference genome with single-molecule technologies. *Nature*, *546*, 524–527. https://doi.org/10.1038/nature22971

Jin, R., Klasfeld, S., Zhu, Y., Fernandez Garcia, M., Xiao, J., Han, S.-K., Konkol, A., & Wagner, D. (2021). LEAFY is a pioneer transcription factor and licenses cell reprogramming to floral fate. *Nature Communications*, *12*, 626. https://doi.org/10.1038/s41467-020-20883-w

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, *596*, 583–589. https://doi.org/10.1038/s41586-021-03819-2

Jun, G., Wing, M. K., Abecasis, G. R., & Kang, H. M. (2015). An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data. *Genome Research*, *25*, 918–925. https://doi.org/10.1101/gr.176552.114

Kelley, D. R. (2020). Cross-species regulatory sequence activity prediction. *PLoS Computational Biology*, *16*, e1008050. https://doi.org/10.1371/journal.pcbi.1008050

Kim, B., Seo, J., Jeon, S., Koo, J., Choe, J., & Jeon, T. (2019). Why are saliency maps noisy? Cause of and solution to noisy saliency maps. *2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*. https://doi.org/10.1109/iccvw.2019.00510

Koo, P. K., Majdandzic, A., Ploenzke, M., Anand, P., & Paul, S. B. (2021). Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks. *PLoS Computational Biology*, *17*, e1008925. https://doi.org/10.1371/journal.pcbi.1008925

Krützfeldt, L.-M., Schubach, M., & Kircher, M. (2020). The impact of different negative training data on regulatory sequence predictions. *PLoS One*, *15*, e0237412. https://doi.org/10.1371/journal.pone.0237412

Lai, X., Blanc-Mathieu, R., GrandVuillemin, L., Huang, Y., Stigliani, A., Lucas, J., Thévenon, E., Loue-Manifel, J., Turchi, L., Daher, H., Brun-Hernandez, E., Vachon, G., Latrasse, D., Benhamed, M., Dumas, R., Zubieta, C., & Parcy, F. (2021). The LEAFY floral regulator displays pioneer transcription factor properties. *Molecular Plant*, *14*, 829–837. https://doi.org/10.1016/j.molp.2021.03.004

Lamesch, P., Berardini, T. Z., Li, D., Swarbreck, D., Wilks, C., Sasidharan, R., Muller, R., Dreher, K., Alexander, D. L., Garcia-Hernandez, M., Karthikeyan, A. S., Lee, C. H., Nelson, W. D., Ploetz, L., Singh, S., Wensel, A., & Huala, E. (2011). The Arabidopsis Information Resource (TAIR): Improved gene annotation and new tools. *Nucleic Acids Research*, *40*, D1202–D1210. https://doi.org/10.1093/nar/gkr1090

Lang, D., Weiche, B., Timmerhaus, G., Richardt, S., Riaño-Pachón, D. M., Corrêa, L. G. G., Reski, R., Mueller-Roeber, B., & Rensing, S. A. (2010). Genome-wide phylogenetic comparative analysis of

plant transcriptional regulation: A timeline of loss, gain, expansion, and correlation with complexity. *Genome Biology and Evolution*, *2*, 488–503. https://doi.org/10.1093/gbe/evq032

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The SEQUENCE ALIGNMENT/Map format and SAMtools. *Bioinformatics*, *25*, 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Li, W., Wong, W. H., & Jiang, R. (2019). DeepTACT: Predicting 3D chromatin contacts via bootstrapping deep learning. *Nucleic Acids Research*, *47*, e60–e60. https://doi.org/10.1093/nar/gkz167

Liu, Q., Xia, F., Yin, Q., & Jiang, R. (2017). Chromatin accessibility prediction via a hybrid deep convolutional neural network. *Bioinformatics*, *34*, 732–738. https://doi.org/10.1093/bioinformatics/btx679

Lu, Z., Marand, A. P., Ricci, W. A., Ethridge, C. L., Zhang, X., & Schmitz, R. J. (2019). The prevalence, evolution and chromatin signatures of plant regulatory elements. *Nature Plants*, *5*, 1250–1259. https://doi.org/10.1038/s41477-019-0548-z

Mamidi, S., Healey, A., Huang, P., Grimwood, J., Jenkins, J., Barry, K., Sreedasyam, A., Shu, S., Lovell, J. T., Feldman, M., Wu, J., Yu, Y., Chen, C., Johnson, J., Sakakibara, H., Kiba, T., Sakurai, T., Tavares, R., Nusinow, D. A., … Kellogg, E. A. (2020). A genome resource for green millet *Setaria viridis* enables discovery of agronomically valuable loci. *Nature Biotechnology*, *38*, 1203–1210. https://doi.org/10.1038/s41587-020-0681-2

Marand, A. P., Chen, Z., Gallavotti, A., & Schmitz, R. J. (2021). A *cis*-regulatory atlas in maize at single-cell resolution. *Cell*, *184*, 3041–3055.e21. https://doi.org/10.1016/j.cell.2021.04.014

Marand, A. P., Zhang, T., Zhu, B., & Jiang, J. (2017). Towards genome-wide prediction and characterization of enhancers in plants. *Biochimica Et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, *1860*, 131–139. https://doi.org/10.1016/j.bbagrm.2016.06.006

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*, 10. https://doi.org/10.14806/ej.17.1.200

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., Radchuk, V., Dockter, C., Hedley, P. E., Russell, J., Bayer, M., Ramsay, L., Liu, H., Haberer, G., Zhang, X.-Q., Zhang, Q., Barrero, R. A., Li, L., Taudien, S., … Stein, N. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature*, *544*, 427–433. https://doi.org/10.1038/nature22043

McCormick, R. F., Truong, S. K., Sreedasyam, A., Jenkins, J., Shu, S., Sims, D., Kennedy, M., Amirebrahimi, M., Weers, B. D., McKinley, B., Mattison, A., Morishige, D. T., Grimwood, J., Schmutz, J., & Mullet, J. E. (2017). The *Sorghum bicolor* reference genome: Improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization. *The Plant Journal*, *93*, 338–354. https://doi.org/10.1111/tpj.13781

Mewes, H. W., Albermann, K., Bähr, M., Frishman, D., Gleissner, A., Hani, J., Heumann, K., Kleine, K., Maierl, A., Oliver, S. G., Pfeiffer, F., & Zollner, A. (1997). Erratum: Overview of the yeast genome. *Nature*, *387*, 737–737. https://doi.org/10.1038/42755

Min, X., Zeng, W., Chen, S., Chen, N., Chen, T., & Jiang, R. (2017). Predicting enhancers with deep convolutional neural networks. *BMC Bioinformatics*, *18*, 478. https://doi.org/10.1186/s12859-017-1878-3

Noguero, M., Atif, R. M., Ochatt, S., & Thompson, R. D. (2013). The role of the DNA-binding One Zinc Finger (DOF) transcription factor family in plants. *Plant Science*, *209*, 32–45. https://doi.org/10.1016/j.plantsci.2013.03.016

Ou, B., Yin, K.-Q., Liu, S.-N., Yang, Y., Gu, T., Wing Hui, J. M., Zhang, L., Miao, J., Kondou, Y., Matsui, M., Gu, H.-Y., & Qu, L.-J. (2011). A high-throughput screening system for *Arabidopsis* transcription factors and its application to Med25-dependent transcriptional regulation. *Molecular Plant*, *4*, 546–555. https://doi.org/10.1093/mp/ssr002

Ouyang, S., Zhu, W., Hamilton, J., Lin, H., Campbell, M., Childs, K., Thibaud-Nissen, F., Malek, R. L., Lee, Y., Zheng, L., Orvis, J., Haas, B., Wortman, J., & Buell, C. R. (2007). The TIGR Rice Genome Annotation Resource: Improvements and new features. *Nucleic Acids Research*, *35*, D883–D887. https://doi.org/10.1093/nar/gkl976

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

Quang, D., & Xie, X. (2016). DanQ: A hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences. *Nucleic Acids Research*, *44*, e107–e107. https://doi.org/10.1093/nar/gkw226

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*, 841–842. https://doi.org/10.1093/bioinformatics/btq033

Schep, A. N., Buenrostro, J. D., Denny, S. K., Schwartz, K., Sherlock, G., & Greenleaf, W. J. (2015). Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Research*, *25*, 1757–1770. https://doi.org/10.1101/gr.192294.115

Schmutz, J., Cannon, S. B., Schlueter, J., Ma, J., Mitros, T., Nelson, W., Hyten, D. L., Song, Q., Thelen, J. J., Cheng, J., Xu, D., Hellsten, U., May, G. D., Yu, Y., Sakurai, T., Umezawa, T., Bhattacharyya, M. K., Sandhu, D., Valliyodan, B., … Jackson, S. A. (2010). Genome sequence of the palaeopolyploid soybean. *Nature*, *463*, 178–183. https://doi.org/10.1038/nature08670

Schmutz, J., McClean, P. E., Mamidi, S., Wu, G. A., Cannon, S. B., Grimwood, J., Jenkins, J., Shu, S., Song, Q., Chavarro, C., Torres-Torres, M., Geffroy, V., Moghaddam, S. M., Gao, D., Abernathy, B., Barry, K., Blair, M., Brick, M. A., Chovatia, M., … Jackson, S. A. (2014). A reference genome for common bean and genome-wide analysis of dual domestications. *Nature Genetics*, *46*, 707–713. https://doi.org/10.1038/ng.3008

Shen, Y., Chen, L.-L., & Gao, J. (2021). CharPlant: A de novo open chromatin region prediction tool for plant genomes. *Genomics, Proteomics & Bioinformatics*, *19*, 860–871. https://doi.org/10.1016/j.gpb.2020.06.021

Shirley, M. D., Ma, Z., Pedersen, B. S., & Wheelan, S. J. (2015). Efficient "pythonic" access to FASTA files using pyfaidx. *PeerJ PrePrints*, *3*, e970v1. https://doi.org/10.7287/peerj.preprints.970v1

Shlyueva, D., Stampfel, G., & Stark, A. (2014). Transcriptional enhancers: From properties to genome-wide predictions. *Nature Reviews Genetics*, *15*, 272–286. https://doi.org/10.1038/nrg3682

Shrikumar, A., Tian, K., Avsec, Ž., Shcherbina, A., Banerjee, A., Sharmin, M., Nair, S., & Kundaje, A. (2020). Technical note on transcription factor motif discovery from importance scores (TF-MoDISco) version 0.5.6.5. In *arXiv* (No. 1811.00416). *arXiv:1811.00416*, https://doi.org/10.48550/arXiv.1811.00416

Stewart, A. J., Hannenhalli, S., & Plotkin, J. B. (2012). Why transcription factor binding sites are ten nucleotides long. *Genetics*, *192*, 973–985. https://doi.org/10.1534/genetics.112.143370

Taberlay, P. C., Kelly, T. K., Liu, C.-C., You, J., De Carvalho, D. D., Miranda, T. B., Zhou, X. J., Liang, G., & Jones, P. A. (2011). Polycomb-repressed genes have permissive enhancers that initiate reprogramming. *Cell*, *147*, 1283–1294. https://doi.org/10.1016/j.cell.2011.10.040

Talevich, E., Invergo, B. M., Cock, P. J., & Chapman, B. A. (2012). Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython. *BMC Bioinformatics*, *13*, 209. https://doi.org/10.1186/1471-2105-13-209

Tange, O. (2018). *GNU Parallel 2018*. Zenodo. https://doi.org/10.5281/zenodo.1146014

Tao, Z., Shen, L., Gu, X., Wang, Y., Yu, H., & He, Y. (2017). Embryonic epigenetic reprogramming by a pioneer transcription factor in plants. *Nature*, *551*, 124–128. https://doi.org/10.1038/nature24300

TensorFlow Developers. (2022). *TensorFlow (Version v2.8.2)*. Zenodo. https://doi.org/10.5281/zenodo.4724125

Tu, X., Mejía-Guerra, M. K., Valdes Franco, J. A., Tzeng, D., Chu, P.-Y., Shen, W., Wei, Y., Dai, X., Li, P., Buckler, E. S., & Zhong, S. (2020). Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nature Communications*, *11*, 5089. https://doi.org/10.1038/s41467-020-18832-8

Tuskan, G. A., DiFazio, S., Jansson, S., Bohlmann, J., Grigoriev, I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A., Schein, J., Sterck, L., Aerts, A., Bhalerao, R. R., Bhalerao, R. P., Blaudez, D., Boerjan, W., Brun, A., Brunner, A., … Rokhsar, D. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). *Science*, *313*, 1596–1604. https://doi.org/10.1126/science.1128691

Van de Weyer, A.-L., Monteiro, F., Furzer, O. J., Nishimura, M. T., Cevik, V., Witek, K., Jones, J. D. G., Dangl, J. L., Weigel, D., & Bemm, F. (2019). A species-wide inventory of NLR genes and alleles in *Arabidopsis thaliana*. *Cell*, *178*, 1260–1272.e14. https://doi.org/10.1016/j.cell.2019.07.038

Washburn, J. D., Mejia-Guerra, M. K., Ramstein, G., Kremling, K. A., Valluru, R., Buckler, E. S., & Wang, H. (2019). Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence. *Proceedings of the National Academy of Sciences*, *116*, 5542–5549. https://doi.org/10.1073/pnas.1814551116

Weber, B., Zicola, J., Oka, R., & Stam, M. (2016). Plant enhancers: A call for discovery. *Trends in Plant Science*, *21*, 974–987. https://doi.org/10.1016/j.tplants.2016.07.013

Wei, Q., & Dunbrack, R. L. (2013). The role of balanced training and testing data sets for binary classifiers in bioinformatics. *PLoS One*, *8*, e67863. https://doi.org/10.1371/journal.pone.0067863

Wrightsman, T., Marand, A. P., Crisp, P. A., Springer, N. M., & Buckler, E. S. (2021). *Modeling chromatin state from sequence across angiosperms using recurrent convolutional neural networks*. Zenodo. https://doi.org/10.5281/zenodo.6699866

Xi, Y., & Li, W. (2009). BSMAP: Whole genome bisulfite sequence MAPping program. *BMC Bioinformatics*, *10*, 232. https://doi.org/10.1186/1471-2105-10-232

Xiao, J., & Wagner, D. (2015). Polycomb repression in the regulation of growth and development in Arabidopsis. *Current Opinion in Plant Biology*, *23*, 15–24. https://doi.org/10.1016/j.pbi.2014.10.003

Yamaguchi, N. (2021). LEAFY, a pioneer transcription factor in plants: A mini-review. *Frontiers in Plant Science*, *12*, 701406. https://doi.org/10.3389/fpls.2021.701406

Yang, R., Jarvis, D. E., Chen, H., Beilstein, M. A., Grimwood, J., Jenkins, J., Shu, S., Prochnik, S., Xin, M., Ma, C., Schmutz, J., Wing, R. A., Mitchell-Olds, T., Schumaker, K. S., & Wang, X. (2013). The reference genome of the halophytic plant *Eutrema salsugineum*. *Frontiers in Plant Science*, *4*, 46. https://doi.org/10.3389/fpls.2013.00046

Yang, X., Yan, J., Zhang, Z., Lin, T., Xin, T., Wang, B., Wang, S., Zhao, J., Zhang, Z., Lucas, W. J., Li, G., & Huang, S. (2020). Regulation of plant architecture by a new histone acetyltransferase targeting gene bodies. *Nature Plants*, *6*, 809–822. https://doi.org/10.1038/s41477-020-0715-2

Zhao, H., Tu, Z., Liu, Y., Zong, Z., Li, J., Liu, H., Xiong, F., Zhan, J., Hu, X., & Xie, W. (2021). PlantDeepSEA, a deep learning-based web service to predict the regulatory effects of genomic variants in plants. *Nucleic Acids Research*, *49*, W523–W529. https://doi.org/10.1093/nar/gkab383

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, *12*, 931–934. https://doi.org/10.1038/nmeth.3547

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Wrightsman, T., Marand, A. P., Crisp, P. A., Springer, N. M., & Buckler, E. S. (2022). Modeling chromatin state from sequence across angiosperms using recurrent convolutional neural networks. *The Plant Genome*, *15,* e20249. https://doi.org/10.1002/tpg2.20249