The Plant Cell, Vol. 25: 2813–2830, August 2013, www.plantcell.org

LARGE-SCALE BIOLOGY ARTICLE

The *Tarenaya hassleriana* Genome Provides Insight into Reproductive Trait and Genome Evolution of Crucifers[W][OPEN]

Shifeng Cheng,[a,1] Erik van den Bergh,[b,1] Peng Zeng,[a] Xiao Zhong,[a] Jiajia Xu,[c] Xin Liu,[a] Johannes Hofberger,[b] Suzanne de Bruijn,[d,e] Amey S. Bhide,[f] Canan Kuelahoglu,[g] Chao Bian,[a] Jing Chen,[a] Guangyi Fan,[a] Kerstin Kaufmann,[e] Jocelyn C. Hall,[h] Annette Becker,[f] Andrea Bräutigam,[g] Andreas P.M. Weber,[g] Chengcheng Shi,[a] Zhijun Zheng,[a] Wujiao Li,[a] Mingju Lv,[c] Yimin Tao,[c] Junyi Wang,[a] Hongfeng Zou,[a,i,j] Zhiwu Quan,[a,i,j] Julian M. Hibberd,[k] Gengyun Zhang,[a,i,j] Xin-Guang Zhu,[c] Xun Xu,[a] and M. Eric Schranz[b,2]

[a] Beijing Genomics Institute, 518083 Shenzhen, China
[b] Biosystematics Group, Wageningen University, 6708 PB Wageningen, The Netherlands
[c] Plant Systems Biology Group, Partner Institute of Computational Biology, Chinese Academy of Sciences/Max Planck Society, Shanghai 200031, China
[d] Molecular Biology Group, Wageningen University, 6708 PB Wageningen, The Netherlands
[e] Institute for Biochemistry and Biology, University of Potsdam, 14476 Potsdam, Germany
[f] Plant Developmental Biology Group, Institute of Botany, Justus-Liebig-University, 35392 Giessen, Germany
[g] Institute of Plant Biochemistry, Center of Excellence on Plant Sciences, Heinrich-Heine-University, D-40225 Duesseldorf, Germany
[h] Department of Biological Sciences, University of Alberta, Edmonton, Alberta, Canada T6G 2E9
[i] State Key Laboratory of Agricultural Genomics, Beijing Genomics Institute, 518083 Shenzhen, China
[j] Key Laboratory of Genomics, Ministry of Agriculture, Beijing Genomics Institute, 518083 Shenzhen, China
[k] Department of Plant Sciences, University of Cambridge, Cambridge CB2 3EA, United Kingdom

The Brassicaceae, including *Arabidopsis thaliana* and *Brassica* crops, is unmatched among plants in its wealth of genomic and functional molecular data and has long served as a model for understanding gene, genome, and trait evolution. However, genome information from a phylogenetic outgroup that is essential for inferring directionality of evolutionary change has been lacking. We therefore sequenced the genome of the spider flower (*Tarenaya hassleriana*) from the Brassicaceae sister family, the Cleomaceae. By comparative analysis of the two lineages, we show that genome evolution following ancient polyploidy and gene duplication events affect reproductively important traits. We found an ancient genome triplication in *Tarenaya* (Th-α) that is independent of the Brassicaceae-specific duplication (At-α) and nested *Brassica* (Br-α) triplication. To showcase the potential of sister lineage genome analysis, we investigated the state of floral developmental genes and show *Brassica* retains twice as many floral MADS (for MINICHROMOSOME MAINTENANCE1, AGAMOUS, DEFICIENS and SERUM RESPONSE FACTOR) genes as *Tarenaya* that likely contribute to morphological diversity in *Brassica*. We also performed synteny analysis of gene families that confer self-incompatibility in Brassicaceae and found that the critical *SERINE RECEPTOR KINASE* receptor gene is derived from a lineage-specific tandem duplication. The *T. hassleriana* genome will facilitate future research toward elucidating the evolutionary history of Brassicaceae genomes.

INTRODUCTION

Studies of the model plant *Arabidopsis thaliana* and its close relatives in the Brassicaceae family have provided fundamental insight into the processes and patterns of plant evolution and function (Koornneef and Meinke, 2010; Hu et al., 2011; Wang et al., 2011). Comparative analyses between Brassicaceae and crop species have had profound influences on plant improvement and production. For example, knowledge about the control and evolution of plant reproductive traits, such as floral and fruit development and self-incompatibility (SI) systems, can be directly related to plant fitness and yield (Tanksley, 2004; Shen et al., 2005). The Brassicaceae have also been a model for understanding the dynamics and impacts of ancient polyploidy (genome doubling), considering that the entire family has undergone a whole-genome duplication (named At-α) and the *Brassica* crops have had an additional genome triplication (Br-α) (Blanc et al., 2003; Thomas et al., 2006; Wang et al., 2011). Genes retained in multiple copies due to these ancient polyploidy events, in addition to more recent tandem duplications, have played important roles in the evolution and regulation of

[1] These authors contributed equally to this work.
[2] Address correspondence to eric.schranz@wur.nl.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: M. Eric Schranz (eric.schranz@wur.nl).
[W] Online version contains Web-only data.
[OPEN] Articles can be viewed online without a subscription.
www.plantcell.org/cgi/doi/10.1105/tpc.113.113480

key traits (Edger and Pires, 2009; Flagel and Wendel, 2009). However, the polyploid history of the Brassicaceae also complicates synteny and evolutionary inferences to distantly related crop species.

To fully exploit the fundamental trait and genome insights garnered from Brassicaceae systems and improve synteny analyses to more distant crops, we report the genome sequencing and analysis of *Tarenaya hassleriana* from the Brassicaceae sister family Cleomaceae. Currently, papaya (*Carica papaya*), a member of the order Brassicales, is the closest relative with a complete genome sequence; however, these two lineages diverged more than 70 to 110 million years ago (Ming et al., 2008; Beilstein et al., 2010). The Cleomaceae is the phylogenetic sister family to the Brassicaceae, with the two lineages having diverged only ~38 million years ago (Schranz and Mitchell-Olds, 2006; Couvreur et al., 2010). Brassicaceae and Cleomaceae share many traits (Hall et al., 2002; Iltis et al., 2011), such as a preponderance of herbaceous species, the same general floral ground plan (four sepals, four petals, six stamens, and two fused carpels), and a replum in the mostly dehiscent fruits, referred to as capsules. There are also a number of key differences. Most of the 300 Cleomaceae species are restricted to the semitropics and arid desert regions and lack a genetic pollen-pistil SI system, whereas most of the 3700 Brassicaceae species largely radiated into cold temperate regions and possess a genetically regulated SI system (Guo et al., 2011). Another striking distinction is in floral symmetry: Cleomaceae have mostly monosymmetric flowers and Brassicaceae have mostly disymmetric flowers (Endress, 1999; Patchell et al., 2011). Cleomaceae also exhibit greater variation in the basic floral plan with increases in stamen number, petal dimorphisms, and stalks to the ovary, whereas Brassicaceae exhibit greater diversity in fruit morphology and dehiscence capabilities (Franzke et al., 2011). Comparative analyses can be used to elucidate the genomic basis of these differences. The Cleomaceae species we sequenced is *T. hassleriana*, often referred to as the spider flower, which is widely grown as an ornamental species and used as an educational model (Marquard and Steinback, 2009). This species was formerly named *Cleome hassleriana* (often erroneously labeled as *Cleome spinosa*), but the genus *Cleome* has undergone recent taxonomic revisions (Iltis and Cochrane, 2007).

Brassicaceae and Cleomaceae have undergone independent ancient polyploidy events. At least five ancient polyploidy events have occurred in the evolutionary history of *Arabidopsis* (Bowers et al., 2003; Van de Peer et al., 2009), four of which are shared with Cleomaceae: ζ near the origin of seed plants (Jiao et al., 2011), ε near the origin of angiosperms (Jiao et al., 2011), the ancient hexaploidy At-γ shared by nearly all eudicots (Jaillon et al., 2007; Vekemans et al., 2012), and At-β restricted to part of the order Brassicales as it is lacking from the papaya genome (Ming et al., 2008). The most extensively studied ancient polyploidy event is the more recent At-α genome duplication (Arabidopsis Genome Initiative, 2000; Bowers et al., 2003; Schnable et al., 2012) and is shared by all Brassicaceae species (Schranz et al., 2012). The crop genus *Brassica* has all the ancient polyploidy events in common with *Arabidopsis* but also has undergone an additional and more recent whole-genome triplication (hexaploidy) event (Br-α) after its split with *Arabidopsis* around ~17 million years ago (Wang et al., 2011). Limited BAC and transcriptome sequencing revealed that *Tarenaya* lacked the At-α event and that it underwent an independent ancient genome triplication (Th-α) (Schranz and Mitchell-Olds, 2006; Barker et al., 2009). Thus, *Tareneya* provides a unique opportunity to contrast genome evolution from a common ancestor comparing three genomic equivalents in *Tarenaya*, two in *Arabidopsis*, and six in *Brassica*, and furthermore to contrast two independent ancient genome triplications (Th-α versus Br-α). We not only compare these polyploidy events and more recent tandem duplication events, but also show how they contributed to the genes regulating key reproductive traits (Van de Peer et al., 2009).

RESULTS

Genome Sequencing and Integration with Physical Map

The *T. hassleriana* genome is relatively small (~290 Mb: 2n = 20) and within the range of sequenced Brassicaceae species: *Schrenkiella parvula* (formerly *Thellungiella parvula*), 140 Mb (Dassanayake et al., 2011); *A. thaliana*, 157 Mb (Bennett et al., 2003); *Arabidopsis lyrata*, 207 Mb (Hu et al., 2011); *Capsella rubella*, 210 Mb (Slotte et al., 2013); *Aethionema arabicum*, 240 Mb (Haudry et al., 2013); *Sisimbrium irio*, 262 Mb (Haudry et al., 2013); *Eutrema salsugineum* (formerly *Thelungiella salsuginea*), 314 Mb (Wu et al., 2012; Yang et al., 2013); *Leavenworthia alabamica*, 316 Mb (Haudry et al., 2013); and *Brassica rapa* 485 Mb (Wang et al., 2011). To generate a high-quality draft genome assembly, we used both sequenced paired-end libraries and constructed a Bacterial Artificial Chromosome (BAC) based whole-genome profiling (WGP) physical map. We used the Illumina next-generation sequencing platform to generate ~70.2 Gb (245X genome-depth) raw data of paired-end reads ranging from 90 to 100 bp (see Supplemental Table 1 online) from seven libraries with various insert sizes (350 to 20 kb). Sequence data were filtered, yielding ~40 Gb of high-quality sequence (~139X coverage) (see Supplemental Table 2 online). Sequences were assembled using SOAPdenovo (Li et al., 2010) (version 2.21) (see Supplemental Table 3 online). We also sequenced and mapped over 4 Gb of transcriptome data to the assembly showing that >94% of the genic regions were covered (see Supplemental Table 4 online).

The physical map was made using the Keygene WGP fingerprinting technique (van Oeveren et al., 2011). We generated 192,000 BAC-based Illumina sequence tags from 19,200 BACs from two libraries (*Eco*RI and *Mse*I) with an average insert size of ~125 kb (giving a total of 32X genome equivalents) (see Supplemental Table 5 online). We identified 87,617 high-quality and unique WGP tag sequences that allowed us to uniquely identify 15,567 BACs (with an average of 40.3 tags per BAC). These were used to build a high-stringency map assembly using modified Finger Printed Contigs (FPC) software (Engler et al., 2003) to generate 786 contigs using data from 9396 BACs (see Supplemental Table 6 online). We integrated the WGP physical map scaffolds with the Short Oligonucleotide Analysis Package (SOAP) sequence scaffolds to produce 77 superscaffolds from the integration of 349 sequence scaffolds (integrating between two and 21 scaffolds per superscaffold with an average of 4.5

Table 1. Summary of the genome sequencing, assembly, and annotation.

Assembly			
	N50 (size/number)	N90 (size/number)	Total sizes
Contigs	21.58 kb/2761	2.7 kb/13591	222 Mb
Scaffolds	551.9 kb/98	64.8 kb/622	256.5 Mb
Superscaffolds	1.26Mb/40	7.4kb/1014	273Mb
Annotation			
	Glean	RNA-Seq supported	Homologous with *Arabidopsis*
# Genes	28917	20337	24245
	LTR	DNA transposons	Total size
TE sizes, Mb (%)	97.28 (38.19)	11.8 (4.62)	110 (43.3)

scaffolds per superscaffold) through Basic Local Alignment Search Tool (BLAST) mapping of the WGP anchors (tags). We evaluated the quality of the integration between physical map and superscaffolds by manually checking the ordering and orientation of connected scaffolds by analyzing collinearity relative to *A. lyrata* (example shown in Supplemental Figure 1 online), confirming that all *Tarenaya* superscaffolds have extensive and extended synteny. We further validated our assembly by comparing it with four previously published Sanger-sequenced BACs (Schranz and Mitchell-Olds, 2006; Navarro-Quezada, 2007), revealing nearly identical assemblies (see Supplemental Figure 2 online). The final assembly statistics of the integrated dataset are summarized (Table 1; see Supplemental Table 7 online). With this integration, the N50 was increased by more than 2.6-fold (N50 = 1.26 Mb) due to the merger of most of the de novo assembled scaffolds into superscaffolds.

Gene Annotation

Gene annotation was conducted using a pipeline that integrates de novo gene prediction, homology-based alignment, and RNA-seq data. In total, we conducted >4 Gb of RNA-seq, of which 77.4% of reads could be confidently mapped onto the genome (see Supplemental Table 8 online). To analyze the overall gene expression patterns and to provide basic gene expression information, transcriptomes were generated for several *T. hassleriana* tissues (see Supplemental Table 9 online). A principal component analysis showed that stamen, root, and seed profiles separate most, a result similar to the pattern detected in *A. thaliana* where these three tissue types separate most over the first two components (see Supplemental Figure 3 online) (Schmid et al., 2005).

The prediction of gene models by various techniques was summarized (see Supplemental Table 10 online), with a large range in the number of predicted genes. To be conservative, we used the models predicted by GLEAN to have a high posterior probability out of the various gene prediction techniques, which resulted in the identification of 28,917 highly supported gene models with an average transcript length of 2216 bp, coding sequence size of 1169 bp, and 5.27 exons per gene, both similar to that observed in *A. thaliana* and *B. rapa* (see Supplemental Figure 4 online). A total of 92.9% of gene models have a homolog match or conserved motif in at least one of the public protein databases, including Swissprot (McMillan and Martin, 2008), 71.1%; the Translated European Molecular Biology Laboratory database (Boeckmann et al., 2003), 92.5%; InterPro (Zdobnov and Apweiler, 2001), 74.9%; the Kyoto Encyclopedia of Genes and Genomes (Kanehisa and Goto, 2000), 55.2%; and Gene Ontology (GO) (Ashburner et al., 2000), 55.7% (see Supplemental Table 11 online), and 97.1% are represented among the public Expressed Sequence Tag (EST) collections or de novo Illumina mRNA-Seq data. In addition to protein-coding genes, we also identified 220 microRNA, 862 tRNA, and 685 small nuclear RNA genes in the *T. hassleriana* genome (see Supplemental Table 12 online). Orthologous clustering of proteomes predicted for *T. hassleriana* and three Brassicaceae species (*A. thaliana*, *A. lyrata*, and *B. rapa*) revealed 15,112 genes in 12,689 families in common (Figure 1). We found that

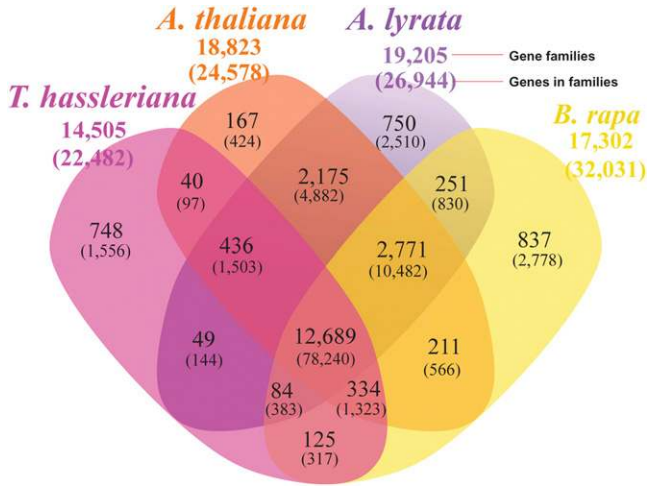

Figure 1. Venn Diagram Illustrating the Shared and Unique Gene Families from *T. hassleriana* (Cleomaceae), *A. thaliana*, *A. lyrata*, and *B. rapa* (Brassicaceae).

In total, we predicted 28,917 well-supported gene models for *T. hassleriana*, of which 22,482 could be placed into one of the 14,505 gene families. A total of 87.5% of gene families found in *Tarenaya* were present in all three Brassicaceae species and 7.4% in one or two Brassicaceae species, and only 5% of gene families were unique to *Tarenaya* with many of these associated genome-specific retrotransposons. Thus, comparative functional and evolutionary analysis of well-characterized *Arabidopsis* and Brassicaceae genes is feasible using *Tarenaya* as an outgroup.

20,926 *T. hassleriana* genes clustered with at least one of the three genomes. Furthermore, 1556 *Tarenaya*-specific genes in 748 families were identified, most of which were enriched for genes of unknown function and for which 34% have EST supported annotation.

To evaluate the status of our genomic assembly, we compared it to the nine published crucifer genomes (Haudry et al., 2013) (see Supplemental Table 13 online). Our *T. hassleriana* assembly has the highest assembled sequence versus expected genome size with an estimated 93% genome size completeness. For comparison, the *B. rapa* assembly only has 51.6% genome coverage. Our assembly also has the largest scaffold N50 (1.26 Mb) of the Illumina-only sequenced genomes (*T. hassleriana*, *Ae. arabicum*, *S. irio*, *Leavenworthia alabamica*, and *B. rapa*). The number of predicted genes is on par with that in the other crucifers and the number of *A. thaliana* orthologs is slightly lower than average, as expected from an organism outside the Brassicaceae. Eighty-eight percent of ultraconserved core eukaryotic genes (Parra et al., 2009) were found, suggesting a slightly lower covered gene space. This might be due to the fact that the percentage of transposable elements (TEs) is relatively high, which can cause difficulties assembling regions where TEs and genes are intermixed.

Because of our high genome sequencing coverage, we were able to identify more than 43% of the 293-Mb *T. hassleriana* genome as being composed of transposons. For comparison, only 31% of the 529 Mb *B. rapa* genome has been identified as transposons (see Supplemental Table 13 online). The discrepancy is because of the great difference in the coverage of the assemblies (93% versus 56%). Thus, the percentage of transposons in *Brassica* is largely underestimated. To examine the contiguity of the genome assembly (causes for gaps between contigs in the scaffolds), we analyzed the distribution of contigs versus transposon long-terminal repeats across the largest 491 scaffolds (see Supplemental Figure 5 online). By doing so, we can demonstrate that regions with a high density of long terminal repeats (LTRs) correspond to smaller contigs in scaffolds and thus generate regions of lower contiguity. Both de novo repeat identification and homology-based methods were applied to predict transposable elements (TEs) (see Supplemental Table 14 online). The majority of repetitive sequences were Class I long-terminal repeat retrotransposons, constituting 36.6% of the genome compared with 27.1% in *B. rapa*. The overall lower percentages of annotated transposons in *Brassica* are likely due to its lower genome sequence coverage (see Supplemental Table 13 online) because the *B. rapa* genome is nearly 200 Mb larger than *Tarenaya*. Most of the repeats were located in the intergenic regions.

Comparative Analysis of Ancient Polyploidy Events

The Brassicaceae-Cleomaceae system allowed us to compare genome evolution after several rounds of independent ancient polyploidy (Figure 2). We confirmed that the Cleomaceae polyploidy event (Th-α) occurred independently of, and more recently than, the Brassicaceae-specific duplication event detected in *Arabidopsis* and *Brassica* (At-α). We also detected the nested *B. rapa* ancient hexaploidy (triplication of the genome) event (Br-α) and show that it is of approximately the same age as Th-α. For all three taxa, we detected the diffuse signal of the older and shared events (At-β, At-γ, ε, and ζ) (Figure 2). The Th-α and Br-α events are of approximately the same age and, as discussed below, represent independent ancient hexaploidy events. Using whole genome intragenomic dot plots of *Tarenaya*, we showed many triplicated blocks (see Supplemental Figure 6 online), and analysis of syntenic depth with QuotaAlign (Tang et al., 2011) shows that 49.4% of genes are found at 3× coverage (see Supplemental Table 15 online). To illustrate the homologous relationships and the evolutionary history of triplicated/duplicated segments in Cleomaceae and Brassicaceae, we integrated intra- and intergenomic analyses (Figure 3). We analyzed synteny relationships

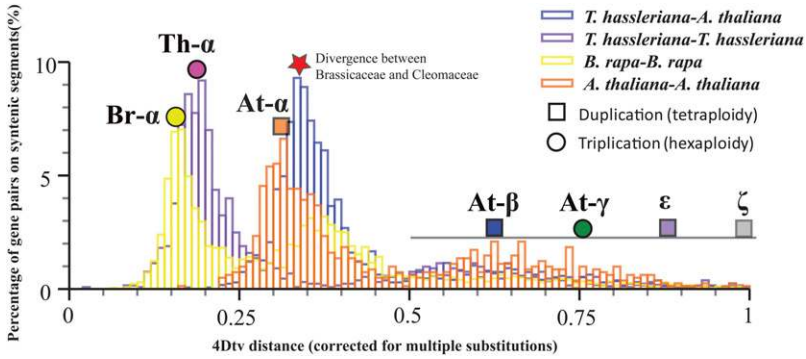

Figure 2. Relative Timing of the Polyploidy Events and Lineage Splitting Based on Divergence of Fourfold Degenerate Sites (4DTv) for Duplicated Genes within *A. thaliana*, *B. rapa*, and *T. hassleriana* and Orthologous Genes between *A. thaliana* and *T. hassleriana*.

All plots detect broad overlapping peaks between 0.5 and 1.0, representing shared older polyploidy events (At-β, At-γ, ε, and ζ). The divergence of the Brassicaceae-Cleomaceae lineages is seen by the differentiation of *Arabidopsis* and *Tarenaya* homologs at the peak centered at ~0.35 (highlighted by red star). The divergence between paralogs from the At-α duplication event occurred slightly after the lineage splitting and is detected by the peaks centered at ~0.3 for both *Arabidopsis* and *Brassica*. The At-α peak is lacking from *Tarenaya*, proving At-α is Brassicaceae specific. Nearly overlapping distributions between 0.15 and 0.25 were detected for *Brassica* and *Tarenaya*, representing the independent Br-α and Th-α ancient hexaploidy events, respectively.

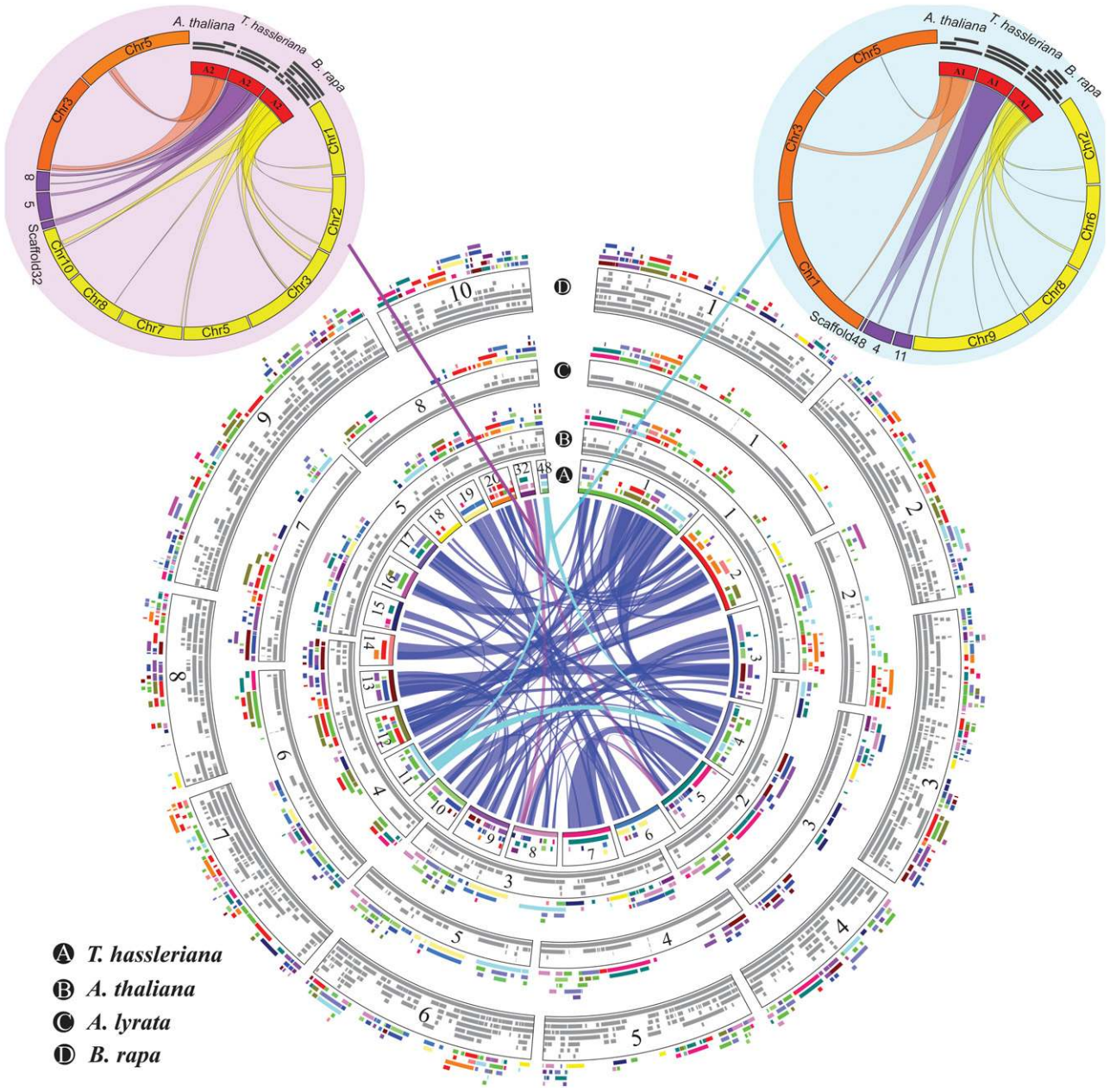

Figure 3. Homologous Genome Blocks within and between Genomes for Cleomaceae and Brassicaceae.

The largest 20 (plus additional two smaller scaffolds) color-coded superscaffolds of *T. hassleriana* are taken as the reference, such that any region homologous to the 22 scaffolds is colored accordingly. A, Self-alignment of *Tarenaya* superscaffolds, with the inner circle showing links of syntenic blocks. Over 47% of the genome is found in three copies, supporting the conclusion that it experienced an ancient hexaploidy event (Th-α = triplication). Rings within a genome (inner gray bars) and blocks homologous to *Tarenaya* (outer color-coded bars) for completed Brassicaceae genomes: B, *A. thaliana*; C, *A. lyrata*; and D, *B. rapa*. The inner gray bars show a clear pattern related to the ancient polyploidy events of the Brassicaceae (At-α = duplication) and nested Brassica-specific lineage (Br-α = triplication). The color-coded outer rings of homology relative to *Tarenaya* show a complex pattern due to the independent polyploidy events between families. The two small insets illustrate examples of the three *Tarenaya* to two *Arabidopsis* to six *Brassica* genome equivalents due to ancient polyploidy events.

both within and between genomes (see Supplemental Figures 6 to 9 online). For *T. hassleriana*, 86, 83, and 85% of the protein-coding genes were homologous to the genes in the *A. thaliana*, *A. lyrata*, and *B. rapa* genomes, respectively (see Supplemental Table 16 online). By making these comparisons of *Tarenaya* versus *A. thaliana*, *A. lyrata*, and *B. rapa* genomes (Figure 3), we found significant, 3:2, 3:2, 3:4, 3:5, and 3:6 homologous patterns, respectively, which is consistent with the polyploid history of the species. To illustrate this pattern, we highlighted two ancestral blocks (A1 and A2) (Figure 3, two small insets). Note that the three

Tarenaya blocks show almost perfect collinearity, whereas one of two *Arabidopsis* regions is broken across two chromosomes, suggesting a Brassicaceae-specific rearrangement(s) after At-α. Since synteny analysis has been extensively performed within Brassicaceae, we also show our results with the collinear blocks color-coded according to the current Brassicaceae conventions (Schranz et al., 2006) (see Supplemental Figure 10 online).

We inferred the putative "A ancestor" (pre-At-α) shared by *A. thaliana* and *A. lyrata*, the "B ancestor" of *B. rapa* (pre-Br-α but post-At-α ancestral genome state), and the "T ancestor" of *T. hassleriana* (pre-Th-α ancestral genome state) (see Methods). We compared our identified homologous replicated blocks within and between genomes in Brassicaceae species and *T. hassleriana* with the results of an earlier analysis of conserved At-α blocks (Blanc et al., 2003; Thomas et al., 2006). First, we reconstructed our version of the pre-At-α ancestor (A ancestor) of *A. thaliana* (version TAIR9), which resulted in 64 ancestral regions involving 19,976 protein-coding genes (see Supplemental Table 15 online). The corresponding relationships to the blocks identified by Wolfe and colleagues (Blanc et al., 2003) (that we refer to as "Wolfe blocks") are illustrated in Supplemental Figure 11 online. We used the same method on the minimized genomes of *A. lyrata*, *B. rapa*, and *T. hassleriana* and generated 61, 71, and 87 conserved ancestral blocks covering 24,373; 25,646; and 20,680 protein-coding genes, respectively, to represent the postulated A, B, and T ancestors (see Supplemental Figures 12 to 14 online). A table listing all genes included in each block for each genome is provided in the supplemental materials (see Supplemental Data Set 1 online). A comparison of the reconstructed T and A ancestor genomes revealed a 1:1 relationship, supporting our conclusion that the ancestral genome of the Brassicaceae and Cleomaceae was conserved before the independent duplication (At-α) and triplication (Th- α) events, respectively. Since *B. rapa* underwent a nested and specific triplication following the At-α, we see a 1:2 pattern when we compare the inferred T to B ancestors. A comparison of conserved ancestral genomic blocks across species is shown in Supplemental Figure 15 online. We then traced the extent of gene retention and fractionation in homologous blocks after polyploidy events. We partitioned the two subgenomes of *Arabidopsis*, three subgenomes of Brassica, and three subgenomes of *T. hassleriana*, respectively, by comparing them with the reference A ancestor of *A. lyrata* (see Supplemental Figures 16 to 19 online). These improvements to understanding genome evolution after independent ancient polyploidy events of Brassicaceae species will facilitate synteny analyses in distant crop species.

Comparative Analysis of Type II MADS Box Genes

The development of the four floral organ types and later the fruits is regulated by Type II MINICHROMOSOME MAINTENANCE1, AGAMOUS, DEFICIENS and SERUM RESPONSE FACTOR (MADS) domain proteins (Smaczniak et al., 2012) as described by the ABCDE model (Theissen, 2001). The types of MADS box genes that regulate development are remarkably well conserved across eudicots, with the molecular mechanisms of their action extensively studied in *Arabidopsis*. We found that the *Tarenaya* genome contains representatives of all the major Type II MADS box genes described in *Arabidopsis* and *Brassica* (see Supplemental Figure 20 online). We concentrated on the retention of the MADS box genes derived from At-α, Br-α, and Th-α and compared this with the polyploid origins of additional duplicates (At-β, At-γ, ε, T [tomato] [Tomato Genome Consortium, 2012], and Pt-α [poplar] [Tuskan et al., 2006]) (Figure 4). Theoretically, the At-α, Br-α, and Th-α events should have given rise to two *Arabidopsis*, six *Brassica*, and three *Tarenaya* gene copies (syntelogs) from a single ancestral gene. Of the 11 MADS box gene clades involved in floral, fruit, and inflorescence development that were likely present in the most recent common ancestor of Brassicaceae and Cleomaceae shown in Figure 4, we found that only three duplicate pairs are in fact maintained in *Arabidopsis* due to At-α: *APETALA1* (*AP1*)/*CAULIFLOWER* (*CAL*) (A-function), *SHATTERPROOF1* (*SHP1*)/*SHP2* (D-function), and *SEPALLATA1* (*SEP1*)/*SEP2* (E-function). Thus, *Arabidopsis* has only three of 11 possible replicates (27.3% syntelog retention). This implies that during early Brassicaceae evolution (before the split of *Arabidopsis-Brassica*), there were 14 gene lineages. From these 14 lineages then there would be 42 possible syntelogs in *Brassica* due to Br-α triplication, of which 28 copies are considered as additional syntelogs. We find a remarkable 19 of these additional gene copies (67.8% syntelog retention). This includes all three possible copies maintained for the following seven Brassica gene families: *SEP4*, *SEP3*, *AGAMOUS-LIKE79* (*AGL79*), *FRUITFULL*, *AP1*, *PISTILLATA* (*PI*), and *SHP1*. The only two genes to return to single copy in *Brassica* after Br-α are *CAL* and *SHP2*. When we also consider the At-α gene retention, then a maximum of four of six gene copies are found in the *SEP1/2* clade, *AP1/CAL* clade, and the *SHP1/2* clade. From the 11 ancestral loci, 33 syntelogs would be expected in *Tarenaya* due to the Th-α triplication, with 22 possible additional syntelogs. However, we only recovered six (27.3% syntelog retention) with no cases where all three possible copies are maintained (we do not count the additional tandem duplicate of Th-*AP3* here, which is discussed below). Thus, we find more than double the syntelog retention in *Brassica* than in *Tarenaya*, despite the fact that both are ancient hexaploids of approximately the same age. The greatest differential in gene copy retention between *Brassica* and *Tarenaya* is for the *AGL79* clade (3 versus 1) and the *SHP1/2* clade (4 versus 1). The *SHATTERPROOF* genes in Brassicaceae regulate various traits during carpel and fruit formation (Colombo et al., 2010). The single-copy nature of the *SHP* homolog in *Tarenaya* is thus notable, since this is the only gene that is duplicated in *Arabidopsis* due to At-α but has returned to single copy in *Tarenaya* (see Supplemental Figure 21 online). In Cleomaceae, fruit morphology is less diverse than in Brassicaceae, and we hypothesize that the retention of *SHP* genes plays an important role in the morphological variability of Brassicaceae.

Comparative Collinearity Analysis of Floral Developmental Regulators

To assess the contributions of ancient polyploidy and tandem gene duplications to floral regulatory gene diversification, we conducted a more detailed analysis of gene synteny and expression patterns. Almost all floral MADS box genes show conserved synteny between Brassicaceae and Cleomaceae. For example, the A-class genes show stable duplicate retention; the loci containing

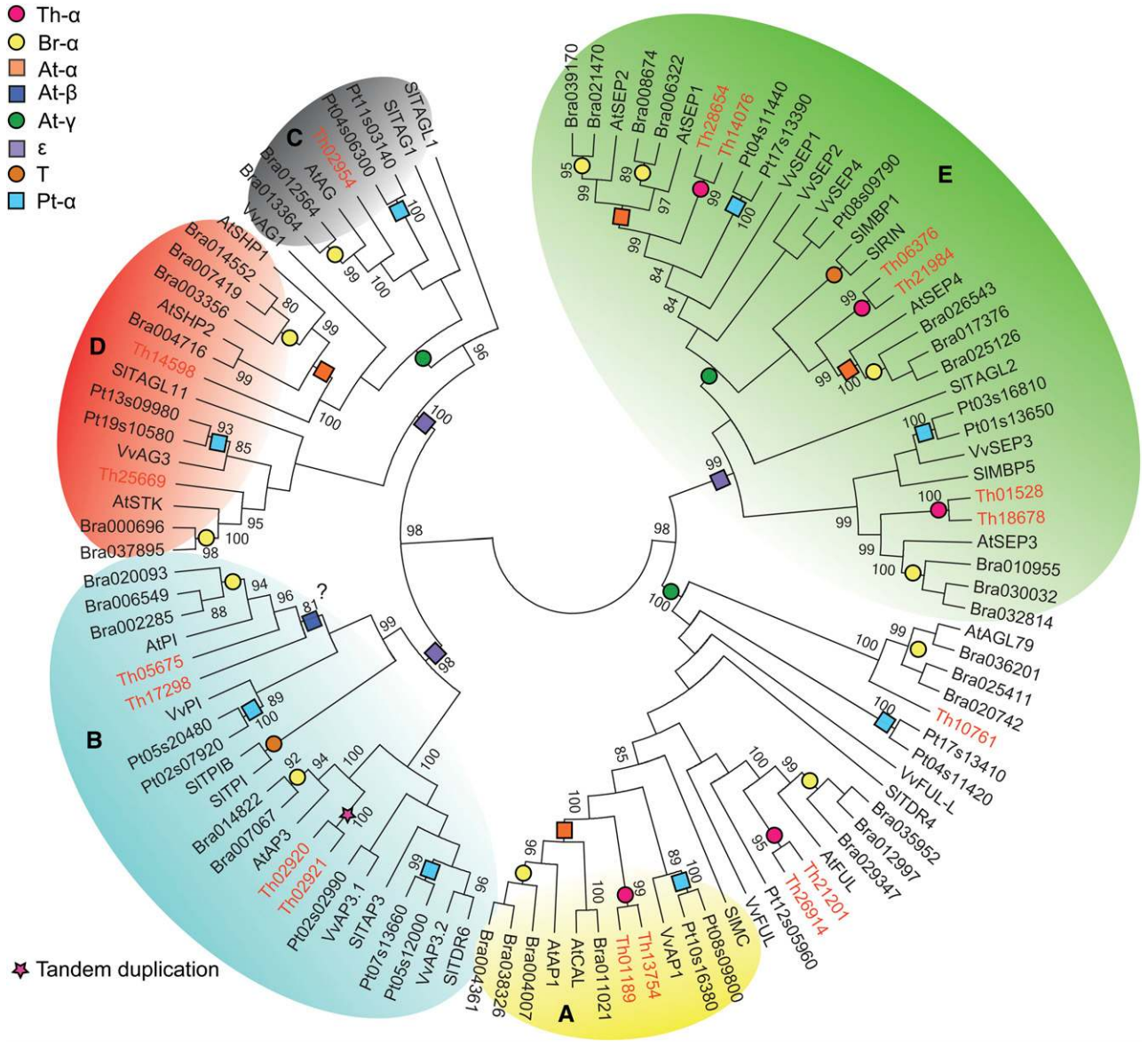

Figure 4. Phylogenetic Tree of Type-II MADS Box Transcription Factor Genes Involved in Floral Organ Specification.

The alignment used to create this tree is available in the online materials (see Supplemental Data Set 2 online). The major floral MADS box genes cluster into five groups corresponding to the five main functional types (AP1-like genes, shown in yellow; AP3/PI-like genes, B-type shown in blue; AG-like genes shown in gray; STK-like genes shown in red; and SEP-like genes in green) according to the ABC(DE) model of floral development. Species included are *A. thaliana* (At), grape (*Vitis vinifera*, Vv), tomato (*Solanum lycopersicum*, Sl), poplar (*Populus trichocarpa*, Pt), *B. rapa* (Br), and *T. hassleriana* (Th). *Tarenaya* genes are indicated in red. The colored squares (duplication events) and circles (triplication events) placed on nodes represent gene lineage expansion(s) that can be associated with particular ancient polyploid events: Th-α, At-α, Br-α, At-β, At-γ, ε, T (identified by tomato genome sequencing), and Pt-α (identified by poplar genome sequencing). Type-II MADS box genes are often retained after ancient polyploidy events. From the tree above, we calculated a 27.3% syntelog (homolog generated by a polyploidy event) retention after At-α, 27.3% syntelog retention after Th-α, and a much higher (67.8%) syntelog retention after Br-α, despite the fact that Th-α and Br-α triplications are of approximately the same age. The *Tarenaya* B-class genes show unusual patterns in that the *AP3* homologs (Ch02920 and Th02921) represent a recent tandem duplication, which is rare for floral MADS box genes, and there are two copies of *PI* homologs that are likely due to At-β with one lineage being lost in Brassicaceae. Shown is a maximum likelihood tree with 1000 replicate bootstrap values, of which the branches with a bootstrap value of >80 are presented, visualized topology only.

the *AP1* and *CAL* homologs in *Tarenaya*, *Arabidopsis*, and *Brassica* are syntenic to one another with little evidence of local rearrangements (see Supplemental Figure 22 online).

Compared with other MADS box genes, the B-class (*PI* and *AP3*) genomic regions show a more dynamic pattern (Figure 5). The split between *PI* and *AP3* is an old duplication due to the angiosperm ε polyploidy event (Figure 4), with almost no detectable collinearity between these regions (Figure 5). Comparison of B-class *Tarenaya* genomic regions with Brassicaceae allowed us to detect two B-class duplication events: A recent *Tarenaya* tandem duplication of *AP3* (Th02920 and Th02921) and an older *PI* duplication, likely due to At-β, which has been lost from Brassicaceae but is still retained in *Tarenaya* (Th17298) (Figure 5; see Supplemental Figures 23 and 24 online).

Strikingly, we also detected two gene transposition events: a Brassicaceae-specific AP3 transposition and a shared transposition event of one *PI* gene before the split of Brassicaceae and Cleomaceae (Figure 5). The Brassicaceae-specific AP3 transposition event also involved the flanking EMBRYO DEFECTIVE1967 (EMB1967) gene containing two conserved domains: the N-terminal region of microspherule protein (MCRS_N) and Forkhead-associated (FHA). Tarenaya AP3 is similarly flanked by a Forkhead-associated protein (Th02919) that has its highest match to EMB1967; however, the orientation of AP3 and the Forkhead genes is inverted between species (see Supplemental Figure 25 online) and is also detected in a distantly related Cleome species. In general, B-class genes are functionally highly conserved across angiosperms, whereas A-class gene function appears to be less conserved (Litt and Kramer, 2010). However, we have shown that it is in fact the B-class genes that have undergone transposition events.

Members of the TEOSINTE BRANCHED1, CYCLOIDEA, and PCF (TCP) gene family play an important role in the transition from polysymmetric to monosymmetric flowers (reviewed in Busch and Zachgo, 2009; Jabbour et al., 2009; Rosin and Kramer, 2009), including monosymmetric Brassicaceae (Busch and Zachgo, 2007; Busch et al., 2012). Cleomaceae floral morphology, especially in petal and stamen position, numbers, and asymmetry, is quite variable. However, the role of Tarenaya TCP homologs in monosymmetry has not yet been fully characterized. We find a pattern of conservation of genomic collinearity around the TCP1 locus between species (see Supplemental Figure 26 online). *Arabidopsis* contains only a single TCP1 locus, as does *A. lyrata*. Due to Br-α, Brassica has three syntenic copies of TCP1. We also can detect the syntenic regions in Brassicaceae species due to At-α (see Supplemental Figure 26 online) but find no At-α derived homologs of TCP genes, suggesting the loss of a TCP1 syntelog occurred early in Brassicaceae evolution. In Tarenaya, we find three genomic regions derived from Th-α, with two copies of TCP1 intact (Th21666 and Th24587) (see Supplemental Figure 26 online). The correlation between multiple copies of TCP members and monosymmetry has been noted across angiosperms (Rosin and Kramer, 2009).

Expression of A- and B-Class Homolog Genes

The expression of *T. hassleriana* homologs of *A. thaliana* major floral regulators was also analyzed with quantitative RT-PCR (qRT-PCR). The two putative *T. hassleriana* homologs of *CAL/AP1* (Th-*CAL/AP1-1* and Th-*CAL/AP1-2*) show similar expression in all bud stages, but Th-*CAL/AP1-1* with only half of the transcript abundance of Th-*CAL/AP1-2*. Th-*CAL/AP1-1* shows highest expression in sepals and ~10 times lower expression in petals. Expression of neither homolog was detectable in stamens, petals, gynoecia, capsules, roots, or leaves. Th-*PI-1*, the homolog of *PI* in *A. thaliana*, is collinear with the Brassicaceae gene order and is expressed mainly in petals and stamens, with less expression in younger and higher expression in older stages (see Supplemental Figure 27 online). The second *PI* homolog of *T. hassleriana*, Th-*PI-2*, is expressed at a much lower rate than Th-*PI-1*, ranging from around 50% of the Th-*PI-1* expression in stamens to only 10% of the Th-*PI-1* expression in late bud states.

The second two putative floral homeotic B-function genes, Th-*AP3-1* and Th-*AP3-2*, are highly similar in coding, 3′, and 5′ untranslated region sequence, and both are homologous to the *AP3* gene in *A. thaliana*. Th-*AP3-1* is expressed throughout the observed stages of bud development. In petals and stamens at anthesis, it is the most highly expressed gene of the MADS box genes analyzed (see Supplemental Figure 27 online). In petals, it is expressed at~200% and in stamens around 400% higher level than Th-*CAL*, Th-*PI-1*, and Th-*AP3-2*. Expression of Th-*AP3-2* in buds is around one-third lower than that of Th-*AP3-1*, and differential expression between both genes is detected in petals and stamens. While Th-*AP3-1* shows higher expression in stamens than in petals, Th-*AP3-2* has stronger expression in petals than in stamens (see Supplemental Figure 27 online). The divergence in expression of the B-class genes, along with the aforementioned gene transpositions, is indicative of the likely role in B-class gene functional differentiation and the regulation of the different floral morphologies between families. We acknowledge that these comparative analyses of *Tarenaya* to Brassicaceae are based on the analysis of the draft genome sequence of a single domesticated accession of one species; thus, future comparative analyses to other Cleomaceae species is needed for validation.

Evolution of the Brassicaceae SI Locus

Many Brassicaceae species possess a pollen-pistil recognition system that confers SI through the rejection of self-pollen (Boyes et al., 1997). This system is based on the interaction of the stigmatically expressed S-receptor kinase (encoded by *SRK*) with a small polymorphic peptide that is coded by the S-locus Cys-rich protein (*SCR*) gene, both located on the so-called S-locus of the genome. Many *SCR* alleles are likely derived from (partial) duplication and/or gene conversion from *SRK* alleles (Koornneef and Meinke, 2010; Guo et al., 2011). The cysteine-rich protein kinase (*CRK*) genes and *ARK* genes that belong to the S-locus are part of a larger family of receptor-like protein-kinases (*RLK*) genes, which have been shown to be mostly involved in oxidative stress and pathogen response (Chen et al., 2004; Wrzaczek et al., 2010). It should be noted that most ecotypes of *A. thaliana* are self-compatible due to pseudogenization of the *SRK* and/or *SCR* genes, a pattern seen in other self-compatible crucifers, but still an exception among Brassicaceae (Nasrallah et al., 2004). *T. hassleriana* does not possess a SI system, but it still contains an S-like locus that contains

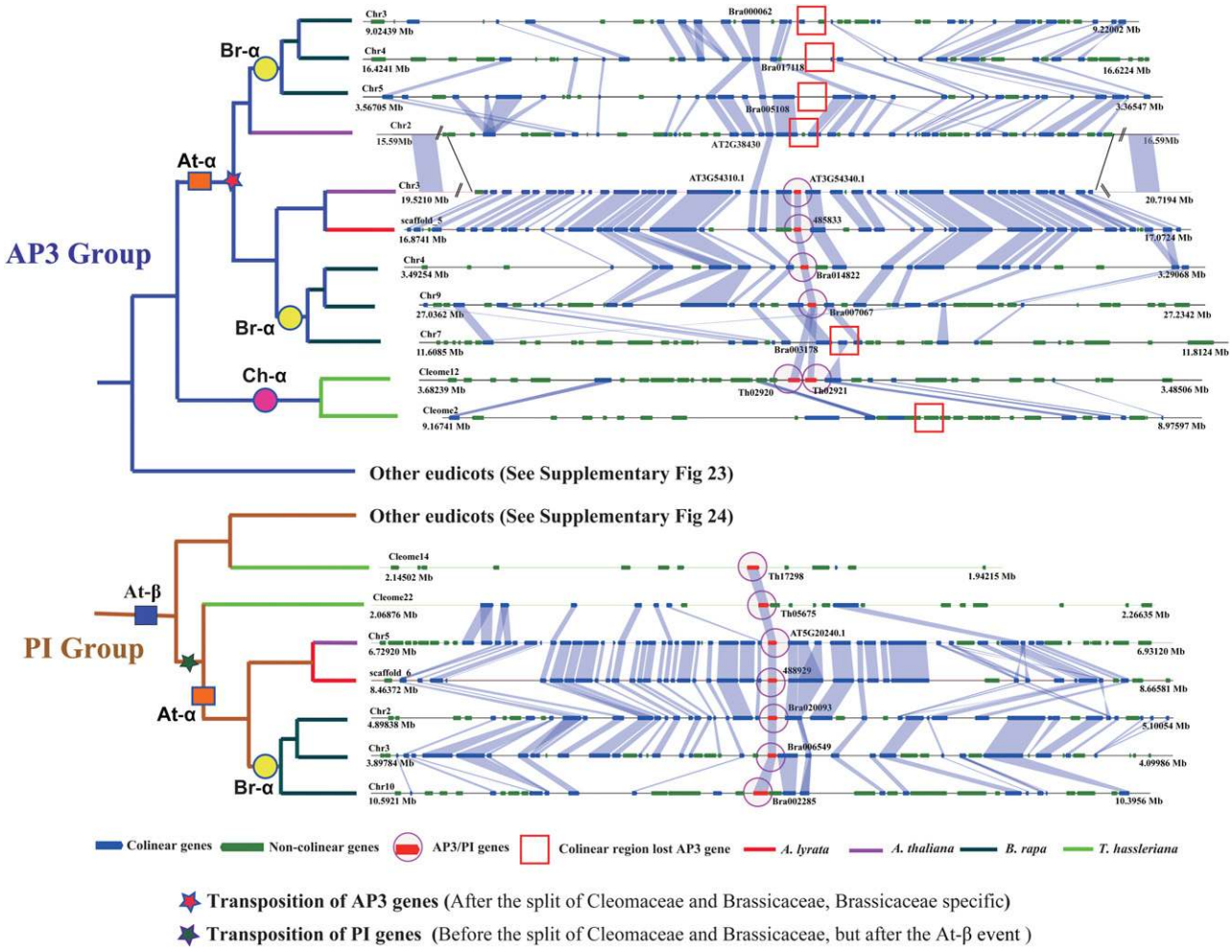

Figure 5. Collinearity Analysis of B-Class Type II MADS Box Gene (*AP3* and *PI*) Homologs Reveals Unusual Patterns of Gene Loss, Lineage-Specific Transpositions, and Local Tandem Duplications.

The placement of ancient polyploid events giving rise to gene duplicates is shown on appropriate nodes (At-β, At-α, Br-α, and Th-α). For the *AP3* group genes in Brassicaceae (shown by red bars), there is only a single locus retained in *A. thaliana* and *A. lyrata* and two retained Brassica syntelogs derived from Br-α. Collinear homoeologous regions derived from At-α are detectable in Brassicaceae genomes; however, the *AP3* syntelogs were lost (regions highlighted in red boxes). *T. hassleriana* has an unusual tandem duplication of *AP3* genes in one of two homoeologous regions derived from Th-α. The *AP3* genes and the neighboring Forkhead gene (*EMB1967*) are the only genes syntenic to the AP3 Brassicaceae region (see Supplemental Figure 25 online). The Cleomaceae AP3 region is syntenic with AP3 regions of all other eudicot genomes analyzed (see Supplemental Figure 23 online). Thus, we conclude that there was a lineage-specific transposition of AP3 and the neighboring Forkhead locus in the Brassicaceae. There is only a single copy of PI genes (red bars) in *A. thaliana* and *A. lyrata* and all three Br-α derived syntelogs in Brassicaceae. There is no detectable homoeologous region in Brassicaceae derived from At-α. In *Tarenaya*, we detected one syntenic *PI* gene and region to the Brassicaceae, but also a second region that is syntenic to other eudicots (see Supplemental Figure 24 online). We conclude that these two *Tarenaya PI* genes were generated due to the At-β ancient duplication event with the subsequent transposition of one locus into the region collinear between Brassicaceae and Cleomaceae and loss of the nontransposed locus from only the Brassicaceae lineage. The differences in genomic context and gene expression (see Supplemental Figure 27 online) may contribute to shifts in floral morphology and symmetry between families.

functional genes, and it is likely that this part of the genome is close to the ancestral state of this locus for Brassicaceae.

SRK and *ARK* on the S-locus are characterized by specific variations on the following protein domain compositions: B_lectin (B), S_locus_glyco (S), PAN-2 (Pa), Pkinase_Tyr (Pk), and Duf3403 (D1) and/or DUF3660 (D2) (Zhang et al., 2011). Using the Pfam database (Punta et al., 2012), we found that the S locus region in *C. rubella*, *B. rapa*, *A. lyrata*, and *A. thaliana* as well as the homologous region in *T. hassleriana* mostly contains *SRK* genes with a B-S-Pa-D1-Pk-D2 protein domain structure, followed by the B-S-Pa-Pk-D1/2 protein domain structure, which is more common across all SRK families (Figure 6; see Supplemental Figure 28 online). Three syntenic regions containing most of the genes of the S-locus were found in *Tarenaya*. One of these

contained a gene with the exact B-S-Pa-Pk-D1/2 protein domain structure: Th11131. Our analysis of domain structure further found that another gene, Th22785, had a B-S-Pa-Pk protein domain structure, an architecture shared with many angiosperm genes but not with SI-specific SRK alleles. We have also found two homologs of the SI-modifier gene, *Pub8*, in two of the three *Tarenaya* syntenic regions. One of the homologs, Th22784, is adjacent to the Th22785 locus. The other homolog, Th25331, is contained in the syntenic region that completely lacks any S_locus_glyco Pfam-containing proteins. Interestingly, we do not find a *Pub8* homolog in close proximity to Th11131. However, based on the alignment of all three regions, we can assume that the single-copy ancestral region contained homologs to *Pub8*, *ARK3*, and *B120* (Figure 6). To ensure that syntenic genes that were not found were not due to gaps in our assembly coverage, we manually confirmed that all relevant regions had no large gaps. In Supplemental Table 17 online, a summary of the coverage in these regions can be found. No large gaps were present except for one 2.5-kb gap in the region around Th22784. Based on the syntenic order of S-locus genes in other species, it is extremely unlikely that a gene is present in this gap, let alone any important candidate gene for this region.

We conclude that the syntenic *Tarenaya* gene Th11131, which is most closely related to *ARK3* with which it shares protein architecture, is similar to the ancestral version of SI locus genes in Brassicaceae. We hypothesize that the origin of the S-locus is due to a local rearrangement/duplication of an *ARK3*-like gene and subsequent expansion and diversification of the S-locus. Since *ARK3* in *Arabidopsis* is not expressed in the stigma, the regulatory domains needed for tissue-specific expression may potentially be derived from the concurrent duplication of elements from neighboring genes.

DISCUSSION

The completed genome of *A. thaliana* was a major milestone in plant biology and has provided a key tool for understanding plant gene function and genome organization (Arabidopsis Genome Initiative, 2000). Subsequently, there has been a great effort and interest to sequence other crucifer species to leverage the knowledge gained from *Arabidopsis* to other species in a comparative context. To date, there are more completed crucifer genomes (including the crop *B. rapa*; Wang et al., 2011) than any

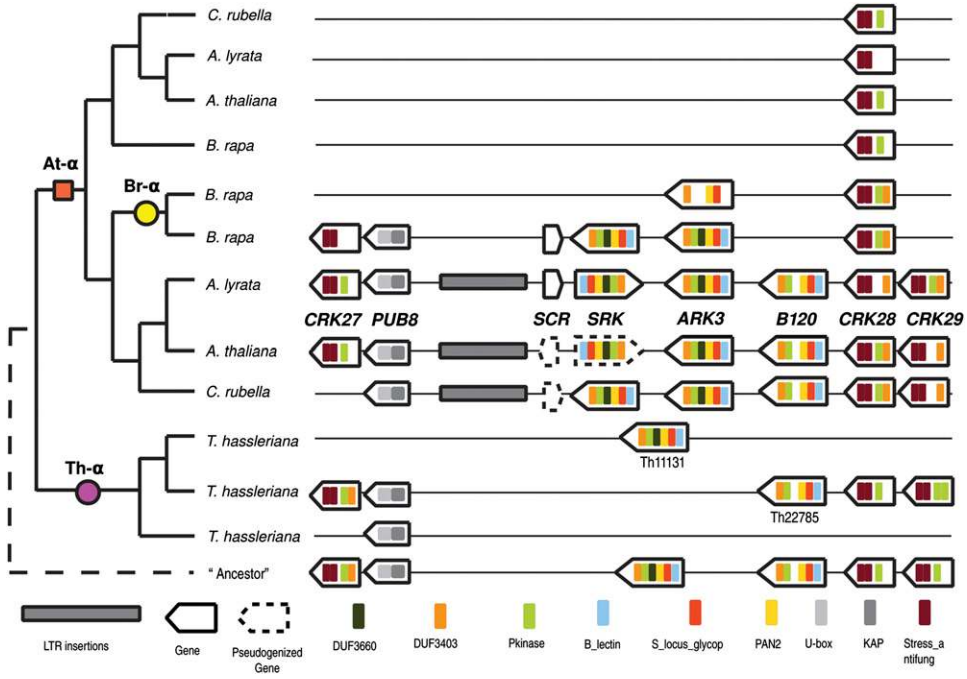

Figure 6. Synteny and Protein Domain Analysis of the Brassicaceae SI-Like Regions with Cleomaceae and Inference of the Ancestral Genomic Region.

All regions presented here are drawn in more detail (including genetic coordinates) in Supplemental Figure 28 online. Genes are marked by block arrows and color coded according to their protein domain composition as listed in the legend. In the case of pseudogenes, the block arrows have a dashed border. To find protein domain composition in pseudogenes, the longest open reading frame was translated in silico (see Methods). Gene orientation is shown by block arrows pointing left for gene orientation toward the 5′ end and pointing right for gene orientation toward the 3′end. The At-α duplication and the Br-α and Th-α triplications have been marked on the tree with an orange box (At-α) and yellow and purple circles (Br-α and Th-α, respectively). Each branch corresponds to a subgenome resulting from such a polyploid event. Theoretically, *B. rapa* should have six subgenomes, but only the regions showing synteny are listed here for clarity. The bottom branch represents a hypothetical layout of this genome region in the common ancestor of these species before the At-α, Br-α, and Th-α polyploid events. From our results, we conclude that an *ARK3*-like gene underwent a Brassicaceae-specific tandem gene duplication generating the key SI receptor *SRK*.

other eudicot plant family, and ambitious plans exist to sequence many more (such as the Brassica Map Alignment Project; Pires et al., 2013). We sequenced *T. hassleriana*, a phylogenetic outgroup of the Brassicaceae sister family, the Cleomaceae. We have shown that the vast majority of genes in *Tarenaya* have clear homologs within Brassicaceae. We provide several examples of how this sister-group genome can be used to elucidate patterns of gene, genome, and trait evolution within the Brassicaceae. Specifically, we focused on independent ancient polyploidy events, floral MADS box, and SI gene evolution. The genome of *T. hassleriana* will facilitate future research into the evolutionary and functional history of *Arabidopsis* genes and pathways.

While it has long been known that there are numerous recent polyploid plants (Jiao et al., 2011), with the arrival of the genomics era, it has become clear that there also is extensive evidence for ancient polyploidy across the tree of life (Kasahara, 2007; Jiao et al., 2011; Murat et al., 2012). Most ancient plant polyploid events that have been identified are ancient tetraploidy events (duplications such as At-α), but there are at least four published genome analyses of ancient plant hexaploidy (triplication events): at the base of the eudicots (At-γ) (Jaillon et al., 2007), in tomato (T) (2012), in Brassica (Br-α) (Wang et al., 2011), and this report of the ancient genome triplication in the *Tarenaya* (Tr-α). The Br-α and Tr-α events are of approximately the same age, allowing us to contrast independent ancient hexaploidy events from closely related lineages. The analysis of the retention of replicated genes (syntelogs) of the Type II MADS box genes provides a compelling example of what can be deduced by the comparison of these independent ancient triplications. We found that *Brassica* retains nearly twice as many Type-II MADS box genes as does *Tarenaya*. Genes retained after polyploidy often are dosage-sensitive gene complexes, whereby interacting partners must be maintained in the proper ratios (Edger and Pires, 2009). Considering the wealth of phenotypic diversity seen in the *Brassica* crops, we hypothesize that this great enrichment of morphological regulators in Brassica derived from Br-α may play a significant role. Also, the recently released genome of *L. alabamica* of the Brassicaceae also revealed an ancient hexaploidy (La-α) (Haudry et al. 2013). Future comparisons of the closely related La-α, Br-α, and Th-α events should be fruitful.

Comparative analyses can also be used to identify important gene transposition and deletion events. Type-II MADS box genes are remarkably resistant to gene transpositions and thus their collinearity is highly conserved across all angiosperms (Type-I MADS box genes are highly prone to transposition, but their functions are less known). When comparing *Tarenaya* to Brassicaceae, we found almost all Type-II MADS box genes are collinear, except for the B-class homologs of *AP3* and *PI*. Specifically, we established that there was a Brassicaceae-specific transposition of *AP3* and a rare tandem duplication of a floral MADS box gene of *AP3* homologs in *Tarenaya*. The transposition of *AP3* also involved a neighboring Forkhead gene, which may be coregulated and important for *AP3* function. We further demonstrate that *Tarenaya* has two homologs of *PI*: one that is syntenic with other eudicots and for which the locus is lost from Brassicaceae and one that is syntenic with Brassicaceae *PI* homologs. Both *PI* and *AP3* in *Tarenaya* have diverged in expression levels. MADS box B-class gene homologs in the *AP3* lineage as well as *TCP* members have been implicated in the establishment of monosymmetric flowers in monocots, including Orchidaceae species (Tsai et al., 2004, 2008; Mondragón-Palomino and Theissen, 2009; Bartlett and Specht, 2010; Preston and Hileman, 2012), *PI* genes have contributed to floral diversification in Asterids (Viaene et al., 2009), and B-class genes have contributed to *Aquilegia* floral diversification (Kramer et al., 2007). Thus, it is possible that both B-class and TCP genes have an impact on floral monosymmetry in Cleomaceae. Furthermore, B-class gene diversification has also been implicated in regulating floral gender shifts (Ackerman et al., 2008) and could similarly have diversified in Cleomaceae.

By comparing Brassicaceae genomes with the *Tarenaya* genome, we established that *SRK* in the S-locus occurred via a local rearrangement/duplication of an *ARK3*-like gene and subsequent expansion and diversification in Brassicaceae. In *Tarenaya*, we can clearly identify syntenic regions that contain homologous functional genes, including an *ARK3* and *CRK* homologs. The exact function of *ARK3* is not known, but based on gene expression analysis, this gene is thought to function during development of the sporophyte, perhaps in processes related to organ maturation, the establishment of growth pattern transitions (Dwyer et al., 1994), and/or involvement in pathogen responses (Pastuglia et al., 2002). However, it is not expressed in the stigma. Further research on Brassicaceae and Cleomaceae *ARK3* homologs is needed to establish the function of these genes. Interestingly, while Cleomaceae species do not have a SI system, many species, including *T. hassleriana*, are polygamous (trimonoecious) and can have flowers on the same inflorescence with different genders: male sterile, female sterile, or complete (Stout, 1923; Cruden and Lloyd, 1995; Machado et al., 2006), providing an alternative mechanism to reduce inbreeding.

Our results demonstrate the utility of the *Tarenaya* genome in complementing Brassicaceae genetic research in efforts to understand the function and evolution of genes and traits. The *Tarenaya* genome sequence will also pave the way for further studies of Cleomaceae traits not found in Brassicaceae, such as the evolution of C_4 photosynthesis (Brown et al., 2005; Marshall et al., 2007).

METHODS

Sample Preparation, Library Construction, Genome Sequencing, and Assembly

The purple-flowered *Tarenaya hassleriana* (Purple Queen) line selected for sequencing (ES1100) was first inbred by hand-pollination and floral bagging for four generations. Earlier generations of this line have previously been used for both BAC library construction and limited BAC sequencing (Schranz and Mitchell-Olds, 2006) and transcriptome sequencing and analysis (Barker et al., 2009); however, the material was referred to as being from *Cleome spinosa*. The two species are morphologically very similar, with only slight differences in stem-spine morphology, pubescence of sepals, ovary and capsules, and flower color; *C. spinosa* has only white flowers and the sepal and ovary are glandular-pubescent, whereas *T. hassleriana* can be white, pink, or purple and the sepals and ovary are glabrous (have no pubescence). Thus, many

commercial seed providers erroneously label their *T. hassleriana* material as *C. spinosa*.

We extracted DNA from leaves of *T. hassleriana* and constructed seven pair-end libraries with insert sizes of 350 bp, 500 bp, 800 bp, 2 kb, 5 kb, 10 kb, and 20 kb. Illumina Hisequation 2000 was then applied to sequence those DNA libraries, and 70.22-Gb raw data were generated. Low-quality reads, reads with adaptor sequences, and duplicated reads were filtered, and the remaining high quality data were used in the assembly. SOAPdenovo2.21 was applied to assemble the genome in the procedure of contig construction, scaffold construction, and gap closure. After gap closure, the assembly was broken down into contigs again according to the position of Ns in the assembly.

WGP and Physical Map Construction

The BAC library was prepared using leaf material of *T. hassleriana*. Two BAC libraries were subsequently generated: the first using *Hin*dIII (CLEH library) and the second using *Eco*RI (CLEE library). Average insert sizes were 145 kb for the CLEH library and 130 kb for the CLEE library. The vector used for library construction was pCC1BAC (Epicentre). For each library, 9600 clones were picked and arrayed into 384 well plates. Together, the two libraries equal ~8.8 genome equivalents (4.6 GE CLEH library and 4.2 GE CLEE library) at an estimated haploid genome size of 300 Mb. WGP was performed according to the methods detailed by van Oeveren et al. (2011). The resulting FPC map was used in further analysis.

FPC Map Assembly and Integration with de Novo Assembled Scaffolds

Sequence-based physical BAC maps were assembled using an improved version of FPC software (Keygene), capable of processing sequence-based BAC fingerprint (WGP) data instead of fragment mobility information as used in the original FPC software (Soderlund et al., 1997). The scaffolds from the SOAPdenovo assembly were then mapped to the physical contigs using nucleotide BLAST (Altschul et al., 1997). Hits were used only when they had a 100% identity match to the anchors. Subsequent filtering was performed to eliminate anchors with multiple hits and to establish superscaffold strand direction. The scaffolds were then ordered according to the mapped anchors and reassembled into superscaffolds.

RNA-Seq

A mixed sample from five tissues (buds, leaves, petioles, stems, and flowers) from *Tarenaya* flowering plants was used to isolate RNA. Total RNA was extracted using Trizol (Invitrogen). The isolated RNA was then treated with RNase-Free DNase and then subsequently with an Illumina mRNA-Seq Prep Kit, following the manufacturer's instructions. The insert size of the RNA libraries was ~200 bp, and the sequencing was done using Illumina GA II. Raw reads were filtered if there were adaptor contaminations and low quality (>10% bases with unknown quality). After filtering, all RNA reads were mapped back to the reference genome using Tophat Version 1.3.3 (Trapnell et al., 2009), implemented with bowtie66 version 0.12.7 (Langmead, 2010), and the transcripts were assembled according to the genome using Cufflinks (version 1.1.0) (Trapnell et al., 2012). Single libraries from eight different tissues were isolated with the Qiagen RNeasy plant mini kit and treated with RNase-free DNase. Illumina TruSeq Libraries were constructed according to the manufacturer's suggestions and sequenced. Raw reads were filtered to remove adaptors and low-quality bases and mapped to the predicted coding sequences using Cufflinks.

Plant Genome Sources

In all analyses where plant genomes are used, source database and version information can be found in Supplemental Table 18 online.

Gene Annotation

Gene models were predicted following several steps: (1) De novo gene prediction. De novo predictions were performed on repeat masked genome assembly. AUGUSTUS (version 2.03) (Stanke and Morgenstern, 2005), GlimmerHMM (version 3.02) (Majoros et al., 2004), and SNAP (version 2.0) were used to perform the de novo annotation. (2) Homology gene prediction. The protein sequences from *Arabidopsis thaliana*, *Brassica rapa*, *Carica papaya*, *Glycine max*, Theobroma cacao, and *Vitis vinifera* were mapped to the *Tarenaya* genome using tBLASTn, by an E-value cutoff of 10^{-5}, and then Genewise (version 2.2.0) (Birney et al., 2004) was used for gene annotation. (3) RNA aided annotation. All the RNA reads were mapped back to the reference genome by Tophat (version 1.0.14) (Trapnell et al., 2009), implemented with bowtie version 0.12.5, and the transcripts were assembled according to the genome using Cufflinks (version 0.8.2) (Trapnell et al., 2012). All the predictions were combined using GLEAN to produce the consensus gene sets.

The tRNA genes were identified by tRNAscan-SE (Lowe and Eddy, 1997). For rRNA identification, the *Arabidopsis* rRNA sequences were first downloaded from the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/nuccore). Then, rRNAs in the database were aligned against the *Tarenaya* genome using BLASTn to identify possible rRNAs. Other noncoding RNAs, including microRNA and small nuclear RNA, were identified using INFERNAL (Nawrocki et al., 2009) by searching against the Rfam database.

Gene Family Clustering

OrthoMCL (version 1.4) (Li et al., 2003) was used with default parameters followed by an all-versus-all BLASTP (E-value $\leq$ 1e-5) process using the protein sequence data sets from six plant species. Splice variants were removed from the data set (the longest protein sequence prediction was usually maintained), and the internal stop codons and incompatible reading frames were filtered. All gene families were classified according to the presence or absence of genes for specific species, and it was determined which gene families were species specific or genus specific.

A total of 184204 sequences from *A. thaliana*, *Arabidopsis lyrata*, *B. rapa*, *C. papaya*, *Vitis vinifera*, and *T. hassleriana* were clustered into 24,591 gene families. Of these, 9395 contained sequences from all six genomes, 2492 from Brassicaceae (*A. thaliana*, *A. lyrata*, and *B. rapa*), 1176 from plants as outgroups only bearing the At-γ event (*C. papaya* and *V. vinifera*), and 748 clusters were specific to *Tarenaya*. Of the 28,917 protein-coding genes predicted for *Tarenaya*, 22,482 were clustered in a total of 14,505 groups. The 748 *Tarenaya*-specific clusters contained 1556 genes, of which 529 have at least one INTERPRO domain. Singletons make up a total of 6435 genes, of which 2926 have at least one INTERPRO domain. Interestingly, many gene families that have decreased in number of genes in *Tarenaya* show bigger genes than others, containing more exons and more TE insertions. However, more Gene Ontology annotations are enriched for these contraction gene families. These results indicated that these contraction gene families may be functionally constrained.

Repeat Annotation

Repeats of the *Tarenaya* genome were identified by a combination of homology-based and de novo approaches. In the homology-based method, databases of known repetitive sequences were used to search against the genome assembly. In this way, RepeatMasker-3.2.9 and RepeatProteinMask software (Chen, 2004) were used to build the homology database and search the repeat sequence.

Furthermore, in the de novo approach, three de novo software packages (Piler-DF-1.0 [Edgar and Myers, 2005], RepeatScout-1.0.5, and LTR-FINDER-1.0.5 [Xu and Wang, 2007]) were used to build a de novo repeat database of the *Tarenaya* genome. RepeatMasker was then used

to identify repeats using both the repeat database that we built in-house and Repbase. Finally, the de novo prediction and the homolog prediction of TEs according to the position in the genome were combined.

Phylogenetic Analysis and Species Divergence Time Estimation

The maximum likelihood phylogenetic tree of *T. hassleriana* and other plant genomes was constructed using whole-genome fourfold degenerate sites among species. *Oryza sativa* and *Sorghum bicolor* were taken as the monocot outgroups. The following steps were taken: First, all of the single-copy gene families were extracted from the OrthoMCL clustering results. Second, multiple sequence alignments were run for each single copy gene family using the protein-coding sequences. Third, for each aligned gene family, the CDS back-translation of the protein multiple alignments was performed from the original DNA sequences using in-house Perl scripts, and then the fourfold degenerate sites of orthologous genes in all single-copy gene families (concatenated into one supergene for each species) were extracted. The branch length represents the neutral divergence rate. The substitution model (GTR+gamma+I) and Mrbayes (Ronquist et al., 2012) were used to reconstruct the phylogenetic tree.

Synteny and Collinearity Analysis

First, a homology search within and between species was performed using BLASTP (E-value threshold 1e-7, top 20 hits). Tandem gene families and weak matches were removed using in-house Perl scripts for further analysis. Tandem gene families were defined as clusters of genes within 10 intervening genes from one another, and the longest model was maintained to represent each family. For the weak matches, only the top BLAST hits were retained by applying a C-score threshold of 0.8 [C-score(A, B) = score (A, B)/max(best score of A, best score of B)] (Putnam et al., 2007).

Then, based on these filtered BLAST results, the whole genome-wide sequence alignments within and between genomes using genes as anchors, which was to search syntenic blocks, were conducted by an in-house pipeline implementing Dynamic Programming (parameters: score_of_match, 50; penalty of mismatch, −5; penalty of indel, −5; penalty_of_extension_indel, −2; block_size, ≥5 gene pairs; gap_between_neighbor_blocks, 30 genes). The running time for each whole-genome sequence alignment using genes as anchors by Dynamic programming was ~5 to ~6 h. At the same time, i-Adhore 3.0 (Proost et al., 2012) (http://bioinformatics.psb.ugent.be/software) was used to identify syntenic and collinearity blocks (gap_size = 30, cluster_gap = 35, q_value = 0.75, prob_cutoff = 0.01, anchor_points = 5, alignment_method = gg4, level_2_only = false, table_type = family) within and between genomes, and all the syntenic blocks identified by i-adhore were contained in our results identified by the Dynamic programming method; however, the latter method is more sensitive and accurate. All the dot plot figures were plotted using a Support Vector Graphics package implemented perl scripts (see Supplemental Figures 1, 2, 6 to 9, and 11 to 14 online).

Ancestral Genome Reconstruction

Based on the paralogous duplicates within each genome, four minimized genomes were independently created for *A. thaliana*, *A. lyrata*, *B. rapa*, and *T. hassleriana*, by condensing local duplications to one gene, removing transposons, and including only genes within blocks defined by retained pairs. Each of the minimized genomes represents the ancestral state and predate the recent polyploidy event. At the same time, these four ancestral state genomes were compared with the Ken Wolfe's 45 ancestral blocks of *A. thaliana* (see Supplemental Figures 11 to 14 online).

Partition of the *T. hassleriana* Genome into Three Subgenomes Following the Recent Polyploidy Event

To avoid the potentially confounding results with the independent ancient polyploidy events, the ancestor genome of *A. lyrata* (A ancestor) was used as the reference, and collinear blocks were identified using i-adhore 3.0 by aligning the four proteomes (*A. thaliana*, *A. lyrata*, *B. rapa*, and *T. hassleriana*) against the A ancestor genome. From the last common ancestor (the A ancestor represents this genome state) of these four species, both *A. thaliana* and *A. lyrata* experienced a whole-genome duplication event (At-α), *B. rapa* experienced one whole-genome duplication event (At-α) and an additional whole-genome triplication event (Br-α), and *T. hassleriana* experienced an independently whole-genome triplication event (Th-α). Therefore, quote ratios of 2:1, 2:1, 3/4/5/6:1, and 3:1 were observed for *A. thaliana*, *A. lyrata*, *B. rapa*, and *T. hassleriana* genomes. For *T. hassleriana*, the triplicated blocks identified were chained into three subgenomes compared with the A ancestor using dynamic programming. The main criteria are that the chained triplicated blocks are (1) nonoverlapping in the *T. hassleriana* genome; (2) have no more than 10% overlap between their orthologous A ancestor regions (results were similar using 0% overlap in A ancestor); and (3) maximize coverage of the *T. hassleriana* genome (annotated gene space). A similar strategy was taken for other genomes based on their polyploidy level from the last recent common ancestor, as shown in Supplemental Figures 16 to 19 online. For *T. hassleriana*, a total of 16,770 (63.2%) *Tarenaya* genes are in the triplicated blocks compared with the A ancestor, 688 (2.6%) genes are in the duplicated blocks, which indicate that another copy was possibly lost after the triplication event, and only 135 (0.5%) have one syntenic ortholog, which means that a very small subset of triplicated genes lost two copies. However, 8913 (33.6%) *Tarenaya* genes are not in any replicated blocks, which indicates species-specific deletion in *A. lyrata* or species-specific gains in *Tarenaya* after their divergence along evolutionary time.

Interproscan and Gene Functional Annotation

Interproscan version 4.5 was used to scan protein sequences against the protein signatures from InterPro (Hunter et al., 2009) (version 22.0) to infer functions for the protein-coding genes. This was done for the entire target proteomes involved in our main text analysis, including *A. thaliana*, *A. lyrata*, *T. hassleriana*, *B. rapa*, *Solanum lycopersicum*, *V. vinifera*, *Prunus persica*, *Populus trichocarpa*, and *C. papaya*. InterPro integrates protein families, domains, and functional sites from different databases: Pfam, PROSITE, PRINTS, ProDom, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, Gene3D, and PANTHER. Interproscan integrates the searching algorithms of all these databases. In total, Interproscan identified 93,038 protein domains of 4733 distinct domain types. Seventy-five percent of the genes (21,829 out of 28,917 genes in total) have been assigned with at least one domain.

Genomic Analysis for Reproductive Traits

For the syntenic and protein domain analysis of SI genes, the genes annotated in *A. thaliana*, *C. rubella*, and *A. lyrata* were used as published in an extensive study into S-locus variation in *Arabidopsis* species (Guo et al., 2011). Homologs were then sought using top BLAST hits of these genes against *T. hassleriana* and *B. rapa*. All homology candidates were confirmed by manual inspection of the alignment in dot plots generated in MAFFT (Katoh et al., 2002) to confirm synteny of candidate regions. After compiling a definitive list of syntenic regions, the genes from these regions were analyzed through the PFAM online protein domain analysis program (Punta et al., 2012). Figure 6 was manually compiled from the results of this program.

qRT-PCR

For the qRT-PCR, total RNA was isolated from roots, leaves, three bud stages (1 to 5 mm, 5 to 10 mm, and 10 to 25 mm length), sepals, petals, stamens, carpels at anthesis, and three stages of siliques (10 mm, 10 to 30 mm, and 30 to 50 mm length) with the GeneJET plant RNA purification

mini kit (Fermentas). First-strand cDNA was synthesized using 500 ng total RNA with the RevertAid H Minus First Strand cDNA synthesis kit (Fermentas) using random hexamer primers.

The qRT-PCR experiments were performed according to the MIQE guidelines (Bustin et al., 2009). Exon spanning primers were generated using PerlPrimer 1.1.21. (Marshall, 2004). A primer efficiency test was performed and the primers were tested with genomic DNA to ensure cDNA specificity. Standard dose response curves were constructed for all genes using serial dilutions (1:50 to 1:50,000) of 10- to 25-mm-long bud cDNA template to calculate amplification efficiency. The qRT-PCR assay was performed with the LightCycler 480 II (Roche) and the data analyzed with the LCS480 1.5.0.39 software. Each reaction was composed of 10 μL of 2× DyNAmo Flash SYBR Green Mastermix (Biozym Scientific), 2 μL each of 10 μM forward and reverse primers, 1 μL water, and 5 μL of 1:100 diluted template cDNA. Each reaction was performed in biological duplicate and technical triplicate along with water and RNA controls for each primer pair. The *GLYCERALDEHYDE-3-PHOSPHATE DEHYDROGENASE C SUBUNIT* and *ELONGATION FACTOR 1-ALPHA* genes served as internal controls. The following PCR program was used: 7 min at 95°C; 45 cycles of 10 s at 95°C, 15 s at 60°C, and 15 s at 72°C, followed by a melting curve of 5 s at 95°C, 1 min at 65°C, and 30 s at 97°C. The quantification cycles (C_q) were calculated according to the second derivative maximum algorithm. The raw C_q data were analyzed according to the Comparitive C_q method ($\Delta\Delta C_q$) (Schmittgen and Livak, 2008). Gene expression was first normalized relative to the expression of the two reference genes in the respective tissues. The expression was further normalized with the expression of the reference genes in 10- to 25-mm long buds, which acted as an interassay calibrator. The relative expression was then calculated with reference to the expression of *T. hassleriana CAL* in stage 3 buds.

Accession Numbers

Sequence data from this article can be found in the Arabidopsis Genome Initiative or GenBank/EMBL databases under the following accession numbers: Th-*CAL/AP1-1*, Th01189; Th-*CAL/AP1-2*, Th13754; Th-*PI-1*, Th05675; Th-*PI-2*, Th17298; Th-*AP3-1*, Th02920; Th-*AP3-2*, Th02921; *EMB196*, AT3G54350; and *TCP1*, AT1G67260. The genomic reads of *T. hassleriana*, as well as RNA sequencing data, have been deposited into the NCBI Short Read Archive under accession numbers SRA058749 and GSM1008474. The information for the raw reads data can be found in Supplemental Table 1 online. The genome sequence and annotation data set have been deposited into NCBI (project ID PRJNA175230 [superscaffolds]; the accession number is AOUI00000000).

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Evaluation of Superscaffold13 Integrated by the Combination of SOAPdenovo Assembled Scaffolds and the Physical Map as an Example of Our Verification Method.

Supplemental Figure 2. Comparison of the New *T. hassleriana* Assembly with Previously Sequenced BACs.

Supplemental Figure 3. Principal Component Analysis of Gene Expression.

Supplemental Figure 4. Comparison of the Distribution of Gene Features (Including mRNA Length, CDS Length, Exon Length, Intron Length) among Several Selected Angiosperms.

Supplemental Figure 5. Graph Representing Contig Length versus LTR (from Retrotransposons) Density in the *T. hassleriana* Assembly.

Supplemental Figure 6. Dot Plot Figure to Show Syntenic Duplicates within the *T. hassleriana* Genome.

Supplemental Figure 7. Dot Plot Figure to Show the Collinear Blocks between *A. thaliana* and *T. hassleriana* Genomes.

Supplemental Figure 8. Dot Plot Figure to Show the Collinear Blocks between *A. lyrata* and *T. hassleriana* Genomes.

Supplemental Figure 9. Dot Plot Figure to Show the Collinear Blocks between *B. rapa* and *T. hassleriana* Genomes.

Supplemental Figure 10. Homologous Blocks within and between Genomes Based on the Chromosome Layout of *A. lyrata*.

Supplemental Figure 11. Dot Plot Figure to Show the Corresponding Homologous Relationship between "Wolfe" Blocks and the 64 Ancestor Genome Blocks of *A. thaliana* Identified in This Article and Named "A Ancestor."

Supplemental Figure 12. Dot Plot Figure to Show the Corresponding Homologous Relationship between the "Wolfe" Blocks and the 61 Ancestor Genome Blocks of *A. lyrata* Identified in This Article, Forming the Alternative "A" Ancestor Used in Subgenome Analyses.

Supplemental Figure 13. Dot Plot Figure to Show the Corresponding Homologous Relationship between the "Wolfe" Blocks and the 71 Ancestor Genome Blocks of *B. rapa* Identified in This Article Named "B Ancestor."

Supplemental Figure 14. Dot Plot Figure to Show the Corresponding Homologous Relationship between "Wolfe" Blocks and the 87 Ancestor Genome Blocks of *Tarenaya hassleriana* Identified in This Article Named "C Ancestor."

Supplemental Figure 15. Multiple Homologous Relationships between the Ancestor Genome Blocks ("Wolfe" Ancestral Blocks, A Ancestral Blocks of *A. thaliana*, A Ancestral Blocks *of A. lyrata*, B Ancestral Blocks *of B. rapa*, C Ancestral Blocks of *T. hassleriana*) and the Reference Genome *A. thaliana*.

Supplemental Figure 16. The Genome of *A. lyrata* Is Partitioned into Two Subgenomes (Based on the At-α Event) by Comparing against the Reference Ancestor Genome of *A. lyrata*.

Supplemental Figure 17. The Genome of *A. thaliana* Is Partitioned into Two Subgenomes (Based on the At-α Event) by Comparing against the Reference Ancestor Genome of *A. lyrata*.

Supplemental Figure 18. The Genome of *B. rapa* Is Partitioned into Three Subgenomes (Based on the Br-α Event) by Comparing against the Reference Ancestor Genome of *A. lyrata*.

Supplemental Figure 19. The Genome of *T. hassleriana* Is Partitioned into Three Subgenomes (Based on the Ch-α Event) by Comparing against the Reference Ancestor Genome of *A. lyrata*.

Supplemental Figure 20. Complete ML Tree of MADS Type II Homologs in All Sequenced Angiosperms.

Supplemental Figure 21. *SHP1*, Two Gene Copies, and Synteny.

Supplemental Figure 22. *AP1/CAL* Gene Copies and Synteny.

Supplemental Figure 23. Synteny Plot of the *Tarenaya* Region Containing Th02920/Th02921 (*AP3* Genes).

Supplemental Figure 24. Synteny Plot of the *Tarenaya* Region Containing Th17298 (*PI* Gene).

Supplemental Figure 25. Simplified Synteny Plots of *AP3* and *PI* Genes in *Tarenaya*.

Supplemental Figure 26. *TCP1* Gene Copies and Synteny.

Supplemental Figure 27. Histogram of Gene Expression in Various *Tarenaya* Tissues.

Supplemental Figure 28. SI Locus Domain–Related Genes in Syntenic Block Context in Brassiaceae and *Tarenaya*.

Supplemental Table 1. Library Construction and Sequencing (Raw data).

Supplemental Table 2. Statistics of the Generation of the Clean Data.

Supplemental Table 3. Summary Statistics of the Preliminary Genome Assembly.

Supplemental Table 4. Evaluation of Gene Region Coverage by RNA Mapping.

Supplemental Table 5. Summary of Results after Deconvolution and Filtering of the Whole-Genome Profiling Tags.

Supplemental Table 6. FPC Assembly Results for the High, Reduced, and Low-Stringency WGP Assemblies.

Supplemental Table 7. Final Contig and Scaffold Median Lengths after Superscaffold Construction through Integration with the Physical Map.

Supplemental Table 8. Summary Statistics of RNA-Seq Sequencing Data and the Alignments Mapping to Genes and Genome.

Supplemental Table 9. Mapping Statistics and Transcript Dynamics for Eight RNA Samples.

Supplemental Table 10. Summary Statistics of the Results from Different Gene Annotation Strategies.

Supplemental Table 11. Summary of Functional Annotations Derived from Interproscan.

Supplemental Table 12. Summary of smRNA Annotation.

Supplemental Table 13. Comparison of *T. hassleriana* Assembly with Nine Other Sequenced Crucifers.

Supplemental Table 14. Classification of Transposable Elements within the *T. hassleriana* Genome.

Supplemental Table 15. Statistics of the Ancestor Genome Construction within *A. thaliana*, *A. lyrata*, *B. rapa*, and *T. hassleriana*.

Supplemental Table 16. Syntenic and Collinearity Analysis between *Tarenaya* and Other Species.

Supplemental Table 17. Assembly Coverage of S-Locus-Like Regions in *T. hassleriana*.

Supplemental Table 18. Plant Genome Source Databases.

Supplemental Data Set 1. List of Syntenic Order of *Tarenaya* Genes in the Ancestor Block Configuration of the Genome.

Supplemental Data Set 2. MADS Box Gene Alignment as Used for the ML Phylogenetic Tree in Supplemental Figure 20.

ACKNOWLEDGMENTS

This work was supported by following funding sources to Beijing Genomics Institute, Shenzhen: State Key Laboratory of Agricultural Genomics, Guangdong Provincial Key Laboratory of core collection of crop genetic resources research and application (2011A091000047), Shenzhen Engineering Laboratory of Crop Molecular design breeding, and National Natural Science Funds for Distinguished Young Scholar (30725008). E.v.d.B., J.H., and M.E.S. were supported by the Netherlands Organization for Scientific Research (NWO VIDI Grant 864.10.001 and NWO Ecogenomics Grant 844.10.006). K.K. wishes to thank the Alexander-von-Humboldt Foundation and the BMBF for support. S.d.B. was supported by an NWO Experimental Plant Sciences graduate school "master talent" fellowship. A.P.M.W. thanks the Deutsche Forschungsgemeinschaft for support (Grants WE 2231/9-1 and EXC 1028). We also thank the co-principal investigators of the Brassicales Map Alignment Project for their support (Rod Wing, J. Chris Pires, Thomas Mitchell-Olds, Detlef Weigel, and S. Stephen Wright).

AUTHOR CONTRIBUTIONS

M.E.S., G.Z., J.M.H., and X.Zhu. designed the project. G.Z., M.E.S., P.Z. and S.C. led the sequencing and analysis. C.S., Z.Z., W.L., M.L., Y.T., J.W., X.X, H.Z. and Z.Q. assisted with sequencing and analysis. J.C. and G.F. did the SOAPdenovo genome assembly. J.C. and X.Zhong. did the annotation. P.Z., S.C., E.v.d.B., M.E.S., and C.B. did the evolutionary analysis. S.C., J.X., E.v.d.B., and J.H. constructed the physical map. M.E.S., S.C., and E.v.d.B. did the reproductive trait evolution analysis. K.K. and S.d.B. did the phylogenetic analysis of MADS box genes. C.K., A.Bräutigam., and A.P.M.W. conducted tissue-specific transcriptomic analyses of *T. hassleriana*. A.Bräutigam. and A.Becker. did the qRT-PCR analysis of MADS box genes. J.C.H. did analysis of the *TCP* gene family. M.E.S., E.v.d.B., and S.C. wrote the article.

Received May 7, 2013; revised July 6, 2013; accepted August 6, 2013; published August 27, 2013.

REFERENCES

Ackerman, C.M., Yu, Q., Kim, S., Paull, R.E., Moore, P.H., and Ming, R. (2008). B-class MADS-box genes in trioecious papaya: Two paleoAP3 paralogs, CpTM6-1 and CpTM6-2, and a PI ortholog CpPI. Planta **227:** 741–753.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. Nucleic Acids Res. **25:** 3389–3402.

Arabidopsis Genome Initiative (2000). Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. Nature **408:** 796–815.

Ashburner, M., et al.; The Gene Ontology Consortium (2000). Gene ontology: Tool for the unification of biology. Nat. Genet. **25:** 25–29.

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 Biol. Evol. **1:** 391–399.

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.

Beilstein, M.A., Nagalingum, N.S., Clements, M.D., Manchester, S. R., and Mathews, S. (2010). Dated molecular phylogenies indicate a Miocene origin for *Arabidopsis thaliana*. Proc. Natl. Acad. Sci. USA **107:** 18724–18728.

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

Birney, E., Clamp, M., and Durbin, R. (2004). GeneWise and Genomewise. Genome Res. **14:** 988–995.

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.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003).

The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. **31:** 365–370.

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

Boyes, D.C., Nasrallah, M.E., Vrebalov, J., and Nasrallah, J.B. (1997). The self-incompatibility (S) haplotypes of *Brassica* contain highly divergent and rearranged sequences of ancient origin. Plant Cell **9:** 237–247.

Brown, N.J., Parsley, K., and Hibberd, J.M. (2005). The future of C4 research—Maize, *Flaveria* or *Cleome*? Trends Plant Sci. **10:** 215–221.

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

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

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

Bustin, S.A., Benes, V., Garson, J.A., Hellemans, J., Huggett, J., Kubista, M., Mueller, R., Nolan, T., Pfaffl, M.W., Shipley, G.L., Vandesompele, J., and Wittwer, C.T. (2009). The MIQE guidelines: Minimum information for publication of quantitative real-time PCR experiments. Clinical Chemistry **55:** 611–622.

Chen, K.G., Fan, B.F., Du, L.Q., and Chen, Z.X. (2004). Activation of hypersensitive cell death by pathogen-induced receptor-like protein kinases from *Arabidopsis*. Plant Mol. Biol. **56:** 271–283.

Chen, N. (2004). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr. Protoc. Bioinformatics **4:**10.

Colombo, M., Brambilla, V., Marcheselli, R., Caporali, E., Kater, M.M., and Colombo, L. (2010). A new role for the SHATTERPROOF genes during *Arabidopsis* gynoecium development. Dev. Biol. **337:** 294–302.

Couvreur, T.L.P., Franzke, A., Al-Shehbaz, I.A., Bakker, F.T., Koch, M.A., and Mummenhoff, K. (2010). Molecular phylogenetics, temporal diversification, and principles of evolution in the mustard family (Brassicaceae). Mol. Biol. Evol. **27:** 55–71.

Cruden, R.W., and Lloyd, R.M. (1995). Embryophytes have equivalent sexual phenotypes and breeding systems: Why not a common terminology to describe them? Am. J. Bot. **82:** 816–825.

Dassanayake, M., Oh, D.-H., Haas, J.S., Hernandez, A., Hong, H., Ali, S., Yun, D.-J., Bressan, R.A., Zhu, J.-K., Bohnert, H.J., and Cheeseman, J.M. (2011). The genome of the extremophile crucifer *Thellungiella parvula*. Nat. Genet. **43:** 913–918.

Dwyer, K.G., Kandasamy, M.K., Mahosky, D.I., Acciai, J., Kudish, B.I., Miller, J.E., Nasrallah, M.E., and Nasrallah, J.B. (1994). A superfamily of S locus-related sequences in *Arabidopsis*: Diverse structures and expression patterns. Plant Cell **6:** 1829–1843.

Edgar, R.C., and Myers, E.W. (2005). PILER: Identification and classification of genomic repeats. Bioinformatics **21** (suppl. 1)**:** i152–i158.

Edger, P.P., and Pires, J.C. (2009). Gene and genome duplications: The impact of dosage-sensitivity on the fate of nuclear genes. Chromosome Res. **17:** 699–717.

Endress, P.K. (1999). Symmetry in flowers: Diversity and evolution. Int. J. Plant Sci. **160** (S6)**:** S3–S23.

Engler, F.W., Hatfield, J., Nelson, W., and Soderlund, C.A. (2003). Locating sequence on FPC maps and selecting a minimal tiling path. Genome Res. **13:** 2152–2163.

Flagel, L.E., and Wendel, J.F. (2009). Gene duplication and evolutionary novelty in plants. New Phytol. **183:** 557–564.

Franzke, A., Lysak, M.A., Al-Shehbaz, I.A., Koch, M.A., and Mummenhoff, K. (2011). Cabbage family affairs: The evolutionary history of Brassicaceae. Trends Plant Sci. **16:** 108–116.

Guo, Y.L., Zhao, X., Lanz, C., and Weigel, D. (2011). Evolution of the S-locus region in Arabidopsis relatives. Plant Physiol. **157:** 937–946.

Hall, J.C., Sytsma, K.J., and Iltis, H.H. (2002). Phylogeny of Capparaceae and Brassicaceae based on chloroplast sequence data. Am. J. Bot. **89:** 1826–1842.

Haudry, A., et al. (2013). An atlas of over 90,000 conserved noncoding sequences provides insight into crucifer regulatory regions. Nat. Genet. **45:** 891–898.

Hu, T.T., et al. (2011). The *Arabidopsis lyrata* genome sequence and the basis of rapid genome size change. Nat. Genet. **43:** 476–481.

Hunter, S., et al. (2009). InterPro: The integrative protein signature database. Nucleic Acids Res. **37** (Database issue)**:** D211–D215.

Iltis, H.H., and Cochrane, T.S. (2007). Studies in the Cleomaceae V: A new genus and ten new combinations for the flora of North America. Novon **17:** 447–451.

Iltis, H.H., Hall, J.C., Cochrane, T.S., and Sytsma, K.J. (2011). Studies in the Cleomaceae I. On the separate recognition of Capparaceae, Cleomaceae, and Brassicaceae. Ann. Mo. Bot. Gard. **98:** 28–36.

Jabbour, F., Nadot, S., and Damerval, C. (2009). Evolution of floral symmetry: A state of the art. C. R. Biol. **332:** 219–231.

Jaillon, O., et al; French-Italian Public Consortium for Grapevine Genome Characterization (2007). The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. Nature **449:** 463–467.

Jiao, Y., et al. (2011). Ancestral polyploidy in seed plants and angiosperms. Nature **473:** 97–100.

Kanehisa, M., and Goto, S. (2000). KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. **28:** 27–30.

Kasahara, M. (2007). The 2R hypothesis: An update. Curr. Opin. Immunol. **19:** 547–552.

Katoh, K., Misawa, K., Kuma, K.i., and Miyata, T. (2002). MAFFT: A novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. **30:** 3059–3066.

Koornneef, M., and Meinke, D. (2010). The development of *Arabidopsis* as a model plant. Plant J. **61:** 909–921.

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.

Langmead, B. (2010). Aligning short sequencing reads with Bowtie. Curr. Protoc. Bioinformatics **11:** 7.

Li, L., and Stoeckert, C.J., Jr., and Roos, D.S. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. Genome Res. **13:** 2178–2189.

Li, R.Q., et al. (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Res. **20:** 265–272.

Litt, A., and Kramer, E.M. (2010). The ABC model and the diversification of floral organ identity. Semin. Cell Dev. Biol. **21:** 129–137.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. **25:** 955–964.

Machado, I., Cristina Lopes, A., Valentina Leite, A., and Virgíniade Brito Neves, C. (2006). *Cleome spinosa* (Capparaceae): Polygamodioecy and pollination by bats in urban and Caatinga areas, northeastern Brazil. Bot. Jahrb. Syst. Pflanzengesch. Pflanzengeogr. **127:** 69–82.

Majoros, W.H., Pertea, M., and Salzberg, S.L. (2004). TigrScan and GlimmerHMM: Two open source ab initio eukaryotic gene-finders. Bioinformatics **20:** 2878–2879.

Marquard, R.D., and Steinback, R. (2009). A model plant for a biology curriculum: Spider flower (*Cleome hasslerana* L.). Am. Biol. Teach. **71:** 235–244.

Marshall, D.M., Muhaidat, R., Brown, N.J., Liu, Z., Stanley, S., Griffiths, H., Sage, R.F., and Hibberd, J.M. (2007). Cleome, a genus closely related to *Arabidopsis*, contains species spanning a developmental progression from C(3) to C(4) photosynthesis. Plant J. **51:** 886–896.

Marshall, O.J. (2004). PerlPrimer: Cross-platform, graphical primer design for standard, bisulphite and real-time PCR. Bioinformatics **20:** 2471–2472.

McMillan, L.E.M., and Martin, A.C.R. (2008). Automatically extracting functionally equivalent proteins from SwissProt. BMC Bioinformatics **9:** 418.

Ming, R., et al. (2008). The draft genome of the transgenic tropical fruit tree papaya (*Carica papaya Linnaeus*). Nature **452:** 991–996.

Mondragón-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. (Lond.) **104:** 583–594.

Murat, F., Van de Peer, Y., and Salse, J. (2012). Decoding plant and animal genome plasticity from differential paleo-evolutionary patterns and processes. Genome Biol. Evol. **4:** 917–928.

Nasrallah, M.E., Liu, P., Sherman-Broyles, S., Boggs, N.A., and Nasrallah, J.B. (2004). Natural variation in expression of self-incompatibility in *Arabidopsis thaliana*: Implications for the evolution of selfing. Proc. Natl. Acad. Sci. USA **101:** 16070–16074.

Navarro-Quezada, A.R. (2007). Molecular Evolution of Tropinone-Reductase-Like and Tau GST Genes Duplicated in Tandem in Brassicaceae. (Munich, Germany: LMU Munich).

Nawrocki, E.P., Kolbe, D.L., and Eddy, S.R. (2009). Infernal 1.0: Inference of RNA alignments. Bioinformatics **25:** 1335–1337.

Parra, G., Bradnam, K., Ning, Z., Keane, T., and Korf, I. (2009). Assessing the gene space in draft genomes. Nucleic Acids Res. **37:** 289–297.

Pastuglia, M., Swarup, R., Rocher, A., Saindrenan, P., Roby, D., Dumas, C., and Cock, J.M. (2002). Comparison of the expression patterns of two small gene families of S gene family receptor kinase genes during the defence response in *Brassica oleracea* and *Arabidopsis thaliana*. Gene **282:** 215–225.

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. Int. J. Plant Sci. **172:** 352–365.

Preston, J.C., and Hileman, L.C. (2012). Parallel evolution of TCP and B-class genes in Commelinaceae flower bilateral symmetry. EvoDevo **3:** 6.

Proost, S., Fostier, J., De Witte, D., Dhoedt, B., Demeester, P., Van de Peer, Y., and Vandepoele, K. (2012). i-ADHoRe 3.0—Fast and sensitive detection of genomic homology in extremely large data sets. Nucleic Acids Res. **40:** e11.

Punta, M., et al. (2012). The Pfam protein families database. Nucleic Acids Res. **40** (Database issue)**:** D290–D301.

Putnam, N.H., et al. (2007). Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science **317:** 86–94.

Ronquist, F., Teslenko, M., van der Mark, P., Ayres, D.L., Darling, A., Höhna, S., Larget, B., Liu, L., Suchard, M.A., and Huelsenbeck, J.P. (2012). MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. Syst. Biol. **61:** 539–542.

Rosin, F.M., and Kramer, E.M. (2009). Old dogs, new tricks: Regulatory evolution in conserved genetic modules leads to novel morphologies in plants. Dev. Biol. **332:** 25–35.

Schmid, M., Davison, T.S., Henz, S.R., Pape, U.J., Demar, M., Vingron, M., Schölkopf, B., Weigel, D., and Lohmann, J.U. (2005). A gene expression map of *Arabidopsis thaliana* development. Nat. Genet. **37:** 501–506.

Schmittgen, T.D., and Livak, K.J. (2008). Analyzing real-time PCR data by the comparative C(T) method. Nat. Protoc. **3:** 1101–1108.

Schnable, J.C., Wang, X., Pires, J.C., and Freeling, M. (2012). Escape from preferential retention following repeated whole genome duplication in plants. Front. Plant Sci. **3:** 94.

Schranz, M.E., Lysak, M.A., and Mitchell-Olds, T. (2006). The ABC's of comparative genomics in the Brassicaceae: Building blocks of crucifer genomes. Trends Plant Sci. **11:** 535–542.

Schranz, M.E., and Mitchell-Olds, T. (2006). Independent ancient polyploidy events in the sister families Brassicaceae and Cleomaceae. Plant Cell **18:** 1152–1165.

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.

Shen, J.X., Fu, T.D., Yang, G.S., Ma, C.Z., and Tu, J.X. (2005). Genetic analysis of rapeseed self-incompatibility lines reveals significant heterosis of different patterns for yield and oil content traits. Plant Breed. **124:** 111–116.

Slotte, T., et al. (2013). The *Capsella rubella* genome and the genomic consequences of rapid mating system evolution. Nat. Genet. **45:** 831–835.

Smaczniak, C., et al. (2012). Characterization of MADS-domain transcription factor complexes in *Arabidopsis* flower development. Proc. Natl. Acad. Sci. USA **109:** 1560–1565.

Soderlund, C., Longden, I., and Mott, R. (1997). FPC: A system for building contigs from restriction fingerprinted clones. Comput. Appl. Biosci. **13:** 523–535.

Stanke, M., and Morgenstern, B. (2005). AUGUSTUS: A web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. **33** (Web Server issue)**:** W465–W467.

Stout, A.B. (1923). Alternation of sexes and intermittent production of fruit in the spider flower (*Cleome spinosa*). Am. J. Bot. **10:** 57–66.

Tang, H., Lyons, E., Pedersen, B., Schnable, J.C., Paterson, A.H., and Freeling, M. (2011). Screening synteny blocks in pairwise genome comparisons through integer programming. BMC Bioinformatics **12:** 102.

Tanksley, S.D. (2004). The genetic, developmental, and molecular bases of fruit size and shape variation in tomato. Plant Cell **16** (suppl.)**:** S181–S189.

Tomato Genome Consortium (2012). The tomato genome sequence provides insights into fleshy fruit evolution. Nature **485:** 635–641.

Theissen, G. (2001). Development of floral organ identity: Stories from the MADS house. Curr. Opin. Plant Biol. **4:** 75–85.

Thomas, B.C., Pedersen, B., and Freeling, M. (2006). Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. Genome Res. **16:** 934–946.

Trapnell, C., Pachter, L., and Salzberg, S.L. (2009). TopHat: Discovering splice junctions with RNA-Seq. Bioinformatics **25:** 1105–1111.

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D.R., Pimentel, H., Salzberg, S.L., Rinn, J.L., and Pachter, L. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat. Protoc. **7:** 562–578.

Tsai, W.C., Kuoh, C.S., Chuang, M.H., Chen, W.H., and Chen, H.H. (2004). Four DEF-like MADS box genes displayed distinct floral morphogenetic roles in *Phalaenopsis* orchid. Plant Cell Physiol. **45:** 831–844.

Tsai, W.C., Pan, Z.J., Hsiao, Y.Y., Jeng, M.F., Wu, T.F., Chen, W.H., and Chen, H.H. (2008). Interactions of B-class complex proteins involved in tepal development in *Phalaenopsis* orchid. Plant Cell Physiol. **49:** 814–824.

Tuskan, G.A., et al. (2006). The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science **313:** 1596–1604.

Van de Peer, Y., Maere, S., and Meyer, A. (2009). The evolutionary significance of ancient genome duplications. Nat. Rev. Genet. **10:** 725–732.

van Oeveren, J., de Ruiter, M., Jesse, T., van der Poel, H., Tang, J.F., Yalcin, F., Janssen, A., Volpin, H., Stormo, K.E., Bogden, R., van Eijk, M.J.T., and Prins, M. (2011). Sequence-based physical mapping of complex genomes by whole genome profiling. Genome Res. **21:** 618–625.

Vekemans, D., Proost, S., Vanneste, K., Coenen, H., Viaene, T., Ruelens, P., Maere, S., Van de Peer, Y., and Geuten, K. (2012). Gamma paleohexaploidy in the stem lineage of core eudicots: Significance for MADS-box gene and species diversification. Mol. Biol. Evol. **29:** 3793–3806.

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. Mol. Biol. Evol. **26:** 2627–2645.

Wang, X.W., et al; Brassica rapa Genome Sequencing Project Consortium (2011). The genome of the mesopolyploid crop species *Brassica rapa*. Nat. Genet. **43:** 1035–1039.

Wing, R.A., Mitchell-Olds, T., Pires, J.C., Schranz, M.E., Weigel, D., and Wright, S. (2013). Brassicales Map Alignment Project (BMAP). http://www.brassica.info/resource/sequencing/bmap.php. Accessed August 22, 2013.

Wrzaczek, M., Brosché, M., Salojärvi, J., Kangasjärvi, S., Idänheimo, N., Mersmann, S., Robatzek, S., Karpiński, S., Karpińska, B., and Kangasjärvi, J. (2010). Transcriptional regulation of the CRK/DUF26 group of receptor-like protein kinases by ozone and plant hormones in *Arabidopsis*. BMC Plant Biol. **10:** 95.

Wu, H.-J., et al. (2012). Insights into salt tolerance from the genome of *Thellungiella salsuginea*. Proc. Natl. Acad. Sci. USA **109:** 12219–12224.

Xu, Z., and Wang, H. (2007). LTR_FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. Nucleic Acids Res. **35** (Web Server issue)**:** W265–W268.

Yang, R., et al. (2013). The reference genome of the halophytic plant *Eutrema salsugineum*. Front. Plant Sci. **4:** 46.

Zdobnov, E.M., and Apweiler, R. (2001). InterProScan—An integration platform for the signature-recognition methods in InterPro. Bioinformatics **17:** 847–848.

Zhang, X., Wang, L., Yuan, Y., Tian, D., and Yang, S. (2011). Rapid copy number expansion and recent recruitment of domains in S-receptor kinase-like genes contribute to the origin of self-incompatibility. FEBS J. **278:** 4323–4337.