A Large Family of Genes That Share Homology with *CLAVATA3*

J. Mark Cock* and Sheila McCormick

Reproduction et Développement des Plantes, Unité Mixte de Recherche 5667, Ecole Normale Supérieure de Lyon, 46 Allée d'Italie, 69364 Lyon cedex 07, France (J.M.C.); and Plant Gene Expression Center, United States Department of Agriculture/Agricultural Research Service-University of California-Berkeley, 800 Buchanan Street, Albany, California 94710 (S.M.)

The receptor kinase superfamily in plants includes a large number of genes, but very little is known about the ligands that bind to these receptors. In the Arabidopsis genome, for example, 340 receptor kinase genes have been identified (The Arabidopsis Genome Initiative, 2000). The proteins encoded by this superfamily share a common overall structure: They are all integral membrane proteins with a single membrane-spanning region and a cytosolic kinase domain. The superfamily can be divided into at least 20 different families based on the structure of the predicted extracellular domains, which differ markedly between families (The Arabidopsis Genome Initiative, 2000; Torii and Clark, 2000). Even within families the extracellular domain can exhibit a high level of variability. For example, within the family of receptor kinases with Leu-rich repeats (LRRs) in their extracellular domains, the number of LRRs present varies from three to more than 20 and the LRRs may be associated with other protein domains.

Several members of the plant receptor kinase gene family have been shown to play key regulatory roles in various aspects of plant development (Torii and Clark, 2000) and defense (Song et al., 1995). In contrast, their mode of action is poorly understood and, until recently, the members of this family have been described as "receptor-like kinases" because none of them had been shown to interact with a ligand. In a recent report, however, Trotochaud et al. (2000) showed that the receptor kinase CLAVATA1 interacts physically with a small secreted peptide encoded by the *CLAVATA3* (*CLV3*) locus. This is supported by genetic evidence that also indicates that the products of *CLV1* and *CLV3* function in close association (Clark et al., 1995; Fletcher et al., 1999). A second candidate for a receptor kinase ligand gene is *SCR*, which has been shown to encode the male component of the self-incompatibility response in *Brassica oleracea* (Schopfer et al., 1999). Based on several lines of data, including the highly polymorphic nature of the *SCR* gene, its expression pattern, and its location at the *S* locus, it is very likely that the SCR peptide binds to the S-locus receptor kinase (SRK) as a ligand.

The extracellular domain of CLV1 contains 21 LRRs, whereas SRK possesses an S domain with 12 conserved cysteines and a PTDT box (for review, see Torii and Clark, 2000). Both *CLV1* and *SRK* are members of large gene families and it was initially hoped that identification of *CLV3* and *SCR* would facilitate the identification of ligands for other receptor kinases of these gene families. However, paralogues of these genes have been more difficult to identify than was expected. For example, initial searches with the sequences of several *SCR* alleles did not identify any significant homologies in the databases (Schopfer et al., 1999). A more recent study, however, has identified a large family of *SCR*-like genes in Arabidopsis (Vanoosthuyse et al., 2001). This family of genes exhibits an extreme level of diversity with only a small number of residues being conserved throughout the family. The high level of sequence diversity is the main reason these genes were not detected in initial database searches. In the study described here, we have searched for genes with similarity to *CLV3*, employing a similar approach to that used to identify the *SCR*-related genes.

Initial database searches with the CLV3 sequence yielded a poor match with the maize (*Zea mays*) embryo-surrounding region (ESR) proteins (high E values, i.e. low probability scores, in the range of 0.15–0.45). ESR genes are expressed in a specific zone of the developing endosperm and, like *CLV3*, encode small, secreted polypeptides (Opsahl-Ferstad et al., 1997). CLV3 and the ESR sequences share a short, conserved region of 14 amino acids but are otherwise unrelated at the sequence level. Two independently isolated, mutant alleles of *CLV3* (*clv3-1* and *clv3-5*) contain a point mutation within this conserved region (encoding clv3^{G75A}) indicating that it is important for gene function (Fletcher et al., 1999). Based on these observations, which indicate that the conserved region is an important functional element of CLV3, and presumably of the ESR proteins, we carried out a database search for related sequences with the same conserved region. This was done, initially, by searching with tBlastn (Zhang and Madden, 1997) and either the entire CLV3 and ESR2 sequences, or their conserved regions, as queries. The newly retrieved sequences then were used to repeat the search in an

* Corresponding author; e-mail Mark.Cock@ens-lyon.fr; fax 334–7272–8600.

iterative manner. The search was terminated when no novel sequences were retrieved. A total of 42 related sequences were identified including 28 genomic sequences and 13 sequences that were represented only by expressed sequence tags (ESTs). We have named these genes *CLE* (for *CLAVATA3/ESR-related)*. The availability of ESTs indicates that the sequences identified correspond to expressed genes. The EST sequences were derived from seven different plant species, including both monocots and dicots (Table I). No homologs were identified from outside the plant kingdom. Based on the ESTs that were identified, the members of this family appear to be expressed in a wide range of tissues (Table I).

In total, 24 *CLE* genes were found in the Arabidopsis genome. They are scattered throughout the genome, in most cases as individual genes although a head-to-tail cluster of four genes was found within a 16.5-kbp region on chromosome 2 (*CLE4* to *CLE7*). Of these genes, *CLE5*, *CLE6*, and *CLE7* are more closely related to each other than to the other members of the family, suggesting that they have arisen by local gene duplication events. *CLE4*, on the other hand, is more similar to *CLE3* on chromosome 1 than it is to the three other members of the cluster. The similarity between *CLE4* and *CLE3* suggests that there has been a recent transfer of one of these genes from one of these loci to the other. A similar phenomenon has been observed for resistance gene clusters in tomato (Parniske and Jones, 1999).

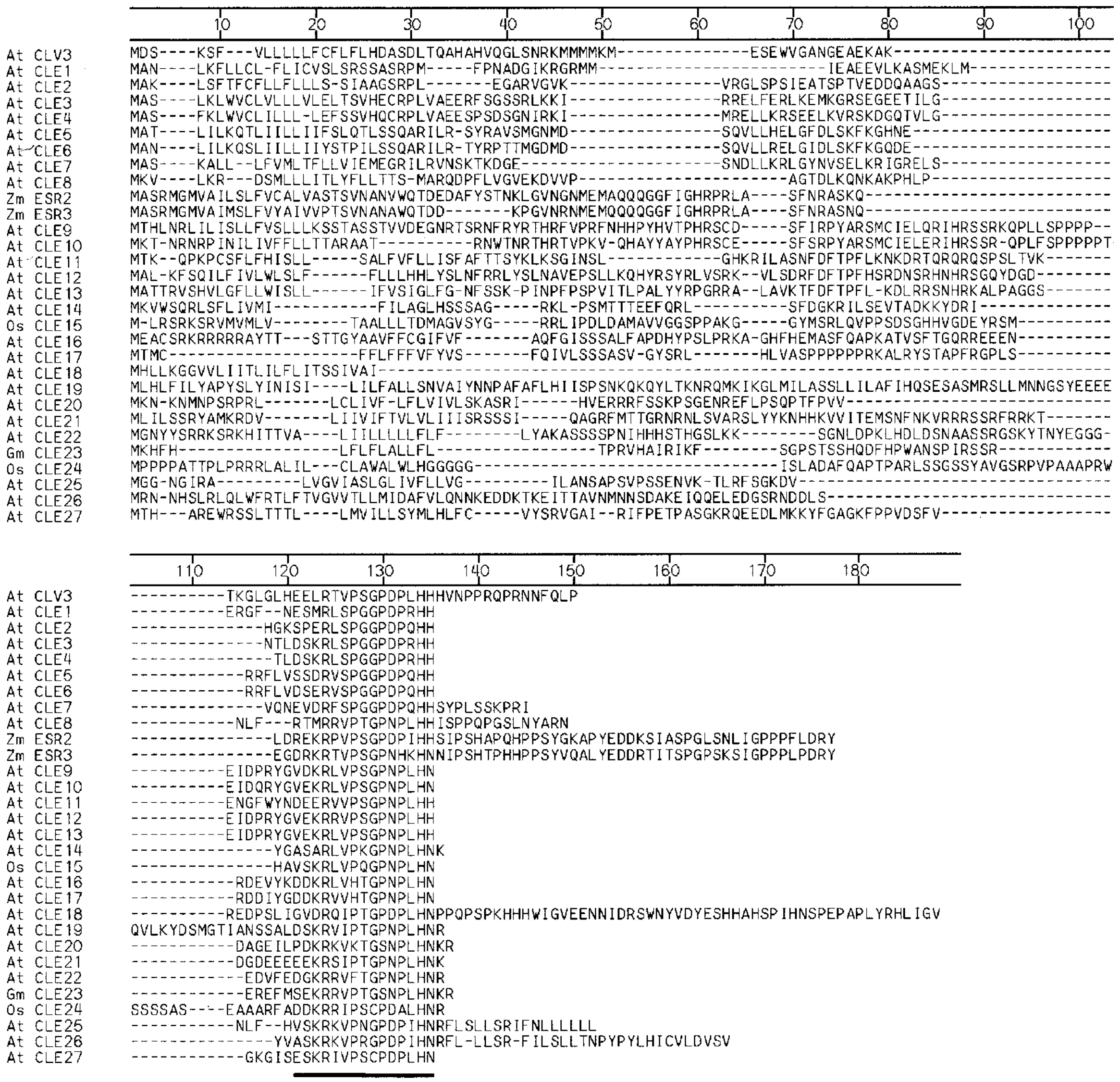
Apart from *CLV3*, where the reading frame is interrupted by two introns (Fletcher et al., 1999), all of the sequences identified here consisted of one exon with a single open reading frame. Of the EST sequences identified, 15 (including four that correspond to Arabidopsis genomic sequences) included the 3′ end of the open reading frame and analysis of these sequences showed that they all contained a stop codon between zero and 13 residues downstream of the conserved region. This supports the gene structures that we propose. In particular, it indicates that additional coding exons do not occur downstream of the sequences identified, at least for the transcript represented by the ESTs.

Twenty-seven of the sequences had complete open reading frames that were supported by good quality sequence data (Fig. 1). All 27 genes contain short open reading frames that are predicted to encode small polypeptides (average molecular mass is 7,674 ± 1,840 D) with hydrophobic regions at the N-terminal end. In all but four cases the hydrophobic N termini are predicted to act as signal peptides (as predicted with SignalP; Nielsen et al., 1997), indicating that the gene products of most of the members of this family are secreted. The four exceptions, CLE16, CLE25, CLE26, and CLE27, are predicted to possess signal anchors. Note, however, that these predictions need to be tested experimentally and it is possible that these four proteins possess non-standard signal peptides.

The majority of the predicted mature polypeptides are highly basic (average pI 9.49 ± 1.57) and hydrophilic throughout their length. Somewhat different hydrophilicity profiles were observed for CLE19 and CLE15, where a short hydrophobic region occurs within the polypeptide, and for CLE25 and CLE26, where predominantly hydrophobic C termini follow the conserved region.

Table I. *ESTs that share sequence similarity with CLAVATA3*

Gene Name	Organism	Tissue of Origin	EST Accession No.	Genomic Sequence Accession No.
ESR3	Maize	Kernel	AW448658	X99970
CLE28	Maize	Mixed adult tissues	BE025272	–
CLE29	Wheat (*Triticum aestivum*)	Spike (before anthesis)	BE498847	–
CLE30	Wheat	Endosperm	BE401912, BE414012, AW448658	–
CLE31	Wheat	Endosperm	BE402259	–
CLE32	Cotton (*Gossypium hirsutum*)	Immature fiber	AW187682	–
CLE16	Arabidopsis	Mixed tissues	AI993471, R30249	AC005560
CLE2	Arabidopsis	Mixed tissues	T44515	AL161548
CLE9	Arabidopsis	7-D seedling (NaCl treated)	U74112	AC013427
CLE4	Arabidopsis	Rosette (4 to 7 weeks)	AI998558	AC005311
CLE33	Tomato (*Lycopersicon esculentum*)	Flower buds	AW929449	–
CLE34	Soybean (*Glycine max*)	Hypocotyl (etiolated)	BE475290	–
CLE23	Soybean	Immature flowers	AI748451, AW119439, BE821206	–
CLE35	*Medicago truncatula*	Root (phosphate starved)	AW329414	–
CLE36	*M. truncatula*	Roots (+*Glomus versiform*)	AW586793	–
CLE37	*M. truncatula*	Root nodules	AL380419	–
CLE38	*M. truncatula*	Root nodules	AL381237, AL381238	–
CLE39	*M. truncatula*	Senescent root nodules	BE999212	–



```
                  10        20        30        40        50        60        70        80        90       100
At CLV3    MDS----KSF---VLLLLLFCFLFLHDASDLTQAHAHVQGLSNRKMMMMKM--------------ESEWVGANGEAEKAK-----------------------
At CLE1    MAN----LKFLLCL-FLICVSLSRSSASRPM-----FPNADGIKRGRMM------------------------IEAEEVLKASMEKLM---------------
At CLE2    MAK----LSFTFCFLLFLLLS-SIAAGSRPL-------EGARVGVK----------------VRGLSPSIEATSPTVEDDQAAGS------------------
At CLE3    MAS----LKLWVCLVLLLVLELTSVHECRPLVAEERFSGSSRLKKI----------------RRELFERLKEMKGRSEGEETILG------------------
At CLE4     MAS----FKLWVCLILLL-LEFSSVHQCRPLVAEESPSDSGNIRKI----------------MRELLKRSEELKVRSKDGQTVLG------------------
At CLE5    MAT----LILKQTLIILLIIFSLQTLSSQARILR-SYRAVSMGNMD----------------SQVLLHELGFDLSKFKGHNE---------------------
At CLE6    MAN----LILKQSLIILLIIYSTPILSSQARILR-TYRPTTMGDMD----------------SQVLLRELGIDLSKFKGQDE---------------------
At CLE7    MAS----KALL--LFVMLTFLLVIEMEGRILRVNSKTKDGE---------------------SNDLLKRLGYNVSELKRIGRELS------------------
At CLE8    MKV----LKR---DSMLLLITLYFLLTTS-MARQDPFLVGVEKDVVP----------------------AGTDLKQNKAKPHLP-------------------
Zm ESR2    MASRMGMVAILSLFVCALVASTSVNANVWQTDEDAFYSTNKLGVNGNMEMAQQQGGFIGHRPRLA----SFNRASKQ--------------------------
Zm ESR3    MASRMGMVAIMSLFVYAIVVPTSVNANAWQTDD-------KPGVNRNMEMQQQQGGFIGHRPRLA----SFNRASNQ--------------------------
At CLE9    MTHLNRLILISLLFVSLLLKSSTASSTVVDEGNRTSRNFRYRTHRFVPRFNHHPYHVTPHRSCD-----SFIRPYARSMCIELQRIHRSSRKQPLLSPPPP--
At CLE10   MKT-NRNRPINILIVFFLLTTARAAT----------RNWTNRTHRTVPKV-QHAYYAYPHRSCE-----SFSRPYARSMCIELERIHRSSR-QPLFSPPPPPT
At CLE11   MTK--QPKPCSFLFHISLL-----SALFVFLLISFAFTTSYKLKSGINSL------------GHKRILASNFDFTPFLKNKDRTQRQRQSPSLTVK-------
At CLE12   MAL-KFSQILFIVLWLSLF------FLLLHHLYSLNFRRLYSLNAVEPSLLKQHYRSYRLVSRK--VLSDRFDFTPFHSRDNSRHNHRSGQYDGD--------
At CLE13   MATTRVSHVLGFLLWISLL------IFVSIGLFG-NFSSK-PINPFPSPVITLPALYYRPGRRA--LAVKTFDFTPFL-KDLRRSNHRKALPAGGS-------
At CLE14   MKVWSQRLSFLIVMI-----------FILAGLHSSSAG-----RKL-PSMTTTEEFQRL----------SFDGKRILSEVTADKKYDRI--------------
Os CLE15   M-LRSRKSRVMVMLV--------TAALLLTDMAGVSYG-----RRLIPDLDAMAVVGGSPPAKG----GYMSRLQVPPSDSGHHVGDEYRSM-----------
At CLE16   MEACSRKRRRRRAYTT---STTGYAAVFFCGIFVF-------AQFGISSSALFAPDHYPSLPRKA-GHFHEMASFQAPKATVSFTGQRREEEN----------
At CLE17   MTMC--------------------FFLFFFVFYVS-------FQIVLSSSASV-GYSRL-------HLVASPPPPPPRKALRYSTAPFRGPLS----------
At CLE18   MHLLKGGVVLIITLILFLITSSIVAI-----------------------------------------------------------------------------
At CLE19   MLHLFILYAPYSLYINISI----LILFALLSNVAIYNNPAFAFLHIISPSNKQKQYLTKNRQMKIKGLMILASSLLILAFIHQSESASMRSLLMNNGSYEEEE
At CLE20   MKN-KNMNPSRPRL-------LCLIVF-LFLVIVLSKASRI------HVERRRFSSKPSGENREFLPSQPTFPVV----------------------------
At CLE21   MLILSSRYAMKRDV-------LIIVIFTVLVLIIISRSSSI-----QAGRFMTTGRNRNLSVARSLYYKNHHKVVITEMSNFNKVRRRSSRFRRKT-------
At CLE22   MGNYYSRRKSRKHITTVA----LIILLLLLFLF--------LYAKASSSSPNIHHHSTHGSLKK--------SGNLDPKLHDLDSNAASSRGSKYTNYEGGG-
Gm CLE23   MKHFH-----------------LFLFLALLFL-----------------TPRVHAIRIKF--------SGPSTSSHQDFHPWANSPIRSSR------------
Os CLE24   MPPPPATTPLPRRRLALIL---CLAWALWLHGGGGG--------------------------------ISLADAFQAPTPARLSSGSSYAVGSRPVPAAAPRW
At CLE25   MGG-NGIRA---------LVGVIASLGLIVFLLVG---------ILANSAPSVPSSENVK-TLRFSGKDV---------------------------------
At CLE26   MRN-NHSLRLQLWFRTLFTVGVVTLLMIDAFVLQNNKEDDKTKEITTAVNMNNSDAKEIQQELEDGSRNDDLS------------------------------
At CLE27   MTH---AREWRSSLTTTL----LMVILLSYMLHLFC-----VYSRVGAI--RIFPETPASGKRQEEDLMKKYFGAGKFPPVDSFV------------------

                110       120       130       140       150       160       170       180
At CLV3    ----------TKGLGLHEELRTVPSGPDPLHHHVNPPRQPRNNFQLP
At CLE1    ----------ERGF--NESMRLSPGGPDPRHH
At CLE2    --------------HGKSPERLSPGGPDPQHH
At CLE3    --------------NTLDSKRLSPGGPDPRHH
At CLE4    ---------------TLDSKRLSPGGPDPRHH
At CLE5    ------------RRFLVSSDRVSPGGPDPQHH
At CLE6    ------------RRFLVDSERVSPGGPDPQHH
At CLE7    --------------VQNEVDRFSPGGPDPQHHSYPLSSKPRI
At CLE8    -----------NLF---RTMRRVPTGPNPLHHISPPQPGSLNYARN
Zm ESR2    ---------------LDREKRPVPSGPDPIHHSIPSHAPQHPPSYGKAPYEDDKSIASPGLSNLIGPPPFLDRY
Zm ESR3    ---------------EGDRKRTVPSGPNHKHNNIPSHTPHHPPSYVQALYEDDRTITSPGPSKSIGPPPLPDRY
At CLE9    ----------EIDPRYGVDKRLVPSGPNPLHN
At CLE10   ----------EIDQRYGVEKRLVPSGPNPLHN
At CLE11   ----------ENGFWYNDEERVVPSGPNPLHH
At CLE12   ----------EIDPRYGVEKRRVPSGPNPLHH
At CLE13   ----------EIDPRYGVEKRLVPSGPNPLHH
At CLE14   ---------------YGASARLVPKGPNPLHNK
Os CLE15   ---------------HAVSKRLVPQGPNPLHN
At CLE16   -----------RDEVYKDDKRLVHTGPNPLHN
At CLE17   -----------RDDIYGDDKRVVHTGPNPLHN
At CLE18   ----------REDPSLIGVDRQIPTGPDPLHNPPQPSPKHHHWIGVEENNIDRSWNYVDYESHHAHSPIHNSPEPAPLYRHLIGV
At CLE19   QVLKYDSMGTIANSSALDSKRVIPTGPNPLHNR
At CLE20   -----------DAGEILPDKRKVKTGSNPLHNKR
At CLE21   -----------DGDEEEEEKRSIPTGPNPLHNK
At CLE22   ------------EDVFEDGKRRVFTGPNPLHNR
Gm CLE23   ------------EREFMSEKRRVPTGSNPLHNKR
Os CLE24   SSSSAS----EAAARFADDKRRIPSCPDALHNR
At CLE25   -----------NLF--HVSKRKVPNGPDPIHNRFLSLLSRIFNLLLLLL
At CLE26   ---------------YVASKRKVPRGPDPIHNRFL-LLSR-FILSLLTNPYPYLHICVLDVSV
At CLE27   ------------GKGISESKRIVPSCPDPLHN
```

Figure 1. Alignment of the deduced polypeptides of the CLE gene family (CLE1 to CLE27) with CLV3, ESR2, and ESR3. The multiple alignment was constructed using Lasergene sequence analysis software (DNASTAR, London) and optimized manually. The conserved region is indicated by a bar. The CLV3 sequence is based on the gene sequence submitted by Flecher et al. (1999; accession no. AF126009) that differs by five substitutions and two indels from the sequence submitted by the Arabidopsis Genome Initiative (Lin et al., 1999; accession no. AC006233). The two deduced CLV3 polypeptides differ only at residue seven (Val or Leu). At, Arabidopsis; Gm, soybean; Os, rice (*Oryza sativa*); Zm, maize.

The conserved region was located at or near the C-terminal end of the majority of the sequences we identified (see Fig. 1). CLV3 and the ESR proteins are among the exceptions that possess additional amino acid residues C terminal to the conserved region. The conserved region contains a number of highly conserved amino acids, including between two and three Pro residues. Secondary structure predictions indicate that it adopts a turn structure and is exposed on the surface of the protein. This is consistent with this region being involved in interactions with other proteins. The Gly^{75} codon that is mutated in *clv3-1* and *clv3-5* is highly conserved in the CLE genes. The only exceptions are *CLE24* and *CLE27*, where a Cys codon occurs at this position.

Although the CLE proteins share an overall resemblance when parameters such as length, charge, and hydrophilicity are considered, at the amino acid sequence level they are highly divergent (Fig. 1). This sequence diversity could be the result of either: (a) the conserved domain being associated with diverse, unrelated (non-homologous) protein domains, or (b) extensive sequence diversification having occurred within the CLE family (or to a combination of these two phenomena). The second possibility would not be unprecedented because a very high level of sequence diversity was also observed for the *SCR*-like (SCRL) gene family in Arabidopsis despite the fact that there is good reason to believe that the SCRL genes are homologous throughout their length (Vanoosthuyse et al., 2001).

There is evidence that processing of secreted signaling polypeptides occurs in plants and it has been pointed out that CLV3 contains a potential dibasic processing site that could be recognized by subtilases (Schaller and Ryan, 1994; Berger and Altmann, 2000). More recent evidence indicates that CLV3 is not processed, however. Anti-CLV3 antibodies detect a protein of approximately the size expected for the secreted, "unprocessed" polypeptide (8.4 kD) in Arabidopsis extracts (Trotochaud et al., 2000). It would, nonetheless, be interesting to determine whether other members of the CLE family are processed. If the more divergent regions of the proteins do not constitute part of the final mature gene product, this might explain, at least in part, the high level of sequence divergence in these regions.

The ESR proteins may represent ligands for LRR receptor kinases that are expressed in the embryo or endosperm, and it is tempting to speculate that the CLE proteins also represent ligands for LRR receptor kinases. Functional analysis of these genes will be required to confirm this. An additional problem will be assigning ligands to specific members of the receptor kinase superfamily. The data currently available indicate that it will not be straightforward to do this, based simply on sequence data. For example, putative ligands have recently been identified for the pollen-specific receptor kinases, which have five to six LRRs (Muschietti et al., 1998). These putative ligands are small Cys-rich proteins, unrelated to CLV3 and the CLE proteins, and are more like SCR, which is also rich in Cys (although there is no obvious homology at the sequence level; Tang et al., 2000). The ligand for another LRR receptor kinase, BRI1, which has 27 LRRs, is thought to be a brassinosteroid (He et al., 2000). Note, however, that BRI1 also has a 70-amino acid "island" in the extracellular domain that is not found in other LRR receptor kinases (Li and Chory, 1997). Therefore, it seems likely that, even within the LRR family of receptor kinases, the ligands will turn out to be structurally diverse. We suggest that a reasonable starting point for assigning candidate ligands to receptors in this family may be to group the receptors according to the number of LRRs in their extracellular domains.

Small genes, like those of the CLE family, are often overlooked by automated annotation programs (Ride et al., 1999; Vanoosthuyse et al., 2001). Of the 28 CLE sequences identified in genome sequence (24 from Arabidopsis and four from rice), only eight were annotated as genes and, of these, only five of the annotations corresponded to those proposed here. The approach we used to identify the CLE genes consisted of identifying a conserved motif, based on the alignment of a limited number of members of the gene family, and using this information to manually select additional gene family members from a pool of potential paralogues identified in a sensitive homology search using tBlastn. The SPH (Ride et al., 1999), SCRL and LCR (Vanoosthuyse et al., 2001) gene families were identified in a similar manner, indicating that this approach may be generally applicable to the identification of families of genes that encode small proteins.

Received February 5, 2001; returned for revision February 22, 2001; accepted March 9, 2001.

LITERATURE CITED

Berger D, Altmann T (2000) Genes Dev **14:** 1119–1131

Clark SE, Running MP, Meyerowitz EM (1995) Development **121:** 2057–2067

Fletcher JC, Brand U, Running MP, Simon R, Meyerowitz EM (1999) Science **283:** 1911–1914

He Z, Wang ZY, Li J, Zhu Q, Lamb C, Ronald P, Chory J (2000) Science **288:** 2360–2363

Li J, Chory J (1997) Cell **90:** 929–938

Muschietti JP, Eyal Y, McCormick S (1998) Plant Cell **10:** 319–330

Nielsen H, Engelbrecht J, Brunak S, von Heijne G (1997) Protein Eng **10:** 1–6

Opsahl-Ferstad HG, Le Deunff E, Dumas C, Rogowsky PM (1997) Plant J **12:** 235–246

Parniske M, Jones JD (1999) Proc Natl Acad Sci USA **96:** 5850–5855

Ride JP, Davies EM, Franklin FCH, Marshall DF (1999) Plant Mol Biol **39:** 927–932

Schaller A, Ryan CA (1994) Proc Natl Acad Sci USA **91:** 11802–11806

Schopfer CR, Nasrallah ME, Nasrallah JB (1999) Science **286:** 1697–1700

Song WY, Wang GL, Chen LL, Kim HS, Pi LY, Holsten T, Gardner J, Wang B, Zhai WX, Zhu L-H et al. (1995) Science **270:** 1804–1806

Tang W, Ezcurra I, Cotter R, Muschietti J, McCormick S (2000) ASCB meeting, abstract no. L68. San Francisco

The Arabidopsis Genome Initiative (2000) Nature **408:** 796–815

Torii KU, Clark SE (2000) *In* M Kreis, JC Walker, eds, Advances in Botanical Research, Thematic Volume: Plant Protein Kinases, Vol 32. Academic Press, London, pp 270–298

Trotochaud AE, Jeong S, Clark SE (2000) Science **289:** 613–617

Vanoosthuyse V, Miege C, Dumas C, Cock JM (2001) Plant Mol Biol **46:** 17–34

Zhang J, Madden TL (1997) Genome Res **7:** 649–656