

 Open access • Posted Content • DOI:10.1101/2020.02.28.970137

## Motif analysis in co-expression networks reveals regulatory elements in plants: The peach as a model — [Source link](#)

Najla Ksouri, Jaime A. Castro-Mondragon, F. Montardit-Tardà, van Helden J ...+2 more authors

**Institutions:** University of Oslo, Aix-Marseille University, European Bioinformatics Institute

**Published on:** 04 Mar 2020 - bioRxiv (Cold Spring Harbor Laboratory)

**Topics:** WRKY protein domain

Related papers:

- [Incorporating Motif Analysis into Gene Co-expression Networks Reveals Novel Modular Expression Pattern and New Signaling Pathways](#)
- [Cis-motifs upstream of the transcription and translation initiation sites are effectively revealed by their positional disequilibrium in eukaryote genomes using frequency distribution curves](#)
- [PRECISE: Software for Prediction of cis-Acting Regulatory Elements](#)
- [A Gibbs sampling method to detect overrepresented motifs in the upstream regions of coexpressed genes.](#)
- [RhizoBindingSites, a Database of DNA-Binding Motifs in Nitrogen-Fixing Bacteria Inferred Using a Footprint Discovery Approach.](#)

Share this paper:    

View more about this paper here: <https://typeset.io/papers/motif-analysis-in-co-expression-networks-reveals-regulatory-7fli0sqqxf>

1 **Motif analysis in co-expression networks reveals regulatory elements**  
2 **in plants:**

3 **The peach as a model**

4

5 **Running title: *In silico* prediction of peach regulatory motifs**

6 Najla Ksouri<sup>1</sup>, Jaime A. Castro-Mondragón<sup>2,3</sup>, Francesc Montardit-Tardà<sup>1</sup>, Jacques  
7 van Helden<sup>2</sup>, Bruno Contreras-Moreira<sup>4,5,6\*</sup> and Yolanda Gogorcena<sup>1\*</sup>

8 <sup>1</sup>Laboratory of Genomics, Genetics and Breeding of Fruits and Grapevine, Estación  
9 Experimental de Aula Dei-Consejo Superior de Investigaciones Científicas, 50059  
10 Zaragoza, Spain.

11 <sup>2</sup>Aix-Marseille Univ, INSERM UMR\_S 1090, Theory and Approaches of Genome  
12 Complexity (TAGC), F-13288 Marseille, France.

13 <sup>3</sup>Current address: Centre for Molecular Medicine Norway (NCMM), Nordic EMBL  
14 Partnership, University of Oslo, 0318 Oslo, Norway.

15 <sup>4</sup>Laboratory of Computational and Structural Biology, Department of Genetics and  
16 Plant Production, Estación Experimental de Aula Dei-Consejo Superior de  
17 Investigaciones Científicas, 50059 Zaragoza, Spain.

18 <sup>5</sup>Fundación ARAID, Zaragoza, Spain

19 <sup>6</sup>Current address: European Molecular Biology Laboratory, European Bioinformatics  
20 Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SD, UK

21 \*Senior authors

22 **One sentence summary**

23 Motifs prediction depends on the promoter size. A proximal promoter region defined as  
24 an interval of -500 bp to +200 bp seems to be the adequate stretch to predict *de novo*  
25 regulatory motifs in peach

26 **Footnotes:**

27 **Authors' contribution**

28 NK, BC-M and YG devised the study objectives, designed the experiment, discussed  
29 data and wrote the manuscript. NK performed the bioinformatics analysis, FM-T  
30 contributed to delimit the proximal promote region. JA-CM aided to prepare the figures  
31 and provided critical feedback. JvH contributed in the critical discussion of results. YG

32 and BC-M contributed the analysis tools and YG conceived the experiment and  
33 supervised the activities. All authors read and approve the manuscript.

#### 34 **Fundings**

35 This work was partly funded by the Spanish Ministry of Economy and Competitiveness  
36 grants AGL2014-52063R, AGL2017-83358-R (MCIU/AEI/FEDER/UE); and the  
37 Government of Aragón with grants A44 and A09\_17R, which were co-financed with  
38 FEDER funds. N. Ksouri was hired by project AGL2014-52063R and now funded by a  
39 PhD fellowship awarded by the Government of Aragón.

#### 40 **Abstract**

41 Identification of functional regulatory elements encoded in plant genomes is a  
42 fundamental need to understand gene regulation. While much attention has been given  
43 to model species as *Arabidopsis thaliana*, little is known about regulatory motifs in  
44 other plant genera. Here, we describe an accurate bottom-up approach using the online  
45 workbench RSAT::Plants for a versatile ab-initio motif discovery taking *Prunus persica*  
46 as a model. These predictions rely on the construction of a co-expression network to  
47 generate modules with similar expression trends and assess the effect of increasing  
48 upstream region length on the sensitivity of motif discovery. Applying two discovery  
49 algorithms, 18 out of 45 modules were found to be enriched in motifs typical of well-  
50 known transcription factor families (bHLH, bZip, BZR, CAMTA, DOF, E2FE, AP2-  
51 ERF, Myb-like, NAC, TCP, WRKY) and a novel motif. Our results indicate that small  
52 number of input sequences and short promoter length are preferential to minimize the  
53 amount of uninformative signals in peach. The spatial distribution of TF binding sites  
54 revealed an unbalanced distribution where motifs tend to lie around the transcriptional  
55 start site region. The reliability of this approach was also benchmarked in *Arabidopsis*  
56 *thaliana*, where it recovered the expected motifs from promoters of genes containing  
57 ChIPseq peaks. Overall, this paper presents a glimpse of the peach regulatory  
58 components at genome scale and provides a general protocol that can be applied to  
59 many other species. Additionally, a RSAT Docker container was released to facilitate  
60 similar analyses on other species or to reproduce our results.

61 **Keywords:** Motif prediction, cis-regulatory elements, *Prunus persica*, Transcription  
62 Factor binding motifs

## 63 1. Introduction

64 Peach [*Prunus persica* (L.) Batsch], a member of *Prunus* genus, is one of the best  
65 genetically characterized species within the Rosaceae family. With a small size diploid  
66 genome ( $2n = 2x = 16$ ; 230 Mbp), and relatively short generation time (2-3 years), peach  
67 has become a model species for fruit genetic studies (Abbott et al., 2002). Obtaining  
68 elite genotypes with broad environmental adaptations and good fruit quality are the  
69 fundamental targets of all *Prunus* breeding programs, since they directly affect the  
70 economical relevance of this crop (Gogorcena et al., 2020). Indeed, previous works  
71 have reported strong affinity between environmental cues and the fruit quality and  
72 aroma (Wong et al., 2016; Tanou et al., 2017). To stand the environmental stimuli and  
73 ensure edible fruit development, a complex re-arrangement of the gene expression  
74 network is required.

75 The modulation of gene expression is a complex process occurring at various levels  
76 from which the transcriptional regulation is the core control code (Petrillo et al., 2014).  
77 The transcription machinery is regulated by an interplay between DNA-binding proteins  
78 called transcription factors (TFs) and cis-regulatory elements (CREs). TFs bind short  
79 sequences known as TF binding sites (TFBS) or motifs located at CREs (e.g.,  
80 promoters, enhancers, silencers). TFs may act as either activators or repressors of gene  
81 expression, leading to dynamic changes of the cellular pathways. For peach, annotation  
82 of TFs is available in the plant transcription factor database (plantTFDB) (Tian et al.,  
83 2019).

84 As of February 2020, plantTFDB v5.0 stores 2780 peach TFs classified into 58 families  
85 (<http://planttfdb.cbi.pku.edu.cn/>). While much is known about TF families, TF-binding  
86 motifs remain elusive. Deciphering the cis-regulatory network has become a  
87 prerequisite toward scoping out the foundations of transcriptional regulation in *P.*  
88 *persica*. The computational exploration of these DNA motifs has been greatly  
89 stimulated by the availability of genomic data and the release of whole genome  
90 sequence assemblies (Verde et al., 2013; Verde et al., 2017). In this context, a variety of  
91 plant motif finders has emerged. Notwithstanding their value, they are hampered by  
92 certain limitations such as, a restricted range of species, Promzea for maize (Liseron-  
93 Monfils et al., 2013), and AthaMap for *Arabidopsis* (Steffens et al., 2005), and limited  
94 analysis capabilities around experimentally defined motifs as PlantCare, (Rombauts et

95 al., 1999) or PlantPAN, (Chang et al., 2008). Thereby, to circumvent these pitfalls, we  
96 have adopted a plant-customized tool for *de novo* motifs discovery, RSAT::Plants  
97 (<http://rsat.eead.csic.es/plants/>). RSAT has both a friendly user interface and command-  
98 line tools for versatile analyses in a wide collection of plants (Nguyen et al., 2018).  
99 Since the analysis of proximal promoter regions is easier in small genomes with short  
100 intergenic regions, most of cis-regulatory motif predictions so far have been conducted  
101 in *Arabidopsis thaliana* (Ma et al., 2012; Korkuc et al., 2014; Cherenkov et al., 2018).

102 In *P. persica* there are only two examples of regulatory motif discovery, in particular  
103 on a set of 350 dehydrin promoter sequences (Zolotarov and Strömviik, 2015) and 30  
104 heat responsive genes (Gismondi et al., 2020). In contrast to these case studies, we  
105 propose a structured bottom-up framework to identify statistically over-represented  
106 motifs on a genome scale. Our probabilistic approach relies on the hypothesis that genes  
107 within co-expressed modules are likely co-regulated by the same TFs. This approach  
108 has been successfully tested in other species, for example in *Arabidopsis thaliana*  
109 (Koschmann et al., 2012; Ma et al., 2013) and maize (Yu et al., 2015). According to  
110 Bianchi et al., 2015, an arbitrary defined segment of 1500 bp upstream of the  
111 transcription start site (TSS) can be considered as the proximal promoter in peach.  
112 However, recent studies about the genomic delimitation of proximal promoters in  
113 *Prunus persica* effectively reduced this region to a window of approximately 500 nt  
114 (Montardit-Tardà, 2018).

115 The proposed approach relies on three fundamentals, i) an accurate definition of co-  
116 expressed gene modules, ii) an assessment of the effect of upstream region length  
117 regarding the effectiveness of motif discovery and, finally iii) disclosing the effect of  
118 splitting the analysis around the TSS site in discovering potential cis-elements. All  
119 together, we demonstrate the utility of our strategy in analyzing genome wide data to  
120 provide insights on gene regulation dynamics across tissues and specific conditions. To  
121 the best of our knowledge, no work has been reported on cis-elements present in *P.*  
122 *persica* on genome wide level, hence the originality of our survey. Additionally, the  
123 predicted motifs from this study can be browsed at ([https://eead-csic-](https://eead-csic-compbio.github.io/coexpression_motif_discovery/peach/)  
124 [compbio.github.io/coexpression\\_motif\\_discovery/peach/](https://eead-csic-compbio.github.io/coexpression_motif_discovery/peach/)), where we provide readers  
125 with direct links to the results, source code and a Docker container to reproduce the  
126 analysis on any other plant species.

## 127 2. Results

### 128 2.1 Identification of differentially expressed transcripts and construction of 129 weighted co-expression network

130 After quality assessment and pseudo-alignment, an expression matrix was generated  
131 from eight peach published transcriptomes, including treated and control samples with  
132 their corresponding biological replicates. Differential analysis yielded 11,335 altered  
133 transcripts using  $Q$ -value  $< 0.01$  and  $|\beta| > 1$  thresholds. The number of differentially  
134 expressed transcripts (DETs) identified in each RNA-seq experiment is listed in **Table**  
135 **1**. Detailed information about quality control, pseudo-alignment and differential  
136 expression analyses is shown in **Table S1**. An overview of our workflow is provided in  
137 **Figure 1**.

138 The WGCNA R-package was adopted to construct an unsigned co-expression network  
139 for 11,335 stress-related transcripts. All samples and DETs were considered in the  
140 network construction, as neither outliers nor transcripts with missing values, were  
141 detected (**Figure S1. A**). Using a dynamic tree cut algorithm, 45 co-expression modules  
142 were retained with size ranging from 29 to 1795 transcripts per module (**Figure S1. C**).  
143 The 45 distinct modules (labeled with different colors) are shown in a dendrogram in  
144 which major tree branches constitute modules and leaves correspond to DETs (**Figure**  
145 **S1. B**).

### 146 2.2 Transcription factor binding site (TFBS) prediction

#### 147 2.2.1 Effect of proximal promoter length on prediction accuracy

148 As a first step towards extracting regulatory signatures, upstream region boundaries  
149 were defined from -1500 bp to +200 bp relative to TSS (Up 1). Six out of 45 modules  
150 were found to display positive signals and higher significance when compared to the  
151 random clusters. Upstream regions of modules (M9, M10, M11, M18, M21 and M41)  
152 matched known core DNA-binding elements corresponding to Myb-like, BZR,  
153 CAMTA, bZip, E2FE, and TF families. Modules with their corresponding regulatory  
154 elements are represented in **Figure 2** and further information is provided in **Table S2**.  
155 Motifs resulting from both oligo and dyad analysis correspond to signatures with strong  
156 confidence estimation. Besides, eight poly (AT)-rich signals were discarded from M1,  
157 M2, M3, M4 and M6 due to their low complexity. Curiously, these (AT) patterns were  
158 also detected in the random clusters and their occurrence seemed to be associated with

159 the size of the module (**Table S3**). For instance, M1 is the largest module with 1795  
160 sequences and (AT)-repetitive signals were detected in 40 out of the corresponding 50  
161 random clusters.

162 Furthermore, when we restricted the motif discovery to the region with [-500 bp, +200  
163 bp] boundaries (Up 2), fifteen modules were found to discern statistically significant  
164 motifs. These were then grouped into 10 TF families as illustrated in **Figure 2** (TCP,  
165 bHLH, BZR, bZip, NAC, WRKY, AP2-ERF, Myb-like, CAMTA and E2FE).

166 An in-depth look at the major changes occurring when trimming the upstream segments  
167 to 500 bp resulted in interesting observations, summarized as follows. Spurious (AT)  
168 rich events considered as low quality predictions were limited to M2 and were replaced  
169 by relevant regulatory elements in M1, M3, M4 and M6 (**Table S2**). Significant signals  
170 buried in the long upstream region (Up 1) were inferred in modules M24, M28 and M43  
171 (**Figure 2, Table S2**). Besides, shortening the upstream promoter region size to 500 bp  
172 enhances the statistical relevance of the predicted motifs, compared to the negative  
173 controls, regardless of the algorithm applied.

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

### 181 **2.2.2 Effect of splitting the promoter region around the TSS on motif** 182 **prediction**

183 Next, due to the difference in nucleotide composition in coding and non-coding regions,  
184 we subdivided the proximal promoter region in two segments around the TSS, with  
185 each interval examined separately: upstream, from -500 bp to 0 bp (Up 3), and  
186 downstream, from 0 to +200 bp (Up 4). Doing so, motifs of two additional TF families  
187 were discovered, BCP in module M1, DOF in modules M7, M9 and M21. In contrast to  
188 BCP sites lying downstream the TSS (Up 4), DOF sites were found across both

189 intervals (see **Figure 2, Table S2**). Intriguingly, an uncharacterized motif was over-  
190 represented in upstream 4 of module M25 requiring further research.

191 In conclusion, a total of 77 TF binding motifs were revealed from the different assessed  
192 promoter regions (**Table S2**). Modules with candidate predicted motifs might be  
193 classified in two types depending on their potentially matching TF. Indeed, across the  
194 four examined upstream tracts, we recognize those with motifs bound by a single TF  
195 family, considered as single TF-driven modules (e.g., M6, M11, M18, M28 and M41).  
196 Conversely, modules having multiple TFBS for several distinct TFs suggest a possible  
197 combinatorial regulation under particular circumstances. However, more evidence is  
198 needed to address this issue. On the other hand, we observed that the majority of cis-  
199 regulatory elements yielded in this study were mainly detected in the upstream region  
200 Up 2, defined from -500 bp to + 200 bp (**Figure 2, Table S2**).

### 201 **2.3 Gene Ontology enrichment**

202 A Gene Ontology analysis was conducted to annotate the potential function of the  
203 gene modules. Thirteen modules were significantly enriched with biological processes  
204 (**Figure 4**). Six GO terminologies were particularly intriguing and will be briefly  
205 described. In modules M1 and M18, transcripts were over-represented respectively in  
206 leaf and root tissues under drought experiment which is in line with the  
207 “photosynthesis” and “response to water” enrichment. Similarly, module M2 was  
208 enriched for “response to stimulus” with high TPM values in fruit tissue at different  
209 ripening stage. Transcripts within M5 were mostly abundant in fruit tissue under cold  
210 stress, in line with the “cold acclimation” enrichment. Not surprisingly, “response to  
211 stress” was over-represented in fruit in module M10 as we are dealing with stress  
212 conditions. Finally, hormonal levels are known to imbalance under stress explaining the  
213 enrichment of “response to hormone stimulus” in M21. Overall, we consider that the  
214 GO enrichment results (**Figure 4.A**) are in harmony with the expression profiles of  
215 transcripts in **Figure 4.B**.

### 216 **2.4 TFs annotation and prediction of their TFBS using footprintDB**

217 The predicted modules were examined for genes encoding TFs. In total 39 annotated  
218 TFs were shortlisted in **Figure 5**. Myb and Myb-like TFs were exclusively expressed in  
219 modules M1 and M2. They were particularly over-represented in fruit and leaf tissues in  
220 agreement with their transcript profiling illustrated in **Figure 4.B**. We hypothesize that



221 Myb factors may act as regulators of drought stress and ripening in peach. In the same  
222 vein, bHLH genes identified in M3 were notably abundant in stigma tissue, which is in  
223 accordance with **Figure 4.B**. NAC and E2FE transcription factors were respectively  
224 annotated in M4 and M41, and their coding genes were repressed among experiments in  
225 all tissues. The WRKY TFs assigned to module M6 were abundant under hyper  
226 hydricity fitting with **Figure 4.B** and suggesting a regulatory function of the WRKY in  
227 such a condition. Module M7 was associated with genes encoding three TFs with  
228 different expression profiles (DOF, bHLH and ERF). Calmodulin binding proteins  
229 identified in M11 and bZip annotated in M18 and M21 were highly abundant among all  
230 experiments indicating that they may be involved in multiple biological processes.

231 Subsequently, we verified whether the disclosed motifs in each module are the actual  
232 binding sites of the aforementioned TFs (**Figure 5**). TFs were individually examined for  
233 their potential DNA-site using footprintDB and results were compared to those derived  
234 from RSAT. Consensus sequences predicted from genes coding TFs showed high  
235 similarity to consensus sequences predicted from modules (**Table 2**). As for instance,  
236 the binding motif “tTTGGCGGGAAA” identified in module M41 is almost identical to  
237 E2FE-predicted site “TTTTGGCGGGAAAA” from the same module. This suggests  
238 that E2FE may modulate gene expression in M41 and “tTTGGCGGGAAA” motif  
239 could be the *bona fide* binding site of this transcription factor.

## 240 **2.5 Motif scanning**

241 To identify the position of transcription factor binding sites (TFBS) in the promoter  
242 region of *P. persica* genes, position-specific scoring matrices (PSSMs) of all candidate  
243 motifs (77) were *in silico* scanned to the long (Up 1) upstream stretch [-1500, +200 bp].  
244 We observed a clear positional bias of the TFBS close to the TSS, more precisely within  
245 the interval [-500 bp, +200 bp], then it progressively declines towards the 5' limit  
246 (**Figure 6**). For motifs detected respectively in Up 1 (yellow color), Up 2 (green) and  
247 Up 3 (blue), sites were notably concentrated upstream the TSS showing a bell-shaped  
248 distribution from -500 bp to +0 bp with a maximum of density around -250 bp.  
249 Conversely, the positional distribution of motifs predicted along the upstream 4 was  
250 biased toward downstream the TSS with the flatter peak reaching its limit at the TSS  
251 (Up 4, purple). Detailed scanning results can be accessed at [https://ead-csic-](https://ead-csic-compbio.github.io/coexpression_motif_discovery/peach)  
252 [compbio.github.io/coexpression\\_motif\\_discovery/peach](https://ead-csic-compbio.github.io/coexpression_motif_discovery/peach). On the other side, (AT)

253 repetitive elements were also scanned to check their relevance, e.g., whether they  
254 correspond to the TATA box. The underlying data included in **Figure S<sub>2</sub>**, showed that  
255 TFBSs of these motifs were remarkably distant to the TSS and were distributed across  
256 the whole proximal region.

## 257 **2.6 Validation of the protocol for *de novo* cis-element discovery**

258 To demonstrate the performance of the motif finding approach, we evaluated the effect  
259 of variable proximal promoter lengths on uncovering true DNA-binding sites in  
260 *Arabidopsis thaliana*. Experimentally proven motifs from a selection of *A. thaliana*  
261 transcription factors belonging to different families were successfully recovered by at  
262 least one algorithm. As summarized in **Figure 7**, JASPAR and *de novo* identified motifs  
263 displayed high consensus similarity. Moreover, in order to refine the comparison, we  
264 annotated the newly reported motifs JASPAR to ensure that they correspond to the TF  
265 family in question. As expected, *de novo* motifs shared the same annotation as the  
266 reference JASPAR motifs, which underlines the predictive performance of the proposed  
267 methodology.

## 268 **3. Discussion**

269 In the present study, transcriptional profiling of eight independent data sets was  
270 conducted to decipher the intricate process of gene regulation in peach and to reveal  
271 meaningful biological signatures. DETs were grouped into 45 co-expression modules  
272 undergoing similar changes in their expression patterns. Unlike conventional clustering  
273 methods (such as k-means and hierarchical clustering), which are based on geometric  
274 distances, WGCNA is a graph-based approach relying on network topology as inferred  
275 from the correlation among expression values (Li et al., 2018). In our hands, the  
276 WGCNA algorithm robustly and accurately defined modules within a complex multi-  
277 condition dataset.

278 Discerning regulatory signals from blocks of co-expressed genes is a common  
279 presumption used to identify functional genomic elements. It has been successfully  
280 applied and approved in various plants species like *Arabidopsis thaliana* (Koschmann et  
281 al., 2012; Ma et al., 2013), *Zea mays* (Yu et al., 2015) and *Hordeum vulgare L.*  
282 (Cantalapiedra et al., 2017). However, little is known about its applicability to woody  
283 species. To our knowledge, this article is the first in which this hypothesis has been  
284 tested in *Prunus persica* genome wide.

285 For each predicted module, two-motif discovery algorithms (oligo and dyad analysis)  
286 were ran to discover significant motifs in the upstream promoter region. As suggested  
287 by Bianchi and colleagues, we initially defined the upstream promoter size as an  
288 interval of [-1500 bp to +200 bp] relative to the TSS (Bianchi et al., 2015). Discovered  
289 motifs with significant poly-(AT) sites were discarded due to their low complexity and  
290 scarcity of information concerning their specific-regulatory function. We reasoned that  
291 low complexity sequences might be linked to repetitive stretches of DNA, extensively  
292 present in plant genomes (Yu et al., 2015). Interestingly, when tuning the promoter  
293 upstream length to a tract of [-500 bp, +200 bp] relative to the TSS, these low  
294 complexity motifs were limited to module M2. It would seem that long upstream  
295 promoter regions unbalance the signal-to-noise ratio exacerbating the identification of  
296 such AT motifs. Along the same lines, we observe a dependence of (AT)-rich sites on  
297 the dataset size. Indeed, AT-low-complexity motifs were only detected in the first six  
298 modules, which contain from 560 to 1795 upstream sequences. In light of these  
299 considerations, we believe that in our study case, they may result in part due to the  
300 properties of DNA sequences (both upstream region length and dataset size) rather than  
301 the performance of the chosen algorithms. In **Table S3**, the results revealed that AT-rich  
302 occurrence in random cluster increases in parallel with the module size.

303 To check whether the AT-rich patterns overlap the TATA boxes, a position scanning  
304 experiment was conducted. It is well documented in plants that a TATA box region lays  
305 between -30 and +35 bp with respect to the TSS (Zhu Qun et al., 1995; Smale, 2001)  
306 However, the scanning results portrayed that peaks were located far from this interval,  
307 confirming that they are distinct signals (**Figure S2**).

308 By limiting the promoter length to a window of -500 bp, new regulatory motifs were  
309 recovered. Additionally, splitting the proximal promoter region into two intervals  
310 around the TSS enabled the discovery of further hidden candidate TF motifs. Such  
311 observations may strengthen our hypothesis that shorter upstream regions improve the  
312 sensitivity motif discovery (from 11 motif sequences identified within Up 1 to 58  
313 sequences identified in Up 3 and Up 4 assessed separately). Defining the upstream  
314 promoter length has been a controversial issue (Kristiansson et al., 2009). If the interval  
315 is too short or too long, the motif of interest may not be captured. Therefore, we reason  
316 that an analysis on regions of variable length would yield a more comprehensive picture  
317 of the complex regulatory code.

318 The spatial distribution of the occurrences of the 77 inferred motifs along the promoter  
319 region is crucial to understand gene regulation in *Prunus persica*. Our findings revealed  
320 that TFBSs are not uniformly dispersed across the promoter but they exhibit a strikingly  
321 mixture of 2 density profiles: while the majority showed bell-shaped distribution at the  
322 interval of [-500 bp, 0 bp], others were diverged downstream the TSS [0 bp, +200 bp]  
323 (**Figure 6**). These findings are similar to those described in *A. thaliana*, with nearly two  
324 thirds of the examined TFBSs within the region from -400 bp to +200 bp (Yu et al.,  
325 2016). TFBSs of bHLH, BZR, TCP and WRKY are particularly concentrated from -500  
326 bp to 0 bp. This denotes a positional binding preference within this proximal region,  
327 which is in agreement with (Yu et al., 2016) reporting that their positional preference is  
328 between -100 bp to -50 bp. On the other hand, bZip, CAMTA, E2FE and Myb-like  
329 exhibited a dual binding distribution with central peaks upstream and downstream the  
330 TSS. A possible explanation of this is that some TFs may display different binding  
331 preferences depending on their TF-specific structure, biological functions or  
332 combinatorial with other TFs. The degree to which the arrangement of motif sites is  
333 associated to their function needs to be further investigated especially that data about  
334 TFBS distribution in plants is only limited to *Arabidopsis thaliana* (Zou et al., 2011;  
335 Yu et al., 2016). According to our findings, we may consider that the boundary from -  
336 500 bp to 0 bp is an adequate region to look for the majority of TFBSs lying in the  
337 proximal promoter region in peach. However, we should keep in mind that proximal  
338 TFBSs could also occur downstream the TSS. Thus, we suggest defining the peach  
339 proximal promoter length as a tract of [-500 bp to +200 bp], analyzing separately the  
340 two regions around the TSS for a better motif coverage. In fact, according to Montardit-  
341 Tardà (2018), differences in the nucleotide composition were found upstream and  
342 downstream the TSS. At this point, we should mention that gene regulation involves a  
343 complex interplay between the proximal (promoter) and distal regulatory regions  
344 located thousands of base pairs away from the TSS (e.g., enhancers) (Li et al., 2019).  
345 Our workflow sheds light mainly on sequence signatures extracted from the proximal  
346 promoter. Thus, it might not be adequate to study distal genomic elements.

347 Furthermore, rather than barely returning a list of significant motifs, our methodology  
348 assigned them to different modules to help shape a clear overview of the peach  
349 regulation code. Overall, we were able to distinguish 18 modules harboring 77 motifs  
350 from 11 TF families: bHLH, bZip, BZR, CAMTA, DOF, E2FE, AP2-ERF, Myb-like,

351 NAC, TCP and WRKY. While some modules, such as M6, M11, M28 and M41, seem to  
352 be driven by a single TF (WRKY, CAMTA and E2FE, respectively), motifs from  
353 different families were annotated in the rest. This can be explained by the fact that some  
354 promoter sequences may encompass multiple TFBSs of perhaps interacting TFs. Indeed,  
355 TFs have been reported to frequently operate in combination (Guo et al., 2018; Kumar  
356 et al., 2018). Combinatorial regulation is required to confer specific responses in a  
357 particular tissue and under a particular stress. Thus, the hypothesis of cooperative  
358 interactions between diverse motifs in peach is worthy to be further investigated.

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

365 The biological significance of modules with significant identified signals was  
366 determined by Gene Ontology analysis and TF annotation. The enriched modules  
367 reflected many biological functions involved in abiotic stress responses such as cold  
368 acclimation, response to stress, response to water and response to hormone (**Figure 4**).  
369 In this context, modules M1 enriched for “photosynthesis” contained candidate Myb  
370 and Myb-related factors. These findings are in line with (Baldoni et al., 2015) reporting  
371 that Myb TF family is known to regulate drought tolerance and the stomatal movements  
372 in plants. bHLH binding sites were mainly disclosed in modules M3, M5 and M7  
373 (**Figure 5**). Associated TFs among those were abundant under various stress conditions  
374 proposing a multi-functional role of bHLH. According to Bianchi et al., (2015), bHLH  
375 factors play a central role in flavonoid biosynthesis and cold acclimation in peach.  
376 Similarly, bZip TFs were found in both M18 and M21 and their transcripts were mainly  
377 over-represented in all experiments. Our results are supported by previous studies  
378 reporting that bZip were induced by various environmental cues. Indeed it was revealed  
379 that they play a pivotal role in responses to cold stress in peach and enhance water use  
380 efficiency in almond-peach rootstocks (Hu et al., 2018). WRKYs putative motifs were  
381 restricted to M6 and were exclusively activated in leaf tissue under hyper-hydricity  
382 (HH) stress. It is well known that HH leads to morphological abnormalities, such as  
383 brittle leaves (Carrillo Bermejo et al., 2017). We speculate that WRKY factors may be

384 implicated in morphological damages produced by HH. Module M11 was found to be a  
385 potential CAMTA-driven module (**Figure 2**), where two genes coding CAMTA were  
386 annotated (**Figure 5**). A previous study in *A. thaliana* demonstrated that cold stress  
387 increases the level of calcium sensed by CAMTA (Doherty et al., 2009). This  
388 perturbation of calcium levels leads to modification of the CAMTA activity that in turn  
389 triggers the induction of cold response genes of the CBF family. For this reason,  
390 CAMTA motifs are of great interest. From the perspective of peach breeding, these  
391 findings may be of great interest, as genes within modules are potential targets for  
392 further experimental validation.

393 Finally, a major drawback of motif discovery approaches is their limited performance.  
394 To tackle this issue we designed a control experiment in which genomic sites detected  
395 by ChIP-seq for 10 *A. thaliana* TFs were analyzed. Comparing the *de novo* predicted  
396 motifs to the corresponding curated motifs in JASPAR we observed a high similarity in  
397 terms of Ncor scores (**Figure 7 and Table S4**). When searching for *in-vivo* validated  
398 motifs, we would ideally expect to get identical predicted motifs. Nonetheless, while  
399 most consensus sequences had high Ncor values  $> 0.8$ , others had lower values. As well,  
400 we observed that the choice of upstream region length affects the performance. In some  
401 cases, particularly Up 1 and Up 3, the expected motif was not even found. Unlike the  
402 results found in peach, examining 4 upstream tracts only returned motifs from the same  
403 query families probably as a consequence of the JASPAR TFBSs profiles being curated.  
404 Taken together, we believe that the proposed workflow is robust enough to be extended  
405 to other species in order to identify reliable regulatory motifs.

#### 406 **4. Conclusion**

407 DNA motif discovery is a primary step for studying gene regulation, however the *in*  
408 *silico* prediction of regulatory motifs is not straightforward. In contrast to previous  
409 surveys that usually assume a fixed promoter length right at the start; this work reports  
410 regulatory elements while testing different upstream sequence intervals. It is among the  
411 first efforts providing a comprehensive collection of *Prunus persica* motifs without a  
412 prior knowledge. By coupling gene expression networks and module analysis, we were  
413 able to extract interpretable information from a large set of noisy data and to reveal  
414 primary candidate TF-target binding sites responding to specific conditions. These  
415 results offer a more complete view of the proximal regulatory signatures in *P. persica*

416 and we believe that it may contribute to address the knowledge gap about the  
417 transcriptional regulatory code in non-model species.

## 418 **5. Materials and methods**

### 419 **5.1 Input data and processing**

420 Eight peach RNA-sequencing datasets were downloaded from the European Nucleotide  
421 Archive (<https://www.ebi.ac.uk/ena>) and were used as raw reads for this project. This  
422 comprehensive dataset includes data of various peach cultivars, from various tissues  
423 (root, leaf, stigma and fruit), different stress conditions and developmental stages. A  
424 detailed list of the project IDs and metadata is provided respectively in **Table 1 and**  
425 **Table S1.A**. The obtained reads were quality-processed and trimmed using FASTQC  
426 v.0.11.5 and Trimmomatic v.0.36 (Bolger et al., 2014), to discard adaptors and low-  
427 quality sequences with mean Phred score ( $Q < 30$ ) and window size of 4:15. The first  
428 nucleotides were then head-cropped to ensure a per-position A, C, G, T frequency near  
429 to 0.25. Following the trimming, only sequences longer than 36 bp were retained for  
430 further analysis. The complete workflow is shown in **Figure 1 (see step 1)**.

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

### 440 **5.2 Transcript-level profiling**

441 Differential expression analysis was conducted with Sleuth R package v.0.29.0  
442 (Pimentel et al., 2017) for each RNA data set separately. The Wald test (WT) was  
443 applied to output abundance files in order to retain the significant expressed transcripts  
444 from each experiment. Samples and their biological replicates from each experiment  
445 were compared with their corresponding control. To reduce the false positives, only  
446 transcripts passing an FDR cutoff  $Q$ -value  $< 0.01$  and beta statistic (approximation of  
447 the Log<sub>2</sub> Fold Change between two tested conditions)  $|\beta| > 1$  were retained. Significant

448 transcripts obtained from each RNA-seq project were merged into a single list with an  
449 assigned mean TPM value for each replicate.

### 450 **5.3 Construction of co-expressed network**

451 Based on the assumption that co-expressed genes may share the same biological  
452 signature, weighted gene co-expression network analysis (WGCNA v.1.61) was  
453 performed to extract clusters of densely interconnected genes named modules  
454 (Langfelder and Horvath, 2008). Samples were firstly clustered to remove outliers and  
455 transcripts with missing entries. A similarity matrix was constructed by performing  
456 pairwise Pearson correlation across all targets. Then an adjacency matrix was built  
457 raising the similarity matrix to a soft power ( $\beta$ ). Here  $\beta$  was set to 7 reaching thus 83%  
458 of the scale free topology fitting index ( $R^2$ ). To minimize the effect of noise, matrix  
459 adjacency was transformed to Topological Overlap Measure (TOM) and its  
460 corresponding dissimilarity matrix (1-TOM) was generated. Finally, modules were  
461 defined using the `cutreeDynamic` function with a minimum module size of 20 targets.  
462 Compared to standard hierarchical clustering, this approach solves the issue of setting  
463 the final number of clusters and arranges the genes based on their topological overlap to  
464 eliminate spurious associations resulting from the correlation matrix.

### 465 **5.4 De novo cis regulatory sequences discovery using RSAT::Plants**

466 Gene modules resulting from network analysis were subjected to an *ab-initio* motif  
467 discovery pipeline using the RSAT::Plants standalone (**Figure 1, step 2**). For each  
468 module, the analysis initiates by generating as negative control 50 random clusters of  
469 the same size for each module as described previously (Contreras-Moreira et al., 2016).  
470 Sequences with four different boundaries around the TSS were retrieved from the genes  
471 in the co-expressed modules, random clusters and *Prunus persica* genome v2. The  
472 upstream sequences were defined as intervals of i) -1.5 kb to +200 bp ii) -500 bp to  
473 +200 bp and iii) two segments around the TSS: -500 bp to 0 and 0 bp to +200 bp. Note  
474 that the 0 to +200 interval corresponds to the 3' UTR region, which is already  
475 downstream. RSAT *peak-motifs* was run under the differential analysis mode, where  
476 module's upstream sequences served as the test set and all upstream sequences from  
477 peach genome were considered as the control set to estimate the background model (a  
478 background model was created for each upstream stretch) (Thomas-Chollier et al.,  
479 2012). Two discovery algorithms were used: i) oligo-analysis, which is based on the



480 over-representation of k-mers in upstream regions, and ii) dyad-analysis, which looks  
481 for over-represented spaced pairs of oligonucleotides (DeFrance et al., 2008). For each  
482 run, up to five motifs were returned per algorithm and were retained to compare their  
483 statistical significance with the 50 random clusters considered as negative control.

484 Candidate motifs were chosen based on their significance (log E-value) compared to  
485 negative control and were subsequently annotated by comparison to the footprintDB  
486 collection of plant curated motifs (<http://floresta.eead.csic.es/footprintdb>) (Sebastian and  
487 Contreras-Moreira, 2014) using the *compare-matrix* tool in RSAT (Nguyen et al., 2018)  
488 requiring a normalized correlation score  $N_{cor} \geq 0.4$ .

489 Finally, selected motifs were scanned along the stretch [-1500 bp, +200 bp] to predict  
490 their corresponding binding site positions, using as background model a Markov chain  
491 of order 1 ( $m=1$ ) and a cutoff  $P$ -value  $\leq 1e^{-4}$ . To ensure the clarity and reproducibility of  
492 this strategy, a repository including the source code, links to the results and a tutorial  
493 explaining how to reproduce a similar analysis on any species is available at  
494 [https://eead-csic-compbio.github.io/coexpression\\_motif\\_discovery/peach](https://eead-csic-compbio.github.io/coexpression_motif_discovery/peach).

## 495 **5.5 Transcription factor prediction and Gene Ontology analysis**

496 Hereafter, the analysis was restricted to modules with significant detected signals.  
497 Firstly, genes coding peach TFs were predicted and classified using the iTAK database  
498 (<http://itak.feilab.net/cgi-bin/itak/index.cgi>, last accessed January 2020). Protein  
499 sequences of TFs were subsequently submitted to footprintDB to predict their  
500 interacting DNA-binding site. To functionally interpret the co-expressed modules, Gene  
501 Ontology (GO) enrichment was conducted on PlantRegMap / PlantTFDB portal v5.0  
502 (<http://planttfdb.gao-lab.org/>, last accessed January, 2020) (Tian et al., 2019).  $P$ -value  
503 of 0.01 was set to retain the significant GO terms.

## 504 **5.6 Validation of the pipeline by detecting *a priori* known motifs in *Arabidopsis*** 505 ***thaliana***

506 To assess the impact of upstream region lengths on the identification of relevant motifs,  
507 we used sets of experimentally validated binding sites of 10 *Arabidopsis thaliana* TF  
508 families. Sequences of the proven sites were downloaded from JASPAR database  
509 (Fornes et al., 2020) and were locally aligned with BLASTN against the *A. thaliana*  
510 TAIR10.42 genome from Ensembl Plants to obtain the closest neighbor genes. The  
511 following parameters were used:  $E$ -value  $\leq 1e^{-5}$ ,  $max\_target\_seqs = 1$ ,  $max\_hsps = 1$

512 query-coverage of 80% and percentage of identity 98%. Upstream sequences of  
513 neighbor genes were obtained with *retrieve-seq* from RSAT::Plants. Similarity between  
514 references (JASPAR) and newly discovered motifs was computed with Ncor score (see  
515 above).

## 516 **Supplemental Data**

517 **Supplemental Table S1. A.** Detailed information about the RNA-seq data used for  
518 differential analysis

519 **Supplemental Table S1. B.** Number of survived and dropped reads after quality  
520 processing and pseudo-aligned reads using kallisto program

521 **Supplemental Table S2.** List of candidate regulatory sites discovered within four  
522 upstream tracts of different lengths. Motifs are represented as IUPAC consensus  
523 sequences. TF match: Transcription factor family of the best match in footprintDB.  
524 Ncor: normalized Pearson correlation varying between 0 and 1.  $Ncor \geq 0.4$  indicates  
525 high confidence annotations. Gray color indicates that no significant motifs were found.

526 **Supplemental Table S3.** List of low complexity motifs considered as false positive  
527 predictions within a boundary from -1500 bp to +200 bp upstream region length. For  
528 each algorithm, sequences are presented both as IUPAC consensus sequences using the  
529 degeneracy code and as sequence logos. Last column indicates the occurrence number  
530 of AT-rich motifs within the 50 random clusters used as negative control. Ps: **W** letter  
531 refers to (A or T) nucleotide and **S** refers to (C or G). Number of sites corresponds to  
532 the occurrence number of a single motif.

533 **Supplemental Table S4.** Similarity of JASPAR motifs (considered as queries) and de  
534 *nov*o predicted dyad motifs in *Arabidopsis thaliana*. Numbers tagged with asterisks  
535 indicate number of peaks recovered by BLASTN (see Methods). The Ncor scores  
536 correspond to JASPAR databases.

537 **Supplemental Figure S1.** Co-expression network analysis. **(A):** Sample clustering to  
538 detect outliers. Sample with the same node color are derived from the same RNA-  
539 experiment. **(B):** Topological overlap measure plot. The different shades of color  
540 signify the strength of the connections between the genes (from white not significantly  
541 correlated to red signifying highly significantly correlated). Modules identified are  
542 colored along both column and row and are boxed. **(C):** Distribution of the module  
543 size.

544 **Supplemental Figure S2.** Positional distribution of AT-rich repetitive motifs along  
545 upstream 1: [-1500 bp, +200 bp].

## 546 Acknowledgements

547 We thank Eric OLO NDELA for his support and help in creating the HTML report and  
548 providing useful feedback. We also thank Claudio Antonio Meneses Araya, Dayan  
549 Sanhueza and Tomás Carrasco for giving us access to the RNA-seq data.

## 550 Tables

551 **Table 1.** Summary of RNA-seq data used as input and the number of differentially  
552 expressed transcripts (DETs) identified in each RNA-seq experiment.

Project ID	References	Experiments	Tissues	Conditions	DETs
PRJNA271307	(Li et al., 2015)	Ripening stage	Fruit	6	2601
PRJNA288567	(Sanhueza et al., 2015)	Cold storage	Fruit	6	6447
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	1118
PRJNA328435	Unpublished	Cold storage	Fruit	2	2963
PRJNA397885	Unpublished	Chilling injury	Fruit	4	2429

553

554 **Table 2.** Similarity comparison between RSAT and footprintDB DNA-binding motif  
 555 predictions. The best predictions in footprintDB were selected in *Arabidopsis thaliana*.  
 556 The TFs grouped in this table are the same labeled with a star in **Figure 5**

Modules	RSAT Consensus	TFs	Gene IDs	FootprintDB Consensus	STAMP E-value
M41	tTTGGCGGGAAA	E2FE	Prupe.5G180000	TTTTGGCGGGAAAA	3e <sup>-138</sup>
M21	GaCACGTGkC	bZip	Prupe.1G455300	ACGTGgc	3e <sup>-20</sup>
M18	tGCCACGTGGC	bZip	Prupe.1G419700	TGACGTGGC	1e <sup>-16</sup>
M18	tGCCACGTGGC	bZip	Prupe.1G434500	CACGTGGC	1e <sup>-127</sup>
M18	tGCCACGTGGC	bZip	Prupe.2G182800	TGCCACGT	8e <sup>-125</sup>
M7	tGCCGACa	AP2-ERF	Prupe.3G157100	TGCCGCC	1e <sup>-49</sup>
M7	tGCCGACa	AP2-ERF	Prupe.7G222700	CCGACA	4e <sup>-47</sup>
M7	CACGTGkCGG	bHLH	Prupe.6G303500	CCACGTGr	2e <sup>-84</sup>
M7	aAAAGTc	DOF	Prupe.6G092600	AAAG	2e <sup>-34</sup>
M6	GaAAAGTCaaa	WRKY	Prupe.4G075400	AAAGTCAA	4e <sup>-63</sup>
M6	GaAAAGTCaaa	WRKY	Prupe.5G106700	aAAAGTCAA	2e <sup>-59</sup>
M4	ttAAGCAAata	NAC	Prupe.1G106100	AAGcAAc	7e <sup>-10</sup>
M4	ttAAGCAAata	NAC	Prupe.7G102000	AAGCAA	9e <sup>-35</sup>
M3	CGaCACGTGtCGGtt	bHLH	Prupe.1G252600	CACGTGA	8e <sup>-15</sup>
M3	CGaCACGTGtCGGtt	bHLH	Prupe.2G190100	CACGTGC	3e <sup>-77</sup>
M3	CGaCACGTGtCGGtt	bHLH	Prupe.6G159200	gCACGTG	5e <sup>-20</sup>
M3	CGaCACGTGtCGGtt	bHLH	Prupe.3G064500	CACGTG	9e <sup>-10</sup>

557

## 558 **Figure Legends**

559 **Figure 1.** Bottom-up framework for *de novo* motif discovery. **Step1:** differential  
 560 expression analysis for transcript detection and extraction of co-expressed modules.  
 561 **Step2:** *de novo* motif detection using the peak-motifs tool from RSAT::Plants. Numbers  
 562 correspond to the different tested upstream tracts, with TSSs anchored on position 0 bp,  
 563 while letters represent tools within peak-motifs. Green and orange boxes label software  
 564 and RSAT tools, respectively.

565 **Figure 2.** Position Specific Scoring Matrix (PSSM) representation of top scored  
 566 discovered motifs per modules, along different upstream lengths. The x-axis

567 corresponds to the four intervals: Up 1: [-1500 bp, +200 bp], Up 2: [-500 bp, -200 bp],  
568 Up 3: [-500 bp, 0 bp] and Up 4 [0 bp, +200 bp]. The y-axis informs about the motif  
569 family revealed per module. Cell colors indicate the statistical significance of the  
570 identified motifs while cell sizes represent the normalized correlation (Ncor). Number  
571 of sites corresponds to the number of sites used to build the PSSM. When motifs from  
572 the same family are identified with both algorithms (oligo and dyad-analysis), or in  
573 different upstream tracts (Up 1, Up 2, Up 3 and Up 4), only the most significant one is  
574 represented in the heatmap. Further details are provided in **Table S3**. An interactive  
575 report with source code is accessible at [https://eead-csic-](https://eead-csic-compbio.github.io/coexpression_motif_discovery/peach/)  
576 [compbio.github.io/coexpression\\_motif\\_discovery/peach/](https://eead-csic-compbio.github.io/coexpression_motif_discovery/peach/)

577 **Figure 3.** Illustrative comparison between predicted motif DEL2 (corresponding to  
578 E2FE transcription factor) within two different upstream promoter lengths: -1500 bp to  
579 +200 bp (**A**) and -500 bp to +200 bp (**B**). The name of the best match among plant  
580 motifs in footprintDB is labeled in red, next to its Ncor (Normalized correlation) value  
581 labeled in blue. The x-axis corresponds to the module of interest (M41) and random  
582 clusters ranked ranked by the most significant motifs. The y-axis corresponds to the  
583 statistical significance  $-\log_{10}(P\text{-value})$ . Number of sites corresponds to the occurrence  
584 number of a single motif. The evidence supporting the putative motifs is Ncor (in blue)  
585 and the significance (black bars) when compared to negative controls (gray bars).

586 **Figure 4.** Functional annotation of relevant gene modules. (**A**): Gene ontology  
587 enrichment. (**B**): Mean transcript abundance profiling in term of transcripts per million  
588 (TPM). The x-axis corresponds the different experimental conditions while the y-axis  
589 indicates the number of differentially transcripts per module. Experiment and tissue  
590 types are highlighted by different colors (see the color key at the bottom of the figure).  
591 Gene profiles along the different conditions are provided at ([https://eead-csic-](https://eead-csic-compbio.github.io/coexpression_motif_discovery/peach/)  
592 [compbio.github.io/coexpression\\_motif\\_discovery/peach](https://eead-csic-compbio.github.io/coexpression_motif_discovery/peach/)). See supplementary **Table S1**  
593 for the abbreviations.

594 **Figure 5.** List of transcription factors within relevant modules. Blue and red squares  
595 indicate transcripts per million while bottom color bars correspond to the tissues types  
596 and different experiments, respectively (See the legend at the right side of the figure).  
597 TFs showing sequence similarity between their footprintDB and RSAT predicted motifs  
598 are labeled with a star.

599 **Figure 6.** Positional distribution of the detected oligo motifs in promoter genes of  
600 *Prunus persica*. Four density distributions were derived from four assessed upstream  
601 regions. Up 1: from -1500 bp to 200 bp, Up 2: from -500 bp to +200 bp, Up 3: from -  
602 500 bp to 0 bp and Up 4 from 0 bp to + 200 bp. The x-axis corresponds to upstream  
603 length in base pairs (bp). The y-axis corresponds to density of captured sites with *P*-  
604 value  $<10 e^{-4}$ . Only oligo motifs are presented here, dyads are provided in the report at  
605 [https://ead-csic-compbio.github.io/coexpression\\_motif\\_discovery/peach](https://ead-csic-compbio.github.io/coexpression_motif_discovery/peach).

606 **Figure 7.** Similarity between JASPAR motifs (considered as queries) and *de*  
607 *novo* predicted oligo motifs found in *Arabidopsis thaliana* along four different upstream  
608 regions. Numbers tagged with a star indicate number of peaks recovered by BLASTN  
609 (see Methods). The Ncor scores correspond to JASPAR databases. Only oligo-analysis  
610 motifs are shown (dyads are available at supplementary Table S4). Upstream 1: [-1500  
611 bp to +200 bp], Upstream 2: [-500 bp to +200 bp], Upstream3: [-500 bp to 0 bp] and  
612 Upstream4: [0 bp to +200 bp]

### 613 **Literature cited**

614 **Abbott AG, Georgi L, Yvergniaux D, Wang Y, Blenda A, Reighard G, Inigo M,**  
615 **Sosinski B** (2002) Peach: The model genome for Rosaceae. *Acta Hort* **575**: 145–  
616 155

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

620 **Baldoni E, Genga A, Cominelli E** (2015) Plant MYB transcription factors: Their role  
621 in drought response mechanisms. *Int J Mol Sci* **16**: 15811–15851

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

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

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

- 628 **Cantalapiedra CP, García-pereira MJ, Gracia MP, Igartua E** (2017) Large  
629 differences in gene expression responses to drought and heat stress between elite  
630 Barley cultivar scarlett and a spanish landrace. *Front Plant Sci* **8**: 1–23
- 631 **Carrillo Bermejo EA, Alamillo MAH, Samuel David GT, Llanes MAK, Enrique C**  
632 **de la S, Manuel RZ, Rodriguez Zapata LC** (2017) Transcriptome, genetic  
633 transformation and micropropagation: Some biotechnology strategies to diminish  
634 water stress caused by climate change in sugarcane. *Plant, Abiotic Stress*  
635 *Responses to Clim. Chang. IntechOpen*, pp 90–108
- 636 **Chang WC, Lee TY, Huang H Da, Huang HY, Pan RL** (2008) PlantPAN: Plant  
637 promoter analysis navigator, for identifying combinatorial cis-regulatory elements  
638 with distance constraint in plant gene groups. *BMC Genomics* **9**: 1–14
- 639 **Cherenkov P, Novikova D, Omelyanchuk N, Levitsky V, Grosse I, Weijers D,**  
640 **Mironova V** (2018) Diversity of cis-regulatory elements associated with auxin  
641 response in *Arabidopsis thaliana*. *J Exp Bot* **69**: 329–339
- 642 **Contreras-Moreira B, Castro-Mondragon JA, Rioualen C, Cantalapiedra CP, Van**  
643 **Helden J** (2016) RSAT::Plants: Motif discovery within clusters of upstream  
644 sequences in plant genomes. *In* R Hehl, ed, *Plant Synth. Promot. Methods Mol.*  
645 *Biol. Humana Press, New York*, pp 279–295
- 646 **Defrance M, Janky R, Sand O, van Helden J** (2008) Using RSAT oligo-analysis and  
647 dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nat Protoc*  
648 **3**: 1589–1603
- 649 **Doherty CJ, Van Buskirk HA, Myers SJ, Thomashow MF** (2009) Roles for  
650 *Arabidopsis* CAMTA transcription factors in cold-regulated gene expression and  
651 freezing tolerance. *Plant Cell* **21**: 972–984
- 652 **Fornes O, Castro-Mondragon JA, Khan A, van der Lee R, Zhang X, Richmond**  
653 **PA, Modi BP, Correard S, Gheorghe M, Baranašić D, et al** (2020) JASPAR  
654 2020: update of the open-access database of transcription factor binding profiles.  
655 *Nucleic Acids Res* **48**: 87–92
- 656 **Franco-Zorrilla JM, López-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R**  
657 (2014) DNA-binding specificities of plant transcription factors and their potential

- 658 to define target genes. *PNAS* **111**: 2367–2372
- 659 **Gismondi M, Daurelio LD, Maiorano C, Monti LL, Lara M V., Drincovich MF,**  
660 **Bustamante CA** (2020) Generation of fruit postharvest gene datasets and a novel  
661 motif analysis tool for functional studies: uncovering links between peach fruit  
662 heat treatment and cold storage responses. *Planta* **251**: 1–18
- 663 **Gogorcena Y, Sánchez G, Moreno-vázquez S, Pérez S, Ksouri N** (2020) Genomic-  
664 based breeding for climate-smart peach varieties. *In* C Kole, ed, *Genome Des.*  
665 *Clim. fruit Crop*. Springer-Nature, pp 291–351
- 666 **Guo J, Chen J, Yang J, Yu Y, Yang Y, Wang W** (2018) Identification,  
667 characterization and expression analysis of the VQ motif-containing gene family in  
668 tea plant (*Camellia sinensis*). *BMC Genomics* **19**: 1–12
- 669 **Howe KL, Contreras-moreira B, Silva N De, Maslen G, Akanni W, Allen J,**  
670 **Alvarez-jarreta J, Barba M, Bolser DM, Cambell L, et al** (2020) Ensembl  
671 Genomes 2020 enabling non-vertebrate genomic research. *Nucleic Acids Res* 1–7
- 672 **Hu P, Li G, Zhao X, Zhao F, Li L, Zhou H** (2018) Transcriptome profiling by RNA-  
673 Seq reveals differentially expressed genes related to fruit development and ripening  
674 characteristics in strawberries (*Fragaria × ananassa*). *Peer J* **6**: 1–25
- 675 **Jiao Y, Shen Z, Yan J** (2017) Transcriptome analysis of peach [*Prunus persica* (L.)  
676 Batsch] stigma in response to low-temperature stress with digital gene expression  
677 profiling. *J Plant Biochem Biotechnol* **26**: 141–148
- 678 **Korkuc P, Schippers JHM, Walther D** (2014) Characterization and identification of  
679 cis-regulatory elements in *Arabidopsis* based on single-nucleotide polymorphism  
680 information. *Plant Physiology* **164**: 181–200
- 681 **Koschmann J, Machens F, Becker M, Niemeyer J, Schulze J, Bulow L, Stahl DJ,**  
682 **Hehl R** (2012) Integration of bioinformatics and synthetic promoters leads to the  
683 discovery of novel elicitor-responsive cis-regulatory sequences in *Arabidopsis*.  
684 *Plant Physiol* **160**: 178–191
- 685 **Kristiansson E, Thorsen M, Tamás MJ, Nerman O** (2009) Evolutionary forces act on  
686 promoter length: Identification of enriched cis-regulatory elements. *Mol Biol Evol*

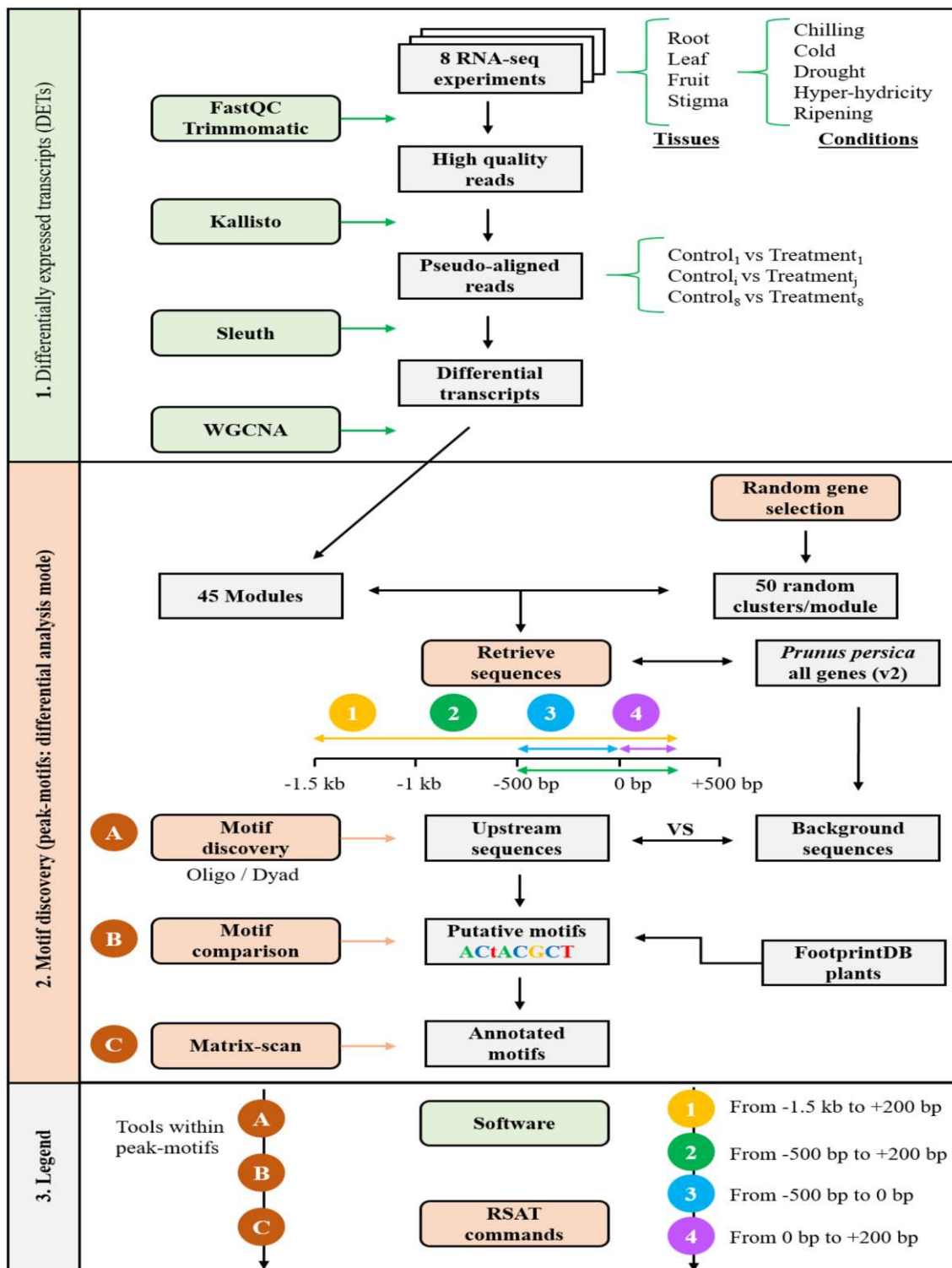


- 687           **26**: 1299–1307
- 688   **Ksouri N, Jiménez S, Wells CE, Contreras-Moreira B, Gogorcena Y** (2016)  
689       Transcriptional responses in root and leaf of *Prunus persica* under drought stress  
690       using RNA sequencing. *Front Plant Sci* **7**: 1–19
- 691   **Kumar N, Dale R, Kemboi D, Zeringue EA, Kato N, Larkin JC** (2018) Functional  
692       analysis of short linear motifs in the plant cyclin-dependent kinase inhibitor  
693       SIAMESE. *Plant Physiol* **177**: 1569–1579
- 694   **Langfelder P, Horvath S** (2008) WGCNA: An R package for weighted correlation  
695       network analysis. *BMC Bioinformatics* **9**: 1–13
- 696   **Li E, Liu H, Huang L, Zhang X, Dong X, Song W, Zhao H, Lai J** (2019) Long-range  
697       interactions between proximal and distal regulatory regions in maize. *Nat Commun*  
698       **10**: 1–14
- 699   **Li J, Zhou D, Qiu W, Shi Y, Yang JJ, Chen S, Wang Q, Pan H** (2018) Application  
700       of weighted gene co-expression network analysis for data from paired design. *Sci*  
701       *Rep* **8**: 1–8
- 702   **Li X, Jiang J, Zhang L, Yu Y, Ye Z, Wang X, Zhou J, Chai M, Zhang H, Arús P, et**  
703       **al** (2015) Identification of volatile and softening-related genes using digital gene  
704       expression profiles in melting peach. *Tree Genet Genomes* **11**: 1–15
- 705   **Liseron-Monfils C, Lewis T, Ashlock D, McNicholas PD, Fauteux F, Strömvik M,**  
706       **Raizada MN** (2013) Promzea: A pipeline for discovery of co-regulatory motifs in  
707       maize and other plant species and its application to the anthocyanin and  
708       phlobaphene biosynthetic pathways and the maize development atlas. *BMC Plant*  
709       *Biol* **13**: 1–17
- 710   **Ma S, Bachan S, Porto M, Bohnert HJ, Snyder M, Dinesh-Kumar SP** (2012)  
711       Discovery of stress responsive DNA regulatory motifs in *Arabidopsis*. *PLoS One*  
712       **7**: 1–13
- 713   **Ma S, Shah S, Bohnert HJ, Snyder M, Dinesh-Kumar SP** (2013) Incorporating motif  
714       analysis into gene co-expression networks reveals novel modular expression  
715       pattern and new signaling pathways. *PLoS Genet* **9**: 1–20

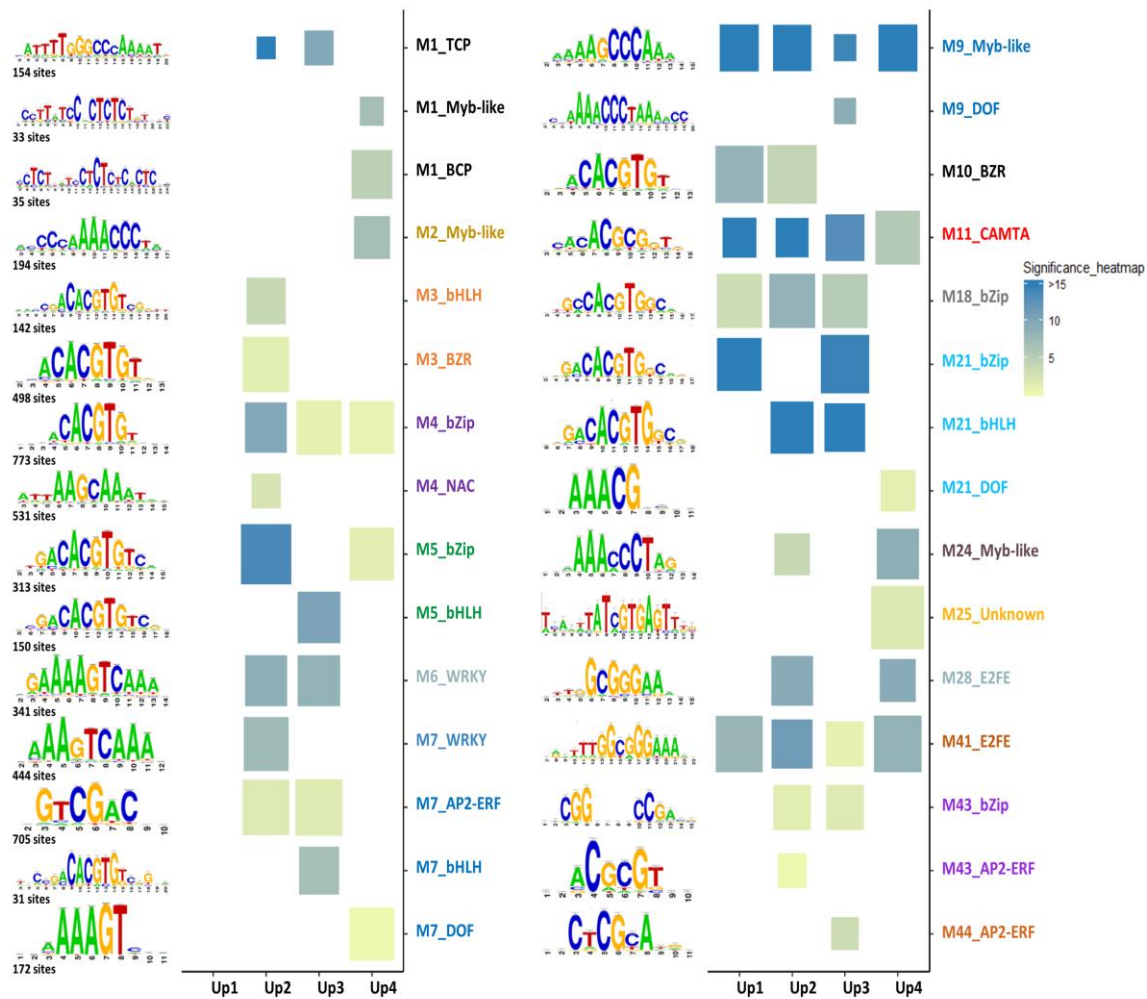
- 716 **Montardit Tardà F** (2018) Genomic delimitation of proximal promoter regions: Three  
717 approaches in *Prunus persica* [http://agris.fao.org/agris-](http://agris.fao.org/agris-search/search.do?recordID=QC2019600125)  
718 [search/search.do?recordID=QC2019600125](http://agris.fao.org/agris-search/search.do?recordID=QC2019600125)
- 719 **Nguyen NTT, Contreras-Moreira B, Castro-Mondragon JA, Santana-Garcia W,**  
720 **Ossio R, Robles-Espinoza CD, Bahin M, Collombet S, Vincens P, Thieffry D,**  
721 **et al** (2018) RSAT 2018: Regulatory sequence analysis tools 20th anniversary.  
722 *Nucleic Acids Res* **46**: 209–214
- 723 **Petrillo E, Godoy Herz MA, Barta A, Kalyna M, Kornblihtt AR** (2014) Let there be  
724 light: Regulation of gene expression in plants. *RNA Biol* **11**: 1215–1220
- 725 **Pimentel H, Bray NL, Puente S, Melsted P, Pachter L** (2017) Differential analysis of  
726 RNA-seq incorporating quantification uncertainty. *Nat Methods* 1–6
- 727 **Rombauts S, Déhais P, Van Montagu M, Rouzé P** (1999) PlantCARE, a plant cis-  
728 acting regulatory element database. *Nucleic Acids Res* **27**: 295–296
- 729 **Sanhueza D, Vizoso P, Balic I, Campos-Vargas R, Meneses C** (2015) Transcriptomic  
730 analysis of fruit stored under cold conditions using controlled atmosphere in  
731 *Prunus persica* cv. “Red Pearl.” *Front Plant Sci* **6**: 1–12
- 732 **Scheelbeek PFD, Tuomisto HL, Bird FA, Haines A, Dangour AD** (2017) Effect of  
733 environmental change on yield and quality of fruits and vegetables: two systematic  
734 reviews and projections of possible health effects. *Lancet Glob Heal* **5**: 21
- 735 **Sebastian A, Contreras-Moreira B** (2014) FootprintDB: A database of transcription  
736 factors with annotated cis elements and binding interfaces. *Bioinformatics* **30**:  
737 258–265
- 738 **Smale ST** (2001) Core promoters: Active contributors to combinatorial gene regulation.  
739 *Genes Dev* **15**: 2503–2508
- 740 **Steffens NO, Galuschka C, Schindler M, Bülow L, Hehl R** (2005) AthaMap web  
741 tools for database-assisted identification of combinatorial cis-regulatory elements  
742 and the display of highly conserved transcription factor binding sites in  
743 *Arabidopsis thaliana*. *Nucleic Acids Res* **33**: 397–402
- 744 **Tanou G, Minas IS, Scossa F, Belghazi M, Xanthopoulou A, Ganopoulos I,**

- 745       **Madesis P, Fernie A, Molassiotis A** (2017) Exploring priming responses involved  
746       in peach fruit acclimation to cold stress. *Sci Rep* **7**: 1–14
- 747       **Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, Van Helden**  
748       **J** (2012) A complete workflow for the analysis of full-size ChIP-seq (and similar)  
749       data sets using peak-motifs. *Nat Protoc* **7**: 1551–1568
- 750       **Tian F, Yang D-C, Meng Y-Q, Jin J, Gao G** (2019) PlantRegMap: charting functional  
751       regulatory maps in plants. *Nucleic Acids Res* **1**: 1–10
- 752       **Verde I, Abbott AG, Scalabrin S, Jung S, Shu S, Marroni F, Zhebentyayeva T,**  
753       **Dettori MT, Grimwood J, Cattonaro F, et al** (2013) The high quality draft  
754       genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity,  
755       domestication and genome evolution. *Nat Genet* **45**: 487–94
- 756       **Verde I, Jenkins J, Dondini L, Micali S, Pagliarani G, Vendramin E, Paris R,**  
757       **Aramini V, Gazza L, Rossini L, et al** (2017) The Peach v2.0 release: high-  
758       resolution linkage mapping and deep resequencing improve chromosome-scale  
759       assembly and contiguity. *BMC Genomics* **18**: 1–18
- 760       **Wong DCJ, Lopez Gutierrez R, Dimopoulos N, Gambetta GA, Castellarin SD**  
761       (2016) Combined physiological, transcriptome, and cis-regulatory element  
762       analyses indicate that key aspects of ripening, metabolism, and transcriptional  
763       program in grapes (*Vitis vinifera* L.) are differentially modulated accordingly to  
764       fruit size. *BMC Genomics* **17**: 1–22
- 765       **Yu C-P, Chen SC-C, Chang Y-M, Liu W-Y, Lin H-H, Lin J-J, Chen HJ, Lu Y-J,**  
766       **Wu Y-H, Lu M-YJ, et al** (2015) Transcriptome dynamics of developing maize  
767       leaves and genomewide prediction of cis elements and their cognate transcription  
768       factors. *Proc Natl Acad Sci* **112**: 2477–2486
- 769       **Yu CP, Lin JJ, Li WH** (2016) Positional distribution of transcription factor binding  
770       sites in *Arabidopsis thaliana*. *Sci Rep* **6**: 1–7
- 771       **Zhu Qun, Dabi T, Lamb C** (1995) TATA box and initiator functions in the accurate  
772       transcription of a plant minimal promoter in vitro. *Plant Cell* **7**: 1681–1689
- 773       **Zolotarov Y, Strömvik M** (2015) De novo regulatory motif discovery identifies

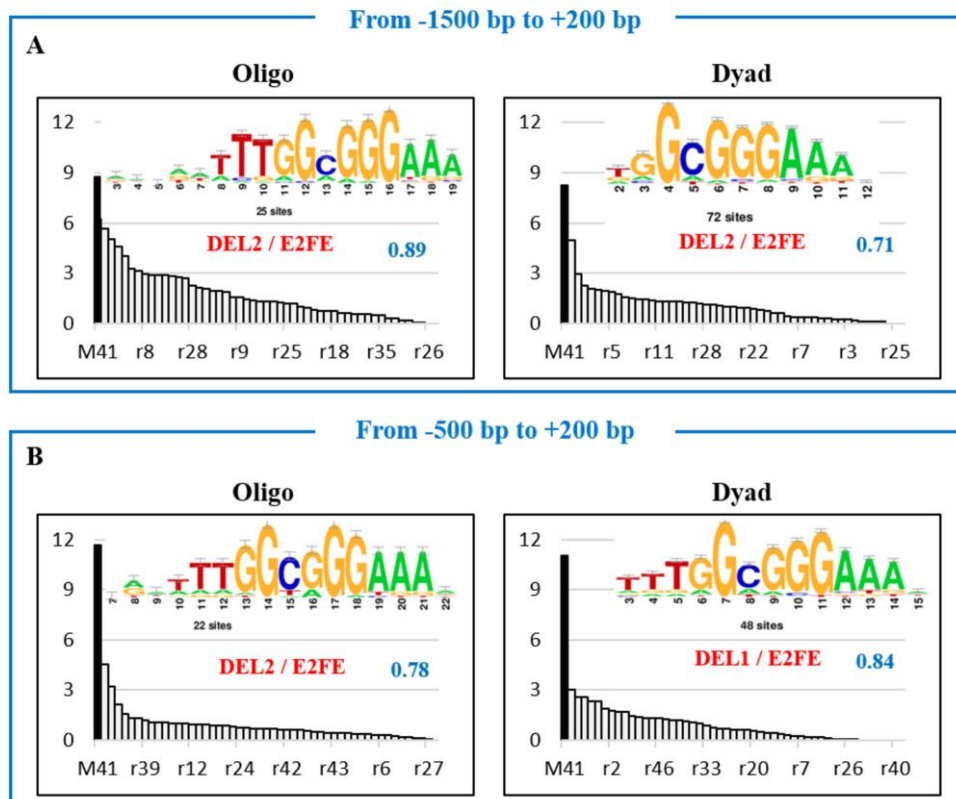
- 774 significant motifs in promoters of five classes of plant dehydrin genes. PLoS One  
775 **10**: 1–19
- 776 **Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, Shiu S-H**  
777 (2011) Cis-regulatory code of stress-responsive transcription in *Arabidopsis*  
778 *thaliana*. PNAS **108**: 14992–14977



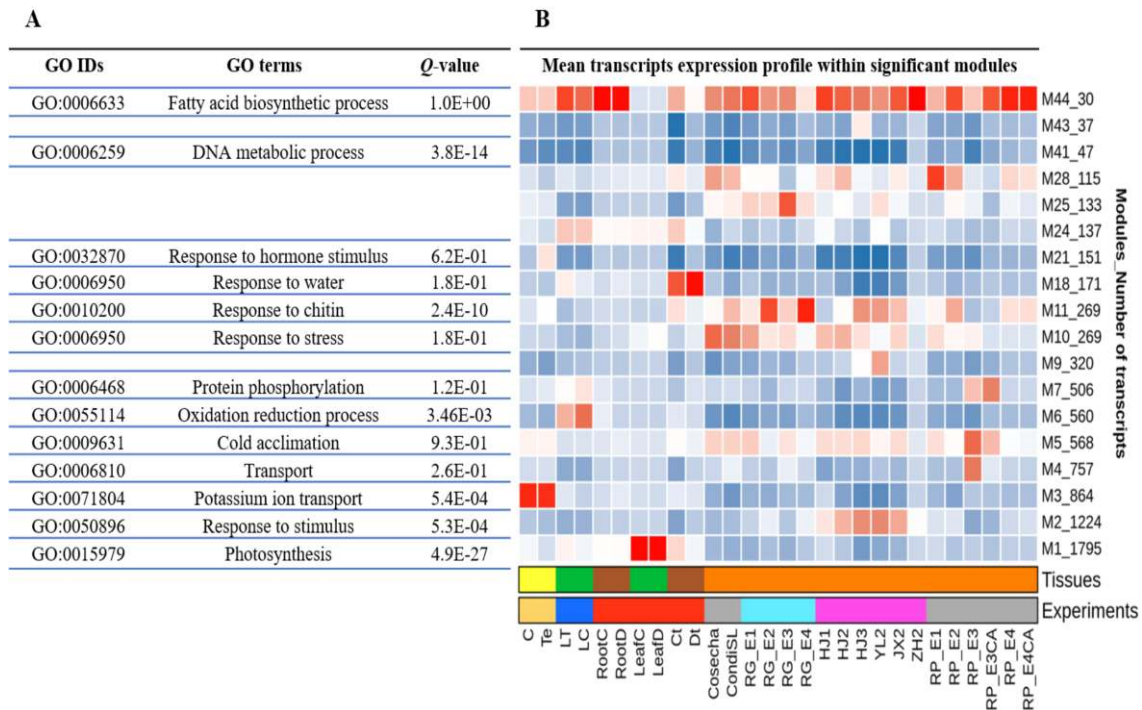
**Figure 1.** Bottom-up framework for *de novo* motif discovery. **Step1:** differential expression analysis for transcript detection and extraction of co-expressed modules. **Step2:** *de novo* motif detection using the peak-motifs tool from RSAT::Plants. Numbers correspond to the different tested upstream tracts, with TSSs anchored on position 0 bp, while letters represent tools within peak-motifs. Green and orange boxes label software and RSAT tools, respectively.



**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: [-1500 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 normalized correlation (Ncor). 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 **Table S3**. An interactive report with source code is accessible at [https://ead-csic-compbio.github.io/coexpression\\_motif\\_discovery/peach/](https://ead-csic-compbio.github.io/coexpression_motif_discovery/peach/)

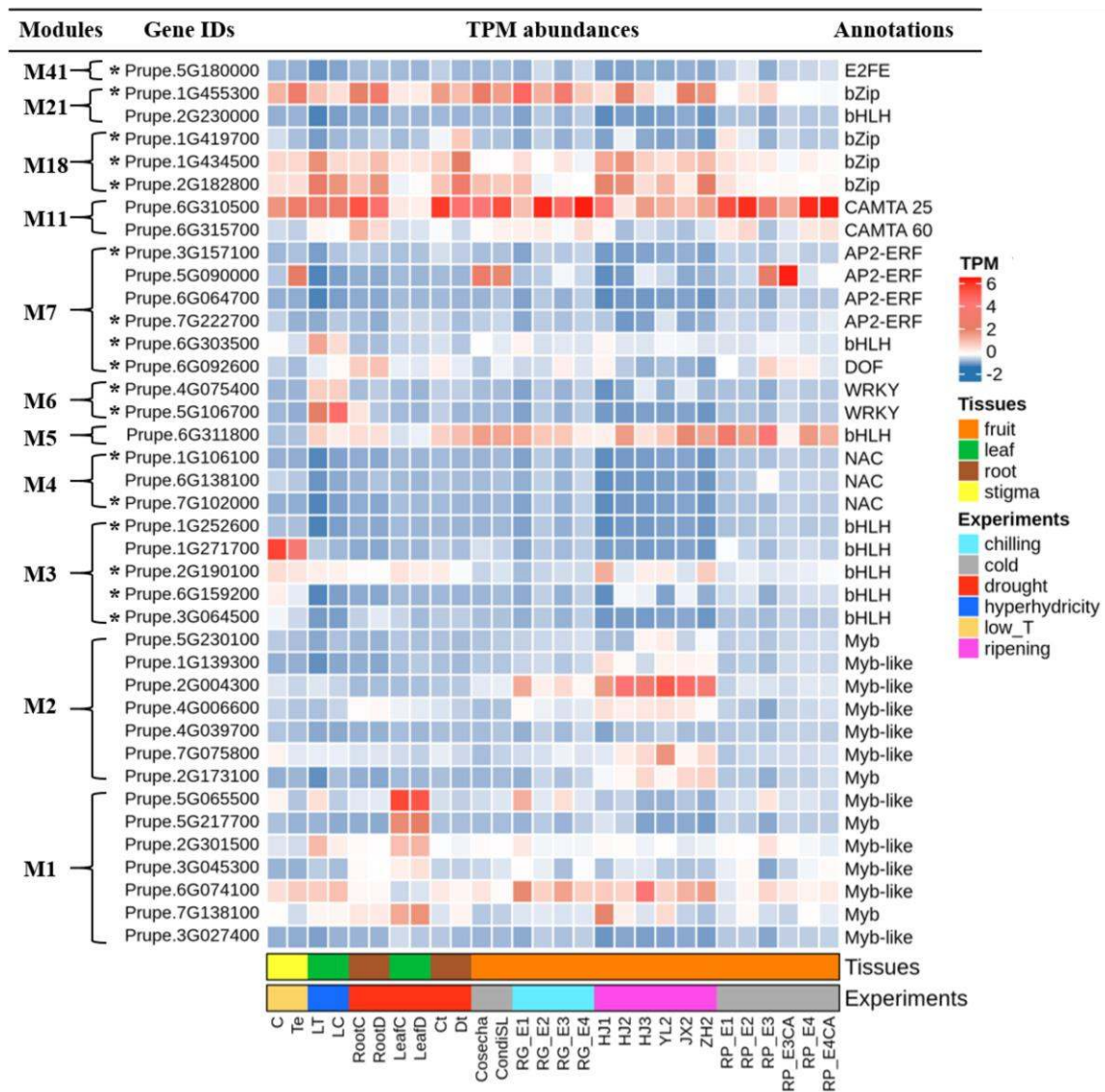


**Figure 3.** Illustrative comparison between predicted motif DEL2 (corresponding to E2FE transcription factor) within two different upstream promoter lengths: -1500 bp to +200 bp (**A**) and -500 bp to +200 bp (**B**). The name of the best match among plant motifs in footprintDB is labeled in red, next to its Ncor (Normalized correlation) 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  $-\log_{10}(P\text{-value})$ . 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 to negative controls (gray bars).

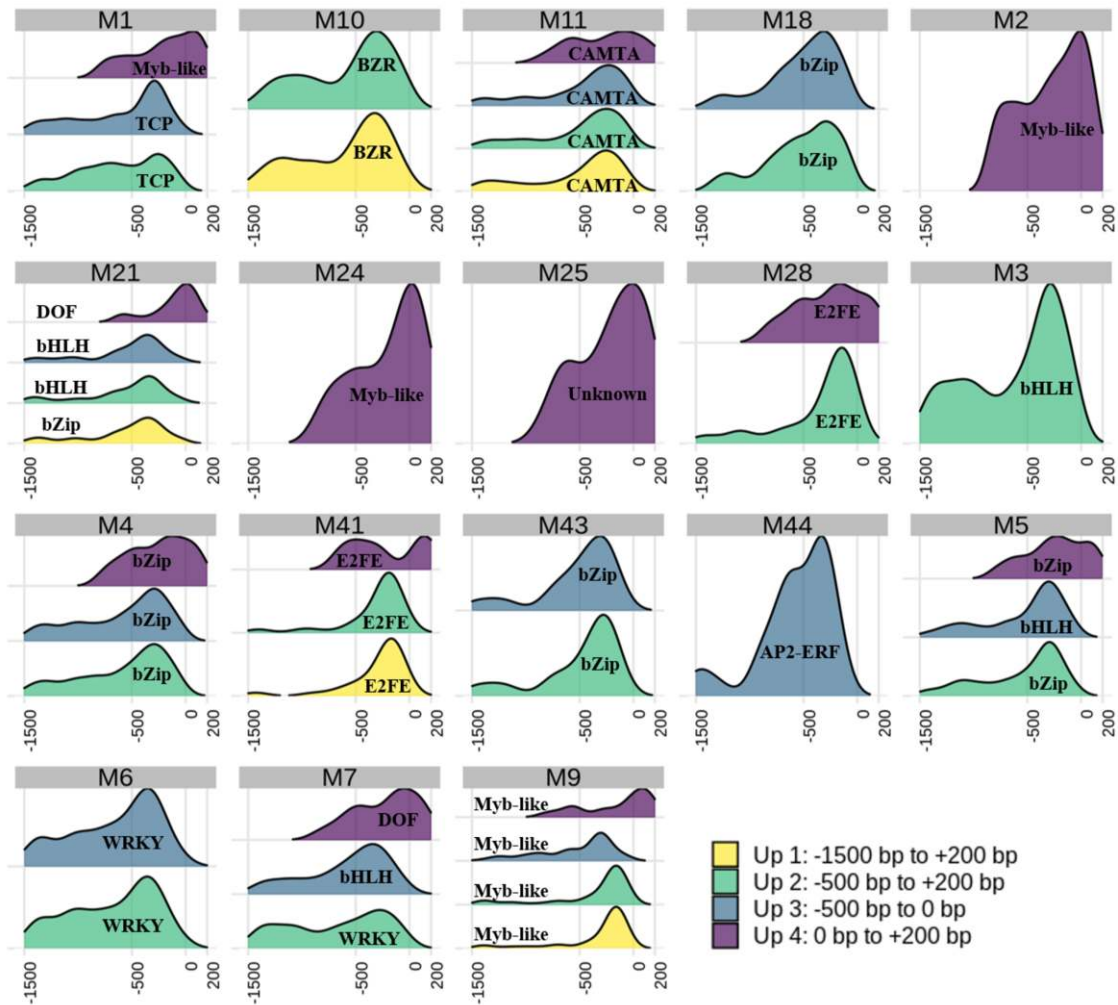


**Figure 4.** Functional annotation of relevant gene modules. **(A):** Gene ontology enrichment. **(B):** Mean transcript abundance profiling in term of transcripts per million (TPM). The x-axis corresponds the different experimental conditions while the y-axis indicates the number of differentially transcripts per module. Experiment and tissue types are highlighted by different colors (see the color key at the bottom of the figure). Gene profiles along the different conditions are provided at ([https://ead-csic.compbio.github.io/coexpression\\_motif\\_discovery/peach](https://ead-csic.compbio.github.io/coexpression_motif_discovery/peach)). See supplementary **Table S1** for the abbreviations.





**Figure 5.** List of transcription factors within relevant modules. Blue and red squares indicate transcripts per million while bottom color bars correspond to the tissues 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.



**Figure 6.** Positional distribution of the detected oligo motifs in promoter genes of *Prunus persica*. Four density distributions were derived from four assessed upstream regions. Up 1: from -1500 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 density of captured sites with  $P$ -value  $< 10 e^{-4}$ . Only oligo motifs are presented here, dyads are provided in the report at [https://ead-csic-compbio.github.io/coexpression\\_motif\\_discovery/peach](https://ead-csic-compbio.github.io/coexpression_motif_discovery/peach).

JASPAR TFBS	JASPAR logos	Parameters	Upstream 1	Upstream 2	Upstream 3	Upstream 4
MA0549.1 BZR 35*	<b>CACGTGG</b>	Significance Logo Ncor	--	3.44 <b>CCACGTGG</b> 0.72	--	8.28 <b>cCACGTGG</b> 0.73
MA0565.1 AP2-B3 106*	<b>TGCATGC</b>	Significance Logo Ncor	--	--	--	6.52 <b>TGCATGCA</b> 0.68
MA0931.1 bZip 339*	<b>GACACGTG</b>	Significance Logo Ncor	32.64 <b>tGCCACGTG</b> 0.64	38.66 <b>tGCCACGTG</b> 0.49	--	49.8 <b>tGACGTGGCG</b> 0.83
MA1197.1 CAMTA 351*	<b>aCCCGTg</b>	Significance Logo Ncor	52.96 <b>ttCACGCG</b> 0.76	66.33 <b>ttCACGCG</b> 0.77	8.32 <b>CaCGCGtC</b> 0.89	66.9 <b>ttCACGCG</b> 0.81
MA1224.1 AP2-ERF 470*	<b>GCCGAC</b>	Significance Logo Ncor	--	10.66 <b>CGCCGaC</b> 0.81	--	20.33 <b>CGCCGaC</b> 0.74
MA1276.1 DOF 223*	<b>aaAAAAGT</b>	Significance Logo Ncor	--	--	--	1.92 <b>AAAAGT</b> 0.75
MA1289.1 TCP 294*	<b>GGGACCAC</b>	Significance Logo Ncor	11.49 <b>tgGGGaCCACt</b> 0.84	17.39 <b>gGGGaCCACt</b> 0.57	--	19.18 <b>tGGGGACCAC</b> 0.74
MA1303.1 WRKY 303*	<b>aaAAGTCAACG</b>	Significance Logo Ncor	14.06 <b>aaataaaGTCaaCGt</b> 0.53	20.46 <b>GTtGACtTtt</b> 0.74	--	25.74 <b>aAAGTCAACG</b> 0.8
MA1355.1 Myb-like 231*	<b>aAACCCTAAtt</b>	Significance Logo Ncor	10.38 <b>aaCCCAatta</b> 0.53	18.89 <b>AaCCCAac</b> 0.72	--	26.85 <b>AACCCTAAtt</b> 0.65
MA1359.1 bHLH 258*	<b>CACGTG</b>	Significance Logo Ncor	34.74 <b>CCACGTGG</b> 0.61	49.94 <b>CacGtGcCaCGTG</b> 0.61	6.12 <b>GaCACGTGtc</b> 0.61	55.41 <b>cCACGTGGc</b> 0.64

**Figure 7.** Similarity between JASPAR motifs (considered as queries) and *de novo* predicted oligo motifs found in *Arabidopsis thaliana* along four different upstream regions. Numbers tagged with a star indicate number of peaks recovered by BLASTN (see Methods). The Ncor scores correspond to JASPAR databases. Only oligo-analysis motifs are shown (dyads are available at supplementary **Table S4**). Upstream 1: [-1500 bp to +200 bp], Upstream 2: [-500 bp to +200 bp], Upstream3: [-500 bp to 0 bp] and Upstream4: [0 bp to +200 bp]

## Parsed Citations

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

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**Baldoni E, Genga A, Cominelli E (2015) Plant MYB transcription factors: Their role in drought response mechanisms. *Int J Mol Sci* 16: 15811–15851**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

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

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

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

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**Cantalapiedra CP, Garcia-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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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 Responses to Clim. Chang. IntechOpen*, pp 90–108**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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 Synth. Promot. Methods Mol. Biol.* Humana Press, New York, pp 279–295**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**Defrance M, Janky R, Sand O, van Helden J (2008) Using RSAT oligo-analysis and dyad-analysis tools to discover regulatory signals in nucleic sequences. *Nat Protoc* 3: 1589–1603**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**Doherty CJ, Van Buskirk HA, Myers SJ, Thomashow MF (2009) Roles for *Arabidopsis* CAMTA transcription factors in cold-regulated gene expression and freezing tolerance. *Plant Cell* 21: 972–984**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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. *PNAS* 111: 2367–2372**

- Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- 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 studies: uncovering links between peach fruit heat treatment and cold storage responses. *Planta* 251: 1–18**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- 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 Des. Clim. fruit Crop*. Springer-Nature, pp 291–351**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- 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**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- 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* 1–7**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- Hu P, Li G, Zhao X, Zhao F, Li L, Zhou H (2018) Transcriptome profiling by RNA-Seq reveals differentially expressed genes related to fruit development and ripening characteristics in strawberries (*Fragaria × ananassa*). *Peer J* 6: 1–25**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- 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**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- Korkuc P, Schippers JHM, Walther D (2014) Characterization and identification of cis-regulatory elements in *Arabidopsis* based on single-nucleotide polymorphism information. *Plant Physiology* 164: 181–200**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- 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**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- 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**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- 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**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- Kumar N, Dale R, Kemboi D, Zeringue EA, Kato N, Larkin JC (2018) Functional analysis of short linear motifs in the plant cyclin-dependent kinase inhibitor SIAMESE. *Plant Physiol* 177: 1569–1579**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- Langfelder P, Horvath S (2008) WGCNA: An R package for weighted correlation network analysis. *BMC Bioinformatics* 9: 1–13**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- Li E, Liu H, Huang L, Zhang X, Dong X, Song W, Zhao H, Lai J (2019) Long-range interactions between proximal and distal regulatory regions in maize. *Nat Commun* 10: 1–14**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- 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**  
Pubmed: [Author and Title](#)  
Google Scholar: [Author Only Title Only Author and Title](#)
- 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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**Ma S, Bachan S, Porto M, Bohnert HJ, Snyder M, Dinesh-Kumar SP (2012) Discovery of stress responsive DNA regulatory motifs in *Arabidopsis*. *PLoS One* 7: 1–13**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

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

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

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

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**Rombauts S, Déhais P, Van Montagu M, Rouzé P (1999) PlantCARE, a plant cis-acting regulatory element database. *Nucleic Acids Res* 27: 295–296**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**Scheelbeek PFD, Tuomisto HL, Bird FA, Haines A, Dangour AD (2017) Effect of environmental change on yield and quality of fruits and vegetables: two systematic reviews and projections of possible health effects. *Lancet Glob Heal* 5: 21**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

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

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

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

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**Tanou G, Minas IS, Scossa F, Belghazi M, Xanthopoulou A, Ganopoulos I, Madesis P, Fernie A, Molassiotis A (2017) Exploring priming responses involved in peach fruit acclimation to cold stress. *Sci Rep* 7: 1–14**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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: 487–94**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**Wong DCJ, Lopez Gutierrez R, Dimopoulos N, Gambetta GA, Castellarin SD (2016) Combined physiological, transcriptome, and cis-regulatory element analyses indicate that key aspects of ripening, metabolism, and transcriptional program in grapes (*Vitis vinifera* L.) are differentially modulated accordingly to fruit size. BMC Genomics 17: 1–22**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

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

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

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

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)

**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*. PNAS 108: 14992–14977**

Pubmed: [Author and Title](#)

Google Scholar: [Author Only](#) [Title Only](#) [Author and Title](#)