DNA Research, 2022, 29(4), 1–15
https://doi.org/10.1093/dnares/dsac029
Advance Access Publication Date: 29 July 2022
Research Article

OXFORD

Research Article

BLSSpeller to discover novel regulatory motifs in maize

Razgar Seyed Rahmani[1,2,†], Dries Decap[2,†], Jan Fostier [2*,‡], and Kathleen Marchal [1,2,3*,‡]

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, Gent, Belgium, [2]Department of Information Technology, IDLab, Ghent University—imec, Gent, Belgium, and [3]Department of Biochemistry, Genetics and Microbiology, University of Pretoria, Pretoria, South Africa

*To whom correspondence should be addressed. Tel. +32 9 331 49 42. Email: jan.fostier@ugent.be (J.F.); kathleen.marchal@ugent.be (K.M.)

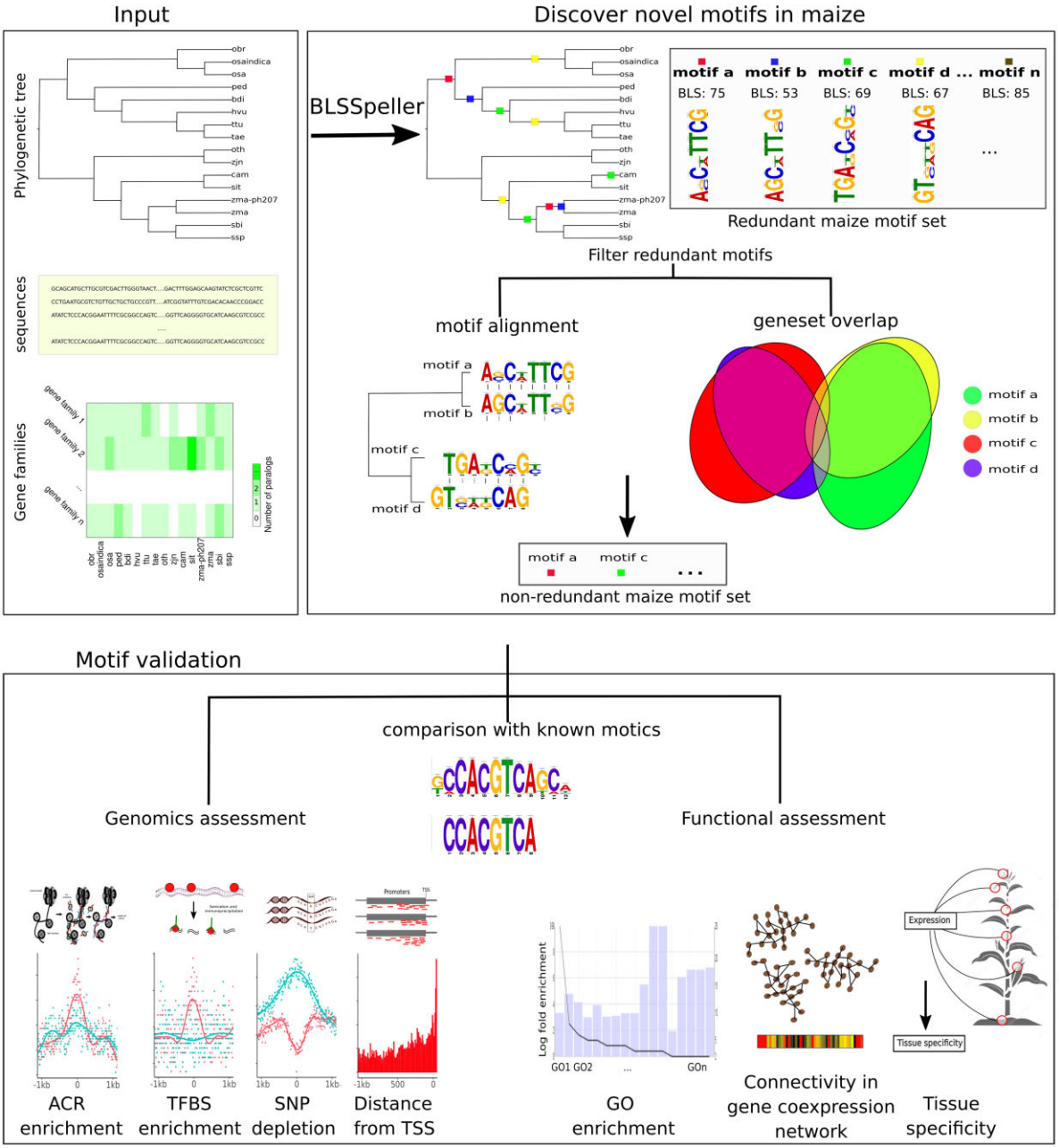
[†]These authors contributed equally to this work.
[‡]These authors should be regarded as Joint Last Authors.

Received 18 February 2022; Editorial decision 21 July 2022

Abstract

With the decreasing cost of sequencing and availability of larger numbers of sequenced genomes, comparative genomics is becoming increasingly attractive to complement experimental techniques for the task of transcription factor (TF) binding site identification. In this study, we redesigned BLSSpeller, a motif discovery algorithm, to cope with larger sequence datasets. BLSSpeller was used to identify novel motifs in *Zea mays* in a comparative genomics setting with 16 monocot lineages. We discovered 61 motifs of which 20 matched previously described motif models in Arabidopsis. In addition, novel, yet uncharacterized motifs were detected, several of which are supported by available sequence-based and/or functional data. Instances of the predicted motifs were enriched around transcription start sites and contained signatures of selection. Moreover, the enrichment of the predicted motif instances in open chromatin and TF binding sites indicates their functionality, supported by the fact that genes carrying instances of these motifs were often found to be co-expressed and/or enriched in similar GO functions. Overall, our study unveiled several novel candidate motifs that might help our understanding of the genotype to phenotype association in crops.

Graphical Abstract



A phylogenetic tree, promoter sequences and gene family membership of 16 monocot lineages, corresponding to 13 species were used as input to identify conserved motifs. The genomic positions for each motif were extracted for maize and redundant motifs were filtered. The resulting motifs were validated by comparing them with known motifs, and by using publicly available genomic and functional data.

Key words: comparative genomics, *de novo* motif discovery, cis-elements, *Zea mays*

1. Introduction

Most of the genetic variation associated with phenotypic variation in plants is located in non-coding regions.[1,2] Hence, uncovering the functional regulatory signals hidden in non-coding regions is crucial to improve our understanding of transcriptional regulation from both fundamental and applied point of view.[3–5] Specific cis-regulatory elements (CRE) are recognized by specific transcription factors (TFs) and play an important role in regulating the rate and

timing of gene expression. Maize is one of the world's most important crops, with a well-studied genome and a large body of experimental data on gene regulation already available, including the experimental identification of TF binding sites,[5,6] and the characterization of active chromatin regions.[7,8] However, like for most crop species, the availability of experimental information is, for technical[7,9,10] and budgetary reasons, limited. Comparing ChIP-seq TF binding site profiles obtained between a wild strain and a strain carrying a mutation in the TF has elucidated binding sites of the studied TF.[11,12] However, those studies are restricted to the profiling of a handful of TFs in maize.[6,13–16] Even one of the most comprehensive large-scale profiling DAP-seq experiments covered only 104 TFs of the ~2,000 annotated TFs in maize, and focused on leaf tissue only.[5] In principle, open chromatin identification methods like ATAC-seq also allow identifying regions with putative regulatory function, but with low resolution (non-specific peaks covering hundreds to a thousand bp) and in a condition-dependent way.[17]

However, experimental information on regulation can be complemented with computational analyses. Comparative genomics to accurately pinpoint the location of functional TF binding sites were already popular in the pre-genomics era[18–21] and have regained scientific interest with the decreasing cost of sequencing and availability of many sequenced genomes.[22–24] Comparative approaches (phylogenetic footprinting, phylogenetic shadowing) that search for motifs that are conserved across genomes have been successfully applied to accurately identify TF binding sites.[25–28] This is particularly true for plants, where TF binding sites are known to be well-conserved across closely related but also more distantly related species and in addition often located in the close neighbourhood of the coding genes.[5,7,26]

Even though many tools for phylogenetic footprinting have been developed, they are not designed to cope with the large numbers of genomes that are currently available. This is unfortunate because finding a motif conserved in a set of sequences by chance decreases with the number of homologous sequences in which the motif is detected and the phylogenetic distances between the sequences in which the motif was found. So in principle, the more sequences can be included during motif detection, the more confidence one can gain in a detected motif. BLSSpeller is a motif detection tool that is unique in exhaustively exploring the full sequence space.[29] The tool also has been redesigned to be able to handle a larger number of input sequences in the comparative analysis. In addition it allows, like many state-of-the-art comparative approaches[19,27,30] to account for the phylogenetic relatedness between the orthologues during its search for conserved motifs and it is able to discover motifs that are conserved in only a subset of the used input sequences. By allowing for an alignment-free search of motifs conserved across species,[29] it can discover binding sites that were relocated during evolution.

BLSSpeller was used to discover novel motifs in *Zea mays*, several of which were supported by complementary sequence-based and/or functional information. Overall, we discovered 20 motifs that perfectly matched previously described motif models in Arabidopsis, together with several yet undescribed motifs generating a useful resource of novel predictions that can complement results from functional genomics studies.

2. Materials and methods

2.1. Datasets used for motif detection

The reference genomes, the structural annotations for 16 species, and orthologous gene families (inferred using the PLAZA integrative method) were downloaded from the PLAZA monocots 4.5.[31] The details about the size of the genome, number of chromosomes, and annotated genes of the species used in this study are provided in Supplementary Table S1. Each gene family consists of all homologous (orthologous and paralogous) genes from all 16 species (41,970 gene families). Gene families with orthologues in less than 3 species were removed, resulting in a total of 21,727 gene families used for motif detection.

To reduce the potential number of false positive motifs, we restricted our search to the 1 kb region upstream of the transcription start site (TSS). For the genes located on the negative strand, the reverse complement of the extracted sequence was considered. Low-complexity and homopolymeric sequences were masked using RepeatMasker[32] prior to performing motif detection.

2.2. BLSSpeller

The core method of our motif detection approach is based on a previous implementation of BLSSpeller.[29] As the original implementation in MapReduce required a lot of intermediate data storage, it was limited in the number of sequences that could be used in the comparative analysis. To overcome this bottleneck and improve its computational performance, BLSSpeller was reimplemented in Apache Spark in order to take advantage of parallel, distributed-memory compute platforms. Compute-intensive parts, e.g., enumerating conserved motifs in a gene family, were implemented in the C++ programming language for efficiency reasons. This reimplementation is 2.3 times faster on average and requires several orders of magnitude less intermediate data storage (Supplementary Table S2), allowing it to use more sequence data. Additional functionality was implemented to identify the location of the conserved motifs in each species.

Below and in Supplementary Fig. S1, we conceptually summarize BLSSpeller. For every gene family (Fig. 1a), a generalized suffix tree, truncated at depth k, is constructed of the promotor sequences of genes within that gene family. The suffix tree is traversed in a depth-first manner to exhaustively enumerate all motifs of a pre-specified length k. The motifs can contain up to a pre-specified number of degenerate characters from the IUPAC alphabet. To this end, information on multiple children of a node in the suffix tree is aggregated. For example, if a degenerate character R (representing A or G) is introduced, child nodes 'A' and 'G' are explored and motif occurrences in both branches are aggregated.

For each length-k motif, the suffix tree reveals in which promoter sequences an instance of that motif appears (Supplementary Fig. S1b). The degree of conservation of the motif within the gene family is expressed by the Branch Length Score (BLS). The BLS takes a value between 0 (motif occurs only in a single genome) and 1 (motif occurs in all species/lineages of the gene family). The BLS is derived from the phylogenetic tree that connects the species/lineages from which the promoter sequences were derived. Each branch in the tree represents a different species and/or lineage represented by a different genomic accession number. The BLS is calculated by finding, in this phylogenetic tree, the minimum spanning tree that connects the relevant subset of species/lineages and summing the weights of the horizontal branches in that tree. If the motif appears in multiple paralogs of the same species/lineage, the branch length to that species/lineage is only accounted for once. If a gene family does not contain a representative orthologue in some of the considered species/lineages, we delete the branches in the tree that correspond to the missing species/lineages and rescale the branch lengths so that the sum of the weights

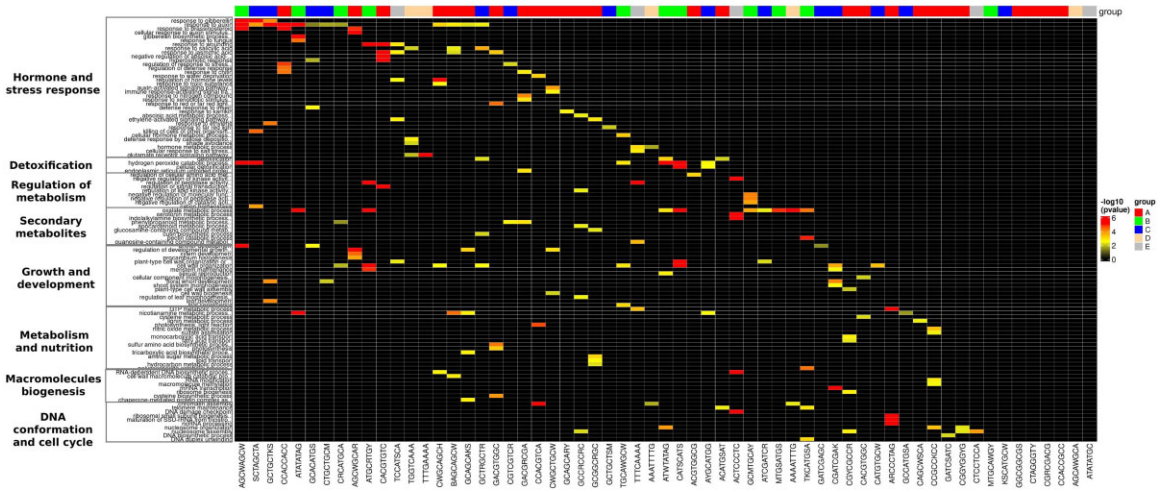


Figure 1. Enriched GO terms (biological process) for the gene sets corresponding to the predicted maize motifs. Each column represents a gene set in maize sharing the indicated motif and each row indicates a biological process that is found enriched amongst the genes in the gene set. The entries indicate the −log 10 *P* values of the GO enrichment. Only the five most significantly enriched GO terms are shown for each gene set. The GO terms were grouped into eight groups based on their biological similarity. The column annotation bar indicates the group membership of each motif as described in 'Prioritization of potential novel motifs in *Zea mays*'.

on the branches of the tree again amounts to 1. Within the tree of a gene family, a higher BLS corresponds to a motif that appears in relatively more species/lineages and/or more distantly related species/lineages. A motif for which the BLS exceeds a predefined BLS threshold in a gene family is said to be *conserved* within that gene family. We use multiple BLS thresholds (i.e., 0.07, 0.13, 0.41, 0.54, 0.65, 0.75, 0.85, and 0.95) (Supplementary Fig. S1b) to also allow the discovery of motifs that are conserved only in a subset of the species/lineages and therefore might have a lower absolute BLS.

Subsequently, we calculate the *recurrence score* (referred to as the 'confidence score' in the original publication[33]) for every motif and BLS threshold (Supplementary Fig. S1d). The recurrence score is calculated as 1 – (*expected recurrence* of the motif at the considered BLS threshold)/(*recurrence* of the observed motif at the considered BLS threshold). Here, the recurrence of a motif equals the total number of gene families in which the motif is conserved. The expected recurrence of a motif is estimated as the median number of gene families in which motifs with the same nucleotide content as the observed motif are conserved. A recurrence score of 0.90 for a given BLS threshold means that the observed motif is conserved (at that BLS threshold) in 10 times as many gene families than expected.

The recurrence and expected recurrence of a motif are computed as follows. As described before, we use the suffix tree to exhaustively enumerate all motifs and we create a binary matrix indicating for each motif (rows of the matrix) whether it meets a certain BLS threshold (columns of the matrix). This procedure is repeated for all gene families (Supplementary Fig. S1c). The matrices of all gene families are aggregated into a single matrix, where each matrix element now corresponds to the recurrence of a certain motif at a certain BLS threshold. To calculate the expected recurrence of a motif, we group all motifs with the same nucleotide content and refer to these as a nucleotide content group. We extend this list by adding a count of 0 for all motif permutations that are not present. Next, per nucleotide content group and per BLS threshold, the expected recurrence is computed as the median value among all motifs in the nucleotide content group (Supplementary Fig. S1c). This median value can be 0.

The software can be obtained at https://bitbucket.org/dries_decap/bls-speller-spark (December 2021, date last accessed). BLSSpeller was used to discover 8 bp conserved motifs with at most three degenerate sites in the promoter sequences of 16 monocot lineages (obtained from 13 species). Considering the fact that the median length of experimentally identified cis-elements is 8 bp,[34] we opted for motifs of length 8, as this gives the best balance between detecting biologically relevant motifs while maintaining computational tractability. To obtain a set of reliable motifs, we selected motifs with a recurrence score of at least 0.9 in at least one of the considered BLS thresholds. This resulted in 1,295 motifs with 2,320,402 instances. It is this core method that has been evaluated in Supplementary Fig. S4.

2.3. Identifying motif instances in *Zea mays*

To identify the instances of these motifs in maize, we consider for each motif the lowest BLS threshold at which the corresponding recurrence score is 0.9 or higher (Supplementary Fig. S1d). We then identify all gene families in which this motif has a BLS that meets the set threshold. The motif instances (genomic locations) in maize are selected. As such, we obtained 1,292 motifs with at least one instance in maize (out of 1,295 motifs identified among all species, 3 motifs have no instances in *Z. mays*).

Because BLSSpeller is an exhaustive approach, it will output highly similar motifs that only differ from each other in a limited number of (degenerate) characters or that have largely overlapping instances. To reduce the level of redundancy, we removed motifs with fewer than 20 or more than 3,000 instances—1% of the motifs with extreme number of instances—as motifs with instances in too few or too many unique genes are either hard to validate or are likely correspond to general TF binding sites (25 motifs were removed). To remove redundancy among the remaining motifs, they were compared in a pairwise manner. The criteria to decide whether two motifs were sufficiently similar to be considered redundant are (i) the pairwise alignment distance between the two motifs and (ii) the degree of overlap among the genes that contain instances of these motifs.

To identify the pairwise alignment distance, the motifs were sorted by their number of degenerate sites (higher to lower) and subjected to an all to all mutual pairwise alignment. The distance between motifs is defined using the following cost model in the pairwise alignment: a match score of 0, a mismatch penalty of 1 and an indel penalty of 1.

The pairwise degree of overlap in genes that contain an instance of two considered motifs is determined as follows:

$$\text{Degree of overlap in genes} = \frac{N_i \cap N_j}{\min(N_i, N_j)},$$

where the numerator indicates the number of genes in maize containing an instance of both motifs *i* and *j*, and the denominator indicates the minimum of the number of genes in maize that contain instances of motifs *i* and *j*, respectively.

Motif '*i*' is considered redundant if its alignment distance to motif '*j*' is less than 5 and the degree of overlap in genes is larger than 0.5. This indicates that the smallest gene set of a motif is contained for at least 50% in the larger gene set of another motif. The threshold of smaller than 5 has been chosen based on the distribution of the similarity observed between random motifs with a similar nucleotide composition as the true motifs. Only 5% of random motifs show a pairwise distance < 5 (Supplementary Fig. S2a). The threshold on the gene set overlap was set at 0.5 and reflects a trade-off between stringency and allowance for a larger than expected overlap between genes (Supplementary Fig. S2b) due to combinatorial regulation. The redundant motifs are removed from the sorted list of motifs in a greedy manner, hereby keeping for each similar set of motifs the most degenerate one (as this one contains most if not all the instances of the other motif). This resulted in 61 non-redundant motifs (referred to as non-redundant motifs hereafter, Supplementary Appendix S1).

For downstream analyses that were performed at the level of the motif instances (overlap of motif instances with chromatin regions and ChIP-seq peaks or when assessing the degree of polymorphism), overlapping instances (minimum overlap in bp = 5) were removed. This is done in order to prevent that the same instance would contribute multiple times to the analysis. Overlapping instances were removed as follows: if two motifs contain overlapping instances, we retain the instance for the motif that covers the highest number of genes and remove the instances of the other motifs that overlapped with the selected motif. This resulted in 50,354 non-overlapping instances for 61 non-redundant motifs.

2.4. Generating random motifs and random instances in *Zea mays*

Here we generated a set of random motifs and their instances to be used in the comparative downstream motif validation analysis. First, the GC content was inferred from the instances of 61 non-redundant motifs in maize. Then, 1,000 random 8-bp *k*-mers were generated using IUPAC DNA codes with the same GC content with at most three degenerate sites. Simulating motifs like this ensures that the random motifs share properties comparable with those of our predicted motifs: i.e., having the same GC content and at most three degenerate sites.

To identify random maize motif instances, we identified all gene families in which a given random motif occurs. Like for the predicted motifs (see Datasets used for motif detection), we only considered gene families that have sequences of at least three different species (at least sequences of three species should be included). From the maize sequences, the random motif instance was extracted irrespective of the BLS score of that family. Subsequently, on the obtained motif instances, the same redundancy filtering criteria mentioned above were applied: random motifs with instances in more than 3,000 or fewer than 20 different maize genes were removed, and overlapping instances were filtered as we did for the predicted motifs. Of the remaining random motifs, 61 random motifs with a pairwise distance higher than 4 to any of the conserved motifs were randomly selected. In this way, the random motifs underwent the same filtering criteria as the predicted motifs while being sufficiently different from any of the predicted motifs.

2.5. Comparing motifs with the Arabidopsis motif compendium

The Arabidopsis DAP-seq motif compendium was obtained from O'Malley[35] and predicted consensus motifs were compared using the Tomtom tool from the MEME suite.[36] An adjusted *P* value smaller than 0.05 was used to determine significant similarity between the compared motifs.

2.6. Tissue specificity (Tau index)

To calculate the tissue specificity index of a gene, we used expression datasets profiled in maize from eight different tissues.[37] RNA-seq raw sequencing reads were downloaded from SRA.[38] Read quality was assessed using FastQC.[39] Remaining adapters, low-quality bases (Phred quality score < 20) and reads shorter than 50 bp were filtered using Trimmomatic.[40] The cleaned reads were aligned to the *Z. mays* B73 AGPv4 genome assembly[41] using the STAR software.[42] FeatureCount[43] was used to quantify expression as count values using the annotation (Zea_mays.B73_RefGen_v4.49.gtf) provided by EnsemblPlants. Only uniquely mapped reads were considered for expression quantification. The count data were normalized for library size and gene length differences using TMM normalization implemented in the edgeR package.[44] Tissue-specific expression of genes was assessed by the Tau index.[45] The Tau index was calculated as follows:

$$\text{Tau} = \frac{\sum_{i=1}^{n}(1-\overline{x}_i)}{n-1}$$

$$\overline{x}_i = \frac{x_i}{\max_{1\le i\le n}(x_i)},$$

where x_i is the expression of a gene in tissue *i* and *n* is the number of tissues in which expression was profiled. The higher the Tau index, the more tissue-specific the expression of the gene is.

2.7. GO enrichment

Functional gene annotation for *Z. mays* B73 AGPv4 was downloaded from www.maizegdb.org. GO enrichment, limited to the 'Biological Process' ontology, was performed using the topGO package.[46] *P* values were adjusted by the BH method using the p.adjust function in R. From the result, the GO terms with more than 1,500 annotated genes or less than 3 significant genes, and adjusted *P* value > 0.01 were removed. After filtering redundant GO terms, the five most significantly enriched GO terms for each cluster are visualized using the pheatmap function in R.[47]

2.8. Co-expression analysis

To perform co-expression analysis, we used the datasets that have been collected and processed by Zhou et al.[48] (Supplementary Table S3). In this study, a comprehensive compilation of RNAseq data

from studies in maize with more than 20 samples spanning different genotypes, tissues, and conditions was made. Lowly expressed (FPKM < 1 in more than 90% of samples), low variance genes, tRNA, and ribosomal genes were filtered out from each dataset. For each dataset, the rank matrix was derived from the gene–gene Pearson correlation matrix calculated on the log2-transformed FPKM values. The rank matrix was transformed to the mutual rank (MR) matrix using the following formula:

$$\mathrm{MR(AB)} = \sqrt{\mathrm{rank}_{(\mathrm{A}\rightarrow\mathrm{B})} * \mathrm{rank}_{(\mathrm{B}\rightarrow\mathrm{A})}},$$

where rank (A → B) is the rank of the correlation of gene B with gene A as compared with its correlation with all other genes.[49] Then the MR matrix was converted to the probability matrix (P) using:

$$P_{(\mathrm{AB})} = \mathrm{e}^{-(\mathrm{MR}_{(\mathrm{AB})}-1)/10}.$$

The P matrix was converted to the co-expression graph after removing P values smaller than 0.05. The connectivity score between nodes (genes) in each network was calculated using the Katz index.[50,51] The Katz index assesses connectivity of two nodes in a graph by exploiting the neighbourhood of the nodes. When calculating the connectivity between two nodes, it considers all paths in the graph that connect the two nodes, but favours shorter paths by assigning them a higher weight. Here we considered all the paths connecting two nodes with a maximum length of three, weighted according to their path length to calculate the similarity between any two nodes in each co-expression network. The Katz index can be calculated and normalized for the *degree* of the connecting genes (normalized Katz index) as follows:

$$\mathrm{Katz\_index} = \alpha(A) + \alpha^2(A)^2 + \alpha^3(A)^3$$

$$\mathrm{normalized\_Katz}_{ij} = \frac{\mathrm{Katz\_index}_{ij}}{\sqrt{d_i * d_j}},$$

where the 'A' is the adjacency matrix, α is a parameter to weight the order of the neighbours (set to 0.3), and 'd' indicates the degree for a gene 'i' and 'j'. The normalized Katz index ranges between 0 and 1 (highly skewed towards zero). A high value indicates high connectivity. For the normalized Katz index, ideally a distribution with an exponential decay is expected for a random gene set. However, distributions of normalized Katz indexes for gene sets sharing a random motif are indicative of some residual connectivity in the co-expression networks which is to be expected, given the large connectivity in the co-expression network (Supplementary Fig. S3). However, we observe that the majority of pairwise connectivity scores for random gene sets fall below 0.05. Therefore, we considered 0.05 as the threshold to distinguish between random and non-random connectivity (see below).

To assess whether genes that share the same motif in maize were more connected in each of the co-expression networks than expected by chance, we first extracted the sets of genes that shared the same motif. A gene set is here defined as a set of genes that share an instance of the same motif. Subsequently, all pairwise Katz indices between genes in a set were assessed and the number of times the Katz index was above a predefined threshold (0.05) was counted, where 0.05 was defined based on the distribution of the Katz score of random gene sets. The obtained count was referred to as the gene set score. For each gene set, 1,000 random gene sets of the same size were obtained by randomly selecting genes from the co-expression network. For each of the random gene sets, the gene set score was calculated as described above. The distribution of these gene set scores obtained for the 1,000 random sets was used to construct the null distribution, which followed a normal distribution after log2 transformation. The parameters of this null distribution were estimated using maximum likelihood. The ccdf (complementary cumulative distribution function) was used to obtain the P value of the observed gene set score given null distribution. A small P value indicates that connectivity between the genes of the gene set corresponding to a certain predicted motif is significantly higher than expected by chance. This analysis was performed for each co-expression network separately.

2.9. Overlap with active chromatin regions and degree of polymorphism

The ACRs, ChIP-seq peaks, and SNPs were downloaded from the following sources: 165,913 non-overlapping 500 bp accessible chromatin regions (ACRs), integrated over more than 50,000 single cells were obtained from Marand et al.[7]; 144,890 non-overlapping TF binding integrated from a ChIP-seq study covering 104 TFs obtained from Tu et al.[5]; SNPs for maize were obtained from Imputed HapMap 3.2.1 (uplifted to B73 AGPv4).[52] For the active chromatin and ChIP-seq regions (ACRs and ChIP-seq), sequences were selected that covered up to 1 kb up and downstream of the centre of the experimental ACR or ChIP seq peak to include the flanking regions. We subsequently assessed which fraction of the total instances of the 61 motifs overlapped with the sequences contained in a window centred at a pre-specified position up or downstream of the active chromatin and ChIP-seq regions. For the plots of the individual motifs (Supplementary Figs S8–S12), we performed the same analysis, but only focusing on the instances of one particular motif.

To assess the fraction of SNPs located at the location of the motif instances and in their flanking regions, we counted the number of SNPs in a sliding window of 8 bp that starts at the location of the motif instance and that moves up to 1 kb up and downstream of motif instance locations and divided this by the total number of SNPs occurring in the entire sequences considered for the above-mentioned analysis. The GenomicRange package[53] was used to count the overlap between the windows that contained SNPs, ChIP-seq or ACR regions and the windows containing motifs and the result was visualized using ggplot2.[54] The same analyses were performed for both predicted and random motifs.

3. Results

3.1. BLSSpeller, a comparative method to perform comparative motif detection

We used BLSSpeller[29] to identify conserved motifs in a comparative genomics setting and validated our predictions by means of publicly available genomics and functional data. To identify motifs in a comparative setting, we obtained from PLAZA[31] gene sequences of 16 monocot lineages derived from 13 species, including two lineages of maize, two lineages of rice, and two lineages of wheat (Supplementary Table S1). We could include multiple accessions of the same species as (i) BLSSpeller accounts for the phylogenetic distance between the input sequences and (ii) for each of these accessions a good genome assembly and annotation was available, allowing to include each of the accessions in the phylogenetic tree. BLSSpeller weighs the contribution of each sequence to the motif detection (so it considers the genomes of the accessions as 'different' species): sequences that are closer in the tree will contribute less

information to the motif detection than sequences from more distant accessions. Genes were grouped into gene families where each gene family can contain both paralogs and orthologues. BLSSpeller was used to search in the promoter sequences of gene families with at least three species for conserved motifs.

Details on BLSSpeller can be found in the Materials and methods. Briefly, BLSSpeller scores, for all possible motifs of a pre-specified length (e.g., 8 bp), their degree of conservation within each gene family, denoted by the BLS.[33] Conceptually, the BLS of a motif expresses the fraction of promoter sequences in a gene family that contain the motif, weighted by the relative phylogenetic distance of the lineages from which they were extracted. When the BLS of a motif within a gene family exceeds a pre-specified BLS threshold, the motif is said to be *conserved* within that gene family. We used different thresholds on the BLS to also enable the identification of motifs that were conserved in only a subset of the species of a gene family. Conservation is soft constrained by allowing for some degeneracy in a motif. The BLS assigns more relevance to conservation in different species/lineages (orthologues) than conservation within a species/lineage (paralogs).

To reduce false-positive predictions, an additional selection criterion is imposed: given that biological processes tend to be regulated by multiple genes, a true motif is more likely to be conserved in multiple gene families. Therefore, we identified as the more reliable motifs, those predictions that were conserved within more gene families than random motifs of the same length and nucleotide composition. The degree to which a predicted motif is conserved within more gene families than a random motif is reflected by its recurrence score (see Materials and methods) and is computed for multiple BLS thresholds. Motifs with a recurrence score of 0.9 or higher for any of the considered BLS thresholds were retained. A recurrence score of 0.9 for a given BLS threshold means that the observed motif is conserved (at that BLS threshold) within 10 times as many gene families than expected. This filtering step largely reduced the number of motif candidates and resulted in a final list of 1,295 motifs.

Supplementary Fig. S4a shows the effect of the filtering based on the recurrence threshold. It shows that the fraction of retained motifs after applying the recurrence-based filtering increases with the threshold on the BLS, indicating that motifs with a lower BLS are less likely to occur significantly more often across families than random motifs and hence that these motifs are indeed more likely to be spurious.

To investigate the impact of the number of paralogs in a gene family on the BLS of the motifs in that family, we plotted per gene family the number of motifs with a high BLS (> 0.95) as a function of the average number of paralogs over species within a gene family, before and after applying the filtering based on the recurrence score threshold (Supplementary Fig. S4, panels b and c). Supplementary Fig. S4b shows that even though the BLS scheme downweighs the impact of paralogs, it is not entirely independent of the average number of paralogs present in gene families as the number of predicted motifs increases with an increase in the average number of paralogs. This is to be expected as a higher number of paralogs implies a larger sequence space and hence a higher probability of detecting by chance a conserved motif with a high BLS. As shown in Supplementary Fig. S4c, recurrence filtering removes many of these likely spurious motifs detected in motif families with a high average number of paralogs.

3.2. Identifying motifs and instances relevant to *Zea mays*

To select from the motifs predicted by BLSSpeller those that are relevant in maize, we assessed for each motif whether gene families exist that contained a motif instance in the corresponding maize sequences. This resulted in a final selection of 1,292 motifs and 2,320,402 instances. However, many of these motifs are redundant as for instance the same motif can be recovered with a different level of degeneracy. Therefore, redundant motifs were removed based on the degree of similarity between the motifs and the degree to which the motifs covered similar genes (see Materials and methods). This resulted in a final list of 61 non-redundant motifs, each of which with at least 20 instances in different maize genes (Supplementary Appendix S1). The average GC content of these motif instances was 62% as compared with an average GC content of 45% for the maize promoter regions. For integration with complementary genomics data (see below), also redundant motif instances were removed, and a set of random motifs and instances was generated that has the same nucleotide content and distribution of the number of instances in maize as the predicted motifs (see Materials and methods).

3.3. Predicted motifs are associated with processes known to be conserved across species

Genes with the same motifs are expected to be co-regulated and hence involved in the same biological processes. To assess whether the gene sets containing the same motif are indeed functionally similar, we performed GO enrichment. For each motif, a representative gene set was compiled by taking for each gene family in which the motif occurred, the genes of maize that contained an instance of the considered motif. The gene sets corresponding to 53 out of the 61 motifs were enriched for at least one biological function (adj. P value < 0.01) (Fig. 1). For the genes sets corresponding to the random motifs this was only true for 15 out of the 61 random motifs (Supplementary Fig. S5). The level of enrichment that was observed for random motifs is higher than what one would expect intuitively. This is mostly due to the fact that we followed a rather conservative approach to simulate the random motifs. We intentionally selected from random motifs those that mimic the predicted motifs with respect to their GC content and the distribution of number of associated genes. Enriched GO terms of the representative gene sets related to hormone and stress response, detoxification, regulation of metabolism, secondary metabolites, growth and development, metabolism and nutrition, macromolecules biogenesis, and DNA conformation and cell cycle (Fig. 1). Although overall, the gene sets of the different motifs covered a variety of biological processes, processes related to 'hormone and stress response' were most frequently found enriched, indicating that, as expected, the genes carrying the predicted motifs are involved in key processes known to be conserved across species.[55–57]

3.4. Validation of the predicted maize motifs by comparing with Arabidopsis motifs

To validate the motifs, we compared the 61 maize motifs with experimentally verified motifs in *Arabidopsis thaliana* obtained from a DAP-seq experiment.[35] The study of O'Malley et al. provides one of the most comprehensive compendia of experimentally generated TF binding sites in plants, covering binding sites for 529 TFs. Our predicted maize motifs match significantly better to known Arabidopsis motifs than random motifs. Of the 61 maize motifs, 20 motifs perfectly match (adj. P value < 0.05) known Arabidopsis motifs, whereas the same was true for three random motifs only (Supplementary Fig. S6). Based on these similarities, the retrieved motifs cover Arabidopsis binding sites for a diverse set of TFs,

including members of C2H2, AP2, bHLH, NLP, ERF, ARE, HMG, and ABI (Supplementary Table S4).

Subsequently, we assessed the functional similarity between our predicted motifs and their corresponding matching motifs in Arabidopsis. Hereto, we assumed the function of the Arabidopsis motif was the same as the function of the TF associated with the motif (TAIR website). For the maize motifs, the function was inferred by performing GO enrichment on the representative gene set (set of genes that share an instance of the same motif) for each of the motifs. Supplementary Table S4 shows that at least for some motifs, a clear similarity can be detected between the inferred functions of the predicted motifs in maize and their matching motifs in Arabidopsis. The large overlap between our predicted motifs and experimentally validated motifs supports the relevance of our predictions.

Although for most of the predicted motifs, a unique one to one match with an Arabidopsis motif was found, the Arabidopsis motif, NLP_tnt.AtNLP4, shows a high similarity with four different maize motifs (GCAGCARY, GCAGCAKS, AGCWGCAR, and BAGCAGCW) (Supplementary Table S4), suggesting these motifs might still be redundant or might represent binding sites for different TFs with similar functions in maize. Indeed the genes having an instance of each of the four aforementioned motifs are all enriched in 'metabolism and development', consistent with the AtNLP4 function in Arabidopsis which involved in cell elongation and response to nitrate.

3.5. Genes sharing instances of the same predicted motif are co-expressed

Assuming that sharing a motif implies co-regulation and hence co-expression, we assessed whether maize genes that have instances of a similar motif also exhibit a similar expression profile. Hereto, we used a previously published expression compendium of maize,[48] comprising 8 developmental stages of the same genotype, 28 tissue-specific datasets sampled from the same tissue from multiple inbred lines, 5 tissue-genotype datasets originating from multiple tissues of specific inbred lines, and 4 datasets from recombinant inbred populations (see Supplementary Table S3 for details).

To measure similarities in co-expression, we built co-expression networks. The different studies in the compendium each capture gene expression associations in different biological contexts. The compendium is therefore highly unbalanced in the number of matching tissues/conditions. To avoid that the co-expression analysis would be biased by this unbalance, we built a co-expression network for the data of each study separately, resulting in 45 co-expression networks. To assess whether sets of genes that share instances of the same motif were also co-expressed, we measured the degree to which they were connected in each of the 45 co-expression networks (see Materials and methods). The results show that for the majority of the gene sets sharing a motif, the genes are significantly more connected than random gene sets of the same size (Fig. 2 and Supplementary Fig. S7). This observation further corroborates the functionality of our predicted motifs.

The degree to which the co-regulated gene sets displayed a connectivity in the co-expression network increased with the number of samples in the dataset for which the co-expression network was built and with the type of samples that were profiled: In contrast to networks derived from tissue profiling experiments, recombinant inbred line (RIL) networks did not capture strong connectivity between the genes sharing a motif, despite the fact that they were derived from experiments with relatively many samples (Fig. 2).

Like comparing with known motifs, the co-expression analysis showed that some motifs might either be variants of one true binding site or represent binding sites of closely related TFs. For example, gene sets with instances of respectively the motif ATWTATAG or ATATATAG in their promoter regions are connected in the same co-expression networks (Fig. 2). This high connectivity in the same co-expression networks indicates that those motifs probably represent the same binding site and should have been merged into a single motif. Similar observations were made for the motifs [AGCWAGCW, TGCAWGCA], [GCTRGCTR, SCTAGCTA], [CRCATGCA, TKCATGSA, ACATGSAT], and [GCTGCTKS, GCAGCAKS].

3.6. Tissue-specific genes have a larger number of predicted motifs

The timing, level, and tissue-specificity of gene expression depends on the presence of CREs that recruit specific TFs in response to tissue demands.[58] We investigated whether an association exists between the degree to which a maize gene displays a tissue specific expression pattern and the number of motifs that were predicted to have an instance in the promoter of the gene. To assess the level of tissue-specific expression of a gene, we used RNA-seq expression data from eight different tissues profiled in maize B73 line.[37] Tissue-specific expression was measured using the Tau index[45] (Materials and methods). Figure 3a shows that in general, the more a gene has instances of different motifs, the higher the gene's degree of tissue-specific expression (higher Tau index). This observation is in line with literature, indicating that tissue-specific genes have evolved numerous unique binding sites during evolution and that the combinatorial effects of several binding sites may be critical to provide the proper response to developmental, condition, and tissues-specific demands.[59,60] Unlike the predicted motifs, for random motifs no relationship between the number of instances of different motifs and the degree of tissue-specific expression (Tau index) was observed (Fig. 3c).

3.7. Location of the predicted motif instances is biased towards the TSS

The region in the close vicinity of the TSS is known to play a central role as binding site for TFs.[61] Therefore, the specific positional relationship between TF binding sites and the TSS has widely been used to predict the validity of TF binding sites.[62–64] To validate the motifs predicted by BLSSpeller, we assessed whether the instances of our predicted motifs in maize were more frequently located around the TSS site than instances of random motifs. Figure 3b shows that this is indeed the case, as predicted motifs are highly enriched in regions close to the TSS, while random motifs are evenly distributed across the 1 kb upstream region (Fig. 3d). This enrichment towards the TSS has been observed for most of the predicted motifs (Supplementary Figs S8–S12). This result further supports the biological validity of our motif predictions in maize.

3.8. Predicted motif instances overlap with open chromatin regions and TF binding sites

To further validate our predicted motifs, we assessed to what extent instances of the predicted motifs were occurring more frequently than expected by chance in nucleosome-depleted regions and/or in regions known to functionally bind TFs. Hereto, we used the data from a study that profiled ACRs in maize at single-cell resolution

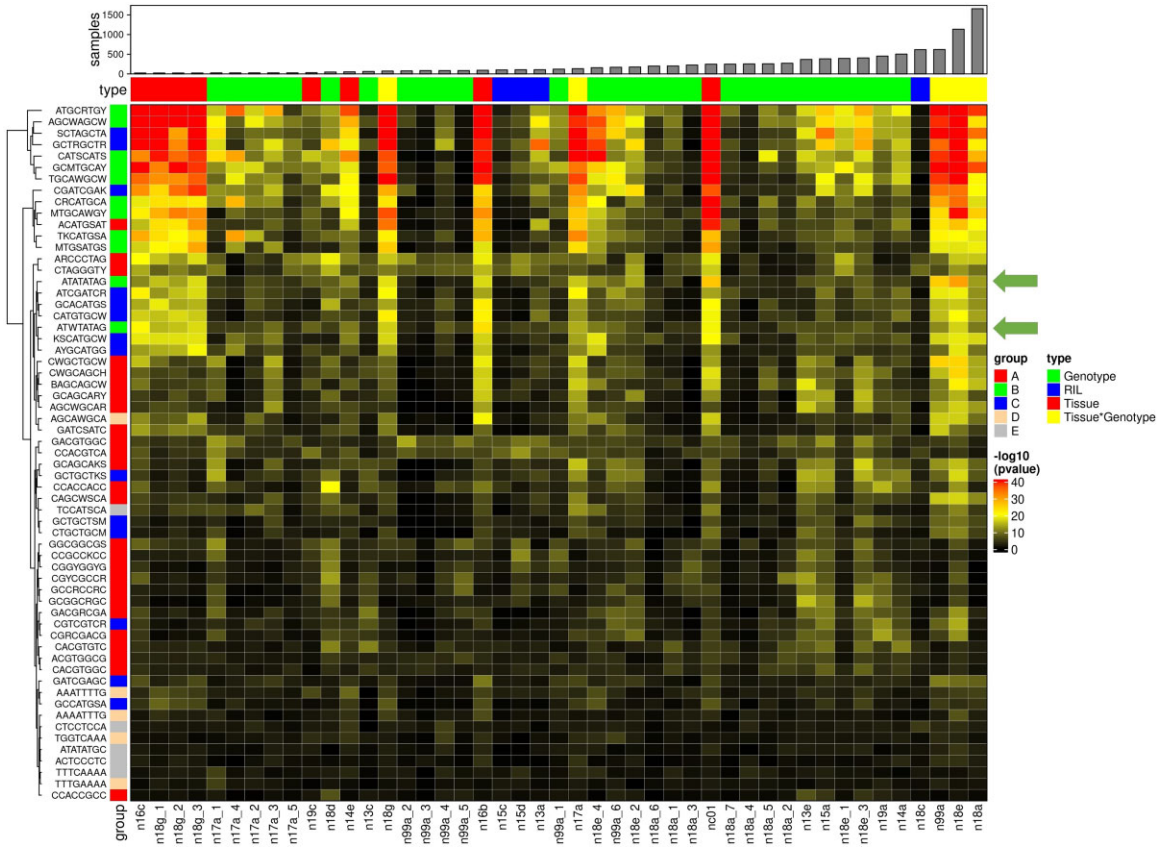


Figure 2. Connectivity of gene sets sharing the same predicted motif in the co-expression networks. Rows display the different motifs for which a gene set is considered. Columns display the different experimental conditions for which a co-expression network was constructed (sorted by sample size). Entries indicate the –log 10 *P* values of observing the same connectivity in a co-expression network by chance as was observed for the gene sets that share the same predicted motif. The coloured annotation bar on columns indicates the type of experimental dataset, and the top annotation bar indicates the sample size of each experimental dataset. The row annotation bar indicates the group membership of each motif described in 'Prioritization of potential novel motifs in *Zea mays*'. The two arrows on the right indicates the two similar motifs (ATWTATAG and ATATATAG) with the same connectivity behaviour.

using more than 50,000 cells from six organs using scATAC-seq,[7] and a comprehensive ChIP-seq study in maize that profiled the putative DNA binding sites for 104 TFs in leaf tissue.[5]

Figure 4a and b shows that a significant overlap exists between our predicted motif instances and respectively the nucleosome-depleted regions and ChIP-bound regions. Both panels show that our predicted motif instances occur more frequently in the peaks of the ACRs and ChIP-seq data than random motif instances. For ACRs, also the random motif instances are enriched near the centres of the experimentally identified ACRs, but not as extreme as what we observe for the predicted motif instances. The random motifs were generated to have sequence properties similar to the ones of the predicted motifs. Because ACRs have a high GC content,[7] the enrichment of random motif instances near the centre of the ACR peaks indicates that the enrichment of the predicted motif instances in the ACRs is partially the result of their GC content. In contrast to what is observed for the ACR peaks, random motif instances are depleted in the proximity of the ChIP-seq binding sites. These observations further confirm the relevance of our predicted motifs.

The same overlap between the predicted motif instances and active chromatin regions and TF binding regions (ACRs and ChIP-seq binding sites), but then for each motif separately, can be seen in Supplementary Figs S8–S12. The results show that for the majority of the predicted motifs, a trend similar to the one displayed in Fig. 4 is observed. The trend observed in Fig. 4a and b is therefore not driven by the pattern of only a few predicted motifs.

3.9. Predicted motifs are under selection

True binding sites tend not to tolerate mutations that interfere with their functionality. To further validate whether this was also true for the predicted motifs we compared, the frequency with which these predicted motifs accumulated SNPs in their binding sites and flanking regions (1 kb up and downstream of the locations of the predicted motif instances). Figure 4c shows that the regions centred around the predicted motif instances are more depleted in SNPs than the regions centred around random motif instances. Supplementary Figs S8–S12 show the same pattern at the level of the individual motifs. The results presented in Supplementary Figs S8–S12 clearly show that some motifs are under clear selective constraints and therefore likely functionally relevant (motifs 1, 4, 5, 8, 9, 10, and 11).

Overall, the degree of polymorphism in the regions flanking the random motif instances was lower than in the regions centred around the random motif instances (Fig. 4c) and drops faster than what was observed for the regions flanking the predicted motif instances. Repeating the analysis using different sets of random motifs showed that this pattern was robust. In general, this decrease in polymorphism for regions flanking of predicted and random motif instances can be explained by the fact that moving away from motifs

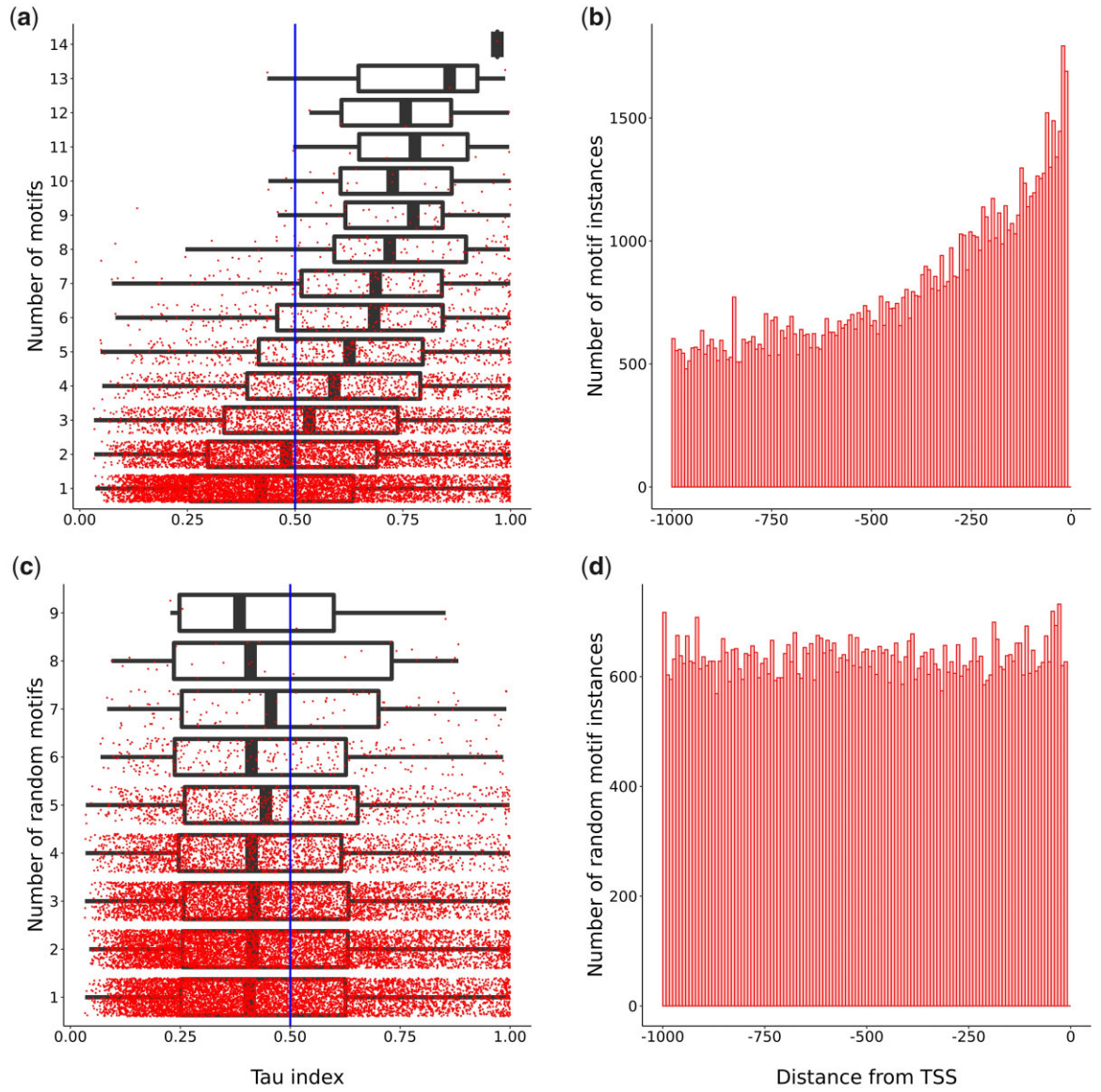


Figure 3. Tissue specificity of the expression of gene sets carrying instances of the predicted motifs in maize and position of the predicted motifs relative to the TSS. Panels (a) (predicted motifs) and (c) (random motifs) show how the tissue specificity of gene expression (Tau index) is associated with the number of motifs that have at least one instance in the gene. X-axis: Tau index, y-axis: number of unique (non-redundant) motifs. Panels (b) and (d) show the positions in maize of the motif instances relative to the TSS for respectively the predicted and random motifs. X-axis: distance from the TSS, y-axis: number of motif instances (61 non-redundant motifs) per 8 bp bins across 1,000 bp upstream of the TSS.

(predicted and random) that are located in intergenic regions, increase the chance of reaching genic regions. In those genic regions, a higher level of selective constraint and hence higher conservation and less polymorphism is expected.[65,66] In principle, both negative and positive selection may play a role in explaining the observed lower rate of decrease in polymorphism for flanking regions of predicted motifs than for flanking regions of random motifs. In case of deleterious mutations in motif regions, negative selection indirectly increases the chance of surviving the SNPs in flanking regions by purging sequences having deleterious mutations in motif regions. Under this scenario, a fitness trade-off might exist between selecting sequences with a suboptimal SNP in the flanking regions versus sequences with deleterious SNP in the predicted binding site: the potential suboptimal effect of SNPs in regions flanking a true binding site can be offset by the beneficial effect of not having a deleterious SNP at the position of a true functional motif. Alternatively, some neutral and suboptimal SNPs in flanking regions would be selected for due to their linkage to beneficial SNPs within the true motif regions (i.e., positive selection). Such compensatory and linkage effects cannot be achieved for random motifs and hence they tolerate fewer SNPs in their flanking regions, which are likely to include some genic regions. Although both mechanisms of respectively negative and positive selection could explain the relatively lower depletion of SNPs in the flanking regions of predicted than of random motifs, the contribution of the negative selection is expected to be much higher as mutations with beneficial effects are rare. Overall, the strong depletion of SNP

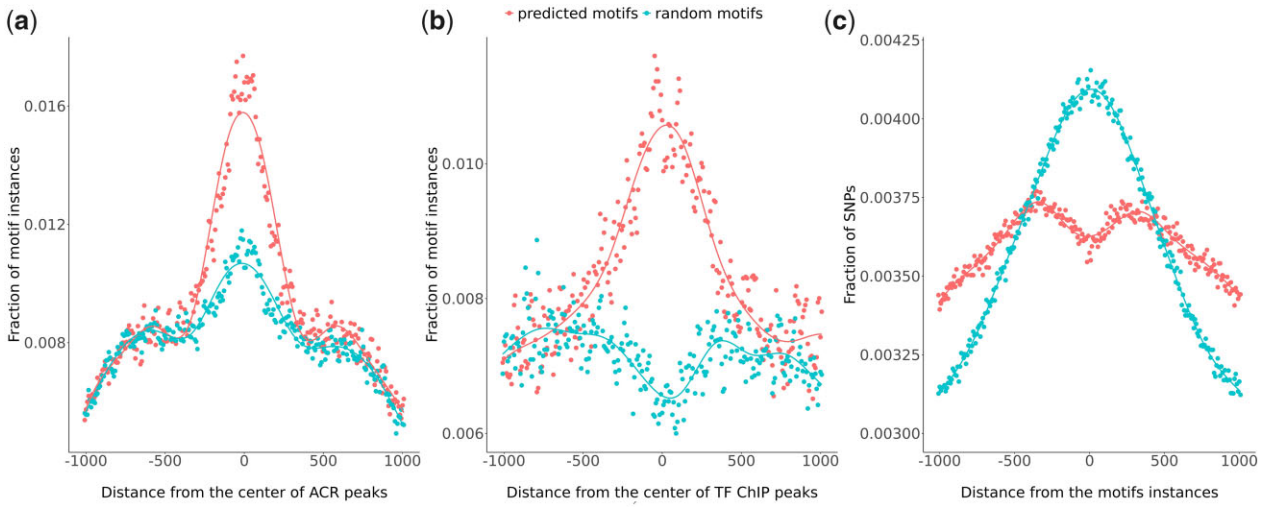


Figure 4. Overlap between the location of the predicted motif instances and respectively ACR, ChIP-seq binding locations and SNPs in maize. (a) Fraction of respectively the predicted and random motif instances that overlap with ACRs and their flanking regions. To include flanking regions, the sequences were extended 1 kb up and downstream of the centre of the open chromatin region. (b) Fraction of respectively the predicted and random motif instances and the ChIP-seq corresponding binding sites of 104 TFs. The ChIP-seq binding sites are extended 1 kb up and downstream of the centre of the binding peak to include flanking regions. (c) Fraction of SNPs that overlap with the location of a motif instance and their surroundings. To include flanking regions, the sequences were extended 1 kb up and downstream of the centre of the location of motif instances of respectively the predicted and random motifs.

close to the predicted motif instances suggests that both natural and artificial selection has resulted in the fixation of variants in the predicted motifs in individuals with favourable phenotypes.

3.10. Prioritization of potential novel motifs in *Zea mays*

Supplementary Figs S8–S12 show the results of the aforementioned genomic and functional assessments at the level of the individual motifs. Overall, most motifs seem to be supported by at least one genomic and/or functional line of evidence or feature. For several predicted motifs, the inferred function based on GO enrichment of the genes sets representative for the motif in maize corresponds to the function of the matching motif-TF pair in Arabidopsis: predicted maize motifs corresponding to gene sets enriched in hormone and stress response, matched the Arabidopsis motifs of known master regulators of response to hormones and stress in plants i.e., C2C2, bHLH34, NLP4, TGA4, ERF13, ABI3, and MYB74 (Supplementary Figs S8–S12). Likewise, motifs of gene sets enriched in growth and development showed high similarity with the motifs of several main players in plant development ERF6, AREB3 bZIP42, GATA20, and ZIM. One predicted motif of which the target gene set was enriched in DNA conformation and cell cycle showed a high similar to the Arabidopsis motif of TBP3 (ARCCCTAG) (Supplementary Fig. S8, motif 1). TBP3 encodes a telomeric DNA binding protein and belongs to single Myb histone gene family, of which the members preferentially bind to double-stranded telomeric repeats.[67]

However, not all predicted motifs could be clearly associated with known TF binding sites in Arabidopsis. To prioritize the most reliable motif predictions, we used additional genomics and functional data to support our findings.

According to their shared supporting features, five groups of predicted motifs were distinguished. The first group of motifs show a high similarity to known Arabidopsis motifs and are particularly strongly supported by all levels of genomic evidence (Supplementary Fig. S8): compared with their flanking regions and random motif instances, the location of the predicted motif instances in this group is strongly centred around the TSS, is depleted for polymorphisms, and shows overlap with active chromatin and ChIP-seq bound regions. These motifs occur in genes that tend to be broadly expressed in many tissues (lower Tau index) and that are enriched in processes that are widely conserved across species such as 'DNA conformation and cell cycle', 'Metabolism and nutrition', and 'Regulation of metabolism' (Supplementary Fig. S8). The more uniform expression behaviour across conditions of these genes might explain why their mutual correlation and hence also connectivity in the co-expression network is more difficult to capture (as their expression might not sufficiently vary across conditions). Figure 5a shows a representative motif from this group of motifs.

Motifs belonging to group 2 also show a high similarity to Arabidopsis motifs and are particularly well supported by the functional evidence and by the enrichment of their motif instances upstream of the TSS (Supplementary Fig. S9). Because their target genes show high variability in expression across conditions (high Tau index), they also show relatively high connectivity in the co-expression networks, representative of several different conditions. Their target genes are highly enriched for 'secondary metabolites', 'hormone and stress response', and 'growth and development', processes for which a high level of polymorphism between individuals has been described.[68,69] The overlap between motif instances and SNPs confirms that indeed the motifs in this group tend to be associated with a rather high level of polymorphism (an increase rather than a depletion of SNPs at the motif position). Such a high level of polymorphism could be associated with natural and artificial selection in maize which benefits from heterosis and is associated with increased fitness. If functionally divergent alleles enable adaptation to different environments, spatially heterogeneous natural selection (balancing selection) might maintain locus-specific polymorphism.[70] This is in line with the fact that genes in this group are enriched for 'secondary metabolites', and 'hormone and stress response' processes of which are responsible for adaptation and are under more relaxed purifying selection or under stronger diversifying selection.[68,71–73] In addition, the less clear support provided by the overlap between motif instances and active chromatin regions

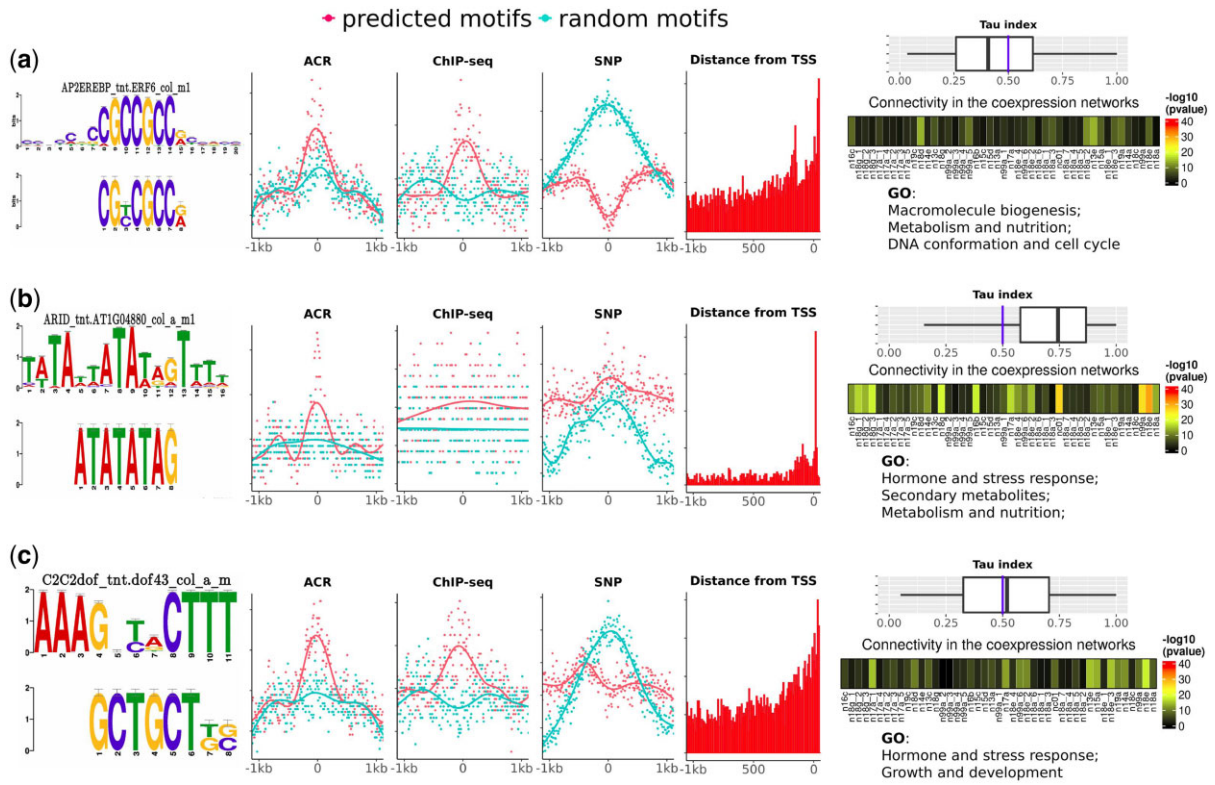


Figure 5. Representative motifs for each group. (a) Group 1: representative of predicted motifs that show a high similarity with a corresponding Arabidopsis motif and that are well-supported by genomic assessments; (b) Group 2: representative of predicted motifs that show a high similarity with a corresponding Arabidopsis motif and that are well-supported by functional evidence; (c) Group 3: novel motifs that do not show any similarity with Arabidopsis motifs, but that are supported by genomics and/or functional evidence. For each row: left panel: most similar Arabidopsis motif (top) and predicted motif (bottom); middle panels: overlap between the motif instances (predicted: red and random: blue) and respectively ACRs, ChIP-seq binding sites, the occurrence of SNPs, and distance from the TSS; right panel: from top to bottom: tissue specificity distribution (Tau index), connectivity in co-expression networks, and GO enrichment for genes having at least one instance of the predicted motif under consideration (A color version of this figure appears in the online version of this article.).

(ACR and ChIP-seq) can be due to the mismatch between the conditions in which the chromatin profiling experiments were performed and the conditions under which the genes carrying the motifs of this group are expressed, which are as shown in Supplementary Fig. S9 rather tissue specific. A representative motif of this group is shown in Fig. 5b.

The third group represents potentially novel motifs that show no match with a corresponding Arabidopsis motif (insignificant *P* value and presence of strong mismatch in at least one position) (Supplementary Fig. S10), but that are well supported by genomic and/or functional assessments. Motifs in this group were sorted based on their Tau index (from least to most tissue specific, Supplementary Fig. S10). Also here, we observe—similar to what we noticed for motifs of groups 1 and 2—that the higher the tissue specificity of the expression of the genes carrying the motif, the more the genes show connectivity in the corresponding tissue-specific co-expression networks and the less they were supported by ACR, ChIP-seq, and SNP evidence. A representative motif of this group is shown in Fig. 5c.

Motifs belonging to group 4 (Supplementary Fig. S11) and group 5 (Supplementary Fig. S12) are less clearly supported by additional evidence. Even though the genes carrying these motifs show enrichment in a certain GO function (groups 4 and 5) and that motifs of group 4 (Supplementary Fig. S11) in addition to this, show similarity to known Arabidopsis motifs, we could not find any support for these predictions through the genomics or expression-based evidence (low connectivity in the co-expression network).

4. Discussion

With the increasing availability of sequenced genomes, comparative analyses to identify TF binding sites are an attractive alternative to complement tedious experimental approaches. In this work, we present BLSSpeller, an exhaustive, alignment-free search for motifs conserved in orthologous sequences of related species. We applied BLSSpeller to identify motifs in *Z. mays* by using a comparative genomics approach using 16 monocot lineages (obtained from 13 species).

Overall, we observed that despite the high 'AT' content of plant promoters, the motifs identified by BLSSpeller are enriched in 'GC'. This is in line with previous studies in plants which also observed that regions in which TF binding sites occur are GC enriched i.e., open chromatin regions[7,74] or experimentally identified binding sites.[5]

Complementing the predictions with publicly available complementary data sources allowed pinpointing several promising motif candidates in *Z. mays*. Overall, we identified several motifs, motifs with a match to known motifs in Arabidopsis, but also novel motif candidates that were supported by additional genomic assessments: the location of the instances of these predicted motifs are occurring more frequently than expected by chance in the regions upstream of the TSS, in active chromatin regions and/or ChIP-seq bound regions. In addition, the instances of these motifs tend to be depleted of SNPs, consistent with the action of purifying selection. Also, functional

assessment could support the motifs: genes that share instances of the same predicted motifs are enriched in similar GO functions and tend to be more often co-expressed than random gene sets (displaying connectivity in the co-expression networks). We observed that there was an inverse relationship between the levels to which predicted motifs were supported by the functional versus the genomic assessments. Genes that were functionally well supported by expression analysis tended to display a more tissue-specific expression and were therefore less supported by the chromatin binding assays conducted under one specific condition. In contrast, processes that were ubiquitously expressed and not condition dependent were less supported by expression connectivity, but well supported by the available chromatin accessibility data. This indicates that the chromatin accessibility landscape varies largely between conditions and is likely a major factor in contributing to tissue-specific expression. It also shows that expression connectivity can more easily be captured for processes that are variably expressed across conditions.

In summary, we show that BLSSpeller, a high-performance motif discovery tool, can successfully be used to predict novel motifs in a comparative setting. Combining such predictions with available genomics and functional data allowed further elucidating transcriptional regulation in *Z. mays*. Although their impact requires further characterization, the provided motifs offer a large and valuable source for further investigation.

Acknowledgement

We thank Hilde Nelissen for her critical insights and useful feedback on the manuscript.

Funding

The work was supported by grants of the Fonds Wetenschappelijk Onderzoek-Vlaanderen (FWO) (3G046318, G.0371.06), UGent BOF (BOF/IOP/2022/045 and 01J06219) and Flanders Innovation & Entrepreneurship (VLAIO, project 'ATHENA', HBC.2019.2528); R.S.R. was supported by a grant from Ministry of Science, Research and Technology, Iran.

Author contributions

Supervision—K.M. and J.F.; method development—D.D. and J.F. Formal analyses and visualization—R.S.R.; validation and application—R.S.R. and K.M.; writing—original draft, R.S.R., D.D., J.F., and K.M.; writing—review and editing, R.S.R., D.D., J.F., and K.M.; funding acquisition—K.M. and J.F.

Conflict of interest

None declared.

Supplementary data

Supplementary data are available at DNARES online.

References

1. Miculan, M., Nelissen, H., Ben, M.H., et al. 2021, A forward genetics approach integrating genome-wide association study and expression quantitative trait locus mapping to dissect leaf development in maize (*Zea mays*), *Plant J.*, **107**, 1056–71.
2. Wallace, J.G., Bradbury, P.J., Zhang, N., Gibon, Y., Stitt, M. and Buckler, E.S. 2014, Association mapping across numerous traits reveals patterns of functional variation in maize, *PLoS Genet.*, **10**, e1004845.
3. Cherry, T.J., Yang, M.G., Harmin, D.A., et al. 2020, Mapping the cis-regulatory architecture of the human retina reveals noncoding genetic variation in disease, *Proc. Natl. Acad. Sci. USA*, **117**, 9001–12.
4. Salvi, S., Sponza, G., Morgante, M., et al. 2007, Conserved noncoding genomic sequences associated with a flowering-time quantitative trait locus in maize, *Proc. Natl. Acad. Sci. USA*, **104**, 11376–81.
5. Tu, X., Mejía-Guerra, M.K., Franco, J.A.V., et al. 2020, Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors, *Nat. Commun.*, **11**, 1–13.
6. Eveland, A.L., Goldshmidt, A., Pautler, M., et al. 2014, Regulatory modules controlling maize inflorescence architecture, *Genome Res.*, **24**, 431–43.
7. Marand, A.P., Chen, Z., Gallavotti, A. and Schmitz, R.J. 2021, A cis-regulatory atlas in maize at single-cell resolution, *Cell*, **184**, 3041–3055.e21.
8. Ricci, W.A., Lu, Z., Ji, L., et al. 2019, Widespread long-range cis-regulatory elements in the maize genome, *Nat. Plants*, **5**, 1237–49.
9. Bartlett, A., O'Malley, R.C., Huang, S.-s.C., et al. 2017, Mapping genome-wide transcription-factor binding sites using DAP-seq, *Nat. Protoc.*, **12**, 1659–72.
10. Gerstein, M.B., Kundaje, A., Hariharan, M., et al. 2012, Architecture of the human regulatory network derived from ENCODE data, *Nature*, **489**, 91–100.
11. Johnson, D.S., Mortazavi, A., Myers, R.M. and Wold, B. 2007, Genome-wide mapping of in vivo protein-DNA interactions, *Science*, **316**, 1497–502.
12. Kheradpour, P. and Kellis, M. 2014, Systematic discovery and characterization of regulatory motifs in ENCODE TF binding experiments, *Nucleic Acids Res.*, **42**, 2976–87.
13. Bolduc, N., Yilmaz, A., Mejia-Guerra, M.K., et al. 2012, Unraveling the KNOTTED1 regulatory network in maize meristems, *Genes Dev.*, **26**, 1685–90.
14. Li, C., Yue, Y., Chen, H., Qi, W. and Song, R. 2018, The ZmbZIP22 transcription factor regulates 27-kD γ-zein gene transcription during maize endosperm development, *Plant Cell*, **30**, 2402–24.
15. Pautler, M., Eveland, A.L., LaRue, T., et al. 2015, FASCIATED EAR4 encodes a bZIP transcription factor that regulates shoot meristem size in maize, *Plant Cell*, **27**, 104–20.
16. Yang, H., Liu, X., Xin, M., et al. 2016, Genome-wide mapping of targets of maize histone deacetylase HDA101 reveals its function and regulatory mechanism during seed development, *Plant Cell*, **28**, 629–45.
17. Yocca, A.E. and Edger, P.P. 2022, Current status and future perspectives on the evolution of cis-regulatory elements in plants, *Curr. Opin. Plant Biol.*, **65**, 102139.
18. Monsieurs, P., Thijs, G., Fadda, A.A., et al. 2006, More robust detection of motifs in coexpressed genes by using phylogenetic information, *BMC Bioinformatics*, **7**, 1–15.
19. Blanchette, M. and Tompa, M. 2003, FootPrinter: a program designed for phylogenetic footprinting, *Nucleic Acids Res.*, **31**, 3840–2.
20. Cliften, P., Sudarsanam, P., Desikan, A., et al. 2003, Finding functional features in Saccharomyces genomes by phylogenetic footprinting, *Science*, **301**, 71–6.
21. Blanchette, M. and Tompa, M. 2002, Discovery of regulatory elements by a computational method for phylogenetic footprinting, *Genome Res.*, **12**, 739–48.
22. Wei, H., Liu, J., Guo, Q., et al. 2020, Genomic organization and comparative phylogenic analysis of NBS-LRR resistance gene family in *Solanum pimpinellifolium* and *Arabidopsis thaliana*, *Evol. Bioinform. Online*, **16**, 1176934320911055.
23. Hou, L., Xie, J., Wu, Y., et al. 2021, Identification of 11 candidate structured noncoding RNA motifs in humans by comparative genomics, *BMC Genomics*, **22**, 1–14.
24. Taboada-Castro, H., Castro-Mondragón, J.A., Aguilar-Vera, A., Hernández-Álvarez, A.J., van Helden, J. and Encarnación-Guevara, S.

2020, RhizoBindingSites, a database of DNA-binding motifs in nitrogen-fixing bacteria inferred using a footprint discovery approach, *Front. Microbiol.*, **11**, 567471.
25. Blanchette, M. 2007, Computation and analysis of genomic multi-sequence alignments, *Annu. Rev. Genomics Hum. Genet.*, **8**, 193–213.
26. Rombauts, S., Florquin, K., Lescot, M., Marchal, K., Rouzé, P. and Van de Peer, Y. 2003, Computational approaches to identify promoters and cis-regulatory elements in plant genomes, *Plant Physiol.*, **132**, 1162–76.
27. Aerts, S. 2012, Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets, *Curr. Top. Dev. Biol.*, **98**, 121–45.
28. Fickett, J.W. and Wasserman, W.W. 2000, Discovery and modeling of transcriptional regulatory regions, *Curr. Opin. Biotechnol.*, **11**, 19–24.
29. De Witte, D., Van de Velde, J., Decap, D., et al. 2015, BLSSpeller: exhaustive comparative discovery of conserved cis-regulatory elements, *Bioinformatics*, **31**, 3758–66.
30. Carmack, C.S., McCue, L.A., Newberg, L.A. and Lawrence, C.E. 2007, PhyloScan: identification of transcription factor binding sites using cross-species evidence, *Algorithms Mol. Biol.*, **2**, 1–17.
31. Van Bel, M., Diels, T., Vancaester, E., et al. 2018, PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics, *Nucleic Acids Res.*, **46**, D1190–6.
32. Chen, N. 2004, Using Repeat Masker to identify repetitive elements in genomic sequences, *Curr. Protoc. Bioinformatics*, **5**, 4.10.11–14.10.14.
33. Stark, A., Lin, M.F., Kheradpour, P., et al.; Berkeley Drosophila Genome Project. 2007, Discovery of functional elements in 12 Drosophila genomes using evolutionary signatures, *Nature*, **450**, 219–32.
34. Reineke, A.R., Bornberg-Bauer, E. and Gu, J. 2011, Evolutionary divergence and limits of conserved non-coding sequence detection in plant genomes, *Nucleic Acids Res.*, **39**, 6029–43.
35. O'Malley, R.C., Huang, S.-S.C., Song, L., et al. 2016, Cistrome and epicistrome features shape the regulatory DNA landscape, *Cell*, **165**, 1280–92.
36. Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L. and Noble, W.S. 2007, Quantifying similarity between motifs, *Genome Biol.*, **8**, R24–9.
37. Kremling, K.A., Chen, S.-Y., Su, M.-H., et al. 2018, Dysregulation of expression correlates with rare-allele burden and fitness loss in maize, *Nature*, **555**, 520–3.
38. Leinonen, R., Sugawara, H., and Shumway, M.; International Nucleotide Sequence Database Collaboration. 2011, The sequence read archive, *Nucleic Acids Res.*, **39**, D19–21.
39. Andrews, S. 2010, FastQC: a quality control tool for high throughput sequence data. https://www.bioinformatics.babraham.ac.uk/projects/fastqc/ (December 2021, date last accessed).
40. Bolger, A.M., Lohse, M. and Usadel, B. 2014, Trimmomatic: a flexible trimmer for Illumina sequence data, *Bioinformatics*, **30**, 2114–20.
41. Jiao, Y., Peluso, P., Shi, J., et al. 2017, Improved maize reference genome with single-molecule technologies, *Nature*, **546**, 524–7.
42. Dobin, A., Davis, C.A., Schlesinger, F., et al. 2013, STAR: ultrafast universal RNA-seq aligner, *Bioinformatics*, **29**, 15–21.
43. Liao, Y., Smyth, G.K. and Shi, W. 2014, featureCounts: an efficient general purpose program for assigning sequence reads to genomic features, *Bioinformatics*, **30**, 923–30.
44. Robinson, M.D., McCarthy, D.J. and Smyth, G.K. 2010, edgeR: a bioconductor package for differential expression analysis of digital gene expression data, *Bioinformatics*, **26**, 139–40.
45. Yanai, I., Benjamin, H., Shmoish, M., et al. 2005, Genome-wide midrange transcription profiles reveal expression level relationships in human tissue specification, *Bioinformatics*, **21**, 650–9.
46. Alexa, A. and Rahnenfuhrer, J. 2010, topGO: enrichment analysis for gene ontology, R Package Version, 2. https://bioconductor.org/packages/2.5/bioc/html/topGO.html (December 2021, date last accessed).
47. R Core Team. R: A language and environment for statistical computing. 2013. https://cran.r-project.org/ (December 2021, date last accessed).
48. Zhou, P., Li, Z., Magnusson, E., et al. 2020, Meta gene regulatory networks in maize highlight functionally relevant regulatory interactions, *Plant Cell*, **32**, 1377–96.
49. Obayashi, T., Hayashi, S., Saeki, M., Ohta, H. and Kinoshita, K. 2009, ATTED-II provides coexpressed gene networks for Arabidopsis, *Nucleic Acids Res.*, **37**, D987–91.
50. Katz, L. 1953, A new status index derived from sociometric analysis, *Psychometrika*, **18**, 39–43.
51. Bonchi, F., Esfandiar, P., Gleich, D.F., Greif, C. and Lakshmanan, L.V. 2012, Fast matrix computations for pairwise and columnwise commute times and Katz scores, *Internet Math.*, **8**, 73–112.
52. Bukowski, R., Guo, X., Lu, Y., et al. 2018, Construction of the third-generation *Zea mays* haplotype map, *Gigascience*, **7**, 1.
53. Lawrence, M., Huber, W., Pages, H., et al. 2013, Software for computing and annotating genomic ranges, *PLoS Comput. Biol.*, **9**, e1003118.
54. Wickham, H. 2011, ggplot2, *WIREs Comp. Stat.*, **3**, 180–5.
55. Berens, M.L., Berry, H.M., Mine, A., Argueso, C.T. and Tsuda, K. 2017, Evolution of hormone signaling networks in plant defense, *Annu. Rev. Phytopathol.*, **55**, 401–25.
56. Katsir, L., Chung, H.S., Koo, A.J. and Howe, G.A. 2008, Jasmonate signaling: a conserved mechanism of hormone sensing, *Curr. Opin. Plant Biol.*, **11**, 428–35.
57. Chater, C., Kamisugi, Y., Movahedi, M., et al. 2011, Regulatory mechanism controlling stomatal behavior conserved across 400 million years of land plant evolution, *Curr. Biol.*, **21**, 1025–9.
58. Kaufmann, K., Pajoro, A. and Angenent, G.C. 2010, Regulation of transcription in plants: mechanisms controlling developmental switches, *Nat. Rev. Genet.*, **11**, 830–42.
59. Vanneste, S., Coppens, F., Lee, E., et al. 2011, Developmental regulation of CYCA2s contributes to tissue-specific proliferation in Arabidopsis, *EMBO J.*, **30**, 3430–41.
60. Siefers, N., Dang, K.K., Kumimoto, R.W., Bynum, W.E. IV, Tayrose, G. and Holt, B.F. III 2009, Tissue-specific expression patterns of Arabidopsis NF-Y transcription factors suggest potential for extensive combinatorial complexity, *Plant Physiol.*, **149**, 625–41.
61. Quandt, K., Frech, K., Karas, H., Wingender, E. and Werner, T. 1995, Matlnd and Matlnspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data, *Nucl. Acids Res.*, **23**, 4878–84.
62. Worsley-Hunt, R., Bernard, V. and Wasserman, W.W. 2011, Identification of cis-regulatory sequence variations in individual genome sequences, *Genome Med.*, **3**, 65–14.
63. Ma, S., Shah, S., Bohnert, H.J., Snyder, M. and Dinesh-Kumar, S.P. 2013, Incorporating motif analysis into gene co-expression networks reveals novel modular expression pattern and new signaling pathways, *PLoS Genet.*, **9**, e1003840.
64. Tabach, Y., Brosh, R., Buganim, Y., et al. 2007, Wide-scale analysis of human functional transcription factor binding reveals a strong bias towards the transcription start site, *PLoS One*, **2**, e807.
65. Clark, R.M., Schweikert, G., Toomajian, C., et al. 2007, Common sequence polymorphisms shaping genetic diversity in *Arabidopsis thaliana*, *Science*, **317**, 338–42.
66. Tatarinova, T.V., Chekalin, E., Nikolsky, Y., et al. 2016, Nucleotide diversity analysis highlights functionally important genomic regions, *Sci. Rep.*, **6**, 1–12.
67. Procházková Schrumpfová, P., Vychodilová, I., Dvořáčková, M., et al. 2014, Telomere repeat binding proteins are functional components of Arabidopsis telomeres and interact with telomerase, *Plant J.*, **77**, 770–81.
68. Warren, A.S., Anandakrishnan, R. and Zhang, L. 2010, Functional bias in molecular evolution rate of *Arabidopsis thaliana*, *BMC Evol. Biol.*, **10**, 1–10.
69. Andolfatto, P., Wong, K.M. and Bachtrog, D. 2011, Effective population size and the efficacy of selection on the X chromosomes of two closely related Drosophila species, *Genome Biol Evol.*, **3**, 114–28.
70. Lee, C.-R. and Mitchell-Olds, T. 2012, Environmental adaptation contributes to gene polymorphism across the *Arabidopsis thaliana* genome, *Mol. Biol. Evol.*, **29**, 3721–8.

71. Zhou, J., Lemos, B., Dopman, E.B. and Hartl, D.L. 2011, Copy-number variation: the balance between gene dosage and expression in *Drosophila melanogaster*, *Genome Biol. Evol.*, **3**, 1014–24.
72. Shi, T., Rahmani, R.S., Gugger, P.F., et al. 2020, Distinct expression and methylation patterns for genes with different fates following a single whole-genome duplication in flowering plants, *Mol. Biol. Evol.*, **37**, 2394–413.
73. Schuster-Böckler, B., Conrad, D. and Bateman, A. 2010, Dosage sensitivity shapes the evolution of copy-number varied regions, *PLoS One*, **5**, e9474.
74. Morton, T., Petricka, J., Corcoran, D.L., et al. 2014, Paired-end analysis of transcription start sites in Arabidopsis reveals plant-specific promoter signatures, *Plant Cell.*, **26**, 2746–60.