# Two large *Arabidopsis thaliana* gene families are homologous to the *Brassica* gene superfamily that encodes pollen coat proteins and the male component of the self-incompatibility response

Vincent Vanoosthuyse, Christine Miege, Christian Dumas and J. Mark Cock*

*Reproduction et Développement des Plantes, UMR 9938 CNRS-INRA-ENSL, Ecole Normale Supérieure de Lyon, 46 allée d'Italie, 69364 Lyon Cedex 07, France*

Author for correspondence (tel: 33 72 72 86 00, fax: 33 04 72 72 86 00, Email: Mark.Cock@ens-lyon.fr)

The *B. oleracea* nucleotide sequence data reported here are available in the EMBL, GenBank and DDBJ Nucleotide Sequence Databases under the accession numbers AJ278643 (for $SCR_3$), AJ278640 (for $SCR_{16}$) and AJ278642 (*SCRL1*). The *A. thaliana* SCRL and LCR predicted protein sequences have been submitted to the Swissprot database under the accession numbers P82620 to P82647 (for A. thaliana SCRL1 to SCRL28), P82716 to P82731 (for LCR1 to LCR16), Q9T0E3 (for LCR17), P82732 to P82735 (for LCR18 to LCR21), Q9M0F3 (for LCR22), P82737 to P82739 (for LCR23 to LCR25), Q9M0F2 (for LCR26), Q9M0F1 (for LCR27), P82743 to P82765 (for LCR28 to LCR50), O80684 (for LCR51), P82766 to P82779 (for LCR52 to LCR65), Q9H8H0 (for LCR66), Q42179 (for LCR67), Q9ZUL7 (for LCR68),

Q39182 (for LCR69), Q41914 (for LCR70), P82781 (for LCR71), Q9ZUL8 (for LCR72), P82782 to P82795 (for LCR73 to LCR86),

*Key words: Arabidopsis thaliana*, *Brassica oleracea*, defensin, pollen coat protein, receptor-like kinase ligand, *S* locus cysteine-rich protein

**Abstract**

The male component of the self-incompatibility response in *Brassica* has recently been shown to be encoded by the *S locus cysteine-rich* gene (*SCR*). *SCR* is related, at the sequence level, to the pollen coat protein (PCP) gene family whose members encode small, cysteine-rich proteins located in the proteo-lipidic surface layer (tryphine) of *Brassica* pollen grains. Here we show that the *Arabidopsis* genome includes two large gene families with homology to *SCR* and to the PCP gene family, respectively. These genes are poorly predicted by gene-identification algorithms and, with few exceptions, have been missed in previous annotations. Based on sequence comparison and an analysis of the expression patterns of several members of each family, we discuss the possible functions of these genes. In particular, we consider the possibility that SCR-related genes in *Arabidopsis* may encode ligands for the *S* gene family of receptor-like kinases in this species.

**Introduction**

In the genus *Brassica*, self-incompatible plants are able to recognise and reject self-pollen, or pollen derived from genetically closely-related plants, when it arrives on the stigma surface (reviewed in Cock, 2000). Recognition of self-pollen is genetically controlled by a single, highly polymorphic locus, the *S* locus. The region of the genome in the vicinity of the *S* locus exhibits an unusually high level of structural polymorphism; comparison of different haplotypes has revealed numerous deletions, insertions and rearrangements extending over distances of several hundred kilobases (Boyes an Nasrallah, 1993; Boyes *et al*., 1997).

Recent work has unequivocally identified both the male and female components of this cell-cell recognition system and these are both encoded by genes at the *S* locus (Schopfer *et al*.,

1999; Takayama *et al*., 2000b). The female component is the *S* locus receptor kinase (SRK), a protein which closely resembles animal receptor kinases and which is composed of an extracellular domain (the *S* domain), a single membrane-spanning domain and a cytoplasmic kinase domain (Stein *et al*., 1991, Delorme *et al*., 1995). *SRK* is expressed specifically in the stigma. A low abundance of *SRK* transcripts has also been detected in anthers but these transcripts correspond to the antisense strand of the *SRK* gene (Cock *et al*., 1997). The male component is predicted to be a small, secreted protein: the *S* locus cysteine-rich protein (SCR, also known as SP11 for *S* pollen 11; Schopfer *et al*., 1999; Takayama *et al*., 2000b). *SCR* is expressed specifically in developing anthers from the unicellular to tricellular stages of microspore development. *In situ* hybridisation has shown that, at early stages, *SCR* is expressed in both the microspores and the tapetum, with transcripts persisting in the developing microspores after tapetal breakdown. This expression pattern suggests that SCR, synthesised in the tapetum, may be transferred to the developing microspores after the tapetum degenerates (as proposed by Heslop-Harrison, 1975), providing a possible explanation for the sporophytic nature of SI in *Brassica*. This is consistent with earlier work that indicated that the male component of the SI response was located in the pollen coat (Stephenson *et al*., 1997).

The *Brassica S* locus also includes the *S* locus glycoprotein gene (*SLG*) which encodes a secreted protein similar to the extracellular domain of SRK (Nasrallah *et al*., 1987). Current evidence indicates that SLG is not required for haplotype-specific recognition of self-pollen (Cabrillac *et al*., 1999, Nishio and Kusaba, 2000; Takasaki *et al*., 2000). Rather, SLG, and a related protein SLR1 (for *S* locus related 1), have been implicated in pollen adhesion (Luu *et al*., 1999) a process that is not influenced by the SI system (Luu *et al*., 1997). *SLR1* is not genetically linked to the *S* locus. Interestingly, SLG and SLR1 have been shown to interact specifically with two members of the pollen coat protein (PCP) family, PCP-AI and SLR1-BP, respectively (Doughty *et al*., 1998; Hiscock *et al*., 1995; Takayama *et al*., 2000a). PCPs are

small, cysteine-rich proteins structurally related to SCR (Doughty *et al*., 2000). Analysis of the expression patterns of three members of this family, *PCP1*, *PCP-A1* and *SLR1-BP*, by in situ hybridisation revealed a gametophytic pattern of transcript accumulation in developing microspores (Stanchev *et al*., 1996; Doughty *et al*., 1998; Takayama *et al*., 2000a).

Although it has not been demonstrated directly, genetic evidence strongly suggests that SCR binds to SRK as a ligand. In this study, we have searched for homologues of *SCR* and the PCP gene family in *Arabidopsis* with the aim of identifying potential ligands for receptor kinases of the *S* gene family. *Arabidopsis* probably does not possess an orthologue of *SRK* (Conner *et al*., 1998) but *SRK* has been shown to be a member of a gene family that includes a number of other putative receptor kinases, both in *Brassica* (Pastuglia *et al*., 1997) and in other species including *Arabidopsis thaliana* (Walker and Zhang, 1990; Walker, 1993; Tobias *et al*., 1992; Dwyer *et al*., 1994; Zhao *et al*., 1994). Moreover, data from genome sequencing indicate that the family of *SRK* paralogues in *Arabidopsis* includes a large number of genes. Considering the tendency for the gene families encoding receptors and the gene families encoding their cognate ligands to evolve in a concerted manner (Fryxell 1996), we reasoned that *Arabidopsis* paralogues of the *Brassica SCR* (and possibly also PCP) genes might encode ligands for the *S* gene family of receptor-like kinases in this species. Here we report the identification of 114 genes in the *Arabidopsis* genome that are predicted to encode small, secreted, cysteine-rich proteins. Sequence comparison has shown that these genes can be readily divided into two gene families and that these two families are most similar to *SCR* and to the PCP family, respectively.

## Materials and methods

*Plant material*

The P57Si ($S_3/S_3$; Delorme *et al*., 1995) and P57Sc ($S_{15}/S_{15}$; Cabrillac *et al*., 1999) lines and the $F_2$ population derived from an $S_3/S_5$ heterozygous plant (Miege *et al*., 1999) have been described. Two *Brassica oleracea* var. *italica* (broccoli) lines, one homozygous for the $S_{16}$ haplotype and the second carrying a mutant form of this haplotype, were a gift from Veronique Ruffio, INRA, Rennes. The mutant $S_{16}$ line carries $SLG_{16}$ but lacks $SRK_{16}$, probably due to a deletion or recombination event (V. Ruffio, I. Fobis-Loisy, T. Gaude and J.M. Cock, unpublished results). These plants are self-compatible on both the female and male side indicating that this event has also affected the male component. The $F_2$ populations segregating for $S_{15}$ and $S_3$ and for $S_{16}$ and the mutant form of the $S_{16}$ haplotype were derived from progeny of crosses between P57Si and P57Sc and between lines homozygous for the wild-type and mutant $S_{16}$ haplotypes, respectively. The *A. thaliana* ecotype used was Columbia-0.

*Cloning of* B. oleracea SCR *and SCRL sequences and genetic mapping*

Sequences corresponding to the $SCR_3$ allele were amplified both from anther cDNA (binuclear to trinuclear microspore stages) of the P57Si line using rapid amplification of cDNA ends – polymerase chain reaction (RACE-PCR; Frohman *et al*. 1988), and from a *Brassica* $S_3/S_3$ anther cDNA library (Cock *et al*., 1997). For these amplifications, two SCR-specific oligonucleotides, SCR1 (5'-ATGAARTCIGCIRTITAYGC-3') and SCR3 (5'-CAYATHCARGARGTIGARGCNAA-3') were used in nested PCR reactions in combination with either RA2 (Frohman *et al*. 1988) or a vector-specific oligonucleotide, for the RACE-PCR or the cDNA library respectively. The products obtained in the two experiments were identical. The 5' end of the $SCR_3$ cDNA was amplified from the anther cDNA library by

nested PCR using vector- and gene-specific oligonucleotides (The gene-specific oligonucleotides were SC6: 5'-ATTTACTAGTCACTATGCAACAACATTGTCCAAC-3' in the first round and $SCR_3$P2: 5'-CCTCCACATGAGCAAAATATCTTCTT-3' in the second round).

The *$SCR_{16}$* allele was amplified by nested RACE-PCR using SCR1 and SCR2 (5'-RTITAYGCITTRTTRTGYTTYATHTT-3'), both in combination with RA2. *SCRL1* was amplified by RACE-PCR from anther cDNA of the P57Sc line using oligonucleotides SCR3 and RA2.

For genetic mapping *$SCR_3$* sequence was amplified with oligonucleotides SCR3P1 (5'-CATTCAATTCAATTTGGGTGGAC-3') and SCR3P2, *$SCR_{16}$* sequence was amplified with oligonucleotides $SCR_{16}$5' (5'-GGAGTGTGGTCGTTTTCGTTTG-3') and $SCR_{16}$3' (5'-CAACAACATCGTCCAATTTTATCTG-3') and *SCRL1* sequence was amplified with oligonucleotides SCC1 (5'-TGCGGAAATAATGGAAAGAGTGC-3') and SCC2 (5'-GACATACACATTTACGACTGGGAGG-3'). Oligonucleotides PK1, PK4 and PK5 (Nishio *et al*., 1997) were used to follow segregation of *$SRK_{16}$* in the population segregating for *$S_{16}$* and a mutant form of this haplotype.

*Sequence analysis*

Database searches were carried out using tBlastn (Zhang and Madden, 1997). Exon sequences in the *Arabidopsis* genes were identified manually in sequence translated with MBS translator (http://mbshortcuts.com/translator/) and optimal intron splice sites were identified using SplicePredictor (Usuka and Brendel, 2000). Multiple alignments and analysis of predicted polypeptides was carried out using Lasergene sequence analysis software (DNASTAR, London, UK). Phylogenetic analysis was carried out using ClustalW (Thompson *et al*., 1994)

and trees were plotted with NJplot (Perrière and Gouy, 1996). Signal peptide prediction was carried out with SignalP (Nielson *et al*., 1997). The GC content of the SCRL gene cluster (Figure 2c) was calculated using the Artemis package (http://www.sanger.ac.uk/Software/Artemis/). Ratios of nonsynonymous to synonymous substitutions were calculated as described by Ota and Nei (1994). Data points from comparisons that gave proportions of observed synonymous or nonsynonymous substitutions (ps or pn) of greater than 0.75, indicating saturation, were eliminated.

*Analysis of gene expression*

Total stigma RNA from *A. thaliana* Columbia 0 was incubated with DNaseI in the presence of a ribonuclease inhibitor (RNAsin; Promega, Madison, USA) for 15 min at 37°C. DNaseI–treated RNA (1 µg) was then reversed transcribed with Superscript II reverse transcriptase (Gibco BRL, Bethesda, USA) using an oligo $dT_{11}$ primer. PCR amplification was carried out in a 50 µl volume under the following conditions: 94° C for 2 min followed by 30 cycles of denaturation at 94°C for 30 sec, annealing at 50°C for 30 sec and extension at 72°C for 40 sec in a GeneAmp PCR system 9700 cycler (Perkin Elmer, New Jersey, NY). For the second series of PCR cycles, 1 µl of the first PCR was re-amplified in a 20 µl volume under the same conditions with 30 PCR cycles. The cDNA was replaced with 20 ng of Columbia 0 genomic DNA or with water in control reactions. *SCRL* and *LCR* sequences were PCR amplified using the following oligonucleotide pairs:

for SCRL1: 5'- TATGGAGTTTTGTTTATGGTT -3'

5'- CTTAACAATCATAATCACATCTAC -3'

for SCRL2: 5'- AATGTGGTGTTTTGTTTATGA -3'

5'- ATAGTAAAGCTTAACAATCATA -3'

for SCRL4: 5'- CTTGTGTCTTATTCTCCCTTC -3'

5'- TTAAATAGACGTTTTCTATGGG -3'

for SCRL5: 5'- CTTGTGTCTTATTCTCCCTTT -3'

5'- TCATATTAGGACAACTACACATC -3'

for SCRL22: 5'- GGATGTGGTGTACTACTTTGT -3'

5'- AAAAAGTTATAAGATTAGGAA -3'

for SCRL23: 5'- ATGAGGTGTACTACTTTGATTAT -3'

5'- TTTTAGTTACACTCATAGGGA -3'

for SCRL27: 5'- CTACCTTGTTCATGGTTTCTT -3'

5'- TGAAAATTTAGCAAGGAGAA -3'

for LCR1: 5'- ATCTTTCAACTGTCATTTACTG -3'

5'- TAGCAACATTAGTAAAGTGGAG -3'

for LCR21: 5'- TATCATGTTCTTATTTTCTCGT -3'

5'- ATTTAAAGGTTATTTCCATGAC -3'

for LCR30: 5'- ACCACCGTTATTGCTATTT -3'

5'- ACAAGAAAAATCAAGCGTT -3'

for LCR69: 5'- ATCTCAGCTGTTCTCATCAT -3'

5'- GAGAATGGGTAGATCAGCAATGT -3'

for LCR75: 5'- GAGCTCTACTACCTCTATGC -3'

5'- AAAATAATGAGAAACAAGAACT -3'

The oligonucleotide sequences were compared with the *Arabidopsis* genome sequence to ensure that they were gene-specific. The products amplified by these oligonucleotides all include intron sequences allowing genomic products to be distinguished from cDNA. In order to ensure that an equal amount of cDNA had been added to each PCR reaction, the abundance of actin cDNA in each sample was compared by PCR amplification using the following

oligonucleotides: 5'-GATTTGGCATCACACTTTCTACA-3' and 5'-GTTCCACCACTGAGCACAATG-3'. All RT-PCR analyses were repeated at least once in independent experiments.

**Results**

*Molecular cloning of two novel* SCR *alleles and an SCR-related sequence from* Brassica oleracea.

Following the recent identification of the *SCR* gene, which encodes the male component of the SI response in *Brassica* (Schopfer *et al*., 1999; Takayama *et al*., 2000b), we were interested in identifying related genes, particularly in the model species *Arabidopsis thaliana*. When this project was initiated, sequence data was available for only a limited number of alleles. As a first step, therefore, we attempted to obtain sequence data from additional *SCR* alleles in order to provide a clearer picture of the extent of polymorphism at this locus and to provide additional sequences for database searching.

Using degenerate oligonucleotides based on published *SCR* alleles, three novel sequences were isolated from *B. oleracea* lines homozygous for the $S_3$, $S_{16}$ and $S_{15}$ haplotypes. Based on mapping data (see below) these sequences were found to represent two new *SCR* alleles and a sequence corresponding to another member of this gene family. They were therefore named $SCR_3$, $SCR_{16}$ and *SCRL1* (for SCR-like1).

*Genetic mapping of* Brassica SCR *and SCRL sequences*

Segregation of $SCR_3$ was analysed in an $F_2$ population of 142 plants descended from an $S_3/S_5$ heterozygote and was compared with the segregation of $SLG_3$ and $SLG_5$ which has previously been determined in this population (Miege *et al*., 1999). No recombination events between $SCR_3$ and $SLG_3$ were detected (data not shown). Thus $SCR_3$ is tightly linked to the *S* locus and is, therefore, highly likely to be an allele of *SCR*.

PCR analysis showed that the $SCR_{16}$ allele was present in a wild-type $S_{16}$ haplotype but absent from a mutant form of this haplotype which has undergone a deletion that also affects the *SRK* gene (data not shown). Plants homozygous for the mutant $S_{16}$ haplotype accept $S_{16}$ pollen and produce pollen that is accepted by plants carrying the wild-type $S_{16}$ haplotype. SI is therefore defective on both the male and female sides in these plants (V. Ruffio, I. Fobis-Loisy, T. Gaude and J.M. Cock, unpublished results). No recombination events were detected between $SCR_{16}$ and $SRK_{16}$ in an $F_2$ population of 20 plants segregating for the $S_{16}$ haplotype and the mutant $S_{16}$ haplotype (data not shown). Taken together these data indicate strongly that $SCR_{16}$ is also an allele of *SCR*. *SCRL1*, on the other hand, was detected in all 19 plants of an $F_2$ population segregating for $S_{15}$ and $S_3$ indicating that it is not linked to the *S* locus and, therefore, corresponds to an *SCR*-related gene.

*Identification of* SCR *and PCP homologues in* Arabidopsis

*SCR*-, *SCRL1*- and PCP-homologous sequences in the *Arabidopsis* genome were identified by iterative searches using the tBlastn programme. Primary searches were carried out using the deduced amino acid sequences of 7 *SCR* alleles, *SCRL1* (this study, accession N° AJ278642) and three PCP genes. The *SCR* alleles used were $SCR_3$ (this study, accession N° AJ278643), $SCR_{16}$ (this study, accession N° AJ278640) and $SCR_6$ (accession N° AAF17503) from *B. oleracea* and $SCR_8$ (accession N° AAF17505), $SCR_9$ (accession N° BAA85458), $SCR_{12}$

(accession N° AB035503) and *SCR*$_{52}$ (accession N° AB035505) from *B. rapa*. *Arabidopsis* genes with homology to the 7 *SCR* and SCRL sequences were, in turn, compared with the database using tBlastn. No novel homologous sequences were detected after the fourth series of iterative searches. Table 1 shows that a total of 28 SCRL (*SCR*-related) sequences were identified in the *Arabidopsis* sequence available in the databases (corresponding to approximately 86 % of the total genome). This search did not detect any SCRL genes in species other than *Brassica* and *Arabidopsis*.

Primary searches for PCP homologues were carried out with *PCP-A1* (accession N° CAA06464), *PCP1* (accession N° X97054) and *PCP2* (accession N° X97055). Initial searches detected genes which most closely resembled the PCP family, as expected, but, by the fourth iterative series of searches, homologues detected were either more closely related to the defensin gene family than to the PCP family or were even more distantly related. The iterative searches were, therefore, not continued after the fourth series. A total of 86 genes, designated LCR for low molecular weight, cysteine-rich, were identified in the four series of iterative searches (Table 2).

None of the 28 SCRL genes had previously been annotated in the databases. Of the 86 LCR genes only 10 had been previously annotated correctly, the majority of these being sequences that are more closely related to the defensin gene family (Broekaert *et al*., 1995) than to the PCP gene family. The remaining 77 members of the LCR family were either incorrectly annotated (7 genes) or were not annotated.

Apart from the 114 genes described here, several additional sequences with similarity to the SCRL (6 sequences) and LCR (7 sequences) gene families were found in the *Arabidopsis* genome. These sequences did not appear to correspond to complete genes, however, and may be pseudogenes. They were not analysed further.

*Genomic organisation of the SCRL and LCR gene families*

Figure 1a shows that the SCRL genes are scattered throughout the *Arabidopsis* genome, with some loci containing closely linked gene clusters (Figure 1b). Several of the LCR genes are also found in clusters (see Figure 1b, for example). In both families, clustered genes tend to be more similar to each other than to the rest of the family suggesting that they have arisen by local gene duplication events.

We examined further the structure of the largest SCRL gene cluster by dot matrix analysis (Figure 2a). The long diagonals to either side of the main diagonal correspond to regions of similarity between the genes of the cluster. In most cases the similarity extends both to the intron and to 5' and 3' non-coding sequences. For example, comparison of *SCRL4* and *SCRL5* revealed two regions of high similarity (77.4%) that extend approximately 360 bp upstream and 90 bp downstream of each coding region (Figure 2b). These two conserved regions are separated by less than 500 bp and it will be interesting to determine whether the upstream regulatory elements of *SCRL5* are located in the short region of non-coding sequence between *SCRL4* and *SCRL5*.

In addition to the regions of similarity between SCRL genes in the cluster, the dot matrix analysis also detected a periodic structure in this region of the chromosome with segments giving multiple short matches preceding each SCRL gene (Figure 2a). These regions were analysed using the MEME algorithm (Bailey and Elkan, 1994) but no significant repeat sequences were detected. Rather, the observed pattern appears to be due to decreased complexity in AT-rich regions located immediately upstream of each SCRL gene (Figure 2c). AT-rich sequences have been found to be associated with several gene duplication events in animals (Hyrien *et al*., 1987; Legouy *et al*., 1989; Murru *et al*., 1990) and it is possible that the

AT-rich regions in the SCRL gene cluster played some role in the gene duplications that created this group of genes.

In an attempt to reconstruct the evolutionary history of the two largest clusters of SCRL and LCR genes, we examined phylogenetic relationships between the genes by the neighbour joining method (Figure 3). In general, no strong correlation was obtained between predicted phylogenetic relationships and physical positions on the chromosome indicating that the evolutionary history of these gene clusters is complex, involving multiple rearrangements. However, the head-to-tail arrangement of *SCRL8, SCRL6, SCRL4* and *SCRL5* correlates with a predicted series of three duplication events suggesting that part of this cluster arose by repeated, sequential gene duplications perhaps due to unequal crossing-over events.

The size of the single intron in the SCRL and LCR genes is variable, 183 ± 180 bp for the SCRL gene family and 274 ± 224 bp for the LCR gene family (Tables 1 and 2). No significant similarity was detected between introns of the different members of the two gene families apart from in very closely related genes such as *SCRL4* and *SCRL5*.

*Structural features of SCRL and LCR predicted polypeptides*

All but one of the SCRL and LCR proteins were predicted to possess cleaved signal peptides although the prediction was weak for some of the LCR proteins and it is possible that these particular proteins are not secreted (Tables 1 and 2). These predictions will need to be confirmed experimentally but they were nonetheless used here to permit comparison of the predicted mature proteins. Based on this analysis, the mature SCRL proteins were predicted to vary between 4.4 and 9.5 kDa (average mass: 7.8 ± 1.0 kDa) and to be basic (average pI of 8.1 ± 0.9), hydrophilic proteins. The predicted mature LCR proteins are smaller on average (4.5 to

11.0 kDa with an average mass of 6.6 ± 1.4 kDa) but are also slightly basic (pI 7.3 ± 1.4) and hydrophilic.

Alignments of both the SCRL (Figure 4) and selected LCR (Figure 5) deduced amino acid sequences revealed a high level of sequence diversity. On the whole, the predicted mature polypeptides share very little sequence homology apart from a small number of conserved residues including eight highly conserved cysteines. Several of the predicted SCRL and LCR polypeptides lack one or more of the conserved cysteines probably as a result of either nucleotide mutations or deletions (SCRL1, SCRL21, SCRL28 and LCR28 for example; Tables 1 and 2). We do not know at present if these proteins are functional but it is worth noting that several functional, allelic forms of *SCR* in *Brassica* lack one of the conserved cysteines (Watanabe *et al*., 2000).

When the SCRL deduced amino acid sequences were compared, the predicted signal peptides were found to be more highly conserved between genes than was the rest of the polypeptide (Figure 4). A similar observation has been made for alleles of *SCR* in *Brassica* (Scopfer et al., 1999; Watanabe et al., 2000). This conservation of the peptide signal sequences was surprising as, in general, signal peptides tend to diverge more rapidly than other protein domains and a similar phenomenon was not observed for the LCR proteins (Figure 5). Moreover, the difference in sequence conservation between these two domains would seem to be due to a different rate of sequence divergence as more recently diverged genes such as *SCRL4* and *SCRL5* (Figure 3) exhibited the same phenomenon (87.5% identity between predicted signal peptides but only 68.4% identity for the rest of the polypeptides). These observations suggest that, during evolution, the two domains of the SCRL proteins are have been under different selection pressures. To analyse this phenomenon further, we calculated the ratio of nonsynonymous to synonymous substitutions in five regions corresponding to the residues between the cysteines of the mature polypeptide (Figure 6a). For this analysis, we

concentrated on the five genes of the largest SCRL gene cluster (*SCRL4* to *SCRL8*) because the high level of similarity between these genes allowed unambiguous identification of homologous codons. When each member of the cluster was compared with each of the other members, a high level of variance was observed particularly when the average ratio of nonsynonymous to synonymous substitutions was high (Figure 6b). To provide a clear representation of this variance, we calculated the average ratio of nonsynonymous to synonymous substitutions for each gene after comparing, individually, with all the genes in the cluster. Figure 6b shows that, in all five genes, the region encoding the domain between the second and third conserved cysteine (denoted region C in Figure 6) has accumulated a large proportion of nonsynonymous substitutions. The ratio consistently exceeded one indicating that positive selection is acting to diversify the sequence in this region. A similar phenomenon was seen for regions D and E but, in this case, the values were not significantly greater than one when the standard deviations were taken into account. Positive selection does not seem to be acting on other regions of the genes that were analysed (regions A and B in Figure 6b). In particular, region A, which is adjacent to the predicted signal peptide cleavage site, appears to be functionally constrained.

*Relationships between the SCRL and LCR gene families in* Arabidopsis *and* SCR *and the PCP family in* Brassica

Despite the high level of diversity of the SCRL and LCR genes, several of their features support the hypothesis that they are homologues of the *SCR, SCRL1* and PCP genes from *Brassica* (in the sense that all these genes probably share a common ancestor). Firstly, both families encode proteins with eight conserved cysteines in similar configurations. Secondly, all of the SCRL/LCR genes encode a short polypeptide of less than 12 kDa. Thirdly, all but one of the

deduced polypeptides include a hydrophobic amino-terminal domain that is predicted to function as a signal sequence. Finally, with the exception of *SCRL8, LCR27* and *LCR71,* all the SCRL/LCR genes are interrupted by a single intron inserted near the 5' end of the coding sequence (Tables 1 and 2). In all cases, the intron is positioned between the first and second nucleotide of a codon, a strong indicator of homology. In *Brassica,* gene structure has been determined for several *SCR* alleles (Schopfer *et al*., 1999; Takayama *et al*., 2000b) and for four PCP genes *(PCP-A1, PCP1*, *PCP2* and *SLR1-BP*; Doughty *et al*., 1998; Stanchev *et al*., 1996; Takayama *et al*., 2000a). All five of these genes contain a single intron in a position that corresponds to those of the SCRL/LCR genes and, again, in all cases the intron is located between the first and second nucleotide of a codon.

The above observations not only indicate homology between the genes identified in *Arabidopsis* and the *SCR* and PCP genes in *Brassica*, they also indicate homology between the SCRL and the LCR gene families in *Arabidopsis*. Nonetheless, despite being distantly related, the *Arabidopsis* SCRL and LCR genes can clearly be recognised as belonging to two distinct sub-families at the sequence level. Figure 7 shows a comparison of the consensus sequences of the deduced SCRL and LCR proteins. Several features distinguish the two families, particularly the position of the fourth conserved cysteine relative to the third and fifth conserved cysteines. The differences between these two families were reflected in the fact that none of the PCP/LCR sequences detected any of the SCRL genes in tBlastn searches.

The LCR family includes 14 genes (*LCR66* to *LCR79*) whose deduced amino acid sequences possess all of the conserved residues typical of defensins (Broekaert *et al*., 1994; Doughty *et al*., 1998). However, based on sequence comparisons, it was not possible to clearly class the members of the LCR family as being either PCP-like or defensin-like. For example, whilst the conserved serine at position 48 (numbering refers to the scale on the alignment in Figure 5) and tryptophane or phenylalanine at position 51 were only found in these 14 genes,

the two conserved glycines at positions 53 and 80 and the conserved glutamic acid at position 75 were more or less conserved in many of the other members of the family. In consequence, all of these genes were grouped together in the LCR family.

*Expression of the SCRL and LCR genes*

None of the SCRL sequences were represented in the *Arabidopsis* EST database. This suggested that, if the members of the family are expressed, transcripts only accumulate to low levels or accumulate with a very restricted pattern of expression. Another possibility is that the genes could be inducible. We analysed the expression patterns of 7 SCRL genes by RT-PCR. Figure 8 shows that two of the genes were expressed specifically in flower buds (*SCRL23* and *SCRL27*) whilst three others (*SCRL4, SCRL5* and *SCRL22*) were expressed in more than one organ: *SCRL22* was expressed in all the organs tested whereas *SCRL4* was expressed in flower buds and roots and *SCRL5* was expressed in flower buds and stems. No *SCRL1* nor *SCRL2* transcripts were detected even after a second series of 30 PCR cycles. Interestingly, different patterns of expression were observed for *SCRL4* and *SCRL5* despite the similarity between these two genes (76.7 % at the amino acid level) and their close association in the largest SCRL gene cluster (Figure 1b). This result suggests that there has been a diversification of function of the genes in this cluster.

Only six of the LCR genes were represented in the EST database. Five of these genes were most similar to defensin genes and included *LCR67* (*PDF1.1)*, *LCR77* (*PDF1.2*) and *LCR70* for which the corresponding ESTs were derived from dry seed, seedling and green silique libraries, respectively. Interestingly, the EST corresponding to the sixth gene, *LCR17*, which is more similar to PCP genes, was derived from a flower bud library. This would be consistent with the gene being expressed in pollen.

The fact that only six LCR genes were represented in the EST database indicated that, like the SCRL genes, the majority of the members of this family are expressed at low levels or with a restricted pattern of expression. We analysed the expression pattern of five members of the LCR gene family by RT-PCR (Figure 8). The sample of LCR genes analysed included two genes that are predicted to encode proteins with all the conserved residues of defensins (*LCR69* and *LCR75)* and three genes that are more similar to *Brassica* PCP genes. *LCR69* and *LCR75* exhibited a very broad pattern of expression with transcripts being detected in all the organs analysed. In contrast, the three PCP-like genes, *LCR1, LCR21* and *LCR30,* were expressed specifically in flower buds.

Penninckx *et al*., (1996) used RT-PCR to determine the expression patterns of two other *Arabidopsis* defensins, *PDF1.1* (*LCR67*) and *PDF1.2* (*LCR77*). *PDF1.1* transcripts were detected in siliques and dry seeds whereas *PDF1.2* transcripts were only detected in leaves inoculated with *Alternaria brassicicola* and were not detected in any uninfected tissues.

## Discussion

### *Identification of the SCRL and LCR gene families*

The aim of this study was to identify genes that potentially encode ligands for the *S* gene family group of receptor-like kinases in *Arabidopsis*. To do this, we searched the *Arabidopsis* genome for homologues of the *Brassica SCR* gene that is thought to encode the ligand for SRK. We also searched for homologues of the PCP gene family because one member of this family, PCP-A1, binds to SLG, a protein that closely resembles the extracellular domain of SRK.

We showed that the *Arabidopsis* genome contains two distinct families of *SCR*-like and PCP-like genes (the SCRL and LCR families, respectively), with each family containing a large number of genes. The existence of these two extensive gene families suggests that the common ancestor of *SCR* and the PCP genes is relatively ancient. This is surprising, considering the fact that SLG and the extracellular domain of SRK are highly similar at the sequence level. Moreover, in several haplotypes *SLG* and *SRK* are more similar to each other than to *SLG* and *SRK* alleles from other haplotypes (Stein *et al*., 1991; Kusaba *et al*., 1997). This concerted evolution of *SLG* and *SRK* was originally taken to indicate that both of these genes were involved in mediating the SI response. However, more recent evidence indicates that *SLG* is not required for self-pollen recognition (Cabrillac *et al*., 1999; Takasaki *et al*., 2000) and *Brassica* plants that lack a functional *SLG* gene have been identified (Nishio and Kusaba, 2000). Dixit *et al*. (2000) have proposed that SLG is involved in the maturation of the SRK protein although it seems to be redundant in this function and this is probably not a haplotype-specific function. Alternatively there is evidence that SLG is involved in pollen adhesion (Luu *et al*., 1999), a process that is not influenced by the SI response (Luu *et al*., 1997). It therefore seems more likely that the sequence similarity between *SLG* and *SRK* does not reflect functional conservation but, rather, is a result of their close proximity at the *S* locus, facilitating the action of a homogenising mechanism such as gene conversion (Cabrillac *et al*., 1999). The fact that *SCR* and *PCP-A1* are members of two distinct gene families would seem to support this model.

*Possible functions of the LCR and SCRL genes*

The LCR gene family was shown to be heterogeneous and to contain genes with similarity to both the PCP and the defensin gene families. The members of this family are highly divergent at the sequence level and exhibit various different patterns of expression. Data

from EST sequencing and from the RT-PCR analysis carried out in this study indicate that the LCR genes that are most similar to the PCP family tend to be expressed specifically in flowers whilst the genes that are more closely related to the defensins tend to be expressed in both floral and vegetative organs. This suggests that the members of this family may carry out more than one type of function in the plant. In *Brassica*, PCPs are thought to play a role in pollination, perhaps as one of the components that mediate pollen adhesion (Luu *et al*., 1999; Doughty *et al*., 2000). Plant defensins have been so named because of their resemblance to the defensin antimicrobial peptides of insects and mammals. Several of the plant defensin genes have been shown to be induced following attack by pathogens (reviewed in Broekaert *et al*., 1995). In addition, PDF1.1 and PDF1.2 from *Arabidopsis,* and Rs-AFP2, a defensin from radish, have been shown to possess potent antifungal activity (Terras *et al*., 1992; Penninckx *et al*., 1996). SIα-2, from wheat, is an inhibitor of α-amylase (Bloch and Richardson, 1991). It is likely that the members of the LCR family in *Arabidopsis* carry out similar roles to the PCPs and defensins identified in other species.

The function of the SCRL genes in *Arabidopsis* is as yet unknown, but based on their similarity to *SCR*, one possibility is that they encode ligands for *S* gene family receptor-like kinases. To date, expression patterns are known for only a small number of *Arabidopsis S* gene family receptor-like kinases. These genes exhibit various different patterns of expression in a wide range of organs including leaves, cotyledons, sepals and specific tissues of the pistil, roots, hypocotyl and petioles (Tobias *et al*., 1992; Dwyer *et al*., 1994; Walker, 1993). In comparison, of the five SCRL genes for which we detected transcripts, two were expressed specifically in flower buds, whilst the others were also expressed in certain vegetative tissues. Although only semi-quantative, our study indicated that all the SCRL genes analysed were expressed at low levels. These expression patterns are not inconsistent with the hypothesis that the SCRL genes encode ligands.

Comparison of different alleles of the *Brassica SCR* gene has shown that the signal peptide is more highly conserved than the mature protein (Scopfer et al., 1999; Watanabe *et al*., 2000). Although less marked, this phenomenon was also noted when the *Arabidopsis* SCRL genes were compared (Figure 4). Further analysis of several members of this family indicated that the regions encoding the amino acids between the second and third conserved cysteines (and perhaps also other regions of the coding sequence) have been subjected to diversifying selection. Examples of genes that have undergone diversifying selection are rare and this phenomenon is almost always associated with genes involved in recognition events. This analysis, therefore, provides further support for the hypothesis that the SCRL proteins function as ligands and, moreover, identifies a region of these proteins that is potentially involved in such interactions. It will be interesting to determine whether the corresponding regions of the *Brassica SCR* gene are also under diversifying selection, as might be predicted for domains determining haplotype specificity.

*Protein structure*

Structural analysis of defensins has shown that the eight conserved cysteine residues form intramolecular disulphide bridges (Bruix *et al*., 1993). Biochemical analysis indicates that this is also the case for PCP-A1 and it has been suggested that the structures of the members of the PCP family resemble those of the defensins (Doughty *et al*., 1998). It is likely that the LCR proteins also have similar structures. The structure of the members of the SCRL family is less clear, particularly as some functional, allelic forms of SCR in *Brassica* lack one of the conserved cysteines. However, analysis of the SCRL family provides some clues as to the residues that may be important for the basic structure of these proteins. For example, in all of the allelic forms of SCR and in all but one of the SCRL proteins there is a single residue

between the 4th and 5th cysteine and a single residue between either the 6th and 7th or the 6th and 8th cysteine residue. In general, the conservation of the cysteine residues in all of these families indicates that they are important structurally. The majority of these proteins are predicted to be secreted and the presence of multiple disulphide bridges could be an important factor contributing to their stability in the extracellular environment.

*Large gene families in* Arabidopsis

In total, we identified 29 *SCR*-like genes and an additional 86 LCR genes that more closely resemble the *Brassica* PCP gene family. The existence of a large family of PCP-related genes in *Arabidopsis* is consistent with a previous report, based on hybridisation of a *PCP1* probe to genomic DNA, that indicated the presence of a large family of 30 to 40 PCP-like genes in the *Brassica* genome (Stanchev *et al*., 1996). Taken together, the SCRL and LCR families constitute well over 100 genes in *Arabidopsis* and it is likely that a large number of more divergent, but nonetheless related, genes exist in the genome. Based on the analysis carried out here, we predict that the superfamily that includes the SCRL and LCR gene families constitutes a significant percentage (as much as 1%) of the genes in the *Arabidopsis* genome. Several other gene families in *Arabidopsis* are estimated to contain over 100 members. These include MYB-related transcription factors (Kranz *et al*., 1998), cytochrome P450 genes (http://drnelson.utmem.edu/Arablinks.html), the AtPCMP family (Aubourg *et al*., 2000) and the receptor-like kinase superfamily.

The majority of the *Arabidopsis* SCRL and LCR genes identified in this study were found in regions of the genome that had already been annotated. These genes had not been recognised as coding sequence, however, and were located in the spaces between the annotated genes. A similar observation was made recently by Ride *et al*. (1999) who identified 32 novel

*Arabidopsis* genes (*SPH1* to *SPH32*) related to the *Papaver S* (self-incompatibility) gene (which is either unrelated, or only very distantly related, to the self-incompatibility genes in *Brassica*). In both cases the gene families are composed of small genes that are poorly predicted by algorithms such as GeneFinder and Grail. These studies highlight the importance of completing preliminary genome annotation with more detailed searches based on homology that take into account conserved features of gene families. The organisation of genes into gene families provides an important starting point for the analysis of gene function by revealing the relationships between genes in the genome.

**Acknowledgements**

We would like to thank Laurent Duret and Gabriel Marais (Université Lyon I) for helpful discussions, Isabelle Fobis-Loisy for mapping the $SCR_{16}$ gene, Hervé Leyral for technical assistance and Veronique Ruffio (INRA, Le Rheu) for the broccoli lines carrying the $S_{16}$ haplotype. We also thank Charlie Scutt and Isabelle Fobis-Loisy for their comments on manuscript. J.M.C. is a member of the Institut National de la Recherche Agronomique.

**References**

Aubourg, S., Boudet, N., Kreis, M., Lecharny, A. 2000. In *Arabidopsis thaliana*, 1% of the genome codes for a novel protein family unique to plants. Plant Mol Biol. 42:603-613.

Bailey, T.L. and Elkan, C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings of the Second International Conference on

Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California,

Bloch, C. Jr. and Richardson, M. 1991. A new family of small (5 kDa) protein inhibitors of insect alpha-amylases from seeds or sorghum (Sorghum bicolar (L) Moench) have sequence homologies with wheat gamma-purothionins. FEBS Lett. 279: 101-104.

Boyes, D.C. and Nasrallah, J.B. 1993. Physical Linkage of the *SLG* and *SRK* genes at the self-incompatibility locus of *Brassica oleracea*. Mol. Gen. Genet. 236: 369-373.

Boyes, D.C., Nasrallah, M.E., Vrebalov, J. and Nasrallah, J.B. 1997. The self-incompatibility (*S*) haplotypes of Brassica contain divergent and rearranged sequences of ancient origin. Plant Cell 9, 237-247.

Broekaert, W.F., Terras, F.R.G., Cammue, B.P.A. and Osborn, R.W. 1995. Plant defensins: Novel antimicrobial peptides as components of the host defense system. Plant Physiol., 108: 1353-1358.

Bruix, M., Jimenez, M.A., Santoro, J., Gonzalez, C., Colilla, F.J., Mendez, E. and Rico, M. 1993. Solution structure of gamma 1-H and gamma 1-P thionins from barley and wheat endosperm determined by 1H-NMR: a structural motif common to toxic arthropod proteins. Biochemistry. 32:715-724.

Cabrillac, D., Delorme, V., Garin, J., Ruffio-Châble, V., Giranton, J.-L., Dumas, D., Gaude, T. and Cock, J.M. 1999. The $S_{15}$ Self-incompatibility Haplotype in *Brassica oleracea* Includes Three *S* Gene Family Members Expressed in Stigmas. Plant Cell, 11: 971-986.

Cock, J.M., Swarup, R. and Dumas, C. 1997. Natural antisense transcripts of the *S* locus receptor kinase gene and related sequences in *Brassica oleracea*. Mol. Gen. Genet., 255: 514-524.

Cock, J.M. 2000. A Receptor Kinase and the Self-Incompatibility Response in *Brassica*. In: Advances in Botanical Research, thematic volume: Plant Protein Kinases, Kreis, M. and Walker, J.C. (eds) Academic Press, London, 32: 270-298.

Conner, J.A., Conner, P., Nasrallah, M.E., Nasrallah, J.B. 1998. Comparative Mapping of the Brassica *S* Locus Region and its Homeolog in Arabidopsis: Implications for the Evolution of Mating Systems in the Brassicaceae. Plant Cell 10: 801-812.

Delorme, V., Giranton, J.L., Hatzfeld, Y., Friry, A., Heizmann, P., Ariza, M.J., Dumas, C., Gaude, T., and Cock, J.M. 1995. Characterisation of the *S* locus genes, *SLG* and *SRK*, of the *Brassica* $S_3$ haplotype: identification of a membrane-localised protein encoded by the *S* locus receptor kinase gene. Plant J. 7: 429-440.

Dixit, R., Nasrallah, M.E. and Nasrallah, J.B. 2000. Post-transcriptional maturation of the S receptor kinase of brassica correlates with Co-expression of the S-locus glycoprotein in the stigmas of two brassica strains and in transgenic tobacco plants. Plant Physiol. 124: 297-312.

Doughty, J., Dixon, S., Hiscock, S.J., Willis, A.C., Parkin, I.A.P. and Dickinson H.G. 1998. PCP-1A, a defensin-like Brassica pollen coat protein that binds the *S* locus glycoprotein, is the product of gametophytic gene expression. Plant Cell 10: 1333-1347.

Doughty, J., Wong, H.Y. and Dickinson H.G. 2000. Cysteine-rich Pollen Coat Proteins (PCPs) and their Interactions with Stigmatic S (Incompatibility) and S-Related Proteins in Brassica: Putative Roles in SI and Pollination. Annals Bot. 85: 161-169.

Dwyer, K.G., Kandasamy, M.K., Mahosky, D.I., Acciai, J., Kudish, B.I., Miller, J.E., Nasrallah, M.E. and Nasrallah, J.B. 1994. A superfamily of *S* locus-related sequences in Arabidopsis: Diverse structures and expression patterns. Plant Cell 6: 1829-1843.

Fryxell, K.J. 1996. The coevolution of gene family trees. Trends Gen. 12: 364-369.

Frohman, M.A., Dush, M.K. and Martin, G.R. 1988. Rapid production of full-length cDNAs from rare transcripts: Amplification using a single gene-specific oligonucleotide primer. Proc Natl Acad Sci USA 85: 8998-9002.

Heslop-Harrison, J. 1975. Incompatibility and the pollen stigma reaction. Ann. Rev. Plant Physiol. 26: 403-425.

Hiscock, S.J., Doughty, J., Willis, A.C. and Dickinson, H.G., 1995. A 7-kDa pollen coating-borne peptide from *Brassica napus* interacts with S-locus glycoprotein and S-locus-related glycoprotein. Planta, 196, 367-374.

Hyrien, O., Debatisse, M., Buttin, G. and de Saint Vincent, B.R. 1987. A hotspot for novel amplification joints in a mosaic of Alu-like repeats and palindromic A + T-rich DNA. EMBO J. 6: 2401-2408.

Kranz, H.D., Denekamp, M., Greco, R., Jin, H., Leyva, A., Meissner, R.C., Petroni, K., Urzainqui, A., Bevan, M., Martin, C., Smeekens, S., Tonelli, C., Paz-Ares, J. and Weisshaar, B. 1998. Towards functional characterisation of the members of the *R2R3-MYB* gene family from *Arabidopsis thaliana*. Plant J. 16: 263-276.

Kusaba, M., Nishio, T., Satta, Y., Hinata, K. and Ockendon, D. 1997. Striking sequence similarity in inter- and intra-specific comparisons of class I SLG alleles from Brassica oleracea and Brassica campestris: Implications for the evolution and recognition mechanisms. Proc. Natl. Acad. Sci. USA 94: 7673-7678.

Legouy, E., Fossar, N., Lhomond, G. and Brison, O. 1989. Structure of four amplified DNA novel joints. Somat Cell Mol Genet. 15: 309-320.

Luu, D.-T., Heizmann, P. and Dumas, C. 1997. Pollen-stigma adhesion in kale is not dependent on the self-(in)compatibility genotype. Plant Phyiol 115: 1221-1230.

Luu, D.-T., Marty-Mazars, D., Trick, M., Dumas, C. and Heizmann, P. 1999. Pollen-Stigma Adhesion in Brassica involves SLG and SLR1 Glycoproteins. Plant Cell 11: 251-262.

Miege, C., Dumas, C. and Cock, J.M. (1999) Identification of a gene linked to the Brassica S (self-incompatibility) locus by differential display. Comptes Rendus de l'Académie des Sciences / Life Sciences, 322: 1051-1060.

Murru, S., Casula, L., Pecorara, M., Mori, P., Cao, A. and Pirastu, M. 1990. Illegitimate recombination produced a duplication within the FVIII gene in a patient with mild hemophilia A. Genomics. 7: 115-118.

Nasrallah, J.B., Kao, T.-H., Chen, C.-H., Goldberg, M.L., and Nasrallah, M.E. 1987. Amino-acid sequence of glycoproteins encoded by three alleles of the *S*-locus of *Brassica oleracea*. Nature 326: 617-619.

Nielsen, H., Engelbrecht, J., Brunak, S. and von Heijne G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering 10: 1-6.

Nishio, T., Kusaba, M., Sakamoto, K., Ockendon, D.J. 1997. Polymorphism of the kinase domain of the *S*-locus receptor kinase gene (*SRK*) in *Brassica oleracea* L. Theor. Appl. Genet. 95: 335-342.

Nishio, T. and Kusaba, M. 2000. Sequence diversity of *SLG* and *SRK* in *Brassica oleracea* L.. Annals Botany 85: 141-146.

Ota, T, and Nei, M. 1994. Variance and covariances of the numbers of synonymous and nonsynonymous substitutions per site. Mol. Biol. Evol. 11: 613-619.

Pastuglia, M., Roby, D., Dumas, C. and Cock, J.M. 1997. Rapid induction by wounding and bacterial infection of an *S* gene family receptor-like kinase gene in *Brassica oleracea*. Plant Cell 9: 49-60.

Penninckx, I.A., Eggermont, K., Terras, F.R., Thomma, B.P., De Samblanx, G.W., Buchala, A., Metraux, J.P., Manners, J.M. and Broekaert, W.F. 1996. Pathogen-induced systemic

activation of a plant defensin gene in Arabidopsis follows a salicylicacid-independent pathway. Plant Cell 8: 2309-23.

Perrière, G. and Gouy, M. 1996. WWW-Query: An on-line retrieval system for biological sequence banks. Biochimie 78: 364-369.

Ride, J.P., Davies, E.M., Franklin, F.C.H. and Marshall, D.F. 1999. Analysis of *Arabidopsis* genome sequence reveals a large new gene family in plants. Plant Mol. Biol. 39: 927-932

Sambrook, J., Fritsch, E.F. and Maniatis, T. 1989. *Molecular cloning: a laboratory manual.* Cold Spring Harbour, NY: Cold Spring Harbour Laboratory Press.

Schopfer, C.R., Nasrallah, M.E. and Nasrallah, J.B. 1999. The male determinant of self-incompatibility in Brassica. Science 286: 1697-700.

Stanchev, B.S., Doughty, J., Scutt, C.P., Dickinson, H., and Croy, R.R. 1996. Cloning of PCP1, a member of a family of pollen coat protein (PCP) genes from *Brassica oleracea* encoding novel cysteine-rich proteins involved in pollen-stigma interactions. Plant J. 10: 303-313.

Stein, J.C., Howlett, B., Boyes, D.C., Nasrallah, M.E., and Nasrallah, J.B. 1991. Molecular cloning of a putative receptor protein kinase gene encoded at the self-incompatibility locus of *Brassica oleracea*. Proc. Natl. Acad. Sci. USA 88: 8816-8820.

Stephenson, A.G., Doughty, J., Dixon, S., Elleman, C., Hiscock, S. and Dickinson, H.G. 1997. The male determinant of self-incompatibility in *Brassica oleracea* is located in the pollen coating. Plant J. 12: 1352-1359.

Takasaki, T., Hatakeyama, K., Suzuki, G., Watanabe, M., Isogai, A. and Hinata, K. 2000. The S receptor kinase determines self-incompatibility in Brassica stigma. Nature 403: 913-916.

Takayama, S., Shiba, H., Iwano, M., Asano, K., Hara, M., Che, F.S., Watanabe, M., Hinata, K. and Isogai, A. 2000a. Isolation and characterization of pollen coat proteins of *Brassica campestris* that interact with *S* locus-related glycoprotein 1 involved in pollen-stigma adhesion. Proc Natl Acad Sci U S A. 97: 3765-3770.

Takayama, S., Shiba, H., Iwano, M., Shimosato, H., Che, F.S., Kai, N., Watanabe, M., Suzuki, G., Hinata, K. and Isogai, A. 2000b. The pollen determinant of self-incompatibility in *Brassica campestris*. Proc Natl Acad Sci U S A. 97: 1920-1925.

Terras, F.R., Schoofs, H.M., De Bolle, M.F., Van Leuven, F., Rees, S.B., Vanderleyden, J., Cammue, B.P. and Broekaert, W.F. 1992. Analysis of two novel classes of plant antifungal proteins from radish (*Raphanus sativus* L.) seeds. J Biol. Chem. 267:15301-15309.

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

Tobias, C.M., Howlett, B. and Nasrallah, J.B. 1992. An *Arabidopsis thaliana* gene with sequence similarity to the *S*-locus receptor kinase of *Brassica oleracea*. Plant Physiol. 99: 284-290.

Usuka, J. and Brendel, V. 2000. Gene structure prediction by spliced alignment of genomic DNA with protein sequences: Increased accuracy by differential splice site scoring. J. Mol. Biol. 297: 1075-1085.

Walker, J.C. 1993. Receptor-like protein kinase genes of *Arabidopsis thaliana*. Plant J. 3: 451-456.

Walker, J.C. and Zhang, R. 1990. Relationship of a putative receptor protein kinase from maize to the *S*-locus glycoproteins of *Brassica*. Nature 345: 743-746.

Watanabe, M, Ito, A, Takada, Y, Ninomiya, C, Kakizaki, T, Takahata, Y, Hatakeyama, K, Hinata, K., Suzuki, G., Takasaki, T., Satta, Y., Shiba, H., Takayama, S. and Isogai, A. 2000. Highly divergent sequences of the pollen self-incompatibility (S) gene in class-I S haplotypes of Brassica campestris (syn. rapa) L. FEBS Lett 473: 139-144.

Zhang, J. and Madden, T.L. 1997. "PowerBLAST”: A new network BLAST application for interactive or automated sequence analysis and annotation." Genome Res. 7: 649-656.

Zhao, Y., Feng, X.-H., Watson, J.C., Bottino, P.J. and Kung, S.-D. 1994. Molecular cloning and biochemical characterization of a receptor-like serine/threonine kinase from rice. Plant Mol. Biol. 26: 791-803.

**Figure legends**

*Figure 1.* Organisation of the SCRL and LCR genes in the *Arabidopsis* genome. a. Schematic representation of the five chromosomes of *A. thaliana* showing the positions of 23 of the SCRL genes. Closely linked genes are separated by a comma. Numbers refer to gene numbers (e.g. 1: *SCRL1*). b. Organisation of the two largest clusters of SCRL and LCR genes. Boxes represent coding regions, arrows indicate the predicted orientation of each gene.

*Figure 2.* Analysis of the largest cluster of SCRL genes. a. Dot matrix comparison of the nucleotide sequence of the SCRL gene cluster region with itself. The positions of the coding regions of the five SCRL genes are indicated by stippled boxes on the diagonal. b. Schematic representation of the gene cluster showing two highly similar regions (stippled boxes) that span the *SCRL4* and *SCRL5* genes. The coding regions of the five SCRL genes are represented by boxes, arrows indicate the predicted orientation of each gene. c. Graphical representation of percent GC content in the region of the cluster based on a window size of 200 bp.

*Figure 3.* Phylogenetic analysis of the two largest clusters of SCRL and LCR genes. Neighbour-joining trees were constructed from manually-optimised alignments of the entire deduced polypeptides. Bootstrap values were calculated from 1000 replicates. The structure of each gene cluster is shown schematically to the right of each phylogenetic tree.

*Figure 4.* Alignment of the deduced polypeptides of the *Arabidopsis* SCRL gene family (At SCRL1-29) with deduced polypeptides of four *SCR* alleles (Bo SCR3, Bo SCR16, Bo SCR6 and Br SCR12) and *SCRL1* (Bo SCRL1) from *Brassica oleracea* or *B. rapa*. Consensus

residues (shared by at least 16 of the aligned sequences) are highlighted and indicated above the scale.

*Figure 5.* Alignment of the deduced polypeptides of 32 representative members of the *Arabidopsis* LCR gene family (At LCR) with the deduced polypeptides of three PCP genes from *Brassica oleracea (*Bo PCP). Consensus residues (shared by at least 16 of the aligned sequences) are highlighted and indicated above the scale.

*Figure 6.* Evidence for positive selection within the SCRL coding sequences. a. Schematic representation of the SCRL gene regions analysed in part b. The regions analysed were located between the conserved cysteine codons as shown. b. Graph showing average ratios of nonsynonymous to synonymous (dn/ds) substitutions for each region (A to E) of the coding sequences of the largest SCRL gene cluster, genes *SCRL4* to *SCRL8* (4 to 8). For several regions we noted a high level of variability in the calculated dn/ds ratio depending on which other member of the cluster a gene was compared with. For this reason, each region (A to E) of each of the five genes was compared with the corresponding region of each of the other members of the gene cluster and an average dn/ds ratio was calculated.

*Figure 7.* Consensus sequences of the deduced, mature SCRL and LCR protein families. c: cysteine; g: glycine; x: any amino acid (subscripts refer to multiple residues, the number of which may be variable as indicated). Cysteines that are conserved in all members of the family are shown in upper case.

*Figure 8.* RT-PCR analysis of the expression patterns of seven SCRL and five LCR genes. Actin gene expression was analysed as a control. After the first series of 30 cycles (left panels),

35 µl of PCR product was analysed (5 µl for the actin control). For the second series of 30 PCR cycles (right panels),1 µl of the products from the first series of amplifications was used as a template and 10 µl of the final product was analysed. Genomic DNA samples were not re-amplified but some of the product from the first series of amplifications was run along with the products of the second amplification for comparison. The positions of DNA size markers are shown to the left in kilobases (kbp). Arrowheads indicate faint bands. S: stem, R: root, L: rosette leaves, B: flower buds, g: genomic DNA, c: no DNA control.