HAL open science

Tuning promoter boundaries improves regulatory motif discovery in nonmodel plants: the peach example

Najla Ksouri, Jaime Castro-Mondragón, Francesc Montardit-Tarda, Jacques van Helden, Bruno Contreras-Moreira, Yolanda Gogorcena

To cite this version:

Najla Ksouri, Jaime Castro-Mondragón, Francesc Montardit-Tarda, Jacques van Helden, Bruno Contreras-Moreira, et al.. Tuning promoter boundaries improves regulatory motif discovery in nonmodel plants: the peach example. Plant Physiology, 2021, 185 (3), pp.1242-1258. 10.1093/plphys/kiaa091 . hal-03243277

HAL Id: hal-03243277

https://amu.hal.science/hal-03243277

Submitted on 31 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - NoDerivatives 4.0 International License

doi:10.1093/plphys/kiaa091 PLANT PHYSIOLOGY 2021: 185: 1242–1258

Plant Physiology

Research Article

Tuning promoter boundaries improves regulatory motif discovery in nonmodel plants: the peach example

Najla Ksouri ,[1] Jaime A. Castro-Mondragón ,[2,3] Francesc Montardit-Tarda,[1] Jacques van Helden ,[2,4] Bruno Contreras-Moreira [5,6,†,‡] and Yolanda Gogorcena [1,*,‡]

1 Laboratory of Genomics, Genetics and Breeding of Fruits and Grapevine, Estación Experimental de Aula Dei-Consejo Superior de Investigaciones Científicas, Zaragoza, Spain
2 Aix-Marseille Univ, INSERM UMR_S 1090, Theory and Approaches of Genome Complexity (TAGC), F-13288 Marseille, France
3 Centre for Molecular Medicine Norway (NCMM), Nordic EMBL Partnership, University of Oslo, 0318 Oslo, Norway
4 CNRS, Institut Français de Bioinformatique, IFB-core, UMS 3601, Evry, France
5 Laboratory of Computational and Structural Biology, Department of Genetics and Plant Production, Estación Experimental de Aula Dei–Consejo Superior de Investigaciones Científicas, Zaragoza, Spain
6 Fundación ARAID, Zaragoza, Spain

*Author for communication: aoiz@eead.csic.es
†Present address: European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK
‡Senior authors.

N.K., B.C.-M., and Y.G. devised the study objectives, designed the experiment, discussed data, and wrote the article. N.K. performed the bioinformatics analyses, and F.M.-T. helped delimit the proximal promoter region. J.A.-C.M. aided in the preparation of figures and provided critical feedback. J.v.H. contributed to the critical discussion of results. Y.G. and B.C.-M. contributed tools, data, and supervised the activities. All authors read and approved the article.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (https://academic.oup.com/plphys) are: Yolanda Gogorcena (aoiz@eead.csic.es), Bruno Contreras-Moreira (bcontreras@eead.csic.es).

Abstract

The identification of functional elements encoded in plant genomes is necessary to understand gene regulation. Although much attention has been paid to model species like Arabidopsis (*Arabidopsis thaliana*), little is known about regulatory motifs in other plants. Here, we describe a bottom-up approach for *de novo* motif discovery using peach (*Prunus persica)* as an example. These predictions require pre-computed gene clusters grouped by their expression similarity. After optimizing the boundaries of proximal promoter regions, two motif discovery algorithms from RSAT::Plants (http://plants.rsat.eu) were tested (oligo and dyad analysis). Overall, 18 out of 45 co-expressed modules were enriched in motifs typical of well-known transcription factor (TF) families (bHLH, bZip, BZR, CAMTA, DOF, E2FE, AP2-ERF, Myb-like, NAC, TCP, and WRKY) and a few uncharacterized motifs. Our results indicate that small modules and promoter window of [−500 bp, +200 bp] relative to the transcription start site (TSS) maximize the number of motifs found and reduce low-complexity signals in peach. The distribution of discovered regulatory sites was unbalanced, as they accumulated around the TSS. This approach was benchmarked by testing two different expression-based clustering algorithms (network-based and hierarchical) and, as control, genes grouped for harboring ChIPseq peaks of the same Arabidopsis TF. The method was also verified on maize (*Zea mays*), a species with a large genome. In summary, this article presents a glimpse of the peach regulatory components at genome scale and provides a general protocol that can be applied to other species. A Docker software container is released to facilitate the reproduction of these analyses.

Received October 1, 2020. Accepted December 7, 2020. Advance access publication January 13, 2021

© The Author(s) 2021. Published by Oxford University Press on behalf of American Society of Plant Biologists.

Open Access

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs licence (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial reproduction and distribution of the work, in any medium, provided the original work is not altered or transformed in any way, and that the work is properly cited. For commercial re-use, please contact journals.permissions@oup.com

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

Introduction

Peach (*Prunus persica* [L.] Batsch), a member of *Prunus* genus, is one of the best genetically characterized species within the Rosaceae family. With a small diploid genome (≃260 Mbp), and relatively short generation time (2–3 years), peach has become a model species for fruit genetic studies (Abbott et al., 2002). Obtaining elite genotypes with broad environmental adaptations and good fruit quality are the fundamental targets of all *Prunus* breeding programs (Gogorcena et al., 2020). To cope with environmental stimuli and ensure edible fruit development, a complex rearrangement of the gene expression network is required.

The modulation of gene expression is a complex process occurring at various levels, of which the transcriptional regulation is the core control code (Petrillo et al., 2014). The transcription machinery works as an interplay between DNA-binding proteins called transcription factors (TFs) and *cis*-regulatory elements (CREs). TFs bind short sequences known as TF-binding sites (TFBSs) located at CREs (e.g. promoters, enhancers, silencers). The different sites recognized by a TF are usually summarized as motifs or matrices. TFs may act as either activators or repressors of gene expression, leading to dynamic changes of the cellular pathways. For peach, annotation of TFs is available in resources such as the database plantTFDB (Tian et al., 2019).

As of November 2020, plantTFDB v5.0 stored 2,780 peach TFs classified into 58 families (http://planttfdb.cbi.pku.edu.cn). However, while much is known about TF families, most TFBSs and motifs are yet to be characterized. Deciphering the *cis*-regulatory network has become a prerequisite toward scoping out the foundations of transcriptional regulation in peach and other plants. The computational exploration of DNA motifs has been greatly stimulated by the availability of genomic data and the release of whole genome sequence assemblies (Verde et al., 2013, 2017). In this context, a variety of plant motif finders has emerged. Notwithstanding their value, they are hampered by certain limitations such as a restricted range of species (Promzea for maize [*Zea mays*] (Liseron-Monfils et al., 2013) and AthaMap for Arabidopsis [*Arabidopsis thaliana*] (Steffens et al., 2005)]; out-of-date databases (PlantCare, last updated in 2002; Lescot et al., 2002) or platforms allowing only the recovery of experimentally defined motifs (PlantPAN; Chang et al., 2008). Thereby, to circumvent these pitfalls, we have adopted a plant-customized tool for *de novo* motif discovery, Regulatory sequence analysis tools (RSAT)::Plants (http://plants.rsat.eu). RSAT has both a friendly user interface and command-line tools for versatile analyses in a wide collection of plants (Nguyen et al., 2018).

Cis-regulatory sequences have been studied in species such as Arabidopsis (Korkuc et al., 2014; Cherenkov et al., 2018), rice (*Oryza sativa*; Tonnessen et al., 2019), and maize (Yu et al., 2015; GalLi et al., 2018). In *P. persica*, there have been so far two motif discovery experiments: (1) a set of 350 dehydrin promoter sequences (Zolotarov and Strömvik, 2015) and (2) 30 heat responsive genes (Gismondi et al., 2020). In contrast to these case studies, we propose a structured bottom-up framework to identify statistically over-represented motifs on a genome scale. Our probabilistic approach relies on the hypothesis that genes within co-expressed modules are likely co-regulated by the same TFs. This approach has been successfully tested in many studies, for example, in Arabidopsis (Koschmann et al., 2012; Ma et al., 2013) and maize (Yu et al., 2015). According to Bianchi et al., 2015, an arbitrary defined segment of 1,500-bp upstream of the transcription start site (TSS) can be considered as the proximal promoter in peach. However, recent studies about the genomic delimitation of proximal promoters in *P. persica* effectively reduced this region to a window of approximately 500 nt (Montardit-Tarda, 2018).

The proposed pipeline, summarized on Figure 1, relies on four key ideas:

(1) an accurate definition of co-expressed gene modules;
(2) the identification of over-represented motifs as compared with a biologically meaningful background model;
(3) an assessment of the effect of upstream region length regarding the effectiveness of motif discovery; and
(4) disclosing the effect of splitting the analysis around the TSS site in discovering potential *cis*-elements.

All together, we demonstrated the utility of our strategy in analyzing genome-wide data to provide insights on gene regulation dynamics across tissues and specific conditions. In addition, the motifs predicted in this study can be browsed at https://eead-csic-compbio.github.io/coexpression_motif_discovery, where we provide readers with direct links to the results, a source code repository and a Docker software container to reproduce the analysis on any other plant species. A step-by-step tutorial for Web users is also available at https://github.com/RSAT-doc/motif_discovery_clusters.

Results

Identification of differentially expressed transcripts and module definition using weighted co-expression networks

After quality assessment and pseudo-alignment, an expression matrix was generated from eight published peach transcriptomes, including treated and control samples with their corresponding biological replicates. Differential analysis yielded 11,335 altered transcripts using $Q < 0.01$ and $|\beta| > 1$ thresholds. The number of differentially expressed transcripts (DETs) identified in each RNA-seq experiment is given in Table 1. Detailed information about quality control, pseudo-alignment and differential expression analyses is available on Table S1.

Expression values of 11,335 stress-related transcripts and 64 samples were used to construct the co-expression modules using the weighted gene co-expression network analysis (WGCNA) package. All samples and DETs were considered in the network construction, as neither outliers nor

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

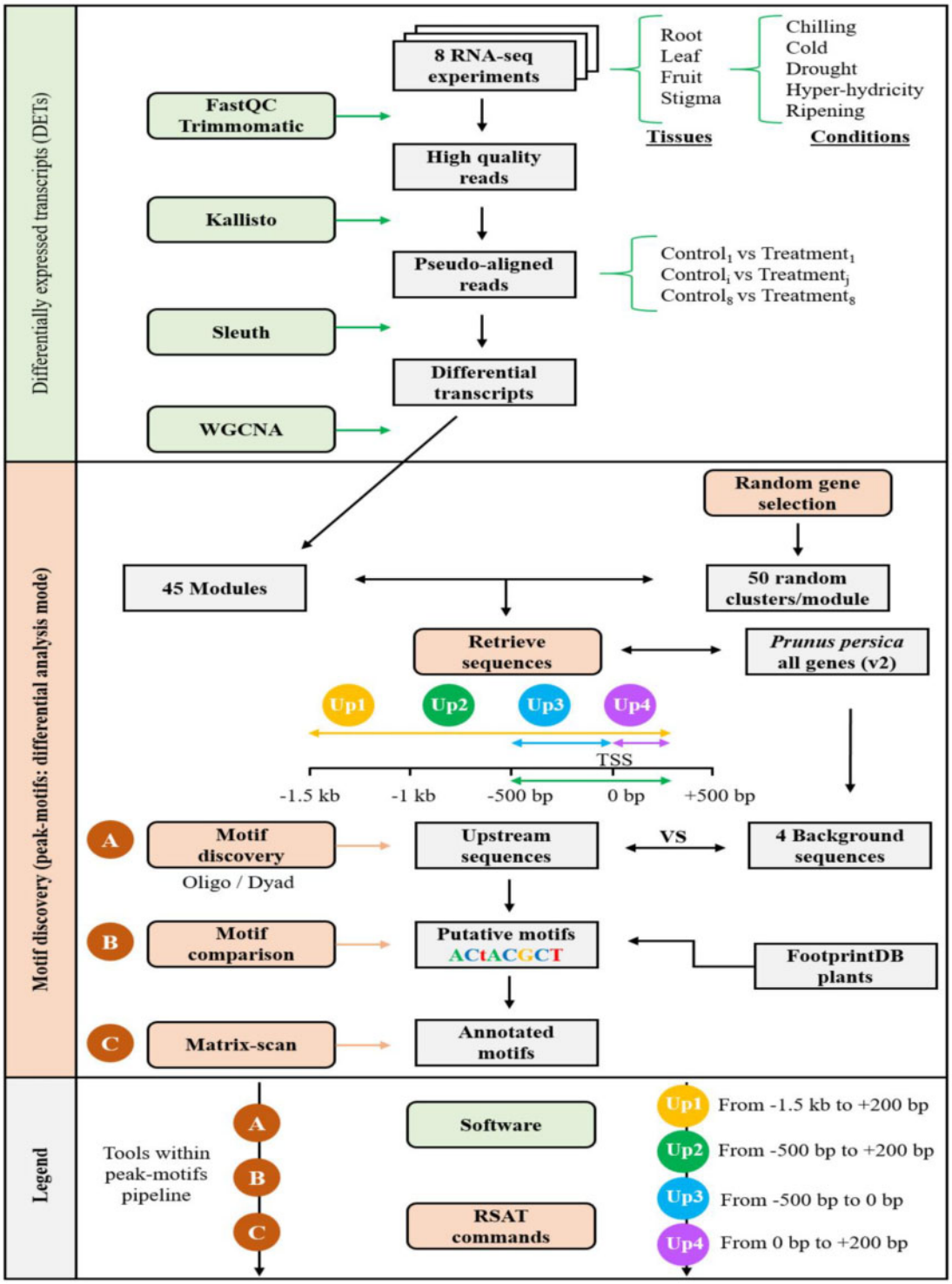

Figure 1 Bottom-up framework for *de novo* motif discovery. Step 1: differential expression analysis for transcript detection and extraction of co-expressed modules. Step 2: *de novo* motif detection using the peak-motifs tool from RSAT::Plants. Numbers correspond to the different tested upstream tracts, with TSSs anchored on position 0 bp, while letters represent tools within peak-motifs. Green and orange boxes label software used and RSAT tools, respectively.

transcripts with missing values were detected (Supplemental Figure S1A). Using a dynamic tree cut algorithm, 45 co-expression modules were retained with size ranging from 29 to 1,795 transcripts per module (Supplemental Figure S1B). The inter-connectivity of genes among modules was visualized using the topological overlap measure (TOM). Highly connected genes are highlighted in red color in Supplemental Figure S1C. Moreover, relatedness between the identified modules was also computed as the module eigengene (ME) measure, also known as the first principal

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

Table 1 Summary of RNA-seq data used as input and the number of DETs identified in each RNA-seq experiment

Project IDs	References	Experiments	Tissues	Conditions	DETs
PRJNA271307	(Li et al., 2015)	Ripening stage	Fruit	6	2,601
PRJNA288567	(Sanhueza et al., 2015)	Cold storage	Fruit	6	6,447
PRJNA248711	(Bakir et al., 2016)	Hyper hydricity	Leaf	2	15
PRJEB12334	(Ksouri et al., 2016)	Drought	Root/Leaf	4	350
PRJNA252780	(Jiao et al., 2017)	Low T°	Stigma	2	406
PRJNA323761	Unpublished	Drought	Root	2	1,118
PRJNA328435	Unpublished	Cold storage	Fruit	2	2,963
PRJNA397885	Unpublished	Chilling injury	Fruit	4	2,429

component of the module. The resulting plots are Supplemental Figure S1, D and E, where modules exhibiting high inter-connectedness are marked by progressively saturated blue and red colors. These findings, together with the membership analyses in Supplemental Figure S2, provide evidence that the resulting modules are consistent and might be biologically meaningful.

Prediction of CREs

Effect of proximal promoter length on prediction accuracy

As a first step toward extracting regulatory signatures, upstream region boundaries were defined from −1,500 to +200 bp relative to TSS (Up 1). Six out of 45 modules contained motifs with higher statistical significance than those detected in random clusters. Upstream regions of modules (M9, M10, M11, M18, M21, and M41) matched known core DNA-binding elements corresponding to Myb-like, BZR, CAMTA, bZip, and E2FE TF families. Modules with their corresponding regulatory elements are represented in Figure 2 and further information is provided in Supplemental Table S2. Motifs resulting from both oligo and dyad analysis correspond to signatures with strong confidence estimation. Note that while oligos are oligonucleotides of 6–8 bp, dyads are pairs of trinucleotides (monads) separated by a spacer of 0–20 bp. Moreover, eight poly (AT)-rich signals were discarded from M1, M2, M3, M4, and M6 due to their low complexity (Supplemental Table S2). Curiously, these (AT) patterns were also detected in the random clusters and their occurrence seemed to be associated with the size of the module (Supplemental Table S3). For instance, M1 is the largest module with 1,795 sequences, and (AT)-repetitive signals were detected in 40 out of the corresponding 50 random clusters of the same size.

When we restricted the motif discovery to the region with [−500 bp, +200 bp] boundaries (Up 2), 15 modules were found with statistically significant motifs. These were then grouped into 10 TF families as illustrated in Figure 2: TCP (Teosinte branched 1 (tb1, *Zea mays* (Maize)), [1] cycloidea (cyc; Antirrhinum majus; Garden snapdragon) [2] and PCF in rice (*Oryza sativa*)), bHLH (Basic helix-loop-helix), BZR (Brassinazole resistant), bZip (Basic Leucine Zipper), NAC (Derived from NAM (no apical meristem)), WRKY (conserved amino acid sequence at the N-terminus of the DNA-binding domain (DBDs)), AP2-ERF (APETALA2-Ethylene Responsive factor), Myb-like (V-myb avian myeloblastosis viral oncogene homolog), CAMTA (Calmodulin binding), and E2FE (Transcription factor E2FE).

Inspection of the major changes occurring when trimming the upstream segments to 500 bp resulted in interesting observations, summarized as follows. Spurious (AT)-rich sequences, considered as low-quality predictions, were limited to M2 and were replaced by relevant regulatory elements in M1, M3, M4, and M6 (Supplemental Table S2). Significant signals buried in the long upstream region (Up 1) were inferred in modules M5, M7, M24, M28, and M43 (Figure 2 and Supplemental Table S2). Besides, shortening the upstream promoter region size to 500 bp enhanced the statistical relevance of the predicted motifs, compared with the negative controls, regardless of the algorithm applied.

Overall, these findings suggest that shortening the upstream region increases the signal-to-noise ratio to detect biologically relevant motifs and, at the same time, reduces the occurrence of low complexity AT-rich motifs. In Figure 3, we illustrate a clear example of this observation. Indeed, with both oligo and dyad analysis, the statistical significance of motif E2FE found in Module M41 (black bars) noticeably increased compared with those identified in random clusters (gray bars). Hence, more significant motif discovery was accomplished in the window of [−500 bp, +200 bp].

Effect of splitting the promoter region around the TSS on motif prediction

Next, due to the difference in nucleotide composition in coding and noncoding regions, we subdivided the proximal promoter region in two segments around the TSS, with each interval examined separately: upstream, from −500 to 0 bp (Up 3), and downstream, from 0 to +200 bp (Up 4). By doing so, motifs of two additional TF families were discovered: BCP in module M1, and DNA-binding with one finger (DOF) in modules M7, M9, and M21. In contrast to BCP sites laying downstream the TSS (Up 4), DOF sites were found across both intervals (Figure 2 and Supplemental Table S2). Intriguingly, an uncharacterized motif was over-represented in Up 4—of module M25 requiring further research.

In total, 77 TF binding motifs were revealed from the scrutinized promoter regions (Supplemental Table S2). Modules with candidate predicted motifs can be classified in two types depending on their potentially matching TF. Indeed,

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

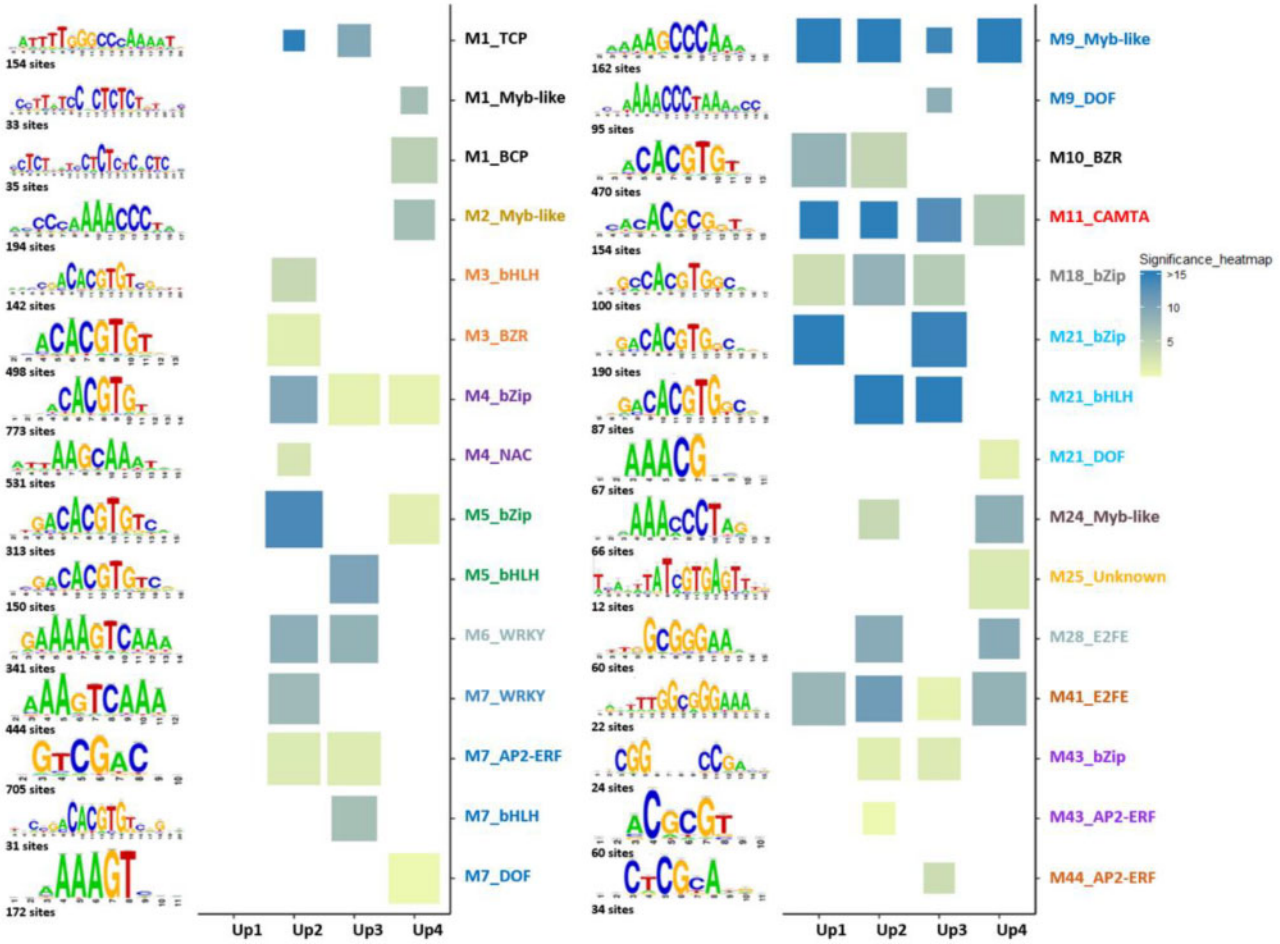

Figure 2 Position-specific scoring matrix (PSSM) representation of top scored discovered motifs per modules along different upstream lengths. The *x*-axis corresponds to the four intervals: Up 1: [−1,500 bp, +200 bp], Up 2: [−500 bp, +200 bp], Up 3: [−500 bp, 0 bp] and Up 4 [0 bp, +200 bp]. The *y*-axis informs about the motif family revealed per module. Cell colors indicate the statistical significance of the identified motifs while cell sizes represent the Ncor. Larger squares indicate high Ncor and thus high confidence annotations. Number of sites corresponds to the number of sites used to build the PSSM. When motifs from the same family are identified with both algorithms (oligo and dyad analysis), or in different upstream tracts (Up 1, Up 2, Up 3, and Up 4), only the most significant one is represented in the heatmap. Further details are provided in Supplemental Table S2. An interactive report with source code is accessible at https://eead-csic-compbio.github.io/coexpression_motif_discovery/peach/.

across the four examined upstream tracts, using both algorithms, we recognized those with motifs bound by a single TF family, considered as single TF-driven modules (e.g. M6, M11, M18, M28, and M41). Conversely, modules having multiple TFBS for several distinct TFs suggest a possible combinatorial regulation under particular circumstances. However, more evidence is needed to address this issue. Moreover, we observed that proximal region Up 2, defined from −500 to + 200 bp, yielded the highest number of significant CREs discovered in this study (Figure 2 and Supplemental Table S2).

Gene ontology enrichment analysis in modules of interest

Gene ontology (GO) analysis was conducted to interpret co-expression modules. For convenience, we present the top enriched biological terms in Table 2, and we provide details about cellular and molecular ontologies in Figure S3. Enriched processes in M1, M3, and M6 are the most informative. Indeed, transcripts in M1 were mostly over-represented in leaf tissue under drought, which is in line with the "photosynthesis" enrichment. Perturbations of ion effluxes are known to be stress-related, likely explaining the enrichment on "potassium ion transport" and "oxidation-reduction process" terms in M3 and M6, respectively. Conversely, "RNA processing," "translation," and "DNA metabolic process" terms, inferred, respectively, in modules M9, M24, and M41, are general terms that indicate a wide range of responses.

TFs annotation and prediction of their TFBS using footprintDB

To verify whether modules with predicted motifs might contain their potential binding TFs, gene-encoding TFs were

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

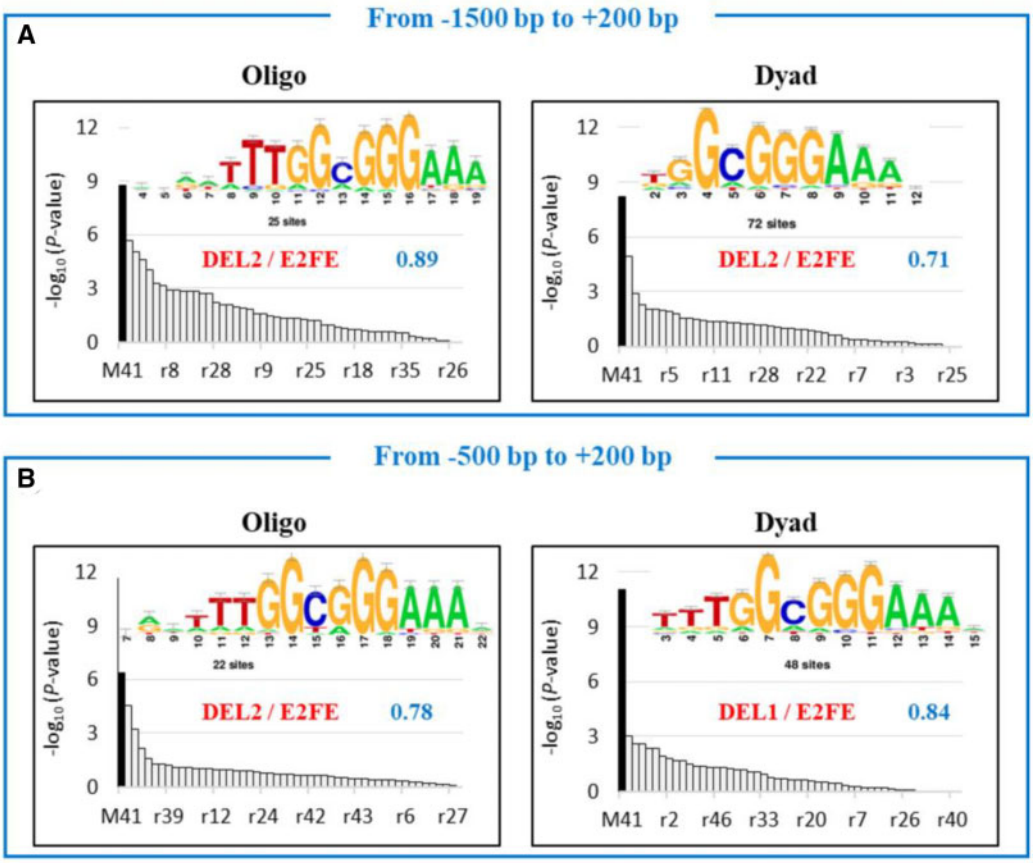

Figure 3 Illustrative comparison between predicted motif DEL2 (corresponding to E2FE transcription factor) along Up 1 and Up 2 upstream regions. A, Up 1: [−1,500 bp to +200 bp]. B, Up 2: [−500 bp to +200 bp] relative to the TSS. The name of the best match among plant motifs in footprintDB is labeled in red, next to its Ncor value labeled in blue. The *x*-axis corresponds to the module of interest (M41) and random clusters ranked by the most significant motifs. The *y*-axis corresponds to the statistical significance −log10 (*P*-value). Number of sites corresponds to the occurrence number of a single motif. The evidence supporting the putative motifs is Ncor (in blue) and the significance (black bars) when compared with negative controls (gray bars).

Table 2 Gene ontology enrichment in co-expression modules using PlantRegMap/PlantTFDB portal v5.0 and the adjusted *P*-value (FDR < 0.05)

Enriched modules	Biological GO IDs	GO terms	FDR values	Significance
M1 (1795)	GO:0015979	Photosynthesis	$4.9e^{-27}$	26.3
M2 (1224)	GO:0050896	Response to stimulus	$5.3e^{-04}$	3.3
M3 (864)	GO:0071804	Potassium ion transport	$5.4e^{-04}$	3.3
M6 (560)	GO:0055114	Oxidation reduction process	$3.5e^{-03}$	2.5
M9 (320)	GO:0006396	RNA processing	$5.8e^{-08}$	7.2
M11 (269)	GO:0010200	Response to chitin	$2.4e^{-10}$	9.6
M21 (151)	GO:1901700	Response to oxygen-compound	$4.0e^{-02}$	1.4
M24 (137)	GO:0006412	Translation	$2.7e^{-12}$	11.6
M41 (47)	GO:0006259	DNA metabolic process	$3.8e^{-14}$	13.4

Here, we only present the top enriched biological processes. More details about cellular and molecular ontologies are provided in Figure S3. Numbers in parenthesis indicate the number of genes per modules.

annotated and shortlisted in Figure 4. Subsequently, they were individually examined for their potential DNA motifs using footprintDB, and the results were compared with those obtained with RSAT. For consistency, control subsets of 50 random TFs selected from outside each module were additionally assessed. Motif-to-motif similarities between footprintDB and RSAT predicted matrices were computed using the Ncor score. Note this is an independent analysis, which does not use the clustered sequences; instead, it uses protein sequence of TFs. Our results revealed that consensus sequences predicted from the module corresponding TFs showed higher similarity to consensus sequences enriched in modules than those predicted from random TFs (Table 3 and Supplemental Figure S4). For instance, the binding motif tTTGGCGGGAAA enriched in module M41 is almost identical to E2FE-predicted site TTTTGGCGGGAAAA for the E2FE

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

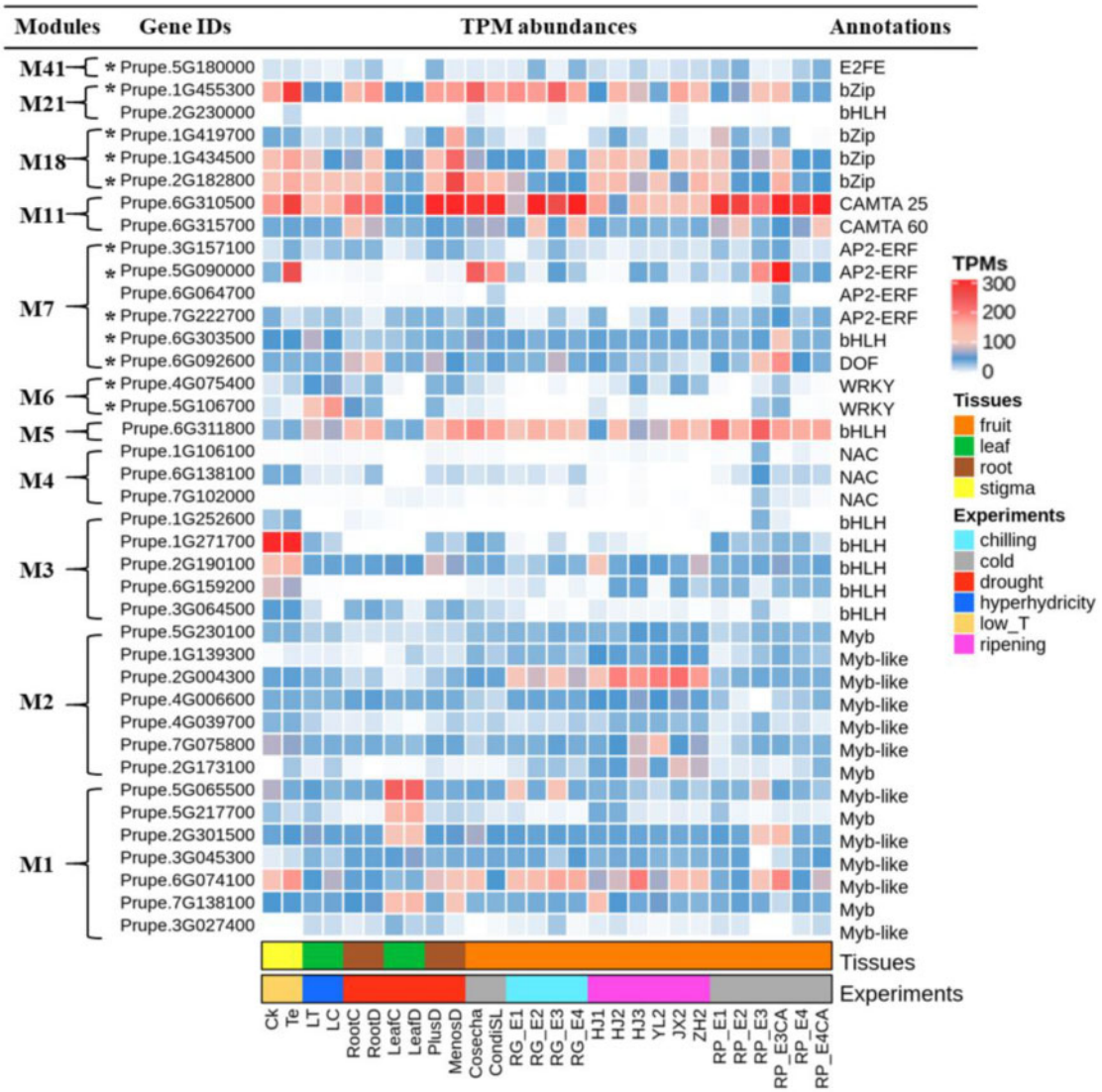

Figure 4 List of transcription factors within relevant modules. Blue and red squares indicate the abundance level of TPMs, while bottom color bars correspond to the tissue types and different experiments, respectively (see the legend at the right side of the figure). TFs showing sequence similarity between their footprintDB and RSAT predicted motifs are labeled with a star (Table 3). See Supplemental Table S6 for the abbreviations.

Table 3 Similarity comparison between RSAT and footprintDB DNA-binding motif predictions (matrix similarities were computed using cor ⩾ 0.7 and Ncor ⩾ 0.5)

Modules	RSAT Consensus	TFs	TF IDs	FootprintDB consensus	STAMP *E*-value
M41	tTTGGCGGGAAA	E2FE	Prupe.5G180000	tTTTTGGCGGGAAAA	$6e^{-119}$
M21	GACACGTGTC	bZip	Prupe.1G455300	ACGTGgc	$3e^{-20}$
M18	GCCACGTGGC	bZip	Prupe.1G419700	TGACGTGGC	$1e^{-16}$
M18	GCCACGTGGC	bZip	Prupe.1G434500	ACGTGGCa	$3e^{-19}$
M18	GCCACGTGGC	bZip	Prupe.2G182800	ACGTGKC	$4e^{-41}$
M7	GCCGACA	AP2-ERF	Prupe.3G157100	CCGaC	$2e^{-35}$
M7	GCCGACA	AP2-ERF	Prupe.5G090000	CCGACAT	$2e^{-64}$
M7	GCCGACA	AP2-ERF	Prupe.7G222700	CACCGACA	$1e^{-47}$
M7	CACGTGk	bHLH	Prupe.6G303500	CACGTGg	$7e^{-34}$
M7	aAAAGTc	DOF	Prupe.6G092600	AwAAAG	$1e^{-34}$
M6	GAAAAGTCAAAa	WRKY	Prupe.4G075400	aAAAGTCAA	$4e^{-63}$
M6	GAAAAGTCAAAa	WRKY	Prupe.5G106700	aAAAGTCAAC	$7e^{-49}$

The best predictions in footprintDB were selected in Arabidopsis. The TFs grouped in this table are the same labeled with a star in Figure 4.

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

TF in that module. This suggests that TF Prupe.5G180000 may modulate gene expression of M41 and that motif could be the *bona fide* binding site of this TF.

Motif scanning

To identify the position of regulatory sites (TFBSs) within proximal promoters of *P. persica* genes, position-specific scoring matrices of all candidate motifs (77) were scanned along Up 1 upstream intervals [−1,500 bp, +200 bp]. We observed a clear positional bias of TFBSs close to the TSS in the interval [−500 bp, +200 bp], progressively declining toward the 5′-end (Figure 5). For motifs detected, respectively, in Up 1 (yellow color), Up 2 (green), and Up 3 (blue), sites were notably concentrated upstream the TSS showing a bell-shaped distribution from −500 to +0 bp with a maximum of density around −250 bp. Conversely, the positional distribution of motifs predicted along Up 4 was biased toward downstream the TSS, with the flatter peak reaching its limit at the TSS (Up 4, purple). Detailed scanning results can be accessed at https://eead-csic-compbio.github.io/coexpression_motif_discovery/peach. Repetitive (AT) elements were also scanned to check their relevance, for example, whether they correspond to the TATA box. The underlying data included in Supplemental Figure S5 showed that TFBSs of these motifs were remarkably distant to the TSS and were distributed across the complete proximal region.

The performance of motif discovery using hierarchical clustering

To demonstrate the efficiency of our methodology, we reproduced the analyses described above with conventional clustering approaches within the clValid R package. After testing nine algorithms, hierarchical clustering (HC) with $k = 26$ gave the best score and was thus selected (see

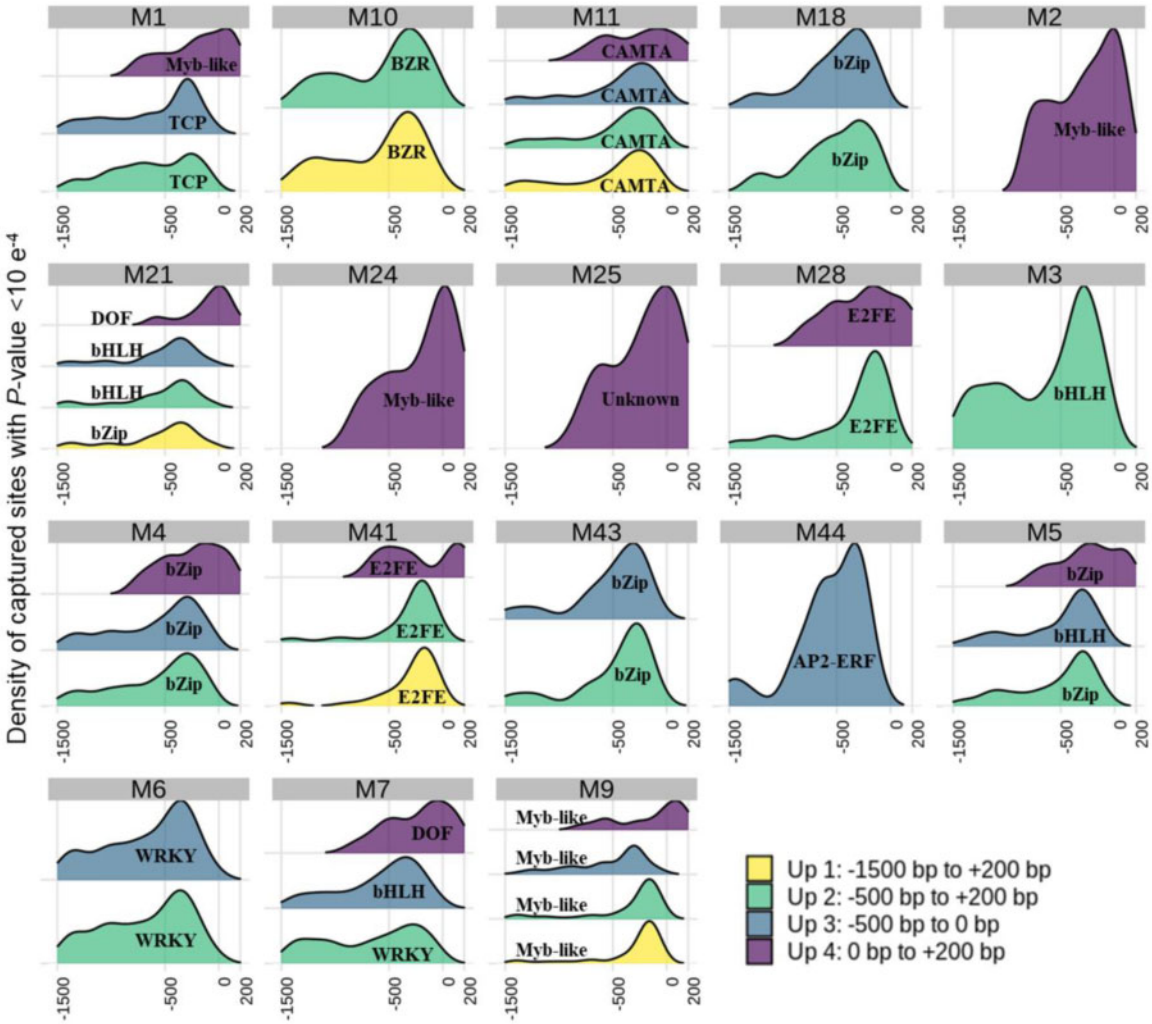

Figure 5 Positional distribution of the detected oligo motifs in promoters of *P. persica*. Four density distributions were derived from four assessed upstream regions. Up 1: from −1,500 bp to +200 bp, Up 2: from −500 bp to +200 bp, Up 3: from −500 bp to 0 bp and Up 4 from 0 bp to + 200 bp. The *x*-axis corresponds to upstream length in base pairs (bp). The *y*-axis corresponds to the density of captured sites with *P*-value $< 10\ e^{-4}$. Only oligo motifs are presented here; dyads are provided in the report at https://eead-csic-compbio.github.io/coexpression_motif_discovery/peach.

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

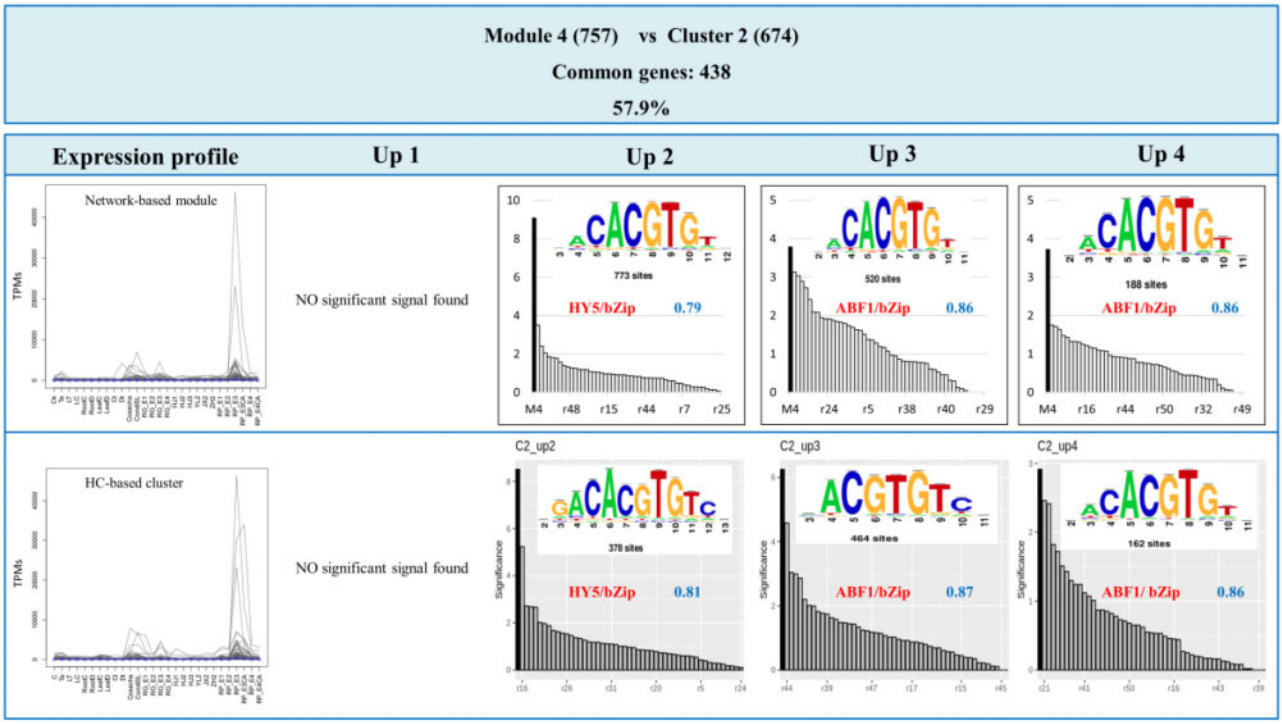

Figure 6 Comparison of putative DNA motifs identified from network-based module M4 and hierarchical cluster C2 along four different upstream regions. Up 1: [−1,500 bp, +200 bp], Up 2: [−500 bp, +200 bp], Up 3: [−500 bp, 0 bp], Up 4: [0 bp, +200 bp]. Motifs are represented by their logo. The name of the best match among plant motifs in footprintDB is labeled in red, next to its Ncor value in blue. Numbers in parenthesis indicate the number of genes in each module/cluster. The percentages correspond to shared genes, calculated using the number of genes in each module as denominator. For each paired network-based module and HC clusters, the gene expression profile is provided. TPMs refer to transcripts per million.

Supplemental Figure S6). Of these, we could match 14 clusters to previously discussed network-based modules having similar expression profile and sharing at least 15% of genes. In Figure 6, we present as an example the comparison of the motifs discovered in module M4 and hierarchical cluster C2, sharing 57.9% of genes. As expected, in this example and in 11/14 of the cases (see Supplemental Figure S7 and Supplemental Table S4), we were able to predict the same motif family in both network and HC strategies, which supports the robustness of the protocol proposed. In most cases, motifs identified in network-based modules were supported by a larger number of sites than those of hierarchical clusters (Supplemental Table S4). Note that a large module (M5) was divided in two clusters (C7 and C20), and in both cases a bZip motif was identified. In modules M25, M28, and M32, the differences in predicted motifs were due to the imbalance in gene numbers between clusters and their respective clusters (see Supplemental Table S4). More details of the hierarchical cluster benchmark can be found in the GitHub repository.

The performance of motif discovery using ChIPseq-based clusters

As a positive control, we tested whether our motif discovery protocol was able to identify the cognate regulatory sequences of genes tagged in ChIPseq experiments. Thus, in this section, we deal with superior quality clusters, as all genes in each considered study are known to be physically bound by a TF. As this type of data is not available in peach, we used public Arabidopsis datasets from 10 different TF families. Using curated data from the JASPAR database, we observed that the experimental motifs were successfully discovered by at least one algorithm in all datasets. The results are summarized in Figure 7A and Supplemental Table S5, where normalized similarity scores (Ncor) were used to compute the similarity between JASPAR and *de novo* motifs. In 8/10 cases Ncor values $\geqslant 0.7$ were obtained, despite the fact that proximal regions of variable length were used instead of ChIPseq peaks. Moreover, the most likely motifs recognized by the ChIPped TFs were estimated from their amino acid sequence with footprintDB (Figure 7A). The observed similarity between the different motifs underlines the predictive performance of the proposed methodology.

The performance of motif discovery on monocots

The goal of this section was to check the efficiency of our protocol on larger genome species. Four clusters of maize co-expressed genes with experimentally confirmed TFBSs were extracted from (Yu et al., 2015) and used to reproduce the analysis (ABI4, E2F1, Myb59, and WRI). Motifs among the first three modules were successfully detected "*de novo*." As shown in Figure 7B, the motifs discovered by RSAT

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

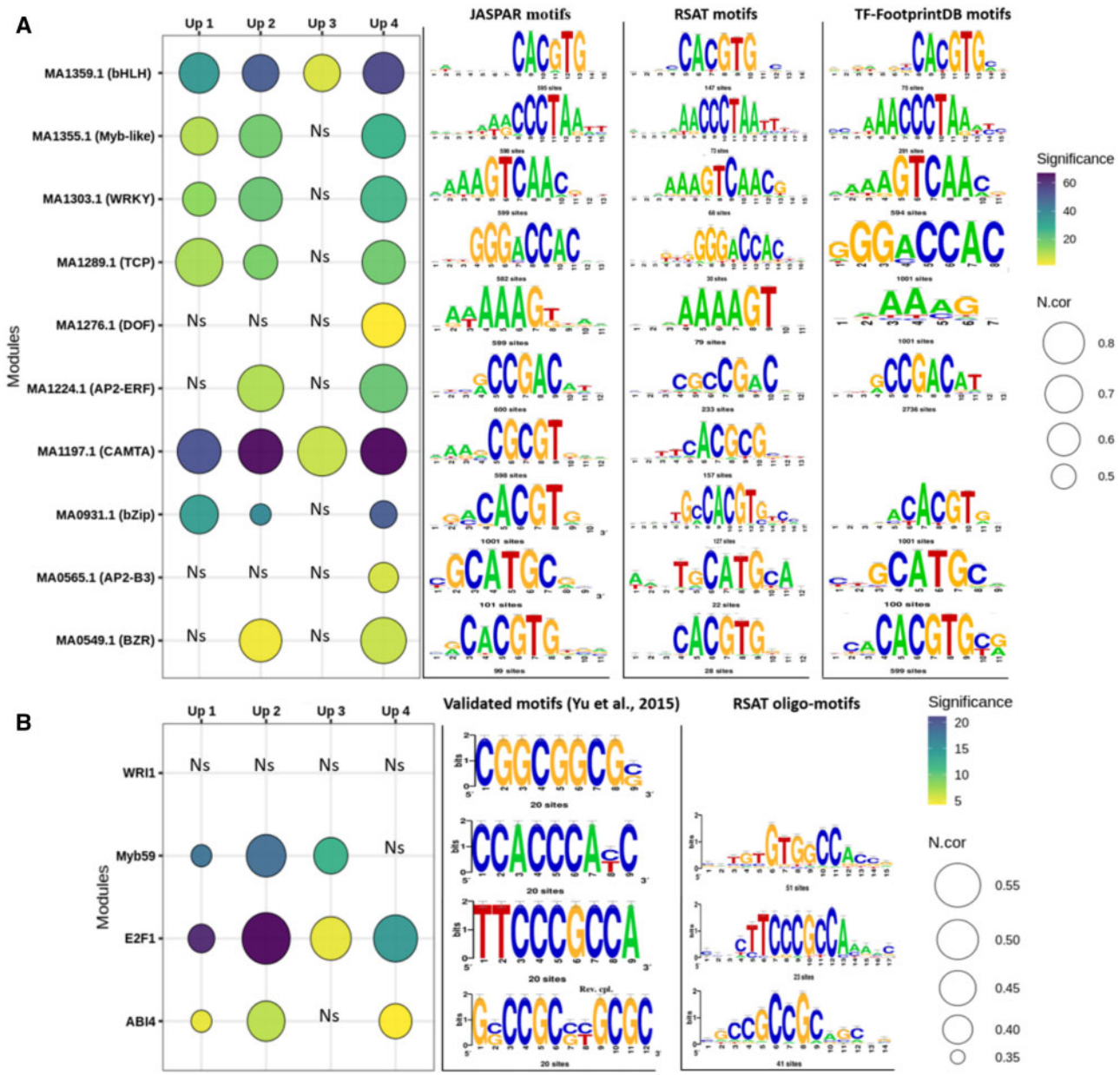

Figure 7 Similarity between experimentally validated motifs and *de novo* predicted oligo-motifs found along four different upstream regions in two different species. A, JASPAR motifs of Arabidopsis (considered as positive controls) were compared with RSAT-predicted motifs. Similarities between JASPAR and RSAT motifs were computed using the Ncor score. When motifs were identified in different upstream tracts (Up 1, Up 2, Up 3, and Up 4), only the most significant one was represented. TF-footprintDB sequence logos correspond to motifs predicted based on the protein sequence of the ChIPped TFs. B, Experimentally validated motifs in *Zea mays* were compared with RSAT predicted motifs. The *x*-axis corresponds to the four intervals: Up 1: [−1,500 bp, +200 bp], Up 2: [−500 bp, +200 bp], Up 3: [−500 bp, 0 bp], and Up 4 [0 bp, +200 bp]. The *y*-axis informs about the modules tested. Circle colors indicate the statistical significance of the identified motifs while the sizes represent the Ncor. Larger circles indicate high confidence annotations. NS stands for nonsignificant signal when compared with the 50 random clusters used as negative control.

showed higher significance within the Up 2 upstream region. Moreover, their logos were similar to those reported by Yu et al. (2015) with better matching scores within Up 2 windows (Ncor). Note that the WRI1 motif was also discovered but was not significant in the context of the negative controls.

Discussion

In this study, transcriptional profiling of eight independent data sets was conducted to decipher the intricate process of gene regulation in peach and to reveal meaningful biological signatures. DETs were grouped into 45 co-expression modules undergoing similar changes in their expression patterns.

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

Unlike conventional clustering methods (such as *k*-means and HC), which are based on geometric distances, WGCNA is a graph-based approach relying on network topology as inferred from the correlation among expression values (Li et al., 2018). In our hands, the WGCNA algorithm robustly and accurately defined modules within a complex multi-condition dataset.

Discerning regulatory signals from blocks of co-expressed genes is a common presumption used to identify functional genomic elements. It has been successfully applied and approved in various plant species like Arabidopsis (Koschmann et al., 2012; Ma et al., 2013), maize (Yu et al., 2015), and barley (*Hordeum vulgare*; Cantalapiedra et al., 2017). However, little is known about its applicability to woody species.

For each predicted module, two-motif discovery algorithms (oligo and dyad analysis) were run to discover significant motifs in proximal promoter regions. As suggested by Bianchi et al., we initially defined the promoter as an interval of [−1,500 bp to +200 bp] relative to the TSS (Bianchi et al., 2015). Discovered motifs with significant poly-(AT) sites were discarded due to their low complexity and scarcity of information concerning their specific regulatory function. We reasoned that low complexity sequences might be linked to repetitive stretches of DNA, extensively present in plant genomes (Yu et al., 2015). Interestingly, when tuning the promoter upstream length to a tract of [−500 bp, +200 bp] relative to the TSS, these low complexity motifs were limited to module M2. It would seem that long upstream promoter regions unbalance the signal-to-noise ratio, exacerbating the identification of such AT motifs. Along the same lines, we observed a dependence of (AT)-rich sites on the dataset size. Indeed, AT-low-complexity motifs were only detected in the first six modules, which contained upstream regions from 560 to 1,795 genes. In light of these considerations, we believe that in our study case, they may result in part due to the properties of DNA sequences (upstream region length, cluster size) rather than the performance of the chosen algorithm. In fact, our results (Supplemental Table S3) revealed that AT-rich occurrence in random clusters increases in parallel with the module size.

To check whether the AT-rich patterns overlap the TATA boxes, a positional scanning experiment was done. It is well documented in plants that a TATA box region lays between −30 and +35 bp with respect to the TSS (Zhu Qun et al., 1995; Smale, 2001). However, the scanning results portrayed that peaks were located far from this interval, confirming that they are distinct signals (Supplemental Figure S5).

Defining the promoter length has been a controversial issue for different reasons (Kristiansson et al., 2009). If the interval is too short or too long, the motif of interest may not be captured. Therefore, we reason that an analysis on regions of variable length would yield a more comprehensive picture of the complex regulatory code. By limiting the promoter length to a window of −500 bp, new regulatory motifs were recovered. Additionally, splitting the proximal promoter region into two intervals around the TSS enabled the discovery of further hidden candidate TF motifs (Figure 2 and Supplemental Table S2). Such observations may strengthen our hypothesis that shorter upstream regions improve the sensitivity of motif discovery (from 11 motif sequences identified within Up1 to 58 sequences identified in Up 3 and Up 4 assessed separately).

The spatial distribution of the occurrences of the 77 inferred motifs along the promoter region is crucial to understand gene regulation in *P. persica*. Our findings revealed that regulatory sites are not uniformly dispersed across the promoter but they exhibit a strikingly mixture of two density profiles: while the majority showed bell-shaped distribution at the interval of [−500 bp, 0 bp], others diverged downstream of the TSS [0 bp, +200 bp] (Figure 5). These findings are similar to those described in Arabidopsis, with nearly two-thirds of the examined TFBSs within the region from −400 to +200 bp (Yu et al., 2016). TFBSs of bHLH, BZR, TCP, and WRKY were particularly concentrated within the interval −500 to 0 bp. This denotes a positional binding preference within this proximal region, which is in agreement with (Yu et al., 2016) reporting that their positional preference is between −200 and 0 bp. On the other hand, bZip, CAMTA, E2FE, and Myb-like exhibited a dual binding distribution with central peaks upstream and downstream of the TSS. A possible explanation of this is that some TFs may display different binding preferences depending on their TF-specific structure, biological functions, or combinations with other TFs. The degree to which the arrangement of regulatory sites is associated to their function needs to be further investigated, especially since that kind of data is mostly limited to Arabidopsis (Zou et al., 2011; Yu et al., 2016). According to our findings, we may consider that the boundary from −500 to 0 bp is an adequate region to look for the majority of TFBS laying in the proximal promoter regions in peach. This region roughly overlaps the window with most potential binding sites predicted on Arabidopsis and rice by (Weirauch et al., 2014), although using *de novo* discovery instead of motif scanning. However, we should keep in mind that proximal TFBSs could also occur downstream the TSS. Thus, we suggest defining the peach proximal promoter length as a tract of [−500 bp to +200 bp], analyzing separately the two regions around the TSS for a better motif coverage. In fact, according to Montardit (2018), differences in the nucleotide composition are observed upstream and downstream the TSS. At this point, we should mention that gene regulation involves a complex interplay between the proximal (promoter) and distal regulatory regions located thousands of base pairs away from the TSS (e.g. enhancers; Li et al., 2019a). Our workflow sheds light mainly on sequence signatures extracted from the proximal promoter. Thus, it might not be adequate to study distal genomic elements.

Furthermore, rather than only returning a list of significant motifs, our methodology assigned them to different modules to help shape a clear overview of the peach regulatory code.

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

Overall, we were able to distinguish 18 modules harboring 77 motifs from 11 TF families: bHLH, bZip, BZR, CAMTA, DOF, E2FE, AP2-ERF, Myb-like, NAC, TCP, and WRKY. Although some modules, such as M6, M11, M28, and M41, seem to be driven by a single TF (WRKY, CAMTA, and E2FE, respectively), motifs from different families were annotated in the rest. This can be explained by the fact that some promoter sequences may encompass multiple TFBSs of perhaps interacting TFs. Indeed, TFs have been reported to frequently operate in combination (Guo et al., 2018). According to Reiter et al. (2017), cooperative binding of multiple TFs enables high-binding specificity and fine-tuning in gene regulation. Cooperative interactions between TFs in peach probably deserve further investigations.

From the inferred list of motifs (Figure 2), we found similar binding sequences potentially perceived by different classes of TFs. For example, motifs tGaCACGTGtc and GaCACGTGkCGg in module M5 are distinct but can be aligned despite different nucleotide frequencies in some positions. We presume that TFs from related families may have similar DNA recognition sequences, as reported for instance by Franco-Zorrilla et al. (2014) for Myb and AP2 TFs.

TFs from gene modules and their expression profile were analyzed. Transcript abundance of Myb TFs annotated in M1 and M2 was variable across the different conditions. This is consistent with MYBs playing various roles in plant development and metabolism (Li et al., 2019b). Regarding bHLH TFs, they were found in four modules with overall low abundance. However, in module M5, Prupe.6G311800 seems to be expressed in the fruit, leaf, and root. On the contrary, Prupe.1G271700 in module M3 is found expressed only in the stigma.

Gene encoding WRKYs were exclusively annotated in M6, and transcripts of (Prupe 5G.106700) were activated in leaf tissue under hyper-hydricity (HH) stress (Figure 4). It is well known that HH leads to morphological abnormalities, such as brittle leaves (Carrillo Bermejo et al., 2017). We speculate that WRKY may be involved in HH damage. In module M7, transcripts of a particularly AP2-ERF factor (Prupe.5G090000), were mainly over-represented in fruit tissue under cold stress. As described by Wang et al., (2017), low temperature leads to higher rate of ethylene production in peach, which triggers the AP2-ERF transcription machinery. Hence, we hypothesize that Prupe.5G090000 could effectively be implicated in peach adaptation to cold by acting as a key regulator of ethylene signaling.

Finally, bZip factors found in both M18 and M21 modules were highly expressed in most conditions, indicating that they may be involved in responses to various environmental cues.

Curiously, Prupe.4G075400 and Prupe.5G106700 (encoding WRKYs in Module M6), Prupe.6G092600 and Prupe.5G09000 (encoding, respectively, DOF and AP2-ERF in M7) and Prupe.5G180000 (corresponding to E2FE in M41) portrayed a great similarity between their individually predicted motif and those inferred in gene modules (Table 3). In other words, the regulatory elements detected are likely the *bona fide* target sites of those TFs. Experimental assays are needed to confirm these predictions.

Although the main contribution of this study is the optimization of promoter intervals for motif discovery, our protocol nonetheless depends on having pre-computed gene clusters on which to carry out the analyses. For this reason, we also tested it with clusters from different sources, not just network-based modules. First, we validated the protocol using a more conventional (HC) approach. The results indicated that in most cases where HC clusters matched network modules the same regulatory signatures were found (Supplemental Figure S7). These findings demonstrate that independently of the adopted clustering algorithm our pipeline performs consistently. In three notable exceptions (M25, M28, M32), we observed that although paired modules and HC clusters shared high percentages of genes, 30.4, 64.7, and 74.1%, respectively, predicted motifs were from distinct families. These results indicate that combining distinct clustering algorithms might improve the ability to detect regulatory signatures.

Nonetheless, we consider that modules derived from co-expression network analyses are ideally suited to the task of *de novo* motif finding in variable upstream region size. In most of the cases, putative motifs identified in gene modules have more sites than those identified in classical hierarchical clusters, indicating high confidence predictions. Concerning the sensitivity, we were able to detect 18 significant elements in modules while only 15 were revealed in HC clusters (Supplemental Table S4). Moreover, for 11,335 DETs, the clValid function required a longer execution time (nearly 48 h) to reveal the best clustering algorithm and the optimum number of clusters. In contrast, only a few minutes were needed to generate 45 modules using the WGCNA R package.

A second validation experiment was a positive control in which we analyzed ideal gene clusters that group genes tagged in 10 different Arabidopsis ChIPseq experiments. Comparing the *de novo* predicted motifs to the corresponding curated motifs in JASPAR we observed a high similarity in terms of Ncor scores (Figure 7A and Supplemental Table S5). Note, however, that the motifs were not identical, as instead of symmetrical chromatin peaks we analyzed proximal promoters of variable length, a setup that is more challenging than standard ChIPseq analysis. In fact, we observed that the choice of upstream region length affects the performance. In some cases, particularly Up 1 and Up 3, the expected motif was not even found. Unlike the results found in peach, examining four upstream tracts only returned motifs from the same query families, probably as a consequence of the JASPAR TFBSs profiles being curated.

Finally, we broadened the validation by including gene clusters from maize. The results in Figure 7B indicate that analysis of Up 2 yielded motifs which were more significant and matched better the experimentally verified motifs reported by Yu et al. (2015). This is consistent with our

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

findings in dicot species (peach and Arabidopsis) reporting a clear dependence between promoter length and prediction accuracy. Although Yu and colleagues cut promoter sequences in the range −1,000 to +200 bp, by applying our protocol we were able to define Up 2 as the most informative interval. Overall, these results suggest that our methodology is robust and can be extended to other plant species.

Conclusion

DNA motif discovery is a primary step for studying gene regulation, however, the *in silico* prediction of regulatory motifs is not straightforward. In contrast to the previous surveys that usually assume a fixed promoter length right at the start, this work reports regulatory elements while testing different upstream sequence intervals. It provides also a comprehensive collection of *P. persica* motifs without prior knowledge. By coupling modules definition and promoter analysis, we were able to extract interpretable information from a large set of noisy data and to reveal primary candidate TF target-binding sites responding to specific conditions. These results offer a more complete view of the proximal regulatory signatures in peach, and we believe that it may contribute to address the knowledge gap about the transcriptional regulatory code in nonmodel species.

Materials and methods

Input data and processing

Eight peach (*P. persica*) RNA-sequencing datasets were downloaded from the European Nucleotide Archive (https://www.ebi.ac.uk/ena) and were used as raw reads for this project. This comprehensive dataset includes data of various peach cultivars under different stress conditions and from various tissues: root ((Ksouri et al., 2016) and PRJNA323761), leaf (Ksouri et al., 2016; Bakir et al., 2016)), stigma (Jiao et al., 2017), and fruit ((Li et al., 2015; Sanhueza et al., 2015), PRJNA328435 and PRJNA397885). A detailed list of the project IDs and metadata are provided, respectively, in Table 1 and Supplemental Table S6. The obtained reads were quality-processed and trimmed using FASTQC v.0.11.5 and Trimmomatic v.0.36 (Bolger et al., 2014), to discard adaptors and low-quality sequences with mean Phred score ($Q < 30$) and window size of 4:15. The first nucleotides were then head-cropped to ensure a per-position A, C, G, T frequency near to 0.25. Following the trimming, only sequences longer than 36 bp were retained for further analysis. An overview of our complete workflow is shown in Figure 1 (see Step 1).

The high-quality reads from each RNA-seq project were quantified separately using the pseudo-aligner kallisto v.0.43.1 for fast and accurate transcripts count and abundance (Bray et al., 2016). Kallisto was run in two steps: (1) a transcriptome index was built from all cDNA transcripts of *P. persica* v2, from Ensembl Plants release 39 (Verde et al., 2017; Howe et al., 2020) and (2) each sample was pseudo-aligned against the index. Transcript-level abundance was estimated and normalized to transcripts per million (TPM) using 100 bootstraps (−b 100) to ascertain the technical variation. For single-end read mode, average fragment length and standard deviation were, respectively, set to (−l 200) and (−s 50).

Transcript-level profiling

Differential expression analysis was conducted with Sleuth R package v.0.29.0 (Pimentel et al., 2017) for each RNA data set separately. The Wald test was applied to the output abundance files to retain the significant expressed transcripts from each experiment. Samples and their biological replicates from each experiment were compared with their corresponding control. To reduce the false positives, only transcripts passing an FDR cutoff Q-value < 0.01 and beta statistic (approximation of the Log_2 Fold Change between two tested conditions) $|\beta| > 1$ were retained. Transcripts from all RNA-seq projects were normalized together using Sleuth function and then merged into a single list with an assigned mean TPM value for each treatment. These values were used in Figure 4. The Sleuth script is available in the GitHub repository.

Network-based detection of co-expressed modules

Based on the assumption that co-expressed genes may share the same biological signature, WGCNA v.1.61 was performed to extract clusters of densely interconnected genes named modules (Langfelder and Horvath, 2008). Transcripts were clustered to remove sample outliers and transcripts with missing entries. A similarity matrix was constructed by performing pairwise Pearson correlation across all targets. Then an adjacency matrix was built raising the similarity matrix to a soft power (β). Here β was set to 7 reaching thus 83% of the scale free topology fitting index (R^2). To minimize the effect of noise, matrix adjacency was transformed to topological overlap measure (TOM). Modules were defined as gene sets with high topological overlap with a minimum module size of 20 targets. Compared with standard HC, this approach solves the issue of setting the final number of clusters and arranges the genes based on their topological overlap to eliminate spurious associations resulting from the correlation matrix. Three diagnostic module functions were evaluated: The ME, intra-connectivity (Kwithin), and Module membership (MM). ME is considered as a representation of gene expression profiles and defined as the first principal component of a given module. Kwithin measures how connected a given gene is with respect to others of the same module, and MM calculates the correlation between gene expression values and ME (Langfelder and Horvath, 2008). MM values close to 1 or −1 indicate genes highly connected to the module.

De novo cis-regulatory sequences discovery using RSAT::Plants

Gene modules resulting from network analysis were subjected to *de novo* motif discovery pipeline using the RSAT::Plants standalone (Figure 1, Step 2). In our tests, this protocol required clusters with at least 15 sequences, as reported by Contreras-Moreira et al. (2016). For each

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

module, 50 random clusters with the same size were generated and used as a negative control as described previously (Contreras-Moreira et al., 2016). Sequences with four different boundaries around the TSS were retrieved from the genes in the co-expressed modules, random clusters and *P. persica* genome v2. The upstream sequences were defined as intervals of (1) −1.5 kb to +200 bp, (2) −500 to +200 bp, and (3) two segments around the TSS: −500 to 0 bp and 0 to +200 bp. Note that the 0 to +200 bp interval corresponds to the 3′UTR region, which is already downstream. RSAT peak-motifs was run under the differential analysis mode, where module's upstream sequences served as the test set and all upstream sequences from the peach genome were used to estimate the background model (Thomas-Chollier et al., 2012). A background model was created for each upstream stretch. Two discovery algorithms were used: (1) oligo analysis, which is based on the over-representation of *k*-mers in upstream regions (Helden et al., 1998) and (2) dyad analysis, which looks for over-represented spaced pairs of oligonucleotides (Helden et al., 2000). For each run, up to five motifs were returned per algorithm and were retained to compare their statistical significance with the 50 random clusters considered as negative control.

Candidate motifs were chosen based on their significance (log *E*-value) compared with negative control and were subsequently annotated by comparison to the FootprintDB collection of plant motifs (http://floresta.eead.csic.es/footprintdb; Sebastian and Contreras-Moreira, 2014) using the *compare-matrix* tool in RSAT (Nguyen et al., 2018). A normalized correlation score Ncor $\geqslant$ 0.4 was set to retain the best match.

Finally, selected motifs were scanned along the stretch [−1,500 bp, +200 bp] to predict their corresponding binding site positions, using as background model a Markov chain of order 1 (m = 1) and a cutoff $P \leqslant 1E^{-4}$.

TF prediction and GO analysis

Hereafter, the analysis was restricted to modules with significant detected signals. First, genes encoding TFs were predicted using the iTAK database (http://itak.feilab.net/cgi-bin/itak/index.cgi, last accessed September 2020). Their protein sequences were subsequently submitted to footprintDB to predict their putative target DNA-binding site (https://github.com/eead-csic-compbio/footprintDBclient). GO functional enrichment analysis was conducted using PlantRegMap/PlantTFDB portal v5.0 (http://planttfdb.cbi.pku.edu.cn, last accessed September 2020) and the adjusted *P*-value (FDR $<$0.05; Tian et al., 2019).

Conventional clustering of genes from expression data

To demonstrate that the proposed protocol is not strictly dependent on co-expression network analyses, we also generated gene clusters with conventional clustering methods in the clValid R package (Brock et al., 2008). Indeed, clValid allows the simultaneous comparison of multiple algorithms in a single function call. Furthermore, it can determine the best clustering method and the optimal number of clusters (k). *K*_min and *K*_max were, respectively, set up to 3 and 50. Hereafter, we use the term "module" to refer to a group of interconnected genes derived from network analysis and the term "cluster" to group of genes resulting from classical clustering. We paired clusters to network modules having similar expression profile and sharing at least 15% of the genes in the module. Clusters below that 15% cutoff were used as a negative control.

Positive control: gene clusters based on ChIPseq peaks

Peak sequences of 10 ChIPseq datasets from *A. thaliana* were downloaded from JASPAR database (Fornes et al., 2020). They were locally aligned with BLASTN against the Arabidopsis TAIR10.42 genome from Ensembl Plants to obtain the closest neighbor genes. The Blast parameters were as follows: *E*-value $\leqslant$ 1e^{-5}, max_target_seqs = 1, max_hsps = 1 query-coverage of 80% and percentage of identity 98%. Similarity between references (JASPAR) and *de novo* discovered motifs was computed with the normalized Ncor score (see above). Additionally, motif predictions based on the protein sequences of the ChIPped TFs were done using footprintDB.

Validation in monocot species

To study the reliability of this methodology on larger genomes, the analysis was widened to maize (*Z. mays*), a monocot with a genome one order of magnitude larger than peach. Four gene clusters with Electrophoretic mobility shift assay (EMSA)-confirmed motifs were retrieved from Yu et al.'s (2015) study (ABI4, E2F1, Myb59, and WRI1). Besides being strongly co-expressed, genes in each set shared common GO terms, indicating that may be involved in the same biological function. According to Yu et al. (2015), to reveal the basis of leaf regulatory network, maize clusters were defined from 22 leaf transcriptomes with developmental time series (from dry seeds to 192 h post imbibition). Putative motifs were predicted within [−1,000 bp, +200 bp] upstream sequences, and TF-TFBS interactions were verified using EMSAs. Among the defined clusters in the original work, herein ABI4, E2F1, Myb59, and WRI1 were selected based on their size. Indeed, to get reliable results, using clusters of at least 15 sequences is recommended. To convert genes from version Zea_mays.B73RefGen_v3 to the current Zea_mays.B73RefGen_v4.46 the gene ID history converter from Ensembl Plants was used (see Table S7).

Accession numbers

Genes referenced in this article can be found at the NCBI web page, under the genome section (https://www.ncbi.nlm.nih.gov/genome/?term=prunus+persica), and were retrieved from the Prunus_persica_NCBIv2 assembly (GCA_000346465.2).

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

Supplemental data

Supplemental Figure S1. Co-expression network analysis.

Supplemental Figure S2. Module membership (MM) raised to power 7 versus intramodular connectivity (Kwithin) plotted separately for each identified module.

Supplemental Figure S3. Gene ontology (GO) analysis of co-expression modules.

Supplemental Figure S4. Comparison of the motifs predicted with footprintDB to the corresponding enriched consensus in the promoters of modules M6, M7, M18, M21, and M41.

Supplemental Figure S5. Positional distribution of AT-rich repetitive motifs along upstream 1 (Up 1): [−1,500 bp, + 200 bp].

Supplemental Figure S6. Plots of the connectivity measure, Dunn, and silhouette indices for the nine tested clustering algorithms.

Supplemental Figure S7. Comparison of discovered DNA motifs identified from network-based modules and hierarchical-based clusters, along four upstream regions.

Supplemental Table S1. Number of surviving and dropped reads after quality processing and pseudo-aligned reads using kallisto (v.0.43.1).

Supplemental Table S2. List of candidate regulatory sites discovered within four upstream tracts of different lengths.

Supplemental Table S3. List of low complexity motifs considered as false positive predictions within a boundary from −1.500 to + 200 bp upstream region length.

Supplemental Table S4. Comparison of motifs identified in network-based modules and hierarchical clusters with similar expression profile.

Supplemental Table S5. Similarity of JASPAR motifs (considered as queries) and *de novo* predicted dyad motifs in Arabidopsis.

Supplemental Table S6. Detailed information about the RNA-seq data used for differential analysis.

Supplemental Table S7. List of gene IDs of maize clusters using the *Zea mays* reference genome *Zea_mays*. B73RefGen_v3 and the corresponding IDs in version *Zea_mays*.B73RefGen_v4.46.

Acknowledgments

We acknowledge support of the publication fee by the CSIC Open Access Publication Support Initiative through its Unit of Information Resources for Research (URICI). We thank Eric OLO NDELA for his support and help in creating the HTML report and providing useful feedback. We also thank Claudio Antonio Meneses Araya, Dayan Sanhueza, and Tomás Carrasco for giving us access to their RNA-seq data.

Funding

This work was partly funded by the Spanish Ministry of Economy and Competitiveness grants AGL2014-52063R, AGL2017-83358-R (MCIU/AEI/FEDER/UE); and the Government of Aragón with grants A44, A08_17R and A09_17R, which were co-financed with FEDER funds. N.K. was funded by project AGL2014-52063R, two months of Erasmus plus traineeship at Aix-Marseille University (AMU; Lab.Technological Advances for Genomics and Clinics (TAGC)) and at present a PhD contract awarded by the Government of Aragón.

Conflict of interest statement: The authors declare no conflict of interest.

References

Abbott AG, Georgi L, Yvergniaux D, Wang Y, Blenda A, Reighard G, Inigo M, Sosinski B (2002) Peach: the model genome for Rosaceae. Acta Hortic **575**: 145–155

Bakir Y, Eldem V, Zararsiz G, Unver T (2016) Global transcriptome analysis reveals differences in gene expression patterns between nonhyperhydric and hyperhydric peach leaves. Plant Genome **9**: 1–9

Bianchi VJ, Rubio M, Trainotti L, Verde I, Bonghi C, Martínez-Gómez P (2015) *Prunus* transcription factors: breeding perspectives. Front Plant Sci **6**: 1–20

Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics **30**: 2114–2120

Bray NL, Pimentel H, Melsted P, Pachter L (2016) Near-optimal probabilistic RNA-seq quantification. Nat Biotechnol **34**: 525–528

Brock G, Pihur V, Datta S, Datta S (2008) clValid: an R Package for cluster validation. J Stat Softw **25**: 1–22

Cantalapiedra CP, García-pereira MJ, Gracia MP, Igartua E (2017) Large differences in gene expression responses to drought and heat stress between elite barley cultivar Scarlett and a Spanish landrace. Front Plant Sci **8**: 1–23

Carrillo Bermejo EA, Alamillo MAH, Samuel David GT, Llanes MAK, Enrique C de la S, Manuel RZ, Rodriguez Zapata LC (2017) Transcriptome, genetic transformation and micropropagation: some biotechnology strategies to diminish water stress caused by climate change in sugarcane. Plant, Abiotic Stress and Responses to Climate Change. IntechOpen, pp 90–108

Chang WC, Lee TY, Huang H Da, Huang HY, Pan RL (2008) PlantPAN: plant promoter analysis navigator, for identifying combinatorial *cis*-regulatory elements with distance constraint in plant gene groups. BMC Genomics **9**: 1–14

Cherenkov P, Novikova D, Omelyanchuk N, Levitsky V, Grosse I, Weijers D, Mironova V (2018) Diversity of *cis*-regulatory elements associated with auxin response in Arabidopsis thaliana. J Exp Bot **69**: 329–339

Contreras-Moreira B, Castro-Mondragon JA, Rioualen C, Cantalapiedra CP, van Helden J (2016) RSAT::Plants: motif discovery within clusters of upstream sequences in plant genomes. *In* R Hehl, ed, Plant Synthetic Promoters Methods in Molecular Biology. Humana Press, New York, pp 279–295

Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe M, Baranašić D, et al. (2020) JASPAR 2020: update of the open-access database of transcription factor binding profiles. Nucleic Acids Res **48**: 87–92

Franco-Zorrilla JM, López-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. Proc Natl Acad Sci **111**: 2367–2372

Galli M, Khakhar A, Lu Z, Chen Z, Sen S, Joshi T, Nemhauser JL, Schmitz RJ, Gallavotti A (2018) The DNA binding landscape of the maize AUXIN RESPONSE FACTOR family. Nat Commun **9**: 1–14

Gismondi M, Daurelio LD, Maiorano C, Monti LL, Lara M V., Drincovich MF, Bustamante CA (2020) Generation of fruit postharvest gene datasets and a novel motif analysis tool for functional

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

studies: uncovering links between peach fruit heat treatment and cold storage responses. Planta **251**: 1–18

Gogorcena Y, Sánchez G, Moreno-vázquez S, Pérez S, Ksouri N (2020) Genomic-based breeding for climate-smart peach varieties. *In* C Kole, ed, Genome Designing of Climate Fruit Crops. Springer-Nature, Cham Switzerland, pp 271–331

Guo J, Chen J, Yang J, Yu Y, Yang Y, Wang W (2018) Identification, characterization and expression analysis of the VQ motif-containing gene family in tea plant (Camellia sinensis). BMC Genomics **19**: 1–12

Helden J v, André B, Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. J Mol Biol **281**: 827–842

van Helden J, Rios AF, Collado-Vides J (2000) Discovering regulatory elements in non-coding sequences by analysis of spaced dyads. Nucleic Acids Res **28**: 1808–1818

Wang K, Yin XR, Zhang B, Grierson D, Xu CJ, Chen KS (2017) Transcriptomic and metabolic analyses provide new insights into chilling injury in peach fruit. Plant Cell Environ **40**: 1531–1551

Howe KL, Contreras-Moreira B, Silva N De, Maslen G, Akanni W, Allen J, Alvarez-Jarreta J, Barba M, Bolser DM, Cambell L, et al. (2020) Ensembl Genomes 2020 enabling non-vertebrate genomic research. Nucleic Acids Res **48**: D689–D695

Jiao Y, Shen Z, Yan J (2017) Transcriptome analysis of peach [Prunus persica (L.) Batsch] stigma in response to low-temperature stress with digital gene expression profiling. J Plant Biochem Biotechnol **26**: 141–148

Korkuc P, Schippers JHM, Walther D (2014) Characterization and identification of *cis*-regulatory elements in Arabidopsis based on single-nucleotide polymorphism information. Plant Physiol **164**: 181–200

Koschmann J, Machens F, Becker M, Niemeyer J, Schulze J, Bulow L, Stahl DJ, Hehl R (2012) Integration of bioinformatics and synthetic promoters leads to the discovery of novel elicitor-responsive *cis*-regulatory sequences in Arabidopsis. Plant Physiol **160**: 178–191

Kristiansson E, Thorsen M, Tamás MJ, Nerman O (2009) Evolutionary forces act on promoter length: Identification of enriched *cis*-regulatory elements. Mol Biol Evol **26**: 1299–1307

Ksouri N, Jiménez S, Wells CE, Contreras-Moreira B, Gogorcena Y (2016) Transcriptional responses in root and leaf of *Prunus persica* under drought stress using RNA sequencing. Front Plant Sci **7**: 1–19

Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics **9**: 1–13

Lescot M, Déhais P, Thijs G, Marchal K, Moreau Y, Van De Peer Y, Rouzé P, Rombauts S (2002) PlantCARE, a database of plant *cis*-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences. Nucleic Acids Res **30**: 325–327

Li E, Liu H, Huang L, Zhang X, Dong X, Song W, Zhao H, Lai J (2019a) Long-range interactions between proximal and distal regulatory regions in maize. Nat Commun **10**: 1–14

Li J, Han G, Sun C, Sui N (2019b) Research advances of MYB transcription factors in plant stress resistance and breeding. Plant Signal Behav **14**: e1613131–9

Li J, Zhou D, Qiu W, Shi Y, Yang JJ, Chen S, Wang Q, Pan H (2018) Application of weighted gene co-expression network analysis for data from paired design. Sci Rep **8**: 1–8

Li X, Jiang J, Zhang L, Yu Y, Ye Z, Wang X, Zhou J, Chai M, Zhang H, Arús P, et al. (2015) Identification of volatile and softening-related genes using digital gene expression profiles in melting peach. Tree Genet Genomes **11**: 1–15

Liseron-Monfils C, Lewis T, Ashlock D, McNicholas PD, Fauteux F, Strömvik M, Raizada MN (2013) Promzea: a pipeline for discovery of co-regulatory motifs in maize and other plant species and its application to the anthocyanin and phlobaphene biosynthetic pathways and the maize development atlas. BMC Plant Biol **13**: 1–17

Ma S, Shah S, Bohnert HJ, Snyder M, Dinesh-Kumar SP (2013) Incorporating motif analysis into gene co-expression networks reveals novel modular expression pattern and new signaling pathways. PLoS Genet **9**: 1–20

Montardit Tardá F (2018) Genomic delimitation of proximal promoter regions: Three approaches in Prunus persica http://agris.fao.org/agris718 search/search.do?recordID=QC2019600125

Nguyen NTT, Contreras-Moreira B, Castro-Mondragon JA, Santana-Garcia W, Ossio R, Robles-Espinoza CD, Bahin M, Collombet S, Vincens P, Thieffry D, et al. (2018) RSAT 2018: Regulatory sequence analysis tools 20th anniversary. Nucleic Acids Res **46**: 209–214

Petrillo E, Godoy Herz MA, Barta A, Kalyna M, Kornblihtt AR (2014) Let there be light: regulation of gene expression in plants. RNA Biol **11**: 1215–1220

Pimentel H, Bray NL, Puente S, Melsted P, Pachter L (2017) Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods **14**:1–6

Reiter F, Wienerroither S, Stark A (2017) Combinatorial function of transcription factors and cofactors. Curr Opin Genet Dev **43**: 73–81

Sanhueza D, Vizoso P, Balic I, Campos-Vargas R, Meneses C (2015) Transcriptomic analysis of fruit stored under cold conditions using controlled atmosphere in *Prunus persica* cv "Red Pearl" Front Plant Sci **6**: 1–12

Sebastian A, Contreras-Moreira B (2014) FootprintDB: a database of transcription factors with annotated *cis*-elements and binding interfaces. Bioinformatics **30**: 258–265

Smale ST (2001) Core promoters: active contributors to combinatorial gene regulation. Genes Dev **15**: 2503–2508

Steffens NO, Galuschka C, Schindler M, Bülow L, Hehl R (2005) AthaMap web tools for database-assisted identification of combinatorial *cis*-regulatory elements and the display of highly conserved transcription factor binding sites in *Arabidopsis thaliana*. Nucleic Acids Res **33**: 397–402

Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J (2012) A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nat Protoc **7**: 1551–1568

Tian F, Yang D-C, Meng Y-Q, Jin J, Gao G (2019) PlantRegMap: charting functional regulatory maps in plants. Nucleic Acids Res **1**: 1–10

Tonnessen BW, Bossa-Castro AM, Mauleon R, Alexandrov N, Leach JE (2019) Shared *cis*-regulatory architecture identified across defense response genes is associated with broad-spectrum quantitative resistance in rice. Sci Rep **9**: 1–13

Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T, Dettori MT, Grimwood J, Cattonaro F, et al. (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. Nat Genet **45**: 487–494

Verde I, Jenkins J, Dondini L, Micali S, Pagliarani G, Vendramin E, Paris R, Aramini V, Gazza L, Rossini L, et al. (2017) The Peach v2.0 release: high-resolution linkage mapping and deep resequencing improve chromosome-scale assembly and contiguity. BMC Genomics **18**: 1–18

Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, et al. (2014) Determination and inference of eukaryotic transcription factor sequence specificity. Cell **158**: 1431–1443

Yu C-P, Chen SC-C, Chang Y-M, Liu W-Y, Lin H-H, Lin J-J, Chen HJ, Lu Y-J, Wu Y-H, Lu M-YJ, et al. (2015) Transcriptome dynamics of developing maize leaves and genomewide prediction of

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021

cis-elements and their cognate transcription factors. Proc Natl Acad Sci **112**: 2477–2486

Yu CP, Lin JJ, Li WH (2016) Positional distribution of transcription factor binding sites in *Arabidopsis thaliana*. Sci Rep **6**: 1–7

Zhu Qun, Dabi T, Lamb C (1995) TATA box and initiator functions in the accurate transcription of a plant minimal promoter in vitro. Plant Cell **7**: 1681–1689

Zolotarov Y, Strömvik M (2015) *De novo* regulatory motif discovery identifies significant motifs in promoters of five classes of plant dehydrin genes. PLoS One **10**: 1–19

Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, Shiu S-H (2011) *Cis*-regulatory code of stress-responsive transcription in *Arabidopsis thaliana*. Proc Natl Acad Sci **108**: 14992–14977

Downloaded from https://academic.oup.com/plphys/article/185/3/1242/6094791 by guest on 31 May 2021