Receptor-like kinases from *Arabidopsis* form a monophyletic gene family related to animal receptor kinases

Shin-Han Shiu and Anthony B. Bleecker*

Department of Botany and Laboratory of Genetics, University of Wisconsin, Madison, WI 53706

Edited by Elliot M. Meyerowitz, California Institute of Technology, Pasadena, CA, and approved July 6, 2001 (received for review March 22, 2001)

Plant receptor-like kinases (RLKs) are proteins with a predicted signal sequence, single transmembrane region, and cytoplasmic kinase domain. Receptor-like kinases belong to a large gene family with at least 610 members that represent nearly 2.5% of *Arabidopsis* protein coding genes. We have categorized members of this family into subfamilies based on both the identity of the extracellular domains and the phylogenetic relationships between the kinase domains of subfamily members. Surprisingly, this structurally defined group of genes is monophyletic with respect to kinase domains when compared with the other eukaryotic kinase families. In an extended analysis, animal receptor kinases, Raf kinases, plant RLKs, and animal receptor tyrosine kinases form a well supported group sharing a common origin within the superfamily of serine/threonine/tyrosine kinases. Among animal kinase sequences, *Drosophila* Pelle and related cytoplasmic kinases fall within the plant RLK clade, which we now define as the RLK/Pelle family. A survey of expressed sequence tag records for land plants reveals that mosses, ferns, conifers, and flowering plants have similar percentages of expressed sequence tags representing RLK/Pelle homologs, suggesting that the size of this gene family may have been close to the present-day level before the diversification of land plant lineages. The distribution pattern of four RLK subfamilies on *Arabidopsis* chromosomes indicates that the expansion of this gene family is partly a consequence of duplication and reshuffling of the *Arabidopsis* genome and of the generation of tandem repeats.

The ability to perceive and process information from chemical signals via cell surface receptors is a basic property of all living systems. In animals, the family of receptor tyrosine kinases (RTKs) mediates many signaling events at the cell surface (1, 2). This class of receptors is defined structurally by the presence of a ligand-binding extracellular domain, a single membrane-spanning domain, and a cytoplasmic tyrosine kinase domain. In plants, receptor-like kinases (RLKs) are a class of transmembrane kinases similar in basic structure to the RTKs (3). In *Arabidopsis* alone, it has been reported that there are more than 300 RLKs (4, 5). In the limited cases where a functional role has been identified for plant RLKs, they have been implicated in a diverse range of signaling processes, such as brassinosteroid signaling via BRI1 (6), meristem development controlled by CLV1 (7), perception of flagellin by FLS2 (8), control of leaf development by Crinkly4 (9), regulation of abscission by HAESA (10), self-incompatibility controlled by SRKs (11), and bacterial resistance mediated by Xa21 (12). Putative ligands for SRK (13, 14), CLV1 (15, 16), BRI1 (17), and FLS2 (18) have recently been identified. Proteins interacting with the kinase domains of RLKs *in vitro* have also been found (19–21).

Plant RLKs can be distinguished from animal RTKs by the finding that all RLKs examined to date show serine/threonine kinase specificity, whereas animal receptor kinases, with the exception of transforming growth factor-β (TGF-β) receptors, are tyrosine kinases. In addition, the extracellular domains of RLKs are distinct from most ligand-binding domains of RTKs identified so far (1, 2). These differences raise the question of the specific evolutionary relationship between the RTKs and RLKs within the recognized superfamily of related eukaryotic serine/threonine/tyrosine protein kinases (ePKs). An earlier phylogenetic analysis (22), using the six RLK sequences available at the time, indicated a close relationship between plant sequences and animal RTKs, although RLKs were placed in the "other kinase" category. A more recent analysis using only plant sequences led to the conclusion that the 18 RLKs sampled seemed to form a separate family among the various eukaryotic kinases (23). The recent completion of the *Arabidopsis* genome sequence (5) provides an opportunity for a more comprehensive analysis of the relationships between these classes of receptor kinases.

To understand the evolution of the RLK family and its relationship with other kinase families and provide a framework to facilitate the prediction of RLK function, we set out to conduct a genome-wide survey of RLK-related sequences in *Arabidopsis.* Through a phylogenetic analysis of the conserved kinase domains, we sought to determine (*i*) whether RLKs belong to a monophyletic group when compared with other ePKs and (*ii*) how the RLKs are related to animal receptor kinases. To investigate the relationship between the evolution of land plants and the expansion of the RLK family, we performed a survey of expressed sequence tags (ESTs) for a variety of organisms. Finally, we looked into the chromosomal distribution of four RLK subfamilies to investigate the potential mechanisms contributing to the expansion of this gene family in *Arabidopsis*.

Materials and Methods

Sequence Selection. *RLKs.* All published plant RLK sequences were retrieved, and their kinase domain sequences were used to conduct batch BLAST analysis (24) for related sequences in Viridiplantae, with an E value cutoff of 1×10^{-10}. The cutoff was chosen based on multiple phylogenetic analyses using data sets generated from cutoff E values of 1×10^{-20}, 1×10^{-10}, and 1. All known RLKs were recovered at 1×10^{-20}; therefore, a more relaxed criterion, 1×10^{-10}, was used to retrieve all potentially related genes. The search results were merged, and redundant sequences were deleted. As of February 2001, more than 900 nonredundant candidates of plant RLKs or related kinases were present in GenBank, and they were used for subsequent phylogenetic analysis. For a complete list of genes in the RLK/Pelle gene family, see supporting information, which is published on the PNAS web site, www.pnas.org. The gene name or accession numbers for RLKs shown in the manuscript are as follows: ARK2 (AAB33486), At2g15300 (AAD26903), At2g19130 (AAD12030), At2g24370 (AAD18110), At2g33580

EVOLUTION

This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: RLK, receptor-like kinase; RLCK, receptor-like cytoplasmic kinases; ePK, eukaryotic protein kinase; RTK, receptor tyrosine kinase; RSK, receptor serine/threonine kinase; APH(3′)III, aminoglycoside kinase III; EST, expressed sequence tag; TGF-β, transforming growth factor-β; LRR, leucine-rich repeat.

*To whom reprint requests should be addressed. E-mail: bleecker@facstaff.wisc.edu.

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

(AAB80675), At2g45340 (AAB82629), At4g11480 (CAB82153), At4g26180 (CAA18124), At4g39110 (CAB43626), BRI1 (AAC49810), CLV1 (AAF26772), ERECTA (AAC49302), HAESA (CAB79651), PR5K (AAC49208), PERK (AAD43169), RPK1 (AAD11518), RKF1 (AAC50043), RKF3 (AAC50045), RKL1 (AAC95351), TMK1 (JQ1674), WAK1 (CAA08794), F1P2.130 (CAB61984), F4I18.11 (T02456), F13F21.28 (AAD43169), F15A17.170 (CAB86081), F17J16.160 (CAB86939), F18L15.120 (CAB62031), F23E13.70 (CAA18124), F23M19.11 (AAD39611), F27K19.130 (CAB80791), MLD14.2 (BBA99679), and T20L15.220 (CAB82765).

Representatives of the eukaryotic protein kinase (ePK) superfamily. Based on Hanks and Hunter (22), plant and animal sequences from each ePK family were chosen. Plant kinases that seemed to be unique to plants were also included in this study (23). Their accession numbers are as follows. *Arabidopsis* sequences: CDC2a (AAB23643), CPK7 (AAB03247), CKI1 (CAA55395), CKA1 (BAA01090), AME2 (BAA08215), MKK3 (BAA28829), MEKK1 (BAA09057), NAK (AAA18853), NPH1 (AAC01753), PVPK-like PK5 (BAA01715), CTR1 (AAA32779), MRK1 (BAA22079), S6K-like PK1 (AAA21142), GSK3β (CAA64408), GSK3ι (CAA68027), SnRK2-like PROKINa (AAA32845), and Tousled (AAA32874); human sequences: CaMK1 (NP003647), CDK3 (NP001249), CK1α1 (NP001883), CK2α (CAB65624), GRK6 (P43250), RK (Q15835), Hunk (NP055401), CLK1 (P49759), MAPK10 (P53779), MAPKK1 (Q02750), MAPKKK1 (Q13233), cAPK (P17612), Raf1 (TVHUF6), c-SRC (P12931), TLK1 (NP036422), and TTK (A42861).

Animal receptor kinases. One representative human receptor tyrosine kinase sequence was selected from each RTK subfamily (1, 2) as follows: AXL (NP001690), DDR (Q08345), EGFR (P00533), EPH (P21709), FGFR2 (P21802), HGFR (P08581), IR (NP000199), KLG-like PTK7 (AAC50484), LTK (P29376), MuSK (AAB63044), PDGFRβ (PFHUGB), RET (S05582), RYK (I37560), TIE (P35590), TRKα (BAA34355), and VEGFR (P17948). Human TGF-β receptors (TGFβR I, P36897; TGFβR II, P37173) were chosen as animal representatives of receptor serine/threonine kinases.

Sequence Annotation, Alignment, and Phylogenetic Analysis. *Delineation of structural domains.* Structural domains of all sequences were annotated according to SMART (25) and Pfam (26) databases. The receptor-like kinase configuration was determined by the presence of putative signal sequences and extracellular domains. Sequences without signal sequences, transmembrane regions, or putative extracellular domains were also included in the analysis. The kinase domain sequences delineated initially according to sequence prediction databases were modified to include missing or exclude excessive flanking sequences according to the subdomain signature of eukaryotic kinases (22).

Alignment of sequences. The sizes of the kinase domains range from 250 to 300 aa. These sequences were compiled and aligned by using CLUSTALX (27). The weighing matrices used were BLOSUM62 or PAM250 with the penalty of gap opening 10 and gap extension 0.2. The alignments generated by these two scoring tables are similar to each other and were manually adjusted according to the subdomain signatures of eukaryotic kinases (22). The alignment for all 610 RLK family members is provided as supporting information.

Optimality criterion and PAUP program parameters. The aligned sequences were analyzed with PAUP (29) based on the Neighbor–Joining method (28), minimal evolution, and maximum parsimony criteria. To obtain the optimal trees, bootstrap analyses were conducted with 100 replicates using the heuristic search option. Two character-weighing schemes used were (*i*) all characters of equal weight and (*ii*) consider the number of nucleotide changes required to change from one amino acid to the other. All other parameters for PAUP were the default values. Because of the difficulty in aligning kinase subdomain X, two character sets were defined with or without kinase subdomain X sequences.

Tree rooting and display. Aminoglycoside kinase (APH(3′)III) from *Staphylococcus* (P00554) (30) and the *Arabidopsis* homolog of RIO1 family kinases (S61006) (31) were used as outgroups in this study. In all analyses, the rooting based on either sequence gave the same results. The numbers associated with each branch represent the bootstrap support, and branches with less than 50% support are collapsed.

Table 1. The proportion of EST records representing RLK/Pelle homologs in various organisms

Organism	Total EST*	RLK homologs	%EST
Porphyra yezoensis	10,185	0	0
C. elegans	109,095	0	0
D. melanogaster	95,211	3	0.003
Chlamydomonas reinhardtii	55,860	0	0
Marchantia polymorpha	1,307	1	0.077
Mosses	9,159	19	0.207
Ceratopteris richardii	2,838	7	0.247
Pinus taeda	21,797	100	0.459
Arabidopsis thaliana	112,467	620	0.551
Glycine max	122,843	704	0.573
Lotus japonicus	26,844	135	0.503
Lycopersicon esculentum	87,680	526	0.6
Oryza sativa	62,390	185	0.297
Triticum aestivum	44,132	178	0.403
Zea mays	73,965	135	0.183

*The searches were conducted based on EST available from GenBank as of Dec. 15, 2000.

Identification of Sequences Representing RLK Homologs. *Genomic sequences.* The kinase domain protein sequences of CLV1 and NAK were used to conduct BLAST searches against the genome sequences of *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, and human. The genomic sequence hits with an E value smaller than 1×10^{-10} were included for further analysis. Phylogenetic trees were constructed with the candidate sequences and the eukaryotic protein kinase representatives. Sequences that fell into the same clade as RLKs and had more than 50% bootstrap support were regarded as RLK homologs. The sequences shown in this analysis are *Caenorhabditis* Pelle-like sequence (CePelle, T23534), *Drosophila* Pelle (DmPelle, Q05652), and human IRAK1 (NP001560).

EST sequences. CLV1 and DmPelle kinase domain sequences were used to conduct BLAST searches against the EST records of organisms listed in Table 1. All EST sequences with E values smaller than 1.0 were retrieved for further analysis. The sequences with E values smaller than 1×10^{-50} were regarded as RLK homologs. The rest of the sequences longer than 300 nucleotides were submitted for batch BLASTX searches against *Arabidopsis* polypeptide records in GenBank. These sequences were regarded as RLK homologs if the top five matches of the BLAST outputs were RLK family kinases.

Results

The Diversity of RLKs in the *Arabidopsis* Genome. As the *Arabidopsis* genome sequencing effort approached completion, we conducted a genome-wide survey of the RLK gene family to gain more understanding of its size and complexity. The kinase domains of 22 different plant RLKs with various extracellular domains were used to search for similar sequences in GenBank polypeptide records of Viridiplantae, including all land plants

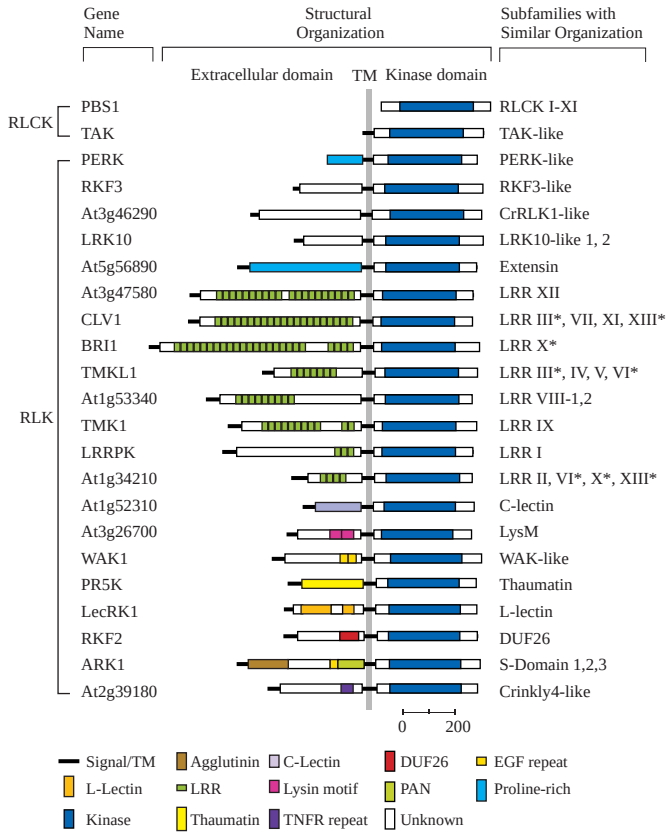

Fig. 1. Domain organization of representative RLKs and RLK-subfamily affiliations. Based on the presence or absence of extracellular domains, members of this gene family are categorized as RLKs or RLCKs. The gray line indicates the position of the membrane-spanning domain. The signal peptides are presumably absent in mature proteins but are displayed to demonstrate their presence in the RLKs. Locus names or MAtDB gene names are provided for the RLK representatives. Domain names are given according to SMART and Pfam databases (25, 26). Subfamilies are assigned based on kinase phylogeny (see supporting information for subfamily assignments for all members of the *Arabidopsis* RLK/Pelle family) and are shown according to the domain organization of the majority of members in a given subfamily. Subfamilies with >30% of members in more than one major extracellular domain category are designated with asterisk. DUF, domain of unknown function; EGF, epidermal growth factor; C lectin, C-type lectin; L-lectin, legume lectin; PAN, plasminogen/apple/nematode protein domain; TM, transmembrane region; TNFR, tumor necrosis factor receptor.

and algae. With the cutoff E value of 1×10^{-10}, more than 900 nonredundant sequences were retrieved. The most recent survey of the completed *Arabidopsis* genome revealed 620 sequences related to RLKs. Ten of these sequences showed greatest sequence similarity to the Raf kinase family. For the remaining 610 *Arabidopsis* sequences, 193 did not have an obvious receptor configuration as determined by the absence of putative signal sequences and/or transmembrane regions (see supporting information). The other 417 genes with receptor configurations can be classified into more than 21 structural classes by their extracellular domains with examples shown in Fig. 1. The sizes of these classes varied greatly. The leucine-rich repeat (LRR) containing RLKs represented the largest group in *Arabidopsis* with 216 genes.

To determine whether RLKs with similar extracellular domains also have similar kinase domains, the polypeptide sequences of the kinase domains of all 620 *Arabidopsis* genes were aligned, and a phylogenetic tree was generated with the Neighbor–Joining method (28) using APH(3′)III as outgroup (see supporting information for the complete alignment). APH(3′)III is a bacterial gene that is thought to be a distant relative of ePK (30). The phylogeny of *Arabidopsis* kinase domain sequences revealed an interesting pattern where the sequences clearly fell into distinct clades (see supporting information for the phylogenetic tree). We have tentatively assigned these natural groups into 44 different RLK subfamilies based on the kinase domain phylogeny (see supporting information for the subfamily assignment). A noteworthy feature of the pattern obtained is that the members within each of the RLK subfamilies tend to have similar extracellular domains, indicating that a single domain-shuffling event may have led to the founding of each of the various RLK subfamilies. For example, the diverse LRR-containing RLKs fell into more than 13 subfamilies based on kinase-domain phylogeny. With few exceptions, the pattern obtained is consistent with the grouping based on the structural arrangement of LRRs and the organization of introns in the extracellular domains of the individual RLKs (data not shown). Phylogenetic trees were also generated using minimum evolution and maximum parsimony criterions. The results were similar to phylogeny generated with the Neighbor–Joining method (data not shown).

The Relationship Between RLKs and Other Families of Protein Kinases from *Arabidopsis*. Despite the similar domain organization between different plant RLKs, the phylogenetic relationships among members of this family have not been thoroughly studied. Members of the RLK family could have arisen independently multiple times from distinct families of ePKs. Alternatively, they could have originated from a single ePK family and have a monophyletic origin. To address this question, we conducted a phylogenetic analysis by using the kinase domain amino acid sequences of representative RLK sequences from each RLK subfamily and representatives from different ePK families found in *Arabidopsis*.

In the phylogeny based on minimal evolution criterion, all RLK representative sequences from *Arabidopsis* formed a well supported clade, indicating that RLKs have a monophyletic origin within the superfamily of plant kinases (Fig. 2). In addition to RLK sequences, this monophyletic group also included kinases with no apparent signal sequence or transmembrane domain, and they were collectively named receptor-like cytoplasmic kinases (RLCKs, Fig. 1). Some of these kinases formed subfamilies distinct from other RLKs, whereas others were embedded within several different RLK subfamilies. To determine whether the monophyletic grouping of the RLK family represented a bias because of the exclusive use of *Arabidopsis* sequences, an extended analysis was conducted using RLK sequences from plants other than *Arabidopsis*. The sequences analyzed all fell into the same clade as *Arabidopsis* RLKs (data not shown).

Among the ePK families found in *Arabidopsis*, Raf kinases were paraphyletic to the RLK family and, together with RLKs, formed a well supported group with a bootstrap value of 98% (Fig. 2). Based on the parsimony criterion, the support for the RLK family and Raf kinases as a monophyletic group was still high at 86% (data not shown). Taken together, these results indicated that Raf kinases are the closest relatives to RLKs among the *Arabidopsis* sequences analyzed.

The Relationships Between Animal Receptor Kinases and Plant RLKs. Animal RTKs and receptor serine/threonine kinases (RSKs) are other families of ePKs with a domain organization similar to that of the plant RLKs. To determine the relationships among these receptor kinase families, we analyzed the phylogenetic relationships between the kinase domain sequences of representative *Arabidopsis* RLKs and animal receptor kinases. *Arabidopsis* and human representatives of other ePK families were also included.

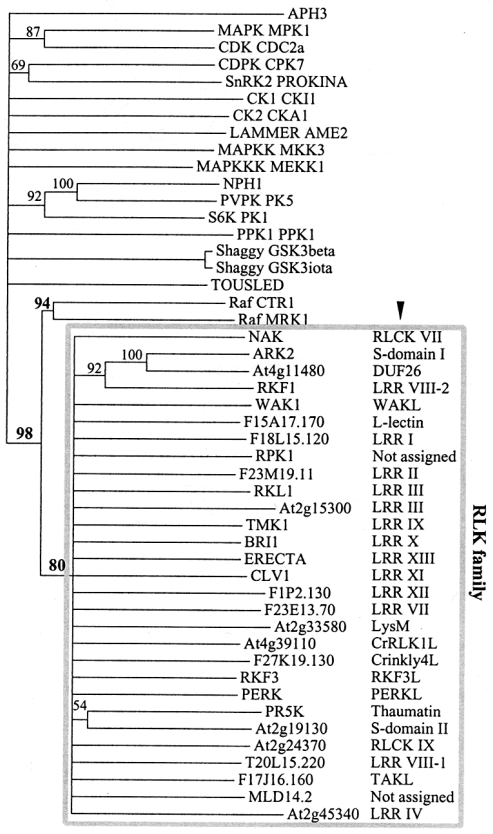

Fig. 2. *Arabidopsis* receptor-like kinases and related kinases form a monophyletic group distinct from all of the other eukaryotic protein kinases found in the *Arabidopsis* genome. The tree was generated with the kinase domain sequences of representative *Arabidopsis* ePKs and RLKs using APH(3′)III as outgroup based on minimal evolution. The bootstrap values are shown at the nodes. The boxed region represents the receptor-like kinase family. The arrowhead indicates the RLK subfamily. The abbreviations used are according to Fig. 1.

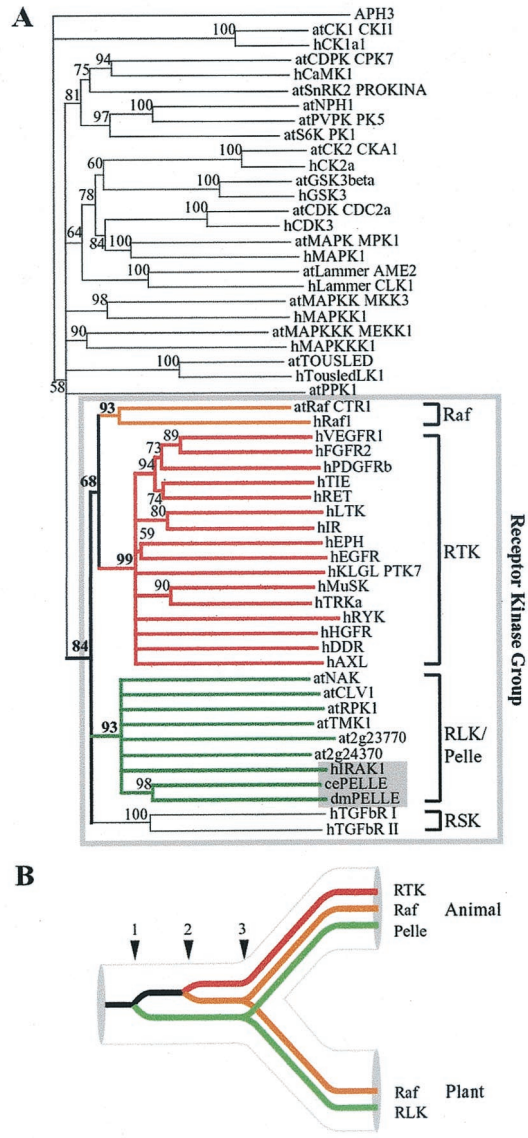

Fig. 3. Human receptor kinases and *Arabidopsis* receptor kinases belong to distinct but related families, and Pelle kinases are the animal homologs of *Arabidopsis* RLKs. (*A*) Plant and animal representatives of ePKs were used in this analysis. The tree is rooted with APH(3′)III based on minimal evolution. It indicates that Raf, RSK, RTK, and RLK form a well supported group distinct from all other ePKs (boxed region). The bootstrap values are shown at the nodes. Animal Pelle kinases (shaded area) are found in the same clade as RLKs. (*B*) The proposed evolutionary relationships between receptor kinase family members are as follows: 1, an ancient duplication event leading to the divergence of RLK/Pelle from RTK/Raf; 2, a more recent gene duplication leading to the divergence of RTK from Raf; and 3, the divergence of plant and animal lineages, resulting in the ancestral sequences that gave rise to the extant receptors and related kinases.

The phylogenetic tree generated based on minimal evolution criterion is shown in Fig. 3. All 16 RTK subfamily representatives and c-SRC formed a well supported group, indicating a monophyletic origin for tyrosine receptor kinases. The sister groups to the RTK family were Raf kinases. Plant RLKs included in this analysis formed another monophyletic group, indicating that RLKs have a distinct origin from that of Raf kinases and animal RTKs. Plant RLKs, Raf kinases, RSKs, and RTKs collectively formed a well supported group with a bootstrap support of 84%. The monophyly of kinases in this group when compared with the other ePK families was also supported by analyses based on maximum parsimony (data not shown). However, the specific relationships between animal RSKs, RTKs, Raf kinases, and plant RLKs were less conclusive because different optimality criteria gave inconsistent results. To investigate whether the results obtained were biased by using only human sequences, we conducted an extended analysis including RTK and RSK sequences from *Caenorhabditis*, *Drosophila*, sponge, and hydra, and we reached the same conclusion (data not shown). Based on these analyses, we defined the monophyletic group that contains the RLK, RTK, RSK, and Raf kinase genes as the receptor kinase group.

Homologs of Plant RLKs in Eukaryotes. To determine whether members of the RLK family are present in organisms other than flowering plants, we first used the kinase domain sequence of CLV1 to search for homologous sequences in the genomes of yeast, *C. elegans*, *D. melanogaster*, and human. No RLK homolog was found in the yeast genome. Five animal homologs of the RLK family were found: the Pelle kinase (DmPelle) in *Drosophila* (40), the Pelle-like kinase (CePelle) in *Caenorhabditis* (T23534), and three IRAKs in human (32). DmPelle, CePelle, and IRAK1 are all cytoplasmic kinases and all fell into the same clade as plant RLKs with strong bootstrap support (Fig. 3, shaded area). A similar search using other RLK kinase sequences yielded the same results (data not shown). Based on this

analysis, we defined the clade containing the plant RLKs and Pelle-like sequences as the RLK/Pelle family.

To broaden the scope of the searches, we used the amino acid sequences of CLV1 and Pelle kinase domains to search the EST database for RLK homologs in 20 different eukaryotes (15 shown in Table 1). Sequences with an E value of less than 1×10^{-50}, a conservative criterion, were regarded as RLK homologs without further examination. The remaining sequences were subjected to BLAST searches and were treated as RLK homologs if the top five matches were known RLKs or RLK homologs. The results of EST searches are shown in Table 1. All seven of the flowering plants, including four dicots and three monocots, have 0.18% to 0.6% of their ESTs representing RLK/Pelle family members. Pines, ferns, and mosses all have a percent EST representation similar to that of flowering plants. With the exception of the three ESTs representing *Drosophila* Pelle kinase, no other organism examined produced ESTs, which could be classified as RLK/Pelle family members.

The Distribution of RLKs on *Arabidopsis* Chromosomes. The size discrepancy of the RLK/Pelle family between plants and animals raises the question on how the expansion of this family occurred in the plant lineages. To address this question, we examined the location of RLKs on the *Arabidopsis* chromosomes. After comparing the location of genes to the phylogeny based on kinase domains, we found that subfamilies differed in their chromosomal distributions. At one extreme, 35 of the 40 members of the DUF26 subfamily were located on chromosome 4 (Fig. 4*A*). At the other extreme, 51 genes representing LRR X, XI, and XIII subfamilies were distributed among all five chromosomes (Fig. 4*B*). In addition, we found that more than 30% of the RLK/Pelle family members in *Arabidopsis* are in tandem repeats with 2 to 19 genes. A closer look at the location of the 38 DUF26 subfamily members on chromosome 4 (including three potential pseudogenes) indicates that 34 of them are in tandem repeats (Fig. 4*C*). The phylogenetic relationships between DUF26 genes in the tandem repeats indicates that at least one intrachromosomal duplication event occurred in the region containing tandem repeats. Taken together, the results suggest that tandem duplication events and large-scale duplications of chromosomes are two of the potential mechanisms responsible for the expansion of the RLK/Pelle family in *Arabidopsis.*

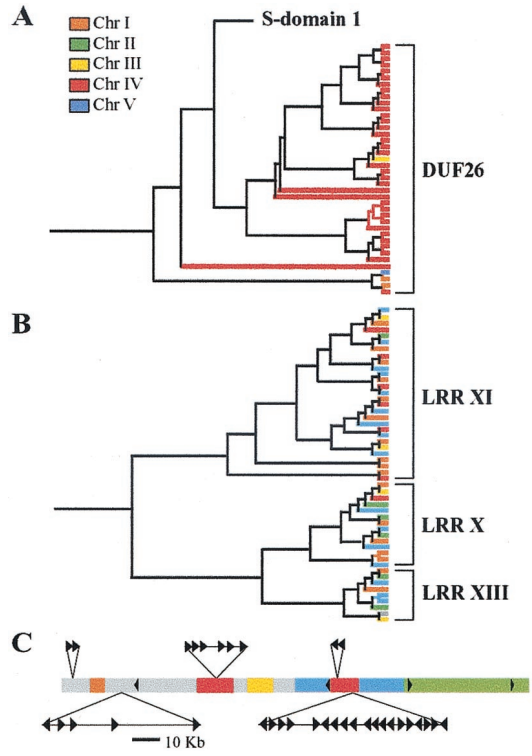

Fig. 4. Distribution of RLKs on *Arabidopsis* chromosomes provides clues for the mechanisms of RLK family expansion. (*A*) The cladogram of the DUF26 subfamily was generated with the kinase domain sequences based on minimal evolution criterion. The color coding on branches indicates the chromosome on which each gene in the subfamily is located. Note that most DUF26 members are located on chromosome 4. (*B*) The cladogram of LRR X, XI, and XIII subfamilies was generated and color-coded in the same manner as *A*. Note that most genes derived from duplication events are located on different chromosomes. (*C*) A detailed depiction of DUF26 distribution on chromosome 4 indicates that tandem duplications and an internal chromosomal duplication may contribute extensively to the expansion of this subfamily. The 10-kb legend is for the expanded region showing tandem repeats. The regions with postulated chromosomal duplications are color-coded according to their similarity to regions on the other chromosomes. The color-coding scheme is the same as *A*. Three potential DUF26 pseudogenes are also included in the diagram.

Discussion

Evolutionary History of the Receptor Kinase Group. Plant RLKs were originally grouped into a single family based on their configuration as transmembrane kinases with serine and threonine specificity. Our analysis provides a phylogenetic basis for the classification of RLKs as a single family in the eukaryotic protein kinase superfamily. Interestingly, 24% of the 610 *Arabidopsis* genes in the RLK/Pelle family analyzed do not have an extracellular domain based on the absence of signal sequences and transmembrane regions. Some of these apparently cytoplasmic kinases form unique subfamilies, whereas others are most closely related to kinases with a receptor topology. The latter may represent ancestral forms that were recruited into the receptor kinase configuration by domain fusion events. Alternatively, some of the soluble kinase forms could be derived from ancestral receptor kinase forms. In any case, it is apparent that kinase domains from the RLK/Pelle family were recruited multiple times by fusion with different extracellular domains to form the subfamilies found in *Arabidopsis*. This notion can be expanded to include the animal RTK and RSK families in the receptor kinase group, which appear to have been formed by recruitment of kinases from the same lineage, distinct from all other ePK families.

Based on the kinase domain phylogeny, a hypothetical sequence of events that occurred in the evolution of the receptor kinase group is proposed in Fig. 3*B*. According to this model, an early gene duplication event led to the founding of two lineages that diversified into the RTK and Raf families on one hand and the RLK/Pelle family on the other. This diversification seems to have occurred before the divergence of plants and animals. In addition, both lineages contain representatives of soluble kinase and transmembrane receptor forms. It should be noted that the soluble Pelle-like and Raf kinases form complexes with cell surface receptors and are responsible for transduction of signals to downstream effectors (33, 34). Perhaps the continual recruitment of this particular lineage of kinase modules was favored during evolution because ancestral forms had already specialized in mediating signaling from transmembrane receptors. Examination of kinases belonging to the receptor kinase group in more primitive eukaryotes may be informative. Whereas fungi such as yeast and *Neurospora* do not appear to have representatives of the receptor kinase group, the slime mold, *Dictyostelium discoideum*, has several examples (data not shown). None of these sequences from slime mold has predicted signal peptide or transmembrane regions, and most of the sequences are dual specificity kinases based on their kinase activities (35), consistent with the possibility that the ancestral form for extant receptor kinases may have been soluble kinases.

Diversification of the Plant RLK/Pelle Family. The small number of representatives of the RLK/Pelle family in animals compared with the much larger number in *Arabidopsis* indicates that the expansion of the plant RLKs occurred after the divergence of plant and animal lineages or that massive gene loss occurred in the animals. A comparison of EST representation with the known total number of RLK/Pelle members in the fully sequenced genomes of *C. elegans*, *D. melanogaster*, and *Arabidopsis* indicated that the EST representation provided a conservative estimate of the total number of family members in the genomes. The lack of RLK/Pelle ESTs in *Porphyra* and *Chlamydomonas* argues that, rather than massive gene loss in the animal genomes examined, the RLK/Pelle family likely underwent expansion after the divergence of animal and plant lineages. Interestingly, all land plants have similar percent representations of RLK/Pelle kinases, suggesting that the size of this gene family may have been similar to the present-day level before the diversification of the land plant lineages. Additional sequence information will be necessary to determine whether all RLK subfamilies found in *Arabidopsis* are equally represented in these other land plant lineages. The early expansion of the RLK/Pelle family could be associated with evolution of multicellularity, as has been suggested for the RTK family in animals (36). Alternatively, the expansion of the family could be associated with the development of the complex array of attributes required for the migration of plant lineages from the aquatic to the terrestrial environment. Examination of RLK/Pelle representation in multicellular green algae such as *Chara* could help to resolve this question.

The monophyletic origin of the RLK/Pelle family implies that the expansion of the family to its present size in *Arabidopsis* was the result of multiple gene-duplication events. Two possible mechanisms for the amplification of this family are suggested by the way members of some subfamilies are distributed on the *Arabidopsis* chromosomes. For example, the DUF26 subfamily is organized in tandem arrays (Fig. 4*C*). These tandem arrays were likely generated by gene duplications resulting from unequal crossing-over as seen in the other gene families such as disease resistance genes (37). Gene duplication is also driven by larger scale duplication events, including polyploidization followed by reshuffling of chromosomal regions (5, 38, 39). Tandem arrays of DUF26 members are located in such duplicated regions on chromosome 4. However, the localization of other DUF26 subfamily members almost exclusively on chromosome 4 suggests that this subfamily expanded after the extensive chromosome duplications and reshuffling identified for multiple regions of all five *Arabidopsis* chromosomes (5, 38, 39).

On the other hand, members of LRR X, XI, and XIII subfamilies are distributed among all five chromosomes, with related genes on each branch of the phylogenetic tree generally located on different chromosomes. These three related subfamilies are of particular interest because they include *Arabidopsis* RLKs with known developmental functions such as BRI1, CLV1, ERECTA, and HAESA (6, 7, 10). The difference in distribution patterns between the DUF26 and these LRR subfamilies could indicate that the LRR subfamilies originally expanded by mechanisms that did not include localized (e.g., tandem) duplications. Given sufficient evolutionary time, several rounds of polyploidization followed by chromosomal rearrangements could produce a given subfamily of the observed size from a single prototypical gene. Alternatively, these LRR subfamilies may have originally expanded via localized duplications that occurred early enough in evolutionary time that extensive chromosome reshuffling could have eliminated linkage between subfamily members. Both proposed mechanisms imply that the LRR X, XI, and XIII subfamilies may have expanded much earlier in time than the DUF26 subfamily. A comparative analysis of RLK subfamilies in other plant lineages should help to resolve this issue.

We thank Micahel Gribskov for advice on sequence retrieval and alignments and Donna Fernandez, Frans Tax, Sara E. Patterson, and Melissa D. Lehti-Shiu for reading the manuscript. This work was supported by Department of Energy Grant DE-FG02-91ER20029 (to A.B.B.) and National Research Initiative Competitive Grants Program/U.S. Department of Agriculture Grant 2000-21469 (to S.-H.S.).

1. van der Geer, P., Hunter, T. & Lundberg, R. A. (1994) *Annu. Rev. Cell Biol.* **10,** 251–337.
2. Robertson, S. C., Tynan, J. A. & Donoghue, D. J. (2000) *Trends Genet.* **16,** 265–271.
3. Walker, J. C. (1994) *Plant Mol. Biol.* **26,** 1599–1609.
4. McCarty, D. R. & Chory, J. (2000) *Cell* **103,** 201–209.
5. Arabidopsis Genome Initiative. (2000) *Nature (London)* **408,** 796–815.
6. Li, J. & Chory, J. (1997) *Cell* **90,** 929–938.
7. Clark, S. E., Williams, R. W. & Meyerowitz, E. M. (1997) *Cell* **89,** 575–585.
8. Gomez-Gomez, L. & Boller, T. (2000) *Mol. Cell* **5,** 1003–1011.
9. Becraft, P. W., Stinard, P. S. & McCarty, D. R. (1996) *Science* **273,** 1406–1409.
10. Jinn, T. L., Stone, J. M. & Walker, J. C. (2000) *Genes Dev.* **14,** 108–117.
11. Stein, J. C., Howlett, B., Boyes, D. C., Nasrallah, M. E. & Nasrallah, J. B. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 8816–8820.
12. Song, W. Y., Wang, G. L., Chen, L. L., Kim, H. S., Pi, L. Y., Holsten, T., Gadner, J., Wang, B., Zhai, W. X., Zhu, L. H., *et al.* (1995) *Science* **270,** 1804–1806.
13. Schopfer, C., Nasrallah, M. & Nasrallah, J. (1999) *Science* **286,** 1697–1700.
14. Takayama, S., Shiba, H., Iwano, M., Shimosato, H., Che, F.-S., Kai, N., Watanabe, M., Suzuki, G., Hinata, K. & Isogai, A. (1999) *Proc. Natl. Acad. Sci. USA* **97,** 1920–1925.
15. Brand, U., Fletcher, J. C., Hobe, M., Meyerowitz, E. M. & Simon, R. (2000) *Science* **289,** 617–619.
16. Trotochaud, A. E., Jeong, S. & Clark, S. E. (2000) *Science* **289,** 613–617.
17. Wang, Z.-Y., Seto, H., Fujioka, S., Yoshida, S. & Chory, J. (2001) *Nature (London)* **410,** 380–383.
18. Gomez-Gomez, L., Bauer, Z. & Boller, T. (2001) *Plant Cell* **13,** 1155–1163.
19. Stone, J. M., Collinge, M. A., Smith, R. D., Horn, M. A. & Walker, J. C. (1994) *Science* **266,** 793–795.
20. Bower, M. S., Matias, D. D., Fernandes-Carvalho, E., Mazzurco, M., Gu, T., Rothstein, S. J. & Goring, D. R. (1996) *Plant Cell* **8,** 1647–1650.
21. Trotochaud, A. E., Hao, T., Wu, G., Yang, Z. & Clark, S. E. (1999) *Plant Cell* **11,** 393–406.
22. Hanks, S. K. & Hunter, T. (1995) *FASEB J.* **9,** 576–596.
23. Hardie, D. G. (1999) *Annu. Rev. Plant Physiol. Plant Mol. Biol.* **50,** 97–131.
24. Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) *Nucleic Acids Res.* **25,** 3389–3402.
25. Schultz, J., Copley, R. R., Doerks, T., Ponting, C. P. & Bork, P. (2000) *Nucleic Acids Res.* **28,** 231–234.
26. Sonnhammer, E. L. L., Eddy, S. R., Birney, E., Bateman, A. & Durbin, R. (1998) *Nucleic Acids Res.* **26,** 320–322.
27. Higgins, D. G., Thompson, J. D. & Gibson, T. J. (1996) *Methods Enzymol.* **266,** 383–402.
28. Saitou, N. & Nei, M. (1987) *Mol. Biol. Evol.* **4,** 406–425.
29. Swafford, D. L. (1998) PAUP*: Phylogenetic Analysis Using Parsimony (*and other materials) (Sinauer, Sunderland, MA).
30. Hon, W.-C., McKay, G. A., Thompson, P. R., Sweet, R. M., Yang, D. S. C., Wright, G. D. & Berghuis, A. M. (1997) *Cell* **89,** 887–895.
31. Leonard, C. J., Aravind, L. & Koonin, E. V. (1998) *Genome Res.* **8,** 1038–1047.
32. Cao, Z., Henzel, W. J. & Gao, X. (1996) *Science* **271,** 1128–1131.
33. O'Neil, L. A. & Greene, C. (1998) *J. Leukocyte Biol.* **63,** 650–657.
34. Daum, G., Eisenmann-Tappe, I., Fries, H.-W., Troppmair, J. & Rapp, U. R. (1994) *Trends Biochem. Sci.* **19,** 474–480.
35. Nuckolls, G. H., Osherov, N., Loomis, W. F. & Spudich, J. A. (1996) *Development (Cambridge, U.K.)* **122,** 3295–3305.
36. Hunter, T. & Plowman, G. D. (1997) *Trends Biochem. Sci.* **22,** 18–22.
37. Ronald, P. C. (1998) *Curr. Opin. Plant Biol.* **1,** 294–298.
38. Blanc, G., Barakat, A., Guyot, R., Cooke, R. & Delseny, M. (2000) *Plant Cell* **12,** 1093–1101.
39. Vision, T. J., Brown, D. G. & Tanksley, S. D. (2000) *Science* **290,** 2114–2117.
40. Shelton, C. A. & Wasserman, S. A. (1993) *Cell* **72,** 515–525.