PCP
PLANT & CELL PHYSIOLOGY

Regular Paper

Genome Re-Sequencing of Diverse Sweet Cherry (*Prunus avium*) Individuals Reveals a Modifier Gene Mutation Conferring Pollen-Part Self-Compatibility

Kentaro Ono[1,4], Takashi Akagi[1,2,4], Takuya Morimoto[1,4], Ana Wünsch[3,]* and Ryutaro Tao[1,]*

[1]Laboratory of Pomology, Graduate School of Agriculture, Kyoto University, Kyoto, 606-8502 Japan
[2]Japan Science and Technology Agency (JST), PRESTO, Kawaguchi-shi, Saitama 332-0012, Japan
[3]Unidad de Hortofruticultura, Centro de Investigación y Tecnología Agroalimentaria de Aragón (CITA), Instituto Agroalimentario de Aragón - IA2 (CITA-Universidad de Zaragoza), Avda, Montañana 930, 50059 Zaragoza, Spain
[4]These authors contributed equally to this work
*Corresponding authors: Ryutaro Tao, E-mail, rtao@kais.kyoto-u.ac.jp; Fax, +81-75-753-6497; Ana Wünsch, E-mail, awunsch@aragon.es; Fax, +34-976-716335.

(Received February 22, 2018; Revised March 23, 2018)

The S-RNase-based gametophytic self-incompatibility (GSI) reproduction barrier is important for maintaining genetic diversity in species of the families Solanaceae, Plantaginaceae and Rosaceae. Among the plant taxa with S-RNase-based GSI, *Prunus* species in the family Rosaceae exhibit *Prunus*-specific self-incompatibility (SI). Although pistil *S* and pollen *S* determinants have been identified, the mechanism underlying SI remains uncharacterized in *Prunus* species. A putative pollen-part modifier was identified in this study. Disruption of this modifier supposedly confers self-compatibility (SC) to sweet cherry (*Prunus avium*) 'Cristobalina'. To identify the modifier, genome re-sequencing experiments were completed involving sweet cherry individuals from 18 cultivars and 43 individuals in two segregating populations. Cataloging of subsequences (35 bp kmers) from the obtained genomic reads, while referring to the mRNA sequencing data, enabled the identification of a candidate gene [*M locus-encoded GST* (*MGST*)]. Additionally, the insertion of a transposon-like sequence in the putative *MGST* promoter region in 'Cristobalina' down-regulated *MGST* expression levels, probably leading to the SC of this cultivar. Phylogenetic, evolutionary and gene expression analyses revealed that *MGST* may have undergone lineage-specific evolution, and the encoded protein may function differently from the corresponding proteins encoded by *GST* orthologs in other species, including members of the subfamily Maloideae (Rosaceae). Thus, MGST may be important for *Prunus*-specific SI. The identification of this novel modifier will expand our understanding of the *Prunus*-specific GSI system. We herein discuss the possible functions of MGST in the *Prunus*-specific GSI system.

Keywords: 'Cristobalina' • cherry • Pollen-part modifier • Self-incompatibility • S-RNase • Subsequence cataloging.

Abbreviations: BWA, Burrows–Wheeler Aligner; DEG, differentially expressed gene; GSI, gametophytic self-incompatibility; GST, glutathione *S*-transferase; MGST, M locus-encoded GST; qRT-PCR, quantitative real-time PCR; RPKM, reads per kilobase per million total reads; SC, self-compatibility; SCK, SC-specific kmer; SFB, *S* haplotype-specific F-box; SFBB, *S* locus F-Box Brothers; SI, self-incompatibility; SLF, *S* locus F-box; TE, transposable element.

Footnotes: The nucleotide sequences reported in this paper have been submitted to the DDBJ database with accession numbers PRJDB6734, LC371283, and LC371380.

Introduction

Self-incompatibility (SI) is a major reproductive strategy of flowering plants to maintain genetic diversity within a species (De Nettancourt 2001). The plant families Rosaceae, Solanaceae and Plantaginaceae share the RNase-based gametophytic SI (GSI) system. Self- and non-self recognition in this system is controlled by a haploblock, called the *S* locus, which encodes the pistil *S* determinant S-RNase (S-RNase) and pollen *S* determinant F-box protein(s). The pollen F-box proteins are named *S* locus F-box (SLF) proteins in the families Solanaceae and Plantaginaceae, *S* locus F-Box Brothers (SFBB) in the subtribe Malinae of the family Rosaceae and *S* haplotype-specific F-box (SFB) in *Prunus* species of the family Rosaceae (McCubbin and Kao 2000, McClure 2009, Sassa et al. 2010, Tao and Iezzoni 2010, Meng et al. 2011). These families use monophyletic homologous S-RNases as the pistil *S* determinant (Igic and Kohn 2006, Morimoto et al. 2015). Because the most recent common ancestor of the three families is the ancestor of 75% of eudicot families, RNase-based SI is believed to be the ancestral state in most eudicots (Igic and Kohn 2006). However, the pollen *S*, *SLF/SFBB/SFB*, evolved in a lineage-specific manner probably because of the rapid birth/death of F-box genes (Xu et al. 2009), which may have triggered the taxa-dependent functional diversification of pollen *S* F-box proteins (Ushijima et al. 2003, Sonneveld et al. 2005, Aguiar et al. 2015, Kubo et al. 2015, Akagi et al. 2016). Several lines of evidence have revealed that the SI recognition mechanism in *Prunus* species, especially regarding the pollen *S* determinant function, is distinct from the

Plant Cell Physiol. 59(6): 1265–1275 (2018) doi:10.1093/pcp/pcy068, Advance Access publication on 4 April 2018, available online at www.pcp.oxfordjournals.org

corresponding mechanism in other taxa that exhibit S-RNase-based GSI (Ushijima et al. 2004, Hauck et al. 2006, Sonneveld et al. 2005, Tao and Iezzoni 2010). Multiple pollen *S* F-box genes in the *S* locus are assumed to be involved in degrading and detoxifying non-self S-RNases, but not self S-RNase, which functions as a cytotoxin in the pollen tube in the families Solanaceae and Plantaginaceae and in the subtribe Malinae (Kubo et al. 2010, Minamikawa et al. 2010, Kakui et al. 2011, Kubo et al. 2015). Meanwhile, a single pollen *S* determinant SFB is believed to mediate the cytotoxicity of the self S-RNase in *Prunus* species (Tao and Iezzoni 2010, Matsumoto and Tao 2016). However, little is known about the molecular basis for the functional diversification of pollen *S* F-box genes.

In addition to the *S* locus SI/self-compatibility (SC) specificity determinant genes, other genes located outside the *S* locus are expressed non-specifically and independently of the *S* haplotype. Thus, characterizing these genes may be important for elucidating the functional diversification and evolution of the S-RNase-based GSI system. These 'self-incompatibility modifier' genes are expressed at various stages of the SI/SC reaction in the S-RNase-based GSI system (Goldraij et al. 2006, Zhao et al. 2010, Entani et al. 2014). In the family Solanaceae, two genes expressed in the pistil, *HT-B* and *120 K*, are candidate pistil modifier genes (Goldraij et al. 2006). Additionally, *PhSSK1* and *PhCUL1*, which are expressed in the pollen tube and encode components of the SCF complex, are probably the pollen-part modifier genes in *Petunia* species (Zhao et al. 2010, Entani et al. 2014). *Prunus* species mutants exhibiting SC presumably caused by mutations to the pollen-part modifier genes have also been described. 'Cristobalina' is a Spanish sweet cherry (*Prunus avium*) cultivar whose SC is probably conferred by a mutation to a modifier gene located at the edge of chromosome 3 (Wünsch and Hormaza 2004, Cachi and Wünsch 2011). Although 'Cristobalina' exhibits SC, pollen tube growth is significantly lower during self-pollination than during cross-pollination (Cachi et al. 2014). Therefore, 'Cristobalina' may be more appropriately described as exhibiting semi-SC. A pollen-part modifier gene conferring SC was also detected in apricot (*Prunus armeniaca*) (Zuriaga et al. 2012, Zuriaga et al. 2013). This modifier gene was localized to apricot chromosome 3, in a genomic region similar to where the 'Cristobalina' modifier gene is located.

In the era of 'omics'-based studies, genome-wide sequencing approaches have been used to identify causal mutations in non-model plant species, even with minimal reference genome information. For example, Illumina gDNA-sequencing (gDNA-Seq) and mRNA-Seq data were assembled to identify the sex-determining gene in *Diospyros* species with no available reference genome information (Akagi et al. 2014). We herein describe our attempts to use large-scale DNA sequencing technologies to identify the pollen-part modifier gene (*M*) in the 'Cristobalina' cultivar. Cataloging of genomic DNA subsequences with mRNA-Seq data from diverse genotypes, including a few individuals in two segregating populations, enabled us to identify the candidate gene. We discuss the lineage-specific evolution of this candidate gene, which may be useful for functionally characterizing the *Prunus*-specific GSI system.

Results

Identification of candidate polymorphisms representing the *M* locus

A co-segregation test using the polymorphisms in a segregating population was effective for detecting the genetic region linked to the *M* locus (**Fig. 1A**). However, an association test using the polymorphisms in diverse randomly selected cultivars was effective for discarding additional polymorphisms that are difficult to sort out in a small segregating population (**Fig. 1B**). We combined these two tests to identify SC-specific genome sequences. The SC-specific genome contigs were obtained by cataloging subsequences from 43 individuals in two segregating populations and 18 sweet cherry cultivars, including 'Cristobalina', based on Illumina gDNA-Seq data (**Fig. 2**; Supplementary Table S1). We cataloged the 35-mer subsequences (kmers) starting with an 'A' nucleotide among the random gDNA-Seq paired-end 150 bp (PE150) or paired-end 100 bp (PE100) reads for 18 SC and 25 SI F_1 individuals (approximately ×64 and ×114 sequence coverage for the SC and SI individuals, respectively) from the crosses between sweet cherry (*P. avium*) cultivars, 'Brooks' (S_1S_9MM) and 'Cristobalina' (S_3S_6Mm) (B×C) (Cachi and Wünsch 2011) and 'Lambert' (S_3S_4MM) and 'Cristobalina' (L×C). A comparison of the SC and SI kmer pools resulted in 31,388 pre-SC-specific kmers (pre-SCKs, coverage ×10 or more) (**Fig. 2A**) that were highly abundant at the bottom edge of chromosome 3 (**Fig. 2B**), which is where the *M* locus has been genetically mapped (Cachi and Wünsch 2011). Next, we compared the SC population-specific kmers with the kmers extracted from 17 sweet cherry cultivars with a homozygous *MM* genotype ('Ambrunes', 'Benishuho', 'Brooks', 'Colt', 'Ferrovia', 'Gassannishiki', 'Hedelfingen', 'Lambert', 'Napoleon', 'Rainier', 'Sam', 'Satonishiki', 'Stella', 'Sue', 'Summit', 'Takasago' and 'Vic') to select the SCKs further (**Fig. 1B**). With this information, the 31,388 pre-SCKs were filtered to 661 SCKs supposedly including candidate polymorphisms at the *M* locus. The paired-end Illumina reads containing SCKs were assembled by the CAP3 assembler (Huang and Madan 1999) to construct the regions surrounding the SCKs, ultimately yielding 235 contigs. Errors were removed from artifacts and recombinant contigs by genotyping and filtering of the polymorphisms via SC and SI mapping of the contigs. Thus, 29 candidate contigs were obtained.

Anchoring the expressed genes surrounding the candidate contigs

The 29 candidate contigs were annotated by blastn analyses using the *Prunus persica* Whole Genome Assembly v2.0 & Annotation v2.1 (v2.0.a1) database (Verde et al. 2013) and the *Prunus avium* Whole Genome Assembly v1.0 & Annotation v1 (v1.0.a1) database (Shirasawa et al. 2017). Of the 29 contigs, only eight overlapped with genic and/or regions flanking a gene (Supplementary Table S2). The mRNA-Seq reads for the pollen grains of 12 SC and 15 SI individuals of the B×C/L×C populations as well as the pollen grains of 18 sweet cherry cultivars (Supplementary Table S1) were mapped to the genes in the sweet cherry genomes (n = 43,673) and to one peach candidate

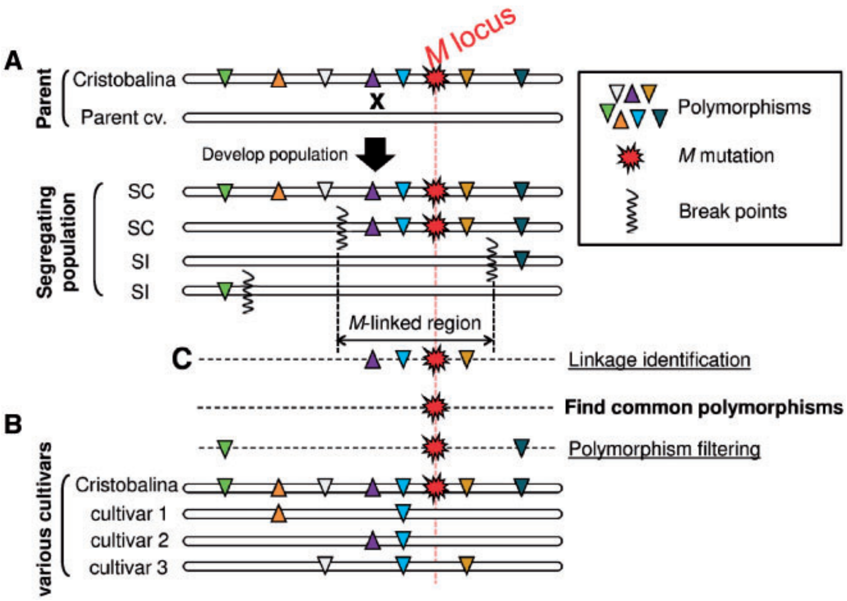

Fig. 1 Schematic representation of the method used to identify the *M* locus. Two genomic approaches were used to filter the candidate polymorphisms in segregating populations (A) and in diverse cultivars (B). (A) A comparison of the genomic pools of segregated SC and SI lines detected the polymorphisms in a genetically limited region (*M*-linked region) in which the length (or number of polymorphisms) depended on break points in the recombinants. (B) Filtering of the 'Cristobalina'-specific polymorphisms by subtracting those in various SI cultivars, which is not completely dependent on the linkage to the *M* locus. (C) Integration of these two approaches resulted in relatively few polymorphisms in the *M*-linked region.

gene to estimate the expression levels in terms of reads per kilobase per million total reads (RPKM) (Supplementary Data S1). Four genes overlapping the candidate genomic contigs were significantly expressed in pollen grains (**Table 1**, RPKM >1). One of the four significantly expressed candidate genes, which encodes a glutathione *S*-transferase (GST) kappa protein (**Table 1**), was located on the bottom edge of chromosome 3, which corresponds to the middle of the SCK peak (**Fig. 3**) and the *M* locus previously identified by genetic mapping (Cachi and Wünsch 2011). The genomic synteny in this region was high between sweet cherry (subgenus *Cerasus*) and peach (subgenus *Amygdalus*), and the orthologous genes present were highly conserved (**Fig. 3B, C**). Although the candidate *GST*-like gene is concatenated to another adjacent *GST*-like gene to form a single gene (Pav.sc0000661.g340.1) in the draft sweet cherry genome database (Shirasawa et al. 2017; http://cherry.kazusa.or.jp/), we confirmed that they are expressed separately (**Fig. 5A;** Supplementary Fig. S1). Thus, they were considered to be independent *GST* genes (Pav.sc0000661.g340.1-1 and Pav.sc0000661.g340.1-2), similar to the corresponding genes in the peach genome. We named the candidate *GST*-like gene (Pav.sc0000661.g340.1-1) *MGST* (*M* locus-encoded GST). Of the genes in the region surrounding *MGST*, only three were substantially expressed similar to *MGST* in pollen grains (RPKM >50) (**Fig. 3D, E**). However, these three genes did not contain SC- or SI-specific polymorphisms in the genic or promoter sequences, consistent with the SCK cataloging results.

Characterization of the *MGST* mutation

Two SCK-derived genomic contigs were detected in the region flanking *MGST* (Supplementary Fig. S2). An alignment of these genomic contigs with reference genomic sequences (Shirasawa et al. 2017) suggested the presence of an insertion in the *MGST* promoter region 280 bp upstream from the start codon (**Fig. 4A**). Sanger7 sequencing of the insertion amplified by a PCR revealed a 1,848 bp putative transposable element (TE) (**Fig. 4A**). This insertion was specifically conserved in the SC individuals in the two F_1 populations and in SC sweet cherry cultivars ('Talegal Ahín' and 'Son Miró') as well as in 'Cristobalina' (**Fig. 4B**). Although 'Stella' is SC, the causal mutation for SC is located in the *S* locus (Ushijima et al. 2004) and thus no PCR amplification specific to the TE insertion was observed. The TE sequences were identical in the three SC cultivars, 'Talegal Ahín', 'Son Miró' and 'Cristobalina'. Considering these cultivars are all from the same region of eastern Spain (Cachi and Wünsch 2014), the allele conferring SC probably evolved recently, and then it dispersed in the surrounding geographical area. Sequences of about 200 bp were commonly amplified by PCR in all F_1 progeny and cultivars regardless of their phenotypes (SC/SI) (data not shown). The 200 bp band appeared to be from an intact *M* allele, suggesting that 'Talegal Ahín' and 'Son Miró' carry a heterozygous *Mm* genotype, similar to 'Cristobalina'.

We examined the *MGST* expression pattern in flower organs (sepal, petal, filament, ovary, pistil and pollen grains) and leaves. Expression levels were higher in pollen grains than in other flower organs for both heterozygous SC ('Cristobalina', *Mm*) and homozygous SI ('Satonishiki', *MM*) cultivars (**Fig. 5A**). However, significant *MGST* expression was detected in the other tested organs with increasing PCR cycles. Thus, *MGST* was defined as a pollen-enriched gene that is also expressed in other flower organs, but at lower levels. To examine the effect

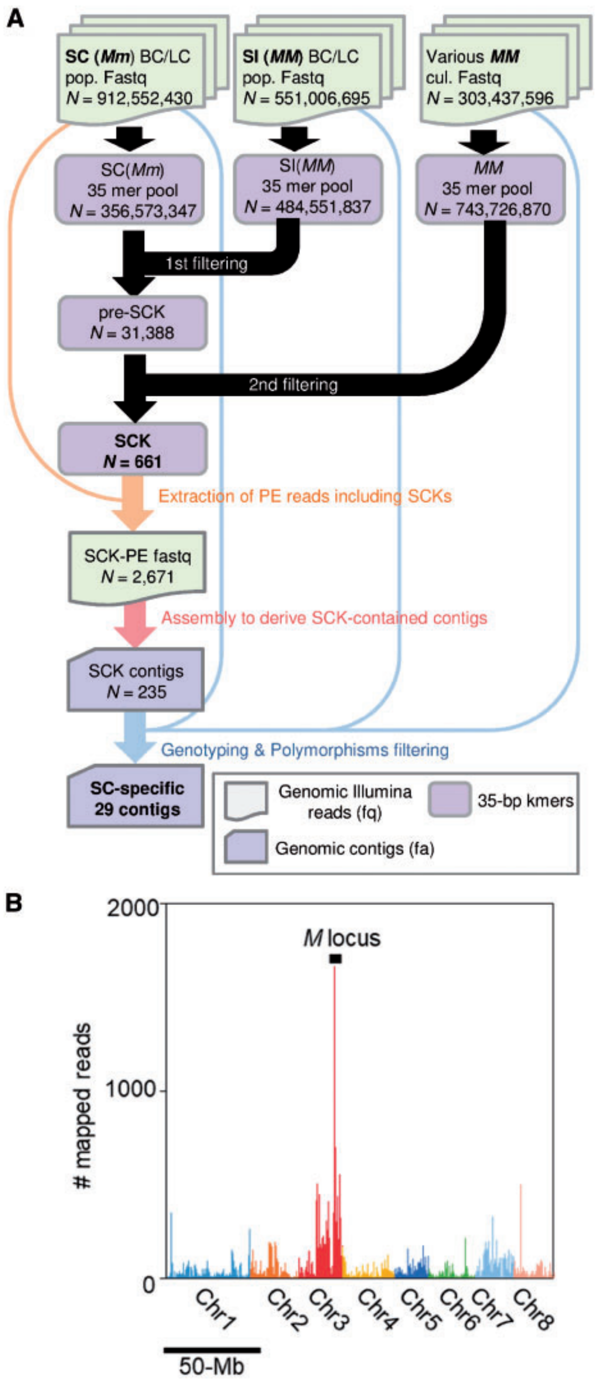
A
SC (*Mm*) BC/LC pop. Fastq N = 912,552,430
SI (*MM*) BC/LC pop. Fastq N = 551,006,695
Various *MM* cul. Fastq N = 303,437,596
SC(*Mm*) 35 mer pool N = 356,573,347
SI(*MM*) 35 mer pool N = 484,551,837
MM 35 mer pool N = 743,726,870
1st filtering
pre-SCK N = 31,388
2nd filtering
SCK N = 661
Extraction of PE reads including SCKs
SCK-PE fastq N = 2,671
Assembly to derive SCK-contained contigs
SCK contigs N = 235
Genotyping & Polymorphisms filtering
SC-specific 29 contigs
Genomic Illumina reads (fq)
35-bp kmers
Genomic contigs (fa)
B
2000
1000
0
# mapped reads
M locus
Chr1
Chr2
Chr3
Chr4
Chr5
Chr6
Chr7
Chr8
50-Mb

Fig. 2 Cataloging of the SC-specific kmers (SCKs) and their genomic distribution. (A) Flowchart of the method used for filtering the SCKs from genomic fastq data for the SC/SI segregating populations and cultivars, and for assembling the contigs including SCKs. Gray, orange and purple boxes represent Illumina fastq data, 35-mer kmers and genomic contigs, respectively. The two independent filtering methods involving segregating populations and various cultivars (**Fig.1**) resulted in the detection of 661 SCKs constituting 29 genomic contigs in which polymorphisms were perfectly co-segregated with the SC/SI phenotypes. (B) Genomic distribution of the reads, including pre-SCKs, in the segregated lines [after the first filtering in (A), *n* = 108,354]. These reads were significantly enriched in the putative *M*-linked region at the bottom of chromosome 3, which is consistent with the results of a previous study (Cachi and Wünsch 2011).

of the TE-like insertion in the promoter region, we compared the *MGST* expression levels in the pollen grains with *MM*, *Mm* and *mm* genotypes. A quantitative real-time PCR (qRT-PCR) analysis of *Mm* ('Cristobalina') and *MM* ('Satonishiki') individuals as well as four *mm* individuals produced from the selfing of 'Cristobalina' plants revealed the *m* allele-specific substantial decrease in *MGST* expression levels and the *M* allele dosage-dependent expression pattern (**Fig. 5B**). Additionally, the *MGST* transcript sequence in 'Cristobalina' included no specific substitutions when compared with those of the analyzed SI cultivars. Other candidate genes (six cherry genes and one peach gene), which were identified based on the assembled SCKs, had none of the potentially disruptive mutations affecting protein sequences or gene expression levels (data not shown).

Whole-genome expression profiling in SC and SI pollen grains

The expression profiles based on the pollen mRNA-Seq data for 12 SC (*Mm*) and 15 SI (*MM*) individuals of the B×C/L×C populations and 18 sweet cherry cultivars were analyzed to detect the differentially expressed genes (DEGs) (Supplementary Data S1). In the sweet cherry genome, DESeq analysis (Anders and Huber 2010) resulted in the identification of 160 DEGs [false discovery rate <0.05, RPKM >1] (Supplementary Table S3). Moreover, *MGST* was one of the DEGs (false discovery rate = 0.048) with an expression ratio (1.97-fold higher in SI than in SC individuals) that fits the theoretical value for the comparison between the *Mm* and *MM* genotypes. This result supported the presumption that the TE-like insertion in the promoter region disrupts *MGST* expression in 'Cristobalina' plants. Enriched Gene Ontology (GO) terms were not detected among the other 159 DEGs ($P > 0.1$, Fisher's exact test), with the peach genome as the background. According to the annotated functions, the DEGs were associated with multiple types of disease resistance gene products, such as NB-ARC domains or chitinase, which are involved in immune responses, and are potentially related to the SI mechanism. The use of individuals with an *Mm* genotype that resulted in both *m* (SC) and *M* (SI) haploid pollen grains made it difficult to analyze the DEGs between *M* and *m* pollen grains thoroughly. Nevertheless, we could observe a considerable up-regulation of gene expression levels in the *m* pollen grains.

Evolution of the *MGST* gene

A phylogenetic tree constructed with *GST*-like genes in the *Arabidopsis thaliana* genome indicated that the *Prunus MGST* gene and its orthologs in eudicots form a single *MGST* clade with AT5G38900 (**Fig. 6A**). In contrast, when *MGST* orthologs nested within the *MGST* clade in diverse representative species were analyzed, two subclades (subclades I and II) were formed with well-supported bootstrap values (95/100 and 87/100, respectively) (**Fig. 6B**). Subclade I comprised *MGST* and its orthologs from two genera, *Prunus* and *Fragaria*, while subclade II included *GST*-like genes from most of the included species. The detection of site-branch-specific positive selection

Table 1 Details regarding candidate genes originating from 29 candidate contigs

Gene name	Originating contig	Genomic location	Descriptions	RPKM[a]	
				SC	SI
Pav_sc0000661.1_g340.1.mk[b]	Contig_13, Contig_194	chr3: 18395989-18401024 bp	GLUTATHIONE S-TRANSFERASE KAPPA (PRUPE_ppa011285mg)	10.2	12.9
Pav_sc0000661.1_g340.1-1 (*MGST*)[b]				52.9	102.8
Pav_sc0000661.1_g340.1-2[b]				14.3	12.6
Pav_sc0000254.1_g670.1.mk	Contig_22	chr2: 16155791-16160496 bp	E3 ubiquitin-protein ligase RLIM-like isoform X1	75.8	60.0
Pav_sc0000254.1_g680.1.mk	Contig_22	chr2: 16155733-16156005 bp		0.0	0.3
Pav_sc0000254.1_g520.1.mk	Contig_305	chr2: 16244198-16250556 bp	E3 ubiquitin-protein ligase RLIM-like isoform X1	25.2	16.0
Pav_sc0000617.1_g440.1.mk	Contig_52	chr4: 12393454-12416603 bp	Polynucleotide 3′-phosphatase ZDP	0.3	0.2
Pav_sc0001710.1_g1630.1.br	Contig_349	chr3: 4128486-4131739 bp		0.0	0.0
Pav_sc0001607.1_g010.1.mk	Contig_31	chr0: 36428484-36429454 bp	Protein trichome birefringence-like 10	0.0	0.0
Prupe.1G105900[c]	Contig_R4	chr1: 8495646-8503110 bp[c]	VESICLE TRANSPORT V-SNARE PROTEIN VTI1-RELATED	16.9	15.5

[a] RPKM values for SC or SI progeny are listed.

[b] Pav_sc0000661.1_g340.1.mk was divided into two genes (*MGST* and Pav_sc0000661.1_g340.1-2) according to the orthologous sequence in the peach genome.

[c] The Contig_R4 sequence was similar to that of a single peach gene (Prupe.1G105900), while it was also similar to the sequence of a non-genic region in the cherry genome sequence.

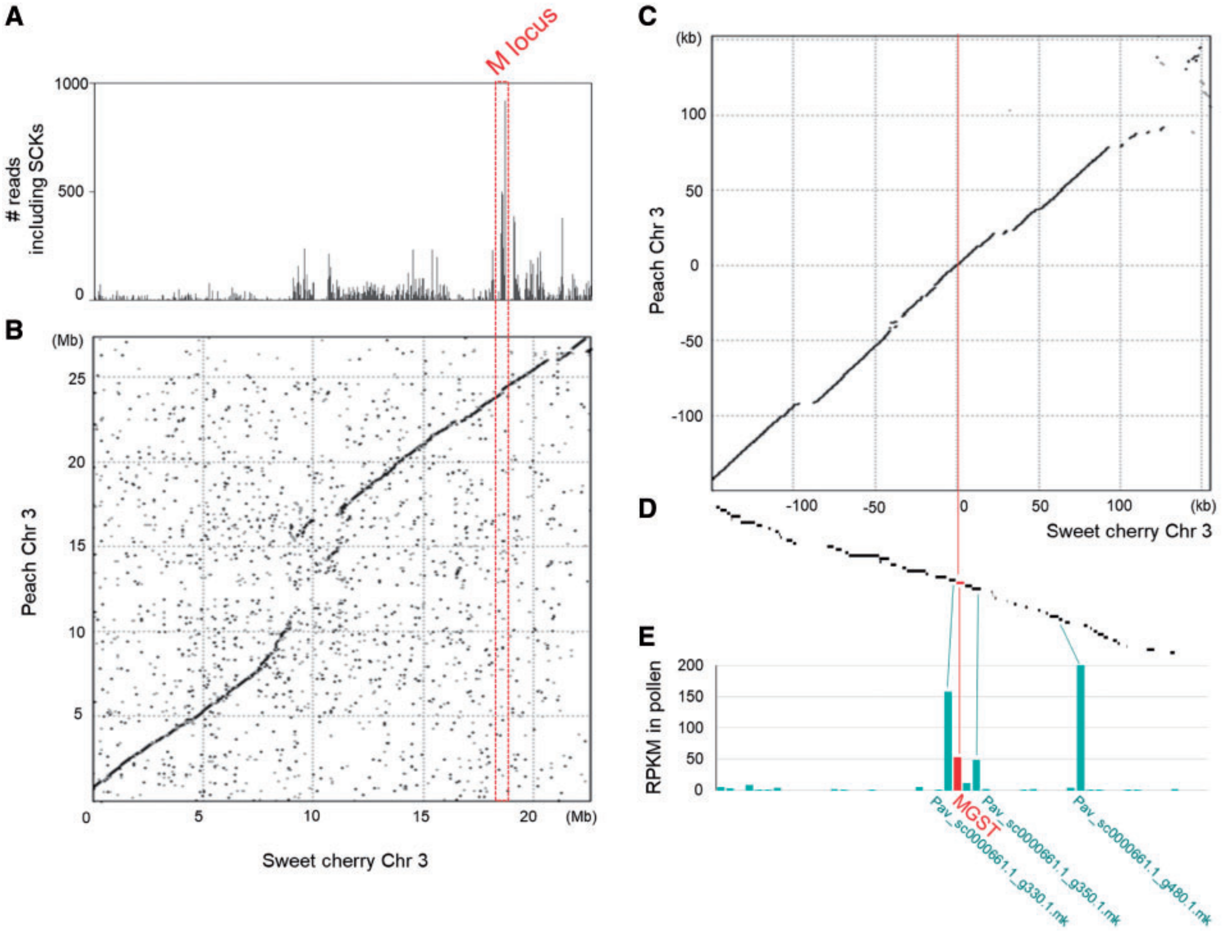

Fig. 3 Characterization of the genomic synteny and the expressed genes in the *M* locus. (A) Distribution of reads containing pre-SCKs in chromosome 3. The hypothetical *M* region is highlighted by a red box. (B) Genomic synteny of chromosome 3 from *P. persica and P. avium*. (C) Genomic synteny of chromosome 3 with a closed up region surrounding the M locus. The genomic sequences surrounding the *M* locus were mostly conserved in two distantly related *Prunus* subgenera, *Amygdalus* and *Cerasus*. (D) Gene models and (E) expression levels provided as RPKM values for the *M* locus and surrounding regions in 'Brooks' pollen grains. The *MGST* gene is indicated in red. Four genes substantially expressed in pollen grains (RPKM >10) were annotated according to the gene models.

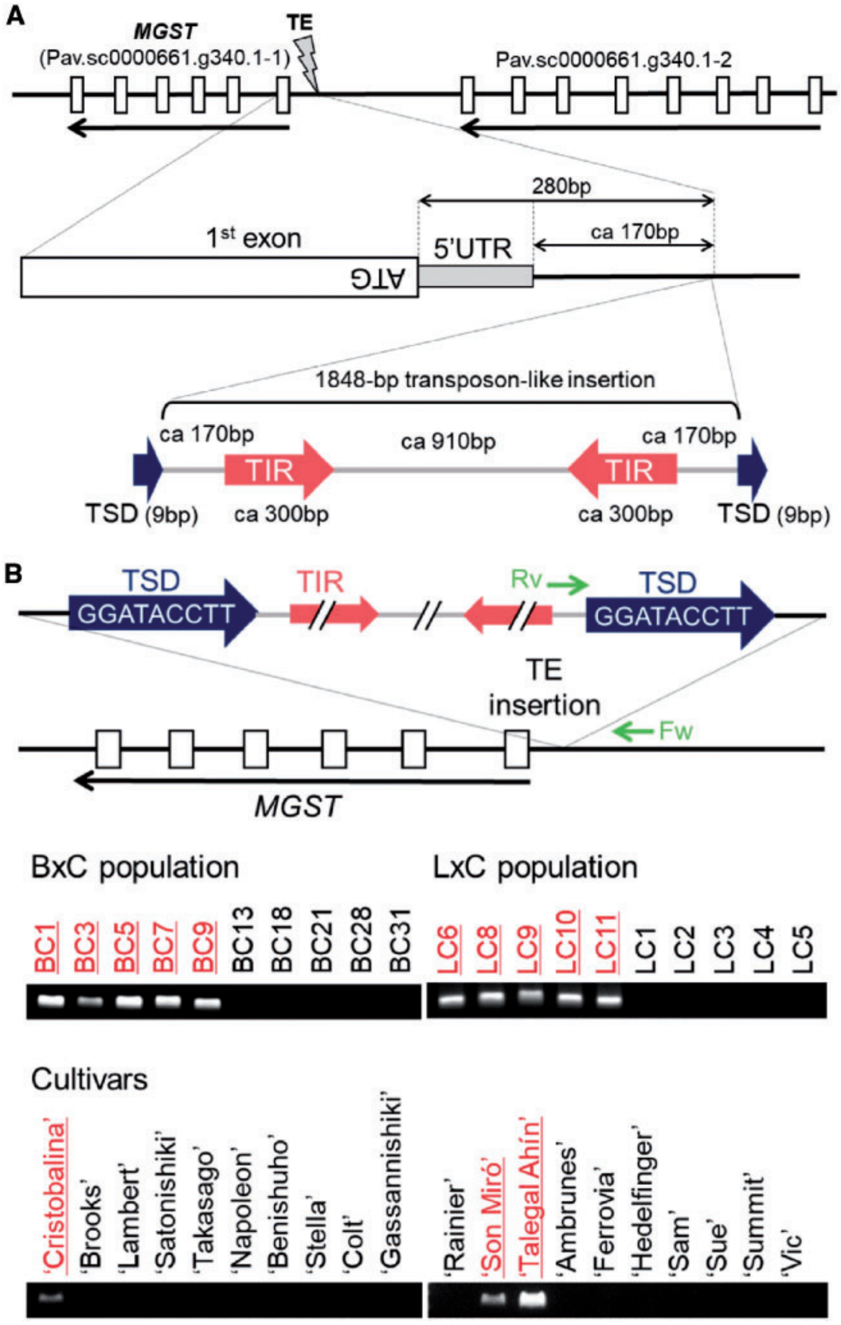

Fig. 4 Characterization of the mutated *MGST* in 'Cristobalina'. (A) Structure of the mutated *MGST* (*m* allele) in 'Cristobalina'. A 1,848 bp transposable element (TE)-like sequence is inserted 280 bp upstream of the *MGST* start codon. The TE-like sequence includes 9 bp target site duplication (TSD) sequences and approximately 300 bp terminal inverted repeats (TIRs), but no transposase gene sequence, indicating that the TE-like insertion is a non-autonomous DNA transposon. (B) Detection of the TE-inserted *MGST* genes in the B×C/L×C populations and various sweet cherry cultivars. A PCR was conducted with primer sets designed to amplify sequences within and flanking the TE (Fw and Rv) indicted by green arrows. The amplicons for the TE-inserted region were detected specifically in SC individuals and cultivars with the *m* allele (underlined in red). Although 'Stella' is SC, the causal mutation for SC is located in the *S* locus (Ushijima et al. 2004).

(evolutionary rate *dN*/*dS* >>1) using PAML (Yang 1998) identified a few sites that were under significant positive selection specific to the branch with *Prunus MGST* (**Fig. 6B**, *P* = 0.0008 against the null hypothesis with *dN*/*dS* = 1, posterior probability = 0.991 and 0.965 for Thr92 and Asp96, respectively, in a Bayes Empirical Bayes test). In contrast, significant positive selection was not detected in the branch with the *Fragaria MGST* ortholog in subclade I (mrna04224.1-v1.0-hybrid). These observations implied that the *Prunus MGST* gene may have undergone a neofunctionalization so the resulting protein functioned differently from the proteins encoded by orthologous *GST* genes in other species.

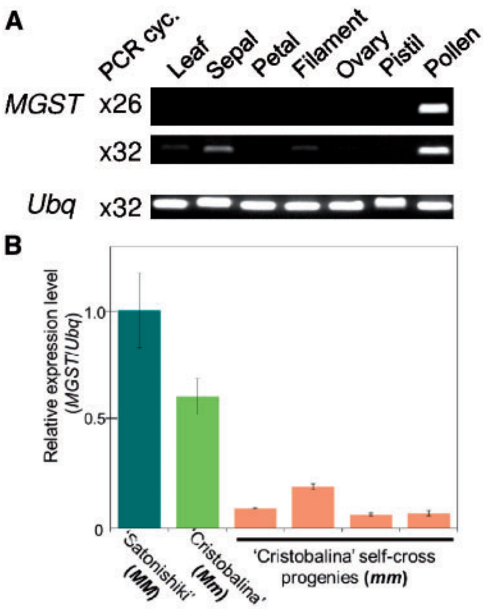

Fig. 5 Expression patterns of the *MGST* gene. (A) The *MGST* expression levels in the leaves and various flower organs of sweet cherry 'Cristobalina' were assessed by a reverse transcription–PCR (RT–PCR). Pollen-enriched *MGST* expression was observed, but the gene was also expressed in the other analyzed tissues. (B) A quantitative RT-PCR was completed to compare the expression levels of the intact (*M*) and mutated (*m*) *MGST* genes. Expression levels were determined using 'Satonishiki' (*MM*), Cristobalina (*Mm*) and four SC individuals with *mm* genotypes (generated by the selfing of 'Cristobalina'). The *MGST* expression levels were approximately 5- to 10-fold lower in *mm* individuals than in 'Satonishiki' (*MM*). The expression level in 'Cristobalina' (*Mm*) was approximately the mean of the expression levels in the *MM* and *mm* genotypes. The *MGST* expression level in 'Satonishiki' was set to 1 for the qRT-PCR analysis. A *Ubiquitin*-like gene [similar to a peach *Ubiquitin* gene (Prupe.4G204200.1)] was used for standardizing the RNA concentrations.

The expression patterns of the orthologs of the *Prunus MGST* gene revealed distinct features among the subclade I genes. Published pollen transcriptomic data for *Prunus*, *Pyrus*, *Fragaria* and *Petunia* species with the S-RNase-based GSI system were obtained from short-read databases and mapped to the orthologous/paralogous *MGST* sequences in the reference genomes to determine RPKM values. The expression levels of the *MGST* orthologs in pollen grains are provided in **Fig. 6C**. The *MGST* genes of *P. avium* and *P. mume*, which are two distantly related *Prunus* species, and the *Fragaria MGST* ortholog in subclade I were highly expressed at similar levels (RPKM >50). In contrast, the *MGST* orthologs in subclade II were expressed at much lower levels (**Fig. 6C**). Similarly, the *Petunia MGST* ortholog expression level was considerably lower than the subclade I gene expression levels.

Discussion

In this study, we confirmed the power of a genome-wide subtraction through cataloging of subsequences (kmers) generated

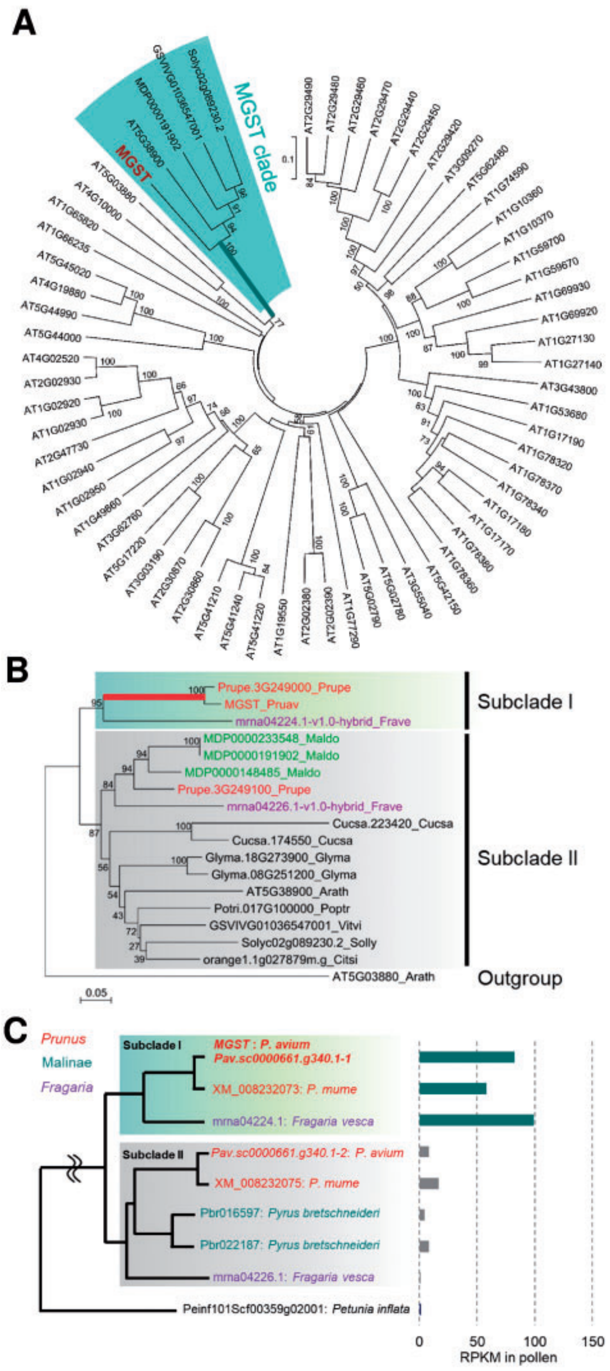

Fig. 6 Evolution of the *MGST* orthologs/paralogs. (A) Phylogenetic tree consisting of the *Prunus MGST* gene and its orthologs in apple (*Malus*×*domestica*), grape (*Vitis vinifera*) and tomato (*Solanum lycopersicum*) as well as the *Arabidopsis thaliana GST*-like genes. The *MGST* clade, for which the divergence (dark green branch) was statistically supported (bootstrap value = 100/100), was probably broadly monophyletic. (B) Evolutionary tree of the *MGST* orthologs in apple (Maldo), wild strawberry (Frave), sweet cherry (Pruav), peach (Prupe), cucumber (Cucusa), soybean (Glyma), poplar (Poptr), *A. thaliana* (Arath), sweet orange (Citsi), grape (Vitvi) and tomato (Solly). The AT5G03880 gene, which was located in the clade adjacent to that containing *A. thaliana GST* genes (A), served as the outgroup. Subclades I and II were statistically supported by bootstrap values (95/100 and 87/100, respectively). The *Prunus MGST* gene was nested in subclade I, which included only a *Fragaria* ortholog among the tested

from large-scale DNA sequencing data. It is noteworthy that this method enabled the efficient identification of a causal mutation in a wild-type gene with only a few individuals in segregating populations and existing cultivars. A similar methodology was used to confine the male-specific region of the Y chromosome and to identify the male determinant in dioecious persimmon (*Diospyros lotus*) without any reference genome sequence information (Akagi et al. 2014). Unlike the relatively large male-specific persimmon genomic region comprising approximately 1 Mbp, the *m* allele-specific region identified in this study is small (about 2 kbp). Thus, this investigation as well as a previous study (Akagi et al. 2014) can be considered as model cases for the application of kmer cataloging in diverse gene discovery studies involving non-model plant species.

The candidate SC-specific polymorphism detected at the *M* locus appeared to down-regulate 'Cristobalina' *MGST* expression to produce the *m* allele. Consistent with our results, the most likely candidate gene for an SC mutation in apricot was the *MGST* ortholog, encoding a disulfide bond A-like oxidoreductase (Supplementary Fig. S3). A putative apricot *M* locus gene was recently identified by combining a map-based cloning approach and mRNA-Seq analysis (Muñoz-Sanz et al. 2017). The *FaSt* (*Falling Stones*) miniature inverted repeat TE inserted in the third exon adversely affected this gene (Muñoz-Sanz et al. 2017). Although the candidate *M* genes in sweet cherry and apricot seemed to be identical, the disruptive mutations were different and appeared to be generated by independent TE insertions. The transposon-like sequences in the *MGST* promoter region in 'Cristobalina' have not been drastically amplified, although there seem to be many transposon relics in the cherry genome (Supplementary Fig. S4). The existence of this transposon-like sequence in the putative *MGST* promoter region is a possible causal mutation inducing SC in 'Cristobalina', 'Talegal Ahín' and 'Son Miró' (Cachi and Wünsch 2014).

The *MGST* gene encodes a GST, but the fact that this gene belongs to the thioredoxin (oxidoreductase) superfamily suggests that in a broader sense, the encoded protein may help stabilize substrate proteins (Montrichard et al. 2009). Thioredoxin-h in *Nicotiana alata* is reportedly localized in the extracellular matrix of the stylar transmitting tract and reduces S-RNases in vitro (Juárez-Díaz et al. 2006). The *Prunus* MGST proteins may have a similar affinity for S-RNases and potentially contribute to the proper folding of S-RNases when they enter pollen tubes. Alternatively, MGST may affect glutathione metabolism in the SI system. Glutathione metabolism and the resulting tryptophan synthesis are likely to be important for plant immune system activities, such as the hypersensitive response against hyphae (Hiruma et al. 2013), which may be associated with the SI reaction.

Fig. 6 Continued
orthologs. The red branch, which corresponds to the divergence of *Prunus MGST*, underwent specific positive selection. (C) Evolution of the *MGST* orthologs in plant species with the S-RNase-based GSI system, including three genera in the family Rosaceae and the genus *Petunia*, and their expression levels in pollen grains. Only the *Prunus MGST* and its *Fragaria* ortholog (mrna04224.1-v1.0-hybrid) were nested in subclade I. Both of these genes were highly expressed (RPKM >50). Genes from the genus *Prunus*, the subtribe Malinae and the genus *Fragaria* are highlighted in red, dark green and purple, respectively, in (B) and (C).

Regardless of the MGST function, it is likely that down-regulated *MGST* expression disrupts the SI reaction, leading to SC in pollen grains with the *m* allele. However, *m* confers only partial SC but not full SC. Cachi et al. (2014) observed that pollen tubes grew more slowly in self-pollen grains with *m* than in non-self-pollen grains with *M*. They also reported that the *M* allele was more common than the *m* allele in the F$_1$ individuals of *MM*×*Mm* semi-compatible crosses sharing one *S* haplotype. For example, there were fewer individuals with the S_3S_3Mm and S_3S_4Mm genotypes than expected from the 'Lambert' (S_3S_4MM)×'Cristobalina' (S_3S_6Mm) cross. Furthermore, there were fewer than expected individuals with the *Mm* genotype from the cross between the completely compatible 'Vic' (S_2S_4MM) and 'Cristobalina' (S_3S_6Mm) (A. Wünsch, unpublished results). These observations suggest that MGST may influence incompatibility reactions as well as the elongation of pollen tubes and/or fertilization in compatible crosses.

The *Fragaria MGST* ortholog (mrna04224.1-v1.0-hybrid) is nested in subclade I with *Prunus MGST*, and was highly expressed in pollen grains, similar to *Prunus MGST*. However, a phylogenetic analysis indicated that the subtribe Malinae has no subclade I *MGST* orthologs nested with *Prunus MGST*, suggesting that MGST is not an essential component of the S-RNase-based GSI system in the family Rosaceae. Furthermore, our evolutionary assessments indicated that ancestral *MGST* was potentially under positive selection, which is specific to the origin of the *Prunus MGST* gene. In other words, the *Prunus* MGST might have undergone a neofunctionalization as part of an adaptive evolution. This implies that a more thorough characterization of the MGST function may help to elucidate the molecular basis of the *Prunus*-specific GSI system.

Materials and Methods

Plant materials and phenotyping of self-(in)compatibility

Several individuals from F$_1$ populations derived from 'Cristobalina' sweet cherry (*P. avium*) were used for gDNA-Seq and mRNA-Seq analyses. Specifically, 18 SC and 25 SI individuals from two F$_1$ populations from the 'Brooks' (*MM*)×'Cristobalina' (*Mm*) (B×C) (Cachi and Wünsch 2011) and 'Lambert' (*MM*)×'Cristobalina' (*Mm*) (L×C) crosses were used. Except for one individual (B×C-0) the SC/SI phenotype of BXC family was previously (Cachi and Wünsch 2011). B×C SC/SI phenotyping was carried out by fruit set assay and pollen tube growth test after self-pollination (Cachi and Wünsch 2011). The SC/SI phenotype of B×C0 was determined only by the pollen tube growth test. L×C was phenotyped for SC/SI by fruit set assay and pollen tube growth test after self-pollination according to the same protocol described in Cachi and Wünsch (2011). Another 17 sweet cherry cultivars with a homozygous *M* (*MM*) genotype ('Ambrunes', 'Benishuho', 'Brooks', 'Colt', 'Ferrovia', 'Gassannishiki', 'Hedelfingen', 'Lambert', 'Napoleon', 'Rainier', 'Sam', 'Satonishiki', 'Stella', 'Sue', 'Summit', 'Takasago' and 'Vic') and one SC cultivar, 'Cristobalina' (*Mm*; Wünsch and Hormaza 2004), underwent gDNA-Seq and mRNA-Seq analyses. Two additional SC cultivars exhibiting SC similar to that of 'Cristobalina', 'Talegal Ahín' and 'Son Miró' (Cachi and Wünsch 2014) were used for genotyping the *M* locus.

Four selected SC accessions from the selfing of 'Cristobalina' (C×C), which were homozygous for the *M* locus (*mm*), were used for the qRT-PCR analysis of candidate genes. The plant materials and their uses are summarized in Supplementary Table S1. All plant materials were grown in the orchards of CITA de Aragon in Zaragoza, Spain, with the exception of eight cultivars ('Benishuho', 'Colt', 'Gassannishiki', 'Napoleon', 'Rainier', 'Satonishiki', 'Stella' and 'Takasago') that were grown in the experimental orchard of Kyoto University in Kyoto, Japan.

Preparation of gDNA-Seq libraries

Genomic DNA was extracted from young leaves using the cetyltrimethylammonium bromide (CTAB) method. Approximately 1.5 µg of genomic DNA was fragmented using NEBNext dsDNA Fragmentase (New England BioLabs; NEB) for 20 min at 37°C and then cleaned using Agencourt AMPure XP (Beckman Coulter Genomics) for a subsequent size selection as described by Akagi et al. (2014). The DNA fragments underwent an end repair using the End Repair Module Enzyme Mix (NEB), with A-base overhangs added with the Klenow fragment (NEB). The end repair and A-base addition were completed in a 50 µl reaction volume, after which the DNA fragments were cleaned using Agencourt AMPure XP and eluted with distilled water. Barcoded NEXTflex adaptors (Bioo Scientific) were ligated to the eluted fragments at room temperature using NEB Quick Ligase (NEB) following the manufacturer's instructions. To remove self-ligated adaptor dimers, libraries were size-selected using Agencourt AMPure XP and eluted from the beads in 20 µl of distilled water. Adaptor-ligated DNA libraries were enriched by a PCR amplification using PrimeStar Max DNA polymerase (TAKARA), with the following program: 98°C for 30 s; 8–12 cycles at 98°C for 10 s, 65°C for 30 s and 72°C for 30 s; 72°C for 5 min. The enriched libraries were purified with Agencourt AMPure XP, and the quality and quantity were assessed using the Agilent BioAnalyzer (Agilent Technologies) and Qubit fluorometer (Invitrogen).

Preparation of the mRNA-Seq libraries

Total RNA was extracted from mature anthers (containing mature pollen grains) using the PureLink Plant RNA Reagent (Thermo Fisher Scientific). Approximately 10 µg of total RNA was used to construct an mRNA library. The mRNA was purified using the Dynabeads mRNA Purification Kit (Life Technologies). Next, cDNA was synthesized with a random hexamer primer and Superscript III reverse transcriptase (Life Technologies). Following a heat inactivation at 65°C for 2 min, second-strand cDNA was synthesized by nick translation using the second-strand buffer (200 mM Tris–HCl, pH 7.0, 22 mM $MgCl_2$ and 425 mM KCl), DNA polymerase I (NEB) and RNaseH (NEB) with an incubation at 16°C for 2.5 h. Double-stranded cDNA was purified using Agencourt AMPure XP with a 1.8:1 (v/v) AMPure:reaction volume ratio. The cDNA was eluted from the beads in 10 µl of distilled water. The resulting double-stranded cDNA was fragmented and used for constructing a library as described for the genomic library. Ten or twelve cycles of PCR enrichment were completed using the same temperatures and times described above.

Illumina sequencing and processing

All libraries were sequenced using the Illumina HiSeq 2500/4000 systems (100 bp or 150 bp paired-end reads; Supplementary Table S1) at the Vincent J. Coates Genomics Sequencing Laboratory at UC Berkeley. The raw sequencing reads were processed using custom Python scripts developed in the Comai laboratory and available online (http://comailab.genomecenter.ucdavis.edu/index.php/Barcoded_data_preparation_tools). Briefly, reads were split based on index information and trimmed based on quality (average Phred sequence quality >20 over a 5 bp sliding window) and to remove contaminating adaptor sequences. A read-length cut-off of 35 bp was applied to both DNA reads.

Cataloging of SC-specific subsequences from gDNA-Seq reads

The method used to identify SCKs is summarized in **Fig. 1A**. The trimmed reads were concatenated to derive SC and SI pools for the B×C and L×C populations as well as the selected cultivars. We cataloged 35 bp subsequences starting with an 'A' nucleotide in the SC and SI pools using custom Python scripts (https://github.com/Comai-Lab/kmer-extract-by-trigger-site) based on a previous study (Akagi et al. 2014). We compared the kmer catalogs of the SC and SI pools and extracted the completely SC-specific kmers ('0' counts in the SI pool) (coverage ≥10; estimated coverage for a haploid = 32) to cover the SC-specific polymorphisms comprehensively. Next, 35 bp subsequences starting with an 'A' nucleotide were cataloged in the SI cultivar pool to filter the SC-specific kmers further. These two steps involving the subtraction of SI kmer pools resulted in the extraction of 661 SC-specific kmers. The paired-end Illumina reads including these SC-specific kmers were assembled using the CAP3 assembler to derive the genomic contigs covering the candidate polymorphisms responsible for the SC phenotype and the surrounding genomic regions.

Linkage tests of the candidate genomic contigs

We conducted a two-step linkage test involving 235 candidate contigs. In the first step, Illumina reads from the individuals of the B×C and L×C populations and 17 *MM* cultivars were mapped to the 235 contigs with the Burrows–Wheeler Aligner (BWA) version 0.7.7. Default parameters (http://bio-bwa.sourceforge.net/) or the –n 0 parameter were applied, which allowed up to approximately 5% and 0% nucleotide mismatches, respectively. The single nucleotide polymorphisms or indels in a contig were integrated to call genotypes for each individual as previously described (Akagi et al. 2014). We then assessed the co-segregation between genotypes and SC/SI phenotypes. In the second step, Illumina reads of the 18 SC progeny were divided into four groups (three B×C populations and one L×C population) to check the commonality of the SCKs among SC progeny. The SCKs existing in more than two of the four groups and the associated contigs were selected.

Identification of the expressed genes surrounding the candidate genomic contigs or located in the *M* locus

The SCK-derived genomic contigs, which were filtered by the linkage/recombination test, were subjected to blastn analysis (e-value <1e-10) using the sweet cherry [*Prunus avium* Whole Genome Assembly v1.0 & Annotation v1 (v1.0.a1) (Shirasawa et al. 2017)] and peach [*Prunus persica* Whole Genome Assembly v2.0 & Annotation v2.1 (v2.0.a1) (Verde et al. 2013)] genomes as references to identify their locations in the *Prunus* genomes. Next, the pollen mRNA-Seq reads for 12 SC and 15 SI individuals of the B×C/L×C populations and for 18 sweet cherry cultivars were mapped to the coding sequences of cherry genes (*n* = 43,673) and one peach candidate gene, which were annotated in the sweet cherry and peach reference genomes. The mapping was completed with the default parameters of BWA version 0.7.7 to calculate the expression level of each gene in pollen grains. The RPKM value of seven cherry genes and one peach candidate gene expressed in pollen grains from 'Cristobalina' F_1 progeny and the RPKM value of *M*-linked genes expressed in 'Brooks' pollen grains are presented in **Table 1** and **Fig. 3E**, respectively. The substantially expressed coding sequences that overlapped or were surrounded by the SCK-derived genomic contigs were considered as candidate *M* genes.

Sequence characterization and genotyping of *MGST*

The *MGST* (Pav.sc0000661.g340.1-1) gene and its 5′-flanking regions, which were anchored by the SCK-derived polymorphic contigs, were amplified by PCR using primers (Supplementary Table S4) designed based on the *P. avium* genomic sequences (Shirasawa et al. 2017). The amplicons for the 5′ promoter region of 'Cristobalina' and the SC individuals from the B×C/L×C populations were larger than those of the SI individuals (data not shown). Moreover, the Sanger sequencing of the amplicons from the SC individuals revealed the existence of a TE-like insertion. The 'Cristobalina' genomic reads were mapped to the TE-like insertion and its surrounding sequences with BWA version 0.7.7, allowing no mismatches to confirm the accuracy of the 1,848 bp TE-like sequence. To call the genotypes of this TE-like insertion across the segregating populations and the sweet cherry cultivars, PCR analyses were conducted using primers designed to amplify fragments within or flanking the TE-like insertion (**Fig. 4B;** Supplementary Table S4). The PCR was completed with the following program: 94°C for 1 min; 35 cycles of 94°C for 30 s, 53°C for 30 s and 72°C for 30 s; 72°C for 5 min.

Analyses of *MGST* evolution

The *A. thaliana* genes encoding GST or annotated with GST-associated GO/PATHWAY/PFAM/PANTHER/KEGG terms were extracted from the TAIR10 database. A blastp search using *P. avium MGST* as the query (e-value <1e-18) was used to extract *Prunus MGST* orthologs from the genomes of apple (*Malus*×*domestica*), wild strawberry (*Fragaria vesca*), cucumber (*Cucumis sativus*), soybean (*Glycine max*), popular (*Populus trichocarpa*), sweet orange (*Citrus sinensis*) and grape (*Vitis vinifera*) available in the Phytozome v12 database (https://phytozome.jgi.doe.gov). Furthermore, the *MGST* orthologs in Chinese white pear (*Pyrus*×*bretschneideri*; Wu et al. 2013), Japanese apricot (*Prunus mume*; Zhang et al. 2012) and *Petunia inflata* (Bombarely et al. 2016) were extracted from their draft genome sequences with blastx using *P. avium MGST* as the query (e-value <1e-20) to map the pollen transcriptomic data (**Fig. 6C**). The protein sequences encoded by these genes were aligned using MAFFT version 7 (https://mafft.cbrc.jp/alignment/server/) (Katoh et al. 2017) and manually pruned with SeaView version 2.4 (http://doua.prabi.fr/software/seaview) (Gouy et al. 2010). Phylogenetic trees were constructed based on the Neighbor–Joining (**Fig. 6A**) and maximum-likelihood (**Fig. 6B**) methods of the MEGA v6 program (http://www.megasoftware.net/reltime) (Tamura et al. 2013). The WAG (Whelan and Goldman 2001) model with invariant sites was used as part of the maximum-likelihood method, with the nearest neighbor interchange applied as the tree-searching heuristic. All sites, including those with missing and gap data, were used to construct the phylogenetic tree. Meanwhile, the Neighbor–Joining method involved the Poisson matrix with gamma-distributed rates (alpha parameter 3) as well as pairwise deletions for missing data. To detect positive selection, the nucleotide sequences of the analyzed genes and an outgroup ortholog (AT5G03880) were subjected to an in-frame alignment using the Pal2Nal server (Suyama et al. 2006). We then detected codon-based site-branch-specific positive selection using PAML. The significance of the positive selection on the foreground branches was evaluated using the likelihood ratio test, with a null hypothesis of $dN/dS = 1$. Site-specific positive selection was assessed by Bayes Empirical Bayes analysis.

We mapped the published Illumina mRNA-Seq reads for pollen grains from *P. avium* (DRX001700), *P. mume* (DRX001701), *F. vesca* (SRX426507), *P.* ×*bretschneideri* (SRX1356151) and *P. inflata* (SRX515117) to the *MGST* ortholog/paralog sequences. The read counts were converted to RPKM values to compare expression levels.

Accession numbers

All sequence data generated during this study have been deposited in appropriate DDBJ databases. Illumina gDNA-Seq and mRNA-Seq reads were submitted to the Short Read Archives database (BioProject ID PRJDB6734), while the MGST genic sequences and the TE-like insertion sequences from *P. avium* 'Cristobalina' were submitted to the GenBank database (IDs LC371238 and LC371380).

Supplementary Data

Supplementary data are available at PCP online.

Funding

This work was supported by the Japan Society for the Promotion of Science [Grants-in-Aid for Scientific Research (A) Nos. 20248004, 24248007 and 15H02431 to R.T.] and by Spanish Government 'Ministerio de Economia y Competitividad' (MINECO) INIA and FEDER funds research proyects [RTA2015-00027-00-00 and RFP2015-00015-00-00].

Acknowledgments

We thank Drs. L. Comai, I. Henry and M. Lieberman (University of California, Davis) for bioinformatics support, and M. Watanabe for assisting with the preliminary experiments.

Disclosures

The authors have no conflicts of interest to declare.

References

Aguiar, B., Vieira, J., Cunha, A.E., Fonseca, N.A., Iezzoni, A. and Van Nocker, S. (2015) Convergent evolution at the gametophytic self-incompatibility system in *Malus* and *Prunus*. *PLoS One* 10: e0126138.

Akagi, T., Henry, I.M., Morimoto, T. and Tao, R. (2016) Insights into the *Prunus*-specific S-RNase-based self-incompatibility system from a genome-wide analysis of the evolutionary radiation of *S* locus-related F-box genes. *Plant Cell Physiol.* 57: 1281–1294.

Akagi, T., Henry, I.M., Tao, R. and Comai, L. (2014) A Y-chromosome-encoded small RNA acts as a sex determinant in persimmons. *Science* 346: 646–650.

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

Bombarely, A., Moser, M., Amrad, A., Bao, M., Bapaume, L., Barry, C.S., et al. (2016) Insight into the evolution of the Solanaceae from the parental genomes of *Petunia hybrida*. *Nat. Plants* 2: 16074.

Cachi, A.M., Hedhly, A., Hormza, J.I. and Wünsch, A. (2014) Pollen tube growth in the self-compatible sweet cherry genotype, 'Cristobalina', is slowed down after self-pollination. *Ann. Appl. Biol*. 164: 73–84.

Cachi, A.M. and Wünsch, A. (2011) Characterization and mapping of non-S gametophytic self-compatibility in sweet cherry (*Prunus avium* L.). *J. Exp. Bot.* 62: 1847–1856.

Cachi, A.M. and Wünsch, A. (2014) Characterization of self-compatibility in sweet cherry varieties by crossing experiments and molecular genetic analysis. *Tree Genet. Genomes* 10: 1205–1212.

De Nettancourt, D. (2001) Incompatibility and Incongruity in Wild and Cultivated Plants. Springer, Berlin.

Entani, T., Kubo, K., Isogai, S., Fukao, Y., Shirakawa, M., Isogai, A., et al. (2014) Ubiquitin–proteasome-mediated degradation of S-RNase in a solanaceous cross-compatibility reaction. *Plant J.* 78: 1014–1021.

Goldraij, A., Kondo, K., Lee, C.B., Hancock, C.N., Sivaguru, M., Vazquez-Santana, S., et al. (2006) Compartmentalization of S-RNase and HT-B degradation in self-incompatible *Nicotiana*. *Nature* 439: 805–810.

Gouy, M., Guindon, S. and Gascuel, O. (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. *Mol. Biol. Evol.* 27: 221–224.

Hauck, N.R., Yaman, H., Tao, R. and Iezzoni, A.F. (2006) Accumulation of nonfunctional S-haplotypes results in the break-down of gametophytic self-incompatibility in tetraploid *Prunus*. *Genetics* 172: 1191–1198.

Hiruma, K., Fukunaga, S., Bednarek, P., Piślewska-Bednarek, M., Watanabe, S., Narusaka, Y., et al. (2013) Glutathione and tryptophan metabolism are required for Arabidopsis immunity during the hypersensitive response to hemibiotrophs. *Proc. Natl. Acad. Sci. USA* 110: 9589–9594.

Huang, X. and Madan, A. (1999) CAP3: a DNA sequence assembly program. *Genome Res.* 9: 868–877.

Igic, B. and Kohn, J.R. (2006) The distribution of plant mating systems: study bias against obligately outcrossing species. *Evolution* 60: 1098–1103.

Juárez-Díaz, J.A., McClure, B., Vázquez-Santana, S., Guevara-García, A., León-Mejía, P., Márquez-Guzmán, J., et al. (2006) A novel thioredoxin h is secreted in *Nicotiana alata* and reduces S-RNase *in vitro*. *J. Biol. Chem.* 281: 3418–3424.

Kakui, H., Kato, M., Ushijima, K., Kitaguchi, M., Kato, S. and Sassa, H. (2011) Sequence divergence and loss-of-function phenotypes of *S locus F-box brothers* genes are consistent with non-self recognition by multiple pollen determinants in self-incompatibility of Japanese pear (*Pyrus pyrifolia*). *Plant J.* 68: 1028–1038.

Katoh, K., Rozewicki, J. and Yamada, K.D. (2017) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. *Brief. Bioinform*. doi.org/10.1093/bib/bbx108.

Kubo, K., Entani, T., Takara, A., Wang, N., Fields, A.M., Hua, A., et al. (2010) Collaborative non-self recognition system in S-RNase-based self-incompatibility. *Science* 330: 796–799.

Kubo, K., Paape, T., Hatakeyama, M., Entani, T., Takara, A., Kajihara, K., et al. (2015) Gene duplication and genetic exchange drive the evolution of S-RNase-based self-incompatibility in *Petunia*. *Nat. Plants* 1: 14005.

Matsumoto, D. and Tao, R. (2016) Distinct self-recognition in the *Prunus* S-RNase-based gametophytic self-incompatibility system. *Hort. J.* 85: 289–305.

McClure, B. (2009) Darwin's foundation for investigating self-incompatibility and the progress toward a physiological model for S-RNase-based SI. *J. Exp. Bot.* 60: 1069–1081.

McCubbin, A.G. and Kao, T.H. (2000) Molecular recognition and response in pollen and pistil interactions. *Annu. Rev. Cell Dev. Biol.* 16: 333–364.

Meng, X., Sun, P. and Kao, T.H. (2011) *S*-RNase-based self-incompatibility in *Petunia inflata*. *Ann. Bot.* 108: 637–646.

Minamikawa, M., Kakui, H., Wang, S., Kotoda, N., Kikuchi, S., Koba, T., et al. (2010) Apple *S* locus region represents a large cluster of related, polymorphic and pollen-specific F-box genes. *Plant Mol. Biol*. 74: 143–154.

Montrichard, F., Alkhalfioui, A., Yano, H., Vensel, W.H., Hurkm, W.J. and Buchanan, B.B. (2009) Thioredoxin targets in plants: the first 30 years. *J. Proteomics* 72: 452–474.

Morimoto, T., Akagi, T. and Tao, R. (2015) Evolutionary analysis of genes for S-RNase-based self-incompatibility reveals S locus duplications in the ancestral Rosaceae. *Hort. J.* 84: 233–242.

Muñoz-Sanz, J.V., Zuriaga, E., Badenes, M.L. and Romero, C. (2017) A disulfide bond A-like oxidoreductase is a strong candidate gene for self-incompatibility in apricot (*Prunus armeniaca*) pollen. *J. Exp. Bot.* 68: 5069–5078.

Sassa, H., Kakui, H. and Minamikawa, M. (2010) Pollen-expressed F-box gene family and mechanism of S-RNase-based gametophytic self-incompatibility (GSI) in Rosaceae. *Sex. Plant Reprod.* 23: 39–43.

Shirasawa, K., Isuzugawa, K., Ikenaga, M., Saito, Y., Yamamoto, T., Hirakawa, H., et al. (2017) The genome sequence of sweet cherry (*Prunus avium*) for use in genomics-assisted breeding. *DNA Res.* 24: 499–508.

Sonneveld, T., Tobutt, K.R., Vaughan, S.P. and Robbins, T.P. (2005) Loss of pollen-*S* function in two self-compatible selections of *Prunus avium* is associated with deletion/mutation of an *S* haplotype-specific F-box gene. *Plant Cell* 17: 37–51.

Suyama, M., Torrents, D. and Bork, P. (2006) PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acid Res.* 34: W609–W612.

Tamura, K., Glen, S., Daniel, P., Alan, F. and Sudhir, K. (2013) MEGA6: molecular evolutionary genetics analysis Version 6.0. *Mol. Biol. Evol.* 30: 2725–2729.

Tao, R. and Iezzoni, A.F. (2010) The *S*-RNase-based gametophytic self-incompatibility system in *Prunus* exhibits distinct genetic and molecular features. *Scientia Hort*. 124: 423–433.

Ushijima, K., Sassa, H., Dandekar, A.M., Gradziel, T.M., Tao, R. and Hirano, H. (2003) Structural and transcriptional analysis of the self-incompatibility locus of almond: identification of a pollen-expressed F-box gene with haplotype-specific polymorphism. *Plant Cell* 15: 771–781.

Ushijima, K., Yamane, H., Watari, A., Kakehi, E., Ikeda, K., Hauck, N.R., et al. (2004) The S haplotype-specific F-box protein gene, *SFB*, is defective in self-compatible haplotypes of *Prunus avium* and *P. mume*. *Plant J.* 39: 573–586.

Verde, I., Abbott, A.G., Scalabrin, S., Jung, S., Shu, S., Marroni, F., et al. (2013) The high-quality draft genome of peach (*Prunus persica*) identifies unique patterns of genetic diversity, domestication and genome evolution. *Nat. Genet.* 45: 487–494.

Whelan, S. and Goldman, N. (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol. Biol. Evol*. 18: 691–699.

Wu, J., Wang, Z., Shi, Z., Zhang, S., Ming, R., Zhu, S., et al. (2013) The genome of the pear (*Pyrus bretschneideri* Rehd.). *Genome Res.* 23: 396–408.

Wünsch, A. and Hormaza, J.I. (2004) Genetic and molecular analysis in 'Cristobalina' sweet cherry, a spontaneous self-compatible mutant. *Sex. Plant Reprod.* 17: 203–210.

Xu, G., Ma, H., Nei, M. and Kong, H. (2009) Evolution of F-box genes in plants: different modes of sequence divergence and their relationships with functional diversification. *Proc. Natl. Acad. Sci. USA* 106: 835–840.

Yang, Z. (1998) Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15: 568–573.

Zhang, Q., Chen, W., Sun, L., Zhao, F., Huang, B. and Yang, W. (2012) The genome of *Prunus mume*. *Nat. Commun*. 3: 1318.

Zhao, L., Huang, J., Zhao, Z., Li, Q., Sims, T.L. and Xue, Y. (2010) The Skp1-like protein SSK1 is required for cross-pollen compatibility in S-RNase-based self-incompatibility. *Plant J.* 62: 52–63.

Zuriaga, E., Molina, L., Badenes, M.L. and Romero, C. (2012) Physical mapping of a pollen modifier locus controlling self-incompatibility in apricot and synteny analysis within the Rosaceae. *Plant Mol. Biol.* 79: 229–242.

Zuriaga, E., Munoz-Sanz, J.V., Molina, L., Gisbert, A.D., Badenes, M.L. and Romero, C. (2013) An S-locus independent pollen factor confers self-compatibility in 'Katy' apricot. *PLoS One* 8: e53947.