Footprints of evolution:

the dynamics of effector genes in the *Phytophthora* genome

Rays Jiang

Promoter:
Prof. dr. ir. P. J. G. M. de Wit
Hoogleraar in de Fytopathologie
Wageningen Universiteit

Co-Promoter:
Dr. ir. F. P. M. Govers
Universitair hoofddocent
Laboratorium voor Fytopathologie
Wageningen Universiteit

Promotiecommissie:
Prof. dr. J. A. M. Leunissen, Wageningen Universiteit
Prof. dr. ir. J. Bakker, Wageningen Universiteit
Prof. dr. M. A. M. Groenen, Wageningen Universiteit
Prof. dr. R. F. Hoekstra, Wageningen Universiteit

Dit onderzoek is uitgevoerd binnen de onderzoekschool Experimental Plant Sciences

Rays Jiang

Footprints of evolution:

the dynamics of effector genes in the *Phytophthora* genome

Proefschrift
ter verkrijging van de graad van doctor
op gezag van de rector magnificus
van Wageningen Universiteit,
Prof. dr. M. J. Kropff,
in het openbaar te verdedigen
op woensdag 29 maart 2006
des namiddags te vier uur in de Aula

Rays H. Y. Jiang (2006)

Footprints of evolution:
the dynamics of effector genes in the *Phytophthora* genome

PhD Thesis Wageningen University, The Netherlands
with summaries in English and Dutch

ISBN 90-8504-359-x

Contents

Chapter 1

General introduction

Late blight and microbiology

Few plant pathogens have changed human history as well as scientific concepts as *Phytophthora infestans*. It caused famine and immigration, and at the same time, the birth of a new scientific discipline.

Potato (*Solanum tuberosum*) was introduced from the New World into Europe by the Spanish Conquistadors around 1570. By the mid 19th century, the planting was widespread and the population of Ireland became entirely dependant on this single crop. The genetic uniformity of the potatoes throughout Europe and the unusual cool summer of 1845 set the stage for the devastating famine caused by potato late blight. As a consequence one million people starved and another million emigrated to North America (Bourke 1993).

Microbes as causal agents of diseases are unknown to the world by then. In 1861, Anton de Bary convincingly showed that a white mold was the cause of the late blight epidemic and not the consequence. This was not only the beginning of the scientific discipline of plant pathology but also the coining of the concept ‘germ caused disease’. A year later, the common belief of ‘Spontaneous Generation’ was replaced by the ‘Germ Theory’ of Louis Pasteur.

The evolving population of *P. infestans* has always been an extraordinary tale to reveal the relationship between microbes and mankind. Native to South or Central America from where its host originated, *P. infestans* spread to the rest of the world with human aid about 150 years ago. The world population of *P. infestans* consisted of a single clonal lineage (US-1) until recently. Again by human involvement in 1980’s, more strains have been spread from Central America and the once uniform population has become complex since (Ristaino 2002).

The path to pathogenicity: gain and loss

The genus *Phytophthora* belongs to the oomycetes that are placed in the kingdom Stramenopila. Together with several other major groups such as animals, fungi, plants, amoebezoa and alveolates, stramenopiles make up the "crown" of the eukaryotic phylogenetic tree by massive radiation (Baldauf et al. 2000; Baldauf 2003).

Stramenopiles include heterotrophic organisms like *Phytophthora* as well as autotrophic ones such as the delicate unicellular diatom and the magnificent giant kelp. The last common ancestor of Stramenopiles was most likely photosynthetic with a plastid captured 1,300 million years ago by secondary endosymbiosis (Yoon et al. 2002; Yoon et al. 2004; Bachvaroff, Puerta, and Delwiche 2005).

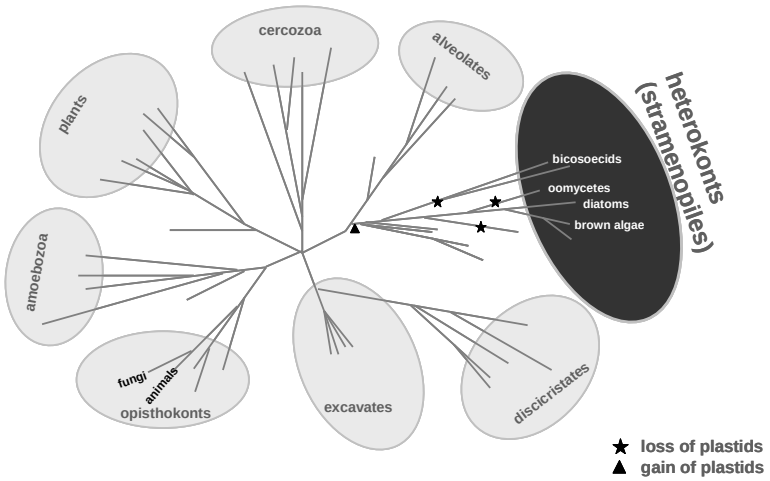
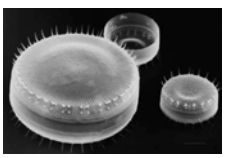


Figure 1. **Evolution of heterotrophism of oomycetes.** The eukaryotic phylogenetic tree was modified from Baldauf (2003). The events of losing plastid are arbitrary.

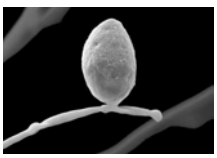
All known oomycetes are heterotrophic, but they probably have a photosynthetic past. Analysis of 5S and 16S rRNA secondary and primary structure indicates that oomycetes are derived from a group of heterokont algae (Wolters and Erdmann 1988). Also recent genome sequence data show that they share photosynthetic genes with autotrophic stramenopiles (Tyler et al., in preparation). Therefore, oomycetes have been through an evolutionary process of 'gain-and-loss' of plastids (Figure 1). Additionally, some genera like *Phytophthora,* have evolved weaponry to attack plants and to earn a parasitic living.

Convergent evolution of plant destroyers

Plant diseases are caused by multitudes of organisms such as bacteria, nematodes and viruses, but the vast majority of plant pathogens are fungi and oomycetes. By occupying similar niches, oomycete pathogens have gone through convergent evolution with the plant pathogenic fungi (Figure 2).



Stephanodiscus niagarae, a diatom in fresh water. (Photo courtesy of Mark Edlund, Science Museum of Minnesota)



Phytophthora infestans, an oomycete © Minister of Public Works and Government Services Canada 2002

Puccinia graminis, a fungus sporulating on a wheat leaf © 2003 RC Evans, Biology Department, Acadia University

Figure 2. **Convergent evolution of oomycetes and fungi.** *Phytophthora* shares phylogenetic affinity with diatoms but shows great morphological resemblance to fungi.

The kingdom Stramenopila is phylogenically distant from the kingdom Fungi which is more closely related to the animal kingdom (Baldauf et al. 2000; Margulis and Schwarts 2000; Baldauf 2003) (Figure 1). The distinct phylogenetic position of oomycetes and fungi explains their fundamental differences in physiology, biochemistry and genetics. For example, *Phytophthora* species do not synthesize sterols but require an exogenous sterol source, and the cell walls of *Phytophtora* are mainly composed of cellulose instead of chitin which is the major component of fungal cell walls. During its life cycle, *Phytophthora* is mainly diploid while fungi are haploid. However, both oomycetes and fungi have filamentous growth, employ similar infection structures such as appressoria and haustoria, and share an extensive repertoire of cell wall degrading enzymes to degrade host tissues (Erwin and Ribeiro 1996; Latijnhouwers, de Wit, and Govers 2003). Convergent evolution has shaped the weaponry of these two major groups of plant pathogens.

A versatile group of pathogens

Oomycetes include organisms with diverse life styles: free living saprophytes, animal parasites, and plant pathogens. Example of destructive animal parasites are *Saprolegnia* and *Aphanomyces* species. They cause infections on fishes and pose serious threat to aquaculture and ornamental fish cultures (Neish and Hughes 1980). Oomycete plant pathogens have various parasitic life styles: the genera *Phytophthora, Pythium* and *Aphanomyces* are necrotropic or hemi-biotrophic pathogens, whereas the genera *Albugio, Peronospora* and *Hyaloperonospora* are obligate biotrophic pathogens which exclusively survive on living plants.

Phytophthora is a genus comprised of over 65 species entirely pathogenic on various plants. The *Phytophthora* taxa form a recently evolved monophyletic group (Cooke et al. 2000; Kroon et al. 2004). However, the geographical presence of *Phytophthora* is widespread, ranging from the tropics to temperate regions (Erwin and Ribeiro 1996).

Many *Phytopthora* species cause severe damage in agriculture, forestry and natural habitats. *P. infestans* (causing potato late blight) and *Phytophthora sojae* (causing soybean root rot) are economically important pathogens. *Phytophthora ramorum* is a recently discovered species destroying the oak trees along the west coast of the USA (Rizzo, Garbelotto, and Hansen 2005). These three pathogens differ in their genome sizes, sexual behavior and host specificity. Their genome sizes range from 65 Mb in *P. ramorum* to 240 Mb in *P. infestans*. The mating system of *P. sojae* is homothallic whereas that of the other two is heterothallic. Many plants can be infected by *P. ramorum*, whereas only few are susceptible to *P. sojae* and *P. infestans*.

The double life of effectors

Many biotropic oomycetes and fungi grow extracellularly in the plant hosts and they commonly secrete a variety of molecules presumably to promote infection (Knogge 1998). These molecules with a potential role in virulence or pathogenicity are termed virulence factors or effectors. According to the guard hypothesis, pathogens use effectors to interact with and modify host targets. Plant hosts can be equipped with resistance proteins that guard the virulence target, detect invasion and initiate defenses (Dangl and Jones 2001). So if plants are able to perceive the effector/host target complex the infection processes may be halted, despite the intrinsic virulence function of effectors. Effectors that trigger plant defense responses are called avirulence factor or elicitors (van't Slot and Knogge 2002).

Elicitor activity can result in an array of defense reactions, such as an oxidative burst, production of phytoalexins and pathogenesis related (PR) proteins. In particular, a form of programmed cell death called the hypersensitive response (HR) is effective to stop obligate biotrophic pathogens from invading plants (Nimchuk et al. 2003).

An effector protein will only show elicitor activity if it betrays the identity of the pathogen to the host. This happens in plants carrying genes that (in)directly recognize the effector. In a few cases the intrinsic function of elicitors is known, such as the metalloprotease activity of AvrPita in *Magnaporthe grisea* (Jia et al. 2000), chitin binding of AVR4 in *Cladosporium fulvum* (van den Burg et al. 2004), the transglutaminase activity of GPE1 in *P. sojae* (Brunner et al. 2002) and sterol binding of elicitin in *Phytophthora cryptogea* (Mikes et al. 1997).

The dual role of effectors has determined their close association with hosts, their intrinsic function can be manipulation, poisoning or counter-defense, but at the same time, their elicitor activity may alarm the plant surveillance systems.

Recognition or evading recogition

Similar to the animal immune system, the active plant defense response will only be triggered after recognition events between the host and pathogen. Perception of general and specific pathogen-associated molecules can lead to broad or specific resistance, respectively.

A very high degree of cultivar-specific resistance is induced by a particular elicitor derived from a pathogen genotype. Flors gene-for-gene model derived from the flax-flax rust system in the early 1940s (Flor 1942), elegantly explained the specificity between host and pathogen genotypes. The hypothesis states that the product of each gene determining resistance (*R* gene) in the host, specifically interacts with the product of a corresponding gene for avirulence (*Avr* gene) in the pathogen. This high specificity

appears to be prevalent in plant-pathogen interactions because the gene-for-gene model has explained the genetics of diseases caused by viruses, bacteria, fungi, oomycetes and nematodes .

Disease resistance to *P. infestans* can be broad or specific. Non-host resistance offers protection at the genus level, whereas race-specific resistance occurs at the cultivar level (Black et al. 1953). Protection by race specific resistance in potato is governed by a large set of *P. infestans Avr* and potato *R* genes, respectively. Eleven major *R* genes were introgressed from *Solanum demissum* into potato to provide resistance against *P. infestans*. According to Flors' hypothesis the 11 *R* genes suggest the presence of 11 corresponding *Avr* genes. Genetic analyses on both the potato and *P. infestans* has confirmed the gene-for-gene model (Black et al. 1953; Mastenbroek 1953; van der Lee et al. 2001).

Unfortunately, typical 'boom and bust' cycles have been observed after *R* gene deployment. Resistance conferred by the *R* genes in the field was rapidly overcome by new races of *P. infestans*. As resistance is lost so quickly, the *Avr* genes presumably undergo rapid changes enabling *P. infestans* to evade detection by the *R* gene introgressed in potato. Host *R* gene imposed selection pressure thus causes adaptive evolution of phytopathogens. Point mutations, deletions, genome rearrangements and gene amplifications may play important roles in the process.

The paces of elicitor evolution

Elicitors form a very heterogeneous group of molecules possibly because of their different intrinsic functions. From an evolutionary perspective, classification can be based on the time point of elicitor emergence. Early evolved elicitors may have homologues across kingdoms. The most prominent example is the family of Nep1-like proteins (NLPs) present in bacteria, fungi and stramenopiles (Qutob, Kamoun, and Gijzen 2002; Pemberton and Salmond 2004). The most recent elicitors occur in only one or few genera and can be viewed as innovations in the group. Most elicitor genes characterized in *Phytophthora* belong to the latter group.

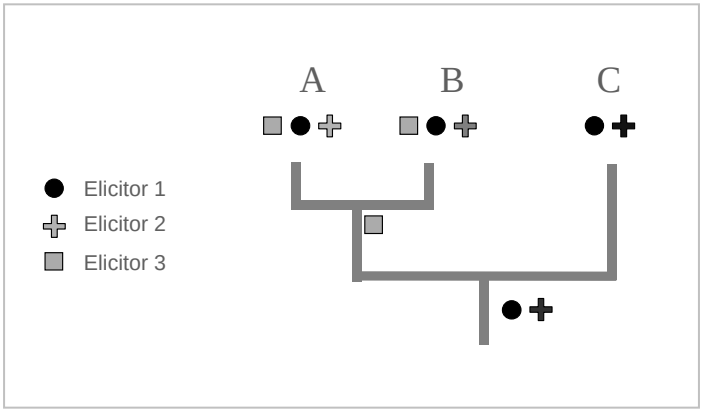


Figure 3. **Difference time of emergence and difference paces of evolution of elicitor genes**. A, B and C represent three species. Elicitor 1 and elicitor 2 occurred earlier than elicitor 3. Elicitor 2 became highly divergent which is indicated by different shading.

Elicitors do not only occur at different time points during evolution but also evolve at different paces (Figure 3). Assuming an elicitor was present in the ancestral *Phytophthora* before speciation, then after evolving into the individual species, the fast changing elicitor will have highly polymorphic sequences (Elicitor 2 in Figure 3). In contrast, the selection constraint will result in sequence conservation of other elicitors (Elicitor 1 in Figure 3). Rapid evolution of some elicitors may be due to the need to evade plant detection systems, as exemplified by the positive selection of *SCR74* genes in *P. infestans* (Liu et al. 2005) and the extreme diversity in *Atr13* in *Hyaloperonospora parasitica* (Allen et al. 2004). On the other hand, conservation of elicitors may be traced back to their indispensable physiological function. For example, the need for sterol carriers for the nearly sterol auxotroph oomycetes may have led to the emergence of the highly conserved elicitins.

The different phylogenetic distribution and level of sequence divergence may reveal possible roles of elicitors in the interaction with host plants. Some of the highly divergent oomycete elicitors, for example, are race specific (i.e. AVR3a and AVR1b of *P. infestans* and *P. sojae*, respectively), and are involved in the highly specific gene-for-gene interactions with major NBS-LRR resistance genes (Shan et al. 2004; Armstrong et al. 2005).

The intriguing G and C

The codon usage bias has caught the curiosity of scientists since the codon was cracked. Codons are universal sets of nucleotide triplets that specify the 20 amino acids in all organisms. The 64 possible combinations lead to the redundancy of codons, and as a result, preference for certain synonymous codons in a species causes codon bias. With the rising of the genome era, large data sets have become available enabling to address this question. Many organisms have been found to have such a bias and *Phytophthora* is no exception. With the current knowledge, two major mechanisms are considered to be responsible for such biases in codon usage: selection pressure and mutation bias (Sharp et al. 1993; Knight, Freeland, and Landweber 2001).

Table 1. **Genomes with extreme base compositions.** Table was adapted from G. Glockner (2000).

Species	Taxonomy	AT content	comment
Plasmodium falciparum	protists (Apicomplexa)	80%	causes malaria
Dictyostelium discoideum	protists (Dictyosteliida)	78%	studies on cell motility cytoskeleton, signal transduction
Tetrahymena thermophila	protists (Ciliates)	73%	studies on DNA rearrangement, chromatin assembly
Brugia malayi	nematodes	75%	causes elephantiasis
Borrelia burgdorferi	bacteria	71%	causes Lyme disease
Chlamydomonas reinhardtii	green plants	~ 37%	studies on photosynthetic pathways

Mutational bias is a global force acting on all sequences whereas selection pressure is more of a local force. Because of the overall action of mutational bias, it differentiates the codon usage between different organisms and leads to species-specific codon bias (Chen et al. 2004). Mutational bias is able to shift the whole genome to an extreme nucleotide composition (Table 1) such as very high AT content (>70%) in the free living protist *Dictyostelium discoideum* (Eichinger et al. 2005), the malaria parasite *Plasmodium falciparum* (Bowman et al. 1999) and the bacterium *Borrelia burgdorferi* causing Lyme disease (Fraser et al. 1997). These genomes have been shaped by mutational bias independently because of their entirely different phylogenetic positions (Table 1). Selective pressure is a local force because it only acts on coding regions. Co-adaptation of codon usage and tRNA content could establish efficient protein expression. Such selective pressure to optimize translation is expected to be stronger for genes expressed at high levels. In many organisms, preference of codons has been correlated to either the abundance of tRNAs or amount of transcripts (Duret and Mouchiroud 1999; Kanaya et al. 1999).

The coding regions of *Phytophthora* have been found to have elevated GC content. High GC3 (GC content at the 3rd codon position) has been reported to cause the increase of GC content and is related to codon bias in *P. infestans* (Jiang et al. 2005; Randall et al. 2005). It is intriguing to investigate whether mutation bias or selection pressure is at work or not.

Mobile elements can be largely viewed as genetic parasites in the genome. In *Drosophila melanogaster, Arabidopsis thaliana, Caenorhabditis elegans, Saccharomyces cerevisiae,* and *Homo sapiens* mobile elements were reported to exhibit overall AT-richness regardless of the GC content of host genomes (Lerat, Capy, and Biemont 2002). However, some mobile elements in the *P. infestans* genome show a high GC content similar to the coding regions in the genome (Jiang et al. 2005). This raises the question whether mimicry of the codon usage is used by the most widespread *Phytophthora* retrotransposons.

The genome era of *Phytophthora*

Molecular genetic linkage maps have been developed for *P. infestans* and *P. sojae* (Whisson et al. 1995; van der Lee et al. 1997; van der Lee et al. 2004). These genetic data are essential for positional cloning of genes linked to particular phenotypes and are valuable in assessing genome re-arrangements. Avirulence genes and mating factor associated genes have been placed on the genetic maps and with the help of molecular markers, physical maps of several regions of the genome were obtained (Judelson, Spielman, and Shattock 1995; van der Lee et al. 2001; Whisson et al. 2001; MacGregor et al. 2002; Whisson et al. 2004).

The draft sequences of the *P. sojae* and *P. ramorum* genomes have been completed at the Joint Genome Institute (JGI) and were released for public access in 2004. These are the first sequenced

oomycetes, and together with the genome of the marine diatom *Thalassiosira pseudonana,* the only genomes available so far in the kingdom Stramenopila. For the late blight pathogen *P. infestans*, a one year pilot study has just been completed at the MIT Broad Institute. In this pilot study, an optimal sequencing strategy for *P. infestans* was determined. Now the genome sequencing is under way and the draft genome sequence is expected to be released in 2006.

To the delight of the genome analysts, large EST data sets were generated from a wide range of tissues and culture conditions in *P. sojae and P. infestans*. Pilot studies (Kamoun et al. 1999b; Qutob et al. 2000) were the incentive for larger projects. Currently, 30,000 *P. sojae* ESTs, defining 7200 unigenes are available (http://staff.vbi.vt.edu/estap). A further 75,000 ESTs from *P. infestans*, defining 18,000 unigenes have been produced by the Syngenta-*Phytophthora* research consortium (Randall et al. 2005).

Expression profiling has been used for displaying overall transcript differences to search for several candidate genes. cDNA-AFLP has been used to analyze stage-specific gene expression (Avrova et al. 2003), to reveal downstream targets of the G-protein signaling pathway (Dong et al. 2004) and to search for candidate of avirulence genes in *P. infestans* (Guo et al. 2006). This relatively cheap and highly sensitive method has generated large amount of data. Computational methods have been applied to link cDNA-AFLP data with the genomic resources to assist rapid annotation and cloning of candidate genes ((Dong et al. 2004;Qin et al. 2001). In addition to cDNA-AFLP, an Affymetrix® array (i.e., the Syngenta custom designed *Phytophthora* GeneChip) has been developed based on the large unigene set of *P. infestans* (Randall et al. 2005). Differences in gene expression between various life stages and culture conditions, and between different strains have yielded many candidate genes to be explored.

The plasticity of the *Phytophthora* genome causes variation between strains in ploidy level and chromosome number, and gives rise to trisomic regions (Tooley 1987; Goodwin et al. 1992; van der Lee et al. 2004). One efficient way to explore this variation is comparative genomic hybridization (Vissers et al. 2005). In *P. infestans*, several gene amplifications have been identified by hybridizing the Affymetrix® array with genomic DNA derived from different strains and one gene amplification has been associated with an avirulence locus (Jiang et al, in preparation).

Comparative genomics has become a powerful approach to reveal the commonalties and differences between genomes as sequence data accumulate. Genomes can be compared at different phylogenetic distances to address different questions (Figure 4). A basic understanding can be gained by genomic comparisons at very long phylogenetic distances. For example, comparing the genomes of yeast, worms, and flies revealed that these metazoans encode many of the same proteins after more than 1 billion years of separation (Rubin et al. 2000).

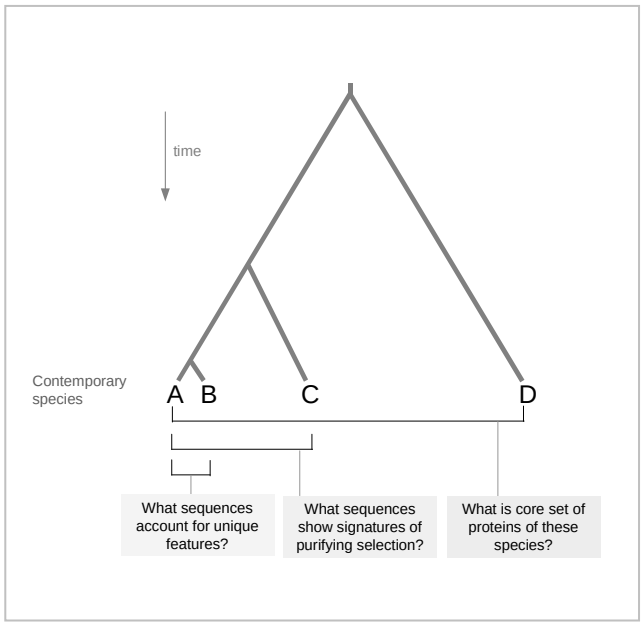


Figure 4. **Comparisons of genomes at different phylogenetic distances are appropriate to address different questions**.
A generalized phylogenetic tree is shown, leading to four different organisms, with A and D representing the most distantly related species. Modified from R. C. Hardison (2003).

Features specifying organisms may be obtained by comparing more closely related species such as *P. sojae* and *P. ramorum*. The phylogenetic distance between these two *Phytophthora* species may be larger than the distance between human and chimpanzee, depicted in Figure 4 as two very closely related species A and B, but smaller than species with a moderate distance (ca. 70-100 million years divergence between e.g. A and C). In plant pathogens, elicitor genes are likely to be under host selection and to undergo genomic rearrangements in order to avoid plant recognition, for example deletion of *Avr* genes as reported for several pathogenic fungi (van Kan, van den Ackerveken, and de Wit 1991; Rohe et al. 1995; Jia et al. 2000; Westerink et al. 2004). Because of their different host ranges, *P. sojae* and *P. ramorum* are expected to have species specific pathogenicity and virulence genes and at those loci their genomes may be divergent (Figure 5).

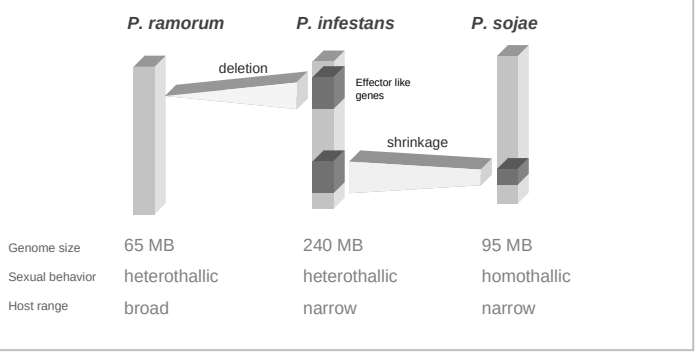


Figure 5. **Effector genes may exhibit gene rearrangement patterns.** The major differences in genome sizes and properties of three *Phytophthora* species are listed.

Functional genomics has been adopted to elucidate the role of *Phytophthora* derived proteins. Computational tools have been developed to mine the sequence data sets for proteins functioning in virulence and pathogenicity. For example, an algorithm called PexFinder was developed to identify genes encoding secreted proteins in EST data bases (Torto et al. 2003). Experimental essays were also established to test the function of *Phytophthora* genes in pathogenesis. Elicitor activity can be tested in *in planta* assays based on necrosis inducing activity that is indicative for their role in plant defense responses (Kamoun et al. 1999a; Qutob, Kamoun, and Gijzen 2002; Torto et al. 2003). Functional analysis of *Phytophthora* genes using gene silencing revealed that INF1 elicitin acts as a determinant of non-host resistance (Kamoun et al. 1998), and that heterotrimeric G-proteins and a bZip transcription factor play important roles in development and virulence (Latijnhouwers and Govers 2003; Latijnhouwers et al. 2004; Blanco and Judelson 2005).

Scope of this thesis

The genome structure, transcriptome and elicitor reservoir of the destructive pathogen *Phytophthora* were analyzed by exploring the genome resources with bioinformatical as well as molecular biological tools.

We used mainly molecular biological methods to identify avirulence genes (Chapter 2 and Chapter 3). With computational methods, we investigated the basic genome features such as base composition (Chapter 8), mobile elements (Chapter 4) and co-linearity between the genomes (Chapter 7). We focused on secreted protein analysis such as sequence diversity (Chapter 7), genome organization (Chapter 4, Chapter 5 and Chapter 6) and patterns of evolution (Chapter 5, Chapter 6 and Chapter 7).

In the potato-*P. infestans* interaction avirulence factors induce resistance with high specificity following the gene-for-gene model. We aimed at cloning *Avr* genes to be able to unravel the interaction mechanism at the molecular level and explain the rapid loss of avirulence after *R* gene deployment. We adopted a cDNA-AFLP based strategy to identify avirulence-associated transcripts by comparing strains with different virulence phenotypes (Chapter 2). To further select *Avr* gene candidates, bioinformatics tools as well as the genetic segregation pattern in a mapping population were used to find transcripts associated with virulence. These transcriptome markers were combined with the previously identified genetic markers and physical contigs spanning the *Avr3b-Avr10-Avr11* locus, and this resulted in identification of genes located at this locus (Chapter 3). The *Avr3b-Avr10-Avr11* locus seems to be extremely dynamic. The avirulence haplotype is comprised of an amplified gene cluster which is absent in a virulent haplotype, resulting in hemizygousity at this *Avr* locus.

Elicitins belong to a group of extracellular elicitor proteins that cause a hypersensitive response (HR) in tobacco. To investigate the diversity and the genome organization of the family members, we performed

database mining and Southern blot analysis (Chapter 4 and Chapter 5). Elicitins were found to be ubiquitous among *Phytophthora* species and they belong to one of the most highly conserved and complex protein families in the *Phytophthora* genus. Many of the family members are clustered in the genome. To understand the evolution of this family, phylogeny construction was performed showing that the formation of the family has occurred in the common ancestor giving rise to *Phytophthora* species. We used bioinformatic tools to classify different groups of elicitins and to assign various functions to the groups.

The elicitor genes and their genomic context were investigated to gain a better understanding of their evolution (Chapter 6). Overall co-linearity was found when large regions of *P. sojae* and *P. ramorum* were compared. However, genome rearrangements such as insertion, deletion and expansion were also found. The hotspots for rearrangements often harbor genes encoding secreted proteins. For example, cutinase genes and a gene family encoding Cys-rich proteins appeared to have expanded in *P. sojae*. Deletion is especially common for genes encoding proteins with the 'RXLR-DEER' motif that occur exclusively in oomycete effectors.

The whole reservoir of secreted proteins was revealed by computational methods. A secretome comprised of over 1000 proteins is present in both *P. sojae* and *P. ramorum* (Chapter 7). We investigated all secreted proteins for their sequence diversity and genome organization. Most secreted proteins belong to families and different families were found to evolve at a different pace. Surface anchored proteins, mating associated factors and ‘RXLR-DEER’ proteins are among the most fast evolving gene families in the genome.

The base composition was analyzed by making use of the complete genome sequence. In coding regions a high GC3 was found and this causes codon bias (Chapter 8). We hypothesize that both selective pressure and mutation bias drive codon bias in *Phytophthora* and we present evidence for the occurrence of both driving forces. Analysis was also performed with the most widespread groups of retrotransposons, and they show high GC3 and a codon bias that is similar to *Phytophthora* genes

References

Allen, R. L., P. D. Bittner-Eddy, L. J. Grenville-Briggs, J. C. Meitz, A. P. Rehmany, L. E. Rose, and J. L. Beynon. 2004. Host-parasite coevolutionary conflict between *Arabidopsis* and downy mildew. Science **306**:1957-1960.

Armstrong, M. R., S. C. Whisson, L. Pritchard, J. I. Bos, E. Venter, A. O. Avrova, A. P. Rehmany, U. Bohme, K. Brooks, I. Cherevach, N. Hamlin, B. White, A. Fraser, A. Lord, M. A. Quail, C. Churcher, N. Hall, M. Berriman, S. Huang, S. Kamoun, J. L. Beynon, and P. R. Birch. 2005. An ancestral oomycete locus contains late blight avirulence gene *Avr3a*, encoding a protein that is recognized in the host cytoplasm. Proc Natl Acad Sci USA **102**:7766-7771.

Avrova, A. O., E. Venter, P. R. J. Birch, and S. C. Whisson. 2003. Profiling and quantifying differential gene transcription in *Phytophthora infestans* prior to and during the early stages of potato infection. Fungal Genetics and Biology **40**:4-14.

Bachvaroff, T. R., S. M. V. Puerta, and C. F. Delwiche. 2005. Chlorophyll c-containing plastid relationships based on analyses of a multigene data set with all four chromalveolate lineages. Mol Biol Evol **22**:1772-1782.

Baldauf, S. L. 2003. The deep roots of eukaryotes. Science **300**:1703-1706.

Baldauf, S. L., A. J. Roger, I. Wenk-Siefert, and W. F. Doolittle. 2000. A kingdom-level phylogeny of eukaryotes based on combined protein data. Science **290**:972-977.

Black, W., C. Mastenbroek, W. R. Mills, and L. C. Peterson. 1953. A proposal for an international nomenclature of races of *Phytophthora infestans* and of genes controlling immunity in *Solanum demissum* derivatives. Euphytica **2**.

Blanco, F. A., and H. S. Judelson. 2005. A bZIP transcription factor from *Phytophthora* interacts with a protein kinase and is required for zoospore motility and plant infection. Mol Microbiol **56**:638-648.

Bourke, A. 1993. The visitation of god? The potato and the great Irish famine. Lilliput Press Ltd, Dublin.

Bowman, S., D. Lawson, D. Basham, D. Brown, T. Chillingworth, C. M. Churcher, A. Craig, R. M. Davies, K. Devlin, T. Feltwell, S. Gentles, R. Gwilliam, N. Hamlin, D. Harris, S. Holroyd, T. Hornsby, P. Horrocks, K. Jagels, B. Jassal, S. Kyes, J. McLean, S. Moule, K. Mungall, L. Murphy, K. Oliver, M. A. Quail, M. A. Rajandream, S. Rutter, J. Skelton, R. Squares, S. Squares, J. E. Sulston, S. Whitehead, J. R. Woodward, C. Newbold, and B. G. Barrell. 1999. The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. Nature **400**:532-538.

Brunner, F., S. Rosahl, J. Lee, J. J. Rudd, C. Geiler, S. Kauppinen, G. Rasmussen, D. Scheel, and T. Nurnberger. 2002. Pep-13, a plant defense-inducing pathogen-associated pattern from *Phytophthora* transglutaminases. EMBO J **21**:6681-6688.

Chen, S. L., W. Lee, A. K. Hottes, L. Shapiro, and H. H. McAdams. 2004. Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci USA **101**:3480-3485.

Cooke, D. E., A. Drenth, J. M. Duncan, G. Wagels, and C. M. Brasier. 2000. A molecular phylogeny of *Phytophthora* and related oomycetes. Fungal Genet Biol **30**:17-32.

Dangl, J. L., and J. D. G. Jones. 2001. Plant pathogens and integrated defence responses to infection. Nature **411**:826-833.

Dong, W. B., M. Latijnhouwers, R. H. Y. Jiang, H. J. G. Meijer, and F. Govers. 2004. Downstream targets of the *Phytophthora infestans* G alpha subunit PiGPA1 revealed by cDNA-AFLP. Mol Plant Pathol **5**:483-494.

Duret, L., and D. Mouchiroud. 1999. Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis, Drosophila,* and *Arabidopsis*. Proc Natl Acad Sci USA **96**:4482-4487.

Eichinger, L., J. A. Pachebat, G. Glockner, M. A. Rajandream, R. Sucgang, M. Berriman, J. Song, R. Olsen, K. Szafranski, Q. Xu, B. Tunggal, S. Kummerfeld, M. Madera, B. A. Konfortov, F. Rivero, A. T. Bankier, R. Lehmann, N. Hamlin, R. Davies, P. Gaudet, P. Fey, K. Pilcher, G. Chen, D. Saunders, E. Sodergren, P. Davis, A. Kerhornou, X. Nie, N. Hall, C. Anjard, L. Hemphill, N. Bason, P. Farbrother, B. Desany, E. Just, T. Morio, R. Rost, C. Churcher, J. Cooper, S. Haydock, N. van Driessche, A. Cronin, I. Goodhead, D. Muzny, T. Mourier, A. Pain, M. Lu, D. Harper, R. Lindsay, H. Hauser, K. James, M. Quiles, M. M. Babu, T. Saito, C. Buchrieser, A. Wardroper, M. Felder, M. Thangavelu, D. Johnson, A. Knights, H. Loulseged, K. Mungall, K. Oliver, C. Price, M. A. Quail, H. Urushihara, J. Hernandez, E. Rabbinowitsch, D. Steffen, M. Sanders, J. Ma, Y. Kohara, S. Sharp, M. Simmonds, S. Spiegler, A. Tivey, S. Sugano, B. White, D. Walker, J. Woodward, T. Winckler, Y. Tanaka, G. Shaulsky, M. Schleicher, G. Weinstock, A. Rosenthal, E. C. Cox, R. L. Chisholm, R. Gibbs, W. F. Loomis, M. Platzer, R. R. Kay, J. Williams, P. H. Dear, A. A. Noegel, B. Barrell, and A. Kuspa. 2005. The genome of the social amoeba *Dictyostelium discoideum*. Nature **435**:43-57.

Erwin, D. C., and O. K. Ribeiro. 1996. *Phytophthora* diseases worldwide. The American Phytopathological Society, St. Paul Minnesota USA.

Flor, H. H. 1942. Inheritance of pathogenicity of *Melampsora lini*. Phytopathology **32**:653-669.

Fraser, C. M., S. Casjens, W. M. Huang, G. G. Sutton, R. Clayton, R. Lathigra, O. White, K. A. Ketchum, R. Dodson, E. K. Hickey, M. Gwinn, B. Dougherty, J. F. Tomb, R. D. Fleischmann, D. Richardson, J. Peterson, A. R. Kerlavage, J. Quackenbush, S. Salzberg, M. Hanson, R. vanVugt, N. Palmer, M. D. Adams, J. Gocayne, J. Weidman, T. Utterback, L. Watthey, L. McDonald, P. Artiach, C. Bowman, S. Garland, C. Fujii, M. D. Cotton, K. Horst, K. Roberts, B. Hatch, H. O. Smith, and J. C. Venter. 1997. Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature **390**:580-586.

Glöckner, G. 2000. Large Scale Sequencing and Analysis of AT Rich Eukaryote Genomes. Current Genomics **1**:289-299.

Goodwin, S. B., L. J. Spielman, J. M. Matuszak, S. N. Bergeron, and W. E. Fry. 1992. Clonal diversity and genetic differentiation of *Phytophthora infestans* populations in Northern and central Mexico. Phytopathology **82**:955-961.

Guo, J., R. H. Y. Jiang, L. Kamphuis, and F. Govers. 2006. A cDNA-AFLP based strategy to identify transcripts associated with avirulence in *Phytophthora infestans*. Fungal Genet Biol. in press

Hardison, R. C. 2003. Comparative genomics. PLoS Biol **1**:E58.

Jia, Y., S. A. McAdams, G. T. Bryan, H. P. Hershey, and B. Valent. 2000. Direct interaction of resistance gene and avirulence gene products confers rice blast resistance. EMBO J **19**:4004-4014.

Jiang, R. H., A. L. Dawe, R. Weide, M. van Staveren, S. Peters, D. L. Nuss, and F. Govers. 2005. Elicitin genes in *Phytophthora infestans* are clustered and interspersed with various transposon-like elements. Mol Genet Genomics **273**:20-32.

Jiang, R. H., B. M. Tyler, S. C. Whisson, A. R. Hardham, and F. Govers. 2006. Ancient origin of elicitin gene clusters in *Phytophthora* genomes. Mol Biol Evol **23**:338-351.

Judelson, H. S., L. J. Spielman, and R. C. Shattock. 1995. Genetic mapping and non-mendelian segregation of mating-type loci in the oomycete, *Phytophthora infestans*. Genetics **141**:503-512.

Kamoun, S., G. Honee, R. Weide, R. Lauge, M. Kooman-Gersmann, K. de Groot, F. Govers, and P. de Wit. 1999a. The fungal gene *AVR9* and the oomycete gene *inf1* confer avirulence to potato virus X on tobacco. Mol Plant-Microbe Interact **12**:459-462.

Kamoun, S., P. Hraber, B. Sobral, D. Nuss, and F. Govers. 1999b. Initial assessment of gene diversity for the oomycete pathogen *Phytophthora infestans* based on expressed sequences. Fungal Genet Biol **28**:94-106.

Kamoun, S., P. van West, V. Vleeshouwers, K. E. de Groot, and F. Govers. 1998. Resistance of *Nicotiana benthamiana* to *Phytophthora infestans* is mediated by the recognition of the elicitor protein INF1. Plant Cell **10**:1413-1425.

Kanaya, S., Y. Yamada, Y. Kudo, and T. Ikemura. 1999. Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene **238**:143-155.

Knight, R. D., S. J. Freeland, and L. F. Landweber. 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol **2**:0010.11-13.

Knogge, W. 1998. Fungal pathogenicity. Current Opinion Plant Biol **1**:324-328.

Kroon, L. P., F. T. Bakker, G. B. Van Den Bosch, P. J. Bonants, and W. G. Flier. 2004. Phylogenetic analysis of *Phytophthora* species based on mitochondrial and nuclear DNA sequences. Fungal Genet Biol **41**:766-782.

Latijnhouwers, M., P. J. de Wit, and F. Govers. 2003. Oomycetes and fungi: similar weaponry to attack plants. Trends Microbiol **11**:462-469.

Latijnhouwers, M., and F. Govers. 2003. A *Phytophthora infestans* G-protein beta subunit is involved in sporangium formation. Eukaryot Cell **2**:971-977.

Latijnhouwers, M., W. Ligterink, V. G. Vleeshouwers, P. van West, and F. Govers. 2004. A Galpha subunit controls zoospore motility and virulence in the potato late blight pathogen *Phytophthora infestans*. Mol Microbiol **51**:925-936.

Lerat, E., P. Capy, and C. Biemont. 2002. Codon usage by transposable elements and their host genes in five species. J Mol Evol **54**:625-637.

Liu, Z., J. I. Bos, M. Armstrong, S. C. Whisson, L. da Cunha, T. Torto-Alalibo, J. Win, A. O. Avrova, F. Wright, P. R. Birch, and S. Kamoun. 2005. Patterns of diversifying selection in the phytotoxin-like *scr74* gene family of *Phytophthora infestans*. Mol Biol Evol **22**:659-672.

MacGregor, T., M. Bhattacharyya, B. Tyler, R. Bhat, A. F. Schmitthenner, and M. Gijzen. 2002. Genetic and physical mapping of *Avrla* in *Phytophthora sojae*. Genetics **160**:949-959.

Margulis, L., and K. V. Schwarts. 2000. Five Kingdoms: an illustrated guide to the phyla of life on earth. W.H. Freeman and company, New, York.

Mastenbroek, C. 1953. Experiments on the inheritance of blight immunity in potatoes derived from *Solanum demissum* Lindl. Euphytica **2**:197-206.

Mikes, V., M. L. Milat, M. Ponchet, P. Ricci, and J. P. Blein. 1997. The fungal elicitor cryptogein is a sterol carrier protein. Febs Letters **416**:190-192.

Neish, G. A., and G. C. Hughes. 1980. Fungal diseases of fishes. T.W.F. Publications, New Jersey, USA.

Nimchuk, Z., T. Eulgem, B. F. Holt, 3rd, and J. L. Dangl. 2003. Recognition and response in the plant immune system. Annu Rev Genet **37**:579-609.

Pemberton, C. L., and G. P. C. Salmond. 2004. The Nep1-like proteins—a growing family of microbial elicitors of plant necrosis. Mol Plant Pathol **5**:353-359.

Qin, L., P. Prins, J. T. Jones, H. Popeijus, G. Smant, J. Bakker, and J. Helder. 2001. GenEST, a powerful bidirectional link between cDNA sequence data and gene expression profiles generated by cDNA-AFLP. Nucleic Acids Res **29**:1616-1622.

Qutob, D., P. T. Hraber, B. W. S. Sobral, and M. Gijzen. 2000. Comparative analysis of expressed sequences in Phytophthora sojae. Plant Physiol **123**:243-253.

Qutob, D., S. Kamoun, and M. Gijzen. 2002. Expression of a *Phytophthora sojae* necrosis-inducing protein occurs during transition from biotrophy to necrotrophy. Plant J **32**:361-373.

Randall, T. A., R. A. Dwyer, E. Huitema, K. Beyer, C. Cvitanich, H. Kelkar, A. M. Fong, K. Gates, S. Roberts, E. Yatzkan, T. Gaffney, M. Law, A. Testa, T. Torto-Alalibo, M. Zhang, L. Zheng, E. Mueller, J. Windass, A. Binder, P. R. Birch, U. Gisi, F. Govers, N. A. Gow, F. Mauch, P. van West, M. E. Waugh, J. Yu, T. Boller, S. Kamoun, S. T. Lam, and H. S. Judelson. 2005. Large-scale gene discovery in the oomycete *Phytophthora infestans* reveals likely components of phytopathogenicity shared with true fungi. Mol Plant Microbe Interact **18**:229-243.

Ristaino, J. B. 2002. Tracking historic migrations of the Irish potato famine pathogen, *Phytophthora infestans*. Microbes and Infection **4**:1369-1377.

Rizzo, D. M., M. Garbelotto, and E. M. Hansen. 2005. *Phytophthora ramorum:* integrative research and management of an emerging pathogen in California and Oregon forests. Annual Rev Phytopathol **43**:309-335.

Rohe, M., A. Gierlich, H. Hermann, M. Hahn, B. Schmidt, S. Rosahl, and W. Knogge. 1995. The Race-Specific Elicitor, Nip1, from the Barley Pathogen, *Rhynchosporium Secalis*, Determines Avirulence on Host Plants of the *Rrs1* Resistance Genotype. EMBO J **14**:4168-4177.

Rubin, G. M., M. D. Yandell, J. R. Wortman, G. L. Gabor Miklos, C. R. Nelson, I. K. Hariharan, M. E. Fortini, P. W. Li, R. Apweiler, W. Fleischmann, J. M. Cherry, S. Henikoff, M. P. Skupski, S. Misra, M. Ashburner, E. Birney, M. S. Boguski, T. Brody, P. Brokstein, S. E. Celniker, S. A. Chervitz, D. Coates, A. Cravchik, A. Gabrielian, R. F. Galle, W. M. Gelbart, R. A. George, L. S. Goldstein, F. Gong, P. Guan, N. L. Harris, B. A. Hay, R. A. Hoskins, J. Li, Z. Li, R. O. Hynes, S. J. Jones, P. M. Kuehl, B. Lemaitre, J. T. Littleton, D. K. Morrison, C. Mungall, P. H. O'Farrell, O. K. Pickeral, C. Shue, L. B. Vosshall, J. Zhang, Q. Zhao, X. H. Zheng, and S. Lewis. 2000. Comparative genomics of the eukaryotes. Science **287**:2204-2215.

Shan, W., M. Cao, D. Leung, and B. M. Tyler. 2004. The *Avr1b* locus of *Phytophthora sojae* encodes an elicitor and a regulator required for avirulence on soybean plants carrying resistance gene *Rps1b*. Mol Plant Microbe Interact **17**:394-403.

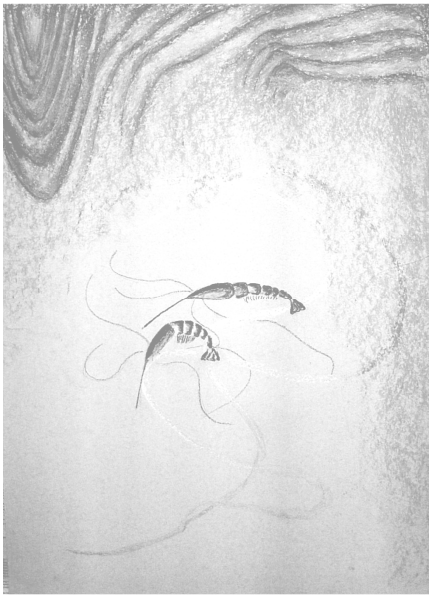
Sharp, P. M., M. Stenico, J. F. Peden, and A. T. Lloyd. 1993. Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans **21**:835-841.

Tooley, P. W., and Therrien, C. D. 1987. Cytophotometric determination of the nuclear DNA content of 23 Mexican and 18 non-Mexican isolates of *Phytophthora infestans*. Experimental Mycology **11**:19-26.

Torto, T. A., S. Li, A. Styer, E. Huitema, A. Testa, N. A. Gow, P. van West, and S. Kamoun. 2003. EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen *Phytophthora*. Genome Res **13**:1675-1685.

van den Burg, H. A., C. A. Spronk, S. Boeren, M. A. Kennedy, J. P. Vissers, G. W. Vuister, P. J. de Wit, and J. Vervoort. 2004. Binding of the AVR4 elicitor of *Cladosporium fulvum* to chitotriose units is facilitated by positive allosteric protein-protein interactions: the chitin-binding site of AVR4 represents a novel binding site on the folding scaffold shared between the

invertebrate and the plant chitin-binding domain. J Biol Chem **279**:16786-16796.
van der Lee, T., I. De Witte, A. Drenth, C. Alfonso, and F. Govers. 1997. AFLP linkage map of the oomycete *Phytophthora infestans*. Fungal Genet Biol **21**:278-291.
van der Lee, T., A. Robold, A. Testa, J. W. van't Klooster, and F. Govers. 2001. Mapping of avirulence genes in *Phytophthora infestans* with amplified fragment length polymorphism markers selected by bulked segregant analysis. Genetics **157**:949-956.
van der Lee, T., A. Testa, A. Robold, J. van 't Klooster, and F. Govers. 2004. High-density genetic linkage maps of *Phytophthora infestans* reveal trisomic progeny and chromosomal rearrangements. Genetics **167**:1643-1661.
van Kan, J. A. L., G. J. M. van den Ackerveken, and P. J. G. M. de Wit. 1991. Cloning and characterization of cDNA of avirulence gene *Avr9* of the fungal pathogen *Cladosporium fulvum*, causal agent of tomato leaf mold. Mol Plant Microbe Interact **4**:52-59.
van't Slot, K. A. E., and W. Knogge. 2002. A dual role for microbial pathogen-derived effector proteins in plant disease and resistance. Crit. Rev Plant Sci **21**:229-271.
Vissers, L. E. L. M., J. A. Veltman, A. Geurts van Kessel, and H. G. Brunner. 2005. Identification of disease genes by whole genome CGH arrays. Human Mol Genet **14**:r1-r9.
Westerink, N., B. F. Brandwagt, P. J. de Wit, and M. H. Joosten. 2004. *Cladosporium fulvum* circumvents the second functional resistance gene homologue at the *Cf-4* locus (*Hcr9-4E*) by secretion of a stable avr4E isoform. Mol Microbiol **54**:533-545.
Whisson, S. C., S. Basnayake, D. J. Maclean, J. A. G. Irwin, and A. Drenth. 2004. *Phytophthora sojae* avirulence genes *Avr4* and *Avr6* are located in a 24 kb, recombination-rich region of genomic DNA. Fungal Genet Biol **41**:62-74.
Whisson, S. C., A. Drenth, D. J. Maclean, and J. A. G. Irwin. 1995. *Phytophthora sojae* avirulence genes, RAPD, and RFLP markers used to construct a detailed genetic linkage map. Mol Plant Microbe Interact **8**:988-995.
Whisson, S. C., T. van derLee, G. J. Bryan, R. Waugh, F. Govers, and P. R. J. Birch. 2001. Physical mapping across an avirulence locus of *Phytophthora infestans* using a highly representative, large-insert bacterial artificial chromosome library. Mol Genet Genomics **266**:289-295.
Wolters, J., and V. A. Erdmann. 1988. Cladistic analysis of ribosomal RNAs--the phylogeny of eukaryotes with respect to the endosymbiotic theory. Biosystems **21**:209-214.
Yoon, H. S., J. D. Hackett, C. Ciniglia, G. Pinto, and D. Bhattacharya. 2004. A molecular timeline for the origin of photosynthetic eukaryotes. Mol Biol Evol **21**:809-818.
Yoon, H. S., J. D. Hackett, G. Pinto, and D. Bhattacharya. 2002. The single, ancient origin of chromist plastids. Proc Natl Acad Sci USA **99**:15507-15512.



Chapter 2

A cDNA-AFLP based strategy to identify transcripts associated with avirulence in *Phytophthora infestans*

Fungal Genetics and Biology 43(2006): 111–123

Jun Guo*, Rays H.Y. Jiang*, Lars G. Kamphuis and Francine Govers

* These authors contributed equally to the work

A cDNA-AFLP based strategy to identify transcripts associated with avirulence in *Phytophthora infestans*

Jun Guo, [a, b, c, #, †] Rays H.Y. Jiang, [a, #] Lars G. Kamphuis, [a, §] and Francine Govers[a, *]

[a] Plant Sciences Group, Laboratory of Phytopathology, Wageningen University, Binnenhaven 5, NL-6709 PD Wageningen and Graduate School Experimental Plant Sciences, The Netherlands
[b] Institute of Vegetables and Flowers, Chinese Academy of Agricultural Sciences, Beijing 100081, China
[c] College of Plant Protection, Northwest A & F University, Yangling Shaanxi 712100, China
[§] Present address: Australian Centre for Necrotrophic Fungal Pathogens, SABC, Murdoch University, Perth, WA 6150, Australia
[*] For correspondence: E-mail Francine.Govers@wur.nl; Tel. +31 317 483 138; Fax +31 317 483 412The GenBank accession numbers for the TDFs described in this paper are DW010060 to DW010198.

[#] These authors contributed equally to the work

Key words

cDNA-AFLP; Transcript profiling; *Phytophthora infestans*; Avirulence gene; BSA

Abstract

Expression profiling using cDNA-AFLP is commonly used to display the transcriptome of a specific tissue or developmental stage. Here cDNA-AFLP was used to identify transcripts in a segregating F1 population of *Phytophthora infestans,* the oomycete pathogen that causes late blight. To find transcripts derived from putative avirulence (*Avr)* genes germinated cyst cDNA from F1 progeny with defined avirulence phenotypes was pooled and used in a Bulked Segregant Analysis (BSA). Over 30.000 transcript derived fragments (TDFs) were screened resulting in 99 *Avr*-associated TDFs as well as TDFs with opposite pattern. With 142 TDF sequences homology searches and database mining was carried out. cDNA-AFLP analysis on individual F1 progeny revealed 100% co-segregation of four TDFs with particular AVR phenotypes and this was confirmed by RT-PCR. Two match the same *P. infestans* EST with unknown sequence and this is a likely candidate for *Avr4.* The other two are associated with the *Avr3b-Avr10-Avr11* locus*.* This combined cDNA-AFLP / BSA strategy is an efficient approach to identify *Avr*-associated transcriptome markers that can complement positional cloning.

Introduction

Many plant-pathogen interactions are governed by specific interactions between pathogen avirulence (*Avr*) genes and corresponding plant resistance (*R*) genes. An interaction where a corresponding pair of *R* gene and *Avr* gene is present and expressed, results in incompatibility and the plant is resistant. When one of the two is inactive or absent, the interaction is compatible and the plant susceptible. This cross talk between host and pathogen was assembled in the gene-for-gene model by Flor (1942), who extracted the concept from his work on the interactions between flax and flax rust. Since the early nineties numerous *R* genes from model plant or crop species have been identified and cloned (Young, 2000; Dangl and Jones, 2001) and, in parallel, many *Avr* genes mainly from fungi and bacteria (White et al., 2000; Luderer and Joosten, 2001; van't Slot and Knogge, 2002). The availability of both a cloned *R* gene and its corresponding cloned *Avr* gene offers exciting opportunities to elucidate the gene-for-gene interaction at the molecular and cellular level. In recent years the guard model has won ground particularly by studies on a few model pathosystems such as the interactions between Arabidopsis or tomato and the bacterial speck pathogen *Pseudomonas syringae,* and tomato and the leaf mold fungus *Cladosporium fulvum* (Innes, 2004; Rooney et al., 2005). In this model R proteins and pathogen effectors (i.e., AVR proteins) are part of a larger dynamic complex. The pathogen effectors target host cell proteins in order to suppress defense responses or elicit susceptible responses. R proteins evolved as a counter-defense and function to monitor the effector targets.

The subject of our studies is *Phytophthora infestans,* the notorious Irish potato famine pathogen and the causal agent of late blight (Govers and Latijnhouwers, 2004). *Phytophthora* species resemble fungi morphologically but in the tree of life they are classified as oomycetes, a unique group of eukaryotes that evolved independently from fungi. Oomycetes include significant pathogens of insects and animals and they are responsible for a wide variety of destructive plant diseases. All *Phytophthora* species (more than 65), and the majority of the *Pythium* species are plant pathogens, and also all downy mildew diseases and white rusts are caused by oomycetes (Agrios, 1997). Oomycetes not only look like fungi, they also behave like fungi and use the same weaponry to attack plants (Latijnhouwers et al., 2003). Similarly, the R proteins that plants use to defeat oomycetes have the same architecture as R proteins that stop fungal invasions (Ballvora et al., 2002; van der Vossen et al., 2003; Gao et al., 2005; Huang et al., 2005) and many oomycete-plant interactions follow the gene-for-gene model. Genetic analyses on host and pathogen have demonstrated that this model also suits the potato-*P. infestans* pathosystem (van der Lee et al., 2001).

Unlike R proteins, the pathogens' AVR proteins or effectors are highly divergent (Luderer and Joosten, 2001; van't Slot and Knogge, 2002). Many of the fungal *Avr* genes were cloned by reverse-genetics using purified elicitor preparations as starting material. For genetically more tractable fungi, like for example *Magnaporthe grisea,* positional cloning appeared to be a suitable approach, and for cloning bacterial *Avr* genes classical bacterial genetics such as genetic complementation proved to be very

efficient (van den Ackerveken and Bonas, 1997; Collmer, 1998). In the case of *Phytophthora* however, *Avr* gene cloning has lagged behind (Tyler, 2001; Tyler, 2002). Because of the (hemi-)biotrophic nature of many oomycete-plant interactions purifying elicitors is difficult and, in our hands attempts to identify race specific elicitors from *P. infestans* were unsuccessful (Alfonso and Govers, 1995). Therefore reverse genetics is not an option. Moreover low DNA transformation efficiencies and relatively large genome sizes hamper complementation or gene tagging approaches. A more suitable approach is positional cloning and recently three oomycete *Avr* genes have been identified starting off with this approach: *Avr1b*-1 from *Phytophthora sojae* (Shan et al., 2004), and *ATR13* and *ATR1*^{NDWsB} from the Arabidopsis downy mildew pathogen *Hyaloperonospora parasitica* (Allen et al., 2004; Rehmany et al., 2005). These two species are homothallic and the number of inbred progeny that was generated was sufficient to obtain recombinants in the *Avr* regions and to identify closely linked markers.

For cloning *Avr* genes in *P. infestans* we also adopted a positional cloning approach and generated high-density maps of chromosomal regions carrying *Avr* genes (van der Lee et al., 2001). In addition a BAC library of a strain carrying six dominant *Avr* genes and suitable for marker landing, is available (Whisson et al., 2001). However, *P. infestans* is heterothallic and the problem we face is the inability to generate large segregating mapping populations. Also the relatively large genome size (245 Mb) reduces the marker density and even with high-density linkage maps (van der Lee et al., 2004) we were not able to generate enough markers for efficient landing. To complement the positional cloning strategy we aimed at generating transcriptome markers. In this study we combined a cDNA-AFLP based strategy with Bulked Segregant Analysis (BSA) to identify *Avr*-associated transcripts. cDNA-AFLP is a relatively simple method to obtain a genome-wide display of differentially expressed genes and it has already been successfully used for gene discovery in *P. infestans* (Avrova et al., 2003; Dong et al., 2004). Many of the known *Avr* genes show a relatively high expression or a stage specific expression in pre-infection stages and therefore we used germinating cysts as starting material for RNA isolation. cDNA-AFLP patterns obtained from pools of strains with identical AVR phenotypes revealed a high number of putative *Avr*-associated transcript derived fragments (TDFs) for each of the four *Avr* loci that were targeted. Subsequently, segregation of the *Avr*-associated TDFs in an F1 mapping population was analyzed resulting in transcriptome markers for two *Avr* loci.

Materials and methods

P. infestans strains and mapping population

The *P. infestans* strains used in this study are two Dutch field isolates of opposite mating type (80029; A1 and 88133; A2) and 18 F1-progeny (designated as cross 71). The cross 71 mapping population was previously described and characterized by Drenth et al (1995) and van der Lee et al. (1997). The nomenclature of genes, gene clusters and phenotypes is according to van der Lee et al. (2001) with one exception; *Avr3* now has the suffix 'b' to indicate that this avirulence gene elicits resistance on plants

carrying resistance gene *R3b* and not *R3a* (Huang et al., 2004). Consequently, an avirulent and virulent phenotype on *R3b* plants is indicated by AVR3b and avr3b, respectively.

P. infestans culture conditions

P. infestans strains were routinely grown at 18 ℃ in the dark on rye agar medium supplemented with 2% sucrose (RSA) (Caten and Jinks, 1968). To obtain germinating cysts for RNA isolation, sporulating mycelium grown on RSA was flooded with ice-cold water and incubated at 4 ℃. At this temperature sporangia release the zoospores into the water. After 4 hours incubation the zoospore suspension was filtered through a 10-µm nylon mesh to remove sporangia and mycelial fragments. Cysts were obtained by vigorous shaking of the zoospore suspension for 2 min. To allow germination the cyst suspension was incubated at 18 ℃ for at least 2 hours. The germination rate and germ tube length were checked with regular time intervals. When more than half of the cysts were germinated and the length of their germ tubes was 4-6 times the diameter of the cysts the tissue was collected by centrifugation (5 min at 3000 g), frozen in liquid nitrogen and stored at -80 ℃.

cDNA-AFLP analysis

RNA isolation, cDNA synthesis and cDNA-AFLP analysis were performed as described previously for *P. infestans* by Dong et al. (2004). Total RNA from germinated cysts was isolated using Trizol (Gibco-BRL) according to the manufacturer's instructions and subsequently purified using phenol-chloroform extraction. Poly A^+ RNA was isolated from 100 µg total RNA with the QIAGEN Oligotex mRNA kit. cDNA was synthesized using oligodT (12-18) and Superscript II reverse transcriptase (Gibco-BRL). The primary template for cDNA-AFLP was prepared in a one-step restriction-ligation reaction in which adapters were ligated to *Apo*I/*Taq*I digested cDNA fragments. The quality of each primary template was checked by performing a PCR on the diluted primary template using primers matching the adapters and by analyzing the PCR products on agarose gel. Based on the intensity on gel the quantity was estimated. Pre-amplification was performed in 25 cycles using primers corresponding to the *Apo*I and *Taq*I adapters without extension (A and T primers as in Dong et al., 2004). The diluted pre-amplification products were used as template for the selective amplification with two selective base extensions at the 3'-end of the primers (A+2 and T+2 primers). The A+2 primers were either labeled by phosphorylating the 5'-end with [γ-^{32}P]ATP for detection of the cDNA-AFLP fragments by autoradiography, or with IRD700 or IRD800 for fluorescence detection using LI-COR Global IR^2 systems. For analysis of the cDNA-AFLP fragments by silver staining the primers were not labeled. Separation of the cDNA-AFLP fragments was performed on 4 to 6% denaturating polyacrylamide gels as described by van der Lee et al. (1997).

Bulked segregant analysis

Bulked segregant analysis (BSA) was performed essentially following the procedure described by Michelmore *et al.* (1991). Ten F1 progeny of the cross 71 mapping population were selected and divided over four pools consisting of 2 or 3 F1 progeny with identical or nearly identical avirulence phenotypes (Table 1). Each phenotype is represented by 4-6 F1 progeny divided over two pools. From the six avirulence genes that segregate in cross 71 *Avr3* (renamed *Avr3b*), *Avr10* and *Avr* 11 are closely linked (van der Lee et al., 2001) and in this study we consider *Avr3b-Avr10-Avr11* as one locus. Primary templates of the 2 or 3 F1 progeny that made up one pool were mixed in equal amounts (based on the quantity and quality check described above) and served as template for the pre-amplification. In the selective amplification all 256 *Apo*I+2 / *Taq*I+2 primer combinations were used. In Fig. 1B, 1C, 1D and 1E the expected patterns for each of the pools are shown.

Table 1 Composition **of BSA pools for selecting *Avr*-associated TDFs.**

		Phenotypes on differentials containing resistance gene*					
Pool	Strain	*R1*	*R3b*	*R10*	*R11*	*R4*	*R2*
1	re11-16	AVR	avr	avr	avr	AVR	AVR
	T15-1	AVR	avr	avr	avr	AVR	AVR
	T30-2	AVR	avr	avr	avr	AVR	AVR
2	D12-2	avr	avr	avr	avr	avr	AVR
	D12-23	avr	avr	avr	avr	avr	AVR
	T35-3	avr	avr	avr	avr	avr	avr
3	D12-17	AVR	AVR	AVR	AVR	avr	avr
	T15-9	AVR	AVR	AVR	AVR	avr	avr
4	T20-2	avr	AVR	AVR	AVR	AVR	AVR
	E12-3	avr	AVR	AVR	AVR	AVR	avr

*AVR and avr indicate avirulence and virulence phenotype, respectively.

Isolation, cloning and sequencing of TDFs

The cDNA-AFLP fragments (i.e. TDFs) of interest were excised from gels using a razor blade. The gel slices were rehydrated in 100 µl of water and incubated at 70 ℃ for 15 min. The eluted fragment was reamplified with the primers with the same two base pair extension as used in the cDNA-AFLP analysis. PCR products were purified using QIAquick PCR purification kit (Qiagen, Hilden) and cloned into pGEM-T Easy (Promega, Madison, WI, USA). Recombinant clones were sequenced by BaseClear (Leiden, The Netherlands) or Shanghai Biotech (Shanghai, China).

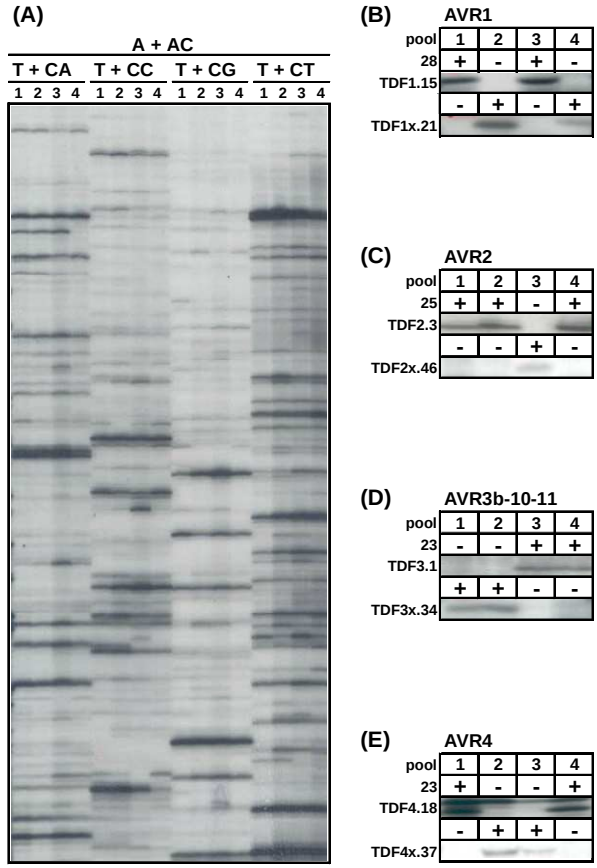


Fig. 1. cDNA-AFLP analysis. (A) Section of autoradiograph showing cDNA-AFLP fingerprints in four BSA pools generated with the indicated primer combinations. For the composition of pools 1, 2, 3, and 4 see Table 1. In panels (B), (C), (D) and (E) the second row shows the number of *Avr*-associated TDF candidates found in this study and the expected cDNA-AFLP patterns in the four pools for TDFs associated with AVR1, AVR2, AVR3b-AVR10-AVR11 and AVR4 phenotypes, respectively. The third row shows examples of candidate TDFs with the expected pattern. The fourth row shows the expected opposite pattern and the fifth row examples. In (B) TDF1.25 was obtained with primer combination A+TG/T+CG and TDF1x.21 with A+GA/T+GT. In (C) TDF2.3 with A+AG/T+TT and TDF2x.46 with A+TG/T+GG. In (D) TDF3.1 with A+AG/T+AC and TDF3x.34 with A+GT/T+CC. In (E) TDF4.18 with A+TG/T+GG and TDF4x.37 with A+GT/T+TA.

DNA sequence analysis and bioinformatics

Sequences were analysed in Vector NTI 8. For BLAST searches we used the NCBI BLAST program and the Standalone-BLAST Version 2.2.3 (Altschul et al., 1997). The *P. infestans* EST databases are accessible at http://www.pfgd.org and http://staff.vbi.vt.edu/estap (Kamoun et al., 1999; Randall et al., 2005). The genomic sequences and annotated protein sequences of *P. sojae* and *P. ramorum* were obtained from the website of the DOE Joint Genome Institute (http://www.jgi.doe.gov/genomes). TDF sequences were searched against GenBank and EST databases by BLASTX and BLASTN, respectively. A GenBank hit was considered to be a homologue if the BLASTX *E* value is less than 1e-3. A TDF was considered to be represented by an EST if the BLASTN identity is equal to or larger than 99%. RT-PCR primers were designed based on the cloned TDF sequence or the EST sequence if the TDF has a corresponding EST. Primer lengths were between 18 bp to 25 bp with melting temperatures higher than 55 ^{o}C in all cases. The primer sequences are available from the authors upon request.

RT-PCR analysis

To remove genomic DNA from RNA preparations, 10 µg total RNA was treated with 4 units RQ1 RNase-free DNase (Promega, Madison, WI) at 37 ℃ for 1 h. The removal of all DNA was verified in a PCR reaction under the same conditions as those used for the RT-PCR reaction, except that the cDNA

synthesis step was omitted. The first-strand cDNA was synthesized using oligo(dT16) and Superscript II reverse transcriptase for 30 min at 40 ℃ (Gibco-BR L). Sequence-specific primers were used in the subsequent PCR with cDNA as template with 30 cycles (30 s at 94 ℃, 30 s at 56-60 ℃ and 60 s at 72 ℃).

Nomenclature of TDFs

The cDNA-AFLP fragments and the clones containing the fragments are named TDF followed by a number that refers to the *Avr* gene for which, according the BSA pattern, the TDF was a candidate. This *Avr*-associated number is then followed by a period and a random clone number. For *Avr3b-Avr10-Avr11* the *Avr*-associated number is 3. In cases where an 'x' is added as suffix the TDF showed an opposite pattern in the BSA. Occasionally, an 's' is added at the very end to indicate that the TDF was selected in the BSA analysis on silver stained gels.

Results and discussion

BSA for selecting transcripts associated with avirulence

BSA was initially developed as a method for rapidly identifying polymorphic DNA markers linked to any specific gene or genomic region (Michelmore et al., 1991). Two bulked DNA samples are generated from a segregating population from a single cross. Each pool, or bulk, contains individuals that are identical for a particular trait or genomic region but arbitrary at all unlinked regions. The two bulks are therefore genetically dissimilar in the selected region but seemingly heterozygous at all other regions. Previously, the *P. infestans* cross 71 mapping population was succesfully used for BSA to identify AFLP markers linked to six *Avr* genes segregating in cross 71 (van der Lee et al., 2001). In the present study we used the same cross 71 mapping population, a similar pool design and the same pool sizes for a BSA approach aimed at selecting transcripts derived from *Avr* genes. Instead of DNA, cDNA of different individuals from the cross was pooled. The phenotypes of the strains that constitute the four BSA pools are listed in Table 1. Anticipating that *Avr* genes are expressed just prior to infection we used RNA isolated from germinating cysts as starting material.

It is logical to combine a BSA approach with an efficient genome-wide transcriptional profiling method. Recently, Dong et al. (2004) described an optimized cDNA-AFLP protocol for *P. infestans* that was based on *in silico* cDNA-AFLP fingerprinting of a large set of *P. infestans* ESTs. The primer combination *Apo*I / *Taq*I and selective amplification using primers with two base extensions resulted in clear transcription profiles that were easy to score. Fig. 1A shows a section of a typical autoradiograph with cDNA-AFLP patterns obtained from the four BSA pools with four primer combinations following the

protocol of Dong et al. (2004). All 256 *Apo*I+2 / *Taq*I+2 primer combinations were used to generate radioactive TDFs that were visualized by autoradiography. A subset of the primer combinations was used to generate unlabeled TDFs and those were visualized by silver staining. Over 30,000 TDFs ranging in size from 40-600 bp were analyzed. Overall the patterns obtained with autoradiography and silver staining were comparable but remarkably some TDFs that were detected with the radioactive primer were not visible as a band on silver stained gels and, *vice versa*, some clear bands on silver stained gels were absent on autoradiographs.

TDFs present in avirulent but not in virulent strains are expected to show up in two pools (or three in the case of *Avr2*) but not in the others. In fact, the pool design included internal controls, for example, an *Avr1* specific transcript should only be present in pool 1 and pool 3 whereas an *Avr4* specific transcript should be present in pool 1 and pool 4 but not in pool 2 nor pool 3. For each of the three *Avr* genes and the *Avr3b-Avr10-Avr11* locus 23 or more TDFs that behaved according to the predicted patterns were detected. In total 99 such *Avr*-associated TDFs were found, some of which were only visible by silver staining. In all cases TDFs with opposite pattern were also found. Examples are shown in Fig. 1B, 1C, 1D and 1E. Although the observed BSA patterns suggest that the TDFs represent genes that are specifically expressed in either avirulent or virulent strains one should bear in mind that also polymorphisms in the *Apo*I or *Taq*I recognition site or in the two base pair extensions may result in differential cDNA-AFLP patterns.

Segregation of Avr-associated TDFs in cross 71

From previous studies in which the segregation of the avirulence phenotypes in cross 71 was analyzed, it was evident that the AVR1, AVR2 and AVR4 phenotypes behave as single dominant traits (Alfonso and Govers, 1995; van der Lee et al., 2001). AVR3b, AVR10 and AVR11 are also dominant but the genes are closely linked (van der Lee et al., 2001). The *Avr3b-Avr10-Avr11* locus might harbour three independent genes but it can not be excluded that the locus contains a single gene that either controls other loci conferring avirulence on *R3b*, *R10* and *R11* plants or that interacts with an uncharacterized R gene shared by *R3b*, *R10* and *R11* plants. Many of the known avirulence factors from plant pathogens are effector proteins that are present in avirulent strains but absent, unstable or mutated in virulent strains (Westerink et al., 2004). Hence, the *Avr*-associated TDFs that were identified in the BSA may all represent candidate *Avr* genes. However, we hypothesize that from each set only TDFs derived from one transcript (or possibly three in the case of *Avr3b-Avr10-Avr11*) can represent the real *Avr* gene. To make a further selection, we performed fluorescent and silver stained cDNA-AFLP analyses on the two parental lines of cross 71 and 18 F1 progeny, and screened for presence or absence of TDFs. For 25 *Avr*-associated TDFs there was segregation in the F1 progeny, 8 of which were associated with *Avr1*, 8 with *Avr2*, 4 with *Avr3b-Avr10-Avr11* and 5 with *Avr4*. Representative patterns are shown in Fig. 2. Two of the 25 were not polymorphic in the parental lines and are thus unlikely candidates to represent an *Avr*

gene. However, for four of the 25 *Avr*-associated TDFs the presence/absence pattern matched exactly with the avirulence phenotypes of the two parental lines and the 18 F1 progeny making them ideal transcriptome markers representing an *Avr* gene. Two are associated with the *Avr3b-Avr10-Avr11* locus (TDF3.1 and TDF3.4), and two with *Avr4* (TDF4.1s and TDF4.2s) (Fig. 2).

None of the *Avr1* and *Avr2* candidates cosegregated with avirulence and it is therefore unlikely that these TDFs are derived from *Avr1* or *Avr2*. Nevertheless, based on the segregation patterns of the remaining 19 TDFs we anticipate that some of them are linked to the *Avr* locus (data not shown). If the polymorphism represents a DNA polymorphism they could be used as markers for fine mapping the *Avr* regions. Alternatively, they could be used for the construction of a transcriptome map (Brugmans et al., 2002).

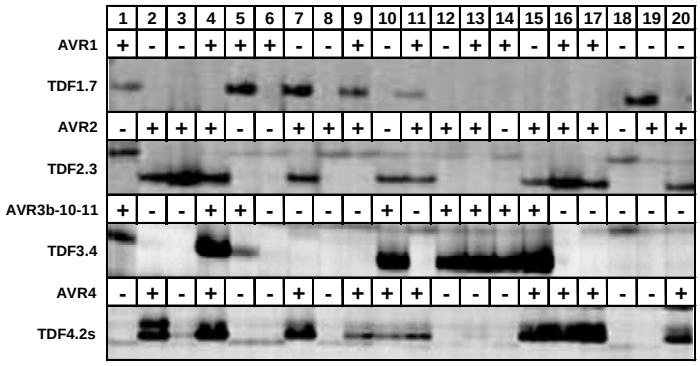


Fig. 2. cDNA-AFLP patterns showing the segregation of four *Avr*-associated TDFs in 18 F1 progeny of cross 71 (lanes 3-20). Lane 1 and 2 show the cDNA-AFLP patterns obtained from the two parental isolates 80029 and 88133, respectively. The avirulence and virulence phenotypes of parents and progeny are indicated by + and -, respectively. TDF1.7 was obtained with primer combination A+AT/T+GA, TDF2.3 with A+AG/T+TT, TDF3.4 with A+AA/T+AG and TDF4.2s with A+TC/T+TC.

TDF cloning and sequencing

To enable further analysis of the TDFs we cloned the majority of the 99 *Avr*-associated TDFs and a number of TDFs with opposite pattern. TDFs were excised from gel, re-amplified, cloned and sequenced. Based on the size of the clone insert and the presence or absence of the expected two-base primer extension in the sequence, it was concluded that 142 TDFs were successfully cloned. Overall, the success rate of cloning was over 94%. GenBank accession numbers of the cloned TDFs and AFLP codes showing the primer extensions and fragment size, are listed in Table 2. The TDF nucleotide sequences were used to design primers for RT-PCR analysis (see *3.4.*) and for similarity searches in various databases (see *3.5.*).

RT-PCR expression analysis of Avr-associated TDFs

In parallel to the segregation analysis of the TDFs in cross 71, expression of cloned TDFs was analyzed by RT-PCR. For 42 TDFs suitable primers were designed and for 38, RT-PCR products were obtained.

Table 2 GenBank accession numbers and AFLP codes showing the primer extensions and fragment size of 142 cloned TDFs.

TDF	AFLP code	Accession	TDF	AFLP code	Accession	TDF	AFLP code	Accession
1.1	A+AA/T+AAs169	DW010060	2.21	A+TG/T+GAs399	DW010177	3x.29	A+GG/T+GGs193	DW010132
1.2	A+AA/T+CTs232	DW010070	2.22	A+TT/T+CGs314	DW010178	3x.30	A+GG/T+GGs120	DW010133
1.3	A+AA/T+GGs345	DW010173	2.23	A+TT/T+CGs312	DW010179	3x.33	A+GT/T+AGs178	DW010134
1.5	A+AT/T+ACs234	DW010078	2.1s	A+CC/T+TAs210	DW010093	3x.34	A+GT/T+CCs137	DW010135
1.6	A+AT/T+CTs71	DW010079	2.2s	A+GG/T+GCs220	DW010095	3x.35	A+GT/T+GTs381	DW010188
1.7	A+AT/T+GAs233	DW010080	2x.8	A+CC/T+CCs93	DW010103	3x.42	A+TA/T+GTs303	DW010189
1.8a	A+AT/T+TTs72	DW010081	2x.39	A+TA/T+ACs346	DW010181	3x.43	A+TA/T+TTs77	DW010137
1.8b	A+AT/T+TTs72	DW010081	2x.46	A+TG/T+GGs254	DW010102	3x.45	A+TA/T+GTs112	DW010138
1.9	A+CA/T+AGs362	DW010174	3.1	A+AG/T+ACs153	DW010104	3x.47	A+TG/T+TGs257	DW010139
1.10	A+CA/T+GGs360	DW010172	3.2	A+AG/T+ACs104	DW010114	3x.48	A+TG/T+AGs227	DW010140
1.11	A+CC/T+CGs152	DW010061	3.3	A+AG/T+TTs242	DW010116	3x.51	A+TT/T+GTs291	DW010190
1.12	A+CC/T+TAs65	DW010062	3.4	A+AA/T+AGs156	DW010117	3x.52	A+TT/T+CCs218	DW010142
1.13	A+TC/T+ACs219	DW010063	3.6	A+AT/T+GAs160	DW010118	4.1	A+AA/T+AGs233	DW010146
1.14	A+TC/T+ACs212	DW010064	3.7	A+AT/T+TCs315	DW010184	4.2	A+AA/T+GGs105	DW010155
1.15	A+TC/T+GTs179	DW010065	3.8	A+AT/T+TCs252	DW010119	4.3	A+AC/T+AGs139	DW010158
1.16	A+CT/T+ATs222	DW010066	3.9	A+AT/T+TCs108	DW010120	4.4	A+AC/T+CGs183	DW010159
1.17	A+CT/T+ATs222	DW010066	3.10	A+GC/T+CAs147	DW010105	4.5	A+AC/T+GTs69	DW010160
1.18	A+CT/T+CTs240	DW010067	3.11	A+GC/T+GAs250	DW010106	4.7	A+AT/T+AAs232	DW010161
1.19	A+GC/T+AGs225	DW010068	3.12	A+GC/T+GTs265	DW010107	4.8	A+TC/T+AAs123	DW010162
1.20	A+GG/T+TAs124	DW010071	3.13	A+GC/T+TAs156	DW010108	4.9	A+CT/T+CGs96	DW010163
1.21	A+GG/T+TTs161	DW010072	3.14	A+GA/T+CAs281	DW010109	4.10	A+GC/T+CCs114	DW010147
1.22	A+GT/T+CAs268	DW010073	3.15	A+GG/T+CCs164	DW010110	4.11	A+GA/T+AGs503	DW010191
1.23	A+TA/T+TGs132	DW010074	3.16	A+GG/T+CCs84	DW010111	4.12	A+GA/T+CGs238	DW010148
1.24	A+TA/T+GTs110	DW010075	3.17	A+GG/T+CCs84	DW010111	4.13	A+GG/T+AGs85	DW010149
1.25	A+TG/T+CGs109	DW010076	3.19	A+GT/T+TAs168	DW010112	4.14	A+GT/T+AGs296	DW010192
1.1s	A+TA/T+GAs150	DW010069	3.20	A+TT/T+AAs347	DW010182	4.15	A+TA/T+GTs354	DW010193
1.2s	A+TC/T+CAs200	DW010077	3.21	A+TT/T+GCs293	DW010183	4.16	A+TG/T+AAs240	DW010150
1x.15	A+GC/T+CTs169	DW010082	3.22	A+TT/T+TTs110	DW010115	4.17	A+TG/T+AGs117	DW010151
1x.21	A+GA/T+GTs137	DW010083	3.1s	A+TA/T+AGs90	DW010113	4.18	A+TG/T+GGs116	DW010152
2.1	A+AG/T+GCs155	DW010084	3x.2	A+AG/T+ATs93	DW010126	4.19	A+TT/T+CAs95	DW010153
2.3	A+AG/T+TTs137	DW010096	3x.4	A+AC/T+ACs257	DW010136	4.20	A+TT/T+GGs117	DW010156
2.4	A+AA/T+AAs251	DW010097	3x.5	A+AT/T+GAs158	DW010141	4.1s	A+TC/T+GAs125	DW010154
2.5	A+AC/T+CCs117	DW010098	3x.6	A+CA/T+ACs163	DW010143	4.2s	A+TC/T+TCs144	DW010157
2.6	A+AC/T+GTs152	DW010099	3x.7	A+CA/T+CCs130	DW010144	4.3s	A+GT/T+TTs180	DW010198
2.7	A+AC/T+TGs183	DW010100	3x.9	A+TC/T+GGs164	DW010145	4x.3	A+AC/T+TTs74	DW010167
2.8	A+AT/T+CTs319	DW010180	3x.10	A+CG/T+AAs169	DW010121	4x.10	A+TC/T+TCs133	DW010164
2.9	A+CA/T+CCs117	DW010101	3x.11	A+CG/T+GAs215	DW010122	4x.12	A+CG/T+TAs196	DW010165
2.10	A+CC/T+GCs101	DW010085	3x.13	A+GC/T+ATs172	DW010123	4x.18	A+GC/T+TTs483	DW010194
2.11	A+TC/T+GCs97	DW010086	3x.14	A+GC/T+CCs333	DW010185	4x.20	A+GA/T+GCs187	DW010166
2.12	A+TC/T+GGs72	DW010087	3x.16	A+GC/T+GCs216	DW010124	4x.31	A+GT/T+ACs364	DW010195
2.13	A+TC/T+ACs97	DW010088	3x.17	A+GC/T+GTs270	DW010125	4x.32	A+GT/T+ACs375	DW010196
2.14	A+CT/T+ATs144	DW010089	3x.19	A+GA/T+CAs400	DW010186	4x.37	A+GT/T+TAs361	DW010197
2.15	A+CT/T+CCs113	DW010090	3x.22	A+GA/T+TCs215	DW010127	4x.38	A+GT/T+TGs216	DW010168
2.16	A+CG/T+GAs412	DW010175	3x.23	A+GG/T+ATs151	DW010128	4x.40	A+TA/T+CTs286	DW010169
2.17	A+CG/T+GGs401	DW010176	3x.24	A+GG/T+CAs167	DW010129	4x.49	A+TT/T+TAs258	DW010170
2.18	A+GA/T+GAs238	DW010091	3x.25	A+GG/T+GGs352	DW010187	4x.50	A+TT/T+TGs206	DW010171
2.19	A+GT/T+ATs274	DW010092	3x.27	A+GG/T+GGs236	DW010130			
2.20	A+TA/T+CGs168	DW010094	3x.28	A+GG/T+CAs218	DW010131			

The primer design was based on the TDF sequence itself or on the sequence of a matching *P. infestans* EST with a sequence similarity higher than 99%. The majority of the RT-PCR products could be visualized on agarose gels but for several the small size of the RT-PCR product or the occurrence of multiple bands with size differences of only a few base pairs required an electrophoresis system with a higher resolution (i.e. polyacrylamide gels). Table 3 lists the 38 TDFs including the amplicon sizes. The RT-PCR analysis included the two parental strains of cross 71 and 9 F1 progeny. Of the 38 TDFs four showed an RT-PCR expression pattern that perfectly matched the avirulence phenotypes in parents and F1 progeny, and these are the same four TDFs that matched in the segregation analysis based on cDNA-AFLP patterns: TDF3.1, TDF3.4, TDF4.1s and TDF4.2s (Table 3). In the avirulent parent and progeny the RT-PCR product was present and in the virulent parent and progeny it was absent. Since both RT-PCR and cDNA-AFLP give this black and white pattern it is very likely that the difference is caused by presence versus absence of mRNA and not by polymorphisms in the sequences. Hence, the genes corresponding to these TDFs seem to be regulated at the transcriptional level.

Several of the other 34 TDFs showed differential RT-PCR patterns but there was no association with the avirulence phenotypes. A substantial number, however, showed no differential expression at all. Again none of the *Avr*-associated TDFs tested by RT-PCR appeared to be a candidate for *Avr1* or *Avr2.*

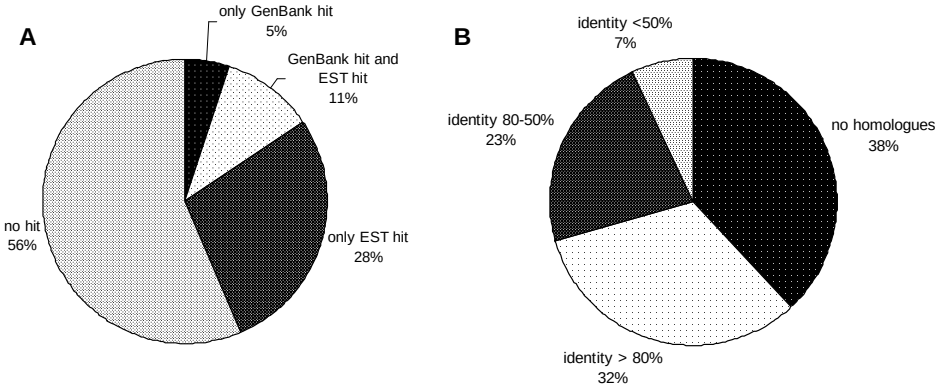


Fig. 3. Percentages of 142 cloned *P. infestans* TDFs with (A) sequence homology in GenBank and *P. infestans* EST databases and (B) homologues in *P. sojae*. Homologues were counted if the BLASTX *E* value was less than 1e-3. In the GenBank sequences the *P. infestans* ESTs deposited in GenBank were not included.

Sequence similarity of TDFs with *Phytophthora* sequences and known sequences

Sequence similarity to known sequences may help in assigning a function to the genes from which the TDFs are derived. All TDF sequences were compared by BLAST algorithm to the NCBI GenBank and *P. infestans* EST databases with an *E*-value cutoff of 1e-03. Of the 142 TDFs 56% had no match at all. A

small percentage (16 %) had a match in GenBank and 39 % had high sequence similarity to *P. infestans* ESTs (Fig. 3A). The *P. infestans* EST database comprises over 75.000 ESTs obtained from cDNA libraries representing a broad range of growth conditions, stress responses, and developmental stages (Randall et al., 2005). It is likely that more TDFs have matching cDNA clones in the EST libraries but because many of the ESTs are only partially sequenced, matching cDNA clones may not always be recognizable. On the other hand cDNA-AFLP is a very sensitive method and is able to detect very low abundance mRNA that may not be present in the EST database.

The 142 TDFs were also BLASTed against the fully sequenced genomes of *P. sojae* and *P. ramorum*. Over one third had no homologues in *P. sojae* and one third had homologues with a similarity higher than 80% (Fig. 3B). The homologues in the *P. sojae* proteome that were assigned to represent the TDFs were subsequently BLASTed against the SwissProt database. A wide range of hits was found, such as proteins that function as phosphatase or kinase but also an ABC transporter, a water channel protein, and molecular motor proteins. As expected, there are also TDFs that do have a match in the *P. sojae* proteome but no hit in SwissProt. Table 3 shows the results for 38 of the 142 cloned TDFs.

Table 3 RT-PCR analysis and sequence similarity of 38 *P. infestans* TDFs

TDF	*P. infestans* EST hit [a]	amplicon size (bp) [b]	RT-PCR [c]	*P. sojae* homologue [d]	*P. ramorum* homologue [d]	SwissProt hit of *P. sojae* homologue	BLAST identity (%)	*E*-value
1.1		117	-	pro135357	pro75828	RB38_HUMAN (P57729) Ras-related protein Rab-38	40	3.00E-33
1.2		190	-	pro135623	pro71960	RDPO_SCHPO (Q05654) Retrotransposable element Tf2 155 kDa protein	29	1.00E-99
1.3	CON_001_13933	304*	-	pro135300	pro75790	PTPJ_HUMAN (Q12913) Protein-tyrosine phosphatase	36	1.00E-36
1.5		202	-					
1.6		38	-					
1.7	CON_003_04202	390*	-	pro140341	pro83108			
1.14		90	-	pro143752	pro87069	RDPO_SCHPO (Q05654) Retrotransposable element Tf2 155 kDa protein	27	2.00E-66
1.22	CON_016_07340	400*	-	pro125097	pro83808	SYM_ARATH (Q9SVN5) Probable methionyl-tRNA synthetase	37	9.00E-21
2.3	CON_001_16821	191*	-	pro129917	pro73127	VTL2_MOUSE (O89116) Vesicle transport v-SNARE protein Vti1-like 2	29	4.00E-21
2.7	CON_001_30638	472*	-	pro109725	pro87143	AQP3_HUMAN (Q92482) Aquaporin 3	43	1.00E-34
2.11		59	-					
2.13		59	-					
2.15	CON_001_14541	380*	-	pro131502	pro84862	GTT2_HUMAN (P30712) Glutathione S-transferase theta 2	33	1.00E-20
3.1		54	yes					
3.3		148	-	pro108156	pro39196	ENGA_RICPR (Q9ZCP6) Probable GTP-binding protein engA	24	1.00E-03
3.4		115	yes	pro133266	pro80794	MYH3_CHICK (P02565) Myosin heavy chain, fast	22	2.00E-07

						skeletal muscle		
3.7		282	-	pro131930	pro74150	TRHY_SHEEP (P22793) Trichohyalin	18	1.00E-11
3.8		172	-					
3.9		62	-					
3.16		52	-					
3.19	CON_002_01106	377*	-	pro133266	pro80794	MYH3_CHICK (P02565) Myosin heavy chain, fast skeletal muscle	22	2.00E-07
3.20	CON_010_06936	490*	-	pro137091	pro85962	NSB1_HUMAN (P82970) Nucleosomal binding protein 1	23	3.00E-04
3x.7		94	-		pro80914			
3x.11		182	-	pro143645	pro80057	MYSJ_DICDI (P54697) Myosin IJ heavy chain	33	1.00E-114
3x.22		182	-	pro131604	pro80644	DSPP_HUMAN (Q9NZW4) Dentin sialophosphoprotein precursor	17	7.00E-09
3x.33	CON_001_10962	236*	-	pro131005	pro85669			
4.1		191	-	pro131094	pro86402	BFR1_SCHPO (P41820) Brefeldin A resistance protein	20	6.00E-17
4.2		67	-					
4.10		80	-	pro138207	pro81288			
4.13		42	-					
4.14	CON_001_33999	271*	-	pro140951	pro72858	CATL_DROME (Q95029) Cathepsin L precursor	40	5.00E-58
4.18		71	-					
4.19	CON_001_29569	452*	-	pro131364	pro73340			
4.20		75	-	pro134550	pro74902	CSK_CHICK (P41239) Tyrosine-protein kinase	28	4.00E-18
4.1s	CON_001_33634	186*	yes	pro109418	pro83335			
4.2s	CON_001_33634	186*	yes	pro109418	pro83335			
4x.3		41	-					
4x.50	CON_014_07231	473*	-	pro138318	pro82098			

[a] *P. infestans* EST hits with *E* value < 1e-50 and identity > 99% are listed..
[b] RT-PCR amplicon size was calculated based on TDF or EST sequence information; * indicates that the primers were designed on the EST sequence.
[c] 'yes' indicates that the RT-PCR polymorphism correlates with the AVR phenotypes of the parents and 9 F1 progeny; - indicates no polymorphism or no correlation with the AVR phenotypes.
[d] *P. sojae* and *P. ramorum* whole genome sequences and the gene annotation at the JGI website (http://www.jgi.doe.gov/genomes) were used for analysis. Genes with BLAST *E* value less than 1e-3 were considered homologues.

TDFs represented by *P. infestans* ESTs and TDFs with *P. sojae* homologues with a variety of putative functions were taken for further data mining and bioinformatics analysis such as *in silico* expression analysis, signal peptide prediction and gene copy number of the homologues in *P. sojae* and *P. ramorum* (Table 4). In the *P. infestans* EST database (Randall et al., 2005) we analyzed the distribution of ESTs representing the TDFs over the various libraries and, based on these numbers, we predicted stage specific expression patterns and expression levels. For example, for TDF3x.34 many ESTs are found in the germinating cysts library and zoospore library but none in a mycelium library. This indicates that the TDF3x.34 gene is specifically expressed at a relatively high level in zoospores and germinating cysts. In contrast, for TDF2.7 there is only one EST in the database, which indicates that this aquaporin-like gene is transcribed at a low level in wall-less zoospores. Of the 55 TDFs for which we found a matching *P. infestans* EST, only 16 have ESTs in germinating cyst stages. Our screening strategy did

not include a stage specific selection but since we used germinating cysts as starting material one would expect to find matching ESTs in that stage. This is true for only one third of the TDFs confirming that cDNA-AFLP is a very sensitive method that can reveal very low abundance transcripts.

Many of the fungal and oomycete elicitors identified to date are small secreted proteins with an even number of cysteine residues that usually form disulfide bridges to stabilize the protein (van't Slot and Knogge, 2002). Another feature typical for oomycete elicitors is the RXLR motif, a motif shared by four oomycete avirulence factors that lack cysteines (Allen et al., 2004; Shan et al., 2004; Armstrong et al., 2005; Rehmany et al., 2005). For the *Avr*-associated TDFs the presence of a signal peptide combined with a particular cysteine signature or an RXLR motif can be indicative for elicitor function. Two proteins representing TDF3x.34 and TDF4x.49 are predicted to be secreted by the program SignalPv2.0 (Nielsen et al., 1997; Nielsen and Krogh, 1998) and are also rich in cysteine residues. The protein represented by TDF4.1s has a homologue in *P. sojae* that is a secreted protein so we anticipate that the full length TDF4.1s protein also has a signal peptide (Table 4). These three proteins have no homology with any known protein but, interestingly, the *P. sojae* homologue of the TDF4.1s protein has an RXLR motif making TDF4.1s a promising candidate for an *Avr* gene.

Table 4 Analysis of *Avr*-associated TDFs using data mining and bioinformatics.

TDF	*P. infestans* EST hit [a]	protein size [b]	signal peptide [c]	Cys spacing pattern [d]	transcripts in *P. infestans* EST database [e]	*P. sojae* homologue [f]	SwissProt BLAST hit of *P. sojae* homologue	putative function	phylogenetic distribution [g]	genes in *P. sojae* [h]	genes in *P. ramorum* [h]
1.7	CON_003_04202	nd	-		ZO(1) SP(1) MY(1)	pro140341		unknown	only in *Phytophthora*	1	1
2.3	CON_001_16821	nd	-		MY(1)	pro129917	VTL2_MOUSE (O89116)	vesicle transporter	other species	1	1
2.7	CON_001_30638	nd	-		ZO(1)	pro109725	AQP3_HUMAN (Q92482	water channel	other species	>10	>10
3.4		nd	-		-	pro133266	MYH3_CHICK (P02565)	cyto-skeleton related	other species	1	1
3x.34	CON_020_07430	159	SP	C-n20-C-n9-C-n8-C	ZO(13) CY(7) SP(1)	pro138143		unknown secreted protein	only in *Phytophthora*	8	3
4.1s	CON_001_33634	>150		none	MY(1)	pro109418		unknown secreted protein	only in *Phytophthora*	2	1
4.1		nd	-		-	pro131094	BFR1_SCHPO (P41820)	ABC transporter	other species	>10	>10
4.20		nd	-		-	pro134550	CSK_CHICK (P41239)	kinase	other species	1	1
4x.49	CON_011_07076	300	SP	C-n28-C-n3-C-n10-C-n17-C-n23-C-n103-C	MY(8)	pro144423		unknown secreted protein	only in *Phytophthora*	>10	>10

[a] *P. infestans* EST hits with *E* value < 1e-50 and identity > 99% are listed.
[b] nd indicates that the protein sequence is incomplete.
[c] SP indicates that a signal peptide is predicted at the N-terminus by the program SignalP (Nielsen et al., 1997; Nielsen and Krogh, 1998).
[d] The proteins with signal peptide were used for cysteine spacing analysis.
[e] The tissue types from which the EST libraries are derived are zoospores (ZO), germinated cysts (CY), sporangia (SP) and mycelia (MY). The numbers in brackets indicate the number of ESTs present in the various libraries (Randall et al., 2005).
[f] *P. sojae* whole genome sequences and the gene annotation at the JGI website (http://www.jgi.doe.gov/genomes) were used for analysis. Genes with BLAST *E* value less than 1e-3 were considered homologues.
[g] Homologues in species other than *Phytophthora* were considered as homologues when the BLAST *E* value was less than 1e-3 and the similarity >30%.
[h] *P. sojae* and *P. ramorum* whole genome sequences and gene annotation at the JGI website (http://www.jgi.doe.gov/genomes) were used for analysis. Genes with BLAST similarity higher than 50% were considered to be members of the same gene family. Numbers indicate the size of the family.

With the exception of one *Phytophthora* elicitor, i.e. NIP1 (Fellbrich et al., 2002; Qutob et al., 2002), all oomycete elicitors and avirulence factors identified to date are unique for oomycetes: there are no homologous in organisms other than oomycetes. This is true for elicitins (Jiang et al., 2006) and the glycoprotein elicitor (gpe) containing pep13 (Sacks et al., 1995; Brunner et al., 2002), two protein families which are ubiquitous in the *Phytophthora* genus and have elicitor activity on a large variety of plant species. This is also true for the four ecotype- or cultivar-specific oomycete avirulence factors with the RXLR motif (Allen et al., 2004; Shan et al., 2004; Armstrong et al., 2005; Rehmany et al., 2005). In contrast to elicitin genes and *gpe* genes, however, none of these four avirulence genes belongs to a conserved gene family. Apart from the conserved RXLR motif they all show high sequence divergence with their homologues in other species and this may be a hallmark for host- or cultivar-specific avirulence genes. To evaluate the likelihood that the TDFs are derived from *Avr* genes we analyzed the phylogenetic distribution and we investigated whether the cloned TDFs belong to a gene family. Of the 88 TDFs that have homologues in *P. sojae* and *P. ramorum,* 25 seem to be unique for *Phytophthora.* Noticeably, all three secreted proteins listed in Table 4 only occur in *Phytophthora.* The homologues of TDF 4x.49 form a large gene family with over 10 members in both, *P. sojae* and *P. ramorum*, and those of TDF3x.34 appear to form a larger family in *P. sojae* than in *P. ramorum*. TDF4.1s has only two weak homologues (BLAST identity < 40%) in *P. sojae* and one weak homologue in *P. ramorum*, which suggests that this gene is of high sequence divergence among *Phytophthora* species.

Conclusions

In this study we demonstrate that combining a bulked segregant analysis strategy with a highly efficient transcriptional profiling method can be very effective in selecting *Avr*-associated transcripts. We focused on four *Avr* genes and for two of these we found TDFs that fulfill all criteria that make the TDF a likely *Avr* candidate. First of all, the TDFs occurred in germinating cysts, a preinfection stage in which an *Avr* gene is most likely to be expressed. Secondly, the TDFs were present in pools consisting of strains having an AVR phenotype but were absent in pools consisting of virulent strains (avr phenotype). Thirdly, segregation of the TDFs in F1 progeny correlated entirely with segregation of the AVR/avr phenotypes and, fourthly, RT-PCR confirmed the *Avr* associated segregation in the F1 progeny.

The two TDFs that were assigned as candidates for *Avr4,* TDF4.1s and TDF4.2s, appear to match to the same *P. infestans* EST contig and the deduced protein is an unknown protein. Since the homologues in *P. sojae* and *P. ramorum* are very divergent the protein seems to be unique for *P. infestans.* The *P. sojae* homologue though, has all the hallmarks of the family of RXLR proteins: a signal peptide, an RXRL motif and high sequence divergence with the other family members. All four oomycete *Avr* genes identified so far, *P. sojae Avr1b*-1 (Shan et al., 2004), *P. infestans Avr3a* (Armstrong et al., 2005) and the two *Hyaloperonospora parasitica* ecotype-specific *Avr* genes, *ATR13* and *ATR1* (Allen et al., 2004;

Rehmany et al., 2005), belong to this RXLR super family and sequencing of the full-length gene represented by TDF4.1s and TDF4.2s showed that this gene is also an RXLR family member (P. van Poppel, J.G and F.G, unpublished). Hence TDF4.1/TDF4.2 gene is a likely candidate for *Avr4.* Functional characterization is in progress.

The two TDFs that associate with the *Avr3b-Avr10-Avr11* locus, TDF3.1 and TDF3.4 are more mysterious. They fulfill all selection criteria but there are no matching *P. infestans* ESTs, and only TDF3.4 has an obvious homologue in the *P. sojae* proteome. These TDFs have recently been used as probes and markers to zoom in on the *Avr3b-Avr10-Avr11* locus and physical mapping showed that *Avr3b-Avr10-Avr11* linked AFLP and the TDFs are located on the same BAC contig (R.H.Y.J., Rob Weide and F.G., unpublished).

For *Avr1* and *Avr2* no candidates were recovered. Previously we used the same mapping population and a similar pooling strategy to identify AFLP markers (van der Lee et al., 2001). Also in that study the selection for *Avr4* and *Avr3b-Avr10-Avr11* linked markers was much more successful. For *Avr2* this could be explained by the fact that it was not included in the BSA, only random markers were selected. In the present study the pooling for *Avr2* was not optimal which may have caused a lower efficiency. For *Avr1,* however, it is unclear why the screening was unsuccessful. In both studies the BSA screening resulted in the highest number of candidates for *Avr1* but just one AFLP marker (van der Lee et al., 2001) and none of the TDFs passed the next, more stringent selection steps.

Previous studies in *P. infestans* have demonstrated that cDNA-AFLP is a powerful technique that complements other expression profiling approaches such as EST sequencing (Avrova et al., 2003; Dong et al., 2004). Here we showed that cDNA-AFLP can be combined with BSA to find transcripts associated with particular phenotypes. Since the *Avr*-linked genetic markers and the *Avr*-associated TDFs were generated from the same mapping population we can now integrate the various gene discovery approaches to identify *P. infestans Avr* genes.

Acknowledgements

We are grateful to Jacqueline Gerdsen for help with cloning, Drs. Qin Ling and Xiaowu Wang for useful suggestions and discussions, Prof. Dongyu Qu for support and encouragement, and Rob Weide for critically reading the manuscript. This work was financially supported by the Wageningen University Interdisciplinary Research and Education Fund (INREF) in the framework of the joint WU-CAAS PhD training programme and by NWO-Aspasia grant 015.000.057. The authors acknowledge Syngenta for access to the Syngenta *Phytophthora* Consortium EST Database prior to public release and the DOE Joint Genome Institute for making *P. sojae* and *P. ramorum* sequence data publicly available.

References

Agrios, G.N., 1997. Plant Pathology, fourth edition. Academic Press, USA

Alfonso, C., Govers, F., 1995. A search for determinants of race-specificity in the *Phytophthora infestans*-potato pathosystem. In: Dowley, L.J., Bannon, E., Cooke, L., Keane, T., O'Sullivan, E. (Eds.), *Phytophthora infestans*. Boole Press Ltd., Dublin, pp, 107-115.

Allen, R.L., Bittner-Eddy, P.D., Grenville-Briggs, L.J., Meitz, J.C., Rehmany, A.P., Rose, L.E., Beynon, J.L., 2004. Host-parasite coevolutionary conflict between Arabidopsis and downy mildew. Science 306, 1957-1960.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389-3402.

Armstrong, M.R., Whisson, S.C., Pritchard, L., Bos, J.I., Venter, E., Avrova, A.O., Rehmany, A.P., Bohme, U., Brooks, K., Cherevach, I., Hamlin, N., White, B., Fraser, A., Lord, A., Quail, M.A., Churcher, C., Hall, N., Berriman, M., Huang, S., Kamoun, S., Beynon, J.L., Birch, P.R., 2005. An ancestral oomycete locus contains late blight avirulence gene *Avr3a*, encoding a protein that is recognized in the host cytoplasm. Proc. Natl. Acad. Sci. USA 102, 7766-7771.

Avrova, A.O., Venter, E., Birch, P.R.J., Whisson, S.C., 2003. Profiling and quantifying differential gene transcription in *Phytophthora infestans* prior to and during the early stages of potato infection. Fungal Genet. Biol. 40, 4-14.

Bachem, C.W., van der Hoeven, R.S., de Bruijn, S.M., Vreugdenhil, D., Zabeau, M., Visser, R.G., 1996. Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. Plant J. 9, 745-753.

Ballvora, A., Ercolano, M.R., Weiss, J., Meksem, K., Bormann, C.A., Oberhagemann, P., Salamini, F., Gebhardt, C., 2002. The *R1* gene for potato resistance to late blight (*Phytophthora infestans*) belongs to the leucine zipper/NBS/LRR class of plant resistance genes. Plant J. 30, 361-371.

Brugmans, B., Fernandez del Carmen, A., Bachem, C.W., van Os, H., van Eck, H.J., Visser, R.G., 2002. A novel method for the construction of genome wide transcriptome maps. Plant J. 31, 211-222.

Brunner, F., Rosahl, S., Lee, J., Rudd, J.J., Geiler, C., Kauppinen, S., Rasmussen, G., Scheel, D., Nurnberger, T., 2002. Pep-13, a plant defense-inducing pathogen-associated pattern from *Phytophthora* transglutaminases. EMBO J. 21, 6681-6688.

Caten, C.E., Jinks, J.L., 1968. Spontaneous variability of single isolates of *Phytophthora infestans*. Can. J. Bot. 46, 329-348.

Collmer, A., 1998. Determinants of pathogenicity and avirulence in plant pathogenic bacteria. Curr. Opin. Plant Biol. 1, 329-335.

Dangl, J.L., Jones, J.D., 2001. Plant pathogens and integrated defence responses to infection. Nature 411, 826-833.

Dong, W.B., Latijnhouwers, M., Jiang, R.H.Y., Meijer, H.J.G., Govers, F., 2004. Downstream targets of the *Phytophthora infestans* G alpha subunit PiGPA1 revealed by cDNA-AFLP. Mol. Plant Pathol. 5, 483-494.

Drenth, A., Janssen, E.M., Govers, F., 1995. Formation and Survival of Oospores of *Phytophthora infestans* under Natural Conditions. Plant Pathol. 44, 86-94.

Fellbrich, G., Romanski, A., Varet, A., Blume, B., Brunner, F., Engelhardt, S., Felix, G., Kemmerling, B., Krzymowska, M., Nurnberger, T., 2002. NPP1, a *Phytophthora* associated trigger of plant defense in parsley and *Arabidopsis*. Plant J. 32, 375-390.

Flor, H.H., 1942. Inheritance of pathogenicity of *Melampsora lini*. Phytopathology 32, 653-669.

Gao, H., Narayanan, N.N., Ellison, L., Bhattacharyya, M.K., 2005. Two classes of highly similar coiled coil-nucleotide binding-leucine rich repeat genes isolated from the Rps1-k locus encode *Phytophthora* resistance in soybean. Mol Plant Microbe Interact 18, 1035-1045.

Govers, F., Latijnhouwers, M., 2004. Encyclopedia of Plant and Crop Science DOI: 10.1081/E-EPCS-120019918. Marcel Dekker Inc., New York, pp, 1-5.

Huang, S., van der Vossen, E.A., Kuang, H., Vleeshouwers, V.G., Zhang, N., Borm, T.J., van Eck, H.J., Baker, B., Jacobsen, E., Visser, R.G., 2005. Comparative genomics enabled the isolation of the *R3a* late blight resistance gene in potato. Plant J. 42, 251-261.

Huang, S.W., Vleeshouwers, V., Werij, J.S., Hutten, R.C.B., van Eck, H.J., Visser, R.G.F., Jacobsen, E., 2004. The *R3* resistance to *Phytophthora infestans* in potato is conferred by two closely linked *R* genes with distinct specificities. Mol. Plant Microbe Interact. 17, 428-435.

Innes, R.W., 2004. Guarding the goods. New insights into the central alarm system of plants. Plant Physiol. 135, 695-701.

Jiang, R.H.Y., Tyler, B.M., Whisson, S.C., Hardham, A.R., Govers, F., 2006. Ancient origin of elicitin gene clusters in *Phytophthora* genomes. Mol. Biol. Evol. in press

Kamoun, S., Hraber, P., Sobral, B., Nuss, D., Govers, F., 1999. Initial assessment of gene diversity for the oomycete pathogen *Phytophthora infestans* based on expressed sequences. Fungal Genet. Biol. 28, 94-106.

Latijnhouwers, M., de Wit, P.J., Govers, F., 2003. Oomycetes and fungi: similar weaponry to attack plants. Trends Microbiol. 11, 462-469.

Luderer, R., Joosten, M.H.A.J., 2001. Avirulence proteins of plant pathogens: determinants of victory and defeat. Mol. Plant Pathol. 2, 355-364.

Michelmore, R.W., Paran, I., Kesseli, R.V., 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc. Natl. Acad. Sci. USA 88, 9828-9832.

Nielsen, H., Engelbrecht, J., Brunak, S., vonHeijne, G., 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Eng. 10, 1-6.

Nielsen, H., Krogh, A., 1998. Prediction of signal peptides and signal anchors by a hidden Markov model. Proc. Int. Conf. Intell. Syst. Mol. Biol. 6, 122-130.

Qutob, D., Kamoun, S., Gijzen, M., 2002. Expression of a *Phytophthora sojae* necrosis-inducing protein occurs during transition from biotrophy to necrotrophy. Plant J. 32, 361-373.

Randall, T.A., Dwyer, R.A., Huitema, E., Beyer, K., Cvitanich, C., Kelkar, H., Fong, A.M., Gates, K., Roberts, S., Yatzkan, E., Gaffney, T., Law, M., Testa, A., Torto-Alalibo, T., Zhang, M., Zheng, L., Mueller, E., Windass, J., Binder, A., Birch, P.R., Gisi, U., Govers, F., Gow, N.A., Mauch, F., van West, P., Waugh, M.E., Yu, J., Boller, T., Kamoun, S., Lam, S.T., Judelson, H.S., 2005. Large-scale gene discovery in the oomycete *Phytophthora infestans* reveals likely components of phytopathogenicity shared with true fungi. Mol. Plant Microbe Interact. 18, 229-243.

Rehmany, A.P., Gordon, A., Rose, L.E., Allen, R.L., Armstrong, M.R., Whisson, S.C., Kamoun, S., Tyler, B.M., Birch, P.R., Beynon, J.L., 2005. Differential recognition of highly divergent downy mildew avirulence gene alleles by RPP1 resistance genes from two Arabidopsis lines. Plant Cell 17, 1839-1850.

Rooney, H.C., van't Klooster, J.W., van der Hoorn, R.A., Joosten, M.H., Jones, J.D., de Wit, P.J., 2005. *Cladosporium* Avr2 inhibits tomato Rcr3 protease required for *Cf-2*-dependent disease resistance. Science 308, 1783-1786.

Sacks, W., Nurnberger, T., Hahlbrock, K., Scheel, D., 1995. Molecular characterization of nucleotide sequences encoding the extracellular glycoprotein elicitor from *Phytophthora megasperma*. Mol. Gen. Genet. 246, 45-55.

Shan, W., Cao, M., Leung, D., Tyler, B.M., 2004. The *Avr1b* locus of *Phytophthora sojae* encodes an elicitor and a regulator required for avirulence on soybean plants carrying resistance gene *Rps1b*. Mol. Plant Microbe Interact. 17, 394-403.

Tyler, B.M., 2001. Genetics and genomics of the oomycete host interface. Trends Genet. 17, 611-614.

Tyler, B.M., 2002. Molecular basis of recognition between *Phytophthora* pathogens and their hosts. Ann. Rev. Phytopathol. 40, 137-167.

van den Ackerveken, G., Bonas, U., 1997. Bacterial avirulence proteins as triggers of plant disease resistance. Trends Microbiol. 5, 394-398.

van der Lee, T., De Witte, I., Drenth, A., Alfonso, C., Govers, F., 1997. AFLP linkage map of the oomycete *Phytophthora infestans*. Fungal Genet. Biol. 21, 278-291.

van der Lee, T., Robold, A., Testa, A., van't Klooster, J.W., Govers, F., 2001. Mapping of avirulence genes in *Phytophthora infestans* with amplified fragment length polymorphism markers selected by bulked segregant analysis. Genetics 157, 949-956.

van der Lee, T., Testa, A., Robold, A., van 't Klooster, J., Govers, F., 2004. High-density genetic linkage maps of *Phytophthora infestans* reveal trisomic progeny and chromosomal rearrangements. Genetics 167, 1643-1661.

van der Vossen, E., Sikkema, A., Hekkert, B.L., Gros, J., Stevens, P., Muskens, M., Wouters, D., Pereira, A., Stiekema, W., Allefs, S., 2003. An ancient *R* gene from the wild potato species *Solanum bulbocastanum* confers broad-spectrum resistance to *Phytophthora infestans* in cultivated potato and tomato. Plant J. 36, 867-882.

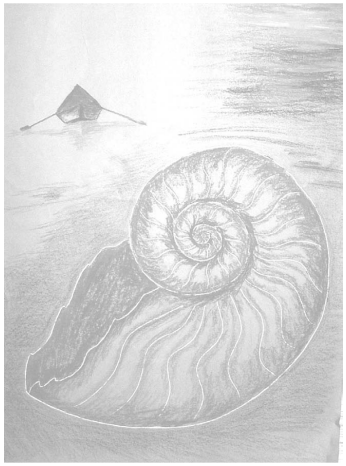
van't Slot, K.A.E., Knogge, W., 2002. A dual role for microbial pathogen-derived effector proteins in plant disease and resistance. Crit. Rev. Plant Sci. 21, 229-271.

Westerink, N., Joosten, M.H.A.J., de Wit, P.J.G.M., 2004. Fungal (A)virulence factors at the crossroads of disease susceptibility and resistance. In: Punja, Z. (Ed.), Fungal Disease Resistance in Plants. Biochemistry, Molecular Biology and Genetic Engineering. Haworth Press, pp, 93-127.

Whisson, S.C., van derLee, T., Bryan, G.J., Waugh, R., Govers, F., Birch, P.R.J., 2001. Physical mapping across an avirulence locus of *Phytophthora infestans* using a highly representative, large-insert bacterial artificial chromosome library. Mol. Genet. Genomics 266, 289-295.

White, F.F., Yang, B., Johnson, L.B., 2000. Prospects for understanding avirulence gene function. Curr. Opin. Plant Biol. 3, 291-298.

Young, N.D., 2000. The genetic architecture of resistance. Curr. Opin. Plant Biol. 3, 285-290.



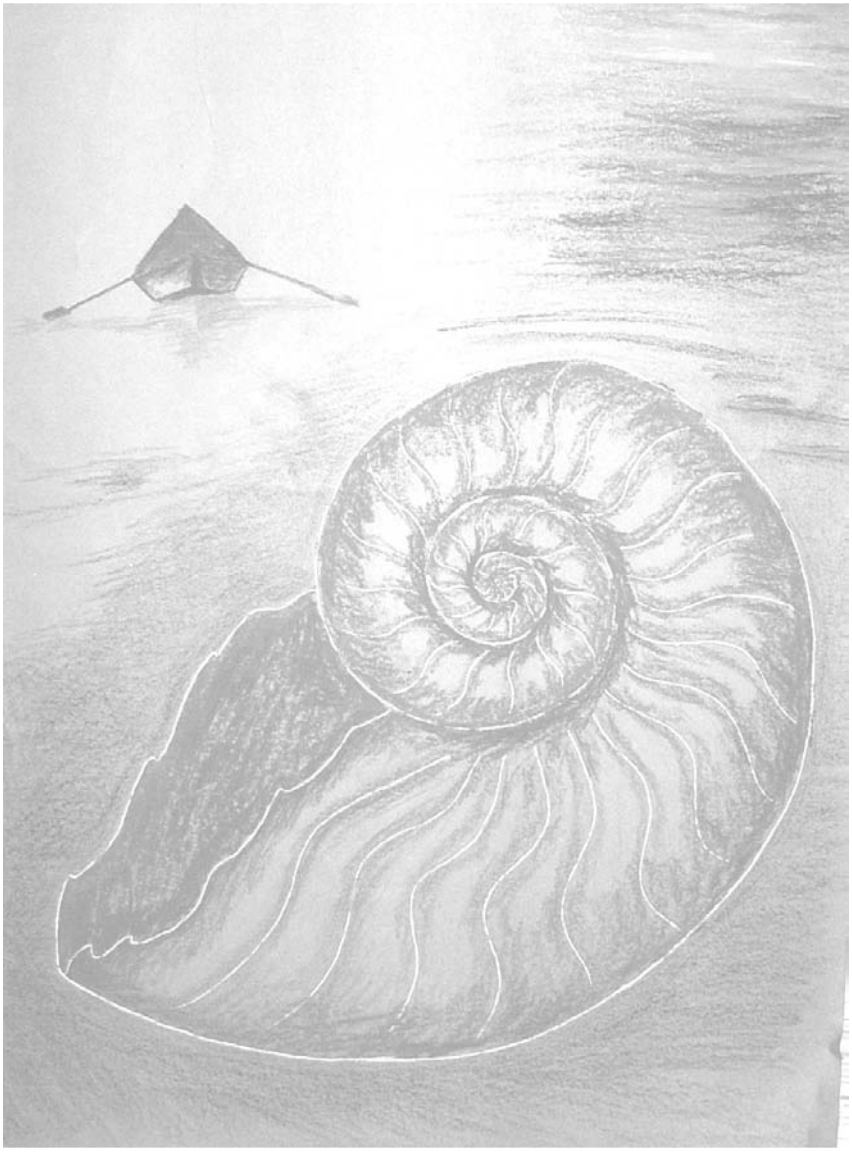
Chapter 3

Amplification generates modular diversity at an avirulence locus in the pathogen *Phytophthora*

submitted

Rays H.Y. Jiang*, Rob Weide*, Peter van de Vondervoort, and Francine Govers

* These authors contributed equally to the work



Amplification generates modular diversity at an avirulence locus in the pathogen *Phytophthora*

Rays H.Y. Jiang[1], Rob Weide[1], Peter van de Vondervoort, and Francine Govers[2]

Laboratory of Phytopathology, Plant Sciences Group, Wageningen University, Binnenhaven 5, NL-6709 PD Wageningen and Graduate School Experimental Plant Sciences, The Netherlands

[1]These authors contributed equally

Keywords

avirulence, gene amplification, *Phytophthora,* modular diversity

Abbreviations

CNV (Copy Number Variation)

CGH (Comparative Genomic Hybridization)

BSA (Bulked Segregant Analysis)

pi3.4 (*Phytophthora infestans* 3.4*)*

pi3.4F^{A} (Phytophthora infestans 3.4 Full length Avirulence associated)

pi3.4F^{V} (Phytophthora infestans 3.4 Full length Virulence associated*)*

pi3.4T (Phytophthora infestans 3.4 Truncated)

pi3.4A (*Phytophthora infestans* 3.4 Amplified*)*

Abstract

The destructive late blight pathogen *Phytophthora infestans* is notorious for its rapid adaptation to circumvent detection mediated by plant resistance (*R*) genes. We performed comparative genomic hybridization on microarrays (array-CGH) in a near genome-wide survey to identify genome rearrangements related to changes in virulence. Six loci with copy number variation were found, one of which involves an amplification co-localizing with a previously identified locus that confers avirulence in combination with either *R* gene *R3b*, *R10* or *R11*. Besides array-CGH, we used three independent approaches to find candidate genes at the *Avr3b-Avr10- Avr11* locus: positional cloning, cDNA-AFLP analysis and Affymetrix® array expression profiling. This resulted in one candidate, *pi3.4* that encodes a protein of 1956 amino acids with regulatory domains characteristic for transcription factors. Amplification is restricted to the 3'end of the full length gene but the amplified copies still contain the hallmarks of a regulatory protein. Sequence comparison showed that the amplification may generate modular diversity and assist in the assembly of novel full length genes via unequal crossing-over. Analyses of *P. infestans* field isolates revealed that the *pi3.4* amplification correlates with avirulence; isolates virulent on *R3b*, *R10* and *R11* plants lack the amplified gene cluster. The ancestral state of *3.4* in the *Phytophthora* lineage is a full length, single copy gene. In *P. infestans*, however, *pi3.4* is a dynamic gene that is amplified and has moved to other locations. Modular diversity could be a novel mechanism for pathogens to quickly adapt to changes in the environment.

Introduction

Microorganisms that are successful as pathogens often have flexible genomes. In bacterial pathogens gene flow caused by horizontal gene transfer is a key event to gain pathogenicity (Schmidt and Hensel 2004) and gene amplifications can lead to increased virulence (Romero and Palacios 1997). In eukaryotic pathogens genomic rearrangements can play a critical role in creating antigenic variation to evade host defence responses (Vanhamme and Pays 1995). Genome plasticity also contributes to the success of many different classes of plant pathogens. For example, loss of elicitor genes to evade host detection has been described in several plant pathogenic fungi (van Kan et al. 1991; Westerink et al. 2004; Rohe et al. 1995; Orbach et al. 2000), and also genomic rearrangements in unstable repeat-rich, subtelomeric regions and transposon insertions generating new virulent alleles have been reported (Orbach et al. 2000; Luderer et al. 2002). Rearrangements within coding regions can also result in changes in virulence. Due to the variation in the number of repeats in the AvrBs3 elicitor family of plant pathogen *Xanthomonas* spp. host and cultivar specificity of these bacteria can change rapidly (Leach and White 1996). Many plant pathogenic bacteria have a type III secretion system (TTSS) that is utilized to inject effectors into plant cells (Alfano and Collmer 2004) and the TTSS effector genes are mostly found on hyper-variable regions. Rearrangements in these regions are often responsible for changes in virulence (Alfano et al. 2000).

In nature plant pathogens with a high evolutionary potential are more likely to overcome genetic resistance in the host plant than those with a low evolutionary potential (McDonald and Linde 2002). Factors that determine evolutionary potential include gene flow, population size, mutation rate and reproduction system (sexual or asexual). *Phytophthora infestans*, the subject of this study, has a high evolutionary potential (McDonald and Linde 2002). *P. infestans* causes late blight, the disease that led to the Irish potato famine in the mid 1840s and resulted in the death or displacement of millions of people (Bourke 1993). Until today late blight continues to cause huge losses worldwide. The genus *Phytophthora* comprises over 65 notorious plant pathogens that not only cause considerable damage to many different commercially grown crops but also to natural vegetations (Nicholls 2004). *Phytophthora* is a fungus-like organism but unrelated to fungi. It is classified as a Stramenopile together with diatoms and brown algae.

A classical strategy to fight plant pathogens in agricultural settings is to breed and grow resistant plants. Natural resistance found in wild relatives can easily be introgressed in crop plants and often provides specific resistance to particular races or pathotypes of a pathogen. This high specificity between a particular host genotype (cultivar) and a particular pathogen genotype (race) is the basis of the gene-for-gene hypothesis launched by Flor and Oort in the 1940's (Flor 1942; Oort 1944) and supported by numerous recent studies focused on the molecular and functional identification of plant resistance genes (*R* genes) and their matching avirulence (*Avr*) genes in the pathogen (van't Slot and Knogge 2002). Resistance breeding in potato (*Solanum tuberosum*) using the Mexican species *Solanum demissum* resulted in eleven independent late blight resistant potato lines showing race-specific resistance (Wastie 1991) and, according to the gene-for-gene hypothesis, the presence of eleven corresponding *Avr* genes in the pathogen is anticipated. Genetic analyses on both potato and *P. infestans* could indeed confirm this (Alkherb et al. 1995; Spielman et al. 1989; van der Lee et al. 2001a). Unfortunately, very soon after the resistant potato lines were exposed to the natural *P. infestans* population new races of *P. infestans* appeared that could easily overcome the resistance (Wastie 1991). Presumably, the *Avr* genes undergo rapid changes so that the pathogen evades detection by the introgressed *R* genes.

P. infestans is absolutely notorious for its ability to change in response to major *R* genes and such adaptive changes may be accelerated by the genome plasticity of *P. infestans*. Among field isolates high levels of genetic variation have been found, not only in Central America, the center of origin of *P. infestans,* but also in North America and Western Europe (Goodwin et al. 1998; Goodwin et al. 1992; Zwankhuizen et al. 1998). Moreover, polyploidy seems to occur in field isolates (Gu et al. 1993; Tooley 1987) as well as aneuploidy and trisomy (Carter et al. 1999; van der Lee et al. 2004). In addition, chromosomal deletions and translocations related to virulence and mating behavior have been described (Judelson 1996; van der Lee et al. 2001b). Also transposon(-like) elements that often contribute to genome plasticity, are very abundant in *P. infestans* (Jiang et al. 2005). Micro-synteny was found between *Phytophthora* species with different genome sizes but *P. infestans* possesses larger intergenic regions as compared to other species and this is mainly due to the abundance of heterogeneous

transposons interspersed between genes. Recent transposition of some retrotransposon families in *P. infestans* possibly led to its large genome size (240 Mb) (Jiang et al. 2006).

In *P. infestans*, we have strong indications that a change in phenotype on *R3b, R10,* and *R11* potato lines is associated with a genome rearrangement or chromosomal deletion (van der Lee et al. 2001b). In previous mapping studies, all based on a cross 71 mapping population (for details see methods), we identified a locus that carries three closely linked dominant *Avr* genes, *Avr3*, *Avr10* and *Avr11* (van der Lee et al. 2001a; van der Lee et al. 2004). *Avr3* has recently been renamed to *Avr3b* to indicate that this *Avr* gene elicits resistance on plants carrying resistance gene *R3b* but not *R3a* (Huang et al. 2004). The *Avr3b-Avr10-Avr11* locus resides in a subtelomeric region, and in the avirulent parent and progeny of cross 71 this locus is hemizygous (van der Lee et al. 2001b). Marker analyses revealed that the avirulent parent has a number of high copy repeat sequences that are absent in the virulent parent. In contrast, the virulent parent and progeny of cross 71 have a chromosomal deletion in the vicinity of the *Avr* locus that could span the *Avr* genes. In a large collection of field isolates that are virulent on potato lines carrying *R3b*, *R10* and *R11* the *Avr3b-Avr10-Avr11*-linked markers are also absent suggesting a similar chromosomal deletion (van der Lee et al. 2001b). To further investigate the correlation between genome rearrangements and the avirulence phenotype, we need a more in depth analysis of the structural organization of the genes residing at this locus.

To clone the *Avr* locus we initially adopted a positional cloning approach. By means of Bulked Segregant Analysis (BSA), several AFLP markers tightly linked to *Avr3b, Avr10* and *Avr11* were obtained (van der Lee et al. 2001a) and these were used to screen a BAC library. The resulting BAC contig partially spanned the locus (Whisson et al. 2001) but repeat sequences at the BAC ends and the lack of recombinants hampered extending the BAC contig by chromosome walking to eventually cover the complete locus. In parallel to positional cloning we performed expression profiling with the aim to identify *Avr*-associated transcripts. cDNA-AFLP was performed on pools of F1 progeny that differed in their virulence phenotypes. The template for cDNA-AFLP was RNA isolated from germinating cysts where *Avr* genes are expected to be expressed. Genome-wide screening resulted in a number of trancript derived fragments (TDFs) that are present in avirulent strains but absent in virulent strains (Guo et al. 2006).

For the present study we used the *Avr3b-Avr10-Avr11* linked AFLP markers (van der Lee et al. Genetics 2001), the initial BAC contig (Whisson et al. 2001) and the *Avr3b-Avr10-Avr11*-associated TDFs (Guo et al. 2006) as starting point for cloning the locus. Screening the BAC library with an *Avr*-associated TDF allowed us to extend the BAC contig. We then selected additional transcriptome markers by screening a custum made *Phytophthora* GeneChip composed of 18,256 unigenes generated from a large scale EST project (Randall et al. 2005). Six markers obtained via three different approaches all landed within a relatively small region of 10 kb that comprises an open reading frame (ORF). Comparison of the avirulent and virulent haplotype revealed a remarkable amplification of truncated copies of the ORF in the avirulent haplotype that could fuction as a source of modules for generating novel full length genes. In eukaryotes, assembly of existing gene modules is a significant mechanism for evolution of novel biological functions (Patthy 2003; Tordai et al. 2005). To determine whether copy number variation

(CNV) is related to virulence, we used the *Phytophthora* GeneChip for array-based comparative genomic hybridization (array-CGH) a method that can detect amplifications and deletions at a genome-wide scale (Vissers et al. 2005; Lucito et al. 2003). Array-CGH revealed CNV at six loci in the *P. infestans* genome, one of which is indeed the *Avr3b-Avr10-Avr11* locus. Comparison with *P. sojae* and *P. ramorum* showed that the *Avr3b-Avr10-Avr11* locus in *P. infestans* is highly rearranged. The biological relevance of the rearrangement and the modular diversity found at the *Avr3b-Avr10-Avr11* locus is discussed.

Results

Isolation of two cDNA-AFLP fragments co-segregating with the *Avr3b-Avr10-Avr11* locus

Previously, we used a cDNA-AFLP-based strategy to identify transcripts in germinating cysts that are associated with avirulence in *P. infestans*. In a bulked segregant analysis (BSA) approach pools of cDNA consisting of F1 progeny of a mapping population segregating for the AVR3b-AVR10-AVR11 phenotype, were screened. Of 23 TDFs that were only present in the avirulent pools but not in the virulent pools segregation in the mapping population was analysed and two, TDF3.1 and TDF3.4, showed 100% association with the AVR3b-AVR10-AVR11 phenotype in the two parental strains and 18 individual F1 progeny (Guo et al. 2006). All TDFs selected by BSA were cloned and sequenced and primers were designed to further confirm the transcript polymorphism by RT-PCR. RNA isolated from germinating cysts from the two parental strains and 9 F1 progeny was analysed. Again TDF3.1 and TDF3.4 showed a polymorphism that correlates with the AVR3b-AVR10-AVR11 phenotype (Fig. 1). RT-PCR with TDF3.4 primers resulted in a 115 bp RT-PCR product in the avirulent strains whereas the product was absent in the virulent strains. RT-PCR with TDF3.1 primers resulted in a clear amplification product of 54 bp in avirulent strains. In contrast, in the virulent strains a less intense band showed up that was also a few bp larger in size. From these results we conclude that in the mapping population TDF3.1 and TDF3.4 co-segregate with the *Avr3b-Avr10-Avr11* locus.

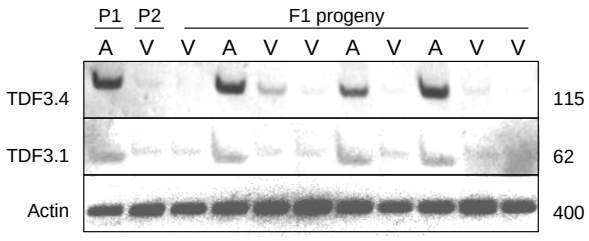


Figure 1. Expression of two candidate TDFs in germinating cysts of different *P. infestans* strains. RT-PCR of TDF3.4 and TDF3.1 on the two parental strains of the mapping population, 80029 (P1) and 88133 (P2), and 9 F1 progeny. The sizes of the amplification products (bp) are on the right. A stands for avirulent and refers to strains with AVR3b-AVR10-AVR11 phenotype, whereas V stands for virulence for strains with the avr3b-avr10-avr11 phenotype.

TDF3.4 and TDF3.1 are located on a Linkage Group VIII BAC contig that comprises the *Avr3b-Avr10-Avr11* locus

Genomic Southern blot analysis was performed with DNA isolated from the two parental strains 80029 and 88133 and digested with *Pst*I and *Eco*RI. TDF3.1 hybridized to two fragments and showed a clear polymorphism in the two strains: the fragments are only present in the genome of the avirulent parent 80029 (Fig. 2A) indicating that the virulent parent lacks a TDF3.1 homologue. In contrast, TDF3.4 hybridized to several restriction fragments with different intensities and with obvious polymorphism between the two parents (Fig 2B) suggesting that TDF3.4 represents a gene family.

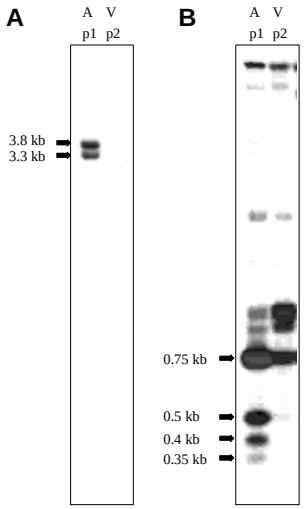


Figure 2. Genomic Southern blot analysis reveals polymorphism of two candidate TDFs in the parental strains 80029 (p1) and 88133 (p2) of a mapping population. The arrows indicate the polymorphic bands. The DNA was hybridized with probes derived from TDF3.1 (*A*) or TDF3.4 (*B*).

Screening of a 10x coverage BAC library of *P. infestans* strain T30-4 with the TDF3.4 probe resulted in over 50 positive BACs. This large number is consistent with the presence of multiple loci in the genome. Strain T30-4 is an F1 progeny of strain 80029 and 88133 that is heterozygous at the *Avr3b-Avr10-Avr11* locus. The dominant allele is inherited from parent 80029 (Whisson et al. 2001). Interestingly, two BACs, 34G01 and 40D03, hybridized much stronger to the TDF3.4 probe than the other positive BACs. BAC34G01 and BAC40D03 together with 38 randomly chosen positive BACs were fingerprinted by restriction analyses, and Southerns blots containing *Bgl*II digested BAC DNA were screened with the TDF3.4 probe. Based on BAC fingerprinting and TDF3.4 hybridization patterns, six BAC contigs could be distinguished (Supplemental Research Data Table S1). When the TDF3.1 probe was used to screen the BAC Southerns blots three BACs, including BAC34G01 and BAC40D03, gave a clear signal. These three BACs belong to contig-I, demonstrating that this contig is shared by both TDF3.4 and TDF3.1.

Previously we identified a minimum tiling path of four BACs partially spanning the *Avr3b-Avr10-Avr11* locus (Whisson et al. 2001). Those BACs were selected by screening 3-dimensional pools of the BAC library with AFLP markers located on Linkage Group VIII and linked to the *Avr3b-Avr10-Avr11* locus (van der Lee et al. 2001a). In that screening we identified two additional BACs that carry the *Avr3b-Avr10-Avr11* co-segregating AFLP marker E+AA/M+CTs239 but they could not be integrated in the

contig comprising the minimal tiling path (Whisson et al. 2001). These BACS were identified as BAC34G01 and BAC40D03, the BACs that in the present study showed a very strong hybridization signal with TDF3.4. As shown in Fig. 3A, four BACs assigned to contig-I, the one sharing TDF3.4 and TDF3.1, can now be bridged with the minimum tiling path (here defined as contig-0) via BAC22O12. The other contig-I BACs did not show up as positives when screening the library with *Avr*-linked AFLP markers nor did they hybridize to TDF3.1. Therefore we assume that these BACs represent the virulent haplotype.

Based on these results we conclude that the genetically identified *Avr3b-Avr10-Avr11* locus on Linkage Groep VIII is covered by BAC contig-I.

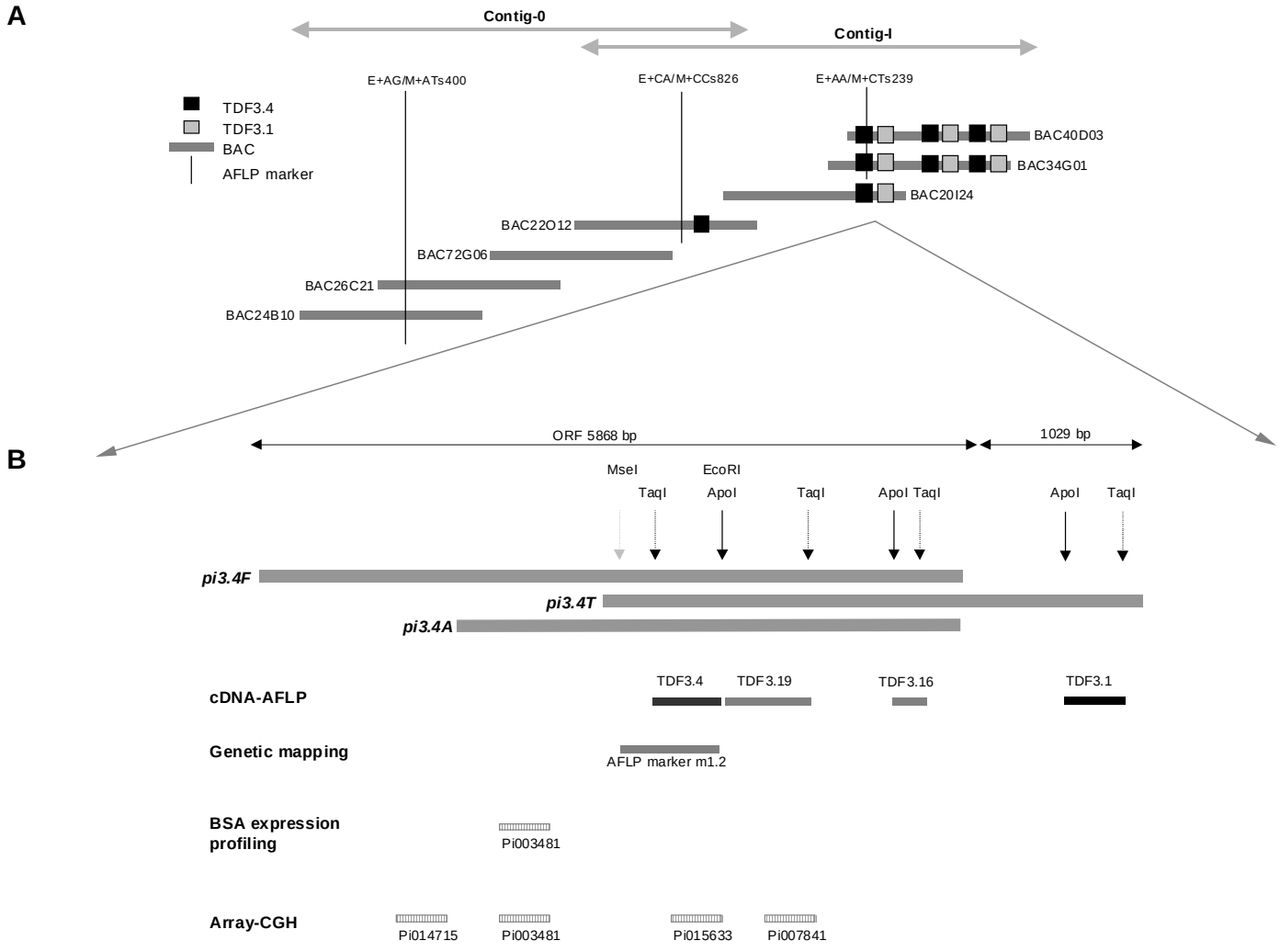


Figure 3. The *Avr3b-Avr10-Avr11* locus on Linkage Group VIII. *(A)* BAC contigs across the *Avr3b-Avr10-Avr11* locus. Contig-0 is the BAC contig previously identified by Whisson et al. (2001) *(B)* Multiple markers resulting from independent approaches (as listed on the left side) are derived from *pi3.4*. m1.2 is the cloned AFLP marker E+AA/M+CTs239. The sizes of BACs, genes and markers are not drawn on scale.

The *Avr3b-Avr10-Avr11* locus contains one full length gene

As shown by the intensity of the hybridization signals, contig-I contains multiple copies of TDF3.4. BAC20I24, however, is the only BAC that is confirmed to contain the *Apo*I restriction site giving rise to TDF3.4 and was therefore selected to clone a putative avirulence gene. Southern blot hybridization showed that this BAC contains at least two copies of TDF3.4 located on two different *Bam*HI fragments of about 7 and 5 kb, respectively. The 7 kb fragment was subcloned and sequence analysis revealed the presence of a single ORF of 5871 bp. The 1.1 kb sequence upstream of the ORF contains a consensus for an oomycete transcription initiation site (McLeod et al. 2004; Pieterse et al. 1994) suggesting a 5'UTR of 547 bp. The *Bam*H1 cloning site is located 25 bp downstream of the stopcodon. Analysis of other overlapping subclones of BAC20I24 showed that the 5 kb *Bam*HI fragment represents a tandem repeat each carrying a TDF3.4 copy. One of the repeats has a 100% match with the TDF3.4 sequence and an ORF of 2877 bp in the same frame as the ORF on the 7 kb fragment ending with the same stopcodon. There are, however, several SNPs when compared to the sequence of the 7 kb fragment and also several in frame stopcodons upstream of TDF3.4. In the 3'UTR sequence there is a polyadenylation site, 1029 bp downstream of the stopcodon, and a sequence that is identical to TDF3.1 at the very end. To distinguish the different copies we named the full length copy on the 7 kb *Bam*HI fragment *pi3.4F (Phytophthora infestans* 3.4 Full length*)* and the second copy that lacks the N-terminus, *pi3.4T* (*Phytophthora infestans* 3.4 Truncated). Other copies are named *pi3.4* unless mentioned otherwise.

Of the six BAC contigs that hybridize to TDF3.4 only BAC contig-I contains the full length copy *pi3.4F*. Hybridization of the BAC Southern blots with a probe derived from the 5' end of the ORF showed that all contig-I BACs were positive except the two multicopy BACs, BAC34G01 and BAC40D02 (Supplemental Research Data Table S1). Because of the diagnostic *Apo*I site we know that *pi3.4F* on BAC20I24 co-segregates with the avirulence phenotype and we gave this particular copy a suffix A for Avirulence associated: $pi3.4F^{A}$. The allelic copy $pi3.4F^{V}$, is derived from the virulent haplotype and sequencing of subclones from BAC15E9 revealed an ORF of exactly the same length as in $pi3.4F^{A}$. Remarkably, the promoter region and the first 4155 nucleotides of the ORF are exactly identical in $pi3.4F^{A}$ and $pi3.4F^{V}$. Only in the last 1716 nucleotides of the ORF are 30 SNPs and immediately downstream of the stopcodon the sequences diverge (Supplemental Research Data Fig S1). $pi3.4F^{V}$ is the only *pi3.4* copy present on BAC15E9 and the only *pi3.4* copy present at this locus in the virulent haplotype.

Multiple markers selected from three independent approaches are derived from gene *Pi3.4*

Analysis of the sequences obtained from the *Avr3b-Avr10-Avr11* locus showed that in addition to TDF3.1 and TDF3.4 two other TDFs, that were among the 23 *Avr*-associated TDFs in the BSA screening, are present in *pi3.4F* and *pi3.4T* (i.e. TDF3.16 and TDF3.19; Guo et al. 2006). Their sequences are identical to sequences in *pi3.4F* and *pi3.4T* with the exception of one SNP between TDF3.19 and *pi3.4F.* TDF3.16 and TDF3.19 were initially ignored because RT-PCR and/or cDNA-AFLP analysis showed no

clear segregation in the F1 progeny. *pi3.4F* should give rise to a large transcript (approx. 6.5 kb) and as a result multiple TDFs can be generated from the same gene. Apart from the sequence alignment the four TDFs can be matched to *pi3.4F* and *pi3.4T* based on the restriction enzyme sites used for cDNA-AFLP (Fig. 3B).

To find additional transcriptome markers for the *Avr3b-Avr10-Avr11* locus a second BSA expression profiling strategy was performed. The pooling strategy was similar to that used for cDNA-AFLP analysis (Guo at al. 2006; Supplemental Research Data Table S2) and pooled RNA samples derived from germinating cysts, were used as probes on a custom designed *Phytophthora* GeneChip. Array sequences with more than two fold higher expression in avirulent pools as compared to virulent pools were selected as *Avr* candidates. Among 18,256 array sequences, three candidates were found of which one is derived from *pi3.4*. The sequence of pi003481 (472 bp) has one SNP when aligned with the *pi3.4F*^{A}. The other two candidates either failed to give the expected polymorphism with RT-PCR or did not co-segregate with the AVR3b-AVR10-AVR11 phenotype and were not pursued further (data not shown).

Previously most of the AFLP markers linked to the *Avr3b-Avr10-Avr11* locus have been cloned (van der Lee et al. 2001b). Sequence comparison showed that the 154 bp sequence of TDF3.4 is contained within the 239 bp sequence of AFLP marker E+AA/M+CTs239 (clone m1.2). The polymorphism at the *Eco*RI site in the genomic DNA that gave rise to a marker on the genetic map overlaps with the *Apo*I site that forms the basis of an expression profiling marker.

In conclusion, one AFLP marker selected by genetic linkage mapping, four TDFs selected by cDNA-AFLP analysis and one array sequence selected by the GeneChip expression array (Table 1) all landed in a relatively small region of less than 10 kb. Thus six candidate sequences obtained by three independent approaches are derived from gene *pi3.4* (Fig. 3B).

Table 1. Multiple markers resulting from BSA expression profiling, genetic mapping and array-CGH are localized in *pi3.4*.

markers	type of fragment	AFLP codes	Size (bp)	selection criteria	confirmation by RT-PCR	aligns to
TDF3.1	cloned cDNA-AFLP fragment	A+AG/T+ACs154	154	differential expression in AVR strains	x	3'UTR of *pi3.4T*
TDF3.16	cloned cDNA-AFLP fragment	A+GG/T+CCs85	85	differential expression in AVR pools		ORF
TDF3.19	cloned cDNA-AFLP fragment	A+GT/T+TAs169	169	differential expression in AVR pools		ORF
TDF3.4	cloned cDNA-AFLP fragment	A+AA/T+AGs158	158	differential expression in AVR strains	x	ORF
m1.2	cloned AFLP fragment	E+AA/M+CTs239	239	co-segregation with AVR phenotypes		ORF
PI003481	EST derived GeneChip array clone	-	472	differential expression in AVR strains and increased copy number in AVR strains		ORF
PI007841	EST derived GeneChip array clone	-	344	increased copy number in AVR strains		ORF
PI014715	EST derived GeneChip array clone	-	1358	increased copy number in AVR strains		ORF
PI015633	EST derived GeneChip array clone	-	191	increased copy number in AVR strains		ORF

The hemizygous *Avr3b-Avr10-Avr11* locus contains numerous *pi3.4* repeats

The *Avr3b-Avr10-Avr11* locus was proposed to be located in a hemizygous region because all linked AFLP markers were in coupling phase with avirulence while markers in repulsion phase were never found (van der Lee et al. 2001a). Moreover m5.1, a cloned fragment of AFLP marker E+AG/M+ATs400 located in contig-0 (Fig. 3A), is absent in the virulent parent and virulent progeny of the cross 71 mapping population (van der Lee et al. 2001b). Similarly, in this study we showed that TDF3.1, located in the 3'UTR of *pi3.4T,* is unique for the avirulent parent demonstrating that this truncated copy of *pi3.4* is absent in the virulent parent (Fig. 3B). These data confirm that the *Avr3b-Avr10-Avr11* locus is hemizygous with deletions in the virulent allele.

On the other hand, the avirulent allele seems to comprise specific repeats. Two cloned AFLP markers, designated m1.2 and m7.1 and segregating with the AVR3b-AVR10-AVR11 phenotype, showed stronger hybridization to genomic DNA of the avirulent parent 80029 and avirulent progeny than to that of virulent parent and progeny (van der Lee et al. 2001b). Also TDF3.4 which is part of m1.2 and located in the coding region of *pi3.4,* showed strong hybridization to a number of genomic DNA fragments in the avirulent parent (Fig. 2B) and avirulent progeny (data not shown). This high intensity of hybridization indicates the presence of *Pi 3.4* repeats in AVR3b-AVR10-AVR11 strains. Since two contig-I BACs (BAC34G01 and BAC40D03) appear to contain multiple copies of *pi3.4* we conclude that the repeats are located at the *Avr3b-Avr10-Avr11* locus. BAC34G01 and BAC40D03 both have an insert size of 160 kb and largely overlap with each other. From the library screening with TDF3.4 as a probe we estimated that the hybridization intensity was around 50 fold higher as compared to other positive BACs, which suggests ca. 50 copies of *pi3.4* at that particular locus. Presence of multiple repeats on these two BACs was demonstrated by restriction digestion with various enzymes resulting in repeat-like patterns with few but very intense bands (data not shown). With the notion that the full length ORF of *pi3.4F* is nearly 6 kb, we anticipate that many truncated or fragmented copies of *pi3.4* are located on BAC34G01 and BAC40D03.

To investigate which part of *pi3.4* is amplified, probes spanning different parts of the gene were designed, i.e., the 5' and 3' end of *pi3.4F*^{A} ORF and the immediate 3'UTR of *pi3.4T* (see material and methods for details). These were hybridized to *Bgl*II digested DNA of BAC34G01 and BAC40D03. With the probe of the 5' end of the ORF no hybridization was found but with the 3' end probe that overlaps with the TDF3.4 sequence, there was a very strong hybridizing fragment of 2.3 kb and the signal was around 50 fold higher than that of other *pi3.4* containing BACs. Also with the 3'UTR probe and TDF3.1 a strong signal was obtained. These results show that at the *Avr3b-Avr10-Avr11* locus truncated copies of *pi3.4* are amplified. We refer to these copies as *pi3.4A* (*Phytophthora infestans* 3.4 Amplified*)*

Amplification identified by array-CGH

To investigate the relationship between gene amplification and avirulence phenotype we used comparative genomic hybridization (CGH). CGH has traditionally been used to investigate gene amplifications or deletions at a genome-wide scale using metaphase chromosomes as targets (Suijkerbuijk et al. 1994). More recently array-based CGH methods were developed that allow genome-wide screening of genomic copy number changes at a much higher resolution and array-CGH is now commonly used to detect genomic aberrations in cancer or to identify disease genes in humans (Vissers et al. 2005). Arrays of genomic fragments or oligonucleotide microarrays are hybridized with total genomic DNA isolated from different individuals or tissues, and differences in hybridization intensity reveal which sequences are amplified or deleted (Lucito et al. 2003).

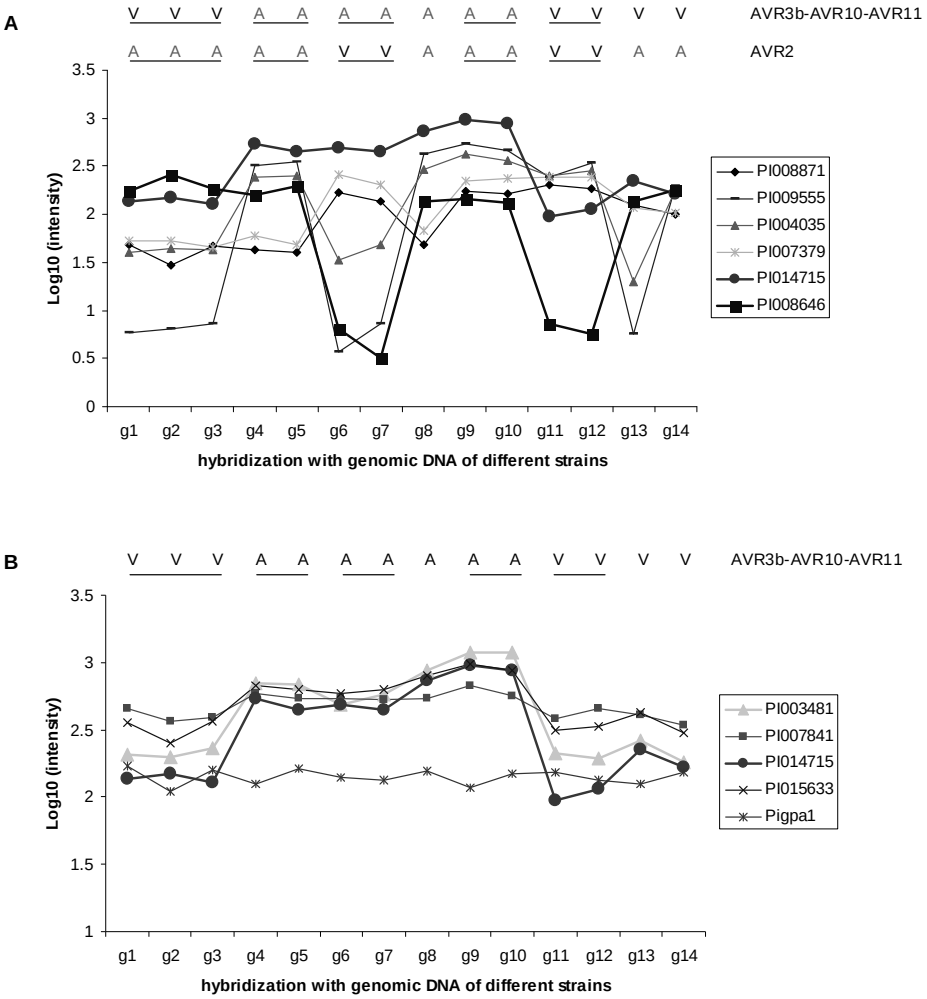


Figure 4. Array-CGH reveals copy number variation (CNV). The *Phytophthora* GeneChip was hybridized with genomic DNA derived from isolates T30-2 (g1, g2, g3), T15-5 (g4, g5), D12-17 (g6, g7), T30-3 (g8), T20-4 (g9, g10), D12-6 (g11, g12), 90128 (g13) and 88069 (g14). The phenotypes of the strains are indicated by V (Virulent) or A (Avirulent). Samples derived from the same strain are underlined. The Y-axis shows the log10 of the hybridization intensity. *(A)* Eight array sequences representing six genes show CNV (standard deviation >0.5). Pi003481 and Pi004035 (not shown here) gave similar patterns as Pi014715 and Pi001829, respectively, and each of these pairs is derived from the same gene. Pi014715 and Pi008646 show correlations with avirulence phenotypes, (AVR3b-AVR10-AVR11 and AVR2, respectively) whereas the other four show no correlation with any known phenotype. For further details see Supplemental Research Data Table S4. *(B)* Four array sequences derived from *pi3.4* show an AVR3b-AVR10-AVR11 correlated CNV. The array sequence representing the single copy gene *Pigpa1* was plotted as control.

In this study we used the custom designed *Phytophthora* GeneChip as the basis and genomic DNA derived from eight *P. infestans* strains, including six F1 progeny from the mapping population with different AVR phenotypes, as hybridization probes. Array-GCH revealed that out of 18,256 array sequences eight sequences representing six genes have large copy number changes (standard deviation > 0.5) (Fig. 4A). As the selection criteria were rather stringent these six loci are expected to have drastic variations in copy numbers in the strains that we analysed. Interestingly, two of the amplified loci could be correlated to AVR phenotypes (Fig. 4A). Amplification of two array sequences, pi003481 and pi014715, was observed in strains with the AVR3b-AVR10-AVR11 phenotype. Interestingly, both array sequences have sequence homology with *pi3.4* demonstrating that also an unbiased approach reveals a drastic amplification of *pi3.4* in avirulent strains. The second amplification was observed with array sequence pi008646 and occurs in strains that have an AVR2 phenotype. This sequence has no homology with any known gene and there is no homologue in *P. sojae* or *P. ramorum*. Moreover, pi008646 did not show an AVR2 associated expression pattern in the BSA expression profiling experiment using the same GeneChip and, as yet, the significance of this amplification and the relation with *Avr2* is not clear.

In addition to pi003481 and pi014715, the GeneChip contains two other sequences with sequence homology to *pi3.4* (pi007841 and pi015633). All four sequences show 97-99% similarity to *pi3.4F*^{A} (Table 2). They are either derived from different *pi3.4.*copies or they represent polymorphisms between strain T30-4 and 88069, the strain from which most of the array sequences are deduced. The behaviour of all four *pi3.4* sequences in the array-CGH is similar with three of the four showing a more than two fold increase in intensity in AVR3b-AVR10-AVR11 strains as compared to avr3b-avr10-avr11 strains (Fig. 4B). The somewhat lower increase seen with pi007841 is possibly due to the fact that only 3 out 13 oligonucleotides on the GeneChip that represent pi007841 are identical to the *pi3.4F*^{A} sequence; the other ten have SNPs or gaps (Table 2). The reason that pi007841 and pi015633 did not show up in the unbiased screening is the difference in stringency.

Table 2. Ratios of hybridization levels in avirulent and virulent strains. For array-GCH, the custom designed *Phytophthora* GeneChip was used. The first four array sequences listed are derived from *pi3.4.*

code of the array sequence	size of the sequence (bp)	polymorphisms compared to *pi3.4F*^{A}	polymorphisms compared to *pi3.4F*^{V}	number of oligo's identical to *pi3.4F*^{A} or with mismatch [a,b]	number of oligo's identical to *pi3.4F*^{V} or with mismatch [a,b]	avirulent/virulent strains (array-CGH)
PI003481 [c]	472	1 SNPs	2 SNPs	12 / 1	13 / 0	3.8
PI014715	1358	27 SNPs	27 SNPs	10 / 3	9 / 4	4.4
PI015633	191	6 SNPs	0 SNPs	7 / 6	13 / 0	2.2
PI007841	344	8 SNPs 26 bp gap, 27 bp can not be aligned	8 SNPs 26 bp gap, 27 bp can not be aligned	3 / 10	2 / 12	1.4
PI000084 (*Pigpa1*) [d]	254	0 (compared to *Pigpa1*)				1.0

[a] for each array sequence, a set of 13 oligos (25 bp) was designed and synthesized on the array
[b] mismatch oligos contain 1-15 mismatching nucleotides as compared to *pi3.4F*^{A} or *pi3.4F*^{V}
[c] PI003481 was also selected by BSA expression analysis on the *Phytophthora* GeneChip
[d] PI000084 represents the single copy gene *Pigpa1*

Since the four identified array sequences are the only ones out of a total of 18,256 array sequences that show distinct AVR3b-AVR10-AVR11 correlated patterns, the gene amplification associated with AVR3b-AVR10-AVR11 seems to be specific for *pi3.4.* One should bear in mind however, that this *Phytophthora* GeneChip was not designed as a genome-wide array. The oligonucleotides on the array are based on a unigene set deduced from over 75,757 ESTs from different tissues and different developmental stages (Randall et al. 2005). It is not known how many more sequences on the Gene-Chip, if any, correspond to this particular region on Linkage Group VIII and it is therefore not possible the delineate the borders of the amplification.

Pi3.4 amplification in *P. infestans* field isolates is associated with AVR3b-AVR10-AVR11 phenotypes

The array-CGH experiment included two *P. infestans* field isolates, 88069 and 90128 that are both virulent on *R3b*, *R10* and *R11* plants. Similar to field isolate 88133 (i.e. the virulent parent of the mapping population) 88069 and 90128 show no amplification of *pi3.4* (Fig. 4A).

Table 3. The presence of AFLP marker E+AA/M+CTs239 derived from *pi3.4* correlates with the AVR3b phenotype in field isolates. The phenotypes on *R1*, *R4*, *R3b*, *R10* and *R11* are indicated by avirulent (A) and virulent (V). The presence or absence of AFLP markers is indicated by + or -, respectively.

isolates	*R1*	*R4*	*R3b*	*R10*	*R11*	E+AA/ M+CTs239	E+GG/ M+CCs251	E+CA/ M+GGs826	E+AG/ M+AAs258	E+TG/ M+CTs240
80029[a]	A	V	A	A	A	+	+	+	+	+
88133[a]	V	A	V	V	V	-	-	-	-	-
T15-2[b]	A	A	V	V	V	-	-	-	-	-
T30-4[b]	V	V	A	A	A	+	+	+	+	+
85152	V	V	V	A	V	-	-	-	-	-
85192	V	V	V	V	V	-	+	+	+	+
86004	V	V	V	V	V	-	+	+	+	+
86022	V	V	V	V	A	-	+	+	+	+
87013	V	V	V	-	-	-	+	+	+	+
87194	V	V	V	V	V	-	-	-	-	-
88039	V	V	V	A	V	-	+	+	+	+
88184	V	V	V	V	V	-	-	-	-	-
88215	A	V	V	V	V	-	-	-	-	-
89005	V	V	V	V	A	-	+	+	+	+
89018	V	V	V	V	V	-	+	+	+	+
89047	V	V	A	A	A	+	+	+	+	+
89140.02	V	V	V	V	V	-	+	+	+	+
89140.10	V	V	V	V	V	-	+	+	+	+
89150.01	V	V	V	V	V	-	+	+	+	+
89153.01	V	V	V	V	V	-	-	-	-	+
89155.03	A	V	A	A	A	+	+	+	+	+
90040	V	V	V	V	V	-	-	-	-	-
90041	V	V	V	V	V	-	-	-	-	-
90101	A	A	V	V	V	-	+	+	+	+
90126	A	V	V	A	V	-	nd	-	-	-
90214.03	V	V	V	A	V	-	+	+	+	+
90222	V	V	V	V	V	-	-	-	-	-
91001	A	A	A	A	A	+	+	nd	+	+
91002	A	A	A	A	A	+	+	+	+	+
91011	A	V	V	A	V	-	+	+	+	+
91012	A	V	V	V	-	-	+	+	+	+

[a] parents of the mapping population
[b] F1 progeny of the mapping population
nd: not determined

To further analyze the association between *pi3.4* and the AVR3b-AVR10-AVR11 phenotype in field isolates, the *pi3.4*-derived AFLP marker E+AA/M+CTs239 (van der Lee et al. 2001a) was analysed in 29 field isolates collected in the Netherlands between 1980 and 1991. This population has highly diverse avirulence phenotypes and comprises many different DNA fingerprint genotypes (Drenth et al. 1994). The presence of E+AA/M+CTs239 correlates with the AVR3b phenotype in five isolates including 80029, the avirulent parent of the mapping population, whereas 24 isolates virulent on *R3b* plants (race 3b) lack the AFLP marker. Also four other AFLP markers that were previously mapped at the *Avr3b-Avr10-Avr11* locus (van der Lee et al. 2001a) and are all located in the contig-0 region, were analysed but none of these showed this strong correlation (Table 3). This demonstrates that also in field isolates a polymorphism that maps in *pi3.4* correlates with the AVR3b phenotype. The five AVR3b isolates are also avirulent on *R10* and *R11* plants. However, from the 24 race 3b isolates seven are avirulent on either *R10* or *R11*. This might suggest that not in all cases *Avr3b*, *Avr10* and *Avr11* are as closely linked as in the mapping population. It should be noted, however, that symptoms on *R10* and *R11* plants are in general less severe than symptoms on *R3b* plants (van der Lee et al. 2001a) and therefore phenotypic scoring for AVR10 and AVR11 may not be as accurate as for AVR3b.

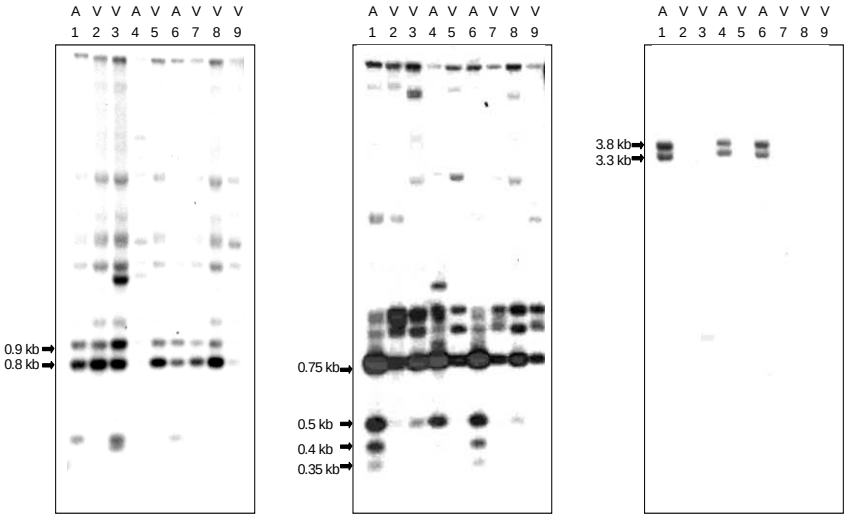


Figure 5. Genomic Southern blot analysis reveals a correlation between *pi3.4* amplification and avirulence in field isolates. Genomic DNA derived from 80029 (1), 88133 (2), 84044 (3), 85005 (4), 85025 (5), 87001(6), 87010 (7), 87177 (8), and 88175 (9) was digested with *Eco*R1 and *Pst*I, and hybridized with probes from *(A)* the 5'end of *pi3.4F*, *(B)* the 3'end of *pi3.4F* and *(C)* the 3'UTR of *pi3.4T*. The probes used for hybridization are described in material and methods.

To investigate whether *pi3.4* is amplified in the Dutch field isolates, a set of 17 isolates with different AVR phenotypes was subjected to genomic Southern blot analysis. Hybridization with a probe from the 5' end of the *pi3.4F* ORF resulted in multiple hybridizing fragments in all isolates with some polymorphisms but there is no correlation with the avirulence phenotype (shown for nine isolates in Fig. 5A). The two strongly hybridizing fragments of 0.8 and 0.9 kb, respectively, are present in all isolates and co-migrate with the fragments in BAC20I24. A probe derived from the 3' end of the ORF also hybridizes to multiple polymorphic fragments (Fig. 5B). The polymorphic bands of 0.4 kb and 0.35 kb are present in the AVR3b-AVR10-AVR11 isolates but absent in virulent isolates. Two additional bands of 0.5

kb and 0.75 kb show a higher level of intensity in avirulent isolates than in virulent ones indicating that also in these field isolates the presence of a *pi3.4* amplification is associated with avirulence. Probes derived from the 3'UTR (3'UTR probe and TDF3.1) of *pi3.4T* both hybridize to the same two fragments of 3.8 kb and 3.3 kb (Fig. 5C) but these are absent in virulent isolates. This shows that also in various field isolates the presence of *pi3.4T* and *pi3.4A* is diagnostic for the AVR3b-AVR10-AVR11 phenotype.

pi3.4F and pi3.4T encode proteins that resemble transcription factors

At the *Avr3b-Avr10-Avr11* locus, we distinguish three different *pi3.4* genes: the full length $pi3.4F^{A}$, the truncated *pi3.4T* and the amplified *pi3.4A*. The virulent haplotype also contains a full lenght *pi3.4* copy at this locus, $pi3.4F^{V}$, but lacks the truncated and the amplified copies. Both $pi3.4F^{A}$ and $pi3.4F^{V}$ contain one intact continuous ORF and encode a protein of 1956 amino acids that has several typical domains and motifs (Fig. 6).

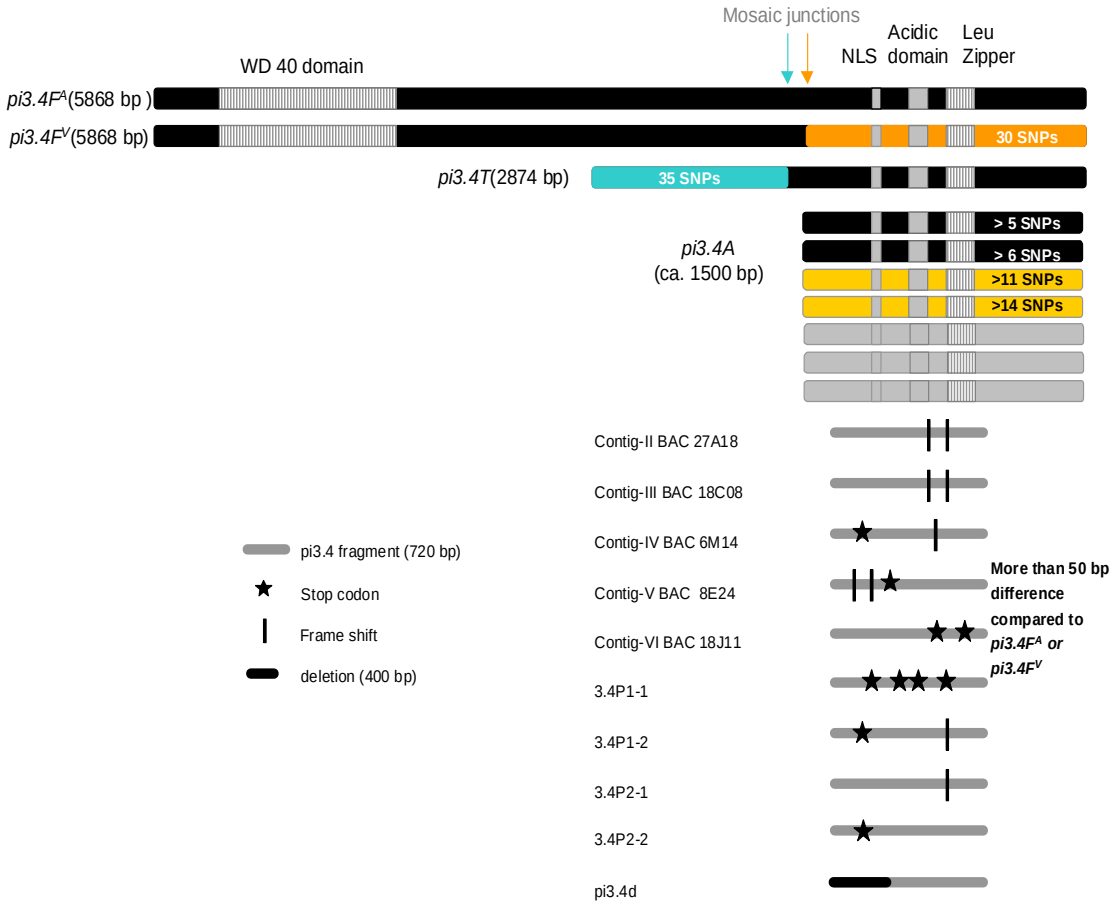


Figure 6. Mosaic structure in Pi3.4F. $pi3.4F^{A}$, *pi3.4T* and *pi3.4A* are copies present in the avirulent haplotype. $pi3.4F^{V}$ is derived from the virulent haplotype. The size of *pi3.4A* is estimated from transcript lenghts. From various *pi3.4A* copies 720 bp were sequenced and the number of SNPs detected in these 720 bp fragments is indicated. Sequences derived from other loci carry frame shifts and stop codons, or have deletions. For alignments see Supplemental Research Data Figures S1 and S2.

The N-terminal part of the protein contains a WD40 domain and at the C terminus there is a Nuclear Localization Signal (NLS) followed by an acidic domain comprised of nearly only D and E residues, and a leucine zipper. The truncated copy, *pi3.4T,* has an ORF with the capacity to produce a protein of 959 amino acids but since this ORF contains several potential start codons downstream of the first ATG we do not know the exact size of the pi3.4T protein. Nevertheless, *pi3.4T* still codes for a protein that includes the NLS, the acidic domain and the leucine zipper. The two *Phytophthora* species that have been fully sequenced each contain one highly similar homologue of *pi3.4F* but no truncated copies. At the protein level the homologues in *P. sojae* (*ps3.4*) and *P. ramorum* (*pr3.4*) show 77% similarity to Pi3.4F whereas the similarity between Ps3.4 and Pr3.4 is 81% (Supplemental Research data Fig S1). There are no homologues known in other species, except in dog (*Canis familiaris*) where one protein (XP_535743) is predicted with a similar domain organization. The identity in the N-terminal part is 30% (252/838) and in the C-terminal part 26% (264/1002). The function of Pi3.4 is unknown but the combination of a NLS, an acidic domain and a leucine zipper often occurs in transcription factors.

Modular diversity at the *Avr3b-Avr10-Avr11* locus results in mosaic structures

pi3.4F^A and pi3.4F^V show a very unequal distribution of polymorphism. The first 4155 nucleotides (1385 amino acids) are identical but within the last 1716 nucleotides there are 30 SNPs causing 10 amino acid changes (Fig. 6 and Supplemental Research data Fig S1). The distribution of SNPs between *pi3.4F^A* and *pi3.4T* on the one hand and *pi3.4F^V* and *pi.3.4T* on the other hand, is remarkable. In the region where *pi3.4F* and *pi3.4T* overlap en where *pi3.4F^A* and *pi3.4F^V* are identical, *pi3.4T* shows 33 SNPs resulting in 6 amino acids changes. However, downstream of the point of diversification between *pi3.4F^A* and *pi3.4F^V*, *pi3.4T* has only 2 SNPs with *pi3.4F^A* but 29 SNPs with *pi3.4F^V*. Mosaic junctions seem to occur in the region surrounding the point of diversification between *pi3.4F^A* and *pi3.4F^V* (Fig. 6). We also analyzed a few *pi3.4A* copies located on the multicopy BACs. Sequence analysis of PCR fragments of 720 bp covering the NLS, the acidic domain and the leucine zipper and representing four different *pi3.4A* copies, revealed no stop codons or frame shifts. Two of the four showed closer sequence similarity to *pi3.4F^A* with 5 and 6 SNPs, and the other two to *pi3.4F^V* with 11 and 13 SNPs (Fig. 6). This suggests that the diversity in the *pi3.4F^V* 3' end was generated from the amplified gene modules via unequal crossing-overs.

pi3.4F, *pi3.4T* and *pi3.4A* are located on BAC contig-I but, in addition, several other *pi3.4* copies are present elsewhere in the genome. To study the sequence diversity of *pi3.4* within the genome, we analysed *pi3.4* fragments that were amplified from BACs belonging to contig-II to VI (Supplemental Research Data Table S1) and *pi3.4* fragments obtained by PCR on genomic DNA. Sequence analsysis of the 720 bp fragments showed that, apart from the intact *pi3.4* copies present at the *Avr3b-Avr10-*

Avr11 locus, all the other copies are pseudo-genes carrying various frameshifts and stop-codon mutations or, in the case of *pi3.4d*, a large deletion of 400 bp (Fig. 6).

Rearrangements in the *3.4F* region in *P. infestans* as compared to *P. sojae* and *P. ramorum*

P. infestans has multiple *pi3.4* copies scattered in the genome, and in AVR3b-AVR10-AVR11 strains there is a *pi3.4* gene amplification at the *Avr3b-Avr10-Avr11* locus. In contrast, *P. sojae* and *P. ramorum,* two species that have been fully sequenced (http://genome.jgi-psf.org/), each have only one homologue of *pi3.4F* in their genome (*ps3.4* and *pr3.4,* respectively) and no truncated copies. Genomic Southern blot analysis confirmed that in *P. sojae, P. ramorum* and several other *Phytophthora* species *3.4* is a single copy gene (data not shown).

Analysis of genomic regions in the vicinity of *ps3.4* and *pr3.4* revealed a high level of co-linearity between *P. sojae* and *P. ramorum* (Fig. 7). A total of 30 orthologous gene pairs can be identified in a region covering 127 kb in *P. sojae* and 93 kb in *P. ramorum*. Only two genes in *P. sojae* and one gene in *P. ramorum* do not have orthologues in this region. The gene order is well conserved except for one reversal. The genes immediately flanking *ps3.4* in *P. sojae* are conserved in *P. ramorum* (see detail in Fig. 7). To be able to investigate whether this co-linearity also exists in *P. infestans* we first searched for the putative orthologues in *P. infestans* using the sequence of the 30 *P. ramorum* orthologues in the region. Among the 18,256 sequences present on the *Phytophthora* GeneChip 20 homologues were assigned based on best BLAST hit with *E* value less than 1e-30 and over 50% matching residues. None of the 20 array sequences, however, showed a gene amplification pattern similar to *pi3.4* (Fig. 7). For a more direct approach we selected four *P. infestans* ESTs that are the homologues of four genes located in the near vicinity of *ps3.4* and *pr3.4* (Fig. 6; Supplemental Research Data Table S3) and used these as probes to screen the BACs containing *pi3.4* copies. The BAC contigs cover around 100-150 kb each and the contig-0/contig-I region is even larger. In *P. ramorum* these four genes are located within a 12 kb region surrounding *pr3.4.* Even when we take into account that the genome of *P. infestans* is less gene dense than that of *P. ramorum* (Jiang et al. 2006), the homologues should still be present on the BAC contigs. None of the four ESTs gave a hybridization signal in the six *pi3.4* BAC contigs, indicating that there is no synteny in any of the six *pi3.4* regions with the *3.4* regions in *P. ramorum* or *P. sojae*. There is, however, synteny between *P. infestans, P. sojae* and *P. ramorum* in other regions. One example is a 100 kb region in *P. infestans* that covers 16 genes including the G-protein-coupled receptor gene *pigcr1,* and is highly conserved in *P. sojae* and *P. ramorum* with respect to genes and gene order (R.H.Y.J., I. Sama and F.G., unpublished data). The absence of synteny at the *P. infestans pi3.4* regions and the fact that *pi3.4* has multiple truncated copies scattered over the genome suggests that the *pi3.4* region has been translocated and is prone to amplification and reshuffling.

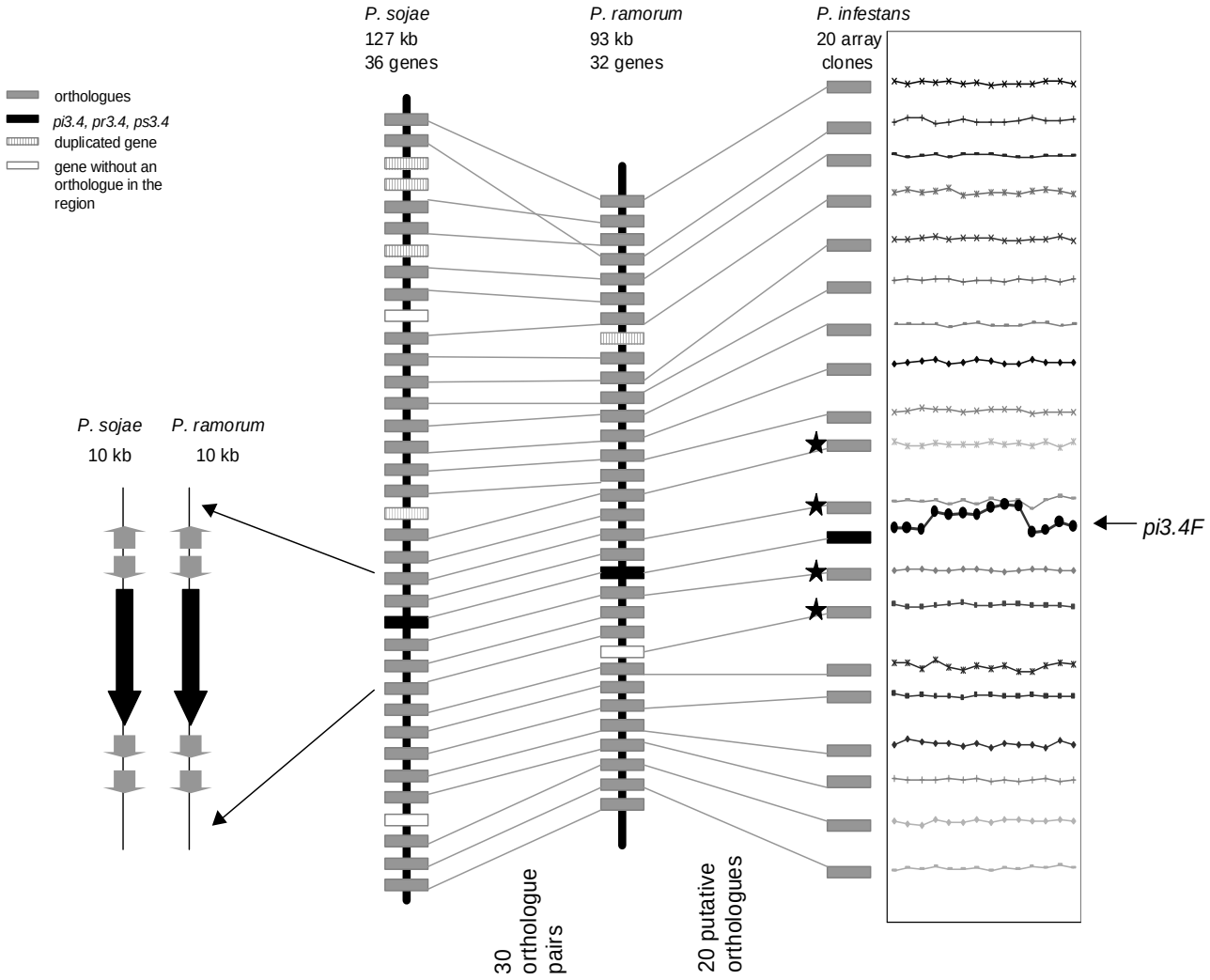


Figure 7. Synteny between *P. sojae* and *P. ramorum* in genomic regions containing the *pi3.4F* homologue. *pr3.4* and *ps3.4* are single copy genes (indicated in black). Among 18,256 *P. infestans* array sequences, 20 homologues of the 30 orthologue pairs are found. Stars indicate the four ESTs used for screening the BAC contig. On the right, the Log10 of the intensity of the array-CGH is plotted next to the homologue in *P. infestans.*

An avirulence locus with a regulatory gene

We have successfully combined transcriptional profiling, genetic mapping, and array-CGH to physically map a complex avirulence locus in *P. infestans* that contains an amplified gene cluster with modular diversity in the amplified gene. Unlike most avirulence loci hitherto identified in plant pathogens the *Avr3b-Avr10-Avr11* locus does not encode a typical secreted elicitor protein. Instead, the *Avr3b-Avr10-Avr11* locus contains genes encoding proteins with a domain organization reminiscent of transcription factors. A full length copy named *Pi3.4F,* is present in the virulent as well as the avirulent haplotype but the truncated copies (*pi3.4T*) are exclusively found in the avirulent haplotype (Fig. 8). Compared to full length, truncated copies lack a WD-40 domain, a domain that is involved in protein-protein interactions but not essential for transcriptional activation. The amplified gene cluster is restricted to isolates with the AVR3b-AVR10-AVR11 phenotype. Upon infection these isolates induce a hypersensitive response (HR) in potato lines carrying the *R3b*, *R10* and *R11* resistance genes and as a result disease development is arrested. The fact that three avirulence genes map genetically at the same locus and that this locus contains a regulatory gene favour a model in which the *Avr3b-Avr10-Avr11* locus is responsible for regulating the expression of other genes. These target genes might encode secreted proteins that function as the AVR factors in the gene-for-gene model and are recognized, directly or indirectly, by the corresponding *R* proteins. In the cross 71 mapping population *Avr3b*, *Avr10* and *Avr11* always cosegregate and our model thus requires that in the parental lines the target genes are either homozygous or consist of two avirulent alleles.

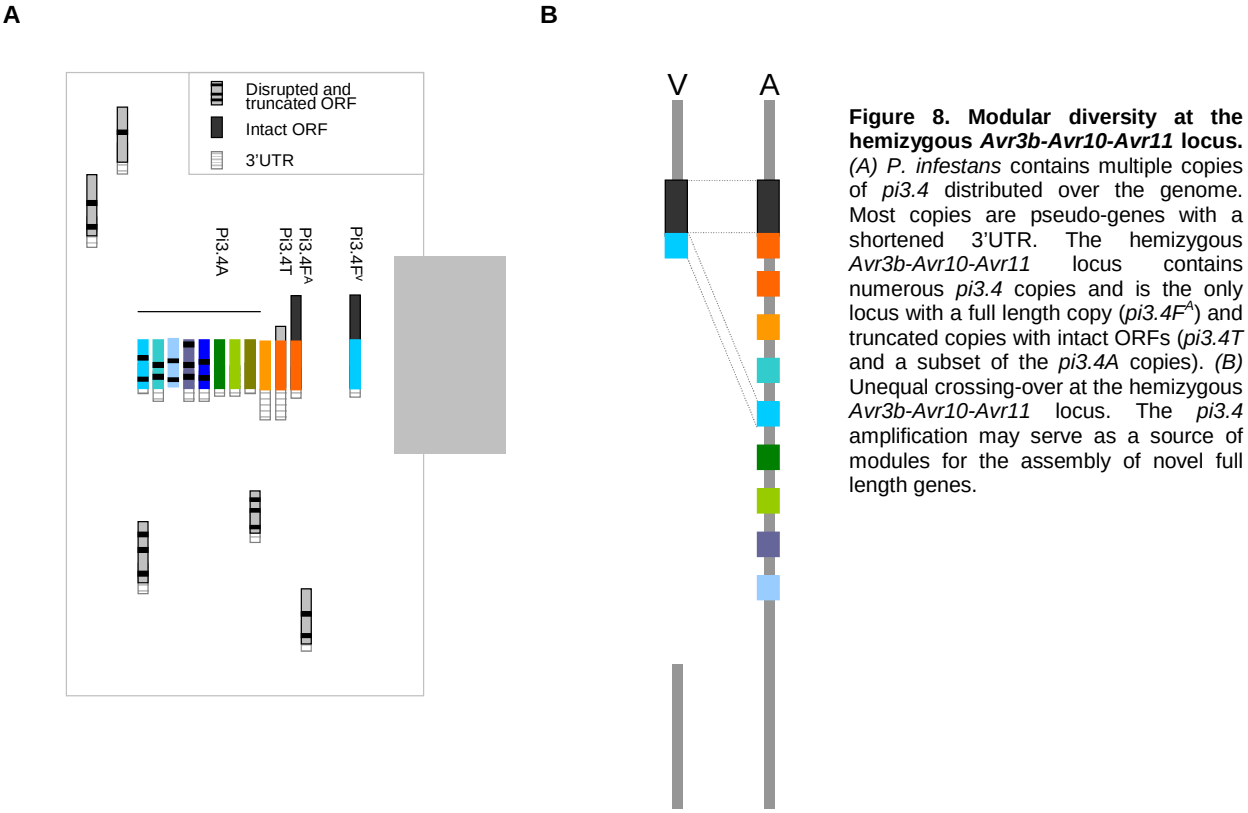


Figure 8. Modular diversity at the hemizygous *Avr3b-Avr10-Avr11* locus. *(A) P. infestans* contains multiple copies of *pi3.4* distributed over the genome. Most copies are pseudo-genes with a shortened 3'UTR. The hemizygous *Avr3b-Avr10-Avr11* locus contains numerous *pi3.4* copies and is the only locus with a full length copy (*pi3.4F*^{A}) and truncated copies with intact ORFs (*pi3.4T* and a subset of the *pi3.4A* copies). *(B)* Unequal crossing-over at the hemizygous *Avr3b-Avr10-Avr11* locus. The *pi3.4* amplification may serve as a source of modules for the assembly of novel full length genes.

To find support for this model we re-examined the results reported by Al-Kherb et al. (1995). This is one of the few studies on inheritance of virulence and avirulence in *P. infestans* and, apart from our own studies (van der Lee et al. 2001a; van der Lee et al. 2001b), the only one including analyses on *Avr10* and *Avr11.* Interestingly, for only one of the nine examined avirulence phenotypes Al Kherb et al. (1995) found indications for a second locus being involved in determining the phenotype. This was observed in two crosses and it concerned *Avr10*. This complies with our model but unfortunately we could not extend these observations to *Avr11* and *Avr3b.* The parents of the two informative crosses were both virulent on *R11* plants and therefore a role for a second locus in regulating *Avr11* could not be determined. In addition, the *R3* plants used by Al Kherb et al. (1995) are from the Black differential set. They carry *R3a* but lack *R3b* (Huang et al. 2004) and hence the AVR3 segregation is not informative for *Avr3b.* In four other crosses Al Kherb et al. (1995) found segregation of both, AVR10 and AVR11, but no indication for linkage between the two *Avr* genes. In view of our model the lack of linkage can be explained when the regulatory locus in each of the parental lines of these four crosses is homozygous, while the target genes are heterozygous and consisting of a virulent and avirulent allele.

Additional support for a regulatory gene at an avirulence locus in oomycetes comes from studies in *P. sojae. P. sojae* has at least three clusters of *Avr* genes (Tyler 2001) one of which comprises the tightly linked *Avr1b* and *Avr1k* (Whisson et al. 1995). *Avr1b* (renamed *Avr1b-1)* was cloned by positional cloning and turned out to be a small secreted protein that has elicitor activity (Rehmany et al. 2005; Shan et al. 2004). However, several isolates carrying an avirulent allele of *Avr1b-1* remain virulent on plants carrying the resistance gene *Rps1b* due to the fact that *Avr1b-1* is not transcribed. A second gene located at the same locus, *Avr1b-2,* is required for accumulation of mRNA of *Avr1b-1* and could encode a transcription factor. At this point we do not know whether the *Avr3b-Avr10-Avr11* locus in *P. infestans* also contains putative target genes. We have no clue what other genes are located in the vicinity of the *pi3.4* genes. Fortunately, genome sequencing of *P. infestans* strain T30-4 is in progress at the Broad Institute and in the future we will be able to scan the *P. infestans* genome for putative target genes. Likely candidates are genes belonging to the RXLR family. They encode a diverse group of secreted proteins that share a RXLR motif which is a hallmark of oomycete effector proteins (Rehmany et al. 2005). The first four cultivar- or ecotype specific avirulence genes that have been identified to date in oomycetes belong to this family, including the *P. sojae Avr1b-1* gene, mentioned above (Shan et al. 2004) and the *P. infestans Avr3a* gene (Armstrong et al. 2005). In plants carrying the corresponding *R* gene, these RXLR proteins elicit a HR but in the absence of the *R* gene the RXLR proteins may be acting as virulence factors. Many proteins secreted by *P. infestans* contain an RXLR domain (Birch et al. 2006). One example is the protein encoded by *ipiO*, a gene that is highly expressed in the periphery of expanding late blight lesions where hyphae invade plant cells (van West et al. 1998). This suggests that IPI-O has a role in pathogenicity and is supported by the observation that IPI-O binds to plant plasma membrane proteins and disrupts cell wall-plasma membrane adhesions (Gouget et al. 2006; Senchou et al. 2004). In bacteria several effector proteins are known to have a dual function in both virulence and avirulence (van't Slot and Knogge 2002) and many effector genes are transcriptionally controlled (Jones 2005; Schell 2000).

The avirulence locus contains a full length copy and truncated copies of *pi3.4*

The fact that two different transcriptional profiling strategies, cDNA-AFLP analysis and GeneChip expression arrays, gave rise to several markers associated with the *Avr3b-Avr10-Avr11* locus suggests extreme expression differences between virulent and avirulent strains. Indeed high levels of mRNA of the truncated size and hybridizing to probes from the truncated copies were present in germinated cysts from avirulent strains but absent in virulent strains. In contrast, full length transcripts were present in all strains but only at very low levels (R.W., R.H.Y.J. and F.G. unpublished data). Presumably, both the full length and truncated transcripts produce proteins but without functional analyses we do not know if these proteins both have a function in determining the avirulent phenotype. With respect to the full length copies the localized polymorphism is remarkable, and raises the possibility that Pi3.4F^{A} targets a different set of genes than Pi3.4F^{V}. The polymorhism, however, is restricted to the C-terminal part and hence the WD40 domain will not discriminate. Also the NLS, the acidic domain and the leucine zipper are identical suggesting interaction with the same DNA elements. It is therefore unlikely that *pi3.4F^{A}* is the gene that is responsible for the phenotype. The truncated copies with the intact ORFs are more attractive candidates since they are unique for avirulent strains. Their C-terminal parts are not very diverse from those in Pi3.4F^{A} or Pi3.4F^{V} suggesting they target the same genes. But why then are strains that carry only *pi3.4F^{V}* not capable of activating the target genes that confer avirulence on *R3a*, *R10* and *R11* plants? One possibility is that the WD40 domain in Pi3.4F binds to a repressor or is itself an intramolecular repressor that prevents Pi3.4F to function as transcriptional activator. Another possibility is that the level of Pi3.4F is far too low to see the effect of activation of the target genes. The gene amplification in avirulent strains may result in an enormous increase of the Pi3.4 protein and thus in a much more efficient activation of the target genes. This can be tested by ectopic (over)expression of a truncated *pi3.4* copy in virulent *P. infestans* strains in which the putative target genes have the suitable genotype (e.g., the virulent parent or virulent progeny of cross 71).

The avirulence locus is a dynamic locus with modular diversity

Physical mapping and sequence analyses of the *Avr3b-Avr10-Avr11* locus revealed multiple copies of *pi3.4* that contain intact ORFs but are slightly different in sequence. In all *P. infestans* isolates that we have analyzed *pi3.4* is a multigene family. In two distant *Phytophthora* species, *P. sojae* and *P. ramorum, 3.4* is a single copy gene that encodes the full length protein. Southern blot hybridization showed that most *Phytophthora* species contain only a single copy of the *3.4* gene with the exception of *Phytophthora mirabilis* and *Phytophthora ipomoea* (P.v.d.V. and F.G. unpublished results). Since the latter two are closely related to *P. infestans* we anticipate that the ancestral state of the *3.4* gene in the *Phytophthora* genus is a single copy gene and that amplification occurred after divergence of the various clades. In *P. infestans* we found several copies outside the *Avr3b-Avr10-Avr11* locus but none of these

has an intact ORF. They all have different mutations suggesting that after the transition from a single copy gene to a multi copy gene family there was no pressure for the various copies to remain functional (Fig. 6).

In *P. sojae* and *P. ramorum* the single copy *3.4* gene is located on scaffolds that are largely syntenic for at least 100 kb and for nearly all genes in this region an orthologue is found in the other species. In contrast, in the *Pi3.4F* region in *P. infestans* the orthologues are absent and apparently, in *P. infestans* (or its predecessor) the *3.4* gene moved from its original position to settle elsewhere in the genome. The fact that we observe amplification of truncated copies at one locus and spread of only truncated copies to other loci suggests that the initial translocation involved the full length *pi3.4* gene. Subsequent amplification by an unkown mechanism must have been the cause of the numerous truncated copies at the *Avr3b-Avr10-Avr11* locus while erroneous recombination and unequal crossing-overs may have resulted in the spread of truncated copies to other loci. It is also possible that the truncated copies moved independently to the other loci without a transfer at the *Avr3b-Avr10-Avr11* locus. To reconstruct these events we need to analyse more *pi3.4*. loci in *P. infestans* isolates of different geographical origin and collected in different time periods. In this study we mainly focused on Dutch field isolates collected in the 1980s and 1990s. All AVR3b-AVR10-AVR11 isolates in this population have the amplification and they have the highest copy number at this locus found so far. In the Mexican isolate TV580 collected in Toluca Valley in the early 1980s, we observed a *pi3.4* amplification at the same locus but based on hybridization intensity we estimate that the copy number is somewhat lower than in the Dutch field isolates. In TV580 the AVR specific TDF3.1 marker is present but, remarkably this marker is completely absent in isolates belonging to the US-1 lineage. These US-1 isolates were collected before 1980, so before the world wide population displacement took place (R.W. and F.G., unpublished results). Nevertheless, among those US-1 isolates there is variation in *pi3.4* copy number but since we have no mapping information we can not link this to a particular locus. The finding that the *pi3.4* amplification occurs in different populations raises the question whether there is any advantage for *P. infestans* to have this amplification. Obviously, for isolates that try to invade plants carrying *R3b*, *R10* or *R11,* the gene amplification results in suicide and the only way to survive is to get rid of the amplification. Unequal crossing-over at the *Avr3b-Avr10-Avr11* locus may facilitate the change from avirulence to virulence and since the full length *pi3.4* copy on the virulent allele has a mosaic structure, it is most likely the result of such an unequeal crossing-over event (Fig. 8). This is supported by the finding that some of the truncated copies from the avirulent haplotype have more sequence similarity to *pi3.4F*^{V} than to *pi3.4F*^{A}. On the other hand, the *pi3.4* amplification may serve as a source of modules for the assembly of novel full length genes, and in turn, this molecular diversity may result in adaptive advantage to the pathogen. Alternatively, the *pi3.4* amplification may result in massive production of regulators that control expression of effector genes. In laboratory infection assays on plants lacking *R3b*, *R10* and *R11* we see no apparent differences in aggressiveness or pathogenicity between strains with and without the *pi3.4* amplification but we can not exclude that in nature, isolates with the *pi3.4* amplification behave different. Under field conditions minor changes could have major effects. Late blight is a polycyclic disease and a

minimal increase in spore density or a slightly shorter infection cycle can speed up the development of the epidemics.

Is the modular diversity responsible for variation?

With array-CGH we found six loci in the *P. infestans* genome that show CNV including *pi3.4.* This analysis was limited to a few Dutch field isolates and therefore we may have missed other loci with CNV. Nevertheless, we expect that the extreme CNV as seen at the *pi3.4*.locus is relatively rare. To date CNV has not been described for any other *Phytophthora* gene despite the fact that *Phytophthora* has many multigene families. Most genes encoding secreted proteins belong to multigene families and often the family members are clustered in the genome (R.H.Y.J. and F.G. unpublished). One example is M96, a family of mating-type specific genes that contains 22 tandemly arrayed copies at a single site (Cvitanich et al. 2005). In contrast to *pi3.4,* M96 is a family of tandemly repeated copies in all investigated *Phytophthora* species and occurs as such in both haplotypes of *P. infestans*. Detailed analysis of the M96 gene family showed that it evolved via concerted evolution but there are various other mechanisms by which gene families can evolve.

Gene amplification is one of the driving forces of genome evolution. It is probably ubiquitous in bacterial genomes and contributes substantially to the prokaryotic genomic plasticity (Romero and Palacios 1997). In eukaryotic cells, gene amplification can greatly help to increase the level of certain gene products required for a specific developmental process (Tower 2004). Gene amplification can also lead to drug resistance in cancer cells, insects and plants (Donn et al. 1984; Field et al. 1988; Lengauer et al. 1998). To our knowledge, amplification of *pi3.4* in *P. infestans* is the first reported case of a gene amplification associated with avirulence in a eukaryotic pathogen. Our data suggests that in *P. infestans* unequal crossing-over is a mechanism to create diversity and to assemble a full length gene from amplified modules. Emergence of novel genes by modular assembly from existing genes is well documented (Patthy 2003; Tordai et al. 2005). Classic examples are the immunoglobulin loci in vertebrates where module amplification and shuffling results in large molecular diversity at the protein level. The *pi3.4* amplification makes the *Avr3b-Avr10-Avr11* locus an attractive locus to study the exact mechanism that is used by *P. infestans* to generate modular diversity and to investige how pathogens can use modular diversity to adapt to their environments.

Methods

***P. infestans* strains, nomenclature and phenotypes**

The *P. infestans* strains used in this study are Dutch field isolates collected in the 1980s and 1990s. Previously two of these stains, 80029 and 88133, were used to generate an F1-progeny suitable for

genetic mapping. This mapping population, designated as cross 71, is described by Drenth et al (1995) and van der Lee et al. (1997). The F1 pogeny strain T30-4 was used to construct a BAC library (Whisson et al. 2001). The nomenclature of genes, gene clusters and phenotypes is according to van der Lee et al. (2001) with one exception; *Avr3* now has the suffix 'b' to indicate that this avirulence gene elicits resistance on plants carrying resistance gene *R3b* and not *R3a* (Huang et al. 2004). Consequently, an avirulent and virulent phenotype on *R3b* plants is indicated by AVR3b but not avr3b, respectively. For this study the phenotype on plants carrying *R3b*, *R10* and *R11* is relevant. Strains 80029 and T30-4 are avirulent on these plants and thus have the AVR3b-AVR10-AVR11 phenotype whereas 88133 is virulent and has the avr3b-avr10-avr11 phenotype.

Nucleic acids manipulation and Southern blot analysis

Genomic *P. infestans* DNA was isolated according to procedures described by Drenth et al. (1995). Isolation of RNA from mycelia, sporangia, zoospores and germinating cysts was performed as described by van West et al. (1998). Southern blot analysis was conducted as described by Drenth et al. (1993). For genomic and BAC Southern blot hybridization various probes were developed that cover different parts of the *pi3.4F* gene. The 5' ORF probe is a 556 bp PCR fragment amplified from BAC20I24 with the primer pair pi3.4F5f (GTGCGCCCACTGTCCAACTGGG) and pi3.4F5r (CCGACAGACAGCGGCTTCCTCG) and covers the region starting 0.12 kb downstream of the startcodon of the *pi3.4F* ORF. The 3' ORF probe is 942 bp PCR fragment amplified from BAC20I24 with the primer pair 3.4gwf (AAGAAACGCGATCTGGATGAATGGG) and 3.4gwr (CAGCTGTAGCAGAGATACGTAAATC) and covers the region starting 4.6 kb downstream of the startcodon of the *pi3.4F* ORF. The 3' UTR probe is derived from a *Bam*H1 sub-clone of BAC20I24 and covers 400 bp immediately downsteam of the stop codon. Also TDF3.1 and TDF3.4 were used as hybridization probes. Nucleic acids manipulations were performed according to standard procedures (Sambrook and Russell 2001). DNA sequencing was done by BaseClear (Leiden, The Netherlands).

BAC library screening, BAC fingerprinting and contig building

The *P. infestans* BAC library was screened as described by Jiang et al. (2005). To obtain BAC fingerprint patterns, 1 µg of BAC DNA was digested with various restriction enzymes and the restriction fragments were visualized by gel electrophoresis as described by Jiang et al. (2005). For contig building fragments from different BACs but sharing identical length were considered as common fragments.

Reverse transcriptase - PCR analysis

For RT-PCR, 10 µg total RNA was treated with 4 units RQ1 RNase-free DNase (Promega, Madison, WI) at 37 ℃ for 1 h to remove genomic DNA. The first-s trand cDNA was synthesized using oligo dT (16) and Superscript II reverse transcriptase for 30 min at 40 ℃ (Gibco-BRL). Sequence-specific primers were used in the subsequent PCR with cDNA as template with 30 cycles (30 s at 94 ℃, 30 s at 56-60 ℃ and

60 s at 72 ℃). The RT-PCR primers that were used are based on the TDF3.1 sequence (GenBank accession DW010104) (TDF3.1f: ACTGCATCACACCATCAG and TDF3.1r: GCCGAACAATAGCTCATG), and the TDF3.4 sequence (GenBank accession DW010117) (TDF3.4f: AGCTGGTTGAAGCGCGAC and TDF3.4r: GGAAGGCCGGAGAGCGTC).

AFLP and cDNA-AFLP

AFLP was performed as described by van der Lee et al. (1997) using the restriction enzyme combination *Eco*RI/*Mse*I and primers with two selective bases. cDNA-AFLP was performed as described by Dong et al (2004) and Guo et al (2006). The nomenclature of the AFLP and cDNA-AFLP markers is according to van der Lee et al. (1997), Dong et al. (2004) and Guo et al. (2006).

The *Phytophthora* GeneChip

The Syngenta custom designed *Phytophthora* GeneChip is an Affymetrix® array containing 19,324 unique sequences of which 18,256 represent unigenes. The sequences were generated from a large scale EST project and represent 75,757 ESTs obtained from libraries representing a wide range of growth conditions, stress responses, and developmental stages (Randall et al. 2005) (http://www.pfgd.org/). Over 82% of the sequences on the GeneChip are from *P. infestans.*

Bulked segregant analysis using the *Phytophthora* GeneChip

For bulked segregant analysis (BSA) RNA of F1 progeny with similar or overlapping AVR phenotypes was pooled. Pools for selecting transcripts associated with avirulence genes were constructed as described by Guo et al. (2006). To probe the custom designed *Phytophthora* GeneChip, a total of 6 RNA samples (four pools and two unpooled samples; Supplemental Research Data Table S2) was used. cDNA synthesis, array hybridization, and intensity normalisation were performed similar to the methods described by Zhu et al. (2001). Array clones showing at least a two-fold induction of hybridization intensity with RNA pools derived from avirulent isolates as compared to RNA pools from virulent isolates were chosen as candidates.

Comparative genomic hybridization using the *Phytophthora* GeneChip

For comparative genomic hybridization (CGH) the custom designed *Phytophthora* GeneChip was hybridized with 14 independent genomic DNA samples representing 6 F1 progeny and two *P. infestans* field isolates, 88069 and 90128, both unrelated to the mapping population. From strain T30-2, three independent genomic DNA samples were isolated from three separate cultures. Similarly from each of the strains T15-5, D12-17, T20-4, D12-6, 88069, and 90128, two independent DNA samples were isolated. Genomic DNA was purified on a continuous cesium chloride-ethidium bromide gradient as described by Sambrook and Russell (2001) and fluorescently labeled using the random priming method

with BioPrime kit (Invitrogen, Carlsbad, CA). In brief, a total of 2 µg of genomic DNA from each sample was used to mix with the 20X random primer solutions and denatured at 99^{o}C for 5 min. Following the immediate cooling to 4^{o}C, 5 µl of 10X dNTP mix with biotin labeled dCTP and 1 µl of Klenow fragments were added to the reaction, and incubate at 37^{o}C for 2 h. Labeled DNA fragments were then assayed by gel electrophoresis and fragments in a size range of 100-200 bp were applied to the GeneChip for hybridization. Reproducible differences in hybridization intensity between samples reflect copy number variations (CNV) in the different strains. Array clones showing at least two fold increase in intensity in one strain as compared to another were considered to have variable copy numbers in different strains. For each array clone, the relative intensity was calculated by dividing the individual absolute intensity by the average intensity of the 14 samples. As an indication of the intensity change within the 14 samples, the standard deviation value was calculated from the 14 relative intensities. Array clones with a standard deviation larger than 0.5 were considered to represent a gene with CNV.

Acknowledgments

This work was financially supported by an Aspasia grant from the Netherlands Organization for Scientific Research (NWO-Aspasia 015.000.057). We kindly thank Syngenta, in particular Tong Zhu, Makoto Ono and George Aux, for making the Syngenta custom designed *Phytophthora* GeneChip available and for performing the GeneChip hybridizations. We are grateful to Jun Guo for help with selecting the TDFs, Pierre de Wit for critically reading the manuscript and Harold Meijer for many helpful suggestions.

References

Alfano, J.R., A.O. Charkowski, W.L. Deng, J.L. Badel, T. Petnicki-Ocwieja, K. van Dijk, and A. Collmer. 2000. The *Pseudomonas syringae Hrp* pathogenicity island has a tripartite mosaic structure composed of a cluster of type III secretion genes bounded by exchangeable effector and conserved effector loci that contribute to parasitic fitness and pathogenicity in plants. *Proc. Natl. Acad. Sci. USA* **97:** 4856-4861.

Alfano, J.R. and A. Collmer. 2004. Type III secretion system effector proteins: double agents in bacterial disease and plant defense. *Annu. Rev. Phytopathol.* **42:** 385-414.

Alkherb, S.M., C. Fininsa, R.C. Shattock, and D.S. Shaw. 1995. The inheritance of virulence of *Phytophthora infestans* to potato. *Plant Pathol.* **44:** 552-562.

Armstrong, M.R., S.C. Whisson, L. Pritchard, J.I. Bos, E. Venter, A.O. Avrova, A.P. Rehmany, U. Bohme, K. Brooks, I. Cherevach, N. Hamlin, B. White, A. Fraser, A. Lord, M.A. Quail, C. Churcher, N. Hall, M. Berriman, S. Huang, S. Kamoun, J.L. Beynon, and P.R. Birch. 2005. An ancestral oomycete locus contains late blight avirulence gene *Avr3a*, encoding a protein that is recognized in the host cytoplasm. *Proc. Natl. Acad. Sci. USA* **102:** 7766-7771.

Bachem, C.W., R.S. van der Hoeven, S.M. de Bruijn, D. Vreugdenhil, M. Zabeau, and R.G. Visser. 1996. Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. *Plant J.* **9:** 745-753.

Birch, P.R., A.P. Rehmany, L. Pritchard, S. Kamoun, and J.L. Beynon. 2006. Trafficking arms: oomycete effectors enter host plant cells. *Trends Microbiol.* **14:** 8-11.

Bourke, A. 1993. *The visitation of god? The potato and the great Irish famine.* Lilliput Press Ltd, Dublin.

Carter, D.A., K.W. Buck, S.A. Archer, T. Van der Lee, R.C. Shattock, and D.S. Shaw. 1999. The detection of nonhybrid, trisomic, and triploid offspring in sexual progeny of a mating of *Phytophthora infestans*. *Fungal Genet. Biol.* **26:** 198-208.

Cvitanich, C., M. Salcido, and H.S. Judelson. 2005. Concerted evolution of a tandemly arrayed family of mating-specific genes in *Phytophthora* analyzed through interspecific and intraspecific comparisons. *Mol Genet Genomics*. DOI 10.1007/s00438-005-0074-8

Donn, G., E. Tischer, J.A. Smith, and H.M. Goodman. 1984. Herbicide-resistant alfalfa cells: an example of gene amplification in plants. *J. Mol. Appl. Genet.* **2:** 621-635.

Drenth, A., E.M. Janssen, and F. Govers. 1995. Formation and Survival of Oospores of *Phytophthora infestans* under Natural Conditions. *Plant Pathol.* **44:** 86-94.

Drenth, A., I.C.Q. Tas, and F. Govers. 1994. DNA-Fingerprinting Uncovers a New Sexually Reproducing Population of *Phytophthora infestans* in the Netherlands. *Eur. J. Plant Pathol.* **100:** 97-107.

Field, L.M., A.L. Devonshire, and B.G. Forde. 1988. Molecular evidence that insecticide resistance in peach-potato aphids (*Myzus persicae Sulz.)* results from amplification of an esterase gene. *Biochem. J.* **251:** 309-312.

Flor, H.H. 1942. Inheritance of pathogenicity of *Melampsora lini*. *Phytopathology* **32:** 653-669.

Goodwin, S.B., C.D. Smart, R.W. Sandrock, K.L. Deahl, Z.K. Punja, and W.E. Fry. 1998. Genetic change within populations of *Phytophthora infestans* in the United States and Canada during 1994 to 1996: Role of migration and recombination. *Phytopathology* **88:** 939-949.

Goodwin, S.B., L.J. Spielman, J.M. Matuszak, S.N. Bergeron, and W.E. Fry. 1992. Clonal diversity and genetic differentiation of *Phytophthora infestans* populations in Northern and central Mexico. *Phytopathology* **82:** 955-961.

Gouget, A., V. Senchou, F. Govers, A. Sanson, A. Barre, P. Rouge, R. Pont-Lezica, and H. Canut. 2006. Lectin receptor kinases participate in protein-protein interactions to mediate plasma membrane-cell wall adhesions in Arabidopsis. *Plant Physiol.* **140:** 81-90.

Gu, W.K., L.J. Spielman, J.M. Matuszak, J.R. Aist, C.J. Bayles, and W.E. Fry. 1993. Measurement of Nuclear DNA Contents of Mexican Isolates of *Phytophthora infestans*. *Mycol. Res.* **97:** 857-860.

Guo, J., R.H.Y. Jiang, L.G. Kamphuis, and F. Govers. 2006. A cDNA-AFLP based strategy to identify transcripts associated with avirulence in *Phytophthora infestans*. *Fungal Genet. Biol.*

Jiang, R.H., A.L. Dawe, R. Weide, M. van Staveren, S. Peters, D.L. Nuss, and F. Govers. 2005. Elicitin genes in *Phytophthora infestans* are clustered and interspersed with various transposon-like elements. *Mol. Genet. Genomics* **273:** 20-32.

Jiang, R.H., B.M. Tyler, S.C. Whisson, A.R. Hardham, and F. Govers. 2006. Ancient origin of elicitin gene clusters in *Phytophthora* genomes. *Mol. Biol. Evol.* **23:** 338-351.

Jones, B.D. 2005. *Salmonella* invasion gene regulation: a story of environmental awareness. *J. Microbiol.* **43 Spec:** 110-117.

Judelson, H.S. 1996. Chromosomal heteromorphism linked to the mating type locus of the oomycete *Phytophthora infestans*. *Mol. Gen. Genet.* **252:** 155-161.

Latijnhouwers, M., W. Ligterink, V.G. Vleeshouwers, P. van West, and F. Govers. 2004. A Galpha subunit controls zoospore motility and virulence in the potato late blight pathogen *Phytophthora infestans*. *Mol. Microbiol.* **51:** 925-936.

Leach, J.E. and F.F. White. 1996. Bacterial avirulence genes. *Annu Rev Phytopathol* **34:** 153-179.

Lengauer, C., K.W. Kinzler, and B. Vogelstein. 1998. Genetic instabilities in human cancers. *Nature* **396:** 643-649.

Lucito, R., J. Healy, J. Alexander, A. Reiner, D. Esposito, M. Chi, L. Rodgers, A. Brady, J. Sebat, J. Troge, J.A. West, S. Rostan, K.C.Q. Nguyen, S. Powers, K.Q. Ye, A. Olshen, E. Venkatraman, L. Norton, and M. Wigler. 2003. Representational oligonucleotide microarray analysis: A high-resolution method to detect genome copy number variation. *Genome Res.* **13:** 2291-2305.

Luderer, R., F.L. Takken, P.J. de Wit, and M.H. Joosten. 2002. *Cladosporium fulvum* overcomes *Cf-2*-mediated resistance by producing truncated AVR2 elicitor proteins. *Mol. Microbiol.* **45:** 875-884.

McDonald, B.A. and C. Linde. 2002. The population genetics of plant pathogens and breeding strategies for durable resistance. *Euphytica* **124:** 163-180.

McLeod, A., C.D. Smart, and W.E. Fry. 2004. Core promoter structure in the oomycete *Phytophthora infestans*. *Eukaryot. Cell* **3:** 91-99.

Michelmore, R.W., I. Paran, and R.V. Kesseli. 1991. Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. *Proc. Natl. Acad. Sci. USA* **88:** 9828-9832.

Nicholls, H. 2004. Stopping the Rot. *PLoS Biol.* **2:** e213.

Orbach, M.J., L. Farrall, J.A. Sweigard, F.G. Chumley, and B. Valent. 2000. A telomeric avirulence gene determines efficacy for the rice blast resistance gene *Pi-ta*. *Plant Cell* **12:** 2019-2032.

Oort, A.J.P. 1944. Onderzoekingen over stuifbrand. II. Overgevoeligheid voor stuifbrand (*Ustilago tritici*). Tijdschr. *Plantenziekten* **50:** 73-106.

Patthy, L. 2003. Modular assembly of genes and the evolution of new functions. *Genetica* **118:** 217-231.

Pieterse, C.M., A.M. Derksen, J. Folders, and F. Govers. 1994. Expression of the *Phytophthora infestans ipiB* and *ipiO* genes in planta and in vitro. *Mol. Gen. Genet*. **244:** 269-277.

Randall, T.A., R.A. Dwyer, E. Huitema, K. Beyer, C. Cvitanich, H. Kelkar, A.M. Fong, K. Gates, S. Roberts, E. Yatzkan, T. Gaffney, M. Law, A. Testa, T. Torto-Alalibo, M. Zhang, L. Zheng, E. Mueller, J. Windass, A. Binder, P.R. Birch, U. Gisi, F. Govers, N.A. Gow, F. Mauch, P. van West, M.E. Waugh, J. Yu, T. Boller, S. Kamoun, S.T. Lam, and H.S. Judelson. 2005. Large-scale gene discovery in the oomycete *Phytophthora infestans* reveals likely components of phytopathogenicity shared with true fungi. *Mol. Plant Microbe Interact.* **18:** 229-243.

Rehmany, A.P., A. Gordon, L.E. Rose, R.L. Allen, M.R. Armstrong, S.C. Whisson, S. Kamoun, B.M. Tyler, P.R. Birch, and J.L. Beynon. 2005. Differential recognition of highly divergent downy mildew avirulence gene alleles by RPP1 resistance genes from two Arabidopsis lines. *Plant Cell* **17:** 1839-1850.

Rohe, M., A. Gierlich, H. Hermann, M. Hahn, B. Schmidt, S. Rosahl, and W. Knogge. 1995. The Race-Specific Elicitor, Nip1, from the Barley Pathogen, *Rhynchosporium Secalis*, Determines Avirulence on Host Plants of the *Rrs1* Resistance Genotype. *Embo Journal* **14:** 4168-4177.

Romero, D. and R. Palacios. 1997. Gene amplification and genomic plasticity in prokaryotes. *Annu. Rev. Genet.* **31:** 91-111.

Sambrook, J. and D.W. Russell. 2001. *Molecular cloning: A laboratory manual*. Cold Spring Harbor Laboratory Press, New York.

Schell, M.A. 2000. Control of virulence and pathogenicity genes of *Ralstonia solanacearum* by an elaborate sensory network. *Annu. Rev. Phytopathol.* **38:** 263-292.

Schmidt, H. and M. Hensel. 2004. Pathogenicity islands in bacterial pathogenesis. *Clin. Microbiol Rev.* **17:** 14-56.

Senchou, V., R. Weide, A. Carrasco, H. Bouyssou, R. Pont-Lezica, F. Govers, and H. Canut. 2004. High affinity recognition of a *Phytophthora* protein by Arabidopsis via an RGD motif. *Cel. Mol. Life Sci.* **61:** 502-509.

Shan, W., M. Cao, D. Leung, and B.M. Tyler. 2004. The *Avr1b* locus of *Phytophthora sojae* encodes an elicitor and a regulator required for avirulence on soybean plants carrying resistance gene *Rps1b*. *Mol. Plant Microbe Interact.* **17:** 394-403.

Spielman, L.J., B.J. McMaster, and W.E. Fry. 1989. Dominance and Recessiveness at Loci for Virulence against Potato and Tomato in *Phytophthora infestans*. *Theor. Appl. Genet.* **77:** 832-838.

Suijkerbuijk, R.F., D.E. Olde Weghuis, M. Van den Berg, F. Pedeutour, A. Forus, O. Myklebost, C. Glier, C. Turc-Carel, and A. Geurts van Kessel. 1994. Comparative genomic hybridization as a tool to define two distinct chromosome 12-derived amplification units in well-differentiated liposarcomas. *Genes Chromosomes Cancer* **9:** 292-295.

Tooley, P.W., and Therrien, C. D. 1987. Cytophotometric determination of the nuclear DNA content of 23 Mexican and 18 non-Mexican isolates of *Phytophthora infestans*. *Exp Mycol* **11:** 19-26.

Tordai, H., A. Nagy, K. Farkas, L. Bányai, and L. Patthy. 2005. Modules, multidomain proteins and organismic complexity. *FEBS J* **272:** 5064-5078.

Tower, J. 2004. Developmental gene amplification and origin regulation. *Annu. Rev. Genet.* **38:** 273-304.

Tyler, B.M. 2001. Genetics and genomics of the oomycete host interface. *Trends Genet* **17:** 611-614.

van der Lee, T., A. Robold, A. Testa, J.W. van 't Klooster, and F. Govers. 2001a. Mapping of avirulence genes in *Phytophthora infestans* with amplified fragment length polymorphism markers selected by bulked segregant analysis. *Genetics* **157:** 949-956.

van der Lee, T., A. Testa, A. Robold, J. van 't Klooster, and F. Govers. 2004. High-density genetic linkage maps of *Phytophthora infestans* reveal trisomic progeny and chromosomal rearrangements. *Genetics* **167:** 1643-1661.

van der Lee, T., A. Testa, J. van 't Klooster, G. van den Berg-Velthuis, and F. Govers. 2001b. Chromosomal deletion in isolates of *Phytophthora infestans* correlates with virulence on R3, R10, and R11 potato lines. *Mol. Plant Microbe Interact.* **14:** 1444-1452.

van Kan, J.A.L., G.J.M. van den Ackerveken, and P.J.G.M. de Wit. 1991. Cloning and characterization of cDNA of avirulence gene *Avr9* of the fungal pathogen *Cladosporium fulvum*, causal agent of tomato leaf mold. *Mol. Plant Microbe Interact.* **4:** 52-59.

van West, P., A.J. de Jong, H.S. Judelson, A.M.C. Emons, and F. Govers. 1998. The *ipiO* gene of *Phytophthora infestans* is highly expressed in invading hyphae during infection. *Fungal Genet. Biol.* **23:** 126-138.

Vanhamme, L. and E. Pays. 1995. Control of gene expression in trypanosomes. *Microbiol. Rev.* **59:** 223-240.

van't Slot, K.A.E. and W. Knogge. 2002. A dual role for microbial pathogen-derived effector proteins in plant disease and resistance. *Crit. Rev. Plant Sci.* **21:** 229-271.

Vissers, L.E., J.A. Veltman, A. Geurts van Kessel, and H.G. Brunner. 2005. Identification of disease genes by whole genome CGH arrays. *Hum. Mol. Genet.* **14 Spec No. 2:** R215-223.

Wastie, R.L. 1991. Breeding for resistance. In *Advances in plant pathology* (eds. D.S. Ingram and P.H. Williams). Academic press limited, London.

Westerink, N., B.F. Brandwagt, P.J. de Wit, and M.H. Joosten. 2004. *Cladosporium fulvum* circumvents the second functional resistance gene homologue at the *Cf-4* locus (*Hcr9-4E*) by secretion of a stable avr4E isoform. *Mol. Microbiol.* **54:** 533-545.

Whisson, S.C., A. Drenth, D.J. Maclean, and J.A.G. Irwin. 1995. *Phytophthora sojae* avirulence genes, RAPD, and RFLP markers used to construct a detailed genetic linkage map. *Molecular Plant. Microbe Interact.* **8:** 988-995.

Whisson, S.C., T. van derLee, G.J. Bryan, R. Waugh, F. Govers, and P.R.J. Birch. 2001. Physical mapping across an avirulence locus of *Phytophthora infestans* using a highly representative, large-insert bacterial artificial chromosome library. *Mol. Genet. Genomics* **266:** 289-295.

Zhu, T., P. Budworth, B. Han, D. Brown, H.S. Chang, G.Z. Zou, and X. Wang. 2001. Toward elucidating the global gene expression patterns of developing Arabidopsis: Parallel analysis of 8 300 genes by a high-density oligonucleotide probe array. *Plant Physiol. Biochem.* **39:** 221-242.

Zwankhuizen, M.J., F. Govers, and J.C. Zadoks. 1998. Development of potato late blight epidemics: disease foci, disease gradients, and infection sources. *Phytopathology* **88**: 754-763.

Supplemental Research Data Table S1

Table S1. BACs containing *pi3.4*. Based on the hybridization patterns with various probes, the BACs are classified in six contigs I to VI. The probes used in the hybridization and the sizes of the *Bgl*II fragments hybridizing to TDF3.4 are listed in the column headings.

contig	BAC	tdf3.4 14 kb	tdf3.4 2.3 kb	tdf3.4 9.5 kb	tdf3.4 ~15 kb	tdf3.4 ~12 kb	tdf3.4 8 kb	tdf3.4 13 kb	tdf3.1	pi3.4F-5' end
I	BAC26D01	x								
I	BAC12I03	x								x
I	BAC37O23	x								x
I	BAC15E09	x								x
I	BAC29J11	x								x
I	BAC43G13	x								x
I	BAC22O12	x								x
I	BAC20I24		x						x	x
I	BAC34G01		x						x	
I	BAC40D03		x						x	
II	BAC09C18			x						
II	BAC27A18			x						
II	BAC45E08			x						
III	BAC03I21				x					
III	BAC18C08				x					
III	BAC13I12				x					
III	BAC22G09				x					
III	BAC27F09				x					
III	BAC40E02				x					
IV	BAC06M14					x				
IV	BAC28E04					x				
IV	BAC09H13					x				
IV	BAC28D13					x				
IV	BAC45G12					x				
IV	BAC28P02					x				
IV	BAC43M19					x				
V	BAC08E24						x			
V	BAC09F16						x			
V	BAC11F03						x			
V	BAC15O02						x			
V	BAC34B12						x			
VI	BAC05F22							x		
VI	BAC18J11							x		

[a]m7.1 is the cloned AFLP maker E+CA/M+GGs826

Supplemental Research Data Table S2

Table S2. Composition of pools used for hybridization of the custom designed *Phytophthora* GeneChip.

	strain	*Avr1*	*Avr3b-Avr10-Avr11*	*Avr4*	*Avr2*
pool1	rE11-16	Aa	aa	Aa	Aa
	T15-1	Aa	aa	Aa	Aa
	T30-2	Aa	aa	Aa	Aa
pool2	D12-2	aa	aa	aa	Aa
	D12-23	aa	aa	aa	Aa
pool3	D12-17	Aa	Aa	aa	aa
	T15-9	Aa	Aa	aa	aa
pool4	T20-2	aa	Aa	Aa	Aa
	E12-3	aa	Aa	Aa	aa
pool 5	T35-3	aa	aa	aa	aa
pool6	T30-2	Aa	aa	Aa	Aa

Supplemental Research Data Table S3

Table S3. *P. infestans* homologues of *P. ramorum* genes located in the *pr3.4* region.

protein in *P. ramorum*	**protein size (aa)**	**distance from *pr3.4***	***P. infestans* homologue**	**uni-gene code**	**BLAST *E* value**	**BLAST similarity**
pro80791	1091	3 kb	PB021B03 EST	NR007K03	2E-56	81%
pro80792	223					
pro80793	129	0.5 kb	PD027G09 EST	NR011A08	7E-60	92%
pro80794-1	196					
pr3.4	2000		*pi3.4F*		2E-89	93%
pro71831	234	3 kb	PB014G05 EST	NR006N11	2E-66	86%
pro80796	222					
pro80796-2	593					
pro80797	1348	8,3 kb	PF034D01 EST	NR015H15	1E-121	95%

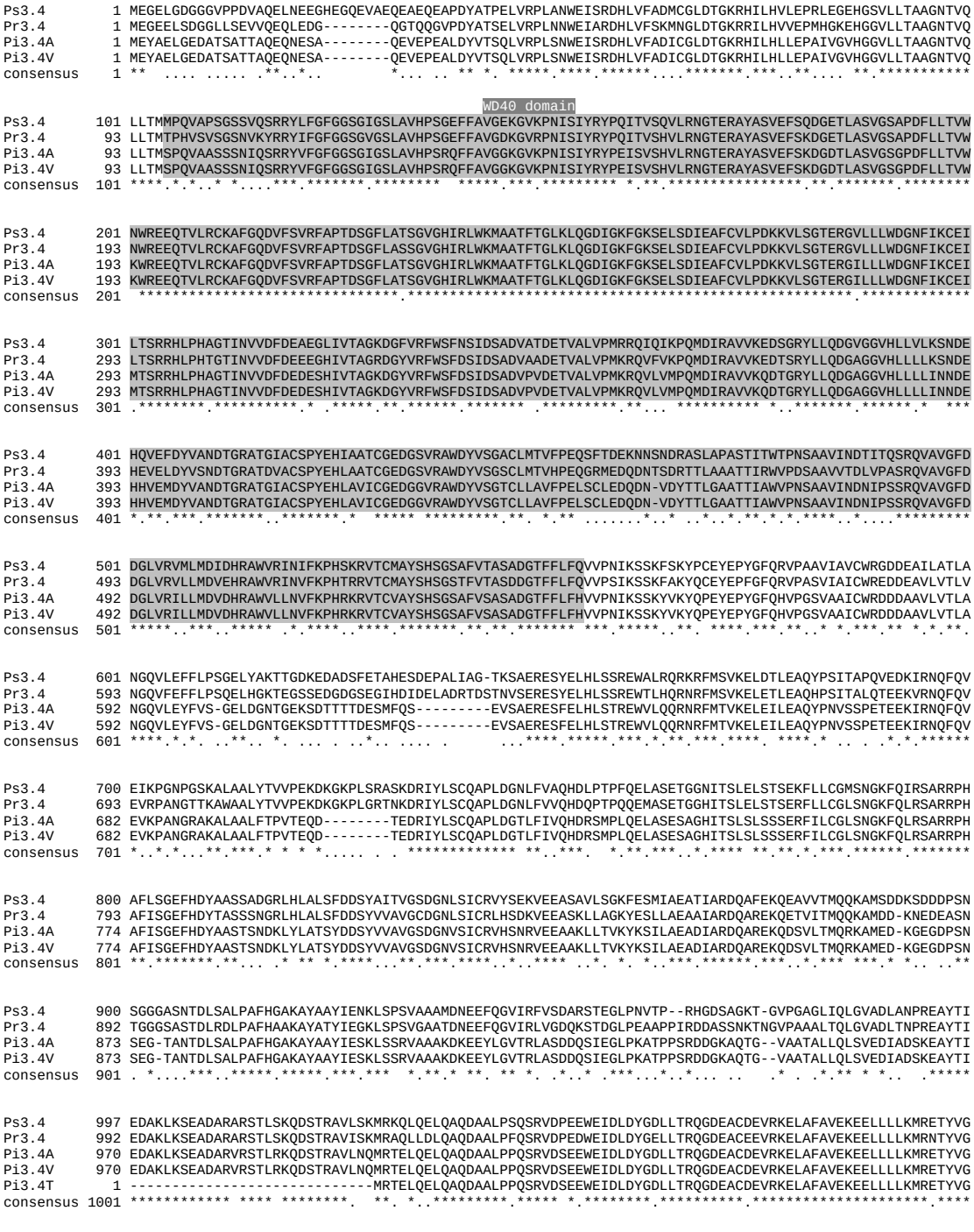
Supplemental Research Data Table S4

Table S4. Eight sequences representing six genes with CNV as shown by array-CGH.

code of the array sequence	size of the sequence (bp)	intensity increase corresponds to an AVR phenotype	SwissProt hit	*E* value	copy number in *P. sojae* [a]	copy number in *P. ramorum* [a]	array-CGH intensity change (from 0 to 1) [b]
PI001829 [c]	715		Fimbrin (P54680)	1E-50	1	1	0.61
PI004035 [c]	855		Fimbrin (P54680)	1E-70	1	1	0.67
PI007379	506		-	-	0	0	0.58
PI008646	612	AVR2	-	-	0	0	0.55
PI008871	483		-	-	0	0	0.53
PI009555	413		-	-	0	0	0.84
PI003481 [d]	472	AVR3b-AVR10-AVR11	-	-	1	1	0.59
PI014715 [d]	1358	AVR3b-AVR10-AVR11	-	-	1	1	0.63

[a] copy number is estimated from BLAST hit with identity percentage > 50% and *E* value < 1e-30
[b] intensity change indicates copy number variation, and the calculation of intensity change is described in material and methods.
[c] PI001829 and PI004035 are derived from the same gene
[d] PI003481and PI014715 are derived from *pi3.4*

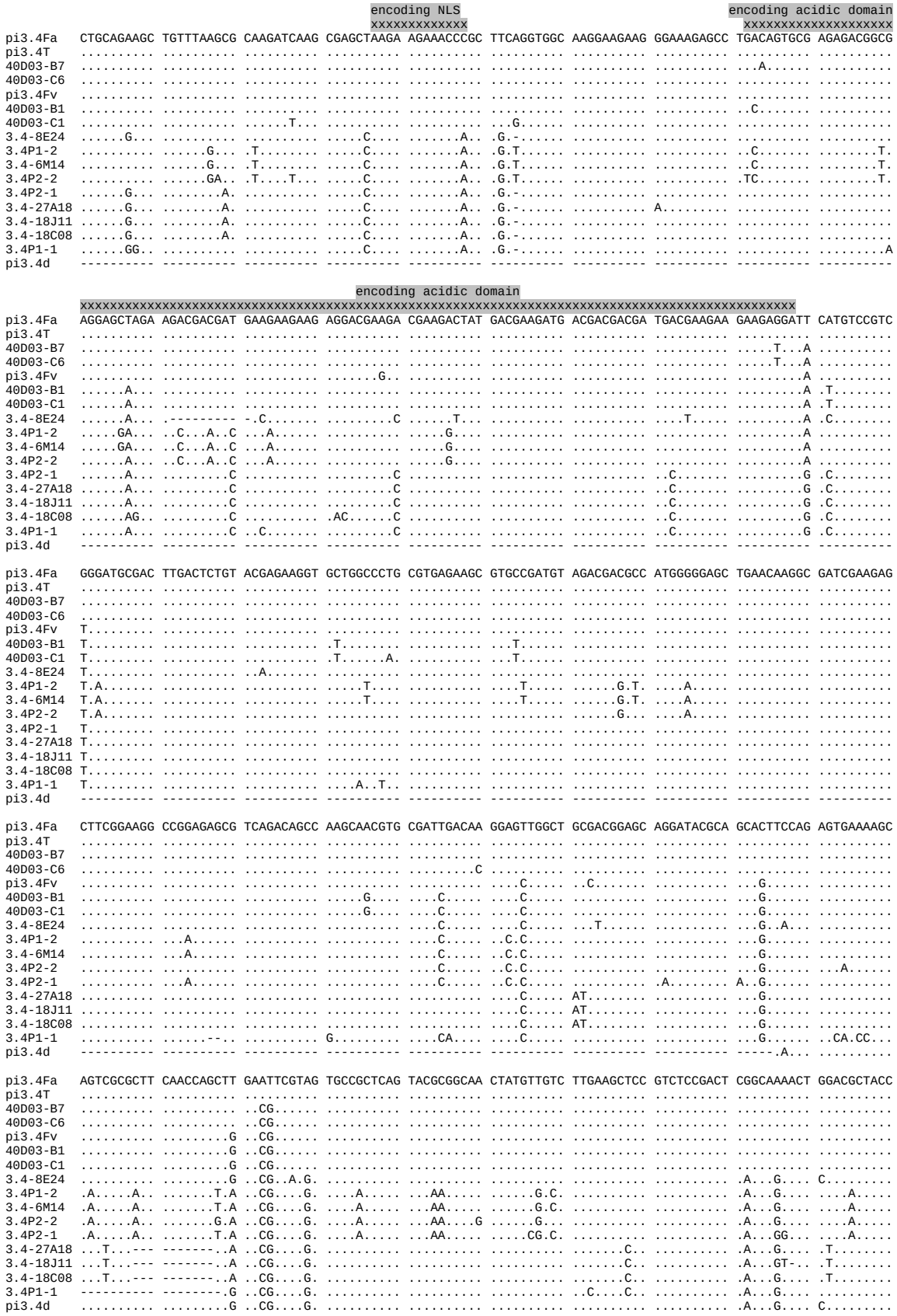
Supplemental Research Data Figure S1



```
Ps3.4     1196 GAGWSPAAQESGGSKSVHGFEARKRLRTERKERLARWRSNKPGEDADDPRDVVAIAFAQRNMGDYKLKTAADYVVPEHQRVNAVKKLRQMALLEESLYSA
Pr3.4     1191 AAGWTVVS-ETGSSKSVHGFEARKRLRTERKDRLTRWMNSKPGEDADDPRDVVAIAFAQRNMGDYKLKTAADYVVPEHQRVNAVKKLRQIALLEESLYAA
Pi3.4A    1169 ASSGPTAVQETSGSKSVHGFEARKQLRTERKERLARWQSNKPGEDADDPRDVVAIAFAQRNMGDYKLKTAPDYVVPEHQRVNAVKKLRQMALLEESLYAA
Pi3.4V    1169 ASSGPTAVQETSGSKSVHGFEARKQLRTERKERLARWQSNKPGEDADDPRDVVAIAFAQRNMGDYKLKTAPDYVVPEHQRVNAVKKLRQMALLEESLYAA
Pi3.4T     171 ASSGPTAVQETSGSKSVHGFEARKRLRTERKERLARWQSNKPGEDADDPRDVVAIAFAQRNMGDYKLKTAPDYVVPEHQRVNAAKKLRQMALLEESLYAA
consensus 1201 .   .....*. .*********** ******.**.**...****************************** *****************.********.*


Ps3.4     1296 RLRFNAKVLELRELKVQLINELRHDQERLVKLRATLAAGDDGSSAQPPPPPSGLSIDLREWPEQRERVSDGDIELFLKEKKVISHSQTQSADANVSMTRL
Pr3.4     1290 RLRFNAKVLELRELKVQLINELRRDQERLIELLVALAAKGSGGGSSAPPAPAGLAVDLREWPEQRERVSDADIELFLREKKVISHSQSQTAGSEISLTRT
Pi3.4A    1269 RLRFNAKVLELRELKVQLINELHRDQERLVELRAALASSG---ATPSSVPSKDLVVDLREWPEQRERVSDADIELFLREQKVISHVQTQTGEADLALTRI
Pi3.4V    1269 RLRFNAKVLELRELKVQLINELHRDQERLVELRAALASSG---ATPSSVPSKDLVVDLREWPEQRERVSDADIELFLREQKVISHVQTQTGEADLALTRI
Pi3.4T     271 RLRFNAKVLELRELKVQLINELHRDQERLVELRAALASSG---ATPSSVSSKDLVVDLREWPEQRERVSDADIELFLREQKVISHVQTQTGEAHLALTRI
consensus 1301 *********************..*****..*...** .. . ....  . . *..*************.******.* ***** *.*...... .**.


Ps3.4     1396 SGAENPHQTSDRSIAAVARALEGPNSTLTAGNEVLTGLTIGHLHLQMEDGAAVASAATVSAPATGSQALGSNGGNFDEDDVSVREYLMRQEALLLERKQR
Pr3.4     1390 GGAEKHHQSSDRSIAAVARAIEGPGSTVSTGSSASVGITIGHLHLQMDDGGAAATATAAPAASTGSSSLGLDGGNFDGDDVAVREYLMRQEALLLERKQR
Pi3.4A    1366 GGAEKRPQSADRSIAAVARALEG-----TSGKEAAKGVTIGHLHLQMDDG--TAAAATASSAHAG----ALNVENMDEDDVVVHEYLMRQEALLLERKQQ
Pi3.4V    1366 GGAEKRPQSADRSIAAVARALEG-----TSGKEAAKGVTIGHLHLQMDDS--TAAAATASSAHAG----ALNVGNMDEDDVVVHEYLMRQEALLLERKQQ
Pi3.4T     368 GGAEKRPQSADRSIAAVARALEG-----TSGKEAAKGVTIGHLHLQMDDG--TAAAATASSAHAG----ALNVENMDEDDVVVHEYLMRQEALLLERKQQ
consensus 1401 .***.. *. **********.**. .....*.....*.*********.*....*.*.... .. *.  .... .* *.***.*.**************


Ps3.4     1496 REIVAFDAAVAHLRREKLALEVGFKQAELRLFTLFSELSLLESFESKENLLSSKLEKSKSEKAQIVAELRDVQELLASKKRELDEWSRQERAVQGEFLTL
Pr3.4     1490 REVVAFDAAVAHLRRERLALGVGFKQAELRLFTLVSELSLLESFESKENLLSNKLEKSKGEKAQIMAELRDVQDLLASKKHELDEWSRQERSVQAEFLML
Pi3.4A    1455 REIASFDAAVAHLRHEKLALEVRFKQAELRLFTLLSELSLLESFESKENLLSSKLEKSKGEKAQIVADLRDVQDLLTIKKRDLDEWARQEKAVQAEFLTL
Pi3.4V    1455 REIASFDAAVAHLRHEKLALEVGFKQAELRLFTLLSELSLLESFESKENLLSSKLEKSKGEKAQIVADLRDVQDLLTIKKRDLDEWARQEKAVQAEFLTL
Pi3.4T     457 REIASFDAAVAHLRHEKLALEVGFKQAELRLFTLLSELSLLESFESKENLLSSKLEKSKGEKAQIVADLRDVQDLLTIKKRDLDEWARQEKAVQAEFLTL
consensus 1501 **.  *********.*.***.* ***********.*****************.******.*****.*.*****.**  **..**** ***..**.***.*

                                                 NLS                         Acidic domain
Ps3.4     1596 VNGGTGSSTTPHPSFSALQKLFKRKIKRAKKKPTASG-KDDGKDGNGQSPESARDGEEQEEEEEEEEDEDEDYDEDDDDDDEEEEDTCPPGCDLTLYEKV
Pr3.4     1590 VNGGTGSTATPHPSFAALQKLFKRKIKRAKKKPTASGGKEEGKDGNGQSPESARDGEEQEEEEEEEEDEDEDYDEDDDDDDEEEEDACPLGCDLTLYEKV
Pi3.4A    1555 VNGGTGSSATPHPSFAALQKLFKRKIKRAKKKP-ASGGKEEGKE-----PDSARDGEELEDDDEEEEDEDEDYDEDDDDDDEEEEDSCPSGCDLTLYEKV
Pi3.4V    1555 VNGGTGSSATPHPSFAALQKLFKRKIKRAKKKP-ASGGKEEGKE-----PDSARDGEELEDDDEEEEDEDEDYDEDDDDDDEEEEDTCPSGCDLTLYEKV
Pi3.4T     557 VNGGTGSSATPHPSFAALQKLFKRKIKRAKKKP-ASGGKEEGKE-----PDSARDGEELEDDDEEEEDEDEDYDEDDDDDDEEEEDSCPSGCDLTLYEKV
consensus 1601 *******..******.****************.***.*..**......*.******* *...********************** **.**********


Ps3.4     1695 LALREKRADVDDAVGELNKAIEELRKAGERQAAKQRAIDKELAATEQDTQQFQSEKQSRFNQLDVVVALSTRQLRCLESPSPNQLNWLLPAQAAGCLVFT
Pr3.4     1690 LALREKRADVDDAMGELNKAIEELRKAGDRQAAKQRAIDKELAATEQDTQQFQSEKQSRFNQLDVVVPLSTRQLRCLEPPSPNQQGWLLSAQAAGCLVFT
Pi3.4A    1649 LALREKRADVDDAMGELNKAIEELRKAGERQTAKQRAIDKELAATEQDTQHFQSEKQSRFNQLEFVVPLSTRQLCCLEAPSPTRQNWTLPAQADGCLVFT
Pi3.4V    1649 LALREKRADVDDAMGELNKAIEELRKAGERQTAKQRAIDKELAATEQDTQQFQSEKQSRFNQLDVVVPLSTRQLCCLEAPSPTRQNWTLPAQADGCLVFT
Pi3.4T     651 LALREKRADVDDAMGELNKAIEELRKAGERQTAKQRAIDKELAATEQDTQHFQSEKQSRFNQLEFVVPLSTRQLCCLEAPSPTRQNWTLPAQAAGCLVFT
consensus 1701 *************.*************.** ****************** ***********  **.****** ***.***   ..* *.*** ******
                                     Leu zipper
Ps3.4     1795 TQAFAALTERIESLQAENKQLRQQFRDLHKQQSVLAREKRLQQETIARSQERSEQLQRLKFGQLVDLEVLDRACDASAVDELQARVKLREVEGERTLRRV
Pr3.4     1790 TRAFSALTERIESLQAENKHLRQQFRDLHKQQSVLAREKRLQQESIARAQERSEQLQRLKFGQLVDLEVLDRACDASSMDELQTRVKLREVEGERALRRV
Pi3.4A    1749 TRAFSALSERIESLQAENKKLRQQFRDLHKQQNVLAKEKRQQQEAIARAQDRSEQLMRLKFGQLVDLEVLDRACDASSVDELQTRVKLREVEGERALRRA
Pi3.4V    1749 TRAFSALSERIESLQAENKKLRQQFRDLHKQQNVLAKEKRQQQEAIARAQDRSEQLMRLKFGQLVDLEVLDRACDASSVDELQTRVKLREVEGERALRRA
Pi3.4T     751 TRAFSALSERIESLQAENKKLRQQFRDLHKQQNVLAKEKRQQQEAIARAQDRSEQLMRLKFGQLVDLEVLDRACDASSVDELQTRVKLREVEGERALRRA
consensus 1801 *.**.**.**********.*********** ***.*** ***.***.*.***** *****************..****.**********.***


Ps3.4     1895 KQMHPQLQHAILQATEQNTALLAQIAALTERQAELERELNRSQGSGSGSGG--DGSGLLSLEDDAALAEREMKERNKLVRLVKLQAREVEALKQEIGLLR
Pr3.4     1890 KQTHPQFQHAILQATEQKTALLTQIAELTERQAELERELNHNQGGGSGSSGGRDGGGLLNMEDDAALAEREMKERNKLVRLVKLQAREVEALKQEIGLLR
Pi3.4A    1849 KQTHPQLQHAILQATEQNTALLAQIAALTERQAELERELNRHQGGGS-SSG--DSGGVLSLDDDEALVEREMKERNKLVRLVKLQAREVEALKQEIGLLR
Pi3.4V    1849 KQTHPQLQHAILQATEQNTALLAQIAALTERQAILERELNRHQGGGS-SSG--DSGGVLSLDDDEVLAEREMKERNKLVRLVKLQAREVEALKQEIGLLR
Pi3.4T     851 KQTHPQLQHAILQATEQNTALLAQIAALTERQAELERELNRHQGGGS-SSG--DSGGVLSLDEDEALVEREMKERNKLVRLVKLQAREVEALKQEIGLLR
consensus 1901 **.***.*********.****.***.****** ******..**.**.*.*  * .*.*...**  * ********************************


Ps3.4     1993 GKDGKVYAPRV
Pr3.4     1990 GKDGKVYAPRG
Pi3.4A    1946 GKDGKVYAPRG
Pi3.4V    1946 GKDGKVYAPRG
Pi3.4T     948 GKDGKVYAPRG
consensus 2001 **********.
```

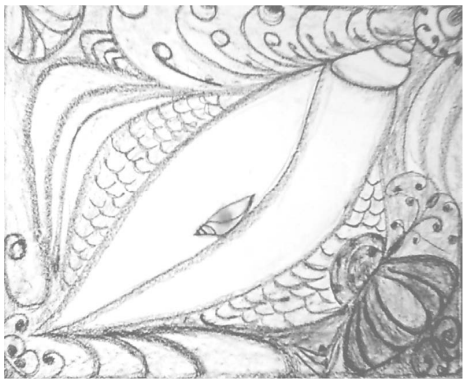
Figure S1. Sequence alignment of Ps3.4, Pr3.4, Pi3.4F^{A}, Pi3.4F^{V} and Pi3.4T. The polymorphisms in grey are between Pi3.4T and Pi3.4F (A and V), in black between Pi3.4F^{V} and both, Pi3.4F^{A} and Pi3.4T, and in red between Pi3.4F^{A} and both, Pi3.4F^{V} and Pi3.4T. The predicted protein domains are shaded.

Supplemental Research Data Figure S2



```
                                                                                    encoding Leu Zipper
                                                               xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
pi3.4Fa   TGCTCAAGCT GATGGCTGCC TCGTGTTCAC CACTCGCGCC TTCTCTGCGT TGTCTGAGCG TATCGAGTCG CTGCAAGCAG AGAACAAGAA GCTGCGTCAA
pi3.4T    .......... .C........ .......... .......... .......... .......... .......... .......... .......... ..........
40D03-B7  .......... .......... .......... .......... .A........ .......... .......... .......... .......... ..........
40D03-C6  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
pi3.4Fv   .......... .......... .......... .......... .......... ....C..... .........T .......... .......... ..........
40D03-B1  .......... .C........ .......... .......... .......... ....C..... .......... .......... .......... ..........
40D03-C1  .......... .C........ .......... .......... .......... ....C..... .......... .......... .......... ..........
3.4-8E24  .......... .C........ ........-- -------A.. A--------- ---....... .......... .......... .......... .......T..
3.4P1-2   ..T....... .C........ ........-- ------.... .......... ....C..... .......... .......T.. .......... ..........
3.4-6M14  ..T....... .C........ ........-- ------.... .......... ....C..... .......... .......T.. .......... ..........
3.4P2-2   ..T....... .C........ .......... .......... .......... ....C.G... .......... .......T.. .......... ..........
3.4P2-1   ..T....... .C........ ........-- ------.... .......... ....C..... ....A..... .......T.. .......... ..........
3.4-27A18 .......... .C........ .T........ .......... .......... .......... ...-...... .......... ....T..... ..........
3.4-18J11 .......... .C........ .......... .......... .......... .......... ...-.....C .......... ....T..... ..........
3.4-18C08 .......... .C........ ....C..... .......... .......... .......... ...-...... .......... ....T..... ..........
3.4P1-1   .......... .C........ .......... .......... .......... .......... .......... .......... ...C...... ..........
pi3.4d    .......... .C........ ........-- -------A.. A--------- ---....... C......... .......... .......... ..........

                      encoding Leu Zipper
          xxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxxx
pi3.4Fa   CAGTTCCGAG ACCTTCACAA ACAGCAGAAC GTGCTTGCTA AGGAGAAGCG TCAGCAGCAA GAAGCCATTG CAAGAGCTCA GGATCGCAGC GAGCAGCTCA
pi3.4T    .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
40D03-B7  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
40D03-C6  .......... .......... .......... .......... .......... .......... .......... .......... .......... ..........
pi3.4Fv   .......... ....C..... G......... .......... .......... ...A...... .......... .......... ...C...... ..........
40D03-B1  .......... ....C..... G......... .......... .......... ...A...... .......... .......... ...C...... ..........
40D03-C1  .......... ....C..... G......... .......... .......... ...A..A... .......... .......... ...C...... ..........
3.4-8E24  .......... ....C..... .......... .......... .......... .......... .......... ......A... ...C...... ..........
3.4P1-2   .......... .......... .......... .......... .......... ...A..A... ....T..... .......... ...C...G.. .....T....
3.4-6M14  .......... .......... .......... .......... .......... ...A..A... ....T..... .......... ...C...G.. .....T....
3.4P2-2   .......... .......... .......... .......... .......... ...A.TA... ....T..... .G........ ...C...G.. .....T....
3.4P2-1   .......... .......... .......... .......... .......... ...A..A... ....T..... .......... ...C...G.. .....T...G
3.4-27A18 .......... ....C..... G......... .......... T......... .......... .......... .......... ...C...... ..........
3.4-18J11 .......... ....C..... G......... .......... T......... .......... R.......G. .......... ...C...... ..........
3.4-18C08 .......... ....C..... G......... .......... T......... .......... .......... .......... ...C...... ..........
3.4P1-1   .......... ....C..... G......... .......... .......... ...A..A... ....T..... .......... ...C...G.. .....T....
pi3.4d    .......... ....C..... .......... .......... .......... .......... .......... ......A... ...C...... ..........

pi3.4Fa   TGCGACTGAA ATTCGGACAG
pi3.4T    .......... ..........
40D03-B7  .......... ..........
40D03-C6  .......... ..........
pi3.4Fv   .......... ..........
40D03-B1  .......... ..........
40D03-C1  .......... ..........
3.4-8E24  .......... ...T......
3.4P1-2   .......... ..........
3.4-6M14  .......... ..........
3.4P2-2   .......... ..........
3.4P2-1   .......... ..........
3.4-27A18 .......... ..........
3.4-18J11 .....A.... ..........
3.4-18C08 .......... ..........
3.4P1-1   .......... ..........
pi3.4d    .......... ...T......
```

Figure S2. Sequence alignment of fragments derived from different *pi3.4* copies with *pi3.4F^A^* (pi3.4Fa), *pi3.4F^V^* (pi3.4Fv) and *pi3.4T* (pi3.4T). The amplified copies from BAC40D03 are 3.4-40D03-B7, 3.4-40D03-C6, 3.4-40D03-B1 and 3.4-40D03-C1. The copies derived from contig-II to contig VI are 3.4-27A18, 3.4-18C08, 3.4-6M14, 3.4-8E24 and 3.4-18J11. The copies obtained by PCR on genomic DNA are 3.4P1-1, 3.4P1-2, 3.4P2-3, 3.4P2-2 and pi3.4d. In the alignment nucleotides substitutions are shown, identical nucleotides are denoted with '.' and deletions are marked with '-'.



Chapter 4

Elicitin genes in *Phytophthora infestans* are clustered and interspersed with various transposon-like elements

Molecular Genetics and Genomics 273(1): 20-32

Rays H.Y. Jiang, Angus L. Dawe, Rob Weide, Marjo van Staveren, Sander Peters, Donald L. Nuss and Francine Govers

Elicitin genes in *Phytophthora infestans* are clustered and interspersed with various transposon-like elements

Rays H.Y. Jiang[1], Angus L. Dawe[2], Rob Weide[1], Marjo van Staveren[3], Sander Peters[3], Donald L. Nuss[4] and Francine Govers[1 *]

[1] Plant Sciences Group, Laboratory of Phytopathology, Wageningen University, Binnenhaven 5, NL-6709 PD Wageningen and Graduate School Experimental Plant Sciences, The Netherlands
[2] Biology Department, New Mexico State University, 234 Foster Hall, Las Cruces, NM 88003, USA
[3] Greenomics, Plant Research International, P.O. Box 16, NL-6700 AA Wageningen, The Netherlands
[4] Center for Biosystems Research, University of Maryland Biotechnology Institute, College Park, Maryland 20742-4450, USA

[*] For correspondence: E-mail Francine.Govers@wur.nl; Tel. +31 317 483 138; Fax +31 317 483 412

The GenBank accession numbers for the sequences described in this paper are AY830090 to AY830111.

Keywords

Class I element – Class II element – Late blight – CHROMO domain

Abstract

Sequencing and annotation of a contiguous stretch of genomic DNA (112.3 kb) of the oomycete plant pathogen *Phytophthora infestans* revealed the order, spacing and genomic context of four members of the *inf* elicitin gene family. Analysis of the GC content at the third codon position (GC3) of six genes encoded in the region, and a set of randomly selected coding regions as well as random genomic regions, showed that high GC3 is a general feature of *Phytophthora* genes that can be exploited to optimize gene prediction programs for *Phytophthora* species. At least one-third of the annotated 112,3 kb *P. infestans* sequence consisted of transposons or transposon-like elements. The most prominent were four Tc3/gypsy and Tc1/copia type retrotransposons and three DNA transposons that belong to the Tc1/mariner, Pogo and PiggyBac groups, respectively. Comparative analysis of other available genomic sequences suggests that ubiquity and heterogeneity of transposable elements are common features of the *Phytophthora infestans* genome.

Introduction

Phytophthora infestans is the causal agent of potato late blight and one of the most devastating plant pathogens known today. *Phytophthora* belongs to the class oomycetes and the genus comprises over sixty species, all notorious pathogens of crop plants, trees and ornamentals (Erwin and Ribeiro 1996). Their growth morphology and dispersal strategy resemble that of fungi and the weaponry that oomycetes and fungi use to attack plants appears to be comparable (Latijnhouwers et al. 2003). In the eukaryotic phylogenetic tree, however, oomycetes are classified as heterokonts together with brown algae and diatoms, and positioned on a branch completely separate from fungi (Baldauf 2003). The distinct phylogenetic position of oomycetes and fungi is manifested in, amongst others, differences in intracellular structures, in cell wall composition, in physiological and biochemical processes and in ploidy level (Erwin and Ribeiro 1996; Kamoun 2003).

It is very likely that evolutionary history also shaped the genes and genomes of oomycetes. Within oomycetes, *Phytophthora* is the most extensively studied genus and data on *Phytophthora* genes, gene structure, gene expression, and repeat elements are steadily accumulating (Kamoun 2003). The EST databases of *P. infestans* (Kamoun et al. 1999; Randall et al. accepted for publication) and *Phytophthora sojae* (Qutob et al. 2000) are the most advanced (https://xgi.ncgr.org/spc; http://www.pfgd.org) and various other smaller scale EST projects are ongoing (e.g. in *Phytophthora nicotianae*; Skalamera et al. 2004). In addition, the genomes of two *Phytophthora* species have been sequenced, *P. sojae* to 9X coverage (genome.jgi-psf.org/sojae1/sojae1.home.html) and *Phytophthora ramorum* to 7X coverage (genome.jgi-psf.org/ramorum1/ramorum1.home.html). *Phytophthora* spp. have flexible genome sizes which generally exceed those of fungi and other microorganisms; the *P. ramorum* genome is 65 Mb whereas the *P. infestans* genome is estimated to be 240 Mb (Kamoun 2003).

To fully explore the available genome and EST databases requires annotation tools and gene prediction programs specifically trained for *Phytophthora.* A first step is the detailed analysis and annotation of relatively small genomic regions. Discovery of typical features of coding regions, patterns of repeat distribution, and diversity and number of transposable elements is instrumental for the design of automated gene prediction programs to be used to scan whole genomes. Here, we present the gene annotation of a 112,3 kb *P. infestans* bacterial artifical chromosome (BAC) sequence. This BAC was selected from a physical contig that spans a number of elicitin genes.

Elicitins belong to a particular class of extracellular proteins produced by *Phytophthora* species. They were first characterized on the basis of their ability to induce defense responses in plants, in particular in *Nicotiana* species, and are thought to act as species-specific avirulence factors, and thus, as determinants of the host range for selected plant-*Phytophthora* interactions (Ricci et al. 1992; Kamoun et al. 1998). Elicitins can act as a sterol-carrier (Mikes et al. 1998), a biological function that seems to be

essential, since *Phytophthora* itself cannot synthesize sterols and must retrieve them from external sources (Hendrix and Guttman 1970).

In *P. infestans,* elicitins are encoded by a complex multigene family (Kamoun et al. 1999). All *inf* elicitin genes encode putative extracellular proteins that share the 98 amino-acid elicitin domain corresponding to the mature canonical INF1 protein, the most abundant extracellular protein. Five *inf* genes encode proteins with an extended C-terminal domain (Kamoun et al. 1999). Preliminary data based on genomic Southern blot hybridizations suggested that several members of the *inf* elicitin gene family are clustered in the genome. Studies in two other *Phytophthora* species also demonstrated clustering of elicitin genes. In both, *Phytophthora cinnamomi* and *Phytophthora cryptogea,* a 6 kb genomic region spans four elicitin genes (Panabieres et al. 1995; Duclos et al. 1998). In *P. infestans* the *inf1* gene is highly expressed in mycelium but not in sporangia and cysts (Kamoun et al. 1997). Most other *inf* genes have a similar expression pattern but the expression levels vary. In an EST library from mycelium, *inf5* and *inf6* were among the four most abundant cDNA clusters suggesting that these *inf* genes are highly active (Kamoun et al. 1999).

The aim of this study was (i) to sequence and annotate a long stretch of genomic DNA of *P. infestans,* (ii) to investigate the order and spacing of the *inf* elicitin genes located on this stretch, and (iii) to examine the genome context of the elicitin gene cluster. Annotation of the sequence revealed many repeats and showed that the elicitin gene cluster is interspersed with transposons and transposon-like elements representing various classes and groups. Analysis of GC percentage and codon usage of coding sequences showed particular characteristic features of *Phytophthora* genes that will be instrumental for gene prediction.

Material and methods

BAC library screening

The *P. infestans* BAC library used for screening is described by Whisson et al. (2001). Screening was done by colony hybridization according to standard procedures and ^{32}P-labelled probes were prepared by the random hexamer method with a random primer labelling kit (Prime-a-Gene®, Gibco-BRL). The *inf* elicitin probes were prepared from EST clones from the *P. infestans* MY EST library described by Kamoun et al. (1999). Alkaline lysis was used to isolate plasmid DNA and the insert was released by digestion with *Eco*RI and *Bam*HI and subsequently purified from the gel. The *inf1* probe was derived from EST clone MY18D10, *inf2A* from MY05C05, *inf2B* from MY02D01, *inf3* from MY19C07, *inf4* from MY11E04, *inf5* from MY01C05 and *inf6* from MY01D04. BACs hybridizing to the *inf* probes were picked and grown in LB containing 12,5 µg/ml chloramphenicol. BAC DNA was isolated by alkaline lysis and digested with restriction enzymes. Fragments were size separated on agarose gels by electrophoresis

and transferred to Hybond N^+ membranes. Hybridization with the individual *inf* probes was performed to confirm the *inf* genes to be located on the BACs.

BAC fingerprinting, contig building and insert size determination

To obtain BAC fingerprint patterns, 1 µg of BAC DNA was digested with 10 U *Hin*dIII in 100 µl reaction buffer for 4 hours at 37 ^{o}C, the digestion products were precipitated by iso-propanol and dissolved in 10 µl TE for gel electrophoresis. For contig building, fragments from different BACs but sharing identical length were considered as common fragments. To determine the BAC insert sizes, BAC DNA was digested with the rare cutting enzyme *Not*I that separates the vector pBeloBAC11 from the insert. The digested DNA was analysed on CHEF (Contour-clamped Homogeneous Electric Field) gels using a CHEF-DR II Pulse Field Gel Electrophoresis Apparatus from BIO-RAD II. The CHEF gels consisted of 1% agarose in 0.5 x TBE, and electrophoresis was performed with 5-15 s switch time (linear ramping), at 13 ^{0}C running temperature in 5 x TBE buffer with 220 V power supply (voltage constant) for 18 hours.

Shotgun cloning, sequencing and sequence assembly

BAC DNA was purified by Plasmid-Safe™ ATP-Dependent DNase (Epicenter) to remove contaminating *E. coli* genomic DNA, subsequently sheared, fractionated and cloned into TOPO® vector as described by the manufacture (TOPO® Shotgun Subcloning Kit Invitrogene). Average insert sizes of the shotgun clones were between 2.5 and 3 kb. Plasmids were manually prepared from cultures of stored colonies using the Qiaprep 8 system (Qiagen) with a vacuum manifold. Each preparation was checked for yield by electrophoresis, and then submitted to the DNA Sequencing Core Facility at the Center for Biosystems Research for analysis on ABI 3100 or 377 (Applied Biosystems) machines. Dual (5' and 3') sequence reads of the cloned fragments were obtained using the M13 forward and reverse primers. Results files were scanned using the SeqMan unit of DNAStar (DNAStar, Inc. Madison, WI) running on a Macintosh G4 computer to monitor the presence of known vector sequence and quality of data prior to further analyses. Shotgun sequences were base-called by the PHRED basecaller and assembled using the Gap4 assembler of the Staden2003 package. Using the PREGAP4 interface, GAP4-assembled sequences were parsed into the GAP4 assembly database (Bonfield et al. 1995). The GAP4 interface and its features were then used for editing and sequence finishing. Consensus calculations with a quality cut-off value of 40 were performed from within GAP4 using a probabilistic consensus algorithm based on expected error rates output by PHRED. By sequencing PCR products using custom designed primers and bridging the ends of contiguous fragments the remaining sequence gaps were closed. Most of the gap-closure sequencing was performed at Greenomics, PRI, using the ABI PRISM Big Dye terminator Cycle Sequencing Ready reaction kit with FS AmpliTaq DNA polymerase (Perkin Elmer) and analysed on an ABI 3730XL DNA Analyser. To verify the assembly, read-pairs were analysed on direction and size using a maximum size spacing of 2.5 kb.

Programs for sequence annotation

Sequences were analyzed in Vector NTI 8. For BLAST searches we used the NCBI BLAST program and the Standalone-BLAST Version 2.2.3 (Altschul et al. 1990). Repeat analysis was done with PiPMaker (Schwartz et al. 2000) and for multiple sequence alignment ClustalX 1.0 was used (Jeanmougin et al. 1998). Phylogenetic tree construction was performed by Molecular Evolutionary Genetic Analysis2.1 (MEGA) (Kumar et al. 1994). Calculation scripts were written in Python2.2 (http://www.python.org).

Genome databases and EST databases

P. infestans and *P. sojae* EST databases are accessible at https://xgi.ncgr.org/spc (Syngenta *Phytophthora* Consortium (SPC) EST sequence databases) (Randall et al. accepted for publication) and http://www.pfgd.org (*Phytophthora* Functional Genomics Database) (previously the *Phytophthora* Genome Consortium database (https://xgi.ncgr.org/pgc)) (Kamoun et al. 1999; Qutob et al. 2000). The sequences from *Fusarium graminearum* and *Aspergillus nidulans* were downloaded from the Broad institute website (http://www.broad.mit.edu/annotation) and the *Blumeria graminis* EST database was downloaded from the Phytopathogenic Fungi and Oomycete EST Database Version 1.4 (Soanes et al. 2002). Random genomic sequences of *P. infestans* and *P. sojae*, produced by the Broad Institute, Cambridge MA, USA (*P. infestans*) and the DOE Joint Genome Institute, Walnut Creek, CA, USA (*P. sojae*), were retrieved from the NCBI trace file archive (http://www.ncbi.nlm.nih.gov/Traces).

Results and discussion

Clustering of elicitin genes on the *P. infestans* genome

From a BAC library that contains 10-fold coverage of the *P. infestans* genome (Whisson et al. 2001) one-third of the BAC colonies was screened with cDNA clones representing seven members of the *inf* elicitin gene family, *inf1*, *inf2A*, *inf2B*, *inf3*, *inf4*, *inf5* and *inf6*. Genomic Southern blot analysis had shown previously that all seven are single copy genes. At least seven BACs hybridized to two or more of the *inf* probes suggesting clustering of the *inf* elicitin genes in the genome. Insert sizes were determined by analyzing *Not*I digests on CHEF gels. By fingerprinting *Hin*dIII digests and hybridization of fingerprint blots with the *inf* probes one physical contig spanning the seven *inf* elicitin genes was obtained (Fig. 1). BAC11A5 carrying four elicitin genes including the canonical *inf1* gene was chosen for further analysis.

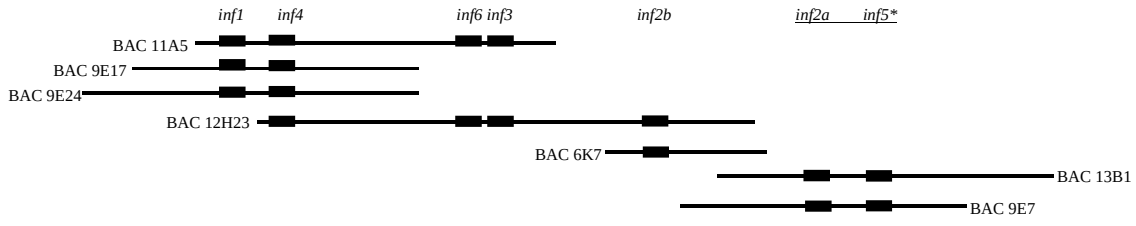


Fig. 1 A *P. infestans* BAC contig of approx. 250 kb containing seven *inf* elicitin genes. Black boxes represent the *inf* genes. * The order of *inf2A* and *inf5* has not been determined. BAC insert sizes are: BAC6K7-45kb, BAC9E7-50 kb, BAC11A5-130 kb, BAC9E24-150 kb, BAC12H23-120 kb, BAC13B1-120 kb.

Assembly and annotation of BAC clone 11A5

BAC11A5 was sequenced using a combination of shotgun and directed sequencing. A total number of 1031 sequence reads was assembled and edited using PHRED to yield one continuous 112.3 kb sequence contig with an average of 5.38-fold coverage (GenBank Accession number AY830090). Annotation of the assembled sequence revealed the four expected elicitin genes, two putative genes and a large number of transposon-like sequences (Fig. 2, Table 1). One of the two putative genes, named *ORF457*, has an open reading frame (ORF) of 1374 bp with no homology to known genes. Its annotation however, is supported by a perfect match (*E* value 0) with EST contig CON_001_15864 compiled of one EST sequence present in a library of mating cultures (Randall et al. accepted for publication). The other putative gene was designated *kre6-like* because its deduced amino acid sequence has homology to the yeast protein KRE6, a beta-glucan synthesis-associated protein (P32486). However, *kre6-like* seems to be a pseudogene with two frame shift mutations that disrupt the ORF.

Table 1 Genes present on BAC11A5 and the GC percentage in each reading frame. The highest percentage is in bold and underlined. The six genes and ORFs are ordered according to their position in the 112.3 kb sequence contig.

genes [a]	frame	protein length (aa)	ORF location (bp)	GC %			number of EST hits	TI distance (bp) [d]	transcriptional initiation site consensus [e] TNSCAWTCTSCAATTTGCW
				frame 1	frame 2	frame 3			
inf1	-1	118	5505-5862	43.70	54.24	**83.90**	318	37	TTCCATTGTGCAATTTGCT
inf4	-2	118	51885-52242	40.34	48.31	**67.80**	13	38	CCTCATTCCGCAATTTCCA
kre6-like [b]	-3	475	60681-62112	45.38	49.89	**50.74**	none	94	TAGCCCACTCTAATTTTCG
ORF457	-1	457	91284-92658	55.90	44.64	**56.24**	1	16	TTCACAGCTCAAACTTGTG
inf3 [c]	-2	188	92880-93447	40.21	68.09	**72.34**	4	30	TCTCACTCTGCAATCTGCT
inf6	2	183	95517-96069	59.24	59.56	**65.57**	352	44	TGCCATTCTCCAATTTGCT

[a] GenBank accession numbers AY830094 (*inf1*), AY830095 (*inf4*), AY830093 (*kre6-like*), AY830097 (*ORF457*), AY830092 (*inf3*), AY830096 (*inf6*)
[b] pseudogene; length is calculated after frameshift correction
[c] pseudogene; length is calculated after frameshift correction and by adding a start codon
[d] distance from the predicted transcription initiation site to the start codon
[e] as described by Pieterse et al. (1994) and McLeod et al. (2004); the residues that match the consensus are shaded

The order of the four elicitin genes on the sequence contig is *inf1, inf6, inf3* and *inf4* and this is in agreement with the order deduced from the physical BAC contig (Fig. 1). All four *inf* genes have perfect EST hits (*E* values 0 with the *P. infestans inf* ESTs mentioned in material and methods). Nevertheless, like *kre6-like*, *inf3* seems to be a pseudogene: it has one frameshift mutation and lacks a start codon. In the *P. infestans* EST database consisting of approx. 75,000 ESTs (Randall et al. accepted for publication) there are only four *inf3* ESTs suggesting that *inf3* is expressed at a lower level than the other six elicitin genes in the contig with 13-352 transcripts. The intergenic region between *inf3* and *ORF457* is rather small (221 bp from the TAA stop codon of *inf3* to the ATG start codon of *ORF457*) and *inf3* and *ORF457* are located in the most gene dense region of the BAC, together with *inf6*. *ORF457* is

29.2 kb away from its other neighbouring gene, *kre6-like* (Fig. 2). The presence of such a gene island illustrates the uneven gene density in the *P. infestans* genome and explains the discrepancy between a projected genomic coding capacity of 13,000 genes based on the overall gene density on BAC11A5 (this study) and the recently predicted 18,000 unigenes based on EST analysis (Randall et al. accepted for publication). Surveys of other genomic regions in *P. infestans* and of the assembled *P. sojae* and *P. ramorum* genome sequences (accessible via www.jgi.doe.gov/) revealed that gene islands are very common in *Phytophthora* species (data not shown).

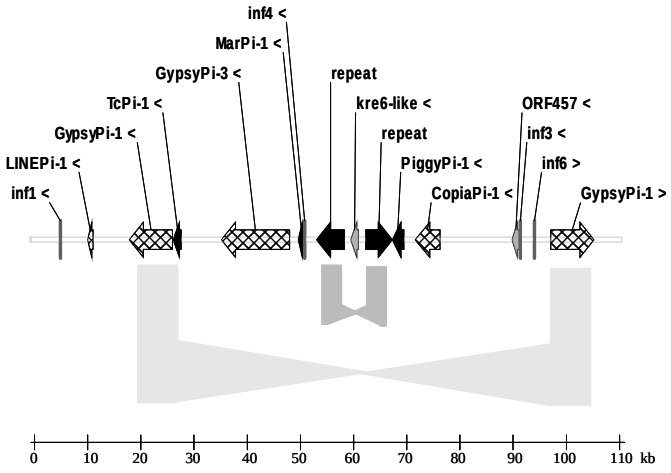


Fig. 2 Genes and mobile elements present on BAC11A5. On this schematic drawing of the 112.3 kb sequence contig (see scale bar) all elements are shown proportionally according to their sizes. The symbols > and < following the codes indicate the orientation of the genes and mobile elements similar to the arrow blocks drawn on the contig itself. The 8.3 kb and 5.3 kb repeats are connected with light and dark gray block arrows, respectively.

The four elicitin genes and the two putative genes all have the 19 nucleotide core promoter consensus sequence that spans the transcription start site in several oomycete genes (Table 1) (McLeod et al., 2004). Neither the elicitin genes nor the two putative genes have introns, which agrees with the observation that the majority of identified *Phytophthora* genes lack introns (Kamoun 2003).

High GC content at the third codon position as a general feature of coding sequences in *Phytophthora*

Phytophthora genes are associated with a high GC content (Qutob et al. 2000; Hraber and Weller 2001). Due to the redundant nature of the codons, the GC rich feature is expected to be more pronounced at the third position of a codon (defined as GC3) (Kamoun and Styer, 2000). The coding regions of the six genes identified in the BAC11A5 sequence contig have an average GC content of 55.67% and this is slightly higher than the overall average GC content of 51.60 % of the 112.3 kb sequence contig. In contrast, the GC3 in the coding regions has an average of 66.10% and in all six genes GC3 is higher

than GC1 and GC2 (Table 1). When the GC content of the BAC11A5 sequence contig, calculated from a sliding window of 300 bp, is plotted against the position, a high GC3 in a coding region is visualized as a 'GC peak' against the average GC content of 51.60%. This is shown in Fig. 3 with a GC plot of the first 12.5 kb of the sequence contig containing *inf1*. The *inf1* ORF resides in frame –1, thus the third position of the codon is in GC frame +1/-3, which gives a GC content peak that precisely correlates with the position of the ORF of *inf1* (position 5507-5864). Scanning the remaining 100 kb of the sequence in a similar way revealed GC peaks at the positions of the *inf3, inf4, inf6* and *kre6-like* genes. In addition, a number of other distinctive peaks were found that correspond to several transposon and retrotransposon-like sequences such as *MarPi-1, GypsyPi-1, GypsyPi-2* and *CopiaPi-1* as described below.

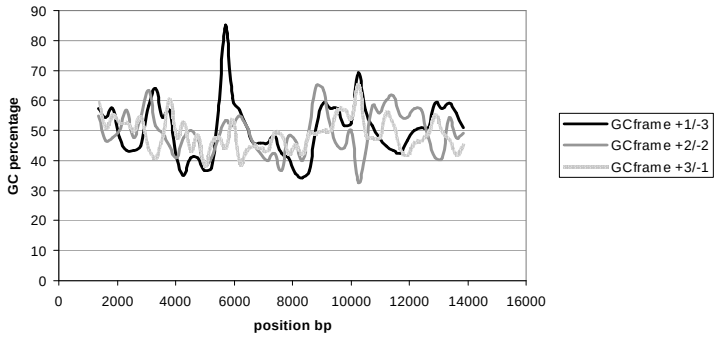


Fig. 3 GC content distribution of 12.5 kb sequence of BAC11A5 in three frames. The GC percentage is plotted against the position on the sequence contig. The GC calculation was done with a sliding window size of 300 bp. The arrow indicates the position of *inf1*. The *inf1* ORF resides in frame -1

To investigate if a high GC3 is a general feature of *P. infestans* genes, we retrieved the sequence of 79 full-length *P. infestans* genes from GenBank and calculated the GC3. They represent a heterogeneous set of genes involved in various biological processes such as metabolism, cellular organization, energy, signal transduction and plant-pathogen interaction. The average GC content of the coding sequences of the genes is 58.15% whereas the average GC3 is 73.05%. With three exceptions the highest GC content is always at the third position of the codons (Fig. 4). Even some genes that have an average GC content of less than 52% have a significantly high GC3 (> 67%), e.g., the G-protein α subunit gene *pigpa1* (AY050536), a microtubal binding protein gene (gi23394381) and the elicitin gene *inf4* (Table 1). The exception is the *ipiB* gene family, a cluster of three linked genes located on a 5.4 kb genomic fragment with very short intergenic regions (Pieterse et al., 1994). The predicted IPI-B proteins have a high glycine content (36.82%), with 51.64% of the glycine residues encoded by GGT and 18.85% by GGA resulting in a high average GC content of ~66%. Hence, GC3 in *ipiB* is lower (~48%) than GC1 (~78%) and GC2 (~71%).

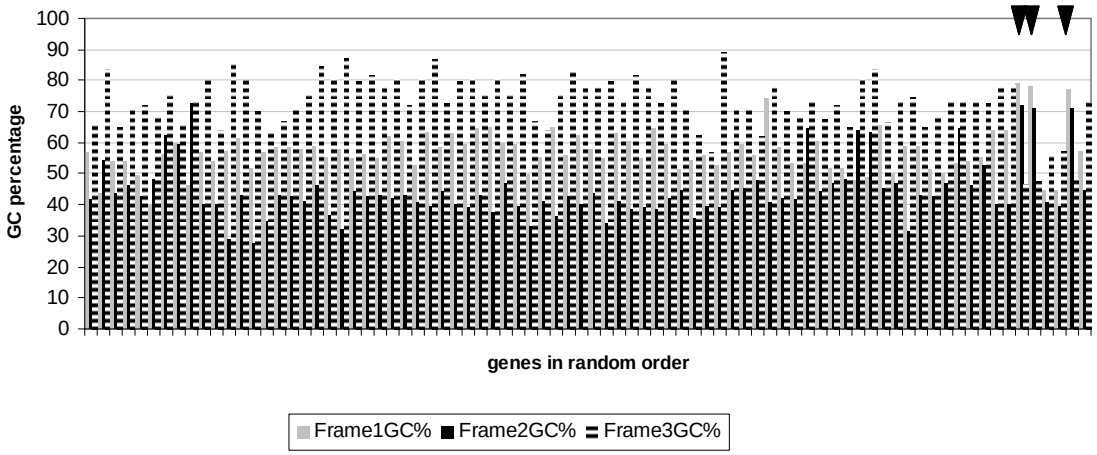


Fig. 4 Visualization of the GC content in three reading frames of 79 *P. infestans* genes retrieved from GenBank. For each gene the GC percentage of the three reading frames is plotted in three bars within one column. The black arrowheads point at the columns representing the three *ipiB* genes.

To evaluate the value of the high GC3 for gene prediction in *Phytophthora* we compared the GC3 in random genomic sequences with that in EST sequences. From the NCBI trace file archive we retrieved a thousand randomly selected genome sequences with of average size of 600 bp from two *Phytophthora* species, *P. infestans* and *P. sojae.* GC3 was calculated from randomly selected frames. For each of the two species we also retrieved 1000 randomly selected good quality EST contigs from EST databases. From these EST sequences GC3 was calculated from the putative ORFs. As shown in Fig. 5, the four sets of sequences give three distinctive peaks, the majority of the genomic sequences of both species have a GC3 lower than 55% whereas in the *P. infestans* EST sequences the GC3 in most cases exceeds 60% and in *P. sojae* ESTs even 80%. To compare this with the situation in other organisms, a similar calculation was performed on 1000 randomly selected ORFs and genome sequences from two fungi, *Fusarium graminearum* and *Aspergillus nidulans*. In both fungi, the GC3 peak of the ORFs is only slightly higher (around 55%) than the GC3 peak of the genome sequences (around 50%).

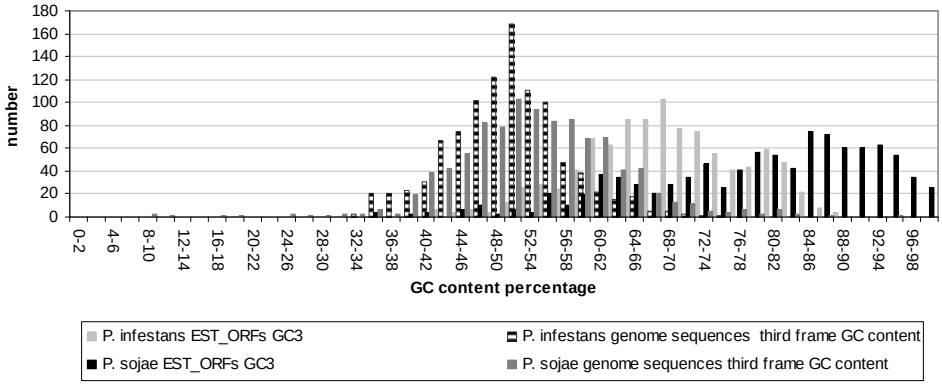


Fig. 5 GC3 of 1000 randomly selected EST ORFs and 1000 randomly selected genomic sequences from *P. infestans* and *P. sojae.* The Y-axis shows the number of EST ORFs or genomic fragments that have the GC3 percentage indicated on the X-axis.

Each species systematically uses certain synonymous codons in coding sequences. A biased usage of GC rich codons is expected in *Phytophthora* genes. To investigate the relationship between GC3 and codon usage, codon usage was calculated for 79 *P. infestans* genes (Supplemental data, Table S1). Except for the stop codon with a preference for 'TAA', and arginine with a slight preference for codons ending with A or T, *P. infestans* prefers to use codons with a G or C at the third position. Codon usage analysis on ORFs derived from 1000 *P. infestans* ESTs and 1000 *P. sojae* ESTs gave similar results: in both species there is a clear preference for codons with a G or C at the third position including the codons for arginine.

The unequal GC distribution combined with a high GC3 in coding regions is a fascinating feature of *Phytophthora* genomes. The finding that GC3 peaks in GC genome scans can reveal the position of genes (as shown for *inf1* in Fig. 3) justifies exploitation of this feature as a gene discovery tool for *Phytophthora.*

Repeat distribution in BAC11A5

A dot plot analysis of the BAC11A5 sequence revealed repeats of different lengths and in different orientations (data not shown). Most prominent were two large repeat units of 8.2 kb and 5.3 kb, respectively (Fig. 2). The 8.2 kb repeat unit is a retrotransposon of the Ty3-Gypsy family that we named *GypsyPi-1.* The two *GypsyPi-1* elements, which have 99% similarity, are in inverted orientations and located 71.8 kb apart. Since each *GypsyPi-1* element has a pair of long terminal repeats (LTRs), there are four copies of the 345 bp LTR in the sequence contig. In the other large repeat unit of 5.3 kb no transposon-related elements could be identified. The 5.3 kb repeats are also in inverted orientation and interspersed by a 4.1 kb fragment that contains the *kre6-like* pseudogene.

In addition to the larger repeats the sequence contig contains a number of smaller tandem repeats and inverted repeats. One 33 bp fragment is repeated twice in tandem with 100% similarity and five inverted repeats ranging in length from 35 bp to 76 bp have similarity levels of 86% to 100%. Interestingly, three out of the five inverted repeats are located in promoter regions. A 35 bp perfect inverted repeat separated by a one nucleotide spacer is located 508 bp upstream of the *inf1* ORF, and similarly, a 45 bp perfect inverted repeat, also with a one nucleotide spacer, is located 675 bp upstream of the *inf4* ORF. Also the promoter region of the *kre6-like* pseudogene contains an inverted repeat. It is located 900 bp upstream of the ORF, 76 bp in length and the similarity between the inverted repeats is 87%. It is conceivable that these palindrome like sequences have functions in regulating gene expression but this remains to be determined.

BAC11A5 has three different Class II transposons

Transposable elements are mobile DNA sequences that can move from one genomic location to the other. They are classified in two groups according to their transposition intermediate. Class I elements

transpose via RNA and reverse transcriptase is needed to convert the RNA intermediate into DNA for new transposition. Class II elements transpose directly as DNA molecules and no RNA intermediate is needed (Feschotte et al. 2002). They are characterized by Terminal Inverted Repeats (TIRs), and by a transposase gene located between the TIR borders.

Distributed in the BAC11A5 sequence contig are three Class II elements. They range in size from 1,5 to 2.5 kb and take up a total of 5.7% of the BAC sequence (Fig. 3, Table 2). Their transposase sequences are highly diverged and based on BLAST homology and phylogenetic analysis these three transposons belong to three different families, i.e., Tc1/mariner, Pogo and PiggyBac. The Tc1/mariner type of transposons (Plasterk 1996) together with Pogo transposons (Tudor et al. 1992; Smit and Riggs 1996) are probably the most widely spread Class II elements. Fig. 6 shows a phylogenetic tree constructed from the transposase regions of various characterized Tc1/mariner and Pogo elements and two of the three Class II elements identified in BAC11A5, named *MarPi-1* and *Tc1Pi-1*, respectively. *MarPi-1* groups with the mariner like transposons (Fig. 6) and the 200 amino acid transposase has the highest BLASTP homology with a putative rice transposase (AC093017_21) (*E* value 2e-08). TIRs expected to flank the *MarPi-1* transposase cannot be found. *Tc1Pi-1* is a Tc1 like transposon (Figure 6). The 375 amino acid transposase obtained, after correcting for a frame shift mutation, shows the highest BLASTP homology to a putative transposase from *Anopheles gambiae* (XP_310448) (*E* value 1e-08). The TIRs of *Tc1Pi-1* are inverted repeats of 120 bp flanking the transposase that show 61.3% similarity. The third Class II element found in BAC11A5 belongs to PiggyBac family and was named *PiggyPi-1*. Members of this family of transposases are mainly found in animals and are related to the transposase of the canonical piggyBac transposon from the moth *Trichoplusia ni* (Sarkar et al. 2003). They have no obvious homology to other transposon families. The deduced 724 amino acid transposase of *PiggyPi-1* shows the highest BLASTP homology to the piggyBac transposable element of *Homo sapiens* (NP 689808.2) (*E* value 9e-15). As with *MarPi-1* no TIRs can be detected.

Table 2 Characteristics of three different Class II elements present on BAC11A5. The GC content calculation was performed on the transposase coding regions. The highest percentage is in bold and underlined.

transposon [a]	group	size (kb)	TIR (bp)	copies in BAC11A5	TIR pair similarity %	GC%			number of EST hits [b]	homologue in *P. sojae* [c]	homologue in *P. ramorum* [c]
						frame 1	frame 2	frame 3			
MarPi-1	mariner	~1.5		1		58.00	44.72	**58.29**	36	no	no
Tc1Pi-1	Tc1	1.5	120	1	63.1	56.68	43.09	**63.82**	0	no	no
PiggyPi-1	piggy	~2.5		1		**50.00**	44.40	43.02	0	no	no

[a] GenBank accession numbers AY830109 (*MarPi-1*), AY830110 (*Tc1Pi-1*), AY830111 (*PiggyPi-1*)
[b] based on BLASTN search against 72,904 *P. infestans* ESTs; *E* value less than e^{-100} and percentage identity higher than 95%
[c] based on BLASTN search against *P. sojae* and *P. ramorum* genome sequences; *E* value less than e^{-100} and percentage identity higher than 80%

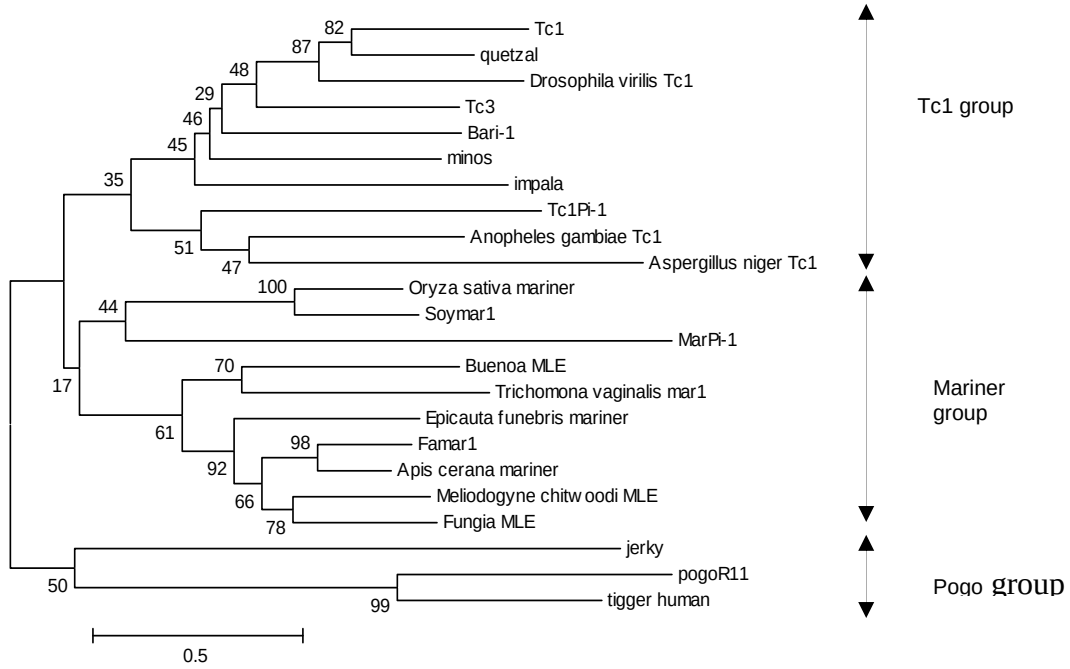


Fig. 6 Phylogram of Class II elements belonging to the Tc1/mariner and Pogo groups. The transposase protein sequences including the conserved DDE regions were used to construct the unrooted phylogram based on Neighbor-Joining analysis. Confidence of groupings was estimated by using 1,000 bootstrap replicates; numbers next to the branching point indicate the percentage of replicates supporting each branch. The sequences used for the phylogenetic tree are as follows: *Anopheles gambiae Tc1,* XP_310448.1 [*Anopheles gambiae*]; *Apis cerana mariner,* BAB86288.1 mariner transposase [*Apis cerana*]; *Aspergillus niger Tc1,* AAB50684.1 putative Tc1-mariner class transposase [*Aspergillus niger*]; *Bari-1,* S33560 transposon-like element Bari-1 [*Drosophila melanogaster*]; *Buenoa MLE,* AAC28142.1 mariner transposase [*Buenoa sp.*]; *Drosophila virilis Tc1,* AAA88882.1 Tc1-like transposase [*Drosophila virilis*]; *Epicauta funebris mariner,* AAC28145.1 mariner transposase [*Epicauta funebris*]; *Famar1,* AAO12863.1 Famar1 transposase [*Forficula auricularia*]; *Fungia MLE*,BAB32436.1 transposase [*Fungia sp. Kusabiraishi*]; *impala,* AAB33090.2 transposase [*Fusarium oxysporum*]; *jerky*, NP_032441.3 jerky [*Mus musculus*]; *Meliodogyne chitwoodi MLE,* CAD26968.1 transposase [*Meloidogyne chitwoodi*]; *minos,* S26856 transposon Minos [*Drosophila hydei*]; *pogoR11,* S20478 transposon pogoR11 [*Drosophila melanogaster*]; *quetzal,* AAB02109.1 transposase [*Anopheles albimanus*]; *Oryza sativa mariner*, AC093017_21 putative transposase [*Oryza sativa*]; *Soymar1,* AAC28384.1 mariner transposase [*Glycine max*]; *Tc1,* P03934 TC1A_CAEEL Transposable element TC1 transposase [*Caenorhabditis elegans*]; *Tc3,* P34257 TC3A_CAEEL Transposable element TC3 transposase [*Caenorhabditis elegans*]; *tigger human,* AAH37869.1 Tigger transposable element derived 4 [*Homo sapiens*] and *Trichomonas vaginalis mar1,* AAP45328.1 mar1 putative transposase [*Trichomonas vaginalis*]. *MarPi-1* and *Tc1Pi-1* are two of the three *P. infestans* Class II elements identified in this study on BLASTN search against *P. sojae* and *P. ramorum* genome sequences; *E* value less than e^{-100} and percentage identity higher than 80%

BAC11A5 has a diverse group of retrotransposons

Class I elements are also called retroelements or retrotransposons because they transpose via a mRNA intermediate with an indispensable activity of reverse transcriptase. Class I elements are classified into LTR (Long Terminal Repeat) retrotransposons, LINEs (Long Interspersed Nuclear Elements) and SINEs (Short Interspersed Nuclear elements) (Baltimore 1985; Echalier 1989; Flavell et al. 1992). BAC11A5 contains three LTR retrotransposons and one LINE.

LTR retrotransposons are specified by their direct long terminal repeats. The LTRs are typically flanking a number of genes among which the two major genes called *gag* and *pol*. A number of proteins such as protease (PR), reverse transcriptase (RT), integrase (INT) and RNaseH (RH) can be encoded by the *pol* gene. Based on the sequence divergence of reverse transcriptases and also the order of RT and INT coding domains, LTR retrotransposons are further divided into two groups, the Ty1/copia and the Ty3/gypsy group (Xiong and Eickbush 1990). Fig. 7 shows a phylogenetic tree constructed from the RT

domains of various identified retroelements and the three LTR elements identified in BAC11A5. Two appear to belong to the Ty3/gypsy group and one belongs to the Ty1/copia group. One of the Ty3/gypsy like elements called *GypsyPi-1*, is located on the 8.2 kb repeat unit described above and hence, two *GypsyPi-1* copies are present in BAC11A5. These two copies share 99.4% similarity at nucleotide level. One copy carries a frame shift mutation in the *pol* gene, the other copy still possesses intact *gag* and *pol* genes. A GAG protein of 328 amino acids together with a 1517 amino acids long POL protein comprised of PR, RT, RH and INT domains are encoded by two ORFs flanked by a pair of LTRs. The second Ty3/gypsy like element, named *GypsyPi-2*, is 13.0 kb in length with a pair of LTRs of 530 bp. The Ty1/copia like retrotransposon, named *CopiaPi-1,* is 4.7 kb in length with LTRs of 220 bp. *CopiaPi-1* shows the typical Ty1/copia element domain order of RT-INT, different from the INT-RT domain order in the two Tc3/gypsy like elements.

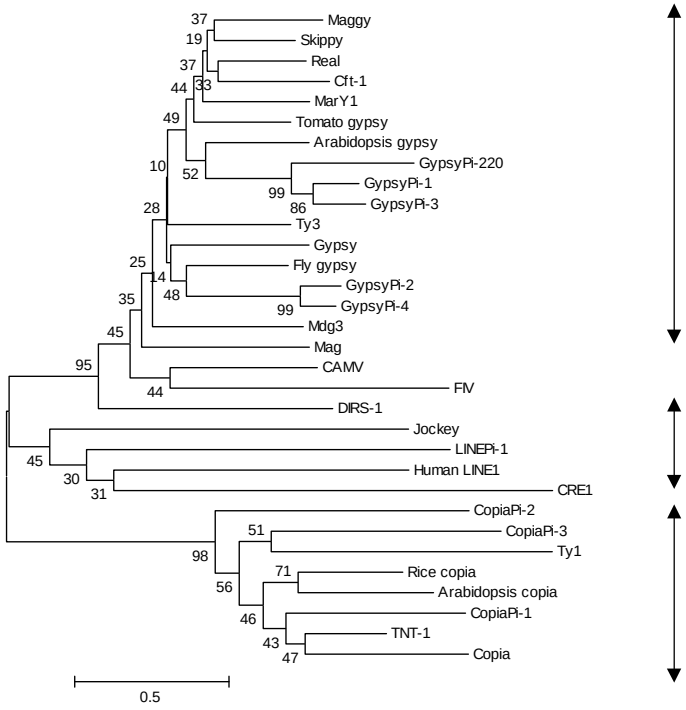


Fig. 7 Phylogram of retroelements. The reverse transcriptase protein sequences were used to construct the unrooted phylogram based on Neighbor-Joining analysis. Confidence of groupings was estimated by using 1,000 bootstrap replicates; numbers next to the branching point indicate the percentage of replicates supporting each branch. On the right different groups of retrotransposons are indicated. Arabidopsis copia, BAB84015.1 polyprotein [*Arabidopsis thaliana*]; Arabidopsis gypsy, AF128395 retrotransposon [*Arabidopsis thaliana*]; CAMV, M90543 reverse transcriptase [*Cauliflower mosaic virus*]; *Cft-1*, AAF21678 pol polyprotein [*Cladosporium fulvum*]; *Copia*, P04146 copia protein [*Drosophila melanogaster*]; *Cre1*, A34728 transposon CRE1 [*Crithidia fasciculata*]; *DIRS-1*, C24785 DIRS-1element [*Dictyostelium discoideum*]; *FIV*, S23820 pol polyprotein [*Feline immunodeficiency virus*]; *Gypsy*, AAB50148, polyprotein [*Drosophila melanogaster*]; *GpysyPi-220*, AF490339 strain 220 gypsy-like retrotransposon [*Phytophthora infestans*]; *human LINE1*, P08547 HUMAN LINE-1 HOMOLOG [*Homo sapiens*] *Jockey*, P21328 mobile element jockey [*Drosophila melanogaster*]; *Mag*, S08405 silkworm transposon mag [*Bombyx mori*]; *Maggy*, AAA33420 polyprotein [*Magnaporthe grisea*]; *MarY1*, BAA78625 polyprotein [*Tricholoma matsutake*]; *Mdg3*, T13798 retrotransposon mdg3 [*Drosophila melanogaster*]; *REAL*, BAA89272 polyprotein Pol [*Alternaria alternata*]; *Rice copia,* AAR88589.1 putative copia-like retrotransposon protein [*Oryza sativa*]; *Skippy*, S60179 retrotransposon skippy [*Fusarium oxysporum*]; *Tnt1*, P10978 Tnt-1element [*Nicotiana tabacum*]; *Tomato gypsy,* T17459 Gypsy-like polyprotein [*Lycopersicon esculentum*]; *Ty1*, B2267 retrotransposon Ty9121 [*Saccharomyces cerevisiae*] and *Ty3*, S69842 Ty3 protein [*Saccharomyces cerevisiae*]. *GypsyPi-1, GypsyPi-2, GypsyPi-3, GypsyPi-4, CopiaPi-1, CopiaPi-2* and *CopiaPi-3* are the seven *P. infestans* retrotransposons identified in this study.

Due to their sequence divergence, LINEs are considered to be the most ancient group of transposable elements (Xiong and Eickbush 1990). LINEs lack the terminal direct repeat and possess a polyadenylation signal in the 3'end of the sequence. In the BAC11A5 sequence contig one LINE of 3.2 kb was identified that was named *LINEPi-1.* As other identified LINEs (Noma et al. 1999; Schmidt 1999), *LINEPi-1* codes for a RT but lacks INT. A protein with endonucleolytic activity is encoded upstream of the RT gene. The 3' polyadenylation signal A_8 is found at 80 bp downstream of the stop codon of the *pol* gene.

The four more-or-less intact retroelements and the LINE cover 33.2% (37.3 kb) of the sequence contig. This percentage underestimates the total fraction of Class I elements in this BAC since a number of fragmented elements are not taken into account.

Distribution of transposable elements in *Phytophthora*

To investigate whether other regions of the genome contain a similar distribution of transposable elements, a number of BAC sequences were retrieved from GenBank and analysed for the presence of retrotransposons. In total 500 kb genomic DNA sequence composed of fragments from five partially sequenced BACs was compiled and fragments without unordered gaps were used for the analysis. The BACs which are derived from the same library and the same strain as BAC11A5 are PI-BAC-14M19 (AC146943), PI-BAC-14P22 (AC146983), PI-BAC-21G17 (AY497062), PI-BAC-25C5 (AC147181) and PI-BAC-26O7 (AC147180) and cover randomly selected regions containing genes with significant homology to a variety of known sequences, e.g., beta-glucosidase, deoxyribose-phosphate aldolase and elongation factor. In the 500 kb fragments two extra copies of *GypsyPi-1* were found each sharing 97% homology at the nucleotide level with the copy found in BAC11A5. In addition, two new Ty3/gypsy like retrotransposons, *GypsyPi-3* and *GypsyPi-4*, and two new Ty1/copia like retrotransposons, *CopiaPi-2* and *CopiaPi-3*, were identified. The features of the seven retroelements and one LINE are summarized in Table 3 and Fig. 8.

Table 3 Characteristics of the different retrotransposons described in this study. The retrotransposons are found either on BAC11A5 or on 500 kb random genomic fragments. The GC content calculation was performed on the reverse transcriptase regions of the retrotransposons. The highest percentage is in bold and underlined.

retroelement [a]	group	size (kb)	LTR (bp)	copies in BAC11A5	copies in random 500 kb	LTR pair similarity %	GC%			number of EST hits [c]	homologue in *P. sojae* [d]	homologue in *P. ramorum* [d]
							frame 1	frame 2	frame 3			
GypsyPi-1	gypsy	8.2	350	2	4	99.7/99.7/100.0/100.0	57.55	41.88	**72.92**	4	yes	yes
GypsyPi-2	gypsy	13.0	530	1	1	97.7	54.61	38.01	**63.47**	1	no	no
GypsyPi-3	gypsy	7.6	314	0	2	94.2/99.4	53.60	42.24	**71.48**	0	yes	yes
GypsyPi-4	gypsy	~12 [b]	540	0	1	98.0	54.51	40.23	**72.56**	1	yes	yes
CopiaPi-1	copia	4.7	220	1	1	91.4	52.41	39.66	**64.01**	0	yes	no
CopiaPi-2	copia	5.7	240	0	1	81.3	**50.89**	43.42	37.86	1	yes	yes
CopiaPi-3	copia	5.5	238	0	1	98.3	**67.94**	36.24	45.80	0	no	no
LINEPi-1	LINE	3.2	none	1	ND	none	50.17	44.07	**52.04**	1	no	no

[a] GenBank accession numbers AY830091 (*GypsyPi-1*), AY830106 (*GypsyPi-2)*, AY830104 (*GypsyPi-3*), AY830107 (*GypsyPi-4*), AY830098 (*CopiaPi-1*), AY830099 (*CopiaPi-2*), AY830100 (*CopiaPi-3),* AY830108 *(LINEPi-1)*
[b] due to a gap in the GenBank sequence the size of *GypsyPi-4* size can not be determined exactly
[c] based on BLASTN search against 72,904 *P. infestans* ESTs; *E* value less than e^{-100} and percentage identity higher than 95%
[d] based on BLASTN search against *P. sojae* and *P. ramorum* genome sequences; *E* value less than e^{-100} and percentage identity higher than 80%

The retroelements identified in this study are not identical to any of the previously described Ty1/copia and Ty3/gypsy elements in *Phytophthora.* Tooley and Garfinkel (1996) identified a number of Ty1/copia like elements in *P. infestans* by degenerate PCR but the partial sequences are not suitable for phylogenetic analysis. By using a similar approach Judelson (2002) obtained partial Ty3/gypsy sequences from several *Phytophthora* species one of which, *P. infestans GypsyPi-220,* was analyzed in more detail. *GypsyPi-220* is closely related to *GypsyPi-1* and *GypsyPi-3* with 56% and 60% similarity, respectively and in the phylogenetic tree these three retroelements seem to form a subclass of *P. infestans* specific Tc3/gypsy like retrotransposons (Fig. 7).

To investigate whether the transposable elements found in BAC11A5 also exist in other *Phytophthora* genomes, NCBI trace files containing 1,533,511 *P. sojae* genome sequences and 898,494 *P. ramorum* sequences were searched using BLASTN. Hits with more than 80% homology and *E* values lower than e-100 were considered to be homologous. None of the DNA transposons nor the LINE element were found in *P. sojae* or *P. ramorum* (Table 2 and 3). In contrast, several LTR retrotransposons were present of both, the Tc3/gypsy class and the Tc1/copia class (Table 3) indicating that these elements are more widely distributed in the genus and invaded the *Phytophthora* genome before speciation. Phylogenetically *P. infestans* is not a close relative of *P. sojae* or *P. ramorum*; the tree species fall in three different clades (Kroon et al. 2004). The Class II elements could well be clade-specific and hence, present in more closely related species within the *P. infestans* clade such as *P. mirabilis* and *P. phaseoli.* This is true for *DodoPi,* a recently described *P. infestans* hAT-like DNA transposon that is not related to the Class II transposons described in this study (Ah Fong and Judelson 2004).

GypsyPi-1 and _GypsyPi-3_ belong to a class of retrotransposons that carry a CHROMO domain

GypsyPi-1 and *GypsyPi-3* are closely related: the similarity of the protein sequence of the reverse transcriptase region is 71%. The C-terminal part of the POL proteins deduced from the *GypsyPi-1* and *GypsyPi-3* sequences both contain a CHROMO (CHRromatin Organization MOdifier) domain (PF00385). The closely related *GypsyPi-220* is truncated at the C-terminus and also lacks the LTRs (Judelson 2002). Domain searches with the protein sequences against the pfam database (Bateman et al. 1999; Bateman et al. 2000) gave hits of this CHROMO domain with an *E* value of 1e-08. CHROMO domains are associated with alteration of the structure of chromatin to the condensed morphology of heterochromatin (Cavalli and Paro 1998). Proteins with CHROMO domains can often modify the structure of chromatin. Examples are the *Drosophila* and human heterochromatin protein Su (HP1) (Aasland and Stewart 1995), the *Drosophila* protein Polycomb (Pc) (Paro and Hogness 1991) and mammalian DNA-binding/helicase proteins (Koonin et al. 1995). A POL protein combined with a CHROMO domain is not unique. A CDART (Conserved Domain Architecture Retrieval Tool) (Geer et al. 2002) search revealed several other retroelements with a CHROMO domain at the C-terminal end of POL, and a similar protein architecture as *GypsyPi-1* and *GypsyPi-3,* a.o., *CfT-1* from the fungus *Cladosporium fulvum* (AAF21678), *Skipper* from *Dictyostelium discoideum* (T14598), a retrotransposon from the fish *Takifugu rubripes* (AAC33526) and two plant retrotransposons (*Oryza sativa* NP_920591

and *Arabidopsis thaliana* NP_683628). It will be interesting to know to what extent this CHROMO domain influences the retro-transposition events, and if they modify the chromatin structure of the host.

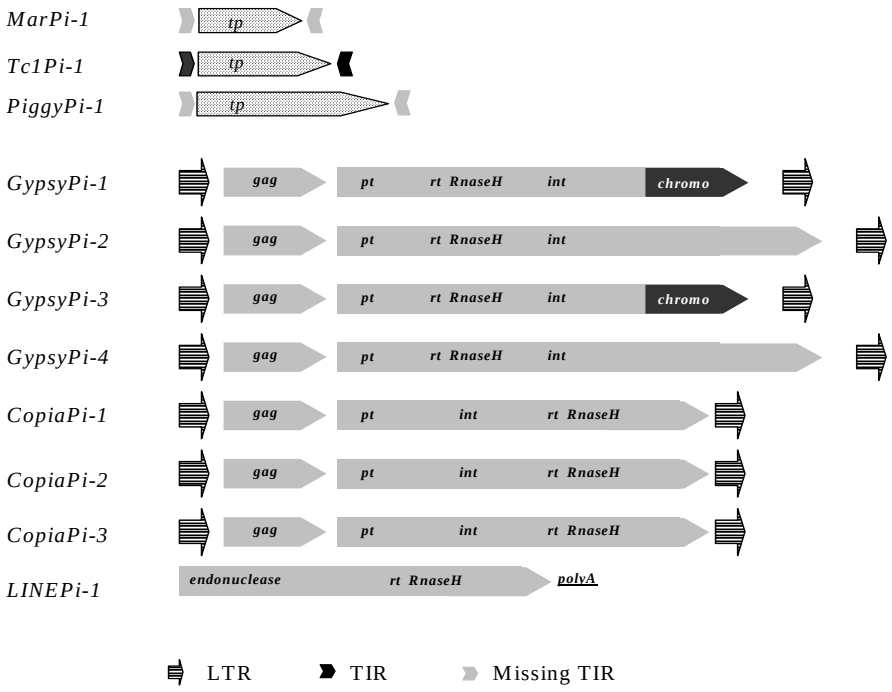


Fig. 8 Schematic representation of mobile elements described in this study. The sizes of the elements are not on scale. They are listed in Table 2 and 3. TIR: Terminal Inverted Repeat, LTR: Long Terminal Repeat, *tp*: transposase, *pt*: protease, *rt:* reverse transcriptase, *int*: integrase and *chrom*: CHROMO domain.

Activity of the transposable elements

The genomic region covered by BAC11A5 seems to be a hotspot for retro- and DNA transposon insertions. However, several of the mobile elements identified in BAC11A5 and in the 500 kb random genomic DNA sequence, carry numerous mutations and deletions indicating lack of activity of the majority of the transposons. Nevertheless, EST database searches show the occurrence of transcripts of a diverse group of mobile element in various developmental stages of the life cycle of *P. infestans*, demonstrating that at least some transposons are actively transcribed. From *GypsyPi-1, GypsyPi-2, CopiaPi-2* and *LINEPi-1* up to four transcripts are found in the EST database with a homology in the range of 95% to 99% at nucleotide level (Table 3). *MarPi-1* even has 36 transcripts with more than 95% homology and distributed over libraries from different developmental stages (Table 2).

As described above, the GC3 of the ORF in *P. infestans* genes is generally high. To determine whether the ORFs present in the transposable elements have the same characteristic we performed a GC analysis of the 800 bp ORF encoding RT in *GypsyPi-1, GypsyPi-2, GypsyPi-3, GypsyPi-4, CopiaPi-1, CopiaPi-2, CopiaPi-3* and *LINEPi-1* and the transposase ORF of *MarPi-1*, *Tc1Pi-1* and *PiggyPi-1*. The

results are shown in Table 2 and 3. All four Tc3/gypsy elements, one Tc1/copia like element, *CopiaPi-1,* and one DNA transposon, *Tc1Pi-1,* have a high GC3 content similar to most *P. infestans* genes. The GC3 content is above 60% and is higher than GC1 and GC2. In contrast, *CopiaPi-2, CopiaPi-3, LINEPi-1, MarPi-1* and *PiggyPi-1* either do not show a high GC3 or GC3 is not the highest of the three.

GypsyPi-1 and *GypsyPi-3* not only show the high GC3 feature, but also have a similar codon usage as *Phytophthora* genes. Codon usage was calculated for the deduced GAG and POL proteins of *GypsyPi-1* and *GypsyPi-3* and the same calculation was performed on several sets of ORFs deduced from 1000 randomly selected ESTs of each *P. infestans* and *P. sojae.* A high correlation was found for the codon usage in *P. infestans* and *P. sojae* ORFs (regression line y = 0.96x + 1.38; R^2=0.92) and in *P. infestans* ORFs and *GypsyPi-1* / *GypsyPi-3* ORFs (regression line y = 0.88x + 3.78; R^2=0.73 (data not shown).

The LTRs of LTR retrotransposons are generated during the replication and integration process as a pair of identical sequences (Boeke and Corces 1989). The divergence of this pair of sequences indicates the time elapsed since the event of transposition: the more divergent the LTR pair is, the longer ago the transposition event occurred. The sequence similarity was calculated for the LTR pairs of the four Tc3/gypsy elements and three Tc1/copia elements. The highest sequence similarity is found in the pairs of the four *GypsyPi-1* copies, ranging from 99.4 to 100.0% whereas the lowest similarity (81.3 %) is found in *CopiaPi-2* (Table 3). This indicates that *GypsyPi-1* was transposed in a more recent past then the other retrotransposons identified in this study. Like *GypsyPi-1, GypsyPi-3* is probably a relatively 'young' retroelement. Both seem to be widespread because four and two copies of *GypsyPi-1* and *GypsyPi-3,* respectively*,* were found in 500 kb random genomic fragments and the sequences of the copies located at different positions have a high similarity in their LTRs as well as coding regions.

It is remarkable that both the 'young' retroelements *GypsyPi-1* and *GypsyPi-3* show a high GC3 and a codon usage that is similar to that of other *P. infestans* genes. Maybe these elements have already resided in *P. infestans* since the early stage of *P. infestans* evolution and gradually acquired the characteristics of host genes allowing a more efficient use of host cellular machinery during replication and transposition. *GypsyPi-1* is indeed transcribed as demonstrated by the identification of *GypsiPi-1* ESTs.

Conclusions

Physical mapping of BACs, BAC sequencing and annotation of a long contiguous stretch of genomic DNA of *P. infestans* showed that members of the elicitin gene family are clustered in the genome but yet dispersed over a large region of 200-250 kb that is invaded by repeats and numerous transposable elements. Two of the four *inf* genes, *inf3* and *inf6* reside on a gene island with one other gene of unknown function*,* whereas the two other *inf* genes, *inf1* and *inf4,* are 46 kb apart, and 86 and 40 kb, respectively, away from the gene island. Comparison of the coding and non-coding sequences showed

that the GC content of the coding regions is slightly higher. More significant was the high GC content of the third base of a codon in an ORF, the GC3, a characteristic feature that can be used in gene prediction programs. In the promoter regions a few putative regulatory elements were found but as yet the relevance of those elements is unknown.

Transposons and retrotransposons are ubiquitous in various kingdoms, such as fungi, plants, ciliates and animals (Kim et al. 1998; Daboussi and Capy 2003). Due to the difference in their mode of transposition, Class I and Class II elements are thought to contribute differently to the genome size (Kumar and Bennetzen 1999). Class II DNA transposons have a 'copy/cut –paste' mode of transposition while Class I retroelements have to go through a RNA intermediate step and can therefore potentially be propagated in large quantities. One third of the BAC11A5 sequence contig consists of Class I retroelements and also other genomic regions in *P. infestans* contain numerous retroelements (this study; Judelson 2002; Tooley and Garfinkel 1996). Without a genome sequence it is, as yet, not possible to calculate the overall percentage of transposon sequences in the *P. infestans* genome but it seems likely that transposons are, at least in part, responsible for the large genome size of 240 Mb.

With two *Phytophthora* genomes sequenced (www.jgi.doe.gov/) comparative genomics is now within reach. Efforts are currently focused on finding distinctive features in genomes of various *Phytophthora* species and developing gene annotation tools. This study was a small scale inventory of genome organization and genome structure in *P. infestans* and a first step into the annotation process. With a large EST repository (Randall et al. accepted for publication) and a survey genome sequence of *P. infestans* in hand (O'Neill et al. 2004) *P. infestans* is ready to enter the genomics arena.

Acknowledgements

We are grateful to Sharmili Mathur for expert technical assistance, Steve Whisson for providing the BAC library and filters, Grardy van den Berg for screening the BAC library, and Pierre de Wit for critically reading the manuscript. This work was financially supported by NWO-Aspasia grant 015.000.057 and USDA Cooperative Agreement #58-8230-6-081. The authors acknowledge Syngenta for access to the Syngenta *Phytophthora* Consortium EST Database and the Broad Institute and the DOE Joint Genome Institute for depositing random genomic sequences of *P. infestans* and *P. sojae*, respectively, in the NCBI trace file archive

References

Aasland R, Stewart AF (1995) The chromo shadow domain, a second chromo domain in heterochromatin-binding protein 1, HP1. Nucleic Acids Res 23: 3168-3173

Ah Fong AM, Judelson HS (2004) The hAT -like DNA transposon *DodoPi* resides in a cluster of retro- and DNA transposons in the stramenopile *Phytophthora infestans*. Mol Genet Genomics 271: 577-585

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. J Mol Biol 215: 403-410

Baldauf SL (2003) The deep roots of eukaryotes. Science 300: 1703-1706

Baltimore D (1985) Retroviruses and retrotransposons - the role of reverse transcription in shaping the eukaryotic genome. Cell 40: 481-482

Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer ELL (1999) Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 27: 260-262

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer ELL (2000) The Pfam protein families database. Nucleic Acids Res 28: 263-266

Boeke JD, Corces VG (1989) Transcription and reverse transcription of retrotransposons. Annu Rev Microbiol 43: 403-434

Bonfield JK, Smith K, Staden R (1995) A new DNA sequence assembly program. Nucleic Acids Res 23: 4992-4999

Cavalli G, Paro R (1998) Chromo-domain proteins: linking chromatin structure to epigenetic regulation. Curr Opin Cell Biol 10: 354-360

Daboussi MJ, Capy P (2003) Transposable elements in filamentous fungi. Annu Rev Microbiol 57: 275-299

Duclos J, Fauconnier A, Coelho AC, Bollen A, Cravador A, Godfroid E (1998) Identification of an elicitin gene cluster in *Phytophthora cinnamomi*. DNA Seq 9: 231-237

Echalier G (1989) Drosophila retrotransposons - interactions with genome. Adv Virus Res 36: 33-105

Erwin DC, Ribeiro OK (1996) *Phytophthora* diseases worldwide. The American Phytopathological Society, St. Paul Minnesota USA

Feschotte C, Jiang N, Wessler SR (2002) Plant transposable elements: Where genetics meets genomics. Nature Rev Genetics 3: 329-341

Flavell AJ, Smith DB, Kumar A (1992) Extreme heterogeneity of Ty1-Copia group retrotransposons in plants. Mol Gen Genetics 231: 233-242

Geer LY, Domrachev M, Lipman DJ, Bryant SH (2002) CDART: protein homology by domain architecture. Genome Res 12: 1619-1623

Hendrix JW, Guttman SM (1970) Sterol or calcium requirement by *Phytophthora parasitica* var. *nicotianae* for growth on nitrate. Mycologia 62: 195-198

Hraber PT, Weller JW (2001) On the species of origin: diagnosing the source of symbiotic transcripts. Genome Biol 2: 37

Jeanmougin F, Thompson JD, Gouy M, Higgins DG, Gibson TJ (1998) Multiple sequence alignment with Clustal X. Trends Biochem Sci 23: 403-405

Judelson HS (2002) Sequence variation and genomic amplification of a family of Gypsy-like elements in the oomycete genus *Phytophthora*. Mol Biol Evol 19: 1313-1322

Kamoun S (2003) Molecular genetics of pathogenic oomycetes. Eukaryot Cell 2: 191-199

Kamoun S, Hraber P, Sobral B, Nuss D, Govers F (1999) Initial assessment of gene diversity for the oomycete pathogen *Phytophthora infestans* based on expressed sequences. Fungal Genet Biol 28: 94-106

Kamoun S, Styer A (2000) An improved codon usage table for *Phytophthora infestans*. http://www.oardc.ohio-state.edu/phytophthora/codon.htm

Kamoun S, van West P, Vleeshouwers VGAA, de Groot KE, Govers F (1998) Resistance of *Nicotiana benthamiana* to *Phytophthora infestans* is mediated by the recognition of elicitor protein INF1. The Plant Cell 10: 1413-1426

Kamoun S, van West P, de Jong AJ, de Groot KE, Vleeshouwers VGAA, Govers F (1997) A gene encoding a protein elicitor of *Phytophthora infestans* is down-regulated during infection of potato. Mol Plant Microbe Interact 10: 13-20

Kim JM, Vanguri S, Boeke JD, Gabriel A, Voytas DF (1998) Transposable elements and genome organization: A comprehensive survey of retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Res 8: 464-478

Koonin EV, Zhou S, Lucchesi JC (1995) The chromo superfamily: new members, duplication of the chromo domain and possible role in delivering transcription regulators to chromatin. Nucleic Acids Res 23: 4229-4233

Kroon LP, Bakker FT, Van Den Bosch GB, Bonants PJ, Flier WG (2004) Phylogenetic analysis of *Phytophthora* species based on mitochondrial and nuclear DNA sequences. Fungal Genet Biol 41: 766-782

Kumar A, Bennetzen JL (1999) Plant retrotransposons. Annu Rev Genet 33: 479-532

Kumar S, Tamura K, Nei M (1994) MEGA - Molecular evolutionary genetics analysis software for microcomputers. Comput Appl Biosci 10: 189-191

Latijnhouwers M, de Wit PJGM, Govers F (2003) Oomycetes and fungi: similar weaponry to attack plants. Trends Microbiol 11: 462-469

McLeod A, Smart CD, Fry WE (2004) Core promoter structure in the oomycete *Phytophthora infestans*. Eukaryot Cell 3: 91-99

Mikes V, Milat ML, Ponchet M, Panabieres F, Ricci P, Blein JP (1998) Elicitins, proteinaceous elicitors of plant defense, are a new class of sterol carrier proteins. Biochem Biophys Res Commun 245: 133-139

Noma K, Ohtsubo E, Ohtsubo H (1999) Non-LTR retrotransposons (LINEs) as ubiquitous components of plant genomes. Mol Gen Genet 261: 71-79

O'Neill K, Zody MC, Karlsson E, Govers F, van der Vondervoort P, Weide R, Whisson S, Birch P, Ma L, Birren B, Fry W, Judelson H, Kamoun S, Nusbaum C (2004) Sequencing the *Phytophthora infestans* Genome: Preliminary Studies. Book of Abstracts of the annual meeting of the NSF *Phytophthora* Molecular Genetics Network, New Orleans LA, USA, May 21-13, 2004, p. 5.

Panabieres F, Marais A, LeBerre JY, Penot I, Fournier D, Ricci P (1995) Characterization of a gene cluster of *Phytophthora cryptogea* which codes for elicitins, proteins inducing a hypersensitive- like response in tobacco. Mol Plant Microbe Interact 8: 996-1003

Paro R, Hogness DS (1991) The Polycomb protein shares a homologous domain with a heterochromatin-associated protein of *Drosophila*. Proc Natl Acad Sci U S A 88: 263-267

Pieterse CMJ, Van West P, Verbakel HM, Brasse P, Van den Berg-Velthuis GCM, Govers F (1994) Structure and Genomic Organization of the *ipib* and *ipio* gene clusters of *Phytophthora infestans*. Gene 138: 67-77

Plasterk RH (1996) The Tc1/mariner transposon family. Curr Top Microbiol Immunol 204: 125-143

Qutob D, Hraber PT, Sobral BWS, Gijzen M (2000) Comparative analysis of expressed sequences in *Phytophthora sojae*. Plant Physiol 123: 243-253

Randall TA, Dwyer RA, Huitema E, Beyer K, Cvitanich C, Kelkar H, Fong AMVA, Gates K, Roberts S, Yatzkan E, Gaffney T, Law M, Testa A, Torto T, Zhang M, Zheng L, Mueller E, Windass J, Binder A, Birch PRJ, Gisi U, Govers F, Gow N, Mauch F, van West P, Waugh M, Yu J, Boller T, Kamoun S, Lam ST, Judelson HS (2004) Large-scale gene discovery in the oomycete *Phytophthora infestans* reveals likely components of phytopathogenicity shared with true fungi. Mol Plant Microbe Interact Accepted for publication

Ricci P, Trentin F, Bonnet P, Venard P, Moutonperronnet F, Bruneteau M (1992) Differential production of parasiticein, an elicitor of necrosis and resistance in tobacco, by isolates of *Phytophthora Parasitica*. Plant Pathology 41: 298-307

Sarkar A, Sim C, Hong YS, Hogan JR, Fraser MJ, Robertson HM, Collins FH (2003) Molecular evolutionary analysis of the widespread *piggyBac* transposon family and related "domesticated" sequences. Mol Genet Genomics 270: 173-180

Schmidt T (1999) LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. Plant Mol Biol 40: 903-910

Schwartz S, Zhang Z, Frazer KA, Smit A, Riemer C, Bouck J, Gibbs R, Hardison R, Miller W (2000) PipMaker--a web server for aligning two genomic DNA sequences. Genome Res 10: 577-586

Skalamera D, Wasson AP, Hardham AR (2004) Genes expressed in zoospores of *Phytophthora nicotianae*. Mol Genet Genomics 270: 549-557

Smit AF, Riggs AD (1996) *Tiggers* and DNA transposon fossils in the human genome. Proc Natl Acad Sci U S A 93: 1443-1448

Soanes DM, Skinner W, Keon J, Hargreaves J, Talbot NJ (2002) Genomics of phytopathogenic fungi and the development of bioinformatic resources. Mol Plant Microbe Interact 15: 421-427

Tooley PW, Garfinkel DJ (1996) Presence of Ty1-copia group retrotransposon sequences in the potato late blight pathogen *Phytophthora infestans*. Mol Plant Microbe Interact 9: 305-309

Tudor M, Lobocka M, Goodell M, Pettitt J, O'Hare K (1992) The *pogo* transposable element family of *Drosophila melanogaster*. Mol Gen Genet 232: 126-134

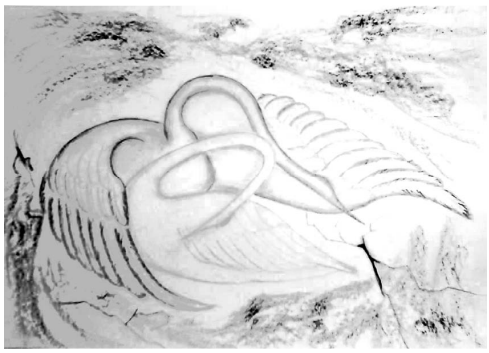
Whisson SC, van der Lee T, Bryan GJ, Waugh R, Govers F, Birch PRJ (2001) Physical mapping across an avirulence locus of *Phytophthora infestans* using a highly representative, large-insert bacterial artificial chromosome library. Mol Genet Genomics 266: 289-295

Xiong Y, Eickbush TH (1990) Origin and evolution of retroelements based upon their reverse-transcriptase sequences. EMBO J 9: 3353-3362

Supplemental material

Table S1 Codon usage and 3rd position GC frequency in 79 *Phytophthora infestans* genes. A total of 28583 codons was counted.

Amino acids	Codon usage %						Codon ends with A or T %	Codon ends with G or C %
*	TAA	TAG	TGA					
*	58.23	24.05	17.72				**75.95**	24.05
A	GCA	GCC	GCG	GCT				
A	13.54	36.31	19.69	30.47			44.00	**56.00**
C	TGC	TGT						
C	72.20	27.80					27.80	**72.20**
D	GAC	GAT						
D	76.41	23.59					23.59	**76.41**
E	GAA	GAG						
E	19.24	80.76					19.24	**80.76**
F	TTC	TTT						
F	78.96	21.04					21.04	**78.96**
G	GGA	GGC	GGG	GGT				
G	15.23	47.47	5.50	31.81			47.03	**52.97**
H	CAC	CAT						
H	80.53	19.47					19.47	**80.53**
I	ATA	ATC	ATT					
I	2.45	68.19	29.35				31.81	**68.19**
K	AAA	AAG						
K	14.29	85.71					14.29	**85.71**
L	CTA	CTC	CTG	CTT	TTA	TTG		
L	6.66	24.21	41.11	14.03	1.64	12.34	22.34	**77.66**
M	ATG							
M	100.00						0.00	**100.00**
N	AAC	AAT						
N	82.00	18.00					18.00	**82.00**
P	CCA	CCC	CCG	CCT				
P	13.93	26.78	34.51	24.79			38.72	**61.28**
Q	CAA	CAG						
Q	21.95	78.05					21.95	**78.05**
R	AGA	AGG	CGA	CGC	CGG	CGT		
R	4.81	3.98	11.94	38.70	5.46	35.09	**51.85**	48.15
S	AGC	AGT	TCA	TCC	TCG	TCT		
S	19.58	9.77	6.61	16.28	37.52	10.24	26.62	**73.38**
T	ACA	ACC	ACG	ACT				
T	10.61	35.99	38.29	15.11			25.73	**74.27**
V	GTA	GTC	GTG	GTT				
V	5.60	29.75	52.00	12.64			18.24	**81.76**
W	TGG							
W	100.00						0.00	**100.00**
Y	TAC	TAT						
Y	87.37	12.63					12.63	**87.37**

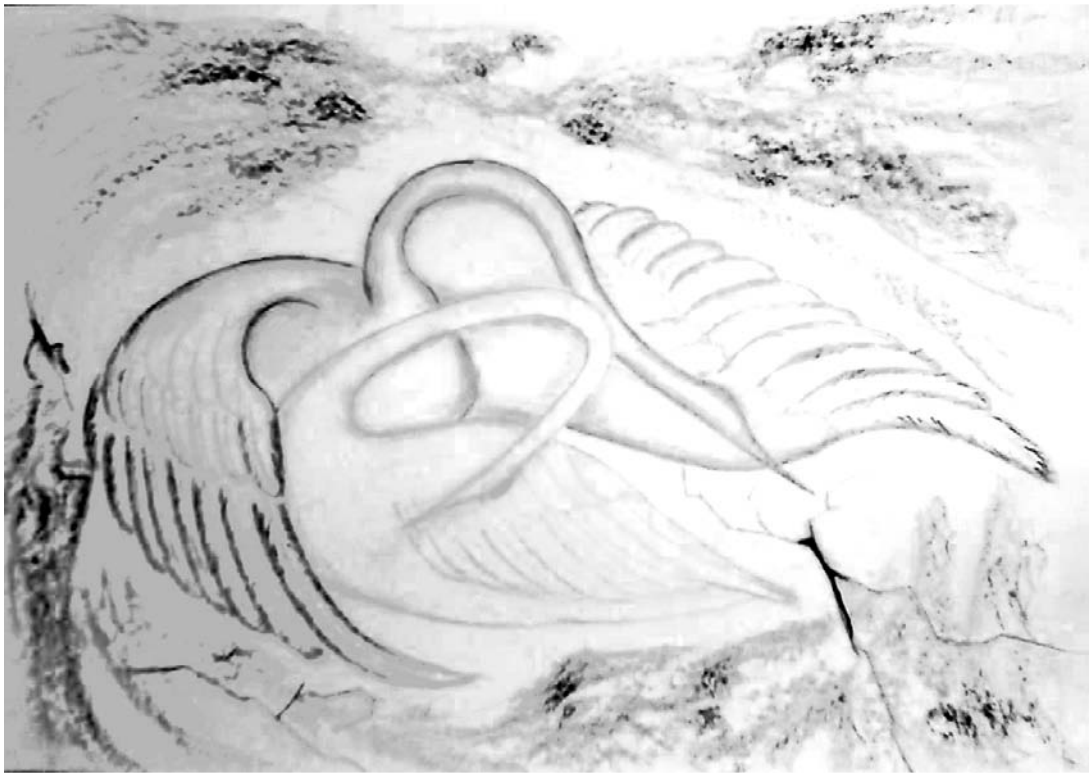


Chapter 5

Ancient origin of elicitin gene clusters in *Phytophthora* genomes

Molecular Biology and Evolution 23(2):338-351

Rays H.Y. Jiang, Brett M. Tyler, Stephen C. Whisson, Adrienne R. Hardham and Francine Govers



Ancient origin of elicitin gene clusters in *Phytophthora* genomes

Rays H.Y. Jiang*, Brett M. Tyler[§], Stephen C. Whisson[#], Adrienne R. Hardham[†] and Francine Govers*

*Laboratory of Phytopathology, Wageningen University, Binnenhaven 5, NL-6709 PD, Wageningen and Graduate School Experimental Plant Sciences, The Netherlands
[§]Virginia Bioinformatics Institute, Virginia Polytechnic and State University, Blacksburg VA, USA
[#] Plant Pathogen Interactions Programme, Scottish Crop Research Institute, Invergowrie, Dundee DD2 5DA, Scotland, UK
[†]Plant Cell Biology Group, Research School of Biological Sciences, Australian National University, Canberra, ACT 2601, Australia

For correspondence: E-mail Francine.Govers@wur.nl; Tel. +31 317 483 138; Fax +31 317 483 412

Keywords

elicitin, *Phytophthora*, molecular phylogeny

Abstract

The genus *Phytophthora* belongs to the oomycetes in the eukaryotic stramenopile lineage and is comprised of over 65 species that are all destructive plant pathogens on a wide range of dicotylodons. *Phytophthora* produces elicitins (ELIs), a group of extracellular elicitor proteins that cause a hypersensitive response in tobacco. Database mining revealed several new classes of elicitin-like (ELL) sequences with diverse elicitin domains in *Phytophthora infestans, Phytophthora sojae*, *Phytophthora brassicae* and *Phytophthora ramorum*. ELIs and ELLs were shown to be unique to *Phytophthora* and *Pythium* species. They are ubiquitous among *Phytophthora* species and belong to one of the most highly conserved and complex protein families in the *Phytophthora* genus. Phylogeny construction with elicitin domains derived from 156 ELIs and ELLs showed that most of the diversified family members existed prior to divergence of *Phytophthora* species from a common ancestor. Analysis to discriminate diversifying and purifying selection showed that all 17 ELI and ELL clades are under purifying selection. Within highly similar ELI groups there was no evidence for positively selected amino acids suggesting that purifying selection contributes to the continued existence of this diverse protein family. Characteristic cysteine spacing patterns were found for each phylogenetic clade. Except for the canonical clade ELI-1, ELIs and ELLs possess C-terminal domains of variable length, many of which have a high threonine, serine or proline content suggesting an association with the cell wall. In addition, some ELIs and ELLs have a predicted GPI (glycosylphosphatidylinositol) site suggesting anchoring of the C-terminal domain to the cell membrane. The *eli* and *ell* genes belonging to different clades are clustered in the genomes. Overall, *eli* and *ell* genes are expressed at different levels and in different life cycle stages but those sharing the same phylogenetic clade appear to have similar expression patterns.

Introduction

The genus *Phytophthora* comprises over 65 phytopathogenic species that cause many economically important diseases and can have devastating effects on natural habitats (Erwin and Ribeiro 1996). *Phytophthora infestans,* also known as the notorious 'Irish potato famine fungus', causes late blight disease on potato and tomato worldwide, and *P. sojae* is responsible for root and stem rot on soybean. Two recently discovered species are *P. ramorum*, the causal agent of 'Sudden Oak Death' (Werres et al. 2001; Rizzo, Garbelotto, and Hansen 2005), and *P. brassicae,* a pathogen on the model plant *Arabidopsis thaliana* (Roetschi et al. 2001; Man in 't Veld et al. 2002). *Phytophthora* belongs to the oomycetes, a diverse group of fungus-like eukaryotes that share phylogenetic similarity with brown algae and diatoms. In the tree of life oomycetes are grouped in the stramenopile lineage that is distant from the plant, animal and fungal lineages (Margulis and Schwarts 2000; Baldauf 2003).

A common feature of many different types of plant pathogens is the secretion of a variety of extracellular effector molecules into the plant apoplast (Van't Slot and Knogge 2002) that are presumed to promote infection of the host plant. Many of these proteins, called elicitors, elicit plant defense responses and, in particular, a form of programmed cell death called the hypersensitive response (HR). In most cases, the defense response benefits the plant, and the response is triggered by the detection of the elicitors by plant defense receptors. In some cases however, elicitation of these responses promotes infection because the pathogen can thrive on the dying plant tissue. *Phytophthora* species ubiquitously secrete a unique class of highly conserved effector molecules named elicitins. Elicitins are wide-spread in *Phytophthora* species and closely related *Pythium* species (Panabieres et al. 1997) but are absent from any other organism studied so far. Hence it is conceivable that they could be responsible for novel mechanisms of interaction with plants.

Molecular cloning and EST database analysis showed that elicitin genes form families in *Phytophthora* species such as *P. infestans* (Kamoun, Lindqvist, and Govers 1997; Kamoun et al. 1999)*, P. sojae* (Mao and Tyler 1996; Qutob et al. 2003)*, P. brassicae* (L. Belbahri and F. Mauch, personal communication) and *Phytophthora cinnamomi* (Duclos et al. 1998). Gene families are considered to arise from chromosomal duplications. In theory, only one copy must maintain the original function whereas the other copies can undergo functional divergence. Remarkably, in *Phytophthora* most of the extracellular proteins described to date are encoded by multigene families (Gotesson et al. 2002; Qutob, Kamoun, and Gijzen 2002; Torto et al. 2003; Liu et al. 2005). Some of these gene families are *Phytophthora* specific whereas others have homologues in plant pathogens in other kingdoms such as fungi and bacteria.

Based on the phylogenetic distribution pattern, elicitor genes can roughly be divided into three groups (fig. 1). Group-A elicitors have homologues across kingdoms. The most prominent example is the NIP or

NPP (Necrosis Inducing Protein *Phytophthora*) family that belongs to the super family of Nep1-like proteins (NLPs) present in bacteria, fungi and stramenopiles (Qutob, Kamoun, and Gijzen 2002; Pemberton and Salmond 2004). Elicitins are typical examples of Group-B elicitors which are conserved but only present in one or a few genera. Group-C elicitors are species specific or highly divergent such as IPI-O (Pieterse et al. 1994), AVR3a (Armstrong et al. 2005) and SCR74 (Liu et al. 2005) in *P. infestans* and AVR1b in *P. sojae* (Shan et al. 2004). The different phylogenetic distributions of elicitors may relate to the role they have in the interaction with host plants. Some of the highly divergent Group-C elicitors, for example, are race specific elicitors (i.e. AVR3a and AVR1b of *P. infestans* and *P. sojae*, respectively), and are involved in highly specific gene-for-gene interactions with major NBS-LRR resistance genes.

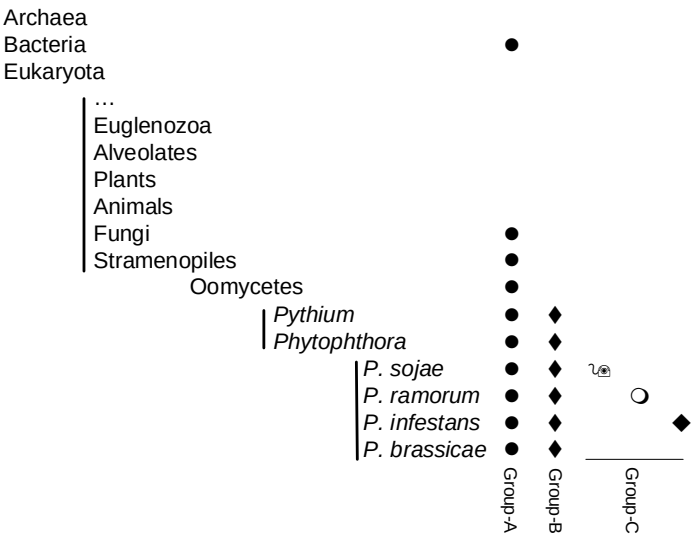


Fig. 1. — Phylogenetic distribution of *Phytophthora* elicitors. Group-A elicitors are distributed across kingdoms, Group-B elicitors are restricted to oomycetes and conserved between species, and Group-C elicitors are species specific or highly divergent between species.

To address the question of how the elicitin gene family evolved and what kind of functions the different family members may have, a thorough investigation of the diversity of the family members is needed. Elicitins all share a conserved domain with a characteristic signature of six cysteine residues that form three distinct disulfide bonds (Fefeu et al. 1997). These features together with the small sizes of the proteins allow efficient mining of EST (Expressed Sequence Tag) databases and genome sequences by PSI-BLAST (Altschul et al. 1997). To investigate the diversity and evolutionary relationship within the elicitin gene family, we searched for new elicitin gene family members by making use of *Phytophthora* EST databases (Randall et al. 2005) and whole genome sequences of *P. sojae* and *P. ramorum* (http://genome.jgi-psf.org). Subsequently, we classified all family members based on sequence diversity and protein motifs and constructed a phylogenetic tree to reveal their evolutionary relationships. We also investigated the molecular evolution within clades by calculating the rate of nonsynonymous to synonymous substitutions (d_N/d_S or ω) and the spatial and temporal expression patterns of various family

members by Northern blot hybridization and transcript counting in EST databases. Finally, we analyzed and compared the genome organization of elicitin genes across *Phytophthora* species

Materials and methods

Genome databases and EST databases

The *P. infestans* and *P. sojae* EST databases are accessible at http://www.pfgd.org and http://staff.vbi.vt.edu/estap and most *Phytophthora* EST sequences are available through GenBank (Kamoun et al. 1999; Qutob et al. 2000). *Blumeria graminis*, *Magnaporthe grisea* and *Cladosporium fulvum* EST databases were downloaded from Phytopathogenic Fungi and Oomycete EST Database Version 1.4 (Soanes et al. 2002) (http://cogeme.ex.ac.uk) and the *Magnaporthe grisea* genome sequence was available at http://www.broad.mit.edu/annotation/fungi/magnaporthe/ (Dean et al. 2005). The genomic sequences and annotated protein sequences of *P. sojae*, *P. ramorum* and of the diatom *Thalassiosira pseudonana* (Armbrust et al. 2004) were obtained from the website of the DOE Joint Genome Institute (http://www.jgi.doe.gov/genomes).

Nucleic acid manipulations

For the isolation of sporangia, zoospores and germinating cysts, *P. infestans* strain NL-88069 was grown at 18 ^{o}C in the dark on rye agar medium supplemented with 2% sucrose (RSA). Tissue collection, RNA isolation and Northern hybridization were performed as previously described by van West et al. (1998). Hybridization screening of the *P. infestans* bacterial artificial chromosome (BAC) library was carried out as described in Whisson et al. (2001). BAC contig building was performed as described by Jiang et al. (2005).

Database mining

All known ELI and ELL sequences of different *Phytophthora* species and *Pythium* species were retrieved from GenBank. The elicitin domains of all ELIs and ELLs were used to construct an elicitin domain database, from which PSSMs (Position Specific Scoring Matrices) were generated. Subsequently, PSI-BLAST was performed by screening EST databases and annotated protein databases with the PSSMs to identify new ELI and ELL candidates with an *E* value cut off of 0.05. PSI-BLAST hits were placed in a candidate data set. BLASTN of each individual candidate was performed against all identified elicitin nucleotide sequences. Hits with *E* value less than 10^{-30} were considered to be the known ELIs or ELLs and discarded, whereas hits with *E* value between 10^{-30} and 0.05 and the characteristic six cysteine signature were considered to be the new ELIs or ELLs and analyzed manually. These were then added to the screening process until no new ELIs or ELLs could be found.

Bioinformatics tools

Sequences were analyzed in Vector NTI 8 package. For BLAST searches we used the NCBI BLAST program and the Standalone-BLAST Version 2.2.3 (Altschul et al. 1997). Multiple sequence alignment was performed by ClustalX 1.8 and for phylogenetic tree construction Molecular Evolutionary Genetic Analysis 2.1 (MEGA) (Kumar et al. 2001) was used. Phylogeny reconstruction of ELI and ELL domains was performed by Neighbor-Joining analysis. Poisson Correction (PC) was chosen as the distance parameter as specified in the program MEGA. The inferred phylogeny was tested by 1,000 bootstrap replicates. Signal peptides were predicted by SignalP 2.0 (Krogh et al. 2001) and transmembrane domain prediction was performed with the program SOSUI (Hirokawa, Boon-Chieng, and Mitaku 1998). For GPI (glycosylphosphatidylinositol) anchor prediction, big-PI Plant Predictor (Eisenhaber et al. 2003) was used. Protein motifs were searched against the Prosite database (Bairoch 1991; Sigrist et al. 2002). Global d_N/d_S ratios were calculated by the fast diversifying/purifying selection detection program with Single Likelihood Ancestor Counting (SLAC) analysis (Pond and Frost 2005). Tests for purifying or diversifying selection were performed with the codeml program in the PAMLv3.14 package (Yang 1997; Yang et al. 2000). Models M0, M1a, M2a, M7, M8 were used for the analysis. Positively selected amino acid sites were assigned based on a probability > 95% with Bayes empirical Bayes (BEB) statistics (Yang, Wong, and Nielsen 2005) in model M2a. Calculation scripts were written in Python 2.2 (http://www.python.org) and are available from the authors upon request.

Results

Elicitins and elicitin-like proteins in *Phytophthora*

Previous studies in *P. infestans*, *P. sojae* and *P. brassicae* showed that elicitins are encoded by complex gene families. In *P. infestans*, seven *inf* genes have been cloned by either low stringency hybridization with heterologous probes, PCR amplification with degenerate primers or random sequencing of cDNAs, and were named *inf1, inf2a, inf2b, inf3, inf4, inf5, inf6* and *inf7* (Kamoun, Lindqvist, and Govers 1997; Kamoun et al. 1997; Kamoun et al. 1999). In *P. sojae*, Qutob et al. (2003) identified an elicitin gene family comprised of *sojA, sojB, soj2, soj3, soj5, soj6,* and three family members with diverse sequences named *soj7, sojX* and *sojY*. In *P. brassicae*, five members of the elicitin gene family have been described (L. Belbahri and F. Mauch, unpublished data). The proteins encoded by *bra1, bra2, bra5* and *bra6* share the highly conserved 98 amino acid elicitin domain while the elicitin domain in BRA7 is more diverse. In many *Phytophthora* species and a few *Pythium* species, one major elicitin has been identified as an abundantly secreted protein and the protein sequences of several of these have been deposited into Genbank (supplementary material table S2). Phylogenetic trees published previously distinguished

different elicitin classes. However, proteins with a diverse elicitin domain were not classified and referred to as elicitin-like (Kamoun, Lindqvist, and Govers 1997; Qutob et al. 2003).
To name elicitins and elicitin-like proteins in a systematic and consistent way, we propose a novel classification system and three letter abbreviations for individual proteins. The elicitins sharing a highly conserved 98 amino acid domain with six cysteine residues and a typical elicitin type cysteine spacing pattern are classified as ELIs, and they are labeled with the first three or four letters of the species name followed by a number, such as INF1, SOJ1, BRA1 and RAM1. Elicitin-like proteins possessing shorter or longer elicitin domains that are more diverse at the sequence level than the ELI elicitin domains, are classified as ELLs. The ELLs of *P. infestans* are named INL, of *P. sojae* SOL, of *P. brassicae* BRL and of *P. ramorum* RAL. Consequently, in this paper we renamed *P. sojae* SOJ7, SOJX and SOJY, *P. infestans* INF7, and *P. brassicae* BRA7 into SOL1A, SOL6, SOL3A, INL1 and BRL1B, respectively.

Diverse elicitin gene family members in *P. infestans*, *P. sojae, P. brassicae* and *P. ramorum*

Mining of 35,266 EST contig sequences comprised of transcripts derived from various developmental stages of *P. infestans* (Randall et al. 2005) revealed 11 new members of the elicitin gene family bringing the total number in *P. infestans* to 19. In a similar way, 13,234 *P. sojae* and 5863 *P. brassicae* EST contig sequences were mined resulting in six and four new members, respectively, resulting in totals of 16 in *P. sojae* and nine in *P. brassicae*. Except for SOJC and SOJ3B, all new members found in the EST libraries are ELLs, resulting in 21 new ELLs in total.

To analyze the diversity of elicitin gene families in *Phytophthora* at the whole genome level, we mined the assembled draft genome sequences of *P. sojae* and *P. ramorum*. In total, 18 SOJ and 39 SOL domains were found in *P. sojae* including the eight SOJs and seven SOLs extracted from the EST databases. In *P. ramorum,* 17 RAM and 31 RAL domains were found. As yet, there is no *P. ramorum* EST database available. A few *ell* genes (*sol2C*, *ral13A*, *sol13A* and *sol13H*) found in the genome sequences encode repeated elicitin domains but it is not known if these genes are active.

To investigate the presence of elicitins in other eukaryotic filamentous plant pathogens, 3021 unique EST sequences of *Blumeria graminis,* 513 unique EST sequences of *Cladosporium fulvum,* and 8821 unique EST sequences and the draft genome sequence of *Magnaporthe grisea* were analyzed using the same mining methods. In addition, the genome sequence of the marine diatom *T. pseudonana* which, similar to *Phytophthora,* belongs to the stramenopile lineage was used for mining. No ELI or ELL sequences were found in any of these species. Other plant pathogenic oomycetes may produce elicitins but to our knowledge this has not been reported.

Phylogenetic reconstruction of ELIs and ELLs

A total of 156 elicitin domains derived from 128 ELIs and ELLs from *P. brassicae, P. infestans, P. ramorum* and *P. sojae* (supplementary material table S1) and several additional ELIs identified in other *Phytophthora* species and two *Pythium* species (supplementary material table S2) were used to construct a phylogenetic tree (fig. 2). Seventeen distinctive clades with high bootstrap values (>60) could be identified and most show bootstrap values higher than 80. Four are ELI clades that together form a distinct branch. The remaining 13 are ELL clades. ELL clades are typically more divergent than ELI clades.

Orthologues are homologues separated by a speciation process, for example INF3 of *P. infestans* and SOJ3A of *P. sojae*. Paralogues are homologues generated by a gene duplication event, such as *P. sojae* SOJ3A and SOJ3B. From the tree it is clear that every clade containing more than two ELIs or ELLs is comprised of orthologues. All clades have ELIs or ELLs derived from *P. sojae* and *P. ramorum* and in most cases genes from these two species are over-represented. This is obviously due to the fact that the complete genome of these two species was sampled, including potential pseudogenes. All four ELI clades have a *P. infestans* member as have eight of the 13 ELL clades. The nine *P. brassicae* BRAs and BRLs are present in seven different clades. Of the 17 clades 13 have members of three or more species and this strongly suggests that the diversity in the elicitin gene family existed before these species evolved.

In the phylogenetic tree shown in figure 2 the ELI-1 clade is the largest with 32 members. However, only thirteen of these belong to the four species that we focus on in this paper. The other nineteen ELI-1 elicitins were identified in various other *Phytophthora* species and in *Pythium vexans* (supplementary material table S2). ELI-1 elicitins are easy to identify: the mature protein is just the 98 amino acid elicitin domain and they are the most abundantly secreted proteins in culture filtrates. The ELI-1 clade includes the previously identified class I-A, class I-B, class II and class Py elicitins (Kamoun, Lindqvist, and Govers 1997; Qutob et al. 2003). This sub-classification can be resolved when only the ELI-1 elicitins are used as input for phylogenetic tree construction (data not shown).

Interestingly, *P. infestans* INF4 has no apparent orthologue in *P. sojae* or *P. ramorum* and is thus not covered by any clade. Also OLI, an elicitin from *Pythium oligandrum* with a highly divergent elicitin domain, is not assigned to any specific clade. SOJ3X and RAM3X share sequence homology with ELI-2 and ELI-3 clades but the apparent orthologues cannot be found. Therefore, neither SOJ3X nor RAM3X is covered by a clade. For similar reasons, no specific clade was assigned to SOL1E or SOL11F.

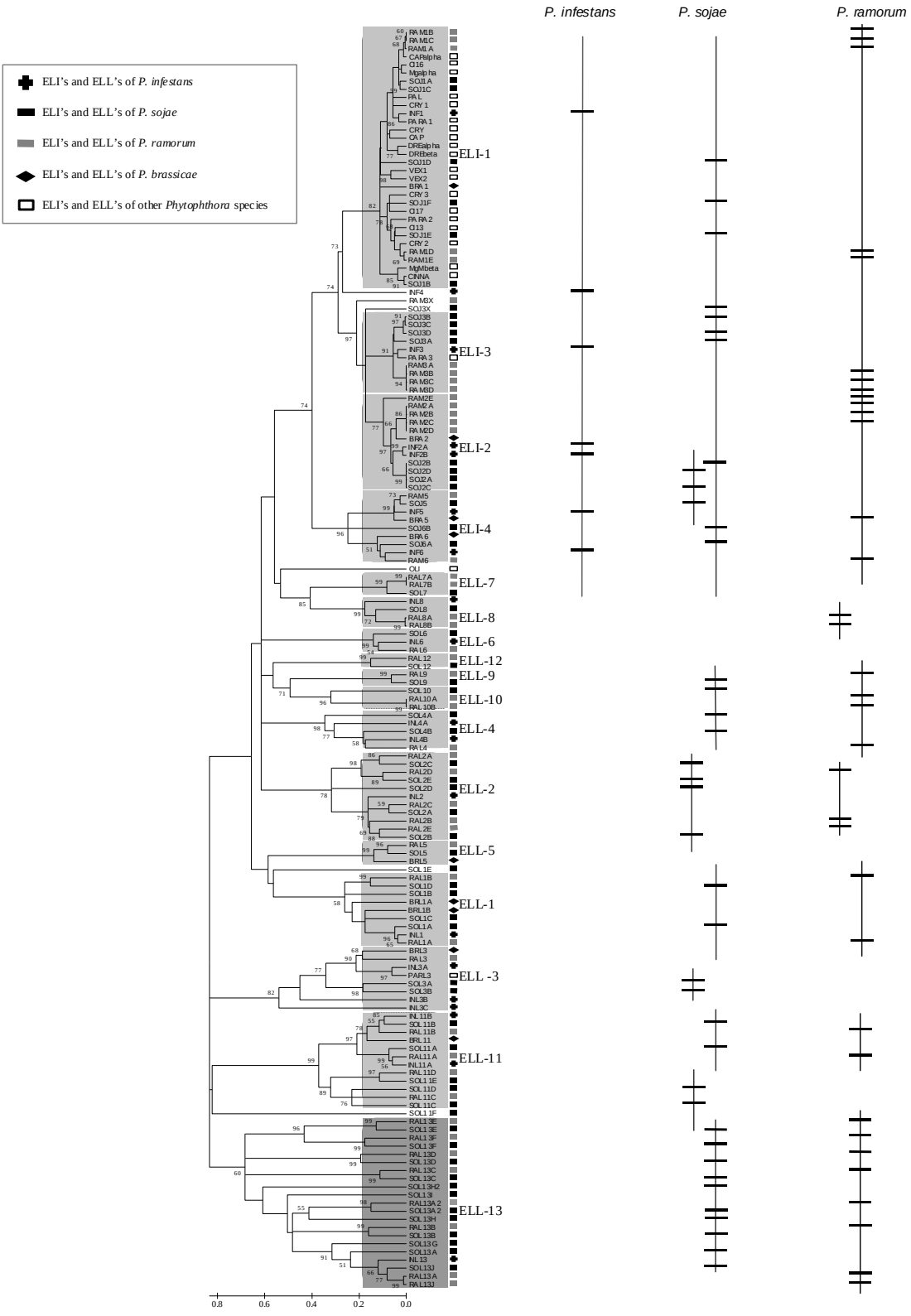


Fig. 2. — Phylogram of ELI and ELL amino acid sequences from *P. brassicae, P. infestans, P. ramorum, P. sojae* and other *Phytophthora* spp. The conserved elicitin domains were used to construct the unrooted phylogram based on Neighbor-Joining analysis. Confidence of groupings was estimated by using 1,000 bootstrap replicates; numbers next to the branching point indicate the percentage of replicates supporting each branch. The shaded blocks show the ELIs or ELLs belonging to the same clade. Clade ELL-13 with the most diverse sequences is shaded with a grid pattern. On the right the physically linked *eli* and *ell* genes are schematically drawn. The length of the lines does not represent the physical distance. A list of ELIs and ELLs derived from other *Phytophthora* species is available as supplementary material.

The 13 ELL clades show large sequence diversity within members of the same clade as well as between members of other ELL clades. The ELL-13 clade is the largest clade (18 members) and comprises ELLs with the most diverse elicitin domains.

ELI and ELL clades are highly conserved across *Phytopthphthora* species and are under purifying selection

The 17 ELI and ELL clades belong to the most conserved elicitors identified in *Phytophthora*. A BLAST search of the *P. sojae* draft genome sequence with *P. infestans* ELIs and ELLs and a set of (putative) elicitors of *P. infestans* resulted in similarity matches with a broad range of *E* values. By plotting the BLAST identity percentages the level of conservation of the genes between *P. infestans and P. sojae* can be visualized (fig. 3). Together with the highly conserved NLP protein NPP1 (a Group-A elicitor), ELIs and ELLs (Group-B elicitors) are more conserved than all other elicitors. Other Group-B elicitors, CRN1 and CRN2 (Torto et al. 2003), are undergoing expansion and gene loss in *Phytophthora* species (R.H.Y. Jiang et al., unpublished data) and they show less sequence conservation than ELIs and ELLs. The Group-C elicitors (SCR74, AVR3a and IPI-O) are highly divergent between *P. infestans* and *P. sojae* and thus belong to the least conserved elicitors. The ELI INL4A is the previously described mating associated factor M-25 (Fabritius, Cvitanich, and Judelson 2002) and in the analysis it appears to be the least conserved elicitin family member. Interestingly, family members of another mating associated secreted protein, M-96 (Fabritius, Cvitanich, and Judelson 2002), are highly divergent (J.H.Y. Jiang, unpublished results).

Table 1
Selection test based on codeml model M2a[a] of PAMLv3.14.

gene (sub)family[b]	number of sequences[c]	positively selected sites[d]	p_0	ω_0	p_1	ω_1	p_2	ω_2	κ[e]	overall ω[f]
eli-1	13	none	0.92	0.08	0.08	1.00	0.00	-	2.08	0.22
eli-2	11	none	1.00	0.09	0.00	1.00	0.00	5.29	2.12	0.14
eli-3	9	none	0.98	0.03	0.00	1.00	0.02	1.83	3.81	0.26
eli-4a	4	none	0.94	0.04	0.04	1.00	0.02	16.76	2.94	0.11
eli-4b	4	none	0.85	0.06	0.15	1.00	0.00	-	2.00	0.16
scr74	20	8	0.73	0.15	0.00	1.00	0.27	9.19	3.35	2.28

[a] The site model allows the ω ratio to vary among sites (among codons or amino acids in the protein) (Yang et al. 2000). $\omega_0 < 1$ and $\omega_2 > 1$ are estimated from the data while $\omega_1 = 1$ is fixed.
[b] For the calculation only the sequences encoding the conserved elicitin domain and signal peptide were included.
[c] For *eli-1* only the sequences derived from *P. brassicae, P. infestans, P. ramorum* and *P. sojae* were used. The pseudo-gene *inf3* was not included in the *eli-3* sequences. *eli-4a* consists of *bra5, ram5, inf5* and *soj5*. *eli-4b* consists of *bra6, inf6, ram6* and *soj6*. A set of 20 randomly selected *scr74* genes (Liu et al. 2005) was used for comparison.
[d] Positively selected amino acid sites were assigned based on a probability > 95% with Bayes empirical Bayes (BEB) statistics (Yang, Wong, and Nielsen 2005).
[e] κ is the estimated transition/transversion rate parameter.
[f] The overall ω value is based on SLAC (Single Likelihood Ancestor Counting) analysis (Pond and Frost 2005).

The survival of orthologues after speciation is due to selection pressure exerted on the genes. For a protein-coding gene, selection is estimated by comparing the rate of nonsynonymous nucleotide substitutions per nonsynosymous sites (d_N, amino acid replacing) and synonymous nucleotide substitution per synosymous sites (d_S, silent). The ratio d_N/d_S, denoted by ω, is used as a measure of selective pressure at the protein level and ω values of 1, < 1 and ω > 1 indicate neutral, purifying and diversifying selection, respectively. By using the fast diversifying/purifying selection detection program DataMonkey with SLAC (Single Likelihood Ancestor Counting) analysis (Pond and Frost 2005), all 17 clades of *elis* and *ells* show overall ω values lower than 1 (table 1) which indicates purifying selection. The highly conserved *eli* clades were also analysed with the codeml program of the PAML (Phylogenetic Analysis by Maximum Likelihood) package developed by Yang et al (1997, 2000). All *eli* clades show ω_0 value ranging from 0.02 to 0.09 and no positively selected sites could be detected (table 1). Moreover, in the highly similar *soj2, soj3, ram2* and *ram* 3 groups within the *eli-2* and *eli-3* clades no positive selection was found. For comparison we subjected *P. infestans scr74* sequences to the same analysis. In contrast to the conserved *elis* and *ells*, *scr74* is a Group-C elicitor with only weak homologues in other *Phytophthora* species. The *scr74* gene family was shown to be under diversifying selection in *P. infestans* (Liu et al. 2005) and 8 positive selection sites identified by Liu et al. (2005) were also detected with the more stringent criteria that we used in this study (table 1). These results suggest that in the different *Phytophthora* species the complex elicitin family is maintained by purifying selection.

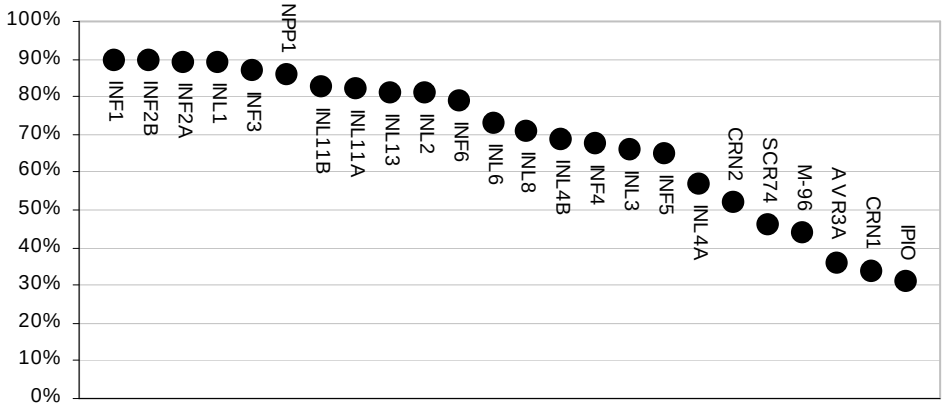


Fig. 3. — Conservation of ELIs and ELLs between *P. infestans* and *P. sojae*. The percent identity of *P. infestans* ELIs and ELLs to *P. sojae* is plotted and indicated on the Y-axis. The identity scores of several previously identified elicitor(-like) genes and gene family members are included for comparison: NPP1 (AAK25828), CRN2 (AAN31502), SCR74 (AAU21463), M-96 (AAN37691), AVR3a (CAI72254), CRN1 (AAN31500) and IPI-O (AAA21422). Signal peptides were ommitted from the analysis. The labels of ELIs and ELLs are positioned below the data points and those of the other elicitor(-like) proteins above the data points.

Elicitin domains show clade-specific cysteine spacing patterns

The elicitin domains of ELIs and ELLs are of variable length but they all contain six cysteine residues at conserved positions. The six cysteines form three disulfide bonds that stabilize the α-helix folded protein (Boissy et al. 1996; Fefeu et al. 1997). Based on the cysteine spacing pattern, the ELIs and ELLs can be

classified in distinct groups and these groups coincide with the classification in clades based on the phylogenetic reconstruction (table 2).

Table 2
The spacing pattern of the six cysteine residues present in the elicitin domain in ELIs and ELLs belonging to different clades.

name [a]	domain size (aa)	clades	cysteine spacing pattern
BRA1, INF1, RAM1A, RAM1B, RAM1C, RAM1D, RAM1E, SOJ1A, SOJ1B, SOJ1C, SOJ1D, SOJ1E, SOJ1F	98	ELI-1	C-23-C-23-C-4-C-14-C-23-C CxxxxxxxxxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxxxxxxxxxC
BRA2, INF2A, INF2B, RAM2A, RAM2B, RAM2C, RAM2D, RAM2E, SOJ2A, SOJ2B, SOJ2C, SOJ2D	98	ELI-2	C-23-C-23-C-4-C-14-C-23-C CxxxxxxxxxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxxxxxxxxxC
INF3, RAM3A, RAM3B, RAM3C, RAM3D, SOJ3A, SOJ3B, SOJ3C, SOJ3D	98	ELI-3	C-23-C-23-C-4-C-14-C-23-C CxxxxxxxxxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxxxxxxxxxC
BRA5, INF5, RAM5, SOJ5, BRA6, INF6, RAM6, SOJ6A, SOJ6B	98	ELI-4	C-23-C-23-C-4-C-14-C-23-C CxxxxxxxxxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxxxxxxxxxC
BRL1A, BRL1B, INL1, RAL1A, RAL1B, SOL1A, SOL1B, SOL1C, SOL1D	85	ELL-1	C-16-C-22-C-4-C-14-C-18-C Cxxx……xxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxx……...xxxxC
INL2, RAL2A, RAL2B, RAL2C, RAL2D, RAL2E, SOL2A, SOL2B, SOL2C, SOL2D, SOL2E	88	ELL-2	C-16-C-22-C-4-C-14-C-21-C Cxxx……xxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxx…..xxxxC
BRL3, INL3A,INL3B,INL3C, RAL3, SOL3A, SOL3B	87,88	ELL-3	C-16-C-22/23-C-4-C-14-C-20-C Cxxx……xxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxx…..xxxxC
INL4A, INL4B, RAL4, SOL4A, SOL4B	92	ELL-4	C-20-C-22-C-4-C-15-C-20-C Cxxxxx…xxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxx…..xxxxC
BRL5, RAL5, SOL5,	89	ELL-5	C-16-C-23-C-4-C-14-C-21-C Cxxx……xxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxx…..xxxxC
INL6, RAL6, SOL6	91	ELL-6	C-17-C-24-C-4-C-14-C-21-C Cxxx……xxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxx…..xxxxC
RAL7A, RAL7B, SOL7	92	ELL-7	C-19-C-23-C-4-C-14-C-21-C Cxxx……xxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxx…..xxxxC
INL8, RAL8A, RAL8B, SOL8	87,93	ELL-8	C-19-C-24-C-4-C-14-C-15/21-C Cxxx……xxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxx……...xxxxC
RAL9, SOL9	91	ELL-9	C-18-C-23-C-4-C-14-C-21-C Cxxx……xxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxx…..xxxxC
RAL10A, RAL10B, SOL10	91	ELL-10	C-18-C-23-C-4-C-14-C-21-C Cxxx……xxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxx…..xxxxC
BRL11, INL11A, INL11B, RAL11A, RAL11B, RAL11C, RAL11D, SOL11A, SOL11B, SOL11C, SOL11D, SOL11E	93,98	ELL-11	C-20-C-25/28-C-4-C-12/14-C-21-C Cxxx……xxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxx…..xxxxC
RAL12, SOL12	92	ELL-12	C-20-C-22-C-4-C-14-C-21-C Cxxx……xxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxxxx…..xxxxC
INL13, RAL13A, RAL13A2, RAL13B, RAL13C, RAL13D, RAL13E, RAL13F, RAL13J, SOL13A, SOL13A2, SOL13B, SOL13C, SOL13D, SOL13E, SOL13F, SOL13G, SOL13H, SOL13H2, SOL13I, SOL13J	75,78,80,81,82	ELL-13	C-20-C-(12-17)-C-4-C-13-C-17/18-C Cxxx……xxxxxCxxxxxxxxxxxxCxxCxxxxxxxCxxx……...xxxxC

[a] The shaded ELIs and ELLs represent ESTs

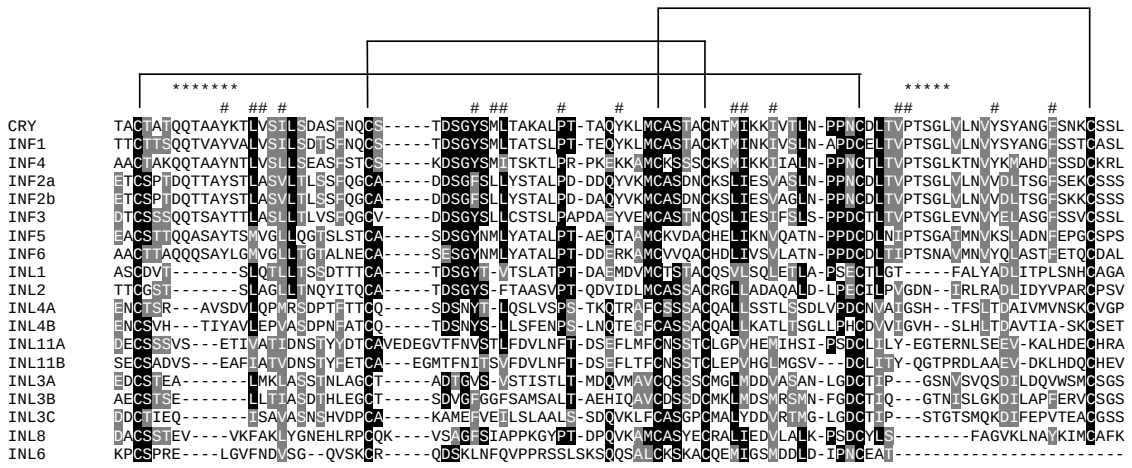


Fig. 4.— Multiple sequence alignment of *P. infestans* ELIs and ELLs with the ELI-1 elicitin CRY from *P. cryptogea*. The sequence alignment was generated from the conserved elicitin domains. From INL6 only an incomplete gene sequence was present in the EST database. The three lines shown above the alignment connect the cysteine residues that form disulfide bonds in CRY. * indicates the residues in CRY that correspond to the gaps in the alignment. # indicate the residues that interact with ergosterol (Boissy et al. 1999).

All four ELI clades fall in one group and have a cysteine spacing pattern of C_1-23-C_2-23-C_3-4-C_4-14-C_5-23-C_6. In contrast, each ELL clade has a typical cysteine spacing pattern. The number of amino acid residues between C_3 and C_4 (i.e. four) is conserved in all groups whereas the two most variable regions are between C_1 and C_2, and between C_5 and C_6. Thus the cysteine spacing pattern of the overall elicitin family can be visualized as C_1-variable-C_2- 23-C_3-4-C_4-14-C_5-variable-C_6. The three disulfide bonds are formed between C_1 and C_5, C_2 and C_4, and C_3 and C_6. The two variable regions correspond to the two alignment gaps in a multiple sequence alignment as shown for *P. infestans* INFs and INLs in figure 4. In the sequence alignment the region between C_2 and C_4 is most conserved.

Most ELIs and ELLs have C-terminal domains with typical repeat structures and GPI anchors

Most ELIs and ELLs are predicted to possess a signal peptide at the N-terminus in front of the conserved elicitin domain and an extended C-terminal domain following the conserved domain (supplementary material table S1). ELIs and ELLs in fourteen of the seventeen clades have C-terminal domains ranging in length from 17 to 291 amino acids whereas ELL-7, ELL-9 and ELL-10 members have shorter C-terminal domains of up to seven amino acids. The majority of the ELI-1 proteins lack a C-terminal domain. Only 11 out of the 32 known ELI-1 elicitins have a short C-terminal tail and in the initial ELI-1 sub-classification these were classified as class II. Hence, most ELI-1 proteins are comprised solely of a signal peptide and the conserved 98 amino acid elicitin domain. In *P. sojae* and *P. ramorum* the ELI-1 clade has several members but *P. infestans* has only one. In contrast, *P. infestans* has a second elicitin without a C-terminal domain, i.e. INF4, which has no orthologues in the other species.

Many of the C-terminal domains appear to have a biased amino acid composition. They are particularly rich in threonine, serine and proline residues and quite often these residues are part of a repeat. Despite

the fact that the phylogenetic tree of the elicitin family (fig. 2) was constructed with only the conserved elicitin domain, the C-terminal domains of ELIs and ELLs show clade-specific features, not only in the amino acid composition but also in the repeat structure. Table S1 (supplementary material) summarizes the features of C-terminal domains of 129 ELIs and ELLs and in figure S1 (supplementary material) for each clade an example of a C-terminal domain is shown with the three most abundant amino acids highlighted. In the C-terminal domains of several of the ELI-2, ELI-3, ELI-4, ELL-1, ELL-2 and ELL-13 proteins more than 40% of the amino acid residues is comprised of threonine and serine. The C-terminal domains of ELI-4 and ELL-8 proteins are rich in proline, an amino acid that does not have a backbone proton and can easily form turns in the secondary protein structure (figure S1, supplementary material). In many of the C-terminal domains repeat units can be recognized such as the 'APSAE' repeat unit in BRA5 and the 'SA' repeats in INL2. They are comprised of two to five amino acids and can be repeated up to 15 times. The presence of several O-GalNAc-glycosylation sites as predicted by the program NetOGlyc 3.1 (Julenius et al. 2005) suggests that the C-terminal domains are glycosylated.

Several classes of ELLs such as ELL-1, ELL-2 and ELL-13 seem to possess hydrophobic regions at the extreme C-terminal end as predicted by the program SOSUI (Hirokawa, Boon-Chieng, and Mitaku 1998). These hydrophobic regions are part of the glycosylphosphatidylinositol (GPI) anchor site predicted by the program big-PI plant predictor (Eisenhaber et al. 2003). In proteins carrying such a motif the hydrophobic C-terminal end is cleaved off from the mature protein and in stead a GPI is added that will anchor the protein to the plasma membrane.

Clustering of *eli* and *ell* genes in the genomes

In *P. infestans,* seven *eli* genes (*inf1, inf2a, inf2b, inf3, inf4, inf5* and *inf6*) and one *ell* gene (*inl1*) were shown to be single copy genes by genomic Southern blot hybridization (data not shown). One physical contig of 250 kb spanning the seven *inf* genes was obtained by BAC library screening and contig building (fig. 5A). Sequencing and annotation of one of the BACs containing four *inf* genes showed that the average spacing between *inf* genes and other genes is 20 kb (Jiang et al. 2005). Also in *P. sojae* and *P. ramorum, eli* genes were found to be clustered in the genome. In *P. sojae*, one contig of 115 kb containing 15 *eli* genes and one small contig with three *eli* genes could be identified (fig. 2 and fig. 5A). In *P. ramorum* a region of 59 kb with 14 *eli* genes was found (fig. 2 and fig. 5A). The other *soj* and *ram* genes were found on other small scaffolds but it cannot be excluded that some genes are on different scaffolds because of gaps in the draft genome sequence.

Not only *eli* genes, but also *ell* genes were found to be clustered in the genomes of *P. sojae* and *P. ramorum.* For example, *sol4a, sol4b, sol9* and *sol10* are located within a 123 kb sequence contig in *P. sojae* while *ral4, ral10a, ral10b* and *ral9* map within a 97 kb sequence contig in *P. ramorum* (fig. 2 and fig. 5B). Another example is the clustering of genes encoding ELLs belonging to clade ELL-13. In *P. sojae* and *P. ramorum*, five *sol13* and six *ral13* genes, respectively, were found to be clustered in a 31

kb sequence contig in both species (fig. 2 and fig. 5C). Other examples of *ell* gene clustering are shown in figure 2. In *P. sojae* and *P. ramorum* often *ell*'s belonging to the same clade are clustered suggesting that these *ell* genes are paralogues that resulted from gene duplication.

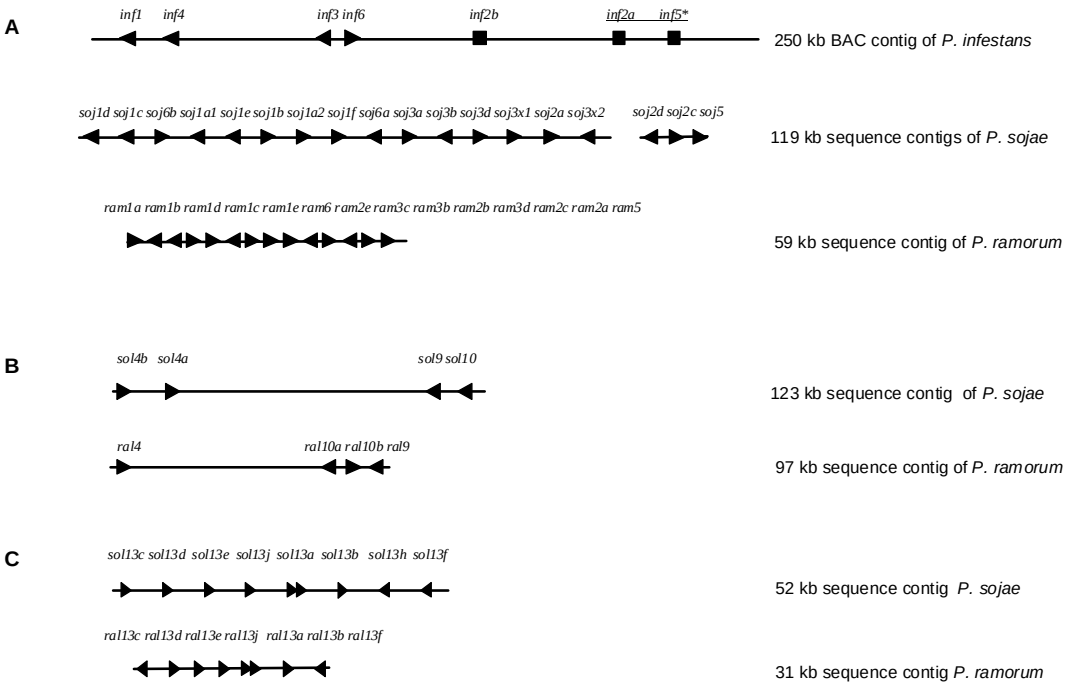


Fig. 5. — Clustering of elicitin genes in *P. ramorum*, *P. infestans* and *P. sojae*. Black arrows indicate *eli* and *ell* genes and their orientations. Black squares indicate genes with unidentified orientation. Thin horizontal lines represent DNA contigs. * The order of *inf2A* and *inf5* has not been determined. *(A) eli* contigs. *(B) ell* contigs containing genes encoding ELL-4, ELL-9 and ELL-10 proteins. *(C) ell13* contigs. *sol13a* and *ral13a* have repeated elicitin domains.

Eli and *ell* genes in the same clade show similar expression patterns

Members of gene families are often differentially expressed in space and time. To see whether that is also true for the different members of the complex elicitin gene family we examined tissue-specific expression patterns as well as expression levels of different family members by transcript counting based on EST databases and by northern blot hybridization.

For *P. infestans* a large collection of over 75,000 ESTs is available (Randall et al. 2005) and for *P. sojae* a collection of 21,282 ESTs is available (B. Tyler et al. unpublished; http://staff.vbi.vt.edu/estap). Transcript counting showed that elicitin genes are differentially expressed in zoospores, sporangia and mycelia of *P. infestans* and *P. sojae* (table 3). These differential expression patterns are specific for a particular phylogenic clade of *eli* or *ell* genes. For example, ESTs from the *inf1* and *soj1* genes in the ELI-1 clade are strongly represented in mycelium libraries, as well as libraries from material that include substantial amounts of mycelia, such as mating cultures (*P. infestans*) and infected plant tissue (*P. sojae*). In contrast, ESTs from genes in the ELL-3 clade are primarily found in zoospore libraries of both *P. infestans* and *P. sojae*. A northern blot with RNA isolated from mycelia, sporangia, zoospores, cysts and germinating cysts from *P. infestans* and hybridized with *inf1, inf2A, inf2B, inf4, inf5, inf6* and *inl1*

confirmed the findings of the transcript counting in *P. infestans* (fig. 6A). The differential expression of ELI-1 and ELL-3 elicitin genes was also observed experimentally in a third species, *P. parasitica*. Northern blot analysis showed that the ELI-1 elicitin *para1* is only expressed in mycelium whereas *parl3,* which belongs to the ELL-3 clade, is expressed in zoospores and germinating cysts (fig. 6B). In *P. sojae*, several *eli* and *ell* genes are strongly expressed during infection of soybean, compared to their expression in mycelia, such as *soj1c*, *soj5* and *soj6*, and especially *sol2a* and *sol2b*.

Table 3
Expression patterns of *eli* and *ell* genes represented by the number of transcripts present in *P. infestans* and *P. sojae* EST databases from different life stages.

Sequence[a]	transcripts per 10,000 in cDNA libraries[b] from				
	Zoospores[c]	sporangia	mating	mycelia	infection
inf1	0.8	8.3	65.6	61.8	na[d]
soj1a1; 1a2[e]	0	na	na	7.8	18.2
soj1b	0	na	na	15.6	2.8
soj1c	0	na	na	6.7	11.2
inf2a	1.7	3.5	17.4	11.3	na
inf2b	1.7	3.5	18	11.6	na
soj2a; 2b; 2d[e]	0	na	na	3.4	1.4
soj2c1; 2c2[e]	0	na	na	12.3	11.2
inf3	0	0	1.9	0.3	na
soj3a	0	na	na	5.6	1.4
soj3c	0	na	na	3.4	1.4
inf4	0.8	0	1.9	2.7	na
inf5	0	3.5	36	27.9	na
soj5	0	na	na	3.4	12.6
inf6	1.7	2.4	96.5	58.8	na
soj6a	1.9	na	na	7.8	14.0
inl1	0	4.7	2.6	4.8	na
sol1a	1.9	na	na	1.1	1.4
inl2	0.8	1.2	0	0.9	na
sol2a	0	na	na	0	11.2
sol2b	0	na	na	0	12.6
inl3a	1.7	0	0	0	na
inl3b	1.7	1.2	0	0	na
inl3c	6.7	1.2	0	0	na
sol3a	34.7	na	na	0	2.8
sol3b	1.9	na	na	0	0
inl4a	0	0	0.6	0	na
inl4b	0	1.2	0	0.6	na
sol5	0	na	na	1.1	0
inl6	0	1.2	0	0	na
sol6	1.9	na	na	0	0
inl8	0	0	0.6	0	na
sol8	0	na	na	0	1.4
sol9	0	na	na	0	1.4
sol10	0	na	na	0	1.4
inl11a	0	0	0	0.3	na
inl11b	1.7	0	0.6	2.1	na
sol11a	0	na	na	2.2	0
sol11b	0	na	na	0	2.8
inl13	0	4.7	0	0.9	na
sol13e	0	na	na	2.2	0

[a] An individual *eli* or *ell* sequence was searched against the databases using BLASTN with *E* value cutoff of 1e-100. The following *P. sojae* genes, identified from the genome sequence, had no matches to ESTs in the database: *soj1d-f; soj3b; soj3d; soj3x; soj6b; sol1b-e; sol2c-e; sol4a-b; sol7; sol11c-f; sol12; sol13a-d; sol13f-j*

[b] The *P. infestans* EST database contained 11,919 sequences from zoospores and germinated cysts, 8470 from sporangia, 5541 from mating cultures and 33,651 from mycelia grown under various conditions. The infection library available for *P. infestans* was too small for a reliable counting. The *P. sojae* EST database contained 5187 sequences from zoospores and germinated cysts, 8949 from mycelia grown under various conditions and 7146 *P. sojae* sequences from infected soybean tissue (Qutob et al. 2000). No libraries from sporangia or mating cultures were available from *P. sojae*.

[c] Zoospore libraries from both species included cDNAs from both free-swimming zoospores and germinated cysts.

[d] na = not available.

[e] Genes *soj1a1* and *soj1a2* encode identical mRNA sequences and so their ESTs were counted together. The same was true for the three genes *soj2a*, *soj2b* and *soj2d*, as well as the two genes *soj2c1* and *soj2c2*.

The expression levels of *eli* and *ell* genes also differ widely. For some, like *inl6,* only one transcript was found in the entire EST database while for others, like e.g. *inf6,* more than 100 transcripts were present. In *P. sojae*, many *soj* and *sol* genes predicted in the genome sequence have no ESTs at all. In *P. infestans inf1, inf5* and *inf6* are among the most abundantly expressed genes in mycelium (Kamoun et al. 1999). Overall, the expression levels of *eli* genes seem to be higher than those of *ell* genes with the exception of *sol3a* which is one of the most strongly expressed genes in *P. sojae* zoospores.

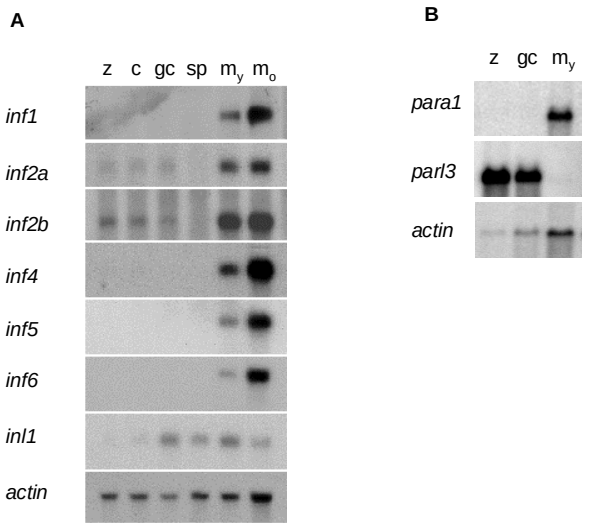


Fig. 6.—*(A)* Autoradiographs of a northern blot containing RNA isolated from *P. infestans* zoospores (z), cysts (c), cysts germinated in water for 2.5 hours (gc), sporangia (s), a young mycelial culture started from sporangia that were allowed to germinate for 20 hours at 18º C (m_y) and old mycelium (m_o) and hybridized with various elicitin *P. infestans inf* and *inl* probes and an actin probe as loading control. *(B)* Autoradiographs of a northern blot containing RNA isolated from *P. parasitica* zoospores (z), germinated cysts (gc) and young mycelium (m_y) and hybridized with *P. parasitica para1* and *parl3* probes, and an actin probe as loading control.

The difference in expression levels between *eli* and *ell* genes can also be related to the difference in the GC3 (third position codon) content in the coding regions. In a genome wide survey of *P. sojae* we found that highly expressed genes usually have a higher GC3 than lowly expressed genes (R.H.Y. Jiang and F. Govers, submitted). In *P. brassicae*, *P. infestans*, *P. ramorum* and *P. sojae*, the average GC3 of 46 *eli*'s is 86%, which is higher than the 75% of 82 *ell*'s. The higher GC3 percentage of *eli* genes agrees with the higher expression levels of *eli* genes as compared to *ell* genes.

Discussion

Elicitins are unique and ubiquitous in *Phytophthora* and *Pythium*

Elicitin genes are present in all examined *Phytophthora* species. The encoded proteins are ubiquitous across the whole *Phytophthora* genus with ELI-1 elicitins as the most abundant component in *Phytophthora* culture filtrates. However, the elicitins seem to be limited to the oomycetes *Phytophthora* and *Pythium.* Searches in GenBank and Pfam protein domain database (Sonnhammer et al. 1998) did not reveal any other organism that has elicitin-like sequences. In this study, a thorough search was

performed on EST databases from a few plant pathogenic ascomycete fungi and on the genome sequence of the rice blast fungus and a marine diatom, but no elicitin-like sequences were found.

In plants there is a group of proteins called nsLTPs (Non-Specific Lipid Transfer Proteins) that in some aspects resemble the elicitins (Blein et al. 2002). Similar to elicitins, nsLTPs are small secreted cysteine-rich proteins that interact with lipids. They facilitate the transfer of lipids between natural or artificial membranes (Kader 1996) and have been implicated to play roles in protection and defense (Buhot et al. 2001). However, the cysteine spacing patterns of these two classes of proteins are very different and therefore it is unlikely that elicitins and nsLTPs share a close phylogenetic relationship. Secreted proteins with a nsLTP-like cysteine spacing pattern cannot be found in the *P. infestans* EST database or the unigene sets of *P. sojae* and *P. ramorum*.

The elicitin gene family is an ancient family within the *Phytophthora* genus

In this study we focused on three *Phytophthora* species that differ from each other in various traits, such as host range (narrow or broad), sexual behaviour (homo- or heterothallic) and genome size (ranging from 65 Mb to 240 Mb). The ITS (rRNA internal transcribed spacer) based phylogenetic tree (Cooke et al. 2000) shows that *P. infestans* and *P. sojae* are located on widely divergent branches, indicating early species diversification during the evolution of the genus. In a phylogenic tree based on β-tubulin and EF-1α sequences *P. brassicae*, *P. infestans*, *P. sojae* and *P. ramorum* also fall in four different clades (Kroon et al. 2004).

The examined ELIs and ELLs were derived from the whole genome sequence of *P. sojae* and *P. ramorum* and from a large *P. infestans* EST collection and a smaller set of *P. brassicae* EST's. Thus the phylogenetic tree should cover all members present in *P. sojae* and *P. ramorum* and the majority of the ELIs and ELLs *in P. infestans*. From *P. brassicae* only a limited number of ELIs and ELLs were available for inclusion in the tree.

From the relationships represented by the phylogenic tree, it is evident that most clades possess family members from the three different species which are well-represented in the data set, indicating that the genes founding these clades had already diverged before the common ancestor gave rise to *P. infestans, P. sojae* and *P. ramorum*. Since in most cases, all clades of ELIs and ELLs are preserved in the three species, we infer that the different clades of these proteins have distinct functions. This hypothesis is supported by the conserved tissue-specific expression of members of some clades, most notably the zoospore-specific expression of ELL-3 genes in three different species. Within individual clades however, there has been extensive divergence of the genes in each species and there is no evidence for diversifying selection.

In *P. infestans*, *P. cinnamomi* and *P. cryptogea* the genes of the ELI-1 clade were reported to form a gene cluster within the genome (Panabieres et al. 1995; Duclos et al. 1998; Jiang et al. 2005). In this

study we showed that also *eli* and *ell* genes belonging to other clades occur in clusters. Within the *eli* clusters few other genes are present (Jiang et al. 2005 and data not shown), and therefore the *eli* gene density can be compared. Interestingly, the gene density is different in the different species.In *P. infestans* the average density in this region is one *eli* per 22.5 kb. In *P. sojae,* the average gene density in a contig containing 15 *eli* genes is one *eli* per 7.7 kb whereas in *P. ramorum* the gene density is even higher: one *eli* per 4.2 kb in a contig containing 14 *eli* genes. The species with the smallest genome size, i.e. *P. ramorum* (65 Mb), shows a higher *eli* density, whereas in the much larger *P. infestans* genome (240 Mb) mobile elements fill the spaces in between *eli* genes (Jiang et al. 2005).

Assuming that the ELIs and ELLs are all members of a single gene family that arose from a series of duplication events, then the primary gene duplication should be an event that occurred in the common ancestor. Subsequent duplications of individual family members then occurred after speciation, for example within the ELI-1 clade for *P. sojae* and *P. ramorum*. In *P. infestans*, the clustered *eli* genes show similar expression patterns with a high level of expression in mycelium, thus it is conceivable that the clustering is important for simultaneous expression of these *eli* genes. The chromatin environment can be crucial for the establishment and maintenance of transcriptional activation or repression via histone and DNA modifications (Lusser 2002). A detailed analysis of the promoter regions and putative regulatory elements may reveal why several members of this ancient gene family are clustered in *Phytophthora* genomes.

The intrinsic function and subcellular location of ELIs and ELLs

The ELI-1 clade elicitins are well known for their HR-inducing activity on *Nicotiana* species. Silencing of the *inf1* gene in *P. infestans* resulted in strains that were able to colonize the non-host *N. benthamiana* and in this interaction INF1 protein apparently acts as a determinant of non-host resistance (Kamoun et al. 1998). Also, members of the ELI-2 clade have HR-inducing activity as demonstrated for INF2A and INF2B (Huitema et al. 2005), and SOJ3 and SOJ6 (Qutob et al. 2003). From the ELL-clades only a few members have been tested for HR-inducing activity (Qutob et al. 2003; J.S. Marshall, A.R. Hardham and F. Govers, unpublished data) but so far no such activity was found. Similar to ELIs, ELLs have a structurally conserved six-cysteine elicitin domain. However, the intervening amino acid sequences in ELLs exhibited more diversity than those in ELIs. The multiple sequence alignment of ELIs and ELLs from *P. infestans*, showed that two regions are completely or partly deleted from ELLs as compared to ELIs (fig. 4) and this also results in the distinct cysteine spacing patterns of the various ELI and ELL clades as shown in table 2. In the 3D structure of cryptogein, an ELI-1 from *P. cryptogea* (Mikes et al. 1997), residues of these variable regions (QQTAAY and PTSGL) are located at the surface. This suggests that the surface exposed residues in ELIs may be variable or deleted due to selection pressure, possibly imposed by host plants.

Since the HR in plants blocks the growth of biotrophic pathogens, the HR-inducing activity is certainly not the primary function of ELIs in *Phytophthora.* Based on biochemical data it is now generally accepted

that the intrinsic biological function of ELIs is related to lipid binding and/or processing. Cryptogein, for example, binds sterols, such as ergosterol, and functions as a sterol-carrier protein (Mikes et al. 1997; Vauthrin et al. 1999), whereas ELI-4 members in *P. capsici* were shown to have phospholipase activity (Nespoulous et al. 1999). *Phytophthora* species cannot synthesize sterols, but still do require sterols for several physiological functions. Expression analysis showed that the ELI-1 members are highly expressed in mycelium, perhaps correlated with the requirement to acquire sterols from the environment during vegetative growth. Since other lifecycle stages such as sporangia, zoospores, and cysts are short-lived and primarily involved in pathogen dispersal and host invasion, the requirement for sterols may be less. The intrinsic functions of ELLs are unknown. The majority of the residues in cryptogein that are known to be involved in sterol binding are divergent in ELLs (fig. 4) thus bringing into question the sterol binding capacity of ELLs. One of the ELL-4 members however, was postulated to function in relation to the sterol-like mating hormones in *Phytophthora.* INL4A is identical is to the previously described mating associated factor M-25 and is encoded by a gene that is specifically expressed and highly induced in mating cultures during sexual development (Fabritius, Cvitanich, and Judelson 2002).

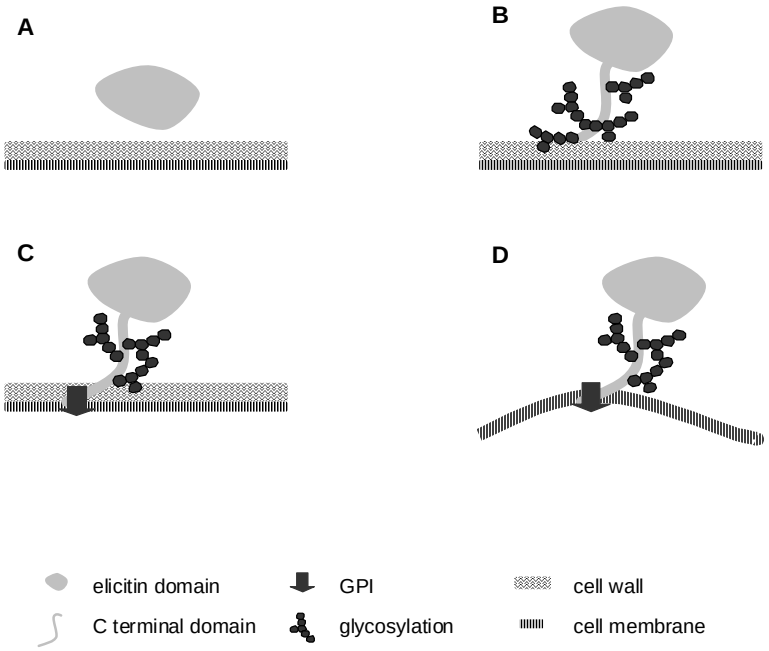


Fig. 7. — Schematic representation of ELIs and ELLs. *(A)* ELIs secreted in the culture filtrate. *(B)* ELIs and ELLs that are hypothesized to be linked to the cell wall by extensive glycosylation of the C-terminal domain. *(C)* ELLs hypothesized to be anchored to the cell membrane by GPI. *(D)* ELLs hypothesized to be anchored to the cell membrane of wall-less zoospores by GPI.

All ELIs and ELLs possess a signal peptide and the majority is likely to be associated with the cell wall or anchored to the cell membrane. Figure 7 gives a schematic representation of the various ELIs and ELLs based on the presence or absence of a C-terminal domain, on typical features in the C-terminal domain and on stage specific expression patterns. The ELI-1 elicitins are extracellular proteins consisting solely of the elicitin domain, with occasionally a short tail, and are abundantly secreted during mycelial growth (fig. 7A). The overrepresentation of threonine and serine residues in the C-terminal domains of several ELIs end ELLs suggests extensive O-glycosylation and linkage to the cell wall (fig. 7B). In fact, INF2A

was demonstrated experimentally to be a cell wall associated protein (V.G.A.A. Vleeshouwers et al., in preparation). In plants proline-rich and hydroxyproline rich glycoproteins are amongst the most extensively characterized cell wall components (Cassab 1998). The C-terminal domains of ELI-4 and ELL-8 members are very proline rich suggesting that also these ELLs are associated with the cell wall. The occurrence of GPI sites in several ELL clades suggests that anchoring to the cell membrane is a common way of ELLs to be tethered to the exterior of the cell (fig. 7C). Also ELL-3 is predicted to be GPI-anchored. Because of the specific expression of the ELI-3 genes in the zoospore stage in which a cell wall is lacking, the mature ELL-3 proteins could be tethered to the cell membrane of the mobile zoospores by the GPI anchor and the putative O-linked glycosylation may be used to coat the zoospore surface with oligosaccharides (fig. 7D).

Acknowledgements

We thank Dinah Qutob, Mark Gijzen and Felipe Arredondo for *P. sojae* cDNA libraries, Ralph Dean for *P. sojae* EST sequencing, Mark Waugh, Chunhong Mao and Sucheta Tripathi for bioinformatics assistance, Jeff Boore and Dan Rokhsar for *P. sojae* and *P. ramorum* genome sequences, Wilko Dijksma and Vivianne Vleeshouwers for performing some hybridizations, Syngenta for access to the Syngenta *Phytophthora* Consortium Database prior to public release, Felix Mauch, Lassaad Belbahri and Jerry Marshall for providing unpublished results, and Pierre de Wit for critically reading the manuscript. This work was financially supported by grants to FG from NWO-Aspasia (015.000.057), to BMT from NRI-USDA Cooperative State Research, Education and Extension Service (00-52100-9684 and 2002-35600-12747) and NSF (MCB-0242131) and by funding to BMT from the Virginia Bioinformatics Institute, the US Department of Energy Joint Genome Institute and the Willie Commelin Scholten Foundation. SW was financially supported by the Scottish Executive Evironment and Rural Affairs Department (SEERAD) and RHYJ recieved a travel grant from the Technology Foundation STW for attending the *Phytophthora* annotation jamboree.

References

Altschul, S. F., T. L. Madden, A. A. Schaffer, J. H. Zhang, Z. Zhang, W. Miller, and D. J. Lipman. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389-3402.

Armbrust, E. V., J. A. Berges, C. Bowler, B. R. Green, D. Martinez, N. H. Putnam, S. G. Zhou, A. E. Allen, K. E. Apt, M. Bechner, M. A. Brzezinski, B. K. Chaal, A. Chiovitti, A. K. Davis, M. S. Demarest, J. C. Detter, T. Glavina, D. Goodstein, M. Z. Hadi, U. Hellsten, M. Hildebrand, B. D. Jenkins, J. Jurka, V. V. Kapitonov, N. Kroger, W. W. Y. Lau, T. W. Lane, F. W. Larimer, J. C. Lippmeier, S. Lucas, M. Medina, A. Montsant, M. Obornik, M. S. Parker, B. Palenik, G. J. Pazour, P. M. Richardson, T. A. Rynearson, M. A. Saito, D. C. Schwartz, K. Thamatrakoln, K. Valentin, A. Vardi, F. P. Wilkerson, and D. S. Rokhsar. 2004. The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. Science **306**:79-86.

Armstrong, M. R., S. C. Whisson, L. Pritchard, J. I. Bos, E. Venter, A. O. Avrova, A. P. Rehmany, U. Bohme, K. Brooks, I. Cherevach, N. Hamlin, B. White, A. Fraser, A. Lord, M. A. Quail, C. Churcher, N. Hall, M. Berriman, S. Huang, S. Kamoun, J. L. Beynon, and P. R. Birch. 2005. An ancestral oomycete locus contains late blight avirulence gene *Avr3a*, encoding a protein that is recognized in the host cytoplasm. Proc. Natl. Acad. Sci. USA **102**:7766-7771.

Bairoch, A. 1991. PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res. **19 Suppl**:2241-2245.

Baldauf, S. L. 2003. The deep roots of eukaryotes. Science **300**:1703-1706.

Blein, J. P., P. Coutos-Thevenot, D. Marion, and M. Ponchet. 2002. From elicitins to lipid-transfer proteins: a new insight in cell signalling involved in plant defence mechanisms. Trends Plant Sci. **7**:293-296.

Boissy, G., E. deLaFortelle, R. Kahn, J. C. Huet, G. Bricogne, J. C. Pernollet, and S. Brunie. 1996. Crystal structure of a fungal elicitor secreted by *Phytophthora cryptogea*, a member of a novel class of plant necrotic proteins. Structure **4**:1429-1439.

Boissy, G., M. O'Donohue, O. Gaudemer, V. Perez, J. C. Pernollet, and S. Brunie. 1999. The 2.1 A structure of an elicitin-ergosterol complex: a recent addition to the Sterol Carrier Protein family. Protein Sci. **8**:1191-1199.

Buhot, N., J. P. Douliez, A. Jacquemard, D. Marion, V. Tran, B. F. Maume, M. L. Milat, M. Ponchet, V. Mikes, J. C. Kader, and J. P. Blein. 2001. A lipid transfer protein binds to a receptor involved in the control of plant defence responses. FEBS Lett. **509**:27-30.

Cassab, G. I. 1998. Plant Cell Wall Proteins. Annu. Rev. Plant Physiol. Plant Mol. Biol. **49**:281-309.

Cooke, D. E., A. Drenth, J. M. Duncan, G. Wagels, and C. M. Brasier. 2000. A molecular phylogeny of *Phytophthora* and related oomycetes. Fungal Genet. Biol. **30**:17-32.

Dean, R. A., N. J. Talbot, D. J. Ebbole, M. L. Farman, T. K. Mitchell, M. J. Orbach, M. Thon, R. Kulkarni, J. R. Xu, H. Pan, N. D. Read, Y. H. Lee, I. Carbone, D. Brown, Y. Y. Oh, N. Donofrio, J. S. Jeong, D. M. Soanes, S. Djonovic, E. Kolomiets, C. Rehmeyer, W. Li, M. Harding, S. Kim, M. H. Lebrun, H. Bohnert, S. Coughlan, J. Butler, S. Calvo, L. J. Ma, R. Nicol, S. Purcell, C. Nusbaum, J. E. Galagan, and B. W. Birren. 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature **434**:980-986.

Duclos, J., A. Fauconnier, A. C. Coelho, A. Bollen, A. Cravador, and E. Godfroid. 1998. Identification of an elicitin gene cluster in *Phytophthora cinnamomi*. DNA Seq. **9**:231-237.

Eisenhaber, B., M. Wildpaner, C. J. Schultz, G. H. Borner, P. Dupree, and F. Eisenhaber. 2003. Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for *Arabidopsis* and rice. Plant Physiol. **133**:1691-1701.

Erwin, D. C., and O. K. Ribeiro. 1996. *Phytophthora* diseases worldwide. The American Phytopathological Society, St. Paul Minnesota USA.

Fabritius, A. L., C. Cvitanich, and H. S. Judelson. 2002. Stage-specific gene expression during sexual development in *Phytophthora infestans*. Mol. Microbiol. **45**:1057-1066.

Fefeu, S., S. Bouaziz, J. C. Huet, J. C. Pernollet, and E. Guittet. 1997. Three-dimensional solution structure of beta cryptogein, a beta elicitin secreted by a phytopathogenic fungus *Phytophthora cryptogea*. Protein Sci. **6**:2279-2284.

Gotesson, A., J. S. Marshall, D. A. Jones, and A. R. Hardham. 2002. Characterization and evolutionary analysis of a large polygalacturonase gene family in the oomycete plant pathogen *Phytophthora cinnamomi*. Mol. Plant Microbe Interact. **15**:907-921.

Hirokawa, T., S. Boon-Chieng, and S. Mitaku. 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics **14**:378-379.

Huitema, E., V. G. Vleeshouwers, C. Cakir, S. Kamoun, and F. Govers. 2005. Differences in intensity and specificity of hypersensitive response induction in *Nicotiana* spp. by INF1, INF2A, and INF2B of *Phytophthora infestans*. Mol. Plant Microbe Interact. **18**:183-193.

Jiang, R. H., A. L. Dawe, R. Weide, M. van Staveren, S. Peters, D. L. Nuss, and F. Govers. 2005. Elicitin genes in *Phytophthora infestans* are clustered and interspersed with various transposon-like elements. Mol. Genet. Genomics **273**:20-32.

Julenius, K., A. Molgaard, R. Gupta, and S. Brunak. 2005. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiol. **15**:153-164.

Kader, J. C. 1996. Lipid-Transfer Proteins in Plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. **47**:627-654.

Kamoun, S., P. Hraber, B. Sobral, D. Nuss, and F. Govers. 1999. Initial assessment of gene diversity for the oomycete pathogen *Phytophthora infestans* based on expressed sequences. Fungal Genet. Biol. **28**:94-106.

Kamoun, S., H. Lindqvist, and F. Govers. 1997. A novel class of elicitin-like genes from *Phytophthora infestans*. Mol. Plant Microbe Interact. **10**:1028-1030.

Kamoun, S., P. van West, V. Vleeshouwers, K. E. de Groot, and F. Govers. 1998. Resistance of *Nicotiana benthamiana* to *Phytophthora infestans* is mediated by the recognition of the elicitor protein INF1. Plant Cell **10**:1413-1425.

Kamoun, S., P. vanWest, A. J. deJong, K. E. deGroot, V. Vleeshouwers, and F. Govers. 1997. A gene encoding a protein elicitor of *Phytophthora infestans* is down-regulated during infection of potato. Mol. Plant Microbe Interact. **10**:13-20.

Krogh, A., B. Larsson, G. von Heijne, and E. L. L. Sonnhammer. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J. Mol. Biol. **305**:567-580.

Kroon, L. P., F. T. Bakker, G. B. Van Den Bosch, P. J. Bonants, and W. G. Flier. 2004. Phylogenetic analysis of *Phytophthora* species based on mitochondrial and nuclear DNA sequences. Fungal Genet. Biol. **41**:766-782.

Kumar, S., K. Tamura, I. B. Jakobsen, and M. Nei. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics **17**:1244-1245.

Liu, Z., J. I. Bos, M. Armstrong, S. C. Whisson, L. da Cunha, T. Torto-Alalibo, J. Win, A. O. Avrova, F. Wright, P. R. Birch, and S. Kamoun. 2005. Patterns of diversifying selection in the phytotoxin-like *scr74* gene family of *Phytophthora infestans*. Mol. Biol. Evol. **22**:659-672.

Lusser, A. 2002. Acetylated, methylated, remodeled: chromatin states for gene regulation. Curr. Opin. Plant Biol. **5**:437-443.

Man in 't Veld, W. A., A. de Cock, E. Ilieva, and C. A. Levesque. 2002. Gene flow analysis of *Phytophthora porri* reveals a new species: *Phytophthora brassicae* sp nov. Eur. J. Plant Pathol. **108**:51-62.

Mao, Y., and B. M. Tyler. 1996. Cloning and sequence analysis of elicitin genes of *Phytophthora sojae*. Fungal Genet. Biol. **20**:169-172.

Margulis, L., and K. V. Schwarts. 2000. Five Kingdoms: an illustrated guide to the phyla of life on earth. W.H. Freeman and company, New York.

Mikes, V., M. L. Milat, M. Ponchet, P. Ricci, and J. P. Blein. 1997. The fungal elicitor cryptogein is a sterol carrier protein. FEBS Lett. **416**:190-192.

Nespoulous, C., O. Gaudemer, J. C. Huet, and J. C. Pernollet. 1999. Characterization of elicitin-like phospholipases isolated from *Phytophthora capsici* culture filtrate. FEBS Lett. **452**:400-406.

Panabieres, F., A. Marais, J. Y. LeBerre, I. Penot, D. Fournier, and P. Ricci. 1995. Characterization of a gene cluster of *Phytophthora cryptogea* which codes for elicitins, proteins inducing a hypersensitive- like response in tobacco Mol. Plant Microbe Interact. **8**:996-1003.

Panabieres, F., M. Ponchet, V. Allasia, L. Cardin, and P. Ricci. 1997. Characterization of border species among *Pythiaceae*: several *Pythium* isolates produce elicitins, typical proteins from *Phytophthora* spp. Mycol. Res. **101**:1459-1468.

Pemberton, C. L., and G. P. C. Salmond. 2004. The Nep1-like proteins—a growing family of microbial elicitors of plant necrosis. Mol. Plant Pathol. **5**:353-359.

Pieterse, C. M., P. van West, H. M. Verbakel, P. W. Brasse, G. C. van den Berg-Velthuis, and F. Govers. 1994. Structure and genomic organization of the *ipiB* and *ipiO* gene clusters of *Phytophthora infestans*. Gene **138**:67-77.

Pond, S. L., and S. D. Frost. 2005. DataMonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics **21**:2531-2533.

Qutob, D., P. T. Hraber, B. W. S. Sobral, and M. Gijzen. 2000. Comparative analysis of expressed sequences in *Phytophthora sojae*. Plant Physiol. **123**:243-253.

Qutob, D., E. Huitema, M. Gijzen, and S. Kamoun. 2003. Variation in structure and activity among elicitins from *Phytophthora sojae*. Mol. Plant Pathol. **4**:119-124.

Qutob, D., S. Kamoun, and M. Gijzen. 2002. Expression of a *Phytophthora sojae* necrosis-inducing protein occurs during transition from biotrophy to necrotrophy. Plant J. **32**:361-373.

Randall, T. A., R. A. Dwyer, E. Huitema, K. Beyer, C. Cvitanich, H. Kelkar, A. M. Fong, K. Gates, S. Roberts, E. Yatzkan, T. Gaffney, M. Law, A. Testa, T. Torto-Alalibo, M. Zhang, L. Zheng, E. Mueller, J. Windass, A. Binder, P. R. Birch, U. Gisi, F. Govers, N. A. Gow, F. Mauch, P. van West, M. E. Waugh, J. Yu, T. Boller, S. Kamoun, S. T. Lam, and H. S. Judelson. 2005. Large-scale gene discovery in the oomycete *Phytophthora infestans* reveals likely components of phytopathogenicity shared with true fungi. Mol. Plant Microbe Interact. **18**:229-243.

Rizzo, D. M., M. Garbelotto, and E. M. Hansen. 2005. *Phytophthora ramorum*: Integrative Research and Management of an Emerging Pathogen in California and Oregon Forests. Annu. Rev. Phytopathol. **43:** 309-335.

Roetschi, A., A. Si-Ammour, L. Belbahri, F. Mauch, and B. Mauch-Mani. 2001. Characterization of an *Arabidopsis-Phytophthora* pathosystem: resistance requires a functional PAD2 gene and is independent of salicylic acid, ethylene and jasmonic acid signalling. Plant J. **28**:293-305.

Shan, W., M. Cao, D. Leung, and B. M. Tyler. 2004. The *Avr1b* locus of *Phytophthora sojae* encodes an elicitor and a regulator required for avirulence on soybean plants carrying resistance gene *Rps1b*. Mol. Plant Microbe Interact. **17**:394-403.

Sigrist, C. J., L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform. **3**:265-274.

Soanes, D. M., W. Skinner, J. Keon, J. Hargreaves, and N. J. Talbot. 2002. Genomics of phytopathogenic fungi and the development of bioinformatic resources. Mol. Plant Microbe Interact. **15**:421-427.

Sonnhammer, E. L. L., S. R. Eddy, E. Birney, A. Bateman, and R. Durbin. 1998. Pfam: multiple sequence alignments and HMM-profiles of protein domains. Nucleic Acids Res. **26**:320-322.

Torto, T. A., S. Li, A. Styer, E. Huitema, A. Testa, N. A. Gow, P. van West, and S. Kamoun. 2003. EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen *Phytophthora*. Genome Res. **13**:1675-1685.

van't Slot, K. A. E., and W. Knogge. 2002. A dual role for microbial pathogen-derived effector proteins in plant disease and resistance. Crit. Rev. Plant Sci. **21**:229-271.

van West, P., A. J. de Jong, H. S. Judelson, A. M. C. Emons, and F. Govers. 1998. The *ipiO* gene of *Phytophthora infestans* is highly expressed in invading hyphae during infection. Fungal Genet. Biol. **23**:126-138.

Vauthrin, S., V. Mikes, M. L. Milat, M. Ponchet, B. Maume, H. Osman, and J. P. Blein.
1999. Elicitins trap and transfer sterols from micelles, liposomes and plant plasma membranes. Biochim. Biophys. Acta. **1419**:335-342.

Werres, S., R. Marwitz, W. Veld, A. De Cock, P. J. M. Bonants, M. De Weerdt, K.
Themann, E. Ilieva, and R. P. Baayen. 2001. *Phytophthora ramorum* sp nov., a new pathogen on *Rhododendron* and *Viburnum*. Mycol. Res. **105**:1155-1165.

Whisson, S. C., T. van derLee, G. J. Bryan, R. Waugh, F. Govers, and P. R. J. Birch. 2001. Physical mapping across an avirulence locus of *Phytophthora infestans* using a highly representative, large-insert bacterial artificial chromosome library. Mol. Genet. Genomics **266**:289-295.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. **13**:555-556.

Yang, Z., R. Nielsen, N. Goldman, and A. M. Pedersen. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. Genetics **155**:431-449.

Yang, Z., W. S. Wong, and R. Nielsen. 2005. Bayes empirical bayes inference of amino acid sites under positive selection. Mol. Biol. Evol. **22**:1107-1118.

Supplementary material

Table S1
ELIs and ELLs in *P. brassicae*, *P. infestans*, *P. ramorum* and *P. sojae*.

clade	name[a]	gene size (nt)[b]	signal peptide size (aa)	domain size (aa)[b]	C-terminal domain size (aa)[b]	Thr and Ser[c]	Pro[d]	repeat[e]	O-glycosylation sites[f]	GPI[g]	synonyms	accession number
ELI-1	BRA1	354	20	98	-	-	-					AAO92423.1
ELI-1	INF1	354	20	98	-	-	-					AAB31120.1
ELI-1	RAM1A	438	20	98	28	10	3	TA(3)	10			DQ229218
ELI-1	RAM1B	354	20	98	-	-	-					DQ229219
ELI-1	RAM1C	354	20	98	-	-	-					DQ229220
ELI-1	RAM1D	354	20	98	-	-	-					DQ229221
ELI-1	RAM1E	354	20	98	-	-	-					DQ229222
ELI-1	SOJ1A	354	20	98	-	-	-				Sojein2[h]	AJ007859
ELI-1	SOJ1B	354	20	98	-	-	-				SOJB[i]	AAO24640.1
ELI-1	SOJ1C	354	20	98	-	-	-				Sojein1[h]	AJ007858
ELI-1	SOJ1D	432	20	98	26	10	3	TA(3)	10			DQ229235
ELI-1	SOJ1E	372	20	98	6	5	-		5			DQ229236
ELI-1	SOJ1F	378	23	98	5	4	-		4			DQ229237
ELI-2	BRA2	615	20	98	87	23	6	AA(6) TT(5)	19			AAO92424.1
ELI-2	INF2A	555	20	98	67	34	5	TT(6) AA(5)	34			AAB94814.1
ELI-2	INF2B	567	20	98	71	33	6	PAAT(3)	33			AAB94815.1
ELI-2	RAM2A	573	20	98	73	39	6	TT(13) APAA(3)	39			DQ229223
ELI-2	RAM2B	573	20	98	73	39	6	TT(13) APAA(3)	39			DQ229224
ELI-2	RAM2C	573	20	98	73	39	6	TT(13) APAA(3)	39			DQ229225
ELI-2	RAM2D	>321	20	>87	-	-	-					DQ229226
ELI-2	RAM2E	564	20	99	69	36	6	TT(11) APAA(3)	36			DQ229227
ELI-2	SOJ2A	582	20	98	76	39	5	AT(10)	39			AAO24641.1
ELI-2	SOJ2B	582	20	98	76	39	5	AT(10)	39			DQ229238
ELI-2	SOJ2C	600	20	98	82	41	6	AT(11)	41			DQ229239
ELI-2	SOJ2D	>399	20	98	>15	12	-	ST(4)	12			DQ229240
ELI-3	INF3	552	15	98	71	44	1	ST(9)	42			AY830092
ELI-3	RAM3A	588	20	98	72	38	2	TT(7) SS(4)	34			DQ229228
ELI-3	RAM3B	612	20	98	80	57	1	TT(11) SS(6) AA(4)	57			DQ229229
ELI-3	RAM3C	615	20	98	81	58	-	TT(10) SS(5) AA(3)	51			DQ229230
ELI-3	RAM3D	>339	20	>74	-	-	-					DQ229231
ELI-3	SOJ3A	573	17	98	76	54	2	TT(12) SS(4) AA(3)	54			DQ229241
ELI-3	SOJ3B	630	20	98	86	60	-	TT(12) AS(7)	60			AAO24642.1
ELI-3	SOJ3C	>550	20	98	>58	41	-	ST(12) AA(4)	40			DQ229242
ELI-3	SOJ3D	630	20	98	86	60	-	TT(12) AS(7)	60			DQ229243
ELI-4	BRA5	537	20	98	61	21	12	TT(5) APSAE(4)	21			AAO92425.1
ELI-4	BRA6	498	20	98	48	18	10	TDAP(4)	18			AAO92426.1
ELI-4	INF5	552	20	98	66	20	12	PAA(5) TT(3)	20			AAL16012.1
ELI-4	INF6	549	20	98	65	23	15	APT(8)	23			AAL16013.1
ELI-4	RAM5	576	20	98	74	27	11	TTA(7)	27			DQ229233
ELI-4	RAM6	879	20	98	175	65	45	PT(40) SA(7) DV(3)	65			DQ229234
ELI-4	SOJ5	555	20	98	67	20	12	PA(11) TT(4)	20			DQ229245

ELI-4	SOJ6A	552	20	98	66	22	16	APT(12)	22			AAO24643.1
ELI-4	SOJ6B	>459	21	>74	-	-	-					DQ229246
ELL-1	BRL1A	570	19	81	90	23	1	ST(3) GI(3)	11			DQ229172
ELL-1	BRL1B	609	18	85	98	41	3	AS(9) TT(3)	40	+	BRA7	AAO92427.1
ELL-1	INL1	597	18	85	96	56	1	SS(9) TT(6) AA(4)	54	+	INF7[j]	AAL16014.1
ELL-1	RAL1A	591	18	85	94	48	1	SA(10) TT(3)	47	+		DQ229202
ELL-1	RAL1B	744	20	85	143	64	3	SA(14) TT(4)	60	+		DQ229203
ELL-1	SOL1A	621	21	85	103	56	1	SS(12) AA(6) TT(4)	54	+	SOJ7[l]	DQ229265
ELL-1	SOL1B	573	21	85	85	32	4	SS(7)	16			DQ229266
ELL-1	SOL1C	423	19	85	37	13	2	SS(3)	8			DQ229267
ELL-1	SOL1D	732	20	84	140	59	4	SA(15)	55	+		DQ229268
ELL-2	INL2	594	-	88	103	45	2	SA(12)	42	+		DQ229180
ELL-2	RAL2A	651	19	88	110	27	18	TPAPT(3)	39			DQ229204
ELL-2	RAL2B	501	17	88	55	40	3	ST(7)	40			DQ229205
ELL-2	RAL2C	1029	-	88	108	50	3	SS(10) AA(6) TV(3)	48	+		DQ229206
ELL-2	RAL2D	513	-	88	81	39	15	PS(6) TT(4)	39			DQ229207
ELL-2	RAL2E	480	20	88	52	24	7	AT(6)	24			DQ229208
ELL-2	SOL2A	609	-	88	106	50	3	AS(9) TT(3)	48	+		DQ229270
ELL-2	SOL2B	477	20	88	51	21	7	AP(6) TT(3)	21			DQ229271
ELL-2	SOL2C	1422	19	88	132	69	18	ST(11) AP(7)	69			DQ229272
ELL-2	SOL2D	1422	19	88	132	69	18	ST(11) AP(7)	69			DQ229272
ELL-2	SOL2E	1422	19	88	132	69	18	ST(11) AP(7)	69			DQ229272
ELL-3	BRL3	>468	-	87	>66	25	-	SS(6)	20	+		DQ229174
ELL-3	INL3A	522	20	87	67	24	0	SS(5) GT(4)	24	+		DQ229181
ELL-3	INL3B	471	21	88	48	17	2	SS(5)	13	+		DQ229182
ELL-3	INL3C	447	19	88	42	16	1	SS(4)	11	+		DQ229183
ELL-3	RAL3	507	20	87	62	23	-	SS(6) VG(3)	21	+		DQ229209
ELL-3	SOL3A	522	20	87	67	26	-	GS(6)	23	+	SOJY[l]	AAO24645.1
ELL-3	SOL3B	327	-	83	-	-	-					DQ229274
ELL-4	INL4A	492	21	93	50	6	1				M-25[k]	AAN37687.1
ELL-4	INL4B	>417	18	92	>29	2	-					DQ229185
ELL-4	RAL4	465	21	92	42	4	1					DQ229210
ELL-4	SOL4A	2106	26	92	55	8	2					DQ229275
ELL-4	SOL4B	471	21	92	44	7	1		1			DQ229276
ELL-5	BRL5	501	21	89	56	16	-	GS(4) IA(3)	5	+		DQ229175
ELL-5	RAL5	510	22	89	58	14	-	SS(4) AG(3)	7			DQ229211
ELL-5	SOL5	525	24	89	61	17	-	VG(3) SS(3)	11	+		DQ229277
ELL-6	INL6	321	22	>85	nd	nd	nd					DQ229186
ELL-6	RAL6	528	27	91	58	15	4		8	+		DQ229212
ELL-6	SOL6	516	22	91	59	15	5		8	+	SOJX[l]	AAO24644.1
ELL-7	RAL7A	369	23	92	4	1	-					DQ229213
ELL-7	RAL7B	369	23	92	4	1	-					DQ229214
ELL-7	SOL7	363	21	92	4	1	-		1			DQ229279
ELL-8	INL8	777	18	87	153	24	22	KPT(6) HY(4) DD(3)	19			DQ229187
ELL-8	RAL8A	741	18	93	135	22	23	PT(10) GKY(3)	19			DQ229215
ELL-8	RAL8B	702	-	93	130	24	22	PTPKP(4) EE(3) DS(3)	21			DQ229216
ELL-8	SOL8	840	18	87	174	26	26	PKPTE(5) DD(5) HY(4)	17			DQ229280
ELL-9	RAL9	351	19	91	7	1	-					DQ229217
ELL-9	SOL9	354	20	91	7	1	-		1			DQ229281
ELL-10	RAL10A	360	25	91	4	1	-					DQ229188

ELL-10	RAL10B	360	25	91	4	1	-					DQ229189
ELL-10	SOL10	564	-	91	2	-	-					DQ229247
ELL-11	BRL11	>423	23	93	>35	13	-	MS(4)	7			DQ229171
ELL-11	INL11A	>660	-	98	>61	19	2	DD(4) GT(3)	13			DQ229177
ELL-11	INL11B	567	20	93	50	12	1	GS(4) AA(3)	10	+		DQ229178
ELL-11	RAL11A	531	18	98	61	19	-	MS(4) DD(4) AA(3)	13	+		DQ229190
ELL-11	RAL11B	540	21	93	66	21	-	GS(5) TA(3)	14	+		DQ229191
ELL-11	RAL11C	417	22	98	18	1	-		2			DQ229192
ELL-11	RAL11D	414	20	98	19	2	-		1			DQ229193
ELL-11	SOL11A	570	22	100	67	23	-	SS(4) DD(4) AA(3)	17	+		DQ229248
ELL-11	SOL11B	534	23	93	62	16	-	SG(4) AA(3)	10	+		DQ229249
ELL-11	SOL11C	408	23	94	18	1	-					DQ229250
ELL-11	SOL11D	1257	29	98	291	55	19	RP(5) AA(5) TD(4) SS(4) VG(3) EE(3)	15			DQ229251
ELL-11	SOL11E	405	17	98	19	1	-		1			DQ229252
ELL-12	RAL12	414	23	92	23	11	-		10			DQ229194
ELL-12	SOL12	399	24	92	17	8	-		4			DQ229254
ELL-13	INL13	>459	18	80	>55	16	3	SL(3) AT(3)	14			DQ229179
ELL-13	RAL13A	1230	21	80	198	62	9	ST(9)	38	+		DQ229195
ELL-13	RAL13A2	1230	21	81	198	62	9	ST(9)	38	+		DQ229195
ELL-13	RAL13B	1128	21	81	275	58	8	AT(7) SS(5) YE(3) NC(3) ID(3)	11	+		DQ229196
ELL-13	RAL13C	411	22	75	40	7	2		2	+		DQ229197
ELL-13	RAL13D	861	25	81	75	26	8	TT(5) AA(3)	23			DQ229198
ELL-13	RAL13E	474	25	80	53	15	-	SA(4)	11			DQ229199
ELL-13	RAL13F	639	22	81	110	55	4	SS(14) AT(5)	52			DQ229200
ELL-13	RAL13J	1041	-	80	202	76	10	AS(9) TT(6)	50	+		DQ229201
ELL-13	SOL13A	1245	19	80	205	65	10	ST(11) AE(4)	42	+		DQ229255
ELL-13	SOL13A2	1245	19	81	205	65	10	ST(11) AE(4)	42	+		DQ229255
ELL-13	SOL13B	1140	24	81	275	62	8	SS(7) AT(5) NC(3)	13			DQ229256
ELL-13	SOL13C	423	32	75	34	5	4		2	+		DQ229257
ELL-13	SOL13D	762	27	81	31	12	2		8			DQ229258
ELL-13	SOL13E	456	25	80	47	12	1	SA(4)	10			DQ229259
ELL-13	SOL13F	609	22	81	100	45	5	SS(8) AA(5) TD(4)	45	+		DQ229260
ELL-13	SOL13G	822	22	78	174	57	18	TT(7) SP(5) CA(3)	36			DQ229261
ELL-13	SOL13H	1044	24	83	44	12	1	LV(3)	4			DQ229262
ELL-13	SOL13H2	1044	24	81	44	12	1	LV(3)	5			DQ229262
ELL-13	SOL13I	477	22	82	54	20	6	PSS(3)	18			DQ229263
ELL-13	SOL13J	663	21	80	120	46	3	SS(11) AG(3)	32	+		DQ229264
ELI	RAM3X	828	19	98	152	89	6	SA(15) TT(9)	86	+		DQ229232
ELI	INF4	354	20	98	-	-	-					AAL16011.1
ELI	SOJ3X	537	20	98	61	30	8	TA(9) SS(3)	30			DQ229244
ELL	SOL1E	600	18	85	94	25	8	GS(7) AA(3)	24	+		DQ229269
ELL	SOL11F	729	18	85	140	55	1	SS(11) AT(6) ME(4)	45	+		DQ229253

[a] Multiple domains derived from the same gene are underlined
[b] > indicates that the sequence is not complete.
[c] The number of Thr and Ser residues in the C-terminal domain.
[d] The number of Pro residues in the C-terminal domain.
[e] Repeats in the C-terminal domain. The number of repeats is listed in brackets.
[f] The number of putative O-glycosylation sites in the C-terminal domain as predicted by NetOGlyc 3.1.
[g] The presence of GPI anchors as predicted by big-PI Plant Predictor.
[h] Mao and Tyler 1996; Becker, Nagel, and Tenhaken 2000.
[i] Qutob et al. 2003.
[j] Kamoun, Lindqvist, and Govers 1997; Kamoun et al. 1997; Kamoun et al. 1999.
[k] Fabritius, Cvitanich, and Judelson 2002.

Supplementary material

Table S2
ELIs and ELLs present in GenBank

gene name	accession number	species
BRA1	gb|AAO92423.1	Phytophthora brassicae
BRA2	gb|AAO92424.1	*Phytophthora brassicae*
BRA5	gb|AAO92425.1	*Phytophthora brassicae*
BRA6	gb|AAO92426.1	*Phytophthora brassicae*
BRA7	gb|AAO92427.1	*Phytophthora brassicae*
CAP	gb|AAK32727.1	*Phytophthora capsici*
CAP	gb|AAP43023.1	*Phytophthora capsici*
CAPalpha	gb|AAO15602.1	*Phytophthora capsici*
CI13	emb|CAB38322.1	*Phytophthora cinnamomi*
CI16	emb|CAB38323.1	*Phytophthora cinnamomi*
CI17	emb|CAB38324.1	*Phytophthora cinnamomi*
CIN	sp|P15569	*Phytophthora cinnamomi*
CINNA	emb|CAB38321.1	*Phytophthora cinnamomi*
CRY	sp|P15570	*Phytophthora cryptogea*
CRY_A1	pir|S49905	*Phytophthora cryptogea*
CRY1	emb|CAA84227.1	*Phytophthora cryptogea*
CRY2	sp|P41803	*Phytophthora cryptogea*
CRY3	sp|P41804	*Phytophthora cryptogea*
DREaplha	gb|AAB22770.1	*Phytophthora drechsleri*
DREbeta	gb|AAB22771.1	*Phytophthora drechsleri*
INF1	gb|AAB31120.1	*Phytophthora infestans*
INF2A	gb|AAB94814.1	*Phytophthora infestans*
INF2B	gb|AAB94815.1	*Phytophthora infestans*
INF4	gb|AAL16011.1	*Phytophthora infestans*
INF5	gb|AAL16012.1	*Phytophthora infestans*
INF6	gb|AAL16013.1	*Phytophthora infestans*
INF7	gb|AAL16014.1	*Phytophthora infestans*
M-25	gb|AAN37687.1	*Phytophthora infestans*
MgMalpha	gb|AAB27563.1	*Phytophthora megasperma*
MgMbeta	gb|AAB27564.1	*Phytophthora megasperma*
PAL	gb|AAP43024.1	*Phytophthora palmivora*
PARA1	gb|P41801	*Phytophthora parasitica*
PARA2	gb|AAK01625	*Phytophthora parasitica*
PARA3	gb|AAB34031	*Phytophthora parasitica*
PARL1	gb|DQ112351	*Phytophthora parasitica*
SOJ2	gb|AAO24641.1	*Phytophthora sojae*
SOJ3	gb|AAO24642.1	*Phytophthora sojae*
SOJ6	gb|AAO24643.1	*Phytophthora sojae*
SOJB	gb|AAO24640.1	*Phytophthora sojae*
SOJX	gb|AAO24644.1	*Phytophthora sojae*
SOJY	gb|AAO24645.1	*Phytophthora sojae*
OLI	emb|CAC39588.1	*Pythium oligandrum*
VEX1	gb|AAB34416.1	*Pythium vexans*
VEX2	gb|AAB34417.1	*Pythium vexans*
seq13	gb|AAE35924.1	*Sequence 13 from patent US 5981843*
seq8	gb|AAE35923.1	*Sequence 8 from patent US 5981843*

Supplementary material

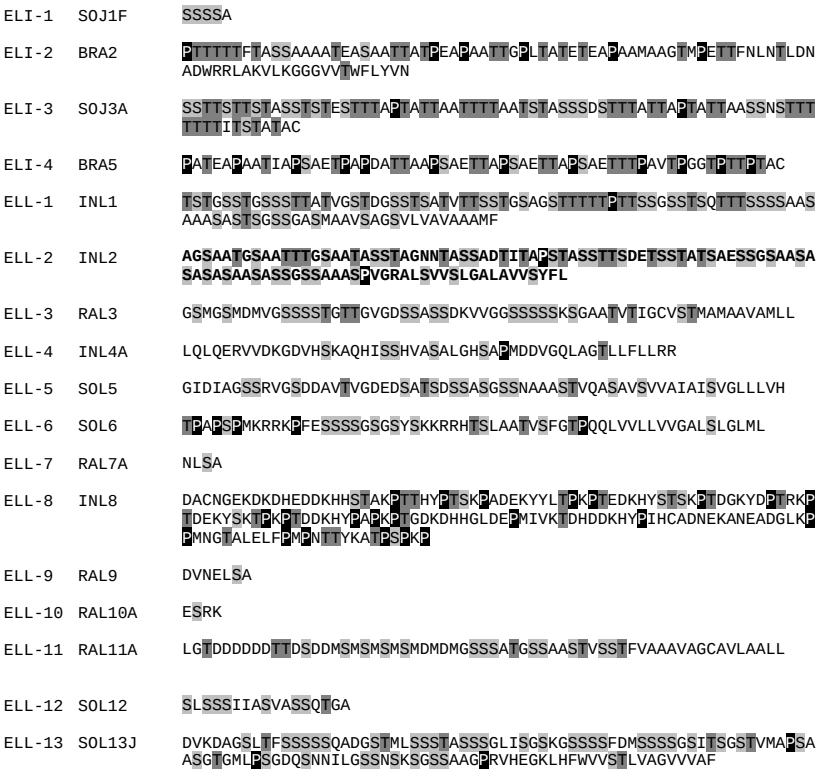
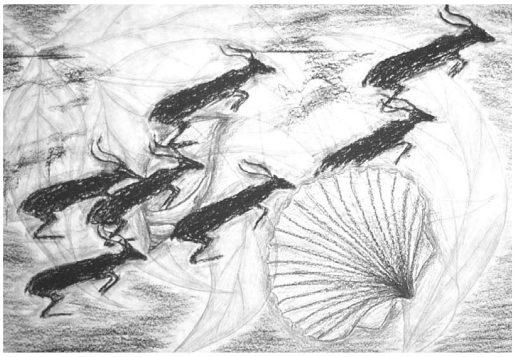


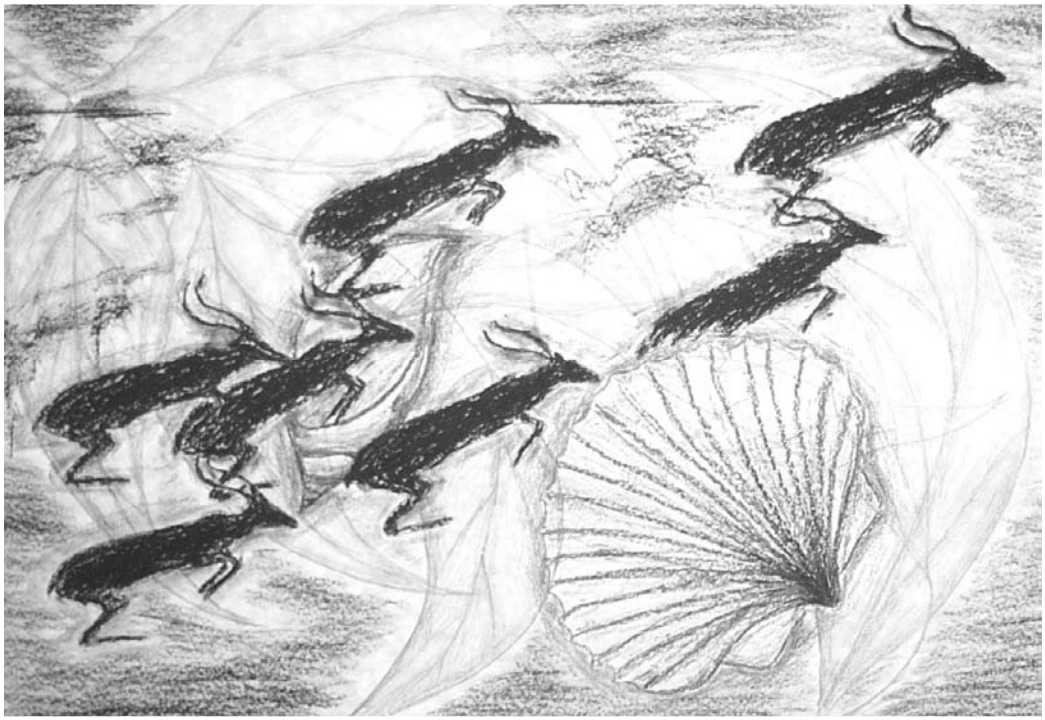
Fig. S1.— C-terminal domains of ELIs and ELLs. The residues Thr (T), Ser (S) and Pro (P) are shaded. For each clade an example of an individual ELI or ELL was randomly chosen.



Chapter 6

Synteny or lack of synteny: comparative analyses of genes encoding secreted proteins in *Phytophthora*

Rays H.Y. Jiang, Brett M. Tyler and Francine Govers



Synteny or lack of synteny: comparative analyses of genes encoding secreted proteins in *Phytophthora*

Rays H.Y. Jiang, Brett M. Tyler* and Francine Govers

Laboratory of Phytopathology, Plant Sciences Group, Wageningen University, Binnenhaven 5, NL-6709 PD Wageningen and Graduate School Experimental Plant Sciences, The Netherlands *Virginia Bioinformatics Institute, Virginia Polytechnic and State University, Blacksburg VA, USA

Key words

Phytophthora, synteny, comparative genomics, elicitor

Alternative title

Effector like genes reside in dynamic regions of the *Phytophthora* genome

Abstract

Co-linearity was found between two *Phytophthora* genomes by comparing in total 1.5 Mb sequence of *Phytophthora sojae* to 0.9 Mb syntenic sequence of *Phytophthora ramorum*. Overall, the gene order in the two species is conserved but genome rearrangements have also occurred. The hotspots for rearrangements often harbor genes encoding secreted proteins including effectors important for interaction with host plants. Among secreted protein genes different evolutionary patterns were found. Elicitin genes that code for a complex family of highly conserved *Phytophthora* specific elicitors show conservation in gene number and order, and are often clustered. In contrast, the race specific elicitor gene *Avr1b* and its homologues are scattered over the genome and often appear to be missing from the syntenic regions. Some gene families encoding secreted proteins were found to be expanded in one species as compared to the other which could be the result of either repeated gene duplications in one species or specific deletions in the other. The different evolutionary patterns revealed by comparative genomics may shed light on the functions of these secreted proteins in the biology and pathology of the two *Phytophthora* species.

Introduction

In all kingdoms of life, comparative genomics has revealed synteny between related species (Hamer et al. 2001; McCouch 2001). Synteny refers to conserved gene order between orthologous chromosomal segments of two or more organisms; a set of matches that connect genes with conserved order is identified as a synteny block. Substantial synteny blocks are maintained in the genomes of rice (*Oryza sativa*) and *Arabidopsis thaliana* 200 million years after separation (Yu et al. 2002). Also across the whole vertebrate radiation, synteny is observed. Birds and mammals evolved separately for over 300 million years but conserved synteny blocks are recognized in the genomes of chicken, mouse and human (International Chicken Genome Sequencing Consortium 2004). Synteny blocks and their breakpoints can be utilized to track the course of genome evolution, such as predicting ancestral chromosomes, deducing lineage specific duplications and inferring the rate of genome rearrangements. For example, the ancestral chromosome architecture in vertebrates can be predicted by comparing synteny blocks across various species (International Chicken Genome Sequencing Consortium 2004). Segmental duplications have pointed out hotspots for mammalian chromosome evolution, because abundant lineage specific duplications were found at the breakpoints of synteny blocks in various genomes (Bailey et al. 2004). Rapid chromosomal rearrangements within the animal kingdom have been inferred from studies in nematodes; an excessive amount of breakpoints was found in the genomes of *Caenorhabditis elegans* and *Caenorhabditis briggsae* despite their relatively short divergence 80–110 million years ago (Mitreva et al. 2005).

When a pathogen is compared with a non-pathogen, synteny breakpoints can be ideal to find pathogenesis related genes. Because of the flexibility required for the interaction with hosts, selection can encourage loss of synteny in regions containing pathogenicity or virulence genes via genome rearrangements. A nearly perfect synteny was found between the bacterial pathogen *Listeria monocytogenes* causing food-borne infections and the closely related non-pathogenic species *Listeria innocua.* However, synteny is broken at a locus required for chemotaxis and motility which may explain the different needs of parasitic and free living life styles (Buchrieser et al. 2003). Also in eukaryotic parasites comparative genomics revealed virulence related genes. Between the human malaria parasite *Plasmodium falciparum* and the rodent malaria parasite *Plasmodium yoelii,* there is striking synteny except for some species-specific genes that are frequently found to trigger host immune responses (Carlton et al. 2005). Three related protozoan pathogens, *Trypanosoma brucei, Trypanosoma cruzi,* and *Leishmania major* that diverged 200 to 500 million years ago, show high levels of synteny. Their synteny breakpoints, however, are associated with expansions of retroelements, structural RNAs and surface antigen gene families (El-Sayed et al. 2005).

To locate the precise breakpoints of synteny, appropriate evolutionary distance is crucial for the comparison. Two genomes should not only be similar enough to be aligned but also divergent enough to yield differences. The fungal rice blast pathogen *Magnaporthe grisea* shares little conservation in gene order with the related bread mold fungus *Neurospora crassa,* and therefore the evolutionary distance is too large for whole genome alignment (Dean et al. 2005). In contrast, the majority of the *Phytophthora*

taxa form a recently evolved monophyletic group (Cooke et al. 2000). These fungal-like organisms are oomycetes that evolved independent from true fungi and comprise many pathogens of crops, trees and ornamentals (Erwin and Ribeiro 1996). Their phylogenetic affinity suggests that different *Phytophthora* species may be at a proper evolutionary distance to search for synteny and synteny breakpoints in their genomes. With the release of the complete genome sequence of two *Phytophthora* species comparative analyses of large genomic regions becomes feasible. One is *Phytophthora sojae,* an economically important species responsible for yield losses in soybean due to stem and root rot. The other is *Phytophthora ramorum,* a recently identified species that causes sudden oak death and is disastrous for the oak trees on the west coast of the USA (Rizzo et al. 2004). The two species have different genome sizes (95 Mb versus 65 Mb), sexual behavior (homothallic versus heterothallic) and host specificity. *P. sojae* exclusively infects soybean whereas *P. ramorum* has a broad host range. Genome rearrangement patterns obtained from comparative genomics may explain aspects of distinct pathogenicity properties of these two individual species.

For phytopathogenic fungi and oomycetes, obvious pathogenicity related factors are extracellular effectors. These proteins are commonly secreted by pathogens into the plant apoplast to promote infection of the host. However, if plants are able to recognize these effectors, defense responses including a form of programmed cell death (the hypersensitive response) can be triggered and infection can be halted (van't Slot and Knogge 2002). Effectors that are able to elicit host responses are termed elicitors. From an evolutionary perspective, elicitors can be classified according to the time of their emergence. Elicitors that are highly conserved across all *Phytophthora* species are likely to have a basal role in the interaction with the host. In contrast, species-specific elicitors probably evolved more recently and are likely to contribute to the unique features exemplified in, e.g., avirulence on specific host plants. In this study, two groups of elicitors of different phylogenetic distribution were investigated: elicitins representing genus-specific elicitors and AVR1b(-like) proteins representing species-specific elicitors.

Elicitins are ubiquitous within the *Phytophthora* genus but also unique for *Phytophthora* and some closely related *Pythium* species. They belong to an extensive gene family composed of elicitin (ELI) and elicitin-like (ELL) genes and are highly conserved between *Phytophthora* species (Jiang et al. 2006). They induce defense responses in plants, in particular in *Nicotiana* species (Ricci et al. 1992; Kamoun et al. 1998) and one of the elicitins, INF1 from the late blight pathogen *Phytophthora infestans,* has been proposed to be a determinant of non-host resistance in *Nicotiana benthamiana* (Kamoun et al. 1998). AVR1b belongs to a group of elicitors that are species-specific (Shan et al. 2004). The avirulence gene *Avr1b* in *P. sojae* does not belong to any highly conserved gene family and lacks highly similar homologs in other species. AVR1b is responsible for the specific gene-for-gene interaction between soybean lines carrying the *Rps1* resistance gene and *P. sojae* strains carrying *Avr1b* (Shan et al. 2004). Thus elicitins and AVR1b differ in various aspects including degree of sequence conservation, phylogenetic distribution and specificity of host interaction. It is not known whether the evolutionary processes that shaped these two types of elicitors are different. One way to answer this question would be a thorough

examination of the genomic locations of the elicitin and *Avr1b(-like)* genes and comparison of these regions between different *Phytophthora* species.

The aims of this study were (i) to examine whether elicitin and *Avr1b(-like)* genes in *P. sojae* and *P. ramorum* are located in syntenic blocks or regions, and if so, (ii) to identify other secreted protein genes and gene families located in the syntenic regions, (iii) to characterize expansion and shrinkage patterns of gene families and (iv) to analyse insertions and deletions in more detail. This study presents a detailed investigation of the local co-linearity between two *Phytophthora* genomes, and discusses the evolutionary dynamics of the different types of effector genes.

Results

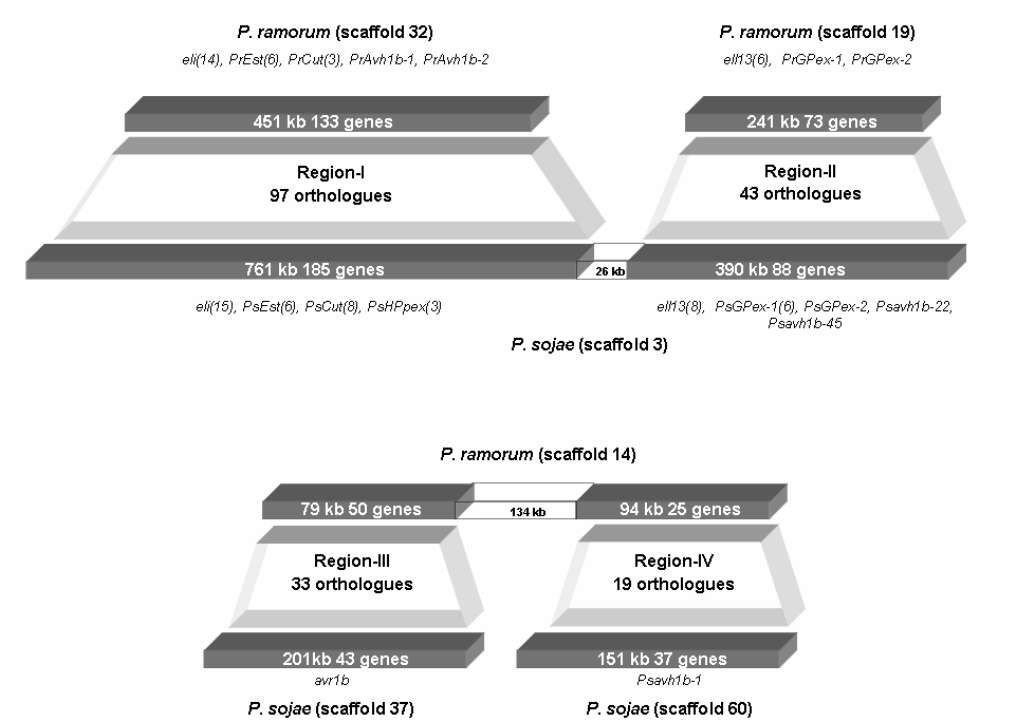
Four regions of synteny between *P. sojae* and *P. ramorum*

This study is focused on genomic regions containing two types of effector genes: elicitin genes, and *Avr1b (-like)* genes. The elicitin gene family is comprised of elicitin (*eli*) and elicitin-like (*ell*) genes, of which many cluster in the genomes (Jiang et al. 2006). For this study an *eli* gene cluster and the *ell13* gene cluster were chosen. For the race specific elicitor gene *Avr1b* and *Avr1b-like* genes two regions were chosen, one containing *PsAvr1b* and the other containing a close homologue designated as *Psavh1b-1* (BLASTP *E* value 4E-63). So in total, four genomic regions spanning the two groups of effector genes were investigated: region-I with the *eli1, eli2, eli3, eli4* gene cluster, region-II with the *ell13* cluster, region-III with *PsAvr1b* and region-IV with *Psavh1b-1* (Fig. 1A). Region-I and region-II are located on the same scaffold in *P. sojae.* Similarly, region-III and region-IV are located on the same scaffold in *P. ramorum*. In addition to the *eli*, *ell* and *Avr1b (-like)* genes, the four regions contain other secreted protein genes and gene families and as such, are representative examples to illustrate the genomic organization and context of secreted protein encoding genes.

Region-I in *P. sojae* is defined as a 761 kb genomic region containing 15 clustered *eli* genes. A total of 185 annotated genes is present in this region. PhIG (Phylogenetically inferred groups) and BLAST analysis revealed that the syntenic region in *P. ramorum* spans a 451 kb sequence that contains 133 annotated genes. In region-I, 97 gene pairs can be assigned as orthologues based on the sharing of best reciprocal BLAST hits. Most of the 97 orthologous pairs show the same gene order and orientation but occasionally reversals of position and orientation do occur; there are 3 pairs with reversed position and 8 pairs with reversed orientation. Region-I covers a wide range of genes encoding proteins involved in various processes such as signal transduction (cAMP-dependent kinase), cell metabolism (Acyl-CoA desaturase) and transport of metabolites (solute carrier). Region-I also contains a number of (putative)

effector genes. In addition to the elicitin genes it contains genes coding for cutinases, histidine/proline (His/Pro) rich secreted proteins and two *P. ramorum Avr1b* homologues.

A



B

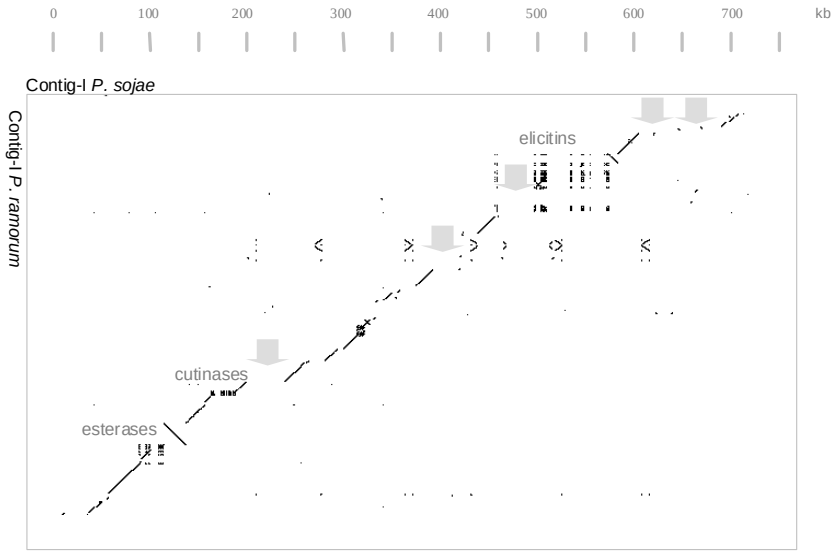


Fig. 1. A, The four syntenic regions used for comparative analysis. The chromosomal regions are shown in dark gray bars and the syntenic regions are shown in light gray blocks The secreted protein genes located in these regions are indicated above or underneath. If there is more than one gene, the number of genes is indicated in brackets. Region-I and region-II are on the same scaffold in *P. sojae* and region-IV and region-IV are on the same scaffold in *P. ramorum*. **B,** A PipMaker dot blot of Region-I reveals co-linearity and re-arrangements between *P. sojae* (plotted horizontal) and *P. ramorum* (plotted verical). The position in the sequence is indicated by the ruler above the plot. Homology was plotted when at least 100 bp align without a gap and with nucleotide identity above 70%. Several major synteny break points are marked by arrows.

Region-II is a 390 kb sequence from *P. sojae* that spans the *ell13* gene cluster and a 241 kb syntenic region in *P. ramorum* was selected using the same approach as described above. In total, 43 gene pairs with best reciprocal BLAST hit were assigned to be orthologues. With the exception of 10 reversals, gene order and orientation are largely conserved. As region-I, region-II contains genes involved in various processes and, in particular, many genes encoding secreted cysteine (Cys) rich proteins.

Region-III and region-IV in *P. sojae* contain *Avr1b* and *PsAvh1b-1*, respectively. For region-III, a total of 33 orthologuous pairs define a 201 kb region *in P. sojae* and a 79 kb region in *P. ramorum* as syntenic. For region-IV, 19 orthologuous pairs define a region of 151 kb in *P. sojae* and 94 kb region *in P. ramorum* as syntenic (Fig. 1A). Few reversals are found in these two regions.

To visualize the co-linearity and gene re-arrangements between these two genomes, Region-I was used for a dot blot by using the program PipMaker (Schwartz et al. 2000). The discontinuous diagonal line shows that Region-I is largely syntenic between *P. sojae* and *P. ramorum* (Fig. 1B). Most other stretches away from the main diagonal line are various mobile elements. The three major gene families can also be visualized as repeated lines. Noticeably, the cutinase gene family shows more copy numbers in *P. sojae* as compared to *P. ramorum*. In several regions synteny is broken as shown by the gap in the blot. Five of such gaps containing annotated genes can be identified (Fig. 1B).

In total, 1.50 Mb sequence in *P. sojae* is compared to 0.86 Mb in *P. ramorum*. The 192 orthologuous gene pairs represent 54% of all the genes present in the four examined regions in *P. sojae* and 68% of the genes defined in the syntenic regions in *P. ramorum.* Apparently more than half of the genes in these particular regions share orthology. Because the four regions are derived from different scaffolds, they belong to four distinct synteny blocks. All rearrangements described in this study occur within the synteny blocks.

Conserved phospholipid synthesis associated genes are scattered in the *eli* gene cluster

ELIs of different *Phytophthora* species share a highly conserved 98 amino acid elicitin domain and in all species *eli* genes are members of complex gene families. In region-I the *eli* gene cluster is comprised of 15 members in *P. sojae* spanning 115 kb and 14 members in *P. ramorum* spanning 59 kb. With the exception of one *eli*, i.e. *ram5*, all *P. ramorum eli* genes have orthologues in the syntenic region in *P. sojae* (Jiang et al. 2006). *Soj5*, the *ram5* orthologue in *P. sojae* is located on a different scaffold together with two other *eli* genes.

Elicitor activity is not the primary function of ELIs in *Phytophthora* because it is likely to be detrimental for the pathogen. The intrinsic biological function of ELIs is generally thought to be related to lipid binding and/or processing. One of the ELIs is known to bind ergosterol and to function as a sterol-carrier protein (Mikes et al. 1997; Vauthrin et al. 1999), whereas another ELI was shown to have phospholipase activity

(Nespoulous et al. 1999). Scattered in between the 15 *eli* genes in *P. ramorum*, there are 5 non-*eli* genes. Two non-eli genes are present elsewhere in *P. sojae*, the other three are conserved in the *P. sojae eli* region. One of these has no hits with known proteins in GenBank but the other two seem to be associated with phospholipid synthesis.

One of the two phospholipid associated genes shows BLASTX homology to an inositol polyphosphate-4-phosphatase IP4P (NP004018) with *E* value 7E-32. Phosphatidylinositol signaling pathways are important for cells to respond to extracellular signals. Phosphatidylinositols mediate intracellular signaling by rapid turnover and generating second messengers (Zhang and Majerus 1998). IP4P is involved in the pathway by removing the phosphate group from the inositol ring of inositol 3,4-bisphosphate (Shearn et al. 2001).
The other phospholipid associated gene shows BLASTX homology (*E* value 1E-109) to the PIG-A protein (Phosphatidylinositol-glycan biosynthesis, class A) (P32363). GPI (glycosylphosphatidylinositol) lipid anchoring is an important post-translational modification of eukaryote proteins. PIG-A is required for synthesis of the early intermediate for the GPI-anchor. It performs its task by using C-terminal domain binding of UDP-GlcNAc whereas the N-terminal domain interacts with the phosphatidylinositol moiety (Miyata et al. 1993; Eisenhaber et al. 2003a). Since these two phospholipid synthesis associated genes are located in between *eli* genes and conserved in both *Phytophthora* species, it is possible that their function is related to the function of ELIs in modifying and interacting with lipids.

Clustering of three different groups of Cys-rich secreted proteins with GPI anchors

The *ell13* genes encode proteins that have a signal peptide, a Cys-rich elicitin domain and a C-terminal tail that may have a GPI anchor (Jiang et al. 2006). In *P. ramorum*, all 6 members of the *ell13* gene family are located in region-II, and in *P. sojae* 8 out of 10 *ell13* genes are located in this region. *sol13G* and *sol 13I* are located on another scaffold. The ELL13 family has a typical Cys spacing pattern of C-20-C-(12-17)-C-4-C-13-C-17/18-C which in three ELI13 proteins occurs twice, SOL13A, RAL13A and RAL13E. A phylogenetic tree based solely on the elicitin domain (Fig. 2) showed that all ELL13 proteins encoded in region-II have orthologues in both species with the exception of SOL13A and SOL13H. Therefore, the expansion of the *ell13* family probably occurred before speciation and is conserved in *P. sojae* and *P. ramorum*.

In the vicinity of the *ell13* cluster there are other genes encoding secreted proteins and, remarkably, they have features in common with ELL13 such as Cys-rich domains and GPI anchors (Fig. 2). We named these genes *GPex* (GPI anchored Putative extracellular protein). *GPex1* forms a family consisting of 6 members in *P. sojae* but there is only a single copy in *P. ra*morum. Three out of the 6 *P. sojae GPex1* genes (*PsGPex1-3, PsGPex1-5* and *PsGPex1-6)* are pseudo-genes carrying frameshift mutations. The other three *PsGPex1* genes do possess an intact ORF, two of which encode a GPI anchor. Moreover, transcripts of *PsGPex1-2* and *PsGPex1-4* are present in the EST database. The Cys-rich domain of the

PGEX1 family members is conserved and has a characteristic Cys spacing pattern of C-C-30-C-16-C-31-C-30-C. When this Cys-rich domain was used to contruct a phylogenetic tree the expansion of *GPex1* in *P. sojae* became apparent (Fig. 2). The six *PsGPex1* genes do not form a single clade and to be so divergent the *PsGPex-1* genes should have evolved at a fast pace. Alternatively, the phylogenetic pattern could be explained by gene loss in *P. ramorum*. GPEX2 also possesses a Cys-rich domain but, in contrast to *GPex1, GPex2* is a single copy gene in both *P. sojae* and *P. ramorum*. PsGPex2 and PrPGex2 are particularly rich in cysteines with the latter lacking a GPI anchor (Fig. 2).

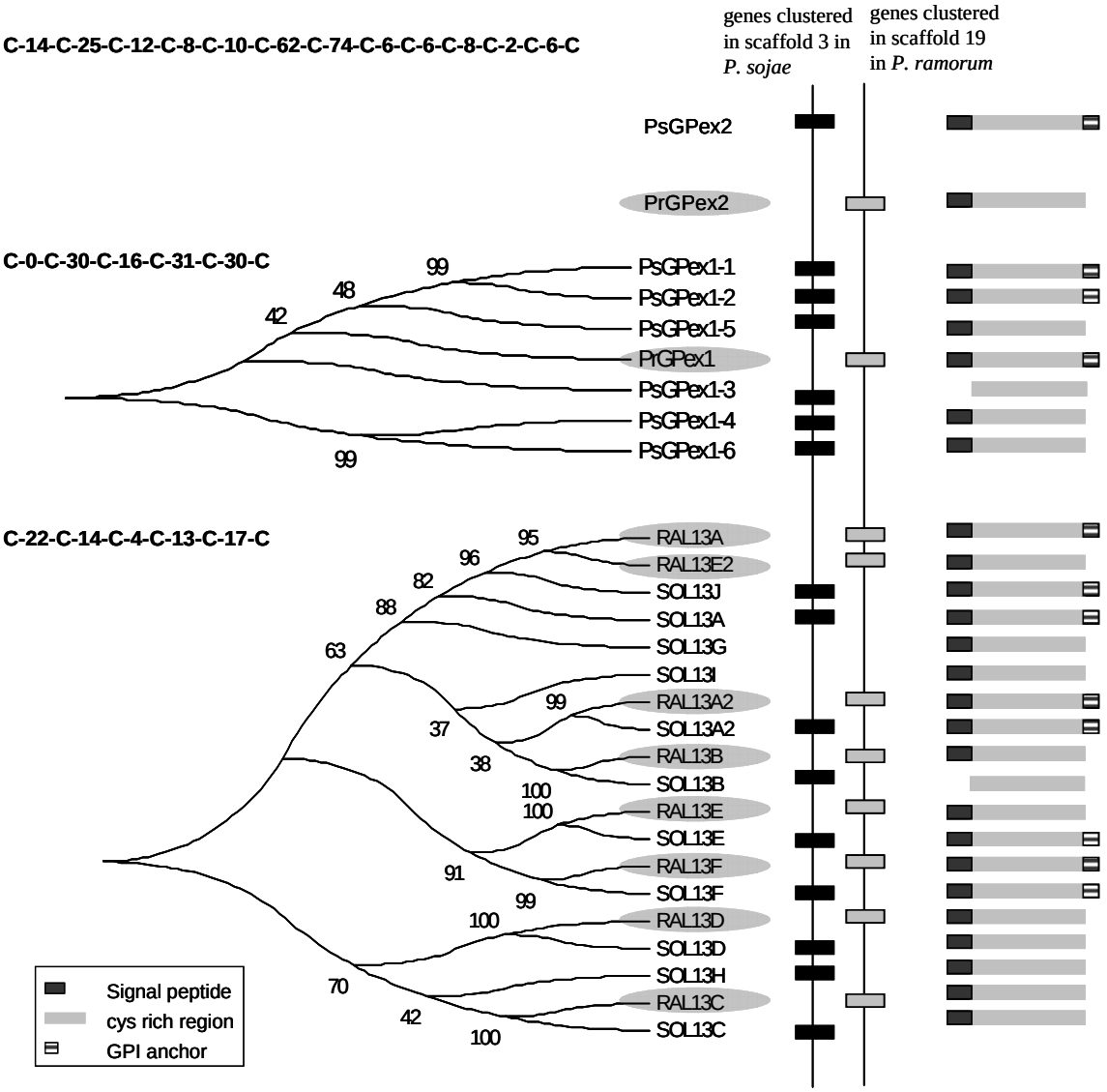


Fig. 2. Clustering of three gene groups encoding Cys rich secreted proteins with a GPI anchor. The phylogenetic trees are based on the Cys rich domains. Proteins shaded in gray are derived from *P. ramorum*. The scaffolds only depict the presence of genes, not the order nor the relative distance. RAL13A and RAL13A2 are two elicitin domains in one protein. The same holds for RAL13E and RAL13E3, and SOL13A and SOL13A2, respectively.

Expansion of the extracelluar cutinase family versus conservation of an intracelluar esterase family

Lipases and esterases constitute a large category of enzymes which possess a wide variety of structurally diverse substrates (Ollis et al. 1992). Two families of this enzyme super-family are found to cluster side by side in region-I. One family is cutinase and the other is a type of esterase. No apparent sequence homology can be found for these two families.

Cutinase is one of the smallest members of the serine hydrolase family and is named after its ability to degrade cutine polymers (Longhi and Cambillau 1999). Region-I contains a gene cluster encoding secreted proteins with a cutinase domain (IPR011150). In *P. ramorum* the gene cluster comprises three genes (*PsCut1-3)* while in *P. sojae* the same region carries 8 cutinase genes (*PsCut3-10*). The expansion, however, is not limited to this region. A total of 14 cutinase genes can be identified in the whole genome of *P. sojae* whereas only 4 can be found in *P. ramorum.* All the cutinase genes have intact ORFs and ESTs derived from *PsCut1, PsCut10, PsCut12 and PsCut13* are present in the *P. sojae* EST database.

A phylogenetic tree was constructed to visualize the relationship between the cutinase genes (Fig. 3A). Two distinct clades can be identified. All the 8 *PsCut* genes and 3 *PrCut* genes that are assigned to the clusters on Region-I fall in clade 1 while clade 2 contains cutinase genes that are located on other scaffolds (scaffold 134 of *P. sojae* and scaffold 11 of *P. ramorum)*. *PsCut1, 2 and 14* are located on two small scaffolds that may be integrated into the gene cluster in clade-I in later sequence assembly releases. Apparently the division of the two clades based on sequence similarity coincides with the different locations of the genes on the genome.

In contrast to the expansion of the extracellular cutinase gene family in *P. sojae*, an esterase gene family shows similar duplication patterns in both species. The esterase gene family is found 50 kb apart from the cutinase family in *P. sojae*. The members of this family encode intracellular proteins because a signal peptide is lacking at the N-terminus. Their closest homologue in the SwissProt database is a monoglyceride lipase (MGLL_RAT Q8R431) with BLAST *E* value 1.8e-16. Six genes encoding proteins with the esterase domain can be found in both *P. sojae* and *in P. ramorum*. The esterase gene in *P. sojae* is designated as *PsEst*, in *P. ramorum* as *PrEst*. Search with interpro domain IPR000379 in annotated proteins showed that these 12 genes are part of a large family of more than a hundred members all possessing esterase domains. Phylogenetic construction of the 12 esterases in region-I showed six groups (Fig. 3B). Each group is comprised of two proteins, one derived from *P. sojae* and the other derived from *P. ramorum*. The six groups indicate that each *PrEst* gene in *P. ramorum* clearly has an orthologue in *P. sojae* except for *PrEst3.* In addition, the gene order and orientation of these esterase genes are also conserved between the two species.

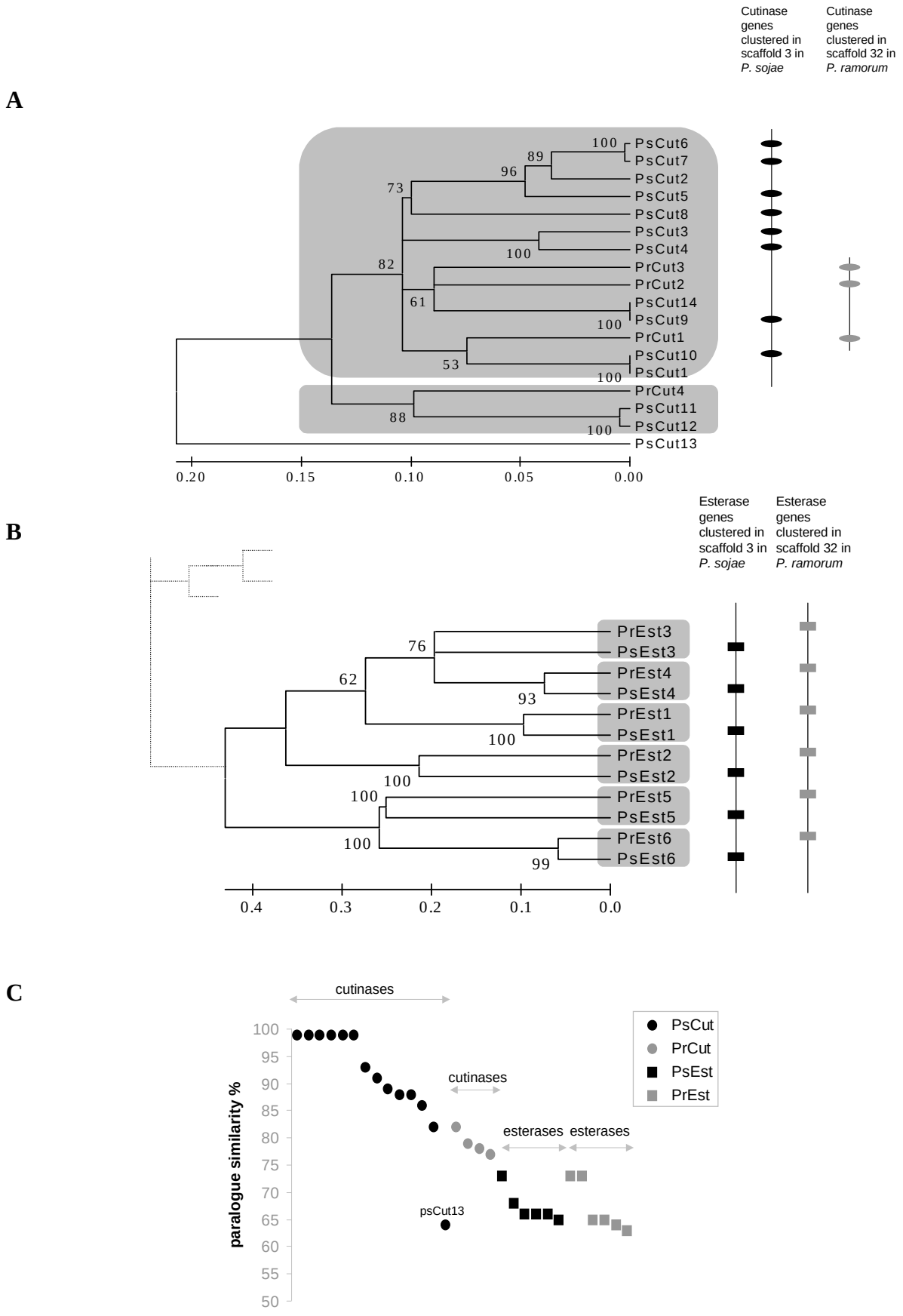


Fig. 3. A, Expansion of cutinase genes in scaffold 3 in *P. sojae.* **B,** Orthologues of each esterase gene in scaffold 32 in *P. ramorum* can be identified in scaffold 3 in *P. sojae.* Only esterase genes clustered in the investigated regions are shown in the phylogenic tree. In **A** and **B** the scaffolds only depict the presence of genes, not the order nor the relative distance. **C,** Similarity of paralogues of cutinases and esterases. Among the cutinase genes only *psCut13* shows a relatively low level of similarity with its paralogue.

In this study, the two neighboring gene families gave distinct evolutionary patterns. One way to explain the pattern is to trace back the last duplication event. Gene duplication after speciation gives rise to similar copies (paralogues) while duplication events prior to speciation should result in divergent genes (orthologues). Given a constant mutation rate, the time of the last duplication can be deduced from the similarity of paralogues. Nearly identical paralogues indicate a very recent duplication and vice versa, highly divergent paralogues indicate an old duplication. Within one genome, cutinase gene family members are more similar to each other than esterase gene family members to each other (Fig. 3C). Recent duplications are indicated by high similarity (above 75%) of cutinase paralogues with the exception of PsCut13. In contrast, all esterases have a paralogue of 60%-75% similarity indicating older duplications. Moreover, the *P. sojae* cutinases have been duplicated more recently than those of *P. ramorum* because the 4 cutinases of *P. ramorum* show lower similarity to their paralogues. Therefore, recent duplications are responsible for most members in the *P. sojae* cutinase family and lead to the expansion pattern.

Insertion of a His-Pro rich secreted protein gene cluster in the genome of *P. sojae*

In addition to the expansion of secreted protein encoding genes, other patterns like insertion/deletion are also identified. In this study, a deletion block is defined as a region containing 3 or more genes that lack homologues in the syntenic sequence. A total of 11 deletion blocks from 20-80 kb can be identified in a 2.4 Mb region. Effector-like genes are often found in these blocks. All 6 *Av1b* like genes are harbored in five of the deletion blocks, and a gene cluster encoding secreted proteins also resides in one of the blocks (Fig. 4A).

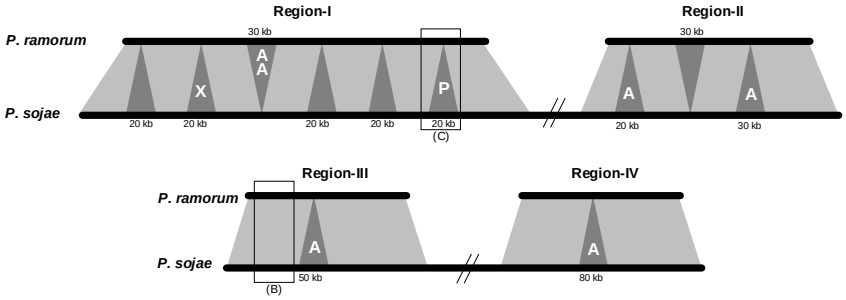
Table 1. Amino acid composition of the PsPHPex proteins. The average protein composition was calculated from 19276 *P. sojae* protein sequences.

name	size (aa)	signal peptide size (aa)	% Cys	% His	% Pro
PsPHPex1	275	18	0	19	44
PsPHPex2*	284	18	0	28	47
PsPHPex3	265	18	0	21	45
average *P. sojae* protein	491	-	2	2	5

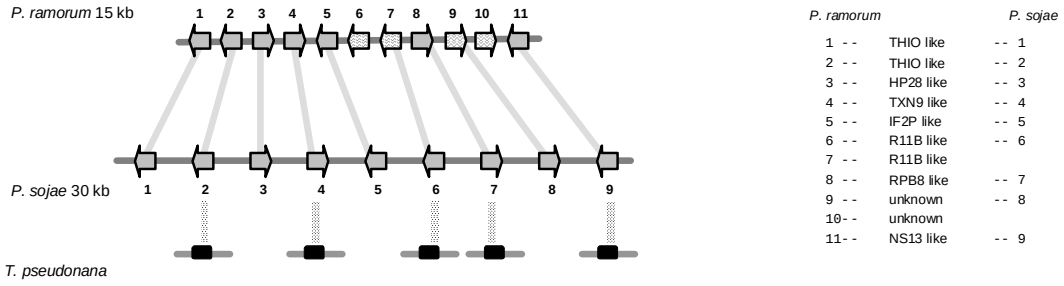
**PsPHPex2* is a pseudo-gene; the protein sequence was obtained after correction of the frameshift mutations.

To illustrate that these regions are largely syntenic despite of the deletions, the 30 kb sequence next to a deletion block is shown in Fig. 6A. For all 9 genes in the region of *P. sojae*, orthologues can be assigned to the genes in the syntenic region of *P. ramorum*. Although two genes have extra copies in *P. ramorum*, the order and orientation of the 9 genes are strictly conserved between the two genomes. When the 9 genes were BLASTed against the diatom *Thalassiosira pseudonana* genome, 6 homologues were found on different scaffolds but no syntenic region could be identified (Fig. 4B).

A



B



C

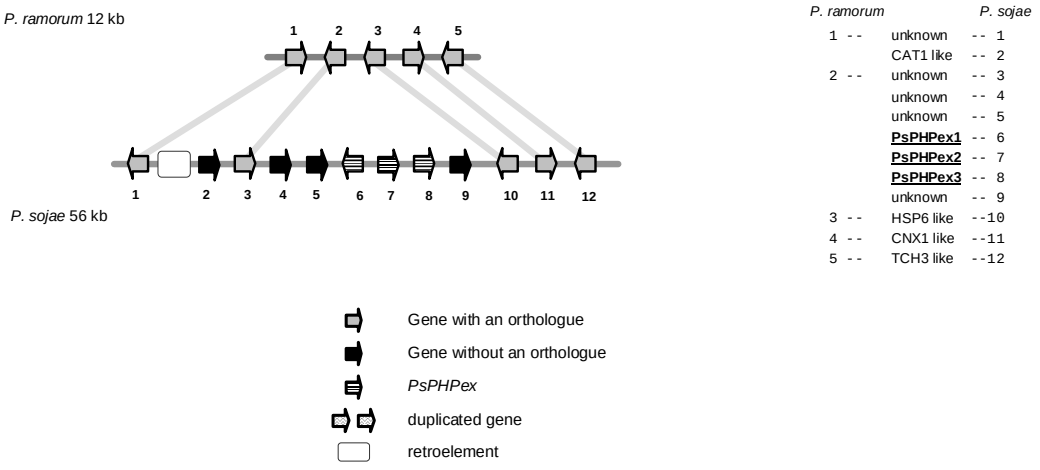


Fig. 4. A, Deletion blocks in the syntenic regions. Dark grey triangles represent deletion blocks. A stands for *Avr1b-like* genes residing in the block and double A refers to two *Avr1b-like* genes. P stands for *PsPHPex* genes and X for unknown secreted protein genes. The rectangles (B) and (C) are enlarged below. **B,** Conservation of gene order and orientation outside a deletion block. The homologues in the diatom *T. pseudonana* are on separate scaffolds. **C,** Insertion of the *PsPHPex* gene cluster in region-I in *P. sojae*. In **B** and **C** the length of the genes and the intergenic regions are not drawn on scale. The genes are named based on the best BLAST hit (*E* value cutoff < 1e-5) in the SwissProt database.

One of the deletion blocks contains a small gene cluster. It is comprised of three genes located 110 kb away from the *eli* cluster in *P. sojae*. The program SignalPV2.0 predicts the three encoded proteins to be secreted. The gene family has no homologues in *P. ramorum* and has unusual amino acid composition. As compared with the average amino acid composition derived from 19276 protein sequences in *P. sojae*, these three proteins have a high content of His (>19%) and Pro (>40%) because the average His and Pro percentage is 2% and 5%, respectively (Table 1). Therefore the three genes were designated as *PsPHPex-1* (*Phytophthora sojae* Pro/His rich *Phytophthora* extracellular protein)*, PsPHPex-2 and PsPHPex-3*. *PsPHPex-2* is a pseudo-gene carrying frameshift mutations whereas the other two have intact ORFs.

The 56 kb genomic region containing the gene cluster was compared with the 12 kb syntenic region of *P. ramorum* (Fig. 4C). The flanking genes of the *PsPHPex* family show conservation of gene order and orientation whereas the *PsPHPex* cluster with a few neighbour genes appear to be inserted in the genome of *P. sojae*, or from the point of view of *P. ramorum*, deleted from the syntenic region.

Avr1b and *Avr1b* homologues appear to be deleted from the syntenic regions

The *Avr1b* and *Avr1b-like* genes belong to a class of genes encoding small secreted proteins with highly divergent sequences. Despite of the sequence divergence, this class of proteins does share common features: they are small in size and lack Cys residues but most importantly, they all bear a 'RXLR-DEER' motif. The 'RXLR' motif is a novel motif possessed by several characterized oomycete effectors (Rehmany et al. 2005); AVR1b-like proteins in particular, have a conserved 'DEER' sequence after the 'RXLR' motif. Therefore this motif is named 'RXLR-DEER'.

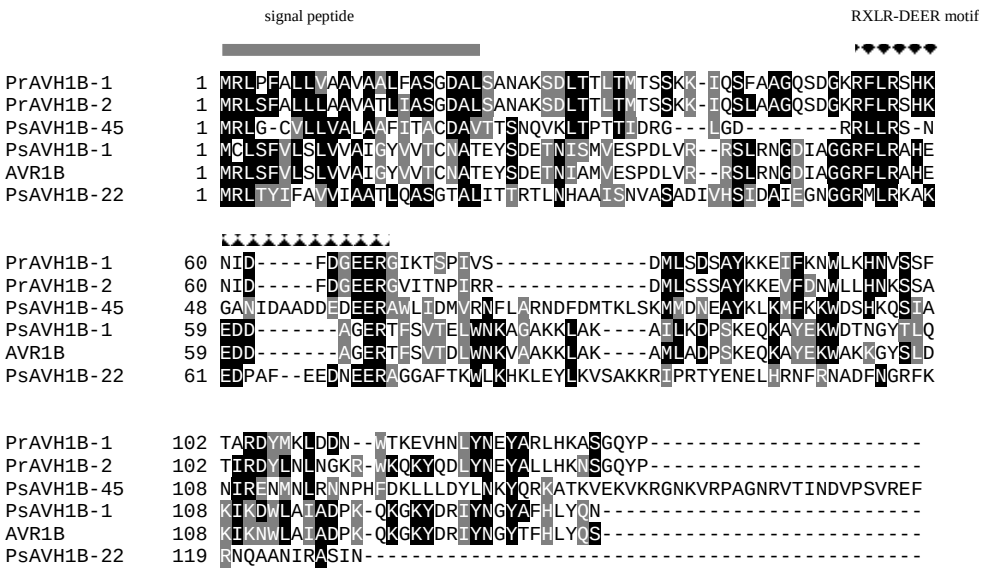


Fig. 5. Sequence alignment of AVR1b and five AVR1b-like proteins sharing the 'RXLR-DEER' motif. The predicted signal peptide is indicated by a gray bar and the 'RXLR-DEER' motif is indicated by dots above the alignment.

gene with an orthologue

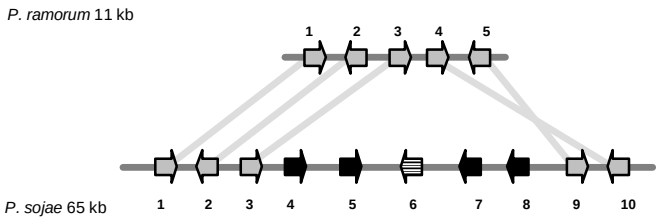
gene without an orthologue

Avr1b or Avr1b like gene

duplicated gene

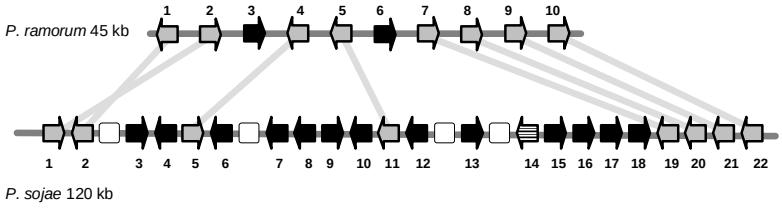
retroelement

A



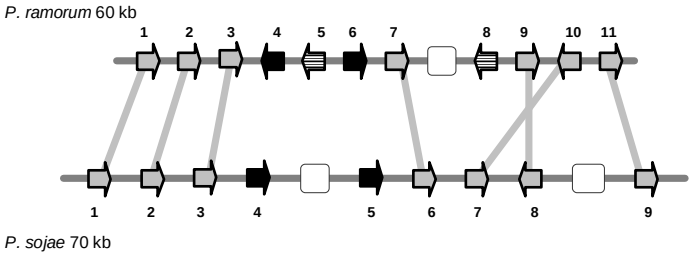
P. ramorum		*P. sojae*
1 --	YBEB like	-- 1
2 --	CALD like	-- 2
3 --	YADC like	-- 3
	unknown	-- 4
	unknown	-- 5
	Avr1b	-- 6
	FAB1 like	-- 7
	RRP4 like	-- 8
4 --	unknown	-- 9
5 --	PUR6 like	--10

B



P. ramorum		*P. sojae*
1 --	unknown	-- 1
2 --	unknown	-- 2
3 --	ATPG like	
	unknown	-- 3
	ABP1 like	-- 4
4 --	SFC1 like	-- 5
	unknown	-- 6
	YL52 like	-- 7
	unknown	-- 8
	unknown	-- 9
	unknown	--10
5 --	unknown	--11
6 --	unknown	
	unknown	--12
	unknown	--13
	PsAvh1b-1	--14
	unknown	--15
	unknown	--16
	unknown	--17
	unknown	--18
7 --	unknown	--19
8 --	unknown	--20
9 --	M3K7 like	--21
10--	CTR1 like	--22

C



P. ramorum		*P. sojae*
1 --	ACOD like	-- 1
2 --	OTC like	-- 2
3 --	DYHB like	-- 3
4 --	unknown	
5 --	**PrAvh1b-1**	
6 --	unkonwn	
	unknown	-- 4
	unknown	-- 5
7 --	AUT1	-- 6
8 --	PrArh1b-2	
9 --	unknown	-- 8
10--	RB2B like	-- 7
11--	TTC1 like	-- 9

Fig. 6. Deletions and re-arrangements of *Avr1b-like* genes in the genome of *P. ramorum* and *P. sojae*. The length of the genes and the intergenic regions are not drawn on scale. The genes are named based on the best BLAST hit (*E* value cutoff < 1e-5) in the SwissProt database. *Avr1b-like* genes are underlined. **A,** Deletion of *Avr1b* from *P. ramorum* (Region-III). **B,** Deletion of *PsAvh1b-1* from *P. ramorum* (Region-IV). **C,** Deletion of *PrAvh1b-1* and *PrArh1b-2* from *P. sojae* (Region-I).

Avr1b and *PsAvh1b-1* are located in region-III and region-IV, respectively. Additionally, two *Avr1b-like* genes (*Psavh1b-22* and *Psavh1b-45)* are located in the vicinity of the *ell13* gene in *P. sojae*, and two *Avr1b-like* genes named *Pravh1b-1* and *Pravh1b-2* are located in the neighborhood of the *eli* cluster *in P. ramorum*. The analysis included the close homologues of *Avr1b* because in *P. sojae*, *Pravh1b-1* is the closest homologue (89.1% similarity at protein level) to *Avr1b*, and in *P. ramorum*, *Pravh1b-1* and *Pravh1b-2* are the closest homologues (23.1% and 22.1% at protein level, respectively) to *Avr1b*.

These six proteins were used as a group of proteins bearing 'RXLR-DEER' motifs to compare their genomic organization between *P. sojae* and *P. ramorum*. The 'RXLR-DEER' motifs are shown in the sequence alignment (Fig. 5). Except for the 'RXLR-DEER' motif, low sequence conservation is found. Remarkably, in all six cases, deletion of the genomic regions containing the *Avr1b-like* sequences is found both in *P. sojae* and *P. ramorum* (Fig. 4). The 6 *Avr1b-like* genes are in 5 of the deletion blocks (Fig. 3A). In *P. ramorum*, homologues of *Avr1b* and *Psavh1b-1* genes are not present. Close examination shows that a 65 kb genomic region containing *Avr1b* is syntenic with a 11 kb region in *P. ramorum*. The 50 kb *Avr1b* region containing 5 genes in *P. sojae* appears to be deleted from *P. ramorum* (Fig. 6B). Similarly, analysis of the *Psavh-1b* region shows that a genomic sequence of 80 kb containing 14 genes including *Psavh1b-1* is missing from the genome of *P. ramorum* (Fig. 6C). In *P. sojae* two other *Avr1b-like* genes, *Psavh1b-22* and *Psavh1b-45,* are located next to the gene cluster *ell13* but in the syntenic region of in *P. ramorum* these two genes are missing.

In *P. ramorum*, *Pravh1b-1* and *Pravh1b-2* are located 30 kb apart and 28 kb away from the *eli* genes. A 60 kb region containing these two genes of *P. ramorum* is syntenic with a 70 kb region in *P. sojae*. A total of 7 pairs of orthologues can be identified in the region. The genes flanking the two *Avr1b* homologues show synteny between the two species. However, the two regions containing *Pravh1b-1* and *Pravh1b-2* are absent from the *P. sojae* genome. A total of 30 kb sequence appears to be deleted from the syntenic region (Fig. 6D).

Discussion

High co-linearity between the genomes of *P. sojae* and *P. ramorum*

The level of synteny reflects the divergence of the compared genomes. In *P. sojae* and *P. ramorum*, the gene order and orientation are largely conserved in the investigated regions, with only 15% of the orthologue pairs showing either position or direction reversal. The overall co-linearity suggests high similarity between the two genomes. This result agrees with the close phylogenetic relationship between *Phytophthora* species (Cooke et al. 2000). At larger phylogenetic distance, for example between *M. grisea* and *N. crassa,* no evidence for extensive regions of conserved synteny was observed. Only fragments containing several co-linear genes could be found between the two filamentous ascomycetes

that separated about 200 million years ago (Hamer et al. 2001; McCouch 2001). This study demonstrates that the evolutionary distance between *Phytophthora* species is appropriate to resolve synteny breakpoints. Mining genes from such breakpoints may result in candidates with functions in pathogenicity or avirulence.

Phytophthora belongs to the Stramenopiles (heterokonts) a taxonomic group that also includes golden-brown algae and diatoms. Diatoms are unicellular, photosynthetic organisms that play a central role in the global carbon cycle. To date one diatom genome has been sequenced, the marine diatom *Thalassiosira pseudonana* (Ambrust et al. 2004). There was no co-linearity found between *Phytophthora* and the marine diatom *T. pseudonana* in the 30kb sequences in region-III. This lack of synteny may be explained by the early divergence of these organisms.

The role of mobile elements

Due to lower selection pressure, sequence similarity in intergenic regions is usually much lower than in coding regions. Mobile elements, however, do not seem to follow the same evolutionary history as genes because their localization and orientation are largely random as compared to those of genes. Most likely, mobile elements have a more recent expansion. For example, in wheat and rice, different retrotransposons invaded the genomes over time repeatedly and created species-, genera-, and family-specific repeats and the genome size difference between these two species is mainly due to occurrence of retrotransposons (Sandhu and Gill 2002). Overall, more mobile elements were found in the genome of *P. sojae* than in the genome of *P. ramorum* (R.H.Y. Jiang, unpublished data). Their activity in *P. sojae* may be responsible for its larger genome size.

Mobile elements can not only mutate genes by transposition, they can also produce genome rearrangements through ectopic recombination between repeats created by their amplification. In many organisms, chromosomal rearrangements such as deletions, duplications, inversions, and translocations mediated by mobile elements have been found (Lonnig and Saedler 2002). In this study, mobile elements were found within and flanking the deletion blocks. Their role in transposition and assisting recombination may have contributed to genome rearrangements to form these deletion blocks. Changes mediated by mobile elements can be useful for natural populations in their adaptation to environmental constraints (Daboussi and Capy 2003). In *M. gri*sea, unstable subtelomeric regions are rich in repeats and mobile elements, and genomic rearrangements in these regions can give rise to new virulent alleles (Orbach et al. 2000). Expansions of retrotransposons and surface antigen gene families are associated with the synteny breakpoints in three related protozoan pathogens (El-Sayed et al. 2005). Facilitated by mobile elements, rearrangements of effector genes in *Phytophthora* may also lead to rapid adaptation to their hosts.

Expansion versus stability

For pathogens, pathogenicity related genes are responsible for the parasitic life style. As compared to genes with house keeping functions, they often evolve at a faster pace. When the rice blast pathogen *M. grisea* was compared to the related bread mold fungus *N. crassa*, several gene families were found expanded in *M. grisea.* Many of the expanded families are associated with pathogenicity such as genes encoding cutinases, subtilisin-like serine proteases and cysteine-rich polypeptides (Dean et al. 2005).

P. sojae and *P. ramorum* exhibit differences in their mating behavior, genome sizes and host specificity. For example, *P. sojae* has a very limited host range whereas *P. ramorum* has a very extended host range. These biological differences may also be reflected in their pathogenicity related genes and the expansion of cutinase genes may be an example of this. Cutinase and esterase gene families are located side by side in the genome, but the localization of the proteins is different; the cutinases are predicted to be secreted whereas the esterases remain inside the cell. The expansion of cutinase genes is in sharp contrast to the neighboring esterase gene family with a highly conserved pattern. Cutinases are produced by phytopathogenic fungi and by pollen to hydrolyse ester bonds in the cutin polymer. In pathogenic fungi, cutinases are secreted by penetration hyphae to assist in entering of the cuticle (Kolattukudy et al. 1995). Disruption of a cutinase gene in *Fusarium solani* f.sp. *pisi* decreased its virulence by more than 50% (Dantzig et al. 1986). Cutinases in *Phytophthora* may also be related to virulence. Adaptation to perform enzymatic tasks at the host surface may lead to frequent duplications of cutinase genes in *Phytophthora*.

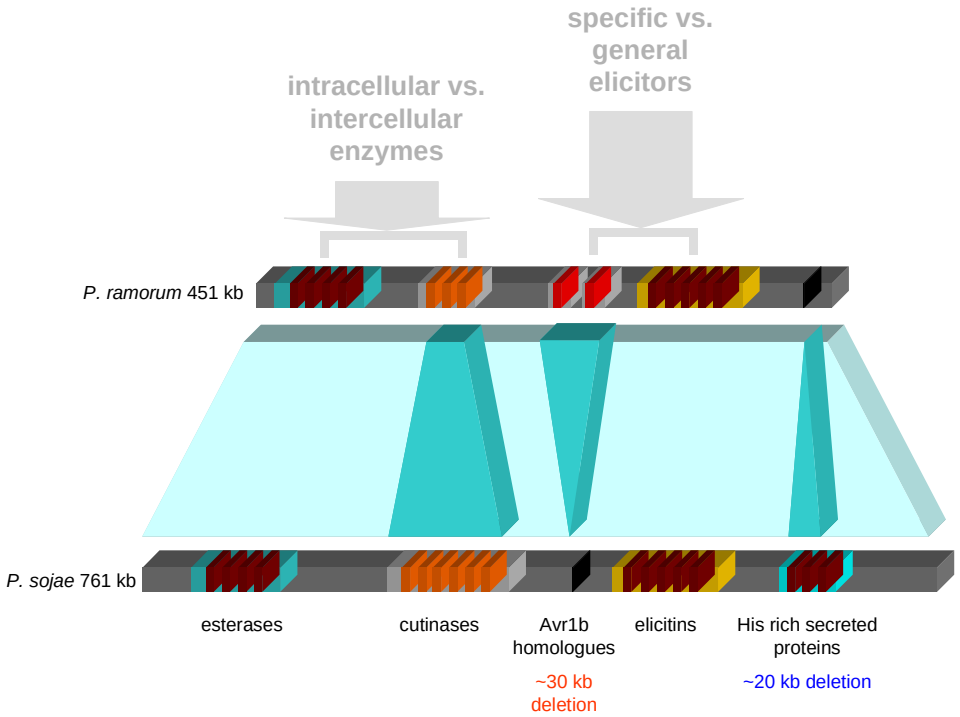


Fig. 7. Schematic representation of the syntenic Region-I in *P. sojae* and *P. ramorum*. Region-I contains esterase, cutinase and elicitin genes. Deletions and expansions are shown by triangular blocks. Genes are indicated by thin blocks. The number of thin blocks does not represent the true gene number.

One single genomic region can have several types of expansion or insertion/deletion patterns. A schematic drawing of region-I is shown in Fig. 7. Mutation and ectopic recombination must have occurred throughout the region and due to natural selection and genetic drift, some changes were lost whereas others were retained. Changes and adaptive advantages or disadvantages can give rise to contrasting evolutionary patterns. Contrasting patterns are illustrated with two groups of genes, those encoding the intracellular esterases and extracellular cutinases, and the ones encoding the general elicitin elicitors and the species-specific elicitors like AVR1b and its homologues (Fig. 7).

Genes encoding GPI anchored proteins are often clustered

GPIs anchor proteins or polysaccharides to a membrane. A wide range of possible roles of GPI membrane anchors has been suggested, including a space-filling role in the cell coating, shedding and turnover of membrane proteins, signal transduction and intracellular targeting (Thomas et al. 1990). GPI-anchored proteins are ubiquitously present in eukaryotes, and also exist in some Archaebacteria such as *Sulfolobus* (Ikezawa 2002). Some pathogens specialize in the synthesis of GPI-anchored proteins. The cell surface of trypanosomes, for example, is enveloped by a dense coat comprised of millions of variant surface glycoproteins (Overath and Engstler 2004).

The ELL13 group is the most divergent group in the elicitin family. Despite of their sequence diversity, the members of this group share common features such as a predicted signal peptide, a Cys rich domain; and in some cases, a predicted GPI anchor. It is remarkable that genes encoding two other families of secreted proteins with Cys-rich domains and GPI anchors are located next to the *ell13* genes. The close physical distance between these three groups of genes is conserved in the two genomes. These three groups belong to a larger group of *Phytophthora* genes all encoding glycosylated GPI anchored proteins. The mobile wall-less zoospores are thought to be covered by GPI anchor proteins and coated with oligosaccharides added by O-linked glycosylation (Jiang et al. 2006). A preliminary search showed that around 100 and 90 genes in *P. sojae* and *P. ramorum*, respectively, code for such genes, and many of them occur in clusters in the genome. Many of these surface glycoproteins contain repeat-like sequences and are evolving rapidly in *Phytophthora.* They may play a role in interactions with host plants or in recognition of gametes (Jiang et al. in preparation).

Highly dynamic RXLR-DEER bearing proteins

Despite of the high co-linearity of the two genomes, some regions turned out to be very flexible and do not show synteny. These regions often contain genes encoding secreted proteins. Breakage of synteny by virulence related genes may be a common theme in the genome evolution of pathogens. For example, synteny between two malaria parasites was found broken with genes interacting with host immune systems (Carlton et al. 2005). In plant pathogens, avirulence genes are able to interact with host resistance genes with high levels of specificity and they can undergo co-evolution with their hosts (Stahl

and Bishop 2000). Avirulence genes can be expected to belong to the most rapidly evolving genes in a pathogen. To date four oomycete *Avr* genes have been cloned, *Avr3a from P. infestans* (Armstrong et al. 2005), *Avr1b* from *P. sojae* (Shan et al. 2004), and *Atr1* (Rehmany et al. 2005) and *Atr13* (Allen et al. 2004) from *Hyaloperonospora parasitica.* Strikingly, all four oomycete AVRs bear a 'RXLR' motif (Rehmany et al. 2005) which suggests a shared novel interaction mechanism with the host. AVR3a, AVR1b and ATR1 also have the trailing 'DEER' motif and thus belong to the 'RXLR-DEER' group.

In the syntenic regions that we analyzed in this study, the six genes encoding proteins bearing a 'RXLR-DEER' motif all appear to be deleted from the other genome, suggesting that the genomic loci containing these genes have been rearranged resulting in loss or gain of these genes. AVR1b is one of the six 'RXLR-DEER' proteins and the avirulence determinant of *P. sojae* on soybean plants carrying the *Rps1b* gene. The ancestral *Avr1b* gene probably has been through high selection pressure during the interaction with plants and as a result is highly divergent from the homologue in *P. ramorum*. Due to selection pressure, the genomic *Avr1b* region may be a 'hotspot' for mutations and rearrangements eventually leading to deletion of the gene. Deletion of the other five *Avr1b-like* genes suggest a similar scenario of host selection pressure.

Material and methods

Genome databases

The genomic sequences and annotated protein sequences of *P. sojae* (version1), *P. ramorum* (version1), and of the diatom *Thalassiosira pseudonana* (Armbrust et al. 2004) were obtained from the website of the DOE Joint Genome Institute (http://www.jgi.doe.gov/genomes).

Orthologue search and determination of synteny

The elicitin and *ell13* sequences were identified by Jiang et al (2006). The *Avr1b* sequence was retrieved from GenBank (AAR05402). The homologues of *Avr1b* were determined by BLAST against whole genome sequences in *P. sojae* and *P. ramorum*. The surrounding genomic regions of the effector like genes were extracted. Subsequently every gene in the investigated region was used to BLAST against all the genes in the other genome; if gene pairs from two genomes shared best reciprocal BLAST hits, they were assigned as orthologues. Mobile elements mostly do not have homologues in the syntenic regions and they were left out from the analysis.

Bioinformatics tools

Sequences were analyzed with the Vector NTI 8 package. Multiple sequence alignment was performed using ClustalX 1.8 and for phylogenetic tree construction Molecular Evolutionary Genetic Analysis 2.1 (MEGA) (Kumar et al. 2001) was used. Phylogeny reconstruction was performed by Neighbor-Joining analysis. Poisson Correction (PC) was chosen as the distance parameter as specified in the program MEGA. The inferred phylogeny was tested by 1,000 bootstrap replicates. Dot blots were made using the program PipMaker (Schwartz et al. 2000). If alignments showed at least 100 bp homology with a nucleotide identity above 70%, the homologous stretch was plotted. For BLAST searches we used the NCBI BLAST program and the Standalone-BLAST Version 2.2.3 (Altschul et al. 1997). Signal peptides were predicted by SignalP 2.0 (Krogh et al. 2001). For GPI (glycosylphosphatidylinositol) anchor prediction, big-PI Plant Predictor (Eisenhaber et al. 2003b) was used. Calculation scripts were written in Python 2.2 (http://www.python.org) and are available from the authors upon request.

Acknowledgements

This work was financially supported by a grant to FG from NWO-Aspasia grant (015.000.057). RHYJ received a travel grant from the Technology Foundation STW for attending the *Phytophthora* annotation jamboree.

Literature cited

Allen, R.L., Bittner-Eddy, P.D., Grenville-Briggs, L.J., Meitz, J.C., Rehmany, A.P., Rose, L.E. and Beynon, J.L. 2004. Host-parasite coevolutionary conflict between *Arabidopsis* and downy mildew. Science 306:1957-60.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Armbrust, E.V., Berges, J.A., Bowler, C., Green, B.R., Martinez, D., Putnam, N.H., Zhou, S.G., Allen, A.E., Apt, K.E., Bechner, M., Brzezinski, M.A., Chaal, B.K., Chiovitti, A., Davis, A.K., Demarest, M.S., Detter, J.C., Glavina, T., Goodstein, D., Hadi, M.Z., Hellsten, U., Hildebrand, M., Jenkins, B.D., Jurka, J., Kapitonov, V.V., Kroger, N., Lau, W.W.Y., Lane, T.W., Larimer, F.W., Lippmeier, J.C., Lucas, S., Medina, M., Montsant, A., Obornik, M., Parker, M.S., Palenik, B., Pazour, G.J., Richardson, P.M., Rynearson, T.A., Saito, M.A., Schwartz, D.C., Thamatrakoln, K., Valentin, K., Vardi, A., Wilkerson, F.P. and Rokhsar, D.S. 2004. The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. Science 306:79-86.

Armstrong, M.R., Whisson, S.C., Pritchard, L., Bos, J.I., Venter, E., Avrova, A.O., Rehmany, A.P., Bohme, U., Brooks, K., Cherevach, I., Hamlin, N., White, B., Fraser, A., Lord, A., Quail, M.A., Churcher, C., Hall, N., Berriman, M., Huang, S., Kamoun, S., Beynon, J.L. and Birch, P.R. 2005. An ancestral oomycete locus contains late blight avirulence gene *Avr3a*, encoding a protein that is recognized in the host cytoplasm. Proc Natl Acad Sci USA 102:7766-71.

Bailey, J., Baertsch, R., Kent, W., Haussler, D. and Eichler, E. 2004. Hotspots of mammalian chromosomal evolution. Genome Biol 5:R23.

Baldauf, S.L. 2003. The deep roots of eukaryotes. Science 300:1703-6.

Buchrieser, C., Rusniok, C., Kunst, F., Cossart, P. and Glaser, P. 2003. Comparison of the genome sequences of *Listeria monocytogenes* and *Listeria innocua*: clues for evolution and pathogenicity. FEMS Immunol Med Microbiol 35:207-13.

Carlton, J., Silva, J. and Hall, N. 2005. The genome of model malaria parasites, and comparative genomics. Curr Issues Mol Biol 7:23-37.

Cooke, D.E., Drenth, A., Duncan, J.M., Wagels, G. and Brasier, C.M. 2000. A molecular phylogeny of *Phytophthora* and related oomycetes. Fungal Genet Biol 30:17-32.

Daboussi, M. J. and Capy, P. 2003. Transposable elements in filamentous fungi. Ann Rev Microbiol 57:275-299.

Dantzig, A.H., Zuckerman, S.H. and Andonov-Roland, M.M. 1986. Isolation of a *Fusarium solani* mutant reduced in cutinase activity and virulence. J Bacteriol 168:911-6.

Dean, R.A., Talbot, N.J., Ebbole, D.J., Farman, M.L., Mitchell, T.K., Orbach, M.J., Thon, M., Kulkarni, R., Xu, J.R., Pan, H., Read, N.D., Lee, Y.H., Carbone, I., Brown, D., Oh, Y.Y., Donofrio, N., Jeong, J.S., Soanes, D.M., Djonovic, S., Kolomiets, E., Rehmeyer, C., Li, W., Harding, M., Kim, S., Lebrun, M.H., Bohnert, H., Coughlan, S., Butler, J., Calvo, S., Ma, L.J., Nicol, R., Purcell, S., Nusbaum, C., Galagan, J.E. and Birren, B.W. 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature 434:980-6.

Eisenhaber, B., Maurer-Stroh, S., Novatchkova, M., Schneider, G. and Eisenhaber, F. 2003a. Enzymes and auxiliary factors for GPI lipid anchor biosynthesis and post-translational transfer to proteins. Bioessays 25:367-85.

Eisenhaber, B., Wildpaner, M., Schultz, C.J., Borner, G.H., Dupree, P. and Eisenhaber, F. 2003b. Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for *Arabidopsis* and rice. Plant Physiol 133:1691-701.

El-Sayed, N.M., Myler, P.J., Blandin, G., Berriman, M., Crabtree, J., Aggarwal, G., Caler, E., Renauld, H., Worthey, E.A., Hertz-Fowler, C., Ghedin, E., Peacock, C., Bartholomeu, D.C., Haas, B.J., Tran, A.-N., Wortman, J.R., Alsmark, U.C.M., Angiuoli, S., Anupama, A., Badger, J., Bringaud, F., Cadag, E., Carlton, J.M., Cerqueira, G.C., Creasy, T., Delcher, A.L., Djikeng, A., Embley, T.M., Hauser, C., Ivens, A.C., Kummerfeld, S.K., Pereira-Leal, J.B., Nilsson, D., Peterson, J., Salzberg, S.L., Shallom, J., Silva, J.C., Sundaram, J., Westenberger, S., White, O., Melville, S.E., Donelson, J.E., Andersson, B., Stuart, K.D. and Hall, N. 2005. Comparative Genomics of *Trypanosomatid* Parasitic Protozoa. Science 309:404-409.

Erwin, D.C. and Ribeiro, O.K. 1996. *Phytophthora* diseases worldwide. The American Phytopathological Society, St. Paul Minnesota USA

Jiang, R. H., Tyler B. M., Whisson S. C., Hardham A. R., and Govers, F. 2006. Ancient origin of elicitin gene clusters in *Phytophthora* genomes. Mol Biol Evol 23:338-351.

Hamer, L., Pan, H., Adachi, K., Orbach, M.J., Page, A., Ramamurthy, L. and Woessner, J.P. 2001. Regions of microsynteny in *Magnaporthe grisea* and *Neurospora crassa*. Fungal Genet Biol 33:137-43.

Hillier, L.W., Miller, W., Birney, E., Warren, W., Hardison, R.C., Ponting, C.P., Bork, P., Burt, D.W., Groenen, M.A., Delany, M.E., Dodgson, J.B., Chinwalla, A.T., Cliften, P.F., Clifton, S.W., Delehaunty, K.D., Fronick, C., Fulton, R.S., Graves, T.A., Kremitzki, C., Layman, D., Magrini, V., McPherson, J.D., Miner, T.L., Minx, P., Nash, W.E., Nhan, M.N., Nelson, J.O., Oddy, L.G., Pohl, C.S., Randall-Maher, J., Smith, S.M., Wallis, J.W., Yang, S.P., Romanov, M.N., Rondelli, C.M., Paton, B., Smith, J., Morrice, D., Daniels, L., Tempest, H.G., Robertson, L., Masabanda, J.S., Griffin, D.K., Vignal, A., Fillon, V., Jacobbson, L., Kerje, S., Andersson, L., Crooijmans, R.P., Aerts, J., van der Poel, J.J., Ellegren, H., Caldwell, R.B., Hubbard, S.J., Grafham, D.V., Kierzek, A.M., McLaren, S.R., Overton, I.M., Arakawa, H., Beattie, K.J., Bezzubov, Y., Boardman, P.E., Bonfield, J.K., Croning, M.D., Davies, R.M., Francis, M.D., Humphray, S.J., Scott, C.E., Taylor, R.G., Tickle, C., Brown, W.R., Rogers, J., Buerstedde, J.M., Wilson, S.A., Stubbs, L., Ovcharenko, I., Gordon, L., Lucas, S., Miller, M.M., Inoko, H., Shiina, T., Kaufman, J., Salomonsen, J., Skjoedt, K., Wong, G.K., Wang, J., Liu, B., Yu, J., Yang, H., Nefedov, M., Koriabine, M., Dejong, P.J., Goodstadt, L., Webber, C., Dickens, N.J., Letunic, I., Suyama, M., Torrents, D., von Mering, C., Zdobnov, E.M., et al. 2004. Sequence and comparative analysis of the chicken genome provide unique perspectives on vertebrate evolution. Nature 432:695-716.

Ikezawa, H. 2002. Glycosylphosphatidylinositol (GPI)-anchored proteins. Biol Pharm Bull 25:409-17.

Kamoun, S., van West, P., Vleeshouwers, V., de Groot, K.E. and Govers, F. 1998. Resistance of *Nicotiana benthamiana* to *Phytophthora infestans* is mediated by the recognition of the elicitor protein INF1. Plant Cell 10:1413-1425.

Kolattukudy, P.E., Rogers, L.M., Li, D., Hwang, C.S. and Flaishman, M.A. 1995. Surface signaling in pathogenesis. Proc Natl Acad Sci USA 92:4080-7.

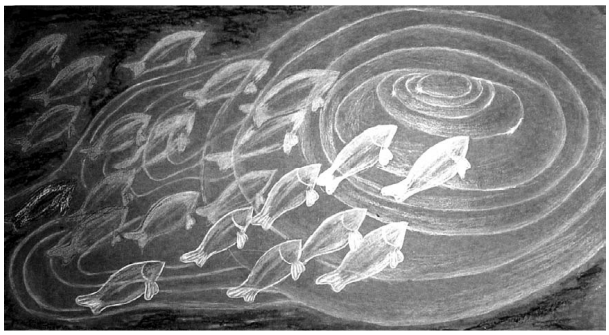
Krogh, A., Larsson, B., von Heijne, G. and Sonnhammer, E.L.L. 2001. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol Biol 305:567-580.

Kumar, S., Tamura, K., Jakobsen, I.B. and Nei, M. 2001. MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17:1244-5.

Longhi, S. and Cambillau, C. 1999. Structure-activity of cutinase, a small lipolytic enzyme. Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids 1441:185-196.
Lonnig, W.-E. and Saedler, H. 2002. Chromosome rearrangements and transposable elements. Annu Rev Genet 36:389-410.
McCouch, S.R. 2001. Genomics and synteny. Plant Physiol 125:152-5.
Mikes, V., Milat, M.L., Ponchet, M., Ricci, P. and Blein, J.P. 1997. The fungal elicitor cryptogein is a sterol carrier protein. FEBS Letters 416:190-192.
Mitreva, M., Blaxter, M.L., Bird, D.M. and McCarter, J.P. 2005. Comparative genomics of nematodes. Trends Genet In Press
Miyata, T., Takeda, J., Iida, Y., Yamada, N., Inoue, N., Takahashi, M., Maeda, K., Kitani, T. and Kinoshita, T. 1993. The cloning of PIG-A, a component in the early step of GPI-anchor biosynthesis. Science 259:1318-20.
Nespoulous, C., Gaudemer, O., Huet, J.C. and Pernollet, J.C. 1999. Characterization of elicitin-like phospholipases isolated from *Phytophthora capsici* culture filtrate. FEBS Lett 452:400-6.
Ollis, D.L., Cheah, E., Cygler, M., Dijkstra, B., Frolow, F., Franken, S.M., Harel, M., Remington, S.J., Silman, I. and Schrag et, a. 1992. The alpha/beta hydrolase fold. Protein Eng 5:197-211.
Orbach, M.J., Farrall, L., Sweigard, J.A., Chumley, F.G. and Valent, B. 2000. A telomeric avirulence gene determines efficacy for the rice blast resistance gene *Pi-ta*. Plant Cell 12:2019-2032.
Overath, P. and Engstler, M. 2004. Endocytosis, membrane recycling and sorting of GPI-anchored proteins: *Trypanosoma brucei* as a model system. Mol Microbiol 53:735-44.
Rehmany, A.P., Gordon, A., Rose, L.E., Allen, R.L., Armstrong, M.R., Whisson, S.C., Kamoun, S., Tyler, B.M., Birch, P.R. and Beynon, J.L. 2005. Differential Recognition of Highly Divergent Downy Mildew Avirulence Gene Alleles by *RPP1* Resistance Genes from Two *Arabidopsis* Lines. Plant Cell 17:1839-50.
Ricci, P., Trentin, F., Bonnet, P., Venard, P., Moutonperronnet, F. and Bruneteau, M. 1992. Differential production of parasiticein, an elicitor of necrosis and resistance in tobacco, by isolates of *Phytophthora Parasitica*. Plant Pathol 41:298-307.
Rizzo, D.M., Garbelotto, M. and Hansen, E.M. 2004. *Phytophthora Ramorum*: Integrative Research and Management of an Emerging Pathogen in California and Oregon Forests. Annu Rev Phytopathol.
Sandhu, D. and Gill, K.S. 2002. Gene-containing regions of wheat and the other grass genomes. Plant Physiol 128:803-11.
Schwartz, S., Zhang, Z., Frazer, K.A., Smit, A., Riemer, C., Bouck, J., Gibbs, R., Hardison, R. and Miller, W. 2000. PipMaker--a web server for aligning two genomic DNA sequences. Genome Res 10:577-86.
Shan, W., Cao, M., Leung, D. and Tyler, B.M. 2004. The *Avr1b* locus of *Phytophthora sojae* encodes an elicitor and a regulator required for avirulence on soybean plants carrying resistance gene *Rps1b*. Mol Plant Microbe Interact 17:394-403.
Shearn, C.T., Walker, J. and Norris, F.A. 2001. Identification of a novel spliceoform of inositol polyphosphate 4-phosphatase type I alpha expressed in human platelets: structure of human inositol polyphosphate 4-phosphatase type I gene. Biochem Biophys Res Commun 286:119-25.
Stahl, E.A. and Bishop, J.G. 2000. Plant-pathogen arms races at the molecular level. Curr Opi Plant Biol 3:299-304.
Thomas, J.R., Dwek, R.A. and Rademacher, T.W. 1990. Structure, biosynthesis, and function of glycosylphosphatidylinositols. Biochemistry 29:5413-22.
van't Slot, K.A.E. and Knogge, W. 2002. A dual role for microbial pathogen-derived effector proteins in plant disease and resistance. Crit Rev Plant Sci 21:229-271.
Vauthrin, S., Mikes, V., Milat, M.L., Ponchet, M., Maume, B., Osman, H. and Blein, J.P. 1999. Elicitins trap and transfer sterols from micelles, liposomes and plant plasma membranes. Biochimica Et Biophysica Acta-Biomembranes 1419:335-342.
Yu, J., Hu, S., Wang, J., Wong, G.K.-S., Li, S., Liu, B., Deng, Y., Dai, L., Zhou, Y., Zhang, X., Cao, M., Liu, J., Sun, J., Tang, J., Chen, Y., Huang, X., Lin, W., Ye, C., Tong, W., Cong, L., Geng, J., Han, Y., Li, L., Li, W., Hu, G., Huang, X., Li, W., Li, J., Liu, Z., Li, L., Liu, J., Qi, Q., Liu, J., Li, L., Li, T., Wang, X., Lu, H., Wu, T., Zhu, M., Ni, P., Han, H., Dong, W., Ren, X., Feng, X., Cui, P., Li, X., Wang, H., Xu, X., Zhai, W., Xu, Z., Zhang, J., He, S., Zhang, J., Xu, J., Zhang, K., Zheng, X., Dong, J., Zeng, W., Tao, L., Ye, J., Tan, J., Ren, X., Chen, X., He, J., Liu, D., Tian, W., Tian, C., Xia, H., Bao, Q., Li, G., Gao, H., Cao, T., Wang, J., Zhao, W., Li, P., Chen, W., Wang, X., Zhang, Y., Hu, J., Wang, J., Liu, S., Yang, J., Zhang, G., Xiong, Y., Li, Z., Mao, L., Zhou, C., Zhu, Z., Chen, R., Hao, B., Zheng, W., Chen, S., Guo, W., Li, G., Liu, S., Tao, M., Wang, J., Zhu, L., Yuan, L. and Yang, H. 2002. A Draft Sequence of the Rice Genome (*Oryza sativa L. ssp. indica*). Science 296:79-92.
Zhang, X. and Majerus, P.W. 1998. Phosphatidylinositol signalling reactions. Semin Cell Dev Biol 9:153-60.

Author-recommended internet resources

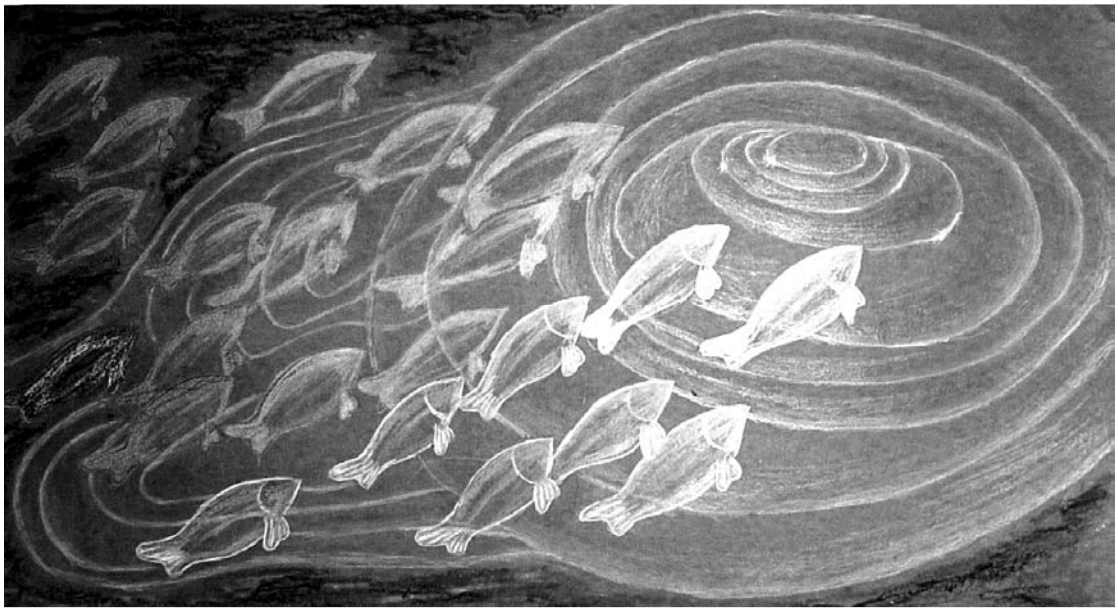
DOE Joint Genome Institute (JGI) *P. sojae* genome database:
http://genome.jgi-psf.org/sojae1
DOE Joint Genome Institute (JGI) *P. ramorum* genome database:
http://genome.jgi-psf.org/ramorum1
The Virginia Bioinformatics Institute (VBI) Microbial Database hosts a range of microbial sequences including the genome sequence and annotation data of *P. sojae* and *P. ramorum*:
http://phytophthora.vbi.vt.edu



Chapter 7

Different paces of evolution in the secretome of *Phytophthora*

Rays H.Y. Jiang, Joe Win, Sophien Kamoun, Brett M. Tyler and Francine Govers



Different paces of evolution in the secretome of *Phytophthora*

Rays H.Y. Jiang[1], Joe Win[2], Sophien Kamoun[2], Brett M. Tyler[3] and Francine Govers[1]

[1]Laboratory of Phytopathology, Plant Sciences Group, Wageningen University, Binnenhaven 5, NL-6709 PD Wageningen and Graduate School Experimental Plant Sciences, The Netherlands
[2]Department of Plant Pathology, Ohio Agricultural Research and Development Center, The Ohio State University, Wooster, OH44691, USA
[3]Virginia Bioinformatics Institute, Virginia Polytechnic and State University, Blacksburg VA, USA

For correspondence: E-mail Francine.Govers@wur.nl; Tel. +31 317 483138; Fax +31 317 483412

Alternative title

The reservoir of secreted protein encoding genes in *Phytophthora sojae and Phytophthora ramorum*

Keywords

secretome, comparative genomics, *Phytophthora*

Abstract

Pathogenic bacteria, fungi and oomycetes all possess a wide range of molecules to interact with their hosts. Proteins secreted by pathogens are of ultimate interest because they might be effector molecules that play important roles in pathogenesis. A bioinformatic approach was used to reveal the reservoir of secreted proteins from all annotated genes in two sequenced *Phytophthora* genomes. A total of 1464 and 1188 putative secreted protein genes from *Phytophthora sojae* and *Phytophthora ramorum,* respectively, were identified and investigated for their sequence diversity, expansion of family members and genome organization. More than 80% of the secreted protein genes form gene families, and many of the families are clustered in the genome. The secretome appears to evolve at a faster pace than the average genome. Individual families within the secretome are evolving at different paces as well. The fast evolving genes include several repeat containing families and the RXLR-DEER super-family that may play different roles in the interaction with the hosts.

Introduction

Microbes secrete numerous proteins to adapt to their environment. Vital functions such as substrate degradation, cell-wall turnover, communication with each other and interaction with hosts are carried out by the microbial secretome (the sum of all secreted proteins). Some micro-organisms are endowed with natural capacity to secrete large quantities of proteins and are commercially exploited for protein production. For example, the secretome of the 'cell factory' *Bacillus subtilis* is estimated to consist of approximately 300 proteins (Tjalsma et al. 2000).

For pathogenic microbes, the secretome performs important functions at the frontline of host-pathogen interactions via a large repertoire of effector molecules to establish efficient colonization and to suppress or evade host defense responses. To deploy effectors to the place of their action, microbes have developed astonishingly diverse secretory machineries. Bacterial pathogens use various mechanisms such as the Sec pathway, autotransporters and type I-IV secretion systems to deliver proteins for adhesion, internalization and host maneuver (Lee and Schneewind 2001). The bacterial type III secretion system is used to inject effectors into host cells of animals and plants. Repertoires of type III effectors in *Pseudomonas syringae* are highly dynamic and varied, and this variation is likely to facilitate adaptation to different hosts (Guttman et al. 2002; Greenberg and Vinatzer 2003). In the protozoan *Plasmodium falciparum,* protein delivery into host erythrocytes is achieved by a novel mechanism that is based on recognition of a conserved pentameric motif behind an eukaryotic signal peptide. This malaria parasite delivers an estimated secretome consisting of 400 proteins via this system into the host cells to carry out virulent and host remodeling tasks (Hiller et al. 2004; Marti et al. 2004). In the rice blast fungus *Magnaporthe grisea***,** the secretion relies on the typical eukaryotic N-terminal signal peptide for governing the protein across the membrane of the endoplasmic reticulum. The *M. grisea* secretome consists of 739 proteins and is approximately twice as large as the secretome of the non-pathogenic fungus *Neurospora crassa*. The larger secretome of *M. grisea* is partly due to expansion of some protein families that are involved in virulence of the fungus (Dean et al. 2005).

In this study we analyzed the secretome of two important plant pathogens, *Phytophthora sojae* and *Phytophthora ramorum. P. sojae* is the causal agent of soybean root rot, a narrow host pathogen that causes around $1-2 billion losses per year worldwide. *P. ramorum* has a broad host range. It is well know as the causal agent of the sudden oak death syndrome that is destroying many oak trees along the west coast of the US (Rizzo et al. 2005). *Phytophthora* is a genus that consists of more than 65 destructive plant pathogens (Erwin and Ribeiro 1996), and belongs to the oomycetes, a stramenopile lineage that evolved independent from plants, animals and fungi (Baldauf 2003). Therefore the pathogenicity of *Phytophthora* must have evolved independently from phytopathogenic fungi despite the morphological similarities between the two (Latijnhouwers et al. 2003). The secretome of *Phytophthora* holds the keys that can unlock some of the unique features of oomycete pathogens. In 2004, the genome sequences of *P. sojae* and *P. ramorum* were released and this offered the opportunity to investigate the complete set genes encoding secreted protein. *P. sojae* and *P. ramorum* have different

genome sizes and a different mating system. *P. sojae* is homothallic with a genome size of 95 Mb and *P. ramorum* is heterothallic, and has a smaller genome of 65 Mb. The comparison of the secretomes of the two pathogens should reveal similarities of oomycete pathogens as well as the differences.

The secretome is a particularly dynamic part of the proteome because it is responsible for many of the interactions with the changing environment. Natural selection is acting on each individual gene, and as a result, sequence variations between two related genomes are not uniformly distributed. The dynamic nature of a secretome can be seen from comparisons between a pathogen and a non-pathogen, between two pathogen species and even between different strains of one pathogen. For example, the comparison between the pathogen *Listeria monocytogenes* and the nonpathogenic species *Listeria innocua* showed that the difference between the two manifests strongest in the secretome (Trost et al. 2005). A proteomics comparison between *Pseudomonas aeruginosa* isolates showed a conserved species-specific core proteome, but the profiles of the secretomes were highly variable and differed between clones (Wehmhoner et al. 2003). When examining and comparing *Phytophthora* secretomes, we expect to discover some of the most dynamic differences between the two genomes.

As *Phytophthora* is a genus of successful plant pathogens, the secretome analysis should offer us insight into the mechanisms and evolution of plant-pathogen interactions. To search for putative virulence factors in the secretome, differences in evolutionary pace can be used as one of the selection criteria. Fast evolving genes can be detected by their high level of sequence divergence resulting from accelerated alteration of gene sequences under diversifying selection.

Here, we report detailed computational analyses to characterize the secretomes in the two *Phytophthora* species, and to compare the two secretomes in search of groups of proteins with different paces of evolution.

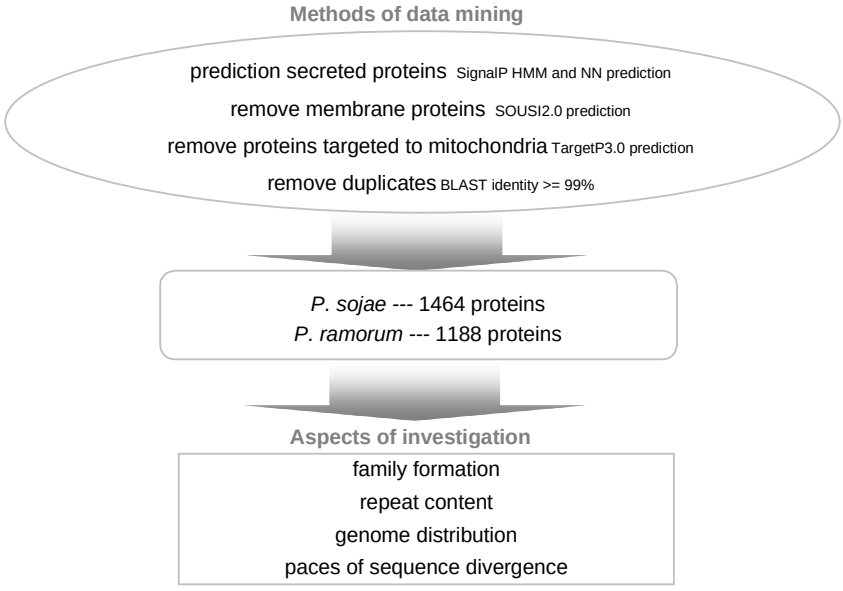
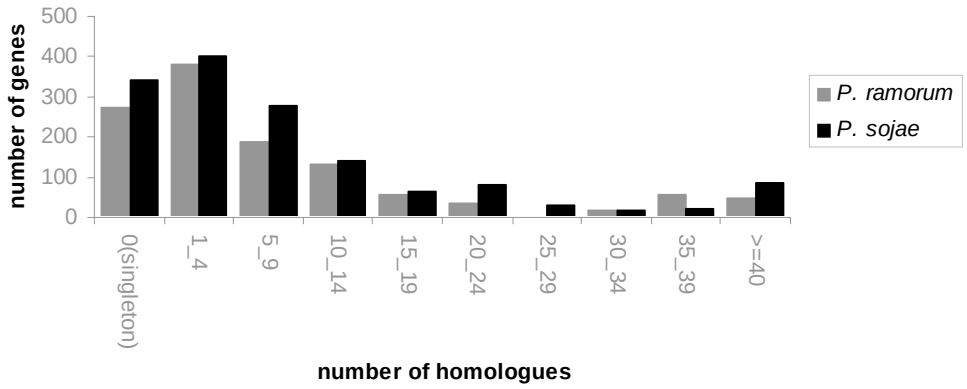


Fig. 1. Outline of data mining approaches

The reservoir of secreted protein encoding genes in *P. sojae* and *P. ramorum*

All of 19,027 annotated genes in the genome of *P. sojae* and 15,743 genes in the genome of *P. ramorum* were subjected to signal peptide prediction and several additional bioinformatics predictions as described in materials and methods and outlined in Fig. 1. A total of 1464 and 1188 proteins were found to constitute the secretome of *P. sojae* and *P. ramorum,* respectively. These proteins range from 100 to 840 amino acids in length and with an average size of 328 and 346 amino acids in *P. sojae* and *P. ramorum*, respectively. In sheer number, *P. sojae* had more secreted protein genes in its genome than *P. ramorum*. In percentage, *spe* (Secreted Protein Encoding) genes accounted for about 8% of the total genes in both genomes. To further characterize these large numbers of *spe* genes, their features such as family formation, repeat content, genome distribution and sequence divergence were investigated.

A



B

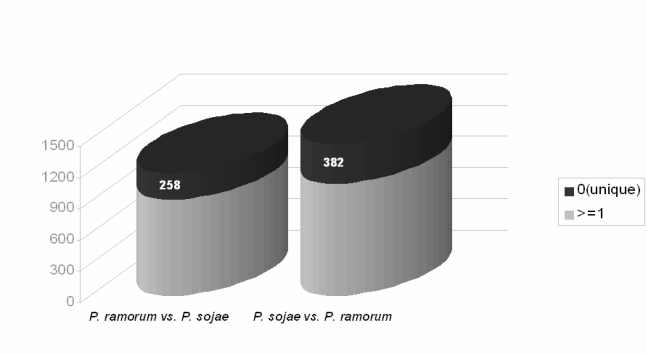


Fig. 2. Family formation of secreted proteins and comparison between *P. sojae* and *P. ramorum*. A, number of homologues of secreted proteins found in the secreted protein data set of *P. sojae* and *P. ramorum*. To assign paralogues, BLASTP searches of each protein against the secreted protein set were performed (*E* value less than 1E-05). **B,** number of homologues of secreted proteins found in the other proteome. To assign homologues, BLASTP searches of each protein against the secreted protein set of the other species were performed (*E* value less than 1E-05).

The majority of *spe* genes form gene families

Gene families are considered to arise from gene duplications. Redundancy resulting from duplication events offers the opportunity to develop new functions before the extra copy is lost due to genetic drift or deleterious mutations (Lynch and Conery 2000). In *Phytophthora* most of the extracellular proteins described to date are encoded by multigene families, such as the family of polygalacturonases, necrosis inducing proteins and elicitins (Gotesson et al. 2002; Qutob et al. 2002; Torto et al. 2002; Jiang et al. 2006).

To investigate the extent of gene family formation, within each species we searched for paralogues of each protein by performing BLASTP searches against the complete set of secreted proteins (*E* value less than 1E-05). By this criteria, the majority of the secreted proteins had homologues within the secretome. Singletons, defined as proteins without a homologue, take up 15% and 18% of the secretomes in *P. sojae* and *P. ramorum*, respectively. The larger secretome size in *P. sojae* (1464 versus 1188 in *P. ramorum*) was accounted for by the larger number of singletons as well as larger family sizes. In both *Phytophthora* species, gene families with 2-5 homologues were most common, whereas larger families are rare. Around 30% of the *spe* genes belong to families with more than 10 homologues and around 10% to families with more than 30 members (Fig. 2A). The presently identified families with more than 10 members are listed in table 1.

Gene families with more than 10 members were manually annotated according to the top BLAST hits in SwissProt and GenBank databases. A wide range of putative functions such as elicitor activity, cell wall degradation and enzyme inhibition was found within the largest families (38 families). However, several of these had no known function (Table 1). The largest family according to this grouping contains around 50 members. Three families bearing the RXLR-DEER motif were amongst the largest families. Other large families with elicitor activities are the NIP family, the elicitin family and the crn (Crinkling and Necrosis-inducing protein) family, that have 52, 21 and 15 members in *P. sojae* respectively (Table 1).

To evaluate the potential role of being a general or specific elicitor, the sequence divergence of proteins between the two secretomes was analyzed. For elicitors, the phylogenetic distribution and level of sequence divergence can be indicative for their function. Elicitors that are ubiquitously distributed in the *Phytophthora* genus have been shown to have elicitor activity on a broad range of host species, such as the elicitins (Jiang et al. 2006). In contrast, elicitors specific to a species are likely to trigger host response with high level specificity, and this ‘uniqueness’ of the elicitor may be a hallmark for the ecotype- or cultivar- specific avirulence genes. To find the number of homologues between the two secretomes, and unique proteins in each secreteome, each protein was BLASTed against the whole genome protein set of the other species. A BLAST hit with an *E* value less than 1E-05 was considered as being homologous, a protein with best BLAST hit *E* value larger than 1E-05 was considered to be unique for that species. There were 382 unique secreted proteins in *P. sojae* and 258 in *P. ram*orum (Fig. 2B). Around 20% of the secreted proteins has 1-4 homologues and around 10% have more than 40 homologues in the proteome of the other species.

Table 1. Secreted protein families in *P. sojae* and *P. ramorum* .

cluster family code[a]	family description	number in *P. ramorum*	number in *P. sojae*	gene cluster in *P. ramorum* [b]	gene cluster in *P. sojae* [b]	family divergence [c]
1 [d]	RXLR-DEER family-I	48	48			++++
2	NIP (IPR008701)	46	52		9 (89 kb) 8 (19 kb)	+
3	unknown	41	46	10 (39 kb)	6 (36 kb) 9 (42 kb)	+
4	unknown	24	26	10 (144 kb)	6 (27 kb)	+
5	RXLR-DEER family-II	21	9			++++
6 [d]	Elicitin (IPR002200)	19	21	11 (60 kb)	10 (115 kb)	+
7	Allergen V5/Tpx-1 (IPR001283)	18	32	5 (6 kb) 5 (46 kb)	6 (29 kb) 9 (95 kb)	+
8	Phosphopantetheine attachment site (IPR006162)	16	23		8 (71 kb)	+
9	Serine protease, trypsin family (IPR001254)	16	25		7 (96 kb)	+
10	unknown	15	7	6 (11 kb)	6 (37 kb) 6 (16 kb)	++
11	contain Gly repeats	15	26		7 (55 kb)	++++
12	Apple domain (IPR000177)	14	16	5 (7 kb)		+
13 [d]	RXLR-DEER family-III	13	6			+++
14	unknown	13	15		6 (50 kb)	+
15	EGF-like (IPR006209)	13	15			++
16	polygalacturonase (Glycoside hydrolase, family 28 IPR000743)	13	22	12 (45 kb)	14 (93 kb)	+
17	contain Pro repeats	12	11			+++
18	Palmitoyl-protein hydrolase (Esterase/lipase/thioesterase IPR000379)	11	5			+
19	Intradiol ring-cleavage dioxygenase (IPR000627)	11	7			+
20	Glycoside hydrolase, family 17 (IPR000490)	11	10			+
21	Pectate lyase (IPR002022)	11	13			+
22	Glycoside hydrolase, family 3 (IPR001764)	11	15			+
23	Tyrosinase (IPR002227)	11	16	5 (42 kb)	7 (72 kb)	+
24	Pectinesterase (IPR000070)	10	15			+
25	Neutral zinc metallopeptidases (IPR006025)	9	10	5 (22 kb)		+
26	Metallo-phosphoesterase (IPR004843)	9	12			+
27	Pectate lyase (IPR002022)	8	19			+
28	unknown	5	10			++
29	Glycoside hydrolase, family 12 (IPR002594)	5	11		7 (58 kb)	++
30	Carbonic anhydrase (IPR001148)	5	11			+
31 [d]	crn like	4	15			++
32	Cutinase (IPR000675)	4	15		9 (25 kb)	+
33 [d]	Kazal type protease inhibitor (IPR002350)	2	14		6 (33 kb)	++++
34 [e]	EHD-I	0	6			++++
35 [e]	EHD-II	6	0			++++
36 [e]	EHD-III	3	5			++++
37 [e]	EHD-M96	9	13			++++
38 [f]	PsGSpex-1	0	17			++++

[a] Protein sequences were clustered into families by all against all BLAST using threshold limit *E* value less than 1E-30. Functional annotation was manually curated by examining BLAST searches to the GenBank NR protein database and SwissProt sequences. Sequence clusters with at least 10 members in either species are shown.

[b] Clustering of more than 5 gene family members (intergenic region less than 50 kb) in one physical region in the genome is listed

[c] A protein lack a highly similar BLAST hit (sequence similarity < 55%) in the other proteome is counted as divergent and a protein having highly similar BLAST hit (sequence similarity > 55%) is counted as conserved. Families comprised of more than 50% divergent members are considered to be highly divergent (++++), families comprised of more than 30% divergent members are considered moderately divergent (+++), families comprised of more than 10% divergent members are counted as slightly divergent (++), and families with more than 90% conserved members are counted as conserved (+).

[d] Part of a larger gene family

[e] Member count includes proteins without signal peptides and corrected pseudogenes

[f] PsGSpex-1 family belongs to the small un-annotated proteins

The secretome is a dynamic part of the proteome

In *P. sojae* and *P. ramorum*, secretomes take up a similar percentage of the entire proteome. Orthologues reflect speciation events in evolution. The number of orthologues shared between species indicates the relationship between the genomes. In the two genomes, 9768 orthologue pairs can be assigned based on best bi-directional BLAST hits (Tyler et al., in preparation). Among the *spe* genes, a total of 572 pairs can be assigned as orthologues. In both species, the percentage of orthologues is higher in the whole genome than in the secretome: 62% of all genes and 48% of the *spe* genes were orthologues in *P. ramorum*. Similar results were found in *P. sojae*, 51% of all genes and 39% of the *spe* genes were orthologues.

Higher percentage of unknown genes is present in the secretome than the proteome. About 20% of the proteins in the proteome have no homology to known proteins and lack an interpro domain (Tyler et al., in preparation), whereas in the secretomes of *P. sojae* and *P. ramorum*, over 40% of the proteins lack homology to characterized domains or sequences.

The percentage of unique genes is lower in the genome than in the secretome. Assuming genes without BLAST hits in the other genomes as unique genes, the *P. sojae* secretome had 26% unique genes whereas the whole genome had 9% unique genes (Tyler et al., in preparation). *P. ramorum* gave similar results: 22% unique genes in the secretome and 4% unique genes in the genome. These numbers indicate that the secretome is evolving faster than the average genome.

Clustering of gene families in the genome

In *Phytophthora,* most previously identified *spe* genes that belong to one family are clustered in the genome. Examples are the *ipiO* family (Pieterse et al. 1994), *Avr3a* and its homologues (Armstrong et al. 2005), the *polygalacturonase* family (Gotesson et al. 2002) and the elicitin family (Jiang et al. 2006). The clustering of gene family members was estimated. In general, *spe* genes appear to be distributed all over the genome. Nearly one half of the total *spe* genes is scattered in various scaffolds while many gene family members form clusters. In this study, genes were deemed to be in a cluster if their intergenic regions are less than 50 kb in length.

In *P. sojae*, 1464 genes are distributed over 222 scaffolds. Among these, 663 (45%) are members of gene families and form clusters in the genome. In *P. sojae*, 14 members of the polygalacturonase gene family were found to be clustered, as were 8 genes encoding NIPs (Necrosis Inducing Protein) and 10 elicitin genes for each family (Fig. 3). Members of these gene families were also found in clusters in *P. ramorum.* In *P. ramorum*, 1188 genes were distributed over 268 scaffolds, 467 (40 %) of these were gene family members and formed clusters in the genome (Fig. 3). 19% and 12% of the *spe* genes form clusters comprised of 5 or more members in *P. sojae* and *P. ramorum*, respectively. The majority of the clusters contains 2-3 genes of one family.

Noticeably, genes encoding proteins bearing the 'RXLR-DEER' motif are mostly scattered over the genome. In *P. sojae*, only 18 'RXLR-DEER' genes appear in small clusters, i.e., 6 clusters comprised of 2 homologues and 2 clusters comprised of 3 homologues. The rest 'RXLR-DEER' genes do not from clusters in the genome.

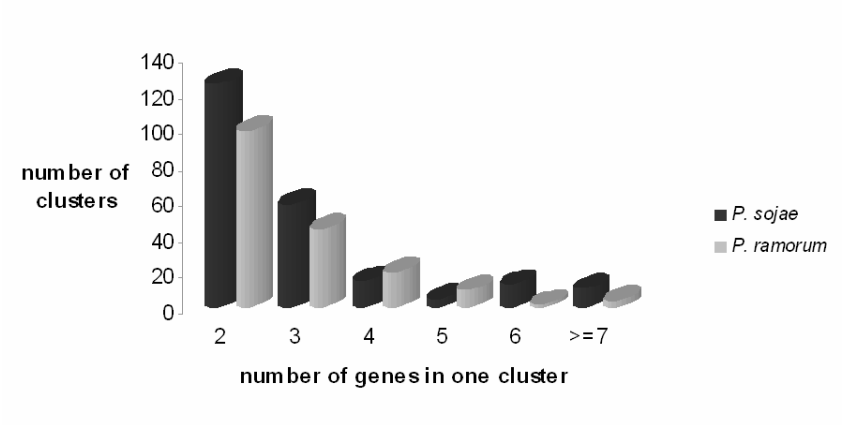


Fig. 3. Clustering of gene family members in the genome. Clustering is considered if the genes are located on the same scaffold and the intergenic regions between genes are less than 50 kb. Genes in one cluster show BLASTN homology (*E* vlaue less than 1E-05) to one another.

Large groups of secreted proteins contain sequence repeats

Repeated sequences occur in many known proteins, and they can have important functions in for example, surface adhesion or molecular interaction (Andrade et al. 2001; Main et al. 2003). In *P. infestans,* surface proteins containing tandem repeats have been previously described such as the Car90 protein (Gornhardt et al. 2000) and some of the elicitins (Jiang et al. 2006). Both Car90 and elicitins typically have simple sequence repeats in the C-terminal part (50-200 amino acids) of the protein. To determine whether such repeats occur in the secreted proteins, we analyzed 100 amino acid of the most C-terminal part of each secreted protein. In 202 *P. sojae* and 187 *P. ramorum* proteins, we found simple amino acids repeats that make up more than 20% of their C-terminal domains. The largest group of repeats is based on Thr or Ser residues while another major type of repeat contained Pro, Asp, Glu, Gly and Asp. The percentage of each type of repeat is shown in Fig. 4.

The most commonly occurring repeats (Thr-Ser based) suggest the presence of O-linked glycosylation sites in the C-terminal domains (Wilson et al. 1991). Thr-Ser rich O-glycosylated domains are found in many cell-wall-associated proteins (Jentoft 1990). Therefore the abundance of this type of repeats indicates that the secretome contains large amounts of surface glycoproteins.

Some secreted proteins are almost entirely composed of repeats. For example, 14 proteins in *P. sojae* and 9 proteins in *P. ramorum* sharing homology with the mating induced protein M96 of *P. infest*ans (Fabritius et al. 2002) have the alternating acidic and basic residue 'HD' repeat in the N terminals, followed by long stretch of 'YG' repeats. On average, the biased residues of Tyr, Gly, Asp, Lys and His make up more than 65% of the mature protein (Table 2).

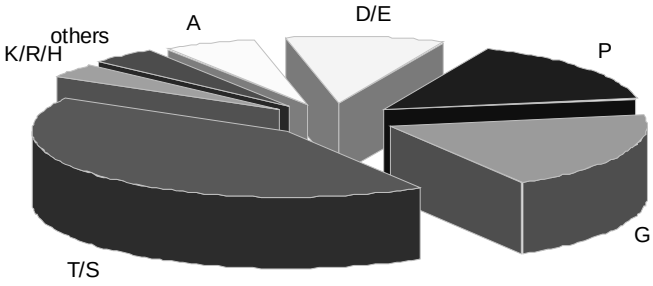


Fig. 4. Distribution of different groups of repeat containing proteins amongst 202 secreted proteins in *P. sojae* . The C-terminal 100 amino acid sequences were used for repeat analysis. Repetitive sequence is counted if a C-terminal domain consists of more than 20% repeats made up of 2-8 residues. The amino acid next to the graph indicates the major composition of the repeat. T/S indicates that the major composition of the repeat is T and/or S. The same holds for D/E and K/R/H. Similar results were obtained from the repeat containing proteins of *P. ramorum*.

Different protein families within the secretome show various degrees of divergence

A large proportion of the proteins is highly conserved between the two secretomes. In total, about 10% of the secreted proteins of *P. sojae* and *P. ramorum* possess homologues with very high sequence similarity (more than 90%). However, large numbers of unique proteins are also present in the secretome. The presence of both unique and highly conserved proteins demonstrates different degrees of sequence divergence of secreted proteins.

The degree of sequence divergence of the total set was visualized by dividing proteins into groups based on the sequence similarity of best BLAST hit in the other genome. The group with sequence similarity less than 35% are considered to be highly divergent and the group with similarity more than 75% is considered to be highly conserved. The group with similarity between 35% and 55% is considered as divergent, and between 55% and 75% as conserved (Fig. 5A). In both species, the majority of the secreted proteins (over 60%) fall in the conserved and highly conserved groups. A minority, 12% and 19% in *P. ramorum* and *P. sojae,* respectively, is highly divergent from the other genome.

Table 2. Three families of divergent secreted proteins.

family	name	protein size (aa)	signal peptide (aa)	C number	D-E%	P%	T-S%	Y%	repeat [a]	GPI [b]	transcripts present in the EST database [c]	protein code [h]
EHD-I [e]	PsEHD-I-1	184	28	0	15	20	27	1	DS(5) PPPKPTN(3) AL(3)			ps_137363
EHD-I	PsEHD-I-2	214	28	0	14	13	28	2	AT(10) PK(8) DD(8)		ZO(1)	ps_137397
EHD-I	PsEHD-I-3	214	28	0	14	13	28	2	AT(10) PK(8) DD(8)		ZO(1)	ps_137398
EHD-I	PsEHD-I-4	214	28	0	14	13	29	2	AT(10) DD(8) PK(7)		ZO(1)	ps_137399
EHD-I	PsEHD-I-5	221	18	0	17	15	29	1	TP(14) AA(5) DSSE(3)		ZO(4)	ps_109021
EHD-I	PsEHD-I-6	206	27	0	23	12	7	3	PA(5) ED(4) LV(3)		ZO(5)NS(1)	ps_109026
EHD-II	PrEHD-II-1	230	21	0	10	6	16	1		+		pr_74324
EHD-II	PrEHD-II-2	235	21	1	11	6	14	1		+		pr_74325
EHD-II	PrEHD-II-3	235	21	1	11	6	14	1		+		pr_74326
EHD-II	PrEHD-II-4	211	20	1	10	6	10	1		+		pr_74342
EHD-II	PrEHD-II-5	210	20	1	10	6	11	1		+		pr_74344
EHD-II [d]	PrEHD-II-6	217	..	2	7	6	18	1		+		pr_86438
EHD-III [f]	PsEHD-III-1	209	21	1	7	12	21	0				ps_144578
EHD-III	PsEHD-III-2	209	20	1	7	12	21	0				ps_132650
EHD-III	PsEHD-III-3	217	23	1	8	12	19	0				ps_135835
EHD-III	PsEHD-III-4	217	23	1	8	12	19	0				ps_135837
EHD-III	PsEHD-III-5	226	23	0	5	13	21	1				ps_135840
EHD-III [g]	PrEHD-III-1	284	26	0	5	13	29	1	TP(16) SK(7) QQ(4) LG(3) AA(3)			pr_86872
EHD-III	PrEHD-III-2	285	26	0	5	12	28	1	TP(16) SK(7) QQ(3) LG(3)			pr_78978
EHD-III	PrEHD-III-3	288	26	0	5	13	27	1	TP(16) SK(7) QQ(4) RA(3)			pr_78980
EHD-M96	M96(AAN37691)	260	23	1	13	4	19	14	YG(13) SS(7) DD(5) RR(4) HH(3)			-
EHD-M96	PrEHD-M96-1	253	20	0	19	4	10	19	YG(21) HD(7) SS(3)			pr_78103
EHD-M96 [d]	PrEHD-M96-2	235	-	0	22	3	12	20	YG(21) HDD(6) RK(4) ES(4) AT(3)			pr_78119
EHD-M96	PrEHD-M96-3	282	20	0	19	3	10	18	YG(24) HD(11) ES(4) AT(3)			pr_78120
EHD-M96	PrEHD-M96-4	190	26	1	18	4	9	21	YG(21) DD(5) EH(3)			pr_80687
EHD-M96	PrEHD-M96-5	206	26	1	19	3	8	20	YG(24) DD(5) EH(3)			pr_80690
EHD-M96	PrEHD-M96-6	189	26	1	19	4	10	20	YG(20) DD(6) EH(3)			pr_80700
EHD-M96	PrEHD-M96-7	275	20	0	20	4	10	18	YG(22) HD(11) ES(4) AT(3)			pr_85978
EHD-M96	PrEHD-M96-8	192	26	1	17	3	9	20	YG(20) DD(5) EH(3)			pr_86930
EHD-M96	PrEHD-M96-9	186	26	1	18	4	10	20	YG(20) DD(5) EH(3)			pr_87748
EHD-M96 [d]	PsEHD-M96-1	199	20	1	21	3	8	19	YG(13) DD(7) RH(4)			ps_126826
EHD-M96 [e]	PsEHD-M96-2	159	-	0	13	5	15	20	YG(13) SA(3)			ps_130869
EHD-M96	PsEHD-M96-3	325	20	0	16	3	9	21	YG(30) KD(13) HH(4)			ps_130870
EHD-M96	PsEHD-M96-4	299	23	0	17	3	11	19	YG(25) KD(13) SS(4) HH(3)			ps_130929
EHD-M96 [g]	PsEHD-M96-5	88	-	3	18	0	0	32	GKDYY(6)			ps_130934
EHD-M96 [f]	PsEHD-M96-6	308	-	0	19	5	10	19	YG(20) DD(9) AT(6) RH(4) SS(3)			ps_139067
EHD-M96	PsEHD-M96-7	305	20	0	19	5	11	18	YG(21) KD(10) AT(6) RH(4) SS(3)			ps_139068
EHD-M96	PsEHD-M96-8	276	20	0	19	3	11	20	YG(24) DD(8) KSK(4) RR(3)			ps_140746
EHD-M96	PsEHD-M96-9	276	20	0	19	3	11	21	YG(25) DKSK(4) RR(3)			ps_140748
EHD-M96	PsEHD-M96-10	275	20	0	21	3	12	18	YG(20) DKSK(7) RR(4)			ps_140749
EHD-M96	PsEHD-M96-11	296	20	0	18	5	11	18	YG(19) HD(7) AT(6) SS(3)			ps_144075
EHD-M96 [e]	PsEHD-M96-12	138	20	0	24	3	6	11	HD(7) YG(5)			ps_144076
EHD-M96	PsEHD-M96-13	299	23	1	17	3	11	18	YG(25) KD(13) SS(4) HH(4)			ps_145010

[a] Repeats were counted from the mature protein without signal peptide. Repeat times are listed in the brackets
[b] The presence of C-terminal GPI anchor is indicated by the + sign
[c] The tissue types that ESTs derived from are indicated by ZO(zoospore), and NS(non-stage specific).The number of transcript is listed in the brackets
[d] incomplete sequence
[e] pseudo-gene with a frame shift
[f] pseudo-gene lack of a start codon
[g] pseudo-gene with a stop codon in the coding region
[h] the protein codes are used in the *P. sojae* and *P. ramorum* annotation. ps stands for *P. sojae* and pr stands for *P. ramorum*.

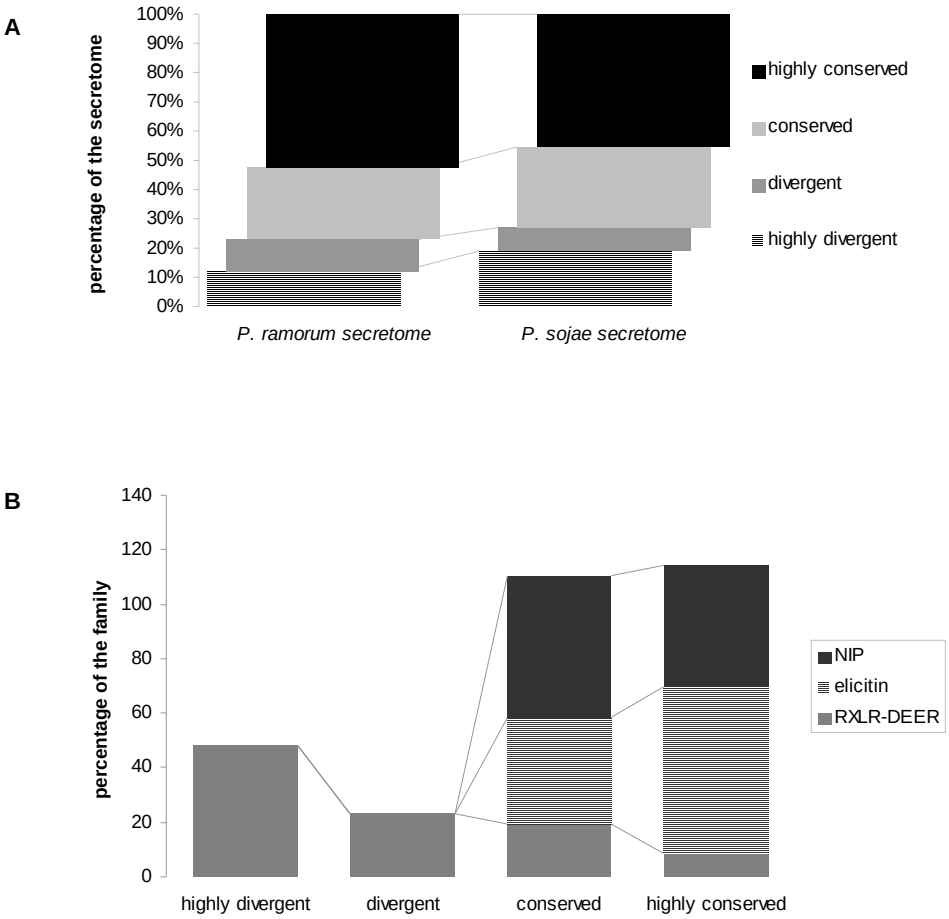


Fig. 5. A, Sequence divergence of secretomes between *P. sojae* and *P. ramorum.* The secretome was divided into 4 groups based on the sequence similarity of best BLAST hit in the other proteome. The highly conserved group has sequence similarity less than 35%. The divergent group has similarity between 35% and 55%, and the conserved group between 55% and 75%. The highly divergent group has sequence similarity larger than 75%. **B**, Sequence divergence of three families between *P. sojae* and *P. ramorum.* The *P. sojae* family members were used to BLAST the proteome of *P. ramorum*. Similar criteria as in **Fig. 5. A** were used to define highly divergent, divergent, conserved and highly conserved family members.

Two major groups of proteins are evolving rapidly in the secetome, namely the repeat containing proteins and the RXLR-DEER proteins. The repeat containing proteins exhibit high sequence divergence; 40% of the *P. sojae* and 30% of the *P. ramorum* repeat containing proteins can be classified as highly divergent. For example, the mating induced protein M96 family belongs to the group of highly divergent repeat containing proteins, therefore it was designated as EHD-M96 (Extracellular Highly Divergent protein - M96) family. The other group of fast evolving proteins bears the RXLR-DEER motif. A total of 103 and 97 RXLR-DEER proteins can be identified in the secretomes of *P. sojae* and *P. ramorum*, respectively (selection criteria are described in material and methods). 49% of the *P. sojae*

and 36% of the *P. ramorum* RXLR-DEER proteins can be classified as highly divergent, only a minority,10%, is highly conserved in both species.

Different families showed different paces of divergence. To illustrate the varied evolving tempo, three groups of secreted proteins, namely, elicitins, NIPs and RXLR-DEER proteins, in *P. sojae* were compared to one another. Elicitin and elicitin-like proteins (21 members, Table 1) were part of a large complex family (Jiang et al. 2006). Proteins sharing homology with the *P. sojae* NPP protein were designated as the NIP group (52 members) (Gijzen et al., unpublished data). A total of 103 proteins sharing the 'RXLR-DEER' motif was designated as RXLR-DEER group. These three groups show two types of distinct divergence patterns (Fig. 5B). All of the NIP and elicitin family members are conserved or highly conserved, whereas about 50% of RXLR-DEER members are highly divergent.

Three novel families of Extracellular Highly Divergent proteins (EHD) show lineage specific expansion patterns

Based on the rate of divergence, protein families can be classified as highly divergent or highly conserved. Highly divergent families are of particular interest because they may lead to the specific features related to plant-pathogen interaction. Three highly divergent protein families are described here to show three types of divergent patterns: one is unique to *P. sojae*, the other unique to *P. ramorum*, and the third is expanded in *P. sojae and P. ramorum* independently. In this study three novel families, EHD-I, EHD-II and EHD-III (Table 1), are described to demonstrate these three distinct divergent patterns.

The family EHD-I was composed of 6 members in *P. sojae* but is not present in the genome of *P. ramorum* or the large EST data set of *P. infestans*. EHD-1 family has a high percentage of Asp, Glu, Thr, Ser as well as Pro in the amino acid composition. Repeats based on biased amino acids can be identified in the sequence such as 'AT' and 'DD'. In the *P. sojae* EST database, EHD-1 transcripts are exclusively found in the zoospore stage, indicating that EHD-1 family is a class of extracellular protein with a zoospore specific expression pattern.

The family EHD-II was unique to *P. ramorum* as it is absent in the genome of *P. sojae* and the ESTs of *P. infestans*. EHD-II members possess hydrophobic regions at the extreme C-terminal. These hydrophobic regions are part of the glycosylphosphatidylinositol (GPI) anchor site predicted by the program big-PI plant predictor (Eisenhaber et al. 2003). The hydrophobic C-terminal end will be cleaved off from the mature protein, and a GPI is added that will anchor the EHD-II mature protein to the plasma membrane.

The family EHD-III has members in both *P. sojae* and *P. ramorum*, but as the phylogenic tree in Fig. 6 shows, the members in *P. sojae* form two distinct clades while the members in *P. ramorum* form another clade. The family EHD-III is also rich in Thr, Ser and Pro.

Members of the three divergent families are clustered in the genome (Fig. 7). All 6 members of the family EHD-I cluster in a 112 kb region whereas 5 members of EHD-II cluster in a 122 kb region. For family EHD-III, three members are clustered in *P. sojae* and two members form a small cluster in *P. ramorum*.

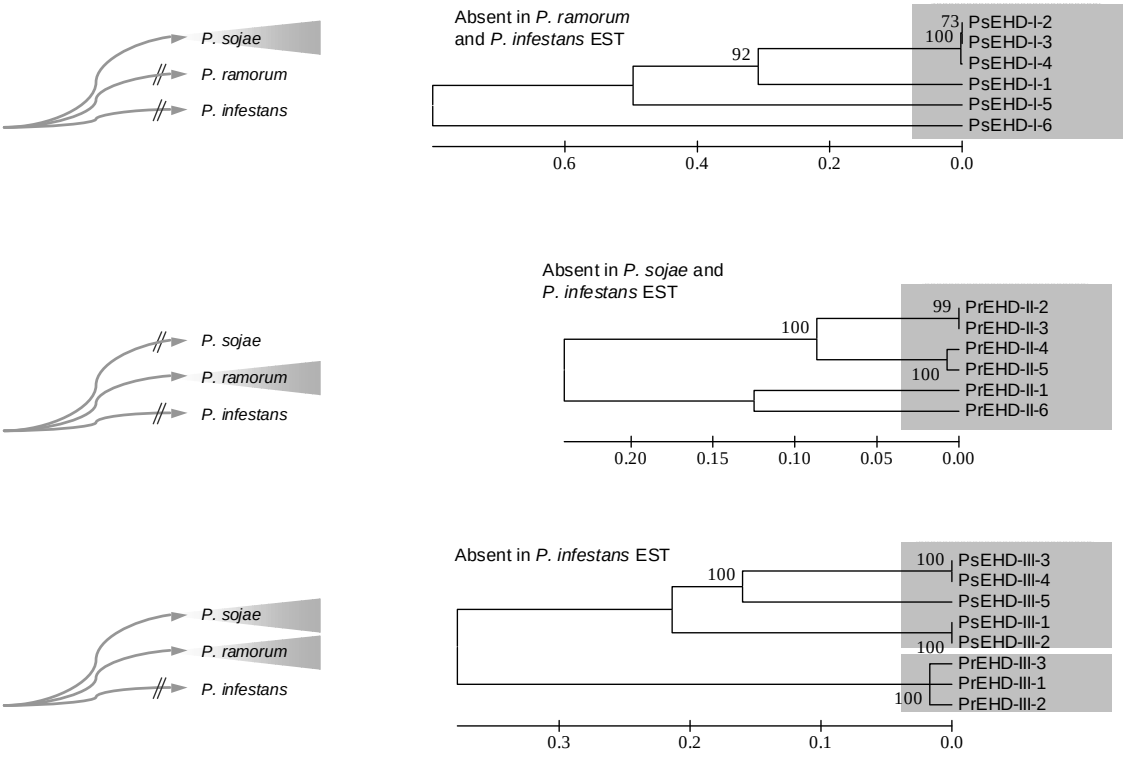


Fig. 6. Three families of highly divergent secreted proteins. PsEHD-I is unique to *P. sojae*, PrEHD-II unique to *P. ramorum*, and PsEHD-III is expanded in *P. sojae and P. ramorum* independently. Phylograms were constructed from EHD amino acid sequences from *P. ramorum* and *P. sojae.* The unrooted phylogram was based on Neighbor-Joining analysis. Confidence of groupings was estimated by using 1,000 bootstrap replicates; numbers next to the branching point indicate the percentage of replicates supporting each branch. The shaded blocks show the EHD familiy members belonging to the same clade.

Remarkably, the three families of highly divergent proteins (EHD I-III) share some structural similarities. None of them is rich in Cys residues which are typically found in fungal extracellular elicitors, all of them possess repeat like sequences and are particularly rich in Thr, Ser and Pro residues and quite often these residues are part of a repeat. The presence of O-GalNAc-glycosylation sites in EHD-I and EHD-III as predicted by the program NetOGlyc 3.1 (Julenius et al. 2005) suggests that these families are glycosylated. The three EHD families are most likely to be associated with the exterior of the cell either via GPI anchoring or glycosylation. They may be representative of a large group of repeat containing surface proteins that are evolving rapidly in *Phytophthora*.

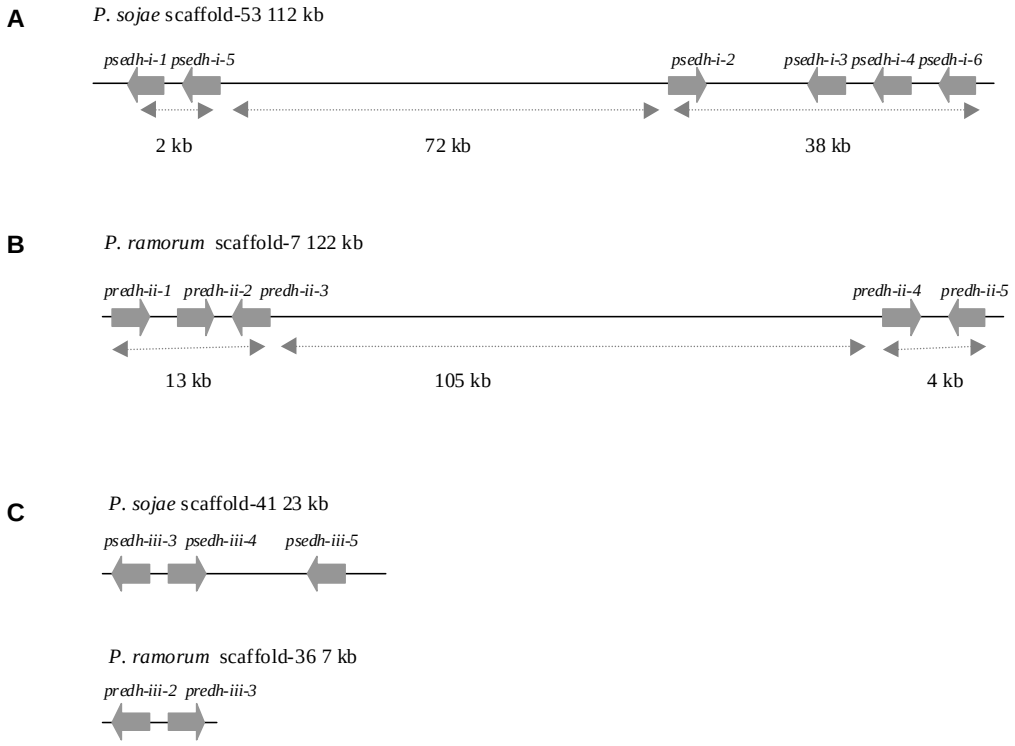


Fig. 7. Clustering of three families of highly divergent secreted protein encoding genes in the genome of *P. sojae* and *P. ramorum*. Grey block arrows indicate *EHD* genes and their orientations. Thin horizontal black lines represent DNA contigs. **A,** family EHD-I **B,** family EHD-II **C**, family EHD-III

The un-annotated small secreted proteins

In plant-microbe interactions, several small proteins secreted by the pathogen were shown to play (a)virulence related roles such as AVR9 (63 amino acids) of *Cladosporium fluvum* (van Kan et al. 1991) and the SCR74 (74 amino acids) of *P. infestans* (Liu et al. 2005). However, the homologue of SCR74 in *P. sojae* has to be manually annotated, which shows limitations of the current automatic annotation pipeline. The annotation of small ORFs is a challenge because of their small sizes and limited codon usage information. Despite of these technical difficulties, a set of putative secreted small proteins is still valuable for *Phytophthora* research to identify previously undocumented subsets of the secretomes.

A set of putative small secreted proteins (ca. 500) for each organism is proposed here in addition to the secretome predicted from the released versions of *P. sojae* and *P. ramorum* proteomes. These small proteins contain signal peptides predicted with high probabilities (mostly > 0.95). Curiously, there is no apparent overlap between two sets of small secreted proteins from *P. ramorum* and *P. sojae*. Only three proteins showed similarities to each other upon bi-directional BLAST searches. This highlights the differences between the two genomes and indicates the fast pace of evolution of this group of proteins.

The small secreted proteins may belong to the key differences between the secretomes *of P. sojae* and *P. ramorum*, and also likely have a functional role in plant-pathogen interaction. This can be illustrated by a 17-member protein family in *P. sojae*, PsGSpex-1 (*P. sojae* Genome derived Putative Small Extracellular protein). Genes encoding PsGSpex-1 on average show a high GC content of 64% that is indicative for *Phytophthora* gene coding capability (Jiang et al. 2006). The PsGSpex-1 proteins typically have 113 amino acids and are highly polymorphic (Fig. 8.). This family is unique to *P. sojae* because it is not present in any predicted protein datasets of *P. ramorum*. It is necessary to survey and sequence the genes encoding this family of secreted proteins in other *P. sojae* isolates to ascertain the presence of host selection pressure.

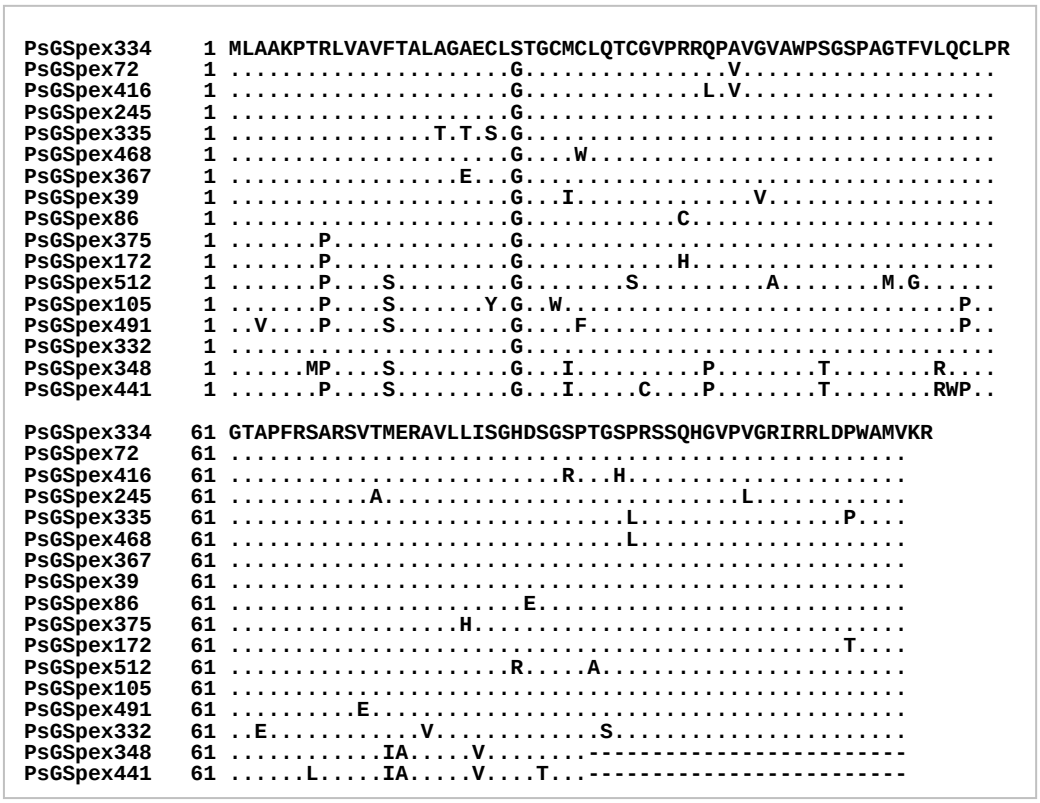


Fig. 8. Multiple sequence alignment of a family of putative small secreted proteins from *P. sojae* showing a remarkable polymorphism. The sequence names (Ps_gpex_XXX) and start positions (1 and 61) are on the left margin. Similar amino acids in the alignment are denoted with '.' and amino acid substitutions are shown in their universal one letter symbols. Deletions are marked with '-'.

Discussion

Availability of the complete genome sequences for two closely related yet distinctive plant pathogens enabled us to reveal prevailing evolutionary processes in the secretome of these important pathogens using bioinformatics techniques. We found that patterns of evolution differ between the genome and secretome of a single pathogen, between the *P. sojae* secretome and *P. ramorum* secretome as well as

between individual families within a secretome despite the fact that the percentages of the secreted protein genes in the transcriptomes are similar in *P. sojae and P. ramorum*.

Difference paces of evolution

The rice blast fungus *M. grisea* has a total of 11,109 genes in the predicted transcriptome, a number similar to that of the non-pathogenic *N. crassa* (10,082 genes). However, the secretome of *M. grisea* is twice the size of *N. crassa* (Dean et al. 2005). Our study compares two closely related pathogen genomes with the secrotomes taking up the same percentage of the coding genes. However, the fact that there is higher percentage of non-orthologous genes and unique genes in the secretome than the percentage in the genome indicates that the secretome has a faster pace of evolution.

P. sojae has a larger genome than *P. ramorum*, which is reflected by the larger sizes both in the total gene number and *spe* gene number. *P. sojae* has higher percentages of non-orthologous genes as well as unique genes compared to *P. ramorum*. The sizes of gene families are also in general larger in *P. sojae* than in *P. ramorum*. Thus the larger secretome of *P. sojae* maybe a result of both evolving 'new' genes and expanding 'old' genes. It is notable that *P. sojae* is a narrow-host pathogen, yet it possesses many larger families of secreted proteins than *P. ramorum*, a broad-host pathogen. It is likely that in *P. sojae* some of the gene family members serve more specialized and/or redundant functions.

Within a secretome, different types of proteins are evolving at different paces because of various selection forces. Genes involved in molecular 'arms races' are often undergoing rapid changes driven by diversifying selection. The 'arms races' are typically waged in host-pathogen and male-female interactions. The mammalian major histocompatibility complex (MHC) genes (Hughes and Nei 1988) and the plant resistance genes (Meyers et al. 2005) are examples of genes under diversifying selection in host-pathogen interactions. The sperm and egg recognition proteins (Lee et al. 1995; Metz et al. 1998; Rooney and Zhang 1999; McCartney and Lessios 2004) and the pollen coating proteins (Fiebig et al. 2004) are also rapidly evolving possibly due to the male-female interactions.

The large number of gene families in *Phytophthora* indicates the importance of duplication in shaping the secretome. Different protein families are diverging at a different pace in the two species (Fig. 9A). In some enzyme families, higher selection pressure may keep large part of the sequence conserved to maintain enzymatic functions. For example, between *P. sojae* and *P. ramorum*, highly conserved families include the NIP family and the cutinase family. On the other hand, highly divergent families might be under host selection pressure to manipulate host or evade resistance. These families are particularly interesting to unravel plant-pathogen interactions and will be discussed more in depth in the following sections. By and large, there are two distinct groups of proteins that are most highly diverged: the repeat containing proteins and the RXLR-DEER proteins (Fig. 9B).

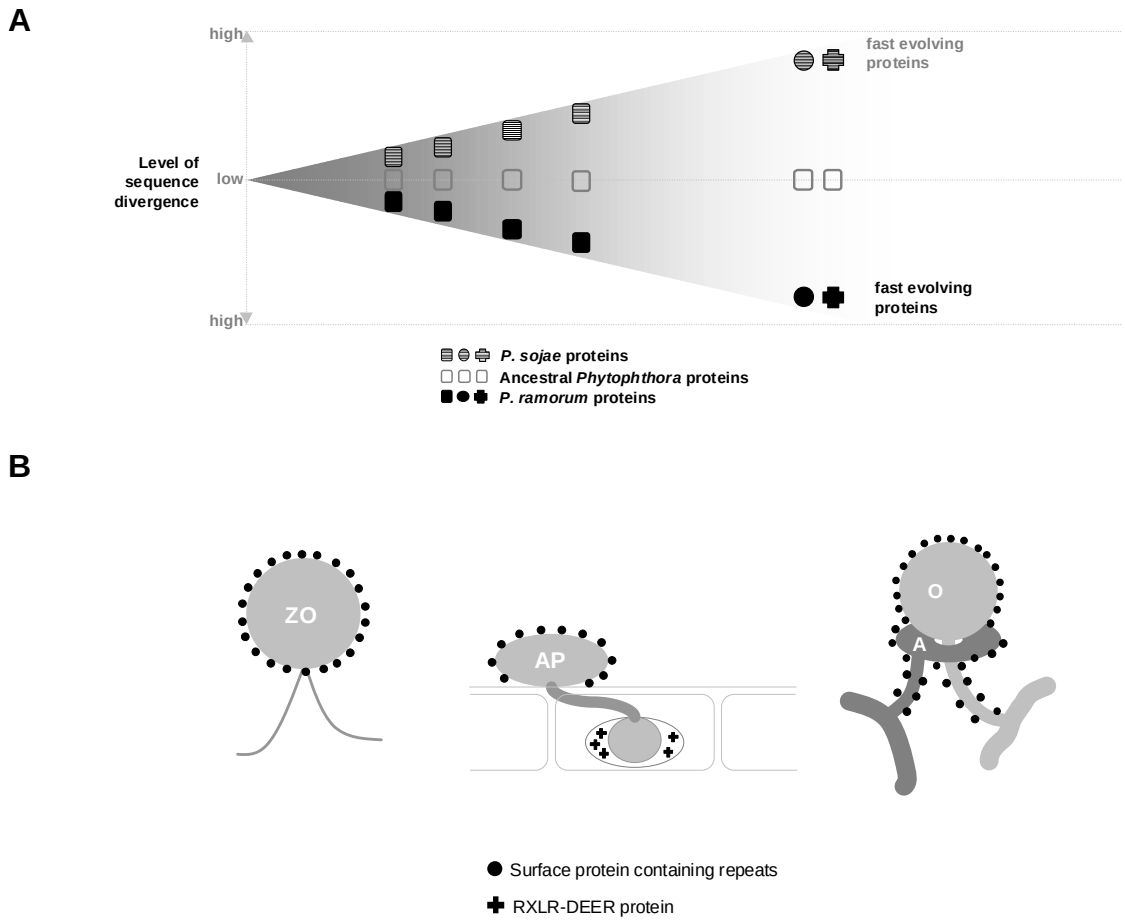


Fig. 9. A, Difference paces of evolution in the secretomes of *P. ramorum* and *P. sojae*. After the common ancestor gave rise to the two *Phytophthora* species, different proteins diverged at different paces. Various levels of conservation are found in the present day *Phytophthora* species. The triangle shading indicates the extent of sequence conservation, dark shading stands for conservation whereas light shading stands for divergence **B,** Surface proteins containing repeats and RXLR-DEER proteins are two major groups of highly divergent proteins. The letters stand for ZO (zoospore) AP (Appressorium) O (Oospore) and A (Antheridium). Surface proteins containing repeats coat various spore structures, and may mediate the interaction between gametes. RXLR-DEER proteins are primarily secreted into the space between membranes of haustoria and host cells.

Clustering of the gene family members

Duplication is the fuel for evolution to generate new genes. Clustering of homologous genes in the genome is mostly a reflection of the past duplication events, or in some cases, concerted evolution at work (Hurles 2004). In *Phytophthora*, many of the gene family members are clustered in the genome. Because one of the major mechanisms to enlarge a family is through unequal crossing-over events, the close physical distance between genes may reflect a relatively recent duplication. This may, for example, account for the expanded cutinase gene family in *P. sojae*.

However, the physical clustering can also be of functional importance rather than a reminiscence of a recent duplication. If members in a gene family perform the same physiological function, a member cannot evolve individually and the whole family will appear to evolve as a single unit. The most well-known example is the coordinated evolution of tandemly repeated ribosomal DNA (rDNA) (Schlotterer and Tautz 1994). Concerted evolution does not only operate in house-keeping genes, but also in rapid evolving genes that have been shown to be important in the interaction between organisms. In gamete recognition, concerted evolution may underlie the mechanism of species-specific fertilization. The abalone egg receptors for the sperm lysin is tandemly repeated and interdependently evolving (Swanson and Vacquier 1998). Concerted evolution may play a role in pathogenesis, the *Coccidioides* surface protein SOWgp is capable of immuno-modulating the host, and its repeated domain is evolving in concert (Johannesson et al. 2005).

The hallmark of concerted evolution is that paralogues are highly similar but orthologues diverge over time (Hurles 2004). The process of homogenization across the family is driven by gene conversions and/or birth-and-death selection (Nei and Rooney 2005). In *Phytophthora,* some of the tightly clustered genes may also evolve as a unit. For example, the EHD-III and EHD-M96 family are highly divergent between species but remain very similar within a genome. Whether they are under concerted evolution or not will be answered by analysis of accurately assembled tandamly repeated sequences. Further research may uncover the role of concerted rapidly evolving families in interaction with plant host or gametes.

Extracellular proteins with repeats are highly divergent

Repeated sequences occur in more than 10% of all known proteins (Marcotte et al. 1999). These proteins are typically non-globular and the internal repetition renders the protein with an enlarged binding area. A wide range of functions such as coating, adhesion and protein-protein interactions is carried out by these proteins (Andrade et al. 2001; Main et al. 2003).

For pathogens, internal repeats of genes may be of importance in adaptation. Higher mutation rate is found in simple sequence repeats (SSR) because of unequal crossing over events and polymerase inadequacy. Due to the flexibility needed to interact with the host, changing SSR in pathogenicity factors may bring adaptive advantages (Li et al. 2004). In pathogenic bacteria, SSRs within genes have been associated with genetic and phenotypic flexibility. Repeats occur in genes encoding surface components mediating adhesion and specific virulence factors such as lipopolysaccharide-modifying enzymes (van Belkum et al. 1998). In *Xanthomonas* species, the AvrBs3 family possesses an internal repeat domain that determines the outcome of an infection. The compatible or incompatible interaction with host plants is specified by the variable number of repeats in AvrBs3 in different pathotypes (Leach and White 1996).

In *P. infestans,* surface proteins containing tandem repeats have been previously described. The Car90 protein shares sequence similarity with human mucins, a group of repeat proteins that protect and lubricate the epithelial surfaces (Moniaux et al. 2001). The *Phytophthora* Car90 protein was proposed to serve a similar role in forming a mucous cover to protect the germlings (Gornhardt et al. 2000). Other groups of proteins containing repeat like sequences in the C terminal domains belong to the elicitin and elicitin like protein family (Jiang et al. 2006). Two of the elicitins, INF2A and INF2B, have been shown to be associated with the cell wall (V. G. A. A. Vleeshouwers et al., unpublished data). In the secretome of the two *Phytophthora* species, large numbers of repeat containing proteins are present, and many of them are found to be fast evolving as shown by the rapid divergence of EHD-I, EHD-III and EHD-M96 families. The cell surface of *Phytophthora* may be coated by some of the proteins; crucial protection, adhesion and interaction processes can be mediated by such exterior proteins. The mating associated EHD-M96 family members may mediate the specific recognition between gametes of the individual *Phytophthora* species (Fig. 9.)

RXLR-DEER superfamily constitutes large group of potential virulence factors

Pathogen produced avirulence proteins are directly or indirectly recognized by host resistance gene product with a high level of specificity (Flor 1942). In contrast to resistance genes, avrirulence genes differ tremendously in their sequence and function (Luderer and Joosten 2001; van't Slot and Knogge 2002). To date four oomycete *Avr* genes have been cloned (Allen et al. 2004; Shan et al. 2004; Armstrong et al. 2005; Rehmany et al. 2005).The cysteine residues often present in fungal AVRs are lacking in all four oomycete AVR's. Strikingly, a RXLR motif was present in all these AVRs suggesting a shared novel mechanism in host-pathogen interaction (Rehmany et al. 2005). A large number of RXLR-DEER proteins is found in the secretomes of *P. sojae* and *P. ramorum*. The RXLR-DEER superfamily consists of over 100 proteins and may include other candidate avirulence proteins.

The antagonism between host and pathogen leads to co-evolutionary arms races (Stahl and Bishop 2000). Both plant and pathogen possess molecules at this antagonistic interface. These proteins involving in the host-pathogen interaction often show fast rate of evolution. In plants, rapid evolution of defense genes was illustrated by the accelerated amino-acid substitution in, e.g., chitinases (Bishop et al. 2000) and resistance proteins (Caicedo and Schaal 2004; Rose et al. 2004). In pathogens, diversifying selection was found for pathogenicity and virulence related genes such as the fungal polygalacturonase genes (Stotz et al. 2000) and the oomycete avirulence gene *ATR13* (Allen et al. 2004). In *Phytophthora*, the RXLR-DEER superfamily is rapidly evolving which is indicative for their involvement in the host-interaction process. This large group of genes may have opened a brand new field to unravel the interaction between *Phytophthora* and plant hosts.

The RXLR-DEER superfamily is highly dynamic; not only the sequences appear to be rapidly evolving, also the genomic regions are undergoing frequent rearrangements. From the 10 largest families in the secretome of *P. ramorum*, two belong to the RXLR-DEER super-family and they are the only fast

evolving genes among the largest families. The genome distribution of RXLR-DEER genes is also different from the other large families. RXLR-DEER family members are mostly scattered in the genome and clustering of more than 3 genes is rare. Despite the fact that the two *Phytophthora* genomesoverall colinear synteny often appears to be interrupted by RXLR-DEER genes (Jiang et al., unpublished data).the genomic loci containing RXLR-DEER genes seem to have been through extensive rearrangements causing loss or gain of genes. We speculate that these features are most likely shaped by selection pressure exerted by plant hosts, and that RXLR-DEER genes may play an important role in virulence and host specificity.

Material and methods

Genome and EST databases

The genomic sequences and annotated protein sequences of *P. sojae* and *P. ramorum* were obtained from the web site of the Joint Genome Institute (JGI), the Department of Energy (http://www.jgi.doe.gov/genomes). The 75,757 EST data set (Randall et al. 2005) was used for *P. infestans* data mining.

Database mining of secreted proteins

The presence of signal peptides was analyzed by the computer program SignalPv2.0 (Nielsen et al. 1997; Nielsen et al. 1999) on all annotated genes to reveal the whole reservoir of secreted proteins of the two *Phytophthor*a species. The signal peptides were predicted by two criteria adapted and described by Torto et al (Torto et al. 2003): (1) SignalP-HMM prediction is positive with a score > 0.9 and (2) SignalP-NN predicts a cleavage site between 10 and 30 amino acids in length. Membrane proteins were further separated from the data set if transmembrane domains are predicted in the C-terminal by the program SOSUIv1.1(Hirokawa et al. 1998). Proteins targeted to mitochondria were removed based on the prediction made by the program TargetPv3.0 (Emanuelsson et al. 2000). Some repeated sequences could not be perfectly assembled in the draft sequence, and several nearly identical genes were found to be artifacts. To have a realistic estimation of the secretome, a gene was considered to be a duplicate and discarded if it showed BLASTN similarity larger or equal to 99% to another gene.

RXLR-DEER protein and C-terminal repeat protein mining

The RXLR-DEER motif was defined as (21aa signal peptide)-X(5-46)-[R|K]X[L|V|I|A][R|K|H]-(X 4-39 including at least 25% D or E) [G|T|E|D][E|D][R|K], X represents any amino acid residue. From the secreted protein data set, proteins bearing such motif were selected using pattern-matching to a regular expression in a Python script. The C-terminal 100 amino acid sequences were extracted from the

secreted protein data set for repeat analysis. A C-terminal domain with more than 20% repeat made of 2-8 residues is counted as a repeat positive sequence.

Bioinformatics tools

Sequences were analyzed in Vector NTI 8 package. For BLAST searches we used the NCBI BLAST program and the Standalone-BLAST Version 2.2.3 (Altschul et al. 1997). Multiple sequence alignment was performed using ClustalX 1.8 (Thompson et al. 1994) and for phylogenetic tree construction Molecular Evolutionary Genetic Analysis 2.1 (MEGA) (Kumar et al. 1994) was used. For GPI (Glycosylphosphatidylinositol) lipid anchor prediction, big-PI Plant Predictor (Eisenhaber et al. 2003) was used. Protein motifs were searched against the Prosite database (Bairoch 1991; Sigrist et al. 2002). Customized calculation scripts were written in Python 2.2 (http://www.python.org) and are available from the authors upon request.

Identification of un-annotated putative small secreted proteins

Small ORFs encoding for 61-120 amino acids were translated from both directions of the genomic sequences of *P. sojae* and *P. ramorum* using the 'getorf' program from EMBOSS (The European Molecular Biology Open Software Suite) (Rice et al. 2000). Amino acid sequences translated from these ORFs were subjected to SignalP v2.0 program using the criteria described above in the section 'Database mining of secreted proteins'. Signal peptide-containing sequences were collected for *P. sojae* and *P. ramorum* in two separate FASTA files (Pearson and Lipman 1988). Amino acid sequences that were predicted to be membrane proteins (determined using TMHMM2 (http://www.cbs.dtu.dk/services/TMHMM/), SOSUI v1.1, and big-PI v3 programs) or mitochondrial proteins (determined using TargetP v3.0 (http://www.cbs.dtu.dk/services/TargetP/) (Emanuelsson et al. 2000)) were removed from the files. The sequences were then searched against the released proteomes of *P. sojae* and *P. ramorum,* and those showing identical hits were removed. Duplicate sequences (>99% identity) were also removed after self-blastp searches for sequences in each file. Blastp searches and hmmpfam (HMMER software, http://hmmer.wustl.edu/) searches for the sequences were performed against the non-redundant (nr) NCBI protein database and the Pfam HMM library (Bateman et al. 2004), respectively.

Acknowledgements

This work was financially supported by a grant to FG from NWO-Aspasia grant (015.000.057). RHYJ received a travel grant from the Technology Foundation STW for attending the *Phytophthora* annotation jamboree.

Literature cited

Allen, R.L., Bittner-Eddy, P.D., Grenville-Briggs, L.J., Meitz, J.C., Rehmany, A.P., Rose, L.E. and Beynon, J.L. 2004. Host-parasite coevolutionary conflict between *Arabidopsis* and downy mildew. Science 306:1957-60.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J.H., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402.

Andrade, M.A., Perez-Iratxeta, C. and Ponting, C.P. 2001. Protein repeats: Structures, functions, and evolution. J Struc Biol 134:117-131.

Armstrong, M.R., Whisson, S.C., Pritchard, L., Bos, J.I., Venter, E., Avrova, A.O., Rehmany, A.P., Bohme, U., Brooks, K., Cherevach, I., Hamlin, N., White, B., Fraser, A., Lord, A., Quail, M.A., Churcher, C., Hall, N., Berriman, M., Huang, S., Kamoun, S., Beynon, J.L. and Birch, P.R. 2005. An ancestral oomycete locus contains late blight avirulence gene *Avr3a*, encoding a protein that is recognized in the host cytoplasm. Proc Natl Acad Sci USA 102:7766-71.

Bairoch, A. 1991. PROSITE: a dictionary of sites and patterns in proteins. Nucleic Acids Res 19 Suppl:2241-5.

Baldauf, S.L. 2003. The deep roots of eukaryotes. Science 300:1703-6.

Bateman, A., Coin, L., Durbin, R., Finn, R.D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E.L., Studholme, D.J., Yeats, C. and Eddy, S.R. 2004. The Pfam protein families database. Nucleic Acids Res 32:D138-41.

Bishop, J.G., Dean, A.M. and Mitchell-Olds, T. 2000. Rapid evolution in plant chitinases: Molecular targets of selection in plant-pathogen coevolution. Proc Natl Acad Sci USA 97:5322-5327.

Caicedo, A.L. and Schaal, B.A. 2004. Heterogeneous evolutionary processes affect *R* gene diversity in natural populations of *Solanum pimpinellifolium*. Proc Natl Acad Sci USA 101:17444-17449.

Dean, R.A., Talbot, N.J., Ebbole, D.J., Farman, M.L., Mitchell, T.K., Orbach, M.J., Thon, M., Kulkarni, R., Xu, J.R., Pan, H., Read, N.D., Lee, Y.H., Carbone, I., Brown, D., Oh, Y.Y., Donofrio, N., Jeong, J.S., Soanes, D.M., Djonovic, S., Kolomiets, E., Rehmeyer, C., Li, W., Harding, M., Kim, S., Lebrun, M.H., Bohnert, H., Coughlan, S., Butler, J., Calvo, S., Ma, L.J., Nicol, R., Purcell, S., Nusbaum, C., Galagan, J.E. and Birren, B.W. 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature 434:980-6.

Eisenhaber, B., Wildpaner, M., Schultz, C.J., Borner, G.H., Dupree, P. and Eisenhaber, F. 2003. Glycosylphosphatidylinositol lipid anchoring of plant proteins. Sensitive prediction from sequence- and genome-wide studies for *Arabidopsis* and rice. Plant Physiol 133:1691-701.

Emanuelsson, O., Nielsen, H., Brunak, S. and von Heijne, G. 2000. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300:1005-16.

Erwin, D.C. and Ribeiro, O.K. 1996. *Phytophthora* diseases worldwide. The American Phytopathological Society, St. Paul Minnesota USA

Fabritius, A.L., Cvitanich, C. and Judelson, H.S. 2002. Stage-specific gene expression during sexual development in *Phytophthora infestans*. Mol Microbiol 45:1057-1066.

Fiebig, A., Kimport, R. and Preuss, D. 2004. Comparisons of pollen coat genes across *Brassicaceae* species reveal rapid evolution by repeat expansion and diversification. Proc Natl Acad Sci USA 101:3286-91.

Flor, H.H. 1942. Inheritance of pathogenicity of *Melampsora lini*. Phytopathology 32:653-669.

Gornhardt, B., Rouhara, I. and Schmelzer, E. 2000. Cyst germination proteins of the potato pathogen *Phytophthora infestans* share homology with human mucins. Mol Plant Microbe Interact 13:32-42.

Gotesson, A., Marshall, J.S., Jones, D.A. and Hardham, A.R. 2002. Characterization and evolutionary analysis of a large polygalacturonase gene family in the oomycete plant pathogen *Phytophthora cinnamomi*. Mol Plant Microbe Interact 15:907-921.

Greenberg, J.T. and Vinatzer, B.A. 2003. Identifying type III effectors of plant pathogens and analyzing their interaction with plant cells. Curr Opi Microbiol 6:20-28.

Guttman, D.S., Vinatzer, B.A., Sarkar, S.F., Ranall, M.V., Kettler, G. and Greenberg, J.T. 2002. A Functional Screen for the Type III (Hrp) Secretome of the Plant Pathogen *Pseudomonas syringae*. Science 295:1722-1726.

Hiller, N.L., Bhattacharjee, S., van Ooij, C., Liolios, K., Harrison, T., Lopez-Estrano, C. and Haldar, K. 2004. A host-targeting signal in virulence proteins Reveals a Secretome in Malarial Infection. Science 306:1934-1937.

Hirokawa, T., Boon-Chieng, S. and Mitaku, S. 1998. SOSUI: classification and secondary structure prediction system for membrane proteins. Bioinformatics 14:378-9.

Hughes, A.L. and Nei, M. 1988. Pattern of nucleotide substitution at major histocompatibility complex class I loci reveals overdominant selection. Nature 335:167-70.

Hurles, M. 2004. Gene Duplication: The Genomic Trade in Spare Parts. PLoS Biology 2:e206.

Jentoft, N. 1990. Why are proteins O-glycosylated? Trends Biochem Sci 15:291-4.

Jiang, R.H., Dawe, A.L., Weide, R., van Staveren, M., Peters, S., Nuss, D.L. and Govers, F. 2005a. Elicitin genes in *Phytophthora infestans* are clustered and interspersed with various transposon-like elements. Mol Genet Genomics 273:20-32.

Jiang, R.H.Y., Tyler, B.M., Whisson, S.C., Hardham, A.R. and Govers, F. 2006. Ancient origin of elicitin gene clusters in *Phytophthora* genomes. Mol Biol Evol 23:338-351.

Johannesson, H., Townsend, J.P., Hung, C.-Y., Cole, G.T. and Taylor, J.W. 2005. Concerted evolution in the repeats of an immunomodulating cell surface protein, SOWgp, of the human pathogenic fungi *Coccidioides immitis* and *C. posadasii*. Genetics in press

Julenius, K., Molgaard, A., Gupta, R. and Brunak, S. 2005. Prediction, conservation analysis, and structural characterization of mammalian mucin-type O-glycosylation sites. Glycobiology 15:153-64.

Kumar, S., Tamura, K. and Nei, M. 1994. MEGA - Molecular evolutionary genetics analysis software for microcomputers. Comput Appl Biosci 10:189-191.

Latijnhouwers, M., de Wit, P.J. and Govers, F. 2003. Oomycetes and fungi: similar weaponry to attack plants. Trends Microbiol 11:462-9.

Leach, J.E. and White, F.F. 1996. Bacterial avirulence genes. Annu Rev Phytopathol 34:153-179.

Lee, V.T. and Schneewind, O. 2001. Protein secretion and the pathogenesis of bacterial infections. Genes Dev. 15:1725-1752.

Lee, Y., Ota, T. and Vacquier, V. 1995. Positive selection is a general phenomenon in the evolution of abalone sperm lysin. Mol Biol Evol 12:231-238.

Li, Y.-C., Korol, A.B., Fahima, T. and Nevo, E. 2004. Microsatellites within genes: structure, function, and evolution. Mol Biol Evol 21:991-1007.

Liu, Z., Bos, J.I., Armstrong, M., Whisson, S.C., da Cunha, L., Torto-Alalibo, T., Win, J., Avrova, A.O., Wright, F., Birch, P.R. and Kamoun, S. 2005. Patterns of diversifying selection in the phytotoxin-like *scr74* gene family of *Phytophthora infestans*. Mol Biol Evol 22:659-72.

Luderer, R. and Joosten, M.H.A.J. 2001. Avirulence proteins of plant pathogens: determinants of victory and defeat. Mol Plant Pathol 2:355-364.

Lynch, M. and Conery, J.S. 2000. The evolutionary fate and consequences of duplicate genes. Science 290:1151-1155.

Main, E.R., Jackson, S.E. and Regan, L. 2003. The folding and design of repeat proteins: reaching a consensus. Curr Opi Struc Biol 13:482-489.

Marcotte, E.M., Pellegrini, M., Yeates, T.O. and Eisenberg, D. 1999. A census of protein repeats. J Mol Biol 293:151-160.

Marti, M., Good, R.T., Rug, M., Knuepfer, E. and Cowman, A.F. 2004. Targeting malaria virulence and remodeling proteins to the host erythrocyte. Science 306:1930-1933.

McCartney, M.A. and Lessios, H.A. 2004. Adaptive evolution of sperm binding tracks egg incompatibility in neotropical sea urchins of the genus *Echinometra*. Mol Biol Evol 21:732-745.

Metz, E.C., Robles-Sikisaka, R. and Vacquier, V.D. 1998. Nonsynonymous substitution in abalone sperm fertilization genes exceeds substitution in introns and mitochondrial DNA. Proc Natl Acad Sci USA 95:10676-10681.

Meyers, B.C., Kaushik, S. and Nandety, R.S. 2005. Evolving disease resistance genes. Curr Opi Plant Biol 8:129-134.

Moniaux, N., Escande, F., Porchet, N., Aubert, J.P. and Batra, S.K. 2001. Structural organization and classification of the human mucin genes. Front Biosci 6:D1192-206.

Nei, M. and Rooney, A.P. 2005. Concerted and birth-and-death evolution of multigene families. Annu Rev Genet 39.

Nielsen, H., Brunak, S. and von Heijne, G. 1999. Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Pro Engi 12:3-9.

Nielsen, H., Engelbrecht, J., Brunak, S. and vonHeijne, G. 1997. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Pro Engi 10:1-6.

Pearson, W.R. and Lipman, D.J. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci USA 85:2444-8.

Pieterse, C.M.J., Van West, P., Verbakel, H.M., Brasse, P., Van den Bergvelthuis, G.C.M. and Govers, F. 1994. Structure and genomic organization of the *ipiB* and *ipiO* gene clusters of *Phytophthora infestans*. Gene 138:67-77.

Qutob, D., Kamoun, S. and Gijzen, M. 2002. Expression of a *Phytophthora sojae* necrosis-inducing protein occurs during transition from biotrophy to necrotrophy. Plant J 32:361-373.

Randall, T.A., Dwyer, R.A., Huitema, E., Beyer, K., Cvitanich, C., Kelkar, H., Fong, A.M., Gates, K., Roberts, S., Yatzkan, E., Gaffney, T., Law, M., Testa, A., Torto-Alalibo, T., Zhang, M., Zheng, L., Mueller, E., Windass, J., Binder, A., Birch, P.R., Gisi, U., Govers, F., Gow, N.A., Mauch, F., van West, P., Waugh, M.E., Yu, J., Boller, T., Kamoun, S., Lam, S.T. and Judelson, H.S. 2005. Large-scale gene discovery in the oomycete *Phytophthora infestans* reveals likely components of phytopathogenicity shared with true fungi. Mol Plant Microbe Interact 18:229-43.

Rehmany, A.P., Gordon, A., Rose, L.E., Allen, R.L., Armstrong, M.R., Whisson, S.C., Kamoun, S., Tyler, B.M., Birch, P.R. and Beynon, J.L. 2005. Differential recognition of highly divergent downy mildew avirulence gene alleles by RPP1 resistance genes from two Arabidopsis lines. Plant Cell 17:1839-50.

Rice, P., Longden, I. and Bleasby, A. 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16:276-7.

Rizzo, D.M., Garbelotto, M. and Hansen, E.M. 2005. *Phytophthora ramorum:* integrative research and management of an emerging pathogen in California and Oregon forests. Ann Rev Phytopathol 43:309-335.

Rooney, A. and Zhang, J. 1999. Rapid evolution of a primate sperm protein: relaxation of functional constraint or positive Darwinian selection? Mol Biol Evol 16:706-710.

Rose, L.E., Bittner-Eddy, P.D., Langley, C.H., Holub, E.B., Michelmore, R.W. and Beynon, J.L. 2004. The maintenance of extreme amino acid diversity at the disease resistance gene, *RPP13,* in *Arabidopsis thaliana*. Genetics 166:1517-1527.

Schlotterer, C. and Tautz, D. 1994. Chromosomal homogeneity of *Drosophila* ribosomal DNA arrays suggests intrachromosomal exchanges drive concerted evolution. Curr Biol 4:777-83.

Shan, W., Cao, M., Leung, D. and Tyler, B.M. 2004. The *Avr1b* locus of *Phytophthora sojae* encodes an elicitor and a regulator required for avirulence on soybean plants carrying resistance gene *Rps1b*. Mol Plant Microbe Interact 17:394-403.

Sigrist, C.J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. and Bucher, P. 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform 3:265-74.

Stahl, E.A. and Bishop, J.G. 2000. Plant-pathogen arms races at the molecular level. Curr Opi Plant Biol 3:299-304.

Stotz, H.U., Bishop, J.G., Bergmann, C.W., Koch, M., Albersheim, P., Darvill, A.G. and Labavitch, J.M. 2000. Identification of target amino acids that affect interactions of fungal polygalacturonases and their plant inhibitors. Physiol Mol Plant Pathol 56:117-130.

Swanson, W.J. and Vacquier, V.D. 1998. Concerted evolution in an egg receptor for a rapidly evolving abalone sperm protein. Science 281:710-2.

Thompson, J.D., Higgins, D.G. and Gibson, T.J. 1994. Clustal-W - improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position- specific gap penalties and weight matrix choice. Nucleic Acids Res 22:4673-4680.

Tjalsma, H., Bolhuis, A., Jongbloed, J.D.H., Bron, S. and van Dijl, J.M. 2000. Signal peptide-dependent protein transport in *Bacillus subtilis*: a genome-based survey of the secretome. Microbiol Mol Biol Rev 64:515-547.

Torto, T.A., Li, S., Styer, A., Huitema, E., Testa, A., Gow, N.A., van West, P. and Kamoun, S. 2003. EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen *Phytophthora*. Genome Res 13:1675-85.

Torto, T.A., Rauser, L. and Kamoun, S. 2002. The *pipg 1* gene of the oomycete *Phytophthora infestans* encodes a fungal-like endopolygalacturonase. Curr Genet 40:385-390.

Trost, M., Wehmhoner, D., Karst, U., Dieterich, G., Wehland, J. and Jansch, L. 2005. Comparative proteome analysis of secretory proteins from pathogenic and nonpathogenic *Listeria* species. Proteomics 5:1544-57.

van Belkum, A., Scherer, S., van Alphen, L. and Verbrugh, H. 1998. Short-sequence DNA repeats in prokaryotic genomes. Microbiol Mol Biol Rev 62:275-293.

van Kan, J.A.L., van den Ackerveken, G.J.M. and de Wit, P.J.G.M. 1991. Cloning and characterization of cDNA of avirulence gene *Avr9* of the fungal pathogen *Cladosporium fulvum*, causal agent of tomato leaf mold. Mol Plant Microbe Interact 4:52-59.

van't Slot, K.A.E. and Knogge, W. 2002. A dual role for microbial pathogen-derived effector proteins in plant disease and resistance. Crit Rev Plant Sci 21:229-271.

Wehmhoner, D., Haussler, S., Tummler, B., Jansch, L., Bredenbruch, F., Wehland, J. and Steinmetz, I. 2003. Inter- and intraclonal diversity of the *Pseudomonas aeruginosa* proteome manifests within the secretome. J Bacteriol 185:5807-5814.
Wilson, I.B., Gavel, Y. and von Heijne, G. 1991. Amino acid distributions around O-linked glycosylation sites. Biochem J 275 (2):529-34.

Author-recommended internet resources

DOE Joint Genome Institute (JGI) *P. sojae* genome database:
http://genome.jgi-psf.org/sojae1
DOE Joint Genome Institute (JGI) *P. ramorum* genome database:
http://genome.jgi-psf.org/ramorum1
The Virginia Bioinformatics Institute (VBI) Microbial Database hosts a range of microbial sequences including the genome sequence and annotation data of *P. sojae* and *P. ramorum*:
http://phytophthora.vbi.vt.edu

Chapter 8

High GC3 and retroelement codon mimicry in *Phytophthora*

submitted

Rays H.Y. Jiang and Francine Govers

High GC3 and retroelement codon mimicry in *Phytophthora*

Rays H.Y. Jiang and Francine Govers

Laboratory of Phytopathology, Plant Sciences Group, Wageningen University, Binnenhaven 5, NL-6709 PD Wageningen and Graduate School Experimental Plant Sciences, The Netherlands

For correspondence: E-mail Francine.Govers@wur.nl; Tel. +31 317 483138; Fax +31 317 483412

Keywords

GC3 – *Phytophthora* - codon bias - retrotransposon

Abstract

The genus *Phytophthora* is entirely comprised of destructive plant pathogens, and is phylogenetically distinct from plants, animals or fungi. *Phytophthora* genes show a strong preference for usage of codons ending with G or C (high GC3). The presence of high GC3 in genes can be utilized to differentiate coding regions from non-coding regions in the genome. We found that both selective pressure and mutation bias drive codon bias in *Phytophthora.* Indicative for selection pressure is the higher GC3 value of highly expressed genes in different *Phytophthora* species. Lineage specific GC increase of non-coding regions is reminiscent of whole genome mutation bias. Heterogeneous retrotransposons exist in *Phytophthora* genomes and many of them vary in their GC content. Interestingly, the most widespread groups of retroelements in *Phytophthora* show high GC3 and a codon bias that is similar to host genes. Apparently, selection pressure has been exerted on the retroelement's codon usage, and such mimicry of host codon bias might be beneficial for the propagation of retrotransposons.

Introduction

In nearly all organisms, the 20 amino acids are specified in universal sets of nucleotide triplets named codons. The redundancy of codons enables species to systematically use certain synonymous codons. Many organisms have been observed to have codon bias. Two major mechanisms are considered responsible for such biases in codon usage: selection pressure and mutation bias (Sharp et al. 1993).

Mutational bias is a global force acting on all sequences. The base composition is constrained by genome-wide mutational processes. In some organisms the whole genome has shifted to extreme GC or AT content by mutational bias. GC content can reach up to more than 60% in the green algae *Chlamydomonas reinhardtii* based on cesium chloride estimate (Scala et al. 2002) and genome sequences analysis (JGI, *C. reinhardtii* sequencing project). The highest AT content was found in the malaria parasite *Plasmodium falciparum* with around 80% AT (Bowman et al. 1999). A shift of the whole genome to an extreme AT content is also observed in the free living protist *Dictyostelium discoideum* (Eichinger et al. 2005) and the bacterium *Borrelia burgdorferi* (Fraser et al. 1997) that causes Lyme disease. These genomes with extreme nucleotide composition have evolved independently because the organisms belong to entirely different phylogenetic groups. Codon usage can also be driven by global mutational bias which creates different trends between species. For example, in a number of species, the overall GC content of the genome shows correlation with species-specific codon bias (Chen et al. 2004).

The other major mechanism is selective pressure that works only on coding sequences via selection exerted on translation processes. Selection on synonymous codon positions could lead to a co-adaptation of codon usage and tRNA content. Efficient protein expression can be established via this mechanism. For genes that are expressed at high levels, such selective pressure to optimize translation is expected to be stronger. In such cases, codon bias is expected to correlate with the quantity of tRNA and expression levels. The abundance of tRNAs has been shown to correspond to preference of codons in organisms such as *Escherichia coli* and *Saccharomyces cerevisiae* (Kanaya et al. 1999). A clear correlation between codon usage and gene expression levels was found in *Caenorhabditis eleg*ans, *Drosophila melanogaster* and *Arabidopsis thaliana* (Duret and Mouchiroud 1999).

Phytophthora is a genus comprised of over 65 phytopathogenic species (Erwin and Ribeiro 1996), which cause severe damage in agriculture, forestry and natural habitats. *Phytophthora infestans* (causing potato late blight) and *Phytophthora sojae* (causing soybean root rot) are economically important crop pathogens. *Phytophthora ramorum* is a recently discovered species destroying woody shrubs and trees including oaks along the west coast of US (Rizzo et al. 2004) and in Europe. *Phytophthora* species morphologically resemble plant pathogenic fungi and convergent evolution has shaped these two major groups of pathogens with similar weaponry to attack plants (Latijnhouwers et al. 2003). *Phytophthora* is placed in the kingdom *Stramenopila* that has evolved distant from plants, animals and fungi (Baldauf

2003; Baldauf et al. 2000; Margulis and Schwarts 2000). The majority of the *Phytophthora* species form a recently evolved monophyletic group (Cooke et al. 2000) which suggests that the basic genome features of *Phytophthora* species would share similarities. *P. sojae* and *P. infestans* have an estimated GC content of around 50% overall in the genome, and around 60% in coding regions (Jiang et al. 2005). In *P. infestans,* high GC3 (GC content at the 3rd codon position) has been reported to correlate with a high GC content and was found to be related to codon bias (Jiang et al. 2005; Randall et al. 2005). An elevated GC content is not found in the phylogenetically related marine diatom *Thalassiosira pseudonana* which has only 48% GC in coding regions (Armbrust et al. 2004).

Interestingly, some mobile elements in *P. infestans* also show high GC content (Jiang et al. 2005). Mobile elements constitute the most abundant genetic material in higher eukaryotes. The relationship between transposons and hosts may be a continuum from extreme parasitism to mutualism (Kidwell and Lisch 2001). On the one hand, mobile elements are parasitic. They enrich their abundance by using the cellular apparatus and at the cost of host energy. On the other hand, mobile elements play a central role in the structure, function and evolution of eukaryotic genomes (Bennetzen 2000; Kazazian 2004) such as providing cis-regulatory sequences, assembling a kinetochore and speeding up protein diversification (Nekrutenko and Li 2001; Topp et al. 2004). To enlarge genome sizes, retroelements are particularly effective because they transpose via a mRNA intermediate synthesized by a reverse transcriptase. The amplification of the mRNA intermediate may rapidly increase retroelement copy number as compared to the 'cut and paste' mode of the DNA transposons. The large genome of *P. infestans* was proposed to be primarily due to insertions of retroelements (Jiang et al. 2005). To understand the relationship between mobile elements and host genomes at a genetic level, investigation of their codon choices can be informative.

Lerat et al. (2002) reported that a lowered GC content is a host-independent characteristic common to all mobile elements. In *C. eleg*ans, *A. thaliana*, *D. melanogaster, S. cerevisiae* and *Homo sapiens,* mobile elements exhibit overall AT-richness despite the fact that the host genomes vary in GC content. It was found that mobile elements exhibit codon usage bias similar to weakly expressed host genes in AT rich genomes like *C. eleg*ans, but show no similarity in GC rich genomes such as *D. melanogaster* (Lerat et al. 2002). However, in *P. infestans*, the sharing of codon bias appears to be related to copy number of the retrotransposons. Two Ty3/Gypsy retrotransposons occur frequently in the genome and share similar codon bias as host genes, whereas other retrotransposons with lower copy number do not share such codon bias (Jiang et al. 2005). This raises the question whether high GC3 is a general feature for the most widespread *Phytophthora* retrotransposons.

With the availability of the whole genome sequences of *P. sojae* and *P. ramorum*, it is feasible to analyze GC3 and codon usage in detail on a genome wide scale. By exploring both EST and genome databases, it is also possible to correlate expression levels with codon bias. The aim of this study is to search for evidence for whole genome mutational bias and/or selection pressure in *Phytophthora*. For that purpose we (1) analyzed the relative synonymous codon usage (RSCU) and GC3 in the two sequenced

Phytophthora genomes, (2) investigated whether the nucleotide composition differs between coding regions and intergenic regions, and (3) analyzed the relationship of codon bias between high copy retrotransposons and that of host genes.

Material and Methods

Genome databases and EST databases

The *P. infestans* EST databases are accessible at http://www.pfgd.org and http://staff.vbi.vt.edu/estap and most *Phytophthora* EST sequences are available through GenBank (Kamoun et al. 1999; Qutob et al. 2000; Randall et al. 2005) and http://phytophthora.vbi.vt.edu/EST. *Blumeria graminis* and *P. sojae* EST databases were downloaded from Phytopathogenic Fungi and Oomycete EST Database Version 1.4 (Soanes et al. 2002) http://cogeme.ex.ac.uk. The genomic sequences and annotated protein sequences of *P. sojae* and *P. ramorum* were obtained from the website of the DOE Joint Genome Institute http://www.jgi.doe.gov/. The annotated genes from *Aspergillus nidulans, Fusarium graminearum, Magnaporthe grisea, Neurospora crassa, Stagonospora nodorum* and *Ustilago maydis* were downloaded from the Broad institute website http://www.broad.mit.edu/annotation.

Bioinformatics tools

Sequences were analyzed in Vector NTI 8 package. For BLAST searches we used the NCBI BLAST program and the Standalone-BLAST Version 2.2.3 (Altschul et al. 1997). Multiple sequence alignment was performed by ClustalX 1.8 and for phylogenetic tree construction Molecular Evolutionary Genetic Analysis 2.1 (MEGA) (Kumar et al. 2001) was used. Phylogeny reconstruction of reverse transcriptase domains of retrotransposons was performed by Neighbor-Joining analysis. Poisson Correction (PC) was chosen as the distance parameter as specified in the program MEGA. The inferred phylogeny was tested by 1,000 bootstrap replicates. The calculation scripts were written in Python 2.2 (http://www.python.org) and are available from the authors upon request.

GC content and codon analysis

GC content of the entire sequence, first, second, and third codon positions (GCaverage, GC1, GC2 and GC3, respectively) were calcuated for each gene. For the GC frame plot, GC content was calculated with a 300 bp sliding window for all six reading frames similar to the methods used by the program FramePlot (Ishikawa and Hotta 1999). Relative Synonymous Codon Usage (RSCU) values were calculated as performed by Sharp et al (1986). For the correlation of codon usage between different gene sets, Trp, Met and three stop codons were excluded from analysis. For the non-coding region GC analysis, sequences between -100 to -200 bp and between -1 to -100 upstream of the start codons were

used. Neighboring coding sequences were excluded from the data set. From each genome, a set of 10,000 sequences was randomly selected to test the differences.

Statistical analysis of the differences between data sets

Statistical analysis was performed with the package SPSS 12.0.1 according to the program instructions. In order to make the data set a better approximation of Normal distribution, a GC percentage was converted into an arcsin value before performing the F-test and T-test. F-test was conducted to determine whether the two samples have different variances, and the one-tailed probability of the variances are not significantly different was calculated. If two samples show equal variances, T-test was performed to determine whether the two datasets have the same mean and the probability that two samples come from data sets with the same mean was calculated. Alternatively, if two samples show unequal variances, Kolmogorov-Smirnov Z test and Mann-Whitney U test were used to determine whether the variable in each of two independent samples comes from the same underlying population. Wald-Wolfowitz Median test was conducted to determine whether there is a difference in median values between the two samples.

Results

High GC3 causes high GC content in *Phytophthora* genes

Two sets of 10,000 ORFs were taken from the whole genome sequences of *P. sojae* and *P. ramorum*, respectively. High GC content was found in the ORFs as previously reported (Jiang et al. 2005; Randall et al. 2005). The average GC percentage is 60.0% for *P. sojae* and 58.6% for *P. ramorum*. In particular, high GC3 was found: GC3 is higher than GC2 or GC1 in more than 90% of the ORFs in both *Phytophthora* species (Table 1). *P. sojae* and *P. ramorum* genes show a GC3 of 76.0% and 73.1%, respectively. For comparison, GC3 was also calculated for several ascomycete and basidiomycete fungi. In *Magnaporthe grisea* and *Neurospora crassa,* genes also have high GC3 with an average value above 65%, and in more than 80% of the genes in both species GC3 is higher than GC1 or GC2. In contrast, in *Blumeria graminis* and *Fusarium graminearum*, less than 50% of the genes have GC3 higher than GC1 or GC2 (Table 1).

To investigate whether high GC3 is a specific feature for coding regions or not, a set of 10,000 randomly selected genomic sequences of 1 kb in size were retrieved from *P. sojae*. These genomic sequences and the 10,000 ORFs were used for GC analysis and the distribution of GC content was plotted. ORFs have a GC content peak around 60% whereas genome sequences show a peak around 50%. These two distinct peaks indicate that coding regions have a higher GC content than average random genomic regions (Fig. 1A). The GC3 of the majority of ORFs is even higher with a peak around 70% (Fig. 1A).

Table 1. GC content and GC3 in *Phytophthora* and several fungal species. Data sets derived from ESTs are shaded. Data derived from *Phytophthora* species are in bold.

		Total ORFs	% ORFs with GC3max [a]	% ORFs with GC > 60 [b]	GC (%) [c]	GC3 (%) [d]
EST data set	*Blumeria graminis* EST derived ORFs	283	11.7	17.0	45.4	43.1
	***P. infestans* EST derived ORFs**	**1000**	**81.1**	**98.6**	**57.8**	**70.7**
	***P. sojae* EST derived ORFs**	**1000**	**98.6**	**99.9**	**62.3**	**83.5**
Genomic data set	***P. sojae* ORFs**	**10000**	**93.7**	**99.6**	**60.0**	**76.0**
	***P. ramorum* ORFs**	**10000**	**90.3**	**99.0**	**58.6**	**73.1**
	Aspergillus nidulans ORFs	1000	53.4	69.6	53.5	58.7
	Fusarium graminearum ORFs	1000	41.9	55.4	51.8	55.6
	Magnaporthe grisea ORFs	1000	82.2	91.0	57.9	68.5
	Neurospora crassa ORFs	1000	82.2	92.1	56.2	65.7
	Stagonospora nodorum ORFs	1000	66.9	81.8	54.8	61.6
	Ustilago maydis ORFs	1000	62.6	86.2	55.6	61.5

[a] percentage of genes having GC3 higher than GC1 or GC2
[b] percentage of genes having GC1, GC2 or GC3 higher than 60%
[c] the average GC content
[d] the average GC3

The higher percentage of GC content in the coding region is caused by high GC3. When GC1, GC2 and GC3 of the 10,000 ORFs were plotted, GC1 and GC2 form two distinct peaks with lower GC content than the peak of GC3 (Fig. 1B). GC1 has a peak around 40% and GC2 has a peak around 60%. When the same analysis was performed with *P. ramorum* genome sequences and ORFs, similar results were obtained (data not shown). We can conclude that on a whole genome scale, *P. sojae* and *P. ramorum* genes on average show a higher GC content than non-coding regions, and this increase in GC content is mainly due to a high GC3 value.

High GC3 and codon bias

The high GC3 of *Phytophthora* genes is a result of the codon bias as previously reported from a small set of *P. infestans* genes (Jiang et al. 2005). To investigate the relationship between GC3 and codon bias in the whole genome, Relative Synonymous Codon Usage (RSCU) values were calculated for codons from two sets of 10,000 genes derived from *P. sojae* and *P. ramorum.* RSCU values are defined as the number of times that a particular codon is observed, relative to the number of times that the codon would be observed in the absence of any codon usage bias (Sharp and Li 1986). In the absence of any codon usage bias, the RSCU value would be 1. A codon that is used less frequently than expected will have a RSCU value lower than 1, and vice versa, higher than 1 for a codon that is used more frequently than expected. In both genomes, with the exception of TGA (stop codon), all codons

with the 3rd position A or T have a RSCU value lower than 1. Except for GGG (Gly), TTG (Leu) and AGG (Arg), all codons ending with C or G have a RSCU value higher than 1 (Table 2).

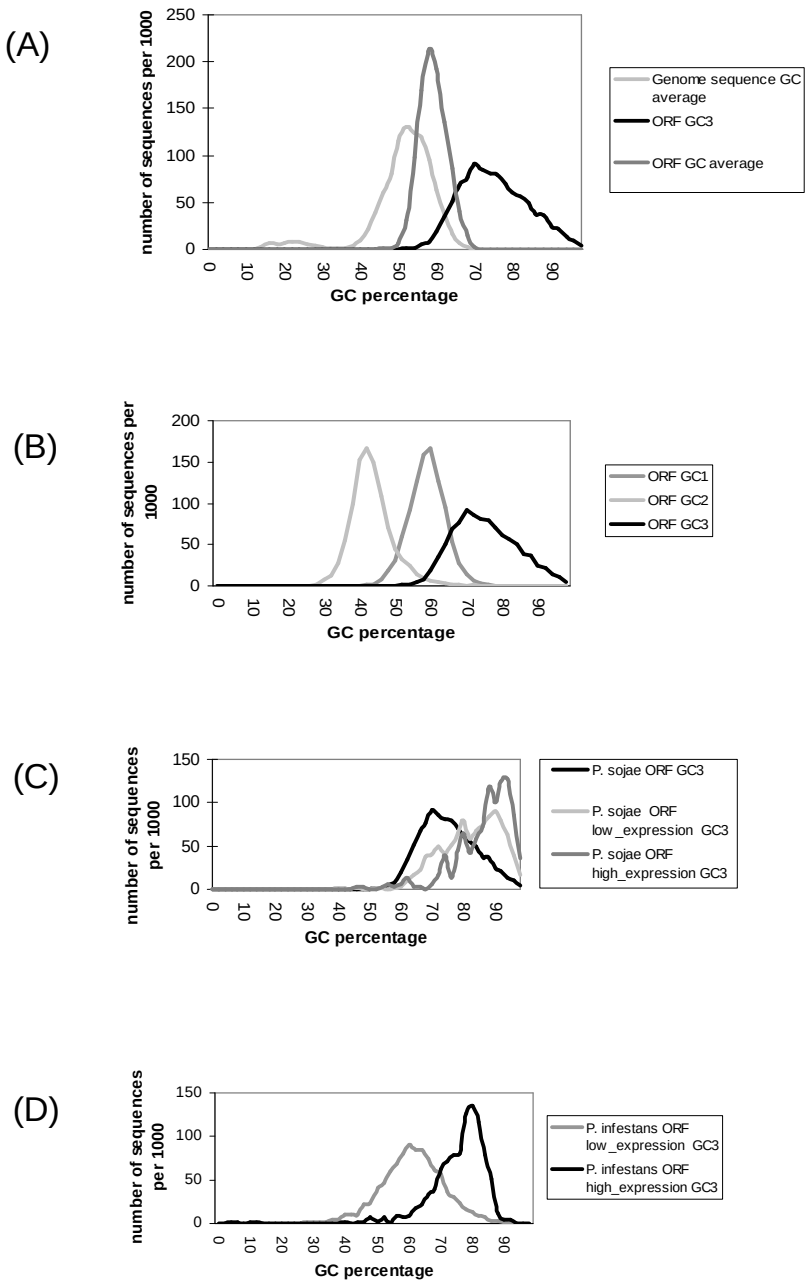


Fig. 1. (A) GC plot of coding regions vs. randomly selected genomic regions in *P. sojae*. Two sets of 10,000 sequences were used for the analysis. (B) GC plot of GC1, GC2 and GC3 of 10,000 *P. sojae* ORFs. (C) GC plot of annotated genes vs. weakly/highly expressed genes in *P. sojae.* 936 more weakly expressed genes and 305 more highly expressed genes were used for the analysis. (D) GC plot of lower vs. higher expressed genes in *P. infestans.* 3000 weakly expressed genes and 700 highly expressed genes were used for the analysis.

The percentage of codons ending with A or T and that of codons ending with C or G is plotted to show the cause of high GC3 of genes (Fig. 2). Except for the stop codon, all the other redundant amino acid codons have a preference for the degenerate 3rd position ending with C or G. A slightly lower GC% was found in *P. ramorum* codons. The preference of high GC3 codons in *P. sojae* and *P. ramorum* agrees with the results from *P. infestans* (Hraber and Weller 2001; Jiang et al. 2005; Randall et al. 2005). Therefore it can be concluded that the high GC3 of *Phytophthora* genes is due to biased usage of codons ending with C or G.

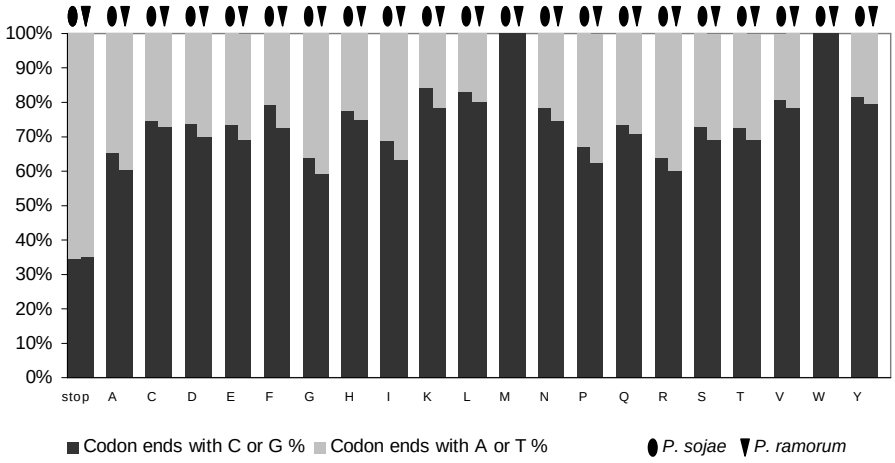


Fig. 2. The percentage of codons ending with C/G and A/T in *P. sojae* and *P. ramorum*. On the X-axis, single letter amino acid codes are used.

Whole genome mutation bias indicated by higher GC content in non-coding regions in *P. sojae*

The two *Phytophthora* species show a different degree of codon bias. On average *P. sojae* genes show higher GC3 than *P. ramorum*. Furthermore, *P. sojae* also shows a more biased RSCU value (Table 2). Nearly for each codon with a RSCU value > 1, *P. so*jae has a larger value than *P. ramorum*, and for codons with a RSCU value < 1, *P. sojae* has a smaller value than *P. ramorum*. Moreover, a higher percentage of codons ending with G or G are used in *P. sojae* as compared to *P. ramorum* (Fig. 2). These observations suggest a higher GC bias in *P. sojae* than in *P. ramorum*.

Whole genome mutation bias was found to lead to species-specific codon biases (Chen et al. 2004) and it may have caused the differences between *P. sojae* and *P. ramorum*. To detect the mutation bias, 500 bp non-coding sequences representing intergenic regions were used for GC3 analysis. Because the characteristic transcription initiation site of oomycetes typically occurs within 100 bp upstream of the start codon (McLeod et al. 2004), non-coding regions (5'UTR and partial promoter) were extracted from upstream promoter sequences ranging from – 1 bp to -100 bp in front of 10,000 ORFs. Intergenic regions were extracted from –100 to –200 bp in front of the same 10,000 ORFs. The average GC of *P. sojae* non-coding regions is 54.1% and *P. ramorum* non-coding regions give an average GC of 53.0%.

Table 2. Relative Synonymous Codon Usage (RSCU) and 3rd position GC frequency in 10,000 *P. sojae* genes and 10,000 *P. ramorum* genes. A total of 10,818,654 codons was counted *in P. sojae* and 10,040,752 in *P. ramorum*.

Amino Acid [a]	**RSCU** [b]					
stop	TAA	TAG	TGA			
ps	**0.9**	**1.0**	**1.1**			
pr	**1.0**	**1.1**	**1.0**			
A	GCA	GCC	GCG	GCT		
ps	**0.5**	1.2	**1.4**	**0.8**		
pr	**0.7**	1.2	**1.3**	**0.9**		
C	TGC	TGT				
ps	1.5	0.5				
pr	1.5	0.5				
D	GAC	GAT				
ps	**1.5**	**0.5**				
pr	**1.4**	**0.6**				
E	GAA	GAG				
ps	**0.5**	**1.5**				
pr	**0.6**	**1.4**				
F	TTC	TTT				
ps	**1.6**	**0.4**				
pr	**1.5**	**0.5**				
G	GGA	GGC	GGG	GGT		
ps	0.8	**1.9**	**0.7**	**0.7**		
pr	0.8	**1.8**	**0.6**	**0.8**		
H	CAC	CAT				
ps	**1.6**	**0.4**				
pr	**1.5**	**0.5**				
I	ATA	ATC	ATT			
ps	0.1	**2.1**	**0.8**			
pr	0.1	**1.9**	**1.0**			
K	AAA	AAG				
ps	**0.3**	**1.7**				
pr	**0.4**	**1.6**				
L	CTA	CTC	CTG	CTT	TTA	TTG
ps	**0.3**	**1.6**	**2.5**	**0.6**	**0.1**	**0.9**
pr	**0.4**	**1.5**	**2.4**	**0.7**	**0.2**	**1.0**
M	ATG					
ps	1.0					
pr	1.0					
N	AAC	AAT				
ps	**1.6**	**0.4**				
pr	**1.5**	**0.5**				
P	CCA	CCC	CCG	CCT		
ps	**0.6**	**1.0**	1.6	**0.7**		
pr	**0.7**	**0.9**	1.6	**0.8**		
Q	CAA	CAG				
ps	**0.5**	**1.5**				
pr	**0.6**	**1.4**				
R	AGA	AGG	CGA	CGC	CGG	CGT
ps	0.4	**0.5**	**0.9**	**2.4**	**1.0**	**0.9**
pr	0.4	**0.4**	**1.0**	**2.3**	**0.9**	**1.1**
S	AGC	AGT	TCA	TCC	TCG	TCT
ps	**1.6**	**0.6**	0.5	1.0	**1.8**	0.6
pr	**1.5**	**0.7**	0.5	1.0	**1.7**	0.6
T	ACA	ACC	ACG	ACT		
ps	**0.5**	**1.1**	1.8	0.6		
pr	**0.6**	**1.0**	1.8	0.6		
V	GTA	GTC	GTG	GTT		
ps	**0.2**	**1.2**	2.1	**0.5**		
pr	**0.3**	**1.0**	2.1	**0.6**		
W	TGG					
ps	1.0					
pr	1.0					
Y	TAC	TAT				
ps	1.6	0.4				
pr	1.6	0.4				

[a] ps (*P. sojae*) pr (*P. ramorum*)
[b] RSCU values differing in *P. sojae* and *P. ramorum* are in bold.

The GC plot shows two overlapping peaks, the peak of *P. sojae* has a slight shift towards high GC content when compared to that of *P. ramorum* (data not shown). Because only the data set of –1bp to –100 bp shows equal variance ($p > 0.001$), T test was conducted and the result showed that the difference of their means is significant (Table 3). Three additional tests which do not presume equal variances of data set were conducted to compare the –100 to –200 intergenic regions. Kolmogorov-Smirnov Z test and Mann-Whitney U test showed significant difference between the *P. sojae* and *P. ramorum* datasets. Wald-Wolfowitz Median test showed significant difference in median values (Table 3). We also used a set of 5000 sequences of 1000 bp non-coding regions, statistically significant difference between *P. sojae* (53.6%) and *P. ramorum* (53.1%) was obtained. When two sets of 5000 randomly chosen non-coding sequences of *P. ramorum* were compared, no significant difference can be detected with various test methods (Table 3).

Table 3. GC content in the non-coding regions in *P. sojae* and *P. ramorum*

Non-coding regions from	GC% average	Standard deviation of GC%	GC% median	p value of F-test [c]	p value of T-test [e]	Kolmogorov-Smirnov Z test [e]	Mann-Whitney U test [f]	Wald-Wolfowitz Median test [g]
***P. ramorum* (-100-200 bp)** [a]	51.8	7.6	52	<0.001	**-**	**<0.001**	**<0.001**	**<0.001**
P. sojae *(-100-200 bp)* [a]	52.4	8.4	53					
***P. ramorum* (-1-100 bp)** [a]	53.0	7.3	54	**0.003**	**<0.001**	**<0.001**	**<0.001**	**<0.001**
P. sojae *(-1-100 bp)* [a]	54.1	7.6	55					
***P. ramorum* (-100-200 bp)** [b]	51.9	7.6	52	**0.38**	0.37	0.79	0.61	0.65
P. ramorum *(-100-200 bp)* [b]	51.8	7.6	52					
***P. ramorum* (-1-100 bp)** [b]	52.9	7.3	54	**0.44**	0.85	0.42	0.51	0.95
P. ramorum *(-1-100 bp)* [b]	53.0	7.3	54					

[a] Regions between -100 to -200 bp and between –1 to –100 upstream of the start codons were used for the analysis. From each genome, a set of 10,000 sequences was randomly selected.

[b] Regions between -100 to -200 bp and between –1 to –100 upstream of the start codons were used for the analysis. 5,000 randomly chosen *P. ramorum* sequences were used to compare the other 5,000 randomly chosen *P. ramorum* sequences.

[c] To make the sample a better approximation of Normal distribution, GC percentages were converted into arcsin values before the F-tests. F-test was conducted to determine whether the two samples have different variances. p value is the one-tailed probability that the variances are NOT significantly different ($p > 0.001$).

[d] GC percentages were converted into arcsin values before the T-tests. T-test was performed to determine whether the two samples have the same mean. p value is the probability that two samples come from data sets with the same mean. Two-tailed T-tests were used for comparing the -1-100 bp regions between *P. sojae* and *P. ramorum*, and for comparing two randomly chosen *P. ramorum* data sets.

[e f] Kolmogorov-Smirnov Z test and Mann-Whitney U test were used to determine whether the variable in each of two independent samples comes from the same underlying population. Equal variance was not presumed in these tests. p value is the probability that two samples come from the same underlying population.

[g] Wald-Wolfowitz Median test was conducted to determine whether there is a difference in median values between the two samples. Equal variance was not presumed in the test. p value is the probability that two samples have the same median values.

As compared to *P. ramorum*, the increased GC content in the non-coding regions in *P. sojae* is in line with the higher GC3 value of coding regions. In summary, *P. sojae* shows an increase in GC content in coding regions, 5'UTR and promoter regions, and the intergenic regions of 1.4%, 1.1% and 0.6%, respectively. The shift of base composition in non-coding regions suggests the global effect of mutation bias.

High GC3 is associated with high levels of expression

For *P. sojae,* both the whole genome sequence and EST sequences are available and this offers the opportunity to analyze the association between high GC3 and expression level. The GC3 analysis showed that EST-derived ORFs show higher GC3 (83.5%) than the ORFs derived from whole genome sequences (76.0%) (Table 1). This difference suggests that expressed genes have higher GC3 than the average genes annotated from the genome. To further analyze the relationship between GC3 and expression, *P. sojae* ESTs were divided into two groups: one containing highly expressed genes and the other containing weakly expressed genes. A total of 1602 unique EST contigs derived from zoospores, mycelium and infection tissues from the Phytopathogenic Fungi and Oomycete EST Database (Soanes et al. 2002) was used. From this data set, 936 contigs with less than 3 ESTs were defined as more weakly expressed and contigs with more than 5 ESTs were defined as more highly expressed. On the GC plot, average genes give a GC3 peak around 70%, the more weakly expressed genes show a GC3 peak round 80 –90%, and the more highly expressed genes show a GC3 peak above 90% (Fig. 1C). The distinct peaks show that expressed genes have a higher GC3 than the annotated genes of the genome, and in particular, the high GC3 feature is more pronounced in genes with high expression levels.

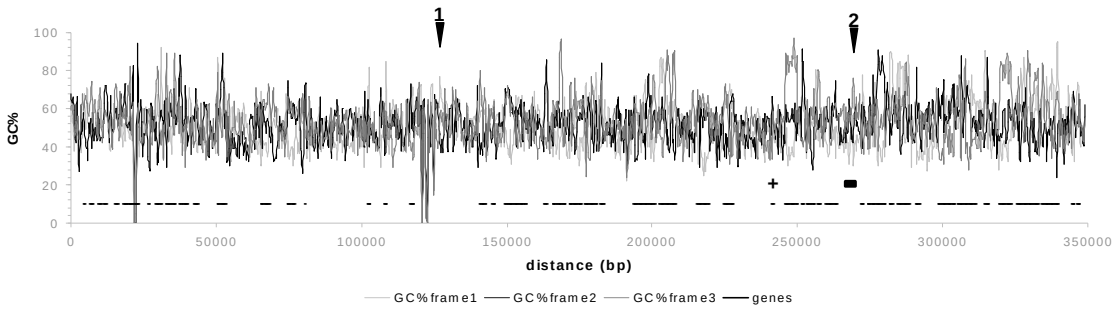
A similar analysis was carried out with a large EST data set generated from a wide range of life stages and culture conditions in *P. infestans* (Randall et al. 2005). 3000 contigs consisting of a single EST were defined as weakly expressed genes and 700 contigs with more than 10 ESTs were defined as highly expressed. The GC3 of weakly expressed genes show a peak around 60% whereas that of the highly expressed genes gives a peak around 80% (Fig. 1D). Therefore the highly expressed genes in *P. infestans* also show higher GC3, which agrees with the results obtained with *P. sojae*.

Genes can be visualized as peaks on a GC plot

Because high GC3 is a feature for *Phytophthora* genes and because random genomic sequences on average have a lower GC content, it should be possible to utilize GC3 values to differentiate coding and non-coding regions in the genome. Previously, *inf1* was shown to appear as a peak in a 12 kb region of *P. infestans* on a GC frame plot (Jiang et al. 2005). To investigate this feature on a larger scale, scaffold 52 (JGI genome release version 1) of *P. ramorum* was used as a whole for a GC frame plot to visualize the position of genes and GC peaks. Scaffold 52 is 350 kb in size and contains 80 annotated genes with various predicted functions such as (de)phosphorylation, transport and DNA repair. On average, genes are 2 kb in size and 3 kb apart from each other. The genes are rather evenly distributed except for two 20 kb gene-poor regions located at around 80 kb and 130 kb from the start. On a genomic GC frame plot, the positions containing the majority of the genes can be shown as GC peaks (Fig. 3A). At a GC threshold of 75%, only 2 GC peaks are found in the intergenic regions (Fig. 3B). Further analysis of these two peak regions revealed that one 9 kb gene is missing in the initial annotation and also one fragmented retrotransposon was not annotated. At a GC threshold of 65%, only one annotated gene

lacks a GC peak, and this is a retrotransposon-like element (Fig. 3A). Therefore, we conclude that on this GC frame plot of 350 kb genomic *P. ramorum* sequence, most GC peaks above 70% represent genes. Analysis of a 110 kb region containing elicitin genes in *P. infestans* also showed similar results with peaks representing genes and some high GC retrotransposons (Jiang et al. 2005). *P. infestans* has a larger genome than *P. ramorum* and the peaks derived from retrotransposons appear more frequently.

(A)



(B)

GC threshold	Genes having peaks above GC threshold	Peaks above GC threshold outside genes	Genes without peaks above GC threshold	Total peaks
65%	79	59	1	139
70%	68	18	12	98
75%	50	2	31	83
80%	36	1	44	81
85%	25	0	55	80

Fig. 3. (A) GC frame plot of scaffold 52 of *P. ramorum.* The GC content was calculated from a scanning window of 300 bp. The positions of 80 annotated genes are indicated underneath the GC graphs. Two sequence gaps gave 0% GC in the plot. Peak 1 and peak 2 are above 75% but no genes were annotated. Peak 1 corresponds to a retroelement. Peak 2 corresponds to an ORF of 9156 bp that was missing in the initial annotation (indicated by a black rectangle). The retroelement indicated by '+' shows no GC peak above 65%. (B) Relationship between GC peaks and 80 genes of scaffold 52 in *P. ramorum*

Retrotransposons have a more variable GC content than genes

The genome of *P. infestans* has heterogeneous groups of retrotransposons. Retrotransposons can exhibit either similar or low GC content as compared to average *Phytophthora* genes. On the GC frame plot of scaffold 52 in *P. ramorum*, several high GC peaks (above 75%) are derived from retrotransposon-like elements and also the only annotated gene without a GC peak above 65% is a retrotransposon. In the whole genome of *P. sojae*, annotated genes with exceptionally low GC content are often retrotransposons. From the 10,000 *P. sojae* ORFs, 16 with GC content less than 50% were selected for further analysis. Except for a few fragmented pseudo-genes, most of them are retrotransposon-like elements (R.H.Y, Jiang., unpublished results).

The previously characterized retrotransposons *GypsyPi-1* (AY830091) and *GypsyPi-3* (AY830104) show high GC3 and a similar codon usage as *P. infestans* genes (Jiang et al. 2005). The level of correlation that is represented by the coefficient of determination (R^2), ranges in value from 0 to 1. If R^2 is 1, there is a perfect correlation; if R^2 is 0, there is no correlation. The correlation of codon usage between *GypsyPi-1/GypsyPi-3* and *P. infestans* genes gave an R^2 value of 0.91 which shows a high level of correlation (Fig. 4A). This similarity is slightly lower than that between *P. infestans* and *P. sojae* genes (R^2 = 0.97) (Fig. 4C). In contrast, another retrotransposon, *CopiaPi-2* (AY830099), does not have high GC3 or similar codon usage as *P. infestans* genes. The regression between *CopiaPi-2* and *P. infestans* genes gives a R^2 value of 0.12 (Fig. 4B). The level of correlation is even lower than that between *P. infestans* and *B. graminis* ORFs (R^2 = 0.18) (Fig. 4D).

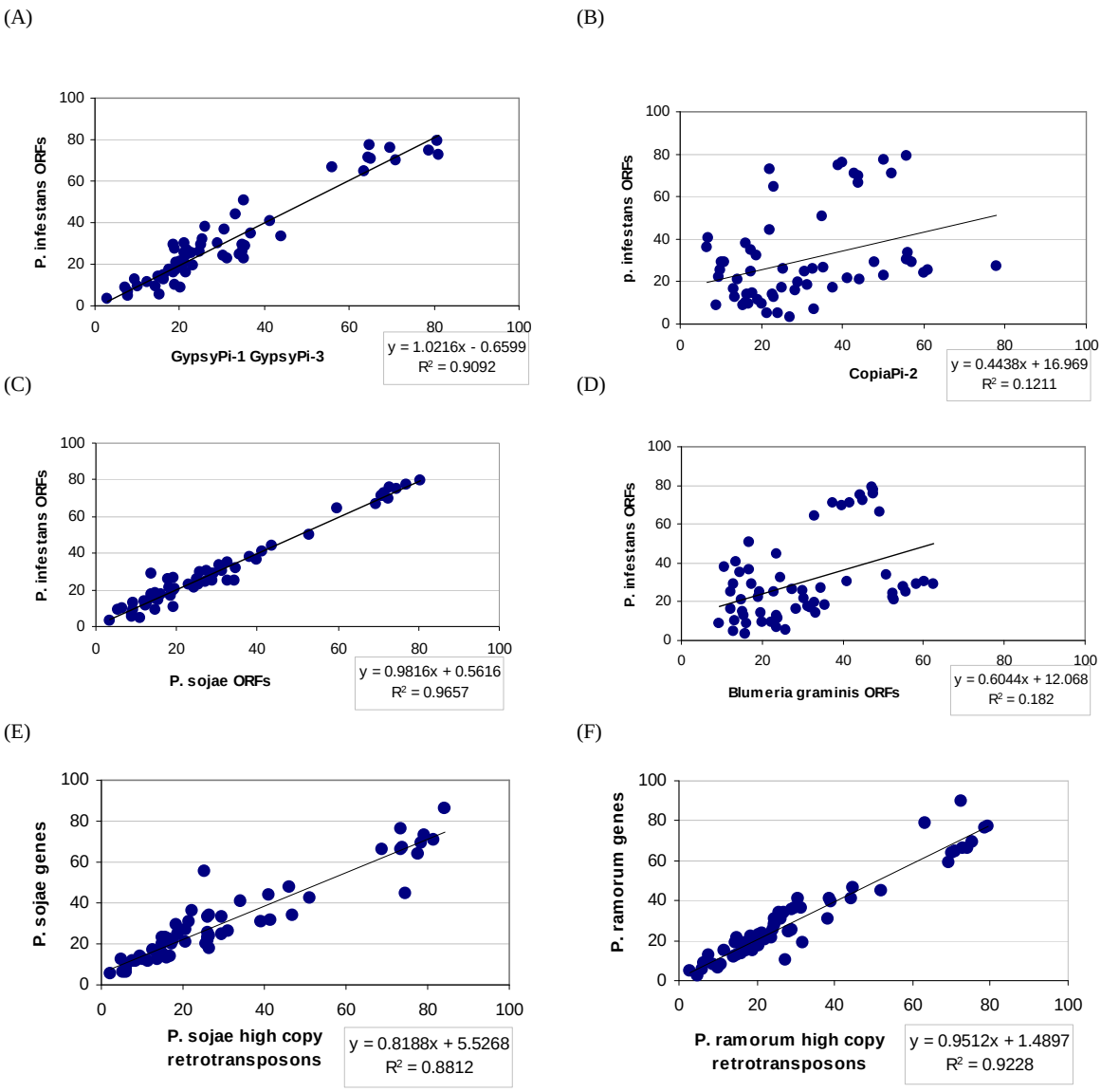
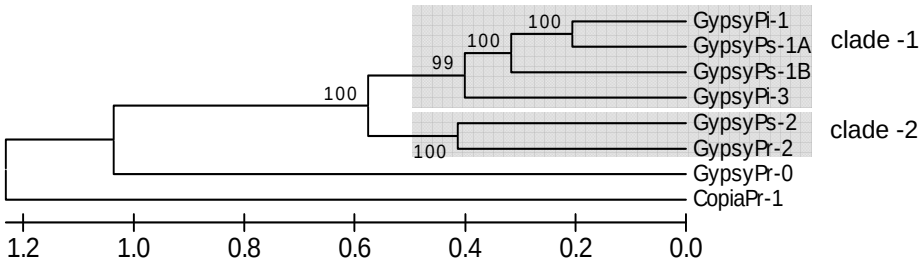


Fig. 4. Correlation of codon usage between different sets of genes. A set of 1000 ORFs from *P. infestans* ESTs was correlated with (A) *GypsyPi-1* and *GypsyPi-3* elements, (B) *CopiaPi-2* element, (C) ORFs derived from *P. sojae* ESTs, and (D) ORFs derived from *B. graminis* ESTs. (E) Correlation between *P. sojae* genes and *P. sojae* high copy retrotransposons. (F) Correlation between *P. ramorum* genes and *P. ramorum* high copy retrotransposons.

The most abundant retrotransposons have a similar codon bias to *Phytophthora* genes

GypsyPi-1 and *GypsyPi-3* of *P. infestans* that both have high GC3, may have undergo recent transposition events. This assumption is based on the high level of sequence similarity between the LTR pairs, but also on the fact that among 7 described retrotransposons, *GypsyPi-1* and *GypsyPi-3* have the highest copy number in 500 kb sequences derived from several regions in the genome. In contrast, the retrotransposon *CopiaPi-2* with low GC3 has less copies than *GypsyPi-1* and *GypsyPi-3* in the regions analyzed (Jiang et al. 2005).

(A)



(B)

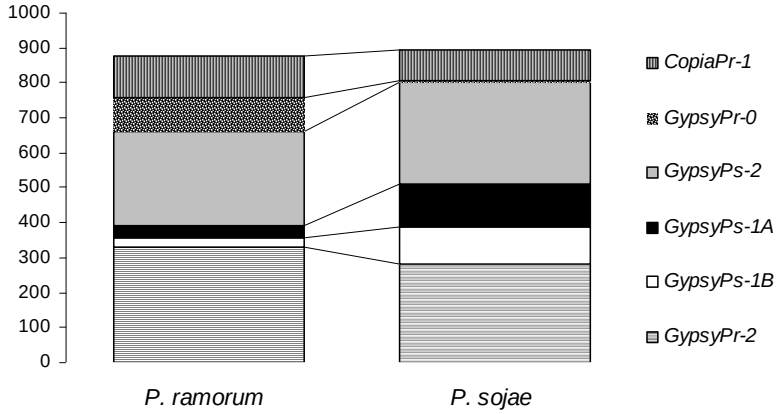


Fig. 5. (A) Phylogenetic tree of the high copy retrotranspons of *Phytophthora*. The reverse transcriptase domains were used to construct the unrooted phylogram based on Neighbor-Joining analysis. Confidence of groupings was estimated by using 1,000 bootstrap replicates; numbers next to the branching point indicate the percentage of replicates supporting each branch. (B) The presence of the six high copy retro-transposons in the genome of *P. sojae* and *P. ramorum*. The Y-axis represents the estimated copy number. Copy number is obtained from BLAST hit with identity percentage > 80% and *E* value < 1e-30.

To investigate whether the higher copy number of retrotransposons is associated with high GC3, several of the most widely spread retrotransposons in the genomes of *P. sojae* and *P. ramorum* were retrieved for analysis. A set of 100 sequences encoding reverse transcriptase domains was extracted from the whole genome sequence *of P. sojae* and 23 unique sequences (BLAST identity < 95%) were found. The copy number was determined by the number of genomic positions with BLAST hit identity of more than 95%. Three sequences with the highest copy number (> 90) were selected, and we assume that these three reverse transcriptases represent three high copy retrotransposons in *P. sojae*. Using a similar approach, three high copy retrotransposons were identified in the *P. ramorum* genome.

These six high copy retrotransposons were subsequently characterized based on BLAST homology. They are all Gypsy like elements with the exception of one Copia like retrotransposon, *CopiaPr-1*. The Gypsy like elements of *P. sojae* were named *GypsyPs-1A, GypsyPs-1B* and *GypsyPs-2*, and of *P. ramorum GypsyPr-2* and *GypsyPr-0.* A phylogenetic tree based on the reverse transcriptase domain was constructed to visualize their relationship (Fig. 5A). Several of the high copy retrotransposons derived from different *Phytophthora* species form one clade. *GypsyPi-1, GypsyPi-3* from *P. infestans* and *GypsyPs-1A, GypsyPs-1B* from *P. sojae* all belong to clade-1. Within clade-1, they typically share 60% BLAST similarity with each other in the reverse transcriptase domain. The other two Gypsy elements *GypsyPs-2* and *GypsyPr-2* from *P. sojae* and *P. ramorum,* respectively, fall in clade-2. Clade-1 and clade-2 therefore represent the most widespread retrotransposons in the *Phytophthora* genomes. The copy number of the retroelements was also estimated by BLASTN hits in the *P. sojae* and *P. ramorum* genomes. Homologous elements were detected with a hit with BLAST *E* value < 1e-30 and identity > 80%. Each of the retrotransposons has a homologue in the other genome and different expansion patterns were found. *GypsyPs-2, GypsyPr2* and *CopiaPr-1* show similar expansions in both genomes, whereas *GypsyPs-1A* and *GypsyPs-1B* are less expanded in *P. sojae* than in *P. ramorum*. The homologue of *GypsyPr-0* is a low copy element in *P. sojae* (Fig. 5B).

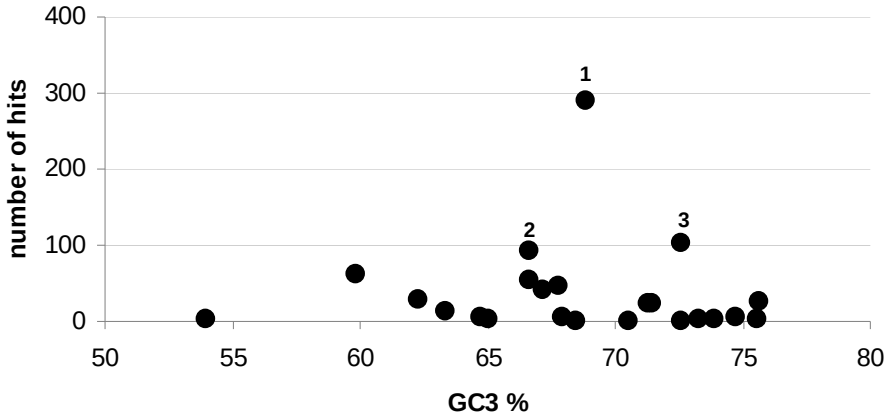


Fig. 6. The relationship between GC3 and copy number of a set of retrotransposons in *P. sojae*. The three high copy retrotransposons are marked with numbers. 1 (*GypsyPs-2*), 2 (*GpysyPs-1A*) and 3 (*GpysyPs-1B*).

From the *P. sojae* genome, the set of 23 sequences coding for reverse transcriptases (3000 bp) were used for GC3 analysis and 21 of them show a higher GC3 value (> 60%) than the average GC content of the genome. Although the increase in GC3 value does not correlate with higher copy numbers (Fig. 6), the three high copy retrotransposons do have a high GC3 value. In *P. sojae*, *GypsyPs-1A, GypsyPs-1B* and *GypsyPs-2* have a GC3 value of 71%, 67% and 69%, respectively. In *P. ramorum*, *GypsyPr-2, GypsyPr-0* and *CopiaPr-1* have a GC3 value of 65%, 65% and 80%, respectively.

The 3000 bp *P. sojae* sequences encoding the reverse transcriptases of the three high copy retrotransposons were used for codon analysis. A high level of correlation of codon usage was found between the high copy retrotransposons and *P. sojae* genes with a R^2 value of 0.88 (Fig. 4E). A similar high level of correlation was found between the high copy retrotransposons and *P. ramorum* genes with a R^2 of 0.92 (Fig. 4F), and this agrees with the high level of correlation found in *P. infestans* (Fig. 4A). We can conclude that in the three *Phytophthora* genomes, high copy retrotransposons share similar codon biases as host genes.

Discussion

Mutational bias is a global force to change base composition whereas selection pressure is a local force acting on coding sequences. In this study we show that both forces participate in shaping codon bias in *Phytophthora*. The majority of the *Phytophthora* species are monophyletic and have evolved recently (Cooke et al. 2000). The close phylogenetic relationship may explain the similar elevated GC content in coding regions in several *Phytophthora* species. However, lineage specific increases of GC content have also occurred. *P. sojae* shows a significantly higher GC content in the non-coding region compared to *P. ramorum*, and so we infer that whole genome mutation bias has shifted the *P. sojae* base composition towards a higher GC content. At the same time, selection pressure can be clearly detected by the correlation of high expression levels and GC content in *P. sojae* and *P. infestans.* Therefore, both mutation bias and selection pressure are at work in the *Phytophthora* genomes, and they drive the codon bias with different emphasis: mutation bias gives rise to differences between species and selection pressure tunes the codon usage to expression levels.

Invasion and replication of retrotransposons can change genome sizes within a short time frame. The difference in genome size between wheat and rice is mainly due to amplification of retrotransposons in the gene-poor regions (Sandhu and Gill 2002). In the last three million years, the maize genome has increased from 1200 Mb to 2400 Mb due to retrotransposon activity (SanMiguel et al. 1998). In *P. infestans,* heterogenous retrotransposons were suggested to be largely responsible for the large genome size of 240 Mb (Jiang et al. 2005). Several of the most abundant retroelements characterized in this study may have invaded the *Phytophthora* genome before speciation due to their presence in different *Phytophthora* species. In addition, lineage specific expansion has occurred. *GpysyPs-1A* and *GypsyPs-1B* appear to have expanded in *P. sojae* but to a lesser extent in *P. ramorum*. *GyspyPr-0* is

expanded in *P. ramorum* but its homologue has only a few copies in *P. sojae*. These expansion patterns may contribute to the different sizes of the *Phytophthora* genomes. A recent burst of *GypsyPi*-1and *GypsyPi-3* activity may have largely contributed to the increase in size of the *P. infestans* genome.

The replication of these genetic parasites involves transcription of the retroelement and synthesis of the reverse transciptase. If a retrotransposon is only under whole genome mutation bias, GC3 should have a similar value as the mean GC content of the genome. Our results show that high copy retrotransposons have an increased GC content similar to coding sequences, which indicates that their codon usage has been driven by selection pressure to optimize protein translation. The transposable elements of *C. elegans, S. cerevisiae, A. thaliana, D. melanogaster* and *H. sapiens* appear to be AT-rich regardless of the base composition of their host genomes (Lerat et al. 2002). The AT bias of retrotransposons was suggested to reflect the AT-rich characteristics of the reverse transcriptase of retroviruses. For example, in some lentiviruses the error prone reverse transcriptase may lead to the preference of G-to-A and C-to-T transition (Zsiros et al. 1999). However, in the case of the three *Phytophthora* genomes, high copy retrotransposons are likely to be primarily under selection pressure similar to that imposed on host genes.

Retrotransposon activity can be disastrous to hosts because it is able to cause massive deleterious mutations. Tight host control of retrotransposon mobilization is needed in any viable species. Multiple steps can be used to control retrotransposon activity, such as reduction of expression, degradation of transcripts by gene silencing and limiting integration (Labrador and Corces 1997). Facing host surveillance systems, successful retrotransposons must be able to handle these control steps in order to spread in the genome. In *Phytophthora*, mimicry of host codon usage can be beneficial for genetic parasitic elements to optimize their production of reverse transcriptase. Utilizing the host cellular machinery efficiently may be one of the strategies for retrotransposons to propagate successfully in the genome.

Acknowledgements

We thank Brett Tyler for helpful suggestions and discussions and the DOE-Joint Genome Institute (http://www.jgi.doe.gov/index.html), and in particular Brett Tyler, Jeff Boore and Dan Rokhsar, for *P. sojae* and *P. ramorum* genome sequences. This work was financially supported by an Aspasia grant from the Netherlands Organisation for Scientific Research (NWO-Aspasia (015.000.057).

References

Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25:3389-3402

Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou SG, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kroger N, Lau WWY, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS (2004) The genome of the diatom *Thalassiosira pseudonana*: Ecology, evolution, and metabolism. Science 306:79-86

Baldauf SL (2003) The deep roots of eukaryotes. Science 300:1703-6

Baldauf SL, Roger AJ, Wenk-Siefert I, Doolittle WF (2000) A kingdom-level phylogeny of eukaryotes based on combined protein data. Science 290:972-977

Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. Plant Mol Biol 42:251-269

Bowman S, Lawson D, Basham D, Brown D, Chillingworth T, Churcher CM, Craig A, Davies RM, Devlin K, Feltwell T, Gentles S, Gwilliam R, Hamlin N, Harris D, Holroyd S, Hornsby T, Horrocks P, Jagels K, Jassal B, Kyes S, McLean J, Moule S, Mungall K, Murphy L, Oliver K, Quail MA, Rajandream MA, Rutter S, Skelton J, Squares R, Squares S, Sulston JE, Whitehead S, Woodward JR, Newbold C, Barrell BG (1999) The complete nucleotide sequence of chromosome 3 of *Plasmodium falciparum*. Nature 400:532-538

Chen SL, Lee W, Hottes AK, Shapiro L, McAdams HH (2004) Codon usage between genomes is constrained by genome-wide mutational processes. Proc Natl Acad Sci USA 101:3480-5

Cooke DE, Drenth A, Duncan JM, Wagels G, Brasier CM (2000) A molecular phylogeny of *Phytophthora* and related oomycetes. Fungal Genet Biol 30:17-32

Duret L, Mouchiroud D (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis, Drosophila,* and *Arabidopsis*. Proc Natl Acad Sci USA 96:4482-7

Eichinger L, Pachebat JA, Glockner G, Rajandream MA, Sucgang R, Berriman M, Song J, Olsen R, Szafranski K, Xu Q, Tunggal B, Kummerfeld S, Madera M, Konfortov BA, Rivero F, Bankier AT, Lehmann R, Hamlin N, Davies R, Gaudet P, Fey P, Pilcher K, Chen G, Saunders D, Sodergren E, Davis P, Kerhornou A, Nie X, Hall N, Anjard C, Hemphill L, Bason N, Farbrother P, Desany B, Just E, Morio T, Rost R, Churcher C, Cooper J, Haydock S, van Driessche N, Cronin A, Goodhead I, Muzny D, Mourier T, Pain A, Lu M, Harper D, Lindsay R, Hauser H, James K, Quiles M, Babu MM, Saito T, Buchrieser C, Wardroper A, Felder M, Thangavelu M, Johnson D, Knights A, Loulseged H, Mungall K, Oliver K, Price C, Quail MA, Urushihara H, Hernandez J, Rabbinowitsch E, Steffen D, Sanders M, Ma J, Kohara Y, Sharp S, Simmonds M, Spiegler S, Tivey A, Sugano S, White B, Walker D, Woodward J, Winckler T, Tanaka Y, Shaulsky G, Schleicher M, Weinstock G, Rosenthal A, Cox EC, Chisholm RL, Gibbs R, Loomis WF, Platzer M, Kay RR, Williams J, Dear PH, Noegel AA, Barrell B, Kuspa A (2005) The genome of the social amoeba *Dictyostelium discoideum*. Nature 435:43-57

Erwin DC, Ribeiro OK (1996) *Phytophthora* diseases worldwide. The American Phytopathological Society, St. Paul Minnesota USA

Fraser CM, Casjens S, Huang WM, Sutton GG, Clayton R, Lathigra R, White O, Ketchum KA, Dodson R, Hickey EK, Gwinn M, Dougherty B, Tomb JF, Fleischmann RD, Richardson D, Peterson J, Kerlavage AR, Quackenbush J, Salzberg S, Hanson M, vanVugt R, Palmer N, Adams MD, Gocayne J, Weidman J, Utterback T, Watthey L, McDonald L, Artiach P, Bowman C, Garland S, Fujii C, Cotton MD, Horst K, Roberts K, Hatch B, Smith HO, Venter JC (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. Nature 390:580-586

Hraber PT, Weller JW (2001) On the species of origin: diagnosing the source of symbiotic transcripts. Genome Biol 2:37

Ishikawa J, Hotta K (1999) FramePlot: a new implementation of the frame analysis for predicting protein-coding regions in bacterial DNA with a high G + C content. FEMS Microbiol Lett 174:251-3

Jiang RH, Dawe AL, Weide R, van Staveren M, Peters S, Nuss DL, Govers F (2005) Elicitin genes in *Phytophthora infestans* are clustered and interspersed with various transposon-like elements. Mol Genet Genomics 273:20-32

Kamoun S, Hraber P, Sobral B, Nuss D, Govers F (1999) Initial assessment of gene diversity for the oomycete pathogen *Phytophthora infestans* based on expressed sequences. Fungal Genet Biol 28:94-106

Kanaya S, Yamada Y, Kudo Y, Ikemura T (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNAs: gene expression level and species-specific diversity of codon usage based on multivariate analysis. Gene 238:143-55

Kazazian HH, Jr (2004) Mobile elements: drivers of genome evolution. Science 303:1626-1632

Kidwell MG, Lisch DR (2001) Perspective: transposable elements, parasitic DNA, and genome evolution. Inter J Organic Evol 55:1-24

Kumar S, Tamura K, Jakobsen IB, Nei M (2001) MEGA2: molecular evolutionary genetics analysis software. Bioinformatics 17:1244-5

Labrador M, Corces VG (1997) Transposable element-host interactions: regulation of insertion and excision. Annu Rev Genet 31:381-404

Latijnhouwers M, de Wit PJ, Govers F (2003) Oomycetes and fungi: similar weaponry to attack plants. Trends Microbiol 11:462-9

Lerat E, Capy P, Biemont C (2002) Codon usage by transposable elements and their host genes in five species. J Mol Evol 54:625-637

Margulis L, Schwarts KV (2000) Five Kingdoms: an illustrated guide to the phyla of life on earth. W.H. Freeman and company, New York

McLeod A, Smart CD, Fry WE (2004) Core promoter structure in the oomycete *Phytophthora infestans*. Eukaryot Cell 3:91-9

Nekrutenko A, Li W-H (2001) Transposable elements are found in a large number of human protein-coding genes. Trends Genetics 17:619-621

Qutob D, Hraber PT, Sobral BWS, Gijzen M (2000) Comparative analysis of expressed sequences in *Phytophthora sojae*. Plant Physiol 123:243-253

Randall TA, Dwyer RA, Huitema E, Beyer K, Cvitanich C, Kelkar H, Fong AM, Gates K, Roberts S, Yatzkan E, Gaffney T, Law M, Testa A, Torto-Alalibo T, Zhang M, Zheng L, Mueller E, Windass J, Binder A, Birch PR, Gisi U, Govers F, Gow NA, Mauch F, van West P, Waugh ME, Yu J, Boller T, Kamoun S, Lam ST, Judelson HS (2005) Large-scale gene discovery in the oomycete *Phytophthora infestans* reveals likely components of phytopathogenicity shared with true fungi. Mol Plant Microbe Interact 18:229-43

Rizzo DM, Garbelotto M, Hansen EM (2004) *Phytophthora Ramorum*: Integrative Research and Management of an Emerging Pathogen in California and Oregon Forests. Annu Rev Phytopathol

Sandhu D, Gill KS (2002) Gene-containing regions of wheat and the other grass genomes. Plant Physiol 128:803-11

SanMiguel P, Gaut BS, Tikhonov A, Nakajima Y, Bennetzen JL (1998) The paleontology of intergene retrotransposons of maize. Nature Genetics 20:43-45

Scala S, Carels N, Falciatore A, Chiusano ML, Bowler C (2002) Genome properties of the diatom *Phaeodactylum tricornutum*. Plant Physiol 129:993-1002

Sharp PM, Li WH (1986) An evolutionary perspective on synonymous codon usage in unicellular organisms. J Mol Evol 24:28-38

Sharp PM, Stenico M, Peden JF, Lloyd AT (1993) Codon usage: mutational bias, translational selection, or both? Biochem Soc Trans 21:835-41

Soanes DM, Skinner W, Keon J, Hargreaves J, Talbot NJ (2002) Genomics of phytopathogenic fungi and the development of bioinformatic resources. Mol Plant Microbe Interact 15:421-7

Topp CN, Zhong CX, Dawe RK (2004) Centromere-encoded RNAs are integral components of the maize kinetochore. Proc Natl Acad Sci USA 101:15986-15991

Zsiros J, Jebbink MF, Lukashov VV, Voute PA, Berkhout B (1999) Biased nucleotide composition of the genome of HERV-K related endogenous retroviruses and its evolutionary implications. J Mol Evol 48:102-111

Chapter 9

General discussion

In the thesis, various genomic features of the destructive plant pathogen *Phytophthora* have been described. In *P. infestans,* gene amplification at an avirulent gene locus was found to be related to avirulence, and amplified sequences were shown to provide modular diversity for the evolution of the avirulence associated gene *pi3.4*. Database mining showed that elicitins are ubiquitous among *Phytophthora* species and belong to one of the most highly conserved and complex protein families in the *Phytophthora* genus. Phylogeny construction indicated that most of the diversified elicitin family members existed prior to divergence of the common ancestor of *Phytophthora*. Comparative genomics revealed synteny between *Phytophthora sojae* and *Phytophthora ramorum* and synteny break points were found to often harbor effector-like genes. Secretome analysis showed that secreted proteins constitute some of the most dynamic parts of the proteome. In particular, the 'RXLR-DEER' type effector genes are rapidly evolving and may play important roles in the interaction with host plants. In this Chapter, findings described in this thesis are discussed from a genomic and evolutionary point of view.

Gene amplification and change of virulence

Gene amplification is one of the driving forces of genome evolution. It probably is ubiquitous in the genomes of bacteria and contributes substantially to the prokaryotic genomic plasticity. The essential role of gene amplification in adaptive evolution has been demonstrated in a wide range of processes, such as antibiotics resistance *in Proteus vulgaris*, tolerance to heavy metals in *Thiobacillus ferrooxidans*, adaptation to scarcity of nutrients in *Salmonella typhimurium* and increased virulence in *Vibrio cholerae* (Romero and Palacios 1997). In eukaryote, gene amplification is frequently observed in cancer development and cancer cells after prolonged treatment with xenobiotics. For example, amplified oncogenes can deregulate cell cycles and amplified transporter genes can lead to drug resistance in cancer cells (Schwab 1994; Lengauer, Kinzler, and Vogelstein 1998). Also in insects and plants, gene amplification has been reported to cause drug resistance, i.e., insecticide and herbicide resistance, respectively (Donn et al. 1984; Field, Devonshire, and Forde 1988).

From the perspective of genome architecture, gene amplification also causes the genome structure to undergo drastic changes. *Streptomyces*, for example, shows a very high genetic instability and gene amplification is part of a cascade of genomic rearrangements. Massive chromosomal deletions can be over 1 Mb in size and such deletions are either caused or triggered by gene amplification (Birch, Hausler, and Hutter 1990; Leblond and Decaris 1994). In humans cancer progression is very frequently associated with impaired genome stability such as gene amplification, large scale deletions and translocations (Lengauer, Kinzler, and Vogelstein 1998).

In *P. infestans*, comparative genomic hybridization showed that a total of six loci possess copy number variation which indicates six 'hotspots' for genome rearrangements and potential loci for virulence related genes (Chapter 3). One is the avirulence locus *Avr3b-Avr10-Avr11* that contains numerous copies of the *pi3.4* gene. The gene amplification of *pi3.4* has led to an increased number of transcripts (unpublished

data), but it is not known whether these transcripts give rise to functional proteins. Undoubtedly the amplification of *pi3.4* also causes genome instability. The amplified region in *P. infestans* is highly rearranged as compared to *P. sojae* and *P. ramorum*. Genetic mapping showed that this locus is hemizygous (van der Lee et al. 2001), and hybridization confirmed that the amplified gene cluster was deleted in various virulent isolates. Therefore, in *P. infestans*, the change in phenotype from AVR3b-AVR10-AVR11 (i.e. avirulent on *R3b*, *R10* and *R11* plants) to avr3b-avr10-avr11 may be due to the deletion of a large region containing the amplified gene cluster.

Apart from the capability of over-expressing genes and changing genome architecture, the gene amplification at the *Avr3b-Avr10-Avr11* locus may also be reminiscent for a novel mechanism to create new genes. The amplified domain can recruit varied modules to generate new mosaic genes. As compared to the process of random mutation, such a mechanism speeds up gene evolution and may help the pathogen to rapidly adapt to its environment.

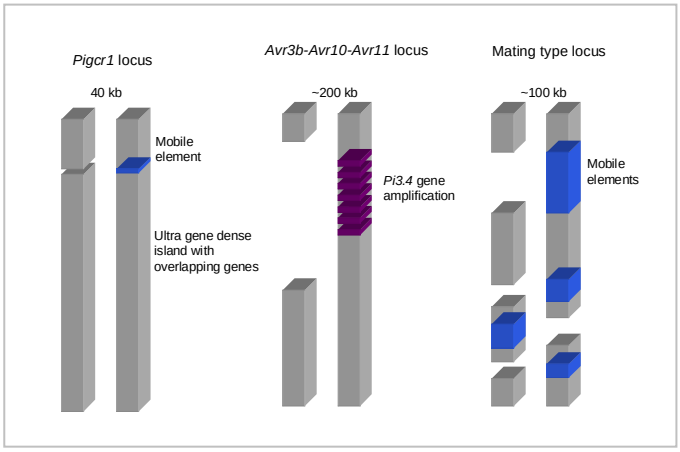


Figure 1. **Three examples of allelic polymorphism in the *P. infestans* genome** The three regions are of different sizes. See text for further details.

The hemizygous *Avr3b-Avr10-Avr11* locus belongs to the few loci with extreme polymorphism so far identified in the genome *of P. infestans.* One other highly polymorphic region is the mating type locus (Ah Fong and Judelson 2004) with large insertions of mobile elements (Figure 1). In comparison, a gene dense region containing the *PiGCR1* gene shows very little polymorphism between the haplotypes (Figure 1). The avirulence and mating type associated loci may belong to the most re-arranged regions in the genome.

Footprints of evolution

Life styles have left footprints of evolution in the genomes. When two genomes show sufficient similarity to be aligned but also enough divergence to detect architectural changes, such footprints can be revealed (Mira, Klasson, and Andersson 2002). Between free-living bacterial species, rearrangements

are frequently found (Perna et al. 2001; Edwards, Olsen, and Maloy 2002). In contrast, some intracellular symbionts hardly undergo any structural changes. For example, two obligate symbionts of aphids exhibit nearly total conservation in their genomes after perhaps 50 million years divergence (Tamas et al. 2002).

In pathogen genomes, hot spots for rearrangements often harbor pathogenesis-related genes. *Listeria monocytogenes* is a pathogen causing food-borne infections. Its genome has a nearly perfect synteny to the closely related non-pathogenic species *Listeria innocua.* However, the synteny is broken at the locus of genes required for chemotaxis and motility (Buchrieser et al. 2003) which may explain part of their life style differences. In another comparison, striking conservation of gene synteny was found between the human and rodent malaria parasite, *Plasmodium falciparum* and *Plasmodium yoelii,* respectively. However, genes triggering host immune responses are frequently found to be species-specific (Carlton, Silva, and Hall 2005).

Differences in selection pressure may explain why genes responsible for pathogenesis are often found to be specific for a lineage or to undergo rapid changes. Genome evolution is primarily shaped by events of mutation, translocation and duplication. Since different genes are under different selection pressure, some of the rearrangements will be favored while others may be selected against. While comparing *P. sojae and P. ramorum*, the sharp contrast in evolutionary patterns observed for some neighboring genes may reveal such selection pressure differences (Chapter 6).

Overall colinearity was found between the genomes of *P. sojae* and *P. ramorum* (Chapter 6). The cutinase and esterase gene families were found to be located side by side in the genome. The encoded proteins have a different subcellular localization, cutinases are secreted whereas esterases are intracellular. In contrast to the esterase genes, cutinase genes were found to be expanded in *P. sojae* (Figure 2). In the same region, also two types of elicitor genes were found, one encoding the general elicitor elicitin and the other encoding a putative specific elicitor. Apparently, the specific elicitor gene was deleted from *P. sojae* (Figure 2). The potential role of cutinases and specific elicitors in early stages of the interaction with the host may have caused selection pressure acting in favor of changes.

Mobile elements are probably important factors that promote rearrangement events in the genome. They can jump from one place to the other along stretches of endogenous DNA and may also assist in unequal crossing overs. For example, transposable elements are proposed to cause extensive genome rearrangements in the genomes of free living bacteria. Whereas the streamlined genomes of endosymbionts may have retained integrity because of lack of mobile elements (Mira, Klasson, and Andersson 2002), the genomes of *Phytophthora* possess heterogeneous groups of mobile elements and many of them are found to border or intersperse rearranged regions. They may contribute to form ‘hotspots’ of rearrangements.

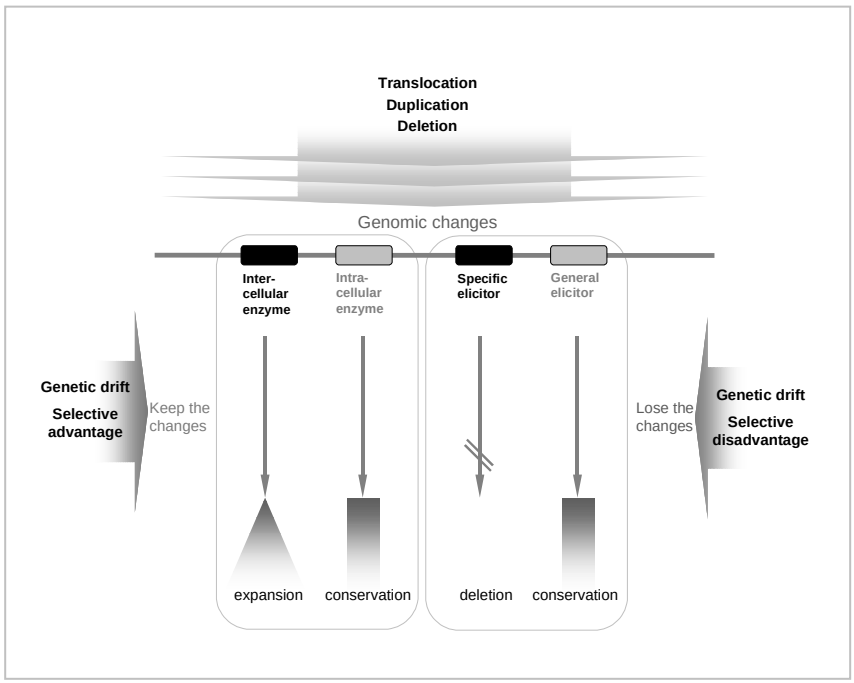


Figure 2. **Different evolutionary patterns of neighboring genes in *P. sojae*.** The ancestral genomic region containing the enzyme and elicitor gene families is exposed to different evolutionary forces indicated by block arrows. Forces to keep and to lose the genome rearrangements are both present.

When cruising down along 760 kb genomic sequence of *P. sojae* and comparing it with the sequence *of P. ramorum*, six large insertion/deletion blocks can be found and three of them harbor elicitor like genes. In this same region, expansion of the cutinase gene family has taken place. This unequal distribution of genome rearrangements showed that the selection on genome structure is not neutral. In particular, some rearrangements leading to a change in the repertoire of pathogenesis-related genes could be preferably maintained during evolution.

Fast paces of evolution – exterior and sex

In *Phytophthora*, not only genome architecture evolves with non-uniform dynamics in various regions; also the secretome evolves at a different pace compared to the average proteome. Moreover, individual protein families also change at different rates within the secretome (Chapter 7).

For a pathogen, the secretome comprises of large number of molecules that the pathogen uses to kill or modulate host cells, and to interact with its hosts. Selection imposed by hosts presumably works particularly on these sets of proteins. Several secreted protein families associated with pathogenicity are expanded in the rice blast pathogen *Magnaporthe grisea* as compared to the bread mold fungus *Neurospora crassa* (Dean et al. 2005). The *Phytophthora* secretome is changing at a faster rate than the overall proteome. For example, in *P. sojae*, around 20% of the secretome is encoded by genes that are unique in this species when compared to *P. ramorum*. In contrast, in the whole *P. sojae* genome, only around 10% of the genes is not present in *P. ramorum*. In the reverse comparison, these percentages are 15% and 4%, respectively.

Protein families within a secretome are diverging at a different pace. Some enzyme family members are highly conserved, possibly due to the functional constraint of their enzymatic performances. During interaction of a pathogen with its host environment, surface proteins have important roles in pathogenesis and eliciting host defenses. For example, in *L. monocytogenes,* surface proteins constitute the major virulence factors and they have essential functions such as facilitating the invasion of eukaryotic cells (Buchrieser et al. 2003). As in many other organisms, the *Phytophthora* cell surface is covered by many proteins. Some will be associated with the cell wall, others are anchored to the cell membrane via post-translational modifications (Chapter 6 and Chapter 7). Several families of cell surface associated proteins were characterized as fast evolving. This may be a result of relaxation of selection pressure, or alternatively, positive selection imposed by biotic and abiotic stresses.

Genes involved in reproduction have been shown to evolve at a fast rate. The most famous example is the human SRY (Sex-determining Region of the Y chromosome) gene. SRY is a transcription factor that acts as a male-dominant trigger for testis differentiation. The extensive variation between mammals implies that SRY structure and function evolve very rapidly, a feature shared by several other genes on the Y chromosome (Graves 2002). Genome wide comparison showed rapid evolution of sex regulators between the nematode *Caenorhabditis elegans* and other nematode species (Stothard and Pilgrim 2003).

In *Phytophthora*, a family of highly divergent proteins named EHD-M96 appears to be fast evolving (Chapter 7). One member of this family is the previously identified mating induced protein M96 of *P. infest*ans (Fabritius, Cvitanich, and Judelson 2002). The EHD-M96 proteins possess many repeats and are likely to play a role in coating sexual spores of *Phytophthora*. In plants, pollen surface proteins containing glycine repeats are undergoing rapid evolution, and this fast evolving pace is primarily due to duplication, deletion, and divergence of repetitive sequences (Fiebig, Kimport, and Preuss 2004). In *Phytophthora,* the extensive repeats in EHD-M96 proteins may play a similar role in the family evolution. Despite the extreme morphological and physiological differences between the various types of spores and gametes that are involved in sexual reproduction (Figure 3), the proteins coating oospores in *Phytophthora* probably belong to the same fast evolving class of proteins that coat pollen grains, eggs and sperms.

A

B

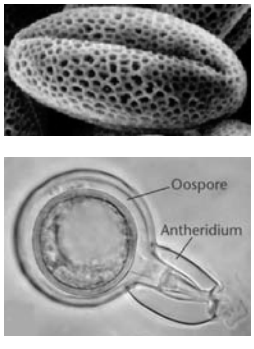


Figure 3. Plant pollen and *Phytophthora* sexual spore. (A) Scanning electron micrograph of a mature *Brassica* pollen grain (Robert et al. 2004) **(B)** Oospore with the remnant of an antheridium. (Photo courtesy of Plant Pathology Section, West Virginia University)

Another mating-induced gene in *Phytophthora* has been proposed to be associated with sterol binding (Fabritius, Cvitanich, and Judelson 2002). Elicitin genes form large complex families in *Phytophthora* species (Chapter 3). In *P. infestans* the most divergent family member of all, *inl4b,* was characterized as a mating-induced gene (Fabritius, Cvitanich, and Judelson 2002). Considering the sterol carrier function of some other family members, INL4B may intercept sterol-based signals or substrates in *Phytophthora*. In fungi, mating partner recognition depends on peptide pheromones and receptors with a high degree of specificity and also these pheromone genes are rapidly evolving (Casselton and Olesnicky 1998). INL4B may acquire specificity for a similar reason.

The wild herds in the genome: RXLR-DEER

Another family of fast evolving genes in the genome that is most likely involved in virulence and host specificity of *Phytophthora* is the RXLR-DEER super-family, which comprises four recently identified ecotype- or cultivar-specific avirulence genes (Chapter 7). Pathogen-derived avirulence genes are able to trigger highly specific host resistance. In fungi, cultivar-specific avirulence proteins share little sequence similarity but many of them possess an even number of cysteines to stabilize proteins in the apoplast (van't Slot and Knogge 2002). To date, four oomycete *Avr* genes have been cloned, two from *Phytophthora* and two from the closely related downy mildew pathogen *Hyaloperonospora parasitica* (Allen et al. 2004; Shan et al. 2004; Armstrong et al. 2005; Rehmany et al. 2005). The commonly present cysteine residues in fungal AVRs are lacking in these AVRs. Strikingly, a RXLR-DEER motif was found in all four oomycete AVRs and it resembles the host targeting motif used by the malaria parasite *Plasmodium* (Rehmany et al. 2005). The occurrence of a common motif in unrelated pathogens with a very different host range, plants and humans, suggests a shared novel mechanism to interact with the eukaryotic host cells.

The sequence of the RXLR-DEER family members is highly divergent. From the 10 largest families in the secretome of *P. ramorum*, two are RXLR-DEER families. They are the only two out of ten that have homologues with low similarity. The sequence divergence in the RXLR-DEER superfamily is evident when comparing the evolutionary pattern with that of another elicitor gene family, i.e., the elicitin gene family (Figure 4). Elicitins form a large and complex family in *Phytophthora*. Phylogeny construction of the elicitin gene family indicated that most of the diversified members existed prior to divergence of the common ancestor of *Phytophthora,* and the family members are highly conserved between the present day species (Chapter 5). In contrast, the members of the RXLR-DEER family differ widely between species, and the ancestral state of the family is not maintained in individual *Phytophthora* species.

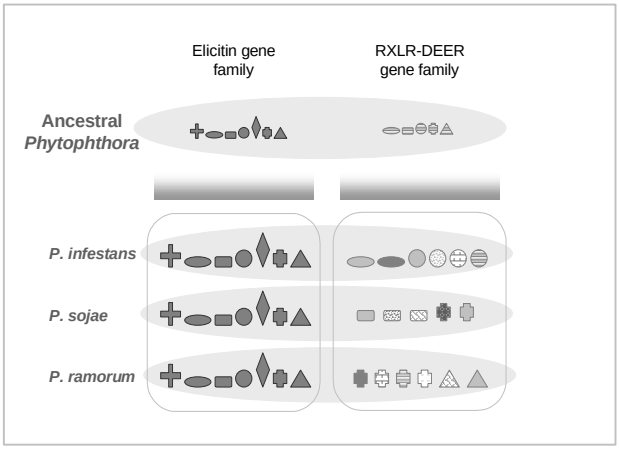


Figure 4. Different patterns of evolution in the elicitin gene family and the RXLR-DEER gene family.
Various forms and shades represent gene family members with sequence divergences. See text for further details.

The genome distribution of RXLR-DEER family members is mostly scattered and clustering of more than three genes is rare. In contrast, members of the other large families often form clusters in the genome. The genomic loci containing RXLR-DEER genes have undergone extensive rearrangements causing loss or gain of genes. In the syntenic regions described in Chapter 6, the six genes encoding proteins with a RXLR-DEER motif all appear to be deleted from the other species. Genomic rearrangements seem to be a rule and synteny is mostly broken by this type of genes. High selection pressure favoring such changes may have led to the divergent sequences of RXLR-DEER genes, and to the dynamics of the loci that harbors these genes. Because four RXLR-DEER members have been demonstrated to be ecotype- or cultivar-specific avirulence factors, we assume that the selection pressure is exerted by the host plants.

Mobile elements and their hosts: genome sizes and choices

The most abundant genetic elements in higher eukaryotes are mobile elements. Their nature was used to be viewed as purely parasitic, but newly emerging data show that mobile elements play a central role in the structure, function and evolution of eukaryotic genomes (Bennetzen 2000; Kazazian 2004) such as providing cis-regulatory sequences, assembling a kinetochore and speed up protein diversification (Nekrutenko and Li 2001; Topp, Zhong, and Dawe 2004). The relationship between transposons and hosts may have to be revised into a continuum from extreme parasitism to mutualism (Kidwell and Lisch 2001).

Retrotransposons are the most abundant mobile elements and are present in all eukaryotic genomes. Retrotransposon insertions are able to increase genome sizes within a short time frame. Comparison of sequence divergence between maize and sorghum showed that the maize genome has increased a staggering 1200 Mb in size in the last three million years due to retrotranposition activity (SanMiguel et al. 1998). In *Phytophthora*, retrotransposons are found to have different expansion patterns in different lineages (Chapter 8). The variation in genome sizes within the *Phytophthora* genus may be due to the difference in activity of transposons present in the various species. *P. infestans* has a large genome (240 Mb) as compared to other *Phytophthora* species, and some of the heterogeneous retrotransposons

found in *P. infestans* may be largely responsible for this expansion in genome size (Chapter 4).

Viruses and retroelements are obligate genetic entities. Their replication hitchhikes on the host cellular machinery. The transposable elements of yeast, worm, fly, human and mustard weed do not appear to mimic their host genes because they are AT-rich regardless of the base composition of their host genomes (Lerat, Capy, and Biemont 2002). Retroviruses often show AT-rich characteristics in their reverse transcriptase and this trend may be reflected in retrotransposons (Zsiros et al. 1999). In *Phytophthora*, however, the codon bias of high copy retrotransposons is similar to that of the host genes. Production of the reverse transcriptase encoded by the retroelements is needed for retrotransposons to replicate, and therefore the translation of this protein will be subjected to similar selection pressure as host transcripts. Selection pressure has pushed up GC3 in host genes of *Phytophthora*, and probably shaped the codon bias of the high copy retrotransposons (Chapter 8).

Future perspectives

P. infestans continues to cause immense economical and environmental damage, and an efficient and imminent solution is not yet available. To date, late blight control still heavily relies on chemical agents. *R*-gene introgression into commercial cultivar demands tremendous breeding efforts and so far *R*-gene-based resistance is not sufficient to prevent late blight epidemics. To solve the problem, more in depth knowledge of the pathogen is needed. Genomics opens a vast new field to explore the world of *Phytophthora*.

This thesis illustrates the power of genome mining and comparative genomics. The plasticity of the pathogen genomes was studied by comparing the variation between different isolates and various species. Copy number variation derived from gene amplification was studied on a whole genome scale with comparative genomic hybridization, and differences in expression profiles between strains were detected by cDNA-AFLP and Affymetrix© arrays. Gene mining revealed the reservoir of secreted protein genes in *Phytophthora*, thereby offering insight into part of the pathogen's repertoire of pathogenesis. Comparative genomics pointed out the hotspots of effector-like genes in the genome and predicted important virulence-related roles for several families of fast evolving genes.

Genome sequencing of *P. infestans* is now in progress at the Broad Institute, MIT. New momentum in late blight research will be gained by exploring the *P. infestans* genome sequences and integrating the genome data with biochemical and physiological data. Late blight researchers should be able to study genes in the context of their genome localization, variation between isolates, divergence across different species and potential contribution to pathogenesis. Analysis of the complete *P. infestans* genome sequences shall answer many questions. For example, what has caused the large genome size of *P. infestans* as compared to other *Phytophthora* species? What is the quantitative contribution of

retrotransposon transposition and gene family expansion in the evolutionary process of genome expansion? What is the size of its sceretome and what are the unique features as compared to the secretomes of other *Phytophthora* species? What constitutes the conserved virulence determinants? Where are the hypervariable genome regions and how many rapidly evolving families are there? What proportion of the proteome can be used as suitable drug targets? And what proportion can be exploited by plant breeders to develop new resistance strategies? From fundamental research to practical applications: the blueprint of *P. infestans* will offer many new opportunities and guidelines to tackle the notorious Irish famine pathogen.

Reference

Ah Fong, A. M., and H. S. Judelson. 2004. The hAT -like DNA transposon DodoPi resides in a cluster of retro- and DNA transposons in the stramenopile *Phytophthora infestans*. Mol Genet Genomics **271**:577-585.

Allen, R. L., P. D. Bittner-Eddy, L. J. Grenville-Briggs, J. C. Meitz, A. P. Rehmany, L. E. Rose, and J. L. Beynon. 2004. Host-parasite coevolutionary conflict between *Arabidopsis* and downy mildew. Science **306**:1957-1960.

Armstrong, M. R., S. C. Whisson, L. Pritchard, J. I. Bos, E. Venter, A. O. Avrova, A. P. Rehmany, U. Bohme, K. Brooks, I. Cherevach, N. Hamlin, B. White, A. Fraser, A. Lord, M. A. Quail, C. Churcher, N. Hall, M. Berriman, S. Huang, S. Kamoun, J. L. Beynon, and P. R. Birch. 2005. An ancestral oomycete locus contains late blight avirulence gene *Avr3a*, encoding a protein that is recognized in the host cytoplasm. Proc Natl Acad Sci USA **102**:7766-7771.

Bennetzen, J. L. 2000. Transposable element contributions to plant gene and genome evolution. Plant Mol Biol **42**:251-269.

Birch, A., A. Hausler, and R. Hutter. 1990. Genome rearrangement and genetic instability in *Streptomyces* spp. J Bacteriol **172**:4138-4142.

Buchrieser, C., C. Rusniok, F. Kunst, P. Cossart, and P. Glaser. 2003. Comparison of the genome sequences of *Listeria monocytogenes* and *Listeria innocua*: clues for evolution and pathogenicity. FEMS Immunol Med Microbiol **35**:207-213.

Carlton, J., J. Silva, and N. Hall. 2005. The genome of model malaria parasites, and comparative genomics. Curr Issues Mol Biol **7**:23-37.

Casselton, L. A., and N. S. Olesnicky. 1998. Molecular genetics of mating recognition in basidiomycete fungi. Microbiol Mol Biol Rev **62**:55-70.

Dean, R. A., N. J. Talbot, D. J. Ebbole, M. L. Farman, T. K. Mitchell, M. J. Orbach, M. Thon, R. Kulkarni, J. R. Xu, H. Pan, N. D. Read, Y. H. Lee, I. Carbone, D. Brown, Y. Y. Oh, N. Donofrio, J. S. Jeong, D. M. Soanes, S. Djonovic, E. Kolomiets, C. Rehmeyer, W. Li, M. Harding, S. Kim, M. H. Lebrun, H. Bohnert, S. Coughlan, J. Butler, S. Calvo, L. J. Ma, R. Nicol, S. Purcell, C. Nusbaum, J. E. Galagan, and B. W. Birren. 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature **434**:980-986.

Donn, G., E. Tischer, J. A. Smith, and H. M. Goodman. 1984. Herbicide-resistant alfalfa cells: an example of gene amplification in plants. J Mol Appl Genet **2**:621-635.

Edwards, R. A., G. J. Olsen, and S. R. Maloy. 2002. Comparative genomics of closely related salmonellae. Trends Microbiol **10**:94-99.

Fabritius, A. L., C. Cvitanich, and H. S. Judelson. 2002. Stage-specific gene expression during sexual development in Phytophthora infestans. Mol Microbiol **45**:1057-1066.

Fiebig, A., R. Kimport, and D. Preuss. 2004. Comparisons of pollen coat genes across *Brassicaceae* species reveal rapid evolution by repeat expansion and diversification. Proc Natl Acad Sci USA **101**:3286-3291.

Field, L. M., A. L. Devonshire, and B. G. Forde. 1988. Molecular evidence that insecticide resistance in peach-potato aphids (Myzus persicae Sulz.) results from amplification of an esterase gene. Biochem J **251**:309-312.

Graves, J. A. M. 2002. The rise and fall of SRY. Trends Genet **18**:259-264.

Kazazian, H. H., Jr. 2004. Mobile elements: drivers of genome evolution. Science **303**:1626-1632.

Kidwell, M. G., and D. R. Lisch. 2001. Perspective: transposable elements, parasitic DNA, and genome evolution. Evolution; Inter J Organic Evol **55**:1-24.

Leblond, P., and B. Decaris. 1994. New insights into the genetic instability of streptomyces. FEMS Microbiol Lett **123**:225-232.

Lengauer, C., K. W. Kinzler, and B. Vogelstein. 1998. Genetic instabilities in human cancers. Nature **396**:643-649.

Lerat, E., P. Capy, and C. Biemont. 2002. Codon usage by transposable elements and their host genes in five species. J Mol Evol **54**:625-637.

Mira, A., L. Klasson, and S. G. Andersson. 2002. Microbial genome evolution: sources of variability. Curr Opin Microbiol **5**:506-512.

Nekrutenko, A., and W.-H. Li. 2001. Transposable elements are found in a large number of human protein-coding genes. Trends Genet **17**:619-621.

Perna, N. T., G. Plunkett, 3rd, V. Burland, B. Mau, J. D. Glasner, D. J. Rose, G. F. Mayhew, P. S. Evans, and a. Gregor et. 2001. Genome sequence of enterohaemorrhagic Escherichia coli O157:H7. Nature **409**:529-533.

Rehmany, A. P., A. Gordon, L. E. Rose, R. L. Allen, M. R. Armstrong, S. C. Whisson, S. Kamoun, B. M. Tyler, P. R. Birch, and J. L. Beynon. 2005. Differential recognition of highly divergent downy mildew avirulence gene alleles by RPP1 resistance genes from two Arabidopsis lines. Plant Cell **17**:1839-1850.

Robert, L. S., E. Foster, M. Lévesque-Lemay, E. Routly, D. Wilkinson, and S. Gleddie. 2004. The potential of limiting transgene flow by modifying the proteins on the surface of pollen grains. PBI Bulletin.

Romero, D., and R. Palacios. 1997. Gene amplification and genomic plasticity in prokaryotes. Annu Rev Genet **31**:91-111.

SanMiguel, P., B. S. Gaut, A. Tikhonov, Y. Nakajima, and J. L. Bennetzen. 1998. The paleontology of intergene retrotransposons of maize. Nature Genet **20**:43-45.

Schwab, M. 1994. Human neuroblastoma: amplification of the N-myc oncogene and loss of a putative cancer-preventing gene on chromosome 1p. Recent Results Cancer Res **135**:7-16.

Shan, W., M. Cao, D. Leung, and B. M. Tyler. 2004. The *Avr1b* locus of *Phytophthora sojae* encodes an elicitor and a regulator required for avirulence on soybean plants carrying resistance gene *Rps1b*. Mol Plant Microbe Interact **17**:394-403.

Stothard, P., and D. Pilgrim. 2003. Sex-determination gene and pathway evolution in nematodes. Bioessays **25**:221-231.

Tamas, I., L. Klasson, B. Canback, A. K. Naslund, A.-S. Eriksson, J. J. Wernegreen, J. P. Sandstrom, N. A. Moran, and S. G. E. Andersson. 2002. 50 million years of genomic stasis in endosymbiotic bacteria. Science **296**:2376-2379.

Topp, C. N., C. X. Zhong, and R. K. Dawe. 2004. Centromere-encoded RNAs are integral components of the maize kinetochore. Proc Natl Acad Sci USA **101**:15986-15991.

van der Lee, T., A. Testa, J. van 't Klooster, G. van den Berg-Velthuis, and F. Govers. 2001. Chromosomal deletion in isolates of *Phytophthora infestans* correlates with virulence on R3, R10, and R11 potato lines. Mol Plant Microbe Interact **14**:1444-1452.

van't Slot, K. A. E., and W. Knogge. 2002. A dual role for microbial pathogen-derived effector proteins in plant disease and resistance. Crit Rev Plant Sci **21**:229-271.

Zsiros, J., M. F. Jebbink, V. V. Lukashov, P. A. Voute, and B. Berkhout. 1999. Biased nucleotide composition of the genome of HERV-K related endogenous retroviruses and its evolutionary implications. J Mol Evol **48**:102-111.

Summary

Phytophthora is a genus comprised of over 65 destructive plant pathogenic species that cause severe damages in agriculture, forestry and natural habitats. Economically important pathogens are *Phytophthora infestans* (causing potato late blight) and *Phytophthora sojae* (causing soybean root and stem rot). A newly discovered species, *Phytophthora ramorum* is destroying oak trees along the west-coast of the USA by causing the Sudden Oak Death syndrome. *Phytophthora* belongs to the oomycetes, which together with phyopathogenic fungi constitute the major groups of plant pathogens. Oomycetes and fungi resemble each other morphologically but belong to different kingdoms, Stramenopila and Fungi, respectively. Convergent evolution has shaped a similar set of weaponry for attacking plants in oomycetes and fungi.

Phytophthora secretes a variety of molecules into the plant apoplast presumably to promote infection. These molecules with a potential role in virulence or pathogenicity are named virulence factors or effectors. Despite their intrinsic virulence functions, effectors may cause failure of infection if plants recognize them and initiate defense responses. Effectors that trigger plant defense responses are called avirulence factors or elicitors. **This thesis** describes the drastic genome rearrangements at an avirulence locus in *P. infestans*, the characterization of effector gene reservoirs in the fully sequenced *P. sojae* and *P. ramorum* genomes, and evolutionary patterns of effector genes.

Potato and *P. infestans* interact according to the gene-for-gene model. As a first step towards unraveling the molecular interaction mechanisms, *P. infestans* avriulence (*Avr*) genes have to be isolated. **Chapter 2** describes a transcriptional profiling strategy to identify avirulence associated transcripts. cDNA-AFLP was used for comparing transcripts in *P. infestans* strains with different virulence phenotypes. A large number of avirulence-associated TDFs (Transcript Derived Fragments) were cloned and sequenced, and EST and genome databases were mined to generate more sequence data. To identify promising candidates, bioinformatic predictions such as the presence of signal peptides, number of cysteine residues and putative virulence functions were used as important selection criteria. **Chapter 3** describes how a combination of transcriptional profiling, and genetic and physical mapping resulted in the characterisation of a complex avirulence locus. Four avirulence associated TDFs were shown to be derived from one gene, named *pi3.4*. Genetic mapping and physical mapping placed *pi3.4* at the *Avr3b-Avr10-Avr11* locus on linkage group VIII. Comparative genomic hybridization (CGH) revealed that this *Avr* locus belongs to one of the six loci in the genome that show copy number variation (CNV). An amplified *pi3.4* gene cluster is present in the avirulent haplotype but absent in the virulent haplotype. Only the 3' half of the *pi3.4* gene is amplified, and this amplification was found to provide diverse modules for assembly of novel full length genes.

Among the proteins secreted by *Phytophthora*, elicitins are produced most abundantly. Elicitins show

elicitor activity by causing a hypersensitive response in tobacco. In **Chapter 4** and **Chapter 5,** the diversity and genome organization of elicitin genes in four *Phytophthora* species are described. Elicitins were found to be encoded by a large complex gene family, and they belong to one of the most highly conserved groups of proteins in the *Phytophthora* genus. Many elicitin (ELI) and elicitin-like (ELL) genes are clustered in the genome. Phylogeny construction indicated that the complex elicitin gene family existed before the ancestral *Phytophthora* species gave rise to the current *Phytophthora* species. Molecular phylogeny classified elicitin family members into 17 different clades, namely 4 ELI clades and 13 ELL clades. Based on expression patterns and bioinformatic predictions, different clades are proposed to possess distinct functions.

The two fully sequenced *Phytophthora* species, *P. sojae* and *P. ramorum*, differ in their genome sizes, sexual behavior and host specificity. Comparative genomics was carried out to gain insight into the evolution of effector genes. In **Chapter 6** the genome organization of potential effector genes is described. Overall co-linearity was found between large genomic regions in *P. sojae* and *P. ramorum*. However, insertions, deletions and expansions revealed some hotspots for genome rearrangements, and such rearrangement hotspots often harbor genes associated with virulence. Contrasting evolutionary patterns were found for neighboring gene families, for example, families encoding extracellular enzymes showed more rearrangements than those encoding intracellular enzymes. Also genes encoding host specific elicitors showed more rearrangements than those encoding general elicitors.

In **Chapter 7** the whole reservoir of secreted proteins present in the proteome was revealed by bioinformatic-predictions. A large secretome comprised of over thousand proteins was found in both *P. sojae* and *P. ramorum.* The majority of secreted protein encoding genes form families and many of them are clustered in the genome. Comparison between secretomes of *P. sojae* and *P. ramorum* showed that different families are evolving at a different pace. The most rapidly evolving families include the surface anchored proteins, mating associated factors and 'RXLR-DEER' proteins. They may play important roles in either host-pathogen interactions or in reproduction.

In **Chapter 8** the base compositions of *P. sojae* and *P. ramorum* were calculated and compared. This information is useful for analyzing basic genome features. The coding regions of *Phytophthora* clearly show high GC3 (3^{rd} position codon usage) and this preference causes codon bias in *Phytophthora* genes. Evolutionary forces such as selective pressure and mutation bias were found to drive codon bias in *Phytophthora.* The higher GC3 value of highly expressed genes in different *Phytophthora* species is indicative for selection pressure, whereas lineage specific GC increase of non-coding regions is reminiscent of whole genome mutation bias. The most widespread groups of mobile elements were retrieved from the genomes and they show a codon bias that is similar to the genes of the host *Phytophthora*.

Finally, the evolutionary implications of the findings presented in this thesis are discussed in **Chapter 9**. For pathogenic organisms, genes encoding effectors are instrumental for interaction with their hosts. As a result, the evolutionary pace of effector genes is in general faster than that of average genes. The results presented in this thesis demonstrate that comparative genomics is a powerful tool to discover these genes and to point out promising candidates responsible for the process of pathogenesis. The ongoing *P. infestans* genome sequencing project will provide new resources for fundamental and applied research, and with the blueprint of *P. infestans* in hand, late bight research will gain momentum.

Samenvatting

Het geslacht *Phytophthora* omvat meer dan 65 verwoestende plantpathogene soorten die ernstige schade toebrengen aan landbouwgewassen en aan planten, struiken en bomen in de natuur. Economisch belangrijke pathogenen zijn onder andere *Phytophthora infestans,* de veroorzaker van de aardappelziekte, en *Phythophthora sojae,* die wortel- en stengelrot op sojaboon veroorzaakt. Een onlangs ontdekte soort, *Phythophthora ramorum*, is verantwoordelijk voor het Sudden Oak Death syndroom en verwoest eikenbomen langs de westkust van de Verenigde Staten. *Phytophthora* behoort tot de oömyceten die, samen met plantpathogene schimmels, de belangrijkste groep plantpathogenen vormen. Morfologisch gezien lijken oömyceten en schimmels op elkaar maar ze behoren tot verschillende rijken, respectievelijk de Stramenopila en de Fungi. Convergente evolutie heeft ertoe geleid dat oömyceten en schimmels een vergelijkbaar wapenarsenaal hebben dat nodig is om planten aan te vallen.

Phytophthora scheidt een breed spectrum aan moleculen uit in de plant apoplast, waarschijnlijk om infectie te bevorderen. Deze moleculen, waarvan verondersteld wordt dat ze een rol spelen bij virulentie of pathogeniteit, worden virulentiefactoren of effectors genoemd. Ondanks hun intrinsieke virulentie functies kunnen effectors ook verantwoordelijk zijn voor het mislukken van een infectie, namelijk als planten de effectors herkennen en vervolgens verdedigingsreacties initiëren. Effectors die verdedigingsreacties bij planten veroorzaken worden avirulentiefactoren of elicitors genoemd. **Dit proefschrift** beschrijft de drastische genoomherschikkingen in een avirulentielocus in *P. infestans*, de karakterisering van het reservoir van effectorgenen in de genomen van *P. sojae* en *P. ramorum* en evolutionaire patronen in effectorgenen.

De interactie tussen aardappel en *P. infestans* verloopt volgens het gen-om-gen model. Als een eerste stap naar het ontrafelen van de moleculaire interactiemechanismen moeten *P. infestans* avirulentiegenen (*Avr* genen) geïsoleerd worden. **Hoofdstuk 2** beschrijft een zogenaamde 'transcriptional profiling' strategie om transcripten, geassocieerd met avirulentie, te identificeren. cDNA-AFLP werd gebruikt om transcripten te vergelijken tussen *P. infestans* stammen met verschillende virulentie fenotypen. Een groot aantal avirulentie-geassocieerde TDF's (Transcript Derived Fragments) werd gekloneerd en gesequenced, en vervolgens werden EST- en genoom-databanken doorgespit om meer sequentiegegevens te verkrijgen. Om veelbelovende kandidaten te identificeren werden selectiecriteria gebruikt die met behulp van bioinformatica getoetst werden, zoals het vóórkomen van een signaalpeptide, het aantal cysteïne residuen en mogelijke virulentiefuncties. **Hoofdstuk 3** beschrijft hoe een combinatie van 'transcriptional profiling' en genetische en fysische kartering leidde tot de karakterisering van een complex avirulentie locus. Vier avirulentie-geassocieerde TDF's bleken afgeleid te zijn van één gen, *pi3.4*. Genetische en fysische kartering plaatste *pi3.4* op het *Avr3b-Avr10-Avr11* locus op koppelingsgroep VIII. 'Comparative Genomic

Hybridization' (CGH) liet zien dat dit *Avr* locus behoort tot een van de zes loci in het genoom waarvan het aantal kopieën variabel is ('copy number variation' - CNV). Een geamplificeerd *pi3.4* gencluster is aanwezig in het avirulente haplotype maar afwezig in het virulente haplotype. Alleen de 3' helft van het *pi3.4* gen is geamplificeerd, en deze amplificatie bleek variabele modules op te leveren die mogelijk gebruikt worden om nieuwe volledige genen samen te stellen.

Van de eiwitten die worden uitgescheiden door *Phytophthora* worden elicitines het meest overvloedig aangemaakt. Elicitines hebben elicitor-activiteit; ze veroorzaken een hypersensitieve reactie in tabaksplanten. In **Hoofdstuk 4** en **Hoofdstuk 5** wordt de diversiteit en genoomorganisatie van elicitinegenen in vier *Phytophthora* soorten beschreven. Het bleek dat elicitines gecodeerd worden door een grote en complexe genfamilie, en dat ze behoren tot een van de meest geconserveerde eiwitgroepen in het *Phytophthora* geslacht. Vele elicitine (ELI) en elicitine-achtige (ELL) genen komen geclusterd voor in het genoom. Fylogenetische analyse gaf aan dat de complexe elicitine genfamilie al bestond voordat uit de oer-*Phytophthora* de huidige *Phytophthora* soorten ontstonden. Met moleculaire fylogenie werden leden van de elicitinefamilie geclassificeerd in 17 verschillende groepen, te weten 4 ELI groepen en 13 ELL groepen. Uit expressiepatronen en voorspellingen met behulp van bioinformatica, kon worden afgeleid dat verschillende groepen verschillende functies uitoefenen.

De twee *Phytophthora* soorten waarvan de genoomsequentie volledig bekend is, *P. sojae* en *P. ramorum*, verschillen in hun genoomgroottes, seksueel gedrag en gastheerspecificiteit. Vergelijkende genoomanalyse werd uitgevoerd om inzicht te verkrijgen in de evolutie van effectorgenen. In **Hoofdstuk 6** wordt de genoom organisatie van potentiële effectorgenen beschreven. Algemene co-lineariteit werd gevonden tussen grote genoom gebieden van *P. sojae* en *P. ramorum*. Desalniettemin brachten inserties, deleties en expansies een aantal 'hotspots' voor genoomherschikkingen aan het licht en zulke hotspots bleken vaak virulentie-geassocieerde genen te herbergen. Contrasterende evolutionaire patronen werden gevonden voor naburige genfamilies waarbij, bijvoorbeeld, families die coderen voor extracellulaire enzymen meer herschikkingen vertoonden dan families die coderen voor intracellulaire enzymen. Ook vertoonden genen die voor gastheerspecifieke elicitoren coderen meer herschikkingen dan genen die voor algemene elicitoren coderen.

In **Hoofdstuk 7** wordt het hele reservoir aan uitgescheiden eiwitten aanwezig in het proteoom geanalyseerd met behulp van bioinformatica. Zowel *P. sojae* als *P. ramorum,* beschikt over een groot secretoom bestaande uit meer dan duizend eiwitten. De meeste genen die voor uitgescheiden eiwitten coderen kunnen gegroepeerd worden in families, en vele komen in het genoom geclusterd voor. Vergelijking van de secretomen van *P. sojae* en *P. ramorum* liet zien dat verschillende families met een verschillende snelheid evolueren. Tot de snelst evoluerende families behoren de eiwitten die aan het celoppervlak verankerd zijn, factoren geassocieerd met paring en "RXLR-DEER" eiwitten. Zij kunnen een belangrijke rol spelen in gastheer-pathogeen interacties of in reproductie.

In **Hoofdstuk 8** zijn de base samenstellingen van de *P. sojae* en *P. ramorum* genomen berekend en vergeleken. Deze informatie is waardevol bij het analyseren van basale eigenschappen van het genoom. De coderende gebieden laten duidelijk een hoog GC3 gehalte zien (GC codon gebruik op de 3^{e} positie), en deze voorkeur voor het gebruik van bepaalde codons in *Phytophthora* genen heeft een zogenaamde 'codon bias' tot gevolg. Gevonden werd dat evolutionaire krachten zoals selectiedruk en tendens tot mutaties de 'codon bias' in *Phytophthora* aanzwengelen. Het hogere GC3 gehalte in *Phytophthora* genen die hoog tot expressie komen duidt op selectiedruk, terwijl toename van het GC gehalte in niet-coderende gebieden in de ene *Phytophthora* soort ten opzichte van de andere, wijst op een mutatie tendens in het hele genoom ('whole genome mutation bias'). De meest wijdverspreide groepen van transposons werden uit het genoom van de beide *Phytophthora* soorten gefilterd en geanalyseerd, en deze vertonen een 'codon bias' die vergelijkbaar is met die van de genen van de gastheer *Phytophthora*.

Tenslotte worden in **Hoofdstuk 9** de evolutionaire implicaties van de bevindingen die in dit proefschrift zijn beschreven besproken. Voor pathogene organismen zijn effectorgenen van belang voor de interactie met hun gastheren. Dit heeft tot gevolg dat effectorgenen over het algemeen sneller evolueren dan de meeste andere genen. De resultaten gepresenteerd in dit proefschrift laten zien dat vergelijkende genoom analyse een krachtig instrument is om deze genen te ontdekken, en om veelbelovende kandidaten verantwoordelijk voor pathogenese aan te wijzen. Op dit moment wordt gewerkt aan het bepalen van de volledige DNA sequentie van het genoom van *P. infestans* en eind 2006 zal de genoomsequentie beschikbaar zijn. Dit geeft een nieuwe impuls aan fundamentele en toegepaste wetenschappers, en met de blauwdruk van *P. infestans* tot onze beschikking zal het aardappelziekte onderzoek aan momentum winnen.

Acknowledgement

Francine Govers has been my mentor. Her devotion to science has shaped my science career. Her extraordinary care is a luxury for a PhD student. At the same time, she always permits me the freedom to explore interesting terrains in science. I remain amazed by her from time to time, probably because so much involvement and dedication do not exist in my own personality.

My office in Wageningen remains a special place in my mind: fruits and nuts, dusts and stains, overgrown plants and insect corpses. From there, everyday I wandered into the world of genome sciences. For 5 years, my good friend Rob Weide shared the office and the complaint from cleaners with me. We nearly buried ourselves with all sorts of messes and we shared many memorable talks about politics, culture and the silliness of life.

Wageningen is an energetic place for me because my friends Pieter van Poppel and Klaas Bouwmeester live there. Work is intrinsically coupled with fun because there is always a coffee break, a casual chat, a nonsense game, an adventure to town or a cynical conversation around the corner. And there are always endless rock concerts, street theatres, cafes and pubs to be visited. Together with other friends Peter van Esse, Ursula Ellendorff and Melvin Bolton, we made the PhD period into a time of celebration.

The extreme friendliness of the research group in Wageningen is very special for me; Peter van de Vondervoort and Harold Meijer are certainly the ones to count on if you have a computer crash, an experiment failure or a twisted ankle.

I would like to thank my promoter Pierre de Wit for his guidance. And I appreciate the time spent together with Ha Tran Thi Thu, Ilona Kars, Ramin Roohparvar, Marco Kruijt, Martijn Staats, Emilie Fradin and Ronnie de Jonge. I would like to thank Sander Schouten, Lute-Harm Zwiers, Peter van Baarlen, Kiona Harbers, Jun Guo, Wubei Dong, Wilco Ligterink, Iziah Sama, Lars Kamphuis, Jacqueline Gerdsen and Andrė van't Slot for their friendship.

My parents keep watching me closely from China. My dad is always on the cutting edge of communication technologies, email, internet phone, web-cam conversation etc. I suspect his ultimate goal is checking up on me as frequently as possible. My brief emails hardly gave any detail of my life in Europe, but mysteriously my parents know the exact amount of snow fall in the Netherlands, the precise flight schedule of my travel, the impact factor of my science publications and the slightest insecurity in my life. It is rather strange that their affection seems to be something so blind, tolerant and unlimited.

Each time at the end of a conversation on phone, my dad can always find a way to remind my partner Arne Zijlstra to take care of me. However I have never got such a reminder to watch out for Arne's well-being - unfair. I always secretly consider that Arne has foolishly chosen me to be his partner. And I am reluctant to admit that his broad knowledge and insane openness have brought perspective and depth in my life. In general, Arne is entirely blind to race, color, nationality or social status. To me, he is tolerant of nearly everything, sadness, despair, anger, or impatience. I only hope my presence brings enjoyment in his life.

Rays Jiang
Jan-2006

Curriculum Vitae

Rays Jiang was born on July 6, 1975 in a communist-setup research institute in Kunming, China. She has been fascinated by life science since her childhood. At Yunnan University (China), she obtained her BSc as Biotechnology major in 1997. In 1998, she studied medical microbiology at Beijing Medical Union University (China). At Wageningen University (the Netherlands), she obtained her MSc as Biotechnology major in 2000. From 2000 to 2005, she carried out her PhD research at the laboratory of Phytopathology, Wageningen University. This thesis summarizes the research results of her PhD program in deciphering microbial genomes. In 2006, with a Netherland Genomics Initiative fellowship, she will explore pathogen genomes in depth at the Virginia Bioinformatics Institute and the Broad Institute in the US.

Training and supervision plan of the Graduate School Experimental Plant Sciences (EPS)

1. Participation in postgraduate courses and workshps

a) EPS Autumn school 'Interaction between plants and attacking organisms' (2000)
b) Course 'Winter school bioinformatics' (2000)
c) Course 'Safe handling with radioactive materials and sources' (2001)
d) Course 'AFLP markers in plant systematics and breeding' (2001)
e) EPS Autumn school 'Disease resistance in plants' (2002)

2. Participation in international meetings

a) Durable disease resistance symposium. Ede, The Netherlands (2000) poster presentation
b) CAAS-WUR autumn workshop; Satellite meeting *Phytophthora*. Beijing, China (2003) oral presentation
c) British Society for Plant Pathology (BSPP) presidential meeting - Plant pathogen genomics. Nottingham, UK (2003) poster presentation
d) *Phytophthora* genome sequence annotation jamboree. DOE Joint Genome Institute (JGI), Walnut Creek, California, USA (2004) Actively contributed to the genome annotation of *Phytophthora*
e) 23rd Fungal Genetics Conference at Asilomar, Pacific Grove, California, USA (2005) poster presentation and oral presentation

3. Participation in national meetings

a) Experimental Plant Science theme 2 Symposium, Amsterdam, The Netherlands (2003) oral presentation
b) CBS/Wageningen Phytopathology symposium, Utrecht, The Netherlands (2003) oral presentation
c) Center for BioSystems Genomics (CBSG) Symposium, Wageningen, The Netherlands. (2004) oral presentation

The research described in this thesis was performed at the laboratory of Phytopathology of Wageningen University and was financially supported by NWO-Aspasia grant 015.000.057.

Paintings used in this thesis: Rays Jiang