PCP
PLANT & CELL PHYSIOLOGY

Exploiting Genomic Features to Improve the Prediction of Transcription Factor-Binding Sites in Plants

Quentin Rivière[1,*], Massimiliano Corso[1,2], Madalina Ciortan[3], Grégoire Noël[4], Nathalie Verbruggen[1,†] and Matthieu Defrance[3,*,†]

[1]Brussels Bioengineering School, Laboratory of Plant Physiology and molecular Genetics, Université Libre de Bruxelles, Brussels 1050, Belgium
[2]INRAE, AgroParisTech, Institut Jean-Pierre Bourgin (IJPB), Université Paris-Saclay, Versailles 78000, France
[3]Interuniversity Institute of Bioinformatics in Brussels, Machine Learning Group, Université Libre de Bruxelles, Brussels 1050, Belgium
[4]Functional and Evolutionary Entomology, Gembloux Agro-Bio Tech, University of Liège, Passage des Déportés 2, Gembloux 5030, Belgium
[†]These authors co-directed this work.

*Corresponding authors: Quentin Rivière, E-mail, qriviere@ulb.be; Matthieu Defrance, E-mail, matthieu.defrance@ulb.be

(Received 12 June 2021; Accepted 6 July 2022)

The identification of transcription factor (TF) target genes is central in biology. A popular approach is based on the location by pattern matching of potential *cis*-regulatory elements (CREs). During the last few years, tools integrating next-generation sequencing data have been developed to improve the performance of pattern matching. However, such tools have not yet been comprehensively evaluated in plants. Hence, we developed a new streamlined method aiming at predicting CREs and target genes of plant TFs in specific organs or conditions. Our approach implements a supervised machine learning strategy, which allows decision rule models to be learnt using TF ChIP-chip/seq experimental data. Different layers of genomic features were integrated in predictive models: the position on the gene, the DNA sequence conservation, the chromatin state and various CRE footprints. Among the tested features, the chromatin features were crucial for improving the accuracy of the method. Furthermore, we evaluated the transferability of predictive models across TFs, organs and species. Finally, we validated our method by correctly inferring the target genes of key TFs controlling metabolite biosynthesis at the organ level in Arabidopsis. We developed a tool—Wimtrap—to reproduce our approach in plant species and conditions/organs for which ChIP-chip/seq data are available. Wimtrap is a user-friendly R package that supports an R Shiny web interface and is provided with pre-built models that can be used to quickly get predictions of CREs and TF gene targets in different organs or conditions in *Arabidopsis thaliana, Solanum lycopersicum, Oryza sativa* and *Zea mays*.

Keywords: *Arabidopsis thaliana* ● *Cis*-regulatory elements ● Flavonoid target genes ● Genomics ● Plant organs ● Predictive modeling ● R package

Introduction

Gene regulation is one of the most fundamental biological phenomena. It explains how, from the same genetic code, a cell can harbor different states, according to the cell cycles and the signals from the environment. For multicellular organisms such as plants, gene regulation is also involved in processes such as cell specialization, organogenesis, growth and aging (Aerts 2012, Spitz and Furlong 2012). Gene regulation encompasses a cascade of regulatory processes that intervene during the flow of genetic information. The control of transcription by RNA polymerase II constitutes the first level of regulation. In order to transcribe a gene, the RNA polymerase II complex needs first to stably bind the DNA upstream in the vicinity of the transcription start site (TSS), in a region called the 'core promoter'. Some core promoters present DNA sequences that are sufficiently attractive, but in most cases the recruitment of the RNA polymerase involves interactions with components that are called transcription factors (TFs) and cofactors (Fuda et al. 2009).

TFs are key regulators of gene expression, characterized by DNA-binding domains that can recognize specific motifs of 6–20 nucleotides. They are proteins that bind to *cis*-regulatory regions located on the 'promoter', upstream of the TSS, near the binding sites of the RNA polymerase, and are classified as repressors or activators depending on whether they favor or block the recruitment of the subunits of the polymerase (Lee et al. 2012). However, the mechanisms of action of transcription are complex. Organisms pack the DNA in highly condensed structures, called 'chromatin', that allow fitting within the space of the cell (prokaryotes) or the nucleus (eukaryotes), making it difficult for regulatory molecules to access and bind to DNA. The action of some TFs consists therefore of triggering or maintaining the opening of the DNA at *cis*-regulatory regions

(Spitz and Furlong 2012). Another source of complexity originates from the existence of additional *cis*-regulatory regions located outside the promoter, such as enhancers and silencers (Lenhard et al. 2012). The latter are located remotely in terms of base pairs from the TSS, upstream, downstream or on the gene body, and interact with the promoter, thanks to the ability of the DNA to form loops that bring two regions closer together. TF activity is crucial to determine the state and identity of a cell and thus to regulate developmental processes and stress responses (Vaquerizas et al. 2009). This activity is dependent on the chromatin state and the TF expression, post-translational modifications and interacting partners, which might be specific to the condition or lineage (Veljkovic and Hansen 2004).

TF target genes can be predicted based on the sequence specificities of the *cis*-regulatory elements that can be recognized by the TF of interest. To identify TF target genes, several challenges have to be addressed: (i) identifying and modeling (as 'motifs') the sequence specificities of TF-binding sites (TFBSs) and (ii) locating and scoring the potential occurrences of motifs along *cis*-regulatory regions, i.e. 'pattern matching' (Aerts 2012). Since the 1980s, intense research efforts have been made in this field. For prokaryotes, efficient and well-performing methods have been obtained (Vuong and Misr 2011), while for most of the multicellular eukaryotes there is still a need for further development. The main difficulty for the latter organisms relies on the length of the *cis*-regulatory regions, which are longer than in prokaryotes (Hardison and Taylor 2012). Because the sequence of the TFBSs is highly variable and too short compared with the length of the regions considered as '*cis*-regulatory', a genome-wide analysis might identify almost all the genes as potential targets of the studied TFs. To restrict the width of the '*cis*-regulatory' regions on which pattern matching is performed, 'cluster' and 'phylogenetic' footprinting methods can be used, as TFs tend (i) to cluster (*cis*-regulatory regions show intervals with a high density of binding sites of the same and/or of different TFs) and (ii) to bind to sites (or clusters) that are evolutionarily conserved. These methods of footprinting still suffer, however, from a high number of false-positive predictions of *cis*-regulatory regions (Aerts 2012).

In recent years, new experimental strategies to study *cis*-regulatory elements on a genome-wide scale have emerged. In particular, ChIP-chip/seq has made it possible to map the binding regions of transcription (co)-factors (Mundade et al. 2014) as well as to study specific marks and variants of 'histones'. In eukaryotes, the histones are proteins that associate with DNA to form the 'chromatin'. In that structure, DNA is wrapped around a succession of yoyo-shaped histone octamers ('nucleosomes'), which can pile up in a closed and condensed structure which makes the DNA inaccessible to transcription (co)-factors (Bonev and Cavalli 2016). Some mechanisms allow unpacking of the structure, depending to a large extent on histone variants and marks (e.g. covalent modifications of the histone tails) (Lawrence et al. 2016) as well on methylation of the cytosine, which might be studied by BiSulfite-seq (BS-seq) (Jones 2012). The control of the DNA accessibility is therefore decisive in the regulation of the binding of *cis*-regulatory elements by TFs. Complementary techniques to ChIP-chip/seq and BS-seq, such as DNase I-seq, Assay for Transposase-Accessible Chromatin using Sequencing (ATAC-seq), Micrococcal Nuclease Digestion with deep Sequencing (MNase-seq), Nucleosome Occupancy and Methylome Sequencing assay (NOMe-seq) or Formaldehyde-Assisted Isolation of Regulatory Elements using Sequencing (FAIRE-seq), have allowed the degree of opening of the DNA to be directly probed (Meyer and Liu 2014).

The greater availability of genomic and epigenomic data paved the way for new bioinformatic methods dedicated to the prediction of TFBSs. An expanding number of tools have been released (Gusmao et al. 2016, Jankowski et al. 2016, Kumar and Bucher 2016, Chen et al. 2017, Liu et al. 2017, Qin et al. 2017, Quang and Xie 2017, Schmidt et al. 2017, Schmidt et al. 2019, Behjati Ardakani et al. 2019, Keilwagen et al. 2019, Li and Guan 2019, Li et al. 2019). Of particular interest is the new footprinting approach, called 'digital genomic' footprinting, which is based on the property of the TFs to protect the *cis*-regulatory elements from cleavage by DNase I. In contrast to 'cluster' and 'phylogenetic' footprinting techniques, digital genomic footprinting takes into account the chromatin state dynamics and therefore the accessibility of the *cis*-regulatory elements across treatments, growth stages or cell types and tissues.

However, in plant species, long-established techniques have not been systematically compared with new methods and, importantly, integrative tools able to combine all these techniques are still lacking (Lai et al. 2019). Therefore, we developed Wimtrap, a tool to predict condition- or organ-specific *cis*-regulatory elements and TF gene targets, with a great flexibility regarding the input data. We used this tool to compare most of the different techniques described above and to evaluate the benefits of combining them. Accuracy of the predictions was obtained based on ChIP-chip/seq data and allowed the validation of Wimtrap. We illustrated the use of our tool with an example highlighting the strength of the condition specificity of the predictions, taking into consideration TFs that control the late steps of flavonoid biosynthesis. Wimtrap is implemented as a fully documented R package (https://github.com/RiviereQuentin/Wimtrap) and Shiny application (https://github.com/RiviereQuentin/WimtrapWeb). We focused mainly on *Arabidopsis thaliana* (L.)—the model species for plant genetics and molecular biology—but extended our work to other plant species. Wimtrap works currently for *A.thaliana* in 10 conditions (organs or growing conditions), *Solanum lycopersicum* in two conditions, and *Oryza sativa* and *Zea mays* in one condition.

Results

Analysis overview

We developed a machine learning approach (**Fig. 1**) to predict *cis*-regulatory elements and TF target genes using information

obtained from motifs of TFBSs, DNA sequence, transcript models, conserved elements and/or epigenetic data. The method is focused on plants, especially on *A.thaliana*, the model species for plant genetics and molecular biology, for which data are the most abundant. The different analyses that we performed, as well as our workflow, can be schematically described as follows.

First of all, based on a literature search and the query of seven specialized databases, we retrieved: (i) the genomic sequences and the transcript models of *A.thaliana*, *S.lycopersicum*, *O.sativa* and *Z.mays*, (ii) the motifs and ChIP-chip/seq data for 57 TFs in the seedlings and flowers of *A.thaliana*, in the ripening fruits of *S.lycopersicum*, in the seedlings of *O.sativa* and in the seedlings of *Z.mays*, (iii) five genomic maps of *cis*-regulatory elements (two in *A.thaliana*, and one each in *S.lycopersicum*, *O.sativa* and *Z.mays*), (iv) five genomic maps of digital genomic footprints (DGFs; one in *A.thaliana* flowers and one each in seedlings of the other three species and (v) 42 chromatin feature-peak data (24 in *A.thaliana* seedlings, three in *A.thaliana* flowers, three in *S.lycopersicum* seedlings, nine in *O. sativa* seedlings and three in *Z.mays* seedlings). These data and information were then integrated into the Wimtrap pipeline, composed of several steps (**Fig. 1**):

Step 1: Motif match location along the genome. Pattern matching analyses were carried out with the motifs of the TFs to obtain the location of the candidate binding sites (= the motif matches). Each candidate was scored according to the fit of the DNA sequence with the motif.

Step 2: Motif match annotation. The motif matches were annotated with features characterizing their genomic context. The distance of the candidate binding sites to the closest transcript was calculated using the TSS as reference. The structure (promoter, coding sequence, etc.) overlapped by the candidate binding sites was also determined. Then, the average signal or the density of peaks/elements of the features related to DNA sequence conservation, DGFs and chromatin state were computed on intervals of $\pm$ 10, $\pm$ 200 or $\pm$500 bp around the potential *cis*-regulatory elements.

Step 3: Motif match labeling and dataset balancing. The motif matches were labeled as 'positive' ('is.TFBS=1') when they were validated by available ChIP-chip/seq data, and as 'negative' ('is.TFBS=0') when not. To avoid an over-representation of the negative candidate binding sites compared with the positive ones, a subset of negative candidate binding sites was randomly selected so that the composition of the dataset changed to 50% of negative and 50% of positive motif matches. Balancing a dataset is a classical approach to overcome the tendency of predictive models to categorize all the instances into the most prevalent class (here, that of the 'negative' candidate binding sites) when the minority class (the 'positive' candidate binding sites) is rarely represented (Kotsiantis et al. 2006).

Step 4: Modeling of motif match classifiers. 'Decision-rule' models, made up of a collection of regression trees, were trained by extreme gradient boosting (Chen and Guestrin 2016). Such models allowed a decision to be made as to whether a candidate TFBS was 'positive' or 'negative' based on the integrated features. Two kinds of models were built: the TF-specific models, based on data from a single TF, and the TF-pooled models, obtained from all the TFs considered in a given organism and condition (seedlings of *A.thaliana*, flowers of *A.thaliana*, ripening fruits of *S. lycopersicum*, seedlings of *O.sativa* or seedlings of *Z.mays*). The TF-specific models were trained with different sets of features in order to compare the predictive potential of existing techniques and assess the benefits of an integrative approach.

Step 5: TFBS prediction. New motif matches, which were not used in the modeling process and were located and annotated as described above (steps 1 and 2), were fed to the binary classifiers. These models classified the candidates as either 'positive' ('is.TFBS=1') or 'negative' ('is.TFBS'=0). The first ones were retained as the predicted TFBSs, while the second ones were filtered out.

Finally, we evaluated Wimtrap. TF-specific models were used to assess the method accuracy, depending on the integrated features, as well as to perform state-of-the art feature importance analyses. TF-pooled models were tested for their transferability across TFs/conditions/organisms, and were applied to a real-world example of a case study. Model accuracies were calculated by computing the area under the ROC (receiver operating characteristic) curve (AUC). For TF-specific models, we proceeded to the 5-fold cross-validation protocol: each TF-specific dataset was split into five sub-datasets. Training and AUC computation were iterated five times, each time using a different sub-dataset for obtaining the ROC curve. For TF-pooled models, we tested such models on TFs and/or condition or organism that were not taken into account in the training.

Performances of TF-specific models according to the integrated features

Based on the 28 TFs studied by ChIP-seq in *A.thaliana* seedlings, we computed the ROC curves of TF-specific models trained with different groups of features, taken individually or in combination (**Fig. 2**). These groups of features were called 'layers', as they represented distinct layers of information that could be added to each other. The layers are the following: (i) motif occurrences and scores, (ii) position related to the transcript model, (iii) DNA sequence conservation, (iv) DGF occurrence and score, and (v) chromatin state.

Each layer corresponds to a given technique. The first layer is related to the 'cluster' footprinting, the second to the tendency of the TFs to be located on the promoters in proximity to the TSS, the third to the 'phylogenetic' footprinting, the fourth to the 'digital genomic' footprinting and the fifth to the association of TFs with a genomic region characterized by an open state of the chromatin.

For each layer of features, we also briefly characterized the association between the *cis*-regulatory elements and the fea-

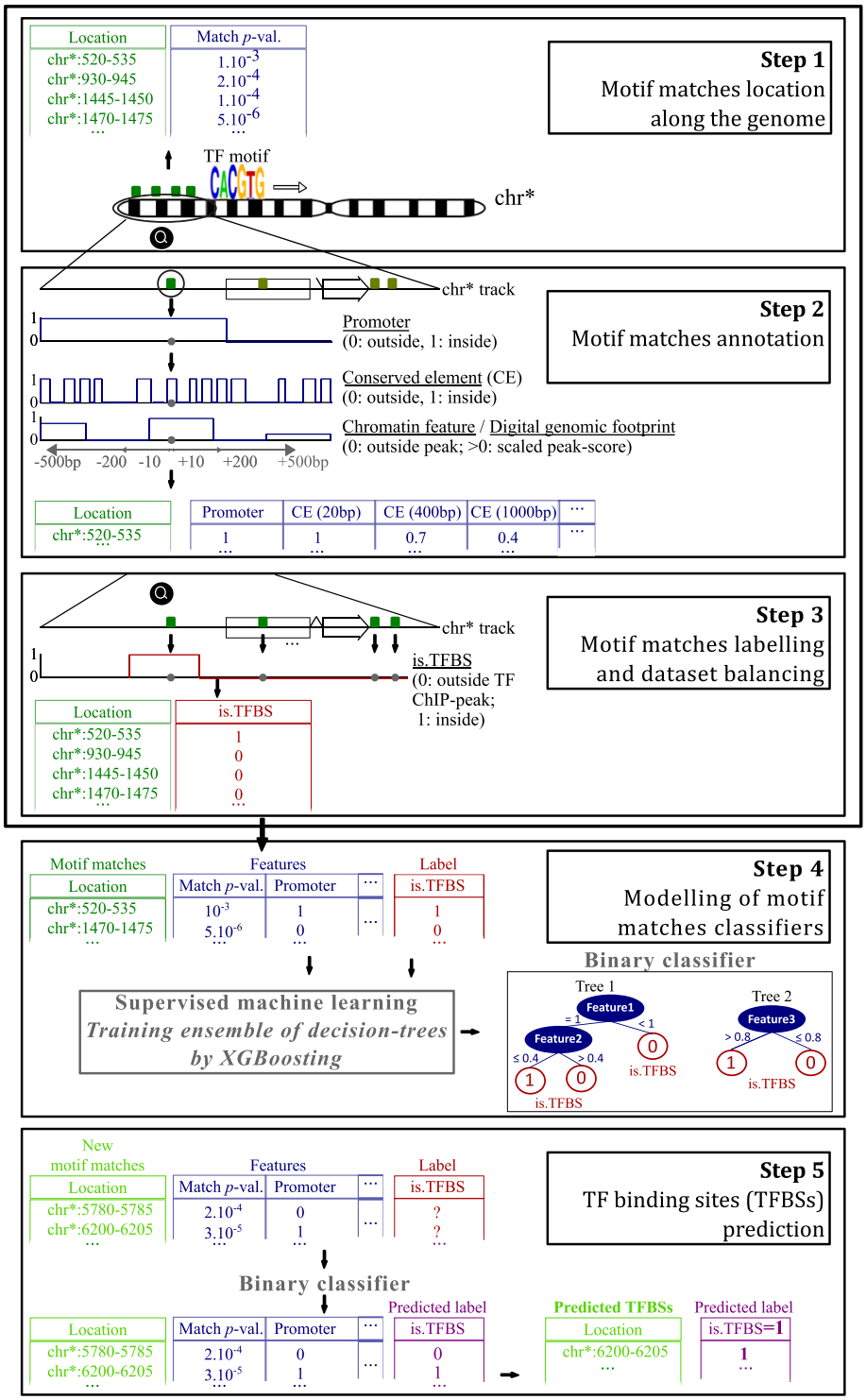

Fig. 1 Methodology workflow diagram. Location, annotation and labeling of candidate TF-binding sites. The candidate binding sites of the TFs included in ChIP peak results are (i) located by pattern matching, (ii) labeled as 'positive' ('is.TFBS = 1′) when they overlap a ChIP peak in the considered condition and 'negative' when not ('is.TFBS = 0) and (iii) annotated with their position on the closest transcript and with the average signal of the genomic data in their neighborhood, on windows of ±10, ±200 and ±500 bp. Machine learning: A binary classifier is trained by extreme gradient boosting (XGBoosting). The latter implements a supervised learning strategy, where an ensemble of decision tree models are trained to predict the label of motif matches based on their match *P*-value and their annotations. Prediction of TF-binding sites: The binary classifier is used to predict the label of a new set of motif matches, located and annotated as described above. The motif matches assigned to 'is.TFBS = 1′ by the classifier are those predicted as TF-binding sites. The other motif matches, assigned to 'is.TFBS = 0′, are left over. PWM, position weight matrix; TF, transcription factor; Chr, chromosome; CE, conserved element; TFBS, TF-binding site.

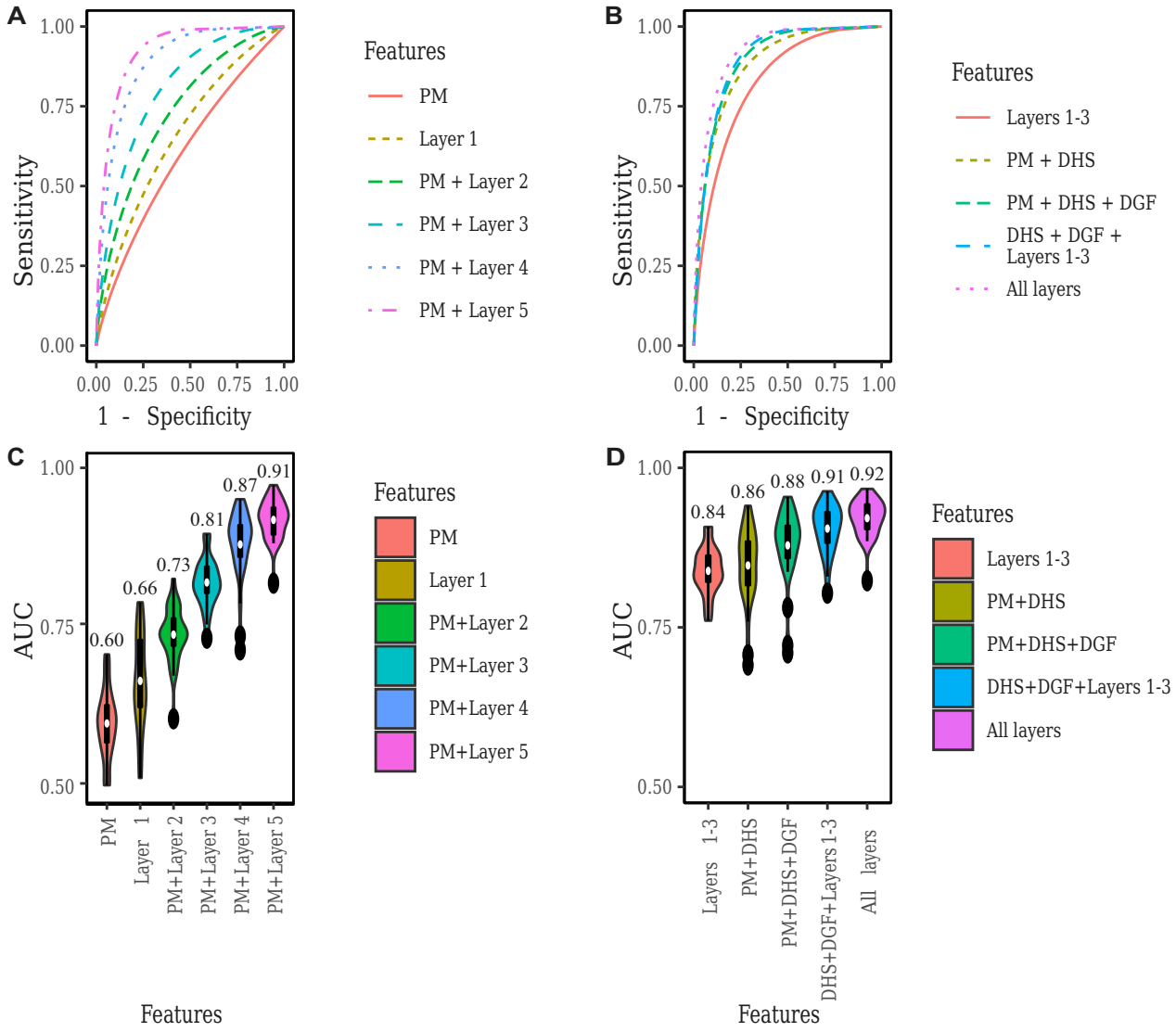

Fig. 2 Predictivity of the layers of features and selected combination of features (A) Mean ROC curve and (B) AUC achieved by internal validation of TF-specific models that integrate, in addition to the *P*-values of the matching score of the PWM matches, the genomic context features that belong to the different layers of features. For each TF, a model is built and evaluated based on a balanced dataset for that factor following the 5-fold cross-validation procedure: the considered dataset is divided into five partitions. Of these, four are considered to build a model and one is used to assess the performance. The operation is repeated five times, in such a way that each partition is retained only once for validation purposes. (C, D) As above, but considering a combination of selected features. PM, pattern matching; DHS, DNase I hypersensitivity; DGF, digital genomic footprint scores. Layer 1 includes the results of pattern matching.

tures. These associations can be visualized in **Supplementary Figs. S1–S5**, in which Pearson's correlation between the features and the 'is.TFBS' label of the potential binding sites (equal to 1 when a potential binding site is 'positive', and to 0 when it is 'negative') was plotted.

Layer 1: Motif occurrence and score. Layer 1 allowed assessment of the pattern matching and the 'cluster' footprinting method as it includes the *P*-value of the PWM (pseudo weight matrix) matches and the number of matches co-occurring in the vicinity of the potential binding sites. Models based solely on pattern matching (scores of the PWM matches) were associated with an average AUC of 0.60 (**Fig. 2C**). Integrating the density of PWM matches on windows of 400 bp or 1,000 bp led to an AUC of 0.66. Features of the layer showed a variable but overall low ability to filter the potential binding sites (**Supplementary Fig. S1**). The *P*-values of the PWM matches exhibited low predictive levels, except for the TFs NAC50 and NAC52.

Layer 2: Position on the gene. Layer 2 allowed evaluation of the rationale behind promoter scanning. Models integrating

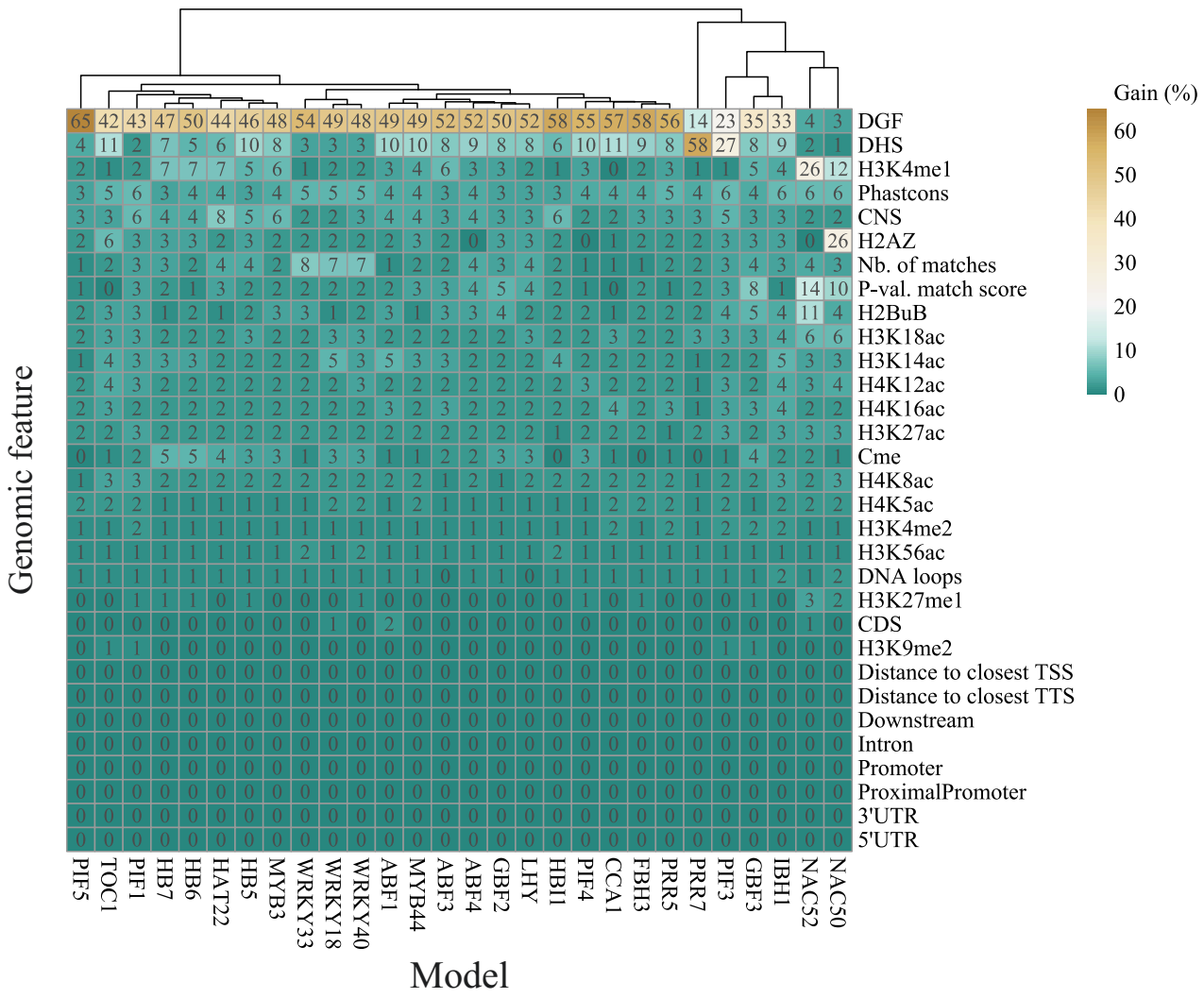

Fig. 3 Importance of the genomic features in the full TF-specific models obtained from TFs studied in seedlings of *Arabidopsis thaliana*. Importance is expressed in terms of gains. Only the features selected in at least one model are shown. The features are ordered according to their average importance amongst the models considered, while the TFs are ordered by hierarchical clustering. For each item of data, the gains associated with the features extracted on windows of 20, 400 and 1,000 bp are summed. DGF, digital genomic footprint; DHS, DNase I hypersensitivity; CNS, conserved non-coding sequence; Nb., number; P-val, *P*-value; Cme, cytosine (DNA) methylation; TSS, transcription start site; TTS, transcription termination site; UTR, untranslated region.

the results of pattern matching with the position on the gene (structure and distance to the closest TSS) reached on average an AUC of 0.73 (**Fig. 2C**). We found that potential binding sites located on the promoter or the 5′-untranslated region (5′-UTR) were more likely to be *cis*-regulatory elements, while those located on the intron or coding sequence were less likely to be so (**Supplementary Fig. S2**). The chance for a PWM match to be a *cis*-regulatory element increased while getting closer upstream to the TSS but suddenly dropped at several base pairs downstream from the TSS. Overall, 49% of the *cis*-regulatory elements were located on the promoter at a maximum of 2,000 bp from the TSS, while 9% were located on the 5′-UTR (**Fig. 4**). The 42% remaining *cis*-regulatory elements were distributed as follows: (i) 18% in the gene body, downstream of the 5′-UTR (i.e. the coding sequence, intron and 3′-UTR), (ii) 8% in the regions downstream of the transcript stop site and (iii) 16% in the intergenic regions.

Layer 3: DNA sequence conservation. We integrated two sets of conserved elements in *A.thaliana*, from which we respectively derived the 'Conserved Non-Coding Sequences' ('CNS') and 'Phastcons' datasets. The first dataset was built by combining the location of non-coding conserved elements predicted by three independent studies (Thomas et al. 2007, Baxter et al. 2012, Haudry et al. 2013), which analyzed the homeologs in *A.thaliana* and the orthologs in the eudicots and the family of the Brassicaceae. The second dataset is composed of

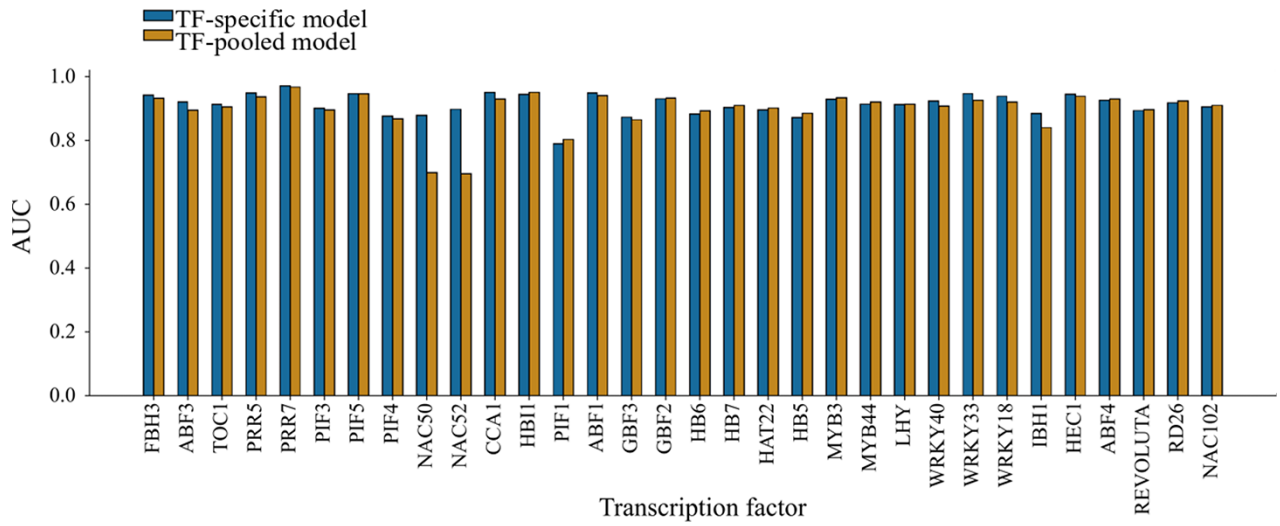

Fig. 4 Comparison of the performance of the TF-specific models with the TF-pooled models For each TF, a model based on the data related to this TF (TF-specific model) and to the other TFs (general model) are compared. The area under the ROC curve (AUC) is evaluated. The features of all the layers are combined to build the models.

scored phylogenetic footprints that have been identified with the 'phastCons' tool (Siepel and Haussler 2005) from the alignment of the coding and non-coding sequences of ortholog genes belonging to 63 monocots and eudicots plants species (Tian et al. 2020). Layer 3 is associated with the 'phylogenetic footprinting' approach. Models combining the results of pattern matching and sequence conservation obtained an average AUC of 0.81. With the 'CNS' dataset, we observed a clear tendency of the *cis*-regulatory elements to be associated with phylogenetic footprints (**Supplementary Fig. S3**). However, some differences across TFs were found. For instance, the binding sites of NAC50 and NAC52 did not tend to be associated with evolutionarily conserved regions. For CCA1, HAT22, MYB44, HB5, HB7, HB6 and LHY, the *cis*-regulatory elements did not tend to be conserved (cf. 20 bp windows) but were associated with highly conserved surrounding regions (cf. 400 bp and 1,000 bp windows). With the second set of conserved elements (named 'Phastcons'), the association between the *cis*-regulatory elements and high degrees of conservation of DNA sequence was generally weak (Pearson's correlation of −0.11 on average).

Layer 4: DGF occurrence and score. Layer 4 was constructed based on the results of a state-of-the-art digital genomic footprinting analysis. Models built on the results of pattern matching and digital genomic footprinting reached an average AUC of 0.87 (**Fig. 2C**). The *cis*-regulatory elements were preferentially located in regions of a high density of DGFs (cf. 400 bp and 1,000 bp windows) (**Supplementary Fig. S4**). NAC50 and NAC52 *cis*-regulatory elements were not associated with DGFs, in contrast to those of the other TFs.

Layer 5: Chromatin state. The integration of 23 chromatin state-related features into the results of pattern matching led to models with an average AUC of 0.91 (**Fig. 2C**). The *cis*-regulatory elements were found to be associated with different chromatin states defined by Sequeira-Mendes et al. (2014) and ranked from 'A' to 'I' according to their degree of DNA opening. The association was positive with the 'B' and 'D' chromatin states and negative with the 'G', 'H' and 'I' states. Sequeira-Mendes et al. (2014) observed that the chromatin states 'B' and 'D' tended to occur on intergenic regions (including promoters and enhancers), the 'G' on introns and coding sequences, and the 'H' and 'I', on heterochromatin (**Supplementary Fig. S5**). When assessing in more detail the individual variables characterizing the chromatin state, the eight features most associated with cis-regulatory elements were, in decreasing order of association: the DNase I hypersensitivity score (DNAseI-hypersensitive sites (DHS); a measure of the opening of DNA), the H3K4me1 histone mark, the methylation of cytosine, the nucleosome density and the H3K27me1, H3K9me2, H3K56ac, H2BuB and H3K18ac histone marks. TFs showed overall homogeneous patterns. However, for four of them, several important features were not associated with *cis*-regulatory elements. This was the case for NAC50 and NAC52, for which a lack of predictivity of the DHS and H3K56ac could be observed, as well as CCA1 and IBH, for which the nucleosoms density, and H3K18ac and H2BuB histone marks were not predictive of *cis*-regulatory elements.

For this layer, we also assessed whether the association of the chromatin state features with the *cis*-regulatory elements depended on the distance to the TSS because differences between the promoters and the enhancers were expected (Sequeira-Mendes et al. 2014) (**Supplementary Fig. S6**). There were five chromatin features for which the signal was on average distinct between positive and negative potential binding sites independently from the distance to the TSS: DHS, H3K4me1, H3K27me1 and H3K9me3. The remaining features showed little association with the *cis*-regulatory elements in the immediate vicinity of the TSS. On distal regions, H2A.Z, H3K56ac

and H4K5ac showed strong associations with binding sites. As regards H3K18ac and H3K27me3, a striking finding was that *cis*-regulatory elements were associated with high or low levels depending on whether the regions were distal or proximal to the TSS (< −2,500 bp for H3K27me3, < −5,000 bp for H3K18ac).

Combination of layers. The combination of the condition-independent layers 1, 2 and 3 allowed us to obtain an average AUC of 0.84. The combination of the whole set of layers led to an average AUC of 0.92.

Restriction of layer 5 to the DHS features only. Finally, we generated ROC curves using only the features related to DNA opening (DHS) to consider the chromatin state. We found that DHS was the feature the most associated with the *cis*-regulatory elements in layer 5 (**Supplementary Fig. S7**). Models based only on pattern matching and DHS showed an average AUC of 0.86 (**Fig. 2D**). Adding only layer 4 to these models led to an average AUC of 0.88, while adding layer 4 together with layers 1, 2 and 3 led to an average AUC of 0.91 (**Fig. 2D**).

Importance of features in the full TF-specific models

We studied the relative importance of the features in the 28 TF-specific models built in *A.thaliana* seedlings (see 'Analysis overview') based on the whole set of features (layers 1–5) ('full' TF-specific models). We considered the gain, a classical metrics for XGBoost models. The gain of a feature is equal to the sum of the gains at each branch that uses this feature to operate a split, divided by the sum of the gains of all the features. XGBoost adds new splits on regression trees depending on the added gain, which reflects the increase of accuracy in a leaf when this leaf is further split into two new ones (Chen and Guestrin 2016). The DGF (layer 4), associated with an average gain of 42%, appeared as the most important feature in the TF-specific models for all the TFs (**Fig. 3**). The other features had on average <10% of gain. The most important features among those were, in decreasing order of importance: DHS (layer 5), H3K4me1 (layer 5), PhastCons (layer 3), CNS (layer 3), H2A.Z (layer 5), number of matches (layer 1) and *P*-value of the PWM match score (layer 1). They accounted for a gain of 24% on average. They were followed by the remaining features of layer 5 (cytosine methylation and the histone marks—except H3K4me1), responsible for a total average gain of 34%, and by the features of layer 2 (position on the gene), which did not bring any gain to the models. The absence of gain associated with the features of layer 2 indicated that they were redundant with other features, most probably with the histone marks and variants, which can be combined to predict the position of different gene structures along the genome (Heyndrickx et al. 2014). We can therefore postulate that such features add similar information when they are integrated into a model already including features related to DNA opening (such as DHS or DGF). This probably explains why removing the histone marks from the full TF-specific models reduced the AUC by only 0.1, from 0.91 to 0.92 (**Fig. 2D**).

Essentially, the analysis of feature importance agreed with that of the performances of the TF-specific models according to the integrated feature, which was presented in the previous section. In fact, DGF and DHS were the most determinant features in increasing the accuracy of the sole pattern matching but could not explain alone the performances of the full models (**Fig. 2D**). In addition, there were only a few noticeable variations across the TF-specific models for the importance of the feature. Among them, for some TFs, DGF was not the most important feature. In PRR7 and PIF3 models, this was the DHS; in the NAC50 model, the H3K4me1 histone mark; and in the NAC52 model, the H2A.Z variant. The *P*-value of the PWM match was only 3% on average, but it increased to 8, 14 and 10% in the models of GBF3, NAC52 and NAC50, respectively, except *PRR7*, *PIF3*, *NAC50* and *NAC52*. However, the overall homogeneity of the results observed across the different TFs justifies testing the performances of TF-pooled models.

Transferability of TF-pooled models

To evaluate the generalization of Wimtrap, we trained general models by pooling data related to all TFs except one and evaluated the performances on the TF that was left over. We then compared for each TF the performance of the general model with the one obtained with its specific model (**Fig. 4**). We applied this approach for each of the selected 28 TFs in Arabidopsis seedlings. Performances of the TF-pooled models and of the TF-specific ones were similar, except for NAC50, NAC52, and IBH1.

We also evaluated the transferability of TF-pooled models across conditions or species. We could build models only from *A.thaliana* flowers, from *S.lycopersicum* ripening fruits, *O.sativa* seedlings and *Z.mays* seedlings as we could not find more than two TF ChIP-chip/seq data for other plant species/conditions. In Arabidopsis seedlings, we assessed a TF-pooled model trained from Arabidopsis flowers, *S.lycopersicum* ripening fruits, *O.sativa* seedlings and *Z.mays* seedlings. The set of features integrated in the models was restricted to the features of layers 1–4 in addition to the DHS and the methylation of cytosine. Indeed, all the genomic data were not available in both the training and tested condition 'organism'. We extracted the epigenetic data related to Arabidopsis seedlings and used the models obtained from Arabidopsis flowers, *S.lycopersicum* ripening fruit, *O.sativa* seedlings and *Z.mays* seedlings, respectively, to predict the binding sites in Arabidopsis seedlings. This allowed us to reach an average AUC of 0.80 with the first model, 0.86 with the second one, 0.68 with the third one and 0.82 with the last one (**Fig. 5**). These were higher values than the average AUC of 0.60 associated with sole pattern matching (**Fig. 2**).

As the model obtained from *O.sativa* showed lower performances than the other models applied to Arabidopsis seedlings, we performed additional analyses to obtain further insights. We built an extended model, based on the above-mentioned features and on seven different chromatin marks. In rice, the chromatin marks were more important than the DHS and the DGF.

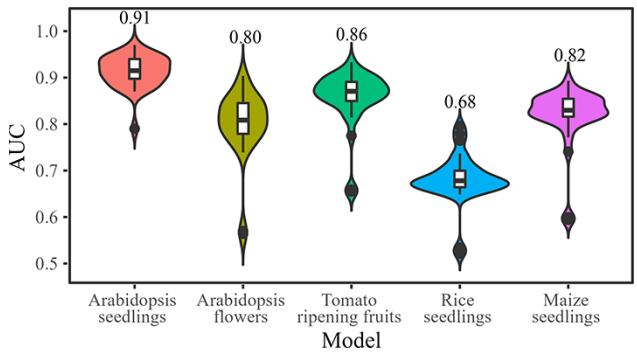

Fig. 5 Performances of models trained on Arabidopsis seedlings, Arabidopsis flowers, tomato ripening fruits, rice seedlings and maize seedlings, and evaluated on the 28 TFs studied in Arabidopsis seedlings The area under the ROC curve is reported. In the Arabidopsis flowers model, the features of layers 1, 2, 3 and 4 are integrated in addition to the DHS and the methylation of the cytosine, while in the tomato ripening fruits model, the features of layers 1 and 3 are in integrated in addition to the Phastcons, the DHS and the methylation of cytosine.

Accordingly, the AUC obtained for predicting *cis*-regulatory elements of TFs in *O.sativa* seedlings increased from 0.76 to 0.84 when the chromatin marks were integrated with the features of layers 1–4, the DHS and the methylation of cytosine (data only presented in text).

Characterization of targets of MBW TFs involved in the regulation of plant flavonoids

In order to test Wimtrap on a real-world application, we proceeded to identify the gene targets and validate the pathways that are controlled by TT2, TT8 and TTG1, which constitute the MYB–bHLH–WD40 (MBW) complex and that control the synthesis and accumulation of secondary metabolites (i.e. phenylpropanoids) in seeds. While TT8 and TTG1 participate in other MBW complexes that control the accumulation of flavonoids in leaves (e.g. TTG1–TT8–PAP1), TT2 is more specific to seeds, despite recent studies suggesting that TT2 also controls heat stress responses in the vegetative parts of Arabidopsis (Jacob et al. 2021). The decision to study the TT2–TT8–TTG1 complex is based on our expertise in phenylpropanoid compounds in plants (Corso et al. 2020, Corso et al. 2021, Alberghini et al. 2022, Boutet et al. 2022) and because the role of this MBW complex in controlling phenylpropanoid metabolism in seeds has been experimentally validated in several studies (see Xu et al. 2015 for a review). Hence, we regarded it as a relevant case study to illustrate how the outputs by Wimtrap can differ according to the tissue considered. We predicted the gene targets of TT2–TT8–TTG1 in seeds, roots and flowers of *A.thaliana*. Even though there were no TF ChIP data in seeds and roots, we could run our tool in these organs because we could get DGF- and DHS-predictive features and could transfer the TF-pooled model trained from seedlings. Using this rationale, six additional conditions were also included in our package for *A.thaliana* (non-hair part of the roots, heat-shocked seedlings, dark-grown seedlings, dark-grown seedlings exposed to 30 min of light, dark-grown seedlings exposed to 3 h of light and dark-grown seedlings exposed to a long day cycle) and one for *S.lycopersicum* (immature fruits).

To perform this analysis, we first predicted the gene targets TT2, TT8 and TTG1. We considered that the gene targets of a TF are the genes whose TSS is the closest to a potential binding site predicted as 'positive' using Wimtrap. We determined the best prediction score threshold to distinguish between 'positive' and 'negative' candidate gene targets based on the 28 TFs studied in *A.thaliana* seedlings. This best threshold was 0.86 on average.

The results highlighted a strong impact of the tissue on the type and number of potential TT2, TT8 and TTG1 gene targets (**Fig. 6A**, **B**). In addition, a higher number of potential targets was identified in seeds and roots for TT2 and TT8, compared with TTG1, while a similar number of targets among the MBW TFs was predicted in flowers (**Fig. 6A**). The Gene Ontology (GO) enrichment analyses revealed a higher number of enriched GO terms in seeds compared with roots and flowers (**Fig. 6B**). Finally, we focused on phenylpropanoid-related GO terms to evaluate whether our predictions could point towards a more significant association of TT2 with phenylpropanoid metabolism in seeds than in other organs, as expected according to the literature (Corso et al. 2021). A higher number of enriched GO terms associated with phenylpropanoids and flavonoids was identified for TT2 compared with TT8 and TTG1, with differences according to the tissue considered in the analyses (**Fig. 6C**). In the seed, four, two and one phenylpropanoid GO terms were identified for TT2, TT8 and TTG1, respectively. In roots, there were less enriched phenylpropanoid GO terms (or with a higher *P*-value) than in seeds for all three TFs. This held true in flowers for TT2 and TT8, while it was the opposite for TTG1, which had a higher number of enriched phenylpropanoid GO terms in flowers than in seeds. Our results agree with the role that is more specific to the seeds of TT2 in the accumulation of phenylpropanoids.

Discussion

An efficient approach to exploit and study genomic features at the location of TF-binding sites

The identification of the transcriptional targets of a TF by an approach based on pattern matching represents a major challenge. An important difficulty consists of building reference datasets. To date, there is still no consensual method to build a reference set of binding sites based on ChIP-seq data (Li et al. 2019). The identification of the ChIP peaks is dependent on the tool and the parameters that were used. Moreover, ChIP peaks do not allow the precise location of the binding sites and they report only stable interactions (Mundade et al. 2014). Another limitation inherent to our study comes from the epigenetic data. Due to their scarcity, we integrated data that were not perfectly fitting with the ChIP-seq data (from seedlings of different ages, grown in different conditions). In spite of that,

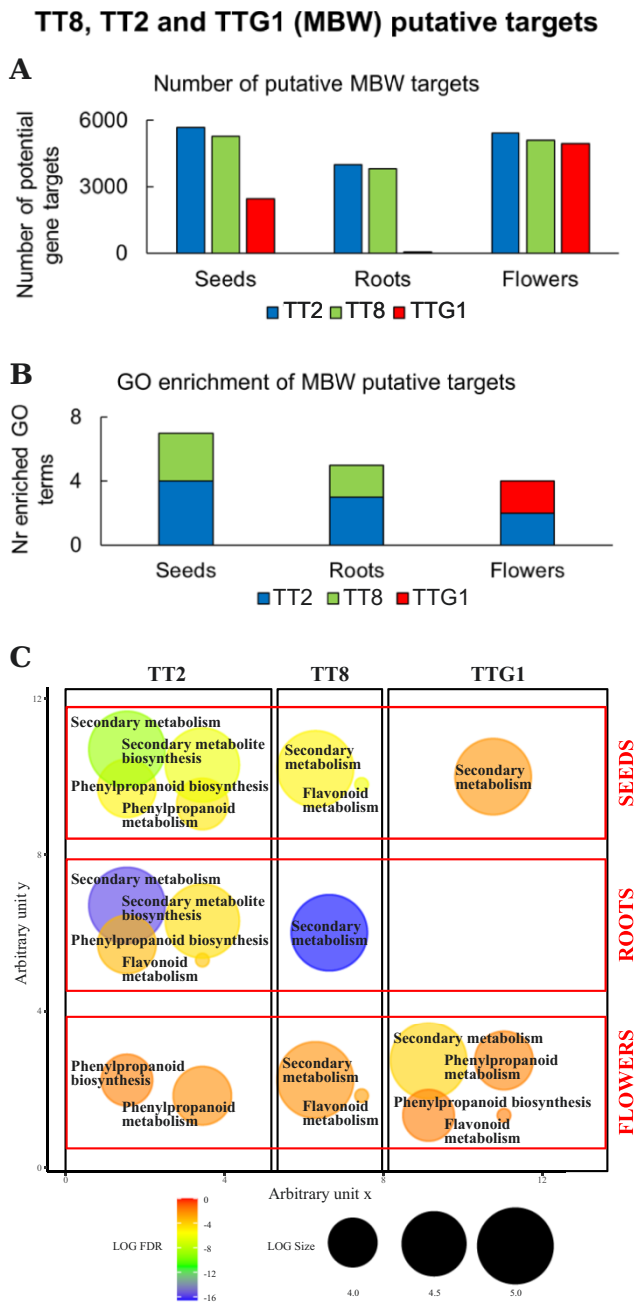

Fig. 6 Prediction of gene targets of components of the MBW complex and associated pathways (A) Number of putative targets predicted for TT2, TT8 and TTG1 using Wimtrap in seeds, roots and flowers. (B) Number of GO-enriched biological process terms among the putative targets of TT2, TT8 and TTG1 in the different organs considered. (C) Bubble chart representing the phenylpropanoid-related GO terms significantly enriched (FDR $\leq$ 0.05) in seeds, roots and flowers among the TT2, TT8 and TTG1 targets predicted in Arabidopsis. The GO terms are clustered according to the TF and the tissue considered. They are represented as circles whose diameter is proportional to the logarithm of their size (i.e. the total number of genes they annotate in the TAIR10 genome) and whose color follows a gradient scale related to the logarithm of their FDR (Fisher's enrichment test).

the results obtained with Wimtrap were consistent among the TFs considered.

We could assess in particular: (i) the predictivity of different layers of genomic features, (ii) the influence of the scale of the considered genomic regions and (iii) the generalization of the models.

Predictivity of layers of features. We obtained high performances of models when predicting TFBSs in Arabidopsis seedlings. Wimtrap highlighted the decisiveness of the features based on the DNase I-seq data [i.e. those related to DNase I-hypersensitive sites (DHS—open regions of DNA) and the digital genomic footprints (DGF)]. Compared with the histone modifications, the DHS present the important advantage of preserving their predictivity independently from the distance to the TSS. The high predictive power of the DHS can also be linked to their ability to identify both active and resting TFBSs (Zhu et al. 2015). They might therefore buffer variations related to the activity of enhancers and promoters across the integrated data, which were obtained from independent studies.

Despite being less predictive than the DHS (included in layer 5) and the DGF (layer 4), the features of layers 1–3, which are related to condition-independent features (results of pattern matching and 'phylogenetic' footprinting and position on the gene), were also shown to be very valuable for significantly improving the performances of pattern matching. Layers 1–3 are therefore 'time and cost-effective' as, in contrast to layer 5, they are already available for numerous plants.

The predictivity of the genomic data might vary according to their quality and/or the approach that was taken to generate them. This is well illustrated with the layer related to DNA sequence conservation, in which the dataset 'Phastcons' appeared less predictive than the 'Conserved Elements' dataset. Indeed, to allow sensitive detection of conserved elements, it is important to restrict the comparison to species that diverged relatively recently (Haudry et al. 2013), but the 'Phastcons' dataset was computed from a wide set of phylogenetically distant eudicots (Tian et al. 2020). This might make the identification of the conserved elements on enhancers very difficult as the divergence is an important source of phenotypic novelties on these *cis*-regulatory regions (Meireles-Filho and Stark 2009, Wittkopp and Kalay 2012).

One advantage of our approach is that it allows the automatic elaboration of decision rules that are more complex than simply retaining all the PWM matches that are located on a promoter or a conserved element, DHS or DGF. We found that the modeling was especially relevant to obtain good performances at predicting binding sites based solely on the condition-independent features of layers 1–3. To a lesser extent, we also found as that our method could improve the results of digital genomic footprinting by integrating features of layers 1–3 and 5.

Multiscale extraction of genomic features. The analysis at different scales of the genomic regions on which the potential binding sites are located (in 20, 400 and 1,000 bp windows) is a characteristic of our method. We obtained important gains in the prediction of potential of the features related to digital genomic footprinting, DNA sequence conservation, number of PWM matches and nucleosome positioning when considering the surrounding context of the potential binding sites and not only their 20 bp genomic location. These improvements might primarily come from the tendency of the TFs to be densely recruited on *cis*-regulatory regions (Aerts 2012, Pott and Lieb 2015), which can be identified from clusters of binding footprints, conserved elements or homotypic PWM matches. As for the special case of the nucleosome positioning data, we suggest that the overlap of a potential binding site with a nucleosome is not predictive because some nucleosomes can be easily moved to make *cis*-regulatory elements accessible (Collings et al. 2013, T. Zhang et al. 2015). However, the density of nucleosomes in the surrounding regions is important as TFs tend to target loosely packed regions of the chromatin.

Our approach allows some technical limitations to be overcome. For instance, evolutionarily conserved binding sites cannot be identified individually but only in clusters due to their short sequences (Haudry et al. 2013). Regarding the digital genomic footprints, it is known that they might be distant by >20 bp from the actual binding site (Neph et al. 2012, Gusmao et al. 2014).

Generalization of the models across TFs and organisms/conditions in plants. The generalizability of the predictive models across TFs and conditions in a given organism opens up a wide range of applications. The pre-existence of ChIP/chip-seq data related to the studied TFs and/or to the studied condition is not necessary. Nevertheless, we must point out that transferring models from one condition to another comes with a cost in terms of performance. This might be related, among others, to differences in quality between the genomic data obtained in the 'training' condition and those obtained in the 'studied' organism condition. We have also to point out that TF-pooled models of NAC50 and NAC52 perform substantially less well than TF-specific models. NAC50 and NAC52 bind the DNA on sites exhibiting a particular palindromic motif and might recruit a demethylase that will cause the silencing of the targeted genes (S. Zhang et al. 2015, Butel et al. 2017, van Rooijen et al. 2020). However, for NAC50 and NAC52, we could still demonstrate a positive association with the histone variant H2A.Z, representing a hallmark of *cis*-regulatory regions (Sequeira-Mendes et al. 2014).

Regarding the generalization of models across organisms, we obtained encouraging results, even though we need to remain cautious. When we transferred the models built from *S.lycopersicum* ripening fruits and *Z.mays* seedlings to *A.thaliana* seedlings, we obtained good performances, although lower than those achieved by the models built from *A thaliana* seedlings. On the other hand, we observed that the *O.sativa* model did not reach high AUC values when applied to *A.thaliana* seedlings. This might be related to a relatively low predictivity power of the DHS, DGF and cytosine methylation data obtained in *O.sativa* seedlings. We observed that the prediction performances of TFBSs in *O.sativa* seedlings were significantly enhanced when data on chromatin marks were added to those of DHS, DGF and cytosine methylation. Wimtrap can therefore help to select the best data available for a given organism and condition. However, further analyses will be needed to understand the differences in predictivity of features across organisms and conditions. This might be due to technical issues or species/condition specificities in the gene regulation mechanisms. In any case, this illustrates the importance of assessing the performances of a model to validate it. It is good practice to start with a model integrating the features of layers 1–3 in addition to the DGF and DHS (if available). If this clearly does not meet the expected AUC levels, then it is worth searching for additional chromatin state features.

A user-friendly and flexible tool

The user of Wimtrap can easily obtain TFBS and gene target predictions in any plant species for which genomic data of layers 1–3 are available and in any condition for which features of layers 4 and 5 can be obtained. Our approach can be fully reproduced with our R package and Shiny interface, with a great flexibility regarding the input data, pattern matching algorithm and machine learning technique. Wimtrap can also be used to compare other genomic regions than just the *cis*-regulatory elements (e.g. transgene/gene, enhancers/promoters, poised enhancers/active enhancers). In addition, pre-integrated models and databases allow the tools for hundreds of TFs to be immediately run for *A.thaliana*, not only in the seedlings and flowers but also in the whole roots, root hairs, seed coats and under several light treatments; for *S.lycopersicum*, not only in ripening fruits but also in immature fruits; and for *O.sativa* seedlings and *Z.mays* seedlings (Tian et al. 2020).

The performance of Wimtrap obviously depends on the genomic features which are provided to the models and, therefore, on the tools that were used to generate such data. When developing Wimtrap, we mainly focused on its flexibility in terms of input data as well as on its being user-friendly. We aimed at making the building of predictive models for new organisms/conditions easy, based on the available data. Here we did not directly compare Wimtrap with existing methods but compared the rationales implemented by a wide range of tools by assessing separately different layers of features. Other valuable resources can be used to predict TF target genes in plants, such as TEPIC 2 or ConsReg (Schmidt et al. 2019, Song et al. 2020). However, TEPIC 2 requires Linux operating systems and ConsReg requires expression data, which might be limiting.

Examples of application of Wimtrap

The activity and function of many TFs are specific to the plant organ and tissue, or to the condition considered (Franco-Zorrilla et al. 2014, Song et al. 2020). This is the case for some TFs belonging to the R2R3 MBW complex, which act synergistically to control the genes involved in the regulation of the late steps of flavonoid and proanthocyanidin biosynthesis and accumulation in seeds. More specifically, the MYB (TT2), bHLH (TT8) and WDR (TTG1) protein complex is active in Arabidopsis seeds, with TT2 and TT8 playing a major role in the complex and are the main TFs controlling flavonoid genes (Lepiniec et al. 2006, Xu et al. 2015, Corso et al. 2020).

As an example of the use of Wimtrap, we showed how novel insights into the biological functions of components of a TF complex can be obtained at the organ level. Compared with roots and flowers, a higher number of enriched GO categories specific to phenylpropanoid metabolism have been identified for TT2 and TT8 target genes in seeds, while no enrichment was observed for TTG1 targets. Previous work highlighted a major role for TT2 and TT8 in the regulation of flavonoid late biosynthetic genes in seeds (Xu et al. 2015). As for TTG1, while its participation in the MBW complex has been demonstrated, less information is available about its regulation and functions (Baudry et al. 2004, Quattrocchio et al. 2006). Hence, a key role for TT2 and TT8 in flavonoid regulation and the major impact of these TFs in seeds have been confirmed. The results obtained for TT2, TT8 and TTG1 highlighted a strong impact of the organ and/or the condition on the prediction of TF target genes (**Fig. 6**). This is an important aspect of Wimtrap.

In conclusion, we developed an effective approach to study the specificities of the plant *cis*-regulatory elements and made available a bioinformatic tool to improve the prediction of TFBSs, which comes with pre-built models for *A.thaliana*, *S.lycopersicum*, *O.sativa* and *Z.mays*. Prediction of potential TFBSs can also be useful for comparing TFBSs of homologous genes, for choosing mutation sites or for inferring potential regulators of co-regulated genes. One of the strengths of such an approach is that it can retrieve *cis*-regulatory elements that are overlooked by ChIP/chip-seq data, as they can only catch stable interactions (Mundade et al. 2014), while TF binding events are often transient (Li et al. 2019). The predictions might be especially relevant when they are confronted with expression data (Rister and Desplan 2010, Li et al. 2019).

In the near future, the advent of new technologies such as the ChIP-exo/ChIP-nexus and ATAC-seq will be beneficial. Peaks of the ChIP-exo/ChIP-nexus are narrower than the ChIP-chip/seq data and therefore allow more accurate identification of the location of binding sites (Welch et al. 2017). It will help us in particular to better decipher the proportion of TF binding events that are due to direct binding (on primary/alternative motifs) and indirect binding. As regards ATAC-seq, it is emerging as a cost-effective alternative to DNase I-seq (Karabacak Calviello et al. 2019). Relevant data about new organisms and/or conditions will soon become available.

Materials and Methods

Data

Data on *A.thaliana* seedlings and flowers, and *S.lycopersicum* ripening fruits were obtained from Arabidopsis RegNet (Heyndrickx et al. 2014), PlantRegMap/PlantTFDB (Jin et al. 2017, Tian et al. 2020), PlantDHS (Zhang et al. 2016), the Gene Expression Omnibus (Clough and Barrett 2016) and Ensembl Plants Biomart (Kinsella et al. 2011) databases. Additional information was retrieved from published articles (Gómez-Porras et al. 2007, Thomas et al. 2007, Baxter et al. 2012, Brandt et al. 2012, Haudry et al. 2013, Nuruzzaman et al. 2013, Fujisawa et al. 2014, Sequeira-Mendes et al. 2014, Zhiponova et al. 2014, Wang et al. 2015, Gaillochet et al. 2017, Ye et al. 2017) (**Supplementary Tables S1–S14**). The filters that were used to query Ensembl Plants Biomart and the Gene Expression Omnibus are described in **Supplementary text S1**. For each species considered, we downloaded the genome sequence and protein-coding transcript models (using the TAIR10 assembly for *A.thaliana*, SL3.0 for *S.lycopersicum*, IRGSP-1.0 for *O.sativa* L. ssp. Japonica and Zm-B73-REFERENCE-NAM-5.0 for *Z.mays* B73). In addition, we obtained 57 TF-ChIP-seq peak files (28 obtained in *A.thaliana* seedlings, 3 in *A.thaliana* flowers, 5 in *S.lycorpersicum* ripening fruits, 4 in *O sativa* L. ssp. Japonica seedlings and 17 in *Z.mays* B73), five sets of conserved elements (2 for *A.thaliana*, 1 for *S.lycopersicum*, 1 for *O.sativa* L. ssp. Japonica and 1 for *Z.mays* B73), five sets of DNase I-seq and BS-seq data (one each for *A.thaliana* seedlings, *A.thaliana* flowers, *S.lycopersicum* ripening fruits, *O.sativa* L. ssp. Japonica seedlings and *Z.mays* B73 seedlings), one partitioning of the genome between nine categories of chromatin stated (one for *A.thaliana* seedlings), two sets of H3K4me3, H3K4me3, H3K36me3, H3K27ac, H3K9ac, H4K12ac and H3K27me3 ChIP-seq data (one each for *A.thaliana* seedlings and *O.sativa* L. ssp. Japonica seedlings) and one set of MNase-seq, H2A.Z, H2BuB18, H3K4me1, H3K4me2, H3K9me2, H3K27me1, H3K14ac, H4K5ac, H3K18ac, H3K56ac, H3T3ph, H4K8ac and H4K16ac ChIP-seq data (for *A.thaliana* seedlings). Furthermore, we directly collected the motifs of 55 of the 57 TFs, either as a PWM or as a logo. Details about the source of the data, the experimental design as well as the data analysis pipeline are provided in **Supplementary Tables S1–S14**. In particular, for ChIP-seq data, the number of samples is comprised between 1 and 4 (2.1 on average $\pm$ 0.95 SD), and the false discovery rate (FDR) is between 10^{-2} and 10^{-5} (0.04 on average $\pm$ 0.02 SD).

Data pre-processing

PWMs. Relevant data were pre-processed to obtain the jaspar raw pfm format (Castro-Mondragon et al. 2022). PWMs could be obtained: (i) directly from the PlantTFDB database (Jin et al. 2017), (ii) by de novo discovery analysis of the ChIP-seq data using peak motifs (Thomas-Chollier et al. 2012) or (iii) by measuring the relative heights of the letters at each position of a consensus sequence or logo, using the arbitrary total count number of 1,000. TFs for which such pre-processing steps were necessary to obtain the PWM are specified in **Supplementary Tables S1, S4, S6, S9** and **S12**.

Gene structures. Basic manipulations using the R packages GenomicRanges (Lawrence et al. 2013) and rtracklayer (Lawrence et al. 2009) were required to obtain the location of the TSS, transcription termination sites (TTS), proximal promoters, promoters, 5′-UTRs, coding sequences (CDS), introns, 3′-UTRs and downstream regions in the BED format (Kent et al. 2002). For the gene structures, we used as input the text files downloaded from the Ensembl Plants Biomart following the procedure detailed in **Supplementary text 1**.

Conserved elements and chromatin states. The conserved non-coding sequences of *A.thaliana* identified by Thomas et al. (2007), Baxter et al. (2012) and Haudry et al. (2013) were merged by union and exported in BED format R using GenomicRanges (Lawrence et al. 2013) and rtracklayer (Lawrence

et al. 2009). The conserved elements of *A.thaliana* and *S.lycopersicum* along with their phastcons scores were downloaded from PlantRegMap as GTF files and directly used as such. The genome partition into nine chromatin states defined by Sequeira-Mendes et al. (2014) was encoded in BED files. Each region was annotated in the 'name' field by the chromatin state (from 'A' to 'I') and in the 'score' field by a dot to indicate to Wimtrap to extract a categorical feature.

ChIP/DNase/BS/MNase peaks. In the majority of cases, results of peak-calling analyses could be obtained from the Gene Expression Omnibus or supporting information of peer-reviewed articles, either in the BED format or in formats that could be easily converted to BED using R or awk (Aho et al. 1988). If applicable, peaks from replicates were then merged by union and the scores were summed on overlapping regions using GenomicRanges (Lawrence et al. 2013) and rtracklayer (Lawrence et al. 2009) R packages. In some cases, only data resulting from signal generation analysis were available. Such data consisted of UCSC tracks defining a signal (the fold change over control) along the genome. These formats were wig, bedGraph and bigWig (Kent et al. 2002). To generate BED files with the location and summit score of peaks based on data encoded in such formats, we applied the *sigWin* function of the CSAR R package (see the code provided in **Supplementary text 2**) (Muiño et al. 2011). The bigWig and bedGraph files needed to be converted to wig files first, with the bigWigToWig or bedGraphToWig UCSC program. The wig files allowed the partitioning of the genome into non-overlapping and scored genomic regions of equal length and equally spaced (=bins). Bins were filtered according to a minimum score threshold. For ChIP-seq data, it was a fold change of 1, except if this threshold resulted in such a high number of bins that it was impossible to load them into the R session. Then, a more stringent threshold was considered: the median of the fold changes. For cytosine methylation, a ratio of methylated cytosine of a minimum of 0.2 was considered. Once the bins were filtered, the scores of the overlapping bins between replicates were summed between replicates, and bins showing a gap <30 bp were subsequently merged. The resulting intervals were finally annotated with the score at the peak summit. Data that required pre-processing with sigWin are specified in **Supplementary Tables S3, S5, S8, S11** and **S14**.

DGFs. The location and scores of DGFs obtained with the footprinting2012 (Neph et al. 2012) tool for *A.thaliana* seedlings could be directly downloaded in BED format from PlantRegMap (Tian et al. 2020). Related data are encoded in BED files. For the *A.thaliana* flowers and *S.lycopersicum* ripening fruits, we reproduced the PlantRegMap analysis pipeline starting from the raw sequences of the reads generated by DNase-seq. The code used to obtain the DGFs for *A. thaliana* flowers is provided in **Supplementary text 3**.

Identification of candidate TF-binding sites

Candidate TFBSs were located by genome scanning against the PWMs using the *matchPWM* function of the Biostrings R package (Pagès et al. 2019). A 1 bp step sliding window was moved all along the genome. The length of the sliding window was set to the length of the considered PWM. At each step, the sequence of the sliding window was aligned to the PWM. Each nucleotide in the sequence was associated with its weight at its corresponding position in the PWM and the sum was operated over these weights. To calculate the *P*-values, we carried on an empirical assessment of the background probability density of the distribution of the match scores. This could be achieved based on random genomic regions due to the low prevalence of actual TFBSs. Sequences of 5,000 bp were thus randomly sampled at a rate of 200 bp by chromosome and were scanned at each base pair on both strands. The resulting match scores were ordered in increasing order and associated with their *P*-value, i.e. the proportion of matches with an equal or superior score.

Our pattern matching approach was compared with FIMO, a popular matching tool (Grant et al. 2011, Jayaram et al. 2016). Using the same *P*-value detection threshold of 10^{-3}, we found that 75% of the PWM matches detected using Wimtrap were also discovered by FIMO. Furthermore, a positive correlation of 0.77 (*P*-value < 2.2×10^{-16}) between the $\log_{10}$ of the *P*-values computed by the two methods was obtained (**Supplementary Fig. S9**). These considerations indicated the accuracy of our method.

Candidate TFBSs were defined as the PWM matches with a *P*-value $\geq 10^{-3}$. This threshold allowed the detection, for the 28 TFs related to *A.thaliana* seedlings, of the most prevalent ('primary') motif on two-thirds on average of the cognate ChIP peaks, which corresponds to previous observations (Heyndrickx et al. 2014) (**Supplementary Figs. S9, S10, Supplementary Table S15**).

Feature construction

Candidate TFBSs were annotated with five layers of features. Layer 1 included the *P*-value of the match score as well as the number of other homotypic matches, i.e. of matches against the same PWM as that of the candidate binding site, occurring at $\pm$ 200 bp and $\pm$ 500 bp from the center of the candidate TFBS. Layer 2 was relative to the position of the candidate TFBS on the gene. It encompassed the distance to the closest TSS and TTS but also as many features as there were gene structures. The structure found at the center of the considered candidate was associated with the score of '1'; the other structures were granted the score of '0'. In the case where several structures overlapped the same potential TFBS, only one structure was left with a score of '1', considering the following rule of preference: Proximal promoter > Promoter/downstream regions > Coding sequence > 5′-untranslated region > 3′-untranslated region > intron. Layers 3–5 included all the other data and were respectively associated with the sequence conservation, the DGF and the chromatin state/opening. Categorical features (cf. the partitioning of the genome of *A.thaliana* into nine functional chromatin states) were extracted by performing 'dummy variable encoding' to create as many variables as there were categories and by assigning the value of '1' to the categories overlapped by the center of the candidate TFBSs and 0 to the others. As for constructing features from 'numerical' data (scored genomic regions) and 'overlapping' data (non-scored genomic regions satisfying a given property)—which represented most of the data of layers 3–5—we calculated the base pair average of each considered features around the PWM matches, on three different scales: on windows of $\pm$ 10, 200 and 1,000 bp from the center of the candidate TFBSs. These represented, respectively, the scale of a *cis*-regulatory element, a ChIP peak and a promoter. Mathematically, our procedure of extraction can be described as follows. Let us consider the extracted data as an ensemble of n genomic regions each defined by their location $\{y_1, y_2, ..., y_n | \ y=$ (chromosome, start, end)$\}$ and by their scores $\{x_1, x_2, ..., x_n\}$ ($x_{1,..,n} = 1$ for overlapping features). Let be $\bar{x}$, the average score on the l bp window defined by the region $w =$ (chromosome, start, end). Considering that $\{z_1, z_2, ..., z_n\}$ are the length of the overlap of each region $\{y_1, y_2, ..., y_n\}$ with w:

$$\bar{x} = \sum_{i=1}^{n} \frac{x_i * \#z_i}{l}$$

The extracted features were scaled between 0 and 1 the features extracted from each TF (to allow the comparison of the same feature in different experiments, conditions or organisms).

Candidate TF-binding site labeling

The candidate binding sites for a given TF were labeled as 'positive', i.e. actual active *cis*-regulatory elements (in a considered condition), if they were overlapping a ChIP peak of the TF (in the condition considered). They were considered as 'negative' if they did not. The so-called 'target' feature was set to '1' for the candidates labeled as active *cis*-regulatory elements and '0' for the others. The length of the ChIP peaks was limited to $\pm$ 200 bp from the peak centers as most of the PWM matches were located in this interval (**Supplementary Table S15**).

Dataset balancing and splitting

Applying the steps described above allowed us to build a master dataset. This master dataset was at first balanced. For each TF, we randomly selected as many 'negative' potential candidate sites as there were 'positive' candidates, using the *sample.int* function of the base package in R, and removed those from the dataset. The selected 'negative' instances were kept in the master dataset and the others were removed. Balancing a dataset is a classical approach to overcome the tendency of binary classifiers to categorize all the instances into the most prevalent class (here, that of the 'negative' potential binding sites) when the minority class (the 'positive' potential binding sites) is rarely represented (Kotsiantis et al. 2006). The master dataset was then split into three TF-pooled datasets, according to the organism and the condition: *A.thaliana* seedlings, *A.thaliana* flowers and *S.lycopersicum* ripening fruits. These datasets were then subdivided into TF-specific datasets.

Machine learning

Models were obtained by machine learning to predict the label of candidate TF-binding sets. The machine learning step was preceded by a selection of the features to integrate in the models. This was based on the pairwise correlations between the features. If two features had a correlation >95%, the feature with the largest mean absolute correlation with the other features was removed. This feature selection was conducted with the caret R package (Kuhn 2020).

To select the algorithm of machine learning, we trained models based on each of the 28 TF-specific datasets generated from *A.thaliana* seedling data. The performances of the models were estimated using the 5-fold cross-validation strategy: each TF-specific dataset is cut into five smaller datasets of equal size. A model is trained with four of the five parts while the area under the ROC curve (AUC) is computed by applying the model on the remaining part. The process is repeated again four times so that each of the five parts is used for computing the AUC. The final AUC of a model is the mean of the five AUCs thus obtained. The AUCs were calculated with the pROC R package (Robin et al. 2011).

Initially, we evaluated algorithms of 'random forest', 'logistic regression' and 'gradient boosting'. Gradient boosting was clearly outcompeting (data not shown). We tested three different algorithms of gradient boosting: CatBoost (Prokhorenkova et al. 2019), LightGBM (Ke et al. 2017) and XGBoost (Chen and Guestrin 2016). The analyses were implemented with the respective packages in R (Dorogush et al. 2018, Chen et al. 2021, Shi et al. 2021). The following hyperparameters were set for all three algorithms: (maximum) depth of the tree = 6, learning rate = 0.3, number of iterations = 100, coefficient at the L2 regularization term of the cost function = 10, proportion of features used at each split selection = 1 and minimum instance in a leaf = 1. Parameters specific to each algorithm were set as follows: for CatBoost, number of split for numerical values = 64; for lightGBM, maximum number of leaves = 2^5 and number of threads = 2; for XGBoost, booster = tree and minimum loss reduction required to make a further partition on a leaf node of the tree = 0. All the other parameters are the default parameters. A mean AUC of 0.925 was achieved with CatBoost, and 0.927 with both lightGBM and XGBoost (**Supplementary Fig. S11**). We selected XGBoost as it is a method that has been well established for several years (Chen and Guestrin 2016). XGBoost is an algorithm which adds the predictions of an ensemble of regression trees. It builds the regression tree successively, each new tree being trained to predict the residuals, i.e. the deviation between the predicted values of the actual values, output by the former tree. Therefore, for a XGBoost model formed of K regression trees:

$$\widehat{y_i} = \Phi(x_i) = \sum_{k=1}^{K} f_k(x_i)$$

Where $\widehat{y_i}$ is the *i*th prediction, obtained by addition of the outputs of the K regression trees, based on the vector of features x_i. The regression trees are defined so that the regularized objective is minimized:

$$\mathcal{L}(\Phi) = \sum_i l(\widehat{y_i}, y_i) + \sum_k \Omega(f_k)$$

Where

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda ||w||^2$$

The function is the loss function which measures the difference between $\widehat{y_i}$, the *i*th prediction, and y_i, the actual *i*th value (= 1 if the *i*th instance is a 'positive' candidate TFBS, = 0 if is a negative one). T is the number of leaves in the tree f, λ the regularization parameter and w is a vector representing all the possible scores that can output f. Ω is a penalty function that allows avoiding overfitting.

Evaluation strategy

The performances of the models were assessed by computing the area under the ROC curve (AUC), which is a valid measure of the accuracy when computed from balanced datasets. Is A candidate TFBS is predicted as 'positive' or 'negative' according to whether its prediction score (output by a XGBmodel based on its annotations with the extracted features) is, respectively, superior or inferior to a certain threshold. The ROC curve plots the sensitivity and the 1–specifity obtained with increasing prediction score thresholds. The sensitivity is equal to TP/(TP + FN) and the specificity to TN/(TN + FP), where TP stands for 'true positive'—the total number of 'positive' candidates predicted as 'positive'—, FN for 'false negative'—the total number of 'positive' candidates predicted as 'negative'—, TN for 'true negative'—the total number of 'negative' candidates predicted as 'negative'—and FP for 'false positive'—the total number of negative candidate predicted as 'positive'. The higher the AUC, the more accurate a model is. An AUC of 1 corresponds to a perfect guess while an AUC of 0.5 corresponds to a random guess.

The performances of the TF-specific models were evaluated as described in the previous section. Models obtained from TF-pooled models were validated in a different way. Two procedures were possible. In the first case, models were built with all but one of the TFs of the dataset. The TF set aside was then used to compute the AUC. This allowed us to estimate the generalization of the models across TFs in a given organism/condition. In the second case, models were trained based on a TF-pooled dataset and were tested on another TF-pooled dataset. This allowed us to study the transferability of the models from one organism/condition to another.

Prediction of the targets of the MBW complex

For each of the 28 TFs studied in *A.thaliana* seedlings, all the protein-coding genes encoded in the genome of *A.thaliana* were annotated with the highest prediction score among their cognate predicted TFBSs. They were then labeled as 'positive' or 'negative' potential gene targets depending on whether their TSS was the closest or not to an occurrence of the motif of the TF on ChIP peaks. The optimal threshold to predict gene targets was determined using the *coords* function of the pROC R package, based on the ROC curves obtained with the 28 TFs studied by ChIP-seq in *A.thaliana* seedlings.

The potential gene targets of the MBW components in *A.thaliana* flowers were obtained with the TF-pooled model trained from the three TFs studied in *A.thaliana* flowers, based on all features of layers 1–4 and on the DHS. For running predictions in roots and seeds, we transferred to these organs the TF-pooled model trained from the 28 TFs studied in *A.thaliana* seedlings, also based on all the features of layers 1–4 in addition to the DHS (data about other features of layer 5 were not available in flowers, seeds and roots). For TT2 and TT8, we determined the genes whose TSS was the closest to an occurrence of their respective motifs (Jacob et al. 2021) with a Wimtrap prediction score ≥ 0.86. For TTG1, we determined the genes whose TSS was the closest of two neighboring motifs—one G-Box close to one AC-rich motif or one MYB motif—maximum distance between the two motifs = 30 bp (Xu et al. 2015)—both with prediction scores ≥ 0.86.

Supplementary Data

Supplementary data are available at *PCP* online.

Data Availability

Wimtrap can be downloaded from Github as a classical R package (https://github.com/RiviereQuentin/Wimtrap) or as a user-friendly R Shiny interface (https://github.com/RiviereQuentin/WimtrapWeb). It is fully documented by a manual, user guide and tutorial video (https://www.youtube.com/watch?v=6371fN7dkak). It allows reproduction of our approach to build new models for other conditions and/or organisms. The data underlying this article are available on GitHub (https://github.com/RiviereQuentin/carepat), as well as the R package (https://github.com/RiviereQuentin/Wimtrap) and R Shiny application (https://github.com/RiviereQuentin/WimtrapWeb).

Rivière_et_al.SuppTextS1-3&SuppFig1-11.pdf is available here (temporary link): https://owncloud.ulb.ac.be/index.php/s/PVGijICtXeTn1Bk.

Rivière_et_al.SuppTables1-14.xlsx is available here (temporary link): https://owncloud.ulb.ac.be/index.php/s/yxN0nT9DwJBQwwu.

Funding

Q.R. is a PhD fellow of the Fonds pour la Formation à la Recherche dans l'Industrie et l'Agronomie (F.R.I.A., Belgium) (references: FC:009155) and was grant-aided in the framework of the project of the Fonds de la Recherche Scientifique PDR T.0085.16. He is also the recipient of the Van Buuren–Jaumotte–Demoulin Prize. The work was published with the support of the Fondation Universitaire of Belgium.

Acknowledgements

We thank the Fonds de la Recherche Scientifique and the Fonds David et Alice Van Buuren for their financial support.

Disclosures

The authors have no conflicts of interest to declare.

References

Aerts, S. (2012) Computational strategies for the genome-wide identification of cis-regulatory elements and transcriptional targets. Curr. Top. Dev. Biol. 98: 121–145.

Aho, A.V., Kernighan, B.W. and Weinberger, P.J. (1988) The AWK Programming Language. Addison-Wesley Publishing Company, Boston.

Alberghini, B., Zanetti, F., Corso, M., Boutet, S., Lepiniec, L., Vecchi, A., et al. (2022) Camelina [*Camelina sativa* (L.) Crantz] seeds as a multi-purpose feedstock for bio-based applications. *Ind. Crops Prod.* 182: 114944.

Baudry, A., Heim, M.A., Dubreucq, B., Caboche, M., Weisshaar, B., Lepiniec, L., et al. (2004) TT2, TT8, and TTG1 synergistically specify the expression of BANYULS and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. *Plant J.* 39: 366–380.

Baxter, L., Jironkin, A., Hickman, R., Moore, J., Barrington, C., Krusche, P., et al. (2012) Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. *Plant Cell* 24: 3949–3965.

Behjati Ardakani, F., Schmidt, F. and Schulz, M.H. (2019) Predicting transcription factor binding using ensemble random forest models. *F1000Research* 7: 1603.

Bonev, B. and Cavalli, G. (2016) Organization and function of the 3D genome. *Nat. Rev. Genet.* 17: 661–678.

Boutet, S., Barreda, L., Perreau, F., Totozafy, J.-C., Mauve, C., Gakière, B., et al. (2022) Untargeted metabolomic analyses reveal the diversity and plasticity of the specialized metabolome in seeds of different *Camelina sativa* genotypes. *Plant J.* 110: 147–165.

Brandt, R., Salla-Martret, M., Bou-Torrent, J., Musielak, T., Stahl, M., Lanz, C., et al. (2012) Genome-wide binding-site analysis of REVOLUTA reveals a link between leaf patterning and light-mediated growth responses: *REVOLUTA ChIP-Seq Analysis*. *Plant J.* 72: 31–42.

Butel, N., Le Masson, I., Bouteiller, N., Vaucheret, H. and Elmayan, T. (2017) sgs1: a neomorphic nac52 allele impairing post-transcriptional gene silencing through SGS3 downregulation. *Plant J.* 90: 505–519.

Castro-Mondragon, J.A., Riudavets-Puig, R., Rauluseviciute, I., Berhanu Lemma, R., Turchi, L., Blanc-Mathieu, R., et al. (2022) JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* 50: D165–D173.

Chen, T., Tong, H., Michael, B., Vadim, K., Yuan, T., Hyunsu, C., et al. (2021) Xgboost: Extreme Gradient Boosting. https://CRAN.R-project.org/package=xgboost (September 19, 2019, date last accessed).

Chen, T. and Guestrin, C. (2016) XGBoost: a scalable tree boosting system. *In* Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD'16. the 22nd ACM SIGKDD International Conference. ACM Press, San Francisco, CA, pp. 785–794.

Chen, X., Yu, B., Carriero, N., Silva, C. and Bonneau, R. (2017) Mocap: large-scale inference of transcription factor binding sites from chromatin accessibility. *Nucleic Acids Res.* 45: 4315–4329.

Chen, T., Tong, H., Michael, B., Vadim, K., Yuan, T., Hyunsu, C., et al. (2017) LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *In Advances in Neural Information Processing Systems 30 (NIP 2017)*. https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/ (September 19, 2019, date last accessed).

Clough, E. and Barrett, T. (2016) The gene expression omnibus database. *In* Statistical Genomics. Edited by Mathé, E. and Davis, S. pp. 93–110. Springer New York, New York.

Collings, C.K., Waddell, P.J. and Anderson, J.N. (2013) Effects of DNA methylation on nucleosome stability. *Nucleic Acids Res.* 41: 2918–2931.

Corso, M., Perreau, F., Mouille, G. and Lepiniec, L. (2021) Specialized metabolites in seeds. *Adv. Bot. Res.* 98: 35–70.

Corso, M., Perreau, F., Mouille, G. and Lepiniec, L. (2020) Specialized phenolic compounds in seeds: structures, functions, and regulations. *Plant Sci.* 296: 110471.

Dorogush, A.V., Ershov, V. and Gulin, A. (2018) CatBoost: Gradient Boosting with Categorical Features Support. *CoRR*, abs/1810.11363. http://arxiv.org/abs/1810.11363 (November 13, 2021, date last accessed).

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

Fuda, N.J., Ardehali, M.B. and Lis, J.T. (2009) Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature* 461: 186–192.

Fujisawa, M., Shima, Y., Nakagawa, H., Kitagawa, M., Kimbara, J., Nakano, T., et al. (2014) Transcriptional regulation of fruit ripening by tomato

FRUITFULL homologs and associated MADS box proteins. *Plant Cell* 26: 89–101.

Gaillochet, C., Stiehl, T., Wenzl, C., Ripoll, J.-J., Bailey-Steinitz, L.J., Li, L., et al. (2017) Control of plant cell fate transitions by transcriptional and hormonal signals. *eLife* 6: e30135.

Gómez-Porras, J.L., Riaño-Pachón, D.M., Dreyer, I., Mayer, J.E. and Mueller-Roeber, B. (2007) Genome-wide analysis of ABA-responsive elements ABRE and CE3 reveals divergent patterns in Arabidopsis and rice. *BMC Genomics* 8: 260.

Grant, C.E., Bailey, T.L. and Noble, W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinformatics* 27: 1017–1018.

Gusmao, E.G. Allhoff, M., Zenke, M. and Costa, I.G. (2016) Analysis of computational footprinting methods for DNase sequencing experiments. *Nat. Methods* 13: 303–309.

Gusmao, E.G. Dieterich, C., Zenke, M. and Costa, I.G. (2014) Detection of active transcription factor binding sites with the combination of DNase hypersensitivity and histone modifications. *Bioinformatics* 30: 3143–3151.

Hardison, R.C. and Taylor, J. (2012) Genomic approaches towards finding cis-regulatory modules in animals. *Nat. Rev. Genet.* 13: 469–483.

Haudry, A., Platts, A.E., Vello, E., Hoen, D.R., Leclercq, M., Williamson, R.J., et al. (2013) An Atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat. Genet.* 45: 891–898.

Heyndrickx, K.S., de Velde, J.V., Wang, C., Weigel, D. and Vandepoele, K. (2014) A functional and evolutionary perspective on transcription factor binding in *Arabidopsis thaliana*. *Plant Cell* 26: 3894–3910.

Jacob, P., Brisou, G., Dalmais, M., Thévenin, J., van der Wal, F., Latrasse, D., et al. (2021) The seed development factors TT2 and MYB5 regulate heat stress response in Arabidopsis. *Genes* 12: 746.

Jankowski, A., Tiuryn, J. and Prabhakar, S. (2016) Romulus: robust multistate identification of transcription factor binding sites from DNase-Seq data. *Bioinformatics* 32: 2419–2426.

Jayaram, N., Usvyat, D. and Martin, A.C.R. (2016) Evaluating tools for transcription factor binding site prediction. *BMC Bioinformatics* 17: 547.

Jin, J., Tian, F., Yang, D.-C., Meng, Y.-Q., Kong, L., Luo, J., et al. (2017) PlantTFDB 4.0: toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* 45: D1040–D1045.

Jones, P.A. (2012) Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* 13: 484–492.

Karabacak Calviello, A., Hirsekorn, A., Wurmus, R., Yusuf, D. and Ohler, U. (2019) Reproducible inference of transcription factor footprints in ATAC-Seq and DNase-Seq datasets using protocol-specific bias modeling. *Genome Biol.* 20: 42.

Keilwagen, J., Posch, S. and Grau, J. (2019) Accurate prediction of cell type-specific transcription factor binding. *Genome Biol.* 20: 9.

Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., et al. (2002) The human genome browser at UCSC. *Genome Res.* 12: 996–1006.

Kinsella, R.J., Kahari, A., Haider, S., Zamora, J., Proctor, G., Spudich, G., et al. (2011) Ensembl BioMarts: a hub for data retrieval across taxonomic space. *Database* 2011: bar030.

Kotsiantis, S., Dimitris, K. and Pintelas, P. (2006) Handling imbalanced datasets: a review. *GESTS Int. Trans. Comput. Sci. Eng.* 30: 1–13.

Kuhn, M. (2020) Caret: Classification and Regression Training. https://CRAN.R-project.org/package=caret (November 18, 2021, date last accessed).

Kumar, S. and Bucher, P. (2016) Predicting transcription factor site occupancy using DNA sequence intrinsic and cell-type specific chromatin features. *BMC Bioinform.* 17: S4.

Lai, X., Stigliani, A., Vachon, G., Carles, C., Smaczniak, C., Zubieta, C., et al. (2019) Building transcription factor binding site models to understand gene regulation in plants. *Mol. Plant* 12: 743–763.

Lawrence, M., Daujat, S. and Schneider, R. (2016) Lateral thinking: how histone modifications regulate gene expression. *Trends Genet.* 32: 42–56.

Lawrence, M., Gentleman, R. and Carey, V. (2009) Rtracklayer: an R package for interfacing with genome browsers. *Bioinformatics* 25: 1841–1842.

Lawrence, M., Huber, W., Pagès, H., Aboyoun, P., Carlson, M., Gentleman, R., et al. (2013) Software for computing and annotating genomic ranges. *PLoS Comput. Biol.* 9: e1003118.

Lepiniec, L., Debeaujon, L., Routaboul, J. M., Baudry, A., Pourcel, L., Nesi, N., et al. (2006) Genetics and biochemistry of seed flovonoids. *Annu. Rev. Plant Biol.* 57(1): 405–430.

Lee, D.J., Minchin, S.D. and Busby, S.J.W. (2012) Activating transcription in bacteria. *Annu. Rev. Microbiol.* 66: 125–152.

Lenhard, B., Sandelin, A. and Carninci, P. (2012) Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.* 13: 233–245.

Li, H. and Guan, Y. (2019) Leopard: fast decoding cell type-specific transcription factor binding landscape at single-nucleotide resolution. *Bioinformatics* 31: 721–731.

Li, H., Quang, D. and Guan, Y. (2019) Anchor: trans-cell type prediction of transcription factor binding sites. *Genome Res.* 29: 281–292.

Liu, S., Zibetti, C., Wan, J., Wang, G., Blackshaw, S., Qian, J., et al. (2017) Assessing the model transferability for prediction of transcription factor binding sites based on chromatin accessibility. *BMC Bioinform.* 18: 355.

Meireles-Filho, A.C. and Stark, A. (2009) Comparative genomics of gene regulation—conservation and divergence of cis-regulatory information. *Curr. Opin. Genet. Dev.* 19: 565–570.

Meyer, C.A. and Liu, X.S. (2014) Identifying and mitigating bias in next-generation sequencing methods for chromatin biology. *Nat. Rev. Genet.* 15: 709–721.

Muiño, J.M., Kaufmann, K., van Ham, R.C., Angenent, G.C. and Krajewski, P. (2011) ChIP-Seq analysis in R (CSAR): an R package for the statistical detection of protein-bound genomic regions. *Plant Methods* 7: 11.

Mundade, R., Ozer, H.G., Wei, H., Prabhu, L. and Lu, T. (2014) Role of ChIP-Seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. *Cell Cycle* 13: 2847–2852.

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., et al. (2012) An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489: 83–90.

Nuruzzaman, M., Sharoni, A.M. and Kikuchi, S. (2013) Roles of NAC transcription factors in the regulation of biotic and abiotic stress responses in plants. *Front. Microbiol.* 4: 248.

Pagès, H., Aboyoun, P., Gentleman, R. and DebRoy, S. (2019) Biostrings: efficient manipulation of biological strings. *R Package Version 2.54.0*.

Pott, S. and Lieb, J.D. (2015) What are super-enhancers? *Nat. Genet.* 47: 8–12.

Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A.V. and Gulin, A. (2019) CatBoost: unbiased boosting with categorical features.

Qin, Q., Feng, J. and Ioshikhes, I. (2017) Imputation for transcription factor binding predictions based on deep learning. *PLoS Comput. Biol.* 13: e1005403.

Quang, D. and Xie, X. (2017) FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. *Genomics* 166: 40–47.

Quattrocchio, F., Baudry, A., Lepiniec, L., Grotewold, E. (2006) The Regulation of Flavonoid Biosynthesis. *In* Edited by Grotewold, E. pp. 97–122. Springer New York, New.

Rister, J. and Desplan, C. (2010) Deciphering the genome's regulatory code: the many languages of DNA. *BioEssays* 32: 381–384.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., et al. (2011) PROC: an open-source package for R and S+ to analyze and compare ROC curves. *BMC Bioinformatics* 12: 77.

Schmidt, F., Gasparoni, N., Gasparoni, G., Gianmoena, K., Cadenas, C., Polansky, J.K., et al. (2017) Combining transcription factor binding affinities with open-chromatin data for accurate gene expression prediction. *Nucleic Acids Res.* 45: 54–66.

Schmidt, F., Kern, F., Ebert, P., Baumgarten, N. and Schulz, M.H. (2019) TEPIC 2—an extended framework for transcription factor binding prediction and integrative epigenomic analysis. *Bioinformatics* 35:nscription factor binding pr 1608–1609.

Sequeira-Mendes, J., Aragüez, I., Peiró, R., Mendez-Giraldez, R., Zhang, X., Jacobsen, S.E., et al. (2014) The functional topography of the *Arabidopsis* genome is organized in a reduced number of linear motifs of chromatin states. *Plant Cell* 26: 2351–2366.

Shi, Y., Ke, G., Soukhavong, D., Lamb, J., Meng, Q., Finley, T., et al. (2021) Lightgbm: Light Gradient Boosting Machine. https://CRAN.R-project.org/package=lightgbm (November 12, 2021, date last accessed).

Siepel, A. and Haussler, D. (2005) Phylogenetic hidden markov models. *In* Statistical Methods in Molecular Evolution. Statistics for Biology and Health. pp. 325–351. Springer-Verlag, New York.

Song, Q., Lee, J., Akter, S., Rogers, M., Grene, R., Li, S., et al. (2020) Prediction of condition-specific regulatory genes using machine learning. *Nucleic Acids Res.* 48: e62.

Spitz, F. and Furlong, E.E.M. (2012) Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.* 13: 613–626.

Thomas, B.C., Rapaka, L., Lyons, E., Pedersen, B. and Freeling, M. (2007) Arabidopsis intragenomic conserved noncoding sequence. *Proc. Natl. Acad. Sci. USA* 104: 3348–3353.

Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D., van Helden, J., et al. (2012) RSAT peak-motifs: motif analysis in full-size ChIP-Seq datasets. *Nucleic Acids Res.* 40: e31.

Tian, F., Yang, D.-C., Meng, Y.-Q., Jin, J. and Gao, G. (2020) PlantRegMap: charting functional regulatory maps in plants. *Nucleic Acids Res.* 48: D1104–D1113.

van Rooijen, R., Schulze, S., Petzsch, P. and Westhoff, P. (2020) Targeted misexpression of NAC052, acting in H3K4 demethylation, alters leaf morphological and anatomical traits in *Arabidopsis thaliana*. *J. Exp. Bot.* 71: 1434–1448.

Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A. and Luscombe, N.M. (2009) A census of human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* 10: 252–263.

Veljkovic, J. and Hansen, U. (2004) Lineage-specific and ubiquitous biological roles of the mammalian transcription factor LSF. *Gene* 343: 23–40.

Vuong, P. and Misr, R. (2011) Guide to genome-wide bacterial transcription factor binding site prediction using OmpR as model. *In* Selected Works in Bioinformatics. Edited by Xia, X. InTech: 41–56.

Wang, C., Liu, C., Roqueiro, D., Grimm, D., Schwab, R., Becker, C., et al. (2015) Genome-wide analysis of local chromatin packing in *Arabidopsis thaliana*. *Genome Res.* 25: 246–256.

Welch, R., Chung, D., Grass, J., Landick, R. and Keleş, S. (2017) Data exploration, quality control and statistical analysis of ChIP-Exo/Nexus experiments. *Nucleic Acids Res.* 45: e145.

Wittkopp, P.J. and Kalay, G. (2012) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat. Rev. Genet.* 13: 59–69.

Xu, W., Dubos, C. and Lepiniec, L. (2015) Transcriptional control of flavonoid biosynthesis by MYB–BHLH–WDR complexes. *Trends Plant Sci.* 20: 176–185.

Ye, H., Liu, S., Tang, B., Chen, J., Xie, Z., Nolan, T.M., et al. (2017) RD26 mediates crosstalk between drought and brassinosteroid signalling pathways. *Nat. Commun.* 8: 14573.

Zhang, S., Zhou, B., Kang, Y., Cui, X., Liu, A., Deleris, A., et al. (2015) C-terminal domains of histone demethylase JMJ14 interact with a pair of NAC transcription factors to mediate specific chromatin association. *Cell Discov.* 1: 15003.

Zhang, T., Zhang, W. and Jiang, J. (2015) Genome-wide nucleosome occupancy and positioning and their impact on gene expression and evolution in plants. *Plant Physiol.* 168: 1406–1416.

Zhang, T., Marand, A.P. and Jiang, J. (2016) PlantDHS: a database for DNase I hypersensitive sites in plants. *Nucleic Acids Res.* 44: D1148–D1153.

Zhiponova, M.K., Morohashi, K., Vanhoutte, I., Machemer-Noonan, K., Revalska, M., Van Montagu, M., et al. (2014) Helix–loop–helix/basic helix–loop–helix transcription factor network represses cell elongation in arabidopsis through an apparent incoherent feed-forward loop. *Proc. Natl. Acad. Sci. USA* 111: 2824–2829.

Zhu, B., Zhang, W., Zhang, T., Liu, B. and Jiang, J. (2015) Genome-wide prediction and validation of intergenic enhancers in arabidopsis using open chromatin signatures. *Plant Cell* 27: 2415–2426.