Genome sequences of two Phytophthora species responsible for Sudden Oak Death and Soybean Root Rot provide novel insights into their evolutionary origins and mechanisms of pathogenesis

Brett M. Tyler (Virginia Bioinformatics Institute), Sucheta Tripathi (Virginia Bioinformatics Institute), Andrea Aerts (JGI), Douda Bensasson (University of Manchester), Paramvir Dehal (JGI), Inna Dubchak (JGI), Matteo Garbelotto (UC Berkeley), Mark Gijzen (Agri-food Canada), Wayne Huang (University of Ohio), Kelly Ivors (North Carolina State University), Rays Jiang (Broad Institute, MIT, Harvard), Sophien Kamoun (Sainsbury Laboratory), Konstantinos Krampis (Virginia Bioinformatics Institute), Kurt Lamour (University of Tennessee), Hayes McDonald (Vanderbilt-Ingram Cancer Center), Monica Medina (UC Merced), Paul Morris (Queen's University), Nik Putnam (JGI), Sam Rash (UC Davis), Asaf Salamov (JGI), Brian Smith (LBNL), Joe Smith (LBNL), Astrid Terry (JGI), Trudy Torto (LBNL), Igor Grigoriev (JGI), Daniel Rokhsar (JGI), Jeffrey Boore (JGI)

Introduction

The approximately 60 species of *Phytophthora* are all destructive pathogens, causing rots of roots, stems, leaves and fruits of a wide range of agriculturally and ornamentally important plants (1). Some species, such as *P. cinnamomi*, *P. parasitica* and *P. cactorum*, each attack hundreds of different plant host species, whereas others are more restricted. Some of the crops where *Phytophthora* infections cause the greatest financial losses include potato, soybean, tomato, alfalfa, tobacco, peppers, cucurbits, pineapple, strawberry, raspberry and a wide range of perennial tree crops, especially citrus, avocado, almonds, walnuts, apples and cocoa, and they also heavily affect the ornamental, nursery and forestry industries. The economic damage overall to crops in the United States by *Phytophthora* species is estimated in the tens of billions of dollars, including the costs of control measures, and worldwide it is many times this amount (1). In the northern midwest of the U.S., *P. sojae* causes $200 million in annual losses to soybean alone, and worldwide causes around $1-2 billion in losses per year. *P. infestans* infections resulted in the Irish potato famine last century and continues to be a difficult and worsening problem for potato and tomato growers worldwide, with worldwide costs estimated at $5 billion per year.

Phytophthora is part of the group oömycetes, which contains many other destructive plant pathogens, in particular the downy mildews and more than 100 species of the genus *Pythium.* These organisms at least double the losses due to *Phytophthora*.

Because oömycetes, including *Phytophthora*, are evolutionarily very distant from "true" fungi (see below), these organisms are very difficult to control because they are unaffected by the majority of fungicides. The difficulty of control is worsened for most *Phytophthora* diseases because the affected plants are attacked underground where they cannot be economically treated. The pending phaseout of soil fumigants such as methyl bromide further exacerbates this problem. A final complication is that many oömycetes appear to have an extraordinary genetic flexibility which enables them to rapidly adapt to and overcome chemical control measures and genetic resistance bred into plant hosts.

In addition to their impacts on agriculture, *Phytophthora* species are highly destructive of forests. *Phytophthora* species currently cause severe damage to both trees and understory vegetation in Australian eucalypt forests (2-4) and to hardwood forests in Europe (primarily to alder, oak and chestnut) (4). In the US, *P. cinnamomi* completely eliminated chestnut trees from the southern Appalachians 50 years before the arrival of chestnut blight, which destroyed the northern chestnut forests (5). On the west coast of the U.S., *Phytophthora* damages Douglas fir forests and threatens the elimination of natural stands of Port Orford Cedar (5).

In California and Oregon, a disease called Sudden Oak Death Syndrome, caused by a new, virulent species, *P. ramorum*, has been destroying coastal oaks since 1996. *P. ramorum* can infect not only coast live oak, California black oak, shreve oak and tan oak, but also a large variety of shrubs that inhabit the oak ecosystems including bay laurel, rhododendron, huckleberry, buckeye, madrone, manzanita, California coffeeberry and toyon (cemarin.ucdavis.edu/index2.html). Destruction of the California live oaks threatens entire

coastal forest ecosystems as this is the keystone species in these ecosystems. Several lines of evidence suggest this pathogen may be introduced in the US. Laboratory tests also reveal that Northern Red Oak and Pin Oak, which are distributed extensively throughout the midwest, northeast and Canada, are highly susceptible to this pathogen. To date the total damage caused by *P. ramorum* stands at around $200 million, including damage to forestry, recreation and tourism.

In some ecosystems, broad host range *Phytophthora* pathogens such as *P. cinnamomi* and *P. ramorum* destroy not only major tree species, but large numbers of understory species, threatening entire ecosystems. For example in southwestern Australia, *P. cinnamomi* has been estimated to directly affect over 2000 of the 9000 native plant species. In North Carolina, 34 genera of hosts are affected, including pines, cypresses, firs, cedars, rhododendrons, mountain laurel, blueberry and bearberry. In the tropics, *P. palmivora* is virtually omnivorous(1).

Oömycetes, including *Phytophthora* species, fall within the kingdom Stramenopiles (6), which constitutes a distinct major branch of the eukaryotic evolutionary tree, as distant from plants, animals and fungi as apicomplexan parasites such as malaria (Fig. 1). In addition to the oömycetes, Stramenopiles include golden-brown algae, diatoms and brown algae such as kelp.

[Fig 1 Eukaryotic Phylogeny based on 18S rRNA Sequences (below) shows tree that highlights positions of Stramenopiles as well as kingdoms involved in secondary endosymbiosis]

Because of the economic impact of *P. sojae*, molecular genetic and genomics studies are most advanced in this pathogen, along with the potato and tomato pathogen *P. infestans* (26, 45). These two species have served as model species for the oömycetes as a whole. Formal genetic analysis is especially facile in *P. sojae* because the species is homothallic and can readily be selfed. Segregation of markers is generally regular except for occasional aberrant isolates (45-49). Backcross series have been developed over up to eight generations (50). An excellent genetic map has been constructed using molecular markers and several avirulence genes have been placed on the map (47, 51, 52). In *P. sojae* the genetic map consists of 376 RFLP, RAPD and AFLP markers that define 21 major linkage groups and 7 minor groups (47, 52). Similar genetic resources have been developed *for P. infestans* (51). Stable transformation of several *Phytophthora* species, including *P. sojae* and *P. infestans*, is possible using hygromycin, neomycin or streptomycin resistance genes fused to oömycete promoter and terminator sequences (26, 53, 54). Transformation of *P. sojae*, *P. infestans* and *P. palmivora* (55) has been achieved by PEG fusion of protoplasts and by particle bombardment (reviewed by (26), and more recently, by electroporation of cell-wall-less zoospores (B. Tyler and F. Govers, unpublished) and *Agrobacterium* transformation (I. Vijn and F. Govers, unpublished). Stable transformation occurs by heterologous integration, often in tandem arrays (26). Gene silencing has proven useful in several *Phytophthora* species (56) (57) (58)

Over 30,000 ESTs from *P. sojae* as well as 7000 ESTs from *P. infestans* have been placed in public databases accessible at http://www.pfgd.org/ and http://www.vbi.vt.edu/~estap. The ESTs include 9512 *P. sojae* sequences obtained from *P. sojae*-infected soybean tissue sampled at 72 hours post-infection. These sequences provide an excellent resource for identifying infection-specific genes. The *P. sojae* ESTs define around 7200 unigenes, comprised of 2801 contigs of two or more ESTs along with around 4400 singletons. The 2801 contigs encompass 7011 ESTs from infected tissue and 16682 from a variety of axenic tissues. Of the 2801 contigs, 432 have a statistically significantly higher proportion of ESTs from infected tissue including 181 found exclusively in the infection library.

Whole Genome Sequencing and Annotation

The *P. sojae* and *P. ramorum* genomes were sequenced using a whole genome shotgun approach, resulting in approximately 9-fold coverage of *P. sojae* and 7-old coverage of P. ramorum. fGenomic DNA of *P. sojae* was extracted from mycelia of strain P6497 [Forster et al 1994] 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 accession number MYA-2949), and has a genotype identical to most P. ramorum isolates from California. The genome sequences have been submitted to Genbank.

The genome size of *P. sojae,* as estimated from the sequence assembly, is approximately 95 Mb, somewhat larger than an earlier estimate of 62 Mb based reassociation kinetics. The genome size of *P. ramorum* was estimated at 65 Mb, consistent with estimates from nuclear staining (Arredondo and Tyler, unpublished). Approximately 0.3% of the nucleotides in the *P. ramorum* sequence appear to be polymorphic, and these polymorphisms may prove useful for tracking the spread of different genetic individuals of *P. ramorum*. In contrast, only 0.07% of the nucleotides in the *P. sojae* sequence are polymorphic. This is consistent with the fact that *P. sojae* is homothallic and selfs readily to produce long-lived oospores, whereas *P. ramorum* is heterothallic and outbreeding. Forster et al (1994) reported that xxxxx loci in *P. sojae* assayed by RFLP analysis all appeared homozygous.

As an aid to the sequence assembly, a physical map of *P. sojae* was constructed by restriction enzyme fingerprinting of BAC clones from two libraries produced by digestion of genomic DNA by HindIII and BamHI respectively. The BAC clones were digested with seven enzymes (HindIII, XhoI, BamHI, BglII, XbaI, ClaI and HaeIII and labeled simultaneously with four fluorescent dyes. Contigs were assembled using FPC software. A total of 7,680 HindIII clones of average size 55 kb and 4,992 BamHI clones of average length 120 kb were fingerprinted. Of these, a total of 8,681 clones were assembled into 257 contigs. Of these, 11 contigs contain 100-200 clones; 47 contigs contain 50-99 clones; 63 contigs contain 35-49 clones; 87 contigs contain 10-24 clones and 49 contigs contain 3-9 clones. The largest contig spans a 2.2 Mb region. A minimum tiling path consisting of 1,400 clones was subjected to BAC end sequencing.

Using the JGI genome annotation pipeline that includes several gene prediction and annotation methods we identified 19,027 genes of *P. sojae* and 15,743 genes of P. ramorum. The majority of gene models (75-80%) are predicted *ab initio* using the program Fgenesh, trained for the genomes of *P. sojae* and *P. ramorum* using available EST sequences. Fgenesh achieved respectively 89% and 83% sensitivity with 88% and 85% specificity in predicting exons. The remaining 20-25% of the models are homology-based models predicted using combination of Fgenesh+ and Genewise and synteny-based modeling using fgenesh2. The latter was used to correct imperfect models of orthologous genes. A set of 9,768 putative pairs of orthologs was identified between *P. sojae* and *P. ramorum* gene models using the criterion of best reciprocal

Blast hits. Approximately 5,000 of the pairs initially had an alignment length that was less than 75% of the gene length. For these pairs, synteny-based gene prediction improved the models, resulting in an increase in the average coverage of alignment between orthologs from 89% to 93%, an increase in the number of pairs with >90% coverage from 5,823 to 6,676 pairs, and a significant drop in the number of pairs with low coverage.

For *P. sojae*, 7,850 ESTs were mapped onto the genomic assembly and used for validation, correction and extension of predicted gene models. At least partial support was found for 7,088 models.

EST sequences are not yet available for *P. ramorum*. Instead, to validate the *P. ramorum* gene predictions, we used Multidimensional Protein Identification Technology (MudPIT) to characterize tryptic fragments of proteins expressed in mycelium and germinating cysts. Of 1,438 peptides matched to scaffold_1 (1.2 Mb), 65% fall within a predicted ORF. Of these 65%, 91% fall within an ORF together with at least one additional supporting peptide. Of the 380 genes predicted in this scaffold, 79 were supported by two peptides. Peptides were considered as expanding the current gene call if they were located within 500 nucleotides of a predicted gene call, and within 1,000 nucleotides of another identified peptide. Ten percent of the peptides fell into this category suggesting that 17% of the gene calls may be expanded. And finally, 3% of the peptides were located within 1,000

nucleotides of each other, but more than 500 nucleotides away from a predicted gene, suggesting 20 new gene models.

Table 1

	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

Based on single-linkage clustering analysis, the majority of predicted genes from both genomes form groups of homologous proteins. Only a small number of genes - 1,755 in *P. sojae* and 624 in *P. ramorum* - did not have a homolog in the other genome when a significance threshold of 1e-8 was used. The overall higher number of predicted genes in *P. sojae* results from greater expansion of many gene families in *P. sojae*. Overall, about 80% of the genes in both genomes

have homology to known proteins or known protein domains, but 1,563 pairs of *Phytophthora* orthologs showed no homology to any other species than *Phytophthora*.

Whole-genome DNA sequence alignment demonstrated a high level of similarity between the two species. 75.8% of all *P. ramorum* and 79.7% of all *P. sojae* exons were covered by the alignment and about 75% of them were conserved at a high level of 70/100 bp. Non-coding regions covered by 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, or else regulatory elements.

The predicted genes were electronically annotated and classified according to Gene Ontology, KOG clusters, and KEGG metabolic pathways using sequence similarity searches. E.C numbers have been assigned to 9,520 and 9,892 genes in P. sojae and P. ramorum, respectively. 3890 and 3830 genes in P. sojae and P. ramorum have KOG assignments. Comparative analysis of annotations shows that gene counts and identities in various functional categories and pathways are very similar and indicates that these are closely related organisms.

Analysis and visualization of gene prediction and annotation for these two genomes are available from JGI Genome Portals (www.jgi.doe.gov/genomes) and at the VBI Microbial Database (http://phytophthora.vbi.vt.edu).

The 9,768 putative pairs of orthologs identified between *P.sojae* and *P.ramorum* gene models are arranged in regions of extended synteny. There are about 250 syntenic regions of four or more orthologous genes, with the longest region containing a string of 195 ortholog pairs.

SNPS and SSRs

A critical need in understanding the epidemiology of *P. ramorum* is the need to be able to distinguish different genetic individuals of *P. ramorum* so that patterns of spread can be traced. However, very little genetic variation can be detected in *P. ramorum* isolates from the US, using conventional techniques such as AFLPs (Ivors KI, Hayden KJ, Bonants PJM, Rizzo DM, Garbelotto M. 2004), presumably because most of the population has derived clonally from a single introduction or a small number of introductions of closely related strains. We examined the *P. ramorum* and also the *P. sojae* genome sequences for regions that may be useful in genetically distinguishing closely related strains. Simple Sequence Repeats (SSRs) or microsatellites have been used for genetic typing of an extensive variety of eukaryotic organisms. A total of 1,000 and 2128 microsatellite loci ranging between 2 to 6 bp in motif length were observed in the genomes of *P. ramorum* and *P. sojae* respectively. As expected, the frequency of trinucleotide repeats in exons was considerably higher when compared to di, tetra- or pentanucleotide repeats. In general, the density of SSRs (bp per Mb) in *P. sojae* is about 1.5 times that of *P. ramorum*. Dinucleotide repeats were the most abundant microsatellite repeats

making up 56% of all repeats followed by trinucleotide repeats at 29%. Single Nucleotide Polymorphisms (SNPs) offer another resource for population genetics. Sequencing of the *P. ramorum* genome identified approximately 83,077 sites at which the genome sequence of this diploid organism is polymorphic. The *P. sojae* genome sequence predicted 37,666 SNPs. The lower level of polymorphism in the *P. sojae* genome is likely due to the fact that *P. ramorum* is heterothallic (outcrossing) whereas *P. sojae* is homothallic (inbreeding). Screening of these SSR and SNP sites is underway to determine if any of these sites are variable enough to detect recent genetic divergence in the *P. ramorum* population that could be used to track patterns of spread.

Transposons

Approximately half of the *P. sojae* genome corresponds to moderately repeated sequences. The two genome sequences reveal that most of the repetitive sequences are comprised of transposons and as well as large rapidly-diverging multigene familes (see below). The transposons include both retro-elements and DNA-mediated elements, which transpose through a "cut and paste" mechanism catalyzed by a transposase, such as *Mariner*/*Tc1* and *hAT*-like transposons (Fong et al., 2004). 50 Mariner-like elements could be identified in the *P. sojae* sequences and 37 in *P. ramorum*.

NUMTS

The two genome sequences were screened for the presence of nuclear mitochondrial DNAs (NUMTs), which originate from translocations of mtDNA to the nucleus (Lopez et al. 1994). In *P. sojae* there were 103 matches to mtDNA with at least 90 % identity, ranging in length from 20 bp to 442 bp. In *P. ramorum* there were only 33 matches of at least 90 % identity, with lengths of from 19 bp to 137 bp. Searches of the two genomes with randomized mitochondrial genome sequences resulted in fewer than six matches, all less than 20 bp in length. NUMTs comprised 16.27×10^{-3} % of the genome of *P. sojae* and 2.28×10^{-3} % of the *P. ramorum* genome. Six NUMTs were recognized to be common to the two *Phytophthora* genomes. In each case the NUMT sequence was identical in the two nuclear genomes and in the two mitochondrial genomes. These NUMTs are fragments from the mitochondrial genes *rrnS*, *nad4*, *atp1*, *cox3*, *trnL* and *trnS* (multiple copies within the nuclear genomes for the latter two). Only the rrnS NUMT occur in regions of the two nuclear genomes that are syntenic (scaffold 6: 112815-113526 from P. sojae and 124: 100002-100713 from P. ramorum). [Lopez JV, Yuhki N, Masuda R, Modi W, O'Brien SJ (1994) J Mol Evol. 2: 174-90]

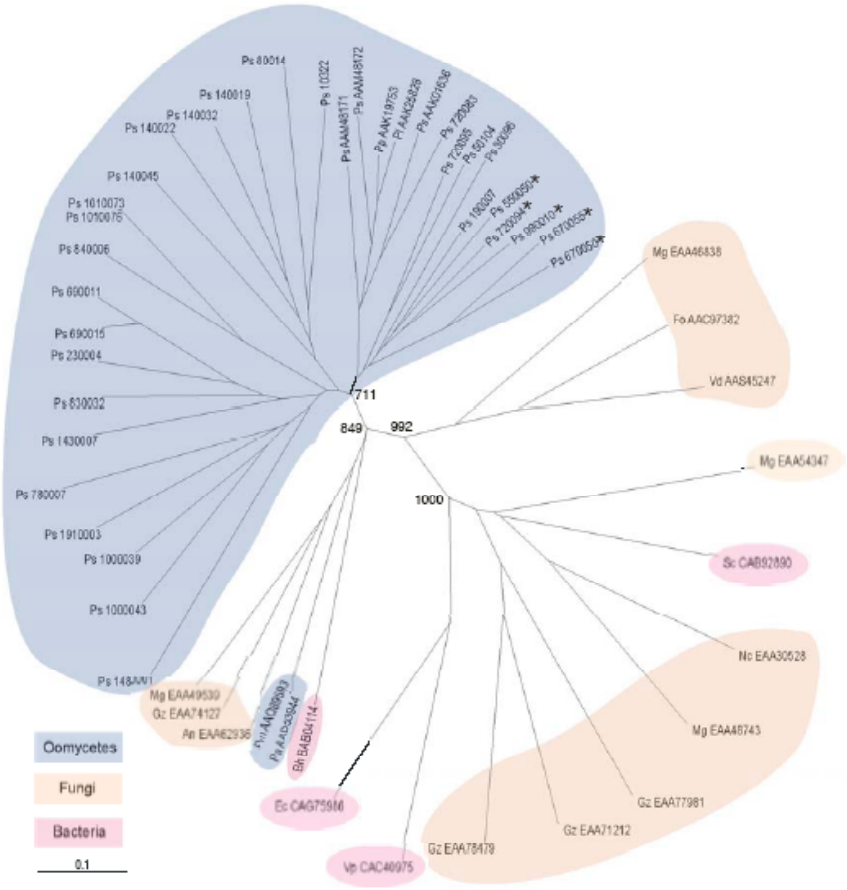

Figure 1. An un-rooted phylogenetic tree constructed from NPP1-like protein sequences from

Acknowledgements

This work was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under Contract No. W-7405-ENG-36.