

Supplementary Materials & Methods

Data collection and preparation

Genome data of nine oomycetes were downloaded (see Table). To define gene families, a similarity search on all oomycete proteins was executed (BLASTP (Altschul et al. 1997); E-value cut-off e-5). Two minimal overlap thresholds were introduced: only hits with at least 25% high scoring pairs (HSPs) coverage of the largest sequence (either hit or query) and 50% alignment coverage of the largest sequence were kept. Also, the order of the HSPs was required to be equal in hit and query. Protein families were built using MCL (Enright et al. 2002), inflation factor 5. Families containing more than two hundred proteins are excluded from subsequent analyses.

Table. Datasets used in the current study and their sources. Abbreviations: prot (proteins sequences), cod (coding sequences), scaf (scaffold sequences).

Genome	Version	Data	Source	Url
<i>Hyaloperonospora arabidopsidis</i>	v8	prot, cod, gff	<i>Saprolegnia genome</i> Sequencing Project, Broad Institute of Harvard and MIT	http://www.broadinstitute.org/
<i>Phytophthora capsici</i>	v11	prot, scaf, gff	JGI, US dept. of Energy	http://genome.jgi.doe.gov
<i>Phytophthora cinnamomi</i> var. <i>cinnamomi</i>	v1	prot, scaf, gff	JGI, US dept. of Energy	http://genome.jgi.doe.gov
<i>Phytophthora infestans</i> strain T30-4	v1	prot, cod, gff	<i>Phytophthora infestans</i> Sequencing Project, Broad Institute of Harvard and MIT	http://www.broadinstitute.org/
<i>Phytophthora parasitica</i> INRA-310	v1	prot, cod, gff	<i>Phytophthora parasitica</i> INRA-310 Sequencing Project, Broad Institute of Harvard and MIT	http://www.broadinstitute.org/
<i>Phytophthora ramorum</i>	v1	prot, cod, gff	<i>Saprolegnia genome</i> Sequencing Project, Broad Institute of Harvard and MIT	http://www.broadinstitute.org/
<i>Phytophthora sojae</i>	v1	prot, cod, gff	<i>Saprolegnia genome</i> Sequencing Project, Broad Institute of Harvard and MIT	http://www.broadinstitute.org/
<i>Pythium ultimum</i>	v1	prot, cod, gff	<i>Saprolegnia genome</i> Sequencing Project, Broad Institute of Harvard and MIT	http://www.broadinstitute.org/
<i>Saprolegnia parasitica</i>	v2	prot, cod, gff	<i>Saprolegnia genome</i> Sequencing Project, Broad institute of Harvard and MIT	http://www.broadinstitute.org/

Defining TE families

All protein sequences were searched for sequence similarity to RepBase Update Transposable Elements (Jurka et al. 2005) using TBLASTN (E-value cut-off E-3) and screened with

TransposonPSI (<http://transposonpsi.sf.net>). Protein families were labeled 'TE family' if at least one member was matched by one or both searching strategies.

2HOM block screening

Each gene on a scaffold was converted into its corresponding protein family, while storing the original gene and protein identifiers. On a scaffold, adjacent genes forming a block of different gene families were collected. If the block AB was succeeded by a gene corresponding to family A, this gene was skipped in order to avoid collecting AB and BA, while B is in fact a single gene. A scaffold string of ACBBAB would give blocks AC, CB and BA. If later the block BC was found, this was stored as a copy of the previously found CB. Likewise, a string containing DEED firstly detects the DE and secondly stores ED as its copy. All blocks with more than one copy in a single species were qualified 2HOM blocks, which might also occur (single or multiple copy) in other species

We assessed the chance to retrieve more or the same number of observed 2HOM blocks residing on the same scaffold by randomly reshuffling the complete genomes of the analyzed *Phytophthora* 10,000 times while keeping the genome structure (number of scaffolds, genes and 2HOM blocks) intact.

Construction of 2HOM block phylogenetic trees

For each 2HOM block with at least four copies across all species, protein sequences were aligned for both genes constituting the block, using MAFFT with the accuracy-oriented E-INS-i method (Kato et al. 2002). RAxML (Stamatakis 2006) was employed to estimate maximum likelihood phylogenetic trees based on the multiple sequence alignments using the substitution matrix WAG under the gamma model. We assessed the robustness with rapid bootstrap analysis conducting 100 bootstrap replicates.

Categorizing and counting individual 2HOM duplications

For each 2HOM block, timing of causal duplication(s) was deduced. The number of duplications underlying a block is based on the occurrences of that block; a triple-copy 2HOM block reflects two duplications. If no orthologous copies of a block exist, the duplication(s) is (are) added to the 'private'-category. If a block has one or more orthologous copies the phylogenetic trees were analyzed, provided that the block has more than three copies, and that trees contained a proper outgroup and were similar for both genes (6-21% of 2HOM blocks with gene trees could not be analyzed). Phylogenetically analysed blocks might add duplications to the 'private'-category and to the 'shared'-category. We also classify duplications as "shared" if they precede speciation of only a subset of *Phytophthora* lineages, so the resulting set of "shared" 2HOM block duplications also includes duplications that are more recent than the last common ancestor of all analyzed *Phytophthora*. Blocks that were present twice in a single species and once in another species could not be included here. In total, 15-23 % of inferred duplications could not be categorized.

K_s-based timing of paralogous and orthologous gene pairs

For each protein family, all possible gene pairs were obtained. Coding sequences were aligned using protein-guided nucleotide sequence alignment with EMBOSS and the TreeBeST tool 'backtrans' (Ponting 2009). The resulting pairwise alignments were used to calculate K_s-values with CODEML (Yang 2007), specifying the equilibrium codon frequencies by the average nucleotide frequencies at the three codon positions (F3x4). For a better representation of full paraneome K_s-distributions, the values were corrected for the fact that not all possible gene pairs within a family reflect a duplication. The reweighting procedure applied here has been previously described by Maere *et al.* (2005). For distributions of orthologs we selected

families with a single copy in each of the species order to avoid including out-paralogs. Values below 0.1 (possible alleles) and higher than 5 (possible saturation effects) were excluded from the distributions.

Phylogenomic inference of ancient gene duplications

For all protein families, neighbour-joining trees were built using QuickTree (Howe et al. 2002) and automatically reconciled with the species tree using NOTUNG (Chen et al. 2000). For branches with weak bootstrap support (<80), rearrangements were performed in order to minimize assumed duplication and loss events. Based on all gene trees, each branch on the species tree was assigned a number of gene duplications and losses.

References

- Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z et al. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* 25:3389-3402.
- Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: A program for dating gene duplications and optimizing gene family trees. *Journal of Computational Biology* 7:429-447.
- Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30:1575-1584.
- Howe K, Bateman A, Durbin R. 2002. QuickTree: Building huge neighbour-joining trees of protein sequences. *Bioinformatics* 18:1546-1547.
- Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O et al. 2005. Repbase update, a database of eukaryotic repetitive elements. *Cytogenetic and Genome Research* 110:462-467.
- Katoh K, Misawa K, Kuma K, Miyata T. 2002. MAFFT: A novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research* 30:3059-3066.
- Ponting C. 2009. TreeBeST: Tree building guided by species tree.
- Stamatakis A. 2006. RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* 22:2688-2690.

Yang Z. 2007. PAML 4: Phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586-1591.