Genomic insights into the recent evolutionary adaptation of snapdragon and the origin of *S*-locus

Sihui Zhu[†,1,2,3], Yu'e Zhang[†,4], Lucy Copsy[5], Dongfeng Zheng[1,2,3], 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, haplotype-resolved genome assembly, HiFi reads, *S*-locus, comparative genomics

Abstract

A multi-allelic *S*-locus, containing pistil *S-RNase* and dozens of *S*-locus F-box (*SLF*), underlies genetic control of self-incompatibility (SI) in the Plantaginaceae. Loss of SI is a prerequisite for the transition from self-incompatible (SI) to self-compatible (SC), which is frequent in flowering plants and influences adaptive potential and genome evolution. However, there have been limited comparative studies on the evolutionary shift from a genome-wide perspective due to lack of high-qualitied genomic data. Snapdragon has been used as an excellent model to study self-incompatibility. Here, through integrating sequencing data from multiple platforms, we present the chromosome-level reference and haplotype-resolved genome assemblies of a self-incompatible *Antirrhinum hispanicum* line, AhS_7S_8. Alongside with draft genome of *Misopates orontium* (common name: lesser snapdragon), comparative genomics reveals that the *A.hispanicum* diverged from its self-compatible cousin *M.orontium* 12.3 million years ago (Mya). Expanded gene families enriched in distinct functional terms implied quite different evolutionary trajectory of outcrossing and selfing species. For the first time, two complete *A.hispanicum S*-haplotypes spanning ~1.2Mb and consist of 32 *SLFs* were reconstructed, while most of the *SLFs* derived from retroelement-mediated proximal or tandem duplication approximately 122 Mya. Moreover, we detected a candidate *cis*-transcription factor associated with regulating

*SLF*s expression, and two miRNAs may control expression of this transcription factor. The genome data in this study provides an excellent resource 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 that prevents inbreeding and promote outcrossing in hermaphrodites flowering plants, thereby maintains genetic diversity and help angiosperms expand into a wide range of habitats.

The molecular mechanisms of SI in eudicots have be thoroughly studied. In Solanaceae, Rosaceae, Rubiaceae, Rutaceae and Plantaginaceae family, SI is controlled by a single polymorphic *S*-locus[1], with a variety of haplotypes. The locus generates pistil *S* determinant (*S*-ribonuclease with cytotoxicity) and pollen *S* determinant (dozens of pollen-specific *S*-locus F-box genes called *SLFs*) that mediate self/non-self recognition. A unique *S*-haplotype consists of specific *SLFs* and a specific *S-RNase*. Successful fertilization only occurs between individuals harboring different *S*-haplotypes[2], and the *S-RNase* can be degraded by non-self *SLF*s protein products collectively[3]. Linked *SLFs* and *S-RNase* are thought to have co-evolved to maintain a stable SI system. However, the origin and evolution of the *S*-locus in the Plantaginaceae remain unknown.

The breakdown of SI results in a mating system change from outcrossing to self-fertilization in angiosperm evolutionary history. SI has been lost frequently during angiosperm diversification, and the transition to self-compatibility (SC) are usually accompanied by accumulation of loss-of-mutations[4].

There has been intense interest in the role of parallel molecular changes underlying the evolutionary shifts to selfing associated with loss of SI[5]. Loss-of-function or loss-of-expression of *S-RNase* can confer self-compatibility (SC) in some Solanaceae species[6] [7], which also observed in *Antirrhinum*[8] (commonly known as snapdragon). Therefore, it has long been served as an ideal system to explore the impact of mating system shifts on evolutionary processes and dynamics of adaptation. However, most previous reports focused on the studies of incompatible genes themselves, few systematic analyses of the evolution of the locus at the whole genome level, and the impact of gene duplication on the shape process of functional locus was not considered. With the advent of the post genome era, accurate and contiguous genome assembly, the critical basis for understanding genomic diversity and evolutionary trajectories, is available now. Yet, genomic region that of great biological interest remain problematic for complete assembly. Due to its high heterozygosity[11] [12] [13], the *S*-locus has always been difficult to assemble. In this study, we assembled a 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'). Also, high-qualitied draft genome of self-incompatible *Antirrhinum linkianum* and self-compatible *Misopates orontium* were present here. These sequence resources could provide us an opportunity to perform comparative genomic analyses to capture consequences of the self-incompatibility breakdown and

obtain insights into the evolutionary genomics basis of loci controlling self-incompatibility.

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 provide opportunity to assemble haplotype-resolved genome sequences of several species with adequate sequence depth[14]. We took the advantage of HiFi data to assemble two haploid genomes of SI-Ah, each contains a *S*-haplotype. In this study, the complete structure and sequence of *S*-locus allowed us to examine how and when the *SLF* genes originated, as well as variation between different *S*-haplotypes. Inter- and intra-specific structural variations were detected among several *S*-haplotypes. 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 snapdragon. A better understanding of SI evolution allows manipulation to reduce mating restriction, with direct or undirect applications ranging from plant conservation to agriculture development[15].

Results

Genome assembly and evaluation

In this study, we combined different sequencing strategies to derive *A.hispanicum* reference genome assembly: Illumina pair-ended reads, PacBio

single-molecule reads, BioNano optical maps, and Hi-C sequencing reads (Fig. S1 and Table S1). The line *A.hispanicum* AhS_7S_8 was estimated to be with genome size of ~473.4Mb and a modest heterozygosity of 0.59% (Fig. S2) based on the 21-mer spectrum. Firstly, 36.0Gb of continue long reads (CLR) and 19.1Gb of circular consensus reads (CCS) were obtained on PacBio platform for primary assembly. Then, a total of 203.3Gb BioNano data and 69.9Gb Hi-C paired-end reads were produced to scaffold the contigs into 8 pseudo-chromosomes. At the end, we obtained three sets of chromosome-level genome assemblies of AhS_7S_8, 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 AhS_7S_8 and self-compatible relative *A.majus* (S_CS_C), 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 gaps and inversions in few chromosomes, which might cause by heterozygous region

collapse in Canu assembly (Fig. S4). To facilitate comparative analyses, we also sequenced the genome of self-incompatible *Antirrhinum linkianum* and self-compatible *Misopates orontium.* The heterozygosity and genome size of *A.linkianum* is very close to that of *A.hispanicum*, while *M.orontium* exhibits heterozygosity rate of 0.21% and approximately 24.1% reduction in genome size .

The results indicated that 95.0-98.7% of conserved genes were completely recovered in these assemblies (Fig. S6 and Table S6). 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 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 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)[16] 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 draft genome of other two species also reached ‘reference’ grade quality (19.3 and 21.9 for *A.linkianum* and *M.orontium*, respectively). 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 assembled with a continuous BioNano molecular cmap as strong evidence (Fig. S7). The *A.hispanicum* 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 approximately 43% of *Antirrhinum* 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 (12.0%) and Ty1/copia (11.6%) 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[17], which therefore should be excepted have lesser TE in selfing species. As we can observed, the total amount of TE in self-compatible *M.orontium* (38.0%) is apparently lower than self-incompatible species (Table S4). 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 *A.hispanicum*, 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 2,112 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 *A.hispanicum* and 105 miRNA genes of 42 families (details listed in Table S9). In addition, 719 transfer RNA genes (tRNAs), 311 small nuclear RNA genes (snRNAs), and 791 ribosomal RNA (rRNAs) genes for *A.hispanicum* 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.

Sequence comparison of assemblies

Genomic variation is also a major source of raw materials contributing to genetic diversity, and may be an important cause of speciation[18]. *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. 2 and Fig. S27-33). 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 method (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 (Fig. S27-33) and the reference genome assembly. We identified 2,876 insertions, 2,888 deletions, 151 inversions and 6 translocations in the S_7-S_8 haplome comparison (Table S14). 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[19]. We found 151 inversions ranging in size from 3,216bp to 3.72Mb, with a mean size of 415.9kb. 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.

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 *A.hispanicum, A.linkianum*, *M.orontium* with another 16 species (Table S10) were clustered into 36,295 orthogroups (Table S11) and used for phylogenetic relationship construction subsequently. Our analysis revealed that *A.hispanicum* and *A.linkianum* split

at around 3.5 Mya [95% CI=1.6-5.3 Mya], which is collided with emergence of modern Mediterranean climate and rapid speciation of *Antirrhinum* lineage, and in agreement with previously reported results using various datasets and methodologies[20]. Before that, *M.orontium* diverged from *Antirrhinum* genus and lost self-incompatibility assuming self-compatible is ancestral state in Plantaginaceae[10]. And the divergence of Plantaginaceae from Lamiales order common ancestor occurred ~55.8 Mya during Paleocene-Eocene transition (Fig. 3A).

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)[21] genome, and *Vitis vinifera* genome which have underwent none large-scale genome duplication after whole-genome triplication (WGT) of AEK[22]. Synteny analysis of the *Antirrhinum* 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 *A.hispanicum* (Fig. 1A). The *Ks* distribution of paranomes displayed two peaks (Fig. 3B), indicating *Antirrhinum* genus have at least experienced another one WGD after the Gamma-WGT. We next performed intergenomic comparisons between *A.hispanicum* and two other species, *S.lycopersicum* and *A.coerulea* (Fig. S13). Given the known lineage-specific WGD in *A.coerulea*[23] and two lineage-

specific WGT in *S.lycopersicum*[24], 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 *A.hispanicum*, *C.rattanii*[25] and *A.linkianum* paranomes showed a distinct peak at 2.21 corresponding to the well-reported gamma-WGT shared by all eudicots (Fig. 3B and S14), as well as an additional peak at ~0.61 corresponding to relatively recent duplication event shared among Plantaginaceae members. Using whole-genome triplication time as ~117 Mya[26] and *Ks*-peak at 2.21, we estimated the younger WGD occurred at ~32.3 Mya, posterior to divergence of *Antirrhinum* and other Lamiales plants, indicating it 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 two *Antirrhinum* species (peak value of *Ks* log-transformed distribution =0.04, Fig. 3C), and T is the divergence time of the two species (3.5Mya). The neutral mutation rate for *Antirrhinum* genus was estimated to be 5.7e-9 (95% CI:3.8e-9 - 1.3e-8) substitution per site per year, which is very close to the result estimated by using genotyping-by-sequencing dataset[20].

Loss of self-incompatibility occurred many times during angiosperms evolution history[27]. Although two species have split up for millions pf years , synteny

block analysis revealed strong collinearity between *A.hispanicum* and *M.orontium* (Fig. S13). And to investigate the evolution trajectory of SI- and SC- snapdragons, gene family analysis was conducted based on orthologous relationship of 19 plant species (Fig. 3A). A total of 445 and 294 gene families were expanded and contracted respectively in *Antirrhinum* lineage. The results also revealed 704 (3,230 genes) expanded and 1,692 (2,956 genes) contracted gene families in *A.hispanicum*, 1,097 (4,787 genes) expanded and 1,392 (4,015 genes) contracted gene families in *A.linkianum* respectively. (Fig. 3A). Moreover, 702 gene families (4,305 genes) were found in expansion, while 1209 families (1,771 genes) were found in contraction in the self-compatible *M.orontium*. These rapidly evolving gene families might give us some clues about consequence of losing self-incompatibility in *M.orontium* lineage.

The GO enrichment analysis for 445 expanded gene families of self-incompatible *Antirrhinum* showed that the significantly enriched GO terms mostly involved ‘binding’, ‘enzyme activity’, and ‘response’ (Fig. 3D). The abundance of transposable elements across whole genome may have contributed to copy number variation of these gene families in *Antirrhinum*.

While in *M.orontium*, enriched GO terms of expanded families mostly related to photosynthesis (Fig. 3E), and KEGG enrichment analysis revealed that the expanded genes were mainly implicated in phenylpropanoid, flavonoid, sesquiterpenoid and triterpenoid biosynthesis (Fig. S16), which may

contribute to the survive of self-compatible species facing environment changes[28] on the way to the evolutionary 'dead-end'. For the complete list of GO enrichment analysis results, see Table S12 and S13.

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

Identification sequence details of the *S*-locus would be the initial step towards understanding self-incompatibility in snapdragon. 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 (Fig. 1A). A cluster of 32 *SLF* (*S*-locus F-box) genes, which match to the structural characteristics of candidate *S*-locus, located in chromosome 8 (Fig. 1A). Gapless *S*-locus of 1.25Mb in *A.hispanicum* and pseudo *S*-locus of 804kb in *A.majus*[29] are both reconstructed with BioNano long-range molecules as support evidence (Fig. S7 and S8), whereas *S*-locus of *A.linkianum* and *M.orontium* were identified by synteny alignment with *A.hispanicum*. Except *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 S16). Additionally, the *S*-locus was surrounded by two copies of organ specific protein at left side (OSP1: *AnH01G33547.01*, OSP2: *AnH01G33548.01*) and five copies of phytocyanin genes at right side (here named PC1: *AnH01G33649.01*, PC2: *AnH01G33650.01*, PC3:

AnH01G33651.01, PC4: *AnH01G33652.01*, PC5: *AnH01G33653.01*) (Fig. 4E). 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 shape process of such a supergene. It has been proved that tandem or proximal duplicates evolved toward functional roles in plant self-defense and adaptation to dramatic changing environments[30]. 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). So, 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 results of stepwise local duplications, they would exhibit less differences and there should be more distant paralogs scattered elsewhere in the genome. Based on classification of gene duplication type, the two OSPs derived from the gamma-WGD followed by a recent tandem duplication (Table S20), while the PCs arise from tandem duplication asynchronously assuming constant substitution rate, both can be support by gene phylogenetic tree topology of their 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 S16). 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 changed much in cultivar *A.majus* after its domestication (Fig. S21).

Besides, 27 of 32 *SLFs* are intron-less, suggesting *SLFs* proliferated 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 122 Mya[31], indicated the *S*-locus structure is a very conserved and long lived remnants[32] [33] survived from long-term balancing selection (Fig. 4B). This observation concludes that the *S*-allelic polymorphisms should be inherited from common ancestors and the differentiations began before the currently extant species formation. Dissection of *SLFs* and flanking genes indicate that the S-locus is still experiencing gene gaining (Table S20).

From the intraspecies similarity matrix heatmap of *SLFs* of SI-Ah, 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. 4C). 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 pistil tissue (Fig. 4D). 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. 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[34], while over-expression of zma-miR408b in maize were reported to respond to abiotic stress to maintain leaf growth and improve biomass[35]. 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 Al-*S*-locus refers *S*-locus of self-incompatible *A.linkianum*. Am-S_C-locus and Mo-S_C-locus which contain no *S-RNase* denotes ψ*S*-locus of self-compatible *A.majus* and *M.orontium*. 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 compared different *S*-haplotype to identify polymorphism of the supergene.

Synteny plot showed structure and gene content of the flanking ends of different *S*-haplotype were well conserved. The S_7-locus has highly collinearity with Am-S_C-locus, S_8 locus and Al-*S*-locus except a small inversion at the start of alignment (Fig. 6A-C). From the comparison between Mo-S_C-locus and S_7-locus, a large deletion with a length of about 556kb could be observed, which involved S-RNase and more than ten SLFs as well as dozens of functional genes in S_7-haplotype. These genes comprise ATPase, Cytokinin, DUF4218, other ones mostly are uncharacterized proteins. This structural variation can be responsible for the loss of self-incompatibility in *M.orontium*. 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. 2), 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. S35) 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. 2), as reported in potato too[36].

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 pistil (Table S18), which is consistent with the situation in crab cactus[33], and they are nearly undetectable in leaf and petal. Those two gene only expressed in pistil of AhS_7S_C and AhS_8S_C respectively as expect, yet at higher level than AhS_7S_8 pistil. 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 AhS_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 AhS_7S_8 (Fig. 7B). We assumed that group2 are necessary for degrade toxic *S-RNase*.

Epigenetic modifications are believed to be crucial in controlling gene expression[37]. To determine whether methylation level is responsible for the difference between two groups, we compared the DNA methylation in the pollen from AhS_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. 7C). 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 obtain chromosome-level genome of self-incompatible *A.hispanicum*. Benefiting from the robust assembly, we can determine the genomic landscape of first self-incompatible snapdragon. Along with draft genomes of other two Plantaginaceae species, *A.linkianum* (2n=16) and *M.orontium* (2n=16), these assemblies would be valuable resources and enable us to understand genome evolution of *Antirrhinum* lineage, gaining knowledge of the proximate and ultimate consequences of losing self-incompatibility[4]. Apart from the shared gamma WGT in all eudicots, *A.hispanicum* 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.linkianum* during the Pliocene Epoch, coinciding with establishment of modern Mediterranean climate in the Mediterranean basin[38]. Mating system shift is ubiquitous throughout angiosperms, the evolutionary consequences of transition have attracted extensive attention for a long time[4 39 40]. The comparative genomic analysis indicated that after losing self-incompatibility, though the genomic architecture of SC- *M.orontium* 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 rate, and produced more plant compounds 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[41]. 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.

Structural variation such as translocations and inversion are considered drivers of genome evolution and speciation[42 43]. 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[44]. 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 single haplotype. Inter-species and intra-species *S*-haplotype comparison revealed continuous dynamic changes exist in snapdragon *S*-locus, and the hyper-divergence may reveal the cause of reproductive isolation and speciation. 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)*, *Antirrhinum linkianum,* and *Misopates orontium* were collected for DNA extraction using cetyltrimethylammonium bromide (CTAB) method. PCR-free 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. For PacBio Sequel II sequencing, high molecular weight (HMW) genomic DNA was extracted to construct SMRTbell libraries with insert size of 20kb, using SMRTbell Express Template Prep Kit 2.0 (PacBio, #100-938-900) and sequenced on PacBio Sequel II system. Another library of 15kb size was sequenced to generate high-qualitied consensus reads (HiFi reads) for *AhS_7S_8*. For BioNano optical maps sequencing, HMW genomic DNA of *AhS_7S_8* were 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).

To obtain a chromosome-level genome, Hi-C library was constructed for

AhS$_7$*S*$_8$. Young leaves were fixed with formaldehyde and lysed, then cross-linked DNA samples were digested with restriction 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 three species, plus pollen and pistil of line *AhS*$_7$*S*$_C$ and *AhS*$_8$*S*$_C$ were collected in the green house for RNA-seq. Two biological replicates were prepared for pistil and pollen. 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 library were generated.

Genome Assembly

To estimate the genome size independent of assembly, we characterized the basic genome feature using high-quality Illumina short reads. Firstly, PE reads were trimmed using Trim_galore (v0.6.1)[45] 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 three species were evaluated using software GenomeScope2[46] .

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)[47] 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[48], only primary contigs were kept for further scaffolding. The draft genome of *A.linkianum* and *M.orontium* were assembled using same analysis process. On the other hand, to obtain haplotype-resolved genome assemblies of AhS_7S_8, ~40X PacBio HiFi reads along with Hi-C data were utilized to separate maternal and paternal allele using FALCON-phase[49]. The resulting two phased contig-level haplotigs were used for downstream analysis independently. The contig assemblies were scaffolded into pseudo-

chromosomes 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 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 utilized to cluster contigs into 8 groups subsequently. Clean Hi-C reads were then mapped to the BioNano-derived scaffolds using BWA-mem[50]. The chromosome-level assembly was furtherly generated using the 3D-DNA analysis pipeline to correct mis-join, order, orientation, and translocation from the assembly[51]. Manual review and refinement of the candidate assembly was performed in Juicerbox Assembly Tools (v1.11.08)[52] 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 used ultra-high-depth PacBio dataset of two offsprings and PBJelly[53] to fill gaps in haplome assemblies, followed by two rounds of polish with gcpp program (https://github.com/PacificBiosciences/gcpp) in pbbioconda package. 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[54] website to annotate automatically and visualize genome maps.

Assemblies evaluation and comparison

First, the continuity and completeness of genome assemblies were benchmarked using long terminal repeat (LTR) Assembly Index (LAI)[16] and BUSCO embryophyta_odb10[55] core genesets, which include 2121 single-copy orthologous genes. Second, the PacBio reads and Illumina PE reads were mapped to respective assemblies using minimap2 and BWA[50], to evaluate correctness. MUMmer package[56] 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 between two phased assemblies were called and visualized using SYRI[57] based on reciprocal alignment from nucmer and show-coords. The effect of SNVs was estimated using SnpEff[58], 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) to build a *de novo* repetitive elements library from the assembled genome sequences independently. GenomeTools suite[59] was used to annotate LTR-RTs with protein HMMs from the Pfam database. In addition, LTR_retriever[60] was used to utilize output of GenomeTools and LTR_FINDER[61] 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[62] 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)[63] with parameters "-div 22 -cutoff 220 -xsmall".

Gene prediction and annotation

The gene prediction was performed using BRAKER2[64] 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[65], then converted and merged to BAM format. The hint files were incorporated with *ab initio* prediction tool AUGUSTUS[66] and GeneMark-EP[67] to train gene models. The predicted genes with protein length < 120 or 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)[68], rRNA and snRNA genes were identified using cmscan program in INFERNAL(v1.1.2)[69] package to search from the Rfam database. Four small-RNA sequencing datasets were collapsed, combining with known miRNA sequence download from mirBase(v22)[70] as

input of miRDP2(v1.1.4)[71] 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[72] 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)[73]. The protein sequences were also submitted to plantTFDB online server[74] to identify transcription factors and assign TFs families. In addition, predicted genes were also annotated biological function using diamond[75] to search against NCBI non-redundant protein and UniProtKB (downloaded on 21st Mar 2021, UniProt Consortium)[76] database with an e-value of 1e-5. COG annotation of genes was performed using eggNOG-mapper[77].

Whole-genome duplication and intergenomic analysis

Syntenic blocks (at least 5 genes per block) of *A.hispanicum* were identified using MCscan (Python version)[78] with default parameters. Intra-genome syntenic relationships were visualized using Circos[79] in Fig1. We also compared *A.hispanicum* genome with several plant genomes, including *Salvia miltiorrhiza*[80], *Solanum lycopersicum*[24], and *Aquilegia coerulea*[23], *Vitis*

vinifera[22]. 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[81] 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[82]. We plotted output data from result files using custom python scripts 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 angiosperm species listed in (Table S11) using OrthoFinder2[83]. Only longest protein sequences were selected to represent each gene model. Rooted species tree inferred by OrthoFinder2 using STRIDE(Specie Tree Root Inference from Duplication Events)[84] and STAG(Species Tree Inference from All Genes)[85] algorithm was used as a starting species tree for downstream analysis.

The species divergence time was estimated using MCMCtree[86] in PAML[82] package with branch lengths estimated by BASEML, with *Amborella* as outgroup. The Markov chain Monte Carlo (MCMC) process consists of

500,000 burn-in iterations and 400,000 sampling iterations. The same parameters were executed twice to confirm the results were robust. Species divergence time for *Amborella trichopoda- Arabidopsis thaliana* (173-199 Mya), *Vitis vinifera-Petunia axillaris* (111-131 Mya) and *Solanum lycopersicum–Solanum tuberosum* (5.23-9.40 Mya) which were obtained from TimeTree database[87] 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[88], 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 with Fisher's exact test and adjusted by the Benjamini-Hochberge method 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[65] to map clean RNA-seq reads to reference genomes with parameters '--quantMode

TranscriptomeSAM --outSAMstrandField intronMotif --alignIntronMax 6000 --alignIntronMin 50'. The transcripts per million (TPM) values were obtained using expectation maximization tool rsem[89]. 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 $\boldsymbol{r} > 0.85$ (Fig. S29), indicating good reproducibility. 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[90] 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[91]. 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[92]. 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[93] with 100 replicates of bootstrap and parameter '-m MF'. And duplicate gene were classified into different categories using DupGen_finder[30] 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 *A.hispanicum* to discover TF binding sites by MEME[94], and Tomtom was used for comparison against JASPAR database[95] 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[100] 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; L.C. provided plant seeds; S.Z. and D.Z. analyzed the data; S.Z., E.C. and Y.X. wrote the manuscript. All authors discussed the results and commented on the final manuscript.

References

1. Hiscock, S. J. & McInnis, S. M. Pollen recognition and rejection during the sporophytic self-incompatibility response: Brassica and beyond.

Trends Plant Sci. **8**, 606–613 (2003).

2. McClure, B. A. & Franklin-Tong, V. Gametophytic self-incompatibility: understanding the cellular mechanisms involved in ‘self’ pollen tube inhibition. *Planta* **224**, 233–245 (2006).

3. Bedinger, P. A., Broz, A. K., Tovar-Mendez, A. & McClure, B. Pollen-Pistil Interactions and Their Role in Mate Selection. *Plant Physiol.* **173**, 79–90 (2017).

4. Igic, B., Lande, R. & Kohn, J. R. Loss of Self-Incompatibility and Its Evolutionary Consequences. *Int. J. Plant Sci.* **169**, 93–104 (2008).

5. Shimizu, K. K. & Tsuchimatsu, T. Evolution of Selfing: Recurrent Patterns in Molecular Adaptation. *Annu. Rev. Ecol. Evol. Syst.* **46**, 593–622 (2015).

6. Golz, J. F., Clarke, A. E., Newbigin, E. & Anderson, M. A relic S-RNase is expressed in the styles of self-compatible Nicotiana sylvestris. *Plant J.* **16**, 591–599 (1998).

7. Markova, D. N. *et al.* Evolutionary history of two pollen self-incompatibility factors reveals alternate routes to self-compatibility within Solanum. *Am. J. Bot.* **104**, 1904–1919 (2017).

8. Yang, Q., Zhang, D., Li, Q., Cheng, Z. & Xue, Y. Heterochromatic and genetic features are consistent with recombination suppression of the self-incompatibility locus in Antirrhinum. *Plant J.* **51**, 140–151 (2007).

9. Schwarz-Sommer, Z., Davies, B. & Hudson, A. An everlasting pioneer:

The story of Antirrhinum research. *Nat. Rev. Genet.* **4**, 655–664 (2003).

10. Zhao, H. *et al.* Origin, loss, and regain of self-incompatibility in angiosperms. *Plant Cell* **34**, 579–596 (2022).

11. Li, W. & Chetelat, R. T. 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 (2015).

12. Yamane, H., Ikeda, K., Ushijima, K., Sassa, H. & Tao, R. 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 (2003).

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

14. Cheng, H. *et al.* Haplotype-resolved assembly of diploid genomes without parental data. *Nat. Biotechnol.* (2022) doi:10.1038/s41587-022-01261-x.

15. Muñoz-Sanz, J. V., Zuriaga, E., Cruz-García, F., McClure, B. & Romero, C. Self-(In)compatibility Systems: Target Traits for Crop-Production, Plant Breeding, and Biotechnology. *Front. Plant Sci.* **11**, 195 (2020).

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

17. Wright, S. I. & Schoen, D. J. Transposon dynamics and the breeding

system. *Genetica* **107**, 139 (1999).

18. Lowry, D. B. & Willis, J. H. A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8**, e1000500 (2010).

19. Ayala, D., Guerrero, R. F. & Kirkpatrick, M. Reproductive isolation and local adaptation quantified for a chromosome inversion in a malaria mosquito. *Evolution* **67**, 946–958 (2013).

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

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

22. Jaillon, O. *et al.* The grapevine genome sequence suggests ancestral hexaploidization in major angiosperm phyla. *Nature* **449**, 463–468 (2007).

23. Filiault, D. L. *et al.* The Aquilegia genome provides insight into adaptive radiation and reveals an extraordinarily polymorphic chromosome with a unique history. *Elife* **7**, e36426 (2018).

24. Sato, S. *et al.* The tomato genome sequence provides insights into fleshy fruit evolution. *Nature* **485**, 635–641 (2012).

25. Frazee, L. J. *et al.* New genomic resources and comparative analyses

reveal differences in floral gene expression in selfing and outcrossing Collinsia sister species. *G3 (Bethesda).* **11**, jkab177 (2021).

26. Jiao, Y. *et al.* A genome triplication associated with early diversification of the core eudicots. *Genome Biol.* **13**, R3 (2012).

27. Barrett, S. C. H. The evolution of plant sexual diversity. *Nat. Rev. Genet.* **3**, 274–284 (2002).

28. Dong, N.-Q. & Lin, H.-X. Contribution of phenylpropanoid metabolism to plant development and plant–environment interactions. *J. Integr. Plant Biol.* **63**, 180–209 (2021).

29. Li, M. *et al.* Genome structure and evolution of Antirrhinum majus L. *Nat. Plants* **5**, 174–183 (2019).

30. Qiao, X. *et al.* Gene duplication and evolution in recurring polyploidization-diploidization cycles in plants. *Genome Biol.* **20**, 1–23 (2019).

31. Igic, B. & Kohn, J. R. Evolutionary relationships among self-incompatibility RNases. *Proc Natl Acad Sci U S A.* **98**, 13167–13171 (2001).

32. Xue, Y., Carpenter, R., Dickinson, H. G. & Coen, E. S. Origin of allelic diversity in antirrhinum S locus RNases. *Plant Cell* **8**, 805–814 (1996).

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

34. Debernardi, J. M., Rodriguez, R. E., Mecchia, M. A. & Palatnik, J. F.

Functional specialization of the plant miR396 regulatory network through distinct microRNA-target interactions. *PLoS Genet.* **8**, e1002419 (2012).

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

36. Zhou, Q. *et al.* Haplotype-resolved genome analyses of a heterozygous diploid potato. *Nat. Genet.* **52**, 1018–1023 (2020).

37. Gibney, E. R. & Nolan, C. M. Epigenetics and gene expression. *Heredity (Edinb).* **105**, 4–13 (2010).

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

39. Goodwillie, C. *et al.* 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 (2010).

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

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

42. Dodsworth, S., Leitch, A. R. & Leitch, I. J. Genome size diversity in angiosperms and its influence on gene space. *Curr. Opin. Genet. Dev.* **35**, 73–78 (2015).

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

44. Lamichhaney, S. *et al.* Structural genomic changes underlie alternative reproductive strategies in the ruff (Philomachus pugnax). *Nat. Genet.* **48**, 84–88 (2015).

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

46. Ranallo-Benavidez, T. R., Jaron, K. S. & Schatz, M. C. GenomeScope 2.0 and Smudgeplot for reference-free profiling of polyploid genomes. *Nat. Commun.* **11**, 1432 (2020).

47. Koren, S. *et al.* Canu: Scalable and accurate long-read assembly via adaptive κ-mer weighting and repeat separation. *Genome Res.* **27**, 722–736 (2017).

48. Guan, D. *et al.* Identifying and removing haplotypic duplication in primary genome assemblies. *Bioinformatics* **36**, 2896–2898 (2020).

49. Kronenberg, Z. N. *et al.* Extended haplotype-phasing of long-read de novo genome assemblies using Hi-C. *Nat. Commun.* **12**, 1935 (2021).

50. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).

51. Olga, D. *et al.* De novo assembly of the Aedes aegypti genome using Hi-C yields chromosome-length scaffolds. *Scienc* **356**, 92–95 (2017).

52. Durand, N. C. *et al.* Juicer Provides a One-Click System for Analyzing Loop-Resolution Hi-C Experiments. *Cell Syst.* **3**, 95–98 (2016).

53. English, A. C. *et al.* Mind the Gap: Upgrading Genomes with Pacific Biosciences RS Long-Read Sequencing Technology. *PLoS One* **7**, e47768 (2012).

54. Tillich, M. *et al.* GeSeq - Versatile and accurate annotation of organelle genomes. *Nucleic Acids Res.* **45**, W6–W11 (2017).

55. Simão, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V. & Zdobnov, E. M. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31**, 3210–3212 (2015).

56. Marçais, G. *et al.* MUMmer4: A fast and versatile genome alignment system. *PLoS Comput. Biol.* **14**, 1–14 (2018).

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

58. Cingolani, P. *et al.* A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin).* **6**, 80–92

(2012).

59. Gremme, G., Steinbiss, S. & Kurtz, S. Genome tools: A comprehensive software library for efficient processing of structured genome annotations. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **10**, 645–656 (2013).

60. Ou, S. & Jiang, N. LTR_retriever: A highly accurate and sensitive program for identification of long terminal repeat retrotransposons. *Plant Physiol.* **176**, 1410–1422 (2018).

61. Xu, Z. & Wang, H. LTR-FINDER: An efficient tool for the prediction of full-length LTR retrotransposons. *Nucleic Acids Res.* **35**, W265–W268 (2007).

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

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

64. Brůna, T., Hoff, K. J., Lomsadze, A., Stanke, M. & Borodovsky, M. BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR genomics Bioinforma.* **3**, lqaa108 (2021).

65. Dobin, A. *et al.* STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).

66. Stanke, M. *et al.* AUGUSTUS: ab initio prediction of alternative transcripts. *Nucleic Acids Res.* **34**, W435–W439 (2006).

67. Brůna, T., Lomsadze, A. & Borodovsky, M. GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genomics Bioinforma.* **2**, 1–14 (2020).

68. Lowe, T. M. & Eddy, S. R. tRNAscan-SE: A Program for Improved Detection of Transfer RNA Genes in Genomic Sequence. *Nucleic Acids Res.* **25**, 955–964 (1997).

69. Nawrocki, E. P. & Eddy, S. R. Infernal 1.1: 100-fold faster RNA homology searches. *Bioinformatics* **29**, 2933–2935 (2013).

70. Kozomara, A., Birgaoanu, M. & Griffiths-Jones, S. MiRBase: From microRNA sequences to function. *Nucleic Acids Res.* **47**, D155–D162 (2019).

71. Kuang, Z., Wang, Y., Li, L. & Yang, X. MiRDeep-P2: Accurate and fast analysis of the microRNA transcriptome in plants. *Bioinformatics* **35**, 2521–2522 (2019).

72. Blum, M. *et al.* The InterPro protein families and domains database: 20 years on. *Nucleic Acids Res.* **49**, D344–D354 (2021).

73. Quevillon, E. *et al.* InterProScan: Protein domains identifier. *Nucleic Acids Res.* **33**, W116–W120 (2005).

74. Jin, J. *et al.* PlantTFDB 4.0: Toward a central hub for transcription factors and regulatory interactions in plants. *Nucleic Acids Res.* **45**,

D1040–D1045 (2017).

75. Buchfink, B., Xie, C. & Huson, D. H. Fast and sensitive protein alignment using DIAMOND. *Nat. Methods* **12**, 59–60 (2014).

76. Bateman, A. *et al.* UniProt: The universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

77. Huerta-Cepas, J. *et al.* EggNOG 5.0: A hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res.* **47**, D309–D314 (2019).

78. Tang, H. *et al.* Synteny and collinearity in plant genomes. *Science* vol. 320 486–488 (2008).

79. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* **19**, 1639–1645 (2009).

80. Song, Z. *et al.* A high-quality reference genome sequence of Salvia miltiorrhiza provides insights into tanshinone synthesis in its red rhizomes. *Plant Genome* **13**, e20041 (2020).

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

82. Yang, Z. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591 (2007).

83. Emms, D. M. & Kelly, S. OrthoFinder: Phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

84. Emms, D. M. & Kelly, S. STRIDE: Species tree root inference from gene duplication events. *Mol. Biol. Evol.* **34**, 3267–3278 (2017).

85. Emms, D. M. STAG: Species Tree Inference from All Genes. *bioRxiv* (2018) doi:10.1101/267914.

86. Yang, Z. & Rannala, B. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. *Mol. Biol. Evol.* **23**, 212–226 (2006).

87. Kumar, S., Stecher, G., Suleski, M. & Hedges, S. B. TimeTree: A Resource for Timelines, Timetrees, and Divergence Times. *Mol. Biol. Evol.* **34**, 1812–1819 (2017).

88. Mendes, F. K., Vanderpool, D., Fulton, B. & Hahn, M. W. CAFE 5 models variation in evolutionary rates among gene families. *Bioinformatics* **36**, 5516–5518 (2020).

89. Li, B. & Dewey, C. N. RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics* **12**, 323 (2011).

90. Edgar, R. C. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).

91. Bray, N. L., Pimentel, H., Melsted, P. & Pachter, L. Near-optimal probabilistic RNA-seq quantification. *Nat. Biotechnol.* **34**, 525–527 (2016).

92. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: A

Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

93. Kozlov, A. M., Darriba, D., Flouri, T., Morel, B. & Stamatakis, A. RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. *Bioinformatics* **35**, 4453–4455 (2019).

94. Bailey, T. L. *et al.* MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* **37**, 202–208 (2009).

95. Fornes, O. *et al.* JASPAR 2020: Update of the open-Access database of transcription factor binding profiles. *Nucleic Acids Res.* **48**, D87–D92 (2020).

96. Li, H. & Wren, J. Toward better understanding of artifacts in variant calling from high-coverage samples. *Bioinformatics* **30**, 2843–2851 (2014).

97. Tischler, G. & Leonard, S. Biobambam: Tools for read pair collation based algorithms on BAM files. *Source Code for Biology and Medicine* (2014) doi:10.1186/1751-0473-9-13.

98. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing -- Free bayes -- Variant Calling -- Longranger. *arXiv Prepr. arXiv1207.3907* (2012).

99. Garrison, E. *Vcflib: A C++ library for parsing and manipulating VCF files*. *GitHub* (2012).

100. Song, Q. *et al.* A reference methylome database and analysis pipeline

to facilitate integrative and comparative epigenomics. *PLoS One* **8**, e81148 (2013).

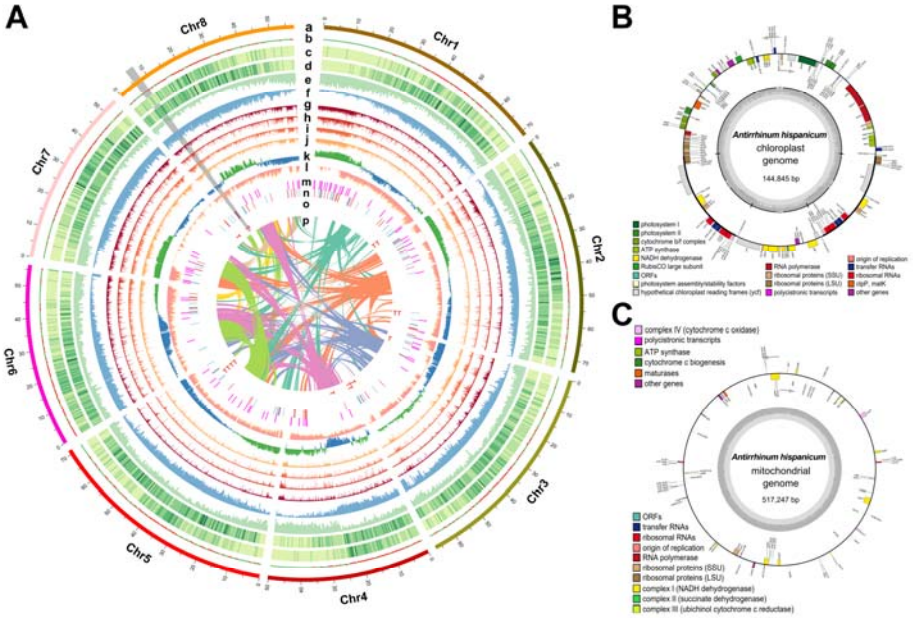


Fig.1 | Genomic features of *A.hispanicum*. A Characteristics of the nuclear genome, the outermost layer is circular ideogram of the eight pseudomolecules. Track **b-j** corresponds to the distribution of the GCcontent (red denote higher than average whole-genome level, and green vice versa), Ty3/Gypsy, Ty1/Copia retrotransposons (higher density shown in darker green), gene density, methylation level, expression level of four tissues (petal, pollen, pistil, 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.

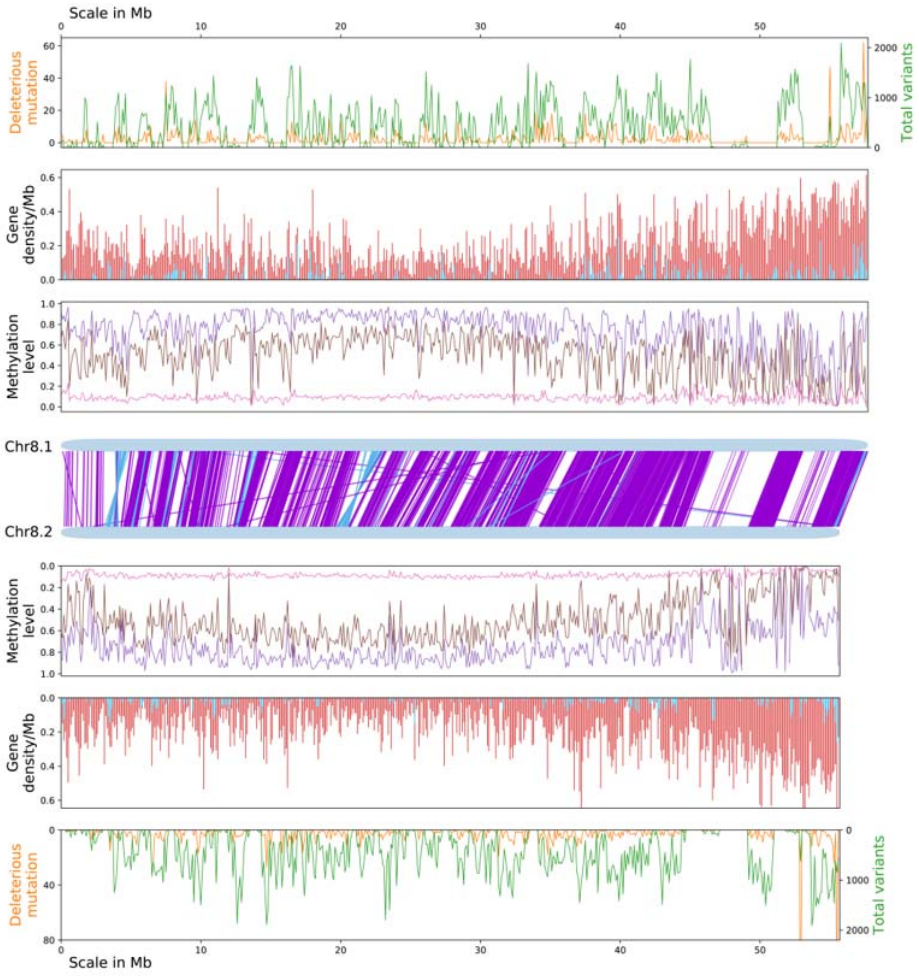


Fig.2 | Alignment of *A.hispanicum* chromosome 8 from two haplomes. The central light blue bars represent homologous chromosome 8 with the purple and dark blue lines indicating the forward and reverse alignments. The distribution of total variants(green), deleterious mutations (orange), annotated genes (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.

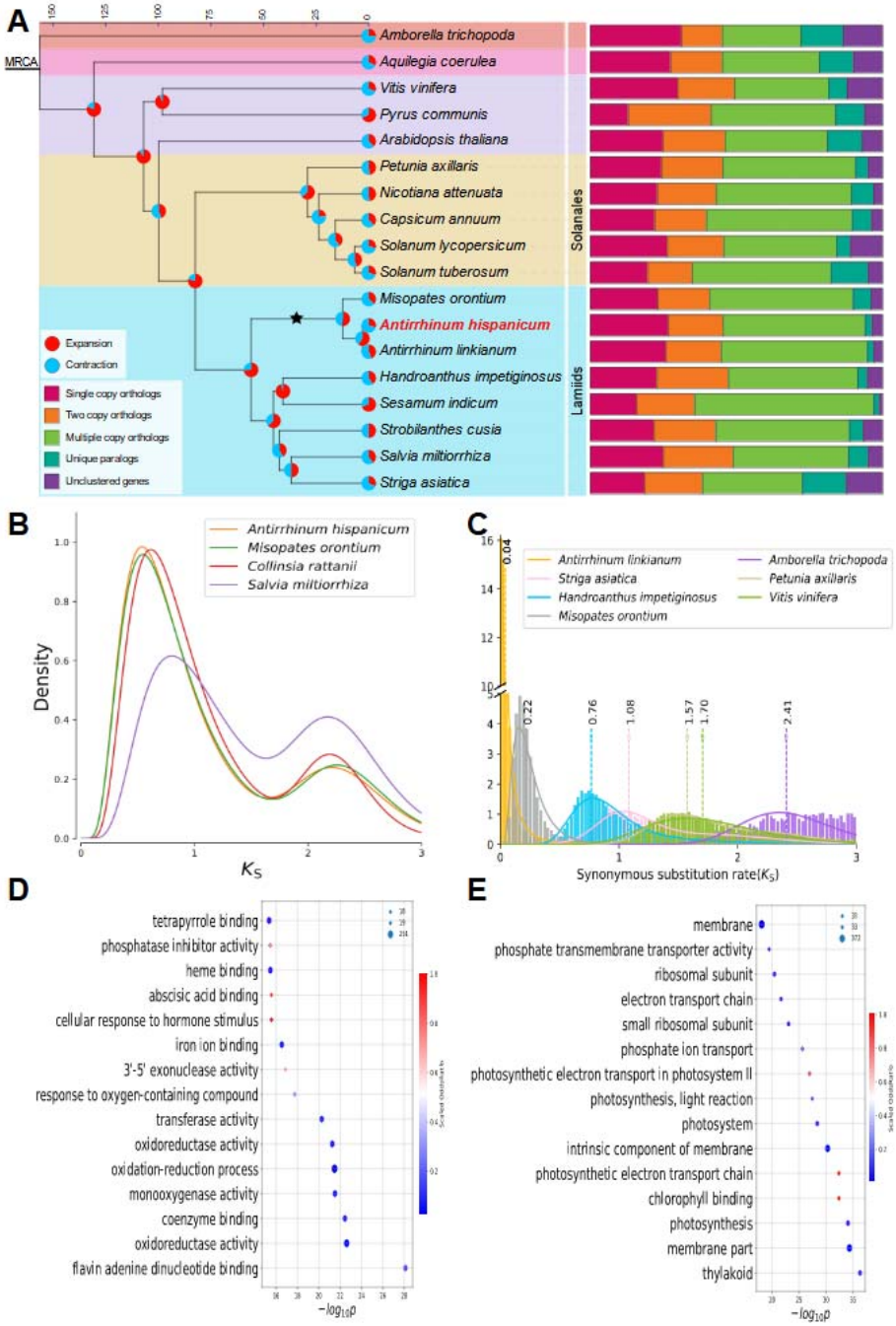


Fig. 3 | Comparative genomics analysis and lineage expanded gene families of *Antirrhinum*. A Phylogeny position of the snapdragon and gene families evolution. The left panel shows phylogenetic tree of 18 species. The

black star denotes the Plantaginaceae-specific WGD. 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 plant species. **B** Paranome *Ks* distribution of three Plantaginaceae members and *Salvia miltiorrhiza*. **C** *Ks* distribution of one-to-one orthologs identified between *A.hispanicum* and other species. The density curves were fitted using Gaussian mixture model, and the peak values marked with dotted dash lines. **D-E** Gene Ontology enrichment analysis of expanded family members in *Antirrhinum* lineage and *M.orontium* branch. The color of circle represents the scaled odds ratio. The size of circle represents the gene number of GO terms.

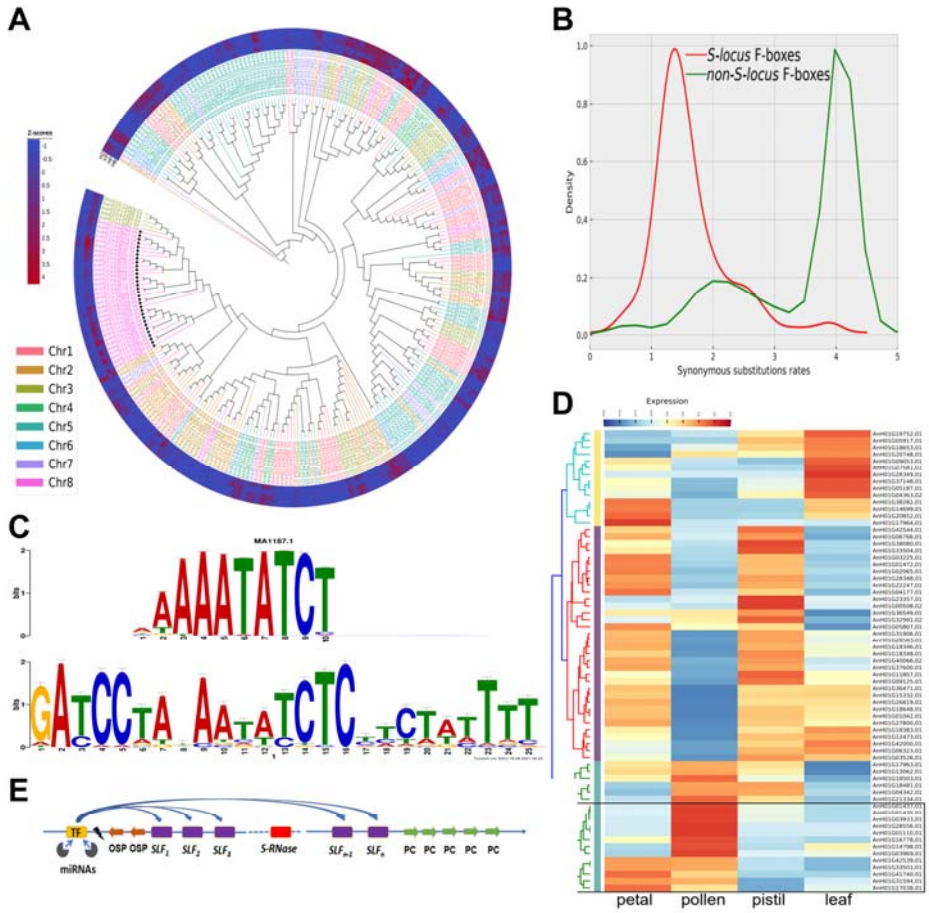


Fig.4 | Sequence analysis and candidate regulator of *SLFs*. A Phylogenetic tree reconstructed based on the amino acid sequences of all F-box genes from *A.hispanicum* genome. The maximal likelihood phylogenetic tree was calculated using raxml-ng with 1000 bootstrap. The color of gene ID denoted the 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. 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 pistil. **E** Schematic regulatory relationship of snapdragon *S*-locus.

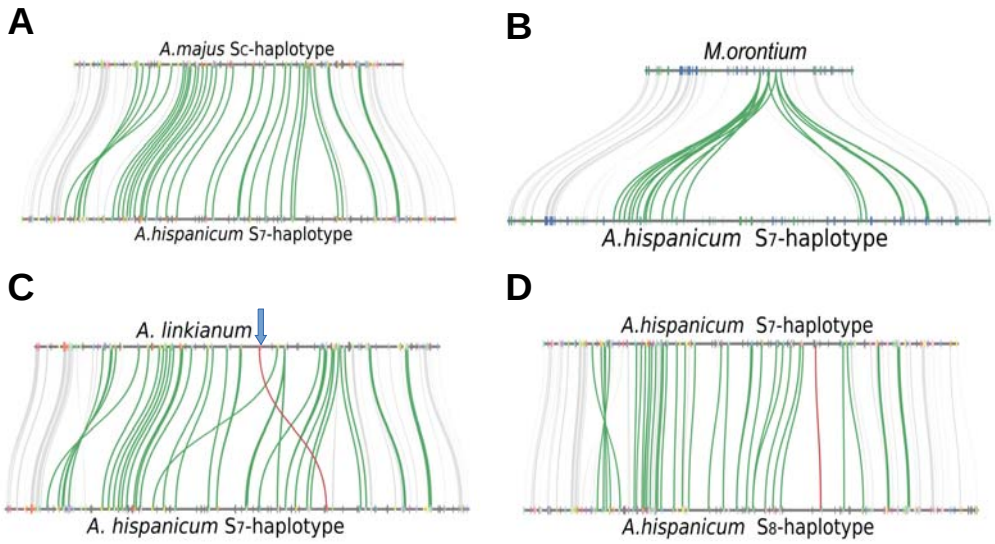


Fig.6 | The comparison of inter- and intra-specific *S*-haplotypes. A-C The micro collinearity patterns of *S*-haplotype in *A.majus*, *M.orontinum*, *A.linkianum*, with S_7-haplotype from *A.hispanicum*. **D** The microsynteny of S_7-haplotype vs. S_8-haplotype. The green curve lines and the red lines 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*.

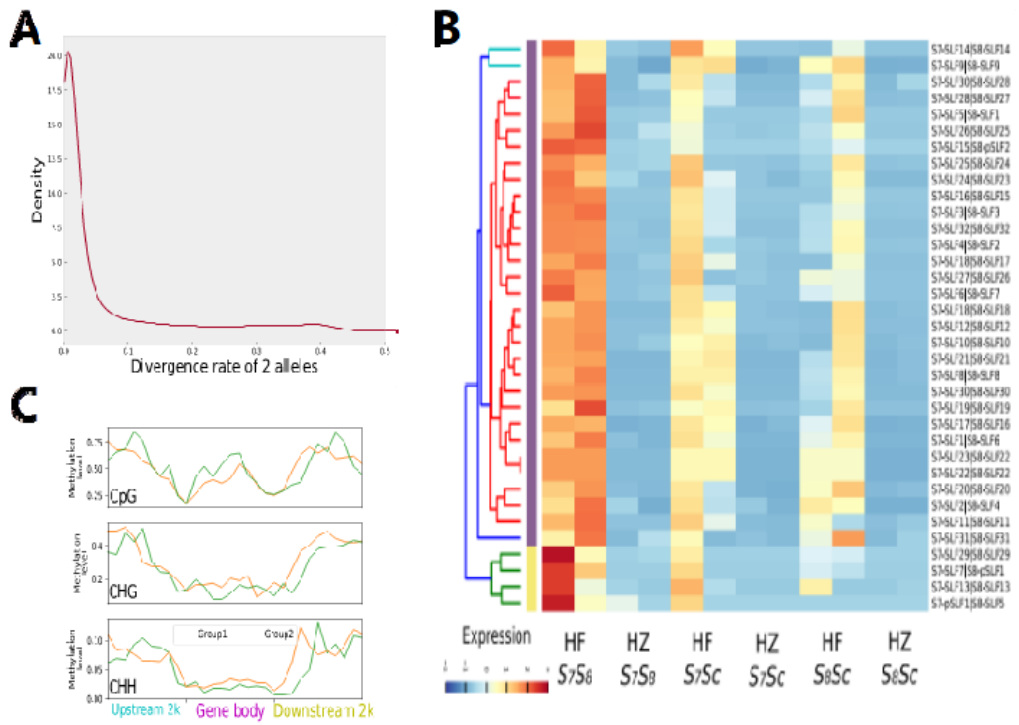


Fig.7 | Allelic expression and methylation profile of *SLFs* for line *AhS$_7$S$_8$*. **A** The distribution of allelic divergence of genes. The divergence was measured using Levenshtein edit distance. **B** Allele expression heatmap of *SLFs* in three genotypes. Every two columns represent allelic gene pairs of two *S*-haplotypes (HF: pollen, HZ: pistil). *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.

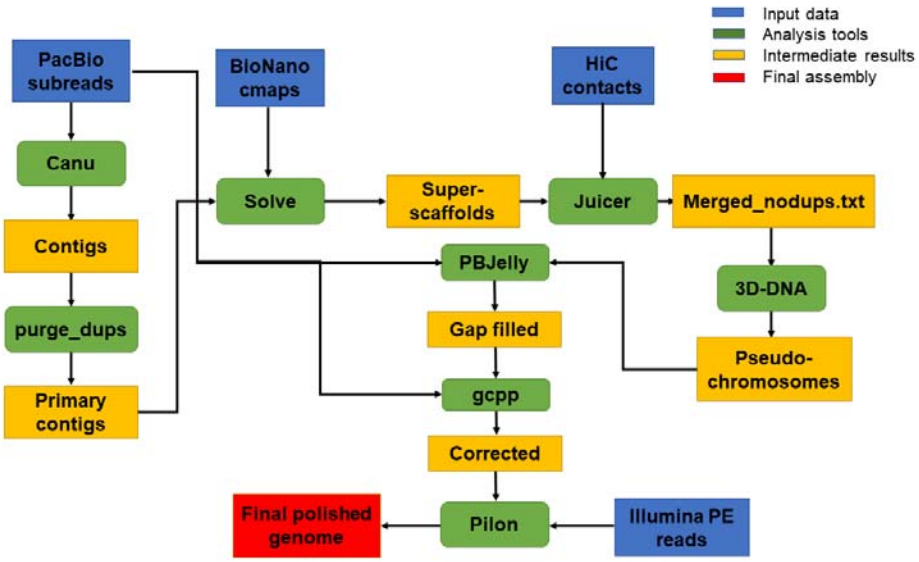


Fig. S1 Analysis flowchart of heterozygous *AhS*$_7$*S*$_8$.

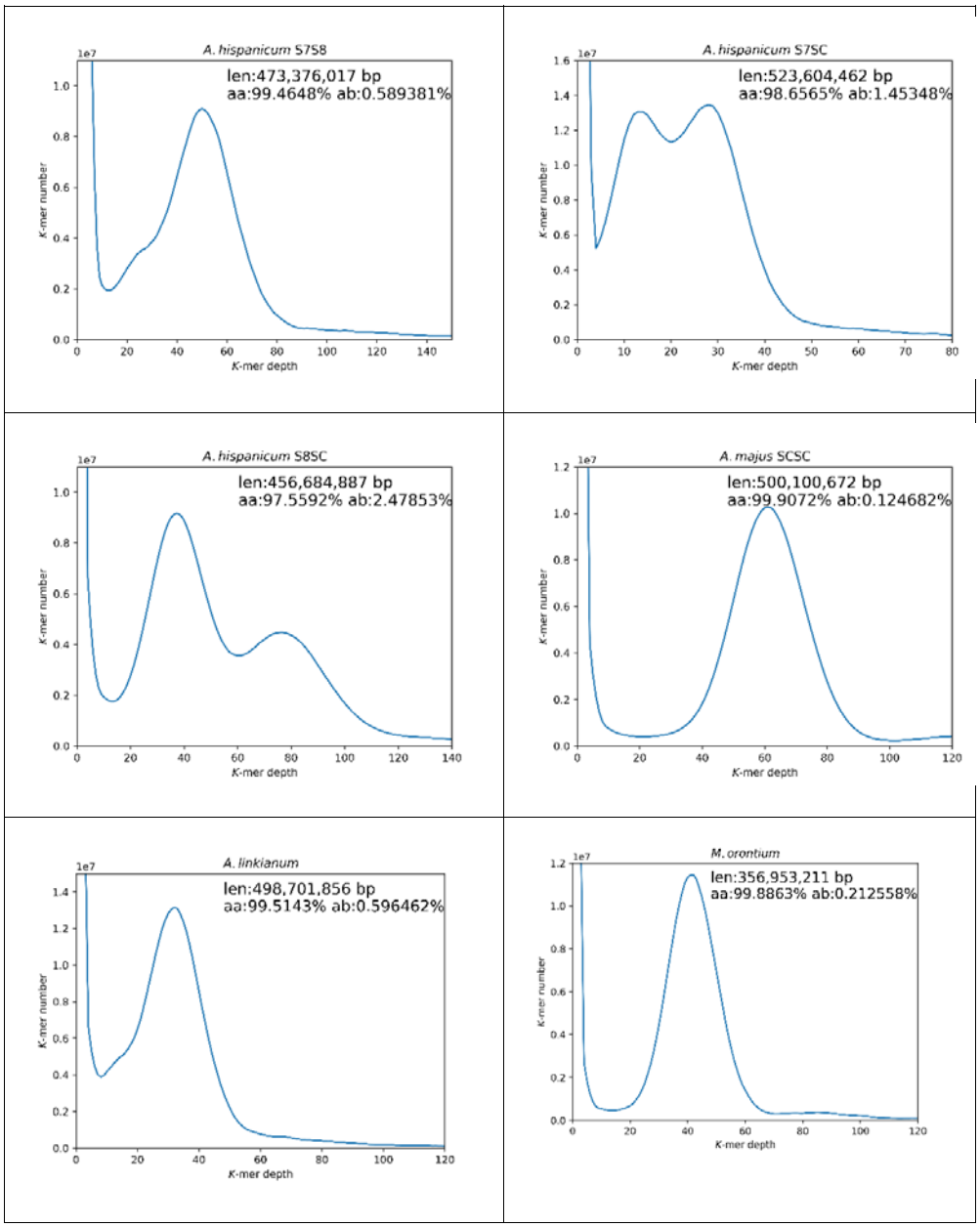


Fig. S2 Estimation of genome size using k-mer spectrum analysis. The frequency histograms of 21-mers were generated from whole genome Illumina reads of 3 *A.hispanicum* lines (AhS_7S_8, AhS_7S_C, AhS_8S_C), self-compaptible *A.majus* (AmS_CS_C), *A.linkianum* and *Misopates orontium*. The first and second peak represent the heterozygous and homozygous regions in genome respectively.

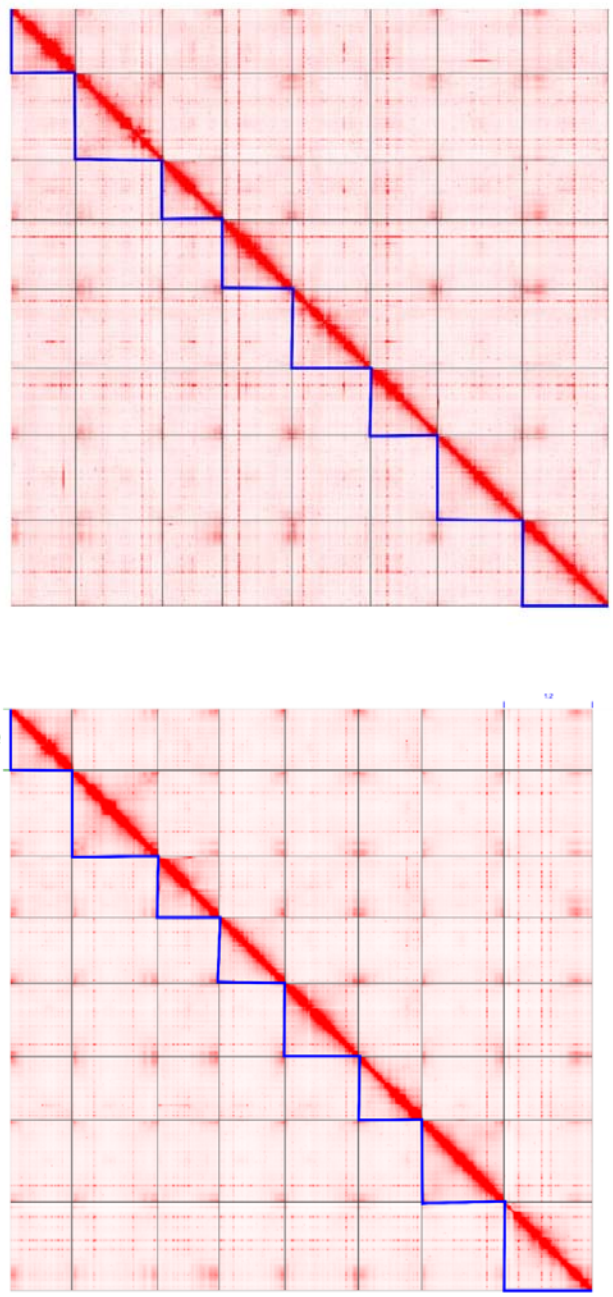


Fig. S3 Hi-C contact map representing chromosome-scale phased genomes of *Antirrhinum hispanicum*. Hi-C contact matrix visualization for the phased genomes, S_7 haplome and S_8 haplome. The pixel intensity represents the Hi-C links at 100 Kb size-windows at logarithmic scale. The blue bold lines separate chromosomes for each phased genome.

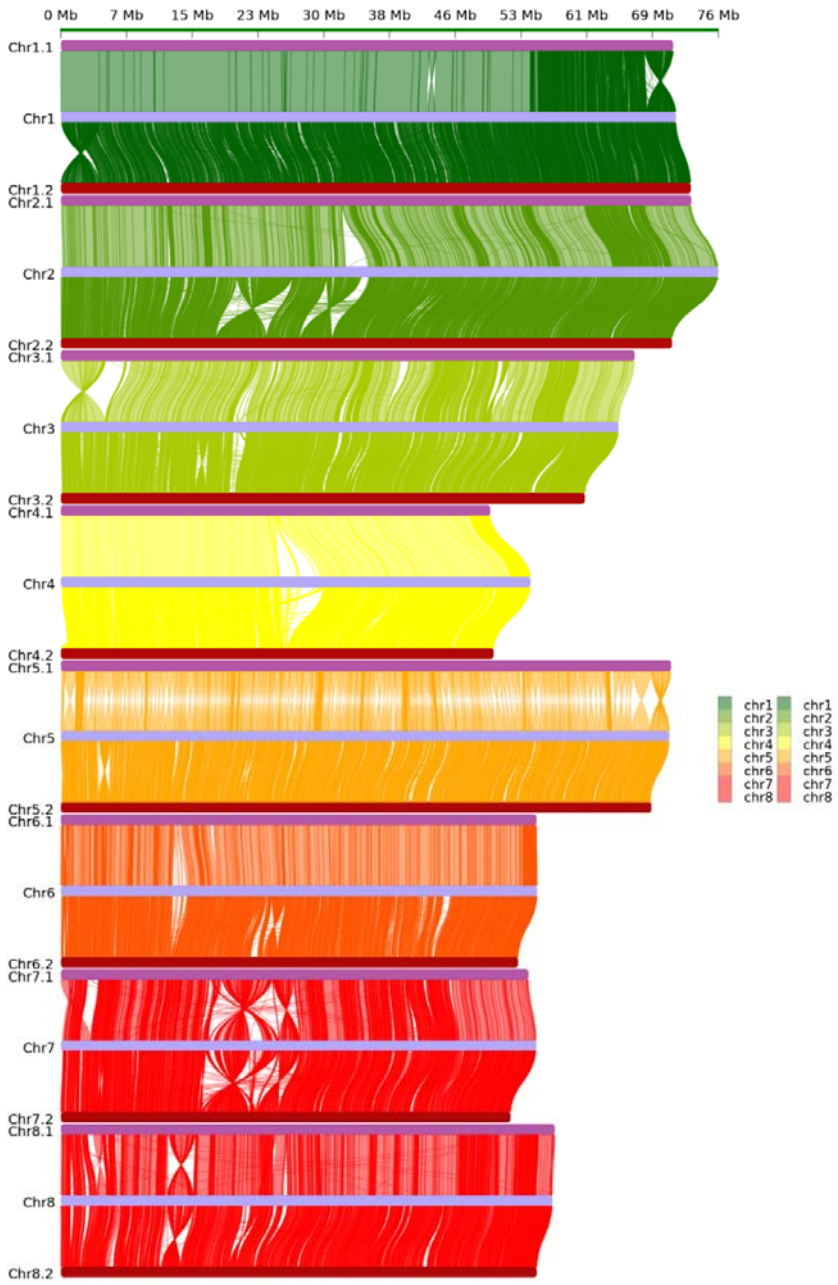


Fig. S4 DNA colinear alignment between two haplomes and consensus assembly of *A.hispanicum*.

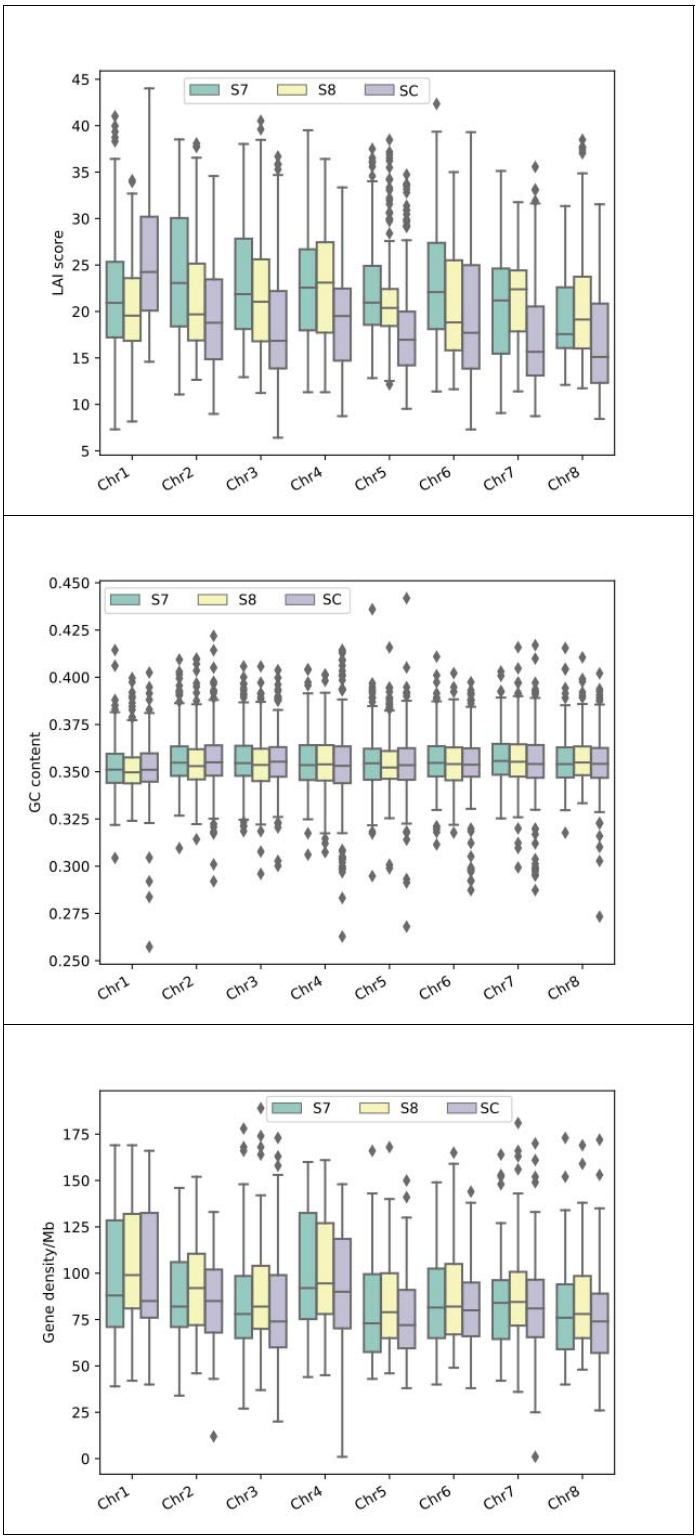


Fig. S5 Boxplot of LTR Assembly Index (a), GC content (b), and gene density (c) for two haploid genomes of *A.hispanicum* and S_C genome of *A.majus*.

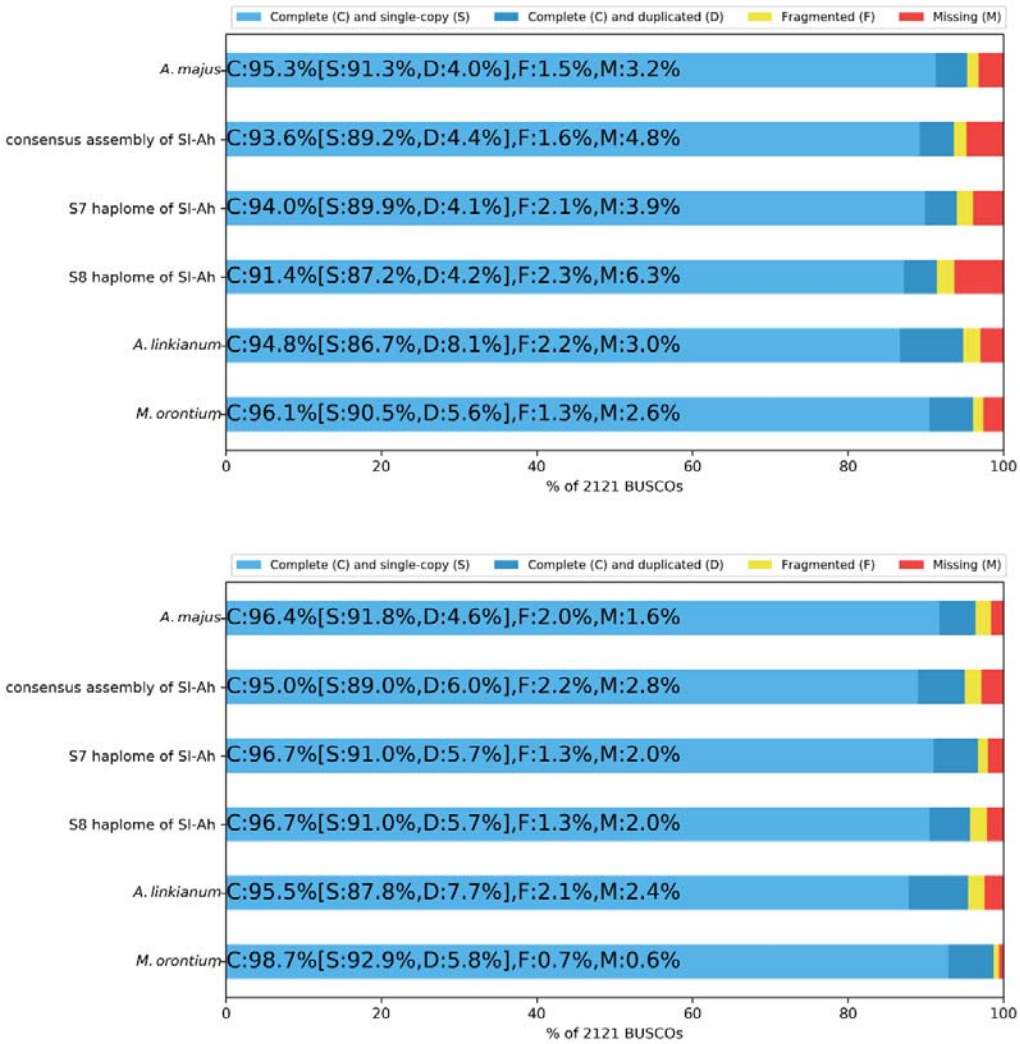


Fig. S6 BUSCO assessment results for *A.majus*, consensus-assembly and two haplomes of SI-Ah, *A.linkianum*, *M.orontium*. The different panels correspond to different running modes:(A) assembled genomes, (B) predicted proteins.

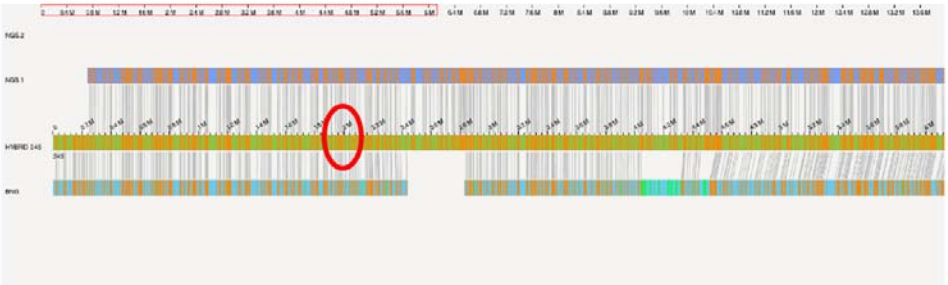


Fig. S7 BioNano evidence for *S*-locus correctness of S_7 haplotype. The top blue panel represents pacbio-assembly reference cmap. The middle panel represents hybrid scaffold consensus cmap. The bottom represents BioNano *de novo* assembly cmap. The red circle denoted *S-RNase*. The grey lines between bars represent consensus restriction enzyme cutting sites.

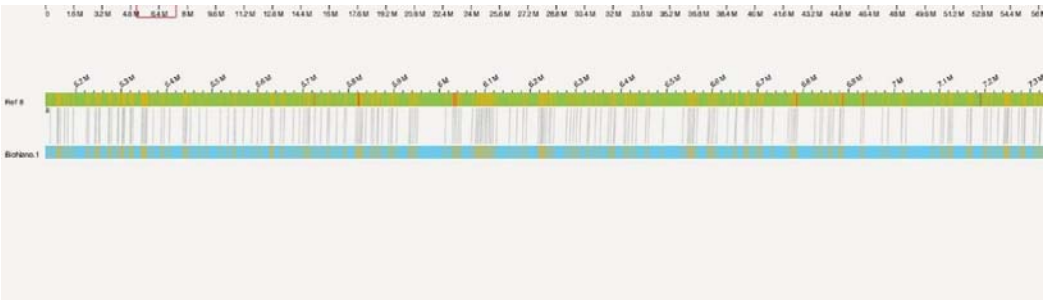


Fig. S8 Bionano evidence for pseudo *S*-locus correctness of S_C haplotype in *A.majus*. The top green panel represents reference cmap. The bottom represents bionano *de novo* assembly cmap. The grey lines between bars represent consensus restriction enzyme cutting sites.

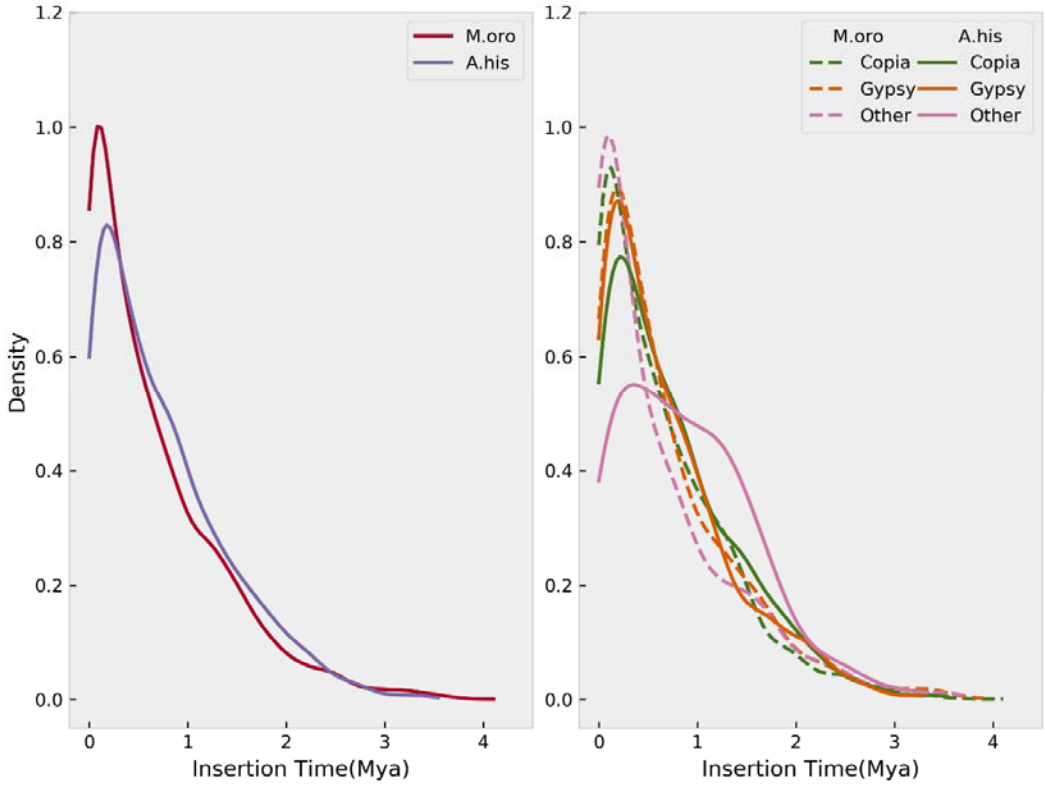


Fig. S9 The density distribution of intact LTR-RTs insertion time. a) The density distribution of insertion time of intact LTR-RTs in self-compatible *M.orontium* and self-incompatible *A.hispanicum* genome. b) The density distribution of insertion time of intact Copia, Gypsy and other unclassified LTR-RTs in *M.orontium* and *A.hispanicum* genomes.

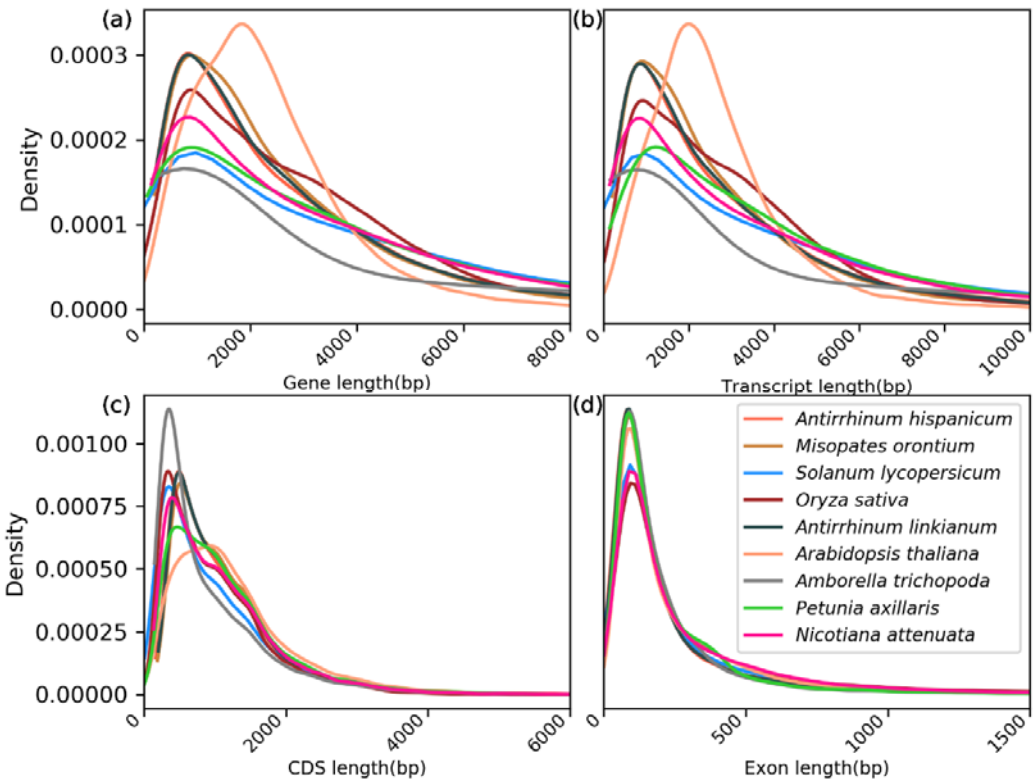


Fig. S10 Comparison of gene length (a), transcript length (b), CDS length (c), and exon length (d) of *A.hispanicum*, *A.linkianum*, *M.orontium* and other six species. The x-axis represents length and the y-axis represents the density of genes. No unexpected differences were observed between *A.hispanicum* and *A.linkianum*. Other species, except *A.thaliana*, show similar genomic feature with snapdragon.

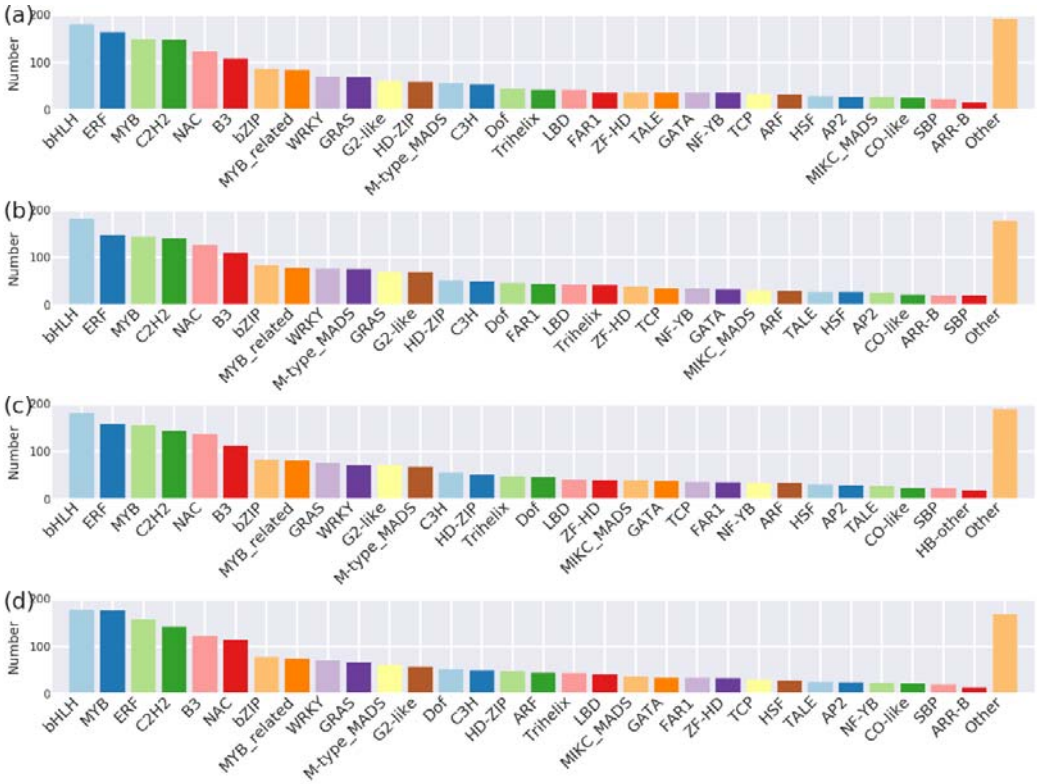
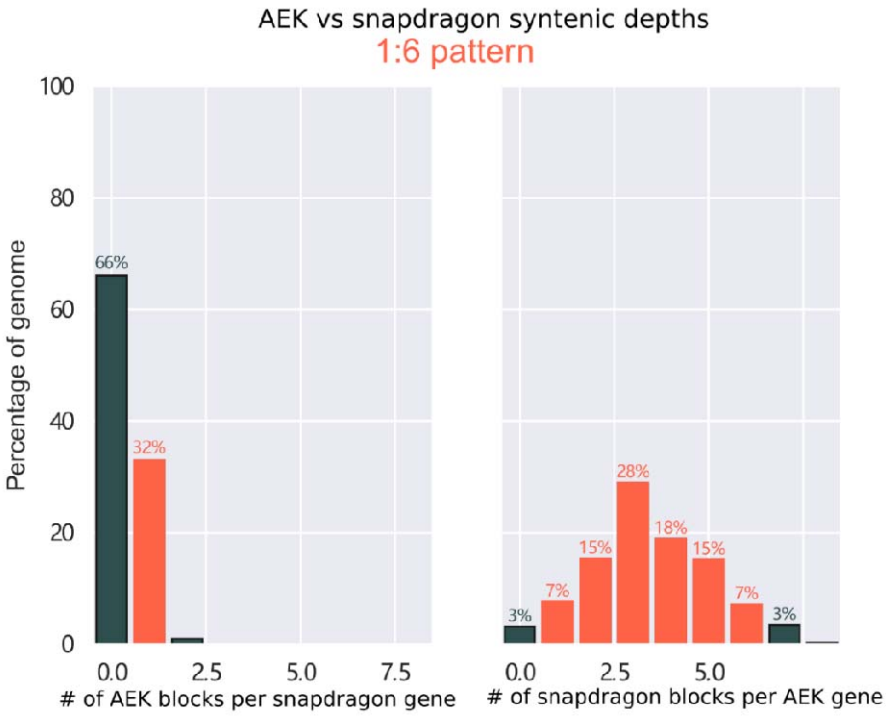


Fig. S11 Families distribution of predicted transcription factors in *A.hispanicum* (a), *A.linkianum* (b), *A.majus*(c) and *M.orontium*(d) genome.



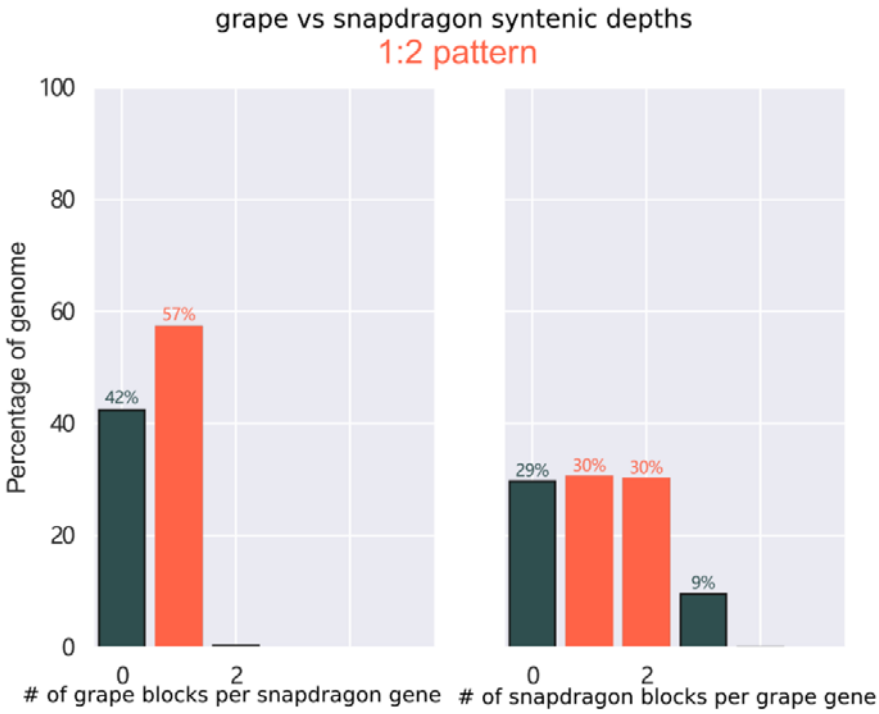


Fig. S12 Ratio of syntenic depth between snapdragon and AEK, grape genome respectively. Syntenic blocks of snapdragon per AEK gene and syntenic blocks of AEK per snapdragon gene are shown indicating a clear 1:6 pattern.

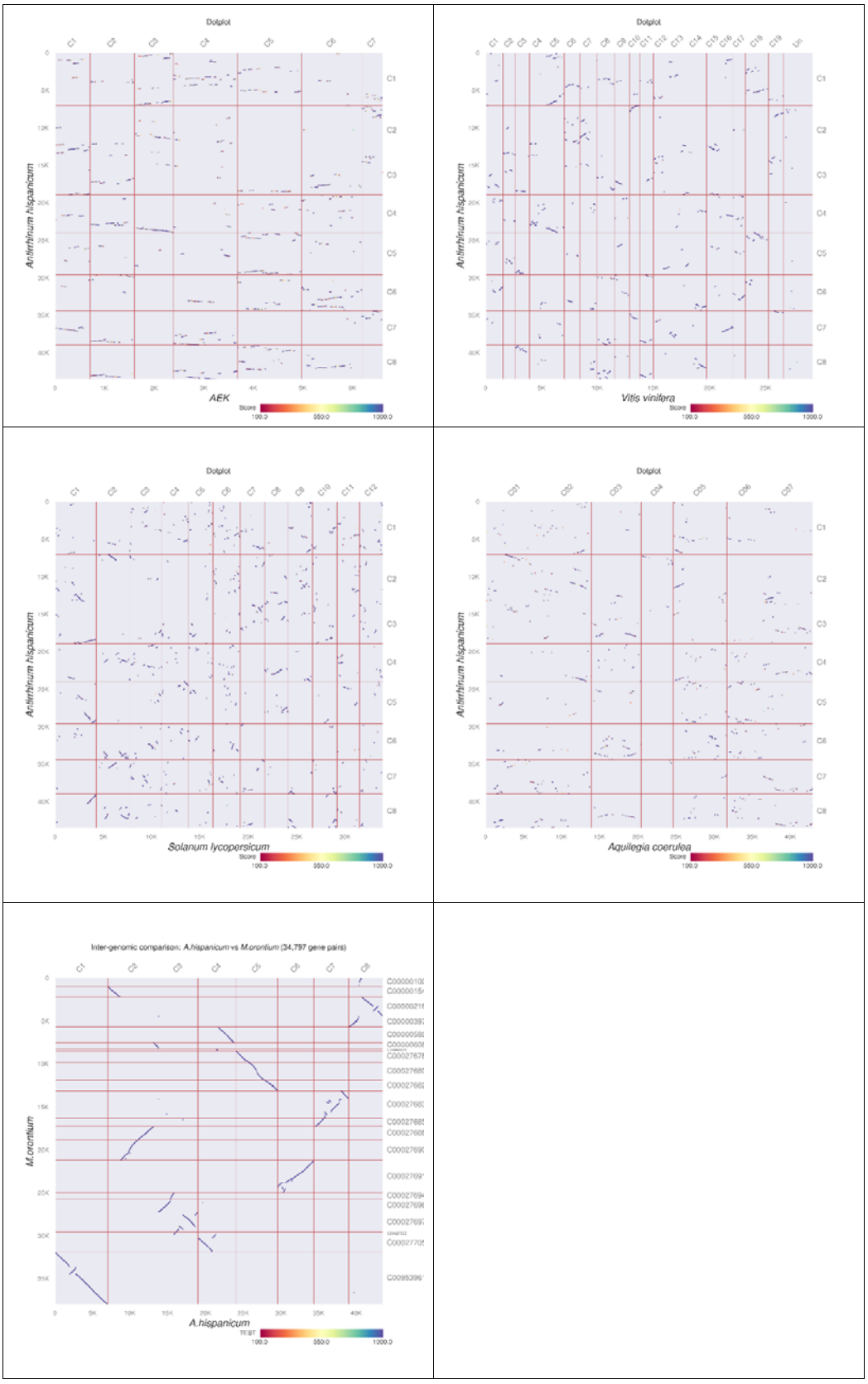


Fig. S13 Inter- and Intra- genomic dot-plot for *Antirrhinum hispanicum* genome and other selected species including ancestral eudicot karyotype (AEK).

Dot-plot based synteny visualization showed a clear 1:6 synteny depth pattern between AEK and *A.hispanicum* genome. Synteny plot between, *A.hispanicum - Vitis vinifera*, 1:2 pattern, *A.hispanicum - Solanum lycopersicum*, 2:3 pattern, *A.hispanicum - Aquilegia coerulea*, 1:3 pattern. Previous studies demonstrated that Aquilegia species experienced a WGD event after the divergence with grape.

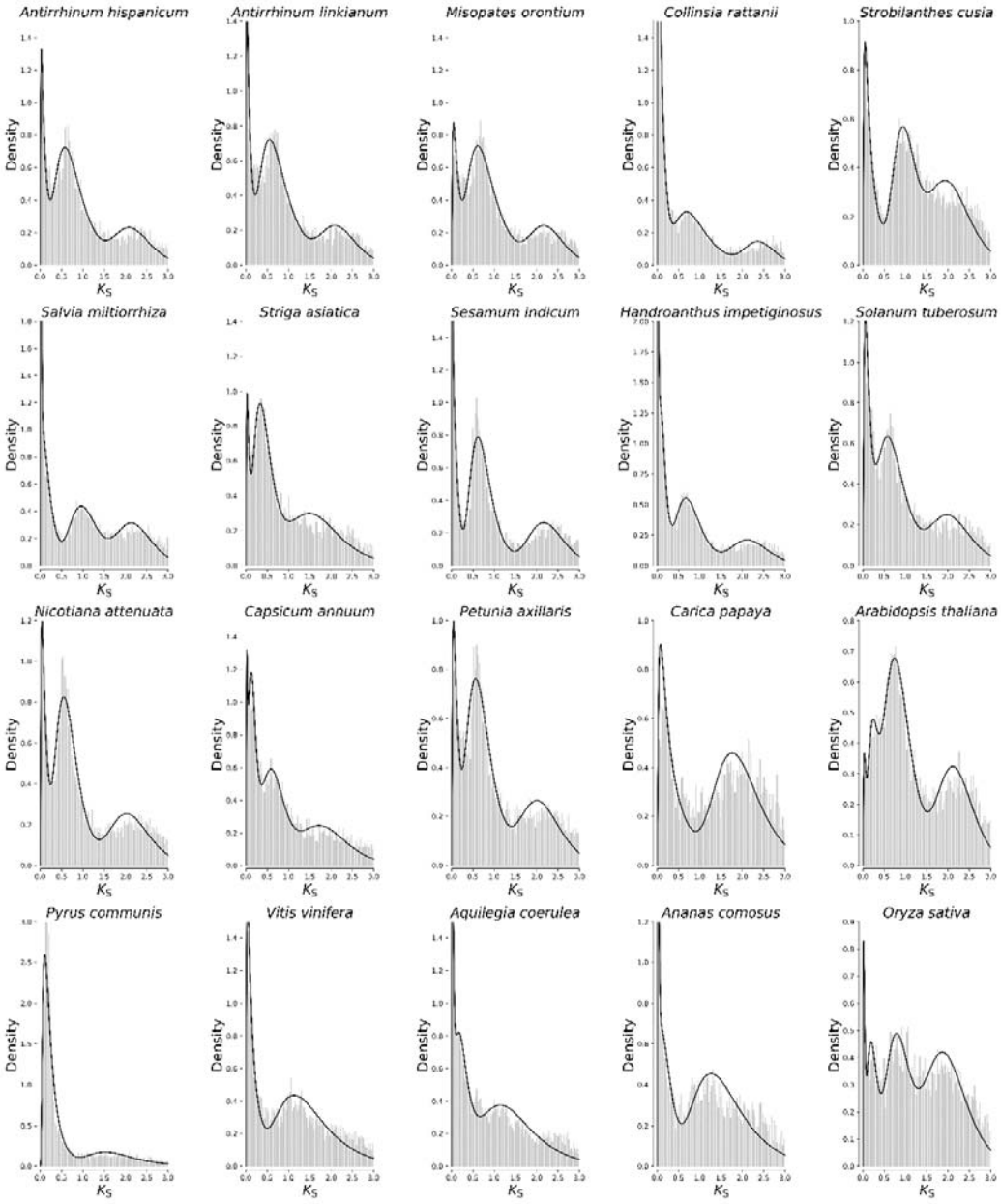


Fig. S14 Density plot of *Ks* (synonymous substitution rate) distribution of paralogous gene pairs for selected plant species. The solid lines were fitted using Gaussian Mixture Model function.

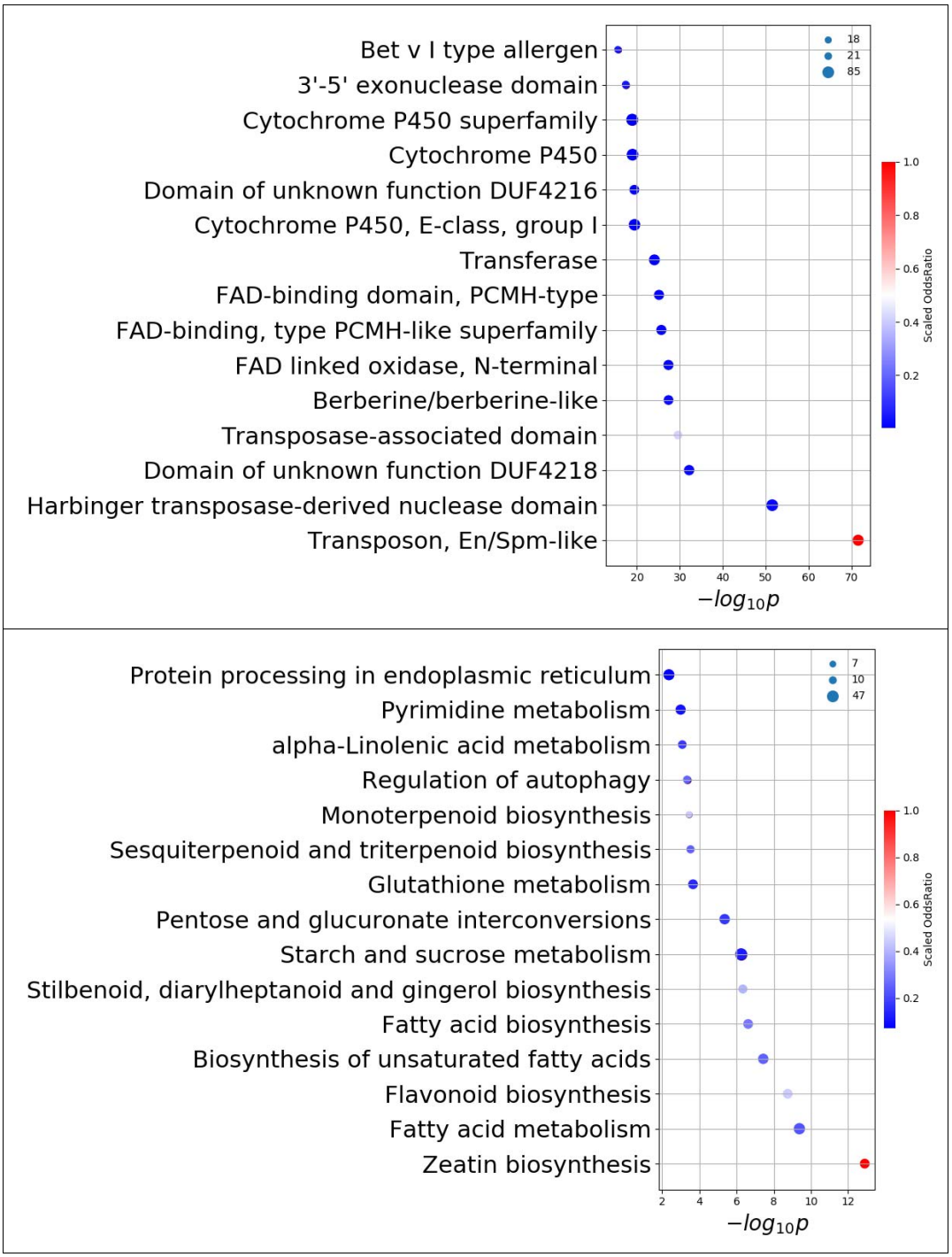


Fig. S15 InterPro, and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of *Antirrhinum* lineage gene members from expanded families. The color of circle represents the scaled odds ratio. The size of circle represents the gene number of the IPR terms or KEGG pathway. The p-value is corrected in hypergeometric test.

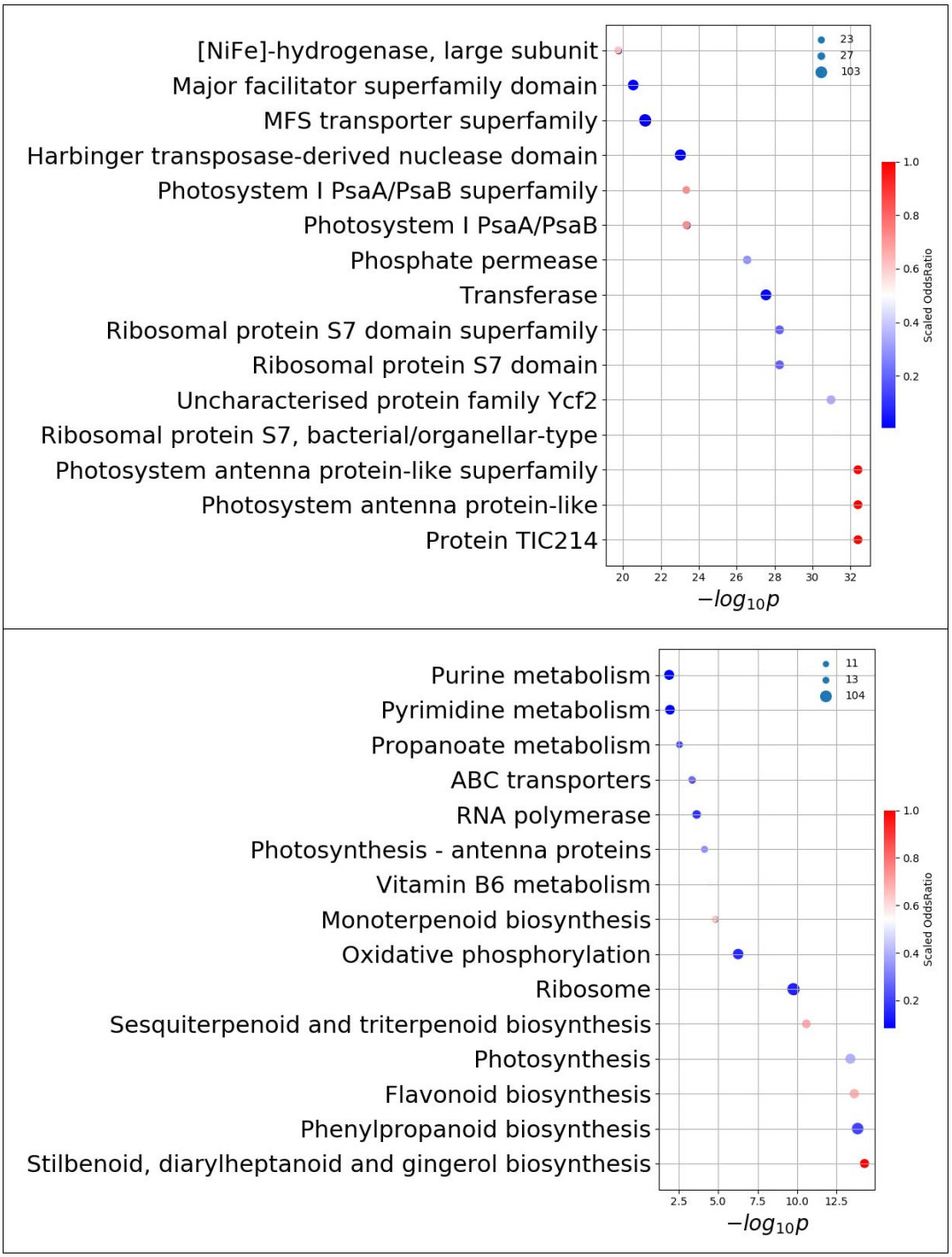


Fig. S16 InterPro, and Kyoto Encyclopedia of Genes and Genomes (KEGG) enrichment analysis of *Misopates orontium* gene members from expanded families. The color of circle represents the scaled odds ratio. The size of circle represents the gene number of the IPR terms or KEGG pathway. The p-value is corrected in hypergeometric test.

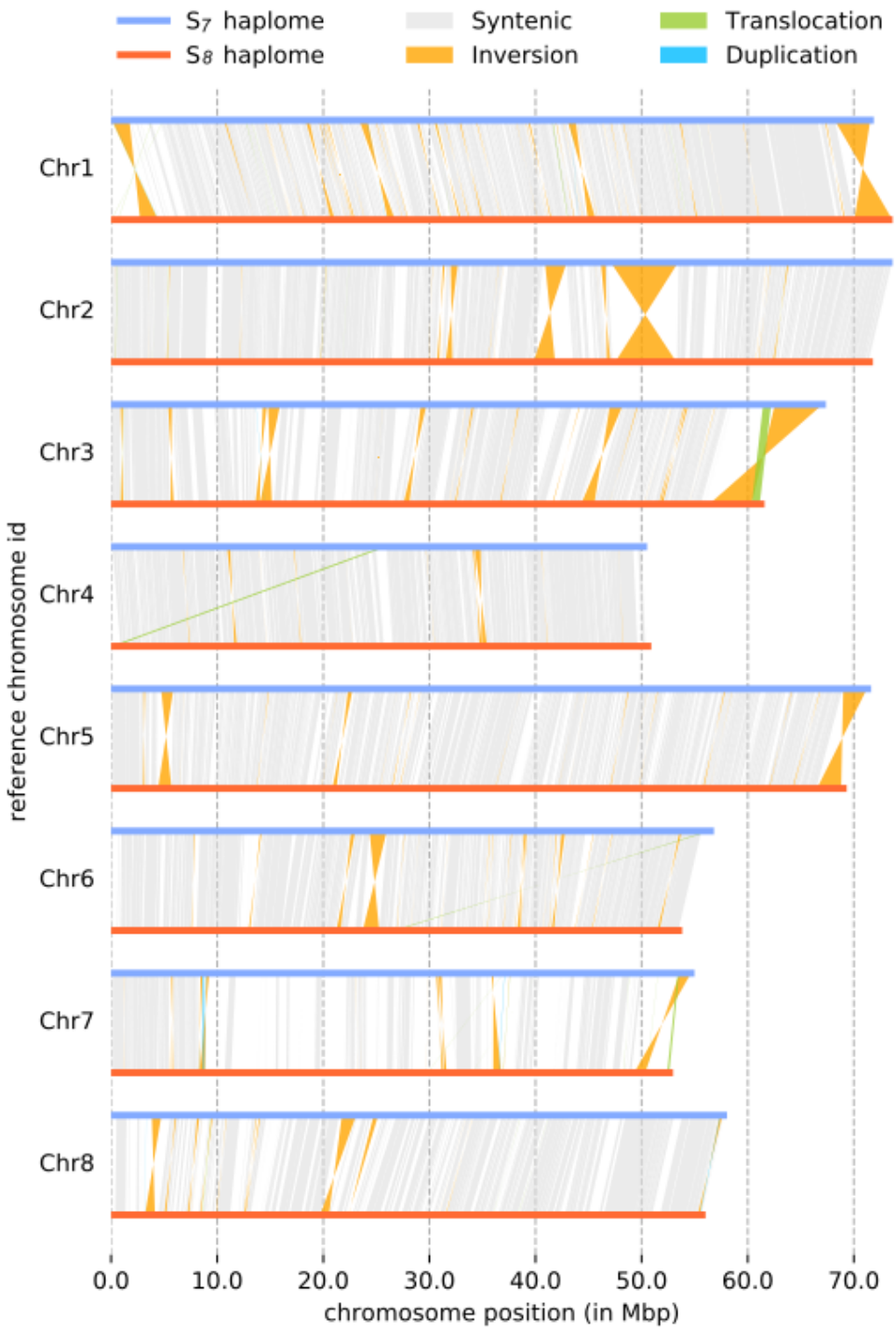


Fig. S17 DNA alignment of two haploid genomes. Structural rearrangements are represented by colored lines (orange for inversions, green for translocations, blue for duplications) connecting regions of two homologous chromosomes (blue and red horizontal lines, respectively); syntenic regions are connected by gray lines.

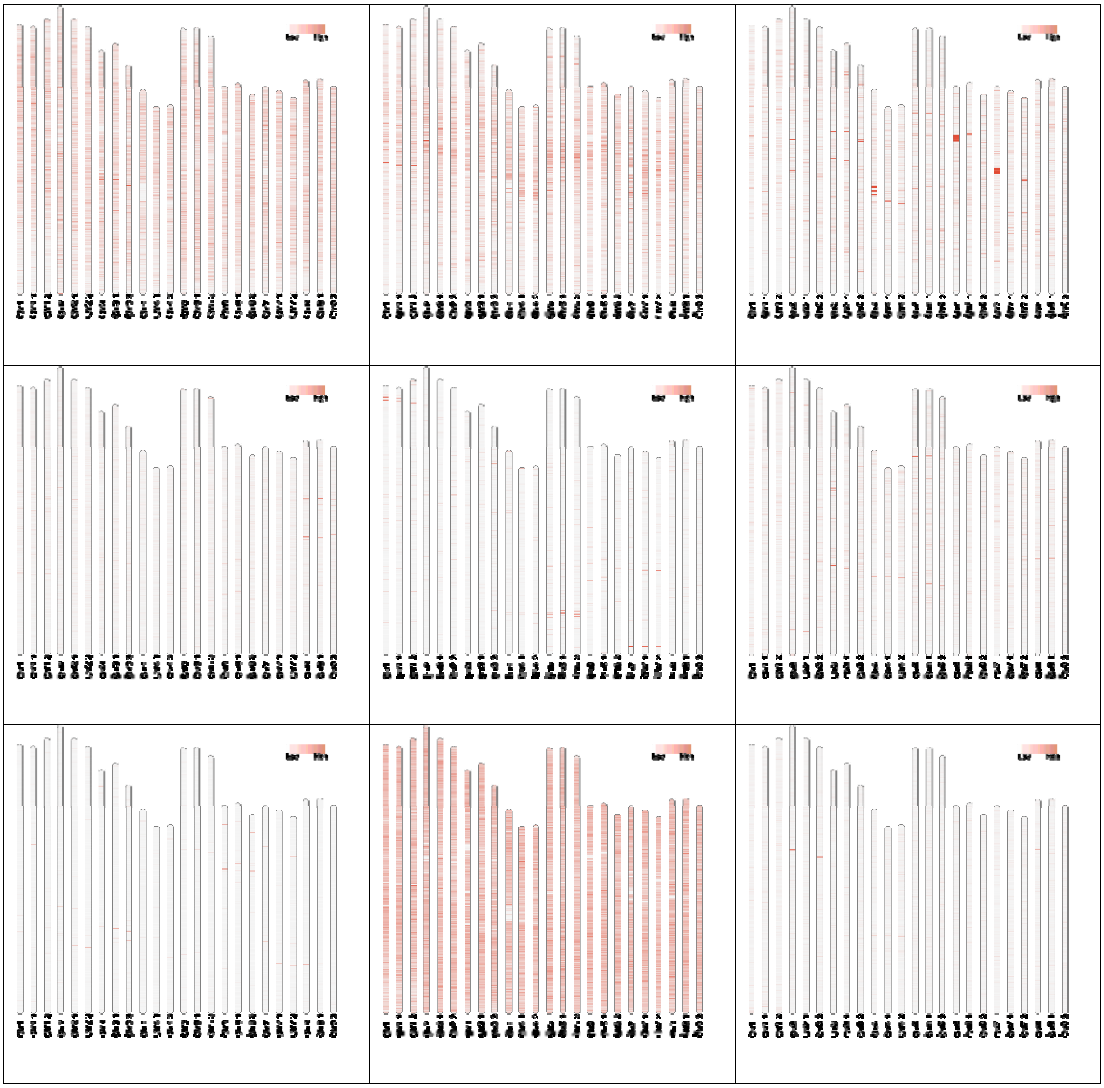


Fig. S18 The distribution heatmap of repeat elements density across three haplomes. The density for each type of elements (Gypsy, Copia, EnSpms, hAT, Helitron, LINE (Long Interspersed Nuclear Element), SINE (Short Interspersed Nuclear Element), MITEs, and MuDRs) was caculated using a non-overlap sliding window approach with window size of 100kb. The ideogram labels ended with '.1' are chromosomes of S_7 haplome, with '.2' are chromosomes of S_8 haplome. The other ideograms denote S_C haplome of homozygous *A.majus*.

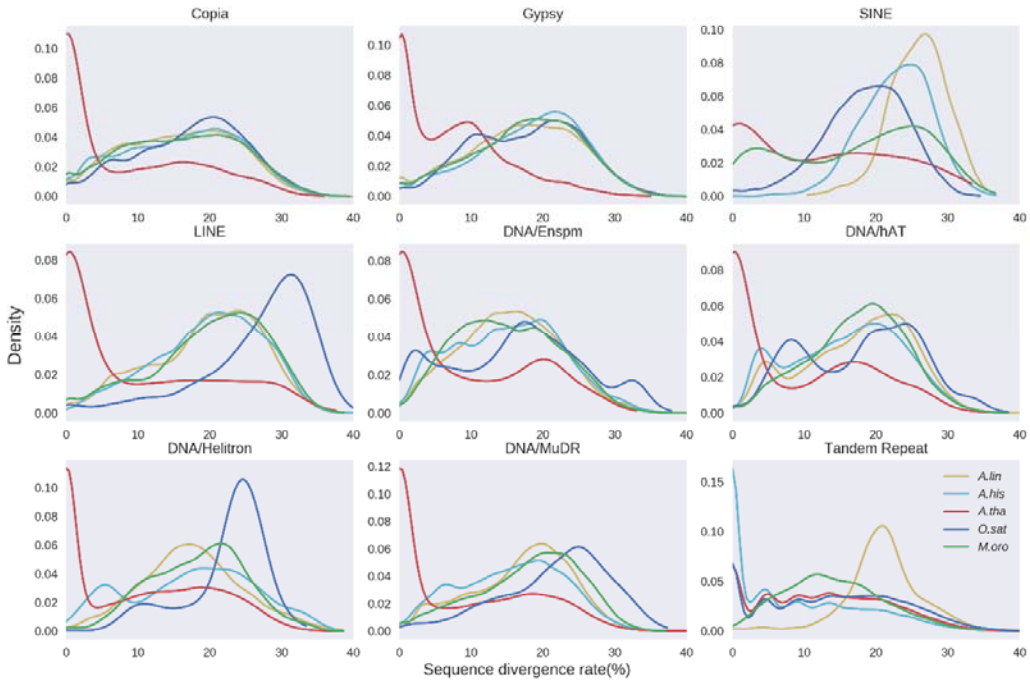


Fig. S19 Denstity plot of divergence rates of different transposable element in different genomes.

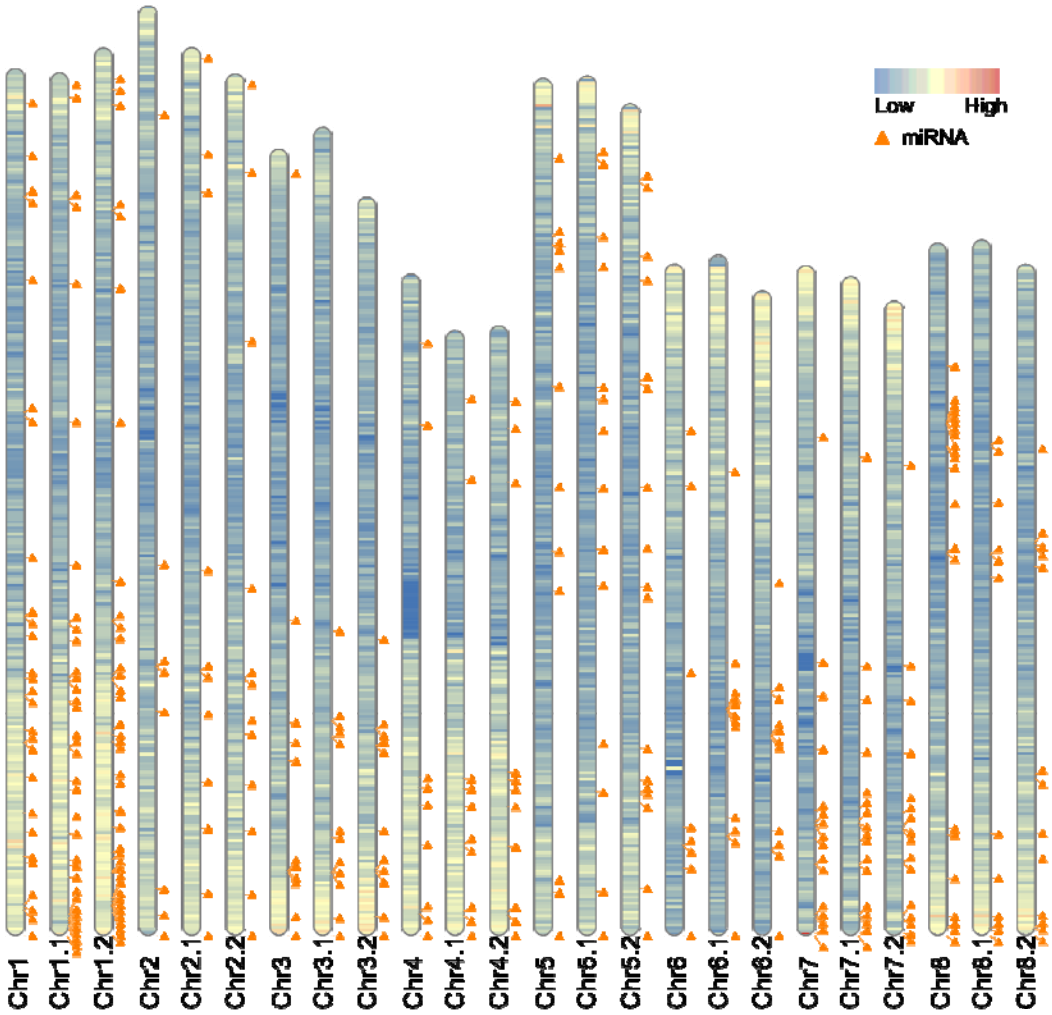


Fig. S20 Gene density heatmap of 3 haplomes (S_7, S_8, and S_C). The ideogram labels ended with '.1' and '.2' denote chromosomes of S_7 haplome and S_8 haplome of SI-Ah respectively. The other ideogram represent SC-Am S_C haplome. Chromosomes were binned into 100kb window to display the gene distribution of assembled chromosomes. Also, the miRNA location of 3 different haplotype genomes were marked along the chromosome using yellow triangle.

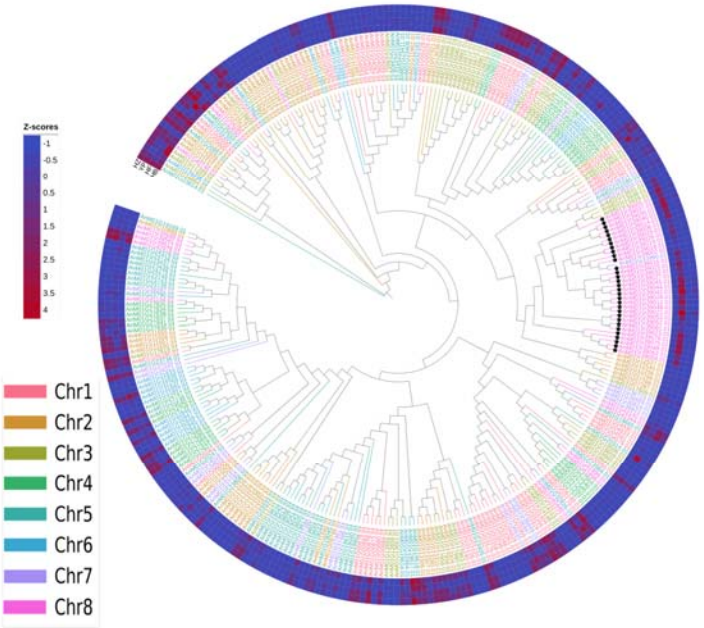


Fig. S21 Phylogenetic tree reconstructed based on the amino acid sequences of all F-box genes from *A.majus* genome. The maximal likelihood tree was calculated using raxml-ng with 1000 bootstrap and GTR+G substitution model. The color of gene ID denoted the chromosome the gene within as indicated in the legend. Heatmap color represent Z-scores derived from RNA-seq expression data for each gene. SLF genes are marked with black star.

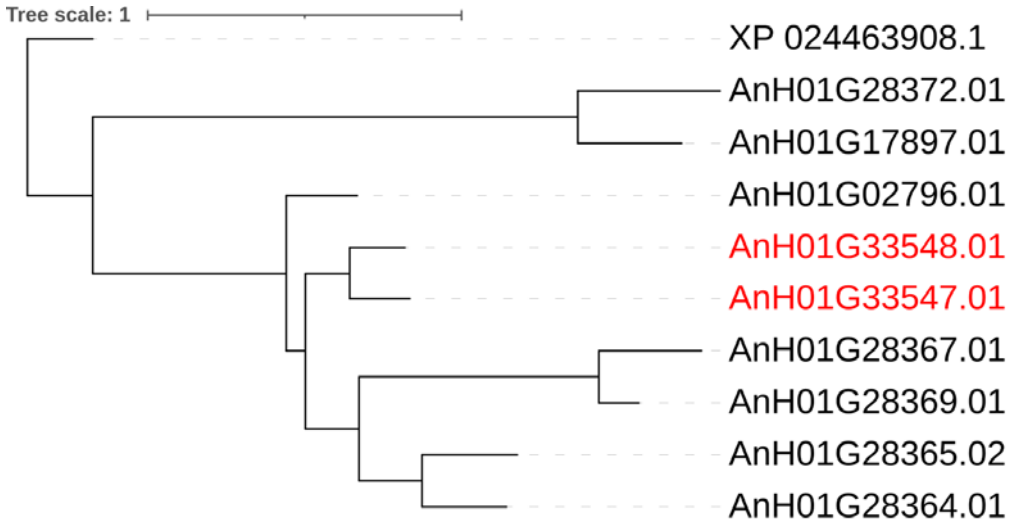


Fig. S22 Phylogenetic tree reconstructed based on the amino acid sequences of all organ specific protein genes from *A.hispanicum* genome with XP_024463908.1 as outgroup, using raxml-ng.

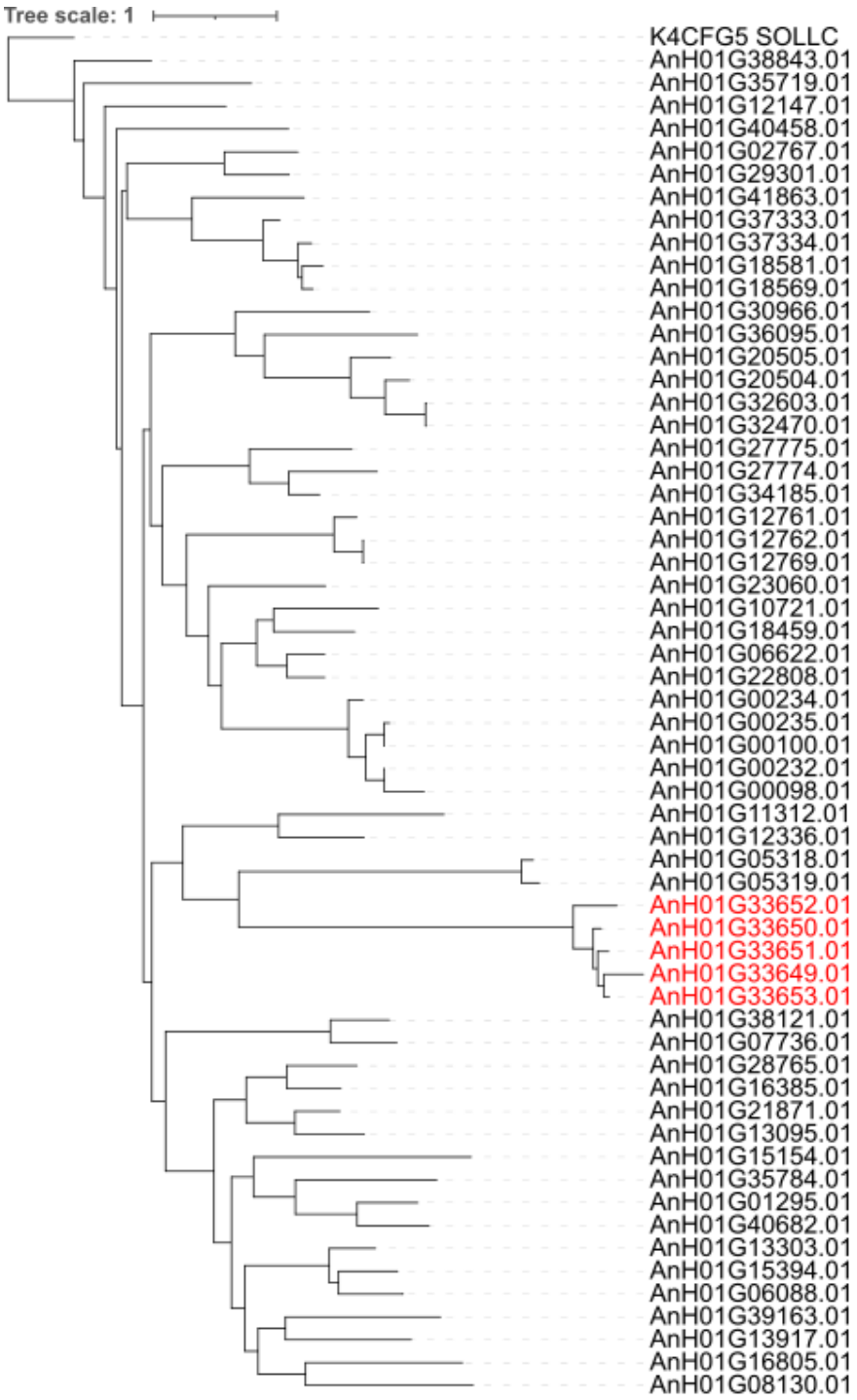


Fig. S23 Phylogenetic tree reconstructed based on the amino acid sequences of all phytocyanin domain like genes from *A.hispanicum* genome with K4CFG5 as outgroup, using raxml-ng.

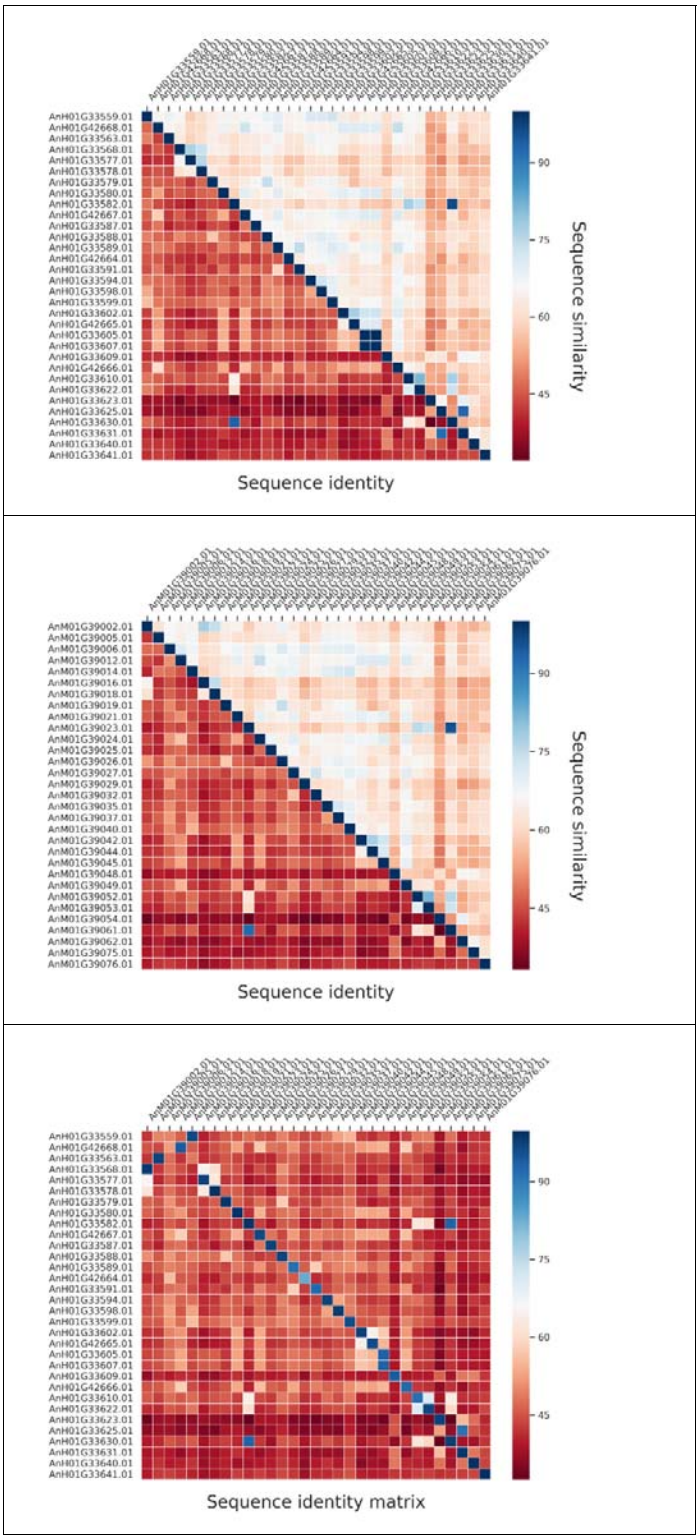


Fig. S24 Heatmap showing sequence identity (lower left triangle) and sequence similarity (upper right triangle) between *SLFs* in SI-Ah, SC-Am, and Ah-vs-Am. The genes in y-axis are ordered by position on the chromosome.

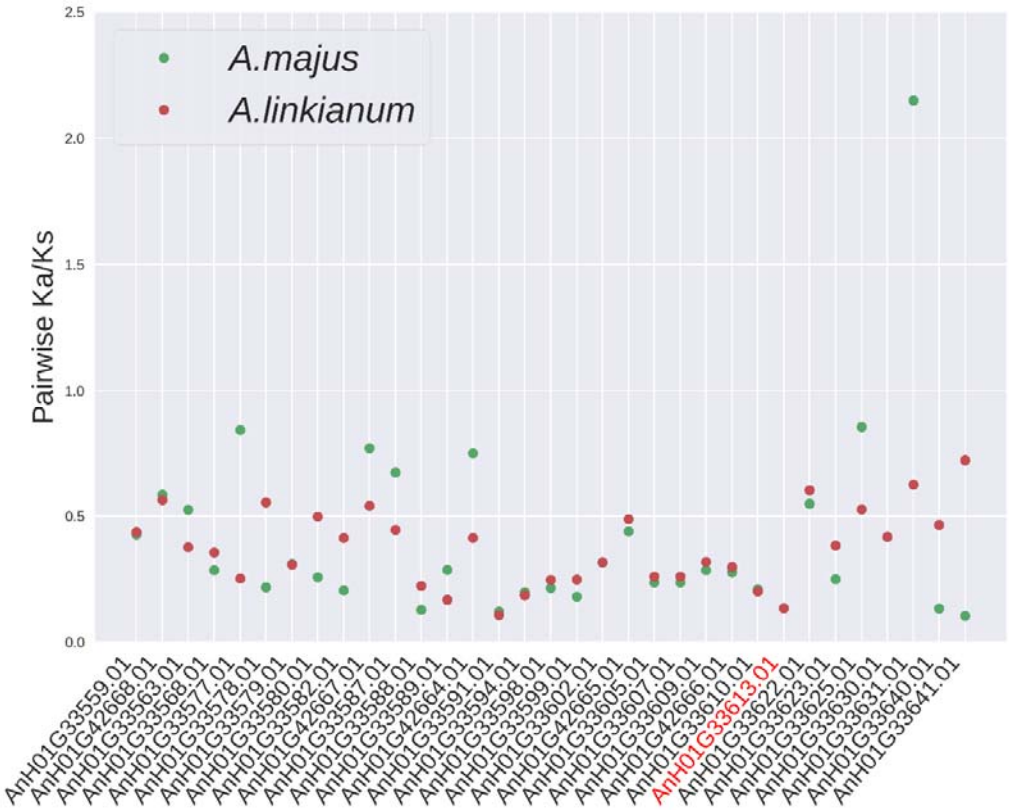


Fig. S25 Pairwise Ka/Ks analysis for all *SLF*s and ribonuclease gene in *S*-locus between *A.hispanicum* and *A.mjus, A.linkianum*. Genes are ordered by genomic coordinates, and label of ribonuclease gene is highlighted in red.

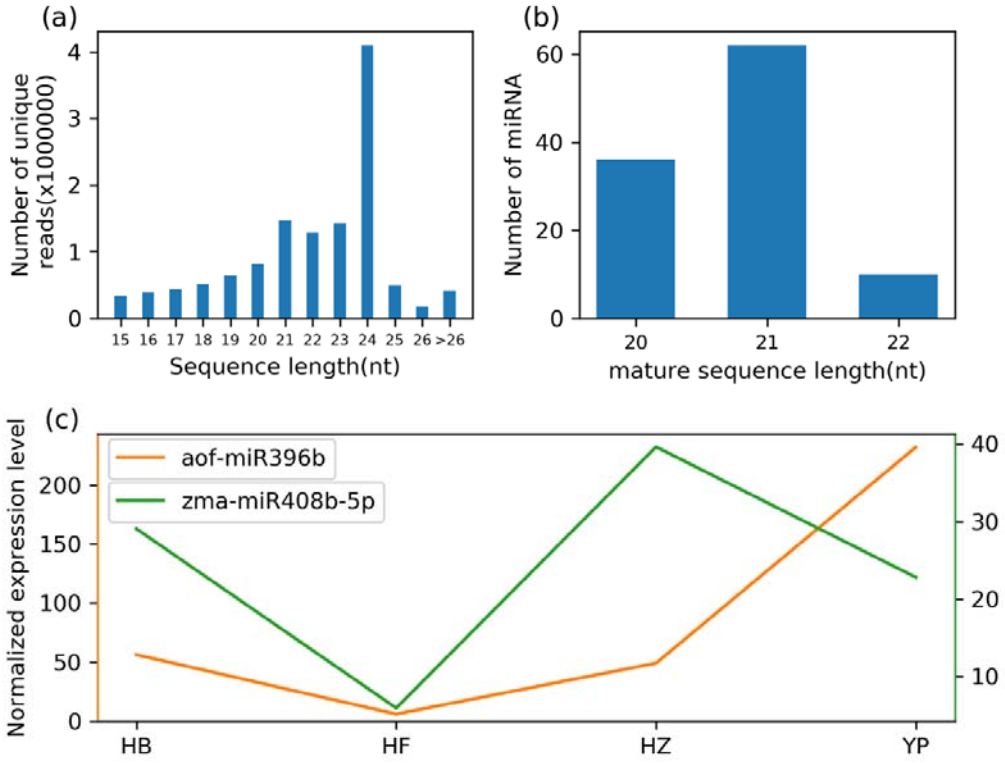


Fig. S26 (a) Length distribution of unique sRNA in four libraries of SI-Ah. Most of the generated reads were 24 and 21 nucleotides. **(b)** Count number of identified miRNAs in SI-Ah. **(c)** Expression level in four tissues of two miRNAs which may play a role in regulate expression of the MYB-related transcription factor, *AnH01G33501.01*.

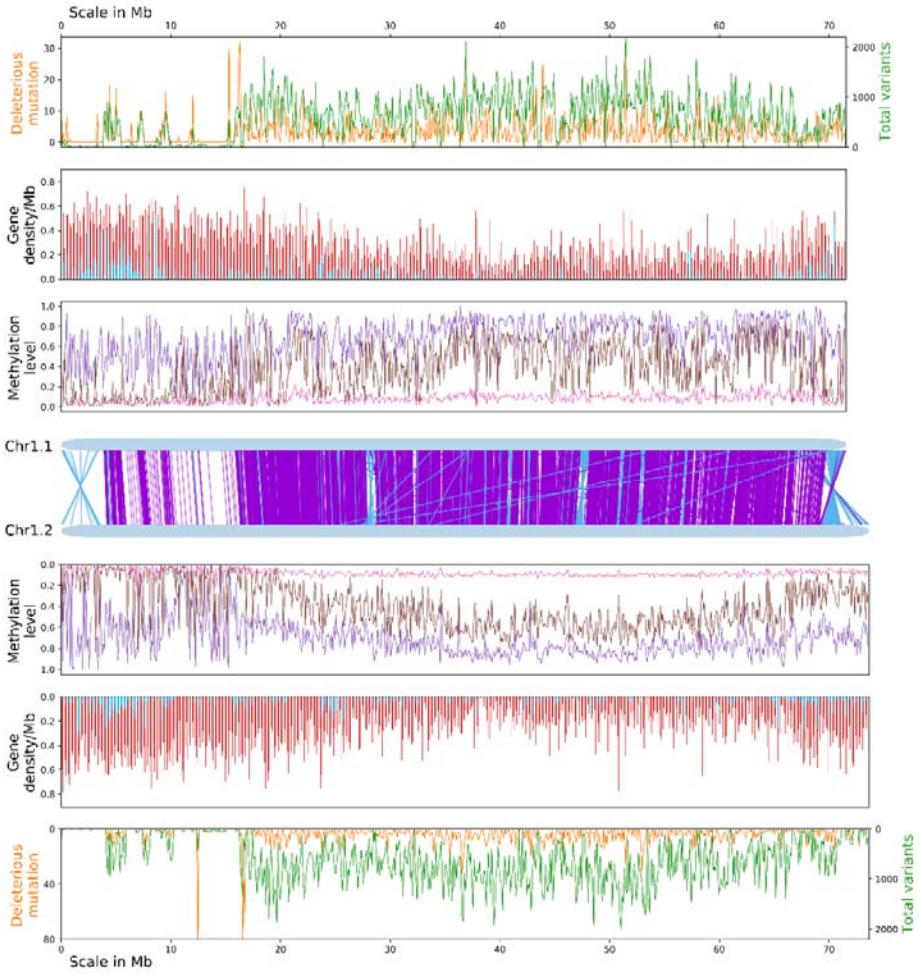


Fig. S27 Alignment of SI-Ah chromosome 1 from two haplomes. The central light blue bars represent homologous chromosome 1 with the purple and dark blue lines indicating the forward and reverse alignments. The distribution of total variants(green), deleterious mutations (orange), annotated genes (pink), 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 windows.

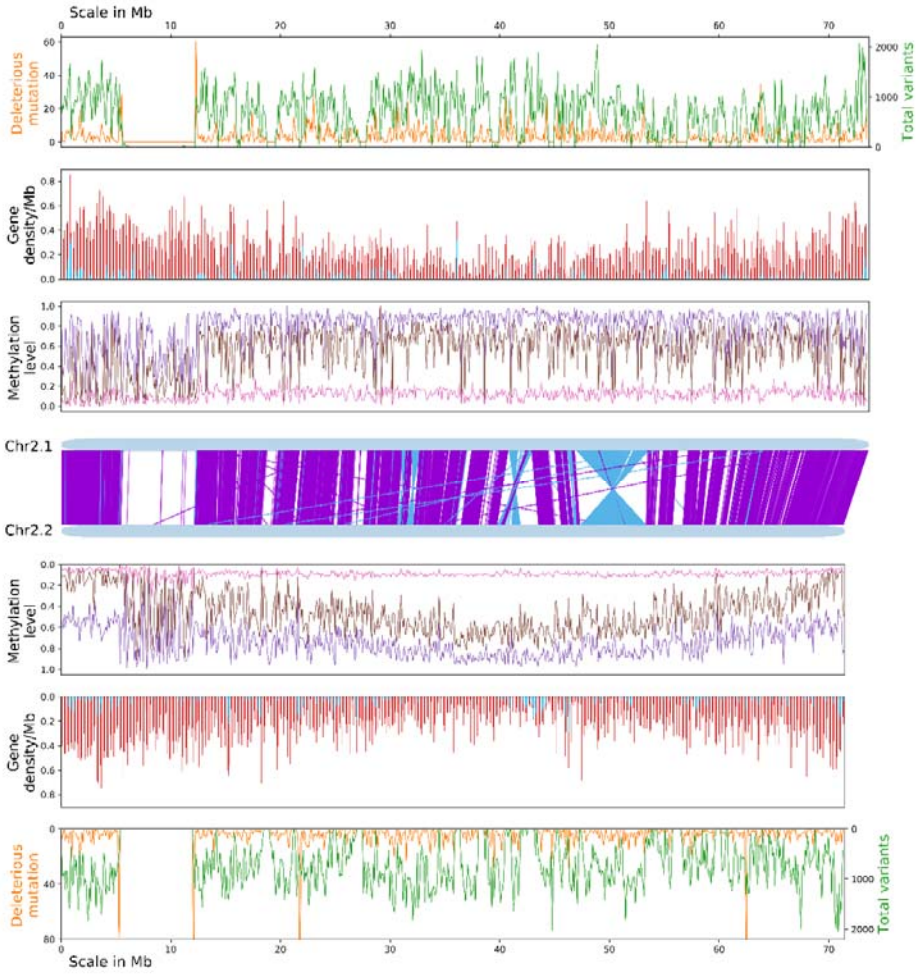


Fig. S28 Alignment of SI-Ah chromosome 2 from two haplomes. The central light blue bars represent homologous chromosome 2 with the purple and dark blue lines indicating the forward and reverse alignments. The distribution of total variants(green), deleterious mutations (orange), annotated genes (pink), 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 windows.

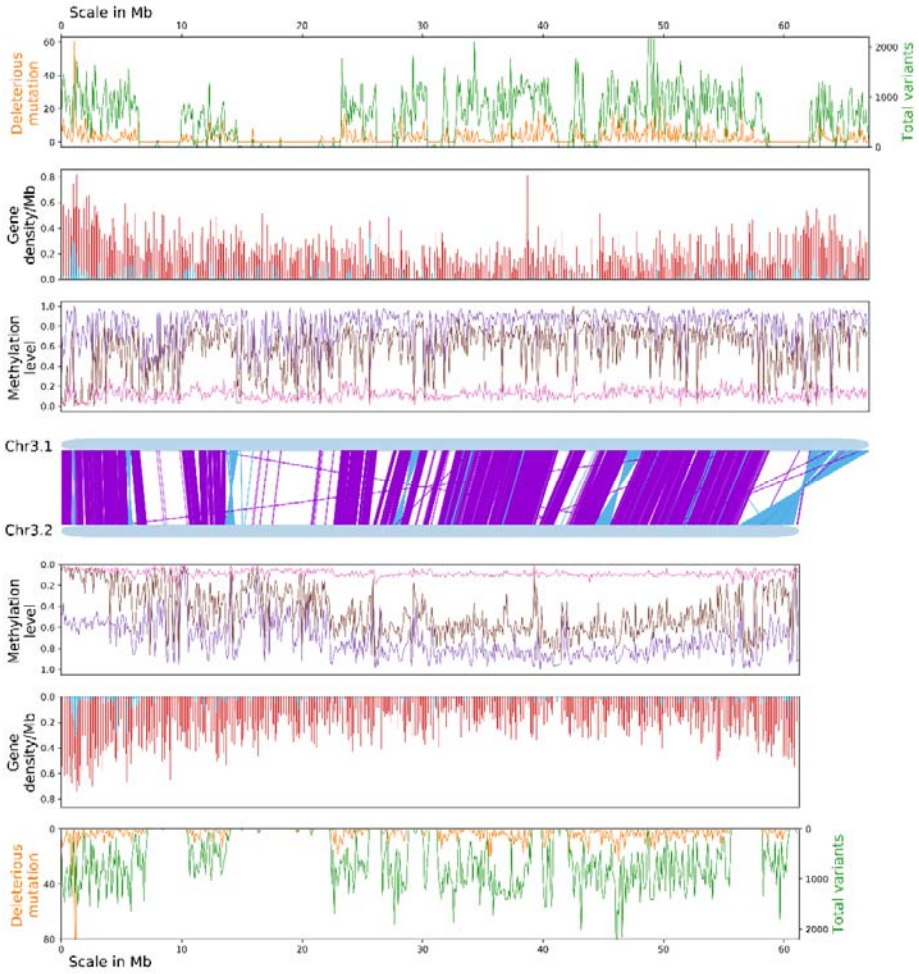


Fig. S29 Alignment of SI-Ah chromosome 3 from two haplomes. The central light blue bars represent homologous chromosome 3 with the purple and dark blue lines indicating the forward and reverse alignments. The distribution of total variants(green), deleterious mutations (orange), annotated genes (pink), 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 windows.

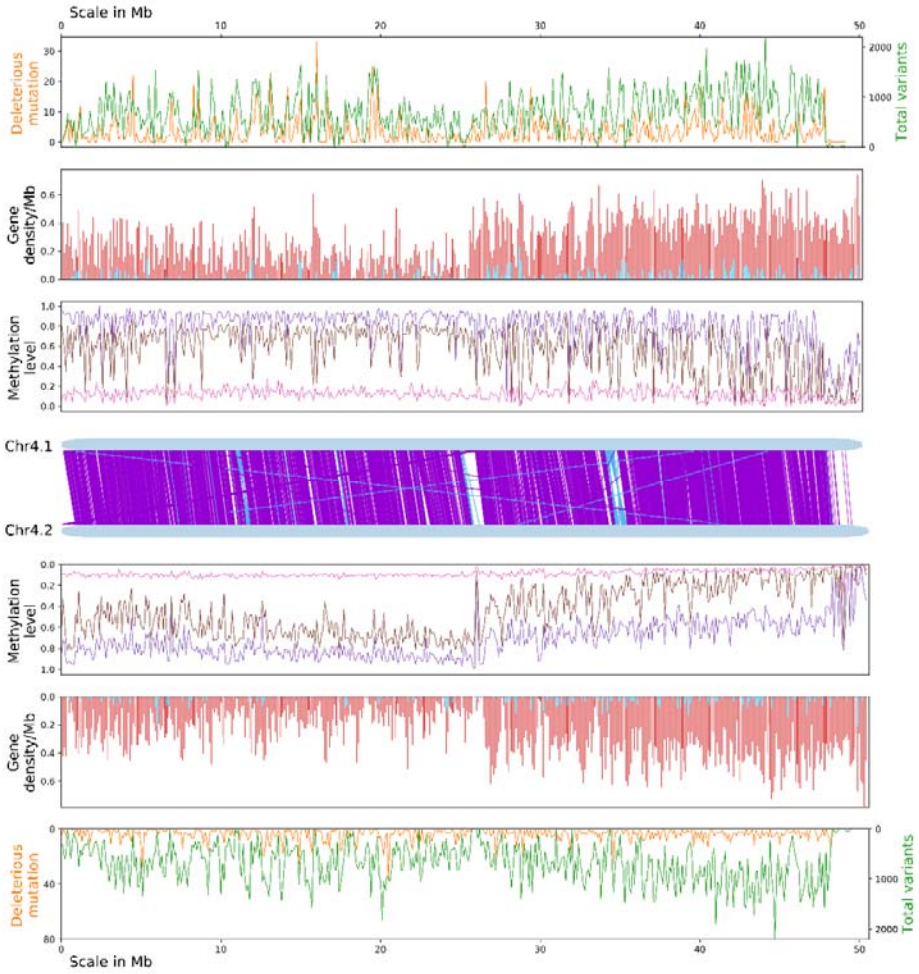


Fig. S30 Alignment of SI-Ah chromosome 4 from two haplomes. The central light blue bars represent homologous chromosome 4 with the purple and dark blue lines indicating the forward and reverse alignments. The distribution of total variants(green), deleterious mutations (orange), annotated genes (pink), 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 windows.

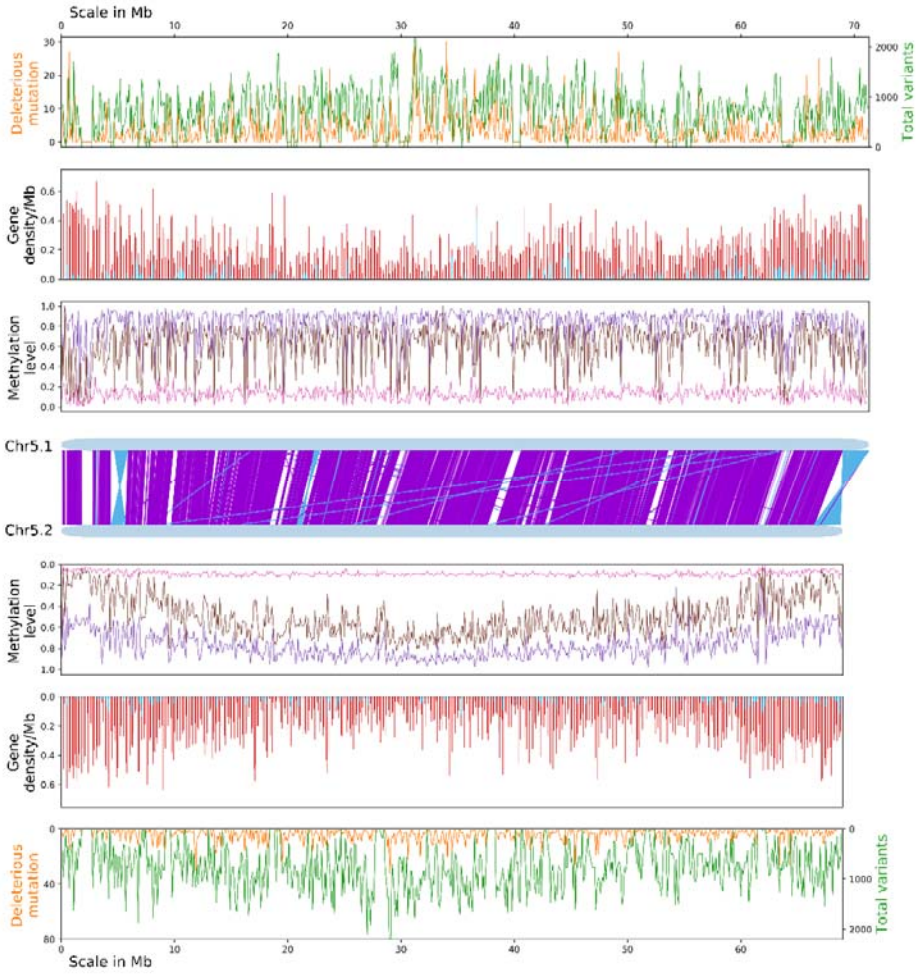


Fig. S31 Alignment of SI-Ah chromosome 5 from two haplomes. The central light blue bars represent homologous chromosome 5 with the purple and dark blue lines indicating the forward and reverse alignments. The distribution of total variants(green), deleterious mutations (orange), annotated genes (pink), 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 windows.

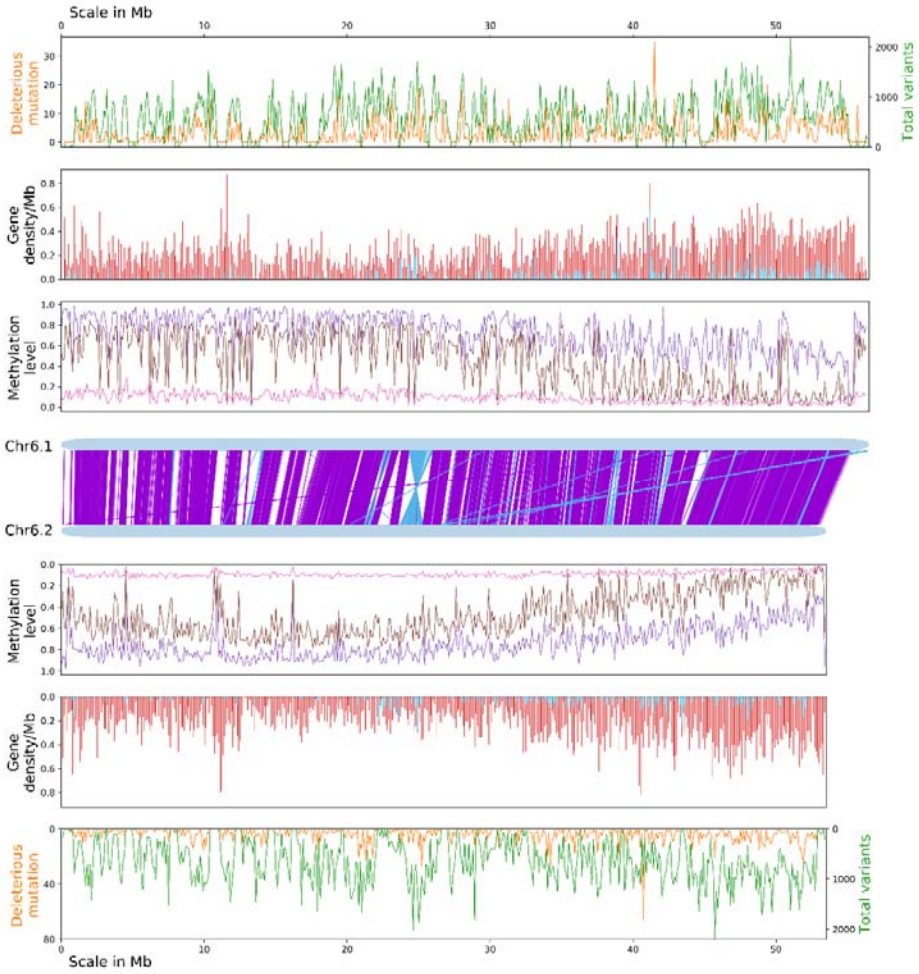


Fig. S32 Alignment of SI-Ah chromosome 6 from two haplomes. The central light blue bars represent homologous chromosome 6 with the purple and dark blue lines indicating the forward and reverse alignments. The distribution of total variants(green), deleterious mutations (orange), annotated genes (pink), 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 windows.

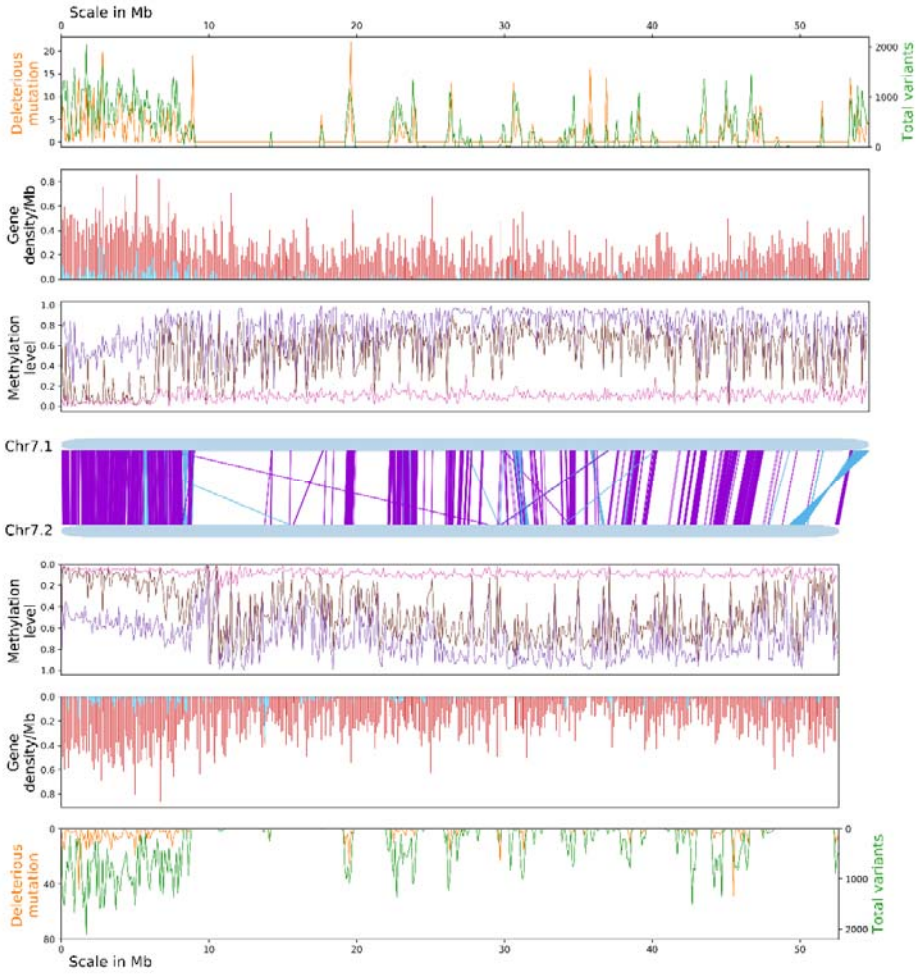


Fig. S33 Alignment of SI-Ah chromosome 7 from two haplomes. The central light blue bars represent homologous chromosome 7 with the purple and dark blue lines indicating the forward and reverse alignments. The distribution of total variants(green), deleterious mutations (orange), annotated genes (pink), 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 windows.

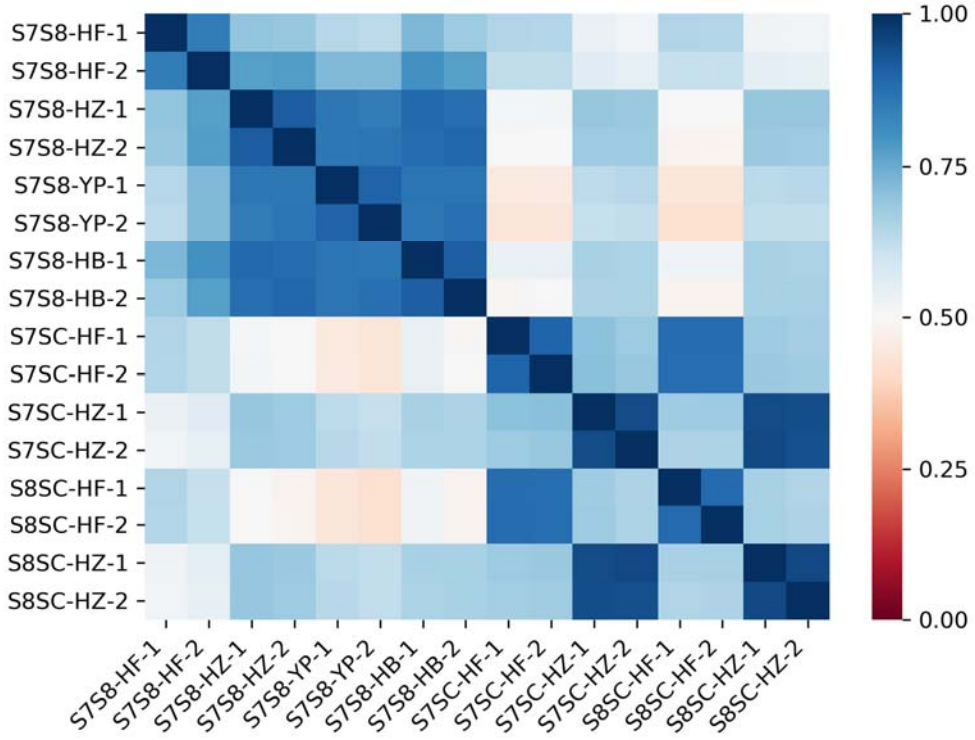


Fig. S34 Spearman correlation heatmap of 16 RNA-seq samples. HB: petal, HF: pollen, YP: leaf, HZ: style.

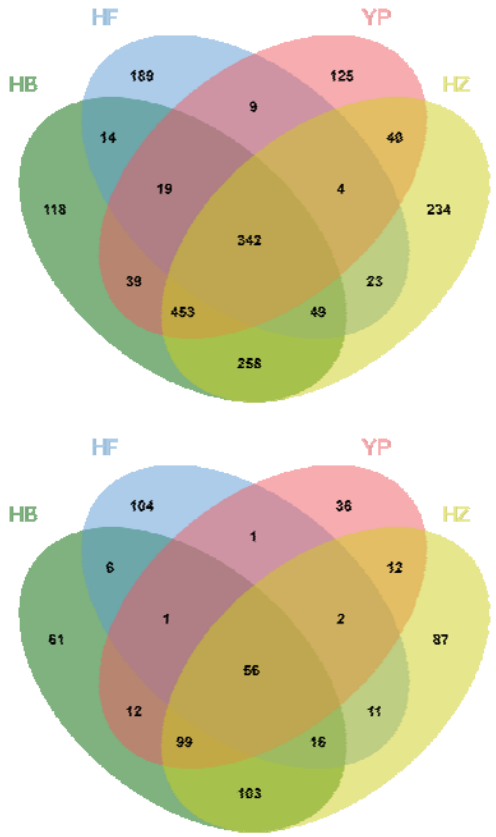


Fig. S35 Venn diagram of imbalanced expressed gene pairs across four tissues (S_7-dominant and S_8-dominant). HB: petal, HF: pollen, YP: leaf, HZ: style.