Maya Watanabe
University of Massachusetts Amherst, Spring 2021

Senior Honors Thesis

MADS-box genes and microProteins: implications for transcription factor evolution and regulation

Abstract

Across flowering plants, floral structure is variant, but most are generally composed of a combination of sepals, petals, stamens, and carpels, the identity of which are specified by MADS-box transcription factor genes. The highly-conserved floral homeotic MADS-box genes of the MIKC-type MADS subfamily play important roles in agriculture as disruption to protein function leads to phenotypic consequences in flower and fruit size. Transcription factors, like the MADS-box family, may be regulated by other factors such as microProteins. MicroProteins (miPs) are small proteins that typically consist of a single protein-protein interaction domain. These families of small proteins act as negative regulators as their expression leads to repression of their related target proteins. While several subfamilies of miPs have been classified according to the family of transcription factors with which they interact, no previous literature exists identifying miPs that regulate the MADS-box family. In order to determine whether microProteins may exist within the MADS-box gene family, I performed searches for potential miPs of the MADS-box genes using Hidden Markov Model (HMM) profile searching and protein domain profiling. From these profiles and subsequent individual gene alignments, I identified 10 strong candidate miPs within the MADS-box gene family that have strong consensus in the protein-protein interaction domains of full-length MADS-box genes.

Acknowledgements

Throughout the research process, I have a great deal of assistance, support, and encouragement.

I would first like to thank my supervisor and mentor Dr. Madelaine Bartlett, whose expertise and depth of knowledge was invaluable during the formulation, execution, and analysis of my project. I am grateful for the safe and supportive spaces of scientific and personal exploration that you foster as well as your generosity in dispensing advice throughout my undergraduate experience.

I would also like to thank my mentor Dr. Jarrett Man without whose training, insight, and dry humor, I would not have been able to embark on this thesis project. Thank you for your technical wisdom, your enduring support, and your infinite patience.

Additionally, I thank all the members of the Bartlett Laboratory for their constant support, feedback, and suggestions. Thank you for always lending an ear and an eye, as well as for being a supportive and inclusive community.

Finally, I could not have completed this project without the support of my friends and family. To my parents, thank you for encouraging me to explore these opportunities. And to my friends, thank you for always lending a sympathetic ear.

Introduction

Families of transcription factors have been characterized to regulate genes involved in a diverse array of plant processes such as floral organ morphology, hormone signaling, response to environmental factors, and stem cell differentiation (Bartlett 2017; Singh et al. 2002; Castelán-Muñoz et al. 2019; Drisch and Stahl 2015). The highly-conserved MADS-box transcription factor (TF) family of genes is responsible for the great diversity of floral structure that is present today (Bartlett 2017).

The MADS-box family names comes from the first four discovered members: the *MINICHROMOSOME MAINTENANCE1 (MCM1)* in yeast, *AGAMOUS (AG)* in *Arabidopsis thaliana*, *DEFICIENS (DEF)* in *Antirrhinum*, and serum response factor *(SRF)* in humans (Ng and Yanofsky 2001). Structurally, all MADS-box genes have a highly conserved MADS-domain (M-domain) that is responsible for binding to the DNA of their target genes as dimers and recognize a "CArG" box motif (CC[A/T]$_6$GG) (Nam et al. 2004, Lai et al. 2019). The MADS-box gene family can phylogenetically be subdivided into the type I (also known as M-type or SRF-like) and type II (also known as MIKC-type or MEF2-like) MADS transcription factors (Alvarez-Buylla et al. 2000). Less-studied, the type I MADS-box genes are typically shorter and encoded by a single exon, but still perform transcriptional regulatory activities (Masiero et al. 2011). The well-studied type II genes, due to their additional domains, are typically longer and are encoded by five to eight exons (Masiero et al. 2011).

The MIKC-type MADS-box genes consist of three domains in addition to the DNA-binding M-domain: the intervening (I) domain, the keratin-like coiled-coil (K), and a C-terminal (C) domain (Nam et al. 2004; Ng and Yanofsky 2001). The weakly conserved I-domain plays a role in both DNA-binding specificity and the facilitation of protein-protein interactions (Lai et al. 2021). The highly conserved K-domain determines oligomerization strength and specificity in the dimerization and tetramization of MADS-box transcription factors (Hugouvieux and Zubieta 2018). Like the I-domain, the C-terminal domain is variable with few conserved structures (Lai et al. 2019). The C-domain does not appear to have consistent functional specificity across genes. Some MADS genes contain C-domains that encode transcriptional activation functions while others participate in protein-protein interaction (Piwarzk et al. 2007; Honma and Goto 2001).

Across angiosperms, floral structure is variant, but most flowers are composed of a combination of sepals, petals, stamens, and carpels, the identities of which are specified by the

MADS-box genes. In *Arabidopsis thaliana* (arabidopsis), the flower is composed of four whorls containing each of the floral organs (Kater et al. 2006). The ABC(DE) homeotic floral model is the most widely used model of classification of floral development genes. In arabidopsis, except for one A-class gene (*APETALA2*), these genes are entirely comprised of MADS-box family genes (Becker and Theißen 2003). The A-E class genes specify sepals in the first whorl, the A-B-E class genes specify petals in the second whorl, the B-C-E class genes specify stamens in the third whorl, the C-E class genes specify carpels in the fourth whorl, and the C-D-E class genes specify ovules (Theißen et al. 2016; Bartlett 2017) (Figure 1).

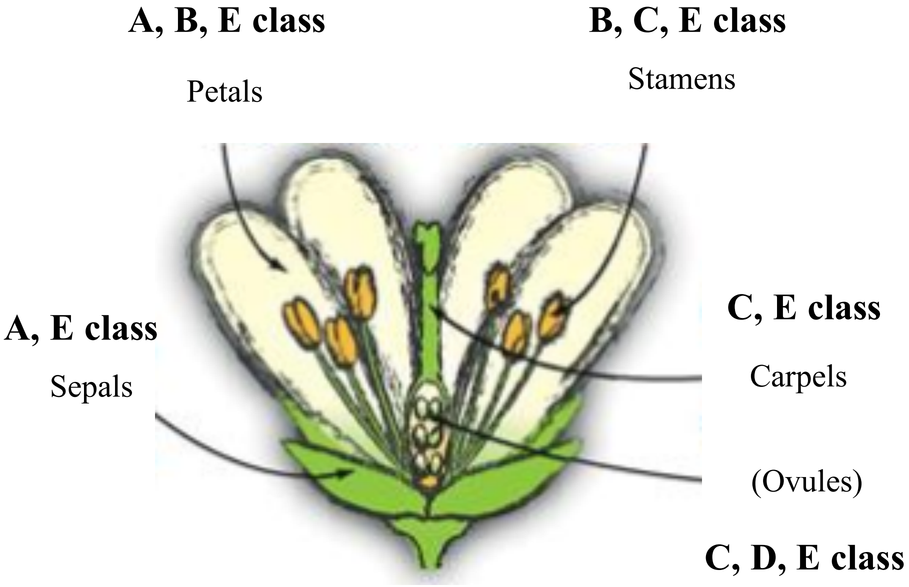

Figure 1. General floral structure and the classes of MADS-box genes that determine organ identity. Modified from Theißen et al. (2016).

The activity of transcription factors, such as those found in the MADS-box family, is tightly regulated. In addition to transcriptional regulation, microProtein-mediated inhibition of protein complex formation is one post-translational way in which protein interactions are regulated (Eguen et al. 2015; Dolde et al. 2018). MicroProteins (miPs) are a class of small proteins that consist of a single protein-protein interaction domain (Bhati et al., 2018). MicroProteins negatively regulate their target proteins in various ways, including competitive inhibition of competent protein complex formation and/or nuclear localization (Eguen et al. 2015; Staudt and Wenkel 2011; Hong et al. 2011; Zhang et al. 2009).

While the inhibitory power of miPs is largely due to the specificity of its protein-protein interactions, the respective protein-protein interaction domains of the miP and its target proteins need not be identical for such negative regulation to occur. The interaction between a protein and a miP with identical protein-protein interaction domains is known as homotypic miP-inhibition while interaction by non-identical (but compatible) protein-protein interaction domains is known as heterotypic miP-inhibition (Bhati et al. 2018). The specification of homotypic or heterotypic inhibition is likely due to the requirement of the target transcription factor to function as a homo- or heterodimer (Graeff and Wenkel 2012). Because of their protein-protein interaction compatibility with target proteins, miP structure is not highly conserved (except that they are small and usually consist of a protein-protein interaction domain).

To date, miPs have been characterized in several of the large transcription factor families. LITTLE ZIPPERS (ZPR) miPs repress the activity of the Class III homeodomain-leucine zipper (HD-ZIPIII) proteins which promote the development of adaxial leaf fates and meristem formation (Wenkel et al. 2007) The four members of the ZPR family (ZPR1/2/3/4) contain a leucine zipper motif similar to that of the HD-ZIPIII proteins (Wenkel et al. 2007). MINI ZINC FINGERS (MIFs) miPs inhibit the activity of zinc-finger-homeodomain protein (ZF-HD) which have been implicated in a large regulatory network of defense and response to environmental stress, development of floral and vegetative organs, and regulation of gametogenesis (Hu and Ma 2006; Takatsuji 1999; Ciftci-Yilmaz and Mittler 2008). The MIF proteins inhibit DNA-binding, inhibit nuclear localization, and to form non-function heterodimers with their target proteins (Hong et al. 2011).

The fabrication of synthetic miPs indicates that miPs are useful tools for specific inhibition of proteins (Eguen et al. 2020). Furthermore, the targeting specificity of these microProtein tools is extremely precise. Seo et al. (2012) demonstrate that overexpression of microProteins leads to phenotypes identical to those of target gene-deficient mutants, that is there are no other phenotypic alterations except those regulated by the target gene. Applied to crop and agricultural bioengineering, the world of microProteins provides precision access to control of proteins in a wide range of important signalling pathways.

Thus far, microProteins have been identified in transcription factor families that regulate meristem formation (Wenkel et al. 2007), defense and environmental response regulatory networks (Ciftci-Yilmaz and Mittler 2008), and inflorescence (Magnani and Hake 2008). This large range of roles indicates that the existence of microProteins is not limited by the function of the proteins that they target. Instead their pervasive presence indicates an effective and important regulatory role that has occurred across protein families and plant species. Thus, although miPs have not been identified within the MADS-box transcription factor family, there is potential for their existence. To investigate the existence of microProteins within the MADS-box gene family, I performed a filter and search process of 13 plant genomes (Straub and Wenkel, 2017; Man et al. 2020) and protein domain identification. From this approach I identified 23 truncated genes as potential miP candidates within the MADS-box family. Subsequent individual gene alignments revealed that 12 of these genes had strong potential to qualify as true miPs.

Materials and Methods

MicroProtein candidate database identification

In this project, I identified potential MADS-box microProteins (miPs) within several plant species spanning land plant diversity. From *Phytozome v12.1* (Goodstein et al., 2012), I obtained and merged the primary transcript peptide annotation databases for the species *Amborella trichopoda, Ananas comosus* (pineapple)*, Arabidopsis thaliana, Daucus carota* (carrot)*, Malus domestica* (apple), *Mimulus guttatus, Marchantia polymorpha, Oryza sativa* (rice)*, Solanum lycopersicum* (tomato), *Physcomitrium patens, Sphagnum fallax,* and *Zea mays* (henceforth called the multi-species database). After merging these individual databases, I used the software *miPFinder v1* (Straub and Wenkel 2017) to filter potential miP candidates from the combined database. Finally, I compiled candidates in a miP candidate database on which I conducted further searches.

MiP gene discovery

My method of gene discovery follows from previous work done by Man et al. 2020. From previous literature, I identified and compiled examples of well-studied and characterized full-length MADS-box genes from the species *A. thaliana*, *O. sativa*, and *Z. mays* (Table 1.) I obtained the primary peptide transcripts of each gene in Table 1 from *Phytozome v12.1* (Goodstein et al., 2012). To increase the list of full-length MADS-box genes to include those from species in the multi-species database, I performed a preliminary search using the genes from Table 1 as search priors. I identified matches in the multi-species database using Hidden Markov Model (HMM) profile searching (Eddy, 2011). With this extended list of full-length MADS-box genes (see

Supplemental Materials), I performed a final search for miPs. This search followed the same procedure as the first full-length gene search but instead of searching within the multi-species database, I searched in miP candidate database. I found a total of 135 truncated MADS-box genes (see Appendix Table A.1).

I inferred final trees using gene alignments generated using *MAFFT v7.313* (Katoh and Standley, 2013) and filtered for homoplastic positions with *Noisy v1.5.12* (Dress et al. 2008). Maximum-likelihood trees with 1,000 bootstrap replicates were inferred using *IQTree v2.1.2* (Nguyen et al. 2015). I visualized final trees using *FigTree v1.4.4* and R.

Table 1. List of full-length MADS-box genes from three plant species

Species	Genes	
A. thaliana	*AP1; AT1G69120.1*	*SEP1; AT5G15800.2*

	AP3; AT3G54340.1	*SEP2; AT3G02310.1*
	PI; AT5G20240.1	*SEP3; AT1G24260.2*
	AG; AT4G18960.1	*SEP4; AT2G03710.1*
	STK; AT4G09960.3	
O. sativa	*OsMADS20; LOC_Os02g49840.1*	*OsMADS58; LOC_Os05g11414.1*
	OsMADS18; LOC_Os07g41370.1	*OsMADS3; LOC_Os01g10504.1*
	OsMADS15; LOC_Os07g01820.1	*OsMADS1; LOC_Os03g11614.1*
	OsMADS14; LOC_Os03g54160.1	*OsMADS7; LOC_Os08g41950.1*
	SUPERWOMAN1; LOC_Os06g49840.1	*OsMADS8; LOC_Os09g32948.1*
	OsMADS4; LOC_Os05g34940.1	*OsMADS5; LOC_Os06g06750.1*
	OsMADS2; LOC_Os01g66030.1	*OsMADS19; LOC_Os02g45770.1*
Z. mays	*ZMM4; Zm00001d034045_P003*	*ZMM2; Zm00001d008882_P001*
	ZMM15; Zm00001d013259_P002	*ZMM25; Zm00001d042591_P002*
	ZMM28; Zm00001d022088_P004	*ZMM23; Zm00001d039434_P001*
	ZMM16; Zm00001d042618_P001	*SILKY1; Zm00001d036425_P002*
	ZMM29; Zm00001d010232_P001	*ZMM8; Zm00001d048082_P001*
	ZMM1; Zm00001d023955_P003	*ZMM14; Zm00001d028217_P001*

Gene domain classification

To assign protein domain classifications to each gene, I detected gene domains from the Pfam database using *HMMER v3.1b2* (Eddy, 2011). I then coded domain hits as follows: 0 - no domain hits; 1 - K-box only; 2 - MADS-domain only; 3 - both K-box and MADS-domain. I then mapped domain hits onto a final tree with SIMMAP (Bollback 2006) using functions from the R packages *ape* and *phytools*. I classified truncated genes with no K-box and no MADS-domain as the most likely candidates for true miPs. To verify these results of the *hmm* detection, I aligned miP genes and visually identified regions of high sequence consensus using *Jalview 2* (Waterhouse et al. 2009).

Individual Alignments

Using the miP classification from the domain mapping, I investigated the domains of the most likely miP candidates. Based on my maximum-likelihood gene tree with 1,000 bootstrap replicates (Figure 1), for each of the miP candidates, I chose the closest full-length homeotic MADS-box gene of the same species (see Supplemental Materials). Those miP candidates without a full-length homolog of the same species were omitted from the alignments. I performed gene alignments in *Jalview 2* (Waterhouse et al. 2009). In these alignments I investigated whether or not the miP candidates had high consensus in the protein-protein interaction domain(s) of their full-length counterparts. Because miPs compete with their full-length paralogous proteins to form non-functional heterodimers, they often share the same or compatible protein-protein interaction domain (Bhati et al. 2019). Evidence of high consensus in the protein-protein interaction domains would indicate strong viability of the miP candidates as true miPs of the MADS-box TF family.

Results

Gene trees reveal potential miP candidates

In order to determine the presence of candidate miPs in my genome searches, I created several MADS-box trees. In Figure 1, I present the gene tree with protein domain presence mapped. The protein domains were detected using *HMMER v3.1b2* and mapped onto the tree in R using the SIMMAP function of the *phytools* package. The domain combinations are color-coded as follows: red, no domain hits; blue, K-box only; green, MADS-domain only; brown both K-box and MADS-domain. I hypothesize that the genes with neither a K-box nor a MADS-domain are most likely to be true miPs rather than MADS-box genes.

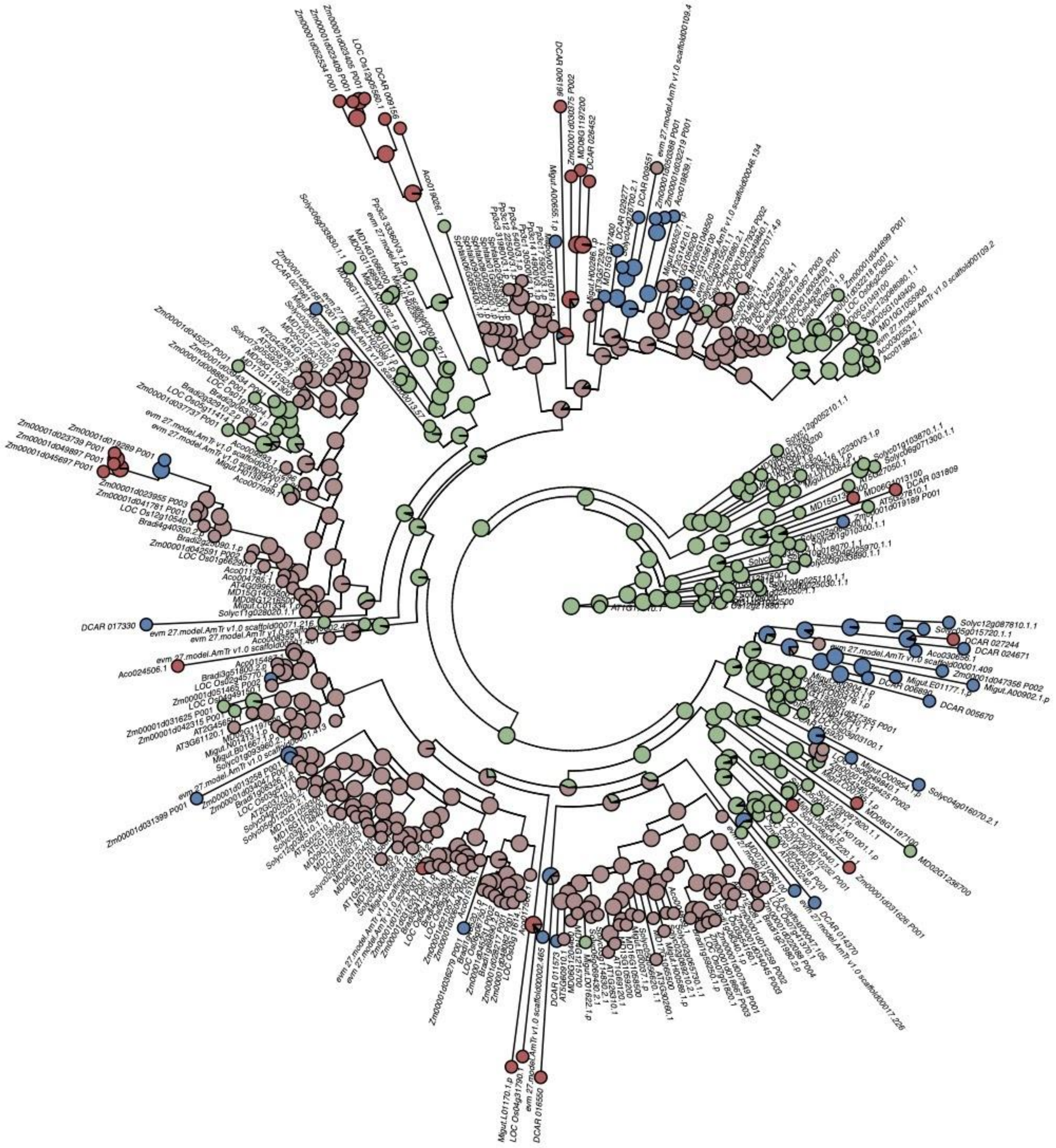

Figure 2. Protein domain mapping using results from HMM domain profiling and Pfam domain database. Red indicates miP candidates (no K-box and no MADS-domain); blue indicates truncated genes (K-box only); green indicates truncated genes (MADS-domain only); brown indicates full-length MADS-box genes (K-box and MADS-domain). Based on this domain profiling, genes with a K-box only (blue), MADS-domain only (green), or no K-box or MADS-domain (red) are most likely candidates for true microProteins.

The domain mapping and ancestral reconstruction of the tree in Figure 2 indicate that there are many potential microProtein candidates related to the MADS-box family. Because miPs likely consist of a protein-protein interaction domain (Eguen et al. 2015), the truncated genes with a K-box only, a MADS-domain only or neither a K-box nor a MADS-domain (respectively blue, green, and red in Figure 2) are considered candidates for miPs. In this project, I chose to investigate only a small subset of this set of genes, the subset of miP candidates without a K-box and without a MADS-domain.

It is also important to note that the tree in Figure 2 appears rooted at a clade of genes mostly consisting of a MADS-domain only. This grouping is likely due to the fact that I used an outgroup of an M-type MADS-box gene, a MADS-family gene consisting of a single MADS-domain and a variable C-terminal domain (Chen et al. 2018; Masiero et al. 2011). This outgroup sampling seems to be affecting the ancestral reconstruction for the rest of the tree. The ancestral reconstructed states that occur at the node where each MADS-box clade branches off indicate that the full-length and miP candidates arose from genes with only a MADS-domain. It is thought that the MADS-box family of genes evolved from a region of topoisomerases IIA subunit A (TOPOIIA-A) and that subsequent gene duplication in the TOPOIIA-A lineage likely gave rise to the distinct type I and type II MADS-box genes present today (Gramzow et al. 2010). Thus, it is unlikely that the true ancestry of the MADS-box family of genes is like that depicted in Figure 2. In order to account for this outgroup sampling, future searches should look closely at this group of genes and re-evaluate their placement in the tree. It is possible that they do not belong in this tree.

In Figure 2, the maximum-likelihood gene tree with 1,000 bootstrap replicates is made up of full-length MADS-genes (gray), homeotic genes from *A. thaliana*, *O. sativa*, and *Z. mays* are shows in the following colors per class: light green, A-class; blue and purple, B-class; red, C-class; dark-green, D-class; pink, E-class. Finally, genes in the color burgundy are all truncated genes which are to be examined to determine if they have potential as true MADS-box miPs.

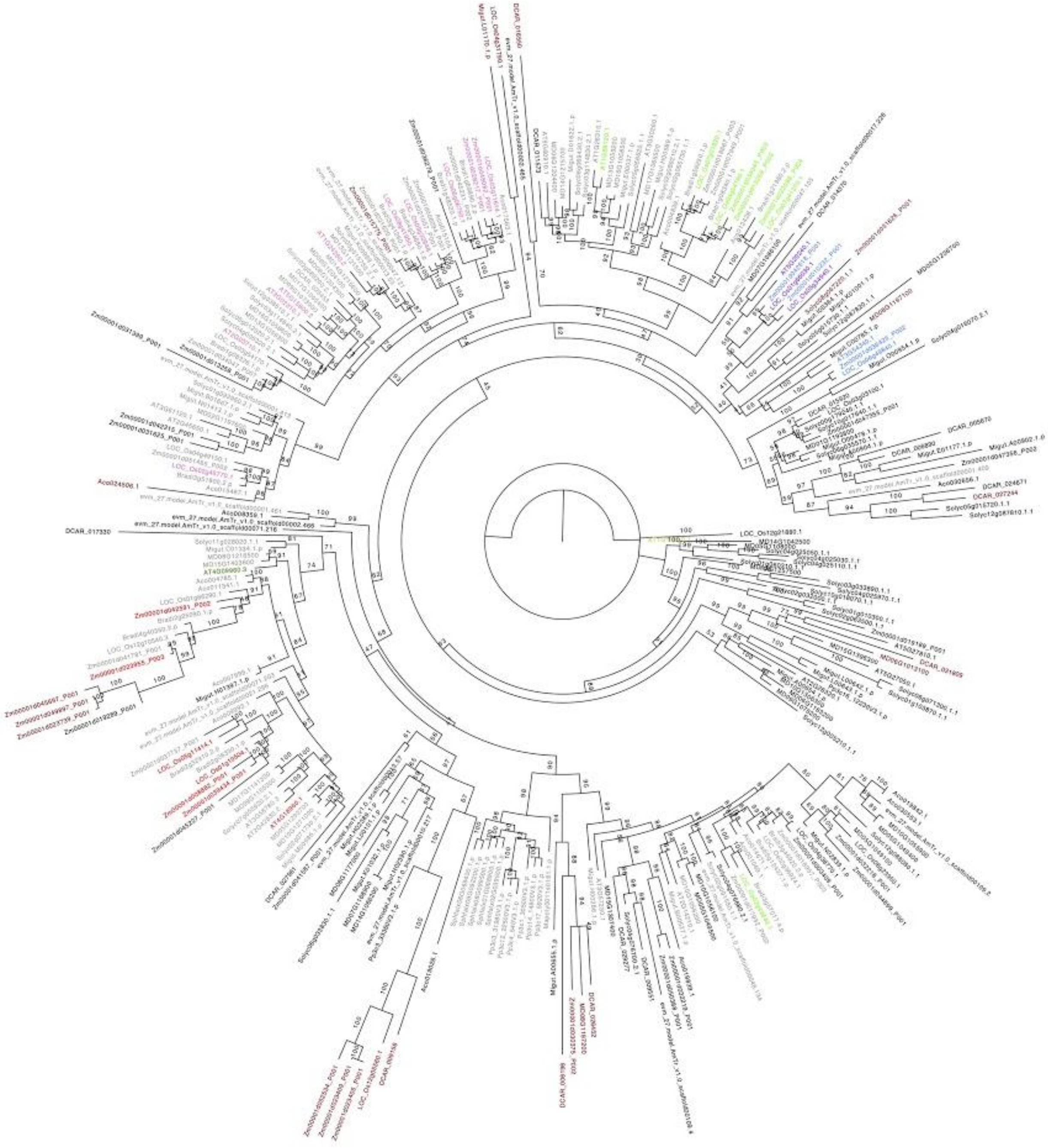

Figure 3. Full phylogeny full length MADS-genes (grey), A-class genes (green), B-class genes (blue and purple), C-class genes (red), D-class genes (dark green), E-class genes (pink), type I MADS-box outgroup (yellow), truncated MADS-box genes (black), and candidate microProteins (burgundy).

Figure 3 shows the same genes presented in Figure 2 with bootstrap support values and the MADS-box clades color-coded for ease of reading. Having decided to focus on only the candidate miPs with no K-box and no MADS-domain, the tree in Figure 3 focuses on highlighting the ABCDE MADS-box clades and the selected candidate miPs within them. From this tree, there is 1 miP candidate sister to the A clade, 3 miP candidates in the B genes, 1 miP candidate in a clade sister to the B-class genes, 12 miP candidates in the C/D clade, 4 in the E clade, and 2 miP candidates in the outgroup. Thus, miP candidates seem to be concentrated in the C/D clade.

MicroProtein candidates

The B (purple and blue), the C (red), and E class clades seem to have the greatest number of candidate miPs (Figure 2). The remaining MADS-box clades do not have clear miP candidates. Table 2 lists all truncated genes without a K-box and MADS-domain (burgundy).

Table 2. Full list of candidate microProteins genes with no detected K-box or MADS-domain.[1]

Aco024506.1
DCAR_006196
DCAR_026452
DCAR_009156
DCAR_016550
DCAR_027244
DCAR_031809
LOC_Os12g05560.1
LOC_Os04g31790.1
MD08G1197200
MD08G1197100
MD06G1013100
Migut.L01170.1.p
Solyc08g067220.1.1
Zm00001d030375_P002
Zm00001d023405_P001
Zm00001d023409_P001
Zm00001d052534_P001
Zm00001d023739_P001
Zm00001d049897_P001
Zm00001d045697_P001
Zm00001d015775_P001
Zm00001d031626_P001

I looked at the subset of miPs without a K-box domain and without a MADS-domain to determine if any have protein-protein interaction domains using an alignment with the closest known homeotic homolog. Table 3 lists each candidate miP, its closest known homeotic MADS-box gene(s), and the domain(s) in which I found the highest consensus from individual alignments.

[1] Based on Figure 2.

Table 3. Domain[2] consensus between miP candidates and closest full-length MADS-box gene(s)[3] and the clade classification.

Candidate microProtein	Closest full-length homolog	Domain consensus	Clade
Aco024506.1	Aco015487.1	-	E
LOC_Os04g31790.1	LOC_Os02g45770.1 (*OsMADS19*)	K, C	E
LOC_Os12g05560.1	LOC_Os05g11414.1	K	C/D
MD08G1197200	MD10G1056200	K	C/D
Migut.L01170.1.p	Migut.K00969.1.p	K, K-C	E
Zm00001d015775_P001	Zm00001d021057	I-K	E
Zm00001d023405_P001	Zm00001d042591_P002	I, I-K,	C/D
Zm00001d023409_P001		-	C/D
Zm00001d052534_P001		-	C/D
Zm00001d023739_P001	Zm00001d023955_P003 (*ZMM1*)	I-K, K-C	C/D
Zm00001d045697_P001		K	C/D
Zm00001d049897_P001		I-K, K-C	C/D
Zm00001d030375_P002	Zm00001d017932_P002	K-C	C/D
Zm00001d031626_P001	Zm00001d042618_P001 (*ZMM16*); Zm00001d010232_P001 (*ZMM29*)	-	B

[2] M, I, K, and C denote their respective MADS-box domains, K-C indicates a region overlapping the K-domain and the C-domain, and I-K indicates a region overlapping the I-domain and the K-domain.

[3] MiP candidate genes were aligned with their closest floral homeotic full-length MADS-box gene(s) (Fig. 3), to determine consensus. Those without a homolog of the same species within the same clade were omitted from the alignment.

The presence of consensus between the full-length MADS-box gene and the miP would indicate evidence for the viability of the candidate as a true miP. Of the 23 original miP candidates without a K-box and without a MADS-domain, 14 had homologs in the same clade and species, and of those 14, 10 showed consensus in protein-protein interaction domains. There were 9 miP candidates (of the original 23) that did not have close homologs in the same species may be miss-annotated or their close homologs may be un-annotated and missing from the gene tree. The M-, I-, and K-domains facilitate interaction between the MADS-box TF and other proteins (Hugouvieux et al. 2018; Lai et al. 2019). There is also evidence that regions spanning both I- and K-domains facilitate protein-protein interaction (Lai et al. 2019). Additionally, the end of the K-domain into the C-domain facilitates tetramerization with target proteins (Song and Chen, 2018). The third column of Table 3 lists consensus between the miP candidate and the four MADS-box domains. The K-domain has high consensus most frequently across all 10 of the miP candidates. As the K-domain plays a role in protein-protein interactions (Lai et al. 2019), this high frequency makes it more likely that these miP candidates are true miPs of the MADS-box family. Finally, the last column of Table 3 provides a general categorization of the miPs into a MADS-box clade. This classification was based on the relationships in Figure 3.

Individual protein alignments reveal strongest miP candidates

To investigate my chosen subset of miP candidates more closely, I made alignments between candidate miPs and their closest full-length homologs. The first miP candidate I examined was *Zm00001d015775_P001* and its closest homolog *Zm00001d021057* (Figure 4). *Zm00001d015775_P001* has high consensus in the I-domain and in a region that spans both the I and K domains of *Zm00001d021057*. The I-domain is involved in dimerization specificity of the transcription factor (Lai et al. 2019; Grimplet et al. 2016). Crystal structures of *SEP3* reveal that overlapping regions in the I- and K-domains also play a role in dimerization and tetramerization (Puranik et al. 2014). Thus, because *Zm00001d015775_P001* seems to share regions that facilitate protein-protein interactions, it is a likely candidate for a true miP.

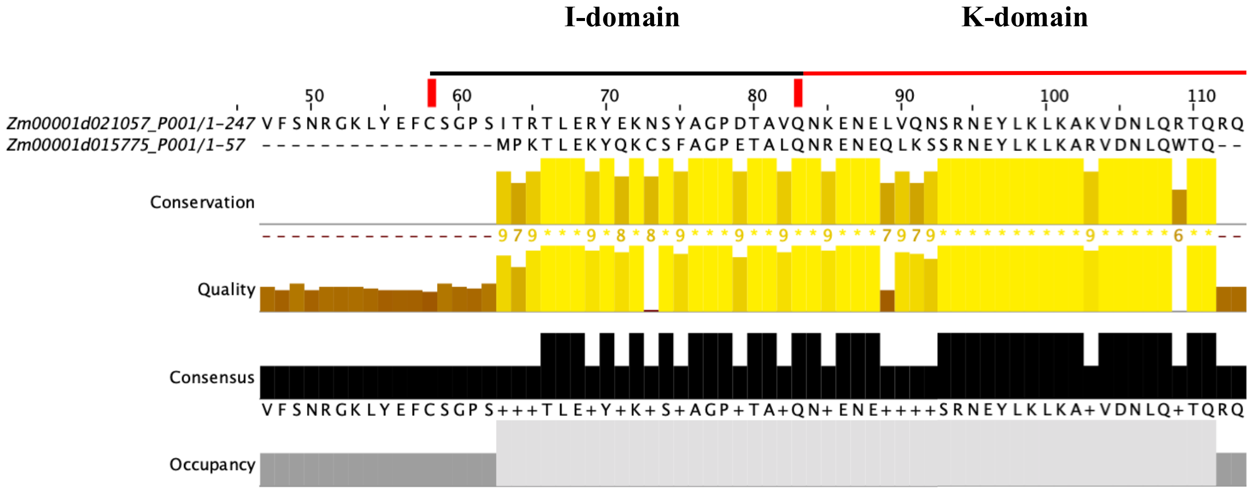

Figure 4. Alignment of miP candidate *Zm00001d015775_P001* and its closest homolog *Zm00001d021057* reveals strong consensus in the I- domain and a region spanning the I and K domains. Domain annotations based on annotation of *ZMM1* by Dong et al. (2019).

The next miP candidate I investigated was *Zm00001d023739_P001* and its closest homolog *Zm00001d023955_P003* (*ZMM1*) (Figure 5). The alignment of the miP candidate and the C-class gene *ZMM1* in Figure 4 shows strong consensus in two regions, one spanning the I- and K-domains and the other spanning the K- and C-domains. Similar to the region spanning the I- and K-domains mediates dimerization and tetramerization, the region spanning the end of the K-domain and the beginning of the C-domain are important for proper floral organ identity specification (Piwarzyk et al. 2007). The K-domain consists of two α-helices, K1 and K2 (Yang and Jack 2004; Hugouvieux et al. 2018). While Piwarzk et al. (2007) used a model of the K-domain marked out in three α-helices, K1, K2, and K3, they showed that the part of the helix in the final region of the K-domain was necessary for the function of AP3 and PI.

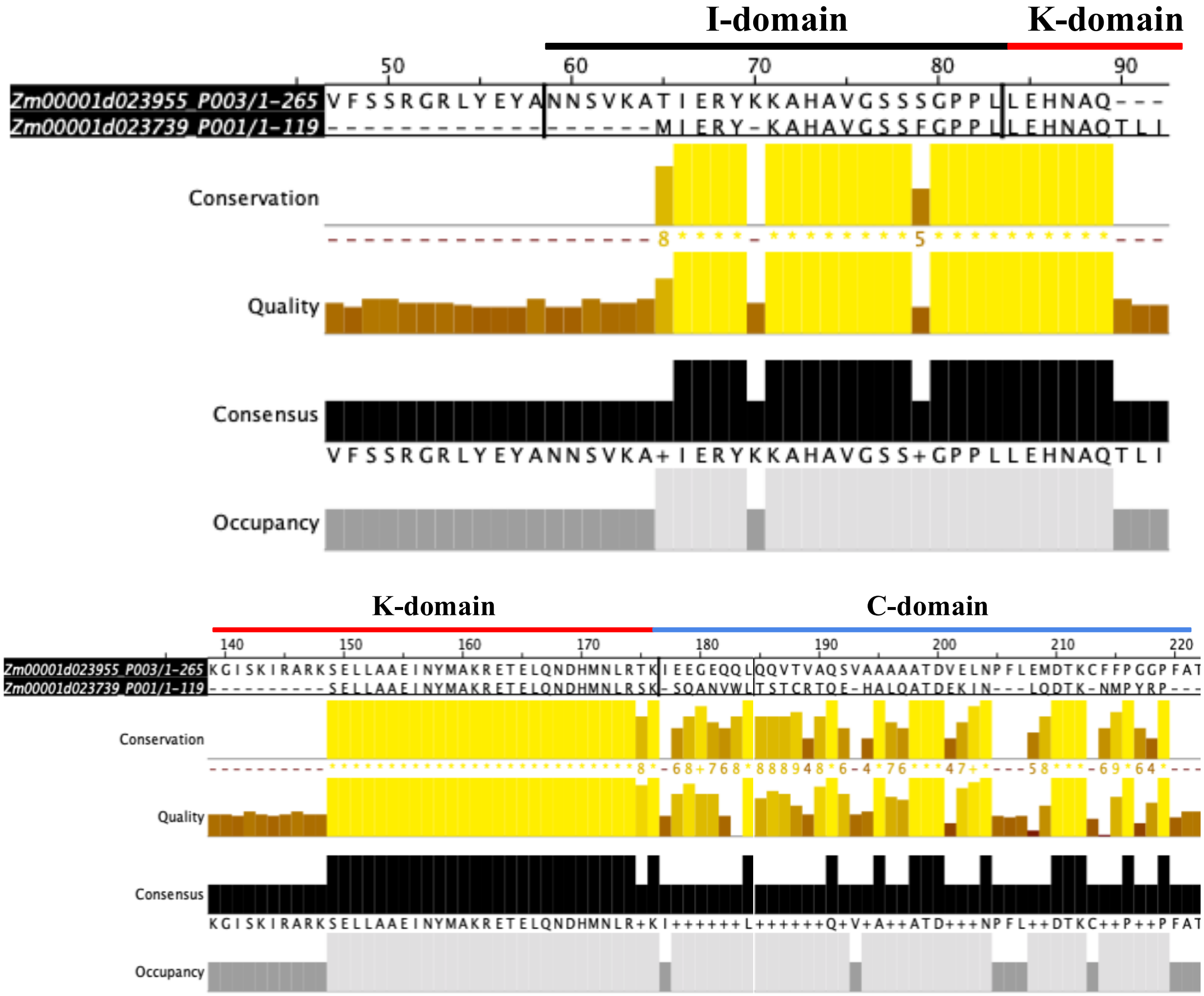
I-domain
K-domain
50
60
70
80
90
Zm00001d023955_P003/1-265
Zm00001d023739_P001/1-119
VFSSRGRLYEYANNSVKATIERYKKAHAVGSSSGPPLLEHNAQ---
-------------------MIERY-KAHAVGSSFGPPLLEHNAQTLI
Conservation
8****-********5**********
Quality
Consensus
VFSSRGRLYEYANNSVKA+IERYKKAHAVGSS+GPPLLEHNAQTLI
Occupancy
K-domain
C-domain
140
150
160
170
180
190
200
210
220
Zm00001d023955_P003/1-265
Zm00001d023739_P001/1-119
KGISKIRARKSELLAAEINYMAKRETELQNDHMNLRTKIEEGEQQLQQVTVAQSVAAAAATDVELNPFLEMDTKCFFPGGPFAT
----------SELLAAEINYMAKRETELQNDHMNLRSK-SQANVWLTSTCRTQE-HALQATDEKIN---LQDTK-NMPYRP---
Conservation
----------**************************8*-68+768*888948*6-4*76***47+*---58***-69*64*---
Quality
Consensus
KGISKIRARKSELLAAEINYMAKRETELQNDHMNLR+KI++++++L++++++Q+V+A++ATD+++NPFL++DTKC++P++PFAT
Occupancy

Figure 5. Alignment of *Zm0001d023739_P001* and *ZMM1* reveal strong consensus in the a region spanning the I- and K-domains (top) and in a region spanning the K- and C-domains (bottom). Domain annotation for *ZMM1* based on amino acid alignments of Dong et al. (2019).

The next miP candidate I investigated was *Zm00001d045697_P001* and its full-length homolog *Zm00001d023955_P003* (*ZMM1*) (Figure 6). Figure 6 shows strong consensus in the K-domain. While the K-domain as a whole mediates protein-protein interactions, Puranik et al. (2014) show that the two homodimers of the SEP3 K-domain associate due to the hydrophobic interactions of the second α-helix, K2. In Figure 7, we see that strong consensus between the two genes occurs primarily toward the end of *ZMM1*'s K-domain where the K2 helix lies. Due to the important role in protein-protein interactions that K2 plays, *Zm00001d045697_P001* is a likely candidate for a true C-class miP.

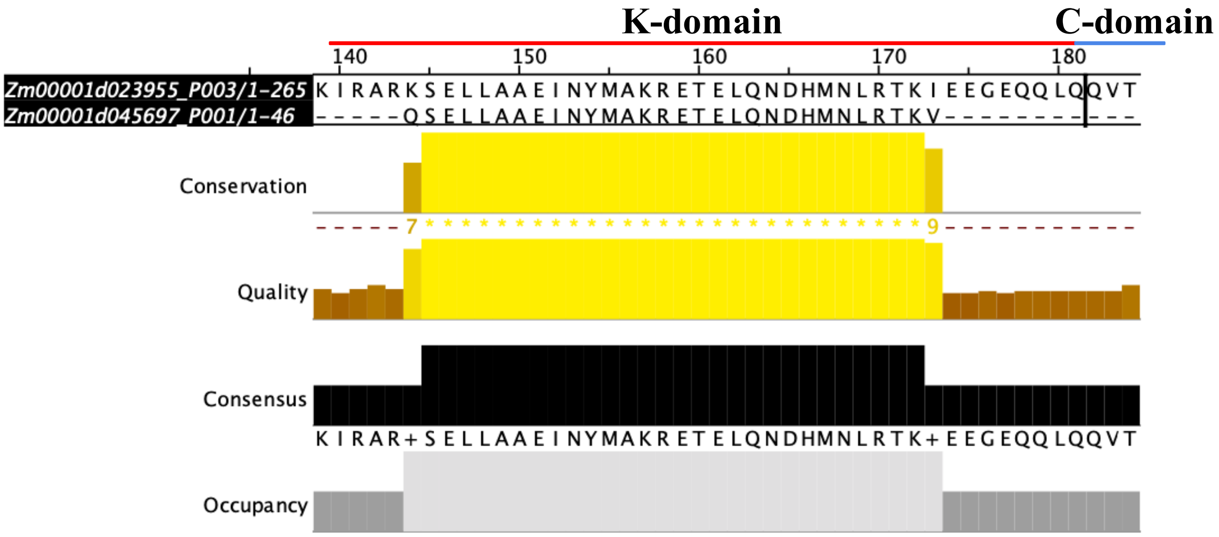

Figure 6. Alignment of *Zm00001d045697_P001* and its full-length homolog *Zm00001d023955_P003* (*ZMM1*) reveals strong consensus in the K-domain. Domain annotation for *ZMM1* based on amino acid alignments of Dong et al. (2019).

The next miP candidate that I investigated was *Zm00001d049897_P001* and its closest full-length homolog *Zm00001d049897_P001* (*ZMM1*) (Figure 7). Figure 7 shows strong consensus in two regions, one spanning the I- and K-domains and the other spanning the K- and C-domains. Similar to the alignment in Figure 4, the miP candidate *Zm00001d049897_P001* has consensus in two protein-protein interaction regions, this makes it a good candidate for a true miP of the MADS-box family.

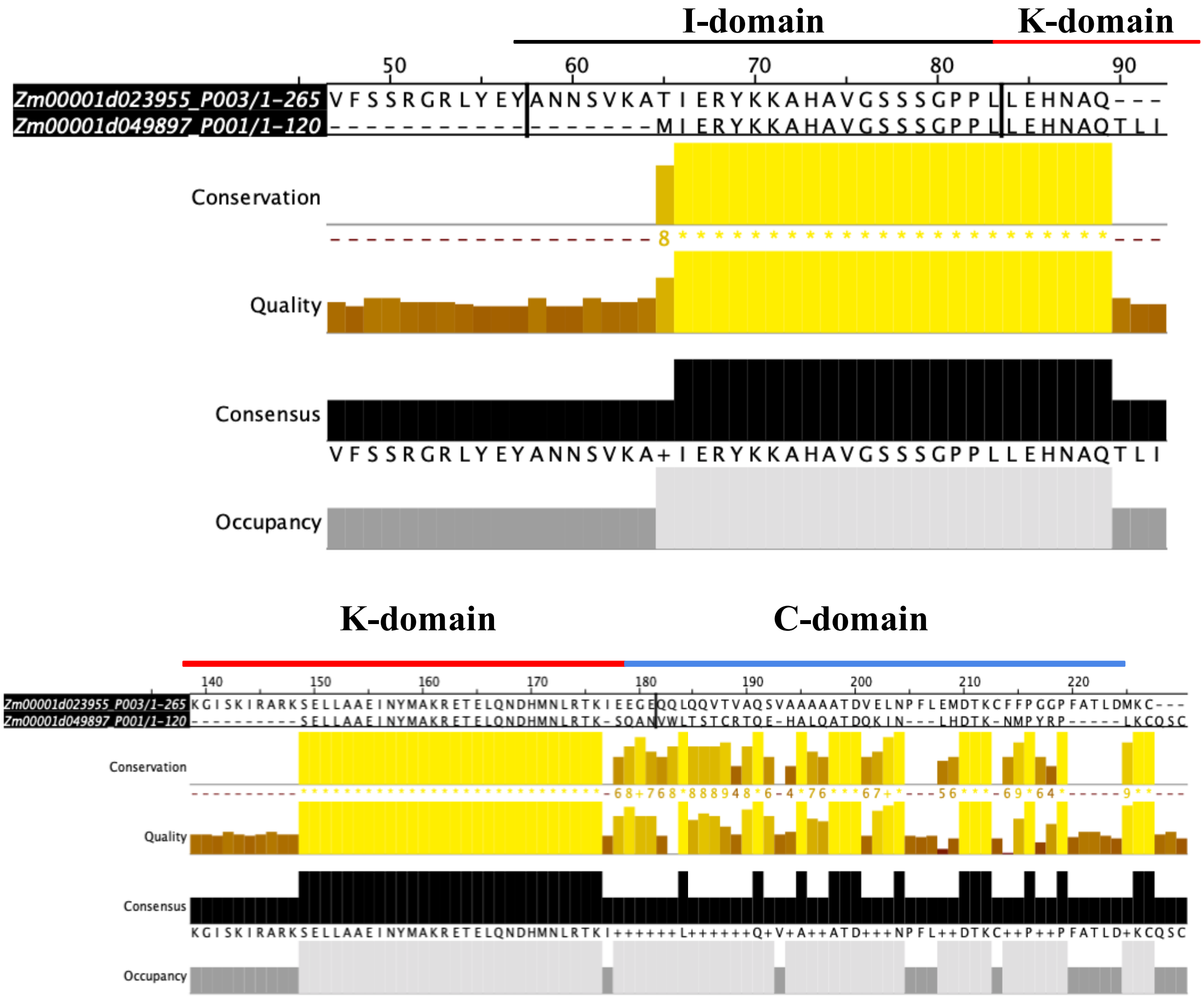

Figure 7. Alignment of *Zmm00001d049897_P001* and *its closest homolog ZMM1* reveal strong consensus in the a region spanning the I- and K-domains (top) and in a region spanning the K- and C-domains (bottom). Domain annotation for *ZMM1* based on amino acid alignments of Dong et al. (2019).

After viewing all these promising alignments, it is important to note that I have chosen to show only 5 of the miP candidates with the highest consensus. These 5 genes all happen to be maize genes. As a note of caution, the maize genome is not as well annotated as other species' genomes (the arabidopsis genome is very well annotated), so although in these alignments, the miP candidates look promising, there is also a possibility of a misannotation in these genes.

Discussion

No previous literature exists identifying microProteins within the MADS-box transcription factor family. However, evidence for the existence of microProteins would indicate a novel regulatory mechanism within the MADS-box transcription factor family. To determine the existence of miPs in the MADS-box family, I used an initial general filtering of gene length (< 140 aa) with the *miPFinder v1* (Straub and Wenkel 2017) and then performed an iterative method of genome searching (Man et al. 2020) followed by domain mapping onto the gene tree consisting of both full-length and truncated MADS-box genes. Using this method, I found 23 miP candidates of particular interest with no K-box domain and no MADS-domain. Of the 23 candidates, 14 had homologs in the same species, and of those 14, 10 had significant consensus with the protein-protein interaction domains of their closest homeotic homolog. The miP candidates without close homologs in the same species may be miss-annotated or their close homologs may be un-annotated. These consensus between protein-protein interaction domains occurred in I, K- domains and overlapped in I-K and K-C domains. The identification of these potential miP within the MADS-box transcription factors has greater implications for the regulatory mechanisms within this important family.

Support from synthetic microProteins

While I cannot definitively classify the miP candidates identified in this research as true MADS-box miPs, previous research in *synthetic* miPs within the MADS-box family shows that MADS-box proteins can be regulated by miPs. Seo et al. (2010) engineered synthetic microProteins targeting the MIKC-type MADS-box gene *AGAMOUS (AG)*, the C-class homeotic MADS-box gene in arabidopsis. The authors engineered eight truncated variations of *AG* consisting of various combinations of all four MIKC domains. Overexpression of the engineered the *AG-K* (K-domain only) was shown to produce disruptions in floral structure similar to those of the *ag-3* knockout mutant. This indicates the negative effects of *AG-K* occur through protein-protein interaction mediated by the K-domain.

Additionally, these authors produced similarly truncated forms of the MIKC-type MADS-box gene *SUPPRESSOR OF OVEREXPRESSOR CONSTANS 1 (SOC1)*. The researchers generated four truncated *SOC1* genes to mimic microProteins to target *SOC1* consisting of a combination of the MADS-, I-, and K-domains (C-terminus was excluded) of the MADS-box gene. These synthetic microProteins consisted of 1) the MADS-domain only (*S-M*), 2) the MADS-domain, I-domain, and K-domain (*S-MIK)*, 3) the second-half of the MADS-domain, the I domain, and the K-domain (*S-IK*), and 4) the K-domain only (*S-K*). The authors showed that transgenic arabidposis plants overexpressing *S-MIK*, *S-IK,* and *S-K* exhibited a delayed flowering phenotype indicating expression of truncated forms of the *SOC1* gene containing the protein-protein interaction domain, the K-domain, suppressed *SOC1* activity. The *S-M* truncated form also showed slightly delayed flowering which is thought to be the result of the competition between MADS-domains of *S-M* and *SOC1*. Additionally, the authors found that these synthetic microProteins inhibited *SOC1-SOC1* homodimerization and that all except the *S-M* prevented *SOC1* nuclear localization.

These results are intriguing as I too show that several miP candidates exhibit strong consensus in the K-domain, M-domain, and overlapping I-K domains (Table 3). While for *AG*, Seo et al. (2012) synthesized a truncated gene spanning the K- and C-domains, they concluded that this synthesized miP did not produce the expected negative regulation. While I have three miP candidates showing strong consensus in a region spanning both the K- and C-domains, these genes also all showed strong consensus in at least one other domain (I, I-K, or K). Thus, due to the presence of I, I-K, or K domains, it is likely that if these are true miPs, they would function as negative regulators of their target MADS-box proteins. Thus, since the miPs with consensus in the protein-protein interaction domains have close homologs in the same clade and species, it is likely that the miP candidates can interact with these close homologs (and potentially others not included in my gene trees) and negatively regulate their function.

MicroProteins and MADS-box clades

Based on the findings of my search and the gene tree created from both full-length and truncated MADS-box genes, the C-D-class clade has 7 strong miP candidates, and the E-class clade has 3 strong miP candidates (miP candidates with consensus in a protein-protein interaction region; Table 3). The A- and B-class clades of genes do not have any miP candidates with consensus in the protein-protein interaction domains.

One explanation for the lack of MADS-box miPs in the A-class clade of genes is that this clade may already be regulated by another family of miPs. In the A-class genes, there appears to an indirect potential method of regulation. In the shoot apical meristem pathway (SAM), the arabidopsis gene *FLOWERING LOCUS T* (*FT*) forms a complex with two basic leucine-zipper transcription factors which leads to the activation of the A-class MADS-box gene *APETALA1* (*AP1*; Andrés et al. 2015). The LITTLE ZIPPER (ZPR) family of proteins which regulate the class III homeodomain-leucine zipper (HD-ZIPIII) proteins (Wenkel et al. 2007). While the leucine-zipper transcription factors that bind to FT are not of the HD-ZIPIII family, there is potential for a similar family of microProteins with a similarly compatible leucine zipper domain that may regulate the formation of the FT-leucine-zipper complex. This regulation in turn would also indirectly regulate the activation of the AP1 protein in the SAM.

In arabidopsis, expression of the floral meristem identity genes *LEAFY* (*LFY*) and *AP1* are required for the activation of the B-class gene *APETALA3* (*AP3*; Lamb et al. 2002). Thus, because another mechanism of regulation exists for this B-class gene, it is possible that this class of MADS-box genes are not regulated by miPs.

While I have found promising candidates of miP regulations, I must also acknowledge the possibility that MADS-box genes may not be regulated by any miPs.

Most frequent consensus in the K-domain

MicroProteins typically contain a protein-protein interaction domain similar to that of the protein family with which they interact (Eguen et al. 2015). However, I note that different amino acid sequences result in similar protein structure. There are two general modes of miP inhibition: 1) homotypic inhibition, in which the protein-protein interaction domains are the same and 2) heterotypic inhibition, in which the protein-protein interaction domains are compatible but not necessarily the same (Eguen et al. 2015). Through my gene alignment and subsequent consensus analyses, I was looking generally for similarity between the amino acid sequences of both the miP candidate and its full-length homolog, that is homotypic inhibition. I did not however look directly at protein structure compatibility. It is possible that the miP candidates for which I deemed amino acid consensus was absent may have compatible protein structure for their full-length homologs, that is heterotypic inhibitory properties. Future work on these miP would benefit for a deeper analysis of protein structure of all the miP candidates.

Domain-classification of the miP candidates that I found in my searches shows that of the 10 strong microProtein candidates, 1 candidate showed consensus strictly in the I-domain, 4 candidates showed consensus in the I-K region, 5 showed consensus in a region strictly in the K-domain, 4 candidates showed consensus in the K-C region, and only 1 gene showed consensus in a region strictly in the C domain (Table 3). None of these genes showed consensus in the M-domain. The M-domain is the DNA-binding domain of the MADS-box gene family, highly conserved across plant and animal species (Lai et al. 2019). However, domain swap experiments in *Arabidopsis* where the M-domain was replaced with that of *SRF* (in yeast) and *MEF2* (in humans) reproduced normal phenotypes of *arabidopsis* MADS-box genes *AP1*, *AP3*, *PI*, and *AG* (Riechman and Meyerowitz, 1997; Lai et al. 2019). The results of this experiment indicate that DNA-binding (through the M-domain) specificity of these genes seems to be independent of their corresponding homeotic properties Riechmann and Meyerowitz 1997). Thus, the M-domains may be largely interchangeable for DNA-binding specificity, other factors may play a role, and other MADS-domains (specifically the I- and K-domains) may contribute to DNA-binding specificity (Lai et al. 2019). *In vivo* experiments demonstrate the both the I- and K-domains do play roles in DNA-binding specificity (Hugouvieux et al. 2018; Lai et al. 2019).

Given this literature, it is possible that because the M-domain may be interchangeable and independent of MADS-box protein DNA-binding *and* function and because the I- and K-domains are known to play roles in both, miP candidate domains are more likely to be similar to these latter domains. Indeed of my 10 strong miP candidates, half showed strong consensus in at least the K-domain and 6 showed strong consensus in a region at least overlapping part of the K-domain. Consensus in the I-domain is less common in my candidates; only 1 gene shows strong consensus in a region strictly in the I-domain, while 4 genes show strong consensus in a region overlapping the I- and K-domains. The I-domain is a small domain that links the M- and K-domains and exhibits more sequence diversity than either of those two domains (Lai et al. 2019). While the I-domain plays a role in dimerization activity, it also stabilizes the M-domain and plays a role in DNA-binding (Lai et al. 2021). Similar to the I-domain, the C-domain is less conserved and has more sequence diversity than either the M- or K-domains. However, it does play a role in the formation of protein complexes and particularly determines the specificity of interactions of MADS-box proteins (Song and Chen 2018). Of the miP candidates, half show strong consensus at least in partial regions of the C-domain. The K-domain is a more highly conserved and defining feature of the type II MADS-box genes (Lai et al. 2019). Because miPs typically regulate their target proteins through protein-protein interaction and the K-domain facilitates a majority of these interactions, it makes sense that most of the miP candidates show strong consensus in this domain over others.

Future Directions

Thus far, using domain mapping and individual gene alignments, my research has identified 10 promising miP candidates and their close homologs. I have shown that these miPs have significant consensus in the protein-protein interaction domains of homeotic MADS-box genes. However, I cannot definitively classify these truncated genes as miPs without further experimentation and validation.

Before beginning *in vivo* experiments, I need to validate the expression patterns and expression timing of both miP genes and their homologs. In order to determine if it is possible for these miPs to negatively regulate their target proteins, I would need to show that they are expressed at the same time and the same place as their full-length homologs. One way to do this is to examine expression atlases of the species' full genomes. For example, all my most promising are from the maize genome, although not the most well annotated, there are databases of gene expression such as the *MaizeGDB* (Portwood et al. 2018).

If expression data supports the likelihood of the miP candidates interacting with full-length MADS-box homologs, *in vivo* experiments would be necessary to confirm the negative regulatory function. I would expect to see that a knock-out of the coexpressed full-length gene would produce the same phenotypic alteration as the over-expression of the microProtein and vice versa. Furthermore, additional *in silico* experiments and research would be necessary for identification of these candidates as true miPs. While I present a preliminary assessment and classification of the miP candidates using gene alignments, a more careful analysis of gene annotations, alignments with homologs in the same species, as well as using the results of RNAseq data would show expression patterns and phylogenetic relationships that would determine if these candidates are viable as true miPs.

The implications of this and future research on MADS-box microProteins is important for both agricultural and horticultural advancement. The potential targeting specificity of these microProtein tools is extremely precise. The synthetic microProteins demonstrate that overexpression of miPs lead to phenotypes identical to those of target gene-deficient mutants, that is there are no other phenotypic alterations except those regulated by the target gene (Seo et al. 2012). Applied to crop and agricultural bioengineering, families of miPs would provide precision access to control of proteins in a wide range of important signalling pathways. The existence of microProteins within the type II MADS-box gene family would indicate a new regulatory mechanism within this important transcription factor family.

References

Alvarez-Buylla, E. R., Pelaz, S., Liljegren, et al. (2000). An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. *Proceedings of the National Academy of Sciences of the United States of America*, *97*(10), 5328–5333.

Andrés, F., Romera-Branchat, M., Martínez-Gallegos, R., et al.. (2015). Floral Induction in Arabidopsis by FLOWERING LOCUS T Requires Direct Repression of *BLADE-ON-PETIOLE* Genes by the Homeodomain Protein PENNYWISE. *Plant Physiology, 169(3)*, 2187-2199. doi: https://doi.org/10.1104/pp.15.00960.

Bartlett, M. E. (2017). Changing MADS-Box Transcription Factor Protein-Protein Interactions as a Mechanism for Generating Floral Morphological Diversity. *Integrative and Comparative Biology*, *57*(6), 1312–1321.

Becker, A., and Theissen, G. (2003). The major clades of MADS-box genes and their role in the development and evolution of flowering plants. *Molecular Phylogenetics and Evolution*, *29*(3), 464–489.

Bhati, K. K., Blaakmeer, A., Paredes, E. B., Dolde, U., Eguen, T., Hong, S.-Y., Rodrigues, V., Straub, D., Sun, B., and Wenkel, S. (2018). Approaches to identify and characterize microProteins and their potential uses in biotechnology. *Cellular and Molecular Life Sciences: CMLS*, *75*(14), 2529–2536.

Bollback, J. P. (2006). SIMMAP: Stochastic character mapping of discrete traits on phylogenies. *BMC Bioinformatics, 7(88)*. doi: https://doi.org/10.1186/1471-2105-7-88.

Castelán-Muñoz, N., Herrera, J., Cajero-Sánchez, W., Arrizubieta, M., Trejo, C., García-Ponce, B., Sánchez, M. de la P., Álvarez-Buylla, E. R., and Garay-Arroyo, A. (2019). MADS-Box Genes Are Key Components of Genetic Regulatory Networks Involved in Abiotic Stress and Plastic Developmental Responses in Plants. *Frontiers in Plant Science*, *10*, 853.

Chen, Y.-T., Chang, C.-C., Chen, C.-W., Chen, K.-C., & Chu, Y.-W. (2018). MADS-Box Gene Classification in Angiosperms by Clustering and Machine Learning Approaches. *Frontiers in Genetics*, *9*, 707.

Ciftci-Yilmaz, S., and Mittler, R. (2008). The zinc finger network of plants. *Cellular and Molecular Life Sciences: CMLS*, *65*(7-8), 1150–1160.

Dolde, U., Rodrigues, V., Straub, D., Bhati, K. K., Choi, S., Yang, S. W., and Wenkel, S. (2018). Synthetic MicroProteins: Versatile Tools for Posttranslational Regulation of Target Proteins. *Plant Physiology*, *176*(4), 3136–3145.

Dong, Q., Wang, F., Kong, J., Xu, Q., Li, T., Chen, L., Chen, H., Jiang, H., Li, C., and Cheng, B. (2019). Functional analysis of *ZmMADS1a* reveals its role in regulating starch biosynthesis in maize endosperm. *Scientific Reports, 9(3253)*. doi: https://doi.org/10.1038/s41598-019-39612-5.

Dress, A. W. M., Flamm, C., Fritzsch, G., Grünewald, S., Kruspe, M., Prohaska, S. J., and Stadler, F. P. (2008). Noisy: Identification of problematic columns in multiple sequence alignments. *Algorithms for Molecular Biology, 3(7)*, https://doi.org/10.1186/1748-7188-3-7.

Drisch, R. C., and Stahl, Y. (2015). Function and regulation of transcription factors involved in root apical meristem and stem cell maintenance. *Frontiers in Plant Science*, *6*, 505.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, *7*(10), e1002195.

Eguen, T., Straub, D., Graeff, M., and Wenkel, S. (2015). MicroProteins: small size-big impact. *Trends in Plant Science*, *20*(8), 477–482.

Eguen, T., Ariza, J. G., Brambilla, V., Sun, B., Bhati, K. K., Fornara, F., and Wenkel, S. (2020). Control of flowering in rice through synthetic microProteins. *Journal of Integrative Plant Biology*, *62*(6), 730–736.

Galimba, K. D., Tolkin, T. R., Sullivan, et al. (2012). Loss of deeply conserved C-class floral homeotic gene function and C- and E-class protein interaction in a double-flowered ranunculus mutant. *PNAS, 109(34)*, E2267-E2275.

Goodstein, D. M., Shu, S., Howson, R., Neupane, R., Hayes, R. D., Fazo, J., Mitros, T., Dirks, W., Hellsten, U., Putnam, N., and Rokhsar, D. S. (2012). Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Research*, *40*(Database issue), D1178–D1186.

Graeff, M., and Wenkel, S. (2012). Regulation of protein function by interfering protein species. *Biomolecular Concepts*, *3*(1), 71–78.

Grimplet, J., Martínez-Zapater, J. M., and Carmona, M. J. (2016). Structural and functional annotation of the MADS-box transcription factor family in grapevine. *BMC Genomics*, *17*, 80.

Hong, S.-Y., Kim, O.-K., Kim, S.-G., Yang, M.-S., and Park, C.-M. (2011). Nuclear import and DNA binding of the ZHD5 transcription factor is modulated by a competitive peptide inhibitor in Arabidopsis. *The Journal of Biological Chemistry*, *286*(2), 1659–1668.

Honma, T., and Goto, K. (2001). Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. *Nature*, *409*(6819), 525–529.

Hu, W., and Ma, H. (2006). Characterization of a novel putative zinc finger gene MIF1: involvement in multiple hormonal regulation of Arabidopsis development. *The Plant Journal: For Cell and Molecular Biology*, *45*(3), 399–422.

Hugouvieux, V., Silva, C. S., Jourdain, A., Stigliani, A., Charras, Q., Conn, V., Conn, S. J., Carles, C. C., Parcy, F., and Zubieta, C. (2018). Tetramerization of MADS family transcription factors SEPALLATA3 and AGAMOUS is required for floral meristem determinacy in Arabidopsis. *Nucleic Acids Research*, *46*(10), 4966–4977.

Hugouvieux, V., and Chloe Zubieta. (2018). MADS transcription factors cooperate: complexities of complex formation. *Journal of Experimental Botany*, *69*(8), 1821–1823.

Kater, M. M., Dreni, L., and Colombo, L. (2006). Functional conservation of MADS-box factors controlling floral organ identity in rice and Arabidopsis. *Journal of Experimental Botany*, *57*(13), 3433–3444.

Katoh, K. and Daron M. Standley. (2013). MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution, 30(4)*, 772-280. doi: https://doi.org/10.1093/molbev/mst010.

Lai, X., Daher, H., Galien, A., Hugouvieux, V., and Zubieta, C. (2019). Structural Basis for Plant MADS Transcription Factor Oligomerization. *Computational and Structural Biotechnology Journal*, *17*, 946–953.

Lai, X., Vega-Leon, R., Hugouvieux, V., et al. (2021). The Intervening Domain is Required for DNA-binding and Functional Identity of Plant MADS Transcription Factors. *bioRXiv*. doi: https://doi.org/10.1101/2021.03.10.434815.

Lamb, R. S., Hill, T. A., Tan, Q. K., and Irish, V. F. (2002). Regulation of APETALA3 floral homeotic gene expression by meristem identity genes. *Development*, *129(9)*, 2079-2086. PMID: 11959818.

Magnani E. and Sarah Hake. (2008). *KNOX* Lost the *OX*: The *Arabidopsis KNATM* Gene Defines a Novel Class of KNOX Transcriptional Regulators Missing the Homeodomain. *Plant Cell, 20(4)*, 875-887. doi: 10.1105/tpc.108.058495.

Man, J., Gallagher, J. P., and Bartlett, M. (2020). Structural evolution drives diversification of the large LRR-RLK gene family. *New Phtyologist, 226(5)*, 1492-1505. doi: https://doi.org/10.1111/nph.16455.

Masiero, S., Colombo, L., Grini, P. E., Schnittger, A., and Kater, M. M. (2011). The emerging importance of type I MADS box transcription factors for plant reproduction. *The Plant Cell*, *23*(3), 865–872.

Nam, J., Kim, J., Lee, S., An, G., Ma, H., and Nei, M. (2004). Type I MADS-box genes have experienced faster birth-and-death evolution than type II MADS-box genes in angiosperms. *Proceedings of the National Academy of Sciences of the United States of America*, *101*(7), 1910–1915.

Ng, M., and Yanofsky, M. F. (2001). Function and evolution of the plant MADS-box gene family. *Nature Reviews. Genetics*, *2*(3), 186–195.

Nguyen, L., Schmidt, H. A., von Haeseler, A., and Minh, B. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution*, *32(1)*, 268-274. doi: https://doi.org/10.1093/molbev/msu300.

Piwarzyk, E., Yang, Y., and Jack, T. (2007). Conserved C-terminal motifs of the Arabidopsis proteins APETALA3 and PISTILLATA are dispensable for floral organ identity function. *Plant Physiology*, *145*(4), 1495–1505.

Portwood, J. L. II, Woodhouse, M. R., Cannon, E. K., et al. (2018). MaizeGDB 2018: the maize multi-genome genetics and genomics database. *Nucleic Acids Res., 47(D1)*, D1146-D1154. doi: 10.1093/nar/gky1046.

Puranik, S., Acaijaoui, S., Conn, S., et al. (2014). Structural Basis for Oligomerization of the MADS Domain Transcription Factor SEPALLATA3 in *Arabidopsis*. *Plant Cell, 26(9),* 3603-3615. doi: https://doi.org/10.1105/tpc.114.127910.

Riechmann, J. L. and E. M. Meyerowitz. (1997). Determination of floral organ identity by Arabidopsis MADS domain homeotic proteins AP1, AP3, PI, and AG is independent of the DNA-binding specificity. *Mol. Biol. Cell., 8(7)*, 1243-1259. doi: 10.1091/mbc.8.7.1243.

Seo, P. J., Hong, S. Y., Ryu, J. Y., Jeong, E. Y., Kim, S. G., Baldwin, I. T., and Park, C. M. (2012). Targeted inactivation of transcription factors by overexpression of their truncated forms in plants. *The Plant Journal: For Cell and Molecular Biology*, *72*(1), 162–172.

Singh, K.., Foley, R. C., and Oñate-Sánchez, L. (2002). Transcription factors in plant defense and stress responses. *Curr. Opin. Plant Biol., 5(5)*, 430-436. doi: 10.1016/s1369-5266(02)00289-3.

Singh, R., Low, E. L., Ooi, L. C., et al. (2013). The oil palm *SHELL* gene controls oil yield and encodes a homologue of SEEDSTICK. *Nature, 500*, 340-344. doi: https://doi.org/10.1038/nature12356.

Song, G. and Quixia Chen. (2018). Overexpression of the MADS-box gene K-domain increases the yield potential of blueberry. *Plant Science, 276*, 22-31. doi: https://doi.org/10.1016/j.plantsci.2018.07.018.

Staudt, A. C., and Wenkel, S. (2011). Regulation of protein function by "microProteins." *EMBO Reports*, *12*(1), 35–42.

Straub, D., and Wenkel, S. (2017). Cross-Species Genome-Wide Identification of Evolutionary Conserved MicroProteins. *Genome Biology and Evolution*, *9*(3), 777–789.

Takatsuji, H. (1999). Zinc-finger proteins: the classical zinc finger emerges in contemporary plant science. *Plant Molecular Biology*, *39*(6), 1073–1078.

Theißen, G., Melzer, R., and Rümpler, F. (2016). MADS-domain transcription factors and the floral quartet model of flower development: linking plant development and evolution. *Development 143(18)*, 3259-3271. doi: https://doi.org/10.1242/dev.134080.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., and Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. *Bioinformatics*, *25*(9), 1189–1191.

Wenkel, S., Emery, J., Hou, B.-H., Evans, M. M. S., and Barton, M. K. (2007). A feedback regulatory module formed by LITTLE ZIPPER and HD-ZIPIII genes. *The Plant Cell*, *19*(11), 3379–3390.

Yang, Y., and Jack, T. (2004). Defining subdomains of the K domain important for protein-protein interactions of plant MADS proteins. *Plant Molecular Biology*, *55*(1), 45–59.

Zhang, L. Y., Bai, M. Y., Wu, J., Zhu, et al. (2009). Antagonistic HLH/bHLH transcription factors mediate brassinosteroid regulation of cell elongation and plant development in rice and Arabidopsis. *The Plant Cell*, *21*(12), 3767–3780.

Appendix

Table A.1 Full list of potential miP genes after secondary search using full-length MADS-box genes as search priors.

Species	Genes	
A. comosus	Aco019026.1	Aco024506.1
	Aco019839.1	Aco030553.1
	Aco019842.1	Aco030656.1
A. thaliana	AT2G26320.1	AT5G27810.1
	AT5G27050.1	
D. carota	DCAR_005670	DCAR_016550
	DCAR_006196	DCAR_017330
	DCAR_006890	DCAR_024671
	DCAR_009156	DCAR_026452
	DCAR_009551	DCAR_027244
	DCAR_011573	DCAR_027961
	DCAR_014370	DCAR_029277
	DCAR_015920	DCAR_031809
A. trichopoda	evm_27.model.AmTr_v1.0_scaffold00002.465	evm_27.model.AmTr_v1.0_scaffold00017.226
	evm_27.model.AmTr_v1.0_scaffold00002.466	evm_27.model.AmTr_v1.0_scaffold00071.216
	evm_27.model.AmTr_v1.0_scaffold00010.217	evm_27.model.AmTr_v1.0_scaffold00109.2
	evm_27.model.AmTr_v1.0_scaffold00013.57	evm_27.model.AmTr_v1.0_scaffold00109.4

O. sativa	LOC_Os03g03100.1	LOC_Os06g23950.1
	LOC_Os04g31790.1	LOC_Os12g05560.1
	LOC_Os04g38770.1	LOC_Os12g21880.1
M. domestica	MD01G1193800	MD08G1197100
	MD02G1236700	MD08G1197200
	MD05G1049100	MD09G1075200
	MD05G1049400	MD10G1055900
	MD05G1049500	MD10G1056100
	MD05G1108000	MD10G1306300
	MD06G1013100	MD13G1257500
	MD06G1163200	MD14G1042500
	MD07G1086100	MD14G1066200
	MD07G1168900	MD15G1307400
	MD08G1177000	MD15G1396300
M. guttatus	Migut.A00654.1.p	Migut.K01032.1.p
	Migut.A00655.1.p	Migut.L00101.1.p
	Migut.A00902.1.p	Migut.L00642.1.p
	Migut.A00904.1.p	Migut.L00643.1.p
	Migut.C00785.1.p	Migut.L01170.1.p
	Migut.E01177.1.p	Migut.N02839.1.p

	Migut.H01397.1.p	Migut.O00478.1.p
	Migut.H02389.1.p	Migut.O00954.1.p
	Migut.H02390.1.p	Migut.K01001.1.p
	Migut.I00364.1.p	
P. patens	Pp3c16_12230V3.1.p	Pp3c3_33360V3.1.p
S. lycopersicum	Solyc00g179240.1.1	Solyc04g076700.2.1
	Solyc01g010300.1.1	Solyc05g015720.1.1
	Solyc01g060310.1.1	Solyc05g015730.1.1
	Solyc01g103870.1.1	Solyc06g033830.1.1
	Solyc02g032000.1.1	Solyc06g035570.1.1
	Solyc02g063500.1.1	Solyc06g071300.1.1
	Solyc03g033890.1.1	Solyc08g067220.1.1
	Solyc04g016070.2.1	Solyc10g017640.1.1
	Solyc04g025030.1.1	Solyc10g018070.1.1
	Solyc04g025050.1.1	Solyc12g005210.1.1
	Solyc04g025110.1.1	Solyc12g087810.1.1
	Solyc04g025970.1.1	Solyc12g087820.1.1
	Solyc04g076680.2.1	Solyc12g088080.1.1
Z. mays	Zm00001d003409_P001	Zm00001d036279_P001
	Zm00001d013258_P001	Zm00001d041587_P001
	Zm00001d015775_P001	Zm00001d042315_P001

	Zm00001d019189_P001 Zm00001d019289_P001 Zm00001d023405_P001 Zm00001d023409_P001 Zm00001d023739_P001 Zm00001d030375_P002 Zm00001d031399_P001 Zm00001d031625_P001 Zm00001d031626_P001 Zm00001d032218_P001 Zm00001d044899_P001 Zm00001d045227_P001 Zm00001d045697_P001 Zm00001d047355_P001 Zm00001d047356_P002 Zm00001d049897_P001 Zm00001d050388_P001 Zm00001d052534_P001 Zm00001d032219_P001

Supplemental Materials

All the following files can be found in the folder *MADS mip search files*.

Full list of all full-length MADS-box genes in the file miP_search/knownMADS_seedSeqs.fasta.

Full list of all full-length and truncated MADS-box genes in the file miP_search/miP_finalSeqs.fasta.

All individual alignments of the 14 miP candidates can be found in the folder individual_alignments.