Evolution of DNA-Binding Sites of a Floral Master Regulatory Transcription Factor

Jose M. Muiño,*[1,2] Suzanne de Bruijn,[3,4] Alice Pajoro,[4] Koen Geuten,[5] Martin Vingron,[1] Gerco C. Angenent,[4,6] and Kerstin Kaufmann*[3]

[1]Computational Molecular Biology, Max Planck Institute for Molecular Genetics, Berlin, Germany

[2]Laboratory of Bioinformatics, Wageningen University, Wageningen, The Netherlands

[3]Institute for Biochemistry and Biology, Potsdam University, Potsdam, Germany

[4]Laboratory of Molecular Biology, Wageningen University, Wageningen, The Netherlands

[5]Laboratory of Molecular Plant Biology, Department of Biology, University of Leuven (KU Leuven), Leuven, Belgium

[6]Bioscience, Plant Research International, Wageningen, The Netherlands

***Corresponding author:** E-mail: muino@molgen.mpg.de; kerstin.kaufmann@uni-potsdam.de.

Associate editor: Stephen Wright

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 (BSs) among plant species have not yet been experimentally determined. Here, we performed comparative ChIP (chromatin immunoprecipitation)-seq experiments of the MADS-domain transcription factor SEPALLATA3 (SEP3) in two closely related *Arabidopsis* species: *Arabidopsis thaliana* and *A. lyrata* which have very similar floral organ morphology. We found that BS 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.

Key words: **MADS-domain transcription factor, *cis*-regulatory evolution, plant development.**

Introduction

Plant development is controlled by transcription factors (TFs), which form complex gene-regulatory networks (Kaufmann, Pajoro, et al. 2010). Genome-wide TF DNA-binding studies revealed that these factors have several thousands of binding sites (BSs) in the *Arabidopsis* genome, and may regulate the expression of many genes directly, likely in combination with other TFs (for review, see Pajoro, Biewers, et al. 2014). 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 BSs across plant species was done indirectly using DNA sequence conservation studies, as 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). Although the study by Haudry et al. (2013) resulted in the recovery of the highest number of TF BSs 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 Brassicacae, conservation of these CNSs was very low, ranging from 0.8% in *Oryza 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.

Article

Recent studies in mammals and insects have characterized the conservation of TF DNA BSs across different species using ChIP (chromatin immunoprecipitation)-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. 2012; 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), whereas 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), whereas *Drosophila* genomes have a much lower content of these elements (Lynch et al. 2011). In plants, E2F BSs may have been amplified by transposon activity in Brassicaceae species (Henaff 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 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 rearrangements, 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 Ma (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*.

Arabidopsis 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 Ma (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 noncoding 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-BSs 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-BSs and potential direct target genes.

Results

Identification of SEPALLATA3 DNA-BSs 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 (fig. 1*A*). This allowed us to use a previously generated antibody against *A. thaliana* SEP3 (AthSEP3) for ChIP experiments (Kaufmann et al. 2009). The heterologous expression of an *A. lyrata SEP3* (Aly*SEP3*) promoter::gene

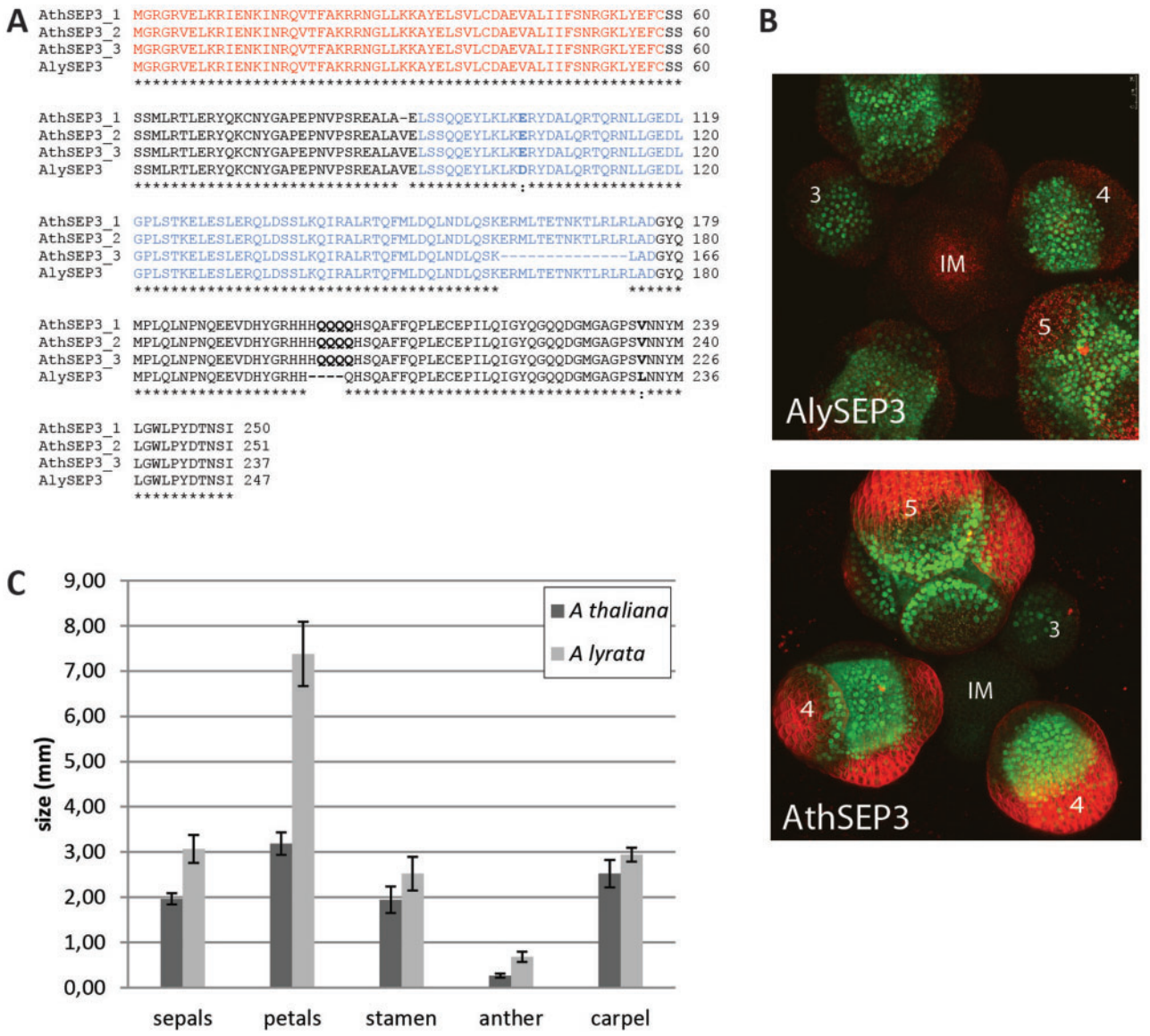

Fig. 1. SEP3 protein sequence and expression conservation. (*A*) Multiple sequence alignment of AthSEP3 splice forms and AlySEP3. Species-specific differences are indicated in bold. The MADS-domain is labelled in red; the K-domain is marked blue. (*B*) Maximum projection confocal images of inflorescence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

fragment fused to GFP was highly similar to that of the *AthSEP3 GFP* reporter gene fusion, supporting the conservation of *SEP3* gene functions in the two species (fig. 1*B*). As DNA-BSs of SEP3 may vary between tissues and developmental stages (Pajoro, Madrigal, et al. 2014), 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) (supplementary fig. S1, Supplementary Material online), 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 with *A. thaliana* (fig. 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 with *A. thaliana*.

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, Muino, et al. 2010) 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 with the other replicate (supplementary fig. S2, Supplementary Material online). 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 *Drosophila 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 3,488 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 reanalyzed using the same approach and the most up-to-date genome version.

ChIP-seq data analysis by CSAR (Muino et al. 2011) revealed a slightly larger number (1.2 ×) 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×). 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 an 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* (fig. 2*A* and *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. 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 *A. thaliana* (fig. 2*C*). 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 Materials and Methods). Comparing this new annotation with the TAIR10 and Araly1 gene annotations, some differences on 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 supplementary fig. S3, Supplementary Material online).

Evolutionary Turnover of SEP3 DNA-BSs

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

Table 1. Number of SEP3 BSs and Potential Target Genes Identified by ChIP-seq.

	Total Number of BSs		
	FDR < 0.05	FDR < 0.01	FDR < 0.005
Arabidopsis thaliana	3,233 (5,466)	2,276 (3,979)	2,043 (3,622)
Arabidopsis lyrata	4,167 (4,184)	2,784 (2,831)	2,137 (2,198)

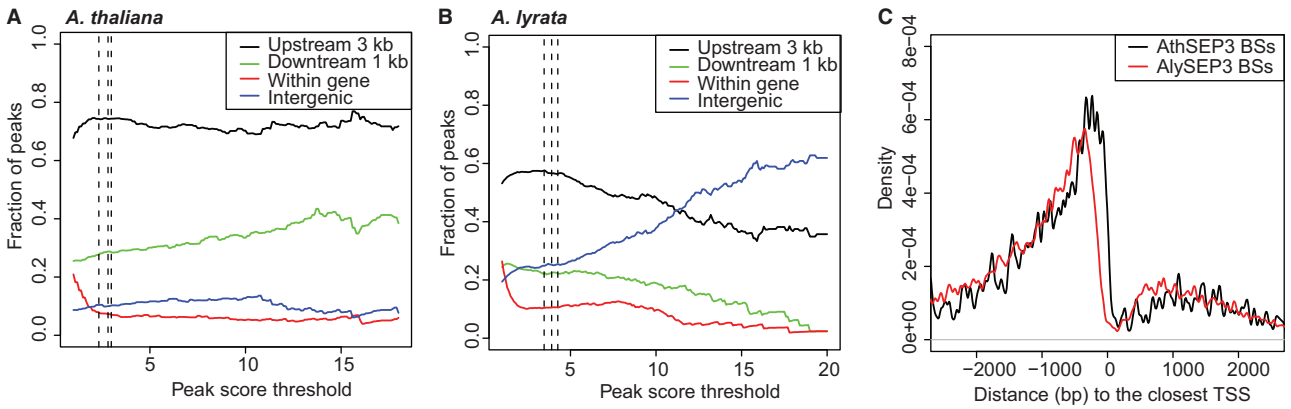

Fig. 2. SEP3 binding relative to genomic features in *Arabidopsis thaliana* and *A. lyrata*. (*A*, *B*) Enrichment of SEP3 BSs within promoters (black line, up to 3 kb upstream of gene start) and downstream regions (green, up to 1 kb downstream of end of gene) with the increase of the ChIP-seq score threshold used. BSs within genes (red line) and peaks in intergenic regions without any neighboring 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).

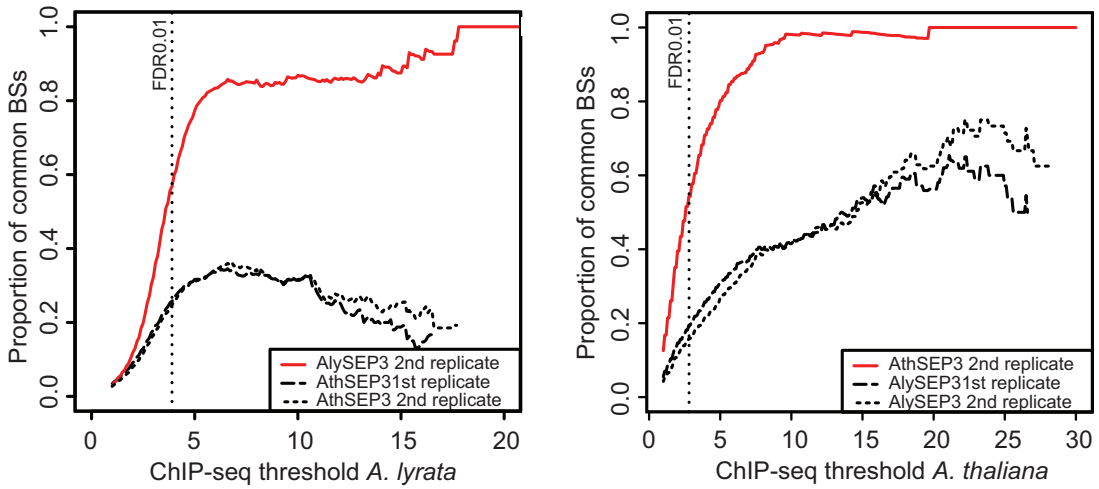

FIG. 3. Proportion of conserved SEP3 BSs between and within species. The plots show the proportion of common BSs between the best AlySEP3 replicate and the top 3,000 BSs of other ChIP-seq data sets (*A*), and the proportion of common BSs between the best AthSEP3 replicate and the top 2,000 BSs of the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

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 (fig. 3). 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 supplementary fig. S2, Supplementary Material online).

Similarly to He et al. (2011), in order to correct for the different numbers of BSs that were identified in the data sets 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 (fig. 3*A*), 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* (fig. 3*A*). In particular, at FDR < 0.01 the proportion of common BSs between species was, on average, 26%, whereas 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 (fig. 3*B*). 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," "postembryonic organ development," "stamen development" and "androecium development," and "floral organ development" ($P < 7 \times 10^{-5}$) were the top five GO categories enriched (supplementary table S1, Supplementary Material online). 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 (supplementary table S2, Supplementary Material online). The BS turnover is as low as 62% (36 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 fig. 4 for some examples).

DNA Sequence and BS Conservation

Next, we studied the relationship between DNA sequence conservation and SEP3 BS conservation. To test how well

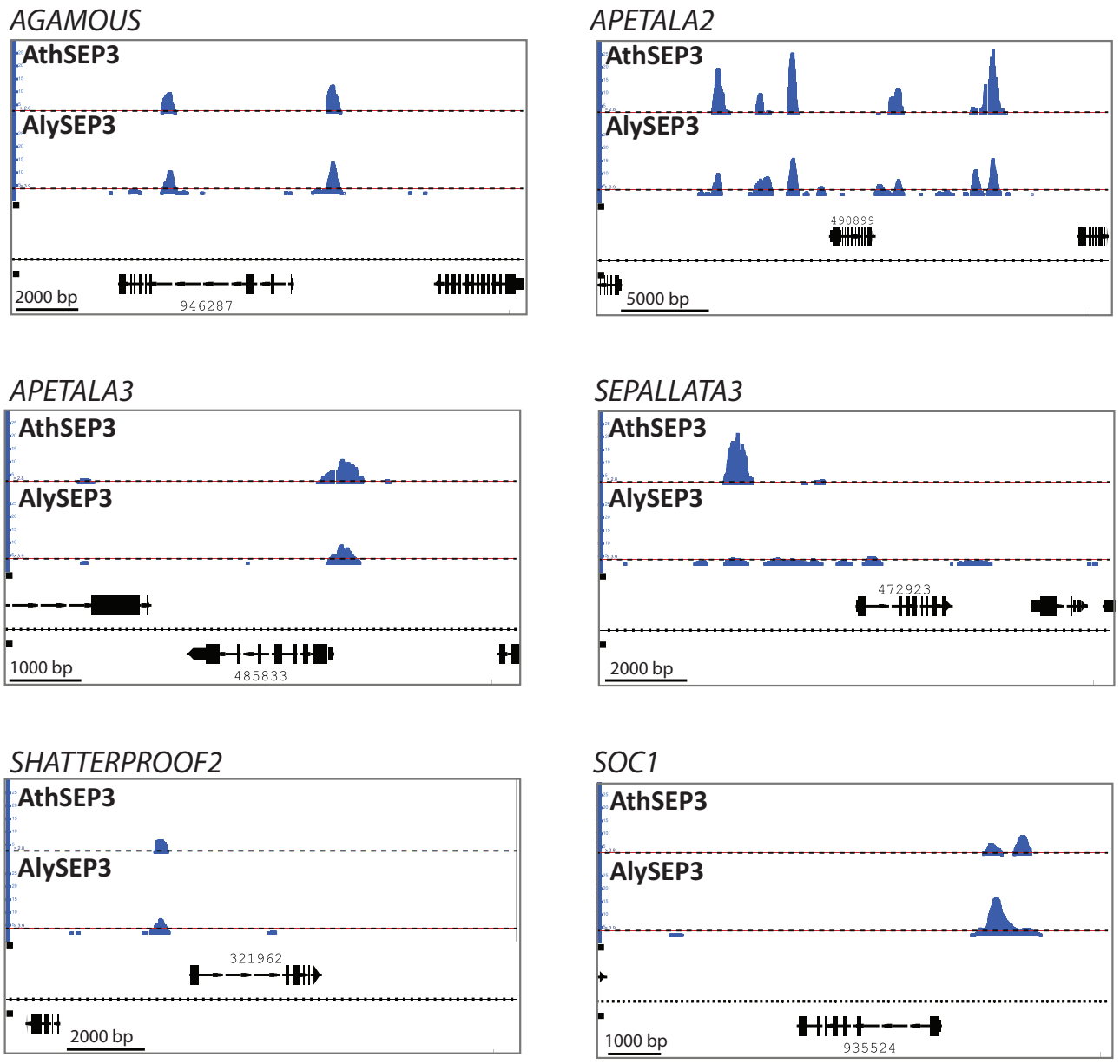

Fig. 4. Conservation of SEP3 DNA-binding and potential direct target genes between *Arabidopsis lyrata* and *A. thaliana*. SEP3 BSs in several homeotic and other key-regulatory gene loci are shown in the aligned genomes. The respective genomic locus of each TF gene in *A. lyrata* is indicated. The horizontal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

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* (fig. 5*A*). We found an enrichment (fig. 5*B*) in regions defined as 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 with the *A. lyrata*-specific BSs ($P < 0.0001$ Fisher's exact test) (fig. 5*B*). The presence of a CArG-box motif in the bound region in both species is also associated with BS conservation (fig. 5*C* and *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 (supplementary fig. S4*A* and *D*, Supplementary Material online). 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 (Muino 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 (supplementary fig. S4*C*, Supplementary Material online). 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

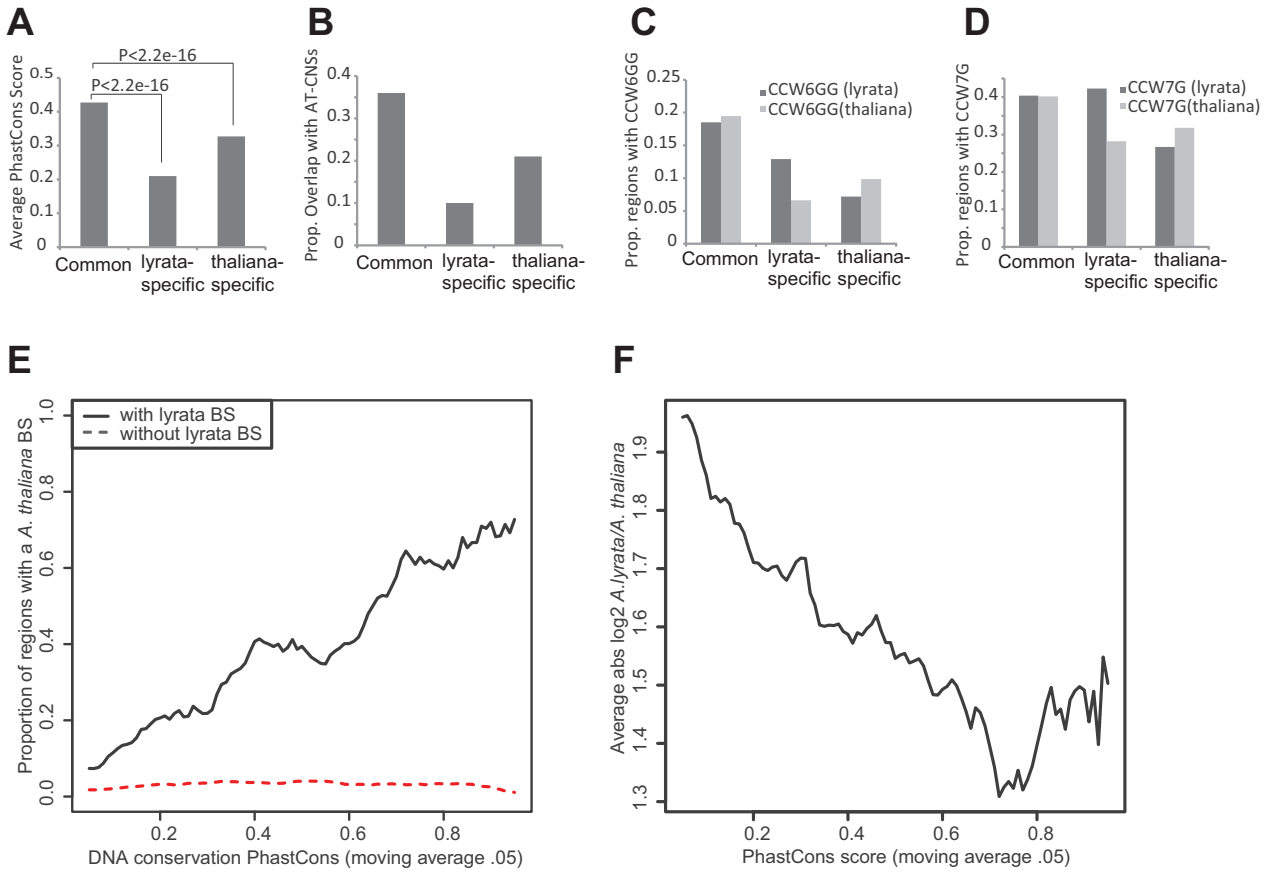

Fig. 5. SEP3 binding conservation versus DNA sequence conservation. (*A*) Average PhastCons conservation score in SEP3-bound regions that are commonly bound in *Arabidopsis thaliana* and *A. lyrata*, as well as in species-specifically bound regions. (*B*) Proportion of common and species-specific SEP3 BSs overlapping with a 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 ($CC[A/T]_6GG$) or ($CC[A/T]_7G$). (*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.

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 (supplementary fig. S4*B*, Supplementary Material online). Quantitative changes in SEP3 occupancy levels are associated with differences in PhastCons scores (fig. 5*F*; Pearson's $r = -0.21$; $P < 2.2 \times 10^{-16}$).

Genomic Position and DNA-BS Conservation. Prompted by the observation that the distribution of SEP3 BS position relative to their potential target genes was different in *A. thaliana* compared with *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 (fig. 6*A*). 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) (fig. 6*A* and supplementary fig. S5, Supplementary Material online).

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; fig. 6*B*, 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 $r = -0.92$; $P < 0.0002$). 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) (fig. 6*B*). 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.

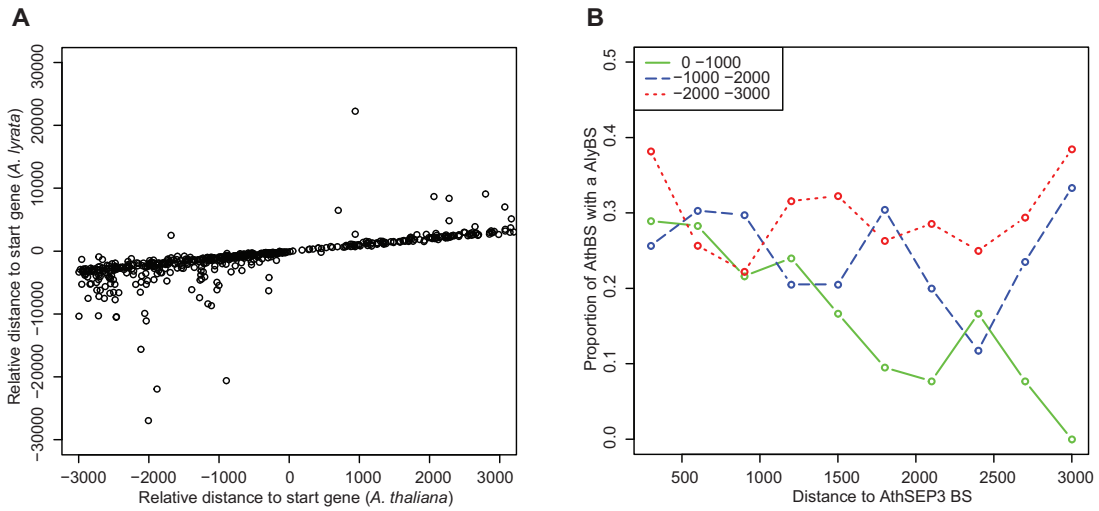

Fig. 6. SEP3 BS conservation versus position conservation. (*A*) AthSEP3 BS relative position to their target gene compared with their orthologous regions in *Arabidopsis lyrata* when the BS is conserved. We only considered the BSs that were common to both species. For different scale of the *y* axis, see supplementary figure S4, Supplementary Material online. (*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.

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 is between 2×10^{10} and 4×10^{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 BSs (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 discard 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 superfamilies of DNA/MuDR and DNA/hAT transposons ($P < 0.005$; hypergeometric test), as well as an uncharacterized repetitive element family (rnd-6_family-174, hereafter abbreviated "*6-174*") (supplementary table S3, Supplementary Material online). Because of the particularly strong enrichment (89 of 169 elements containing an 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 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 (supplementary fig. S6, Supplementary Material online). 72% of all *6-174* sequences that have a significant SEP3 BS possess at least one perfect CArG box of the consensus $CC[A/T]_6GG$, whereas only 54% of all those sequences without an SEP3 BS possess a CArG box. CArG box sequences of type $CC[A/T]_7G$ are not enriched in *6-174* sequences with SEP3 BSs. In *A. thaliana*, there are only nine *6-174* elements. None of them shows an SEP3 BS in our data, neither do they contain a perfect CArG-box motif ($CC[A/T]_6GG$ or $CC[A/T]_7G$). 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 (supplementary table S4, Supplementary Material online; $P < 0.05$).

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 an

SEP3 BS, the proportions significantly change ($P < 0.032$; Chi-square test) to 12% (more highly expressed in *A. lyrata*) and nonsignificantly change ($P < 0.45$; Chi-square test) to 17% (more highly expressed in *A. thaliana*), which indicates that TEs containing an 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 versus DNA-BS Conservation

Following 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 an SEP3 BS, we observed that the presence of an SEP3 BS in the *A. thaliana* ortholog negatively depends on the strength of the purifying selection of protein sequences of the orthologous genes (fig. 7*B*, black line). Specifically, when the *A. thaliana* paralog has an SEP3 BS, only 32% (8/25) of *A. thaliana* orthologs with $K_a/K_s < 0.1$ also have an 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 an SEP3 BS, the presence of an SEP3 BS in the *A. thaliana* ortholog is independent of purifying selection (fig. 7*B*, red line). Specifically, when the *A. thaliana* paralog does not have an SEP3 BS, 30% (31/102) of *A. thaliana* orthologs with $K_a/K_s < 0.1$ have an 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 an 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 an 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*. Data sets were generated in three biological replicates, and showed a high level of reproducibility (supplementary fig. S7, Supplementary Material online, $R \approx 0.98$). Quantitative comparison of the floral transcriptomes from the two species showed that the majority of orthologous gene pairs showed similar levels of expression. In total, 2,454 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

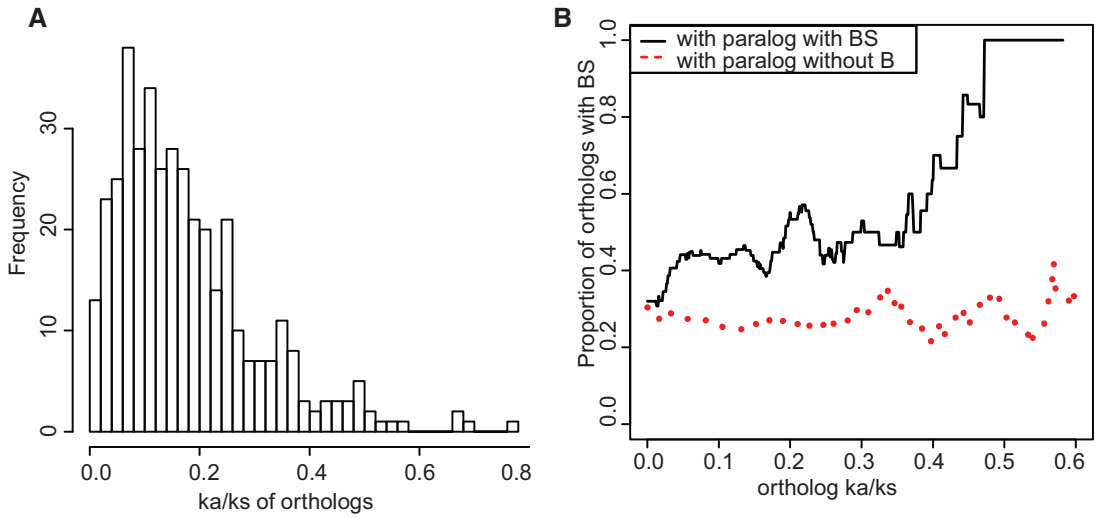

FIG. 7. SEP3 binding conservation versus divergence time and the impact of purifying selection of associated genes. (*A*) Distribution of the K_a/K_s values for orthologous genes in *Arabidopsis thaliana* and *A. lyrata*. (*B*) Proportion of orthologous genes with conserved BS depending on their K_a/K_s values when at least one *A. thaliana* paralog has conserved regulation (continuous 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.

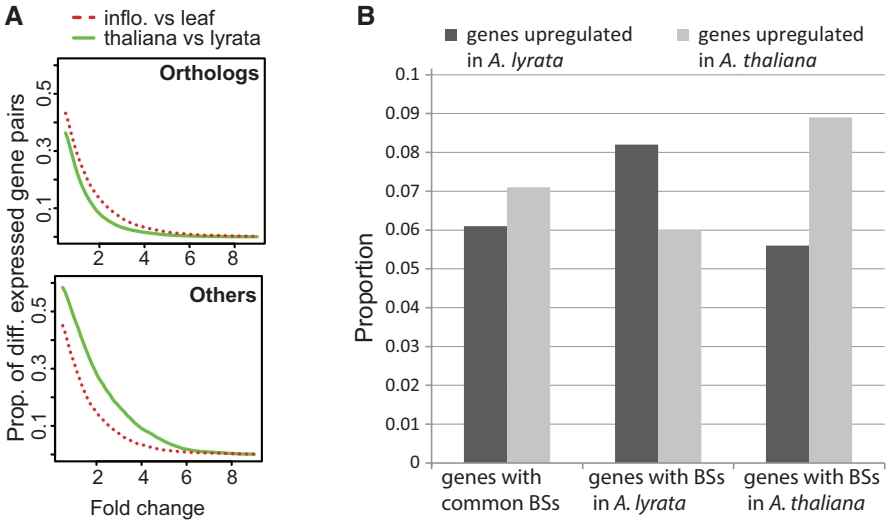

Fig. 8. Evolution of gene expression versus changes in SEP3 binding. (*A*) Overall gene expression comparison in leaves and inflorescences of *Arabidopsis lyrata*, and in inflorescences of *A. thaliana* versus *A. lyrata*. The analysis was done for orthologs genes and for paralog genes. (*B*) Proportion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

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 (fig. 8*A*). 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 BSs (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 (supplementary fig. S8, Supplementary Material online).

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 (fig. 8*B*). In contrast, orthologs with species-specific BSs have a slightly higher proportion of differentially expressed genes (fig. 8*B*). Genes with higher occupancy levels of SEP3 in *A. thaliana* tend to be more strongly expressed in this species (supplementary fig. S8, Supplementary Material online). 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 an 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 (supplementary table S5, Supplementary Material online). 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 (supplementary table S5, Supplementary Material online), 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.

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. Nonaligned 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 Ma (Hu et al. 2011), which is in a similar time range of *D. melanogaster* and its closest relatives (~2.5–30 Ma), or of human and macaque (~20 Ma). 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 (supplementary fig. S4, Supplementary Material online). However, many BSs are occupied in a species-specific manner despite presence of a conserved CArG box in the other species (fig. 5*C* and *D* and supplementary fig. S4*A*, Supplementary Material online). 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), whereas *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 (Henaff 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 is mostly due to small deletion/insertion (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 (fig. 6*B*) 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 (0–1 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, nonfunctional 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, Wellmer, et al. 2010; Wuest et al. 2012; ÓMaoiléidigh et al. 2013; Pajoro, Madrigal, et al. 2014). 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; fig. 6). BSs located in distal promoter regions (1–3 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 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 TEs (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, whereas 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 coevolution 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 coevolution 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 supplementary table S5, Supplementary Material online; 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.

Materials and Methods

Plant Growth

Arabidopsis 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 8/4 °C day/night under short day (12 h day, 12 h 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 (16 h day, 8 h 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. 2012).

Confocal Scanning Laser Microscopy

GFP tagged protein localization was observed trough CSLM on Leica SPE DM5500 upright microscope using a ACS APO 40×/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 scanning electron microscopy 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 stereomicroscope. 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, PA). Observations were made using a JEOL JSM-6360 microscope at the Department of Biology, KU Leuven.

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, Muino, et al. 2010). Sequencing libraries for AlySEP3 ChIP-seq were generated using Genome Analyzer IIx, HiSeq2000 or Miseq, see supplementary table S6, Supplementary Material online, 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 data sets were submitted to Gene Expression Omnibus (GEO) (accession number GSE63464). Sequences in FASTQ format were mapped to the unmasked *A. thaliana* 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 $>$ 1 Mb). 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 five 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 supplementary table S6, Supplementary Material online). 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 nonoverlapping windows of size 1 kb 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 Data Analysis

Directional RNA-seq libraries were generated and sequenced in triplicates for *A. thaliana* inflorescences, and *A. lyrata* inflorescences and leaves (supplementary table S6, Supplementary Material online). 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 with *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 Reannotation of A. lyrata and A. thaliana Genomes

Because the gene annotation of *A. lyrata* and *A. thaliana* may be of different quality, we have reannotated 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 genome, 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 transcripts were detected in the scaffold 1–9.

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, last accessed February 29, 2012), 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 differentially expressed between *A. lyrata* and *A. thaliana*. This information was also used to link candidate target genes of SEP3 in both plant species. List of orthologous gene pairs obtained by the method "Best-Hits-and-Inparalogs

family" was downloaded from PLAZA dicot 3.0 (Proost et al. 2015). We consider as paralogous pairs of genes, these homologous gene pairs 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 (February 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 organism, 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 was 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 $\pm$100 bp 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.

GO 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 (June 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 data sets 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

Supplementary figures S1–S8 and tables S1–S6 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors thank Professor Ya-Long Guo (State Key Laboratory of Systematic and Evolutionary Botany, Chinese Academy of Sciences) for providing them with the assembled data sets of TE insertions from *Arabidopsis thaliana* and *A. lyrata*. K.K. thanks the Alexander-von-Humboldt foundation and the BMBF. S.d.B. received a Netherlands Organisation for Scientific Research (NWO) Experimental Plant Sciences graduate school "master talent" fellowship. Data generated in this project were submitted to GEO with id: GSE63463. The authors declare no competing financial interest.

References

Abel C, Clauss M, Schaub A, Gershenzon J, 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.

Airoldi CA, Davies B. 2012. Gene duplication and the evolution of plant MADS-box transcription factors. *J Genet Genomics*. 39:157–165.

Anders S, Huber W. 2010. Differential expression analysis for sequence count data. *Genome Biol.* 11:R106.

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

Behrens S, Vingron M. 2010. Studying the evolution of promoter sequences: a waiting time problem. *J Comput Biol.* 17:1591–1606.

Bennett MD, Leitch IJ, Price HJ, Johnston JS. 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. *Ann Bot.* 91:547–557.

Bennetzen JL, Ma J, Devos KM. 2005. Mechanisms of recent genome size variation in flowering plants. *Ann Bot.* 95:127–132.

Biggin MD. 2011. Animal transcription networks as highly connected, quantitative continua. *Dev Cell.* 21:611–626.

Caris PL, Geuten KP, Janssens SB, Smets EF. 2006. Floral development in three species of *Impatiens* (Balsaminaceae). *Am J Bot.* 93:1–14.

Castillo-Davis CI, Hartl DL, Achaz G. 2004. cis-Regulatory and protein evolution in orthologous and duplicate genes. *Genome Res.* 14:1530–1536.

Cusanovich DA, Pavlovic B, Pritchard JK, Gilad Y. 2014. The functional consequences of variation in transcription factor binding. *PLoS Genet.* 10:e1004226.

de Koning AP, Gu W, Castoe TA, Batzer MA, Pollock DD. 2011. Repetitive elements may comprise over two-thirds of the human genome. *PLoS Genet.* 7:e1002384.

de Souza FS, Franchini LF, Rubinstein M. 2013. Exaptation of transposable elements into novel cis-regulatory elements: is the evidence always strong? *Mol Biol Evol* 30:1239–1251.

Dehal P, Boore JL. 2005. Two rounds of whole genome duplication in the ancestral vertebrate. *PLoS Biol.* 3:e314.

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

Feschotte C. 2008. Transposable elements and the evolution of regulatory networks. *Nat Rev Genet.* 9:397–405.
Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I. 2004. VISTA: computational tools for comparative genomics. *Nucleic Acids Res.* 32:W273–W279.
Guan Y, Guo J, Li H, Yang Z. 2013. Signaling in pollen tube growth: crosstalk, feedback, and missing links. *Mol Plant.* 6:1053–1064.
Haudry A, Platts AE, Vello E, Hoen DR, Leclercq M, Williamson RJ, Forczek E, Joly-Lopez Z, Steffen JG, Hazzouri KM, et al. 2013. An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. *Nat Genet.* 45:891–898.
Hawkins JS, Proulx SR, Rapp RA, Wendel JF. 2009. Rapid DNA loss as a counterbalance to genome expansion through retrotransposon proliferation in plants. *Proc Natl Acad Sci U S A.* 106:17811–17816.
He Q, Bardet AF, Patton B, Purvis J, Johnston J, Paulson A, Gogol M, Stark A, Zeitlinger J. 2011. High conservation of transcription factor binding and evidence for combinatorial regulation across six *Drosophila* species. *Nat Genet.* 43:414–420.
Heinz S, Romanoski CE, Benner C, Allison KA, Kaikkonen MU, Orozco LD, Glass CK. 2013. Effect of natural genetic variation on enhancer selection and function. *Nature* 503:487–492.
Henaff E, Vives C, Desvoyes B, Chaurasia A, Payet J, Gutierrez C, Casacuberta JM. 2014. Extensive amplification of the E2F transcription factor binding sites by transposons during evolution of *Brassica* species. *Plant J.* 77:852–862.
Hollister JD, Smith LM, Guo YL, Ott F, Weigel D, Gaut BS. 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, Goto K. 2001. Complexes of MADS-box proteins are sufficient to convert leaves into floral organs. *Nature* 409:525–529.
Hu TT, Pattyn P, Bakker EG, Cao J, Cheng JF, Clark RM, Fahlgren N, Fawcett JA, Grimwood J, Gundlach H, et al. 2011. The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. *Nat Genet.* 43:476–481.
Hupalo D, Kern AD. 2013. Conservation and functional element discovery in 20 angiosperm plant genomes. *Mol Biol Evol.* 30:1729–1744.
Johnson R, Gamblin RJ, Ooi L, Bruce AW, Donaldson IJ, Westhead DR, Wood IC, Jackson RM, Buckley NJ. 2006. Identification of the REST regulon reveals extensive transposable element-mediated binding site duplication. *Nucleic Acids Res.* 34:3862–3877.
Kaufmann K, Muino JM, Jauregui R, Airoldi CA, Smaczniak C, Krajewski P, Angenent GC. 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, Muino JM, Osteras M, Farinelli L, Krajewski P, Angenent GC. 2010. Chromatin immunoprecipitation (ChIP) of plant transcription factors followed by sequencing (ChIP-SEQ) or hybridization to whole genome arrays (ChIP-CHIP). *Nat Protoc.* 5:457–472.
Kaufmann K, Pajoro A, Angenent GC. 2010. Regulation of transcription in plants: mechanisms controlling developmental switches. *Nat Rev Genet.* 11:830–842.
Kaufmann K, Wellmer F, Muino JM, Ferrier T, Wuest SE, Kumar V, Serrano-Mislata A, Madueno F, Krajewski P, Meyerowitz EM, et al. 2010. Orchestration of floral initiation by APETALA1. *Science* 328:85–89.
Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. 2013. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biol.* 14:R36.
Kunarso G, Chia NY, Jeyakani J, Hwang C, Lu X, Chan YS, Ng HH, Bourque G. 2010. Transposable elements have rewired the core regulatory network of human embryonic stem cells. *Nat Genet.* 42:631–634.
Li HJ, Xue Y, Jia DJ, Wang T, Hi DQ, Liu J, Cui F, Xie Q, Ye D, Yang WC. 2011. POD1 regulates pollen tube guidance in response to micropylar female signaling and acts in early embryo patterning in *Arabidopsis*. *Plant Cell* 23:3288–3302.
Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. 2009. SOAP2: an improved ultrafast tool for short read alignment. *Bioinformatics* 25:1966–1967.
Lynch M, Bobay LM, Catania F, Gout JF, Rho M. 2011. The repatterning of eukaryotic genomes by random genetic drift. *Annu Rev Genomics Hum Genet.* 12:347–366.
Maere S, Heymans K, Kuiper M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics* 21:3448–3449.
Maston GA, Evans SK, Green MR. 2006. Transcriptional regulatory elements in the human genome. *Annu Rev Genomics Hum Genet.* 7:29–59.
Maumus F, Quesneville H. 2014. Ancestral repeats have shaped epigenome and genome composition for millions of years in *Arabidopsis thaliana*. *Nat Commun.* 5:4104.
Moghe GD, Shiu SH. 2014. The causes and molecular consequences of polyploidy in flowering plants. *Ann N Y Acad Sci.* 1320:16–34.
Muino JM, Kaufmann K, van Ham RC, Angenent GC, Krajewski P. 2011. ChIP-seq Analysis in R (CSAR): an R package for the statistical detection of protein-bound genomic regions. *Plant Methods* 7:11.
Muino JM, Smaczniak C, Angenent GC, Kaufmann K, van Dijk AD. 2014. Structural determinants of DNA recognition by plant MADS-domain transcription factors. *Nucleic Acids Res.* 42:2138–2146.
Ni X, Zhang YE, Negre N, Chen S, Long M, White KP. 2012. Adaptive evolution and the birth of CTCF binding sites in the *Drosophila* genome. *PLoS Biol.* 10:e1001420.
Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.
ÓMaoiléidigh DS, Wuest SE, Rae L, Raganelli A, Ryan PT, Kwasniewska K, Das P, Lohan AJ, Loftus B, Graciet E, et al. 2013. Control of reproductive floral organ identity specification in *Arabidopsis* by the C function regulator AGAMOUS. *Plant Cell* 25:2482–2503.
Pajoro A, Biewers S, Dougali E, Leal Valentim F, Mendes MA, Porri A, Coupland G, Van de Peer Y, van Dijk AD, Colombo L, et al. 2014. 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 JM, Matus JT, Jin J, Mecchia MA, Debernardi JM, Palatnik JF, Balazadeh S, Arif M, et al. 2014. Dynamics of chromatin accessibility and gene regulation by MADS-domain transcription factors in flower development. *Genome Biol.* 15:R41.
Pelaz S, Ditta GS, Baumann E, Wisman E, Yanofsky MF. 2000. B and C floral organ identity functions require SEPALLATA MADS-box genes. *Nature* 405:200–203.
Pertea M, Pertea GM, Antonescu CM, Chang TC, Mendell JT, Salzberg SL. 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nat Biotechnol.* 33:290–295.
Proost S, Pattyn P, Gerats T, Van de Peer Y. 2011. Journey through the past: 150 million years of plant genome evolution. *Plant J.* 66:58–65.
Proost S, Van Bel M, Vaneechoutte D, Van de Peer Y, Inze D, Mueller-Roeber B, Vandepoele K. 2015. PLAZA 3.0: an access point for plant comparative genomics. *Nucleic Acids Res.* 43:D974–D981.
Rodríguez-Mega E, Piñeyro-Nelson A, Gutierrez C, García-Ponce B, Sánchez MDLP, Zluhan-Martínez E, Álvarez-Buylla ER, Garay-Arroyo A. 2015. Role of transcriptional regulation in the evolution of plant phenotype: a dynamic systems approach. Dev Dyn. 244:1074–1095.
Schmidt D, Schwalie PC, Wilson MD, Ballester B, Goncalves A, Kutter C, Brown GD, Marshall A, Flicek P, Odom DT. 2012. Waves of retrotransposon expansion remodel genome organization and CTCF binding in multiple mammalian lineages. *Cell* 148:335–348.
Sicard A, Lenhard M. 2011. The selfing syndrome: a model for studying the genetic and evolutionary basis of morphological adaptation in plants. *Ann Bot.* 107:1433–1443.
Smaczniak C, Immink RG, Muino JM, Blanvillain R, Busscher M, Busscher-Lange J, Dinh QD, Liu S, Westphal AH, Boeren S, et al. 2012. Characterization of MADS-domain transcription factor complexes in *Arabidopsis* flower development. *Proc Natl Acad Sci U S A.* 109:1560–1565.
Smit AFA, Hubley R, Green P. 1996–2010. RepeatMasker Open-3.0 Available from: http://www.repeatmasker.org.

Smit AFA, Hubley R. 2008–2015. RepeatModeler Open-1.0 Available from: http://www.repeatmasker.org.

Smyth DR, Bowman JL, Meyerowitz EM. 1990. Early flower development in *Arabidopsis*. *Plant Cell* 2:755–767.

Stefflova K, Thybert D, Wilson MD, Streeter I, Aleksic J, Karagianni P, Brazma A, Adams DJ, Talianidis I, Marioni JC, et al. 2013. Cooperativity and rapid evolution of cobound transcription factors in closely related mammals. *Cell* 154: 530–540.

Stone JR, Wray GA. 2001. Rapid evolution of cis-regulatory sequences via local point mutations. *Mol Biol Evol.* 18:1764–1770.

Urbanus SL, de Folter S, Shchennikova AV, Kaufmann K, Immink RG, Angenent GC. 2009. In planta localisation patterns of MADS domain proteins during floral development in *Arabidopsis thaliana*. *BMC Plant Biol.* 9:5.

Van de Velde JV, Heyndrickx KS, Vandepoele K. 2014. Inference of transcriptional networks in *Arabidopsis* through conserved noncoding sequence analysis. *Plant Cell* 26:2729–2745.

Villar D, Flicek P, Odom DT. 2014. Evolution of transcription factor binding in metazoans—mechanisms and functional implications. *Nat Rev Genet.* 15:221–233.

Wright SI, Le QH, Schoen DJ, Bureau TE. 2001. Population dynamics of an Ac-like transposable element in self- and cross-pollinating arabidopsis. *Genetics* 158:1279–1288.

Wuest SE, O'Maoileidigh DS, Rae L, Kwasniewska K, Raganelli A, Hanczaryk K, Lohan AJ, Loftus B, Graciet E, 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.

Zhong S, Lin Z, Fray RG, Grierson D. 2008. Improved plant transformation vectors for fluorescent protein tagging. *Transgenic Res.* 17:985–989.