Patterns of Diversifying Selection in the Phytotoxin-like *scr74* Gene Family of *Phytophthora infestans*

Zhenyu Liu, Jorunn I. B. Bos,* Miles Armstrong,† Stephen C. Whisson,† Luis da Cunha,* Trudy Torto-Alalibo,* Joe Win,* Anna O. Avrova,† Frank Wright,† Paul R. J. Birch,† and Sophien Kamoun**

*Department of Plant Pathology, The Ohio State University, Ohio Agricultural Research and Development Center, Wooster, Ohio; †Plant-Pathogen Interaction Programme, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, United Kingdom

Phytophthora infestans, the organism responsible for the Irish famine, causes late blight, a re-emerging disease of potato and tomato. Little is known about the molecular evolution of *P. infestans* genes. To identify candidate effector genes (virulence or avirulence genes) that may have co-evolved with the host, we mined expressed sequence tag (EST) data from infection stages of *P. infestans* for secreted and potentially polymorphic genes. This led to the identification of *scr74*, a gene that encodes a predicted 74-amino acid secreted cysteine-rich protein with similarity to the *Phytophthora cactorum* phytotoxin PcF. The expression of *scr74* was upregulated approximately 60-fold 2 to 4 days after inoculation of tomato and was also significantly induced during early stages of colonization of potato. The *scr74* gene was found to belong to a highly polymorphic gene family within *P. infestans* with 21 different sequences identified. Using the approximate and maximum likelihood (ML) methods, we found that diversifying selection likely caused the extensive polymorphism observed within the *scr74* gene family. Pairwise comparisons of 17 *scr74* sequences revealed elevated ratios of nonsynonymous to synonymous nucleotide-substitution rates, particularly in the mature region of the proteins. Using ML, all 21 polymorphic amino acid sites were identified to be under diversifying selection. Of these 21 amino acids, 19 are located in the mature protein region, suggesting that selection may have acted on the functional portions of the proteins. Further investigation of gene copy number and organization revealed that the *scr74* gene family comprises at least three copies located in a region of no more than 300 kb of the *P. infestans* genome. We found evidence that recombination contributed to sequence divergence within at least one gene locus. These results led us to propose an evolutionary model that involves gene duplication and recombination, followed by functional divergence of *scr74* genes. This study provides support for using diversifying selection as a criterion for identifying candidate effector genes from sequence databases.

Introduction

Oomycetes, also known as water molds, form a diverse group of fungus-like eukaryotic microorganisms that are distantly related to fungi but are more closely related to diatoms and brown algae in the Stramenopiles (or heterokonts), one of several major eukaryotic kingdoms (Sogin and Silberman 1998; Baldauf et al. 2000; Margulis and Schwartz 2000). *Phytophthora infestans*, the most notorious and destructive oomycete, was responsible for the Irish potato famine in the nineteenth century. This species remains a devastating pathogen, causing late blight, a reemerging disease of potato and tomato (Fry and Goodwin 1997*a*, 1997*b*; Birch and Whisson 2001; Schiermeier 2001; Smart and Fry 2001; Ristaino 2002; Shattock 2002; Nicholls 2004). It is estimated that the late blight disease causes multibillion-dollar losses in potato and tomato production worldwide (Fry and Goodwin 1997*a*, 1997*b*). Although *P. infestans* is a pathogen of great economic importance, the molecular mechanisms of pathogenicity and host specificity are not well understood. Furthermore, mechanisms of molecular evolution of *P. infestans* genes and gene families are largely unknown.

The neutral theory of molecular evolution maintains that most molecular polymorphisms within a species and most molecular divergence between species are driven by random fixation of selectively neutral mutations (Kimura 1983). By investigating the prevalence of nucleotide polymorphism and divergence, it is possible to obtain considerable insight into the evolutionary processes that shaped a particular genomic region (Hudson 1993). In the last decade, the genetic architecture of polymorphisms within a species and species divergence have been widely studied (Hughes, Ota, and Nei 1990; Karl and Avise 1992; Orr and Coyne 1992; Berry and Kreitman 1993; Haag and True 2001; Wu 2001). Many researchers have focused on identifying genes and finding genomic regions of functional importance on which selection has acted, thus helping to unravel the evolutionary genetic basis of ecological diversification. For example, diversifying selection (also known as positive selection) can be an indicator of genomic regions that contain genes or gene families of functional importance.

The most reliable indicator of diversifying selection at the molecular level is a higher nonsynonymous nucleotide-substitution rate (d_N) than synonymous nucleotide-substitution rate (d_S) between two protein-coding DNA sequences (ratio $\omega = d_N / d_S > 1$) (Li, Wu, and Luo 1985; Nei and Gojobori 1986; Ina 1995; Yang and Bielawski 2000). Based on this criterion, statistical methods, such as the approximate method (also known as the counting method) and the maximum likelihood (ML) method, have been developed and implemented into computer software packages for detecting diversifying selection (Yang and Bielawski 2000). On the basis of such methods, a number of genes involved in defense systems or immunity, genes involved in evading defense systems or immunity, and toxin protein genes have been shown to be under diversifying selection (Stahl and Bishop 2000; Yang and Bielawski 2000).

Key words: diversifying selection, *Phytophthora infestans*, virulence, avirulence, cysteine-rich, host-microbe interactions.

E-mail: kamoun.1@osu.edu.

Mol. Biol. Evol. 22(3):659–672. 2004
doi:10.1093/molbev/msi049
Advance Access publication November 17, 2004

In plant pathogen interactions, resistance is often regulated by recognition of pathogen molecules by the plant. This is illustrated by the gene-for-gene concept, which implies that an avirulence (*Avr*) gene from the pathogen is recognized directly or indirectly by a matching resistance (*R*) gene from the plant, resulting in recognition of the pathogen and activation of plant defense mechanisms (Dangl and Jones 2001). Diversifying selection in genes encoding proteins that function at the interface of attack and defense in host-pathogen antagonism, such as the *Avr* and *R* genes, is likely to reflect an "arms race" co-evolution (Thomas and Stephen 1999, 2000; Stahl and Bishop 2000). The rationale is that natural selection driven by a co-evolutionary arms race is likely to leave a signature at the molecular level. Thus, evolutionary analyses of defense or attack (virulence) genes can provide insight into how plants and pathogens co-evolve under the "arms race" model, and the extent to which co-evolutionary interactions shape the present genetic variation in plant and pathogen populations (Stahl and Bishop 2000).

P. infestans research has entered the genomics era. Current genomic resources include expressed sequence tags (ESTs) from a variety of developmental and infection stages, as well as sequences of selected regions of the genome (Kamoun 2003). A number of data-mining and functional strategies have been developed to exploit the sequence resources. For example, Torto et al. (2003) developed an algorithm to identify putative extracellular effector proteins from EST data sets. Bos et al. (2003) described a strategy to identify candidate *Avr* genes based on the assumption that these genes exhibit significant sequence variation within populations of the pathogen. Accumulation of structural genomic resources, genome sequences for *P. infestans*, and the availability of appropriate statistical methodologies provide the opportunity to investigate patterns of diversifying selection in effector proteins from *P. infestans*. Effector proteins are molecules produced by plant pathogens to manipulate biochemical and physiological processes in their host plants by promoting infection (virulence genes) or by triggering defense responses (*Avr* genes) (Torto et al. 2003). Based on the assumption that evidence of diversifying selection in effector genes could reflect an "arms race" co-evolution between the host and the pathogen, we hypothesized that identifying *P. infestans* genes under diversifying selection will augment other criteria to help us select candidate effector genes important in virulence and host specificity.

In this study, we mined EST data from infection stages of *P. infestans* for secreted and potentially polymorphic genes. One class of genes, identified by Torto et al. (2003), encodes secreted small cysteine-rich (SCR) proteins, a feature reminiscent of the products of *Avr* genes from plant pathogenic fungi and oomycetes (van't Slot and Knogge 2002; Bittner-Eddy et al. 2003). One of these genes, *scr74*, encodes a predicted protein with significant similarity to PcF, a 52 amino acid phytotoxic necrosis-inducing protein secreted by *Phytophthora cactorum* (Orsomando et al. 2001). Further characterization of the *scr74* gene suggested that it is upregulated during colonization of tomato and potato by *P. infestans* and forms a highly polymorphic gene family. We investigated the molecular evolution of the *scr74* genes by means of the approximate method of Nei and Gojobori (1986), which calculates the average ω ratio across all the amino acid sites. In addition, we used the ML method to identify particular amino acid residues on which diversifying selection has acted (Nielsen and Yang 1998; Yang and Bielawski 2000). Results showed that diversifying selection likely caused the extensive polymorphism observed within the *scr74* gene family. Based on this and additional analyses of gene copy number and organization, we propose an evolutionary model that involves duplication followed by functional divergence of *scr74* genes. This study provides support for using diversifying selection as a criterion for identifying candidate effector genes from sequence databases.

Materials and Methods

Phytophthora infestans Strains and Culture Conditions

P. infestans isolate 90128 (A2 mating type, race 1.3.4.7.8.9.10.11) and 88069 (A1 mating type, race 1.3.4.7) were routinely cultured at 18°C on rye agar medium supplemented with 2% sucrose (Caten and Jinks 1968). For RNA extraction, plugs of mycelium were transferred to modified Plich medium (Kamoun et al. 1993) and grown for 2 to 3 weeks before harvesting. Non-sporulating mycelium and germinated cysts were obtained, and potato plant inoculations were carried out as described by Avrova et al. (2003). Other isolates that were used include 19 different U.S. isolates of *P. infestans* representing 10 clonal lineages (US930468 [US1 clonal lineage]; US930258 [US1]; US940501 [US1]; US920165 [US1]; US940507 [US1]; US940330 [US7]; US960022 [US7]; US990004 [US6]; US940480 [US8], US940504 [US13]; US940502 [US14]; US940289 [US16]; US970045 [US17]; US970001 [US17]; US970015 [US17]; US990025 [US11]; US980008 [US11]; US980066 [US11]; US940494 [US12]) (Goodwin et al. 1998), strain IPO-0 (US1, race 0) (Vleeshouwers et al. 2000), and strain T30-4, which was used for construction of the bacterial artificial chromosome (BAC) library (Whisson et al. 2001).

RNA Manipulations, Northern Blot Analysis, and Real-Time Reverse Transcriptase Polymerase Chain Reaction (RT-PCR)

Total RNA isolation from *P. infestans* mycelium and from infected tomato leaves, as well as northern blot hybridizations were carried out as described by Huitema et al. (2003). Total RNA extraction and cDNA synthesis for *P. infestans* mycelium, sporangia, zoospores, germinating cysts and uninfected and infected potato cultivar Bintje leaves, and SYBR green real-time RT-PCR assays were carried out as described by Avrova et al. (2003). Real-time RT-PCR primers for the constitutively expressed *P. infestans* control gene *actA*, and for the *in planta*–induced gene *calA*, are given in Avrova et al. (2003). Primers for *scr74* were 5′-CCACGATTGCTGTGGTAAAAGTT-3′ and 5′-TCGCTGTGGTTTGGAATCTAGA-3′, and amplified a 72-bp fragment.

Table 1
Distribution of *scr74* sequences among *Phytophthora infestans* isolates

P. infestans isolate (clonal lineage)	SCR74 Allele																			
	A10	A11	B3a	B3b	B7	B10	C3a	C3b	C4	C9	C10	D1	D2	D4	D5	D6	E5	E6	E11	F12	G1
US940507 (US1)	x	x													x						
US960022 (US7)			x		x	x															
US990004 (US6)				x		x	x		x	x	x										
US940480 (US8)						x		x									x	x	x	x	
US970015 (US17)			x									x	x	x	x	x					
IPO-0 (US1)																					x
88069 (NA)															x						
90128 (NA)						x															

DNA Manipulations and Southern Blot Analysis

Total DNA samples from the 19 *P. infestans* isolates listed above were kindly provided by Dr. C. Smart (Cornell University) and were used for Southern blot analysis. DNA samples (15 μg each) were digested with *Hind* III restriction enzyme, and separated by gel electrophoresis on a 1% agarose gel in TBE buffer and transferred to Hybond N^+ membranes (Amersham Biosciences Corp, Piscataway, N.J.) according to the manufacturer's instructions. Southern blot hybridizations were conducted at 65°C in Modified Church Buffer (0.36 M Na_2HPO_4, 0.14 $MNaH_2PO_4$, 1 mM ethylene diamine tetraacetic acid [EDTA], and 7% sodium dodecyl sulfate [SDS]). Filters were washed at 55°C in 1 × SSC/0.5% SDS, and 0.5 × SSC/0.5% SDS (Sambrook and Russell 2001). To reveal the hybridizing bands, membranes were exposed to a phosphor imager screen (Molecular Dynamics Storm 840 Phosphor Imager). Hybridizations to the *P. infestans* BAC library, BAC DNA isolation, BAC end-sequencing, and Southern blotting of BAC clones were performed as described by Whisson et al. (2001).

Hybridization Probes for Southern and Northern Blot Analysis

DNA inserts from cDNA clones of *scr74* and *actA* were gel-purified after digestion and used as probes for Southern and northern blot hybridizations. The probes were radiolabeled with α^{32}P-dATP using a random primer labeling kit (Invitrogen, San Diego, Calif.).

Primer Design, PCR Amplification, and Sequencing

A pair of oligonucleotide primers SCR74-FCla (5′-GGAAATCGATCCGGTCATCGTCACTACTCAAC-AGCTCG–3′) and SCR74-RNot (5′- GGAAGCGGCCG-CTTCATTCATTTGATTATCACTGTATCTC–3′) were designed for the amplification of a 304-bp fragment containing the entire open reading frame (ORF) of *scr74*. The fragments were cloned in pGR106 (Lu et al. 2004) using the *Cla*I and *Not*I restriction enzymes. Five *P. infestans* isolates were used for PCR amplifications (table 1). Polymerase chain reaction amplifications and DNA sequencing were performed as described earlier (Bos et al. 2003). The sequences described here were deposited in GenBank (accession numbers AY723699–AY723725).

Sequence Analysis

PexFinder and signal peptide predictions were carried out following the methods of Torto et al. (2003). Similarity searches were performed locally on Intel Linux and Mac OSX workstations. Search programs included BLAST (Altschul et al. 1997), and the similarity search programs implemented in the BLOCKS (Henikoff et al. 2000), Pfam (Bateman et al. 2002), SMART (Letunic et al. 2002), and InterPro (Apweiler et al. 2001) Web sites. Sequence data analysis and interpretation were performed as described by Bos et al. (2003). Base calling was performed with the algorithm phred (Ewing and Green 1998; Ewing et al. 1998). Only sequences with phred Q values higher than 20 were retained for analysis. Sequences were aligned, and ambiguous calls were checked against chromatograms using Sequencher 4.1 (Gene Codes Corp.). Two observations suggest that the sequences we generated are of high quality. First, sequences identical to those of ESTs PC015G09 and PH011A12 were recovered from five independent PCR clones. Second, the observed polymorphisms were unevenly distributed and were essentially localized in the ORF.

Databases

We examined several sequence databases, including publicly available GenBank nonredundant databases and dBEST (Karsch-Mizrachi and Ouellette 2001), the *Phytophthora* Functional Genomics Database (PFGD, www.pfgd.org), and the Syngenta *Phytophthora* Consortium (SPC) database, a proprietary database of Syngenta Inc. containing ca. 75,000 ESTs from *P. infestans*.

Diversifying Selection Analyses

The rate of nonsynonymous nucleotide substitutions per nonsynonymous site (d_N) and the rate of synonymous nucleotide substitutions per synonymous site (d_S) across all the amino acids sites in pairwise comparisons between nucleotide sequences were estimated using the approximate method of Nei and Gojobori (1986) implemented in the YN00 program in the PAML software package (Yang 1997).

To identify which SCR74 amino acids have been affected by diversifying selection, we used maximum likelihood models of codon substitution that allow for heterogeneous selection pressures among sites along the protein (Nielsen and Yang 1998; Yang et al. 2000; Yang and Bielawski 2000). Analyses were done with the computer

program CODEML in the PAML package (Yang 1997). This method consists of two major steps. The first step uses the likelihood ratio test (LRT) to test for diversifying selection by comparing a null model with an alternative one that accounts for sites under diversifying selection. The six models recommended by Yang et al. (2000) were tested. They were null models, M0, M1, M7, corresponding to alternative models, M3, M2, M8, respectively. Twice the difference in log likelihood ratio between a null model and an alternative model was compared with a chi-squared (χ^2) distribution with degrees of freedom equaling the difference in the numbers of parameters estimated from the pair of models. The likelihood ratios of the two models test whether an alternative model fits the data better than the null model. The second step identifies amino acids under diversifying selection by using the empirical Bayes theorem, as implemented in CODEML, to calculate the posterior probability that a particular amino acid belongs to a given selection class (neutral, deleterious or advantageous) (Yang 1997). Amino acid sites with a high posterior probability for an advantageous class of sites ($\omega > 1$) were deemed more likely to be under diversifying selection.

Analysis of Genetic Recombination

Open reading frames of *scr74* nucleotide sequence were translated into amino acid sequences, and multiple alignments were conducted using the CLUSTAL-X program (Thompson et al. 1997). Evidence for genetic recombination was sought with the SplitsTree program that uses the Split Decomposition method (Huson 1998). The difference in sum of squares (DSS) statistics (McGuire and Wright 2000) implemented in the TOPALi program (Milne et al. 2004) was used to further investigate recombination. Phylogenetic trees were constructed with the Neighbor-Joining method based on Jukes–Cantor distances as implemented in TOPALi (Milne et al. 2004). Parametric bootstrapping using the DSS statistic was used to compare tree topologies (Goldman, Anderson, and Rodrigo 2000).

Results

A *P. infestans* cDNA Isolated from Infected Tomato Leaves Encodes a Secreted Small Cysteine-Rich Protein with Similarity to a Phytotoxin

We analyzed an EST data set generated from tomato leaves 3 days after infection with *P. infestans* using BLASTN searches against a 0.7X whole genome shotgun sequence of *P. infestans* and all publicly available tomato sequences. Among the ESTs examined, 241 of 2808 showed more than 90% identity to *P. infestans* sequences but less than 90% identity to tomato sequences and were hypothesized to originate from the pathogen. These ESTs were then annotated by similarity and motif searches against public databases and using PexFinder to identify cDNAs encoding extracellular proteins (Torto et al. 2003). One EST, PC015G09, showed significant similarity to the necrosis-inducing protein PcF secreted by *P. cactorum* (Orsomando et al. 2001) and the candidate *Avr* gene *scr91* (Bos et al. 2003). Full-length sequencing of the corresponding cDNA revealed an ORF of 222 bp, corresponding to a predicted translated product of 74 amino acids containing 8 cysteines. SignalP (Nielsen et al. 1997; Nielsen and Krogh 1998) analysis of the predicted protein identified a 21-amino acid signal peptide with a significant mean S value of 0.849. Polymerase chain reaction amplifications with primers based on the cDNA sequence were successful with *P. infestans* genomic DNA but not with tomato. Based on these analyses, we propose that the analyzed cDNA is from *P. infestans* and encodes a putative 74-amino acid secreted and cysteine-rich protein, with features reminiscent of several *Avr* genes from plant pathogenic fungi. We designated the cDNA small cysteine-rich protein 74 or *scr74*.

The *scr74* Gene Is Upregulated During Infection of Tomato and Potato by *P. infestans*

To determine whether the *scr74* gene is upregulated during infection, we used northern blot analysis to detect *scr74* mRNA in a time course of *P. infestans* infection of its host plant tomato. Total RNA was isolated from leaves of tomato 0, 1, 2, 3, and 4 days post-inoculation (dpi) with *P. infestans* isolate 90128 and from *P. infestans* mycelium grown in liquid rye-sucrose medium. A northern blot containing these samples was hybridized with probes made from *P. infestans scr74* and the constitutive gene *actA* (Unkles et al. 1991). During the interaction, *scr74* transcripts were first detected at two dpi and reached maximal levels at 3 and 4 dpi (fig. 1*A*). In contrast to *actA, scr74* resulted in significantly stronger hybridization signals in the time course compared to mycelium, suggesting that the *scr74* gene is upregulated during infection of tomato by *P. infestans* (fig. 1*A*). Quantification of the hybridization signals indicated that *scr74* is upregulated approximately 60-fold at 2 to 4 dpi compared to mycelium.

Real-time RT-PCR analysis was used to investigate the expression of *scr74* using cDNA templates derived from *P. infestans* mycelium grown in pea broth, sporangia, zoospores, germinating cysts, uninfected potato cv. Bintje, and infected Bintje (12, 24, 33, 48, 56, and 72 hpi). The *actA* gene from *P. infestans* was used as a constitutively expressed endogenous control, and expression was also compared to *calA*, a gene known to be upregulated in germinating cysts and infected plants (Avrova et al. 2003). Expression of *scr74* in different samples was compared to the level of its expression in a calibrator sample, which was cDNA from mycelium and was assigned the value of 1.0. As expected, the *calA* gene was upregulated approximately fivefold in germinating cysts and 18-fold at 48 hpi (data not shown; Avrova et al. 2003). The *scr74* gene was upregulated 380-fold in germinating cysts and showed elevated levels of expression throughout infection (fig. 1*B*). Repeated amplifications, on independent occasions with different cDNA samples, resulted in similar expression profiles.

scr74 Belongs To A Polymorphic Gene Family in *P. infestans*

We used the *scr74* sequence to search the SPC database, containing ca. 75,000 ESTs from *P. infestans*.

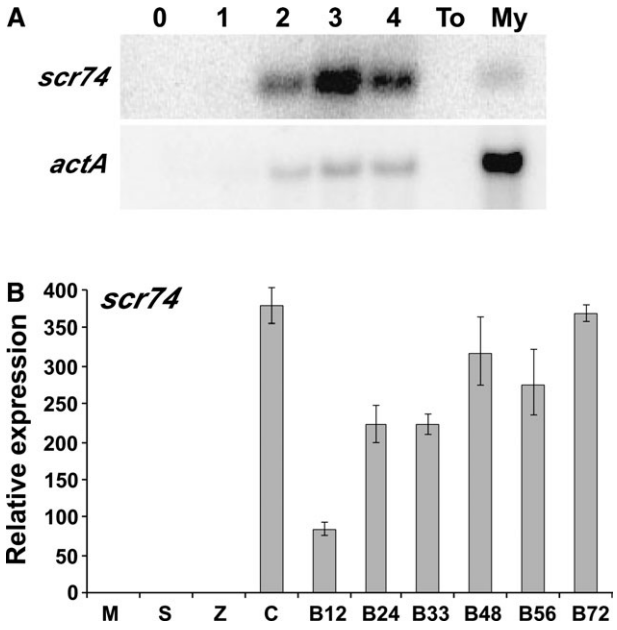

FIG. 1.—Time course expression of *scr74* during infection of tomato and potato by *P. infestans*. (*A*) Total RNA isolated from infected leaves of tomato, 0, 1, 2, 3, and 4 days after inoculation (0–4), from noninfected tomato leaves (To), and from *P. infestans* mycelium grown in synthetic medium (My) was sequentially hybridized with probes from the *scr74* and the actin A (*actA*) gene. (*B*) Expression levels of *scr74* in sporangia (S), zoospores (Z), germinating cysts (C), and at 12, 24, 33, 48, 56, and 72 hpi of susceptible cv. Bintje (B12, B24, B33, B48, B56, and B72) relative to those in mycelium (M) and normalized relative to *actA* expression levels. The tomato experiment was performed with *P. infestans* isolate 90128, and the potato one with isolate 88069.

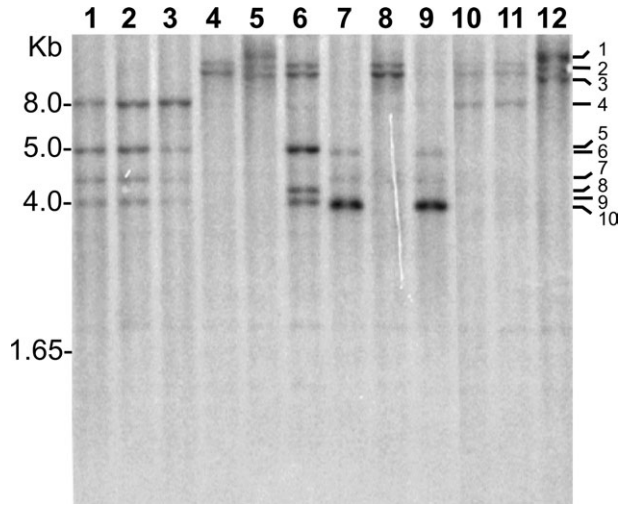

FIG. 2.—Occurrence of *scr74* sequences in *P. infestans* isolates. Southern blot analysis of genomic DNA from *P. infestans* isolates. DNA was digested with *Hind*III and hybridized with an *scr74* probe. Lane 1 contains DNA from US930468 (US1 clonal lineage); lane 2, US930258 (US1); lane 3, US940501 (US1); lane 4, US920165 (US1); lane 5, US940507 (US1); lane 6, US940330 (US7); lane 7, US960022 (US7); lane 8, US990004 (US6); lane 9, US940480 (US8); lane 10, US940504 (US13); lane 11, US940502 (US14); lane 12, US940289 (US16); lane 13, US970045 (US17); lane 14, US970001 (US17); lane 15, US970015 (US17); lane 16, US990025 (US11); lane 17, US980008 (US11); lane 18, US980066 (US11); lane 19, US940494 (US12). Positions of molecular length markers are given in kilobases on the left. The 10 different fragments identified are indicated on the right. Isolates US940507, US960022, US990004, US940480, and US970015 were used for PCR amplification and sequencing.

We identified one *scr74*-like EST, PH011A12, generated from germinating cysts, a pre-infection stage of *P. infestans*. Similar to PC015G09, the sequence of the cDNA corresponding to PH011A12 revealed a 222-bp ORF corresponding to a predicted translated product of 74 amino acids. However, the two predicted proteins differed in eight amino-acids. This result prompted us to use Southern hybridization for a preliminary investigation of the copy number of *scr74*-like sequences among 19 different *P. infestans* isolates representing 10 US clonal genotypes (Goodwin et al. 1998) (fig. 2). The *scr74* probe hybridized to at least 10 different genomic DNA fragments in 12 of 19 isolates (fig. 2). The number of hybridizing bands varied between isolates ranging from two to five, indicating a polymorphic gene family.

scr74 Sequences Are Highly Polymorphic in *P. infestans*

To examine sequence polymorphism in the *scr74* genes, we used PCR amplification with primers flanking the ORF, followed by DNA sequencing. We selected five *P. infestans* isolates (US940507, US960022, US990004, US940480, and US970015) representing five US clonal lineages and all 10 different fragments detected by Southern blots (fig. 2). In addition, we included DNA from strain IPO-0, a well-established lab strain (Vleeshouwers et al. 2000). Direct sequencing of amplicons obtained from genomic DNA of the six isolates resulted in mixed sequences, suggesting that the primers amplified multiple alleles or paralogs of *scr74*. Therefore, we cloned the amplicons and generated high-quality sequences (phred Q $>$ 20) of the 304 bp inserts of 45 different clones. In total, 21 different sequences encoding 19 predicted amino acid sequences were obtained (table 1). Polymorphisms were detected in 32 of the 304 nucleotides. Most of the polymorphic sites (31/225) were in the coding sequences, whereas only 1 site of the 79 untranslated region (UTR) nucleotides sequenced was polymorphic. Both sequences SCR74-C4 and E6 contained premature stop codons and were excluded from further analysis. Sequences SCR74-B10 and D5 were obtained from five independent amplicons and were identical to ESTs PC015G09 and PH011A12, respectively.

Multiple alignments of the 17 predicted SCR74 amino acid sequences revealed that eight conserved cysteines define the *scr74* gene family signature (fig. 3, the nucleotide multiple alignment is available online in the Supplementary Material). No differences in length (74 amino acids) were observed. Each member of the SCR74 family was predicted to contain a signal peptide (positions 1 to 21), and an extracellular mature protein of 53 amino acids (positions 22 to 74, the last residue of SCR74; fig. 3). Amino acid sequences of the SCR74 family were highly polymorphic. A total of 21 polymorphic amino acid sites were identified. Interestingly, 19 of 21 polymorphic sites were located in the mature protein, whereas only 2 of 21 were in the signal peptide, suggesting that amino acid variation was more frequent in the mature protein than in the signal peptide. Amino acid substitutions involved different chemical classes of amino acids, such as hydrophobic amino acids,

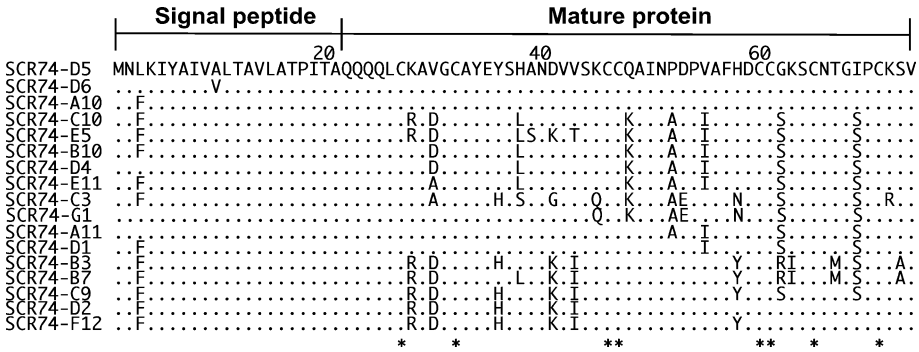

FIG. 3.—Multiple sequence alignment of 17 SCR74 amino acid sequences from *Phytophthora infestans*. Single-letter amino acid codes were used. Identical amino acids are indicated by dots. The eight conserved cysteine residues are indicated by asterisks. Residue numbers are denoted above the sequences. The predicted signal peptide starts from the first residue and consists of 21 residues. The mature protein ranges from the twenty-second to the last residue.

including phenylalanine (F), leucine (L), and alanine (A), and charged amino acids, including arginine (R), lysine (K), and histidine (H).

scr74 Genes Are Under Diversifying Selection

To characterize the selection pressures underlying sequence diversification in the SCR74 family, we calculated d_N and d_S across the entire ORF sequences. We found that d_N value was greater than d_S ($\omega = d_N / d_S > 1$) in 99 of 136 pairwise sequence comparisons of 17 complete nucleotide sequences of *scr74* (fig. 4*A*, see also the online Supplementary Material for the complete data set). These results provide evidence that diversifying selection has acted on the *scr74* gene family.

Ratios of d_N to d_S Are Elevated in the Mature Protein Region of SCR74

Diversifying selection typically acts on particular domains or amino acid sites within a given protein. To test for deviation in the substitution pattern of different regions of SCR74, we calculated d_N and d_S separately for the signal peptide and the mature protein regions. We found that d_N exceeded d_S for 123 of 136 possible pairwise comparisons of 17 nucleotide sequences of the mature protein region of SCR74 (fig. 4*B*). In contrast, d_N was greater than d_S in only 28 of 136 pairwise comparisons for the signal peptide region of SCR74 (fig. 4*B*). These results suggest that the mature protein is more divergent than the signal peptide, and that diversifying selection is probably an important evolutionary force in shaping sequence variation of the mature SCR74 protein.

SCR74 Amino Acid Sites Under Diversifying Selection

To detect the particular amino acid sites under diversifying selection in the *scr74* gene family, we applied three pairs of ML models of codon substitution: M3/M0, M8/M7, and M2/M1 (Nielsen and Yang 1998; Yang et al. 2000). The discrete model M3 with three site classes suggested that about 23% of the amino acid sites were under diversifying selection with $\omega_1 = 5.542$, whereas about 16% of amino acid sites were under strong diversifying selection with $\omega_2 = 14.711$ (table 2). The LRT for comparing M3 with M0 is $2\Delta L = 2 \times [-638.61-(-619.35)] = 38.52$, which is greater than the χ^2 critical value (13.28 at the 1% significance level, with degrees of freedom = 4; table 2). This indicates that the discrete model M3 fits the data significantly better than the neutral model M0, which does not allow for the presence of diversifying selection sites with $\omega > 1$. We then used the empirical Bayes theorem to identify 21 amino acid sites implicated as being under diversifying selection with greater than 99% confidence under the discrete model M3 (table 2). We plotted the position of the 21 diversifying selection sites in SCR74 (fig. 5). Interestingly, about 90% (19 of 21) of amino acid sites were located in the mature SCR74 protein, whereas only about 10% (two of 21) of amino acid sites were in the signal peptide. Again, this suggests that sites under diversifying selection occur more frequently in the mature protein of SCR74.

We also performed the LRT between the null model M7 (beta) and the alternative model M8 (beta+ω). The model M8 showed that about 68% of sites had ω from a U-shaped beta distribution, and about 32% of sites were under strong diversifying selection with $\omega = 10.737$. The difference between model M7 and model M8 was statistically significant, as indicated by the LRT: $2\Delta L = 2 \times [-643.20-(-619.46)] = 47.48$, which is greater than the χ^2 critical value (9.21 at 1% significance level, with degrees of freedom = 2; table 2). Thus, model M8 fitted the data significantly better than model M7. Under model M8, using the empirical Bayes theorem, we identified the same sites under diversifying selection as the ones identified under model M3, although the confidence level of diversifying selection sites identified varied from 73% to greater than 99% (table 2).

The selection model M2 did not identify any sites under diversifying selection. The probable reason is that the neutral model M1 failed to account for sites with $0 < \omega < 1$ that occur in the SCR74 data set (Yang et al. 2000). Thus, the small proportion of sites with $\omega > 1$ in the SCR74 data set were incorrectly added to the class of neutral sites with $\omega = 1$ using this model (Yang et al. 2000).

scr74 Genes Are Clustered in the *P. infestans* Genome

To further investigate gene copy number and organization, *scr74* was hybridized to a *P. infestans* BAC library constructed from strain T30-4 (Whisson et al.

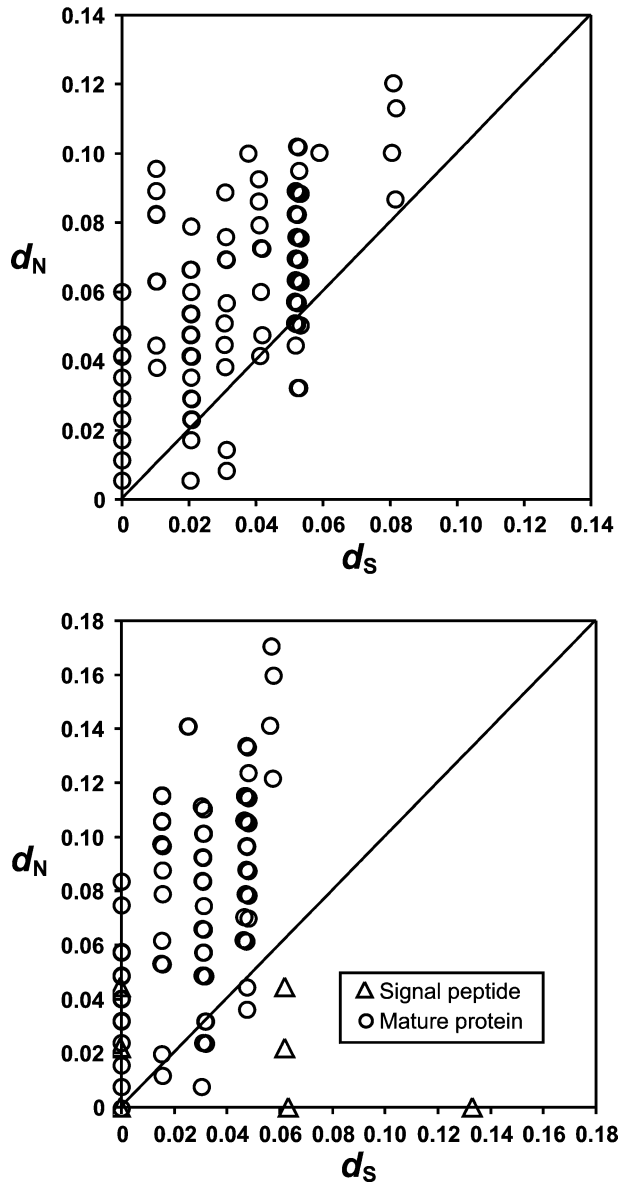

FIG. 4.—Pairwise comparison of nucleotide substitution rates in 17 SCR74 sequences from *Phytophthora infestans*. (*A*) Comparison for the whole SCR74 sequences. (*B*) Comparison for signal peptide and mature protein regions of SCR74. The nonsynonymous (d_N) and synonymous (d_S) substitution rates were estimated using the approximate method of Nei and Gojobori (1986) with the PAML software. The diagonal line indicates $d_N = d_S$, meaning neutral selection. Points above the line represent diversifying selection with $\omega = d_N / d_S > 1$. Points below the line indicate purifying selection with $\omega = d_N / d_S < 1$.

2001). Twelve hybridizing BAC clones were identified. Given that the BAC library was estimated to represent 10-fold genome coverage, 12 hybridizing clones suggest that *scr74* is either a single-copy gene or is present as a tightly clustered gene family in this strain. Southern hybridization of *scr74* to the 12 clones, restriction digested with *Hind*III, is presented in figure 6. Three clones, 11G3, 12O12, and 42H10, contained two hybridizing restriction fragments, suggesting the presence of two copies of *scr74*. Clone 42H10 was partially sequenced at the MIT Broad Institute (GenBank accession number AC147005). A 31,592 bp contig from this sequence was found to contain two copies of *scr74* sequence separated by 24,608 bp between ATG start codons. Two additional groups of BAC clones showed single hybridizing restriction fragments in two discrete size classes. These may represent a third and fourth copy of the *scr74* gene, or alleles of a third copy. Clone 19M21, containing the larger size class of hybridizing restriction fragment, has also been partially sequenced (83 kb of an estimated total size of 130 kb) at the Broad Institute (accession number AC147508). No sequences similar to *scr74* were found in the 19M21 partial sequence. A greater sequence coverage of the clone, or targeted sequencing of regions between existing contigs, may be required to locate this *scr74*-like gene copy. Clone 64I10 contains the remaining size class of hybridizing restriction fragment. This BAC was sub-cloned and sequenced at SCRI, and a third sequence variant of *scr74* was obtained. Southern analysis, BAC end-sequencing, and PCR were used as described previously (Whisson et al. 2001) to show that all 12 BACs are contiguous. These analyses also indicated that at least three copies of the *scr74* gene are clustered in a 300-kb region of the *P. infestans* T30-4 genome (results not shown).

Phylogenetic Analysis of the SCR74 Family Suggests Recombination Contributed to Sequence Divergence

We investigated whether recombination has contributed to the evolution of the *scr74* gene family. Detecting evidence of recombination in a short sequence alignment (225 bp) and with low average pairwise sequence divergence (0.06 substitutions per position) was difficult because of the low signal in the data. We used the Split Decomposition method (Huson 1998), which uses all columns in the alignment and is therefore most likely to find evidence. We also included sequences *scr74*-C4 and E6, despite the fact they are predicted to encode truncated proteins. The SplitsTree program (Huson 1998) tests the percentage fit to a phylogenetic tree topology (100% fit implies no evidence of recombination, with 80% set as the threshold for significant evidence). The fit (45.5%) for our alignment was significant. The SplitsTree network diagram (fig. 7*A*) indicated that there are six non-recombinant sequences falling into three groups (one containing C3; one containing E5, C4, and E6; and one containing B3 and B7) and a group of 13 sequences that showed evidence of recombination.

We also analyzed the sequences found in the BACs. The BAC sequence from clone 64I10 aligned most closely with sequences E5, C4, and E6 (three sequences which do not appear to have resulted from recombination). The stop codon in sequence C4 should result in a translated product of 24 amino acids (barely longer than the cleaved signal peptide), whereas E6 would be truncated after 71 amino acids. Genomic sequence from 64I10 contains both stop codons, and sequence E5 contains neither. The two nucleotide substitutions that result in the stop codons are the only polymorphisms found among these four sequences, suggesting that this copy of *scr74* is a pseudogene. The two copies of *scr74* found within the sequence of BAC 42H10 (42H10-1/42H10-2) are both predicted to encode

Table 2
Likelihood Ratio Test Results

Model	Estimates of Parameters	InL[a]	Diversifying Selection Sites[b]	Model Comparison	2ΔL[c]	χ^2 Critical Value (1%)	Degrees of Freedom
M0: one ratio	$\omega = 2.093$	−638.61	Not allowed				
M1: neutral	$P_0 = 0.458$, $P_1 = 0.542$	−638.25	Not allowed				
M2: selection	$P_0 = 0.458$, $P_1 = 0.542$, **$P_2 = 0.00003$, $\omega_2 = 0.00002$**	−638.25	None	M1 vs. M2	0	9.21	2
M3: discrete	$P_0 = 0.607$, $P_1 = 0.229$, **$P_2 = 0.164$,** $\omega_0 = 0.00001$, $\omega_1 = 5.542$, **$\omega_2 = 14.711$**	−619.35	**3 F 10 A 28 K 30 V 36 Y 38 H 39 A 41 D 43 V 45 K 48 Q 52 P 53 D 55 I 58 H 62 S 63 K 67 T 69 S 72 K 73 S**	M0 vs. M3	38.52	13.28	4
M7: beta	$P = 1.791$, $q = 0.001$	−643.20	Not allowed				
M8: beta+ω	$P_0 = 0.681$, $P = 23.072$, $q = 99(\infty)$, **$P_1 = 0.319$, $\omega = 10.737$**	−619.46	**3 F** 10 A **28 K 30 V 36 Y 38 H** 39 A **41 D 43 V** 45 K 48 Q 52 P 53 D **55 I 58 H 62 S** 63 K 67 T **69 S** 72 K 73 S	M7 vs. M8	47.48	9.21	2

NOTE.—Twice the difference in log likelihood ratio between a null model and an alternative model is compared with a χ^2 distribution with degrees of freedom will test whether an alternative model fits the data better than the null model. For example, the likelihood ratio test statistic for comparing M3 with M0 is $2\Delta L = 2 \times [-638.61 - (-619.35)] = 38.52$ that is much greater than the χ^2 critical value (13.28 at 1% significance level with degrees of freedom = 4). This indicated that the discrete model M3 significantly better fitted the data than the neutral model M0. So diversifying selection sites identified by the model M3 are significant. Both model M3 and model M8 identified the same diversifying selection sites.

[a] InL = log likelihood value.

[b] Amino acid sites inferred to be under diversifying selection with a probability >99% are in bold; 73% − 90% are underlined.

[c] Likelihood ratio test: $2\Delta L = 2(\text{InL}_{\text{alternative hypothesis}} - \text{InL}_{\text{null hypothesis}})$.

full-length proteins. Sequence 42H10-1 is identical to C3 (a nonrecombinant), whereas 42H10-2 is most similar to sequences D4 and B10. This suggests that the detected recombination has likely arisen at the 42H10-2 locus.

In addition, we used a window approach to detect likely recombination. We used the DSS method (McGuire and Wright 2000) in the TOPALi program (Milne et al. 2004) with a range of half-window sizes. The only significant result followed splitting the SCR74 alignment into two halves (i.e., half-window of 122 bp) which was significant at $P < 0.05$ based on a parametric bootstrapping test to compare two tree topologies (Goldman, Anderson, and Rodrigo 2000), but using a DSS statistic rather than Log Likelihood.

We compared phylogenetic trees constructed using the first and second halves of the alignment to see if we could infer recombinants (fig. 7*B* and 7*C*). The automatic detection algorithm in TOPALi failed to recover recombinants, presumably because of low signal in the data. However, visual inspection confirmed that the six sequences [(B3, B7); (E6, C4, E5); and C3] lacking evidence of recombination in the SplitsTree diagram appeared to maintain their relative positions, while a considerable number of topology shifts have occurred with respect to the remaining sequences. This pattern is consistent with recombination having been active in at least one of the *scr74* loci.

Discussion

In this study, we describe, characterize, and investigate molecular evolution and structural organization of the

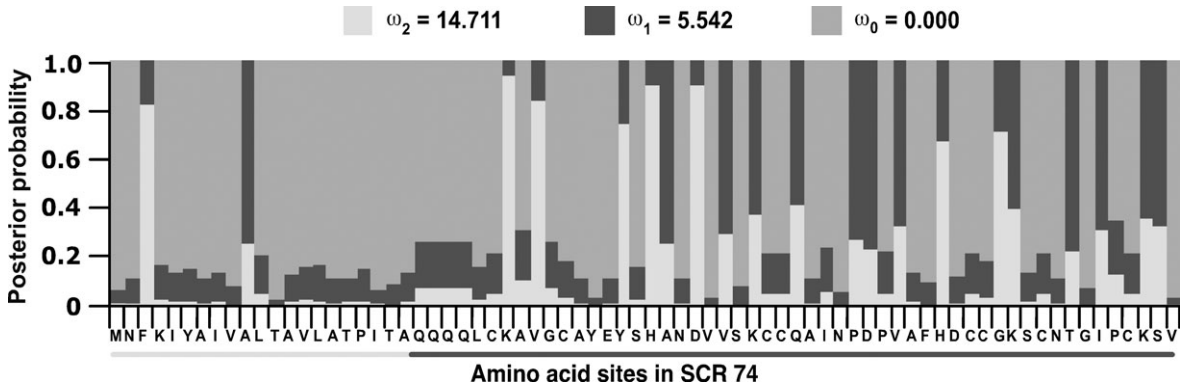

FIG. 5.—Posterior probabilities for site classes estimated under the discrete model M3 in PAML along the SCR74 protein sequence. ML estimates of probabilities and ω ratios for the three site classes are $P_0 = 0.607$, $P_1 = 0.229$, and $P_2 = 0.164$, and $\omega_0 = 0.000$ (gray), $\omega_1 = 5.542$ (dark gray), and $\omega_2 = 14.711$(light gray). Amino acid sites having higher posterior probabilities for site classes ω_1 or ω_2 are potentially under diversifying selection. For example, the posterior probabilities at the site three (F) are 0.000, 0.179, and 0.821, indicating that this site is likely to be under diversifying selection. Signal peptides are underlined in light gray. Mature proteins are underlined in dark gray.

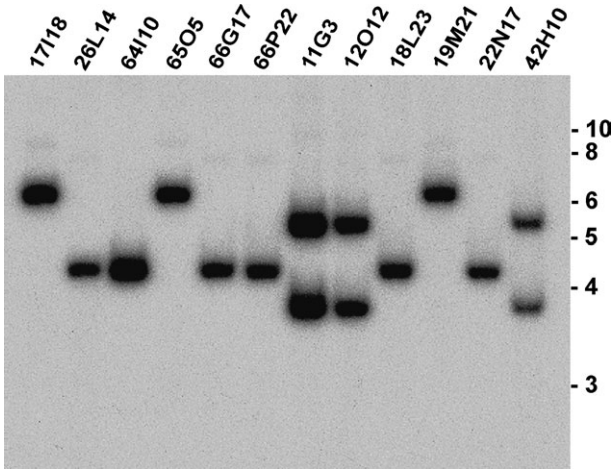

FIG. 6.—Hybridization of *scr74* to 12 bacterial artificial chromosome (BAC) clones digested with *Hin*dIII. BAC clone names are indicated on top. Molecular weight marker sizes (kb) are given to the right of the figure.

highly polymorphic *scr74* gene family. This family was identified from the potato late blight pathogen *P. infestans* and was predicted to encode secreted cysteine-rich proteins with similarity to the phytotoxin necrosis-inducing protein PcF of *P. cactorum* (Orsomando et al. 2001). The *scr74* gene was initially identified by computer-aided mining of ESTs from infection stages of *P. infestans* to possess features reminiscent of pathogen effector molecules based on the criteria of Torto et al. (2003) and Bos et al. (2003). Further molecular characterization of *scr74* showed that these genes are widely distributed and form a highly polymorphic gene family with at least 21 different sequences identified. Also, expression of *scr74* was upregulated in germinating cysts and during infection of tomato and potato by *P. infestans*. Based on prevalent models of plant-pathogen co-evolution, some effectors, notably those with *Avr* function, are predicted to exhibit significant sequence variation within populations of the pathogen (van't Slot and Knogge 2002; Bos et al. 2003; Dodds et al. 2004). The polymorphic nature and expression pattern of *scr74*, along with structural features such as secretion and similarity to a phytotoxin, suggest that these genes may encode pathogen effector proteins and may play a role in the infection process, perhaps as an Avr protein.

Unfortunately, using the method described by Bos et al. (2003), we could not perform association genetic analyses to assess the likelihood that *scr74* is an *Avr* gene. All isolates examined produced mixed amplicons, and DNA hybridization experiments indicated the probable presence of *scr74* paralogs in all isolates. The occurrence of multiple *scr74*-like sequences in the *P. infestans* genome confounds association studies because it is not possible to determine unambiguously the array of *scr74* sequences of a given isolate by PCR amplification. Detailed characterization of *scr74* genomic loci will help to design improved strategies for isolate genotyping.

Evolutionary analyses revealed that the *P. infestans* *scr74* gene family exhibits an unusual pattern of evolution and is likely to be under diversifying selection as detected by both the approximate method of Nei and Gojobori (1986) and the ML method (Nielsen and Yang 1998; Yang et al. 2000; Yang and Bielawski 2000). In most proteins, neutral and purifying selection are thought to be major evolutionary forces, with a high proportion of amino acid sites conserved as a result of structural and functional constraints (Li 1997; Golding and Dean 1998). Under these circumstances, the approximate method should not be sensitive enough to detect diversifying selection because it averages ω ratios over all sites of the protein (Yang and Bielawski 2000). Nevertheless, using the approximate method, we detected diversifying selection across the entire *scr74* sequence. This is likely because of the large number of highly divergent *scr74* sequences (Yang and Bielawski 2000) and the relatively small and simple structure of these genes.

Compared to the approximate method, the ML method developed by Yang and collaborators (Yang et al. 2000; Yang and Bielawski 2000) is more sensitive for detecting diversifying selection and can also identify the particular amino acids sites under diversifying selection. We also obtained significant support for diversifying selection in the SCR74 family using two of three models implemented in the ML method. Interestingly, all 21 polymorphic amino acid sites were identified as being under diversifying selection although at different confidence levels.

We also found higher d_N to d_S ratios in the mature protein region than in the signal peptide region of SCR74. Moreover, 19 of 21 amino acid sites under diversifying selection (90%) are located in the mature protein region. Remarkably, most polymorphic nucleotide sites, 27 out of a total of 32 polymorphisms or a 16.7% polymorphism rate, were in the mature protein region of the ORF. The signal peptide region and the sequenced portion of the UTRs accounted for only four and one polymorphic sites, respectively. The nucleotide polymorphism rates in the signal peptide region and the UTRs were also lower than in the mature protein region at 6.3 % and 1.3 %, respectively. This rapid accumulation of nucleotide changes and amino acid replacements in the mature protein region indicates that diversifying selection has been acting mainly on this portion of the protein. Thus, diversifying selection may have acted on the domain of SCR74 that is directly related to its biological activity, and it may have shaped functional diversity in this protein family. Similar observations have been made for other genes under diversifying selection, such as murine β-Defensins (Morrison et al. 2003) and toxin genes in the venomous gastropod *Conus* (Thomas and Stephen 1999, 2000).

How did the *scr74* genes evolve to result in divergent, rapidly accumulated nucleotide changes in sequences corresponding to the mature protein but remain relatively conserved in the signal peptide and UTR sequences? Gene duplication, followed by functional divergence of duplicated genes, is an important evolutionary force for the emergence of new gene function (Stephens 1951; Nei 1969; Ohno 1970; Ohta 1980, 1993). Goodman (1976), Goodman et al. (1987) first reported that the rate of amino acid substitutions was accelerated following duplication of hemoglobin genes into α and β hemoglobins and suggested that this acceleration was caused by natural

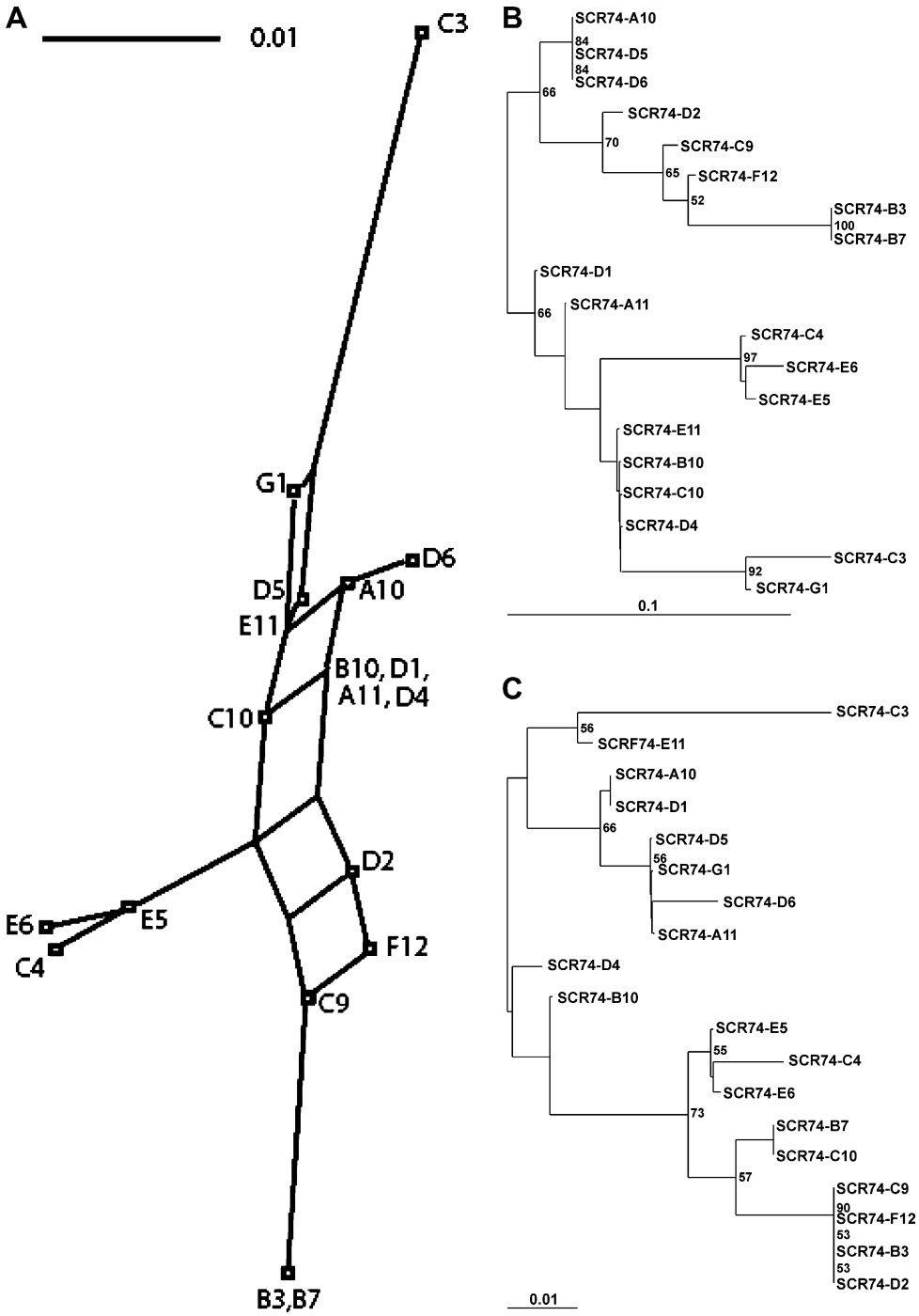

FIG. 7.—Evidence for genetic recombination between *scr74* copies. (*A*) Output from SplitsTree. The data are not consistent with a simple tree and are therefore represented by a tree-like network. (*B, C*) Parts of the output are from the TOPALi software package for detecting mosaic sequences and enabling comparisons of tree topologies for different portions of a gene sequence. Panels *B* and *C* correspond to Neighbor-Joining trees generated with the first and second halves of the SCR74 protein alignment, respectively. Note the lack of congruence between the two trees. The scale bars represent weighted sequence divergence. Bootstrap values higher than 50% are shown at the nodes.

selection. In fact, it has been debatable whether rapid evolution (acceleration in the rate of amino acid substitutions) in gene families following gene duplication occurs by diversifying selection or relaxation of functional constraints due to gene redundancy (Kimura 1983; Li 1985; Ohta 1993, 1994). Based on our analyses, we propose an evolutionary model that involves gene duplication followed by functional divergence of *scr74* genes. We found at least three copies of *scr74* to be clustered in a region of the *P. infestans* T30-4 genome of less than 300 kb, suggesting that gene duplication may have occurred. Moreover, we also detected genetic recombination in at least one of the *scr74* gene loci. Both diversifying selection and relaxation of selective constraints may have played

complementary roles in promoting sequence and functional divergence following gene duplication and recombination. This explanation is consistent with those for the adaptive evolution of primate murine β-Defensin gene family in vertebrates (Morrison et al. 2003) and pancreatic ribonuclease genes in a leaf-eating monkey (Zhang, Zhang, and Rosenberg 2002). We are currently further analyzing *P. infestans* genomic clones containing *scr74* genes to gain more insights into the role of gene duplication and recombination in the molecular evolution of this family.

What is driving the diversifying selection observed in *scr74* genes? The co-evolutionary "arms race" model states that adaptation and counter-adaptation between host and pathogen or parasite drive their antagonistic co-evolution (Dawkins and Krebs 1979). Co-evolution of host-pathogen or host-parasite is thought to generate the evolutionary forces that shape the genes involved in these interactions. Recently, a number of genes involved in host-pathogen antagonistic co-evolution have been revealed to be under diversifying selection, resulting in accelerated amino acid substitutions in sites that determine recognition by the host or the pathogen (Stahl and Bishop 2000). Several toxins and their counteracting detoxifying enzymes, as well as hydrolytic enzymes and their corresponding inhibitors, are thought to be involved in antagonistic co-evolution, and diversifying selection acting on these molecules has been documented (Leckie et al. 1999; Thomas and Stephen 1999, 2000; Bishop, Dean, and Mitchell-Olds 2000; Stotz et al. 2000). For example, several studies showed that diversifying selection has acted on hypervariable solvent-exposed residues of the leucine-rich repeat (LRR) region of some plant disease resistance R proteins from *Arabidopsis*, lettuce, tomato, rice, and flax (Parniske et al. 1997; Meyers et al. 1998; Wang et al. 1998; Noel et al. 1999; Ellis, Dodds, and Pryor 2000; Mondragon-Palomino et al. 2002). Diversifying selection in *R* genes is thought to reflect an "arms race" in plant-pathogen co-evolution in order to select novel resistance specificities (Endo, Ikeo, and Gojobori 1996; Stahl and Bishop 2000; Yang 2002). Interestingly, diversifying selection was recently detected in the flax rust *Avr* genes *AvrL567* that are recognized by the flax *L5, L6*, or *L7 R* genes, lending additional support for an arms race model of gene-for-gene evolution (Dodds et al. 2004).

Here we show that the *scr74* gene family of the oomycete plant pathogen *P. infestans* is under diversifying selection. Although the nature of the selective pressures remains unclear, we propose that diversifying selection acting on the mature SCR74 protein has resulted in functionally important intraspecific polymorphisms. SCR74 is a secreted protein with similarity to a necrosis-inducing protein and has many hallmarks of an effector protein that may play a role in the infection process. The *scr74* genes are also significantly upregulated during infection of host plants. Altogether, this suggests that the selective forces that shaped *scr74* evolution might be related to host-pathogen co-evolution. Future functional analyses, such as the *in planta* expression assays described by Huitema et al. (2004), combined with site-directed mutagenesis, will be necessary to determine the nature and significance of the adaptive changes, as well as to dissect the functional basis of adaptive evolution in *scr74*.

We did not detect the patterns of polymorphisms and diversifying selection observed in the SCR74 family in elicitins, a well-studied family of secreted cysteine-rich proteins of *Phytophthora* that has been implicated in host specificity (Kamoun, Lindqvist, and Govers 1997; Kamoun et al. 1997; Qutob et al. 2003). In repeated analyses using the approximate and ML methods, we found no evidence of positive selection in elicitin sequences from *P. infestans* and *Phytophthora sojae* (unpublished data). This suggests that distinct selective forces shaped the SCR74 and elicitin families throughout the evolution of *Phytophthora*. Slow rates of evolution in elicitins are consistent with the view that these proteins are recognized by ancient broad-spectrum plant genes and are implicated in species-level or non-host resistance (Kamoun 2001).

This study provides support for using diversifying selection as an additional selection criterion for candidate effector genes from EST databases, and it therefore augments previously defined criteria, such as secretion and intraspecific polymorphism (Bos et al. 2003; Torto et al. 2003). In the future, accumulation of cDNA and genomic sequences from plant and animal pathogens will yield more opportunities to investigate patterns of diversifying selection in effector gene families. Ultimately, analyses of diversifying selection will help us to establish functional connections between pathogen effectors and host defense processes, and to provide insights into the molecular basis of pathogen-host co-evolution.

Acknowledgments

We thank Ian Holford and Xiaodong Bai for help with computer-related problems; Shujing Dong and Diane Kinney for technical assistance; Chris Smart and Pieter van West for providing *P. infestans* DNA; Larry Madden, Guo-liang Wang, and Elisabeth Mueller for useful suggestions; and Stephen J. Gould for inspiration. This work was supported by Syngenta Biotechnology, by the National Science Foundation ([NSF] Plant Genome Research Program grant DBI-0211659), and by the Scottish Executive, Environment and Rural Affairs Department. We thank the Syngenta *Phytophthora* Consortium for access to sequences of *P. infestans*. Salaries and research support were provided by state and federal funds appropriated to the Ohio Agricultural Research and Development Center, the Ohio State University.

Literature Cited

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389–3402.

Apweiler, R., T. K. Attwood, A. Bairoch et al. (25 co-authors). 2001. The InterPro database, an integrated documentation resource for protein families, domains, and functional sites. Nucleic Acids Res. **29**:37–40.

Avrova, A. O., E. Venter, P. R. J. Birch, and S. C. Whisson. 2003. Profiling and quantifying differential gene transcription

in *Phytophthora infestans* prior to and during the early stages of potato infection. Fungal Genet. Biol. **40**:4–14.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. E. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science **290**:972–977.

Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer. 2002. The Pfam protein families database. Nucleic Acids Res. **30**:276–280.

Berry, A., and M. Kreitman. 1993. Molecular analysis of an allozyme cline: alcohol dehydrogenase in *Drosophila melanogaster* on the east coast of North America. Genetics **134**:869–893.

Birch, P. R. J., and S. Whisson. 2001. *Phytophthora infestans* enters the genomics era. Mol. Plant Pathol. **2**:257–263.

Bishop, J. G., A. M. Dean, and T. Mitchell-Olds. 2000. Rapid evolution in plant chitinases: molecular targets of selection in plant-pathogen coevolution. Proc. Natl. Acad. Sci. USA **97**:5322–5327.

Bittner-Eddy, P. D., R. L. Allen, A. P. Rehmany, P. Birch, and J. L. Beynon. 2003. Use of suppression subtractive hybridization to identify downy mildew genes expressed during infection of *Arabidopsis thaliana*. Mol. Plant Pathol. **4**: 501–507.

Bos, J. I. B., M. Armstrong, S. C. Whisson, T. A. Torto, M. Ochwo, P. R. J. Birch, and S. Kamoun. 2003. Intraspecific comparative genomics to identify avirulence genes from *Phytophthora*. New Phytologist **159**:63–72.

Caten, C. E., and J. L. Jinks. 1968. Spontaneous variability of single isolates of *Phytophthora infestans*. I. Cultural variation. Can. J. Bot. **46**:329–347.

Dangl, J. L., and J. D. Jones. 2001. Plant pathogens and integrated defence responses to infection. Nature **411**:826–833.

Dawkins, R., and J. R. Krebs. 1979. Arms races between and within species. Proc. R. Soc. Lond. B Biol. Sci. **205**:489–511.

Dodds, P. N., G. J. Lawrence, A. M. Catanzariti, M. A. Ayliffe, and J. G. Ellis. 2004. The *Melampsora lini AvrL567* avirulence genes are expressed in haustoria and their products are recognized inside plant cells. Plant Cell **16**:755–768.

Ellis, J., P. Dodds, and T. Pryor. 2000. The generation of plant disease resistance gene specificities. Trends Plant Sci. **5**: 373–379.

Endo, T., K. Ikeo, and T. Gojobori. 1996. Large-scale search for genes on which positive selection may operate. Mol. Biol. Evol. **13**:685–690.

Ewing, B., and P. Green. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res. **8**:186–194.

Ewing, B., L. Hillier, M. C. Wendl, and P. Green. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res. **8**:175–185.

Fry, W. E., and S. B. Goodwin. 1997*a*. Re-emergence of potato and tomato late blight in the United States. Plant Dis. **81**:1349–1357.

———. 1997*b*. Resurgence of the Irish potato famine fungus. Bioscience **47**:363–371.

Golding, G. B., and A. M. Dean. 1998. The structural basis of molecular adaptation. Mol. Biol. Evol. **15**:355–369.

Goldman, N., J. P. Anderson, and A. G. Rodrigo. 2000. Likelihood-based tests of topologies in phylogenetics. Syst. Biol. **49**:652–670.

Goodman, M. 1976. Protein sequences in phylogeny. Pp. 141–59 *in* F. J. Ayala, ed. Molecular evolution. Sinauer Associates, Sunderland, Mass.

Goodman, M., J. Czelusniak, B. F. Koop, D. A. Tagle, and J. L. Slightom. 1987. Globins: a case study in molecular phylogeny. Cold Spring Harbor Symp. Quant. Biol. **52**:875–90.

Goodwin, S. B., C. D. Smart, R. W. Sandrock, K. L. Deahl, Z. K. Punja, and W. E. Fry. 1998. Genetic change within populations of *Phytophthora infestans* in the United States and Canada during 1994 to 1996: role of migration and recombination. Phytopathology **88**:939–949.

Haag, E. S., and J. R. True. 2001. Perspective: from mutants to mechanisms? Assessing the candidate gene paradigm in evolutionary biology. Evolution **55**:1077–1084.

Henikoff, J. G., E. A. Greene, S. Pietrokovski, and S. Henikoff. 2000. Increased coverage of protein families with the blocks database servers. Nucleic Acids Res. **28**:228–230.

Hudson, R. R. 1993. Levels of DNA polymorphism and divergence yield important insights into evolutionary processes. Comment. Proc. Natl. Acad. Sci. USA **90**:7425–7426.

Hughes, A. L., T. Ota, and M. Nei. 1990. Positive Darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major histocompatibility complex molecules. Mol. Biol. Evol. **7**:515–524.

Huitema, E., V. G. A. A. Vleeshouwers, D. M. Francis, and S. Kamoun. 2003. Active defence responses associated with non-host resistance of *Arabidopsis thaliana* to the oomycete pathogen *Phytophthora infestans*. Mol. Plant Pathol. **4**:487–500.

Huitema, E., J. I. Bos, M. Tian, J. Win, M. E. Waugh, and S. Kamoun. 2004. Linking sequence to phenotype in *Phytophthora*-plant interactions. Trends Microbiol. **12**: 193–200.

Huson, D. H. 1998. SplitsTree: analyzing and visualizing evolutionary data. Bioinformatics **14**:68–73.

Ina, Y. 1995. New methods for estimating the numbers of synonymous and nonsynonymous substitutions. J. Mol. Evol. **40**:190–226.

Kamoun, S., M. Young, C. Glascock, and B. M. Tyler. 1993. Extracellular protein elicitors from *Phytophthora*: host-specificity and induction of resistance to fungal and bacterial phytopathogens. Mol. Plant-Microbe Interact. **6**:15–25.

Kamoun, S., H. Lindqvist, and F. Govers. 1997. A novel class of elicitin-like genes from *Phytophthora infestans*. Mol. Plant-Microbe Interact. **10**:1028–1030.

Kamoun, S., P. van West, A. J. de Jong, K. de Groot, V. Vleeshouwers, and F. Govers. 1997. A gene encoding a protein elicitor of *Phytophthora infestans* is down-regulated during infection of potato. Mol. Plant-Microbe Interact. **10**:13–20.

Kamoun, S. 2001. Nonhost resistance to *Phytophthora*: novel prospects for a classical problem. Curr. Opin. Plant Biol. **4**:295–300.

Kamoun, S. 2003. Molecular genetics of pathogenic oomycetes. Eukaryotic Cell **2**:191–199.

Karl, S. A., and J. C. Avise. 1992. Balancing selection at allozyme loci in oysters: implications from nuclear RFLPs. Science **256**:100–102.

Karsch-Mizrachi, I., and B. F. Ouellette. 2001. The GenBank sequence database. Methods Biochem. Anal. **43**:45–63.

Kimura, M. 1983. The neutral theory of molecular evolution. Cambridge University Press, Cambridge, U.K.

Leckie, F., B. Mattei, C. Capodicasa, A. Hemmings, L. Nuss, B. Aracri, G. De Lorenzo, and F. Cervone. 1999. The specificity of polygalacturonase-inhibiting protein (PGIP): a single amino acid substitution in the solvent-exposed β-strand/β-turn region of the leucine-rich repeats (LRRs) confers a new recognition capability. EMBO J. **18**:2352–2363.

Letunic, I., L. Goodstadt, N. J. Dickens, T. Doerks, J. Schultz, R. Mott, F. Ciccarelli, R. R. Copley, C. P. Ponting, and P. Bork.

2002. Recent improvements to the SMAT domain-based sequence annotation resource. Nucleic Acids Res. **30**:242–244.
Li, W. H., C. I. Wu, and C. C. Luo. 1985. A new method for estimating synonymous and nonsynonymous rates on nucleotide substitution considering the relative likelihood of nucleotide and codon changes. Mol. Biol. Evol. **2**:150–174.
Li, W. H. 1985. Accelerated evolution following gene duplication and its implications for the neutralist-selectionist controversy. Pp. 333–352 *in* Ohta, T., and Aoki, K., eds. Population genetics and molecular evolution. Japan Scientific Societies Press, Tokyo, Japan.
Li, W. H. 1997. Molecular evolution. Sinauer Associates, Sunderland, Mass.
Lu, R., I. Malcuit, P. Moffett, M. T. Ruiz, J. Peart, A. J. Wu, J. P. Rathjen, A. Bendahmane, L. Day, and D. C. Baulcombe. 2004. High throughput virus-induced gene silencing implicates heat shock protein 90 in plant disease resistance. EMBO J. **3**:5690–5699.
Margulis, L., and K. V. Schwartz. 2000. Five kingdoms: an illustrated guide to the phyla of life on earth. W. H. Freeman and Co., New York, N. Y.
McGuire, G., and F. Wright. 2000. TOPAL 2.0: improved detection of mosaic sequences within multiple alignments. Bioinformatics **16**:130–134.
Meyers, B. C., K. A. Shen, P. Rohani, B. S. Gaut, and R. W. Michelmore. 1998. Receptor-like genes in the major resistance locus of lettuce are subject to divergent selection. Plant Cell **10**:1833–1846.
Milne, I., F. Wright, G. Rowe, D. F. Marshal, D. Husmeier, and G. McGuire. 2004. TOPALi: Software for automatic identification of recombinant sequences within DNA multiple alignments. Bioinformatics **20**:1806–1807.
Mondragon-Palomino, M., B. C. Meyers, R. W. Michelmore, and B. S. Gaut. 2002. Patterns of positive selection in the complete NBS-LRR gene family of *Arabidopsis thaliana*. Genome Res. **12**:1305–1315.
Morrison, G. M., C. A. M. Semple, F. M. Kilanowski, R. E. Hill, and J. R. Dorin. 2003. Signal sequence conservation and mature peptide divergence within subgroups of the murine β-Defensin gene family. Mol. Biol. Evol. **20**:460–470.
Nei, M. 1969. Gene duplication and nucleotide substitution in evolution. Nature **221**:40–42.
Nei, M., and T. Gojobori. 1986. Simple methods for estimating the number of synonymous and nonsynonymous nucleotide substitutions. Mol. Biol. Evol. **3**:418–426.
Nicholls, H. 2004. Stopping the rot. PLoS Biol. **2**:891–895.
Nielsen, H., J. Engelbrecht, S. Brunak, and G. von Heijne. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. **10**:1–6.
Nielsen, H., and A. Krogh. 1998. Prediction of signal peptides and signal anchors by a hidden Markov model. Pp. 122–130 *in* Proceeding of the Sixth International Conference on Intelligent Systems for Molecular Biology (ISMB 6). AAAI Press, Menlo Park, Calif.
Nielsen, R., and Z. Yang. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. Genetics **148**:929–936.
Noel, L., T. L. Moores, E. A. Van Der Biezen, M. Parniske, M. J. Daniels, J. E. Parker, and J. D. Jones. 1999. Pronounced intraspecific haplotype divergence at the RPP5 complex disease resistance locus of *Arabidopsis*. Plant Cell **11**:2099–2112.
Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag, New York.
Ohta, T. 1980. Evolution and variation of multigene families (Lecture notes in biomathematics, vol. 37). Springer-Verlag, New York.
Ohta, T. 1993. Pattern of nucleotide substitutions in growth hormone-prolactin gene family: a paradigm for evolution by gene duplication. Genetics **134**:1271–1276.
Ohta, T. 1994. Further examples of evolution by gene duplication revealed through DNA sequence comparisons. Genetics **138**:1331–1337.
Orr, H. A., and J. A. Coyne. 1992. The genetics of adaptation: a reassessment. Am. Nat. **140**:725–742.
Orsomando, G., M. Lorenzi, N. Raffaelli, M. D. Rizza, B. Mezzetti, and S. Ruggieri. 2001. Phytotoxic protein PcF, purification, characterization, and cDNA sequencing of a novel hydoxyproline-containing factor secreted by the strawberry *pathogen Phytophthora cactorum*. J. Biol. Chem. **276**:21578–21584.
Parniske, M., K. E. Hammond-Kosack, C. Golstein, C. M. Thomas, D. A. Jones, K. Harrison, B. B. Wulff, and J. D. Jones. 1997. Novel disease resistance specificities result from sequence exchange between tandemly repeated genes at the Cf-4/9 locus of tomato. Cell **91**:821–832.
Qutob, D., E. Huitema, M. Gijzen, and S. Kamoun. 2003. Variation in structure and activity among elicitins from *Phytophthora sojae*. Mol. Plant Pathol. **4**:119–124.
Ristaino, J. B. 2002. Tracking historic migrations of the Irish potato famine pathogen, *Phytophthora infestans*. Microbes Infect. **4**:1369–1377.
Sambrook, J., and D. W. Russell. 2001. Molecular cloning. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, N. Y.
Schiermeier, Q. 2001. Russia needs help to fend off potato famine, researchers warn. Nature **440**:1011.
Shattock, R. C. 2002. *Phytophthora infestans*: populations, pathogenicity and phenylamides. Pest Manage. Sci. **58**:944–950.
Smart, C. D., and W. E. Fry. 2001. Invasions by the late blight pathogen: renewed sex and enhanced fitness. Biol. Invas. **3**:235–243.
Sogin, M. L., and J. D. Silberman. 1998. Evolution of the protests and protistan parasites from the perspective of molecular systematics. Int. J. Parasitol. **28**:11–20.
Stahl, E. A., and J. G. Bishop. 2000. Plant-pathogen arms races at the molecular level. Curr. Opin. Plant Biol. **3**:299–304.
Stephens, S. G. 1951. Possible significance of duplication in evolution. Adv. Genet. **4**:247–265.
Stotz, H. U., J. G. Bishop, C. W. Bergmann, M. Koch, P. Albersheim, A. G. Darvill, and J. M. Labavitch. 2000. Identification of target amino acids that affect interactions of fungal polygalacturonases and their plant inhibitors. Physiol. Mol. Plant Pathol. **56**:117–130.
Thomas, F. D., and R. P. Stephen. 1999. Molecular genetics of ecological diversification: duplication and rapid evolution of toxin genes of the venomous gastropod *Conus*. Proc. Natl. Acad. Sci. USA **96**:6820–6823.
Thomas, F. D., and R. P. Stephen. 2000. Evolutionary diversification of multigene families: allelic selection of toxins in predatory cone snails. Mol. Biol. Evol. **17**:1286–1293.
Thompson, J. D., T. J. Gibson, F. Plewniak, F. Jeanmougin, and D. G. Higgins. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25**:4876–4882.
Torto, T. A., S. Li, A. Styer, E. Huitema, A. Testa, N. A. Gow, P. van West, and S. Kamoun. 2003. EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen *Phytophthora*. Genome Res. **13**:1675–1685.
Unkles, S. E., R. P. Moon, A. R. Hawkins, J. M. Duncan, and J. R. Kinghorn. 1991. Actin in the oomycetous fungus *Phytophthora infestans* is the product of several genes. Gene **100**:105–112.

van't Slot, K. A. E., and W. Knogge. 2002. A dual role for microbial pathogen-derived effector proteins in plant disease and resistance. Crit. Rev. Plant Sci. **21**:229–271.

Vleeshouwers, V. G. A. A., W. van Dooijeweert, F. Govers, S. Kamoun, and L. T. Colon. 2000. The hypersensitive response is associated with host and nonhost resistance to *Phytophthora infestans*. Planta **210**:853–864.

Wang, G. L., D. L. Ruan, W. Y. Song et al. (12 co-authors). 1998. Xa21D encodes a receptor-like molecule with a leucine-rich repeat domain that determines race-specific recognition and is subject to adaptive evolution. Plant Cell **10**:765–779.

Whisson, S. C., T. van der Lee, G. J. Bryan, R. Waugh, F. Govers, and P. R. J. Birch. 2001. Physical mapping across an avirulence locus of *Phytophthora infestans* using a high representative, large insert bacterial artificial chromosome library. Mol. Genet. Genomics **266**:289–295.

Wu, C. I. 2001. The genic view of the process of speciation. J. Evol. Biol. **14**:851–865.

Yang, Z., and J. P. Bielawski. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol. Evol. **15**:496–503.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555–556.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431–449.

Yang, Z. 2002. Inference of selection from multiple species alignments. Curr. Opin. Genet. Dev. **12**:688–694.

Zhang, J., Y. Zhang, and H. Rosenberg. 2002. Adaptive evolution of a duplicated pancreatic ribonuclease gene in a leaf-eating monkey. Nat. Genet. **30**:411–415.

Brian Golding, Associate Editor

Accepted November 8, 2004