

1. Selection of the *P. infestans* strain to be sequenced
2. Isolation of *P. infestans* genomic DNA
3. Genome sequencing and assembly
4. *P. infestans* genome annotation
 - 4.1 The *P. infestans* gene annotation pipeline
 - 4.2 Gene prediction using Orthosearch
 - 4.3 Targeted annotation of specific gene families and pseudogenes
 - 4.4 Annotation of the *Phytophthora* Crinkler protein family and domain structures
 - 4.5 Annotation of RXLR effectors
 - 4.6 Analysis of protein-coding genes
5. Analysis of repeats
 - 5.1 Repeat identification and curation of mobile element reference sequences
 - 5.2 Estimation of autonomous helitron abundance among diverse genomes
6. Computation of blocks of conserved gene order among *Phytophthora* genomes
7. Functional analyses of Crinkler genes and C-terminal domains
 - 7.1 Plant growth conditions
 - 7.2 Plasmid construction
 - 7.3 Potato virus X (PVX) agroinfection assays
 - 7.4 CRN2 deletion analyses reveal a C-terminal effector domain
 - 7.5 Phenotypic characterization of diverse CRN C-terminal domains *in planta*
8. Generation of an optical map for the *P. infestans* genome

- 8.1 Genomic DNA preparation for optical mapping
 - 8.2 Optical mapping of *P. infestans* genome
 - 8.3 Map comparison between optical maps and sequence scaffolds
9. Generation and analysis of a Full-Length Normalized cDNA library of *P. infestans*
10. Analysis of gene expression using NimbleGen microarrays

Supplementary Tables

Table S1: Protein-coding gene structure statistics

Table S4: Albatross vs. Pi-1 Gypsy LTR retrotransposon clades and contribution to genome size.

Table S5: Estimate of the abundance of autonomous helitrons across diverse eukaryotes

Table S6: Protein families implicated in plant pathogenesis

Table S-A2: Predictions of RXLR genes according to method

Table S-A3: Sources of external genomic data

Additional supplemental tables and data sets are provided as additional files.

1. Selection of the *P. infestans* strain to be sequenced

We selected the *P. infestans* isolate T30-4 for sequencing because it was the strain used for construction of the existing bacterial artificial chromosome (BAC) library¹. T30-4 is one of the offspring of the cross that was used to build the existing AFLP-based genetic map for *P. infestans*^{2,3}. Out of the cross offspring, T30-4 had been selected for the BAC library because it contained the avirulence allele for every known virulence/avirulence locus segregating in the cross⁴. This choice allows us to capture the maximum number of avirulence genes in the reference sequence, and also allows for the possibility of correlating the sequence with the genetic map. However, during analysis of the data, we noted that the sequenced isolate, from the BAC library stock (now designated T30-4(SW)), differs from the original cross isolate (designated T30-4(FG)). Both isolates are offspring of the same cross, and the sequenced genome does contain the avirulence allele for all segregating virulence/avirulence genes that have been cloned.

2. Isolation of *P. infestans* Genomic DNA

The *P. infestans* isolate T30-4 was grown for two weeks in liquid Plich medium² at 18°C in the dark. Mycelia were washed with sterile H₂O, snap frozen in liquid N₂ and freeze dried. Genomic DNA was extracted using a modified method of⁵. Mycelia (~100 mg) were ground to a fine powder under liquid nitrogen, mixed with 10 mL of extraction buffer (0.2 M Tris pH8.5, 0.25 M NaCl, 25 mM EDTA, 0.5% SDS), 7 mL Tris-equilibrated phenol and 3 mL of chloroform, incubated at room temperature for 1 hour, then centrifuged at 6,000g for 15 min. The aqueous phase was extracted with an equal volume of chloroform and centrifuged at 6,000g for 15 min. Isopropanol (0.6 volumes) was added to the aqueous phase, mixed gently, and then precipitated on ice for 30 min. The DNA was centrifuged at 6,000g for 15 min, washed with 70% ethanol, dried and resuspended in 700 µL of H₂O.

3. Genome sequencing and assembly

Based on low coverage pilot whole genome shotgun sequencing and complete sequencing of 10 BACs, we expected the *P. infestans* genome to be highly repetitive. Because this can cause problems for assembly, we opted for a strategy that involved generating a large fraction of our sequence coverage from clones that provided high physical coverage (long inserts). This strategy allowed us to achieve long-range connectivity of scaffolds (contigs ordered and oriented but including gaps) even if long stretches of sequence could not be assembled due to their repetitiveness.

Whole genome shotgun was performed in libraries of 4 types: 4 kb plasmids, 10 kb plasmids, 40 kb Fosmids, and 80 kb BACs. Contribution of data from each library type is given in Supplementary Table S-A1.

Whole genome assembly was performed using Arachne⁶. The total length of the assembled contigs is 190,133,476 bases (b) in 18288 contigs. The contig N50 is 44,484 b with 899 contigs of this size or larger (N50 is the weighted median contig size, computed as the size of the contig for which 50% of bases are assembled in contigs of that size or larger). The total length of scaffolds (including spanned gaps) is 228,543,505 b, in 4,921 scaffolds. This is the total length of scaffolds including Ns. Some sequence gaps are estimated by spanning read pairs to have very small or even negative size (undetected overlap) but are represented by longer runs of Ns to cleanly separate the contigs. The estimated total length of the assembled scaffolds is only 225,599,947 b. The N50 scaffold is 1,570,906 b, with 38 scaffolds of this size or larger. Scaffolds of 200 kilobases (kb) or larger contain 79.4% of all assembled bases.

4. *P. infestans* genome annotation

4.1 The *P. infestans* gene annotation pipeline

Protein-coding gene structures were initially automatically annotated using a combination of gene models predicted from the gene finding programs Orthosearch (see Supplemental Information section 5.2) and GeneID⁷. Among the overlapping sets of gene predictions,

our automated annotation method uses heuristics such as splice agreement with ESTs and relative overlap with the BLAST hits to choose the prediction most in accord with the evidence. At loci where supporting protein or EST homology was lacking and both Orthosearch and GeneID predicted genes, the Orthosearch cross-genome homology based models were favored over the GeneID *ab initio* predictions due to the overall better accuracy of Orthosearch in calling *P. infestans* genes in comparison to GeneID. Genes encoding proteins found homologous to known mobile element encoded proteins, or exhibiting evidence of transposition and found to encode novel mobile elements, were excluded from the gene set, with few exceptions. Targeted annotation methods for specific gene families, including RXLR and Crinkler families and pseudogenes are described below. RNA genes were predicted using both tRNA-scanSE⁸ and Rfam⁹.

4.2 Gene prediction using Orthosearch

Orthosearch is a comparative gene caller that works by identifying the set of valid gene models (those with intact ORFs and valid splicing) that maximize a score based on conservation of their encoded protein sequences. Unlike other comparative gene callers, Orthosearch does not incorporate a single-genome *ab initio* model (beyond the requirement for gene model validity); therefore, its predictive power comes entirely from conservation. Orthosearch calls genes on one genome (the “target”) using one or more genomes of related species (“informants”) and a set of pairwise alignments between the target and each informant. Orthosearch uses these alignments to divide the target genome into loci of consistent order and orientation of alignment between the target and some or all of the informants (regions of “conserved synteny”). Within each locus, a set of seed gene models is generated on the target. Each model is then iteratively projected to the informant genomes and back to the target, generating a valid gene model on each genome with maximal shared coding sequence. The final score of each model is computed based on the protein identity among the orthologous gene models generated. The set of genes predicted at a locus consists of the non-overlapping (on the target genome) models that maximize the total score for the locus. The Orthosearch algorithm and parameters will be described in greater detail elsewhere (R.E.H, manuscript in preparation).

For annotation of the *P. infestans* genome, we used the genomes of *P. sojae* and *P. ramorum*¹⁰ as informants and generated alignments using PatternHunter¹¹.

4.3 Targeted annotation of specific gene families and pseudogenes

Gene families corresponding to CRN proteins, NPP1-like, elicitor-like proteins, small secreted cysteine-rich proteins (SCRs), and transporters were targeted for manual review and gene model editing using Argo (<http://www.broad.mit.edu/annotation/argo/>). Evidence for identifying genes and editing exon boundaries was derived from protein and EST alignments, with focus provided by related proteins of known function. In the case of NPP1-like proteins, elicitors, and transporters, matches to corresponding Pfam¹² profiles were leveraged for genome loci identification, and GeneWise¹³ (run with the –pseudo option) assisted in pseudogene identification. Identification and annotation of CRN family members was challenged by the great complexity of this protein family, and so additional methods were required (see below). In the course of manual gene annotation, putative pseudogenes were identified as gene products that appeared to be disrupted, either due to frameshifts, intervening stop codons, or mobile element insertions, and were annotated correspondingly. Genes extending into a sequencing gap or truncated due to the bounds of a scaffold were annotated as 5' or 3' partials.

4.4 Annotation of the *Phytophthora* Crinkler protein family and domain structures

Alignment of the 16 previously known CRN proteins¹⁴ revealed a highly conserved N-terminal region followed by highly diverse C-terminal regions. A sequence logo illustrating this conserved N-terminal region is shown as Supplementary Figure S20. This hallmark of the CRN family was leveraged to discover all instances of CRN genes within each of the three *Phytophthora* genomes (*P. sojae*, *P. ramorum*, and *P. infestans*). A PSI-BLAST profile was created for this conserved N-terminal region and searched against the *Phytophthora* genomes using PSI-TBLASTN¹⁵. In addition, an HMM profile was created and searched against the 6-frame translation of the genomes using HMMER

¹⁶. Matches within the genome were manually interrogated using Argo, and gene models were manually constructed. The CRN gene annotations were manually reviewed at least twice during the course of our analysis of the CRN gene family, and adjustments were made based on our expanding evidence for the extent of CRN C-terminal domain diversity (see below). As in the case of the other gene families manually reviewed, CRN genes disrupted or severely truncated due to an intervening stop codon were annotated as pseudogenes.

As part of this work, a novel sequence analysis pipeline (named PSI-Domainer) was developed to identify protein domains conserved among CRN family proteins. First, the complete collection of manually annotated CRN proteins was divided into overlapping peptides of length 50 amino acids (aa) spaced by 10 aa. The k-mer (size 3) frequency within the original CRN protein database was computed, and each overlapping peptide was scored based on its k-mer abundance. The single peptide with the highest sum k-mer score was chosen as a domain seed candidate. This peptide domain seed candidate was searched against the original CRN protein database using PSI-BLAST (-z 8000000 -j 10 -P 1 -s T -e 1e-3 -h 1e-3 -v 10000 -b 10000 -F F). The coordinates of matches were captured and all peptides in the overlapping peptide database that overlap in coordinates to matches derived from the PSI-BLAST of the domain seed candidate were removed from the overlapping protein database. All regions of CRN proteins matching the domain seed candidate were assigned a label according to the domain seed and PSI-BLAST match score. This entire process cycled, starting with choosing the next domain seed candidate from among the remaining overlapping peptides, and continued to cycle until the overlapping peptide database was exhausted. The end result of this cyclical fully automated analysis was a labeling of related regions of CRN proteins according to homologous domain seeds.

The seed labels assigned to CRN proteins were sometimes found to overlap. We removed overlap among adjacent labels by preferentially selecting a label as follows. A directed graph was constructed with domain labels as vertices, and edges directed between labels according to their non-overlapping order within CRN proteins. For each

series of labels within each CRN protein, all labels i preceding and non-overlapping a label j were connected to form the graph. All nodes i preceding and overlapping node j were counted as colliding nodes. The graph of connected nodes was then pruned of edges connecting nodes that are found to collide more frequently than to connect. Each node is assigned a value corresponding to the number of times the corresponding label was assigned to CRN protein regions. The directed graph is then searched exhaustively from each node to find the path (set of connected nodes/labels) with the cumulative highest abundance among the CRN proteins. Labels corresponding to this path were preferentially selected and overlapping labels on each CRN protein were removed. The nodes corresponding to this highest scoring path were removed from the graph and the next highest scoring path was computed. This graph parsing, preferential label selection, and overlapping label removal cycled until all nodes were removed from the graph. The selected nodes provided a preliminary annotation of the domain structure for each CRN protein.

The ordered labels of seed-matching regions of each CRN protein were further refined to group collinearly labeled regions into more complete domain designations (at this stage, multiple non-overlapping adjacent seed labels may correspond to a single domain having the length of multiple seeds). A new graph was constructed based on the non-overlapping labels obtained from the previous step, and unbranched paths within the graph were grouped as fragments of a larger domain and replaced by a new single domain label. A preliminary assignment of domains within CRN proteins was reported as the final output of PSI-Domainer.

PSI-Domainer is completely automated, and although useful for gaining an initial glimpse into the domain organization of this family, was found to be inadequate to describe the full complexity of the CRN domain organization, as it reported multiple domain labels that correspond to the same overall clade of sequence homology and yielded weakly resolved domain boundaries. Accordingly, rather than using PSI-Domainer to annotate the domain structure, we utilized it simply to highlight those

regions of CRN proteins that share sequence homology, and then manually curated domains and domain boundaries using a semi-automated process described below.

First, PSI-Domainer was run based on the complete set of CRN protein sequences. The single most abundant domain reported in the final output of PSI-Domainer was chosen for manual curation. These domain sequence regions, along with the directly neighboring domains, were extracted from the CRN proteins, multiply aligned using MUSCLE¹⁷, and viewed using Jalview¹⁸. Sequence conservation was manually evaluated within Jalview, and domain boundaries were clipped accordingly. An HMM profile was constructed based on the manually curated protein domain, and searched back against the set of CRN proteins, masking any matching protein region having an E-value $\leq 1e-5$. PSI-Domainer was then executed again, this time using the proteins masked for occurrences of the previous domain, and the manual domain curation process was repeated. This cyclic semi-automated procedure continued until all regions of homology between CRN proteins were accounted for by curated protein domains.

The full collection of HMMs representing these manually curated protein domains was searched back against the original CRN proteins ($E \leq 1e-5$), and all domain instances were identified. The few overlapping domains found at this search stringency were resolved by coalescing related domains into a single HMM profile and revisiting the boundary assignment from within Jalview. Remaining CRN protein regions lacking homology to other CRN proteins and exceeding 100 aa in length were assigned singleton domain status. Consensus CRN domain peptide sequences are provided in Supplementary Table S11. The complete set of annotated CRN genes across *Phytophthora* are provided in Supplementary Table S10. The domain organization of annotated pseudogenes was assigned by BLASTX ($E \leq 1e-5$) homology of the pseudogene nucleotide sequence to peptides corresponding to the annotated CRN protein domains.

4.5 Annotation of RXLR effectors

Several bioinformatic strategies have been previously employed for identifying candidate RXLR effectors within genome sequences, relying on matches to an HMM profile or to a sequence pattern (regular expression or ‘regex’). The least restrictive criteria required the presence of a signal peptide within the first 30 residues followed by an RXLR pattern match between residues 30 and 60¹⁹. Extensions to this pattern to include a downstream EER motif that is present in most but not all validated RXLR effectors have also been employed, requiring an exact match to the EER residues, or a loose match, allowing E->D and R->K replacements, while still requiring an exact match to the upstream RXLR motif^{19,20}. Parallel searches for effectors using HMM profiles based on alignments of the RXLR-EER region derived from a reference set of RXLR effectors were also performed in both studies, either to provide a greater specificity for effector identification, or as a way to complement the extended regular expression-based searches. Depending on the method or combination of methods used, hundreds to over a thousand candidate effectors can be identified. By imposing highly restrictive criteria, such as a suitably scoring HMM profile match, or a highly restrictive RXLR-EER pattern match, we expect to miss the identification of known or high quality effector candidates encoding slight variations within this motif and those lacking the EER motif altogether. By loosening the criteria, such as by dropping the EER motif requirement, or further loosening the already liberal RXLR motif, we expect to encounter unacceptable numbers of false positives. By combining the already established search methods along with additional attributes of candidate peptides, such as family membership, coding likelihood, motif conservation, and signature of rapid evolutionary pattern, we hope to further refine the list of candidate RXLR effectors, as we primarily focus on the annotation of such genes in the *P. infestans* genome.

We reapplied the Win and Whisson methods to sequenced Stramenophile genomes, including the diatoms (*T. pseudonana* and *P. tricornutum*) as outgroups for comparison. All open reading frames (ORFs) encoding at least 70 amino acids from start to stop codon were extracted and signal peptides were predicted using SignalP 3.0²¹. Within a set of alternative overlapping ORFs within the same reading frame, the single ORF with the highest scoring signal peptide (HMM prob ≥ 0.9) or encoding the longest ORF was

selected. These ORFs with predicted signal peptides were subsequently analyzed using the following methods:

- Win method: a signal peptide within residues 1-30 followed by an RXLR motif within residues 30-60 ¹⁹.
- HMM: a search of Leighton Pritchard's HMM profile for the RXLR motif ('cropped.hmm'), retaining matches having a bit score > 0 ²⁰.
- Regex: allow for a signal peptide between residues 10-40, followed by the RXLR motif within the next 100 residues, followed by the EER motif, allowing D and K replacements to E and R, respectively
 $(\backslash w\{10,40\}\backslash w\{1,96\}R\backslash wLR\backslash w\{1,40\}[ED][ED][KR])$ ²⁰.

Counts for the number of peptides meeting the stated motif criteria are provided in Supplementary Table S-A2.

The HMM and the regex identified an equal number of candidate effectors in *P. infestans* (395 each), with a striking amount of overlap between them; the intersection of 326 *P. infestans* peptides having both an HMM match and the extended pattern match provides an approximate lower bound for the RXLR effector repertoire in *P. infestans*. The approximate upper bound could total over a thousand based on applying the more liberal method of Win *et al.* ¹⁹. We attempted to arrive at a more robust estimate of the RXLR effector arsenal by examining additional properties of these predicted peptides, excluding those that are unlikely to be protein-coding or lack features that can be considered hallmarks of RXLR effector families.

In addition to predicted signal peptides and motif identifications using the methods described above, we examined the protein coding score of gene predictions based on a 5th order Markov Model (as described by Parra *et al.* ⁷) and homologies to known functional RXLR effectors (BLASTP, E≤1e-10). We additionally considered homologies to other predicted secreted proteins encoding variations of the RXLR motif (tentative false negatives), and homologies to non-secreted unlikely effector proteins (tentative false positives). Clusters of homologous predicted peptides were leveraged as a resource in this

analysis, and computed by gathering all predicted ORFs with homology (BLASTP, $E \leq 1e-10$) to ORFs with predicted signal peptides, followed by clustering using TribeMCL²². Pairs of ORFs within clusters sharing at least 50% protein identity in a MUSCLE global alignment were further analyzed for evidence of positive selection (as previously described¹⁹). Secreted RXLR effector candidates were selected for annotation based any of the following individual or aggregate criteria:

- a) encoding an RXLR motif identified by both the HMM search and Regex methods (831: {*P. infestans* (Pi) 326, *P. ramorum* (Pr) 246, *P. sojae* (Ps) 225, *H. parasitica* (Hp) 34}). [Individual attribute]
- b) exhibiting sequence homology to any previously known functional RXLR effector (414: {Pi 236, Pr 95, Ps 71, Hp 12}). [Individual attribute]
- c) belonging to a TribeMCL cluster where $\geq 50\%$ of members are RXLR-EER (defined by criteria 'a' above) effector candidates (155 families containing 599 members: {Pi 262, Pr 162, Ps 152, Hp 23}). [Aggregate attribute]
- d) all secreted peptide members of families where $\geq 50\%$ of members share homology with reference effectors. (50 families containing 257 members {Pi 173, Ps 53, Pr 21, Hp 10}). [Aggregate attribute]
- e) all secreted members of Win-method RXLR enriched ($\geq 50\%$) families where $\max(\omega) > 1$ and coding score ≥ 20 (51 families containing 367 members: {Pi 154, Pr 127, Ps 74, Hp 12}). [Aggregate attribute]
- f) all RXLR candidates defined by the Win-method matched by ESTs and having a coding score ≥ 20 : includes 163 peptides {Pi: 118, Ps: 45}. Note that EST libraries were limited to *P. infestans* and *P. sojae* based on the public availability of such sequences. [Aggregate attribute]
- g) all secreted members of Win-method RXLR enriched ($\geq 50\%$) families having an average coding score of at least 20 and containing at least one member supported by ESTs (55 families containing 344 members: {Pi 175, Ps 84, Pr 78, Hp 7}). [Aggregate attribute]

The resulting set of proteins fulfilling at least one of our selection criteria totals 1192 RXLR effector candidates with the organism distribution of {Pi 488, Pr 309, Ps 335, Hp 60}.

Further interrogation of the *P. infestans* genome and manual refinement of the RXLR effectome by project collaborators led to the actual annotation of 563 *P. infestans* RXLR candidates. The additional candidates were selected based on the occurrence of conserved C-terminal W-Y-L motifs and inferred from their overall sequence conservation with previously published effectors²³. The number of RXLR effectors annotated for the *P. ramorum* and *P. sojae* genomes is largely consistent with those previously published¹⁰. All automated analyses described above predicted a substantially larger repertoire of RXLR candidates in *P. infestans* as compared to *P. sojae* and *P. ramorum*, consistent with our assertion that *P. infestans* encodes an expanded repertoire of RXLR effectors.

The diatom genomes do harbor a small number of ORFs identified as having signal peptides and matching RXLR motifs, but they are more than an order of magnitude less abundant than in the three *Phytophthora* species, and none show sequence homology to known oomycete RXLR effectors. Therefore, given current knowledge of RXLR effectors, we do not confidently annotate any such gene among diatoms. The complete set of annotated *Phytophthora* RXLR effectors are provided in Supplementary Table S7.

Conserved domains identified within RXLR effectors were identified using PSI-Domainer. In contrast to the CRN domain identification, RXLR domains were identified via a fully automated execution of the PSI-domainer software, exhaustively identifying all conserved domains among RXLR effectors.

4.6 Analysis of protein-coding genes

Protein family and ortholog clustering were performed using Tribe-MCL²² and Ortho-MCL²⁴, respectively, using matching protein pairs (BLASTP, $E \leq 1e-10$). Tests for diversifying selection were performed on clustered protein pairs as previously described

¹⁹. Signal peptides were predicted using signalP 3.0 ²¹. Analyses of targeted protein families were described separately above (e.g. RXLRs, CRNs). Protein domains and product names were assigned based on sequence homologies to known proteins or Pfam ¹² domains scoring above the trusted cutoff scores.

Gene functional categories were computed by searching annotated proteins against the NCBI COG and KOG databases ²⁵ using BLASTP ($E \leq 1e-10$) and transitively assigning functional categories from the single best match. Functional enrichments and deficiencies for genes identified as the core *Phytophthora* orthologs were computed using the Fisher one-tail exact tests using the R software as described ²⁶.

Analysis of localized gene density was performed by plotting the abundance of genes according to 5' and 3' flanking intergenic lengths. Intergenic lengths were binned according to log(length) and plotted as a 2-dimensional heatmap using Excel. Of course, only the subset of genes flanked by neighboring genes at both termini could be included in this analysis. The source of gene annotations for organisms other than *P. infestans* are provided in Supplementary Table SA-3.

5. Analysis of Repeats

5.1 Repeat identification and curation of mobile element reference sequences

A library of repetitive element family consensus sequences was constructed *de novo* using RepeatScout ²⁷ with a seed size of 13. RepeatScout was applied separately to each of the *Phytophthora* genomes, and instances of repetitive regions were identified using RepeatMasker ²⁸. We quantified the total repeat content of each genome was performed by summing all bases matching repeat family consensus sequences from the corresponding RepeatScout libraries. The set of repetitive elements defined by RepeatScout includes all nearly identical repetitive sequence regions found at least three times in the genome, including transposable elements, as well as recently duplicated paralogous genes and other dispersed duplicated regions.

To gain biological insight into the composition of the repeat regions of the genome, we analyzed the most abundant repetitive sequences. We extracted regions of the genome containing high copy repetitive elements, including flanking genomic regions, multiply aligned the genome segments, and analyzed the multiple alignment to define conserved features of the element (a semi-automated or manual process). This often sharply delineated mobile element insertion points and allowed the identification of target site duplications (TSD). Further inspection of curated repeat instances identified other features consistent with canonical structures of mobile elements, such as terminal inverted repeats (TIR) for DNA transposons and long terminal repeats (LTR) of retro-transposons. Sequence homology to families of known transposon ORFs based on PSI-TBLASTN matches to profiles of transposon ORFs or matches to known transposon proteins in a collection of (retro-)transposon ORFs (Transposon-PSI (<http://transposonpsi.sf.net>)) provided evidence for transposon family classification (ie. hAT, Gypsy, etc). Reference sequences for the mobile elements identified are provided in Supplementary Table S3. Instances of these elements in the genome were identified using RepeatMasker, and full-length elements were designated as those instances matching at least 80% of the reference sequence length.

Additional transposable elements in high abundance but exhibiting few recent insertions were discovered during our analysis of large *P. infestans*-specific or ortholog-deficient protein families. Our genome annotation effort attempts to exclude gene predictions that show homology to known transposons, but those that correspond to novel elements or elements of known transposon families that are highly diverged are not automatically recognized as such and persist in initial annotated gene sets. Our analysis of large protein families, specifically examples of recent paralog duplications within large families, identified elements recently (retro-)transposed. Evidence of recent insertion, including TSDs and LTRs or ITRs, supported the reclassification of corresponding gene family members as diverged or novel families of transposable elements. Reference element sequences were added to our curated mobile element library for *P. infestans*, and gene families corresponding to the newly identified mobile element were removed from the

protein-coding gene annotation set of *P. infestans*. Protein coding genes previously annotated for *P. sojae* and *P. ramorum*¹⁰ were likewise removed from their gene sets prior to subsequent comparative analyses with *P. infestans*.

The relative abundance of mobile element types among the three *Phytophthora* species was estimated using Transposon-PSI ($E \leq 1e-5$). Collinear PSI-BLAST matches along the genome separated by at most 1 kb were chained, and overlapping profile homologies with lesser scores were excluded. Total numbers of (chained) matches according to mobile element family were reported to reflect element abundance.

5.2 Estimation of autonomous helitron abundance among diverse genomes

A combination of PSI-TBLASTN¹⁵ and GeneWise¹³ were utilized to identify candidate remnants of helitron protein-coding regions in the genome. Helitron protein sequences were retrieved from NCBI based on BLASTP homology to known helitrons (and helentrons) from *Arabidopsis*, *C. elegans*, and Zebrafish. A multiple alignment of reference helitron proteins was generated using MUSCLE, manually examined and trimmed using Jalview, and utilized to generate both a PSI-BLAST profile and HMM profile. The PSI-BLAST profile was searched against a diverse collection of genomes previously shown to exhibit evidence of helitron activity, autonomous or non-autonomous²⁹, or found to encode helitron-like proteins as determined by the earlier BLASTP search of the 'nr' database at NCBI. First, genomic loci showing sequence similarity to helitron proteins were identified by a PSI-TBLASTN ($E \leq 1$) search of the helitron PSI-BLAST profile against the genome sequences. High-scoring segment pairs (HSPs) found collinear between the PSI-BLAST profile and genomic scaffold were chained together, allowing a maximum of 1 kb between collinear HSPs and no more than 20% overlap of either HSP length. The regions of the genome corresponding to the chained HSPs were extracted from the genome, including a flanking 500 bp, and a helitron ORF gene model was constructed for each locus using GeneWise and the helitron HMM profile (using the -pseudo option). The GeneWise model coordinates for each region were transformed to the full genome coordinates, and multiple GeneWise predictions found to overlap the same genomic locus were resolved to retain only the

longest of the overlapping models. The number of remaining models was counted to provide an estimate of the number of remnants of autonomous helitrons in the genome. Note, for *P. infestans*, the number of helitrons we report in the main manuscript is based on a high stringency search of nucleotide sequences of identified full-length elements. The protein-based homology search methods described here are more appropriate for the cross-genome comparisons, and should yield greater sensitivity for detecting more ancient autonomous helitrons. The sources for the genomes examined are provided in Supplementary Table S-A3.

6. Computation of blocks of conserved gene order among *Phytophthora* genomes

Regions of conserved collinear gene order between the three *Phytophthora* genomes were computed using DAGchainer³⁰. Instead of scoring the absolute distance between genes in bases, only the relative order of the genes along each scaffold was considered (DAGchainer parameters: -e -1 -Z 3 -D 10000 -x 1 -g 1 -A 6). Also, rather than using all blast matches between genes as anchors for collinear blocks, only matches between orthologous gene pairs were utilized, including co-orthologs, as defined by OrthoMCL²⁴. Collinear blocks were defined between each pair of the three *Phytophthora* genomes. The few candidate regions of segmental genome duplication, defined as those regions in one genome that exhibit conserved gene order to two or more blocks in another genome, were resolved to remove any overlap between blocks. Resolving block overlap involved one of several operations: insertion of a smaller block into a larger block either incorporating a local inversion or fragmenting the larger block to accommodate a rearrangement or disregarding a candidate region of segmental genome duplication. The operation (for example: insertion of a block within a larger block with removal of intervening overlapping collinear gene pairs or retention of the larger block and dismissing the smaller block) was chosen to be that which optimized the sum of the scores of the genes retained in collinear blocks (using the DAGchainer E-value based scoring function³⁰). The range of collinear gene blocks reported for each genome corresponds to the union of blocks obtained from the pairwise comparisons to the other genomes.

7. Functional analyses of Crinkler Genes and C-terminal domains

7.1 Plant growth conditions

Nicotiana benthamiana plants were grown under greenhouse conditions at 22-25°C under high light intensity prior to and throughout the experiments.

7.2 Plasmid construction

pGR106 constructs containing CRN1 and CRN2 (pGR106-CRN1 and pGR106-CRN2) were described previously³¹. cDNAs of all other tested paralogs were cloned using the approach described by Torto *et al.*³¹.

7.3 Potato virus X (PVX) agroinfection assays

Plants were maintained under greenhouse conditions with a temperature of 22-25°C and high light intensity throughout the experiments. Multiple independent experiments using side-by-side wound-inoculations with recombinant *Agrobacterium tumefaciens* GV3101 (carrying pGR106-cDNA) were conducted as previously described³¹. Leaf material was harvested when plants showed symptoms of systemic infection (12-14 dpi), frozen in liquid nitrogen and stored at -80°C.

7.4 CRN2 deletion analyses reveal a C-terminal effector domain

We performed deletion analysis to define the effector domain of CRN2. Gene specific primers were designed to enable amplification of gene fragments with desired length, introduce proper start and stop codons when necessary and generate appropriate flanking restriction sites. The deleted gene constructs were cloned into pGR106, an *A. tumefaciens* binary Potato virus X vector^{31,32}, and assayed in *N. benthamiana*. Recombinant *A. tumefaciens* colonies were toothpick inoculated side by side on mature *N. benthamiana* leaves. Inoculations were repeated for each construct and development of local necrosis was monitored and recorded. Deletion of the signal peptide did not abolish cell death-

inducing activity. Additional deletion analysis revealed that up to 172 N-terminal and 49 C-terminal amino acids could be removed before loss of activity occurred. Similar analyses were carried out by deleting C-terminal portions of CRN2. These results show that a C-terminal domain of CRN2 spanning a 234 amino acid region (173–407) is sufficient for cell death induction.

7.5 Phenotypic characterization of diverse CRN C-terminal domains *in planta*

Crinkler (CRN) effector domains were PCR amplified using cDNA generated from total RNA of *P. infestans* T30-4 strain (corresponding CRN targets and phenotypes are listed in Supplementary Table S-A4). Amplified fragments were digested with either ClaI/NotI or AscI/NotI restriction enzymes and ligated into pGR106 vector digested with same enzyme combinations. Assays for the necrosis phenotype on mature *N. benthamiana* leaves were performed as described above for the CRN2 deletions.

8. Generation of an Optical Map for the *P. infestans* Genome

8.1 Genomic DNA preparation for optical mapping

The optical map was generated from a sibling of the T30-4 isolate derived from the same cross (T30-4(FG)). *P. infestans* protoplasts were washed three times with PBS buffer to remove the storage buffer and glycerol, and lysed in TE pH 8.0 containing 5 mM EGTA and 1 mg/ml proteinase K by heating the protoplast suspension to 65° C for 30 min. The DNA solution was then incubated at 37° C overnight to ensure full digestion of proteins from the lysed protoplasts and the autodigestion of excess proteinase K. T7 DNA (final concentration ~30 pg/μl) was added to the genomic DNA solution as a sizing standard.

8.2 Optical mapping of *P. infestans* genome

A whole genome shotgun optical mapping approach^{33 34 35 36} was used to construct a high-resolution whole genome physical map of the *P. infestans* genome using the restriction enzyme SpeI to guide genome sequence assembly. We acquired ~1089-fold coverage of the genome as single DNA molecule maps and assembled high quality maps

into 56 optical map contigs ranging from 884 kb to 16,442 kb in size spanning 404.88 Mb; this span was reduced through joining homologous regions. Contigs were assembled by the map assembler³⁷ using divide and conquer and iterative assembly strategies for distributing the computational load^{35,38,39}. Within the 56 optical map contigs, there are 42,310 single DNA molecule maps with a total mass of 20,177.74 Mb, which corresponds to ~85-fold coverage of the estimated genome size 237 Mb⁴⁰. Out of the 56 optical map contigs, there are 16 finished chromosome optical contigs containing both chromosome ends, and each of the two ends has at least 5 single molecule maps terminating at similar positions; there are 12 other optical map contigs with a single chromosome end, whereas the other 28 optical map contigs lack clear chromosome ends. Most of the optical map contigs without chromosome ends include large restriction fragments (>100 kb) or small repeats that challenge the identification of single molecule maps spanning these regions.

Heterozygosity and chromosomal rearrangements impact the optical maps for diploid organisms like *P. infestans*. Accordingly, we analyzed the 56 optical map contigs for homologous restriction patterns using local alignment with the gap insertion option enabled^{41 38 42}. The map alignment software allows for both local and global optical map pairwise alignments. A global alignment expects global similarity between the two maps and works well if the two maps to be compared are very similar. However, global alignment approaches will fail to detect local significant instances of variation, such as large insertions or deletions. Our approach for extracting multiple high-scoring alignments is based on an efficient linear scaling approach of Huang and Miller⁴³. We generate confidence scores using an approach similar to that used by Waterman and Vingron⁴⁴ for sequence alignments.

Map alignment analysis found 24 optical map contigs sharing homologous restriction patterns including 6 optical map contigs with both chromosome ends, 6 optical map contigs with one chromosome end, and 12 with no chromosome end. In total there are 32 optical map contigs for the haploid genome of *P. infestans* spanning 237.43 Mb including 10 finished chromosome optical map contigs. Consequently the genome size should be

larger than 237 Mb because 22 chromosome optical map contigs remain unfinished. These missing regions are most likely comprised of high copy repeat sequences.

8.3 Map comparison between optical maps and sequence scaffolds

We compared optical maps with the *in silico* maps generated by virtual digestion of the current sequence scaffolds with SpeI using map alignment software. 285 (~193 Mb) / 4921 (total) sequence scaffolds having 3 or more SpeI restriction fragments (not including ends and fragments less than 0.4 kb) were tested for alignment; 96 scaffolds presented 159 aligned stretches (~65%) spanning 125.83 Mb.

9. Generation and analysis of a Full-Length Normalized cDNA library of *P. infestans*

A Full-length Normalized cDNA Library was generated using mixed developmental stages: mycelia and germinated cyst. Mycelia were collected from cultures of two and three weeks in liquid Plich Medium at 18°C in the dark. Pellet was washed 3 times with sterile water, vacuum dried and grinded to a fine powder with liquid nitrogen. Total RNA from mycelia powder was extracted according to the TRIzol reagent manufacturer's instructions (Invitrogen) and store at -70°C. Zoospores release from sporangia was induced by adding cold water solution to 14 days rye sucrose agar cultures in plates of 140mm. 15ml of zoospores solution was harvest after 2h of incubation at 4°C. Zoospores were concentrated and put in incubation horizontally at 18°C. After 16h of incubation germinated cyst were collected by centrifugation and lysed in 2ml of TRIzol Reagent (Invitrogen). Isolation proceeded as indicates by the manufactures, with some modifications. RNA pellet was washed with 75% ethanol, dried and dissolve in The RNA Storage solution buffer (Ambion) and store at -70°C. The final mix of total RNA contained 91% isolated from mycelia and 11% isolated of germinated cyst of *P. infestans* T30-4. The Normalize Full-Length cDNA library was constructed by DNAFORM service (<http://www.dnaform.jp>) in seven steps: 1) 1st strand cDNA synthesis by using trehalose-thermostabilized wildtype M-MuLV reverse transcriptase using oligo (dT) primer; 2) Full-length cDNA selection of mRNA/cDNA hybrids by Cap-Trapper selection method as described by ⁴⁵; 3) 2nd strand cDNA synthesis based on 5' Adaptor

ligated to the recovered full-length cDNAs; 4) Library Normalization by removal of the most abundant cDNAs using streptavidin-coated beads. 5) Cloning of double stranded Normalized cDNAs into phage Lambda-FLC III; 6) Production of a plasmid stock from the phage in accordance with the protocol for cDNA library Cre mediated in vitro excision; 7) Transformation of Plasmid DNA into T1 bacteriophage resistant *E. coli* DH10B.

Paired-end sequencing of 10,752 cDNA clones was performed at the Broad Institute using conventional Sanger sequencing technology. Paired reads of these cDNAs are available in dbEST with accessions in the range GR284383- GR301386. Low quality ends were trimmed using Lucy⁴⁶, and vector removed using the seqclean software (<http://compbio.dfci.harvard.edu/tgi/software/>) and the UniVec database (<ftp://ftp.ncbi.nih.gov/pub/UniVec/>), yielding 10,561 clones with high quality vector-trimmed sequences. The cDNA sequences were mapped to the genome using GMAP⁴⁷, and only the single top scoring match was reported. Of the 10,561 cDNA clones, 10,325 have at least one paired read that maps to the genome with at least 90% identity. Of the 236 cDNAs that did not map to the assembly, 44 were identified as likely contaminants by their high percentage sequence identity to known bacterial proteins (from organisms such as *E. coli*) or to plant proteins (from potato or tomato). Based on these values, we estimate that the genome assembly is approximately $10325/(10561 - 44) = 98.2\%$ complete, and therefore we estimate approximately $((1/0.982 - 1) * 17797) = 326$ genes to be missing from the assembly.

10. Analysis of gene expression using NimbleGen microarrays

Total RNA samples were extracted from mycelia of *P. infestans* T30-4 growing for 10-12 days in three media types: Pea, V8 juice, and Rye Sucrose agar. Mycelia were harvested with a sterile blade, ground in liquid nitrogen and frozen. Powder tissue was homogenized with RLT buffer containing β -mercaptoethanol from Qiagen (Cat.No. 74904) proceeding with manufacture's protocol. In addition to mycelia stage, RNA was also extracted during infection of *P. infestans* T30-4 on potato. Zoospores stage was induced by adding cold water to the culture and incubating for 2-3 hours at 4°C. Each

potato leaf was infected with 4 drops containing zoospores and incubated at 18°C in containers with a wet paper towel to promote humidity. We collected infected leaf discs at 2, 3, 4 and 5 days post inoculation, ground in liquid nitrogen and froze for processing with a RNAeasy Plant Mini Kit (Qiagen, Cat No. 74904). All biological samples were generated with replicates. RNA quality and integrity were checked prior to cDNA synthesis using the Bioanalyzer (Agilent 2100). NimbleGen microarray services were utilized for cDNA preparations and subsequent chip hybridizations. Microarray probe intensities were quantile normalized across all chips⁴⁸, and the ‘tukey.biweight’ function of the Bioconductor ‘affy’ library was applied to sets of probe intensities to compute robust average gene intensities⁴⁹. To exclude from the analysis probe sequences that correspond to heavily repetitive sequences and that could extensively cross-hybridize, the probes were aligned to the genome using SSAHA⁵⁰ (parameters: -maxGap 5 -maxInsert 5 -mp 50 -pf -da 0) and only those probes matching the genome at most five times were retained in our analysis and subsequently utilized for quantile normalization and computing target gene intensities. Analysis of gene expression was performed using the MultiExperiment Viewer (MeV)⁵¹. Log2 transformed array intensity values were analyzed for differential gene expression using the t-test, as implemented in MeV, assuming equal variances. For this study, only the array targets corresponding to annotated *P. infestans* genes were analyzed. T-tests were performed comparing two groups: Group A, consisting of sample replicates for mycelia grown in Pea, V8, and RS agar, and Group B, consisting of replicates for one of the days post-infection. False discovery rates were addressed by computing q-values for each test⁵². Such tests were performed for each pair of replicates post-infection, and all significantly ($p \leq 0.05$, $q \leq 0.05$) differentially expressed genes were reported; those genes exhibiting at least two-fold variation in gene expression between averaged media and infected potato sample replicates are described (Supplementary Table S12).

Table S1: Protein-coding gene structure statistics

	<i>P. infestans</i>	<i>P. sojae</i>	<i>P. ramorum</i>
Genes			
Length mean (median)	1,523 (1,182) bp	1,613 (1,239) bp	1,625 (1,251) bp
Genes with introns	67 %	56 %	53%
Exons			
Mean number per gene	2.8	2.6	2.6
Length mean (median)	472 (221) bp	536 (273) bp	553 (270) bp
Mean G+C content	53.1 %	59 %	57 %
Introns			
Length mean (median)	125 (73) bp	124 (78) bp	123 (74)
Mean G+C content	46 %	51 %	49 %

Table S4: Albatross vs. Pi-1 Gypsy LTR retrotransposon clades and contribution to genome size*

	Pi-1	Albatross
Total bp masked	22345764	32196223
% of contig bp	11.70%	16.86%
% syntenic bp	9.45%	7.51%
% non-syntenic bp	13.39%	23.89%
non-syntenic enrichment	1.417	3.181
Number FL-elements*	1147	336
non-syntenic FL-enrichment	0.93	2.1

*a minimum match length of 80% of the reference sequence was required for full-length instance classification.

Table S5: Estimate of the abundance of autonomous helitrons across diverse eukaryotes

Group	Organism	Count of helitron ORF models (remnants of helitrons, including pseudogenized ORFs)
<i>Phytophthora</i>	<i>P. infestans</i>	698
	<i>P. sojae</i>	28
	<i>P. ramorum</i>	56
<i>Insecta</i>	<i>Drosophila melanogaster</i>	0
	<i>Drosophila yakuba</i>	9
	<i>Aedes aegypti</i>	8
	<i>Nasonia vitripennis</i>	136
Invertebrates	<i>Caenorhabditis elegans</i>	6
	<i>Nematostella vectensis</i>	0
Vertebrate	<i>Danio rerio</i>	46
Plants	<i>Arabidopsis thaliana</i>	54
	<i>Oryza sativa</i>	109
Fungi	<i>Aspergillus nidulans</i>	3

Table S6: Protein families implicated in plant pathogenesis*

	<i>P. infestans</i>	<i>P. sojae</i>	<i>P. ramorum</i>	<i>T.pseudonana</i> (diatom)	<i>P. tricornutum</i> (diatom)
Aspartyl proteases	12	13	14	5	6
Serine carboxypeptidases	24	21	19	6	10
Cysteine proteases	33	29	35	16	13
Glycosyl hydrolases	157	190	173	31	22
Pectin esterases	11	19	13	0	0
Pectate lyases	30	32	30	0	0
Lipases	19	27	17	22	17
Phospholipases	36	31	28	18	11
Protease inhibitors, all	38	26	18	11	5
Cytochrome P450s	19	25	24	7	7
ABC transporters	156	176	179	50	52
NPP1-like proteins (necrosis-inducing proteins)	27	39	59	0	0
Elicitin-like proteins	40	57	50	0	0
PcF/SCR-like	16	8	1	0	0
Crinklers (CRN-family)	196	100	19	0	0
RXLR effectors	563	350 [†]	350 [†]	0	0

* counts of annotated pseudogenes are omitted

[†] counts of RXLR effectors reported in ¹⁰. We annotated 335 (*P. sojae*) and 309 (*P. ramorum*) RXLR effectors (Supplementary Table S7).

Table S-A2: Predictions of RXLR genes according to method

Organism	# secreted ORFs	Win method	HMM total	Regex total	Either HMM or Regex
<i>P. infestans</i>	23094	1071	395	395	464
<i>P. sojae</i>	18795	613	297	285	357
<i>P. ramorum</i>	14317	508	304	291	349
<i>H. parasitica</i>	8314	188	60	64	90
<i>T. pseudonana</i>	7417	93	7	13	20
<i>P. tricornutum</i>	5828	86	2	15	17

Table S-A3: Sources of external genomic data

Organism	Genome data source
<i>Aedes aegypti</i>	ftp.vectorbase.org
<i>Arabidopsis thaliana</i>	ftp.arabidopsis.org
<i>Aspergillus nidulans</i>	broad.mit.edu
<i>Caenorhabditis elegans</i>	sanger.ac.uk
<i>Danio rerio</i>	genome-archive.cse.ucsc.edu
<i>Drosophila melanogaster</i>	ftp.flybase.net
<i>Drosophila yakuba</i>	ftp.flybase.net
<i>Fusarium graminearum</i>	broad.mit.edu
<i>Homo sapiens</i>	ftp.ncbi.nih.gov
<i>Hyaloperonospora parasitica</i>	genome.wustl.edu
<i>Laccaria bicolor</i>	genome.jgi-psf.org
<i>Magnaporthe grisea</i>	broad.mit.edu
<i>Mus musculus</i>	genome.ucsc.edu
<i>Nasonia vitripennis</i>	ftp.hgsc.bcm.tmc.edu
<i>Oryza sativa</i>	ftp.ncbi.nih.gov
<i>Phaeodactylum tricornutum</i>	genome.jgi-psf.org
<i>Phytophthora ramorum</i>	genome.jgi-psf.org
<i>Phytophthora sojae</i>	genome.jgi-psf.org
<i>Plasmodium falciparum</i>	broad.mit.edu
<i>Thalassiosira pseudonana</i>	genome.jgi-psf.org
<i>Ustilago maydis</i>	broad.mit.edu

References for Supplementary Information

1. Whisson, S.C. et al. Physical mapping across an avirulence locus of *Phytophthora infestans* using a highly representative, large-insert bacterial artificial chromosome library. *Mol Genet Genomics* 266, 289-95 (2001).
2. Van der Lee, T., De Witte, I., Drenth, A., Alfonso, C. & Govers, F. AFLP Linkage Map of the Oomycete *Phytophthora infestans*. *Fungal Genet Biol* 21, 278-91 (1997).
3. van der Lee, T., Testa, A., Robold, A., van 't Klooster, J. & Govers, F. High-density genetic linkage maps of *Phytophthora infestans* reveal trisomic progeny and chromosomal rearrangements. *Genetics* 167, 1643-61 (2004).
4. van der Lee, T., Robold, A., Testa, A., van 't Klooster, J.W. & Govers, F. Mapping of avirulence genes in *Phytophthora infestans* with amplified fragment length polymorphism markers selected by bulked segregant analysis. *Genetics* 157, 949-56 (2001).
5. Raeder, U. & Broda, P. Rapid preparation of DNA from filamentous fungi. *Letters in Applied Microbiology*, 1-17 (1985).
6. Jaffe, D.B. et al. Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* 13, 91-6 (2003).
7. Parra, G., Blanco, E. & Guigo, R. GeneID in *Drosophila*. *Genome Res* 10, 511-5 (2000).
8. Lowe, T.M. & Eddy, S.R. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res* 25, 955-64 (1997).
9. Griffiths-Jones, S. et al. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res* 33, D121-4 (2005).
10. Tyler, B.M. et al. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* 313, 1261-6 (2006).
11. Ma, B., Tromp, J. & Li, M. PatternHunter: faster and more sensitive homology search. *Bioinformatics* 18, 440-5 (2002).
12. Finn, R.D. et al. The Pfam protein families database. *Nucleic Acids Res* 36, D281-8 (2008).
13. Birney, E., Clamp, M. & Durbin, R. GeneWise and Genomewise. *Genome Res* 14, 988-95 (2004).
14. Win, J., Kanneganti, T.D., Torto-Alalibo, T. & Kamoun, S. Computational and comparative analyses of 150 full-length cDNA sequences from the oomycete plant pathogen *Phytophthora infestans*. *Fungal Genet Biol* 43, 20-33 (2006).
15. Altschul, S.F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* 25, 3389-402 (1997).
16. Eddy, S.R. Profile hidden Markov models. *Bioinformatics* 14, 755-63 (1998).

17. Edgar, R.C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* 32, 1792-7 (2004).
18. Clamp, M., Cuff, J., Searle, S.M. & Barton, G.J. The Jalview Java alignment editor. *Bioinformatics* 20, 426-7 (2004).
19. Win, J. et al. Adaptive evolution has targeted the C-terminal domain of the RXLR effectors of plant pathogenic oomycetes. *Plant Cell* 19, 2349-69 (2007).
20. Whisson, S.C. et al. A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature* 450, 115-8 (2007).
21. Bendtsen, J.D., Nielsen, H., von Heijne, G. & Brunak, S. Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol* 340, 783-95 (2004).
22. Enright, A.J., Kunin, V. & Ouzounis, C.A. Protein families and TRIBES in genome sequence space. *Nucleic Acids Res* 31, 4632-8 (2003).
23. Jiang, R.H., Tripathy, S., Govers, F. & Tyler, B.M. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. *Proc Natl Acad Sci U S A* 105, 4874-9 (2008).
24. Li, L., Stoeckert, C.J., Jr. & Roos, D.S. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* 13, 2178-89 (2003).
25. Tatusov, R.L. et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* 4, 41 (2003).
26. Rivals, I., Personnaz, L., Taing, L. & Potier, M.C. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* 23, 401-7 (2007).
27. Price, A.L., Jones, N.C. & Pevzner, P.A. De novo identification of repeat families in large genomes. *Bioinformatics* 21 Suppl 1, i351-8 (2005).
28. Smit, A., Hubley, R. & Green, P. RepeatMasker Open-3.0. (1996-2004).
29. Kapitonov, V.V. & Jurka, J. Helitrons on a roll: eukaryotic rolling-circle transposons. *Trends Genet* 23, 521-9 (2007).
30. Haas, B.J., Delcher, A.L., Wortman, J.R. & Salzberg, S.L. DAGchainer: a tool for mining segmental genome duplications and synteny. *Bioinformatics* 20, 3643-6 (2004).
31. Torto, T.A. et al. EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen *Phytophthora*. *Genome Res* 13, 1675-85 (2003).
32. Lu, R. et al. High throughput virus-induced gene silencing implicates heat shock protein 90 in plant disease resistance. *EMBO J* 22, 5690-9 (2003).
33. Lim, A. et al. Shotgun optical maps of the whole *Escherichia coli* O157:H7 genome. *Genome Res* 11, 1584-93 (2001).
34. Lin, J. et al. Whole-genome shotgun optical mapping of *Deinococcus radiodurans*. *Science* 285, 1558-62 (1999).
35. Zhou, S. et al. Shotgun optical mapping of the entire *Leishmania major* Friedlin genome. *Mol Biochem Parasitol* 138, 97-106 (2004).
36. Zhou, S. et al. Whole-genome shotgun optical mapping of *Rhodobacter sphaeroides* strain 2.4.1 and its use for whole-genome shotgun sequence assembly. *Genome Res* 13, 2142-51 (2003).

37. Valouev, A., Schwartz, D.C., Zhou, S. & Waterman, M.S. An algorithm for assembly of ordered restriction maps from single DNA molecules. *Proc Natl Acad Sci U S A* 103, 15770-5 (2006).
38. Zhou, S. et al. Validation of rice genome sequence by optical mapping. *BMC Genomics* 8, 278 (2007).
39. Zhou, S., Jerschleb, J. & Schwartz, D.C. *A single molecule system for whole genome analysis*, (Elsevier B.V., Amsterdam, 2007).
40. Tooley, P.W. & Therrien, C.D. Determination of the Nuclear DNA Content of 23 Mexican and 18 Non-Mexican Isolates of *Phytophthora infestans*. *Experimental Mycology* 11, 19-26 (1987).
41. Valouev, A. et al. Alignment of optical maps. *J Comput Biol* 13, 442-62 (2006).
42. Zhou, S. et al. Single-molecule approach to bacterial genomic comparisons via optical mapping. *J Bacteriol* 186, 7773-82 (2004).
43. Huang, X. & Miller, W. A time-efficient, linear space local similarity algorithm. *Adv. Appl. Math.* 12, 337-357 (1991).
44. Waterman, M. & Vingron, M. Sequence comparison significance and poisson approximation. *Statistical Science*, 401-418 (1994).
45. Carninci, P. et al. Balanced-size and long-size cloning of full-length, cap-trapped cDNAs into vectors of the novel lambda-FLC family allows enhanced gene discovery rate and functional analysis. *Genomics* 77, 79-90 (2001).
46. Chou, H.H. & Holmes, M.H. DNA sequence quality trimming and vector removal. *Bioinformatics* 17, 1093-104 (2001).
47. Wu, T.D. & Watanabe, C.K. GMAP: a genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics* 21, 1859-75 (2005).
48. Bolstad, B.M., Irizarry, R.A., Astrand, M. & Speed, T.P. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics* 19, 185-93 (2003).
49. Reimers, M. & Carey, V.J. Bioconductor: an open source framework for bioinformatics and computational biology. *Methods Enzymol* 411, 119-34 (2006).
50. Ning, Z., Cox, A.J. & Mullikin, J.C. SSAHA: a fast search method for large DNA databases. *Genome Res* 11, 1725-9 (2001).
51. Saeed, A.I. et al. TM4: a free, open-source system for microarray data management and analysis. *Biotechniques* 34, 374-8 (2003).
52. Storey, J.D. & Tibshirani, R. Statistical significance for genomewide studies. *Proc Natl Acad Sci U S A* 100, 9440-5 (2003).