Signatures of adaptation to obligate biotrophy in the *H. arabidopsidis* genome

Supporting Online Material

Table of Contents

1. MATERIALS AND METHODS

Isolation of genomic DNA

Asexual sporangiophores were collected by gently shaking sporulating leaves in water. Spores were then collected by centrifugation and DNA was isolated as described in (*1*).

Plasmid and Fosmid library construction

Plasmid library construction: To generate genomic fragments, each DNA sample was processed in a Hydroshear machine (GeneMachines). Linkers (BstXI/EcoRI-Invitrogen) were added to the ends of the repaired fragments (Lucigen DNA Terminator Kit), according to manufacturer's directions. Using gel electrophoresis, one size fraction was selected for isolation and subcloning: 4.0-4.3 kb. A second gel electrophoresis of each isolated fraction was used to remove excess linkers. The genomic fragments were retrieved from the matrix by degrading the agar with agarase, followed by phenol extraction. The DNA was concentrated using ethanol precipitation. The DNA fragments from the isolated size fraction were subcloned into plasmid vector pOTw13. The vector was first linearized with BstXI and the 1.7 kb vector fragment was gel purified to eliminate the SacB stuffer fragment. The vector fragment was dephosphorylated and purified by preparative gel electrophoresis. Ligation was performed according to manufacturer's directions (New England Biolabs), and the *Escherichia coli* host, DH10B-T1 phage-resistant, was transformed with the chimeric plasmid DNAs. Transformed cells were plated for production using chloramphenicol and IPTG selection.

Fosmid library construction: A fosmid library was constructed using an arabinose-inducible Szybalski vector (*2*) known as "pCC" (pCopyControl) from Epicentre Technologies (Madison, WI). Briefly, the genomic DNA was sheared to the required 40 kb size range. The DNA was end-repaired to blunt, 5'-phosphorylated ends, and the end-repaired DNA was size resolved by low melting point agarose gel electrophoresis. The blunt-end DNA was recovered from the low melting point agarose and was ligated into the CopyControlpCC1FOS vector. The ligated DNA was packaged into EPI300-T1R plating cells and grown overnight. Glycerol stocks were prepared, ready for induced growth, plating, picking and DNA extraction. Fosmids prepared by Lucigen according to their company protocol also were used.

Clone growth: The pcc01 fosmid trays were plated on a LB+Sucrose+chloramphenicol +IPTG+X gal media. The colonies were picked and transferred into a LB + 80% glycerol + chloramphenicol + glucose media using a Qpix robot into a 384 well tray and incubated at 37°C for 24 hours. The picked 384 well tray became the master copy, which was used for a replication step into a new 384 well tray. A Biomek robot dispensed cells from the master tray into LB+80% glycerol+chloramphenicol+glucose media in the 384 well replication tray. The replication tray was incubated 24 hours at 37°C. Both master and replication copies were retained for future use. A final 384 well growth tray, containing LB+chloramphenicol media plus an auto-induction solution, was loaded with cells from the replication copy. The auto induction solution allowed for an increased cell growth, with a concomitant increase in DNA yield before purification. This growth tray was incubated for 18 hours at 37°C. The plasmid vector pOTw13 cells were plated on chloramphenicol + sucrose media. The colonies were then picked into a 2xYT + chloramphenicol media with a Qpixrobot into a 384 archive well tray and incubated at 37°C for 24 hours. The 384 well archive tray was then sent for DNA purification.

DNA purification: The DNA purification step followed the same method for both pOTw13 and pcc01. The final growth tray was placed on a Biomek robot for the archive prep method. Cells from the growth tray were added to a cycle plate for the 5' end sequence and a cycle plate for the 3' end sequence. Both cycle plates rested on magnetic plates. After the cell addition, a solution containing lysis solution and magnetic beads was dispensed into the wells. The released DNA attached to the magnetic beads. The beads were then pulled to the bottom of the wells by the magnetic plate. This allowed for the remaining media and cell debris to be aspirated from the wells. Three ethanol washes were performed and the tray was dried.

Sequencing reactions: For each direction of the DNA to be read, the appropriate primers plus big dye for each vector were added to the plate. The cycle plates were placed in thermal-cyclers for 35 cycles of 30 seconds at 95°C, 15 seconds at 50°C, and 2 minutes at 60°C. The reactions were precipitated with a 100 ml 100% EtOH/1 ml NaOAc mixture, followed by a 70% ethanol wash. The pellets were then dried.

The processed sequencing reactions were re-hydrated and were loaded onto ABI3730XL capillary sequencing robots. The samples were run on 36 cm arrays with a run time of 60 minutes per 96 well quadrant. As sequencer runs were completed, the data was transferred to an Oracle database. For all of the above activities, sample processing and tracking were facilitated by a bar code system that is linked to the Oracle database.

BAC library construction and analysis

Library construction. The *Hpa* BAC library was constructed with BAC vector pBeloBACII as described previously (*3*). The library contains 7996 cloned inserts, with an average insert size of 80 kb.

BAC end sequencing. BAC clones were processed and sequenced using standard 384 well protocols (Sanger Centre, UK). Separate sequencing reactions using SP6 and T7 primers were performed for each BAC clone, resulting in a total of 13,071 sequence traces. Raw trace reads are available to download from the ENSEMBL trace server repository ftp site, ftp://ftp.ensembl.org/pub/traces/hyaloperonospora_parasitica_at/.

BAC clone sequencing. BAC clones were grown, processed and sequenced using standard 384 well protocols, as according to the standard Sanger sequencing pipeline (Sanger Centre, UK). The finished sequences are available for download from the EMBL Nucleotide Sequence Database (http://www.ebi.ac.uk/embl/).

BAC fingerprinting. BAC DNA was extracted, transferred to 384 well plates, and restriction digested with *Hin*dIII for 2 hours at 37°C. Products were separated on 1% agarose gels, run at 85 V for 16 hours, and stained with Vistra Green for 45 minutes. Bands were visualized on a Typhoon scanner with 526SP emission filter, and images stored digitally.

BAC fingerprint contig assembly. The raw gel images were processed with IMAGE software (http://www.sanger.ac.uk/Software/Image/) to generate an output of normalized band values, sizes and gel traces that were analysed in an FPC fingerprint database (*4*). FPC bins and orders clones on the basis of shared bands. The fingerprint assembly was performed using a cutoff of 1e-09 and a fixed match tolerance of 7 generating 713 contigs from fingerprints of 6181 clone inserts. Manual merging of contigs brought the contig number to 708.

Assembly of Sanger shotgun reads

The genome was sequenced to a total of 9.5X phred Q20 redundancy from plasmid, fosmid and BAC end sequences. The combined sequence reads were assembled using the PCAP software (*5*). PCAP was then used to process the overlaps (bdocs, bclean), calculate the layout (bcontig) to generate the consensus sequence, using default parameters. Using this dataset, a round of automated sequence improvement (pre-finishing) was done and 23,855 of 32,122 pre-finishing reads were incorporated into the initial assembly. After the initial assembly with PCAP, we modified the read pair constraints iteratively after recalculating the statistics on read pair distances from the assembly. While the contiguity improved, the assembly was still fragmented. The largest supercontigs were used for the N50 calculations.

The initial PCAP assembly, consisting of only plasmid end sequences, contained 1,053,419 reads, yielding more than 8 fold coverage for an estimated 70 Mb shotgun assembly. A total of 1,014,758 reads was assembled using the PCAP software (*5*). Assembly quality assessment accounted for read depth, chimeric reads, repeat content, cloning bias, and G/C content. The final PCAP assembly (version 7.0.1), which included all plasmid sequences, 8346 BAC end sequences, 25,516 fosmid end sequences, plus pre-finishing reads was performed. This yielded 9.5 fold phred Q20 sequence coverage. Additional filtering following assembly removed contigs less than 2000 bases, as well as potential contaminants. 5354 contigs (including a large number of singletons) were removed by this process, with 5473 contigs and 1842 supercontigs remaining. Approximately 98-99% of the 67 Mb shotgun assembly is covered.

The sequence of version 7.0.1 has been deposited on NCBI's Entrez Genome Project website, http://www.ncbi.nlm.nih.gov/sites/entrez, with the Genome Project ID 30969, WGS accession numbers [ABWE01000001-ABWE01005422] and project accession number [ABWE00000000].

Merging of full length BAC sequences with the Sanger Shotgun assembly

The BAC sequences were matched to v7.0.1 using Blat. Where the length of the BAC sequence differed from the length of the spanned assembly by less that 1%, the BAC sequence was automatically substituted for that region of the assembly. Where the BAC joined two contigs, the BAC sequence was automatically used to join the contigs if the replaced sequences differed no more than 1% with the BAC sequence. All other matches were reviewed manually. 30 BACs were not easily integrated due to substantial unresolvable errors in the Sanger assembly. These BAC sequences were appended to the assembly, and should be considered the authoritative assembly of the relevant regions. We calculated that this introduced 2.4 Mb of additional redundancy into the assembly. All short Sanger assembled contigs entirely contained within full length BAC sequences were removed from the assembly.

Illumina genome and cDNA sequencing

DNA extraction. Genomic DNA was extracted from *Hpa* conidospores from infected *A. thaliana* Ws *eds*1-1 plants using a Nucelon PhytoPure DNA extraction kit using the default protocol followed by a phenol/chloroform extraction and isopropanol precipitation.

Illumina DNA library preparation and sequencing. The non-paired-end libraries were sequenced on the Illumina GA1 platform using 120bp inserts. The paired-end libraries were sequenced on the Illumina GA2 platform using 400bp (±10%) inserts. The protocol used was the same as the manufacturer's protocol except that the purification of the ligation of the Illumina adapters were performed on a 5% polyacrylamide gel and the library validation was performed a 6% polyacrylamide gel. The base calling was done on the Illumina GAP v1.0 pipeline for all but 2 lanes (which were performed on the GAP v1.3 pipeline). All the raw sequences have been deposited in the NCBI short read archive: http://www.ebi.ac.uk/ena/data/view/ERP000272.

Quality checking the Illumina preparation and sequencing. The libraries were sequenced on a single lane initially for quality checking after which the decision to sequence further lanes was made. For both the paired-end and non-paired-end sequencing runs, a PhiX control lane was also run to eliminate mechanical error. The raw reads generated from the Illumina Pipeline included errors in the form of PCR duplicate reads, adapter contamination and *Xanthamonas* contamination. Contamination was dealt with through post analysis filtering through sequence homology analysis. PCR duplicates accounted for 9% of the generated *Hpa* Emoy2 reads and were removed before analysis to avoid spurious heterozygosity detection.

Illumina cDNA sequencing. *Hpa* RNA was extracted from infected leaves of 7 days post inoculation *A. thaliana* Ws *eds1-1* using TRI-REAGENT according to protocol (Sigma). RNA was resuspended in diethylpyrocarbonate (DEPC) treated water. RNAse inhibitors (RNAseguard, Promega) were added and samples were DNAse treated (RNAse free, Roche). RNA was re-extracted with phenol/chloroform, ethanol precipitated and resuspended in DEPC treated water. First and second strand cDNA syntheses were performed using the default protocol from the Creator SMART cDNA Library Construction KitTM (Clontech). After the last amplification step, cDNA was phenol/chloroform extracted followed by isopropanol precipitation. The cDNA was then normalized using Duplex-specific nuclease (Evrogen) according to default protocol. The normalized cDNA was than prepared to be sequenced on the Illumina GA1 platform using 120bp inserts with a 35 bp read length (Kemen et al., unpublished).

The base calling was done on the Illumina GAP v1.0 pipeline. All the raw sequences have been deposited in the NCBI short read archive: http://www.ebi.ac.uk/ena/data/view/ERP000272.

Integration of the Sanger and Illumina assemblies

Illumina sequencing of Emoy2 DNA yielded 35x coverage with paired-end reads, and 11x coverage with non-paired-end reads. The Velvet algorithm (*6*) was used to assemble the paired-end Illumina reads. The resultant assembly was 56.9 Mb. The longest contig was 603,164 bp and the mean contig length was 2980 bp.

To integrate the Velvet assembly with the Sanger assembly (including full length BACs) we developed a four-stage pipeline. In stage 1, the Illumina reads were matched to the Sanger assembly using MAQ (http://maq.sourceforge.net/maq-man.shtml) to identify sequencing errors in the Sanger assembly that could be confidently corrected using the Illumina data. In stage 2 we used methods similar to those described (*7*) to use Velvet assembled short reads to span gaps in the Sanger assembly. Reads within 250 bp of gaps and ends of contigs were identified by MAQ and assembled using Velvet. The assemblies were then matched back to the Sanger assembly using Blat, and integrated into the assembly, replacing the corresponding region of the Sanger assembly. In stage 3 we identified and removed regions of the Sanger assembly not covered by Illumina paired end reads that showed homology to possible contaminants. In stage 4, 14 scaffolds >2 kb that were unique to the Velvet assembly (largest was 6.2 kb), were appended to the assembly, creating the 86.1 Mb v8.3.

We used dnadiff (*8*) to compare the Velvet assembly to the v8.3 assembly. 99.7% of the Velvet assembly aligns to the current v8.3 assembly while 97.3% of the v8.3 assembly aligns to the Velvet assembly. The difference between the size of the Velvet assembly and the v8.3 assembly is due to 27.2 Mb of sequence being collapsed into 5.9 Mb of scaffolds in the Velvet assembly and 2 Mb of 'N's captured through the larger Sanger paired end reads.

V8.3 assembly statistics for the H. arabidopsidis genome

Release version	Release date	WGS plasmid reads	WGS fosmid reads	BAC ends	Illumina paired end reads	Total input reads	Size (Mb)	Coverage	Major scaffolds (>2 kb)	N50 scaffold number
8.3	September 2009	1080646	25516	13071	56727498	57846731	82	45X	1783	75

Genome size estimation

The total length of the v8.3 assembly was 82.1 Mb. To independently estimate the total genome size, we conducted statistical analyses of the coverage provided by the Illumina reads and by the Sanger reads.

To estimate the genome size from the coverage provided by the Illumina reads, 67 Mb of high quality assembled Sanger sequence from version 7 of the assembly was matched using MAQ to 2393125128 bp of sequence from Illumina paired-end reads (66475698 total reads from six lanes). The Illumina read coverage at each nucleotide position (67509127 positions) was calculated and the frequency of positions with each level of coverage was plotted (Figure S1A). A Gaussian curve was fit to the main peak of the distribution by least squares and used to obtain the mean of the distribution (23.93). The genome size was estimated by dividing the total length of the Illumina reads by the mean coverage: 2393125132/23.93 = 100.0 Mb.

To estimate the genome size from the unassembled Sanger read data, the coverage of every read was calculated from a Blastn all-versus-all search of the trimmed random shotgun reads. For each read we counted the number of matches with > 95% identify over the length of the match, with a minimum overlap of 30 nt. The average coverage at the nucleotide positions defined by each read can then obtained from the following formula:

$$C = 1 + M*R/[(R+L)-2*O]$$

Where M = the number of matches
R = average length of all the trimmed trace files (720 nt)
L = length of each individual query sequence
O = minimum overlap required to call a match (i.e. 30 nt)

The frequencies of reads with different coverages were then binned and plotted to identify a peak corresponding to the single copy sequences (Figure S1B). A Gaussian curve was fit to the single copy peak by least squares and used to obtain the mean of the distribution (8.39). The genome size was estimated by dividing the total length of the Sanger reads by the mean coverage: 1140851*720 /8.39 = 97.9 Mb.

The close agreement of the two statistical estimates suggests that the actual genome size (mean of the two estimates = 99 Mb) is significantly larger than the assembled length of 82 Mb.

An explanation for the discrepancy is suggested by the second prominent peak in Figure S1B. The presence of this peak suggests that there are a large number of sequences in the genome that are greater than 95% identical and have an average copy number of around 3. Such sequences would very likely have been assembled as single copy sequences by the assembly software. Plotting the sequence coverage provided by the Sanger reads against the assembled genome did not reveal any contigs or long segments of the assembly with elevated coverage (not shown), ruling out the presence of large triplicated regions or chromosomes. The reads with elevated copy number also did not correspond to contaminants such as bacteria or *Arabidopsis*. The reads with elevated copy number did not correspond to gene models from *Hpa*, suggesting that the repeats were largely confined to non-genic regions. The plot of the Illumina read coverage did not identify a sharp peak of triplicated sequences, but rather a long tail corresponding to high copy coverage. The different shapes of the two plots likely resulted from the fact that the Illumina reads were much shorter than the Sanger reads, and a perfect match was required for the Illumina reads, compared to a 95% match for the Sanger reads.

Sanger EST sequencing

Two cDNA libraries were constructed using total RNA extracted from spores of *Hpa* isolate Emoy2. For library Hp_ENSC (*Hpa* Emoy2, Normalised, Spore-derived cDNA), cDNA was cloned into pExpress 1 (NotI-EcoRV) vector, with an average insert size of 1.65 kb (library constructed by Express Genomics Inc, Maryland, USA). Library Hp_EFNS (*Hpa* Emoy2 Full-Length Enriched, Normalised Spore total RNA) was constructed using plasmid cloning vector pBlueScript II SK+, is enriched for cDNAs carrying complete 5'-ends, with an average insert size of 1 kb (Vertis Biotechnologie AG, Freising, Germany). Both libraries were plated and clones containing cDNA sequences were prepped and sequenced with standard 384 well plate protocols (Sanger Centre, UK), using M13FWD and M13REV primers or M13 and T7 primers. From the Hp_ENSC library, 4500 sequences were generated, and 32,153 sequences from the Hp_EFNS library, derived from a total of 18432 clones. All raw trace sequences are available to download from the ENSEMBL trace server repository ftp site, ftp://ftp.ensembl.org/pub/traces/hyaloperonospora_parasitica_at/. Raw traces were processed using a locally installed semi-automated pipeline, based on Trace2dbest (*9*). Sequences corresponding to vector, adaptor, poly A/T tails, or low sequence quality (Phred score < Q20) were screened and removed from the raw traces. The cDNA sequences were assembled into 13364 unigenes using TGICL software on a linux server. Unigenes were aligned to genomic scaffolds using Exonerate, running the Est2genome model. Alignment scores exceeding 2000 were considered significant matches.

454 cDNA sequencing.

Hpa RNA was extracted from infected leaves 3 days post inoculation of 3 week-old *A. thaliana* Ws *eds*1-1 using a protocol adapted from (*10*). RNA was resuspended in DEPC treated water. RNAse inhibitors (RNAseguard, Promega) were added and samples were DNAse treated (RNAse free, Roche). RNA was re-extracted with phenol/chloroform, ethanol precipitated and resuspended in DEPC treated water. First and second strand cDNA syntheses were performed using the default protocol from the Creator SMART cDNA Library Construction KitTM (Clontech). After the last amplification step Proteinase K digestion was performed with the whole of the reaction and not just with half as in the Creator SMART protocol. cDNA was phenol/chloroform extracted and ethanol precipitated using 1.3 µg glycogen. For positive selection of *Hpa* cDNAs, 4 µg of genomic DNA, genomified using the GenomiPhi Kit (GE Healthcare), digested and biotinylated (Rougon-Cardaso, unpublished, thesis) were mixed with the target (cDNA synthesized from RNA of infected leaves) and ethanol precipitated. The mixture was resuspended in 10 µl of sterile hybridization buffer after which the driver and target were denatured at 95^{o}C for 10 minutes and hybridized for 36 hours at 66^{o}C. The biotinylated DNA was captured by Streptavidin coated beads (Magnasphere Paramagnetic Beads, Promega). Hybrids were recovered using a protocol adapted from (*11*). Positively selected cDNA was digested with Sfi I and Mse I restriction enzymes. Oligonucleotide fragments and salt were removed by spin-column chromatography through a Sephadex G-25 resin (Roche) after each digestion. The following primers were ligated to form adapters 454A and 454B: Biot-SfiAdaptor454Aoverhang, Biotin-AGCCTCCCTCGCGCCATCAGATTA; SfiAdapter454Acomp, PO_4-TCTGATGGCGCGAGGGAGGC; Mse-TOP, TACTGAGCGGG CTGGCAAGGC; Mse-BOT, GCCTTGCCAGCCCGCTCAG.

400 ng of cDNA were ligated with 300 ng adapter 454A and 300 ng of adapter 454B. Biotinylated fragments were hybridized to 20 µl Magnasphere Paramagnetic Beads (Promega) pre-washed as specified by the manufacturer and pre-incubated with blocking agents. Beads with hybridized cDNA were washed 4 times with 0.1xSSC and captured with a magnet (Promega) and supernatant was discarded. After preparation of cDNA with 454 adaptors attached the sample was sent to 454 Life Sciences (Branford, Connecticut, USA) for further processing and sequencing with 454 GS-Flx technology.

The 266500 returned 454 sequenced reads were filtered for oomycete ribosomal protein genes and *A. thaliana* contamination, resulting in 61,327 passing ESTs.

Gene prediction

Gene predictions in *Hpa* were done using a combination of *ab initio* methods and whole genome BLASTX to the NCBI nr database.

Augustus (*12*), a generalized Hidden Markov Model (HMM) method that uses hints from external sources, was downloaded from http://augustus.gobics.de/. It was configured and retrained using 1724 genes assembled from cDNA sequences from *Hpa* using the PASA pipeline (*13*), and additional hints were provided through the assembled Illumina cDNA data.

GeneZilla, formerly TIGRSCAN (*14*), a generalized Hidden Markov Model (HMM) method, was obtained from http://www.genezilla.org/. It was reconfigured and retrained using cDNA sequences from *Hpa.*

SNAP**.** The *ab initio* gene finding program Semi-HMM-based Nucleic Acid Parser (SNAP) was obtained from http://homepage.mac.com/iankorf (*15*) and installed onto Mac OS X. Species-specific parameter estimation was performed using a training set derived from 100 manually curated *Hpa* genes with full-length EST support, and a HMM was built from these parameters. The SNAP program was run with the HMM and genome sequence as inputs, from which the gene predictions were generated.

Merging gene predictions**.** After predictions were made by the individual methods, the gene models were compared with each other. Where gene models were identical, the GeneZilla model was kept (arbitrary). Overlapping gene models from different prediction methods that varied in splice sites were quality checked to retain the best model from all the predictions for that locus. Non-overlapping gene models from all methods were again passed through QC including hexamer analysis and coding potential analysis to determine the false positives and true positives. The false positives were discarded. The overlapping and non-overlapping datasets were merged into the final gene model call. Gene IDs were assigned to these newly created gene models on the basis of their location on the supercontigs.

cDNA support for **H. arabidopsidis** *transcript models*

Sanger EST unigenes from spore cDNA were aligned to *Hpa* genomic scaffolds using Exonerate, running the Est2genome model. The Illumina tags from infected tissue were aligned to the *A. thaliana* TAIR9 genome assembly and the TAIR9 transcript models, using MAQ. 28% of the Illumina cDNA aligned to the *Hpa* v8.3 assembly and 12% aligned to the v8.3 transcript models. 52% of the cDNA that aligned to the v8.3 assembly also aligned to the transcript models. By comparison, 36% of the cDNA aligned to the TAIR9 genome assembly and 17% of the cDNA aligned to the TAIR9 transcript models. 45% of the cDNA that aligned to the genome assembly also aligned to the transcript models. The majority of the cDNA that aligned to the TAIR9 genome assembly but not the TAIR9 transcript models aligned to untranslated regions of the genes. From this we conclude that the v8.3 transcript models are fairly robust and the sizes of the untranslated regions of transcripts compared to the corresponding coding regions are smaller in *Hpa* than in *A. thaliana*.

CEGMA pipeline

A subset of 458 core eukaryotic genes (CEGs) was additionally annotated using a local installation of the CEGMA pipeline (*16*). These genes are conserved across six eukaryotic species (*A. thaliana*, *Homo sapiens, Drosophila melanogaster, Caenorhabditis elegans, Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*).

Automated annotation of gene models

Predicted gene models were translated into protein sequences, and submitted to a semi-automated annotation pipeline. BLASTX, SignalP, TMHMM, InterProScan analysis were run, the results parsed using in-house scripts, and stored in a custom-built database.

Repeat analysis

Since the public repeat library for oomycete genomes is very limited, a *de novo* repeat family identifier and modeling package called Repeat Modeler (http://www.geospiza.com/media-v4/posters/PAG2005.pdf) was used to generate a repeat Library database. This resulted in the creation of an extensive, uncurated library of putative *Hpa* repeats. This database, along with repbase (*17*), was searched to generate repeats using RepeatMasker (*18*) (http://www.repeatmasker.org). Whole genome *de novo* repeat finder RepeatModeler was run over the genomes of *P. sojae* v4.0 (http://vmd.vbi.vt.edu) and *H. arabidopsidis* v8.3. RepeatModeler runs as a wrapper around three other de novo repeat finders, RECON (*19*), RepeatScout (*20*), and trfinder (*21*).

Metabolic Pathway Analysis.

Pathway annotation for *Hpa* was done using KAAS (KEGG automated annotation server). Predicted protein sequences were submitted to KAAS for assigning a KO (KEGG Orthology) identifier. The query sequences were blasted against the KEGG Genes reference database, with homologs selected on the basis of their BLAST score. Homologs above a certain threshold were identified as ortholog candidates based on BLAST score as well as bidirectional best hit information. Ortholog candidates were divided into KO groups according to the annotation of KEGG GENES database. Finally the assignment score was calculated based on likelihood and heuristics for each KO group. Then, the K ID of the KO group with the highest score was assigned to the query sequence. Once all KO IDs were assigned (essentially the gene products linked to the KEGG pathways), a pathway diagram was constructed. Pathway maps were generated choosing the non-organism specific option.

Mass spectrometry and proteomics

Intercellular washing fluid isolation. 2 ½ wk *old A. thaliana* accession Landsberg *erecta* plants were infected with *Hpa* isolate Cala2 by spray inoculation (75 spores/µl). At 3 days post inoculation plants were harvested and vacuum infiltrated with 0.3 M Mannitol. Infiltrated plants were washed three times with dH_2O and dried. Intercellular washing fluids (IWF) were retrieved by centrifugation at 1200 *x g* for 20 minutes at 4˚C. IWFs were dialyzed overnight at 4˚C against 5 L dH_2O and lyophilized. Protein was resolubilized in dH_2O and protein concentration was determined according to Bradford (*22*)

Protein identification**.** Proteins were separated by SDS-PAGE followed by excision of individual bands. Each band was *in gel* digested with trypsin and peptides were extracted. Tryptic peptides were analysed by nanoLC-MS-MS. Three independent replicas were analyzed. Spectra were assigned against the IPI *Arabidopsis* database v3.14 by the MASCOT® search algorithm (Matrix Science, Boston, MA, USA) and outputs were loaded into Scaffold® (proteome software, Portland, OR, USA). Spectra not assigned to *Arabidopsis* proteins were extracted and subsequently searched against translated databases of *Hpa* genome assembly v3.0 and predicted gene models using MASCOT®. Spectra assigned to genomic sequences with no corresponding gene model were annotated manually.

Phylogenetic analysis**.** EGL 12 endoglucanase domains and pectinesterase domains of all orthologs were identified according to interpro (http://www.ebi.ac.uk/InterProScan). Signal peptides were predicted by SignalP (http://_www.cbs.dtu.dk/services/SignalP). EGL 12 endoglucanase domains and pectinesterase domains of all orthologs were identified according to interpro (http://www.ebi.ac.uk/InterProScan). Sequences of NLPs from *P. sojae* and *P. ramorum* were downloaded from http://genome.jgi-psf.org/, using a e-20 compared to NIP1 from *P. sojae* as a cut-off. Pseudogenes and fragmented ORFs were removed, resulting in 10 genes from *Hpa*, 29 from *P. sojae*, and 40 from *P. ramorum*. Codon aligments were done using the RevTrans webserver (http://www.cbs.dtu.dk/services/RevTrans/), setting the alignment mode to dialign-T. Maximum Likelihood (ML) analysis was carried out using RAxML (*23*) on the webserver at CIPRIS (http://8ball.sdsc.edu:8889/cipres-web/Home.do) with likelihood search and estimation of invariable sites. All other parameters were set to default values. As the rapid bootstrapping algorithm (*24*) can lead to deviation in the range of several percent, the analysis was repeated five times and the bootstrap values obtained were averaged. Minimum Evolution (ME) analysis was done using Mega 4.0 (*25*), using the Tamura-Nei substitution model. All other parameters were set to default values. For testing tree robustness, 1000 bootstrap replicates were carried out. Maximum Parsimony (MP) analysis was carried out using Mega 4.0, with all parameters set to default values, again performing 1000 bootstrap replicates. For Bayesian analysis (BA), MrBAYES (*26*), version 3.1.2 was used. Four incrementally heated chains were run for 10 million generations implementing the GTR-I-G substitution model and 4 gamma categories, with a sampling frequency of 1000. The first 5000 trees were discarded, and the remaining 5000 trees were used to compute a majority rule consensus tree to reveal the posterior probabilities.

Synteny analysis

Synteny between *Hpa*, *P. sojae*, *P. ramorum*, and *P. infestans* was determined using the algorithm PHRINGE (http://oomycetes-public.genomeprojectsolutions-databases.com; (*27*).

2. ADDITIONAL RESULTS AND DISCUSSION

Endoglucanase 12 and Pectinase genes

Mass spectrometry confirmed that *HaEGL12-1*, *HaEGL12-2*, and *HaPect1* proteins are abundant in apoplastic fluid extracted from *Arabidopsis* leaves colonized by *Hpa* (Figs. S3 & S4). Accordingly, quantitative RT-PCR demonstrated that *HaEGL12-1*, *HaEGL12-2*, and *HaPect1* are induced during infection, while *HaEGL12-3*, *HaPect2*, and *HaPect3* are not (Fig. S5). The remaining *Hpa* enzymes are highly divergent from the orthologous genes in *P. sojae* and *P. ramorum* and showed a high degree of substitution pattern heterogeneity compared to most other pettiness (Fig. S8).

Zoospore-associated genes

The genes associated with zoospore structure are all unlinked in *Phytophthora*, so a single genome rearrangement would be insufficient to explain their loss in *Hpa*. Instead, such genes likely degenerated to the point where their detection is no longer possible. This resembles the process of evolution proposed for bacterial symbionts having reduced genomes, in which mutations first converted genes to pseudogenes, followed by their total disintegration (*28*). To further explore the mechanism of gene loss in *Hpa,* we assessed whether genes flanking a flagellar gene in *P. infestans* had maintained linkage in *Hpa*. Linkage was preserved in only a few instances, and in those cases no similarity was detected between the flagellar gene from *P. infestans* and sequences located between the flanking genes in *Hpa*. Moreover, these regions were reduced in size in *Hpa* compared to *P. infestans.* This suggests that the mechanism of gene loss likely involved a combination of base changes and small deletions.

Missing genes for nitrate and sulfate assimilation

The genes for nitrate, nitrite and sulfite reductases could not be found (Fig. S10A; Table S3) despite the 9.5-fold coverage of Sanger genome sequence, 46-fold coverage of Illumina genome sequence, 36,663 Sanger ESTs and 61,327 454 ESTs. It is theoretically possible that the three genes nevertheless escaped sequencing, but it is far more likely that they are simply missing from the *Hpa* genome. This supported in the case of the nitrate and nitrite reductases by the fact that those two genes are adjacent in the *Phytophthora* genome sequences whereas the syntenic region in the *Hpa* genome is simply missing the two genes, together with an adjacent nitrate transporter (Fig. S10B). The same triplet of nitrate assimilation genes also occurs as a cluster in saprophytic fungi but is deleted in the obligately parasitic rust fungi *Melampsora populina-larici* and *Puccinia graminis f.sp. tritici* (F. Martin, personal communication) and the obligately parasitic powdery mildew fungi *Blumeria graminis*, *Erysiphe pisi*, and *Golovinomyces orontii* (P. Spanu, personal communication). *Hpa* presumably retains the phosphoadenosine phosphosulfate reductase to recycle ADP from excess phosphoadenosine phosphosulfate not required for sulfation of peptides, lipids and carbohydrates. The rust fungi are likewise missing genes for sulfite reductase while retaining genes for phosphoadenosine phosphosulfate reductase.

3. SUPPLEMENTARY FIGURES

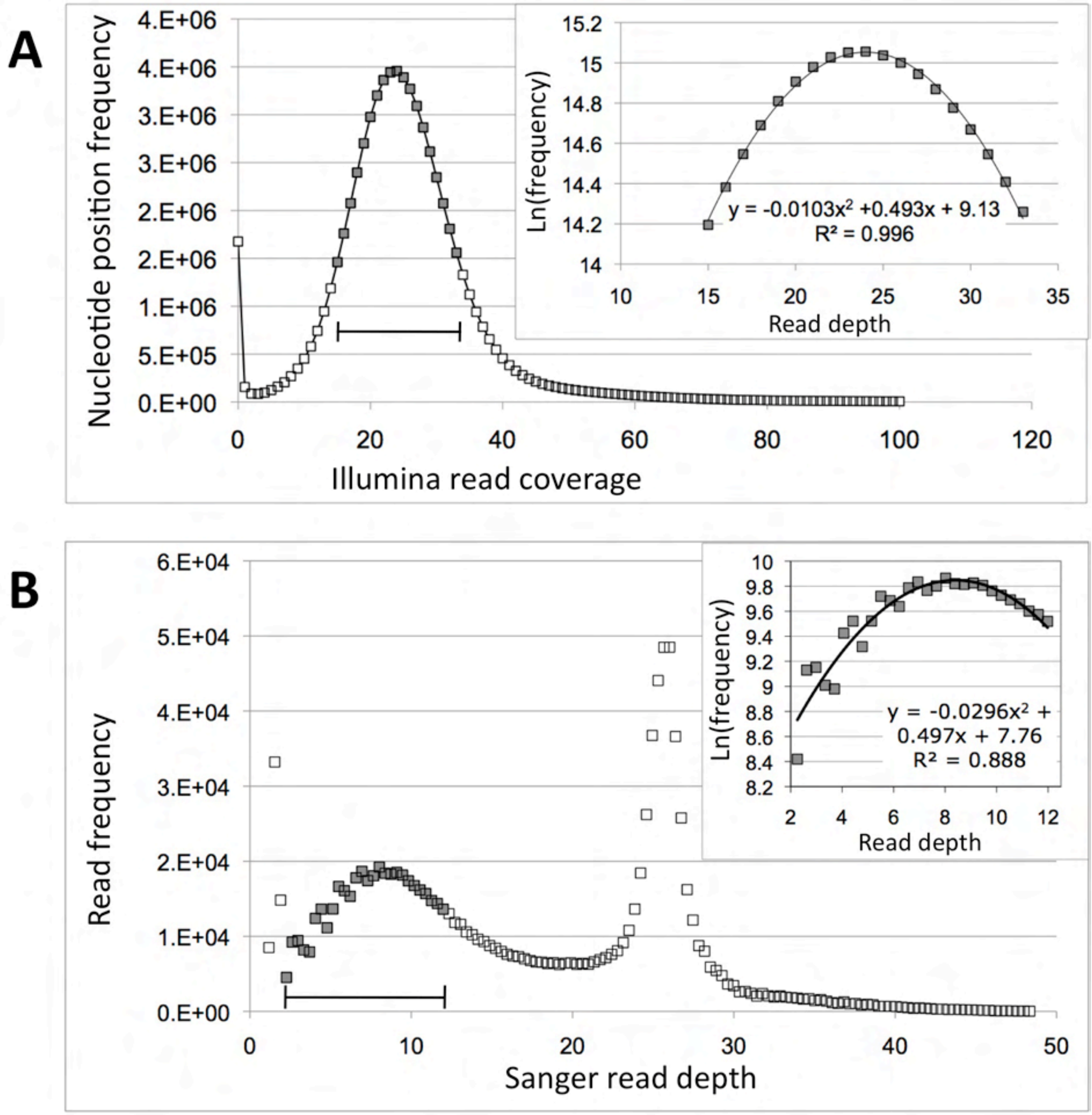


Fig. S1. Genome size estimation from Illumina and Sanger read coverage

(A) Frequency of nucleotide positions in the Sanger assembly with given Illumina read coverage. (B) Frequency of Sanger reads with given Sanger read coverage. In both (A) and (B), to obtain the mean coverage of the single copy sequences, a Gaussian curve was fitted to the main peak (indicated by shaded points and horizontal bar) by fitting a quadratic function to the natural-log-transformed frequency data (inset). From each fitted quadratic function $ax^2 + bx + c$, the mean of each Gaussian distribution was obtained as –b/2a.

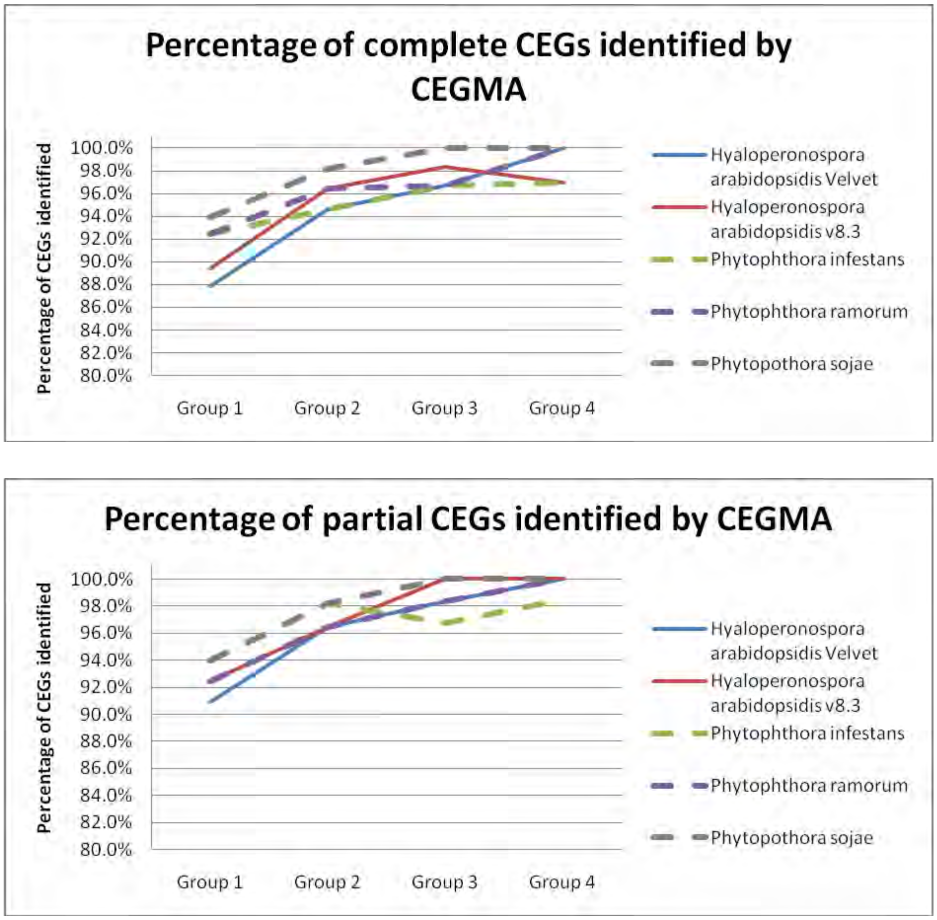
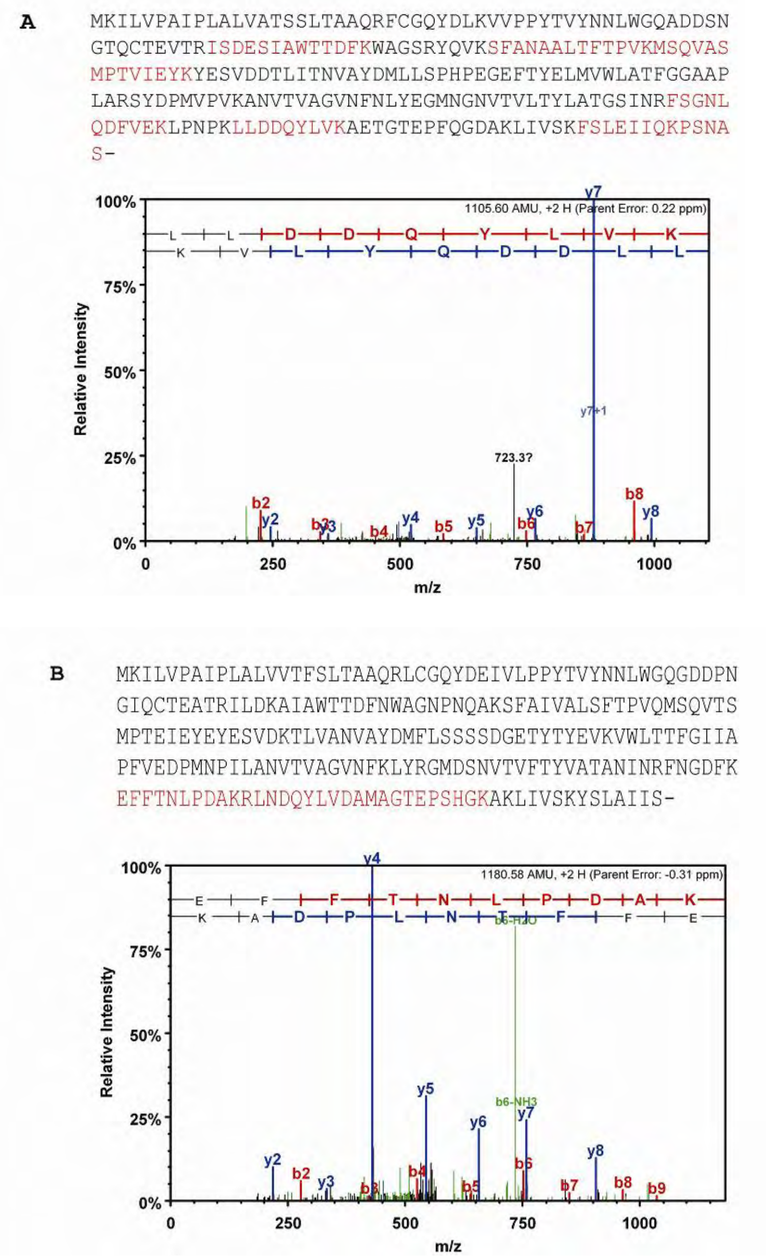


Fig. S2. Identification of single copy core eukaryotic orthologous genes (CEGs) by the CEGMA pipeline

~95% of the CEGs were identified in the *Hpa* Emoy2 Velvet and v8.3 assemblies. This is comparable to the number of CEGs identified in *P. infestans* (95%), *P. ramorum* (96%) and *P. sojae* (98%). There were 4 CEGs partially identified in the Velvet, v8.3 and *P. infestans* assemblies, and one partially identified in *P. ramorum*.

The CEGs are split into 4 groups with Group 1 being the least conserved between organisms, and Group 4 being the most conserved between organisms.



A

```
MKILVPAIPLALVATSSLTAAQRFCGQYDLKVVPPYTVYNNLWGQADDSN
GTQCTEVTRISDESIAWTTDFKWAGSRYQVKSFANAALTFTPVKMSQVAS
MPTVIEYKYESVDDTLITNVAYDMLLSPHPEGEFTYELMVWLATFGGAAP
LARSYDPMVPVKANVTVAGVNFNLYEGMNGNVTVLTYLATGSINRFSGNL
QDFVEKLPNPKLLDDQYLVKAETGTEPFQGDAKLIVSKFSLEIIQKPSNA
S-
```

B

```
MKILVPAIPLALVVTFSLTAAQRLCGQYDEIVLPPYTVYNNLWGQGDDPN
GIQCTEATRILDKAIAWTTDFNWAGNPNQAKSFAIVALSFTPVQMSQVTS
MPTEIEYEYESVDKTLVANVAYDMFLSSSSDGETYTYEVKVWLTTFGIIA
PFVEDPMNPILANVTVAGVNFKLYRGMDSNVTVFTYVATANINRFNGDFK
EFFTNLPDAKRLNDQYLVDAMAGTEPSHGKAKLIVSKYSLAIIS-
```

***Fig. S3. Sequences of* H. arabidopsidis *extracellular glucanases identified by mass spectrometry of proteins extracted from intracellular spaces of* Arabidopsis *leaves infected with* H. arabidopsidis**

(A) *HaEGL12*-1 (VMD ID #808599); **(B)** *HaEGL12-2* (VMD ID #814377). All identified peptides are highlighted in red. The annotated MS/MS spectrum is that of peptide LLDDQYLVK. Red and blue highlighted amino acids represent identified B- and Y-ions respectively.

A

```
MKIFVPSAVALISIMTMIVANAITCSGPHARVEPPAGALVVDAAANPKYK
NSFRTLAGAVNKLDLSSGNQQTIFILSGLYKEKVTIPFLNGPLVLQGATC
DATSYAKNQVTIAQATAQKNLPNDVTNDRNALTSTVLFKSNNVKVYNLNI
ANTAGNVGQAVAVTVDGENYGFYGCDLRGYQDTLLTKKGKQLYAKSRITG
AVDFIFGLEAAVWCERCDIESIGAGCITANGRSSNGSNSYYVFNHARVYG
SKGSLVGKTFLGRPWRPYARVVFQNSELSNVVNAAGWSKWNGQSPDHVHF
REFKNVGPGAAKRERAAFSQQLTQAVSINQILGDNYKTQSWVDLAYL-
```

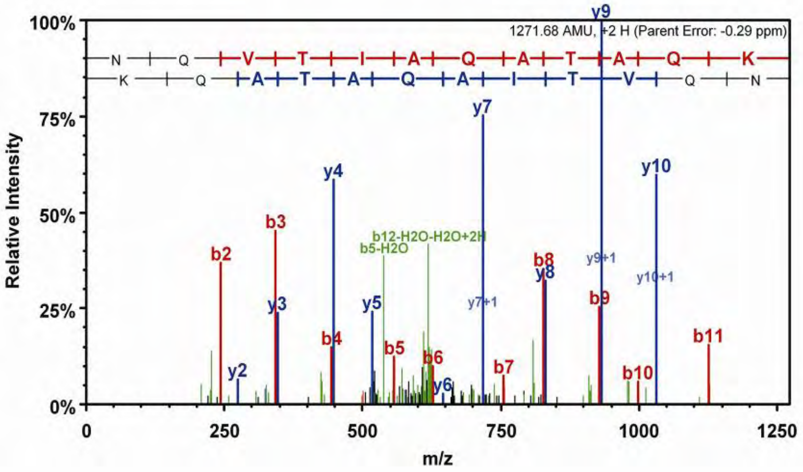


***Fig. S4. Sequence of HaPect1 protein identified by mass spectrometry of proteins extracted from intracellular spaces of* Arabidopsis *leaves infected with* H. arabidopsidis**

All identified peptides are highlighted in red. The annotated MS/MS spectrum is that of peptide NQVTIAQATAQK. Red and blue highlighted amino acids represent identified B- and Y-ions respectively.

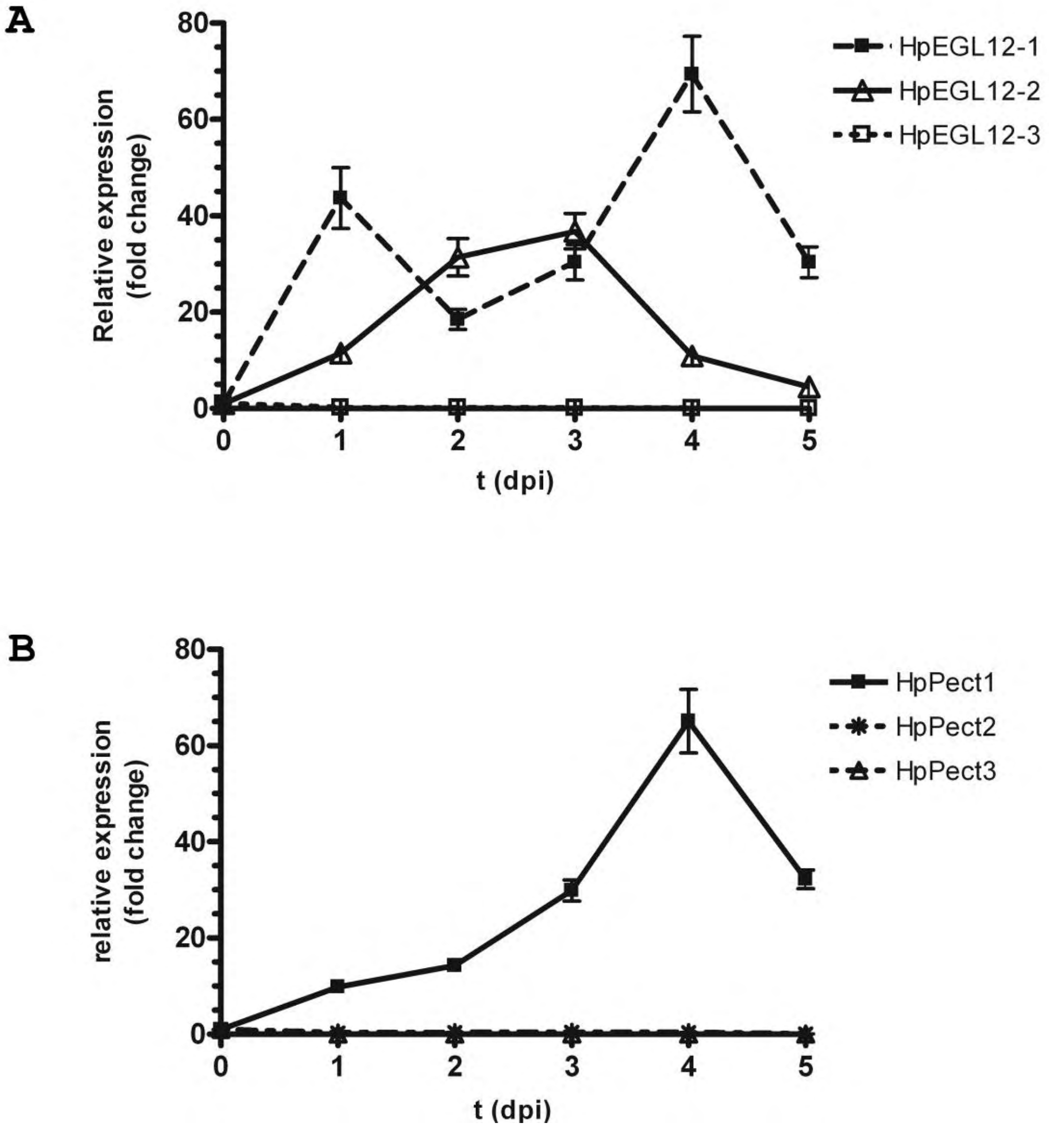


***Fig. S5. Abundance of transcripts from HaEGL12 and HaPect genes during infection of* Arabidopsis *by* H. arabidopsidis**

Expression was normalized against *Hpa* Actin2 and expression at t= 0 dpi (days post infection) is set at 1.0. **(A)** *HpEGL12-1*, -2, and -3. **(B)** *HpPect1*, -2, and -3.

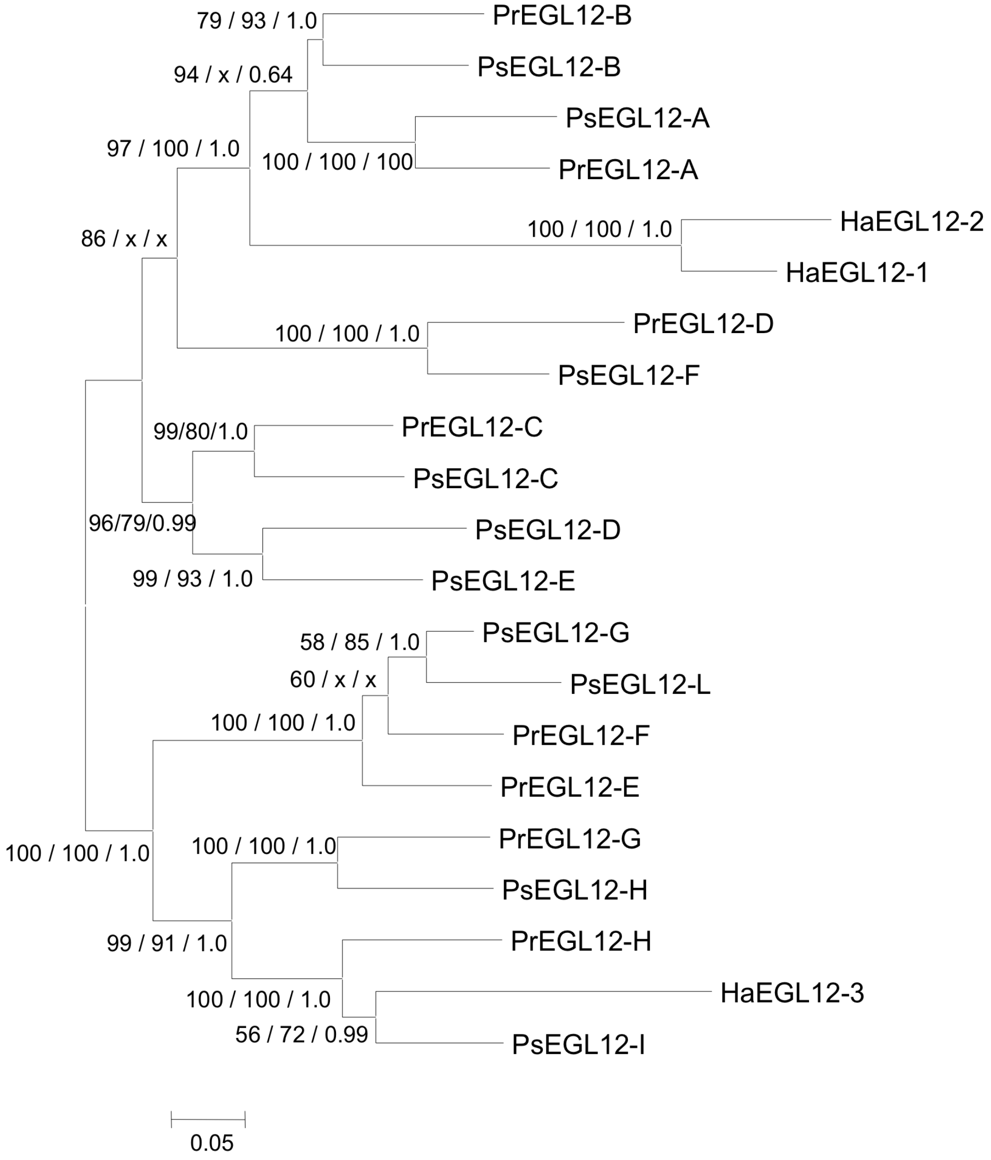


***Fig. S6. Minimum Evolution phylogenetic reconstruction of GH12 endoglucanase genes from* H. arabidopsidis (Ha)*,* P. sojae (Ps)*, and* P. ramorum (Pr)**

Values on branches are support in Minimum Evolution, Maximum likelihood and Bayesian analysis, respectively; x denotes conflicting topology.

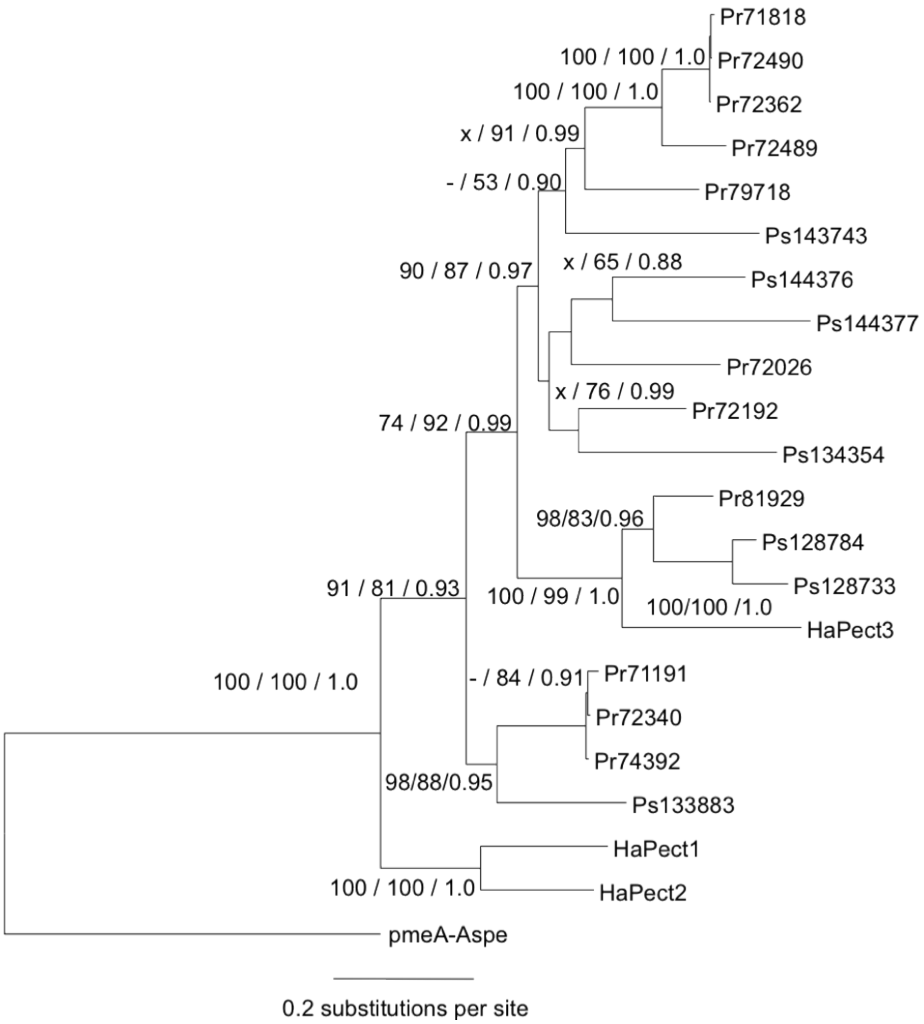
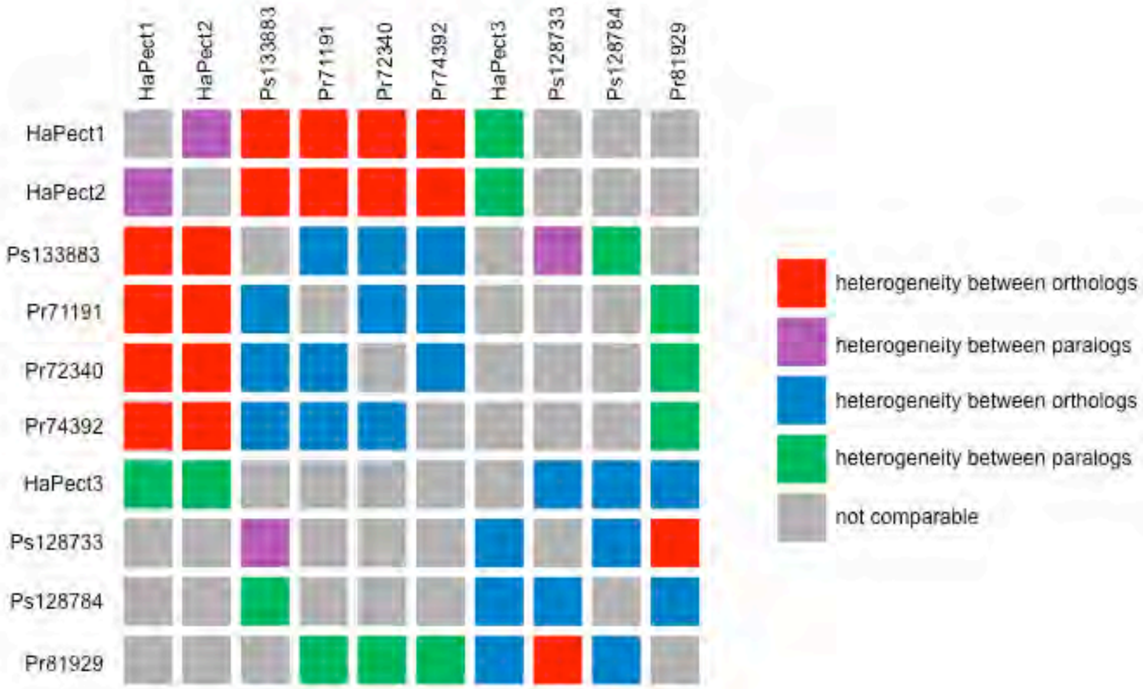


Fig. S7. Minimum Evolution phylogenetic reconstruction of pectin methyl esterase genes from H. arabidopsidis (Ha), P. sojae (Ps), _and_ P. ramorum (Pr)

Values on branches are support in Minimum Evolution, Maximum likelihood and Bayesian analysis, respectively, x denotes conflicting topology.



	HaPect1	HaPect2	Ps133883	Pr71191	Pr72340	Pr74392	HaPect3	Ps128733	Ps128784	Pr81929
HaPect1	⬜	🟪	🟥	🟥	🟥	🟥	🟩	⬜	⬜	⬜
HaPect2	🟪	⬜	🟥	🟥	🟥	🟥	🟩	⬜	⬜	⬜
Ps133883	🟥	🟥	⬜	🟦	🟦	🟦	⬜	🟪	🟩	⬜
Pr71191	🟥	🟥	🟦	⬜	🟦	🟦	⬜	⬜	⬜	🟩
Pr72340	🟥	🟥	🟦	🟦	⬜	🟦	⬜	⬜	⬜	🟩
Pr74392	🟥	🟥	🟦	🟦	🟦	⬜	⬜	⬜	⬜	🟩
HaPect3	🟩	🟩	⬜	⬜	⬜	⬜	⬜	🟦	🟦	🟦
Ps128733	⬜	⬜	🟪	⬜	⬜	⬜	🟦	⬜	🟦	🟥
Ps128784	⬜	⬜	🟩	⬜	⬜	⬜	🟦	🟦	⬜	🟦
Pr81929	⬜	⬜	⬜	🟩	🟩	🟩	🟦	🟥	🟦	⬜

🟥 heterogeneity between orthologs

🟪 heterogeneity between paralogs

🟦 heterogeneity between orthologs

🟩 heterogeneity between paralogs

⬜ not comparable

Fig. S8. Codon substitution pattern heterogeneity for orthologs of H. arabidopsidis pectin methyl esterase genes (two orthologous gene groups).

Substitution pattern heterogeneity was computed using MEGA4.0 (codon translation, complete deletion). Significance threshold was set to p = 0.05

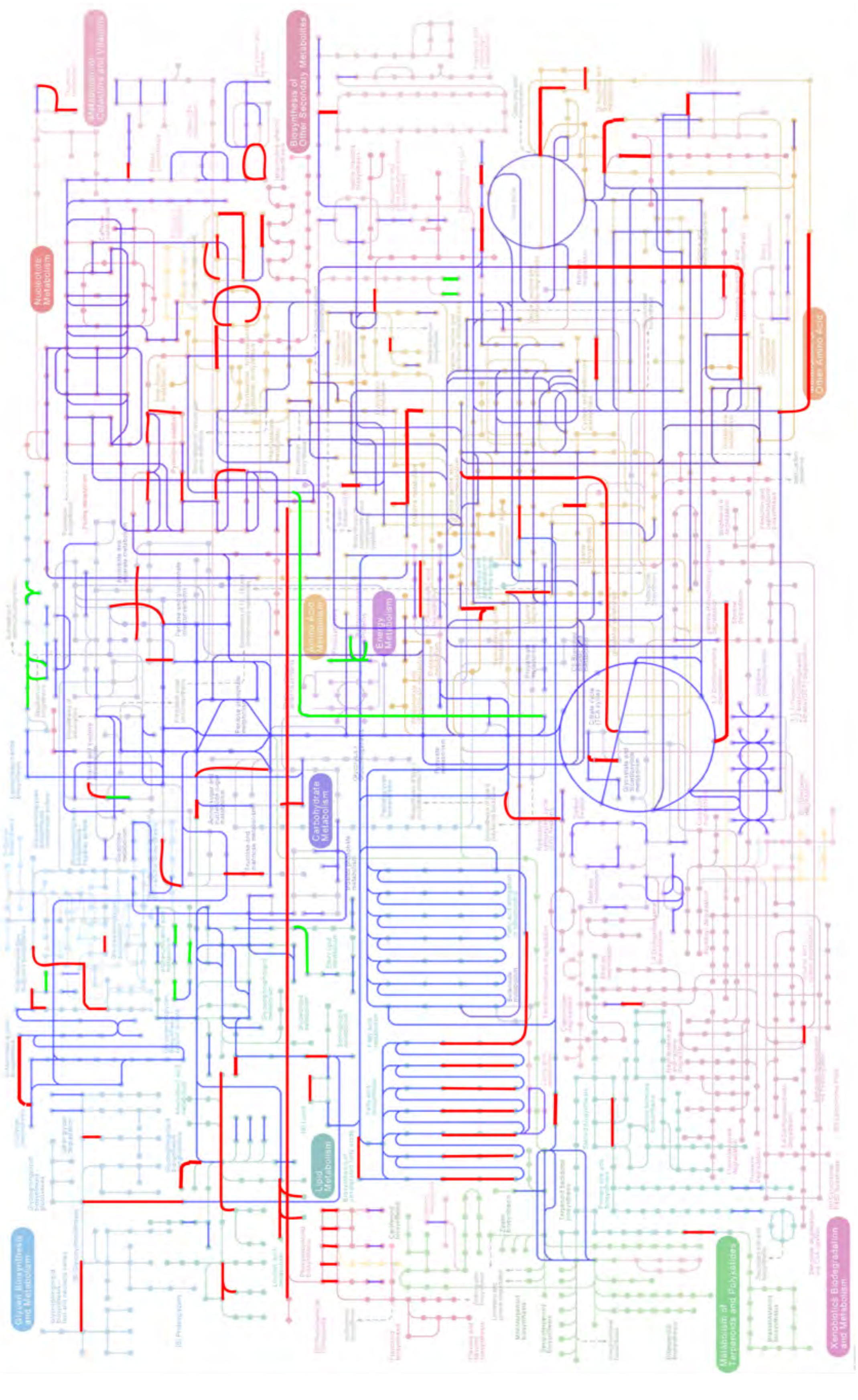

Glycan Biosynthesis and Metabolism
Metabolism of Cofactors and Vitamins
Biosynthesis of Other Secondary Metabolites
Nucleotide Metabolism
Other Amino Acid
Amino Acid Metabolism
Energy Metabolism
Carbohydrate Metabolism
Citrate cycle (TCA cycle)
Lipid Metabolism
Metabolism of Terpenoids and Polyketides
Xenobiotics Biodegradation and Metabolism

***Fig. S9. Metabolic pathways (KEGG Atlas Mapping) in* H. arabidopsidis (Hpa), P. infestans, P. sojae *and* P. ramorum**

Gene products (circles) are joined by enzyme-catalyzed pathways (lines). Components highlighted in blue are present in *Hpa* and at least one *Phytophthora* species. Components in red are absent in *Hpa* but present in at least 2 out of three *Phytophthora* species. Components in green are present in *Hpa* but are absent in the *Phytophthora* species. Greyed circles/lines indicate components absent in all four species.

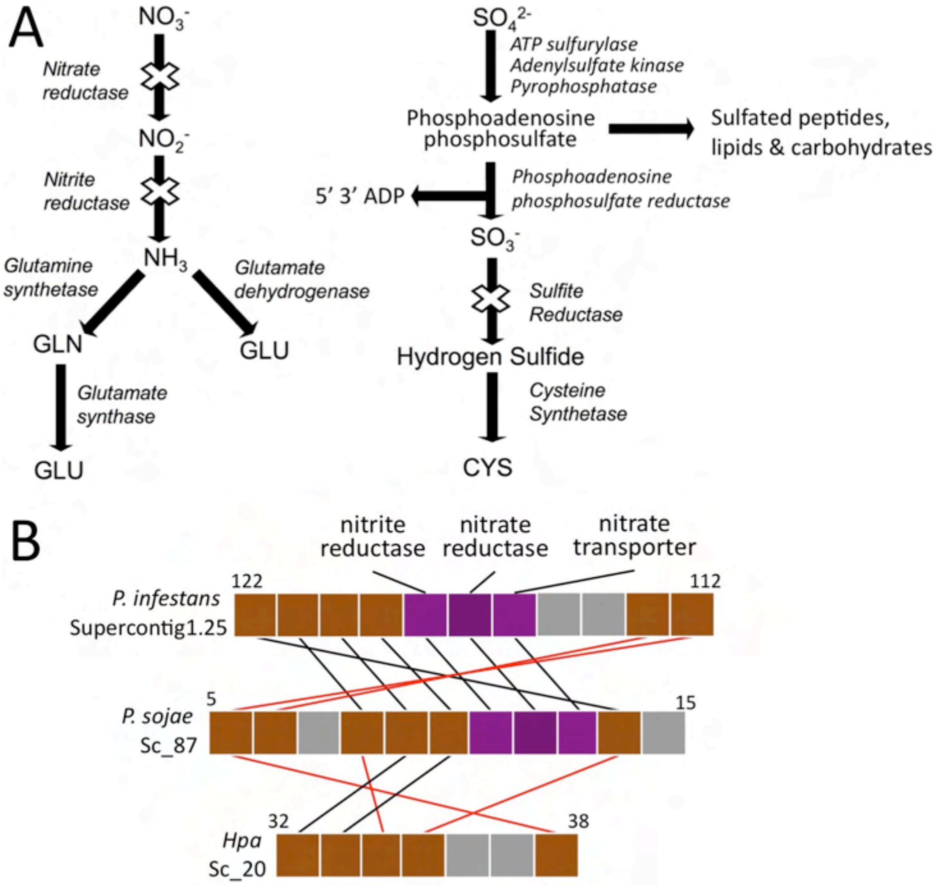


Fig. S10. Genes for nitrate and sulfate metabolism are present in **Phytophthora** ***but absent in*** **H. arabidopsidis.**

(A) Pathways for assimilation of N and S. Enzyme names are italicized. All enzymes shown are present in one or two copies in *Phytophthora sojae*, *P. ramorum*, and *P. infestans*. White crosses designate genes that are absent in *Hpa*. Gene IDs for all species are provided in Table S3.

(B) Nitrate and nitrite reductase genes are adjacent in *Phytophthora* genomes but missing from the syntenic location in the Hpa genome. Regions shown are from *P. sojae* version 1.1, *P. infestans* version 1.0 and *Hpa* version 6. Hpa region spans spanning scaffold_20:236914-269510 (v6) and SuperContig73:223284-255880 (v8.3.2) Colored boxes show order of gene models. Non-coding DNA is not represented. Dark brown, orthologs; gray, non-conserved genes; purple, cluster of nitrate assimilation genes specific to *Phytophthora* genomes. Black lines join syntenic genes with the same orientation; red lines join genes with reversed orientations.

4. SUPPLEMENTARY TABLES

***Table S1. Repeat elements in the genomes of* H. arabidopsidis *and* P. sojae**

Repeat Elements	Percentage (Hpa)	Number of Bases (Hpa)	Percentage (Ps)	Number of Bases (Ps)
rRNA	0.03%	23939	-	-
DNA/Harbinger	0.03%	29980	-	-
DNA/TcMar-	0.06%	49711	-	-
LINE/RTE-BovB	0.07%	57545	-	-
DNA/TcMar-Fot1	0.07%	55487	-	-
DNA/hAT-Ac	0.08%	67429	-	-
DNA/hAT-Charlie	-	-	0.04%	27294
DNA/hAT-Tag1	-	-	0.10%	73899
DNA/hAT-Tip100	-	-	0.05%	40226
Low_complexity	0.09%	72015	0.12%	92418
DNA/TcMar-Tc2	0.08%	65004	-	-
LINE/L1	0.16%	137070	0.12%	92735
DNA/TcMar-Tc1	0.20%	166679	-	-
DNA	0.23%	193196	-	-
DNA/Maverick	0.31%	259632	1.19%	921034
RC/Helitron	0.31%	256711	-	-
LTR/Gypsy-Cigr	0.37%	310080	-	-
Simple_repeat	0.38%	313567	0.46%	358990
SINE	-	-	0.18%	137266
DNA/MuDR	0.44%	362138	1.0%	587625
LINE/L1-Tx1	0.48%	400228	0.07%	56638
LINE/R1	0.50%	414864	-	-
LINE/telomeric	0.53%	441344	-	-
LINE/Penelope	-	-	0.05%	42415
DNA/TcMar-Pogo	0.89%	737215	0.09%	73683
LTR/Ngaro			0.15%	115201
LTR/Copia	3.60%	3000437	1.99%	1542992
LTR/Gypsy	16.5%	13563599	6.60%	5120500
Unknown Repeats	17.7 %	14531071	12.51%	9711997
Other	56.72%	46542701	75.53%	58621816
Total	100%	82051642	100%	77616729

Table S2. Copy numbers of annotated* H. arabidopsidis *genes for phospholipid signalling enzymes, secondary metabolite biosynthesis, and ABC transporters, compared to* Phytophthora *genomes.

Gene product	***H. arabidopsidis***	***P. sojae***	***P. ramorum***
CORE SET PHOSPHOLIPID SIGNALING ENZYMES	40	59	59
SECONDARY METABOLITE BIOSYNTHESIS			
Nonribosomal peptide synthetases	1	4	4
Polyketide synthases	1	1	1
Cytochrome P450's	10	30	24
ABC TRANSPORTERS			
Total	68	140	135
PDR (ABCG-full)	13	51	48
ABCG-half	12	22	21
MDR (ABCB)	5	9	7
MRP (ABCC)	9	21	22

Table S3. ***Genes IDs for nitrogen and sulphur assimilation enzymes in* Phytophthora *species and* H. arabidopsidis**

	P. sojae[1]	*P. ramorum*[1]	*P. infestans*[2]	*H. arabidopsidis*[1]
Nitrate reductase	Ps140563	Pr71442	PITG_13012.1	none
Nitrite reductase	Ps140562	Pr76696	PITG_13013.1	none
Nitrate transporter[3]	Ps140564	Pr76698	PITG_13011.1	none
Glutamine synthetase	Ps109140 Ps109139	Pr72153 Pr72154	PITG_14180.1 PITG_14179.1	Ha802420
Glutamate synthase (NADH)	Ps135530	Pr72102	PITG_07380.1	Ha805196
Glutamate synthase (Ferridoxin)	Ps130831	Pr78125	PITG_12037.1 PITG_16280.1	Ha812981
Glutamate dehydrogenase	Ps108919	Pr71959	PITG_07671.1	Ha805610; Ha806617
ATP sulfurylase Adenylsulfate kinase Pyrophosphatase	Ps112102	Pr79353	PITG_04010.1	Ha813786
Phosphoadenosine phosphosulfate reductase	Ps156997	Pr74880	PITG_04601.1	Ha809449
Sulfite reductase	Ps139493 Ps139488	Pr71878 Pr81882	PITG_19263.1 PITG_18187.1	none
Cysteine synthetase	Ps109172 Ps109175	Pr71225 Pr71224	PITG_12727.1 PITG_12725.1	Ha814750

[1] Gene IDs from the VBI Microbial Database vmd.vbi.vt.edu.
[2] Gene IDs from the Broad Institute Database http://www.broadinstitute.org
[3] Annotation is by similarity (Blastp = 2E-84) to experimentally characterized *Porphyra yezoensis* nitrate transporter (BAG70346.1)

5. CURRENT AND OTHER AFFILIATIONS

Current

Baxter, Laura: Systems Biology Centre, Warwick University, Coventry, CV4 7AL, UK

Bittner-Eddy, Peter: Department of Microbiology, University of Minnesota, Minneapolis MN, 55455, USA

Chibucos, Marcus: Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, MD 21201

Downton, Polly: Systems Biology Doctoral Training Centre, University of Warwick, Coventry, CV4 7AL, UK

Rodgers, Jane: The Genome Analysis Centre, The Genome Analysis Centre, Norwich Research Park, Colney, Norwich, NR4 7UH,UK.

Rougon-Cadoso, Alejandra: Laboratorio Nacional de Genómica para la Biodiversidad, CINVESTAV Irapuato, Mexico

Ryden, Peter: Institute of Food Research, Colney, Norwich, NR4 7UA U.K.

Studholme, David: Biosciences, College of Life and Environmental Sciences University of Exeter EX4 4QD, UK

Other

Jones, Jonathan: Founder and Science Advisory Board member of Mendel Biotechnology Inc., 3935 Point Eden Way Hayward, California 94545-3720USA

6. ADDITIONAL SOURCES OF SUPPORT

Boot, Nico: as for van den Ackerveken

Cabral, Adriana: as for van den Ackerveken

Govers, Francine: Centre for BioSystems Genomics NGI/NWO, The Netherlands.

Jiang, Rays: as for Govers

Judelson, Howard: United States Department of Agriculture-National Institute of Food and Agriculture

Kamoun, Sophien: Gatsby Charitable Foundation

Meijer, Harold: Dutch Technology Foundation STW, which is the applied science division of NWO, and the Technology Programme of the Ministry of Economic Affairs in the framework of a VIDI project

Rogers, Jane: Wellcome Trust

Rougon, Alejandra: Consejo Nacional de Ciencia y Tecnología (National Board of Science and Technology), Mexico

Thines, Marco: Research funding program LOEWE of the Ministry of Science, Research and the Arts of Hessen, as well as from the Max-Planck Society.

van den Ackerveken, Guido: VIDI innovational research grant from the Netherlands Organization for Scientific Research (NWO), by the European Research Area-Network (ERA-NET) Plant Genomics and by the Centre for BioSystems Genomics (CBSG), which is part of the Netherlands Genomics Initiative and NWO.

7. REFERENCES CITED

1. A. P. Rehmany *et al.*, *Fungal Genet Biol* 30, 95 (2000).
2. J. Wild, Z. Hradecna, W. Szybalski, *Genome Research* 12, 1434 (2002).
3. A. P. Rehmany *et al.*, *Fungal Genet Biol* 38, 33 (2003).
4. C. Soderlund, S. Humphray, A. Dunham, L. French, *Genome Research* 10, 1772 (2000).
5. X. Huang *et al.*, *Genome Research* 13, 2164 (2003).
6. D. R. Zerbino, E. Birney, *Genome Res* 18, 821 (2008).
7. S. Ossowski *et al.*, *Genome Res* 18, 2024 (2008).
8. S. Kurtz *et al.*, *Genome Biol* 5, R12 (2004).
9. J. Parkinson *et al.*, *Bioinformatics* 20, 1398 (2004).
10. J. L. White, J. M. Kaper, *J Virol Methods* 23, 83 (1989).
11. S. Bashiardes *et al.*, *Nat Methods* 2, 63 (2005).
12. M. Stanke, O. Schoffmann, B. Morgenstern, S. Waack, *BMC Bioinformatics* 7, 62 (2006).
13. B. J. Haas *et al.*, *Nucleic Acids Res* 31, 5654 (2003).
14. W. H. Majoros, M. Pertea, S. L. Salzberg, *Bioinformatics* 20, 2878 (2004).
15. I. Korf, *BMC Bioinformatics* 5, 59 (2004).
16. G. Parra, K. Bradnam, I. Korf, *Bioinformatics* 23, 1061 (2007).
17. J. Jurka *et al.*, *Cytogenet Genome Res* 110, 462 (2005).
18. A. F. A. Smit, R. Hubley, P. Green. *RepeatMasker Open-3.0*. (1996-2004).
19. Z. Bao, S. R. Eddy, *Genome Res* 12, 1269 (2002).
20. A. L. Price, N. C. Jones, P. A. Pevzner, *Bioinformatics* 21 Suppl 1, i351 (2005).
21. G. Benson, *Nucleic Acids Res* 27, 573 (1999).
22. M. M. Bradford, *Anal Biochem* 72, 248 (1976).
23. A. Stamatakis, *Bioinformatics* 22, 2688 (2006).
24. A. Stamatakis, P. Hoover, J. Rougemont, *Syst Biol* 57, 758 (2008).
25. K. Tamura, J. Dudley, M. Nei, S. Kumar, *Mol Biol Evol* 24, 1596 (2007).
26. J. P. Huelsenbeck, F. Ronquist, *Bioinformatics* 17, 754 (2001).
27. P. S. Dehal, J. L. Boore, *BMC Bioinformatics* 7, 201 (2006).
28. F. J. Silva, A. Latorre, A. Moya, *Trends in Genetics* 17, 615 (2001).

8. LIST OF 134 HIGH-CONFIDENCE RXLR EFFECTOR CANDIDATE GENES

The following table lists 134 predicted RXLR genes/proteins that are considered “high-confidence” candidates, based on bioinformatic criteria. The standardized names are to be used by the general community. Coordinates of each gene in v8.3 are listed.