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

Modelling and Predicting Plant TF Binding Sites

Corresponding authors: M. Defrance, Interuniversity Institute of Bioinformatics in Brussels, Machine Learning Group, Université libre de Bruxelles, 1050 Brussels, Belgium; Q. Rivière, Brussels Bioengineering School, Laboratory of Plant Physiology and molecular Genetics, Université libre de Bruxelles, 1050 Brussels, Belgium.

Subject areas: regulation of gene expression, new methodology

0 black and white figure, 6 colour figures, 0 table, 3 supplementary texts, 11 supplementary figures, 14 supplementary tables

Supplementary data are available online

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

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

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

Modelling and Predicting Plant TF Binding Sites

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

[1]Brussels Bioengineering School, Laboratory of Plant Physiology and molecular Genetics, Université Libre de Bruxelles, 1050 Brussels, Belgium

[2]Université Paris-Saclay, INRAE, AgroParisTech, Institut Jean-Pierre Bourgin (IJPB), 78000, Versailles, France

[3]Interuniversity Institute of Bioinformatics in Brussels, Machine Learning Group, Université Libre de Bruxelles, 1050 Brussels, Belgium

[4]Functional and Evolutionary Entomology, Gembloux Agro-Bio Tech, University of Liège, Passage des Déportés 2, 5030 Gembloux, Belgium
† These authors co-directed the work
***Correspondance:** Matthieu Defrance, matthieu.defrance@ulb.be; Quentin Rivière, qriviere@ulb.be.

***Correspondence:**
Matthieu Defrance
matthieu.defrance@ulb.be
Quentin Rivière
qriviere@ulb.be

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

Modelling and Predicting Plant TF Binding Sites

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

[1]Brussels Bioengineering School, Laboratory of Plant Physiology and molecular Genetics, Université Libre de Bruxelles, 1050 Brussels, Belgium

[2]Université Paris-Saclay, INRAE, AgroParisTech, Institut Jean-Pierre Bourgin (IJPB), 78000, Versailles, France

[3]Interuniversity Institute of Bioinformatics in Brussels, Machine Learning Group, Université Libre de Bruxelles, 1050 Brussels, Belgium

[4] Functional and Evolutionary Entomology, Gembloux Agro-Bio Tech, University of Liège, Passage des Déportés 2, 5030 Gembloux, Belgium
✝ These authors co-directed the work
***Correspondance:** Matthieu Defrance, matthieu.defrance@ulb.be; Quentin Rivière, qriviere@ulb.be.

Abstract

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 performances 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 to learn decision rule models 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 cis-regulatory element 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 TF 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 a 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, genomics, predictive modelling, R package, Plant organs, flavonoid target-genes

1 Introduction

Gene regulation is one of the most fundamental biological phenomena. It explains how, from the same genetic code, a cell can harbour different states, according to the cell cycles and the signals from the environment. For multi-cellular organisms as plants, gene regulation is also involved in processes such as cell specialization, organogenesis, growth, and ageing (Aerts, 2012; Spitz and Furlong, 2012). Gene regulation encompasses a cascade of regulatory processes that intervene all along with 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 at 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 attractive enough, but in most cases, the recruitment of the RNA-polymerase involves interactions with components that are called transcription factors and cofactors (Fuda *et al.*, 2009).

Transcription factors (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 the TSS, nearby the binding sites of the RNA-polymerase, and are classified as repressors or activators depending on whether they favour the recruitment of the subunits of the polymerase or block it (Lee *et al.*, 2012). However, the mechanisms of action of the transcription are complex. Organisms pack the DNA in highly condensed structures, called 'chromatin', that allow fitting with the space of the cell (prokaryotes) or the nucleus (eukaryotes), which makes it difficult for regulatory molecules to access and bind to DNA. The action of some TFs consists therefore in 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 recognised by the TF of interest. To identify TF-target genes several challenges have to be addressed: (1) identifying and modelling (as 'motifs') the sequence specificities of TF binding sites, and (2) 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 performant 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 TF-binding sites is highly variable and too short compared to 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 is performed the pattern-matching, 'cluster' and 'phylogenetic' footprinting methods can be used, as TFs tend (a) to cluster (cis-regulatory regions show intervals with a high density of binding sites of the same and/or of different TFs), (b) to bind to sites (or clusters) that are evolutionary 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-scale have emerged. In particular, ChIP-chip/seq 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 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 DNAseI-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 to directly probe the degree of opening of the DNA (Meyer and Liu, 2014).

The greater availability of genomic and epigenomic data paved the way for new bioinformatic methods dedicated to the prediction of TFs binding sites. An expanding number of tools have been released (Gusmao *et al.*, 2016; Jankowski *et al.*, 2016; Kumar and Bucher, 2016; X. Chen *et al.*, 2017; Schmidt *et al.*, 2017; Schmidt *et al.*, 2017; Qin and Feng, 2017; Quang and Xie, 2017; Liu *et al.*, 2017; Li *et al.*, 2019; Behjati Ardakani *et al.*, 2019; Li and Guan, 2019; Keilwagen *et al.*, 2019). Of particular interest is the new footprinting approach, called the 'digital genomic' footprinting (DGF), which is based on the property of the TFs to protect the cis-regulatory elements from cleavage by the DNAseI. Contrary to 'cluster' and 'phylogenetic' footprinting techniques, the DGF takes into account the chromatin state dynamics and therefore that of 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 to new ones 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-seq/chip 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 *Arabidopsis thaliana* in 10 conditions (organs or growing conditions), *Solanum lycopersicum* in two conditions, and *Oryza sativa*, and *Zea mays* in one condition.

2 Results

2.1 Analysis overview

We developed a machine-learning approach (Figure 1) to predict cis-regulatory elements and TF target genes using information obtained from TFBS motifs, 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 were performed and the workflow can be schematically described as follows. Based on literature search and the query of 7 specialized databases, we retrieved:

- the genomic sequences and the transcript models of *A. thaliana*, *S. lycopersicum*, *O. sativa*, and *Z. mays*;
- 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*;
- 5 genomic maps of cis-regulatory elements (2 in *A. thaliana*, 1 in *S. lycopersicum*, 1 in *O. sativa,* 1 in *Z. mays*);
- 5 genomic maps of digital genomic footprints (DGFs) (1 in *A. thaliana* seedlings, 1 in *A. thaliana* flowers, 1 in *S. lycopersicum* seedlings, 1 in *O. sativa* seedlings, 1 in *Z. mays* seedlings);
- 42 chromatin feature-peak data (24 in *A. thaliana* seedlings, 3 in *A. thaliana* flowers, 3 in *S. lycopersicum* seedlings, 9 in *O. sativa* seedlings, 3 in *Z. mays* seedlings);

These data and information were then integrated into the Wimtrap pipeline, composed of several steps (Figure 1):

Step 1: Candidate TF binding sites location. Pattern-matching analyses were carried out with the motifs of the TFs to obtain the location of the potential binding sites. Each potential binding site was scored according to the fit of the DNA sequence with the motif.

Step 2: Candidate TF binding sites annotation. The potential binding sites were annotated with features characterizing their genomic context. The distance of the potential binding sites to the closest transcript was calculated using the TSS as reference. The structure (promoter, coding sequence, …) overlapped by the potential 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 ± 10 bp, ± 200 bp, or ±500bp around the potential cis-regulatory elements.

Step 3: Candidate TF binding site labelling and dataset balancing. The potential binding sites were labelled as 'positive' when they were validated by available ChIP-chip/seq data and as 'negative' if not. To avoid an overrepresentation of the negative potential binding sites compared to the positive ones, a subset of negative potential binding sites was randomly selected so that the composition of the dataset turned to 50% of negative and 50% of positive binding sites. 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' potential binding sites) when the minority class (the 'positive' potential binding sites) is rarely represented (Sotiris Kotsiantis *et al.*, 2006).

Step 4: Modelling of binary classifiers of candidate TF binding sites. 'Decision-rule' models, made of a collection of regression trees, were trained by extreme gradient boosting (Chen and Guestrin, 2016). Such models allowed to decide whether a candidate TF binding site is 'positive' or 'negative' based on the integrated features. Two kinds of models were built: the TF-specific ones, based on data from a single TF, and the TF-pooled ones, 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: Evaluation of the accuracy of the binary classifiers. Model accuracies were assessed by computing the area under the ROC (Receiver Operating Characteristic) curve (AUC). For TF-specific models, we proceeded to the 5-cross validation protocol: each TF-specific dataset was split into 5 sub-datasets. Training and AUC computation were iterated 5 times, using each time 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.

2.2 Performances of TF-specific models according to the integrated features

Based on the 28 TFs studied by ChIP-seq in *Arabidopsis thaliana* seedlings, we computed the ROC curves of TF-specific models trained with different groups of features, taken individually or in combination (Figure 2). These groups of features were called 'layers', as they represented distinct layers of information that could be added to each other's. The layers are the following: (1) Motif occurrences and scores, (2) Position related to the transcript model, (3) DNA sequence conservation, (4) Digital genomic footprint (DGF) occurrence and score, and (5) Chromatin state.

Each layer corresponds to a given technique. The 1st layer is related to the 'cluster' footprinting; the 2nd to the tendency of the TFs to be located on the promoters in proximity to the TSS; the 3rd, to the 'phylogenetic' footprinting; the 4th, to the 'digital genomic' footprinting; and the 5th, 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 features. These associations can be visualized in Supplementary figures 1-5, in where Pearson's correlation between the features and the binarized label of the potential binding sites (equal to 1 when a potential binding site is 'positive', and 0 when it is 'negative') is plotted.

2.2.1 Layer 1: Motif occurrence and score. Layer 1 allowed assessing the pattern-matching and the 'cluster' footprinting method as it includes the p-value of the PWM 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 to an average AUC of 0.60 (Figure 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 figure 1). The p-values of the PWM matches exhibited low predictive levels, except for the TFs NAC50 and NAC52.

2.2.2 Layer 2: Position on the gene. Layer 2 allowed evaluating the rationale behind promoter scanning. Models integrating the results of pattern-matching with the position on the gene (structure, distance to closest transcription start site) reached on average an AUC of 0.73 (Figure 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 (Supplementary figure 2). The chance for a PWM match to be a cis-regulatory element increased while getting upstream closer to the TSS but suddenly dropped at several bp downstream from the TSS. Overall, 49% of the cis-regulatory elements were located on the promoter at maximum – 2 000 bp from the TSS, while 9% were located on the 5'UTR (Figure 4). The 42% remaining cis-regulatory-elements were distributed as follows: (1) 18% in the gene body, downstream of the 5'UTR (i.e. coding sequence, intron, 3'UTR), (2) 8% in the regions downstream to the transcript stop site, and (3) 16% in the intergenic regions.

2.2.3 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 analysed the homeologs in *A. thaliana* and the orthologs in the eudicots and the family of the Brassicaceae. The second dataset is composed of 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 figure 3). 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 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 in average).

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

2.2.5 Layer 5: Chromatin state. The integration of 23 chromatin state-related features to the results of pattern-matching led to models with an average AUC of 0.91 (Figure 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' ones. 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 the heterochromatin (Supplementary figure 5). When assessing more in details the individual variables characterizing the chromatin state, the 8 features the most associated with cis-regulatory elements were, by decreasing order of association: the DNase-I hypersensitivity score (DHS – measure of the opening of DNA), the H3K4me1 histone mark, the methylation of the cytosine, the nucleosomes density and the H3K27me1, H3K9me2, H3K56ac, H2BuB, and H3K18ac histone marks. TFs showed overall homogeneous patterns. However, for 4 of them, several important features were not associated to cis-regulatory elements. This was the case of 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 nucleosomes 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 figure 6). There were 5 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 for H3K18ac and H3K27me3, a striking point came from that cis-regulatory elements were associated with high or low levels depending on whether the regions were distal or proximal to the TSS (< -2500bp for H3K27me3, < -5000bp for H3K18ac).

2.2.6. Combination of layers. The combination of the layers 1, 2 and 3, conditions-independent, 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.

2.2.7. Restriction of the 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 DHS was the feature the most associated with the cis-regulatory elements in the layer 5 (Supplementary figure 7). Models based only on pattern-matching and DHS showed an average AUC of 0.86 (Figure 2D). Adding the sole layer 4 to these models led to an average AUC of 0.88, while adding the layer 4 together with the layers 1, 2, 3 led to an average AUC of 0.91 (Figure 2D).

2.3 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 “2.1 Analysis overview”) based on the whole set of features (layers 1 to 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 (Chen and Guestrin, 2016). The DGF (layer 4), associated to an average gain of 42%, appeared as the most important feature in the TF-specific models for all the TFs, except *PRR7*, *PIF3*, *NAC50* and *NAC52* (Figure 3). In PRR7- and PIF3-models, the most important feature was the DHS; in NAC50-model, the H3K4me1 histone mark; and in NAC52-model, the H2A.Z variant. The other features got in average less than 10% of gain. The most important features following the DGF were, by 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). We observed some important variations across the TF-specific models in terms of importance of those features. For instance, while the gain for the p-value of the PWM match score was 3% in average, it raised to 8, 14 and 10% in the models of GBF3, NAC52 and NAC50, respectively. The features related to the layer 2 (position on the gene) were the least important features of the models for all TFs.

2.4 Transferability of TF-pooled models

To evaluate the generalization of Wimtrap, we trained general models by pooling data related to all TFs but one and evaluated the performances on the TF that was leftover. We then compared for each TF the performance of the general model to the one obtained with its specific model (Figure 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 2 TF ChIP-chip/seq-data for other plant species/condition. 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 integrating the features of the layers 1-4 in addition to the DHS and the methylation of the cytosine. Indeed, all the genomic data were not available both in 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 (Figure 5). These were higher values than the average AUC of 0.60 associated with sole pattern-matching (Figure 2).

As the model obtained from *O. sativa* showed lower performances than the other models applied to Arabidopsis seedlings, we performed additional analyses to get further insights. We built an extended model, based on the abovementioned features and on 7 different chromatin marks. In rice, the chromatin marks were more important than the DHS and the DGF. Accordingly, the AUC obtained at predicting cis-regulatory elements of TFs in *O. sativa* seedlings increased from 0.76 to 0.84 when the chromatin marks were integrated to the features of the layers 1-4, the DHS and the methylation of the cytosine (Data only presented in text).

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

As an example of application, Wimtrap was used to identify and validate the pathways that are potentially controlled by TT2, TT8 and TTG1 seeds, roots and flowers of *Arabidopsis thaliana*. Even though there were no TF-ChIP data in seeds and roots, predictions were obtained 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, 6 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 3h of light, dark-grown seedlings exposed to a long day cycle) and 1, for *S. lycopersicum* (immature fruits).

To perform this analysis, we predicted at first the gene targets TT2, TT8 and TTG1. We considered that the gene targets of a TF are the genes whose the 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 in average.

The results highlighted a strong impact of the tissue on the type and number of potential TT2, TT8 and TTG1 gene targets (Figures 6 A and B). In addition, a higher number of potential targets has been identified in seeds and roots for TT2 and TT8, compared to TTG1, while a similar number of targets among the MBW TFs has been predicted in flowers (Figure 6A). The GO enrichment analyses revealed a higher number of enriched GO terms in seeds compared to roots and flowers (Figure 6B). Finally, a higher number of enriched GO terms associated to phenylpropanoids and flavonoids was identified for TT2 and TT8, compared to TTG1, with differences according to the tissue considered in the analyses (Figure 6C). In the seed, four and three phenylpropanoid-GO-terms were identified for TT2 and TT8, respectively, while none was predicted for TTG1. Similar results were obtained in roots, while in flowers two enriched phenylpropanoid-GO-terms were highlighted for both TT2 and TTG1.

3 Discussion

3.1 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 to locate with precision the location of the binding sites and they report only for 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, 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.

3.1.1 Predictivity of layers of features. We obtained high performances of models when predicting TFs binding sites in *Arabidopsis* seedlings. Wimtrap highlighted the decisiveness of the features based on the DNAseI-seq data (i.e. those related to the DNAseI hypersensitive sites [DHS – open regions of DNA] and the digital genomic footprints [DGF]). Compared to 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 poised TF binding sites (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 less predictive than the DHS (included in layer 5) and the DGF (layer 4), the features of the layers 1-3, which are related to condition-independent features (results of pattern-matching and 'phylogenetic' footprinting and position on the gene) were shown to be also very valuable for significantly improving the performances of pattern-matching. Layers 1-3 are therefore 'time and cost-effective' as, contrary to layer 5, they are already available for numerous plant.

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 the DNA sequence conservation, in which the dataset 'Phastcons' appeared less predictive than the 'Conserved Elements' one. 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 modelling was especially relevant to get good performances at predicting binding sites based solely on the condition-independent features of the 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 the layer 1-3 and 5.

3.1.2 Multi-scale extraction of genomic features. The analysis at different scales of the genomic regions on which are located the potential binding sites (on 20bp, 400bp and 1000bp windows) is a characteristic of our method. We obtained important gains in the prediction of potential of the features related to the 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 very 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 accessible cis-regulatory elements (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 to overcome some technical limitations. 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 more than 20 bp from the actual binding site (Neph *et al.*, 2012; Gusmao *et al.*, 2014).

3.1.3 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 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 that transferring models from a condition to another comes with a cost in terms of performances. 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 NAC50 and NAC52, for which TF-pooled models are substantially less performing than the TF-specific ones. 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 very by the models built from *A. thaliana* seedlings. On the other hand, we observed than 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 performances of prediction of TFBS in *O. sativa* seedlings were significantly enhanced when data about chromatin marks were added to the 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 of predictivity of features across organisms and conditions. This might be due to technical issues or species/condition-specificities in the gene regulation mechanisms.

3.2 An user-friendly and flexible tools

The user of Wimtrap can easily get 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 the layer 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 kinds of genomic regions than the cis-regulatory elements (e.g. transgene/gene, enhancers/promoters, poised enhancers/active enhancers). In addition, pre-integrated models and databases allow to immediately run the tools for hundreds of TFs 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 the ripening fruits but also in the immature fruits; and for O. sativa seedlings and Z. mays seedlings. (Tian et al., 2020).

The performances of Wimtrap depends obviously 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 user-friendship. We aimed at making easy the building of predictive models for new organism/condition, based on the available data. Here we did not directly confront Wimtrap to 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, p.2; Song *et al.*, 2020). However, TEPIC 2 requires Linux operating systems and ConsReg, expression data, which might be limiting.

3.3 Examples of application of Wimtrap

The activity and function of many TFs is 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-MYB/bHLH/WD40 (MBW) complex, which act synergistically to control the genes involved in the regulation of the late steps of flavonoid and proanthocyanidin (PA) 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 that play 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 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 to 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 works highlighted a major role of TT2 and TT8 in the regulation of the flavonoid late biosynthetic genes in seeds (Xu et al., 2015). As for TTG1, while it has been demonstrated its participation to the MBW complex, 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 on flavonoid regulation and the major impact for these TFs in seeds has been confirmed. The results obtained on TT2, TT8 and TTG1 highlighted a strong impact of the organ and/or the condition on the prediction of TF-target genes (Figure 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 predictions of TF binding sites, which comes with pre-built models for *A. thaliana*, *S. lycopersicum*, *O. sativa*, and *Z. mays*. Prediction of potential TF binding sites can also be useful for comparing TF binding sites 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 next future, the advent of new technologies such as the ChIP-exo/ChIP-nexus and the ATAC-seq will be beneficial. Peaks of ChIP-exo/ChIP-nexus are narrower than the ChIP-chip/seq data and allow therefore to identify more accurately 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 bindings (on primary/alternative motifs) and indirect bindings. As for the ATAC-seq, it is emerging as a cost-effective alternative to the DNAseI-seq (Karabacak Calviello *et al.*, 2019). Relevant data about new organisms and/or conditions will be soon available.

4 Materials & Methods

4.1 Data

A. thaliana seedlings and flowers, and *S. lycopersicum* ripening fruits data were obtained from Arabidopsis RegNet (Heyndrickx *et al.*, 2014), PlantRegMap/PlantTFDB (Jin *et al.*, 2017; Tian et al., 2020), PlantDHS (Zhang *et al.*, 2016), Gene Expression Omnibus (Clough and Barrett, 2016), Ensembl Plants Biomart (Kinsella *et al.*, 2011) databases. Additional information were retrieved from published articles (Thomas *et al.*, 2007; Gómez-Porras *et al.*, 2007; Brandt *et al.*, 2012; Baxter *et al.*, 2012; Nuruzzaman *et al.*, 2013; Haudry *et al.*, 2013; Fujisawa *et al.*, 2014; Zhiponova *et al.*, 2014; Sequeira-Mendes *et al.*, 2014; Wang *et al.*, 2015; Ye *et al.*, 2017; Gaillochet *et al.*, 2017) (Supplementary tables S1-S8). The filters that were used to query Ensembl Plants Biomart and 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 *Zea mays* B73). In addition, we obtained 57 TF-ChIP-seq peak files (28 obtained in *A. thaliana* seedling, 3 in *A. thaliana* flowers, 5 in *S. lycorpersicum* ripening fruits, 4 in *O. sativa* L. ssp. Japonica seedlings and 17 in *Zea mays* B73), 5 sets of conserved elements (2 for *A. thaliana*, 1 for *S. lycopersicum*, 1 for *O. sativa* L. ssp. Japonica, and 1 for *Zea mays* B73), 5 sets of DNAseI-seq and BS-seq data (1 for *A. thaliana* seedlings, 1 for *A. thaliana* flowers, 1 for *S. lycopersicum* ripening fruits, 1 for *O. sativa* L. ssp. Japonica seedlings, and 1 for *Zea mays* B73 seedlings), 1 partitioning of the genome between 9 categories of chromatin stated (1 for *A. thaliana* seedlings), 2 sets of H3K4me3-, H3K4me3-, H3K36me3-, H3K27ac-, H3K9ac-, H4K12ac-, H3K27me3-ChIP-seq data (1 for *A. thaliana* seedlings and 1 for *O. sativa* L. ssp. Japonica seedlings), and 1 set of MNase-seq, H2A.Z-, H2BuB18-, H3K4me1-, H3K4me2-, H3K9me2-, H3K27me1-, H3K14ac-, H4K5ac-, H3K18ac-, H3K56ac-, H3T3ph-, H4K8ac-, and H4K16ac-ChIP-seq data (1 for *A. thaliana* seedlings). Furthermore, we directly collected the motifs of 55 of the 57 TFs, either as Pseudo Weight Matrix (PWM) or a logo. Details about the source of the data, the experimental design as well as the data analysis pipeline are provided in supplementary tables 1-14. In particular, for ChIP-seq data, the number of samples is comprised between 1 and 4 (2.1 in average ± 0.95 standard deviation) and the FDR, between 10^{-2} and 10^{-5} (0.04 in average ± 0.02 standard deviation).

4.2 Data pre-processing

4.2.1. 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 1000. TFs for which such pre-processing steps were necessary to obtain the PWM are specified in Supplementary tables 1, 4, and 6.

4.2.2. 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'UTR, coding sequences (CDS), introns, 3'UTR 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 S1.

4.2.3. 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 9 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.

4.2.4. ChIP/DNAse/BS/MNase-peaks

In the majority of cases, results of peak-calling analyses could be obtained from 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 the wig, the bedGraph and the 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 provide 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 partition 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 minimum 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 inferior to 30bp 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 S1-S8.

4.2.5. DGFs

The location and scores of digital genomic footprints obtained with footprinting2012 (Neph *et al.*, 2012) tool for *A. thaliana* seedlings could be directly downloaded in BED format from PlantRegMap (Tian *et al.*, 2019). 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 S3.

4.3 Identification of candidate TF binding sites

Candidate TF bindings sites 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 to 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 low prevalence of actual TF binding sites. Sequences of 5 000 bp were thus randomly sampled at a rate of

200 bp by chromosome and were scanned at each bp on both strands. The resulting match scores were ordered in the increasing order and associated to their p-value, i.e. the proportion of matches with an equal or superior score.

Our pattern-matching approach was compared to 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 figure 9). These considerations indicated the accuracy of our method.

Candidate TF-binding sites were defined as the PWM matches with a p-value equal or higher to 10^{-3}. This threshold allowed the detection, for the 28 TFs related to *A. thaliana* seedlings, the most prevalent ('primary') motif on 2/3 in average of the cognate ChIP-peaks, which corresponds to previous observations (Heyndrickx *et al.*, 2014) (Supplementary figures 9 and 10, Supplementary table 9).

4.4 Feature construction

Candidate TF binding sites were annotated with 5 layers of features. The 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 than that of the candidate binding site, occurring at ± 200 bp and ± 500 bp from the centre of the candidate TF binding site. The layer 2 was relative to the position of the candidate TF binding site 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 centre of the considered candidate was associated to the score of '1'; the other structures were granted with the score of '0'. In the case where several structures overlapped a same potential TF binding site, only one structure was left with the 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 to the sequence conservation, the DGF and the chromatin state/opening. Categorical features (c.f. the partitioning of the genome of *A. thaliana* into 9 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 TF binding sites and 0 to the other ones. 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 the layers 3-5, we calculated the base-pair average of each considered features around the PWM matches, on three different scales: on windows of ± 10 bp, 200 bp and 1 000 bp from the center of the candidate TF binding sites. 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 consider the extracted data as an ensemble of *n* genomic regions defined each by their location $\{y_1, y_2, ..., y_n \mid y = (chromosome, start, end)\}$ and by their scores $\{x_1, x_2, ..., x_n\}$ ($x_{1,...,n} = 1$ for overlapping features). Let be $\overline{x}$, 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*:

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

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).

4.5 Candidate TF binding sites labelling

The candidate binding sites for a given TF were labelled 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 labelled as active cis-regulatory elements and '0' for the other ones. The length of the ChIP-peaks was limited to ± 200 bp from the peak centers as most of the PWM matches were located in this interval (Supplementary table 9).

4.6 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 than there was 'positive' ones, using the *sample.int* function of the base package in R and removed those from the dataset. The selected 'negative' instance were kept in the master dataset and the other ones 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 (Sotiris Kotsiantis *et al.*, 2006). The master dataset was then split into 3 TF-pooled datasets, according to the organism and the condition: *A. thaliana* seedlings, *A. thaliana* flowers and *Solanum lycopersicum* ripening fruits. These datasets were then subdivided into TF-specific datasets.

4.7 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 higher than 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* seedlings data. The performances of the models were estimated using the 5-cross validation strategy: each TF-specific dataset is cut into 5 smaller datasets of equal size. A model is trained with 4 of the 5 parts while the area under the ROC curve (AUC) is computed by applying the model on the remaining part. The process is repeated again 4 times so that each of the 5 parts is used for computing the AUC. The final AUC of a model is the mean of the 5 AUCs thus obtained. The AUCs were calculated with the pROC R package (Robin *et al.*, 2011).

In the first places, we evaluated algorithms of 'random forest', 'logistic regression' and 'gradient boosting'. Gradient boosting was clearly outcompeting (data not shown). We tested 3 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). 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, 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 area under the curve (AUC) of 0.925 was achieved with CatBoost, and 0.927 with lightGBM and XGBoost as well (Supplementary figure 11). We selected XGBoost as it is a well-established method since several years (Chen and Guestrin, 2016). XGBoost is an algorithm which adds the predictions of an ensemble of regression trees. It builds successively the regression tree, 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:

$$\hat{y}_i = \Phi(x_i) = \sum_{k=1}^{K} f_k(x_i)$$

Where $\hat{y}_i$ is the ith prediction, obtained by addition of the outputs of the K regression trees $\{f_1, f_2, \dots, f_k\}$, based on the vector of features x_i. The regression trees are defined so that the regularized objective is minimized:

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

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

The function is the loss function which measures the difference between $\hat{y}_i$, the ith prediction, and y_i, the actual ith value (= 1 if the ith instance is a 'positive' candidate TF binding sites, = 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 over-fitting.

4.8 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. This threshold is A candidate TF binding site 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’. Higher is the AUC, more accurate is a model. 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 the TFs of the dataset but one. 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.

4.9 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 TF-binding sites. They were then labelled 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 3 TFs studied in *A. thaliana* flowers, based on all features of the layers 1 to 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, based also on all the features of the layers 1 to 4 in addition to the DHS (data about other features of the layer 5 were not available in flowers, seeds and roots). For TT2 and TT8, we determined the genes whose the TSS was the closest of an occurrence of their respective motifs (Jacob *et al.*, 2021) with a Wimtrap prediction score ≥ 0.86. For TTG1, we determined the genes whose the TSS was the closest of 2 neighbouring motifs - 1 G-Box close to 1 AC-Rich- or 1 MYB-motif – maximum distance between the 2 motifs = 30bp (Xu *et al.*, 2015) - with both a prediction scores ≥ 0.86.

5 Data availability

Wimtrap can be downloaded from Github as a classical R package (https://github.com/RiviereQuentin/Wimtrap) or as an 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 to reproduce 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

6 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.

7 Acknowledgments

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

8 Disclosures

The authors have no conflicts of interest to declare.

9 References

Aerts, S. (2012) 'Computational Strategies for the Genome-Wide Identification of Cis-Regulatory Elements and Transcriptional Targets'. In *Current Topics in Developmental Biology*. Elsevier, pp. 121–145. DOI: 10.1016/B978-0-12-386499-4.00005-7.

Aho, A.V., Kernighan, B.W. and Weinberger, P.J. (1988) *The AWK Programming Language*. Addison-Wesley Publishing Company Available at: https://books.google.be/books?id=53ueQgAACAAJ.

Arvey, A. *et al.* (2012) 'Sequence and Chromatin Determinants of Cell-Type-Specific Transcription Factor Binding'. *Genome Research*, 22(9), pp. 1723–1734. DOI: 10.1101/gr.127712.111.

Baudry, A. *et al.* (2004) 'TT2, TT8, and TTG1 Synergistically Specify the Expression of *BANYULS* and Proanthocyanidin Biosynthesis in *Arabidopsis Thaliana*'. *The Plant Journal*, 39(3), pp. 366–380. DOI: 10.1111/j.1365-313X.2004.02138.x.

Baxter, L. *et al.* (2012) 'Conserved Noncoding Sequences Highlight Shared Components of Regulatory Networks in Dicotyledonous Plants'. *The Plant Cell*, 24(10), pp. 3949–3965. DOI: 10.1105/tpc.112.103010.

Behjati Ardakani, F., Schmidt, F. and Schulz, M.H. (2019) 'Predicting Transcription Factor Binding Using Ensemble Random Forest Models'. *F1000Research*, 7, p. 1603. DOI: 10.12688/f1000research.16200.2.

Bonev, B. and Cavalli, G. (2016) 'Organization and Function of the 3D Genome'. *Nature Reviews Genetics*, 17(11), pp. 661–678. DOI: 10.1038/nrg.2016.112.

Boyle, A.P. *et al.* (2011) 'High-Resolution Genome-Wide in Vivo Footprinting of Diverse Transcription Factors in Human Cells'. *Genome Research*, 21(3), pp. 456–464. DOI: 10.1101/gr.112656.110.

Brandt, R. *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*'. *The Plant Journal*, 72(1), pp. 31–42. DOI: 10.1111/j.1365-313X.2012.05049.x.

Budden, D.M. *et al.* (2014) 'Predicting Expression: The Complementary Power of Histone Modification and Transcription Factor Binding Data'. *Epigenetics & Chromatin*, 7(1), p. 36. DOI: 10.1186/1756-8935-7-36.

Butel, N. *et al.* (2017) '*Sgs1* : A Neomorphic *Nac52* Allele Impairing Post-Transcriptional Gene Silencing through *SGS3* Downregulation'. *The Plant Journal*, 90(3), pp. 505–519. DOI: 10.1111/tpj.13508.

Castro-Mondragon, J.A. *et al.* (2022) 'JASPAR 2022: The 9th Release of the Open-Access Database of Transcription Factor Binding Profiles'. *Nucleic Acids Research*, 50(D1), pp. D165–D173. DOI: 10.1093/nar/gkab1113.

Chen, T. *et al.* (2021) *Xgboost: Extreme Gradient Boosting*. Available at: https://CRAN.R-project.org/package=xgboost.

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. San Francisco, California, USA: ACM Press, pp. 785–794. DOI: 10.1145/2939672.2939785.

Chen, X. *et al.* (2010) ‘A Dynamic Bayesian Network for Identifying Protein-Binding Footprints from Single Molecule-Based Sequencing Data’. *Bioinformatics*, 26(12), pp. i334–i342. DOI: 10.1093/bioinformatics/btq175.

Chen, X. *et al.* (2017) ‘Mocap: Large-Scale Inference of Transcription Factor Binding Sites from Chromatin Accessibility’. *Nucleic Acids Research*, 45(8), pp. 4315–4329. DOI: 10.1093/nar/gkx174.

Clough, E. and Barrett, T. (2016) ‘The Gene Expression Omnibus Database’. In Mathé, E. and Davis, S. (eds.) *Statistical Genomics*. New York, NY: Springer New York, pp. 93–110. DOI: 10.1007/978-1-4939-3578-9_5.

Collings, C.K., Waddell, P.J. and Anderson, J.N. (2013) ‘Effects of DNA Methylation on Nucleosome Stability’. *Nucleic Acids Research*, 41(5), pp. 2918–2931. DOI: 10.1093/nar/gks893.

Cuellar-Partida, G. *et al.* (2012) ‘Epigenetic Priors for Identifying Active Transcription Factor Binding Sites’. *Bioinformatics*, 28(1), pp. 56–62. DOI: 10.1093/bioinformatics/btr614.

Dorogush, A.V., Ershov, V. and Gulin, A. (2018) ‘CatBoost: Gradient Boosting with Categorical Features Support’. *CoRR*, abs/1810.11363. Available at: http://arxiv.org/abs/1810.11363.

Ferrari, K.J. *et al.* (2014) ‘Polycomb-Dependent H3K27me1 and H3K27me2 Regulate Active Transcription and Enhancer Fidelity’. *Molecular Cell*, 53(1), pp. 49–62. DOI: 10.1016/j.molcel.2013.10.030.

Franco-Zorrilla, J.M. *et al.* (2014) ‘DNA-Binding Specificities of Plant Transcription Factors and Their Potential to Define Target Genes’. *Proceedings of the National Academy of Sciences*, 111(6), pp. 2367–2372. DOI: 10.1073/pnas.1316278111.

Fuda, N.J., Ardehali, M.B. and Lis, J.T. (2009) ‘Defining Mechanisms That Regulate RNA Polymerase II Transcription in Vivo’. *Nature*, 461(7261), pp. 186–192. DOI: 10.1038/nature08449.

Fujisawa, M. *et al.* (2014) ‘Transcriptional Regulation of Fruit Ripening by Tomato FRUITFULL Homologs and Associated MADS Box Proteins’. *The Plant Cell*, 26(1), pp. 89–101. DOI: 10.1105/tpc.113.119453.

Gaillochet, C. *et al.* (2017) ‘Control of Plant Cell Fate Transitions by Transcriptional and Hormonal Signals’. *ELife*, 6, p. e30135. DOI: 10.7554/eLife.30135.

Gómez-Porras, J.L. *et al.* (2007) ‘Genome-Wide Analysis of ABA-Responsive Elements ABRE and CE3 Reveals Divergent Patterns in Arabidopsis and Rice’. *BMC Genomics*, 8(1), p. 260. DOI: 10.1186/1471-2164-8-260.

Grant, C.E., Bailey, T.L. and Noble, W.S. (2011) ‘FIMO: Scanning for Occurrences of a given Motif’. *Bioinformatics*, 27(7), pp. 1017–1018. DOI: 10.1093/bioinformatics/btr064.

Gusmao, E.G. *et al.* (2016) ‘Analysis of Computational Footprinting Methods for DNase Sequencing Experiments’. *Nature Methods*, 13(4), pp. 303–309. DOI: 10.1038/nmeth.3772.

Gusmao, E.G. *et al.* (2014) ‘Detection of Active Transcription Factor Binding Sites with the Combination of DNase Hypersensitivity and Histone Modifications’. *Bioinformatics*, 30(22), pp. 3143–3151. DOI: 10.1093/bioinformatics/btu519.

Hardison, R.C. and Taylor, J. (2012) ‘Genomic Approaches towards Finding Cis-Regulatory Modules in Animals’. *Nature Reviews Genetics*, 13(7), pp. 469–483. DOI: 10.1038/nrg3242.

Haudry, A. *et al.* (2013) ‘An Atlas of over 90,000 Conserved Noncoding Sequences Provides Insight into Crucifer Regulatory Regions’. *Nature Genetics*, 45(8), pp. 891–898. DOI: 10.1038/ng.2684.

Hesselberth, J.R. *et al.* (2009) ‘Global Mapping of Protein-DNA Interactions in Vivo by Digital Genomic Footprinting’. *Nature Methods*, 6(4), pp. 283–289. DOI: 10.1038/nmeth.1313.

Heyndrickx, K.S. *et al.* (2014) 'A Functional and Evolutionary Perspective on Transcription Factor Binding in *Arabidopsis Thaliana*'. *The Plant Cell*, 26(10), pp. 3894–3910. DOI: 10.1105/tpc.114.130591.

Jacob, P. *et al.* (2021) 'The Seed Development Factors TT2 and MYB5 Regulate Heat Stress Response in Arabidopsis'. *Genes*, 12(5), p. 746. DOI: 10.3390/genes12050746.

Jankowski, A., Tiuryn, J. and Prabhakar, S. (2016) 'Romulus: Robust Multi-State Identification of Transcription Factor Binding Sites from DNase-Seq Data'. *Bioinformatics*, 32(16), pp. 2419–2426. DOI: 10.1093/bioinformatics/btw209.

Jayaram, N., Usvyat, D. and R. Martin, A.C. (2016) 'Evaluating Tools for Transcription Factor Binding Site Prediction'. *BMC Bioinformatics*, 17(1), p. 547. DOI: 10.1186/s12859-016-1298-9.

Jin, J. *et al.* (2017) 'PlantTFDB 4.0: Toward a Central Hub for Transcription Factors and Regulatory Interactions in Plants'. *Nucleic Acids Research*, 45(D1), pp. D1040–D1045. DOI: 10.1093/nar/gkw982.

Jones, P.A. (2012) 'Functions of DNA Methylation: Islands, Start Sites, Gene Bodies and Beyond'. *Nature Reviews Genetics*, 13(7), pp. 484–492. DOI: 10.1038/nrg3230.

Kähärä, J. and Lähdesmäki, H. (2015) 'BinDNase: A Discriminatory Approach for Transcription Factor Binding Prediction Using DNase I Hypersensitivity Data'. *Bioinformatics*, 31(17), pp. 2852–2859. DOI: 10.1093/bioinformatics/btv294.

Kaplan, T. *et al.* (2011) 'Quantitative Models of the Mechanisms That Control Genome-Wide Patterns of Transcription Factor Binding during Early Drosophila Development' Barsh, G.S. (ed.). *PLoS Genetics*, 7(2), p. e1001290. DOI: 10.1371/journal.pgen.1001290.

Karabacak Calviello, A. *et al.* (2019) 'Reproducible Inference of Transcription Factor Footprints in ATAC-Seq and DNase-Seq Datasets Using Protocol-Specific Bias Modeling'. *Genome Biology*, 20(1), p. 42. DOI: 10.1186/s13059-019-1654-y.

Ke, G. *et al.* (2017) 'LightGBM: A Highly Efficient Gradient Boosting Decision Tree'. In *Advances in Neural Information Processing Systems 30 (NIP 2017)*. Available at: https://www.microsoft.com/en-us/research/publication/lightgbm-a-highly-efficient-gradient-boosting-decision-tree/.

Keilwagen, J., Posch, S. and Grau, J. (2019) 'Accurate Prediction of Cell Type-Specific Transcription Factor Binding'. *Genome Biology*, 20(1), p. 9. DOI: 10.1186/s13059-018-1614-y.

Kent, W.J. *et al.* (2002) 'The Human Genome Browser at UCSC'. *Genome Research*, 12(6), pp. 996–1006. DOI: 10.1101/gr.229102.

Kinsella, R.J. *et al.* (2011) 'Ensembl BioMarts: A Hub for Data Retrieval across Taxonomic Space'. *Database*, 2011(0), pp. bar030–bar030. DOI: 10.1093/database/bar030.

Kuhn, M. (2020) *Caret: Classification and Regression Training*. Available at: https://CRAN.R-project.org/package=caret.

Kumar, S. and Bucher, P. (2016) 'Predicting Transcription Factor Site Occupancy Using DNA Sequence Intrinsic and Cell-Type Specific Chromatin Features'. *BMC Bioinformatics*, 17(S1), p. S4. DOI: 10.1186/s12859-015-0846-z.

Lai, X. *et al.* (2019) 'Building Transcription Factor Binding Site Models to Understand Gene Regulation in Plants'. *Molecular Plant*, 12(6), pp. 743–763. DOI: 10.1016/j.molp.2018.10.010.

Lawrence, M. *et al.* (2013) 'Software for Computing and Annotating Genomic Ranges' Prlic, A. (ed.). *PLoS Computational Biology*, 9(8), p. e1003118. DOI: 10.1371/journal.pcbi.1003118.

Lawrence, M., Daujat, S. and Schneider, R. (2016) 'Lateral Thinking: How Histone Modifications Regulate Gene Expression'. *Trends in Genetics*, 32(1), pp. 42–56. DOI: 10.1016/j.tig.2015.10.007.

Lawrence, M., Gentleman, R. and Carey, V. (2009) 'Rtracklayer: An R Package for Interfacing with Genome Browsers'. *Bioinformatics*, 25(14), pp. 1841–1842. DOI: 10.1093/bioinformatics/btp328.

Lee, D.J., Minchin, S.D. and Busby, S.J.W. (2012) ‘Activating Transcription in Bacteria’. *Annual Review of Microbiology*, 66(1), pp. 125–152. DOI: 10.1146/annurev-micro-092611-150012.

Lenhard, B., Sandelin, A. and Carninci, P. (2012) ‘Metazoan Promoters: Emerging Characteristics and Insights into Transcriptional Regulation’. *Nature Reviews Genetics*, 13(4), pp. 233–245. DOI: 10.1038/nrg3163.

Li, H. and Guan, Y. (2019) *Leopard: Fast Decoding Cell Type-Specific Transcription Factor Binding Landscape at Single-Nucleotide Resolution*. Bioinformatics DOI: 10.1101/856823.

Li, H., Quang, D. and Guan, Y. (2019) ‘Anchor: Trans-Cell Type Prediction of Transcription Factor Binding Sites’. *Genome Research*, 29(2), pp. 281–292. DOI: 10.1101/gr.237156.118.

Liu, S. *et al.* (2017) ‘Assessing the Model Transferability for Prediction of Transcription Factor Binding Sites Based on Chromatin Accessibility’. *BMC Bioinformatics*, 18(1), p. 355. DOI: 10.1186/s12859-017-1769-7.

Luo, K. and Hartemink, A.J. (2013) ‘Using DNase Digestion Data to Accurately Identify Transcription Factor Binding Sites’. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 80–91.

Meireles-Filho, A.C. and Stark, A. (2009) ‘Comparative Genomics of Gene Regulation—Conservation and Divergence of Cis-Regulatory Information’. *Current Opinion in Genetics & Development*, 19(6), pp. 565–570. DOI: 10.1016/j.gde.2009.10.006.

Meyer, C.A. and Liu, X.S. (2014) ‘Identifying and Mitigating Bias in Next-Generation Sequencing Methods for Chromatin Biology’. *Nature Reviews Genetics*, 15(11), pp. 709–721. DOI: 10.1038/nrg3788.

Muiño, J.M. *et al.* (2011) ‘ChIP-Seq Analysis in R (CSAR): An R Package for the Statistical Detection of Protein-Bound Genomic Regions’. *Plant Methods*, 7(1), p. 11. DOI: 10.1186/1746-4811-7-11.

Mundade, R. *et al.* (2014) ‘Role of ChIP-Seq in the Discovery of Transcription Factor Binding Sites, Differential Gene Regulation Mechanism, Epigenetic Marks and Beyond’. *Cell Cycle*, 13(18), pp. 2847–2852. DOI: 10.4161/15384101.2014.949201.

Natarajan, A. *et al.* (2012) ‘Predicting Cell-Type-Specific Gene Expression from Regions of Open Chromatin’. *Genome Research*, 22(9), pp. 1711–1722. DOI: 10.1101/gr.135129.111.

Neph, S. *et al.* (2012) ‘An Expansive Human Regulatory Lexicon Encoded in Transcription Factor Footprints’. *Nature*, 489(7414), pp. 83–90. DOI: 10.1038/nature11212.

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’. *Frontiers in Microbiology*, 4. DOI: 10.3389/fmicb.2013.00248.

O’Connor, T.R. and Bailey, T.L. (2014) ‘Creating and Validating Cis-Regulatory Maps of Tissue-Specific Gene Expression Regulation’. *Nucleic Acids Research*, 42(17), pp. 11000–11010. DOI: 10.1093/nar/gku801.

Pagès, H. *et al.* (2019) ‘Biostrings: Efficient Manipulation of Biological Strings’. *R Package Version 2.54.0*.

Piper, J. *et al.* (2013) ‘Wellington: A Novel Method for the Accurate Identification of Digital Genomic Footprints from DNase-Seq Data’. *Nucleic Acids Research*, 41(21), pp. e201–e201. DOI: 10.1093/nar/gkt850.

Pique-Regi, R. *et al.* (2011) ‘Accurate Inference of Transcription Factor Binding from DNA Sequence and Chromatin Accessibility Data’. *Genome Research*, 21(3), pp. 447–455. DOI: 10.1101/gr.112623.110.

Pott, S. and Lieb, J.D. (2015) ‘What Are Super-Enhancers?’ *Nature Genetics*, 47(1), pp. 8–12. DOI: 10.1038/ng.3167.

Prokhorenkova, L. *et al.* (2019) ‘CatBoost: Unbiased Boosting with Categorical Features’.

Qin, Q. and Feng, J. (2017) ‘Imputation for Transcription Factor Binding Predictions Based on Deep Learning’ Ioshikhes, I. (ed.). *PLOS Computational Biology*, 13(2), p. e1005403. DOI: 10.1371/journal.pcbi.1005403.

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 DOI: 10.1101/151274.

Quattrocchio, F. *et al.* (2006) 'The Regulation of Flavonoid Biosynthesis'. In *The Science of Flavonoids*. pp. 97–122.

Raj, A. *et al.* (2015) 'MsCentipede: Modeling Heterogeneity across Genomic Sites and Replicates Improves Accuracy in the Inference of Transcription Factor Binding' Zheng, D. (ed.). *PLOS ONE*, 10(9), p. e0138030. DOI: 10.1371/journal.pone.0138030.

Rister, J. and Desplan, C. (2010) 'Deciphering the Genome's Regulatory Code: The Many Languages of DNA'. *BioEssays*, 32(5), pp. 381–384. DOI: 10.1002/bies.200900197.

Robin, X. *et al.* (2011) 'PROC: An Open-Source Package for R and S+ to Analyze and Compare ROC Curves'. *BMC Bioinformatics*, 12, p. 77.

van Rooijen, R. *et al.* (2020) 'Targeted Misexpression of NAC052, Acting in H3K4 Demethylation, Alters Leaf Morphological and Anatomical Traits in Arabidopsis Thaliana' Griffiths, H. (ed.). *Journal of Experimental Botany*, 71(4), pp. 1434–1448. DOI: 10.1093/jxb/erz509.

Schmidt, F. *et al.* (2017) 'Combining Transcription Factor Binding Affinities with Open-Chromatin Data for Accurate Gene Expression Prediction'. *Nucleic Acids Research*, 45(1), pp. 54–66. DOI: 10.1093/nar/gkw1061.

Schmidt, F. *et al.* (2019) 'TEPIC 2—an Extended Framework for Transcription Factor Binding Prediction and Integrative Epigenomic Analysis' Berger, B. (ed.). *Bioinformatics*, 35(9), pp. 1608–1609. DOI: 10.1093/bioinformatics/bty856.

Sequeira-Mendes, J. *et al.* (2014) 'The Functional Topography of the *Arabidopsis* Genome Is Organized in a Reduced Number of Linear Motifs of Chromatin States'. *The Plant Cell*, 26(6), pp. 2351–2366. DOI: 10.1105/tpc.114.124578.

Sherwood, R.I. *et al.* (2014) 'Discovery of Directional and Nondirectional Pioneer Transcription Factors by Modeling DNase Profile Magnitude and Shape'. *Nature Biotechnology*, 32(2), pp. 171–178. DOI: 10.1038/nbt.2798.

Shi, Y. *et al.* (2021) *Lightgbm: Light Gradient Boosting Machine*. Available at: https://CRAN.R-project.org/package=lightgbm.

Siepel, A. and Haussler, D. (2005) 'Phylogenetic Hidden Markov Models'. In *Statistical Methods in Molecular Evolution*. Statistics for Biology and Health. New York: Springer-Verlag, pp. 325–351. DOI: 10.1007/0-387-27733-1_12.

Song, Q. *et al.* (2020) 'Prediction of Condition-Specific Regulatory Genes Using Machine Learning'. *Nucleic Acids Research*, 48(11), pp. e62–e62. DOI: 10.1093/nar/gkaa264.

Sotiris Kotsiantis, Dimitris Kanellopoulos., and Panayiotis Pintelas (2006) 'Handling Imbalanced Datasets: A Review'. *GESTS International Transactions on Computer Science and Engineering*, 30.

spicuglia, salvatore. and Vanhille, L. (2012) 'Chromatin Signatures of Active Enhancers'. *Nucleus*, 3(2), pp. 126–131. DOI: 10.4161/nucl.19232.

Spitz, F. and Furlong, E.E.M. (2012) 'Transcription Factors: From Enhancer Binding to Developmental Control'. *Nature Reviews Genetics*, 13(9), pp. 613–626. DOI: 10.1038/nrg3207.

Sung, M.-H. *et al.* (2014) 'DNase Footprint Signatures Are Dictated by Factor Dynamics and DNA Sequence'. *Molecular Cell*, 56(2), pp. 275–285. DOI: 10.1016/j.molcel.2014.08.016.

Thomas, B.C. *et al.* (2007) 'Arabidopsis Intragenomic Conserved Noncoding Sequence'. *Proceedings of the National Academy of Sciences*, 104(9), pp. 3348–3353. DOI: 10.1073/pnas.0611574104.

Thomas-Chollier, M. *et al.* (2012) 'RSAT Peak-Motifs: Motif Analysis in Full-Size ChIP-Seq Datasets'. *Nucleic Acids Research*, 40(4), pp. e31–e31. DOI: 10.1093/nar/gkr1104.

Tian, F. *et al.* (2019) 'PlantRegMap: Charting Functional Regulatory Maps in Plants'. *Nucleic Acids Research*, p. gkz1020. DOI: 10.1093/nar/gkz1020.

Tsai, Z.T.-Y., Shiu, S.-H. and Tsai, H.-K. (2015) 'Contribution of Sequence Motif, Chromatin State, and DNA Structure Features to Predictive Models of Transcription Factor Binding in Yeast' Ioshikhes, I. (ed.). *PLOS Computational Biology*, 11(8), p. e1004418. DOI: 10.1371/journal.pcbi.1004418.

Vaquerizas, J.M. *et al.* (2009) 'A Census of Human Transcription Factors: Function, Expression and Evolution'. *Nature Reviews Genetics*, 10(4), pp. 252–263. DOI: 10.1038/nrg2538.

Veljkovic, J. and Hansen, U. (2004) 'Lineage-Specific and Ubiquitous Biological Roles of the Mammalian Transcription Factor LSF'. *Gene*, 343(1), pp. 23–40. DOI: 10.1016/j.gene.2004.08.010.

Vuong, P. and Misr, R. (2011) 'Guide to Genome-Wide Bacterial Transcription Factor Binding Site Prediction Using OmpR as Model'. In Xia, X. (ed.) *Selected Works in Bioinformatics*. InTech. DOI: 10.5772/24321.

Wang, C. *et al.* (2015) 'Genome-Wide Analysis of Local Chromatin Packing in *Arabidopsis Thaliana*'. *Genome Research*, 25(2), pp. 246–256. DOI: 10.1101/gr.170332.113.

Welch, R. *et al.* (2017) 'Data Exploration, Quality Control and Statistical Analysis of ChIP-Exo/Nexus Experiments'. *Nucleic Acids Research*, 45(15), pp. e145–e145. DOI: 10.1093/nar/gkx594.

Wittkopp, P.J. and Kalay, G. (2012) 'Cis-Regulatory Elements: Molecular Mechanisms and Evolutionary Processes Underlying Divergence'. *Nature Reviews Genetics*, 13(1), pp. 59–69. DOI: 10.1038/nrg3095.

Wolfe, K.H. *et al.* (1989) 'Date of the Monocot-Dicot Divergence Estimated from Chloroplast DNA Sequence Data.' *Proceedings of the National Academy of Sciences*, 86(16), pp. 6201–6205. DOI: 10.1073/pnas.86.16.6201.

Won, K.-J., Ren, B. and Wang, W. (2010) 'Genome-Wide Prediction of Transcription Factor Binding Sites Using an Integrated Model'. *Genome Biology*, 11(1), p. R7. DOI: 10.1186/gb-2010-11-1-r7.

Xu, W., Dubos, C. and Lepiniec, L. (2015) 'Transcriptional Control of Flavonoid Biosynthesis by MYB–BHLH–WDR Complexes'. *Trends in Plant Science*, 20(3), pp. 176–185. DOI: 10.1016/j.tplants.2014.12.001.

Yardımcı, G.G. *et al.* (2014) 'Explicit DNase Sequence Bias Modeling Enables High-Resolution Transcription Factor Footprint Detection'. *Nucleic Acids Research*, 42(19), pp. 11865–11878. DOI: 10.1093/nar/gku810.

Ye, H. *et al.* (2017) 'RD26 Mediates Crosstalk between Drought and Brassinosteroid Signalling Pathways'. *Nature Communications*, 8(1), p. 14573. DOI: 10.1038/ncomms14573.

Zhang, S. *et al.* (2015) 'C-Terminal Domains of Histone Demethylase JMJ14 Interact with a Pair of NAC Transcription Factors to Mediate Specific Chromatin Association'. *Cell Discovery*, 1(1), p. 15003. DOI: 10.1038/celldisc.2015.3.

Zhang, T., Marand, A.P. and Jiang, J. (2016) 'PlantDHS: A Database for DNase I Hypersensitive Sites in Plants'. *Nucleic Acids Research*, 44(D1), pp. D1148–D1153. DOI: 10.1093/nar/gkv962.

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 Physiology*, 168(4), pp. 1406–1416. DOI: 10.1104/pp.15.00125.

Zhiponova, M.K. *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'. *Proceedings of the National Academy of Sciences*, 111(7), pp. 2824–2829. DOI: 10.1073/pnas.1400203111.

Zhu, B. *et al.* (2015) 'Genome-Wide Prediction and Validation of Intergenic Enhancers in Arabidopsis Using Open Chromatin Signatures'. *The Plant Cell*, 27(9), pp. 2415–2426. DOI: 10.1105/tpc.15.00537.

10 Legends to main figures and tables

Figure 1 Methodology workflow diagram *Data Gathering:* The Wimtrap pipeline starts with a step of data gathering from the literature and various specialized databases in order to obtain, for a given organism and condition, TF ChIP-peak results and related PWMs, genome sequences, transcript models and additional genomic data related to sequence conservation, digital footrprinting and/or chromatin state. Most of the data consist of genomic maps, i.e. of location of peaks/elements that are optionally scored. *Location, labelling and annotation of candidate TF-binding sites:* The potential binding sites of the TFs included in ChIP-peak results are (i) located by pattern-matching, (ii) labelled as 'positive' when they overlap a ChIP-peak in the considered condition, 'negative' when not and (iii) annotated with their position on the closest transcript and with the average signal of the genomic data in their neighbourhood, on windows of ±10bp, ±200bp and ±500bp. *Machine learning:* Predictive models are trained by extreme gradient boosting (XGBoosting). Two strategies are used to train a model: either the algorithm is fed with the data of only one TF and obtain a 'TF-specific model'; either with the data of all TFs studied in a given condition and organism and we obtain a 'TF-pooled model'. *Evaluation:* The TF-specific models obtained from *Arabidopsis thaliana* seedlings serve to evaluate the accuracy of the models according to the genomic data that they integrated and to assess the feature importance. Model performances are assessed with the area under the curve (AUC); the feature importance, by the gain. The TF-pooled models are used to evaluate the transferability of the models between organism/condition/TF and to illustrate an example of application of Wimtrap. SI: Supporting information: PWM: Position Weight Matrix; TF: Transcription factor; DGF: Digital Genomic Footprint; Chr: Chromosome; bp: base pair.

Figure 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 transcription factor, a model is built and evaluated based on a balanced data set for that factor following the 5-fold cross validation procedure: the considered data set is divided into 5 partitions. Among these, 4 are considered to build a model and 1 is used to assess performances. The operation is repeated 5 times, in such a way that each partition is retained only once for validation purposes. (C, D) Idem but considering combination of selected features. PM: Pattern-Matching; DHS: DNAseI-hypersensitivity; DGF: Digital Genomic Footprint scores. The layer 1 includes the results of pattern-matching.

Figure 3 Importance of the genomic features in the full TF-specific models obtained from transcription factors 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 data, the gains associated to the features extracted on windows of 20bp, 400bp and 1000bp are summed. DGF: Digital genomic footprint; DHS: DNAseI-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.

Figure 4 Comparison of the performance of the TF-specific models with the TF-pooled models For each transcription factor, a model based on the data related to this transcription (TF-specific model) and to the other transcription factors (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.

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

Figure 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) Enriched GO-terms related to metabolites and phenylpropanoid metabolism among the putative targets of TT2, TT8 and TTG1, in the different organs considered.