RNase-based Self-incompatibility in Cacti

Karolis Ramanauskas[a] **and Boris Igić**[a,1]

[a]University of Illinois at Chicago, 840 W Taylor St, M/C 067, Chicago, IL, 60607, U.S.A.

This manuscript was compiled on February 16, 2021

Approximately one half of all flowering plants express genetically-based physiological mechanisms that prevent self-fertilization. One such mechanism, termed RNase-based self-incompatibility, employs ribonucleases as the pistil-component. Although apparently widespread, it has only been characterized in a handful of scattered families, partly due to the difficulties presented by life history traits of many plants, which complicate genetic research. One such group is the cactus family, where many species are known to express self-incompatibility but whose underlying mechanisms remain unknown. We demonstrate the utility of a candidate-based RNA-seq approach, combined with some unusual features of loci that effect self-incompatibility, which can be used to uncover the genetic basis of this mechanism. Specifically, we assemble transcriptomes from the crab cactus, *Schlumbergera truncata*, and interrogate them for tissue-specific expression of candidate genes, their structural characteristics, correlation with expressed phenotype(s), and phylogenetic placement. In each case, the results are consistent with operation of the RNase-based self-incompatibility mechanism in Cactaceae. The finding yields additional evidence for ancestry of this mechanism in nearly all eudicots and a clear path to better practices for conservation of one of the most charismatic plant families.

Plant evolution | Self-incompatibility | Breeding systems | Molecular evolution | RNA-seq

Flowering plants display an astonishing array of morphological, physiological, and life history traits. Perhaps the most striking feature of this group is their propensity for highly dynamic evolution of breeding systems—suits of traits that affect the frequency of self- and cross-fertilization. Changes in breeding systems have profound consequences for all other traits, because they modify effective population sizes, opportunity for recombination, the distribution of genetic variation in space and time (1, 2). These changes alter the prospects for evolution across the entire genome (3, 4). Unlike animal species, 95% of flowering plant individuals are co-sexual, with hermaphrodite flowers, providing ample opportunity for self-pollination and self-fertilization. And yet, relatively few predominantly self-fertilize.

Nearly one-half of angiosperms instead strictly avoid self-fertilization (5). They most commonly obligately cross-fertilize by deploying a genetic system, broadly termed 'self-incompatibility' (SI), which causes otherwise fertile hermaphrodites to recognize and reject their own pollen. A variety of molecular genetic mechanisms can effect such SI responses. Among the four best-understood mechanisms, three are expressed in relatively small groups, each with a distinct genetic basis, indicating that the basic phenomenon of SI is non-homologous and convergent, with different genetic causes generating dozens independently evolved occurrences across angiosperms (6, 7). These convergent systems enforce SI in Brassicaceae, Papaveraceae, and Primulaceae, each employing unrelated gene families in pistil and pollen (8–10). One mechanism, however, is widespread and possibly homologous across eudicots (6, 11). Thus far, it is known to occur in five families (Solanaceae, Plantaginaceae, Rubiaceae, Rosaceae, and Rutaceae), each of which employ a member of the T2/S-RNase gene family (S-RNases) as the stylar components, and several S-linked F-box-containing proteins (SLFs) as the pollen components of the self-recognition and rejection response (12–16).

The significance of the RNase-based SI (RSI) mechanism is underscored by the fact that it is widely distributed, present in both rosid and asterid clades within eudicots. This group originated well over 100 Mya and contains approximately two-thirds of all flowering plant species (17). Lineages with RSI are also associated with an increased diversification rates (18). Despite the clear evidence that the mechanism is homologous across eudicots, some doubts remain about orthology of individual components (19, 20). Iterations in different families show significant variation in molecular details, most critically the mode of pollen tube rejection—whether it proceeds by self-recognition (*Prunus*) or non-self-recognition (as in all other cases), as well as the role and number of pollen-expressed genes (21). Understanding the distribution of SI and its genetic mechanisms across flowering plants can furnish critical weight of evidence about the nature of trait homology of this system (6, 11, 13, 14) and its influence on diversification rates (18). Synthesis of such knowledge can be used to precisely detect convergence—independent re-evolution of alternative SI mechanisms—help identify when and were it occurred, its frequency, and help infer the factors that cause evolutionary innovation. Finally, a better understanding of SI mechanisms allows manipulations to reduce mate limitation and increase

Significance Statement

Most flowering plant species have co-sexual individuals whose own pollen often lands on their stigmas. And yet, approximately one half are self-incompatible—they cannot self-fertilize. Such individuals express a genetic mechanism that sorts incoming pollen, rejecting all pollen tubes that match a particular genomic region, the S-locus. One mechanism is particularly widespread. It relies on proteins termed ribonucleases (RNases). Here, we show that this mechanism operates among cacti, including the cultivated crab cactus, *Schlumbergera truncata*. Our finding implies that RNase-based self-incompatibility may be far more common than is generally appreciated. Understanding the distribution of this phenomenon is critical for determining the causes and consequences of variation in plant mating patterns. Its characterization is also broadly useful for a variety of applications, including plant breeding and, of particular interest to us, conservation.

K.R. & B.I. designed research; K.R. & B.I. performed research; K.R. & B.I. analyzed data; and K.R. & B.I. wrote the paper.

The authors declare no conflicts of interest.

[1]To whom correspondence should be addressed. E-mail: boris@uic.edu

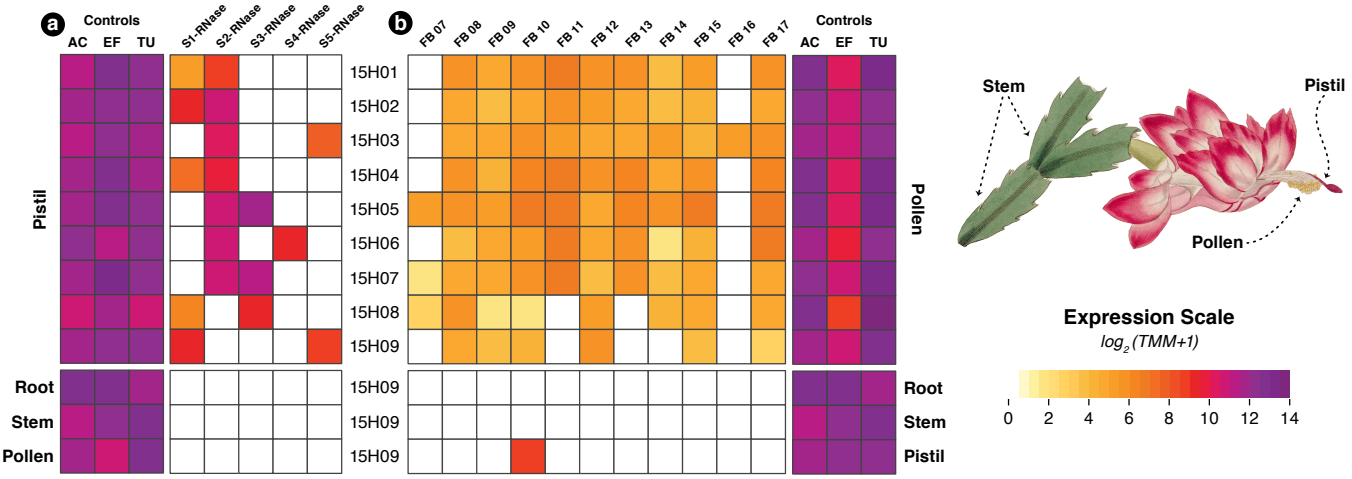

Fig. 1. Summary of candidate gene (S-RNase and SLF) expression data for nine *Schlumbergera truncata* individuals. Following tissue-specific transcriptome assembly, we identified putative S-RNase and SLF genes with secondary sequence features and phylogenetic placement (following 24). Individual plant accessions are shown in rows between the two figure panels. Control genes: AC, actin; EF, elongation factor EF-1 alpha; TU, alpha-tubulin. (a) Pistil-expressed putative S-RNase (S1-S5) follows the expected pattern. Each genotype is heterozygous (expresses two S-RNase alleles). The expression of each predicted allele level is fairly high, in all cases in top-10% among pistil-expressed genes. Their expression is undetectable in root, stem, and pollen. Subsequent crossing experiments confirm that these sequences segregate with the S-phenotypes, and successfully predict crossing ability from predicted genotypes. (b) Pollen-expressed of putative SLF. We expect at least a dozen SLF genes to be expressed in each S-haplotype, genetically linked with one S-RNase allele, in a functional RSI mechanism (with non-self recognition; 21). Each S-haplotype is also expected to express most—if not all—SLF genes, except for the one that recognizes its own linked S-RNase. Their expression is undetectable in root, stem, or pistil, except in the case of FB_10 (recovered in 15H09 pistils; see text for details). Although this set of genes shows a variety of features broadly consistent with SLFs, we lack a genome assembly of the S-locus from any Cactaceae species, and our limited expression data include aggregated pollen grains that express the content of both parental S-haplotypes.

fruit set, with direct applications ranging from plant conservation (22) to agriculture (23).

Uncovering the molecular genetic basis of SI is generally complicated, expensive, and time-consuming, which may explain the relative paucity of data across angiosperms, and its discovery in only a handful of plant families since 1986. In particular, wild tropical species and many neglected or underutilized crops are not amenable to classic forward genetic approaches because of long generation times, inconvenient stature, or difficult plant care. Some unusual features of RSI can, however, be harnessed to overcome the usual obstacles present in uncovering the genetic bases of traits. First, the wide phylogenetic distribution of RSI indicates that T2/S-RNases are clear pistil-part expression candidates in eudicots with unknown basis of SI (24). Second, their expression is known to be highly tissue-specific, narrowly confined to styles within pistils, where S-RNases are highly expressed (13). Likewise, the pollen-expressed gene candidates are highly tissue-specific (21, 25). Third, each individual with a functional RSI system is obligately heterozygous at the S-locus, and negative frequency-dependent selection maintains many S-alleles, because plants with rare S-alleles are compatible with many more potential mates (26). Therefore, any hunt for such genes ought to recover heterozygous individuals with highly divergent alleles. Finally, simple glasshouse crossing experiments can be used to confirm the association of putative S-RNase genotypes with the predicted crossing phenotypes (27). For example, cross-pollination between individuals with matching S-locus genotypes will not result in fruit set. Informed by such features, we may be able to uncover the genetic basis of SI with transcriptome sequencing and candidate genes, coupled with simple experimental design.

Here, we describe how an easily generalized approach yielded the discovery of RSI in cacti (Cactaceae). Specifically, we employ common transcriptome-generating methods to find that pistils of *Schlumbergera truncata* express highly polymorphic and obligately heterozygous S-RNases, which are closely related to S-RNases from other plant families with this SI system. Moreover, the inferred S-locus genotypes predict the crossing phenotypes, as assessed by hand-pollinations, and we find many pollen-expressed genes that strongly resemble known SLFs. With the recent finding of RSI in Rutaceae (16), we add to the weight of evidence that this mechanism is likely more common across core eudicots than generally appreciated. A high proportion of cacti–including many rare, threatened, and endangered species—express SI (22, 28). Our candidate-based RNA-seq approach for discovery of this mechanism is simple, inexpensive, and easily applied. It holds especially great promise for determining the genetic basis of SI in woody perennials, particularly trees and other species with long juvenile phases. Below, we provide evidence that RSI operates in cacti, explore the resulting insights into the evolution of this widespread SI system, and the clear path to discovery of new instances of SI across flowering plants.

Results

Pistil expression of S-RNases. Sequencing of *Schlumbergera truncata* pistil RNA resulted in a median transcriptome assembly of 30,842 genes and 51,414 isoforms (range 16,949–42,717 and 24,392–73,047, respectively; Table S1). Most importantly, using non-invasive sampling, relying on as few as one pistil per individual, we reliably recovered candidate RNases, a total of five distinct sequences that display the salient features of functional S-RNases (24).

The recovered S-RNases candidates are highly polymorphic and obligately heterozygous (Figure 1a). The total coding region sequences range from 639 to 666 bp, similar to other S-RNases. Each contains highly conserved residues, closely resembling the structure of functional S-RNases, including active site histidines present among across the T2/S-RNase protein

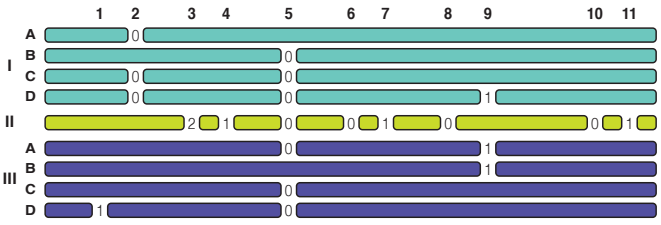

Fig. 2. Intron positions in the known S-RNases from Cactaceae, relative to the other members of the plant T2/S-RNase gene family. This gene family found across land plants is generally divided into three phylogenetic classes, largely congruent with intron presence-absence patterns at 11 intron positions. S-RNases from the cacti display the pattern IIIa. They contain two introns: one at overall position 5 (phase 0; not interrupting a codon), which is common for this gene family, and another at position 9 (phase 1; interrupting a codon after the first position), previously unknown in functional S-RNases, but shared with Class I T2/S-RNases. Other S-RNases characterized to date display the intron patterns IIIc (many families), and those from *Prunus* (Rosaceae) display the pattern IIId (24).

family. One of the highly conserved features of functional S-allele sequences is their strongly basic isoelectric point. The candidates we recovered are each strongly basic (pI range 8.66–9.02), within the range of those expressed in other families with RSI. The gene structure, however, displays an interesting departure. It contains two introns, of which one is unique among S-RNases characterized to date. While the first is shared by nearly all members of this protein family (Figure 2; position 5), a second one is absent among known S-RNases, but common among non-S-RNase proteins in this gene family (position 9). S-RNases from *Prunus* also possess a unique intron (position 1), and it is seems that intron gains and losses, though relatively rare and reasonably phylogenetically informative, are more common than previously thought (6). Finally, the five putative S-RNases are both highly expressed and tissue-specific. Each S-allele we recovered is among the top-10% most highly expressed genes in the pistil, but appears entirely absent in other tissues examined (pollen, petal, stem, and root; Figure 1a).

We additionally examined the pistil transcriptome of a single individual of *Matucana madisoniorum*, as well as the published genome (29) of *Stenocereus thurberi* (`Sthu_v1.3`, GenBank assembly accession: `GCA_002740465.1`). In both individuals, we recovered two putative S-RNases apiece, with expression and/or sequence features that closely match those recovered from *Schlumbergera truncata*. *Stenocereus thurberi* alleles, named S_1 and S_7, are found on short assembled genome contigs `NCQT01146464.1` (2,776 bp) and `NCQT01107209.1` (39,932 bp), respectively. The allele S_7 was found at the 5'–end of the contig and, based on the alignment with the S_1 allele, it is missing ca. 93 bp at the 5'–end of the molecule. As we did not confirm their functional identity in subsequent phenotyping experiments, we continue to refer to the genes we uncovered in *Matucana* and *Stenocereus* as "putative" S-alleles or S-RNases, reflecting a considerably lower collective weight of evidence.

S-RNase sequences are expected to display extreme persistent polymorphism, the result of strong and persistent negative frequency-dependent selection (26, 30). We find that pairwise amino acid sequence identity among five *Schlumbergera truncata* S-RNase alleles ranges from 50.5% to 70.1%. This is well within the range found in more intensely sampled species. For example, *Solanum chilense* (Solanaceae), where amino acid identity ranges between 30%-96% (31). Despite the small sample of alleles, as expected, we find a well-supported pattern

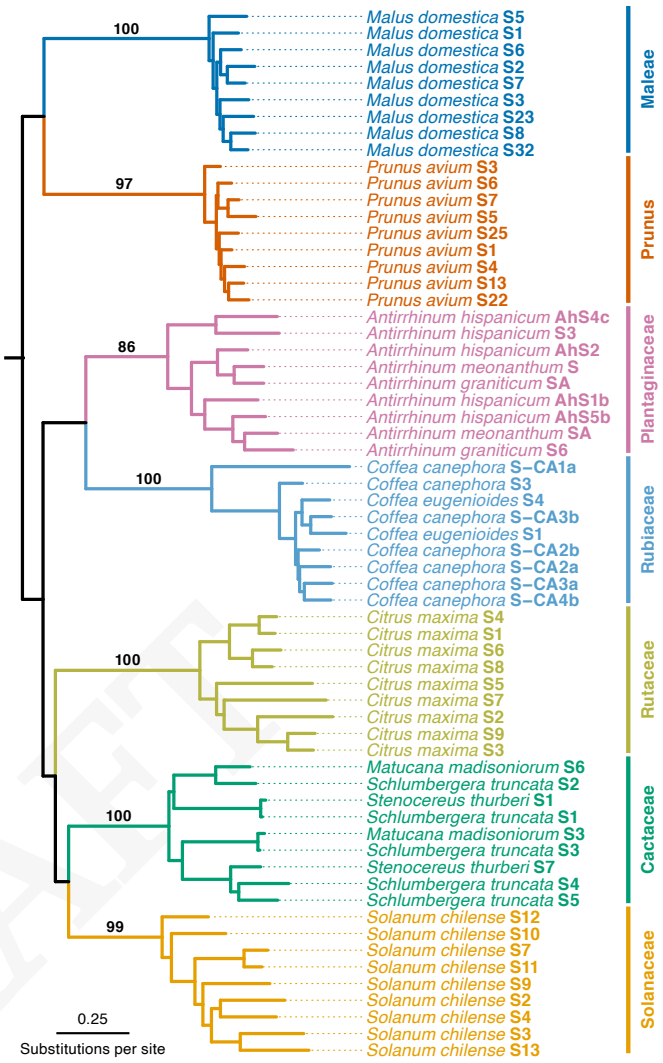

Fig. 3. A gene tree of S-RNases. Previously characterized alleles from Solanaceae, Plantaginaceae, Rubiaceae, Rutaceae, Rosaceae are shown alongside Cactaceae S-RNase alleles uncovered in the present study. Cactaceae sequences follow the naming scheme from Figure 1. Alleles from *Schlumbergera*, *Matucana*, and *Stenocereus* display trans-generic polymorphism, a common feature of SI systems. (A gene tree with all sampled plant T2/S-RNase gene family members, along with branch support is available in Supplementary Information Figure S1.

of trans-generic polymorphism. One of the two *Matucana madisoniorum* S-RNase alleles is closely related (95.3% identical) to *Schlumbergera truncata* allele S_3, and the other is most similar to the *Schlumbergera truncata* allele S_2 (72.9% AA sequence identity; allele named S_6). Likewise, one of the two *Stenocereus thurberi* S-RNase alleles is closely related to *Schlumbergera truncata* allele S_1 (96.6% identical), while the other was most similar to the *Schlumbergera truncata* allele S_5 (67.3% AA sequence identity; allele named S_7).

Pollen expression of putative SLF genes. Pollen RNA-seq assembly of *Schlumbergera truncata* resulted in a median transcriptome assembly of 28,580 genes and 46,638 isoforms (range 18,571–35,689 and 27,729–58,662, respectively; Table S1). It is comparable to the assemblies recovered in *Petunia inflata* pollen (23,731 and 25,522 genes; 32). Plant genomes can contain hundreds of F-box domain containing genes (33, 34), but only a small fraction of these is expressed in pollen. We re-

stricted our search of F-box domain containing genes by using only previously described S-locus F-box sequences as BLAST queries. Among the assembled genes, we recovered a total of twenty transcripts that contain F-box motifs and identified eleven as putative SLFs (see Phylogenetic relationships among F-boxes). The coding sequence regions of these eleven genes range from 1230 to 1341 bp, similar to other SLFs (Table SX) and none were found in other tissues, with the exception of `FB_10` which was recovered from the pistils of 15H03 and 15H09 (Figure 1b, Table SX). It is unclear, however, whether this is a result of contamination. *Schlumbergera truncata* makes copious amounts of pollen and we did not emasculate plants used in our experiments. The rest were expressed exclusively in pollen with maximum TMM ranges of 38–126 and 0–6 in pollen and pistils, respectively (Table SX).

S-RNase genotypes predict crossing phenotypes. A functional link (prediction of crossing success), that is the confirmation of an association between putative S-RNase genotypes and the expected crossing phenotypes, is critical piece of evidence for inference of RSI in textitSchlumbergera truncata. In each case, crossing results are consistent with the expectation that the RNase sequences we obtained from represent functional alleles under single-locus gametophytic SI.

First, we recovered relatively few alleles in our sample of *Schlumbergera truncata*—five alleles (six S-genotypes) from nine unique individuals (Figure 1). This relative dearth of diversity at the S-locus is unsurprising. Although crab cacti are commonly grown as ornamental plants, they appear to have their origins in cultivation from relatively few progenitor plants (35). Such restriction of unique genotypes proved helpful for confirming crossing relationships (Figure 4). All crosses involving individuals with matching S-allele genotypes failed. On the other hand, all crosses involving individuals with differing S-allele genotypes succeeded, albeit with variable rates of fruit set and seed set. Some variation is expected due to subtle environmental conditions, for example. Except for 15H06, specifically selected because of a reported occurrence of SC, no other individuals set fruit following self-fertilization. Intriguingly, this self-compatible plant is also the only one that expresses S-RNase allele S_4 (genotype S_2S_4; Figure 1). The expression level of this S-RNase allele is within the range of functional alleles and it does not possess easily identifiable as loss-of-function mutations (e.g., loss of RNase activity, premature stop codons, frameshift mutations). Taken together, our data point to the possibility that the breakdown of SI in this case may be caused by pollen-part mutations, which warrants further work. Any pointed interpretation of this data is tempered by the fact that we recovered the smallest transcriptome assembly from this plant. We found approximately half of the genes/isoforms in 15H06 pollen compared to the median pollen transcriptome assembly.

Phylogenetic relationships among S-RNases. Gene phylogenies may not form strictly bifurcating relationships when recombination and gene conversion cause homoplasy, especially over hundred-million-year time scales (24). Nevertheless, information regarding gene relationships among putative S-RNases, especially at shallower time scales, appears to be reasonably well-preserved. This idea is supported by comparisons with data independent of DNA sequence alignments used in phylogenetic inference, such as intron presence/absence and isoelectric

♀ \ ♂	S_1S_2	S_1S_3	S_1S_5	S_2S_3	S_2S_4	S_2S_5
S_1S_2	0/24	2/2	4/4	17/17	3/3	4/7
S_1S_3	1/1	0/3	1/3	1/1		
S_1S_5	4/4	1/1	0/4	4/5	1/2	1/3
S_2S_3	12/18	1/1	4/4	0/34	5/5	3/5
S_2S_4	1/1			4/4	5/6	1/1
S_2S_5	4/5	4/4		3/3	1/3	0/3

Fig. 4. Crossing relationships among unique genotypes identified by RNA-seq. Identical dam × sire genotype crossing ought to result in failure to set fruit (pollen alleles are recognized and rejected if they match alleles in maternal genotype). Else, when the parental plants possess non-identical genotypes, crosses ought to be successful. The individual with genotype S_2S_4 (15H06) was included in the study because it is self-fertile. Our data match genotype-phenotype mapping expectations, indicating that the identified sequences are functional S-alleles.

points (6). Therefore, we also examined the inferred phylogenetic relationships among our set of candidates to accumulate additional weight of evidence that these genes are functional S-RNases.

As in nearly all previous studies, our sample of members of the T2/S-RNase family involved in causing pistil-part SI response forms a closely related "class" of RNases, termed "Class III RNases," which share conserved structural features and isoelectric points. To date, all members of this protein class uncovered within each plant family are monophyletic (Figures 3 and S1), although this inference is complicated by the presence of paralogous pseudogenes, loss-of-function mutants, and deep divergences.

Intriguingly, even our small sample of S-RNases from *Schlumbergera* and putative S-RNases from *Matucana* and *Stenocereus* is sufficient to recover S-alleles from different genera that are more closely related to one another than to alleles within the same genus. This common S-locus feature, termed trans-generic polymorphism is caused by the action of negative frequency-dependent selection, a form of diversifying selection on mating type loci (30). Such selection preserves ancestral polymorphism, and it has been uncovered in all cases for multiallelic SI systems recovered thus far. Moreover, trans-generic polymorphism provides unparalleled evidence that the RSI mechanism has been transmitted in an unbroken line of ancestry, from the most recent common ancestor to the extant species (18) within all families where it is found in eudicots, now including Cactaceae. Finally, the finding of RSI in Cactaceae expands the known distribution of this mechanism to Caryophyllales, an order of plants distantly related to rosid (Rosales and Sapindales) and asterid lineages (Solanales, Lamiales, Gentianales), where it is known to operate.

Phylogenetic relationships among putative S-locus F-box–containing proteins. Plant genomes contain up to dozens of T2/S-RNases, but hundreds of genes with the N-terminal F-box motif (34). This necessarily complicates the detection of putative SLFs and related functional inference (32, 36, 37). The SLFs across plant families do not form a single clade, with many SLF-like genes braking the monophyly. Within families, however, SLFs form largely monophyletic groups. We find a similar pattern in the twenty *Schlumbergera truncata* pollen-expressed F-box domain containing genes, eleven of these (`FB_07`–`FB_17`) belong to a clade of closely related sequences, most of them exclusively expressed in pollen (Figure S1).

Although each S-haplotype is expected to contain many linked F-box containing genes, expression levels of SLFs are

much lower than S-RNases, and our extraction methods and coverage were relatively low. Consequently, we cannot confidently infer linked sequences (haplotype subtraction; and leave it for future work). The eleven putative SLFs also appear highly tissue-specific, except for `FB_10`, which is also found in pistil RNA. Nevertheless, with these limitations in mind, all mutually incompatible individuals with identical S-RNase genotypes (15H01, 15H02, & 15H04 vs. 15H05 & 15H07) show similar putative-SLF expression (Figure 1b). Together with their counterparts from organ pipe cactus (*Stenocereus thurberi*) these genes form a monophyletic group (Figure S2). Although we suspect that extensive gene conversion and recombination among these tandem-duplicated paralogs may undermine the assumptions of strictly bifurcating underlying process, they also appear relatively closely related to SLF genes in other families.

Discussion

The combined evidence from RNA-seq assemblies, analyses of gene expression, phylogenetic relationships, and crossing data indicates that *Schlumbergera truncata* enforces SI with a genetic mechanism found across eudicots. SI is found in dozens of genera in the cactus family (28). It may be underpinned by the same RSI mechanism in most, if not all instances within the family. The molecular components of SI are often conserved on smaller phylogenetic scales and members of each family usually rely on the same mechanism (**?**). Interestingly, a suite of superficial traits, such as the number of pollen nuclei and stigmatic surface traits, which are commonly employed as proxies for molecular mechanisms of SI, fail to accurately predict either the genetic mode or molecular mechanism in Cactaceae. Specifically, cacti exhibit dry papillate stigmas and trinucleate pollen, both of which are broadly associated with the sporophytic genetic mode of SI. Instead, at least *Echinopsis, Hatiora, Rhipsalidopsis,* and *Schlumbergera* exhibit the gametophytic mode (35), and we find it effected by RSI in *Schlumbergera*.

Our recovery of putative S-RNase alleles in *Matucana madisoniourum* and *Stenocereus thurberi*, and the preserved ancestral polymorphism implies that this mechanism was also continually present since the common ancestor of nearly all Cactaceae. Transmission of such ancient polymorphism in an unbroken line of descent to present-day species is expected under persistent negative frequency-dependent selection (38). The resulting pattern of preserved trans-generic polymorphism among S-RNase alleles reflects constant presence of SI. This pattern is common within all plant families known to express RSI, and it can be broadly utilized in study of paleopopulation demography and macroevolutionary inference (30, 38).

The finding of RSI in the sixth eudicot family (Figure 5), is also consistent with the hypothesis that the common ancestor of nearly all eudicots utilized RNases as the pistil component of an ancestral SI mechanism (6, 11). Cactaceae, which represent one of the most spectacular arid-adapted radiations on Earth, are members of Caryophyllales, an order of over 10,000 species of core eudicots.Some members, such as Cactaceae and Chenopodiaceae, were long-known express gametophytic SI, while others (Caryophyllaceae, Nyctaginaceae) express sporophytic SI (28). Therefore, it appears that both the conservation of the RSI system as well as the propensity for the origin of novel, non-homologus, systems had been previously underestimated.

It is all but certain now that a similar and homologous system, with orthologous RNase and F-box-containing specificity-determining components, operated more than hundred million years ago. Several important questions, however, remain. While it is clear that this mechanism displays considerable variation across families where it is found, we do not understand its capacity for change since the early Cretaceous Period. For instance, polyploidization seems to obligately cause self-compatibility in most species with RSI, including our focal species *Schlumbergera truncata* (40, 41), but a few exceptions exist, particularly among the plant family Rosaceae (reviewed in 42, 43). Furthermore, it operates in most species as a non-self-recognition system (pollen-part SLF genes defend the pollen tube by defusing all S-RNases except those within their own haplotype), including the apple genus *Malus* (21). But in closely related stonefruit genus *Prunus*, orthologous molecular components instead cause self-recognition (pollen-part SLF genes defend the pollen tube by defusing S-RNases linked to their own haplotype), which is a deviation of genetic and operational significance (44). Whether the mechanism in cacti proceeds by self- or non-self-recognition remains presently unclear, but the increased pace of discovery of genes that underpin RSI in this and other families clearly marks the beginning of a new and accelerated comparative phase in the study of the causes and consequences of breeding systems in plants.

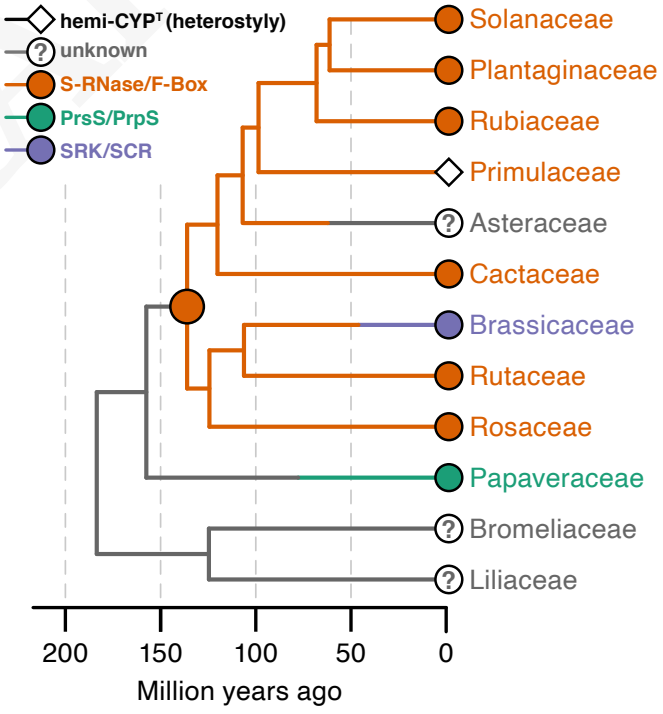

Fig. 5. The distribution of SI systems in selected eudicots and monocots. Colored branches mark the plausible ordering of the evolution of SI systems (as identified in the legend). Stars mark families in which we first identified S-RNase-based self-incompatibility (SI). A steadily increasing weight of evidence shows that this form of SI is homologous across Pentapetalae (most eudicots) and that it was employed by the common ancestor of roughly two-thirds of all land plants.

A baseline expectation of a high propensity for divergence among homologous expression of RSI can be gleaned from previous studies as well as our present findings concerning

Schlumbergera truncata. For decades, families with trinucleate pollen and dry stigmas—including Cactaceae—were thought unlikely to express any form of gametophytic SI, including RSI (45). We find RSI in precisely such a family, in apparent departure from this strong angiosperm-wide correlation. S-RNase alleles from Cactaceae also possess an extra intron, not found in other S-RNases characterized to date. Therefore, even in absence of a complete understanding of functional genetic operation of this system in cacti, we can conclude that the RSI system can broadly undergo a great deal of alteration in its constituent components and associated traits.

We do not provide the benchmark standard evidence used in molecular biology, an *in vivo* functional assay, for determination of causal relationship between genotype and phenotype. Recombinant S-RNase inhibition of incompatible pollen tube growth *in vivo* or *in vitro* would strongly reinforce our findings, as would similar generation of pollen tube growth following introduction of recombinant SLF genes (e.g., 25). And yet, our approach has some important advantages.

Many plants are ill-suited for classical molecular genetic approaches because their juvenile phase exceeds require decade(s). Even species the relatively amenable to manipulation, such as the widely cultivated *Schlumbergera*, require two years from germination to flowering (28). Our RNA-seq-based characterization and correlation of genotype and phenotype expression is simple, inexpensive, and easily applied. The present findings are directly relevant for hundreds of species of cacti in precarious existence, where it can be deployed as a part of effective germplasm conservation program (22, 28). More generally, it can be applied in many other families and it holds great promise for determining the genetic basis of SI in understudied long-lived perennials, especially tropical trees and many species with long juvenile phases. In this way, it can lower the barriers to—and greatly accelerate the pace of—discovery of operational SI mechanisms across angiosperm families, far above the present two families per decade, since ca. 1986.

Materials and Methods

Study Organism. *Schlumbergera truncata* (Haw.) Moran, commonly known as "crab cactus" or "false Christmas cactus," is perhaps the most widely cultivated species of cacti (46). Designated by the International Union for Conservation of Nature (IUCN) as Vulnerable (47), its wild populations have a highly restricted range across the slopes of Brazilian Atlantic forests at the southern edge of the tropics, at elevations of 700–1000 m. This species is easily grown from cuttings and it has been extensively hybridized across Western Europe, since the early 19th century. It produces a rich array of flower colors and forms (35). Although there are dozens of named varieties, nearly all commercially grown plants are of ultimately uncertain provenance. Our samples were sourced from a haphazardly collected personal collection (B.I.; accessions 15H01–15H09). Individuals of this species are generally SI, displaying the gametophytic genetic mode of SI, in which a pistil-expressed genotype recognizes (and rejects) matching pollen-expressed S-allele haplotype (28). One individual, accession 15H06 was specifically included in the study because it was reported to display self-compatibility (SC). SC is also known to occasionally occur in mutant individuals of otherwise SI species (5, 48)

Additionally, we were granted access to one individual of *Matucana madisoniorum* (Hutchison) G.D.Rowley from the Huntington Botanical Gardens and Library (Pasadena, CA; accession 42118). We were interested in this species, because it is designated by the IUCN as Critically Endangered, restricted to a single population within an area no greater than 16 km^2 (49). S-locus genotyping in such rare species could provide crucial information for conservation efforts, because mate limitation due to S-allele scarcity can drive or reinforce population declines (e.g. 50).

RNA Sequencing. Pistils (without ovaries), pollen, stem (cladode), and root tissues were collected and immediately submerged in 1.5 ml of RNA*later*™ solution (Invitrogen). Submerged samples were held at room temperature for thirty minutes and then moved to a -80° C freezer for storage. Approximately 100 mg of tissue was ground to a fine powder in 1.5 ml tubes submerged in liquid nitrogen. Total RNA was isolated using Total RNA Mini Kit (Plant kit; IBI Scientific, Cat. No. IB47341) following manufacturer's instructions. We assessed RNA concentration and purity with a NanoDrop™ Lite Spectrophotometer (Thermo Scientific). The twenty-three samples used in this study were sequenced as part of a larger sequencing effort which consisted of four separate sequencing runs and included additional samples from other plant species. Sequencing libraries were prepared using the KAPA Stranded mRNA-Seq (Roche), and these libraries were sequenced on a single lane of Illumina HiSeq 4000 or Illumina NovaSeq 6000 platform (paired-end 150 bp reads) at the Duke University Center for Genomic and Computational Biology. The number of resulting read pairs (for the twenty-three samples presented here) ranged from 4,148,932 to 9,618,084 with a median of 6,363,556 and average of 6,293,553 (Table S1).

Transcriptome Assemblies. Raw paired-end Illumina reads were first processed using Rcorrector v1.0.4 (51) to correct for random sequencing errors. Then, reads were trimmed with Trimmomatic v0.39 (52) to remove any read containing bases with Phred scores lower than 20, low quality reads less than 50 bp long, and any adapter or other Illumina-specific sequences that were still present. The remaining reads were filtered with Kraken 2 (53) to remove Small and Large Subunit ribosomal RNA (SILVA database) (54) and contaminating reads (minikraken2_v2 database). Additionally, we used custom-built databases, derived from RefSeq libraries: UniVec_Core, viral, mitochondrion, plastid, plasmid, archaea, bacteria, protozoa, human, and fungi to minimize the number of contaminating and non-nuclear reads. Only paired reads were used for transcriptome assemblies. *Schlumbergera truncata* filtered reads were combined across all samples into a single RNA-seq data set. We conducted a *de novo* transcriptome assembly using Trinity v2.8.5 (55) to generate a single reference transcriptome assembly for *Schlumbergera truncata*. The same assembly protocol was followed for the single pistil sample of *Matucana madisoniorum*.

We identified transcripts of candidate genes (S-RNases and SLFs) and expression control/reference genes (EF: Elongation factor 1-alpha 1; TU: alpha-tubulin; and AC: actin) in the transcriptome assemblies with the kakapo v0.7.3-dev pipeline (http://flightless.one). Transcript quantification was performed using RSEM (56). For differential expression analysis and trimmed mean of M-values (TMM, mean of log-expression ratios (57)) calculation SARTools v1.7.3 (58) with edgeR v3.30.2 in R v4.0.2 (59, 60) was used.

We examined the four published draft genomes of cacti (29), but we only found the S-locus partly assembled in the draft genome of organ pipe cactus, *Stenocereus thurberi*. This is not unusual, because the S-locus is generally difficult to assemble, being heterozygous, highly polymorphic, and replete with repetitive elements. We extracted all organ pipe genomic contigs with putative S-RNase or SLF-like sequences and used them to compare intron locations with those inferred from immature mRNA (with introns) in crab cactus transcriptomes. Finally, we calculated isoelectric point values using the ProteinAnalysis tool in Biopython v1.78 (61). Signal peptide sequences are not included in pI calculations.

Phylogenetic Analyses. Additional T2/S-RNase and F-box domain containing sequences were obtained using BLAST by searching GenBank nucleotide collection (nr/nt) database and the genomes of *Arabidopsis thaliana* (Brassicaceae), *Coffea eugenioides* (Rubiaceae), *Prunus avium* (Rosaceae), *Spinacia oleracea* (Amaranthaceae), *Solanum chilense* (Solanaceae), and *Stenocereus thurberi* (Cactaceae) (29) downloaded from NCBI FTP server (ftp://ftp.ncbi.nlm.nih.gov). The genome of *Citrus maxima* (Rutaceae) (62) was downloaded from the "Citrus Genome Database" website (https://citrusgenomedb.org).

Sequence alignments were constructed using DECIPHER v2.16.1 in R v4.0.2 (60, 63). The coding regions of the 453 F-box domain containing sequences were translation-aligned using the `AlignTranslation` function. The alignment of the 153 T2/S-RNase sequences was constructed in two stages. First, the coding regions of sequences belonging to Classes I, II, and III were aligned separately using the `AlignTranslation` function, then the three resulting alignments were profile aligned using the `AlignProfiles` function.

Phylogenetic relationships for T2/S-RNase and F-box domain-containing sequences were inferred using RAxML v8.2.12 (64) with general

time-reversible model and Gamma-distributed among-site rate variation (option `-m GTRGAMMA`). In order to evaluate a diverse set of tree topologies and to reduce the chances of getting trapped in local likelihood maxima, the following strategy was used. For each sequence set, RAxML was run one thousand times. For each run, a random starting tree was used (`-d`), and subtree pruning re-grafting (SPR) moves were randomized (`-f t`). Additionally, the value for the "initial rearrangement setting" (`-i`) was chosen at random (with equal draw probabilities) from a list of values: 2, 4, 6, 8, 12, 16, 20, 24, 32, 48. The tree with the maximum likelihood (least negative log-likelihood score) was selected as the best tree. The log-likelihood scores of the best T2/S-RNase and F-box domain containing gene trees were -56941.297900 and -440386.747201, respectively. Finally, we performed one thousand ordinary (not "fast") non-parametric bootstrap analyses for each dataset (`-b`). For visual clarity and in view of space considerations, we display a subsampled T2/S-RNase gene tree in the main body of the paper (Figure 3) and relegate the full versions of both gene trees (T2/S-RNase and F-Box genes) with hundreds of tips and branch support to the Supplementary Information (Figures S1 and S2).

Controlled Crosses and Phenotyping. We performed controlled crosses to confirm the association between an individual plant's SI function (recognition phenotype) and putative S-RNase allele sequences (genotype). Prior data indicates that this species displays gametophytic SI with stylar inhibition (28, 35). The molecular genetic basis of the response remained unknown, so our objective centered on determining the genotypes at the putative S-RNases and whether they predicted the crossing phenotypes, given the action of gametophytic SI. For example, we expect crosses with a plant bearing the genotype S_1S_2 at the S-locus to be reciprocally sterile with all identical genotypes, and reciprocally fertile when crossed with any other genotype.

Our study plants differed in size and condition, as well as flowering periods, which is common for this species (28). A full diallel crossing design was not possible, because the limited number of co-flowering individuals and available flowers. Nevertheless, we self-fertilized each plant and conducted as many cross-pollination as was practical—in all cases using all flowers for either crossing or tissue collection. In total, we conducted 181 bidirectional cross- and self-fertilization attempts. Each cross was labeled with a jewelry tag and scored 12–14 days later. As expected, all individuals display SI—fail to set fruit following self-fertilization—except for one (15H06) specifically obtained because it was reported to be SC. Furthermore, we expected crosses to fail when putative S-allele genotypes matched for crossed individuals (when fully incompatible, as in 15H01/15H02/15H04 & 15H05/15H07), and to succeed when the genotypes were not matched. The authors independently performed genotyping (K.R.) and phenotyping experiments (B.I.), and compared the outcomes after both experiments were completed. Finally, we used the assembled transcriptome to confirm that the individuals that share the S-locus genotypes are distinct genets (unique genetic individuals) and not ramets (clones) by examining their genotypes at segregating sites over 50 most highly expressed single-copy genes.

Data Accessibility. Web service, supplementary information, and GenBank.

ACKNOWLEDGMENTS. We are grateful to The Huntington Library, Art Museum, and Botanical Gardens for providing access to *Matucana madisoniorum*, encouragement and advice, particularly Emma Goldberg, Alex Harkness, Josh Kohn, John Trager, Dylan Hannon, Jim Folsom, Jenn Yost, and Matt Ritter. We received generous advice and help with RNA sequencing from from Nicolas Devos (Duke University), and critical plant materials from Ken Olsen (15H06 "Sandy-O"). We are grateful that despite limited preliminary data this work was supported by the National Science Foundation grant NSF-DEB-1655692 to B.I.

1. KE Holsinger, Reproductive systems and evolution in vascular plants. *Proc. Natl. Acad. Sci.* **97**, 7037–7042 (2000).
2. SCH Barrett, Mating strategies in flowering plants: the outcrossingselfing paradigm and beyond. *Philos. Transactions Royal Soc. London. Ser. B: Biol. Sci.* **358**, 991–1004 (2003).
3. M Uyenoyama, The evolution of breeding systems in *Evolutionary Genetics: From Molecules to Morphology*, eds. R Singh, C Krimbas. (Cambridge University Press, Cambridge), pp. 300—316 (2000).
4. SI Wright, S Kalisz, T Slotte, Evolutionary consequences of self-fertilization in plants. *Proc. Royal Soc. B: Biol. Sci.* **280**, 20130133 (2013).
5. AR Raduski, EB Haney, B Igić, The expression of self-incompatibility in angiosperms is bimodal. *Evolution* **66**, 1275–1283 (2012).
6. B Igić, JR Kohn, Evolutionary relationships among self-incompatibility RNases. *Proc. Natl. Acad. Sci.* **98**, 13167–13171 (2001).
7. B Igić, R Lande, JR Kohn, Loss of Self-Incompatibility and Its Evolutionary Consequences. *Int. J. Plant Sci.* **169**, 93–104 (2008).
8. JB Nasrallah, TH Kao, ML Goldberg, ME Nasrallah, A cDNA clone encoding an S-locus-specific glycoprotein from Brassica oleracea. *Nature* **318**, 263–267 (1985).
9. HC Foote, et al., Cloning and expression of a distinctive class of self-incompatibility (S) gene from Papaver rhoeas L. *Proc. Natl. Acad. Sci.* **91**, 2265–2269 (1994).
10. J Li, et al., Genetic architecture and evolution of the S locus supergene in Primula vulgaris. *Nat. Plants* **2**, 16188 (2016).
11. JE Steinbachs, KE Holsinger, S-RNase–mediated Gametophytic Self-Incompatibility is Ancestral in Eudicots. *Mol. Biol. Evol.* **19**, 825–829 (2002).
12. MA Anderson, et al., Cloning of cDNA for a stylar glycoprotein associated with expression of self-incompatibility in Nicotiana alata. *Nature* **321**, 38–44 (1986).
13. Y Xue, R Carpenter, HG Dickinson, ES Coen, Origin of allelic diversity in antirrhinum S locus RNases. *The Plant Cell* **8**, 805–814 (1996).
14. H Sassa, et al., Self-incompatibility (S) alleles of the rosaceae encode members of a distinct class of the T2/S ribonuclease superfamily. *Mol. Gen. Genet. MGG* **250**, 547–557 (1996).
15. E Asquini, et al., S-RNase-like Sequences in Styles of Coffea (Rubiaceae). Evidence for S-RNase Based Gametophytic Self-Incompatibility? *Trop. Plant Biol.* **4**, 237–249 (2011).
16. M Liang, et al., Evolution of self-compatibility by a mutant Sm-RNase in citrus. *Nat. Plants* **6**, 131–142 (2020).
17. DE Soltis, PS Soltis, PK Endress, PE Soltis, MW Chase, *Phylogeny and Evolution of Angiosperms* ed. DE Soltis. (Sinauer Associates Incorporated, Sunderland, MA), (2005).
18. EE Goldberg, et al., Species Selection Maintains Self-Incompatibility. *Science* **330**, 493–495 (2010).
19. B Aguiar, et al., Convergent Evolution at the Gametophytic Self-Incompatibility System in Malus and Prunus. *PLOS ONE* **10**, e0126138 (2015).
20. D Matsumoto, R Tao, Distinct Self-recognition in the Prunus S-RNase-based Gametophytic Self-incompatibility System. *The Hortic. J.* **85**, 289–305 (2016).
21. K Kubo, et al., Collaborative Non-Self Recognition System in S-RNase–Based Self-Incompatibility. *Science* **330**, 796–799 (2010).
22. MdC Mandujano, I Carrillo-Angeles, C Martinez-Peralta, J Golubov, Reproductive Biology of Cactaceae in *Desert Plants*, ed. KG Ramawat. (Springer-Verlag), pp. 197—230 (2010).
23. JV Muñoz-Sanz, E Zuriaga, F Cruz-García, B McClure, C Romero, Self-(in)compatibility Systems: Target Traits for Crop-Production, Plant Breeding, and Biotechnology. *Front. Plant Sci.* **11**, 195 (2020).
24. K Ramanauskas, B Igić, The evolutionary history of plant T2/S-type ribonucleases. *PeerJ* **5**, e3790 (2017).
25. P Sijacic, et al., Identification of the pollen determinant of S-RNase-mediated self-incompatibility. *Nature* **429**, 302–305 (2004).
26. S Wright, The Distribution of Self-Sterility Alleles in Populations. *Genetics* **24**, 538–52 (1939).
27. EM East, AJ Mangelsdorf, A New Interpretation of the Hereditary Behavior of Self-Sterile Plants. *Proc. Natl. Acad. Sci.* **11**, 166–171 (1925).
28. TH Boyle, The Genetics of Self-Incompatibility in the Genus *Schlumbergera* (Cactaceae). *J. Hered.* **88**, 209–214 (1997).
29. D Copetti, et al., Extensive gene tree discordance and hemiplasy shaped the genomes of North American columnar cacti. *Proc. Natl. Acad. Sci.* **114**, 12003–12008 (2017).
30. TR Ioerger, AG Clark, TH Kao, Polymorphism at the self-incompatibility locus in Solanaceae predates speciation. *Proc. Natl. Acad. Sci.* **87**, 9732–9735 (1990).
31. B Igic, WA Smith, KA Robertson, BA Schaal, JR Kohn, Studies of self-incompatibility in wild tomatoes: I. S-allele diversity in Solanum chilense Dun. (Solanaceae). *Heredity* **99**, 553 (2007).
32. JS Williams, JP Der, CW dePamphilis, Th Kao, Transcriptome Analysis Reveals the Same 17 S-Locus F-Box Genes in Two Haplotypes of the Self-Incompatibility Locus of Petunia inflata. *The Plant Cell* **26**, 2873–2888 (2014).
33. JM Gagne, BP Downes, SH Shiu, AM Durski, RD Vierstra, The F-box subunit of the SCF E3 complex is encoded by a diverse superfamily of genes in Arabidopsis. *Proc. Natl. Acad. Sci.* **99**, 11519–11524 (2002).
34. G Xu, H Ma, M Nei, H Kong, Evolution of F-box genes in plants: Different modes of sequence divergence and their relationships with functional diversification. *Proc. Natl. Acad. Sci.* **106**, 835–840 (2009).
35. TH Boyle, Identification of self-incompatibility groups in *Hatiora* and *Schlumbergera* (Cactaceae). *Sex. Plant Reproduction* **16**, 151–155 (2003).
36. C Xu, et al., Identification of a canonical SCF-SLF complex involved in S-RNase-based self-incompatibility of *pyrus* (Rosaceae). *Plant Mol. Biol.* **81**, 245–257 (2013).
37. JS Williams, L Wu, S Li, P Sun, TH Kao, Insight into S-RNase-based self-incompatibility in Petunia: recent findings and future directions. *Front. Plant Sci.* **6**, 41 (2015).
38. B Igić, L Bohs, JR Kohn, Historical inferences from the self-incompatibility locus. *New Phytol.* **161**, 97–105 (2004).
39. T Paape, et al., A 15-Myr-Old Genetic Bottleneck. *Mol. Biol. Evol.* **25**, 655–663 (2008).
40. JS Miller, DL Venable, Polyploidy and the Evolution of Gender Dimorphism in Plants. *Science* **289**, 2335–2338 (2000).
41. C Parks, TH Boyle, Variation in ploidy level, fertility, and breeding behavior in cultivated *schlumbergera* (Cactaceae). *Acta Hortic.* **623**, 341–350 (2003).
42. JL Stone, Molecular Mechanisms Underlying The Breakdown Of Gametophytic Self-incompatibility. *The Q. Rev. Biol.* **77**, 17–32 (2002).
43. BK Mable, Polyploidy and self-compatibility: is there an association? *New Phytol.* **162**, 803–811 (2004).
44. S Fujii, Ki Kubo, S Takayama, Non-self- and self-recognition models in plant self-incompatibility. *Nat. Plants* **2**, 16130 (2016).
45. JL Brewbaker, The distribution and phylogenetic significance of binucleate and trinucleate pollen grains in the angiosperms. *Am. J. Bot.* **54**, 1069–1083 (1967).
46. EF Anderson, *The Cactus Family*. (Timber Press (Portland, Oregon, U.S.A.)), (2001).
47. N Taylor, D Zappi, Schlumbergera truncata. *The IUCN Red List Threat. Species* **2017**, e.T152554A121599528 (2017).

48. DP Brown, Self-Fertilization of a *Schlumbergera truncata* Cultivar: Breakdown of Self-incompatibility. *Haseltonia* **16**, 54–56 (2011).
49. C Ostalaza, J Roque, Matucana madisoniorum. *The IUCN Red List Threat. Species* **2013**, e.T152396A632099 (2013).
50. JW Busch, DJ Schoen, The evolution of self-incompatibility when mates are limiting. *Trends Plant Sci.* **13**, 128–136 (2008).
51. L Song, L Florea, Rcorrector: efficient and accurate error correction for Illumina RNA-seq reads. *GigaScience* **4**, 48 (2015).
52. AM Bolger, M Lohse, B Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).
53. DE Wood, J Lu, B Langmead, Improved metagenomic analysis with Kraken 2. *Genome Biol.* **20**, 257 (2019).
54. C Quast, et al., The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
55. MG Grabherr, et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* **29**, 644–652 (2011).
56. B Li, CN Dewey, RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinforma.* **12**, 323 (2011).
57. MD Robinson, A Oshlack, A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol.* **11**, R25 (2010).
58. H Varet, L Brillet-Guéguen, JY Coppée, MA Dillies, SARTools: A DESeq2- and EdgeR-Based R Pipeline for Comprehensive Differential Analysis of RNA-Seq Data. *PLOS ONE* **11**, e0157022 (2016).
59. MD Robinson, DJ McCarthy, GK Smyth, edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
60. R Core Team, *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, Vienna, Austria), (2020).
61. PJA Cock, et al., Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
62. X Wang, et al., Genomic analyses of primitive, wild and cultivated citrus provide insights into asexual reproduction. *Nat. Genet.* **49**, 765–772 (2017).
63. ES Wright, DECIPHER: harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinforma.* **16**, 322 (2015).
64. A Stamatakis, RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).