Supporting Online Material for

Phytophthora Genome Sequences Uncover Evolutionary Origins and Mechanisms of Pathogenesis

Brett M. Tyler,* Sucheta Tripathy, Xuemin Zhang, Paramvir Dehal, Rays H. Y. Jiang, Andrea Aerts, Felipe D. Arredondo, Laura Baxter, Douda Bensasson, Jim L. Beynon, Jarrod Chapman, Cynthia M. B. Damasceno, Anne E. Dorrance, Daolong Dou, Allan W. Dickerman, Inna L. Dubchak, Matteo Garbelotto, Mark Gijzen, Stuart G. Gordon, Francine Govers, Niklaus J. Grunwald, Wayne Huang, Kelly L. Ivors, Richard W. Jones, Sophien Kamoun, Konstantinos Krampis, Kurt H. Lamour, Mi-Kyung Lee, W. Hayes McDonald, Mónica Medina, Harold J. G. Meijer, Eric K. Nordberg, Donald J. Maclean, Manuel D. Ospina-Giraldo, Paul F. Morris, Vipaporn Phuntumart, Nicholas H. Putnam, Sam Rash, Jocelyn K. C. Rose, Yasuko Sakihama, Asaf A. Salamov, Alon Savidor, Chantel F. Scheuring, Brian M. Smith, Bruno W. S. Sobral, Astrid Terry, Trudy A. Torto-Alalibo, Joe Win, Zhanyou Xu, Hongbin Zhang, Igor V. Grigoriev, Daniel S. Rokhsar, Jeffrey L. Boore

*To whom correspondence should be addressed. E-mail: bmtyler@vt.edu

Published 1 September 2006, *Science* **313**, 1261 (2006)
DOI: 10.1126/science.1128796

This PDF file includes:

SUPPLEMENTARY ON-LINE MATERIAL

Tyler et al. *Phytophthora* Genome Sequences Uncover Evolutionary Origins and Mechanisms of Pathogenesis. Science Vol. 313, No. 5791, September 1, 2006.

MATERIALS AND METHODS

Whole genome sequencing.

Genomic DNA of *P. sojae* was extracted from mycelia of strain P6497 (*1*) that has been used extensively for genetic and genomic studies, including production of ESTs and BAC libraries. *P. ramorum* DNA was obtained from strain Pr-102 (ATCC MYA-2949) that was isolated from *Quercus agrifolia* (coast live oak). Pr-102 has a genotype identical to most *P. ramorum* isolates from California (*2*).

The whole genome shotgun data used in the *P. sojae and P. ramorum* assemblies are summarized in Table S1. Paired end sequences from small insert plasmids (~2-4 kb), medium insert (~8 kb) plasmids and large insert (~36 kb) fosmids were generated as described (*3, 4*). Passed lanes are lanes which produce more than 100 bases at quality score of at least Q20 (*3*). Quality and vector trimming were performed as described (*3*). Raw shotgun reads are available for download at the NCBI Trace Archive http://www.ncbi.nlm.nih.gov/Traces/trace.cgi or at the JGI websites http://www.jgi.doe.gov/Psojae and http://www.jgi.doe.gov/Pramorum

Table S1 Shotgun Sequencing Statistics. Trimming includes the removal of nucleotides that are vector sequence and those that are of low quality.

	Total lanes run	Total untrimmed nucleotides	Total trimmed nucleotides	Mean trimmed lengths
P. sojae				
2-4 kb clones	704,125	609,329,668	450,224,087	639
8 kb clones	743,679	721,208,694	416,073,304	559
35 kb clones	105,984	125,162,278	66,541,960	628
Total	1,553,788	1,455,700,640	932,839,351	600
P. ramorum				
2-4 kb clones	398,310	425,954,247	268,536,760	674
8 kb clones	439,224	474,123,232	237,820,719	541
35 kb clones	67,872	78,585,909	46,105,985	679
Total	905,406	978,663,388	552,463,464	610

Genome assembly.

Reads passing the primary quality and vector screens ("passing reads") were assembled into scaffolds by means of JAZZ, a modular suite of tools for large shotgun assemblies that incorporates both read-overlap and read-pairing information (*3*). In the presence of allelic polymorphism, we accepted lower scoring read overlaps when they were corroborated by read

pair constraints. Details of the assembly method were similar to those for pufferfish (*3*) and the sea squirt *(5)*.

The *P. sojae* assembly includes 1,029,163 high quality sequencing reads assembled into 5,577 contiguously assembled segments (contigs) with a total length of 78.0 million base pairs (Mb). These are linked by paired-end constraints into 1,810 scaffolds spanning 86 Mb. The estimated genome size is 95 Mbp, The difference between the assembled sequence and the estimated genome size represents unassembled reads, largely due to unresolved repeats in the genome, and to the characteristics of heterochromatin. This is typical of the draft genome sequences of eukaryotes when using the whole genome shotgun approach. Half the assembly is in the largest 218 contigs, each of which is longer than 105.7 kb and in 54 scaffolds, each longer than 463 kb.

The *P. ramorum* assembly includes 502,201 high quality sequencing reads assembled into 7,588 contigs with a total length of 54.4 Mb. These are linked by paired-end constraints into 2,576 scaffolds spanning 66.6 Mb. Half the assembly is in the largest 277 contigs, each longer than 47.5 kb and in 63 scaffolds, each longer than 308 kb. The estimated genome size is 65 Mbp.

Table S2 Genome Assembly Statistics. The approximate number of nucleotides in the assemblies is estimated by multiplying the number of lanes incorporated by the average trimmed read length. Fold-coverage is calculated based on the size of the assembly.

Summary	Total lanes run	Total lanes passed	Overall fold-coverage	Lanes in assembly	Approx. # nts in assembly	Assembly fold-coverage
P. sojae	1,553,788	1,419,739	9.0	1,029,163	617,497,800	7.9
P. ramorum	905,406	815,983	7.7	502,201	306,342,610	5.6

Files containing unassembled reads in FASTA format can be downloaded from the JGI's web portal (http://genome.jgi-psf.org/sojae1/sojae1.download.ftp.html and http://genome.jgi-psf.org/ramorum1/ramorum1.download.ftp.html).

Physical Mapping

A total of 8,681 clones were assembled into 257 BAC contigs, of which 11 BAC contigs contained 100-200 clones, 47 contained 50-99 clones, 63 contained 35-49 clones, 87 contained 10-24 clones, and 49 contained 3-9 clones. A minimum tiling path consisting of 1,440 clones was subjected to BAC end sequencing. Alignment of the BAC contigs and the sequence scaffolds using the BAC end sequences resulted in a consensus physical map consisting of 60 "super-scaffolds" encompassing 207 sequence scaffolds, 238 BAC contigs, and a total of 73 Mb of the 86 Mb of assembled *P. sojae* DNA sequences (*6*).

BAC library construction.

Two libraries from *Phytophthora sojae* strain P6497 were constructed with BAC vectors pBeloBACII (*7*) and pECBAC1 as previously described (*8, 9*). High-molecular-weight DNA of *P. sojae* was partially digested with *Hin*d III (for pBeloBACII) or *Bam*HI (for pECBAC1) and subjected to size selections on a pulsed-field gel. DNA fragments between 100-300 kb were recovered, ligated to *Hin*d III or *Bam*H I digested, dephosphorylated pBeloBAC11 or pECBAC1,

and transformed into *E. coli* DH10B cells. The white transformant clones grown on selective medium containing chloramphenicol, IPTG, and X-gal were arrayed into 384-well microplates.

BAC fingerprinting

BAC DNA was isolated, transferred into 96-well microtiter plates, digested and labeled with a fingerprinting kit with some modifications (*10*). The fingerprinting kit contained six 6-bp endonucleases (*Bam*H I, *Bgl* II, *Xba* I, *Cla* I, *Hin*d III and *Xho* I) to generate a sufficient number of bands for the smaller-insert clones, plus one 4-bp endonuclease (*Hae* III), and the SnaPshot Multiplex Ready Reaction Mix (Applied Biosystems). The *Bam*H I and *Bgl* II fragment ends were labeled with ddGTP-dR110 (blue); the *Xba* I and *Cla* I ends with ddCTP-dTAMRA (yellow); the *Hin*d III ends with ddATP-dR6G (green); and the *Xho* I ends with ddTTP-dROX (red). *Hae* III was used to cut the large 6-bp enzyme fragments into smaller fragments so that they could be fractionated on a capillary sequencer with a range of 35 - 500 bp. The raw fingerprints in the window ranging from 35 to 500 bases were collected with the GeneScan V3.70 and the ABI 3100 data collection V1.0.1.

BAC Contig Assembly

The raw fingerprint data were edited and converted into the FingerPrinted Contig (FPC) band data as described (*10*). Fifteen datasets corresponding to the four individual colors and their possible 2-, 3- and 4-way combinations were prepared. BAC contigs were assembled from each dataset and analyzed (*10*) using the software FPC V 6.2 (*11, 12*). The dataset generated with *Xho* I, which had an average number of 35.3 bands per clone, was found to yield the largest contig assembly with fewest questionable clones and therefore, chosen for contig map construction. BAC contig editing, contig merging, and singleton addition were conducted as previously described (*9, 10*). The physical map has not yet been anchored to the *P. sojae* genetic map, which consists mostly of unsequenced RAPD and AFLP markers (*13*).

Gene Prediction

Using the Joint Genome Institute (JGI) genome annotation pipeline that includes several gene prediction and annotation methods (*3, 5*), we predicted and annotated 19,027 genes in the genome of *P. sojae* and 15,743 genes in the genome of *P. ramorum* (Table S3). The majority of gene models (75-80%) were predicted *ab initio* using the program FGENESH (*14*), trained for the genomes of *P. sojae* and *P. ramorum* using available EST sequences. In predicting exons, FGENESH achieved, respectively, 89% and 83% sensitivity (fraction of correctly detected true exons) and 88% and 85% specificity (fraction of true exons among all predicted exons). The remaining 20-25% of the models are homology-based, predicted using a combination of FGENESH+ (www.softberry.com) and Genewise (*15*), and synteny-based modeling using FGENESH2 (www.softberry.com). The latter was used to correct imperfect models of orthologous genes. 9,768 pairs of genes were identified as corresponding one-to-one between the two genomes using the criterion of reciprocal best BLASTp matches. 7,850 EST unigenes from *P. sojae* (*16*) have been mapped onto the genomic assembly of *P. sojae* and used for validation, correction and extension of predicted gene models. More than 90% of the unigenes were represented in gene models and more than 95% in the genome sequence.

Table S3. Support for predicted genes in the *P. sojae* and *P. ramorum* genome sequences.

Gene Model Information	***P. sojae***	***P. ramorum***
Total number of gene models	19,027	15,743
FGENESH (*ab initio*)	15,195	12,008
FGENESH+ (homology)	1,345	1,112
Genewise (homology)	1,089	1,264
FGENESH2 (synteny)	1,398	1,359
Complete models	17,291	13,538
Model support		
ESTs	7,088	N/A
genomic conservation	14,722	11,270
homology to known proteins	14,909	13,013
protein domain	11,733	9,982
proteomics	N/A	4,275

Proteomic analysis of *Phytophthora ramorum* proteins.

ESTs are not yet available for *P. ramorum*. Instead, to validate the *P. ramorum* gene predictions, we used Multidimensional Protein Identification Technology (MudPIT) (*17, 18*) to collect tandem mass spectra from tryptic fragments of proteins expressed in mycelium and germinating cysts. Of 51,464 peptides analyzed and matched to the gene model database, 78% fell within predicted gene calls, providing strong support for 3,150 of the 16,066 predicted gene calls. An additional 1,125 gene calls could be expanded based on matches to peptides inferred from open reading frames within 200 bp of a gene model. Finally, the presence of 279 new models was inferred based on clusters of at least three peptides located within 1,000 nucleotides of each other but more than 200 nucleotides from an existing gene model.

Freeze dried samples of mycelium, grown in clarified V8 juice broth (160 ml filtered V8 juice, 3 g $CaC0_3$, 840 mls water) or germinating cysts were disrupted with a bead beater, lysed, fractionated by centrifugation, and digested with trypsin (*17*). Peptides were separated and analyzed using Multidimensional Protein Identification Technology (MudPIT) as described previously using a ThermoFinnigan LTQ (*19*). Peptide tandem spectra were searched against a six frame translation protein database (minimum 35 amino acids stop-to-stop) using DBDigger (*20*) employing the MASPIC scorer. Resulting peptide identifications were sorted and filtered using DTASelect (*21*) and the peptides visualized with the predicted gene calls using the Artemis program (Sanger Institute).

Annotation of Gene Models

The predicted genes were electronically annotated and classified according to the Gene Ontology (*22*), Clusters of Eukaryotic Orthologous Groups (KOG clusters) (*23*), and Kyoto Encyclopedia of Gene and Genomes (KEGG) metabolic pathways (*24*). Enzyme Commission (E.C.) numbers have been assigned to 9,520 and 9,892 genes in *P. sojae* and *P. ramorum*, respectively, and 3,890 and 3,830 genes in *P. sojae* and *P. ramorum,* respectively, have KOG assignments. Manual curation of gene models and annotation was carried out at a one week workshop in

August 2004, and is ongoing at the community annotation web site at http://phytophthora.vbi.vt.edu (*25*). About 21% of the genes in both genomes (4,100/19,027 and 3,636/15,743 respectively) have similarity to known proteins at a BLASTx bit score of 150, while an additional 57% (10,981/19,027 and 8,977/15,743 respectively) contain matches to known protein motifs identified by InterProScan. 1,563 pairs of genes common to the two species show no recognizable homology to any species other than *Phytophthora*. Comparative analysis of annotations shows that gene counts and identities in various functional categories and pathways are very similar between these closely related organisms. Hybridization kinetics suggested that approximately 50% of the *P. sojae* genome is moderately repetitive (*26*). In concert with this, the genome contains numerous open reading frames with similarities both to retrotransposons as well as Mariner-like transposable elements that transpose via a DNA intermediate. There are also numerous large multigene families (see below), as well as simple sequence repeats useful for population genetics studies of particular importance to the study of *P. ramorum* (*27, 28*). Analysis and presentation of gene predictions for these two genomes are available from JGI Genome Portals (http://www.jgi.doe.gov/Psojae and http://www.jgi.doe.gov/Pramorum) and at the VBI Microbial Database (http://phytophthora.vbi.vt.edu).

Inferring orthology and gene colinearity (synteny) relationships.

DNA Sequence Similarity

To assess the extent of conservation between the *P.sojae* and *P.ramorum* genomes at the nucleotide level sequences were aligned using VISTA computational framework previously applied to large eukaryotic genomes (*29*). We utilized an efficient combination of global and local alignment approaches. The procedure includes running Shuffle-LAGAN glocal chaining algorithm (*30*) followed by Shuffle-LAGAN alignment of the intervals of conserved synteny to detect small-scale rearrangements.

Whole-genome alignment demonstrated a high level of similarity between the two species. 75.8% and 79.7% of the length of all *P. ramorum* and *P. sojae* predicted exons are covered by the alignment and about 97% of base pairs in these exon alignments belong to intervals with a high level of conservation (above 70%/100 bp). 68.3% of all aligned *P. ramorum* sequence and 65.4% of *P. sojae* sequence are conserved at the 70%/100bp level. Non-coding regions within these alignments have a high percentage of intervals (about 17%) highly conserved between the two species. These intervals could be coding regions not predicted by current techniques, incidental similarity due to recent divergence, or perhaps regulatory elements.

No evidence for large scale duplications was observed within either genome, though many examples of short regions of tandem duplication were observed within multigene families. The absence of large scale duplications in the *P. ramorum* genome argues strongly against the hypothesis that *P. ramorum* was formed by the recent hybridization of two other *Phytophthora* species (*31*).

The constructed genome-wide pairwise alignments can be downloaded from http://pipeline.lbl.gov/downloads.shtml and are accessible for browsing and various types of analysis through the VISTA browser at http://pipeline.lbl.gov/ linked to the VISTA portal page

http://genome.lbl.gov/vista and the related JGI portals.

Protein Sequence Similarity

Although a few of the analyses reported in this paper estimated orthology from reciprocal bi-directional best BLAST matches (principally in Table 1 of the main text), the well recognized limitations of this method led us to develop a truly phylogenetic approach to analyzing these data and presenting them to the scientific community. This was accomplished through the use of the PhIGs (Phylogenetically Inferred Groups; http://PhIGs.org) pipeline (*32*). In brief this consists of five steps: (A) an all-against-all BLASTP (*33*) analysis, considering all genes of both species of *Phytophthora* and many other completely sequenced genomes; (B) global alignment of the gene pairs using CLUSTALW (*34*) and distance calculation using the JTT matrix and the protdist program from PHYLIP (*35*); (C) iterative, hierarchical clustering of genes into gene families using a graph-based method that respects the evolutionary relationships among the organisms; (D) multiple sequence alignment for each cluster using the ClustalW (*34*) program; (E) creation of phylogenetic trees using the quartet puzzling, maximum likelihood method implemented in the TREE-PUZZLE (*36*) program using the JTT model of amino acid substitution and a gamma distribution of rates over eight rate categories with 10,000 puzzling steps to assess reliability. Each of these phylogenetic trees is available by searching on keywords or by similarity searching to Hidden Markov Models and the trees are built into the genome browsers so that each can be invoked by selecting the gene model using a web browser.

Colinearity maps between *P. sojae* and *P. ramorum* were created with the PhIGs synteny viewer, which is available from the PhIGs website (http://PhIGs.org). The viewer creates a view of the relative physical positions of orthologs across selected genomes. The maps are generated by selecting a genomic span from one species as the reference and the other species as the query. All identified orthologous genes between the selected genome and the query genome are then aligned according to their positions in their respective sequence scaffolds. The diagram in Fig. 2A of the main text was derived directly from the output of the PhIGs viewer.

SNP Identification.

Single nucleotide polymorphisms (SNPs) were identified from the raw genomic sequence data using the following four criteria: (i) bases must have PHRED quality scores >20, (ii) there must be 2 or more alternate bases at a site, represented by at least two sequence reads each (iii) there can be no more than 12 reads underlying the site, and (iv) there can be no additional SNPs within 100 bp up- or down-stream (to eliminate mis-assembled repeat sequences). For *P. ramorum* and *P. sojae*, 13,643 and 499 SNPs were predicted, respectively. At sites where SNPs exist in each species, the consensus sequence reported is based on the more common haplotype represented by high quality nucleotides in the shotgun reads.

Identifying *Phytophthora* genes with a potential photosynthetic ancestry.

Two complementary approaches were used to identify genes that potentially originated from a photosynthetic endosymbionts. Both approaches looked for genes with unusually high matches to sequences from a red alga, which is considered the most likely origin of the endosymbiont of

stramenopile algae, or to sequences from a cyanobacterium, which is the likely origin of the chloroplast genome of both red algae and of green algae and multicellular plants.

The first approach used portions of the "PhIGs" pipeline (http://PhIGs.org) developed for whole genome analysis (*32*) to find BLAST matches to cyanobacterial and red algal sequences. Cyanobacterial matches were defined as the subset of orthologous genes shared by the *Phytophthora* spp. and the diatom, *Thalassiosira pseudonana* which had stronger BLAST matches to a cyanobacterial gene than to any gene from Archaea, Eubacteria (minus cyanobacteria) or Opisthokonts. Red algal matches were defined as the subset of orthologous genes shared by *Phytophthora*, the diatom genes and the nucleomorph of the cryptophyte *Guillardia theta* (*37*) that have stronger BLAST matches to the genome of the red alga, *Cyanidioschyzon merolae* (*38*), than to green plants, Opisthokonts, Archaea, or Eubacteria. For each set of genes identified in this screen, we created a phylogenetic tree using a maximum likelihood approach as implemented in TREE-PUZZLE (*36*) to evaluate the likelihood of an endosymbiont origin. Genes of putative cyanobacterial origin were identified as the subset of genes in the chromalveolates that appear as the sister to a cyanobacterial gene on the tree. The search for genes that might have been transferred from the nucleus of the red alga that had been engulfed in the secondary endosymbiosis was done similarly. In this case, the comparisons were to the genomes of the nucleomorph of the cryptophyte *Guillardia theta* (*37*), the red alga *Cyanidioschyzon merolae* (*38*), and to many plants, fungi, animals, Archaea, and Eubacteria.

The second approach used normalized Smith-Waterman alignment scores (detailed below) to identify candidate endosymbiont genes whose similarity to genes of red or green plants was statistically significantly greater than to genes of organisms with no known photosynthetic ancestry (opisthokonts and amebozoa). This approach considered not only matches to cyanobacteria, plastid genomes and red algae, but also matches to green plant genes. The rationale for inclusion of green plant genes is that the only complete genome sequence available for a red alga is from *Cyanidioschyzon merolae* (*38*) that is an extremophile with a very streamlined genome. The putative red algal secondary endosymbiosis responsible for the chromalveolates has been estimated to have occurred about 1,300 million years ago, only 200 million years after the split of the red and green algae (*39*). Therefore we hypothesized that in some cases, a green plant sequence might be less diverged from the version of the endosymbiont sequence found in the oömycetes than from the *C. merolae* sequence, or might have been retained when the *C. merolae* sequence had been lost. Smith-Waterman alignment scores were obtained using a TimeLogic DeCypher system with the BLOSUM62 scoring matrix, gap opening penalty = 11, and gap extension penalty = 1. Every *Phytophthora* gene was used as a query in a Smith-Waterman search against a set of fully sequenced "donor" genomes might have genes matching the endosymbiont and against a set of fully sequenced "control" genomes which have no known photosynthetic ancestry. The "donor" genomes were nuclear genomes of *Cyanidioschyzon merolae,* and the green plants *Arabidopsis thaliana, Oryza sativa,* and *Chlamydomonas reinhardtii* plus chloroplast and plastid genomes of the green plants *Arabidopsis thaliana*, *Chlamydomonas reinhardtii*, *Nicotiana tabacum* and *Oryza sativa*, the red algae *Cyanidioschyzon merolae*, *Cyanidium caldarium, Porphyra purpurea* and *Gracilaria tenuistipitata*, the glaucophyte *Cyanophora paradoxa*, the euglenoids *Euglena gracilis* and *Euglena longa*, the cryptophyte *Guillardia theta*, the diatom *Odontella sinensis* and the apicomplexan *Eimeria tenella*. The "control" genomes were the animals *Homo sapiens, Mus*

musculus, *Drosophila melanogaster*, and *Caenorhabditis elegans*, the fungi *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Magnaporthe grisea, Neurospora crassa*, *Aspergillus fumigatus*, and *Ustilago maydis*, and the amebozoa *Dictyostelium discoideum* and *Entamoeba histolytica*. The evolutionary relationships of many of these organisms are summarized in Figure S1. To correct for the varying evolutionary distances of the donor and control genomes from *Phytophthora*, we standardized the alignment scores as follows. For every *Phytophthora* gene, the best alignment score from among the four donor genomes was chosen. Then for each of the 12 control genomes, the ratio (best donor score/ best score for that control genome) was calculated, resulting in 12 distributions of score ratios. Each of the ratio distributions was normalized by log-transforming the ratios, subtracting the mean of the log transformed ratios and dividing by the standard deviation, creating 12 sets of z scores. Next, for every *Phytophthora* gene, the *minimum* z-score of the 12 was selected, a conservative approach which identifies the narrowest gap between a donor match and a control match. A second z score transformation was then performed, to normalize the distribution of the minimum z scores. To determine what z-score should be considered biologically significant, the whole procedure was repeated using the *Dictyostelium discoideum* genome as the subject of the search (eliminating it from the control list) in order to estimate the range of z-scores to be found in an organism of no photosynthetic ancestry.

Figure S2 shows the number of genes of putative cyanobacterial or red algal origin in the diatom *Thalassiosira pseudonana* and in the two *Phytophthora* species. Thirty of the most convincing candidates obtained from the BLAST and Smith-Waterman searches are shown in Table S4. In selecting these genes, we placed particular emphasis on well-conserved genes that contain clear orthologs among the opisthokonts or amebozoa, and which have clearly defined functional annotations, in order to rule out artifacts caused by gene loss among the opishthokont and amebozoan lineages. Absence from the *T. pseudonana* genome did not disqualify a candidate, as the sequence may have been missed in the draft sequence or the gene may have been lost from that genome. Fifteen of the genes encoded proteins with a predicted mitochondrial location in *P. sojae* and *P. ramorum*, and of these fifteen the matching plant and/or algal proteins had predicted chloroplast location. A more extensive listing of candidates and their evaluation will be published elsewhere.

Although *Phytophthora* species synthesize lysine via the di-amino-pimelate pathway found in plants and bacteria, rather than the amino-adipate pathway found in fungi, we don't consider the lysine biosynthetic enzymes to have a likely phototroph origin because the di-amino-pimelate pathway is also found in the amebozoan *Dictyostelium discoideum*, and because the best matches to bacterial enzymes lie in the firmicutes and actinobacteria rather than the cyanobacteria. Neither *Phytophthora* species has the gene duplication of glyceraldehyde-3-phosphate dehydrogenase C isozyme that was reported to be common to the chromalveolates (*40, 41*), but this could be due to gene loss in the *Phytophthora* lineage or the gene could be missing from the two draft sequences.

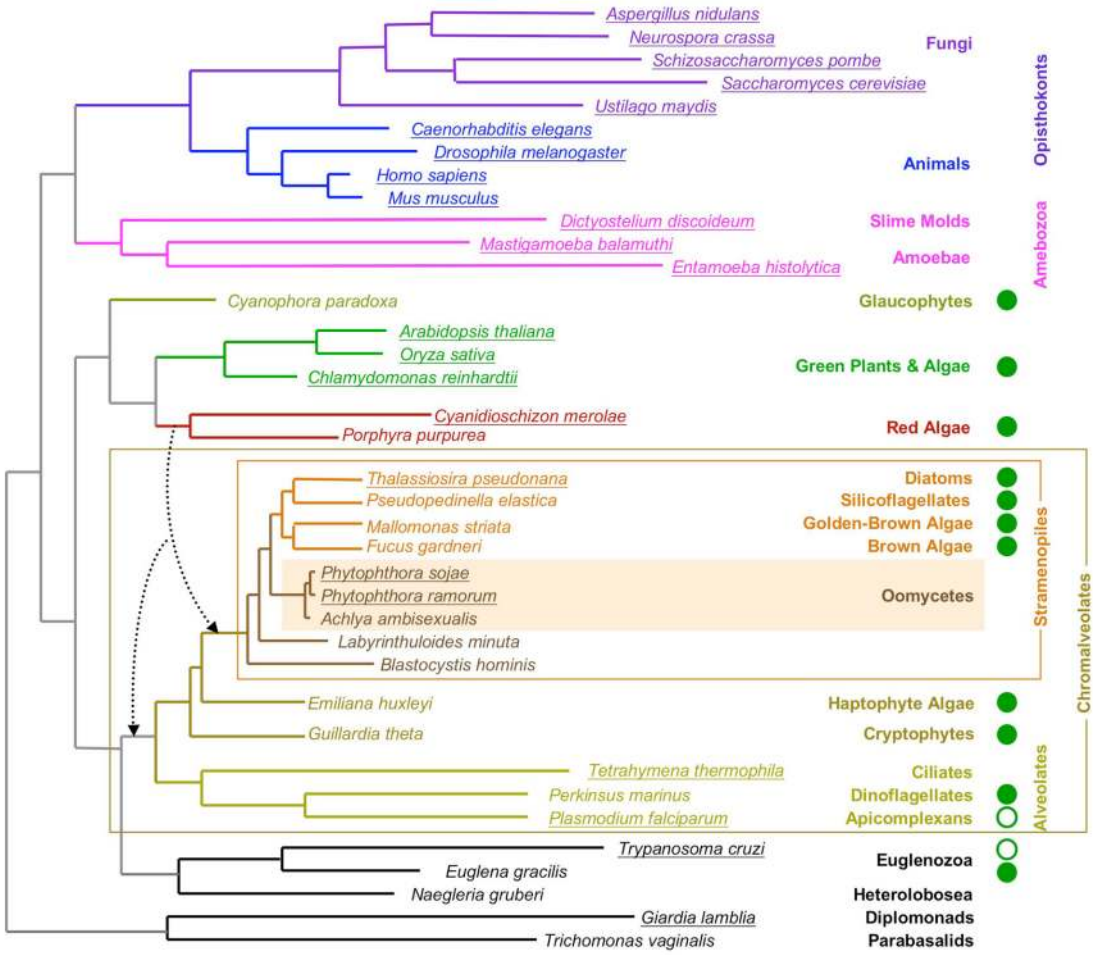

Fig. S1. Schematic phylogenetic tree of the eukaryotes. The tree is adapted from that of Baldauf *et al.* (*42*) that is based on a concatenation of six highly conserved proteins. Several species (mostly stramenopiles) were added to the tree by reference to the 18S rRNA trees of Sogin and Silberman (*43*). Complete genome sequences are available for the underlined species. Filled green circles on the right indicate photosynthetic species. Open green circles indicate species with vestigial plastids of photosynthetic origin. The dotted arrows indicate hypothetical events in which an ancient red algal endosymbiont might have been acquired by an ancestor of the chromalveolates (left arrow) or of the stramenopiles alone (right arrow).

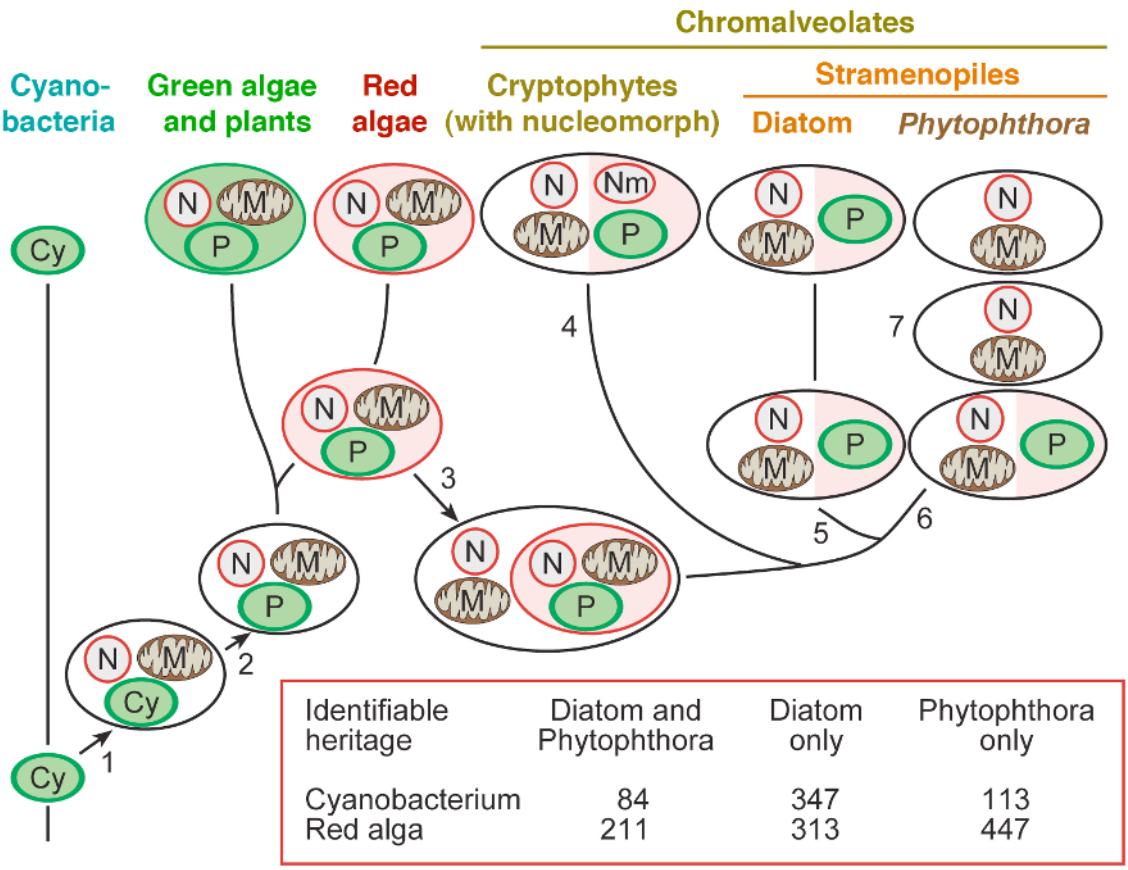

Fig. S2. Genes putatively transferred from photosynthetic endosymbionts. The chromalveolates have a complex pattern of symbioses, providing opportunities for gene transfer between multiple intracellular compartments. Events are reconstructed on this evolutionary tree by numeral: (1) A cyanobacterium becomes an endosymbiont of an early eukaryote, then (2) adapts to become a plastid, forming a lineage that gives rise to green algae and plants, glaucophytes, and red algae. (3) A red alga is engulfed by the ancestor of the chromalveolates, creating a cell with five intracellular compartments, the nucleus and mitochondria of the chromalveolate, plus the nucleus, mitochondria, and plastids of the red algae, followed by the loss of the red algal mitochondria (although this is poorly understood). (4) The degenerated red alga nucleus is retained in some lineages and is termed a nucleomorph. (5) The nucleomorph is lost in the diatom lineage. (6) The nucleomorph is separately lost in the *Phytophthora* lineage. (The inference that these losses were separate is based on each lineage having many unique gene transfers, although the alternative is possible that these occurred in the common ancestor, followed by very large amounts of gene loss.) (7) The plastid is lost in the *Phytophthora* lineage. (The order of events 6 and 7 are uncertain.) Genes that were transferred to the nuclear genomes of the diatom and the *Phytophthora* independently and in their common ancestor are presented in the lower right corner. Intracellular compartments are indicated by letters: Cy, cyanobacterium; N, nucleus; M, mitochondria; P, plastid; and Nm, nucleomorph.

Table S4. Examples of *P. sojae* and *P. ramorum* genes potentially originating from a photosynthetic endosymbiont

.

					Best BLASTp Score						
***P. sojae* GeneID**	***P. ramorum* GeneID**	**Annotation**	**Pathway**	**Tar-get**	**Plastid**	**Cyano-bacteria**	**Red alga**	**Green plant**	**Diatom**	**Opith/A meb**	**Z score**
Cyanobacteria top match											
108148	72019	Cobalamin-independent methionine synthase II	methionine	cyt	none	**0E+00**	1E-156	2E-164	none	3E-162	0.2
108389	54177	prolyl oligopeptidase II	unknown	cyt	none	**7E-165**	9E-01	1E-152	2E-150	6E-70	2.7
108956	75281	2-Isopropylmalate synthase	leucine	mito^{p}	none	**4E-153**	1E-128	1E-147	1E-127	3E-33	3.7
109497	54068	threonine dehydratase	leu, ile, val	mito^{p}	none	**1E-142**	1E-134	6E-122	8E-140	2E-137	0.2
123952	79142	anthranilate synthase	tryptophan	mito*^{p}	2E-28	**5E-137**	6E-91	6E-114	7E-109	6E-84	0.8
108458	38584	NCAIR mutase	purine	mito^{p}	none	**1E-58**	none	2E-28	9E-46	3E-03	3.4
116252	74880	Phosphoadenosine phosphosulfate reductase	methionine, cysteine	cyt	none	**3E-53**	4E-08	7E-16	5E-09	2E-23	2.5
109158	51635	Uroporphyrin III methyltransferase	porphyrin	mito*^{p}	2E-31	**4E-50**	1E-27	4E-34	9E-16	1E-25	0.4
156701	95818	tRNA (guanine-N(7)-)-methyltransferase-like	tRNA	mito^{p}	none	**1E-36**	7E-26	1E-28	6E-22	9E-14	1.6
156385*	80275	similar to Phosphatidate cytidylyltransferase (8 family members in *P. sojae*)	phospholipid	cyt	none	**9E-23**	4E-14	1E-13	6E-13	3E-04	>4
Red alga top match											
136278	87801	Ketol-acid reductoisomerase	leu, ile, vla	mito^{p}	none	3E-24	**1E-163**	1E-158	6E-166	2E-38	3.6
142774	80380	Phosphoserine aminotransferase	serine	mito^{p}	none	3E-100	**1E-134**	none	7E-62	4E-04	>4
108405	72085	asparaginyl tRNA synthetase	tRNA	mito^{p}	none	3E-123	**1E-126**	7E-121	5E-123	1E-106	0.7
109393	75838	SAICAR synthetase	purine	mito*^{p}	none	1E-06	**1E-101**	7E-99	1E-92	5E-43	1.6
119553	72293	glucokinase	glucose catab.	cyt	none	3E-59	**7E-67**	2E+00	2E-64	9E-01	3.0
135234	75742	Histidinol-phosphate/aromatic aminotransferase.	histidine	cyt	none	7E-12	**2E-62**	4E-06	none	5E-04	>4
133425	78949	zinc carboxypeptidase A	unknown	cyt	none	none	**5E-43**	none	9E-61	7E-05	2.8
132772	79657	cAMP-binding mitochondrial solute carrier	unknown	cyt	none	none	**1E-63**	none*	none*	none*	2.8
137179	45002	acyl-carrier-protein reductase	unknown	mito^{p}	none	2E-48	**5E-49**	3E-28	8E-24	5E-16	2.7
137240	77863	similar to sulfur transferase + methyl transferase fusion	unknown	mito	none	*2E-30*	**8E-65**	*3E-23*	4E-46	*1E-33*	2.5
142125	86425	probable nucleoside phosphorylase	unknown	cyt	none	3E-12	**6E-36**	none	none	9E-10	2.8
155781	54215	Ribonuclease HII	unknown	mito^{m}	none	2.E-29	**1E-37**	2E-08	2E-33	7E-13	2.5
Green plant top match											
140563	71442	Nitrate reductase	nitrate util.	cyt	none	1E-04	1E-172	**0E+00**	1E-128	2E-157	0.6
108585	71783	6-phospohogluconate dehydrogenase	pentose phosphate	cyt	none	6E-149	3E-136	**9E-179**	0E+00	2E-126	1.0
155429	83828	aspartate kinase+homoserine dehydrogenase fusion	Lysine, glycine, serine,threonine	cyt	none	6E-25	1E-122	**1E-135**	7E-47	7E-42	3.2
112240	73217	Galactolactone oxidase	ascorbate	mito^{m}	none	1E-24	1E-110	**8E-132**	2E-84	9E-43	3.2
137005	85610	Cobalamin synthesis protein.	cobalamin	cyt	none	6E-79	1E-70	**1E-80**	1E-31	7E-51	2.1
127943	78464	tRNA dihydrouridine synthase	tRNA	cyt	none	6E-62	1E-64	**1E-70**	1E-39	3E-31	3.1
109065	72218	major facilitator superfamily (53 family members in *P. sojae*)	unknown	cyt	none	1E-47	2E-42	**2E-48**	6E-13	8E+00	>4
128553	82990	Prephenate dehydratase family	phenylalanine	mito*^{p}	none	1E-08	4E-28	**3E-30**	2E-17	2E-13	1.7

Gene IDs are from the JGI and VBI databases. Intracellular location of the gene's product was predicted from TargetP (*44*) analysis of the *P. sojae* protein, in some cases (marked with an asterisk) utilizing start codons differing from those predicted by the gene model; in this column, the superscript ^{p} indicates that the cellular location of the matching protein in plants and/or algae was predicted by TargetP to be either the plastid or the plastid and mitochondria, the superscript ^{m} indicates a predicted mitochondrial location only for the matching protein(s). BLASTp analysis was carried out using TimeLogic DeCypher sequence comparison accelerators. Z-scores refer to normalized ratios of Smith-Waterman alignment scores between the best match to a photosynthetic organism and the best match to a non-photosynthetic organism (see text of the Supplementary Information). Opith/Ameb refers to the best match to an opisthokont or amoebazoan.

Identification of the Avh gene superfamily.

The members of the Avh superfamily were identified by recursive tBLASTn searches and Hidden Markov Model (HMM) searches. The tBLASTn searches initially used as queries the sequences of the four *Avr1b*-1 alleles (*45*) (GenBank accession numbers AF449622, AF449621, AF449624 and AF449625), the sequence of the *Avr1b*-1 paralog *Avh1b*-1 (now renamed *Avh*1; AF449626) and the *Avr3a* gene from *P. infestans* (*46*) (CAI72254.1). All *P. ramorum* and *P. sojae* gene models constructed by the JGI were searched using the PAM70 substitution matrix, an expectation limit of 10 and no filtering of low complexity sequences. The hits obtained were manually curated for the presence of a signal peptide, using the SignalP algorithm (*47*), and the presence of a typical *Phytophthora* codon usage (*48-50*). Very short ORFs, ORFs with very low complexity sequences, and ORFs with matches only in the secretory leader were eliminated. Incorrect gene models (usually abnormal introns or fusions to neighboring genes) were also corrected. The remaining ORFs were designated Avh genes and used as queries for fresh searches. This process was repeated until no new Avh genes were recovered, at which point approximately 170 Avh genes had been identified in *P. sojae*. The RXLR and dEER motifs (*51, 52*) of the 170 *P. sojae* Avh genes were then aligned and used to construct a hidden Markov model of the region surrounding the two motifs, using the software HMMer (*53*). The HMM model was used to search the six-frame translation of both genome sequences, yielding another 80 Avh genes after manual curation. The set of 250 Avh genes was then used to query the entire genome sequences of the two species by tBLASTn (*33*), using an automated script. After careful manual review of all candidates, a total of 350 Avh genes were identified in each species. The precise number of Avh genes is somewhat uncertain. Many obvious pseudogenes were found, containing high quality secretory leaders and RXLR-dEER motifs, but with stop codons and/or frameshifts interspersed. In some cases however, it was uncertain whether an Avh candidate was a pseudogene, for example if the encoded protein seemed very short. In some cases, Avh genes were identified as such despite the presence of a single frameshift mutation, on the basis that the frameshift could be due to a sequencing error. A complete listing of the Avh genes and their characterization will be reported elsewhere. Fig. S3 shows the positions and consensus sequences of the RXLR and dEER motifs and the sequences surrounding them in the Avh genes.

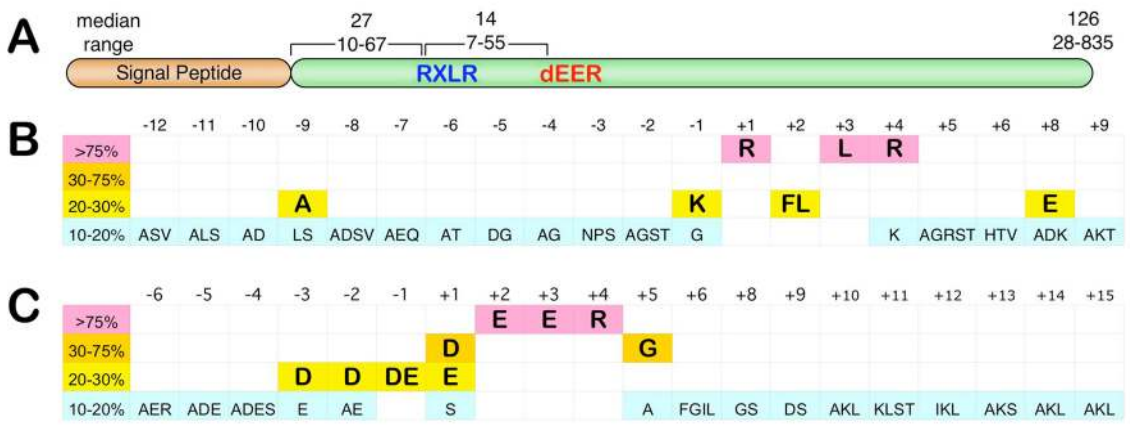

Fig. S3. Characterization of a superfamily of 700 *P. sojae* and *P. ramorum* Avh genes related to oömycete avirulence genes. (**A**) Summary of the structure of the genes. Numbers indicate amino acid residues. None of the encoded proteins contain di-sulfide bonds. (**B**) and (**C**) consensus sequences of RXLR and dEER amino acid motifs, respectively, and the regions immediately surrounding them. Percentages are for the individual amino acids noted.

SUPPORTING TEXT

Absence of genes encoding secondary metabolite toxins

Fungal plant pathogens, distantly related evolutionarily, but similar in some traits, utilize secondary metabolites as toxins, most notably polyketides and non-ribosomal peptides, and the pathogens' genomes frequently contain 20-30 sets of biosynthetic genes for each type of metabolite (*54, 55*). In contrast, we were unable to identify any polyketide synthase genes whatsoever in *P. sojae* or *P. ramorum* and only four pairs of orthologous non-ribosomal peptide synthetase genes. Randall et al (*48*) also failed to find polyketide synthase sequences in a *P. infestans* EST collection.

Further Analysis of the Genome Sequences

Further analysis of the genome sequences will be published elsewhere, describing extensive variation in nuclear mitochondrial DNA content between the genomes of *P. sojae* and *P. ramorum* (56), targeted gene mutation in *Phytophthora* (57), genome wide analysis of phospholipid signalling genes in *Phytophthora* (58), the repertoire of transfer RNA genes and codon usage bias in the genomes of *P. sojae* and *P. ramorum* (50), comparative analysis of *Phytophthora* genes encoding secreted proteins (59), identification of cell wall-associated proteins from *P. ramorum* (60), an integrated BAC and Genome Sequence Physical Map of *P. sojae* (6) and a functional screen to characterize the secretomes of eukaryotic pathogens and their hosts in planta (61).

ADDITIONAL SUPPORTING TABLES

Table S5. Gene IDs for sequences used for tree building in Fig. 1B and 1C of the main text. Numbers prefaced by "GeneID" are from the DOE JGI Genome portal at http://genome.jgi-psf.org. All others are GenBank accession numbers.

Gene family	Organism	GenBank or JGI Gene ID
2-isopropylmalate synthase		
	Arabidopsis thaliana	AAG52882.1
	Cyanidioschyzon merolae	CMQ337C
	Helicosporidium sp	AAU93936.1
	Methanocaldococcus jannaschii	AAB99199.1
	Neurospora crassa	CAE76195.1
	Nostoc sp. PCC 7120	BAB76539.1
	Phytophthora ramorum	GeneID 75281
	Phytophthora sojae	GeneID 108956
	Pseudomonas aeruginosa	AAG07179.1
	Saccharomyces cerevisiae	CAA90522.1
	Schizosaccharomyces pombe	O59736
	Thalassiosira pseudonana	GeneID 139551
	Thermoplasma volcanium	BAB60074.1
	Ustilago maydis	XP_760303.1
NCAIR mutase		
	Arabidopsis thaliana AIR carboxylase	NP_181305.2
	Chlamydomonas reinhardtii AIR carboxylase	GeneID Chlre3l77990
	Chlamydomonas reinhardtii NCAIR mutase	GeneID Chlre3l72789
	Cyanidioschyzon merolae AIR carboxylase	CME023C
	Drosophila melanogaster AIR carboxylase (*Ade*7)	NP_572826.1
	Methanopyrus kandleri NCAIR mutase	AAM01890.1
	Nostoc sp. PCC 7120 NCAIR mutase (*cpm*A)	BAB75584.1
	Phytophthora ramorum NCAIR mutase	GeneID 38584
	Phytophthora ramorum AIR carboxylase	GeneID 41522
	Phytophthora sojae NCAIR mutase	GeneID 108458
	Phytophthora sojae AIR carboxylase	GeneID 116989
	Saccharomyces cerevisiae AIR carboxylase	CAA99327.1
	Thalassiosira pseudonana AIR carboxylase	GeneID 102418
	Thalassiosira pseudonana NCAIR mutase	GeneID 118813
	Trichodesmium erythraeum NCAIR mutase	ZP_00673723.1
	Trichodesmium erythraeum AIRC carboxylase	ZP_00674123.1

SUPPORTING REFERENCES

1. H. Förster, B. M. Tyler, M. D. Coffey, Mol. Plant-Microbe Interact. 7, 780 (1994).
2. K. L. Ivors, K. J. Hayden, P. J. M. Bonants, D. M. Rizzo, M. Garbelotto, Mycol. Res. 108, 378–392 (2004).
3. S. Aparicio *et al.*, Science 297, 1301 (2002).
4. J. C. Detter *et al.*, Genomics 80, 691 (2002).
5. P. Dehal *et al.*, Science 298, 2157 (2002).
6. X. Zhang *et al.*, Mol. Plant-Microbe Interact. in press (Dec 2006).

7. H. Shizuya *et al*., Proc Natl Acad Sci U S A 89, 8794 (1992).
8. H.-B. Zhang, Construction and manipulation of large-insert bacterial clone libraries: Manual (Texas A&M University, College Station, Texas, 2000), pp.
9. C. Ren *et al*., in Handbook of Plant Genome Mapping: Genetic and Physical Mapping K. Meksem, G. Kahl, Eds. (Wiley-VCH Verlag GmbH, Weinheim, Germany, 2005) pp. 173-213.
10. Z. Xu *et al*., Nucleic Acids Res 33, e50 (2005).
11. V. Pampanwar *et al*., Plant Physiol 138, 116 (2005).
12. C. Soderlund, S. Humphray, A. Dunham, L. French, Genome Res 10, 1772 (2000).
13. K. J. May *et al*., Fungal Genet. Biol. 37, 1 (2002).
14. A. A. Salamov, V. V. Solovyev, Genome Res 10, 516 (2000).
15. E. Birney, M. Clamp, R. Durbin, Genome Res 14, 988 (2004).
16. T. Torto-Alalibo *et al*., Mol. Plant-Microbe Interact., in press (Dec 2006).
17. M. P. Washburn, D. Wolters, J. R. Yates, 3rd, Nat Biotechnol 19, 242 (2001).
18. D. A. Wolters, M. P. Washburn, J. R. Yates, 3rd, Anal Chem 73, 5683 (2001).
19. W. H. McDonald, R. Ohi, D. Miyamoto, T. J. Mitchison, J. R. I. Yates, Int J Mass Spectrometry 219, 245 (2002).
20. D. L. Tabb, C. Narasimhan, M. B. Strader, R. L. Hettich, Anal Chem 77, 2464 (2005).
21. D. L. Tabb, W. H. McDonald, J. R. Yates, 3rd, J Proteome Res 1, 21 (2002).
22. The Gene Ontology Consortium, Genome Res 11, 1425 (2001).
23. E. V. Koonin *et al*., Genome Biol. 5, R7 (2004).
24. M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, M. Hattori, Nucl. Acids Res. 32, D277 (2004).
25. S. Tripathy, V. N. Pandey, B. Fang, F. Salas, B. M. Tyler, Nucl. Acids Res. 34, D379–D381 (2006).
26. Y. Mao, B. M. Tyler, Exp. Mycol. 15, 283 (1991).
27. K. Ivors *et al*., Molecular Ecology (2006).
28. N. J. Grünwald, S. Tripathy, K. Ivors, K. H. Lamour, in preparation (2006).
29. K. A. Frazer, L. Pachter, A. Poliakov, E. M. Rubin, I. Dubchak, Nucleic Acids Research 32, W273 (2004).
30. M. Brudno *et al*., Bioinformatics 19, 54i (2003).
31. D. M. Rizzo, M. Garbelotto, E. M. Hansen, Annu Rev Phytopathol 43, 309 (2005).
32. P. S. Dehal, J. L. Boore, BMC Bioinformatics, 7:201 (2006).
33. S. F. Altschul *et al*., Nucleic Acids Res 25, 3389 (1997).
34. J. D. Thompson, D. G. Higgins, T. J. Gibson, Nucleic Acids Res 22, 4673 (1994).
35. J. Felsenstein. (Department of Genome Sciences, University of Washington, Seattle, 2004).
36. H. A. Schmidt, K. Strimmer, M. Vingron, A. von Haeseler, Bioinformatics 18, 502 (2002).
37. S. Douglas *et al*., Nature 410, 1091 (2001).
38. M. Matsuzaki *et al*., Nature 428, 653 (2004).
39. H. S. Yoon, J. D. Hackett, C. Ciniglia, G. Pinto, D. Bhattacharya, Mol Biol Evol 21, 809 (2004).
40. N. M. Fast, J. C. Kissinger, D. S. Roos, P. J. Keeling, Mol Biol Evol 18, 418 (2001).
41. J. T. Harper, P. J. Keeling, Mol Biol Evol 20, 1730 (2003).
42. S. L. Baldauf, A. J. Roger, I. Wenk-Siefert, W. F. Doolittle, Science 290, 972 (2000).

43. M. L. Sogin, J. D. Silberman, Int. J. Parasitol. 28, 11 (1998).
44. O. Emanuelsson, H. Nielsen, S. Brunak, G. v. Heijne, J. Mol. Biol. 300, 1005 (2000).
45. W. Shan, M. Cao, D. Leung, B. M. Tyler, Mol. Plant Microbe Interact 17, 394 (2004).
46. M. R. Armstrong *et al*., Proc Natl Acad Sci U S A 102, 7766 (2005).
47. J. D. Bendtsen, H. Nielsen, G. v. Heijne, S. Brunak, J. Mol. Biol. 340, 783 (2004).
48. T. A. Randall *et al*., Mol Plant Microbe Interact 18, 229 (2005).
49. R. H. Y. Jiang, F. Govers, J. Molec. Evol., in press (2006).
50. S. Tripathy, B. M. Tyler, Mol. Plant-Microbe Interact., in press (Dec 2006).
51. P. R. Birch, A. P. Rehmany, L. Pritchard, S. Kamoun, J. L. Beynon, Trends Microbiol 14, 8 (2006).
52. A. P. Rehmany *et al*., The Plant Cell, 17(6), 1839–1850 (2005).
53. S. Eddy. (Center for Genome Sciences, Washington University School of Medicine, St. Louis, MO 63108, 2003).
54. T. J. Wolpert, L. D. Dunkle, L. M. Ciuffetti, Annu. Rev. Phytopathol. 40, 251–285 (2002).
55. R. Dean *et al*., Nature 434, 980 (2005).
56. K. Krampis, B.M. Tyler and J. Boore, Mol. Plant-Microbe Interact., in press (Dec 2006).
57. K. H. Lamour, L. Finley, O. Hurtado-Gonzales, D. Gobena, M. Tierney, and H.J.G. Meijer, Mol. Plant-Microbe Interact., in press (Dec 2006).
58. H.J.G. Meijer, F. Govers, Mol. Plant-Microbe Interact., in press (Dec 2006).
59. R. H.Y. Jiang, B.M. Tyler, F. Govers Mol. Plant-Microbe Interact., in press (Dec 2006).
60. H. J. G. Meijer, P. J. I. van de Vondervoort, Q.Y. Yin, C.G. de Koster, F. Govers, P. W. J. de Groot, Mol. Plant-Microbe Interact., in press (Dec 2006).
61. S.-J. Lee, B.S. Kelley, C.M.B. Damasceno, B. St. John, B.-S. Kim, B.-D. Kim, J.K.C. Rose Mol. Plant-Microbe Interact., in press (Dec 2006).