





# The Snapdragon Genomes Reveal the Evolutionary Dynamics of the S-Locus Supergene

Sihui Zhu <sup>†,1,2,3</sup> Yu'e Zhang <sup>†,4</sup> Lucy Copsy,<sup>5</sup> Qianqian Han <sup>3,4</sup> Dongfeng Zheng,<sup>1,2,3</sup> Enrico Coen,<sup>5</sup> and Yongbiao Xue <sup>\*,1,2,3,4</sup>

<sup>1</sup>National Genomics Data Center & CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing, China

<sup>2</sup>China National Center for Bioinformation, Beijing, China

<sup>3</sup>University of Chinese Academy of Sciences, Beijing, China

<sup>4</sup>State Key Laboratory of Plant Cell and Chromosome Engineering, Institute of Genetics and Developmental Biology, and the Innovation Academy of Seed Design, Chinese Academy of Sciences, Beijing, China

<sup>5</sup>John Innes Centre, Norwich, United Kingdom

<sup>†</sup>These authors contributed equally to this work.

\*Corresponding author: E-mail: ybxue@genetics.ac.cn.

Associate editor: Rebekah Rogers

## Abstract

The genus *Antirrhinum* has been used as a model to study self-incompatibility extensively. The multi-allelic S-locus, carrying a pistil S-RNase and dozens of S-locus F-box (SLF) genes, underlies the genetic control of self-incompatibility (SI) in *Antirrhinum hispanicum*. However, there have been limited studies on the genomic organization of the S-locus supergene due to a lack of high-quality genomic data. Here, we present the chromosome-level reference and haplotype-resolved genome assemblies of a self-incompatible *A. hispanicum* line, *AhS<sub>7</sub>S<sub>8</sub>*. For the first time, 2 complete *A. hispanicum* S-haplotypes spanning ~1.2 Mb and containing a total of 32 SLFs were reconstructed, whereas most of the SLFs derived from retroelement-mediated proximal or tandem duplication ~122 Mya. Back then, the S-RNase gene and incipient SLFs came into linkage to form the pro-type of type-1 S-locus in the common ancestor of eudicots. Furthermore, we detected a pleiotropic cis-transcription factor (TF) associated with regulating the expression of SLFs, and two miRNAs may control the expression of this TF. Interspecific S-locus and intraspecific S-haplotype comparisons revealed the dynamic nature and polymorphism of the S-locus supergene mediated by continuous gene duplication, segmental translocation or loss, and TE-mediated transposition events. Our data provide an excellent resource for future research on the evolutionary studies of the S-RNase-based self-incompatibility system.

**Key words:** snapdragon, s-locus, supergene, evolutionary genomics, SLFs.

## Introduction

Self-incompatibility (SI) is a molecular recognition system that prevents inbreeding and promotes outcrossing in hermaphroditic flowering plants, thereby maintaining genetic diversity and helping angiosperms expand into a wide range of habitats. SI is found in diverse taxa of flowering plants, including monocotyledons and dicotyledons. The molecular mechanisms of SI in eudicots have been extensively studied for decades. In many species, SI is controlled by a single supergene named the S-locus, with various haplotypes (Takayama and Isogai 2005), carrying linked female and male specific S-determinants. In Brassicaceae and Papaveraceae, harboring a self-recognition SI system, the female and male S-genes are coevolving polymorphic proteins (Sato et al. 2002; Wheeler et al. 2009). In Solanaceae, Rosaceae, Rubiaceae, Rutaceae, and Plantaginaceae, the S-locus, also called type1 S-locus (Zhao et al. 2022), generates the pistil S-determinant (S-ribonuclease with cytotoxicity,

termed S-RNase) and the pollen S-determinants (S-locus F-box genes called SLFs or SFBB) that mediate non-self-recognition (McClure et al. 1989; Sassa et al. 1996; Xue et al. 1996; Sijacic et al. 2004; Asquini et al. 2011; Liang et al. 2020). In a type-1 S-RNase-based SI system, a unique S-haplotype consists of multiple specific SLFs/SFBB and a specific S-RNase, whereas a recombination event between them may lead to a nonfunctional S-haplotype (Fujii et al. 2016). Phylogenetic studies of SLFs and S-RNase in Solanaceae and Plantaginaceae showed no evidence of coevolution, with SLFs having a much shorter evolutionary history (Newbigin et al. 2008).

Supergenes controlling complex phenotypes have long inspired both empirical and theoretical studies on them. The type-1 S-locus supergene identified in many plant lineages, including *Petunia* (Wu et al. 2020), *Solanum* (Li and Chetelat 2015), and *Citrus* (Zhang et al. 2015; Liang et al. 2020), vary in size, sequence diversity, number of

© The Author(s) 2023. Published by Oxford University Press on behalf of Society for Molecular Biology and Evolution.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial License (<https://creativecommons.org/licenses/by-nc/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited. For commercial re-use, please contact [journals.permissions@oup.com](mailto:journals.permissions@oup.com)

Open Access

genes, and mechanism of recombination repression (Gutiérrez-Valencia et al. 2021). In citrus, each S-haplotype comprising an *S-RNase* and ~9 *SLFs* spanned 198 to 370 kb, with 117 citrus *SLFs* being clustered into 12 types (Liang et al. 2020). In *Solanum tuberosum*, 13 *SLFs* are specifically located in the subcentromeric region of chromosome 1, covering 14.6 Mb in this Solanaceae species (Gebhardt et al. 1991). In *Rosa chinensis* chromosome 3, the S-locus and its flanking region containing 30 *SFBB/SLFL* (also called F-box) and an *S-RNase* occupied the region from 1.5 to 43.7 Mb (Vieira et al. 2021). However, no complete S-locus of any Plantaginaceae species has been published so far. An S-haplotype with greater fitness has all the *SLFs* required to express and detoxify all the *S-RNase* in a population. It should be noted that in almond, some F-box genes, not involved in SI specificity determination due to its little allelic sequence polymorphism and expression in pistil, are also found in the vicinity of the *S-RNase* gene (Ushijima et al. 2003). Therefore, searching for the S-genes based on homology in other species as widely used common method for now is obviously problematic. To determine all the *SLFs* and *RNases* involved in SI recognition, the systematic approach integrating genome sequences and transcriptome data simultaneously is necessary in postgenomic era. Also, in plant SI species, most previous reports focused on the discussions of the S-genes themselves. The phylogenetic analyses of *RNase* and *SLFs* gene superfamily of four exemplar families (Solanaceae, Rosaceae, Rutaceae, and Plantaginaceae) revealed that *S-RNase* and *SLFs* both have a single origin in the majority of core eudicots (Steinbachs and Holsinger 2002; Newbigin et al. 2008; Vieira et al. 2008; Zhao et al. 2022). Few systematic analyses of the origin and evolution of the S-locus and the impact of gene duplication on the shaping process of the functional locus were not considered, presenting a critical knowledge gap. Reconstructing the evolutionary history of a supergene is necessary to understand how such a functional module originates and spreads in diverse taxa (Hager et al. 2022; Potente et al. 2022). Yet, sequenced genomes of self-incompatible species are still limited, thereby hampering the further insight of this important supergene, especially the identification of the authentic S-genes and understanding of how multiple male S-genes proliferated in non-self-recognition SI system, as well as of how the whole supergene evolved in different lineages.

With the advent of the flourish sequencing technologies and accurate and contiguous genome assembly, the critical basis for understanding genomic diversity is easily accessible now. In this study, we assembled a chromosome-level reference genome with a total length of ~495 Mb of self-incompatible *A. hispanicum* (*AhS<sub>7</sub>S<sub>8</sub>*, heterozygous genotype at the S-locus, hereafter referred to as “SI-Ah”). Also, high-quality draft genomes of self-incompatible *Antirrhinum linkianum* and self-compatible *Misopates orontium* (common name: lesser snapdragon) were present here for comparative analyses. We also took the advantage of HiFi data and Hi-C reads to assemble the haploid genomes

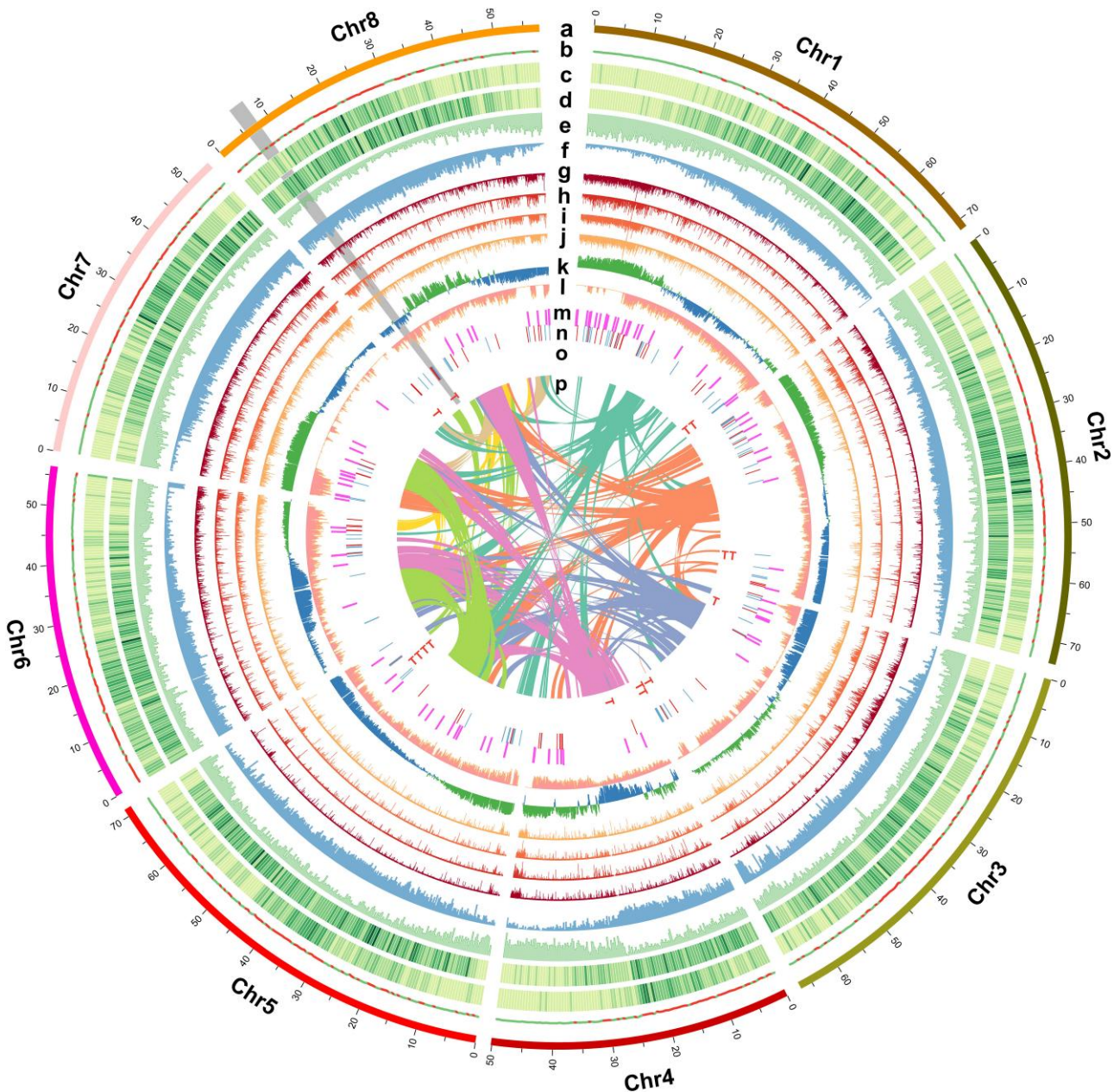
of SI-Ah, each carrying a copy of the S-haplotype. The complete structure and sequence details of the *Antirrhinum* S-haplotypes allowed us to inspect how and when the multiple *SLF* genes proliferated, as well as a candidate TF potentially coregulating the expression of the pollen S-determinant genes. Moreover, the syntenic relationships of the S-locus from five type-1 *S-RNase*-based self-incompatible species confirmed their ancient origin in the common ancestor of Rosids and Asterids and revealed the evolutionary dynamics of S-locus in different lineages. The genomic data and analysis reported in this work are of great value for understanding the genetic bases of adaptive characteristic of species and provide insights into the evolution of the S-locus supergene.

## Results

### Genome Assembly and Evaluation

The S-genotype of a self-incompatible Spanish snapdragon, *A. hispanicum*, has been determined as *S<sub>7</sub>S<sub>8</sub>* by PCR (supplementary fig. S1, Supplementary Material online). We combined different sequencing strategies to derive *A. hispanicum* reference genome assembly: Illumina pair-ended reads, PacBio single-molecule reads, BioNano optical maps, and Hi-C sequencing reads. Based on the 21-mer spectrum using Illumina reads, *AhS<sub>7</sub>S<sub>8</sub>* was estimated to have a haploid genome size of ~473.4 Mb and a modest heterozygosity of 0.59%. Firstly, 36.0 Gb of continuous long reads (CLR) and 19.1 Gb of circular consensus reads (CCS) were obtained on PacBio platform for contig-level assemblies using different assemblers separately. A total of 203.3 Gb of BioNano data and 69.9 Gb of Hi-C paired-end reads were generated to enhance the assembly contiguity. In the end, we obtained three sets of chromosome-level genome assemblies of *AhS<sub>7</sub>S<sub>8</sub>*, a consensus one with mosaic sequence structure (fig. 1) and two haplotype-resolved assemblies each carrying a S-haplotype (*S<sub>7</sub>* haplome with *S<sub>7</sub>*-haplotype and *S<sub>8</sub>* haplome with *S<sub>8</sub>*-haplotype) (supplementary figs. S3 and S4, Supplementary Material online). In this study, we used the consensus assembly to represent *A. hispanicum* for subsequent analyses unless otherwise specified. To facilitate comparative analyses, we also sequenced and assembled the genomes of self-incompatible *A. linkianum* and self-compatible *M. orontium*. The heterozygosity and genome size of *A. linkianum* gave very close result to *A. hispanicum*, whereas *M. orontium* exhibited a heterozygosity rate of 0.21% and ~24.1% reduction in genome size (table 1).

The BUSCO assessment results indicated that 95.0–98.7% of conserved genes were recovered in these assemblies (table 1). In addition, the LAI score of the two haplomes exceeded 20, demonstrating a “golden” standard level, and the draft genome of the other two species also reached “reference” grade quality (19.3 and 21.9 for *A. linkianum* and *M. orontium*, respectively) (Ou et al. 2018). Cumulatively, the results above suggested the reliability



**Fig. 1.** Genomic features of *A. hispanicum*. The outermost layer is a circular ideogram of the eight pseudomolecules. Tracks **b–j** correspond to the distribution of the GC content (red denotes higher than average whole-genome level and green vice versa), Ty3/Gypsy, Ty1/Copia retrotransposons (higher density shown in darker green), gene density, methylation level, and expression level of four tissues (petal, pollen, pistil, and leaf), respectively, all calculated in 250-kb nonoverlap window. Track **k** corresponds to A/B compartments inferred by using Hi-C data. Track **l** corresponds to SNP/InDel density distribution. Track **m** corresponds to the location of miRNA genes. Track **n** corresponds to the location of all F-box genes; blue and red represent forward and reverse strands, respectively. Track **o** marker “T” corresponds to the location of ribonuclease T2 genes. Track **p** corresponds to interchromosome syntenic blocks. The gray sector marked S-locus in Chr8.

of all the nuclear genome assemblies. In addition, circular chloroplast and mitochondrial genomes were also retrieved from the contigs, with a length of 144.8 and 527.4 kb, respectively (supplementary fig. S11, Supplementary Material online).

Repetitive elements and gene models in all the genome assemblies were predicted following the same analysis pipeline, respectively, as described in the Materials and Methods section. Repetitive elements accounted for

~43% of the *Antirrhinum* genome (table 1), and long terminal repeats (LTRs) are the most prominent member of transposons families, covering ~23% of the nucleus genome (fig. 1). Gene models of all genome assemblies were predicted following the same analysis pipeline as described in the Materials and Methods section. A total of 39,258–43,844 protein-coding genes were predicted for each of the soft-masked assemblies (table 1). Nearly 90% of the predicted genes could be assigned to a term in at least

**Table 1.** Statistics of the Genome Assemblies.

Statistic	<i>A. hispanicum</i>			<i>A. linkianum</i>	<i>Misopates orontium</i>
	Consensus assembly	S <sub>7</sub>	S <sub>8</sub>		
Total assembly size (Mb)	495.3	502.6	487.7	519.0	365.7
Heterozygosity (%)	0.59	—	—	0.60	0.21
Max. contig length (Mb)	1.8	48.6	49.4	9.9	50.7
Min. contig length (kb)	20.0	17.6	17.4	11.9	10.0
Contig N50 length (Mb)	0.2	11.3	14.1	2.1	19.6
GC content	35.6	35.5	35.4	35.6	35.7
Repeat element (%)	43.8	45.8	44.7	47.5	38.0
Protein-coding genes (n)	42,667	43,844	43,770	43,022	39,258
Mean transcript length (bp)	3,023	2,781	2,834	2,961	2,896
BUSCO (%)	95.0	96.7	95.7	95.5	98.7
LAI	19.8	21.1	20.3	19.3	21.9

one functional database, suggesting the solidity of the whole-genome annotation results.

### Whole-Genome Duplication in Plantaginaceae

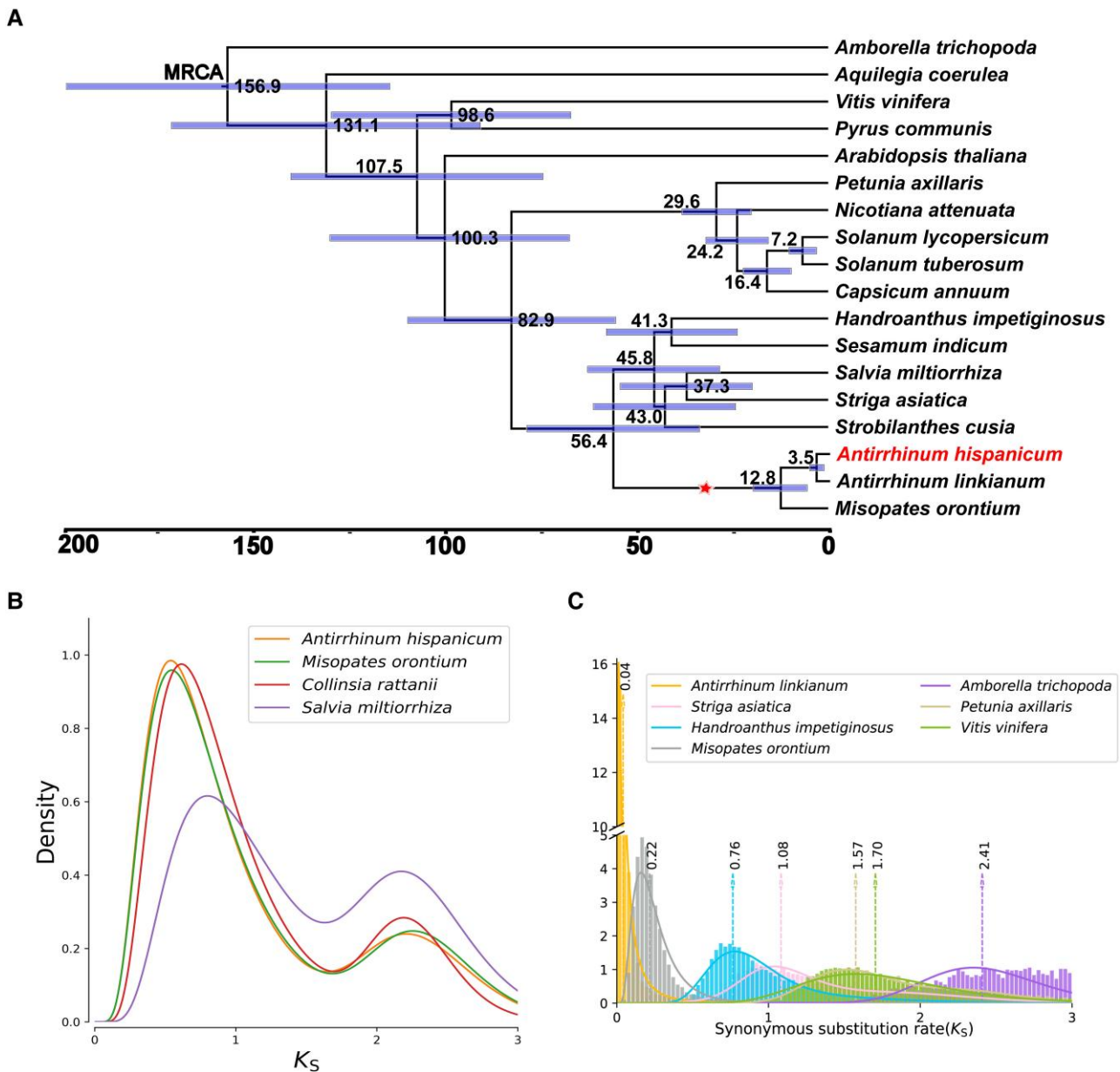
Proteomes from *A. hispanicum*, *A. linkianum*, and *M. orontium*, together with 15 other species, were clustered into 36,295 orthogroups and used for phylogenetic relationship construction subsequently (fig. 2A and supplementary table S11, Supplementary Material online). The genus *Antirrhinum* and *Misopates*, belonging to the family Plantaginaceae, formed a monophyletic group that diversified recently. The species tree indicated that the divergence of Plantaginaceae from Lamiales order common ancestor occurred ~55.8 Mya during the Paleocene–Eocene transition (fig. 2A). *Antirrhinum hispanicum* and *A. linkianum* split at around 3.5 Mya (95% confidence interval [CI] = 1.6–5.3 Mya), which collided with the emergence of modern Mediterranean climate and rapid speciation of *Antirrhinum* lineage (Vargas et al. 2009). To identify and approximately date the WGD events in *Antirrhinum*, we first constructed the paranomes of several species (supplementary fig. S15, Supplementary Material online). The synonymous substitution rate (Ks) distribution plot of paranomes displayed two peaks (fig. 2B), indicating *Antirrhinum* have at least experienced another one WGD after the gamma-whole-genome triplication (WGT) (Jiao et al. 2012). The distributions plots of Ks for *A. hispanicum*, *Collinsia rattanii* (Frazee et al. 2021), and *A. linkianum* paranomes showed a distinct peak at 2.21 echoing the well-reported gamma-WGT that occurred ~117 Mya in the common ancestors of all eudicots (Jiao et al. 2012) (fig. 2B and supplementary fig. S15, Supplementary Material online) and an additional peak at ~0.61 corresponding to a recent duplication event shared among Plantaginaceae members. Therefore, the timing of the younger WGD was estimated to occur at ~32.3 Mya, posterior to the divergence of *Antirrhinum* and other Lamiales plants, indicating a Plantaginaceae-specific WGD event.

The neutral mutation rate of the *Antirrhinum* genus was inferred using the formula  $\mu = D/2T$ , where  $D$  is the evolutionary distance of two *Antirrhinum* species (peak value of Ks log-transformed distribution = 0.04, fig. 2C) and  $T$  is the divergence time of the two species (3.5 Mya). The neutral

mutation rate was estimated to be  $5.7e^{-9}$  (95% CI:  $3.8e^{-9}$ – $1.3e^{-8}$ ) substitution per site per year, which is very close to the result estimated by using genotyping-by-sequencing data set (Otero et al. 2021).

### Position, Structure, and Flanking Genes of the S-Locus Supergene

Although functional genes at the *Antirrhinum* S-locus have attracted much attention for a long time (Xue et al. 1996; Lai et al. 2002; Qiao et al. 2004), the complete genomic structure and full gene members of the supergene is still unknown. Identifying sequence details of the S-locus supergene would be the initial step toward understanding its evolution. The *Antirrhinum* S-locus was found to locate in the peri-telomeric region on the short arm of chromosome 8 (Yang et al. 2007), as opposed to the subcentromeric area in *Solanaceous* species (Ten Hoopen et al. 1998). A total of 275 S-locus-like F-box genes (SLFs) were distributed over the entire *A. hispanicum* genome (fig. 1), whereas 370 and 258 copies were in the genomes of *A. linkianum* and *M. orontium*. A cluster of 32 SLFs (S-locus F-box), which matches the structural characteristics of the candidate S-locus, are located in chromosome 8 (fig. 1). The gapless S-locus of 1.25 Mb in *A. hispanicum* and pseudo-S-locus of 804 kb in self-compatible *Antirrhinum majus* (Li et al. 2019) were supported with BioNano long-range molecules as strong evidence (supplementary figs. S17 and S18, Supplementary Material online), and the S-loci of *A. linkianum* and *M. orontium* were identified by synteny alignment with *A. hispanicum*. For the sake of simplicity, we use “S<sub>7</sub>-haplotype” and “S<sub>8</sub>-haplotype” to refer to the two alleles of S-locus in *AhS<sub>7</sub>S<sub>8</sub>* and their SLF genes as “S<sub>7</sub>-SLF” and “S<sub>8</sub>-SLF”, whereas S<sub>AI</sub>-haplotype refers to the S-locus of self-incompatible *A. linkianum*. There are 32 SLFs annotated in S<sub>7</sub>-haplotype (S<sub>7</sub>-SLF1 ~ S<sub>7</sub>-SLF32) and S<sub>8</sub>-haplotype (S<sub>8</sub>-SLF1 ~ S<sub>8</sub>-SLF32), respectively. Their position in chromosome and sequence information are listed in supplementary table S16, Supplementary Material online. Am-S<sub>C</sub>-haplotype and Mo-S<sub>C</sub>-haplotype which contain no S-RNase denote  $\psi$ S-locus of self-compatible *A. majus* and *M. orontium*. The S-RNase gene was located in roughly the middle of the locus and surrounded by the pollen



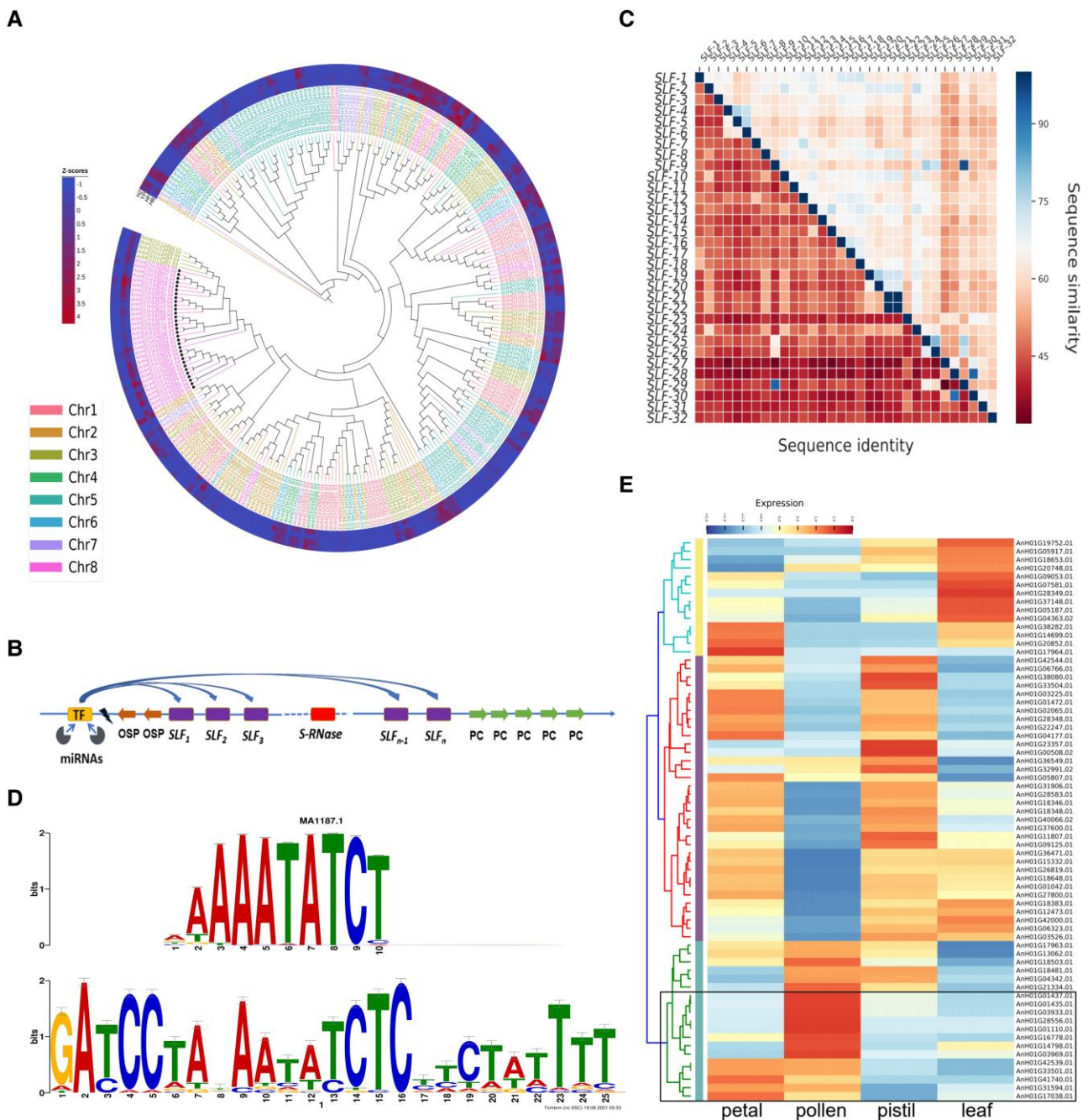
**Fig. 2.** Dating the Plantaginaceae-specific WGD event. (A) The phylogenetic tree of 18 species including 3 species sequenced in this study. The red star denotes the Plantaginaceae-specific WGD. The blue bar indicates 95% highest posterior density interval. (B) Parame  $K_s$  distribution of three Plantaginaceae members and Lamiaceae *Salvia miltiorrhiza* to identify WGD events. Three Plantaginaceae members have the same two WGD peaks, with the older one corresponding to the gamma-WGT occurred ~117 Mya in the ancestor of core eudicots and the younger one a Plantaginaceae-specific WGD. (C)  $K_s$  distribution of one-to-one orthologs identified between *Antirrhinum hispanicum* and other species. The density curves were fitted using Gaussian mixture model, and the peak values were marked with dotted dash lines.

S-genes. Except for *SLFs* and *S-RNase*, the S-locus also contains genes encoding G-beta repeat, matE, zinc finger, cold acclimation, auxin-responsive proteins, and dozens of genes with unknown or TE-related functions (supplementary table S19, Supplementary Material online). Additionally, the S-locus was flanked by two copies of organ-specific protein on the left side (OSP1, *AnH01G33547.01*, and OSP2, *AnH01G33548.01*) and five tandem copies of phytocyanin genes on the right side (here named PC1, *AnH01G33649.01*; PC2, *AnH01G33650.01*; PC3, *AnH01G33651.01*; PC4, *AnH01G33652.01*; and PC5, *AnH01G33653.01*) (fig. 3B). In Mo-S<sub>C</sub>-haplotype, only four homologous PCs were found,

and the tandem structure was interrupted by several other genes. The *A. hispanicum* genome assemblies are, to the best of our knowledge, one of the most contiguous and complete de novo plant genomes covering the complex S-locus supergene to date.

### The S-Locus Expanded via Tandem/Proximal Duplication

There is no doubt that gene duplication plays a crucial role in the shaping process of such a complex supergene. Duplicated genes were classified into five categories:



**Fig. 3.** Identification of a candidate regulator of *SLFs*. (A) Phylogenetic tree reconstructed based on the amino acid sequences of all F-box genes from *Antirrhinum hispanicum* genome. The ML phylogenetic tree was calculated using RAxML-NG with 100 bootstrap replicates. The color of gene ID denoted the chromosome and gene within as indicated in the legend. Heatmap color represent Z-scores derived from RNA-seq expression data for each gene. *SLF* genes are marked with black stars. (B) Schematic structure and regulatory relationship of snapdragon *S*-locus. (C) Heatmap showing sequence identity (lower left triangle) and sequence similarity (upper right triangle) between *SLFs* in SI-Ah. The genes in the x-axis and y-axis are ordered by physical position on the chromosome. (D) The bottom panel represents motif found by scanning upstream 1000 bp of 32 *SLFs* using MEME, and the top panel is a Tomtom alignment for annotation of the putative transcription factor binding site. (E) The cluster heatmap of expression matrix of all MYB-related transcription factor family members in SI-Ah genome. The columns represent four tissues. The legend shows the range of  $\log_2(\text{TPM} + 0.5)$ , and TPM is mean value of two biological replicates. Genes marked in black box may be involved in controlling the expression of *SLFs*, since they are expressed in pollen but not in pistil.

whole-genome duplication (WGD), tandem duplication (TD), proximal duplication (PD), transposed duplication (TRD), and dispersed duplication (DSD). It has been proved that TD and PD evolved toward functional roles in plant self-defense and adaptation to dramatically

changing environments (Qiao et al. 2019). To inspect how multiple *SLFs* were generated and proliferated, in the first place, we built the phylogenetic tree of 275 *SLFLs* (fig. 3A). So, if all the *SLFs* originated via WGDs or large segmental duplication, their closest paralogs should

colocalize elsewhere in the genome and have high sequence similarity with the S-locus F-boxes. On the contrary, if *SLFs* were the results of stepwise local duplications, they would exhibit less divergences, and there should be more distant paralogs scattered elsewhere in the genome. The phylogenetic tree of snapdragon *SLFs* proved that the situation is the latter one. The result of duplicates pattern classification revealed that 24 and 6 *SLFs* are PD and TD, respectively, and the other two *SLFs* are dispersed duplicates (supplementary table S19, Supplementary Material online). Besides, 27 of the 32 *SLFs* are intron-less, suggesting *SLFs* proliferated most likely through the retroelement-mediated way. The pairwise sequence identity of 32 *SLF* paralogs ranged from 33.1% to 77.3%. From the intraspecies similarity matrix heatmap of 32 *SLFs*, we can also observe that the closer the physical distance between two *SLFs*, the higher the sequence similarity and identity (fig. 3C).

Based on the gene duplication pattern analyses, the two flanking OSPs were derived from the gamma-WGD (Jiao et al. 2012) followed through recent tandem duplication (supplementary table S21, Supplementary Material online), whereas the PCs arise from tandem duplication asynchronously assuming a constant substitution rate; both can be supported by the gene phylogenetic tree topology of their paralogs (supplementary figs. S20 and S21, Supplementary Material online). Three pseudogenes at the right end of the S-locus derived from two *SLFs* lost their function because of the presence of premature stop codons. Dissection of *SLFs* and flanking genes indicates that the *Antirrhinum* S-locus experiences gene gaining and loss constantly (supplementary tables S21 and S22, Supplementary Material online).

### *SLFs* Were Coregulated by a Versatile Transcript Factor

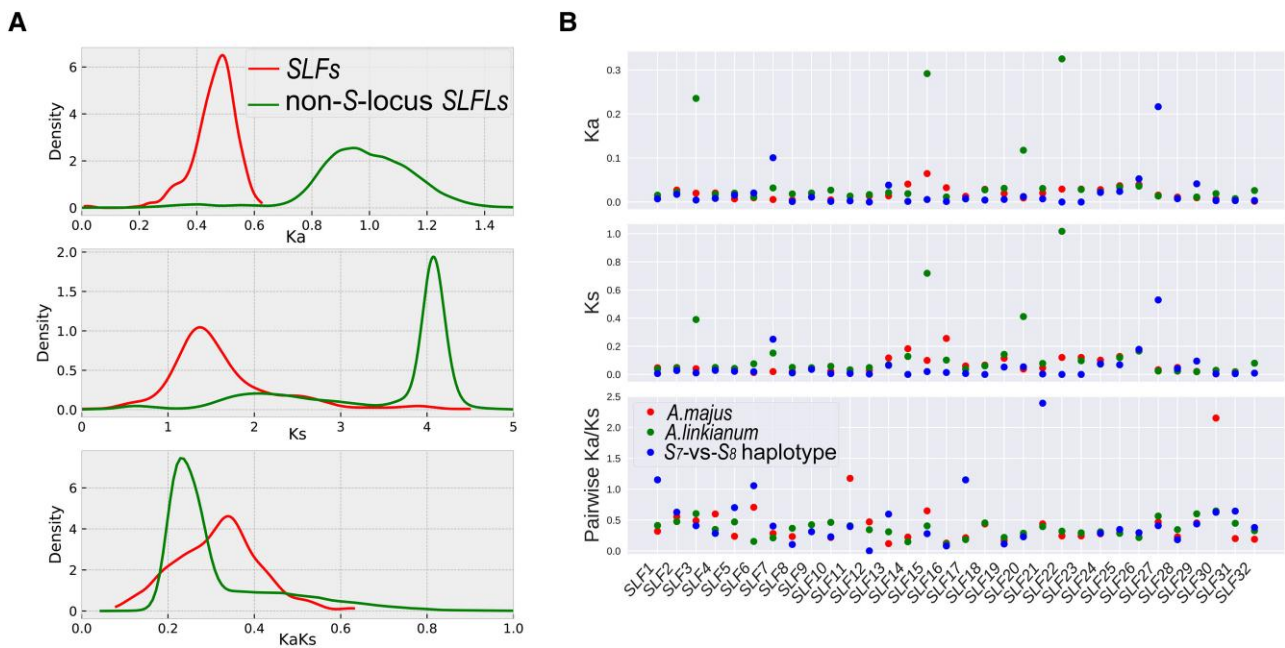
The phylogenetic tree, along with the expression heatmap of 275 *SLFLs*, indicated *SLFs* not only share a closer relationship but also display similar pollen-specific expression patterns, which distinguished from their paralogs in the genome (fig. 3A). This phenomenon has not changed much in cultivar *A. majus* after its domestication (supplementary fig. S19, Supplementary Material online). Inspired by operons in prokaryotes, a regulator controlling the expression of all the tandem *SLF* members was expected. Using MEME suit analysis, we identified a significant motif (GATCCTAXAATATCTC) located upstream region of the *SLFs*, and the motif annotation implied that *SLFs* might be coregulated by an MYB-related family TF (fig. 3D). By examining the expression levels of all the MYB-related family members in the snapdragon genome, *AnH01G01437.01*, *AnH01G01435.01*, *AnH01G03933.01*, *AnH01G28556.01*, *AnH01G01110.01*, *AnH01G16778.01*, *AnH01G14798.01*, *AnH01G03969.01*, *AnH01G42539.01*, *AnH01G33501.01*, *AnH01G41740.01*, *AnH01G31594.01*, and *AnH01G17038.01* were expressed in pollen (and petal) but barely in pistil tissue (fig. 3E). Among them, *AnH01G33501.01* is

located about 1.2 Mb upstream of the S-locus, making it a potential regulator involved in activating the expression of the *SLFs*. Interestingly, this MYB TF also called *RADIALIS* (*RAD*) is proved to control dorsoventral asymmetry in *Antirrhinum* by interplaying with other TFs (Corley et al. 2005; Baxter et al. 2007). The genetic linkage between the S-locus and other functional genes underlying important development traits manifested the pleiotropy of the MYB TF. Moreover, the analysis of small RNA-seq libraries from 4 tissues identified 108 known miRNA genes of 42 families of *A. hispanicum* (details listed in supplementary table S9, Supplementary Material online). By examining miRNA target prediction results, two candidate miRNAs (*aof-miR396b* and *zma-miR408b-5p*) were revealed to bind with different regions of *AnH01G33501.01*, and their expression patterns exhibited a negative correlation with the *AnH01G33501.01* (supplementary fig. S10 and table S17, Supplementary Material online). The miR396 family member, *aof-miR396b*, was believed to influence GRF TF expression (Debernardi et al. 2012), whereas the overexpression of *zma-miR408b* in maize was reported to respond to abiotic stress to maintain leaf growth and improve biomass (Aydinoglu 2020). The results implied that two miRNAs might also play a vital role in controlling self-incompatibility.

### *SLFs* Are Younger Than Non-S-Locus F-Boxes

Usually, TD and PD genes had qualitatively higher nonsynonymous substitution rate ( $K_a$ )/ $K_s$  ratios and lower  $K_s$  value than gene derived from the other duplication types (Liu et al. 2022; Wang et al. 2022). The value of  $K_a$ ,  $K_s$ , and  $K_a/K_s$  of pairwise *SLFs* and non-S-locus F-box genes was calculated, and the  $K_s$  of *SLF* paralogs are much lower than those of non-S-locus F-box paralogs (fig. 4A). The  $K_a/K_s$  values of all F-box gene pairs in snapdragon were all below one, but the *SLFs* exhibited higher  $K_a/K_s$  ratios than non-S-locus F-boxes, indicating higher selection pressure. We also observed an *SLF* of  $S_7$ - and  $S_8$ -haplotype shared identical coding sequence ( $S_7$ -*SLF*<sub>22</sub> and  $S_8$ -*SLF*<sub>22</sub>), which could be caused by genetic exchange (Kubo et al. 2015). Therefore, we used GENECONV (Sawyer 1989) to detect signal of genetic exchange among allelic *SLFs*. Based on the sequence alignments of  $S_7$ -haplotype,  $S_8$ -haplotype, and Am- $S_C$ -haplotype from self-compatible *A. majus* (Li et al. 2019), we found eight cases of significant genetic pairwise exchanges in four *SLF* alleles (supplementary table S15, Supplementary Material online). These results indicate that genetic exchanges in snapdragon may also partly contribute to the functional conservation of *SLFs*, but not as frequent as in Solanaceae species which possessed a much larger S-locus (Kubo et al. 2015).

A total of 125 *SLFs* from 4 S-haplotype ( $S_{AI}$ -haplotype of *A. linkianum*,  $S_7$ ,  $S_8$ , Am- $S_C$ -haplotype) were clustered into 31 subgroups (or types, as defined in Kubo et al. [2010]), suggesting diversification of allelic *SLFs* within each type followed the divergence of genera (supplementary fig. S22, Supplementary Material online).  $S_7$ -*SLF*<sub>22</sub> and  $S_7$ -*SLF*<sub>23</sub>



**Fig. 4.** The inter- and intraspecific comparisons among *Antirrhinum* SLFs. (A) The Ka, Ks, and Ka/Ks distribution of SLFs and other non-S-locus SLFLs. (B) Pairwise Ka, Ks, and Ka/Ks comparison for all SLFs between  $S_7$ -haplotype with  $S_{AI}$ -haplotype of *Antirrhinum linkianum* and  $A_m$ - $S_C$ -haplotype of *Antirrhinum majus* and  $S_8$ -haplotype. Genes are ordered by genomic coordinates. Gene pairs with  $K_s = 0$  were removed to avoid misinterpretation.

have the same coding sequence, whereas  $S_8$ - $SLF_6$  has no correspondence in the other  $S_7$ -haplotype, indicating that the generation of new SLF types might have continued after speciation. The estimation of  $K_s$ ,  $K_a$ , and  $K_a/K_s$  between allelic SLFs showed that interallelic  $K_a$  and  $K_s$  values of all the types (fig. 4B) were much lower than that among paralogous SLFs. These results indicated that alleles of each SLF type are much younger than the SLF paralogs, too. The  $K_s$  value of interspecific SLFs is slighter lower than that of intraspecific SLFs, implying some SLFs have experienced rapid sequence divergence after speciation. In addition, the values of the  $K_a/K_s$  of most allelic SLFs below one are a signature of purifying selection that is likely to maintain necessary protein functions to detoxify S-RNase (fig. 4B).

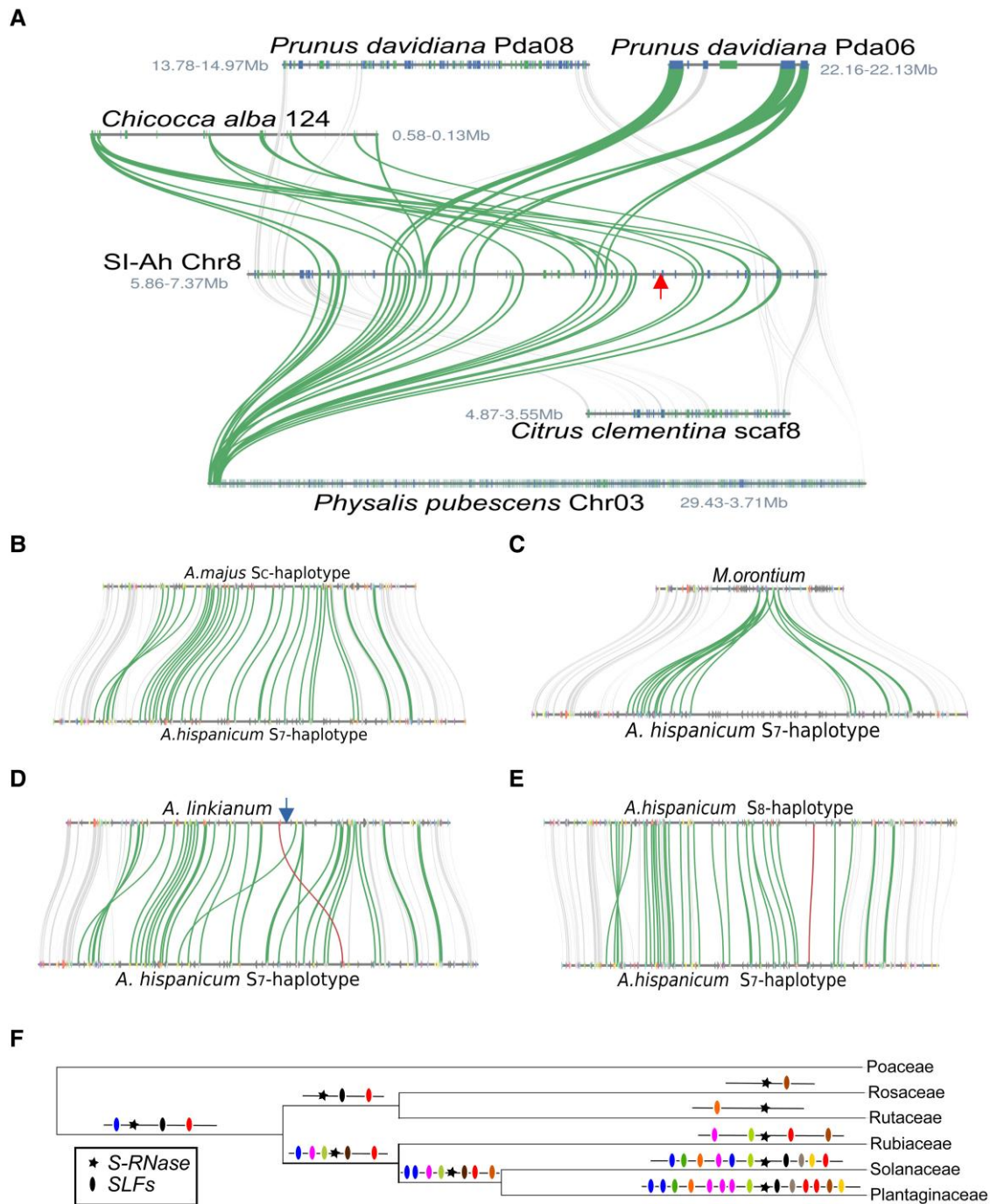
With the estimated *Antirrhinum* mutation rate in previous section, the main peak value of SLF paralog  $K_s$  distribution (ranging from 1.35 to 1.40) corresponding 122 Mya (fig. 4A) indicated that the most SLFs have ancient origin. According to phylogenetic analyses of class III S-RNases, the type-1 S-RNase-based SI system has evolved only once, before the split of the Asteridae and Rosidae, about 120 Mya (Ilgic and Kohn 2001; Vieira et al. 2008; Zhao et al. 2022). This coincidence suggests that recruitment of multiple SLF copies must be very vital for the establishment of S-RNase-based SI system and subsequent expansion of angiosperms.

### Evolution History of the *Antirrhinum* S-Locus

The origin and evolutionary history of the type-1 S-locus, composed of S-RNase and SLF genes, have not been well

studied extensively yet. Therefore, we performed microsynteny analyses between *A. hispanicum* and other type-1 S-RNase-based self-incompatible species (Franklin-Tong and Franklin 2003; McClure and Franklin-Tong 2006; Vieira et al. 2008; Zhao et al. 2022). A comparison of microsynteny revealed the presence of the S-locus remnant in some species, including *Chiococca alba* (Rubiaceae) (He et al. 2022), *Citrus clementina* (Rutaceae) (Wu et al. 2014), *Physalis pubescens* (Solanaceae) (Lu et al. 2021), and *Prunus davidiana* (Rosaceae) (Cao et al. 2022), but not in the syntenic regions of species outside Eudicotyledoneae, such as *Oryza sativa* (Poaceae). The distribution pattern also suggested that the S-locus emerged early right after the divergence of Eudicotyledoneae from other Mesangiosperms (Vieira et al. 2008). In *P. davidiana*, the core part of whole S-locus was translocated to a different chromosome, and only four F-boxes are still there occupying a drastically contracted region (fig. 5A). Among the four genes, two of them have four and three orthologous SLFs in *A. hispanicum* implying that the multiple copies arose after species divergence. The same contraction was also observed in *C. alba* except with more F-box genes. Whereas in the citrus genome, no syntenic region with *A. hispanicum* S-locus was detected. However, in *P. pubescens*, a Solanaceae member phylogenetically closer with *Antirrhinum*, almost all SLFs were lined up in the same order with *Antirrhinum* and well conserved. No gene corresponding to the S-RNase in *A. hispanicum* was found in those species, which could be caused by loss or translocation of S-RNase or mis-assembly/annotation. For example, *pf03G063230.1* of *P. pubescens* at Chr03:106.7 Mb is





**FIG. 5.** Evolution of the *Antirrhinum* S-locus supergene. (A) Microsynteny of the S-locus between *Antirrhinum hispanicum* with other four species (*Chiococca alba*, *Citrus clementina*, *Physalis pubescens*, and *Prunus davidiana*). The red arrow denotes the *S-RNase* of *A. hispanicum*. (B–D) The micro collinearity patterns of S-haplotype in *Antirrhinum majus*, *Misopates orontium*, and *Antirrhinum linkianum*, with *S7*-haplotype from *A. hispanicum*. The Tam downstream 11 kb of *S-RNase* is marked by a blue arrow, and the TE is on the same strand with *S-RNase*. (E) The microsynteny of *S7*-haplotype versus *S8*-haplotype. The green curve lines and the red lines denote *SLFs* and *RNase* gene, respectively; other genes are colored gray. (F) Possible formation process of the extant *Antirrhinum* S-locus. The color of *SLFs* represents newly evolved ones to deal with the evolution of the *S-RNase* with new specificity.

homologous to *S-RNase* in *A. hispanicum* with amino acid sequence identity of 37.8%.

Therefore, it would be reasonable to infer that primitive *SLFs* appeared in the common ancestor of Rosids and Asterids (Igic and Kohn 2001; Steinbachs and Holsinger 2002). Then, accompanied by speciation, *SLFs* genes

underwent loss and gain at different rates in different lineages (fig. 5F). If the new *SLFN* copy can recognize more nonself *S-RNase* (Xue et al. 1996; Kubo et al. 2010; Fujii et al. 2016), then it is advantageous for carrying this new duplication and spreading through the population by selection. Gene acquisition was most likely through

tandem or proximal duplications from ancestor copies, and the losses mainly involved large fragment deletion.

We also tried to make a comparison of *S*-haplotypes from the same and close species to identify polymorphism of the supergene. Synteny plot showed structure and gene content of the flanking ends of different *S*-haplotypes were well conserved. The *S*<sub>7</sub>-haplotype has high collinearity with Am-*S*<sub>C</sub>-haplotype, *S*<sub>8</sub>-haplotype, and *S*<sub>AI</sub>-haplotype except a small inversion at the start of alignment (fig. 5B–E). From the comparison between Mo-*S*<sub>C</sub>-haplotype and *S*<sub>7</sub>-haplotype, a large deletion with a length of about 556 kb could be observed, which involved *S*-RNase and more than ten *SLFs* as well as dozens of functional genes in *S*<sub>7</sub>-haplotype. These genes include ATPase, cytokinin, DUF4218, and also several uncharacterized proteins. This structural variation can be responsible for the loss of self-incompatibility in *M. orontium*. Also, one flanking phyto-cyanin gene in the right of Mo-*S*<sub>C</sub>-haplotype was lost after splitting from *Antirrhinum*. A small inversion, which may involve with recombination suppression to maintain the linkage of the pistil and the pollen *S*-determinant genes, can be observed between the *S*<sub>7</sub> and *S*<sub>8</sub>-haplotype comparison. Several neighbor genes of *S*<sub>7</sub>-RNase have no corresponding allelic counterpart in *S*<sub>8</sub>-haplotype (fig. 5E), suggesting a divergent evolution history in different regions of the *S*-locus. The relative position of *S*-RNase in *S*<sub>AI</sub>-haplotype being different from *A. hispanicum* (fig. 5D) indicated that the changing physical position *S*-RNase does not impact its function and the dynamic change of the polymorphic *S*-locus in *Antirrhinum* has been very active. And we did observe a nearby CACTA transposable element in the vicinity of *A. linkianum* *S*-RNase, *Tam*, which is a very classical kind of TE in *Antirrhinum* and may lead to the gene transposition (Sommer et al. 1985).

Together, these results revealed the dynamic nature of the *S*-locus supergene during evolution and speciation in Eudicotyledoneae, which involves continuous gene duplication, segmental translocation or loss, and TE-mediated transposition.

### Allelic Expression and DNA Methylation Profiling

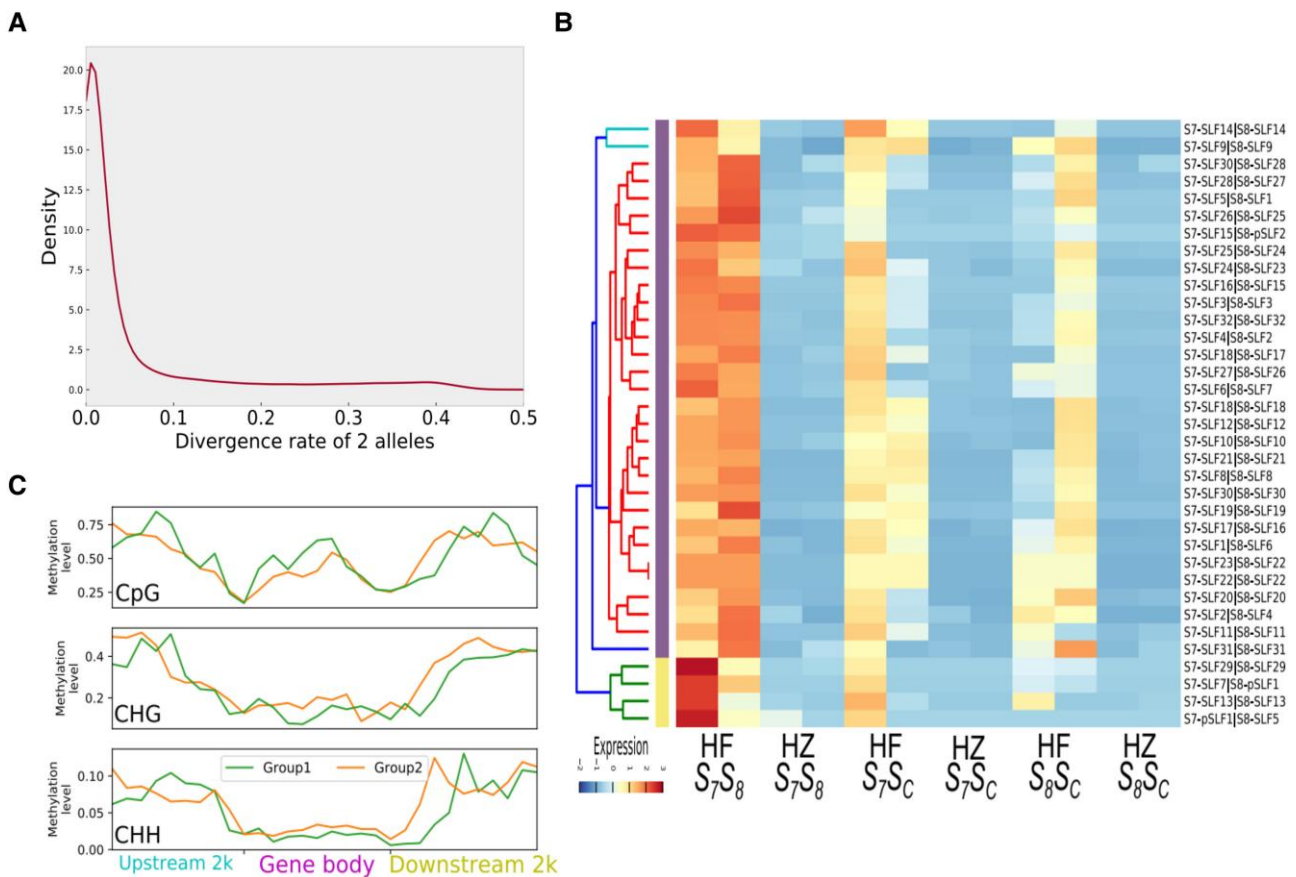
Since haplotype-resolved assembly is available and gene order is highly consistent between two haplomes (supplementary fig. S6, Supplementary Material online), we can determine two alleles of a physical locus and investigate allelic gene expression directly. Based on the syntenic and homologous relationship, 79,398 genes (90.6% of all predicted genes) were found to have homologs on the counterpart haplotype. Most allelic genes displayed a low level of sequence dissimilarity (fig. 6A). To understand the expression profile of allelic genes, RNA-seq data sets were analyzed using allele-aware quantification tools with combined coding sequences of two haplomes as reference. In the survey across 4 different tissues, most expressed loci did not exhibit significant variance, and only 2,504 gene pairs (3.2% of total genes) displayed significant

expression imbalance between two alleles ( $|\text{fold-change}| \geq 2$ ,  $P \leq 0.01$ ), suggesting that most alleles were unbiasedly expressed in *A. hispanicum* genome.

We analyzed four WGBS libraries to profile the methylation state of two haplomes. About 51.6–59.5 million PE reads were generated for each sample (supplementary table S2, Supplementary Material online). About 92.4–92.9% of total reads can be properly mapped to the haplomes. Reads for each sample were mapped to the snapdragon chloroplast genome, with 99% of the conversion rate higher confirming reliability of the experiments. Methycytosines were identified in CpG, CHH, and CHG contexts. A large fraction of cytosines in CpG (~80.6%) and CHG (~55%) contexts were methylated, but a smaller fraction of cytosines in CHH (~8%) context were methylated. No significant methylation level variations were observed between two haplomes, similar to that reported in potato (Zhou et al. 2020).

After examining *SLFs* and *S*-RNase expression of two haplotypes, we found that *S*<sub>7</sub>-RNase and *S*<sub>8</sub>-RNase expressed indiscriminately and reasonably high in *AhS*<sub>7</sub>*S*<sub>8</sub> pistil (supplementary table S18, Supplementary Material online), which is consistent with the situation in crab cactus (Ramanauskas and Igić 2021), and they are nearly undetectable in leaf and petal. Those two genes are only expressed in pistils of *AhS*<sub>7</sub>*S*<sub>C</sub> and *AhS*<sub>8</sub>*S*<sub>C</sub> respectively as we expected, yet at a higher level than that of *AhS*<sub>7</sub>*S*<sub>8</sub> pistil. Since the gene sequences did not change after one crossing experiment, it is very likely that expression of RNase was changed through other mechanisms. As for the expression profiles of *SLFs* in pollen tissue of three different genotypes, *S*<sub>7</sub>-*SLFs* and *S*<sub>8</sub>-*SLFs* did show an overall decline in two offspring samples but tended to be reduced generally. Those 32 *S*-genes could be divided into 2 groups. Group 1 consists of four genes, which were highly expressed in pollen of *AhS*<sub>7</sub>*S*<sub>8</sub> with obvious tissue specificities, but barely expressed in two hybrid progenies, and these genes displayed allelic bias. Another 28 *SLFs* in group 2 maintain expression patterns in 3 genotypes, but the expression level in two hybrid progenies is slightly lower than that of *AhS*<sub>7</sub>*S*<sub>8</sub> (fig. 6B). We assumed that group 2 is necessary for degrading all the toxic *S*-RNase in the population which the sequenced individual grew.

Epigenetic modifications are believed to be crucial in controlling gene expression (Gibney and Nolan 2010). To determine whether methylation level is responsible for the difference between the two groups, we compared the DNA methylation in the pollen from *AhS*<sub>7</sub>*S*<sub>8</sub>. We calculated the average DNA methylation rates for the *SLF* gene regions and found that DNA methylation level in the gene body is higher than in flanking regions, and the highest methylation rate of mCs was in the CpG context, followed by CHG and CHH (fig. 6C). The average methylation rates of two groups in the CpG, CHG, and CHH are comparable in the gene body and flanking regions. Thus, the similar levels and patterns of methylation of two groups of *SLFs* may not significantly explain the variance in expression. Therefore, the difference may be caused by other kinds



**FIG. 6.** Allelic expression and methylation profile of *SLFs* for *AhS<sub>7</sub>S<sub>8</sub>* of *Antirrhinum hispanicum*. (A) The distribution of allelic divergence of genes. The divergence was measured using Levenshtein edit distance. (B) Allelic expression heatmap of *SLFs* in three genotypes. Every two columns represent allelic gene pairs of two *S*-haplotypes (HF, pollen; HZ, pistil). The *SLFs* start with “p” indicated pseudo-F-box genes lack of F-box domain or F-box protein interaction domain, but they are the best hits of the counterpart F-box alleles. (C) Methylation level within group 1 (higher expressed) and group 2 (lower expressed) *SLFs* and their 2-kb flanking (upstream and downstream) regions in CpG, CHG, and CHH context. Each region was divided into ten bins of equal size, and the methylation level in each bin is displayed in the line graph.

of factors. Thus, our data show that some genes grouped into one cluster can be regulated differentially, irrespective of their relative position within the locus.

## Discussion

Sequence analyses of the *S*-genes have been extensively conducted in Solanaceae and Rosaceae (McClure 2004; Williams et al. 2014; Kubo et al. 2015; Wu et al. 2020; Vieira et al. 2021; Lv et al. 2022); however, similar studies on Plantaginaceae species are still limited (Lai et al. 2002; Qiao et al. 2004). Prior this work, the first *S<sub>C</sub>*-haplotype of cultivar *A. majus*, which conferred self-compatibility due to loss of *S-RNase*, has been reported (Li et al. 2019). However, owing to inevitable heterozygosity and difficulty in genome assemblies, research about self-incompatible *Antirrhinum* *S*-loci has been lagging behind. In the current study, we utilized multiple sequencing technologies and the genetic strategy to obtain haplotype-resolved chromosome-level genomes of the SI *A. hispanicum*. Along with high-quality genomes of other two Plantaginaceae species, *A. linkianum* ( $2n = 16$ ) and *M.*

*orontium* ( $2n = 16$ ), these assemblies would provide an opportunity to understand the genome evolution of *Antirrhinum* lineage.

Genetic architecture and the evolutionary history of supergenes have long fascinated biologists. Especially, benefited from the long-range sequencing information, three complete *Antirrhinum* *S*-haplotypes consisting of dozens of *S*-genes were successfully reconstructed in this study. The contiguity and completeness of the genome assemblies, together with solid annotation, gave rise to the possibility to identify all the *S*-genes in *Antirrhinum* and propose a model of the origin of the most widespread type-1 *S*-locus. We annotated 32 *SLFs* and an *S-RNase* in every *S*-haplotype from self-incompatible *Antirrhinum* species. Our phylogenetic observation revealed 31 types of *SLF* in *Antirrhinum*, and genetic exchange among allelic *SLFs* may contribute to generation of new *SLF* types, although the frequency seemed to be much lower than that observed in *Petunia* (Kubo et al. 2015). *Antirrhinum* underwent the Plantaginaceae-specific WGD, but the extant *SLFs* were mainly proliferated via retrotransposon-mediated tandem and proximal duplication at a certain

time in the past, and the flanking functional genes of the *S*-locus are still expanding gradually. These duplicates were necessary intermediate toward targeting to a larger repertoire of *S*-RNase and had adaptive advantages in mating with potential partners.

Our results indicated that *SLF* members within this gene cluster appear to share common *cis*-regulatory elements, ensuring a spatially coordinated expression of protein products to degrade nonself *S*-RNase protein product. The physical position and pollen-specific expression pattern of *AnH01G33501* suggested that it may be a candidate TF regulating downstream *SLFs* expression. Interestingly, in other literatures, the TF is called *RAD* which get involved in the generation of floral dorsoventral asymmetry (Corley et al. 2005; Baxter et al. 2007). Our work here revealed the potential role in controlling self-incompatibility and hidden pleiotropic effect of the *RAD* gene. Also, it implied neighboring genes surrounding the *S*-locus may be more important and complex than previously thought. Moreover, two miRNAs were considered to act as regulators to control the expression of this TF and play a vital role in controlling self-incompatibility (Finnegan et al. 2011). In the future, it should be possible to affirm the inferences with more experimental evidence and epigenetic sequencing data.

Through comparing the *Ka/Ks* values of *SLFs* and non-*S*-locus *F*-boxes, we found that the *SLFs* were younger and faced higher selection pressure than their paralogs in snapdragon. With *Ks* estimation as a proxy of time, we found that ancestral *SLFs* had an ancient origin colliding with the emergence of *S*-RNase-based SI system reported in previous studies (Ilgic and Kohn 2001; Zhao et al. 2022). During the long evolutionary process, different ancestral *SLFs* faced different fate (deletion, expansion, contraction, and pseudogenization) in different lineages. Frequent and rapid acquisition of new *SLFs* was adapted to the allelic diversity of *S*-RNase (Xue et al. 1996) and contributed to the complexity of the extant *Antirrhinum* *S*-locus supergene. Thus, the snapdragon *S*-locus supergene has the charm of both old and young. Still, the timing of origination of *S*-RNase is undetermined, which were given to the following two possible explanations: (1) *RNase* emerged as an *S*-gene before the diversification of eudicots, and subsequent gene loss (e.g., *C. alba* scaffold124) or translocation (*P. pubescens* Chr03) of *S*-RNase occurred in many evolutionary entities, and (2) *RNase* have not come into tight linkage with incipient *SLFs* to form a functional *S*-locus which controlled self-incompatibility in the common ancestor of eudicots until the divergence of Plantaginaceae species. By comparing *Ks* values of ancient *SLFs* and *S*-RNase in Rosoideae species, Lv et al. concluded that ancestral *SLFs* originated before the split of grape and Rosaceae ancestor and after the split, the ancestral *S*-RNase emerged in the Rosaceae common ancestor (Lv et al. 2022). Based on these results and principle of parsimony, hypothesis (1) would be the most plausible model to illustrate the origin and evolution of the *S*-locus supergene. The *RNase* along with several ancestral *SLFs* formed the

pro-type of the extant functional *S*-locus structure in the common ancestor of eudicots.

Structural variations such as translocations and inversions are considered drivers of genome evolution and speciation (Bowers et al. 2003; Dodsworth et al. 2015). These rearrangements have likely contributed to the divergence of these species because they can lead to reproductive isolation if they are large enough to interfere with chromosome pairing and recombination during meiosis (Lamichhaney et al. 2016). By comparing sequence of two *S*-haplotypes, we speculated such an inversion at this locus may be involved in recombination suppression, thus maintaining the integrity and specificity of a haplotype. In *Pyrus pyrifolia*, sequence comparison of neighbor region of *S*<sub>2</sub>- and *S*<sub>4</sub>-RNase revealed that different *S*-haplotypes can show significant variation in the position and orientation of the *SFBBs* relative to the *S*-RNase (Okada et al. 2011). Interspecific and intraspecific of *Antirrhinum* *S*-haplotype comparison revealed continuous dynamic changes exist in the *S*-locus supergene, and the hyper-divergence may imply the cause of reproductive isolation and speciation.

It is of note that, based on the transcriptome data sets, the expression levels of 2 groups of 32 *SLF* genes in both *S*<sub>7</sub>- and *S*<sub>8</sub>-haplotypes showed certain fluctuation which is not caused by methylation level differences. The mechanism behind the expression variation and their roles in determining pollen specificity is not clear yet. In *Petunia*, the number of *SLF* types is less than that of *S*-RNase alleles, therefore ruling out one-to-one interaction between an *SLF* type and an *S*-RNase allele. With empirical interaction data of *SLF* and *S*-RNase (Kubo et al. 2010), Kubo used model simulation to infer that 16–20 expressed *SLFs* in 1 haplotype are saturated for recognizing ~40 allelic *S*-RNase targets (Kubo et al. 2015). However, we failed to conduct such analyses in *Antirrhinum* due to the lack of adequate data. Still, the findings in this study enriched a better understanding of the *S*-locus from a genome-wide perspective.

Ultimately, our study provides abundant genomic resources for insights on the ornamental flowering plant *Antirrhinum*. A combination of advanced sequencing technologies, comparative genomics, and multi-omics data sets will help to further decipher the *S*-locus evolution, thereby expediting the processes of horticulture and genetic research in the future. With more plant whole-genome sequences available, relationships between different SI systems will also become comparable easily, and the mystery of how this supergene evolved would be uncovered.

## Materials and Methods

### Sample Collection, Library Construction, and Sequencing

All plant samples were grown and collected from the green house. *Antirrhinum hispanicum* line *AhS*<sub>7</sub>*S*<sub>8</sub> originated from Spain and was obtained from the Gatersieben

collection in Germany and maintained by vegetative propagation (Hammer et al. 1990). *Antirrhinum linkianum* was collected and identified from Cape St. Vincent near Sagres in Western Algarve, Portugal, by Dr Rosemary Carpenter in 2001 while on a botanical walking holiday. *Misopates orontium* was collected from Sacavem near Lisbon in Portugal before 1970. Two F1 progenies (*AhS<sub>7</sub>S<sub>C</sub>* and *AhS<sub>8</sub>S<sub>C</sub>*), derived from an interspecific cross between *AhS<sub>7</sub>S<sub>8</sub>* and *A. majus* (*S<sub>C</sub>S<sub>C</sub>*) (Xue et al. 1996; Li et al. 2019), were also used for library construction and high-throughput sequencing (for more details, see [supplementary information, Supplementary Material](#) online).

### Genome Assembly, Evaluation, and Annotation of Three Species

For *AhS<sub>7</sub>S<sub>8</sub>*, three chromosome-level assemblies of *AhS<sub>7</sub>S<sub>8</sub>* were generated by utilizing sequencing data from multiple platforms, and we named them consensus assembly SI-Ah (the mosaic one), *S<sub>7</sub>* haplome, and *S<sub>8</sub>* haplome encompassing *S<sub>7</sub>*-haplotype and *S<sub>8</sub>*-haplotype, respectively ([supplementary fig. S5, Supplementary Material](#) online). Contig-level genomes of *A. linkianum* and *M. orontium* were assembled. Repetitive sequences in the assembled genomes were identified using the same pipeline as described in [supplementary materials, Supplementary Material](#) online. The gene prediction was performed using BRAKER2 (Brůna et al. 2021) annotation pipeline with parameters “–etpmode –softmasking,” which integrates RNA-seq data sets and evolutionarily related protein sequences, as well as ab initio prediction results. The predicted genes with protein length < 120 or transcripts per million (TPM) < 0.5 in all RNA-seq samples were removed. The gene boundary of all the SLFs included in the S-locus was examined in Genome browser and manually curated ([supplementary table S15, Supplementary Material](#) online). Pseudogenes were annotated based on the homology search with the predicted protein sequences using PseudogenePipeline (Zou et al. 2009). To perform functional annotation, protein sequences were searched against the multiple protein databases as described in [supplementary materials, Supplementary Material](#) online. The full details about genome assemblies, quality evaluation, and gene prediction can be found in the [supplementary materials, Supplementary Material](#) online.

### Whole-Genome Duplication and Intergenomic Analysis

Syntenic blocks (at least five genes per block) of *A. hispanicum* were identified using MCscan (Python version) with default parameters. Intragenome syntenic relationships were visualized using Circos (Krzywinski et al. 2009) in [figure 1](#). We also compared *A. hispanicum* genome with several plant genomes, including *Salvia miltiorrhiza*, *Solanum lycopersicum*, *Aquilegia coerulea*, and *Vitis vinifera*. Dotplots for genome pairwise synteny were generated using the command “python -m jcvi.graphics.dotplot –cmap = Spectral –diverge = Spectral.”

For the Ks plots, we used wgd package (Zwaenepoel and Van De Peer 2019) with inflation factor of 1.5 and maximum likelihood estimation times of 3 for reciprocal best hits search and Markov Cluster Algorithm clustering. The Ks for paralogs and orthologs were calculated using codeml program of PAML package (Yang et al. 2007). We plotted output data from result files using custom Python scripts with Gaussian mixture model function to fit the distribution (one to five components) and determined peak Ks values. Estimated mean peak values were used for dating WGD events.

### Comparative Genomics and Divergence Time Estimation

Orthologous genes relationships were built based on the predicted proteomes deprived from consensus assembled *A. hispanicum* and other angiosperm species listed in [supplementary table S11, Supplementary Material](#) online using OrthoFinder2 (Emms and Kelly 2019). Only the longest protein sequences were selected to represent each gene model. Rooted species tree inferred by OrthoFinder2 using STRIDE (Emms and Kelly 2017) and STAG algorithm was used as a starting species tree for downstream analysis. The species divergence time was estimated using MCMCtree in PAML (Yang et al. 2007) package with branch lengths estimated by BASEML, with *Amborella* as outgroup. The Markov chain Monte Carlo (MCMC) process consists of 500,000 burn-in iterations and 400,000 sampling iterations. The same parameters were executed twice to confirm the results were robust. Species divergence time for *Amborella trichopoda*–*Arabidopsis thaliana* (173–199 Mya), *V. vinifera*–*Petunia axillaris* (111–131 Mya), and *S. lycopersicum*–*S. tuberosum* (5.23–9.40 Mya) which were obtained from TimeTree database (Kumar et al. 2017) was used to calibrate the estimation model and constrained the root age <200 Mya.

### Gene Expression Analysis

Besides aiding gene model prediction, RNA-seq data sets were also used for quantification of transcripts. For expression analysis, we used STAR (Dobin et al. 2013) to map clean RNA-seq reads to reference genomes with parameters “–quantMode TranscriptomeSAM –outSAMstrandField intronMotif –alignIntronMax 6,000 –alignIntronMin 50.” The TPM values were obtained using expectation maximization tool rsem (Li and Dewey 2011). The reproducibility of RNA-seq samples was assessed using the Spearman correlation coefficient. Samples from the same tissue display a strong correlation with a Pearson’s correlation coefficient of  $r > 0.85$  ([supplementary fig. S16, Supplementary Material](#) online), indicating good reproducibility. Due to haplotype-resolved assembly and the gene structure annotation of the *A. hispanicum* genome, allelic genes from the same locus can be identified using a synteny-based strategy and the identity-based method. Reciprocal best hits between two haplomes were identified using MCSCAN (Tang et al. 2008) at first, and then, genes not in the syntenic

block were searched against coding sequence counterparts to fill up the allelic relationship table. We used MUSCLE (Edgar 2004) to align coding sequences of two allelic genes and calculated Levenshtein edit distance to measure allelic divergence based on the alignments. The divergence rate was defined as the number of edit distances divided by the total length of aligned bases. Note that we used the Python editdistance library (<https://github.com/roy-ht/editdistance>) to implement the string distance calculation.

Allelic transcripts quantification was conducted using cleaned RNA-seq data sets and the allele-aware tool kallisto v.0.46.0 (Bray et al. 2016). We applied this software to obtain the read counts and TPM of allelic genes from both haplotypes. Differential expression analysis of allelic genes was performed using R package edgeR (Robinson et al. 2010). Cutoff criteria of allele imbalanced expressed genes were set as adjusted  $P < 0.01$ , false discovery rate  $< 0.01$ , and  $|\log_2(\text{FC})| > 1$ .

### Phylogenetic Analyses of Genes

By searching Interproscan annotation results, protein sequences annotated with both PF00646 (F-box domain) and PF01344 (Kelch motif) or PF08387 (FBD domain) or PF08268 (F-box associated domain) were considered as S-like SLF genes (SLFL). The RNase T2 domain (PF00445) was used to identify candidate S-RNase gene, and the SLFL genes located in the S-locus region of snapdragon were considered potential SLFs. Organ-specific protein and paralogs were selected using PF10950 as the keyword whereas PF02298 for plastocyanin gene. Protein sequence alignments were constructed using MUSCLE (Edgar 2004) and manually checked. The ML phylogenetic gene trees were constructed using raxml-ng (Kozlov et al. 2019) with 100 replicates of bootstrap and parameter “-model JTT + G.” The gene trees were submitted to iTOL web server for visualization and annotation (Letunic and Bork 2021).

### Duplicated Gene Pair Identification and Ka and Ks Calculation

Paralogous genes were classified into different categories (WGD, TD, PD, TRD, and DSD) using DupGen\_finder pipeline (Qiao et al. 2019) with parameters “-d 15.” The Ka and the Ks of gene pairs were calculated using codeml program of PAML package (Yang et al. 2007).  $K_s = 0$  may lead to misinterpretation of genes with positive selection because of unlimited Ka/Ks ratio; therefore, we eliminated the comparisons with  $K_s = 0$ .

### Cis-Regulatory Element Analysis of Predicted SLF Promoters

Cis-regulatory elements are specific DNA sequences that are located upstream of gene-coding part and involved in regulation of genes expression by binding with TFs. We explored the upstream 2000 bp sequences of 32 SLFs of *A. hispanicum* to discover TF binding sites using MEME tool (Bailey et al. 2009) with the following

parameters: maximum number of different motifs, 5, and minimum motif width of 6 bp. Tomtom was used for the comparison against JASPAR database (Fornes et al. 2020) of the discovered motif.

### Microsynteny Analyses of Type-1 S-Locus

We queried the PlaBi database (<https://www.plabipd.de/>) and downloaded all the genomes with available annotation gff/gtf files belonging to type-1 S-RNase-based self-incompatible Rubiaceae, Rutaceae, Solanaceae, and Rosaceae. First, whole-genome-scale synteny analyses were performed between *A. hispanicum* and all these species using MCscan (Python version) implemented in jvci package (Tang et al. 2008). Only those harboring syntenic regions with snapdragon S-locus were kept for further microsynteny. At last, *C. alba*, *C. clementina*, *P. pubescens*, and *P. davidiana* were selected to represent Rubiaceae, Rutaceae, Solanaceae, and Rosaceae, respectively. Microsynteny analyses focused on the S-locus among selected species were then performed using jvci. The allelic relationship of SLFs from four *Antirrhinum* S-haplotype was also built in jvci based on sequence alignment score.

### Whole-Genome Bisulfite Sequencing Analysis

The raw WGBS reads were processed to remove adapter sequences and low-quality bases using Trim\_galore with default parameters. The cleaned reads were then mapped to the two haplotypes using abismal command from MethPipe (Song et al. 2013) package with parameters “-m 0.02.” All reads were mapped to the chloroplast genome (normally unmethylated) of snapdragon to estimate bisulfite conversion rate. The nonconversion ratio of chloroplast genome Cs to Ts was considered as a measure of experimental error rate.

All cytosines of sequencing depth  $\geq 5$  were seen as authentic methylcytosine sites. Methylation level at every single methylcytosine site was estimated as a probability based on the ratio of methylated to the total reads mapped to that loci. Methylation level in genes and 2-kb flanking regions was determined using custom Python scripts. Gene body refers to the genomic region from start to stop codon coordinates in the gff file. Each gene and its flanking regions were partitioned into ten bins of equal size, and average methylation level in each bin was calculated by dividing the reads indicating methylation by the total reads observed in the respective bin.

### Supplementary material

Supplementary data are available at *Molecular Biology and Evolution* online.

### Acknowledgments

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences

(XDB27010302) and the National Natural Science Foundation of China (32030007).

## Author Contributions

Y.X. designed and supervised the research; L.C. provided the plant seeds; Y. Z. and Q.H. performed the experiments and prepared plant samples for sequencing; S.Z. analyzed the data; S.Z. and D.Z. contributed to data visualization; and S.Z., Y.X., and E.C. wrote the manuscript. All authors discussed the results and commented on the final manuscript.

## Data Availability

All raw genome sequencing data sets, assembled genome sequences, and predicted gene models have been deposited at the GSA database (Wang et al. 2017) in the National Genomics Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, and China National Center for Bioinformation, under accession numbers PRJCA006918 that are publicly accessible at <http://ngdc.cncb.ac.cn/gsa>. RNA-seq data sets have been deposited under accession ID CRA005238.

## Code Availability

The custom codes used in this study are available at <https://github.com/Sihuihzhu/snapdragon>.

## References

- Asquini E, Gerdol M, Gasperini D, Igic B, Graziosi G, Pallavicini A. 2011. S-RNase-like sequences in styles of coffee (Rubiaceae). Evidence for S-RNase based gametophytic self-incompatibility? *Trop Plant Biol.* **4**:237–249.
- Aydinoglu F. 2020. Elucidating the regulatory roles of microRNAs in maize (*Zea mays* L.) leaf growth response to chilling stress. *Planta* **251**:38–38.
- Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME Suite: tools for motif discovery and searching. *Nucleic Acids Res.* **37**:W202–W208.
- Baxter CE, Costa MM, Coen ES. 2007. Diversification and co-option of RAD-like genes in the evolution of floral asymmetry. *Plant Journal* **52**:105–113.
- Bowers JE, Chapman BA, Rong J, Paterson AH. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. *Nature* **422**:433–438.
- Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-Seq quantification. *Nat Biotechnol.* **34**:525–527.
- Brůna T, Hoff KJ, Lomsadze A, Stanke M, Borodovsky M. 2021. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *Nucleic Acids Res.* **3**:lqaa108.
- Cao K, Peng Z, Zhao X, Li Y, Liu K, Arus P, Fang W, Chen C, Wang X, Wu J, et al. 2022. Chromosome-level genome assemblies of four wild peach species provide insights into genome evolution and genetic basis of stress resistance. *BMC Biol.* **20**:139.
- Corley SB, Carpenter R, Copsey L, Coen E. 2005. Floral asymmetry involves an interplay between TCP and MYB transcription factors in *Antirrhinum*. *Proc Natl Acad Sci USA* **102**:5068–5073.
- Debernardi JM, Rodriguez RE, Mecchia MA, Palatnik JF. 2012. Functional specialization of the plant miR396 regulatory network through distinct microRNA–target interactions. *PLoS Genet.* **8**:e1002419.
- Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, Batut P, Chaisson M, Gingeras TR. 2013. STAR: ultrafast universal RNA-Seq aligner. *Bioinformatics* **29**:15–21.
- Dodsworth S, Leitch AR, Leitch IJ. 2015. Genome size diversity in angiosperms and its influence on gene space. *Current Opinion in Genetics & Development* **35**:73–78.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**:1792–1797.
- Emms DM, Kelly S. 2017. STRIDE: Species Tree Root Inference from Gene Duplication Events. *Molecular Biology and Evolution* **34**:3267–3278.
- Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biology* **20**:425.
- Finnegan EJ, Liang D, Wang M-B. 2011. Self-incompatibility: Smi silences through a novel sRNA pathway. *Trends in Plant Science* **16**:238–241.
- Fornes O, Castro-Mondragon JA, Khan A, Van Der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghie M, Baranašić D, et al. 2020. JASPAR 2020: update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**:D87–D92.
- Franklin-Tong NV, Franklin FC. 2003. Gametophytic self-incompatibility inhibits pollen tube growth using different mechanisms. *Trends Plant Sci.* **8**:598–605.
- Frazee LJ, Rifkin J, Maheepala DC, Grant A-G, Wright S, Kalisz S, Litt A, Spigler R. 2021. New genomic resources and comparative analyses reveal differences in floral gene expression in selfing and outcrossing *Collinsia* sister species. *G3: Genes, Genomes, Genetics* **11**:jkab177.
- Fujii S, Kubo KI, Takayama S. 2016. Non-self- and self-recognition models in plant self-incompatibility. *Nat Plants.* **2**:16130.
- Gebhardt C, Ritter E, Barone A, Debener T, Walkemeier B, Schachtschabel U, Kaufmann H, Thompson RD, Bonierbale MW, Ganai MW, et al. 1991. RFLP maps of potato and their alignment with the homoelogenous tomato genome. *Theoretical and Applied Genetics* **83**:49–57.
- Gibney ER, Nolan CM. 2010. Epigenetics and gene expression. *Heredity (Edinb).* **105**:4–13.
- Gutiérrez-Valencia J, Hughes PW, Berdan EL, Slotte T. 2021. The genomic architecture and evolutionary fates of supergenes. *Genome Biol Evol.* **13**:1–19.
- Hager ER, Harringmeyer OS, Wooldridge TB, Theingi S, Gable JT, McFadden S, Neugeboren B, Turner KM, Jensen JD, Hoekstra HE. 2022. A chromosomal inversion contributes to divergence in multiple traits between deer mouse ecotypes. *Science* **377**:399–405.
- Hammer K, Knüpfper S, Knüpfper H. 1990. Das Gaterslebener Antirrhinum-Sortiment. *Kulturpflanze* **38**:91–117.
- He Z, Feng X, Chen Q, Li L, Li S, Han K, Guo Z, Wang J, Liu M, Shi C, et al. 2022. Evolution of coastal forests based on a full set of mangrove genomes. *Nature ecology & evolution* **6**:738–749.
- Igic B, Kohn JR. 2001. Evolutionary relationships among self-incompatibility RNases. *Proc Natl Acad Sci USA* **98**:13167–13171.
- Jiao Y, Leebens-mack J, Ayyampalayam S, Bowers JE, McKain MR, McNeal J, Rolf M, Ruzicka DR, Wafula E, Wickett NJ, et al. 2012. A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**:R3.
- Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A. 2019. RAXML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**:4453–4455.
- Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA. 2009. Circos: An information aesthetic for comparative genomics. *Genome Research* **19**:1639–1645.
- Kubo K, Entani T, Takara A, Wang N, Fields AM, Hua Z, Toyoda M, Kawashima S-i, Ando T, Isogai A, et al. 2010. Collaborative non-

- self recognition system in S-RNase-based self-incompatibility. *Science* **330**:796–799.
- Kubo KI, Paape T, Hatakeyama M, Entani T, Takara A, Kajihara K, Tsukahara M, Shimizu-Inatsugi R, Shimizu KK, Takayama S. 2015. Gene duplication and genetic exchange drive the evolution of S-RNase-based self-incompatibility in *Petunia*. *Nat Plants* **1**:14005.
- Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Molecular Biology and Evolution* **34**:1812–1819.
- Lai Z, Ma W, Han B, Liang L, Zhang Y, Hong G, Xue Y. 2002. An F-box gene linked to the self-incompatibility (S) locus of *Antirrhinum* is expressed specifically in pollen and tapetum. *Plant Mol Biol*. **50**: 29–41.
- Lamichhane S, Fan G, Widemo F, Gunnarsson U, Thalmann DS, Hoepfner MP, Kerje S, Gustafson U, Shi C, Zhang H, et al. 2016. Structural genomic changes underlie alternative reproductive strategies in the ruff (*Philomachus pugnax*). *Nature Genetics* **48**:84–88.
- Letunic I, Bork P. 2021. Interactive Tree of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.* **49**:W293–W296.
- Li W, Chetelat RT. 2015. Unilateral incompatibility gene *ui1.1* encodes an S-locus F-box protein expressed in pollen of *Solanum* species. *Proc Natl Acad Sci USA* **112**:4417–4422.
- Li B, Dewey CN. 2011. RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**:323.
- Li M, Zhang D, Gao Q, Luo Y, Zhang H, Ma B, Chen C, Whibley A, Ye Z, Cao Y, et al. 2019. Genome structure and evolution of *Antirrhinum majus* L. *Nat Plants* **5**:174–183.
- Liang M, Cao Z, Zhu A, Liu Y, Tao M, Yang H, Xu Q Jr, Wang S, Liu J, Li Y, et al. 2020. Evolution of self-compatibility by a mutant sm-RNase in citrus. *Nat Plants* **6**:131–142.
- Liu Z, Zhang Y, Altaf MA, Hao Y, Zhou G, Li X, Zhu J, Ma W, Wang Z, Bao W. 2022. Genome-wide identification of myeloblastosis gene family and its response to cadmium stress in *Ipomoea aquatica*. *Front Plant Sci.* **13**:979988.
- Lu J, Luo M, Wang L, Li K, Yu Y, Yang W, Gong P, Gao H, Li Q, Zhao J, et al. 2021. The *Physalis floridana* genome provides insights into the biochemical and morphological evolution of *Physalis* fruits. *Hortic Res.* **8**:244.
- Lv S, Qiao X, Zhang W, Li Q, Wang P, Zhang S, Wu J. 2022. The origin and evolution of RNase T2 family and gametophytic self-incompatibility system in plants. *Genome Biol Evol.* **14**:evac093.
- McClure B. 2004. S-RNase and SLF determine S-haplotype-specific pollen recognition and rejection. *Plant Cell* **16**:2840–2847.
- McClure BA, Franklin-Tong V. 2006. Gametophytic self-incompatibility: understanding the cellular mechanisms involved in “self” pollen tube inhibition. *Planta* **224**:233–245.
- McClure BA, Haring V, Ebert PR, Anderson MA, Simpson RJ, Sakiyama F, Clarke AE. 1989. Style self-incompatibility gene products of *Nicotiana glauca* are ribonucleases. *Nature* **342**:955–957.
- Newbigin E, Paape T, Kohn JR. 2008. RNase-based self-incompatibility: puzzled by pollen S. *Plant Cell* **20**:2286–2292.
- Okada K, Tonaka N, Taguchi T, Ichikawa T, Sawamura Y, Nakanishi T, Takasaki-Yasuda T. 2011. Related polymorphic F-box protein genes between haplotypes clustering in the BAC contig sequences around the S-RNase of Japanese pear. *J Exp Bot.* **62**: 1887–1902.
- Otero A, Fernández-Mazueros M, Vargas P. 2021. Evolution in the model genus *Antirrhinum* based on phylogenomics of topotypic material. *Front Plant Sci.* **12**:631178.
- Ou S, Chen J, Jiang N. 2018. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* **46**:e126.
- Potente G, Léveillé-Bourret É, Yousefi N, Choudhury RR, Keller B, Diop SI, Duijsings D, Pirovano W, Lenhard M, Szövényi P, et al. 2022. Comparative Genomics Elucidates the Origin of a Supergene Controlling Floral Heteromorphism. *Molecular Biology and Evolution* **39**:274.
- Qiao X, Li Q, Yin H, Qi K, Li L, Wang R, Zhang S, Paterson AH. 2019. Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol.* **20**:1–23.
- Qiao H, Wang F, Zhao L, Zhou J, Lai Z, Zhang Y, Robbins TP, Xue Y. 2004. The F-box protein AhSLF-S<sub>2</sub> controls the pollen function of S-RNase-based self-incompatibility. *Plant Cell* **16**:2307–2322.
- Ramanauskas K, Igić B. 2021. RNase-based self-incompatibility in cacti. *New Phytologist* **231**:2039–2049.
- Robinson MD, McCarthy DJ, Smyth GK. 2010. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**:139–140.
- Sassa H, Nishio T, Koyama Y, Hirano H, Koba T, Ikehashi H. 1996. Self-incompatibility (S) alleles of the Rosaceae encode members of a distinct class of the T2/S ribonuclease superfamily. *Mol Gen Genet.* **250**:547–557.
- Sato K, Nishio T, Kimura R, Kusaba M, Suzuki T, Hatakeyama K, Ockendon DJ, Satta Y. 2002. Coevolution of the S-locus genes SRK, SLG and SP11/SCR in *Brassica oleracea* and *B. rapa*. *Genetics* **162**:931–940.
- Sawyer S. 1989. Statistical tests for detecting gene conversion. *Mol Biol Evol.* **6**:526–538.
- Sijacic P, Wang X, Skirpan AL, Wang Y, Dowd PE, McCubbin AG, Huang S, Kao TH. 2004. Identification of the pollen determinant of S-RNase-mediated self-incompatibility. *Nature* **429**:302–305.
- Sommer H, Carpenter R, Harrison BJ, Saedler H. 1985. The transposable element Tam3 of *Antirrhinum majus* generates a novel type of sequence alterations upon excision. *Molecular & general genetics: MGG* **199**:225–231.
- Song Q, Decato B, Hong EE, Zhou M, Fang F, Qu J, Garvin T, Kessler M, Zhou J, Smith AD. 2013. A reference methylome database and analysis pipeline to facilitate integrative and comparative epigenomics. *PLoS One* **8**:e81148.
- Steinbachs JE, Holsinger KE. 2002. S-RNase-mediated gametophytic self-incompatibility is ancestral in eudicots. *Mol Biol Evol.* **19**: 825–829.
- Takayama S, Isogai A. 2005. Self-incompatibility in plants. *Annu Rev Plant Biol.* **56**:467–489.
- Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. *Science* **320**: 486–488.
- Ten Hoopen R, Harbord RM, Maes T, Nanninga N, Robbins TP. 1998. The self-incompatibility (S) locus in *Petunia hybrida* is located on chromosome III in a region, syntenic for the Solanaceae. *Plant J.* **16**:729–734.
- Ushijima K, Sassa H, Dandekar AM, Gradziel TM, Tao R, Hirano H. 2003. Structural and transcriptional analysis of the self-incompatibility locus of almond: identification of a pollen-expressed F-box gene with haplotype-specific polymorphism. *Plant Cell* **15**:771–781.
- Vargas P, Carrió E, Guzmán B, Amat E, Güemes J, Guzman B, Amat E, Vargas P, Carrió E. 2009. A geographical pattern of *Antirrhinum* (Scrophulariaceae) speciation since the Pliocene based on plastid and nuclear DNA polymorphisms. *J Biogeogr.* **36**:1297–1312.
- Vieira J, Fonseca NA, Vieira CP. 2008. An S-RNase-based gametophytic self-incompatibility system evolved only once in eudicots. *J Mol Evol.* **67**:179–190.
- Vieira J, Pimenta J, Gomes A, Laia J, Rocha S, Heitzler P, Vieira CP. 2021. The identification of the Rosa S-locus and implications on the evolution of the Rosaceae gametophytic self-incompatibility systems. *Sci Rep.* **11**:3710.
- Wang Y, Song F, Zhu J, Zhang S, Yang Y, Chen T, Tang B, Dong L, Ding N, Zhang Q, et al. 2017. GSA: Genome Sequence Archive. *Genomics Proteomics Bioinformatics* **15**:14–18.
- Wang H, Wan HT, Wu B, Jian J, Ng AHM, Chung CY, Chow EY, Zhang J, Wong AOL, Lai KP, et al. 2022. A chromosome-level assembly of the Japanese eel genome, insights into gene duplication and chromosomal reorganization. *Gigascience* **11**:giac120.
- Wheeler MJ, De Graaf BHJ, Hadjiosif N, Perry RM, Poulter NS, Osman K, Vátovec S, Harper A, Franklin FCH, Franklin-Tong VE. 2009.



- Identification of the pollen self-incompatibility determinant in *Papaver rhoeas*. *Nature* **459**:992–995.
- Williams JS, Der JP, dePamphilis CW, Kao TH. 2014. Transcriptome analysis reveals the same 17 *S-locus F-box* genes in two haplotypes of the self-incompatibility locus of *Petunia inflata*. *Plant Cell* **26**:2873–2888.
- Wu GA, Prochnik S, Jenkins J, Salse J, Hellsten U, Murat F, Perrier X, Ruiz M, Scalabrin S, Terol J, *et al.* 2014. Sequencing of diverse mandarin, pummelo and orange genomes reveals complex history of admixture during citrus domestication. *Nat Biotechnol.* **32**:656–662.
- Wu L, Williams JS, Sun L, Kao TH. 2020. Sequence analysis of the *Petunia inflata* *S-locus* region containing 17 *S-locus F-box* genes and the *S-RNase* gene involved in self-incompatibility. *Plant Journal* **104**:1348–1368.
- Xue Y, Carpenter R, Dickinson HG, Coen ES. 1996. Origin of allelic diversity in antirrhinum *S* locus RNases. *Plant Cell* **8**:805–814.
- Yang Q, Zhang D, Li Q, Cheng Z, Xue Y. 2007. Heterochromatic and genetic features are consistent with recombination suppression of the self-incompatibility locus in *Antirrhinum*. *Plant Journal*. **51**:140–151.
- Zhang S, Ding F, He X, Luo C, Huang G, Hu Y. 2015. Characterization of the 'Xiangshui' lemon transcriptome by de novo assembly to discover genes associated with self-incompatibility. *Mol Genet Genomics.* **290**:365–375.
- Zhao H, Zhang Y, Zhang H, Song Y, Zhao F, Ye Z, Zhu S, Zhang H, Zhou Z, Guo H, *et al.* 2022. Origin, loss, and regain of self-incompatibility in angiosperms. *Plant Cell* **34**:579–596.
- Zhou Q, Tang D, Huang W, Yang Z, Zhang Y, Hamilton JP, Visser RGFF, Bachem CWBB, Buell CR, Zhang Z, *et al.* 2020. Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat Genet.* **52**:1018–1023.
- Zou C, Lehti-Shiu MD, Thibaud-Nissen F, Prakash T, Buell CR, Shiu SH. 2009. Evolutionary and expression signatures of pseudogenes in *Arabidopsis* and rice. *Plant Physiol.* **151**:3–15.
- Zwaenepoel A, Van de Peer Y, Hancock J. 2019. wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* **35**:2153–2155.