New genomic resources provide insights into the recent evolutionary adaptation of *Antirrhinum* lineage and multiomics investigation of *S*-locus

Sihui Zhu[†,1,2,3], Yu'e Zhang[†,4], Dongfeng Zheng[1,2,3], Lucy Copsy[5] , Enrico Coen[5], Yongbiao Xue[*,1,2,3,4]

1 National Genomics Data Center & CAS Key Laboratory of Genome Sciences and Information, Beijing Institute of Genomics, Chinese Academy of Sciences, Beijing 100101, China

2 China National Center for Bioinformation, Beijing 100101, China

3 University of Chinese Academy of Sciences, Beijing 100049, China

4 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 100101, China

5 John Innes Centre, Norwich NR47UH, UK.

† These authors contributed equally to this work.

* Correspondence: Yongbiao Xue (ybxue@genetics.ac.cn)

Competing Interest Statement: The authors declare no competing interest.

Keywords: snapdragon, self-incompatibility, genome sequencing, HiFi reads, *S*-locus, comparative genomics

Abstract

Multi-allelic *S*-locus, containing pistil *S-RNase* and dozens of *S*-locus F-box (*SLF*), is known to take responsibility for genetic control of self-incompatibility (SI) in Plantaginaceae. Loss of SI is a prerequisite for mating system transition from self-incompatible to self-compatible, which is very frequent in flowering plants. The transition to selfing is excepted to have noticeable effects on adaptive potential and genome evolution. However, there have been limited comparative studies on the evolutionary shift and *S*-locus from genome-wide perspective due to lack of high-quality genomic data to date. The genus *Antirrhinum* has long been used as model plant to study self-incompatibility, here, through integrating sequencing data from multiple platforms, we present the chromosome-level reference and haplotype-resolved genome assembly of an *A.hispanicum* line, AhS_7S_8. Comparative genomics approach revealed that the self-incompatible *A.hispanicum* diverged from its self-compatible relative *A.majus* approximately 3.6 million years ago (Mya) just after Zanclean flood. The expanded gene families enriched in distinct functional terms implied quite different evolutionary trajectory of outcrossing and selfing snapdragon, and contributed to the adaptation of snapdragon with different mating system. Particularly, we successfully reconstructed two complete *S*-haplotypes spanning ~1.2Mb, each consists of 32 *SLF*s and a *S-RNase*, as well as a pseudo *S*-haplotype from self-compatible *A.majus*. Multiple copies of *SLF*s derived from proximal

or tandem duplication via retroelement-mediated way at a very early time, announcing the ancient history of *S*-locus. Furthermore, we detected a candidate *cis*-transcription factor associated with regulating *SLF*s expression, and two miRNAs may control expression of this TF in a higher position. In general, our abundant genomic data provide excellent resources to study mating system transition of snapdragon, and facilitate the functional genomics and evolutional research of *S-RNase* based self-incompatibility system.

Introduction

Self-incompatibility (SI) is a molecular recognition system used by angiosperms to prevent inbreeding and promote outcrossing, thereby maintaining genetic diversity to increase stress tolerance and survive in the constantly changing harsh environments. Despite the fact that SI systems are widespread in flowering plants, loss of SI, often accompanied by a shift to self-crossing, has occurred many times in different lineages independently(Stebbins 1950) (Stebbins 1957) (Zhao et al. 2022). The transition is favored by selection when short-term advantages of selfing, such as reproductive assurance and lower energy use to produce selfed seed, outweighs the costs of inbreeding depression(DYER 1877). However, Stebbins claimed that self-fertilization may be an evolutionary dead end because it may result in the reduction of population genetic diversity and consequently preclude adaptation to new or changing environments in the longer term(Stebbins 1957). While the underlying evolutionary mechanism remains unclear, Stebbins suggested decreased adaptive potential would lead to higher species extinction rate, which is supported by theoretical modeling(Glémin and Ronfort 2013).

The molecular mechanisms of SI in dicotyledons are thoroughly studied. In Solanaceae, Rosaceae, Rubiaceae, Rutaceae and Plantaginaceae family, SI is controlled by a single Mendelian locus, called *S*-locus(Hiscock and McInnis 2003), with a variety of haplotypes. This locus generates two types of

products(Tao and Iezzoni 2010): pistil *S* determinant (*S*-ribonuclease with cytotoxicity) and pollen *S* determinant (dozens of pollen-specific *S*-locus F-box genes called *SLFs*). A unique *S*-haplotype is consist of specific *SLFs* and specific *S-RNase*, and successful fertilization only occurs between individuals harboring different *S*-haplotypes(McClure and Franklin-Tong 2006). The *S-RNase* can be degraded by non-self *SLF*s protein products collectively of a given *S*-haplotype(Bedinger et al. 2017). *SLFs* and *S-RNase* of different haplotypes exhibit polymorphism, and they are excepted to have co-evolved and be physically linked to maintain a stable SI system. However, the origin and evolution of *S*-locus in Plantaginaceae remain unknown.

There has been intense interest in the role of parallel molecular changes underlying the evolutionary shifts to selfing associated with loss of SI(Shimizu and Tsuchimatsu 2015). Loss-of-function or loss-of-expression of *S-RNase* can confer self-compatibility (SC) in some Solanaceae species(Golz et al. 1998) (Markova et al. 2017), which also observed in *Antirrhinum*(Yang et al. 2007). The genus *Antirrhinum* (Plantaginaceae), commonly known as snapdragon, includes more than 20 species across the Mediterranean basin(Schwarz-Sommer et al. 2003), and self-incompatible in this genus is ancestral state while *A.majus* is a SC one(Zhao et al. 2022). Thus the genus serves as an ideal system to explore the impact of mating system shifts on evolutionary processes and dynamics of adaptation.

Accurate and contiguous genome assembly is fundamental basis to

understand the genomic diversity and evolution trajectory of a species comprehensively. Yet, it is frequently constrained by heterochromatin and complex region in plant genomes. Due to its inherent heterozygosity and long-spanning in different species(Li and Chetelat 2015) (Yamane et al. 2003) (Fujimoto et al. 2006), *S*-locus has been a notorious nuisance to assemble with completeness and continuity. In this study, we assembled a high-quality chromosome-level genome reference with total length ~480Mb of self-incompatible *Antirrhinum hispanicum* (AhS_7S_8, heterozygous genotype at *S-locus*, hereafter referred to as 'SI-Ah') and improved previous-released assembly of self-compatible *A.majus* (AmS_CS_C, hereafter referred to as 'SC-Am') with help of advanced sequencing technology. Along with its cultivated sister, this outcrossing snapdragon genome could provide us an opportunity to perform comparative genomic analyses and decipher causes and proximal consequences of the self-incompatibility breakdown.

Recently, Pacific Biosciences (PacBio) introduced high fidelity (HiFi) sequencing technology to generate long reads (>15 kb) with high base accuracy comparable to Illumina data. HiFi reads has provided an opportunity to assemble haplotype-resolved genome sequences of several species with adequate sequence depth, displaying its great potential in resolving regular non-model species genome assembly. We also took the advantage of HiFi data to achieve two haploid genomes of SI-Ah, each contains a *S*-haplotype. In this study, the complete structure and sequence of *S*-locus also enabled us

to examine how and when the *SLF* genes originated unbiasedly, as well as variation between different *S*-haplotypes. Inter- and intra-specific structural variations were detected among serval *S*-haplotypes, revealing its characteristic of being highly polymorphic.

The genomic data and analysis reported in this work are of great value for understanding genetic bases of adaptive characteristic of species, and provide further insights into the evolution of SI in *Antirrhinum*. A better understanding of SI evolution allows manipulations to reduce mating restriction, with direct or undirect applications ranging from plant conservation to agriculture development(Muñoz-Sanz et al. 2020).

Results

Genome assembly and evaluation

In this study, we combined different sequencing strategies to derive *A.hispanicum* reference genome assembly: Illumina pair-ended (PE) reads, PacBio single-molecule reads, BioNano optical maps, and Hi-C library (Fig. S1 and Table S1). To characterize the basic genome feature of *A.hispanicum*, PE reads of line AhS_7S_8 were used to generate 21-mer spectrum for genome survey. The genome size was estimated to be ~473Mb and a modest heterozygosity of 0.59% (Fig. S2), consistent with reproductive mechanisms of this outcrossing species. After that, the genome was assembled with PacBio continue long reads (CLR) and circular consensus reads (CCS) using

Canu(Koren et al. 2017) and Falcon-phase(Chin et al. 2016) separately. Then the contigs were scaffolded with BioNano cmaps and Hi-C data sequentially. At the end, we obtained three sets of chromosome-level genome assemblies, a consensus assembly with mosaic sequence structure (Fig. 1A) and two haplotype-resolved assemblies each contains a copy of *S*-locus (S_7 haplome with S_7-haplotype and S_8 haplome with S_8-haplotype). Yet long stretch of gaps exists in S_8-haplotype of S_8 haplome and cannot be filled up directly with sequencing data of AhS_7S_8 due to the heterochromatic characteristics of *S*-locus. Therefore, to overcome the difficulties in assembly of the heterozygous *S*-locus, we applied a genetic approach using a cross between SI-Ah and SC-Am, generating two F1 (AhS_7S_C and AhS_8S_C). Since the F1 individuals have haploid chromosome sets from AhS_7S_8, their sequencing data were used for error correction and gap filling to improve haplome assemblies. Comparing *A.hispanicum* phased genome sequences with the consensus assembly revealed generally good alignment except some assembly gaps and inversions in few chromosomes, which might due to heterozygous region collapse in Canu assembly (Fig. S4).

In addition, with new BioNano optical maps and Hi-C data of SC-Am, we also improved previously released genome scaffolds(Li et al. 2019) into eight pseudo-chromosomes. Size of four genome assemblies were very close to the estimation from *k*-mer spectrum analysis. The length of homologous chromosomes ranged from 49.7 to 77.1 Mb (Table S3).

The completeness of genome assemblies were benchmarked using BUSCO embryophyta_odb10(Simão et al. 2015) core genesets, which include 2121 single-copy orthologous genes. The results indicated that 91.4-95.3% of conserved genes were completely recovered in the four assemblies (Fig. S6 and Table S5). S_7 haplome and S_8 haplome were combined and evaluated under the protein mode, and 98.6% (2,090) of BUSCO genes were complete, while 92.2% (1,955) being duplicated. 30,956 *Antirrhinum* EST records downloaded from NCBI (2021-04-23) were mapped to the four genome assemblies, with nearly 90% of which can be mapped (Table S5). PE reads from the DNA and RNA libraries from this study were mapped to the genome assemblies using BWA-mem(Li and Durbin 2009) and STAR(Dobin and Gingeras 2015) respectively. The mapping rate ranged from 70% to 94% (properly mapped or uniquely mapped), as listed in Table S5. In addition, LTR Assembly Index(LAI)(Ou et al. 2018) based on LTR-RTs was calculated to evaluate genome assembly continuity. The LAI values of the two haplomes exceed 20, which demonstrates a ‘golden’ standard level, and the other two (the consensus assembly of SI-Ah and SC-Am) also reached ‘reference’ grade quality. Taken together, S_7 haplome outperforms S_8 haplome on sequence continuity and the consensus assembly, but not as good as them in accuracy (Table S5).

Cumulatively, these results above suggested the reliability of all the genome assemblies. Particularly, *S*-locus of self-incompatible snapdragon which spans

more than 1.2Mb was perfectly reconstructed with a continuous BioNano molecular cmap as strong evidence (Fig. S7). The *A.hispanicum* reference and haplotype-resolved genome assemblies are, to the best of our knowledge, one of the most contiguous and complete *de novo* plant genomes covered complex *S*-locus to date.

Genome annotation

Repetitive elements accounted for more than 43% of the snapdragon genome (Table S4), and long terminal repeats (LTRs) being the largest member of transposons families, covering ~23% of the nucleus genome (Fig. 1A), of which, the *Ty3/gypsy* (SI-Ah: 12.0%; SC-Am: 12.4%) and *Ty1/copia* (SI-Ah: 11.6%; SC-Am: 12.1%) are the most abundant transposable elements (TEs). DNA transposons comprises ~12% of the *Antirrhinum* genomes (Table S4). According to theoretical modeling, if the effects of TE insertion are mostly recessive or co-dominant, increased homozygosity would result in more efficient elimination(Wright and Schoen 1999), which therefore should be excepted have lesser TE in selfing species. On the other hand, natural selection against slightly deleterious TE transposition could be comprised in selfing species, because of reduced effective population size(Charlesworth and Wright 2001). In this scenario, a transition would lead to TE content increase.

The insertion time of intact LRT-RTs in *Antirrhinum* based on the divergence

of two flanking terminal repeat sequences revealed a recent burst of LTR-RTs at 0.3 Mya, and indicated that most old LTR-RTs have been lost (Fig. S9). The younger insertions have not accumulated enough substitution in two flanking repeats, making this type of genome more difficult to assemble correctly without long-range information.

Gene models of all genome assemblies were predicted followed same analysis pipeline as described in method. For consensus assembly of SI-Ah, 42,667 protein-coding genes were predicted from the soft-masked genome, 89.3% of which could be annotated in at least one functional database, and other assemblies gave close genome annotation results (Table S7). This can be further evidence to support that all the genomes were well assembled and annotated. The average transcript and CDS length of protein-coding genes were 3,023 bp and 1,146bp, respectively. The average exon and intron lengths were 232 bp and 479 bp, with 4.8 exons per gene on average. The length distribution of CDSs, exons, and introns of snapdragon show similar pattern to other plant species (Fig. S10), which confirmed the credibility of gene prediction results. The GC content is distributed unevenly in most pseudochromosomes and tends to be higher in gene-poor regions. The density of protein-coding genes on the pseudo-chromosomes was highest in distal regions and declined in centromeric regions (Fig. 1A), suggesting that gene distribution is not random. As excepted, the pattern of gene presence is opposite to that of LTR-RT elements and methylation level. A total of 2112

transcription factors belong to 58 families were annotated, including bHLH, ERF, MYB, C2H2, NAC, and B3 families with largest number of members (Fig. S11).

Analysis of small-RNAseq libraries from four tissues identified 108 known miRNA genes of 42 families of SI-Ah and 105 miRNA genes of 42 families (details listed in Table S9 and S10 respectively). In addition, 719 transfer RNA genes (tRNAs), 311 small nuclear RNA genes (snRNAs), and 791 ribosomal RNA (rRNAs) genes for SI-Ah were predicted (Table S8).

In land plants, the chloroplast DNA is a small circular molecule but is nevertheless a vital one because it codes for at least 100 proteins involved in important metabolic process such as photosynthesis. The complete circular *A.hispanicum* chloroplast genome is a typical quadripartite structure spanning 144.8kb, and comprises of four parts: long single copy section (LSC), inverted repeats (IRB), small single copy section (SSC), and inverted repeat A (IRA). Annotation of the chloroplast genome revealed a total of 125 genes, 92 of which are protein coding genes, 6 of which are rRNA genes, and 27 of which are tRNA genes. The mitochondrial genome was 517.4kb in length, harboring at least 40 protein-coding genes, 5 rRNA genes, and 33 tRNA genes. The mitogenome also contains similar sequences of chloroplast DNA. The landscape of chloroplast and mitochondrial genomes are presented in (Fig. 1B and 1C).

Taken together, three parts of snapdragon genome (nuclear, chloroplast,

mitochondria) are of high completeness and correctness, making them a valuable resource for comparative and evolutionary studies.

Comparative genomic analysis

We employed the consensus-assembly to represent *A.hispanicum* in comparative genomic analysis to infer its phylogenetic position and estimate species divergence time. Proteome from SI-Ah and SC-Am, as well as two monocots, 16 eudicots, and Amborella (Table S11) were clustered into 41,209 orthogroups (Table S12) and used for species tree construction subsequently. Our analysis revealed that SI-Ah and SC-Am split from their common ancestor at around 3.6 Mya [95% CI=1.7-6.0 Mya], which is collided with emergence of modern Mediterranean climate and in agreement with previously reported results(Otero et al. 2021) using various datasets and methodologies. And the divergence of *Antirrhinum* genus from other members in Lamiales order occurred ~59.8 Mya (Fig. 2A) during Paleocene-Eocene transition.

Whole-genome duplication (WGD) is known as a major driving force in plant evolution, as it provides abundant genetic materials. To identify WGD events in *Antirrhinum*, we first compared the *Antirrhinum* genome with the reconstructed ancient eudicot (AEK)(Murat et al. 2017) genome, and *Vitis vinifera* genome which have underwent none large-scale genome duplication after whole-genome triplication (WGT) of AEK(Jaillon et al. 2007). Synteny

analysis of the SI-Ah with AEK and *Vitis vinifera* genomes showed synteny depths ratio of 1:6 and 1:2 respectively (Fig. S12), and large-scale segmental duplications can be observed from the self-alignment of SI-Ah (Fig. 1A). The Ks distribution of paranomes displayed two peaks (Fig. 2D), suggesting *Antirrhinum* genus have at least experienced one WGD event after the gamma-WGT. We next performed intergenomic comparisons between *A.hispanicum* and two other species, *Solanum lycopersicum* and *Aquilegia coerulea* (Fig. S13). Given the known lineage-specific WGD in *Aquilegia coerulea*(Filiault et al. 2018) and two lineage-specific WGT in *Solanum lycopersicum*(Sato et al. 2012), our results further confirmed two rounds of WGD in snapdragon.

To approximately date the WGD in snapdragon, the distributions of synonymous substitutions per site (*Ks*) among paralogous genes within a genome were examined. The distributions plots of synonymous substitutions per synonymous site for SI-Ah, *Collinsia rattanii*(Frazee et al. 2021) and *Antirrhinum linkianum* paranomes showed a distinct peak at 2.23 corresponding to the well-reported gamma WGT shared by all eudicots (Fig. 2D and S14), as well as an additional peak at ~0.78 corresponding to relatively recent duplication event shared among Plantaginaceae members. Using whole-genome triplication time as ~117 Mya(Jiao et al. 2012) and Ks-peak at 2.23, we estimated the younger WGD occurred at ~40.9 Mya, posterior to divergence of *Antirrhinum* and other Lamiales plants,

demonstrating a Plantaginaceae-specific WGD event. The *Antirrhinum* lineage neutral mutation rate was inferred using the formula $\mu=D/2T$, where D is the evolutionary distance of SI-Ah and SC-Am (peak value of Ks log-transformed distribution =0.04, Fig. 2C), and T is the divergence time of the two species (3.6Mya, 95% CI=1.7 - 6.0Mya). The neutral mutation rate for *Antirrhinum* genus was estimated to be 5.6e-9 (95% CI:3.3e-9 - 1.2e-8) substitution per site per year, which is very close to the result estimated by using genotyping-by-sequencing dataset(Otero et al. 2021).

Mating system change from outcrossing to selfing occurred many times during angiosperms evolution history(Barrett 2002). To investigate the genome evolution trajectory of two *Antirrhinum* species, gene family analysis was conducted based on orthologous relationship of SI-Ah, SC-Am and other 19 plant species (Fig. 2A). A total of 2,137 and 1,387 gene families were expanded and contracted respectively in *Antirrhinum* lineage. The results also revealed 789 (3,978 family members) expanded and 689 (1,768 family members) contracted gene families in SI-Ah, 574 (3,858 family members) expanded and 623 (1,618 family members) contracted gene families in SC-Am respectively. Among the changed families, 193 (108 expansions and 85 contractions) in SI-Ah and 213 (130 expansions and 83 contractions) in SC-Am orthogroups were rapidly evolving gene families (Fig. 2A). These rapidly evolving gene families might give us some clues about consequence of losing self-incompatibility in *Antirrhinum* lineage.

The GO enrichment analysis for 789 expanded gene families of SI-Ah indicated that the significantly enriched GO terms including transferase activity (GO:0016740), pollination (GO:0009856), recognition of pollen (GO:0048544), pollen-pistil interaction (GO:0009875) and protein kinase activity (GO:0004672) (Fig. 3C), supporting its outcrossing behaviors. While in SC-Am, enriched GO terms of expanded families mostly related to response to endogenous stimulus (GO:0009719), hormone binding (GO:0042562), and signaling transducer activity (GO:0060089) (Fig. 3D), which may contribute to resilience of self-compatible species facing biotic and abiotic stresses. According to gene emerging theory, this may suggest the self-compatible specie underwent fitness optima change(Colbourne et al. 2011), and shift of mating system make the species more sensitive to adapt external environmental changes. As we can observed, these changes can be reflected in the morphological differences between two species, the outcrossing one exhibit significantly reduced flower size (Fig. 2B). For the complete list of GO enrichment analysis results, see Table S13 and S14.

Therefore, although the two species maintain the stable genomic architecture (Fig. 3A), significant differences have already accumulated within genome respectively for adaptations during the past few million years after separation.

Besides, 87 gene families have significantly expanded in the *Antirrhinum* lineage, including auxin responsive protein (some genes of this family have been reported to regulate flower opening and closure(Ke et al. 2018)),

cytochrome P450 domains (CYP), FAD binding domain, F-box domain, major pollen allergen Bet V1 signature, NB-ARC (a domain-containing disease resistance protein), protein phosphatase 2C, and UDP-glucoronosyl and UDP-glucosyl transferase (Fig. 3B). The abundance of transposable elements across whole genome may have contributed to copy number variation of these gene families in *Antirrhinum*.

Sequence comparison of haplomes

Genomic variation is also a major source of raw materials contributing to genetic diversity, and may be an important cause of speciation(Lowry and Willis 2010). *A.hispanicum* genome harbor loads of heterozygous sites for its outcrossing nature. To provide an accurate evaluation of the divergence between two haplomes, we identified polymorphisms between eight homologous chromosome pairs (Fig. 4 and Fig. S15-21). Based on the DNA alignments of two haplotypes, 18,283 highly similar segments (95% identity with length great than 5kb) were detected, covering 35.1% of S_7 haplome and 36.2% of S_8 haplome. Between these syntenic blocks, we found 2,374,480 SNPs, 236,929 insertion, and 212,829 deletion. The overall intragenomic divergence was estimated to be 0.576%, which is consistent with the heterozygosity rate predicted using *kmer*-based approaches (Fig. S2). The small variants were examined in genome regions with different putative function (intergenic, upstream and downstream, exon, intron, and splice sites).

Regulatory regions harbored the most abundant SNVs (48.2%), followed by intergenic regions (38.4%), genic regions (13.0%). 36,857 SNP/InDels are assumed to have deleterious impact on the protein.

We also tried to detect large-scale structural variants (SVs, >50bp) between two phased genomes (S_7 haplome and S_8 haplome, Fig. 5 and Fig. S27-33) and two reference genome assemblies (SI-Ah and SC-Am, Fig. 3A). We identified 2,876 insertions, 2,888 deletions, 151 inversions and 6 translocations in the S_7-S_8 haplome comparison (Table S15), which are generally less than those identified in the SI-Ah versus SC-Am alignments. It is evident that deletions and insertions are the dominant variations, accounting for most of SVs in interspecies and intraspecies alignments. Subsequently, we analyzed the intersection of SVs with various classed of genic and intergenic functional elements. In the haplome alignments, 4,714 SVs fall into intergenic regions and intron locations, with only a few SVs overlaps with coding regions, revealing a conserved functional synteny of the snapdragon genomes. Particularly, inversion is one of the most intriguing variants and has been proved to be involved with recombination repression and reproductive isolation of animals(Ayala et al. 2013). We found 151 inversions ranging in size from 3,216bp to 3.72Mb, with a mean size of 415.9kb. In the comparison SI-Ah and SC-Am, a minor inversion spanning 238.5kb were found upstream of *S-RNase* on chromosome 8. Although amounts of SVs can be observed, it is hard to determine whether they are

individual specific or species generally. Nevertheless, those observed sequence polymorphisms could be valuable in investigation of allele-specific expression, epigenetic regulation, genome structure in the evolutionary analysis of snapdragon.

In addition, we attempted to compare three haplomes by examining the potential genomic characteristic difference, such as distribution of various TE families, GC-contents, and gene density, etc. However, no distinct pattern was found between any homologous chromosomes (Fig. S4, S18-S20).

Position, structure, and flanking genes of the *S*-locus

Identification sequence details of the *S*-locus would be the initial step towards understanding SI in *Antirrhinum*. The *Antirrhinum S*-locus was found to locate in peri-telomeric region on the short arm of chromosome 8, as opposed to sub-centromeric area in *Solanaceous* species. F-box genes are discovered throughout the entire genome, with 275 copies altogether. A cluster of 32 *SLF* (*S*-locus F-box) genes, which match to the candidate *S*-locus, located in chromosome 8 (Fig. 1A). Gapless *S*-locus of 1.25Mb in SI-Ah and pseudo *S*-locus of 804kb in SC-Am are both reconstructed with BioNano long-range molecules as support evidence (Fig. S7 and S8). Except those *SLFs*, the *S*-locus also contains G-beta repeat, matE, zinc finger, cold acclimation, auxin responsive protein, and dozens of genes with unknown function (Table S19). Additionally, the *S*-locus was surrounded by two copies of organ specific

protein at left side (OSP1, OSP2) and five copies of phytocyanin genes at right side (here named PC1, PC2, PC3, PC4, PC5). We also searched paralogs for OSP and PC gene in the snapdragon genome.

The contiguity and completeness of the genome assembly, together with solid annotation, give rise to the possibility of proposing an origin model of *S*-locus. There is no doubt that gene duplication plays a key role in evolution of such a supergene. To inspect how multiple *SLFs* were generated and proliferated, *Ks* distribution of pairwise *SLFs* and non-*SLF* F-box genes were compared, and the synonymous substitutions rates of *SLF* paralogs are much lower than non-*SLF* F-box paralogs (Fig. 4B). If all the *SLFs* originated via WGDs or large segmental duplication, their paralogs should colocalize elsewhere in the genome and have similar divergence with *S*-locus genes. On the contrary, if *SLFs* were the result of stepwise local duplications, they should exhibit less differences and there should be distant paralogs scattered elsewhere in the genome. Based on classification of gene duplication type, the two OSPs derived from WDG, while the PCs arise from tandem duplication asynchronously assuming constant substitution rate, both can be support by gene phylogenetic tree topology of paralogs (Fig. S22, S23). As for *SLFs*, 24 and 6 of them are proximal duplication and tandem duplication genes respectively, and other 2 *SLFs* are dispersed duplication derived (Table S19). The phylogenetic tree along with expression heatmap of 275 F-box genes indicated *SLFs* not only share closer relationship, but also display similar

pollen-specific expression pattern which distinguished from other F-box genes (Fig. 4A). This phenomenon has not even changed much in SC-Am after millions of years (Fig. S21). It has been proved that tandem or proximal duplicates evolved toward functional roles in plant self-defense and adaptation to dramatic changing environments(Qiao et al. 2019).

Besides, 27 of 32 *SLFs* are intronless, suggesting *SLFs* proliferate most likely through retroelement-mediated way. With the estimated *Antirrhinum* mutation rate, the peak value of *SLFs* paralogs Ks distribution (range from 1.35 to 1.40) corresponding 122Mya(Igic and Kohn 2001), indicated the *S*-locus structure is a very conserved and long-lived remnants (Xue et al. 1996) (Ramanauskas and Igić 2017) survived from long-term balancing selection (Fig. 5B).This observation comes to the conclusion that the S-allelic polymorphisms should be inherited from common ancestors and the pro cess of differentiations should begin before the currently extant species format ion. Dissection of *SLFs* and flanking genes indicate that the *S*-locus is still experiencing gene gains (Table S21).

From the intraspecies similarity matrix heatmap of *SLFs* of SI-Ah and SC-Am, we can also observe the divergence among paralogs are much higher than interspecific orthologous pairs (Fig. S24). The pairwise sequence identity of 32 *SLF* paralogs ranged from 33.1 to 77.3%. And the closer the physical distance between two SLFs, the higher the sequence similarity and identity. In addition, the values of the non-synonymous/synonymous substitutions rate

ration (Ka/Ks) of nearly all *SLFs* from different allele below one is a signature of purifying selection to maintain protein function to detoxify *S-RNase* collectively (Fig. S25).

Using MEME suit analysis, we identified a significant motif (GATCCTAXAATATCTC) located upstream of most *SLFs*, and the motif annotation implied *SLFs* are likely to be co-regulated by a MYB-related family transcription factor (Fig. 5C). By examining expression of all MYB-related family members in the 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 but barely in style tissue (Fig. 5D). Among them, *AnH01G33501.01* is located about 1.2Mb upstream of *S*-locus, making it a potential regulator involved in activating *SLFs* expression and subsequently pollen-pistil recognition. The orthologous gene *AnM01G38924.01* was also found specifically expressed in pollen of SC-Am, which suggesting the maintenance of linkage between these genes controlling SI as a genetic toolkit during the SI to SC transition. Moreover, 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 express pattern exhibited negative correlation with the TF (Fig. S26). The miR396 family member, aof-miR396b, was believed to influence GRF transcription factors

expression(Debernardi et al. 2012), while over-expression of zma-miR408b in maize were reported to respond to abiotic stress to maintain leaf growth and improve biomass(Aydinoglu 2020). The results suggest those two miRNAs may also play an important role in self-incompatibility.

S-haplotype comparative analysis

For the sake of simplicity, we use 'S_7-locus' and 'S_8-locus' to refer *S*-locus from the S_7 and S_8 haplotype contain *S-RNase* and *SLF*s respectively, and their *SLF* genes as 'S_7-*SLF*' and 'S_8-*SLF*', while S_C-locus which contains no *S-RNase* denotes ψ*S*-locus of SC-Am. There are 32 *SLFs* annotated in the region of S_7-locus (S_7-*SLF1* ~ S_7-*SLF32*) and S_8-locus (S_8-*SLF1* ~ S_8-*SLF32*) respectively. Their position in chromosome and sequences information were listed in Table S16. We also compared the S_C-locus with two *S*-haplotype of SI-Ah and *S*-haplotype of self-incompatible species *A.linkianum* (referred as Al-*S*-locus). All of the *S*-haplotype were aligned to the S_C-locus of *A.majus* to assess structural differences and to identify *S*-locus specific features (Fig. 6). Synteny plot showed structure and gene content of the flanking ends of different *S*-haplotype were well conserved. The S_C-locus has highly collinearity with S_7-locus (Fig. 6B), S_8 locus and Al-*S*-locus except a small inversion at the start of S_C-S_8 alignment (Fig. 6A and 6C). From the comparison between S_C-locus and S_7&S_8-locus, a large deletion with a length

of about 120kb including *RNase* could be observed, which involved 7 genes and 8 genes in S_7-haplotype and S_8-haplotype respectively. These genes comprise DNase I-like, zinc-binding domain, DUF4218, other ones mostly are uncharacterized proteins. This structural variation can be responsible for the loss of self-incompatibility. Whereas between S_7 and S_8 locus, a small inversion could be observed, which may involve with recombination repression, and serval neighbor genes of S_7-*RNase* have no corresponding orthologs in S_8 haplotype (Fig. 6D), suggesting divergent evolution history in different regions of *S*-locus. The relative position of *S-RNase* in Al-*S*-locus being different from *A.hispanicum*, indicated that the changing physical position *S-RNase* does not impact its function and the dynamic change of 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 may contribute to the gene transposition.

Allelic expression and DNA methylation profiling

Since haplotype-resolved assembly is available and gene order is highly consistent between two haplomes (Fig. 4), we can determine two alleles of a physical locus and investigate allelic gene expression directly. Base on synteny and homology relationship, 79,398 genes (90.6% of all predicted genes) were found to have homologs on the counterpart haplotype (see “Method”). Most allelic genes displayed low level of sequence dissimilarity

(Fig. 7A).

To understand the expression profile of allelic genes, RNA-seq datasets were analyzed using allele-aware quantification tools with combined coding sequences of two haplomes as reference. In the survey across four different tissues, most expressed loci did not exhibit large variance, only 2,504 gene pairs (3.2% of total genes, see Fig. S28) displayed significant expression imbalance between two alleles (|fold-change| ≥ 2, $p \leq 0.01$), suggesting that most alleles were unbiasedly expressed in the SI-Ah genome.

We also analyzed four WGBS libraries to profile methylation state of two haplomes. About 51.6-59.5 million PE reads were generated for each sample (Table S2). 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 too, the conversion rate higher than 99% 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 (Fig. 4), as reported in potato too(Zhou et al. 2020).

After examining *SLFs* and *S-RNase* expression of two haplotypes, we found that S_7-*RNase* and S_8-*RNase* expressed indiscriminately and fairly high in AhS_7S_8 style (Table S18), which is consistent with the situation in crab

cactus(Ramanauskas and Igić 2021), and they are nearly undetectable in leaf and petal. Those two gene only expressed in style of *Ah*S_7S_C and *Ah*S_8S_C respectively as expect, yet at higher level than *Ah*S_7S_8 style. Since the gene sequences did not change after one crossing experiment, it is very likely that *RNase* expression was changed through other mechanisms. As for the expression profiles of *SLFs* in pollen of three different genotypes, S_7-*SLFs* and S_8-*SLFs* did show an overall decline in two offspring samples but tend to be reduced generally. Those genes could be divided into two groups. Group1 consists of 4 genes, were highly expressed in pollen of *Ah*S_7S_8 with obvious tissue-specificities, but barely expressed in two hybrid progenies, and these genes displayed allelic bias. Another 28 *SLFs* in cluster 2 maintain express pattern in three genotypes, but the expression level in two hybrid progenies are slightly lower than *Ah*S_7S_8 (Fig. 6b). We assumed that group2 are necessary for degrade toxic *S-RNase*.

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 two groups, we compared the DNA methylation in the pollen from *Ah*S_7S_8. We calculated the average DNA methylation rates for the *SLFs* gene regions and found that DNA methylation in gene body are higher than flanking region 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 gene body and flanking regions. Thus, the similar levels and patterns of methylation of two groups *SLFs* may not significantly explain the variance in expression. Therefore, the difference may be caused by other kind of epigenetic factors.

Discussion

Assembling genomes of self-incompatible angiosperms has always been a challenge due to inevitable heterozygosity and large portion of repetitive sequences. In current study, we utilized multiple sequencing technologies and genetic strategy to assemble genome of self-incompatible *A.hispanicum*, and updated assembly of its cultivated relative *A.majus*. This is a vast improvement and enable us to understand genome evolution of *Antirrhinum* lineage, gaining knowledge of the proximate and ultimate consequences of losing self-incompatibility(Igic et al. 2008). Benefiting from the robust assembly, we can determine the detailed genomic landscape of *Antirrhinum*. Apart from the shared gamma whole-genome triplication in eudicots, *A.hispanicum* genome underwent a Plantaginaceae-specific WGD event dated back to the Eocene, which may have been involved in large scale chromosomal rearrangement and TEs proliferation. The phylogenomic analysis clearly reveals that *A.hispancium* split from its relative *A.maj*us during the Pliocene Epoch, coinciding with establishment of modern Mediterranean

climate in the Mediterranean basin(Vargas et al. 2009). Mating system shift is ubiquitous throughout angiosperms, the evolutionary consequences of transition have attracted extensive attention for a long time(Igic et al. 2008) (Goodwillie et al. 2010) (Willi and Määttänen 2010). The comparative genomic analysis indicated that after losing self-incompatibility, though the genomic architecture of SC-Am has not changed much, the gene contents have undergone certain fluctuations. To increased environmental tolerance, the self-compatible organism recruits alternative biological metabolic pathways to improve survival and reproduce rate, and assembled more genes to cope with various external stimuli, since it is no longer possible for self-compatible species through recombination to generate new variants to adapt volatile environments.

Owning to long-range sequencing information, complete *S*-locus consist of dozens of genes was fully reconstructed, our results indicated that *SLF* members within this gene cluster share common *cis*-regulatory elements, ensuring a spatially coordinated expression of protein products to degrade *S-RNase*. Additionally, the physical position and expression pattern of *AnH01G33501* suggested that it may be the candidate transcription factor regulates downstream *SLFs* expression and provide targets for molecular marker selection of controlling mating reproduction. Perhaps such marker can be extrapolated to other *S-RNase* based self-incompatible economic species, such as potato and pear. Moreover, two miRNAs were considered to act as

regulator to control expression of this transcription factor(Finnegan et al. 2011). Moreover, based on calculation of Ks of *SLFs*, we found that the ancestral *SLF* had an ancient origin. These findings further enriched our understanding of *S*-locus from a higher perspective, and indicated the other members at this locus are far more important than previously thought. Duplication type analysis of *SLFs* revealed that all of them originated earlier than two WGDs in Plantaginaceae, yet the flanking genes of *S*-locus are still expanding gradually. The *S*-locus has charm of both old and young.

Chromosome-level structural variation such as translocations and inversion are considered drivers of genome evolution and speciation(Dodsworth et al. 2015) (Bowers et al. 2003). 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 pairing and recombination during meiosis(Lamichhaney et al. 2015). By comparing sequence of two *S*-haplotypes, we speculated such an inversion at this locus may be involved in recombination suppression, thus maintain the integrity and specificity of every haplotype. Inter-species and intra-species *S*-haplotype comparison revealed continuous dynamic changes exist in snapdragon *S*-locus. Ultimately, our study provides unprecedented insights into the genome dynamics of this ornamental flower and offers abundant genomic resource for studies on *Antirrhinum*. A combination of advanced sequencing technologies, comparative genomics, and multi-omics datasets will help decipher the

genetic mechanisms of self-incompatibility, thereby expediting the processes of horticulture and genetic research in the future.

Materials and methods

Sample collection and library construction

All plant samples were grown and collected from warm house. For *de novo* assemblies, fresh young leaf tissues of 3 *Antirrhinum hispanicum* lines (*AhS_7S_8, AhS_7Sc,* and *AhS_8Sc)* and *Antirrhinum linkianum* were collected for DNA extraction using cetyltrimethylammonium bromide (CTAB) method. Among them, *AhS_7S_8* and *A.linkianum* are self-incompatible, whereas the other two are partial self-incompatible. PCR-free short paired-end DNA libraries of ~450bp insert size for each individual were prepared and then sequenced on an Illumina HiSeq4000 platform, following the manufacturer's instructions (Illumina).

For PacBio Sequel II sequencing, high molecular weight (HMW) genomic DNA were extracted to construct two independently SMRTbell libraries with insert size of 20kb and 15kb respectively, using SMRTbell Express Template Prep Kit 2.0 (PacBio, #100-938-900) and sequenced on PacBio Sequel II system. The latter library was sequenced to generate high-qualitied consensus reads (HiFi reads) using ccs software in PacBio analysis toolkit (https://github.com/PacificBiosciences) with default parameters. In addition, two 20kb PacBio libraries were also constructed and sequenced for *AhS_7Sc*,

AhS_8Sc and *A.linkianum*.

For BioNano optical maps sequencing, HMW genomic DNA of *AhS_7S_8* and *AmS_CS_C* was extracted using BioNanoPrepTM Plant DNA Isolation Kit (BioNano Genomics), following BioNano Prep Plant Tissue DNA Isolation Base Protocol (part # 30068). Briefly, fresh leaves were collected and fixed with formaldehyde, then followed by DNA purification and extraction. Prepared nuclear DNA was labeled using BspQ1 as restriction enzyme with BioNano Prep Labeling Kit (Berry Genomics, Beijing). The fluorescently labelled DNA was stained at room temperature and then loaded onto a Saphyr chips to scan on the BioNano Genomics Saphyr System by the sequencing provider Berry Genomics Corporation (Beijing, China). Hi-C libraries were constructed for *AhS_7S_8* and *AmS_CS_C* with protocol developed in previous study(Rao et al. 2014). Young leaves were fixed with formaldehyde and lysed, then cross-linked DNA samples were digested with restriction endonuclease enzyme DpnII overnight. After repairing sticky-ends and reversing the cross-linking, DNA fragments were purified, sheared, enriched and sequenced on Illumina NovaSeq6000 under paired-end mode. Four representative tissues, including leaf, petal, pollen, and stamen, of line *AhS_7S_8* and *AmS_CS_C*, plus pollen and style of line *AhS_7S_C* and *AhS_8S_C* were collected in the green house for RNA-seq. Two biological replicates were prepared for each tissue. Strand-specific RNA-seq libraries were prepared following the manufacturer's recommended standard protocol, and then sequenced on Illumina platform under paired-end

mode. At least 20 million paired-end reads were generated for each library. Small RNAs were separated from total RNA described above using PAGE gel, then followed adapter ligation, reverse transcription, PCR product purification, and library quality control. Samples were sequenced using Nextseq500 (50 bp single-end) to yield 20 million reads per sample. Same tissues were also used for whole-genome bisulfite sequencing. Genomic DNA was extracted and fragmented into ~450bp, then treated with bisulfite using the EZ DNA Methylation-Gold Kit according to manufacturer's instructions. The libraries were sequenced on a HiSeq4000 system and at least 50 million PE-reads for each sample were generated.

Genome Assembly

To estimate the genome size independently of assembly, we characterized the basic genome feature using high-quality Illumina short reads of *A.hispanicum*. Firstly, PE reads were trimmed using Trim_galore (v0.6.1)(Krueger 2015) with parameters "-q25 --stringency 3" to remove low-quality bases and adapters. Based on the 21-mer spectrum derived from clean data, heterozygosity rate and haploid genome length of four snapdragon lines were evaluated using software GenomeScope2(Ranallo-Benavidez et al. 2020) .

For line AhS_7S_8, a consensus assembly consist of collapsed sequences, and haplotype-resolved assemblies consist of scaffolds from phased allele, were generated. Firstly, we removed reads of poor quality (RQ < 0.8) and short

length (< 1kb) in PacBio subreads from the 20kb library before assembly. The contig-level assembly was performed on filtered subreads using Canu package (v1.9)(Koren et al. 2017) with parameters “genomeSize=550m corMinCoverage=2 corOutCoverage=200 ‘batOptions□=□-dg 3 -db 3 -dr 1 -ca 500 -cp 50’”. After that, alternative haplotig sequences were removed using purge_dups(Guan et al. 2020), only primary contigs were kept for further scaffolding. The draft genome of *A.linkianum* was assembled using same analysis process. On the other hand, to obtain haplotype-resolved genome assemblies, ~40X PacBio HiFi reads was assembled using FALCON-Unzip(Chin et al. 2016) firstly and then integrate Hi-C data to separate maternal and paternal allele using FALCON-phase(Kronenberg et al. 2018). The resulting two phased contig-level haplotigs were used for downstream analysis independently. These three contig assemblies were scaffolded into pseudo-chromosome following same pipeline described below (Fig. S1).

Low-quality BioNano optical molecular maps with length shorter than 150kb or label density less than 9 per 100kb were removed. After that, the BioNano genome maps combining with contigs were fed into BioNano Solve hybrid scaffolding pipeline v3.6 (BioNano genomics) to produce accordingly hybrid scaffolds under non-haplotype aware mode. When there were conflicts between the sequence maps and optical molecular maps, both were cut at the conflict site and assembled with parameters “-B2 -N2”. Followed by BioNano scaffolding, Hi-C data was incorporated to cluster contigs into 8 groups

subsequently. Raw Hi-C reads were removed of adapters and trimmed for low-quality bases. Clean reads were then mapped to the BioNano-derived scaffolds using BWA-mem(Li and Durbin 2009). The chromosome-level assembly was furtherly generated using the 3D-DNA analysis pipeline to correct mis-join, order, orientation, and translocation from the assembly(Dudchenko et al. 2017). Manual review and refinement of the candidate assembly was performed in Juicerbox Assembly Tools (v1.11.08)(Durand et al. 2016) for interactive correction. Then the long pseudo-chromosomes were regenerated using script “run-asm-pipeline-post-review.sh -s finalize --sort-output --bulid-gapped-map” in 3D-DNA package with reviewed assembly file as input.

So far, three chromosome-level assemblies of AhS_7S_8 were constructed, and we named them as consensus-assembly SI-Ah (the mosaic one), S_7 haplome and S_8 haplome (contains S_7 haplotype and S_8 haplotype respectively). The sequences of two haplomes were named as Chr*.1 and Chr*.2 to distinguish homologous chromosomes. After manual curation, we utilized ultra-high-depth PacBio dataset and program PBJelly(English et al. 2012) to fill gaps in haplome assemblies, followed by two rounds of polish with HiFi reads and gcpp program (https://github.com/PacificBiosciences/gcpp) in pbbioconda package. Finally, Illumina DNA-seq reads were mapped to the polished assemblies for small-scale error correction using Pilon(Walker et al. 2014).

Moreover, AmS_CS_C assembly published in previous study(Li et al. 2019) was further improved with help of new BioNano and Hi-C sequencing data generated in this study. Since AmS_CS_C is a self-compatible inbred cultivar and display highly homozygosity, this *A.majus* consensus assembly was also referred as S_C haplome.

The organellar genome of *Antirrhinum* were identified using BLASTn against the *A.hispanicum* contigs with plant plastid sequences as reference (https://www.ncbi.nlm.nih.gov/genome/organelle/). After circularizing the most confidential hits, we reconstructed a complete circular mitochondrial genome of 517,247bp and a circular chloroplast genome of 144,858bp. The organellar genome sequences were submitted to CHLOROBOX(Tillich et al. 2017) website to annotate automatically and visualize genome maps.

Assemblies comparison and variation detection

MUMmer package(Marçais et al. 2018) was used to inspect the difference between pair-wise genome assemblies. SNV (SNPs and InDels) were called using show-snps with parameters '-rlTHC', then parsed and collapsed adjacent mutation/variation with our custom python script. While structural variation (SVs) including duplication, deletion, inversion, and translocation were called and visualized using SYRI(Goel et al. 2019) based on reciprocal alignment from nucmer and show-coords. The effect of SNVs was estimated using SnpEff(Cingolani et al. 2012), and SVs were annotated with commands

‘bcftools annotate’.

Repetitive DNA elements discovery

To identify repetitive elements within each genome sequences, firstly we used RepeatModeler(open-1.0.11)(Smit and Hubley 2015) to build a *de novo* repetitive elements library from the assembled genome sequences independently. GenomeTools suite(Gremme et al. 2013) was used to annotate LTR-RTs with protein HMMs from the Pfam database. In addition, LTR_retriever(Ou and Jiang 2018) was used to utilize output of GenomeTools and LTR_FINDER(Xu and Wang 2007) to generate whole-genome LTR annotation. Parameters of LTR_FINDER and LTRharvest were set requiring minimum and maximum LTR length of 100 bp and 7 kb.

Each of these libraries were classified with RepeatClassifier, followed by combining and removing redundancy using USERACH (https://www.drive5.com/usearch/). Then the unclassified sequences library was analyzed using BLASTX to search against transposase and plant protein databases to remove putative protein-coding genes. Unknown repetitive sequences were submitted to CENSOR (https://www.girinst.org/censor/index.php) to get further classified and annotated. *De-novo* searches for miniatures inverted repeat transposable element (MITEs) used MITE_Hunter(Han and Wessler 2010) software with

parameters “-w 2500 -n 5 -L 80 -I 80 -m 2 -S 12345678 -P 1”. Finally, all genome assemblies were soft-masked in company with the repeat library using RepeatMasker (open-4-0-9)(Smit et al. 2013) with parameters “-div 22 -cutoff 220 -xsmall”.

Gene prediction and annotation

The gene prediction was performed using BRAKER2(Hoff et al. 2019) annotation pipeline with parameters ‘--etpmode --softmasking’, which integrates RNA-seq datasets and evolutionary related protein sequences, as well as *ab initio* prediction results. Clean RNA-seq reads were aligned to the genome assemblies using STAR(Dobin and Gingeras 2015), then converted and merged to BAM format. The hint files were incorporated with *ab initio* prediction tool AUGUSTUS(Stanke et al. 2006) and GeneMark-EP(Brůna et al. 2020) to train gene models. The predicted genes with protein length < 120 and TPM < 0.5 in any RNA-seq sample were removed. The *SLF* gene models included in the *S*-locus were examined in Genome browser and manually curated.

The tRNA genes were annotated by homologous searching against the whole genome using tRNAscan-SE(v2.0)(Lowe and Eddy 1997), rRNA and snRNA genes were identified using cmscan program in INFERNAL(v1.1.2)(Nawrocki and Eddy 2013) package to search from the Rfam database. Four small-RNA sequencing datasets were collapsed, combining with known miRNA sequence

download from mirBase(v22)(Kozomara et al. 2019) as input of miRDP2(v1.1.4)(Kuang et al. 2019) to identify miRNA genes across the genome. Next, miRNA mature sequences were extracted to predict potential target gene using miRANDA (http://www.microrna.org) by searching against all predicted full-length cDNA sequences.

To perform functional analysis of the predicted gene models, protein sequences were search against the InterPro consortium databases(Blum et al. 2021) including PfamA, PROSITE, TIGRFAM, SMART, SuperFamily, and PRINTS as well as Gene Ontology database and pathways (KEGG, Reactome) databases using InterProScan (v5.51-85.0)(Quevillon et al. 2005). The protein sequences were also submitted to plantTFDB online server(Jin et al. 2017) to identify transcription factors and assign TFs families. In addition, predicted genes were also annotated biological function using diamond(Buchfink et al. 2014) to search against NCBI non-redundant protein and UniProtKB (downloaded on 21st Mar 2021, UniProt Consortium)(Bateman et al. 2021) database with an e-value of 1e-5. COG annotation of genes was performed using eggNOG-mapper(Huerta-Cepas et al. 2019) with a threshold of 10^{-5}.

Whole-genome duplication and intergenomic analysis

Syntenic blocks (at least 5 genes per block) of *A.hispanicum* were identified using MCscan(Tang et al. 2008) (Python version) with default parameters.

Intra-genome syntenic relationships were visualized using Circos(Krzywinski et al. 2009) in Fig1. We also compared *A.hispanicum* genome with several plant genomes, including self-compatible *A.majus*(Li et al. 2019), *Salvia miltiorrhiza*(Song et al. 2020), *Solanum lycopersicum*(Sato et al. 2012), and *Aquilegia coerulea*(Filiault et al. 2018), *Vitis vinifera*(Jaillon et al. 2007). Dotplots for genome pairwise synteny was 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 ML estimation times of 3 for reciprocal best hits search and MCL clustering. The synonymous substitution rate (Ks) for paralogs and orthologs were calculated using codeml program of PAML package(Yang 2007). We plotted output data from result files using a custom python script with Gaussian mixture model function to fit the distribution (1-5 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 20 angiosperm species listed in (Table S11) using OrthoFinder2(v2.4.1)(Emms and Kelly 2019). Only longest protein sequences were selected to represent for each gene model. Rooted species tree inferred by OrthoFinder2 using

STRIDE(Specie Tree Root Inference from Duplication Events)(Emms and Kelly 2017) and STAG(Species Tree Inference from All Genes)(Emms 2018) algorithm was used as a starting species tree for downstream analysis.

The species divergence time was estimated using MCMCtree(Yang and Rannala 2006) in PAML(Yang 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-Oryza sativa* (173-199 Mya), *Vitis vinifera-Petunia axillaris* (111-131 Mya) and *Solanum lycopersicum–Solanum tuberosum* (5.23-9.40 Mya) which were obtained from TimeTree database(Kumar et al. 2017) were used to calibrate the estimation model, and constrained the root age <200 Mya.

The determine the gene family expansion and contraction, orthologous genes count table and phylogenetic species tree topology inferred by OrthoFinder2 were taken into latest CAFE5(Mendes et al. 2020), which employed a random birth and death model to determine expansion and contraction in gene families of given node. Corresponding p value was provided for each node and branch of phylogeny tree and cutoff 0.05 was used to identify gene families undergo expanded or contraction significantly at a specific lineage or species. KEGG and GO enrichment analysis of expanded gene family members were performed using TBtools(Chen et al. 2020) to identify

significantly enriched terms.

Gene expression analysis

Apart from aiding gene model prediction, RNA-seq datasets were also used for transcripts quantification. For expression analysis, we used STAR(Dobin and Gingeras 2015) to map clean RNA-seq reads to reference genome of SI-Ah and SC-Am respectively with parameters '--quantMode TranscriptomeSAM --outSAMstrandField intronMotif --alignIntronMax 6000 --alignIntronMin 50'. The transcripts per million (TPM) values were obtained using expectation maximization tool rsem(Li and Dewey 2011). The reproducibility of RNA-seq samples was assessed using spearman correlation coefficient. Samples from same tissue display strong correlation with a Pearson's correlation coefficient of $r > 0.85$ (Fig. S29), indicating good reproducibility. Thus, these RNA-seq data were reliable for downstream analysis.

Owing to haplotype-resolved assembly and the gene structure annotation of *A.hispanicum* genome, allelic gene from a same locus can be identified using a synteny-based strategy along with identity-based method. Reciprocal best hit between two haplotypes were identified using MCSCAN at first, and genes not in synteny block were search 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 then calculated Levenshtein edit distance to measure allelic divergence. The divergence rate was defined as

the number of edit distance divided by the total length of aligned bases.

Allelic transcripts quantification was conducted using cleaned RNA-seq datasets and allele-aware tool kallisto v.0.46.0(Bray et al. 2016). We applied this software to obtain the expression levels read counts and TPM of genes of both haplotypes. Differential expression analysis of allelic genes was performed using R package edgeR(Robinson et al. 2009). Cutoff criteria of allele imbalanced expressed genes were set as adjusted *P* value < 0.01, false discovery rate < 0.01 and $|\log_2(FC)|>1$.

Phylogenetic analyses of genes

F-box genes in whole genome were identified by searching interproscan annotation results. 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 F-box genes. And the F-box genes located in the *S*-locus region of snapdragon were considered as potential SLFs. Organ specific protein and paralogs were selected using PF10950 as keyword, while PF02298 for plastocyanin gene. Sequence alignments were constructed using MUSCLE and manually checked. The maximum likelihood phylogenetic gene trees were constructed using raxml-ng(Kozlov et al. 2019) with 100 replicates of bootstrap and parameter '-m MF'. And duplicate gene were classified into different categories using DupGen_finder(Qiao et al. 2019) with parameters '-d 15'.

Cis-regulatory element analysis of *SLFs* 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 transcription factors (TFs). Thus, we explored the upstream 2000bp sequences of 32 *SLFs* of SI-Ah to discover TF binding sites by MEME(Bailey et al. 2009), and Tomtom was used for comparison against JASPAR database(Fornes et al. 2020) of the discovered motif.

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 whole-genome bisulfite sequencing reads were mapped to the two haplomes 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 non-conversion ratio of chloroplast genome Cs to Ts was considered ad a measure of error rate.

Each cytosine of sequencing depth ≥5 were seen as true *methylcytosines* 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 2kb flanking regions was determined using Python scripts. Gene body refers to the genomic sequence 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 total reads observed in the respective bin.

Data availability

All raw genome sequencing datasets, assembled genome sequences, and predicted gene models have been deposited at the GSA database in the National Genomics Data Center, Beijing Institute of Genomics (BIG), Chinese Academy of Sciences, under accession numbers PRJCA006918 that are publicly accessible at http://bigd.big.ac.cn/gsa. RNA-seq datasets have been deposited under accession id CRA005238.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (32030007) and the Strategic Priority Research Program of the Chinese Academy of Sciences (XDB27010302).

Conflict of interests

The authors declare no competing interests.

Author contributions

S.Z. and Y.X. designed research; Y.Z. prepared plant samples for sequencing; S.Z. analyzed the data; S.Z. and Y.X. wrote the paper. All authors discussed the results and commented on the final manuscript.

References

Ayala D, Guerrero RF, Kirkpatrick M. 2013. Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. *Evolution (N. Y).* 67:946–958.

Aydinoglu F. 2020. Elucidating the regulatory roles of microRNAs in maize (Zea mays L.) leaf growth response to chilling stress. *Planta* 251:38.

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.* 37:202–208.

Barrett SCH. 2002. The evolution of plant sexual diversity. *Nat. Rev. Genet.* 3:274–284.

Bateman A, Martin MJ, Orchard S, Magrane M, Agivetova R, Ahmad S, Alpi E, Bowler-Barnett EH, Britto R, Bursteinas B, et al. 2021. UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* 49:D480–D489.

Bedinger PA, Broz AK, Tovar-Mendez A, McClure B. 2017. Pollen-Pistil Interactions and Their Role in Mate Selection. *Plant Physiol.* [Internet]

173:79–90. Available from: https://pubmed.ncbi.nlm.nih.gov/27899537

Blum M, Chang HY, Chuguransky S, Grego T, Kandasaamy S, Mitchell A, Nuka G, Paysan-Lafosse T, Qureshi M, Raj S, et al. 2021. The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.*

Bowers JE, Chapman BA, Rong J. 2003. Unravelling angiosperm genome evolution by phylogenetic analysis of chromosomal duplication events. 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, Lomsadze A, Borodovsky M. 2020. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics Bioinforma.*

Buchfink B, Xie C, Huson DH. 2014. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods*.

Charlesworth D, Wright SI. 2001. Breeding systems and genome evolution. *Curr. Opin. Genet. Dev.* 11:685–690.

Chen C, Chen H, Zhang Y, Thomas HR, Frank MH, He Y, Xia R. 2020. TBtools: An Integrative Toolkit Developed for Interactive Analyses of Big Biological Data. *Mol. Plant* [Internet] 13:1194–1202. Available from: https://www.sciencedirect.com/science/article/pii/S1674205220301878

Chin CS, Peluso P, Sedlazeck FJ, Nattestad M, Concepcion GT, Clum A, Dunn C, O'Malley R, Figueroa-Balderas R, Morales-Cruz A, et al. 2016.

Phased diploid genome assembly with single-molecule real-time sequencing. *Nat. Methods*.

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).*

Colbourne JK, Pfrender ME, Gilbert D, Thomas WK, Tucker A, Oakley TH, Tokishita S, Aerts A, Arnold GJ, Basu MK, et al. 2011. The ecoresponsive genome of Daphnia pulex. *Science (80-.).* 331:555–561.

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, Gingeras TR. 2015. Mapping RNA-seq Reads with STAR. :1–19.

Dodsworth S, Leitch AR, Leitch IJ. 2015. Genome size diversity in angiosperms and its influence on gene space. *Curr. Opin. Genet. Dev.* [Internet] 35:73–78. Available from: https://www.sciencedirect.com/science/article/pii/S0959437X15001057

Dudchenko O, Batra SS, Omer AD, Nyquist SK, Hoeger M, Durand NC, Shamim MS, Machol I, Lander ES, Aiden AP, et al. 2017. De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Science (80-.).*

Durand NC, Shamim MS, Machol I, Rao SSP, Huntley MH, Lander ES, Aiden EL. 2016. Juicer Provides a One-Click System for Analyzing Loop-Resolution

Hi-C Experiments. *Cell Syst.*

DYER WTT. 1877. The Effects of Cross and Self-Fertilisation in the Vegetable Kingdom. *Nature* [Internet] 15:329–332. Available from: https://doi.org/10.1038/015329a0

Edgar RC. 2004. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*

Emms DM. 2018. STAG: Species Tree Inference from All Genes. *bioRxiv*.

Emms DM, Kelly S. 2017. STRIDE: Species tree root inference from gene duplication events. *Mol. Biol. Evol.*

Emms DM, Kelly S. 2019. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.*

English AC, Richards S, Han Y, Wang M, Vee V, Qu J, Qin X, Muzny DM, Reid JG, Worley KC, et al. 2012. Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS One*.

Filiault DL, Ballerini ES, Mandáková T, Aköz G, Derieg NJ, Schmutz J, Jenkins J, Grimwood J, Shu S, Hayes RD, et al. 2018. The Aquilegia genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *Elife*.

Finnegan EJ, Liang D, Wang M-B. 2011. Self-incompatibility: Smi silences through a novel sRNA pathway. *Trends Plant Sci.* 16:238–241.

Fornes O, Castro-Mondragon JA, Khan A, Van Der Lee R, Zhang X, Richmond PA, Modi BP, Correard S, Gheorghe 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.

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 (Bethesda).* [Internet] 11:jkab177. Available from: https://pubmed.ncbi.nlm.nih.gov/34014319

Fujimoto R, Okazaki K, Fukai E, Kusaba M, Nishio T. 2006. Comparison of the genome structure of the self-incompatibility (S) locus in interspecific pairs of S haplotypes. *Genetics* 173:1157–1167.

Gibney ER, Nolan CM. 2010. Epigenetics and gene expression. *Heredity (Edinb).* [Internet] 105:4–13. Available from: https://doi.org/10.1038/hdy.2010.54

Glémin S, Ronfort J. 2013. Adaptation and maladaptation in selfing and outcrossing species: new mutations versus standing variation. *Evolution* 67:225–240.

Goel M, Sun H, Jiao W-B, Schneeberger K. 2019. SyRI: finding genomic rearrangements and local sequence differences from whole-genome assemblies. *Genome Biol.* 20:277.

Golz JF, Clarke AE, Newbigin E, Anderson M. 1998. A relic S-RNase is expressed in the styles of self-compatible Nicotiana sylvestris. *Plant J.* 16:591–599.

Goodwillie C, Sargent RD, Eckert CG, Elle E, Geber MA, Johnston MO, Kalisz S, Moeller DA, Ree RH, Vallejo-Marin M, et al. 2010. Correlated evolution of mating system and floral display traits in flowering plants and its implications for the distribution of mating system variation. *New Phytol.* 185:311–321.

Gremme G, Steinbiss S, Kurtz S. 2013. Genome tools: A comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinforma.*

Guan D, Guan D, McCarthy SA, Wood J, Howe K, Wang Y, Durbin R, Durbin R. 2020. Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics*.

Han Y, Wessler SR. 2010. MITE-Hunter: A program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic Acids Res.* 38:1–8.

Hiscock SJ, McInnis SM. 2003. Pollen recognition and rejection during the sporophytic self-incompatibility response: Brassica and beyond. *Trends Plant Sci.* 8:606–613.

Hoff KJ, Lomsadze A, Borodovsky M, Stanke M. 2019. Whole-genome annotation with BRAKER. In: Methods in Molecular Biology.

Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, Mende DR, Letunic I, Rattei T, Jensen LJ, et al. 2019. EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* 47:D309–

D314.

Igic B, Kohn JR. 2001. Evolutionary relationships among self-incompatibility RNases. 2001.

Igic B, Lande R, Kohn JR. 2008. Loss of Self-Incompatibility and Its Evolutionary Consequences. *Int. J. Plant Sci.* [Internet] 169:93–104. Available from: https://doi.org/10.1086/523362

Jaillon O, Aury JM, Noel B, Policriti A, Clepet C, Casagrande A, Choisne N, Aubourg S, Vitulo N, Jubin C, et al. 2007. The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature*.

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. :1–14.

Jin J, Tian F, Yang DC, Meng YQ, Kong L, Luo J, Gao G. 2017. PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.*

Ke M, Gao Z, Chen J, Qiu Y, Zhang L, Chen X. 2018. Auxin controls circadian flower opening and closure in the waterlily. *BMC Plant Biol.* [Internet] 18:143. Available from: https://pubmed.ncbi.nlm.nih.gov/29996787

Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. 2017. Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res.*

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* [Internet] 35:4453–4455. Available from: https://pubmed.ncbi.nlm.nih.gov/31070718

Kozomara A, Birgaoanu M, Griffiths-Jones S. 2019. MiRBase: From microRNA sequences to function. *Nucleic Acids Res.*

Kronenberg ZN, Hall RJ, Hiendleder S, Smith TPL, Sullivan ST, Williams JL, Kingan SB. 2018. FALCON-Phase: Integrating PacBio and Hi-C data for phased diploid genomes. *bioRxiv*.

Krueger F. 2015. Trim Galore!: A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files. *Babraham Inst.*

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 Res.*

Kuang Z, Wang Y, Li L, Yang X. 2019. MiRDeep-P2: Accurate and fast analysis of the microRNA transcriptome in plants. *Bioinformatics*.

Kumar S, Stecher G, Suleski M, Hedges SB. 2017. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.*

Lamichhaney S, Fan G, Widemo F, Gunnarsson U, Thalmann DS, Hoeppner MP, Kerje S, Gustafson U, Shi C, Zhang H, et al. 2015. Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nat. Genet.* [Internet] 48:84–88. Available from:

http://dx.doi.org/10.1038/ng.3430

Li B, Dewey CN. 2011. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* 12.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*.

Li M, Zhang D, Gao Q, Luo Y, Zhang H, Ma B, Chen C, Whibley A, Zhang Yu'e, Cao Y, et al. 2019. Genome structure and evolution of Antirrhinum majus L. *Nat. Plants*.

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. U. S. A.* 112:4417–4422.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res.*

Lowry DB, Willis JH. 2010. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* 8.

Marçais G, Delcher AL, Phillippy AM, Coston R, Salzberg SL, Zimin A. 2018. MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* 14:1–14.

Markova DN, Petersen JJ, Yam SE, Corral A, Valle MJ, Li W, Chetelat RT. 2017. Evolutionary history of two pollen self-incompatibility factors reveals alternate routes to self-compatibility within Solanum. *Am. J. Bot.* 104:1904–

1919.

McClure BA, Franklin-Tong V. 2006. Gametophytic self-incompatibility: understanding the cellular mechanisms involved in “self” pollen tube inhibition. *Planta* 224:233–245.

Mendes FK, Vanderpool D, Fulton B, Hahn MW. 2020. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* [Internet] 36:5516–5518. Available from: https://doi.org/10.1093/bioinformatics/btaa1022

Muñoz-Sanz JV, Zuriaga E, Cruz-García F, McClure B, Romero C. 2020. Self-(In)compatibility Systems: Target Traits for Crop-Production, Plant Breeding, and Biotechnology. *Front. Plant Sci.* [Internet] 11:195. Available from: https://www.frontiersin.org/article/10.3389/fpls.2020.00195

Murat F, Armero A, Pont C, Klopp C, Salse J. 2017. Reconstructing the genome of the most recent common ancestor of flowering plants. *Nat. Genet.* 49:490–496.

Nawrocki EP, Eddy SR. 2013. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics*.

Otero A, Fernández-Mazuecos M, Vargas P. 2021. Evolution in the Model Genus Antirrhinum Based on Phylogenomics of Topotypic Material. *Front. Plant Sci.* 12.

Ou S, Chen J, Jiang N. 2018. Assessing genome assembly quality using the LTR Assembly Index (LAI). *Nucleic Acids Res.* 46:e126.

Ou S, Jiang N. 2018. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.*

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.

Quevillon E, Silventoinen V, Pillai S, Harte N, Mulder N, Apweiler R, Lopez R. 2005. InterProScan: Protein domains identifier. *Nucleic Acids Res.*

Ramanauskas K, Igić B. 2017. The evolutionary history of plant T2/S-type ribonucleases. *PeerJ* 2017.

Ramanauskas K, Igić B. 2021. RNase-based self-incompatibility in cacti. *New Phytol.* 231:2039–2049.

Ranallo-Benavidez TR, Jaron KS, Schatz MC. 2020. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.*

Rao SSP, Huntley MH, Durand NC, Stamenova EK. 2014. Article A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell* [Internet] 159:1665–1680. Available from: http://dx.doi.org/10.1016/j.cell.2014.11.021

Robinson MD, McCarthy DJ, Smyth GK. 2009. edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*.

Sato S, Tabata S, Hirakawa H, Asamizu E, Shirasawa K, Isobe S, Kaneko T, Nakamura Y, Shibata D, Aoki K, et al. 2012. The tomato genome sequence

provides insights into fleshy fruit evolution. *Nature*.

Schwarz-Sommer Z, Davies B, Hudson A. 2003. An everlasting pioneer: The story ofantirrhinum research. *Nat. Rev. Genet.* 4:655–664.

Shimizu KK, Tsuchimatsu T. 2015. Evolution of Selfing: Recurrent Patterns in Molecular Adaptation. *Annu. Rev. Ecol. Evol. Syst.* [Internet] 46:593–622. Available from: https://doi.org/10.1146/annurev-ecolsys-112414-054249

Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. 2015. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* 31:3210–3212.

Smit A, Hubley R. 2015. RepeatModeler open-1.0.

Smit A, Hubley R, Green P. 2013. RepeatMasker Open-4.0. 2013-2015 . *http://www.repeatmasker.org*.

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.

Song Z, Lin C, Xing P, Fen Y, Jin H, Zhou C, Gu YQ, Wang J, Li X. 2020. A high-quality reference genome sequence of Salvia miltiorrhiza provides insights into tanshinone synthesis in its red rhizomes. *Plant Genome*.

Stanke M, Keller O, Gunduz I, Hayes A, Waack S, Morgenstern B. 2006. AUGUSTUS: A b initio prediction of alternative transcripts. *Nucleic Acids Res.*

Stebbins GL. 1950. Variation and Evolution in Plants. Columbia University

Press Available from: https://doi.org/10.7312/steb94536

Stebbins GL. 1957. Self Fertilization and Population Variability in the Higher Plants. *Am. Nat.* [Internet] 91:337–354. Available from: http://www.jstor.org/stable/2458946

Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH. 2008. Synteny and collinearity in plant genomes. *Science (80-.).*

Tao R, Iezzoni AF. 2010. The S-RNase-based gametophytic self-incompatibility system in Prunus exhibits distinct genetic and molecular features. *Sci. Hortic. (Amsterdam).* [Internet] 124:423–433. Available from: https://www.sciencedirect.com/science/article/pii/S0304423810000415

Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S. 2017. GeSeq - Versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.*

Vargas P, Carrió E, Guzmán B, Amat E, Güemes J. 2009. A geographical pattern of Antirrhinum (Scrophulariaceae) speciation since the Pliocene based on plastid and nuclear DNA polymorphisms. *J. Biogeogr.* 36:1297–1312.

Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, Cuomo CA, Zeng Q, Wortman J, Young SK, et al. 2014. Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One*.

Willi Y, Määttänen K. 2010. Evolutionary dynamics of mating system shifts in Arabidopsis lyrata. *J. Evol. Biol.* 23:2123–2131.

Wright SI, Schoen DJ. 1999. Transposon dynamics and the breeding system. *Genetica* [Internet] 107:139. Available from: https://doi.org/10.1023/A:1003953126700

Xu Z, Wang H. 2007. LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.*

Xue Y, Carpenter R, Dickinson HG, Coen ES. 1996. Origin of allelic diversity in antirrhinum S locus RNases. *Plant Cell*.

Yamane H, Ikeda K, Ushijima K, Sassa H, Tao R. 2003. A pollen-expressed gene for a novel protein with an F-box motif that is very tightly linked to a gene for S-RNase in two species of cherry, Prunus cerasus and P. avium. *Plant Cell Physiol.* 44:764–769.

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 J.* 51:140–151.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*

Yang Z, Rannala B. 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.*

Zhao H, Zhang Yue, Zhang Hui, Song Y, Zhao F, Zhang Yu'e, Zhu S, Zhang Hongkui, 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 RGF, Bachem CWB, Robin Buell C, Zhang Z, et al. 2020. Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat. Genet.* [Internet] 52:1018–1023. Available from: http://dx.doi.org/10.1038/s41588-020-0699-x

Zwaenepoel A, Van De Peer Y. 2019. Wgd-simple command line tools for the analysis of ancient whole-genome duplications. *Bioinformatics* 35:2153–2155.

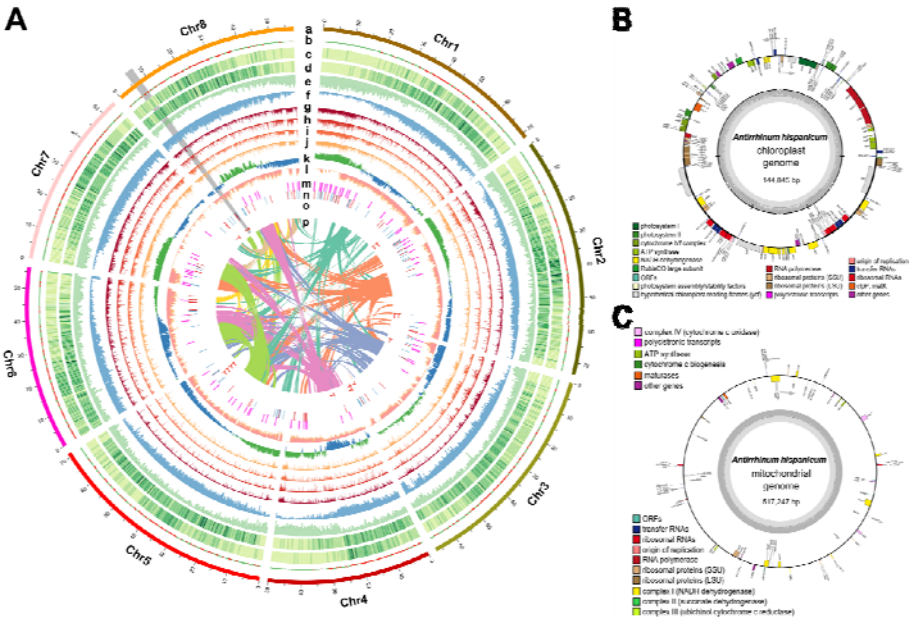

Fig.1 | Genomic landscape of *A.hispanicum*. A Characteristics of the nuclear genome, the outermost layer is circular ideogram of the eight pseudomolecules. Track b-j corresponds to the distribution of the GCcontent (red denote higher than average whole-genome level, and green vice versa), Ty3/Gypsy, Ty1/Copia retrotransposons, gene density, methylation level, expression level of four tissues (petal, pollen, style, and leaf) respectively, all calculated in 250kb non-overlap 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 location of miRNA genes. Track n corresponds to location of all F-box genes, blue and red represent forward and reverse strand respectively. Track o marker “T” corresponds to location of ribonuclease T2 genes. Track p corresponds to inter-chromosome syntenic

blocks. The grey sector marked S-locus in Chr8. **B** The circular map of *A.hispanicum* chloroplast genome. The legend describes the family of genes in the genome. The inner circle depicts nucleotide content (GCcontent in dark gray and ATcontent in light gray). Inverted repeats (IRA and IRB) along with large (LSC) and small (SSC) single copy regions are indicated on the inner circle. The outer circle shows annotated genes that are color-coded by function. Genes located on the outside of the circle are transcribed in a clockwise direction, and those on the inside of the circle are transcribed counterclockwise. **C** The circular gene map of *A.hispanicum* mitochondrial genome. Both figures of circular genomes were generated using OGDRAW.

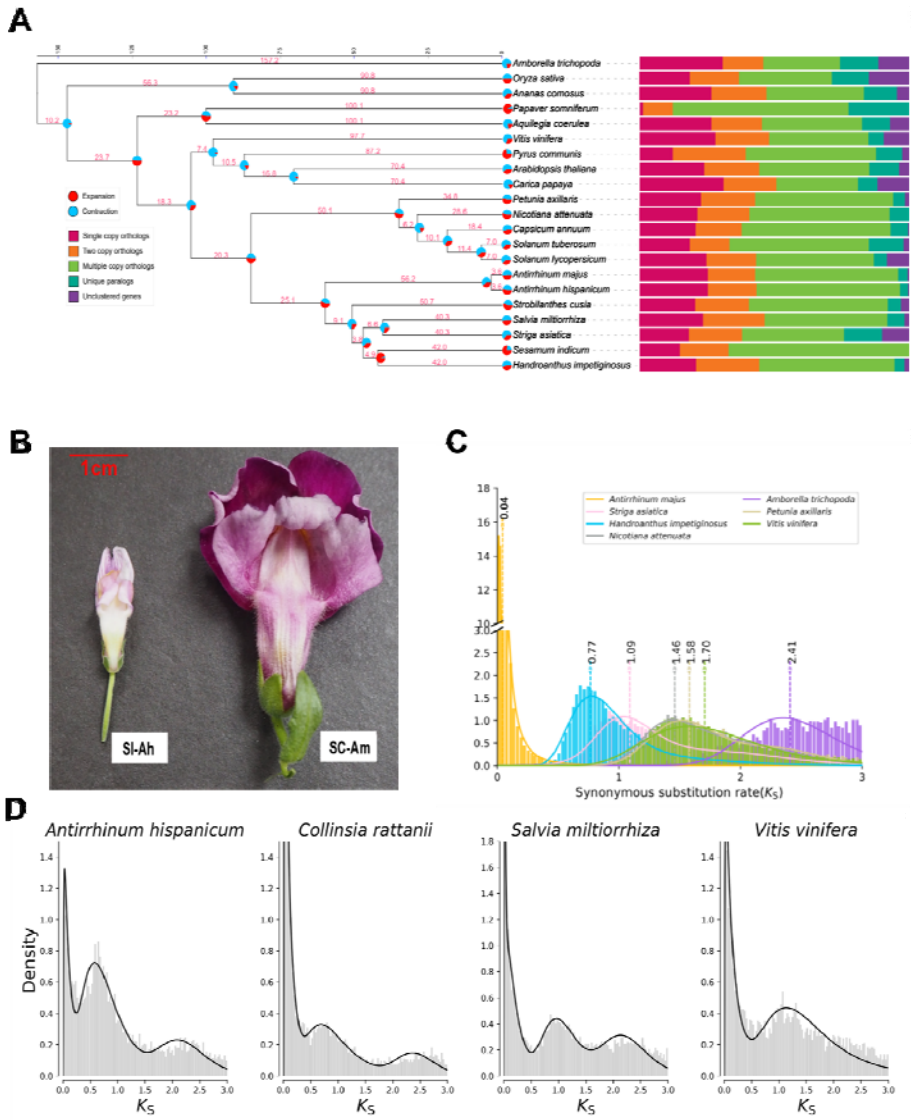

Fig.2 | Comparative genomics analysis and lineage expanded gene families of *Antirrhinum*. A Phylogeny position of the snapdragon and gene families evolution. The left panel shows phylogenetic tree of 21 species. The numbers on the branched denote scaled divergence time of each branch (MYA, million years ago). The pie diagram on each branch of the tree represents the proportion of gene families undergoing expansion (red) or

contraction (blue) events. The right panel shows the distribution of single copy, two copy, multiple copy, unique paralogs, and other orthologs in 21 plant species. **B** Flower size and shape differ between SI-Ah and SC-Am. **C** Ks distribution of one-to-one orthologs identified between SI-Ah and other species. The density curves were fitted using Gaussian mixture model, and the peak values marked with dotted dash lines. **D** Density plot of Ks distribution of paralogous gene pairs for SI-Ah, *C.rattanii*, *S.miltiorrhiza*, and grape. *C.rattanii* is another Plantaginaceae species. The solid lines were fitted using Gaussian mixture model.

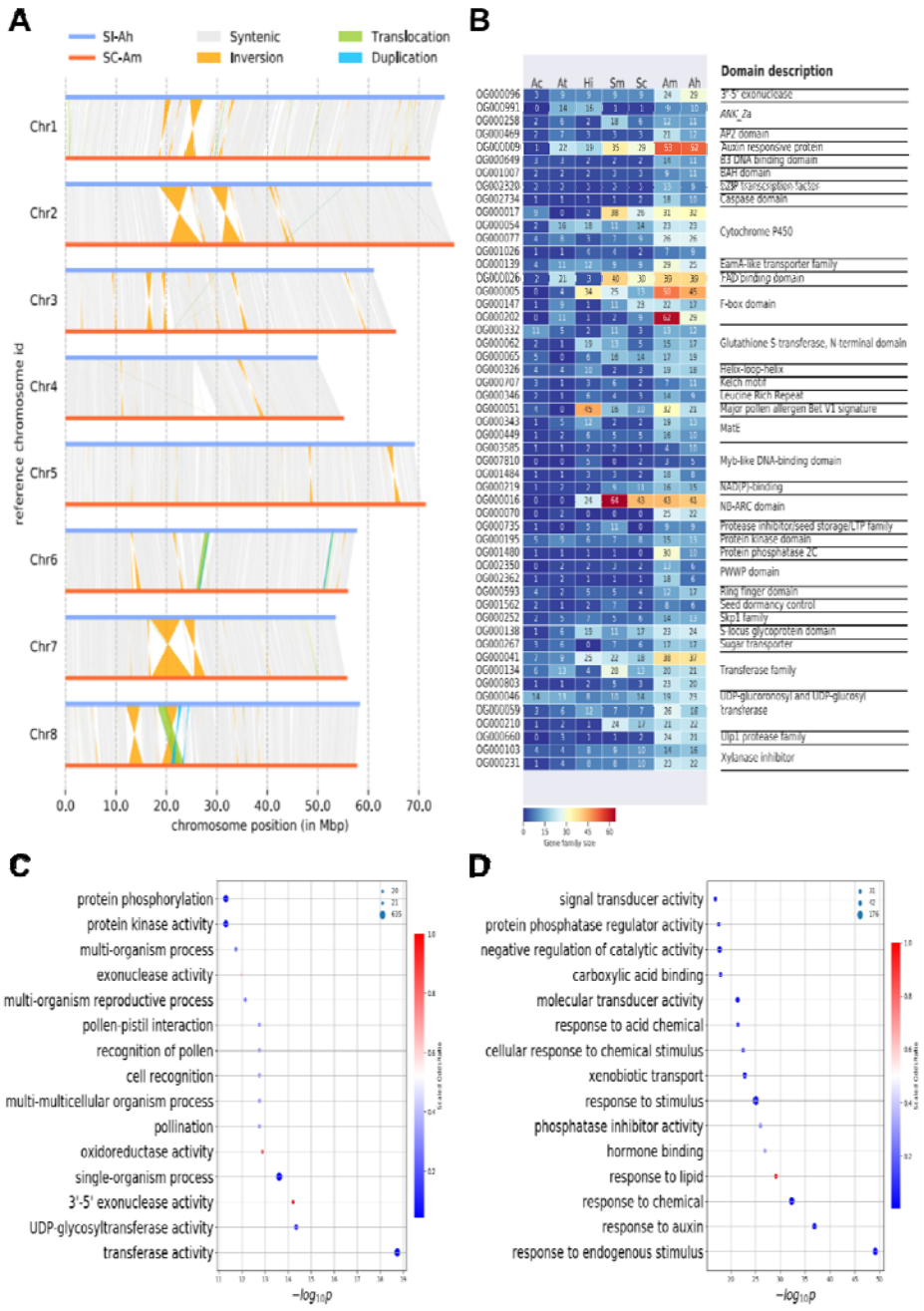

Fig.3 | Similarities and differences between SI-Ah and SC-Am. A DNA alignment between two snapdragon genomes. Structural rearrangements are represented by colored lines (orange for inversions, green for translocations, blue for duplications) connecting regions of two homologous chromosomes;

syntenic regions are connected by gray lines. **B** Functional annotation for the rapidly evolving expanded gene families in *Antirrhinum* lineage. Left panel shows orthogroup ID and the gene family size among 7 species and right panel shows the functional annotation for each gene family. Ac: *Ananas comosus*; At: *Arabidopsis thaliana*; Hi: *Handroanthus impetiginosus*; Sm: *Salvia miltiorrhiza*; Sc: *Strobilanthes cusia*; Am: *Antirrhinum majus*; Ah: *Antirrhinum hispanicum*. **C-D** Gene Ontology enrichment analysis of expanded family members in SI-Ah and SC-Am. The color of circle represents the scaled odds ratio. The size of circle represents the gene number of the GO terms.

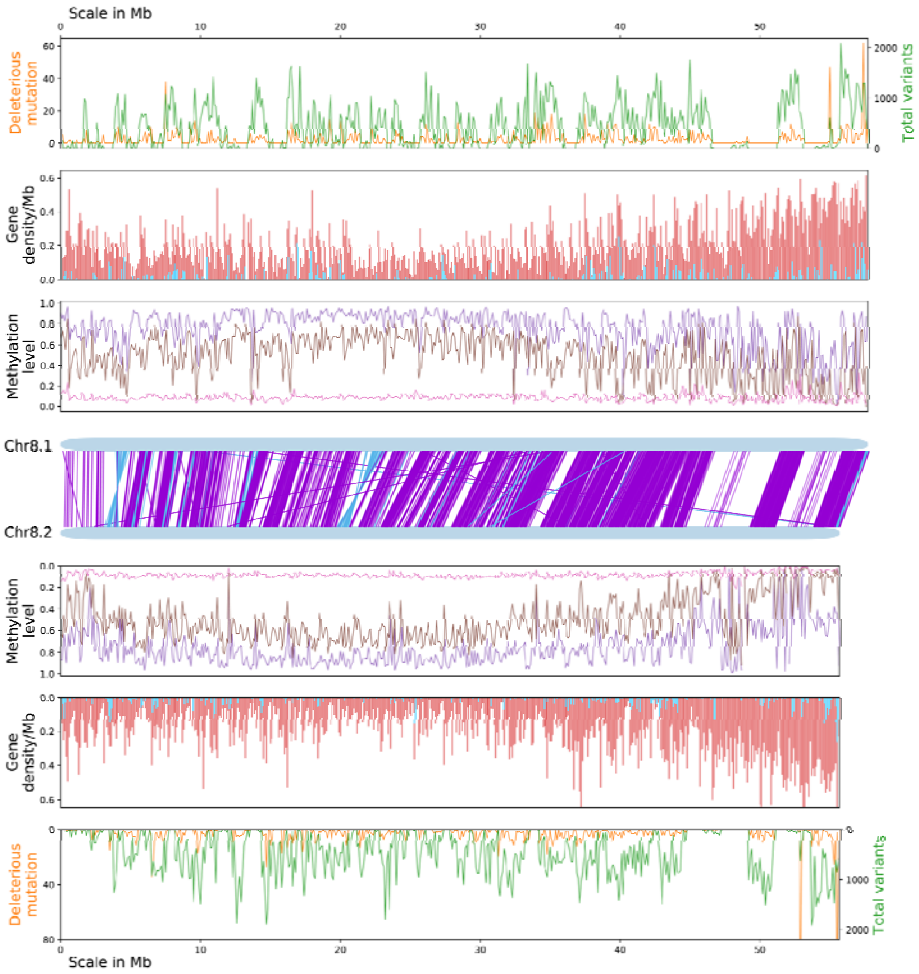

Fig.4 | Alignment of SI-Ah chromosome 8 from two haplomes. The central light blue bars represent homologous chromosome 8 with the purple and dark blue lines indicating the forward and reverse alignments. The distribution of total variants(green), deleterious mutations (orange), annotated genes (pink), imbalanced expressed alleles (blue), methylation level of three contexts are arranged symmetrically for each haplotype. The methylation level of methylated sites in CpG, CHG and CHH contexts are indicated in purple,

brown, and magenta respectively. All of the numbers were determined in 100 kb non-overlap windows.

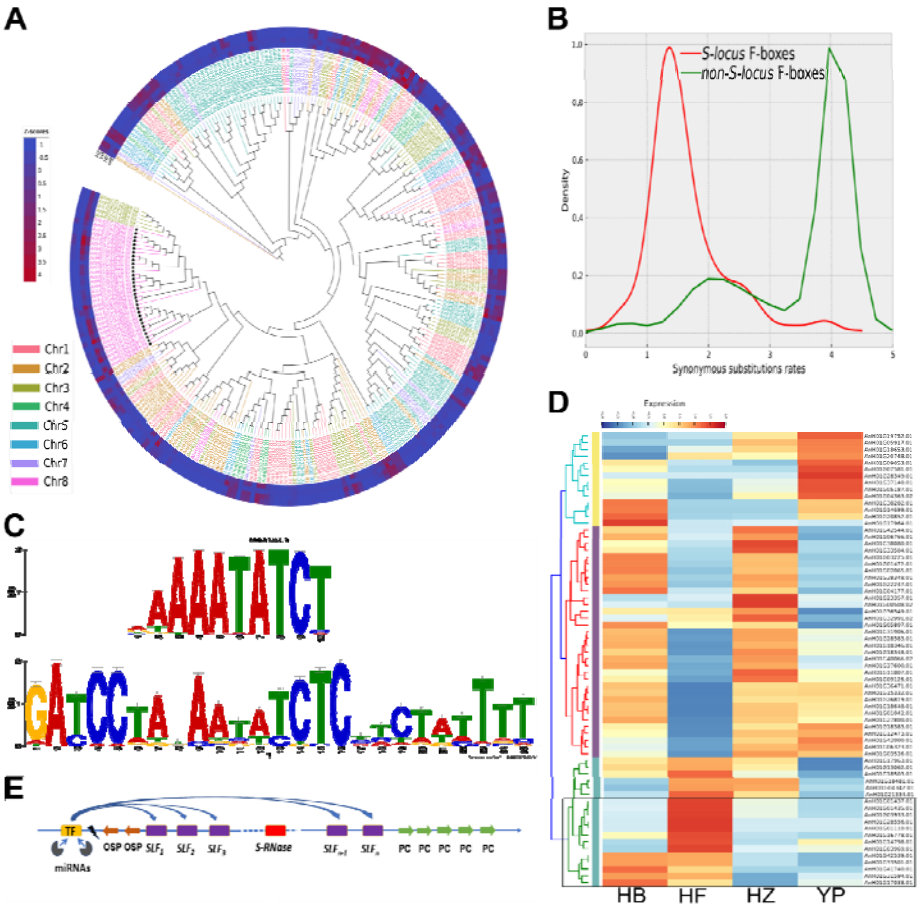

Fig.5 | Sequence analysis and candidate regulator of *SLFs*. A Phylogenetic tree reconstructed based on the amino acid sequences of all F-box genes from *A.hispanicum* genome. The maximal likelihood phylogenetic tree was calculated using raxml-ng with 1000 bootstrap. The color of gene ID denoted the chromosome the 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** The red line is synonymous substitution rate distribution of 32 *SLFs*, while the green one denotes other non-*SLF* F-box genes in the genome. **C** The bottom panel represents motif found by

scanning upstream 1000bp of 32 *SLFs* using MEME, and the top panel is a Tomtom alignment for annotation the putative transcription factor binding site. **D** The cluster heatmap of expression matrix of all MYB-related transcription factor family members in SI-Ah genome. The columns represent 4 tissues (HB: petal, HF: pollen, HZ: style, YP: leaf). The legend shows the range of log2(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 style. **E** Schematic regulatory relationship of snapdragon *S*-locus.

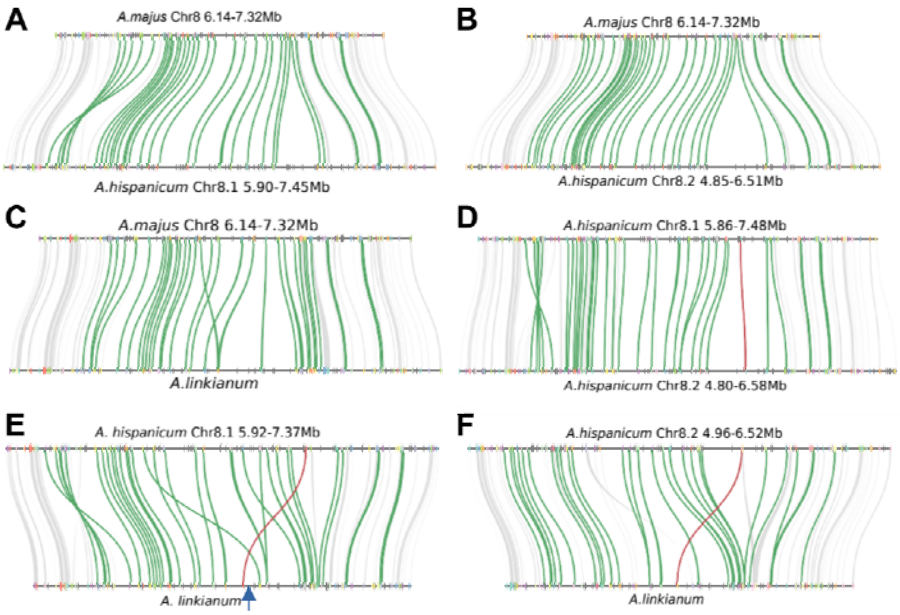

Fig.6 | The comparison of inter- and intra-specific *S*-haplotypes. A-C The micro collinearity patterns of S_C-haplotype with S_7-haplotype, S_8-haplotype, and Al-*S*-haplotype from self-incompatible *A.linkianum*. **D-F** The microsynteny of S_7-haplotype vs. S_8-haplotype, S_7-haplotype vs. Al-*S*-haplotype, and S_8-haplotype vs. Al-*S*-haplotype. The green curve lines and the red lines denote *SLFs* and *RNase* gene respectively, other genes are colored grey. The Tam downstream 11kb of *S-RNase* is marked by a blue arrow, and the TE is on the same strand with *RNase*.

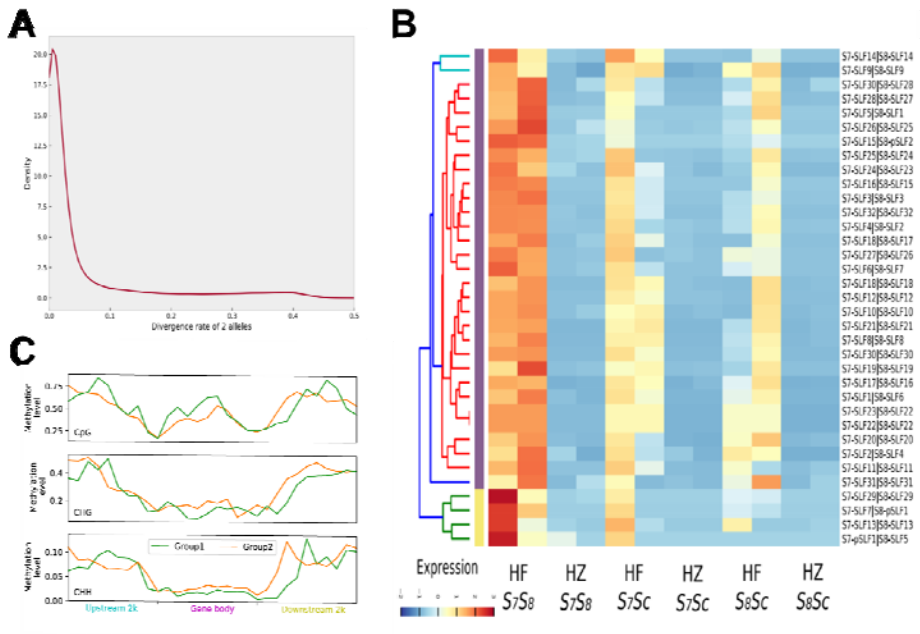

Fig.7 | Allelic expression and methylation profile of *SLFs* for line *AhS$_7$S$_8$*. **A** The distribution of allelic divergence of genes. The divergence was measured using Levenshtein edit distance. **B** Allele expression heatmap of *SLFs* in three genotypes. Every two columns represent allelic gene pairs of two *S*-haplotypes(HF: pollen, HZ: style). *SLF* starts with 'p' indicated pseudo-F-box gene lack of F-box domain or F-box protein interaction domain, but they are the best hit of the counterpart F-box allele. **C** Methylation level within group1 (higher expressed) and group2 (lower expressed) *SLFs* and their 2kb flanking (upstream and downstream) regions in CpG, CHG, and CHH context. Each region was divided into 10 bins of equal size and the methylation level in each bin is displayed in the line graph.