Genome-wide identification of preferentially located motifs in gene-proximal regions provides new insights into *cis*-regulatory sequences in plants

Julien Rozière[1,2,3,✉], **Cécile Guichard**[1,2], **Véronique Brunaud**[1,2], **Marie-Laure Martin**[1,2,4,✉], **and Sylvie Coursol**[3,✉]

[1]Université Paris-Saclay, CNRS, INRAE, Université Evry, Institute of Plant Sciences Paris-Saclay (IPS2), 91190, Gif sur Yvette, France
[2]Université de Paris, Institute of Plant Sciences Paris-Saclay (IPS2), 91190, Gif sur Yvette, France
[3]Université Paris-Saclay, INRAE, AgroParisTech, Institut Jean-Pierre Bourgin (IJPB), 78000, Versailles, France
[4]Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA-Paris, 75005, Paris, France

The identification of *cis*-regulatory elements controlling gene expression is an arduous challenge that is being actively explored to discover the key genetic factors responsible for traits of agronomic interest. In this regard, *in silico* predictive methods have the advantage of being of high-resolution and independent of experimental conditions. Here, we expanded the preferentially located motif (PLM) detection method to comprehensively analyze the gene-proximal regions of *Arabidopsis thaliana* and *Zea mays*. We identified three groups of PLMs in each region for each species and highlighted conserved PLMs in both species, particularly in the 3'-proximal region. Moreover, we showed that PLMs occurred at both transcription factor and microRNA binding sites in each group with specific positional distributions for each region. Although the majority of PLMs were unassigned PLMs, it is noteworthy that some were supported by MNase-defined cistrome occupancy analysis. Enrichment analysis also revealed that unassigned PLMs provide functional predictions distinct from those inferred by the other PLMs. Using PLMs that occured at transcription factor binding sites, we inferred the regulatory network of a poorly characterized *Z. mays*-specific gene family, paving the way for further characterization of its individual members.

Gene expression | Proximal cis-regulatory elements | PLMdetect predictive method | Regulatory network | Plant

Correspondence: ***julien.roziere@inrae.fr***
Correspondence may also be addressed to Marie-Laure Martin (marie_laure.martin@agroparistech.fr) and Sylvie Coursol (sylvie.coursol@inrae.fr).

Introduction

Plants, as sessile organisms, need to adapt to local constraints such as bacteria, fungi and pest presence, as well as environmental changes. One of the fundamental drivers of their adaptation is the activation or repression of gene transcription (1–5). These processes are tuned by many regulatory factors, including transcription factors (TFs) that bind to specific DNA TF binding sites (TFBSs) (6–9).

Efforts to characterize TFBSs at the genome-wide scale rely mostly on the *in vivo* chromatin immunoprecipitation followed by sequencing (ChIP-seq) method (10) or the *in vitro* DNA affinity purification (DAP-seq) method (11). Nevertheless, these two TF-centered methods are limited by the resolution of the data (fragments of about 100 bp), the difficulty of analyzing many TFs with specific antibodies (ChIP-seq) or the lack of consideration of potential co-factors and the genomic context (DAP-seq). In the 2020 ReMap update (12), only 372 transcriptional regulators were thus collected out of the 1,533 encoded by *Arabidopsis thaliana* (13). More globally the dependence on experimental conditions does not allow the identification of a complete cistrome for each TF studied. Chromatin structure profiling methods, including assay for transposase-accessible chromatin with high-throughput sequencing (ATAC-seq) (14), DNase I hypersensitive sites sequencing (DNase-seq) (15) or more recently MNase-defined cistrome-occupancy analysis (MOA-seq) (16) also depend on experimental conditions and may present resolution limitations to identify TFBSs.

Alternative and complementary approaches to identify potential TFBSs include *in silico* predictive methods (17–19). These DNA-centered methods have several advantages: they are capable to rapidly analyze a large number of potential TFBSs, identify *de novo* potential *cis*-regulatory elements without *a priori*, produce the highest-resolution footprints of DNA binding sites, and last but not least, are not dependent on experimental conditions. In this regard, PLMdetect (Preferentially Located Motif detection) (20) currently detects known *cis*-regulatory elements that are over-represented at a specific location relative to the transcription start site (TSS) in *A. thaliana* and are therefore referred to as preferentially located motifs (PLMs) (21–24). However, to date, no large-scale functional PLMdetect-based studies have been performed on plant proximal regions, including the 3'-untranslated region that also contribute to transcriptional regulation (1, 2, 25).

Here, we expanded the PLMdetect method to comprehensively analyze *de novo* the 5'- and 3'-proximal regions of

genes from *A. thaliana* and *Zea mays*. We sought to determine how their differences in genome content and architecture would be reflected in the characteristics of their PLMs in both proximal regions, including their promoters and untranslated transcribed regions. We identified key features of PLMs in each gene-proximal region and characterized similarities and differences in PLMs between the two plant species. Using this knowledge, we highlighted known MOA-seq TF footprints and inferred a gene regulatory network whose topology allowed further characterization of an unannoted gene family.

MATERIALS AND METHODS

Genomic datasets. TAIR10 (29) and B73v4.39 (30) genomes and their annotations were considered to extract the 5'- and 3'-gene proximal sequences of *A. thaliana* and *Z. mays* genes, respectively.

Overview of the expanded PLMdetect method. *De novo* identification of PLMs in 5'- and 3'-proximal regions of genes from *A. thaliana* and *Z. mays* included the three steps described below (Figure 1A).

Extraction of the 5'-proximal region. Only genes with an annotated 5-'UTR region (in GFF3/GTF file) were kept. Indeed, the first base of a gene without an annotated 5'-UTR region corresponds to the start codon which would bias the counting step of the PLM detection. Genes on reverse strand were reverse-complemented to analyze all sequences in the same orientation. Extracted sequences corresponded to the intervals [-1000;+500] and [-1500;+500] bp relative to the TSS for *A. thaliana* and *Z. mays*, respectively. In total, 19,736 and 25,848 genes were analyzed for *A. thaliana* and *Z. mays*, respectively.

Extraction of the 3'-proximal region. Similarly to what has been done for the 5'-gene-proximal region sequences, annotation of the transcription termination site (TTS) was ensured by filtering genes without a 3'-UTR region annotated. Extracted sequences were [-500;+1000] and [-500;+1500] bp with respect to the TTS for *A. thaliana* and *Z. mays*, respectively. To standardize the PLM detection step, genes on forward strand were reverse-complemented. Taking in consideration only annotated 3'-UTR, 20,573 and 25,199 genes were processed for *A. thaliana* and *Z. mays*, respectively.

Creation of a de novo motif file. Every non-polymorphic DNA 4-mers to 8-mers was generated representing 87,296 motifs. Among these motifs, 256, 1,024, 4,096, 16,384, and 65,536 had a length of 4, 5, 6, 7 and 8 bp, respectively.

Detection of the PLMs. Detection of the PLMs was based on (20). Briefly, the gene-proximal sequence file and the motif file were given as inputs. To determine if a motif was a PLM, the method counted the number of the motif occurrences for each position to get a motif distribution. First, a linear regression was computed with a confidence interval of 99% on a learning region defined as the first (last) 500 bp for the 5' (3')-gene-proximal region. Second, the linear regression was extended to the studied region defined as the gene-proximal region excluding the learning region. If the distribution exceeded the confidence interval in this studied region, the motif was considered as a PLM. This motif was characterized by (i) its preferential position, defined as the position where the distribution exceeded the regression line and the maximum high was reached, (ii) a functional window, defined as the part of the studied region where the motif distribution exceeded the confidence interval, and (iii) a score defined as the difference between the high of the peak and the upper bound of the confidence interval at the preferential position.

TFBS and microRNA resources and assignment.

Resource databases. TFBSs (676 total) were extracted from JASPAR Plant 2020 (26) and ChIP-seq of 104 maize leaf TFs (27). MicroRNAs (miRNAs) were obtained from psRNATarget (28) and only those from *A. thaliana* (427 total) and *Z. mays* (321 total) were kept.

TFBS and miRNA assignment. We first assigned TFBSs to PLMs for both species in each region separately using the TOMTOM web tool (31). Euclidean distance was used as a comparison function with a q-value threshold at 0.05 and the complete scoring option deselected. PLMs were also compared to the top 1% of k-mers of the 104 maize TFs (27) by considering only exact matches. Because miRNAs regulate genes by sequence complementarity (32) and are molecules from 19 to 22 nt, we only considered our 8 bp PLMs for this comparison. If a PLM was exactly found in a miRNA, it was assigned as a miPLM.

Functional annotation. Functional annotation of genes from both species was based on MapMan X4 (33) and Gene Ontology (GO) from PLAZA 4.5 (34). Functional enrichment analysis of genes containing an identified PLM (Supplementary Datasets) was performed by comparing the relative occurrence of each term to its relative occurrence in a reference list for each region and species using a hypergeometric test with the R function phyper. These reference lists consisted of all genes considered for PLM detection in each gene-proximal region in both species as described in the *Extraction* sections. *P*-values were adjusted by the Benjamini-Hochberg (BH) procedure to control the False Discovery Rate (FDR). An enriched term had its adjusted *P*-value lower than 0.05.

Comparative analysis of PLMs and MOA-seq motifs. *Z. mays* 5'-PLMs were compared to published MOA-seq motifs (16) using TOMTOM with the same parameters as described for TFBSs assignment and according to the following two criteria: (1) both sequence types had to be identical (TOMTOM q-value < 0.05) and (2) the position of the MOA-seq motif had to be within the functional window of the PLM extended by 50 bases upstream and downstream.

Inference and topology analysis of the HOM04M002476 gene family regulatory network.

Identification of TF-genes regulatory modules. A matrix of dimension 37x64 was generated with TF families in rows and HOM04M002476 genes in columns. Links between TF families and target genes were established when tPLMs, and thus associated TFs, were identified for a given target gene. These links were indicated by ones in the matrix. A zero indicated the absence of a tPLM associated with the TF family in the target gene. Gene and TF modules were obtained using latent block model (LBM) with the R package blockmodels (35).

Enrichment analysis of TF modules. Functional enrichment of each TF module was performed by comparing the relative occurrence of each term to its relative occurrence in the list of genes encoding TFs in *A. thaliana* (2,208 genes) and *Z. mays* (2,164 genes) using a hypergeometric test with the R phyper function. *P*-values were adjusted by the BH procedure to control the FDR. An enriched term had its adjusted *P*-value lower than 0.05.

RESULTS

***De novo* PLM identification in gene-proximal regions of *A. thaliana* and *Z. mays*.** To analyze *de novo* the gene-proximal regions of *A. thaliana* and *Z. mays*, we used the PLMdetect method as summarized in Figure 1. Following extraction of 5'- and 3'-gene proximal region sequences, creation of a *de novo* motif file, and *de novo* PLM detection, we plotted the distribution of the PLM scores obtained for each targeted regions of both plant species (Supplementary Figure S1). Only PLMs with a score > 2 were investigated. This led to the identification of 6,998 and 9,768 PLMs with a score > 2 in the 5'-gene-proximal regions (referred to as 5'-PLMs) of *A. thaliana* and *Z. mays*, respectively (Figure 2A, Supplementary Table S1A and 1C). Regarding the 3'-gene-proximal regions, 7,447 and 6,639 PLMs with a score > 2 (referred to as 3'-PLMs) were detected in *A. thaliana* and *Z. mays*, respectively (Figure 2B, Supplementary Table S1B and 1D).

Based on sequence similarities, 1,182 motifs were common between the 5'- and the 3'-motifs in *A. thaliana*, while 952 motifs were common in *Z. mays*. We also showed that *A. thaliana* and *Z. mays* shared 1,539 5'-motifs and 1,889 3'-motifs. Taking the preferential position as an additional criterion revealed that *A. thaliana* and *Z. mays* shared 1,063 5'-PLMs and 1,677 3'-PLMs (Figure 2A and 2B). It is worth noting that 98% of these PLMs were located in the 200 bases surrounding the TSS or the TTS.

Inspection of the preferential position of the PLMs revealed three groups in each targeted region and species with similar distribution patterns (Figure 3). In the 5'-gene-proximal region, groups 1 and 2 were localized upstream of the TSS, while group 3 was localized on the TSS. In the 3'-gene-proximal region, groups 1 and 2 were localized upstream of the TTS, while group 3 was localized on the TTS. It is worth noting that 5'-PLMs in *Z. mays* were more spread out along the sequence than those of *A. thaliana*. Additionally, each group had specific nucleotide content. Group 1 was composed of A, T, C and G nucleotides in equal proportions in both species. In contrast, group 2 was composed predominantly of A/T (74% and 64% in *A. thaliana* and *Z. mays*, respectively) in agreement with previous observations reporting TATA and TATA-like boxes in this region (36). In the case of group 3, we found that the GC content of the 5'-PLMs differed between the two species (37 % of GC in *A. thaliana vs* 55% in *Z. mays*), in agreement with the report of GC-rich genes in monocot species (37, 38) and recent promoter comparisons using *A. thaliana* and the three cereal species brachypodium, wheat and barley (39).

In the case of the 3'-gene-proximal region, we also noticed that 3,877 and 2,130 3'-PLMs detected in the [-40;+10] bp interval relative to the TTS (which corresponds to the end of group 2 and the whole group 3) in *A. thaliana* and *Z. mays*, respectively, showed similarities to the *cis*-elements that guide the cleavage and polyadenylation molecular complex (CPMC) essential for mRNA biogenesis (40). Indeed, they were A/T-rich (over 78%) in both *A. thaliana* and *Z. mays* similarly to the far upstream element (FUE) and near upstream element (NUE). Furthermore, 3'-PLMs localized 10 bases upstream and downstream of the TTS were composed of sequences rich in T (42% in both species) $>$ A (38% in *A. thaliana* and 34% in *Z. mays*) $>$ C (11% in *A. thaliana* and 14% in *Z. mays*) $>$ G (9% in *A. thaliana* and 11% in *Z. mays*), in agreement with the known proportions of nucleotides in the cleavage element (CE) (40).

Positional distribution of TFBS PLMs in proximal regions. We used the JASPAR Plant 2020 database (26) and top 100 k-mers from ChIP-seq data of 104 maize TFs (27) to determine if TFBSs were present among the identified 5'- and 3'-PLMs (Figure 1B). We found that 13.5% of 5'-PLMs and 9.1% of 3'-PLMs of *A. thaliana* showed similarities to our TFBS resource and were therefore referred to as tPLMs (Figures 2C and 2D, Table 1, Supplementary Table S2A-2B). In *Z. mays*, 9.6% of 5'-PLMs and 8.8% of 3'-PLMs corresponded to tPLMs (Figures 2E and 2F, Table 1, Supplementary Table S2C-2D). It is worth noting that, in *A. thaliana*, 5'-tPLMs were more localized in group 1, whereas 3'-tPLMs were more localized in group 3 (Table 1 and Figure 3). In *Z. mays*, 5'-tPLMs were more localized upstream and downstream of the identified groups, while 3'-tPLMs followed the same behavior as in *A. thaliana* with greater localization in group 3. Altogether, these results revealed a strong localization of 5'-tPLMs in the interval between 200 and 50 bp upstream of the TSS. In contrast, 3'-tPLMs mainly localized in the TTS region in both species.

We next investigated how the different TF families were distributed in each proximal region. Among the 47 TF families listed in our reference, 39 and 40 were susceptible to bind to 5'-tPLMs in *A. thaliana* and *Z. mays*, respectively (Figure 4). Similar ranges were observed for the 3'-gene-proximal region with 35 and 37 TF families susceptible to bind to 3'-tPLMs in *A. thaliana* and *Z. mays*, respectively. We also observed that 35 and 37 TF families overlapped between both regions studied in *A. thaliana* and *Z. mays*, respectively. Interestingly, four TF families (SWIM-type zinc fingers, CAMTA, LHY and ARR) specifically targeted the 5'-tPLMs in *A. thaliana*, while three distinct TF families (HSF

factors, E2F, and BBR-BPC) specifically targeted 5'-tPLMs in *Z. mays*. Comparing *A. thaliana* and *Z. mays* revealed one (CAMTA) and two (HSF factors and E2F) TF families specifically associated with 5'-tPLMs in *A. thaliana* and *Z. mays*, respectively. For the 3'-gene-proximal region, we observed that the BBR-BPC TFs were specific to *A. thaliana*, while the SWIM-type zinc fingers, LHY and ARR TFs were specific to *Z. mays*.

All TF families were not similarly distributed in each gene-proximal region. Indeed, some TF families, such as the MYB TFs, had tPLMs in all three groups previously identified for each region and species studied (Figure 4). In contrast, some TF families were specific to one group. This is in particular the case of the Trihelix TFs which was only present in group 1 in the 5'-proximal region of *A. thaliana*. Other TF families, such as the G2-like TFs, were likely to target different number of PLM groups according to the region and species considered. Finally, it is noteworthy that some TF families, such as MYB and MYB-related TFs, seemed to share very similar distributions.

PLMs occur at miRNA binding sites in proximal regions. It is well known that miRNAs can target transcripts with sequence complementarity (32), thus inducing their degradation. Moreover, it was already described in *Brassica* that miRNA methylates the promoter region of *SP11* gene to silence it (41). Consequently, we used the plant small RNA target analysis server psRNATarget (28) to determine whether some of the detected PLMs might be involved in the binding of miRNAs (Figure 1B). We identified 545 and 315 5'-PLMs which can be targeted by miRNAs (referred to as 5'-miPLMs) in *A. thaliana* and *Z. mays*, respectively (Figures 2C and 2E, Supplementary Table S3A and 3C). We also identified 693 and 210 3'-miPLMs for *A. thaliana* and *Z. mays*, respectively (Figures 2D and 2F, Supplementary Table S3B and 3D), of which 95 and 28 were found to be similar to the 5'-miPLMs identified in each species, respectively. Furthermore, comparison of miPLMs from *A. thaliana* with those identified in *Z. mays* revealed 15 common motifs in the 5'-gene-proximal region and 15 additional distinct common motifs in the 3'-gene-proximal region.

We then examined how miPLMs were distributed in both gene-proximal regions. Surprisingly, more than half of the 5'-miPLMs of *A. thaliana* were located in groups 1 and 2 (Figure 3 and Table 2). In contrast, the 5'-miPLMs of *Z. mays* were overwhelmingly found outside the groups. Inspecting the 3'-miPLMs distribution, we found that, in *A. thaliana*, they were more localized in group 3 than in the other two groups. In contrast, in *Z. mays*, half of the 3'-miPLMs were found in group 1 (Figure 3).

Additionally, we showed that 53 5'-miPLMs and 37 3'-miPLMs corresponded to tPLMs in *A. thaliana* (Figure 2C and 2D, Supplementary Table S1A and 1B). These 90 tPLMs/miPLMs were targeted by 156 and 50 TFs belonging to 17 and 14 TF families in the 5'- and 3'-gene-proximal regions, respectively (Supplementary Table S3E and 3F). WRKY, Basic helix-loop-helix factors (bHLH), Basic leucine zipper factors (bZIP) represented the three major TF families identified in the 5'-gene-proximal region, while C2H2 zinc finger factors, Myb-related and HD-ZIP factors were the three major TF families identified in the 3'-gene-proximal region. Similar observations were made in *Z. mays* in which 19 5'-miPLMs and 15 3'-miPLMs corresponded to tPLMs (Figure 2E and 2F, Supplementary Table S1C and 1D). These 34 tPLMs/miPLMs were targeted by 79 and 97 TFs belonging to 13 and 15 TF families in the 5'- and 3'-gene-proximal regions, respectively (Supplementary Table S3G and 3H). bHLH, bZIP and BZR represented the three major TF families identified in the 5'-gene-proximal region, while bHLH, bZIP and TCP domain were the major TF families in the 3'-gene-proximal region.

Finally, we sought to determine which sequence of miPLMs was homologous to that of miRNAs, knowing that the latter are composed of different parts that do not all have the same role (42) (Figure 1B). We observed that 5'-miPLMs of *A. thaliana* showed more frequent homologies to the 5' seed region (positions 2-8) of the miRNAs, whereas those of *Z. mays* showed more frequent homologies to the compensatory 3' end of miRNAs (Figure 5). For 3'-miPLMs, homologies appeared to be more frequent in the center of the miRNA, surrounding the cleavage site in both *A. thaliana* and *Z. mays*. It is worth noting that in both species the 5'-miPLMs located upstream of the three detected groups exhibited higher homology frequencies in the 5' seed region and the 3' compensatory end, as shown by the bimodality of the density curves. On the other hand, the 5'-miPLMs located downstream of the three groups of PLMs had a higher frequency of homology in the 5' seed region and the cleavage site in *A. thaliana* and *Z. mays*, respectively. Furthermore, the 5'-miPLMs belonging to the three groups showed close homology distributions in *Z. mays*. In contrast, the 5'-miPLMs belonging to group 2 in *A. thaliana* differed from the other two groups in having a higher homology frequency in the 5' seed region, whereas those of groups 1 and 3 had a higher frequency in the bases surrounding the cleavage site. The 3'-miPLMs of *A. thaliana* also had similar homology distributions regardless of miPLM location. In contrast, in *Z. mays*, miPLMs belonging to none of the three groups (i.e., upstream or downstream) had more frequent homologies in the 5' seed region, whereas those belonging to any of the three groups had higher homology frequencies in the bases surrounding the cleavage site. Overall, analysis of sequence homology between miPLMs and miRNAs provided evidence that each gene-proximal region exhibited specific properties, suggesting that miPLMs have distinct functions depending on the region and species considered.

Unassigned PLMs provide functional predictions distinct from those inferred by tPLMs and miPLMs. Comparison of each *de novo* PLM with resources of TFBS and miRNA binding sites revealed that more than 79% of the identified PLMs were unassigned PLMs (referred to as uPLMs) in both plant species (Figure 1B, 2C-2F and Supplementary Table S1). Meanwhile, 11.5% (642/5,559) and 6.9% (589/8,528) of the 5'-uPLMs detected in *A. thaliana* and *Z. mays*, respectively, were similar to RNA polymerase II bind-

ing sites (26) (Supplementary Table S4).

In *A. thaliana*, more than one-third of the 5'-uPLMs were localized in group 1, while more than one-third of the 3'-uPLMs were localized in group 3 (Table 3). Nevertheless, it is noteworthy that the density of the 5'-uPLMs showed a peak in group 2 (Figure 3). In *Z. mays*, the 5'-uPLMs were preferentially localized upstream and downstream of the three groups detected, showing again a greater dispersion than that observed in *A. thaliana*. The 3'-uPLMs of *Z. mays* had a more balanced distribution among the different groups and downstream part than the 3'-uPLMs of *A. thaliana*. Therefore, no one group stood out more than another.

To go further, a systematic enrichment analysis was carried out (Figure 1C). In both species, the 5'- and 3'-uPLMs-containing gene sets constituted two highly differentiated populations in terms of their biological processes or MapMan categories relative to the other PLM classes (Figure 6, Supplementary Table S5 and S6). Furthermore, comparing 5'- and 3'-uPLMs-containing gene sets revealed specific terms associated with each of the two sets (Supplementary Figure S2 and Table S7). It is worth noting that "cellular response to ethylene stimulus" was one of the five most enriched terms in the 3'-uPLMs-gene set of both *A. thaliana* and *Z. mays*. Furthermore, terms related to "Cell wall organisation" were present in the five most enriched MapMan terms for the *A. thaliana*-uPLMs-gene set in both regions, although the protein classes identified were different. As expected, each species had also terms enriched specifically for each uPLMs-gene set (Supplementary Figure S2 and Table S7). For the 5'-uPLMs-gene set, these terms were mainly related to "cell killing" in *A. thaliana*, while they were associated with "transposition" and antibiotic metabolic/catabolic processes in *Z. mays* (Supplementary Table S7G-7H). For the 3'-uPLMs-gene set, specific enriched terms were once again related to "cell killing" in *A. thaliana*, while they were mainly related to "translation" in *Z. mays*.

Recently, a MOA-seq strategy on developping maize ears led to the identification of 215 small ($<$ 30 bp) TF footprints collectively distributed over 100,000 non-overlapping binding sites accross the genome (16). Given the relatively small size of these footprints and their remarkable clustering within the 100 bp proximal to the promoters, we examined them in terms of sequence and position (Figure 1D). We found that 85 of these 215 TF footprints significantly matched with 203 of our motifs. Considering the position of these motifs (plus or minus 50 bases upstream and downstream of the corresponding PLM functional window), 30% of them covered 94 PLMs (Supplementary Table S8), including 24 tPLMs, 14 miPLMs and 60 uPLMs. Interestingly, 24 GO terms were enriched for all types of PLM-containing gene sets, including "transposition" which appeared to be the most enriched GO term among those specific to the 5'-uPLMs containing gene sets of *Z. mays* compared to those of *A. thaliana* (Supplementary Table S9). Together, the analysis of GO/MapMan terms and MOA-seq TF footprints demonstrate that uPLMs represent novel proximal *cis*-regulatory players.

Regulatory network inference and sub-families identification for a poorly characterized, *Z. mays*-specific gene family. Taking into account the contribution of PLMs, we thought to use them to infer gene regulatory networks and go further and deeper into the characterization of some gene families. We focused on a poorly characterized, *Z. mays*-specific gene family (referred to as HOM04M002476 by PLAZA) defined only by the GO term "transposition" (Supplementary Table S7G). This gene family consists of 65 genes, 64 of which were considered in the detection of 5'-PLMs (Supplementary Table S10A). Using the 5'-tPLMs detected for all these 64 genes, we investigated the TF-target gene relationships (Figure 1E). A total of 545 tPLMs were associated with 416 TFs belonging to 37 distinct TF families. Among them, AP2/ERF domain, Myb-related and WRKY were the three most abundant TF families (Supplementary Table S10).

Clustering based on LBM revealed three modules of target genes referred to as G1, G2 and G3 with 7, 3 and 54 members, respectively. It also revealed three modules of TFs referred to as TF1, TF2 and TF3 with 6, 380 and 30 TFs belonging to 2, 29 and 6 TF families, respectively (Figure 7A, Supplementary Table S10). Moreover, genes belonging to module G1 were regulated by TFs from modules TF1 (2/2 families), TF2 (16/29 families) and TF3 (1/6 family). It is worth noting that the SWIM-type zinc finger TF family of module TF1 was specific to genes from G1 module. Genes belonging to module G2 were also regulated by TFs belonging to all three modules, including the BZR TF family of module TF1, all TF families of module TF2, and the HSF factors and G2-like TF families of module TF3 (Figure 7A, Supplementary Table S10). In contrast, genes belonging to module G3 were only regulated by TFs from modules TF2 and TF3. Furthermore, the CXC, CPP and SBP-type zinc finger TF families of module TF3 covered specifically genes from G3 module.

To elucidate the potential involvement of this gene regulatory network in biological processes, GO and MapMan enrichment analyses of the TF1, TF2 and TF3 modules were performed. As expected, terms enriched in a common way in all three modules were related to the regulation of transcription (Supplementary Table S11). Furthermore, in module TF1, the most specifically enriched terms were related to brassinosteroid responses (Figure 7B, Supplementary Table S11A-11B). In module TF2, except for terms related to transcriptional regulation, the most enriched terms was related to “response to stimulus” (Figure 7B, Supplementary Table S11C-11D). Finally, in module TF3 the most enriched terms were mainly related to methylation, response to nutrient levels and primary root development (Figure 7B, Supplementary Table S11E-11F).

DISCUSSION

The identification of *cis*-regulatory motifs controlling gene expression is a real challenge that is being actively explored to discover the key genetic factors responsible for traits of agronomic interest. The genome-wide and *de novo* PLMde-

tect approach reported here allowed us to provide functional insights into the regulatory role of motifs overrepresented at a specific location to the TSS and the TTS in the 5'- and 3'-proximal regions, respectively, of genes from *A. thaliana* and *Z. mays*.

The preferential position of PLMs is a key feature for a better understanding of the organization of the 5'-gene proximal region. By taking into account the preferential position of PLMs, the three-group structure revealed in the two species with distantly related genomes echoes and enriches established knowledge of the 5'-gene-proximal region. Thus, the densities of 5'-uPLMs in both species were higher in the core promoter region corresponding to group 2 and known to be the locus of many regulatory events (21, 43–45). This result suggests that many 5'-uPLMs might represent potential core promoter players in both species (Figure 8).

Omitting potential annotation errors, it is also worth noting that the localization of these motifs including the core promoter was less constrained in *Z. mays* compared to *A. thaliana*, as recently reported for TATA-boxes at varying distances from TSS in *Z. mays* (46). Similarly, we observed that the dispersion of tPLMs in group 1 remained more important in *Z. mays* than in *A. thaliana*, indicating that putative TFBSs have also a less constrained preferential localization in *Z. mays* than in *A. thaliana*. Together, these data suggest that the 5'-proximal genomic context may be less constrained in *Z. mays* than in *A. thaliana*. This might be related to the richness of the *Z. mays* genome in transposable elements compared to that of *A. thaliana* (47).

To date, relatively few studies have examined the distribution of TFBSs content in promoter regions in plants (48–50). According to studies in *A. thaliana* and *P. persica*, the majority of TFBSs localize in a region approximately 50 to 200 bp upstream of the TSS, which is similar to our group 1 (48, 49). Similarly, we observed that 5'-tPLMs were more localized in group 1 than in the other two groups in *A. thaliana*. Furthermore, we showed that the distribution of the associated TF families was similar to that described in *A. thaliana* and *P. persica* (48, 49). In contrast, in *Z. mays*, the 5'-tPLMs were more localized upstream and downstream the three identified groups. Surprisingly, our *Z. mays* TF family distribution showed no overlap with that reported in wheat (50). Thus, it would be interesting to validate these predicted PLMs by functional assay to establish what proportion of them are actually bound by TFs.

The size selected to search for PLMs (maximum 8 bp) was not optimal for detecting miRNA binding since miRNAs regulate gene expression by sequence complementarity on about twenty bases (not tested for computational constraints). Nevertheless, we were able to show that 5'-miPLMs had a maximum density downstream of the TSS in both species (Figure 8), which is consistent with the main mode of action of miRNAs and supports our approach and findings.

Conserved PLMs between species highlight the importance of the 3'-gene proximal region in genome structure. Our finding about the existence of conserved PLMs beween *A. thaliana* and *Z. mays* suggests that the closer we get to the genes, the more the context, including *cis*-regulatory elements (here given by tPLMs), are conserved between species. It is noteworthy that this context seems to be more conserved in the 3'-gene proximal region than in the 5'-gene proximal region, emphasizing the importance of the 3'-gene-proximal region in the genomic structure.

The 3'-proximal region of genes plays a key role in their expression, although it remains poorly studied in plants (25, 40, 51). It is noteworthy that the density maxima observed for tPLMs and uPLMs in both species were reached in the *cis*-elements that guide the CPMC and overlapped with groups 2 and 3. Therefore, it is quite possible that the 3'-PLMs detected in these portions of DNA sequence constitute a catalog of NUE, FUE and CE (Figure 8). In support of this hypothesis, we observed that the AATAAA motif (and its complementary reverse), which is the key site involved in polyadenylation and is extremely conserved in mammals and somewhat less in plants, was located between 10 and 20 bases upstream of the TTS. Furthermore, the nucleotide percentages of 3'-PLMs detected in these regulatory regions are consistent with known proportions of nucleotides in these *cis*-elements. Together, these data support the idea that 3'-PLMs may constitute an accurate catalog of CPMC-guiding *cis*-elements.

In this respect, the presence of 3'-tPLMs in this catalog (around 11%) and more generally in the whole region, opens new mechanistic perspectives on the role of TFs in the 3'-gene proximal region. First, these TFs may act as activators or repressors of the transcriptional machinery (Figure 8). Second, by binding to tPLMs located in the FUE/NUE and CE regions, these TFs could impact pre-mRNAs length and thus mRNA stability by influencing the choice of an alternative polyadenylation site at the end of transcripts (25, 40, 51). Finally, concerning the 3'-miPLMs, it is interesting to note that their density differs in the two species. In *A. thaliana*, the density peak is located in the tens of bases preceding the TTS (group 3). This distribution is reminiscent of the regulatory mechanism identified for the AT3G17185 gene for which a miR390 binding site is located between two alternative polyadenylation sites (52). Therefore, the strong presence of miPLMs at the UTR end suggests that this type of regulation may be more recurrent than previously described. In contrast, in *Z. mays*, the maximum density is found more than 300 bases upstream of the TTS (group 1). This result suggests that there is less regulation of miRNAs between two alternative polyadenylation sites and that complementation binding within the transcript remains the preferred mechanism in *Z. mays*.

The multiple mechanisms underlying the presence of uPLMs. Our finding on the existence of conserved PLMs between *A. thaliana* and *Z. mays* also revealed that the proportion of conserved uPLMs (up to 25 %) was of the same order of magnitude as that of conserved tPLMs (up to 30 %). Therefore, uPLMs may constitute a context that needs to be conserved. In this regard, it is worth noting that the uPLM genomic significance is already supported by experi-

mental MOA-seq data that validate some of the uPLMs identified. This raises the question of what mechanisms underlie the presence of uPLMs.

First, as mentionned earlier, some uPLMs may be players in the core promoter or polyadenylation process (Figure 8). Second, some uPLMs may be non-annotated binding sites. Indeed, we showed that the 104 maize TF ChIP-seq data (27) contributed 16% and 33% more tPLMs for *A. thaliana* and *Z. mays*, respectively, compared to the JASPAR Plant 2020 database that was updated prior to the release of these ChIP-seq data. Similar to the post-transcriptional regulation by miRNAs, RNA-binding proteins (53, 54) are major players that can potentially bind PLMs at the transcriptional level. Consequently, there is no doubt that future resources will supplement the assignment of uPLMs.

Finally, uPLMs may be motifs that are not directly involved in the binding of TFs but that play a crucial role in the correct binding of these regulators to neighboring TFBSs (Figure 8) (55–57). This concept of "flanking sequence context" appears extremely relevant because of the nature of PLMs, which are constrained motifs at a distance from genes. This idea also raises many questions about the existence and role of tPLMs/uPLMs associations, with tPLMs bound by TFs and uPLMs serving as essential context sequences for the formation of the DNA-TF complex. Additional analyses and integration with other *in vivo* information will be key to advance functional tests needed to ascertain all the hypotheses generated from knowledge of these PLMs.

PLMs are a resource for functional annotation of genomes. The messages carried by the PLMs-containing gene sets were extremely diverse. For example, comparing the 5'-uPLMs-containing gene sets to the 3'-uPLMs-containing gene sets in *A. thaliana* revealed specific enrichments for proteins involved in cell wall organization but belonging to two different classes. Thus, we showed that part of the genes encoding alpha-expansin are characterized by 5'-uPLMs signals localized after the TSS, while part of the genes encoding acyl omega-hydroxylases are characterized by 3'-uPLMs signals localized in group 1 and 2.

Functional enrichments also allowed the characterization of group 2 5'-uPLM signals in some genes involved in reactive oxygen metabolic processes. These signals were composed of more than 72% of uPLMs similar to RNA polymerase II binding sites, in agreement with previous observations showing a link between TATA-box and stress response, including the response to H_2O_2 (21, 46, 58).

Additionally, specific enrichments of the 3'-uPLMs-containing gene sets in group 1 compared to other groups revealed an involvement of these genes in ethylene response in both species. This means that part of the genes involved in the ethylene response are characterized by uPLMs signals in the -450 to -200 bases relative to the TTS in *A. thaliana* and *Z. mays*. These signals are further supported by the fact that the uPLMs' sequences remain very conserved (CGTCG and its reverse-complementary CGACG for *A. thaliana*; ACGCCCAC / GGGCGTCC and its reverse-complementary GGACGCCC for *Z. mays*) between the two species.

Together, this knowledge offers a promising avenue to explore the function of orphan genes in both species (59, 60). To do so, we devised a PLMs-based approach to functionally annotate a gene family. This enabled us to show that the HOM04M002476 family can be characterized by three structural gene groups potentially involved in several biological processes, including lipid response, stimulus response, brassinosteroid signaling pathway, response to nutrient levels and plant development. Therefore, this computational approach paves the way for pinpointing the function of this gene family in these distinct processes in future follow-up studies.

CONCLUSION

In conclusion, this genome wide and *de novo* study revealed several classes of PLMs conserved in the 5'- and 3'-gene-proximal regions of *A. thaliana* and *Z. mays*, including that of uPLMs that represent novel proximal *cis*-regulatory players. Further experiments are needed to determine the mechanisms underlying the presence of uPLMs as well as the relative importance of tPLMs and uPLMs as *cis*-regulatory elements controlling gene expression. Meanwhile, our results have broader implications for future characterization of orphan genes in plants.

ACKNOWLEDGEMENTS

We thank all members of the 'Genomic Networks' (GNet) and 'Biomass Quality and Interactions with Drought' (QUALIBIOSEC) teams past and present. We also thank Hank W. Bass for sharing some works prior to publication and Hervé Vaucheret for very helpful discussions.

FUNDING

This work was supported by the Plant2Pro® Carnot Institute in the frame of the PLMViewer program. Plant2Pro® is supported by ANR (agreement #19 CARN 0024 01). The IPS2 and IJPB laboratories benefit from the support of Saclay Plant Sciences-SPS (ANR-17-EUR-0007). J.R. is supported by a PhD fellowship of the Doctoral School 'Structure and Dynamics of Living Systems' from the University Paris-Saclay.

BIBLIOGRAPHY

1. Azodi,C.B., Lloyd,J.P. and Shiu,S.-H. (2020) The cis-regulatory codes of response to combined heat and drought stress in Arabidopsis thaliana. *NAR Genomics and Bioinformatics*, **2**.
2. Zhou,P., Enders,T.A., Myers,Z.A., Magnusson,E., Crisp,P.A., Noshay,J., Gomez-Cano,F., Liang,Z., Grotewold,E., Greenham,K., *et al.* (2021) Prediction of conserved and variable heat and cold stress response in maize using cis-regulatory information. *Plant Cell*, 10.1093/plcell/koab267.
3. Alonge,M., Wang,X., Benoit,M., Soyk,S., Pereira,L., Zhang,L., Suresh,H., Ramakrishnan,S., Maumus,F., Ciren,D., *et al.* (2020) Major Impacts of Widespread Structural Variation on Gene Expression and Crop Improvement in Tomato. *Cell*, **182**, 145-161.e23.
4. Waters,A.J., Makarevitch,I., Noshay,J., Burghardt,L.T., Hirsch,C.N., Hirsch,C.D. and Springer,N.M. (2017) Natural variation for gene expression responses to abiotic stress in maize. *The Plant Journal*, **89**, 706–717.
5. Liu,S., Li,C., Wang,H., Wang,S., Yang,S., Liu,X., Yan,J., Li,B., Beatty,M., Zastrow-Hayes,G., *et al.* (2020) Mapping regulatory variants controlling gene expression in drought response and tolerance in maize. *Genome Biology*, **21**, 163.
6. Li,X., Zhu,C., Yeh,C.-T., Wu,W., Takacs,E.M., Petsch,K.A., Tian,F., Bai,G., Buckler,E.S., Muehlbauer,G.J., *et al.* (2012) Genic and nongenic contributions to natural variation of quantitative traits in maize. *Genome Res.*, **22**, 2436–2444.

7. 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 Genetics*, **10**, e1004845.
8. Fagny,M., Kuijjer,M.L., Stam,M., Joets,J., Turc,O., Rozière,J., Pateyron,S., Venon,A. and Vitte,C. (2021) Identification of Key Tissue-Specific, Biological Processes by Integrating Enhancer Information in Maize Gene Regulatory Networks. *Front. Genet.*, **11**, 606285.
9. Zemlyanskaya,E.V., Dolgikh,V.A., Levitsky,V.G. and Mironova,V. (2021) Transcriptional regulation in plants: Using omics data to crack the cis-regulatory code. *Curr Opin Plant Biol*, **63**, 102058.
10. Robertson,G., Hirst,M., Bainbridge,M., Bilenky,M., Zhao,Y., Zeng,T., Euskirchen,G., Bernier,B., Varhol,R., Delaney,A., *et al.* (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods*, **4**, 651–657.
11. O'Malley,R.C., Huang,S.C., Song,L., Lewsey,M.G., Bartlett,A., Nery,J.R., Galli,M., Gallavotti,A. and Ecker,J.R. (2016) Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. *Cell*, **165**, 1280–1292.
12. Chèneby,J., Ménétrier,Z., Mestdagh,M., Rosnet,T., Douida,A., Rhalloussi,W., Bergon,A., Lopez,F. and Ballester,B. (2020) ReMap 2020: a database of regulatory regions from an integrative analysis of Human and Arabidopsis DNA-binding sequencing experiments. *Nucleic Acids Res.*, **48**, D180–D188.
13. Riechmann,J.L., Heard,J., Martin,G., Reuber,L., Jiang,C., Keddie,J., Adam,L., Pineda,O., Ratcliffe,O.J., Samaha,R.R., *et al.* (2000) Arabidopsis transcription factors: genome-wide comparative analysis among eukaryotes. *Science*, **290**, 2105–2110.
14. Buenrostro,J.D., Giresi,P.G., Zaba,L.C., Chang,H.Y. and Greenleaf,W.J. (2013) Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods*, **10**, 1213–1218.
15. Song,L. and Crawford,G.E. (2010) DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb Protoc*, 2010, pdb.prot5384.
16. Savadel,S.D., Hartwig,T., Turpin,Z.M., Vera,D.L., Lung,P.-Y., Sui,X., Blank,M., Frommer,W.B., Dennis,J.H., Zhang,J., *et al.* (2021) The native cistrome and sequence motif families of the maize ear. *PLoS Genet*, **17**, e1009689.
17. Nguyen,N.T.T., Contreras-Moreira,B., Castro-Mondragon,J.A., Santana-Garcia,W., Ossio,R., Robles-Espinoza,C.D., Bahin,M., Collombet,S., Vincens,P., Thieffry,D., *et al.* (2018) RSAT 2018: regulatory sequence analysis tools 20th anniversary. *Nucleic Acids Res.*, **46**, W209–W214.
18. Bailey,T.L. and Elkan,C. (1995) Unsupervised learning of multiple motifs in biopolymers using expectation maximization. *Mach Learn*, **21**, 51–80.
19. Saad,C., Noé,L., Richard,H., Leclerc,J., Buisine,M.-P., Touzet,H. and Figeac,M. (2018) DiNAMO: highly sensitive DNA motif discovery in high-throughput sequencing data. *BMC Bioinformatics*, **19**.
20. Bernard,V., Lecharny,A. and Brunaud,V. (2010) Improved detection of motifs with preferential location in promoters. *Genome*, **53**, 739–752.
21. Bernard,V., Brunaud,V. and Lecharny,A. (2010) TC-motifs at the TATA-box expected position in plant genes: a novel class of motifs involved in the transcription regulation. *BMC Genomics*, **11**, 166.
22. Frei dit Frey,N., Garcia,A.V., Bigeard,J., Zaag,R., Bueso,E., Garmier,M., Pateyron,S., de Tauzia-Moreau,M.-L., Brunaud,V., Balzergue,S., *et al.* (2014) Functional analysis of Arabidopsis immune-related MAPKs uncovers a role for MPK3 as negative regulator of inducible defences. *Genome Biol*, **15**, R87.
23. Martínez,F., Arif,A., Nebauer,S.G., Bueso,E., Ali,R., Montesinos,C., Brunaud,V., Muñoz-Bertomeu,J. and Serrano,R. (2015) A fungal transcription factor gene is expressed in plants from its own promoter and improves drought tolerance. *Planta*, **242**, 39–52.
24. Bueso,E., Muñoz-Bertomeu,J., Campos,F., Brunaud,V., Martínez,L., Sayas,E., Ballester,P., Yenush,L. and Serrano,R. (2014) ARABIDOPSIS THALIANA HOMEOBOX25 Uncovers a Role for Gibberellins in Seed Longevity1[C][W]. *Plant Physiol*, **164**, 999–1010.
25. Mayr,C. (2019) What Are 3' UTRs Doing? *Cold Spring Harb Perspect Biol*, **11**, a034728.
26. Fornes,O., Castro-Mondragon,J.A., Khan,A., van der Lee,R., Zhang,X., Richmond,P.A., Modi,B.P., Correard,S., Gheorghe,M., Baranašić,D., *et al.* (2019) JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 10.1093/nar/gkz1001.
27. Tu,X., Mejía-Guerra,M.K., Franco,J.A.V., Tzeng,D., Chu,P.-Y., Shen,W., Wei,Y., Dai,X., Li,P., Buckler,E.S., *et al.* (2020) Reconstructing the maize leaf regulatory network using ChIP-seq data of 104 transcription factors. *Nat Commun*, **11**, 1–13.
28. Dai,X., Zhuang,Z. and Zhao,P.X. (2018) psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res.*, **46**, W49–W54.
29. Lamesch,P., Berardini,T.Z., Li,D., Swarbreck,D., Wilks,C., Sasidharan,R., Muller,R., Dreher,K., Alexander,D.L., Garcia-Hernandez,M., *et al.* (2012) The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res.*, **40**, D1202–D1210.
30. Jiao,Y., Peluso,P., Shi,J., Liang,T., Stitzer,M.C., Wang,B., Campbell,M.S., Stein,J.C., Wei,X., Chin,C.-S., *et al.* (2017) Improved maize reference genome with single-molecule technologies. *Nature*, **546**, 524–527.
31. Gupta,S., Stamatoyannopoulos,J.A., Bailey,T.L. and Noble,W.S. (2007) Quantifying similarity between motifs. *Genome Biol*, **8**, R24.
32. Bartel,D.P. (2009) MicroRNA Target Recognition and Regulatory Functions. *Cell*, **136**, 215–233.
33. Thimm,O., Bläsing,O., Gibon,Y., Nagel,A., Meyer,S., Krüger,P., Selbig,J., Müller,L.A., Rhee,S.Y. and Stitt,M. (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. *Plant J*, **37**, 914–939.
34. Van Bel,M., Diels,T., Vancaester,E., Kreft,L., Botzki,A., Van de Peer,Y., Coppens,F. and Vandepoele,K. (2018) PLAZA 4.0: an integrative resource for functional, evolutionary and comparative plant genomics. *Nucleic Acids Res.*, **46**, D1190–D1196.
35. Leger,J.-B. (2016) Blockmodels: A R-package for estimating in Latent Block Model and Stochastic Block Model, with various probability functions, with or without covariates. arXiv:1602.07587 [stat].
36. Joshi,C.P. (1987) An inspection of the domain between putative TATA box and translation start site in 79 plant genes. *Nucleic Acids Res.*, **15**, 6643–6653.
37. Clément,Y., Fustier,M.-A., Nabholz,B. and Glémin,S. (2014) The bimodal distribution of genic GC content is ancestral to monocot species. *Genome Biol Evol*, **7**, 336–348.
38. Sundararajan,A., Dukowic-Schulze,S., Kwicklis,M., Engstrom,K., Garcia,N., Oviedo,O.J., Ramaraj,T., Gonzales,M.D., He,Y., Wang,M., *et al.* (2016) Gene Evolutionary Trajectories and GC Patterns Driven by Recombination in Zea mays. *Front Plant Sci*, **7**, 1433.
39. Peng,F.Y., Hu,Z. and Yang,R.-C. (2016) Bioinformatic prediction of transcription factor binding sites at promoter regions of genes for photoperiod and vernalization responses in model and temperate cereal plants. *BMC Genomics*, **17**, 573.
40. Bernardes,W.S. and Menossi,M. (2020) Plant 3' Regulatory Regions From mRNA-Encoding Genes and Their Uses to Modulate Expression. *Frontiers in Plant Science*, **11**, 1252.
41. Tarutani,Y., Shiba,H., Iwano,M., Kakizaki,T., Suzuki,G., Watanabe,M., Isogai,A. and Takayama,S. (2010) Trans-acting small RNA determines dominance relationships in Brassica self-incompatibility. *Nature*, **466**, 983–986.
42. Ossowski,S., Schwab,R. and Weigel,D. (2008) Gene silencing in plants using artificial microRNAs and other small RNAs: Engineering small RNA-mediated gene silencing. *The Plant Journal*, **53**, 674–690.
43. Grosschedl,R. and Birnstiel,M.L. (1980) Identification of regulatory sequences in the prelude sequences of an H2A histone gene by the study of specific deletion mutants in vivo. *PNAS*, **77**, 1432–1436.
44. Molina,C. and Grotewold,E. (2005) Genome wide analysis of Arabidopsis core promoters. *BMC Genomics*, **6**, 25.
45. Yamamoto,Y.Y., Ichida,H., Abe,T., Suzuki,Y., Sugano,S. and Obokata,J. (2007) Differentiation of core promoter architecture between plants and mammals revealed by LDSS analysis. *Nucleic Acids Res.*, **35**, 6219–6226.
46. Jores,T., Tonnies,J., Wrightsman,T., Buckler,E.S., Cuperus,J.T., Fields,S. and Queitsch,C. (2021) Synthetic promoter designs enabled by a comprehensive analysis of plant core promoters. *Nat. Plants*, **7**, 842–855.
47. Stitzer,M.C., Anderson,S.N., Springer,N.M. and Ross-Ibarra,J. (2021) The genomic ecosystem of transposable elements in maize. *PLOS Genetics*, **17**, e1009768.
48. Yu,C.-P., Lin,J.-J. and Li,W.-H. (2016) Positional distribution of transcription factor binding sites in Arabidopsis thaliana. *Sci Rep*, **6**, 25164.
49. Ksouri,N., Castro-Mondragón,J.A., Montardit-Tardà,F., Helden,J. van, Contreras-Moreira,B. and Gogorcena,Y. (2020) Motif analysis in co-expression networks reveals regulatory elements in plants: The peach as a model. *bioRxiv*
50. Chen,Z.-Y., Guo,X.-J., Chen,Z.-X., Chen,W.-Y. and Wang,J.-R. (2017) Identification and positional distribution analysis of transcription factor binding sites for genes from the wheat fl-cDNA sequences. *Bioscience, Biotechnology, and Biochemistry*, **81**, 1125–1135.
51. Srivastava,A.K., Lu,Y., Zinta,G., Lang,Z. and Zhu,J.-K. (2018) UTR dependent control of gene expression in plants. *Trends Plant Sci*, **23**, 248–259.
52. Hunt,A. (2012) RNA Regulatory Elements and Polyadenylation in Plants. *Frontiers in Plant Science*, **2**, 109.
53. Lee,K. and Kang,H. (2016) Emerging Roles of RNA-Binding Proteins in Plant Growth, Development, and Stress Responses. *Mol Cells*, **39**, 179–185.
54. Cho,H., Cho,H.S. and Hwang,I. (2019) Emerging roles of RNA-binding proteins in plant development. *Curr Opin Plant Biol*, **51**, 51–57.
55. Crocker,J., Abe,N., Rinaldi,L., McGregor,A.P., Frankel,N., Wang,S., Alsawadi,A., Valenti,P., Plaza,S., Payre,F., *et al.* (2015) Low affinity binding site clusters confer hox specificity and regulatory robustness. *Cell*, **160**, 191–203.
56. Stampfel,G., Kazmar,T., Frank,O., Wienerroither,S., Reiter,F. and Stark,A. (2015) Transcriptional regulators form diverse groups with context-dependent regulatory functions. *Nature*, **528**, 147–151.
57. Stringham,J.L., Brown,A.S., Drewell,R.A. and Dresch,J.M. (2013) Flanking sequence context-dependent transcription factor binding in early Drosophila development. *BMC Bioinformatics*, **14**, 298.
58. Miller,G., Shulaev,V. and Mittler,R. (2008) Reactive oxygen signaling and abiotic stress. *Physiol Plant*, **133**, 481–489.
59. Zaag,R., Tamby,J.P., Guichard,C., Tariq,Z., Rigaill,G., Delannoy,E., Renou,J.-P., Balzergue,S., Mary-Huard,T., Aubourg,S., *et al.* (2015) GEM2Net: from gene expression modeling to -omics networks, a new CATdb module to investigate Arabidopsis thaliana genes involved in stress response. *Nucleic Acids Res.*, **43**, D1010–D1017.
60. Portwood,J.L.,II, Woodhouse,M.R., Cannon,E.K., Gardiner,J.M., Harper,L.C., Schaeffer,M.L., Walsh,J.R., Sen,T.Z., Cho,K.T., Schott,D.A., *et al.* (2019) MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Res.*, **47**, D1146–D1154.

Table 1. Number of tPLMs detected in the different groups. Us: region located upstream the three groups; 1, 2 and 3: group 1, 2 and 3; Ds: region located downstream the three groups; NA: region located between the groups when they are not juxtaposed.

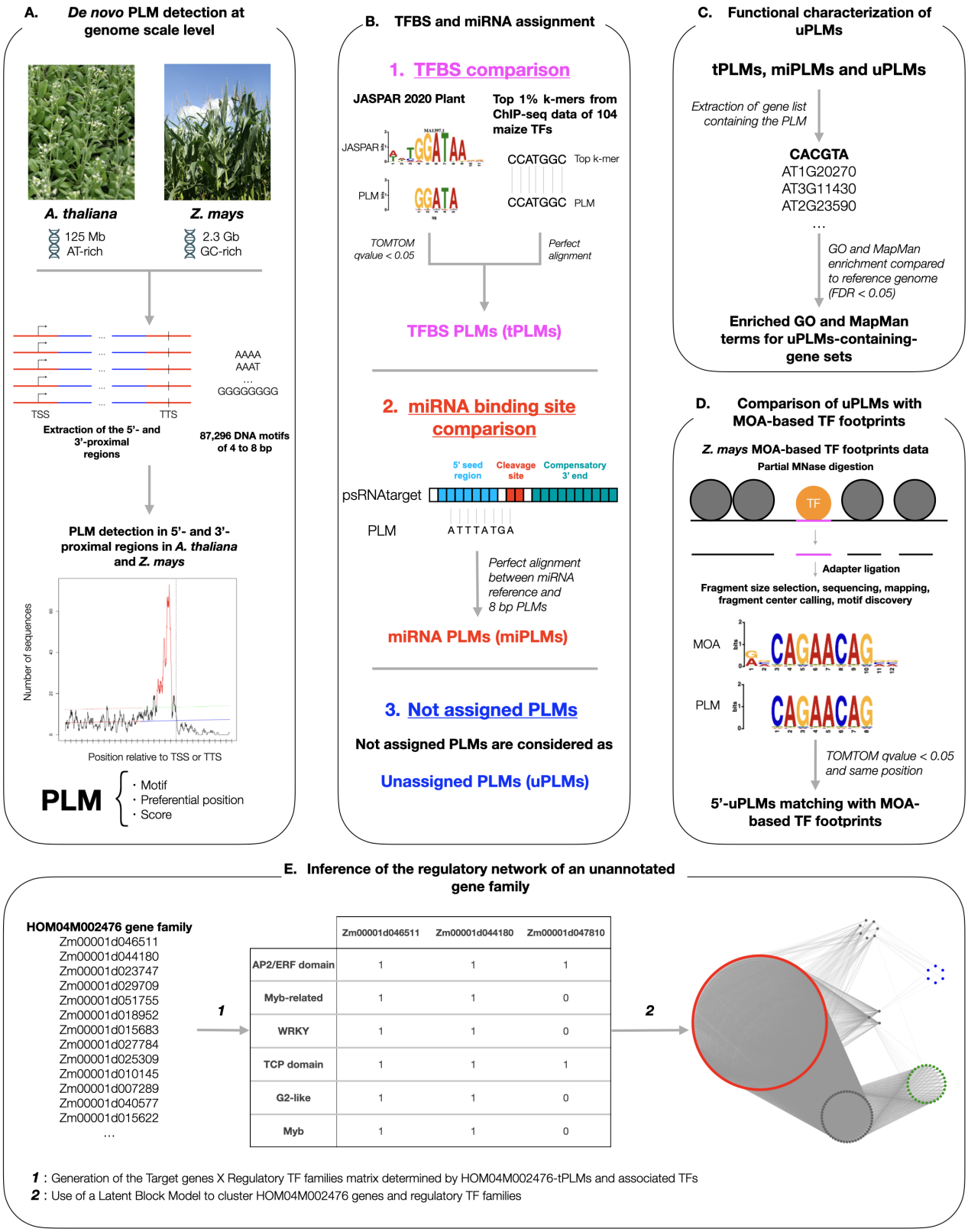
	A. thaliana		*Z. mays*	
	5'	**3'**	**5'**	**3'**
Us	48 (5%)	1 (0.1%)	402 (43%)	6 (1%)
1	621 (66%)	108 (16%)	148 (16%)	121 (21%)
2	103 (11%)	61 (9%)	77 (8%)	67 (11%)
3	80 (8%)	399 (59%)	75 (8%)	233 (40%)
Ds	95 (10%)	77 (11%)	242 (26%)	102 (17%)
NA	–	33 (5%)	–	58 (10%)
Total	947	679	944	587

Table 2. Number of miPLMs detected in the different groups. Us: region located upstream the three groups; 1, 2 and 3: group 1, 2 and 3; Ds: region located downstream the three groups; NA: region located between the groups when they are not juxtaposed.

	A. thaliana		*Z. mays*	
	5'	**3'**	**5'**	**3'**
Us	38 (7%)	0 (0%)	111 (35%)	0 (0%)
1	150 (28%)	146 (21%)	13 (4%)	105 (50%)
2	129 (24%)	134 (19%)	19 (6%)	24 (11%)
3	93 (17%)	257 (37%)	51 (16%)	17 (8%)
Ds	135 (25%)	64 (9%)	121 (38%)	39 (19%)
NA	–	92 (13%)	–	25 (12%)
Total	545	693	315	210

Table 3. Number of uPLMs detected in the different groups. Us: region located upstream the three groups; 1, 2 and 3: group 1, 2 and 3; Ds: region located downstream the three groups; NA: region located between the groups when they are not juxtaposed.

	A. thaliana		*Z. mays*	
	5'	**3'**	**5'**	**3'**
Us	302 (5%)	7 (0.1%)	3,443 (40%)	57 (1%)
1	1,884 (34%)	1,431 (23%)	383 (4%)	1,179 (20%)
2	1,284 (23%)	1,226 (20%)	1,174 (14%)	953 (16%)
3	702 (13%)	2,535 (41%)	793 (9%)	1,575 (27%)
Ds	1,387 (25%)	518 (8%)	2,735 (32%)	1,285 (22%)
NA	–	395 (6%)	–	808 (14%)
Total	5,559	6,112	8,528	5,857



	Zm00001d046511	Zm00001d044180	Zm00001d047810
AP2/ERF domain	1	1	1
Myb-related	1	1	0
WRKY	1	1	0
TCP domain	1	1	1
G2-like	1	1	0
Myb	1	1	0

Fig. 1. The flowchart summarizing the key steps that allowed *de novo* and genome-wide quantification, characterization and exploitation of PLMs in gene-proximal regions of *A. thaliana* and *Z. mays*. **Step A**: *de novo* detection of PLMs in 5'- and 3'-proximal regions of genes in both plant species. Proximal sequences were extracted and scanned for a set of 4 to 8 bp long motifs. **Step B**: Comparison of each *de novo* PLM with resources of TFBS (i.e. JASPAR Plant 2020 (26) and top 1% k-mers from ChIP-seq data of 104 maize TFs (27)) and miRNA binding sites (i.e. psRNATarget (28)), leading to the identification of PLMs matching with TFBS (referred to as tPLMs) and/or miRNA binding sites (referred to as miPLMs). Unassigned PLMs were named uPLMs. **Step C**: Functional characterization of uPLMs-containing gene sets with GO and MapMan enrichments relative to the genome. **Step D**: Validation of 5'-uPLMs in *Z. mays* through comparison with MOA-based TF footprints data (16). **Step E**: Inference of the regulatory network of an unannotated gene family using 5'-tPLMs. For each gene, TFs able to bind to 5'-tPLMs were retrieved to build the Target genes X Regulatory TF families matrix. A Latent Block Model was then used on this matrix to cluster the genes and TF families into groups.

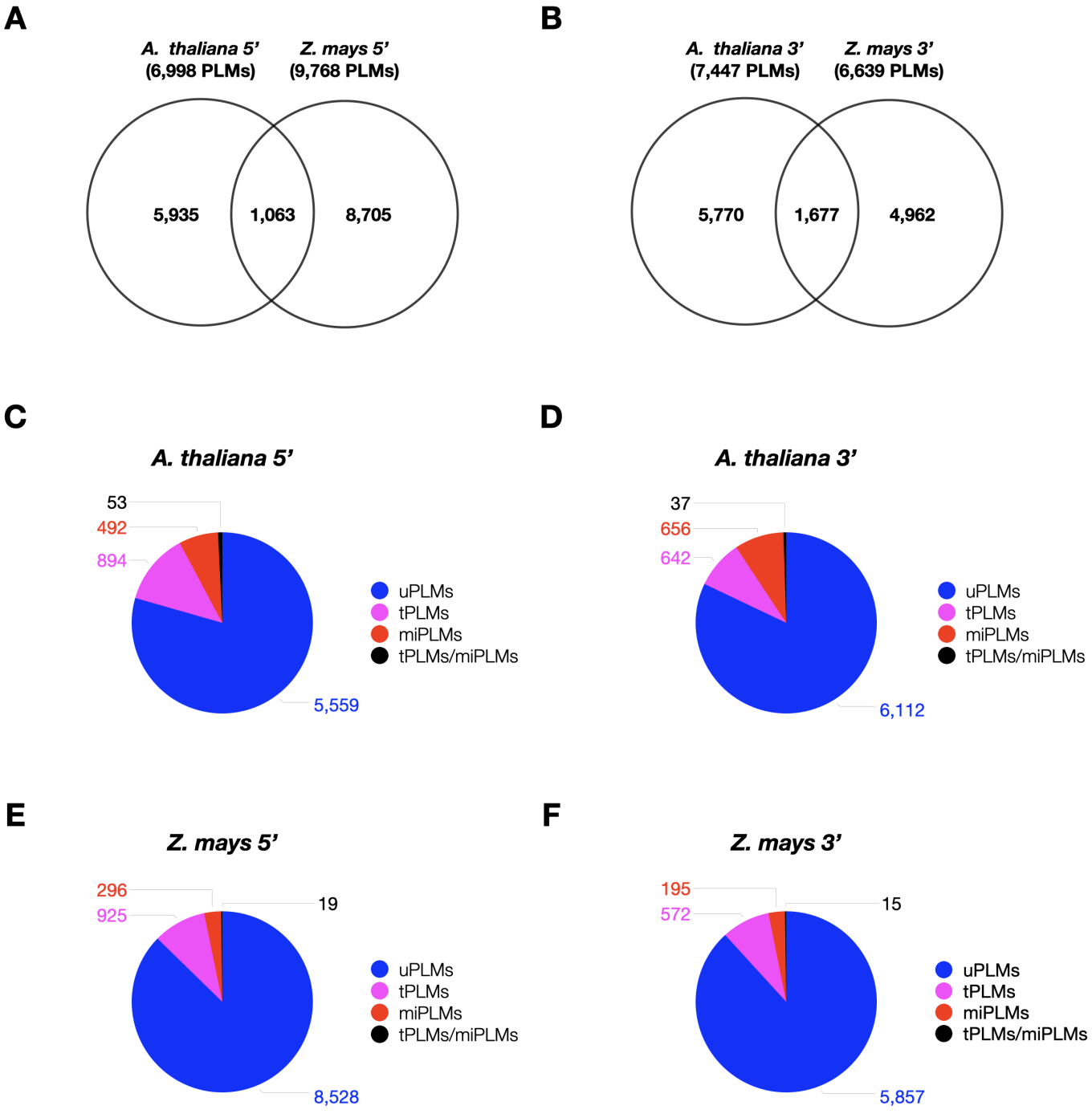


Fig. 2. Characterization of PLM content in gene-proximal regions of *A. thaliana* and *Z. mays*. **(A-B)** Venn diagram showing the extent of overlap between 5'-PLMs **(A)** or 3'-PLMs **(B)** for *A. thaliana* and *Z. mays*. **(C-E)** Dissection of PLM types identified in the 5'-gene-proximal region of *A. thaliana* **(C)** and *Z. mays* **(E)**. **(D-F)** Dissection of PLM types identified in the 3'-gene-proximal region of *A. thaliana* **(D)** and *Z. mays* **(F)**.

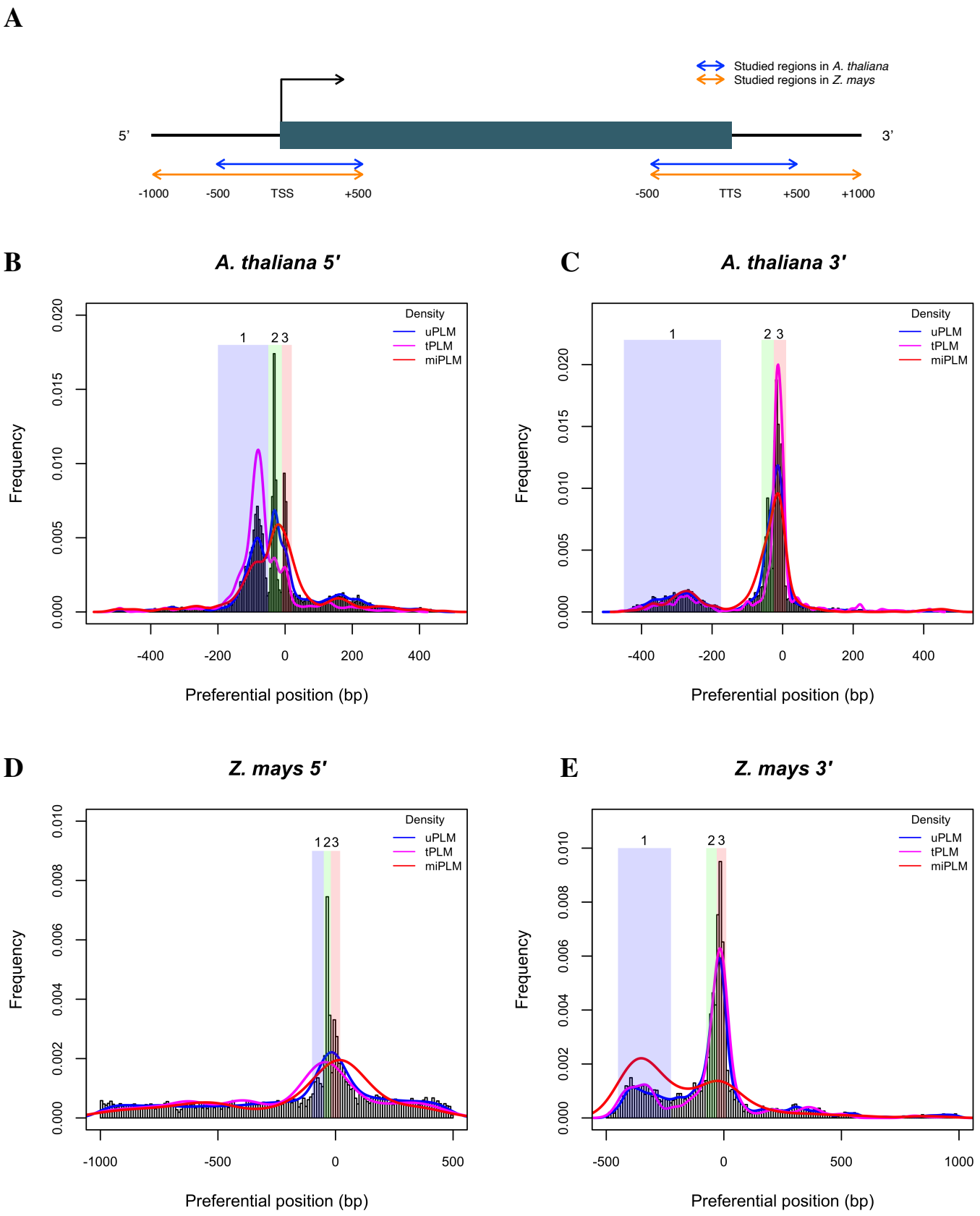


Fig. 3. Frequency of PLMs based on their preferential positions. (**A**) Schema of the studied regions with respect to the gene in *A. thaliana* and *Z. mays*. (**B**) 5'-PLMs in *A. thaliana*. Group 1: [-200;-50[; Group 2: [-50;-10[; Group 3: [-10;+20[. (**C**) 3'-PLMs in *A. thaliana*. Group 1:]-450;-175]; Group 2:]-60;-25]; Group 3:]-25;+10]. (**D**) PLMs in 5'-gene-proximal region in *Z. mays*. Group 1: [-100;-50[; Group 2: [-50;-20[; Group 3: [-20;+20[. (**E**) PLMs in 3'-gene-proximal region in *Z. mays*. Group 1:]-450;-225]; Group 2:]-75;-30]; Group 3:]-30;+10 bp].

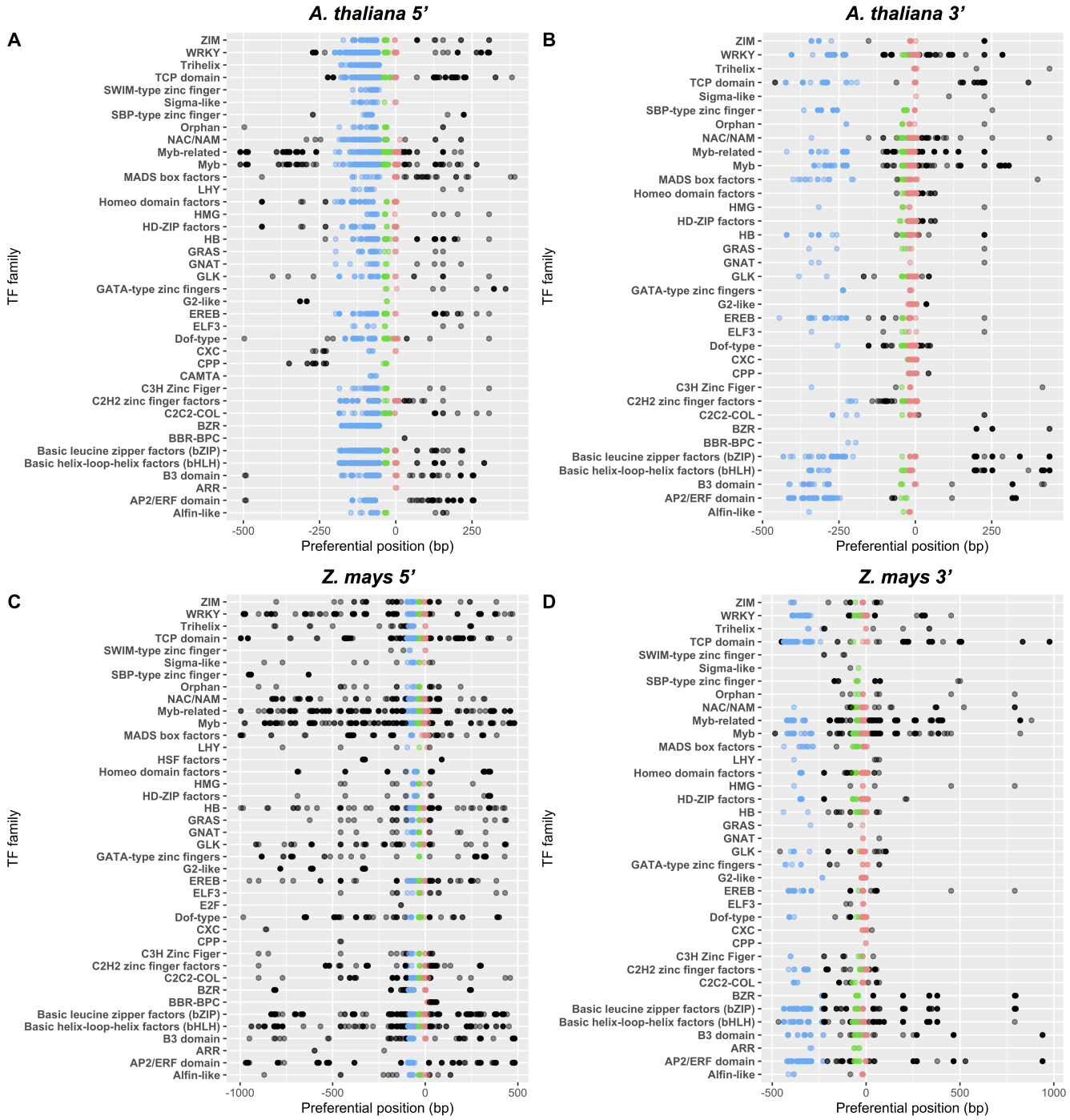


Fig. 4. tPLM preferential positions per TF family. tPLMs in 5'-gene-proximal (**A**) and 3'-gene-proximal regions (**B**) in *A. thaliana*. tPLMs in 5'-gene proximal (**C**) and 3'-gene proximal regions (**D**) in *Z. mays*. Blue, green and red points correspond to the tPLMs in groups 1, 2 and 3, respectively. The black points correspond the tPLMs that do not belong to any group. The opacity of the points is relative to the number of tPLMs at that position.

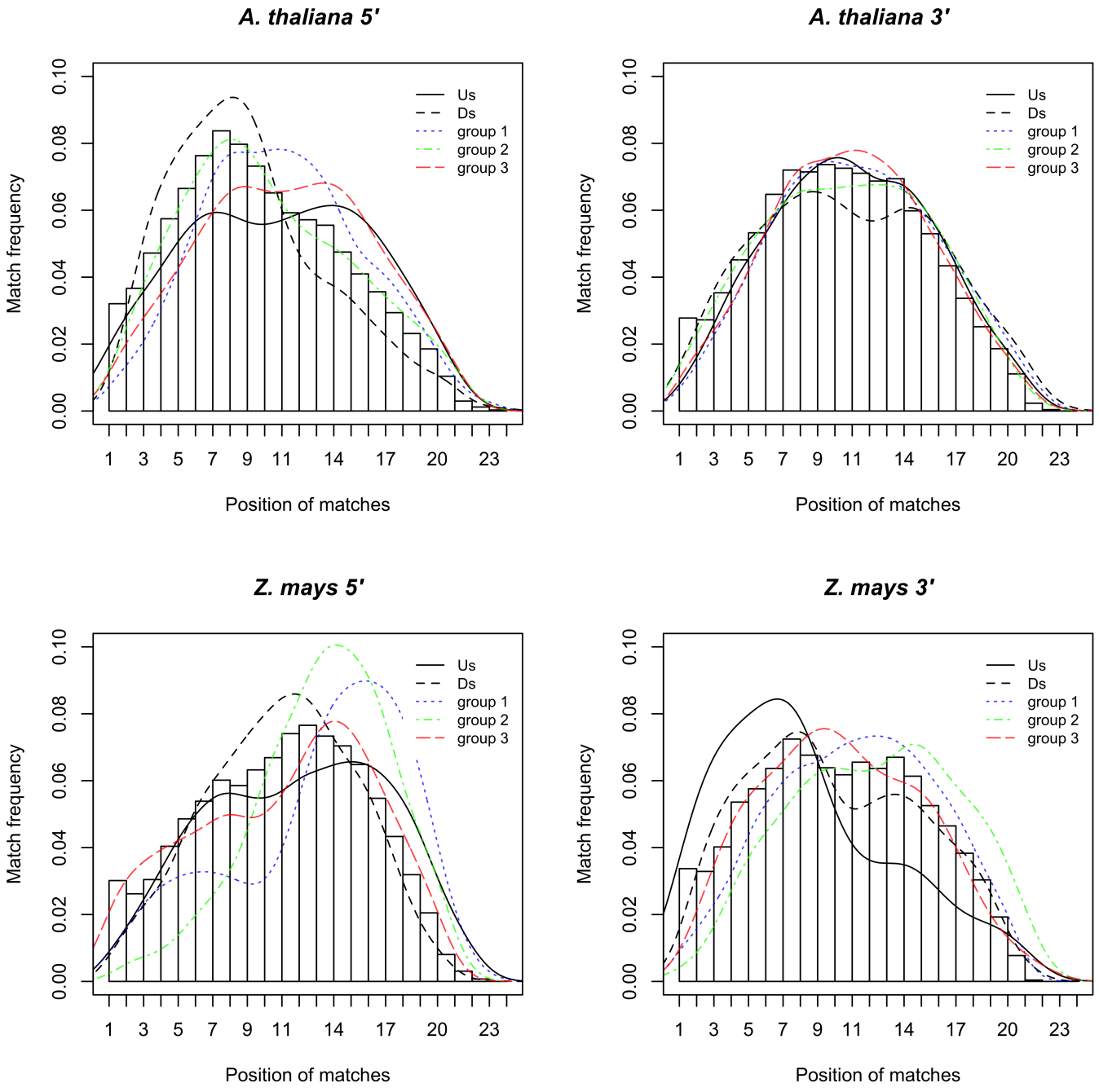


Fig. 5. Frequency of miRNA bases covered by miPLMs in gene-proximal regions of *A. thaliana* and *Z. mays*. The color curves indicate the densities of matched-ribonucleotide positions depending on whether the miPLM that matches belongs to one of the PLM groups (groups 1, 2 or 3) or is located upstream (Us) or downstream (Ds) of these groups.

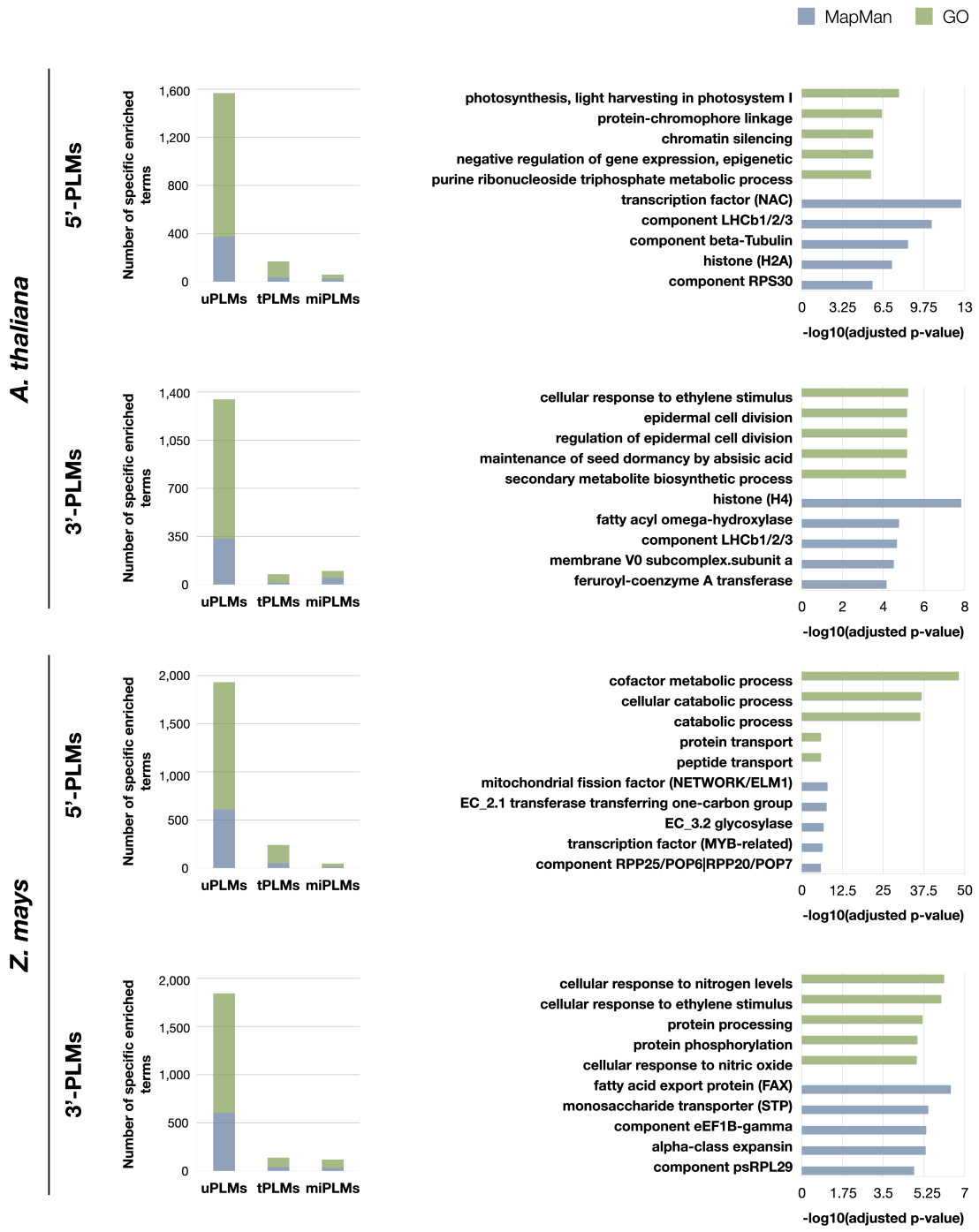


Fig. 6. GO and MapMan terms enriched specifically for each type of PLMs-containing gene sets. Values for GO and Mapman terms are shown in green and blue, respectively. On the left: histograms of the number of GO and MapMan terms enriched specifically for each type of PLMs-containing gene sets in the two species studied. On the right: the 5 most enriched terms specifically for uPLMs. The bar values of the histograms indicate the -log(adjusted p-values) for each term. MapMan terms have been truncated at the maximum precision level. The corresponding integer MapMan terms are as follows: **transcription factor (NAC)**: RNA biosynthesis.transcriptional regulation.transcription factor (NAC); **component LHCb1/2/3**: Photosynthesis.photophosphorylation.photosystem II.LHC-II complex.component LHCb1/2/3; **component beta-Tubulin**: Cytoskeleton organisation.microtubular network.alpha-beta-Tubulin heterodimer.component beta-Tubulin; **histone (H2A)**: Chromatin organization.histones.histone (H2A); **component RPS30**: Protein biosynthesis.ribosome biogenesis.small ribosomal subunit (SSU).SSU proteome.component RPS30; **histone (H4)**: Chromatin organisation.histones.histone (H4); **fatty acyl omega-hydroxylase**: Cell wall organisation.cutin and suberin.cuticular lipid formation.fatty acyl omega-hydroxylase; **membrane V0 subcomplex.subunit a**: Solute transport.primary active transport.V-type ATPase complex.membrane V0 subcomplex.subunit a; **feruroyl-coenzyme A transferase**: Cell wall organisation.cutin and suberin.alkyl-hydrocinnamate biosynthesis.feruroyl-coenzyme A transferase; **mitochondrial fission factor (NETWORK/ELM1)**: Cell cycle organisation.organelle division.mitochondrion and peroxisome division.mitochondrial fission factor (NETWORK/ELM1); **EC_2.1 transferase transferring one-carbon group**: Enzyme classification.EC_2 transferases.EC_2.1 transferase transferring one-carbon group; **EC_3.2 glycosylase**: Enzyme classification.EC_3 hydrolases.EC_3.2 glycosylase; **transcription factor (MYB-related)**: RNA biosynthesis.transcriptional regulation.MYB transcription factor superfamily.transcription factor (MYB-related); **component RPP25/POP6|RPP20/POP7**: RNA processing.ribonuclease activities.RNA-dependent RNase P complex.component RPP25/POP6|RPP20/POP7; **fatty acid export protein (FAX)**: Lipid metabolism.lipid trafficking.fatty acid export protein (FAX); **monosaccharide transporter (STP)**: Solute transport.carrier-mediated transport.MFS superfamily.SP family.monosaccharide transporter (STP); **component eEF1B-gamma**: Protein biosynthesis.translation elongation.eEF1 aminoacyl-tRNA binding factor activity.eEF1B eEF1A-GDP-recycling complex.component eEF1B-gamma; **alpha-class expansin**: Cell wall organisation.cell wall proteins.expansin activities.alpha-class expansin; **component psRPL29** Protein biosynthesis.organelle machinery.plastidial ribosome.large ribosomal subunit proteome.component psRPL29.

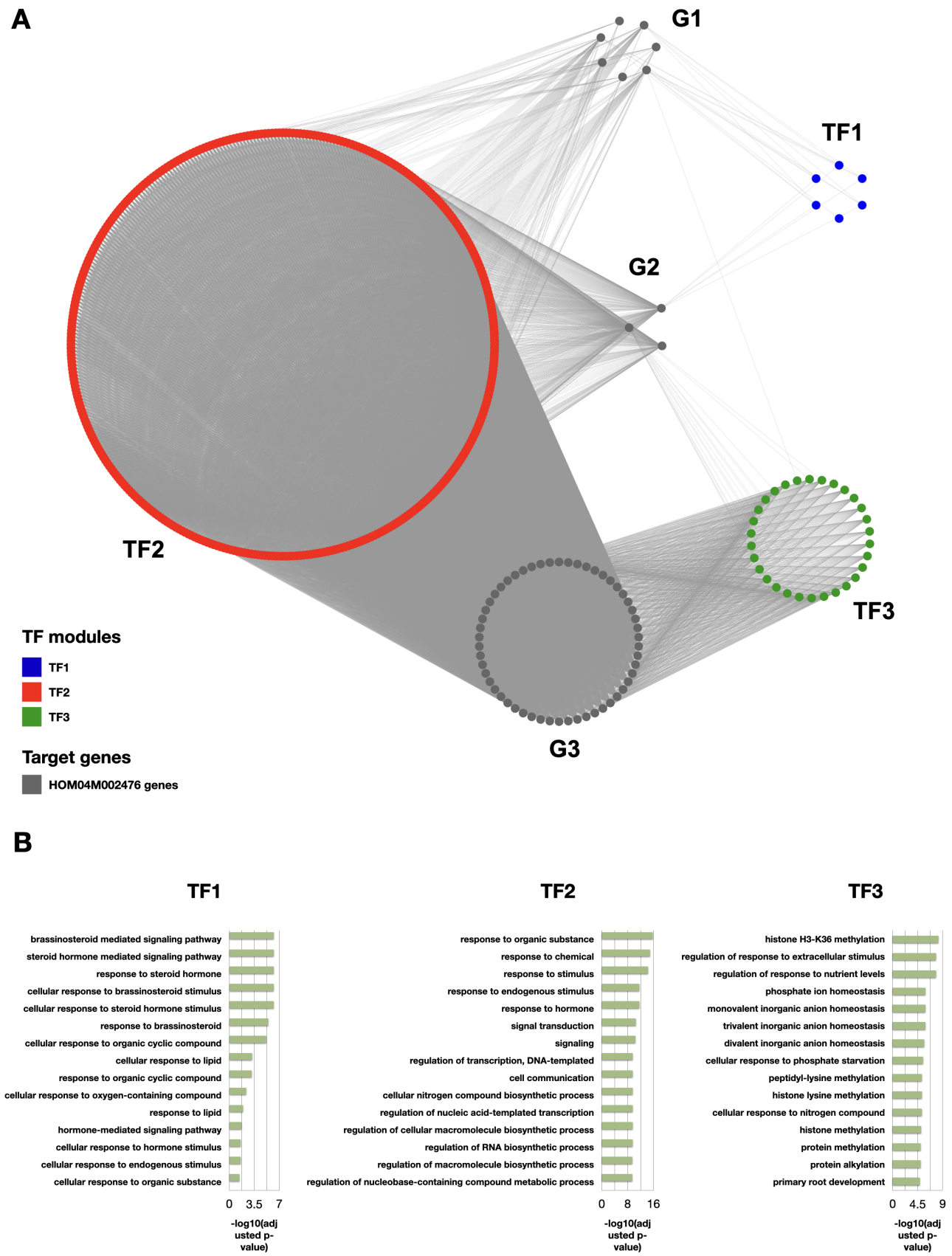


Fig. 7. Regulatory network and sub-families identification for the HOM04M002476 gene family of *Z. mays*. **(A)** Infered regulatory network of HOM04M002476 gene family. **G1, G2, G3** correspond to the three HOM04M002476 gene groups identified by the LBM procedure. They are highlighted in gray. **TF1, TF2, TF3** correspond to the three TF groups identified by the LBM procedure that could be involved in the regulation of HOM04M002476 genes. They are highlighted in blue, red or green. **(B)** Specific GO Biological Process term enrichment of TF1, TF2 and TF3 groups. Histograms represent the 20 most enriched GO Biological Process terms specifically for each TF group compared to the others with their -log10 of the adjusted *P*-values with the Benjamini-Hochberg procedure.

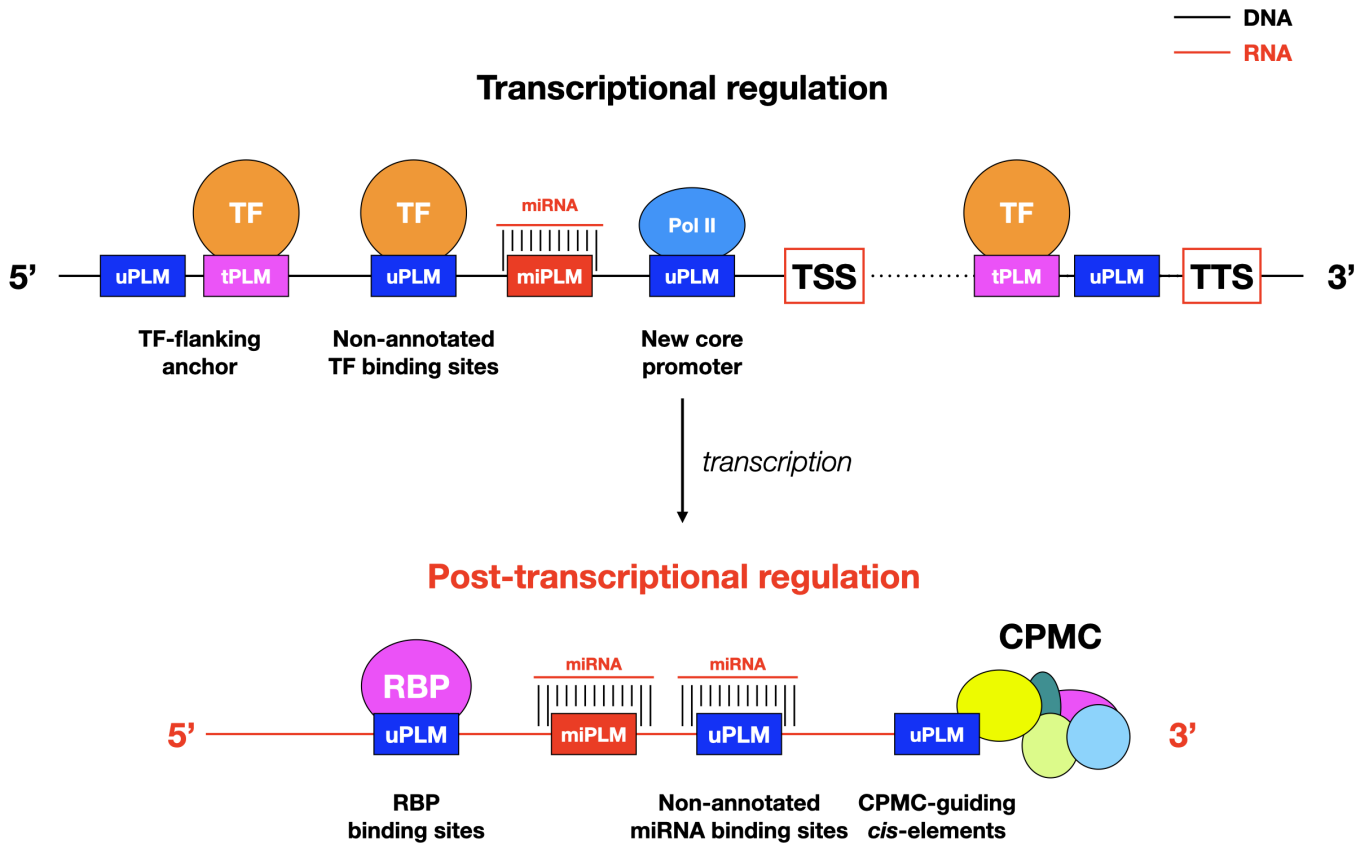


Fig. 8. Proposed models for transcriptional and post-translational control by PLMs in plants. The models integrate only the results presented here. Possible interactions between the illustrated components and the diverse array of other intermediates such as enhancers and long-distance chromatin interaction sites, remain to be investigated. CPMC: cleavage and polyadenylation molecular complex; Pol II: RNA polymerase II; RBP: RNA-binding proteins.