Supporting Information

Richards et al. 10.1073/pnas.1105100108

SI Materials and Methods

Gene-by-Gene Phylogenetic Analysis of Four Oomycete Genomes to Identify Cases of Horizontal Gene Transfer (HGT). The predicted proteome of four oomycetes genomes, *Phytophthora ramorum*, *Phytophthora sojae*, *Phytophthora infestans*, and *Hyaloperonospora parasitica* (also named *Hyaloperonospora arabidopsidis*) (1–3), were used for gene-by-gene phylogenetic analysis using a bioinformatic protocol described previously (4). The protocol followed a multistep process. Each oomycete genome was treated separately. First, for each genome we identified and removed all candidate transposable elements by comparison with Repbase (5), a database of eukaryotic repetitive elements using tBLASTn with e-value cutoff 10^{-20}. Second, we selected the remaining protein sequences and clustered them into genes of closely related groups across the genome of origin (i.e., identifying recent gene duplications) using OrthoMCL (6) with an e-value cutoff 10^{-20} and an inflation value 1.5. Next, we identified and removed all oomycete cluster groups that were only found in oomycete genomes by performing BLASTp (7) searches of a representative sequence from each cluster group against the genomes in our local database (Table S1); proteins that had no hits other than oomycetes were removed. This left 11,434 cluster groups ready for phylogenetic analysis identifying 3,014, 3,018, 3,233, and 2,169 cluster groups from *P. ramorum*, *P.sojae*, *P. infestans*, and *H. parasitica*, respectively. Because the 11,434 gene sets were identified using a parallel process for all four genomes, this generated some four-way redundancy within the 11,434 gene sets. We did not attempt to remove this redundancy and instead analyzed all 11,434 gene sets. We took this approach to use multiple starting seeds for each gene dataset (where possible) and to control for human error in the manual tree selection stages (described below).

We next generated a fast-ML phylogeny (8) for all 11,434 cluster groups using a bespoke gene-by-gene phylogeny pipeline (4). Briefly, this process consists of a series of PERL scripts, which automatically constructed phylogenetic trees for each sequence cluster group identified. The phylogenies were calculated from taxon sampling using a custom-built MySQL database (www.mysql.com) containing the complete genome project-derived, predicted proteome sequence from 795 species, representing a wide diversity of eukaryotes and prokaryote taxa (Table S1). Each candidate sequence was compared against sequences in the database using BLASTp (7) and the best-similarity hits from each species extracted (using the e-value 10^{-20} gathering threshold). These sequences were aligned using Muscle (9), conserved regions from this alignment were sampled using GBlocks (10), and phylogenetic trees constructed using PhyML (8) with a WAG (11) + Γ + I substitution model (Γ + I parameters estimated by PhyML).

This process generated many phylogenies that were unresolved, either because taxon/sequence sampling was too narrow or because highly divergent sequences and paralogues were sampled, limiting resolution of the tree topology. In cases in which highly divergent sequences or too many paralogues limited tree resolution, we adjusted the sampling threshold using 10^{-30} and 10^{-40} to exclude divergent branches or to minimize paralogue sampling. In cases in which the taxon/sequence sampling was too narrow, we increased the threshold to 10^{-10} and 10^{-5} to sample additional members of the gene family. We then repeated the phylogenetic analyses pipeline for these data sets.

All 11,434 phylogenies were manually inspected for tree topologies that suggested fungi–oomycete gene transfer in either direction. An HGT topology is defined as an oomycete sequence branching within a clade of fungal sequences, or vice versa.

Phylogenetic Analysis of Putative HGT Gene Families. Candidate HGT phylogenies totalling 51 separate cluster groups were recovered demonstrating either oomycetes branching within the fungi or vice versa. For each candidate HGT we checked, the genome sampling encompassed all available data by comparison with the GenBank nr database, GenBank EST database, and the Taxonomically Broad EST database (12). Specific attention was made to check for additional sequence data from the *Ectocarpus* and *Blastocystis* sequence data available in GenBank, the genome project of the Diatom *Fragilariopsis cylindrus* available at the Department of Energy genome portal, the genome of the oomycete fish parasite *Saprolegnia parasitica* available at the Broad Institute's genome portal, and the de novo-generated genome sequence of the sister group to the oomycetes, the free-living osmotrophic protist *Hyphochytrium catenoides* (description of sequencing protocol is outlined below). Additional sequences were added to the alignments as required. This process was facilitated using the sequence management for phylogeny programs Refgen and Treenamer (13). In many cases there were several representatives of the gene family in a single fungal or oomycete genome, with some of these genes often having large sections of the amino acid sequence missing relative to the gene family alignment. This is most likely the product of incomplete assembly and poor gene prediction, specifically intron/exon boundaries during automatic annotation of the genomes—meaning that some predicted protein sequences from the genomes are incomplete. Where the presence of these putatively incomplete sequences did not significantly alter the taxonomic representation of the gene, incomplete sequences were excluded from the alignment. Where the taxon sampling was vital for the phylogenetic analysis, the genome sequence data were manually edited and the ORF repredicted.

Each alignment was manually edited and masked to remove gaps and ambiguous alignment positions using the alignment program Seaview (14). All gene alignments are available at http://cogeme.ex.ac.uk/hgt/Oomycete_fungi_HGT_alignment_files.zip.

For each candidate HGT alignment, we identified the optimal model for phylogenetic analysis using Modelgenerator (15). RAxML v 7.2.6 (16) analysis was then used to assess topology and bootstrap support via our easyRAx script (http://projects.exeter.ac.uk/ceem/easyRAx.html). RAxML- and Modelgenerator-predicted models were generally the same, but where they were not, Modelgenerator analyses were used (Table S2). The best-scoring RAxML tree was determined with the PROTMIX method, starting with 10 randomized maximum parsimony trees. Statistical support was evaluated with 100 bootstrap replicates.

The manual alignment checks combined with the second round of phylogenetic analysis demonstrated that 10 of the alignments did not provide strong evidence for HGT, either because the amended taxon sampling demonstrated an alternative tree topology that did not support the HGT hypothesis or because the bootstrap support and tree resolution was too weak to infer HGT.

Four of the HGT gene families were found only in fungi and oomycetes, and therefore HGT is inferred on the basis of taxon sampling only. This process left 37 datasets with phylogenetic support for HGT ready for alternative topology comparison tests (see below).

Testing Phylogenetic Support for Each HGT Using Alternative Tree Topology Tests. For the 37 gene families with phylogenetic trees suggesting HGT we identified and labeled the major fungal and oomycetes taxonomic groups relative to the HGT with a taxon code [i.e., Pezizomycotina labeled 1, Saccharomycotina 2, Taphrinomycotina 3, Basidiomycota 4, "other fungi (including paralogs with or without resolution)" 5, and Oomycetes 6] (Figs. S1.1–S1.30). We then systematically calculated phylogenetic trees with constrained monophyly of these groups and collective groupings (i.e., 1, 2, 3, 4, 1–2, 1–3, 1–4, 1–5, and 6). This process identified 88 constraint groups for the 37 alignments. Searches for the maximum likelihood (ML) trees, constrained or unconstrained, were performed using RAxML v 7.2.6. It was found that a bootstrap analysis gave better trees in most cases, and so 100 boostraps were used. Searches were carried out using the PROT-CAT-WAG-F model, as a two-step process (default in RAxML 7.2.6.). All of the trees, unconstrained and constrained, from a given alignment were grouped into a single file and the site log likelihoods calculated using RAxML. The output from this analysis was used as input to Consel version 0.1k (17) and the trees from each alignment compared using the approximately unbiased (AU) test (18).

The process identified 21 gene families for which the phylogenetic data suggest horizontal gene transfer between fungi and the oomycetes (Table S3; 20 from fungi to oomycetes and 1 from oomycetes to fungi) and where topology comparison test could reject the monophyly of the donor group [AU test: 1 phylogeny at <0.1 (borderline), 5 at <0.05, 3 at <0.01, 12 at <0.001; Table S3], demonstrating a complex pattern of transfer between these distantly related eukaryotic microbes.

In a further nine cases the AU test could not reject monophyly of the fungi. However, these putative HGTs are included here because in each case the HGT hypothesis was supported by moderate to strong bootstrap support, and/or the taxon sampling of the gene family was restricted to fungi, oomycetes, and a few prokaryote groups, suggesting HGT on of the basis of taxon distribution. The evidence for HGT is noted on a case-by-case basis in Table S3.

Analysis to Determine Whether Any Putative HGT Candidates Are the Product of DNA Contamination. Four of the identified HGTs were only present in a single oomycete genome. It is therefore possible that these genes may be annotated as oomycete genes but may be the product of DNA contamination, potentially from fungi, during a genome-sequencing project. To investigate the four single species HGTs, we identified genes adjacent to the putative HGT on genome contigs and generated phylogenies using the pipeline described above. In all four cases we could provide evidence that the gene of putative HGT ancestry was surrounded by oomycete genes of vertical inheritance, suggesting that the HGT derived gene was located on the genome of the oomycete and physically linked to native oomycete genes (Table S6).

Genome Sequencing of *Hyphochytrium*. An *H. catenoides* isolate (ATCC 18719) was inoculated onto Emerson YpSs agar (4 g yeast extract, 15 g soluble starch, 1 g dipotassium phosphate, 0.5 g magnesium sulfate, and 20 g agar dissolved in 1 L deionized water; medium was boiled for 1 min with agitation to dissolve the powder and sterilized by autoclaving at 121 °C for 15 min) and incubated at 25 °C for 2 wk. Colonies were removed, transferred to YpSs medium (made to the YpSs agar recipe but with the agar omitted) and incubated at 25 °C, with agitation, for 3 wk. Biomass was harvested by filtration through micracloth. It was then washed with ultrapure water, frozen in liquid nitrogen, and ground to a powder using a pestle and mortar previously sterilized with 0.1 M NaOH to render them DNA and RNase free. The resulting powder was split into two aliquots, half for DNA and half for RNA extraction. RNA was extracted using a LiCl RNA extraction protocol. Ground mycelium was added to equal volumes of extraction buffer [0.1 M LiCl, 0.1 M Tris (pH 8) with HCl, 10 mM EDTA, and 1% SDS, made up to volume with ddH_2O] and phenol and mixed by inverting for 1 min. To this, 0.5 volumes of CIA (24:1 chloroform/iso-amyl alcohol) was added and mixed by inversion for 30 s. The sample was centrifuged at 4 °C, 9,500 × *g* for 30 min and the upper phase transferred to a sterile tube. To this, 1 volume of 4 M LiCl was added and the tube left on ice overnight. Centrifugation was then carried out at 4 °C, 9,500 × *g* for 20 min, the supernatant removed, and the pellet washed in 70% ethanol and resuspended in Diethylpyrocarbonate water. An equal volume of phenol: CIA was added and the tube vortexed and centrifuged at 4 °C, 16,000 × *g* for 10 min. The aqueous phase was carefully recovered and 2 volumes of 100% ethanol and 0.1 volumes of 3M sodium acetate added. The RNA was left to precipitate overnight at −20 °C. RNA was recovered by centrifugation at 4 °C, 10,000 × *g* for 20 min, the supernatant removed, and the pellet washed with 70% ethanol. Finally, RNA was resupended in 300 μL ultrapure water. Total RNA was quantified and the purity checked using an ND-1000 spectrophotometer (Thermo Fisher Scientific).

DNA was extracted using the Cambio UltraCleansoil DNA kit according to the manufacturer's alternative protocol for maximum yields and quantified in an ND-1000 spectrophotometer. DNA was checked for both eukaryotic and prokaryotic contamination using SSU rDNA PCR in an MJ mini personal thermal cycler. Each 50-μL reaction contained 2 μL of each primer (10 $pM\mu L^{-1}$), 25 μL of Master Mix (Promega, containing 3 mM $MgCl_2$, 400 μM of each dNTP, and 50 U/mL of Taq DNA polymerase), 19 μL of PCR water, and 2 μL of a 1/1,000 dilution of template DNA. For universal eukaryotic SSU rRNA amplification primers 1F (5′CTGGTTGATCCTGCCAG-3′) and 1520R (5′-CTGCAGGTTCACCTA-3′) and the following cycling conditions were used: initial denatration at 95 °C for 5 min, followed by 30 identical cycles of denaturation at 95 °C for 1 min, annealing at 57 °C for 1 min, and extension at 72 °C for 1.5 min, with a final extension at 72 °C for 10 min (19). For universal prokaryotic SSU rRNA amplification primers PA (5′-AGAGTTTGATCCTGGCTCAG-3′) and PH (5′-AAGGAGGTCATCCAGCCGCA-3′) (20) and the following cycling conditions were used, with *Escherichia coli* DNA acting as a positive control: initial denaturation of 94 °C for 5 min, followed by 30 cycles of denaturation at 94 °C for 1 min, annealing at 55 °C for 1 min, and extension at 72 °C for 2 min, with a final extension step of 72 °C for 10 min. Successful amplification was checked by agarose gel electrophoresis on a 0.8% agarose gel, run at 110 V for 45 min. The prokaryotic PCR was negative, whereas the eukaryotic SSU PCR resulted in a clean band of appropriate size. The 1F-1520R PCR product was purified using the Wizard SV gel and PCR clean-up system (Promega) and sequenced externally on both strands by Cogenics (Essex). Chromatograms were checked by eye for inconsistencies, which could be the result of a multitemplate amplification before the derived sequence was used as the seed in a BLASTn search on the National Center for Biotechnology Information BLAST server, suggesting that the DNA was derived from a pure culture of the target microbe.

We sequenced the *Hyphochytrium* genome using two approaches: (*i*) 454 FLX Titanium (Roche) and (*ii*) Illumina GA2 paired-end 76-bp sequencing. The DNA was prepared for both sequencing platforms using the standard protocols. Five micrograms of DNA was fragmented by nebulization. Fragmented DNA was analyzed using a Bioanalyzer (Agilent Technologies) to ensure that the majority of the fragments were between 350 and 1,000 bp. The purified fragmented DNA was processed according to the 454 FLX Titanium Library construction kit and protocol (Roche Applied Science) to ligate adaptors specific to the Titanium sequencing chemistry. The resulting single-stranded DNA library was assessed for size distribution using a Bioanalyzer (Agilent Technologies). Library fragments were added to emul-

sion PCR beads at a ratio of 1:1 to emPCR at the optimal ratio of 1.5 DNA copy per beads and amplified according to the manufacturer's instructions (Roche Applied Science), and a full picotitre plate was sequenced. This generated 1,219,849 sequence reads resulting in 416,601,381 bp of sequence.

For Illumina GA2 sequencing we used a Bioruptor shearing device to obtain 500-bp fragments (this figure is inclusive of the adaptor sequences). V2 Cluster generation kits were used along with v4 SBS sequencing kits. Two lanes of Illumina GA2 paired-end 76-bp sequencing was performed, yielding a total of 63,082,628 reads. Reads with adaptor present were removed. Additionally, any reads without Q > 20 across >90% of bases were removed. After this filtering process, 45,453,312 reads were left in total.

The *Hyphochytrium* genome sequence was assembled using the Velvet assembler (0.7.63) (reference http://genome.cshlp.org/content/18/5/821) and resulted in 311,983 contigs spanning 85,813,724 bp at a mean coverage of 22×. Kmer size of 41 was used along with expected kmer coverage of 8 and coverage cutoff of 3 as determined by the Velvet Optimizer 2.1.7 script bundled with Velvet. At this sampling level the Lander-Waterman model (21) predicts the chance of a given base being absent from the sequence reads as less than 1 in 1 billion. However, this is in the ideal case of even coverage, which is never achieved in practice. In addition, the N50 contig length of the assembly was 611 bp. It was therefore possibile that genes would be missed in homology searches owing to the short contig sizes.

To check genome coverage further we sequenced *Hyphochytrium* cDNA using 454 FLX Titanium (Roche) chemistry. The resulting sequence reads were assembled using the Newbler assembler (Roche) on default settings and yielded 31,368 contigs. These contigs were compared with the *Hyphochytrium* assembly using BLASTn. In this simple analysis more than 99% of all cDNAs matched the assembly with >99% identity and e-value <1xe-10. The remaining contigs were either rich in simple sequence repeats or were very short (<100 bp) or of relatively low raw sequence quality. This strongly indicates that there are very few genic sequences missing from the draft assembly. We then used the CEGMA (22) (Core Eukaryotic Genes Mapping Approach) core conserved gene dataset to investigte recovery of core eukaryotic genes from the draft *Hyphochytrium* assembly. Using CEGMA 248 and the 458 genes only 17 and 28 genes, respectively, appear absent from the *Hyphochytrium* assembly using an e-value cutoff of 1e-04 and a sequence identity threshold of 90%. A more stringent cutoff of 1e-10 yielded 51 and 88 missing genes, respectively. The CEGMA core dataset is based on comparison of a range of different eukaryotic organisms (e.g., animals, plants, alveolates) but does not include any close relatives of *Hyphochytrium*. Nevertheless, depending upon the stringency used the gene recovery rate for the CEGMA dataset was between 80% and 94%.

Hyphochytrium genome sequence data have been submitted to Sequence Read Archive for the raw data (SRP004821), and the assembled genome has been archived at the European Bioinformatics Institute (Genome Project: 61035).

We then searched the assembly for candidate homologs of the 34 oomycete HGTs in the *Hyphochytrium* genome using a tBLASTn approach. Of the 34 oomycete HGTs, in seven cases candidate homolog sequences were identified in *Hyphochytrium*. In five cases the multiple alignment analyses demonstrated that the *Hyphochytrium* sequences were distant relatives, or too divergent to be true homologs, and were therefore not relevant to the putative HGT and excluded from further analysis. In the remaining two cases we added the *Hyphochytrium* genes to the alignments and recalculated the phylogenetic analysis. In both cases the *Hyphochytrium* genes were distantly related paralogues to the oomycetes HGTs branching separately from the fungal/oomycetes HGT branches (Figs. S1.5 and S1.16), demonstrating that the *Hyphochytrium* genes have a different evolutionary derivation and that all 34 fungal-derived HGTs are absent from *Hyphochytrium* genome assembly.

Putative Functional Annotations of the HGT Gene Families and Comparison with *Phytophthora* Metabolic Network Analysis. To annotate putative functions for all 34 HGT gene families, a representative of each HGT gene was used for BLAST (7) and PFAM HMM homology searches (23). This process was used to assign a putative function to each HGT gene family. Each putative annotation was checked to exclude false annotations present in GenBank. To investigate putative cellular locations of each HGT-encoded protein, the complete HGT gene families, including all duplicate forms from all four oomycete genomes, were recovered and subject to WoLFPSORT (24), SignalP (25), and TMHMM (26) analysis for evidence of N-terminal secretion motifs and/or transmembrane domains. The results of these annotations are listed in Table S5. Evidence for an N-terminal secretion motif is only listed in Table S5 when both WoLFPSORT (24) and SignalP (25) methods suggest that the protein character is present.

To investigate where the oomycete HGT proteins putatively function within the metabolic network of the *Phytophthora* species, we sought to reconstruct the oomycete metabolic network. Complete proteome sequence files of *P. ramorum*, *P. sojae*, and *P. infestans* (proteins. fasta files) were recovered from the Broad Institute *Phytophthora* database. We used Kaas (Kegg automated annotation server) to project the downloaded proteome sequences onto Kegg's collection of metabolic pathways (27), using BDH (bidirectional hit) orthology search (similarity index threshold: 60) to explore the Kaas reference genome set (enriched with available fungi and protozoa genomes). Initially this analysis was performed to check and correct the annotations for the 34 HGT gene families (discussed above). Then we used this process to investigate where the HGT candidates fitted into the oomycete metabolic network map. This network analysis often demonstrated partial or missing pathways. In these cases we exploited a minimal threshold (28) in a combination with an SDH (single-direction hit) method to recheck specific components of pathways putatively connected to the HGT functions. The BDH and SDH methods were completed for all three *Phytophthora* predicted proteomes. We also used BLASTn and tBLASTn additionally to query the Broad Institute *Phytophthora* databases for the missing functions by known protein and gene sequences to reconfirm incomplete pathways. Finally, we used the SEED database (29) "Compare regions" tool to investigate synteny of each HGT gene family across distantly related taxa to help define evidence of potential functional linkage of putative homologs of the HGT genes.

Using microarray data for the *P. infestans* transcriptome, we identified evidence of altered transcript abundance both before and after infection. To do this we used publicly available data sets (2) recording the expression of genes in *P. infestans* over a 5-d time course of a potato infection to identify genes previously classified as "genes induced or repressed during infection." These candidate genes were identified by comparing expression intensities derived from samples of infected potato tissue to baseline expression intensities provided by samples derived from mycelium growing in axenic culture (Pea, V8, or RS agar). The data were published as part of the *P. infestans* genome project (2). Normalized transcript abundance values between the two samples were compared using the Student *t* test. Of the 18 HGT gene families present in the *P. infestans* genome, 6 HGT gene families showed evidence of up-regulation *in planta* [2 gene families at <0.1 (borderline), 2 at <0.05, and 2 at <0.01; Table S5].

1. Tyler BM, et al. (2006) *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313:1261–1266.
2. Haas BJ, et al. (2009) Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* 461:393–398.
3. Baxter L, et al. (2010) Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. *Science* 330:1549–1551.
4. Richards TA, et al. (2009) Phylogenomic analysis demonstrates a pattern of rare and ancient horizontal gene transfer between plants and fungi. *Plant Cell* 21:1897–1911.
5. Jurka J, et al. (2005) Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet Genome Res* 110:462–467.
6. Li L, Stoeckert CJ Jr., Roos DS (2003) OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res* 13:2178–2189.
7. Altschul SF, et al. (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res* 25:3389–3402.
8. Guindon S, Gascuel O (2003) A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* 52:696–704.
9. Edgar RC (2004) MUSCLE: A multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* 5:113.
10. Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* 17:540–552.
11. Whelan S, Goldman N (2001) A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. *Mol Biol Evol* 18:691–699.
12. O'Brien EA, et al. (2007) TBestDB: A taxonomically broad database of expressed sequence tags (ESTs). *Nucleic Acids Res* 35(Database issue):D445–D451.
13. Leonard G, Stevens JR, Richards TA (2009) REFGEN and TREENAMER: Automated sequence data handling for phylogenetic analysis in the genomic era. *Evol Bioinform Online* 5:1–4.
14. Galtier N, Gouy M, Gautier C (1996) SEAVIEW and PHYLO_WIN: Two graphic tools for sequence alignment and molecular phylogeny. *Comput Appl Biosci* 12:543–548.
15. Keane TM, Creevey CJ, Pentony MM, Naughton TJ, Mclnerney JO (2006) Assessment of methods for amino acid matrix selection and their use on empirical data shows that ad hoc assumptions for choice of matrix are not justified. *BMC Evol Biol* 6:29.
16. Stamatakis A (2006) RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688–2690.
17. Shimodaira H, Hasegawa M (2001) CONSEL: For assessing the confidence of phylogenetic tree selection. *Bioinformatics* 17:1246–1247.
18. Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51:492–508.
19. Lefèvre E, et al. (2007) Unveiling fungal zooflagellates as members of freshwater picoeukaryotes: evidence from a molecular diversity study in a deep meromictic lake. *Environ Microbiol* 9:61–71.
20. Edwards U, Rogall T, Blöcker H, Emde M, Böttger EC (1989) Isolation and direct complete nucleotide determination of entire genes. Characterization of a gene coding for 16S ribosomal RNA. *Nucleic Acids Res* 17:7843–7853.
21. Lander ES, Waterman MS (1988) Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics* 2:231–239.
22. Parra G, Bradnam K, Korf I (2007) CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23:1061–1067.
23. Marchler-Bauer A, et al. (2005) CDD: A Conserved Domain Database for protein classification. *Nucleic Acids Res* 33(Database issue):D192–D196.
24. Horton P, et al. (2007) WoLF PSORT: Protein localization predictor. *Nucleic Acids Res* 35(Web Server issue):W585–W587.
25. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340:783–795.
26. Sonnhammer EL, von Heijne G, Krogh A (1998) A hidden Markov model for predicting transmembrane helices in protein sequences. *Proc Int Conf Intell Syst Mol Biol* 6: 175–182.
27. Moriya Y, Itoh M, Okuda S, Yoshizawa AC, Kanehisa M (2007) KAAS: An automatic genome annotation and pathway reconstruction server. *Nucleic Acids Res* 35(Web Server issue):W182–W185.
28. Bateman A, et al. (2004) The Pfam protein families database. *Nucleic Acids Res* 32 (Database issue):D138–D141.
29. Overbeek R, et al. (2005) The subsystems approach to genome annotation and its use in the project to annotate 1000 genomes. *Nucleic Acids Res* 33:5691–5702.

Figs. S1.1–S1.30. Phylogenetic evidence for fungi–oomycete HGTs. Phylogenies are calculated as described in *SI Materials and Methods*. Each figure contains a key describing the figure notations. Note that Figs. S1.5 and S1.16 are the datasets with representation from *Hyphochytrium*

Figs. S1.1–S1.30

Table S1. List of genomes included in the phylogenetic pipeline database used in this study

Table S1

Table S2. Model parameters used for phylogenetic analysis estimated by Modelgenerator (RaxML has a preset α parameter for the Γ distribution, so no α value is given, etc.)

Table S2

Table S3. Phylogeny topology comparison tests to confirm branching position of HGT within donor group

Table S3

HGT in red demonstrate transfer in opposite direction. Transfers in gray are not supported by alternative topology comparison test, or alternatively the HGT hypothesis is based on taxon distribution data only, suggesting a weaker standard of support for the HGT. Sequences for *P. ramorum* and *P. sojae* are available from the Department of Energy, Joint Genome Institute website. Protein sequences can be obtained using the protein ID and searching at http://genome.jgi-psf.org/pages/search-for-genes.jsf?organism=Phyra1_1 for *P. ramorum* and http://genome.jgi-psf.org/pages/search-for-genes.jsf?organism=Physo1_1 for *P. sojae*.

Table S4. Evidence of conserved introns across donor and recipient groups in eight HGT gene families

Table S4

Of the 30 HGTs supported by phylogenetic data, we searched for conserved introns in the HGT gene families of the recipient taxa. We found 15 cases in which the transferred gene family had one or more conserved intron. Of these 15 HGT families, 8 conserved introns were also present in the donor groups, suggesting that the pattern of HGT is from eukaryote-to-eukaryote and not as a eukaryotic-to-prokaryotic-to-eukaryotic transfer.

Table S5. Annotation of putative HGT candidates from fungi to oomycetes

Table S5

The HGT shaded in red demonstrates a putative transfer in opposite direction (oomycete to fungi). Transfers in gray are not supported by AU alternative topology comparison test (1), or alternatively the HGT hypothesis is based on taxon distribution data only, suggesting a weaker standard of support for the HGT. Evidence for secretion was identified using both the SignalP and WolfP-Sort methods, as described in *SI Materials and Methods*. Presence of a putative N-terminal secretion motif is only suggested when both methods agree. Sequences for *P. ramorum* and *P. sojae* are available from the Department of Energy, Joint Genome Institute website. Protein sequences can be obtained using the protein ID and searching at the Department of Energy Joint Genome Institute website: http://genome.jgi-psf.org/pages/search-for-genes.jsf?organism=Phyra1_1 for *P. ramorum* and http://genome.jgi-psf.org/pages/search-for-genes.jsf?organism=Physo1_1 for *P. sojae*. Overlap between this analysis and the results reported by Richards et al. (2) is also listed in column 3. In this previous study we identified an additional three tentative HGTs [one Aconitase genes family and two divergent paralogues of the Esterase/Lipase gene family; Fig. S3 of the 2006 paper (2)]. In the 2006 study we could not provide strong evidence either to support or refute the HGT hypothesis. These three HGTs were not recovered in this analysis.

1. Shimodaira H (2002) An approximately unbiased test of phylogenetic tree selection. *Syst Biol* 51:492e508.
2. Richards TA, Dacks JB, Jenkinson JM, Thornton CR, Talbot NJ (2006) Evolution of filamentous plant pathogens: gene exchange across eukaryotic kingdoms. *Curr Biol* 16:1857–1864.

Table S6. Check for contaminant sequences by phylogeny of genes adjacent to HGT

Table S6

Other Supporting Information Files

Dataset S1