Gene duplication and genetic exchange drive the evolution of S-RNase based self-incompatibility

Ken-ichi Kubo¹, Timothy Paape², Masaomi Hatakeyama²³, Tetsuyuki Entani¹, Akie Takara¹, Kie Kajihara¹, Mai Tsukahara¹, Rie Shimizu-Inatsugi², Kentaro K. Shimizu² & Seiji Takayama¹

¹Graduate School of Biological Sciences, Nara Institute of Science and Technology, Ikoma 630-0192, Japan.
²Institute of Evolutionary Biology and Environmental Studies, University of Zurich, CH-8057 Zurich, Switzerland.
³Functional Genomics Center Zurich, CH-8057 Zurich, Switzerland.

Self-incompatibility (SI) systems in flowering plants distinguish self and non-self pollen to prevent inbreeding. While all other SI systems studied to date rely on the self-recognition between each single male- and female-determinants, the Solanaceae plants has a non-self recognition system that functions through the detoxification of non-self female-determinants of S-ribonucleases (S-RNases), expressed in pistils, by multiple male-determinants of S-locus F-box proteins (SLFs), expressed in pollen. However, little is known about how many SLF components constitute such a non-self recognition system and how they evolve. Here we conducted large-scale next-generation sequencing and genomic PCR and identified 16–20 SLFs in each S-haplotype in SI Petunia, for a total of 168 SLF sequences. We predicted the target S-RNases of SLFs by assuming that a particular S-allele must not have a conserved SLF that recognizes its own S-RNase, and validated them by transformation experiments. A simple mathematical model showed that 16–20 SLF sequences would be adequate to recognize the vast majority of target S-RNases. We found evidence of gene conversion events, which we suggest are essential to constitute a non-self recognition system and as well as contributed to self-compatible mutations.

SI is a genetically controlled reproductive barrier in angiosperms that allows the pistil to reject self (genetically-related) pollen and accept non-self (genetically-unrelated) pollen¹⁴. In most cases, this self/non-self discrimination is controlled by male- and
female-specificity determinants (pollen-\textit{S}, pistil-\textit{S}) encoded by multi-allelic genes at the \textit{S}-locus. Because \textit{pollen-S} and \textit{pistil-S} are tightly linked to each other at the \textit{S}-locus, combinations of these alleles are considered \textit{S}-haplotypes.

Two main types of SI system exist: self recognition and non-self recognition\textsuperscript{3}. Although SI species in Brassicaceae and in Papaveraceae differ in their determinant proteins, both possess self recognition that relies on the interactions between highly polymorphic molecules, a ligand and a kind of receptor, derived from a single \textit{S}-haplotype\textsuperscript{1,2}. In such a system, suppression of recombination between \textit{pollen-S} and \textit{pistil-S} results in corresponding shapes of phylogenetic trees of alleles (often called co-evolutionary relationships in a narrow sense)\textsuperscript{5}.

Self-incompatibility in Solanaceae, Plantaginaceae, and Maloideae of Rosaceae are non-self recognition systems\textsuperscript{3}. In these families, pistil-\textit{S} is a secreted ribonuclease termed \textit{S-RNase}, which exerts cytotoxic effects that inhibit the elongation of self-pollen tubes by degrading RNA\textsuperscript{1-4}; consequently, the SI system in these families is referred to as \textit{S-RNase–based SI}. The pollen-\textit{S} is a set of F-box protein(s), termed \textit{S}-locus F-box (SLF\textsuperscript{1-4}, also called \textit{S}-haplotype-specific F-box, SFB\textsuperscript{6}, or \textit{S}-haplotype-specific F-box brothers, SFBB\textsuperscript{7} in Rosaceae), and function as a component of the SCF (Skp1–Cullin1–F-box)–type E3 ubiquitin ligase which generally mediates ubiquitination of target proteins for degradation by the 26S proteasome\textsuperscript{8}. Previously, we proposed that \textit{S-RNase–based SI} in Solanaceae is a collaborative non-self recognition system, in which the product of each SLF interacts with a subset of non-self \textit{S-RNases}, and the products of multiple SLF types are required for the entire suite of non-self \textit{S-RNases} to be collectively recognized and detoxified\textsuperscript{9}.

In contrast to the co-evolutionary relationships observed in the specificity determinants in Brassicaceae and Papaveraceae, \textit{S-RNase} and SLFs derived from Solanaceae and Maloideae exhibited no corresponding allele phylogenies. One possible explanation for this observation is that \textit{S-RNases} and SLFs each proliferate by different mechanisms, giving them the appearance of different evolutionary histories despite tight linkage and co-inheritance as a single haplotype. Increasing the repertoire of SLF genes that constitute \textit{pollen-S} would be advantageous, as this would increase the number of potential mating partners by allowing pollen to recognize and detoxify more non-self \textit{S-RNases}, whereas an increase in diversity of the \textit{S-RNase} genes would have the opposite effect by allowing new \textit{S-RNases} to escape detoxification by the existing
repertoire of SLF proteins. This pattern more resembles disease recognition and pathogen evasion models than other SI systems. Such factors may have caused the differences in the evolutionary diversification of these genes, but the underlying details remain unclear. In order to characterize the evolutionary history of the entire non-self recognition locus, we conducted a large-scale identification of SLFs from many S-haplotypes.

RESULTS

Petunia pollen-S consists of approximately 18 SLF types

We identified SLF genes from eight SI haplotypes (S5, S7, S9, S10, S11, S17, S19 and S22) and two self-compatible (SC) haplotypes (Sm and S0m; see Online Methods) using a combination of next-generation sequencing (NGS) and PCR techniques. Initially, we constructed expressed sequence tag (EST) libraries from male reproductive organs of lines homozygous for each S-haplotype except S10, S22 and Sm, and then identified SLF-related sequences from these EST libraries. Next, we conducted RT-PCR and RACE-PCR to fill gaps and obtain full-length sequences. We then cloned whole coding sequences of novel candidate SLFs by genomic PCR to confirm the absence of assembly errors in all identified (by Sanger sequencing). PCR reactions were applied to all lines including those homozygous for S10, S22 and Sm-haplotypes to identify additional and undetected SLFs through NGS. The expression of all identified SLFs in anther was confirmed by semi-quantitative RT-PCR. Finally, we identified 16 (in S11-haplotype) to 20 (in S19-haplotype) SLF-related sequences per haplotype, for a total of 168 sequences (180 sequences including 12 pseudogenes were listed in Supplementary Table 1). Based on their phylogenetic relationships, we classified them into 18 types (named Type 1 to 18 SLFs). When the sequences from more than three SI S-haplotypes were collapsed into one clade, we classified them as a novel type. Ungrouped sequences that belonged to none of these 18 types were tentatively named FBXs (Fig. 1, Supplementary Fig. 1 and Supplementary Table 1).

Previous analyses demonstrated genetic linkage among the already-known alleles of type 1–6 SLFs and their cognate S-RNases. We analyzed the linkage of newly isolated SLFs using gene-specific primer pairs (Supplementary Table 2). Examination of 48 plants with segregating S-haplotype revealed no recombination between SLFs and their
cognate S-RNases (Supplementary Fig. 2). We also confirmed male reproductive organ–specific expression profiles of newly isolated SLFs and FBXs using RT-PCR (Supplementary Fig. 3). These results indicated that we have identified strong candidates for novel pollen-S components.

**Diversity and deletion of SLFs predict target S-RNase**

Variations among allelic sequences within each type of SLFs can be classified into two types of polymorphisms: copy number variation and amino-acid sequence polymorphism. As for copy-number variations, 0–2 genes of each type of SLF were identified in each S-haplotype: for example, no type-9 SLF was identified in the S₁₀-haplotype, whereas two copies of type-1 SLFs were detected in the S₇-, S₁₇-, and S₁₀-haplotypes (Fig. 2). Regarding amino-acid sequence polymorphism, we observed alleles with high sequence conservation as well as those with relatively moderate sequence conservation. For example, in type 3, seven alleles have high sequence conservation (99.4–87.4% identities), while two alleles (S₇-SLF₃, S₁₁-SLF₃B) have moderate sequence conservation (76.5–72.0%) (Fig. 2, Supplementary Table 3).

Assuming the collaborative non-self recognition model⁹, a functional Sₓ-haplotype must not encode a SLF that recognizes and detoxifies its own Sₓ-RNase. This can be achieved by having either a diverged or deleted allele of the SLF type that recognizes the Sₓ-RNase. This logic predicts that, if the Sₓ-haplotype encodes no highly conserved allele of type-n SLF (SLFn), the conserved SLFn would recognize the “non-self” Sₓ-RNase.

As for type-3 SLFs, S₇-haplotype has only one relatively diverged allele (S₇-SLF₃) (Fig. 2a). S₁₁-haplotype also has such relatively diverged allele (S₁₁-SLF₃B), but it also has conserved one (S₁₁-SLF₃). Therefore, the model predicts that S₇-RNase is the target of the conserved SLF₃. Indeed, our previous transgenic experiments showed that S₁₁-SLF₃ targeted S₇-RNase⁹. In our current study, we further verified interactions by a similar transgenic approach showing that another conserved SLF₃, S₅-SLF₃, recognizes S₇-RNase (for details of transgenic experiments to assay S-RNase–SLF interaction in vivo, see Fig. 3, Supplementary Fig. 5, Supplementary Tables 4 and 5). Furthermore, we confirmed that S₇-SLF₃ and S₁₁-SLF₃B did not recognize S₇-RNase (Supplementary Fig. 5, Supplementary Table 4).

As the case of deletion, we focused on the type-9 SLF clade, where S₁₁₀ and
$S_{19}$-SLF9 was absent (Fig. 2b). The model predicts that $S_{10}$- and $S_{19}$-RNase could be the target of conserved type-9 SLFs. We tested the $S_{19}$-RNase and found that it was indeed targeted by two alleles of type-9 SLFs that we tested (S$_9$-SFL9A and S$_{11}$-SLF9; Supplementary Fig. 6, Supplementary Tables 4 and 5). Additionally, we found that $S_9$, $S_5$, and $S_{10}$-haplotypes lack type-2, type-14, and type-6 SLFs, respectively. Among these, predicted interaction between conserved type-2 SLFs and S$_9$-RNase consisted with our previous results.

We should emphasize that the predictive method does not exclude the possibility that conserved SLF alleles, can act on additional target S-RNases. In the type-1 SLF clade, six out of eight $S$-haplotypes had a highly conserved SLF allele, whereas $S_{17}$- and $S_{22}$-haplotypes had only relatively diverged SLF1s (Fig. 2c). The model predicts that both $S_{17}$-RNase and $S_{22}$-RNase are the targets of conserved SLF1. We previously tested four conserved SLF1 alleles ($S_5$, $S_7$, $S_9$, $S_{11}$) and showed that all of them targeted $S_{17}$-RNase. Our new experiment confirmed that $S_{22}$-RNase is also the target of a conserved SLF1, $S_7$-SLF1 (Supplementary Fig. 7, Supplementary Tables 4 and 5). In addition, our previous experiments showed that two of the conserved SLF1s ($S_5$ and $S_7$) targeted $S_9$-RNase, whereas two others ($S_9$ and $S_{11}$) did not, implying that an evolutionary change in specificity could also occur with very limited amino acid substitutions. Such a change is consistent with maintenance of normal SI function in $S_9$-haplotype.

Collectively, we found the following pattern: $S_x$-RNase is a target of SLFn if the $S_x$-allele of SLFn is diverged or deleted. This predictive approach is very useful for identifying the target S-RNase(s) of each type of SLF. Actually, among eight SI $S$-haplotypes analyzed in this study, we could predict the responsible SLF types for seven S-RNases, and five ($S_7$, $S_9$, $S_{17}$, $S_{19}$, and $S_{22}$-RNases) among them are actually shown to interact with the predicted SLFs with experimental evidence. Because there are more than 40 $S$-haplotypes in Petunia, it is not surprising that a conserved SLF was found in all of the eight surveyed $S$-haplotypes in the majority of the SLF types (types 4, 5, 10, 11, 13 and 16, Supplementary Fig. 1), and are likely to target S-RNases of other unsurveyed $S$-haplotypes. In comparisons of SLF sequences belonging to the same types, most of the allelic sequences were shown to be highly conserved with identities higher than 90% (Supplementary Table 3). The results of the in vivo assay described above suggested that most of these highly conserved SLFs function as pollen-$S$. This is in stark
contrast to other self-recognition SI systems, such as *Brassica* and *Papaver*, where self/non-self discrimination depends on S-haplotype-specific interactions between highly diverged pollen-S and pistil-S1,2.

**Solanum** S-loci contain orthologous SLF-like paralogs

It is not obvious how SLFs and S-RNases came to constitute a genetic unit at a single S-locus during evolution of the S-RNase–based-SI system considering the much lower diversity among SLFs relative to S-RNases11. We conducted phylogenetic analysis of SLFs and S-RNases including those of other genera of Solanaceae by first exploring SLF-orthologs in the whole-genome databases of tomato (*Solanum lycopersicum*) and potato (*Solanum tuberosum*)12,13. We identified 37 and 66 SLF-like F-box sequences in these two species (Online Methods and Supplementary Table 6). These candidates include both SLF and other SLF-like sequences. Phylogenetic trees including sequences we identified as well as published pollen-S–related sequences in other S-RNase–based SI species (Fig. 4a, Supplementary Fig. 8 and Supplementary Table 7) show the Petunia SLFs cluster into a single monophyletic group together with 13 genes from tomato and 14 from potato (this subclade is referred as the Solanaceae SLF clade; Fig. 4a, Supplementary Fig 8). All of the tomato and potato genes belonging to the Solanaceae SLF clade are specifically located within the repeat-rich, subcentromeric regions in chromosome 1 of the assembled *S. lycopersicum* and *S. tuberosum* genome, consistent with the genetically mapped locations of the S-loci in these species (Fig. 4c)14,15. These Solanaceae SLFs are distributed within 17.9 Mb in tomato and 14.6 Mb in potato, suggesting that the S-loci of the Solanaceae are very large, about two to three orders of magnitude larger than the Brassicaceae S-locus (28–110 kb)16.  

Allelic SLFs are much younger than S-RNase alleles and than SLF types

The SLF types in *Petunia* and other SLFs derived from different genera in Solanaceae are distributed throughout the Solanaceae SLF clade (Fig. 4a and Supplementary Fig. 8c), suggesting that major diversification of the types predated separation of genera in Solanaceae. This is consistent with the extensive trans-specific polymorphism found in Solanaceae S-RNases17,18 (Fig. 4b, Supplementary Fig. 9 and Supplementary Table 8). However, there are small subclusters specifically derived from certain genera, *e.g.*, the type-3/-13, type-4/-12 and type-9/-10 clusters, suggesting that generation of new SLF
types might have continued after separation of genera. The branch depth of S-RNases and SLFs suggest that the timing of proliferation of SLF types rather than individual SLF alleles (Figs. 1a and 3b and Supplementary Fig. 8b) is similar to that of the S-RNase alleles (Figs. 1b and 3b and Supplementary Fig. 9b). Allelic SLFs belonging to each type diversify only at terminal branches of the tree and there is no pair of closely related SLF sequences between different genera of Solanaceae, indicating diversification of allelic SLFs within each type followed the divergence of genera. These results suggest that each genus has a similar number of SLF types but there is no one-to-one correspondence of SLF types among these genera possibly due to evolutionary turnover of SLF types.

Estimates of synonymous and non-synonymous substitution rates (\(K_s\) and \(K_a\), respectively) between SLFs and S-RNases showed that inter-allelic \(K_a\) and \(K_s\) values of each SLF type (\(K_a = 0.000–0.090; K_s = 0.001–0.303\)) were much lower than the values for the S-RNases (\(K_a = 0.400; K_s = 0.850\)) in Petunia (Supplementary Table 9). However, intra-haplotypic \(K_a\) and \(K_s\) values of SLFs exhibited ranges similar to those of the S-RNases (\(K_a = 0.321–0.349; K_s = 0.747–0.762\)). These values were similar in Solanum. These results indicate that alleles of each SLF type are much younger than the S-RNase alleles and than the SLF types.

**Genetic exchanges of SLFs have occurred repeatedly**

Our findings seem to conflict with completely suppressed recombination among SLFs and S-RNases thought necessary to maintain all SI systems. While it is clear that linkage at the S-locus is generally necessary to maintain individual haplotypes over large genomic regions (e.g. 15 Mb), we suspect that sharing of SLFs among S-haplotypes through genetic exchange has occurred repeatedly, at least until relatively recently. Supporting this speculation, some sets of genes share complete identity among several alleles; e.g. the \(S_{7r}\) and \(S_{19r}\)-alleles of type-1 SLFs share completely identical nucleotide sequences whereas the alleles of the corresponding S-RNases are quite different (47.5% amino-acid identity; Supplementary Table 3). These findings support genetic exchange among SLFs and the extremely low level of polymorphism among SLF alleles cannot be explained solely by purifying selection on amino-acid replacement as both \(K_s\) and \(K_a\) are low.

In order to detect a statistical signal of genetic exchange among SLFs, we used
LDhat\textsuperscript{19} and GENECONV\textsuperscript{20}. Genetic exchange was estimated on alignments that contained \textit{SLF}s from self-compatible (SC) mutant haplotypes and again on alignments with these sequences removed. Both approaches detected genetic exchange among alleles in type-3, -9 and -10 \textit{SLF}s (Supplementary Tables 10 and 11). When \textit{SLF}s from SC haplotypes were included, we found many more cases of significant pairwise exchanges between \textit{SLF} alleles, most strongly within type-9 \textit{SLF}s, as well as among type-9 and -10 \textit{SLF}s, which are closely related sister groups (Supplementary Tables 11). Several \textit{SLF} types and pairs of \textit{SLF}s within types exhibit significant influences from genetic exchange. In total we found significant exchange among 36 pairs of \textit{SLF}s. This is a conservative estimate of genetic exchange at the \textit{S}-locus because the described approach focused on recombination breakpoints within the coding sequence of a particular \textit{SLF} type but cannot detect those in the intergenic regions that would result in the exchange of entire \textit{SLF}(s). Overall, these results indicate that genetic exchange might play a role in conservation of \textit{SLF} function.

\textbf{Gene conversion contributed to evolution of SC haplotypes}

In the above analysis, we detected gene conversion in self-compatible \textit{S}\textit{0m} and \textit{S}\textit{m}-haplotypes more prevalently than self-incompatible haplotypes (Supplementary Table 11). To investigate the relationship between breakdown of \textit{SI} and recombination among \textit{SLF} genes, we compared \textit{SLF}s between the pollen-side SC haplotype and its ancestral SI haplotype.

\textit{S}\textit{C2}-haplotype is a pollen-side SC mutant of the \textit{S}\textit{17}-haplotype derived from an SC/SI-mixed natural population of \textit{Petunia axillaris}\textsuperscript{21} (Supplementary Fig. 10a). Our data showed that the \textit{S}\textit{C2}-haplotype shares a \textit{S-RNase} and \textit{SLF}s with the \textit{S}\textit{17}-haplotype, but also contains an additional \textit{SLF1} (\textit{S}\textit{C2}-\textit{SLF1C}) identical to the “conserved”-type \textit{S7}/\textit{S19}-\textit{SLF1} (Supplementary Figs. 11a, b), suggesting that this duplication should be the reason for the breakdown of \textit{SI}. In addition to \textit{S}\textit{C2}-haplotype, we newly identified an additional pollen-side SC mutant of \textit{S}\textit{22}-haplotype, designated \textit{S}\textit{22m}-haplotype (Supplementary Fig. 10b). \textit{S}\textit{22m}-haplotype shares a \textit{S-RNase} with \textit{S}\textit{22}-haplotype, but also contains an additional \textit{SLF1} (\textit{S}\textit{22m}-\textit{SLF1B}) yet again identical to \textit{S7}/\textit{S19}/\textit{S}\textit{C2}-\textit{SLF1} (Supplementary Figs. 11a, b). Our previous transgenic experiments\textsuperscript{9} and newly performed ones (Fig. 2 and Supplementary Fig. 5) indicated that the presence of this common additional \textit{SLF1} was sufficient for the breakdown of \textit{SI} in \textit{S}\textit{C2}- and \textit{S}\textit{22m}-pollen.
This suggests that the genetic exchange was responsible for the evolution of self-compatibility (see discussion). The existence of a shared SLF1 among four different haplotypes further represents evidence of recent inter-haplotypic SLF-gene exchange.

Interestingly, in addition to type-1 SLF, we found that these four S-haplotypes also share one common type-8 SLF (S7/1/S19/1/S22m/SLF8) (Supplementary Fig. 11c and 12). This allowed us to compare the phylogenies between type-1 and type-8 SLFs, and we found that these types showed similar topologies (Supplementary Fig. 13). This result suggests that the SLF1–SLF8 linkage unit might have been transferred among different S-haplotypes over evolutionary time. We conclude that genetic exchange among S-haplotypes and some linkage units has occurred repeatedly and contributes to the evolution of both SC and SI S-haplotypes.

**Mathematical models suggest that 16-20 SLFs would be adequate for non-self recognition**

The number of SLF types is much less than that of predicted S-RNase alleles (40 or more)\(^4\), thus one-to-one interactions between a SLF type and a S-RNase allele is not possible. Rather, some SLF types should interact with multiple S-RNase allelic variants, while some S-RNases can be recognized by multiple SLF types. To estimate whether the 16–20 SLF genes we identified here would be adequate for non-self recognition, we compiled the empirical data of the SLF and S-RNase interactions\(^9,22,23\) including data presented in this study (Supplementary Table 12) and developed simple models. Among the 129 tested combinations of SLF and S-RNase, 24 showed positive interactions, thus a SLF would recognize 18.6% S-RNases on average. If we pose a simple assumption that target S-RNase repertoires for each SLF are independent and there are \(n\) SLFs, the proportion of \((1-0.186)^n\) S-RNases cannot be recognized. Thus, \(n\) SLFs can recognize \((1-(1-0.186)^n)\) proportion of S-RNase alleles. With \(n = 16–20\), the probability of recognition approaches saturation (Fig. 5a), which would be sufficient for this self-incompatible system to work. Next we relaxed the assumption so that each SLF can recognize a different proportion of target S-RNases. Because experimental data are already available from 50% (9/18) of the SLF types (see Online Methods), we used Monte-Carlo simulation with bootstrapping for these data. Again, we found similarly that the interactions become saturated by 16–20 SLF types (Fig. 5b). The models suggest that the previously identified eight SLF types may not constitute an efficient
non-self recognition system, but that 16–20 SLFs on each haplotype would be adequate to recognize the vast majority of S-RNase targets if not all, which is estimated to be about 40 alleles in Petunia. The results also support the validity of the number of SLFs identified in Petunia. We note that the recognition rate from these models may be considered as minimum estimates, because different SLF types may tend to recognize different S-RNases since overlapping targets may not be favored by selection. The upper limit of the number of SLF types should be constrained by factors such as the strength of inbreeding depression and the proportion of self-pollen deposited on a stigma in natural population, birth-and-death rate of SLF types and effective population size. These simple models suggest that we have identified the majority of the genetic components of this non-self recognition system.

DISCUSSION
Co-evolving genes for self/non-self recognition systems in are notoriously difficult to study because they typically involve inter-organism dynamics such as disease resistance and virulence and either involve quantitative phenotypes (with epistatic or pleiotropic effects) or are lethal to one or both organisms. Plant pathogen recognition is governed by hundreds of duplicated R-genes that detect numerous pathogens. Autoimmune disease phenotypes by self-recognition may be observed as Dobzhansky-Muller incompatibility in hybrids, because an R-protein from a parent often recognizes self-protein derived from another parent. Solanaceae SI genetic systems provide a unique opportunity to study and model the evolutionary dynamics of co-evolving genes that largely resemble disease recognition and detoxification mechanisms, where duplicated SLF proteins need to recognize diverse non-self S-RNases but not the self S-RNase.

Our exhaustive search for new SLF genes indicated that 16 to 20 SLFs were present in each of ten haplotypes of Petunia. Although only eight SLF types were reported previously, phylogenetic analysis showed that these could be classified into 18 major SLF types with occasional allelic duplication or deletion in each type. Our mathematical models (Fig. 5) suggests that 16–20 SLFs on each haplotype would be adequate to recognize the vast majority of about 40 S-RNase targets.

Our phylogenetic observations suggested that the origins of the SLF types are as ancient as the origin of alleles of S-RNases, whereas alleles of each type of SLFs are
much younger than the S-RNases. This is in clear contrast to self-recognition SI systems, in which the pollen-S and pistil-S show distinct co-evolution in nucleotide substitutions\(^5\). The generation of new SLF types via inter-haplotype genetic exchange provides an explanation for the phylogenetic pattern of SLFs in the collaborative non-self recognition system. Three possible consequences are postulated in Fig. 6a.

In the first scenario, a new SLF acquired by genetic exchange could recognize more effectively wider range of ‘non-self’ S-RNases. In such a case, the SLF acquisition could confer an advantage for outcrossing, and would be fixed in the population (evolution from \(S_{1a^-}\) to \(S_1\)-haplotypes in Fig. 6a). This SLF would spread rapidly over multiple \(S\)-haplotypes in the population and form a new SLF type with short branches in the phylogenetic tree (\(S_1\)-SLF\(n\) in Fig. 6b).

In the second scenario, newly acquired SLF(s) could recognize an endogenous ‘self’ \(S\)-RNase as a specific target, inducing breakdown of SI (evolution from \(S_{2a^-}\) to \(S_{2m^-}\)-haplotypes in Fig. 6a). Most such SC \(S\)-haplotypes should be lost by short term disadvantage (inbreeding depression), and show a short branch in a phylogenetic tree (\(S_{2m^-}\)-SLF\(n\) in Fig. 6b). Occasionally, self-compatible alleles are found in natural SC/SI-mixed populations of Petunia\(^29\), and could be fixed by selective forces such as mate limitation or automatic transmission advantage and by escaping rejection from all \(S\)-haplotypes in outcrossing\(^24,26,30-33\). Although it has long hypothesized that recombination and/or gene conversion between different \(S\)-alleles could induce self-compatibility, no clear example has been found. In natural and domesticated populations of Brassicaceae possessing a one-to-one self-recognition system, loss-of-function mutations of pollen-S, pistil-S or modifiers, rather than recombination, have been shown to be responsible for self-compatibility\(^34-36\). Here we provide experimental evidence that the acquisition of a \(SLF1\) by gene conversion has lead to the evolution of two SC haplotypes (\(S_{C2}\) and \(S_{22m}\)). The self-compatibility in pollen-S is also consistent with the theory suggesting that mutations in male components are more likely to spread in natural populations\(^26,33\).

In the third scenario, some SC haplotypes generated in the second scenario could restore SI phenotype by having mutations in the responsible SLF or its target S-RNase (evolution from \(S_{3a^-}\) to \(S_{7^-}\) or \(S_{7^-}\)-haplotypes in Fig. 6a). Accumulation of mutations in the acquired SLF could regenerate an SI \(S\)-haplotype. As we used in the prediction of target S-RNases, a particular SI \(S\)-haplotype must not have the SLF that recognizes its
own S-RNase, thus the \( SLF \) allele should be lost or diverged, exhibiting a relatively long branch in phylogenetic tree (\( S_{3-RNase} \) in Fig. 6b). Mutations in the target \( S-RNase \) could also regenerate a novel SI S-haplotype. Such an S-RNase mutation that escapes from the recognition by an SLF should have a risk leading to the female sterility. In the collaborative non-self recognition system, the redundant S-RNase recognition by multiple SLFs could reduce the risk and might support the evolution of new S-haplotypes (\( S_{4-RNase} \) in Fig. 6b).

In Solanaceae, the \( S \)-locus is located in a subcentromeric, repeat-rich, and low gene-density region, in which recombination is strongly suppressed\(^{37,38}\). Until recently, it has remained unclear how \( SLFs \) evolved despite being located in such an inactive genomic region. However, recently crossover recombination (\( i.e., \) homologous recombination) was shown to be fully suppressed at centromeres, while non-crossover recombination (\( i.e., \) gene-conversion) is not\(^{39}\), therefore, genetic exchange among \( SLFs \) could also be possible. Other studies suggested that intragenic recombination has also contributed to the diversity of \( S-RNases \)\(^{40,41}\), although the frequency seemed to be rare. Alternatively, RNA-mediated genetic exchange (retrotransposition)\(^{42}\) may have also contributed because the \( Petunia \) \( S \)-locus is rich in retrotransposons\(^{43}\) and no \( SLF \) gene thus far identified contains introns. However, our finding of linkage unit \( SLF1-SLF8 \) in several S-haplotypes suggests the former event is more likely based on the distribution of this linkage unit.

Our comprehensive identification of \( SLFs \) will provide useful data for characterizing the molecular evolution of S-RNase-based SI, a topic that has been debated for many years. In contrast to collaborative non-self recognition model, traditional models of SI assumed self-recognition by a single gene product\(^{26,44,45}\). The findings described here indicate that \( pollen-S \) and \( pistil-S \) have undergone different and complex patterns of molecular evolution and apparently depend on the genetic exchange of \( SLFs \) as well as sequential accumulation of base substitution. In the future, it should be possible to explain the molecular evolution of S-specificity in populations using simulations based on the molecular model we proposed here.

**METHODS**

**Plant materials.** We used lines of \( S_{5}, S_{7}, S_{9}, S_{10}, S_{11}, S_{22}, S_{0m}, S_{m}, \) and
$S_{22m}$-haplotypes from *Petunia hybrida* and the lines of $S_{17}$-, $S_{19}$- and $S_{c2}$-haplotypes from *P. axillaris* (for details, see Supplementary information).

**EST database for male-reproductive organ of Petunia.** For preparation of cDNA libraries and NGS, see Supplementary information.

Based on the results of next-generation sequencing, we constructed EST databases using GENETYX-MAC (ver. 16.0.6).

**Identification of novel SLF genes.** Local BLAST (NCBI BLAST ver. 2. 2. 24) search against EST databases was carried out with GENETYX-MAC, using *Petunia* type 1–6 SLFs $^9$ as queries. We extracted hits with evaluation ($E$) values less than $10^{-20}$, and assembled them using ATSQ (ver. 5.1.3, GENETYX) with the following parameters: matching percentage, 50%; minimum matching number, 10; and group capacity, 5000. All assembled sequences were manually checked. To isolate additional SLFs from S-haplotypes unanalyzed by NGS and to confirm the full length coding sequences, we performed RACE and genomic PCR. Primers are listed in supplementary table 2 and reaction conditions are described previously $^9$. As for newly isolated SLFs, to confirm the absence of assembly errors, entire coding regions were amplified by genomic PCRs and cloned into the pGEM-T easy vector (Promega). At least eight independent clones for each SLF fragment were sequenced on both strands by Sanger's method using an ABI 3130xl Genetic Analyzer (Applied Biosystems).

**Identification of S-related genes in Solanum.** BLAST-searches for SLFs and S-RNases in potato and tomato were carried out at Spud DB $^{12}$ and at Sol Genomic Network $^{13}$ using databases ‘PGSC DM1-3 pseudomolecules (v4.03)’ and ‘ITAG annotation Release 2.4 predicted CDS (SL2.50)’, respectively. *Petunia SLFs* and *S-RNase* from $S_7$-haplotype were used as queries. Hits with $E$-values < $10^{-20}$ were extracted as candidates.

Preliminary analysis on tomato identified only eight SLF-like sequences, including four pseudogenes with premature stop codons, far fewer compared with those in potato and *Petunia*. A supposed reason was that the line used for whole genome analysis of potato is a SI line, *S. tuberosum* Phureja DM $^{12,46}$, whereas that of tomato is a SC line, *S. lycopersicum* cv Heinz 1706 (Ref. 13). Considering the possibility that the most SLFs
have broken down in tomato, we BLAST-searched against scaffold sequences on chromosome 1, and identified additional five unannotated \textit{SLF}-like pseudogenes, named Solic01\_pseudo1–5.

To assess the authenticity of the extracted sequences as \textit{SLF}-related F-box genes, we evaluate the motif composition of deduced products by using Simple Modular Architecture Research Tool (SMART)\textsuperscript{47}. Sequences judged to have neither F-box nor FBA motif were eliminated as false positives. Identified \textit{SLF}-related and \textit{S-RNase}-related genes are listed in Supplementary Table 6. As for potato gene name, ‘PGSC0003DMG’ is omitted from each potato gene ID to simplify, and ‘STchx’ is attached to indicate the gene location on \textit{S. tuberosum} chromosome x in this paper (see Fig. 4).

\textbf{Phylogenetic analysis.} Deduced amino-acid sequences or coding DNA sequences of \textit{F-box} and \textit{RNase} genes were aligned with the ClustalW algorithm, using MEGA5 (ver. 2.2, Ref. 48). For pseudogenes, frame shifts or premature stops were removed manually. Based on the alignments, phylogenetic trees were constructed by the neighbor-joining method using MEGA5.

Codon-by-codon alignments of coding sequences of each type of \textit{SLF}s or \textit{S-RNase}s were constructed by using MEGA5. Based on these alignments, synonymous and non-synonymous substitution rates ($K_s$ and $K_a$) were calculated using DnaSP (ver. 5.10.1, Ref. 49), and recombination and gene conversion were evaluated by LDhat (ver. 2.1, Ref. 19) and GENECONV (ver. 1.81, Ref. 20).

\textbf{Estimation of S-RNase proportion recognized by SLFs.} We applied Bernoulli process to the estimation of the proportion of S-RNases recognized by \textit{n} SLF types, $P_S(n)$, assuming that each SLF recognizes S-RNase allelic variant independently with the same probability $P_R$ as expressed by the following equation (1):

\begin{equation}
P_S(n) = 1 - (1 - P_R)^n
\end{equation}

where $P_R$ is the overall recognition rate and \textit{n} is the number of SLF types. The overall recognition rate $P_R$ is defined by the following equation (2):

\begin{equation}
P_R = \frac{N_p}{N_T}
\end{equation}
where $N_P$ is the number of positive interactions between S-RNase and SLF allelic variants and $N_T$ is the total number of tested interactions. We excluded the positive interaction S_{22}-RNase and S_{7}-SLF1, because the positive interaction was predicted and S_{22}-RNase was experimentally tested only with S_{7}-SLF1, thus this could bias the estimation of the overall recognition rate. Remaining dataset would be considered to represent random samples of the overall recognition rate between S-RNase and SLF allelic variants. Note that $N_P = 24$, $N_T = 129$ are actually assigned in this study, so that $P_R$ is nearly equals to 0.186 (Supplementary Table 13a). The confidential interval $C_I$ is calculated as follows:

$$C_I = t \frac{u}{\sqrt{N_T}}$$

where $u = \sqrt{\sum_{i=1}^{N_T} (p_i - P_R)^2 / (N_T - 1)}$

and $P_i \in \{0, 1\}$, 0 and 1 indicates the negative and positive interaction, respectively, in the combination of each S-RNase and SLF allelic variant interaction. $t = 1.97867$ is in this calculation assigned as the $t$ value of the Student's $t$ inverse cumulative distribution function at the 95th percentile of both sides of the $t$-distribution for the degrees of freedom $N_T-1$.

We also conducted Monte-Carlo simulation in order to consider the difference of recognition rates among SLF types. For details, see Supplementary information.

**Accession codes.** DNA sequences of newly isolated SLFs and S-RNases have been deposited in the DNA Data Bank of Japan (DDBJ) under accessions AB932964 to AB933144. See Supplementary Table 1 and 8 for correspondence between gene names and accession nos.

**Note added in proof**

During revision of this article, Williams et al. reported the independent identification of 17 SLF types in two S-haplotypes of *Petunia inflata*. Their data supported our observation indicating that 16–20 SLFs are sufficient for non-self recognition system in
Petunia SI.

Reference


Correspondence and requests for materials should be addressed to K.K.S. (kentaro.shimizu@ieu.uzh.ch) or S.T. (takayama@bs.naist.jp).

ACKNOWLEDGMENTS
We thank H. Takatsuji at National Institute of Agrobiological Sciences for *P. hybrida* cv. Mitchell and W138; S. Saha at Cornell University for directions on *Solanum* genomic information; M. Iwano, H. Shiba, H. Shimosato-Asano, Y. Wada, K. Murase, M. Kakita, E. Miura, H. Kakui, T. Tsuchimatsu and M. Robinson for discussion or technical advice; and F. Kodama, M. Okamura, E. Mori, Y. Goto, H. Kikuchi, and Functional Genomics Center Zurich for technical assistance. We thank T.-h. Kao at Pennsylvania State University for adapting a unified nomenclature for the same new SLF genes before publication. This work was supported in part by a Grant-in-Aid for Scientific Research on Innovative Areas (23113002) and by Grants-in-Aid for Scientific Research (21248014, 25252021) from the Ministry of Education, Culture, Sports, Science and Technology of Japan awarded to S.T. This work was also supported in part by Swiss National Science Foundation (31003A_140917) and by URPP Evolution in Action of University of Zurich awarded to K.K.S., Marie-Heim Vögtlin grant awarded to R.S.-I. and by Plant Fellows awarded to T.P.

AUTHOR CONTRIBUTIONS

COMPLETING FINANCIAL INTERESTS
The authors declare no competing financial interests.
FIGURE LEGENDS

**Figure 1** Phylogenies of SLFs and S-RNases from *Petunia*. (a, b) Neighbor-joining phylogenetic trees of deduced amino-acid sequences of SLF (a) and S-RNase genes (b) were created with MEGA5 (Ref. 48). Both trees are shown in the same scale; the bar for each tree indicates the number of amino-acid substitutions per site. *PpS4-Fbox0* (a) or *PpS4-RNase* (b) was used as an outgroup. Numbers on the branches indicate bootstrap values > 50% with 1,000 trials. To simplify, subgroups (allelic SLFs) within single SLF type are showed in a compressed representation (black triangles). The full tree of SLFs appears in Supplementary Fig. 1.

**Figure 2** Relationships between phylogenies of SLFs and SLF/S-RNase interactions. (a–c) Neighbor-joining phylogenetic trees of deduced amino acid sequences of type-3 (a), type-9 (b) and type-1 (c) SLFs. For creation and explanation of phylogenetic trees, see legend of Fig. 1. Two-headed arrows indicate positive interactions between SLFs and S-RNases, which lead to pollen acceptance, demonstrated by *in vivo* assay. Gray dotted lines indicate negative interactions, which don’t lead to pollen acceptance. Red and blue characters indicate relatively diverged SLFs and their cognate S-RNases, which are targeted by conserved SLFs.

**Figure 3** Target S-RNase analysis of *S5*-SLF3. Compatibility was judged by monitoring the pollen tubes (arrowhead)⁹. (a) *S5S9*-heterozygote with *S5–SLF3* retained SI. Similar results were obtained for heterozygotes with *S11–*, *S17–*, and *S19*-haplotypes (Supplementary Fig. 5), suggesting *S5*-SLF3 recognizes none of these S-RNases. (b) Transformants *S5S7/S5–SLF3* (T) exhibited breakdown of SI. (c, d) Reciprocal crosses with *S5S7* (WT) suggested that the pollen lost SI. (e) PCR-genotyping of F₁-progeny from WT × T. (f) Schematic explanation of the results in (e). Among four genotypic pollen from T, only *S7*-pollen with *S5–SLF3* successfully fertilized, suggesting that *S5*-SLF3 detoxifies *S7*-RNase. Bars = 200 µm.

**Figure 4** Phylogenies and pericentromeric localization of Solanaceae SLFs and S-RNases. (a, b) Phylogenies of SLFs (a) and S-RNases (b). Symbols indicate the gene is derived from genus: *Petunia* (pink-circle), *Solanum* (blue-triangle), and *Nicotiana*
(green-square). Allelic SLFs collapsed into each type clades are denoted by black triangles. (c) Chromosome 1 of Solunum. Red arrows, S-locus regions; thick blue lines, pericentromeric regions; red lines, SLFs; blue line, S-RNase; black lines, F-box genes outside of SLF clade; ψ, pseudogenes. Potato S-RNase is localized to chromosome 1, but is not mapped due to a probable assembly error. For gene name abbreviation, see Methods.

Figure 5 Estimation of the proportion of S-RNase recognized by multiple SLF types. (a) The result of a Bernoulli process applied to the estimation, based on the assumption that each SLF type recognizes allelic S-RNases independently with the same probability. Error bars indicate confidential intervals. (b) The result of a Monte-Carlo simulation conducted to consider the difference of proportion of S-RNase recognition among SLF types. Error bars indicate standard deviation. Blue and purple horizontal lines indicate 95% and 100% of S-RNase allelic products, respectively. Dotted vertical lines indicate $n = 16, 18, \text{ and } 20$.

Figure 6 Model for the evolution of SLFs. (a) Recombination events and their expected consequences. Possible evolutionary scenarios for three ancient haplotypes ($S_{1a}$, $S_{2a}$ and $S_{3a}$) are indicated. Ovals and boxes indicate SLFs and S-RNases, respectively. Red ovals indicate genetic-exchanged or integrated SLFn whose product recognizes $S_{2a}$- and $S_{3a}$-RNases but not $S_{1}$-RNase. Lighter colors indicate extinct haplotypes. (b) Predicted phylogenies of allelic SLFn (upper) and S-RNases (lower) of extant SI S-haplotypes ($S_{1}$, $S_{2}$, $S_{3}$ and $S_{4}$) and SC S-haplotype ($S_{2m}$), whose evolutionary scenarios are simulated in (a).
Fig. 1
Fig. 5

Number of SLF types, \( n \)

Proportion of S-RNases recognized by SLF type, \( P(n) \)
Ancient haplotype

Extant haplotype

Genetic exchange or integration

S1a
S2a
S3a

S1a
S1
S2
S2m
S3
S4

S2m-SLFn
S2-SLFn
S3-SLFn
S2-SLFn
S4-SLFn
Sx-SLFm

S1-RNase
S2-RNase
S3-RNase
S4-RNase

S2m-RNase

Fig. 6