The Plant Cell, Vol. 26: 2873–2888, July 2014, www.plantcell.org © 2014 American Society of Plant Biologists. All rights reserved.

Transcriptome Analysis Reveals the Same 17 *S-Locus F-Box* Genes in Two Haplotypes of the Self-Incompatibility Locus of *Petunia inflata*[W]

Justin S. Williams,[a] Joshua P. Der,[b] Claude W. dePamphilis,[b,c] and Teh-hui Kao[a,c,1]

[a] Department of Biochemistry and Molecular Biology, Pennsylvania State University, University Park, Pennsylvania 16802
[b] Department of Biology, Pennsylvania State University, University Park, Pennsylvania 16802
[c] Intercollege Graduate Degree Program in Plant Biology, Pennsylvania State University, University Park, Pennsylvania 16802

ORCID IDs: 0000-0002-0195-5509 (J.S.W.); 0000-0001-9668-2525 (J.P.D.)

***Petunia* possesses self-incompatibility, by which pistils reject self-pollen but accept non-self-pollen for fertilization. Self-/non-self-recognition between pollen and pistil is regulated by the pistil-specific *S-RNase* gene and by multiple pollen-specific *S-locus F-box* (*SLF*) genes. To date, 10 *SLF* genes have been identified by various methods, and seven have been shown to be involved in pollen specificity. For a given *S*-haplotype, each SLF interacts with a subset of its non-self S-RNases, and an as yet unknown number of SLFs are thought to collectively mediate ubiquitination and degradation of all non-self S-RNases to allow cross-compatible pollination. To identify a complete suite of *SLF* genes of *P. inflata*, we used a de novo RNA-seq approach to analyze the pollen transcriptomes of S_2-haplotype and S_3-haplotype, as well as the leaf transcriptome of the S_3S_3 genotype. We searched for genes that fit several criteria established from the properties of the known *SLF* genes and identified the same seven new *SLF* genes in S_2-haplotype and S_3-haplotype, suggesting that a total of 17 *SLF* genes constitute pollen specificity in each *S*-haplotype. This finding lays the foundation for understanding how multiple *SLF* genes evolved and the biochemical basis for differential interactions between SLF proteins and S-RNases.**

INTRODUCTION

Due to the accrual of recessive mutations, the consequences of inbreeding can be detrimental to many organisms, including plants. The sessile nature of plants makes it particularly important for them to possess mechanisms that prevent or minimize inbreeding. Many bisexual flowering plants have adopted an intraspecific reproductive strategy, termed self-incompatibility (SI), which allows pistils of a plant to reject pollen from the same plant or plants that are genetically related (self-pollen) and to only accept pollen from genetically distinct individuals for fertilization (non-self-pollen) (de Nettancourt, 2001). In the genus *Petunia* (Solanaceae), the highly polymorphic *S*-locus determines the self versus non-self discrimination between pollen and pistil, based on whether or not the *S*-haplotype of pollen is carried by the pistil. The *S*-locus houses the female determinant and male determinant genes (Iwano and Takayama, 2012; Wang and Kao, 2012). *S-RNase* encodes the female determinant (Lee et al., 1994; Murfett et al., 1994) and is expressed specifically in the transmitting cells of the pistil (Cornish et al., 1987; Anderson et al., 1989). Mature S-RNase is abundantly present in the transmitting tract of the upper third segment of the style, where it is taken up by both self-pollen and non-self-pollen tubes after they have penetrated the stigma (Luu et al., 2000; Goldraij et al., 2006). As the ribonuclease activity of S-RNase is necessary for S-RNase to inhibit self-pollen tube growth (Huang et al., 1994), an allelic variant of S-RNase produced by pistils carrying a specific *S*-haplotype most likely exerts its cytotoxic effect via degradation of RNAs in the cytosol of self pollen tubes (carrying a matching *S*-haplotype with the pistil).

Unlike the female determinant, the male determinant is encoded by multiple intronless genes, named *S-locus F-box1* (*SLF1*), *SLF2*, etc. To date, 10 *SLF* genes (*SLF1* to *SLF10*) have been identified in *Petunia* (McCubbin et al., 2000; Y. Wang et al., 2003, 2004; Hua et al., 2007; Kubo et al., 2010), and seven of them (all except *SLF7*, *SLF9*, and *SLF10*) have been shown to be involved in pollen specificity by a transgenic functional assay (Sijacic et al., 2004; Kubo et al., 2010; Williams et al., 2014). However, the precise number of *SLF* genes required for pollen specificity of any *S*-haplotype is unknown. A conventional F-box protein is a component of a class of E3 ubiquitin ligase, named the SCF complex, which also contains Skp1, Cullin1, and Rbx1 (Bai et al., 1996; Stone and Callis, 2007). SCF, in conjunction with E1 (ubiquitin-activating enzyme) and E2 (ubiquitin-conjugating enzyme), catalyzes transfer of polyubiquitin chains to protein substrates recognized by the specific F-box protein component of the complex, marking them for degradation by the 26S proteasome (Vierstra, 2009). SLF proteins were thought to function as conventional F-box proteins to mediate ubiquitination and degradation of S-RNases (Qiao et al., 2004; Hua and Kao, 2006; Hua and Kao, 2008; Kubo et al., 2010). Indeed, three SLF proteins produced by S_2 pollen of *Petunia inflata*, S_2-SLF1, S_2-SLF4, and S_2-SLFx (not among the previously identified 10 SLF proteins), have recently been shown by coimmunoprecipitation to each be in an SCF complex, which contains a pollen-specific cullin

[1] Address correspondence to txk3@psu.edu.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Teh-hui Kao (txk3@psu.edu).
[W] The online version of this article contains Web-only data.
www.plantcell.org/cgi/doi/10.1105/tpc.114.126920

(named CUL1-P), a pollen-specific Skp1-like protein (named SSK1), and an Rbx1 protein (named RBX1) (Li et al., 2014). For the seven SLF proteins whose function in pollen specificity has been established, each is capable of interacting with only a subset of its non-self S-RNases examined, and none of them can interact with their respective self-S-RNase (Sijacic et al., 2004; Kubo et al., 2010; Williams et al., 2014). According to the collaborative non-self-recognition model, the function of SLF proteins produced by pollen of a given *S*-haplotype is to collectively counter the cytotoxic effect of all the non-self S-RNases that the pollen tube may encounter during cross-pollinations (Kubo et al., 2010). That is, the SCF^{SLF} complexes collaboratively detoxify all non-self S-RNases by mediating their ubiquitination and degradation, and the interaction specificity for a given SCF^{SLF} complex is determined by the type of SLF in the complex.

The finding that multiple SLF proteins are required for pollen specificity raises the questions: (1) How many SLF proteins constitute the pollen specificity determinant? (2) Do different *S*-haplotypes employ the same number of SLF proteins for their pollen specificity? (3) What is the biochemical basis for the differential interactions between SLF proteins and S-RNases? (4) What mechanisms are responsible for the generation of multiple *SLF* genes? Answering these questions will contribute to a better understanding of the evolution, operation, and maintenance of the SI system in *Petunia* (and likely in other species that possess the same type of SI system).

Several methods have been used to identify *SLF* genes in *Petunia*. The first two *SLF* genes identified, named *A113* and *A134* (renamed *SLF9* and *SLF10*), were identified using RNA differential display to search for pollen-expressed genes of *P. inflata* that showed *S*-haplotype-specific sequence polymorphism (McCubbin et al., 2000; Wang et al., 2003). The first *SLF* gene (*SLF1*; formerly named *PiSLF*$_2$) of *P. inflata* shown to be involved in pollen specificity was identified by first screening a BAC library of the S_2S_2 genotype using the S_2-*RNase* gene as probe, followed by chromosome walking to extend the region covered by the BAC clones to 328 kb and subsequent sequencing of this contig (Sijacic et al., 2004; Y. Wang et al., 2004). Four more *SLF* genes of *P. inflata* were identified by screening pollen cDNA libraries using S_2-*SLF1* (without the coding sequence for the F-box domain) as probe (Hua et al., 2007). These four genes were initially named *SLFLa*, *SLFLb*, *SLFLc*, and *SLFLd*, as they were considered to be *SLF*-like genes and not involved in pollen specificity (Hua et al., 2007). Finally, primers were designed based on the sequences of all the above-mentioned *SLF*/*SLFL* genes for 5′, 3′ RACE (rapid amplification of cDNA ends) to identify their orthologs and additional *SLF* genes in *Petunia hybrida* and *Petunia axillaris* (Kubo et al., 2010). Three additional *SLF* genes, named *SLF4*, *SLF5*, and *SLF6*, were identified, and *SLFLa*, *SLFLb*, *SLFLc*, and *SLFLd* have been renamed *SLF7*, *SLF8*, *SLF2*, and *SLF3*, respectively (Kubo et al., 2010; Williams et al., 2014).

To determine all the *SLF* genes involved in pollen specificity, a more systematic approach is required. *P. inflata* currently lacks a publicly available genome sequence; however, advances in high-throughput sequencing and de novo transcriptome assembly methods have made it possible to reconstruct the complete set of expressed gene sequences in non-model organisms. We sequenced the transcriptomes of two pollen haplotypes, S_2 and S_3, from S_2S_2 and S_3S_3 plants, respectively, and of leaf from S_3S_3 plants as a means of discovering a complete suite of *SLF* genes in *P. inflata*. The S_2S_2 and S_3S_3 plants homozygous at the *S*-locus were obtained from bud-selfing S_2S_3 plants to circumvent SI (Ai et al., 1990).

We sequenced 101-bp paired-end reads from a strand-specific RNA-seq library using the Illumina HiSequation 2000 platform and assembled separate de novo transcriptomes for each pollen haplotype and S_3S_3 leaf using the Trinity RNA-seq pipeline (Grabherr et al., 2011). Examining pollen transcriptomes of two *S*-haplotypes allowed us to assess whether any *SLF* gene identified shows allelic sequence polymorphism and whether pollen of different *S*-haplotypes contain the same suite of *SLF* genes. Sequencing the leaf transcriptome allowed us to identify and focus our analysis on genes that are exclusively expressed in pollen but not in leaves. More than 35 Gbp of total sequence data were generated and assembled with Trinity, producing three assemblies with 45,500, 41,409, and 51,961 unique coding sequences (unigenes) in S_2-pollen, S_3-pollen, and S_3S_3-leaf transcriptomes, respectively. We searched our assemblies for ultraconserved orthologs (UCOs; Fulton et al., 2002; Wu et al., 2006; Kozik et al., 2008) to estimate the extent of gene detection and used unigene accumulation curves (Der et al., 2011) to determine that the depth of sequencing was sufficient to detect nearly all of the expressed genes in the transcriptome. These results indicate that there was ~85% gene coverage in the S_2-pollen and S_3-pollen assembles, and 96% gene coverage in the S_3S_3-leaf assemblies, and that ~99% of the genes were discovered using <25% of the reads available.

We identified the same seven novel *SLF* genes expressed in pollen of the S_2-haplotype and S_3-haplotype, bringing the total number of *SLF* genes in these two *S*-haplotypes to 17. To confirm the presence of these genes, we performed PCR using genomic DNA as template and primers designed for each candidate *SLF* gene identified. These genes were also cloned from S_5S_5, $S_{6a}S_{6a}$, and $S_{12}S_{12}$ genetic backgrounds of *P. inflata* to provide additional evidence of allelic sequence diversity and to identify useful sequence polymorphism for downstream phylogenetic analyses. We further performed *S*-locus linkage analyses on four of the seven novel *SLF* genes, using primers specific to the S_2-allele and S_3-allele of each gene, and showed that they are all tightly linked to the *S*-locus.

RESULTS

Construction and Sequencing of RNA-Seq Libraries and Assembly of Sequencing Reads

All seven previously identified *Petunia SLF* genes (*SLF1* to *SLF6*, and *SLF8*) that have been shown to be involved in pollen specificity are exclusively expressed in pollen (Sijacic et al., 2004; Kubo et al., 2010; Williams et al., 2014). To identify additional *SLF* genes of the S_2-haplotype and S_3-haplotype of *P. inflata*, we separately isolated total RNA from pollen of S_2S_2 and S_3S_3 plants for transcriptome analysis. To increase the likelihood of complete gene discovery, we used two biological replicates for each pollen

S-haplotype. In addition, total RNA from leaf tissue of S_3S_3 plants was isolated for transcriptome analysis to identify those transcripts that are only present in the pollen transcriptomes. These five RNA samples were used to generate strand-specific RNA-seq libraries (Borodina et al., 2011) following the Illumina TruSeq v2 sample preparation protocol (see Methods). The workflow for subsequent transcriptome analysis and *SLF* gene discovery is shown in Supplemental Figure 1.

The five RNA-seq libraries were sequenced using the Illumina HiSequation 2000 platform, and the reads from each library were quality trimmed (see Methods). The processed reads from the biological replicates of each pollen *S*-haplotype were combined, resulting in 100.5 bp average read length and 13.4 Gbp total sequence data for S_2 pollen, 100.5 bp average read length and 15.1 Gbp total sequence data for S_3 pollen, and 100.5 bp average read length and 7.0 Gbp total sequence data for S_3S_3 leaf (Supplemental Table 1). These reads were assembled using Trinity software (Grabherr et al., 2011) to yield three separate assemblies: 45,500 unigenes with an average length of 798 bp for S_2 pollen, 41,409 unigenes with an average length of 749 bp for S_3 pollen, and 51,961 unigenes with an average length of 747 bp for S_3S_3 leaf (Table 1). Protein coding sequences were predicted using ESTScan (Iseli et al., 1999; Lottaz et al., 2003) and redundant sequences were filtered using Usearch (Edgar, 2010; see Methods) to produce the final S_2-pollen assembly containing 25,522 unigenes with an average length of 704 bp, S_3-pollen assembly containing 23,731 unigenes with an average length of 675 bp, and S_3S_3-leaf assembly containing 32,559 unigenes with an average length of 610 bp (Table 1). Because 5′- and 3′-untranslated region (UTR) sequences are valuable in primer design for PCRs, both the raw Trinity assemblies and ESTScan processed assemblies (referred to as filtered unigene sets) were maintained for downstream analysis and gene discovery (Supplemental Figure 1).

Table 1. Assembly Statistics

	Unprocessed Trinity Assembly	Processed Assembly
S_2-pollen assembly		
Total number of unigenes	45,500	25,522
Mean unigene length (bp)	798	704
Median unigene length (bp)	477	453
Smallest unigene length (bp)	201	102
Largest unigene length (bp)	13,448	12,606
Total assembly length (bp)	36,437,377	17,961,003
S_3-pollen assembly		
Total number of unigenes	41,409	23,731
Mean unigene length (bp)	749	675
Median unigene length (bp)	457	435
Smallest unigene length (bp)	201	102
Largest unigene length (bp)	9,200	8,700
Total assembly length (bp)	30,762,424	16,021,869
S_3S_3-leaf assembly		
Total number of unigenes	51,961	32,559
Mean unigene length (bp)	747	610
Median unigene length (bp)	538	462
Smallest unigene length (bp)	201	102
Largest unigene length (bp)	7,877	6,591
Total assembly length (bp)	38,834,278	19,855,188

Assessing Transcriptome Coverage and Unigene Discovery

Given that the main goal of this work was to use pollen transcriptomes to identify all candidate *SLF* genes involved in pollen specificity, it is important that the S_2-pollen and S_3-pollen transcriptomes used for gene discovery be as complete as possible. For species whose complete genome sequence is not available, such as *P. inflata*, determining the extent of transcriptome coverage can be difficult. Here, we used UCO detection rates to evaluate the quality of these data (Fulton et al., 2002; Wu et al., 2006; Kozik et al., 2008; Der et al., 2011). UCOs are defined as a set of putatively conserved single-copy genes expected to be present across a broad range of eukaryotes. The coding sequences of UCOs in *Arabidopsis thaliana* were used as tBLASTn queries into both the Trinity assemblies and the filtered unigene sets. We considered a UCO to have been discovered in an assembly if a high-scoring segment pair (HSP) was found in this assembly (defined by a deduced alignment length of at least 30 amino acids with an expectation value of at least 1e-10). Two distinct sets of UCOs were screened. The first set was the list of UCO genes highlighted by the Compositae genome project (http://compgenomics.ucdavis.edu/compositae_reference.php; Kozik et al., 2008) and the second set was the conserved othologous set II (COSII) from SolGenomics (http://solgenomics.net/documents/markers/cosii.xls; Fulton et al., 2002; Wu et al., 2006). Of the 357 genes on the CompGenomics UCO list, the percentages of genes discovered in the pollen assemblies were lower than those of the corresponding leaf assemblies; 68.6% were discovered in S_2 pollen, 66.7% in S_3 pollen, and 91.3% in S_3S_3 leaf. In the SolGenomics COSII gene set, we found that of the 220 genes on the list, 72.7% were discovered in S_2-pollen, 67.3% in S_3-pollen, and 96.4% in S_3S_3-leaf Trinity assemblies (Supplemental Table 2).

The lower percentages of gene discovery in the pollen assemblies than in the leaf assemblies could be due to the fact that the UCOs in both Compositae and Solanaceae COSII lists contain genes that are not expressed in pollen. To address this possibility, we examined 6608 genes in *Arabidopsis* shown by microarray analysis to be expressed in pollen (Schmidt et al., 2011; see Methods); we found that 196 were present on the Compositae list and 107 were present on the Solanaceae COSII list. We thus modified the Compositae and Solanaceae COSII lists by removing the UCOs that are not expressed in *Arabidopsis* pollen and used the remaining UCOs (Supplemental Data Set 1) as query sequences to search the S_2-pollen and S_3-pollen transcriptome data. The results of this analysis showed that ~82% of the Compositae UCOs were discovered in S_2-pollen and S_3-pollen Trinity assemblies, and 87.9 and 85.0% of the Solanaceae COSII UCOs expressed in *Arabidopsis* pollen were discovered in S_2-pollen and S_3-pollen Trinity assemblies, respectively (Supplemental Table 3). We then used the subset of *Arabidopsis* genes shown to be expressed in pollen (totaling 6608) to query the S_2-pollen and S_3-pollen assemblies and found that 5621 (85.1%) were discovered in the S_2-pollen Trinity assembly and 5555 (84.1%) were discovered in the S_3-pollen Trinity assembly (Supplemental Table 3); the percentages were similar to those of UCO discovery in the S_2-pollen and S_3-pollen assemblies.

To confirm that the depth of sequencing was sufficient (i.e., an increase in reads would not likely increase the number of unigenes discovered), we generated unigene accumulation plots for each assembly (raw Trinity assemblies and filtered unigene sets; Supplemental Figure 2), using a random sampling method of total reads plotted against the number of unigenes detected in the assembly (Der et al., 2011). We then used these plots to determine the percentage of total reads at 90, 95, 99, and 100% unigenes discovered. The results showed that 11% or less of total reads corresponded to 95% unigene discovery, and 25% of total reads or less corresponded to 99% unigene discovery (Supplemental Table 4), indicating that additional sequencing would not enhance gene discovery.

Assessing Purity and Tissue Specificity of Pollen and Leaf Transcriptomes

Due to the highly sensitive nature of next-generation driven RNA-seq, contamination of sequences from unrelated sources is a potential issue. To identify any possible contaminating sequences, we first annotated our assemblies using BLASTp to query the translated amino acid sequences of the filtered unigene set into the National Center for Biotechnology Information (NCBI) nonredundant (nr) database. The top 10 hits from the nr database were used to characterize each unigene. These BLASTp results were then used to distribute the sequences taxonomically by using a lowest common ancestor (LCA) analysis in the MEtaGenome Analyzer (MEGAN, v4.70.4) (Huson et al., 2011). Of the total unigenes classified in each assembly, at least 96% were found to be taxonomically assigned to green plants, while <3% of assigned sequences were classified as originating from unrelated taxa, including fungi, bacteria, and viruses (Supplemental Table 5).

It was also important to determine whether non-pollen-expressed genes were present in our pollen transcriptomes, as this might lead to misidentification of *SLF* genes in the pollen assemblies. It was equally important to determine whether the leaf transcriptome assembly was contaminated with pollen sequences since these data were used as a control for identifying pollen-specific genes (i.e., genes expressed in pollen but not in leaves). To assess the possible presence of leaf transcripts in the pollen assemblies, we used the coding sequences of the *rbcL* genes of *P. hybrida* and *P. axillaris* for the large subunit of Rubisco and four genes of *P. axillaris* encoding photosystem II (PSII) proteins to query all three unprocessed transcript assemblies with BLASTn. For the S_3S_3-leaf assembly, an HSP length of 1478 bp matched *rbcL* of *P. axillaris* with 99.9% identity, an HSP length of 691 bp matched *rbcL* of *P. hybrida* with 99.7% identity, and four HSPs of 97 to 1512 bp each matched one of the four PSII genes with 100% identity (Supplemental Table 6). For the S_2-pollen assembly, an HSP length of 206 bp matched *rbcLs* of *P. hybrida* and *P. axillaris* with 99.5 and 100% identity, respectively, but no unigenes matched any of the four PSII genes. The S_3-pollen assembly did not have any hit to *rbcL* or the four PSII genes. Although the S_2-pollen assembly did contain a unigene that matched *rbcL* at nearly 100% identity, its length was much shorter than that of the S_3S_3-leaf assembly. This, coupled with the lack of hits to any of the four *PSII* genes, would indicate that leakage of non-pollen transcripts was very low in the pollen transcriptomes. We also used BLAST to query pistil-specific S_2-*RNase* and S_3-*RNase* into S_2-pollen, S_3-pollen, and S_3S_3-leaf assemblies. We found no significant matches in the pollen assemblies, and only HSP with lengths of 373 bp at 87% (S_3-*RNase*) and 89% identity (S_2-*RNase*) in the S_3S_3-leaf assembly, further indicating the tissue specificity of all the assemblies.

To further examine the tissue specificity of the S_2-pollen, S_3-pollen, and S_3S_3-leaf assemblies, as well as a proof of concept in discovering new types of *SLF* genes, we used the nine *SLF* genes previously identified in the S_2-haplotype (*SLF1*, *SLF3*, *SLF4*, *SLF5*, *SLF6*, *SLF7*, *SLF8*, *SLF9*, and *SLF10*) and the five *SLF* genes previously identified in the S_3-haplotype (*SLF1*, *SLF5*, *SLF6*, *SLF9*, and *SLF10*) to query the unprocessed S_2-pollen and S_3-pollen Trinity assemblies by BLASTn. The results showed that all but *SLF4* and *SLF9* were matched at >98% identity in both S_2-pollen and S_3-pollen assemblies. In the S_2-pollen assembly, the most similar unigene to S_2-*SLF4* was 95% identical at the nucleotide sequence level, with a region of 100% identity spanning 359 bp. We concluded that this unigene resulted from a misassembly of S_2-*SLF4* and another highly similar novel *SLF* (later named S_2-*SLF12*), as no further evidence of this gene was found (see the section "Acquiring Full-Length Candidate *SLF* Sequences"). Moreover, this misassembled transcript was only found in the S_2-pollen assembly. The unigenes in the S_2-pollen and S_3-pollen assemblies most similar to S_2-*SLF9* and S_3-*SLF9* were 93% identical with their respective *SLF9* alleles; however, they matched S_2-*SLF10* and S_3-*SLF10* at 100 and 99% identity, respectively. The absence of *SLF9* in the S_2-pollen and S_3-pollen assemblies was most likely due to assembly error caused by the very high sequence similarity (93% identity) between *SLF9* and *SLF10*.

SLF2 (formerly named Pi-*SLFLc*; Hua et al., 2007) was initially identified from the S_1-haplotype and had not been confirmed in the S_2-haplotype or S_3-haplotype. Using the sequence of S_1-*SLF2* as BLAST query, this gene was identified from the unprocessed S_2-pollen and S_3-pollen assemblies. Using the full-length sequence of S_1-*SLF7* (formerly named *SLFLa*-S_1) and the partial coding sequence of S_2-*SLF7* (formerly named Pi-*SLFLa*-S_2) as BLAST queries, we found several matching full-length unigenes that corresponded to *SLF7* in both S_2-pollen and S_3-pollen assemblies. Using the sequences of S_2-*SLF3* and S_2-*SLF8* as BLAST queries, full-length S_3-*SLF3* and S_3-*SLF8* sequences were found in the S_3-pollen assembly. Interestingly, when using S_2-*SLF4* as a BLAST query into the S_3-pollen assembly, no significant match was found. However, using primers designed in the internal coding regions of S_2-*SLF4* (all PCR primers and conditions listed in Supplemental Table 7), the full-length S_3-*SLF4* sequence was acquired by 5′ and 3′ RACE using S_3 pollen cDNA as template. Subsequently, primers were designed to amplify the full-length coding sequence and were used to amplify, clone, and sequence S_3-*SLF4* using S_3S_3 genomic DNA as template. All the *SLF* genes that were detected in this transcriptome analysis, but had not been previously described in the S_2-haplotype or S_3-haplotype, were confirmed by PCR cloning using S_2S_2 or S_3S_3 genomic DNA as template and sequenced.

The absence of *SLF4* and *SLF9* in the S_2-pollen and S_3-pollen assemblies might suggest that the assemblies were incomplete and/or that expression of these *SLF* genes was too low to acquire enough reads to assemble the appropriate transcript(s). To assess the expression levels of these *SLF* genes, we mapped back reads to the unigenes assembled by Trinity and to the genes missing from the assemblies. That is, the coding sequences of S_2-*SLF4* and S_2-*SLF9* were added to the S_2-pollen Trinity assembled unigenes, and the coding sequences of S_3-*SLF4* and S_3-*SLF9* were added to the S_3-pollen Trinity assembled unigenes. Using perl scripts included in the Trinity assembly package (Haas et al., 2013; http://trinityrnaseq.sourceforge.net/) in conjunction with Bowtie (Langmead et al., 2009) and RSEM (Li and Dewey, 2011), reads from each biological replicate were mapped back separately to determine the transcripts per million (TPM) value. The average TPM values of S_2-*SLF9* and S_3-*SLF9* were 1.10 and 3.11, respectively (Table 2), suggesting that *SLF9* is indeed expressed, albeit weakly, and present in both pollen assemblies. This finding supports our hypothesis that the high degree of sequence similarity of *SLF9* with *SLF10* in both S_2 and S_3 pollen, coupled with the low transcript level of *SLF9* in both *S*-haplotypes prevented successful assembly of its transcripts in S_2-pollen and S_3-pollen assemblies. *SLF4* had an average TPM value of 3.11 in the S_2-pollen assembly and an average TPM value of 0.07 in the S_3-pollen assembly, suggesting that the relative expression of *SLF4* in comparison to the other *SLF* genes was very low in both S_2 and S_3 pollen. To provide a frame of reference, the TPM values of *actin* in the S_2-pollen and S_3-pollen assemblies are also shown in Table 2. These findings support our method of novel *SLF* gene discovery by examining the transcriptome of pollen of two different *S*-haplotypes, as different types of *SLF* genes at the polymorphic *S*-locus differ substantially in their expression levels (with average TPM values ranging from 1.10 to 167.72 for the 10 *SLF* genes in S_2 pollen and 0.07 to 69.75 for the same 10 *SLF* genes in S_3 pollen) (Table 2).

To further validate the tissue specificity of the S_3S_3-leaf assembly, we used the 10 *SLF* genes of the S_3-haplotype as queries for BLAST searches into this assembly. Only S_3-*SLF7* matched with a 100% identity match of 240 bp in the F-box domain. However, as this gene (formerly named Pi-*SLFLa*) was previously shown to be pollen specific (Hua et al., 2007), this matched sequence was most likely derived from a gene whose F-box domain has a sequence identical to that of S_3-*SLF7*. These results and those mentioned above indicate that the pollen and leaf assemblies were tissue specific and that both the S_2-pollen and S_3-pollen assemblies can be used to discover as yet unknown *SLF* genes.

Discovery of Potential Novel *SLF* Genes

Several criteria have been used in previous studies to determine whether a gene is likely to be involved in pollen specificity: (1) pollen-specific expression, (2) an F-box domain in the deduced amino acid sequence, (3) existence in multiple *S*-haplotypes and allelic sequence polymorphism, and (4) linked to the *S*-locus (Sijacic et al., 2004; Kubo et al., 2010). Conventional F-box proteins typically contain protein-protein interaction motifs, such as leucine-rich repeats (LRRs), kelch repeats, or WD repeats, in their C-terminal domains (Kuroda et al., 2002; Lechner et al., 2006); however, no canonical protein-protein interaction motifs have been found among SLF proteins. To ensure that all possible candidate *SLF* genes were identified, we adopted a two-pronged approach: (1) using all 10 *SLF* genes (*SLF1* to *SLF10*) of the S_2-haplotype and S_3-haplotype to form a *Petunia SLF* profile hidden-markov model (HMM), which was then used to query unprocessed S_2-pollen and S_3-pollen Trinity assemblies (retaining hits to these queries for downstream analyses); and (2) using known *SLF* genes from multiple species that possess Solanaceae-type SI as queries for BLAST into unprocessed S_2-pollen and S_3-pollen Trinity assemblies. Both strategies allowed us to expedite the process of *SLF* gene discovery by quickly yielding candidate genes for downstream analyses. While HMM will most likely yield a more concise list of candidate *SLF* genes, BLAST queries can provide additional support for the candidate *SLF* genes identified by the first strategy, resulting in an exhaustive search for all possible *SLF* genes.

For the first strategy, we used all known *SLF* genes from the S_2-haplotype and S_3-haplotype (a total of 20 nucleotide sequences) to produce an *SLF* gene profile HMM by HMMER (Eddy, 2011). Using this *SLF* gene profile, we queried both S_2-pollen and S_3-pollen Trinity assemblies to yield a list of candidate *SLF* genes. Using an e-value cutoff threshold of 1e-10, a total of 66 and 79 candidate *SLF* genes were retained from the

Table 2. Relative Expression of Ten *SLF* Genes from S_2-Pollen and S_3-Pollen Assemblies as Determined by TPM

	S_2 Pollen			S_3 Pollen		
Gene	Biological Replicate 1	Biological Replicate 2	S_2-Pollen Replicate Average	Biological Replicate 1	Biological Replicate 2	S_3-Pollen Replicate Average
SLF1	36.26	26.89	31.57	4.28	7.77	6.03
SLF2	15.82	13.09	14.45	19.03	4.20	11.61
SLF3	70.53	83.44	76.98	75.17	53.42	64.29
SLF4	2.95	3.27	3.11	0.14	0.00	0.07
SLF5	68.63	33.76	51.19	41.42	29.44	35.43
SLF6	13.20	7.90	10.55	6.09	5.38	5.73
SLF7	30.60	35.58	33.09	48.37	22.55	35.46
SLF8	19.11	19.99	19.55	56.64	28.45	42.54
SLF9	1.17	1.03	1.10	2.15	4.07	3.11
SLF10	154.27	181.18	167.72	69.56	69.95	69.75
Actin	2671.86	2609.70	2640.78	2187.82	1640.65	1914.23

S_2-pollen and S_3-pollen assemblies, respectively. These unigene sequences were in turn used as queries by BLASTn (e-value of 1e-10 and only keeping one HSP per query) into the raw Trinity S_3S_3-leaf assembly. The unigenes of the S_2-pollen and S_3-pollen assemblies with percent identity greater than 98% and more than 90-bp alignment length were considered non-pollen specific, and they were removed from the list of the candidate *SLF* unigenes. The unigenes that were <98% identical, or had alignment length <90 bp to the S_3S_3-leaf assembly, as well as the unigenes without a match from the S_3S_3-leaf assembly, were retained for further analysis. From this analysis, 60 unigenes from the S_2-pollen assembly and 71 unigenes from the S_3-pollen assembly were found not to be present in the S_3S_3-leaf assembly. To determine which of these unigenes were most likely *SLF* related, we used discontiguous megaBLAST against the Nucleotide Collection (nr/nt) database as a means of annotating these unigenes, and the results were analyzed using BLAST2Go software (Conesa et al., 2005). From the top 10 BLAST results to each of these unigenes, we found that 35 and 34 unigenes contained hits to *Petunia SLF* genes and/or *Nicotiana DD* genes (predicted *SLF* genes in *Nicotiana alata*; Wheeler and Newbigin, 2007), while the remaining unigenes contained no hits to *SLF* genes; many were annotated as containing kelch repeats, WD repeats, or LRRs. Since no known SLF has been shown to contain these motifs, we removed these unigenes from the list of candidate *SLF* genes. The sequence identification numbers for each of these steps are shown in Supplemental Data Set 2. From these *SLF* candidates, we further removed unigenes matching the 10 known *SLF* genes (*SLF1* to *SLF10*), yielding a total of 17 unigenes from each of the S_2-pollen and S_3-pollen assemblies, representing potentially novel *SLF* genes. The total numbers of unigenes for these assemblies identified from each step are shown in Table 3.

For the second strategy, we used publicly available *SLF* genes and *SLF-like* genes from *Petunia*, *Prunus*, *Nicotiana*, and *Antirrhinum*, totaling 105 sequences (Supplemental Data Set 3), to query each raw Trinity assembly by BLASTn with low stringency parameters (e-value of 100, and up to 50 target sequences retained for each query) to identify as many candidate *SLF* genes as possible. Many duplicate HSPs resulted from these queries, and they were dereplicated based on the unigene sequence ID. From a query into the S_2-pollen and S_3-pollen assemblies, 1459 and 1474 unigenes, respectively, were retained. The nucleotide sequences were queried by BLASTn into the S_3S_3-leaf Trinity assembly, and using the same criteria for HSP specificity as used in the first strategy, 1032 unigenes from the S_2-pollen assembly and 1043 unigenes from the S_3-pollen assembly were found to not be present in the S_3S_3-leaf assembly. These large lists of unigenes made downstream analyses impractical at this stage; thus, we performed NCBI discontiguous megaBLAST to annotate these unigenes. Sequences of interest were determined by their annotated sequence description, using keywords including F-box, *S*-locus, *SLF*, *SFB*, and *DD* genes, in BLAST2Go software (Conesa et al., 2005). From this search, 88 (S_2-pollen assembly) and 92 (S_3-pollen assembly) total candidate F-box genes were returned; similar to the first strategy, multiple genes annotated as containing kelch repeats, WD repeat, and LRR F-box genes were discovered by examining the top ten BLAST HSP results to each of the candidate F-box genes and thus were not considered likely *SLF* candidates. The deduced 129-amino acid sequences of four S_2-pollen unigenes (S2P_comp59029_c0_seq1, S2P_comp59029_c4_seq1, S2P_comp59029_c4_seq4, and S2P_comp59029_c4_seq5) not only contained an F-box domain, but also a reverse-transcriptase-like (RVT_3) domain; we thus removed them from the list of *SLF* candidates. From these queries, we found that 27 and 24 unigenes in the S_2-pollen and S_3-pollen assemblies, respectively, contained hits to *Petunia SLF* genes and/or *Nicotiana DD* genes, while the remaining unigenes contained no hits to *SLF* or *SLF*-related genes. These genes were compared with the previously known *SLF* sequences, and a total of 17 and 18 novel *SLF* candidates were identified from the S_2-pollen and S_3-pollen assemblies, respectively (Table 3; Supplemental Data Set 4). The additional novel *SLF* candidate from the S_3-pollen assembly found from this strategy but not in the first strategy was later determined to not be an *SLF*, but an *SLFlike* gene (see next section).

Table 3. Number of Unigenes Reported as Candidate Novel *SLF* Genes at Each Step of Analysis

	S_2-Pollen	S_3-Pollen
Strategy 1 HMM profiles		
Total unigenes returned	66	79
Unigenes not found in S_3S_3-leaf assembly	60	71
SLF annotated unigenes	35	34
Novel candidate *SLF* unigenes	17	17
Novel candidate *SLF* unigenes (without misassemblies)	15	17
Strategy 2 BLAST query		
Total unigenes returned	1459	1474
Unigenes not found in S_3S_3-leaf assembly	1032	1043
SLF annotated unigenes	27	24
Novel candidate *SLF* unigenes	17	18
Novel candidate *SLF* unigenes (without misassemblies)	15	18

Obtaining Full-Length Sequences of *SLF* Candidate Genes

The candidate *SLF* unigenes identified for each *S*-haplotype by both strategies were compared. A unigene identified in the S_2-pollen assembly (S2P_comp58780_c0_seq1), previously mentioned as similar to S_2-*SLF4* (see section "Assessing Purity and Tissue Specificity of Transcriptomes"), showed sequence similarity to both *SLF4* and a novel *SLF*, based on alignment of their sequences (Supplemental Figure 3). That this unigene was assembled from reads corresponding to two different *SLF* genes was confirmed by the failure to amplify its sequence from S_2S_2 genomic DNA with SLF4FW and SLF12REV primers (Supplemental Table 7); this unigene was thus removed from the list of candidate *SLF*s. Another unigene (S2P_comp60249_c0_seq1) from the S_2-pollen assembly showed high sequence similarity with S_2-*SLF6*. In an alignment between these two sequences (Supplemental Figure 4), S2P_comp60249_c0_seq1 was found to align with 275 bp at the 3′ end of S_2-*SLF6*, with the exception of 3 bp differences and a 14-bp gap. Due to the high sequence similarity between these two sequences, we were unable to design primers specific to this unigene for genomic DNA walking. To assess the authenticity of this unigene, we designed primers SLF6gapFW and SLF6gapREV for PCR (Supplemental Table 7); however, no fragment was amplified from S_2S_2 genomic DNA, suggesting that this unigene most likely resulted from misassembled transcripts. Both misassembled unigenes from the S_2-pollen assembly (S2P_comp58780_c0_seq1 and S2P_comp60249_c0_seq1) were removed from the list of candidate *SLF* unigenes, resulting in the identification of 15 and 18 novel *SLF* candidates from the S_2-pollen and S_3-pollen assemblies, respectively (Table 3; Supplemental Data Sets 2 and 4). All of the novel candidate genes identified by the second strategy were also found by the first strategy in S_2-pollen assemblies; however, one unigene from the S_3-pollen assembly (S3P_comp310798_c0_seq1) found by the second strategy was not discovered by the first strategy (Supplemental Table 8; for clarity, both the unigene names and their corresponding *SLF* gene names, *SLF11* to *SLF17*, are shown). This S_3-pollen unigene S3P_comp310798_c0_seq1 was annotated as being similar to *N. alata SLF-like* gene *DD7-S_2* (later annotated as *SLFLike2*, but not a true *SLF* gene; -see section "Phylogenetic Analysis of Nine Novel SLF Candidate Genes"). The unigenes listed in this table represent the most likely candidates of novel *SLF* genes; however, only two unigenes (S2P_comp43811_c0_seq1 of S_2 pollen and S3P_comp44213_c0_seq1 of S_3-pollen) contained predicted full-length coding sequences. Moreover, multiple unigenes showed high sequence similarity, indicating that they might correspond to a single novel *SLF* gene.

To obtain full-length coding sequences for the unigenes that were not full length, primers were designed for both 5′ and 3′ genomic DNA walking (Supplemental Table 7). From the sequences generated by genomic walking, additional primers were designed in the 5′ and 3′ UTRs, or in the coding sequence containing the predicted start and stop codons (Supplemental Table 7). Using these primers for PCR on genomic DNA from the S_2S_2 and S_3S_3 genotypes, we confirmed the presence of these full-length novel genes. After obtaining the full-length unigenes, we found that multiple unigenes initially discovered from the first and second strategies actually originated from the same gene. Six full-length genes of the 15 unigenes in the S_2-pollen assembly were novel, and eight full-length genes of the 17 unigenes in the S_3-pollen assembly were novel (Supplemental Table 8). Three of the full-length genes identified in the S_3-pollen assembly (later named *SLF15*, *SLF16*, and *SLFLike2*) were not found in the S_2-pollen assembly, and one of the full-length genes identified in the S_2 pollen assembly (later named *SLF13*) was not found in the S_3 pollen assembly. To determine whether the genes not identified from the S_2-pollen or S_3-pollen assembly were present in the genome of the other *S*-haplotype, we used the previously mentioned primers designed in the UTRs and genomic DNA of the S_2S_2 or S_3S_3 genotype as template for PCR, to clone and sequence the resulting fragments. All the novel candidate *SLF* genes, except *SLFLike2*, were successfully obtained. *SLFLike2* was not amplified from the S_2S_2 genomic DNA. In total, eight full-length novel candidate *SLF* genes were discovered for the S_2-haplotype, and nine full-length novel candidate *SLF* genes were discovered for the S_3-haplotype.

We used the same approach described in the section "Assessing Purity and Tissue Specificity of Transcriptomes" to assess whether the expression levels of the *SLF* genes not identified in the S_2-pollen or S_3-pollen assembly were too low to acquire enough reads for assembling a full-length sequence segment. We compared expression levels of these *SLF* genes, in terms of TPM, by mapping back reads to their full-length sequences and the Trinity output. As shown in Table 4, the average TPM values of *SLF15* and *SLF16*, both of which were initially discovered in the S_3-pollen assembly but not in the S_2-pollen assembly, were 0.62 and 0.19, respectively, lower than those (5.59 to 30.4) of all the other novel *SLF* genes identified in the S_2-pollen assembly. Similarly, the average TPM value of S_3-*SLF13* was 0.65, lower than those (2.68 to 89.24) of all the other novel *SLF* genes identified in the S_3-pollen assembly. These results, coupled with the successful PCR amplification of genomic fragments, suggest that *SLF15* and *SLF16* are also present in the S_2-haplotype and that *SLF13* is also present in the S_2-haplotype.

Cloning and Analysis of Three Additional Alleles of Novel *SLF* Candidate Genes

For each of the seven functionally confirmed *SLF* genes, multiple alleles have been identified and sequenced (Sijacic et al., 2004; Hua et al., 2007; Kubo et al., 2010; Williams et al., 2014). Among these *SLF* genes, sequence data of *SLF1* are available for the largest number of alleles, 20 (S_1-S_{15}, S_{6a}, S_{17}, and S_{20}-S_{22}). We thus used MEGA6 software (Tamura et al., 2013) to

Table 4. Relative Expression of Seven Novel *SLF* Genes from S_2-Pollen and S_3-Pollen Assemblies as Determined by TPM

	S_2 Pollen			S_3 Pollen		
Gene	Biological Replicate 1	Biological Replicate 2	S_2-Pollen Replicate Average	Biological Replicate 1	Biological Replicate 2	S_3-Pollen Replicate Average
SLF11	29.43	31.37	30.40	88.77	89.72	89.24
SLF12	16.37	16.76	16.56	10.16	1.55	5.85
SLF13	0.60	27.37	13.98	0.95	0.35	0.65
SLF14	3.89	7.29	5.59	4.25	1.14	2.69
SLF15	0.56	0.68	0.62	4.09	1.28	2.68
SLF16	0.37	0.00	0.19	6.43	3.56	4.99
SLF17	33.67	16.08	24.87	12.26	5.29	8.77
Actin	2671.86	2609.70	2640.78	2187.82	1640.65	1914.23

perform a nucleotide and amino acid p-distance analysis to determine the range of allelic sequence identity for *SLF1*, so that we could use this criterion to assess whether *SLF* sequences from different *S*-haplotypes corresponded to alleles of the same gene. The 20 alleles of *SLF1* ranged from 92.0 to 100% identical at the nucleotide sequence level and 85.8% to 100% identical at the deduced amino-acid sequence level (Supplemental Table 9).

For the nine novel *SLF* candidate genes, we used the primers initially designed to clone their full-length sequences in the S_2-haplotype and S_3-haplotype to perform PCRs on genomic DNA isolated from plants homozygous for the S_5-haplotype, S_{6a}-haplotype, and S_{12}-haplotype. The PCR fragments were cloned and sequenced. For each *SLF* candidate, we performed the p-distance analyses on the nucleotide sequences of the DNA fragments obtained from these five *S*-haplotypes, S_2, S_3, S_5, S_{6a}, and S_{12}, as well as on their deduced amino acid sequences (Supplemental Table 10). For all these nine *SLF* candidate genes, the range of nucleotide sequence identity and the range of amino acid sequence identity were comparable to the corresponding range of *SLF1* (Supplemental Table 9), suggesting that, like *SLF1*, they all showed *S*-allele-specific sequence polymorphism.

We also performed the p-distance analysis between the S_2-allele and S_3-allele of eight of the nine *SLF* candidate genes (*SLF11* to *SLF17* and *SLFLike1*) and the 10 previously identified *SLF* genes (*SLF1* to *SLF10*) (Supplemental Table 11), between all the above-mentioned 18 *SLF* genes/*SLF* candidate genes of the S_2-haplotype (Supplemental Table 12), and between these 18 *SLF* genes/*SLF* candidate genes and *SLFLike2* of S_3-haplotype (Supplemental Table 13). For the S_2-allele and S_3-allele of *SLF11* to *SLF17* and *SLFLike1*, the ranges of allelic sequence identity at both the nucleotide and amino acid sequence levels were similar to those of the 10 previously identified *SLF* genes (*SLF1* to *SLF10*) (Supplemental Table 11). For pairwise sequence comparison of different types of *SLF* genes/*SLF* candidate genes of the S_2-haplotype (Supplemental Table 12), as well as for the S_3-haplotype (Supplemental Table 13), the degrees of nucleotide and amino acid sequence identities between all except *SLF9*, *SLF10*, *SLFLike1*, and *SLFLike2*, were similar. *SLFLike1* and *SLFLike2* were ~10 to 20% less similar to all the previously identified *SLF* genes and all the other *SLF* candidate genes. *SLF9* and *SLF10* showed a high degree of sequence identity (~92% at the nucleotide sequence level and ~86% at the amino acid sequence level); however, the nucleotide sequence identity between their respective S_2-allele and S_3-allele was higher (96.5% for *SLF9* and 98.7% for *SLF10*) than between their respective paralogs (e.g., S_2-*SLF9* and S_2-*SLF10*), suggesting that they are distinct, paralogous genes.

Phylogenetic Analysis of Nine Novel *SLF* Candidate Genes

We next performed phylogenetic analysis of S_2-, S_3-, S_5-, S_{6a}-, and S_{12}-alleles of eight of the nine novel *SLF* candidate genes (all except *SLFLike2*); S_3-, S_5-, S_{6a}-, and S_{12}-alleles of *SLFLike2* (whose transcript was not found in the S_2-pollen assembly and whose sequence could not be amplified from S_2S_2 genomic DNA); S_2-allele and S_3-allele of the 10 previously identified *SLF* genes (*SLF1* to *SLF10*) of *Petunia*, nine full-length *Nicotiana DD* genes, and representative *SLF* genes from *Antirrhinum* (SLF), *Prunus* (SFBa-d), and *Pyrus* (ppSFBB). We produced separate phylogenetic trees using a bootstrapped maximum likelihood approach, based upon ClustalW (Thompson et al., 1994) alignments of both the coding sequences and deduced amino acid sequences (Figures 1A and 1B; Supplemental Data Sets 8 and 9) using RaxML (Stamatakis, 2014). From these results, we could define five separate clades, supported by >90% of 1000 bootstrap replicates. One of these clades contained all the 10 previously identified *Petunia SLF* genes, seven of the nine novel *SLF* candidate genes identified in this work, and eight *N. alata DD* genes (denoted as the Solanaceae *SLF* clade). This finding provides strong support that these seven *SLF* candidate genes are indeed *SLF* genes; thus, they were named *SLF11* to *SLF17* (see Supplemental Table 8 for their corresponding unigene IDs). Four other clades were present, a *Prunus SFB* clade, a *Pyrus SFBB* clade, an *Antirrhinum SLF* clade, and an intermediate clade containing the other two of the nine novel *SLF* candidate genes identified in this work. The same phylogenetic relationship was observed at the amino acid sequence level (Figure 1B). Interestingly, the *N. alata DD* gene, *DD7*, was also present in this intermediate clade; however, no *SLF* previously shown to be involved in SI was present. Thus, these two *SLF* candidate genes were named *SLFLike1* and *SLFLike2*. Nucleotide BLAST results of these intermediate sequences showed that they were similar to *CPR30-like* genes, a class of F-box genes found to be similar to *SLF* genes (L. Wang et al., 2004).

Sequence Comparison of 17 *SLF* Genes

Previously, the normed variability index (NVI) analysis (Kheyr-pour et al., 1990) was used to compare allelic sequences of the same type of SLF, including *Petunia* SLF1 (Hua et al., 2007) and *Prunus* SFB (Nunes et al., 2006). With the identification of the seven novel *SLF* genes, we used this analysis to examine the sequences across the 17 SLF proteins for both the S_2-haplotype and S_3-haplotype. As allelic sequence diversity of most of the previously known SLF proteins is low (<15%), by analyzing a more diverse set of sequences, we might be able to identify regions involved in interactions with S-RNases that are common among all SLFs. The results of the NVI analysis for SLF1 to SLF17 of the S_2-haplotype and SLF1 to SLF17 of the S_3-haplotype are shown in Figure 2A. Not surprisingly, the window-average plot for the 17 SLF proteins of the S_2-haplotype was very similar to that of the S_3-haplotype, with the exception of two regions near amino acid residues 80 and 340. Interestingly, alternating levels of variability and similarity repeated over a span of 20 to 30 amino acid residues. This observation was also made from alignments of the deduced amino acid sequences of all 17 SLF proteins of the S_2-haplotype and all 17 SLF proteins of the S_3-haplotype (Supplemental Figures 5A and 5B). To correlate this pattern of amino acid similarity and variability with predicted SLF structural features, we used VisCoSe (Spitzer et al., 2004) to produce a consensus sequence among the 17 SLF proteins of the S_2-haplotype based on the alignment shown in Supplemental Figure 5A. This consensus sequence was used to predict the secondary structure and solvent accessibility, using I-TASSER (Roy et al., 2010). Three types of secondary structures were predicted from these analyses: coils (C), sheets (S), and helices

(H). In addition, the solvent accessibility was ranked from 0 to 9, where 0 represents a completely buried residue and 9 represents a residue with access to solvent (Supplemental Data Set 5). To determine whether the NVI values were correlated to any of the predicted secondary structures, we averaged the NVI of residues that reside in coils, sheets, or helices and found that the average NVI value for residues in coils was higher (−0.28) than those in helices (−0.41) and sheets (−0.47). Using a similar approach with solvent accessibility values, we found that coiled regions were also rated as more solvent accessible (coils, 2.51; sheets, 0.57; helices, 1.55). Using a four-residue averaged window NVI analysis, we plotted the position of these predicted secondary structures. One region stood out: a coil that spanned from consensus residues 253 to 258. Within the predicted coil spanning residues 253 to 257, position 257 matched an NVI value of 1 (completely divergent) and corresponded to a region within the alignment (Supplemental Figure 5A) with many gaps, indicating that this region is subject to frequent deletion/insertion variation. Similarly, the predicted helix and coils at the end of the C-terminal domain (residues 415 to 417; Supplemental Data Set 5) matched NVI values of 1 (completely divergent); this is most likely due to the varying lengths of the amino acid sequences of different types of SLF proteins. The higher solvent accessibility (Supplemental Data Set 5), high sequence variability (Figure 2B), and their placement in the C-terminal domain suggest that these regions are likely involved in S-RNase recognition.

S-Locus Linkage Analyses

A gene involved in pollen specificity must be located at the *S*-locus, as assessed by the segregation of the gene with the *S-RNase* gene from the same *S*-haplotype (Sijacic et al., 2004; Hua et al., 2007; Kubo et al., 2010; Williams et al., 2014). As the seven novel *SLF* genes were identified from both the S_2-haplotype and S_3-haplotype, we raised 30 progeny plants from bud-selfing an S_2S_3 plant for testing the linkage of these genes to the *S*-locus. Only if the pattern of inheritance of both alleles of an *SLF* gene followed that of its corresponding *S-RNase* (for example, cosegregation of S_2-*SLF11* and S_2-*RNase*), would the gene be considered to be linked to the *S*-locus. This analysis requires that the allelic sequences of each gene contain regions that are sufficiently divergent to allow design of PCR primers to distinguish between them. For *SLF11*, *SLF12*, *SLF13*, and *SLF16*, we were able to design S_2-allele- and S_3-allele-specific primers; however, we were not able to do so for the other three novel *SLF* genes (*SLF14*, *SLF15*, and *SLF17*) due

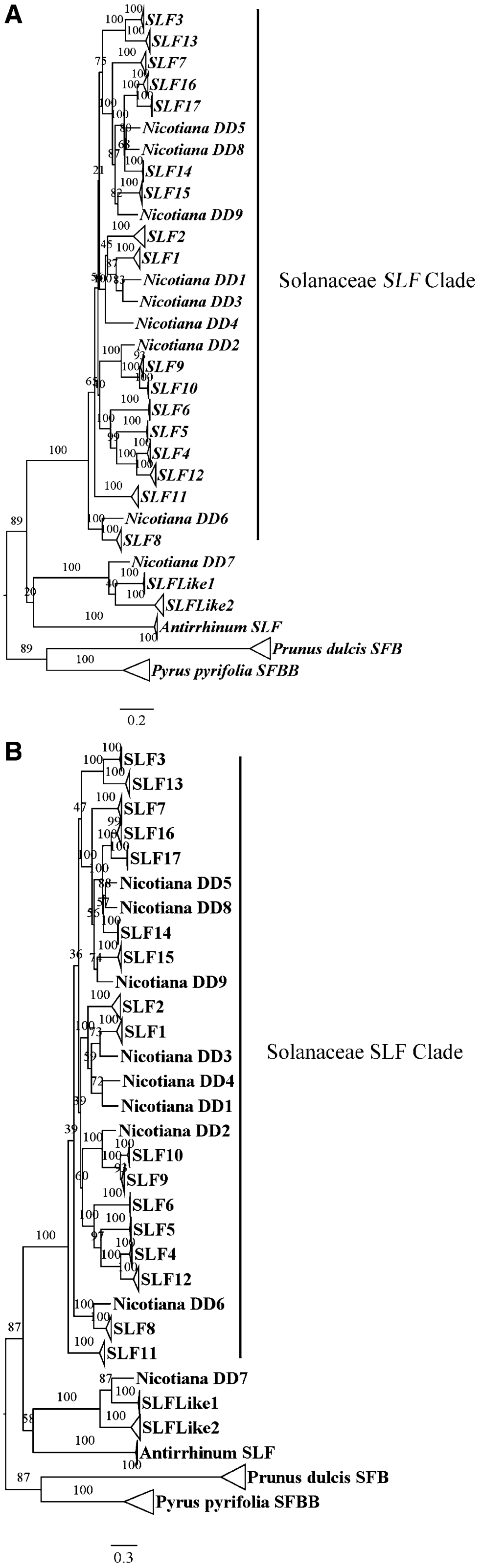

Figure 1. Phylogenetic Analysis of Novel *SLF* Candidate Genes of *P. inflata*.

Sequences of all novel *SLF* candidate genes identified in this work, the S_2-allele and S_3-allele of the 10 previously known *SLF* genes of *P. inflata*, *Nicotiana DD* genes, and representative *SLF* genes from *Prunus*, *Pyrus*, and *Antirrhinum*, were used to generate phylogenies by the maximum likelihood method, using their respective coding sequences **(A)** and deduced amino acid sequences **(B)**. Bootstrap support values of 1000 replicates are indicated as percentages. The Solanaceae *SLF* clade is marked by a black bar to the right of the sequence names. The bars depict the branch lengths, measured as the number of substitutions per site. The ClustalW alignments used to generate the coding sequence and deduced amino acid phylogenies are available online in Supplemental Data Sets 8 and 9, respectively.

to the high degree of allelic sequence identity between the S_2-allele and S_3-allele (Supplemental Table 7). Using genomic DNA from the 30 progeny plants as template for PCR, we found that all of the above-mentioned four novel *SLF* genes were linked to the *S*-locus (Supplemental Figure 6). We also examined *SLFLike1* and found it to be linked to the *S*-locus (Supplemental Figure 6).

DISCUSSION

Prior to the transcriptome analysis performed in this work, 10 *SLF* genes of *Petunia* had been reported, seven (*SLF1* to *SLF3* and *SLF7* to *SLF10*) first identified in *P. inflata* (Y. Wang et al., 2003, 2004; Sijacic et al., 2004; Hua et al., 2007) and three (*SLF4* to *SLF6*) first identified in *P. hybrida* (Kubo et al., 2010). Among them, seven (*SLF1* to *SLF6*, and *SLF8*) have been shown by a transgenic functional assay to be involved in pollen specificity (Sijacic et al., 2004; Kubo et al., 2010; Williams et al., 2014). In this work, we examined the S_2-pollen, S_3-pollen, and S_3S_3-leaf transcriptomes assembled de novo from next-generation RNA-seq data in order to identify all the *SLF* genes that collectively encode the pollen specificity determinant in the S_2-haplotype and S_3-haplotype. To strengthen the claim that we identified most, if not all, of the *SLF* genes at the *S*-loci of these two haplotypes, we performed multiple analyses to assess assembly quality and coverage. Using an UCO discovery method of coverage estimation (Fulton et al., 2002; Wu et al., 2006; Kozik et al., 2008; Der et al., 2011), based upon gene expression data from *Arabidopsis* pollen (Schmidt et al., 2011), we found that ~85% of UCOs were discovered in both S_2-pollen and S_3-pollen assemblies (Supplemental Table 3). To determine whether the depth of sequencing was sufficient, we generated a unigene accumulation curve (Der et al., 2011), which indicates that additional sequencing would not aid gene discovery, as 99% of unigenes were discovered with only 25% or less of the total reads in each assembly (Supplemental Figure 2 and Supplemental Table 4).

As a means of determining the level of contaminating sequences in these transcriptomes, we used an LCA analysis using MEGAN software (Huson et al., 2011) to taxonomically classify each predicted protein-encoding unigene and found that <3% of the assigned sequences originate from taxa not related to *Petunia* (Supplemental Table 5). Coupled with the results from querying chloroplast-specific genes into each assembly (Supplemental Table 6), this indicates that there is little contamination in the S_2-pollen, S_3-pollen, or S_3S_3-leaf transcriptomes.

In an initial proof of concept, we queried the 10 known *Petunia* *SLF*s (*SLF1* to *SLF10*) into the S_2-pollen and S_3-pollen assemblies. In the S_2-pollen transcriptome, all but *SLF9* was found (*SLF4* was misassembled), including S_2-*SLF2* and S_2-*SLF7*, which had previously been identified in the S_1-haplotype, but not in the S_2-haplotype. In the S_3-pollen transcriptome, all except *SLF4* and *SLF9* were discovered, including those *SLF*s that had not previously been identified in the S_3-haplotype (*SLF2*, *SLF3*, *SLF7*, and *SLF8*). Using S_3S_3 genomic DNA as template, *SLF4* was successfully cloned and subsequently sequenced. As previously mentioned, *SLF* genes with a high degree of sequence identity could lead to misassembly of unigenes and affect gene discovery; however, *SLF9* and *SLF10* appear to be the only pair that show a high degree of sequence identity (93%). The absence of assembled transcripts for *SLF4* from the initial analysis of the S_3-pollen transcriptome prompted us to map back reads to determine relative transcript levels of the 10 previously identified *SLF* genes. This analysis confirmed the presence of *SLF4* transcripts in the S_3-pollen transcriptome and revealed that expression of *SLF4* was much lower (average TPM = 0.07) than that of the other nine *SLF* genes (TPM values from 3.11 to 69.75) (Table 2). Similarly, for the seven novel *SLF* genes identified in this work (see discussion below), the absence of assembled transcripts for *SLF15* and *SLF16* in the S_2-pollen transcriptome and for *SLF13* in the S_3-pollen transcriptome led us to examine the expression levels of these genes (Table 4), and the results suggest that this was due to very low expression levels of *SLF15* (TPM = 0.62) and *SLF16* (TPM = 0.19) in S_2 pollen and *SLF13* (TPM = 0.65) in S_3 pollen. It is of note that, based on the transcriptome data, the expression levels of the 17 *SLF* genes in both S_2 and S_3 pollen vary over ~1000-fold. This raises the question as to whether this wide range of differences in expression has any biological relevance.

Our strategy to find novel *SLF* genes was to first query these pollen transcriptomes with a HMM profile built from the sequences of the 10 previously identified *SLF* genes of the S_2-haplotype and S_3-haplotype and BLAST/query these pollen transcriptomes at the nucleotide sequence level with these known *SLF* sequences. The novel, full-length *SLF* candidate genes identified in the S_2-haplotype and S_3-haplotype were confirmed by PCR cloning, using genomic DNA of the respective *S*-haplotypes as template. We then used several criteria established based on the properties of the known *SLF* genes to assess the candidate genes. The criteria include presence in the pollen assemblies but not in the leaf assembly, presence of multiple alleles, and placement in a monophyletic clade with all other known *Petunia SLF* genes. We identified the same seven novel *SLF* genes in both the S_2-haplotype and S_3-haplotype that fit all these criteria. For four of them, we have further shown that they are linked to the *S*-locus. One way to test the function of these novel SLF proteins is to use the well-established transgenic assay (Sijacic et al., 2004; Kubo et al., 2010; Sun and Kao, 2013; Williams et al., 2014) to determine whether they interact with any of their non-self S-RNases. For example, if expression of S_2-*SLF11* in pollen of the S_x-haplotype (S_x being any *S*-haplotype other than S_2) causes breakdown of SI in transgenic S_x pollen, this would suggest that S_2-SLF11 interacts with S_x-RNase to mediate its ubiquitination and degradation, allowing the transgenic S_x pollen to be accepted by S_x-carrying pistils. We could then conclude that *SLF11* is one of the *SLF* genes that constitute pollen specificity. It should be noted that, although the same suite of *SLF* genes was found to be expressed in both S_2 pollen and S_3 pollen in this work, there may be differences in the suite of *SLF* genes expressed in pollen of other *S*-haplotypes. For example, some *S*-haplotype may have a larger number of *SLF* genes due to having more types of *SLF* proteins that can detoxify the same non-self S-RNase (i.e., having a higher degree of *SLF* gene redundancy). However, it is not likely that there are large differences, judging from the finding that *SLF1* to *SLF6* are present in all six or seven *S*-haplotypes examined (Kubo et al., 2010) and all the seven *SLF* genes (*SLF11* to *SLF17*) identified in this work are present in all five *S*-haplotypes examined.

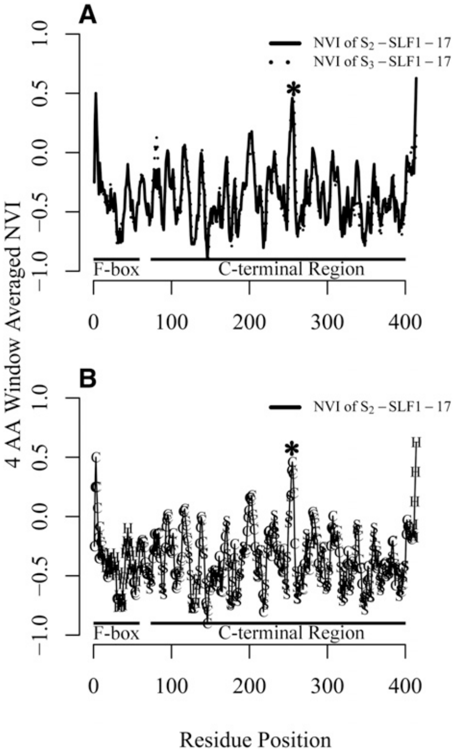

Figure 2. Sequence Variability and Predicted Secondary Structure and Solvent Accessibility of 17 SLF Proteins of *P. inflata*.

(A) The NVI for 17 SLF proteins of the S_2-haplotype, S_2-SLF1 to S_2-SLF17, was produced from alignment of their deduced amino acid sequences and averaged over a 4-amino acid window (thick line); the NVI for the same 17 SLF proteins of the S_3-haplotype, S_3-SLF1 to S_3-SLF17, was similarly produced (indicated with dots).

(B) Secondary structures and solvent accessibility, predicted from the consensus sequence among 17 SLF proteins of the S_2-haplotype (S_2-SLF1 to S_2-SLF17), are superimposed on the NVI plot of these proteins. Residues were predicted to form coils, sheets, and helices (indicated as C, S, and H, respectively). A predicted coil region (residue position 253 to 258) with divergent NVI values and accessibility to solvent is denoted with an asterisk. The predicted F-box domain and C-terminal region are marked by black bars underneath the plotted NVI values.

Using an NVI approach, we analyzed the level of variability among the deduced amino acid sequences of the 17 SLF proteins of the S_2-haplotype and among those of the 17 SLF proteins of the S_3-haplotype (Figure 2A). To understand the pattern of similarity and variability seen in this analysis, we also observed the predicted secondary structure and predicted solvent accessibility of these amino acid residues (Figure 2B; Supplemental Data Set 5). Analyzing the NVI and solvent accessibility values by their corresponding predicted secondary structure places emphasis on coiled regions as being more divergent and more accessible to solvent. In this analysis, we pinpoint two regions in particular that contain an aligned position with an NVI value of 1 (completely divergent).

If all 17 SLF proteins are shown to be involved in pollen specificity, is 17 a reasonable number of the total SLF proteins that constitute pollen specificity for an *S*-haplotype? In other words, would 17 different SLF proteins collectively have the capacity of interacting with all possible non-self S-RNases to detoxify them? In *Petunia*, 32 distinct *S*-haplotypes have been characterized (Sims and Robbins, 2009), but this most likely is the minimum number, as it is difficult, if not impossible, to know that all the haplotypes of any wild species have been exhaustively surveyed. From the limited interaction relationships determined to date, the largest number of S-RNases with which an SLF protein can interact is 5: S_2-SLF1 interacts with S_1-RNase, S_3-RNase, S_7-RNase, S_{12}-RNase, and S_{13}-RNase (Sijacic et al., 2004; Sun and Kao 2013; Williams et al., 2014). Assuming that all the other 16 SLF proteins are each capable of detoxifying five different S-RNases and that there are only 32 *S*-haplotypes in *Petunia*, this would suggest that at least two SLFs would be capable of detoxifying any given S-RNase. In this case, if an SLF protein loses its ability to interact with a non-self S-RNase due to mutation in its gene, pollen would still remain cross compatible with pistils producing this non-self S-RNase. This functional redundancy, or fail-safe mechanism, is supported by the results reported by Sun and Kao (2013). They used an RNA interference approach to suppress the expression of *S_2-SLF1* in pollen of *S_2S_2* transgenic plants and found that transgenic pollen remained compatible with *S_3*- *S_7*- and *S_{13}*-carrying pistils, indicating that S_2-SLF1 is not the only SLF protein that can interact with and detoxify S_3-RNase, S_7-RNase, and S_{13}-RNase.

Two *SLF*-like genes, named *SLFLike1* and *SLFLike2*, were identified along with the seven novel *SLF* genes; however, they do not fall in the *Petunia SLF* clade. Interestingly, *SLFLike1* is also linked to the *S*-locus, as the inheritance of *S_2-SLFLike1* and *S_3-SLFLike1* in 30 progeny plants from a bud-selfed *S_2S_3* plant precisely followed that of *S_2-RNase* and *S_3-RNase*, respectively. This is not the first time that an *SLF-like* gene was found to be linked to the *S*-locus; in an analysis of *SLF* candidate genes in *Nicotiana*, *DD1*, *DD4*, and *DD10* were found to be expressed in tissues other than pollen, yet were also linked to the *S*-locus (Wheeler and Newbigin, 2007). Both the coding sequences and deduced amino acid sequences of *DD1* to *DD9* were included in our phylogenetic analyses of the novel *Petunia SLF* genes, and all except *DD7* were found to fall within the *Petunia SLF* clade. *DD7* was found to form a clade with *SLFLike1* and *SLFLike2*. Interestingly, *S_2-SLFLike2* is not present in the S_2-haplotype, as we did not find its transcript in the *S_2*-pollen transcriptome assembly and failed to amplify genomic DNA using primers designed based on *S_3-SLFLike2*. Analysis of additional alleles of *SLFLike2* showed that *S_{6a}-SLFlike2* and *S_{12}-SLFlike2* encode full-length proteins, but a stop codon occurs at the 139th codon in *S_5-SLFLike2*, effectively truncating more than half of the C-terminal domain. These results, coupled with the finding that all other *DD* genes analyzed are present within the monophyletic clade of currently known *Petunia SLF* genes, suggest that *DD7*, *SLFLike1*, and *SLFLike2* may be vestiges of ancestral *SLF* genes or transcribed pseudogenes.

These transcriptome data have provided a wealth of information, the usefulness of which goes beyond *SLF* gene discovery. For example, these data have already proven useful for generating a protein database for identifying the components (SSK1, Cullin1-P, and RBX1) of the SLF-containing complex and for identifying the SLF proteins that interact with Pi-SSK1 (Li et al., 2014). One such SLF, previously referred to by Li et al. (2014) as

SLFx, is one of the seven novel SLF proteins identified in this work and has been renamed SLF13. Moreover, with the identification of the seven novel *SLF* genes, there are a total of 17 *SLF* sequences that can be used as molecular markers for delimiting the *S*-locus. Using the previously constructed S_2S_2 BAC library (McCubbin et al., 2000; Y. Wang et al., 2004), we can screen this library by PCR with primers designed based on the 17 *SLF* genes and the *S-RNase* gene to isolate BAC clones that span the contiguous region of the *S*-locus that contains all the genes involved in pollen or pistil specificity. The *S*-locus of *Petunia* is located in a subcentromeric region rich in highly repetitive sequences, making whole-genome sequencing of a particular *S*-haplotype difficult to de novo assemble the *S*-locus sequences (Y. Wang et al., 2004). This could be circumvented by sequencing individual BAC clones containing one or more of the *SLF* genes, as there should be much fewer repetitive sequence elements contained in each BAC clone (with an ~100- to ~200-kb genomic fragment). Moreover, next-generation sequencing using larger sequence fragments and longer sequence reads would increase the likelihood of a successful assembly by sequencing through repetitive elements. This type of sequence data would provide the most accurate molecular characterization of the *Petunia S*-locus, which has so far only been defined genetically, and reveal the relative location of the *SLF* genes and the *S-RNase* gene within the *S*-locus. These sequence data would also reveal the presence of any *SLF* genes that might have been missed in the transcriptome analyses performed in this work and/or any unexpressed *SLF* genes. An additional benefit of the molecular characterization of the S_2-locus is the identification of the promoter regions of each *SLF* gene; analysis of these regions coupled with expression results may give insight into the different relative expression of the *SLF* genes (Tables 2 and 4).

Finally, establishing the relationship between each of the 17 SLF proteins of the S_2-haplotype and S_3-haplotype and their respective non-self S-RNases will help elucidate the biochemical basis for the differential interactions between SLF and S-RNase by identifying the amino acid residues that are conserved among the SLF proteins that interact with common S-RNases, but divergent among the SLF proteins that do not.

METHODS

Plant Material

Six genotypes of *Petunia inflata* plants were used: S_2S_2, S_3S_3, and S_2S_3 were described by Ai et al. (1990), and S_5S_5, $S_{6a}S_{6a}$, and $S_{12}S_{12}$ were described by Sun and Kao (2013) and Williams et al. (2014). S_2S_2, S_3S_3, and S_2S_3 plants used in the *S*-locus linkage analysis were progeny of a bud-selfed S_2S_3 plant. All plant genotypes were determined by PCR using primers designed based on the nucleotide sequences of S_2-*RNase*, S_3-*RNase*, S_5-*SLF1*, S_{6a}-*RNase*, and S_{12}-*RNase*, as described by Sun and Kao (2013) and Williams et al. (2014). All the PCR primers used in this work and PCR conditions are listed in Supplemental Table 7. DNAzol (Invitrogen) was used to isolate genomic DNA from leaf tissues according to the manufacturer's protocol. All plants were raised under controlled greenhouse conditions.

Isolation of Pollen and Leaf Total RNA, Library Construction, and Sequencing

Anthers of 15 flowers from late stage 5, as described by Lee et al. (1996), were collected from two biological replicates of both S_2S_2 and S_3S_3 plants, for a total of four samples. To minimize contamination of sporophytic tissues, pollen was separated from anthers by vortexing in 500 µL sterile RNase-free ice-cold double-distilled water, and this suspension was immediately removed to a new 1.5 mL microcentrifuge tube and centrifuged at 12,000*g* for 1 min at 4°C. The supernatant (excess water) was removed, and total RNA was isolated using Trizol reagent (Invitrogen). S_3S_3-leaf RNA was isolated from 0.1 g of leaf tissue using Trizol. To eliminate possible genomic DNA contamination, total RNA was processed through on-column DNase I (Qiagen) treatment. The quality of total RNA was assessed using an Agilent Bioanalyzer 2100, and total RNA was quantified using Qubit Fluorometric Quantitation (Invitrogen). Total RNA concentrations before entering the library preparation were: 320 ng/µL (leaf), 271 ng/µL (S_2S_2 replicate 1), 434 ng/µL (S_2S_2 replicate 2), 658 ng/µL (S_3S_3 replicate 1), and 255 ng/µL (S_3S_3 replicate 2). RNA-seq libraries were built in-house, following the Illumina TruSeq RNA protocol modified for dUTP strand specificity (Borodina et al., 2011). Transcriptome libraries were size selected using Pippin Prep (Sage Science), digested with uracil-DNA glycosylase (UDG; 20 µL reaction containing 17.5 µL cDNA, 2 µL 10× UDG buffer, 0.5 units UDG [1 unit UDG/1 µg of DNA] at 37°C, 15 min), and enriched by 18 cycles of PCR amplification (Illumina TruSeq protocol).

Sequence Quality Processing, Transcriptome Assembly, and Validation

Illumina HiSequation 2000 technology was leveraged in a 101-bp pair-end approach, and all operations involving the Illumina sample pipeline were performed in the Stephan Schuster lab (The Pennsylvania State University). Raw reads were initially processed according to the Illumina HiSequation 2000 pipeline (CASAVA v1.8.2). All reads generated were deposited in NCBI's Sequence Read Archive. These sequences were then processed using CLC Assembly Cell (Version 3.1.1) for quality trimming (using options –r –f 33). Reads processed for quality were assembled separately by tissue type (i.e., S_2 pollen, S_3 pollen, and S_3S_3 leaf) de novo by Trinity (release 2012-10-05; using options–kmer_method jellyfish–max_memory 200–SS_lib_type RF–CPU 35). All data derived from high-throughput sequencing are associated with BioProject PRJNA244357. These Transcriptome Shotgun Assembly projects have been deposited at GenBank under the accessions GBDQ00000000, GBDR00000000, and GBDS00000000. The versions described in this article are the first versions GBDQ01000000, GBDR01000000, and GBDS01000000 corresponding to S_2-pollen, S_3-pollen, and S_3S_3-leaf assemblies, respectively. To retrieve initial read and assembly statistics, FastQC v1.10.1 (Andrews, 2010) and Usearch-sortbylength (Edgar, 2010) output were used. To predict coding sequences, ESTScan software v2.1 (Iseli et al., 1999; Lottaz et al., 2003) was used to process the raw assembly using default parameters and the *Arabidopsis thaliana* scoring matrix. Subsequently, Usearch v5.2.32 and v6.0.307 (Edgar, 2010) were used to dereplicate these predicted coding sequences by substring, prefix, and suffix (as described online, http://www.drive5.com/usearch/manual/). This filtered unigene set was annotated using BLASTp against the NCBI nr database (e-value = 1e-10). The top 10 hits from each query were kept for further analysis. Functional categories and domains were identified from the BLAST results using BLAST2Go with the default settings (Conesa et al., 2005).

To evaluate if sufficient sequencing depth was achieved to detect all expressed transcript sequences, a unigene accumulation curve was generated (Der et al., 2011). This curve, analogous to a species accumulation curve in ecological biodiversity assessment (a type of rarefaction), plots the total number of unigenes detected as a function of sampling effort (number of reads sequenced). As sequencing depth increases, the number of new unigenes detected in the sample increases at a slower rate as complete sampling (unknown total number of unigenes) is approached. For each data set, reads were mapped back to the assembly and the number of reads per unigene was used as input for a custom script (available at https://github.com/DocDer/scripts/blob/master/accumulation_curve.pl) to randomly sample reads and record the total number of unigenes detected. The shape of this curve will be affected by two features of the assembly. First, variation in read depth among transcripts will decrease the rate of increase along the left-hand side of the curve (highly abundant sequences will often be sampled repeatedly before rare transcripts are detected); second, the asymptotic height of the curve is limited by the total number of unigenes assembled from the data set. Sufficient sampling is determined by evaluating whether saturation has been achieved as sequencing effort is increased.

The UCO (Kozik et al., 2008) gene detection analysis was performed using two separate UCO lists. The Compositae list, consisting of 357 eukaryotic UCO sequences from *Arabidopsis* (http://compgenomics.ucdavis.edu/compositae_reference.php) and the Solanaceae COSII (Conserved Ortholog Set II) data set, consisting of 2869 sequences (http://solgenomics.net/documents/markers/cosii.xls) conserved in four Solanaceae species (*Solanum lycopersicum*, *Solanum pennellii*, *Solanum tuberosum*, and *Capsicum annuum*), as well as in *Arabidopsis* and *Coffea canephora* (Fulton et al., 2002; Wu et al., 2006). From the Solanaceae COSII UCOs, only those genes that were single copy and conserved among all species listed were maintained for analyses, totaling 220 genes. These UCO lists were compared with genes found to be expressed in *Arabidopsis* pollen; Schmidt et al. (2011) list pollen expressed genes in three data sets, only genes that were shown to be pollen expressed in all three data sets were used here. UCOs present in this list of pollen-expressed genes were used as queries for tBLASTn into S_2-pollen, S_3-pollen, and S_3S_3-leaf unigenes. A hit by BLAST was defined if the HSP was at least 30 amino acids (90 bp) in length and had an e-value of 1e-10 (Der et al., 2011). The accession numbers of all *Arabidopsis* sequences were obtained from TAIR (Lamesch et al., 2012; http://www.arabidopsis.org/doc/about/tair_terms_of_use/417). The accession numbers of all sequences used are listed in Supplemental Data Set 1.

To assess the contaminating sequences from species not related to *Petunia*, the ESTScan predicted coding sequences (see above) from each assembly were queried into the NCBI nr protein database using BLASTp (e-value 1e-10), with the top 10 hits retained. These results were used in MEGAN v4.70.4 using the LCA analysis (Huson et al., 2011). The predicted coding sequences were taxonomically classified if the following criteria were met: at least three BLAST hits were of a bit score 75 or greater, and these hits are not more than 10% from the best bit score.

Identification of Novel *SLF* Candidate Genes Using HMMER and BLAST

The nucleotide sequences of the S_2-allele and S_3-allele of the previously identified 10 *SLF* genes of *P. inflata* (S_2-*SLF1* to S_2-*SLF10* and S_3-*SLF1* to S_3-*SLF10*) were aligned using ClustalW (Thompson et al., 1994) by codon in MEGA6 (Tamura et al., 2013) with default settings. From this alignment, an *SLF* HMM profile using HMMER (Eddy, 2011) was generated, which was subsequently used to query the S_2-pollen and S_3-pollen Trinity assemblies using HMMER. Hits to the *SLF* HMM profile were ranked by e-value; a cutoff threshold e-value of 1e-10 was implemented to avoid analysis of divergent/non-*SLF*-related sequences. These nucleotide sequences were used to query the S_3S_3-leaf Trinity assembly using BLASTn, and any sequences with hits (defined as 90-bp alignment length at 98% identity) were removed. Discontiguous MegaBLAST (NCBI, http://blast.ncbi.nlm.nih.gov/Blast.cgi) with default parameters was used to annotate the remaining sequences, which were analyzed using BLAST2Go software with default parameters. All unigenes containing at least one hit to a known *SLF* gene (including *Nicotiana SLF* genes, named *DD* genes; Wheeler and Newbigin, 2007) were retained. The previously identified *SLF* genes were queried against these retained unigenes using BLASTn; hits with at least an e-value of 1e-10 and 98% identity were determined to be known *SLF* genes and removed. The remaining unigenes were determined to be candidates for novel *SLF* genes and retained for further analysis. Similarly, a BLASTn approach was used to independently confirm the results obtained from the first strategy, by querying the previously identified *SLF* genes from multiple species (the accession numbers of all sequences listed in Supplemental Data Set 3) using lax parameters (e-value of 100 and up to 50 target sequences). These unigenes were dereplicated and queried against the S_3S_3-leaf assembly, and hits to the S_3S_3-leaf assembly were removed in the same manner as the HMM results. The remaining unigenes were annotated using Discontigous MegaBLAST, which were analyzed by BLAST2Go software with default settings. All unigenes with at least one hit to known *SLF* genes from *Petunia* or to *Nicotiana DD* genes were held for further analysis (Supplemental Data Set 4). Known *SLF* sequences were removed from these unigenes in the same way as HMM results.

Cloning and Sequencing of Full-Length Novel *SLF* Candidate Genes

Genomic DNA walking (APAgene Gold Walking Kit) was employed according to the manufacturer's protocol to acquire full-length novel *SLF* candidate genes. Primers for genomic DNA walking were designed from the corresponding unigene. The partial sequence of S_3-*SLF4* was initially cloned by 5′, 3′ RACE (SMARTer RACE cDNA amplification kit) and subsequently cloned from S_3S_3 genomic DNA using primers designed from the 5′ and 3′ ends containing predicted start and stop codons. These primers and PCR conditions are listed in Supplemental Table 7. Confirmation of full-length contiguous *SLF* candidate genes was performed by PCR, with primers based on 5′ and 3′ UTR and coding regions from the unprocessed Trinity assembly and genomic DNA walking results. Genomic DNA from S_2S_2, S_3S_3, S_5S_5, $S_{6a}S_{6a}$, and $S_{12}S_{12}$ homozygous plants of *P. inflata* were used as templates for PCR. The resulting PCR products (full-length *SLF* genes) were ligated into cloning vector pGEM T-Easy (Promega) and transformed into Stellar competent cells of *Escherichia coli* (Clontech). Plasmids were purified from 5 mL overnight cultures by miniprep using NucleoSpin plasmid kits (Macherey-Nagel) and sequenced at the Penn State Genomics Core Facility.

Identification of *SLF* Genes Not Detected in the S_2-Pollen or S_3-Pollen Assembly

SLF4, *SLF9*, *SLF15*, and *SLF16* were not detected in the S_2-pollen assembly, and *SLF4*, *SLF9*, and *SLF13* were not detected in the S_3-pollen assembly. The full-length coding sequences from each of these genes were added to their respective reference transcript files. Trinity perl script alignReads.pl (Haas et al., 2013; http://trinityrnaseq.sourceforge.net/) in combination with Bowtie (Langmead et al., 2009) (with the following additional options: –aligner bowtie–SS_lib_type RF–prep_rsem–num_top_hits 1) were used to map reads from individual biological replicates back to their respective reference transcripts. To estimate relative transcript abundance, we used only properly mapped read pairs in conjunction with RSEM (Li and Dewey, 2011), including rsem-prepare-reference (with options –no-polyA–transcript-to-gene-map). The transcript-to-gene-map file assigned all unigenes matching a particular *SLF* sequence to that *SLF*, as determined by HMMER, BLAST, and sequencing results. Unigenes corresponding to actin were identified by tBLASTn, using *Nicotiana* actin; the full-length deduced amino acid sequences matched at >98% identity. TPM values were determined using rsem-calculate-expression (with options –bam–paired-end–no-qualities–calc-ci) using the properly mapped pair output from Trinity alignReads.pl.

Sequence Divergence, Phylogenetic, and Normed Variability Index Analyses

To determine sequence identity between alleles of the same type of *SLF* gene (previously identified and the novel *SLF* candidate genes), and sequence identity between different types of these *SLF* genes and novel *SLF* candidate genes of the S_2-haplotype or S_3-haplotype, nucleotide sequences and deduced amino acid sequences were first aligned using ClustalW (Thompson et al., 1994; default parameters) and analyzed by p-distance analysis from MEGA6 software (Tamura et al., 2013; using pairwise deletion).

To generate the phylogeny, *SLF* sequences from *Prunus*, *Pyrus*, *Antirrhinum*, *Nicotiana*, and *Petunia* (all accession numbers are listed in Supplemental Data Set 6) were first aligned using ClustalW (Thompson et al., 1994) provided by MEGA6 (Tamura et al., 2013) (default parameters were used; coding sequences were aligned by codon). From these alignments, RaxML (Stamatakis, 2014) was used to generate maximum likelihood phylogeny with bootstrap support using 1000 replicates. Phylogeny derived from coding sequences was generated (-lnL = 35,233.79) using raxmlHPC-PTHREADS-SSE3 (options: –f a –p 12,345 –x 12,345 -# 1000 –m GTRCAT) with the GTRCAT substitution model. Phylogeny derived from deduced amino acid sequences was generated (-lnL = 22,728.75) using raxmlHPC-PTHREADS-SSE3 (options: –f a –p 12,345 –x 12,345 -# 1000 –m PROTCATAUTO) and using the best-scoring substitution model, DUMMY2. These phylogenetic trees were subsequently edited for clarity using FigTreev1.4 (Rambaut, 2008).

Normed variability index analyses were performed according to Kheyr-pour et al. (1990) with the deduced amino acid alignments of S_2-*SLF1* to S_2-*SLF17* and S_3-*SLF1* to S_3-*SLF17*, produced using ClustalW (Thompson et al., 1994) and MEGA6 (Tamura et al., 2013) (default parameters). These alignments were also used to produce the consensus sequence of S_2-SLF1 to S_2-SLF17, using VisCoSe (Spitzer et al., 2004). This consensus sequence was submitted to I-TASSER (Roy et al., 2010; http://zhanglab.ccmb.med.umich.edu/I-TASSER/) to predict secondary structure and solvent accessibility.

S-Locus Linkage Assay

A *P. inflata* plant of the S_2S_3 genotype (Ai et al., 1990) was self-pollinated at an immature bud stage 2, as described by Lee et al. (1996), and the resulting seeds were germinated to produce a segregating population of S_2S_2, S_3S_3, and S_2S_3 genotypes. Genomic DNA was isolated from leaf tissues of each of 30 randomly chosen progeny plants, and the *S*-genotype of each plant was determined by PCR using primers specific for S_2-*RNase* and primers specific for S_3-*RNase*. For analysis of cosegregation of *SLF11*, *SLF12*, *SLF13*, *SLF16*, and *SLFLike1* with *S-RNase*, primers specific for the S_2-allele and primers specific for the S_3-allele of each gene were designed, and PCRs were performed using the genomic DNA samples from the 30 progeny plants. All primers and the corresponding PCR conditions are listed in Supplemental Table 7.

Accession Numbers

The accession numbers for all of the sequence data referenced in this article are listed in Supplemental Data Set 7.

Supplemental Data

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

Supplemental Figure 1. Workflow of *SLF* Gene Discovery from Sample Collection to Sequencing, Assembly, and Downstream Analyses.

Supplemental Figure 2. Unigene Accumulation Curves of Trinity Assemblies and ESTScan Processed Assemblies.

Supplemental Figure 3. Alignment of Nucleotide Sequences of S_2-*SLF4*, S_2-*SLF12*, and Unigene S2P_comp58780_c0_seq1.

Supplemental Figure 4. Alignment of Nucleotide Sequences of S_2-*SLF6* and Misassembled Unigene S2P_comp60249_c0_seq1.

Supplemental Figure 5. Alignments of Deduced Amino Acid Sequences of 17 *SLF* Genes from Both the S_2-Haplotype and S_3-Haplotype.

Supplemental Figure 6. *S*-Locus Linkage Analysis of Four Novel *SLF* Genes and an *SLF-Like* Gene.

Supplemental Table 1. Sequencing Statistics.

Supplemental Table 2. Initial UCO Analysis.

Supplemental Table 3. *Arabidopsis* Pollen Expressed UCO Analysis.

Supplemental Table 4. Read Accumulation Plot Statistics.

Supplemental Table 5. Least Common Ancestor Analysis.

Supplemental Table 6. Chloroplast-Specific Gene Expression Analysis.

Supplemental Table 7. Primer Sequences and PCR Conditions.

Supplemental Table 8. Novel *SLF* Genes and Corresponding Unigene IDs.

Supplemental Table 9. Nucleotide and Amino Acid Sequence Identities of 20 Alleles of *SLF1*.

Supplemental Table 10. Allelic Nucleotide and Amino Acid Sequence Identities for Seven Novel *SLF* Candidate Genes and Two *SLF-Like* Genes.

Supplemental Table 11. Allelic Nucleotide and Amino Acid Sequence Identities for Ten Previously Identified *SLF* Genes and Eight Novel *SLF* Candidate Genes.

Supplemental Table 12. Pairwise Comparison of Nucleotide and Amino Acid Sequence Identities of Ten Previously Identified *SLF* Genes and Eight Novel *SLF* Candidate Genes from the S_2-Haplotype.

Supplemental Table 13. Pairwise Comparison of Nucleotide and Amino Acid Sequence Identities of Ten Previously Identified *SLF* Genes and Nine Novel *SLF* Candidate Genes from the S_3-Haplotype.

Supplemental Data Set 1. Accession Numbers of UCOs.

Supplemental Data Set 2. Novel *SLF* Candidate Genes Identified by Strategy 1.

Supplemental Data Set 3. *SLF* Genes Used in Identifying *SLF* Genes by Strategy 2.

Supplemental Data Set 4. Novel *SLF* Candidate Genes Identified by Strategy 2.

Supplemental Data Set 5. I-TASSER Secondary Structure and Solvent Accessibility Prediction Data.

Supplemental Data Set 6. *SLF* Genes Used to Generate Phylogeny.

Supplemental Data Set 7. Accession Numbers of All Sequence Data Referenced in This Article.

Supplemental Data Set 8. ClustalW Alignment of Nucleotide Sequences Used to Produce the Coding Sequence Phylogeny.

Supplemental Data Set 9. ClustalW Alignment of Deduced Amino Acid Sequences Used to Produce the Amino Acid Sequence Phylogeny.

ACKNOWLEDGMENTS

We thank the Stephan Schuster Lab for performing next-generation sequencing, Paula Ralph for help in building RNA-seq libraries, Eric Wafula and Loren Honaas for general bioinformatics advice, Anthony Omeis for greenhouse management, Hongli Hao for general laboratory help, and Ryan Hewitt for help with bioinformatics process automation. This work was supported by Grant IOS-1146182 from the National Science Foundation to T.-h.K.

AUTHOR CONTRIBUTIONS

J.S.W. performed all the experiments. J.P.D. and J.S.W. planned and ran the bioinformatics analyses. J.S.W., J.P.D., C.W.D., and T.-h.K. designed the experiments. J.S.W., J.P.D., and T.-h.K. wrote the article. All authors contributed to the design of the study and approved the final article.

Received April 30, 2014; revised June 25, 2014; accepted July 9, 2014; published July 28, 2014.

REFERENCES

Ai, Y., Singh, A., Coleman, C.E., Ioerger, T.R., Kheyr-Pour, A., and Kao, T.-h. (1990). Self-incompatibility in *Petunia inflata*: isolation and characterization of cDNAs encoding three *S*-allele-associated proteins. Sex. Plant Reprod. **3:** 130–138.

Anderson, M.A., McFadden, G.I., Bernatzky, R., Atkinson, A., Orpin, T., Dedman, H., Tregear, G., Fernley, R., and Clarke, A.E. (1989). Sequence variability of three alleles of the self-incompatibility gene of *Nicotiana alata.* Plant Cell **1:** 483–491.

Andrews, S. (2010). FastQC: A quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Bai, C., Sen, P., Hofmann, K., Ma, L., Goebl, M., Harper, J.W., and Elledge, S.J. (1996). *SKP1* connects cell cycle regulators to the ubiquitin proteolysis machinery through a novel motif, the F-box. Cell **86:** 263–274.

Borodina, T., Adjaye, J., and Sultan, M. (2011). A strand-specific library preparation protocol for RNA sequencing. Methods Enzymol. **500:** 79–98.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. (2005). Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21:** 3674–3676.

Cornish, E.C., Pettitt, J.M., Bonig, I., and Clarke, A.E. (1987). Developmentally controlled expression of a gene associated with self-incompatibility in *Nicotiana alata.* Nature **326:** 99–102.

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

Der, J.P., Barker, M.S., Wickett, N.J., dePamphilis, C.W., and Wolf, P.G. (2011). De novo characterization of the gametophyte transcriptome in bracken fern, *Pteridium aquilinum*. BMC Genomics **12:** 99.

Eddy, S.R. (2011). Accelerated profile HMM searches. PLOS Comput. Biol. **7:** e1002195.

Edgar, R.C. (2010). Search and clustering orders of magnitude faster than BLAST. Bioinformatics **26:** 2460–2461.

Fulton, T.M., Van der Hoeven, R., Eannetta, N.T., and Tanksley, S.D. (2002). Identification, analysis, and utilization of conserved ortholog set markers for comparative genomics in higher plants. Plant Cell **14:** 1457–1467.

Goldraij, A., Kondo, K., Lee, C.B., Hancock, C.N., Sivaguru, M., Vazquez-Santana, S., Kim, S., Phillips, T.E., Cruz-Garcia, F., and McClure, B. (2006). Compartmentalization of S-RNase and HT-B degradation in self-incompatible *Nicotiana.* Nature **439:** 805–810.

Grabherr, M.G., et al. (2011). Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat. Biotechnol. **29:** 644–652.

Haas, B.J., et al. (2013). *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nat. Protoc. **8:** 1494–1512.

Hua, Z., and Kao, T.H. (2006). Identification and characterization of components of a putative petunia *S*-locus F-box-containing E3 ligase complex involved in S-RNase-based self-incompatibility. Plant Cell **18:** 2531–2553.

Hua, Z., and Kao, T.H. (2008). Identification of major lysine residues of S(3)-RNase of *Petunia inflata* involved in ubiquitin-26S proteasome-mediated degradation in vitro. Plant J. **54:** 1094–1104.

Hua, Z., Meng, X., and Kao, T.H. (2007). Comparison of *Petunia inflata S*-Locus F-box protein (Pi SLF) with Pi SLF like proteins reveals its unique function in S-RNase based self-incompatibility. Plant Cell **19:** 3593–3609.

Huang, S., Lee, H.S., Karunanandaa, B., and Kao, T.H. (1994). Ribonuclease activity of *Petunia inflata* S proteins is essential for rejection of self-pollen. Plant Cell **6:** 1021–1028.

Huson, D.H., Mitra, S., Ruscheweyh, H.-J., Weber, N., and Schuster, S.C. (2011). Integrative analysis of environmental sequences using MEGAN4. Genome Res. **21:** 1552–1560.

Iseli, C., Jongeneel, C.V., and Bucher, P. (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc. Int. Conf. Intell. Syst. Mol. Biol. **1999:** 138–148.

Iwano, M., and Takayama, S. (2012). Self/non-self discrimination in angiosperm self-incompatibility. Curr. Opin. Plant Biol. **15:** 78–83.

Kheyr-pour, A., Bintrim, S.B., Ioerger, T.R., Remy, R., Hammond, S., and Kao, T.-h. (1990). Sexual plant reproduction sequence diversity of pistil *S*-proteins associated with gametophytic self-incompatibility in *Nicotiana alata.* Sex. Plant Reprod. **3:** 88–97.

Kozik, A., Matvienko, M., Kozik, I., van Leeuwen, H., Van Deynze, A., and Michelmore, R.M. (2008). Eukaryotic ultra conserved orthologs and estimation of gene capture in EST libraries [abstract]. Plant and Animal Genomes Conference **16:** P6.

Kubo, K., Entani, T., Takara, A., Wang, N., Fields, A.M., Hua, Z., Toyoda, M., Kawashima, S., Ando, T., Isogai, A., Kao, T.H., and Takayama, S. (2010). Collaborative non-self recognition system in S-RNase-based self-incompatibility. Science **330:** 796–799.

Kuroda, H., Takahashi, N., Shimada, H., Seki, M., Shinozaki, K., and Matsui, M. (2002). Classification and expression analysis of Arabidopsis F-box-containing protein genes. Plant Cell Physiol. **43:** 1073–1085.

Lamesch, P., et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic Acids Res. **40:** D1202–D1210.

Langmead, B., Trapnell, C., Pop, M., and Salzberg, S.L. (2009). Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol. **10:** R25.

Lechner, E., Achard, P., Vansiri, A., Potuschak, T., and Genschik, P. (2006). F-box proteins everywhere. Curr. Opin. Plant Biol. **9:** 631–638.

Lee, H.S., Huang, S., and Kao, T. (1994). S proteins control rejection of incompatible pollen in *Petunia inflata.* Nature **367:** 560–563.

Lee, H.-s., Karunanandaa, B., Mccubbin, A., Gilroy, S., and Kao, T. (1996). PRK1, a receptor-like kinase of *Petunia inflata*, is essential for postmeiotic development of pollen. Plant J. **9:** 613–624.

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

Li, S., Sun, P., Williams, J.S., and Kao, T.H. (2014). Identification of the self-incompatibility locus F-box protein-containing complex in *Petunia inflata.* Plant Reprod **27:** 31–45.

Lottaz, C., Iseli, C., Jongeneel, C.V., and Bucher, P. (2003). Modeling sequencing errors by combining Hidden Markov models. Bioinformatics **19** (suppl. 2)**:** ii103–ii112.

Luu, D.T., Qin, X., Morse, D., and Cappadocia, M. (2000). S-RNase uptake by compatible pollen tubes in gametophytic self-incompatibility. Nature **407:** 649–651.

McCubbin, A.G., Wang, X., and Kao, T.H. (2000). Identification of self-incompatibility (*S*-) locus linked pollen cDNA markers in *Petunia inflata.* Genome **43:** 619–627.

Murfett, J., Atherton, T.L., Mou, B., Gasser, C.S., and McClure, B.A. (1994). S-RNase expressed in transgenic *Nicotiana* causes *S*-allele-specific pollen rejection. Nature **367:** 563–566.

Nunes, M.D.S., Santos, R.A., Ferreira, S.M., Vieira, J., and Vieira, C.P. (2006). Variability patterns and positively selected sites at the gametophytic self-incompatibility pollen *SFB* gene in a wild self-incompatible *Prunus spinosa* (Rosaceae) population. New Phytol. **172:** 577–587.

Qiao, H., Wang, H., Zhao, L., Zhou, J., Huang, J., Zhang, Y., and Xue, Y. (2004). The F-box protein AhSLF-S_2 physically interacts with S-RNases that may be inhibited by the ubiquitin/26S proteasome pathway of protein degradation during compatible pollination in *Antirrhinum.* Plant Cell **16:** 582–595.

Rambaut, A. (2008). FigTree v1.1.1: Tree figure drawing tool. http://tree.bio.ed.ac.uk/software/figtree/.

Roy, A., Kucukural, A., and Zhang, Y. (2010). I-TASSER: a unified platform for automated protein structure and function prediction. Nat. Protoc. **5:** 725–738.

Schmidt, A., Wuest, S.E., Vijverberg, K., Baroux, C., Kleen, D., and Grossniklaus, U. (2011). Transcriptome analysis of the *Arabidopsis* megaspore mother cell uncovers the importance of RNA helicases for plant germline development. PLoS Biol. **9:** e1001155.

Sijacic, P., Wang, X., Skirpan, A.L., Wang, Y., Dowd, P.E., McCubbin, A.G., Huang, S., and Kao, T.H. (2004). Identification of the pollen determinant of S-RNase-mediated self-incompatibility. Nature **429:** 302–305.

Sims, T.L., and Robbins, T.P. (2009). Gametophytic self-incompatibility in *Petunia*. In *Petunia:* Evolutionary, Developmental and Physiological Genetics. T. Gerats and J. Strommer, eds (New York: Springer), pp. 85–106.

Spitzer, M., Fuellen, G., Cullen, P., and Lorkowski, S. (2004). VisCoSe: visualization and comparison of consensus sequences. Bioinformatics **20:** 433–435.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics **30:** 1312–1313.

Stone, S.L., and Callis, J. (2007). Ubiquitin ligases mediate growth and development by promoting protein death. Curr. Opin. Plant Biol. **10:** 624–632.

Sun, P., and Kao, T.H. (2013). Self-incompatibility in *Petunia inflata*: the relationship between a self-incompatibility locus F-box protein and its non-self S-RNases. Plant Cell **25:** 470–485.

Tamura, K., Stecher, G., Peterson, D., Filipski, A., and Kumar, S. (2013). MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. Mol. Biol. Evol. **30:** 2725–2729.

Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994). CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22:** 4673–4680.

Vierstra, R.D. (2009). The ubiquitin-26S proteasome system at the nexus of plant biology. Nat. Rev. Mol. Cell Biol. **10:** 385–397.

Wang, L., Dong, L., Zhang, Y., Zhang, Y., Wu, W., Deng, X., and Xue, Y. (2004). Genome-wide analysis of *S*-Locus F-box-like genes in *Arabidopsis thaliana.* Plant Mol. Biol. **56:** 929–945.

Wang, N., and Kao, T.H. (2012). Self-incompatibility in *Petunia*: a self/nonself-recognition mechanism employing *S*-locus F-box proteins and S-RNase to prevent inbreeding. Wiley Interdiscip. Rev. Dev. Biol. **1:** 267–275.

Wang, Y., Tsukamoto, T., Yi, K.-W., Wang, X., Huang, S., McCubbin, A.G., and Kao, T.H. (2004). Chromosome walking in the *Petunia inflata* self-incompatibility (*S*-) locus and gene identification in an 881-kb contig containing S_2-RNase. Plant Mol. Biol. **54:** 727–742.

Wang, Y., Wang, X., McCubbin, A.G., and Kao, T.H. (2003). Genetic mapping and molecular characterization of the self-incompatibility (*S*) locus in *Petunia inflata.* Plant Mol. Biol. **53:** 565–580.

Wheeler, D., and Newbigin, E. (2007). Expression of 10 S-class *SLF-like* genes in *Nicotiana alata* pollen and its implications for understanding the pollen factor of the S locus. Genetics **177:** 2171–2180.

Williams, J.S., Natale, C.A., Wang, N., Li, S., Brubaker, T.R., Sun, P., and Kao, T.-H. (2014). Four previously identified *Petunia inflata* S-locus F-box genes are involved in pollen specificity in self-incompatibility. Mol. Plant **7:** 567–569.

Wu, F., Mueller, L.A., Crouzillat, D., Pétiard, V., and Tanksley, S.D. (2006). Combining bioinformatics and phylogenetics to identify large sets of single-copy orthologous genes (COSII) for comparative, evolutionary and systematic studies: a test case in the euasterid plant clade. Genetics **174:** 1407–1420.

Transcriptome Analysis Reveals the Same 17 *S-Locus F-Box* Genes in Two Haplotypes of the Self-Incompatibility Locus of *Petunia inflata*

Justin S. Williams, Joshua P. Der, Claude W. dePamphilis and Teh-hui Kao

Plant Cell 2014;26;2873-2888; originally published online July 28, 2014;
DOI 10.1105/tpc.114.126920