14-12-2018

A novel machine learning approach to study the transcription factor binding sites of MADS-box proteins

Author: Jarno Persoon (940501648060)
Programme(s): Plant Biotechnology (MPB), Bio-informatics (MBF)
Supervisor: dr. Aalt-Jan van Dijk
Examiner: dr. ir. Dick de Ridder
Institution: Wageningen University Research, department Bio-informatics
Thesis Period: May – December 2018

Abstract

Gene regulation is mediated by the binding of a transcription factor (TF) to a transcription factor binding site (TFBS). This binding is influenced by different factors such as the sequence composition of the TFBS, epigenetic regulation, DNA shape and protein interactions. The TFBS of the MADS-box protein family has been extensively studied and is highly similar among different members of this family, however the MADS transcription factors regulate different target genes. So far, the precise relationship between the MADS-box proteins and their TFBSs remains unknown. To gain more insights into this relationship we trained multiple random forest models that classify the binding of a TF to a DNA region found by chromatin immunoprecipitation sequencing (ChIP-seq). The features of these models are related to DNA-properties, Gene Ontology (GO)-terms and the sharing of peak regions between TFs. We trained separate models for the MADS TFs: AGAMOUS (AG), APETALA1 (AP1), APETALA3 (AP3), FLOWERING LOCUS C (FLC), PISTILLATA (PI), SEPALLATA (SEP3), SUPPRESOR OF OVEREXPRESSION OF CO 1 (SOC1) and SHORT VEGETATIVE PHASE (SVP). Interestingly, these models show that GO-terms do not have much predictive value for any MADS-box protein. On the other hand, features describing the DNA-properties and sharing of peak regions between TFs have a high predictive value. From these feature groups, the features describing the motif sequences in a peak region and the sharing of peak regions between MADS-box proteins have high feature importance's. Together these results indicate, that MADS-box proteins share similar functions and that the regulation of their gene targets is mediated by specific motif sequences and interactions between other TFs. Additionally, we present an effective novel machine learning approach to study TFBSs by training a random forest model with different features related to the TFBS.

Introduction

Precise regulation of gene expression during flowering and floral organ development is vital for plant reproduction. The MADS-box transcription factors (TFs) have an essential role in these processes and regulate gene expression by binding to a transcription factor binding site (TFBS) located in a promoter region [1]. Despite the extensive research that has already been conducted, the precise mechanism by which MADS-box TFs regulate their target genes is not known [2–9]. Extending previous research with machine learning techniques could provide new insights. In this way, we can increase our knowledge regarding these proteins and benefit both fundamental and applied research areas.

The MADS-box TF protein family consists of 107 members in the *Arabidopsis* genome and can be further subdivided into two types based on their structure. The type I MADS-box proteins encompass more than half of the *Arabidopsis* MADS-box proteins and contain a conserved MADS DNA-binding domain. Further research still needs to be conducted to unravel the function of the type I MADS-box proteins, in contrast the type II MADS-box proteins are studied in more detail [10]. They are characterized by four conserved protein domains: MADS (M), intervening (I), keratin-like (K) and C-terminal domain (C) [11,12]. The MADS-domain is a DNA-binding domain with a high affinity for the CArG-box, defined by the consensus sequence [CC(A/T)$_6$GG]. The other domains are mainly involved in protein-protein interactions, in which the K-domain facilitates dimerization between MADS-box proteins. The C-domain is considered to play a role in the formation of multimeric MADS-box protein complexes [12]. The formation of complexes between MADS-box proteins and other types of proteins, is one of the important underlying properties of the MADS-box proteins [13–15]. In the formation of these complexes the MADS-box protein SEPALLATA (SEP) seems to play an important role. SEP acts as a connecting component in a wide range of multimeric complexes and is involved in the looping of DNA [14,16]. The formation of MADS-box proteins complexes has been partly described in the floral quartet model, in which specific combinations of MADS-box hetero- or homo-dimers specify the formation of different whorls [17]. Besides the interactions described in the floral quartet model other interactions between non-MADS-box TF, co-repressors and chromatin remodeling factors have been found [13,14,18]. The formation of these complexes is important in the determination of DNA-binding specificity and the regulation of target genes [19].

The binding of a TF to a TFBS dictates which biological processes a TF is involved, therefore it is of great importance that these TFBSs are studied in detail. Chromatin immunoprecipitation sequencing (ChIP-seq) is one of the techniques to study the affinity of a TF towards a TFBS *in vivo* [20]. For the MADS-box proteins different ChIP-seq datasets are available and a recent meta-analysis was performed by Aerts *et al.*. In this analysis ChIP-seq data of the MADS-box proteins: AGAMOUS (AG), APETALA1 (AP1), APETALA3 (AP3), FLOWERING LOCUS C (FLC), PISTILLATA (PI), SEPALLATA (SEP3), SUPPRESOR OF OVEREXPRESSION OF CO 1 (SOC1) and SHORT VEGETATIVE PHASE (SVP) were combined and processed in uniform manner [2–9]. Their analysis showed that all of the TFs had a high affinity for CArG-boxes. Nonetheless, found motifs that contained a CArG-box were not identical and partly specified by an extra di- or tri-nucleotide addition to the CArG-box [2]. Additionally, other studies have shown that DNA shape; which describes the physical structure of DNA, plays an important role in the specification of TF binding [21,22]. For instance, the A-tract is specific for the CArG-boxes of the MADS-box proteins [21]. Although the TFBSs of the previous mentioned MADS-box proteins share a similar motif, they are not involved in the same processes. As can be seen for the MADS-box proteins AP1, AP3, PI and SEP3, which are mainly known for their role during floral organ development. This role has been extensively described in the genetic ABC(D)E model, in which every letter represents a group of genes involved in the development of different whorls [23,24]. In contrast, the MADS-box proteins FLC, SVP and SOC1 are involved in the flowering of the plant [25–28]. The fact that these MADS-box proteins differ in function and that the CArG-box is conserved raises the question which other features, beside the motif sequence, plays an important role in the binding of a MADS-box protein to a DNA-sequence [2]. Traditionally TFBSs are described using a Position Weight Matrix (PWM) that is constructed with features describing base pair occurrence alone. However, a recent study showed that not all of the variance in a TFBS is described in a PWM [29]. Improved methods included different features related to: base pair dependencies, DNA accessibility, DNA shape and chemical structures [22,30–32]. By taking these features into account these methods are not able to describe a TFBS in more detail and also be more accurate in their TFBS prediction [33]. Therefore it is of utmost importance to consider different features while studying TFBSs.

The field of machine learning is able to combine and determine the importance of multiple features related to the TFBS [34]. In this study we applied a random forest (RF) classification model to predict the relevance of different features related to the TFBS site of the MADS-box proteins AG, AP1, AP3, FLC, PI, SEP3, SOC1 and SVP. Processed ChIP-seq data from these TFs were already available from the study of Aerts *et al*. [2]. To extend the analysis of Aerts *et al.* we trained multiple RF models and predicted binding of a TF to a DNA region found by ChIP-seq with features related to: DNA properties (motifs, DNA shape, Local Composition complexity (LCC), melting temperature, GC-content and methylation profile), Gene Ontology (GO)-terms, shared peak regions between the MADS-box proteins and shared regions between other TFs. Our analysis shows that the binding affinity of each MADS-box protein to a peak region, is determined by present motifs and/or sharing of the peak region with other proteins. This confirms previous research which showed that the specificity of the MADS-box proteins is determined by interaction with other proteins [13–16]. Additionally we found new candidate proteins that possibly interact with our MADS-box proteins. We obtained models for each MADS-box TF with an AUC ranging between 0.5-0.8, which indicates that we are able to explain some of the binding affinity of the MADS TFs with the selected features. In summary this study provides insights into which features are important in this prediction and present a new approach in predicting TF binding with ChIP-seq data.

Material and Methods

Features and datasets were generated in Python 2.7. Analysis of the dataset was conducted with Python 3.0 in Jupyter Notebook v4.4.0. Scripts can be found in supplementary 7.

Generation of binding and non-binding dataset

Binding and non-binding datasets were created with the datasets of Aerts *et al.* [2]. This dataset consisted out of eight combined ChIP-seq datasets from the TFs: AG (n = 897), AP1 (n = 789), AP3 (n = 1237), FLC (n = 59), PI (n = 2156), SEP3 (n = 4447), SOC1 (n = 301) and SVP (n = 445) (16-22). For each TF an additional set of non-binding peaks was created. Throughout this report we consider the peaks of the original dataset as binding peaks. To generate a non-binding peak set, all of the binding peaks were pooled and peaks were randomly selected from this pool. To prevent that the same peaks of the binding set occurred in the non-binding set, we excluded the binding peaks from the non-binding peak pool (figure 1). Non-binding peak sets were constructed in such way that the size equals the number of binding peaks. After the non-binding sets were created, the eight different datasets were split into a train and a test set using a fivefold cross-validation (CV). CV was performed with the function Kfold from the package sklearn v0.20.0 [35].

Calculation of the features

For each peak in the binding and non-binding dataset different features were calculated (figure 1). We classified the different features in feature groups based on their properties, see supplementary 1 to see which feature belongs to which group.

Motifs

FASTA files were masked for low complexity DNA and interspersed repeats with RepeatMasker v4.0.8, as specified in Aerts *et al.* [2,36]. From these files we used the train sets of the binding peaks to search for motifs with the tool MEME-ChIP v5.0.3 (for settings, see *Aerts et al.*) [2,37]. Found motifs were filtered on central enrichment, with the help of CentriMO (automatically ran in the MEME-ChIP pipeline) [38]. In CentriMO we defined centrally enriched motifs as motifs with an adjusted p-value lower than 0.05 and a E-value lower than 0.05. Additionally, we included secondary motifs found by SpaMO (automatically ran in the MEME-ChIP pipeline) [39]. SpaMO conducts a spaced motif analysis, in which the null hypothesis is that the occurrence of the secondary motif with primary motif is by chance. Motifs with a p-value lower than 0.05 were included into the analysis. The secondary and centrally enriched motifs were searched with FIMO in the train and test set of the binding and non-binding peaks. Motifs with a q-value lower than 0.05 were recognized as occurring motifs in the dataset. The motif feature was expressed as a binary vector, in which one indicates motif present and zero indicates motif not present.

Found motifs with an E- and q-value lower than 0.05 were assigned to a TF family with the tool TOMTOM v5.0.3, in which we searched for target motifs in the organism *Arabidopsis thaliana* with the data of the DNA affinity purification sequencing (DAP-seq) experiment of O'Malley *et al.* [40,41].

DNA-shape

The DNA-shape was calculated with the R library DNAshapeR v3.8 in R v3.5.1 [42]. Every DNA-shape was calculated per base pair. To implement DNA-shape into our model we calculated the mean, minimum, maximum, standard deviation (SD) and mode for each DNA shape in every peak region with the NumPy package for Python.

Local Composition complexity (LCC), melting temperature and GC-content

LCC, melting temperature and GC-content were calculated with the package Biopython v1.72 [43]. These features were calculated for each peak region.

Methylation

Methylation of a certain peak region was predicted with the previously found different methylated regions (DMR) in the study of Kawakatsu *et al.* [44]. In this study 1107 methylomes of different *A.thaliana* ecotypes were analysed and DMRs in different regions of the *A.thaliana* genomes were found. These DMRs were classified into three groups: CG-DMR (DMRs only in CG context), CH-DMR (DMRs in CHG/CHH context) and C-DMR (DMRs in CG-context and CHG/CHH context). For each group it was determined which percentage of the peak region was located in such region. The found percentage was used as features in the classification.

Shared peak regions

Shared peak regions between MADS-box proteins were determined by comparing the peak regions of the MADS-box proteins with each other. This was done by converting the FASTA files into a bed file and intersect the files with the intersection function of bedtools v2.25.0 [45]. The same procedure has been applied between the peak regions of the different MADS-box proteins against additional DAP-seq and ChIP-seq dataset of O'Malley *et al.* and Heyndrickx *et al.* [40,46]. Since all of the peak regions of the MADS-box proteins had a range of 500bp we adjusted the peak regions from these experiments to a range of 500bp. If a peak shared a region with another peak this was registered in a binary vector, in which one indicates that the peak regions are shared and zero indicates that peaks regions that are not shared.

Quantification of shared peak regions and motifs

Shared peak regions were quantified by counting the occurrence of overlap per base position and adding these counts, resulting in a histogram. A similar thing was done for the motifs, in which we counted the occurrence of a specific motif per base found by FIMO. See supplementary SX for results.

GO-terms

GO-terms were assigned to a ChIP-seq peak, by searching for the most nearby gene related to that peak region. To search for a nearby gene, a GFF file containing all gene locations of the *Arabidopsis* genome, was converted to a bed file. The same was done for all FASTA files containing the peak regions of our dataset. The two BED files were intersected with each other, with the intersection function of bedtools v2.25.0 [45]. For the found genes we searched the corresponding GO-terms with the Python package GOATOOLS v0.8.9 [47]. This was done with the script wr_hier.py included in the package. With the help of this script we up-propagated the GO-term till the highest level. As input we used the obo file of *A.thaliana*, retrieved from the GO consortium (downloaded at 03-10-2018) [48]. Obtained GO-terms in the binding peak set used for training, were used as feature for the analysis. Presence of GO-terms were expressed in a binary vector, in which one indicates that the peak region is associated with that GO-term and zero indicates that the GO-term is not associated with that peak region.

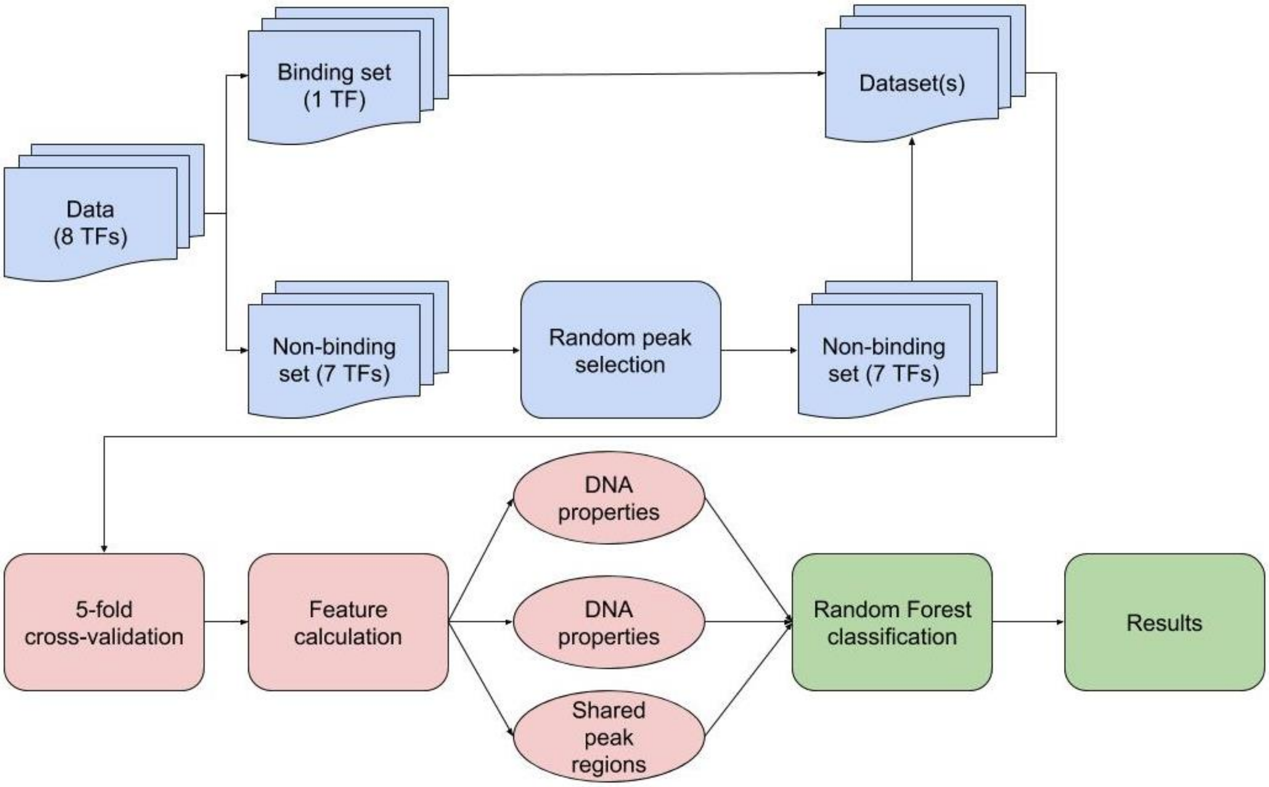

***Figure 1 Flow diagram of material methods**. First a binding and non-binding dataset was created for each dataset (blue). With these datasets a fivefold CV was performed and features were calculated per CV (red). The features were used to train RF models to classify binding from non-binding peaks (green).*

Next to linking the most nearby genes with GO-terms, we also performed a GO-enrichment analysis with the tool AgriGO. For this tool we applied the default settings, in which a Singular Enrichment Analysis (SEA) is applied on the input genes, followed by a Benjamini–Yekutieli correction, to correct for multiple testing [49,50].

Data analysis

Dataframes containing the previous mentioned features were used to train different RF models in Jupyter Notebook with the sklearn package for each TF separately (figure 1). Before the RF models were trained, the feature describing overlap with SEP3, and the feature describing overlap of each MADS TF with itself, were excluded from the models. Additionally, we excluded the feature FLM from the FLC model, because they are closely related to each other. The optimal parameters of the models were determined with the grid search of sklearn [35]. To perform the grid search, we used a range of fifty till thousand in steps of fifty to determine the optimal size of the forest (n_estimators). For the minimal leaf sample size, we took a size of one twentieth of the total amount of peaks in the dataset (binding and non-binding peaks included). In the grid search function, the obtained parameters were cross-validated with a fivefold CV. The parameters found in this way were used in the RF classifier and the RF was trained on the different train sets. Performance of the different classifiers was calculated by first constructing a response operation curve (ROC) and calculating the area under the curve (AUC) of the ROC curve. This was done with the functions available in sklearn [35].

Results

Model performances varies among different feature groups of different TFs

In the study of Aerts *et al.* it was found that MADS-box proteins: AG, AP1, AP3, FLC, PI, SEP3, SOC1 and SVP share a highly similar TFBS. Despite this, each of these proteins is necessary for the regulation of different processes. To define what discriminates these TFs in their binding, we continued the previously mentioned analysis, by training multiple RF models. These models classified ChIP-seq peak regions into binding and non-binding peaks. Throughout this report we define a binding peak, as a peak that occurs in the original dataset and a non-binding peak as any other peak of the seven remaining MADS-box TFs. These binding and non-binding peaks were used for our classification model, to distinguish peaks from one MADS-box TF from the other seven TFs. Binding of a TF to a peak region was predicted with features related to: DNA properties, GO-terms, sharing peak regions of the TF with other TFs in the dataset itself and/or sharing regions with other TF factors (see supplementary S1 for the features in these groups). Training was done for each feature group separate and with multiple combinations of the feature groups. For the trained models we performed a five-fold CV and calculated the AUC for each model. From the calculated AUC we observed that the performances of the different models vary among different MADS-box proteins (figure 2A). For the GO-term models we can observe that the AUC for every model is around 0.5, which equals prediction of binding by guessing. On the other hand, the models containing the DNA-features have an AUC of ~0.6 for most of the TFs and even ~0.7 for SVP. Additionally, the features in this group related to motif sequences, have the highest feature importance's. The range of the AUC of the models describing shared peak regions between the MADS-box proteins themselves ranges from ~0.5-0.8, implying that the importance of shared peak regions between MADS-box proteins is specific for each MADS. The MADS-box proteins AG, AP3, FLC, SOC1 and SVP had a high AUC for the models using features describing shared peak regions, whereas AP1, PI and SEP3 had a low AUC for these models. For the shared peak regions between other TFs, an AUC in the range of ~0.5-0.7 was found. Compared to the other TFs, the TFs SOC1 and SVP had a high AUC, which suggests that these TFs have some unique interactions with other TFs, which are not shared by the other MADS-box proteins. Furthermore, it was observed that the combination of different kind of features did not improve the AUC of the models.

Next to model performance, we also studied the feature importance's, by training a model containing all features. This model was trained on the whole dataset, instead of a separate train and test set. The importance's among features are distributed in most of the models, although some features stand more out than others (figure 2B). This was observed across different TFs and for features related to DNA properties, overlapping regions between MADS-box TFs or overlapping regions between other TFs. Furthermore, the motif features seem to be the most important features in the feature group DNA properties. It is notable that the model performance of the different feature groups complements the feature importance. In general, we observed across the different CVs that the top two features are the most robust in the model with all the features included. Together with the model performances we can observe that the properties of the DNA and sharing regions between MADS-box proteins contribute the most in the classification of the different MADS.

Motif occurrence differs between binding and non-binding dataset

The RF models that predicted binding of a TF based on DNA properties varied in AUC among different TFs. Furthermore, we found that most of the performances were determined by features that described the motifs. Therefore, we counted the occurrence of each motif from our model in the binding and non-binding datasets separately, which helped us determine which motifs are specific for a MADS-box TF. Our analysis showed

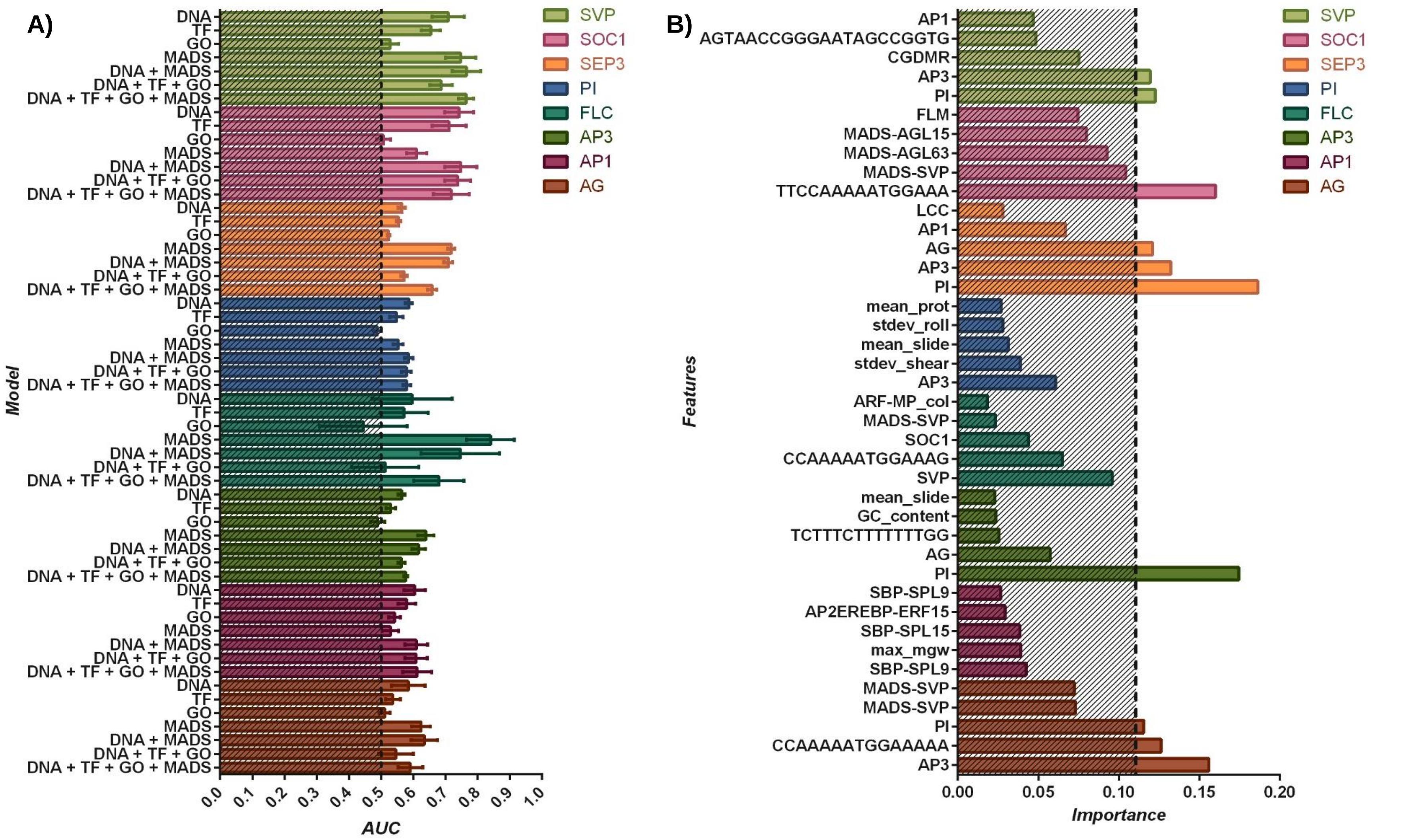

Figure 2 Model performance and feature importance of the different datasets. *A) AUC of RF classification models, the y-axis show which feature groups are included into the model. The x-axis of the plot shows the average AUC per model across the CVs, in which the error bars represent the SD. The dotted line indicates an AUC of 0.5. B) Top five of most important features per TF, the dotted line indicates the boundary in which the features were the most constant.*

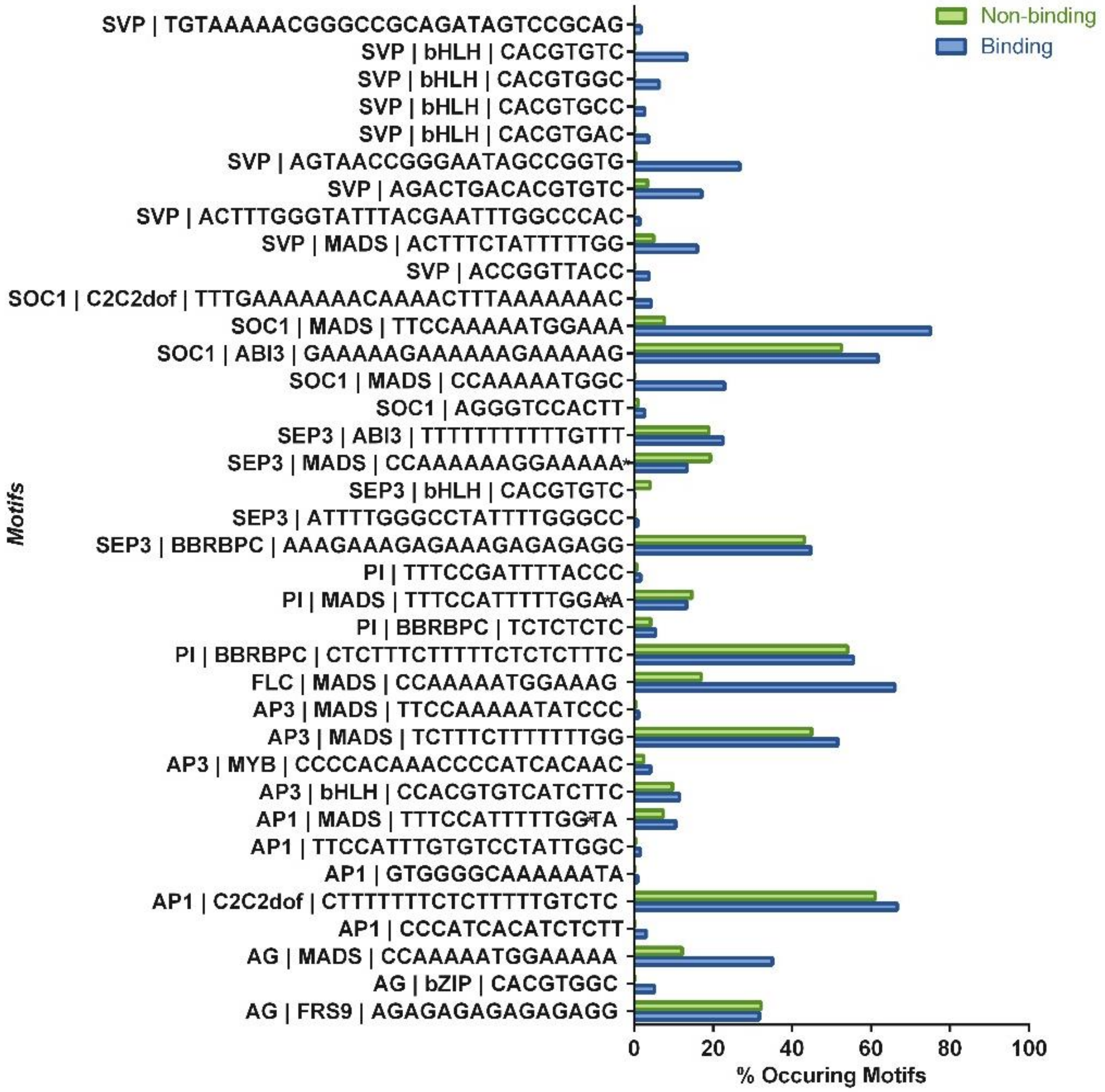

Figure 3 Motif occurrence in binding and non-binding peaks. *Motif occurrence between binding and non-binding varies. On the x-axis the percentage of the total amount of peaks in the dataset containing the motif is shown. On the y-axis the different motifs are shown in which the label information is described as follows: TF name* **|** *TF family* **|** *motif consensus sequence.*

that occurrence of motifs containing a CArG-box differed greatly between the non-binding and binding peaks in the datasets of AG, FLC, SOC1 and SVP (figure 3). From these motifs, the CArG-box motifs of AG and SOC1 and multiple non-CArG-box motifs of SVP seem to be important for our classifiers (figure 2B). To assign the non-CArG- and CArG-box motifs to a TF family we used TOMTOM [41]. From this analysis we observed that motifs belonging to the TF family basic helix loop helix (bHLH), BARLEY B RECOMBINANT/ BASIC PENTACYSTEINE (BBRBPC) and C2C2dof occurred among multiple TFs. Additionally, most of the other found motifs were not unique for the binding dataset. In exception to the motifs found in datasets of SVP and SOC1. Both of these TF bear unique motifs that occur in more than 20% in the binding dataset and in very small amounts or even fully absent in the non-binding dataset.

The location of the motifs in the peak region were also analysed, this was done by scoring the occurrence of the motifs found by FIMO. Despite for selecting of centrally enriched motifs, some found motifs deviated from the middle of the peak (supplementary S2). Nonetheless, it seems that most of the CArG-boxes are centred in the peak region, whereas non-CArG-box motifs are scattered around the peak region.

Overlap between different MADS TF determines some specificity

Our RF models describing the overlapping peak regions between MADS-box proteins, are able to indicate possible interactions between MADS-box proteins. From our models we learned that the MADS-box proteins AG, AP3, FLC, SOC1 and SVP had a high AUC. In contrast the TFs AP1, PI and SEP3 had a low AUC, indicating that our overlapping features are less capable of describing the binding of these TF to a ChIP-seq region. To further define the importance of specific MADS-box proteins in our model we trained the same models based on the whole dataset, instead of a specific train and test set. During the training of the models we excluded the feature SEP3, since most of the peaks in the non-binding dataset consisted out of peaks originating from the SEP3 dataset. In the results we can observe that the peak sharing between, AG-PI, AG-AP3, AP3-PI, FLC-SVP, SEP3-PI, SOC1-PI, SVP-PI, had an importance of 0.4 or higher (figure 4A). In the previously mentioned combinations, the first TF refers to the dataset and the second TF refers to the feature of that particular dataset. Interestingly, combinations with higher feature importance's are specific for a certain dataset (e.g. between SVP-FLC, the importance of FLC is low), this indicates that possible interactions are specific per TF. To further determine whether some of these interactions were already found in the literature, we annotated our results with the physical interactions stated in the BioGRID database [51,52]. From this annotation we can learn that the obtained physical interactions do not fully match with our results. Although the interactions AG-AP3, AG-PI, AP3-PI, FLC-SVP, SEP3-PI were found to physically interact and had a feature importance above 0.4. Besides physical interaction, other interactions may also occur that are described by the absence of the peaks. To exclude whether this was the case, we counted the amount of overlapping regions between the other TFs (figure 4B). In general, the features

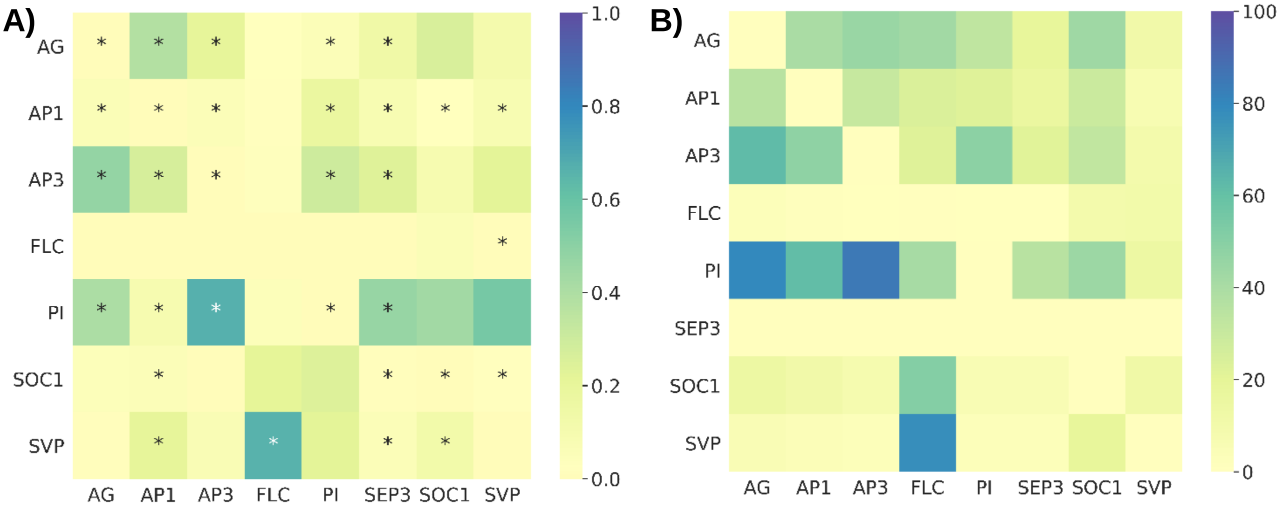

Figure 4 Feature importance and quantity of the MADS-box shared regions features. *In this figure the vertical axes describe the features and the horizontal axes describe the datasets. A) Importance of features describing the overlapping regions. B) The percentage of overlapping peaks in the binding datasets. *Physical interaction found in the BioGRID database [51, 52].*

seem to have the same percentage of overlapping peaks in each dataset, except the peaks between FLC-SOC1 and FLC-SVP appear to be higher. Furthermore, it seems that feature importance is determined by presence of overlapping regions.

Shared regions between other TFs seem to be important for TF prediction

Our models that predicted binding based on shared regions between different MADS-box proteins gave high AUC scores. Therefore, we compared the peak regions of our MADS-box proteins with the peak regions of TFs from two other studies [40,46]. The AUC of these models differed and were especially high for the TFs SOC1 and SVP. To further define the importance of specific TFs in our model we trained the same models based on the whole dataset, instead of a specific train and test set. Features with an importance of 0.1 or higher across all MADS TFs, were considered as most important. From our analysis we observed that the feature importance's (figure 4A) in these models were not high. The highest feature importance was observed in the features related to the other MADS-box proteins AGAMOUS LIKE (AGL)15, AGL63 and FLC, indicating that a possible interaction mainly exists between the TF family members itself. Furthermore, we found that the feature importance of AGL15 in the DAP- and ChIP-seq experiment were almost identical. However, if we count the amount of overlapping regions (figure 4B) between the TFs we can observe that it is considerably less in the DAP-seq experiment. On top of this we also annotated our results, with the physical interactions in the BioGRID database, which showed that the physical interactions found in the BioGRID database did not complement our predicted interactions (e.g. SEP3-FLOWERING LOCUS M (FLM), AP1-SQUAMOSA PROMOTER BINDING PROTEIN-LIKE 9 (SPL9), AG-AGL63, SOC1-AGL63) [51,52]. At last, it seems that the previous mentioned combinations across the different CVs were not consistent, indicating that some of these predictions are highly dependent on some specific peaks in the dataset or that some of the features are highly correlated with each other.

From our model we learned which TFs are possibly important in defining whether a MADS-box protein binds to a peak region. Nonetheless, non-important features are still able to indicate a possible interaction between our MADS-box protein and the TFs from the studies of Heyndrickx *et al.* and O'Malley *et al.* [40,46]. Therefore, we counted the amount of overlapping peaks of these features (figure 4C) and we observed that the TF family C2C2dof had a high amount of sharing regions between our MADS-box proteins. Furthermore, we retrieved most of the features with high feature importance's back in the DNA regions with more than 40% of overlapping.

Gene regulation of different MADS are highly similar

The model performances showed that GO-terms do not contribute to an increase in model prediction. To see what the cause of this low model performance was we performed a GO-enrichment analysis with the tool AgriGO [49,50]. From this analysis we can observe that the different enriched GO-terms of the different TFs are highly alike (supplementary S9). Although some different GO-terms are present between the different TFs, this difference was not high enough to generate a distinct classification between binding and non-binding peaks. Moreover if we calculate the percentage of the GO-terms assigned to a dataset, we observe the same. The percentage of GO-terms are similar among the different datasets (supplementary S5).

Discussion

In this study we examined the TFBSs of eight different MADS-box proteins: AG, AP1, AP3, FLC, PI, SEP3, SOC1 and SVP. Previously it was found, that these TFs shared highly similar motifs, while they are involved in different processes [2]. To determine what defines the binding specificity of these TFs, we trained different RF models with features related to: DNA properties, GO-terms,

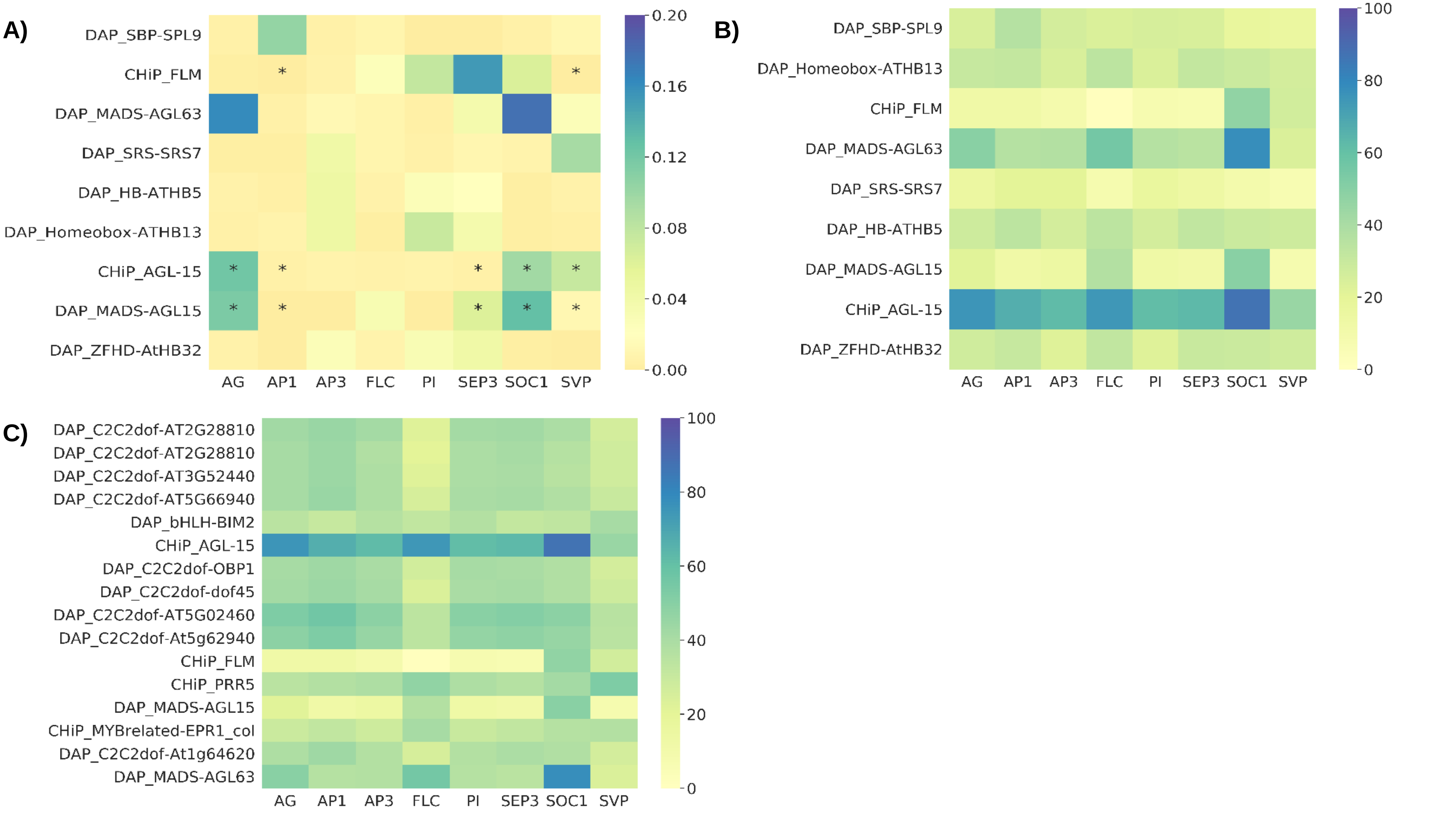

Figure 4 Feature importance and quantity of the TF shared regions features. In this figure the vertical axes describe the features and the horizontal axes describe the datasets. A) Importance of features describing the overlapping regions. B) The percentage of overlapping peaks in the binding datasets of the most important features. C) The percentage of overlapping peaks in the binding dataset, in which overlapping occurred in more than 40% of the feature. *Physical interaction found in the BioGRID database [51,52].

sharing peak regions of the TF with other MADS TFs and sharing of peak regions between other TFs. The performances of these models were different among the MADS-box TFs, indicating that the binding properties of the MADS-box TFs cannot be explained by one feature group. Interestingly, the differences are mostly present in feature groups describing DNA properties and the overlapping of peak regions between TFs. These findings mean that both DNA-TF and TF-protein interactions are important in the binding of MADS-box proteins to TFBSs [19].

Models that only consisted of features related to DNA properties, varied in their performances. These performances were mainly influenced by features defining the presence of a motif. In our analysis both CArG- and non-CArG-box motifs were found, this is not surprising since the presence of these motifs has been stated in other studies [2–9]. However, in our study we show that absence or presence of both CArG- and non-CArG-box motifs partly defines the binding of a specific MADS-box protein in a ChIP-seq peak region. Interestingly, some CArG-box motifs occur more often in the binding set and less in the non-binding set, which indicates a difference in CArG-box motif specificity among the MADS-box TFs. In addition, the finding of non-CArG-box motifs suggested possible interactions of the MADS-box proteins between TFs of other TF families. These interactions can be important in the regulation of different biological processes [53]. An important consideration is that the occurrence of multiple motifs may be a property of the ChIP-seq technique. This is mainly due the fact that the technique is not able to distinguish single TFs from TF complexes. As a consequence sequences related to other TFs may be found, which is likely for MADS-box proteins, since the nature of MADS-box proteins is to form di-/multimeric complexes [15,16,20,24]. In our study we expanded the traditional ChIP-seq analysis, by studying TF-TF interactions. Therefore the property of the ChIP-seq technique to take multimeric complexes into account is in our advantage.

In contrast to the motif features, other features describing the DNA properties of a peak region had low feature importance's. This low contribution can be explained by three different reasons: first, the variance of the features could be too sparse; second, the variation of these features could be mainly located in or around the motif region, which is considerably smaller than the peak region; lastly, the peak regions of all the MADS-box proteins may be too alike in DNA-sequence composition. Therefore, it is impossible to classify all of the peaks correctly with this set of features. To better understand the relationship between MADS-box proteins and the DNA properties in a peak region, different models could be trained. For example, instead of ChIP-seq regions, motifs could be classified by using DNA properties as features. In this way, specific DNA properties might be linked to a motif. Another interesting experiment would be to study the DNA-shape in more detail: although these studies are mostly done with Systematic evolution of ligands by exponential enrichment (SELEX) or Protein Binding Micro-array (PBM) data, it has been shown that DNA-shape is able to improve the prediction of motif sequences [19,22,29,54–56]. To study the DNA shape of different motifs with ChIP-seq, the lack of specificity can be overcome by focussing on the regions where di-/multimeric complexes are located and identify specific shapes in these complexes.

Models that consisted of feature groups describing the overlapping regions between other TFs, had a high AUC. Particularly the models that described the overlapping regions between MADS-box proteins showed interactions stated in the literature [7,13,57]. However, not all physical interactions in the BioGRID database, were found to be in important in our RF classification model [51,52]. The low feature importance of these TFs can be caused by a similar overlapping pattern between our MADS-box TFs. This seems reasonable since previous research and our research confirmed that these particular TFs have a high amount of shared regions between each other [58]. Furthermore, some models have a low AUC for the MADS-box proteins, specifying that the overlapping pattern is unique for some TFs. Next to already known interactions, we also predicted interactions (PI-SVP and PI-SOC1), that were not observed experimentally in the study of Folter *et al.* [10]. An explanation for the predicted interactions between PI-SVP and PI-SOC1, is the overrepresentation of the PI peaks in the non-binding dataset. During the creation of the non-binding set, peaks are randomly picked from a pool, consisting of peaks from other MADS-box TFs. The TFs are represented in different quantities in the pools, causing an unbalanced representation of the TFs in the non-binding dataset. For this reason, we excluded SEP3 as a feature from our analysis, but the results point out that PI should also be excluded from our analysis.

Next to the overlapping regions between the MADS-box TFs, we also studied the overlapping regions of the MADS-box TFs with TFs from other studies [40,46]. From these results, only the interaction between AGL15 and the other MADS-box proteins have been confirmed in literature [57,59]. Other interactions were predicted by our model (e.g. FLM-SEP3, AP1-SPL9), but no physical interactions among these have been shown in the literature before. Moreover, the study of Folter *et al.* has shown that no physical interaction exists between our MADS-box proteins and AGL63, while an interaction between AG, SVP and AGL63 was predicted in our model [10,60]. The possible interactors are involved in different processes such as: root growth (ATHB13), flower related developmental processes (SHI-RELATED SEQUENCE 7 (SRS7)), AGL63, AGL15, ATHB32), flowering (FLM) and/or other plant development related processes (ATHB13, SPL9, ATHB5) [61–70]. Besides these functions we found that these proteins are expressed during the same developmental stages and in the same tissues (56, supplementary 6). It is important to take into consideration that most of the AUCs of this particular model were low, which means that there is not much predictive power in the data from these proteins to predict MADS-box protein binding specificity. Besides aforementioned findings, we also observed that TFs from the C2C2dof family, known for their wide array of functions, often co-occurred with our peak regions (60). Moreover, non-CArG-box motifs related to this TF family were found in our analysis, indicating a possible interaction between the MADS-box proteins and the TFs of C2C2dof family.

From our GO-term model we learned that the whole dataset cannot be assigned to specific gene functions. By focussing on peak areas of certain protein complexes instead of peak areas of single MADS proteins we may be able to find new complex compositions related to certain GO-terms, providing new insights into the function of MADS-box proteins. Additionally, we may be able to identify new TF complexes not found in literature before. We did a preliminary experiment, in which a classifier was trained to distinguish overlapping regions for a pair of MADS proteins from non-overlapping regions. Interestingly some signal seems to be present, but further research needs to be conducted to substantiate this (supplementary S6).

In our study we found possible interactions, between MADS-box proteins and other TFs, which have not been described in literature before. However, most of these TFs are not known to interact with the MADS-box proteins from our dataset. An explanation for this finding, are the conditions of the *in vitro* DAP-seq experiment of O'Malley *et al.*, in this experiment single TF are tested on DNA isolated form leaf tissue introducing two biases [40]. First, the MADS-box proteins are involved in processes that are located in the floral organs. Therefore they are mainly expressed during plant development and in the mature floral tissue, hereby possibly binding to different targets in leaf tissue (supplementary X) [71]. Second, MADS-box proteins act in protein complexes, while the DAP-seq study only focusses on a single TF per interaction. Another explanation for our findings, is that the proteins do not directly interact with our MADS-box proteins and cannot be found by experiments describing interaction between two TFs (for example yeast two hybrid). Nonetheless, this interaction may be present in a complex of TFs via a so called bridging protein. On the other hand it may be involved in the inhibition of a TF complex formation. Therefore no physical interaction between the TFs is present, but rather a genetic interaction. To decide whether these proteins have a function during development more research has to be conducted. To confirm whether the suggested interactions are valid, these interactions should be further studied *in vitro* or *in planta*.

To further research the TFBSs, different things can be done, e.g. models could be further improved by adding more features or adjusting the dataset. Possible features that could be added are features describing: the distance between motifs or by adding features describing other elements e.g. tandem repeats and CpG islands [19,72,73]. Next to adding features, the dataset can also be expanded by adding more ChIP-seq data from other MADS-box proteins. Furthermore, adding SELEX-data to the model may improve model performances, because SELEX-data specifically describes the interaction of a TF *in vitro*. Therefore, such data is more able to describe the affinity of a TF to a specific DNA shape and sequence. The motifs obtained with SELEX-data can be included in our model and might provide an improved prediction for our TFs. In addition to changing the type of data, we can also consider to use another algorithm to find more accurate motifs. The disadvantage of MEME-ChIP is that it does not take dependencies into account, while it has been shown that not all of the variance present in a motif is explained in a single base pair [19,29]. Therefore we are excluding important information in determining our motif sequences, the use of another algorithm may overcome this problem [22,74].

In summary, we show that different features are able to predict the binding of a MADS-box protein to a specific ChIP-seq peak region. From these features, the features describing the motifs and sharing peak regions between TFs have the highest importance's. This suggests that the regulation of MADS gene targets is mediated by specific motif sequences and interactions between MADS and other TFs. In addition to these findings, we present a novel approach to investigate a TFBS of a TF, by modelling TFBSs we are able to obtain insights into TF interactions and their behaviour.

Acknowledgements

Aalt-Jan, for supervising me during my thesis and giving me the opportunity to work on this project. It was great to share thoughts with each other and thank you for the things you have learned me. Gerco Angenent, for the biological insights into the results. Hernando Suarez Duran and Kirby Dick for the feedback they gave me on my report. At last, I would like to thank everyone in the bio-informatic department for their support.

References

1. Messenguy, F., Dubois, E., Bruxelles, L. De & Gryzon, A. E. Role of MADS box proteins and their cofactors in combinatorial control of gene expression and cell development. **316,** 1–21 (2003). doi:10.1016/S0378-1119(03)00747-9
2. Aerts, N., Bruijn, S. De, Mourik, H. Van, Angenent, G. C. & Dijk, A. D. J. Van. Comparative analysis of binding patterns of MADS-domain proteins in Arabidopsis thaliana. 1–16 (2018). doi:10.1186/s12870-018-1348-8
3. Maoiléidigh, D. S. Ó. *et al.* Control of Reproductive Floral Organ Identity Specification in Arabidopsis by the C Function Regulator AGAMOUS. **25,** 2482–2503 (2013). doi:10.1105/tpc.113.113209
4. Pajoro, A. *et al.* Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. (2014). doi:10.1186/gb-2014-15-3-r41
5. Kaufmann, K. *et al.* Orchestration of Floral Initiation by APETALA1. 85–90 (2010). doi:10.1126/science.1185244
6. Wuest, S. E. *et al.* Molecular basis for the specification of floral organs by. (2012). doi:10.1073/pnas.120707510
7. Mateos, J. L. *et al.* Combinatorial activities of SHORT VEGETATIVE PHASE and FLOWERING LOCUS C define distinct modes of flowering regulation in

Arabidopsis. 1–23 (2015). doi:10.1186/s13059-015-0597-1

8. Jauregui, R. *et al.* Target Genes of the MADS Transcription Factor SEPALLATA3 : Integration of Developmental and Hormonal Pathways in the Arabidopsis Flower. **7,** (2009). doi:10.1371/journal.pbio.1000090
9. Immink, R. G. H. *et al.* Characterization of SOC1 ' s Central Role in Flowering by the Identification of Its Upstream and Downstream Regulators 1. **160,** 433–449 (2012).doi:10.1104/pp.112.202614
10. Folter, S. De *et al.* Comprehensive Interaction Map of the Arabidopsis MADS Box Transcription Factors. **17,** 1424–1433 (2005). doi:10.1105/tpc.105.031831.1
11. Kaufmann, K. & Melzer, R. MIKC-type MADS-domain proteins : structural modularity , protein interactions and network evolution in land plants. **347,** 183–198 (2005). doi:10.1016/j.gene.2004.12.014
12. Folter, S. De *et al.* Molecular and Phylogenetic Analyses of the Complete MADS-Box Transcription Factor Family in Arabidopsis : New Openings to the MADS World Lucie Pa r. **15,** 1538–1551 (2003). doi:10.1105/tpc.011544
13. Smaczniak, C., Immink, R. G. H., Muiño, J. M., Blanvillain, R. & Busscher, M. Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. (2012). doi:10.1073/pnas.1112871109
14. Melzer, R. Reconstitution of ' floral quartets ' in vitro involving class B and class E floral homeotic proteins. **37,** 2723–2736 (2009). doi:10.1093/nar/gkp129
15. Melzer, R. & Verelst, W. The class E floral homeotic protein SEPALLATA3 is sufficient to loop DNA in ' floral quartet ' - like complexes in vitro. **37,** 144–157 (2009). doi:10.1093/nar/gkn900
16. Immink, R. G. H. *et al.* SEPALLATA3 : the ' glue ' for MADS box transcription factor complex formation. **10,** (2009). doi:10.1186/gb-2009-10-2-r24
17. Saedler, G. T. and H. Floral quartets. *Nature* **409,** (2001). doi:10.1038/35054172
18. Honma, T. & Goto, K. Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. **409,** 525–529 (2001). doi:10.1038/35054083
19. Smaczniak, C., Muiño, J. M., Chen, D., Angenent, G. C. & Kaufmann, K. Differences in DNA Binding Specificity of Floral Homeotic Protein Complexes Predict Organ-Specific Target Genes. **29,** 1822–1835 (2017). doi:10.1105/tpc.17.00145
20. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10,** 669–680 (2009). doi:10.1038/nrg2641
21. Muiño, J. M. *et al.* Structural determinants of DNA recognition by plant MADS-domain transcription factors. **42,** 2138–2146 (2014). doi:10.1093/nar/gkt1172
22. Zhou, T. *et al.* Quantitative modeling of transcription factor binding specificities using DNA shape. 1–6 (2015). doi:10.1073/pnas.1422023112
23. Theißen, G. Development of floral organ identity: stories from the MADS house. 75–85 (2001). doi:10.1016/S1369-5266(00)00139-4
24. Kater, M. M., Dreni, L. & Colombo, L. Functional conservation of MADS-box factors controlling floral organ identity in rice and Arabidopsis. **57,** 3433–3444 (2018). doi:10.1093/jxb/erl097
25. Becker, Annette, G. T. The major clades of MADS-box genes and their role in the development and evolution of flowering plants. **29,** 464–489 (2003). doi:10.1016/S1055-7903(03)00207-0
26. Moon, J. *et al.* The SOC1 MADS-box gene integrates vernalization and gibberellin signals for flowering in Arabidopsis. (2003). doi:10.1046/j.1365-313X.2003.01833.x
27. Tao, Z. *et al.* Genome-wide identification of SOC1 and SVP targets during the floral transition in Arabidopsis. 549–561 (2012). doi:10.1111/j.1365-313X.2012.04919.x
28. Michaels, S. D. & Amasino, R. M. FLOWERING LOCUS C Encodes a Novel MADS Domain Protein That Acts as a Repressor of Flowering. **11,** 949–956 (1999). doi:10.1105/tpc.11.5.949
29. Rube, H. T., Rastogi, C., Kribelbauer, J. F., Bussemaker, H. J. & Rube, H. T. A unified approach for quantifying and interpreting DNA shape readout by transcription factors. 1–16 (2018). doi:10.15252/msb.20177902

30. Zabet, N. R. & Adryan, B. Estimating binding properties of transcription factors from genome-wide binding profiles. **43,** 84–94 (2015). doi:10.1093/nar/gku1269

31. Meysman, P. *et al.* Use of structural DNA properties for the prediction of transcription-factor binding sites in Escherichia coli. **39,** (2011). doi:10.1093/nar/gkq1071

32. Broos, S. *et al.* PhysBinder : improving the prediction of transcription factor binding sites by flexible inclusion of biophysical properties. **41,** 531–534 (2013). doi:10.1093/nar/gkt288

33. Khamis, A. M. *et al.* A novel method for improved accuracy of transcription factor binding site prediction. **46,** (2018). doi:10.1093/nar/gky237

34. Libbrecht, M. W. & Noble, W. S. Machine learning applications in genetics and genomics. *Nat. Publ. Gr.* **16,** 321–332 (2015). doi:10.1038/nrg3920

35. Pedregosa, F., Weiss, R. & Brucher, M. Scikit-learn : Machine Learning in Python. **12,** 2825–2830 (2011).

36. A.F.A. Smit, R. H. & P. G. RepeatMasker. (1990).

37. Machanick, P. & Bailey, T. L. MEME-ChIP : motif analysis of large DNA datasets. **27,** 1696–1697 (2011). doi:10.1093/bioinformatics/btr189

38. Bailey, T. L. & Machanick, P. Inferring direct DNA binding from ChIP-seq. **40,** 1–10 (2018). doi:10.1093/nar/gks433

39. Whitington, T., Frith, M. C., Johnson, J. & Bailey, T. L. Inferring transcription factor complexes from ChIP-seq data. **39,** (2018). doi:10.1093/nar/gkr341

40. Malley, R. C. O. *et al.* Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. **165,** 1280–1292 (2016). doi:10.1016/j.cell.2016.04.038

41. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. **8,** (2007). doi:10.1186/gb-2007-8-2-r24

42. Li, J. *et al.* Expanding the repertoire of DNA shape features for genome-scale studies of transcription factor binding. **45,** 12877–12887 (2017). doi:10.1093/nar/gkx1145

43. Cock, P. J. A. *et al.* Biopython : freely available Python tools for computational molecular biology and bioinformatics. **25,** 1422–1423 (2009). doi:10.1093/bioinformatics/btp163

44. Kawakatsu, T. *et al.* Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. 492–505 (2016). doi:10.1016/j.cell.2016.06.044

45. Quinlan, A. R. & Hall, I. M. BEDTools : a flexible suite of utilities for comparing genomic features. **26,** 841–842 (2010). doi:10.1093/bioinformatics/btq033

46. Heyndrickx, K. S., Velde, J. Van De, Wang, C., Weigel, D. & Vandepoele, K. A Functional and Evolutionary Perspective on Transcription Factor Binding in Arabidopsis thaliana. **26,** 3894–3910 (2014). doi:10.1105/tpc.114.130591

47. Klopfenstein, D. V *et al.* GOATOOLS : A Python library for Gene Ontology analyses. 1–17 (2018). doi:10.1038/s41598-018-28948-z

48. Consortium, G. O. The Gene Ontology (GO) project in 2006. **34,** 322–326 (2006). doi:10.1093/nar/gkj021

49. Du, Z., Zhou, X., Ling, Y., Zhang, Z. & Su, Z. agriGO : a GO analysis toolkit for the agricultural community. **38,** 64–70 (2010). doi:10.1093/nar/gkq310

50. Tian, T. *et al.* agriGO v2 . 0 : a GO analysis toolkit for the agricultural community , 2017 update. **45,** 122–129 (2017). doi:10.1093/nar/gkx382

51. Stark, C. *et al.* BioGRID : a general repository for interaction datasets. **34,** 535–539 (2006).

52. Chatr-aryamontri, A. *et al.* The BioGRID interaction database : 2017 update. **45,** 369–379 (2017). doi:10.1016/j.pbi.2016.09.003

53. Bemer, M., Dijk, A. D. J. Van, Immink, R. G. H. & Angenent, G. C. Cross-Family Transcription Factor Interactions : An Additional Layer of Gene Regulation. *Trends Plant Sci.* **22,** 66–80 (2017). doi:10.1016/j.tplants.2016.10.007

54. Stormo, G. D. Modeling the specificity of protein-DNA interactions. **1,** 115–130 (2013). doi:10.1007/s40484-013-0012-4

55. Rohs, R. *et al.* The role of DNA shape in protein – DNA recognition. *Nature* **461,** 1248–1253 (2009). doi:10.1038/nature08473

56. Stormo, G. D. & Roy, B. DNA Structure Helps Predict Protein Binding. *Cell Syst.* **3,**

216–218 (2016). doi:10.1016/j.cels.2016.09.004

57. Riechmann, J. L., Krizek, B. A. & Meyerowitz, E. M. Dimerization specificity of Arabidopsis MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA, and AGAMOUS. **93,** 4793–4798 (1996).

58. Yan, W., Chen, D. & Kaufmann, K. Molecular mechanisms of floral organ specification by MADS domain proteins. *Curr. Opin. Plant Biol.* **29,** 154–162 (2016). doi:10.1016/j.pbi.2015.12.004

59. Hill, K., Wang, H. & Perry, S. E. A transcriptional repression motif in the MADS factor AGL15 is involved in recruitment of histone deacetylase complex components. 172–185 (2008). doi:10.1111/j.1365-313X.2007.03336.

60. Trigg, S. A. *et al.* CrY2H-seq : a massively-multiplexed assay for deep coverage interactome mapping. **14,** 819–825 (2017). doi:10.1038/nmeth.4343

61. Silva, A. T., Ribone, P. A., Chan, R. L., Ligterink, W. & Hilhorst, H. W. M. A Predictive Coexpression Network Identifies Novel Genes Controlling the Seed-to-Seedling Phase Transition in Arabidopsis thaliana. **170,** 2218–2231 (2016). doi:10.1104/pp.15.01704

62. Tan, Q. K. & Irish, V. F. The Arabidopsis Zinc Finger-Homeodomain Genes Encode Proteins with Unique Biochemical Properties That Are Coordinately Expressed during Floral Development 1. **140,** 1095–1108 (2006). doi:10.1104/pp.105.070565.1

63. Johannesson, H., Wang, Y., Hanson, J. & Engström, P. The Arabidopsis thaliana homeobox gene ATHB5 is a potential regulator of abscisic acid responsiveness in developing seedlings. 719–729 (2003). doi:10.1023/A:1022567625228

64. Erdmann, R., Gramzow, L., Melzer, R., Becker, A. & Group, P. E. GORDITA (AGL63) is a young paralog of the Arabidopsis thaliana B sister MADS box gene ABS (TT16) that has undergone neofunctionalization. 914–924 (2010). doi:10.1111/j.1365-313X.2010.04290.x

65. Ambrose, B. A. The Arabidopsis B-sister MADS-box protein , GORDITA , represses fruit growth and contributes to integument development. 203–214 (2010). doi:10.1111/j.1365-313X.2010.04139.x

66. Capovilla, G., Symeonidi, E., Wu, R. & Schmid, M. Contribution of major FLM isoforms to temperature- dependent flowering in Arabidopsis thaliana. **68,** 5117–5127 (2017). doi:10.1093/jxb/erx328

67. Zheng, Q., Zheng, Y., Ji, H., Burnie, W. & Perry, S. E. Gene Regulation by the AGL15 Transcription Factor Reveals Hormone Interactions in Somatic Embryogenesis. **172,** 2374–2387 (2016). doi:10.1104/pp.16.00564

68. Yu, S. *et al.* Gibberellin Regulates the Arabidopsis Floral Transition through miR156-Targeted SQUAMOSA PROMOTER BINDING – LIKE Transcription Factors. **24,** 3320–3332 (2012). doi:10.1104/pp.16.00564

69. Cui, L., Shan, J., Shi, M., Gao, J. & Lin, H. The miR156-SPL9-DFR pathway coordinates the relationship between development and abiotic stress tolerance in plants. 1108–1117 (2014). doi:10.1111/tpj.12712

70. Fridborg, I., Kuusk, S., Robertson, M. & Sundberg, E. The Arabidopsis Protein SHI Represses Gibberellin Responses in Arabidopsis and Barley 1. (2001). doi:10.1104/pp.010388.sponse

71. Winter, D. *et al.* An ‘Electronic Fluorescent Pictograph’ browser for exploring and analyzing large-scale biological data sets. *PLoS One* **2,** e718 (2007). doi:10.1371/journal.pone.0000718

72. Chen, Y., Wen, Y. & Chang, W. AtPAN : an integrated system for reconstructing transcriptional regulatory networks in Arabidopsis thaliana. *BMC Genomics* **13,** 85 (2012). doi:10.1186/1471-2164-13-85

73. Chow, C. *et al.* PlantPAN 2 . 0 : an update of plant promoter analysis navigator for reconstructing transcriptional regulatory networks in plants. **44,** 1154–1160 (2016). doi:10.1093/nar/gkv1035

74. Eggeling, R., Roos, T., Myllymäki, P. & Grosse, I. Inferring intra-motif dependencies of DNA binding sites from ChIP-seq data. *BMC Bioinformatics* 1–15 (2015). doi:10.1186/s12859-015-0797-4. doi:10.1186/s12859-015-0797-4

Supplementary

Supplementary S1: Feature composition

In table S1 the feature composition of different feature groups is shown. Features were estimated as described in the methods section.

Table S1: Composition of feature groups

Feature group	Features
DNA properties	-The mean, mode, minimum, maximum and standard deviation of the DNA shapes roll, helical twist, major groove width, propeller twist, tilt, stretch, buckle, slide, shift, electro potential, shear, rise, opening. - Melting temperature - LCC - GC-content - CGDMR, CHDMR, CDMR - A binary feature indicating whether a motif is present in a peak region
Shared regions of MADS-box proteins	- A binary feature indicating whether one of the peak regions is shared with the peak regions of AG, AP1, AP3, FLC, PI, SEP3, SOC1 and/or SVP.
Shared regions of other TFs	- A binary feature indicating whether one of the peak regions is shared with the peak regions of the different TFs mentioned in the article of O'Malley *et al.* or Heyndrickx *et al.* (44, 45)
GO-terms	- A binary feature indicating whether the peak region is associated with the GO-term(s).

Supplementary S2: Shared regions

In the folder MADS and TF, figures are present that describe the frequency of the overlap per base position in the peak regions of the MADS. This is described in the upper part of the figure. In the lower part of the figure, the same is described for the motif frequency (see also material and methods). An example of the histogram is shown in figure S1.

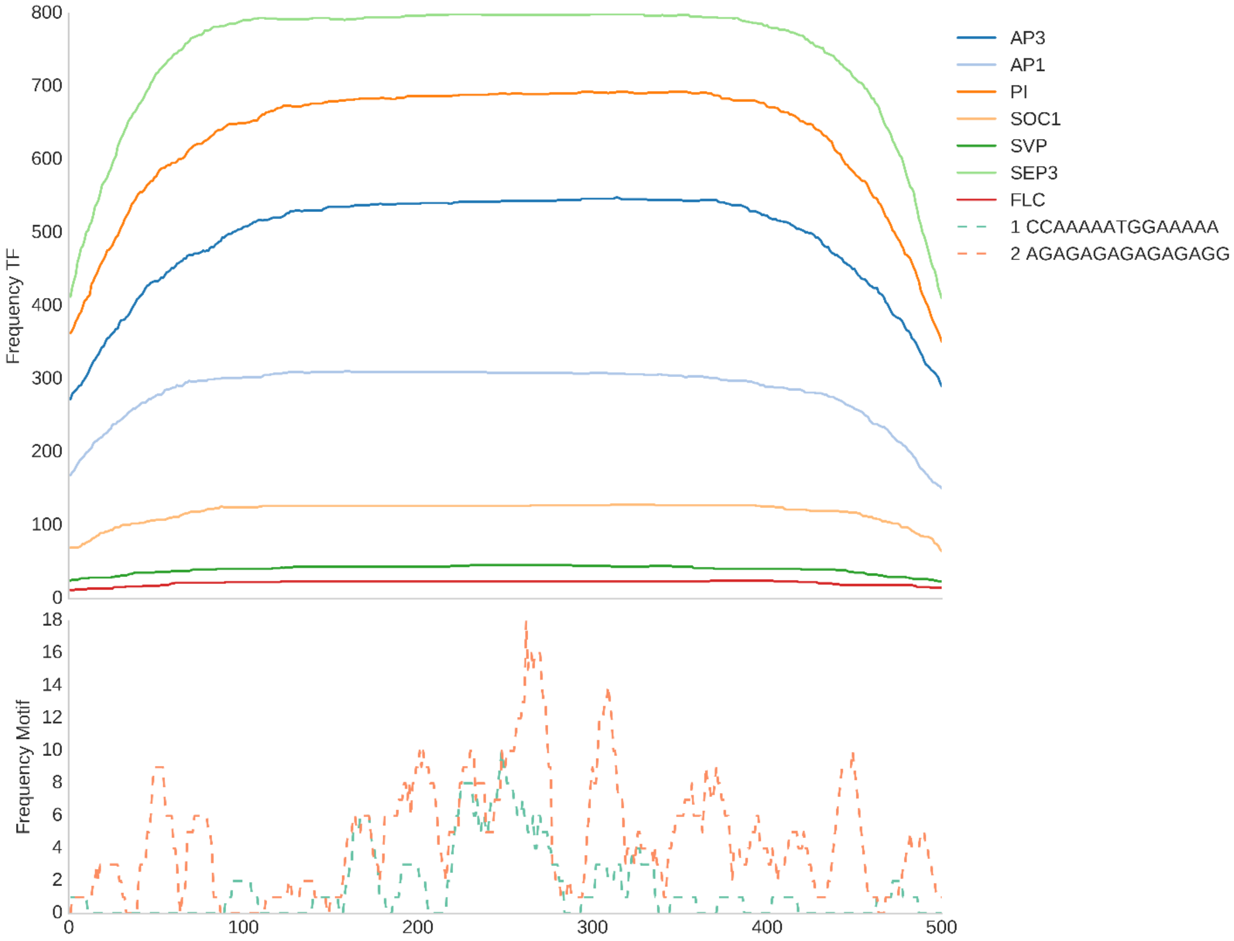

Figure S1 Shared regions AG. *Upper part describes the frequency in how often a shared region occurs in the peak regions of AG, overlap is shown per base. The lower part describes the frequency of the found motifs in the peak regions.*

Supplementary S3: Shared regions MADS

In the table S2, physical interactors found in the BioGRID database are shown. Bold and underlined TFs indicate interactions between the MADS-box proteins itself. Bold TFs indicate a physical interaction with a TF that did not occur in our dataset.

Table S2: physical interactors of the MADS-box proteins

TF	Gene ID	Interactions (physical)
AG	AT4G18960	**SEP3**, FLR1, SEP1, SEP2(AGL2), **AP3**, **PI** , **AG**, AGL6, AGL13, SEP1, SHP1, UBQ3, SPL8, REF6, CHR17, CHR11, CHR4, AGL21, AGL97, AT1G48150, AGL39, AGL24, VSP1, **AGL15**, AGL16, AGL8, **AP1**, BEL1
AP1	AT1G69120	**SEP3**, **SVP**, SEU, AGL24, **AGL20 (SOC1)**, **AP1**, VIT1, **AG**, SYD, SPL8, REF6, CHR17, CHR11, CHR4, ARF2, RPL, BLH1, KNAT3, LUH, INO80, BRM, AGL8, **MAF1(FLM)**, AGL97, AGL92, AGL86, **AP3**, AGL39, AGL21, AGL16, **AGL15**, AGL6, SEP4, SEP1, **PI**, TT16, AT1G48150
AP3	AT3G54340	**PI**, **AP3**, **AG**, **AP1**, SEU, **SEP3**, CHR4, CHR11, AGL13
FLC	AT5G10140	SIZ1, **SVP**, PFT1, AGL16, AT1G54440
PI	AT5G20240	**AP3**, **PI**, **AG**, SEP1(AGL2), **AP1**, CSN5A, AT1G69690, **SEP3**, CHR4, CHR11, CHR17, GRF3, CERK1
SEP3	AT1G24260	**AG**, **AP1**, SHP1, SHP2, STK, **SEP3**, TT16, **AP3**, LUH, REF6, CHR17, CHR11, CHR4, PKL, **SVP**, **PI**, **AGL15**, AGL24**, AGL20 (SOC1)**, AGL16, AGL8, AGL6, SEU, CAL, BEL1
SOC1	AT2G45660	**SEP3**, **AGL15**, SHP1, SEP4, SEP2, SHP2, AGL6, AGL12, AGL13, AGL14, AGL16, AGL17, AGL19, **SVP**, AGL42, **AGL20 (SOC1)**, AT5G47590. AGL71, AGL21, AGL24, TPR2, TPR3, AT2G24550, AGL44, RID3, AT5G04600, BAC089, RR14, SEP1, CAL, TPL, AGL8, SAP18, **AP1**
SVP	AT2G22540	AGL6, AGL21, **AGL20 (SOC1)**, **SEP3**, SEP1, CAL, **AGL15**, AGL16, ATJ3, **FLC**, TFL2, **AP1**, **MAF1 (FLM)**

***** Bold and underlined TF indicate interactions between the MADS-box proteins itself

***** Bold TF indicate a physical interaction with a TF that did not occur in our dataset

Supplementary S4: Pilot experiment of overlapping regions

Background
In our study we showed that overlapping DNA regions between MADS-box proteins found with ChIP-seq data was an important feature in our RF model. This raised the question whether these peaks differed from non-overlapping peaks. To research whether it is possible to answer this question with our dataset, we conducted a pilot experiment, where we compare overlapping peaks with non-overlapping peaks.

Material and methods
Scripts can be found in supplementary S7.

The overlapping regions were selected for the following combinations of datasets: AG-AP1, AG-AP3, AG-FLC, AG-PI, AP3-PI and AG-SEP3. Overlapping regions from these datasets were treated as binding peaks (per combination), while non-overlapping regions were treated as non-binding peaks. Both sets were created in such way that they had the same number of peaks. During non-binding peak selection, the same amount of peaks were taken from both sets. In case one of the sets did not have enough peaks to meet this criterion, the remaining peaks were taken from one of the datasets (see figure S2). These datasets were trained with RF models, as described in the material and methods of the main report.

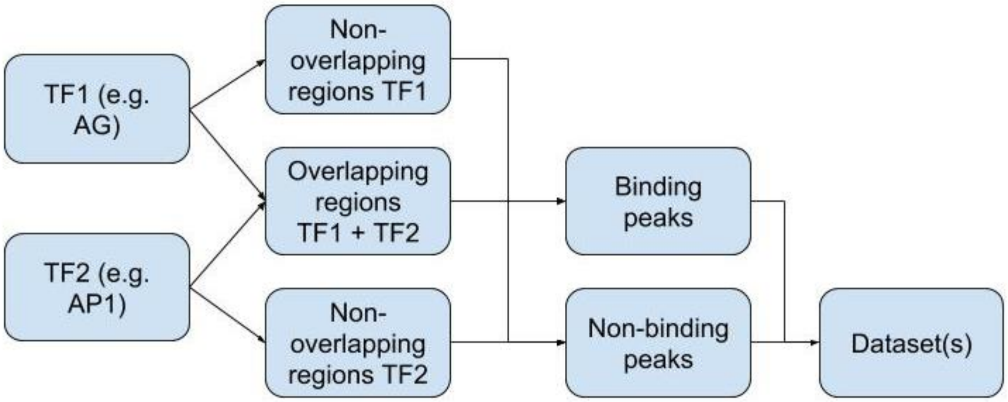
TF1 (e.g. AG)
TF2 (e.g. AP1)
Non-overlapping regions TF1
Overlapping regions TF1 + TF2
Non-overlapping regions TF2
Binding peaks
Non-binding peaks
Dataset(s)

***Figure S2 Flowchart of binding and non-binding peak selection, of overlapping regions.** The overlapping regions between two TFs are considered as binding peaks. The remaining peaks are considered as non-binding peaks. These non-binding peaks are selected as balanced as possible. Binding and non-binding peaks are combined into one dataset.*

Results/Discussion
Results can be found in the Jupyter Notebooks in the supplementary folder S8.

From the results we can observe that the models with the shared region features of the MADS-box proteins have the highest overall performances. These models consist of five features and we have to take into consideration that these features may be biased. The main reason for this is that they are possibly able to describe the nature of each dataset of origin, rather than the overlapping peak region. We can observe that the model with the features describing the overlap between the TF of other studies, has an high performance in the dataset combination AG-FLC (44, 45). Interestingly, this only occurs for this TF combination and not for the others. Moreover, the AP2 TFs seems to have a high feature importance in this model. These preliminary results are able to show that it is possible to gain extra information from the current dataset, if we focus specifically on overlapping regions between MADS-box proteins. Therefore it would be interesting to study these sites in more detail. Like this we can possibly find new interactors related to the TFBS and specific TF complexes. However we have not found specific GO-terms in our current trained models, nonetheless that does not mean that other models of other TF combination do not have a high performance for these GO-terms. To conclude this, further analysis should be conducted.

Supplementary S5: Heatmap of GO-term occurrence

To determine which GO-terms are associated with our peaks, we searched for the most nearby gene of that peak. For the found gene, we searched for the corresponding GO-term and included this GO-term in our model. Figure S3 shows the combined top thirty of the most occurring GO-terms for the class biological processes for the different datasets. The same is shown in S4 and S5, but than for the GO-classes molecular function and cellular component. GO-term occurrence is expressed into a percentage, in which the amount of GO-terms are divided by the amount of peaks in our dataset and multiplied by one hundred.

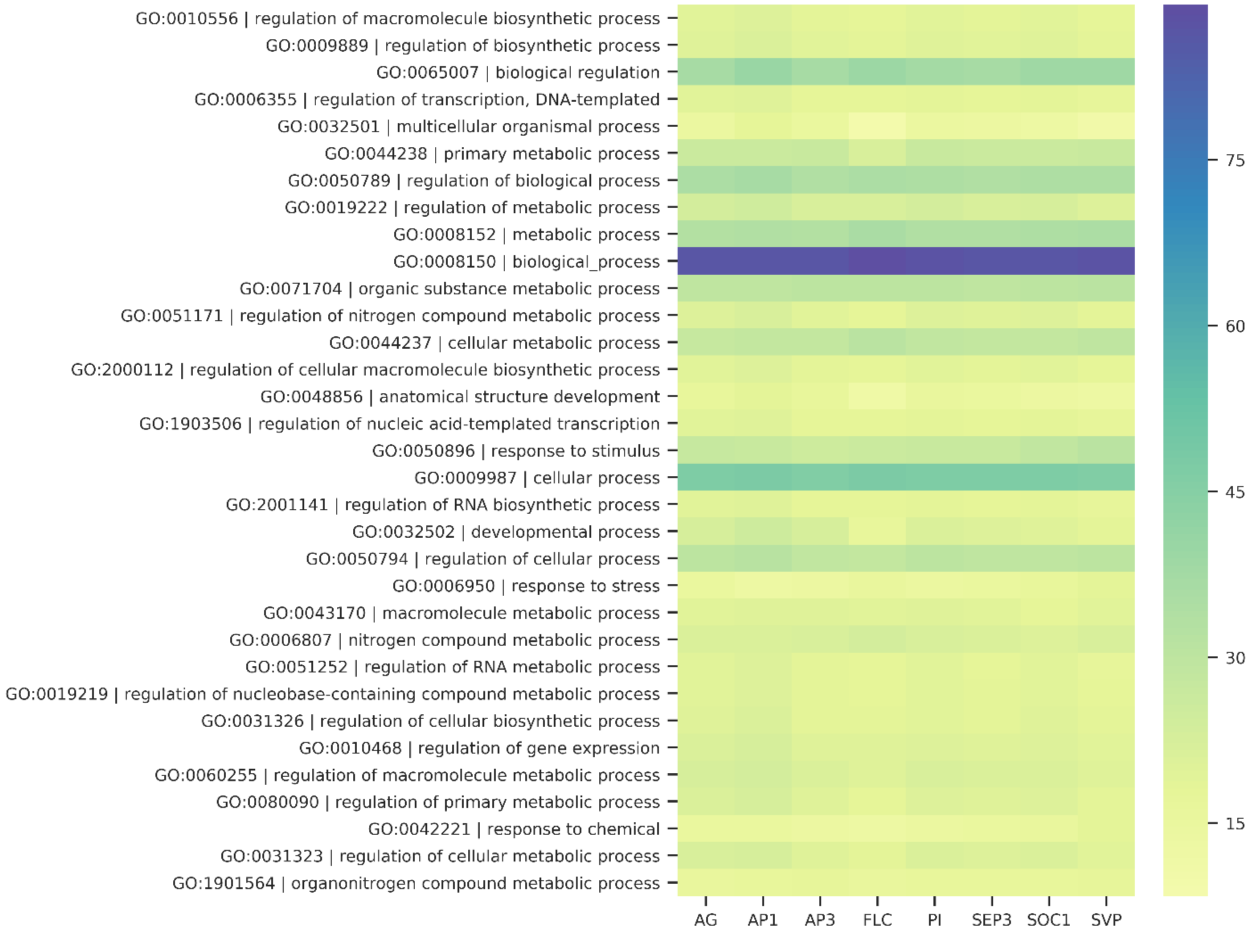

Figure S3 Heatmap occurrence GO-terms biological processes. This heatmap shows the percentage of the most occurring GO-terms (y-axis) per TF (x-axis).

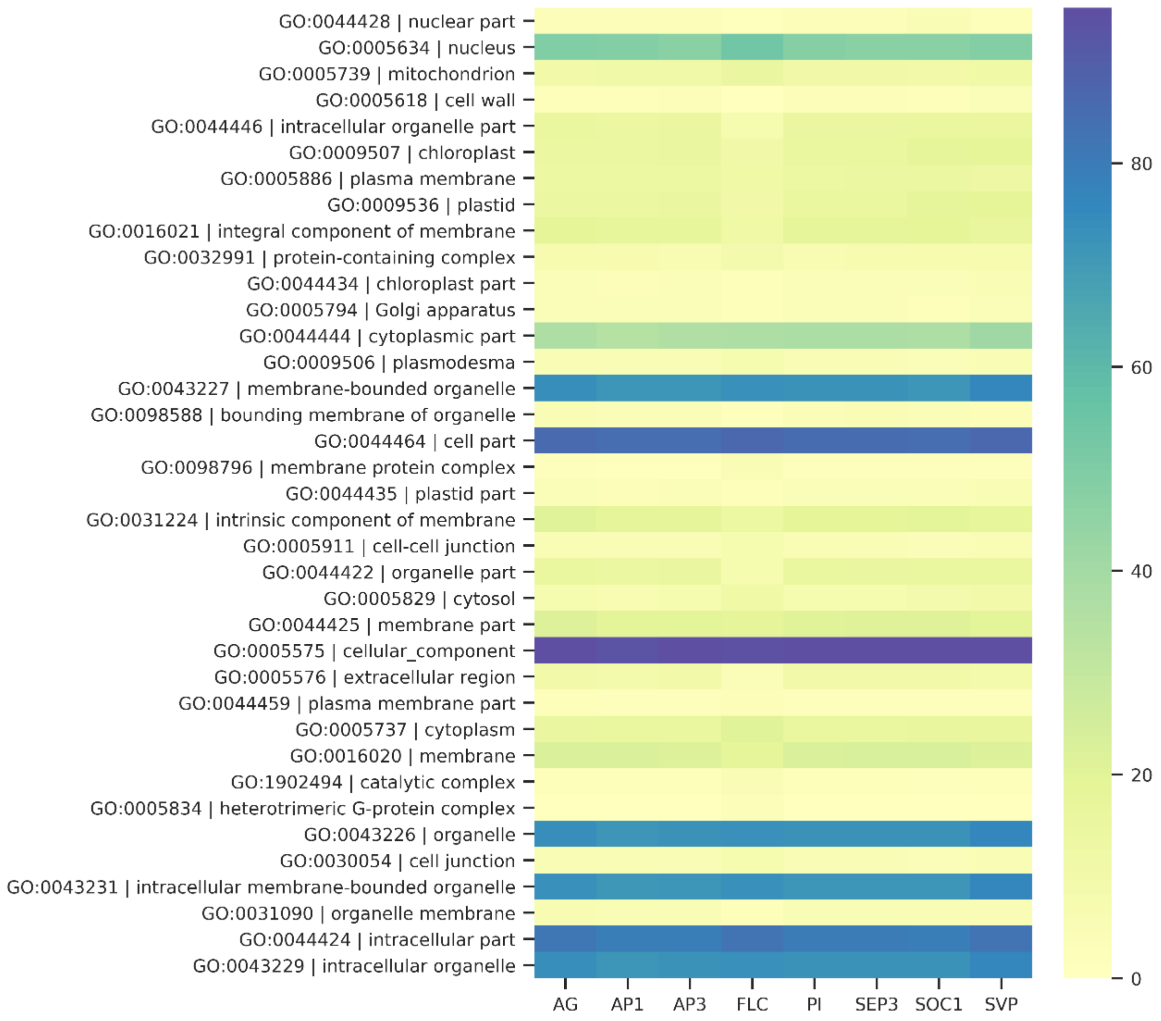

Figure S4 Heatmap occurrence GO-terms molecular function. This heatmap shows the percentage of the most occurring GO-terms (y-axis) per TF (x-axis).

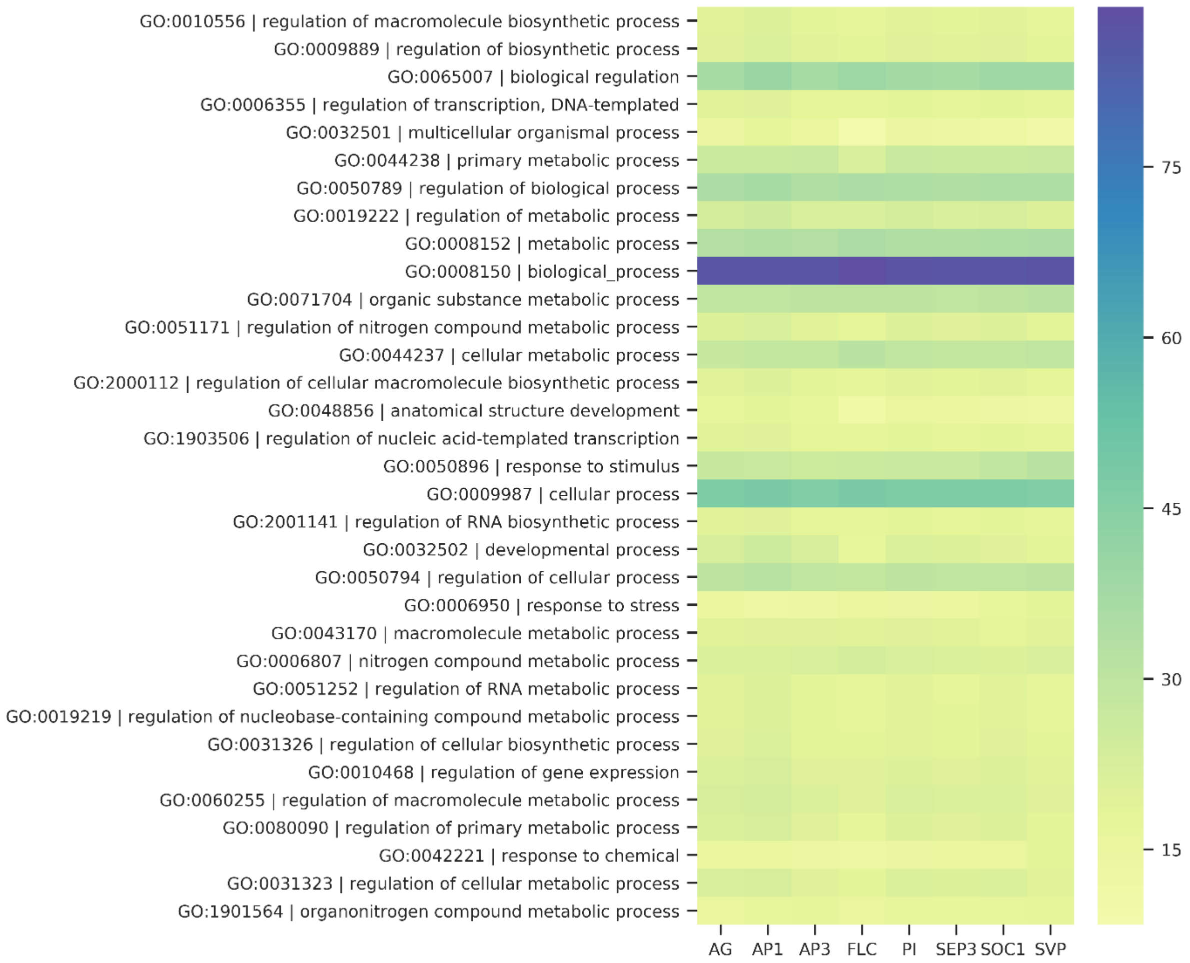

Figure S5 Heatmap occurrence GO-terms cellular component. This heatmap shows the percentage of the most occurring GO-terms (y-axis) per TF (x-axis).