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

This manuscript was automatically generated on November 11, 2021.

Authors

- **Travis Wrightsman**
  0000-0002-0904-6473 · twrightsman
  Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA 14853 · Funded by NSF Graduate Research Fellowship (DGE-1650441); USDA-ARS · CRediT Roles: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing - original draft, Writing - review & editing

- **Alexandre P. Marand**
  0000-0001-9100-8320
  Department of Genetics, University of Georgia, Athens, GA, USA 30602 · Funded by NSF Postdoctoral Fellowship in Biology (DBI-1905869) · CRediT Roles: Formal Analysis, Methodology, Resources, Supervision, Writing - review & editing

- **Peter A. Crisp**
  0000-0002-3655-0130
  School of Agriculture and Food Sciences, University of Queensland, Brisbane, QLD 4072, Australia · Funded by Australian Research Council (ARC) Discovery Early Career Award (DE200101748) · CRediT Roles: Resources, Formal Analysis, Writing - review & editing

- **Nathan M. Springer**
  0000-0002-7301-4759
  Department of Plant and Microbial Biology, University of Minnesota, Saint Paul, MN, USA 55108 · Funded by NSF IOS-1934384 · CRediT Roles: Methodology, Resources, Writing - review & editing

- **Edward S. Buckler**
  0000-0002-3100-371X
  Section of Plant Breeding and Genetics, Cornell University, Ithaca, NY, USA 14853; Institute for Genomic Diversity, Cornell University, Ithaca, NY, USA 14853; Agricultural Research Service, United States Department of Agriculture, Ithaca, NY, USA 14853 · Funded by USDA-ARS · CRediT Roles: Conceptualization, Funding acquisition, Methodology, Supervision, Writing - review & editing

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 leaf ATAC-seq data from 12 angiosperm species to predict the chromatin accessibility 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 accessible chromatin regions 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 held out model on a multi-tissue scATAC 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 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 cross-species "a2z" model it is now feasible to predict the chromatin accessibility and methylation landscape of any angiosperm genome.

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 [1,2]. Current methods to assay chromatin accessibility are highly environment-specific and relatively expensive compared to 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 [3]. Regions that lack DNA methylation are known to be stable over developmental time and overlap significantly with ACRs in plants with larger genomes [4], 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 non-coding variant effects for use in contexts such as breeding. Plants also provide an excellent system to study the genetic basis of adaptation [5]. 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 mRNA abundance [6,7,8], chromatin state [9,10,11], and transcription factor (TF) binding [12] 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 utilizing 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 [13], pioneer factors [14], and chromatin looping mechanisms

[15]. Because of these differences, plant-specific chromatin accessibility models are likely to be necessary.

We know that transcription factor binding sites are strongly conserved across evolutionary time [16, 17] and highly enriched in ACRs [18]. Certain deep learning model architectures, such as convolutional neural networks (CNN), have already been shown effective for predicting chromatin accessibility within species by recognizing important motifs [9,10] and their spatial relationships [19]. Previous work [17,20] 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. DanQ [10] is a recurrent CNN that has already been shown to be able to more accurately predict a number of genomic labels, including chromatin accessibility and DNA methylation, in the human genome than standard CNNs like DeepSEA [9].

Here, we train DanQ to predict chromatin accessibility using leaf ATAC-seq data from 12 angiosperm species [13], 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 5 previously-published grasses [4]. Using a maize single-cell ATAC (scATAC) accessibility atlas [21], 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 important across angiosperms for predicting chromatin accessibility in leaves or methylation state. Our pan-angiosperm chromatin state models are an important stepping stone towards a better understanding of gene regulation and adaptation.

Results

Recurrent convolutional neural networks 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 basepair windows because higher window sizes showed diminishing returns on performance while decreasing our effective resolution (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 postive 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 cross-species accessibility models and 35,652 training regions per species (356,520 total) for the methylation models, split evenly between classes.

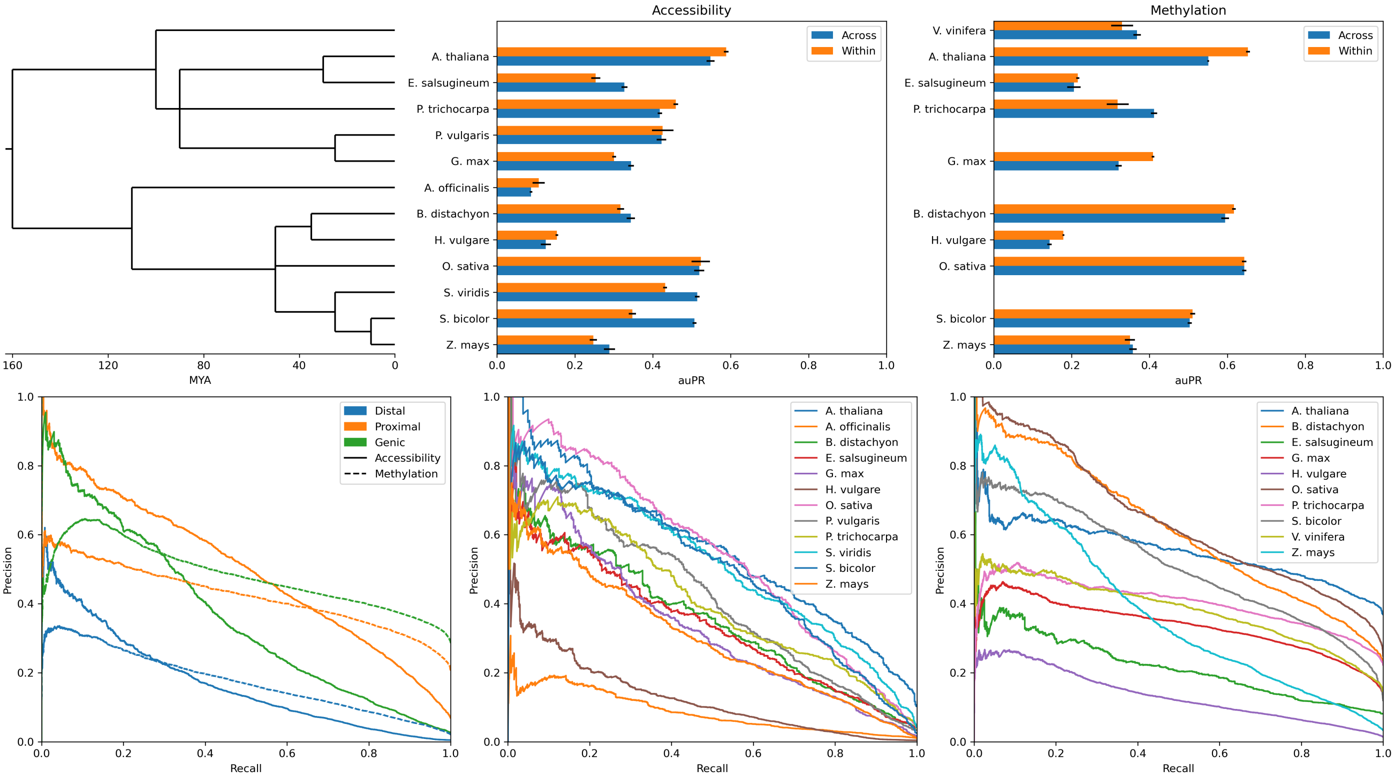

Figure 1: Performance of the cross-species chromatin state classifiers. The top middle and top right show the mean and standard error (due to variability in the stochastic model parameter initialization processes) of the auPR for the accessibility and methylation models, respectively, per species for 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.

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, area under the precision-recall curve (auPR) to the within-species model (Figure 1, top middle and top right). While auPR across species varies substantially, they are also within the range of those observed in the original DanQ and DeepSEA human models and superior to the bag-of-kmers model in *Zea mays* (Figure S3). We also see that both within-species and across-species performance decreases as genome size increases (Figure S4). When comparing the accessibility and hypomethylation models, we see the same trends in performance for each species.

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 (2kb), 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 towards the proximal class and unmethylated regions towards the genic class (Figure S5). In particular, *Hordeum vulgare* 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 (Figure S6). We see a much different result in the methylation model, which shows only a slight bias towards 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 *Z. mays*. 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 (Figure S7), 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 *Saccharomyces cerevisiae* and a *Homo sapiens* GM12878 cell line [22]. We see the plant-trained model has some ability (Figure S8) to predict chromatin accessibility in yeast (auPR 0.21), if not human cell-lines (auPR 0.02).

Leaf-trained models struggle to predict cell type-specific accessible chromatin regions

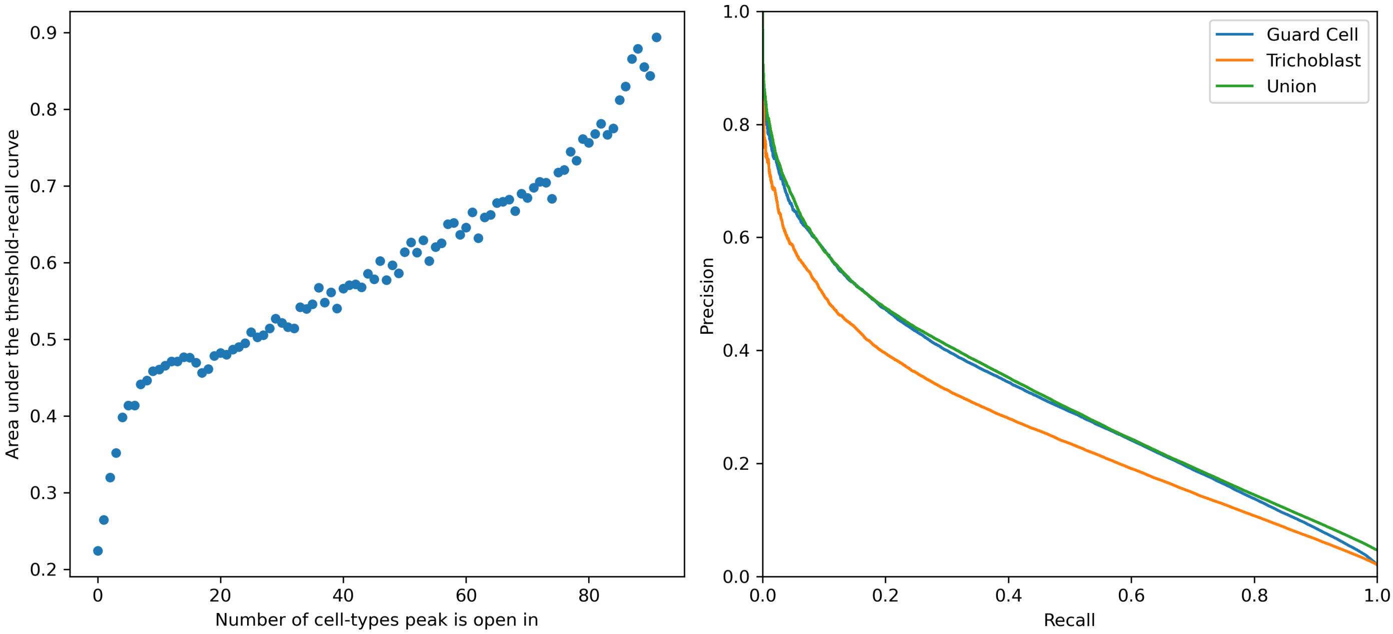

Figure 2: Cross-cell type performance of the *Zea mays* accessibility model. The left plot shows the area under the threshold-recall curve for each set of peaks grouped by the number of cell types they are accessible in. The right plot shows the precision-recall curves for peaks accessible in the guard cell (best) and trichoblast (worst) cell types, as well as peaks open in any cell type (union).

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 [21] as a multi-cell type test set for our single-tissue model. Using a model trained on every species with ATAC-seq data except *Z. mays*, 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 a largest portion of the training data (Figure S9). 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).

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 *Z. mays* and *Arabidopsis thaliana* test sets using their respective held-out models. While TF-MoDISco qualitatively identified many important motifs (Figure S10), 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.

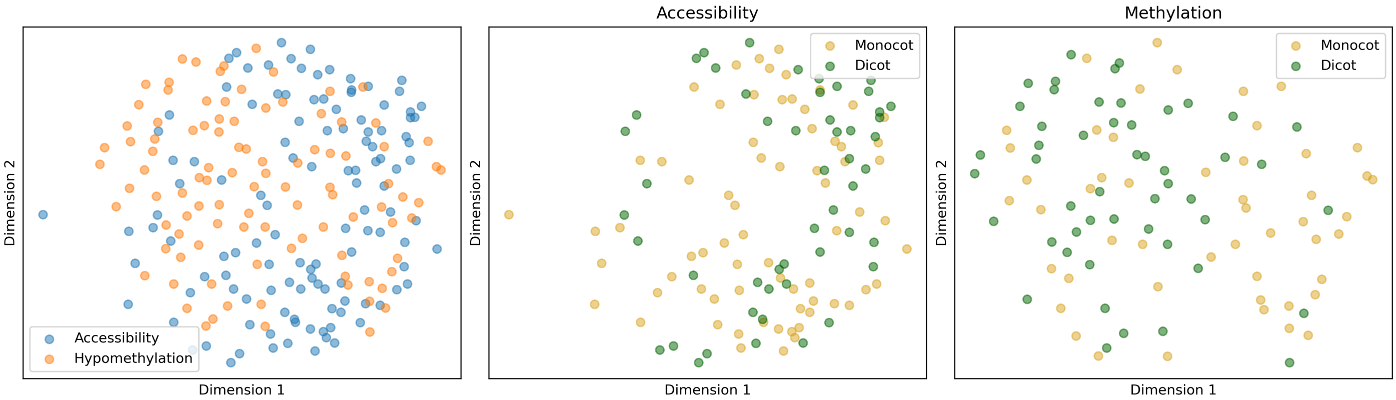

Figure 3: Multidimensional scaling of the high-effect medoid kmer distance matrix across all species and chromatin feature model combinations. Each point is a high-effect kmer in a given species and chromatin feature combination.

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 kmer occlusion method. Using a kmer size of 10bp, representing a common estimate of core binding site length, we ran a kmer occlusion to get effect sizes for each kmer in the test set, binned kmers into “high-effect” and “null-effect”, and then scanned them for matches to JASPAR 2020 CORE *plantae* [23] binding motifs. For our accessibility models, we see that about 20-40% of high-effect kmers match with JASPAR motifs while our methylation models generally seem to have poor matching between JASPAR motifs and high-effect kmers (Figure S12). To look at how similar the high-effect kmers were between chromatin features and species, we used k-medoids to get a subset of representative kmers and then visualized the distances between them using multidimensional scaling. Surprisingly, the high-effect kmers across species and chromatin features cluster together, with slight separation between methylation and accessibility (Figure 3, left). However, there is no separation between species (Figure S11) nor monocots and dicots (Figure 3, middle/right) for either chromatin feature.

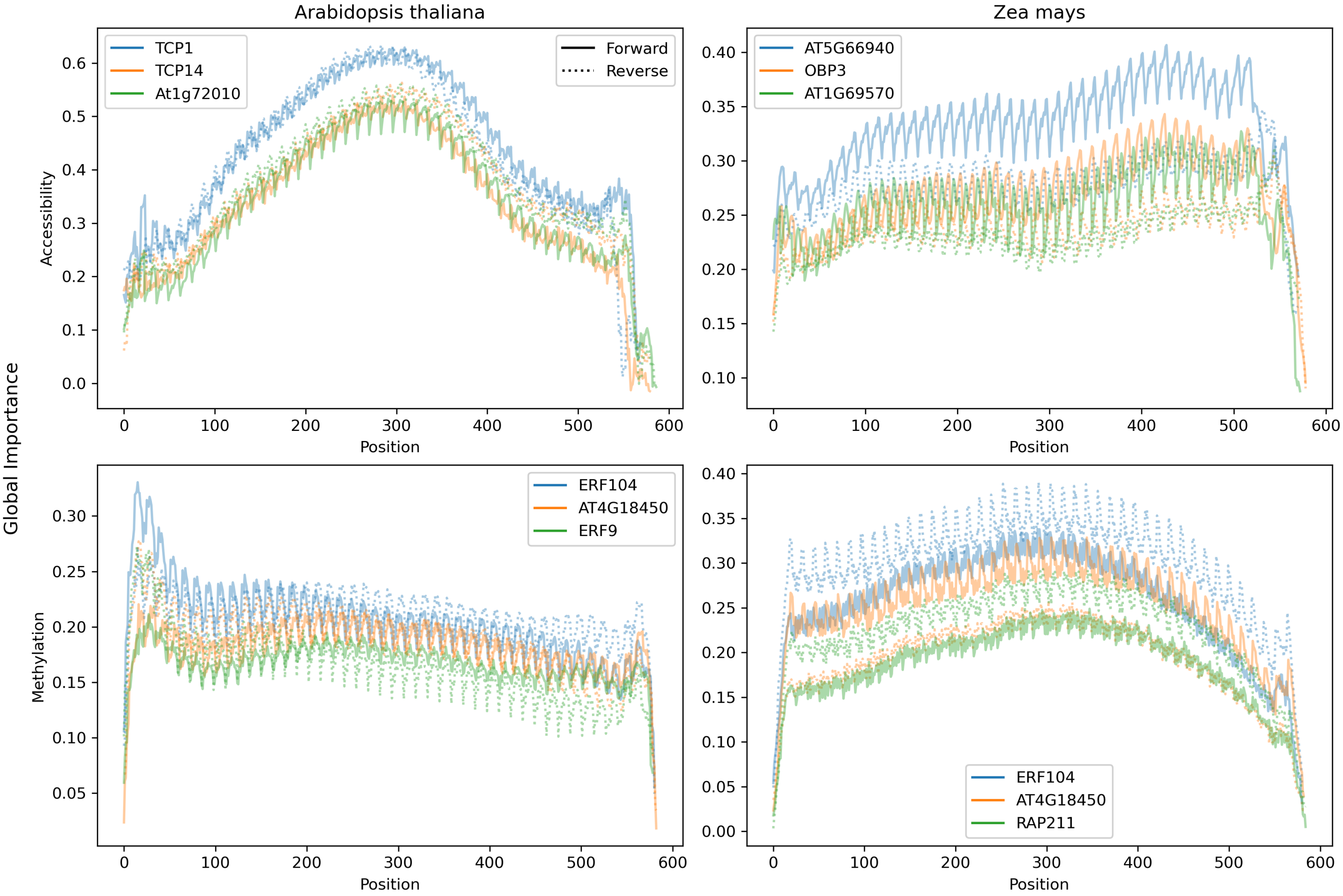

Figure 4: Positional Global Importance Analysis plots for *A. thaliana* (left) and *Z. mays* (right) accessibility (top) and methylation (bottom). The solid and dotted lines represent the importance scores for the positive and negative strand, respectively. Only the top three JASPAR motifs ranked by the maximum global importance across the sequence were plotted.

To understand which known biological motifs were being recognized as important to the model, we used a recently-developed model interpretation method known as Global Importance Analysis (GIA) [24]. First, we ranked JASPAR motifs by their max global importance across all positions for each model (Table 1) and see both species-specific and common TFs across the models. One of the most remarkable observations is that the top 10 motifs in the *A. thaliana* model are all from the TCP family. The *Z. mays* accessibility model also ranked TCP motifs in the top 10 but behind Dof-type motifs. The *A. thaliana* and *Z. mays* 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 *A. thaliana* accessibility (Figure 4, top left) and methylation (bottom left) as well as *Z. mays* 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 *A. thaliana* accessibility model shows a clear bias towards the center of the accessible regions for the top three TFs while the other models are not as consistent.

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

Rank	Accessibility A. thaliana	Accessibility Z. mays	Methylation A. thaliana	Methylation Z. mays
1	TCP1 (TCP)	AT5G66940 (Dof-type)	ERF104 (AP2/ERF)	ERF104 (AP2/ERF)

Rank	Accessibility A. thaliana	Accessibility Z. mays	Methylation A. thaliana	Methylation Z. mays
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)

Discussion

We have shown that recurrent CNNs, DanQ in particular, are an effective architecture on which to base cross-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 towards better cross-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 [25] 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 [26] or small RNA silencing [27], will likely surpass current methods but their cross-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 undercalled ATAC-Seq peaks that are open in very specific cell-types, since 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 maps to trade longer computational times for reduced noise [28] in effect estimates. Particularly since eukaryotic TF binding sites are known to be degenerate [29], SNP 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 MDS plots for each chromatin feature is not too surprising. The cross-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 max pooling or LSTM 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 [30] manner. Given some of them are relatively simple motifs, it is possible that these matches are surrogates for important non-binding 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 preferrable depending on the use case. Since positive observations are created by extending from the midpoint, the effect of TFs that bind to the center of accessible or unmethylation 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 unmethylation regions are not aligned since 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 [31,32,33] plant pioneer transcription factors (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, since they are known [34] to be involved in chromatin remodeling. What role the Dof-type TFs play in accessibility is still unclear due to the wide variety of roles they play [35]. The shared top two motifs between the methylation models have evidence that they are involved in plant pathogen response [36,37]. Knowing that plant immunity genes are among the most variable [38], 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 due to association with the few causal TFs in the same family. While 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 about 1,500 [39], 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 generalizing across tissues as well as species. With the release of highly-accurate protein-folding models such as AlphaFold2 [40], 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 [12,41] 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 towards more biologically-relevant architectures.

Materials and Methods

Software environment

The software environment for the experiments was managed by conda (v4.10.3). Packages were downloaded from the conda-forge [42] and bioconda [43] channels. Software versions not explicitly mentioned in the methods are defined in the conda environment files in the companion code repository on Zenodo.

Raw data

The angiosperm ATAC-seq peaks [13] were downloaded from NCBI GEO accession GSE128434. Genomes and annotations for *Arabidopsis thaliana* (TAIR10) [44], *Eutrema salsugineum* (v1.0) [45], *Phaseolus vulgaris* (v1.0) [46], *Glycine max* (Wm82.a2.v1) [47], *Brachypodium distachyon* (v3.0) [48], *Oryza sativa* (v7.0) [49], *Setaria viridis* (v1.0) [50], *Populus trichocarpa* (v3.0) [51], and *Sorghum bicolor* (v3.1 and v3.1.1) [52] were downloaded from Phytozome. Reference genomes and annotations for *Zea mays* (AGPv4.38) [53] and *Hordeum vulgare* (IBSC_v2) [54] were downloaded from Ensembl Plants. The genome and annotation for *Asparagus officinalis* (v1.1) [55] was downloaded from the Asparagus Genome Project website. Unmethylated regions (UMRs) for the grasses were downloaded from the supplemental information of Crisp *et al.* 2020 [4]. For the unmethylated regions, the *Z. mays* AGPv4 genome and annotation was downloaded from MaizeGDB. The *Vitis vinifera* genome and annotation (Genoscope.12X) [56] were downloaded from the Genoscope website.

JASPAR 2020 Core *Plantae* [23] 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 [22] were downloaded from NCBI GEO accession GSE66386. The yeast (sacCer3 April 2011) [57] and human (hg19) [58] genomes were downloaded from NCBI. Maize scATAC-seq peaks [21] were downloaded from NCBI GEO accession GSE155178. Genome files were indexed using samtools [59].

UMR calling on non-grass species

UMR analysis on the non-grass species was performed as per Crisp *et al.* 2020. Briefly, sequencing reads were trimmed and quality checked using Trim galore! (0.6.4_dev), powered by cutadapt (v1.18) [60] and fastqc (v0.11.4). For all libraries, 20bp were trimmed from the 5′ ends of both R1 and R2 reads and aligned with bsmap (v2.74) [61] to the respective genomes with the following parameters: -v 5 to allow 5 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 [62] fixsam, sorted, and then indexed. Picard MarkDuplicates [63] was used to remove duplicates, BamTools filter to remove improperly paired reads, and bamUtil clipOverlap [64] to trim overlapping reads so as to only count cytosines once per sequenced molecule in a pair for PE 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 summarised into non-overlapping 100bp windows tiling the genome. WGBS pipelines are available on GitHub. To identify unmethylated regions, each 100bp 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 [4].

Training data preprocessing

Interval manipulation was done using a combination of the GNU coreutils, gawk, and bedtools [65]. 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 (2kb), and as distal if neither genic nor proximal. Previous work [20,66] 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 basepairs long and have more than five positive observations. We then downsampled the remaining observations to obtain a training set for the across-species models with a balanced representation of species and target class. 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 [67] and pyfaidx [68]

Training and evaluating the DanQ architecture

The DanQ architecture was implemented using the keras [69] API of tensorflow [70]. 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. Since 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 (auROC). Performance metrics were measured using scikit-learn [71] and curves were plotted using matplotlib [72]. Each model was trained three times to obtain an estimate of variability in performance due to the stochastic nature of the model variable initialization. For comparison between models we used the first of the three trained models.

The bag-of-kmers model was trained and tested independently on the within-species *Z. mays* accessibility and methylation training data using code adapted from Tu *et al.* 2020 [12] and compared to the within-species *Z. mays* accessibility and methylation models. For the two-step masked model comparison we masked the *Z. mays*-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 [73] 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 *Z. mays*-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 (accessibile/inaccessible or unmethylated/methylated). The “exp” accessibility model changed the activation function on the convolutional layer from ReLU to exponential. The “all_v_AtZm” accessibility model was tested on *A. thaliana* and *Z. mays* and trained on the rest of the angiosperm species.

The dendrogram in Figure 1 was plotted using the Phylo package of Biopython [74].

Analysis of maize scATAC-Seq data

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 CPM (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 [21]. Accessibility was predicted using the *Z. mays*-held-out model.

TF-MoDISco and kmer occlusion

We ran TF-MoDISco [75] with a sliding window size of 15bp, a flank size of 5bp, and a target seqlet FDR of 0.15. For converting seqlets to patterns, we set “trim_to_window_size” to 15bp, “initial_flank_to_add” to 5bp, and specified a final minimum cluster size of 60.

The kmer-occlusion method involves masking (replacing with N’s) a sliding kmer across each sequence in a given model’s test set. The difference between the model’s masked and unmasked prediction is the kmer’s “effect size”. We ran the kmer-occlusion method with a kmer size of 10bp on all species and chromatin feature pairs. The top 5% accessibility- or methylation-reducing kmers per species and chromatin feature were classified as “high-effect” kmers. We performed an all-by-all global alignment of the high-effect kmers per species and chromatin feature using Biopython’s pairwise aligner [67]. Using the alignment distance matrix, we clustered these high-effect kmers into 100 representative kmers using k-medoids [76]. We took the 100 medoid kmers for each species and chromatin feature pair and did another all-by-all global alignment to create another distance matrix. The embedded kmer coordinates were created using the MDS function in scikit-learn’s manifold package. High-effect kmers were matched to JASPAR 2020 CORE *plantae* motifs using FIMO [77] and a q-value threshold of 0.05.

Positional Global Importance Analysis

Global importance analysis (GIA) [24] 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 600bp sequences. The 600bp 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. GNU parallel [78] was used to speed up the pGIA analysis.

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

Manuscript

This manuscript was formatted with Manubot [79].

Acknowledgements

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 analysis.

References

1. **Towards genome-wide prediction and characterization of enhancers in plants**
Alexandre P Marand, Tao Zhang, Bo Zhu, Jiming Jiang
Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms (2017-01) https://doi.org/gkzgb9
DOI: 10.1016/j.bbagrm.2016.06.006 · PMID: 27321818

2. **Plant Enhancers: A Call for Discovery**
Blaise Weber, Johan Zicola, Rurika Oka, Maike Stam
Trends in Plant Science (2016-11) https://doi.org/ggzsjj
DOI: 10.1016/j.tplants.2016.07.013 · PMID: 27593567

3. **Transposable element influences on gene expression in plants**
Cory D Hirsch, Nathan M Springer
Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms (2017-01) https://doi.org/gd8kfs
DOI: 10.1016/j.bbagrm.2016.05.010 · PMID: 27235540

4. **Stable unmethylated DNA demarcates expressed genes and their cis-regulatory space in plant genomes**
Peter A Crisp, Alexandre P Marand, Jaclyn M Noshay, Peng Zhou, Zefu Lu, Robert J Schmitz, Nathan M Springer
Proceedings of the National Academy of Sciences (2020-09-22) https://doi.org/gmvz6r
DOI: 10.1073/pnas.2010250117 · PMID: 32879011 · PMCID: PMC7519222

5. **Evolutionary genetics of plant adaptation**
Jill T Anderson, John H Willis, Thomas Mitchell-Olds
Trends in Genetics (2011-07) https://doi.org/b7vnjs
DOI: 10.1016/j.tig.2011.04.001 · PMID: 21550682 · PMCID: PMC3123387

6. **Evolutionarily informed deep learning methods for predicting relative transcript abundance from DNA sequence**
Jacob D Washburn, Maria Katherine Mejia-Guerra, Guillaume Ramstein, Karl A Kremling, Ravi Valluru, Edward S Buckler, Hai Wang
Proceedings of the National Academy of Sciences (2019-03-19) https://doi.org/ggzr4h
DOI: 10.1073/pnas.1814551116 · PMID: 30842277 · PMCID: PMC6431157

7. **Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks**
Vikram Agarwal, Jay Shendure
Cell Reports (2020-05) https://doi.org/ggw7fr
DOI: 10.1016/j.celrep.2020.107663 · PMID: 32433972

8. **Effective gene expression prediction from sequence by integrating long-range interactions**
Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, David R Kelley
Cold Spring Harbor Laboratory (2021-04-08) https://doi.org/gjpx5v
DOI: 10.1101/2021.04.07.438649

9. **Predicting effects of noncoding variants with deep learning–based sequence model**
Jian Zhou, Olga G Troyanskaya

Nature Methods (2015-08-24) https://doi.org/gcgk8g
DOI: 10.1038/nmeth.3547 · PMID: 26301843 · PMCID: PMC4768299

10. **DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences**
Daniel Quang, Xiaohui Xie
Nucleic Acids Research (2016-06-20) https://doi.org/f8v4wj
DOI: 10.1093/nar/gkw226 · PMID: 27084946 · PMCID: PMC4914104

11. **Cross-species regulatory sequence activity prediction**
David R Kelley
PLOS Computational Biology (2020-07-20) https://doi.org/gg645k
DOI: 10.1371/journal.pcbi.1008050 · PMID: 32687525 · PMCID: PMC7392335

12. **Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors**
Xiaoyu Tu, María Katherine Mejía-Guerra, Jose A Valdes Franco, David Tzeng, Po-Yu Chu, Wei Shen, Yingying Wei, Xiuru Dai, Pinghua Li, Edward S Buckler, Silin Zhong
Nature Communications (2020-10-09) https://doi.org/gh33cj
DOI: 10.1038/s41467-020-18832-8 · PMID: 33037196 · PMCID: PMC7547689

13. **The prevalence, evolution and chromatin signatures of plant regulatory elements**
Zefu Lu, Alexandre P Marand, William A Ricci, Christina L Ethridge, Xiaoyu Zhang, Robert J Schmitz
Nature Plants (2019-11-18) https://doi.org/ggd822
DOI: 10.1038/s41477-019-0548-z · PMID: 31740772

14. **LEAFY, a Pioneer Transcription Factor in Plants: A Mini-Review**
Nobutoshi Yamaguchi
Frontiers in Plant Science (2021-07-05) https://doi.org/gmdgvv
DOI: 10.3389/fpls.2021.701406 · PMID: 34290727 · PMCID: PMC8287900

15. **Three-dimensional chromatin packing and positioning of plant genomes**
Ezgi Süheyla Doğan, Chang Liu
Nature Plants (2018-07-30) https://doi.org/gd9ndt
DOI: 10.1038/s41477-018-0199-5 · PMID: 30061747

16. **The transcription regulatory code of a plant leaf**
Xiaoyu Tu, María Katherine Mejía-Guerra, Jose AValdes Franco, David Tzeng, Po-Yu Chu, Xiuru Dai, Pinghua Li, Edward S Buckler, Silin Zhong
Cold Spring Harbor Laboratory (2020-04-22) https://doi.org/ggwrtw
DOI: 10.1101/2020.01.07.898056

17. **Prediction of gene regulatory enhancers across species reveals evolutionarily conserved sequence properties**
Ling Chen, Alexandra E Fish, John A Capra
PLOS Computational Biology (2018-10-04) https://doi.org/gfdkd4
DOI: 10.1371/journal.pcbi.1006484 · PMID: 30286077 · PMCID: PMC6191148

18. **Transcriptional enhancers: from properties to genome-wide predictions**
Daria Shlyueva, Gerald Stampfel, Alexander Stark
Nature Reviews Genetics (2014-03-11) https://doi.org/f3swzg
DOI: 10.1038/nrg3682 · PMID: 24614317

19. **Base-resolution models of transcription-factor binding reveal soft motif syntax**
Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, Julia Zeitlinger
Nature Genetics (2021-02-18) https://doi.org/gh4tbk
DOI: 10.1038/s41588-021-00782-6 · PMID: 33603233

20. **The impact of different negative training data on regulatory sequence predictions**
Louisa-Marie Krützfeldt, Max Schubach, Martin Kircher
Cold Spring Harbor Laboratory (2020-07-28) https://doi.org/gg7ww5
DOI: 10.1101/2020.07.28.224485

21. **A cis-regulatory atlas in maize at single-cell resolution**
Alexandre P Marand, Zongliang Chen, Andrea Gallavotti, Robert J Schmitz
Cell (2021-05) https://doi.org/gjwwb4
DOI: 10.1016/j.cell.2021.04.014 · PMID: 33964211

22. **Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions**
Alicia N Schep, Jason D Buenrostro, Sarah K Denny, Katja Schwartz, Gavin Sherlock, William J Greenleaf
Genome Research (2015-11) https://doi.org/f7xzhg
DOI: 10.1101/gr.192294.115 · PMID: 26314830 · PMCID: PMC4617971

23. **JASPAR 2020: update of the open-access database of transcription factor binding profiles**
Oriol Fornes, Jaime A Castro-Mondragon, Aziz Khan, Robin van der Lee, Xi Zhang, Phillip A Richmond, Bhavi P Modi, Solenne Correard, Marius Gheorghe, Damir Baranašić, … Anthony Mathelier
Nucleic Acids Research (2019-11-08) https://doi.org/ggrsnn
DOI: 10.1093/nar/gkz1001 · PMID: 31701148 · PMCID: PMC7145627

24. **Global importance analysis: An interpretability method to quantify importance of genomic features in deep neural networks**
Peter K Koo, Antonio Majdandzic, Matthew Ploenzke, Praveen Anand, Steffan B Paul
PLOS Computational Biology (2021-05-13) https://doi.org/gksp3k
DOI: 10.1371/journal.pcbi.1008925 · PMID: 33983921 · PMCID: PMC8118286

25. **Polycomb repression in the regulation of growth and development in Arabidopsis**
Jun Xiao, Doris Wagner
Current Opinion in Plant Biology (2015-02) https://doi.org/f63kzz
DOI: 10.1016/j.pbi.2014.10.003 · PMID: 25449722

26. **Polycomb-Repressed Genes Have Permissive Enhancers that Initiate Reprogramming**
Phillippa C Taberlay, Theresa K Kelly, Chun-Chi Liu, Jueng Soo You, Daniel D De Carvalho, Tina B Miranda, Xianghong J Zhou, Gangning Liang, Peter A Jones
Cell (2011-12) https://doi.org/bw7b3j
DOI: 10.1016/j.cell.2011.10.040 · PMID: 22153073 · PMCID: PMC3240866

27. **Small RNAs and transposon silencing in plants**
Hidetaka Ito
Development, Growth & Differentiation (2012-01) https://doi.org/cd64vd
DOI: 10.1111/j.1440-169x.2011.01309.x · PMID: 22150226

28. **Why are Saliency Maps Noisy? Cause of and Solution to Noisy Saliency Maps**
Beomsu Kim, Junghoon Seo, Seunghyeon Jeon, Jamyoung Koo, Jeongyeol Choe, Taegyun Jeon

Institute of Electrical and Electronics Engineers (IEEE) (2019-10) https://doi.org/gm3w6h
DOI: 10.1109/iccvw.2019.00510

29. **Why Transcription Factor Binding Sites Are Ten Nucleotides Long**
Alexander J Stewart, Sridhar Hannenhalli, Joshua B Plotkin
Genetics (2012-11-01) https://doi.org/f4d632
DOI: 10.1534/genetics.112.143370 · PMID: 22887818 · PMCID: PMC3522170

30. **Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq**
CD Arnold, D Gerlach, C Stelzer, LM Boryn, M Rath, A Stark
Science (2013-01-17) https://doi.org/f3sn33
DOI: 10.1126/science.1232542 · PMID: 23328393

31. **Embryonic epigenetic reprogramming by a pioneer transcription factor in plants**
Zeng Tao, Lisha Shen, Xiaofeng Gu, Yizhong Wang, Hao Yu, Yuehui He
Nature (2017-10-25) https://doi.org/gcf5hp
DOI: 10.1038/nature24300 · PMID: 29072296

32. **LEAFY is a pioneer transcription factor and licenses cell reprogramming to floral fate**
Run Jin, Samantha Klasfeld, Yang Zhu, Meilin Fernandez Garcia, Jun Xiao, Soon-Ki Han, Adam Konkol, Doris Wagner
Nature Communications (2021-01-27) https://doi.org/gmdgvt
DOI: 10.1038/s41467-020-20883-w · PMID: 33504790 · PMCID: PMC7840934

33. **The LEAFY floral regulator displays pioneer transcription factor properties**
Xuelei Lai, Romain Blanc-Mathieu, Loïc GrandVuillemin, Ying Huang, Arnaud Stigliani, Jérémy Lucas, Emmanuel Thévenon, Jeanne Loue-Manifel, Laura Turchi, Hussein Daher, … François Parcy
Molecular Plant (2021-05) https://doi.org/gmdgvs
DOI: 10.1016/j.molp.2021.03.004 · PMID: 33684542

34. **Regulation of plant architecture by a new histone acetyltransferase targeting gene bodies**
Xueyong Yang, Jianbin Yan, Zhen Zhang, Tao Lin, Tongxu Xin, Bowen Wang, Shenhao Wang, Jicheng Zhao, Zhonghua Zhang, William J Lucas, … Sanwen Huang
Nature Plants (2020-07-13) https://doi.org/gm52f8
DOI: 10.1038/s41477-020-0715-2 · PMID: 32665652

35. **The role of the DNA-binding One Zinc Finger (DOF) transcription factor family in plants**
Mélanie Noguero, Rana Muhammad Atif, Sergio Ochatt, Richard D Thompson
Plant Science (2013-08) https://doi.org/f437jw
DOI: 10.1016/j.plantsci.2013.03.016 · PMID: 23759101

36. **Flg22 regulates the release of an ethylene response factor substrate from MAP kinase 6 in Arabidopsis thaliana via ethylene signaling**
G Bethke, T Unthan, JF Uhrig, Y Poschl, AA Gust, D Scheel, J Lee
Proceedings of the National Academy of Sciences (2009-04-29) https://doi.org/b3n92d
DOI: 10.1073/pnas.0810206106 · PMID: 19416906 · PMCID: PMC2683104

37. **A High-Throughput Screening System for Arabidopsis Transcription Factors and Its Application to Med25-Dependent Transcriptional Regulation**
Bin Ou, Kang-Quan Yin, Sai-Nan Liu, Yan Yang, Tren Gu, Jennifer Man Wing Hui, Li Zhang, Jin Miao, Youichi Kondou, Minami Matsui, … Li-Jia Qu
Molecular Plant (2011-05) https://doi.org/cm8w9q

DOI: 10.1093/mp/ssr002 · PMID: 21343311

38. **A Species-Wide Inventory of NLR Genes and Alleles in Arabidopsis thaliana**
Anna-Lena Van de Weyer, Freddy Monteiro, Oliver J Furzer, Marc T Nishimura, Volkan Cevik, Kamil Witek, Jonathan DG Jones, Jeffery L Dangl, Detlef Weigel, Felix Bemm
Cell (2019-08) https://doi.org/ggfvc6
DOI: 10.1016/j.cell.2019.07.038 · PMID: 31442410 · PMCID: PMC6709784

39. **Genome-Wide Phylogenetic Comparative Analysis of Plant Transcriptional Regulation: A Timeline of Loss, Gain, Expansion, and Correlation with Complexity**
Daniel Lang, Benjamin Weiche, Gerrit Timmerhaus, Sandra Richardt, Diego M Riaño-Pachón, Luiz GG Corrêa, Ralf Reski, Bernd Mueller-Roeber, Stefan A Rensing
Genome Biology and Evolution (2010) https://doi.org/fs6km4
DOI: 10.1093/gbe/evq032 · PMID: 20644220 · PMCID: PMC2997552

40. **Highly accurate protein structure prediction with AlphaFold**
John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, … Demis Hassabis
Nature (2021-07-15) https://doi.org/gk7nfp
DOI: 10.1038/s41586-021-03819-2 · PMID: 34265844 · PMCID: PMC8371605

41. **Effective gene expression prediction from sequence by integrating long-range interactions**
Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R Ledsam, Agnieszka Grabska-Barwinska, Kyle R Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, David R Kelley
Nature Methods (2021-10-04) https://doi.org/gm2wv4
DOI: 10.1038/s41592-021-01252-x · PMID: 34608324 · PMCID: PMC8490152

42. **The conda-forge Project: Community-based Software Distribution Built on the conda Package Format and Ecosystem**
Conda-Forge Community
Zenodo (2015-07-12) https://doi.org/gmzdsn
DOI: 10.5281/zenodo.4774216

43. **Bioconda: sustainable and comprehensive software distribution for the life sciences**
Björn Grüning, Ryan Dale, Andreas Sjödin, Brad A Chapman, Jillian Rowe, Christopher H Tomkins-Tinch, Renan Valieris, Johannes Köster, The Bioconda Team
Nature Methods (2018-07-02) https://doi.org/gd2xzp
DOI: 10.1038/s41592-018-0046-7 · PMID: 29967506

44. **The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools**
Philippe Lamesch, Tanya Z Berardini, Donghui Li, David Swarbreck, Christopher Wilks, Rajkumar Sasidharan, Robert Muller, Kate Dreher, Debbie L Alexander, Margarita Garcia-Hernandez, … Eva Huala
Nucleic Acids Research (2012-01) https://doi.org/cc3nr3
DOI: 10.1093/nar/gkr1090 · PMID: 22140109 · PMCID: PMC3245047

45. **The Reference Genome of the Halophytic Plant Eutrema salsugineum**
Ruolin Yang, David E Jarvis, Hao Chen, Mark A Beilstein, Jane Grimwood, Jerry Jenkins, ShengQiang Shu, Simon Prochnik, Mingming Xin, Chuang Ma, … Xiangfeng Wang
Frontiers in Plant Science (2013) https://doi.org/gkzg3t
DOI: 10.3389/fpls.2013.00046 · PMID: 23518688 · PMCID: PMC3604812

46. **A reference genome for common bean and genome-wide analysis of dual domestications**

Jeremy Schmutz, Phillip E McClean, Sujan Mamidi, GAlbert Wu, Steven B Cannon, Jane Grimwood, Jerry Jenkins, Shengqiang Shu, Qijian Song, Carolina Chavarro, … Scott A Jackson
Nature Genetics (2014-06-08) https://doi.org/f57qhm
DOI: 10.1038/ng.3008 · PMID: 24908249 · PMCID: PMC7048698

47. **Genome sequence of the palaeopolyploid soybean**
Jeremy Schmutz, Steven B Cannon, Jessica Schlueter, Jianxin Ma, Therese Mitros, William Nelson, David L Hyten, Qijian Song, Jay J Thelen, Jianlin Cheng, … Scott A Jackson
Nature (2010-01) https://doi.org/cf4xb5
DOI: 10.1038/nature08670 · PMID: 20075913

48. **Genome sequencing and analysis of the model grass Brachypodium distachyon**
The International Brachypodium Initiative
Nature (2010-02) https://doi.org/d6n7pw
DOI: 10.1038/nature08747 · PMID: 20148030

49. **The TIGR Rice Genome Annotation Resource: improvements and new features**
S Ouyang, W Zhu, J Hamilton, H Lin, M Campbell, K Childs, F Thibaud-Nissen, RL Malek, Y Lee, L Zheng, … CR Buell
Nucleic Acids Research (2007-01-03) https://doi.org/cwnws4
DOI: 10.1093/nar/gkl976 · PMID: 17145706 · PMCID: PMC1751532

50. **A genome resource for green millet Setaria viridis enables discovery of agronomically valuable loci**
Sujan Mamidi, Adam Healey, Pu Huang, Jane Grimwood, Jerry Jenkins, Kerrie Barry, Avinash Sreedasyam, Shengqiang Shu, John T Lovell, Maximilian Feldman, … Elizabeth A Kellogg
Nature Biotechnology (2020-10-05) https://doi.org/gjzqtb
DOI: 10.1038/s41587-020-0681-2 · PMID: 33020633 · PMCID: PMC7536120

51. **The Genome of Black Cottonwood, Populus trichocarpa (Torr. &amp; Gray)**
GA Tuskan, S DiFazio, S Jansson, J Bohlmann, I Grigoriev, U Hellsten, N Putnam, S Ralph, S Rombauts, A Salamov, … D Rokhsar
Science (2006-09-15) https://doi.org/c7hs34
DOI: 10.1126/science.1128691 · PMID: 16973872

52. **The <i>Sorghum bicolor</i> reference genome: improved assembly, gene annotations, a transcriptome atlas, and signatures of genome organization**
Ryan F McCormick, Sandra K Truong, Avinash Sreedasyam, Jerry Jenkins, Shengqiang Shu, David Sims, Megan Kennedy, Mojgan Amirebrahimi, Brock D Weers, Brian McKinley, … John E Mullet
The Plant Journal (2017-12-28) https://doi.org/gmz6s3
DOI: 10.1111/tpj.13781 · PMID: 29161754

53. **Improved maize reference genome with single-molecule technologies**
Yinping Jiao, Paul Peluso, Jinghua Shi, Tiffany Liang, Michelle C Stitzer, Bo Wang, Michael S Campbell, Joshua C Stein, Xuehong Wei, Chen-Shan Chin, … Doreen Ware
Nature (2017-06-12) https://doi.org/gbhcq6
DOI: 10.1038/nature22971 · PMID: 28605751 · PMCID: PMC7052699

54. **A chromosome conformation capture ordered sequence of the barley genome**
Martin Mascher, Heidrun Gundlach, Axel Himmelbach, Sebastian Beier, Sven O Twardziok, Thomas Wicker, Volodymyr Radchuk, Christoph Dockter, Pete E Hedley, Joanne Russell, … Nils Stein
Nature (2017-04-26) https://doi.org/f95vb9

DOI: 10.1038/nature22043 · PMID: 28447635

55. **The asparagus genome sheds light on the origin and evolution of a young Y chromosome**
Alex Harkess, Jinsong Zhou, Chunyan Xu, John E Bowers, Ron Van der Hulst, Saravanaraj Ayyampalayam, Francesco Mercati, Paolo Riccardi, Michael R McKain, Atul Kakrana, … Guangyu Chen
Nature Communications (2017-11-02) https://doi.org/gcjdtf
DOI: 10.1038/s41467-017-01064-8 · PMID: 29093472 · PMCID: PMC5665984

56. **The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla**
The French–Italian Public Consortium for Grapevine Genome Characterization
Nature (2007-08-26) https://doi.org/ckfnh2
DOI: 10.1038/nature06148 · PMID: 17721507

57. **Erratum: Overview of the yeast genome**
HW Mewes, K Albermann, M Bähr, D Frishman, A Gleissner, J Hani, K Heumann, K Kleine, A Maierl, SG Oliver, … A Zollner
Nature (1997-06-12) https://doi.org/cwhbkg
DOI: 10.1038/42755 · PMID: 9169865

58. **Modernizing Reference Genome Assemblies**
Deanna M Church, Valerie A Schneider, Tina Graves, Katherine Auger, Fiona Cunningham, Nathan Bouk, Hsiu-Chuan Chen, Richa Agarwala, William M McLaren, Graham RS Ritchie, … Tim Hubbard
PLoS Biology (2011-07-05) https://doi.org/djgd3t
DOI: 10.1371/journal.pbio.1001091 · PMID: 21750661 · PMCID: PMC3130012

59. **Twelve years of SAMtools and BCFtools**
Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, Heng Li
GigaScience (2021-02-16) https://doi.org/gjxzc9
DOI: 10.1093/gigascience/giab008 · PMID: 33590861 · PMCID: PMC7931819

60. **Cutadapt removes adapter sequences from high-throughput sequencing reads**
Marcel Martin
EMBnet.journal (2011-05-02) https://doi.org/gdh7xt
DOI: 10.14806/ej.17.1.200

61. **BSMAP: whole genome bisulfite sequence MAPping program**
Yuanxin Xi, Wei Li
BMC Bioinformatics (2009-07-27) https://doi.org/cbrc35
DOI: 10.1186/1471-2105-10-232 · PMID: 19635165 · PMCID: PMC2724425

62. **The Sequence Alignment/Map format and SAMtools**
H Li, B Handsaker, A Wysoker, T Fennell, J Ruan, N Homer, G Marth, G Abecasis, R Durbin, 1000 Genome Project Data Processing Subgroup
Bioinformatics (2009-06-08) https://doi.org/ff6426
DOI: 10.1093/bioinformatics/btp352 · PMID: 19505943 · PMCID: PMC2723002

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

64. **An efficient and scalable analysis framework for variant extraction and refinement from population-scale DNA sequence data**
Goo Jun, Mary Kate Wing, Gonçalo R Abecasis, Hyun Min Kang
Genome Research (2015-06) https://doi.org/f7dz2d
DOI: 10.1101/gr.176552.114 · PMID: 25883319 · PMCID: PMC4448687

65. **BEDTools: a flexible suite of utilities for comparing genomic features**
Aaron R Quinlan, Ira M Hall
Bioinformatics (2010-03-15) https://doi.org/cmrms3
DOI: 10.1093/bioinformatics/btq033 · PMID: 20110278 · PMCID: PMC2832824

66. **The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics**
Qiong Wei, Roland L Dunbrack
PLoS ONE (2013-07-09) https://doi.org/f5bpzf
DOI: 10.1371/journal.pone.0067863 · PMID: 23874456 · PMCID: PMC3706434

67. **Biopython: freely available Python tools for computational molecular biology and bioinformatics**
PJA Cock, T Antao, JT Chang, BA Chapman, CJ Cox, A Dalke, I Friedberg, T Hamelryck, F Kauff, B Wilczynski, MJL de Hoon
Bioinformatics (2009-03-20) https://doi.org/d7zwd2
DOI: 10.1093/bioinformatics/btp163 · PMID: 19304878 · PMCID: PMC2682512

68. **Efficient "pythonic" access to FASTA files using pyfaidx**
Matthew D Shirley, Zhaorong Ma, Brent S Pedersen, Sarah J Wheelan
PeerJ (2018-01-12) https://doi.org/gfzprs
DOI: 10.7287/peerj.preprints.970v1

69. **Keras: the Python deep learning API** https://keras.io/

70. **TensorFlow**
TensorFlow Developers
Zenodo (2021-08-12) https://doi.org/gm2ghn
DOI: 10.5281/zenodo.4724125

71. **Scikit-learn: Machine Learning in Python**
Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, … Édouard Duchesnay
Journal of Machine Learning Research (2011) https://jmlr.csail.mit.edu/papers/v12/pedregosa11a.html

72. **Matplotlib: A 2D Graphics Environment**
John D Hunter
Computing in Science & Engineering (2007) https://doi.org/drbjhg
DOI: 10.1109/mcse.2007.55

73. **Pybedtools: a flexible Python library for manipulating genomic datasets and annotations**
Ryan K Dale, Brent S Pedersen, Aaron R Quinlan
Bioinformatics (2011-12-15) https://doi.org/bps7ds
DOI: 10.1093/bioinformatics/btr539 · PMID: 21949271 · PMCID: PMC3232365

74.

Bio.Phylo: A unified toolkit for processing, analyzing and visualizing phylogenetic trees in Biopython
Eric Talevich, Brandon M Invergo, Peter JA Cock, Brad A Chapman
BMC Bioinformatics (2012-08-21) https://doi.org/gb8t75
DOI: 10.1186/1471-2105-13-209 · PMID: 22909249 · PMCID: PMC3468381

75. **Technical Note on Transcription Factor Motif Discovery from Importance Scores (TF-MoDISco) version 0.5.6.5**
Avanti Shrikumar, Katherine Tian, Žiga Avsec, Anna Shcherbina, Abhimanyu Banerjee, Mahfuza Sharmin, Surag Nair, Anshul Kundaje
arXiv (2020-05-01) https://arxiv.org/abs/1811.00416

76. **NumPy / SciPy Recipes for Data Science: k-Medoids Clustering**
Christian Bauckhage
Unpublished (2015) https://doi.org/gksp3j
DOI: 10.13140/2.1.4453.2009

77. **FIMO: scanning for occurrences of a given motif**
Charles E Grant, Timothy L Bailey, William Stafford Noble
Bioinformatics (2011-04-01) https://doi.org/fcp52k
DOI: 10.1093/bioinformatics/btr064 · PMID: 21330290 · PMCID: PMC3065696

78. **Gnu Parallel 2018**
Ole Tange
Zenodo (2018-04-27) https://doi.org/gmzd4j
DOI: 10.5281/zenodo.1146014

79. **Open collaborative writing with Manubot**
Daniel S Himmelstein, Vincent Rubinetti, David R Slochower, Dongbo Hu, Venkat S Malladi, Casey S Greene, Anthony Gitter
PLOS Computational Biology (2019-06-24) https://doi.org/c7np
DOI: 10.1371/journal.pcbi.1007128 · PMID: 31233491 · PMCID: PMC6611653

Supplementary Information

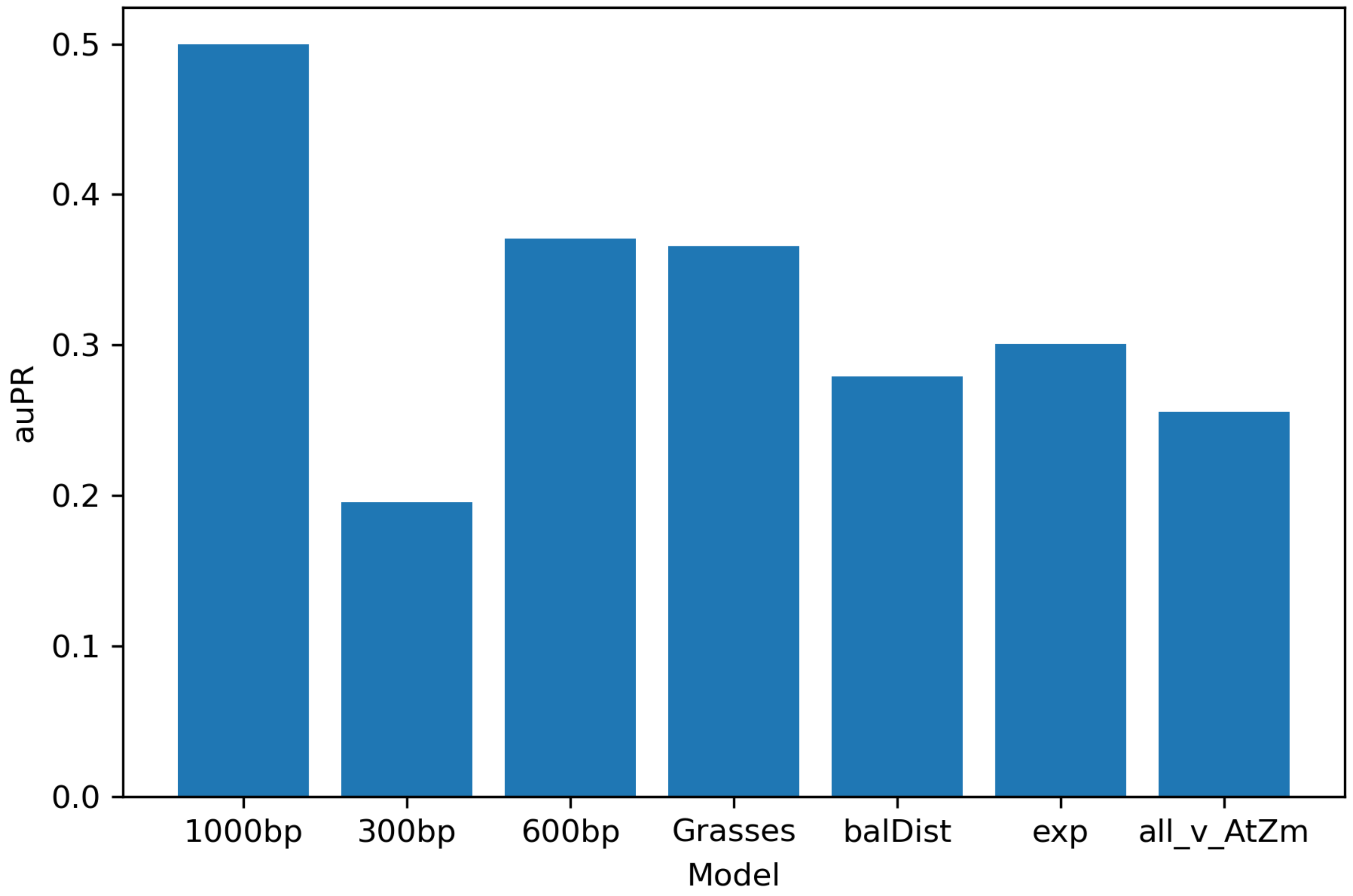

Figure S1: auPR of different across-species accessibility model configurations. From left to right: 1000bp windows, 300bp windows, 600bp windows, training and testing within only grasses, using a training set balanced on both accessibility and distance class, exponential activation on the convolutional layer, and testing on Arabidopsis and Maize while training on the rest.

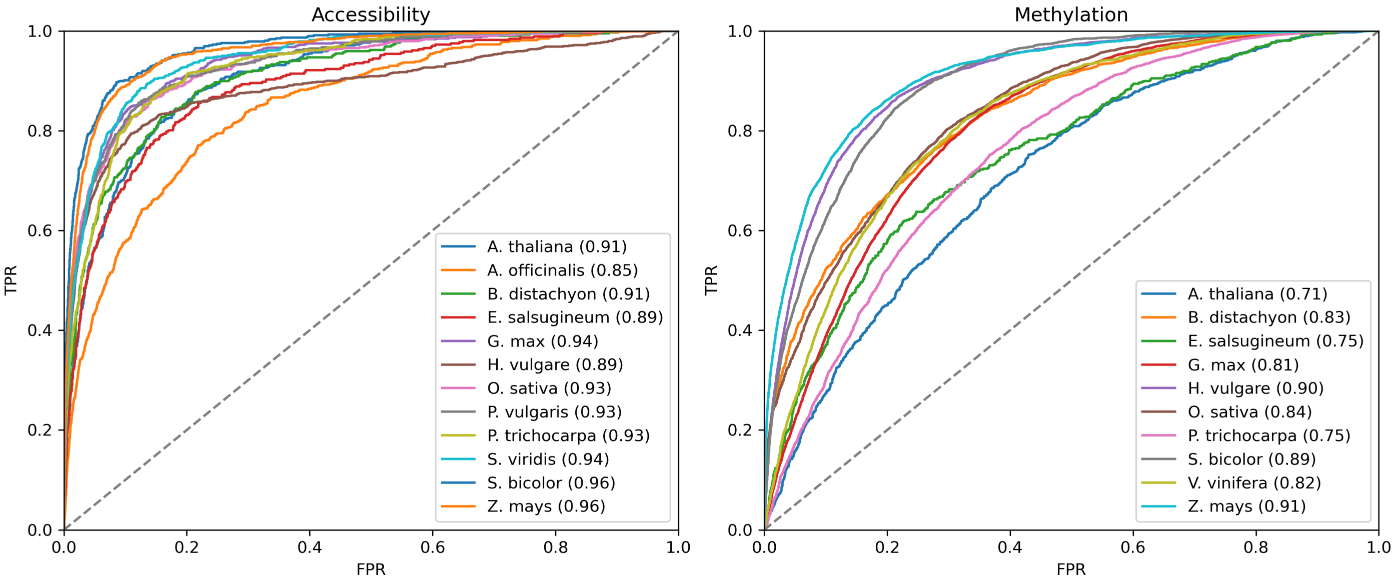

Figure S2: Receiver operating characteristic curves of the across-species models per hold-out species. The gray dashed line is the baseline expectation for a random classifier. The area under the curve is given inside the parentheses for each species in the legend.

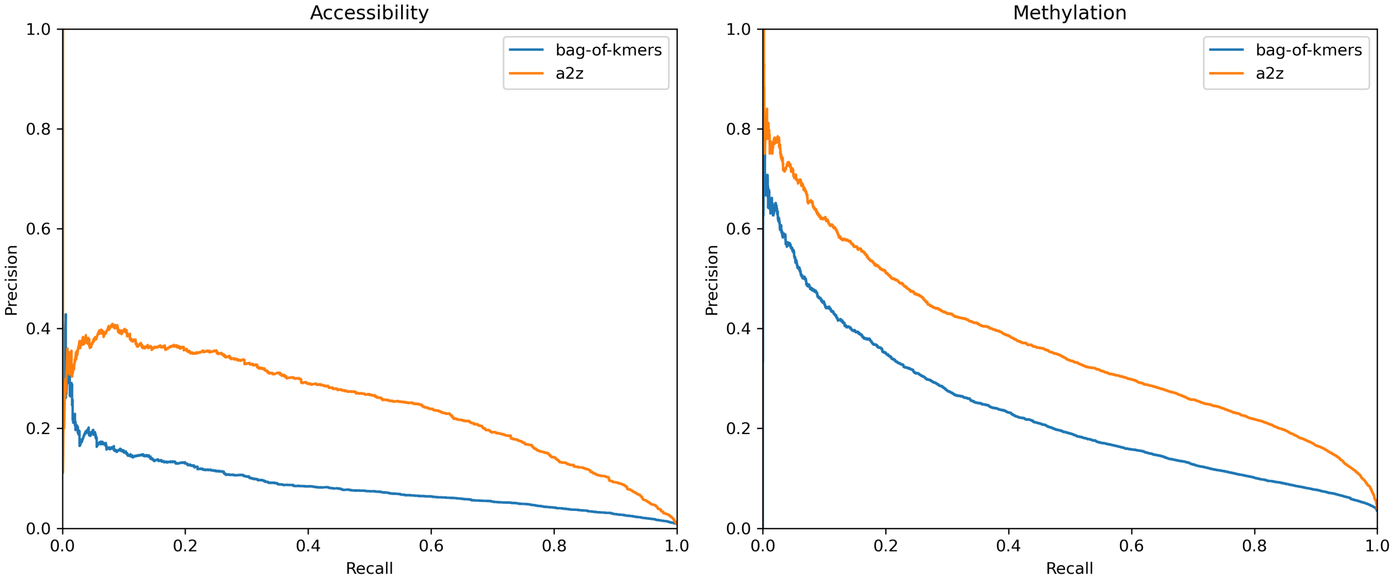

Figure S3: Precision-recall curve comparison of across-species a2z model with the bag-of-kmer model in *Z. mays*.

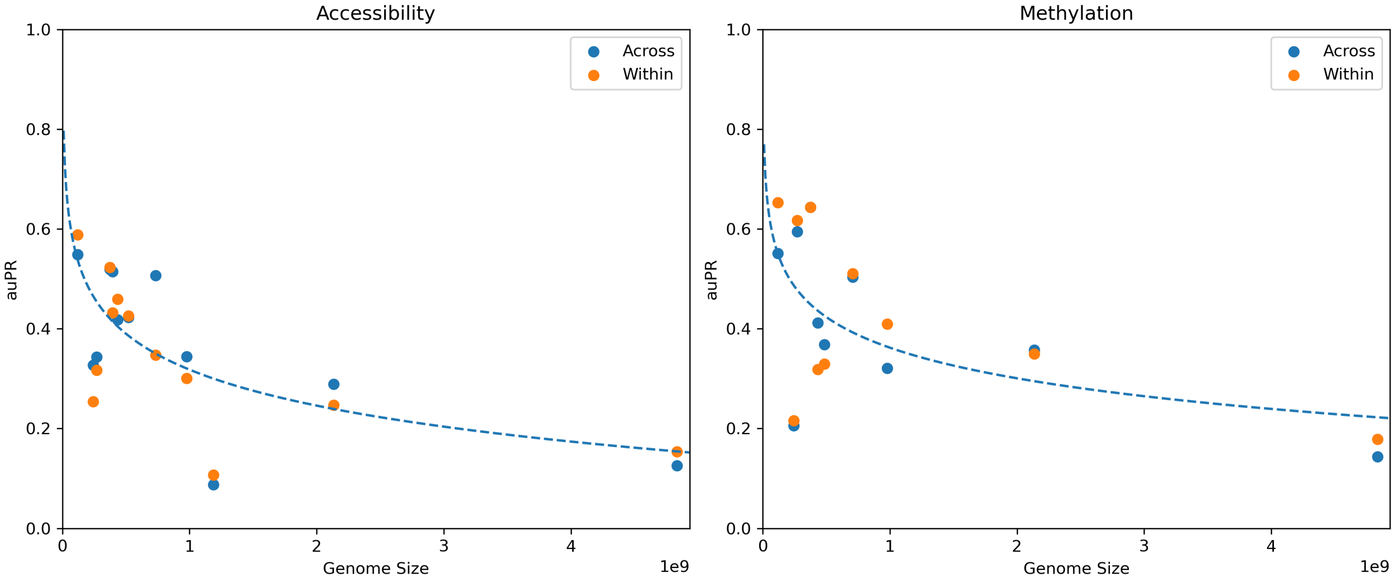

Figure S4: auPR of both model configurations with respect to genome size. The dashed line is an exponential fit to the across-species model auPR values.

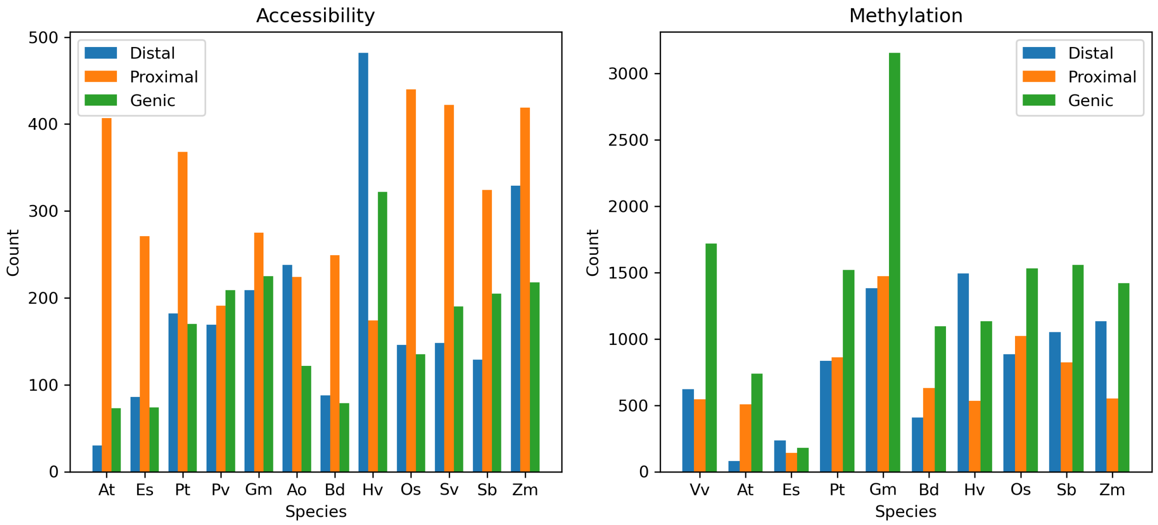

Figure S5: Counts of accessible regions in the across-species test sets by distance class and species.

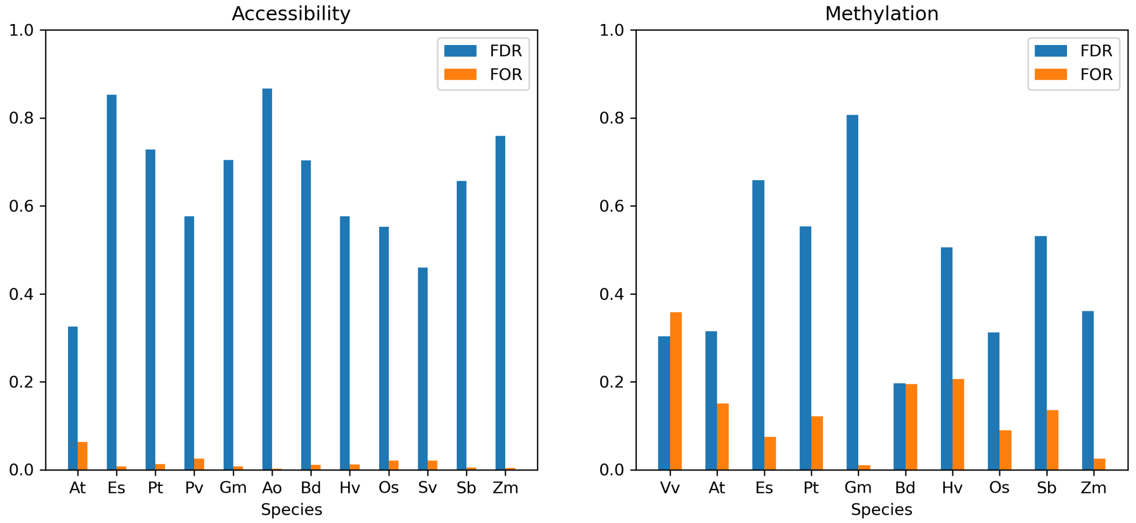

Figure S6: Comparison of false discovery rate (FDR) versus false omission rate (FOR) between models.

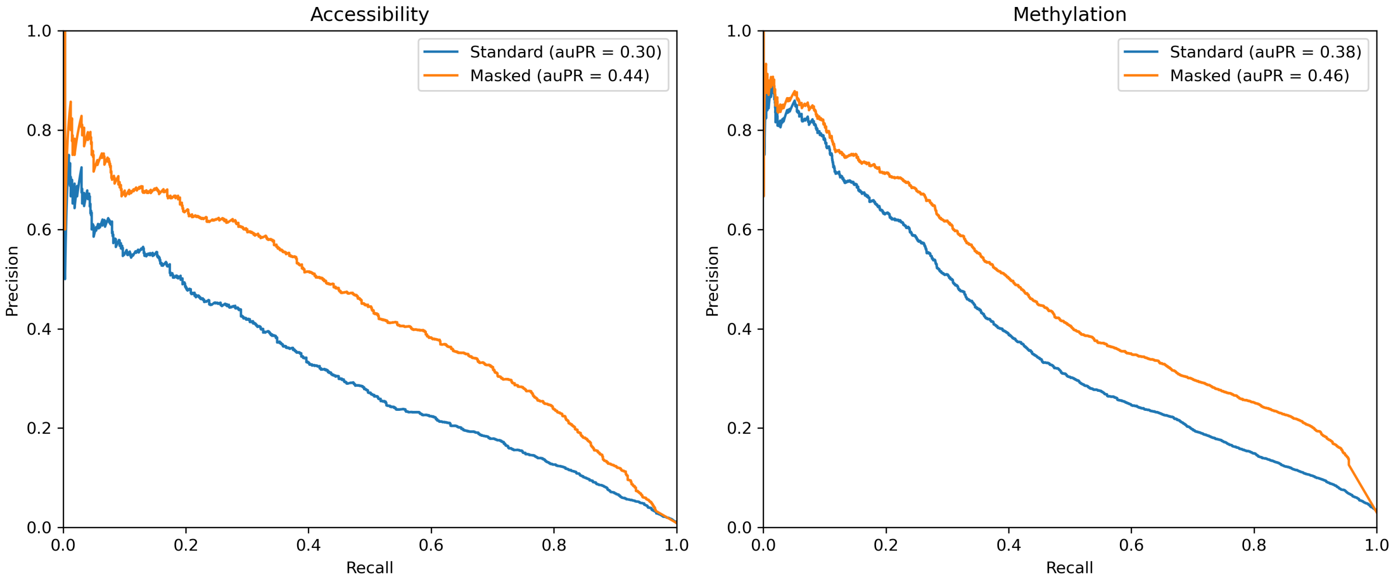

Figure S7: Precision-recall curve comparison between a2z model and the repeat-masked a2z model in *Z. mays*.

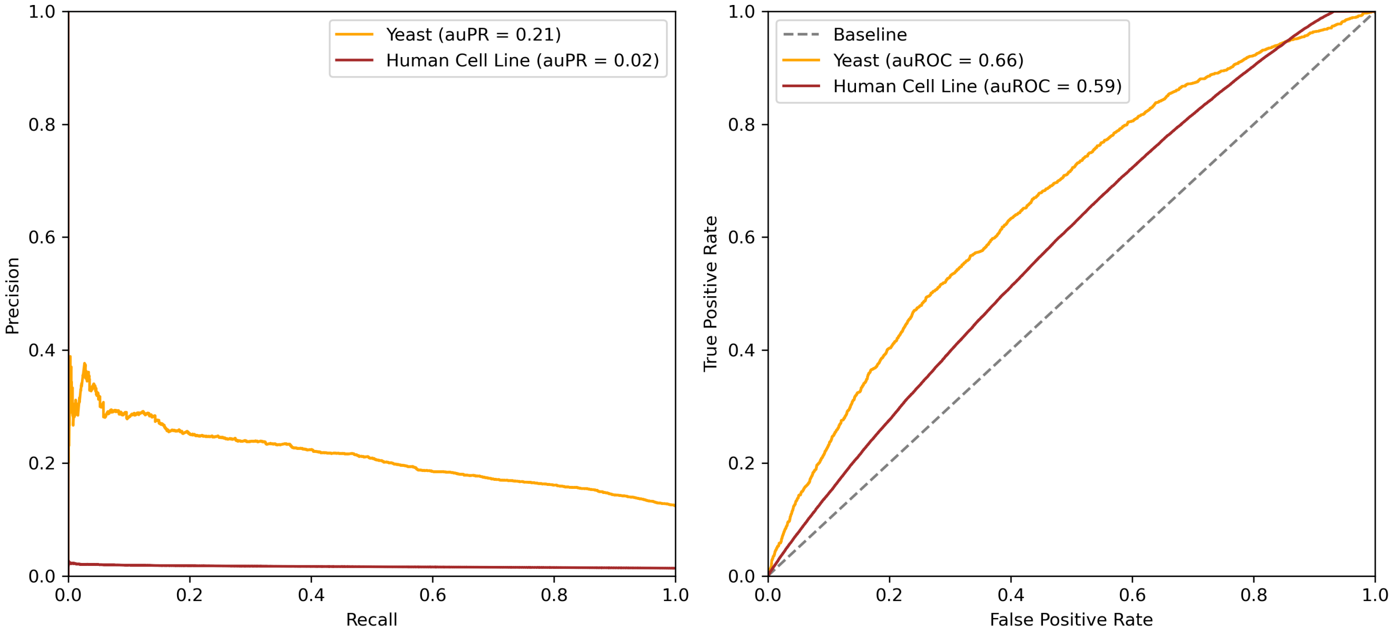

Figure S8: Precision-recall and receiver operating characteristic curves of an angiosperm-trained a2z model on yeast and a human cell line.

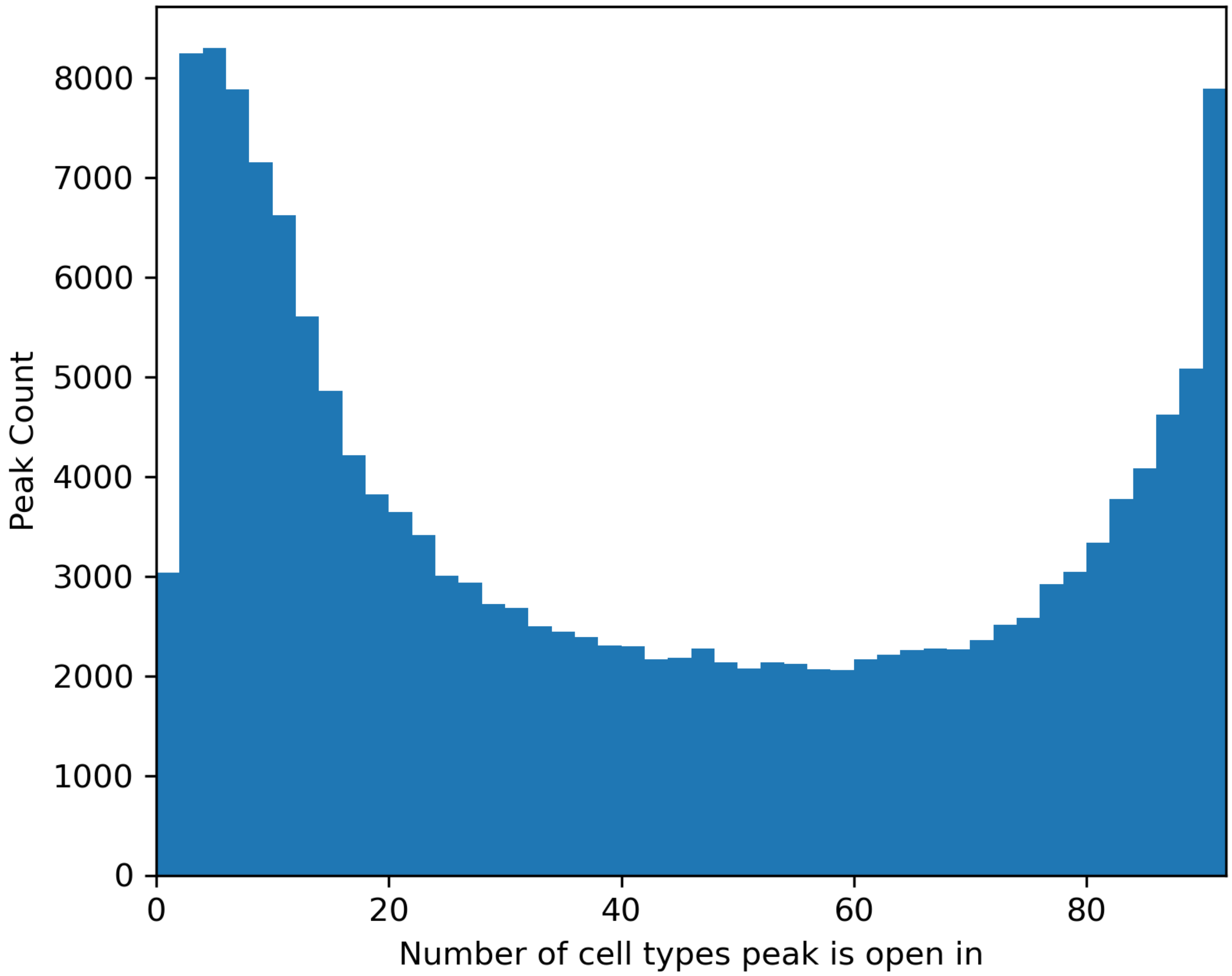

Figure S9: Cell type specificity of maize scATAC-Seq peaks.

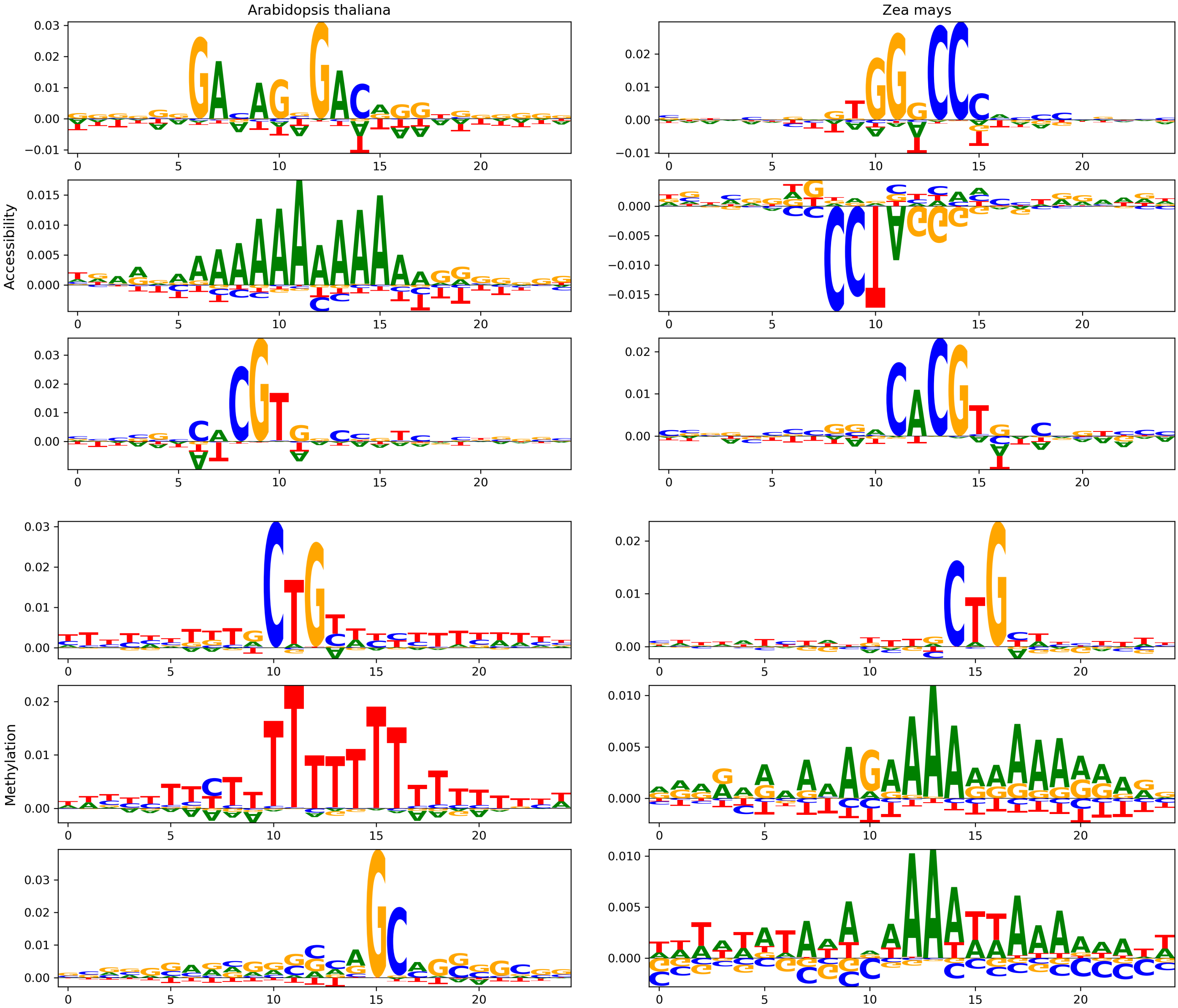

Figure S10: Top 3 TF-MoDISco patterns for four a2z models. *A. thaliana* is the left column, *Z. mays* is the right column. Accessibility is the top row and methylation is the bottom row. Within each species and chromatin feature combination the patterns are ranked from top to bottom by the number of supporting seqlets for that pattern.

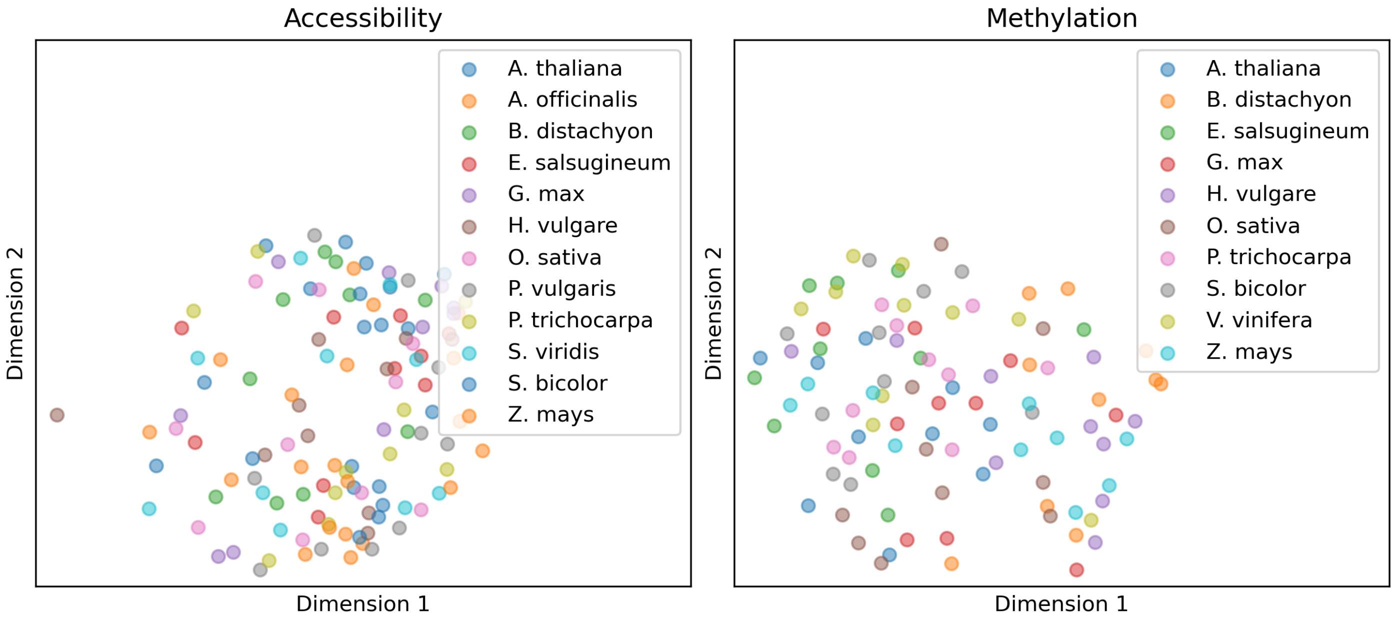

Figure S11: Multidimensional scaling of the top 10 high-effect medoid kmers for each species and chromatin feature model combination, colored by species.

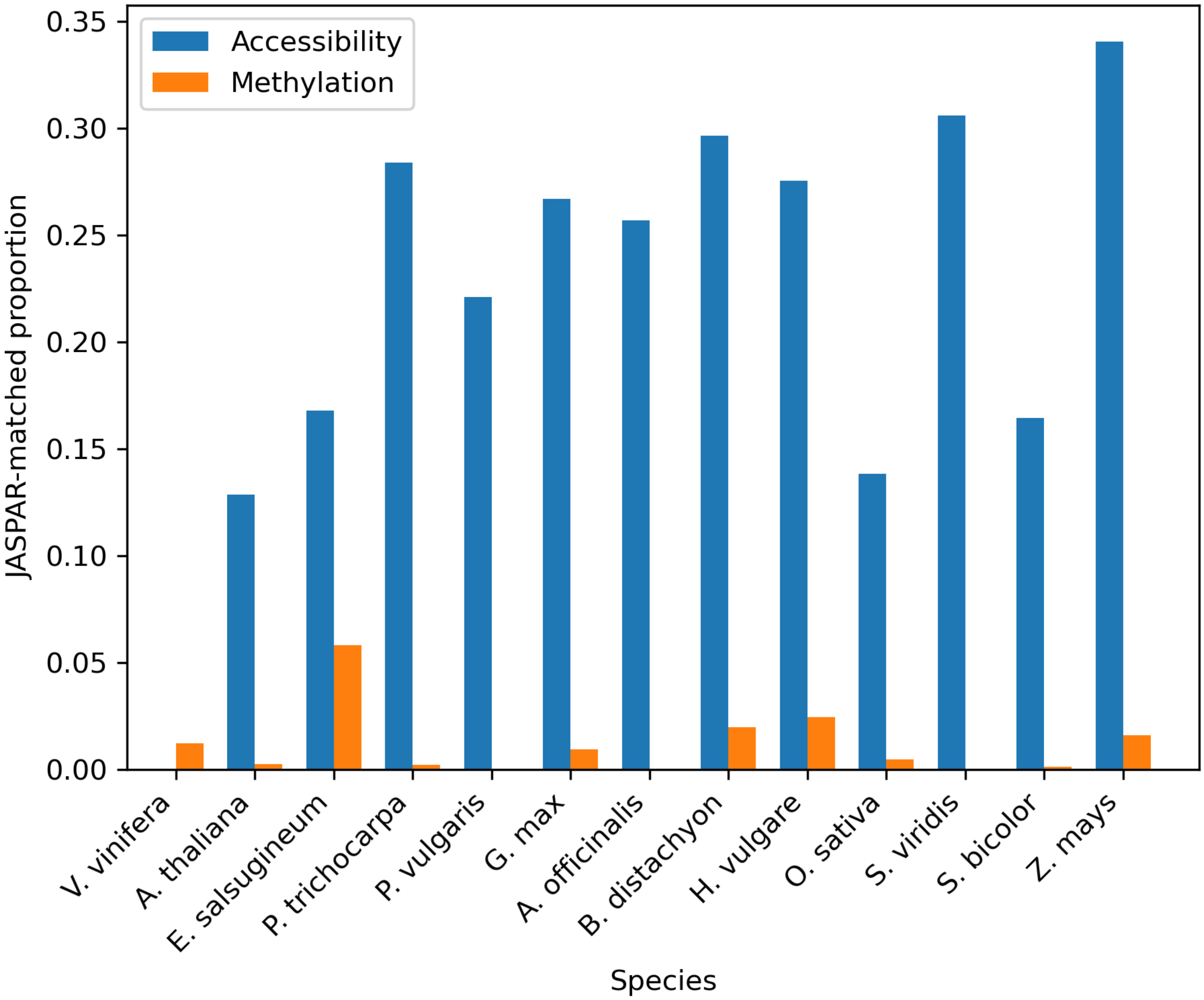

Figure S12: Proportion of high-effect kmers that significantly (q-value <= 0.05) matched JASPAR CORE *plantae* motifs with FIMO, grouped by species and chromatin feature.