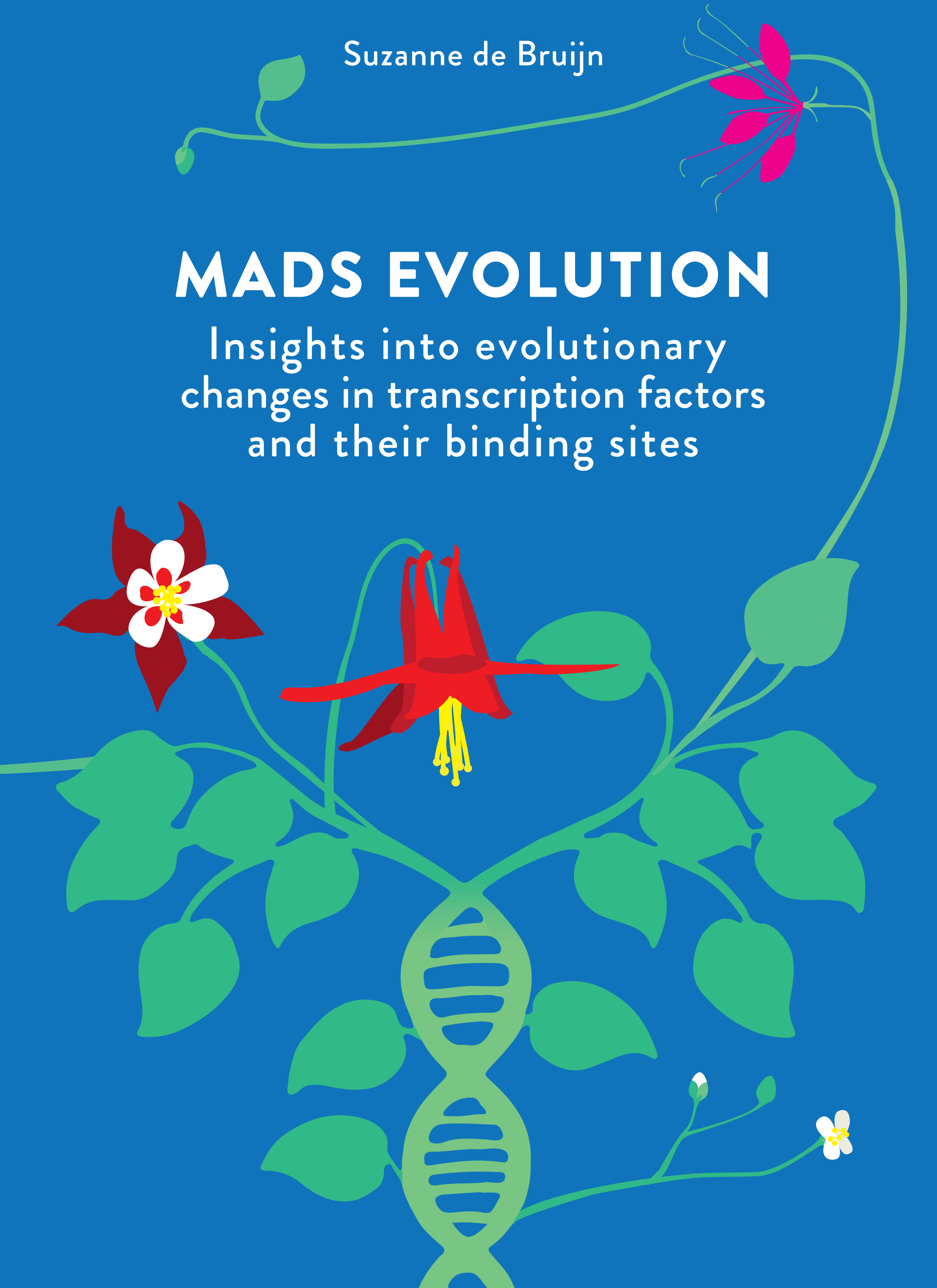
Suzanne de Bruijn

MADS EVOLUTION

Insights into evolutionary changes in transcription factors and their binding sites



Propositions

1. Subtle changes in interaction-affinities of transcription factors can lead to different gene regulation.
   (this thesis)

2. Transcription factor binding sites evolve faster in plants than in animals.
   (this thesis)

3. Examining the traits associated with domestication, we can conclude that humans domesticated themselves.

4. Lack of a mutant phenotype should lead to the conclusion that a more detailed examination is necessary, not that the gene acts redundantly.

5. To understand their own research, biologists need to become bioinformaticians as well.

6. Although we think of English as a universal language, cultural differences do influence the way we understand each other.

7. Too many rules do lead to the inactivation of common sense.

Propositions belonging to the thesis, entitled:

MADS evolution

insights into evolutionary changes in transcription factors and their binding sites

Suzanne de Bruijn
Wageningen, 20 November 2017

MADS evolution

Insights into evolutionary changes in transcription factors and their binding sites

Suzanne de Bruijn

Thesis committee

Promotor
Prof. Dr Gerco C. Angenent
Personal chair at the Laboratory of Molecular Biology
Wageningen University & Research

Co-promotor
Prof. Dr Kerstin Kaufmann
Personal chair at the Institute for Biology
Humboldt University of Berlin, Germany

Other members
Prof. Dr Eric Schranz, Wageningen University & Research
Prof. Dr Ronald Koes, University of Amsterdam
Prof. Dr Annette Becker, Justus-Liebig-Universität Gießen, Germany
Dr Barbara Gravendeel, University of Leiden/Naturalis

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences.

MADS evolution

Insights into evolutionary changes in transcription factors and their binding sites

Suze-Annigje de Bruijn

Thesis
Submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of Rector Magnificus
Prof. Dr A.P.J. Mol,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 20 November 2017
at 1.30 p.m. in the Aula.

Suzanne de Bruijn

MADS evolution. Insights into evolutionary changes in transcription factors and their binding sites

196 pages.

PhD thesis, Wageningen University, Wageningen, the Netherlands (2017)
With references, with a summary in English

ISBN 978-94-6343-670-0
DOI https://doi.org/10.18174/421346

Contents

Chapter 1..7
Introduction. How floral organs are specified

Chapter 2..27
Plant 'Evo-devo' goes genomic: from candidate genes
to regulatory networks

Chapter 3..39
Evolution of DNA-binding sites of a floral master
regulatory transcription factor

Chapter 4..69
Variation in Transcription Factor Binding Sites and
DNA accessibility between *Arabidopsis* species and ecotypes

Chapter 5..97
Evolution of PISTILLATA paralogs in *Tarenaya hassleriana*

Chapter 6..123
Specifying a novel floral organ in *Aquilegia*

Chapter 7..143
General discussion and perspectives

References.. 157
Summary.. 181

Popular article.. 185
"Research into flowers: comparing apples and oranges?"
Acknowledgments..188
List of publications 192
Curriculum Vitae...193
EPS Education statement... 194

CHAPTER 1

Introduction: How floral organs are specified

Angiosperms, or flowering plants, are one of the most diverse and successful groups of plants, with an estimated number of species between 250.000 and 425.000 (Crane et al., 1995; Bowman, 1997; Krizek and Fletcher, 2005; Jiao et al., 2011; Schranz et al., 2012). They originated 150-190 MYA (Magallon et al., 2015), and since then dominated most ecosystems. This group of plants is characterized by several new innovations, among them the flower and the fruit which gave angiosperms their name (angiosperms=enclosed seeds). These novelties increased the speed and robustness of the reproduction cycle. Many angiosperms use pollinators and seed dispersers, and the co-evolutionary changes that occurred between plants and pollinators helped creating the great diversity seen in angiosperm flowers (Chanderbali et al., 2016). The sudden rise in angiosperm species is known as Darwin's abominable mystery (Friedman, 2009).

Humans do not only enjoy the beauty of the flowers, but we also rely heavily on them for human nutrition, since flowers give rise to grains, vegetables, fruits and spicy seeds (Chanderbali et al., 2016). Most of the plants we eat have been domesticated, and the process of breeding for improved crops is ongoing. To aid this process, it is of great importance to understand the evolution of the angiosperm flower, and the underlying changes in gene regulation.

After introducing the angiosperm phylogenetic tree, I will discuss the major mechanisms of genome evolution. Then I will show how a simple genetic model underlies a wealth of different floral morphologies, and discuss the variations of this model that makes it generally applicable. Although the molecular basis of floral organ specification has been studied extensively over the last three decades, we still don't know how this knowledge relates to the final morphology of the flower. I will discuss some of the challenges that lay ahead in this field, before identifying the topics I have examined in this thesis.

A family tree in which some members are more alike than others

When the angiosperms emerged, they did not only quickly radiate into a huge number of species occupying a wide range of habitats, but they also evolved a variety of different floral morphologies. Fossil evidence indicates that early angiosperm flowers were small and simple, without a differentiated perianth (Endress, 1994; Crane et al., 1995; Friis et al., 2001). Indeed, the sister species to all other angiosperms, *Amborella trichopoda (*the only species in the order Amborellales*)* displays small and simple flowers. Only measuring a few mm, these unisexual flowers consist of variable numbers of reproductive organs and tepals, which are not clearly differentiated (Buzgo et al., 2004). The next order in the angiosperm family tree is the Nymphaeales, containing the water lilies. Species in this order exhibit a range of floral morphologies, from small and simple, to large and showy flowers. Together with the small clade of Austrobaileyales, these two orders make up the ANA-grade (Amborellales, Nymphaeales, Austrobaileyales) of basal angiosperms (Mathews and Donoghue, 1999; Parkinson et al., 1999; Qiu et al., 1999; Soltis et al., 1999; Soltis and Soltis, 2004).

Besides these basal angiosperms, there are five well-supported monophyletic groups of angiosperms: the Chloranthales, Magnoliids, Ceratophyllales, the monocots and the eudicots. The relation between these groups is not well resolved, but several studies propose a sister relationship between the monocots and the eudicots (Crane et al., 1995; Moore et al., 2007). These two clades together, the monocots and the eudicots, represent 97% of all angiosperms (Moore et al., 2007). Members of these clades have distinct but standardized flowers. The monocots, containing 22% of all angiosperm species, generally have flowers in which each whorl contains three organs (trimerous) (Crane et al., 1995; Soltis and Soltis, 2004). The eudicots form the largest group of angiosperms, containing approximately 75% of all angiosperm species (Drinnan et al., 1994; Crane et al., 1995). This clade consists of the basal eudicots and the core eudicots, which diverged around 115 MYA (Sanderson et al., 2004). The core eudicots have a standardized “bauplan” typically consisting of four whorls of floral organs. They usually show a clear differentiation of the perianth in sepals and petals, with four or five organs per whorl. This leads to a standard organization of a whorl of sepals, followed by petals, stamens and an innermost whorl of carpels (Zahn et al., 2005; Soltis and Soltis, 2016). While their organ numbers and organization are relatively standardized, core eudicot flowers do show extensive morphological differences in colour, shape and size (Chanderbali et al., 2016).

There are seven major eudicot clades, of which the asterids and the rosids are the largest, each containing about one third of all angiosperms (Soltis and Soltis, 2004). Even though the basal angiosperms and the Magnoliids represent only a few percent of all angiosperms, they do exhibit a tremendous amount of floral diversity. They do not only show a huge diversity in both number and arrangement of floral organs, but they also have a more flexible ‘bauplan’. Many of these species show variability in floral organ numbers within a single plant (Crane et al., 1995; Zahn et al., 2005; Chanderbali et al., 2016).

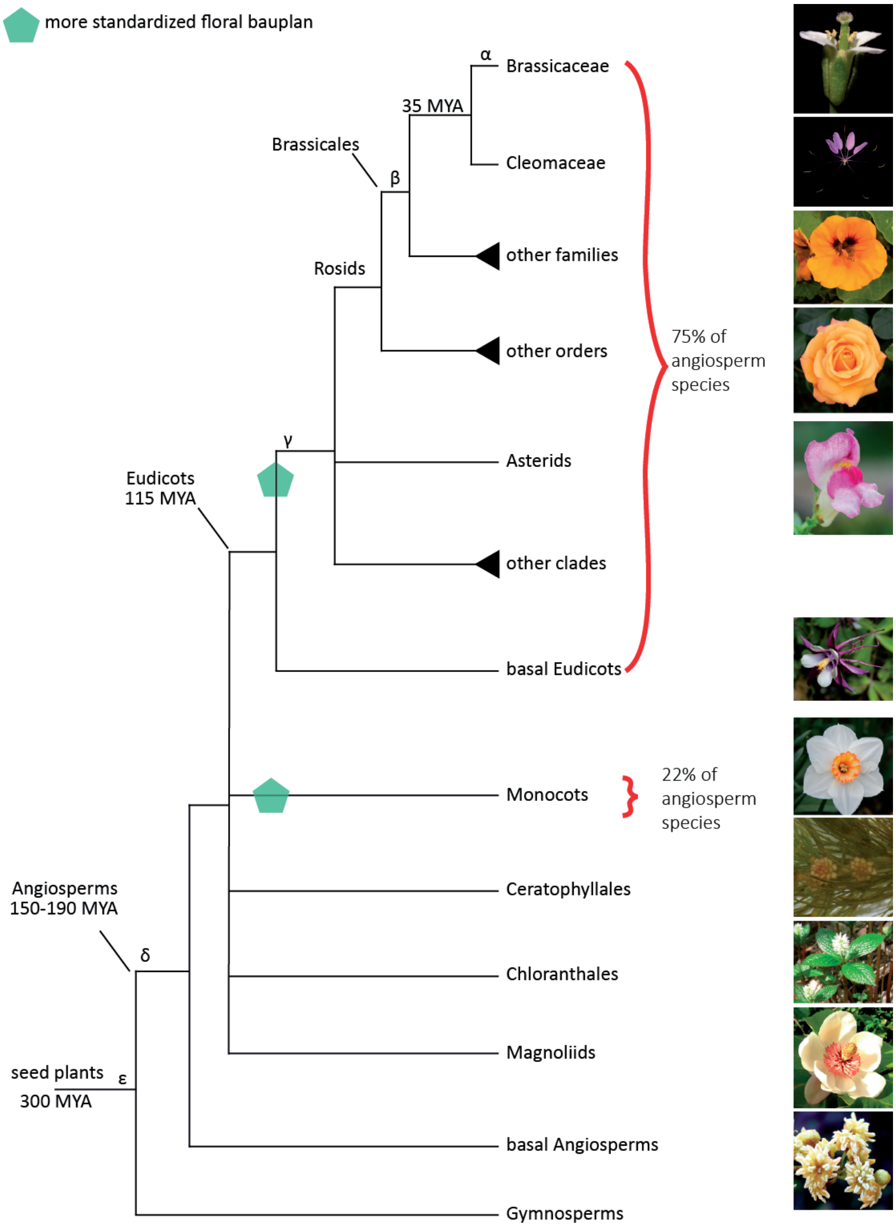


Figure 1: Schematic angiosperm phylogenetic tree. *Mentioned are all genome duplications in the lineage leading to the genus Arabidopsis. On the right side, representative flowers of the corresponding angiosperm families are shown, as example of the wealth of floral morphology. From top to bottom: A. thaliana (Brassicaceae), T. hassleriana (Cleomaceae), Tropaeolum majus (Tropaeolaceae, Brassicales), Rosa, (Rosaceae, rosids), Antirrhinum majus (asterids), Aquilegia (basal eudicot), Narcissus (Monocot), Ceratophyllum demersum, male flower (Ceratophyllales), Chloranthus japonicus, (Chloranthales) Magnolia (magnoliids), Amborella trichopoda, male flower (basal angiosperm).*

Extensive research has been done to understand the genetic basis of floral organ specification, but most of this has been focussed on a few species within the eudicot clades of rosids (*Arabidopsis thaliana*) and asterids (*Antirrhinum majus* and the Solanaceae) and on some monocot species (maize, rice). For evolutionary studies however, it is important to look beyond these model systems. Interesting species from an evolutionary perspective are species that are closely related to the model species, as well as species belonging to sister families. For instance, there is quite some interest in *Arabidopsis lyrata*, being the closest relative of *Arabidopsis thaliana*. Also the Cleomaceae, as sister family to the Brassicaceae is gaining interest (Cheng et al., 2013). Other important species are those that are occupying an informative position in the phylogenetic tree. Examples are *Amborella trichopoda*, as the sister clade to the rest of the angiosperms (Buzgo et al., 2004), and the Ranunculales, which includes the floral model genus *Aquilegia*, because of their position as sister clade to all other eudicots. The Ranunculales are not only interesting for their phylogenetic position, but also show a striking diversity of floral architecture. Besides differences in number, morphology and phyllotaxy of floral organs, the Ranunculales also contain species that developed novel organs such as nectar spurs and staminodia (Kramer, 2009; Becker, 2016).

Origin of genetic diversity

In general, changes to the genome must underlie changes in morphology. These changes can be single nucleotide polymorphisms, small insertions or deletions (indels), genomic rearrangements and duplications of (part of) the genome.

The effect of a mutation depends on whether it is in a coding- or non-coding region. Mutations in a coding region can change the protein sequence, and thereby its function, while mutations in non-coding sequences will not change protein sequences, but might change the expression of the gene by mutating *cis*-regulatory elements (CREs).

In case of transcription factors (TFs), mutations that affect their expression (by changing CREs) are more likely to be fixed than mutations that are in the coding sequence (Carroll, 2008; Stern and Orgogozo, 2008, 2009). TFs act by regulating the expression of many genes, and therefore, changes in TF function are likely to have an impact on many or all of its target genes. This will create pleiotropic effects that are most likely disadvantageous for the organism. In contrast, changes in CREs will not change the function of a TF and may have a more subtle effect by modifying only its expression pattern. To change a developmental pathway, the expression of many genes needs to be changed in concert. This can be achieved by changing the expression of the upstream TF, which consequently affects the entire set of downstream targets. This hypothesis is explained in more detail in chapter two.

Gene duplication enables protein diversification without resulting in pleiotropic or disastrous effects. Immediately after duplication, the two paralogs will be identical, which releases the evolutionary constraint on these genes. However, having two fully redundant paralogs is not believed to be a stable situation and therefore it is unlikely that two identical genes are retained after duplication (Nowak et al., 1997). Instead, it was postulated that there are three options for long-term fate of paralogs: subfunctionalization, neofunctionalization or pseudogenization (Ohno, 1970; Force et al., 1999). Pseudogenization will occur when a gene acquires deleterious mutations that render it non-functional. This normally happens quickly, in the first few million years after the duplication. However, calculations of half-lives of duplicated genes vary greatly between species; whereas the average half-life in fungi is 1 MY (Million Years) and animal species have an average half-life of 4 MY, in *Arabidopsis* the half-life is calculated to be 17.3 MY on average (Lynch and Conery, 2003). It is not known whether a long half-life is common for all plant species, or that *Arabidopsis* is an exception (Moore and Purugganan, 2005). Some genes however, only pseudogenize after a long time, having persisted for hundreds of millions of years before pseudogenizing (Zhang, 2003; Zou et al., 2009). When both paralogs are retained, one or both genes might gain a new function, this is known as neofunctionalization. Another possibility is that the ancestral gene function is divided between the two paralogs, which is called subfunctionalization. Subfunctionalization can occur at the level of protein function and/or at the level of spatiotemporal expression, with changes in expression being more likely (Huminiecki and Wolfe, 2004; Wapinski et al., 2007). As deleterious mutations happen more often than beneficial ones, it is suggested that rapid subfunctionalization is the first thing that happens after duplication, after which neofunctionalization may also play a role (He and Zhang, 2005).

In plants, the main mechanism of gene duplication is polyploidization, or whole genome duplication (WGD). WGDs are common in the history of plants, and all seed plants underwent at least one WGD (Lawton-Rauh, 2003; Cui et al., 2006; Barker et al., 2009; Soltis et al., 2014). Flowering plants went through a genome duplication before the origin of the seed plants (the ε duplication), as well as a genome duplication before the origin of angiosperms (the δ duplication) (see **Figure 1**) (Bowers et al., 2003; Jiao et al., 2011; Li et al., 2015). Besides these ancient duplications, most angiosperms also went through additional, more recent duplications (Blanc and Wolfe, 2004b; Paterson et al., 2010). For instance, *Arabidopsis* is thought to have three additional duplications (Simillion et al., 2002; Blanc et al., 2003). *Arabidopsis* experienced a genome duplication that was Brassicaceae-specific (α duplication), as well as one that was shared with all Brassicales apart from *Carica papaya* (β duplication)(Ming et al., 2008). Furthermore, there has been a genome triplication before the core eudicots (γ triplication) (**Figure 1**) (Jiao et al., 2012). Several other plant lineages also experienced lineage-specific duplications. Monocots went through two monocot-specific WGD, and it has been estimated that at least 50 independent ancient WGDs are distributed across the angiosperm phylogeny (Jiao et al., 2011; Soltis et al., 2014). In several families, like the Asteraceae, Brassicaceae, Cleomaceae, and Fabaceae, ancient WGDs seem associated with an increase in plant diversity (Soltis et al., 2009; Schranz et al., 2012). It seems therefore, that genome duplications are a potential driving force for evolutionary diversity (Lawton-Rauh, 2003; Moore and Purugganan, 2005). Interestingly, key angiosperm innovations like seeds and flowers coincided with genome duplications. These innovations led to radiation of species, and ultimately led to the dominance of the seed- and angiosperms (Jiao et al., 2011; Soltis and Soltis, 2016).

After a WGD, species often turn back to being diploid in a process called diploidization, which involves genomic rearrangements and deletions. In this process, many of the duplicated genes are lost (Wolfe, 2001). However, some classes of genes are preferentially retained after WGDs. It has been shown that a major reason for preferential gene retention is to keep balanced protein levels relative to interaction partners. This is known as the dosage balance hypothesis, which states that after a whole genome duplication, paralogs can be retained to achieve similar concentrations as their (also duplicated) interaction partners (Freeling and Thomas, 2006). Indeed, preferentially retained classes of genes often act in concert with partners, and include gene classes such as ribosomal proteins, protein kinases, and TF (Blanc and Wolfe, 2004a; Conant and Wolfe, 2008; Paterson et al., 2010). This retention due to dosage balance ensures that genes are not lost quickly, which therefore gives these genes more time, and therefore more opportunity to evolve.

Preferential retention of TFs after a WGD means there is an increased number of TFs present in the genome, which over time will evolve new functions. This expansion of TF families after WGDs might be involved in the evolution of morphological complexity. One family of TFs that expanded greatly due to genome duplications is the MADS-domain TF family (Irish, 2003; Geuten et al., 2011). This family is involved in many developmental processes in the life cycle of plants, including in flower and fruit specification.

MADS-domain proteins

The MADS-domain TF family is characterized by the DNA-binding MADS-domain. This domain is named after the first identified MADS-domain TFs: *MINICHROMOSOME MAINTENANCE 1* (*MCM1)* in yeast (Passmore et al., 1988), *AGAMOUS* (*AG*) in *Arabidopsis thaliana*, *DEFICIENS* (*DEF*) in *Antirrhinum majus* and *SERUM RESPONSE FACTOR* (*SRF)* in humans (Norman et al., 1988; Schwarz-Sommer et al., 1990). The MADS-domain TF family is one of the larger TF families in plants; angiosperms have around 100 of these TFs, with 107 genes being found in *Arabidopsis* (Parenicova et al., 2003; Gramzow and Theissen, 2010).

In plants, MADS-domain proteins can be divided into two groups: type I and type II (Alvarez-Buylla et al., 2000). Type I MADS-domain proteins have a single conserved domain, the MADS-domain, and these proteins are usually encoded by a single exon. The function of these proteins only started to be elucidated in the last decade, and they seem to act mainly in plant reproduction (Masiero et al., 2011).

The type II MADS-domain TFs have been studied for a long time. They are involved in all aspects of development, with functions in roots, floral transition and, most famously, floral development (Smaczniak et al., 2012a). These type II MADS-box genes have a different structure than the class I proteins, as they are generally encoded by seven exons. Besides the MADS-domain, class II proteins contain additional domains: following the MADS-domain they have an intervening (I-) domain, a conserved K-domain and a divergent C-terminal domain, which led to these genes also being classified as MIKC-type MADS TFs.

The different domains of MIKC-type TF have specific functions. The MADS-domain is needed for DNA-binding and dimerization (Krizek and Meyerowitz, 1996a), and the I-domain is involved in determining dimerization specificity (Riechmann et al., 1996b). As the I-domain shows relatively high within-subfamily conservation, it might be involved in subfamily-specific functions (Kaufmann et al., 2005). The K-domain is named after keratin, as it shows structural similarities to the coiled-coil domain of keratin. This domain consists of three amphiphatic helices that are an important oligomerization surface for dimerization and tetramerization (Yang and Jack, 2004; Kaufmann et al., 2005; Melzer and Theissen, 2009; Melzer et al., 2009; Puranik et al., 2014). The C-terminal domain is the least conserved domain of MIKC-type TFs (Munster et al., 1997), but is more conserved within each subclade of MIKC-proteins. Although this domain might be needed for higher-order complex formation, experiments to test this gave contradictory results (Munster et al., 1997).

MIKC TFs bind to DNA as dimers, and all bind to similar sequences, called CArG-boxes with the consensus sequence $CC[A/T]_6GG$ (Huang et al., 1993). Although they all bind similar sequences, different members of the TF family do bind to slightly different sequences *in vitro* (Riechmann et al., 1996a; Smaczniak et al., 2017). The importance of these differences for *in vivo* functioning is questioned however. It has been shown that the functional specificity of the floral MADS TF is determined by the MI- or the IK domains (depending on which subclade of MIKC-MADS TF) (Krizek and Meyerowitz, 1996a). Experiments that swapped the plant MADS-domain for the MADS-domain of a human TF also suggested that the *in vivo* specificity does not reside in the MADS-domain, despite its importance for DNA binding. Although these chimeric proteins will adopt the *in vitro* DNA-specificity of the protein that donated the MADS-domain, they still complement the mutant phenotype of the IKC-donor protein (Riechmann and Meyerowitz, 1997).

Besides binding the DNA as dimers, MADS-domain TFs can also form tetramers that bind to two CArG-boxes, and thereby looping the DNA (Egea-Cortines et al., 1999; Theissen and Saedler, 2001; Mendes et al., 2013). In addition, is has been shown that they have interactions with a range of other proteins, among them members of other TF families and chromatin remodelers (Smaczniak et al., 2012b).

Although MADS-domain TFs can be found in all eukaryotes, in protists, fungi and animals this TF family has only a few members. In some plant lineages however, this TF family has expanded considerably. Whereas in extant green algae there is only one MADS-domain TF present, in eudicots this TF family comprises at least 100 members (Gramzow and Theissen, 2010). The largest expansion occurred during the evolution of seed plants. The common ancestor of seed plants had at least four MIKC-type MADS-domain TFs, while *Arabidopsis* has 45 and in some other eudicot species even a larger number of MIKC-type genes have been found (Munster et al., 1997; Krogan and Ashton, 2000; Theissen et al., 2000; Parenicova et al., 2003). This increase is partially due to whole-genome-duplications (Krogan and Ashton, 2000). These duplications coincided with the origin of new plant structures, like seeds and flowers. As MADS-domain TFs play fundamental roles in specifying these structures, it seems that duplication of these genes is highly correlated with the origin of these structures (Theissen et al., 2000).

Flower development, one model fits all?

At the end of the eighties, scientists started to use mutants that showed homeotic conversions of floral organs, to study floral development (Haughn and Somerville, 1988; Komaki et al., 1988; Bowman et al., 1989; Hill and Lord, 1989; Kunst et al., 1989; Bowman et al., 1991). This work was mainly done on *Arabidopsis thaliana* and *Antirrhinum majus*, and the mutants that were studied fell in three classes, each class displaying phenotypes in two adjacent whorls. These studies led to a model for floral specification: the ABC-model.

The model

The first class of mutants (A class) affected whorl one and two, consisting of sepals and petals, respectively. In *Antirrhinum*, the *ovulata* (*ovu*) mutant fell in this class, which showed sepal to carpel and petal to stamen conversions (Carpenter and Coen, 1990). In *Arabidopsis*, two mutants were found; the *apetala2* (*ap2)* mutant showing a sepal to carpel transformation, and petals are either absent or are staminoid in nature (Bowman et al., 1989; Kunst et al., 1989; Bowman et al., 1991). The other *Arabidopsis* mutant, *ap1*, has bracts instead of sepals, and either no organs in the second whorl, new floral buds or mosaics of leaf- and stamen-like tissue (Irish and Sussex, 1990; Mandel et al., 1992). A second class (B class) of mutants affected whorl two and three, where normally petals and stamens form. This class of mutants shows homeotic conversion of petals into sepals, and stamens into carpels (Bowman et al., 1989; Sommer et al., 1990). Two mutants were found with this phenotype, *apetala3* (*ap3)* and *pistillata (pi)* in *Arabidopsis*, and *deficiens* (*def)* and *globosa* (*glo)* in *Antirrhinum* (Tröbner et al., 1992; Goto and Meyerowitz, 1994). The third class (C class) of mutants affected whorls three and four; in this class of mutants, stamens are converted into petals and the carpels are replaced by a new flower bud in a reiterative way, indicating a function in meristem determinacy (Bowman et al., 1989; Carpenter and Coen, 1990; Yanofsky et al., 1990). Except *AP2*, the genes underlying these mutant phenotypes encode TFs belonging to the MIKC-type MADS-domain TFs (Sommer et al., 1990; Yanofsky et al., 1990).

Based on the fact that these phenotypes always affected two adjacent whorls, and because of the genetic interactions between these genes, a model for floral specification was proposed, termed the ABC-model (**Figure 2A**) (Haughn and Somerville, 1988; Bowman et al., 1991; Coen and Meyerowitz, 1991). This model stipulates that the three classes of genes together specify four different types of floral organs. A-function by itself would give rise to sepals, whereas A- in combination with B-function determines petal identity. B- and C- functions together specify stamens, whereas C-function alone gives rise to carpels. In addition, A- and C-function genes inhibit each other's expression.

However, ectopic expression of the ABC genes did not convert leaves into floral organs, showing that the ABC genes are not sufficient to specify flowers. It was shown that another class of MADS-box genes, the *SEPALLATA (SEP)* genes, was also needed for floral development. *Arabidopsis* has four *SEP* genes; single mutants do not show a phenotype, but in the triple mutant of *sep1/2/3* all floral organs are transformed into sepals, whereas the quadruple *sep1/2/3/4* mutant consists of leaves in a floral (whorled) phyllotaxy (Pelaz et al., 2000; Ditta et al., 2004). These genes therefore encompass a function that is needed in all floral whorls, and was termed the E-class function (Theissen, 2001). The D-function was already assigned to genes fulfilling a role in ovule specification (Angenent et al., 1995; Colombo et al., 1995). Whereas overexpressing of combinations of the ABC-class genes was not sufficient to convert leaves into floral organs, adding ectopic expression of *SEPALLATA3* (*SEP3*) led to the conversion of leaves into floral organs. For example, the combination of B- and C class genes with *SEP3* did lead to conversion of leaves into staminoid organs. These experiments showed that the E-class function is needed for floral organ specification, and that expression of a combination of ABCE-class genes is sufficient to specify floral organs (Honma and Goto, 2001; Pelaz et al., 2001).

Taking it to the molecular level

The next question was how this genetic model functions at the molecular level. *In vitro* studies showed that not all necessary combinations of proteins form heterodimers, refuting the idea that the ABCE-class genes specify floral organs by acting in different combinations of heterodimers (Riechmann et al., 1996b). Egea-Cortines et al. (1999) showed that MADS-domain TFs can form multimeric protein complexes. Based on these results, it was suggested that two dimers interact with each other, resulting in a tetrameric complex. This hypothesis that MADS-domain TFs act in organ-specific tetramers was formalized as the "floral quartet" model (Honma and Goto, 2001; Theissen and Saedler, 2001). LC-MS based complex isolation experiments from plants confirmed the interaction predicted by the floral quartet model *in planta* (Smaczniak et al., 2012b), strengthening this molecular model.

Modifications to the ABCE-model

Since the ABCE-model was based on data from only two species, the core eudicots *Arabidopsis thaliana* and *Antirrhinum majus* (Schwarz-Sommer et al., 1990; Coen and Meyerowitz, 1991; Weigel and Meyerowitz, 1994), an intriguing question was whether the model could be generalized for all angiosperms.

The first problem found with the model was the A-function. It appeared that even *Antirrhinum* does not have a gene that specifies the A-function; the A-class mutant *ovulata* was actually a dominant mutant that mimicked the phenotype due to ectopic expression of the C-class gene *PLENA* (Bradley et al., 1993; Lönnig and Saedler, 1994). Although *Arabidopsis* does have genes affecting the first two whorls, the phenotypes of these mutants are not identical to the phenotype that is predicted by the ABC-model. It was suggested that these *Arabidopsis* phenotypes might be caused by incorrectly specified floral meristems (Litt, 2007; Causier et al., 2010a). Therefore, the A-function may not be specified by specific genes, but is a default state of the floral meristem. As a result, it was proposed to modify the ABCE model to the (A)BCE-model. In this model, the B- and C-class functions are identical to the original ABCE-model. The A-function however, is modified and encompasses the function of specifying floral meristem identity, and to specify the expression boundaries of B- and C-class genes (Causier et al., 2010a).

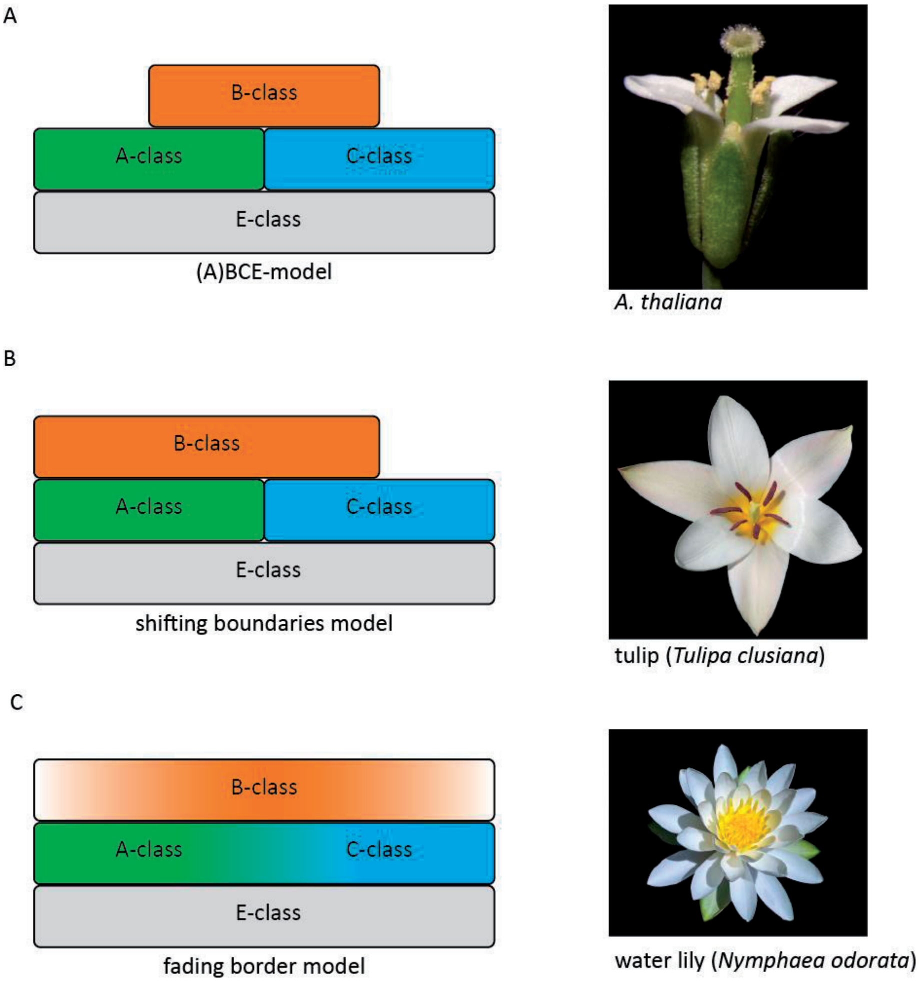


Figure 2: Schematic representation of the (A)BCE-model. A: *The (A)BCE-model;* ***B:*** *the shifting borders model;* ***C:*** *The fading borders model. Examples of flowers obeying these models are shown at the right side.*

This modified model is very similar to a model proposed in the early days of floral organ specification research, which suggested only two functions to specify floral organs (Schwarz-Sommer et al., 1990; Litt, 2007; Causier et al., 2010a).
This modified (A)BCE-model seems directly applicable in all other eudicots (Bowman, 1997; Soltis et al., 2007; Causier et al., 2010a; Litt and Kramer, 2010). There are species-specific differences however, partially due to gene duplications and subsequent subfunctionalization (Vandenbussche et al., 2004; de Martino et al., 2006; Geuten and Irish, 2010; Sharma and Kramer, 2013).

In monocots, orthologs of the B- and C-class genes were also found. Studies to determine whether these orthologs are conserved in function were complicated due to differences in floral organ identities in monocots. The first monocots that were examined were the grasses maize and rice. Grasses have a peculiar perianth that consist of palea/lemma in whorl one and lodicules in whorl two, instead of sepals and petals like eudicot flowers. Nevertheless, the B- and C-class genes seem to be conserved in function, with the C-class genes specifying stamens and carpels, and the B-class genes are needed to specify lodicules in the second whorl, and stamens in the third whorl (Schmidt et al., 1993; Chung et al., 1995; Mena et al., 1996; Kang et al.; Ambrose et al., 2000; Nagasawa et al., 2003; Whipple et al., 2004). Although the perianth of other monocots bears more resemblance to the perianth of eudicot flowers, their perianth is not always clearly differentiated. Instead of a whorl of sepals and a whorl of petals, they often have two whorls of “tepals”. To adhere to the (A)BCE-model it was hypothesized that if these tepals are petaloid in morphology, they should also express B-class genes. This shift of B-class expression to the outer whorl was indeed found for several (but not all) monocots with petaloid tepals (Kanno et al., 2003; Nakamura et al., 2005; Otani et al., 2016). To accommodate these changes in expression, the “shifting boundaries model” was proposed (**Figure 2B**) (van Tunen et al., 1993; Bowman, 1997). This shifting boundary model is not restricted to the monocots, and mostly these shifts of expression affect the B-class genes (Theissen and Melzer, 2007).

Basal angiosperms also possess orthologs of the B- and C-class genes. However, they do not seem to be expressed in the same clearly defined domains as in eudicots and monocots. Instead of sharp borders, the expression patterns are often more fuzzy and overlapping. These fuzzy expression patterns in the basal angiosperms can be correlated with the gradual changes in organ morphology/identity that are displayed by these species. Often, basal angiosperms do not show several clearly distinct organs types, but instead display a range of chimeric organs gradually transitioning between organ types. This gradual change of expression of the (A)BCE-genes is incorporated into a modified version of the model called the “fading border” model (**Figure 3C**) (Buzgo et al., 2004; Soltis et al., 2007).

The B- and C-class genes are not only conserved in angiosperms, but homologs of these genes have also been found in gymnosperms. Like in angiosperms, in gymnosperms these genes are also involved in the specification of reproductive structures. C-class homologs are expressed in both male and female structures, whereas B-class genes are expressed in male structures only (Winter et al., 1999; Theißen and Becker, 2004). It is therefore hypothesized that these genes already specified reproductive structures in the common ancestor of angiosperms and gymnosperms.

Although some modifications are made to the original ABC-model, in general the model appeared to be generally applicable in angiosperms. It appears that the B- and C-functions are widely conserved in angiosperms. However, the expression of these genes has evolved from broad and fuzzy in early angiosperms to clear defined expression patterns in more derived lineages. It is suggested that this restriction in expression is due to the establishment of positive autoregulation of the genes in the (A)BCE-model. Positive autoregulation can amplify expression level differences, which may lead to sharp borders of expression (Buzgo et al., 2004; Theissen and Melzer, 2007).

The extraordinary B-class genes

Among the floral MADS-domain proteins, the B-class proteins are a special case. The B-class does not seem to have pleiotropic effects. Whereas C-class genes are involved in meristem termination as well as organ specification, the sole function of the B-class genes is specifying petals and stamens. This absence of pleiotropic effects suggests that they are allowed to evolve more freely, and could therefore contribute significantly to floral morphological diversity. Indeed, these genes seem to evolve faster than the other MIKC-type MADS-domain protein lineages (Purugganan et al., 1995; Purugganan, 1997). They are also an exception in the floral MADS world as they act as obligate heterodimers. Last but not least, the B-class genes are regulated by the B-class proteins in a positive autoregulatory loop. It is fascinating that B-class genes got involved in petal specification, as the petal is a new organ that was not present in the last common ancestor with gymnosperms. Also, B-class genes are sometimes recruited to specify a new type of organ, the staminodium, found in a whorl between the stamens and carpels (Kramer et al., 2007).

The B-class function is specified by members of the *DEFICIENS*/*GLOBOSA* (*DEF*/*GLO*) clade, named after the first members that were molecularly characterized; *DEFICIENS* (Sommer et al., 1990) and GLOBOSA (Tröbner et al., 1992) in *Antirrhinum*. Their orthologs in *Arabidopsis* are *APETALA3* (*AP3*) and *PISTILLATA* (*PI*), respectively (Jack et al., 1992; Goto and Meyerowitz, 1994).

The ancestor of the B-class genes was duplicated before the origin of the angiosperms, which gave rise to the *PI*- and paleo*AP3*-lineage (Kramer et al., 1998; Kim et al., 2004; Hernández-Hernández et al., 2007). At the origin of the core eudicots, paleo*AP3* underwent another duplication, which gave rise to *euAP3* and *TOMATO MADS BOX GENE6* (*TM6)* (Kramer et al., 1998; Causier et al., 2010b). Besides several diagnostic amino acids, these three gene lineages can be clearly distinguished by their C-terminal motifs.

Genes in the *PI* lineage have a PI motif in their C-termini, which shows very strong conservation (Kramer et al., 1998). *AP3* genes have another motif in their C-termini, the PI-derived motif. In addition, the *AP3* genes from monocots, basal angiosperms and basal eudicots have their own motif, which is called the paleoAP3 motif (Kramer et al., 1998). Core eudicots have the euAP3 motif instead, which evolved from the paleoAP3 motif through a frameshift mutation. This frameshift occurred after the gene duplication that gave rise to the eu*AP3* and *TM6* lineages (Vandenbussche et al., 2003). This (paleo)AP3 motif is lost in the *PI* lineage, possibly through a single truncation event (Kramer et al., 1998). The fact that these motifs are conserved suggests that they are critical for the function of the proteins, and a fair amount of research has focussed on the function of these motifs. However, this effort has remained inconclusive so far. Complementation experiments have been done with proteins with C-terminal deletions or swapped motifs, mutants that miss the C-terminus have been studied, and protein-protein interaction studies were performed. All these experiments have given conflicting results, either showing that the motifs are necessary (Tzeng and Yang, 2001; Lamb and Irish, 2003; Lange et al., 2013; Mao et al., 2015), or that these C-terminal motifs are dispensable for protein function (Whipple et al., 2004; Berbel et al., 2005; Rijpkema et al., 2006; Piwarzyk et al., 2007; Benlloch et al., 2009; Causier et al., 2010b).

Expression and regulation of B-class genes

Genes of the Eu*AP3* and *PI* lineages are usually only expressed in petals and stamens (Jack et al., 1992; Schwarz-Sommer et al., 1992; Tröbner et al., 1992; Goto and Meyerowitz, 1994). However, early in flower development *PI* transcripts can also be found in the carpel primordia in *Arabidopsis* (Tröbner et al., 1992; Goto and Meyerowitz, 1994), while at that stage, *AP3* is also expressed at low levels in sepals in *Arabidopsis* (Weigel and Meyerowitz, 1993), and in both sepals and carpels in *Antirrhinum* (Schwarz-Sommer et al., 1992; Goto and Meyerowitz, 1994). However, even though the initial expression patterns are slightly different, later expression is found in whorl two and three only, and the AP3 and PI proteins can only be detected in domains where both genes are expressed (Krizek and Meyerowitz, 1996b). *AP3* and *PI* expression in petals and stamens is fairly standard in all species examined, especially in the eudicots. Although there are several examples of duplications followed by sub- or neofunctionalization leading to altered gene expression patterns (Bowman, 1997; Zahn et al., 2005), in general the combined expression of the paralogs is in petals and stamens.

It appears that both *PI* and *AP3* are regulated in two steps. First, they are activated independently, but after this initial activation, expression is maintained high through an autoregulatory feedback loop (Tröbner et al., 1992; Goto and Meyerowitz, 1994; Jack et al., 1994; Krizek and Meyerowitz, 1996b; Hill et al., 1998; Tilly et al., 1998; Honma and Goto, 2000). This autoregulation of B-class genes is conserved across the core eudicots (Becker, 2016).

Evolution of obligate heterodimerization

AP3 and PI are interesting among the MADS-domain proteins, as they form obligate AP3-PI heterodimers in core eudicots (Schwarz-Sommer et al., 1992; Tröbner et al., 1992; Riechmann et al., 1996b), which is necessary for DNA binding as well as for the translocation into the nucleus (McGonigle et al., 1996). It is interesting that these proteins act as obligate heterodimers, as the genes coding for these proteins originated from a duplication event. Outside of the core-eudicots however, homodimerization of *AP3*, and/or *PI* can also be found (Tzeng and Yang, 2001; Kramer et al., 2007). This leads to the question how and when obligate heterodimerization evolved.

As mentioned, the *AP3* and *PI* lineages originate from a gene duplication before the origin of the angiosperms, a duplication that is not shared with the gymnosperms. Gymnosperms have B-class genes which can form homodimers (Sundstrom and Engstrom, 2002; Winter et al., 2002), indicating that homodimerization is likely the ancestral state. This means that obligate heterodimerization is a derived feature, and this is probably due to compensatory mutations in both *AP3* and *PI* (Puranik et al., 2014).

The fact that the B-class proteins of *Amborella* (sister clade to all angiosperms) are already capable of forming DEF-GLO heterodimers indicates that heterodimerization was already established at the base of the angiosperms (Melzer et al., 2014). It is not certain however, whether this heterodimerization was already obligatory. We assume that DEF-GLO heterodimers are necessary to specify petals and stamens. Therefore, it is interesting to notice that homodimerization is still possible *in vitro* in several angiosperm lineages, for instance some basal angiosperms, monocots and basal eudicots (Tzeng et al., 2004; Kramer et al., 2007; Melzer et al., 2014; Bartlett et al., 2016). However, whether these homodimers also exist *in planta* and have a biological function remains an open question.

Examples of evolution of B-class gene duplicates

The duplications generating *AP3*/*PI* and later *AP3*/*TM6* are not the only duplications of the B-class genes. In fact, family-specific duplications of both *AP3*- and *PI*-lineages occur with high frequency (Kramer et al., 1998). When both paralogs are retained after a duplication, they will most likely subfunctionalize. How the ancestral protein function is divided between the paralogs however, is specific for each duplication event. Besides subfunctionalization, these duplications can also lead to neofunctionalization, generating morphological novelty or even new types of organs. Below, I provide some examples that illustrate the different possibilities after gene duplication.

The Solanaceae have two paralogs of both B-class genes. They possess an eu*AP3* gene as well as a *TM6* gene, which originated before the radiation of the higher eudicots (Kramer et al., 1998). Also, Solanaceae have two *PI* paralogs as a result of a duplication at the origin of the core asterids (Viaene et al., 2009).

Petunia hybrida did not seem to follow the (A)BCE-model, as the *Phdef* mutant showed a homeotic conversion in the second whorl only (van der Krol et al., 1993). This led to a detailed study of Petunia's B-class genes. It was found that the *Phdef* mutant did not have a third whorl phenotype, because in whorl three *PhTM6* acts redundantly with *PhDEF*. However, *PhTM6* is not expressed in petals, which are therefore specified by *PhDEF* alone (Vandenbussche et al., 2004; Rijpkema et al., 2006). Petunia also has two *GLO* genes, but their function is largely redundant in both stamens and petals, since homeotic conversions are only observed in the double mutant (**Figure 3A**) (Vandenbussche et al., 2004).

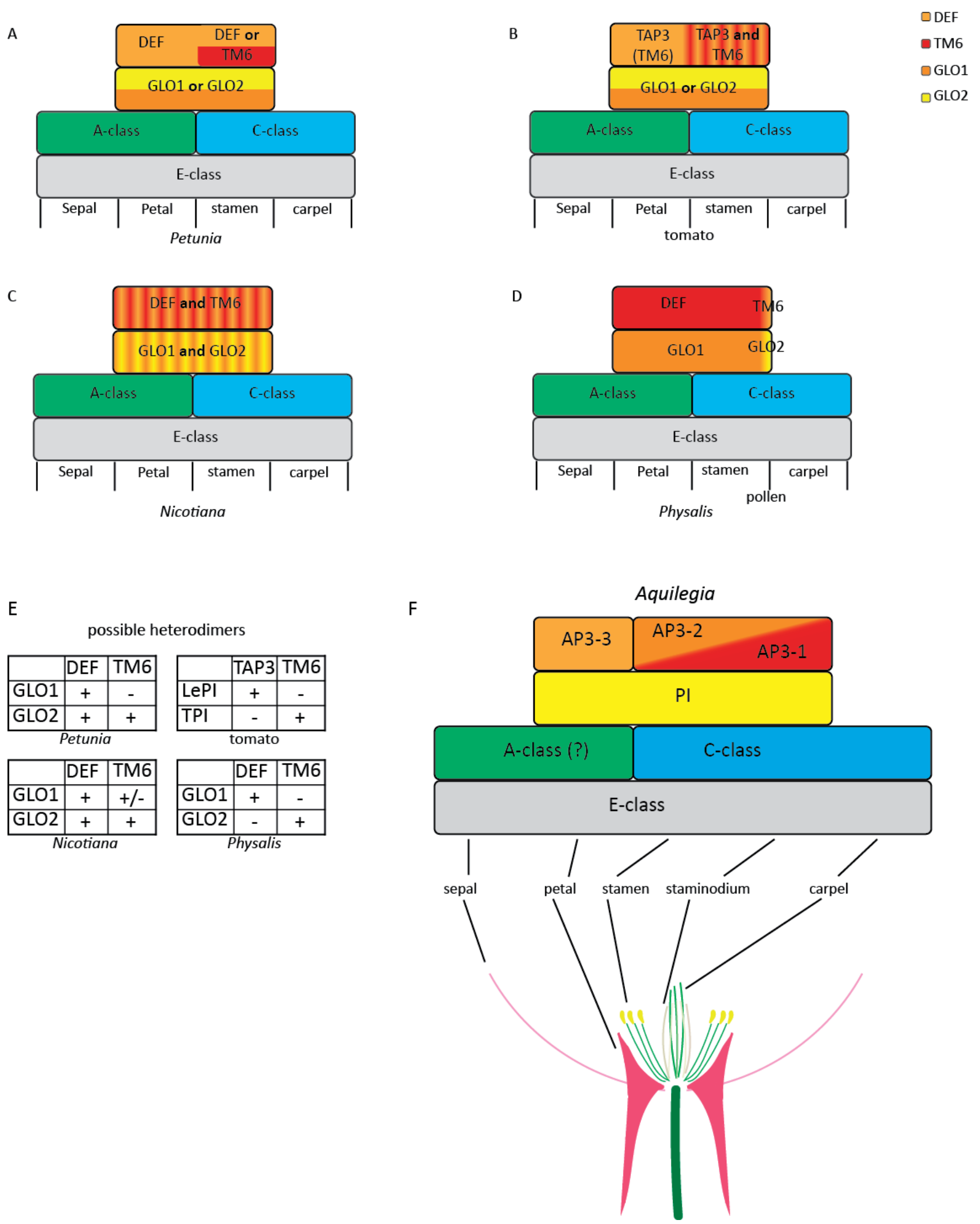
Although all Solanaceae contain two paralogs in both the *AP3* and *PI* clades, they did not subfunctionalize in the same way in all species. In tomato, *TM6* is not only expressed in stamens, like *PhTM6* in petunia, but also in petals. However, *TM6* and *euAP3* genes are not redundant in tomato: the single *tap3* mutant shows homeotic conversions of petals and stamens, whereas the *tm6* mutant shows homeotic conversion of stamens only (de Martino et al., 2006). Similarly to petunia, the *GLO* paralogs are largely redundant (**Figure 3B**) (Geuten and Irish, 2010).

Contrary to tomato and petunia, in *Nicotiana benthamiana* the *GLO* genes are not redundant, as both *glo1* and *glo2* single mutants show homeotic transformations. However, the *glo1glo2* double mutant does have a stronger phenotype (Geuten and Irish, 2010). The *Nicotiana benthamiana DEF/TM6* genes behave similarly to the tomato homologs, with both genes affecting petals as well as stamens, and the double mutant having a stronger phenotype (**Figure 3C**) (Liu et al., 2004; Geuten and Irish, 2010).

These differences in subfunctionalization are relatively subtle compared to the situation in another Solanaceae species, *Physalis floridiana*. In *Physalis*, both corolla and stamen identity are specified by *DEF* and *GLO1* only. The other paralogs, *TM6* and *GLO2* seem to only have a function in pollen maturation (**Figure 3D**) (Zhang et al., 2014; Zhang et al., 2015). Subfunctionalization also occurred at the level of protein interactions. In petunia and *Nicotiana*, TM6 interacts more strongly (*Nicotiana*) or exclusively (petunia) with GLO2 in yeast assays (Vandenbussche et al., 2004; Geuten and Irish, 2010). In tomato and *Physalis*, the situation is more extreme, with each paralog only having one possible dimerization partner (**Figure 3E**) (Leseberg et al., 2008; Zhang et al., 2014; Zhang et al., 2015).

Aquilegia is a genus of plants in the basal eudicot family of Ranunculaceae, which has an interesting floral morphology. Starting at the outside, *Aquilegia* has five petaloid sepals, five spurred petals, then several whorls of stamens, followed by two whorls of a fifth type of organ, staminodia, and centrally a whorl of carpels. Staminodia are sterile organs, and in *Aquilegia* they are typically colourless organs, consisting of lamina linked to a central filament. Duplications early in the evolution of the Ranunculaceae led to three different lineages of *AP3*, designated *AP3-1*, *AP3-2* and *AP3-3* (Kramer et al., 1998; Kramer et al., 2003). It is speculated that these duplications in the *AP3* lineage might be linked to the origin of the staminodia.

The three *AP3* paralogs have acquired distinct expression patterns during the evolution of the *Aquilegia* flower (Kramer et al., 2007). While *AP3-3* is petal specific, both *AP3-1* and *AP3-2* are expressed in stamen and staminodia primordia early in development.



	DEF	TM6
GLO1	+	-
GLO2	+	+

Petunia

	TAP3	TM6
LePI	+	-
TPI	-	+

tomato

	DEF	TM6
GLO1	+	+/-
GLO2	+	+

Nicotiana

	DEF	TM6
GLO1	+	-
GLO2	-	+

Physalis

Figure 3: Examples of subfunctionalization. *In **A-D**: the subfunctionalization of the B-class genes in several Solanaceae species. In **E**: the subfunctionalization at the protein-protein-interaction level. In **F**: neo- and subfunctionalization in Aquilegia.*

At later stages, however, *AP3-1* becomes specifically expressed in the staminodia, while *AP3-2* is expressed in stamens (Kramer et al., 2007). Gene knock-down experiments of *AP3-1* showed an aberrant staminodia phenotype, whereas in the *AP3-2* knock-down plants the stamens were affected (Sharma and Kramer, 2013). Silencing of both *AP3-1* and *AP3-2* simultaneously however, showed a more severe phenotype with both staminodia and stamens being converted into carpel-like structures. These data indicate that *Aquilegia AP3* genes have subfunctionalized as well as undergone neofunctionalization. The three paralogs subfunctionalized in their expression pattern. However, neofunctionalization has likely also taken place, as *AP3-1* and *AP3-2* specify different organs, with the *AP3-1* specifying a novel organ, the staminodia (**Figure 3F**). It is interesting to note that the petal specificity seems to be conserved among most Ranunculaceae, and that there is a correlation between loss of *AP3-3* expression and loss of petals in this family (Zhang et al., 2013).

Floral evolution from a genomics perspective

The evolution of floral MADS-domain proteins has been crucial during flower evolution. However, the only function these TFs have is regulating the expression of other genes. Consequently, the only way they can modify morphology, is through their target genes. This means that mutations in MADS-domain TFs only have an effect on morphology if they lead to differences in DNA binding site recognition or target gene regulation. TFs can change the sites they bind to through changes in their DNA binding specificity and/or affinity. Changes in interactions with other proteins will change the complexes in which the TF operates, and therefore might also modify the set of target genes, or the way these genes are regulated.

Even when the DNA binding specificity and/or affinity of a TF does not change, the set of genes regulated by that TF can evolve through changes in these target genes. Mutations in *cis*-regulatory elements (CREs) might modify binding sites of TFs, thereby either generating new targets or abolishing existing targets (reviewed in chapter 2).

Binding to the DNA does not only depend on the presence of CREs, but also on the chromatin context. The DNA itself can be modified, most famously by methylation. In addition, DNA is wrapped around histones to form chromatin, and this chromatin can be modified. Several varieties of histones exist (Deal and Henikoff, 2011a), and the histones can be modified in several ways, by e.g. methylation or acetylation. These modifications play a role in transcriptional regulation. Some modifications are associated with closed, non-transcribed regions, whereas other modifications are linked to transcriptionally active regions (Pfluger and Wagner, 2007; Ha, 2013; Iwasaki and Paszkowski, 2014). These modifications can change the accessibility of the DNA, and thereby influence TF binding, without modifying the underlying DNA sequence. Some of these chromatin modifications can function as epigenetic memory, and are stably transmitted to daughter cells. This memory is mainly maintained by the action of the Polycomb groub (PcG) and the trithorax group (Trxg) complexes (Bratzel and Turck, 2015; Iglesias and Cerdán, 2016). This kind of epigenetic memory has been shown to play a part in several processes in plants (Iwasaki and Paszkowski, 2014). One well-studied process in which epigenetic memory plays a role is vernalization. In *Arabidopsis*, the MADS-domain TF *FLOWERING LOCUS C (FLC)* is an important repressor of flowering time. *FLC* is expressed in the vegetative phase and thereby represses flowering. However, in a period of prolonged cold, the *FLC* locus gradually acquires trimethylation at lysines 9 and 27 of histone H3 (H3K9 and H3K27) (Bastow et al., 2004), which represses *FLC* expression, and thereby enables flowering.

To understand and predict floral morphology, we should focus on several aspects. We need to know how mutations in TFs can change their function. TFs only function through regulation of their downstream target genes, which means we also need to know which genes are regulated by a TF, and how this set of target genes changes during evolution. In order to predict TF binding sites, we will need a better understanding of how transcriptional regulation works, and which factors are involved. Addressing all these different aspects should help us to understand the evolution of floral morphologies.

Aim and outline of this thesis

The main question I was aiming to answer in this study is how floral organ developmental programs have evolved to generate morphological differences between species. I addressed this question from different angles with an emphasis on molecular and evolutionary aspects of the transcription factors involved.

Chapter 2 introduces this topic and we argue that evolution of *cis*-regulatory elements (CREs) is a driving force of morphological evolution. Changes in CREs can change regulatory networks, which could lead to morphological changes. We discuss how evolution of CREs and regulatory networks can be studied on a genome-wide scale, using emerging sequencing-techniques. We also hypothesize that in case of the flower, these changes will occur downstream of the master regulators of the (A)BCE-model. In **chapter 3,** we tested this hypothesis by comparing binding sites of the major floral regulator SEP3 between the two closely related species *Arabidopsis thaliana* and *Arabidopsis lyrata*. These species diverged about 10 million years ago, and show similar flower morphology, with the major difference being the flower size. Surprisingly, we discovered that only a relatively small proportion of the SEP3 binding sites are conserved between these two species. We found that some of the differences are due to sequence divergence, but many binding sites are species-specific, although the underlying sequence is preserved. We did note however, that binding sites linked to genes involved in floral development showed higher conservation than binding sites linked to genes with different functions. In **Chapter 4** we continue on this topic, but now focus on the chromatin environment, and how much influence this has on evolutionary changes. We compared regions of open chromatin between two ecotypes of *A. thaliana*, as well as with *A. lyrata*.

Genome duplication events are a common phenomenon in plant evolution and also B-class homeotic genes have been duplicated in many species. How this affects their function and molecular properties of the paralogous transcription factors is described in **Chapter 5**. We focused on *Tarenaya hassleriana,* belonging to a sister family of the Brassicaceae, which has two copies of both B-class genes. Whereas the *AP3* duplication is very recent, the *PI* duplication is more ancient. As both paralogs are still expressed, we were interested if there is any sub- or neofunctionalization of the two paralogs. Another example of a species with multiple *AP3* copies is *Aquilegia*, which evolved a new type of floral organ, the staminodia. We studied different molecular properties of these paralogs and tried to identify differences among the copies (**Chapter 6)**. To conclude, we discuss the results obtained in this thesis and their implications for the field of evolutionary developmental biology in **Chapter 7**.

Acknowledgments

Some of the pictures in this introduction are taken by others, and fall under a Creative Common license. In figure 1, pictures are taken by Adampauli (Tropaeolum majus), Georges Seguin (Rosa), David Marquina Reyes (CC BY-NC-ND-2.0) (Antirrhinum majus), JJ Harrison (Aquilegia), David Albeck (Narcissus), Christian Fischer (Ceratophyllum), Qwert1234 (Chloranthus japonicus) and Scott Zona (Amborella). In figure 2, the photograph of the tulip is made in the Chanticleer garden by Derek Ramsey.

CHAPTER 2

Plant 'Evo-devo' goes genomic: from candidate genes to regulatory networks

Suzanne de Bruijn
Gerco C. Angenent
Kerstin Kaufmann

Trends in Plant Sciences, 2012, issue 8, pages 441-447

Abstract

Plant development gives rise to a staggering complexity of morphological structures with different shapes, colors and functions. Understanding the evolution of control mechanisms that underlie developmental processes provides insights into causes of morphological diversity and is therefore of great interest for biologists. New genomic resources and techniques allow for the first time to assess the evolution of developmental regulatory networks at a global scale. Here we address the question how comparative regulatory genomics can be used to reveal the evolutionary dynamics of control networks linked to morphological evolution in plants.

Current approaches in plant 'evo-devo' research

The fascinating complexity of plant morphologies has inspired generations of scientists, from Johann Wolfgang von Goethe and Charles Darwin to contemporary biologists, and understanding the molecular basis of morphological evolution is one of the core questions (Vergara-Silva, 2003). The morphology of multicellular organisms is determined during development, which is heavily controlled by transcription factors (TFs), making these important targets for selection in evolution. In plants, several TFs have been linked to the divergence of morphological traits. Most classical research in evolutionary developmental biology ('evo-devo') has focused on the comparative analysis of 'candidate genes', linking the functional evolution of these genes to morphological diversification. For example, members of the MADS and TCP TF families (see Glossary) have been associated with the evolution of floral organs and floral symmetry, respectively (Theissen et al., 2000; Rosin and Kramer, 2009). Classic evo-devo approaches include mutant analysis, heterologous mutant complementation, comparative gene expression studies, and phylogenetic reconstruction. These approaches have limitations because morphological changes are often likely to be linked to mutations in more than one gene, even when comparing closely related species. On a different level, concerted changes in several traits are often required to create novel organ morphologies with selective advantage. For example, pollinator shifts require (correlated) changes in the color, shape and size of floral organs (Cronk and Ojeda, 2008; Wu et al., 2008).

In contrast to candidate gene approaches, genetic mapping and analysis of quantitative trait loci (QTLs) provide more comprehensive insights into the basis of heritable phenotypic variation at low taxonomic levels (Mackay et al., 2009). A nice example is a study on the selfing syndrome in the genus *Capsella* (Sicard et al., 2011). Using a cross between the outcrossing species *Capsella grandiflora* and the inbreeding species *Capsella rubella*, several QTLs affecting different aspects of flower morphology were identified, showing a complex genetic basis for evolutionary divergence of these traits in the genus *Capsella*. That variation in phenotypic traits can be due to several loci, each with moderate or small effect, has been shown in additional recent studies (reviewed in (Mackay et al., 2009)). However, changes in individual loci can trigger selection shifts and thereby result in divergence of evolutionary trajectories; for example pollinator preference in the genus *Mimulus* was shown to be driven by a major genomic locus (Schemske and Bradshaw, 1999). Although the QTL approach provides a hypothesis of the genetic architecture that underlies a phenotypic trait, identifying the exact molecular basis (specific mutation or gene(s)) of a QTL requires fine mapping and subsequent functional analysis. Conversely, a QTL may contain many genes that contribute to the phenotype of interest. More recently, QTL approaches making use of genome-wide expression data or other omics-type data, also referred to as genetical genomics, have been established (Joosen et al., 2009). However, mapping-based methods remain restricted to studying variation at low taxonomic levels, and have analytical limitations (e.g. genetic marker density). To study the evolution of developmental regulatory networks, there is the need to develop alternative genome-wide approaches that take into account the possibly complex basis of variation in morphological traits.

The number of genomic sequences available is increasing, opening new avenues for evo-devo research. Comparative genomics contributes to understanding of patterns of gene duplication and gene content in seed plants (Jiao et al., 2011; Lee et al., 2011) and has also been used to study sequence conservation among *Arabidopsis* populations and related species (Cao et al., 2011; Gan et al., 2011; Hu et al., 2011). Whole-genome sequencing can be used to improve organism and gene phylogenies that are crucial for establishing hypotheses on character evolution in 'evo-devo' research. Understanding the evolution of gene regulation in development requires information beyond genome sequences, such as experimental data on transcriptional regulation and gene expression variation (Yant, 2012). Therefore, 'top-down' comparative regulatory genomics approaches should be used to identify candidate genes potentially linked to morphological diversification between species.

Glossary

Cis-regulatory elements (CREs): Collections of transcription factor binding sites and other non-coding DNA that are sufficient to facilitate transcription in a defined spatial and/or temporal expression domain.

ChIP-seq: Chromatin-immunoprecipitation followed by sequencing. A technique to determine *in vivo* DNA-bound regions of a protein at genome-wide scale.

DNAseI-seq: treatment of isolated chromatin with DNAseI. DNAseI cuts accessible DNA, and the released fragments are then sequenced. This gives an indication of the chromatin state and, if sequenced deep enough, can reveal protein-binding sites.

Heterochronic: A change in the timing of expression.

Heterotopic: A change in the place of expression.

MADS-box TFs: A family of TFs, present in all groups of eukaryotes. Named after its founding members MCM1 (*Saccharomyces cerevisiae*), AGAMOUS (*Arabidopsis thaliana*), DEFICIENS (*Antirrhinum majus*) and SRF1 (*Homo sapiens*).

Pleiotropic: Influencing more than one trait due to multiple functions of a gene (e.g., a gene that is involved in the growth of both a leaf and a petal).

RNA-seq: Digital quantification of transcriptomes (mRNA) by next-generation sequencing.

SELEX (Systematic evolution of ligands by exponential amplification): procedures for the identification of representative sets of ligands for a protein. In the case of DNA-binding proteins, the protein is mixed with a pool of double-stranded, randomized oligonucleotides. Protein-DNA complexes are recovered and the bound DNA is amplified by PCR, and subjected to a new round of selection. DNA fragments are sequenced to reveal the binding specificity of the protein.

TCP TFs: A family of plant transcription factors. The family is named after Tb1 (*Zea mays* L.), CYCLOIDEA (*Antirrhinum majus*) and Pcf1 (*Oryza sativus*).

Morphological evolution and divergence through *cis*-regulatory elements

Gene expression divergence, caused by mutations in *cis*-regulatory elements (CREs), is an important driving force of morphological evolution (Carroll, 2008; Wittkopp and Kalay, 2012). Heterochronic or heterotopic changes of gene expression can lead to recruitment of genes or gene-regulatory modules to function in a new context or location, ultimately resulting in changes in plant morphology. For example, the evolution of inflorescence architectures in *Solanum* species has been linked to heterochronic shifts in developmental gene expression (Park et al., 2012). Heterotopic shifts of gene expression of floral organ identity genes have for instance been implicated in the evolution of floral organ morphologies (Kanno et al., 2003; He and Saedler, 2005).

CREs usually contain several TF binding sites (TFBSs), which are relatively short and can be degenerate, and so may evolve more easily than the more constrained protein-coding regions. That extensive variation in sequence and position of TFBSs can form the basis of evolutionarily conserved regulatory interactions was elegantly demonstrated recently for the regulation of the floral homeotic *AGAMOUS* gene by LEAFY in different flowering plant species (Moyroud et al., 2011). This example also shows that more experimental data are needed to understand general evolutionary TFBS turnover, TFBS flexibility linked to conserved gene regulation (Weirauch and Hughes, 2010), and how changes in gene expression are achieved at the molecular level by CRE mutations.

Although within species most mutations that cause morphological variation have been found in protein-coding regions, morphological differences between species are often caused by mutations in noncoding regions, indicating that natural selection over longer time periods leads to fixation of mutations with more subtle and specific effects (Carroll, 2008; Stern and Orgogozo, 2008, 2009; Jones et al., 2012). Several examples show the importance of mutations in CREs in plant domestication (Doebley et al., 2006; Konishi et al., 2006; Chen et al., 2007; Studer et al., 2011) and in natural evolution (Bharathan et al., 2002; Hay and Tsiantis, 2006; Uchida et al., 2007). A striking example is the evolution of tissues involved in fruit shattering by modification of the expression of orthologs of the *Arabidopsis* (*Arabidopsis thaliana*) homeobox TF *REPLUMLESS* (*RPL*). It was found that the same point mutation in a conserved CRE was selected during rice domestication (Konishi et al., 2006) and during evolution of the Brassicaceae family (Arnaud et al., 2011) (**Figure 1**). Cases of independent fixation of the same regulatory mutations have also been reported in the animal field, suggesting that some genes and genomic positions are more prone to mutation (with phenotypic consequences) than are others, probably due to a selective advantage (Stern and Orgogozo, 2009; Chan et al., 2010). A classic example for the importance of gene expression changes during crop domestication is the *teosinte branched1* (*tb1*) locus in maize (*Zea mays*), where a higher expression level of a TCP transcription factor gene caused a dramatic increase in apical dominance, due to a transposon insertion in its promoter. This transposon insertion predates maize domestication, indicating that domestication acted on existing variation rather than on new mutations (Studer et al., 2011). Taken together, these findings demonstrate the role of gene expression variation in natural evolution and domestication. They also point towards different underlying mutations, ranging from single nucleotide polymorphisms (SNPs) to insertion of transposable elements.

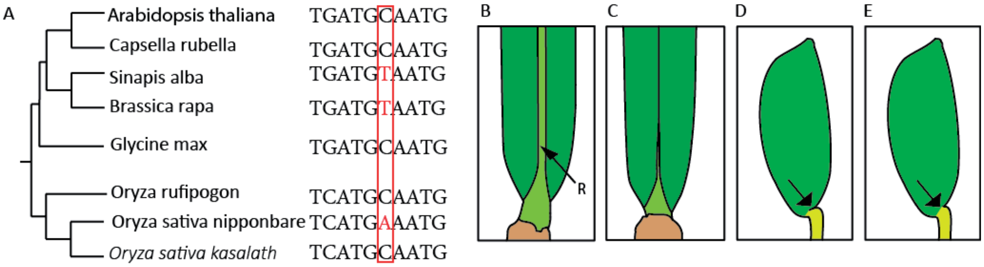

A
Arabidopsis thaliana TGATGCAATG
Capsella rubella TGATGCAATG
Sinapis alba TGATGTAATG
Brassica rapa TGATGTAATG
Glycine max TGATGCAATG
Oryza rufipogon TCATGCAATG
Oryza sativa nipponbare TCATGAAATG
Oryza sativa kasalath TCATGCAATG
B
C
D
E
R

Figure 1: A recurrent point-mutation in a cis-regulatory element controlling the expression of REPLUMLESS (RPL) orthologs in Brassicaceae and rice has been associated with the changes in morphological structures implicated in seed dispersal. A: *Phylogenetic tree and alignment of cis-regulatory sequence in the RPL promoter. A point mutation in the same position (red frame) is found in species/cultivars with reduced development of seed-dispersal structures (B-E) fruit phenotypes:* ***B:*** *Arabidopsis thaliana,* ***C:*** *Brassica rapa,* ***D:*** *Oryza sativa kasalath,* ***E:*** *Oryza sativa nipponbare. Arrow indicates dehiscence zone. Abbreviation: R, replum.*

How does regulation of gene expression evolve?

To understand the evolution of gene expression, one needs to elucidate how CREs originate and diversify during evolution. Potentially, CREs can evolve *de novo* by accumulation of mutations. ‘Co-option’ of existing CREs, which acquire new TF binding sites, can generate new expression domains or result in repression in certain locations or conditions. Changes in gene expression can also result from loss of binding sites by point mutations or indels (Carroll, 2008; Rebeiz et al., 2011; Wittkopp and Kalay, 2012). The gain of new expression features is rare relative to changes in the timing or level of gene expression, the expansion or restriction of spatial expression domains, or the loss of expression features (Prud'homme et al., 2007). In line with this, studies in animals suggest that changes in gene regulation evolve mostly by modification of existing CREs, and genomic sequences are primed to evolve new expression patterns by the presence of latent enhancers (Rebeiz et al., 2011; Wittkopp and Kalay, 2012).

Interestingly, not all TF binding events seem to affect gene expression (Li et al., 2008; Schmidt et al., 2010). It is possible that apparently non-functional binding sites regulate gene expression only in certain genetic backgrounds, populations or environmental conditions. The presence of these sites might also reflect the evolutionary dynamics of gene regulation: Binding sites may represent ‘raw material’ for evolutionary ‘tinkering’ or remnants of past functions. However, although the evolutionary role of *cis*-regulatory mutations is well supported by many examples, little is known about the upstream regulators that bind to evolutionarily diversified DNA sequence elements, revealing an important gap in our understanding of developmental regulatory network evolution in plants.

New approaches to studying regulatory network evolution in plants

Physical interactions between TFs and specific DNA sites form an important molecular basis of developmental regulatory networks. Therefore, understanding the evolutionary dynamics of these physical interactions can reveal the molecular 'rewiring' of networks linked to changes in development and plant morphology.

Genetic networks are robust, due to multiple feedback, feedforward and cross-regulatory mechanisms. This makes it probable that mutations in multiple genes are needed to 'rewire' these networks, and thereby change their outcome at the morphological level. A study of the gene-regulatory network controlling root stele development found that morphological phenotypes were associated with mutations in only 16% of the TFs tested, whereas molecular or expression phenotypes were identified for 65% (Brady et al., 2011). Accordingly, the transcriptional network can be affected in a TF mutant despite the absence of a mutant phenotype. Therefore, mutations in compensatory by-passes or changes in the expression of several genes may also be needed to allow the creation of a new steady state of the network and a robust change in morphology.

The availability of genome sequence information for an increasing number of plant species opens new ways to study the evolution of developmental regulatory networks (**Table 1**) (Yant, 2012). Sequencing and alignment of genomes from different ecotypes or species, alone or in combination with RNA-seq, allows to predict CREs underlying regulatory differences (Gan et al., 2011; Jones et al., 2012), but this will usually not identify the nature of the upstream TFs binding to these elements. Bioinformatic predictions of TFs that bind to certain CREs will benefit from more complete experimental datasets on TF binding in model species. Experimental methods are now available that allow a more direct comparison of TFBSs among species: potential TFBSs can be predicted using DNA-binding models determined by *in vitro* methods such as systematic evolution of ligands by exponential amplification (SELEX) (Moyroud et al., 2011). This method is highly versatile, because it can also be used to study the DNA-binding specificity of TFs in plant species that are not experimentally amenable. A limitation is that not all predicted binding sites in the genome may be accessible to the TF *in vivo*, due to a closed chromatin structure. Alternatively, additional factors may be required to modulate DNA-binding specificity of a TF in the plant cell. Chromatin structure can be assessed with DNAse I-seq, which reveals 'active' *cis*-regulatory regions in the genome with an open chromatin structure, and the method can be used to map differences in CRE activity between individuals, populations or species (Hesselberth et al., 2009; Degner et al., 2012). However, not all functional TF binding sites may be associated with DNAse I hypersensitive sites (Li et al., 2011). To determine TF binding sites at a genome-wide scale *in vivo*, ChIP-seq experiments can be performed. There are a few studies where ChIP-seq was used to compare binding sites of a TF between different animal or yeast species (Tuch et al., 2008; Schmidt et al., 2010; He et al., 2011). The goal of these studies was to study TFBS evolution, not to couple TFBS changes to variation in developmental programmes. ChIP-seq and SELEX-based methods will help in the study of the evolutionary variation of TFBSs at genomic scale in plants. However, given that not all changes in TFBSs will affect gene regulation, these studies should be combined with comparative gene expression analysis, for example by comparative RNA-seq (Brawand et al., 2011). To study evolution of developmental pathways, data from different experimental approaches, such as comparative ChIP-seq and RNA-seq, should be integrated. The relevance of individual CREs and genes that are identified by the genome-wide approaches can be further assessed by detailed gene expression studies/CRE mutagenesis, mutant analysis or virus-induced gene silencing (VIGS) (Becker and Lange, 2010) and heterologous reporter gene and/or complementation assays. This enables changes in (active) TFBSs to be linked to changes in morphology in a top-down approach. Therefore, identifying non-conserved target genes of developmental TFs will provide insights into the mechanisms underlying morphological evolution. In addition to identifying non-conserved regulatory interactions, the combination of methods will also enable the identification of the evolutionarily stable regulatory 'core circuitry'.

ChIP-seq experiments can be performed in every species with a sequenced genome. Amenability for transformation is not a pre-requisite, because antibodies can be raised against the native proteins. However, if a TF functions in specific cell-types, tissues or developmental stages, enrichment strategies may need to be used to generate enough plant material for ChIP-seq or RNA-seq experiments. Most of these methods, such as INTACT (isolation of nuclei tagged in specific cell types) (Deal and Henikoff, 2010) and fluorescence-activated cell sorting (FACS) (Birnbaum et al., 2003), will require the generation of transgenic plants, but laser dissection microscopy (for RNA-seq) can be used in a non-transgenic setting (Torti et al., 2012). Tissue-sampling is also a problem when homologous organs or tissues that are being studied are morphologically very different. Therefore, careful consideration of developmental stages and tissues is crucial to be able to identify primary differences in developmental gene regulation, which ultimately result in morphological diversification. Besides changes in transcriptional regulation, modifications in posttranscriptional regulation, for example linked to changes in micro-RNA target genes, can contribute to developmental regulatory network evolution. Also at this regulatory level, genome-wide approaches can be used for a comparison between species (Pasquinelli, 2012).

Which levels in the gene-regulatory hierarchy are most informative?

Developmental regulatory networks have a complex 'hierarchical' structure (**Figure 2**), with multiple feedback and feedforward loops that enable stable developmental decisions (Davidson and Erwin, 2006; Kaufmann et al., 2010a). Not all levels in the regulatory hierarchy are equally likely to contribute to morphological diversification, and they evolve at different speeds (Davidson and Erwin, 2006). Mutations that affect general 'upstream' regulators (usually highly connected nodes in a network) of developmental or cellular gene expression are more likely to have pleiotropic effects, therefore these mutations tend to reduce fitness (Stern and Orgogozo, 2009). By contrast, genes that execute cellular responses downstream of the so called input–output genes (also called 'intermediate regulators') (**Figure 2**) often act together with other genes in a concerted fashion in basic cellular functions. The expression of

Table 1: Genome-wide methods for the characterization of developmental regulatory networks.

Method	Aim/applications	Limitations[1]	Refs
ChIP-seq	Characterization of TF binding sites and other DNA-binding proteins in vivo	Requires antibody against native protein (or a transformable species)	(Kaufmann et al., 2010b)
		Technically challenging	
		Requires substantial amounts of plant material	
		Association of binding events with transcriptional response requires additional information (-> combine with RNA-seq)	
RNA-seq	Expression level analysis by sequencing; Characterization of organ- and/or tissue-specific transcriptomes and gene expression levels	No direct information about underlying regulatory mutations resulting in changes in gene expression between species	(Wang et al., 2009)
		Reflects the combined effect of transcriptional and posttranscriptional regulation	
	Does not require a full genome sequence	For cell specific transcriptome analysis, additional methods such as INTACT or laser microdissection may be needed	
SELEX	In vitro assay for the characterization of DNA-binding specificity of TFs and CRE prediction using genome sequence information	Generation of realistic DNA-binding models can be challenging	(Moyroud et al., 2011)
		Some of the predicted DNA binding sites may be inaccessible in vivo due to chromatin structure (-> combine with DNAse I seq)	(Wang et al., 2011)
		Association of binding events with transcriptional response requires additional information (-> combine with RNA-seq)	
DNAse I-seq	Characterization of chromatin accessibility *in vivo* using high throughput sequencing	No direct indication on relevance for transcriptional regulation	(Degner et al., 2012)
		No information on the nature of the TFs that bind in an accessible region (-> combine with SELEX or ChIP-seq)	(Pique-Regi et al., 2011)
	Footprinting of protein-DNA interactions	Not all TF binding sites may be in ‘accessible’ chromatin	
		Method not yet frequently used in plants	

[1] Limitations of individual methods can be overcome by combining different methods

several of those genes needs to be modulated in a coordinated manner. These response modules are under combined control of input–output genes, which usually encode TFs (Stern and Orgogozo, 2008). By changing the expression of input-output genes, organ morphologies can be modulated in a specific context.

It has been proposed that input–output genes represent hotspots for evolution (Stern and Orgogozo, 2009). In line with this, some types of TF were repeatedly (although not exclusively) recruited to modify organ morphologies in a certain manner in plants. One example is the heterotopic expression of KNOX TFs resulting in dissected leaf development (Bharathan et al., 2002; Hay and Tsiantis, 2006, 2010). Another example is the recurrent recruitment of CYC-type TCP TFs in generating monosymmetric flowers across distant eudicot lineages (Busch and Zachgo, 2009), which might be linked to an ancestral dorsal expression domain in floral meristems that was selectively expanded and/or switched to later stages of organ development in monosymmetric taxa (Preston and Hileman, 2009; Busch et al., 2012).

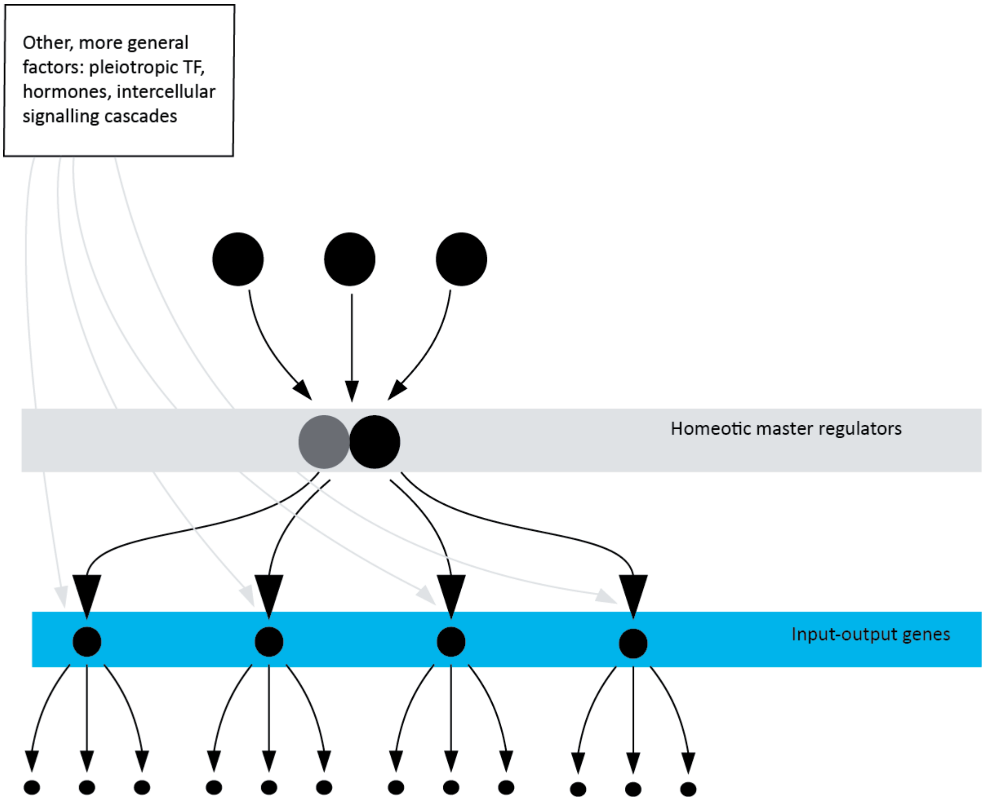

Other, more general factors: pleiotropic TF, hormones, intercellular signalling cascades
Homeotic master regulators
Input-output genes

Figure 2: The hierarchical structure of gene regulatory networks, illustrated for floral organ development. *Floral organ identities are specified by floral homeotic master regulatory TFs, which directly modulate the expression of hundreds of genes in order to generate floral organ morphologies. TF genes are overrepresented among the direct targets. According to this model, many of these TF genes represent intermediary 'input–output' genes, which in turn control the formation of specific organ morphologies. Given the high number of direct target genes of the homeotic master regulators (MADS-box TFs) in floral organ development, they may also participate directly in the regulation of genes at the lowest level of the regulatory hierarchy (not shown here for simplification).*

Because input–output genes are targets of developmental master regulatory TFs, understanding the evolutionary dynamics of the direct target gene repertoire of the 'masters' will be particularly informative for understanding the evolution of gene-regulatory networks underlying morphological diversification. At the subsequent level in the hierarchy, the divergence in target genes of input–output genes will be interesting to assess for a full understanding of regulatory network divergence (see Busch and Zachgo (2007)) for an example).

Flower development is one of the best characterized developmental processes in plants. Although floral organs can have different morphologies, the basic organ 'types' and the molecular mechanisms controlling specification of floral organ identities are largely conserved (Theissen and Melzer, 2007). According to the ABCE model, floral organ identities are specified by the combinatorial action of A-, B-, C- and E- class TFs. These master-regulatory TFs have thousands of DNA-binding sites in the genome, and directly regulate the expression of a variety of regulatory genes that are important for growth, shape and structure of different organs (Kaufmann et al., 2009; Kaufmann et al., 2010c). Given the tight link between organ identity and organ morphology, we propose that the wide variety of flower morphologies seen in nature might be linked to changes in direct targets of the ABCE TFs during flower evolution. The fact that a floral homeotic mutant in a species can be complemented with an ABCE factor from another species makes it indeed possible that evolutionary diversification of floral organ morphologies occurs ‘downstream’ of the master regulators. This hypothesis can be tested by future comparative regulatory genomics approaches.

Macroevolution of regulatory networks and the origin of novel plant morphologies

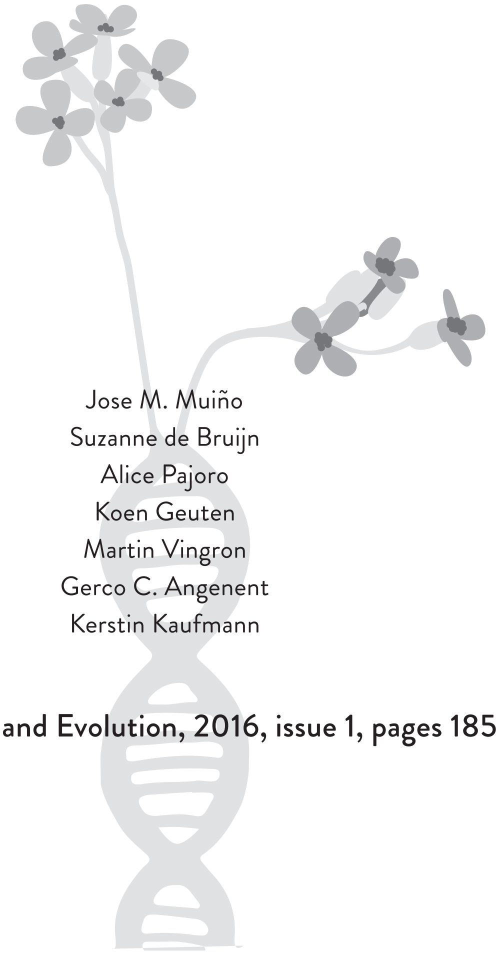
Not every modification in a developmental regulatory network is equally likely to contribute to ‘macro-evolutionary’ creation of body plans or novel organ types with distinct morphologies. In addition to *cis*-regulatory mutations, the evolution and diversification of TF families with key roles in plant development suggest that specific changes in protein functions also contributed to the elaboration and diversification of plant development. TF gene duplications followed by protein diversification increase the functional ‘repertoire’ of an organism. That mutations in the DNA-binding domain play a role in TF evolution across long evolutionary time periods has been exemplified in studies on LEAFY protein function in land plants (Maizel et al., 2005). Understanding how TFs changed their functional specificity during evolution, and how these changes contributed to the recruitment of regulatory modules and networks to novel functions over longer evolutionary time-scales, is another major open question, which also should be addressed in the future by comparative regulatory genomics approaches.

To address the question of how apparently novel organ types originated during evolution, gene expression can provide support for homology to other organs. For example, expression of meristem identity TFs suggests that tendrils in grapevine (*Vitis vinifera*) originated from reproductive meristems (Calonje et al., 2004). At genome-wide level, transcriptome profiling can be used to reveal more general evolutionary relationships in developmental programs of different organs or developmental stages; for example, testing the ‘hourglass’ model of animal embryo development (Domazet-Loso and Tautz, 2010; Kalinka et al., 2010; Brawand et al., 2011). In the plant field, gene expression profiles of floral organs in angiosperms and a non-flowering seed plant (a cycad) were used to test homology between different organ types, addressing the question of the origin of the angiosperm flower (Chanderbali et al., 2010). In another study, comparison of floral organ transcriptomes from the basal eudicot *Eschscholzia californica* and *A. thaliana* revealed (with the exception of MADS-box genes) extensive variation in the expression of genes with roles in flower development between the two species (Zahn et al., 2010). In the future, it will be interesting to discern the molecular changes in developmental regulatory networks that contributed to generate new morphological structures, such as flowers.

In summary, we strongly believe that comparative regulatory genomics approaches will greatly contribute to our understanding of the molecular complexity underlying morphological evolution in plants on short and long evolutionary time-scales.

CHAPTER 3

Evolution of DNA-binding sites of a floral master regulatory transcription factor



Jose M. Muiño
Suzanne de Bruijn
Alice Pajoro
Koen Geuten
Martin Vingron
Gerco C. Angenent
Kerstin Kaufmann

Molecular Biology and Evolution, 2016, issue 1, pages 185-200

Abstract

Flower development is controlled by the action of key regulatory transcription factors of the MADS-domain family. The function of these factors appears to be highly conserved among species based on mutant phenotypes. However, the conservation of their downstream processes is much less well understood, mostly because the evolutionary turnover and variation of their DNA binding sites (BS) among plant species have not yet been experimentally determined.

Here, we performed comparative ChIP (Chromatin-ImmunoPrecipation)-seq experiments of the MADS-domain transcription factor SEPALLATA3 (SEP3) in two closely related *Arabidopsis* species: *A. thaliana* and *A. lyrata* which have very similar floral organ morphology. We found that binding site conservation is associated with DNA sequence conservation, the presence of the CArG-box BS motif and on the relative position of the BS to its potential target gene. Differences in genome size and structure can explain that SEP3 BSs in *A. lyrata* can be located more distantly to their potential target genes than their counterparts in *A. thaliana*. In *A. lyrata*, we identified transposition as a mechanism to generate novel SEP3 binding locations in the genome. Comparative gene expression analysis shows that the loss/gain of BSs is associated with a change in gene expression. In summary, this study investigates the evolutionary dynamics of DNA BSs of a floral key-regulatory transcription factor, and explores factors affecting this phenomenon.

Introduction

Plant development is controlled by transcription factors (TFs), which form complex gene-regulatory networks (Kaufmann et al., 2010a). Genome-wide TF DNA-binding studies revealed that these factors have several thousands of binding sites in the *Arabidopsis* genome, and may regulate the expression of many genes directly, likely in combination with other TFs (for review, see Pajoro et al. (2014a)). Given the important role of developmental processes in environmental adaptation of plants, there is a need to understand the molecular basis of natural variation at the level of developmental gene regulation.

Until now, estimation of TF DNA binding sites (BSs) across plant species was done indirectly using DNA sequence conservation studies, since the only *in vivo* genome-wide profiles of TF DNA BSs were available for *A. thaliana*. Recent studies have focused on identifying conserved noncoding sequences (CNSs) among distantly related flowering plant species (Hupalo and Kern, 2013), within the Brassicaceae family (Haudry et al., 2013), among eudicots (Baxter et al., 2012; Van de Velde et al., 2014) and in more targeted species comparisons (see Haudry et al. (2013) for additional references). While the study by Haudry et al. (2013) resulted in the recovery of the highest number of TF binding sites based on genome-wide TF DNA-binding data in *A. thaliana*, Van de Velde et al. (2014) showed a higher specificity of BS recovery. However, the fraction of recovered BSs varies widely between different TFs. For example, approximately 34%, 15% and 8% of all BSs of the *Arabidopsis* MADS-domain TFs PISTILLATA, APETALA1 and APETALA3, respectively, were successfully predicted in the study of Van de Velde et al. (2014). Haudry et al. (2013) found that although most Brassicaceae genomes contained homologs for more than 75% of the *A. lyrata* CNSs identified by Haudry et al. (2013), the early branching *A. arabicum* genome had homologs for only 38%, and outside Brassicaceae, conservation of these CNSs was very low, ranging from 0.8% in *O. sativa* to 3.4% in *Carica papaya*, which suggest that their *A. lyrata* CNSs show a high turnover rate outside the Brassicaceae lineages. However, as noticed by the authors, an important fraction (75-fold enrichment) of these CNSs seems to represent small noncoding RNAs, not only TF DNA BSs.

Recent studies in mammals and insects have characterized the conservation of TF DNA BSs across different species using ChIP-seq approaches (see Villar et al. (2014) for a review). This offers a direct way to experimentally measure TF DNA BS turnover. Although the number of species and TFs studied are very limited at this moment, it appears that the turnover rate of BSs seems to be different depending on the group of species studied. Developmental TF BSs show higher conservation between *Drosophila* species compared with mammals when considering similar evolutionary distances (Villar et al., 2014). In *Drosophila* species, it seems that there is a stronger association between BSs conservation and regulatory function (Biggin, 2011; He et al., 2011) than in mammals (Schmidt et al., 2010; Stefflova et al., 2013).

Evolutionary mechanisms that drive regulatory diversification are poorly understood. Theoretical models show that BSs can arise on relatively short time-scales upon accumulation of base-pair substitutions (Stone and Wray, 2001). However, recent TF ChIP-seq comparative studies indicate that sequence changes in the TF binding motif only provide an explanation for a minority (12-40%) of TF BS variation (Villar et al., 2014). This proportion increases when sequence changes in BSs of interacting TFs within close distance of the motif are considered. For example, whereas 40% of mice strain-specific PU.1 binding can be linked to a sequence change in their DNA binding sequence, an additional 15% can be explained by mutations in proximal CEBPα or AP-1 binding motifs (Heinz et al., 2013). This suggests that the conservation of DNA-binding of a given TF is also affected by disruption of the binding motifs of other TFs belonging to the same complex.

Besides mutation, another mechanism to create new TF BSs is transposition. The contribution of transposition to BS variation seems to depend on the species studied. In mammals, there are clear examples of BSs that were copied/moved by transposons (e.g. (Johnson et al., 2006; Schmidt et al., 2012), while in *Drosophila,* an association between transposon activity and BS variation has not been detected yet (Ni et al., 2012). This can be related with the fact that mammalian genomes are rich in transposable elements (TEs) (de Koning et al., 2011), while *Drosophila* genomes have a much lower content of these elements (Lynch et al., 2011). In plants, E2F BS may have been amplified by transposon activity in Brassicaceae species (Hénaff et al., 2014).

Although computational prediction of TF BSs allows estimating the extent of regulatory divergence between species, the evolutionary turnover of TF BSs among plant species has not yet been experimentally determined on a genome-wide basis. This is important as many examples are known where changes in *cis*-regulation are causal for organismal diversity (reviewed in Rodríguez-Mega et al. (2015)). To understand the evolutionary dynamics of TF BS at a genome-wide scale, we therefore need *in vivo* experimental approaches to study TF BSs in different species.

In contrast to animals, plants underwent frequent polyploidization events, resulting in a high level of duplication in plant genomes. Duplications are normally followed by genomic re-arrangements, frequent gene loss and plant lineage-specific functional gene diversification (see, e.g. (Airoldi and Davies, 2012; Moghe and Shiu, 2014)). For example, the *A. thaliana* genome has gone through two rounds of whole genome duplication after divergence from *C. papaya* 70 million years ago (Proost et al., 2011). How polyploidization affects *cis*-regulatory evolution is still largely unexplored. For the reasons mentioned above, we performed the first comparison of BSs of a developmentally important TF at genome-wide scale between the two closely related plant species *A. thaliana* and *A. lyrata*.

A. lyrata is a member of the Brassicaceae family and a close relative of the model plant species *A. thaliana*. The two species diverged about 10 million years ago (Hu et al., 2011). The genome of *A. lyrata* has a size of around 200 Mb (close to the family average; N=8), and is therefore significantly larger (60%) than that of *A. thaliana* (~125 Mb; N=5) (Bennett et al., 2003; Hu et al., 2011) . The *A. thaliana* genome size reduction can be largely attributed to deletions in non-coding DNA and transposons, whereas the number of protein-coding genes is only 20% higher in *A. lyrata* than in *A. thaliana* (*A. lyrata*: 32,670; *A. thaliana*: 27,025). An overall sequence identity of 80% allows alignment of the two genomes, and orthologs can be readily identified due to the largely syntenic gene arrangements (Hu et al., 2011).

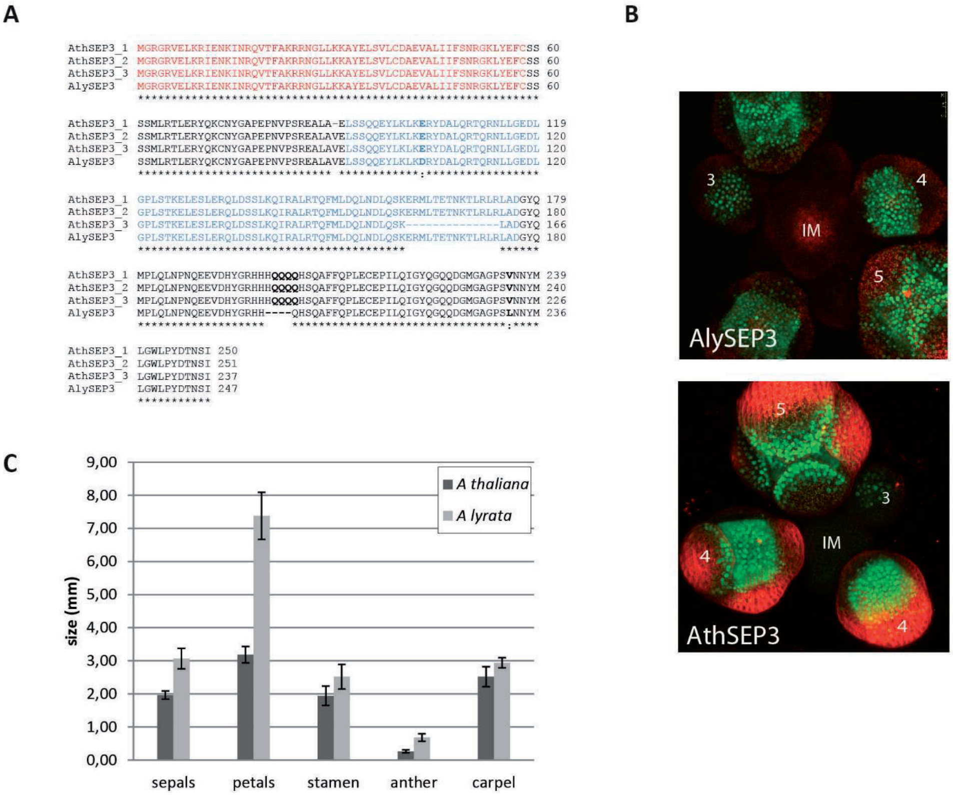
Although the overall morphology of flowers is similar between *A. thaliana* and *A. lyrata*, specific differences exist that are linked to the different mating strategies (*A. lyrata* – outcrossing, insect-pollinated*; A. thaliana* – selfing). Moreover, petals are larger in *A. lyrata* and produce benzenoids (Abel et al., 2009).

We were interested in how differences in genome size and floral organ morphologies between *A. thaliana* and *A. lyrata* are reflected in the evolution of gene regulation. Therefore, we chose to compare DNA-binding sites of the floral MADS-domain TF SEPALLATA3 (SEP3) at genome-wide scale in these two species using ChIP-seq experiments. SEP3 is a key mediator of higher-order protein complex formation of floral homeotic MADS-domain TFs, and therefore an important master regulator of flower development (Pelaz et al., 2000; Honma and Goto, 2001). We also quantified floral gene expression variation between the two species using comparative mRNA-seq. We analyzed the impact of speciation on the evolutionary conservation of SEP3 DNA-binding sites and potential direct target genes.

Results

Identification of SEPALLATA3 DNA-binding sites in two *Arabidopsis* species.

The protein sequences of the *A. thaliana* and *A. lyrata* SEP3 orthologs are identical in the DNA-binding part of the MADS-domain, and also show a high level of identity in other parts of the protein (**Figure 1A**). This allowed us to use a previously generated antibody against *A. thaliana* SEP3 (AthSEP3) for chromatin immunoprecipitation (ChIP) experiments (Kaufmann et al., 2009). The heterologous expression of an *A. lyrata SEP3* (Aly*SEP3*) promoter::gene fragment fused to GFP was highly similar to that of the Ath*SEP3* GFP reporter gene fusion, supporting the conservation of *SEP3* gene functions in the two species (**Figure 1B**). As DNA-BSs of SEP3 may vary between tissues and developmental stages (Pajoro et al., 2014b), we performed a staging of *A. lyrata* flower development using scanning electron microscopy, similar to a previous study on *A. thaliana* (Smyth et al., 1990). The results showed that meristem and early organ development in *A. lyrata* is similar to the development of *A. thaliana* as previously reported (Smyth et al., 1990) (**Figure S1**), allowing us to harvest tissue with similar composition for our ChIP experiments. We found that petal growth was enhanced after stage 11 of *A. lyrata* flower development, resulting in enlarged petals in *A. lyrata* compared to *A. thaliana* (**Figure 1C**). Also the relative growth of anthers and carpels differs to some extent, especially during later stages of flower development. Anthers in *A. lyrata* are larger compared to *A. thaliana*.



```
AthSEP3_1  MGRGRVELKRIENKINRQVTFAKRRNGLLKKAYELSVLCDAEVALIIFSNRGKLYEFCSS 60
AthSEP3_2  MGRGRVELKRIENKINRQVTFAKRRNGLLKKAYELSVLCDAEVALIIFSNRGKLYEFCSS 60
AthSEP3_3  MGRGRVELKRIENKINRQVTFAKRRNGLLKKAYELSVLCDAEVALIIFSNRGKLYEFCSS 60
AlySEP3    MGRGRVELKRIENKINRQVTFAKRRNGLLKKAYELSVLCDAEVALIIFSNRGKLYEFCSS 60
           ************************************************************

AthSEP3_1  SSMLRTLERYQKCNYGAPEPNVPSREALA-ELSSQQEYLKLKERYDALQRTQRNLLGEDL 119
AthSEP3_2  SSMLRTLERYQKCNYGAPEPNVPSREALAVELSSQQEYLKLKERYDALQRTQRNLLGEDL 120
AthSEP3_3  SSMLRTLERYQKCNYGAPEPNVPSREALAVELSSQQEYLKLKERYDALQRTQRNLLGEDL 120
AlySEP3    SSMLRTLERYQKCNYGAPEPNVPSREALAVELSSQQEYLKLKDRYDALQRTQRNLLGEDL 120
           ***************************** ************:*****************

AthSEP3_1  GPLSTKELESLERQLDSSLKQIRALRTQFMLDQLNDLQSKERMLTETNKTLRLRLADGYQ 179
AthSEP3_2  GPLSTKELESLERQLDSSLKQIRALRTQFMLDQLNDLQSKERMLTETNKTLRLRLADGYQ 180
AthSEP3_3  GPLSTKELESLERQLDSSLKQIRALRTQFMLDQLNDLQSK--------------LADGYQ 166
AlySEP3    GPLSTKELESLERQLDSSLKQIRALRTQFMLDQLNDLQSKERMLTETNKTLRLRLADGYQ 180
           ****************************************              ******

AthSEP3_1  MPLQLNPNQEEVDHYGRHHHQQQQHSQAFFQPLECEPILQIGYQGQQDGMAGPSVNNYM 239
AthSEP3_2  MPLQLNPNQEEVDHYGRHHHQQQQHSQAFFQPLECEPILQIGYQGQQDGMAGPSVNNYM 240
AthSEP3_3  MPLQLNPNQEEVDHYGRHHHQQQQHSQAFFQPLECEPILQIGYQGQQDGMAGPSVNNYM 226
AlySEP3    MPLQLNPNQEEVDHYGRH----QHSQAFFQPLECEPILQIGYQGQQDGMAGPSLNNYM 236
           *******************    *******************************:****

AthSEP3_1  LGWLPYDTNSI 250
AthSEP3_2  LGWLPYDTNSI 251
AthSEP3_3  LGWLPYDTNSI 237
AlySEP3    LGWLPYDTNSI 247
           ***********
```

***Figure 1**: **SEP3 protein sequence and expression conservation.** **A**: Multiple sequence alignment of AthSEP3 splice forms and AlySEP3. Species-specific differences are indicated in bold. The MADS-domain is labelled in red; the K-domain is marked blue. **B**: Maximum projection confocal images of inflorescence and young floral meristems of A. thaliana plants harbouring either pAlySEP3::AlySEP3-GFP or pAthSEP3::AthSEP3-GFP constructs. **C**: Sizes of mature floral organs in both species.*

Inflorescence material with floral buds up to stage 10-11 was harvested from *A. lyrata*, in order to use tissues that are morphologically as comparable as possible to the ones that we previously used in *A. thaliana* SEP3 ChIP-seq experiments. ChIP-seq was performed as described previously for *A. thaliana* (Kaufmann et al., 2009; Kaufmann et al., 2010b) in two biological replicates, using a mock-IP (pre-immune serum) as control. Analysis of the two biological replicates showed high level of reproducibility measured as number ($\log_{10}$) of mapped reads per 1-kb window (R= 0.84), as well as proportion of common BSs compared to the other replicate (**Figure S2**). This reproducibility is in the same range as other comparative ChIP-seq studies (He et al., 2011). For example the proportion of common BSs among different *D. melanogaster* replicates was 74% when considering the top 3,488 Twist BSs (He et al., 2011) which is comparable to 62% when using the top 3488 SEP3 BSs. For further analysis we focused on the replicate with higher statistical power, as measured by the number of BSs detected. Previous SEP3 ChIP-seq experiments from *A. thaliana* (Kaufmann et al., 2009) were re-analyzed using the same approach and the most up-to-date genome version.

ChIP-seq data analysis by CSAR (Muiño et al., 2011) revealed a slightly larger number (1.2 x) of SEP3 BSs in the *A. lyrata* genome (2,784; FDR<0.01) compared with the *A. thaliana* genome (2,276; FDR<0.01) (**Table 1**) which could be explained by the larger mappable genome size of *A. lyrata* (1.2 x). With the parameters used for read mapping during the ChIP-seq analysis, the length of the mappable nuclear genome used was 109 Mb for *A. thaliana* and 133 Mb for *A. lyrata*. However, *A. thaliana* shows a larger number of potential SEP3 target genes (3,979; FDR<0.01) than *A. lyrata* (2,831; FDR<0.01). We considered a gene as potential target of SEP3 when a SEP3 BS (FDR<0.01) is located within the 3 kb upstream and 1 kb downstream region of that gene. The larger number of potential target genes in *A. thaliana* is related to the fact that *A. thaliana* has a more compact genome, with an average distance of 3,334 bp between the start of genes, whereas *A. lyrata* shows a larger average distance (6,186 bp); therefore, a given BS in *A. thaliana* is more likely to be in proximity of more than one gene. SEP3 BSs in *A. lyrata* are located more often in intergenic regions (defined as regions not overlapping with the 3 kb upstream and 1 kb downstream of any gene) than in *A. thaliana* (**Figure 2A,B**), suggesting that *cis*-regulatory regions in *A. lyrata* may be found more distal from the start of the gene than in the compact *A. thaliana* genome.

Table 1: Number SEP3 BSs and potential target genes identified by ChIP-seq

	Total number of BSs (potential target genes) for different FDR thresholds		
	FDR < 0.05	FDR < 0.01	FDR < 0.005
A. thaliana	3233 (5466)	2276 (3979)	2043 (3622)
A. lyrata	4167 (4184)	2784 (2831)	2137 (2198)

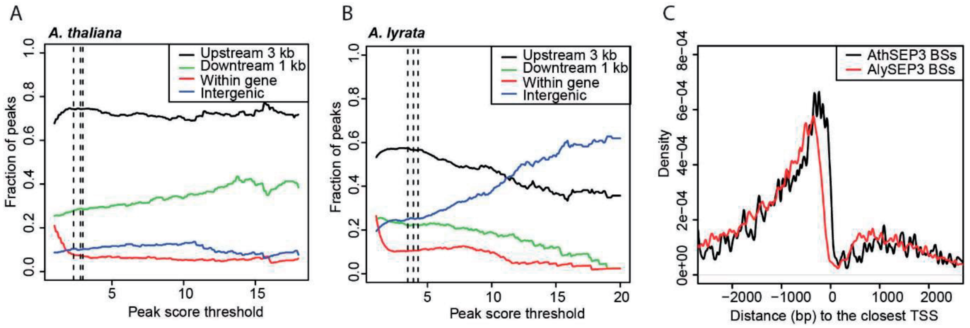


Figure 2: SEP3 binding relative to genomic features in Arabidopsis thaliana and A. lyrata. *(A, B) Enrichment of SEP3 BSs within promoters (black line, up to 3 kb upstream of gene start) and downstream regions (green, up to 1 kb downstream of end of gene) with the increase of the ChIP-seq score threshold used. BSs within genes (red line) and peaks in intergenic regions without any neighbouring gene (blue line) are also shown in the graph. Dotted vertical lines indicate FDR 0.05, 0.01 and 0.001, respectively.* ***C:*** *Distance of SEP3 BSs to the start of the closest gene in A. lyrata (red) and A. thaliana (black).*

Even within promoters (which we define as regions up to 3 kb upstream of the start of the gene), BSs in *A. lyrata* are located at larger distances to the start of the closest gene than in *A. thaliana* (**Figure 2C**).

To be sure that these results are not an artefact of the potentially different quality of the gene annotation used for each species, we created a new, *ab initio,* gene annotation using our inflorescence RNA-seq gene expression data (see Material & Methods). Comparing this new annotation with the TAIR10 and Araly1 gene annotations, some differences in the position of the start of the gene were found. For *A. lyrata,* 11% of the genes among the targets of SEP3 showed a difference in the start position of the gene larger than 500 bp, for *A. thaliana* this proportion was 6%. However, these differences do not affect the general results obtained with the TAIR10 and Araly1 gene annotation that are reported in **Figure 2** (see **Figure S3**).

Evolutionary turnover of SEP3 DNA-binding sites

To study the evolutionary history of individual SEP3-bound genomic regions and to get an estimate of the global BS turnover, we identified pairs of orthologous genomic regions in *A. lyrata* and *A. thaliana*. For this, we made use of the aligned genomes of the two species (Frazer et al., 2004; Dubchak et al., 2009) and used only alignments identified as orthologous regions (total size, 80 Mb). In total, 98% (2,229/2,276) of all SEP3-bound regions in *A. thaliana* and 83 % (2,313/2,784) of SEP3-bound regions in *A. lyrata* reside in detected orthologous genomic regions between both species. To study the level of evolutionary BS turnover between the two species, we focused on BSs located in alignable genomic regions, and took into account the level of reproducibility between independently generated biological replicates. Analogous to the comparative *Drosophila* ChIP-seq study by (He et al., 2011), we compared overlap of SEP3 BSs between biological replicates and between the two *Arabidopsis* species depending on the ChIP-seq score threshold (**Figure 3**).

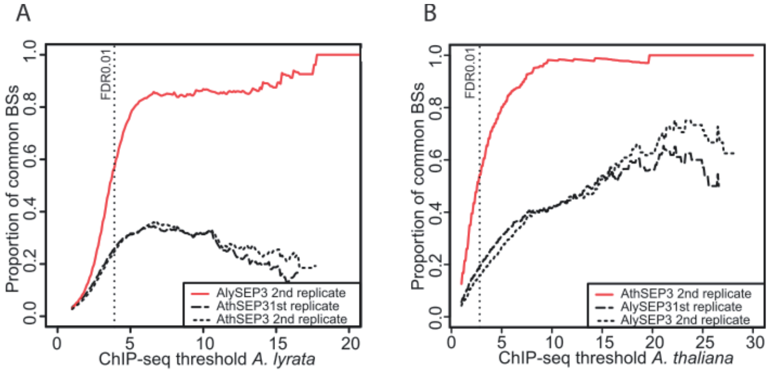


Figure 3: Proportion of conserved SEP3 BSs between and within species. *The plots show the proportion of common BSs between the best AlySEP3 replicate and the top 3,000 BSs of other ChIP-seq datasets (**A**), and the proportion of common BSs between the best AthSEP3 replicate and the top 2,000 BSs of the other ChIP-seq datasets (**B**). Only BSs located in regions that are alignable with the other species were considered.*

We found that at FDR 0.01, the overlap between biological replicates was limited, but reached levels greater than 80% (*A. lyrata*) and greater than 90% (*A. thaliana*) at higher score thresholds. This confirms a good reproducibility between the biological replicates (see also **Figure S2**). Similarly to He et al. (2011), in order to correct for the different numbers of BSs that were identified in the datasets at any FDR threshold, we compared the proportion of common BSs using a fixed number of total BSs. For example, for the AlySEP3 replicate with the largest statistical power, we detected near 3,000 BSs at FDR<0.01 (**Table 1**). Therefore, we calculated the proportion of common BSs compared with the top 3,000 BSs identified in the other AlySEP3 biological replicate (**Figure 3A**), and with the top 3,000 BSs identified in each AthSEP3 replicate. We found that only maximal 35% of the AlySEP3 BSs are conserved within any of the two *A. thaliana* replicates. This fraction is significantly lower than the reproducibility between biological replicates in *A. lyrata* (**Figure 3A**). In particular, at FDR<0.01 the proportion of common BSs between species was, on average, 26%, while among AlySEP3 replicates it was 60%. If we consider a threshold at which the proportion of common BS between replicates is 90%, we obtained a proportion of conservation of 21% between species. Similar conservation ratios are obtained if the top 2,000 BSs are used instead of the 3,000 top BSs. For example, at a proportion of common BSs of 90% between replicates, we obtained a proportion of conservation of 20% between species.

In a similar manner, we studied the BS reproducibility and conservation using the best AthSEP3 ChIP-seq replicate as reference (**Figure 3B**). As the number of BSs that was detected was approximately 2,000 at FDR<0.01, we estimated the proportion of conservation with the top 2,000 BSs in the other samples. Here, we found that at FDR<0.01 the proportion of common BSs between species was, on average, 18%, and among AthSEP3 replicates was 75%. If we consider a higher threshold, with 90% of common BSs between replicates, then we obtain a proportion of conservation of 21%.

We then looked at the function of genes located in the vicinity of the common 529 BSs (at FDR<0.01). Among the potential target genes, there was an enrichment (BINGO, (Maere et al., 2005)) of gene ontology (GO) terms related to the main function of SEP3 when compared with all potential target genes near the 2,229 AthSEP3 BSs. In particular, “negative regulation of developmental process”, “post-embryonic organ development”, “stamen development”, ”androecium development”, and “floral organ development” ($p<7 \times 10^{-5}$) were the top five GO categories enriched (**Table S1**). Regarding TF families, MADS-box, GRAS and TCP families were the only families significantly enriched (hypergeometric test; p<0.05, only families with more than two members were considered) among the targets of the common 529 BSs when compared with all AthSEP3 targets (**Table S2**). The BS turnover is as low as 62 % (36 out of 58) when we only consider AthSEP3 BSs near a MADS-box, GRAS or TCP TF gene. This is significantly lower (p<0.012, Chi-Square test) than when considering all AthSEP3 BSs (76%). Therefore, our data indicate a high turnover of SEP3 BSs in general, but BSs near target genes potentially related to the core function of SEP3 show a lower turnover. Indeed, SEP3 BSs near major homeotic and other flower developmental key-regulatory loci are largely conserved (see **Figure 4** for some examples).

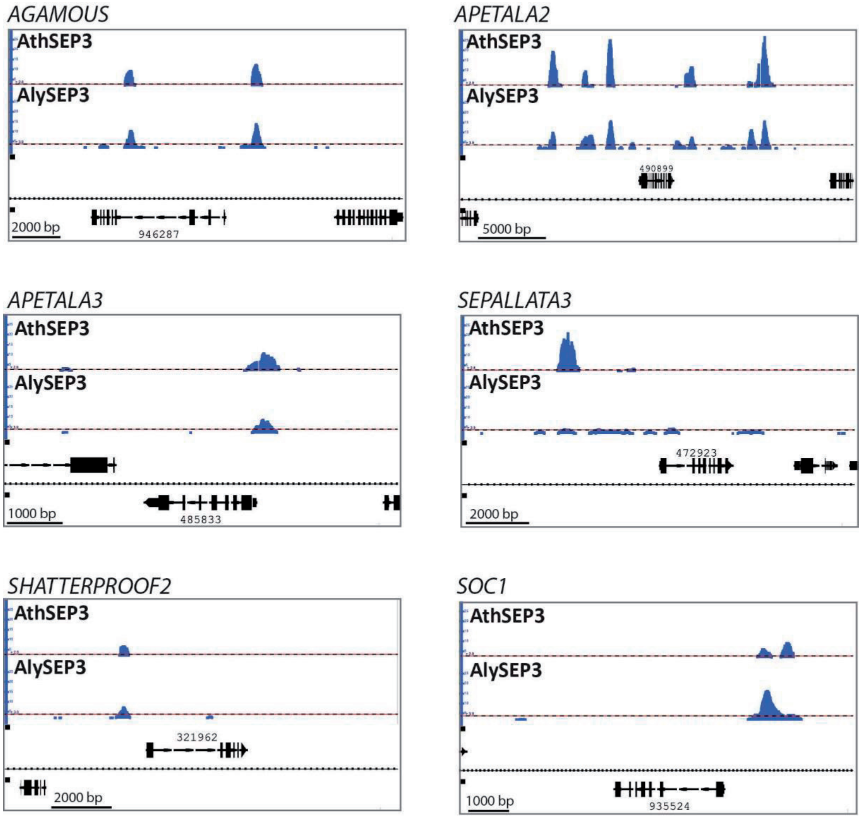


Figure 4: Conservation of SEP3 DNA-binding and potential direct target genes between A. lyrata and A. thaliana. *SEP3 BSs in several homeotic and other key-regulatory gene loci are shown in the aligned genomes. The respective genomic locus of each TF gene in A. lyrata is indicated. The horizontal dotted line indicates the FDR<0.01 threshold.*

DNA sequence and binding site conservation

Next, we studied the relationship between DNA sequence conservation and SEP3 BS conservation. To test how well the general level of DNA sequence conservation correlates with conservation of TF binding, we used PhastCons scores as a measure of conservation. The score of a given region represents the probability of belonging to a conserved element and ranges between 0 and 1. We obtained the PhastCons scores from nine Brassicaceae genomes from (Haudry et al., 2013). We found that the average PhastCons scores were significantly higher in genomic regions that were commonly bound by SEP3 in *A. thaliana* and *A. lyrata* than in regions that were bound specifically in either *A. lyrata* or *A. thaliana* (**Figure 5A**). We found an enrichment (**Figure 5B**) in regions defined as conserved non-coding sequences (CNSs) by Haudry et al. (2013) among the conserved SEP3 BSs compared with the *A. thaliana*-specific BSs (p<0.0001 Fisher's exact test) or compared to the *A. lyrata*-specific BSs (p<0.0001 Fisher's exact test) (**Figure 5B**). The presence of a CArG-box motif in the bound region in both species is also associated with BS conservation (**Figure 5C-D**). The CArG box sequences of *A. thaliana*-specific BSs contain more mutations, deletions, and insertions in their "orthologous" nonbound sequences in *A. lyrata* than CArG-boxes in BSs that are conserved between both species (**Figure S4A** and **D**). Previously, it has been described that the length of the A-tract region inside of the CArG-box motif is important for SEP3 DNA binding (Muiño et al., 2014). Indeed, the length of the A-tract inside of the CArG motif was more often maintained in conserved BSs than in species-specific BSs (**Figure S4C**).

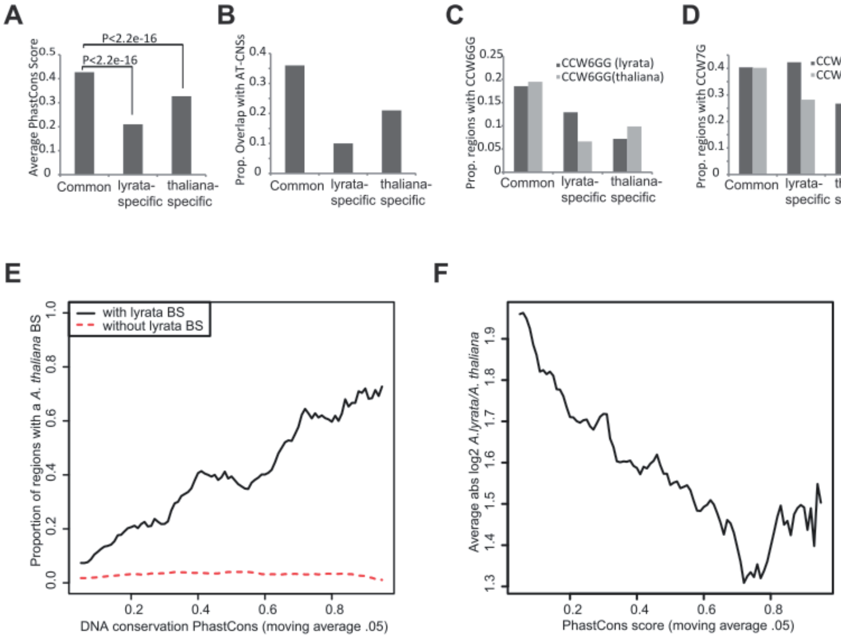


***Figure 5: SEP3 binding conservation vs. DNA sequence conservation**. **A**: Average PhastCons conservation score in SEP3-bound regions that are commonly bound in A. thaliana and A. lyrata, as well as in species-specifically bound regions. **B**: Proportion of common and species-specific SEP3 BSs overlapping with a conserved non coding sequence (CNS) defined by (Haudry, et al. 2013). **C-D:** Proportion of genomic regions with conserved or species-specific SEP3 binding that contain sequences matching the 'perfect' CArG box consensus (CC[A/T]$_6$GG) or (CC[A/T]$_7$G). **E:** Proportion of regions that are bound in A. thaliana out of the regions that are significantly bound (FDR<0.01) or not-bound (FDR>0.01) in A. lyrata (continuous line vs. dash line), depending on the PhastCons score. **F:** Quantitative changes in SEP3 binding levels depending on the PhastCons score in regions with a BS in at least one species. Regions with low PhastCons scores show larger quantitative changes in the SEP3 ChIP-seq score between both species than regions with higher PhastCons scores. Graphs **E** and **F** were calculated using moving average (window size 0.05). abs = absolute.*

The distribution of mutations along the CArG-box region is not uniform. The C/G nucleotides on the border of the motif, as well as some positions within the [A/T] rich core and certain surrounding positions are more often mutated in the *A. thaliana*-specific BSs than in the common BS regions (**Figure S4B**). Quantitative changes in SEP3 occupancy levels are associated with differences in PhastCons scores (**Figure 5F**; Pearson's r=-0.21; p<2.2 x 10^{-16}).

Genomic position and DNA binding site conservation

Prompted by the observation that the distribution of SEP3 BS position relative to their potential target genes was different in *A. thaliana* compared to *A. lyrata,* we studied how BS "relocation" may affect BS conservation. To our surprise, we detected a high variability in the position of SEP3 BSs relative to their potential target genes. Even when we considered only the 529 SEP3-bound regions common to both species at FDR 0.01, the relative positions to their potential target genes show a large variation (**Figure 6A**). Conserved AthSEP3 BSs located in promoter regions tend to be located further upstream in *A. lyrata* (-1.5 kb on average), meanwhile the ones located downstream the start of a gene tend to be located further downstream in *A. lyrata* (707 bp on average) (**Figure 6A, Figure S5**).

Our data show that BS conservation depends on the conservation of the location relative to the start of the target gene. When the BS was located originally in the core promoter region (1 kb upstream; **Figure 6B**, green line), the BS conservation measured as proportion of AthSEP3 BSs conserved in *A. lyrata* inversely depends on the extent to which their position has changed in *A. lyrata* (Pearson's r = -0.92; p<0.0002).

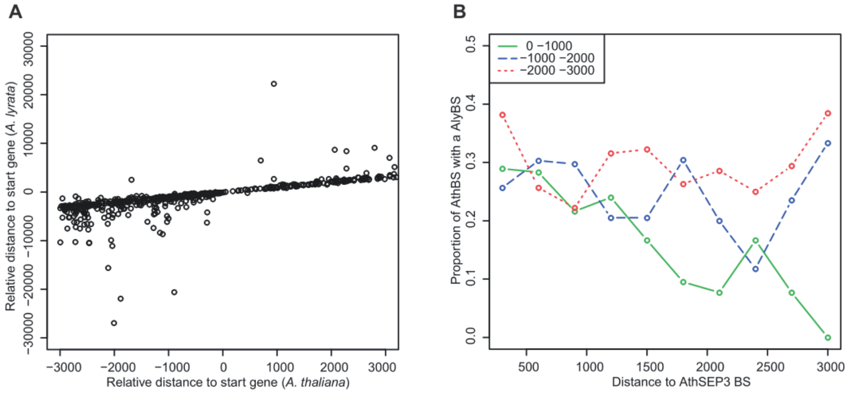


Figure 6. SEP3 BS conservation vs. position conservation. A: *AthSEP3 BS relative position to their target gene compared to their orthologous regions in Arabidopsis lyrata when the BS is conserved. We only considered the BSs that were common to both species. For different scale of the y-axis, see* ***Figure S4. B:*** *Proportion of AthSEP3 BSs that is conserved with A. lyrata depending on the location of the AthSEP3 BS relative to the start of the gene and depending on the distance to the AthSEP3 BS. The x-axis shows the distance between the AthSEP3 BS to its orthologous region in A. lyrata; 0 indicates that both regions are located in the same position relative to the TSS of their gene, a value of, for example 500 bp means that the orthologous region in A. lyrata is 500 bp upstream of the A. thaliana region relative to the gene.*

However when the AthSEP3 BS is located further upstream, changes in relative position to the target gene seem not to significantly affect BS conservation (Pearson's r = -0.16 p<0.64 for the region 1-2 kb, and r=0.09 p<0.80 for the region 2-3 kb) (**Figure 6B**). This suggests that in the case of BSs that are located in the core promoter, the position plays an important role in functionality, for example, due to required direct interactions with the basic transcriptional machinery. More distant SEP3 BSs seem to be more flexible in position.

Generation of new SEP3 DNA-binding events by transposition

Given that many SEP3 binding locations were species-specific, we were interested in potential mechanisms by which BSs may arise during short periods of evolutionary time. Although BSs can originate *de novo* by DNA sequence mutations, this is a slow process. It has been estimated that the time for a particular 10 bp motif to emerge *de novo* by mutation in a 1 kb promoter to be between 2 x 10^{10} and 4 x 10^{10} generations (Behrens and Vingron, 2010). An alternative mechanism is transposition of BS regions; transposons that harbor TF BSs can, potentially, 'amplify' their particular sequence to generate *cis*-regulatory elements in new locations. Later, these *cis*-regulatory elements can evolve to regulate nearby genes, although only few studies so far have demonstrated such a mechanism for the origin of novel 'functional' TF binding sites (de Souza et al., 2013). Recent ChIP-seq experiments on stem-cell regulatory TFs in humans and mice support this idea (Kunarso et al., 2010). The genome of *A. lyrata* shows a high number of transposons and transposon activity. About 50 % of the genomic sequence that is not present in *A. thaliana* encodes transposons (Hu et al., 2011). Despite our stringent mapping approach of the sequence reads, which discards reads that map to several genomic locations, we identified 307 AlySEP3 BSs for which the maximum ChIP-seq score position resides in TEs or other repetitive sequences. In contrast, only 16 AthSEP3 BSs reside in these elements. In *A. lyrata,* the BSs are specifically overrepresented in some types of elements, such as the super-families of DNA/MuDR and DNA/hAT transposons ($p<0.005$; hypergeometric test), as well as an uncharacterized repeat element family (rnd-6_family-174, hereafter abbreviated '*6-174*') (**Table S3**). Because of the particularly strong enrichment (89 out of 169 elements containing a SEP3 BS), we investigated the family *6-174* further. We found that sequences of this family are tightly associated with Long Terminal Repeat (LTR)/Copia retrotransposons in the genome: 96 out of 169 *6-174* members are directly adjacent to such a transposon, and all but 7 are located within a distance of less than 200 bp to an LTR/Copia type transposon. Multiple, largely conserved CArG boxes are frequently identified in sequences of the *6-174* family (**Figure S6**). 72 % of all *6-174* sequences that have a significant SEP3 BS possess at least one perfect CArG box of the consensus CC$[A/T]_6$GG, whereas only 54 % of all those sequences without a SEP3 BS possess a CArG box. CArG box sequences of type CC$[A/T]_7$G are not enriched in *6-174* sequences with SEP3 BSs. In *A. thaliana* there are only nine *6-174* elements. None of them shows a SEP3 BS in our data, neither do they contain a perfect CArG-box motif (CC$[A/T]_6$GG or CC$[A/T]_7$G). The outgroups *Capsella rubella* and *C. papaya* have none of these elements in their genomes. This indicates that the creation of these new BSs by the element *6-174* was a recent process and specific to *A. lyrata*. When studying the genes that are associated with transposons or other repetitive elements that have SEP3 BSs compared with genes associated with transposons or other repetitive elements without SEP3 BSs, we found an overrepresentation of genes involved in ‘embryo development’, ‘meristem structural organization’, and ‘anatomical structure arrangement’ among others (**Table S4**; p<**0.05**).

Among all genes with a TE inserted in their 3 kb upstream region in *A. lyrata*, 21% were significantly (FDR<0.05; foldchange > 0) more highly expressed in *A. lyrata* compared with *A. thaliana* inflorescences, whereas 16% were more highly expressed in *A. thaliana* (FDR<0.05; foldchange < 0). When we only consider TEs carrying a SEP3 BS, the proportions significantly change (p<0.032; Chi-square test) to 12% (more highly expressed in *A. lyrata*) and non-significantly change (p<0.45; Chi-square test) to 17% (more highly expressed in *A. thaliana*), which indicates that TEs containing a SEP3 BS may have a different effect in gene expression than TEs without any SEP3 BS. However, more experimental data are needed to assess the impact of SEP3 BSs located in the transposons on gene regulation.

Protein sequence evolution vs. DNA-BS conservation

Following Susumu Ohno (Ohno, 1970), after a duplication event, the retained duplicates may 1) diverge in function (neofunctionalization), and therefore one of the duplicated genes will retain the ancestral function, whereas the other duplicated gene may be relieved from purifying selection, allowing it to develop a novel function; 2) different functions or regulatory patterns of an ancestral gene might be split over the different paralogs (subfunctionalization); and 3) duplication may preserve the ancestral function in both duplicates, thereby introducing redundancy and/or increasing activity of the gene (gene dosage). Using the information from the SEP3 ChIP-seq data, we wanted to study the relation of TF regulation conservation and functional conservation of proteins (measured by the strength of purifying selection) in this context.

To approach this question, we only considered AlySEP3 target genes with one ortholog and at least one paralog in the *A. thaliana* genome (395 *A. lyrata* genes). when the *A. thaliana* paralog has a SEP3 BS, we observed that the presence of a SEP3 BS in the *A. thaliana* ortholog negatively depends on the strength of the purifying selection of protein sequences

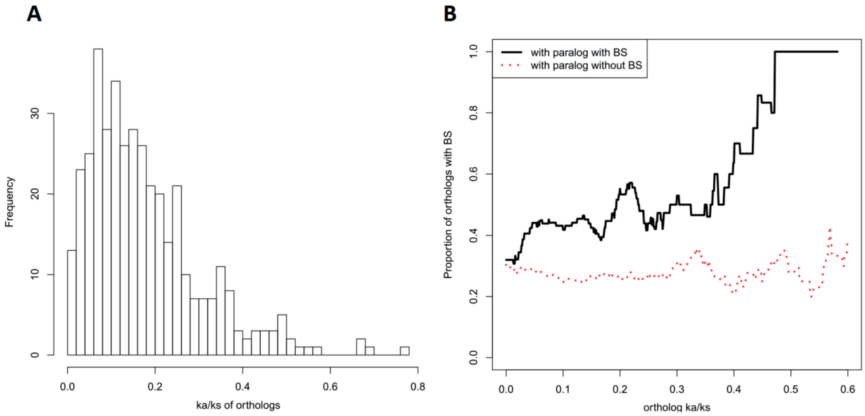


Figure 7. SEP3 binding conservation vs divergence time and the impact of purifying selection of associated genes. A: *Distribution of the K_a / K_s values for orthologous genes in Arabidopsis thaliana and A. lyrata.* ***B:*** *Proportion of orthologous genes with conserved BS depending on their K_a / K_s values when at least one A. thaliana paralog has conserved regulation (continues line) or not (dashed line). To estimate this proportion, a moving average was employed using overlapping windows of size 0.2. Only AlySEP3 target genes with orthologs in A. thaliana and at least one paralog in the A. thaliana genome were considered.*

of the orthologous genes (**Figure 7B**, black line). Specifically, when the *A. thaliana* paralog has a SEP3 BS, only 32% (8/25) of *A. thaliana* orthologs with $K_a/K_s < 0.1$ also have a SEP3 BS, whereas this proportion significantly increases to 57% (23/40) for orthologs with $K_a/K_s > 0.1$ ($P < 0.039$, Fisher's exact test). On the other hand, when the *A. thaliana* paralog does not have a SEP3 BS, the presence of a SEP3 BS in the *A. thaliana* ortholog is independent of purifying selection (**Figure 7B**, red line). Specifically, when the *A. thaliana* paralog does not have a SEP3 BS, 30% (31/102) of *A. thaliana* orthologs with $K_a/K_s < 0.1$ have a SEP3 BS, and this proportion does not change significantly ($P<0.21$, Fisher's exact test) for orthologs with $K_a/K_s >0.1$ (25%, 58 of 228 have a BS). In summary, considering *A. lyrata* genes with a SEP3 BS, there is a tendency to have less purifying selection at the level of protein sequence when both the *A. thaliana* ortholog and its paralog have a BS (potential regulatory conservation), compared with the situation when only one of them (the ortholog or the paralog) has a SEP3 BS (potential regulatory divergence).

Cross-species comparison of floral transcriptomes and potential direct target genes

To compare the gene expression levels of developing flowers in the two species, we generated directional mRNA-seq data of the same type of tissues as was used for the ChIP-seq experiments in *A. thaliana* and *A. lyrata*. Datasets were generated in three biological replicates, and showed a high level of reproducibility (**Figure S7** R ≈ 0.98). Quantitative comparison of the floral transcriptomes from the two species showed that the majority of orthologous gene pairs showed similar levels of expression. 2,454 out of 18,166 (14%) gene pairs were significantly differently expressed (FDR<0.05; abs(log2ratio)>1.5) among the floral transcriptomes of the two species. Combined with expression data from leaves generated with the same directional mRNA-seq protocol, we found that differences in expression of orthologous genes between tissues of the same species are higher than the changes in the same tissues of two closely related species. In contrast, paralogs show higher expression differences between species than between tissues (**Figure 8A**). This suggests that orthologous gene pairs are evolutionarily constrained to maintain their tissue-specific expression patterns, whereas paralogs can evolve lineage-specific expression patterns (and possibly lineage-specific functions). This trend is enhanced for genes with SEP3 binding sites (3 kb upstream of the start to 1 kb downstream of the end of the gene) in both species: orthologs show less expression differences, whereas paralogs tend to be more differentially expressed (**Figure S8**).

Next, we studied changes of gene expression associated with loss or gain of SEP3 BSs. We found that orthologous genes with SEP3 BSs in both species tend to have a conserved expression (**Figure 8B**). In contrast, orthologs with species-specific BSs have a slightly higher proportion of differentially expressed genes (**Figure 8B**). Genes with higher occupancy levels of SEP3 in *A. thaliana* tend to be more strongly expressed in this species (**Figure S8**). These data suggest that loss or gain of SEP3 BSs can be associated with changes in gene expression, and they support the idea that SEP3 mainly acts as an activator of gene expression (Kaufmann et al., 2009). Nevertheless, most orthologous genes (irrespective of whether they have a SEP3 BS) have a more similar expression level in the two species based on our data than when comparing different tissues of the same species.

Finally, we were interested in the functional annotation of genes that were commonly or species-specifically bound by SEP3. Among the target genes that had at least one BS in both species, we found that various GO categories related to floral organ development, meristematic growth and hormonal responses were enriched (**Table S5**). This suggests that the core-regulatory functions of SEP3 in the two species are conserved. However, we also found that specific GO categories were enriched in a species-specific fashion. Among the GO categories that are specifically enriched for *A. lyrata*-specific SEP3 target genes, there are several categories related to cell wall morphogenesis, cell wall modifications, pollen tube growth and pollination. Among the GO categories enriched for *A. thaliana*-specific SEP3 target genes, there are several categories related with RNA interference (**Table S5**), as for example with genes such as *DEFECTIVE IN MERISTEM SILENCING 3*, *ARGONAUTE 10* and *KRYPTONITE*. Whether differences in SEP3 binding to specific target gene promoters are causal to phenotypic divergence of the two species requires future investigation.

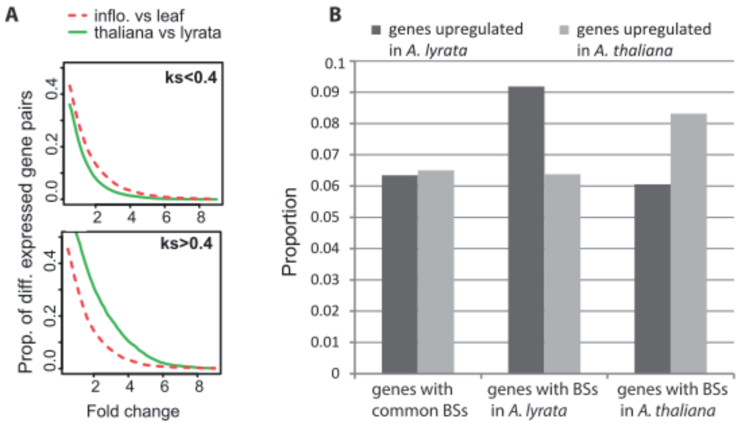


Figure 8. Evolution of gene expression versus changes in SEP3 binding. A: *Overall gene expression comparison in leaves and inflorescences of Arabidopsis lyrata, and in inflorescences of A. thaliana vs. A. lyrata. The analysis was done for orthologous genes and for paralogous genes.* ***B:*** *Proportion of genes with common or species-specific SEP3 BSs that have a higher expression either in A. lyrata or in A. thaliana inflorescences.*

Discussion

In this work, we compare the DNA binding landscapes of the floral master regulatory TF SEP3 in two closely related plant species as a first step to experimentally study the evolutionary dynamics of functional *cis*-regulatory elements in plants. We found that the level of SEP3 BS conservation between the species considered was on average 21%, around four times lower compared with the level of conservation between biological replicates when considering a ChIP-seq threshold in such a way that the proportion of common BSs between biological replicates was 90%. BS conservation was estimated only from the proportion of the genome that can be aligned between both species. Non-aligned regions represent regions present only in one species, or regions with a DNA sequence that is too divergent to be aligned. BSs in such regions (2% for AthSEP3 BSs, and 17% for AlySEP3 BSs) are therefore likely to be not conserved, suggesting that the true genome-wide proportion of conserved BSs may be slightly lower than our estimate. On the other hand, differences in tissue sampling or spatiotemporal variation of SEP3 binding in the two species could lead to an overestimation of BS variation.

Arabidopsis lyrata and *A. thaliana* diverged approximately 10 million years ago (Hu et al., 2011), which is in a similar time range of *D. melanogaster* and its closest relatives (~2.5-30 million years ago), or of human and macaque (~20 million years ago). The variation of BSs detected between *A. lyrata* and *A. thaliana* is negatively correlated with DNA conservation and the conserved presence of perfect CArG-boxes. Not all positions of the motif seem to have the same importance (**Figure S4**). However, many BSs are occupied in a species-specific manner despite presence of a conserved CArG box in the other species (**Figure 5C, D and Figure S4A**). This could be related to the fact that DNA sequence residues outside the core CArG box can contribute to BS functionality, and that binding of other (cooperatively acting)

TFs or chromatin structure is perturbed in the other species, as has been indicated by studies in animals (for review, see Villar et al. (2014)). Variability of BSs is affected by transposon activity. The possibility that TF BSs can be generated by transposition has been described previously (Feschotte, 2008). For example, there is evidence of CTCF and REST BS amplification in mammals associated with TEs (Johnson et al., 2006; Schmidt et al., 2012). The importance of this mechanism seems to be dependent on the group of species studied, since for example, in *Drosophila,* no examples of association between transposon activity and BS variation have been identified yet (Ni et al., 2012). This could be related to the fact that mammalian genomes are rich in TEs (de Koning et al., 2011), while *Drosophila* genomes have a much lower content of these elements (Lynch et al., 2011).There is also evidence that TEs can move the BSs of cell-cycle and developmental regulator E2F in plants (Hénaff et al., 2014). Therefore, transposition in *A. lyrata* could explain part of the observed *A. lyrata*-specific BSs. We detected a much higher proportion of BSs located in transposons in *A. lyrata* than in *A. thaliana*, supporting the idea that in *A. lyrata* transposition is a more important mechanism creating new BSs than in *A. thaliana*, which is in line with the fact that transposon activity is much higher in *A. lyrata* than in *A. thaliana* (Wright et al., 2001; Hollister et al., 2011).

The difference in genome size between *A. lyrata* and *A. thaliana* is not only due to the deletion/insertion of large genomic regions, but mostly due to small deletions/insertions (indels) (Hu et al., 2011). These indels may change the position of *cis*-regulatory elements relative to their potential target gene. Indeed, we observed a significant variation (**Figure 6B**) in the relative position of orthologous BSs to their orthologous candidate target gene. We also observed that a change in the relative position is negatively associated with the conservation of the BS, when the BS is originally located in close proximity to the start of the gene (0 to 1 kb upstream region). This may be one mechanism for creating gene regulatory diversity in plants, where, in contrast to mammals, genome expansion and reduction events are relatively frequent (Bennetzen et al., 2005; Dehal and Boore, 2005; Hawkins et al., 2009).

But how can a low level of SEP3 BS conservation agree with functional conservation? Interestingly, BS conservation is not uniform across the genome. One of the main functions of SEP3 is to control the specification and development of floral organs. Potential direct target genes involved in floral organ development (and related ontology terms; see Results section) show higher BS conservation than genes with other functions. This can explain how the plant can tolerate the low BS conservation: essential target genes for the function of the TF are often conserved, but there is a higher rate of BS turnover in other regions that are not essential for the (main) function of the TF. It is also possible that many BSs are without any regulatory function, but are rather a byproduct of evolution, non-functional at this moment in the context of gene regulation. Based on previous results, it is known that usually only for a subset of genes with BSs of floral MADS-domain TFs, a regulatory function based on gene expression profiling experiments can be identified (Kaufmann et al., 2010c; Wuest et al., 2012; Ó'Maoiléidigh et al., 2013; Pajoro et al., 2014b). In fact, this is a common phenomenon and may result from a combination of experimental limitations (e.g. range of tested conditions), and lack of regulatory activity of certain BSs. For example, a recent study on knockdowns of 59 TFs in human cell culture found that the global effect on gene expression was very low, as the median proportion of potential target genes with altered expression was 9.2% among the 59 TFs studied (Cusanovich et al., 2014).

We found that the relative position of the BS to its candidate target gene affects the conservation of the BS. However, this correlation was only significant for BSs located in the proximal promoter regions (up to 1 kb upstream start of the gene; **Figure 6**). BSs located in distal promoter regions (1 to 3 kb upstream) did not show a significant correlation. This indicates that BSs located up to 1 kb upstream of the start of the gene depend on some type of interaction with the TSS in order to exert their role in gene regulation. On the other hand, BSs located 1-3 kb upstream of the start of the gene do not show this correlation, suggesting that their function seems to be independent of their relative position to the start of the gene. The classical definition of enhancers identifies these elements as insensitive to changes in position and orientation relative to the start of the gene (Maston et al., 2006). Therefore, the BSs which are independent in their position relative to the start of the gene can be considered as being located in enhancer elements, whereas the ones that depend on their relative position to the start of the gene will be located in the core promoter.

The possibility that TF BSs can be generated by transposition has been described previously (Feschotte, 2008). For example, there is evidence of CTCF and REST BS amplification in mammals associated with transposable elements (Johnson et al., 2006; Schmidt et al., 2012). But how gradual or fast is this process? We found that the family of repetitive elements 'rnd-6_family174' alone has contributed to the creation of 89 new SEP3 BSs in *A. lyrata*. This family of repetitive elements is highly amplified specifically in the *A. lyrata* genome where 169 locations of this family can be identified in the genome, in contrast to 9 in the *A. thaliana* genome, and none in *Cap. rubella* and *C. papaya* genomes. Therefore, the most likely model to explain this observation is that there was a 'burst' of LTR/Copia transposition events that amplified this family of repetitive elements at some moment after the divergence of *A. thaliana* and *A. lyrata,* which led to the amplification of SEP3 BSs in the *A. lyrata* genome. Indeed, the evolution of repetitive families follows a 'burst and decay' model (Maumus and Quesneville, 2014) with the proliferation of identical copies that with time accumulate mutations and deletions until they are distinct in sequence from the original copies. When the repetitive element carried a *cis*-regulatory element, this mechanism of multiplication has the consequence of increasing the number of BSs, and therefore the regulatory diversity, in a short period of time, which could be advantageous for the plant to adapt to new conditions.

In our study, we found evidence of the importance of subfunctionalization in the evolution of SEP3 binding after duplication of its target genes. When both paralogs conserved the SEP3 binding there are low levels of purifying selection acting on the protein sequence of the target gene, and likely this will allow for functional diversification of the proteins (neofunctionalization). When only one of the paralogs conserved the ancestral SEP3 binding, there is higher purifying selection and likely the paralogs will keep similar protein function, although their regulation and perhaps their tissue specificity may vary (subfunctionalization). This is in line with the study of (Castillo-Davis et al., 2004) in *Caenorhabditis elegans* that reports that selection can act independently on gene regulation and protein sequence for duplicated genes, while prior to a duplication event the selection on gene regulation and protein sequence is weakly coupled. An important difference between *A. thaliana* and *A. lyrata* is the phenotypes associated with the reproduction strategy. Evolutionary transition from outcrossing to selfing is usually linked with smaller flower size, closer opening angles of petals, lower pollen-to-ovule ratio, reduced separation between anthers and stigma (herkogamy), and less nectar and scent production (Sicard and Lenhard, 2011). The molecular basis of the correlated evolution of these floral phenotypes is still unclear. Most interestingly, it is unknown whether there is a common molecular mechanism that explains this co-evolution of floral characteristics. Because SEP3 is a master regulator of floral development, it is possible that there is an association between changes in its target gene networks and the co-evolution of the phenotypes associated with the mating strategy. Among the *A. lyrata*-specific targets, we detected an interesting enrichment in GO terms: cell wall loosening (which can be associated with organ growth), pollen tube growth and pollination among others (see **Table S5**; Results section). Among the genes that are involved in pollen tube growth and pollination are genes such as *POLLEN DEFECTIVE IN GUIDANCE 1 (POD1)* (Li, et al. 2011) as well as *ROP-INTERACTIVE CRIB MOTIF-CONTAINING PROTEIN 3 (RIC3) and ROP1 ENHANCER 1 (REN1),* which have functions in ROP1 Rho GTPase-dependent pollen tube growth (Guan et al., 2013). In future research, it will be interesting to study to which extent the observed differences in SEP3 binding are causally associated with the alternative mating systems in the two species.

Material and methods

Plant growth

A. lyrata ssp. *lyrata* plants were grown on soil under standard long-day conditions. After germination the plants were vernalized for 7-10 weeks at 8 °C, and then transferred to 20 °C, standard long-day conditions. Alternatively, plants were vernalized for 7 weeks at 8 °C /4 °C day/night under short day (12h day, 12h night), and then transferred to 20 °C, standard long-day conditions. *Arabidopsis thaliana* plants were grown on rock-wool in a growth chamber with standard long-day conditions (16h day, 8h night).

Reporter GFP construct of AlySEP3 promoter and genomic locus

AlySEP3 with upstream region was amplified using primers Fw 5’-CTTGACTAGCCCACAACACTTC-3’ and R 5’-AATAGAGTTGGTGTCATAAGGTAACC-3’. The polymerase chain reaction fragments were cloned into the GATEWAY vector pCR8/GW/TOPO from Invitrogen and transferred through LR reaction into the destination vector AM884381 (pGREEN-GW-eGFP; (Zhong et al., 2008). Expression vector was introduced into *A. thaliana* ecotype *Col-0* by floral dip transformation. Transformant plants were selected on MS medium with BASTA. For comparison, we used previously generated pSEP3::SEP3-GFP plants (4.1 kb promoter, (Smaczniak et al., 2012b).

Confocal scanning laser microscopy

GFP tagged protein localization was observed trough CSLM on Leica SPE DM5500 upright microscope using a ACS APO 40x/1.15 oil lens and using the LAS AF 1.8.2 software. GFP was excited with the 488-nm line of an Argon ion laser. Confocal image acquisition was performed essentially as described in (Urbanus et al., 2009), with the GFP emission filtered with a 505-530 nm band pass filter, and chloroplast autofluorescence with a bandwidth of 650 nm (long pass filter). Image processing and three-dimensional projections were performed using the LAS AF 1.8.2 software package.

Scanning electron microscopy

The SEM procedures were essentially as in (Caris et al., 2006). Plant material was fixed in FAA (40% formalin, acetic acid, 70% alcohol, 5:5:90) and buds were dissected in 70 % ethanol under a stereo-microscope. Dehydration was through a series of 70 % ethanol, a mixture (1:1) of 70 % ethanol and DMM (dimethoxymethane) each for 5 min. and pure DMM for 20 min. The samples were critical point dried using liquid CO_2 in a BAL-TEC CPD030 (BAL-TEC AG, Balzers, Liechtenstein). The material was mounted onto stubs and gold-coated with a sputter coater (SPI Supplies, West Chester, Pennsylvania, USA). Observations were made using a JEOL JSM-6360 microscope at the Department of Biology, KULeuven.

ChIP-seq data generation and analysis

Publically available AthSEP3 ChIP-seq data (Kaufmann et al., 2009) was downloaded from GEO (GSE14600). In particular, SRR016810 was as IP sample (first replicate), SRR016813 as IP sample (second replicate), and SRR016812 as control sample. Generation of ChIP samples and preparation of Illumina sequencing libraries were performed on *A. lyrata* inflorescences essentially as described previously (Kaufmann et al., 2010b). Sequencing libraries for AlySEP3 ChIP-seq were generated using Genome Analyzer IIx, HiSeq2000 or Miseq, see Table S6 for a summary of number reads generated. Low-quality reads from libraries sequenced with Hiseq2000 were removed as this is not done automatically as for the Genome Analyzer IIx. The sequence datasets were submitted to Gene Expression Omnibus (GEO) (accession number GSE63464). Sequences in FASTQ format were mapped to the unmasked *A. thalian*a genome (TAIR10) or to the *A. lyrata* genome (Araly1) depending of the origin of the library using SOAPv2 (Li et al., 2009). A maximum of two mismatches and no gaps were allowed, and reads were iteratively trimmed from the 5` end until mapped or their length fell below 31 nt. Only uniquely mapped reads were retained. Sequence reads mapping to the plastid and mitochondrial genomes were eliminated. For *A. lyrata*, only reads mapping to the 9 longest scaffolds were retained (scaffold length > 1Mb). The R package CSAR was used for peak calling for each biological replicate independently with default parameter values except for *backg*, which was set to 5 for all the analyses except AlySEP3 ChIP-seq replicate 1 which was set to 14. A value of 5 for the parameter *backg* indicates that regions having less than 5 reads mapped in the control were set to 5 to avoid false-positive results due to the low coverage of the control in some regions. We set the value of *backg* in AlySEP3 ChIP-seq replicate 1 analysis to 14 because the higher coverage of these libraries (see Table S6). FDR thresholds were estimated by permutation of reads between IP samples and controls using CSAR for each biological replicate independently and using default parameter except *backg* which was set to 5. Reproducibility of the biological replicates was estimated taking counting number of mapped reads (log2) in non-overlapping windows of size 1kbp it gives a high Pearson correlation coefficient for *A. lyrata* (r=0.842). Only the biological replicate showing a higher enrichment on BSs near start of the gene was used for further analysis. Candidate target genes were defined as genes containing a significant (FDR<0.01) binding event in the region between 3 kb upstream and 1 kb downstream of the annotated gene. Gene annotation was obtained from Phytozome v8.0, only annotation denoted as "mRNA" was used, and only these loci defined as primary transcripts.

RNA preparation for RNA-seq

RNA was prepared from *A. lyrata* tissue samples using the Invitrap spin plant RNA mini kit (Stratec) according to the manufacturer's instructions. RNA concentrations were determined using a Nanodrop ND-1000 spectrophotometer (Thermo Scientific).

RNA-seq analysis

Directional RNA-seq libraries were generated and sequenced in triplicates for *A. thaliana* inflorescences, and *A. lyrata* inflorescences and leaves (**Table S6**). For each library independently reads were mapped to the transcriptome sequence of corresponding organism. We downloaded the sequences of the primary transcript from Phytozome version 8.0, file: "Athaliana_167_TAIR10.transcript_primaryTranscriptOnly.fa" for *A. thaliana* and "Alyrata_107_transcript_primaryOnly.fa" for *A. lyrata*. Read mapping was done with SOAPv2 with default parameter values. Reads mapping to more than one transcript or to the mitochondrial or chloroplast transcriptomes were discarded. Only reads mapping to the forward strand of the transcript were used for further analysis. Because orthologous genes may have different length in *A. thaliana* compared to *A. lyrata*, read count values were normalized by transcript length (as number of reads per kilobase). After that, transcripts with normalized by transcript length count values lower than 10 were set to 10 to avoid any false positive due to the low number of counts (close to zero) in some transcripts. Later, the R package Deseq (Anders and Huber, 2010) was used with default parameters to detect differential expression.

Gene re-annotation of *A. lyrata* and *A. thaliana* genomes

Because the gene annotation of *A. lyrata* and *A. thaliana* may be of different quality, we have re-annotated *ab initio* these two genomes using our RNA-seq expression data. In particular, we mapped the three inflorescence RNA-seq biological replicates to their corresponding genome using TopHat (version 2.0.14; (Kim et al., 2013)), the previous gene annotation of each genome was not used for the mapping. Later, we used StringTie (version 1.0.2; (Pertea et al., 2015)) to reconstruct *ab initio* the transcriptome of both genomes, this is, without use the previous information about the gene annotation. For *A. thaliana* 23,739 transcripts were detected ion the nuclear genome, meanwhile for *A. lyrata* 30,793 transcript were detected in scaffold 1 to 9.

Linking *A. thaliana* and *A. lyrata* genomic data

To link *A. thaliana* and *A. lyrata* genes, we download pairs of homologous *A. thaliana-A. lyrata* genes together with their estimated K_s and K_a values from the Plant Genome Duplication Database (PGDD; http://chibba.agtec.uga.edu/duplication), homologous with K_s =-1 were removed. A K_s value= -1 means that no estimation of K_s was possible to obtain. This information was used to link the expression values of genes that after will be tested for differential expression between *A. lyrata* and *A. thaliana*.

This information was also used to link candidate target genes of SEP3 in both plant species. Lists of orthologous gene pairs obtained by the method "Best-Hits-and-Inparalogs family" were downloaded from PLAZA dicot 3.0 (Proost et al., 2015). We consider as paralogous pair of genes, these homologous pairs of genes obtained from PGDD that were not classified as orthologous gene pairs. To link binding events between both species, we downloaded whole genome alignments of *A. lyrata* and *A. thaliana* from the VISTA software (2/12/2013). We only use dual monotonic alignments, this is, alignments of orthologous (best bidirectional hits) regions (Dubchak et al., 2009). Using the position and strand of the alignments in both organisms, we can calculate the genomic position of a given *A. thaliana* nucleotide in *A. lyrata* and vice versa, when an alignment covers the region of interest. We use this property to translate the ChIP-seq score values for one organism to the other and vice versa at single nucleotide position, and therefore to generate ChIP-seq profiles in wig format that could be represented in one desired genome. We also used this property to translate the position of significant BSs from one species to the other. Then, we linked *A. thaliana* candidate BSs to their *A. lyrata* counterpart when the position of the maximum ChIP-seq score value between both BSs was less than 300 bp. When no BS was found in *A. lyrata*, it was reported as missing in *A. lyrata* using the value NA. The same method was applied to link *A. lyrata* candidate BSs to *A. thaliana* regions. Both lists were added together and only one pair of *thaliana*–*lyrata* BSs were kept when found to be duplicated.

DNA sequence and CARG-box motif conservation

PhastCons scores were obtained from (Haudry et al., 2013), the Phastcons score of a given region represents the probability of belonging to a conserved element and therefore ranges between 0 and 1. They were calculated by (Haudry et al., 2013) from the whole-genome alignments of nine Brassicaceae genomes. We associated PhastCons score to a given BS region as the average phastCons score on the +/-100bp region around the position of the maximum ChIP-seq score value of the significant BS.

We identify a BS as containing a CARG-box motif, if the region 250 bp around the position of the maximum ChIP-seq score value contains the motif CCW_6GG without any mismatch.

Gene ontology term enrichment analysis

BINGO version 2.44 was used to detect GO term enrichment. Because the gene annotation and ontologies used by default by BINGO date from August 2010, we have updated our version of BINGO with annotation and ontology files downloaded from www.geneontology.org (on 6/25/2014)

Transposon analysis

A database of TE insertions from *A. thaliana* and *A. lyrata* was obtained from (Hu et al., 2011). It contains assembled parallel datasets of TE insertions from *A. thaliana* (TAIR8) and *A. lyrata* (Araly1) genome using RepeatModeler (Smit 2008-2010). This follows in the identification of 1,152 repeat units. We used this library to annotate *A. thaliana* (TAIR10) and *A. lyrata* (Araly1) using RepeatMasker version 4.0.3 (Smit 1996-2010). Simple repeats were discarded from further analysis.

Supplementary material

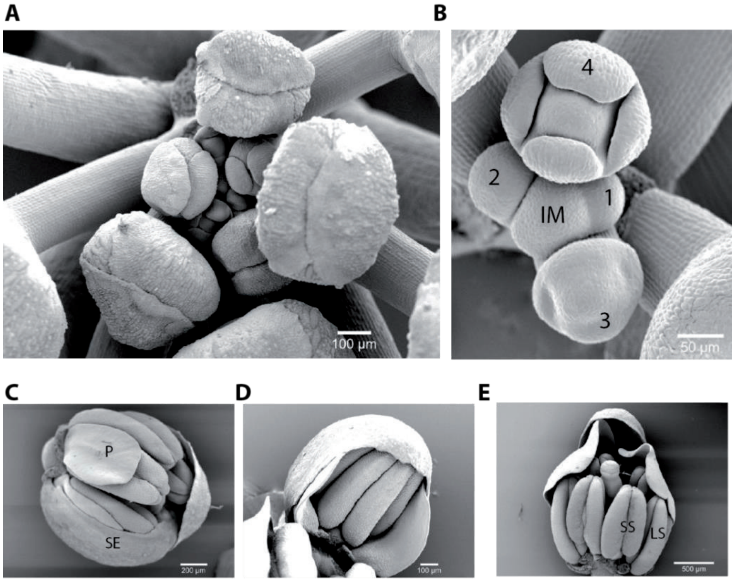


Figure S1: Scanning electron microscopy images of early flower developmental stages in A. lyrata. A: *Transverse view of the apex of an inflorescence in which the oldest flower has reached stage 7-8.* ***B:*** *Inflorescence meristem (IM) and early floral meristems up to stage 4.* ***C-E****: Flowers of stages 8/9 to 12. At stage 12, the petal length exceeds that of the long stamens, in contrast to A. thaliana.*

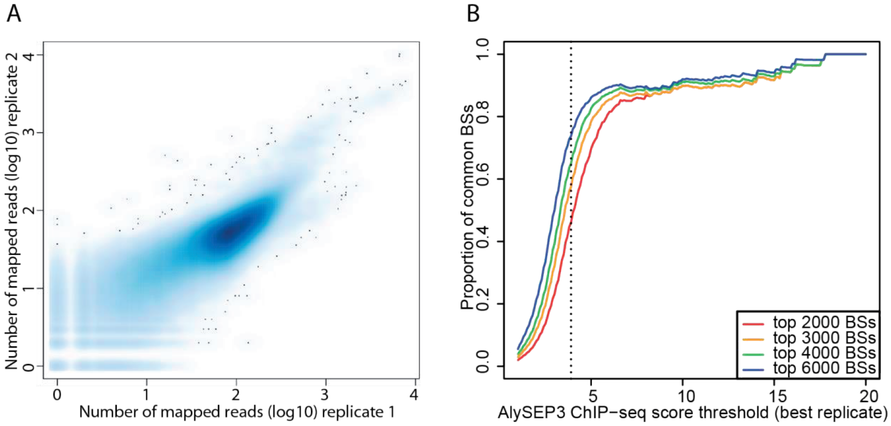


Figure S2*:* ***Reproducibility of AlySEP3 ChIP-seq replicates****.* ***A:*** *Scatterplot of number of mapped reads (log$_{10}$) for each biological replicate using non overlapping windows of size 1 kb. Only scaffolds 1 to 9 were considered.* ***B:*** *Proportion of AlySEP3 BSs identified in replicate 1 that are in common with AlySEP3 replicate 2, for different ChIP-seq score thresholds. Vertical dashed line indicates ChIP-seq score threshold for FDR<0.01 (2,784 BSs identified at this threshold).*

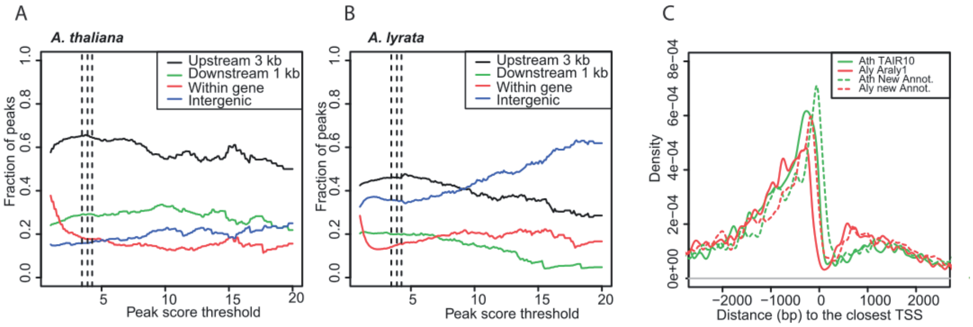


Figure S3: SEP3 binding relative to genomic features in A. thaliana and A. lyrata using the gene annotation generated in this manuscript A, B: *Enrichment of SEP3 BSs within promoters (black line, up to 3 kb upstream of gene start) and downstream regions (green, up to 1 kb downstream of end of gene) with the increase of the ChIP-seq score threshold used. BSs within genes (red line) and peaks in intergenic regions without any neighbouring gene (blue line) are also shown in the graph. Dotted vertical lines indicate FDR 0.05, 0.01 and 0.001, respectively.* ***C:*** *Distance of SEP3 BSs to the start of the closest gene in A. lyrata (red) and A. thaliana (black), when using the TAIR10 and Araly1 annotation (continuous line), and the gene annotation generated in this manuscript (discontinuous lines). In Figure S3A-C, for TAIR 10 and Araly1 gene annotation, only gene models that overlap a minimum of 1bp with genes identified by the annotation generated in this manuscript were used (see Material and Methods). We notice that, as average, the distance of SEP3 BSs to the start of the gene is shorter when using the RNA-seq based annotation than when using the TAIR10 and Araly1 annotation, the most likely explanation is that RNA-seq based annotation is able to detect the 5' UTR more easily that a sequence homology based gene annotation.*

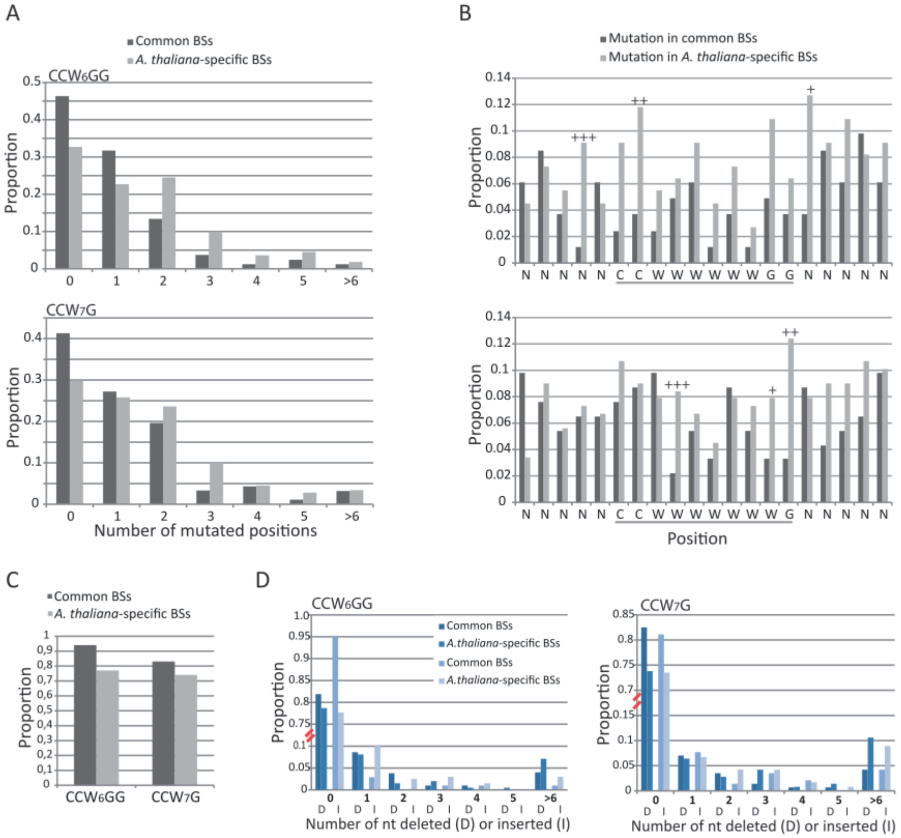


Figure S4: Mutations within CArG box sequences of conserved and non-conserved SEP3 BSs. *The analysis was performed for the canonical CArG boxes of type CCW$_6$GG and CCW$_7$G in orthologous positions in the A. thaliana and A. lyrata genomes based on genome alignment.* ***A:*** *Proportion of nucleotide positions with one or more nucleotide substitutions in the core CArG box sequences within conserved and A. thaliana-specific SEP3 BSs. CArG boxes with insertions and deletions were not considered here.* ***B:*** *Proportion of mutated nucleotides in each position of the central CArG box core and neighboring nucleotide positions. The three most mutated positions in A. thaliana-specific BSs compared to common BSs are indicated by +++, ++ and + in decreasing order; CArG boxes with insertions and deletions were not considered here.* ***C:*** *Conservation of A-tract length in the central [A,T] rich core of the CArG box in common and in A. thaliana-specific BSs in the A. lyrata genome. An A-tract element was defined with the motif A_mT_n, where n+m>3; only CArG boxes with an A-tract (n+m>3) were considered.* ***D:*** *Proportion of CArG box sequences with different numbers of inserted (I) and deleted (D) nucleotides (nt) in common and A. thaliana-specific SEP3 BSs.*

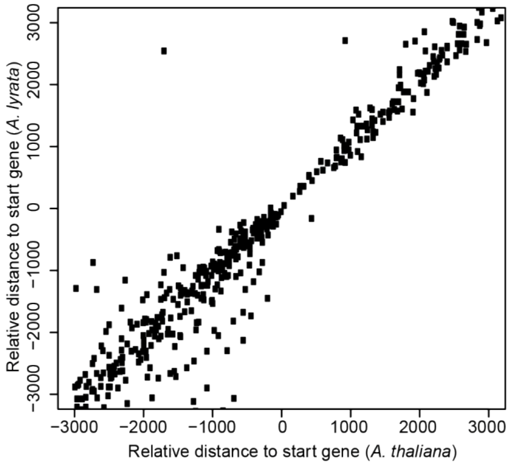


***Figure S5**: Relative position of AthSEP3 BSs to their target gene compared to their orthologous regions in A. lyrata when the BS is conserved. We only considered the BSs that were common to both species. Related to Figure 5, but a different scale was used here for the Y axis for alternative visualization.*

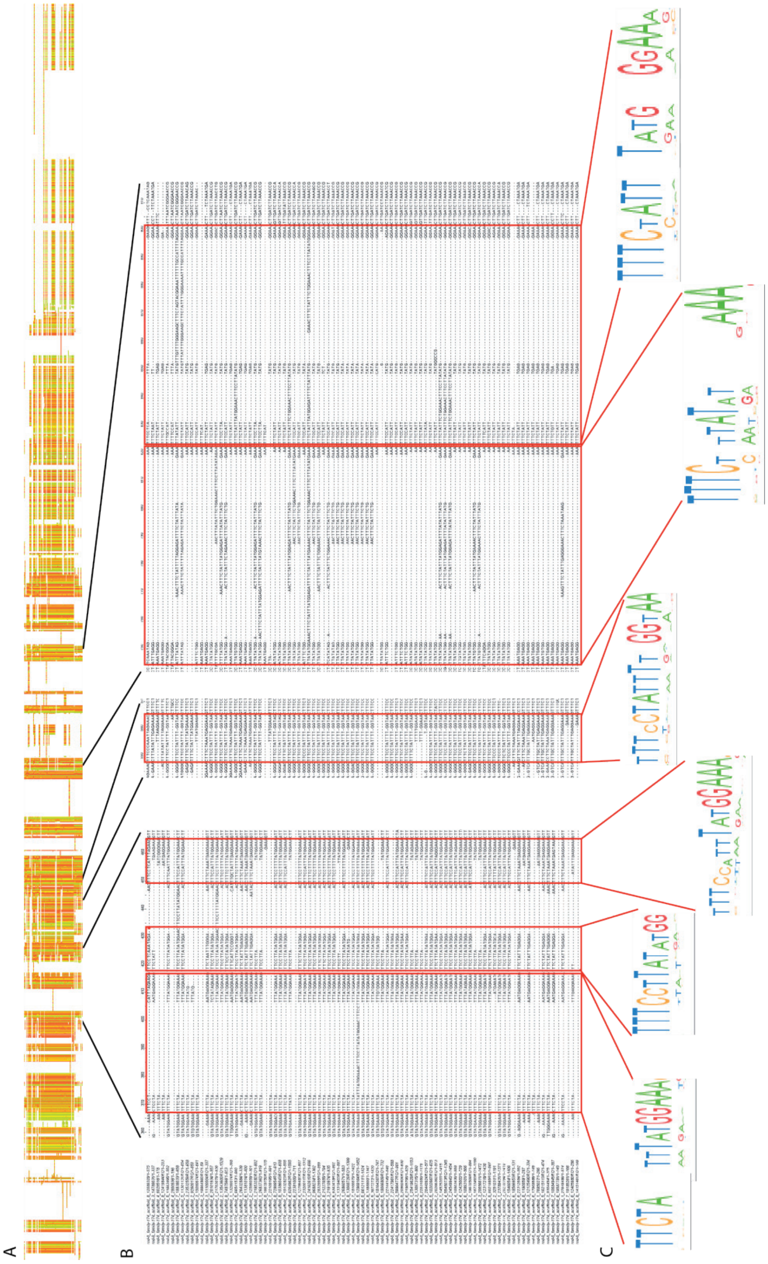


***Figure S6**: Alignment of selected region of 6-174 repetitive sequences with several CArG-box-like sequence motifs. An alignment of the 89 transposon sequences containing a SEP3 binding site was generated using Muscle (http://www.ebi.ac.uk/Tools/msa/muscle/). **A:** Overview of this alignment (made with JProfileGrid2; http://www.profilegrid.org). **B**: Parts of the alignment containing CArG-boxes, with CArG-boxes outlined in red (visualized in Jalview; http://www.jalview.org/). **C**: Motifs of these CArG-boxes (generated by Jalview).*

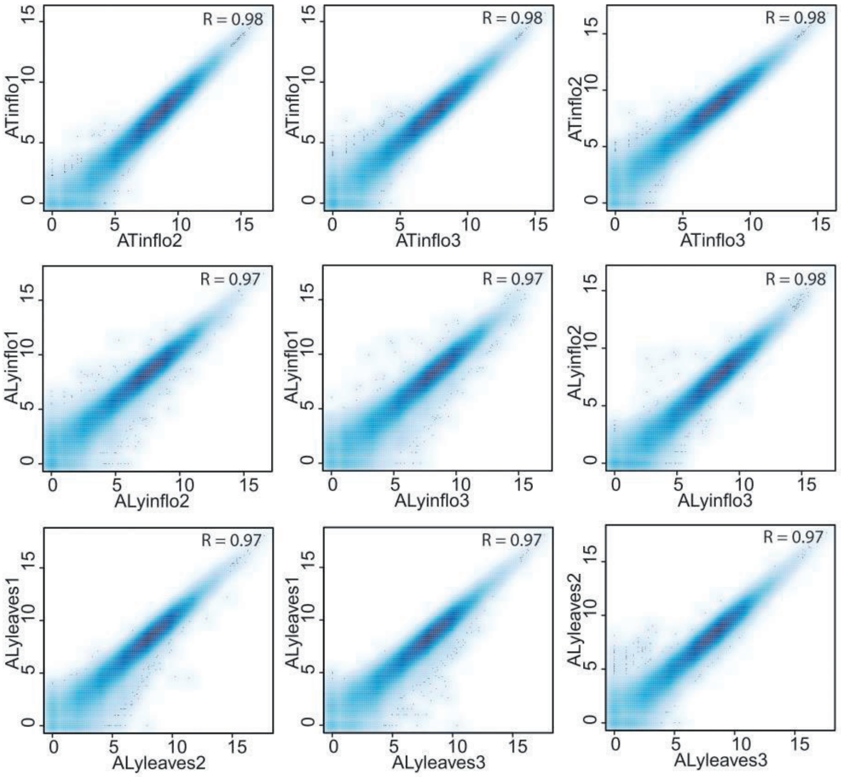


Figure S7: Reproducibility of RNA-seq replicates. *The graphs represent raw counts of reads (log$_2$) before normalization.*

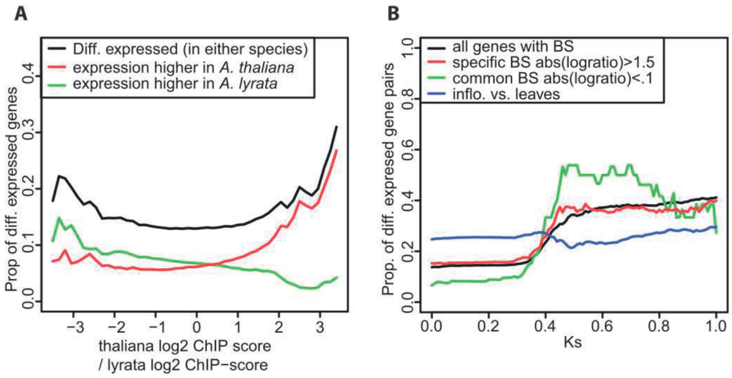


Figure S8: A: *Proportion of generally differentially expressed genes (black line), genes with higher expression in A. thaliana (red line) and genes with higher expression in A. lyrata (green line) depending on the ratio of ChIP-seq scores of [thaliana]/[lyrata]. In A we only consider orthologous gene pairs.* ***B:*** *Proportion of differentially expressed genes pairs depending on Ks. Quantitative changes in ChIP-seq scores in homologous genes were calculated as the maximum ChIP-seq score in the 3kb upstream 1kb downstream region of the gene. Later, these scores were normalized using quantile normalization between A. lyrata and A. thaliana genes The different lines indicate: all genes with SEP3 BSs in at least one of the species (black), genes with a quantitative ChIP-seq score log2 fold-change bigger than 1.5 (red), and genes a quantitative ChIP-seq score log2 fold-change smaller than 0.1 (green). As a control, the proportion of differentially expressed genes with a SEP3 BS in at least one species between A. lyrata leaves and inflorescences is shown (blue line).*

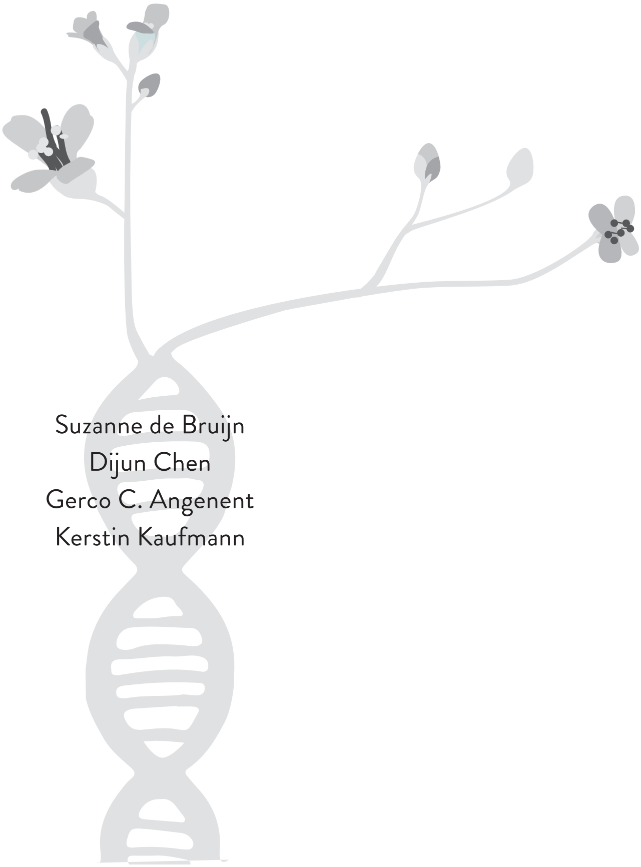
Supplementary tables S1–S6 are available at Molecular Biology and Evolution online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We wish to thank Prof. Ya-Long Guo (State Key Laboratory of Systematic and Evolutionary Botany, Chinese Academy of Sciences) for providing us with the assembled datasets of TE insertions from *A. thaliana* and *A. lyrata*. KK wishes to thank the Alexander-von-Humboldt foundation. SdB received a Netherlands Organisation for Scientific Research (NWO) Experimental Plant Sciences graduate school ‘master talent’ fellowship. Data generated in this project was submitted to GEO with id: GSE63463.

CHAPTER 4

Variation in Transcription Factor Binding Sites and DNA accessibility between *Arabidopsis* species and ecotypes



Suzanne de Bruijn
Dijun Chen
Gerco C. Angenent
Kerstin Kaufmann

Abstract

The (A)BCE-model of floral development is well-known and widely conserved among flowering plants. However, we know less about the pathways downstream of the transcription factors that underlie this model. Even though we start to identify downstream targets, we do not know yet how these targets of the major floral transcription factors evolve and contribute to generate natural variation. A first study on the evolution of binding sites showed a large amount of transcription factor binding site divergence for the floral regulator SEPALLATA3 between two closely related *Arabidopsis* species.

Here, we studied the level of binding site divergence within a species. We performed ChIP-seq experiments for SEP3 in different *A. thaliana* ecotypes. Our results show that there is a large overlap between the binding sites observed in these two ecotypes. This overlap was substantially larger than the binding site conservation found between *A. thaliana* and *A. lyrata*. Furthermore, peak intensity of the ecotype-specific binding sites we found was lower than that of the common binding sites, and showed less sequence conservation. In addition, we performed DNAse-seq assays to obtain profiles of open chromatin regions, and confirm previous findings indicating that the accessibility of chromatin may influence binding site selection.

Introduction

Complex Gene Regulatory Networks (GRNs) underlie plant development (Kaufmann et al., 2010a). These networks are expected to have evolved in species-specific ways to accommodate changes in floral morphology. In floral development, the (A)BCE-model of floral development is widely conserved, but little is known about the downstream targets and pathways. Although in a few model species (primarily in *A. thaliana*) these targets start to be elucidated (Kaufmann et al., 2009; Kaufmann et al., 2010c; Wuest et al., 2012; Ó'Maoiléidigh et al., 2013; Pajoro et al., 2014b), little is known about the molecular processes generating variation in the regulation of these targets during flower development.

How gene regulation evolves has mainly been investigated indirectly through sequence conservation studies. Several studies looked at conserved non-coding sequences in plants and tried to identify *cis*-regulatory elements (Baxter et al., 2012; Haudry et al., 2013; Hupalo and Kern, 2013; Van de Velde et al., 2014). However, none of these studies is capable of correctly predicting all transcription factor binding sites (TFBSs). The study by Haudry et al. (2013) recovered the most TFBSs on a genome-wide scale, as their CNSs cover 35% of TFBSs, when compared with a set of ChIP-seq data from 13 *A. thaliana* TF. Another study recovered a lower amount of TFBS (25%) when compared with ChIP-seq data, but did achieve higher specificity (Van de Velde et al., 2014). The results are depending on the transcription factor (TF) analyzed though; for instance they recovered 35% of the binding sites for the MADS-domain TF PISTILLATA, but only 8% of the APETALA3 binding sites (Van de Velde et al., 2014). These analyses indicate that TF binding is not determined by DNA sequence only. That TFBS selection is not only depending on sequence is also shown experimentally. Determination of TFBSs in different developmental stages of flower development showed differences in bound genes between developmental stages (Pajoro et al., 2014b). A large amount of TFBS divergence was also found in a comparison between closely related *Arabidopsis* species. The binding profiles of the MADS-domain TF SEPALLATA3 (SEP3), a major regulator of flowering, revealed a substantial number of species-specific binding events between *A. thaliana* and *A. lyrata*. These species-specific binding events could not be fully explained by divergence of DNA sequence (chapter 3) (Muiño et al., 2016).

TFs often do not act on their own, but instead are part of bigger protein complexes, which may influence the selection of their binding sites (Slattery et al., 2011; Smaczniak et al., 2012b; Bemer et al., 2017). Another factor influencing TF binding is the local organization of the DNA. DNA is tightly packed into chromatin, which may make some potential binding sites physically unavailable. Therefore, including data on the chromatin state in TFBS prediction models could improve the accuracy of these models. Chromatin state has been interrogated by mapping of different epigenetic modifications, such as DNA methylation and histone modifications (Roudier et al., 2011). Another method to discover regions of open chromatin is DNAse-seq (Hesselberth et al., 2009). This method relies on open chromatin being accessible for DNAse I digestion, and sequencing of small digested fragments leads to a chromatin accessibility profile.

We previously analyzed evolution of TFBSs between the two closely related *Arabidopsis* species *A. thaliana* and *A. lyrata,* which proved to differ substantially in TFBSs of SEP3 (Chapter 3) (Muiño et al., 2016). Although *A. thaliana* and *A. lyrata* diverged only 10 MYA, they do exhibit substantial differences in their genome. With 200 Mb the genome of *A. lyrata* is larger than the genome of *A. thaliana* (125 Mb) (Bennett et al., 2003; Hu et al., 2011). This is largely due to small deletions in non-coding sequences, although the two species also differ in number of genes, with *A. lyrata* possessing 20% more genes than *A. thaliana* (*A. lyrata*: 32,670; *A. thaliana*: 27,025) (Hu et al., 2011). The two species also show different genome organization, as is clear from the difference in chromosome numbers; *A. thaliana* has five chromosomes whereas *A. lyrata* has eight chromosomes, which is the ancestral state in Brassicaceae. These differences between the genomes of *A. thaliana* and *A. lyrata* are substantial, and may play a role in the divergence of SEP3 TFBS between these species.

A. thaliana is widespread throughout the northern hemisphere , and harbors a large amount of natural variation. Sequencing the genomes of 1,135 natural inbred lines revealed 10,707,430 biallelic SNPs and 1,424,879 small-scale indels (up to 40 bp), an average of one genome variant (SNPs or small indels) every 10 bp. This means that most genes contain at least one sequence variant, which likely changes protein function, with 17,692 having at least one high-impact variant. On average 440 genes per accession were predicted to harbor a sequence variant leading to inactivation of the gene. However, this is likely an overestimation, as it does not take into account compensatory mutations (Gan et al., 2011; Long et al., 2013). This analysis of 1,135 accessions showed that besides a group of accessions with extreme pair-wise divergences (termed relicts), the accessions are clustered broadly corresponding to geographical origin (Consortium et al., 2016).

To study SEP3 TFBS within a species, we focused on *A. thaliana* ecotypes. We chose to compare the widely used accession *Col-0* with the accession *Agu-1*. *Agu-1* was sampled from the Iberian Peninsula, a region in which *A. thaliana* accessions show a high percentage of region- and accession-specific SNPs (Cao et al., 2011). In contrast to *Col-0*, *Agu-1* needs a vernalization period before flowering. In addition, *Agu-1* has darker leaves and numerous axillary inflorescences. The flower and inflorescence phenotypes however, are indistinguishable from *Col-0*.

Here, we evaluated the extent of TFBS divergence by analyzing SEP3 binding profiles in two different *A. thaliana* ecotypes, *Col-0* and *Agu-1*. We find that, although there are differences in TFBS between *Col-0* and *Agu-1*, the TF binding profiles of these *A. thaliana* ecotypes strongly resemble each other. We also assessed whether DNA accessibility plays a role in TFBS selection by analyzing the open (active) regions of DNA during flower development. We find that the state of the chromatin may be associated with the divergence of TFBSs between *A. thaliana* and *A. lyrata*, but that it is negligible between the two ecotypes of *A. thaliana* (*Col-0* and *Agu-1*).

Results

Transcription factor binding sites of SEP3 in different *A. thaliana* ecotypes

We compared TFBS profiles of the major floral regulator SEPALLATA 3 (SEP3) between two different *A. thaliana* ecotypes by chromatin immunoprecipitation followed by high-throughput sequencing (ChIP-seq). ChIP-seq experiments were performed on inflorescence tissue with floral buds up to floral stage 10-11 of the two ecotypes *Col-0* and *Agu-1*. Experiments were performed in two biological replicates for each ecotype, and a mock-ChIP (pre-immune serum) dataset was generated as control. The biological replicates show good reproducibility, as shown by Spearman correlations of mapped reads (ρ=0.98 for *Col-0*, ρ=0.97 for *Agu-1*), as well as overlap in number of peaks between replicates (**Figure S1, S2).** Previously generated *A. lyrata* SEP3 ChIP-seq data (Muiño et al., 2016) were reanalyzed in the same way. This analysis shows that *A. lyrata* and *A. thaliana* differ substantially in SEP3 TFBS, whereas, as expected, the SEP3 binding profiles of *A. thaliana* ecotypes are more similar to each other (**Figure S2**). For the remainder of the analyses, we used the replicate with the highest statistical power for each sample (measured by the number of significant TFBS).

As threshold for each dataset we used a peak height where the overlap between replicates starts to plateau, while creating datasets with a similar number of TFBS (see **Figure S2**). Using this threshold, we obtain 5339 peaks for *Col-0*, and 5449 peaks for *Agu-1*. We observe around 70% common SEP3 binding sites between the two ecotypes, and a similar overlap in genes linked to these TFBSs (**Figure 1A, 1B**). The TFBSs show a similar distribution over genomic features for both ecotypes. Most TFBSs are positioned in the intergenic region (40-45%, half of which are located in the 1 kb upstream of genes), followed by the 5’ UTR and the coding sequence. In contrast, the SEP3 binding events in the *A. lyrata* genome are clearly located more distal from the genes; ~70% of TFBSs in *A. lyrata* are located in intergenic regions, and almost half of the total number of TFBSs are located more than 1 kb upstream of the closest gene (**Figure 1C**).

Although the overlap in SEP3 binding sites between ecotypes is substantially higher than between *A. thaliana* and *A. lyrata*, we still observe only about 70% overlap. However, just looking qualitatively at numbers seems to exaggerate the differences between the two ecotypes. We analyzed peak scores, and found that SEP3 peaks common for both ecotypes are on average higher than ecotype-specific peaks (**Figure 2A**), indicating stronger binding of the common binding sites. Read densities show the same pattern, with higher read density for peaks common between the ecotypes (**Figure 1D**). It must be noted that in ecotype-specific binding peaks an increase can be found in read density in the other ecotype. This indicates that at least some of the ecotype-specific peaks might be present in the other ecotype, but that these peaks fall below the significance threshold (**Figure 1D**). Together, these data indicate that the common binding sites between the ecotypes are stronger bound, and therefore more likely to result in changes in gene expression.

There is extensive sequence variation between *A. thaliana* accessions (Consortium et al., 2016). We therefore examined whether sequence variation might underlie differences in TFBSs, by analysis of the presence of SNPs and small insertions/deletions (indels). There are indeed less SNPs in the commonly SEP3 bound regions than in the ecotype-specific binding events (**Figure 2B**). However, the sequence underlying common TFBS also contains variation, and the difference in SNP density between common and ecotype-specific peaks, although significant, is small. When we analyzed whether there are differences in underlying CArG-boxes, the binding motif of MADS-domain TFs, we did not find any significant difference between the common and ecotype-specific peaks (**Figure S3**).

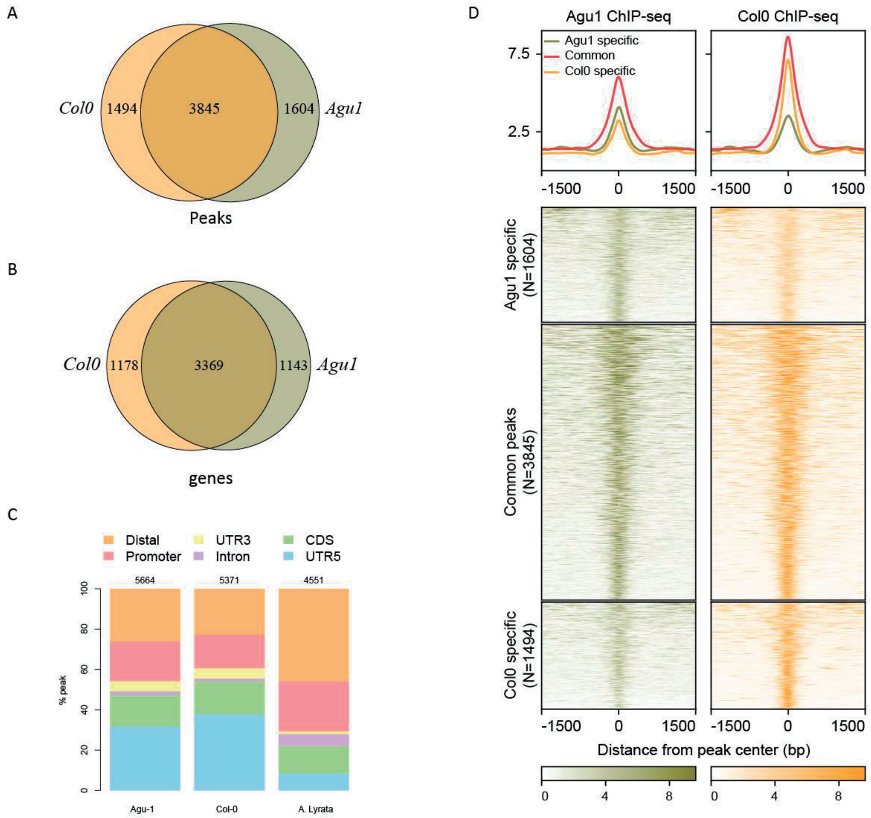


Figure 1: Comparison of ChIP-seq data of two A. thaliana ecotypes. A: *Overlap of ChIP-seq peaks in Col-0 and Agu-1.* ***B:*** *Overlap in potential SEP3 target genes between the Col-0 and Agu-1 ecotypes.* ***C:*** *Genomic distribution of SEP3 ChIP-seq peaks. Promoter is defined as the 1 kb upstream of the TSS. distal regions are all other intergenic regions* ***D:*** *Sequence read intensity around the peaks present in the merged dataset Col-0 and Agu-1 peaks.*

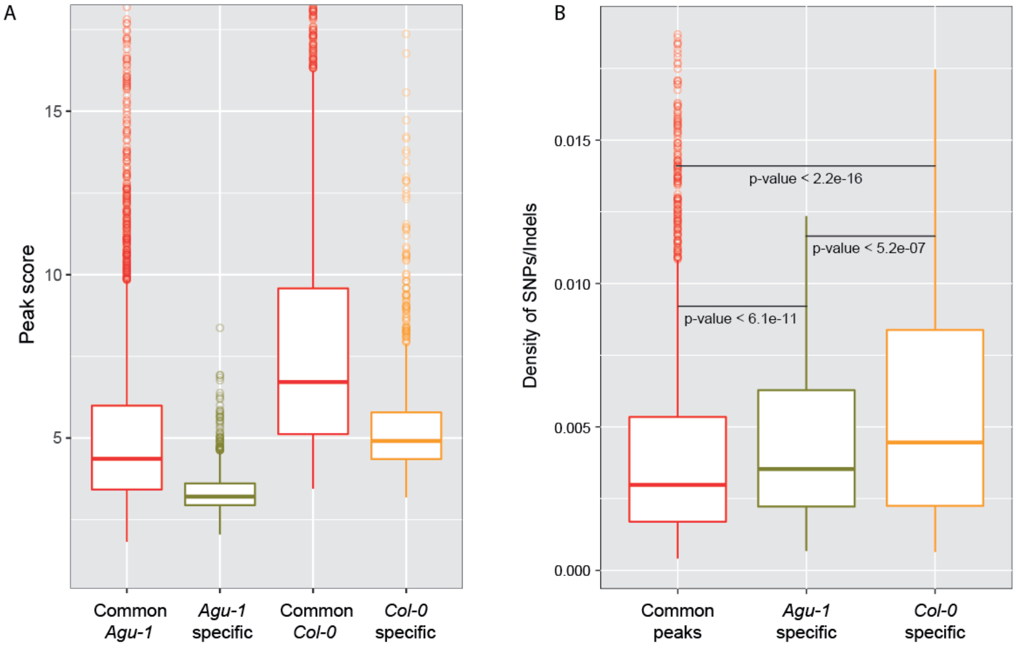


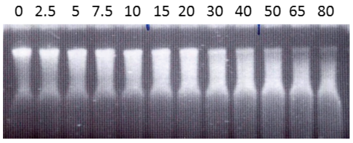
Figure 2: Peak score and density of SNPs in peak regions. A: *Peak height of the common peaks and the ecotype-specific peaks in each sample.* ***B:*** *Density of SNPs/indels under the ChIP peaks, shown as density of SNPs/indels per basepair.*

DNAse-seq assays

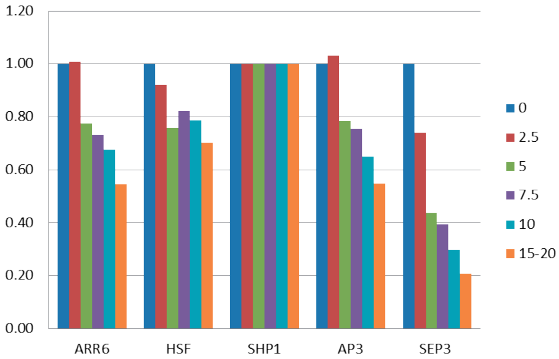
To analyze whether there are other factors that may influence TF binding, we assessed the accessibility of the chromatin (Pajoro et al., 2014b). To identify open regions of DNA, we performed DNAse-seq assays on inflorescences of the *A. thaliana* ecotypes *Col-0* and *Agu-1*, and of *A. lyrata*. Inflorescences with floral buds up to stage 10-11 were harvested, to obtain tissue similar to that used for the SEP3 ChIP-seq experiments. Nuclei extraction was performed to obtain the chromatin. Subsequently, this chromatin was digested with a range of DNAse concentrations, and then the sample with optimum digestion was selected. For optimum digestion, we aim for a smear of small DNA fragments, whereas the majority of the DNA is still intact (**Figure 3A**).

Besides this optical quantification of DNAse digestion, we also analyzed the degree of digestion by qPCR. Quantification of DNAse digestion by qPCR is performed using markers on regions of DNA known to be open during floral development (positive markers), and known closed regions of DNA (negative markers). The sample with the optimal amount of digestion will be the sample where the open regions are digested as much as possible (lower values in the qPCR compared to the undigested sample), whereas the closed regions are not digested yet (qPCR markers for closed regions remain similar to the value for the undigested sample). To normalize the data, we used one of the negative qPCR markers with a very small amplicon size compared to the other markers (**Figure 3B, C**). Using these quantification assays, we selected the optimal digestion for each sample. We performed the DNAse-seq assays in triplicate for *A. lyrata* and both *A. thaliana* ecotypes, *Col-0* and *Agu-1*. Purified DNA was digested as a control for sequence bias of the DNAse I enzyme. After library preparation, the fraction of the libraries with insert sizes between 50-150 bp was selected for high-throughput sequencing.

A



B



C

gene	Size (bp)
ARR6 (neg)	220
HSF (neg)	211
SHP1 (neg)	60
SEP3 (pos)	176
AP3 (pos)	138

Figure 3: Analysis of amount of digestion in our DNAse-seq assays. A: *Digestion of chromatin with different amounts of DNAse, ranging from 0 to 80 units.* ***B:*** *qPCR analysis on the differential digested samples. The different samples are digested with different amounts of DNAse units. For this sample, 5U of DNAse was selected to be sequenced, as this sample showed substantial digestion for one of the positive controls. The SHP1 qPCR was used to normalize the other values.* ***C:*** *The length of the amplicons obtained for each qPCR marker. SEP3 and AP3 amplification is used as positive controls for open chromatin regions.*

Sequencing data

High-throughput sequencing resulted in 4-24 million mapped reads per replicate (see **table S1** for a summary of the sequencing and mapping data). As there was good reproducibility between the replicates (see **Figure S4**), we pooled the replicates, resulting in 25-45 million reads per sample. We calculated the Signal Proportion Of Tags, or SPOT scores (Sullivan et al., 2015), for each replicate as well as the merged samples (**Figure 4**). SPOT scores are a measure of how many reads fall within DNAse hypersensitive sites (DHSs), and are therefore a measure of the signal to noise ratio. For our datasets, the SPOT scores are between 0.46 and 0.6. This is higher than some of the scores for existing datasets and within the same range of some other datasets (Sullivan et al., 2014; Sullivan et al., 2015). However, our control data, consisting of digested naked DNA, also show high SPOT scores (0.25-0.41). This is unexpected, as the assumption is that naked DNA is digested randomly by DNAse I and should have very low SPOT scores.

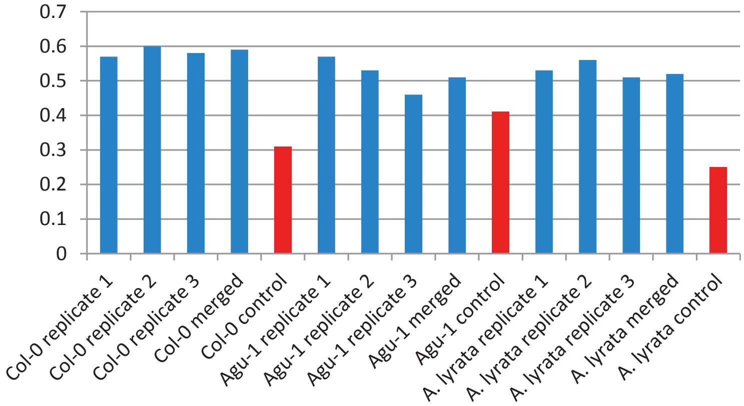


Figure 4: Signal-to-noise ratio in the different DNAse-seq samples. *SPOT scores for all our DNAse-seq samples. Col-0 and Agu-1 samples are mapped onto the Arabidopsis thaliana Col-0 genome (TAIR10); the A. lyrata samples are mapped onto the lyrata genome (JGI v1.0).*

To further assess the quality of our data, we examined the distribution of reads along the chromosomes. DHSs are expected to be correlated with open chromatin; therefore the centromers, being tightly packed regions, are expected to be devoid of DNAse-seq reads. This is indeed what was seen in previous data (Zhang et al., 2012b; Zhang et al., 2012a; Pajoro et al., 2014b; Cumbie et al., 2015) (**Figure S5**). In our dataset however, we do not see a clear dip in sequencing reads around the centromers (**Figure 5**). Although our *Agu-1* dataset shows a slight depletion in reads around the centromer, this pattern is not clearly present in our *Col-0* dataset (**Figure 5**). This analysis is not possible for our *A. lyrata* data because we cannot map reads to the centromeric region (**Figure S6**) due to the poor assembly of the *A. lyrata* genome in these regions.

This unusual distribution of reads along the chromosome led us to investigate the position of reads relative to genes.

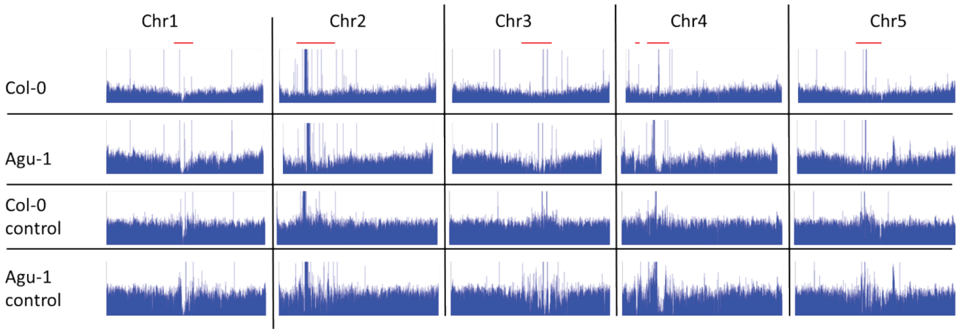


Figure 5: Chromosomal distribution of DNAse-seq reads (DHS). *Centromers are indicated with a red line (centromer position taken from Zhang et al.(Zhang et al., 2012a)).*

Strangely, the majority of the reads can be found in the gene body, increasing in intensity towards the 3’ end of the gene, and there is a depletion of reads at the Transcription Start Site (TSS) and the Transcription End Site (TES) (**Figure 6A, B, C**). This distribution is different from previously generated data in our lab, where the majority of the signal is at the TSS and the TES (**Figure 6D**). Other published datasets also show most of the signal located around the TSS (Zhang et al., 2012a; Sullivan et al., 2014; Cumbie et al., 2015). Remarkably, the same pattern we observe for our samples is also present in the control datasets (digested naked DNA), where we expect an even distribution over the whole genic region. This indicates that our control samples may not have been completely devoid of proteins.

Next, to take a closer look at our data, we inspected the regions around some selected genes (**Figure 7**) and compared our data with other datasets of DHSs in inflorescences and floral developmental stage-specific tissues (Zhang et al., 2012a; Pajoro et al., 2014b). This analysis indicates that for some genes, the DHS landscape of our data (blue) is similar to the published datasets (in green and red)(see in **Figure 7A, B**). However, for other genes (see **Figure 7C, D**), our dataset appears very different from the published data. Taken together, the distributions of reads, as seen at the chromosomal and gene level, raise doubts about the quality of our DNAse-seq assays.

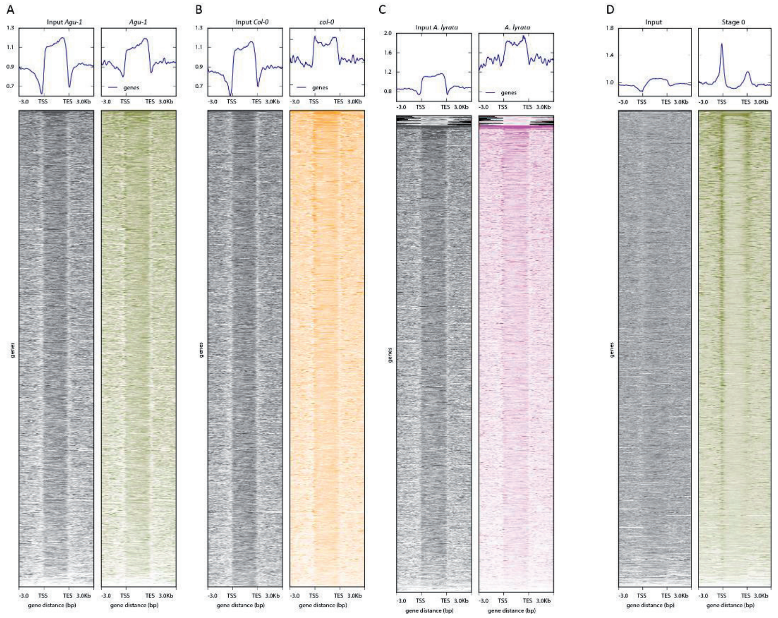


Figure 6: Gene centered distribution of reads. *It can be seen that our data and our control data are very similar. Left is control, right is sample.* ***A:*** *Agu-1.* ***B:*** *Col-0.* ***C:*** *A. lyrata.* ***D:*** *Data from Pajoro et al. 2014 (Pajoro et al., 2014b), showing very little enrichment in the control sample. TSS=transcription start site; TES=transcription end site.*

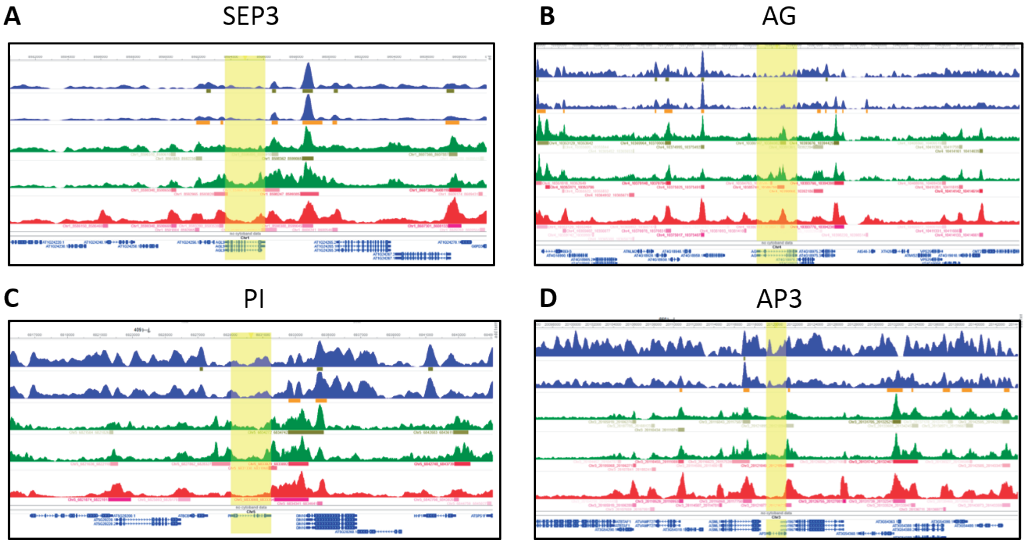


Figure 7: Examples of DHS landscapes. *Shown are the obtained reads from the different samples (not normalized with the control) From top to bottom: our Agu-1 inflorescence dataset (blue), our Col-0 inflorescence dataset (blue), different stages of development (day 0 and day 8) of Pajoro et al. (green)(Pajoro et al., 2014b) and the A. thaliana Col-0 whole inflorescence dataset of Zhang et al. (red) (Zhang et al., 2012a). All datasets were re-analyzed in the same way. The gene itself is highlighted with yellow.* ***A:*** *SEPALLATA3* ***B:*** *AGAMOUS* ***C:*** *PISTILLATA* ***D:*** *APETALA3.*

Regions of open chromatin

Next, we performed peak calling on our DNAse-seq samples to obtain DNAse Hypersensitive sites (DHSs). As we had doubts about the quality of our control samples, peak calling was performed using two different methods: HOTSPOT without using the control dataset, and MACS2 using the control. Peak calling was done for each of the replicates, as well as the merged sample of all three replicates. The number of peaks found with each method are listed in **Table S2**.The final datasets consisted of peaks that were present in the merged dataset as well as in one of the replicates. Interestingly, MACS2 and HOTSPOT gave vastly different numbers of peaks for each of the samples. Hotspot (without the control) resulted in the most peaks for our *Agu-1* sample, whereas for *Col-0* and *A. lyrata* substantially more peaks were found with MACS (with control). We therefore decided to combine the datasets found with each method to obtain the final set of peaks for each sample, see **Table 1**.

We obtained 12987 significant DHSs for our *Col-0* dataset, whereas our datasets for *Agu-1* and *A. lyrata* consisted of 9270 and 10129 DHSs, respectively. We also linked each DHS with the gene positioned closest to it for further analyses (**Table 1**).

Table 1: Final peak numbers and associated genes. Each peak 3kb upstream to 1kb downstream of a gene is linked to that gene. Peaks were called with an FDR of 0.01 using MACS and HOTSPOT.

	Number of DHSs	DHS associated genes
A. thaliana Col-0	12973	11864
A. thaliana Agu-1	9270	7594
A. lyrata	10129	9205

Even though the overall read density shows a depletion of signal around the TSS (**Figure 6**), when we analyzed the distribution of the significant DHSs relative to genes, they are positioned close to the TSS, where DHSs are expected (**Figure 8A**). This distribution does not depend on which peak-calling method was used, as both methods gave similar results (**Figure S7**). This indicates that although the raw data (mapped reads) show an unexpected distribution, the final distribution of our DHSs seems to be similar to published datasets and to what is expected for DHSs. Taking a more detailed look, we found that the majority of peaks in all three samples are positioned in the 1 kb upstream promoter region, with 68-78% found in this region (**Figure 8B**). However, there are differences between the three samples. The distribution of DHS among different genomic components in our *Col-0* data is very similar to previously published data, even though those data originate from different types of tissue (see **Figure S8**)(Pajoro et al., 2014b). However, the DHSs from both *Agu-1* and *A. lyrata* show a different distribution. For the *A. lyrata* sample, we find DHSs on average further away from the TSS than in the *A. thaliana* samples. This is also seen by the larger number of DHSs in more distal regions of the genome. As *A. lyrata* has a larger genome with larger intergenic regions, this was expected. Surprisingly however, *Agu-1* also seems to have less DHSs in the 1 kb upstream promoter of the genes compared to *Col-0*, with 68% of the *Agu-1* peaks found in this region compared with 78% of the *Col-0* DHS. Instead, a larger fraction of DHSs in *Agu-1* are found in the region 1-3kb upstream of the gene. This is unexpected as ecotypes are supposed to be very similar in genome size (Johnston et al., 2005). That changes in the genome sequence underlie these differences is unlikely, because both *Agu-1* and *Col-0* DNAse-seq samples were mapped to the same *A. thaliana Col-0* reference genome. Possibly, differences in quality of the datasets may cause at least some of these differences.

Correlation of DHSs with gene expression data

DHSs are linked to transcriptional activation, and are therefore expected to correlate with expression levels. This is indeed what we saw in data previously generated in our lab (unpublished data (Pajoro et al., 2014b)), as well as what was seen in a published study performed in rice (Zhang et al., 2012b). Therefore, we calculated the correlation of previously obtained RNA expression data (of *A. thaliana Col-0* and *A. lyrata* (Muiño et al., 2016)) with our DHS datasets as a quality control. This analysis shows moderate correlation between our DNAse-seq datasets and expression levels (**Figure 9A, B, D**), with higher correlations for *Col-0* and *A. lyrata* (Pearson's r of 0.67 and 0.5, respectively) than for *Agu*-1 (Pearson's r=0.26).

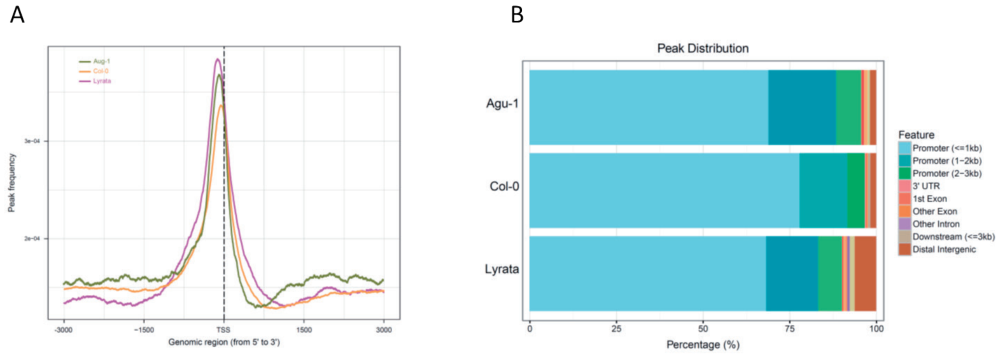


Figure 8: Genomic distribution of our DNAse-seq peaks. A: *Position of DHS peaks compared to genes. Yellow=Col-0; Green=Agu-1; purple=A.lyrata* ***B:*** *Genomic distribution of DHS sites across different genomic features.*

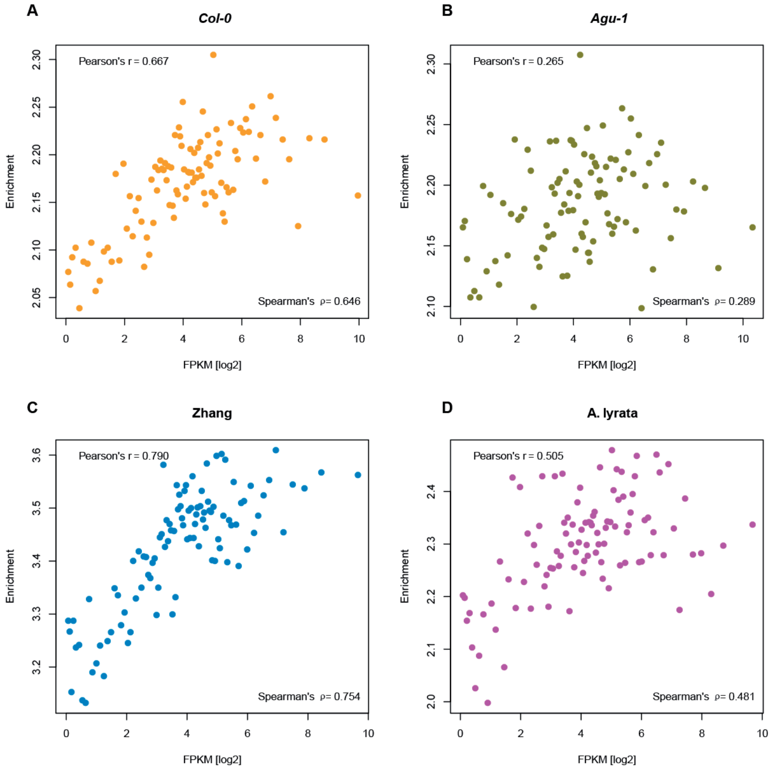


Figure 9: Correlation of DHSs with RNA expression levels*. Scatter plot comparing gene expression and chromatin accessibility. Genes were binned into percentiles (N=100) based on their expression level, and the mean gene expression level (x-axis) and mean DHS peak score (y-axis) of each bin were plotted.* ***A:*** *Col-0;* ***B:*** *Agu-1;* ***C.*** *published data from Zhang et al;* ***D:*** *A. lyrata with A. lyrata expression data. In A, B and C, the DHSs were correlated with expression data from Col-0.*

The differences in correlations most likely indicate differences in data quality. However, it needs to be noted that both *Agu-1* and *Col-0* DHSs are compared with expression data from *Col-0*, neglecting possible differences in gene expression between these ecotypes. As a comparison, we performed the same analysis for a published *A. thaliana* inflorescence dataset (Zhang et al., 2012a) with our RNA-seq data. This analysis gave a correlation between DHSs and expression of Pearson's r= 0.79 (**Figure 9C**). This is higher than the correlations we obtained, indicating that our DHS data might not be of optimal quality.

Overlap with published data

As a last quality control, we compared our DHS peak dataset for *Col-0* with published datasets. We find that between 70-86% of our *Col-0* peaks are also present in those published datasets (**Table 2**). Unsurprisingly, the overlap with these published *Col-0* datasets is lower for *Agu-1*, being only 59-64% (**Table S3**). Those percentages are in a similar range as published comparisons of DNAse-seq data, which found that 70–74 % overlap was found between DNAse-seq datasets of *A. thaliana* inflorescences in different labs (Zhang et al., 2012a; Cumbie et al., 2015). However, that study required 80% of a peak to overlap with a peak from the other dataset, whereas our analysis required only 1 bp. Another consideration is that our dataset contains substantially less DHSs than the datasets we are using as comparison. Indeed, the percentage of overlap between our dataset and the published dataset seems to be mainly dependent on the amount of DHSs in the other dataset. Although the sample that should be most similar to our data (inflorescence (Zhang)) shows the highest percentage of overlap with our data, this dataset also contains the highest number of DHSs (**Table 2**).

In conclusion, it seems that our datasets (in particular the *Agu-1* and *A. lyrata* datasets) are of less quality than the published DHS datasets for *A. thaliana Col-0*. Nevertheless our datasets allow a comparison of DHSs between different ecotypes and different species.

Table 2: Comparison between the *Col-0* DHS data generated here with published DNAse-seq data (Zhang et al., 2012a; Pajoro et al., 2014b). Our *Col-0* dataset contains 12973 DHSs.

dataset	total peaks in dataset	number of peaks of dataset present in our *col-0* data	number of peaks of our *col-0* data present in this dataset	% peaks of our *col-0* data present in this dataset
leaf (Zhang)	20226	10506	10267	79.1%
inflorescence (Zhang)	23715	11413	11133	85.8%
inflorescence day 0 (Pajoro)	19054	9922	9858	76.0%
inflorescence day 2 (Pajoro)	15646	9085	8978	69.2%
inflorescence day 4 (Pajoro)	16334	9501	9362	72.2%
inflorescence day 8 (Pajoro)	19352	9815	9813	75.6%

Comparison of DHSs between ecotypes and species

A pairwise comparison of DHS profiles shows more overlap between the two *A. thaliana* ecotypes than between the *A. thaliana* ecotypes and *A. lyrata* (**Figure 10A**). This pattern is also visible for the genes that are linked to these DHS (**Figure 10B**). A comparison between all three samples indeed shows that the overlap between the *A. thaliana* ecotypes is larger than between either *A. thaliana* ecotype and *A. lyrata* (**Figure 10C, D**). This is expected, as ecotypes should be more similar to each other than to a different species. Interestingly, even though the overlap between ecotypes is larger than the overlap between *A. thaliana* and *A. lyrata*, there are some DHSs that are shared between *A. lyrata* and only one of the *A. thaliana* ecotypes. This overlap is larger between *A. lyrata* and *Col-0* than between *A. lyrata* and *Agu-1*. As our *Col-0* dataset has 40% more DHSs than our *Agu-1* dataset, this is partially due to the difference in size of our datasets. Differences in data quality between the samples may also cause some of the differences. There is also a significant portion of sample-specific DHSs, varying from 21% (*Agu-1*) to 54% (*A. lyrata*).

Next, we analyzed whether the differences in chromatin accessibility that we observed between ecotypes can be linked to divergence in underlying DNA sequence. We first used published PhastCons scores that were generated using nine Brassicaceae species (Haudry et al., 2013). PhastCons scores are a measure of DNA sequence conservation, which range between 0 (no conservation) and 1 (complete conservation). Unexpectedly, the *Agu-1* specific DHSs have higher PhastCons scores than DHSs that are common between both ecotypes (**Figure 11A**). When we look at the distribution of PhastCons scores associated with our DHS data we see that, although the *Agu-1* scores are significantly higher, they do show a large range of values (**Figure 11B**). We assessed sequence divergence with a second method, using *A. thaliana* SNP data (Consortium et al., 2016). Only SNPs/indels between *Col-0* and *Agu-1* were considered in this analysis. Analysis of the density of SNPs under the DHSs shows no significant difference in SNP density between the common DHSs and *Agu-1* specific sites. In contrast, the DHSs specific to *Col-0* are higher in SNP density (**Figure 11C**). As species-specific DHS would be expected to show higher sequence variation, it is surprising that the *Agu-1* DHSs do not show a higher SNP density than the common DHSs. Two factors may contribute to this surprising result. First, we seem to have differences in data quality between our datasets. A second factor to consider is the method of data analysis. Both *Col-0* and *Agu-1* samples were mapped onto the *Col-0* reference genome. Although two mismatches were allowed, it may still be that we missed some *Agu-1* DHSs that are located in regions with very diverged sequences.

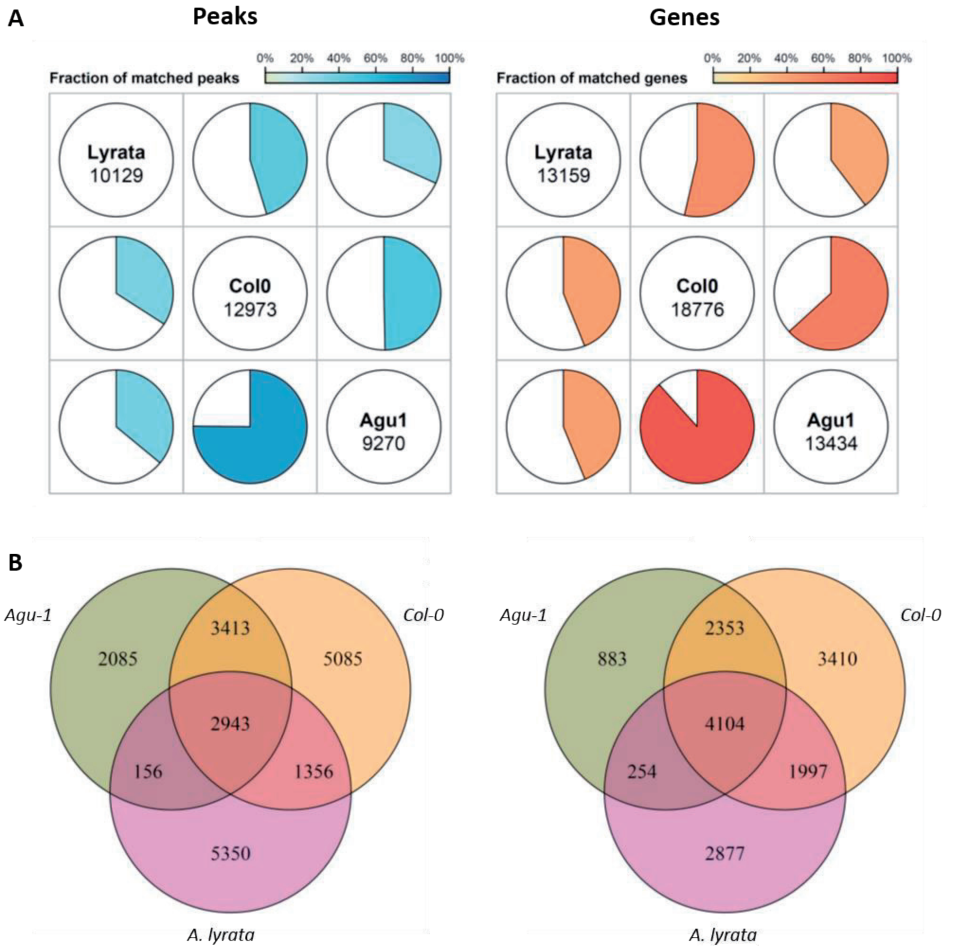


***Figure 10: The fraction of conserved peaks between the different samples**. Pairwise comparisons between DHSs (**A**) and linked genes (**B**); the top row is compared to the A. lyrata dataset, the middle row is against the Col-0 dataset and the bottom row is against the Agu-1 dataset. **C:** Venn-diagram of overlap in peaks of all thee DHS datasets. Overlapping peaks are defined as having at least one bp in common. **D:** Venn-diagram of overlap in DHS-linked genes between our three samples.*

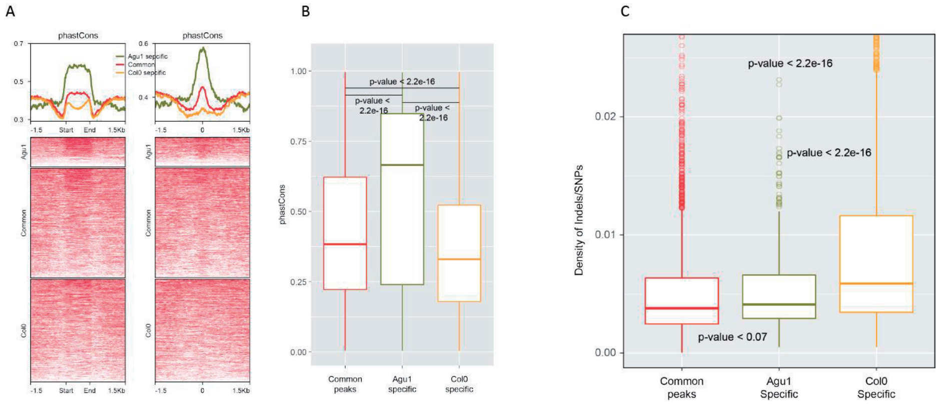


Figure 11: Sequence conservation of DHSs. ***A:*** *Average PhastCons scores of DHS peaks. On the left, "start" and "end" indicate the boundaries of the DHS peak. On the right, 1.5 kb on both sides of the peak center (0) are shown.* ***B:*** *Distribution of PhastCons scores for the Agu-1 specific, the Col-0 specific and the common DHSs.* ***C:*** *Density of SNPs/indels under the DHS peaks.*

Finally, to examine whether our observed differences in SEP3 binding sites could be associated with differences in DNA accessibility, we correlated the SEP3 ChIP-seq data with the DHS data. This analysis shows which percentage of the ChIP-seq peaks falls within a DHS (**Table 3**), and reveals that there is a correlation between SEP3 binding sites and DHSs, as the overlap of *A. lyrata* SEP3 BS is higher with the *A. lyrata* DHSs then with the *A. thaliana* DHSs. Similarly, the *A. thaliana* SEP3 binding sites overlap better with the *A. thaliana* DHSs than with the *A. lyrata* DHSs. Interestingly, this analysis does not show any significant differences between *Agu-1* and *Col-0:* SEP3 binding sites of either ecotype show a very similar overlap with each DHS dataset. The overlap between SEP3 binding and our *Col-0* DHS dataset is significantly higher than with our *Agu-1* DHS dataset. However, this difference can be explained by the different size of the DHS datasets (our *Col-0* DHS dataset contains ~40% more DHSs than our *Agu-1* DHS dataset), as well as differences in quality between our DNAse-seq datasets. Remarkably, a significant fraction of SEP3 binding sites does not overlap with any DHS. This is not unexpected, as MADS-domain TF binding outside of DHS has been reported before (Pajoro et al., 2014b). It is unclear whether MADS-domain TFs do not need open chromatin to bind, or whether this can be explained by a dilution of signal as a result of using samples consisting of mixed tissues.

Table 3: Correlation of DHS data with SEP3 binding sites. Shown is the percentage of SEP3 binding sites that overlaps with a DHS. For the *A. lyrata* data, the peaks were converted to *A. thaliana* genome coordinates. DHS and ChIP peaks are defined as overlapping when they share at least 1 bp.

	Agu-1 DHS	Col-0 DHS	*A. lyrata* DHS
Col-0 ChIP	39.19%	62.06%	29.77%
Agu-1 ChIP	41.74%	62.27%	30.86%
A. lyrata ChIP	16.01%	28.28%	42.81%

Discussion

In this work, we analyzed whether there is variation in binding sites of SEP3, a regulator of flower development, between *A. thaliana* ecotypes. We found a substantial level of overlap in SEP3 binding sites between the *A. thaliana* ecotypes *Agu-1* and *Col-0*, which share 70% of their SEP3 TFBS. However, there are indications that this may be an underestimation of the similarity of the SEP3 binding profiles between these ecotypes. We found that SEP3 binding sites that are common to both ecotypes are on average higher than TFBSs specific for one of the ecotypes. In addition, when we analyzed the read density under the ecotype-specific bound regions, we found indications that a large fraction of these TFBSs might be present in the other ecotype as well, but remains below the threshold. We show that SEP3 TFBSs that are conserved between ecotypes are slightly more conserved in underlying DNA sequence, which was expected. Previously, we performed a comparison of SEP3 TFBS profiles between *A. thaliana* (*Col-0*) and *A. lyrata*, which showed substantial TFBS divergence (Chapter 3)(Muiño et al., 2016). Here, our analyses show that SEP3 TFBS profiles are substantially more conserved between *A. thaliana* ecotypes than between *A. thaliana* and *A. lyrata*.

Our comparison between *A. thaliana* ecotypes, as well as a previous study between *A. thaliana* and *A. lyrata* show that some variation in SEP3 TFBSs can be explained by divergence of the underlying DNA sequence (Muiño et al., 2016). To assess other factors that might underlie TFBS evolution, as well as to gain more insight into the evolution of gene regulation at different evolutionary timescales, we analyzed chromatin accessibility. Open regions of chromatin during floral development were examined by DNAse-seq assays in the *A. thaliana* ecotypes *Agu-1* and *Col-0*, as well as in *A. lyrata*. When we inspected the overlap of DHSs with SEP3 bound genomic regions, we found that there is better overlap of the SEP3 binding pattern with the DHS data generated in the same species, than with the DHS dataset of the other species. For instance, the DHS dataset from *A. lyrata* shows a higher overlap with the SEP3 TFBS profile found in *A. lyrata* than the SEP3 TFBS profile of *A. thaliana*. Interestingly, we do not see this pattern for the *A. thaliana* ecotypes, indicating that chromatin accessibility does not have a strong impact on SEP3 TFBS divergence between ecotypes.

Further analysis of the DHS profiles showed that, as expected, *A. thaliana* ecotypes resemble each other more closely than *A. thaliana* and *A. lyrata* do. Interestingly however, we also find overlap between *A. lyrata* and a single *A. thaliana* ecotype, as well as a substantial amount of lineage-specific DHSs, even within *A. thaliana*. We investigated whether sequence divergence could contribute to this DHS divergence between ecotypes. Unexpectedly, the common DHS were not more conserved in sequence than the ecotype-specific DHSs; instead, *Agu-1* specific DHSs are as conserved (SNPs) or even more conserved (PhastCons scores) in sequence than the DHSs common to both ecotypes. However, these conclusions have to be treated with caution, as there are some concerns about the data quality of our DHS datasets. First, our control sample of digested purified DNA does not show the expected random digestion pattern, but instead resembles our samples in distribution of sequencing reads. Even though the controls showed this unexpected pattern, they were still used to normalize the data in one of the peak-calling methods (MACS2). This means that the control samples did influence the final results. To obtain our final DHS datasets, we used two methods of peak-calling, and merged the obtained datasets. This unusual strategy of merging the datasets obtained with different peak-calling methods was performed to obtain datasets with comparable numbers of DHSs. This was necessary as the used peak-calling methods gave vastly different numbers of DHSs for our different samples, with HOTSPOT (without the control) obtaining more DHSs for *Agu-1*, and MACS2 (using the control) generating larger numbers of DHSs for *Col-0* and *A. lyrata*. Our final datasets show the expected genomic distribution, a substantial correlation with expression data and 70-85% overlap with published DHS datasets. However, the issues with our raw data, as well as our unusual peak-calling strategy call for caution in interpreting our obtained DHS profiles.

One explanation for our unexpected distribution of read density could be that, although we carefully assessed the amount of digestion in each sample, we did not obtain the most optimal level of DNAse digestion to detect open chromatin regions. Our protocol for DNAse treatment involves digestion of the DNA, followed by selection of small fragments, library preparation, and sequencing. In this protocol, the right amount of digestion is particularly important; if over-digested, the most open regions of the chromatin will be completely digested (and therefore not sequenced). Instead, you might get peaks in less open regions or no peaks at all. Under-digestion will lead to a large portion of open chromatin regions remaining intact or poorly digested. These fragments will therefore be too long, and will be selected against in the size-selection step, resulting in less signal from these regions (see **Figure 12** for a visual representation). The fragment size-selection step is another critical component of our protocol. We aimed for open regions of DNA, devoid of nucleosomes. To ensure we would not sequence DNA fragments that were occupied by nucleosomes, we selected for insert-sizes between 50-150 bp. However, it may be that either our DNAse digestion or our selected fragment size was not optimal, and therefore we did not obtain optimal data quality.

To circumvent the problems of determining optimal digestion and size-selection, a different protocol may be used. In this protocol, after DNA-digestion, a biotin adapter is ligated to all obtained DNA fragments. Subsequently, the DNA is cut with MmeI, for which the restriction recognition site is present in the primer, and which cuts several base pairs into the DNA fragment. This MmeI digestion is followed by the ligation of the second adapter (Song and Crawford, 2010; Zhang et al., 2012a; Cumbie et al., 2015). This method assessed chromatin accessibility by a single DNAse cut, instead of requiring two cuts within a certain distance, like our protocol. Although this might lead to some false positives due to random shearing of the DNA, this method is more robust to variation in level of digestion. An additional advantage is that this method circumvents the size-selection step, eliminating one parameter influencing quality of DNAse-seq data. Another method to analyze accessibility of the DNA is ATAC-seq (Assay for Transposase Accessible Chromatin using sequencing). ATAC-seq uses Tn5 transposase to cut the DNA and simultaneously ligate sequencing adapters to the obtained

DNA fragments. Because Tn5 will preferably target open regions of DNA, sequencing the obtained fragments gives the DNA accessibility profile (Buenrostro et al., 2013).

We observed a high level of TFBS conservation between two ecotypes of *A. thaliana* for the developmental regulator SEP3. However, not all TFBSs are conserved between these ecotypes. It would be interesting to elucidate which genes contain an ecotype-specific SEP3 BS in their promoter, and whether these binding sites cause a difference in expression level of these genes. For this analysis expression data from *Agu-1* would be needed. SEP3 binding profiles and expression data from other *A. thaliana* ecotypes would also help to elucidate how much ecotype-specific binding there is. Whereas we observe a large overlap between SEP3 binding in *A. thaliana* ecotypes, SEP3 binding sites are quite diverged between *A. thaliana* and *A. lyrata* (Muiño et al., 2016). It would be interesting to analyze whether this large divergence is due to the many genome rearrangements *A. thaliana* went through after diverging from *A. lyrata*. *A. thaliana* and *A. lyrata* have different genome sizes (125 vs 200 Mb) and a different number of chromosomes (N=5 vs N=8) (Bennett et al., 2003; Hu et al., 2011). In contrast, there are other Brassicaceae species that, although more distant to *A. lyrata*, share a more similar genome. Examples are *Capsella rubella* and *Cardamine hirsuta.* Both of these species share a very similar genome structure with *A. lyrata* (both N=8; genome size of 220 and 225 Mb respectively) (Johnston et al., 2005; Slotte et al., 2013). To examine whether the differences in genome between *A. thaliana* and *A. lyrata* are partially responsible for the large divergence in SEP3 binding, it would be interesting to compare the *A. lyrata* SEP3 binding profile with SEP3 binding profiles in these Brassicaceae species.

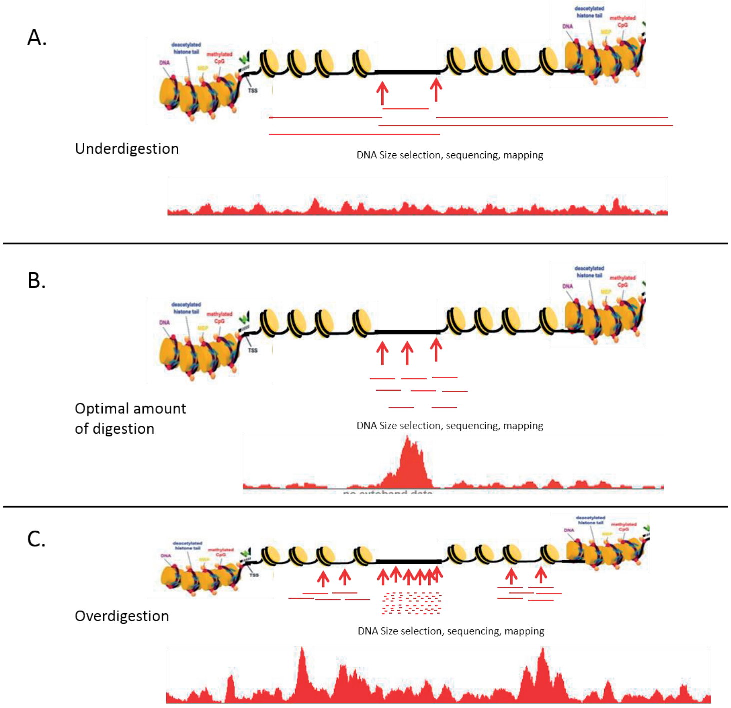

A.
Underdigestion
deacetylated histone tail
methylated CpG
DNA
TSS
DNA Size selection, sequencing, mapping
B.
Optimal amount of digestion
DNA Size selection, sequencing, mapping
C.
Overdigestion
DNA Size selection, sequencing, mapping

Figure 12: The amount of digestion is critical for the success of the DNAse-seq assay. A: *Under-digestion will lead to a low number of peaks, as many open chromatin regions will be poorly digested, and therefore not sequenced.* ***B:*** *The optimal level of digestion will release small-size DNA fragments from open chromatin regions, while leaving the surroundings sequences intact.* ***C:*** *Over-digestion will digest the open regions into fragment sizes smaller than our selected size-range. Regions of chromatin that are slightly less open might be digested to fall in our selected fragment size. This leads to a different pattern of peaks, where the most open regions of chromatin might not be detected anymore.*

Materials and methods

Plant material

Col-0 for inflorescences were grown under standard long-day conditions on soil (16h/8h) at 20 °C day/16 °C night. *Agu-1* (Aguaron-1 (N76409)) and *A. lyrata* plants were grown under long day (16h/8h) at 20 °C day/16 °C night, vernalized for 8 weeks (4 degree during the night, 8 during the day, 12/12h), afterwards back to long days and 20 °C day/16 °C night.

ChIP-seq experiments

Generation of SEP3 ChIP samples of *A. thaliana Col-0* and *Agu-1* inflorescences was performed essentially as described previously (Kaufmann et al., 2010b), with one modification. Instead of fixing the tissue immediately following harvesting, material was frozen immediately and fixed after grinding of the tissue. Fixation was performed using 20ml MC buffer including 0.5% formaldehyde. Material was incubated with fixative for 5 min on ice. Fixation was stopped 2.5M glycine before continuing with the nuclei isolation. Libraries were prepared with the Rubicon Genomics Thruplex® DNA-seq kit, following the manufacturer's instruction. Libraries were sequenced on the Illumina HiSeq2000.

DNAse-seq experiments

DNAse treatment was done as in (Pajoro et al., 2014b). 0.6 g of tissue was used for *A. thaliana Col-0*, *Agu-1* and *A. lyrata* inflorescences. Three biological replicates were done for each ecotype/species. As control, gDNA was treated with DNAseI in the same way as the other samples (1X control per species). 50-300 bp fragments were selected from agarose gel, and this fraction was used for library preparation. Libraries were made with the Rubicon Genomics Thruplex® DNA-seq kit, following the manufacturer's instruction. Following library preparation, the samples were size-selected (200-300 bp) from gel and purified using the Qiagen Minelute gel extraction kit. Libraries were sequenced on the Illlumina HiSeq2000.

ChIP-seq data analysis

We followed the ChIP-seq data analysis guidelines (Landt et al., 2012; Bailey et al., 2013) recommended by the ENCODE project and have developed an analysis pipeline consisting of quality control, read mapping, peak calling, assessment of reproducibility among biological replicates, and peak annotation to reprocess all raw data in a standardized and uniform manner (Chen and Kaufmann, 2017). Specially, the quality of the raw data (FASTQ files) was evaluated with the FastQC program (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads were then mapped to the *A.thaliana* genome (TAIR10) using Bowtie (version 1.1.2) with parameters "--threads 8 -n 2 -m 10 -k 1 --best --chunkmbs 256 -q". Redundant reads were removed using Picard tools (v2.60; http://broadinstitute.github.io/picard/). Peak-calling was performed using MACS2 (version 2.1.0). Duplicated reads were not considered (--keep-dup=1) during peak calling in order to achieve a better specificity (Bailey et al., 2013). The "--mfold" parameter was set as "2-20" to build the model. A relaxed threshold of p-value (p-value ≤ 1e-2) was suggested in order to enable the correct computation of IDR (irreproducible discovery rate) values (Landt et al., 2012). Following the recommendations for the analysis of self-consistency and reproducibility between replicates (https://sites.google.com/site/anshulkundaje/projects/idr, (Li et al., 2011)). Peaks across replicates with an IDR ≤ 0.1 were retained. ChIP-seq data were visualized in the WashU Epigenome Browser (Zhou et al., 2011).

Wiggle tracks were generated using deepTools (Ramirez et al. 2014); read coverage was normalized to 1x sequencing depth (also known as reads per genomic content, RPGC).

DNase-seq data analysis

Potential peaks (called DNase I hypersensitive sites, DHSs) were called using Hotspot without providing an input file, and MACS2 for which an input file was provided. For MACS2, the parameters “--nomodel --shift -100 --extsize 200” were used. Data obtained with MACS2 and Hotspot were merged. Peak calling was performed for each of the replicates, as well as the merged sample of all three replicates. The final dataset consisted of peaks that were present in the merged dataset as well as in one of the replicates.

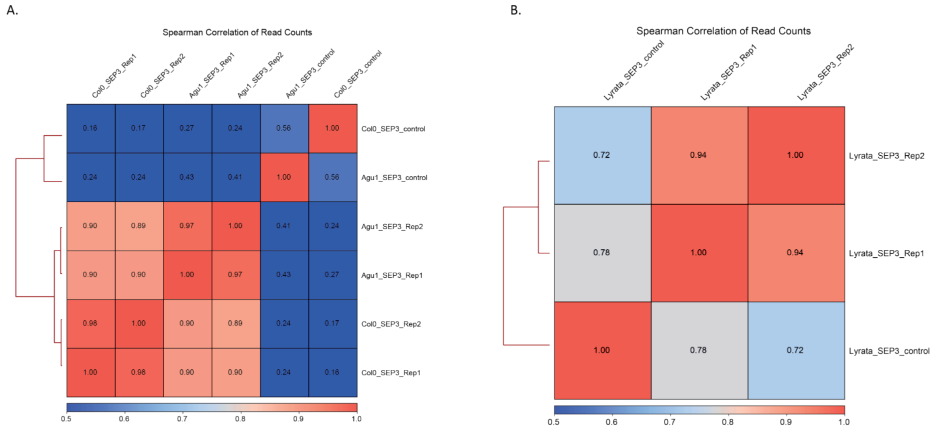
Peak analysis

All the peak-based analyses (including peak overlapping, merging and summary) were performed using BEDTools (v2.25.0) (Quinlan and Hall 2010).

Statistical analysis and data visualization

If not specified, all statistical analyses and data visualization were done in R.

Supplementary material



A.

	Col0_SEP3_Rep1	Col0_SEP3_Rep2	Agu1_SEP3_Rep1	Agu1_SEP3_Rep2	Agu1_SEP3_control	Col0_SEP3_control
Col0_SEP3_control	0.16	0.17	0.27	0.24	0.56	1.00
Agu1_SEP3_control	0.24	0.24	0.43	0.41	1.00	0.56
Agu1_SEP3_Rep2	0.90	0.89	0.97	1.00	0.41	0.24
Agu1_SEP3_Rep1	0.90	0.90	1.00	0.97	0.43	0.27
Col0_SEP3_Rep2	0.98	1.00	0.90	0.89	0.24	0.17
Col0_SEP3_Rep1	1.00	0.98	0.90	0.90	0.24	0.16

B.

	Lyrata_SEP3_control	Lyrata_SEP3_Rep1	Lyrata_SEP3_Rep2
Lyrata_SEP3_Rep2	0.72	0.94	1.00
Lyrata_SEP3_Rep1	0.78	1.00	0.94
Lyrata_SEP3_control	1.00	0.78	0.72

Figure S1: Reproducibility of read counts for our SEP3 ChIP-seq experiment. *Reproducibility is calculated using Spearman correlations for non-overlapping windows of 10 kb.* ***A:*** *Our newly generated Col-0 and Agu-1 SEP3 ChIP-seq data.* ***B:*** *Newly analyzed A. lyrata ChIP-seq data from (Muiño et al., 2016).*

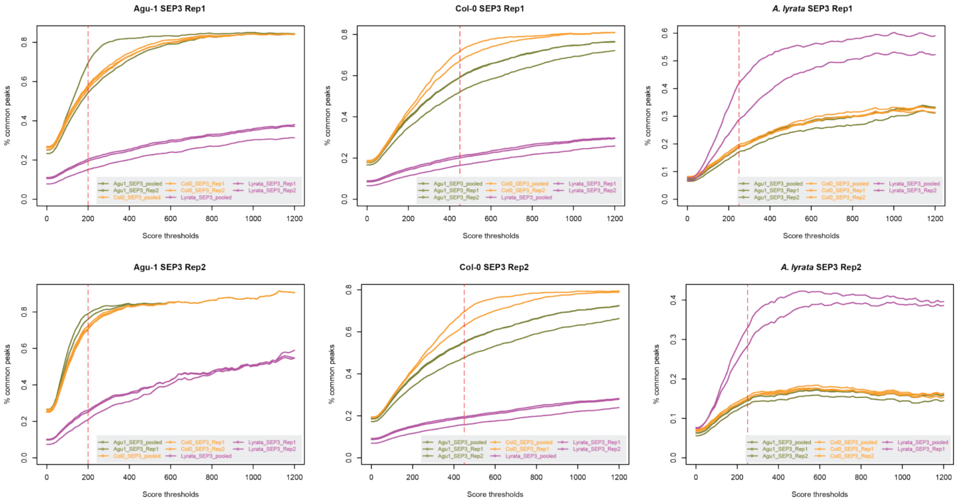


Figure S2: *Overlap in number of peaks depending on peak height.*

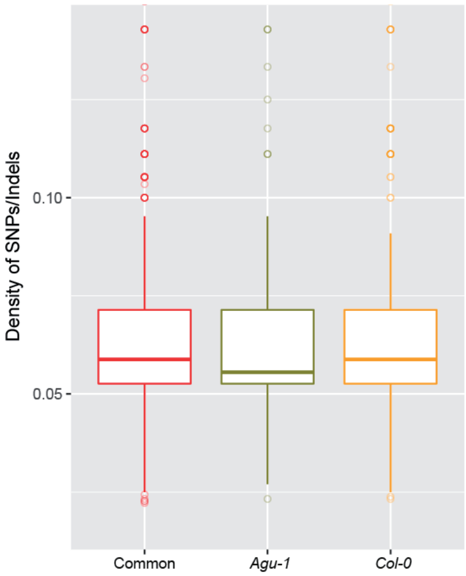


Figure S3: SNPs/indels in CArG-boxes underlying our SEP3 ChIP-seq binding sites.

Table S1: Sequencing/mapping of DNAse-seq data. Data are shown for each replicate, as well as the merged sample.

	Sample	Raw	Mapped/used	% in use
Arabidopsis thaliana **(TAIR 10)**	**Col-0 replicate 1**	20,534,172	5,845,703	28.47%
	Col-0 replicate 2	28,475,081	14,407,135	50.60%
	Col-0 replicate 3	69,917,571	23,558,077	33.69%
	Col-0 merged	118,926,824	43,810,915	36.84%
	Col-0 control	42,618,606	17,513,673	41.09%
	Agu-1 replicate 1	17,970,796	4,817,783	26.81%
	Agu-1 replicate 2	45,636,924	17,912,740	39.25%
	Agu-1 replicate 3	35,165,000	14,043,180	39.94%
	Agu-1 merged	98,772,720	36,773,703	37.23%
	Agu-1 control	36,528,152	12,449,515	34.08%
Arabidopsis lyrata **(JGI v1.0)**	***A. lyrata*** **replicate 1**	11,046,273	4,164,389	37.70%
	A. lyrata **replicate 2**	30,938,855	6,588,333	21.29%
	A. lyrata **replicate 3**	37,702,315	14,416,913	38.24%
	A. Lyrata **merged**	79,687,443	25,169,635	31.59%
	A. Lyrata **control**	45,624,101	17,439,276	38.22%

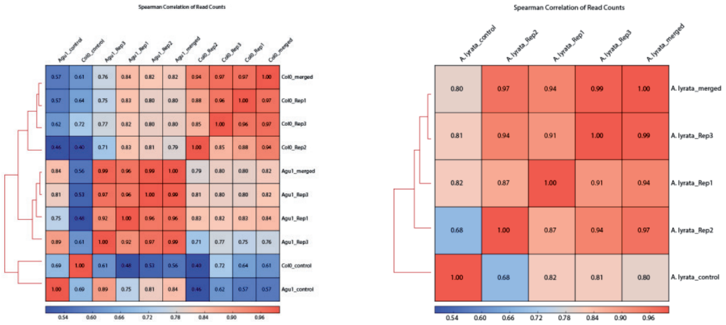


Figure S4: Reproducibility of DNAse-seq data. *Spearman correlation coefficients were calculated between samples, using non-overlapping windows of 5 kb. Pairwise comparisons are shown in the heatmap.*

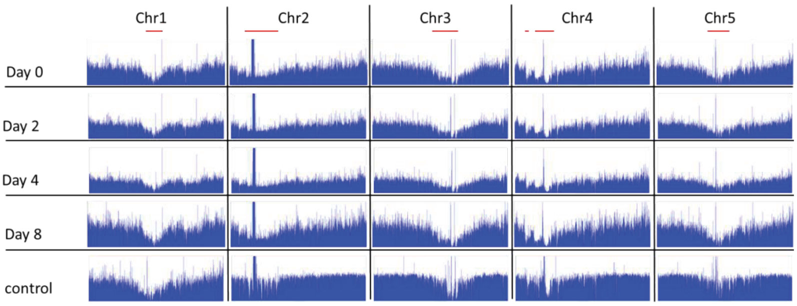


Figure S5: Chromosomal distribution of DNAse-seq data from Pajoro et al 2014 (Pajoro et al., 2014b). *The data were reanalysed in the same way as our dataset. The centromer is indicated with a red bar.*

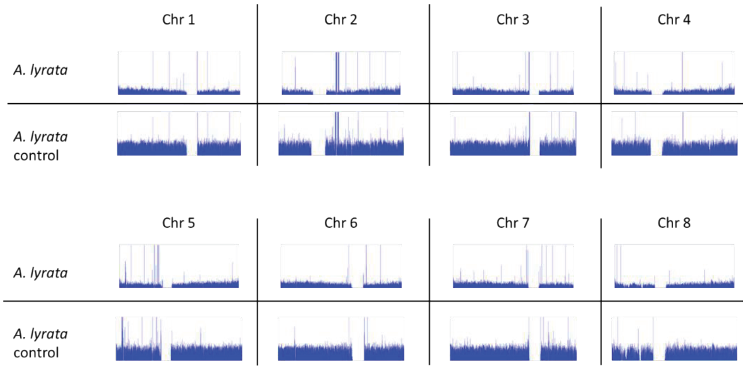


Figure S6: Chromosomal distribution of A. lyrata DNAse-seq reads. *As centromeric sequences are largely missing from the reference genome, it is not possible for sequences to be mapped to those regions.*

Table S2: Called DHS peaks in each replicate and merged sample with different peak-calling methods.

	Sample	Hotspot	MACS2	Merged	% H in M	% M in H
Arabidopsis thaliana **(TAIR 10)**	**Col-0 replicate 1**	3626	6099	6284	94.76%	52.07%
	Col-0 replicate 2	6006	12572	12661	98.52%	42.94%
	Col-0 replicate 3	4975	11453	11708	94.87%	37.22%
	Col-0 merged	7766	13672	14021	95.51%	48.20%
	Agu-1 replicate 1	6528	6325	7644	79.76%	74.42%
	Agu-1 replicate 2	6405	7472	8993	76.22%	60.49%
	Agu-1 replicate 3	6397	3711	7351	43.08%	69.79%
	Agu-1 merged	12159	4572	11996	38.89%	93.18%
Arabidopsis lyrata **(JGI v1.0)**	***A. lyrata*** **replicate 1**	1501	2675	2920	83.68%	40.93%
	A. lyrata **replicate 2**	2437	7019	7167	93.93%	29.51%
	A. lyrata **replicate 3**	2455	8555	8816	89.37%	23.50%
	A. lyrata **merged**	3580	15629	15951	91.01%	19.21%

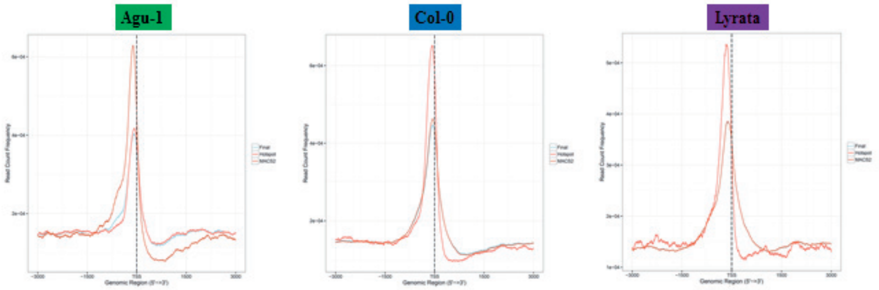


Figure S7: Genomic distribution of DHS peaks. *Both peak calling methods (MACS and HOTSPOT) found the same localization close to the TSS. Shown are the distribution for MACS (), HOTSPOT () as well as the final dataset (blue).*

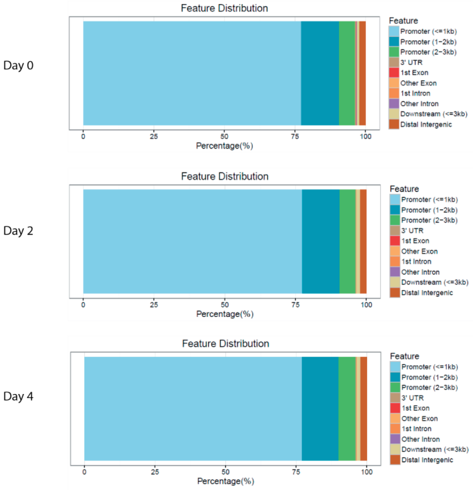


Figure S8: Distribution of DHS peaks in Pajoro et al 2014, data at different time points *(Pajoro et al., 2014b). All three timepoints resemble each other closely.*

Table S3: Comparison between the *Agu-1* DHS data generated here with published DNAse-seq data (Zhang et al., 2012a; Pajoro et al., 2014b). Our *Agu-1* dataset contains 9270 DHSs.

dataset	total peaks in dataset	number of peaks of dataset present in our *Agu-1* data	number of peaks of our *Agu-1* data present in this dataset	% peaks of our *Agu-1* data present in this dataset
leaf (Zhang)	20226	5392	5482	59.1%
inflorescence (Zhang)	23715	5850	5974	64.4%
inflorescence day 0 (Pajoro)	19054	5522	5734	61.9%
inflorescence day 2 (Pajoro)	15646	5203	5358	57.8%
inflorescence day 4 (Pajoro)	16334	5411	5573	60.1%
inflorescence day 8 (Pajoro)	19352	5357	5581	60.2%

Acknowledgments

This work was supported by EpiGeneSys with a grant for a small-scale collaborative project, and by a Sofja-Kovalevskaja project (KK).

CHAPTER 5

Evolution of PISTILLATA paralogs in *Tarenaya hassleriana*

Suzanne de Bruijn
Tao Zhao
Jose M. Muiño
Johanna Müschner
Gerco C. Angenent
Kerstin Kaufmann

Abstract

In angiosperms, floral development is specified by the fairly conserved (A)BCE-model. This model describes how three classes of MADS-domain proteins act in a combinatorial way to specify the four floral organ types. However, there is still a remarkable amount of diversity in floral morphology. One of the mechanisms suggested to contribute to this diversity is duplication of floral MADS-domain transcription factors. Although gene duplication is often followed by loss of one of the copies, sometimes both copies are retained. If both copies are retained they will initially be redundant, providing the freedom for one of the paralogs to change function. This way, gene duplication can lead to subfunctionalization or neofunctionalization.

One duplication of the floral regulator *PISTILLATA* (*PI*) occurred by transposition at the base of the Brassicales. This led to two *PI* paralogs, the original copy, which is conserved in genomic position in most of the angiosperms, and the transposed copy. Interestingly, some Brassicales species, such as the Brassicaceae, only retained the new, transposed copy, whereas others, such as species in the Cleomaceae, retained both copies of *PI*.

Here, we examined both of these *PI* paralogs in the Cleomaceae species *Tarenaya hassleriana*. We find that the two *ThPI* paralogs have very similar expression patterns. However, they may have diverged in function, as only one of these ThPI proteins was able to act heterologously in the first whorl of *A. thaliana* flowers. In addition, we observed differences in protein complex formation between the two paralogs, and there are subtle differences in the DNA-binding specificity of the two ThPI paralogs. Sequence analysis shows that most of the sequence divergence between the two paralogs seems to have emerged in a common ancestor of the Cleomaceae and the Brassicaceae. It is tempting to speculate that the duplication of *PISTILLATA* lead to different properties in the Brassicaceae-specific *PI*, which may contribute to the typical cross-like flower morphology observed in Brassicaceae.

Introduction

Gene duplication is a basis for evolutionary novelty, because selection pressure is temporarily less after a duplication event, allowing one or both of the duplicates to evolve in function. Following a gene duplication event, there are several scenarios for the fate of the newly obtained paralogs. Often, one of the paralogs is quickly lost (Lynch and Conery, 2000). In the case that both paralogs are retained, they might either divide the original function between the two paralogs (subfunctionalization) and/or obtain new functions (neofunctionalization) (Ohno, 1970; Force et al., 1999). Different molecular mechanisms have been proposed to explain how this is achieved (Conant and Wolfe, 2008; Andersson et al., 2015). These different mechanisms are not mutually exclusive, and often several mechanisms are acting upon a paralogous pair of genes simultaneously or consecutively (He and Zhang, 2005; Conant et al., 2014).

In plants, whole genome duplications (WGDs) are a common phenomenon, and all angiosperms have undergone at least one WGD (Lawton-Rauh, 2003; Cui et al., 2006; Soltis et al., 2014). WGDs are implied as a driving force behind the huge increase in the number of plant species (Moore and Purugganan, 2005; Tank et al., 2015). Crucially, several key innovations, such as seeds and flowers, coincided with WGDs (Jiao et al., 2011; Li et al., 2015; Soltis and Soltis, 2016). Interestingly, gene loss after WGDs is not uniform, with some classes of genes being preferentially retained, among which transcription factors (TFs) (Blanc and Wolfe, 2004a; Conant and Wolfe, 2008; Paterson et al., 2010). One of the families of TFs that have seen a preferential retention of its genes is the family of MADS-domain TFs (Irish, 2003; Geuten et al., 2011). Members of this family of TFs are involved in virtually all stages of plant development (Smaczniak et al., 2012a) and are well-known for their crucial roles in flower development (Causier et al., 2010a). They specify the identity of the four different floral organ types, according to the combinatorial (A)BCE model (Haughn and Somerville, 1988; Coen and Meyerowitz, 1991; Theissen, 2001; Causier et al., 2010a). This model and the proteins that fulfil the A-, B-, C- and E-function are generally conserved throughout the angiosperms (Smaczniak et al., 2012a). However, many plant lineages retained multiple copies of these genes after duplication events (Vandenbussche et al., 2004; de Martino et al., 2006; Mondragon-Palomino and Theissen, 2008; Geuten and Irish, 2010; Sharma and Kramer, 2013).

The first floral MADS-box gene paralogs studied in detail were *PLENA* (*PLE*) and *FARINELLI* (*FAR*) from *Antirrhinum majus*. *PLE* is the C-function gene in *Antirrhinum,* with mutants showing homeotic conversions of stamen to petal, and carpel to sepaloid organs, as well as loss of floral determinacy (Bradley et al., 1993). The closely related *FAR* gene was expected to have a floral C-function as well based on gene sequence, expression pattern and protein-protein interactions. However, the only phenotype *far* mutants exhibit is partial male sterility. Nevertheless, the double *far ple* mutant has a more severe phenotype (more complete homeotic conversions, and more organs inside the 4th whorl) than single *ple* mutants, indicating that *FAR* is partially redundant with *PLE*. These data suggest that *PLE* and *FAR* have subfunctionalized (Davies et al., 1999).

Arabidopsis also retained paralogous pairs from the C-, as well as the A- and E-classes of MADS-box genes. The paralogous gene pairs show different degrees of divergence. The paralogs of the C-class gene *AGAMOUS (AG)*, the *SHATTERPROOF*s (*SHP1* and *2*), are not involved in flower development, but play a role in fruit development instead (Liljegren et al., 2000). In contrast, the four *SEPALLATA* paralogs (*SEP1-4*, E-class) are largely redundant (Pelaz et al., 2000; Ditta et al., 2004).

The B-function is fulfilled by two genes, *APETALA3/PISTILLATA (AP3/PI)* and *DEFICIENS/GLOBOSA (DEF/GLO)* in the model species *Arabidopsis thaliana* and *Antirrhinum*, respectively*.* These two genes are closely related, resulting from a duplication before the origin of the angiosperms (Kramer et al., 1998). Within the angiosperms these genes underwent additional duplications. Although *A. thaliana* has only one copy of each B-class gene lineage, several other plant lineages have retained paralogs of the B-class genes after duplication (Kramer et al., 1998; Stellari et al., 2004; Viaene et al., 2009). For example, all Solanaceae species have two AP3-like genes (*AP3* and *TM6*), as well as two *GLO* paralogs. These paralogs have subfunctionalized in species-specific ways (Vandenbussche et al., 2004; de Martino et al., 2006; Geuten and Irish, 2010). A similar pattern is seen in the asterids, where a duplication in the basal asterids led to two *PI* paralogs that show species-specific differences in expression patterns (Viaene et al., 2009).

B-class genes do not only specify petal and stamen identity, but can also be involved in determining the morphology of these organs. For instance, in *Petunia hybrida* a *PI* paralog is required for the fusion of stamens to the corolla tube (Vandenbussche et al., 2004). Another example of involvement of B-class genes in morphology is provided by orchids. Orchids possess a perianth that consists of three morphologically distinct types of tepals, and it has been shown that these different tepal morphologies are specified by different combinations of the three/four *DEF* paralogs that are present in orchids (Mondragon-Palomino and Theissen, 2009; Mondragón-Palomino and Theißen, 2011).

B-class genes might even be able to specify novel floral organs. An example is presented by *Aquilegia*, a basal eudicot that displays an additional type of organ in a whorl between the stamens and carpels, called the staminodia. The specification of this new organ is linked to duplications in the *AP3* lineage (Kramer et al., 1998; Kramer et al., 2003; Kramer et al., 2007; Sharma and Kramer, 2013).

The position of a gene within the genome can be biologically relevant, as genes are dependent on their genomic context for expression. Gene expression is regulated by *cis*-regulatory elements (CREs), which can be dispersed over long distances, even spanning several genes. Epigenetic marks also play a role in regulating gene expression, and as these marks are often deposited in big “blocks”, the epigenetic state of a gene might be dependent on its position in the genome (Dewey, 2011). Interestingly, several studies have shown that after gene duplication, the original gene is more evolutionary constrained in sequence than the copy (Dewey, 2011). For these reasons, exceptionally strong conservation of gene order could indicate that the genomic context of a gene is important for its function and/or regulation (Duran et al., 2009), and is often used as a proxy for the conservation of gene function (Dewey, 2011).

This preservation of gene order in different species is called synteny (Drillon and Fischer, 2011; Irimia et al., 2013; Tekaia, 2016). Synteny can be maintained across hundreds of millions of years, with 90% of the genome being syntenic between human and mice (90 MYA) (Mudge et al., 2005). However, in plants synteny is generally less conserved than in animals. This is due to the fact that in plants several rounds of WGD have occurred and the subsequent process of gene loss and genome rearrangements has blurred syntenic relationships (Timms et al., 2006; Tang et al., 2008; Irimia et al., 2013). Still, extensive genome collinearity can be found between closely related species, and plant species that are more diverged still show microsynteny of small genomic regions (of several genes) (Yan et al., 2004; Timms et al., 2006; Lescot et al., 2008).

In comparative genomics, synteny is used to elucidate true orthologs from other homologous genes. Therefore, synteny can provide information about the evolution of gene families. One family for which synteny analysis has helped unravel its evolutionary history is the family of MADS-box genes (Ruelens et al., 2013; Zhao et al., 2017). One MADS-box gene that displays very conserved synteny is the floral B-class gene *PISTILLATA* (*PI*). The synteny of this gene is retained between the sister species of all angiosperms, *Amborella* and almost all other angiosperm species, with the notable exception of the Brassicaceae family (Cheng et al., 2013; Zhao and Schranz, 2017).

Tarenaya hassleriana belongs to the Cleomaceae, which is a sister family to the Brassicaceae (Hall et al., 2002). *T. hassleriana* is interesting for comparative studies, as the genome is available, and this species diverged from the Brassicaceae relatively recently (35 million years ago) (Cheng et al., 2013). This means that *T. hassleriana* is relatively closely related to the well-established model species *A. thaliana*. In contrast to the cross-like Brassicaceae flowers, different Cleomaceae species exhibit quite diverse floral morphologies (Patchell et al., 2011). *T. hassleriana's* basic floral bauplan (4 sepals, 4 petals, 6 stamens and 2 fused carpels) is similar to *A. thaliana*, but in contrast to the radially symmetric *Arabidopsis* flowers, the flowers of Cleomaceae species are monosymmetric (Patchell et al., 2011; Cheng et al., 2013). *T. hassleriana* has two paralogs of the B-function gene *PI* (Cheng et al., 2013). These *PI* paralogs are probably derived from the At-β-duplication at the origin of the Brassicales, ca. 70 MYA (Ming et al., 2008; Cheng et al., 2013).

We questioned whether the genomic location of *PI* could influence floral morphology, as the synteny of *PI* is conserved throughout the angiosperms, with the exception of the Brassicaceae. *T. hassleriana* has two *PI* paralogs, and this species is closely related to *A. thaliana*, and therefore may be an evolutionary intermediate between the Brassicaceae and the other eudicots. Here, we investigated how these two *PI* paralogs in *T. hassleriana* diverged from each other focusing on expression patterns and several functional features of the two TFs. To test for functional divergence, we performed heterologous expression experiments, and tested both proteins for their specificity in interacting with DNA as well as with other TFs. Although the two paralogs did not diverge in expression pattern, we did observe differences in the biochemical properties of the two paralogous genes. The paralogs are not equally competent in acting heterologously as B-class genes in *Arabidopsis*. Protein interaction studies indicate that although the two PI paralogs behave similar in dimer formation, they have different affinities for interaction partners in the floral quartet model. This model suggests that the MADS-domain TFs underlying the (A)BCE-model exert their function as tetramers (Theissen and Saedler, 2001). These data indicate that both PI paralogs diverged from each other in their biochemical properties, which could imply divergence in gene function. This might have interesting implications for the functional evolution of *PI* genes in the Brassicaceae.

Results

Phylogenetic analysis shows that one of the Tarenaya *PI* paralogs clusters with the Brassicaceae

Previously, it was found that *T. hassleriana* possesses two copies of each B-class gene, *APETALA3* and *PISTILLATA* (Cheng et al., 2013). Interestingly, the *PI* paralogs are in different genomic environments. One of the *ThPI* paralogs shares conserved synteny with the Brassicaceae-specific *PIs* (*ThPI-1*), whereas the other *PI* paralog (*ThPI-2*) is syntenic with the other eudicots (**Figure 1** (Cheng et al., 2013)). The two *PI* paralogs, Th*PI-1* (215 AA) and Th*PI-2* (214 AA), are highly divergent in sequence, sharing only 62% protein sequence identity (68% at the nucleotide level) (**Figure 2A**). Here we present a more detailed phylogeny of *PI* (**Figure 2B**), which shows that the *ThPI* paralog that is syntenic with the Brassicaceae *PI* genes clearly clusters with the Brassicaceae *PI*-clade. The other paralog, *ThPI-2,* is positioned between the Brassicaceae-clade and the eudicots-specific clade. The phylogenies shown are made with Maximum-likelihood, but the Neighbor-Joining algorithm produced similar results. *ThPI-1* clusters with the Brassicaceae, indicating that *ThPI-1* resembles the Brassicaceae *PI* genes in sequence. This means that a substantial part of the sequence divergence observed between *ThPI-1* and *ThPI-2* arose in *ThPI-1* before the split between the Brassicaceae and Cleomaceae. That *ThPI-1* evolved in sequence in a common ancestor of the Brassicaceae and Cleomaceae is also seen in the C-terminal part of the sequence. It has been shown that the N-terminal part of the PI motif (see **Figure 2**) contains signature amino acids able to distinguish between the Brassicaceae and eudicot-specific PIs (Lange et al., 2013). Like the Brassicaceae PI orthologs, ThPI-1, but not ThPI-2, misses the first 2 AA of the PI motif. In addition, ThPI-1 (but not ThPI-2) also resembles the Brassicaceae PI proteins in having a C-terminal extension of six amino acids compared to other eudicot PIs (Lange et al., 2013).

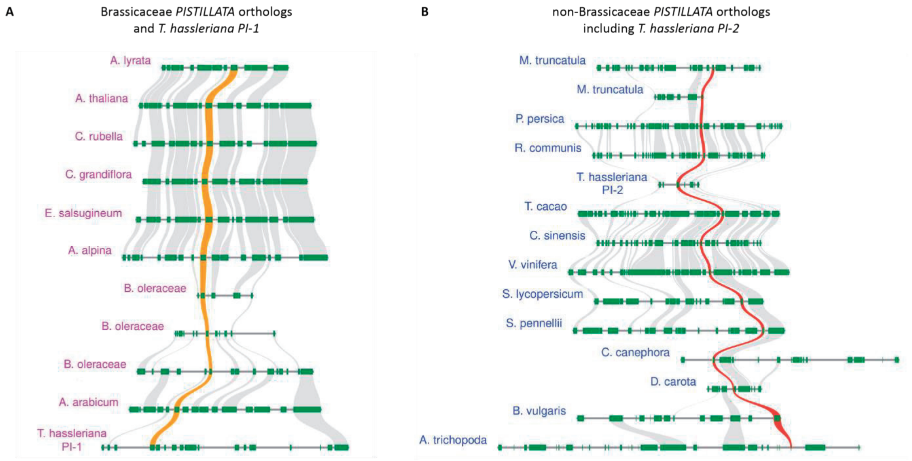


Figure 1: Synteny of both PI paralogs of T. hassleriana. *PI paralogs are shown with an orange (Brassicaceae) or red (non-Brassicaeae) trace. Other syntenic genes are linked in grey.* ***A:*** *Synteny of ThPI-1 with the Brassicaceae PI orthologs. Shown are the Brassicaceae type I genera Arabidopsis (A. lyrata and A. thaliana) and Capsella (C. rubella and C. grandiflora); the Brassicaceae type II species Eutrema salsugineum, Arabis alpina, Brassica oleraceae (3 paralogs), the basal Brassicaceae Aethionema Arabicum and PI-1 of T. hassleriana.* ***B:*** *Synteny of ThPI-2 with non-Brassicaceae angiosperms. Shown are 7 rosid species (Medicago truncatula (2 paralogs), Prunus persica, Ricinus communis, T. hassleriana, Theobroma cacao, Citrus sinensis and Vitis vinifera), 3 asterid species (Solanum lycopersicum, Solanum pennelli and Coffea canephora), one Caryophyllales (Beta vulgaris) and the sister species to all other angiosperms, Amborella (A. trichopoda).*

Once again, this indicates that these changes to the protein sequence of PI occurred before the Brassicaceae-Cleomaceae split, when there were two paralogs of PI present in the genome of the common ancestor of these families. Interestingly, both paralogs of *PI* are also retained in the Cleomaceae species *Gynandropsis gynandra*, hinting that both copies might be functional (**Figure 2A, B**). In contrast, the two *AP3* paralogs of *T. hassleriana* originate from a tandem duplication that is thought to be recent (Cheng et al., 2013), as the two paralogs are highly similar (**Figure 2C, D, S1**). We found that this tandem duplication is not present in *G. gynandra,* which supports the idea that this tandem duplication happened recently, as the Gynandropsis-Tarenaya split is no more than 13.7 MYA (van den Bergh et al., 2014).

The two *ThPI* paralogs did not diverge in expression pattern

It is known that the genomic context of a gene may have an influence on its expression. As the *T. hassleriana PI* paralogs have different genomic environments, we investigated whether they diverged from each other in expression pattern. It was previously shown that both *PI* paralogs in *T. hassleriana* are expressed during flower development (Cheng et al., 2013). Here, we investigate the expression patterns of these genes in more detail, using RNA *in situ* hybridization.

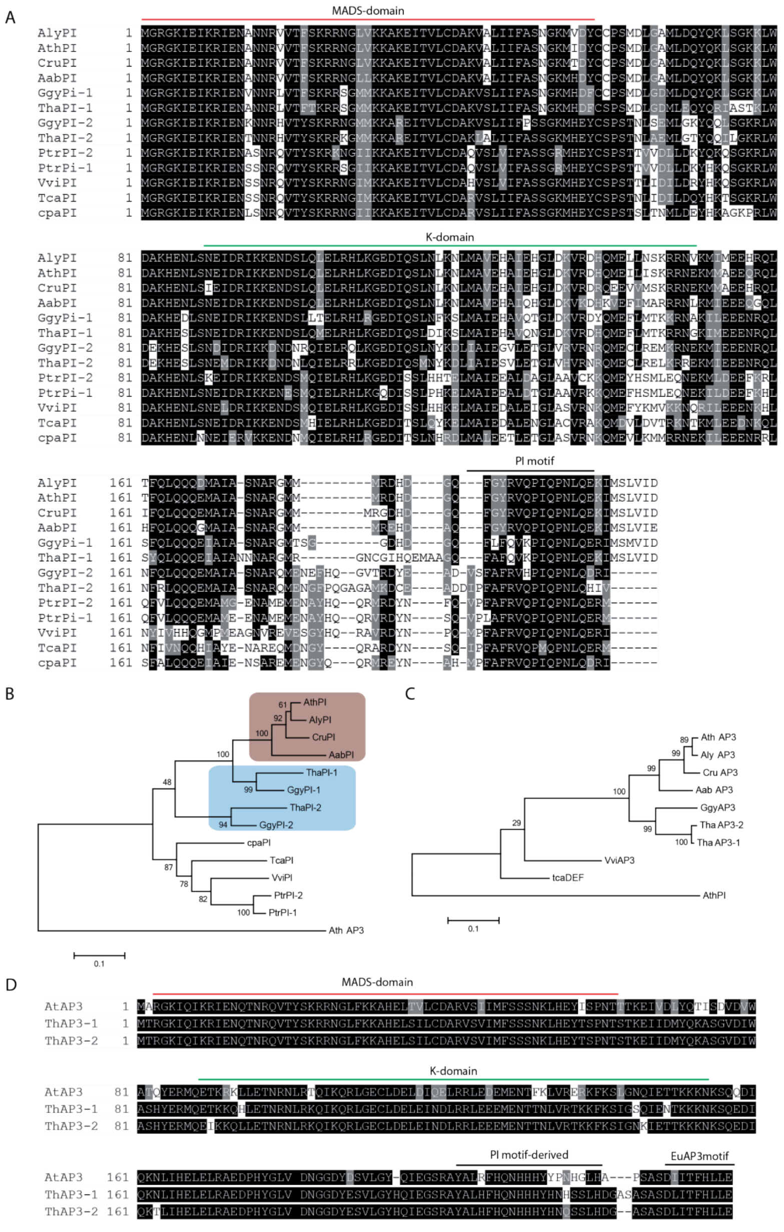


Figure 2: Phylogenetic and sequence analysis of the T. hassleriana B-class genes. A: *Alignment of PI orthologs from several species.* ***B:*** *Maximum-likelihood phylogeny of PI orthologs. The PI orthologs belonging to Brassicaceae species are indicated in beige, the Cleomaceae PI orthologs in blue.* ***C:*** *Maximum-likelihood phylogeny showing the position of the ThAP3 paralogs.* ***D:*** *Alignment of ThAP3 paralogs with AtAP3. The MADS-domain, the K-domain and the lineage-specific C-terminal motifs are indicated in A and D. Abbreviations: Ath=A. thaliana; Aly=A. lyrata; Cru=Capsella rubella; Aab=Aethionema Arabicum; Tha=T. hassleriana; Ggy=G. gynandropsis; Cpa=Carica papaya; Tca=Theobroma cacao; Vvi=Vitis vinifera; Ptr=Populus trichocarpa.*

We designed probes for *ThAP3*, *ThPI-1* and *ThPI-2.* The *ThAP3* probe cross-hybridizes with transcripts from both *ThAP3* paralogs, because the extremely high similarity between these genes did not allow for the design of specific probes (see **Figure S1**). The *ThAP3* probe covered part of the K-domain and the C-terminus of the mRNA, as well as the 3' UTR. The probes for the *PI* paralogs only cover the C-terminal part of the mRNA and the 3'UTR (which we determined using 3'RACE), not the K-domain. Although these two *PI* probes share only 65 % similarity at nucleotide level (longest continuous stretch of identical sequence is 14 bp), we cannot exclude some cross-hybridization with mRNA from the other paralog.

Early during development, when only sepal primordia are present (comparable to *A. thaliana* floral stage 3-4 (Smyth et al., 1990)), *ThAP3* was found expressed in the cells that will give rise to whorl two and three (**Figure 3A and 3I**), and at later developmental stages, when primordia of all organs are formed, expression of *AP3* is specific to petal and stamen primordia (**Figure 3E and 3M**). The expression patterns of *ThPI-1* and *ThPI-2* resemble each other closely. During early stages, expression of both genes can be seen in cells that will give rise to whorl two and three (**Figure 3B, C, J and K**) and, later during development, expression is seen in developing petal and stamen primordia (**Figure 3F, G, H and O**). These data show that *ThPI-1* and *ThPI-2* are both expressed in petals and stamens, with no detectable differences in expression pattern between the two paralogs. It seems therefore that both *PI* paralogs did not diverge in spatial and temporal expression patterns. They are however reported to differ in their level of expression, the expression of *ThPI-2* being lower than the expression of *ThPI-1* (**Figure S5**)(Cheng et al., 2013; Kulahoglu et al., 2014).

Heterologous expression of *Th*PI paralogs in *A. thaliana* gives different phenotypes

Although the two *ThPI* paralogs do not seem to have diverged in expression pattern, they did diverge quite substantially in sequence. We therefore hypothesized that they might have evolved in their function. As a first test for protein function, we expressed both *ThPI* paralogs constitutively in wildtype *A. thaliana*. As a control, we also created lines with *AtPI* overexpression. Constitutive overexpression of the native *PI* in *A. thaliana* has been reported to lead to partial conversion of sepals to petals, due to low expression of *AP3* in the outer whorl (Krizek and Meyerowitz, 1996b). We analyzed twelve *35S:AtPI* lines, and obtained sepal to petal conversion for three of these lines (**Figure 4B**). Expression studies indicated that lines showing a modified phenotype had the highest level of transgene expression (**Figure S2A**). For *ThPI-1*, the paralog that is most similar to the Brassicaceae *PI*, the two lines with the highest transgene expression level (out of 13 lines) displayed homeotic conversions of sepals, very similar to the *35S:AtPI* lines (**Figure 4C, S2B**). This indicates that ThPI-1 is capable of performing similar functions as AtPI in the first whorl of *A. thaliana*. For ThPI-2*,* although we analyzed 14 35S:*ThPI-2* lines, we did not observe an aberrant phenotype in any of these lines (**Figure 4D**). This could indicate that ThPI-2 is unable to specify petals in the first whorl of *A. thaliana*, suggesting that ThPI-2 is functionally different from AtPI. As ThPI-1 did induce a phenotype, it can be concluded that the two *T. hassleriana* proteins are biochemically different. However, it cannot be ruled out that the transgene expression levels in our 14 lines were not high enough to induce a homeotic transformation of the first whorl organs (**Figure S2C, D**), even though we analyzed a similar number of lines for *35S:ThPI-2* as we did for *35S:AtPI* and *35S:ThPI-1*.

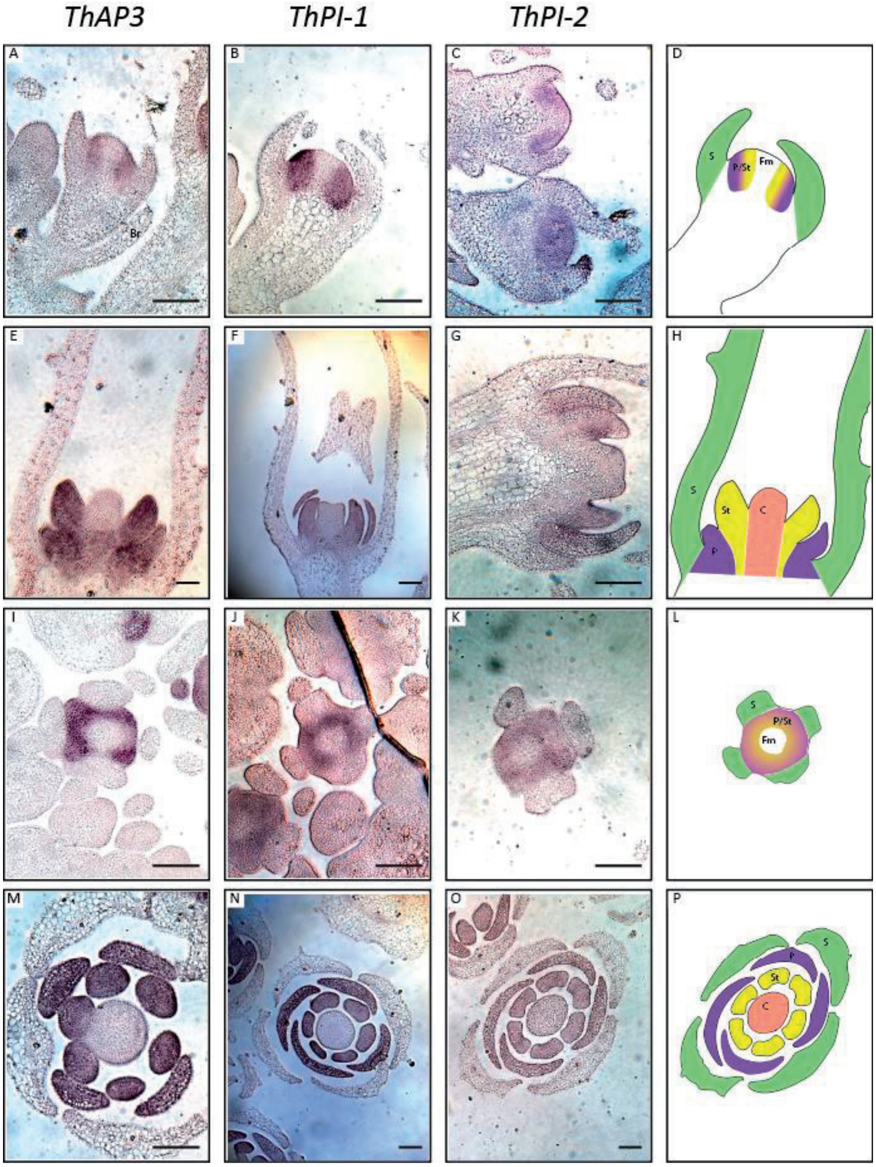


Figure 3: Expression patterns of T. hassleriana B-class genes. *Expression patterns of the ThAP3 paralogs (**A**, **E**, **L**, **M**), ThPI-1(**B**, **F**, **J**, **N**) and ThPI-2 (**C**, **G**, **K**, **O**) as determined by RNA in situ hybridization. Expression was determined in early developmental stages before organ primordia were formed (**A**-**C**, **I**-**K**) as well as later during development (**E**-**G**, **M**-**O**). Schematics of the different developmental stages and planes are shown in (**D**, **H**, longitudinal) and (**L**, **P**, cross). Scale bar=1mm.*

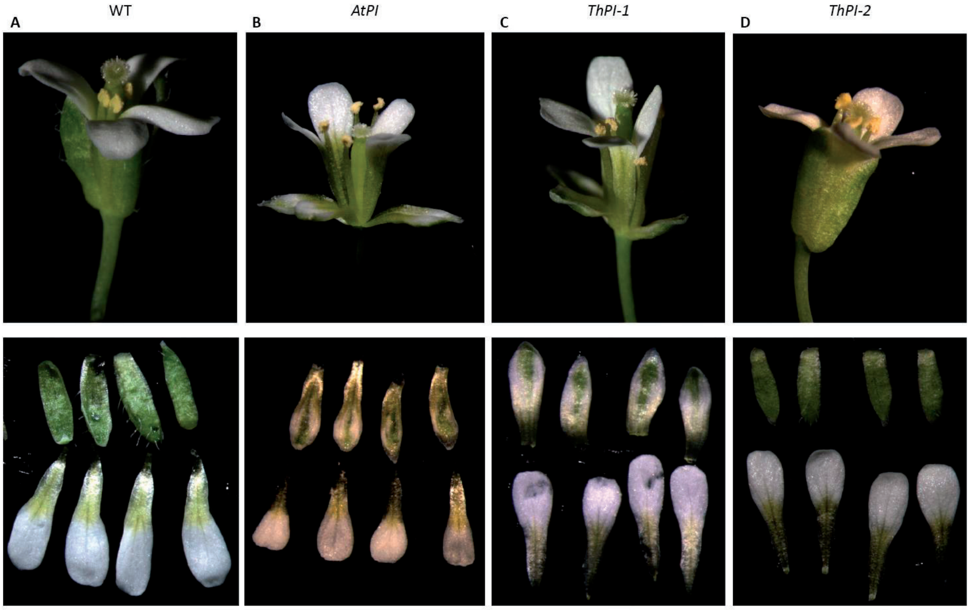


Figure 4: Heterologous expression of the ThPI paralogs in A. thaliana*. A:** *a wildtype (WT) A. thaliana flower.* ***B: *a 35S::AtPI flower, which shows the phenotype obtained by constitutive expression of the native PI. Note the change in orientation of the first whorl organs.* ***C:*** *a 35S::PI-1 flower, showing homeotic conversion of sepals to petals.* ***D:*** *a 35S::ThPI-2 flower, showing no aberrant phenotype. Top row shows whole flower, bottom row shows dissected sepals (top) and petals (bottom).*

The ThPI paralogs differ in their ability to form protein-protein interactions

The heterologous expression assay indicates that the *ThPI* paralogs might be functionally different. We therefore studied the properties of the encoded proteins *in vitro*. As MADS-domain TFs function as part of protein complexes (Pellegrini et al., 1995; Theissen and Saedler, 2001; Smaczniak et al., 2012b), we tested whether the two ThPI paralogs have different capabilities to form DNA-binding protein complexes. Differences in protein complex formation can be relevant, as divergence in protein-protein interactions may lead to divergent gene regulation. TFs need to bind DNA to exert their function, therefore DNA-binding protein complexes were analyzed using Electrophoretic Mobility Shift Assays (EMSAs), a well-established method to study DNA-binding MADS-domain protein-complexes (Tröbner et al., 1992; Riechmann et al., 1996b).

Initially we analyzed interactions between the two ThAP3 and the two ThPI proteins. We could not detect DNA-binding by homodimers of any of the four B-class proteins (**Figure 5A**). This is not surprising, as AP3 and PI form obligate heterodimers in the majority of the eudicots (Riechmann et al., 1996b; Winter et al., 2002; Melzer et al., 2014). We did detect all four possible ThAP3-ThPI heterodimers (**Figure 5A**), indicating that there has been no subfunctionalization at the dimerization level. Interactions between B-class paralogs have been studied in more species. Whereas in some species subfunctionalization at the dimerization level has not been observed (Roque et al., 2013; Gong et al., 2016; Roque et al., 2016), in other species, for instance in the Solanaceae, only specific B-class dimer combinations are possible (Leseberg et al., 2008; Geuten and Irish, 2010). As the ThAP3 paralogs are highly similar to each other, it is not surprising that we did not find subfunctionalization at the dimerization level. Interestingly, ThPI-2 containing dimers migrate slower through the gel than ThPI-1 containing dimers, even though they have similar molecular masses (24.72 vs 24.95 kDa) and charges (**Table S1**).

Although there are no apparent differences between ThPI-1 and ThPI-2 in their ability to form heterodimers with ThAP3 paralogs, it is possible they have different abilities to form larger protein complexes. According to the floral quartet model, B-class proteins act in tetramers with other MADS-domain TFs (Honma and Goto, 2001; Theissen and Saedler, 2001). To determine whether the ThPI paralogs differ in higher order complex formation, we investigated their ability to form complexes with members of other homeotic protein classes. According to the floral quartet model, we expect the B-class proteins to interact with a SEPALLATA (SEP) protein and APETALA1 (AP1) in petals, whereas a complex with one AGAMOUS (AG) and one SEP protein should specify stamens. *T. hassleriana* has one *AG* gene and two genes each for *SEP1/2*, *SEP3* and *SEP4* (Cheng et al., 2013). Focusing on the stamen-specific complex, we analyzed whether the ThPI paralogs interact differently with ThAG and the two ThSEP3 paralogs. SEP3 was chosen as the SEPALLATA protein, as it is suggested to be the most active SEP in *A. thaliana,* based on the number of different protein-interactions it is forming (Immink et al., 2009). First we compared complex formation for all four different B-class heterodimers. As expected, the two ThAP3 paralogs behaved similar in these experiments (**Figure S3A**). However, for the two ThPI paralogs differences in complex formation were observed. Whereas one higher-order complex (beside a dimer complex) was observed for combinations containing ThPI-1, two tetrameric complexes were observed when ThPI-2 was present. This pattern was found for both ThSEP3 paralogs (**Figure S3A**). We studied the composition of these different complexes in more detail for one of the ThSEP3 paralogs (Th1528) (**Figure 5B, S3B, C**). Using dropout experiments, it could be concluded that the single tetrameric complex observed with ThPI-1 consists of ThAP3/ThPI-1/ThAG/ThSEP3, which is the expected complex for stamen-specification. A similar complex (ThAP3/ThPI-2/ThAG/ThSEP3) was observed with ThPI-2 (**Figure 5B**, marked with an asterisk). However, when ThPI-2 is present, a second tetrameric complex (upper band) is observed. This other complex does not contain any B-class proteins, but instead consists of only ThSEP3 and ThAG. The fact that a ThAG/ThSEP3 tetramer is formed in addition to a ThAG/ThSEP3/ThAP3/ThPI-2 complex suggests that a fraction of ThSEP3/ThAG dimers bind to each other, instead of to a ThAP3/ThPI-2 dimer. Although we did not test for differences in protein levels, these data indicate that there are differences between the two ThPIs in affinity to form a complex with ThAG and ThSEP3. The affinity of ThAG/ThSEP3 for ThPI-2 is lower than for ThPI-1, because for ThPI-1 all ThAG/ThSEP3 dimers are incorporated into a ThAG/ThSEP3/ThAP3/ThPI-1 complex.

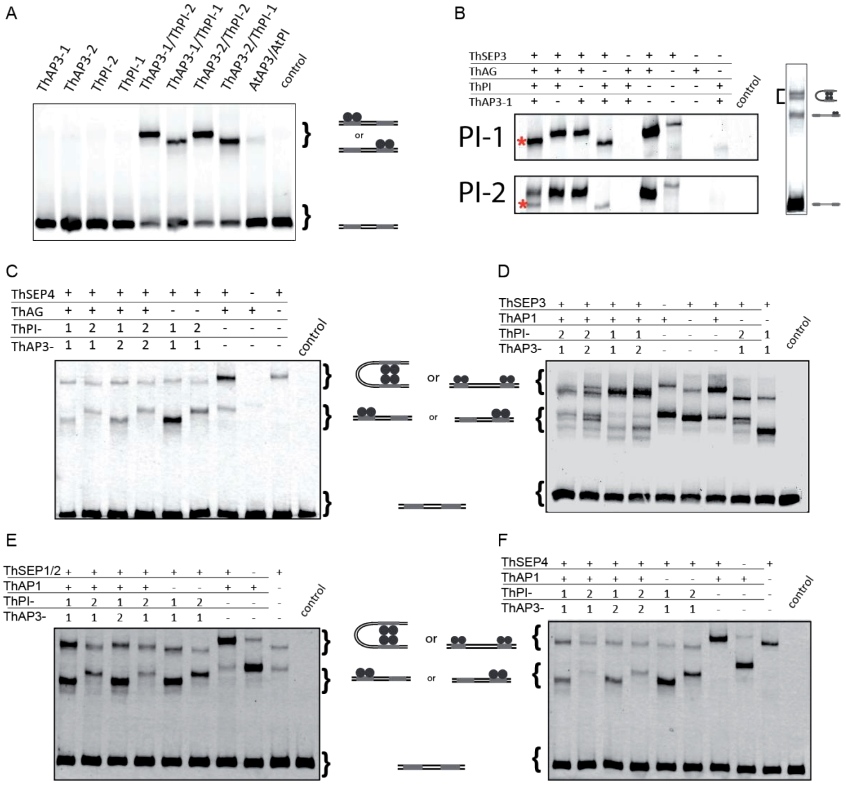


Figure 5: Ability of T. hassleriana B-class genes to form DNA-binding protein-complexes. *In **(A)**, incubations of homo- and heterodimerization of AP3 and PI with DNA probe. **B:** Complexes formed with ThAG, ThSEP3 (Th01528), ThAP3-1 and either of the two ThPI paralogs. The figure shows only the higher order complexes (tetramers); in the figure at the right this part compared to the whole gel is indicated. In **(C)**, EMSA for protein-complex formation with ThAG and a ThSEP4 paralog (Th21984). **(D, E, F)** EMSAs testing the interaction of the B-class dimers with a ThAP1 paralog (Th13754) and different ThSEP paralogs: ThSEP3 (Th1528) (**D**), ThSEP1/2 (Th2854) (**E**), and ThSEP4 (Th21984) (**F**). For all experiments, a promoter fragment from the A. thaliana SEP3 promoter was used as probe (Smaczniak et al., 2012a). The control is an empty-vector control, in which no protein production is expected.*

Summarizing, both ThPI paralogs are capable of forming a complex with ThAG and ThSEP3. However, the data suggest that they do so with different affinities, as ThPI-1 shows a higher affinity for this complex than ThPI-2.

We next investigated if ThPI-2 has a lower affinity than ThPI-1 to form tetramers in general, or whether it is specific for certain protein combinations. We therefore first analyzed tetramer formation with ThAG and a different ThSEP paralog. Interestingly, a single tetrameric complex was observed for both of the ThPI paralogs when a ThSEP4 paralog (Th21984) was used (**Figure 5C**). This indicates that the lower affinity to form a ThAG/ThSEP/ThAP3/ThPI complex with ThPI-2 than with ThPI-1 may not be a general feature for ThPI-2.

Next, we studied combinations of the B-class proteins with one ThSEP and one of the ThAP1 paralogs, a combination that is expected for petal formation **(Figure 5D, E, F)**. interestingly, using ThSEP3, we see a single complex when ThPI-1 is present, whereas two complexes are formed when ThPI-2 is present. Similar to what we have seen for combinations with ThAG, it seems that the higher complex observed for combinations with ThPI-2 may not contain the B-class proteins, as it runs at the same height as ThAP1 homotetramers. When we tested the interaction of the B-class paralogs with ThAP1 and ThSEP4, we obtain a single complex for either of the ThPI paralogs, again indicating that there is no difference in complex formation with ThSEP4. When we examine combinations of the B-class proteins with ThAP1 and another SEP, ThSEP1/2, we observe a single complex for each protein combination. However complexes containing ThPI-1 show a different gel shift than combinations with ThPI-2. Differences in gel shift indicate that these complexes will likely have a different protein composition, but we did not study these differences in detail.

From these experiments it can be concluded that ThPI-1 and ThPI-2 are biochemically different, as they show differences in their affinities to form higher-order complexes. ThPI-2 has a lower affinity for certain higher order complexes than ThPI-1. However, this does depend on the interaction partners, as different ThSEP paralogs gave different results. We can conclude that the ThPI paralogs (as well as the ThSEP paralogs) are diverged in their ability to form protein-protein interactions.

DNA-binding specificity

In the EMSA experiments a single DNA probe is used and the interaction of this probe with the various protein complexes can be tested. However, it is also possible that the two ThPI paralogs differ in their binding specificity and/or affinity to certain DNA sequences. To analyze this, we used SELEX-seq (Systematic Evolution of Ligands by EXponential enrichment followed by deep sequencing) (Jolma et al., 2010) to test whether there are any differences in DNA-binding specificity between the two ThPI paralogs. We performed SELEX-seq experiments on ThAP3/ThPI heterodimers using a custom-made *A. thaliana* AP3 antibody, which recognized the *T. hassleriana* AP3 paralogs (see **Figure S4**).

Good enrichment of bound sequences was obtained for the heterodimers ThPI-1/ThAP3-1 and ThPI-2/ThAP3-1, in SELEX round 8 and 5 respectively, as shown by EMSAs (**Figure 6A, B**). We sequenced these SELEX rounds, and obtained ~0.3 million reads for PI-1/AP3-1 and ~30 million reads for PI-2/AP3-1, with a percentage of perfect CArG-boxes ($CC[A/T]_6GG$) of 12,6% and 14% respectively, indicating that we indeed have good enrichment of ThAP3/ThPI bound sequences. We calculated relative affinities of the heterodimers for each 10 bp sequence (k-mer), and compared these between the two different ThPI heterodimers. This shows that there are differences in DNA-binding specificity between the two different ThPI proteins (**Figure 6C**). For each heterodimer, we used the top 0.1% of K-mers with the highest affinity to perform a motif search using MEME. For both ThPI paralogs we find a motif resembling a CArG-box (**Figure 6D**). For each heterodimer, the motif shows two conserved cytosines in the beginning of the motif, whereas at the 3’ end the first guanine is less conserved than the second. However, the motifs are slightly different from each other. The A-rich stretch in the middle is slightly different for both motifs. In addition, although both motifs show an extension of the motif 3' of the CArG-box, these are slightly different, with the extension of the ThPI-1 motif consisting of two adenines, whereas for ThPI-2 this motif is a thymine followed by two adenines. Taken together, these data suggest that there might be subtle differences in DNA-binding specificities between the two different ThPI paralogs.

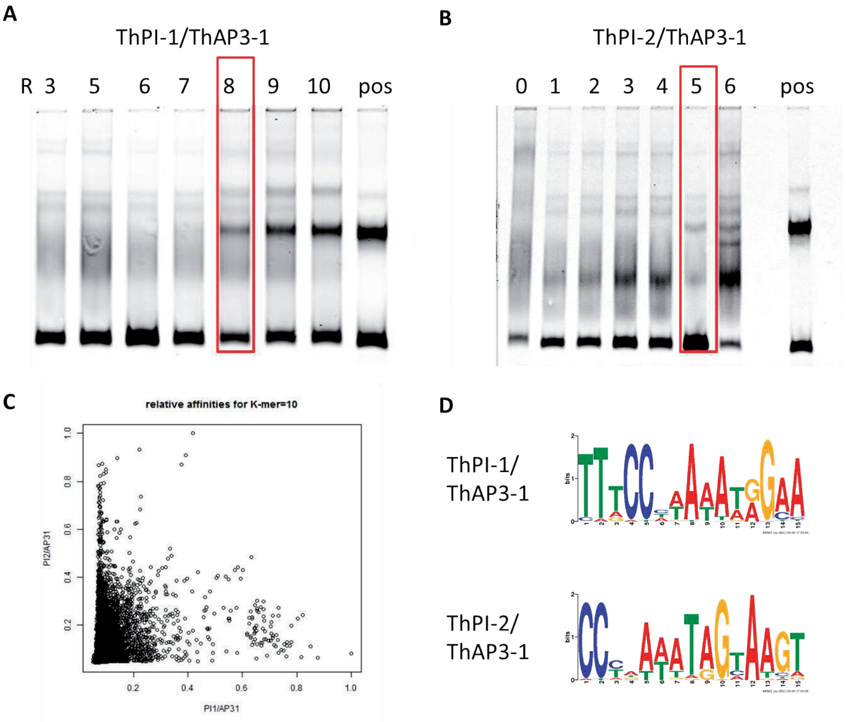


Figure 6: DNA-binding specificities as determined by SELEX-seq*. (****A, B****) EMSAs showing enrichment of bound sequences in different SELEX rounds.* ***C:*** *Dotplot comparing the relative affinities between ThPI-1/ThAP3-1 and ThPI-2/ThaP3-1.* ***D:*** *Motifs obtained for each of the PI/AP3-1 heterodimers. Motif discovery was performed on the most recurring 40N sequence for each of the 0.1% k-mers with the highest affinity.*

Discussion

T. hassleriana has two *PI* paralogs that probably resulted from the β-duplication around 70 million years ago at the base of the Brassicales (**Figure 1**, (Ming et al., 2008; Cheng et al., 2013)). These two paralogs share 62% protein identity. This amount of sequence divergence falls within the range observed for functional B-class paralogs in other species. In the Solanaceae, the GLO paralogs have 63-70% protein identity (around 108 million years old (Bremer et al., 2004; Viaene et al., 2009)), and the paralogs MtPI and MtNGL9 in *Medicago truncatula* share 73% protein identity (duplication occurred around 39 million years ago) (Benlloch et al., 2009; Roque et al., 2016). The basal eudicot *Aquilegia* has three paralogs of

AP3, which share about 60% protein identity. These paralogs originated from two duplications, of which the older one is estimated to be around 120 MYA (Kramer et al., 2003; Kramer et al., 2007; Moore et al., 2007; Sharma and Kramer, 2013).

PI paralogs diversified biochemically

We analyzed whether both *ThPI* paralogs diverged in expression pattern, but did not find any difference in their spatiotemporal expression patterns. It was reported however, that these genes differ in their level of expression (Cheng et al., 2013).

Although the *ThPI* paralogs did not subfunctionalize in expression pattern, we did find functional differences between the two proteins. In our heterologous expression experiment in *A. thaliana*, only ThPI-1, but not ThPI-2, was able to homeotically transform sepals into petaloid structures. This indicates that the proteins may have different functions, although this only was tested in a heterologous system so far. Subsequently, we performed two *in vitro* assays to determine whether the TF protein properties are different: EMSA to determine protein complex specificity and SELEX-seq to investigate the DNA binding specificity. The EMSA results show that ThPI-1 has a higher affinity for some tetrameric complexes than ThPI-2. However, this does not only depend on the PI paralog, as we obtain different results depending on the interaction partners, most importantly the different *Th*SEP paralogs. Although *SEP* genes show extensive redundancy, in some species they display differences in expression patterns and protein-protein interactions, and examples of non-redundant roles for *SEP* genes in flower development have been reported (Ferrario et al., 2003; Malcomber and Kellogg, 2005; Cui et al., 2010; Ruokolainen et al., 2010; Pan et al., 2014). Whether the differences we found between both ThPI paralogs in interactions with the ThSEP paralogs could be significant for flower development depends on whether the expression of these *ThPI* and Th*SEP* genes overlap. Although we were able to isolate all six *T. hassleriana SEP* genes from inflorescences (we cloned all six *SEP* coding sequences), and know that there are differences in expression level in mature flowers (**Figure S6** (Kulahoglu et al., 2014)), we do not have detailed spatiotemporal expression information of the different *SEP* genes throughout floral development. To evaluate whether the observed differences in protein-protein-interactions could make a difference *in planta*, it would be informative to elucidate expression patterns of all *SEP* paralogs. According to the Floral Quartet-model (Theissen and Saedler, 2001), a specific tetramer is formed in each type of floral organ, which binds to two adjacent binding sites in regulatory regions of target genes. The composition of the tetramer determines in part the specificity for a particular target sequence. If however target genes are also controlled by heterodimers alone, then the specificity of the AP3/PI dimers for DNA should be different between the two PI paralogs, to be able to regulate different genes.

We determined DNA-binding specificity for the *T. hassleriana* AP3-1/PI-1 and AP3-1/PI-2 heterodimers *in vitro* using SELEX-seq, and found slightly different binding motifs for the two AP3-1/PI heterodimers. Both PI/AP3-1 heterodimers bind to CArG-boxes, as expected for

MADS-domain proteins. As far as we know, these are the first *in vitro* DNA-binding data for an AP3/PI dimer of any species. The only available *in vitro* data is a study in which they studied binding affinity of *A. thaliana* AP3/PI to different DNA probes using EMSAs (Riechmann et al., 1996a), The only DNA-binding motif for AP3/PI which is published is determined using ChIP-seq of the *A. thaliana* AP3/PI heterodimer (Wuest et al., 2012). The motifs we obtained for the *T. hassleriana* PI paralogs are more similar to each other than to this *A. thaliana* motif, with especially the cytosine on position 1 and 2 of the CArG-box being more conserved in our *T. hassleriana* motifs than in this published *A. thaliana* motif. However, these differences between the *A. thaliana* motif and our *T. hassleriana* motifs might be due to differences in methods used to obtain these motifs. SELEX determines the DNA-binding specificity of the heterodimer to unmethylated DNA *in vitro*. In contrast, ChIP-seq is an *in vivo* method, where sequences might be bound indirectly, DNA might be methylated, and the AP3/PI heterodimer is likely part of a larger protein complex. Both DNA methylation and interaction with cofactors can influence DNA-binding specificity of TFs (Slattery et al., 2011; O'Malley et al., 2016). Interestingly, we did see subtle differences in specificity between two *T. hassleriana* ThAP3-1/ThPI heterodimers. This may indicate that these paralogs could regulate different targets. That paralogous TFs can exhibit differences in binding specificity has also been shown for another plant TF, LEAFY (LFY). LFY is an important regulator of floral identity, and is present as a single-copy TF in most plant species, with the exception of the gymnosperms. Gymnosperms have two paralogs, LFY and NEEDLY (NLY). SELEX-seq experiments on these paralogous proteins from *Welwitschia mirabilis* showed that LFY and NLY have different, although overlapping DNA-binding specificities (Moyroud et al., 2017).

The differences we observed in DNA-binding specificity should be experimentally validated. This could be done *in vitro*, for instance with quantitative EMSAs. To determine whether these differences are relevant *in vivo*, it would be interesting to perform ChIP-seq experiments with these ThPI paralogs, to determine whether they bind to different sites in the genome.

Although we found differences between the ThPI paralogs in protein-protein interactions and in DNA-binding specificity, we did not investigate whether these TF properties between the ThPI paralogs lead to divergence in function. Published data from *PI* duplications in other species show a range of evolutionary possibilities. In some cases, the genes are redundant, as is the case for the petunia and tomato *PI* paralogs. In *Nicotiana benthamiana*, the situation is slightly different as both *PI* genes are necessary for petal and stamen specification (Vandenbussche et al., 2004; Geuten and Irish, 2010). In the Solanaceae species *Physalis floridiana*, as well as in *Medicago truncatula*, the paralogs diverged more substantially, as only one of the *PI* paralogs seems necessary for petal and stamen specification (Benlloch et al., 2009; Zhang et al., 2014; Zhang et al., 2015; Roque et al., 2016). However, at least for the *Medicago truncatula PI* paralogs, it was shown they were both still under purifying selection, arguing against one paralog being in the process of becoming a pseudogene (Roque et al., 2016).

To elucidate how the Th*PI* paralogs evolved in function, functional studies need to be done in *T. hassleriana*. In the absence of mutants, this type of functional data can be obtained using Virus Induced Gene Silencing (VIGS). A first attempt, using Tobacco Rattle Virus (TRV) as vector failed, possibly because *T. hassleriana* is not a good host. Alternatively, transformation could be used to generate CRISPR/CAS9 mutants, however, transformation protocols first need to be developed for *T. hassleriana.*

Importance of synteny

The synteny of the *ThP*I paralogs is interesting: generally, the genomic location of *PI* is conserved throughout the angiosperms. However, the duplication that led to the *ThPI* paralogs transposed one of the *PI* copies into a different genomic location. Whereas *ThPI-2* shares very conserved synteny with *PI* orthologs from the rest of the eudicots, *ThPI-1* is situated in a different genomic location, which it shares with the Brassicaceae. Whether or not this transposition of *ThPI-1* influenced the regulation or function of the gene is an interesting question.

Although closely related plant species show extensive genome colinearity (Consortium, 2005; Cannon et al., 2006; Timms et al., 2006; Lescot et al., 2008; Jung et al., 2009; Hu et al., 2011), plant species that are more diverged do not show large amounts of synteny conservation. However, microsynteny of small genomic regions (of several genes) can be found between distant plant lineages, with examples found even between rice and *Arabidopsis* (diverged 200 Mya) (Yan et al., 2004; Mudge et al., 2005; Timms et al., 2006). Interestingly, conservation of microsynteny is not uniform over the genome (Gebhardt et al., 2003; Yan et al., 2004; Mudge et al., 2005). This might indicate that synteny is more important for certain genomic regions, and possibly certain genes. A published example of conserved synteny is the *ovate* gene (important for fruit shape in tomato), of which the synteny has been conserved for at least 125 million years, between coffee, tomato and grape (Guyot et al., 2012). Although this conservation was stronger than expected, the functional importance of this observation has not been proven. Both B-class genes show extreme synteny conservation. *PI* is conserved in synteny in most angiosperms, except the Brassicaceae. The Cleomaceae are an intermediate form, having one *PI* paralog that is syntenic with most other angiosperms, and the other one shares its position with the Brassicaceae. Intriguingly, a similar situation is observed for *AP3*, which is located in a different genomic position in the Brassicaceae (not Cleomaceae), compared with the conserved location in other angiosperms. Although the fact that both *PI* and *AP3* in the Brassicaceae are in a different genomic region than these B-class genes from the rest of the eudicots is fascinating, whether this actually has functional importance is unknown.

Future directions

Here, we show that the proteins encoded by the two *ThPI* paralogs possess overlapping and distinct TF properties. Both paralogs have similar expression patterns, but the "transposed" copy ThPI-1 seems to have an higher affinity to form certain protein complexes than ThPI-2, and only ThPI-1 caused a phenotype in our heterologous expression experiments. However, we did not reveal whether these differences have an impact on floral morphology in *T. hassleriana*. Functional studies in the species could elucidate if and how these paralogs diverged in function during flower development.

A more general question is what effect, if any, the transposition of *PI* had on gene function. The transposition of *PI* likely occurred at the origin of the Brassicales, around 70 Mya (Cheng et al., 2013). We know that in the Brassicaceae only the new paralog is retained, whereas the Cleomaceae retained both copies. Not only *T. hassleriana*, but also the distantly related Cleomaceae species *G. gynandra* has retained both copies of *PI* (**Figure S4**) (Kulahoglu et al., 2014; Patchell et al., 2014). Interestingly, the "original" and the transposed *PI* paralog diverged in sequence, with the transposed paralog (Brassicaceae *PI*) containing a less conserved PI-motif, and a six amino acids extension compared with the *PI* orthologs of most other eudicots. It would be interesting to create a detailed phylogeny for *PI* in the Brassicales, analyzing more Cleomaceae and Brassicaceae species, as well as species in the other families within the Brassicales. Such a phylogeny could help answering questions about the evolutionary history of *PI*, such as when exactly the transposition took place, when one copy was lost in the Brassicaceae, and whether this copy was lost in more Brassicales families. A more detailed phylogeny could also be used to analyze whether the selection pressure on both paralogs was similar after the duplication.

Fascinatingly, also the genes encoding the obligate heterodimerization partner of PI, AP3, shows high synteny conservation throughout the angiosperms, but not in the Brassicaceae. We know that this transposition occurred after the split between the Brassicaceae and Cleomaceae. Therefore, it would also be interesting to generate a more detailed phylogenetic tree for *AP3* in the Brassicaceae and its closest relatives.

There are some differences in the AP3-PI heterodimerization between eudicot species. For instance, *A. thaliana* AP3/PI heterodimers cannot be observed in Y2H experiments when the full-length proteins are used, although a ternary complex is possible when SEP3 is added (Yang et al., 2003; Immink et al., 2009). In contrast, full-length AP3/PI from petunia and tomato do show interaction in yeast-two-hybrid experiments (Vandenbussche et al., 2004; Leseberg et al., 2008). In addition, in the gel-shift assays performed in this study, the signal obtained for the *A. thaliana* dimer is not as strong as that obtained for the *T. hassleriana* B-class heterodimers, again indicating that the *A. thaliana* B-class proteins behave different from B-class proteins from other species. It would be interesting to compare DNA-binding properties of dimers from a wide range of Brassicales species to pinpoint when these differences originated, and whether they are correlated with the transposition events of *AP3* and *PI*.

Because expression of a gene depends on its genomic environment, the transposition of *ThPI-1* to a new, Brassicaceae-specific location could have led to changes in expression. However, across the angiosperms the expression pattern of *PI* is generally conserved, also in the Brassicaceae. The fact that the expression pattern of *PI* in eudicots is conserved does not mean that this expression pattern is regulated in a conserved way. This is exemplified by *A. thaliana AP3* and *PI* themselves; although they share very similar expression patterns, they do not share similar non-coding regions, and are thought to be activated independently (although expression of both genes is maintained by auto-activation through the AP3/PI dimer) (Jack et al., 1992; Schwarz-Sommer et al., 1992; Goto and Meyerowitz, 1994; Jack et al., 1994; Krizek and Meyerowitz, 1996b; Riechmann et al., 1996b; Hill et al., 1998; Honma and Goto, 2000).
Several studies report about the activity of the promoters of both *AP3* and *PI* in *A. thaliana* and how their expression is maintained. These studies gave detailed information of which promoter fragments are necessary for which part of the expression pattern. Furthermore, some transcription factors were implicated in the regulation of both B-class genes (Hill et al., 1998; Honma and Goto, 2000), but how the expression of these genes is exactly initiated is still enigmatic. Whether the regulation of these genes is conserved throughout the eudicots is also unknown. It could be that the transposition of *PI* led to a different regulation of this gene in the Brassicaceae. Therefore, it is interesting to investigate whether the transposition changed anything in the regulation of *PI*. This could initially be investigated using *in silico* promoter studies, to analyze whether the promoter of the transposed *PI* is significantly different from the “conserved” *PI* promoter. Experimentally, a transgenic approach could be used where promoters of *PI* genes from different species are fused to a reporter gene, and transformed into *A. thaliana*. Even better would be to study the regulation of each PI paralog in the endogenous species by deletion of promoter fragments using CRISPR/CAS9, but this requires an efficient transformation protocol. These experiments together could shine light on the consequences of losing the conserved synteny of *PI* in the Brassicaceae.

Materials and methods

Plant growth

Tarenaya hassleriana was grown in the greenhouse with an average of 22 °C/day and 18 °C/night. Humidity was around 50%.

Arabidopsis thaliana was grown at 20 °C on rockwool under standard long day (18h/6h) conditions.

Alignments and phylogeny

To calculate percentage identities and similarities between the paralogs, http://imed.med.ucm.es/Tools/sias.html was used, with standard settings.

Alignments were made using Muscle. ML phylogenies with 1000x bootstrap were made using Mega6 and the default settings. Boxshade was used for the shading of the alignments.

Sequences used were: *Arabidopsis thaliana*, At3G54340 and At5G20240; *Arabidopsis lyrata*, XM_002877924 And XM_002871885; *Capsella rubella*, XM_006292532 and XM_006288594; *Aethionema arabicum*, AA1026G00001 and AA8G00136, genome version V2.5; *Tarenaya hassleriana*, Th2v17263, Th2v17264, modified Th2v21500 and Th2v23456 (genome version 5); *gynandropsis gynandra* Ggy15517, Ggy19834 and Ggy29007 (genome version V3, unpublished) ;*Carica Papaya*, EF562500; *Theobroma Cacao*, XM_007017619 and XM_007019158; *Populus trichocarpa*, XM_002300928 and XM_002307424; *Vitis Vinifera*, EF418603 and NM_001280946.

RNA isolation and cDNA synthesis

RNA was isolated from Tarenaya inflorescences using the RNeasy plant mini kit (Qiagen) according to the manufacturer's instructions, followed by DNAse treatment (Turbo DNA-free, Ambion). cDNA was made using the RevertAid H Minus first strand cDNA synthesis kit (Fermentas) and a custom primer (5"GGCCAGGCGTCGACTAGTACTTTTTTTTTTTTTTTT 3").

RNA *In situ* hybridization

3'RACE was used to determine the sequence of the 3'UTR. Fragments were obtained by PCR using the 3'RACE primer (GGCCACGCGTCGACTAGTAC) and a gene-specific primer, followed by a PCR with a nested gene-specific primer (primers see **Table 1**). The obtained fragments were cloned into PCR®2.1 TOPO® (ThermoFischerScientific) and sequenced.

RNA *In Situ* hybridization was performed as in (Nardmann et al., 2007). Sequences downstream of the MADS-domain were used as probe (primers used can be found in **Table 1**). These sequences were cloned into PCR2.1® TOPO® (ThermoFischer Scientific) under the T7 promoter and used to prepare digoxigenin-labelled RNA probes. Pictures were taken with a Nikon Optiphot microscope using Nomarsky microscopy, and processed with Photoshop CC 2015.

Table 1: Primers used for the RNA *in situ* hybridisation experiment. Primers to obtain the 3'UTR as well as to generate the *in situ* RNA probes are shown.

in situ probes	Fw	R
AP3	CTCTCCATTCTCTGCGACGCTAG	CATCAAGCTAGGTTTTTCAACTCC
PI-1	GCTCTCCTTCAATGGATCTTGGTG	CACTTATGTCCAAGTCCTTGCAGAG
Pi-2	GATCACTGTTCTATGCGACGCC	GAAACACGCAACGAACCTTGTC
3"RACE	first PCR	nested PCR
ThAP3-1	CTCACTACGAAAGGATGCAAGAGAC	GAAGTTTAAATCGATTGGCAGCC
ThAP3-2	CCTCTCACTACGAAAGGATGCAG	CGATTGGCAATAAAATTGAAACC
ThPI-1	GAGCAGTATCAAAGGATCGCC	GGCCATAGAGCACGCAGTCC
ThPI-2	GAGATGTTGGGCACTTATCAGC	CAAAAGCCTAATCGCCATAGAGAG

EMSAs

T. hassleriana genes were amplified from cDNA and cloned into pSPUTK (primers shown in **Table 2**). Proteins were synthesized using the TnT® SP6 High-Yield Wheat Germ Protein Expression System (Promega) according to the manufactures instructions, using a total of 60 ng plasmid/µl reaction. Proteins for interaction assays were always co-translated, using equimolar amounts of the different plasmids. EMSAs were performed as described in (Smaczniak et al., 2012b) with minor modifications. The fluorescent dye DY-682 was used to label the oligonucleotides. Labelled oligonucleotides were produced in a PCR using vector-specific DY-682-labelled primers, and purified from agarose gel. The binding mix was modified by replacing the glycerol with loading dye. This modification changed the concentration EDTA from 1.2 to 2.2 mM, and added 1 mM Tris-Hcl (pH 7.5), 6.5% sucrose and 0.03% Orange G. For the higher-order complexes, a 4.75% gel was used.
Gel-shifts were visualized with the LiCor Odyssey at 700 nm.

SEP3 probe (pGEM-T sequence underlined):
5'CATGGCCGCGGGATTTTGACGATAACTCCATCTTTCTATTTTGGGTAACGAGGTCCCCTTCCCATTACGTCTTGACGTGGACCCTGTCCGTCTATTTTTAGCAGAATCACTAGTGCGGCCGC-3';

Table 2: Primers used to generate pSPUTK constructs used for *in vitro* protein production.

pSPUTK cloning		
gene	fw	R
AP3 (both AP3-1 and AP3-2)	GATAGATCTATGACGAGGGGAAAGATTCAG	GATAGATCTTCATTCGAGCAAGTGGAAGG
PI-1	TTACCATGGGGAGAGGAAAGATAGAG	ATTATCGATCAGTCGATGACCAAAGACATGATC
PI-2	AATCCATGGGAAGAGGGAAGATAGAGATCAAAAG	TTTATCGATCAGACGATGTGTTGTAAATTGGGC
Th2954 (AG)	ACGGCGTACCAAACGGAGTTG	TTACACTAACTGAAGTGGAGTGTG
Th1528 (SEP3)	CATGCCATGGGAAGAGGTCGTGTTGAG	AAGATCGATCAATTGTTGTCATAAGGTAACCAAC
Th18678 (SEP3)	CATGCCATGGGGAGAGGTCGAGTTG	AAGATCGATCAATTGTTGTCGTAAGGTAACCAAC
Th21984 (SEP4)	ATGGGAAGAGGGAAAGTGGAGC	TCAGATCATCCAGCCGTGGAA
Th2854 (SEP1/2)	ATGGGGAGGGGTAGGGTTG	TCAGAGCATCCAACCAGGG
Th13754 (AP1)	ATGGGAAGGGGAAGGGTTCAG	TTATGTGAAGCAGCCAAGGTTGCAATC

Overexpression

Coding sequence was cloned (under the 35S promoter) into pB7WG2 via PCR8. Primers used in **Table 3**. Constructs were transformed into *A. thaliana col-0* by floral dip.

RNA was prepared from leaves of all transgenic lines using the Invitrap spin plant RNA mini kit (Stratec) according to the manufacturer's instructions. cDNA was prepared using the iScript cDNA synthesis kit from Biorad, according to the manufacturer's instructions. Expression levels were determined by qPCR.

Table 3: Primers used for heterologous expression experiment.

	F	R
ATPI	ATGGGTAGAGGAAAGATCGAG	TCAATCGATGACCAAAGACATAATC
PI-1	ATGGGGAGAGGAAAGATAGAG	TCAGTCGATGACCAAAGACATG
PI-2	ATGGGAAGAGGGAAGATAGAG	TCAGACGATGTGTTGTAAATTGG
qPCR AtPI	GATCATGATGGGCAGTTTGGATATAG	TCGATGACCAAAGACATAATCTTTTCC
qPCR ThPI-1	GACATCCAGTCCCTGGACATC	TCCGTTTCTACGTTTCGTC
qPCR ThPI-2	GACATCCAATCTATGAACTAT	CTCCCTTCTCTTCAATTCT
qPCR TIP41 (reference)	GTGAAAACTGTTGGAGAGAAGCAA	TCAACTGGATACCCTTTCGCA

Supplemental material

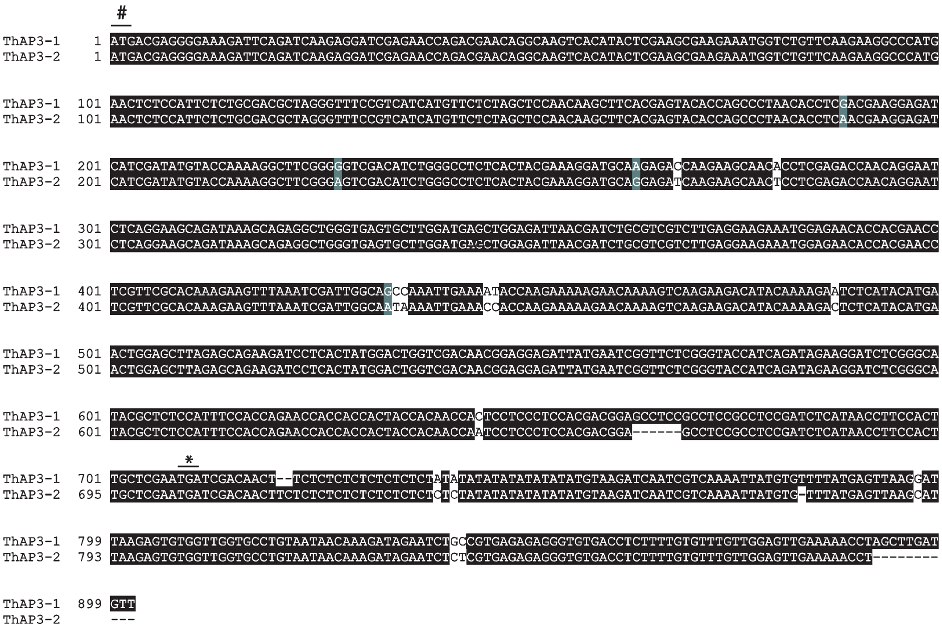


Figure S1: Nucleotide alignment of T. hassleriana AP3 paralogs. *Both coding sequence and 3' UTR are shown. # indicates the startcodon, whereas the * indicates the stopcodon.*

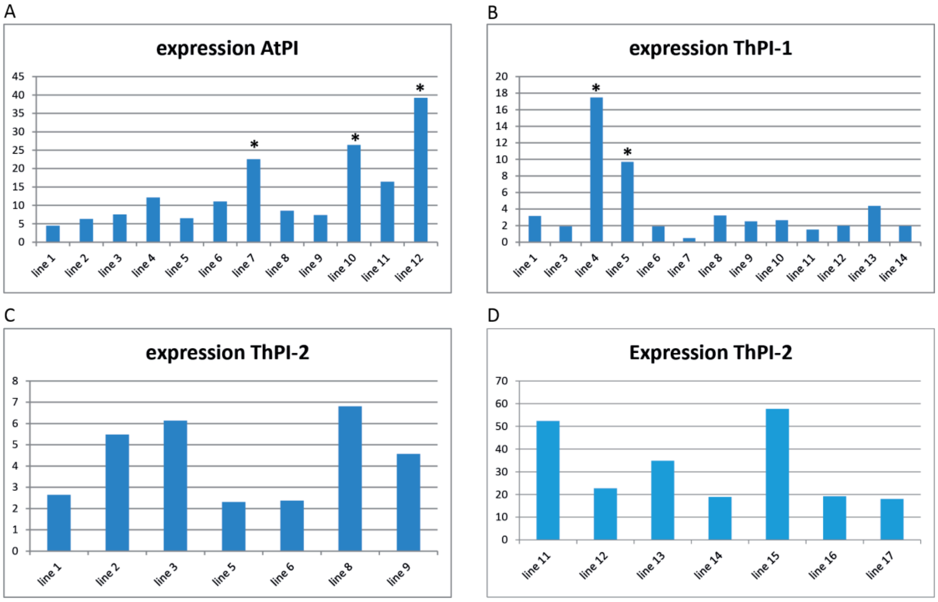


Figure S2: qPCR derived expression levels of the transgenic PI in our overexpression lines. *Expression measured in leaves,* and calculated relative to a reference gene (TIP41). Lines that showed an overexpression phenotype are indicated with an asterisk. ***A:*** *AtPI lines.* ***B:*** *ThPI-1 lines.* ***C:*** *ThPI-2 lines.* ***D:*** *ThPI-2 lines, experiments done in a different lab.*

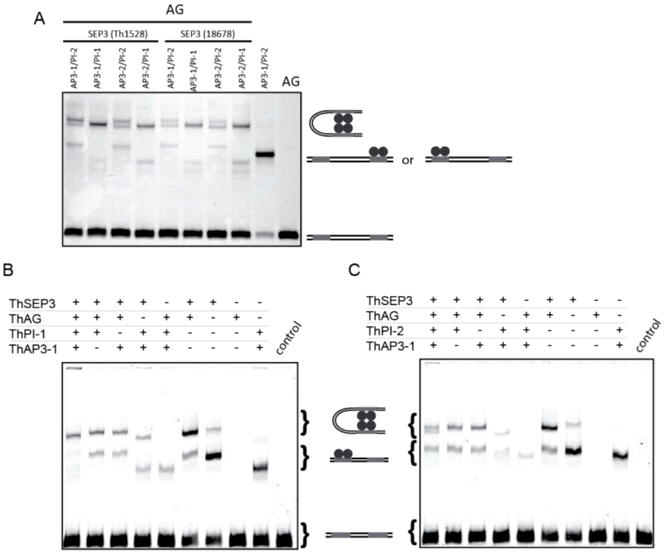


Figure S3: EMSAs to test for higher order complexes containing AP3/PI heterodimers. A: *combinations of each of the four heterodimers with AG and one of the two SEP3 paralogs (Th1528 on the left, Th18678). The two SEP3 paralogs gave similar results. We studied the interaction of AG and one of the SEP3 paralogs (Th1528) and the B-class heterodimers in more detail for AP3-1/PI-1* ***(B)*** *and AP3-1/P-2* ***(C)*** *(see also Figure 3B).*

Table S1: Estimates of the isoelectric point of the four *T. hassleriana* B-class proteins.

	http://pepcalc.com/		http://isoelectric.ovh.org/calculate.php	expasy compute PI/MW
	Ph 7 charge	isoelectric point	isoelectric point	isoelectric point
AP3-1	3.5	8.54	7.73	8.39
AP3-2	4.3	8.91	7.96	8.69
PI-1	4.4	9.14	8.04	8.93
PI-2	5.2	9.12	8.12	9.01

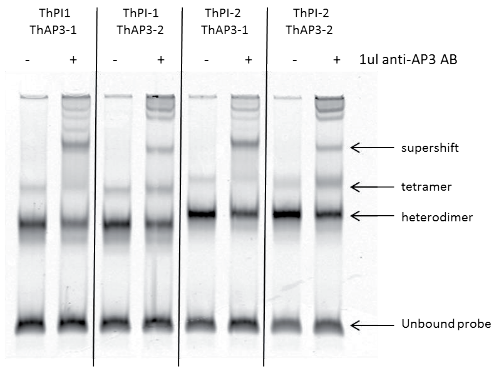


***Figure S4: Arabidopsis thaliana anti-AP3 antibody (AB) recognizes the Tarenaya hassleriana AP3 paralogs**. Recognition of all four B-class heterodimers by the A. thaliana AP3 antibody was assessed on EMSA. For each heterodimer, a supershift of the complex is observed when the AB is added (right) compared to the no AB control (left).*

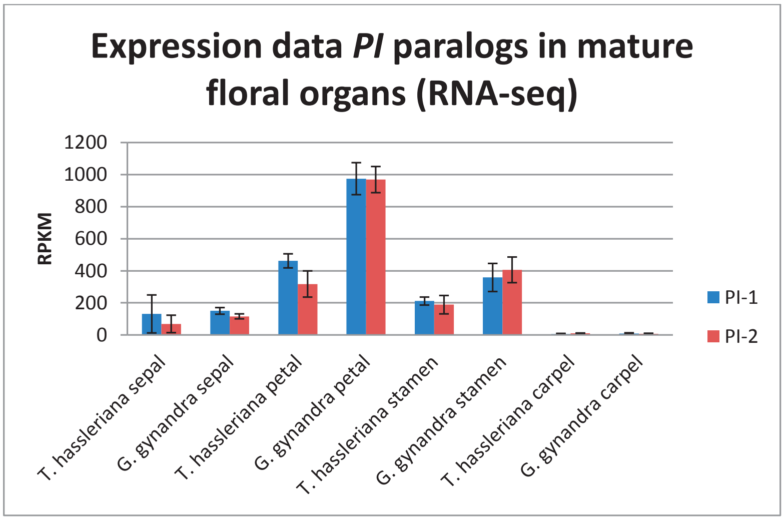


***Figure S5: Expression levels of PI-1 and PI-2, according to RNA-seq data of different mature floral organs.** RNA-seq data obtained from (Kulahoglu et al., 2014).*

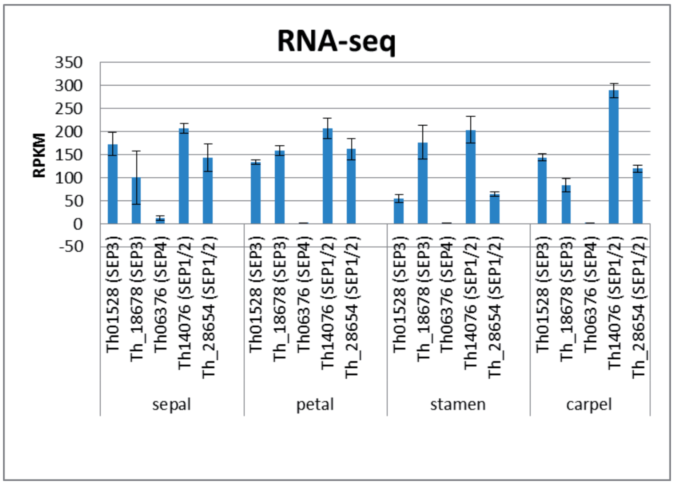
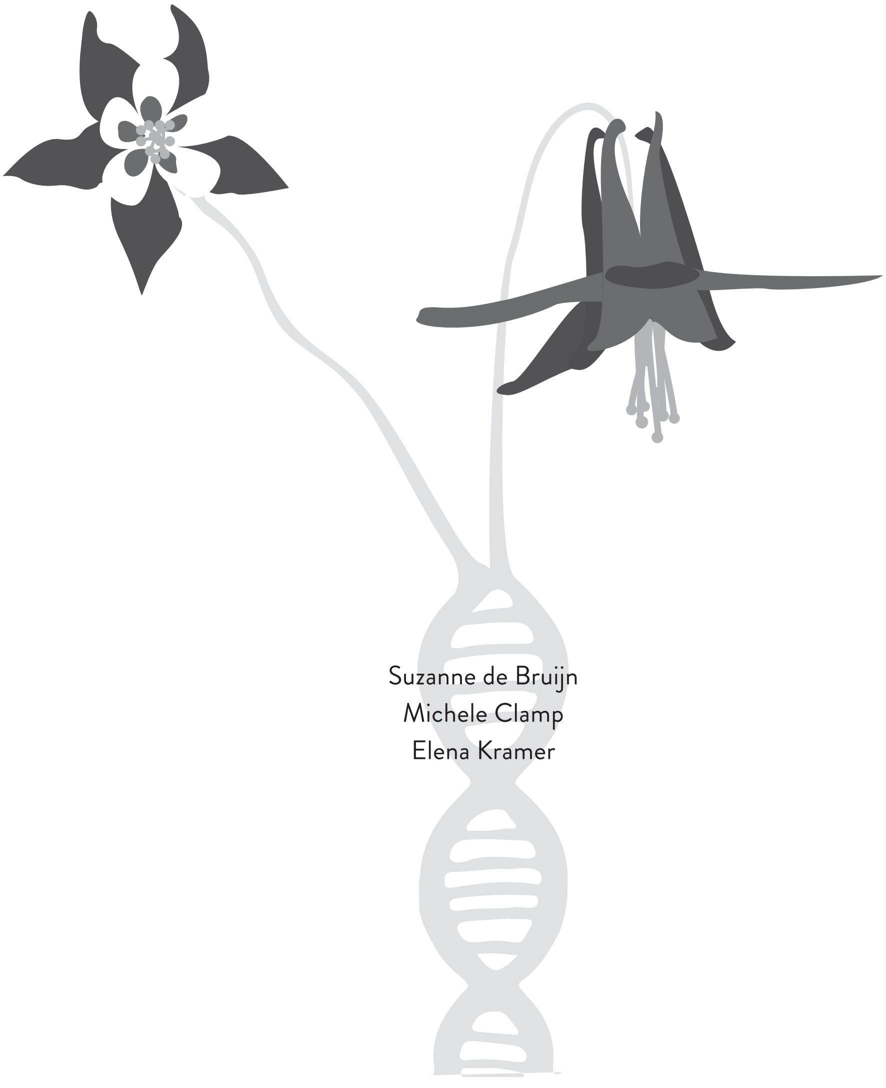


Figure S6: SEP paralog expression data in mature flowers. *Th21984, (a SEP4 paralog) was not present in the dataset. The other SEP4 paralog is hardly expressed. SEP3 and SEP1/2 are both expressed, but expression levels differ between paralogs and between organs. Data from (Kulahoglu et al., 2014).*

CHAPTER 6

Specifying a novel floral organ in *Aquilegia*



Suzanne de Bruijn
Michele Clamp
Elena Kramer

Abstract

Most flowers consist of four different types of organs: sepals, petals, stamens and carpels. These different organs are specified by different combinations of the floral master transcription factors, as described in the (A)BCE-model. Although this model can account for the loss of floral organs, it does not account for the specification of additional organ types. The flower of the basal eudicot *Aquilegia* contains a fifth floral organ type, the staminodium, which is positioned between the stamens and carpels. *Aquilegia* has three paralogs of the floral regulator *APETALA3* (*AP3*). These paralogs show sub- and neofunctionalization, and have been strongly implicated in the specification of the staminodia. Of the three paralogs, *AqAP3-1* specifies staminodia, *AqAP3-2* is needed for proper stamen development, whereas *AqAP3-3* is petal-specific. However, exactly how AqAP3-1 and AqAP3-2 can specify different organs remains unknown as there are no obvious differences in interaction partners. Here, we analyzed whether AqAP3-1 and AqAP3-2 diverged from each other biochemically. We found that these paralogs differ in their DNA-binding specificity, as AqAP3-1 binds to a broader range of sequences than AqAP3-2 and AqAP3-3. We also find differences in protein-complex formation between AqAP3-1 and AqAP3-2, although these differences are subtle, and involve the affinity to form complexes instead of the capability to form specific complexes. In summary, we showed that the paralogs AqAP3-1 and AqAP3-2 are different in their properties to interact with DNA and other transcription factors. The observed differences in interactions between AqAP3-1 and AqAP3-2 may have led to differences in target gene selection between the two paralogs, and hence, to the specification of different floral organs. We suggest several experiments to elucidate whether the differences between AqAP3-1 and AqAP3-2 indeed lead to differences in DNA binding sites.

Introduction

The morphological variation seen among flowers is striking, exhibiting all kind of shapes, colors and sizes. Nevertheless most flowers have the same body plan, consisting of four types of organs: starting from the outside of the flower, a whorl of sepals, followed by petals, male stamens and, in the center of the flower, the female pistil composed of carpels. The specification of these four standard organ identities is described by the ABC model of floral development (Haughn and Somerville, 1988; Bowman et al., 1991; Coen and Meyerowitz, 1991). This model specifies that different combinations of A-, B- and C- functions specify the four different floral organ types. A-function alone specifies sepals, the combination of A- and B-function gives rise to petals, B- and C- function together leads to stamen formation whereas C-function alone specifies carpels. Each of these functions is performed by specific lineages of MADS-box transcription factor (TF) genes (Sommer et al., 1990; Yanofsky et al., 1990; Tröbner et al., 1992; Goto and Meyerowitz, 1994). How the different TFs interact with each other to fulfil their function is described by the floral quartet model (Honma and Goto, 2001; Theissen and Saedler, 2001). Although A-function has turned out to be more complicated than proposed in the original model, B- and C-functions are broadly conserved, and in general this model holds well across angiosperms (Causier et al., 2010a).

Although the B- and C-functions are performed by conserved classes of MADS-box genes, these genes have experienced duplication events in many lineages across the angiosperms (Kramer et al., 1998; Kramer et al., 2004). It is postulated that after a duplication, genes are either lost, or undergo sub- and/or neofunctionalization (Ohno, 1970; Force et al., 1999). Subfunctionalization occurs when the two new paralogs divide the ancestral function, whereas gaining a function that was not present in the ancestral gene is called neofunctionalization. For the B- and C-class genes, it has been shown that, although different patterns of subfunctionalization have occurred in particular lineages, collectively the paralogs generally encompass the ancestral function (Ambrose et al., 2000; Nagasawa et al., 2003; Ferrario et al., 2004; Vandenbussche et al., 2004; Whipple et al., 2004; Zahn et al., 2005; Zahn et al., 2006; Geuten and Irish, 2010; Yellina et al., 2010; Dreni et al., 2011; Hands et al., 2011; Sharma et al., 2011; Sharma and Kramer, 2013).

The basal eudicot genus *Aquilegia* (Ranunculaceae) is used to study the evolution of novel floral morphologies. The phylogenetic position of this genus as sister to the core eudicots, as well as a recent (1-3 million years ago (MYA)) adaptive radiation event, makes this genus useful for evolutionary studies (Hodges and Kramer, 2007; Kramer, 2009). However, the genus is also interesting for developmental studies as the flower features several morphological features that are not present in other plant model species. The perianth of these pentamerous, actinomorphic flowers is of interest because the colorful sepals are petaloid in nature, and the petals have a nectar spur (**Figure 1A**). In addition, these flowers have a novel, fifth type of organ, the staminodium. The staminodia are positioned between the several whorls of stamens and the carpels (Munz, 1946; Tucker and Hodges, 2005). Clearly different from the stamens (**figure 1B**), the staminodia are flattened, sterile organs consisting of a central midrib with ruffled laminae extending to either side (Kramer et al., 2007). Although they arise in two whorls of five organs, all ten staminodia interlock during development to form one continuous sheath around the carpel (**figure 1C and 1D**) (Tucker and Hodges, 2005). These organs are very consistent in position, number, and morphology, indicating that they indeed present a novel type of floral organ (Tucker and Hodges, 2005; Kramer, 2009). This new, fifth, organ type seems to have originated recently, in the ancestor of the closely related genera *Aquilegia*, *Semiaquilegia* and *Urophysa* (~6 MYA) (Kramer, 2009; Sharma et al., 2014). Their function is unknown; however, as they stay attached to the developing fruits long after the other organs fall of, they may be involved in protection of the developing fruit (Kramer, 2009).

Although the ABC-model can easily accommodate loss of floral organ types, it does not account for gaining them. Staminodia are thought to be derived from stamens (Munz, 1946). As *Aquilegia* has been shown to have multiple paralogs of the B-class *APETALA3 (AP3)* genes, it was hypothesized that sub- and neofunctionalization of these loci may have allowed the incorporation of a novel organ into the floral bauplan. *Aquilegia* has three copies of *AP3* and one copy of the other B-class gene, *PISTILLATA* (*PI*) (Kramer et al., 2003). The three *AqAP3* paralogs originated through two duplication events at the base of the Ranunculales (70-120 mya) (Kramer et al., 2003; Sharma et al., 2011). The fact that three copies of *AqAP3* have been retained over a long time suggested that they may have subfunctionalized and/or acquired new functions. Functional studies and analysis of expression patterns have shown that this is indeed the case. *AqAP3-3* has subfunctionalized to be petal specific (Kramer et al., 2007; Sharma et al., 2011). *AqAP3-1* and *AqAP3-2* are both expressed in stamen and staminodium primordia. However, from the moment carpel primordia initiate, the expression of *AqAP3-1* and *AqAP3-2* narrows so that, at later stages of development, *AqAP3-1* is specifically expressed in the staminodia, while *AqAP3-2* is expressed only in stamens (Kramer et al., 2007). Functional analyses using Virus-Induced Gene Silencing (VIGS) have shown that *AqAP3-1* is mainly necessary for staminodia formation, as knock-downs of this paralog partially transform staminodia into carpels (**Figure 1E**). However, sometimes the innermost stamens are also affected, with weak carpelloid traits arising in the anthers (Sharma and Kramer, 2013). In *AqAP3-2* knockdown flowers, only the stamens are affected, which show severe anther reduction and development of trichomes, which indicate transition to carpels, while the staminodia are hardly affected (**Figure 1F)**. Knocking down *AqAP3-1* and *AqAP3-2* together has an additive effect, as these flowers show homeotic transformations of both stamens and staminodia into carpels (Sharma and Kramer, 2013). These phenotypes indicate that *AqAP3-1* and *AqAP3-2* are both needed for proper stamen development, whereas staminodia are specified by *AqAP3-1* alone. Thus, although the *AP3* gene duplication events predate the origin of staminodia by 90-100 my, the paralogs have indeed facilitated the evolution of a new floral organ identity in *Aquilegia*.

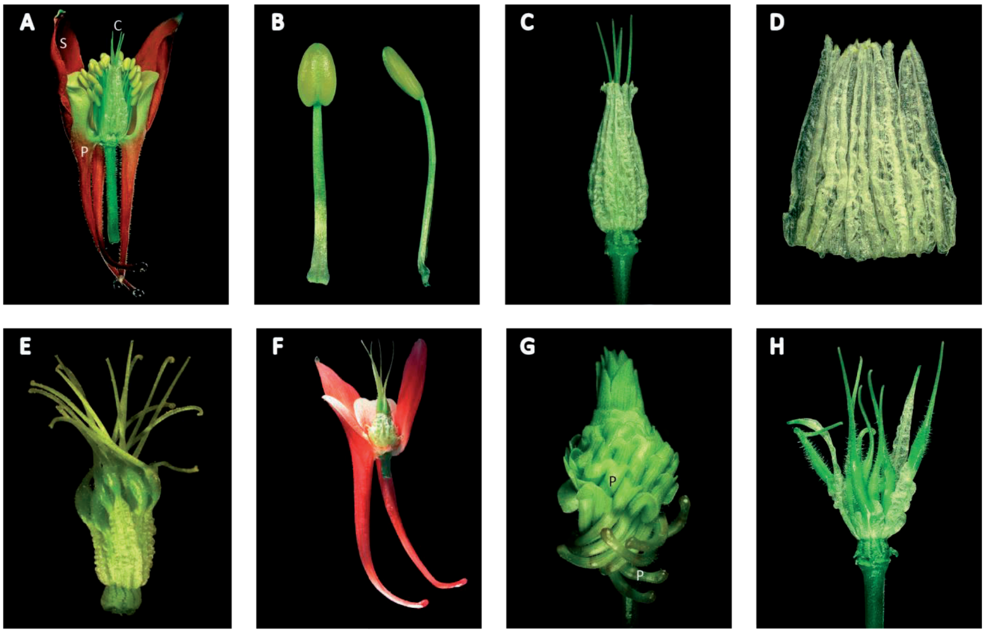


***Figure 1: Floral morphology of Aquilegia**. **A:** WT flower; **B:** WT stamen; **C:** Staminodia, surrounding the carpels; **D:** Staminodia sheath; **E:** Aqap3-1 phenotype, in which staminodia turn carpelloid; **F:** staminodia in an Aqap3-2 silenced flower. Stamens are removed; **G:** Phenotype of AqAG1 knockdown showing stamen to (spurred) petal transformations. Sepals are removed; **H:** Staminodia are transformed into stamens in severe AqAG2 knockdown flowers; sepals, petals and stamens are removed to show staminodia. In **A** and **F**, some sepals and petals are removed to show the inner organs. In **G**, sepals are removed. S=sepals, P=petals, C=carpels.*

The question remains however, as to how AqAP3-1 is capable of specifying a new organ. Although AqAP3-2 and AqAP3-3 are also specifying different organs, this can be explained by the differential expression of their interaction partners (C- versus A-class function). In contrast, there are no obvious differences in interaction partners for AqAP3-2 and AqAP3-1. One of the interaction partners specified by the ABC-model would be the C-class protein AGAMOUS (AG). Interestingly, *Aquilegia* also has two *AG* paralogs (Kramer et al., 2004), but these genes have similar expression patterns in stamens and staminodia (E. M. Kramer, unpublished data). Intriguingly, these *AqAG* paralogs do have distinct functions, as shown by gene knock-down phenotypes. Whereas a reduction of *AqAG1* expression leads to homeotic stamen to petal transformations (**Figure 1G**), knockdown of *AqAG2* leads to green, underdeveloped stamens as well as homeotic transformations of staminodia, which are partially transformed into carpels (**Figure 1H**). This *AqAG2* phenotype of staminodia to carpel transformations resembles *AqAP3-1* knockdown flowers (E. M. Kramer, unpublished data).

Although it seems that the duplications in the *AP3* lineage indeed are involved in the specification of the staminodia, it is still unclear what happens at the molecular level. The fact that *AqAP3-1* and *AqAP3-2* specify different organs despite the fact that there are no obvious differences in the presence of interaction partners suggests that these genes may be biochemically different. Here, we investigate whether there are indeed biochemical differences between these proteins, by analyzing their DNA-binding specificity as well as their capability to interact with the C-class proteins *in vitro*.

Results

The three *Aquilegia AP3* paralogs diverged from one another along their whole sequence, including the MADS-domain (**Figure 2**). As the MADS-domain is needed for DNA-binding and dimerization (Krizek and Meyerowitz, 1996a), we hypothesized that these sequence changes may have led to differences in interaction specificity between the three paralogous *Aq*AP3 proteins. To analyze whether the three AqAP3 paralogs exhibit differences in DNA-binding specificity, we performed Systematic Enrichment of Ligands by EXponential enrichment, followed by high-throughput sequencing (SELEX-seq). This approach consists of incubating *in vitro* translated protein with a dsDNA library, followed by an immunoprecipitation step with an anti-hemagglutinin (HA) antibody. For this experiment, we cloned the coding sequence of all three *Aquilegia AP3* paralogs and *AqPI* into an expression vector. We also created constructs to produce AqAP3 proteins that are tagged with 3xHA at the C-terminus. DNA-binding ability of these *in vitro* produced proteins was analyzed with electrophoretic mobility shift assays (EMSAs) on a known DNA probe (**Figure 3**). Tagged- and non-tagged AqAP3 paralogs gave the same results, although the 3x-HA-tag does raise the gel shift (**Figure 3A, 3B**). These results indicate that the 3xHA tag is not interfering with DNA-binding or heterodimerization. As expected, all three AqAP3/AqPI heterodimers are binding DNA. Whereas none of the three AqAP3 paralogs bind as a homodimer to the DNA probe in this assay, we did observe homodimers for AqPI. However, it seems that the AqPI homodimer is not formed in the presence of AqAP3 (**Figure 3A, B**), as the AqPI-complex disappears when AqAP3 is added to the reaction. This can clearly be observed using the 3xHA tagged AqAP3 paralogs, as the AqPI/AqAP3-HA tagged heterodimers are migrating through the gel slower than AqPI-homodimers (**Figure 3B**). This suggests that although AqPI is capable of homodimerization, it interacts preferentially with AqAP3 (**Figure 3**). Although all three AqAP3 paralogs form heterodimers with AqPI, our experiment does show some differences between the three AqAP3 copies. AqAP3-1/AqPI dimers migrate through the gel slower than the heterodimers with the other two AqAP3 paralogs, which cannot be explained by size or charge of the proteins (**Table S1**). For all three AqAP3/AqPI combinations, a dimer as well as a larger gel shift can be observed, although these larger complexes differ in strength depending on the AqAP3 paralog. These larger complexes could either be AqAP3/AqPI tetramers or two dimers binding to the same DNA molecule, as there are two CArG-boxes present in the used probe. In summary, we find DNA-binding for all three AqAP3/AqPI heterodimers, which is not influenced by the presence of the 3x-HA-tag. We therefore used the 3x-HA-tagged AqAP3 paralogs to perform SELEX-seq on the different AqAP3/AqPI heterodimers.

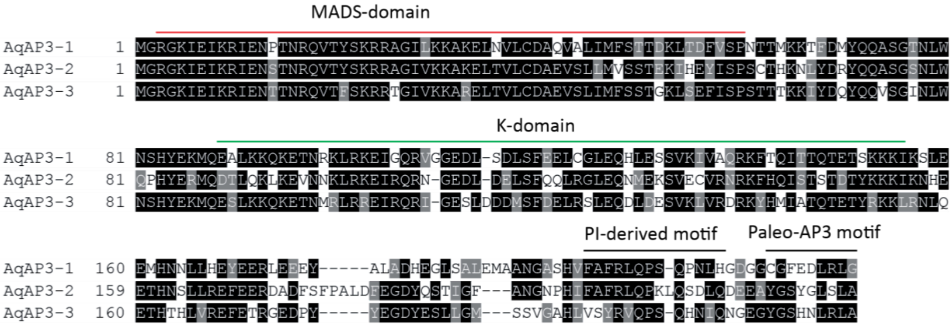

MADS-domain
AqAP3-1 1 MGRGKIEIKRIENPTNRQVTYSKRRAGILKKAKELNVLCDAQVALIMFSTTDKLTDFVSPNTTMKKTFDMYQQASGTNLW
AqAP3-2 1 MGRGKIEIKRIENSTNRQVTYSKRRAGIVKKAKELTVLCDAEVSLLMVSSTEKIHEYISPSCTHKNLYDRYQQASGSNLW
AqAP3-3 1 MGRGKIEIKRIENTTNRQVTFSKRRTGIVKKARELTVLCDAEVSLIMFSSTGKLSEFISPSTTTKKIYDQYQQVSGINLW
K-domain
AqAP3-1 81 NSHYEKMQEALKKQKETNRKLRKEIGQRVGGEDL-SDLSFEELCGLEQHLESSVKIVAQRKFTQITTQTETSKKKIKSLE
AqAP3-2 81 QPHYERMQDTLQKLKEVNNKLRKEIRQRN-GEDL-DELSFQQLRGLEQNMEKSVECVRNRKFHQISTSTDTYKKKIKNHE
AqAP3-3 81 NSHYEKMQESLKKQKETNMRLRREIRQRI-GESLDDDMSFDELRSLEQDLDESVKLVRDRKYHMIATQTETYRKKLRNLQ
PI-derived motif
Paleo-AP3 motif
AqAP3-1 160 EMHNNLLHEYEERLEEEY-----ALADHEGLSALEMAANGASHVFAFRLQPS-QPNLHGDGGCGFEDLRLG
AqAP3-2 159 ETHNSLLREFEERDADFSFPALDFEGDYQSTIGF---ANGNPHIFAFRLQPKLQSDLQDEEAYGSYGLSLA
AqAP3-3 160 ETHTHLVREFETRGEDPY-----YEGDYESLLGM---SSVGAHLVSYRVQPS-QHNIQNGEGYGSHNLRLA

Figure 2: The three APETALA3 paralogs of Aquilegia. *Alignment of the predicted protein sequences of the three Aquilegia AP3 paralogs shows sequence divergence along the length of the protein. The MADS- and K-domains are indicated. The position of these domains is based on the definition of the domains in A. thaliana AP3 (Jack et al., 1992).*

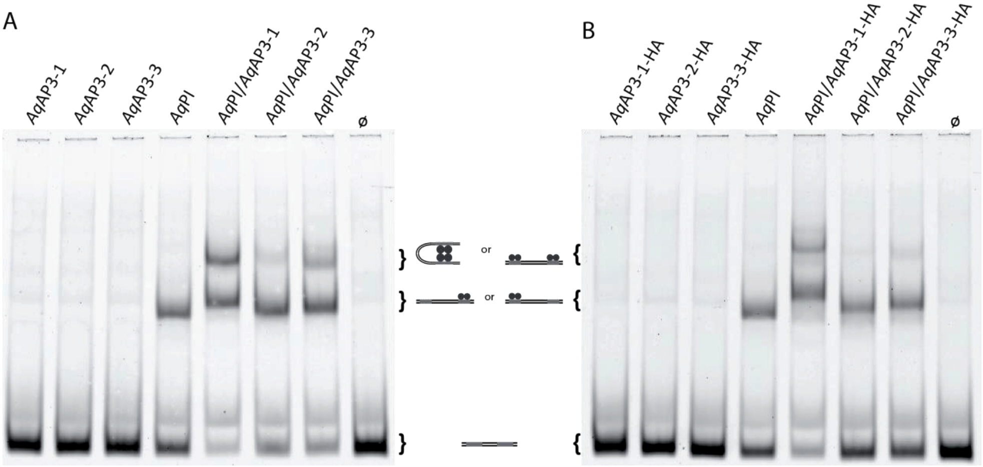


Figure 3: Ability of Aquilegia B-class proteins to form DNA-binding complexes. *EMSAs were used to assess DNA-binding complexes for the three AqAP3 paralogs and AqPI. We analysed both homo- and heterodimerization of the B-class proteins.* ***A:*** *EMSAs using coding sequences for the AqAP3 paralogs and AqPI.* ***B:*** *EMSA determining dimerization of the 3x-HA tagged AqAP3 paralogs. Model representations of the possible formed complexes are shown next to the gel. Protein produced from the empty vector was used as negative control. The DNA probe used is part of the A. thaliana SEP3 promoter and contains two CArG-boxes.*

DNA-binding affinities of B-class heterodimers

We performed SELEX-seq to determine relative DNA-binding affinities of each B-class heterodimer. This experiment consists of incubation of the protein of interest with a dsDNA library. These dsDNA libraries consist of a region of randomized nucleotides, flanked by a barcode for multiplexing, and PCR primer sites for amplification and sequencing purposes. After incubation of proteins with the dsDNA library, TF-DNA complexes were immunoprecipitated using anti-HA antibodies. Subsequently, the obtained DNA molecules are isolated and amplified. Using the obtained DNA as input, the SELEX cycle is repeated until enough enrichment is obtained.

We performed SELEX-seq for all three *Aquilegia* AP3/PI heterodimers in duplicate. After several rounds of SELEX, the enrichment of sequences bound by the heterodimer in each round was analyzed through EMSAs, which show that we obtain visible enrichment of bound sequences in round 4/5 (see **Figure 4A**). Based on these EMSAs, round five was selected for sequencing of all experiments. Between 5.7 and 9.2% of the sequenced libraries contains a perfect CArG-box (CC[A/T]$_6$GG) (**Figure 4B**), the consensus binding site for MADS-domain proteins (Huang et al., 1993). This is in the same range as seen for previous SELEX data (Smaczniak et al., 2017), and suggests that our samples are indeed enriched for DNA sequences bound by AqAP3/AqPI heterodimers.

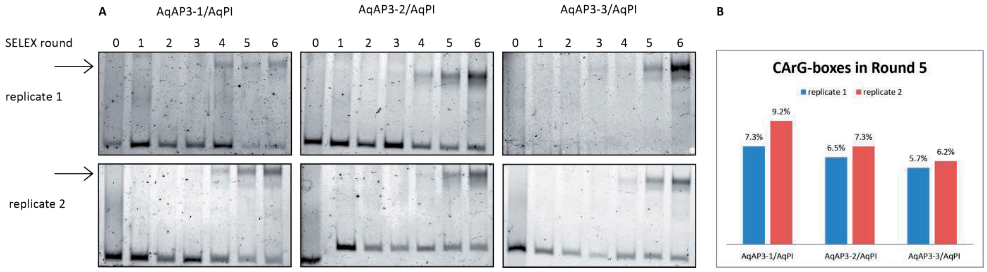


Figure 4: Enrichment of bound DNA-molecules in our SELEX experiments. A: *EMSAs using the output of each SELEX round as probe, showing visible enrichment of bound DNA-molecules starting in round 4/5 of our SELEX experiments (indicated by an arrow).* ***B:*** *Percentage of sequences obtained in SELEX round five that contains a perfect (CC[A/T]$_6$GG) CArG-box.*

We estimated relative affinities of our AqAP3/AqPI heterodimers for each possible 12-bp sequence (12-mer) by comparing the frequencies of these 12-mers in the sequenced round 5 (R5) of the SELEX experiments with the frequency of these 12-mers in the original libraries (R0). The relative affinities for each 12bp sequence are plotted in a heatmap (**Figure 5**). The heatmap shows that the replicates of AqAP3-2/AqPI and AqAP3-3/AqPI cluster together, whereas the AqAP3-1/AqPI SELEX replicates clearly cluster separately. The finding that the replicates of AqAP3-2/AqPI and AqAP3-3/AqPI cluster together suggests that there is no difference in DNA-binding specificity between AqAP3-2 and AqAP3-3 (**Figure 5**). It seems however, that AqAP3-1 is diverged in sequence specificity from the other two AqAP3 paralogs. To look at the differences in DNA-binding in more detail, we divided the sequences present in the heatmap in several clusters. One of these clusters (cluster E) shows high relative affinities for all three AqAP3/AqPI heterodimers. The other four clusters (cluster A-D) all show higher affinity for AqAP3-1 than for the other two AqAP3 paralogs. These data indicate that although the three *Aquilegia* AP3 paralogs share the same DNA binding sites, AqAP3-1 binds to a broader range of sequences than AqAP3-2 and AqAP3-3.

For each of the five different sequence clusters we prepared logos based on multiple alignments. The motif obtained for each cluster resembles a CArG-box, but instead of the consensus sequence CC[A/T]$_6$GG, they have a longer A/T-rich stretch, and only a single C/G on either end (C[A/T]$_8$G). However, there are subtle differences between the motifs from the different sequence clusters. The group of sequences that is most highly bound by all three AqAP3 paralogs (motif E) shows a CArG-box with the consensus CTATATATAG, with the positions in the middle (position 9 and 10 in motif E) showing more variation than the other positions. This motif is also larger than a typical CArG-box, as on both sides there are an additional three base pairs that are conserved.

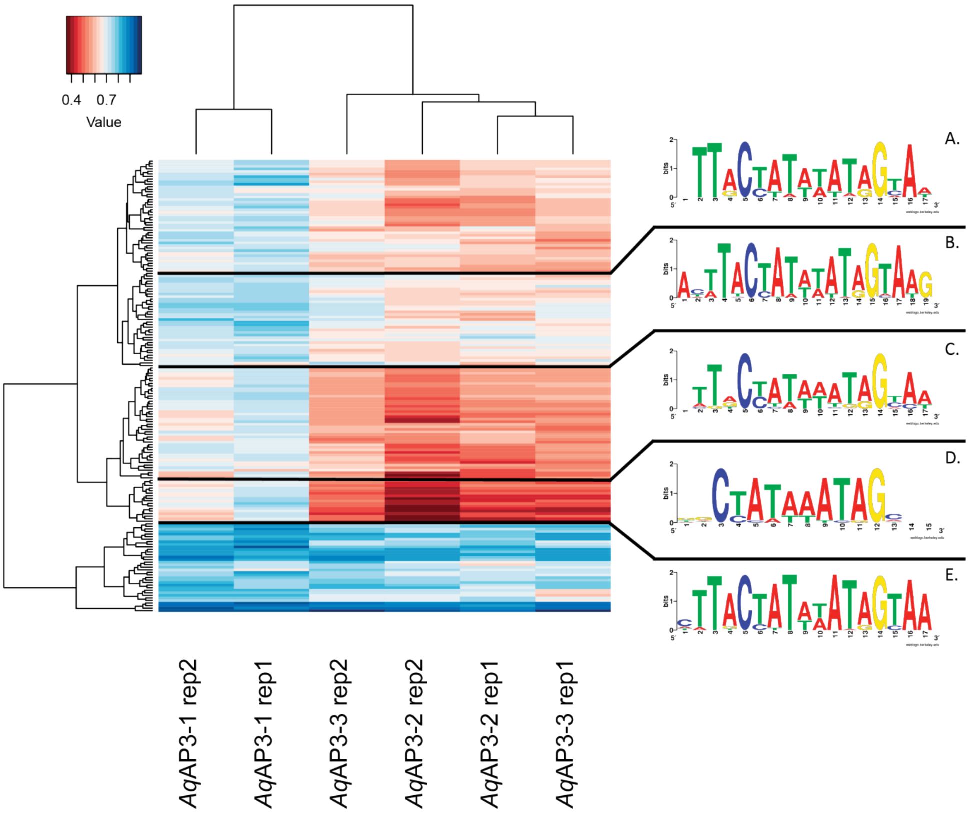


Figure 5: relative DNA-binding affinities of the three Aquilegia AP3 paralogs. *Heatmap of relative affinities calculated for 12-mers based on round five of the SELEX experiments. Plotted is each 12-mer that has a relative affinity of at least 0.7 in any of the samples. Sequence logos of each cluster were made from multiple alignments.*

Both the A/T-rich sequence in the CArG-box and the extensions on both sides gradually become less conserved between the different clusters, with the sequences that are least bound by AqAP3-2 and AqAP3-3 (motif D) having a slightly different A-tract in the middle of the CArG-box (consensus CTATAAATAG), and misses the three nucleotide extension on either side of the CArG-box. These motifs indicate that all three AqAP3 paralogs bind to the same motif, but that AqAP3-1 also binds to more diverged variations of this motif. It seems therefore that AqAP3-1 binds to a broader range of sequences than AqAP3-2 and AqAP3-3. These data show clearly that, although subtle, there are differences in DNA-binding specificity between AqAP3-1 on the one hand, and AqAP3-2/AP3-3 on the other.

Differences in protein-protein interactions

According to the floral quartet model, the proteins in the (A)BCE-model do not act in dimers, but in tetrameric complexes (Theissen and Saedler, 2001). The *Aquilegia* AP3 paralogs do not only differ in their MADS-domain, but also show sequence polymorphisms in the K-domain (**Figure 2**), the domain necessary for tetramerization (Yang and Jack, 2004; Kaufmann et al., 2005; Melzer and Theissen, 2009; Melzer et al., 2009; Puranik et al., 2014). Therefore, changes in the K-domain could lead to differences in protein-protein interactions. To determine whether differences in protein-protein interactions between AqAP3-1 and AqAP3-2 underlie the specification of staminodia, we investigated higher-order complex formation. The staminodia whorl is positioned between the stamen and carpels, which both need the C-class protein AGAMOUS (AG) for correct development. We therefore focused on the interaction of the B- and C-class proteins with the addition of SEPALLATA3 (E-class protein). *Aquilegia* has two AG paralogs, *Aq*AG1 and *Aq*AG2, which as described in the introduction, have distinct effects on the staminodial whorl when silenced.

Using EMSAs, we observed two different higher order complexes for the combination AqAP3/AqPI/AqSEP3/AqAG, regardless of which AqAP3 and which AqAG paralog was present (**Figure 6A**). The relative intensities of these two complexes however, are influenced by the AqAP3 and AqAG paralog analyzed. We always observe more of the higher complex (marked with *) relative to the lower complex (marked with ^) with AqAP3-1, compared to AqAP3-2. We also see differences between the AqAG paralogs. Presence of AqAG2 results in more of the higher complex (*) compared to when AqAG1 is present. In conclusion: we do observe two different DNA-binding complexes for combinations of AqAP3/AqPI/AqSEP3/AqAG regardless of which AqAP3 and AqAG paralog is present. The relative prominence of each complex however, depends on the presence of different AqAP3 and AqAG paralogs.

We attempted to understand the composition of the two observed complexes. We performed dropout experiments where we compared the complexes formed with all four proteins with several other combinations of two or three proteins. It seems that the lower band contains all four proteins (AqAP3/AqPI/AqSEP3/AqAG), as none of the other combinations tested give the same gel shift. The higher complex runs at the same height as both AqSEP3/AqAG and AqSEP3.

However, as AqSEP3 by itself gives a fainter band than the one observed for all four proteins, this higher complex is most likely AqSEP3/AqAG. This means that although we observe two different complexes, only the lower one of these complexes contains AqAP3/AqPI. These observations imply that the two AqAP3 paralogs (and the two AqAG paralogs) are not forming different tetrameric complexes. Instead, it seems that they show different affinities to form protein-protein complexes. A lower affinity of AqSEP3/AqAG for AqAP3/AqPI compared to its affinity for itself will lead to more of the AqSEP3/AqAG tetramer being formed. The relative higher amount of AqSEP3/AqAG tetramer compared to an AqAP3/AqPI/AqSEP3/AqAG complex when AqAP3-1 is used than when AqAP3-2 is present indicates that AqAP3-1/AqPI has a lower affinity to form an AqAP3/AqPI/AqSEP3/AqAG tetramer than AqAP3-2.

AqSEP3 is assumed to be the major *SEPALLATA* gene in *Aquilegia*, since expression of *AqSEP1* and *AqSEP2* is only detected in sepals. However, this is based on expression data from mature organs (Kramer et al., 2007). As detailed expression patterns during development are unknown for the *Aquilegia SEPALLATA* genes, we also tested the formation of B-, C-, E-class complexes with AqSEP2. This combination with AqSEP2 shows a slightly different picture compared with AqSEP3 (**Figure 6B**). There are still two possible complexes, obtained for combinations of AqAP3/AqPI/AqSEP2/AqAG when AqAG2 is present, whereas only the lower complex is observed when AqAG1 is present. However, when AqSEP2 fulfils the role of E-class protein, there is no visible difference anymore between AqAP3-1 and AqAP3-2. This indicates that these AqSEPALLATA paralogs are not equivalent in protein-complex formation, and may not be redundant. Combined, our protein-protein interaction assays suggest that there are subtle differences between AqAP3-1 and AqAP3-2 in the formation of tetrameric complexes. The same was found for the AqAG, and the AqSEP paralogs.

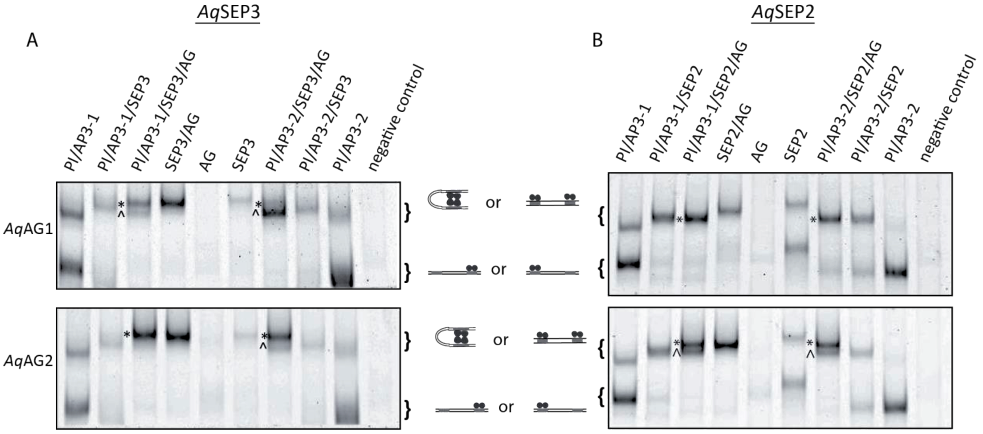


*Figure 6: **Tetrameric complexes formed by AqAP3/AqPI with AqSEP and AqAG paralogs**. EMSAs show two different complexes (marked with * and ^) for combinations of AqAP3/AqPI/AqSEP/AqAG. Several other protein combinations are run on the same EMSA to determine the composition of these two complexes. **A:** AqSEP3 is used as SEP protein. **B:** AqSEP2 is used as SEP protein. DNA probe used is a fragment of the A. thaliana SEP3 promoter containing two CArG-boxes.*

However, the difference between AqAP3-1 and AqAP3-2 seem to be only the affinities with which they form different tetrameric complexes. In total, there are subtle differences in protein complex formation between the AqAP3 paralogs. However, it is difficult to predict which complexes will be formed *in vivo*, as this will most likely depend not only on differences in affinity, but also differences in protein levels.

Structural differences between AqAP3 paralogs

We see differences in both DNA-binding specificity and tetramerization between the AqAP3 paralogs. We next performed structure predictions for each of the three paralogs to attempt to elucidate how the observed differences are generated. The MADS-domain of the three paralogs show a very similar backbone (**Figure 7**). There are some differences in side chains however, most of these are in the betasheet, which is needed for dimerization. The α-helix that interacts with the DNA is more conserved (Pellegrini et al., 1995; Tan and Richmond, 1998; Huang et al., 2000). Although the K-domain of AqAP3-1 seems to be different in structure from AqAP3-2 and AqAP3-3 (**Supplemental Figure 1**), this part of the protein could not be modelled with great confidence.

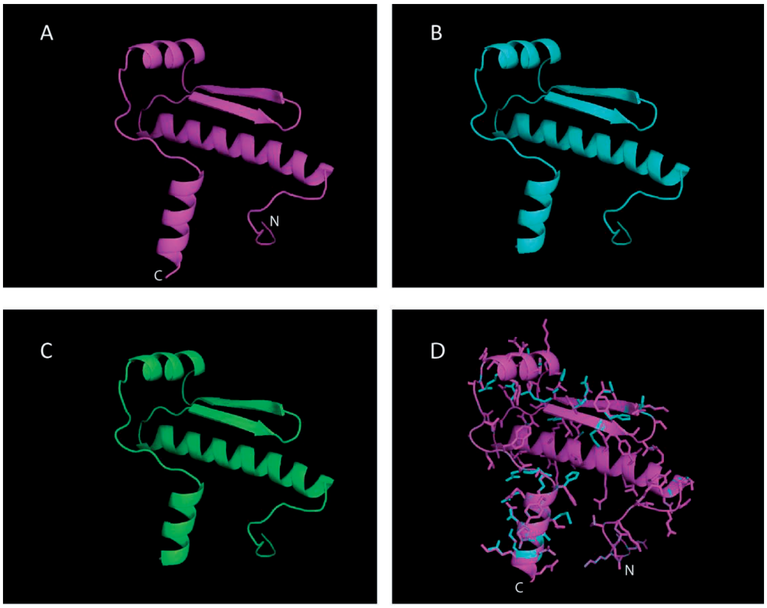


Figure 7: Modelled structures of the three AqAP3 paralogs. *Structures are modelled using Phyre2. (**A, B, C**) MADS-domain of AqAP3-1, -2, -3 respectively. **D**: overlap of the first 90 AA (MADS-domain) of the three AqAP3 paralogs. This shows that the backbone is basically identical. AqAP3-1=magenta, AqAP3-2=cyan, AqAP3-3=green. Proteins were modelled using Phyre2.*

There are no conserved synapomorphies in any of the Ranunculaceae AP3 lines

We show that there are differences in interactions with DNA and proteins between *Aquilegia* AP3-1 and the other AqAP3 paralogs, with AqAP3-1 being more promiscuous in DNA-binding. We wanted to investigate whether these differences could be a general feature for orthologous Ranunculaceae AP3-1 proteins. If promiscuity is a general feature of AP3-1, there could be AP3-1 lineage-specific amino acids in the MADS-domain. To analyze whether there are synapomorphies, we assembled motifs for each of the three AP3 paralogs using sequences from several Ranunculaceae species (**Figure 8**). The motifs we obtained for the partial MADS-domain are very similar for the three AP3 lineages. In the I- and K-domain there are more differences. In the I-domain, position 1, 25 and 26 seem less conserved in AP3-1 compared to the other paralogs, as do position 21 and 42 in the K-domain. However, there are no conserved synapomorphies for AP3-1, or either of the other paralogs.

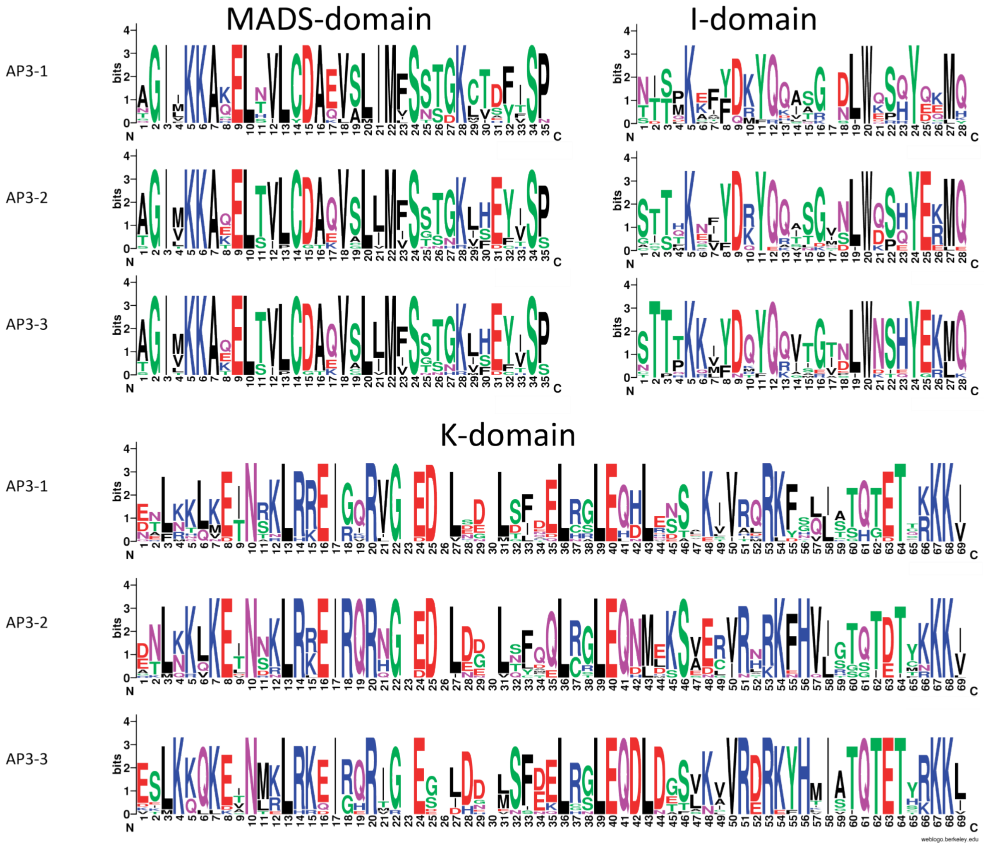


Figure 8: Motifs of the Ranunculaceae AP3s paralogs. *Motifs of the three different AP3 paralogs were calculated from multiple alignments using Weblogo. Motifs are shown for a partial M-domain, the I-domain and the K-domain. M- and K- domains are defined as in figure 2. AP3 paralogs of several Ranunculaceae species were used.*

This means that we cannot conclude anything about the evolution of the specific AP3-1 DNA-binding profile based on sequence conservation. It may be that the DNA-binding profile we find for AqAP3-1 is not general for Ranunculales AP3-1 but instead may have evolved later, for instance in the lineage leading to staminodia-containing taxa (*Aquilegia+ Semiaquilegia+ Urophysa*). It might also be that all AP3-1 proteins share the promiscuity in DNA-binding we found for *Aquilegia* AP3-1, but that this is obtained by a combination of different amino acid changes. To conclude anything about the evolution of the DNA-binding promiscuity of AqAP3-1, as well as what causes these differences between AqAP3 paralogs, we would need to determine the DNA-binding affinities for AP3-1 proteins from additional Ranunculaceae species.

Discussion

Although it is known that the Aq*uilegia* paralogs AP3-1 and AP3-2 specify different organs, it is not yet known how they achieve this. In order to achieve this developmental function, these two paralogs need to regulate different genes. Here we analyzed whether differences in interactions with DNA and/or with other transcription factors could play a role in target gene selection. We analyzed DNA-binding affinities of the AqAP3/AqPI heterodimers, as well as differences in protein-protein interactions.

Using EMSAs, we found that all three AqAP3 paralogs interact with AqPI, as was previously shown with Y2H-experiments (Kramer et al., 2007). Interestingly, AqAP3-1/AqPI shows a larger gel shift than heterodimers containing either of the two other AqAP3 paralogs, which does not seem to be explained by size or charge of the proteins and is possibly caused by another conformation of the AqAP3-1/AqPI dimer. When we studied the DNA-binding affinities of the three different heterodimers, we find that all three AqAP3/AqPI heterodimers bind to a very similar binding motif, which resembles a CArG-box (CC[A/T]$_6$GG), but does not completely adhere to the consensus. Instead of the two cytosines at the beginning of the motif, and the two guanines at the end, we find a motif with a single cytosine at the beginning and a single guanine at the end, with an A/T-rich stretch of eight bases in between (C[A/T]$_8$G). This is in contrast to the SELEX-logos obtained using MADS-domain dimers from *Arabidopsis* (SEP3-AG; AP1-SEP3; AG-AG), where more perfect CArG boxes were found as preferred binding sites, illustrating that different MADS-domain dimers show different sequence binding specificities (Smaczniak et al, 2017). In addition, the motif we find is longer than a classical CArG-box, as the motif is extended by three base pairs (TAA) on both sides. Interestingly, however, we did observe that AqAP3-1/AqPI binds to a broader range of sequences than either AqAP3-2/AqPI or AqAP3-3/AqPI. It seems therefore, that AqAP3-2 and AqAP3-3 have similar DNA-binding specificity, whereas AqAP3-1 binds to similar sites but is more promiscuous.

Until now, few AP3/PI binding sites have been published. The only published AP3/PI motif is based on *A. thaliana* ChIP-seq data, and is a CArG-box with two cytosines and two guanines, although these four bases are not all completely conserved. Similar to the motifs we found for the *Aquilegia* AP3 paralogs, this *A. thaliana* motif exhibits an addition of three base pairs after the CArG-box, in this case three adenosines (Wuest et al., 2012). In Chapter 5 of this thesis, we used SELEX to obtain binding sites for two AP3/PI heterodimers from a species closely related to *A. thaliana*: the Cleomaceae species *Tarenaya hassleriana*. Like the *A. thaliana* motif, these binding motifs also adhere to the CArG-box consensus (although the cytosines and guanines are not completely conserved), and show extensions of the bound motif on both sides of the CArG-box. These differences in DNA-specificity observed for AP3/PI heterodimers from different species are interesting. However, it needs to be taken into account that the motifs were obtained with different experimental methods and with different methods of analysis. The motifs we find here for the basal eudicot *Aquilegia* AP3/PI heterodimer are clearly different from the motifs obtained for the AP3/PI heterodimers from the core eudicots species *A. thaliana* and *T. hassleriana*. Interestingly, the paleoAP3 lineage underwent a duplication at the base of the core eudicots that resulted in the paralogous *euAP3*-lineage and *TM6* clades (Kramer et al., 1998). Both the *A. thaliana* and the *T. hassleriana AP3* paralogs belong to the *euAP3*-lineage, whereas the three *AqAP3* paralogs all belong to the *paleoAP3* lineage. To assess whether the duplication leading to *euAP3/TM6* might have influenced DNA-binding specificity, it would be interesting to analyze whether there are differences in DNA-binding between a broader sampling of euAP3 and TM6 representatives, as well as additional paleoAP3 homologs.

According to the floral quartet model, AP3/PI act together with other floral MADS-domain transcription factors to specify floral organs (Theissen and Saedler, 2001). We analyzed whether there were differences in tetramer formation between AqAP3-1 and AqAP3-2 with the *Aquilegia* AG paralogs, as AG is supposed to be an interaction partner of AP3/PI in stamens, and the AqAG paralogs are also expressed in staminodia. As both AqAP3-1 and AqAG2 are involved in the specification of staminodia, we hypothesized that AqAP3-1 and AqAG2 may interact exclusively. However, that is not what we observed. Both paralogs of AqAP3 and AqAG formed the same complex, but in addition, a complex of presumably AqAG/AqSEP was formed with different intensities depending on the paralog of AqAP3 and AqAG present. These data mean that there are some differences in complex formation between the AqAP3 paralogs, but these differences seem to be more in affinity to form protein complexes rather than in which complexes can be formed. Interestingly, these results are consistent with parallel studies of protein interactions detected in yeast (L. Holappa, E. M. Kramer, unpublished data).

An interesting observation is that there are differences in complex formation with different SEP paralogs. *Aquilegia* has three major SEPALLATAs, AqSEP3 and two genes from the LOFSEP clade, AqSEP1 and AqSEP2, which originated from a duplication at the base of the Ranunculales (Soza et al., 2016). We know that in mature organs, AqSEP3 is ubiquitously expressed, while AqSEP1 and AqSEP2 are expressed primarily in sepals (Kramer et al., 2007). Whether the different SEP paralogs in *Aquilegia* could play a role in staminodia specification is partially depending on their expression during development. However, detailed expression patterns of these genes during floral organ development are not yet known.

In a close relative of *Aquilegia*, *Thalictrum*, the expression of the *SEP* paralogs has been studied. In *Thalictrum*, *SEP3* is most highly expressed. However, *SEP1* and *SEP2* are also expressed in all floral organs, although *SEP1* expression is higher in sepals and stamens (Soza et al., 2016). Interestingly, when these *SEP* genes in *Thalictrum* are silenced, homeotic conversions and chimeric floral organs can be observed. These data indicate that the *SEP* genes in Ranunculaceae can play a role in organ specification and/or organ boundary formation. Whether the SEP paralogs play a role in staminodium-specification could be analyzed by detailed expression studies as well as silencing experiments in *Aquilegia*.
Although our data only show subtle differences in protein complex formation, we cannot exclude that the AqAP3-1 and AqAP3-2 have different interaction partners. We only tested for protein tetramers. However, it is known that MADS-domain TF do not only form quartets, but that they interact with a wide range of proteins (Smaczniak et al., 2012b).

We found here that AqAP3-1 is indeed biochemically different from the other two AqAP3 paralogs. The AqAP3-1/AqPI heterodimer migrates through the EMSA gel slower than the other two AqAP3/AqPI heterodimers. AqAP3-1 also shows differences in DNA-binding specificity, as it binds to a broader range of sequences than the other AqAP3 paralogs. In addition, we observed subtle differences in the formation of protein complexes *in vitro* between AqAP3-1 and AqAP3-2. These data led us to speculate that AqAP3-1 may be different in structure from the other two AqAP3 paralogs. We performed protein structure prediction for the three different paralogs. Although the MADS-domain of the three AqAP3 paralogs looked very similar, the predicted structure for the AqAP3-1 K-domain looked different from the other AqAP3 paralogs. It must be noted, however, that the K-domain could not be predicted with high confidence. Interestingly, using sequences from the Ranunculaceae and Berberidaceae, it was shown that the I+K-domain of AP3-1 is under lower purifying selection than these domains of AP3-2 and AP3-2 (Sharma et al., 2011). It would be interesting to compare structures of the three AP3 paralogs in more detail, but for this analysis the protein structures need to be elucidated.

A remaining intriguing question is how AP3-1 evolved to specify staminodium identity. Although AP3-1 originated from an ancient duplication (70-120 MYA), staminodia evolved only recently, around 6 MYA in the ancestor of *Aquilegia*, *Semiaquilegia* and *Urophysa*. It would be interesting to analyze whether the changes in interactions we observed are coinciding with the emergence of staminodia. A related question is what the functions of AP3-1 and AP3-2 are in non-staminodia species, in which both these paralogs seem commonly expressed in petals and stamens (Rasmussen et al., 2009; Sharma et al., 2011). Examination of Ranunculaceae sequences of the three paralogous *AP3* lineages did not reveal any distinct conserved synapomorphies. It could still be that the differences we observe in DNA-binding between AqAP3-1 vs. AqAP3-2/3 is a general feature for the Ranunculaceae AP3-1, and that they are generated by a combination of different mutations. To test this, DNA-binding affinities for AP3 paralogs of a range of Ranunculaceae species need to be determined. However, it is interesting to note that *AqAP3-1* is evolutionarily more closely related to *AqAP3-2*, being derived from a tandem duplication.

The differences in interactions we observed between AqAP3-1 and AqAP3-2 in this study are only relevant for stamen and staminodia specification if they lead to differences in gene expression between these two organs. Although comparison of expression profiles between stamens and staminodia are interesting, these profiles do not show how the differences are generated. A more direct approach to elucidate whether AqAP3-1 and AqAP3-2 regulate different genes would be to perform ChIP-seq on the different AqAP3 paralogs. This is an elaborate experiment however, as it optimally requires the generation of *ap3* mutants that express tagged AP3 proteins from their native promoter or high-quality antibodies specific for each AqAP3 paralog. Another interesting question is whether the differences in the AqAP3 paralog present are enough to specify different organs, or that other factors, such as the availability of interaction partners or the accessibility of DNA, play a role as well. This could be analyzed by swapping the expression of *AqAP3-1* and *AqAP3-2*. That is, express *AP3-2* from the *AP3-1* promoter (in an *ap3-1* mutant) or vice versa, and analyze the obtained plants for homeotic transformations of staminodia or stamens. A combination of these experiments could hopefully fully elucidate how AqAP3-1 and AqAP3-2 can specify different floral organs.

Material and methods

EMSAs

Aquilegia genes were amplified from cDNA and cloned into pSPUTK (primers shown in table 1). 3xHa-tagged versions of the genes were created in two subsequent PCR reactions with different primers, before cloning them into pSPUTK. Proteins were synthesized using the TnT® SP6 High-Yield Wheat Germ Protein Expression System (Promega) according to the manufactures instructions, using a total of 100 ng plasmid/µl reaction. Proteins for interaction assays were always co-translated, using equimolar amounts of the different plasmids. EMSAs were performed as described in (Smaczniak et al., 2012b) with minor modifications. The fluorescent dye CY5 was used to label the oligonucleotides. Labelled oligonucleotides were produced in a PCR using vector-specific CY5-labelled primers, and purified from agarose gel. Gel-shifts were visualized with the GE Typhoon Trio.

Table 1: primers used to clone genes into pSPUTK

	F	R
AqcoeAP3-1	5' CATGCCATGGGAAGAGGAAAGATTGAG 3'	5' ccatcgaTCATCCTAGTCGGAGATCTTC 3'
AqcoeAP3-2	5' CATGCCATGGGGAGAGGAAAGATTGAG 3'	5' ccatcgaTCATGCCAGACTTAAACC 3'
AqcoeAP3-3	5' CATGCCATGGGAAGAGGAAAGATTGAG 3'	5' ccatcgaTTAAGCAAGTCGCAAATTGTG 3'
AqcoePI	5' CATGCCATGGGGAGAGGAAAGATTGAG 3'	5' ccatcgatCTATTTACTCTCCTGTAAATTAGGC 3'
AqcoeSEP3	5' CATGCCATGGGAAGAGGAAGAGTTG 3'	5' ccatcgaTCAACCCAACCAACCTTGC 3'
AqcoeSEP2	5' cctagatctATGGGGAGAGGAAAAGTAG 3'	5' cctagatcTCAAACCATCCAACCAGGG 3'
AqcoeAG1	5' CATGCCATGGGAAGAGGAAAGATTG 3'	5' ccatcgaTTACCCAAGTTGAAGTGTCG 3'
AqcoeAG2	5' CATGCCATGGGAAGAGGAAAGATTG 3'	5' ccatcgaTCAACACAGTTGGAGAGC 3'

	F	R 1st PCR
AqcoeAP3-1-3xHA-tag	5' CATGCCATGGGAAGAGGAAAGATTGAG 3'	5' CTGGAACATCGTATGGGTATCCTAGTCGGAGATCTTCGAATCC 3'
AqcoeAP3-2-3xHA-tag	5' CATGCCATGGGGAGAGGAAAGATTGAG 3'	5' CTGGAACATCGTATGGGTATGCCAGACTTAAACCATATG 3'
AqcoeAP3-3-3xHA-tag	5' CATGCCATGGGAAGAGGAAAGATTGAG 3'	5' CTGGAACATCGTATGGGTAAGCAAGTCGCAAATTGTGGG 3'
R primer for 2nd PCR for HA-tags	5' ccatcgatctaAGCGTAATCTGGAACATCGTATGGGTAAGCGTAATCTGGAACATCGTATGGGTAAGCGTAATCTGGAACATCGTATGGGTA 3'	

A. Thaliana SEP3 EMSA probe (pGEM-T sequence underlined):
5’CATGGCCGCGGGATTTTGACGATAACTCCATCTTTCTATTTTGGGTAACGAGGTCCCCTTCCCATTACGTCTTGACGTGGACCCTGTCCGTCTATTTTTAGCAGAATCACTAGTGCGGCCGC-3’;

SELEX-seq

SELEX was essentially performed a described before (Smaczniak et al., 2017). The dsDNA libraries contained 40 random nucleotide fragments flanked by specific barcodes that allowed for multiplexing in high-throughput sequencing. The dsDNA libraries contained all necessary features required for direct sequencing with an Illumina Genome Analyzer (Jolma et al., 2010). Proteins were synthesized using TNT SP6 Quick Coupled Transcription/Translation System (Promega) following the manufacturer’s instructions in a total volume of 20 µl. The binding reaction mix was prepared essentially as described previously for EMSA experiments (Smaczniak et al., 2012b) and contained 20 µl of in vitro-synthesized proteins and 50-100 ng of dsDNA in a total volume of 120 µl. The binding reaction was incubated for 1 h at 21 °C followed by 1 h immunoprecipitation with 20 µl anti-HA antibodies coupled to magnetic beads (ThermoScientific) in a thermomixer at 21 °C with constant mixing at 700 rpm. After immunoprecipitation, beads were washed 5 times with 150 µl of binding buffer without salmon-sperm DNA, rinsed once with 500 µl of 1xTE and bound DNA was eluted with 50 µl 1X TE by incubation in a thermomixer for 20 min at 90 °C with full mixing speed. Following this incubation, magnetic beads were immobilized and the supernatant containing the eluated DNA was transferred to a new tube. DNA fragments were amplified with 5 to 11 cycles of PCR with SELEX round-specific primers (Jolma et al., 2010) and the total amplicon was used in the subsequent SELEX round. The amplification efficiency was checked on an agarose gel. Samples for sequencing were amplified, cut out from agarose gel and purified using the Qiaquick Gel Extraction Kit (Qiagen). Different libraries were multiplexed by mixing in an equimolar amounts and sequencing was performed on the HiSeq 2000 (Illumina).

SELEX analysis was performed using the SELEX R package version 1.8.0 to build Markov Models of k-mer frequencies and thus determine which k-mers were enriched in the R5 dataset compared to R0. Relative affinities were calculated for K-mers of length 12, and the background frequencies were modelled with a 6th-order hidden Markov model.

Heatmap was made with a custom script, using Manhattan clustering. Motifs were made of multiple alignments (Muscle) using Weblogo (*http://weblogo.berkeley.edu/*) (Crooks et al., 2004).

Motif generation

AP3 Sequence alignments from (Sharma et al., 2011) were used to make logo's using Weblogo (*http://weblogo.berkeley.edu/*) (Crooks et al., 2004).

Supplementary material

Table S1: Calculations of charge and isoelectric points.

	Length (AA)	Size (kD)	Pepcalc		http://isoelectric.ovh.org/		Expasy pi/mw
			Ph7 charge	Isoelectric point	Isoelectric point	Ph7.4 charge	Isoelectric point
AP3-1	224	25.51	0.5	7.28	6.79	-1.6	7.06
AP3-2	226	26.36	1.5	7.83	7.21	-0.5	7.72
AP3-3	221	25.86	4.6	9.37	8.4	3.4	9.02

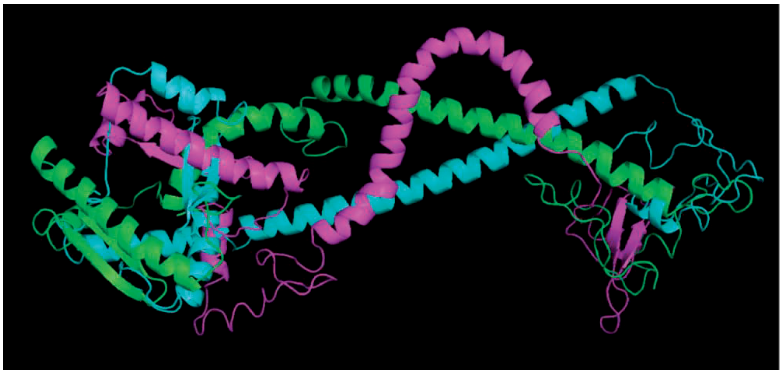


Figure S1. Modelled structures of the three full-length AqAP3 paralogs. *Overlap of the modelled structures shows that the K-domain of AqAP3-1 might have a different structure than AqAP3-2 and AqAP3-3, as the K-domain of AP3-1 has a bend in the α-helix. AqAP3-1=magenta, AqAP3-2=cyan, AqAP3-3=green. Proteins were modelled using Phyre2.*

CHAPTER 7

General discussion and perspectives

Life on earth shows an incredible diversity. One of the intriguing questions in biology is how different morphologies of higher organisms are generated. At the end of the 20th century, advances in understanding the molecular pathways underlying development led to the field of evolutionary developmental biology, or “evo-devo”. Evo-devo tries to understand how developmental programs have evolved to lead to changes in morphology.

In the early days of evo-devo it was observed that a limited set of molecular pathways is used to build different body plans, as exemplified by the homeotic homeobox (HOX) genes in animals (McGinnis et al., 1984; McGinnis and Krumlauf, 1992). In plants, similar conserved mechanisms were found, for example the (A)BCE-model of flower organ specification (Haughn and Somerville, 1988; Coen and Meyerowitz, 1991; Theissen, 2001; Causier et al., 2010a). This model describes which organ types are formed in which position in the flower, and is generally conserved throughout the angiosperms. Even though flowers of most plants have the same basic floral bauplan, as specified by the (A)BCE-model, they still exhibit a large range of different floral morphologies. How these different morphologies are generated at the molecular level remains intriguing (Vergara-Silva, 2003). Early studies in plant evo-devo research mainly focused on candidate-gene approaches, for instance transcription factor (TF) genes from the MADS and TCP families (for reviews see: (Theissen et al., 2000; Rosin and Kramer, 2009). Expression patterns and phylogeny of these candidate genes were analyzed, whereas their functions were studied with heterologous expression approaches (Irish and Benfey, 2004; Di Stilio et al., 2017).

New technologies changed evo-devo research

Possibilities for functional studies in different species expanded with the development of Virus-induced gene silencing (VIGS), which is being used for targeted gene knock-down. This method is now optimized in multiple species, and does not require establishment of stable transformation procedures (Burch-Smith et al., 2004; Senthil-Kumar and Mysore, 2011). To identify genes that underlie morphological differences between closely related species, QTL (Quantitative trait locus) analysis has been used (Irish and Benfey, 2004; Mackay et al., 2009), whereas GWAS (genome wide association studies) has been used for variation within species (Brachi et al., 2011). Another advancement that opened up new directions for evo-devo research is next-generation sequencing. At present, plant genomes and transcriptomes are available from species throughout the plant kingdom. ~100 plant genomes have been sequenced and new bioinformatics methods are being developed to obtain more information from these genomes. For instance, comparative genomics has been used to elucidate whole-genome-duplications in the history of plants, and to trace the history of large gene families (Jiao et al., 2011; Lee et al., 2011; Li et al., 2015; Zhao et al., 2017). Comparative genomics has also been used to find genes linked with certain traits, for instance transitions from perennial to annual life style in the Brassicaceae (Heidel et al., 2016), and the evolution of terpenoid biosynthesis in flowering plants (Hofberger et al., 2015). Next-generation sequencing also permitted the development of a whole new array of genome-wide experimental methods.

Among these are RNA-seq and ChIP-seq. DNAse-seq is used to discover open, active, regions of chromatin (Song and Crawford, 2010). Other new methods are SELEX-seq (selective evolution of ligands by exponential enrichment) and DAP-seq (DNA affinity purification) (Jolma et al., 2010; O'Malley et al., 2016), which allows building of biophysical models of the DNA-binding specificity of transcription factors. Most of these experiments are not limited to certain model species, but can be used for any species of interest.

In this thesis, we have studied different aspects of evo-devo using several genome-wide and sequencing methods, among which ChIP-seq, DNAse-seq and SELEX-seq. We assessed the divergence of TF binding sites (TFBSs) of the major floral regulator SEPALLATA3 (SEP3) between closely related *Arabidopsis* species, as it is known that changes in *cis*-regulatory elements (CREs) play a role in creating morphological diversity (Carroll, 2008; Wittkopp and Kalay, 2012). Gene duplications can also create morphological novelty (Cui et al., 2006; Jiao et al., 2011; Soltis and Soltis, 2016), and we studied the consequences of gene duplication by assessing possible differences between paralogous pairs of the B-class TFs in two different species. Differences in TF function between two paralogs point to the occurrence of sub- or even neo-functionalization.

Divergence of transcription factor binding profiles

Changes in CREs are one possibility to alter gene regulatory networks and thereby create morphological diversity. As we argue in chapter 2, we would expect changes that impact morphology to occur most frequently in the regulation of genes that are immediately downstream of the floral master regulators that fulfil the (A)BCE functions. From earlier ChIP-seq experiments in *A. thaliana* we learned that the master regulators in flower development have thousands of binding sites in the genome, and likely regulate hundreds of genes (Kaufmann et al., 2009; Kaufmann et al., 2010c; Wuest et al., 2012; Ó'Maoiléidigh et al., 2013; Pajoro et al., 2014b). To assess possible changes in gene regulatory networks downstream of the floral regulators, we performed a genome-wide comparison of binding sites of the floral regulator SEP3 in inflorescences between two closely related *Arabidopsis* species, *A. thaliana* and *A. lyrata*. These two species are very similar in flower morphology, but the outcrossing *A. lyrata* has larger flowers, with relatively larger petals (see chapter 3).

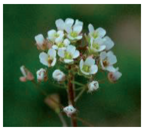
Surprisingly, we found substantial diversity in the SEP3 binding sites between *A. thaliana* and *A. lyrata*, with only around 26% TFBSs conserved (Chapter 3). This prompted us to look at variation within a species. In chapter 4, we analysed the divergence of SEP3 binding between two *A. thaliana* ecotypes. These binding profiles showed more overlap, as ~70% of the TFBSs are common between the two ecotypes. Taken together, these data show that there are differences in SEP3 binding between *A. thaliana* ecotypes as well as between *A. thaliana* and *A. lyrata*. Not surprisingly, the difference between the two species is substantially larger than between the ecotypes. We further analyzed which factors may have caused the difference in SEP3 binding sites. In some cases, these differences in binding events between the species correlated with changes in underlying DNA sequence (see chapter 3 and 4). However, a large proportion of TFBSs that is different between the species shows no obvious divergence in sequence. Another factor that plays a role in TFBS divergence is transposition. We observed that several of the *A. lyrata* SEP3 TFBSs were positioned within transposons. Transposons enable a move of TFBSs to new locations in the genome, a phenomenon that has been reported previously (Feschotte, 2008). As a consequence, transposition of a TFBS-containing transposable element may lead to a change in transcriptional regulation of genes at the new location. That transpositions may create morphological diversity has been shown in several studies. One example is the famous white-to-black color change in moths during the industrial revolution. This was caused by a transposon insertion that caused upregulation of a gene (Van't Hof et al., 2016). There are also examples in plants. In maize, a transposon insertion in the promoter of *TEOSINTE BRANCHED1* (*TB1*) has led to the upregulation of this gene, and subsequently to an increase of apical dominance (Studer et al., 2011). However, whether these changes in gene expression are caused by the presence of a CRE in the transposon is unknown.

Another factor that influences TF binding to DNA is the accessibility of potential binding sites in the genome. DNA in the nucleus is packed into chromatin, and modifications of the histones or the DNA that is wrapped around the histones make some regions more accessible than others. To analyze whether this accessibility may play a role in TFBS selection, we performed DNAse-seq experiments on inflorescences of both *A. thaliana* ecotypes and *A. lyrata* and compared these data with our SEP3 binding profiles. We found that the overlap of open chromatin regions of *A. thaliana* is higher with SEP3 TFBSs of *A. thaliana* than with the SEP3 TFBSs of *A. lyrata*, regardless of the *A. thaliana* ecotype, and vice versa. We did not see this pattern between ecotypes. This suggests that DNA accessibility may contribute to the differences in binding sites between species, but not or not significantly between ecotypes. It would be interesting to select species-specific TFBSs in *A. thaliana* and *A. lyrata* that do not show underlying sequence divergence, and assess whether a difference in DNA accessibility could explain the differences in binding behavior. In addition, it would be interesting to look at DNA methylation and histone marks, to analyze how the difference in DNA accessibility between species is derived for selected candidate genes. It has been shown in seedlings that methylation patterns between *A. thaliana*, *A. lyrata* and *Capsella rubella* differ from each other. Although the difference in genomic structures is the most important factor for this divergence, there are also differences in gene body methylation (Seymour et al., 2014).

Genome sequence and structure divergence between *Arabidopsis* species and ecotypes

The large difference we observed between the SEP3 binding profiles of *A. thaliana* and *A. lyrata*, in contrast to the smaller difference between *A. thaliana* ecotypes, may simply be because the ecotypes diverged from each other more recently. However, it is interesting to note that in animal species with similar divergence times, TFBS profiles seem to deviate slower than we observed between *Arabidopsis* species (Schmidt et al., 2010; He et al., 2011; Stefflova et al., 2013; Villar et al., 2014). It has been suggested that the extent of TFBS divergence might be dependent on the structure of the different genomes, as well as the level of selective constraint on these genomes (Stefflova et al., 2013). Plant genomes tend to evolve faster than animal genomes (Panchy et al., 2016). It is therefore interesting to have a closer look at how much the genomes of the plant species we studied have diverged from each other.

Table 1: Other possible lineage I Brassicaceae species that may be used for comparative analysis. These species have sequenced genomes, and are already used to study variation at the genus level. *pictures of Capsella rubella: Jan van der straaten; Cardamine hirsuta: Aelwin/Wikipedia; A. lyrata: Alfred Cook*

species	*A. thaliana*	*A. lyrata*	*Capsella rubella*	*Cardamine hirsuta*
Diverged from *A. thaliana*	--	10 MYA	14 MYA	32 MYA
Genome size (Mb)	125	207	220	225
Protein-coding genes	27,025	32,670	26,521	29,458
chromosomes	5	8	8	8
Breeding system	Selfing	outcrossing	selfing	Selfing
Causes difference in genome size with *A. thaliana*	(Centromeric regions 14 Mb)	~8 Mb of size difference due to extra centromers Hundreds of thousands small indels	Half the genome consists of centromeric and repetitive regions	Centromeric regions 78.9 Mb
Phenotype				
genome	(Arabidopsis Genome, 2000)	(Hu et al., 2011)	(Slotte et al., 2013)	(Gan et al., 2016)

For *A. thaliana*, the 1001 genomes project has sequenced many different accessions, in an effort to quantify the amount of natural variation present in this species (Ossowski et al., 2008; Cao et al., 2011; Long et al., 2013; Schmitz et al., 2013; Consortium et al., 2016; Kawakatsu et al., 2016). Based on 1,135 different accessions, 10,707,430 biallelic SNPs and 1,424,879 small-scale indels (up to 40 bp) are reported. This represents on average one sequence variant every 10 bp (Consortium et al., 2016). Compared with the *A. thaliana Col-0* reference genome, one-third of protein-coding genes are predicted to be disrupted in at least one accession, although re-annotation of each genome revealed that alternative gene models often restore coding potential (Gan et al., 2011). *A. thaliana* accessions also differ in the number and position of transposons, with any two *A. thaliana* accessions predicted to differ in 200-300 TE insertions (Quadrana et al., 2016). The genomic differences between the species *A. thaliana* and *A. lyrata* are more substantial. The genomes of these two species show >80% sequence identity, and most (90%) of the genome is syntenic, making it possible to align the genomes.

However, there are some clear differences between these two genomes as well. Their size differs substantially, as the genome of *A. thaliana* is 125 Mb whereas *A. lyrata* has a genome of 207 Mb. Comparison of these genomes show that there have been several (10) major rearrangements, among them three chromosomal fusions that led to the difference in chromosome number (five for *A. thaliana; A. lyrata* has eight, which is probably the ancestral number in Brassicaceae). Most of the difference in genome size however, is due to hundreds of thousands of small indels (Hu et al., 2011). Summarized, although these species are closely related and their sequences can still be aligned, substantial rearrangements and sequence divergence in the genomes of *A. thaliana* and *A. lyrata* have occurred in a species-specific manner. To assess whether this reorganization contributed to the differences in TFBS profiles between the two species, it would be interesting to compare TF binding profiles between species that have a more similar genome. There are several other Brassicaceae species with a sequenced genome that could be used for this kind of comparisons, as they are more similar in genome structure to *A. lyrata*. Species that could be used for experiments like this are for instance C. *rubella* and C*ardamine hirsuta* (see **Table 1**).

Divergence of paralogs after gene duplication also plays a role in creating morphological diversity

Evolutionary changes are thought to occur more often in regulatory elements than in coding sequences, as changes in proteins often have pleiotropic effects. Nevertheless, mutations in TFs that create morphological differences do occur. For instance, mutations in TFs functional in skeletal development are involved in the evolution of flightless birds on the Galapagos islands (Burga et al., 2017). One way by which TFs can change function to alter morphology is through gene duplications; after a duplication there are two copies, which temporarily lessens the selection pressure on both paralogs. Plants underwent frequent genome duplications and some of the major evolutionary innovations, like seeds and flowers are thought to coincide with genome duplications. In this thesis, we studied the divergence between paralogs of the floral homeotic B-class genes. Throughout the angiosperm phylogeny, there are frequent

duplication events of both *APETALA3* (*AP3)* and *PISTILLATA* (*PI)* (Chung et al., 1995; Kramer et al., 1998; Kramer et al., 2003; Stellari et al., 2004; Vandenbussche et al., 2004; Jaramillo and Kramer, 2007; Viaene et al., 2009; Bartlett and Specht, 2010). In this thesis, we studied two cases of B-class paralogs. The first paralogous pair is the *PI* paralogs in *T. hassleriana*, a close relative of *Arabidopsis* (Chapter 5). We chose to study this paralogous pair as they show an interesting difference in their location in the genome. *PI* orthologs show a strong conservation of synteny across the eudicots, with the notable exception of the Brassicaceae (Zhao and Schranz, 2017). Intriguingly, *T. hassleriana* has two *PI* paralogs, one which is syntenic with most *PI* orthologs, whereas the other is syntenic with the Brassicaceae *PI* (Chapter 5, (Cheng et al., 2013)). We hypothesized that this difference in genomic location may have influenced floral morphology in the Brassicaceae. Therefore these paralogs in *T. hassleriana* are of particular interest, as they may provide an 'intermediate' situation. Here, we show that this *PI* transposition did not lead to differences in expression pattern between the two *PI* copies. However, our data suggest that these PI paralogs are biochemically different as they do show differences in interaction with other TFs, and possibly differences in DNA-binding specificity. Whether the observed changes could lead to morphological differences should be assessed using functional studies in the future.
In chapter 6, we analyzed the three paralogous AP3 proteins of *Aquilegia*. Flowers of this basal eudicot have a fifth type of organ, which is positioned between stamens and carpels, termed the staminodium. Sub- and neofunctionalization of the *AP3* paralogs led to the incorporation of this new organ into the floral bauplan of *Aquilegia* (Kramer et al., 2007). It has been shown that *AqAP3-1* is necessary for the specification of staminodia. However, it is unknown how gene regulatory networks downstream of Aq*AP3-1* evolved to lead to a novel organ morphology. In this thesis, we show that the AP3 paralogs of *Aquilegia* are biochemically different, as we observed differences in protein-protein interactions as well as DNA-binding specificity.

Both of the pairs of paralogous B-class genes studied in this thesis exhibit differences in interactions with DNA as well as with other TFs. Changes in interaction behavior between paralogous proteins have been shown before. For example, paralogs of the *A. thaliana* AUXIN RESPONSE FACTORS (ARFs) recognize very similar DNA sequences. However, one ARF dimer binds to two sites on the DNA, and the different ARF paralogs prefer different spacing distances between these sites (Boer et al., 2014). A similar observation has been made for the four SEPALLATA proteins of *A. thaliana*. The floral quartet model specifies that MADS-TF can bind as quartets to two CArG-boxes (Theissen and Saedler, 2001). *In vitro* experiments have shown that homotetramers of the different SEP proteins prefer a different length of spacer between the two CArG-boxes (Jetha et al., 2014). Changes in protein-protein interaction also occur. The most famous example of a difference in protein-interactions is FARINELLI of *Antirrhinum*. One of the differences in sequence between FARINELLI and its paralog PLENA (and AG in *A. thaliana*) is a single amino acid insertion (Q173). This single amino acid polymorphism led to a difference in interaction with the SEP proteins. Overexpression of AtAG with or without this glutamine residue in *A. thaliana* had different results, most likely due to different expression patterns of the differentially interacting SEP proteins (Airoldi et al., 2010). This shows that differences in DNA-binding and protein interaction properties between paralogous TF may lead to differences in morphology, although it may depend on the expression pattern of the possible interaction partners.

With respect to the divergence between paralogs, it is interesting to note that SEP3, which we used for our analysis of TF binding profile divergence (Chapters 3 and 4), has three paralogs in *Arabidopsis*. The four *SEP* genes of *A. thaliana* are regarded as largely redundant, although they do have slightly different expression patterns and they are shown to have slightly different behaviors *in vitro* (Pelaz et al., 2000; Ditta et al., 2004; Jetha et al., 2014). We showed that the SEP3 orthologs in *A. thaliana* and *A. lyrata* are very similar in sequence, and that there are no differences in the sequence of the MADS-domain (chapter 3). However, these SEP3 orthologs, as well as the orthologs of the other *SEP* genes, do show subtle differences in other parts of the protein sequence between *A. thaliana* and *A. lyrata*. Formally, we cannot rule out that these sequence differences influence the DNA-binding specificity of the SEP3 orthologs, e.g. by affecting the formation of DNA-binding protein complexes. The four SEP genes are expected to be redundant, but there may be differences in DNA-binding between the paralogs (Jetha et al., 2014). As there is a chance that these possible differences in SEP TFBS profiles are species-specific, it would be interesting to do a more comprehensive comparative ChIP-seq approach, where all four SEP paralogs are interrogated in both *A. thaliana* and *A. lyrata*, and the complete SEP binding profile is compared between species. This would answer whether the collective SEP TFBS profiles are more similar between *A. thaliana* and *A. lyrata* than the SEP3 TFBS profile. However, this is a challenging experiment.

It is intriguing to see that paralogous proteins partly diverge from each other with respect to *in vitro* interaction capacity. However, whether the observed differences between the paralogs lead to differences in morphological outcome depend on whether these biochemical differences lead to a differences in target genes.

Linking observed differences in TFs and their DNA binding profiles to changes in floral morphologies

Although changes in TFs, or the evolution of TFBSs may be important for generating morphological differences, this is only possible if these changes result in differential gene expression of downstream genes. Comparing genome-wide expression data of different organs or different species will result in a set of differentially expressed genes. However, likely not all differentially expressed genes will make a contribution to morphology. Selecting candidate genes that are responsible for the evolution of floral morphology may be more successful when expression data are combined with TF binding profiles of the floral homeotic regulators. Studying candidate genes in detail may not only show how gene regulatory networks evolved to change morphology, but may also increase our knowledge of flower development by identifying more genes with functions in this process.

Even when combining ChIP-seq experiments with expression studies, targets need to be chosen carefully to obtain high-confidence candidate genes. For instance, the SEP3 binding profiles of *A. thaliana* and *A. lyrata* are so diverged that this comparison leads to many potential candidates. In addition, SEP3 plays roles in development of all floral organs throughout flower development, which makes it more complicated to link candidate genes to phenotypes in a specific floral organ. One option could be to analyze binding profiles of other homeotic MADS-domain TFs. For instance, the B-class genes are active in petals and stamens, whereas the C-class TF are involved in stamens and carpels only. Analyzing divergence in binding sites of any of these TFs limits possible morphological changes to these specific organs. Which species are selected for the comparison is also of importance. We showed that the divergence of SEP3 binding profiles between two different *Arabidopsis* species is very large. *A. thaliana* ecotypes also show differences in SEP3 binding, but the larger overlap between these datasets makes selection of suitable candidate genes more amenable. That *A. thaliana* harbors enough natural variation in floral morphology for these type of comparisons has been shown by QTL analysis for petal shape on different *A. thaliana* ecotypes, which resulted in several loci that were not previously implicated in controlling floral organ development (Abraham et al., 2013). I expect that most differences in morphology are due to differences in spatiotemporal expression patterns of genes, not in complete absence of expression in one of the analyzed species. Therefore, these types of experiments could result in more high-confidence candidate genes when time-course experiments would be performed.

Mechanisms of transcriptional regulation

A change in morphology can only occur when gene regulation is altered. We know that the spatiotemporal expression pattern of a gene is determined by the combined activity of several DNA-binding TFs that can act in a cooperative, competing or redundant manner. However, we do not have sufficient knowledge about the mechanisms that lead to changes in gene expression.

The first aspect of transcriptional regulation is binding of the TF to the DNA. Although for more and more TFs the DNA-binding specificity is determined, we still cannot reliably predict TFBSs. One difficulty is that TFs often act as heterodimers, which can influence their preferred binding motif. For instance, the floral regulators SEP3 and AG have distinct DNA binding preferences, and an intermediate between these preferences is observed for SEP3/AG heterodimers (Smaczniak et al., 2017). Interaction partners may even alter DNA-binding specificity, as has been shown for the *Drosophila* HOX proteins. whereas homodimers of all eight HOX paralogs share the same consensus site (Noyes et al., 2008), they acquire novel, paralog-specific DNA-binding specificities in the presence of the cofactor Extradenticle-homothorax (Slattery et al., 2011). To make prediction efforts even more difficult, TFs do not always seem to bind to the highest affinity sites in the genome (Tanay, 2006). DNA methylation also complicates TFBS prediction further, as it has been shown that DNA methylation can influence TF binding

(O'Malley et al., 2016). Sequences outside a TF binding site may also play a role in TFBS selection. Examining a large amount of TFs, It has been shown that the nucleotide composition flanking a bound motif is different from the one around an unbound motif, with a GC content that matches the GC content of the TF motif (White et al., 2013; Dror et al., 2015). Something similar was shown for MADS-domain TFs: although the central dinucleotides and the sequences flanking the CArG-box play the most important role in determining TF binding, regular spacing of certain motifs called A-tracts outside of the bound motif also seems to play a role in SEP3 binding site selection (Muiño et al., 2014). These observed sequence differences outside the bound motif may indicate differences in DNA structure, as it has been shown that the 3D conformation of the DNA can influence TF-DNA binding (Levo and Segal, 2014; Muiño et al., 2014; Slattery et al., 2014).

Binding of a TF to DNA does not automatically lead to altered transcriptional regulation, as a large fraction of binding events do not lead to expression differences (Li et al., 2008; Kaufmann et al., 2009; Schmidt et al., 2010; Cusanovich et al., 2014; Slattery et al., 2014). The outcome of TF binding may be dependent on other interaction partners. An example comes from the MADS-domain TFs SHATTERPROOF1 and 2 (SHP1 and 2) and SEEDSTICK (STK) in *A. thaliana* silique formation. Although the proteins are fully interchangeable, these TFs are differentially expressed, and a different availability of interaction partners leads to different regulatory potential. STK inhibits lignification in the seed abscission zone by interacting with the transcriptional co-repressor SEUSS, whereas the SHPs promote lignification in the valve margins, where SEUSS is not expressed (Balanza et al., 2016).

TFs might also influence gene regulation in indirect ways, leading to a delay in expression changes. One example is the activation of *KNUCKLES (KNU)* expression by AGAMOUS. AG binds to the *KNU* promoter but does not induce a direct change in gene expression. Instead, it blocks a polycomb response element in the promoter, which prohibits the recruitment of more repressive H3K27me3 histone marks to the *KNU* coding region. As a consequence, the repressive H3K27me3 mark will be sequentially diluted with every round of cell division, until there is not enough of the mark to repress *KNU* anymore, at which point *KNU* will be expressed (Sun et al., 2009; Sun et al., 2014).

These examples show that TF function may be influenced by the DNA-binding specificity of the TF itself, interaction of the TF with other proteins, as well as their exact mode of action. Availability of interaction partners, as well as chromatin structure also play a role. All these components make transcriptional regulation a very complicated, and not-well understood process. To understand how transcriptional regulation evolved between species, we will need a better understanding of transcriptional regulation in general.

Thoughts on ChIP-seq methodology

We observed substantial differences between *A. lyrata* and *A. thaliana* SEP3 binding profiles, and discussed several possible biological causes for this TFBS divergence. However, a factor that we did not discuss is the nature of the experiment itself. ChIP-seq experiments revealed that floral MADS-domain TFs bind to thousands of sites in the genome (Kaufmann et al., 2009; Kaufmann et al., 2010c). ChIP-seq experiments combined with expression studies have also shown that only a fraction of these binding events leads to changes in gene expression (Li et al., 2008; Kaufmann et al., 2009; Kaufmann et al., 2010c; Schmidt et al., 2010).

Interestingly, peak score seems to be a reliable indicator of functional binding, at least in homogenous samples (Slattery et al., 2014). However, our *Arabidopsis* samples of whole inflorescences are far from homogeneous and contain many cell types in different developmental stages. Therefore, the peak score reflects partly the number of cells in which the TF is expressed rather than only the binding strength. Interestingly, in our *A. thaliana* ecotype comparison of SEP3 DNA-binding profiles, we observed that the TFBS common between the ecotypes were likely to have a higher peak score than the TFBS present in a single ecotype only (chapter 4). This may indicate that stricter thresholds for determination of peaks should be used to select TFBSs more likely to regulate gene expression.

In our ChIP-seq experiments, we assign a TFBS to the closest gene, based on the one-dimensional genome sequence. However, this is not necessarily the closest gene in the three-dimensional nuclear environment. Hi-C-based experiments show that the *A. thaliana* genome is dominantly arranged in small interactive regions that vary in size between 2-50 kb (Feng et al., 2014; Wang et al., 2015). Although these interacting regions are small, they can still contain several genes. It may be that assigning a TFBS to the nearest gene on the linear genome is not always correct. Another interesting notion in our data, as well as published data, is that often only a fraction of TF-bound genomic regions contains the consensus motif for the TF that is assessed. It may be that some of the regions identified by ChIP-seq are not actually directly bound, but are the consequence of indirect binding, as we know that TFs can act in large complexes. Both of these observations may contribute to the notion that not all TFBSs lead to changes in gene expression.

New methods may influence evo-devo studies

One of the complications of using genome-wide methods in floral evo-devo studies is that they are often done with mixed tissues, which can dilute the signal. Several studies compared expression data from different stages, but for these experiments different samples were often harvested based on bud size, or several developmental stages were pooled together (Singh et al., 2013; Wang et al., 2014). Early flower developmental stages are difficult to harvest, as these are too small to dissect manually. In *A. thaliana*, this problem has been overcome using a floral induction system which generated a large number of synchronized buds (Wellmer et al., 2006). This system has been used for ChIP-seq experiments at a single early developmental stage, as well as several time-course experiments (Kaufmann et al., 2010c; Wuest et al., 2012; Ó'Maoiléidigh et al., 2013; Pajoro et al., 2014b). Although this induction system is not available in other species yet, it would be useful to introduce it at least in other Brassicaceae species.

Several methods to select specific cell-types have been developed. One method is INTACT (isolation of nuclei tagged in specific cell types), in which nuclei of specific cell types are labelled with biotin and purified using streptavidin-coated beads. However, this method relies on transgenics, making it only possible in species amenable to transformation (Deal and Henikoff, 2011b). Laser capture microdissection (LCM) is another method to select specific cells. This method can be used in any species, but is very laborious (Emmert-Buck et al., 1996). Recently, other methods to increase spatiotemporal resolution are being developed. For example, high-resolution gene expression patterns can be obtained from tissue sections by placing them on an array with reverse transcription primers in spots of 100 um, each with a position-specific barcode. After sequencing, these barcodes are used to link the expression data to their original tissue (Giacomello et al., 2017). Other methods to increase spatiotemporal resolution of genome-wide experiment would benefit the field of evo-devo.

Another bottleneck in evo-devo research is functional analysis of genes, i.e. how divergence of genes and genomes affects the development of the plant. Mutants can be a powerful source to study the functions of genes. Although for *A. thaliana* there are mutant databases, for other model species mutant databases are often lacking or not as extensive. TILLING (Targeting Induced Local Lesions in Genomes) is a generic method to find mutants after mutagenesis in basically every species, but is laborious, as many plants need to be screened (Wang et al., 2012). One method that causes a revolution in functional gene analysis is CRISPR/CAS9. CRISPR/CAS9 is originally a bacterial antiviral system, which uses a guide RNA to guide the CAS9 nuclease to the complementary DNA sequence, and CAS9 then cuts the DNA. This system can be used for targeted mutagenesis (Jinek et al., 2012; Cong et al., 2013; Lozano-Juste and Cutler, 2014). Although to use this system a species need to be amenable to transformation, theoretically it could be used in any species to created specific mutants. Not only mutations in protein-coding sequences can be generated, but also specific regulatory elements may be deleted to study details of gene regulation.

The field of floral evo-devo came a long way since discovering that the (A)BCE-model is largely conserved. With the development and use of new techniques, we will be able to elucidate more and more details about floral developmental pathways, as well as how these pathways may be modified to generate different morphologies.

REFERENCES

Abel, C., Clauss, M., Schaub, A., Gershenzon, J., and Tholl, D. (2009). Floral and insect-induced volatile formation in Arabidopsis lyrata ssp. petraea, a perennial, outcrossing relative of A. thaliana. Planta **230,** 1-11.

Abraham, M.C., Metheetrairut, C., and Irish, V.F. (2013). Natural Variation Identifies Multiple Loci Controlling Petal Shape and Size in Arabidopsis thaliana. PLoS One **8,** e56743.

Airoldi, C.A., and Davies, B. (2012). Gene duplication and the evolution of plant MADS-box transcription factors. Journal of genetics and genomics = Yi chuan xue bao **39,** 157-165.

Airoldi, C.A., Bergonzi, S., and Davies, B. (2010). Single amino acid change alters the ability to specify male or female organ identity. Proc Natl Acad Sci U S A **107,** 18898-18902.

Alvarez-Buylla, E.R., Pelaz, S., Liljegren, S.J., Gold, S.E., Burgeff, C., Ditta, G.S., Ribas de Pouplana, L., Martinez-Castilla, L., and Yanofsky, M.F. (2000). An ancestral MADS-box gene duplication occurred before the divergence of plants and animals. Proc Natl Acad Sci U S A **97,** 5328-5333.

Ambrose, B.A., Lerner, D.R., Ciceri, P., Padilla, C.M., Yanofsky, M.F., and Schmidt, R.J. (2000). Molecular and Genetic Analyses of the Silky1 Gene Reveal Conservation in Floral Organ Specification between Eudicots and Monocots. Molecular cell **5,** 569-579.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol **11,** 1.

Andersson, D.I., Jerlström-Hultqvist, J., and Näsvall, J. (2015). Evolution of New Functions De Novo and from Preexisting Genes. Cold Spring Harbor Perspectives in Biology **7**.

Angenent, G.C., Franken, J., Busscher, M., van Dijken, A., van Went, J.L., Dons, H.J., and van Tunen, A.J. (1995). A novel class of MADS box genes is involved in ovule development in petunia. Plant Cell **7,** 1569-1582.

Arabidopsis Genome, I. (2000). Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature **408,** 796-815.

Arnaud, N., Lawrenson, T., Ostergaard, L., and Sablowski, R. (2011). The same regulatory point mutation changed seed-dispersal structures in evolution and domestication. Curr Biol **21,** 1215-1219.

Balanza, V., Roig-Villanova, I., Di Marzo, M., Masiero, S., and Colombo, L. (2016). Seed abscission and fruit dehiscence required for seed dispersal rely on similar genetic networks. Development **143,** 3372-3381.

Barker, M.S., Vogel, H., and Schranz, M.E. (2009). Paleopolyploidy in the Brassicales: Analyses of the Cleome Transcriptome Elucidate the History of Genome Duplications in Arabidopsis and Other Brassicales. Genome Biology and Evolution **1,** 391-399.

Bartlett, M., Thompson, B., Brabazon, H., Del Gizzi, R., Zhang, T., and Whipple, C. (2016). Evolutionary dynamics of floral homeotic transcription factor protein-protein interactions. Mol Biol Evol.

Bartlett, M.E., and Specht, C.D. (2010). Evidence for the involvement of Globosa-like gene duplications and expression divergence in the evolution of floral morphology in the Zingiberales. New Phytol **187,** 521-541.

Bastow, R., Mylne, J.S., Lister, C., Lippman, Z., Martienssen, R.A., and Dean, C. (2004). Vernalization requires epigenetic silencing of FLC by histone methylation. Nature **427,** 164-167.

Baxter, L., Jironkin, A., Hickman, R., Moore, J., Barrington, C., Krusche, P., Dyer, N.P., Buchanan-Wollaston, V., Tiskin, A., and Beynon, J. (2012). Conserved noncoding sequences highlight shared components of regulatory networks in dicotyledonous plants. The Plant Cell **24,** 3949-3965.

Becker, A. (2016). Tinkering with transcription factor networks for developmental robustness of Ranunculales flowers. Annals of Botany.

Becker, A., and Lange, M. (2010). VIGS--genomics goes functional. Trends Plant Sci **15,** 1-4.

Behrens, S., and Vingron, M. (2010). Studying the evolution of promoter sequences: a waiting time problem. Journal of Computational Biology **17,** 1591-1606.

Bemer, M., van Dijk, A.D., Immink, R.G., and Angenent, G.C. (2017). Cross-Family Transcription Factor Interactions: An Additional Layer of Gene Regulation. Trends Plant Sci **22,** 66-80.

Benlloch, R., Roque, E., Ferrandiz, C., Cosson, V., Caballero, T., Penmetsa, R.V., Beltran, J.P., Canas, L.A., Ratet, P., and Madueno, F. (2009). Analysis of B function in legumes: PISTILLATA proteins do not require the PI motif for floral organ development in Medicago truncatula. Plant J **60,** 102-111.

Bennett, M.D., Leitch, I.J., Price, H.J., and Johnston, J.S. (2003). Comparisons with Caenorhabditis (approximately 100 Mb) and Drosophila (approximately 175 Mb) using flow cytometry show genome size in Arabidopsis to be approximately 157 Mb and thus approximately 25% larger than the Arabidopsis genome initiative estimate of approximately 125 Mb. Annals of botany **91,** 547-557.

Bennetzen, J.L., Ma, J., and Devos, K.M. (2005). Mechanisms of recent genome size variation in flowering plants. Annals of botany **95,** 127-132.

Berbel, A., Navarro, C., Ferrandiz, C., Canas, L.A., Beltran, J.P., and Madueno, F. (2005). Functional conservation of PISTILLATA activity in a pea homolog lacking the PI motif. Plant Physiol **139,** 174-185.

Bharathan, G., Goliber, T.E., Moore, C., Kessler, S., Pham, T., and Sinha, N.R. (2002). Homologies in leaf form inferred from KNOXI gene expression during development. Science **296,** 1858-1860.

Biggin, M.D. (2011). Animal transcription networks as highly connected, quantitative continua. Dev Cell **21,** 611-626.

Birnbaum, K., Shasha, D.E., Wang, J.Y., Jung, J.W., Lambert, G.M., Galbraith, D.W., and Benfey, P.N. (2003). A gene expression map of the Arabidopsis root. Science **302,** 1956-1960.

Blanc, G., and Wolfe, K.H. (2004a). Functional divergence of duplicated genes formed by polyploidy during Arabidopsis evolution. Plant Cell **16,** 1679-1691.

Blanc, G., and Wolfe, K.H. (2004b). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. Plant Cell **16,** 1667-1678.

Blanc, G., Hokamp, K., and Wolfe, K.H. (2003). A Recent Polyploidy Superimposed on Older Large-Scale Duplications in the Arabidopsis Genome. Genome Res **13,** 137-144.

Boer, D.R., Freire-Rios, A., van den Berg, Willy A.M., Saaki, T., Manfield, Iain W., Kepinski, S., López-Vidrieo, I., Franco-Zorrilla, Jose M., de Vries, Sacco C., Solano, R., Weijers, D., and Coll, M. (2014). Structural Basis for DNA Binding Specificity by the Auxin-Dependent ARF Transcription Factors. Cell **156,** 577-589.

Bowers, J.E., Chapman, B.A., Rong, J., and Paterson, A.H. (2003). Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events **422,** 433-438.

Bowman, J.L. (1997). Evolutionary conservation of angiosperm flower development at the molecular and genetic levels. Journal of Biosciences **22,** 515-527.

Bowman, J.L., Smyth, D.R., and Meyerowitz, E.M. (1989). Genes directing flower development in Arabidopsis. The Plant Cell **1,** 37-52.

Bowman, J.L., Smyth, D.R., and Meyerowitz, E.M. (1991). Genetic interactions among floral homeotic genes of Arabidopsis. Development **112,** 1-20.

Brachi, B., Morris, G.P., and Borevitz, J.O. (2011). Genome-wide association studies in plants: the missing heritability is in the field. Genome Biol **12,** 232.

Bradley, D., Carpenter, R., Sommer, H., Hartley, N., and Coen, E. (1993). Complementary floral homeotic phenotypes result from opposite orientations of a transposon at the plena locus of antirrhinum. Cell **72,** 85-95.

Brady, S.M., Zhang, L., Megraw, M., Martinez, N.J., Jiang, E., Yi, C.S., Liu, W., Zeng, A., Taylor-Teeples, M., Kim, D., Ahnert, S., Ohler, U., Ware, D., Walhout, A.J., and Benfey, P.N. (2011). A stele-enriched gene regulatory network in the Arabidopsis root. Molecular systems biology **7,** 459.

Bratzel, F., and Turck, F. (2015). Molecular memories in the regulation of seasonal flowering: from competence to cessation. Genome Biol **16,** 1-14.

Brawand, D., Soumillon, M., Necsulea, A., Julien, P., Csardi, G., Harrigan, P., Weier, M., Liechti, A., Aximu-Petri, A., Kircher, M., Albert, F.W., Zeller, U., Khaitovich, P., Grutzner, F., Bergmann, S., Nielsen, R., Paabo, S., and Kaessmann, H. (2011). The evolution of gene expression levels in mammalian organs. Nature **478,** 343-348.

Bremer, K., Friis, E.M., and Bremer, B. (2004). Molecular phylogenetic dating of asterid flowering plants shows early Cretaceous diversification. Syst Biol **53,** 496-505.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nature methods **10,** 1213-1218.

Burch-Smith, T.M., Anderson, J.C., Martin, G.B., and Dinesh-Kumar, S.P. (2004). Applications and advantages of virus-induced gene silencing for gene function studies in plants. The Plant Journal **39,** 734-746.

Burga, A., Wang, W., Ben-David, E., Wolf, P.C., Ramey, A.M., Verdugo, C., Lyons, K., Parker, P.G., and Kruglyak, L. (2017). A genetic signature of the evolution of loss of flight in the Galapagos cormorant. Science **356**.

Busch, A., and Zachgo, S. (2007). Control of corolla monosymmetry in the Brassicaceae Iberis amara. Proc Natl Acad Sci U S A **104,** 16714-16719.

Busch, A., and Zachgo, S. (2009). Flower symmetry evolution: towards understanding the abominable mystery of angiosperm radiation. Bioessays **31,** 1181-1190.

Busch, A., Horn, S., Muhlhausen, A., Mummenhoff, K., and Zachgo, S. (2012). Corolla monosymmetry: evolution of a morphological novelty in the Brassicaceae family. Mol Biol Evol **29,** 1241-1254.

Buzgo, M., Soltis, P., xa, S, Soltis, D., xa, and E. (2004). Floral Developmental Morphology of <em>Amborella trichopoda</em> (Amborellaceae). International Journal of Plant Sciences **165,** 925-947.

Calonje, M., Cubas, P., Martinez-Zapater, J.M., and Carmona, M.J. (2004). Floral meristem identity genes are expressed during tendril development in grapevine. Plant Physiol **135,** 1491-1501.

Cannon, S.B., Sterck, L., Rombauts, S., Sato, S., Cheung, F., Gouzy, J., Wang, X., Mudge, J., Vasdewani, J., Schiex, T., Spannagl, M., Monaghan, E., Nicholson, C., Humphray, S.J., Schoof, H., Mayer, K.F.X., Rogers, J., Quétier, F., Oldroyd, G.E., Debellé, F., Cook, D.R., Retzel, E.F., Roe, B.A., Town, C.D., Tabata, S., Van de Peer, Y., and Young, N.D. (2006). Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes. Proceedings of the National Academy of Sciences **103,** 14959-14964.

Cao, J., Schneeberger, K., Ossowski, S., Gunther, T., Bender, S., Fitz, J., Koenig, D., Lanz, C., Stegle, O., Lippert, C., Wang, X., Ott, F., Muller, J., Alonso-Blanco, C., Borgwardt, K., Schmid, K.J., and Weigel, D. (2011). Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet **43,** 956-963.

Caris, P.L., Geuten, K.P., Janssens, S.B., and Smets, E.F. (2006). Floral development in three species of Impatiens (Balsaminaceae). American journal of botany **93,** 1-14.

Carpenter, R., and Coen, E.S. (1990). Floral homeotic mutations produced by transposon-mutagenesis in Antirrhinum majus. Genes Dev **4,** 1483-1493.

Carroll, S.B. (2008). Evo-devo and an expanding evolutionary synthesis: a genetic theory of morphological evolution. Cell **134,** 25-36.

Castillo-Davis, C.I., Hartl, D.L., and Achaz, G. (2004). cis-Regulatory and protein evolution in orthologous and duplicate genes. Genome Res **14,** 1530-1536.

Causier, B., Schwarz-Sommer, Z., and Davies, B. (2010a). Floral organ identity: 20 years of ABCs. Seminars in cell & developmental biology **21,** 73-79.

Causier, B., Castillo, R., Xue, Y., Schwarz-Sommer, Z., and Davies, B. (2010b). Tracing the evolution of the floral homeotic B- and C-function genes through genome synteny. Mol Biol Evol **27,** 2651-2664.

Chan, Y.F., Marks, M.E., Jones, F.C., Villarreal, G., Jr., Shapiro, M.D., Brady, S.D., Southwick, A.M., Absher, D.M., Grimwood, J., Schmutz, J., Myers, R.M., Petrov, D., Jonsson, B., Schluter, D., Bell, M.A., and Kingsley, D.M. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. Science **327,** 302-305.

Chanderbali, A.S., Berger, B.A., Howarth, D.G., Soltis, P.S., and Soltis, D.E. (2016). Evolving Ideas on the Origin and Evolution of Flowers: New Perspectives in the Genomic Era. Genetics **202,** 1255-1265.

Chanderbali, A.S., Yoo, M.J., Zahn, L.M., Brockington, S.F., Wall, P.K., Gitzendanner, M.A., Albert, V.A., Leebens-Mack, J., Altman, N.S., Ma, H., dePamphilis, C.W., Soltis, D.E., and Soltis, P.S. (2010). Conservation and canalization of gene expression during angiosperm diversification accompany the origin and evolution of the flower. Proc Natl Acad Sci U S A **107,** 22570-22575.

Chen, K.Y., Cong, B., Wing, R., Vrebalov, J., and Tanksley, S.D. (2007). Changes in regulation of a transcription factor lead to autogamy in cultivated tomatoes. Science **318,** 643-645.

Cheng, S., van den Bergh, E., Zeng, P., Zhong, X., Xu, J., Liu, X., Hofberger, J., de Bruijn, S., Bhide, A.S., Kuelahoglu, C., Bian, C., Chen, J., Fan, G., Kaufmann, K., Hall, J.C., Becker, A., Brautigam, A., Weber, A.P., Shi, C., Zheng, Z., Li, W., Lv, M., Tao, Y., Wang, J., Zou, H., Quan, Z., Hibberd, J.M., Zhang, G., Zhu, X.G., Xu, X., and Schranz, M.E. (2013). The Tarenaya hassleriana genome provides insight into reproductive trait and genome evolution of crucifers. Plant Cell **25,** 2813-2830.

Chung, Y.-Y., Kim, S.-R., Kang, H.-G., Noh, Y.-S., Park, M.C., Finkel, D., and An, G. (1995). Characterization of two rice MADS box genes homologous to GLOBOSA. Plant Science **109,** 45-56.

Coen, E.S., and Meyerowitz, E.M. (1991). The war of the whorls: genetic interactions controlling flower development. Nature **353,** 31-37.

Colombo, L., Franken, J., Koetje, E., van Went, J., Dons, H.J., Angenent, G.C., and van Tunen, A.J. (1995). The petunia MADS box gene FBP11 determines ovule identity. Plant Cell **7,** 1859-1868.

Conant, G.C., and Wolfe, K.H. (2008). Turning a hobby into a job: how duplicated genes find new functions. Nature reviews. Genetics **9,** 938-950.

Conant, G.C., Birchler, J.A., and Pires, J.C. (2014). Dosage, duplication, and diploidization: clarifying the interplay of multiple models for duplicate gene evolution over time. Current Opinion in Plant Biology **19,** 91-98.

Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., and Zhang, F. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. Science **339,** 819-823.

Consortium, G., Electronic address, m.n.g.o.a.a., and Genomes, C. (2016). 1,135 Genomes Reveal the Global Pattern of Polymorphism in Arabidopsis thaliana. Cell **166,** 481-491.

Consortium, T.R.C.S. (2005). Sequence, annotation, and analysis of synteny between rice chromosome 3 and diverged grass species. Genome Res **15,** 1284-1291.

Crane, P.R., Friis, E.M., and Pedersen, K.R. (1995). The origin and early diversification of angiosperms **374,** 27-33.

Cronk, Q., and Ojeda, I. (2008). Bird-pollinated flowers in an evolutionary and molecular context. J Exp Bot **59,** 715-727.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res **14,** 1188-1190.

Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., Albert, V.A., Ma, H., and dePamphilis, C.W. (2006). Widespread genome duplications throughout the history of flowering plants. Genome Res **16,** 738-749.

Cui, R., Han, J., Zhao, S., Su, K., Wu, F., Du, X., Xu, Q., Chong, K., Theißen, G., and Meng, Z. (2010). Functional conservation and diversification of class E floral homeotic genes in rice (Oryza sativa). The Plant Journal **61,** 767-781.

Cumbie, J.S., Filichkin, S.A., and Megraw, M. (2015). Improved DNase-seq protocol facilitates high resolution mapping of DNase I hypersensitive sites in roots in Arabidopsis thaliana. Plant Methods **11,** 42.

Cusanovich, D.A., Pavlovic, B., Pritchard, J.K., and Gilad, Y. (2014). The functional consequences of variation in transcription factor binding. PLoS Genet **10,** e1004226.

Davidson, E.H., and Erwin, D.H. (2006). Gene regulatory networks and the evolution of animal body plans. Science **311,** 796-800.

Davies, B., Motte, P., Keck, E., Saedler, H., Sommer, H., and Schwarz-Sommer, Z. (1999). PLENA and FARINELLI: redundancy and regulatory interactions between two Antirrhinum MADS-box factors controlling flower development. Embo J **18,** 4023-4034.

de Koning, A.J., Gu, W., Castoe, T.A., Batzer, M.A., and Pollock, D.D. (2011). Repetitive elements may comprise over two-thirds of the human genome. PLoS Genet **7,** e1002384.

de Martino, G., Pan, I., Emmanuel, E., Levy, A., and Irish, V.F. (2006). Functional analyses of two tomato APETALA3 genes demonstrate diversification in their roles in regulating floral development. Plant Cell **18,** 1833-1845.

de Souza, F.S., Franchini, L.F., and Rubinstein, M. (2013). Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? Molecular biology and evolution**,** mst045.

Deal, R.B., and Henikoff, S. (2010). A simple method for gene expression and chromatin profiling of individual cell types within a tissue. Dev Cell **18,** 1030-1040.

Deal, R.B., and Henikoff, S. (2011a). Histone variants and modifications in plant gene regulation. Current opinion in plant biology **14,** 116-122.

Deal, R.B., and Henikoff, S. (2011b). The INTACT method for cell type-specific gene expression and chromatin profiling in Arabidopsis thaliana **6,** 56-68.

Degner, J.F., Pai, A.A., Pique-Regi, R., Veyrieras, J.B., Gaffney, D.J., Pickrell, J.K., De Leon, S., Michelini, K., Lewellen, N., Crawford, G.E., Stephens, M., Gilad, Y., and Pritchard, J.K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. Nature **482,** 390-394.

Dehal, P., and Boore, J.L. (2005). Two rounds of whole genome duplication in the ancestral vertebrate. PLoS Biol **3**.

Dewey, C.N. (2011). Positional orthology: putting genomic evolutionary relationships into context. Briefings in Bioinformatics **12,** 401-412.

Di Stilio, V.S., Melzer, R., and Hall, J.C. (2017). Editorial: A Broader View for Plant EvoDevo: Novel Approaches for Diverse Model Systems. Frontiers in Plant Science **8,** 61.

Ditta, G., Pinyopich, A., Robles, P., Pelaz, S., and Yanofsky, M.F. (2004). The SEP4 gene of Arabidopsis thaliana functions in floral organ and meristem identity. Curr Biol **14,** 1935-1940.

Doebley, J.F., Gaut, B.S., and Smith, B.D. (2006). The molecular genetics of crop domestication. Cell **127,** 1309-1321.

Domazet-Loso, T., and Tautz, D. (2010). A phylogenetically based transcriptome age index mirrors ontogenetic divergence patterns. Nature **468,** 815-818.

Dreni, L., Pilatone, A., Yun, D., Erreni, S., Pajoro, A., Caporali, E., Zhang, D., and Kater, M.M. (2011). Functional Analysis of All AGAMOUS Subfamily Members in Rice Reveals Their Roles in Reproductive Organ Identity Determination and Meristem Determinacy. The Plant Cell **23,** 2850-2863.

Drillon, G., and Fischer, G. (2011). Comparative study on synteny between yeasts and vertebrates. Comptes Rendus Biologies **334,** 629-638.

Drinnan, A.N., Crane, P.R., and Hoot, S.B. (1994). Patterns of floral evolution in the early diversification of non-magnoliid dicotyledons (eudicots). In Early Evolution of Flowers, P.K. Endress and E.M. Friis, eds (Vienna: Springer Vienna), pp. 93-122.

Dror, I., Golan, T., Levy, C., Rohs, R., and Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment in transcription factor binding across diverse protein families. Genome Res **25,** 1268-1280.

Dubchak, I., Poliakov, A., Kislyuk, A., and Brudno, M. (2009). Multiple whole-genome alignments without a reference organism. Genome Res **19,** 682-689.

Duran, C., Edwards, D., and Batley, J. (2009). Genetic Maps and the Use of Synteny. In Plant Genomics: Methods and Protocols, J.P. Gustafson, P. Langridge, and D.J. Somers, eds (Totowa, NJ: Humana Press), pp. 41-55.

Egea-Cortines, M., Saedler, H., and Sommer, H. (1999). Ternary complex formation between the MADS-box proteins SQUAMOSA, DEFICIENS and GLOBOSA is involved in the control of floral architecture in Antirrhinum majus. Embo J **18,** 5370-5379.

Emmert-Buck, M.R., Bonner, R.F., Smith, P.D., Chuaqui, R.F., Zhuang, Z., Goldstein, S.R., Weiss, R.A., and Liotta, L.A. (1996). Laser capture microdissection. Science **274,** 998-1001.

Endress, P.K. (1994). Floral structure and evolution of primitive angiosperms: Recent advances. Plant Systematics and Evolution **192,** 79-97.

Feng, S., Cokus, S.J., Schubert, V., Zhai, J., Pellegrini, M., and Jacobsen, S.E. (2014). Genome-wide Hi-C analyses in wild-type and mutants reveal high-resolution chromatin interactions in Arabidopsis. Molecular cell **55,** 694-707.

Ferrario, S., Immink, R.G., and Angenent, G.C. (2004). Conservation and diversity in flower land. Curr Opin Plant Biol **7,** 84-91.

Ferrario, S., Immink, R.G.H., Shchennikova, A., Busscher-Lange, J., and Angenent, G.C. (2003). The MADS Box Gene FBP2 Is Required for SEPALLATA Function in Petunia. The Plant Cell **15,** 914-925.

Feschotte, C. (2008). Transposable elements and the evolution of regulatory networks. Nature Reviews Genetics **9,** 397-405.

Force, A., Lynch, M., Pickett, F.B., Amores, A., Yan, Y.-l., and Postlethwait, J. (1999). Preservation of Duplicate Genes by Complementary, Degenerative Mutations. Genetics **151,** 1531-1545.

Frazer, K.A., Pachter, L., Poliakov, A., Rubin, E.M., and Dubchak, I. (2004). VISTA: computational tools for comparative genomics. Nucleic acids research **32,** W273-W279.

Freeling, M., and Thomas, B.C. (2006). Gene-balanced duplications, like tetraploidy, provide predictable drive to increase morphological complexity. Genome Res **16,** 805-814.

Friedman, W.E. (2009). The meaning of Darwin's "abominable mystery". American journal of botany **96,** 5-21.

Friis, E.M., Pedersen, K.R., and Crane, P.R. (2001). Fossil evidence of water lilies (Nymphaeales) in the Early Cretaceous. Nature **410,** 357-360.

Gan, X., Stegle, O., Behr, J., Steffen, J.G., Drewe, P., Hildebrand, K.L., Lyngsoe, R., Schultheiss, S.J., Osborne, E.J., Sreedharan, V.T., Kahles, A., Bohnert, R., Jean, G., Derwent, P., Kersey, P., Belfield, E.J., Harberd, N.P., Kemen, E., Toomajian, C., Kover, P.X., Clark, R.M., Ratsch, G., and Mott, R. (2011). Multiple reference genomes and transcriptomes for Arabidopsis thaliana. Nature **477,** 419-423.

Gan, X., Hay, A., Kwantes, M., Haberer, G., Hallab, A., Ioio, R.D., Hofhuis, H., Pieper, B., Cartolano, M., Neumann, U., Nikolov, L.A., Song, B., Hajheidari, M., Briskine, R., Kougioumoutzi, E., Vlad, D., Broholm, S., Hein, J., Meksem, K., Lightfoot, D., Shimizu, K.K., Shimizu-Inatsugi, R., Imprialou, M., Kudrna, D., Wing, R., Sato, S., Huijser, P., Filatov, D., Mayer, K.F., Mott, R., and Tsiantis, M. (2016). The Cardamine hirsuta genome offers insight into the evolution of morphological diversity. Nat Plants **2,** 16167.

Gebhardt, C., Walkemeier, B., Henselewski, H., Barakat, A., Delseny, M., and Stüber, K. (2003). Comparative mapping between potato (Solanum tuberosum) and Arabidopsis thaliana reveals structurally conserved domains and ancient duplications in the potato genome. The Plant Journal **34,** 529-541.

Geuten, K., and Irish, V. (2010). Hidden variability of floral homeotic B genes in Solanaceae provides a molecular basis for the evolution of novel functions. Plant Cell **22,** 2562-2578.

Geuten, K., Viaene, T., and Irish, V.F. (2011). Robustness and evolvability in the B-system of flower development. Ann Bot **107,** 1545-1556.

Giacomello, S., Salmén, F., Terebieniec, B.K., Vickovic, S., Navarro, J.F., Alexeyenko, A., Reimegård, J., McKee, L.S., Mannapperuma, C., Bulone, V., Ståhl, P.L., Sundström, J.F., Street, N.R., and Lundeberg, J. (2017). Spatially resolved transcriptome profiling in model plant species. Nature Plants **3,** 17061.

Gong, P., Ao, X., Liu, G., Cheng, F., and He, C. (2016). Duplication and Whorl-Specific Down-Regulation of the Obligate AP3-PI Heterodimer Genes Explain the Origin of Paeonia lactiflora Plants with Spontaneous Corolla Mutation. Plant & cell physiology.

Goto, K., and Meyerowitz, E.M. (1994). Function and regulation of the Arabidopsis floral homeotic gene PISTILLATA. Genes Dev **8,** 1548-1560.

Gramzow, L., and Theissen, G. (2010). A hitchhiker's guide to the MADS world of plants. Genome Biol **11,** 1-11.

Guan, Y., Guo, J., Li, H., and Yang, Z. (2013). Signaling in pollen tube growth: crosstalk, feedback, and missing links. Molecular plant **6,** 1053-1064.

Guyot, R., Lefebvre-Pautigny, F., Tranchant-Dubreuil, C., Rigoreau, M., Hamon, P., Leroy, T., Hamon, S., Poncet, V., Crouzillat, D., and de Kochko, A. (2012). Ancestral synteny shared between distantly-related plant species from the asterid (Coffea canephora and Solanum Sp.) and rosid (Vitis vinifera) clades. BMC Genomics **13,** 103-103.

Ha, M. (2013). Understanding the chromatin remodeling code. Plant Science **211,** 137-145.

Hall, J.C., Sytsma, K.J., and Iltis, H.H. (2002). Phylogeny of Capparaceae and Brassicaceae Based on Chloroplast Sequence Data. American journal of botany **89,** 1826-1842.

Hands, P., Vosnakis, N., Betts, D., Irish, V.F., and Drea, S. (2011). Alternate transcripts of a floral developmental regulator have both distinct and redundant functions in opium poppy. Ann Bot **107,** 1557-1566.

Haudry, A., Platts, A.E., Vello, E., Hoen, D.R., Leclercq, M., Williamson, R.J., Forczek, E., Joly-Lopez, Z., Steffen, J.G., Hazzouri, K.M., Dewar, K., Stinchcombe, J.R., Schoen, D.J., Wang, X., Schmutz, J., Town, C.D., Edger, P.P., Pires, J.C., Schumaker, K.S., Jarvis, D.E., Mandakova, T., Lysak, M.A., van den Bergh, E., Schranz, M.E., Harrison, P.M., Moses, A.M., Bureau, T.E., Wright, S.I., and Blanchette, M. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions **45,** 891-898.

Haughn, G.W., and Somerville, C.R. (1988). Genetic control of morphogenesis in Arabidopsis. Dev. Genet. **9,** 73-89.

Hawkins, J.S., Proulx, S.R., Rapp, R.A., and Wendel, J.F. (2009). Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. Proceedings of the National Academy of Sciences **106,** 17811-17816.

Hay, A., and Tsiantis, M. (2006). The genetic basis for differences in leaf form between Arabidopsis thaliana and its wild relative Cardamine hirsuta. Nat Genet **38,** 942-947.

Hay, A., and Tsiantis, M. (2010). KNOX genes: versatile regulators of plant development and diversity. Development **137,** 3153-3165.

He, C., and Saedler, H. (2005). Heterotopic expression of MPF2 is the key to the evolution of the Chinese lantern of Physalis, a morphological novelty in Solanaceae. Proc Natl Acad Sci U S A **102,** 5779-5784.

He, Q., Bardet, A.F., Patton, B., Purvis, J., Johnston, J., Paulson, A., Gogol, M., Stark, A., and Zeitlinger, J. (2011). High conservation of transcription factor binding and evidence for combinatorial regulation across six Drosophila species. Nat Genet **43,** 414-420.

He, X., and Zhang, J. (2005). Rapid subfunctionalization accompanied by prolonged and substantial neofunctionalization in duplicate gene evolution. Genetics **169,** 1157-1164.

Heidel, A.J., Kiefer, C., Coupland, G., and Rose, L.E. (2016). Pinpointing genes underlying annual/perennial transitions with comparative genomics. BMC Genomics **17,** 921.

Heinz, S., Romanoski, C., Benner, C., Allison, K., Kaikkonen, M., Orozco, L., and Glass, C. (2013). Effect of natural genetic variation on enhancer selection and function. Nature **503,** 487-492.

Hénaff, E., Vives, C., Desvoyes, B., Chaurasia, A., Payet, J., Gutierrez, C., and Casacuberta, J.M. (2014). Extensive amplification of the E2F transcription factor binding sites by transposons during evolution of Brassica species. The Plant Journal **77,** 852-862.

Hernández-Hernández, T., Martínez-Castilla, L.P., and Alvarez-Buylla, E.R. (2007). Functional Diversification of B MADS-Box Homeotic Regulators of Flower Development: Adaptive Evolution in Protein–Protein Interaction Domains after Major Gene Duplication Events. Molecular Biology and Evolution **24,** 465-481.

Hesselberth, J.R., Chen, X., Zhang, Z., Sabo, P.J., Sandstrom, R., Reynolds, A.P., Thurman, R.E., Neph, S., Kuehn, M.S., Noble, W.S., Fields, S., and Stamatoyannopoulos, J.A. (2009). Global mapping of protein-DNA interactions in vivo by digital genomic footprinting. Nature methods **6,** 283-289.

Hill, J.P., and Lord, E.M. (1989). Floral development in Arabidopsis thaliana: a comparison of the wild type and the homeotic pistillata mutant. Canadian Journal of Botany **67,** 2922-2936.

Hill, T.A., Day, C.D., Zondlo, S.C., Thackeray, A.G., and Irish, V.F. (1998). Discrete spatial and temporal cis-acting elements regulate transcription of the Arabidopsis floral homeotic gene APETALA3. Development **125,** 1711-1721.

Hodges, S.A., and Kramer, E.M. (2007). Columbines. Curr Biol **17,** R992-994.

Hofberger, J.A., Ramirez, A.M., van den Bergh, E., Zhu, X., Bouwmeester, H.J., Schuurink, R.C., and Schranz, M.E. (2015). Large-Scale Evolutionary Analysis of Genes and Supergene Clusters from Terpenoid Modular Pathways Provides Insights into Metabolic Diversification in Flowering Plants. PLoS One **10,** e0128808.

Hollister, J.D., Smith, L.M., Guo, Y.L., Ott, F., Weigel, D., and Gaut, B.S. (2011). Transposable elements and small RNAs contribute to gene expression divergence between Arabidopsis thaliana and Arabidopsis lyrata. Proc Natl Acad Sci U S A **108,** 2322-2327.

Honma, T., and Goto, K. (2000). The Arabidopsis floral homeotic gene PISTILLATA is regulated by discrete cis-elements responsive to induction and maintenance signals. Development **127,** 2021-2030.

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

Hu, T.T., Pattyn, P., Bakker, E.G., Cao, J., Cheng, J.F., Clark, R.M., Fahlgren, N., Fawcett, J.A., Grimwood, J., Gundlach, H., Haberer, G., Hollister, J.D., Ossowski, S., Ottilar, R.P., Salamov, A.A., Schneeberger, K., Spannagl, M., Wang, X., Yang, L., Nasrallah, M.E., Bergelson, J., Carrington, J.C., Gaut, B.S., Schmutz, J., Mayer, K.F., Van de Peer, Y., Grigoriev, I.V., Nordborg, M., Weigel, D., and Guo, Y.L. (2011). The Arabidopsis lyrata genome sequence and the basis of rapid genome size change. Nat Genet **43,** 476-481.

Huang, H., Mizukami, Y., Hu, Y., and Ma, H. (1993). Isolation and characterization of the binding sequences for the product of the Arabidopsis floral homeotic gene AGAMOUS. Nucleic Acids Res **21,** 4769-4776.

Huang, K., Louis, J.M., Donaldson, L., Lim, F.L., Sharrocks, A.D., and Clore, G.M. (2000). Solution structure of the MEF2A–DNA complex: structural basis for the modulation of DNA bending and specificity by MADS-box transcription factors. Embo J **19,** 2615-2628.

Huminiecki, L., and Wolfe, K.H. (2004). Divergence of spatial gene expression profiles following species-specific gene duplications in human and mouse. Genome Res **14,** 1870-1879.

Hupalo, D., and Kern, A.D. (2013). Conservation and Functional Element Discovery in 20 Angiosperm Plant Genomes. Molecular Biology and Evolution **30,** 1729-1744.

Iglesias, F.M., and Cerdán, P.D. (2016). Maintaining Epigenetic Inheritance During DNA Replication in Plants. Frontiers in Plant Science **7**.

Immink, R.G.H., Tonaco, I.N.A., de Folter, S., Shchennikova, A., van Dijk, A.D.J., Busscher-Lange, J., Borst, J.W., and Angenent, G.C. (2009). SEPALLATA3: the 'glue' for MADS box transcription factor complex formation. Genome Biol **10**.

Irimia, M., Maeso, I., Roy, S.W., and Fraser, H.B. (2013). Ancient cis-regulatory constraints and the evolution of genome architecture. Trends in Genetics **29,** 521-528.

Irish, V.F. (2003). The evolution of floral homeotic gene function. Bioessays **25,** 637-646.

Irish, V.F., and Sussex, I.M. (1990). Function of the apetala-1 gene during Arabidopsis floral development. Plant Cell **2,** 741-753.

Irish, V.F., and Benfey, P.N. (2004). Beyond Arabidopsis. Translational biology meets evolutionary developmental biology. Plant Physiol **135,** 611-614.

Iwasaki, M., and Paszkowski, J. (2014). Epigenetic memory in plants. Embo J **33,** 1987-1998.

Jack, T., Brockman, L.L., and Meyerowitz, E.M. (1992). The homeotic gene APETALA3 of Arabidopsis thaliana encodes a MADS box and is expressed in petals and stamens. Cell **68,** 683-697.

Jack, T., Fox, G.L., and Meyerowitz, E.M. (1994). Arabidopsis homeotic gene APETALA3 ectopic expression: Transcriptional and posttranscriptional regulation determine floral organ identity. Cell **76,** 703-716.

Jaramillo, M.A., and Kramer, E.M. (2007). Molecular evolution of the petal and stamen identity genes, APETALA3 and PISTILLATA, after petal loss in the Piperales. Molecular phylogenetics and evolution **44,** 598-609.

Jetha, K., Theißen, G., and Melzer, R. (2014). Arabidopsis SEPALLATA proteins differ in cooperative DNA-binding during the formation of floral quartet-like complexes. Nucleic Acids Research **42,** 10927-10942.

Jiao, Y., Wickett, N.J., Ayyampalayam, S., Chanderbali, A.S., Landherr, L., Ralph, P.E., Tomsho, L.P., Hu, Y., Liang, H., Soltis, P.S., Soltis, D.E., Clifton, S.W., Schlarbaum, S.E., Schuster, S.C., Ma, H., Leebens-Mack, J., and dePamphilis, C.W. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature **473,** 97-100.

Jiao, Y., Leebens-Mack, J., Ayyampalayam, S., Bowers, J.E., McKain, M.R., McNeal, J., Rolf, M., Ruzicka, D.R., Wafula, E., Wickett, N.J., Wu, X., Zhang, Y., Wang, J., Zhang, Y., Carpenter, E.J., Deyholos, M.K., Kutchan, T.M., Chanderbali, A.S., Soltis, P.S., Stevenson, D.W., McCombie, R., Pires, J.C., Wong, G.K.-S., Soltis, D.E., and dePamphilis, C.W. (2012). A genome triplication associated with early diversification of the core eudicots. Genome Biol **13,** 1-14.

Jinek, M., Chylinski, K., Fonfara, I., Hauer, M., Doudna, J.A., and Charpentier, E. (2012). A Programmable Dual-RNA–Guided DNA Endonuclease in Adaptive Bacterial Immunity. Science **337,** 816-821.

Johnson, R., Gamblin, R.J., Ooi, L., Bruce, A.W., Donaldson, I.J., Westhead, D.R., Wood, I.C., Jackson, R.M., and Buckley, N.J. (2006). Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. Nucleic acids research **34,** 3862-3877.

Johnston, J.S., Pepper, A.E., Hall, A.E., Chen, Z.J., Hodnett, G., Drabek, J., Lopez, R., and Price, H.J. (2005). Evolution of genome size in Brassicaceae. Ann Bot **95,** 229-235.

Jolma, A., Kivioja, T., Toivonen, J., Cheng, L., Wei, G., Enge, M., Taipale, M., Vaquerizas, J.M., Yan, J., Sillanpaa, M.J., Bonke, M., Palin, K., Talukder, S., Hughes, T.R., Luscombe, N.M., Ukkonen, E., and Taipale, J. (2010). Multiplexed massively parallel SELEX for characterization of human transcription factor binding specificities. Genome Res **20,** 861-873.

Jones, F.C., Grabherr, M.G., Chan, Y.F., Russell, P., Mauceli, E., Johnson, J., Swofford, R., Pirun, M., Zody, M.C., White, S., Birney, E., Searle, S., Schmutz, J., Grimwood, J., Dickson, M.C., Myers, R.M., Miller, C.T., Summers, B.R., Knecht, A.K., Brady, S.D., Zhang, H., Pollen, A.A., Howes, T., Amemiya, C., Broad Institute Genome Sequencing, P., Whole Genome Assembly, T.,

Baldwin, J., Bloom, T., Jaffe, D.B., Nicol, R., Wilkinson, J., Lander, E.S., Di Palma, F., Lindblad-Toh, K., and Kingsley, D.M. (2012). The genomic basis of adaptive evolution in threespine sticklebacks. Nature **484,** 55-61.

Joosen, R.V., Ligterink, W., Hilhorst, H.W., and Keurentjes, J.J. (2009). Advances in genetical genomics of plants. Current genomics **10,** 540-549.

Jung, S., Jiwan, D., Cho, I., Lee, T., Abbott, A., Sosinski, B., and Main, D. (2009). Synteny of Prunus and other model plant species. BMC Genomics **10,** 76.

Kalinka, A.T., Varga, K.M., Gerrard, D.T., Preibisch, S., Corcoran, D.L., Jarrells, J., Ohler, U., Bergman, C.M., and Tomancak, P. (2010). Gene expression divergence recapitulates the developmental hourglass model. Nature **468,** 811-814.

Kang, H.-G., Jeon, J.-S., Lee, S., and An, G. (1998). Identification of class B and class C floral organ identity genes from rice plants. Plant Mol Biol **38,** 1021-1029.

Kanno, A., Saeki, H., Kameya, T., Saedler, H., and Theissen, G. (2003). Heterotopic expression of class B floral homeotic genes supports a modified ABC model for tulip (Tulipa gesneriana). Plant Mol Biol **52,** 831-841.

Kaufmann, K., Melzer, R., and Theissen, G. (2005). MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. Gene **347,** 183-198.

Kaufmann, K., Pajoro, A., and Angenent, G.C. (2010a). Regulation of transcription in plants: mechanisms controlling developmental switches. Nature reviews. Genetics **11,** 830-842.

Kaufmann, K., Muino, J.M., Osteras, M., Farinelli, L., Krajewski, P., and Angenent, G.C. (2010b). Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). Nature protocols **5,** 457-472.

Kaufmann, K., Muino, J.M., Jauregui, R., Airoldi, C.A., Smaczniak, C., Krajewski, P., and Angenent, G.C. (2009). Target genes of the MADS transcription factor SEPALLATA3: integration of developmental and hormonal pathways in the Arabidopsis flower. PLoS Biol **7,** e1000090.

Kaufmann, K., Wellmer, F., Muino, J.M., Ferrier, T., Wuest, S.E., Kumar, V., Serrano-Mislata, A., Madueno, F., Krajewski, P., Meyerowitz, E.M., Angenent, G.C., and Riechmann, J.L. (2010c). Orchestration of floral initiation by APETALA1. Science **328,** 85-89.

Kawakatsu, T., Huang, S.S., Jupe, F., Sasaki, E., Schmitz, R.J., Urich, M.A., Castanon, R., Nery, J.R., Barragan, C., He, Y., Chen, H., Dubin, M., Lee, C.R., Wang, C., Bemm, F., Becker, C., O'Neil, R., O'Malley, R.C., Quarless, D.X., Genomes, C., Schork, N.J., Weigel, D., Nordborg, M., and Ecker, J.R. (2016). Epigenomic Diversity in a Global Collection of Arabidopsis thaliana Accessions. Cell **166,** 492-505.

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S.L. (2013). TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol **14,** 1.

Kim, S., Yoo, M.J., Albert, V.A., Farris, J.S., Soltis, P.S., and Soltis, D.E. (2004). Phylogeny and diversification of B-function MADS-box genes in angiosperms: evolutionary and functional implications of a 260-million-year-old duplication. American journal of botany **91,** 2102-2118.

Komaki, M.K., Okada, K., Nishino, E., and Shimura, Y. (1988). Isolation and characterization of novel mutants of Arabidopsis thaliana defective in flower development. Development **104,** 195-203.

Konishi, S., Izawa, T., Lin, S.Y., Ebana, K., Fukuta, Y., Sasaki, T., and Yano, M. (2006). An SNP caused loss of seed shattering during rice domestication. Science **312,** 1392-1396.

Kramer, E., Di Stilio, V., and Schlüter, P. (2003). Complex Patterns of Gene Duplication in the APETALA3 and PISTILLATA Lineages of the Ranunculaceae. International Journal of Plant Sciences **164,** 1-11.

Kramer, E., Jaramillo, M., and Di Stilio, V. (2004). Patterns of Gene Duplication and Functional Evolution During the Diversification of the *AGAMOUS* Subfamily of MADS Box Genes in Angiosperms. Genetics **166,** 1011-1023.

Kramer, E.M. (2009). Aquilegia: A New Model for Plant Development, Ecology, and Evolution. Annual Review of Plant Biology **60,** 261-277.

Kramer, E.M., Dorit, R.L., and Irish, V.F. (1998). Molecular evolution of genes controlling petal and stamen development: duplication and divergence within the APETALA3 and PISTILLATA MADS-box gene lineages. Genetics **149,** 765-783.

Kramer, E.M., Holappa, L., Gould, B., Jaramillo, M.A., Setnikov, D., and Santiago, P.M. (2007). Elaboration of B gene function to include the identity of novel floral organs in the lower eudicot Aquilegia. Plant Cell **19,** 750-766.

Krizek, B.A., and Meyerowitz, E.M. (1996a). Mapping the protein regions responsible for the functional specificities of the Arabidopsis MADS domain organ-identity proteins. Proceedings of the National Academy of Sciences **93,** 4063-4070.

Krizek, B.A., and Meyerowitz, E.M. (1996b). The Arabidopsis homeotic genes APETALA3 and PISTILLATA are sufficient to provide the B class organ identity function. Development **122,** 11-22.

Krizek, B.A., and Fletcher, J.C. (2005). Molecular mechanisms of flower development: an armchair guide. Nature reviews. Genetics **6,** 688-698.

Krogan, N.T., and Ashton, N.W. (2000). Ancestry of plant MADS-box genes revealed by bryophyte (Physcomitrella patens) homologues. New Phytologist **147,** 505-517.

Kulahoglu, C., Denton, A.K., Sommer, M., Mass, J., Schliesky, S., Wrobel, T.J., Berckmans, B., Gongora-Castillo, E., Buell, C.R., Simon, R., De Veylder, L., Brautigam, A., and Weber, A.P. (2014). Comparative transcriptome atlases reveal altered gene expression modules between two Cleomaceae C3 and C4 plant species. Plant Cell **26,** 3243-3260.

Kunarso, G., Chia, N.-Y., Jeyakani, J., Hwang, C., Lu, X., Chan, Y.-S., Ng, H.-H., and Bourque, G. (2010). Transposable elements have rewired the core regulatory network of human embryonic stem cells. Nat Genet **42,** 631-634.

Kunst, L., Klenz, J.E., Martinez-Zapater, J., and Haughn, G.W. (1989). AP2 Gene Determines the Identity of Perianth Organs in Flowers of Arabidopsis thaliana. The Plant Cell **1,** 1195-1208.

Lamb, R.S., and Irish, V.F. (2003). Functional divergence within the APETALA3/PISTILLATA floral homeotic gene lineages. Proc Natl Acad Sci U S A **100,** 6558-6563.

Lange, M., Orashakova, S., Lange, S., Melzer, R., Theissen, G., Smyth, D.R., and Becker, A. (2013). The seirena B class floral homeotic mutant of California Poppy (Eschscholzia californica) reveals a function of the enigmatic PI motif in the formation of specific multimeric MADS domain protein complexes. Plant Cell **25,** 438-453.

Lawton-Rauh, A. (2003). Evolutionary dynamics of duplicated genes in plants. Molecular phylogenetics and evolution **29,** 396-409.

Lee, E.K., Cibrian-Jaramillo, A., Kolokotronis, S.O., Katari, M.S., Stamatakis, A., Ott, M., Chiu, J.C., Little, D.P., Stevenson, D.W., McCombie, W.R., Martienssen, R.A., Coruzzi, G., and Desalle, R. (2011). A functional phylogenomic view of the seed plants. PLoS Genet **7,** e1002411.

Lescot, M., Piffanelli, P., Ciampi, A.Y., Ruiz, M., Blanc, G., Leebens-Mack, J., da Silva, F.R., Santos, C.M., D'Hont, A., Garsmeur, O., Vilarinhos, A.D., Kanamori, H., Matsumoto, T., Ronning, C.M., Cheung, F., Haas, B.J., Althoff, R., Arbogast, T., Hine, E., Pappas, G.J., Sasaki, T., Souza, M.T., Miller, R.N., Glaszmann, J.-C., and Town, C.D. (2008). Insights into the Musa genome: Syntenic relationships to rice and between Musa species. BMC Genomics **9,** 58.

Leseberg, C.H., Eissler, C.L., Wang, X., Johns, M.A., Duvall, M.R., and Mao, L. (2008). Interaction study of MADS-domain proteins in tomato. J Exp Bot **59,** 2253-2265.

Levo, M., and Segal, E. (2014). In pursuit of design principles of regulatory sequences. Nature reviews. Genetics **15,** 453-468.

Li, R., Yu, C., Li, Y., Lam, T.-W., Yiu, S.-M., Kristiansen, K., and Wang, J. (2009). SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics **25,** 1966-1967.

Li, X.Y., Thomas, S., Sabo, P.J., Eisen, M.B., Stamatoyannopoulos, J.A., and Biggin, M.D. (2011). The role of chromatin accessibility in directing the widespread, overlapping patterns of Drosophila transcription factor binding. Genome Biol **12,** R34.

Li, X.Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C.L., Chu, H.C., Ogawa, N., Inwood, W., Sementchenko, V., Beaton, A., Weiszmann, R., Celniker, S.E., Knowles, D.W., Gingeras, T., Speed, T.P., Eisen, M.B., and Biggin, M.D. (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol **6,** e27.

Li, Z., Baniaga, A.E., Sessa, E.B., Scascitelli, M., Graham, S.W., Rieseberg, L.H., and Barker, M.S. (2015). Early genome duplications in conifers and other seed plants. Science Advances **1,** e1501084.

Liljegren, S.J., Ditta, G.S., Eshed, Y., Savidge, B., Bowman, J.L., and Yanofsky, M.F. (2000). SHATTERPROOF MADS-box genes control seed dispersal in Arabidopsis. Nature **404,** 766-770.

Litt, A. (2007). An Evaluation of A-Function: Evidence from the APETALA1 and APETALA2 Gene Lineages. International Journal of Plant Sciences **168,** 73-91.

Litt, A., and Kramer, E.M. (2010). The ABC model and the diversification of floral organ identity. Seminars in cell & developmental biology **21,** 129-137.

Liu, Y., Nakayama, N., Schiff, M., Litt, A., Irish, V.F., and Dinesh-Kumar, S.P. (2004). Virus Induced Gene Silencing of a DEFICIENS Ortholog in Nicotiana Benthamiana. Plant Mol Biol **54,** 701-711.

Long, Q., Rabanal, F.A., Meng, D., Huber, C.D., Farlow, A., Platzer, A., Zhang, Q., Vilhjalmsson, B.J., Korte, A., Nizhynska, V., Voronin, V., Korte, P., Sedman, L., Mandakova, T., Lysak, M.A., Seren, U., Hellmann, I., and Nordborg, M. (2013). Massive genomic variation and strong selection in Arabidopsis thaliana lines from Sweden. Nat Genet **45,** 884-890.

Lönnig, W.-E., and Saedler, H. (1994). The homeotic Macho mutant of Antirrhinum majus reverts to wild-type or mutates to the homeotic plena phenotype. Molecular and General Genetics MGG **245,** 636-643.

Lozano-Juste, J., and Cutler, S.R. (2014). Plant genome engineering in full bloom. Trends Plant Sci **19,** 284-287.

Lynch, M., and Conery, J.S. (2000). The evolutionary fate and consequences of duplicate genes. Science **290,** 1151-1155.

Lynch, M., and Conery, J.S. (2003). The evolutionary demography of duplicate genes. Journal of structural and functional genomics **3,** 35-44.

Lynch, M., Bobay, L.-M., Catania, F., Gout, J.-F., and Rho, M. (2011). The repatterning of eukaryotic genomes by random genetic drift. Annual review of genomics and human genetics **12,** 347.

Mackay, T.F., Stone, E.A., and Ayroles, J.F. (2009). The genetics of quantitative traits: challenges and prospects. Nature reviews. Genetics **10,** 565-577.

Maere, S., Heymans, K., and Kuiper, M. (2005). BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. Bioinformatics **21,** 3448-3449.

Magallon, S., Gomez-Acevedo, S., Sanchez-Reyes, L.L., and Hernandez-Hernandez, T. (2015). A metacalibrated time-tree documents the early rise of flowering plant phylogenetic diversity. New Phytol **207,** 437-453.

Maizel, A., Busch, M.A., Tanahashi, T., Perkovic, J., Kato, M., Hasebe, M., and Weigel, D. (2005). The floral regulator LEAFY evolves by substitutions in the DNA binding domain. Science **308,** 260-263.

Malcomber, S.T., and Kellogg, E.A. (2005). SEPALLATA gene diversification: brave new whorls. Trends Plant Sci **10,** 427-435.

Mandel, M.A., Gustafson-Brown, C., Savidge, B., and Yanofsky, M.F. (1992). Molecular characterization of the Arabidopsis floral homeotic gene APETALA1. Nature **360,** 273-277.

Mao, W.-T., Hsu, H.-F., Hsu, W.-H., Li, J.-Y., Lee, Y.-I., and Yang, C.-H. (2015). The C-Terminal Sequence and PI motif of the Orchid (Oncidium Gower Ramsey) PISTILLATA (PI) Ortholog

Determine its Ability to Bind AP3 Orthologs and Enter the Nucleus to Regulate Downstream Genes Controlling Petal and Stamen Formation. Plant and Cell Physiology **56,** 2079-2099.

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,** 865-872.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet. **7,** 29-59.

Mathews, S., and Donoghue, M.J. (1999). The root of angiosperm phylogeny inferred from duplicate phytochrome genes. Science **286,** 947-950.

Maumus, F., and Quesneville, H. (2014). Ancestral repeats have shaped epigenome and genome composition for millions of years in Arabidopsis thaliana. Nature communications **5**.

McGinnis, W., and Krumlauf, R. (1992). Homeobox genes and axial patterning. Cell **68,** 283-302.

McGinnis, W., Garber, R.L., Wirz, J., Kuroiwa, A., and Gehring, W.J. (1984). A homologous protein-coding sequence in drosophila homeotic genes and its conservation in other metazoans. Cell **37,** 403-408.

McGonigle, B., Bouhidel, K., and Irish, V.F. (1996). Nuclear localization of the Arabidopsis APETALA3 and PISTILLATA homeotic gene products depends on their simultaneous expression. Genes Dev **10,** 1812-1821.

Melzer, R., and Theissen, G. (2009). Reconstitution of 'floral quartets' in vitro involving class B and class E floral homeotic proteins. Nucleic Acids Res **37,** 2723-2736.

Melzer, R., Verelst, W., and Theissen, G. (2009). The class E floral homeotic protein SEPALLATA3 is sufficient to loop DNA in 'floral quartet'-like complexes in vitro. Nucleic Acids Res **37,** 144-157.

Melzer, R., Härter, A., Rümpler, F., Kim, S., Soltis, P.S., Soltis, D.E., and Theißen, G. (2014). DEF- and GLO-like proteins may have lost most of their interaction partners during angiosperm evolution. Annals of Botany.

Mena, M., Ambrose, B.A., Meeley, R.B., Briggs, S.P., Yanofsky, M.F., and Schmidt, R.J. (1996). Diversification of C-Function Activity in Maize Flower Development. Science **274,** 1537-1540.

Mendes, M.A., Guerra, R.F., Berns, M.C., Manzo, C., Masiero, S., Finzi, L., Kater, M.M., and Colombo, L. (2013). MADS Domain Transcription Factors Mediate Short-Range DNA Looping That Is Essential for Target Gene Expression in Arabidopsis. The Plant Cell **25,** 2560-2572.

Ming, R., Hou, S., Feng, Y., Yu, Q., Dionne-Laporte, A., Saw, J.H., Senin, P., Wang, W., Ly, B.V., Lewis, K.L., Salzberg, S.L., Feng, L., Jones, M.R., Skelton, R.L., Murray, J.E., Chen, C., Qian, W., Shen, J., Du, P., Eustice, M., Tong, E., Tang, H., Lyons, E., Paull, R.E., Michael, T.P., Wall, K., Rice, D.W., Albert, H., Wang, M.L., and Zhu, Y.J. (2008). The draft genome of the transgenic tropical fruit tree papaya (Carica papaya Linnaeus). Nature **452**.

Moghe, G.D., and Shiu, S.H. (2014). The causes and molecular consequences of polyploidy in flowering plants. Ann N Y Acad Sci **1320,** 16-34.

Mondragon-Palomino, M., and Theissen, G. (2008). MADS about the evolution of orchid flowers. Trends Plant Sci **13,** 51-59.

Mondragon-Palomino, M., and Theissen, G. (2009). Why are orchid flowers so diverse? Reduction of evolutionary constraints by paralogues of class B floral homeotic genes. Ann Bot **104,** 583-594.

Mondragón-Palomino, M., and Theißen, G. (2011). Conserved differential expression of paralogous DEFICIENS- and GLOBOSA-like MADS-box genes in the flowers of Orchidaceae: refining the 'orchid code'. The Plant Journal **66,** 1008-1019.

Moore, M.J., Bell, C.D., Soltis, P.S., and Soltis, D.E. (2007). Using plastid genome-scale data to resolve enigmatic relationships among basal angiosperms. Proceedings of the National Academy of Sciences **104,** 19363-19368.

Moore, R.C., and Purugganan, M.D. (2005). The evolutionary dynamics of plant duplicate genes. Curr Opin Plant Biol **8,** 122-128.

Moyroud, E., Monniaux, M., Thevenon, E., Dumas, R., Scutt, C.P., Frohlich, M.W., and Parcy, F. (2017). A link between LEAFY and B-gene homologues in Welwitschia mirabilis sheds light on ancestral mechanisms prefiguring floral development. New Phytol.

Moyroud, E., Minguet, E.G., Ott, F., Yant, L., Pose, D., Monniaux, M., Blanchet, S., Bastien, O., Thevenon, E., Weigel, D., Schmid, M., and Parcy, F. (2011). Prediction of regulatory interactions from genome sequences using a biophysical model for the Arabidopsis LEAFY transcription factor. The Plant cell **23,** 1293-1306.

Mudge, J., Cannon, S.B., Kalo, P., Oldroyd, G.E., Roe, B.A., Town, C.D., and Young, N.D. (2005). Highly syntenic regions in the genomes of soybean, Medicago truncatula, and Arabidopsis thaliana. BMC Plant Biol **5,** 15.

Muiño, J.M., Hoogstraat, M., van Ham, R.C., and van Dijk, A.D. (2011). PRI-CAT: a web-tool for the analysis, storage and visualization of plant ChIP-seq experiments. Nucleic Acids Res **39,** W524-527.

Muiño, J.M., Smaczniak, C., Angenent, G.C., Kaufmann, K., and van Dijk, A.D.J. (2014). Structural determinants of DNA recognition by plant MADS-domain transcription factors. Nucleic Acids Research **42,** 2138-2146.

Muiño, J.M., de Bruijn, S., Pajoro, A., Geuten, K., Vingron, M., Angenent, G.C., and Kaufmann, K. (2016). Evolution of DNA-Binding Sites of a Floral Master Regulatory Transcription Factor. Mol Biol Evol **33,** 185-200.

Munster, T., Pahnke, J., Di Rosa, A., Kim, J.T., Martin, W., Saedler, H., and Theissen, G. (1997). Floral homeotic genes were recruited from homologous MADS-box genes preexisting in the common ancestor of ferns and seed plants. Proc Natl Acad Sci U S A **94,** 2415-2420.

Munz, P. (1946). Aquilegia; The cultivated and wild Columbines. (the Bailey Hortorium of the new york state college of Agriculture at Cornell University).

Nagasawa, N., Miyoshi, M., Sano, Y., Satoh, H., Hirano, H., Sakai, H., and Nagato, Y. (2003). SUPERWOMAN1 and DROOPING LEAF genes control floral organ identity in rice. Development **130,** 705-718.

Nakamura, T., Fukuda, T., Nakano, M., Hasebe, M., Kameya, T., and Kanno, A. (2005). The modified ABC model explains the development of the petaloid perianth of Agapanthus praecox ssp. orientalis (Agapanthaceae) flowers. Plant Mol Biol **58,** 435-445.

Nardmann, J., Zimmermann, R., Durantini, D., Kranz, E., and Werr, W. (2007). WOX gene phylogeny in Poaceae: a comparative approach addressing leaf and embryo development. Mol Biol Evol **24,** 2474-2484.

Ni, X., Zhang, Y.E., Nègre, N., Chen, S., Long, M., and White, K.P. (2012). Adaptive evolution and the birth of CTCF binding sites in the Drosophila genome. PLoS Biol **10,** e1001420.

Norman, C., Runswick, M., Pollock, R., and Treisman, R. (1988). Isolation and properties of cDNA clones encoding SRF, a transcription factor that binds to the c-fos serum response element. Cell **55,** 989-1003.

Nowak, M.A., Boerlijst, M.C., Cooke, J., and Smith, J.M. (1997). Evolution of genetic redundancy **388,** 167-171.

Noyes, M.B., Christensen, R.G., Wakabayashi, A., Stormo, G.D., Brodsky, M.H., and Wolfe, S.A. (2008). Analysis of Homeodomain Specificities Allows the Family-wide Prediction of Preferred Recognition Sites. Cell **133,** 1277-1289.

O'Malley, Ronan C., Huang, S.-shan C., Song, L., Lewsey, Mathew G., Bartlett, A., Nery, Joseph R., Galli, M., Gallavotti, A., and Ecker, Joseph R. (2016). Cistrome and Epicistrome Features Shape the Regulatory DNA Landscape. Cell **165,** 1280-1292.

Ó'Maoiléidigh, D.S., Wuest, S.E., Rae, L., Raganelli, A., Ryan, P.T., Kwaśniewska, K., Das, P., Lohan, A.J., Loftus, B., and Graciet, E. (2013). Control of reproductive floral organ identity specification in Arabidopsis by the C function regulator AGAMOUS. The Plant Cell **25,** 2482-2503.

Ohno, S. (1970). Evolution by Gene Duplication.

Ossowski, S., Schneeberger, K., Clark, R.M., Lanz, C., Warthmann, N., and Weigel, D. (2008). Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res **18,** 2024-2033.

Otani, M., Sharifi, A., Kubota, S., Oizumi, K., Uetake, F., Hirai, M., Hoshino, Y., Kanno, A., and Nakano, M. (2016). Suppression of B function strongly supports the modified ABCE model in Tricyrtis sp. (Liliaceae). Scientific Reports **6,** 24549.

Pajoro, A., Biewers, S., Dougali, E., Leal Valentim, F., Mendes, M.A., Porri, A., Coupland, G., Van de Peer, Y., van Dijk, A.D., Colombo, L., Davies, B., and Angenent, G.C. (2014a). The (r)evolution of gene regulatory networks controlling Arabidopsis plant reproduction: a two-decade history. J Exp Bot **65,** 4731-4745.

Pajoro, A., Madrigal, P., Muino, J.M., Matus, J.T., Jin, J., Mecchia, M.A., Debernardi, J.M., Palatnik, J.F., Balazadeh, S., Arif, M., O'Maoileidigh, D.S., Wellmer, F., Krajewski, P., Riechmann, J.L., Angenent, G.C., and Kaufmann, K. (2014b). Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. Genome Biol **15,** R41.

Pan, Z.-J., Chen, Y.-Y., Du, J.-S., Chen, Y.-Y., Chung, M.-C., Tsai, W.-C., Wang, C.-N., and Chen, H.-H. (2014). Flower development of Phalaenopsis orchid involves functionally divergent SEPALLATA-like genes. New Phytologist **202,** 1024-1042.

Panchy, N., Lehti-Shiu, M., and Shiu, S.-H. (2016). Evolution of Gene Duplication in Plants. Plant Physiol **171,** 2294-2316.

Parenicova, L., de Folter, S., Kieffer, M., Horner, D.S., Favalli, C., Busscher, J., Cook, H.E., Ingram, R.M., Kater, M.M., Davies, B., Angenent, G.C., and Colombo, L. (2003). Molecular and phylogenetic analyses of the complete MADS-box transcription factor family in Arabidopsis: new openings to the MADS world. Plant Cell **15,** 1538-1551.

Park, S.J., Jiang, K., Schatz, M.C., and Lippman, Z.B. (2012). Rate of meristem maturation determines inflorescence architecture in tomato. Proc Natl Acad Sci U S A **109,** 639-644.

Parkinson, C.L., Adams, K.L., and Palmer, J.D. (1999). Multigene analyses identify the three earliest lineages of extant flowering plants. Curr Biol **9,** 1485-1488.

Pasquinelli, A.E. (2012). MicroRNAs and their targets: recognition, regulation and an emerging reciprocal relationship. Nature reviews. Genetics **13,** 271-282.

Passmore, S., Maine, G.T., Elble, R., Christ, C., and Tye, B.K. (1988). Saccharomyces cerevisiae protein involved in plasmid maintenance is necessary for mating of MAT alpha cells. Journal of molecular biology **204,** 593-606.

Patchell, M.J., Roalson, E.H., and Hall, J.C. (2014). Resolved phylogeny of Cleomaceae based on all three genomes. Taxon **63,** 315-328.

Patchell, M.J., Bolton, M.C., Mankowski, P., and Hall, J.C. (2011). Comparative Floral Development in Cleomaceae Reveals Two Distinct Pathways Leading to Monosymmetry. International Journal of Plant Sciences **172,** 352-365.

Paterson, A.H., Freeling, M., Tang, H., and Wang, X. (2010). Insights from the comparison of plant genome sequences. Annual review of plant biology **61,** 349-372.

Pelaz, S., Tapia-López, R., Alvarez-Buylla, E.R., and Yanofsky, M.F. (2001). Conversion of leaves into petals in Arabidopsis. Current Biology **11,** 182-184.

Pelaz, S., Ditta, G.S., Baumann, E., Wisman, E., and Yanofsky, M.F. (2000). B and C floral organ identity functions require SEPALLATA MADS-box genes. Nature **405,** 200-203.

Pellegrini, L., Tan, S., and Richmond, T.J. (1995). Structure of serum response factor core bound to DNA. Nature **376,** 490-498.

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.-C., Mendell, J.T., and Salzberg, S.L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nature biotechnology **33,** 290-295.

Pfluger, J., and Wagner, D. (2007). Histone modifications and dynamic regulation of genome accessibility in plants. Current Opinion in Plant Biology **10,** 645-652.

Pique-Regi, R., Degner, J.F., Pai, A.A., Gaffney, D.J., Gilad, Y., and Pritchard, J.K. (2011). Accurate inference of transcription factor binding from DNA sequence and chromatin accessibility data. Genome Res **21,** 447-455.

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 Physiol **145,** 1495-1505.

Preston, J.C., and Hileman, L.C. (2009). Developmental genetics of floral symmetry evolution. Trends Plant Sci **14,** 147-154.

Proost, S., Pattyn, P., Gerats, T., and Van de Peer, Y. (2011). Journey through the past: 150 million years of plant genome evolution. The Plant Journal **66,** 58-65.

Proost, S., Van Bel, M., Vaneechoutte, D., Van de Peer, Y., Inzé, D., Mueller-Roeber, B., and Vandepoele, K. (2015). PLAZA 3.0: an access point for plant comparative genomics. Nucleic Acids Research **43,** D974-D981.

Prud'homme, B., Gompel, N., and Carroll, S.B. (2007). Emerging principles of regulatory evolution. Proc Natl Acad Sci U S A **104 Suppl 1,** 8605-8612.

Puranik, S., Acajjaoui, S., Conn, S., Costa, L., Conn, V., Vial, A., Marcellin, R., Melzer, R., Brown, E., Hart, D., Theissen, G., Silva, C.S., Parcy, F., Dumas, R., Nanao, M., and Zubieta, C. (2014). Structural basis for the oligomerization of the MADS domain transcription factor SEPALLATA3 in Arabidopsis. Plant Cell **26,** 3603-3615.

Purugganan, M.D. (1997). The MADS-box floral homeotic gene lineages predate the origin of seed plants: phylogenetic and molecular clock estimates. Journal of molecular evolution **45,** 392-396.

Purugganan, M.D., Rounsley, S.D., Schmidt, R.J., and Yanofsky, M.F. (1995). Molecular evolution of flower development: diversification of the plant MADS-box regulatory gene family. Genetics **140,** 345-356.

Qiu, Y.L., Lee, J., Bernasconi-Quadroni, F., Soltis, D.E., Soltis, P.S., Zanis, M., Zimmer, E.A., Chen, Z., Savolainen, V., and Chase, M.W. (1999). The earliest angiosperms: evidence from mitochondrial, plastid and nuclear genomes. Nature **402,** 404-407.

Quadrana, L., Silveira, A.B., Mayhew, G.F., LeBlanc, C., Martienssen, R.A., Jeddeloh, J.A., and Colot, V. (2016). The Arabidopsis thaliana mobilome and its impact at the species level. Elife **5,** e15716.

Rasmussen, D.A., Kramer, E.M., and Zimmer, E.A. (2009). One size fits all? Molecular evidence for a commonly inherited petal identity program in Ranunculales. American journal of botany **96,** 96-109.

Rebeiz, M., Jikomes, N., Kassner, V.A., and Carroll, S.B. (2011). Evolutionary origin of a novel gene expression pattern through co-option of the latent activities of existing regulatory sequences. Proc Natl Acad Sci U S A **108,** 10036-10043.

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

Riechmann, J.L., Wang, M., and Meyerowitz, E.M. (1996a). DNA-binding properties of Arabidopsis MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA and AGAMOUS. Nucleic Acids Res **24,** 3134-3141.

Riechmann, J.L., Krizek, B.A., and Meyerowitz, E.M. (1996b). Dimerization specificity of Arabidopsis MADS domain homeotic proteins APETALA1, APETALA3, PISTILLATA, and AGAMOUS. Proc Natl Acad Sci U S A **93,** 4793-4798.

Rijpkema, A.S., Royaert, S., Zethof, J., van der Weerden, G., Gerats, T., and Vandenbussche, M. (2006). Analysis of the Petunia TM6 MADS box gene reveals functional divergence within the DEF/AP3 lineage. Plant Cell **18,** 1819-1832.

Rodríguez-Mega, E., Piñeyro-Nelson, A., Gutierrez, C., García-Ponce, B., Sánchez, M.D.L.P., Zluhan-Martínez, E., Álvarez-Buylla, E.R., and Garay-Arroyo, A. (2015). Role of transcriptional regulation in the evolution of plant phenotype: a dynamic systems approach. Developmental Dynamics **244,** 1074-1095.

Roque, E., Serwatowska, J., Cruz Rochina, M., Wen, J., Mysore, K.S., Yenush, L., Beltrán, J.P., and Cañas, L.A. (2013). Functional specialization of duplicated AP3-like genes in Medicago truncatula. The Plant Journal **73,** 663-675.

Roque, E., Fares, M.A., Yenush, L., Rochina, M.C., Wen, J., Mysore, K.S., Gomez-Mena, C., Beltran, J.P., and Canas, L.A. (2016). Evolution by gene duplication of Medicago truncatula PISTILLATA-like transcription factors. J Exp Bot.

Rosin, F.M., and Kramer, E.M. (2009). Old dogs, new tricks: regulatory evolution in conserved genetic modules leads to novel morphologies in plants. Developmental biology **332,** 25-35.

Roudier, F.C.O., Ahmed, I., Rard, C.B.E., Sarazin, A., Mary-Huard, T., Cortijo, S., Bouyer, D., Caillieux, E., Duvernois-Berthet, E., Al-Shikhley, L., Giraut, L.E.N., Drevensek, S.E.P., Barneche, F.E.D., Rozier, S.D.E., Brunaud, V.E.R., Aubourg, S.E.B., Schnittger, A., Bowler, C., Martin-Magniette, M.-L., Robin, S.E.P., Caboche, M., and Colot, V. (2011). Integrative epigenomic mapping defines four main chromatin states in Arabidopsis. Embo J **30**.

Ruelens, P., de Maagd, R.A., Proost, S., Theißen, G., Geuten, K., and Kaufmann, K. (2013). FLOWERING LOCUS C in monocots and the tandem origin of angiosperm-specific MADS-box genes. Nature Communications **4,** 2280.

Ruokolainen, S., Ng, Y.P., Albert, V.A., Elomaa, P., and Teeri, T.H. (2010). Large scale interaction analysis predicts that the Gerbera hybrida floral E function is provided both by general and specialized proteins. BMC Plant Biol **10,** 129.

Sanderson, M.J., Thorne, J.L., Wikström, N., and Bremer, K. (2004). Molecular evidence on plant divergence times. American journal of botany **91,** 1656-1665.

Schemske, D.W., and Bradshaw, H.D., Jr. (1999). Pollinator preference and the evolution of floral traits in monkeyflowers (Mimulus). Proc Natl Acad Sci U S A **96,** 11910-11915.

Schmidt, D., Schwalie, P.C., Wilson, M.D., Ballester, B., Gonçalves, Â., Kutter, C., Brown, G.D., Marshall, A., Flicek, P., and Odom, D.T. (2012). Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. Cell **148,** 335-348.

Schmidt, D., Wilson, M.D., Ballester, B., Schwalie, P.C., Brown, G.D., Marshall, A., Kutter, C., Watt, S., Martinez-Jimenez, C.P., Mackay, S., Talianidis, I., Flicek, P., and Odom, D.T. (2010). Five-vertebrate ChIP-seq reveals the evolutionary dynamics of transcription factor binding. Science **328,** 1036-1040.

Schmidt, R.J., Veit, B., Mandel, M.A., Mena, M., Hake, S., and Yanofsky, M.F. (1993). Identification and molecular characterization of ZAG1, the maize homolog of the Arabidopsis floral homeotic gene AGAMOUS. The Plant Cell **5,** 729-737.

Schmitz, R.J., Schultz, M.D., Urich, M.A., Nery, J.R., Pelizzola, M., Libiger, O., Alix, A., McCosh, R.B., Chen, H., Schork, N.J., and Ecker, J.R. (2013). Patterns of population epigenomic diversity. Nature **495,** 193-198.

Schranz, M.E., Mohammadin, S., and Edger, P.P. (2012). Ancient whole genome duplications, novelty and diversification: the WGD Radiation Lag-Time Model. Curr Opin Plant Biol **15,** 147-153.

Schwarz-Sommer, Z., Huijser, P., Nacken, W., Saedler, H., and Sommer, H. (1990). Genetic Control of Flower Development by Homeotic Genes in Antirrhinum majus. Science **250,** 931-936.

Schwarz-Sommer, Z., Hue, I., Huijser, P., Flor, P.J., Hansen, R., Tetens, F., Lonnig, W.E., Saedler, H., and Sommer, H. (1992). Characterization of the Antirrhinum floral homeotic MADS-box gene deficiens: evidence for DNA binding and autoregulation of its persistent expression throughout flower development. Embo J **11,** 251-263.

Senthil-Kumar, M., and Mysore, K.S. (2011). New dimensions for VIGS in plant functional genomics. Trends Plant Sci **16,** 656-665.

Seymour, D.K., Koenig, D., Hagmann, J., Becker, C., and Weigel, D. (2014). Evolution of DNA Methylation Patterns in the Brassicaceae is Driven by Differences in Genome Organization. PLoS Genet **10,** e1004785.

Sharma, B., and Kramer, E. (2013). Sub- and neo-functionalization of APETALA3 paralogs have contributed to the evolution of novel floral organ identity in Aquilegia (columbine, Ranunculaceae). New Phytol **197,** 949-957.

Sharma, B., Guo, C., Kong, H., and Kramer, E.M. (2011). Petal-specific subfunctionalization of an APETALA3 paralog in the Ranunculales and its implications for petal evolution. New Phytol **191,** 870-883.

Sharma, B., Yant, L., Hodges, S.A., and Kramer, E.M. (2014). Understanding the development and evolution of novel floral form in Aquilegia. Curr Opin Plant Biol **17,** 22-27.

Sicard, A., and Lenhard, M. (2011). The selfing syndrome: a model for studying the genetic and evolutionary basis of morphological adaptation in plants. Annals of Botany **107,** 1433-1443.

Sicard, A., Stacey, N., Hermann, K., Dessoly, J., Neuffer, B., Baurle, I., and Lenhard, M. (2011). Genetics, Evolution, and Adaptive Significance of the Selfing Syndrome in the Genus Capsella. Plant Cell.

Simillion, C., Vandepoele, K., Van Montagu, M.C.E., Zabeau, M., and Van de Peer, Y. (2002). The hidden duplication past of Arabidopsis thaliana. Proceedings of the National Academy of Sciences **99,** 13627-13632.

Singh, V.K., Garg, R., and Jain, M. (2013). A global view of transcriptome dynamics during flower development in chickpea by deep sequencing. Plant Biotechnology Journal **11,** 691-701.

Slattery, M., Zhou, T., Yang, L., Dantas Machado, A.C., Gordân, R., and Rohs, R. (2014). Absence of a simple code: how transcription factors read the genome. Trends in Biochemical Sciences **39,** 381-399.

Slattery, M., Riley, T., Liu, P., Abe, N., Gomez-Alcala, P., Dror, I., Zhou, T., Rohs, R., Honig, B., Bussemaker, H.J., and Mann, R.S. (2011). Cofactor Binding Evokes Latent Differences in DNA Binding Specificity between Hox Proteins. Cell **147,** 1270-1282.

Slotte, T., Hazzouri, K.M., Agren, J.A., Koenig, D., Maumus, F., Guo, Y.L., Steige, K., Platts, A.E., Escobar, J.S., Newman, L.K., Wang, W., Mandakova, T., Vello, E., Smith, L.M., Henz, S.R., Steffen, J., Takuno, S., Brandvain, Y., Coop, G., Andolfatto, P., Hu, T.T., Blanchette, M., Clark, R.M., Quesneville, H., Nordborg, M., Gaut, B.S., Lysak, M.A., Jenkins, J., Grimwood, J., Chapman, J., Prochnik, S., Shu, S., Rokhsar, D., Schmutz, J., Weigel, D., and Wright, S.I. (2013). The Capsella rubella genome and the genomic consequences of rapid mating system evolution. Nat Genet **45,** 831-835.

Smaczniak, C., Immink, R.G.H., Angenent, G.C., and Kaufmann, K. (2012a). Developmental and evolutionary diversity of plant MADS-domain factors: insights from recent studies. Development **139,** 3081-3098.

Smaczniak, C., Muiño, J.M., Chen, D., Angenent, G.C., and Kaufmann, K. (2017). Differences in DNA-binding specificity of floral homeotic protein complexes predict organ-specific target genes. The Plant Cell.

Smaczniak, C., Immink, R.G., Muino, J.M., Blanvillain, R., Busscher, M., Busscher-Lange, J., Dinh, Q.D., Liu, S., Westphal, A.H., Boeren, S., Parcy, F., Xu, L., Carles, C.C., Angenent, G.C., and Kaufmann, K. (2012b). Characterization of MADS-domain transcription factor complexes in Arabidopsis flower development. Proc Natl Acad Sci U S A **109,** 1560-1565.

Smyth, D.R., Bowman, J.L., and Meyerowitz, E.M. (1990). Early flower development in Arabidopsis. Plant Cell **2,** 755-767.

Soltis, D.E., Visger, C.J., and Soltis, P.S. (2014). The polyploidy revolution then...and now: Stebbins revisited. American journal of botany **101,** 1057-1078.

Soltis, D.E., Chanderbali, A.S., Kim, S., Buzgo, M., and Soltis, P.S. (2007). The ABC model and its applicability to basal angiosperms. Ann Bot **100,** 155-163.

Soltis, D.E., Albert, V.A., Leebens-Mack, J., Bell, C.D., Paterson, A.H., Zheng, C., Sankoff, D., Depamphilis, C.W., Wall, P.K., and Soltis, P.S. (2009). Polyploidy and angiosperm diversification. American journal of botany **96,** 336-348.

Soltis, P.S., and Soltis, D.E. (2004). The origin and diversification of angiosperms. American journal of botany **91,** 1614-1626.

Soltis, P.S., and Soltis, D.E. (2016). Ancient WGD events as drivers of key innovations in angiosperms. Curr Opin Plant Biol **30,** 159-165.

Soltis, P.S., Soltis, D.E., and Chase, M.W. (1999). Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. Nature **402,** 402-404.

Sommer, H., Beltran, J.P., Huijser, P., Pape, H., Lonnig, W.E., Saedler, H., and Schwarz-Sommer, Z. (1990). Deficiens, a homeotic gene involved in the control of flower morphogenesis in Antirrhinum majus: the protein shows homology to transcription factors. Embo J **9,** 605-613.

Song, L., and Crawford, G.E. (2010). DNase-seq: A High-Resolution Technique for Mapping Active Gene Regulatory Elements across the Genome from Mammalian Cells. Cold Spring Harbor Protocols **2010,** pdb.prot5384.

Soza, V.L., Snelson, C.D., Hewett Hazelton, K.D., and Di Stilio, V.S. (2016). Partial redundancy and functional specialization of E-class SEPALLATA genes in an early-diverging eudicot. Developmental biology **419,** 143-155.

Stefflova, K., Thybert, D., Wilson, M.D., Streeter, I., Aleksic, J., Karagianni, P., Brazma, A., Adams, D.J., Talianidis, I., Marioni, J.C., Flicek, P., and Odom, D.T. (2013). Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. Cell **154,** 530-540.

Stellari, G.M., Jaramillo, M.A., and Kramer, E.M. (2004). Evolution of the APETALA3 and PISTILLATA lineages of MADS-box-containing genes in the basal angiosperms. Mol Biol Evol **21,** 506-519.

Stern, D.L., and Orgogozo, V. (2008). The loci of evolution: how predictable is genetic evolution? Evolution; international journal of organic evolution **62,** 2155-2177.

Stern, D.L., and Orgogozo, V. (2009). Is genetic evolution predictable? Science **323,** 746-751.

Stone, J.R., and Wray, G.A. (2001). Rapid evolution of cis-regulatory sequences via local point mutations. Molecular biology and evolution **18,** 1764-1770.

Studer, A., Zhao, Q., Ross-Ibarra, J., and Doebley, J. (2011). Identification of a functional transposon insertion in the maize domestication gene tb1. Nat Genet **43,** 1160-1163.

Sullivan, A.M., Bubb, K.L., Sandstrom, R., Stamatoyannopoulos, J.A., and Queitsch, C. (2015). DNase I hypersensitivity mapping, genomic footprinting, and transcription factor networks in plants. Current Plant Biology **3–4,** 40-47.

Sullivan, Alessandra M., Arsovski, Andrej A., Lempe, J., Bubb, Kerry L., Weirauch, Matthew T., Sabo, Peter J., Sandstrom, R., Thurman, Robert E., Neph, S., Reynolds, Alex P., Stergachis, Andrew B., Vernot, B., Johnson, Audra K., Haugen, E., Sullivan, Shawn T., Thompson, A., Neri Iii, Fidencio V., Weaver, M., Diegel, M., Mnaimneh, S., Yang, A., Hughes, Timothy R., Nemhauser, Jennifer L., Queitsch, C., and Stamatoyannopoulos, John A. (2014). Mapping and Dynamics of Regulatory DNA and Transcription Factor Networks in A. thaliana. Cell Reports **8,** 2015-2030.

Sun, B., Xu, Y., Ng, K.-H., and Ito, T. (2009). A timing mechanism for stem cell maintenance and differentiation in the Arabidopsis floral meristem. Genes Dev **23,** 1791-1804.

Sun, B., Looi, L.-S., Guo, S., He, Z., Gan, E.-S., Huang, J., Xu, Y., Wee, W.-Y., and Ito, T. (2014). Timing Mechanism Dependent on Cell Division Is Invoked by Polycomb Eviction in Plant Stem Cells. Science **343**.

Sundstrom, J., and Engstrom, P. (2002). Conifer reproductive development involves B-type MADS-box genes with distinct and different activities in male organ primordia. Plant J **31,** 161-169.

Tan, S., and Richmond, T.J. (1998). Crystal structure of the yeast MAT[alpha]2/MCM1/DNA ternary complex. Nature **391,** 660-666.

Tanay, A. (2006). Extensive low-affinity transcriptional interactions in the yeast genome. Genome Res **16,** 962-972.

Tang, H., Bowers, J.E., Wang, X., Ming, R., Alam, M., and Paterson, A.H. (2008). Synteny and collinearity in plant genomes. Science **320**.

Tank, D.C., Eastman, J.M., Pennell, M.W., Soltis, P.S., Soltis, D.E., Hinchliff, C.E., Brown, J.W., Sessa, E.B., and Harmon, L.J. (2015). Nested radiations and the pulse of angiosperm diversification: increased diversification rates often follow whole genome duplications. New Phytol **207,** 454-467.

Tekaia, F. (2016). Inferring Orthologs: Open Questions and Perspectives. Genomics Insights **9,** 17-28.

Theißen, G., and Becker, A. (2004). Gymnosperm Orthologues of Class B Floral Homeotic Genes and Their Impact on Understanding Flower Origin. Critical Reviews in Plant Sciences **23,** 129-148.

Theissen, G. (2001). Development of floral organ identity: stories from the MADS house. Curr Opin Plant Biol **4,** 75-85.

Theissen, G., and Saedler, H. (2001). Plant biology. Floral quartets. Nature **409,** 469-471.

Theissen, G., and Melzer, R. (2007). Molecular Mechanisms Underlying Origin and Diversification of the Angiosperm Flower. Annals of Botany **100,** 603-619.

Theissen, G., Becker, A., Di Rosa, A., Kanno, A., Kim, J.T., Münster, T., Winter, K.-U., and Saedler, H. (2000). A short history of MADS-box genes in plants. Plant Mol Biol **42,** 115-149.

Tilly, J.J., Allen, D.W., and Jack, T. (1998). The CArG boxes in the promoter of the Arabidopsis floral organ identity gene APETALA3 mediate diverse regulatory effects. Development **125,** 1647-1657.

Timms, L., Jimenez, R., Chase, M., Lavelle, D., McHale, L., Kozik, A., Lai, Z., Heesacker, A., Knapp, S., Rieseberg, L., Michelmore, R., and Kesseli, R. (2006). Analyses of Synteny Between *Arabidopsis thaliana* and Species in the Asteraceae Reveal a Complex Network of Small Syntenic Segments and Major Chromosomal Rearrangements. Genetics **173,** 2227-2235.

Torti, S., Fornara, F., Vincent, C., Andres, F., Nordstrom, K., Gobel, U., Knoll, D., Schoof, H., and Coupland, G. (2012). Analysis of the Arabidopsis Shoot Meristem Transcriptome during Floral Transition Identifies Distinct Regulatory Patterns and a Leucine-Rich Repeat Protein That Promotes Flowering. Plant Cell **24,** 444-462.

Tröbner, W., Ramirez, L., Motte, P., Hue, I., Huijser, P., Lonnig, W.E., Saedler, H., Sommer, H., and Schwarz-Sommer, Z. (1992). GLOBOSA: a homeotic gene which interacts with DEFICIENSin the control of Antirrhinum floral organogenesis. Embo J **11**.

Tuch, B.B., Galgoczy, D.J., Hernday, A.D., Li, H., and Johnson, A.D. (2008). The evolution of combinatorial gene regulation in fungi. PLoS Biol **6,** e38.

Tucker, S.C., and Hodges, S.A. (2005). Floral Ontogeny of Aquilegia, Semiaquilegia, and Enemion (Ranunculaceae). International Journal of Plant Sciences **166,** 557-574.

Tzeng, T.-Y., and Yang, C.-H. (2001). A MADS Box Gene from Lily (Lilium Longiflorum) is Sufficient to Generate Dominant Negative Mutation by Interacting with PISTILLATA (PI) in Arabidopsis thaliana. Plant and Cell Physiology **42,** 1156-1168.

Tzeng, T.Y., Liu, H.C., and Yang, C.H. (2004). The C-terminal sequence of LMADS1 is essential for the formation of homodimers for B function proteins. J Biol Chem **279,** 10747-10755.

Uchida, N., Townsley, B., Chung, K.H., and Sinha, N. (2007). Regulation of SHOOT MERISTEMLESS genes via an upstream-conserved noncoding sequence coordinates leaf development. Proc Natl Acad Sci U S A **104,** 15953-15958.

Urbanus, S.L., de Folter, S., Shchennikova, A.V., Kaufmann, K., Immink, R.G., and Angenent, G.C. (2009). In planta localisation patterns of MADS domain proteins during floral development in Arabidopsis thaliana. BMC Plant Biol **9,** 5.

Van't Hof, A.E., Campagne, P., Rigden, D.J., Yung, C.J., Lingley, J., Quail, M.A., Hall, N., Darby, A.C., and Saccheri, I.J. (2016). The industrial melanism mutation in British peppered moths is a transposable element. Nature **534,** 102-105.

Van de Velde, J., Heyndrickx, K.S., and Vandepoele, K. (2014). Inference of transcriptional networks in Arabidopsis through conserved noncoding sequence analysis. The Plant Cell **26,** 2729-2745.

van den Bergh, E., Külahoglu, C., Bräutigam, A., Hibberd, J.M., Weber, A.P.M., Zhu, X.-G., and Eric Schranz, M. (2014). Gene and genome duplications and the origin of C4 photosynthesis: Birth of a trait in the Cleomaceae. Current Plant Biology **1,** 2-9.

van der Krol, A.R., Brunelle, A., Tsuchimoto, S., and Chua, N.H. (1993). Functional analysis of petunia floral homeotic MADS box gene pMADS1. Genes Dev **7,** 1214-1228.

van Tunen, A.J., Eikelboom, W., and Angenent, G.C. (1993). FLORAL ORGANOGENESIS IN TULIPA. Flowering Newsletter**,** 33-38.

Vandenbussche, M., Theissen, G., Van de Peer, Y., and Gerats, T. (2003). Structural diversification and neo-functionalization during floral MADS-box gene evolution by C-terminal frameshift mutations. Nucleic Acids Research **31,** 4401-4409.

Vandenbussche, M., Zethof, J., Royaert, S., Weterings, K., and Gerats, T. (2004). The duplicated B-class heterodimer model: whorl-specific effects and complex genetic interactions in Petunia hybrida flower development. Plant Cell **16,** 741-754.

Vergara-Silva, F. (2003). Plants and the Conceptual Articulation of Evolutionary Developmental Biology. Biology and Philosophy **18,** 249-284.

Viaene, T., Vekemans, D., Irish, V.F., Geeraerts, A., Huysmans, S., Janssens, S., Smets, E., and Geuten, K. (2009). Pistillata—Duplications as a Mode for Floral Diversification in (Basal) Asterids. Molecular Biology and Evolution **26,** 2627-2645.

Villar, D., Flicek, P., and Odom, D.T. (2014). Evolution of transcription factor binding in metazoans-mechanisms and functional implications. Nature reviews. Genetics **15,** 221-233.

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

Wang, H., You, C., Chang, F., Wang, Y., Wang, L., Qi, J., and Ma, H. (2014). Alternative splicing during Arabidopsis flower development results in constitutive and stage-regulated isoforms. Frontiers in Genetics **5**.

Wang, J., Lu, J., Gu, G., and Liu, Y. (2011). In vitro DNA-binding profile of transcription factors: methods and new insights. The Journal of endocrinology **210,** 15-27.

Wang, T.L., Uauy, C., Robson, F., and Till, B. (2012). TILLING in extremis. Plant Biotechnol J **10,** 761-772.

Wang, Z., Gerstein, M., and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. Nature reviews. Genetics **10,** 57-63.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi **449,** 54-61.

Weigel, D., and Meyerowitz, E.M. (1993). Activation of floral homeotic genes in Arabidopsis. Science **261,** 1723-1726.

Weigel, D., and Meyerowitz, E.M. (1994). The ABCs of floral homeotic genes. Cell **78,** 203-209.

Weirauch, M.T., and Hughes, T.R. (2010). Conserved expression without conserved regulatory sequence: the more things change, the more they stay the same. Trends in Genetics **26,** 66-74.

Wellmer, F., Alves-Ferreira, M., Dubois, A., Riechmann, J.L., and Meyerowitz, E.M. (2006). Genome-Wide Analysis of Gene Expression during Early Arabidopsis Flower Development. PLoS Genet **2,** e117.

Whipple, C.J., Ciceri, P., Padilla, C.M., Ambrose, B.A., Bandong, S.L., and Schmidt, R.J. (2004). Conservation of B-class floral homeotic gene function between maize and Arabidopsis. Development **131,** 6083-6091.

White, M.A., Myers, C.A., Corbo, J.C., and Cohen, B.A. (2013). Massively parallel in vivo enhancer assay reveals that highly local features determine the cis-regulatory function of ChIP-seq peaks. Proceedings of the National Academy of Sciences **110,** 11952-11957.

Winter, K.U., Becker, A., Munster, T., Kim, J.T., Saedler, H., and Theissen, G. (1999). MADS-box genes reveal that gnetophytes are more closely related to conifers than to flowering plants. Proc Natl Acad Sci U S A **96,** 7342-7347.

Winter, K.U., Weiser, C., Kaufmann, K., Bohne, A., Kirchner, C., Kanno, A., Saedler, H., and Theissen, G. (2002). Evolution of class B floral homeotic proteins: obligate heterodimerization originated from homodimerization. Mol Biol Evol **19,** 587-596.

Wittkopp, P.J., and Kalay, G. (2012). Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. Nature reviews. Genetics **13,** 59-69.

Wolfe, K.H. (2001). Yesterday's polyploids and the mystery of diploidization. Nature reviews. Genetics **2,** 333-341.

Wright, S.I., Le, Q.H., Schoen, D.J., and Bureau, T.E. (2001). Population dynamics of an Ac-like transposable element in self-and cross-pollinating Arabidopsis. Genetics **158,** 1279-1288.

Wu, C.A., Lowry, D.B., Cooley, A.M., Wright, K.M., Lee, Y.W., and Willis, J.H. (2008). Mimulus is an emerging model system for the integration of ecological and genomic studies. Heredity (Edinb) **100,** 220-230.

Wuest, S.E., O'Maoileidigh, D.S., Rae, L., Kwasniewska, K., Raganelli, A., Hanczaryk, K., Lohan, A.J., Loftus, B., Graciet, E., and Wellmer, F. (2012). Molecular basis for the specification of floral organs by APETALA3 and PISTILLATA. Proc Natl Acad Sci U S A **109,** 13452-13457.

Yan, H.H., Mudge, J., Kim, D.J., Shoemaker, R.C., Cook, D.R., and Young, N.D. (2004). Comparative physical mapping reveals features of microsynteny between Glycine max, Medicago truncatula, and Arabidopsis thaliana. Genome **47,** 141-155.

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

Yang, Y., Fanning, L., and Jack, T. (2003). The K domain mediates heterodimerization of the Arabidopsis floral organ identity proteins, APETALA3 and PISTILLATA. The Plant Journal **33,** 47-59.

Yanofsky, M.F., Ma, H., Bowman, J.L., Drews, G.N., Feldmann, K.A., and Meyerowitz, E.M. (1990). The protein encoded by the Arabidopsis homeotic gene agamous resembles transcription factors. Nature **346,** 35-39.

Yant, L. (2012). Genome-wide mapping of transcription factor binding reveals developmental process integration and a fresh look at evolutionary dynamics. American journal of botany **99,** 277-290.

Yellina, A.L., Orashakova, S., Lange, S., Erdmann, R., Leebens-Mack, J., and Becker, A. (2010). Floral homeotic C function genes repress specific B function genes in the carpel whorl of the basal eudicot California poppy (Eschscholzia californica). Evodevo **1,** 13.

Zahn, L.M., Leebens-Mack, J., DePamphilis, C.W., Ma, H., and Theissen, G. (2005). To B or Not to B a flower: the role of DEFICIENS and GLOBOSA orthologs in the evolution of the angiosperms. J Hered **96,** 225-240.

Zahn, L.M., Leebens-Mack, J.H., Arrington, J.M., Hu, Y., Landherr, L.L., dePamphilis, C.W., Becker, A., Theissen, G., and Ma, H. (2006). Conservation and divergence in the AGAMOUS subfamily of MADS-box genes: evidence of independent sub- and neofunctionalization events. Evol Dev **8,** 30-45.

Zahn, L.M., Ma, X., Altman, N.S., Zhang, Q., Wall, P.K., Tian, D., Gibas, C.J., Gharaibeh, R., Leebens-Mack, J.H., Depamphilis, C.W., and Ma, H. (2010). Comparative transcriptomics among floral organs of the basal eudicot Eschscholzia californica as reference for floral evolutionary developmental studies. Genome Biol **11,** R101.

Zhang, J.-S., Li, Z., Zhao, J., Zhang, S., Quan, H., Zhao, M., and He, C. (2014). Deciphering the Physalis floridana Double-Layered-Lantern1 Mutant Provides Insights into Functional Divergence of the GLOBOSA Duplicates within the Solanaceae. Plant Physiol **164,** 748-764.

Zhang, J. (2003). Evolution by gene duplication: an update. Trends in Ecology & Evolution **18,** 292-298.

Zhang, R., Guo, C., Zhang, W., Wang, P., Li, L., Duan, X., Du, Q., Zhao, L., Shan, H., Hodges, S.A., Kramer, E.M., Ren, Y., and Kong, H. (2013). Disruption of the petal identity gene APETALA3-3 is highly correlated with loss of petals within the buttercup family (Ranunculaceae). Proc Natl Acad Sci U S A **110,** 5074-5079.

Zhang, S., Zhang, J.S., Zhao, J., and He, C. (2015). Distinct subfunctionalization and neofunctionalization of the B-class MADS-box genes in Physalis floridana. Planta **241,** 387-402.

Zhang, W., Zhang, T., Wu, Y., and Jiang, J. (2012a). Genome-wide identification of regulatory DNA elements and protein-binding footprints using signatures of open chromatin in Arabidopsis. Plant Cell **24,** 2719-2731.

Zhang, W., Wu, Y., Schnable, J.C., Zeng, Z., Freeling, M., Crawford, G.E., and Jiang, J. (2012b). High-resolution mapping of open chromatin in the rice genome. Genome Res **22,** 151-162.

Zhao, T., and Schranz, M.E. (2017). Network approaches for plant phylogenomic synteny analysis. Curr Opin Plant Biol **36,** 129-134.

Zhao, T., Holmer, R., de Bruijn, S., Angenent, G.C., van den Burg, H.A., and Schranz, M.E. (2017). Phylogenomic Synteny Network Analysis of MADS-Box Transcription Factor Genes Reveals Lineage-Specific Transpositions, Ancient Tandem Duplications, and Deep Positional Conservation. The Plant Cell **29,** 1278-1292.

Zhong, S., Lin, Z., Fray, R.G., and Grierson, D. (2008). Improved plant transformation vectors for fluorescent protein tagging. Transgenic research **17,** 985-989.

Zou, C., Lehti-Shiu, M.D., Thibaud-Nissen, F., Prakash, T., Buell, C.R., and Shiu, S.H. (2009). Evolutionary and expression signatures of pseudogenes in Arabidopsis and rice. Plant Physiol **151,** 3-15.

SUMMARY

Angiosperms are one of the most successful plant lineages, and show an incredible amount of diversity, for instance in flower morphology. All these morphologically different flowers are specified by a single mechanism. Floral organ specification is described by the conserved (A)BCE-model. This model describes how different combinations of the (A-), B-, C- and E-class functions specify different floral organs. Both the mechanism, as well as the transcription factors (TFs) fulfilling these different roles are conserved throughout the angiosperms. Most of these conserved transcription factors are members of the MADS-domain family, which is one of the larger families of plant transcription factors, and members of this family bind to DNA as dimers. How the transcription factors that play different roles in the (A)BCE-model interact molecularly to specify the different floral organs was subsequently described in the floral quartet model, which states that the floral MADS-domain transcription factors form different tetramers that specify the different floral organs.

Although the (A)BCE-model was shown to be conserved throughout the flowering plants, how this model can generate the large amount of morphological diversity seen in flowers has not been elucidated yet. Possibly, downstream target genes that are controlled by these master regulators are crucial for establishing shapes and sizes of the flower and its organs. In this thesis, we examined how major floral homeotic transcription factors as well as their binding sites in the genome evolved, and how this could lead to altered gene regulatory networks.
Often gene regulatory networks are modified by changes in *cis*-regulatory elements, which affect the expression of the associated gene. Although examples of changes in *cis*-regulatory elements have been incidentally shown, new genomic techniques allow to assess the evolution of *cis*-regulatory elements on a genome-wide scale. In **Chapter 2**, we explain how comparative regulatory genomics can be used to reveal changes in regulatory networks that are linked to morphological evolution. We also hypothesize that these changes have occurred downstream of the master regulators of the (A)BCE-model. These floral TFs have thousands of binding sites (BS) in the genome, and regulate the expression of hundreds of genes.

We followed the comparative approach as proposed in chapter 2, and analyzed the evolution of binding sites of the floral regulator SEPALLATA3 (E-class protein), between the closely related species *Arabidopsis thaliana* and *Arabidopsis lyrata* (**Chapter 3**). These species have very similar floral organ morphology, although there are differences in floral organ size. Surprisingly, we found relatively little overlap in SEP3 binding site profiles between these species. We show correlation between conserved binding sites and DNA sequence conservation, although sequence divergence cannot explain all binding site divergence. We also found that the position of the transcription factor binding sites (TFBS) relative to its potential target gene plays a role, as TFBS in the first 1 kb of the promoter are more likely to be conserved if their relative location is conserved. Our analysis revealed that transposition can also play a role in TFBS evolution, as we found an abundant transposon in *A. lyrata* harbouring a SEP3 binding site. Although it is intriguing that properties of TFs could change substantially between closely related species with similar flower morphologies, we observed that binding sites linked to genes involved in floral development showed higher conservation.

In **Chapter 4** we continue our comparative analyses of SEP3 binding and analyze binding site evolution within species. We compared SEP3 binding between *A. thaliana* ecotypes. Our results show that there is a large overlap between the binding sites in these two ecotypes. This overlap was substantially larger than the binding site conservation found between *A. thaliana* and *A. lyrata*. In addition, we analyzed whether chromatin accessibility may play a role in TFBS evolution, by generating profiles of open chromatin regions. Correlations between SEP3 BS and open chromatin regions confirm that the accessibility of chromatin may influence TFBS selection.

Another way to alter gene regulatory networks is to change the TFs themselves. However, a modification in a TF, which modifies the DNA binding behavior is likely to have deleterious effects on plant fitness, as this likely dramatically alters the transcriptional regulation of downstream target genes. Gene duplications can circumvent these negative effects. Directly after a duplication event both paralogs will be redundant, providing the freedom for one of the paralogs to change function. This way, gene duplication can lead to sub- or neofunctionalization of the paralogs. During plant evolution, genome duplication events are a common phenomenon. These duplications are thought to have contributed to several major changes in plant form, as they coincide with innovations such as seeds and flowers. One of the TF families that retained genes after genome duplications is the family of MADS-domain TF. The TFs necessary for floral organ specification have been duplicated several times, and in many species throughout plant history. These include the B-class genes, necessary for petal and stamen specification. There are two B-class genes: *APETALA3* (*AP3*) and *PISTILLATA* (*PI*), and both have paralogous genes in many species. We analyzed cases of B-class duplication in two different species, and examined how the paralogous genes diverged from each other (**Chapter 5 and 6**).

We first studied the paralogous *PISTILLATA* genes of *Tarenaya hassleriana* (**Chapter 5**). *T. hassleriana* belongs to the Cleomaceae family, which is a sister family of the Brassicaceae. The genomic location of *PI* is very conserved among the angiosperms. Interestingly, the *PI* duplication at the base of the Brassicales led to one of the *PI* paralogs being in a new genomic location. Whereas the Cleomaceae retained both the "old" and the "new" copy, the Brassicaceae only retained the *PI* in the new genomic location. This may mean that the Cleomaceae are a kind of intermediate between species with *PI* at the conserved genomic location, and the Brassicaceae with the moved *PI*. We examined these two *PI* paralogs in *T. hassleriana*. Sequence analysis shows that most of the sequence divergence between the two paralogs seems to have emerged in a common ancestor of the Cleomaceae and the Brassicaceae. We found that the genes have similar expression patterns, but diverged in their functionality. That these paralogs may differ in function was shown by heterologous experiments, in which only one of these genes was able to make homeotic changes in the first whorl of *A. thaliana*. In addition, we observed differences between these proteins in protein-protein interaction capabilities, as well as subtle differences in their DNA-binding specificity.

The other paralogous B-class genes we studied were the *AP3* paralogs in the basal eudicot *Aquilegia* (**Chapter 6**). Instead of the usual four floral organ types, *Aquilegia* flowers have a fifth organ type, termed the staminodium, which is positioned between stamens and carpels. The staminodia are sterile, flattened organs, that are fused to form a continuous sheet around the carpels. *Aquilegia* has three paralogs of the B-class gene *APETALA3*, and it had been previously shown that these paralogs sub- and neofunctionalized to accommodate the specification of the fifth type of floral organ. *AqAP3-3* is expressed specifically in petals. Although *AqAP3-1* specifies staminodia and *AqAP3-2* is needed for proper stamen development, these paralogs do have partially overlapping expression patterns. Exactly how AqAP3-1 and AqAP3-2 can specify different organs remains unknown. We analyzed differences between these three AP3 paralogs in DNA-binding specificity as well as whether they differ in their interaction with other MADS-domain transcription factors. We found that these paralogs differ in their DNA-binding specificity, as AqAP3-1 binds to a broader range of sequences than AqAP3-2 and AqAP3-3. We also observed differences in protein-complex formation between AqAP3-1 and AqAP3-2, although these differences were subtle, and involved the affinity between proteins to form complexes rather than the specificity to form specific complexes.

For future research, it would be interesting to assess how changes in transcription factor properties affect gene regulatory networks. Another line of research would be to study targets of the major floral MADS-domain transcription factors to identify new players in floral development, as well as examine how the gene regulatory networks evolve downstream of the (A)BCE-model to generate differences in morphology.

Research into flowers: comparing apples and oranges?

Almost everyone associates flowers with certain emotions. For example, on valentine's Day they can't get in enough red roses, and we associate them with love. And as soon as the first snowdrops and crocuses are out, we feel that spring has come. All these flowers differ in terms of color, shape and size, and yet they are all designed in the same way, with petals, sepals, stamen and a pistil.

How is it possible that the blueprint for all flowers is the same and yet the flowers all look so different? In my research I am going to investigate that question. The master regulators specify where a petal comes and where a pistil is located are conserved, and are the same in each species. However, these regulators turn other genes on or off; which genes those are differ in each plant. This entire blueprint of regulators and targets is called a network, and the details of the network are different in each species. You can compare this to two companies that both have the same management yet each produce something completely different.

Changing networks

I want to know what changes in a network to form a flower so that it looks different. To be exact, how does the network that is controlled by conserved factors change? Does the function of genes change, and does the plant look different as a result? Or do the genes remain the same but carry out their task at a different time or place? Again, you can compare this to a company. The manager stays the same but he can still arrive at a different result, for example by giving employees different tasks or by employing new staff. By looking at how these kinds of networks in plants are designed, and how they have changed in different plant species, I want to find out more about the evolution of flowers.

The networks in plants (but off course also in animals) are formed by genes and their products, proteins. Each gene is a link in the network and each protein has a specific function. To change the appearance of a flower, something in this network has to change. However, we don't know yet what has to change. You may have to alter the activity of 100 genes to have a flower look different. It could also be that you can't turn a gene on or off at all but that you can make a bit more or a bit less of the protein. If we understand how the networks in flowers change, we will better understand how other networks in organisms work.

Locks and keys

We don't know how many changes are needed in a network to change a flower. What we do know is how the network can change. Given a different function, a protein can bring about changes in the network. Another option is that the protein is no longer made, or that it is

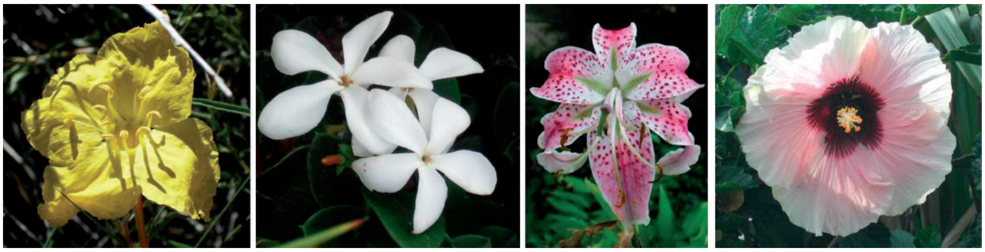


Figure 1| the blueprint for all flowers is the same and yet the flowers all look so different.

made earlier or later on, or in a different quantity. Now how can you turn genes on or off? A gene is made up of two parts. There is a part that is translated into a protein and a part that regulates where and when the gene is translated. This part is called the promoter. The promoter is therefore like an on/off switch for the gene. proteins that ensure that the gene is translated bind onto the gene's promoter (the switch). These regulators are called transcription factors.

You can compare transcription factors and promoters to keys and locks. A transcription factor is a skeleton key that fits a number of locks. A promoter can be compared to a lock. A transcription can only turn a gene on or off if it recognizes the gene; if the key fits the lock, as it were. Because there are several keys and even more locks, you can get an entire network. If one of the locks is changed, the key no longer fits. This changes the network and hence the outcome of the network: the shape of the flower. You can compare this to the hierarchy in a company; all employees have a specific job but it is the boss who tells them when and how to do their job exactly. After a reorganization, a manager may suddenly manage different people, or employees may be given different tasks.

Fishing out DNA

As said above, transcription factors and promoters can be compared to keys and locks. There are a couple of transcription factors that are specific to flowers. These have remained the same in the various plant species. However, the promoters (the locks) are different in each species. As a results, there are certain genes in a species that cannot be turned on or off with the key, while the key will fit the lock in another species. I am studying how these locks have changed. I do this by extracting DNA with the transcription factors attached from plants and cutting the DNA into pieces. I then fish out the transcription factors, the keys, from that DNA mixture. Bits of DNA stay attached to these keys, and these are the locks. What I actually do is fish out bits of DNA and then decide to which genes these bits belong. I do this not in a single plant species but in various species. By comparing the genes I fish out, I can see which locks have remained the same in the various plants, and which locks have changed. That way, I know which links in the network have changed.

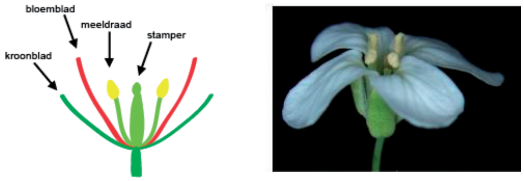


Figure 2| flowers are all designed in the same way, with sepals, petals, stamens and a pistil. The various plant organs are clearly shown in Arabidopsis (photo).

Useful?

Research into flowers. What's the point exactly? This study teaches us how the networks in plants can change. We can use this knowledge to make rice more nutritious, for instance, or to make potatoes immune to a particular fungus. But we can, of course, also use this knowledge to enlarge the number of flower varieties!

The NPC (Netherlands proteomics centre) challenged PhD students to write a popular scientific article about their research. Eight PhD students participated in the contest. During the 2012 NPC Progress Meeting the three winners were awarded the Popular Science Award. As one of the winners, this article was published in the Magazine 'NPC HighLights'.

Acknowledgments

The road to a PhD title is quite a journey, in my case rather literally. Although this journey was very personal, I could not have done it all by myself.

Kerstin, thank you for accepting me as a PhD student. I learned a lot from your supervision. Also, although I was not always happy getting my writings back full with red marks, your comments did improve the writing in this thesis!

Gerco, thank you for all your supervision and an endless amount of interesting discussions. Whenever I felt pessimistic about my work, a discussion with you helped me to get enthusiastic again. When I was not sure what to do, I could always ask you for advice, regardless of in which country I was at the moment. I am sure that without your help, this thesis would not have been finished!

I also must acknowledge Richard and Eric; although not playing any official supervising role, both of you were always willing to help me, and I am thankful for all the feedback you gave on my research, as well as the input about protein-protein interactions (Richard) and Tarenaya and genome evolution/synteny (Eric).

PDS is a very enthusiastic group with the motto “work hard, play hard”, and the atmosphere in the group was one of the reasons why I wanted to do my PhD here! PDS is an environment where you are happy to go to work, even if experiments do not go to plan (and if you have any problems, the coffee break is the place to go!). I must thank everyone in the lab for making the time spent on my PhD fun.

Jenny, thanks for helping me with practical things when I first came to Wageningen, such as finding a place to live!

Alice and Cezary, both of you helped me a lot whenever I had technical questions, whether it was with EMSAs, Westerns, ChIP or DNAse experiments. I not only learned a lot from you, but also enjoyed doing fun stuff together, whether it was in Wageningen, your visits to Potsdam or on one of the trips we did. I was happy to be part of important events in your life (think: weddings!), and hope we will keep in touch!

The journey to a PhD title is easier if you don’t go through it alone, and I was therefore very happy to have Leonie and Hilda starting their PhD journey at the same time as I started mine. Leonie, thanks for all the talks where we figured out that maybe we weren’t doing so bad (and all the chocolate; chocolate makes everything better!), as well as for all the times I was allowed to crash on your couch! Hilda, being part of the same group we could share stories about our research and how things were going. Especially the last year, we had a lot of discussions about “what’s next”, and I was also very happy we could consult with each other about the very last practical things that had to be done to get this thesis printed! Thanks girls, for going through this together!

Sam, Vera, Anneke, Lena, Manjunath, Han, Rufang, Baojian and Mengfan, thanks for the company, the nice talks, nice cocktails ;-), and off course the great (chinese/indian) food!

With a large group, it is difficult to thank everyone personally, but I will try! thanks to Martijn en Tjitske, the best clusteruitje-organizers in the world! Thanks to Jacqueline, with whom I always chat too much. Michiel, thank you for all the help, all the chats in the lab, and for being as critical about afternoon radio programs as I am (for a long journey like a PhD, you need good music ;-)). Froukje, we still have several outstanding invitations to Grenoble! Rumyana, thanks for all the nice chats, scientific or otherwise. Patricia, for some fun chats in the lab (and paella!). Thanks Ruud, Steven, Kim, Marian, Jan and Mieke (o, sorry!), for fun times, scientific advice and lots of chats! Thanks Wilma, for not only helping me find free computers when I needed to finish my thesis, but also with figuring out numerous alternate professions ;-), and sharing the not-so-silent room with me!

Eveline and Fabia, I enjoyed the time we were in the lab together, and visiting you when we were in Brazil made our holiday extra special!

Special thanks goes to my paranymphs. Marco, throughout my PhD you were always willing to help me, whether it was with experiments, sending me samples/protocols whenever I needed something and was in a different lab, or personally, for instance with a warm meal when I was busy moving! Thanks for all this, and for helping me on the very last day of my PhD by being my paranymph!

Suraj, we didn't spent a lot of time working together in the lab, but we did do a lot of fun stuff. I admire the way you are always active and organize a lot of fun activities, and I am sure you made a lot of our foreign visitors feel welcome! We also did our fair share of travelling together, and I am pretty sure we will keep doing this. A visit to India maybe? ;-) I am happy you are my paranymph, and am sure that you will be finished very soon too!

The next leg of my PhD journey came with the move to Potsdam, where I became part of a new group. We quickly formed a team that helped and supported each other. A 15h drive from Nice to Potsdam, including a Italian dinner on the sea, was a great way to get a group closer together (I'll never forget that one)!

Thank you Christopher and Julia. we were all in the same position, which was great for discussions, whether it was about scientific problems or just PhD-life in general. Your support helped me to get this thesis finished. Especially Julia, my roommate in the "late"-office ;-)! I am sure you two will be getting to the end soon too (and I WILL share the cake with you two ;-))!

Wenhao, you were always in the lab (literally), and I enjoyed your company, as well as all the good Chinese food you made, and the times we spent at your place with Huaxia and Sunny. I am sure we'll keep in touch, and look forward to a tour through China! ;-)

Johanna, thanks for helping me with some of my experiments, and for all the fun. I still remember making glühwein in the lab. And, most importantly: thanks for keeping the lab clean and running! ;-)

Dijun, we only spent a little bit of time together in the lab. However, without all your help and analyses I could not have written chapter 4. Thank you for that!

Auch will ich den Damen im Gewächshaus der Uni Potsdam danken, weil sie immer gut für meine Planzen gesorgt haben.

Katrin, you started in haus 20 around the same time I did, and we spend more time together after the move to haus 29. It was always fun to get down to the first floor for a chat with you, to share crazy stories of what was happening, and to go to the gym together. Your support during my time in Potsdam meant a lot, and I am happy you come to my defence to celebrate with me!

My PhD involved yet more travelling when I was awarded a Fulbright scholarship to spent six months in the Cambridge, MA, USA (Thank you Fulbright Center!). Elena, thank you for welcoming me in your lab. I enjoyed my time in the group, learned a lot from our discussions, and the work in your lab resulted in a nice chapter in my thesis (chapter 6)!
Claire, Molly and Minya, thanks for welcoming me, all the help in the lab, as well as our movie nights! Also Cheng-Chiang and Rui, thanks for making my time in the lab more enjoyable!
Lynn, it was always fun chatting with you, I am grateful for you sharing some of your Y2H knowledge with me, and Diego and I very much enjoyed the tour you and your mom gave us!
Christy, thanks for all the help in the lab at the beginning, and I really enjoyed all our coffee dates. They made my stay in Cambridge even nicer, and I am sure that all that coffee and our conversations greatly contributed to getting this thesis finished!

Off course, I cannot forget Oscar and Gracie, who not only guarded our office, but also were great in cheering me up after failed experiments!

Michele, thanks for all the help with the SELEX analysis.
Marsha, I was happy I found a room in your place, and was always happy to have some company when I came home!

One country that I did not visit extensively during this PhD journey is bioinformatics-land. José, thanks for all the bioinformatics, as well as the discussions we had. You played a big part in the work described in chapter 3, but also helped me with several of the other chapters. Thank you!
Aalt-Jan, thank you for stepping in with some last-minute SELEX analysis work!

I also need to thank some people who were not directly related to this thesis, but still made this whole period a lot more fun!
Thanks Ute, Tessa, Pascal and Alex! All the fun times with you definitely got me through this PhD! Also thanks for listening to me complain about it every now and again (and again, and again) ;-)!
Thanks to Krysztof and Magda, for all the game nights! We should keep on doing those every now and again!
Irene en Jupp, jullie hebben geen actieve rol gehad in dit proefschrift, maar zijn wel belangrijk voor me. En bij jullie op bezoek komen, en met kaffee mit kuchen kletsen over de verschillen tussen ossies en wessies was altijd leuk!

Volevo anche ringraziare la famiglia di Diego, in particolare i suoi genitori. Nonostante io parli poco l'italiano, il che rende la comunicazione difficile, mi avete accolto con affetto nella vostra famiglia. Sono molto felice che possiate essere presenti alla mia cerimonia di dottorato per condividere con me questo giorno speciale!

Opa en oma: Jullie waren niet alleen altijd geïnteresseerd in hoe mijn werk ging, maar hadden ook altijd een slaapplek voor me als ik weer eens op reis was. Dit maakte mijn werk een stuk makkelijker, en gezellig!

Mama, je hebt misschien niet letterlijk geholpen met de inhoud van dit boekje, maar je hebt me absoluut geholpen om het af te krijgen. Ik denk dat Opa gezegd zou hebben dat deze titel ook een beetje van jou is!

Last but absolutely not least, Diego! You joined my journey called “life” 6 years ago in San Diego, and since then we have done some travelling together, even living in several different countries. Thanking you for your help with this thesis, I do not even know where to start...you always believed in me, encouraged me, and gave me the time to work on my thesis. But you also literally helped with the content of this thesis, working in the same lab, and even teaching me some of the techniques I used in this thesis. It is safe to say that without you, this thesis would have looked very different! I hope that from now on we will see less of the (German) highways, and that we will spend many, many moons doing fun stuff together!

Cheers,

Suzanne

List of Publications

de Bruijn S, Angenent GC, Kaufmann K, 2012, Plant ‘evo-devo goes genomic: from candidate genes to regulatory networks, **Trends in Plant science**

Cheng S, van den Bergh E, Zeng P, Zhong X, Xu J, Liu X, Hofberger J, **de Bruijn S**, Bhide AS, Kuelahoglu C, Bian C, Chen J, Fan G, Kaufmann K, Hall JC, Becker A, Bräutigam A, Weber AP, Shi C, Zheng Z, Li W, Lv M, Tao Y, Wang J, Zou H, Quan Z, Hibberd JM, Zhang G, Zhu XG, Xu X, Schranz ME, 2013, The Tarenaya hassleriana Genome Provides Insight into Reproductive Trait and Genome Evolution of Crucifers, **Plant Cell**

Muiño JM, **de Bruijn S**, Pajoro A, Geuten K, Vingron M, Angenent GC, Kaufmann K, 2015, Evolution of DNA-binding sites of a floral master regulatory transcription factor, **Molecular Biology and Evolution**

Zhao T, Holmer R, **de Bruijn S**, Angenent GC, van den Burg HA, Schranz ME,, 2017, Phylogenomic Synteny Network Analysis of MADS-Box Transcription Factor Genes Reveals Lineage-Specific Transpositions, Ancient Tandem Duplications and Deep Positional Conservation, **Plant Cell**

Aerts N, **de Bruijn S**, van Mourik H, Angenent GC, van Dijk ADJ, comparative Analysis of Binding patterns of MADS-domain Proteins in *Arabidopsis thaliana*, **Submitted**

Curriculum Vitae

Suzanne (Suze-Annigje) de Bruijn was born on May 10th 1986 in Amsterdam. After finishing high school in 2004, she started university where she studied the Bachelor Life Science and Technology, a joint programme by the University of Leiden and the Delft University of Technology (TU Delft).

Having a preference for the life science part, she continued her education with a Master in Biomolecular Science at the VU in Amsterdam. Here, she got interested in studying plants, and she did an internship studying DNA methylation in petunia. She also did an internship at the University of California, San Diego (USA), where she used molecular genetics for phylogenetics and fine mapping, working on the species *Jatropha curcas* and Maize. In addition, she wrote a thesis (literature study) about intercellular movement of plant transcription factors. She finished her Master in 2011 *Cum laude*.

During the last phase of her studies, Suzanne successfully participated in the Master talent programme from EPS-NWO, which resulted in a PhD project under the supervision of Kerstin Kaufmann and Gerco Angenent. She worked on the evolution of floral MADS domain transcription factors and their binding sites. This work focussed on closely related *Arabidopsis* species, as well as *Tarenaya hassleriana*. The work for this thesis was performed at Wageningen University as well as at the University of Potsdam in Germany. In addition, Suzanne won a Fulbright scholarship that allowed her to spend six months in the lab of Elena Kramer, Harvard University (USA) working on MADS domain proteins in the genus *Aquilegia*. All this work is presented in this thesis.

Apart from her studies, Suzanne has been a passionate kayaker. She managed to combine her studies and her sport successfully. The international water sport venue “Bosbaan” is very close to the VU university in Amsterdam. During her time in San Diego, she made use of the great training facilities/conditions offered. In 2010 Suzanne spent a month in Australia to solely focus on her sport. Between 2003 and 2010 she has competed in several International Kayak Sprint championships.

Education Statement of the Graduate School

Experimental Plant Sciences

Issued to: **Suze-Annigje (Suzanne) de Bruijn**
Date: **20 November 2017**
Group: **Laboratory of Molecular Biology & BU Bioscience**
University: **Wageningen University & Research**

1) Start-up phase	*date*
► **First presentation of your project**	
Title: Evolutionary dynamics of MADS-box TF binding sites and its potential impact on evolution	08 Dec 2011
► **Writing or rewriting a project proposal**	
Title: Evolutionary dynamics of Chromatin accessibility, Epigenesys	Apr 2013
► **Writing a review or book chapter**	
Title: Plant 'evo-devo' goes genomic: from candidate genes to regulatory networks, Trends in Plant Science 2012, DOI: 10.1016/j.tplants.2012.05.002	13 Jun 2012
► **MSc courses**	
Laboratory use of isotopes	
Subtotal Start-up Phase	*11.5 credits**

2) Scientific Exposure	*date*
► **EPS PhD student days**	
EPS PhD student day, Amsterdam, NL	30 Nov 2012
EPS PhD student day, Leiden, NL	29 Nov 2013
EPS PHD student day (Get2gether), Soest, NL	29-30 Jan 2015
► **EPS theme symposia**	
EPS theme 4 'Genome Biology', Wageningen, NL	09 Dec 2011
EPS theme 1 'Developmental Biology of Plants', Wageningen, NL	19 Jan 2012
EPS theme 1 'Developmental Biology of Plants', Leiden, NL	17 Jan 2013
EPS theme 1 'Developmental Biology of Plants', Leiden, NL	08 Jan 2015
EPS theme 1 'Developmental Biology of Plants', Wageningen	21 Jan 2016
► **National meetings (e.g. Lunteren days) and other National Platforms**	
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	04-05 Apr 2011
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	02-03 Apr 2012
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	22-23 Apr 2013
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	14-15 Apr 2014
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	13-14 Apr 2015
Annual meeting 'Experimental Plant Sciences', Lunteren, NL	11-12 Apr 2016
► **Seminars (series), workshops and symposia**	
Workshop: nuclear organization and gene regulation, Utrecht, NL	09 Nov 2011
Seminar: 'Genomic regulatory mechanisms detected by ChIP-exo', Frank Pugh, Utrecht, NL	14 Jun 2012
Seminar: 'Strong selection on the genes controling complex traits in complex environments', Tom Mitchell-Olds, Wageningen, NL	10 Dec 2012
Seminar: 'Chromatin dynamics and cell fate in plants, from genetics to epigenetics', Christel Carles, Wageningen, NL	16 Jan 2013
Seminar: 'Origin and consequences of genetic and epigenetic variation in Arabidopsis thaliana', Detlef Weigel, Wageningen, NL	27 Feb 2013
Seminar: 'Tracing the evolutionary history of plant MADS-box genes', Koen Geuten, Potsdam, Germany	04 Apr 2013
Workshop ' Molecular Genetics of Plant Development and Stress Response, Potsdam, Germany	17 Apr 2013
Seminar: 'Epigenetic control of cell fate in plants', Daniel Schubert, Berlin, Germany	05 Jul 2013
Seminar: 'Dissecting chromatin function in Arabidopsis', Lars Hennig, Potsdam, Germany	09 Oct 2013
Seminar: '(Epi) genetic regulatory mechanisms underlying male gamete functions in Arabidopsis and Physcomitrella', Jorg D. Becker, Potsdam, Germany	11 Nov 2013
Seminar: 'The 5gbp barley genome- everything changes with access to genome sequence", Nils Stein, Potsdam, Germany	18 Nov 2013
Seminar: 'Gene regulatory networks underlying flower development', Frank Wellmer, Potsdam, Germany	09 Jan 2014
Seminar: 'Temperature sensing in plants', Phil Wigge, Potsdam, Germany	30 Apr 2014
Seminar: 'A functional and evolutionary perspective on transcription factor binding in A. thaliana', Klaas Vandepoele, Potsdam, Germany	08 Oct 2014
Seminar: 'Plant-pollinator interactions and the genetics of speciation', Cris Kuhlemeier, Potsdam, Germany	01 Jun 2015
Seminar: 'Mechanodevo- how do plants read their own shape', Olivier Hamant, Wageningen, NL	16 Mar 2016
Seminar: 'Perianth evolution in angiosperms', Sophie Nadot, Wageningen, NL	20 May 2016
Seminar: 'Ants, plants and bacteria: symbiosis as a driver of evolutionary diversity', Corrie Moreau, Cambridge, MA, USA	10 Nov 2016
Seminar: 'Floral adaptations to plant breeding systems', Michael Lenhard, Cambridge, MA, USA	20 Jan 2017

Continued on next page

Seminar: 'Why poorly known taxa are critical to understanding animal evolution', Casey Dunn, Cambridge, MA, USA	23 Feb 2017
Seminar: 'Are there any asexual eukaryotes? Evidence of sexuality and atypical meiosis in bdelloid rotifers', Matthew Meselson, Cambridge, MA, USA	20 Apr 2017
Seminar: 'How to tame a fox {and breed a dog}; a siberian tale of jump started evolution', Lee Dugatkin, Cambridge, MA, USA	21 Apr 2017
Symposium: Stomata; evolution, development and physiology, 12th annual Harvard plant biology symposium, Boston, MA, USA	08-09 May 2017
► **Seminar plus**	
► **International symposia and congresses**	
Workshop 'Molecular Mechanisms Controlling Flower Development', Maratea, Italy	14-17 Jun 2011
International Conference "Genomic Basis of Evo-Devo", Dublin, Ireland	23-26 Jun 2012
Workshop on molecular mechanisms controlling flower development, Presqu'ile de Giens, France	08-12 Jun 2013
Workshop on molecular mechanisms controlling flower development, Parador de Aiguablava, Spain	07-11 Jun 2015
► **Presentations**	
Poster: Evolutionary dynamics of plant MADS-box transcription factor binding sites, Dublin, Ireland	23-26 Jun 2012
Poster: Evolutionary dynamics of MADS-box transcription factor binding sites and potential target genes in flower development, France	08-12 Jun 2013
Poster: Evolution of B-class paralogs controlling flower development in T. hassleriana, Spain	07-11 Jun 2015
Poster: evolution of PISTILLATA paralogs in Tarenaya hassleriana,Boston, MA, USA	09 May 2017
Talk: 'Molecular Genetics of Plant Development and Stress Response' 'evolution of gene regulation by MADS-domain factors, Potsdam, Germany	17 Apr 2013
Talk: Evolutionary dynamics of transcription factor (SEP3) binding sites in flower development, Lunteren, NL	15 Apr 2014
► **IAB interview**	
Meeting with a member of the International Advisory Board of EPS	05 Jan 2015
► **Excursions**	
Company visit, Genetwister	19 Sep 2014
Subtotal Scientific Exposure	*20.7 credits**

3) In-Depth Studies	date
► **EPS courses or other PhD courses**	
5th International PhD school in Plant Development, Siena, Italy	25-28 Jun 2012
Course 'Current Trends in Phylogenetics', Wageningen, NL	22-26 Oct 2012
Course 'Perl Programmierung fur Biologen', Potsdam, Germany	Jul 2013
Course 'Introductory in Multivariate Statistics with R', Potsdam, Germany	Sep 2013
EPS course 'Transcription factors and Transcriptional regulation', Wageningen, NL	17-19 Dec 2013
Online course 'Bioinformatics Methods II', Coursera	Mar-Apr 2014
Course 'Comparative Methods in Genome Annotation', Potsdam, Germany	29-30 Apr 2014
School of physics “Integrated structural and cell biology, from molecules to cells and organisms: thinking out of the box”, les Houches, France	08 Jul-01 Aug 2014
► **Journal club**	
Member literature discussion group Wageningen, NL	Oct 2011-Febr 2013
Member literature discussion group Potsdam, Germany	Mar 2013-Oct 2016
► **Individual research training**	
Scanning Electron Microscopy (Koen Geuten, KU Leuven)	19-20 Dec 2011
Subtotal In-Depth Studies	*17.7 credits**

4) Personal development	date
► **Skill training courses**	
Techniques for writing and presenting a scientific paper, Wageningen, NL	05-08 'Feb 2013
Project and time management, Wageningen, NL	Apr-Jun 2012
Popular science writing, Netherlands proteomics centre	13&27 Jan 2013
ExPectationS, EPS Career day, Wageningen, NL	01 Feb 2013
EMBO YIP PhD course, Heidelberg, Germany	01-06 Dec 2014
Insight out, the conference for women in science, Ede	24 May 2016
► **Organisation of PhD students day, course or conference**	
► **Membership of Board, Committee or PhD council**	
Subtotal Personal Development	*5.7 credits**

TOTAL NUMBER OF CREDIT POINTS*	**55.6**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS

** A credit represents a normative study load of 28 hours of study.*

The research described in this thesis was financially supported by an NWO-EPS graduate school master talent grant, an EpiGeneSys (EU) grant (for small scale collaborative projects), and a Fulbright scholarship (sponsored by Elsevier).

Financial support from Wageningen University for printing this thesis is gratefully acknowledged.

Cover design by the author and Fenna Schaap

Printed by Proefschriftmaken.nl