Plant Molecular Biology **54:** 727–742, 2004.

Chromosome walking in the *Petunia inflata* self-incompatibility (*S*-) locus and gene identification in an 881-kb contig containing *S_2-RNase*

Yan Wang[1], Tatsuya Tsukamoto[2], Ki-wan Yi[2], Xi Wang[2,4], Shihshieh Huang[3], Andrew G. McCubbin[2,5] and Teh-hui Kao[1,2,*]

[1]*Intercollege Graduate Degree Program in Plant Physiology, 403 Althouse Lab, The Pennsylvania State University, University Park, PA 16802, USA;* [2]*Department of Biochemistry and Molecular Biology, 403 Althouse Lab, The Pennsylvania State University, University Park, PA 16802, USA (*author for correspondence; e-mail txk3@psu.edu);* [3]*Mystic Research, Monsanto Company, 62 Maritime Drive, Mystic, CT 06355, USA;* [4]*Present address: Department of Molecular, Cellular and Developmental Biology, University of Michigan, Ann Arbor, MI 48109, USA;* [5]*Present address: School of Biological Sciences, Washington State University, Pullman, WA 99164, USA*

Received 6 January 2004; accepted in revised form 13 April 2004

Key words: cDNA selection, chromosome walking, *Petunia inflata*, self-incompatibility, *S*-locus contigs

Abstract

Self-incompatibility (SI) in the Solanaceae, Rosaceae and Scrophulariaceae is controlled by the polymorphic *S* locus, which contains two separate genes encoding pollen and pistil determinants in SI interactions. The *S-RNase* gene encodes the pistil determinant, whereas the pollen determinant gene, named the pollen *S* gene, has not yet been identified. Here, we set out to construct an integrated genetic and physical map of the *S* locus of *Petunia inflata* and identify any additional genes located at this locus. We first conducted chromosome walking at the S_2 locus using BAC clones that contained either *S_2-RNase* or one of the nine markers tightly linked to the *S* locus. Ten separate contigs were constructed, which collectively spanned 4.4 Mb. To identify additional genes located at the S_2 locus, a 328-kb region (part of an 881-kb BAC contig) containing *S_2-RNase* was completely sequenced. Approximately 76% of the region contained repetitive sequences, including transposon-like sequences. Other than *S_2-RNase*, an F-box gene, named *$PiSLF_2$* (S_2-allele of *P. inflata S*-locus F-box gene), was the only predicted gene whose deduced amino acid sequence was similar to the sequences of known proteins in the database. Two different cDNA selection methods were used to identify additional genes in the 881-kb contig; 11 groups of cDNA clones were identified in addition to those for *S_2-RNase* and *$PiSLF_2$*. RT-PCR analysis of expression profiles and PCR analysis of BAC clones and genomic DNA confirmed that seven of these 11 newly identified genes were located in the 881-kb contig.

Abbreviations: GSI, gametophytic self-incompatibility; PFGE, pulsed-field gel electrophoresis; RT-PCR, reverse transcriptase PCR; SI, self-incompatibility

Introduction

Self-incompatibility (SI) is a reproductive trait adopted by many flowering plants to avoid inbreeding and achieve outcrosses (de Nettancourt, 2001). In most cases, a single polymorphic locus, the *S* locus, controls SI interactions between pollen and the pistil. Genetic studies have revealed two major types of SI, gametophytic SI (GSI) and sporophytic SI (SSI). For GSI, recognition and rejection of self-pollen is determined by the *S* genotype of the haploid pollen, whereas for SSI, this is determined by the *S* genotype of the pollen-producing parent. In the case of GSI, growth of pollen tubes that carry an *S* haplotype identical to one of the *S* haplotypes carried by the pistil is inhibited in the style.

To date, the gene encoding the pistil determinant of SI interactions has been identified in four of the families that possess GSI. Three of these families (Rosaceae, Scrophulariaceae and Solanaceae) employ the same gene, *S-RNase*, in the recognition and rejection of self-pollen (Lee *et al*., 1994; Murfett *et al*., 1994; Xue *et al.,* 1996). Determining how S-RNases mediate *S*-haplotype-specific inhibition of pollen tube growth requires the identification of the pollen *S* gene. A number of approaches have been used to attempt to achieve this end (for review, see Kao and Tsukamoto, in press). Recently, a number of F-box genes located close to *S-RNase* have been identified in several species of the Rosaceae (Entani *et al*., 2003; Ushijima *et al*., 2003) and in *Antirrhinum hispanicum* of the Scrophulariaceae (Lai *et al*., 2002; Zhou *et al*., 2003). Among them, the gene named variously *SLF* (*S* locus F-box) or *SFB* (*S* haplotype-specific F-box) is a potential candidate for encoding the pollen determinant because, in addition to being closest to *S-RNase*, it is specifically expressed in pollen and shows allelic sequence diversity. However, whether this gene is the pollen *S* gene remains to be determined.

To characterize the *S* locus of the Solanaceae and identify the pollen *S* gene by map-based cloning, we previously carried out genetic mapping of the *S* locus of *P. inflata* by recombination analysis of 1205 segregating plants using 13 *S*-linked pollen-expressed genes as markers (Wang *et al*., 2003). Recombination events were detected between four (*3.16*, *G211*, *G212* and *G221*) of the marker genes and *S-RNase*, but none of the crossovers disrupted the normal SI behavior. Based on the recombination frequencies of these four marker genes, the *P. inflata S* locus was mapped to within a 0.25-cM region delimited by markers *3.16* and *G221*, and all the genes required for SI specificity are located within this region. No recombination was found between the other nine marker genes (*3.2*, *3.15*, *A113*, *A134*, *A181*, *A301*, *G261*, *X9* and *X11*) and *S-RNase*. Sequence analysis and/or genomic DNA blot analysis showed that these nine marker genes had a very low level of allelic sequence diversity, making them unlikely candidates for the pollen *S* gene (Wang *et al*., 2003).

In this work, we first set out to construct a physical map of the *S* locus. We conducted chromosome walking from multiple sites of the S_2 locus represented by the previously isolated S_2S_2 BAC clones that contained either S_2-*RNase* or one of the nine marker genes tightly linked to the *S* locus (McCubbin *et al*., 2000b). Since the *S* locus is located in a sub-centromeric region (Entani *et al*., 1999), the presence of highly repetitive sequences made the chromosome walking challenging. Nonetheless, we have obtained 10 separate contigs, each containing either S_2-*RNase* or one of the nine marker genes, and these contigs collectively span ca. 4.4 Mb. We then focused our search for the pollen *S* gene on an 881-kb contig that contained S_2-*RNase*. Two different approaches were used to achieve this goal. First, a 328-kb region containing S_2-*RNase* was completely sequenced. Second, cDNA selection was used to analyze the entire 881-kb BAC contig. Among the genes identified, the most interesting one is a pollen-expressed F-box gene, named *PiSLF*.

Materials and methods

Pooling of the S_2S_2 *BAC library*

The previously constructed S_2S_2 BAC library of *P. inflata* (McCubbin *et al*., 2000b) was used in chromosome walking. This library contained 68 736 clones and was stored in 179 384-well plates. These plates were divided into 10 sets, with set nos. 1–9 each containing 18 plates and set no. 10 containing 17 plates. For each set, the clones from the same row number of all the plates were combined to generate 16 row pools (designated row pool nos. A to P), and the clones from the same column number of all the plates were combined to generate 24 column pools (designated column pool nos. 1–24). All the clones in the same plate were combined to generate plate pools (designated plate pool nos. 1–179). Therefore, a total of 179 plate pools, 160 row pools and 240 column pools were obtained for the BAC library. BAC DNA was prepared from each of the 579 pools as described by McCubbin *et al.* (2000b) and was dissolved in 200 μl of TE.

Screening of the BAC library by PCR

All the 179 plate pools were screened by PCR using appropriate primers designed based on the terminal end sequence(s) of a particular BAC clone used as a starting point for chromosome walking. For each of the positive plate pools identified, the row pools and column pools of the set containing the positive plate were screened by PCR, using the same primer pair, to identify the positive clone(s). PCR was conducted under standard conditions using 1 μl of BAC DNA as template, and the PCR products were analyzed on 1% agarose gels.

Isolation, fingerprinting and blotting of BAC DNA

BAC DNA was isolated from positive clones as described above and dissolved in 35 μl of H_2O. The BAC DNA (10 μl) was then digested with *Bam*HI, and the digests were fractionated by pulsed-field gel electrophoresis (PFGE) using conditions optimized for separations of 1–50-kb DNA fragments (McCubbin *et al.*, 2000b). Overlapping BAC clones were identified by comparing their restriction patterns with that of the starting BAC clone. The fractionated BAC DNA was transferred to a positively charged Biodyne B nylon membrane (Life Technologies), and the DNA blot was hybridized with a radiolabeled probe as described by Wang *et al.* (2003).

Thermal asymmetric interlaced (TAIL-) PCR

The arbitrary degenerate primers (AD1 and AD3), specific nested primer sets (PS1, PS2 and PS3, and PT1, PT2 and PT3), and TAIL-PCR procedure were the same as described by Liu and Whittier (1995), except that 1 μl of 500-fold diluted BAC DNA was used in the primary PCR.

Sequencing of BAC clones and sequence assembly

BAC DNA was isolated as described above from three overlapping BAC clones (120K17, 114G8 and 145J16) that constituted a 328-kb contig. The BAC DNA of each clone was sheared using a Hydroshear apparatus (GeneMachines), and DNA fragments of 1–2 and 4–5 kb were isolated by preparative agarose gel electrophoresis and used for the construction of two separate libraries. Sequencing and sequence assembly were performed essentially as described at www.sciencemag.org/cgi/content/full/294/5550/2323/DCI.

Preparation of P. inflata S_2S_2 C_0t*-1 DNA for cDNA selection*

Genomic DNA was isolated from leaves of S_2S_2 plants using DNAzol reagent (Life Technologies). Approximately 1.3 mg of C_0t-1 DNA was prepared from 6.5 mg of genomic DNA following the protocol of Zwick *et al.* (1997). The C_0t-1 DNA was dissolved in TE at a concentration of 2.0 μg/μl and stored at −20 °C.

Preparation of a cDNA pool from leaves, pistils and pollen for cDNA selection

Total RNA was separately isolated from young leaves, pistils of open flowers, and pollen of open flowers using TRIzol reagent (Invitrogen). Poly(A)$^+$ RNA was isolated from 1 mg of the total RNA using the PolyATract mRNA Isolation System IV (Promega). cDNA was synthesized from a pool of leaf, pistil and pollen poly(A)$^+$ RNA (1.7 μg each) using the cDNA synthesis kit of Takara Bio Inc. The double-stranded cDNA was purified using the QIAquick PCR Purification Kit (Qiagen Inc.), precipitated with ethanol, and dissolved in 20 μl of H_2O.

Two complementary 5′-phosphorylated primers, 5′p-TCGAGAATTCTGGATCCTC-3′ (Oligo 1) and 5′p-GAGGATCCAGAATTCTCGAGTT-3′ (Oligo 2), were mixed in an equal molar ratio, denatured at 100 °C for 10 min, and slowly cooled to room temperature. The PCR amplification cassette was ligated to the cDNA according to the protocol of Simmons and Lovett (1999). The ligated cDNA was purified as described above, and eluted with 100 μl of 10 mM Tris–HCl (pH 8.5).

The cassette-ligated cDNA was amplified (1st PCR) in a 100 μl reaction mixture containing 3 μl cDNA (ca. 20 ng), 2 μM linker primer (Oligo 1), 1× PCR buffer, 0.25 mM dNTPs, 4 units of Taq DNA polymerase, and 0.11 units of Pfu DNA polymerase (Stratagene). A hot PCR method (Parimoo, 1997) was adopted here. Briefly, the reaction cocktail without dNTPs was heated at 94 °C for 3 min, and then held at 80 °C when 2.5 μl of 10 mM dNTPs was added. Standard PCR continued for 30 cycles, with each cycle consisting of denaturation at 94 °C for 30 s, annealing at 55 °C for 30 s, and extension at 72 °C for 2 min. After the final cycle, the extension continued at 72 °C for an additional 10 min. The PCR product was purified and eluted as described above. The eluted cDNA was separately used in nucleolink tube-based and membrane-based cDNA selection.

cDNA selection in an 881-kb BAC contig

cDNA selection was conducted in an 881-kb contig using eight of the 10 overlapping BAC clones (all except 120M2 and 139M11) shown in Figure 2 and a pool of leaf, pistil and pollen cDNA. This contig encompassed the 328-kb region whose sequence was determined. Two different cDNA selection methods, membrane-based selection (Parimoo, 1997) and nucleolink tube-based selection (Childs *et al.*, 2001), were used, with either salmon sperm DNA or C_0t-1 DNA of *P. inflata* S_2S_2 genotype serving as a blocking reagent.

Nucleolink tube-based cDNA selection using salmon sperm DNA as a blocking reagent was performed according to Childs *et al.* (2001) with some modifications. BAC DNA (4 μg) was digested in a 100 μl reaction mixture containing 20 units of *Eco*RI and 5 μl RNase Cocktail (Amicon) at 37 °C overnight. The digested DNA was extracted with phenol and chloroform, precipitated with ethanol, and dissolved in 50 μl of H_2O. To 4 μl of the *Eco*RI digested BAC DNA (ca. 250 ng), 63.5 μl of H_2O was added, and the DNA was denatured at 100 °C for 5 min and quickly chilled on ice. The DNA was mixed with 7.5 μl ice-cold 0.1 M 1-methylimidazole (1-Melm; Sigma) and 25 μl ice-cold 40 mM 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide (EDC; Sigma) in 10 mM 1-Melm, and the mixture was quickly transferred to a Nucleolink tube (Nunc). The tube was sealed with the Biomek aluminum foil lid (Beckman). The DNA was covalently bound to the tube by incubating the tube at 50 °C for 7 h, and the tube was washed to remove unbound BAC DNA as described by Childs *et al.* (2001).

The first round of selection followed the procedure of Childs *et al.* (2001) except that 2 μl (ca. 100 ng) denatured cDNA was used in hybridization. After hybridization and washes, 100 μl H_2O was added to the tubes and the tubes were heated at 98 °C for 5 min to release the hybridized cDNA. cDNA was amplified (2nd PCR) as described in the 1st PCR except that the reaction was carried out in a 50 μl reaction containing 20 μl selected cDNA, 2 units of Taq DNA polymerase, and 0.04 units of Pfu DNA polymerase. Three microliters of the 2nd PCR product was amplified again (3rd PCR) using the same conditions as in the 1st PCR. After amplification, the PCR product was purified and used for the second round of selection. The second round of selection was conducted according to Childs *et al.* (2001) except that ca. 100 ng of cDNA was used in hybridization. The selected cDNA was released and amplified (4th PCR) using the same conditions as described in the 2nd PCR. The 4th PCR product was purified as described above and used for ligation.

Membrane-based cDNA selection was performed essentially as described by Parimoo (1997) with some minor modifications. Approximately 15 ng of each *Eco*RI-digested BAC DNA sample (without carrier DNA) was immobilized onto small pieces (2.5 × 2.5 mm) of Biodyne B nylon membrane. The first round of selection was performed using 2 μl amplified cDNA (ca.100 ng) and either C_0t-1 DNA of *P. inflata* S_2S_2 genotype (at a final concentration of 50 ng/μl in prehybridization solution and 25 ng/μl in hybridization solution) or salmon sperm DNA (at a final concentration of 200 ng/μl in both prehybridization and hybridization solutions) as a blocking reagent. After the primary selection, hybridized cDNA was released and amplified twice (2nd PCR and 3rd PCR) using the same conditions as described in the nucleolink tube-based cDNA selection. cDNA (80 ng) purified from the 3rd PCR was used in the second round of selection as described in the primary selection. The selected cDNA was released and subjected to two consecutive rounds of PCR amplifications (4th and 5th PCRs) using the same conditions as in the 2nd and 3rd PCRs. The 5th PCR product was then purified and used for ligation.

DNA blot analysis of effectiveness of cDNA selection

cDNA samples (0.5 μl each) obtained before selection and after the first and second rounds of selection were fractionated on 1% agarose gels. The DNA blot was prepared and hybridized with a radiolabeled probe as described by Wang *et al.*

(2003), except that no salmon sperm DNA was added to the hybridization solution when a mixture of 18S rDNA and 26S rDNA was used as a probe.

Colony lift hybridization and dot blot analyses of cDNA clones obtained by cDNA selection

The cDNA selected from each of the eight BAC clones was ligated to the pGEM-T Easy vector, and the ligated cDNA was transformed into *E. coli* (XL1-Blue) cells. The recombinant clones obtained for each BAC clone were transferred into a 384-well plate with each well containing 80 μl LB freezing buffer (McCubbin *et al.*, 2000b) plus 100 μg/ml ampicillin. The clones from each plate were spotted onto a Biodyne B membrane, which had been placed onto the surface of an LB agar plate containing 100 μg/ml ampicillin. The colony lifts were prepared according to the procedure of Woo *et al.* (1994), and were hybridized with a radiolabeled probe as described by Wang *et al.* (2003).

For dot blot analysis, ca. 51 clones randomly chosen from each 384-well plate were separately amplified by colony PCR. PCR was performed under standard conditions using 1 μl bacterial suspension as template. For each PCR product, 1 μl each was spotted onto eight Biodyne B membranes, with each membrane containing the PCR products from the same 407 clones. The membranes were air dried, denatured with 0.5 N NaOH, 1.5 M NaCl for 5 min, neutralized with 1.5 M NaCl, 0.5 M Tris–HCl (pH 7.5) for 5 min, and rinsed with 2× SSC for 5 min. Each blot was first hybridized with DNA from one of the eight BAC clones to confirm that every cDNA clone was selected from its corresponding BAC clone, following the procedure of Wang *et al.* (2003). The blots were then separately hybridized with 18S and 26S rDNA, *S_2-RNase*, *$PiSLF_2$* and a 6.4-kb retrotransposon fragment to eliminate those cDNA clones that were derived from these genes. The retrotransposon fragment was isolated from 114G8 by PCR and it corresponded to 132794–138695 bp of the 328-kb sequence (see 'Sequencing of BAC clones 120K17, 114G8 and 145J16' of Results). For each BAC clone, one of the 'non-positive' cDNA clones was randomly chosen for sequencing and then used as a probe to hybridize with the dot blot to eliminate other cDNA clones derived from the same gene. After hybridization, another non-positive clone was randomly chosen for sequencing and hybridization. This step was repeated until all the 407 cDNA clones were analyzed.

RT-PCR analysis of expression of genes identified by cDNA selection

Total RNA was separately isolated from young leaves, young flower buds (0.5 cm in size), mature pistils of flowers 1–2 days before anthesis, and mature pollen of open flowers using TRIzol reagent. To remove any contaminating genomic DNA, the RNA samples were treated with RQ1 RNase-Free DNase (Promega). cDNA was synthesized from 5 μg of total RNA using SuperScript II RNase H^- Reverse Transcriptase (Invitrogen) and oligo(dT) as a primer. Standard PCR was performed using 0.5 μl of cDNA as template. For each gene analyzed, the corresponding cDNA clone (1 ng) and BAC DNA (0.5 ng), as well as *P. inflata S_2S_2* genomic DNA (0.5 μg), were used as positive controls, and all the PCR products were fractionated on 2% agarose gels.

The sequences of the primers used and the expected sizes of the RT-PCR products for the genes analyzed are listed in supplementary Table 1. *S_2-RNase* and an actin gene of *P. inflata* were used as controls. The primers for these two genes were designed so that the PCR products from genomic DNA contained an intron and were thus larger than the corresponding RT-PCR products.

Table 1. Ten separate contigs of the *P. inflata* S_2 locus.

Contig	Size (kb)
S_2-RNase[a]	881
3.2	165
3.15	449
A113	239
A134	594
A181 + *X9*[a,b]	710
A301[a]	587
G261	142
X9	45
X11[c]	611

[a] No BAC clones overlapping with one end of the contig were found.

[b] Contig contains both *A181* and *X9*.

[c] No BAC clones overlapping with either end of the contig were found.

Results

Chromosome walking in the S_2*-locus region of* P. inflata

Chromosome walking was initiated from multiple sites represented by the previously isolated BAC clones that contained either *S_2-RNase* or one of the nine marker genes tightly linked to the *S* locus. For each BAC clone used, both terminal ends (ca. 700 bp) were sequenced and a pair of PCR primers for each end fragment was designed. The BAC DNA prepared from each of the 179 plate pools of the *S_2S_2* BAC library was then used for PCR screening. For each positive plate pool identified, all the row pools and column pools of the set containing the positive plate were screened by PCR to identify the positive clone(s). The results for the screening using BAC clone 120K17 are shown in Figure 1.

After all the positive clones were identified, they, along with the starting BAC clone, were separately digested with *Bam*HI, and the digests were fractionated by PFGE. The overlapping clones were confirmed based on similarity between their fingerprint patterns and that of the starting BAC clone. Both 5′ and 3′ end fragments of the overlapping clone which extended farthest from the starting BAC clone were then isolated by TAIL-PCR, and used as probes to hybridize separately with a BAC DNA blot containing digests of the starting BAC clone and all the overlapping BAC clones. This further confirmed that the clone chosen for next round of walking overlapped with the starting BAC clone, and allowed the identification of its correct end for further walking. The correct end was sequenced and a primer pair was designed for the next round of library screening.

Because of the presence of highly repetitive sequences in the *S* locus, 10% or higher of the plate pools were found to be positive for more than 75% of the BAC clones used for PCR screening. In these cases, we used several strategies to identify unique or low-copy sequences for designing primers to distinguish the true positive pools from those false pools. These included comparison of the terminal end sequence of a BAC clone with the 328-kb sequence (see Figure 2 and the next section) containing *S_2-RNase*, further sequencing of the terminal end, subcloning of the terminal end fragment for further sequencing, and restriction digestion of PCR products of pools.

Ten separate contigs, collectively spanning 4423 kb, were constructed for *S_2-RNase* and each of the nine marker genes (Table 1). Only one of the contigs contained more than one of the marker genes; it contained both *A181* and *X9*. The 881-kb contig containing *S_2-RNase* (with ca. 180 kb upstream and ca. 700 kb downstream of *S_2-RNase*) is shown in Figure 2. Further walking to fill the gaps between contigs was not successful, because BAC clones overlapping with one or both ends of some of the contigs could not be found and because both ends of the other contigs were very rich in

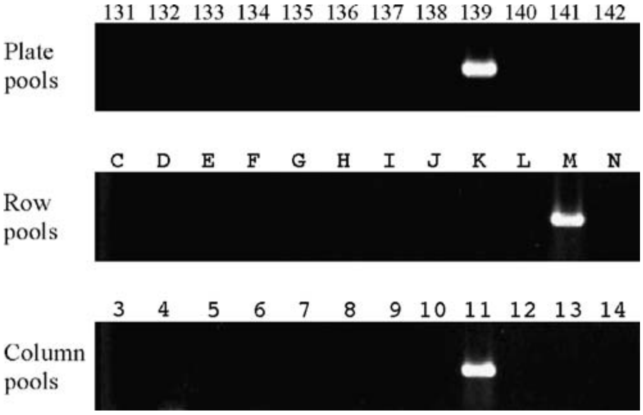

Figure 1. Representative results of PCR-based screening of the *S_2S_2* BAC library. Plate pools nos. 49–179 were screened by PCR using a primer pair designed based on the 5′ end sequence of BAC clone 120K17. One positive plate pool (no. 139) was identified. The row pools and column pools of the plate set no. 8, which contains plate no. 139, were then screened by PCR. The row pool no. M and column pool no. 11 were found to be positive pools. Therefore, the positive clone was located at 139M11.

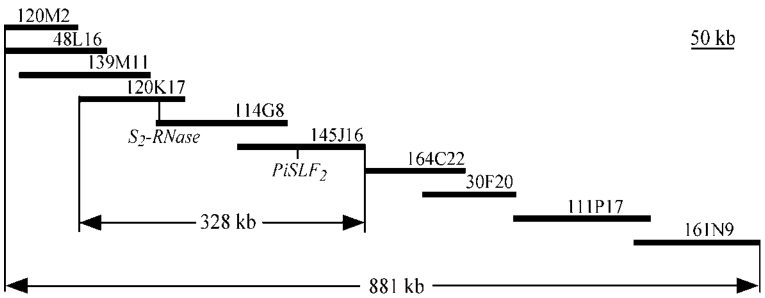

Figure 2. Schematic representation of an 881-kb BAC contig. The locations of *S_2-RNase*, *PiSLF$_2$*, and the completely sequenced 328-kb region are indicated. The lines are drawn to scale.

highly repetitive sequences (Table 1). Because the nine marker genes used in chromosome walking were genetically mapped to the *S* locus (Wang *et al.*, 2003), their contigs, along with the 881-kb contig containing *S_2-RNase*, should be located within the S_2-locus region defined by two recombination breakpoints, one between *3.16* and *S-RNase*, and the other between *G221* and *S-RNase*. Therefore, the 4.4-Mb genomic region collectively spanned by these 10 contigs should reflect the minimum physical size of the S_2 locus delimited by *3.16* and *G221*.

Sequencing of BAC clones 120K17, 114G8 and 145J16

To identify the pollen *S* gene and additional genes located near the *S-RNase* gene, we completely sequenced three overlapping BAC clones, 120K17, 114G8 and 145J16 (Figure 2). The entire sequence (328 473 bp; accession no. AY136628) contained 90 kb of the upstream region and 238 kb of the downstream region of *S_2-RNase*. The details of the sequence analysis described below are graphically presented in supplementary Figure 1.

This 328-kb region had a GC content of 40.95%. RepeatMasker (http://ftp.genome.washington.edu/cgi-bin/RepeatMasker) was used to identify interspersed repeats and low-complexity sequences, similar to those in the *Arabidopsis* database. A total of 48 transposon-like sequences (14.78% of the 328-kb sequence) were identified, and all of them were truncated when compared with those of *Arabidopsis*. Forty-five of them were Gypsy-type and Copia-type LTR (long terminal repeat) retroelements, and three were HAT-type DNA elements. A total of 27 simple repeats and 58 low-complexity sequences were identified, and they represented 1.42% of the 328-kb sequence. The 328-kb sequence was then compared with itself using PipMaker (http://www.cse.psu.edu/pipmaker) to confirm the repetitive sequences found by RepeatMasker and to identify additional repeats present only in the 328-kb region. In total, ca. 76% of this region contained repetitive sequences, including transposon-like sequences. The repetitive sequences and unique sequences were not evenly distributed. Many repetitive sequences were clustered in the 73.5-kb region from the 5′ end of 120K17, whereas a long stretch of unique sequence was found in the region from 73.5 to 93.0 kb, where *S_2-RNase* (at 90 kb from the 5′ end) is located. In the rest of the 328-kb region, repetitive sequences were scattered and interspersed with unique sequences.

The 328-kb sequence was further analyzed using several software packages. AutoPredLTR (http://ricegaas.dna.affrc.go.jp/index.html; Sakata *et al.*, 2002) predicted 21 pairs of direct LTRs, which are shown in supplementary Table 2. GENSCAN (http://bioweb.pasteur.fr/seqanal/interfaces/genscan.html) predicted 50 genes, and their deduced amino acid sequences were used to search the non-redundant protein sequence database using BLASTP (http://www.ncbi.nlm.nih.gov/BLAST/). The results are summarized in Table 2. The deduced amino acid sequences of 33 of the 50 predicted genes showed similarity ($E \geq e^{-4}$) to the sequences of known proteins in the database; however, all but two were highly similar to those of gag-pol polyproteins of various retroelements or to those of putative transposases of *Arabidopsis* DNA elements. The exceptions were Gene 12 and Gene 38. The last two exons (exons 8 and 9) of Gene 12 corresponded to the two exons (located from 89 623 to 89 861 and from 89 968 to

Table 2. Summary of genes predicted by GENSCAN in the 328-kb sequence.

Predicted gene	Peptide size (aa)/exons	Putative homologue, accession number (*E* value)
Gene 1	897/9	*Oryza sativa* putative retroelement, BAC65423 (e^{-128})
Gene 2	3405/20	*Arabidopsis* putative retroelemen, NP_174802 (0)
Gene 3	2248/9	*Oryza sativa* putative retroelement, AAP52850 (7e^{-71})
Gene 4	755/3	*Oryza sativa* putative retroelement, AAP52850 (e^{-180})
Gene 5	1591/4	*Oryza sativa* putative retroelement, AAP52850 (0)
Gene 6	237/2	*Oryza sativa* putative retroelement, CAE02466 (7e^{-35})
Gene 7	1320/5	*Oryza sativa* putative retroelement, CAE02466 (0)
Gene 8	133/1	
Gene 9	292/5	*Oryza sativa* putative retroelement, CAD40072 (e^{-18})
Gene 10	190/2	*Arabidopsis* Mutator-like transposase, AAD31079 (e^{-30})
Gene 11	1690/14	*Arabidopsis* putative retroelement, NP_680252 (0)
Gene 12[a]	901/9	*Petunia inflata* S_2-RNase, AAG21384 (e^{-98})
Gene 13	1451/9	*Arabidopsis* putative transposase, AAD24567 (e^{-138})
Gene 14[a]	103/1	
Gene 15	123/3	
Gene 16	739/3	*Oryza sativa* putative retroelement, CAD37108 (0)
Gene 17	1067/12	*Oryza sativa* putative retroelement, AAP53706 (e^{-85})
Gene 18	580/3	*Oryza sativa* putative retroelement, AAP51828 (9e^{-36})
Gene 19	2068/9	*Oryza sativa* putative retroelement, AAP52850 (0)
Gene 20	354/5	
Gene 21	1659/10	*Oryza sativa* putative retroelement, AAP52850 (0)
Gene 22	135/3	
Gene 23	193/4	*Oryza sativa* putative retroelement, CAE05407 (e^{-16})
Gene 24[a]	198/1	
Gene 25	1081/3	*Zea mays* retroelement, AAM94350 (0)
Gene 26	1402/8	*Hordeum vulgare* retroelement, AAK94516 (e^{-4})
Gene 27[a]	146/3	*Arabidopsis* retroelement, NP_173464 (7e^{-12})
Gene 28	197/4	
Gene 29	333/8	*Solanum tuberosum* callus EST, CK247373 (3e^{-5})
Gene 30[a]	171/2	
Gene 31	817/5	*Petunia hybrida* unknown protein, AAQ72728 (e^{-31})
Gene 32	983/5	*Oryza sativa* putative retroelement, CAE02877 (3e^{-35})
Gene 33[a]	85/2	*Lotus japonicus* EST, AV419137 (8e^{-4})
Gene 34	278/6	
Gene 35	494/7	*Arabidopsis* putative retroelement, AAF79348 (8e^{-51})
Gene 36	63/2	
Gene 37	2063/5	*Zea mays* retroelement, AAD20307 (0)
Gene 38[a]	389/1	*Antirrhinum hispanicum* SLF-S2, CAC33022 (2e^{-39})
Gene 39	608/3	*Oryza sativa* putative retroelement, CAE02308 (e^{-59})
Gene 40	281/3	*Gossypium hirsutum* retroelement, AAP43918 (2e^{-19})
Gene 41	1950/15	*Oryza sativa* putative retroelement, AAP52384 (0)
Gene 42	791/12	*Lycopersicon pennellii* pollen EST, BG140262 (e^{-13})
Gene 43[a]	489/1	*Nicotiana tabacum* BY-2 EST, BP135940 (3e^{-4})
Gene 44	198/1	*Arabidopsis* putative retroelement, BAB02630 (2e^{-57})
Gene 45	1684/3	*Zea mays* retroelement, AAM94350 (0)
Gene 46[a]	426/5	*Vicia faba* putative retroelement, BAA22787 (e^{-114})
Gene 47	356/1	*Oryza sativa* putative retroelement, CAE02308 (5e^{-43})
Gene 48[a]	101/1	
Gene 49	279/3	
Gene 50[a]	1301/16	*Oryza sativa* putative retroelement, BAB08213 (9e^{-48})

[a]Located in regions of unique sequences.

90 542 bp in the 328-kb sequence) of *S_2-RNase*. The predicted open-reading frame of Gene 38 was located from 251 869 to 253 038 bp, 161 kb downstream of *S_2-RNase*, and it was most similar to *AhSLF-S_2* (*S_2*-allele of *A. hispanicum S*-locus F-box gene; Zhou *et al.,* 2003). Gene 38 was

named *PiSLF*$_2$ (*S*$_2$-allele of *P. inflata S*-locus F-box gene) because, like *AhSLF*$_2$, its deduced amino acid sequence contained an F-box domain at the N-terminus.

BLASTP searches showed that one of the 50 predicted genes (Gene 31) was most similar to an unknown protein of *P. hybrida*. For the 16 predicted genes for which no putative homologs were found through BLASTP searches, their predicted coding sequences were used in the BLASTN and TBLASTX searches of the EST database. Four of them (Genes 29, 33, 42 and 43) showed similarity ($E \leq 8e^{-4}$) to EST sequences in the database. The remaining 12 showed no similarity to any EST sequences: eight were located in regions of repetitive sequences, whereas the other 4 (Genes 14, 24, 30 and 48) were located in regions of unique sequences.

Identification of additional genes in the S$_2$*-locus region by cDNA selection*

To examine the authenticity of the non-transposon-like genes predicted by GENSCAN in the 328-kb region and to identify additional genes in the regions flanking the 328-kb region, cDNA selection was conducted in the 881-kb contig (Figure 2) using a membrane-based method and a nucleolink-tube based method. We first tested the effectiveness of these two methods by DNA blot analysis using *S*$_2$*-RNase* (located in two overlapping BAC clones, 120K17 and 114G8) and the predicted *PiSLF*$_2$ gene (located in BAC clone 145J16) as probes. DNA blots containing cDNAs selected from these three BAC clones by the membrane-based method, as well as DNA blots containing cDNAs selected from these three BAC clones and BAC clone 48L16 by the nucleolink tube-based method, were separately hybridized with three probes: a mixture of 18S and 26S rDNA, *PiSLF*$_2$, and *S*$_2$*-RNase*. The hybridization results for the former two probes are shown in Figure 3. After two rounds of selection by either method, the level of cDNA for 18S and 26S rRNA was drastically reduced or virtually undetectable. In contrast, after the first round and second round of selection, the level of *PiSLF*$_2$ cDNA selected from 145J16 was significantly enriched. Moreover, as expected, only the cDNAs selected from 145J16 hybridized to *PiSLF*$_2$. Similarly, the level of *S*$_2$*-RNase* cDNA selected from 120K17 and 114G8 was significantly enriched, and only the cDNAs selected from these two BAC clones hybridized to *S*$_2$*-RNase* (results not shown). These results taken together showed that both cDNA selection methods were effective in selecting the cDNA only from

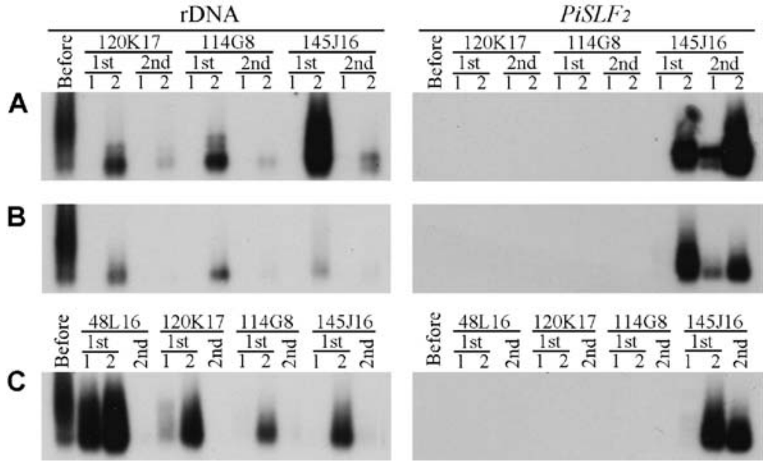

Figure 3. DNA blot analysis of cDNAs selected by the membrane-based or nucleolink tube-based method. (A, B) Blots containing cDNAs before and after selection from three overlapping BAC clones (120K17, 114G8 and 145J16) using the membrane-based method. Either C_0t-1 DNA isolated from *P. inflata* S_2S_2 genotype (A) or salmon sperm DNA (B) was used as a blocking reagent. (C) Blot containing cDNAs before and after selection from four overlapping BAC clones (48L16, 120K17, 114G8 and 145J16) using the nucleolink tube-based method, with salmon sperm DNA as a blocking reagent. Numbers '1' and '2' under '1st' and '2nd' indicate the first and second of two consecutive PCR amplifications after first and second rounds of cDNA selection. The blots were separately hybridized with radiolabeled *PiSLF*$_2$ and a mixture of 18S and 26S rDNA.

the BAC clone(s) that contained the corresponding gene. Moreover, the selection of *PiSLF*$_2$ cDNA suggests that the predicted *PiSLF*$_2$ gene was indeed expressed.

Colony lift hybridization was performed to further examine the effectiveness of the two cDNA-selection methods and different blocking reagents. The colony lifts containing cDNA clones selected from 120K17, 114G8 and 145J16 were separately hybridized with S_2-*RNase* and *PiSLF*$_2$. The hybridization results are summarized in Table 3. The membrane-based method with the C_0t-1 DNA as a blocking reagent was most effective in selecting the 'correct' cDNA and blocking repetitive sequences present in the BAC clones. For example, when 120K17 was used for selection, this combination yielded the largest number (54) of cDNA clones for S_2-*RNase*. This optimal condition was then used in cDNA selection of the remaining five BAC clones of the 881-kb contig.

Dot blot analysis in conjunction with sequencing was used to analyze 407 cDNA clones, with ca. 51 cDNA clones randomly chosen from those selected from each of the eight BAC clones. To determine whether a cDNA clone chosen for sequencing contained repetitive sequences, the clone was used as a probe to hybridize with a dot blot containing DNA of all the eight BAC clones used in selection. The sequence was also compared with itself, its reverse complement sequence, and the 328-kb sequence. A cDNA clone was determined to contain repetitive sequences, if it hybridized to two or more non-overlapping BAC clones, its sequence showed high similarity to at least two different regions of the 328-kb sequence, and/or it contained direct or inverted repeats. All the cDNA sequences were also used in BLASTX searches of the protein database, and BLASTN and TBLASTX searches of the EST database to determine whether they showed similarity to transposons or known genes.

The results of the dot blot analysis of the 407 clones are summarized in Table 4. The cDNA selection was effective because all these clones were

Table 3. Number and percentage (in parentheses) of positive cDNA clones for S_2-*RNase* and *PiSLF*$_2$ obtained under different cDNA selection conditions.

	BAC 120K17				BAC 114G8				BAC 145J16			
	Nucleolink tube		Membrane		Nucleolink tube		Membrane		Nucleolink tube		Membrane	
Gene	Salmon DNA	C_0t-1 DNA	Salmon DNA	C_0t-1 DNA	Salmon DNA	C_0t-*1* DNA	Salmon DNA	C_0t-*1* DNA	Salmon DNA	C_0t-1 DNA	Salmon DNA	C_0t-1 DNA
S_2- *RNase*	27 (7.0%)	Not tested	4 (1.0%)	54 (14.1%)	7 (1.8%)	Not tested	7 (1.8%)	51 (13.3%)	0 (0)	Not tested	0 (0)	0 (0)
PiSLF$_2$	0 (0)	Not tested	0 (0)	0 (0)	0 (0)	Not tested	0 (0)	0 (0)	1 (0.3%)	Not tested	2 (0.5%)	37 (9.6%)

Table 4. Dot blot analysis of 407 cDNA clones selected from eight S_2S_2 BAC clones comprising the 881-kb contig.

BAC clone	Total cDNA clones analyzed	No of cDNA clones containing repetitive sequences[a] (% of total)	No of cDNA clones containing entirely unique or low-copy sequences (% of total)
48L16	48	48 (100%)	0 (0)
120K17	61	28 (46%)	33 (54%)
114G8	53	11 (21%)	42 (79%)
145J16	52	17 (33%)	35 (67%)
164C22	48	36 (75%)	12 (25%)
30F20	48	2 (4%)	46 (96%)
111P17	48	46 (96%)	2 (4%)
161N9	49	37 (77%)	11 (23%)

[a]Including retrotransposon-like sequences.

indeed selected from their corresponding BAC clones and none were derived from rDNA. At least 23% of the cDNA clones selected from each BAC clone contained entirely unique or low-copy sequences, except for those selected from 48L16 (0%) and 111P17 (4%). These results further confirm that the C_0t-1 DNA used was effective in blocking the repetitive sequences of the BAC clones.

Excluding the cDNA clones for *S_2-RNase* and *PiSLF$_2$*, a total of 62 different cDNA clones were sequenced in the process of dot blot analysis. Forty-seven (76%) contained repetitive sequences or were derived from retrotransposons, and the other 15 contained entirely unique or low-copy sequences. The 15 clones were classified into 11 groups based on their sequences. These 11 groups of cDNA clones, as well as cDNA clones for *S_2-RNase* and *PiSLF$_2$*, are listed in Table 5. BLASTX searches of the protein database and TBLASTX searches of the EST database revealed that three of the 11 groups (3-A12, 6-F19 and 6-G22) showed high similarity ($E \leq 8e^{-14}$) and the remaining eight did not show any significant similarity to known proteins or EST sequences in the database.

Among the 11 groups of cDNA clones, only two were derived from the 328-kb region: 2-C2 (114 bp) selected from 120K17, and 3-A12 (834 bp) selected from 114G8 and 145J16. Neither cDNA clone corresponded to any of the non-transposon-like genes predicted by GENSCAN (Table 2), suggesting that these predicted genes might not be real genes. 2-C2, corresponding to 75 707–75 820 bp of the 328-kb sequence, showed no significant similarity to any sequence in the GenBank. The five cDNA clones in the 3-A12 group collectively spanned the region of 210 769–211 602 bp. The longest open-reading frame, 347–556 bp, was preceded and followed by a stop codon. BLASTX searches of the protein database and TBLASTX searches of the EST database showed that the deduced amino acid sequence of this open-reading frame was highly similar to a region (amino acid residues 211–274) of an *Arabidopsis* putative GTPase (accession number AAK96878; $E = 2e^{-21}$) and to a tomato EST (accession number BG127259; $E = 5e^{-30}$). However, the deduced amino acid sequences of the flanking regions (1–346 and 557–834 bp) showed no significant similarity to any known protein sequences. These results suggest that 3-A12 was likely derived from a pseudogene that contained a truncated *GTPase* coding sequence. Therefore, excluding *S_2-RNase*, only two new genes, *PiSLF$_2$* and *2C2*, were found in the 328-kb region.

RT-PCR analysis of expression of the genes identified by cDNA selection

For 11 of the 12 new genes (except for *PiSLF* whose characterization is reported in Sijacic *et al.*, in press) identified by cDNA selection, RT-PCR was performed to examine their expression in

Table 5. Summary of 13 groups of cDNA clones obtained from the 881-kb contig by cDNA selection.

cDNA group	Size (bp)	Source of BAC clone	No of positive clones (% total)[a]	Putative homologue, accession number (*E* value)
2-C2	114	120K17	7 (12%)	
S_2-RNase		120K17, 114G8	45 (40%)	*Petunia inflata* S_2-RNase, AY136628
3-A12	834[b]	114G8, 145J16	29 (28%)	*Arabidopsis* putative GTPase, AAK96878 ($2e^{-21}$)
PiSLF$_2$		145J16	15 (29%)	*Antirrhinum hispanicum* SLF-S_2, CAC33022 ($2e^{-39}$)
5-F13	254	164C22	9 (19%)	
5-G24	219	164C22	3 (6%)	
6-F19	315	30F20	1 (2%)	*Nicotiana benthamiana* EST, CK290013 (e^{-20})
6-G7	293	30F20	44 (92%)	
6-G22	284	30F20	1 (2%)	*Capsicum annuum* putative NBS/LRR resistance protein, AAM47598 ($8e^{-14}$)
7-F24	334	111P17	2 (4%)	
8-A7	328	161N9	7 (14%)	
8-A8	208	161N9	2 (4%)	
8-A19	267	161N9	2 (4%)	

[a] Determined by dot blot analysis.

[b] Contig size of five different cDNA clones.

young flower buds, leaves, mature pistils, and mature pollen. For these 11 genes, as well as S_2-*RNase* and an actin gene used as controls, no bands were detected when RT-PCR was conducted without reverse transcriptase (data not shown). Moreover, for the actin gene, RT-PCR yielded only one band of the expected size (611 bp) in all the tissues examined, and for S_2-*RNase*, RT-PCR yielded only one band of the expected size (341 bp) in flower buds and pistils (Figure 4). These results suggest that none of the RNA samples were contaminated with any genomic DNA.

The intensities of the RT-PCR products for the actin gene in all the tissues examined were similar, consistent with the ubiquitous nature of this gene (Figure 4). For 2-C2, 3-A12, 5-F13, 5-G24, 6-F19, 8-A7 and 8-A19, only one band of the expected size was detected. Moreover, for each of these cDNA groups, the size of the RT-PCR product was similar to those of the PCR products from the corresponding cDNA and BAC clones, and from the S_2S_2 genomic DNA. These results confirm that all these cDNAs were derived from their corresponding genes in the respective BAC clone(s), and

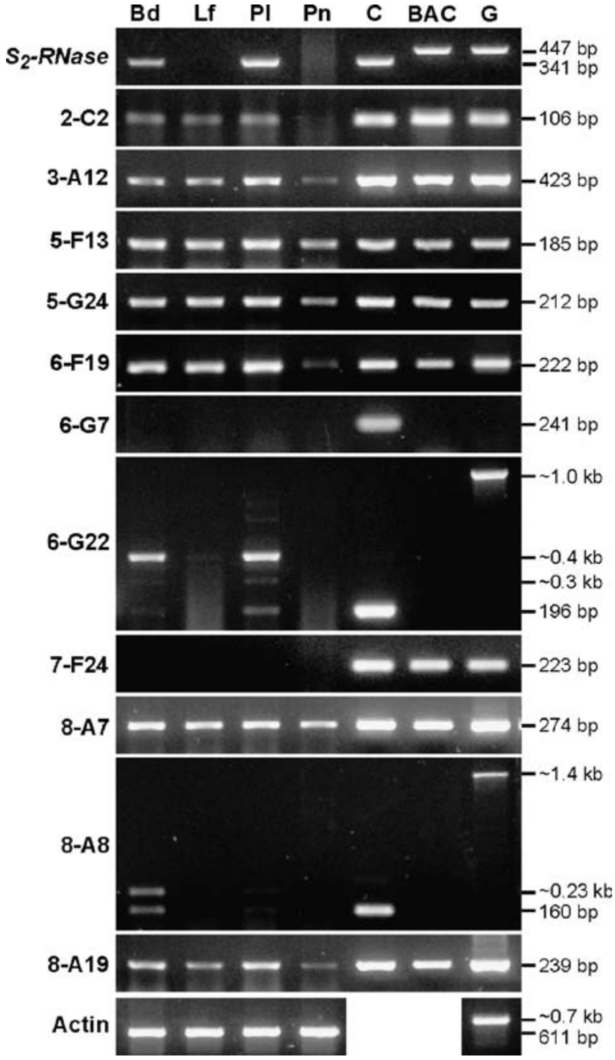

Figure 4. RT-PCR analysis of expression of 11 genes identified by cDNA selection. The primer sequences and expected RT-PCR product sizes are shown in supplementary Table 1. The S_2-*RNase* gene and the actin gene were included as controls. Bd, 0.5-cm flower bud; Lf, leaf; Pl, mature pistil; Pn, mature pollen; C, cDNA clone as a positive control; BAC, BAC clone as a positive control; G, S_2S_2 genomic DNA as a positive control.

that none of the regions amplified contained any intron. The 5-F13, 5-G24 and 8-A7 genes were expressed at similar levels in all the tissues examined, whereas the 2-C2, 3-A12, 6-F19 and 8-A19 genes appeared to be more highly expressed in young flower buds, leaves and mature pistils than in mature pollen.

As mentioned earlier, 3-A12 was likely to be expressed from a pseudogene encoding part of GTPase. However, since the 834-bp cDNA was assembled from five overlapping cDNA clones, it was necessary to confirm the validity of the assembly. Three primer pairs (shown in supplementary Table 1) were designed for separate amplification of the 5′ flanking region along with the coding region (Figure 4), the 3′ flanking region, and nearly the full-length of the 834-bp fragment. RT-PCR using these three primer pairs revealed the same expression pattern (Figure 4; data not shown). Moreover, the size of each RT-PCR product was similar to those of the PCR products of 114G8 and S_2S_2 genomic DNA (Figure 4; data not shown), suggesting that the 834-bp cDNA was transcribed as one unit and its gene did not contain any intron in this region.

Both 6-G22 and 8-A8 had a similar expression pattern as *S_2-RNase*, which was expressed in mature pistils and young flower buds, but not in leaves or pollen. For 6-G22, three bands were detected in the RT-PCR products of flower buds and mature pistils: one strong band (ca. 0.4 kb) and two weak bands (196 bp and ca. 0.3 kb). The 196-bp fragment matched the PCR product of the 6-G22 cDNA clone. Only one band (ca. 1.0 kb) was observed in the genomic DNA control. For 8-A8, two different bands (160 bp and ca. 0.23 kb) were detected in the RT-PCR products of flower buds and pistils, and only the smaller one matched the PCR product of the 8-A8 cDNA clone. Like 6-G22, only one band (ca. 1.4 kb) larger than the RT-PCR products was detected in the genomic DNA control. Although 6-G22 and 8-A8 cDNA fragments strongly hybridized to the respective BAC clones, 30F20 and 161N9, from which they were selected (results not shown), no band was detected from either BAC clone by PCR (Figure 4). These results suggest that 6-G22 and 8-A8 might have been selected from the respective BAC clones due to their sharing sequence similarity with some related sequences.

For 7-F24, no RT-PCR products were detected in any of the tissues examined, even though the 7-F24 cDNA clone, BAC clone 111P17 and genomic DNA all produced one band of the expected size (223 bp). It is possible that the expression level of the corresponding gene was too low to be detected. For 6-G7, only the cDNA clone control produced one band of the expected size. The same result was obtained using another pair of primers for 6-G7. One possible explanation is that the 6-G7 cDNA clone had accumulated many mutations as a result of five rounds of PCR amplification, so that the PCR primers, designed based on the sequence of the cDNA clone, failed to anneal to the authentic cDNA, BAC DNA, or genomic DNA.

Discussion

In this work, we conducted chromosome walking in the S_2-locus region of *P. inflata* and determined, based on the sizes of the 10 separate BAC contigs assembled, that this locus is at least 4.4 Mb in size. To identify the genes that are close to *S_2-RNase*, we completely sequenced a 328-kb region containing *S_2-RNase* and performed cDNA selection in an 881-kb contig encompassing the 328-kb region. A total of 12 new genes were identified, and the expression patterns of 11 of them (see Sijacic *et al.*, in press for the analysis of *PiSLF*) were analyzed by RT-PCR. Eight (*2-C2*, *5-F13*, *5-G24*, *6-F19*, *7-F24*, *8-A7*, *8-A19*, and *PiSLF*) were confirmed to be *bona fide* genes located in the 881-kb region.

Genomic organization of the P. inflata S *locus*

Sequence analysis of the *S*-locus region has previously been reported for several self-incompatible species, *Brassica campestris* and *B. napus* (Brassicaceae), *A. hispanicum* (Scrophulariaceae), and *Prunus dulcis* and *P. mume* (Rosaceae) (Cui *et al.*, 1999; Suzuki *et al.*, 1999; Lai *et al.*, 2002; Entani *et al.*, 2003; Ushijima *et al.*, 2003). The regions analyzed range from 63 to 88 kb, and thus, the 328-kb sequence we have analyzed represents by far the most extensive sequence information available for the locus that controls SI. Among these *S* loci analyzed, only the Solanaceae *S* locus is known to be located in the sub-centromeric region (Entani *et al.*, 1999). This may explain why the *S* locus of *P. inflata* (>4.4 Mb) is much larger than that of members of the other families. For example, recombination analysis in combination with physical mapping and pollination analysis has delimited the S_8 locus of *B. campestris* to a 50-kb region, which contains both the male and female determinant genes (Casselman *et al.*, 2000). Moreover, the S^c locus of *P. dulcis* has been estimated by genomic DNA blot analysis to be ca. 70 kb, because the sequence within this region is highly divergent between different *S* haplotypes and the sequences flanking this region are similar between different *S* haplotypes (Ushijima *et al.,* 2001). Lastly, comparison of genomic sequences of the S_1 and S_7 loci of *P. mume* has identified a highly polymorphic region in both loci, 27 kb in the S_1 locus and 15 kb in the S_7 locus, which is flanked by highly conserved regions (Entani *et al*., 2003). However, the sizes of the *S* loci in these species of the Rosaceae have not been genetically defined.

We previously mapped the *S* locus of *P. inflata* to within a 0.25-centiMorgan region, which contained all the genes for determining SI specificity as demonstrated by pollination analysis (Wang *et al*., 2003). Since we have determined the physical size of the S_2 locus to be at least 4.4 Mb, the ratio of the physical distance to the genetic distance for this region is at least 17.6 Mb/cM. This is much higher than the genome average of 750 kb/cM for tomato (Tanksley *et al*., 1992), but is similar to 25 Mb/cM in the centromere of tomato chromosome 12 where the *jointless-2* gene is located (Budiman *et al*., 2004). However, the ratio for the centromeres of five chromosomes of *Arabidopsis* (with a genome size of 125 Mb; *Arabidopsis* Genome Initiative, 2000) is only 2.1–7.6 Mb/cM (Copenhaver *et al*., 1999), presumably as a result of the considerably smaller genome of this species.

The 328-kb region contains a much higher percentage (ca. 76% of total) of highly repetitive sequences (including transposon-like sequences) than the 10–15% estimated for the tomato genome (Ganal *et al*., 1988). It is interesting to note that the repetitive sequences in the 328-kb region are not evenly distributed. They are clustered in the first 73.5-kb region (upstream from S_2-*RNase*), absent in the region between 73.5 and 93.0 kb where S_2-*RNase* is located, and dispersed in the rest of the region (93.0–328 kb; see supplementary Figure 1).

The 328-kb region also contains a higher percentage of transposon-like sequences (mainly retrotransposons) than the *S*-locus regions of the other species sequenced. Of the 50 genes predicted by GENSCAN in the 328-kb region, the putative identity of 33 was determined by similarity of their deduced amino acid sequences to known proteins in the database; all but two of these 33 predicted genes showed high similarity to transposons. In contrast, only one of the 14 predicted genes in the 64-kb region of the S_1 locus of *P. mume* was highly similar to transposons (Entani *et al.,* 2003). This difference could reflect the centromeric location of the Solanaceae *S* locus, because a large number of retrotransposon sequences are also present in the centromeric regions of *Arabidospis* (*Arabidopsis* Genome Initiative, 2000).

The two predicted genes not encoding transposon-like proteins are S_2-*RNase* and *PiSLF*$_2$. cDNA selection also identified these two genes in the 328-kb region, as well as two other genes (*2-C2* and *3-A12*) that were not predicted by GENSCAN. Nine additional genes were identified by cDNA selection outside the 328-kb region but within the 881-kb contig. Therefore, a total of 13 genes (including S_2-*RNase*) were identified in the 881-kb contig by cDNA selection. However, one (*3-A12*) was found to be a pseudogene by sequence analysis, and three (*6-G7*, *6-G22* and *8-A8*) were found less likely to be located in the 881-kb contig by PCR analysis. Thus, a total of nine *bona fide* genes (the S_2-*RNase* gene and the eight new genes identified in this work) are located in the 881-kb contig. These results suggest that this *S*-locus region is deficient in genes, with a gene density of one gene per 98 kb. If we assume that the *Petunia* genome (1158 Mb; Bennet and Leitch, 1995) contains a similar number of genes as that (ca. 35000) estimated for the tomato genome (Van der Hoeven *et al*., 2002), the average gene density of the *Petunia* genome would be one gene per 33 kb, considerably higher than that found in the 881-kb region. In tomato, the centromeric heterochromatic regions constitute ca. 77% of the chromosomal DNA (Peterson *et al*., 1996). If the *Petunia* genome has the same content of centromeric heterochromatic regions and one gene per 98 kb is the average gene density in these regions, then the average gene density in euchromatic regions would be 10 kb/gene.

In summary, all the features of the 881-kb contig (i.e., low gene density, high percentage of repetitive and transposon-like sequences) and the immense size of the *S*-locus region where recombination is suppressed are consistent with the centromeric location of the Solanaceae *S* locus.

Physiological roles of the newly identified genes

Of the nine genes located in the 881-kb contig, all except 7-F24 are expressed in the pistil and/or pollen (Figure 4; see Sijacic *et al.*, in press for the expression pattern of *PiSLF*). This is reminiscent of the 76-kb region of the S^9 locus of *B. campestris*, where all 11 genes identified are expressed in reproductive tissues (Suzuki *et al.,* 1999). It would be of interest to determine whether any of these genes is involved in reproductive processes. This is possible, considering that genes that control floral traits have been mapped to the *S* locus in tomato (Bernacchi and Tanksley, 1997).

Previously, the large-scale sequencing of the *S*-locus region of the Rosaceae and Scrophulariaceae has revealed a pollen-specific F-box gene named *AhSLF* in *A. hispanicum* (Lai *et al.*, 2002; Zhou *et al.*, 2003), *PdSFB* in *P. dulcis* (Ushijima *et al.*, 2003) and *PmSLF* in *P. mume* (Entani *et al.*, 2003). This gene is close to the *S-RNase* gene (e.g., *AhSLF* is ca. 9 kb from the S_2-*RNase* gene), is specifically expressed in pollen/anthers, and shows allelic sequence diversity. It is thus interesting that we have identified, by direct sequencing and cDNA selection, an F-box gene, *PiSLF*, which is located 161 kb from the *S-RNase* gene of the S_2 locus. Two additional *S*-linked F-box genes of *P. inflata*, named *A113* and *A134*, have previously been identified by mRNA differential display (McCubbin *et al.*, 2000a; Wang *et al.*, 2003); however, their physical distance from the *S-RNase* gene has not been determined in any *S*-genotype. Since *A113* and *A134* are located in two separate contigs that do not overlap with the 881-kb contig, both genes are at least 250 kb and could be up to 4.4 Mb from S_2-*RNase* (Table 1; data not shown). We have recently used a transgenic approach to show that *PiSLF* indeed encodes the pollen determinant of SI (Sijacic *et al.*, in press).

Acknowledgements

We thank Anthony H. Omeis for the greenhouse work and Jiong Wang for the routine lab maintenance. T.T. was partially supported by a Research Fellowship for Young Scientists from the Japan Society for the Promotion of Science. This work was supported by grants from the National Science Foundation (grants IBN-9982659 and IBN-0235176 to T.-h.K).

References

Arabidopsis Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana.* Nature 408: 796–815.

Bennet, M.D. and Leitch, I.J. 1995. Nuclear DNA amounts in angiosperms. Ann. Bot. (Lond.) 76: 113–176.

Bernacchi, D. and Tanksley, S.D. 1997. An interspecific backcross of *Lycopersicon esculentum* × *L. hirsutum*: linkage analysis and a QTL study of sexual compatibility factors and floral traits. Genetics 147: 861–877.

Budiman, M.A., Chang, S.B., Lee, S., Yang, T.J., Zhang, H.B., De Jong, H. and Wing, R.A. 2004. Localization of *jointless-2* gene in the centromeric region of tomato chromosome 12 based on high resolution genetic and physical mapping. Theor. Appl. Genet. 108: 190–196.

Casselman, A.L., Vrebalov, J., Conner, J.A., Singhal, A., Giovannoni, J., Nasrallah, M.E. and Nasrallah, J.B. 2000. Determining the physical limits of the *Brassica S* locus by recombinational analysis. Plant Cell 12: 23–33.

Childs, K.L., Klein, R.R., Klein, P.E., Morishige, D.T. and Mullet, J.E. 2001. Mapping genes on an integrated sorghum genetic and physical map using cDNA selection technology. Plant J. 27: 243–255.

Copenhaver, G.P., Nichel, K., Kuromori, T., Benito, M.-I., Kaul, S., Liu, X., Bevan, M., Murphy, G., Harris, B., Parnell, L.D., McCombie, W.R., Martienssen, R.A., Marra, M. and Preuss, D. 1999. Genetic definition and sequence analysis of *Arabidopsis* centromeres. Science 24: 2468–2474.

Cui, Y., Brugiere, N., Jackman, L., Bi, Y.-M. and Rothstein, S.J. 1999. Structural and transcriptional comparative analysis of the *S* locus regions in two self-incompatible *Brassica napus* lines. Plant Cell 11: 2217–2231.

de Nettancourt, D. 2001. Incompatibility and Incongruity in Wild and Cultivated Plants. Springer-Verlag, Berlin.

Dowd, P.E., McCubbin, A.G., Wang, X., Verica, J.A., Tsukamoto, T., Ando, T. and Kao, T.-h. 2000. Use of *Petunia* as a model for the study of Solanaceous type self-incompatibility. Ann. Bot. 85(Suppl A): 87–93.

Entani, T., Iwano, M., Shiba, H., Che, F.-S., Isogai, A. and Takayama, S. 2003. Comparative analysis of the self-incompatibility (*S*-) locus region of *Prunus mume*: identification of a pollen-expressed F-box gene with allelic diversity. Genes Cells 8: 203–213.

Entani, T., Iwano, M., Shiba, H., Takayama, S., Fukui, K. and Isogai, A. 1999. Centromeric localization of an S-RNase gene in *Petunia hybrida* Vilm. Theor. Appl. Genet. 99: 391–397.

Ganal, M.W., Lapitan, N.L.V. and Tanksley, S.D. 1988. A molecular and cytogenetic survey of major repeated DNA sequences in tomato (*Lycopersion esculentum*). Mol. Gen. Genet. 213: 262–268.

Kao, T.-h. and Tsukamoto, T. in press. The molecular and genetic bases of S-RNase-based self-incompatibility. Plant Cell.

Lai, Z., Ma, W., Han, B., Liang, L., Zhang, Y., Hong, G. and Xue, Y. 2002. An F-box gene linked to the self-incompatibility (*S*) locus of *Antirrhinum* is expressed specifically in pollen and tapetum. Plant Mol. Biol. 50: 29–42.

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

Liu, Y.-G. and Whittier, R.F. 1995. Thermal asymmetric interlaced PCR: automatable amplification and sequencing of insert end fragments from P1 and YAC clones for chromosome walking. Genomics 25: 674–681.

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

McCubbin, A.G., Zuniga, C. and Kao, T.-h. 2000b. Construction of a binary bacterial artificial chromosome library of *Petunia inflata* and the isolation of large genomic fragments linked to the self-incompatibility (*S*-) locus. Genome 43: 820–826.

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

Parimoo, S. 1997. cDNA selection with YACs. Mol. Biotechnol. 8: 255–268.

Peterson, D.G., Price, H.J., Johnston, J.S. and Stack, S.M. 1996. DNA content of heterochromatin and euchromatin in tomato (*Lycopersicon esculentum*) pachytene chromosomes. Genome 39: 77–82.

Sakata, K., Nagamura, Y., Numa, H., Antonio, B.A., Nagasaki, H., Idonuma, A., Watanabe, W., Shimizu, Y., Horiuchi, I., Matsumoto, T., Sasaki, T. and Higo, K. 2002. RiceGAAS: an automated annotation system and database for rice genome sequence. Nucleic Acids Res. 30: 98–102.

Sijacic, P., Wang, X., Skirpan, A.L., Wang, Y., Dowd, P.E., McCubbin, A.G., Huang, S. and Kao, T.-h. in press. Identification of the pollen determinant of S-RNase-mediated self-incompatibility. Nature.

Simmons, A.D. and Lovett, M. 1999. Direct cDNA selection using large genomic DNA targets. Meth. Enzymol. 303: 111–126.

Suzuki, G., Kai, K., Hirose, T., Fukui, K., Nishio, T., Takayama, S., Isogai, A., Watanabe, M. and Hinata, K. 1999. Genomic organization of the *S* locus: identification and characterization of genes in *SLG*/*SRK* region of S^9 haplotype of *Brassica campestris* (syn. *rapa*). Genetics 153: 391–400.

Tanksley, S.D., Ganal, M.W., Prince, J.P., de Vicente, M.C., Bonierbale, M.W., Broun, P., Fulton, T.M., Giovannoni, J.J., Grandillo, S., Martin, G.B., Messerguer, R., Miller, J.C., Miller, L., Paterson, A.H., Pineda, O., Röder, M.S., Wing, R.A., Wu, W. and Young, N.D. 1992. High-density molecular linkage maps of the tomato and potato genomes. Genetics 132: 1141–1160.

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

Ushijima, K., Sassa, H., Tamura, M., Kusaba, M., Tao, R., Gradziel, T.M., Dandekar, A.M. and Hirano, H. 2001. Characterization of the *S*-locus region of almond (*Prunus dulcis*): analysis of a somaclonal mutant and a cosmid contig for an *S* haplotype. Genetics 158: 379–386.

Van der Hoeven, R., Ronning, C., Giovannoni, J., Martin, G. and Tanksley, S. 2002. Deductions about the number, organization, and evolution of genes in the tomato genome based on analysis of a large expressed sequence tag collection and selective genomic sequencing. Plant Cell 14: 1441–1456.

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

Woo, S.-S., Jiang, J., Gill, B.S., Patterson, A.H. and Wing, R.A. 1994. Construction and characterization of a bacterial artificial chromosome library of *Sorghum bicolor*. Nucleic Acids Res. 22: 4922–4931.

Xue, Y., Carpenter, R., Dickinson, H.G. and Coen, E.S. 1996. Origin of allelic diversity in *Antirrhinum S* locus RNases. Plant Cell 8: 805–814.

Zhou, J., Wang, F., Ma, W., Zhang, Y., Han, B. and Xue, Y. 2003. Structural and transcriptional analysis of *S*-locus F-box genes in *Antirrhinum*. Sex. Plant Reprod. 16: 165–177.

Zwick, M.S., Hanson, R.E., McKnight, T.D., Islam-Faridi, M.N., Stelly, D.M., Wing, R.A. and Price, H.J. 1997. A rapid procedure for the isolation of C_0t-1 DNA from plants. Genome 40: 138–142.