



# **Exploring Evolution and Biology of Oomycetes: Integrative and Comparative Genomics**





Michael F Seidl (2013)  
Exploring Evolution and Biology of Oomycetes: Integrative and Comparative Genomics  
PhD thesis, Utrecht University  
Printed by: Proefschriftmaken.nl || Uitgeverij BOXPress  
ISBN: 978-90-393-6929-6





# **Exploring Evolution and Biology of Oomycetes: Integrative and Comparative Genomics**

Evolutie en Biologie van Oömyceten: Data Integratie en  
Vergelijkende Genoomanalyse  
(met een samenvatting in het Nederlands)

Evolution und Biologie von Oömyceten: Daten Integration und  
Vergleichende Genomanalyse  
(mit einer Zusammenfassung in deutscher Sprache)

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit Utrecht op gezag van de rector magnificus, prof.dr. G.J. van der Zwaan, ingevolge het besluit van het college voor promoties in het openbaar te verdedigen op maandag 6 mei 2013 des middags te 2.30 uur

door

Michael Franziskus Seidl

geboren op 13 oktober 1982 te Darmstadt, Duitsland



Promotoren: Prof.dr. P. Hogeweg  
Prof.dr. F.P.M. Govers

Co-promotoren: Dr. B. Snel  
Dr. A.F.J.M. van den Ackerveken

This thesis was accomplished with financial support from the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research.



# CONTENTS

<b>General Introduction</b>	<b>7</b>
Introduction . . . . .	8
The Group of Oomycetes . . . . .	9
Comparative Genomics in Oomycetes . . . . .	13
References . . . . .	22
<b>A Domain-Centric Analysis of Oomycete Plant Pathogen Genomes Reveals Unique Protein Organization</b>	<b>29</b>
Abstract . . . . .	30
Introduction . . . . .	31
Results . . . . .	32
Discussion. . . . .	45
Material & Methods . . . . .	50
Acknowledgments . . . . .	53
Supplementary Material. . . . .	54
References . . . . .	56
<b>Reconstruction of Oomycete Genome Evolution Identifies Differences in Evolutionary Trajectories Leading to Present-Day Large Gene Families</b>	<b>61</b>
Abstract . . . . .	62
Introduction . . . . .	63
Material & Methods . . . . .	64
Results . . . . .	65
Discussion. . . . .	72
Conclusions. . . . .	78
Acknowledgments . . . . .	78
Supplementary Material. . . . .	78
References . . . . .	91
<b>Bioinformatic Inference of Specific and General Transcription Factor Binding Sites in the Plant Pathogen <i>Phytophthora infestans</i></b>	<b>97</b>
Abstract . . . . .	98
Introduction . . . . .	99
Material & Methods . . . . .	100
Results . . . . .	104
Discussion. . . . .	112
Acknowledgments . . . . .	116
Supplementary Material. . . . .	116
References . . . . .	117
<b>A Predicted Functional Gene Network for the Plant Pathogen <i>Phytophthora infestans</i> as a Framework for Genomic Biology</b>	<b>121</b>
Abstract . . . . .	122



Introduction . . . . .	123
Results & Discussion . . . . .	124
Conclusions . . . . .	136
Material & Methods . . . . .	137
Acknowledgments . . . . .	141
Supplementary Material. . . . .	142
References . . . . .	145
<b>Summarizing Discussion</b>	<b>149</b>
Introduction . . . . .	150
Oomycete Genome Sequences . . . . .	150
Genome Annotation Quality Affects Comparative Analysis . . . . .	153
Complex Omics Data Accessibility for Oomycete Biologists . . . . .	156
Comparative Genomics in Oomycetes as a Blueprint for Other Taxa . . . . .	157
Concluding Remarks . . . . .	158
References . . . . .	158
<b>Summary</b>	<b>161</b>
<b>Samenvatting</b>	<b>163</b>
<b>Zusammenfassung</b>	<b>165</b>
<b>Curriculum Vitae</b>	<b>167</b>
<b>List of Publications</b>	<b>169</b>
<b>Acknowledgments</b>	<b>171</b>





# General Introduction

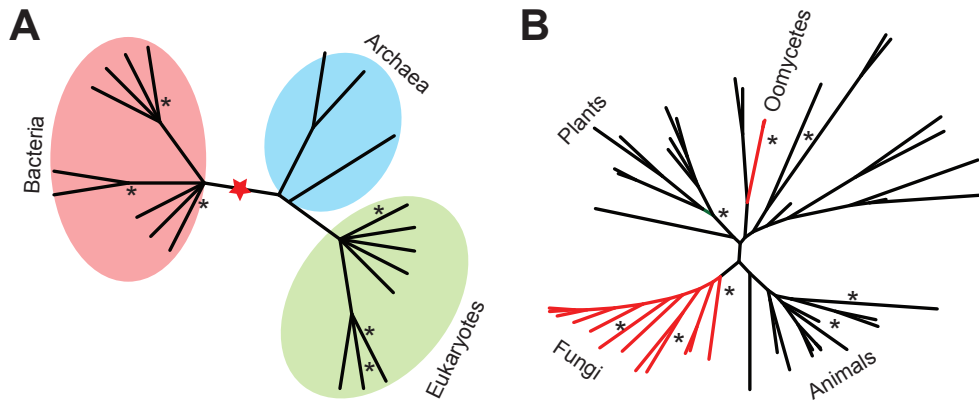
1



## INTRODUCTION

Domestication of animals and plants played a major role in the rise and establishment of human civilizations. People were always dependent on a continuous supply of food, however the tremendous growth in human population makes securing global food resources one of the crucial goals of current and future generations (Baker et al. 1997). Diseases that destroyed livestock and crops have had catastrophic effects causing starvation and huge economic losses (Baker et al. 1997). The ability to successfully infect economically and ecologically important species has evolved several times independently within the tree of life and disease-causing organisms include diverse groups such as bacteria, fungi and nematodes (Figure 1-1A).

A fascinating collection of these pathogens is united in the taxonomic class of oomycetes (Figure 1-1B and Figure 1-2). Since the first description of one of its members in the 19<sup>th</sup> century, these eukaryotic organisms have been in the research focus thereby initiating the scientific field of 'phytopathology'. Even though some members developed to model organisms for host-pathogen interactions, experimental work with this group of organisms in general has shown to be challenging (Govers & Gijzen 2006). Consequently, comprehensive characterization of the (molecular) biology of oomycetes is lagging behind other eukaryotic model organisms. Six years ago, oomycetes entered the genomics era by collaborative sequencing efforts that led to the publication of whole genome sequences of two of its members (Tyler et al. 2006). Since then, many more oomycete genome sequences together with accompanying transcriptomics data became available. To retrieve biological insights from the wealth of these complex data



**Figure 1-1 Independent evolution of pathogenicity in the tree of life.**

(A) Pathogenicity (indicated with asterisk) evolved (at least) in two of the three domains of life (possible root indicated with red star); in bacteria, e.g. proteobacteria such as *Pseudomonas syringae*, and in eukaryotes. (B) The ability to successfully infect a host (indicated with asterisk) has evolved several times independently in eukaryotes, e.g. in morphologically diverse organisms such as fungi and nematodes. Even though oomycetes share apparent similarities with fungi, they are phylogenetically unrelated. Whereas fungi are closely related to metazoans, phylogenetic analyses unambiguously group oomycetes together with non-pathogenic algae in the group Stramenopiles.



and to overcome the experimental limitations inherent to oomycetes, integrative and comparative bioinformatic approaches are crucial. Comparative genomics, especially within oomycetes, but also with fungi and other eukaryotes, is needed to elucidate the genome evolution of these organisms, to identify specific features that characterize these species and to shed light on the function of as yet uncharacterized genes thereby complementing traditional experimental work. For the research described in this thesis, we have exploited the growing number of omics data in oomycetes and applied comparative genomic and integrative bioinformatic approaches to study their biology and evolution.

In this introductory chapter (chapter 1), we briefly touch upon the current knowledge of the phylogeny and biology of oomycetes and describe the use of comparative genomics as a crucial tool to elucidate genome evolution and function. We then introduce the experimental chapters of this thesis, the first two of which focus on the evolution of genes and their encoded proteins. In chapter 2 we studied protein domains, the building blocks of proteins, and in chapter 3 the evolution of genomes, more specifically the gene content, in oomycetes. The next chapter (chapter 4) centers on the regulation of gene expression by identifying *cis*-regulatory DNA motifs and chapter 5 on large-scale functional associations between genes and/or proteins. The analyses presented in this thesis highlight the merit of comparative genomics and lead to the establishment of (testable) hypotheses on the evolution, biology and the function of as yet uncharacterized gene products in oomycetes. This is summarized in the last chapter (chapter 6), in which we discuss recurring themes, especially in the light of related work, and give an outlook how comparative genomics will further aid oomycete research.

## THE GROUP OF OOMYCETES

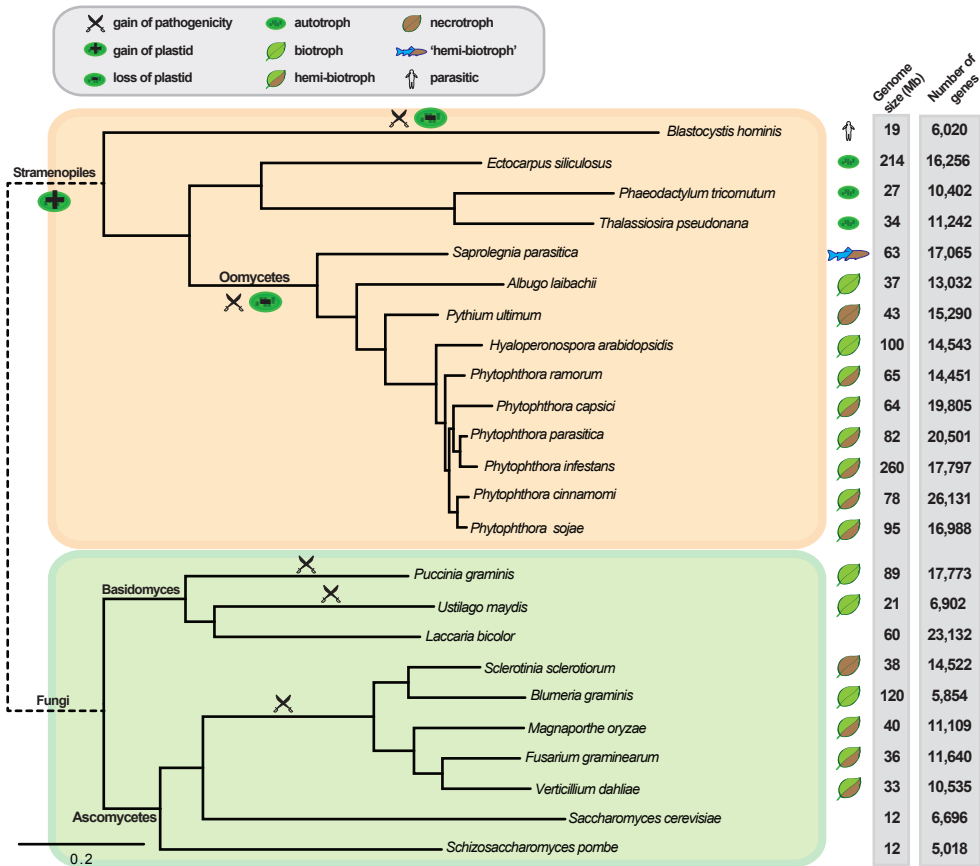
### Oomycetes - an Intriguing Ensemble of Cosmopolitan Organisms

Oomycetes are an intriguing ensemble of more than 1,500 species comprising saprophytes - organisms that feed on dead organic material - and pathogens of animals as well as plants (Govers & Gijzen 2006; Thines & Kamoun 2010). Historically, oomycetes or 'water molds' have received considerable attention since a member of the genus *Phytophthora* significantly impacted human history (Fisher et al. 2012). *Phytophthora infestans* is the causal agent of potato late blight, a disease that led to starvation and death of millions of people during the Irish potato famine and enforced one of the largest immigration waves in human history reducing the population in Ireland from 8.2 to 5.8 million in 20 years (Duncan 1999). The name *Phytophthora*, Greek for 'plant destroyer', was coined in the 19<sup>th</sup> century by the German mycologist Heinrich Anton de Bary (Large 1940). De Bary studied infected potato plants and showed that *P. infestans* was the cause of late blight. Even today, *P. infestans* is a constant threat to global potato and tomato production, estimated to cause losses of up to US\$3 billion annually (Duncan 1999; Pennisi 2010).

Since the 19<sup>th</sup> century and the work of de Bary, *P. infestans* has always been a catalyst for advances in studies of plant pathogens and the interaction with their hosts (Judelson & Blanco 2005). However, *P. infestans* is only one of many pathogens within the oomycetes: the most basal branching oomycetes are parasites of marine algae and diatoms such as the brown algae pathogen *Eurychasma dicksonii* (Grenville-Briggs et al. 2011). Therefore oomycetes most likely evolved from an aquatic environment that they shared with their hosts (Thines & Kamoun 2010). Parasitism on (land) plants, as seen for *P. infestans*, has most likely evolved several times independently in major groups of oomycetes, most prominently within the group of Peronosporales (reviewed in Thines & Kamoun 2010), which besides *Phytophthora* also contains the downy mildews. The genus *Phytophthora* comprises over 100 described species (Kroon et al. 2012), several of which have severe impact on agriculture and ecosystems (for some examples see Figure 1-2). Most *Phytophthora* species have a hemi-biotrophic lifestyle characterized by an initial biotrophic phase that subsequently switches to a necrotrophic phase at the end of the infection cycle. Members of this genus display a huge diversity in their host specificity ranging from the soybean pathogen *Phytophthora sojae* that infects only one host to *Phytophthora ramorum* that is a severe threat to a broader variety of more than 100 plant species (Grunwald et al. 2008). Whereas some oomycetes exhibit similar host ranges and lifestyles compared to *Phytophthora*, others show distinct lifestyles or colonize hosts other than plants. The downy mildew *Hyaloperonospora arabidopsidis* as well as the white rust *Albugo laibachii* are two examples of exclusively *Arabidopsis thaliana* parasitizing oomycetes that evolved their obligate biotrophy lifestyle independently (Figure 1-2). In contrast, *Pythium* species such as *Pythium ultimum* are wide-range saprophytes and opportunistic phytopathogens mainly exhibiting a necrotrophic lifestyle. *Saprolegnia parasitica* is a fish pathogen and a growing problem in aquaculture threatening fish production as one of the emerging major sources of animal protein for human nutrition (Jiang et al. 2013). Exhibiting a wide variety of different lifestyles and occupying diverse ecological niches all around the world makes oomycetes, next to fungi, one of the most prominent and diverse group of eukaryotic pathogens.

### Oomycetes and their Phylogenetic Relationship to Other (Pathogenic) Taxa

The ability to successfully colonize a host and the distinct mode of this relationship, i.e. if the pathogen exhibits a biotrophic or necrotrophic lifestyle, evolved independently within oomycetes and fungi (Figure 1-1B). For many decades, it was assumed that oomycetes and fungi are phylogenetically related because they share several notable similarities ranging from their osmotrophic feeding behavior, mode of reproduction, to hyphal growth (Money 1998; Money et al. 2004). Even though de Bary had already observed some morphological differences between fungi and oomycetes (Large 1940), it was biochemistry that found distinct differences, e.g. in the constitution of the cell wall (Werner et al. 2002; Latijnhouwers et al. 2003). Moreover, molecular phylogenetics placed oomycetes unambiguously together with marine non-pathogenic autotrophic organisms such as the photosynthetic brown algae and diatoms within the phylum Stramenopiles (Baldauf et al. 2000) (Figure 1-2). Recent molecular data suggest that oomycetes as a group are relatively old; they likely diverged from their non-pathogen-



**Figure 1-2 Phylogenetic relationship between Stramenopiles and fungi.**

Predicted phylogenetic relationship between a (non-exhaustive) selection of sequenced fungal and oomycete pathogens together with some of their non-pathogenic relatives. The phylogeny is based on 65 core genes and corroborates the established dichotomies with high support (>95% bootstrap). The possible acquisition and subsequent independent loss of a plastid as well as the independent gain of pathogenicity mechanisms is displayed on the most likely branches for the given selection of species. The sequenced genomes display a huge diversity in size, ranging from 12 Mb in fungi up to 260 Mb in oomycetes, and in number of predicted genes.

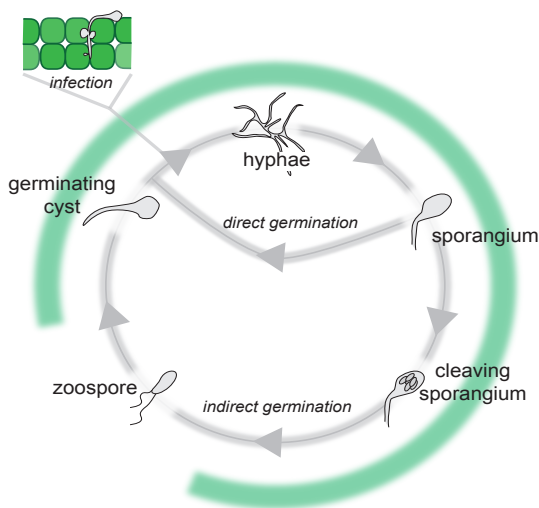
ic relatives around 1,000 million years ago (Bhattacharya et al. 2009) and are slightly younger than fungi but of similar age as animals.

Whereas the common ancestry of Stramenopiles is generally accepted, their relation with other species is still controversial (Baurain et al. 2010). According to the Chromalveolate hypothesis, Stramenopiles are grouped with other chlorophyll-c containing lineages such as Cryptophytes, Alveolates, and Haptophytes into a single monophyletic supergroup (Cavalier-Smith 1999; Keeling 2009). This grouping has been justified under the assumption that the last common ancestor of these species acquired the

plastid via a single secondary endosymbiosis event and inherited it strictly vertically. Subsequently, oomycetes and other plastid-lacking species independently lost these plastids, a hypothesis that seems to be supported by the identification of remnants of plastid-derived genes in their genomes (Andersson & Roger 2002; Tyler et al. 2006; Maruyama et al. 2009). The competing serial eukaryotic-eukaryotic endosymbiosis hypothesis, however, proposes the independent acquisition of plastids through higher order endosymbiosis and, depending on the precise point of acquisition, no or less secondary losses are needed to explain the obvious lack of plastids in several lineages (Cavalier-Smith et al. 1994; Archibald 2009; Baurain et al. 2010).

### A Snapshot of the Biology and Pathology of Oomycetes

Oomycetes are tightly associated with their respective hosts because most of the distinct life stages take place and develop during host colonization, exemplified by the life cycle of *Phytophthora* (Figure 1-3). After a biflagellated zoospore, the free-swimming asexual spore, detects host tissue by chemotaxis, as has been shown for some *Phytophthora* species such as *P. sojae*, it adheres and encysts (Judelson & Blanco 2005). Subsequently, this cyst starts to germinate, forming an appressorium that allows a tight contact to the host cell. This serves as a starting point for hyphal growth subsequent to, in the case of plant pathogenic oomycetes, biochemical and/or mechanical breakdown of the cell wall (Randall et al. 2005; Tyler et al. 2006; Haas et al. 2009). Hyphae spread through the living host tissue, forming specialized infection and feeding structures called haustoria (Hohl & Stössel 1976). Similar to fungi, oomycetes release proteins that interfere with or suppress immune responses in the host, such as the oomycete specific RXLR and Crinkler effectors, and enzymatically degrade organic compounds such as glycoside hydrolases for nutrition uptake (reviewed by Stassen & Van den Ackerveken 2011). After a certain time of vegetative growth, typically 4-7 days for *P. infestans*, new sporangia are formed to allow dispersal and colonization of other hosts. Sporangia



**Figure 1-3 The life cycle of *Phytophthora***

The (asexual) life cycle of *Phytophthora* is mainly associated with the colonization of the host. Phases on or within the host are indicated in green. The infection of the host is initiated directly by a germinating sporangium or indirectly by a germinating cyst.

emerge from the termini of hyphae and can either germinate and infect directly or develop in a zoosporangium that releases zoospores that initiate colonization and infection of a new host. In *P. infestans* and other species sporangia can be blown away to disperse over large distances.

Historically, reaching back to de Bary, these processes have been studied microscopically by observations of morphological alteration such as zoospore formation (Large 1940). However, this has been gradually substituted by studies applying in-depth molecular and biochemical tools such as DNA transformation (Judelson et al. 1991) and gene silencing (Kamoun et al. 1998; van West et al. 1999, also reviewed by e.g. Birch & Whisson 2001; Kamoun 2003). However, these techniques have limitations (Govers & Gijzen 2006) and focus on the function of a single gene and its role in the complex biological/developmental processes, e.g. the contribution of *Pks1* (Xiang et al. 2009) or *Pigpb1* (Latijnhouwers & Govers 2003) in spore formation. In the last few years, especially since the availability of the genome sequences of the first oomycetes (Tyler et al. 2006) and the application of bioinformatics, the research on all aspects of oomycete biology has noticeably gained velocity, accelerating molecular genetics and molecular biology and thereby changing the focus to a more system-wide understanding of oomycete biology.

## COMPARATIVE GENOMICS IN OOMYCETES

### Oomycetes Entering the (Gen-) Omics Era

Since the late 90's of the last century, biology has undergone a tremendous transition facilitated by the availability of technologies to sequence full genomes. The first sequenced organism was the parasitic bacterium *Haemophilus influenzae* (Fleischmann et al. 1995), but the number of fully sequenced genomes was increasing continuously, reaching milestones with the availability of the human genome (Venter et al. 2001; Lander et al. 2001) and genomes of eukaryotic model organisms such as *Saccharomyces cerevisiae* (Goffeau et al. 1996) and *A. thaliana* (Arabidopsis Genome Initiative 2000). This increase in the number of sequenced genomes - 3,707 so far - is even further accelerated with the emergence of new technologies such as next generation sequencing (NGS) (information gained from Genomes OnLine Database on 22.10.2012).

Seven years after the release of the *H. influenzae* genome sequence, the genome sequences of two pathovars of the plant pathogenic bacterium *Xanthomonas* became available (da Silva et al. 2002). Rapidly afterwards, the first eukaryotic plant pathogen was sequenced, the hemi-biotrophic fungus *Magnaporthe grisea* (Dean et al. 2005), and just a year later, the genome sequences of two hemi-biotrophic oomycete plant pathogens, *P. sojae* and *P. ramorum*, were published (Tyler et al. 2006). The availability of these genomes marked a milestone in the emergence of large-scale descriptive data in oomycetes, initiated by the sequencing of expressed sequence tags several years

earlier (Kamoun et al. 1999; Qutob et al. 2000; Randall et al. 2005; Torto-Alalibo et al. 2005). Subsequently, additional omics data including transcriptomics in the form of microarrays (Judelson et al. 2008; Haas et al. 2009), RNA-Seq (Lévesque et al. 2010; Ye et al. 2011; Links et al. 2011; Kunjeti et al. 2012; Savory et al. 2012; Jiang et al. 2013), proteomics (Savidor et al. 2006) and phospho-proteomics (Resjö et al. unpublished) became available. These data are accompanied by many more sequenced genomes representing distinct taxa within oomycetes, e.g. Peronosporales and Pythiales, different biological lifestyles, e.g. biotrophic downy mildews and hemi-biotrophic *Phytophthora*, and host ranges, e.g. the animal pathogen *S. parasitica* (Figure 1-2), and complemented with recently available NGS data (e.g. Grenville-Briggs et al. 2011; Ye et al. 2011; Stassen et al. 2012). The wealth of these diverse data opens a treasure trove. Below we will further discuss initial analyses, and their influence and contribution to our understanding of the biology of oomycetes.

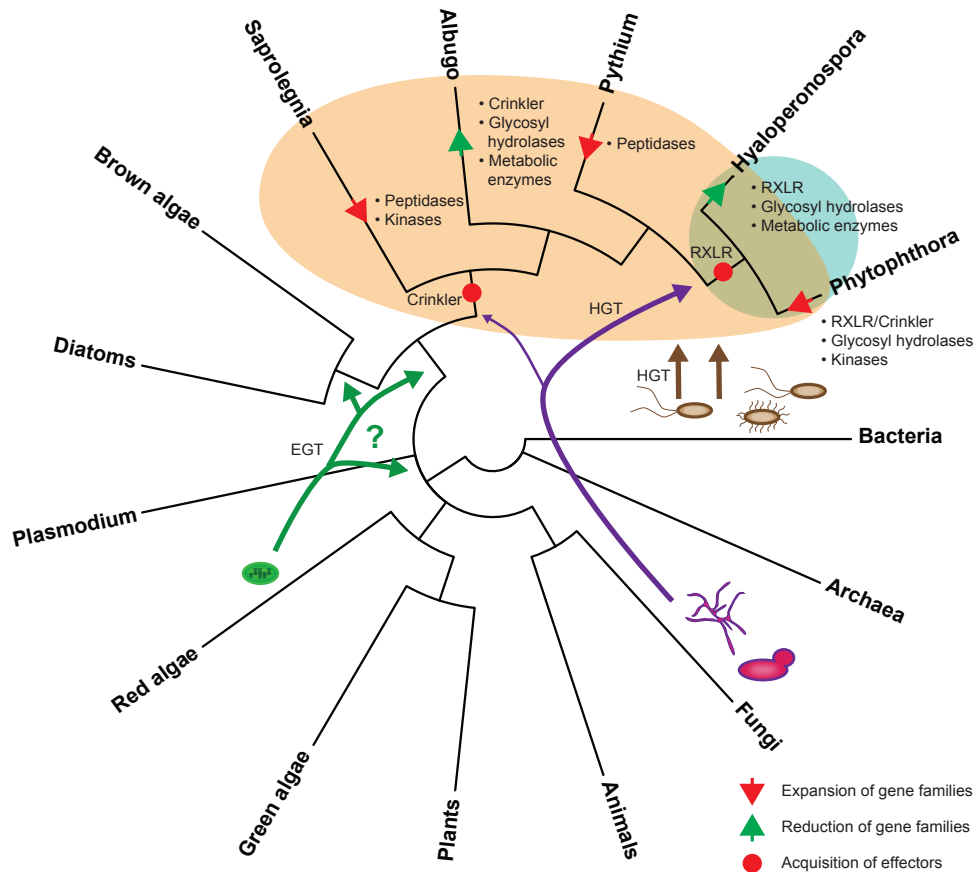
### **Comparative Genomics Enables a Comprehensive Study of Genomes and their Content**

The increasing amount of these complex data sets and their integration enabled comparative studies between two or more oomycetes, their non-pathogenic relatives, and other pathogenic eukaryotes such as fungi or eukaryotes. In general, these comparative analyses provide substantially more information compared to analyses merely based on a single genome because they allow the identification of shared, sometimes called core, and specific genes, possibly linking those to the biology.

Comparative genomics traditionally focused on approaches to gradually close the gap between the increasing amount of sequence data and the lack of experimental functional characterization for the majority of genes (Bork et al. 1998). This has also been a major focus in oomycetes. Recently, however, the focus has moved to more fundamental questions related to genome evolution and physiology of oomycetes and more specifically to the emergence of similarities and differences that account for biological characteristics such as lifestyle and infection (see e.g. Judelson 2012). Comparative genomics in oomycetes (and other species) therefore complements traditional experimental work and will lead to the establishment of (testable) hypotheses on the evolution, biology, and specifically, on the function of as yet uncharacterized gene products.

### **Comparative Genomics in Oomycetes – Application and Initial Insights**

Numerous comparative studies in bacteria and eukaryotes have indicated that multiple and opposing processes shape the genome structure and gene content of genomes. These processes include genome expansions altering the non-coding and/or coding regions of the genome, genome reductions and horizontal gene transfer (e.g. Koonin 2005). The availability of a growing number of sequenced oomycete genomes enables us to approach and answer fundamental questions on the evolution of genome structure and content.



**Figure 1-4 Evolutionary events altering the gene content of oomycetes.**

Acquisition of genes via horizontal gene transfer (HGT) and acquisition of a plastid via endosymbiosis and subsequent endosymbiotic gene transfer (EGT) are indicated. Whereas the transfer from bacteria is found in all species within oomycetes (orange group), HGT from fungi has mainly been observed in the last common ancestor of Peronosporales (turquoise group). The acquisitions of effector genes, expansion and reduction of gene families discussed in the main text are indicated on the respective branches.

Oomycetes display a tremendous variation in their genome size ranging from 37 Mb in *A. laibachii* (Kemen et al. 2011) up to 260 Mb in *P. infestans* (Haas et al. 2009). The expansion of genome sizes in certain taxa seems to be associated with the presence of transposable elements. The genome of *P. infestans* has 74% repetitive DNA content, whereas the percentage in closely related *Phytophthora* such as *P. ramorum* is considerably lower (28%) (Haas et al. 2009). Differences in genome sizes are not limited to oomycetes; they occur also in other lineages. The non-pathogenic brown alga *Ectocarpus siliculosus*, a sister taxon to the pathogenic oomycetes, has also a considerably larger genome (214 Mb) compared to most of its close relatives. This genome expansion, in contrast to that in oomycetes and in the pathogenic fungus *Blumeria graminis* (Spanu

et al. 2010), does not seem to be repeat driven (only 23% of the genome contains repeated regions), but might be influenced by an unusually high number of introns (up to 7 per gene) and long 3' untranslated regions (Cock et al. 2010). The occurrence of a considerable number of repetitive elements in the genomes of *Phytophthora*, especially in *P. infestans*, most likely caused the evolution of a bipartite genome that is characterized by (i) stable gene-rich regions (small intergenic distance) that harbor the core genes and (ii) dynamic gene-poor, transposon and repeat-rich regions that are enriched for fast evolving gene families encoding proteins with predicted functions in pathogenicity (Haas et al. 2009; Raffaele et al. 2010). The mechanism underlying this partitioning has not yet been fully understood, but has been attributed to the high abundance of transposable elements in the genomes and in particular in gene-sparse regions. Interestingly, it has already been known for a while that eukaryotic genomes have a non-random gene order, mainly clustering co-expressed genes and genes encoding proteins involved in the same process (reviewed by Hurst et al. 2004). Crombach and Hogeweg (2007) showed that random genomes possessing transposable elements to mediate chromosomal rearrangements will over time evolve to a structured genome with islands of core genes and of genes with functions in adaptation. These authors argue that transposable elements mediated rearrangements and structured genomes allow rapid adaptation to new, changing environments, a scenario that can easily be visualized for pathogenic species adapting to new hosts or escaping recognition by the host.

Genomes and their gene content, i.e. the abundance and nature of protein coding genes, differ greatly in oomycetes. Initial analyses focusing merely on the abundance of protein-coding genes in the plant pathogens *P. ramorum* and *P. sojae* identified large gene families that were not evident in their non-pathogenic relatives such as diatoms (Tyler et al. 2006; Martens et al. 2008; Judelson & Ah-Fong 2010). These expansions include gene families encoding ABC transporters, kinases, extracellular peptidases or plant material degrading hydrolases (Figure 1-4), but are especially apparent for genes encoding RXLR and Crinkler proteins (Tyler et al. 2006; Jiang et al. 2008; Haas et al. 2009; Schornack et al. 2010), two classes of recently described oomycete-specific effector proteins that translocate into the host cytoplasm. RXLR effectors have been mainly identified within Peronosporales and some of them, e.g. *RXLR29* in *H. arabidopsidis* (Cabral et al. 2011), have been shown to suppress host defenses. In contrast, Crinklers are evolutionarily older and experimentally tested Crinklers seem to trigger cell death (Schornack et al. 2010). Recently, Crinkler 8 (CRN8) has been shown to contain an active kinase domain that is necessary for cell death activity and therefore represents the first characterized example of a biochemically active effector (van Damme et al. 2012). Additional studies focusing on recently sequenced oomycetes such as the obligate biotroph *H. arabidopsidis* or the fish pathogen *S. parasitica* identified distinct expanded or reduced gene families. *S. parasitica* contains an expanded repertoire of peptidases, in particular peptidases of the C1 type, and different types of eukaryotic kinases, several of which have predicted membrane domains and therefore might be cell surface receptors involved in signaling (Jiang et al. 2013). In contrast *H. arabidopsidis*, but also other biotrophic oomycetes such as *A. laibachii*, display a remarkable reduction of pathogenicity-associated gene families and the loss of biosynthetic pathways (Baxter et al.



2010; Kemen et al. 2011) (Figure 1-4). In particular, several components of the inorganic nitrate assimilation pathway such as nitrate reductase, an enzyme that catalyzes the nitrate reduction for acquisition, have been lost (Baxter et al. 2010; Kemen et al. 2011). Similar reductions have been observed in biotrophic pathogenic fungi and have been hypothesized as being the consequence of the obligate biotroph lifestyle and the acquisition of nitrate from the host (Spanu et al. 2010).

In chapter 2 of this thesis, we describe our approach to study the expansion of protein families in plant-pathogenic Peronosporales. It is based on analyzing the individual functional building blocks of proteins, so called protein domains. Protein domains are independent functional, structural and evolutionary units and as such, a major focus of bioinformatics and biochemical research (Rossmann et al. 1974; Orengo et al. 1997; Vogel et al. 2004). We analyzed the expansion of protein domains and their implication to the biology of oomycetes by systematically characterizing the domain composition predicted in the proteomes of four sequenced Peronosporales, five fungal plant pathogens and 58 diverse eukaryotes. We detected nearly 250 expanded domains in fungal and oomycete plant pathogens and successfully linked a substantial part of these domains to pathogenicity. We highlight how comparative genomics can put gene products that have not yet been linked to pathogenicity into their functional context.

Initial studies of the genome content of oomycetes focused merely on the abundance of gene families and the quantitative changes relative to related taxa (Tyler et al. 2006; Martens et al. 2008; Haas et al. 2009). This, however, does not yet take into account different evolutionary scenarios that are needed to fully explain the gene content of extant oomycetes, such as gene gain, gene duplication and gene loss. In chapter 3, we present a phylogenomic study that describes the fundamental evolutionary dynamics that shaped the gene content of oomycetes. By reconciliation of nearly 19,000 individual gene trees with a reliable species phylogeny we determined the patterns of gene gains, duplications and losses. Thereby, we identified major transition points in the evolution of pathogenic oomycetes. The branch leading to *Phytophthora* displays an extraordinary number of duplications that happened in a constrained window of time. Martens and Van de Peer (2010), who used alternative methods to date the age of duplications, made similar a observation. These authors argue that this observation together with the presence of short blocks of homologous genes points to a whole-genome duplication in the last common ancestor of *Phytophthora*. This conclusion, however, is highly controversial, also because of the lack of larger intra-species synteny. Recent evidence taking into account additional oomycetes as well as a critical reevaluation of the initial evidence of homologous blocks of similar age seems to favor alternative scenarios and makes a whole genome duplication in the last common ancestor of *Phytophthora* unlikely (Van Hooff, unpublished). Another point of interest is the difference in the expansion and contraction patterns of distinct functional classes of genes, e.g. transcription factors. Studies on the evolution of distinct gene families (chapter 3) rather than just quantitative differences (chapter 2) allow the reconstruction of the exact mode of evolution that shaped (large or over-abundant) gene families such as glycoside hydrolases in oomycetes. Our results highlight the merit of phylogeny-based

analysis over traditional quantitative- or parsimony-based studies to decipher the mode of evolution in organisms.

### **The Genomes of Oomycetes Contain Novel Gene Fusions, Mainly Involved in Signaling**

Comparative studies of eukaryotes and bacteria have revealed a considerable amount of genes that result from gene fusions. Gene fusion occurs if two, often adjacent, genes are transcribed and translated as a single protein, often maintaining the distinct function of the individual genes. The first gene fusion between two metabolic genes in oomycetes was described in 1997 (Unkles et al. 1997). Since then, comparative studies in oomycetes, in particular in *Phytophthora*, have identified a considerable repertoire of novel gene fusions involving genes with roles in metabolism, but especially in signaling (Bakthavatsalam et al. 2006; Meijer & Govers 2006; Morris et al. 2009). The kinome of *P. infestans* harbors an intriguing ensemble of eukaryotic protein kinases with accompanying domains that form oomycete-specific domain compositions and are therefore likely the results of novel gene fusion (or recombination) (Judelson & Ah-Fong 2010). Another notable example is the specific fusion of the phosphatidylinositol 3-phosphate-binding zinc finger domain (FYVE) together with the phosphatidylinositol kinase (Meijer & Govers 2006).

In chapter 2, we describe how comparative studies of protein domains and the occurrence of domain pairs (two adjacent domains encoded in a single gene) can help identifying novel domain combinations. Applying this approach to 67 eukaryotes (including four oomycetes), we identified a large repertoire of novel domain combinations in oomycetes. Many of the involved domains have a role in signaling and regulation, e.g. histone modification. Moreover, many of the domains involved are highly abundant in oomycetes, e.g. FYVE zinc finger and kinases. This indicates that oomycetes, or at least the analyzed plant pathogenic species, encode proteins that rewire existing signaling networks in a novel way that is distinct from other eukaryotes.

Recently, it has been proposed that the genome organization of oomycetes might facilitate the acquisition of viable fusions (Judelson 2012). Many genes in the genome of oomycetes, especially genes predicted not to be involved in pathogenicity, have short intergenic distances so that fusions can easily occur due to the loss of a stop codon. The availability of more genome sequences together with genome-wide RNA-Seq data to test potential fusion genes will further enhance the identification, validation and subsequent characterization of these genes and their role in the biology of oomycetes.

### **The Genomes of Oomycetes are Chimeric, Containing Traces From Multiple Divergent Sources**

The gene content of many prokaryotes and eukaryotes is not of homogeneous origin but rather displays a chimeric ancestry likely originating from horizontal or endosymbiotic gene transfer (HGT or EGT) (Figure 1-4). Indeed, analyses of several oomycete

genomes revealed the chimeric ancestry of many of its nuclear encoded genes from both eukaryotic and prokaryotic sources (Tyler et al. 2006; Richards et al. 2006; Morris et al. 2009; Richards et al. 2011). Of special interest are genes with functions involved in osmotrophy that have been acquired by cross-kingdom transfer from fungi (Richards et al. 2006; 2011). Richards and colleagues (2011) identified 33 HGTs in plant pathogenic oomycetes that had a fungal ancestry. Many gene families that originated from the HGT events are predicted to encode secreted proteins. After subsequent duplications, these gene families account for a large proportion of the secretome of plant pathogenic oomycetes (Richards et al. 2011). Interestingly, 13 out of the 33 HGT gene families are involved in transport, breakdown and remodeling of sugar, of which 9 encode secreted polysaccharide hydrolases with a possible role in osmotrophy. These fungi-derived HGTs might thereby specifically facilitate the successful adaptation to plants as hosts because osmotrophy has evolved before the acquisition of these gene families in oomycetes (Richards et al. 2011).

Another major contribution to the gene content is the acquisition of genes from endosymbiotic origin (Figure 1-4). In oomycetes, this is of particular interest because their closest relatives are autotrophic photosynthetic diatoms. Together with oomycetes they form the Stramenopile lineage which is, at least according to the Chromalveolate hypothesis, united with other chlorophyll-*c* containing lineages in a monophyletic supergroup. This is based on the assumption that a single shared secondary endosymbiosis event involving a red alga gave rise to plastids observed in several taxa. Genes from the initially acquired plastid might therefore be transferred via EGT to the nucleus, in the case of oomycetes prior to its secondary loss, thus contributing to the gene content of extant organisms. Indeed, the analysis of the genome sequence of *P. ramorum* and *P. sojae* identified several genes with likely red algae origin and this, together with other evidence, seemingly confirms the Chromalveolate hypothesis (Andersson & Roger 2002; Tyler et al. 2006). However, recent data derived from genome sequences of plastid harboring and lacking Stramenopiles, seem to favor a more complex scenario involving a later acquisition of the plastid by the autotrophic Stramenopiles (Stiller et al. 2009; Baurain et al. 2010). Stiller and colleagues (2009) determined that the fraction of observed genes with red algae origin within the genomes of pathogenic Stramenopiles is too low to significantly support the earlier acquisition of the plastid. They highlight the importance of carefully evaluating chimeric phylogenetic signals within genomes. The identified genes with red algae origin in oomycetes are therefore most likely not the result of an early endosymbiosis event within Stramenopiles, but likely point to acquisition of these genes via alternative routes such as ancient HGT. The availability of many more genomes of pathogenic fungi, oomycetes, their non-pathogenic autotrophic sister taxa, but especially of more red algae genomes, together with dedicated research will most certainly shed additional light of the complex evolutionary history of many gene families and the contribution of HGT and EGT to the genome content of extant species.

### ***Cis*-regulatory Elements are Inferred Using Omics Data in *Phytophthora***

Comparative genomics not only provides insights into the evolution of oomycete

genomes, their gene content and its function, but also serves as a starting point to identify and study other functional regions within their genomes (see e.g. Hardison 2003). Of special interest are conserved DNA regions that interact with transcription factors involved in the regulation of expression of genes during different phases of development and infection.

Whereas extensive studies on eukaryotic model organisms and human identified a magnitude of functional DNA elements upstream of coding genes (Singer et al. 1990; Kutach & Kadonaga 2000; Majewski & Ott 2002; Müller et al. 2007; Yang et al. 2007; Hahn & Young 2011; Hoskins et al. 2011) this knowledge in oomycetes is rather limited. In eukaryotes, the basic transcriptional activity is determined by the eukaryotic core promoter that consists of different combinations of functional regions, e.g. the TATA-box and the Initiator (Inr) element, as well as proximal elements such as the common CCAAT-box. Pre-whole genome studies on a small set of oomycete genes identified few non-canonical TATA-box elements (Judelson et al. 1992) and a modified eukaryotic Inr-element with a conserved downstream region called FPR (Pieterse et al. 1994; McLeod et al. 2004), which has not yet been described as an important functional region in other eukaryotes. In addition, very few developmental-specific DNA elements, mostly involved in regulation of sporulation genes, have been identified (Tani & Judelson 2006; Xiang et al. 2009). *P. sojae* for example seems to encode up to 900 transcription factors (Rayko et al. 2010). This leaves a considerable gap when considering the number of so far determined potential transcription factors that are encoded in oomycete genomes.

In chapter 4, we describe the first genome-wide survey of potentially active *cis*-regulatory elements in three *Phytophthora* species with a main focus on the late blight pathogen *P. infestans*. We applied an *in silico* approach that uses gene co-expression and conservation between related species to define functional DNA elements, a method that has been successfully applied in other eukaryotic species (van Noort & Huynen 2006; Vandepoele et al. 2006). The availability of whole-genome sequences and accompanying gene expression data also allowed us to apply this integrative and comparative genomics methodology in *Phytophthora*. We identified several highly abundant motifs with similarity to common eukaryotic promoter elements, which have not yet been described for oomycetes in this quantity. Moreover, we identified several candidates of potentially functional DNA motifs that occur upstream of particular groups of genes such as effectors or transcription factors, thereby providing the first step to a more comprehensive description of transcriptional regulation in oomycetes.

### **Comparative Genomics Guides the Determination of Functional Associations Between Proteins**

Historically, research on genes and proteins has mainly focused on their evolution, their individual function and their contribution to the biology. However, *in vivo*, proteins rarely act solitarily; instead they either directly or indirectly associate with other proteins to synergistically perform complex functions. These interactions can be experimentally determined in a large-scale, ideally genome-wide, fashion using *in vivo* and

*in vitro* techniques such as yeast-two hybrid (see e.g. Fromont-Racine et al. 1997; Ito et al. 2001) or tandem affinity purification (see e.g. Gavin et al. 2002; 2006). These techniques are labor intensive and expensive and therefore mainly eukaryotic model organisms have been subjected to this type of studies. However, *in silico* analyses (e.g. Jansen et al. 2003; Lee et al. 2004; 2010) that integrate distinct genomic and transcriptomic data sets provide alternative strategies to determine the functional associations between proteins in other organisms.

These studies often apply Bayesian frameworks to integrate heterogeneous data into a single unified network that describes all determined associations between proteins (Jansen et al 2003; Lee et al. 2004; 2010). The advantage of this methodology is that each data source adds a distinct level of evidence to each functional linkage between two proteins, correcting for differences in the quality of individual data sources. Gene co-expression for example is only a moderate predictor for functional associations between proteins and therefore linkage based on co-expression should add less confidence to a prediction than other more reliable predictors. Different types of functional omics data have been integrated to predict functional associations. These comprise transcriptomic data for co-expression (e.g. Hughes et al. 2000; Gollub et al. 2003), conserved co-expression (Teichmann & Babu 2002; van Noort et al. 2003), phylogenetic co-occurrence (Pellegrini et al. 1999) and interolog mapping. Interolog mapping describes the transfer of associations, typically protein-protein interactions, from one organism to another. The underlying assumption is the conservation of associations, i.e. proteins that have been experientially shown to be associated, will maintain this association in the other species, too (Walhout et al. 2000). Phylogenetic co-occurrence predicts functional associations between proteins based on the similarity of their phylogenetic profiles, i.e. the presence or absence of genes, across a large amount of divergent species (Pellegrini et al. 1999). It assumes that associated partners should either be gained or lost together, since a single protein cannot perform the synergistic function. Conservation of co-expression across related species has been shown to significantly enhance the predictive power of co-expression towards functional association (Teichmann & Babu 2002; van Noort et al. 2003).

So far none of the above described experiments or bioinformatics network interference studies to comprehensively determine functional associations between proteins has been performed within oomycetes. Yeast-two hybrid technologies were applied to characterize interactions between effector proteins of the downy mildew *H. arabidopsidis* with the known actors in the *A. thaliana* immune system (Mukhtar et al. 2011), but only a subset of the effector proteins was included in this study. Other experimentally determined protein-protein interactions or functional associations are as yet not available for other oomycetes.

Over the last decade, a considerable amount of diverse large-scale functional omics data for several oomycetes became available (e.g. Judelson 2012; see above). In chapter 5, we outline our approach to utilize these available data to predict the first comprehensive functional association network in the late blight pathogen *P. infestans*.

We used a Bayesian approach to assess the merit and integrate four different data sets (co-expression, conserved co-expression, phylogenetic co-occurrence and interolog mapping) providing linkage for a considerable amount of proteins in *P. infestans*. These associations can be used as a framework to obtain additional biological knowledge from available and upcoming large-scale omics data. Using microarray data and the novel predicted associations, we identified and studied the functional module of up-regulated genes early during sporulation. Next to known genes involved in this developmental process, we observed many novel candidates that can now be linked to these important processes in the life cycle of *P. infestans*. Thereby, we highlight how the derived functional association network provides an essential addition to the growing number of genomic resources for *P. infestans*.

## REFERENCES

- Andersson JO, Roger AJ. 2002. A cyanobacterial gene in nonphotosynthetic protists—an early chloroplast acquisition in eukaryotes? *Curr. Biol.* 12:115–119.
- Arabidopsis* Genome Initiative. 2000. Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature.* 408:796–815.
- Archibald JM. 2009. The puzzle of plastid evolution. *Curr. Biol.* 19:R81–8.
- Baker B, Zambryski P, Staskawicz B, Dinesh-Kumar SP. 1997. Signaling in plant-microbe interactions. *Science.* 276:726–733.
- Bakthavatsalam D, Meijer HJG, Noegel AA, Govers F. 2006. Novel phosphatidylinositol phosphate kinases with a G-protein coupled receptor signature are shared by *Dictyostelium* and *Phytophthora*. *Trends Microbiol.* 14:378–382.
- 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.
- Baurain D et al. 2010. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Mol. Biol. Evol.* 27:1698–1709.
- Baxter L et al. 2010. Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. *Science.* 330:1549–1551.
- Bhattacharya D, Yoon HS, Hedges BS, Hackett JD. 2009. Eukaryotes (Eukaryota). In: *The timetree of life*. Hedges, BS & Kumar, S, Editors. Oxford University Press: New York.
- Birch PR, Whisson SC. 2001. *Phytophthora infestans* enters the genomics era. *Mol Plant Pathol.* 2:257–263.
- Bork P et al. 1998. Predicting function: from genes to genomes and back. *J. Mol. Biol.* 283:707–725.
- Cabral A et al. 2011. Identification of *Hyaloperonospora arabidopsidis* transcript sequences expressed during infection reveals isolate-specific effectors. *PLoS ONE.* 6:e19328.
- Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J. Eukaryot. Microbiol.* 46:347–366.
- Cavalier-Smith T, Allsopp MT, Chao EE. 1994. Chimeric conundra: are nucleomorphs and chromists monophyletic or polyphyletic? *Proc. Natl. Acad. Sci. U.S.A.* 91:11368–11372.
- Cock JM et al. 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature.* 465:617–621.
- Crombach A, Hogeweg P. 2007. Chromosome rearrangements and the evolution of genome structuring and

- adaptability. *Mol. Biol. Evol.* 24:1130–1139.
- da Silva ACR et al. 2002. Comparison of the genomes of two *Xanthomonas* pathogens with differing host specificities. *Nature*. 417:459–463.
- Dean RA et al. 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*. 434:980–986.
- Duncan J. 1999. *Phytophthora*-an abiding threat to our crops. *Microbiology Today*.
- Fisher MC et al. 2012. Emerging fungal threats to animal, plant and ecosystem health. *Nature*. 484:186–194.
- Fleischmann RD et al. 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 269:496–512.
- Fromont-Racine M, Rain JC, Legrain P. 1997. Towards a functional analysis of the yeast genome through exhaustive two-hybrid screens. *Nat. Genet.* 16:277–282.
- Gavin A-C et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*. 415:141–147.
- Gavin A-C et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature*. 440:631–636.
- Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B. 1996. Life with 6000 genes. *Science*. 274:546, 563–7.
- Gollub J et al. 2003. The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.* 31:94–96.
- Govers F, Gijzen M. 2006. *Phytophthora* genomics: the plant destroyers' genome decoded. *Mol. Plant Microbe Interact.* 19:1295–1301.
- Grenville-Briggs L et al. 2011. A molecular insight into algal-oomycete warfare: cDNA analysis of *Ectocarpus siliculosus* infected with the basal oomycete *Eurycyrtos dicksonii*. *PLoS ONE*. 6:e24500.
- Grunwald NJ, Goss EM, Press CM. 2008. *Phytophthora ramorum*: a pathogen with a remarkably wide host range causing sudden oak death on oaks and ramorum blight on woody ornamentals. *Mol. Plant Pathol.* 9:729–740.
- Haas BJ et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature*. 461:393–398.
- Hahn S, Young ET. 2011. Transcriptional regulation in *Saccharomyces cerevisiae*: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics*. 189:705–736.
- Hardison RC. 2003. Comparative genomics. *PLoS Biol.* 1:E58.
- Hohl HR, Stössel P. 1976. Host–parasite interfaces in a resistant and a susceptible cultivar of *Solanum tuberosum* inoculated with *Phytophthora infestans*: tuber tissue. *Can. J. Bot.* 54:900–912.
- Hoskins RA et al. 2011. Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Res.* 21:182–192.
- Hughes TR et al. 2000. Functional discovery via a compendium of expression profiles. *Cell*. 102:109–126.
- Hurst LD, Pál C, Lercher MJ. 2004. The evolutionary dynamics of eukaryotic gene order. *Nat. Rev. Genet.* 5:299–310.
- Ito T et al. 2001. A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc. Natl. Acad. Sci. U.S.A.* 98:4569–4574.
- Jansen R et al. 2003. A Bayesian networks approach for predicting protein–protein interactions from genomic data. *Science*. 302:449–453.
- Jiang RHY, Tripathy S, Govers F, Tyler BM. 2008. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. *Proc. Natl. Acad. Sci. U.S.A.* 105:4874–4879.
- Jiang RHY et al. 2013. Distinctive expansion of potential virulence genes in the genome of the oomycete fish

- pathogen *Saprolegnia parasitica*. PLoS Genetics. In press
- Judelson HS. 2012. Dynamics and innovations within oomycete genomes: insights into biology, pathology, and evolution. *Eukaryotic Cell*. 11:1304–1312.
- Judelson HS et al. 2008. Gene expression profiling during asexual development of the late blight pathogen *Phytophthora infestans* reveals a highly dynamic transcriptome. *Mol. Plant Microbe Interact*. 21:433–447.
- Judelson HS, Ah-Fong AMV. 2010. The kinome of *Phytophthora infestans* reveals oomycete-specific innovations and links to other taxonomic groups. *BMC Genomics*. 11:700.
- Judelson HS, Blanco FA. 2005. The spores of *Phytophthora*: weapons of the plant destroyer. *Nat Rev Microbiol*. 3:47–58.
- Judelson HS, Tyler BM, Michelmore RW. 1992. Regulatory sequences for expressing genes in oomycete fungi. *Mol. Gen. Genet*. 234:138–146.
- Judelson HS, Tyler BM, Michelmore RW. 1991. Transformation of the oomycete pathogen, *Phytophthora infestans*. *Mol. Plant Microbe Interact*. 4:602–607.
- Kamoun S. 2003. Molecular genetics of pathogenic oomycetes. *Eukaryotic Cell*. 2:191–199.
- Kamoun S, Hrabec 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, van West P, Vleeshouwers V, de Groot KE, Govers F. 1998. Resistance of *Nicotiana benthamiana* to *Phytophthora infestans* is mediated by the recognition of the elicitor protein INF1. *Plant Cell*. 10:1413–1426.
- Keeling PJ. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. *J. Eukaryot. Microbiol*. 56:1–8.
- Kemen E et al. 2011. Gene gain and loss during evolution of obligate parasitism in the white rust pathogen of *Arabidopsis thaliana*. *PLoS Biol*. 9:e1001094.
- Koonin EV. 2005. Orthologs, paralogs, and evolutionary genomics. *Annu. Rev. Genet*. 39:309–338.
- Kroon LPNM, Brouwer H, de Cock AWAM, Govers F. 2012. The genus *Phytophthora* anno 2012. *Phytopathology*. 102:348–364.
- Kunjeti SG et al. 2012. RNA-Seq reveals infection-related global gene changes in *Phytophthora phaseoli*, the causal agent of lima bean downy mildew. *Mol Plant Pathol*. 13:454–466.
- Kutach AK, Kadonaga JT. 2000. The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Mol. Cell. Biol*. 20:4754–4764.
- Lander ES et al. 2001. Initial sequencing and analysis of the human genome. *Nature*. 409:860–921.
- Large EC. 1940. *The Advance of the fungi*. Johnathan Cape Ltd, London.
- Latijnhouwers M, de Wit PJGM, Govers F. 2003. Oomycetes and fungi: similar weaponry to attack plants. *Trends Microbiol*. 11:462–469.
- Latijnhouwers M, Govers F. 2003. A *Phytophthora infestans* G-protein beta subunit is involved in sporangium formation. *Eukaryotic Cell*. 2:971–977.
- Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY. 2010. Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nat. Biotechnol*. 28:149–156.
- Lee I, Date SV, Adai AT, Marcotte EM. 2004. A probabilistic functional network of yeast genes. *Science*. 306:1555–1558.
- Lévesque CA et al. 2010. Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire. *Genome Biol*. 11:R73.
- Links MG et al. 2011. De novo sequence assembly of *Albugo candida* reveals a small genome relative to other biotrophic oomycetes. *BMC Genomics*. 12:503.



- Majewski J, Ott J. 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res.* 12:1827–1836.
- Martens C, Van de Peer Y. 2010. The hidden duplication past of the plant pathogen *Phytophthora* and its consequences for infection. *BMC Genomics.* 11:353.
- Martens C, Vandepoele K, Van de Peer Y. 2008. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proc. Natl. Acad. Sci. U.S.A.* 105:3427–3432.
- Maruyama S, Matsuzaki M, Misawa K, Nozaki H. 2009. Cyanobacterial contribution to the genomes of the plastid-lacking protists. *BMC Evol. Biol.* 9:197.
- McLeod A, Smart CD, Fry WE. 2004. Core promoter structure in the oomycete *Phytophthora infestans*. *Eukaryotic Cell.* 3:91–99.
- Meijer HJG, Govers F. 2006. Genomewide analysis of phospholipid signaling genes in *Phytophthora* spp.: novelties and a missing link. *Mol. Plant Microbe Interact.* 19:1337–1347.
- Money NP. 1998. Why oomycetes have not stopped being fungi. *Mycological research.* 102:767–768.
- Money NP, Davis CM, Ravishankar JP. 2004. Biomechanical evidence for convergent evolution of the invasive growth process among fungi and oomycete water molds. *Fungal Genet. Biol.* 41:872–876.
- Morris PF et al. 2009. Multiple horizontal gene transfer events and domain fusions have created novel regulatory and metabolic networks in the oomycete genome. *PLoS ONE.* 4:e6133.
- Mukhtar MS et al. 2011. Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science.* 333:596–601.
- Müller F, Demény MA, Tora L. 2007. New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors. *J. Biol. Chem.* 282:14685–14689.
- Orengo CA et al. 1997. CATH—a hierarchic classification of protein domain structures. *Structure.* 5:1093–1108.
- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.* 96:4285–4288.
- Pennisi E. 2010. Armed and dangerous. *Science.* 327:804–805
- Pieterse CM et al. 1994. Structure and genomic organization of the *ipiB* and *ipiO* gene clusters of *Phytophthora infestans*. *Gene.* 138:67–77.
- Qutob D, Hraber PT, Sobral BW, Gijzen M. 2000. Comparative analysis of expressed sequences in *Phytophthora sojae*. *Plant Physiol.* 123:243–254.
- Raffaele S et al. 2010. Genome evolution following host jumps in the Irish potato famine pathogen lineage. *Science.* 330:1540–1543.
- Randall TA et al. 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. doi: 10.1094/MPMI-18-0229.
- Rayko E, Maumus F, Maheswari U, Jabbari K, Bowler C. 2010. Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. *New Phytol.* 188:52–66.
- Richards TA et al. 2011. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proc. Natl. Acad. Sci. U.S.A.*
- Richards TA, Dacks JB, Jenkinson JM, Thornton CR, Talbot NJ. 2006. Evolution of filamentous plant pathogens: gene exchange across eukaryotic kingdoms. *Curr. Biol.* 16:1857–1864.
- Rossmann MG, Moras D, Olsen KW. 1974. Chemical and biological evolution of nucleotide-binding protein. *Nature.* 250:194–199.
- Savidor A et al. 2006. Expressed peptide tags: an additional layer of data for genome annotation. *J. Proteome Res.* 5:3048–3058.

- Savory EA et al. 2012. mRNA-Seq analysis of the *Pseudoperonospora cubensis* transcriptome during cucumber (*Cucumis sativus* L.) infection. *PLoS ONE*. 7:e35796.
- Schorneck S et al. 2010. Ancient class of translocated oomycete effectors targets the host nucleus. *Proc. Natl. Acad. Sci. U.S.A.* 107:17421–17426.
- Singer VL, Wobbe CR, Struhl K. 1990. A wide variety of DNA sequences can functionally replace a yeast TATA element for transcriptional activation. *Genes Dev.* 4:636–645.
- Spanu PD et al. 2010. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science*. 330:1543–1546.
- Stassen JH, Van den Ackerveken G. 2011. How do oomycete effectors interfere with plant life? *Curr Opin Plant Biol.* 14:1–8. 14:1–8.
- Stassen JHM et al. 2012. Effector identification in the lettuce downy mildew *Bremia lactucae* by massively parallel transcriptome sequencing. *Mol Plant Pathol.*
- Stiller JW, Huang J, Ding Q, Tian J, Goodwillie C. 2009. Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? *BMC Genomics*. 10:484.
- Tani S, Judelson H. 2006. Activation of zoosporogenesis-specific genes in *Phytophthora infestans* involves a 7-nucleotide promoter motif and cold-induced membrane rigidity. *Eukaryotic Cell*. 5:745–752.
- Teichmann SA, Babu MM. 2002. Conservation of gene co-regulation in prokaryotes and eukaryotes. *Trends Biotechnol.* 20:407–10
- Thines M, Kamoun S. 2010. Oomycete-plant coevolution: recent advances and future prospects. *Curr Opin Plant Biol.* 13:427–433.
- Torto-Alalibo T et al. 2005. Expressed sequence tags from the oomycete fish pathogen *Saprolegnia parasitica* reveal putative virulence factors. *BMC Microbiol.* 5:46.
- Tyler BM et al. 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science*. 313:1261–1266.
- Unkles SE, Logsdon JM, Robison K, Kinghorn JR, Duncan JM. 1997. The *tigA* gene is a transcriptional fusion of glycolytic genes encoding triose-phosphate isomerase and glyceraldehyde-3-phosphate dehydrogenase in oomycota. *J. Bacteriol.* 179:6816–6823.
- van Damme M et al. 2012. The Irish potato famine pathogen *Phytophthora infestans* translocates the CRN8 kinase into host plant cells. *PLoS Pathog.* 8:e1002875.
- van Noort V, Huynen MA. 2006. Combinatorial gene regulation in *Plasmodium falciparum*. *Trends Genet.* 22:73–78.
- van Noort V, Snel B, Huynen MA. 2003. Predicting gene function by conserved co-expression. *Trends Genet.* 19:238–242.
- van West P, Kamoun S, van 't Klooster JW, Govers F. 1999. Internuclear gene silencing in *Phytophthora infestans*. *Mol. Cell.* 3:339–348.
- Vandepoele K, Casneuf T, Van de Peer Y. 2006. Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. *Genome Biol.* 7:R103.
- Venter JC et al. 2001. The sequence of the human genome. *Science*. 291:1304–1351.
- Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. 2004. Structure, function and evolution of multi-domain proteins. *Curr. Opin. Struct. Biol.* 14:208–216.
- Walhout AJ et al. 2000. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. *Science*. 287:116–122.
- Werner S et al. 2002. Chitin synthesis during in planta growth and asexual propagation of the cellulosic oomycete and obligate biotrophic grapevine pathogen *Plasmopara viticola*. *FEMS Microbiol. Lett.* 208:169–173.

- Xiang Q, Kim KS, Roy S, Judelson HS. 2009. A motif within a complex promoter from the oomycete *Phytophthora infestans* determines transcription during an intermediate stage of sporulation. *Fungal Genet. Biol.* 46:400–409.
- Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E. 2007. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene.* 389:52–65.
- Ye W et al. 2011. Digital gene expression profiling of the *Phytophthora sojae* transcriptome. *Mol. Plant Microbe Interact.* 24:1530–1539.





# A Domain-Centric Analysis of Oomycete Plant Pathogen Genomes Reveals Unique Protein Organization

# 2

Michael F Seidl<sup>1,2</sup>, Guido Van den  
Ackerveken<sup>2,3</sup>, Francine Govers<sup>2,4</sup>, and  
Berend Snel<sup>1,2</sup>

<sup>1</sup>Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Utrecht, The Netherlands

<sup>2</sup>Centre for BioSystems Genomics, Wageningen, The Netherlands

<sup>3</sup>Plant-Microbe Interactions, Department of Biology, Utrecht University, Utrecht, The Netherlands

<sup>4</sup>Laboratory of Phytopathology, Wageningen University, Wageningen, The Netherlands

*Plant Physiol.* **155**:628–644 (2011)  
Copyright American Society of Plant Biologists



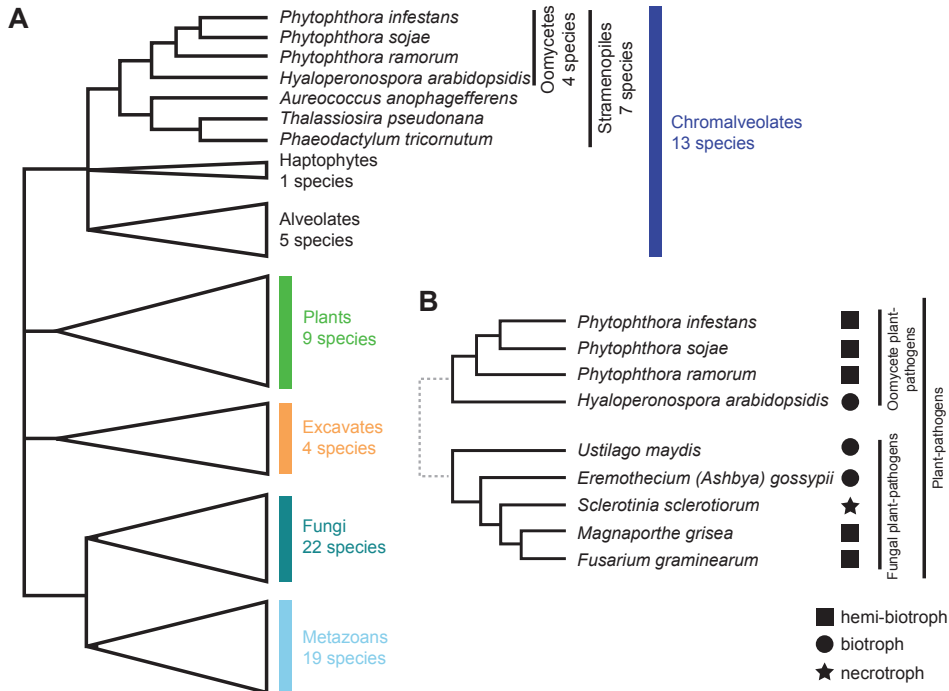
## ABSTRACT

Oomycetes comprise a diverse group of organisms that morphologically resemble fungi but belong to the stramenopile lineage within the supergroup of chromalveolates. Recent studies have shown that plant pathogenic oomycetes have expanded gene families that are possibly linked to their pathogenic lifestyle. We analyzed the protein domain organization of 67 eukaryotic species including four oomycete and five fungal plant pathogens. We detected 246 expanded domains in fungal and oomycete plant pathogens. The analysis of genes differentially expressed during infection revealed a significant enrichment of genes encoding expanded domains as well as signal peptides linking a substantial part of these genes to pathogenicity. Overrepresentation and clustering of domain abundance profiles revealed domains that might have important roles in host-pathogen interactions but, as yet, have not been linked to pathogenicity. The number of distinct domain combinations (bigrams) in oomycetes was significantly higher than in fungi. We identified 773 oomycete-specific bigrams, with the majority composed of domains common to eukaryotes. The analyses enabled us to link domain content to biological processes such as host-pathogen interaction, nutrient uptake, or suppression and elicitation of plant immune responses. Taken together, this study represents a comprehensive overview of the domain repertoire of fungal and oomycete plant pathogens and points to novel features like domain expansion and species-specific bigram types that could, at least partially, explain why oomycetes are such remarkable plant pathogens.

## INTRODUCTION

Oomycetes are a diverse group of organisms that live as saprophytes or as pathogens of plants, insects, fish, vertebrates, and microbes (Govers & Gijzen 2006). The numerous plant pathogenic oomycete species cause devastating diseases on many different host plants and have a huge impact on agriculture. A prominent example is *Phytophthora infestans*, the causal agent of late blight of potato (*Solanum tuberosum*) and tomato (*Solanum lycopersicum*) and responsible for the Irish potato famine in the 19<sup>th</sup> century. Plant pathogenic oomycetes include a large number of different species that vary in their lifestyle, from obligate biotrophic and hemibiotrophic to necrotrophic. In addition, they show great differences in host selectivity, ranging from broad to very narrow (Erwin & Ribeiro 1996; Agrios 2005). Oomycetes have morphological features similar to filamentous fungi, and the two groups exploit common infection structures and mechanisms (Latijnhouwers & Govers 2003). Together with diatoms, brown algae, and golden-brown algae, oomycetes are classified as stramenopiles, a lineage that is united with alveolates in the supergroup of chromalveolates (Baldauf et al. 2000; Yoon et al. 2002). The monophyly of this supergroup, however, is under debate (Baurain et al. 2010). The genomes of oomycetes sequenced so far are variable in size and content, ranging from 65 Mb in *Phytophthora ramorum* to 240 Mb in *P. infestans* (Haas et al. 2009), and only include plant pathogenic species. Analysis of these genomes revealed that several gene families facilitating the infection process are expanded (Martens et al. 2008). Extreme examples are gene families encoding cytoplasmic effector proteins such as RXLR effectors, which share the host cell-targeting motif RXLR and suppress defense responses in the host, and the necrosis-inducing proteins classified as Crinklers (Crn; Haas et al. 2009). To date, a few oomycete genomes have been sequenced, and this enables a comprehensive comparison of genomic features present in oomycetes, fungi, and other eukaryotic species such as gene families and protein domains. Experimentally derived functional knowledge of the majority of gene products in oomycetes in a comparable depth as for model species like *Saccharomyces cerevisiae* and *Arabidopsis thaliana* will likely not be accessible in the near future. Hence, comparative genomics provides an important framework to functionally characterize oomycete gene products and generate hypotheses on the basic cellular functions as well as the complex interactions of these plant pathogens with their hosts and environment.

In this study, we focus on protein domains because these are the basic functional, evolutionary, and structural units that shape proteins (Rossmann et al. 1974; Orengo et al. 1997; Vogel et al. 2004). Domains function independently in single-domain proteins or synergistically in multidomain proteins (Doolittle 1995; Vogel et al. 2004; Bashton & Chothia 2007). Accordingly, some domains always occur with a defined set of functional partners, whereas others are highly versatile and form combinations of two consecutively occurring domains (also called bigrams) with different N- or C-terminal partners (Marcotte et al. 1999; Basu et al. 2008). Here, we analyzed the domain repertoire predicted from the genome sequences of 67 eukaryotic species and compared filamentous plant pathogens with other eukaryotes with a special emphasis on oomycetes. We show how differences in the domain repertoire of oomycetes, especially in the ex-



**Figure 2-1 Phylogenetic relationships of the analyzed species.**

(A) The major eukaryotic groups considered in the analysis and the number of species represented in every group. For the exact species used in the analysis, see Supplemental Table S2-1. The tree is adapted from Simpson and Roger (2004) and incorporates the phylogeny for the stramenopiles based on Blair et al. (2008). (B) Fungal and oomycete plant-pathogenic species used in this analysis. The plant pathogens include species with different lifestyles, indicated by the symbol following the species name. The phylogeny for the fungi is based on James et al. (2006).

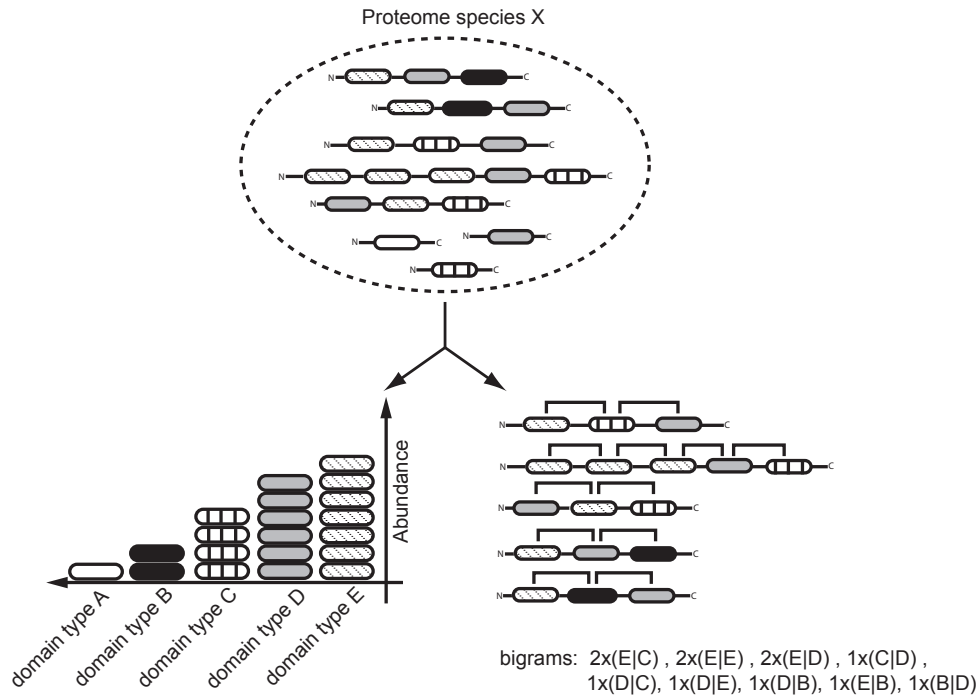
pansion of certain domain families and the formation of species-specific bigram types, can be linked to the biology of this group of organisms. This allowed the generation of candidate sets of proteins and domains that are likely to play roles in the lifestyle of oomycetes or their interaction with plants.

## RESULTS

### The Domain Repertoire of Oomycete Plant Pathogens and its Comparison with Other Eukaryotes

We analyzed the domain architecture of the predicted proteomes in 67 eukaryotes covering all major groups of the eukaryotic tree of life with the exception of the super-group Rhizaria (Figure 2-1A; Supplementary Table S2-1). We included seven strameno-





**Figure 2-2 Description of different metrics used in this study.**

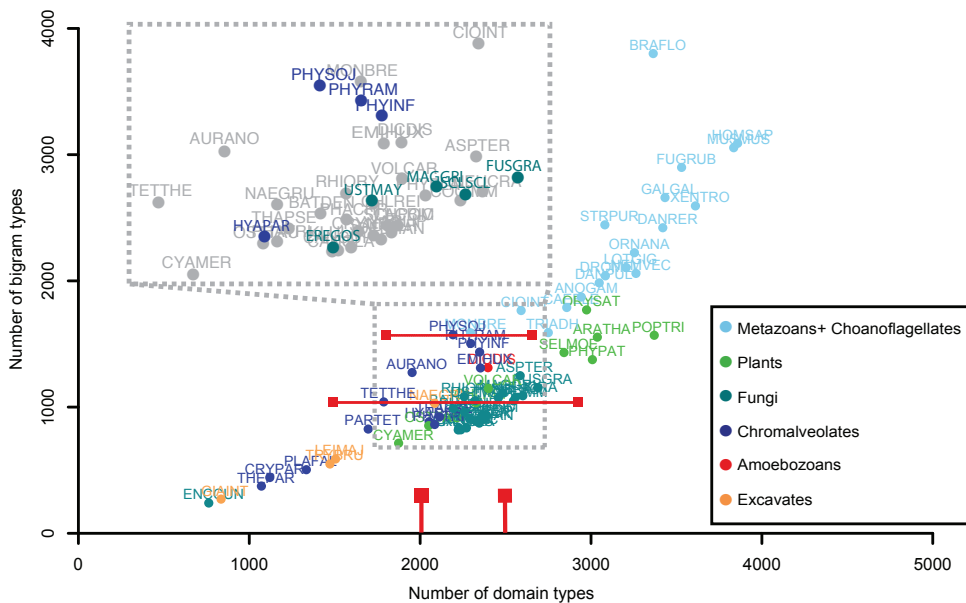
In the example shown, we observe five different domain types. The abundance of a domain type is defined as the number of occurrences of the individual entity within the species (e.g. domain type B has an abundance of two). The versatility is defined as the number of different direct adjacent N- or C-terminal neighbors. We distinguish between N- and C-terminal partners (e.g. the versatility of domain type C is three). A bigram is a set of two directly adjacent domains, and we also consider two entities of the same domain a bigram (e.g. we observe nine different bigram types in the proteome, of which three have an abundance of two (right panel)).

piles, four of which are plant pathogenic oomycetes, namely the obligate biotrophic downy mildew *Hyaloperonospora arabidopsidis* and three hemibiotrophic *Phytophthora* species. The selection also contained five fungal plant pathogens, including rice (*Oryza sativa*) blast fungus (*Magnaporthe grisea*) and corn (*Zea mays*) smut (*Ustilago maydis*), both species with a (hemi)biotrophic lifestyle comparable to the oomycete plant pathogens used in the analysis (Figure 2-1B).

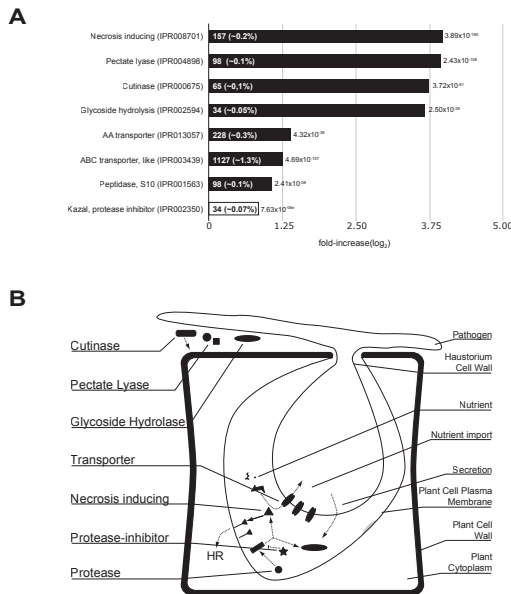
The domain architecture of all 1,250,996 predicted proteins in the 67 eukaryotic genomes was analyzed using HMMER (Eddy 1998) and a local Pfam-A database (Finn et al. 2010). Overall, 59% (737,851) of all proteins have one or more predicted domains. We detected a total of 1,464,807 domains in all species, 80,180 within the stramenopiles and 51,030 in oomycetes.

In order to characterize the domain repertoire of eukaryotes, we used two metrics:

the number of domain types and the number of different combinations of adjacent domains, also called bigrams (Figure 2-2). In total, 13,994 bigram types were identified in the 67 eukaryotic genomes, consisting of 6,356 different domain types. As described by Basu et al. (2008), the number of bigram types increases superlinearly relative to the number of domain types, with the highest numbers in multicellular organisms (Figure 2-3). We observed separate clusters for metazoans, fungi, and plants (including land plants and mosses). Oomycetes and fungi have similar numbers of domain types, ranging from 2,000 to 2,500; however, oomycetes, in particular *Phytophthora* species, contain significantly more bigram types. The three analyzed *Phytophthora* species appeared to have approximately 50% more bigram types compared with other organisms that have similar numbers of domain types (Figure 2-3;  $P = 0.0019$ , by one-sided Wilcoxon rank-sum test). This even holds when we apply a more conservative approach by discarding all domain and bigram types that occur once in each predicted proteome (Supplementary Figure S2-1A). We observed that the number of domain types as well as the number of bigram types increases with proteome size and reaches saturation for larger proteomes (Supplementary Figure S2-1, B and C; Cosentino Lagomarsino et al. 2009).



**Figure 2-3 Dependence of the number of domain and bigram types observed in the analyzed species.** The average number of different bigrams of species that have between 2,000 and 2,500 different domain types is indicated with the bottom horizontal red bar. The top horizontal red bar indicates the average number of different bigrams for *Phytophthora* species. The full species names corresponding to the abbreviations can be found in Supplemental Table S2-1. A magnification of the area encompassing the oomycete and fungal plant pathogens is shown; the species of interest are highlighted. The dots are colored according to the major eukaryotic groups as indicated in the text box.



**Figure 2-4 Overrepresentation of selected, well-described domains involved in plant-pathogen interaction and establishing or maintaining infection.**

(A) The log<sub>2</sub>-fold overrepresentation of the domains in plant pathogens is shown in the bar chart. The absolute number of occurrences in plant pathogens and the percentage of all predicted domains in plant pathogens are displayed in the bars, and the corrected P values are shown at the tip of the bars. The fold overrepresentation and the P value for the Kazal protease inhibitor domain were based on the overrepresentation in oomycetes compared with plant pathogens (indicated by the white bar and asterisks). (B) The overrepresented domains described in (A) are depicted in their possible cellular role during infection of the plant host.

Although oomycetes and in particular *Phytophthora* species contain a similar number of domain types as fungi, they have a larger predicted proteome (Supplementary Figure S2-1B). However, they contain more bigram types than fungi but less than other species with predicted proteomes of similar size (e.g. *Drosophila melanogaster*; Supplementary Figure S2-1C).

### Domain Overrepresentation Provides a Snapshot of Pathogen-Host Interaction

Apart from a wide and abundant repertoire of domains related to transposable elements (Haas et al. 2009), the most abundant domain types in oomycetes are similar to those in other eukaryotes (Supplementary Table S2-2). Hence, absolute domain abundance alone is not indicative enough to correlate domains to the lifestyle of both fungal and oomycete plant pathogens. Instead, we identified domains that are overrepresented in plant pathogens relative to other eukaryotes (Figure 2-1B).

Our analysis inferred 246 overrepresented domains in plant pathogens that are observed in 24,970 proteins ( $P < 0.001$ , by Fisher's exact test; a selection of well-described overrepresented domains is depicted in Figure 2-4A; Supplementary Table S2-3). Since we analyzed the expansion in plant pathogens at the level of a group rather than an individual species, domains that are reported as being expanded in the group are not necessarily expanded in all species of the group or may even be absent (Supplementary Table S2-3). For example, secreted proteins encoding carbohydrate-binding family 25 domains (IPR005085) are only found in *Phytophthora* species and not in fungal plant pathogens, whereas secreted proteins containing the Cys-rich domain (CFEM;

IPR008427) are only observed in fungal pathogens (Kulkarni et al. 2003).

Many proteins involved in host-pathogen interaction are secreted in the apoplast or, like the RXLR effector proteins, translocated into host cells following their secretion from the pathogen (Haas et al. 2009). Hence, we also predicted the presence of potential N-terminal signal peptide sequences in the whole proteomes of the analyzed species. The combined secretome encompasses 100,521 potentially secreted proteins, of which 11,352 are predicted in plant pathogens (Supplementary Figure S2-2). Approximately 20% (2,478) of these proteins contain overrepresented domains; hence, proteins containing overrepresented domains are 1.85-fold enriched in the predicted secretome of the analyzed plant pathogens ( $P = 2.57 \times 10^{-231}$ , by Fisher's exact test).

Oomycete proteins with significantly expanded domains are prime candidates for being pathogenicity associated. To assess this hypothesis, we tested if *P. infestans* genes that are differentially expressed during infection of the potato host are enriched for the aforementioned expanded domains. For this, we utilized NimbleGen microarray data that include genome-wide expression levels of *P. infestans* genes at different days post inoculation (dpi) of potato leaves as well as from mycelium grown *in vitro* on different media (Haas et al. 2009). We identified in total 1,584 genes that are significantly induced or repressed in *P. infestans* during infection (differentially expressed for at least one of the time points 2–5 dpi) compared with those grown *in vitro* (three different growth media;  $P < 0.05$ ,  $q < 0.05$ , by t test; Supplementary Table S2-4A). Of the 1,584 differentially expressed genes, 259 encode proteins containing significantly expanded domains (Supplementary Table S2-4B), which is 1.2-fold more than expected ( $P = 8.8 \times 10^{-5}$ , by Fisher's exact test). Moreover, 44 of these 259 genes also encode proteins with a predicted signal peptide, which is a significant enrichment (1.8-fold;  $P = 4.38 \times 10^{-5}$ , by Fisher's exact test). The majority (41) of these 44 genes are differentially expressed early in infection (2 dpi; Figure 2-5A). All genes differentially expressed at 3 dpi are also differentially expressed at 2 dpi (Figure 2-5, A and B). Consequently, the 44 differentially expressed genes coding for proteins with both predicted signal peptides as well as overrepresented domains are promising candidates for pathogenicity-associated proteins, of which several will be discussed in detail below.

For several groups of overrepresented domains, a direct or indirect role in host-pathogen interaction and/or plant pathogen lifestyle has already been hypothesized or demonstrated (Dean et al. 2005; Tyler et al. 2006; Haas et al. 2009). Nearly 18% of the 246 overrepresented domains belong to three groups of domains: (1) hydrolase domains; (2) domains involved in substrate transport over membranes, such as the general ATP-binding cassette (ABC) transporter-like domain (IPR003439) but also more specialized transporters of sulfate (IPR011547) and amino acids (IPR004841/IPR013057); and (3) domains present in peptidases, such as the metalloprotease-type M28 domain (IPR007484) found in many secreted proteins. Of the hydrolases, which encompass 9% of the overrepresented domains, the majority is present in enzymes that hydrolyze glycosidic bonds. An example is the glycoside hydrolase (GH) family 12 domain (IPR002594). This domain is observed 34 times in plant pathogens, which overall



contain 91,747 domains, and 43 times in all eukaryotes, which have a total of 1,464,807 domains, and hence is 12.62-fold (3.66  $\log_2$ -fold) enriched in the plant pathogens. This domain is mainly observed in secreted proteins (27 out of 34; SignalP prediction). The majority (79%) of the GH-12 domains are found in oomycete plant pathogens, and the expression of two of these hydrolase genes in *P. infestans* (PITG\_08944 and PITG\_16991) is significantly induced during infection of potato (Figure 2-5; Supplementary Table S2-4). In total, 33 differentially expressed genes during plant infection in *P. infestans* encode proteins that contain GH domains, including GH-17 (IPR000490) in endo-1,3- $\beta$ -glucosidases and GH-81 (IPR005200) in  $\beta$ -1,3-glucanases as well as several members of GH-28 (IPR000743), a domain involved in soft rotting of host tissues and described in both fungal and bacterial plant pathogens (He & Collmer 1990; Ruttkowski et al. 1990). Twenty-eight *P. infestans* genes coding for domains involved in transmembrane transport are differentially expressed during plant infection (Supplementary Table S2-4). Examples of genes encoding domains involved in substrate transport over the membrane are PITG\_04307, which encodes an ABC-2-type transporter (IPR013525), PITG\_12808, which encodes an amino acid transporter (IPR013057), as well as PITG\_22087, a gene encoding both ABC-like (IPR003439) and ABC-2-type domains (Supplementary Table S2-4). Extracellular degrading enzymes like cutinases contain an overrepresented domain (IPR000675;  $P = 3.72 \times 10^{-61}$ ). This domain is observed 65 times in plant pathogenic species, corresponding to a 13.3-fold (3.73  $\log_2$ -fold) enrichment (Figure 2-4A). In total, 61 proteins in plant pathogens predicted to possess this domain are potentially secreted. Another overrepresented domain that is present in secreted proteins and involved in maceration and soft rotting of plant tissue is the pectate lyase (IPR004898). This domain is 15.34-fold (3.94  $\log_2$ -fold) enriched in plant pathogens and mainly found in oomycetes. Five genes in *P. infestans* encode this domain as well as a predicted N-terminal signal peptide and are differentially expressed (Figure 2-5).

### Novel Candidate Domains Significantly Expanded in Plant Pathogens

Next to domains that were already directly or indirectly implied in host-pathogen interaction, we identified novel candidates that are also expanded in plant pathogens, several of which are encoded in *P. infestans* genes differentially expressed during infection of the host. Genes encoding the significantly expanded alcohol dehydrogenase (zinc binding; IPR013149) as well as a GroES-like alcohol dehydrogenase (IPR013154) domains are ubiquitous in all analyzed eukaryotes, and also the combination of these two domains is present in all species with only a few exceptions. Nine of these genes in *P. infestans* are induced during infection (Supplementary Table S2-4). Sixty-five genes in plant pathogens encode proteins with FAD-linked oxidase (IPR006094) and berberine/berberine-like (BBE) domains (IPR012951), of which three out of six in *P. infestans* are induced during infection (PITG\_02928, PITG\_02930, and PITG\_20764). The BBE domain is involved in the biosynthesis of the alkaloid berberine (Facchini et al. 1996). The genes encode a predicted N-terminal signal peptide, although molecular analysis of proteins containing these domains in plants indicated that at least some of these are not secreted but instead are targeted to specialized vesicles (Amann et al. 1986; Kutchan & Dittrich 1995; Facchini et al. 1996). Moreover, Moy et al. (2004) observed induced expression

of a soybean (*Glycine max*) gene (BE584185) shortly after infection with *Phytophthora sojae* containing these two domains. A recent analysis from Raffaele et al. (2010) focusing solely on the secretome in *P. infestans* corroborates our results and also concludes that proteins with BBE and FAD-linked oxidase domains are candidate virulence factors. Three genes encoding secreted metallophosphoesterases (IPR004843; PITG\_20454, PITG\_07720, and PITG\_10322) show induced gene expression. These metallophosphoesterase domains are found in phosphatases and hence are involved in the regulation of protein activity, since they work as antagonists of kinase activity.

For approximately 6% of all overrepresented domains, no or limited functional information is available in Pfam. These are the so-called DUFs: domains of unidentified function. Given their expansion in plant pathogens and the fact that other overrepresented domains are known to function in diverse aspects of plant-pathogen interactions, these DUFs are also likely to play a role in the lifestyle of plant pathogens and hence are promising targets for further experimental validation (Supplementary Table S2-3). Secreted proteins containing a combination of two overrepresented DUFs, DUF2403 (IPR018807) and DUF2401 (IPR018805), are exclusively found in fungi and in oomycetes, with the majority (approximately 75%) in oomycetes. The N-terminal DUF2403 contains a Gly-rich region without further functional annotation, whereas five highly conserved Cys residues characterize the C-terminal DUF2401. Proteins containing both DUFs have been characterized in *S. cerevisiae* and in *Candida albicans* as being covalently linked to the cell wall (Terashima et al. 2002; Yin 2005; Klis et al. 2009). Another overrepresented DUF within plant pathogens and mainly found in oomycetes is DUF953 (IPR010357). This domain is present in several eukaryotic proteins with thioredoxin-like function, and two genes in *P. infestans* containing this domain are differentially expressed during infection (PITG\_07008 and PITG\_07010). DUF590 (IPR007632), which is ubiquitous in nearly all eukaryotes, is observed in proteins containing eight putative transmembrane helices. These proteins exhibit calcium-activated ion channel activity and are involved in diverse biological processes (Yang et al. 2008). The *P. infestans* gene PITG\_06653 that contains the DUF590 domain is differentially expressed during infection, and this provides further support for a role in host-pathogen interaction. The exemplified DUFs as well as other overrepresented domains with less or no functional annotation are interesting candidates for further functional studies to decipher their precise role in plant pathogens.

### Domain Overrepresentation in Oomycete Plant Pathogens

Since the previous analysis grouped both fungal and oomycete plant pathogens, domains specifically enriched in oomycetes were not directly discernible. Hence, we compared the relative domain abundance predicted in plant pathogens (Figure 1B) with the aim to identify domains specifically enriched in oomycetes. Of the 75 domains that are overrepresented in oomycetes, 20 are not observed in any fungal plant pathogen and therefore can be considered oomycete specific within plant pathogens (Supplementary Table S2-5). In general, the abundance of expanded domains in *Phytophthora* species is higher than in *H. arabidopsidis*. A well-described example is the NPP1 do-

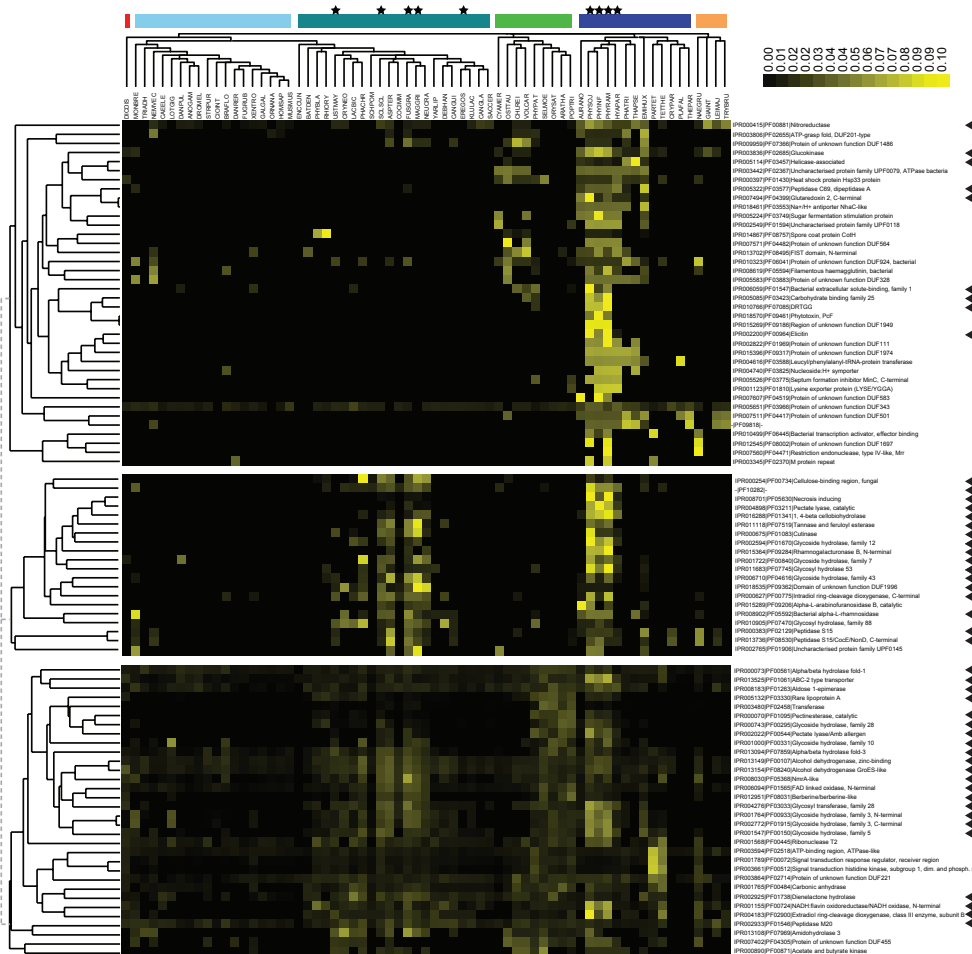
main (IPR008701) that is present in secreted (SignalP: 122) necrosis-inducing proteins. It shows a significant overrepresentation in oomycetes (1.68-fold [0.75  $\log_2$ -fold] enriched), in particular in *Phytophthora* species, but is also observed 10 times in fungal plant pathogens as well as in a few cases in nonpathogenic fungi as noted before (Gijzen & Nürnberger 2006). Four *P. infestans* genes encoding this domain are induced early during infection (2–3 dpi), whereas a single gene (PITG\_18453) is induced late (5 dpi). Several peptidases (e.g. containing the peptidase S1/S6 and C1A domains) are overrepresented compared with other plant pathogens. S1/S6 (IPR001254; 1.6-fold [0.74  $\log_2$ -fold]) is predicted in 91 proteins, of which 67 have a predicted secretion signal, while C1A (IPR000668; 1.79-fold [0.85  $\log_2$ -fold]) is predicted in 78 proteins, of which 31 are potentially secreted. C1A is present in several eukaryotic species, but within the plant pathogenic group it is exclusively found in oomycetes.

Several secreted protease inhibitors of the Kazal family containing the Kazal I1 (IPR002350) and Kazal-type (IPR011497) domains are significantly expanded in oomycetes and are within the group of analyzed plant pathogens specific to oomycetes. This suggests that they provide an increased level of protection of the pathogen against host-encoded defense-related proteases (Tyler et al. 2006). Another domain that is oomycete specific within the plant pathogens is the Na/Pi cotransporter (IPR003841) involved in the uptake of phosphate. Several other transporters that have already been described as being overrepresented in plant pathogens (e.g. the ABC-2-type transporters) are significantly expanded within oomycete plant pathogens, since these species are the major contributors to the overall abundance of this domain in plant pathogens. The abundance of predicted Ser/Thr-like kinase domains (IPR017442) compared with other plant pathogenic species is surprisingly high, and this domain is specifically expanded in the *Phytophthora* species. Even if several expanded domains are observed in both oomycete as well as fungal plant pathogens, the exploration of domains primarily expanded in oomycetes (e.g. certain transporter families and defense- and signaling-related domains) highlights functional entities that discriminate between these groups of plant pathogens.

### **Clustering of Abundance Profiles Reveals Additional Potential Pathogenicity Factors**

We extended the set of candidate domains that might be important for host-pathogen interaction beyond overrepresented domains by searching for additional domains that show presence, absence, and expansion profiles similar to overrepresented domains, since these domains are likely to be functionally linked or involved in similar biological processes (Pellegrini et al. 1999). We calculated a normalized profile of domain abundance and clustered similar abundance profiles using hierarchical clustering (Supplementary data S2-1). Several clusters contained a mix of significantly overrepresented domains and domains whose expansion in plant pathogens is not significant. We exemplify this with three clusters that contain 20% of all overrepresented domains in plant pathogens (Figure 2-6).





**Figure 2-6 Average linkage clustering of normalized domain profiles using Spearman rank correlation as a distance measurement.**

The species tree for all eukaryotic species is depicted on top, with the color code of their supergroup as introduced in Figure 2-1. Plant pathogens are marked with stars, and the arrowheads highlight domains identified as overrepresented in plant pathogens.

In the first cluster (Figure 2-6), domains are mainly expanded in oomycete plant pathogens. The abundance of some domains in plant pathogens is too low to be identified as being overrepresented. For example, the PcF domain (IPR018570), which is present in a small, approximately 50-amino acid necrosis-inducing protein found in various *Phytophthora* species (Orsomando et al. 2001; Liu et al. 2005), was not identified in the initial overrepresentation analysis. Also in this cluster is the sugar fermentation stimulation domain (IPR005224), which is mainly found in bacteria and involved in the regulation of maltose metabolism (Kawamukai et al. 1991). In this first cluster, we observed a high number (approximately 40%) of domains without functional characteriza-

tion that are mainly present in bacteria. An example is DUF1949 (IPR015269), a domain that is only found in the three analyzed *Phytophthora* species. This domain is observed in functional uncharacterized bacterial proteins like YIGZ in *Escherichia coli* K12 and adopts a ferredoxin-like fold (Park et al. 2004). The *Phytophthora* and bacterial proteins containing DUF1949 also contain a second, N-terminal uncharacterized protein family, UPF (UPF00029, IPR001498). This domain is also found in the human protein Impact and is conserved from bacteria to eukaryotes (Okamura et al. 2000). The *P. infestans* gene (PITG\_00027) containing both domains is induced early in infection (Supplemental Table S2-4B). Since these DUFs cluster with overrepresented domains, they are promising candidates for further study.

The domains in the second cluster mainly show an expansion of the abundance in both fungal and oomycete plant pathogens. This cluster contains, for example, cell wall-degrading domains like cutinases, pectate lyases, and other hydrolases and also the NPP1 domain that is found in necrosis-inducing proteins. The glycosyl hydrolase family 88 comprises unsaturated glucuronyl hydrolases thought to be involved in bio-film degradation and is mainly found in bacteria and fungi (Itoh et al. 2006). Interestingly, homologs are also observed in plant pathogenic bacteria (e.g. *Pectobacterium atrosepticum*), in fungi (e.g. *M. grisea*), and in all three *Phytophthora* species.

The third cluster contains domains that are not exclusively found in plant pathogens but have a broader abundance profile. This cluster includes a variety of overrepresented hydrolases, epimerases, and the ABC-2-type transporter domain (IPR013525) that is observed nearly 500 times in plant pathogenic species. Another domain that is found in this cluster is the dienelactone hydrolase domain (IPR002925), observed in all plant pathogens and also in other eukaryotic species, with a high abundance in plants as well as in fungi. This domain hydrolyzes dienelactone to maleylacetate in bacteria (Pathak et al. 1991) and is also detected in a putative 1,3:1,4- $\beta$ -glucanase from *P. infestans* that is proposed to be involved in cell wall metabolism (McLeod et al. 2003).

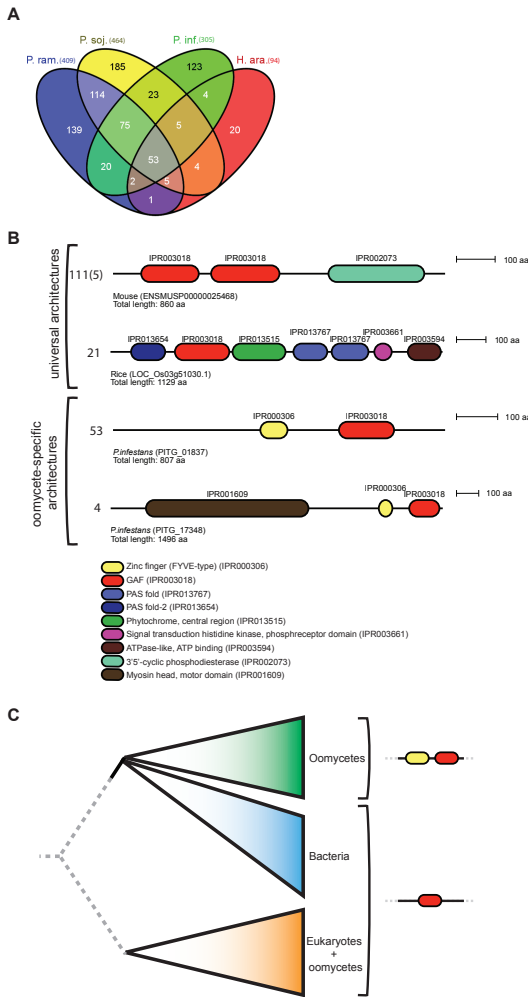
### Quantification of Oomycete-Specific Bigrams

Domains generally do not act as single entities in proteins but rather synergistically with other domains in the same protein or with domains in interacting proteins (Park et al. 2001; Vogel et al. 2004). Domains involved in signaling, sensing, and generic interactions are versatile and form combinations with several different partner domains (Supplementary Table S2-6). As described by others (Vogel et al. 2005), we observed that the versatility of domains is proportional to their abundance (Supplementary Figure S2-3). Hence, we applied a weighted bigram frequency that corrects for abundance to detect domains that are promiscuous or prone to form combinations with different partners (Basu et al. 2008). The average number of promiscuous domains in oomycetes is 424 and in *Phytophthora* is 464. This is higher than the average number of promiscuous domains (357) in all other species (Supplementary Table S2-7).

We observed that oomycetes have a higher number of bigram types than species

with a comparable number of domain types (Figure 2-3). We identified in total 13,994 different bigram types throughout the 67 analyzed species. The majority of these bigram types (i.e. 7,724, or 55.2%) are predicted in only a single species. In oomycetes, bigram types formed by domains that are associated with transposable elements showed a high abundance (Supplementary Tables S2-8 and S2-9). We identified 1,107 bigram types occurring exclusively in plant pathogens, the majority of which (773) are only observed in the analyzed oomycetes (Supplementary Table S2-10). These oomycete-specific bigram types are identified in total 1,511 times in 1,375 predicted proteins. Of the 773 oomycete-specific bigram types, 53 are present in all oomycetes (Figure 2-7A). The biggest overlap in oomycete-specific bigram types is observed between the *Phytophthora* species, especially between *P. ramorum* and *P. sojae*. A recent analysis of domain combination in *P. ramorum* and *P. sojae* already revealed several proteins involved in metabolism and regulatory networks containing novel bigrams (Morris et al. 2009). We additionally observed in total 43 bigram types that are shared either between *P. infestans* and *P. sojae* or between *P. infestans* and *P. ramorum*. However, the majority of oomycete-specific bigrams (467) are specific for a single species. The number of oomycete-specific bigram types highly exceeds the number of oomycete-specific domain types (41). Interestingly, only six of the oomycete-specific domains participate in forming the specific bigrams. Therefore, common domain types form the majority of the observed species-specific domain combinations, emphasizing the importance of novel domain combinations rather than novel domain types as a source for species-specific functionality. Even when we selectively look at the bigrams that occur at least twice in the same proteome or once in at least two different proteomes, we still observe 320 bigram types that are specific to oomycetes and occur in 982 predicted proteins.

Approximately 8% of the proteins containing an oomycete-specific bigram have a predicted secretion signal (9.2% of all oomycete proteins contain a predicted secretion signal). An example that is observed in a secreted putative Cys protease present in all analyzed oomycetes is the combination of the peptidase C1A domain (IPR000668) and the ML domain (IPR003172). The ML domain is known to be involved in lipid binding and innate immunity and has been observed in plants, fungi, and animals (Inohara & Nuñez 2002). The proteins containing this bigram also have an N terminal cathepsin inhibitory domain (IPR013201) that is often found next to the peptidase C1A domain and prevents access of the substrate to the binding cleft (Groves et al. 1996). Another bigram that is found in secreted proteins predicted in the analyzed *Phytophthora* species is the combination of the carbohydrate-binding domain family 25 (IPR005085; CBM25) with a GH-31 domain (IPR000322) as well as the tandem combination of CBM25 domains N terminal to the glycosyl hydrolase domain. The presence of the secreted CBM25 and GH-31 combination has recently been noted in *Pythium ultimum* (Lévesque et al. 2010). We further tried to elucidate the presence of RXLR or Crn motifs in proteins containing oomycete-specific bigrams. We predicted the presence of one of these motifs using individual HMMER models for both the RXLR and the Crn motif (see Materials and Methods). We overall predicted 746 proteins containing an RXLR and 99 proteins with a Crn motif. None of these proteins is predicted to contain an oomycete-specific bigram type.



**Figure 2-7 Quantification of oomycete-specific bigrams**

(A) Venn diagram depicting the presence of oomycete-specific bigram types in the analyzed oomycete proteomes and indicating the number of shared bigram types between different proteomes. The total number of oomycete-specific bigram types in each proteome is shown in parentheses. The Venn diagram was produced using Venny (Oliveros, 2007). (B) Domain architecture of example proteins containing a GAF domain. The top two architectures resemble common protein architectures: the cGMP-dependent 3',5'-cyclic phosphodiesterase (observed 111 times in eukaryotes and five times in oomycetes) and phytochrome A (observed 21 times in eukaryotes). The bottom two architectures depict oomycete-specific architectures: the FYVE-GAF fusion is observed 53 times independent of other domains, and the myosin motor head in combination with the FYVE-GAF fusion is observed four times, a single copy in each of the oomycetes included in this study. aa, amino acids. (C) Simplified evolutionary tree based on the phylogenetic analysis of the GAF domain in prokaryotes and eukaryotes. GAF domains from proteins with a FYVE-GAF fusion are exclusively found to be close to bacterial GAF domains. Other oomycete proteins that only contain the GAF domain without the FYVE domain also cluster with other eukaryotic sequences.

The most abundant oomycete-specific bigram type that occurs in 64 proteins is a combination of the phosphatidylinositol 3-phosphate-binding zinc finger (FYVE type) and the GAF domain. The presence of this oomycete-specific bigram in *P. ramorum* and *P. sojae* has been noted before (Morris et al. 2009). The GAF domain is described as one of the most abundant domains in small-molecule-binding regulatory proteins (Zoraghi et al. 2004). It is present in a large number of different proteins with a wide range of cellular functions, such as gene regulation (Aravind & Ponting 1997) and light detection and signaling (Sharrock & Quail 1989; Montgomery & Lagarias 2002). A typical eukaryotic domain composition involving the GAF domain is N terminal to the 3' 5'-cyclic phosphodiesterase domain found in phosphodiesterases that regulate pathways with cyclic nucleotide-monophosphate as second messengers (Sharrock & Quail 1989; Martinez et

al. 2002). This organization is observed in total 111 times, and five times in oomycetes (Figure 2-7B). The GAF-FYVE bigram is either observed as a single bigram (in 53 proteins) or in combination with other domains (in 11 proteins), for example with myosin (Richards & Cavalier-Smith 2005). In *P. infestans*, two genes (PITG\_07627 and PITG\_09293) encoding proteins with this combination are induced early during infection of the plant (Supplementary Table S2-4B). A phylogenetic analysis of the GAF domain in eukaryotes and prokaryotes showed that all GAF domains in oomycetes that are involved in the fusion with FYVE exclusively cluster with prokaryotic GAF domains, whereas other GAFs also cluster with eukaryotes. Hence, this suggests a horizontal gene transfer from bacteria to oomycetes of those GAF domains that are involved in the fusion with FYVE (Figure 2-7C; see Materials and Methods). The FYVE-type zinc finger is not identified in prokaryotic species; hence, we suggest two independent events, namely a horizontal gene transfer of the GAF domain from bacteria to oomycetes and subsequently a fusion to the zinc finger domain. Horizontal gene transfer seems to play an important role in the evolution of eukaryotes (Keeling & Palmer 2008), and recent evidence indicates that these events also have a significant contribution to the genome content of protists and oomycetes, as they received genetic material from different sources (Richards & Talbot 2007; Martens et al. 2008; Morris et al. 2009). Because GAF domains are known to be involved in many different cellular processes, we can only speculate about the biological function of proteins harboring the GAF-FYVE bigram. A possible function is the targeting of proteins to lipid layers by the zinc finger domain in response to second messengers sensed by the GAF domain.

Several domains involved in the phospholipid signaling were found to be overrepresented in the filamentous plant pathogens and in particular in oomycetes. These included the phosphatidylinositol 3-/4-kinase, PIK (IPR000403), the phosphatidylinositol 4-phosphate 5-kinase domain, PIPK (IPR002498), as well as the phosphatidylinositol 3-phosphate-binding FYVE. Novel domain compositions in proteins involved in phospholipid signaling and metabolism in *Phytophthora* species have been reported previously (Meijer & Govers 2006). Signaling domains like the FYVE and the PIK, as well as domains like the IQ-calmodulin-binding domain (IPR000048) and the phox-like domain (IPR001683), form highly abundant oomycete-specific bigram types (Supplementary Table S2-10). Moreover, other domains, like the Ser/Thr protein kinase-like (IPR017442), pleckstrin homology (IPR001849), and DEP (IPR000591) domains, are involved in several oomycete-specific bigram types (e.g. the DEP-Ser/Thr protein kinase-like domain fusion is predicted in the proteomes of all analyzed oomycetes). Additionally, domains that are components of the histone acetylation-based regulatory system form oomycete-specific bigrams, such as the AP2 (IPR001471) and the histone deacetylase (IPR000286) domain combination (Iyer et al. 2008), which is observed in *P. ramorum* as well as in *P. sojae*.

## DISCUSSION

We predicted the domain repertoire encoded in the genomes of four oomycete

plant pathogens and compared it with a broad variety of eukaryotes spanning all major groups, including several fungal plant pathogens that have a similar morphology, lifestyle, and ecological niche as oomycete plant pathogens. We quantified and examined domain properties observed in oomycetes and especially emphasized differences and common themes within fungal and oomycete plant pathogens and their probable contribution to a pathogenic lifestyle.

We observed that oomycete plant pathogens, in particular *Phytophthora* species, have significantly higher numbers of unique bigram types compared with species with a similar number of domain types (Figure 2-3). However, oomycetes also have on average 50% more predicted genes than most of the analyzed fungi, but at the same time they encode a comparable number of domain types and hence exhibit similar domain diversity (Supplementary Figure S2-1B). The high number of genes observed in oomycetes suggests enlarged complexity compared with fungi, which is not directly obvious from the domain diversity but instead from the number of unique bigram types (Supplementary Figure S2-1C). This observation has two possible explanations: (1) the larger number of genes predicted from oomycete genomes provides the flexibility to form new domain combinations based on a limited set of already existing domains that are in quantities similar to fungi; (2) the domain models that cover specific domains are incomplete and therefore do not provide the required sensitivity for oomycete genomes. Hence, we would underestimate the number of observable domain types (and to a certain extent the number of predicted bigram types). Additionally, oomycetes, especially *Phytophthora* species, are no longer following the observed trend that organisms with a higher number of genes (proteins) contain a larger number of domain types. Consequently, they are shifted when comparing the number of predicted domain and bigram types. Nevertheless, both possible explanations and the observed numbers allow us to conclude that oomycete genomes, especially *Phytophthora* species, harbor a large repertoire of genes encoding different bigram types compared with species of comparable complexity and, in the case of filamentous fungi, even similar morphology.

Oomycetes and fungal plant pathogens seem to be very similar to other eukaryotes with respect to absolute domain abundance (Supplementary Table S2-2), and this metric is hence not sufficiently indicative to correlate domains directly or indirectly with the pathogenic lifestyle. Therefore, we predicted overrepresented domains in plant pathogens and identified 246 domains that are significantly expanded (Supplementary Table S2-3). Proteins containing overrepresented domains are significantly enriched in the predicted secretome of the analyzed plant pathogens, corroborating the idea that expanded domain families are involved in host-pathogen interaction and that these proteins are mainly acting in the extracellular space. It has to be noted that the presence of a predicted signal peptide does not necessarily mean that these proteins are found extracellularly, since some proteins are retained in the endoplasmic reticulum/Golgi and hence are not secreted (Bendtsen et al. 2004).

Since we anticipate that proteins that are directly involved in host-pathogen interaction are differentially regulated upon infection, we utilized the NimbleGen microarray

data of *P. infestans* (Haas et al. 2009) and identified 259 induced/repressed genes encoding proteins containing overrepresented domains. Genes containing overrepresented domains are significantly enriched within the set of differentially expressed genes containing a predicted domain. Moreover, this subset contains a significantly higher abundance of genes with a predicted N-terminal signal peptide than expected. These observations highlight and corroborate the initially emerging link between domain expansion and host-pathogen interaction.

The majority of the 246 expanded domains are present in proteins that are involved in general carbohydrate metabolism, nutrient uptake, signaling networks, and suppression of host responses and hence might contribute to establishing and maintaining pathogenesis (Figure 2-4). The variety of overrepresented domains involved in substrate transport over membranes is of special interest. Filamentous plant pathogens and especially oomycetes exhibit a complex and expanded repertoire of these domains, enabling them to absorb nutrients from their environment and host. The expression of *P. infestans* genes encoding ABC-2-like transporters, amino acid transporters, and Na/Pi cotransporter is induced early in infection of the plant, suggesting that these proteins act during the biotrophic phase of infection. Several other genes encoding proteins with a predicted extracellular localization are induced during infection and contain overrepresented domains. For example, three *P. infestans* genes encoding the predicted N-terminal signal peptide as well as FAD-linked oxidase and BBE domains are induced during infection. The BBE domain is involved in the biosynthesis of the alkaloid berberine (Facchini et al. 1996). Moy et al. (2004) showed that a soybean homolog of this gene is induced after infection with *P. sojae*. Molecular studies of proteins containing BBE domains in plants have indicated that several proteins containing these domains are in fact not secreted but instead targeted to specific alkaloid biosynthetic vesicles where the proteins accumulate (Amann et al. 1986; Kutchan & Dittrich 1995; Facchini et al. 1996). The expansion of domain families with potential direct or indirect roles in host-pathogen interaction in filamentous plant pathogens strongly suggests adaptation to their lifestyle at the genomic level.

In addition to known domains, the set of overrepresented domains also revealed domains that, as yet, have not been implicated in pathogenicity nor are functionally characterized. An example is the DUF953 domain, which, within plant pathogens, is mainly found in oomycetes. This domain is observed in eukaryotic proteins with a thioredoxin-like function, and *P. infestans* genes encoding these domains are differentially expressed during infection. The significant expansion of these domains in plant pathogens, and the fact that other well-described domains with a function in plant pathogenicity are also overrepresented, make proteins encoding poorly described but expanded domains interesting candidates to decipher their role in filamentous plant pathogens in general and oomycetes in particular.

We determined domain overrepresentation on the basis of species groups (plant pathogens and oomycetes) rather than on the level of individual species. We are aware that, as a consequence of this approach, we might have identified domains as being

overrepresented in one group even if they do not need to be present or expanded in all the members (Supplementary Tables S3 and S5). Hence, we might falsely extrapolate the functional role of a domain in a subset of species to the whole group (e.g. a domain that is exclusively found in plant pathogenic fungi and not in oomycetes would still be overrepresented in the plant pathogenic group). Especially when comparing oomycete with fungal plant pathogens, the dominant expansion of domain families within *Phytophthora* species over families in *H. arabidopsidis* might bias the inferred overrepresented domain (Supplementary Table S2-5). Since we in general want to identify candidate domains that might be directly or indirectly involved in host-pathogen interaction, either at the level of filamentous plant pathogens or oomycetes, we think our group-based approach is appropriate to establish a set of candidate proteins and domains.

Moreover, the clustering of presence, absence, and expansion patterns of domains known or implicated to be involved in a plant pathogenic lifestyle with domains that have no known or direct connection to host-pathogen interactions aids in expanding this set of novel candidate domains (Figure 2-6). For example, DUF1949 is within our species selection exclusively found in *Phytophthora* species and adopts a ferredoxin-like fold. The N-terminal region of proteins containing this domain shows similarity to another domain (UPF00029) that has been found in the human Impact protein. The *P. infestans* gene containing both domains is induced early during infection of the plant, providing additional, independent evidence for the possible role of genes encoding this uncharacterized domain in host-pathogen interaction. However, domains that are also abundant in nonpathogenic species (e.g. other stramenopiles) might not be related to or only indirectly involved in pathogenicity. Hence, the exact nature of the contribution of these domains to pathogenesis or to general lifestyle requires more in-depth experimental studies of the candidate domains and genes predicted to contain these functional entities.

Protein domains generally do not act as single entities but in synergy with other domains in the same protein or with other domains in interacting proteins. We identified 773 oomycete-specific bigrams, of which 53 are observed in all analyzed oomycetes (Figure 2-7A; Supplementary Table S2-10). Based on our species selection, we cannot conclude that the oomycete-specific bigrams are common to all oomycetes, since they might only be specific for plant pathogenic oomycetes or even for the selected oomycetes analyzed in this study. The majority of the 773 bigrams, however, are specific for a subset of the tested oomycete species or even a single species. The 320 bigram types that are observed in more than a single species or twice in the same proteome are observed in 982 predicted proteins. These bigrams are less likely to be the result of a wrong gene annotation and include already well-described examples of oomycete-specific domain combinations, such as the FYVE-PIK bigram observed in *Phytophthora* phosphatidylinositol kinases (Meijer & Govers 2006), the AP2-histone deacetylase bigram that is specifically found in *P. ramorum* and *P. infestans* (Iyer et al. 2008), and the myosin head domain-FYVE bigram as well as the FYVE-GAF bigram found in myosin proteins in all analyzed oomycetes (Richards & Cavalier-Smith 2005). Still, some of



the bigrams could be artificial due to false negatives or false positives in the domain predictions. The remaining, species-specific bigrams could be the result of artificial fusion of genes due to wrong gene annotation or an actual biological signal in one of the analyzed oomycete species. The derived results are not only dependent on the quality of the genome sequences of the analyzed oomycetes but also on that of the other eukaryotes. Wrong predictions of bigrams in these species would lead to false negatives in oomycetes. Hence, the number of derived oomycete-specific bigrams is only an approximation, and the true set of oomycete-specific bigrams needs to be further analyzed. Recent analyses of the underlying molecular mechanisms of domain gain in animals have shown that in fact gene fusion, tightly linked with gene duplication, is the major mechanism that shaped novel protein architecture (Buljan et al. 2010; Marsh & Teichmann 2010). The contributions of this mechanism in forming lineage- or even species-specific bigrams in oomycetes and the probable role of the flexible genomes have to be further analyzed. The bigrams presented here form a comprehensive starting point for an in-depth bioinformatic and experimental analysis of promising gene families coding novel domain combinations.

Common domain types form the majority of the observed oomycete-specific bigrams, emphasizing the importance of novel combinations rather than novel domain types as a source for species-specific functionality. Only a minority of proteins containing oomycete-specific bigrams are secreted, and none of these proteins is predicted to contain a RXLR or Crn motif. We are aware that the total number of predicted proteins containing the RXLR or Crn motif is lower than reported in other studies where those were predicted using multiple complementary methods (Haas et al. 2009). However, when directly comparing the number of proteins predicted to contain the RXLR motif by HMMER alone, the reported numbers are similar to our predictions. Together with the observation that RXLR proteins do not contain known Pfam-A domains in the C-terminal domain (Haas et al. 2009), our data are not in conflict with RXLR protein predictions from previous studies. Of the known Crn genes in *P. infestans*, 40% do not encode a secretion signal (Haas et al. 2009); hence, these sequences are not considered in the prediction of Crn motifs in our analysis and explain the discrepancy between the previously reported numbers and our predictions. Haas et al. (2009) have reported a huge number of different C-terminal structures in *P. infestans* Crns that contained up to 36 different domains, of which 33 are not described in Pfam. Several of these domains induce necrosis in plants. Since we focused in our analysis exclusively on Pfam domains, we did not expect to find these proteins containing specific bigrams.

The majority of proteins containing oomycete-specific bigrams seem to be functional in the pathogen cytoplasm. Moreover, domains involved in mediation between macromolecules or lipids (e.g. the FYFE or the phox-like domain) as well as signaling domains (e.g. Ser/Thr kinase-like or the DEP domain) are highly abundant in oomycete-specific bigrams. Ser/Thr kinase domain-like is overrepresented in oomycetes compared with fungal plant pathogens and is particularly expanded within the *Phytophthora* species (Supplementary Table S2-5). This expanded repertoire together with the high abundance of this domain in oomycete-specific bigrams strongly suggests that oomycetes

have the capacity to recombine existing signaling pathways in a novel and complicated network that is distinct from other eukaryotes. This might also be true for other interaction networks, since several domains mediating interactions between macromolecules (e.g. DNA-binding zinc finger [IPR007087] or protein-protein interaction like WW/Rsp5/WWP [IPR001202]) are also highly abundant in oomycete-specific bigrams. Whether this reflects a general phenomenon in all oomycetes, specific for the plant pathogenic species analyzed in this study, or only for *Phytophthora* species, can only be answered when more oomycetes, including saprophytes and pathogens with different hosts, are sequenced.

We outlined a complex but comprehensive picture of the domain repertoire of filamentous plant pathogens focusing on oomycetes and showed how differences compared with other eukaryotes are reflecting the biology of these groups of organisms. Especially the expansion of certain domain families is directly linked with the lifestyle of oomycete plant pathogens and allowed the generation of a set of candidate domains likely to play important roles in the interaction with the plant host. Proteins containing overrepresented domains are enriched in the predicted secretome of the analyzed species. Moreover, the expression analysis of genes encoding domains during infection of the plant revealed a significant enrichment of genes encoding overrepresented domains within the differentially expressed genes. Furthermore, we observed a significantly higher than expected abundance of genes encoding a signal peptide within the set of differentially expressed genes containing expanded domains. This added additional, independent evidence for the biological significance of our observations. Furthermore, oomycete genomes encode a set of proteins containing oomycete-specific domain combinations that are formed by common domain types and include several domains involved in signaling and/or mediation of interactions between macromolecules. Oomycetes, therefore, might possess altered regulatory and signaling networks that differ from other eukaryotes. If the described and discussed differences in the domain repertoire of oomycetes have a direct influence on plant pathogenicity or are generally useful in these organisms needs to be analyzed further. Nevertheless, they provide promising starting points that will aid our understanding of the biology of oomycetes in general and plant pathogens in particular.

## MATERIAL & METHODS

### Species Used in the Analysis

In the performed analysis, 67 eukaryotic species representing four of the five eukaryotic supergroups (excluding Rhizaria) were considered (Figure 2-1A; for species abbreviations, see Supplementary Table S2-1). We used the predicted best model proteomes for all subsequent analyses.

### Identification of Domain Composition

We predicted the domain repertoire of all proteins encoded in the diverse genomes using hmmpfam (HMMER package version 2.3.2) and a local Pfam-A database (version 23). We applied a domain model-specific gathering cutoff and used HMM models that are optimized to search for full-length entities in the query sequence.

In order to obtain the non-overlapping domain architecture of multidomain proteins, we resolved overlapping domains according to certain rules. We defined two domains as overlapping if more than 10% of the predicted domain locations were overlapping (based on the relative length of the domains). If, in the case of overlapping domains, the e-value difference was larger than 5 (on a  $-\log_{10}$  scale), we kept the domain with the highest e-value. In cases where the difference was smaller, we kept the longest model. If both overlapping models had the same length, we considered differences in e-value and bit score. In the case of the Pfam-based predictions for 15 proteins, the applied rules did not resolve overlapping entities. Therefore, we considered the Conserved Domain Database (version 2.16) superfamily annotation, which automatically clusters domain entities that resemble evolutionarily related domains. If both domains corresponded to the same family, we choose one entity.

Based on the nonoverlapping domain architecture, we derived different metrics for each proteome. We counted the abundance for every domain and the resulting number of different domain types per analyzed proteome. We defined domain bigrams as two consecutively located domains in a single protein. We discriminated between reciprocal domain pairs, so that the bigram (A|B) is not identical to (B|A), and took repeating domains into account, such as (A|A). Based on the set of bigrams, we also determined the versatility of all individual domains in a given proteome, which is defined by the number of different direct N- and C-terminal partners, also including reciprocal and self-repeated pairs.

### Prediction of Secreted Proteins

Secreted proteins were predicted using SignalP (version 3.0; Bendtsen et al. 2004) in combination with TMHMM (version 2.0; Krogh et al. 2001). We restricted the analysis to the first 70 amino acids of the protein and accepted signal peptide predictions if both the neural network and the HMM implemented in SignalP predicted the presence of a signal peptide under default parameters. Moreover, we declined predicted signal peptides if TMHMM predicted more than one transmembrane region in the protein. If only a single transmembrane helix was predicted and the predicted region was overlapping with the SignalP prediction for more than 10 amino acids and positioned within the first 35 amino acids from the start, we included the protein in the set of secreted proteins.

### Domain Overrepresentation

Domain overrepresentation was calculated using a one-sided Fisher's exact test. The

derived P values were Bonferroni-corrected for multiple testing by multiplying the P value with the number of conducted tests. The corrected P values were compared with an  $\alpha = 0.001$  to infer domain overrepresentation. For the overrepresented domains in oomycete plant pathogens compared with fungal plant pathogens, we considered domains that occur at least once in a single plant pathogen but nevertheless could also occur in other eukaryotic species.

### Gene Expression Analysis of *Phytophthora infestans*

We extracted NimbleGen expression data of *P. infestans* during infection of potato (*Solanum tuberosum*) 2 to 5 dpi from the Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>). The setup and initial analysis of the NimbleGen data are described by Haas et al. (2009). The  $\log_2$ -transformed and mean-centered array intensities were analyzed for differential expression using Multiexperiment Viewer (Saeed et al. 2006). The t tests were conducted between two groups (group A, different media types; group B, replicates for one of the days post inoculation). The test was applied for each day after inoculation, and significant up-/ down-regulated genes were reported applying a P values cutoff of 0.05. False discovery rates were addressed using R and the qvalue package by computing q values for each of the comparisons and subsequently applying a q value cutoff of 0.05 (Storey & Tibshirani 2003). Visualization of the heat maps was done using R and the Bioconductor package utilizing Spearman correlation as a distance measurement and hierarchical clustering (average linkage; Gentleman et al. 2004). Gene expression intensities relative to the average expression intensities in media types (V8, RS, Pea) were computed in R.

### Clustering of Domain Profiles

We created abundance profiles for each domain based on the abundance in each individual proteome. We excluded domains that were only identified in a single species. The rows (domains) were multiplied by a scaling factor so that the sum of squares was 1, and subsequently the columns (species) were normalized in the same way. We performed a hierarchical clustering (average linkage) of the profiles using the Spearman correlation matrix as a distance measurement. The normalization and clustering were performed using Cluster (Eisen et al. 1998), and the visualization was done using TreeView (<http://rana.lbl.gov/EisenSoftware.htm>).

### Domain Promiscuity

We calculated the domain promiscuity for every domain in the analyzed species based on weighted bigram frequency (Basu et al. 2008). We took a relatively moderate cutoff for determining promiscuous domains; every domain with a higher promiscuity score than a domain that is only present once in the genome and is participating in one bigram type is called promiscuous.

### Prediction of the RXLR and Crn Motifs in Oomycetes

We identified the presence of the RXLR motif in all predicted proteins in the analyzed oomycetes using three different HMMER models (R.H.Y. Jiang, personal communication). The first model was created using *P. ramorum* and *P. sojae* RXLRs and included the RXLR motif itself and 10 amino acids downstream and upstream of the motif. The two other models were based separately on RXLRs from *P. infestans* and *H. arabidopsidis* and included 10 amino acids upstream from the RXLR motif and five amino acids downstream of the DEER motif. We used HMMER (hmmsearch) with an e-value cutoff of 10 and subsequently combined all predictions. Furthermore, we demanded the presence of a predicted signal peptide (SignalP) cleavage site within the first 30 amino acids of the protein, the gap between the cleavage site and the start of the motif to be 30 or less, the start of the motif to be within the first 100 amino acids of the protein, and the starting position of the RXLR motif to be downstream of the cleavage site. For the identification of the Crn LFLAK motif, we used a HMMER model of that region (B.J. Haas, personal communication) and the same sequence demands as for the RXLRs.

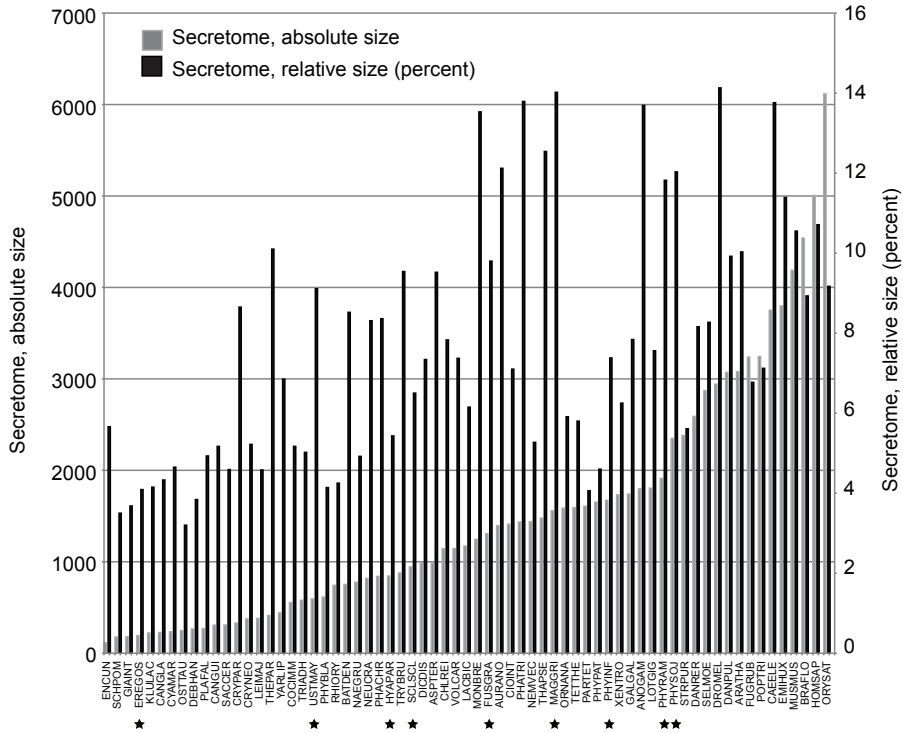
### Phylogenetic Analysis of the GAF Domain

We derived all sequences containing a GAF domain from the selected proteomes and extracted the amino acid sequence of the domain based on the start and end points of the domain model. We conducted a similarity search with the extracted domains using BLASTP (version 2.2.20) with an e-value cutoff of  $1 \times 10^{-5}$  and a low-complexity filter against a set of 295 bacterial predicted proteomes (downloaded from the National Center for Biotechnology Information ftp server on January 27, 2009). In the homologs that were obtained, domains were predicted using hmmpfam as described above. Subsequently, prokaryotic GAF domains were extracted and aligned together with the eukaryotic domains using mafft (version 6.713b) with the local alignment strategy (Kato et al. 2002). A phylogenetic tree was constructed with RAxML (version 7.0.4) using the GAMMA model of rate heterogeneity and the WAG amino acid substitution matrix (Stamatakis 2006).

## ACKNOWLEDGMENTS

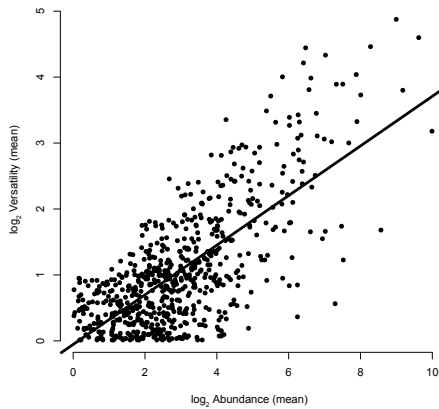
We thank Lidija Berke, John van Dam, and Jos Boekhorst for fruitful discussion and comments on the manuscript as well as Rui Peng Wang for support with the *P. infestans* gene expression data. We also thank Harold J.G. Meijer for discussion of fusion proteins in *P. infestans*, Rays H.Y. Jiang for providing the RXLR-HMMER model, and Brian J. Haas for the Crn LFLAK-HMMER model. Some of the sequence data and annotation were produced by the U.S. Department of Energy Joint Genome Institute (<http://www.jgi.doe.gov>), the Broad Institute of Harvard and the Massachusetts Institute of Technology (<http://www.broadinstitute.org>), or the Stanford Genome Technology Center (<http://med.stanford.edu/sgtc/>) in collaboration with the user community (for detailed information, see Supplementary Table S2-1).





**Figure S2-2** The predicted secretome of the 67 analyzed eukaryotic species.

The absolute size as well as the percentage of the predicted secretome is displayed. The analyzed plant pathogens are indicated with a star and species abbreviations are shown in Supplementary Table S2-1.



**Figure S2-3** The average abundance and versatility of all observed domains in a  $\log_2$ - $\log_2$  plot.

The graph displays a linear positive correlation between the abundance and versatility of different domains (regression line in black). Domains that are highly abundant and do not have a high number of different N- or C-terminal partners are shown in the lower right sector of the plot. Domains that show an uneven distribution of versatility in the examined species might have a low average versatility, even if they have many different partners in some species

## REFERENCES

- Agrios GN. 2005. Plant Pathology. 5th ed. Academic Press, New York.
- Amann M, Wanner G, Zenk MH. 1986. Intracellular compartmentation of two enzymes of berberine biosynthesis in plant cell cultures. *Planta*. 167:310–320.
- Aravind L, Ponting CP. 1997. The GAF domain: an evolutionary link between diverse phototransducing proteins. *Trends Biochem. Sci.* 22:458–459.
- 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.
- Bashton M, Chothia C. 2007. The generation of new protein functions by the combination of domains. *Structure*. 15:85–99.
- Basu MK, Carmel L, Rogozin IB, Koonin EV. 2008. Evolution of protein domain promiscuity in eukaryotes. *Genome Res.* 18:449–461.
- Baurain D et al. 2010. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Mol. Biol. Evol.* 27:1698–1709.
- Bendtsen JD, Nielsen H, Heijne von G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 340:783–795.
- Blair JE, Coffey MD, Park S-Y, Geiser DM, Kang S. 2008. A multi-locus phylogeny for *Phytophthora* utilizing markers derived from complete genome sequences. *Fungal Genet. Biol.* 45:266–277.
- Buljan M, Frankish A, Bateman A. 2010. Quantifying the mechanisms of domain gain in animal proteins. *Genome Biol.* 11:R74.
- Cosentino Lagomarsino M, Sellerio AL, Heijning PD, Bassetti B. 2009. Universal features in the genome-level evolution of protein domains. *Genome Biol.* 10:R12.
- Dean RA et al. 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*. 434:980–986.
- Doolittle RF. 1995. The multiplicity of domains in proteins. *Annu. Rev. Biochem.* 64:287–314.
- Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics*. 14:755–763.
- Eisen MB, Spellman PT, Brown PO, Botstein D. 1998. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U.S.A.* 95: 14863–14868
- Erwin DC, Ribeiro OK. 1996. *Phytophthora* diseases worldwide. American Phytopathological Society (APS Press).
- Facchini PJ, Penzes C, Johnson AG, Bull D. 1996. Molecular Characterization of Berberine Bridge Enzyme Genes from Opium Poppy. *Plant Physiol.*
- Finn RD et al. 2010. The Pfam protein families database. *Nucleic Acids Res.* 38:D211–22. doi: 10.1093/nar/gkp985.
- Gentleman RC et al. 2004. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.* 5:R80.
- Gijzen M, Nürnberger T. 2006. Nep1-like proteins from plant pathogens: recruitment and diversification of the NPP1 domain across taxa. *Phytochemistry*. 67:1800–1807.
- Govers F, Gijzen M. 2006. *Phytophthora* genomics: the plant destroyers' genome decoded. *Mol. Plant Microbe Interact.* 19:1295–1301.
- Groves MR et al. 1996. The prosequence of procaricain forms an alpha-helical domain that prevents access to the substrate-binding cleft. *Structure*. 4:1193–1203.
- Haas BJ et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature*. 461:393–398.



- He SY, Collmer A. 1990. Molecular cloning, nucleotide sequence, and marker exchange mutagenesis of the exo-poly-alpha-D-galacturonidase-encoding *pehX* gene of *Erwinia chrysanthemi* EC16. *J. Bacteriol.* 172:4988–4995.
- Inohara N, Nuñez G. 2002. ML - a conserved domain involved in innate immunity and lipid metabolism. *Trends Biochem. Sci.* 27:219–221.
- Itoh T, Hashimoto W, Mikami B, Murata K. 2006. Substrate recognition by unsaturated glucuronyl hydrolase from *Bacillus* sp. GL1. *Biochem. Biophys. Res. Commun.* 344:253–262.
- Iyer LM, Anantharaman V, Wolf MY, Aravind L. 2008. Comparative genomics of transcription factors and chromatin proteins in parasitic protists and other eukaryotes. *Int. J. Parasitol.* 38:1–31.
- James TY, et al (2006) Reconstructing the early evolution of fungi using a six-gene phylogeny. *Nature* 443: 818–822.
- Katoh K, Misawa K, Kumar S, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.
- Kawamukai M et al. 1991. Nucleotide sequence and characterization of the *sfs1* gene: *sfs1* is involved in CRP\*-dependent mal gene expression in *Escherichia coli*. *J. Bacteriol.* 173:2644–2648.
- Keeling PJ, Palmer JD. 2008. Horizontal gene transfer in eukaryotic evolution. *Nat. Rev. Genet.* 9:605–618.
- Klis FM, Sosinska GJ, de Groot PWJ, Brul S. 2009. Covalently linked cell wall proteins of *Candida albicans* and their role in fitness and virulence. *FEMS Yeast Res.* 9:1013–1028.
- Krogh A, Larsson B, Heijne von G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305:567–580.
- Kulkarni RD, Kelkar HS, Dean RA. 2003. An eight-cysteine-containing CFEM domain unique to a group of fungal membrane proteins. *Trends Biochem. Sci.* 28:118–121.
- Kutchan TM, Dittich H. 1995. Characterization and mechanism of the berberine bridge enzyme, a covalently flavinylated oxidase of benzophenanthridine alkaloid biosynthesis in plants. *J. Biol. Chem.* 270:24475–24481.
- Latijnhouwers M, de Wit PJGM, Govers F. 2003. Oomycetes and fungi: similar weaponry to attack plants. *Trends Microbiol.* 11:462–469.
- Lévesque CA et al. 2010. Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire. *Genome Biol.* 11:R73.
- Liu Z et al. 2005. Patterns of diversifying selection in the phytotoxin-like *scr74* gene family of *Phytophthora infestans*. *Mol. Biol. Evol.* 22:659–672.
- Marcotte EM et al. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science.* 285:751–753.
- Marsh JA, Teichmann SA. 2010. How do proteins gain new domains? *Genome Biol.* 11:126.
- Martens C, Vandepoele K, Van de Peer Y. 2008. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proc. Natl. Acad. Sci. U.S.A.* 105:3427–3432.
- Martinez SE, Beavo JA, Hol WGJ. 2002. GAF domains: two-billion-year-old molecular switches that bind cyclic nucleotides. *Mol. Interv.* 2:317–323.
- McLeod A, Smart CD, Fry WE. 2003. Characterization of 1,3-beta-glucanase and 1,3;1,4-beta-glucanase genes from *Phytophthora infestans*. *Fungal Genet. Biol.* 38:250–263.
- Meijer HJG, Govers F. 2006. Genomewide Analysis of Phospholipid Signaling Genes in *Phytophthora* spp.: Novelty and a Missing Link. *Mol. Plant Microbe Interact.* 19:1337–1347.
- Montgomery BL, Lagarias JC. 2002. Phytochrome ancestry: sensors of bilins and light. *Trends Plant Sci.* 7:357–366.
- Morris PF et al. 2009. Multiple horizontal gene transfer events and domain fusions have created novel regu-

- latory and metabolic networks in the oomycete genome. *PLoS ONE*. 4:e6133.
- Moy P, Qutob D, Chapman BP, Atkinson I, Gijzen M. 2004. Patterns of gene expression upon infection of soybean plants by *Phytophthora sojae*. *Mol. Plant Microbe Interact.* 17:1051–1062.
- Okamura K et al. 2000. Comparative genome analysis of the mouse imprinted gene *Impact* and its nonimprinted human homolog *IMPACT*: toward the structural basis for species-specific imprinting. *Genome Res.* 10:1878–1889.
- Oliveros JC (2007) VENNY: An Interactive Tool for Comparing Lists with Venn Diagrams. <http://bioinfogp.cnb.csic.es/tools/venny/index.html> (October 7, 2010)
- Orengo CA et al. 1997. CATH—a hierarchic classification of protein domain structures. *Structure*. 5:1093–1108.
- Orsomando G et al. 2001. Phytotoxic protein PcF, purification, characterization, and cDNA sequencing of a novel hydroxyproline-containing factor secreted by the strawberry pathogen *Phytophthora cactorum*. *J. Biol. Chem.* 276:21578–21584.
- Park F et al. 2004. Crystal structure of YIGZ, a conserved hypothetical protein from *Escherichia coli* K12 with a novel fold. *Proteins*. 55:775–777.
- Park J, Lappe M, Teichmann SA. 2001. Mapping protein family interactions: intramolecular and intermolecular protein family interaction repertoires in the PDB and yeast. *J. Mol. Biol.* 307:929–938.
- Pathak D, Ashley G, Ollis D. 1991. Thiol protease-like active site found in the enzyme dienelactone hydrolase: localization using biochemical, genetic, and structural tools. *Proteins*. 9:267–279.
- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.* 96:4285–4288.
- Raffaele S, Win J, Cano LM, Kamoun S. 2010. Analyses of genome architecture and gene expression reveal novel candidate virulence factors in the secretome of *Phytophthora infestans*. *BMC Genomics*. 11:637.
- Richards TA, Cavalier-Smith T. 2005. Myosin domain evolution and the primary divergence of eukaryotes. *Nature*. 436:1113–1118.
- Richards TA, Talbot NJ. 2007. Plant parasitic oomycetes such as *Phytophthora* species contain genes derived from three eukaryotic lineages. *Plant Signal Behav.* 2:112–114.
- Rossmann MG, Moras D, Olsen KW. 1974. Chemical and biological evolution of nucleotide-binding protein. *Nature*. 250:194–199.
- Ruttkowski E et al. 1990. Cloning and DNA sequence analysis of a polygalacturonase cDNA from *Aspergillus niger* RH5344. *Biochim. Biophys. Acta.* 1087:104–106.
- Saeed AI et al. 2006. TM4 microarray software suite. *Meth. Enzymol.* 411:134–193.
- Sharrock RA, Quail PH. 1989. Novel phytochrome sequences in *Arabidopsis thaliana*: structure, evolution, and differential expression of a plant regulatory photoreceptor family. *Genes Dev.* 3:1745–1757.
- Simpson AGB, Roger AJ (2004). The real ‘kingdoms’ of eukaryotes. *Curr Biol* 14: R693–R696.
- Stamatakis A. 2006. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*. 22:2688–2690.
- Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100:9440–9445.
- Terashima H et al. 2002. Sequence-based approach for identification of cell wall proteins in *Saccharomyces cerevisiae*. *Curr. Genet.* 40:311–316.
- Tyler BM et al. 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science*. 313:1261–1266.
- Vogel C, Bashton M, Kerrison ND, Chothia C, Teichmann SA. 2004. Structure, function and evolution of multi-domain proteins. *Curr. Opin. Struct. Biol.* 14:208–216.
- Vogel C, Teichmann SA, Pereira-Leal J. 2005. The relationship between domain duplication and recombina-

- tion. *J. Mol. Biol.* 346:355–365.
- Yang YD et al. 2008. TMEM16A confers receptor-activated calcium-dependent chloride conductance. *Nature.* 455:1210–1215.
- Yin QY. 2005. Comprehensive proteomic analysis of *Saccharomyces cerevisiae* cell walls: Identification of proteins covalently attached via glycosylphosphatidylinositol remnants or mild alkali-sensitive linkages. *Journal of Biological Chemistry.* 280:20894–20901.
- Yoon HS, Hackett JD, Pinto G, Bhattacharya D. 2002. The single, ancient origin of chromist plastids. *Proc. Natl. Acad. Sci. U.S.A.* 99:15507–15512.
- Zoraghi R, Corbin JD, Francis SH. 2004. Properties and functions of GAF domains in cyclic nucleotide phosphodiesterases and other proteins. *Mol. Pharmacol.* 65:267–278.





# Reconstruction of Oomycete Genome Evolution Identifies Differences in Evolutionary Trajectories Leading to Present- Day Large Gene Families

3

Michael F Seidl<sup>1,2</sup>, Guido Van den  
Ackerveken<sup>2,3</sup>, Francine Govers<sup>2,4</sup>, and  
Berend Snel<sup>1,2</sup>

<sup>1</sup>Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Utrecht, The Netherlands

<sup>2</sup>Centre for BioSystems Genomics, Wageningen, The Netherlands

<sup>3</sup>Plant-Microbe Interactions, Department of Biology, Utrecht University, Utrecht, The Netherlands

<sup>4</sup>Laboratory of Phytopathology, Wageningen University, Wageningen, The Netherlands

*Genome Biol Evol.* **4**:199–211 (2012)  
Copyright Oxford University Press

## ABSTRACT

The taxonomic class of oomycetes contains numerous pathogens of plants and animals but is related to nonpathogenic diatoms and brown algae. Oomycetes have flexible genomes comprising large gene families that play roles in pathogenicity. The evolutionary processes that shaped the gene content have not yet been studied by applying systematic tree reconciliation of the phylome of these species. We analyzed evolutionary dynamics of ten Stramenopiles. Gene gains, duplications, and losses were inferred by tree reconciliation of 18,459 gene trees constituting the phylome with a highly supported species phylogeny. We reconstructed a strikingly large last common ancestor of the Stramenopiles that contained ~10,000 genes. Throughout evolution, the genomes of pathogenic oomycetes have constantly gained and lost genes, though gene gains through duplications outnumber the losses. The branch leading to the plant pathogenic *Phytophthora* genus was identified as a major transition point characterized by increased frequency of duplication events that has likely driven the speciation within this genus. Large gene families encoding different classes of enzymes associated with pathogenicity such as glycoside hydrolases are formed by complex and distinct patterns of duplications and losses leading to their expansion in extant oomycetes. This study unveils the large-scale evolutionary dynamics that shaped the genomes of pathogenic oomycetes. By the application of phylogenetic based analyses methods, it provides additional insights that shed light on the complex history of oomycete genome evolution and the emergence of large gene families characteristic for this important class of pathogens.

## INTRODUCTION

Recent comparative genome analyses of Stramenopiles have facilitated initial insights into the evolution and lifestyle of the individual species within this lineage and in particular of pathogenic oomycetes (Tyler et al. 2006; Martens et al. 2008; Haas et al. 2009; Gobler et al. 2011; Seidl et al. 2011). The extensive Stramenopile lineage comprises species that cover diverse ecological niches and lifestyles ranging from photosynthetic diatoms and brown algae to filamentous heterotrophic oomycetes. According to the controversial Chromalveolate hypothesis, Stramenopiles are grouped together with other chlorophyll-*c* containing lineages such as Cryptophytes, Alveolates, and Haptophytes into one monophyletic supergroup (Cavalier-Smith 1999; Keeling 2009), sometimes also referred to as CASH. This grouping has been rationalized on the hypothesis that the last common ancestor (LCA) of these lineages acquired its plastid from a single initial event of secondary endosymbiosis with a red alga that has been subsequently inherited strictly vertically. Consequently, plastid-lacking species within CASH lineages have lost their plastids secondarily and independently. The competing serial eukaryotic–eukaryotic endosymbiotic (SEEE) hypothesis proposes an independent spread of plastids within CASH lineages, and hence, dependent on the time point of acquisition, no secondary losses are needed to explain the lack of plastids in several taxa throughout all lineages (Cavalier-Smith et al. 1994; Archibald 2009; Baurain et al. 2010).

The plastid-lacking oomycetes are saprophytes or pathogens of plants and animals with huge economic as well as ecological impact (Govers & Gijzen 2006). Well known are the notorious late blight pathogen *Phytophthora infestans* that infects both tomato and potato and the animal pathogen *Saprolegnia parasitica* that causes saprolegniasis, for example, in salmon. Within the oomycetes studied so far, the genomes of *Phytophthora* spp. have by far the largest genomes, ranging from 65 up to 240 Mb (Supplementary Figure S3-1A). This broad variation in genome sizes is also observed among fungi, many of which are pathogens that exploit infection strategies similar to oomycetes (Latijnhouwers et al. 2003). Within Ascomycetes, for example, the rice blast fungus *Magnaporthe grisea* has a relatively small genome (38 Mb, ~12,000 predicted genes), whereas the recently sequenced genome of the obligate biotrophic powdery mildew fungus *Blumeria graminis* is considerably larger (~100 Mb); the expansion is mainly caused by transposable elements (Spanu et al. 2010; Duplessis et al. 2011). It has been speculated that *Phytophthora* spp. might have undergone a whole-genome duplication or at least several large-scale duplications. That, together with their divergent repertoire of transposable elements, probably contributed to the increased genome size and gene content of the *Phytophthora* spp. (Jiang et al. 2005; Haas et al. 2009; Martens & Van de Peer 2010).

Oomycete pathogens have a large and diverse repertoire of expanded gene families (Tyler et al. 2006; Haas et al. 2009; Baxter et al. 2010; Lévesque et al. 2010; Seidl et al. 2011). These mainly encode proteins that are secreted and implied to be directly or indirectly involved in pathogenicity, such as the NEP1-like proteins (Gijzen & Nürnberger 2006) or glycoside hydrolases (Ospina-Giraldo et al. 2010; Seidl et al. 2011). Two notable classes of highly abundant genes that are identified in several oomycete genom-

es encode secreted proteins characterized by the presence of either the RXLR or the LXLFLAK (Crinkler) motif (Whisson et al. 2007; Dou et al. 2008; Jiang et al. 2008; Haas et al. 2009). These motifs, located in the N-terminal region of the mature protein, play a role in translocation of the proteins from the apoplast to the cytoplasm of the host cell; however, the process is not yet fully understood (Govers & Bouwmeester 2008; Kale et al. 2010; Stassen & Van den Ackerveken 2011).

Initial analyses of the evolution of several pathogenic oomycetes led to the identification of large gene families. However, the individual contributions and the exact sequence of different evolutionary processes such as gene gains, duplications, and losses that caused the enormous increase in gene families sizes are still unknown. We studied these dynamics, and also the general evolution of the gene content, by a phylogenetic approach that reconciled 18,459 individual gene trees that constitute the phylome of Stramenopiles with a reliable species phylogeny. This systematic and comprehensive analysis of the evolutionary events is now feasible because several genomes of oomycetes and their sister lineages have been sequenced, a substantial increase to previous studies. We have utilized the predicted proteomes of six pathogenic oomycetes and four nonpathogenic Stramenochromes (Supplementary Material and Methods S3-1), a sublineage within the Stramenopiles (Patterson 1999). The six oomycetes comprise the fish pathogen *S. parasitica* and five plant pathogens: the necrotrophic wide host range pathogen *Pythium ultimum*, the obligate downy mildew pathogen of *Arabidopsis* *Hyaloperonospora arabidopsidis*, and three *Phytophthora* species, *P. infestans*, *P. sojae*, and *P. ramorum*. The latter two cause stem and root rot on soybean and sudden oak death, respectively. The four aquatic photosynthetic Stramenochromes include the brown alga *Ectocarpus siliculosus*, the golden-brown alga *Aureococcus anophagefferens*, and two diatoms: *Phaeodactylum tricorutum* and *Thalassiosira pseudonana*. Our phylogeny-based approach resulted in an overview of the fundamental evolutionary dynamics underlying major transition points in the evolution of pathogenic oomycetes and how these differences are reflected in the expansion and contraction pattern of distinct functional classes, such as transcription regulation or carbohydrate metabolism. Moreover, we were able to elucidate the evolutionary history of large gene families in oomycetes, such as glycoside hydrolases and peptidases. These families show distinct evolutionary trajectories that caused their abundance in extant taxa, an observation that would not have been possible solely on parsimony- or abundance-based methods. This, together with our other results, highlights the needs for an advanced phylogeny-based analysis of the expansion of large gene families in the future.

## MATERIAL & METHODS

To define protein families in the ten analyzed Stramenopiles, we created a sparse network based on Blast (Altschul et al. 1990) all-versus-all sequence similarity search (e-value cut-off:  $1 \times 10^{-3}$ ). Spurious connections between short segments of similarity were removed, and the network was portioned into families using the Markov clustering algorithm (Van Dongen 2000; Enright et al. 2002). The presence of transposable



elements in the proteomes was predicted by two independent methods and families containing at least one identified transposable element were removed.

A maximum likelihood phylogenetic tree was inferred using RAxML (Stamatakis 2006) (v7.0.4) with a gamma model of heterogeneity and Whelan and Goldman amino acid substitution matrix. A phylogenetic marker was created by concatenation of individual alignments of single-copy families derived by mafft (Katoh et al. 2002) (L-INS-I algorithm). The robustness of the topology was assessed by 1,000 bootstrap replicates. Relative divergence times within the Stramenopiles were estimated with BEAST under a strict clock model (Drummond & Rambaut 2007). The age prior for the last Stramenopile common ancestor (LSCA) was arbitrarily set to 100. We ran ten independent chains with 4,000,000 generations and subsequently averaged the estimates on the relative divergence times. The probability of the deviation between the observed and the expected number of evolutionary events at each branch was assessed by Poisson distribution.

We aligned the individual protein families, subsequently constructed RAxML maximum likelihood trees and assessed the robustness of these with 100 bootstrap replicates. We used NOTUNG (Chen et al. 2006; Durand et al. 2006) (v2.6; 1.5 duplication and 1 loss cost) to reconcile these trees with the species phylogeny. Uncertainties in the protein tree topology were assessed and weakly supported branches (<80% bootstrap support) were rearranged to minimize duplication/loss costs. Orthologous groups (OGs) were formed based on duplications at the LSCA. Consequently, each OG represents a single gene at LSCA or at the point of gain. All OGs are deposited under [http://bioinformatics.bio.uu.nl/michael/index\\_supplementary.html](http://bioinformatics.bio.uu.nl/michael/index_supplementary.html).

Individual OGs were functionally annotated by transfer of clusters of orthologous groups (COG) classification from eggNOG (Muller et al. 2010), by functional annotation of chloroplast-associated proteins via gene ontology utilizing Blast2GO (Conesa et al. 2005), and by prediction of secretion signals and/or of host-cell translocation motifs (RxLR/ LxLFLAK) or based on differential expression of the encoding genes during infection of the host. The prediction of signature Pfam domains identified OGs containing glycoside hydrolases and peptidases. Significant over- or underrepresentation of evolutionary events was assessed using Fisher's exact test, and multiple testing correction was applied.

Complete information regarding all methods and material used for the analyses is reported in Supplementary Material and Methods S3-1.

## RESULTS

### Protein Family Assignment

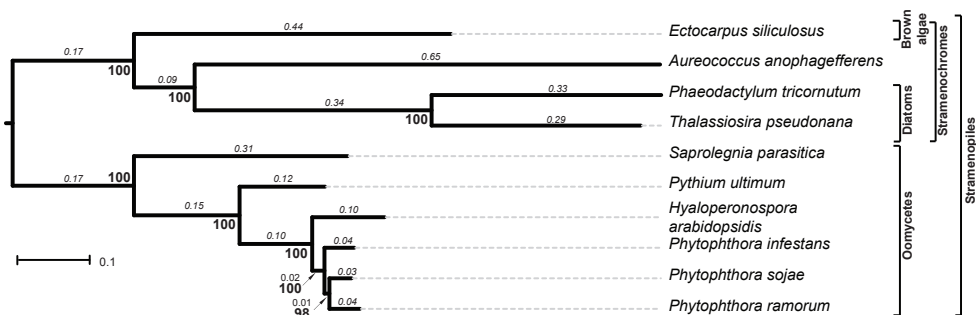
To systematically study the evolutionary dynamics of protein families in ten

Stramenopile species, we classified the combined set of 148,744 predicted proteins into families (Materials and Methods). In total, 18,979 families were formed, and for 27,342 single sequences (singletons), no homology could be established.

Filtering for transposable elements resulted in the removal of 7,905 proteins representing 519 families and 267 singletons. Stramenopiles, in particular oomycetes and the brown alga *E. siliculosus*, contain a large and diverse repertoire of transposable elements (Jiang et al. 2005; Tyler et al. 2006; Haas et al. 2009; Cock et al. 2010). Relics of those have been observed in high abundance in the predicted proteomes and would have biased our analysis (Seidl et al. 2011). In total, this resulted in 45,535 families including 27,075 singletons (Supplementary Figure S3-2A). Other large-scale studies conducted in closely related phyla revealed a comparable number of singletons per genome (Supplementary Figure S3-2B) (see e.g., Martens et al. 2008; Cock et al. 2010). However, a direct comparison is not feasible because different species sets were used in the other studies. The remaining 18,459 multisequence families were used for tree reconciliation.

### Species Phylogeny Utilizing Concatenated Single-Copy Genes

The quality of tree reconciliation is highly dependent on a correct species phylogeny. Furthermore, individual gene trees do not necessarily reflect the true relationship between species. In order to elucidate a reliable species phylogeny, we concatenated multiple families of single-copy genes, that is, families with only one member in each of the ten species included in this study (Figure 3-1). We concatenated alignments of 189 single-copy families and inferred the species phylogeny using a maximum likelihood approach implemented in RAxML (Stamatakis 2006). The robustness was assessed by 1,000 bootstrap replicates. The obtained species phylogeny is highly supported with bootstrap values >95% for all nodes. It mostly resembles the known topology of the tree of life, clearly separating the pathogenic oomycetes from the nonpathogenic Stramenochromes. However, the exact relationships within the genus *Peronosporales*



**Figure 3-1 Phylogeny of the analyzed Stramenopiles.**

Maximum likelihood phylogeny of the analyzed Stramenopiles based on 189 concatenated marker families (branch lengths in “substitutions per site” are displayed in italics). The robustness of the topology was assessed using 1,000 bootstrap replicates (bold numbers).

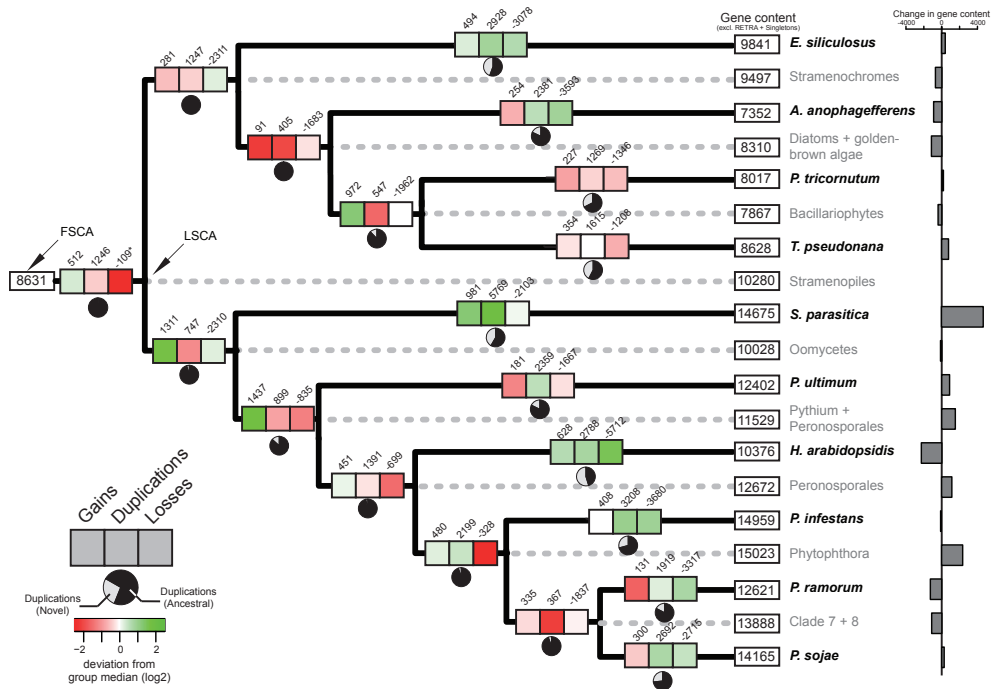
contradict previous studies that either grouped *P. sojae* and *P. infestans* (Blair et al. 2008) or proposed the paraphyly of *Phytophthora* by grouping *P. infestans* as a sister taxa to *H. arabidopsidis* (Runge et al. 2011). Our phylogenetic analysis revealed a closer relationship between *P. ramorum* and *P. sojae*, and we show that this topology is more parsimonious in reconciliation of evolutionary events; hence, it was used for all further analyses (Supplementary Figure S3-3A).

### Systematic Tree Reconciliation Guides Genome Reconstruction

We obtained a comprehensive and dynamic picture of Stramenopile genome evolution by projecting gene gains, duplications, as well as losses onto the species phylogeny (Figure 3-2). For each of the 18,459 families, we inferred maximum likelihood trees, reconciled these with the predicted species phylogeny of Stramenopiles, and subsequently formed 19,596 OGs that represent single genes either at the LSCA or at the respective point of gain of individual OGs. To appropriately describe the evolutionary events that can affect an OG, these groups can also contain genes that descend from a duplication event subsequent to the reference speciation event (LSCA in our case), so called in-paralogs: these genes are related to the other genes within the group with respect to the reference speciation event (LSCA) and are hence orthologous (Fitch 2000; Sonnhammer & Koonin 2002). Consequently, an OG can reflect single-copy orthologs, but also more complex 1:n, n:m relationships, and is used as such throughout the manuscript.

Over 50% of OGs are present in the LSCA. We found homologs outside of Stramenopiles for 95% (~9,750) of these groups, and hence, they predate the LSCA. Based on our data set, the reconstructed genome of the LSCA contained at least 10,280 genes and is consequently remarkably large compared with the genome content of the Stramenochromes. The genes present in the LSCA are enriched for basic cellular functions, like transcription and translation. It is striking to see that of the remaining gains, 30% is observed at the LCA of oomycetes and the LCA of the Pythium + Peronosporales clade (1,311 and 1,437, respectively); the highest number of gene gain observed at any branch ( $P$  value < 0.01, one-sided Wilcoxon rank sum test). This demonstrates that gains, accompanied by duplications, have caused the increase in genome content of pathogenic oomycetes.

Despite the fact that Stramenochromes, unlike pathogenic oomycetes, show only small net changes in the number of encoded proteins (Figure 3-2), their genomes are not static. Similar to oomycete genomes, they are in constant flux: High numbers of duplications are balanced by an equally high number of losses. The contribution of individual duplications and losses on the same branch and the effect on the size of the OG could not have been observed with parsimony-based methods because many of these duplications and losses occur in the same OG on the same branch. Globally, we observed an average of 1.77 duplication and 2.06 losses per OG; however, only few OGs contribute to the majority of evolutionary events (e.g. members of the major facilitator superfamily or amino acid transporters).



**Figure 3-2 Projected evolutionary events on the Stramenopile phylogeny.**

The number of evolutionary events, that is, gene gains, duplications, and losses, are projected onto each branch of the phylogeny. Pie charts indicate the relative contribution of novel or ancestral OGs to the total number of duplications. The heat map highlights the deviation of the number of events from the median of the class (gains, duplications, or losses). Predicted gene content of the ancestors, LSCA and first Stramenopile common ancestor, as well as of the extant taxa (excluding singletons and transposable elements) is displayed in terminal boxes, whereas the calculated change in gene content, that is, change in the number of genes per branch, is shown by bar charts.

To assess whether the observed duplication or loss events per individual branch deviate from the expected number, we calculated the relative frequency of these events per branch. Hence, we inferred branch length by estimating the relative divergence time of Stramenopiles using BEAST (Drummond & Rambaut 2007) and artificially dating the LSCA to 100 units of time (Supplementary Figure S3-3B). We predicted the position of the root by adding the ciliate *Paramecium tetraurelia* as an outgroup species. Based on the cumulative branch lengths (Supplementary Figure S3-3B) and the duplication and loss events (Figure 3-2), we estimated the relative frequency of duplications and losses to be 67 and 78 per unit of time, respectively. We contrasted the observed number of duplications/losses with expected numbers based on the global frequency and the length of the individual branch. The probability that the observed events deviate from the expectations was calculated using Poisson distribution. The abundance of observed duplications and losses significantly deviate from the expected number of events at each branch (Supplementary Figure S3-4). Within the Peronosporales clade, duplications and losses are significantly higher than expected (Supplementary Figure S3-4;

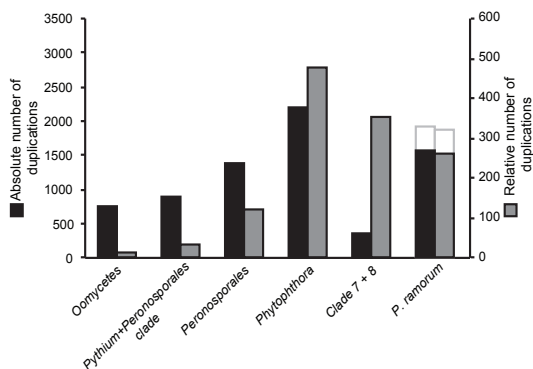
duplications up to a maximum of  $\sim 7$ -fold;  $2.83 \log_2$  fold), indicating an increased turnover of gene families in this clade. Interestingly, also at the LCA of Stramenochromes as well as the LCA of diatoms/golden-brown algae, the abundance of losses is significantly higher than expected, pointing to the contraction of OGs within the Stramenochromes.

A notable example of genome contraction is observed in the downy mildew *H. arabidopsidis*. An accumulation of losses is accompanied by a lower number of duplication events. It is the only branch in the phylogeny where the majority of duplications occurs in lineage-specific groups. Hence, the *H. arabidopsidis* genome encodes a unique repertoire of expanded OGs, while at the same time, ancestral OGs, that is, OGs that were already gained before the point of duplications, were either completely lost or contracted in size.

The increased genome content of the extant oomycetes is mainly caused by three events: gains, continuous duplications at internal branches of the species phylogeny, and a high number of duplications at branches leading to the extant taxa, for example, *P. infestans*, *S. parasitica*, and *P. ultimum*. Duplications at the LCAs are in general of lower abundance and affect ancestral OGs. A notable exception is the observed accumulation of duplications at the LCA of *Phytophthora* spp. ( $2.83$ -fold ( $\log_2$ ) higher than expected) (Figure 3-2; Supplementary Figure S3-4); this is 1.5 times higher compared with the other duplications at internal branches. The increased number of duplications is even more pronounced when considering the relative number of duplication events per branch instead of the absolute abundance and hence points to a major duplication event in the evolution of the *Phytophthora* genus (Figure 3-3 and Supplementary Figure S3-5).

### Differences in the Evolutionary Dynamics of Biologically Distinct Functional Classes

OGs can be assigned to functional classes by projecting the biological function of its individual proteins to the entire OG. We formed broad classes of functionally related OGs by transferring functional annotations from homologs based on the COG functional classification schema (Tatusov et al. 1997) and from predictions, for example, signal



**Figure 3-3** Number of duplication events in *P. ramorum*.

Absolute and relative numbers of duplication events for *P. ramorum* and all its ancestors. The absolute number of duplications is displayed in black, whereas the relative number of duplications (per unit of time) is shown in dark grey. The light grey bar represents the abundance of duplications (absolute and relative) including duplications occurring in lineage-specific OGs in *P. ramorum*.

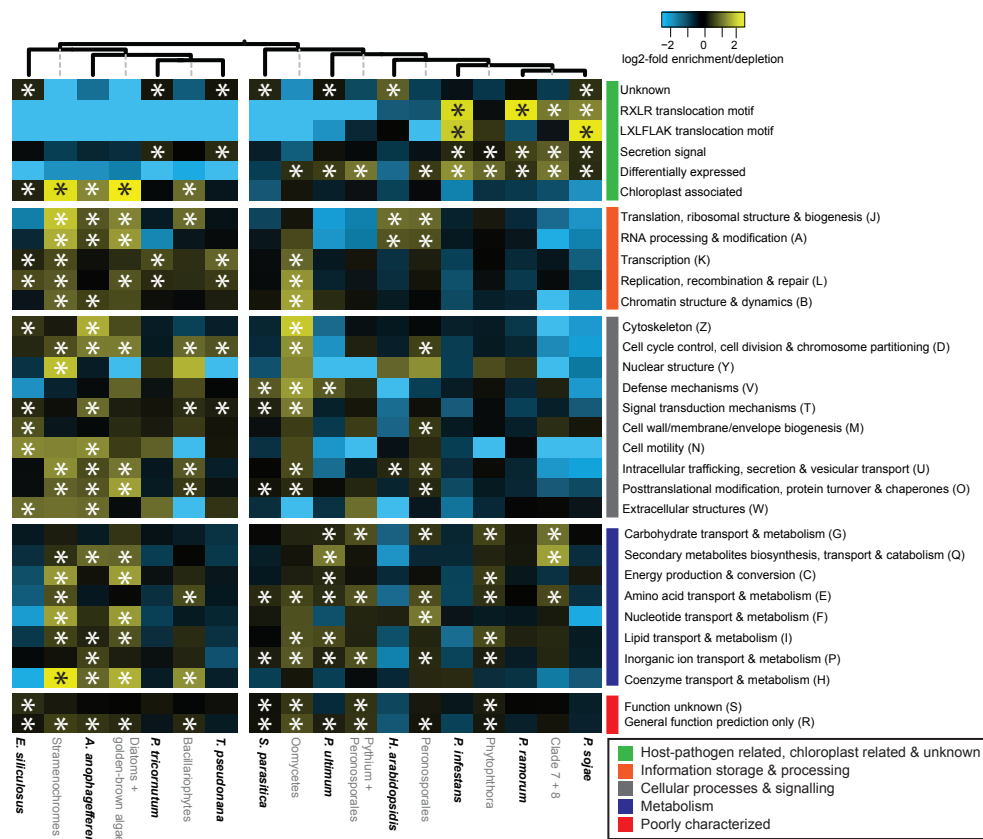
peptides or host cell translocation motifs (RXLR and LXLFLAK) (Supplementary Material and Methods S3-1). These broad functional classes behave strikingly different with respect to their evolutionary pattern of expansion (duplications) and contraction (losses): They are either significantly overrepresented or significantly underrepresented at various points in the evolution of Stramenopiles (Figure 3-4 and Supplementary Figure S3-6).

Overall, OGs belonging to COG 'information processing and storage' or 'cellular processes and signaling' have significantly more duplications at the LCA of oomycetes and within the Stramenochromes than at other branches. In contrast, OGs implied in host-pathogen interaction such as functional classes that contain RXLR and LXLFLAK motifs, secretion signals, as well as genes differentially expressed during infection of the host, predominantly expand within pathogenic oomycetes, both on internal as well as external branches. OGs containing predicted secreted proteins significantly expand at the LCA of *Phytophthora* spp. and throughout the genus, even though the analyzed Stramenopiles do not differ in absolute and relative size of the predicted secretomes (Supplementary Figure S3-1B).

Pathogenicity is not the only characteristic that discriminates the analyzed oomycetes and Stramenochromes, because Stramenochromes are plastid-harboring photosynthetic active organisms. This lifestyle difference is clearly reflected in the observed evolutionary pattern of OGs containing proteins with functional association to the chloroplast (Figure 3-4; Supplementary Figure S3-7). Like the pathogenicity related OGs, these OGs are also highly dynamic in their evolution: They significantly expand at the LCAs within the Stramenochromes as well as at the branch leading to *A. anophagefferens* and significantly contract at the terminal branches and at the LCA of the diatoms. Interestingly, even though oomycetes do not harbor any plastids, we observed a considerable number of genes within the oomycete genomes that belong to ~450 different chloroplast-associated OGs (Supplementary Figure S3-7). At the same time, as expected, losses of chloroplast-associated OGs are enriched at the LCA of oomycetes (Supplementary Figure S3-8).

Notably, OGs related to signal transduction, defense and also transcription predominantly expand early in evolution. It has been previously noted that in prokaryotes, the major changes in regulation of transcription and signal transduction often occur at the origins of major lineages (Cordero et al. 2008). Our observations suggest a similar expansion within the Stramenopile lineage, which may hold true for other eukaryotes as well.

Moreover, OGs characterized as metabolism-related are enriched for duplications at all internal branches throughout the Stramenopiles. Interestingly, OGs related to carbohydrate as well as amino acid transport and metabolism significantly expand at the LCA of oomycetes or throughout the clade. Glycoside hydrolases belong to the class of CAZymes (carbohydrate-active enzymes), which contains proteins involved in synthesis and breakdown of carbohydrates that are found, for example, in the cell wall of both



**Figure 3-4 Differences in evolutionary trajectories between distinct functional classes.**

Over- or underrepresentation of duplication events at distinct branches of the species phylogeny observed for different functional classes (abbreviation for COG classes are displayed behind the description). The heat map shows the fold ( $\log_2$ ) enrichment/depletion in duplications (saturating at -2 and 2). Significance of the overrepresentation/ underrepresentation was assessed using a Fisher's exact test ( $P < 0.05$ ), and multiple testing correction was addressed using a false discovery rate ( $q < 0.05$ ). Significant enrichment is indicated by asterisk; for both significant enrichment and depletion, see also supplementary Figure 3-6A.

pathogen and host. It has been shown before that glycoside hydrolases are abundant in oomycetes and that the majority of those are potentially secreted (>50%) (Tyler et al. 2006; Ospina-Giraldo et al. 2010; Seidl et al. 2011); however, the evolutionary history of expansion has so far not been uncovered.

### Evolutionary Dynamics of Glycoside Hydrolases

Our systematic analysis of the evolution of glycoside hydrolases revealed that individual OGs that are highly abundant in plant pathogenic oomycetes exhibit distinct evolutionary trajectories (Figure 3-5A and Supplementary Figure S3-9). Ninety-four OGs

are predicted to contain glycoside hydrolases; they cover a total of 1,005 proteins of which the majority (85%) is present in oomycetes (e.g., 179 in *P. infestans* and 214 in *P. sojae*) (Figure 3-5A). The repertoire of glycoside hydrolases in oomycetes is dominated by a few large OGs such as, for example, exo-beta-1,3-glucanase (glycoside hydrolase family 17, GH17); >60% of all glycoside hydrolases in oomycetes belong to only ten OGs. The high abundance of proteins within OGs is not due to isolated duplication events on single branches but is instead caused by consecutive duplications along the internal branches of the oomycete phylogeny. In addition to the high abundance of lineage-specific duplications that are partially balanced by losses, we observed a pronounced accumulation of duplications at the LCA of Peronosporales and especially at the LCA of *Phytophthora* (66 duplication events).

Extracellular hydrolases like the exo-beta-1,3-glucanase OG199 and OG225 are examples of OGs that are expanded in oomycetes and lost in Stramenochromes (Figure 3-5B). The expansion in *P. sojae* and *P. ramorum* within OG199 is mainly caused by lineage-specific expansion as well as early duplications followed by subsequent losses in *H. arabidopsidis* and *P. infestans*. In contrast, the expansion of OG225 is dominated by consecutive duplications that occur late in evolution, mainly at the LCA of Peronosporales, the LCA of *Phytophthora*, and lineage specific within *P. sojae*. These duplications are balanced by subsequent losses in all extant Peronosporales. Even though these OGs share similar biological functions, their high abundance, especially in the *Phytophthora* spp., is caused by different evolutionary trajectories.

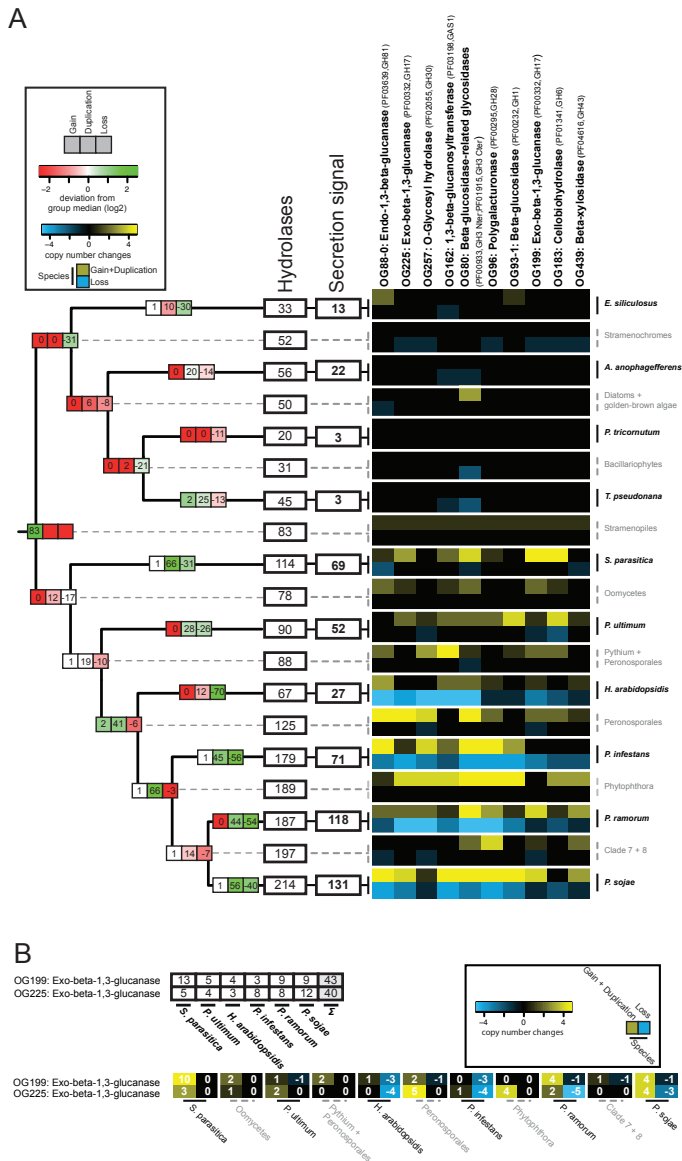
These OGs do not only differ in their individual evolutionary trajectories but also the whole repertoire of glycoside hydrolases displays a different global pattern of expansion and contraction compared with other functional classes. Another class of highly abundant enzymes in pathogenic oomycetes that have a potential role during infection are peptidases (Tyler et al. 2006; Haas et al. 2009). Whereas the LSCA contains only few glycoside hydrolases (33% of the repertoire observed in *P. sojae*), many peptidases are already present at the LSCA (225 OGs), and the repertoire of the extant taxa is either of similar size or reduced (Supplementary Figure S3-10). Nevertheless, these peptidase OGs are not static but in constant flux. We demonstrated that pathogenicity related functional classes evolve along different, even opposing trajectories, while still resulting in the observed high abundance in the present day pathogenic oomycetes.

## DISCUSSION

What are the evolutionary events that caused the expansion of OGs in pathogenic oomycetes, and when and how did the dynamic processes that shaped the genome content of these species take place? To address these questions, we systematically studied evolutionary events directly inferred from phylogenetic analysis and tree reconciliation.

Initial work on gene family evolution in Stramenopiles and in particular in pathogenic oomycetes has been limited to a few species and was based on parsimony methods to





**Figure 3-5 Global and local pattern of expansion and contraction of OGs containing glycoside hydrolase.** (A) The reconciled evolutionary events are projected on the species phylogeny as well as the total abundance of hydrolases at each taxon (ancestral and extant). Heat maps on the different branches display the deviation from the median number of events (i.e., gains, duplications, or losses). The expansion and contraction pattern of the ten largest OGs is displayed next to the phylogeny by a heat map (expansion: yellow; contraction: blue; abundance of duplications/losses saturating at -4 and 4). (B) The number of proteins of two glycoside hydrolase families (OG199 and OG225) in individual species is shown in the table. A heat map displays the expansion and contraction pattern of these two families throughout oomycetes (expansion: yellow; contraction: blue; abundance of duplications/losses saturating at -4 and 4).

reconstruct gain and losses of gene families (Martens et al. 2008; Cock et al. 2010). The expansion of families was inferred based on differences in the presence/absence and abundance pattern between species (Tyler et al. 2006; Martens et al. 2008; Haas et al. 2009; Baxter et al. 2010; Lévesque et al. 2010; Seidl et al. 2011). These analyses already provided initial insights into the genome evolution and led to the identification of large gene families that are implied to play a role in host-pathogen interaction. However, the evolutionary trajectories, that is, the patterns of gene gain, duplications, and losses that caused this abundance were not yet systematically analyzed. This study is an additional step toward uncovering these dynamics by a comprehensive phylogenetic analysis and subsequent tree reconciliation of ten Stramenopiles including six pathogenic oomycetes revealing the patterns of gene gains, duplications, and losses that caused this large gene families.

We reconciled the phylome constituted by 18,459 individual protein trees sampled from ten Stramenopiles with a species phylogeny derived by concatenating 189 single-copy genes (Figure 3-1). The species phylogeny is highly supported and mainly resembles the known topology of the tree of life. It should be noted that the exact topology of the three *Phytophthora* spp. contradicts the topology published by Blair et al. (2008) that suggested a close association of *P. sojae* with *P. infestans*. However, these authors also tested alternatives and concluded that they could not significantly reject the topology in which *P. sojae* and *P. ramorum* are closely associated, a grouping that we predict in this study with high support. The number of evolutionary events derived by reconciliation with the topology proposed by Blair et al. (2008) is higher (2,900 events), and hence, our topology is more parsimonious (Supplementary Figure S3-3A). In most cases, reconciliation with either topology did not result in major differences, whereas in some cases, the numbers of evolutionary events are even more pronounced with the topology proposed by Blair et al. (2008), for example, in the case of the accumulation of duplications at the LCA of *Phytophthora* spp. Recently, Runge et al. (2011) proposed a topology that places *H. arabidopsidis* as a sister taxon to *P. infestans*. It has been previously indicated that some clades of *Phytophthora* are paraphyletic with respect to the downy mildews (Cooke et al. 2000; Göker et al. 2007); however, our reconstructed species phylogeny groups all three analyzed *Phytophthora* spp. in a single cluster. The number of evolutionary events derived by tree reconciliation with the topology proposed by Runge et al. (2011) is much (~7,200 events) higher than our more parsimonious topology (Supplementary Figure S3-11). The disagreement between our topology and the two alternatives does not mean that these alternatives are wrong. Nevertheless, we preferred to use the phylogeny that was reconstructed from our concatenated alignment containing 189 loci. When reconciling a large number of gene families, this topology is the most parsimonious and hence conservative, therefore further supporting our choice.

A comprehensive and dynamic picture of the genome evolution in Stramenopiles was obtained by projecting gene gains, duplications, and losses that were derived by reconciliation of the phylome onto the species phylogeny (Figure 3-2). Our analysis demonstrates that throughout evolution, the genomes of Stramenopiles are not static

but in constant flux; a dynamic that is at least partially disguised by parsimonious-based methods when duplications and losses occurred in the same OG at the same branch. Whereas the genome content of Stramenochromes is of comparable size to the LSCA, genomes of pathogenic oomycetes have been growing by gains and by continuous duplications on both the internal as well as the terminal branches. The LSCA is large and contained ~10,000 genes of which the majority predate the LSCA.

Some of these genes might have not transferred vertically but instead descended from a horizontal gene transfer (HGT). Consequently, we may overestimate the number of genes in the LSCA, introduce unnecessary losses in the derived lineages, and underestimate gains in internal branches. So far, there are only few comprehensive studies that have investigated the fraction of HGTs in Stramenopiles from origins, such as bacteria or eukaryotes (Richards et al. 2006; 2011; Richards & Talbot 2007; Morris et al. 2009). A recent analysis of HGT between fungi and oomycetes has revealed 33 high-confidence HGTs that together contributed to up to ~8% of the secretome of *P. ramorum* and hence to plant parasitic mechanisms of oomycetes (Richards et al. 2011). Indeed, one of their discussed examples, a sugar transporter called *AraJ* (Richards et al. 2006; 2011), is annotated as ancestral (gained at the LSCA or before) in our analysis. More quantitatively, if we consider all OGs that consistently have their best blast hits to eukaryotes or bacteria as potential sources of HGT, only a minority (excluding singletons because these are not considered in our reconstruction) is specific to either oomycetes or Stramenochromes (Supplementary Figure S3-12). These are the only cases where an erroneous placement of the gains at the LSCA could influence our results because OGs that have members in both lineages will be invariably placed at the LSCA. These numbers are of course upper limits because real losses of ancestral OGs at either ancestor of the two lineages also occur or are included in the reported numbers (Supplementary Figure S3-12). Consequently, the quantitative influence of these events to our analysis is marginal, even though it highlights the mosaic nature of the analyzed species.

The interpretation of the inferred gene content of LSCA and the genome evolution of Stramenopiles also depends on the contribution of the plastid to their gene content. If the LCA contained a plastid, as proposed by the Chromalveolate hypothesis (Cavalier-Smith 1999; Keeling 2009), then our estimated size of the LSCA as well as the derived evolutionary events do not change (Figure 3-2). However, our results would be affected if the acquisition of the plastid by the photosynthetic Stramenochromes occurred after the speciation of oomycetes as suggested by the SEEE hypothesis (Cavalier-Smith et al. 1994; Archibald 2009; Baurain et al. 2010). If the plastid endosymbiosis mainly affected chloroplast-associated genes, we would slightly overestimate the size of the LSCA by 295 genes (2.8%) and an equivalent number of losses and gains at the branches leading to oomycetes and Stramenochromes (Supplementary Figure S3-7). However, if the plastid endosymbiosis contributed a wide array of cellular functions to the Stramenochrome ancestor, we would overestimate the size of the LSCA by up to 2,300 genes (Figure 3-2). This number has to be seen as the upper limit because we obtained it by assuming that every OG that we inferred to be lost at the branch leading to oomycetes has descended from the plastid endosymbiosis (Figure 3-2). In contradiction to the SEEE

hypothesis, we observed 432 OGs that are chloroplast-associated and retained in the genomes of both nonphotosynthetic oomycetes and Stramenochromes since the LSCA (Supplementary Figure S3-7). Similarly, 88 and 14 oomycete-specific OGs have their best blast hits in green and red algae genomes, respectively (Supplementary Figure S3-12). These results, together with studies by others (Andersson & Roger 2002; Tyler et al. 2006; Maruyama et al. 2009), seem to slightly favor the early acquisition of the plastid before the speciation of Stramenochromes and oomycetes. However, recent molecular data support a more complex scenario and later acquisition of the plastid thereby rejecting the Chromalveolate hypothesis (Stiller et al. 2009; Baurain et al. 2010; Felsner et al. 2011; Woehle et al. 2011). Nevertheless, our results do not change dramatically and are hence independent of the precise history of the plastid. Dedicated future research, also facilitated by additional genomes from related lineages, will gather additional evidence for either of the two hypotheses and thereby shed light on this controversially discussed event and hence also on our reconstructions.

The massive accumulation of duplications at the LCA of *Phytophthora* spp. points to a large-scale duplication event (Figure 3-3; Supplementary Figure S3-5). It has been postulated that the accumulation of duplications at a constrained point in time can be indicative for duplications that affect either large parts of the genome or the whole genome (McLysaght et al. 2002; Jaillon et al. 2004; Kellis et al. 2004; Jiao et al. 2009). This accumulation of duplication events was already observed earlier by Martens and colleagues who used an independent method to time the age of paralogs in *Phytophthora* spp. (Martens & Van de Peer 2010). The usage of additional outgroup species allows us to more precisely estimate the time of these events, which seem to have happened after the speciation of *H. arabidopsidis* and before the radiation of *Phytophthora* spp. Nevertheless, the usage of the less parsimonious topology of the analyzed Peronosporales proposed by Runge et al. (2011) introduces an accumulation of duplications at the LCA of Peronosporales (Supplementary Figure S3-11). Hence, if this proposed topology is correct, it is tempting to speculate that the analyzed Peronosporales shared this large-scale duplication event. Considering our predicted topology, such an earlier timing of this event could also be possible; the genome contraction of *H. arabidopsidis* might lead to the loss of both duplicates and hence at least partially obscure events happening at the LCA of Peronosporales. Nevertheless, neither the analysis performed by Martens and colleagues nor ours is able to elucidate the exact mode of expansion because of the lack of long-distance intra-species co-linearity of genes. Alternative scenarios, such as segmental duplications that occurred at a constrained point in time followed by reorganization, are at least equally likely, especially given the observed dynamics in genome organization of *Phytophthora* spp. and the genome contraction in *H. arabidopsidis*. Independent of the underlying mechanism, this coordinated expansion of gene families marks a major transition point in their evolution. Together with subsequent lineage-specific losses, the expansion could be the driving force of the speciation and adaptation to different hosts within the *Phytophthora* genus (or even within the Peronosporales); a process that has been proposed before for other organisms, such as yeast (Kellis et al. 2004) or plants (Jiao et al. 2009).

The number of duplications and losses events at each branch is determined by tree reconciliation. This procedure is not only dependent on a reliable species phylogeny, but also on the alignment as well as the gene tree, or in this case, protein tree. Erroneously inferred protein trees, either based on inaccurate alignments or due to biases in the tree predictions itself, will artificially increase the number of duplications at internal branches and losses at terminal branches of the tree. To address if the incorporation of low-quality alignments in our analysis interferes with our main results, we divided the families into high-quality and low-quality alignments (see Supplementary Material and Methods S3-1 and Supplementary Figure S3-15). If we remove the 477 families and their derived OGs that have a low-quality alignment in our analysis, we observe that the absolute numbers of evolutionary events decrease as the analysis is now based on less data (Supplementary Figure S3-15). More importantly, the relative numbers and the major trends observed in our analysis, such as the accumulation of duplication in the common ancestor of *Phytophthora* spp., are independent of the exclusion of the lower quality alignments (Supplementary Figure S3-15). Consequently, our results are robust to the possible bias introduced by the retention of the full set of families. To reduce the possible bias in the tree prediction and to apply an explicit model of evolution, we used a maximum likelihood method to predict the tree topology of the protein trees. More importantly, we used NOTUNG (Chin et al. 2005; Durand et al. 2006) for tree reconciliation that allows to explicitly address this uncertainty in protein trees. NOTUNG allows the rearrangement of weakly supported parts of the tree topology to reduce the evolutionary events needed for reconciliation while keeping strongly supported parts fixed. Throughout this study, we used a bootstrap support of >80% to indicate strongly supported clades of the protein trees. Hence, parts of the tree topology that are not supported with a bootstrap of at least 80% are rearranged to minimize evolutionary events. When we compared the results derived with >80% cutoff to a less conservative cutoff of >60%, leading to less rearrangement, we indeed observed more duplications at the internal branches and more losses at terminal branches, especially within oomycetes (Supplementary Figure S3-13). When we applied an even stricter cutoff of >90%, which resulted in more rearrangement, some duplications at the internal branches, for example, at the LCA of Stramenochromes and especially at the LCA of Peronosporales, were removed; consequently, fewer losses in the terminal taxa were introduced (Supplementary Figure S3-13B). Regardless of the choice of the cutoff (60%, 80%, or 90%), the changes in the abundance of the reconciled evolutionary events did not interfere with our global results indicating the robustness of our framework to this bias.

Our results are directly dependent on the availability, quality, and completeness of the predicted proteomes derived from the various sequenced genomes. The robustness of gene annotation has been observed to have only small effects on the analysis of gene family losses in related species (Martens et al. 2008). In general, more sequenced genomes of closely related oomycetes, preferably sister taxa to the already existing genomes, would enable a more precise timing of the duplication events, especially at the terminal branches. Moreover, our analyses are currently limited to pathogenic oomycetes. Including sequenced genomes of saprophytic species would elucidate whether evolutionary events at the LCA of oomycetes are specific to pathogenic

oomycetes or are instead a general pattern for all oomycetes.

## CONCLUSIONS

We systematically analyzed the genome evolution of pathogenic oomycetes by reconciliation of the Stramenopile phylome with a highly supported species phylogeny. Our analysis uncovered that oomycete genomes, emanating from a common ancestor of Stramenopiles that had a rather large genome encoding for ~10,000 genes, were growing by continuous duplications that predominantly affected ancestral OGs. The massive accumulation of duplication events at the LCA of the *Phytophthora* genus suggests a large-scale duplication event that predates the speciation and hence might be driving the adaptive radiation within this genus. Different functional classes have distinct evolutionary trajectories: not only between classes but also within a single class. Different evolutionary trajectories are proposed to lead to the observed abundance of pathogenicity-related functional classes, for example, glycoside hydrolases and peptidases, an observation that was not yet apparent by previous analyses. Consequently, we unveiled both large-scale evolutionary processes that shape the genomes of extant oomycetes as well as the complex evolution trajectories that lead to highly abundant gene families in this important class of pathogens.

## ACKNOWLEDGMENTS

We thank Lidija Berke, Like Fokkens, Adrian Schneider, and Gabino F. Sanchez-Perez for helpful discussions and comments on the manuscript and Stefan Zoller for technical support. Some of the sequence data and annotation were provided by the US Department of Energy Joint Genome Institute (<http://www.jgi.doe.gov>), the Broad Institute of Harvard, and the Massachusetts Institute of Technology (<http://www.broadinstitute.org>) in collaboration with the user community (for detailed information, see Supplementary Material and Methods S3-1). This project was financed by the Centre for Bio-Systems Genomics (CBSG) which is part of the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research.

## SUPPLEMENTARY MATERIAL

Due to the amount of data, some of the Supplementary additional files, Supplementary Figures S3-6, S3-9 and S3-13, are only accessible online at *Genome Biology and Evolution* (<http://www.gbe.oxfordjournals.org/>).

## Supplementary Material & Methods S3-1

### *Gene family assignment*

We acquired the predicted protein sequences from ten Stramenopiles:

Species	Version	Source	URL	Reference
<i>Ectocarpus siliculosus</i>	v1	BOGAS	<a href="http://bioinformatics.psb.ugent.be/genomes/">http://bioinformatics.psb.ugent.be/genomes/</a>	(Cock et al. 2010)
<i>Aureococcus anophagefferens</i>	v1	JGI	<a href="http://genome.jgi-psf.org/">http://genome.jgi-psf.org/</a>	(Gobler et al. 2011)
<i>Phaeodactylum tricornutum</i>	v2	JGI	<a href="http://genome.jgi-psf.org/">http://genome.jgi-psf.org/</a>	(Bowler et al. 2008)
<i>Thalassiosira pseudonana</i>	v3	JGI	<a href="http://genome.jgi-psf.org/">http://genome.jgi-psf.org/</a>	(Armbrust et al. 2004)
<i>Hyaloperonospora arabidopsidis</i>	v8.3	VBI	<a href="http://vmd.vbi.vt.edu/">http://vmd.vbi.vt.edu/</a>	(Baxter et al. 2010)
<i>Saprolegnia parasitica</i>	v1	BROAD	<a href="http://www.broadinstitute.org/scientific-community/data">http://www.broadinstitute.org/scientific-community/data</a>	NA
<i>Pythium ultimum</i>	v4	BROAD	<a href="http://www.broadinstitute.org/scientific-community/data">http://www.broadinstitute.org/scientific-community/data</a>	(Lévesque et al. 2010)
<i>Phytophthora infestans</i>	v1	BROAD	<a href="http://www.broadinstitute.org/scientific-community/data">http://www.broadinstitute.org/scientific-community/data</a>	(Haas et al. 2009)
<i>Phytophthora ramorum</i>	v1	BROAD	<a href="http://www.broadinstitute.org/scientific-community/data">http://www.broadinstitute.org/scientific-community/data</a>	(Tyler et al. 2006)
<i>Phytophthora sojae</i>	v1	BROAD	<a href="http://www.broadinstitute.org/scientific-community/data">http://www.broadinstitute.org/scientific-community/data</a>	(Tyler et al. 2006)

To define protein families, we created a sparse network of nodes (proteins) connected by edges (sequence similarity) by conducting a blastp (Altschul et al. 1990) all-vs.-all sequence similarity search (e-value cutoff  $1 \times 10^{-3}$ , enabled soft filtering). We subsequently removed edges that were formed between proteins due to short segments of similarity thereby eliminating spurious connections within the network. We removed edges between proteins if the matched area was  $\leq 50\%$  or the ‘actual-matching’ area was  $\leq 20\%$  of either the query or the subject. The matching area is defined as the area from the start position of the first segment to the end position of the last segment and the ‘actual-matching’ area as the sum of the covered area by each individual segment. Subsequently, we partitioned the resulting network into protein (gene) families using the markov clustering algorithm (MCL) (Van Dongen 2000; Enright et al. 2002) with an inflation value of 3.0.

### *Detection of transposable elements*

The presence of transposable elements or their signatures within the predicted proteomes of the analyzed species was assessed using two complementary methods; (i) by screening for the presence of 86 signature domains and the MULE transposase domain within predicted proteins (Zdobnov et al. 2005) and (ii) by screening for sequences that show similarity to position-specific scoring matrices for several families of transposable elements. For (i) we predicted the domain repertoire for all proteins using hmmer3 (Eddy 1998) and a local Pfam database (v24) (Finn et al. 2010) applying a domain model specific cutoff (gathering cutoff). For (ii) we used TransposonPSI (Haas BJ, <http://transposonpsi.sourceforge.net/>) to scan the predicted proteomes for the presence of different families of transposable elements. Subsequently, we removed all families containing predicted proteins that have one or more signature domains that are specific for

transposable elements or exhibit similarity to transposable elements.

#### *Phylogenetic analysis*

We constructed a phylogenetic tree of the analyzed Stramenopiles using 189 families whose members occur in a single copy gene in each of the ten species. Each set of ten single copy genes was first aligned using mafft (Katoh et al. 2002) (v6.713b, L-INS-I algorithm) and the aligned sequences were subsequently concatenated. We removed columns with more than 80% gaps. Furthermore, we removed adjacent divergent positions both up- and downstream of the gap-position until a column with a median of pair-wise BLOSUM62 scores  $\geq 0$  was found. The resulting alignment was used to infer a maximum likelihood phylogenetic tree using RAXML (Stamatakis 2006) (v7.0.4) with gamma model of heterogeneity and estimated alpha parameter (-PROTGAMMA) as well as a WAG amino acid substitution matrix. The robustness of the tree topology was assessed using 1,000 bootstrap replicates. To address if few long alignments dominate the concatenated alignment, we removed all families whose alignments (after removing gaps as outlined above) exceeded a length cutoff that was empirically defined by the length distribution of the 189 single copy families (Supplementary Figure S3-14A). This length cutoff was set to be the 3rd quartile + 0.5 \* inter quartile range of the length distribution and yielded 168 families that were subsequently concatenated. The phylogenetic tree was inferred as described above and robustness was assessed using 1,000 bootstrap replicates (Supplementary Figure S3-14B). The tree topology as well as the bootstrap support for the individual branches is identical to the predicted topology that was based on the full set of 189 single copy markers.

To estimate the relative divergence times of the analyzed Stramenopiles, we inferred the phylogeny including the ciliate *Paramecium tetraurelia* (v1.41, ParameciumDB (Aury et al. 2006; Arnaiz et al. 2007; Arnaiz & Sperling 2011)) as explicit outgroup. We identified 35 single copy families in Stramenopiles and *P. tetraurelia* and utilized these as a concatenated marker that was prepared and analyzed as described above. The relative divergence times were estimated with BEAST (Drummond & Rambaut 2007) under strict clock assumption and a gamma model of site heterogeneity (invariant sites + 4 gamma categories; WAG substitution matrix). We used a defined tree topology and starting branch lengths derived from the beforehand maximum likelihood analysis. Furthermore, we set the age of the last common ancestor of Stramenopiles arbitrary to 100. We ran ten independent chains, each containing 4,000,000 generations of which we sampled every 400 generations. The resulting posterior distributions for parameter estimates were manually assessed using Tracer (v1.5). Subsequently, maximum credibility trees were calculated with TreeAnnotator (1.6.1) after removing 10% burn-in. The estimated branch lengths were averaged over the ten chains. The probability to observe less, equal or more than the abundance of evolutionary events given the expectation values at each individual branch was assessed by Poisson distribution. The expected values were estimated using the global relative frequency of duplications/losses.



### *Reconstruction of gene family evolution*

We reconstructed the evolutionary history of protein families (excluding singletons) to monitor macro-evolutionary events like duplications and losses along the species tree. The sequences of the gene families were aligned using different alignment algorithms similar to a strategy outlined by Muller and colleagues (2010b). We used mafft (v6.713b; L-INS-I, E-INS-I and default parameters) (Katoh et al. 2002) and muscle (v3.7; with default parameter) (Edgar 2004) to align the protein sequences. Moreover, we corrected all alignments with rascal (Thompson et al. 2003) and subsequently assessed the alignment quality with norMD (v1.3) (Thompson et al. 2001). Per individual family the highest scoring alignment out of the refined and original alignments was chosen. We constructed phylogenetic trees using RAxML (Stamatakis 2006) (PROTGAMMA, WAG) for families >3, excluding families >500 members, and the robustness of the trees were assessed using 100 bootstrap replicates.

We reconciled the protein trees with the species tree of Stramenopiles using NOTUNG (Chen et al. 2000; Durand et al. 2006) (modified v2.6, personal communication). The trees were reconciled using a cost of 1.5 for a duplication event and 1 for a loss event. Subsequently, the tree was rooted so that the number of duplication and loss events are minimized. Furthermore, NOTUNG allows the rearrangement of the gene tree topology on weak branches to account for errors in the gene tree that would lead to bias in the derived number of evolutionary events. Weakly supported branches (bootstrap <80%) were rearranged to minimize the number of evolutionary events, while at the same time strongly supported topologies remained intact. Furthermore, we created orthologous groups by dividing families based on duplications occurring at the last common ancestor of Stramenopiles. Consequently, each orthologous group represents a single gene either in the last common ancestor of Stramenopiles or at the point of gain. Subsequently, we used maximum parsimony to project the derived evolutionary events of all orthologous groups, including species-specific groups, on the species phylogeny of Stramenopiles.

To assess the contribution of potential low quality alignments to the evolutionary events we subdivide the set of alignments (optimal score per family) into high quality and low quality subsets. Low quality alignments are defined by a norMD score of <0.75 (Supplementary Figure S3-15), i.e. families with a norMD score of <0.6 (127) and those which exceed this cutoff by 25% (477). The norMD score cutoff of 0.6 was proposed to be of high quality and hence more reliable (Thompson et al. 2001; Muller et al. 2010b). Subsequently, we projected the evolutionary events based on the derived OGs of the high and low quality families onto the species phylogeny (Supplementary Figure S3-15).

To further elucidate the origin of each individual orthologous group we searched for homologs utilizing best hits identified by a blast search (e-value cutoff  $1 \times 10^{-3}$ ; enabled low complexity filtering; query & coverage filtering as described above) against a local version of the eggNOG database(v2) (Muller et al. 2010a). If a homolog of a group in eggNOG could be identified we assumed an origin before the last common ancestor of

Stramenopiles and introduced, if necessary, subsequent losses.

#### *Functional annotation of OGs*

We projected functional annotation to each individual OG using five independent methods: (i) COG functional classification (Tatusov et al. 1997) was assigned by identification of homologs in the eggNOG database utilizing best hits identified by blast search (e-value cutoff  $1 \times 10^{-3}$ ; enabled low complexity filter) (Muller et al. 2010a). If a protein was consistently assigned to a functional class derived from homologs in eggNOG and the OG contained >30% proteins with the identical classification, the functional annotation was projected to the whole OG. (ii) The presence of potentially secreted proteins within an OG was predicted using SignalP (Bendtsen et al. 2004) (v3.0) in combination with TMHMM (v2.0) (Krogh et al. 2001). Secretion signal within the first 70 amino acids of a protein was accepted if both the neural network as well as the HMM consistently predicted the presence of the motif. Signal peptides were rejected if TMHMM predicted more than a single transmembrane region within the protein or a single region that overlapped with the SignalP prediction for less than 10 amino acids or was positioned outside the first 35 amino acids. OGs that contain >30% proteins with a predicted secretion signal were annotated as secreted OG. (iii) Host-cell translocation motifs were predicted using hmmer3 (Eddy 1998) and manually created HMM-profiles of the RXLR and the LXLFLAK motif (R.H.Y. Jiang, personal communication). Next to the RXLR/LXLFLAK motif itself, we also demanded the presence of a predicted secretion signal within the first 30 amino acids, the gap between the RXLR/LXLFLAK motif to the secretion signal to be  $\leq 50$  amino acids and the RXLR/LXLFLAK motif to start within the first 100 amino acids. OGs that contained >30% proteins with a predicted RXLR or LXLFLAK motif were annotated. (iv) Gene expression data of *P. infestans* during infection of the host were acquired from the Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo/>) (Haas et al. 2009). Differentially expressed genes were identified as described elsewhere (Haas et al. 2009; Seidl et al. 2011). OGs containing significantly differentially expressed *P. infestans* genes were annotated as potentially differentially expressed during host-pathogen interaction. (v) Chloroplast associated proteins were identified by transferring the gene ontology (GO) annotation using Blast2GO (default parameters) (Conesa et al. 2005). All proteins that received GO annotations that could be traced back to GO:0015979 (photosynthesis), GO:0009536 (plastid) or GO:0009507 (chloroplast) were annotated as chloroplast associated. OGs containing these proteins were annotated as potentially chloroplast associated. Significantly enriched GO terms of individual proteins present in OGs predicted to be lost at the LCA of oomycetes were defined by BiNGO (version 2.44) (Maere et al. 2005). Significantly enriched GO terms were summarized by removing redundancies using REVIGO (default settings) (Supek et al. 2011). (vi) All OGs that could not be annotated with one of the four methods described above were classified as 'unknown'.

The significant over-/underrepresentation of evolutionary events for individual functional classes at each branch of the phylogeny was assessed by applying a Fisher's exact test (significance level of  $< 0.05$ ). Multiple-testing was addressed using false dis-

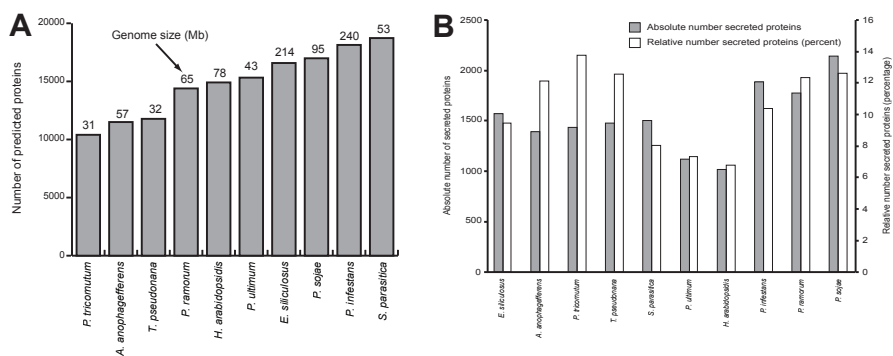
covery rates calculated by the qvalue package (Storey & Tibshirani 2003) and a q-value significance level of 0.05 was applied.

OGs defined as glycoside hydrolase or peptidase were annotated based on the presence of one or more signature domains acquired from the Pfam database which contains in total 72 different Pfam domains for glycoside hydrolases and 171 for peptidases (Finn et al. 2010). We annotated OGs that contained >30% proteins that have on of these predicted signature domain as defined by hmmer3 (gathering cutoff) (Eddy 1998). Additional annotation was transferred based on identified homologs in the eggNOG database (Muller et al. 2010a) (see above).

### Distribution of best blast hits

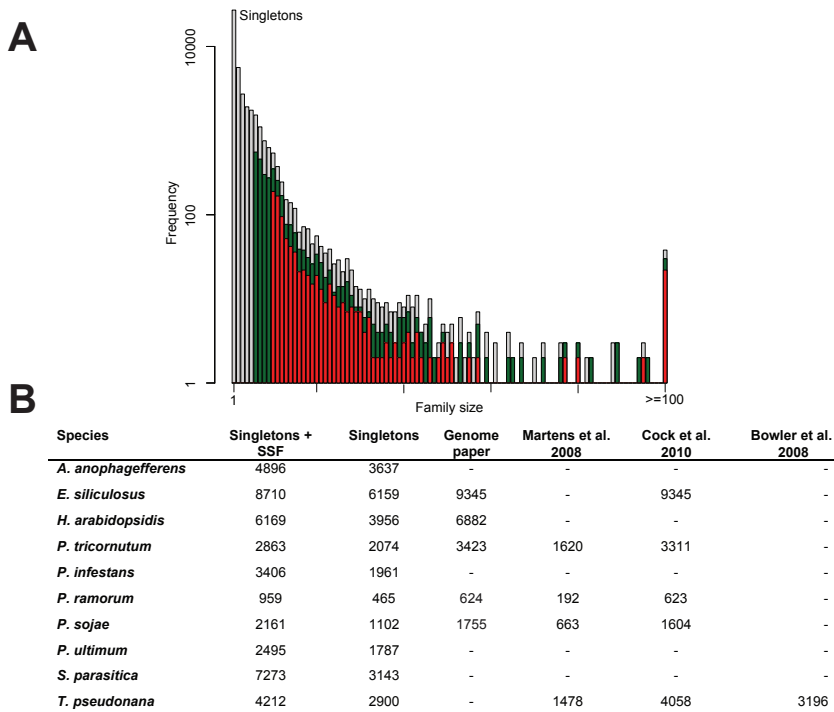
To elucidate the phylogenetic affinity of the OGs to different group of organisms, we searched for the best blast hit (e-value cutoff  $1 \times 10^{-3}$ ; enabled low complexity filter; query & coverage filtering as described above) of each protein that comprises the individual OG. These searches were conducted against the eggNOG database as well as individual proteomes of several algae species (the effective length of the database was fixed for the blast search): the red alga *Cyanidioschyzon merolae* (<http://merolae.biol.s.u-tokyo.ac.jp/>) and the green algae *Volvox carteri* (v2; <http://genome.jgi-psf.org/>), *Ostreococcus tauri* (v2; <http://genome.jgi-psf.org/>) and *Micromonas pusilla* (v3; <http://genome.jgi-psf.org/>). An OG was considered to be affine to a certain group or subgroup of species if >50% of its containing proteins consistently had their best blast hit within this group.

### Supplementary Figures

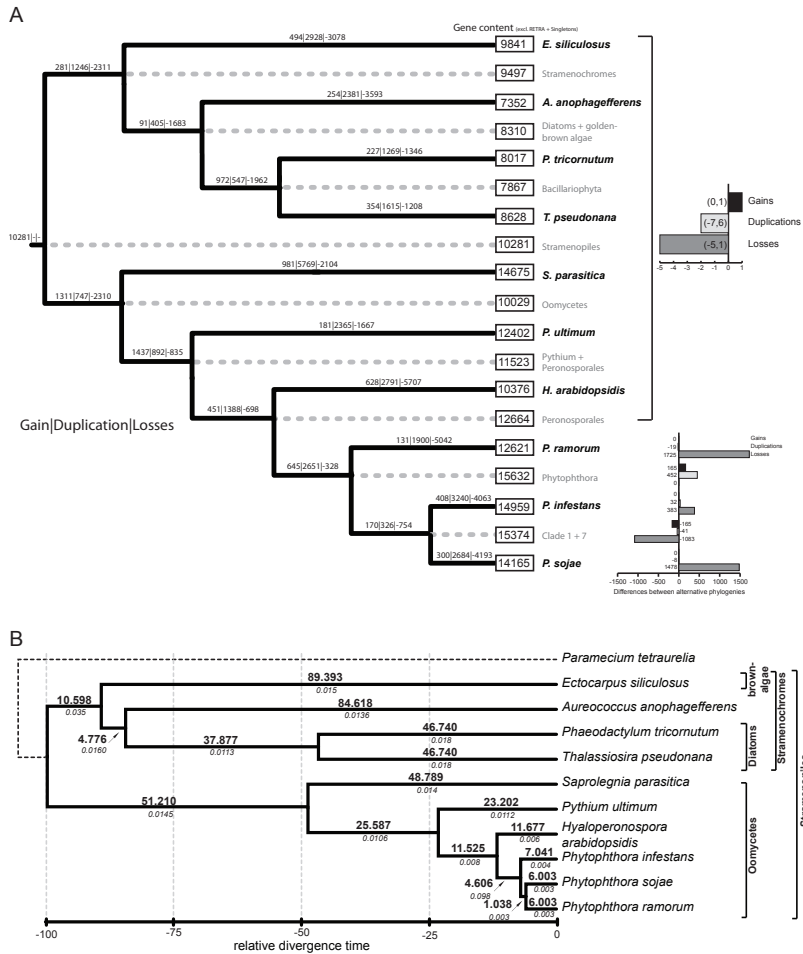


**Figure S3-1 Genome and proteome size of the analyzed Stramenopiles**

(A) The estimated genome sizes in Mb and the number of predicted proteins, (B) and the absolute (grey) and relative (white) number of predicted secreted proteins in the proteomes of the ten Stramenopiles analyzed in this study.

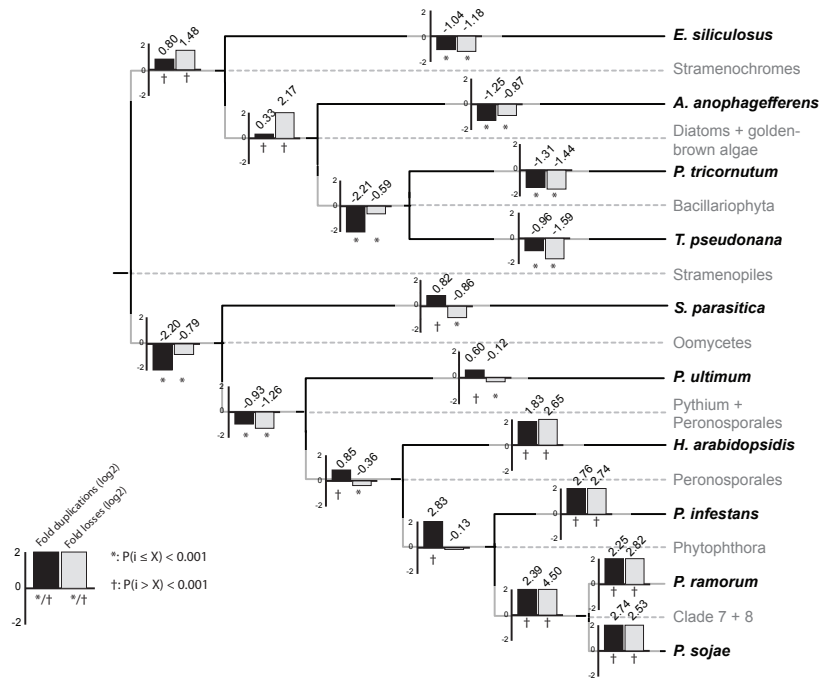


**Figure S3-2 Distribution of gene family sizes and comparison of the number of singletons to other studies.** (A) The size distribution of gene families formed by our analysis is displayed in grey (y-axis in  $\log_{10}$ -scale). Families that have at least a single copy in all oomycetes or in all Stramenopiles are displayed in green or red, respectively. (B) Number of singletons per analyzed Stramenopiles as well as the number of sequences in species-specific families is shown in the table. Reported numbers of singletons in the same species identified in other studies are reported, either by the respective genome paper or by comparative genome analysis.



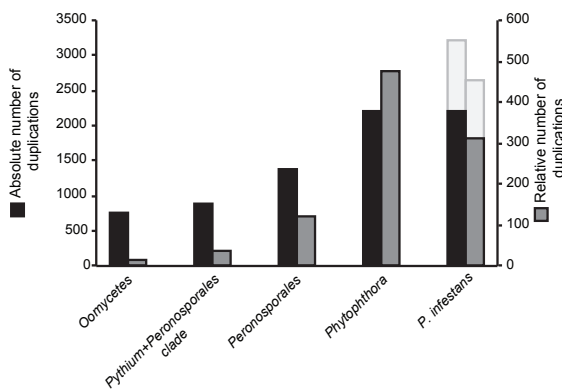
**Figure S3-3 Ultrametric phylogeny of the analyzed Stramenopiles and deviation from alternative species phylogeny.**

(A) Evolutionary events reconciled and projected on the alternative species phylogeny proposed by Blair et al (2008). Numbers of gene gains, duplications and losses are indicated along the branches and the estimated genome size is displayed in the boxes. The bar plots on the right show the differences between the number of evolutionary events determined by tree reconciliation with our and the alternative species phylogeny; positive numbers indicate and increase of evolutionary events when considering the alternative phylogeny (black bar – gain, light-grey – duplication, grey – loss). Events that differ outside of the *Phytophthora* spp. due to the alternative, optimal reconciliation of a single OG are averaged and the range is displayed in brackets. (B) Ultrametric tree of Stramenopiles derived by a maximum likelihood analysis of 35 concatenated single copy marker families identified in the ten Stramenopiles and the ciliate *Paramecium tetraurelia*. *P. tetraurelia* is included as an explicit outgroup to the analyzed Stramenopiles. Averaged divergence times per branch (in bold and standard deviation in *italic*) are reported relative to an arbitrary age of the last common ancestor of Stramenopiles set at 100.



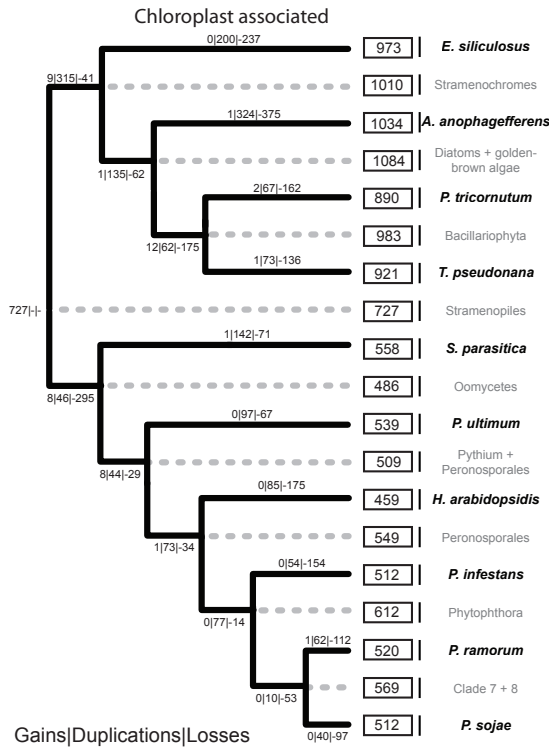
**Figure S3-4 Deviation of observed evolutionary events to the expected number of events at each individual branch.**

Deviation of observed evolutionary events from the expected number of events at each individual branch. Bar charts indicate the fold ( $\log_2$ ) enrichment/depletion of duplications/losses at each branch (fold ( $\log_2$ ) displayed above the bars). Significance of the deviation from the expectation is described by the cumulative probability derived from a Poisson-distribution, either for  $P(i \leq X)$  (\*) or  $P(i > X)$  (†).



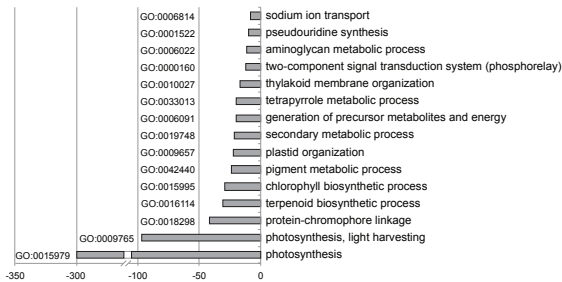
**Figure S3-5 Absolute and relative numbers of duplication events for *P. infestans* and its ancestors.**

Absolute and relative numbers of duplication events for *P. infestans* and its ancestors. The absolute number of duplications is displayed in black, whereas the relative number of duplications per relative divergence time of the branch is shown in grey. The light-grey bar indicates the abundance including duplications of lineage specific OGs.



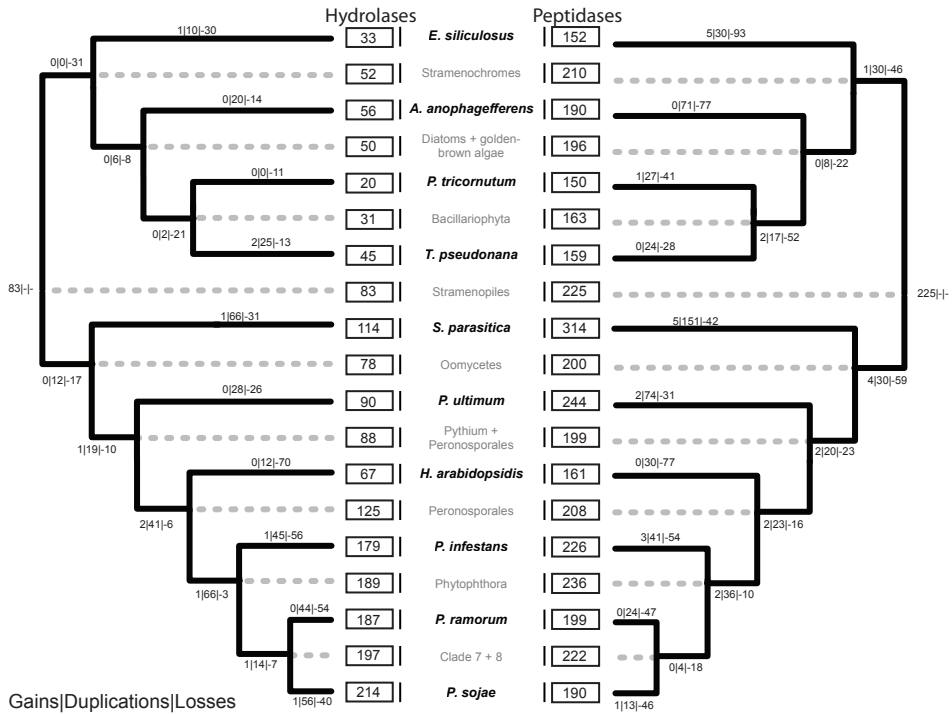
**Figure S3-7 Projected evolutionary events of OGs containing genes associated to chloroplast function on the Stramenopile phylogeny.**

Projected evolutionary events of OGs containing genes associated to chloroplast function (according to GO annotation) on the Stramenopile phylogeny. The number of evolutionary events, i.e. gene gains, duplications and losses, are projected to each branch.



**Figure S3-8 Gene ontology term enrichment of OGs that are lost at the LCA of oomycetes.**

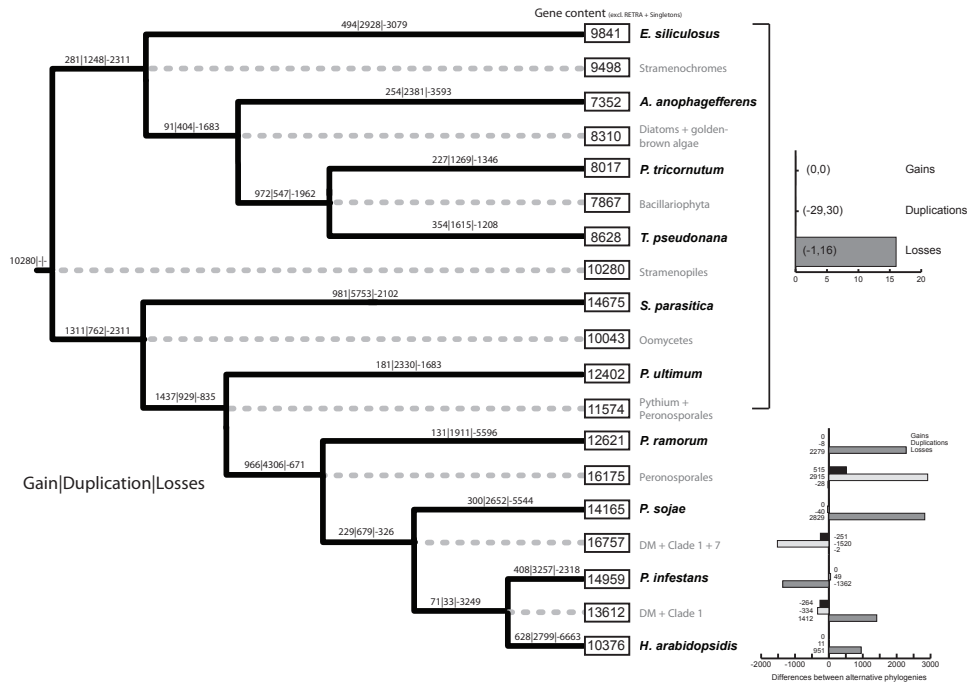
Gene ontology term enrichment of OGs that are lost at the LCA of oomycetes. The fifteen most significant terms as identified by BiNGO (Maere et al. 2005) are displayed by bar charts ( $\log_{10}$  of the corrected p-value). Beforehand, redundant gene ontology terms were summarized using REVIGO (Supek et al. 2011).



**Figure S3-10 Comparison of the reconciled evolutionary events for 94 glycoside hydrolase and 225 peptidase OGs.**

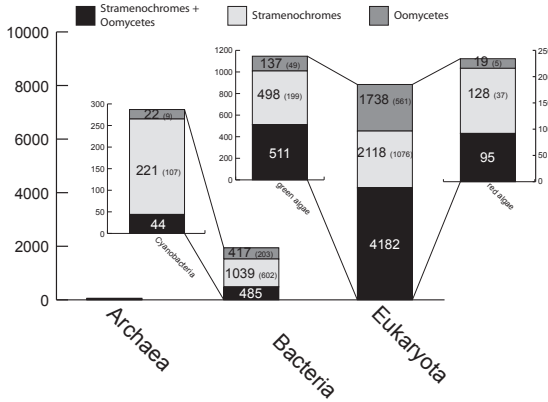
Comparison of the reconciled evolutionary events for 94 glycoside hydrolase (left) and 225 peptidase (right) OGs. The reconciled evolutionary events and the abundance at each taxon (ancestral/extant) are projected onto the species phylogeny.





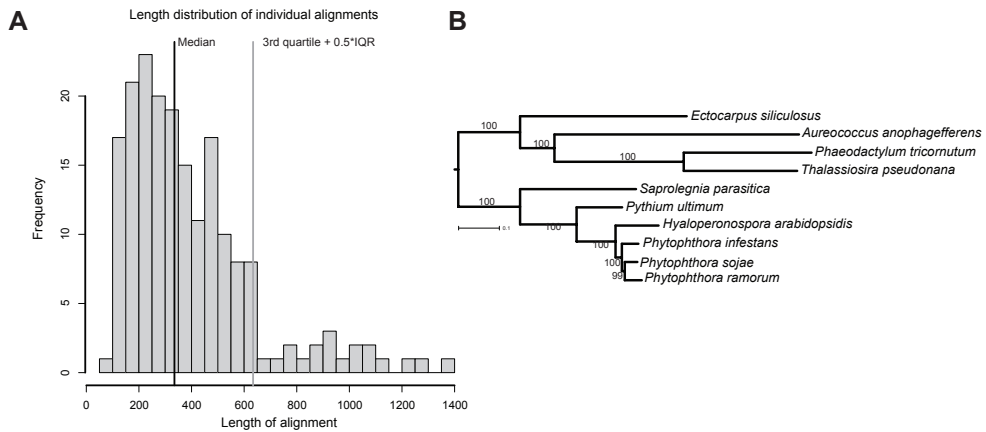
**Figure S3-11** Estimated evolutionary events for alternative species phylogeny proposed by Runge et al. (Runge et al. 2011) and deviation from our results.

Evolutionary events reconciled and projected on the alternative species phylogeny proposed by Runge et al. (Runge et al. 2011). Numbers of gene gains, duplications and losses are indicated along the branches and the estimated genome size is displayed in the boxes. The bar plot on the right shows the differences between the number of evolutionary events determined by tree reconciliation with our and the alternative species phylogeny; positive numbers indicate an increase of evolutionary events when considering the alternative phylogeny (black bar – gain, light-grey – duplication, grey – loss). Events that differ outside of the Peronosporales due to the alternative, optimal reconciliation of a single OG are averaged and the range is displayed in brackets.



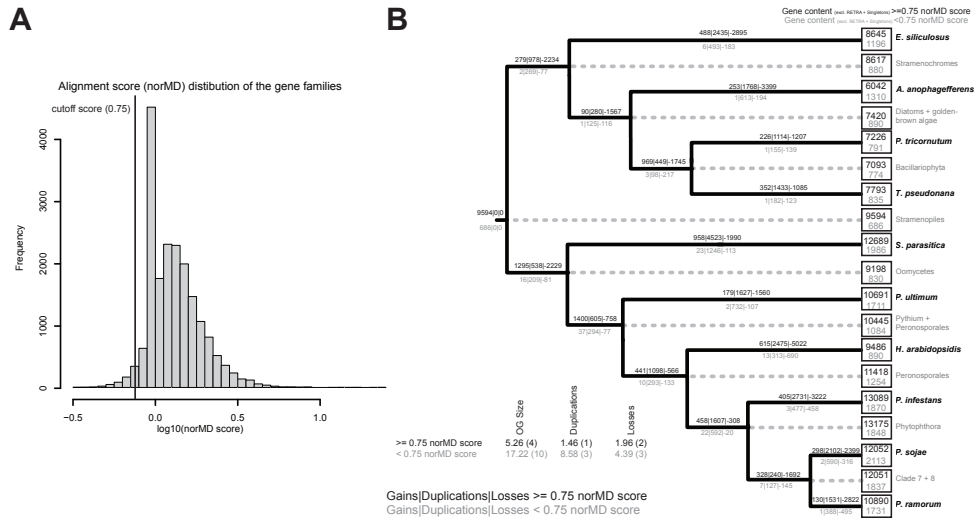
**Figure S3-12 Fraction of the best blast hits of all analyzed OGs to different species groups.**

Number and distribution of the best blast hits of all analyzed OGs to Eukaryota, Bacteria or Archaea. Additionally, best hits to Cyanobacteria as well as red and green algae are indicated as a subset within the Bacteria and the Eukaryota, respectively. OGs were separated into three different groups based on the distribution of the Stramenopiles species: (i) OGs that are found in both Stramenochromes and oomycetes (black), (ii) OGs only in observed Stramenochromes (light-grey) and (iii) OGs only observed in oomycetes (grey). Numbers of singletons that are part of the OGs are indicated in brackets.



**Figure S3-14 Length distribution of individual alignments used to construct the Stramenopile species phylogeny.**

(A) Histogram depicting the length distribution of the 189 individual alignments (after removing gaps and adjacent divergent positions, see additional file 2). Median of the alignment lengths is shown with black vertical line and the 3rd quartile + 0.5\*inter quartile range (IQR) is displayed with the grey vertical line. (B) Species phylogeny of the ten analysed Stramenopiles based on a concatenated marker of 168 single copy families. These concatenated marker excluded all alignments that were longer than the 3rd quartile + 0.5\* IQR of the length distribution of all alignments. The predicted species phylogeny and the support for each individual branch do not differ from the phylogeny that is based on the complete marker.



**Figure S3-15 Distribution of the norMD alignment scores of the analyzed families and the impact of low quality alignments to our results.**

(A) Distribution of the norMD alignment scores of the analyzed families. The vertical black line indicates the applied norMD score cutoff for low quality alignments of 0.75 and alignment scores are reported in log<sub>10</sub>-scale. (B) Projected evolutionary events on the Stramenopile phylogeny divided for OGs derived by high quality alignments (black) and low quality alignments (grey). The number of evolutionary events, i.e. gene gains, duplications and losses, are projected to each branch of the phylogeny. The predicted gene content of the ancestors and of the extant taxa (excluding singletons and transposable elements) is displayed in terminal boxes. Average OG size (median in brackets), number of duplications and losses of high and low quality derived OGs are shown in the table next to the phylogeny.

## REFERENCES

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

Andersson JO, Roger AJ. 2002. A cyanobacterial gene in nonphotosynthetic protists—an early chloroplast acquisition in eukaryotes? *Curr. Biol.* 12:115–119.

Archibald JM. 2009. The puzzle of plastid evolution. *Curr. Biol.* 19:R81–8.

Armbrust EV et al. 2004. The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism. *Science.* 306:79–86.

Arnaiz O, Cain S, Cohen J, Sperling L. 2007. ParameciumDB: a community resource that integrates the *Paramecium tetraurelia* genome sequence with genetic data. *Nucleic Acids Res.* 35:D439–44.

Arnaiz O, Sperling L. 2011. ParameciumDB in 2011: new tools and new data for functional and comparative genomics of the model ciliate *Paramecium tetraurelia*. *Nucleic Acids Res.* 39:D632–6.

Aury J-M et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature.* 444:171–178.

Baurain D et al. 2010. Phylogenomic evidence for separate acquisition of plastids in cryptophytes, haptophytes, and stramenopiles. *Mol. Biol. Evol.* 27:1698–1709.

Baxter L et al. 2010. Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis*

- genome. *Science*. 330:1549–1551.
- Bendtsen JD, Nielsen H, Heijne von G, Brunak S. 2004. Improved prediction of signal peptides: SignalP 3.0. *J. Mol. Biol.* 340:783–795.
- Blair JE, Coffey MD, Park S-Y, Geiser DM, Kang S. 2008. A multi-locus phylogeny for *Phytophthora* utilizing markers derived from complete genome sequences. *Fungal Genet. Biol.* 45:266–277.
- Bowler C et al. 2008. The *Phaeodactylum* genome reveals the evolutionary history of diatom genomes. *Nature*. 456:239–244.
- Cavalier-Smith T. 1999. Principles of protein and lipid targeting in secondary symbiogenesis: euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. *J. Eukaryot. Microbiol.* 46:347–366.
- Cavalier-Smith T, Allsopp MT, Chao EE. 1994. Chimeric conundra: are nucleomorphs and chromists monophyletic or polyphyletic? *Proc. Natl. Acad. Sci. U.S.A.* 91:11368–11372.
- Chen K, Durand D, Farach-Colton M. 2000. NOTUNG: a program for dating gene duplications and optimizing gene family trees. *J. Comput. Biol.* 7:429–447.
- Chen Z-Q et al. 2006. The essential vertebrate ABCE1 protein interacts with eukaryotic initiation factors. *J. Biol. Chem.* 281:7452–7457.
- Chin C-S, Chuang JH, Li H. 2005. Genome-wide regulatory complexity in yeast promoters: separation of functionally conserved and neutral sequence. *Genome Res.* 15:205–213.
- Cock JM et al. 2010. The *Ectocarpus* genome and the independent evolution of multicellularity in brown algae. *Nature*. 465:617–621.
- Conesa A et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 21:3674–3676.
- Cooke D, Drenth A, Duncan J, Wagels G, Brasier C. 2000. A molecular phylogeny of *Phytophthora* and related oomycetes. *Fungal Genet. Biol.* 30:17–32.
- Cordero OX, Snel B, Hogeweg P. 2008. Coevolution of gene families in prokaryotes. *Genome Res.* 18:462–468.
- Dou D et al. 2008. RXLR-mediated entry of *Phytophthora sojae* effector Avr1b into soybean cells does not require pathogen-encoded machinery. *Plant Cell*. 20:1930–1947.
- Drummond AJ, Rambaut A. 2007. BEAST: Bayesian evolutionary analysis by sampling trees. *BMC Evol. Biol.* 7:214.
- Duplessis S et al. 2011. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proc. Natl. Acad. Sci. U.S.A.* 108:9166–9171.
- Durand D, Halldórsson BV, Vernet B. 2006. A hybrid micro-macroevolutionary approach to gene tree reconstruction. *J. Comput. Biol.* 13:320–335.
- Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics*. 14:755–763.
- Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* 32:1792–1797.
- Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30:1575–1584.
- Felsner G et al. 2011. ERAD components in organisms with complex red plastids suggest recruitment of a preexisting protein transport pathway for the periplastid membrane. *Genome Biol Evol.* 3:140–150.
- Finn RD et al. 2010. The Pfam protein families database. *Nucleic Acids Res.* 38:D211–22.
- Fitch WM. 2000. Homology a personal view on some of the problems. *Trends Genet.* 16:227–231.
- Gijzen M, Nürnberger T. 2006. Nep1-like proteins from plant pathogens: recruitment and diversification of the NPP1 domain across taxa. *Phytochemistry*. 67:1800–1807.
- Gobler CJ et al. 2011. Niche of harmful alga *Aureococcus anophagefferens* revealed through ecogenomics.

- Proc. Natl. Acad. Sci. U.S.A. 108:4352–4357.
- Govers F, Bouwmeester K. 2008. Effector trafficking: RXLR-dEER as extra gear for delivery into plant cells. *Plant Cell*. 20:1728–1730.
- Govers F, Gijzen M. 2006. *Phytophthora* genomics: the plant destroyers' genome decoded. *Mol. Plant Microbe Interact*. 19:1295–1301.
- Göker M, Voglmayr H, Riethmüller A, Oberwinkler F. 2007. How do obligate parasites evolve? A multi-gene phylogenetic analysis of downy mildews. *Fungal Genet. Biol.* 44:105–122.
- Haas BJ et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature*. 461:393–398.
- Jaillon O et al. 2004. Genome duplication in the teleost fish *Tetraodon nigroviridis* reveals the early vertebrate proto-karyotype. *Nature*. 431:946–957.
- Jiang RHY et al. 2005. Elicitin genes in *Phytophthora infestans* are clustered and interspersed with various transposon-like elements. *Mol. Genet. Genomics*. 273:20–32.
- Jiang RHY, Tripathy S, Govers F, Tyler BM. 2008. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. *Proc. Natl. Acad. Sci. U.S.A.* 105:4874–4879.
- Jiao Y et al. 2009. A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. *Nat. Genet.* 41:258–263.
- Kale SD et al. 2010. External lipid PI3P mediates entry of eukaryotic pathogen effectors into plant and animal host cells. *Cell*. 142:284–295.
- Katoh K, Misawa K, Kumar S, Miyata T. 2002. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res.* 30:3059–3066.
- Keeling PJ. 2009. Chromalveolates and the evolution of plastids by secondary endosymbiosis. *J. Eukaryot. Microbiol.* 56:1–8.
- Kellis M, Birren BW, Lander ES. 2004. Proof and evolutionary analysis of ancient genome duplication in the yeast *Saccharomyces cerevisiae*. *Nature*. 428:617–624.
- Krogh A, Larsson B, Heijne von G, Sonnhammer EL. 2001. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.* 305:567–580.
- Latijnhouwers M, de Wit PJGM, Govers F. 2003. Oomycetes and fungi: similar weaponry to attack plants. *Trends Microbiol.* 11:462–469.
- Lévesque CA et al. 2010. Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire. *Genome Biol.* 11:R73.
- Maere S, Heymans K, Kuiper M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*. 21:3448–3449.
- Martens C, Van de Peer Y. 2010. The hidden duplication past of the plant pathogen *Phytophthora* and its consequences for infection. *BMC Genomics*. 11:353.
- Martens C, Vandepoele K, Van de Peer Y. 2008. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proc. Natl. Acad. Sci. U.S.A.* 105:3427–3432.
- Maruyama S, Matsuzaki M, Misawa K, Nozaki H. 2009. Cyanobacterial contribution to the genomes of the plastid-lacking protists. *BMC Evol. Biol.* 9:197.
- McLysaght A, Hokamp K, Wolfe KH. 2002. Extensive genomic duplication during early chordate evolution. *Nat. Genet.* 31:200–204.
- Morris PF et al. 2009. Multiple horizontal gene transfer events and domain fusions have created novel regulatory and metabolic networks in the oomycete genome. *PLoS ONE*. 4:e6133.
- Muller J, Szklarczyk D, Julien P, Letunic I, Roth A, Kuhn M, et al. 2010a. eggNOG v2.0: extending the evolu-

- tionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations. *Nucleic Acids Res.* 38:D190–5.
- Muller J, Creevey CJ, Thompson JD, Arendt D, Bork P. 2010b. AQUA: automated quality improvement for multiple sequence alignments. *Bioinformatics.* 26:263–265.
- Ospina-Giraldo MD, Griffith JG, Laird EW, Mingora C. 2010. The CAZyome of *Phytophthora* spp.: a comprehensive analysis of the gene complement coding for carbohydrate-active enzymes in species of the genus *Phytophthora*. *BMC Genomics.* 11:525.
- Patterson D. 1999. The Diversity of Eukaryotes. *American Naturalist.* 154:96–124.
- Richards TA et al. 2011. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. *Proc. Natl. Acad. Sci. U.S.A.*
- Richards TA, Dacks JB, Jenkinson JM, Thornton CR, Talbot NJ. 2006. Evolution of filamentous plant pathogens: gene exchange across eukaryotic kingdoms. *Curr. Biol.* 16:1857–1864.
- Richards TA, Talbot NJ. 2007. Plant parasitic oomycetes such as *Phytophthora* species contain genes derived from three eukaryotic lineages. *Plant Signal Behav.* 2:112–114.
- Runge F et al. 2011. The inclusion of downy mildews in a multi-locus-dataset and its reanalysis reveals a high degree of paraphyly in *Phytophthora*. *IMA Fungus.* 2:163–171.
- Seidl MF, Van den Ackerveken G, Govers F, Snel B. 2011. A domain-centric analysis of oomycete plant pathogen genomes reveals unique protein organization. *Plant Physiol.* 155:628–644.
- Sonnhammer ELL, Koonin EV. 2002. Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet.* 18:619–620.
- Spanu PD et al. 2010. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science.* 330:1543–1546.
- Stamatakis A. 2006. RAXML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics.* 22:2688–2690.
- Stassen JH, Van den Ackerveken G. 2011. How do oomycete effectors interfere with plant life? *Curr Opin Plant Biol.* 14:1–8.
- Stiller JW, Huang J, Ding Q, Tian J, Goodwillie C. 2009. Are algal genes in nonphotosynthetic protists evidence of historical plastid endosymbioses? *BMC Genomics.* 10:484.
- Storey JD, Tibshirani R. 2003. Statistical significance for genomewide studies. *Proc. Natl. Acad. Sci. U.S.A.* 100:9440–9445.
- Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE.* 6:e21800.
- Tatusov RL, Koonin EV, Lipman DJ. 1997. A genomic perspective on protein families. *Science.* 278:631–637.
- Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O. 2001. Towards a reliable objective function for multiple sequence alignments. *J. Mol. Biol.* 314:937–951.
- Thompson JD, Thierry JC, Poch O. 2003. RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics.* 19:1155–1161.
- Tyler BM et al. 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science.* 313:1261–1266.
- Van Dongen S. 2000. A cluster algorithm for graphs. Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam.
- Whisson SC et al. 2007. A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature.* 450:115–118.
- Woehle C, Dagan T, Martin WF, Gould SB. 2011. Red and problematic green phylogenetic signals among thousands of nuclear genes from the photosynthetic and apicomplexa-related *Chromera velia*. *Genome*

Biol Evol. 3:1220–1230.

Zdobnov EM, Campillos M, Harrington ED, Torrents D, Bork P. 2005. Protein coding potential of retroviruses and other transposable elements in vertebrate genomes. *Nucleic Acids Res.* 33:946–954.







# 4

## **Bioinformatic Inference of Specific and General Transcription Factor Binding Sites in the Plant Pathogen *Phytophthora infestans***

Michael F Seidl<sup>1,2</sup>, Rui-Peng Wang<sup>1,3</sup>, Guido  
Van den Ackerveken<sup>2,4</sup>, Francine Govers<sup>2,5</sup>,  
Berend Snel<sup>1,2</sup>

<sup>1</sup>Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

<sup>2</sup>Centre for BioSystems Genomics, P.O. Box 98, 6700 AB Wageningen, The Netherlands

<sup>3</sup>Centre for Integrative Bioinformatics VU (IBIVU), VU University Amsterdam, De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands

<sup>4</sup>Plant-Microbe Interactions, Department of Biology, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

<sup>5</sup>Laboratory of Phytopathology, Wageningen University, P.O. Box 8123, 6700 EE Wageningen, The Netherlands

*PLoS ONE* 7(12):e51295 (2012)



## ABSTRACT

Plant infection by oomycete pathogens is a complex process. It requires precise expression of a plethora of genes in the pathogen that contribute to a successful interaction with the host. Whereas much effort has been made to uncover the molecular systems underlying this infection process, mechanisms of transcriptional regulation of the genes involved remain largely unknown.

We performed the first systematic *de novo* DNA motif discovery analysis in *Phytophthora*. To this end, we utilized the genome sequence of the late blight pathogen *Phytophthora infestans* and two related *Phytophthora* species (*P. ramorum* and *P. sojae*), as well as genome-wide *in planta* gene expression data to systematically predict 19 conserved DNA motifs. This catalog describes common eukaryotic promoter elements whose functionality is supported by the presence of orthologs of known general transcription factors. Together with strong functional enrichment of the common promoter elements towards effector genes involved in pathogenicity, we obtained a new and expanded picture of the promoter structure in *P. infestans*. More intriguingly, we identified specific DNA motifs that are either highly abundant or whose presence is significantly correlated with gene expression levels during infection. Several of these motifs are observed upstream of genes encoding transporters, RXLR effectors, but also transcriptional regulators. Motifs that are observed upstream of known pathogenicity-related genes are potentially important binding sites for transcription factors. Our analyses add substantial knowledge to the as yet virtually unexplored question regarding general and specific gene regulation in this important class of pathogens. We propose hypotheses on the effects of *cis*-regulatory motifs on the gene regulation of pathogenicity-related genes and pinpoint motifs that are prime targets for further experimental validation.

## INTRODUCTION

Oomycetes are an important class of eukaryotic pathogens that have severe ecological and economic impact (Govers & Gijzen 2006), which only recently entered the genomic era (Judelson 2007; 2012). The genus *Phytophthora* contains several well-known species such as the potato and tomato late blight pathogen *Phytophthora infestans* (Haas et al. 2009), the stem and root pathogen of soybean *Phytophthora sojae* (Tyler et al. 2006; Tyler 2007) and the sudden oak death pathogen *Phytophthora ramorum* (Tyler et al. 2006; Grunwald et al. 2012). The genome sequence of these pathogens facilitated insights into the large repertoire of proteins involved in the interaction with the host (Stassen & Van den Ackerveken 2011). For example, proteins containing the amino-acid motifs RXLR and LXLFLAK (Crinkler) belong to two distinct classes of effectors that are targeted to the inside of the plant cell presumably to promote infection of the host (Whisson et al. 2007; Jiang et al. 2008; Haas et al. 2009). Elicitins (ELIs) are proteins that elicit defense responses and induce necrosis whereas the related elicitin-like proteins (ELLS) do not exhibit such an activity (Jiang et al. 2006). Present hypotheses on the functions of ELLS are still inconclusive, but some members seem to be associated with the cell wall or the cell membrane (Jiang et al. 2006). Genes encoding effectors and also other proteins that are involved in the host-pathogen interaction require a precise spatial and temporal expression to facilitate the successful colonization of the host.

There is rich and continuously expanding knowledge on the regulation of the spatio-temporal expression of genes in human and in eukaryotic model organisms such as yeast and fruit fly (e.g. Singer et al. 1990; Kutach & Kadonaga 2000; Majewski & Ott 2002; Müller et al. 2007; Yang et al. 2007; Hahn & Young 2011; Hoskins et al. 2011). In eukaryotes, regulation of transcription is accomplished by the complex interplay of several elements. These include DNA motifs in the upstream regions of genes (*cis*-regulatory elements), which are bound by diverse transcription factors, and the remodeling of the chromatin structure. Elements in proximity to the transcription start site (TSS) include the eukaryotic core promoter elements as well as specific regulatory elements. The basic transcriptional activity is determined by the eukaryotic core promoter, which is typically present within 70 nucleotides (nt) surrounding the transcription start site and directs the mediator complex, general transcription factors, and the RNA polymerase II (RNA Pol II) into a functional pre-initiation complex (Verrijzer & Tjian 1996; Roeder 1998; Woychik & Hampsey 2002). The core promoter in many eukaryotes consists of different combinations of functional DNA motifs: the transcription factor-B recognition element (BRE), followed by the TATA-box, the initiator (Inr) (located at or around the TSS), and the downstream promoter elements (DPE). The CCAAT-box, another common eukaryotic promoter element, mainly occurs upstream of the core promoter elements. In contrast to other eukaryotes, oomycetes seem to lack canonical TATA-box elements (Judelson et al. 1992). However, many genes have an Inr-element that resembles the general eukaryotic Inr-element (Pieterse et al. 1994; McLeod et al. 2004); an element that is sufficient to direct the accurate transcription in the absence of other elements (Purnell et al. 1994; Javahery et al. 1994; McLeod et al. 2004). Interestingly, oomycetes have a flanking promoter region (FPR) downstream of the Inr-element that has not yet

been described as an important functional region in other eukaryotes (McLeod et al. 2004). Our knowledge on specific promoter elements in oomycetes is limited: The upstream regions of the sporulation-specific genes *Cdc14* and *Pks1* contain, next to the Inr- or Inr/FPR- element, specific but distinct elements that are required for correct gene expression (Ah-Fong et al. 2007; Xiang et al. 2009). Additionally, a short (7 nt) motif named cold-box mediates temperature-induced expression of zoosporogenesis-specific genes (Tani & Judelson 2006). This small number of experimentally characterized *cis*-regulatory elements in *Phytophthora* is in sharp contrast to the abundance of the predicted genes encoding the diversity of transcription factors in *Phytophthora* and related non-pathogenic species (Supplementary Table S4-1A) (Rayko et al. 2010). This raises questions about the nature and abundance of the accompanying and not yet described *cis*-regulatory elements in the genomes of *Phytophthora* spp.

To expand our knowledge on the transcriptional regulation in *Phytophthora* spp., we systematically inferred and analyzed DNA motifs. We adopted *in silico* methodologies that have been successfully applied to other eukaryotic pathogens, such as the malaria parasite *Plasmodium falciparum* (van Noort & Huynen 2006), and plants (Vandepoele et al. 2006). It is assumed that co-expressed genes share similar *cis*-regulatory motifs (Roth et al. 1998) and that functional motifs are conserved both within and between species to a higher extent than non-functional DNA. With the availability of genomic and transcriptomic data of several *Phytophthora* spp. (Tyler et al. 2006; Haas et al. 2009) similar methodologies can now also be applied to analyze *cis*-regulatory motifs in these important plant pathogens. We combined the upstream regions of co-expressed genes in *P. infestans* with the upstream regions of their orthologs in *P. sojae* and *P. ramorum* and predicted in total 19 motifs. The analysis of this repertoire revealed a complex picture of the *Phytophthora* promoter and allowed the identification of biologically relevant motifs. Several of these motifs are predicted upstream of genes encoding known effector genes or transcriptional regulators, e.g. Myb-like transcription factors. These motifs thus represent interesting candidates for further experimental validation. Hence, our study represents the first systematic characterization of *cis*-regulatory elements in *Phytophthora* spp. and expands our knowledge on the regulation of gene expression in this important class of pathogens.

## MATERIAL & METHODS

### Identification of Co-expressed *P. infestans* Genes

We retrieved NimbleGen microarray data of *P. infestans* containing three *in vitro* stages (different media types) and four *in planta* stages (Haas et al. 2009) from GEO (Barrett & Edgar 2006). The initial analysis and summary of the NimbleGen data has been described by Haas et al. (Haas et al. 2009). Differentially expressed genes during *in planta* growth were identified using t-tests between two groups (group A, different media types; group B replicates for a single data point post inoculation). The tests

were independently applied for each day after inoculation and genes were deemed significantly differentially expressed (up- and down-regulated) with a p-value cutoff of 0.05. False discovery rates were assessed by computing q-values (q-value cutoff of 0.05) for each comparison. Subsequently, the identified significantly differentially expressed genes were clustered based on their expression profiles, i.e. intensities relative to the average expression intensity in growth media, using Spearman correlation coefficient utilizing the Markov clustering algorithm (version 09-308, 1.008, inflation 5) (Van Dongen 2000; Enright et al. 2002). The cutoff for the correlation coefficient was empirically determined by computing the distribution of Spearman correlation coefficients between 1,000 randomly drawn *P. infestans* genes. The correlation coefficient cutoff was determined by the 95 percent quantile, corresponding to value of 0.86. Single, non-clustered genes were discarded before further analysis.

#### **Identification of Orthologs and Extraction of the 1 kb Upstream Regions in *Phytophthora* spp.**

We identified orthologs (exclusive in-paralogs within *P. infestans*) of all predicted proteins in the analyzed *Phytophthora* spp. using OrthoMCL (version 2.0; default settings; e-value cutoff  $1 \times 10^{-5}$ ) (Li et al. 2003). OrthoMCL covers the vast majority of the predicted proteome by grouping on average 84 percent of the predicted proteins into orthologous groups, ranging from 77 percent for *P. infestans* to 91 percent in *P. ramorum*. Subsequently, we combined the upstream regions of co-expressed *P. infestans* genes (clusters with size  $\geq 2$ ) with their orthologs in *P. ramorum* and *P. sojae* (inclusive in-paralogs) and used these to identify conserved DNA motifs. The upstream region per gene was defined as the 1,000 nt upstream of the translation start site 'ATG' as annotated by the coding sequence. Upstream sequences without an associated annotated coding gene were discarded. If a coding gene occurred within the 1,000 nt, the upstream region was truncated. For genes located on the negative strand the extracted DNA sequence was converted to its reverse complement. The upstream regions were filtered for the remnants of non-annotated genes by similarity search against the NCBI nr database (downloaded 24.10.2011, blastx (Altschul et al. 1990); e-value cutoff  $1 \times 10^{-3}$ ) and the presence of transposable elements identified by TransposonPSI and against the Repbase database (Jurka et al. 2005) (downloaded 19.01.2012, blastn; e-value cutoff  $1 \times 10^{-3}$ ). Subsequently, all significant hits within the sequences were masked from all further analyses. Furthermore, we tried to reduce the number of false positives during motif prediction by removing highly similar upstream regions as defined by 95 percent identity over an area of at least 50 percent of the length of the informative sequences (one of the sequences was retained).

#### **Identification of DNA Motifs Within Clusters of Co-expressed *P. infestans* Genes and their Orthologs in *P. ramorum* and *P. sojae***

DNA motifs in the upstream regions of different clusters of co-expressed *P. infestans* genes and their orthologs (clusters with size  $\geq 5$ ) were identified using the expectation maximization algorithm implemented in MEME (version 4.6.1; e-value cutoff 1) (Bailey

& Elkan 1994). We applied the zoops model allowing for zero or single occurrence of a motif per upstream region, inclusion of the reverse complement DNA strand in the motif identification, a motif length between 4-16 nt, maximally 30 distinct motifs per cluster of co-expressed genes and an empirical 3<sup>rd</sup> order background Markov model based on the upstream region of all *Phytophthora* spp. genes (this background model is also used for all other analyses).

Similar motifs were clustered into families based on their pairwise similarity using the Markov clustering algorithm (inflation 2). Combined motif logos were produced using Weblogo 3 (Crooks et al. 2004). The genome-wide abundance of each motif-family was predicted per individual motif constituting the motif-family and the combined motif using FIMO (part of the MEME/MAST package) (Grant et al. 2011). FIMO calculates a score for each position within the searched sequence based on the position-specific frequency matrix of the *ab initio* determined motifs. These scores are transformed to p-values and subsequently to q-values to address false discovery rates due to multiple testing. We applied a q-value cutoff 0.1 to define the genome-wide abundance for each motif. The location of the motifs in the upstream regions is displayed for the first 1,000 nt using bins of the size 50. To account for shorter upstream regions due to coding genes within the first 1,000 nt, the abundance was weighted accordingly. Similarity to known motifs was assessed using Tomtom (e-value 0.5; min overlap between motifs 3) (Gupta et al. 2007) against the JASPAR Core and JASPAR PolIII database.

To estimate the evolutionary conservation of the identified motifs, we calculated a conservation score that is based on the network-level conservation principle (Pritsker et al. 2004; Elemento & Tavazoie 2005). Assuming that the global gene expression between two closely related species is largely conserved, the network-level conservation principle requires that most of the target sites, i.e. the DNA motifs, are retained. Therefore, we identified the presence of each motif in the upstream regions of orthologous groups between two of the *Phytophthora* spp. (as determined by OrthoMCL groups, see above). We subsequently calculated the number of cases where both orthologous groups maintained the motif and assessed the significance of the overlap (Fisher exact test, conservation scores are reported as the  $-\ln$ ). The values were compared to a set of randomized motifs (the column of each identified motif was shuffled twenty times); the poly-C motif-6 was excluded for this and all subsequent analyses. As expected, the majority of these motifs did not yield any significant hits against the *Phytophthora* upstream regions. Based on the motifs with significant hits we chose the 95 percent quantile as a conservation cutoff, corresponding to a p-value of 0.04. Applying this cutoff to the set of motif families yields a conserved subset that exceeds this score between *P. infestans* and at least one of the other *Phytophthora*.

The identified motifs, their genome wide abundance, their conservation score and location (global as well as per individual gene) are accessible as 'Supplementary data S4-1'.

### Correlation of Conserved DNA Motifs with Gene Expression Levels upon Infection

Functional *cis*-regulatory motifs are DNA elements that modulate the expression of genes upon binding of a transcription factor. They were identified in *P. infestans* by searching for motifs where their presence within the upstream regions significantly correlates with expression levels of the downstream genes similar to the approach outlined by Bussemaker and colleagues (2001). We searched the upstream region of each of the differentially expressed genes for the binding of one of the individual members of the motif-family using FIMO (default settings, no q-value computation) (Grant et al. 2011). For each motif, we retrieved the maximum score per motif-family; the score per hit is defined by the sum of the entries of the position specific scoring matrix. Subsequently, the maximum score is scaled based on the length of the highest scoring motif and the scores for each motif was rescaled in the range [0,10] resulting in a scoring matrix with the dimensions of the number of differentially expressed genes times the number of motifs. Significantly correlated motif scores with the expression level at one of the three different time points (2-4 dpi), expressed as the  $\log_2$ -fold change compared to the growth media, were identified by forward variable selection as implemented in R and multiple testing correction was applied to the p-values by computing q-values (false discovery rate). Motifs with a q-value < 0.01 were deemed significant. For each motif in each condition a 'time course value' (T-value) was calculated: the correlation between the motif score and the expression level at each time point (growth media+2-5 dpi) was transformed into a T-value by multiplying the correlation (r) with the square root of the number of genes (G) ( $T=r*\sqrt{G}$ ) (van Noort & Huynen 2006).

### Functional Annotation of Genes in the Three Analyzed *Phytophthora* spp.

Genes in the analyzed *Phytophthora* spp. were functionally annotated using BLAST2GO algorithm (default parameters) (Conesa et al. 2005). Functional enrichment of GO terms of genes sharing predicted motifs was conducted with the BiNGO package 2.44 (default parameters) (Maere et al. 2005) included in Cytoscape 2.8.1 (Smoot et al. 2011). Significantly enriched GO terms were summarized by removing redundancies using REVIGO (similarity cutoff 0.5) (Supek et al. 2011). Moreover, additional annotation for genes such as RXLRs, Crinklers, elicitors, and elicitor-like was added based on the annotation provided by Haas et al. (2009), Jiang et al. (2006) and the BROAD website (<http://www.broadinstitute.org/>). Significance of this overrepresentation was assessed using Fisher exact test (p-value cutoff 0.05).

### Identification of Known Transcription Factors Binding Common Eukaryotic DNA Elements

Known transcription factors that bind to common eukaryotic promoter elements were identified by determining orthologs of the human genes (proteins) in oomycetes using OrthoMCL (version 2.0; default settings; e-value cutoff  $1 \times 10^{-5}$ ) (Li et al. 2003). The version and source of the nineteen proteomes included in this analysis are shown in Supplementary Table S4-1C. In the case of CBF-B, OrthoMCL clustered the human

gene solitarily and the orthologs of the *Arabidopsis thaliana* *CBF-B* gene were reported.

### Description of the Transcription Factor Repertoire in *Phytophthora* spp.

We predicted the repertoire of potential transcription factors in the proteomes of the three analyzed *Phytophthora* spp. and four non-pathogenic sister taxa (Supplementary Table S4-1C) using Pfam models that describe DNA binding sites. The majority of models have been obtained from DBD (Wilson et al. 2008) and some, e.g. Myb-like DNA binding domain, have been added manually (see Supplementary Table S4-1A for details). Domains were identified using HMMER3, applying the gathering cutoff (Eddy 1998).

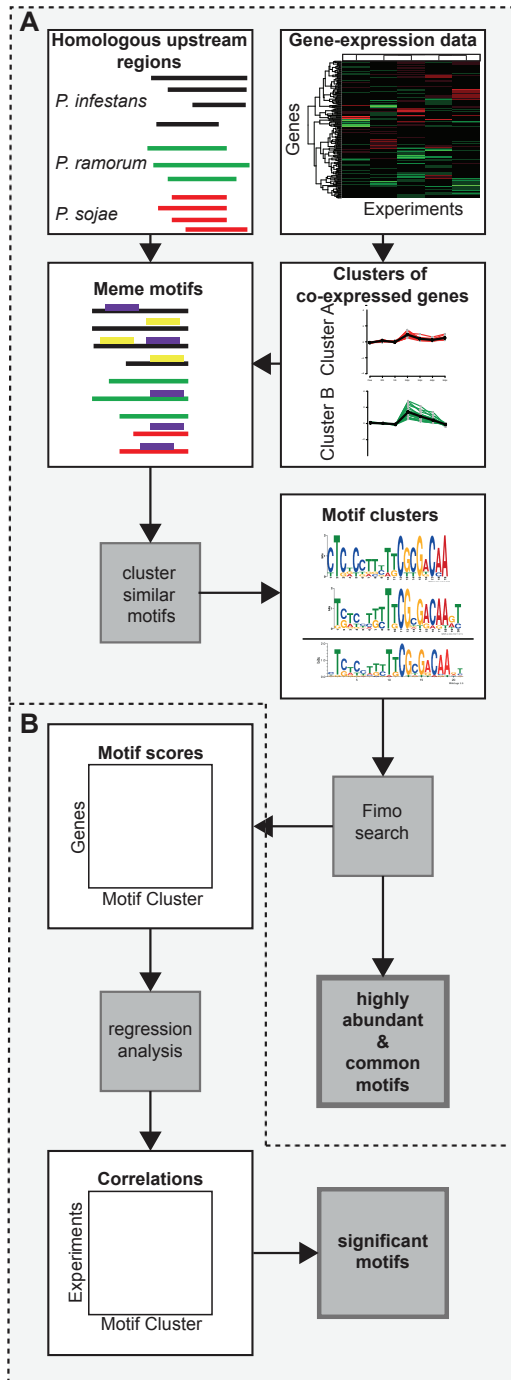
## RESULTS

### Identification of Conserved DNA Motifs in Promoters of *Phytophthora* genes

To predict potential *cis*-regulatory elements in the upstream regions of *Phytophthora* genes, we assumed that co-expressed genes are co-regulated by shared *cis*-regulatory elements (Roth et al. 1998). In total, 1,667 differentially expressed *P. infestans* genes were selected from NimbleGen microarray data of *in vitro* growth (three plant extract media) and *in planta* growth (four conditions) 2-5 days post inoculation (dpi) of potato plants (Haas et al. 2009). The first three conditions (2-4 dpi) coincide with the formation of haustoria, specialized infection structures that are formed inside the plant cells. The later stage of infection (5 dpi) corresponds to necrotrophic growth on dead plant material where the expression of many genes show similar expression profiles to growth in plant extract media (Haas et al. 2009). By clustering the expression profiles of the differentially expressed *P. infestans* genes using Spearman correlation and a graph based clustering algorithm (MCL) (Van Dongen 2000), we obtained 159 groups of co-expressed genes (Figure 4-1A; Material & Methods). For each gene within the co-expressed cluster we identified orthologs in two related species (*P. sojae* and *P. ramorum*) and filtered the upstream regions for remnants of transposable elements (see Material & Methods).

Within 136 co-expressed clusters, we identified 80 motifs representing putative regulatory DNA elements in the upstream regions of co-expressed *P. infestans* genes and their orthologs. Similar motifs, especially common eukaryotic DNA elements, were identified in different clusters of co-expressed genes. Hence, we grouped the total of 80 motifs into 24 distinct motif families (called 'motifs' throughout the remainder of the manuscript), based on the assumption that all motifs within a family represent a binding site for a specific DNA binding protein or complex. To enrich our results for conserved functional DNA motifs, these were filtered by applying an evolutionary conservation filter between *P. infestans* and at least one of the other *Phytophthora* yielding 19 conserved DNA motifs for which the genome-wide abundance was determined using FIMO (Data Supplementary S4-1; Material & Methods).





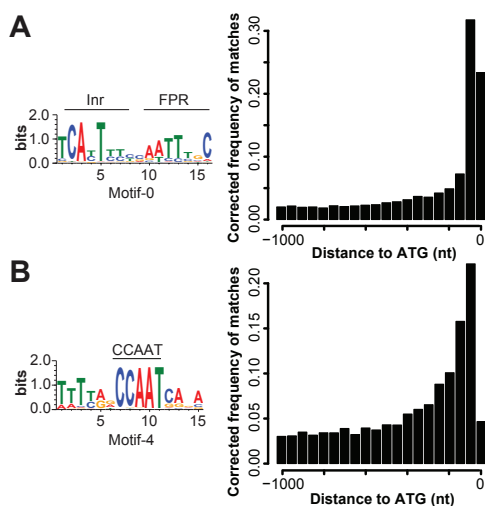
**Figure 4-1 Analysis pipeline used to identify conserved DNA motifs in the three analyzed *Phytophthora* spp.**

(A) Co-expressed *P. infestans* genes were identified and their upstream regions were combined with the ones from orthologous genes in *P. sojae* and *P. ramorum*. In total 80 motifs were identified and similar motifs were grouped into 24 motif families of which 19 remained after conservation filtering. These were automatically and manually inspected for similarity to known eukaryotic promoter elements. (B) To further assess the biological relevance of the motifs, scores describing the occurrence of motifs in each individual upstream region were assigned. The motif score was correlated with the gene expression level of the downstream genes; an approach similar to the one applied by Bussemaker and colleagues (2001). Subsequently, motifs that have a significant correlation with the expression level of genes during infection were identified ( $q < 0.01$ ).

### Promoters of *Phytophthora* Contain Common Eukaryotic Promoter Elements in High Abundance

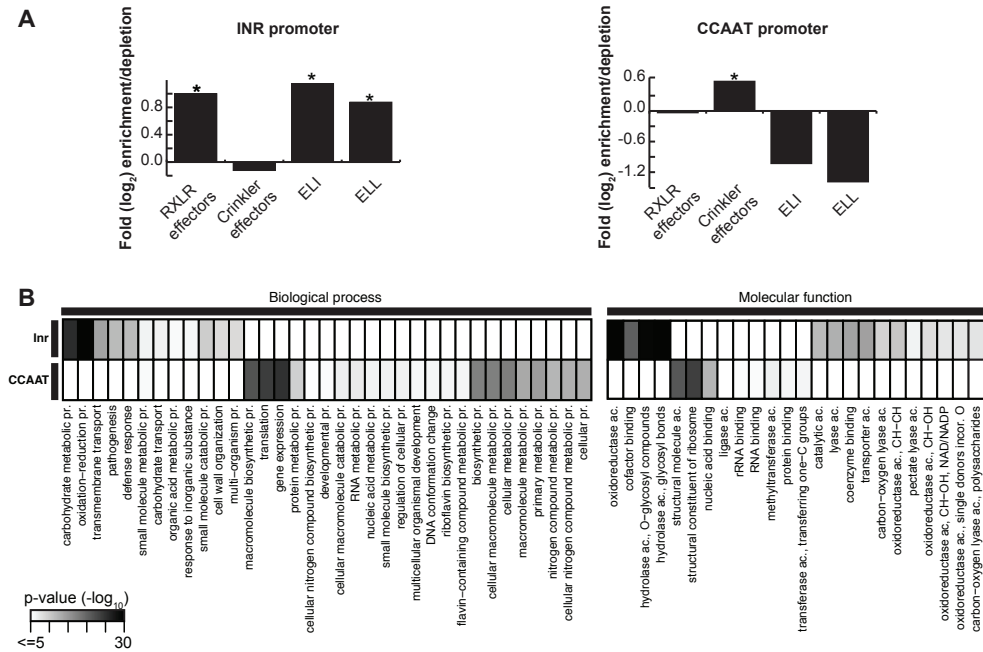
To validate our method, we first surveyed the 19 obtained motifs for similarity to known eukaryotic promoter elements. Pre-genome analyses of the upstream regions of a small set of oomycete genes have identified a Inr/FPR-element as a core promoter element (Pieterse et al. 1994; McLeod et al. 2004). Indeed, our *in silico* approach recovered the previously described oomycete-specific Inr/FPR element (motif-0). In the set of 1,493 *P. infestans* genes included in the motif search it occurs 652 times. Genome-wide, the Inr/FPR-element is the most abundant motif (Material & Methods): It is predicted in 18,138 upstream regions of all annotated genes in the three analyzed *Phytophthora* spp., and in 6,511 or 37 percent of all *P. infestans*. It has a distinct localization at a median of 81 nt upstream of the translation start site (TLS) (Figure 4-2A). In other eukaryotes, the transcription factors TAF1 and TAF2 are associated with the transcription factor II D complex during the initiation of transcription at the Inr-element (Chalkley & CP 1999; Müller et al. 2007). *Phytophthora* spp. and also other oomycetes have TAF1 and TAF2 orthologs (Supplementary Table S4-1B), suggesting the association of these transcription factors with the Inr/FPR-element and further supporting its role as a ubiquitous core promoter element in oomycetes.

Among the 19 automatically derived motifs, motif-4 is significantly similar to the eukaryotic CCAAT-box, also named NFYA- or CBF-B binding box (Figure 4-2B & Supplementary Table S4-2). This common eukaryotic DNA element was so far only reported for a few *Phytophthora* genes (Judelson & Michelmore 1989). In the three *Phytophthora* spp., we predicted the CCAAT-box in the upstream regions of 8,225 genes, 3,418 of which are from *P. infestans*. It is primarily localized at 192 nt upstream of the TLS. When the CCAAT-box co-occurs with the Inr/FPR-elements (3,321 genes), these motifs are approximately 180 nt apart (80nt for the 25th percentile). Interestingly, we found more



**Figure 4-2 Common eukaryotic promoter elements in *Phytophthora*.**

Sequence motif of the (A) Inr/FPR-elements and the (B) CCAAT-box identified in the upstream regions of *Phytophthora* spp. genes. The location of the motif in relation to the TLS is indicated by bar charts (bin size 50 nt). The frequency of the motif per bin was weighted according to the underlying length distribution of the upstream regions.



**Figure 4-3** Enrichment/depletion of functional classes in the set of genes with Inr/FPR- or CCAAT-box elements.

(A)  $\log_2$ -fold enrichment/depletion displayed for four classes of genes (RXLR, Crinkler, ELI and ELL) predicted to contain the Inr/FPR- or the CCAAT-box in their promoter sequence. (B) Overrepresentation of GO functional annotation of genes that contain the Inr/FPR- or the CCAAT-box elements in their promoter sequence. Heat map shows the  $-\log_{10}$ (p-value) of the significant enrichments detected by BINGO (Maere et al. 2005). Non-redundant GO terms (see Material & Methods) with a  $-\log_{10}$ (p-value) > 5 are displayed (see Supplementary Table S4-3A for the full list).

occurrences of the CCAAT-box on the negative strand than on the positive strand (4,310 vs. 3,915), consistent with the observation that the CCAAT-box is found in both orientations (Mantovani 1998; Maity & De Crombrugge 1998). The CCAAT-box binding factor is a heterotrimeric protein complex composed of CBF-A, CBF-B and CBF-C (Kim et al. 1996). We found orthologs of all three CBF-encoding genes in all *Phytophthora* and in other oomycetes species analyzed (Supplementary Table S4-1B), showing additional support for a function of this motif in the regulation of gene expression in oomycetes.

### Enrichment of Distinct Functional Classes in the Sets of Genes with Common Eukaryotic Promoter Elements

To assess whether the described common eukaryotic promoter elements are observed upstream of distinct set of genes, we searched for enrichment of functional Gene Ontology categories as well as other classes associated with host-pathogen interaction, e.g. RXLR effector genes. The sets of genes of which the upstream region contains the Inr/FPR-element or the CCAAT-box are enriched for different functional

categories (Figure 4-3 & Supplementary Table S4-3A). The set with the Inr/FPR-element is highly enriched for RXLR effector, ELI- and ELL genes, and also other genes with predicted functions in pathogenesis, carbohydrate metabolism, glycoside hydrolysis-, oxidoreductase, lyase- or transporter activity; many have a predicted extracellular localization (Supplementary Table S4-3A). Strikingly, 869 of the 1107 predicted RXLR effector genes in the three *Phytophthora* spp. contain the Inr/FPR-element in their upstream regions including several up-regulated RXLR effectors (Table 4-1). In contrast, the set of genes with promoters that contain the CCAAT-box is depleted of RXLR effector, ELI and ELL genes and enriched for Crinkler genes (160 out of 600 Crinkler genes). Furthermore, the CCAAT-box set is enriched for genes encoding proteins with a predicted intracellular localization, as well as gene products involved in gene expression, translation, reproduction and developmental- or metabolic processes. The surprisingly strong adjustment of common eukaryotic promoter elements, such as the Inr/FPR, towards pathogenicity and the strong, opposing functional enrichment of genes regulated by either CCAAT-box or Inr/FPR-element is yet another striking example of successful genome adaptation towards pathogenicity within *Phytophthora*.

### Candidate *Cis*-regulatory Elements that Correlate with Gene Expression Levels upon Infection

To further assess the functional significance of the 19 motifs, we correlated the gene expression levels of the differentially expressed genes with the occurrence of the motifs with a regression-based approach similar to the one described by Bussemaker and colleagues (Bussemaker et al. 2001) (Figure 4-1B; Material & Methods). Four motifs show

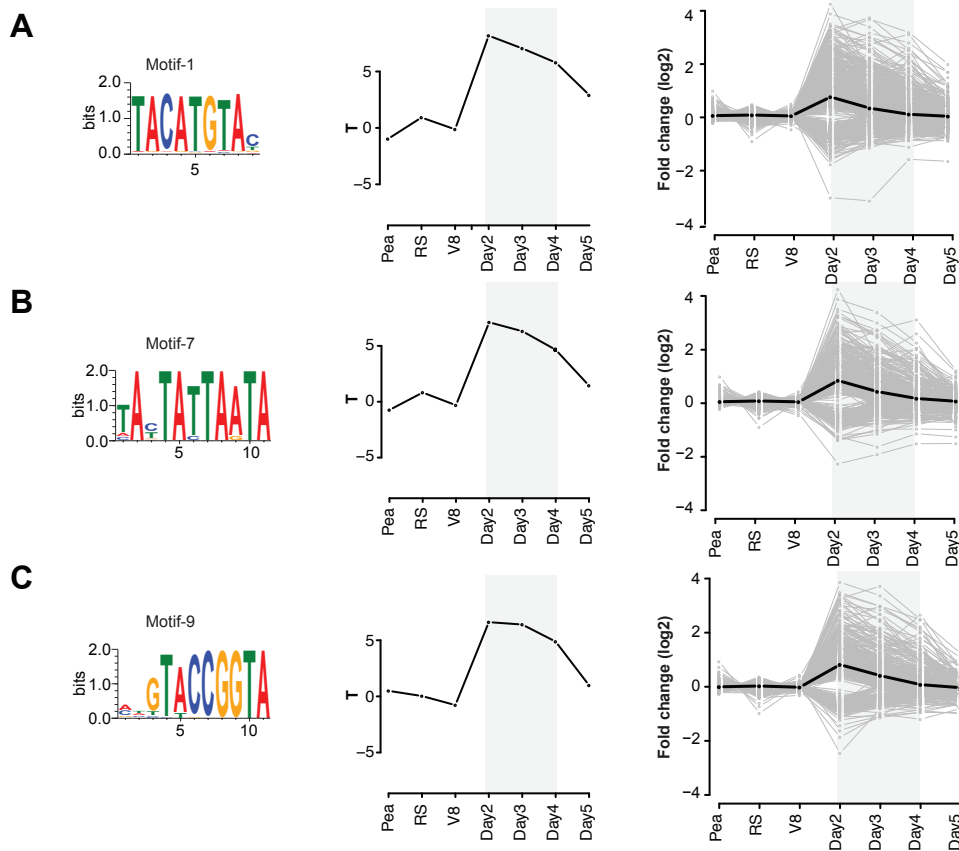
**Table 4-1 Motifs in the upstream regions of a subset of differentially expressed *P. infestans* genes.**

Motif	Gene Description	Gene ID	Position <sup>§</sup>	Sequence*	Fold (log <sub>2</sub> )			
					2dpi	3dpi	4dpi	5dpi
Motif-0	avr4	PITG_07387	-36	cgagc <b>TCA</b> GTCTTCAATTCTCcttt	2.06	1.42	0.90	0.37
Motif-0	Pred. RXLR effector	PITG_00821	-62	cagca <b>TCA</b> TCTTCAACTCGCaacac	0.80	0.23	-0.002	0.12
Motif-0	Pred. RXLR effector	PITG_02860	-49	cacc <b>TCA</b> TCTTCAATTCTTcgact	1.68	1.97	1.15	0.26
Motif-0	Pred. RXLR effector	PITG_12057	-24	agtag <b>TCA</b> TTCGCTTCTGCaggtg	2.27	1.65	1.15	0.30
Motif-1	avr1 family	PITG_16663	-389	cgaac <b>TAC</b> ATGTATATccccgc	1.42	0.99	0.38	0.07
Motif-1	avr2 family	PITG_08278	-119	ctcag <b>TAC</b> ATGTAAccccgc	0.98	0.96	0.42	0.39
Motif-1	Mannitol dehydrogenase	PITG_00972	-352	acatg <b>TAC</b> ATGTATtaata	2.06	2.67	0.85	-0.11
Motif-1	Polygalacturonase	PITG_21247	-108	gatgg <b>TAC</b> ATGTACacggg	0.67	0.01	-0.21	0.13
Motif-3	Pred. RXLR effector	PITG_09218	-108	ttgaa <b>TG</b> CAAATACTAAGTCAaactg	2.48	1.89	1.15	-0.59
Motif-3	avrblb2 family	PITG_20303	-181	aagtc <b>T</b> ACTAATATCAAGTCgatt	2.02	2.03	1.34	0.23
Motif-3	Myb-like transcription factor	PITG_00513	-709	agtta <b>ACT</b> TGTTTTATGTAAGtccca	0.73	0.58	0.33	0.32
Motif-3	Aldose1-epimerase	PITG_14720	-279	aagta <b>TAC</b> AGAAGTCAAGTCAaatga	1.86	0.96	0.55	-0.50

Predicted motifs in the upstream region of a selected subset of differentially expressed *P. infestans* genes. The fold (log<sub>2</sub>) expression change is compared to the average expression level in growth media.

<sup>§</sup>Position of the start of the motif relative to the translation start site (TLS) of the gene.

\*Sequence of the motif (capital letters and bold) and 5 nt upstream and downstream of the motif (small letters) are shown for *P. infestans* genes.



**Figure 4-4 Three *cis*-regulatory elements that correlate with gene expression levels during infection.**

Nucleotide conservation of (A) motif-1, (B) motif-7 and (C) motif-9 is displayed as sequence logos. The T-values for each motif are displayed for each data point as well as gene expression of all differentially expressed genes that contributed to the correlation (see Material & Methods) are displayed.

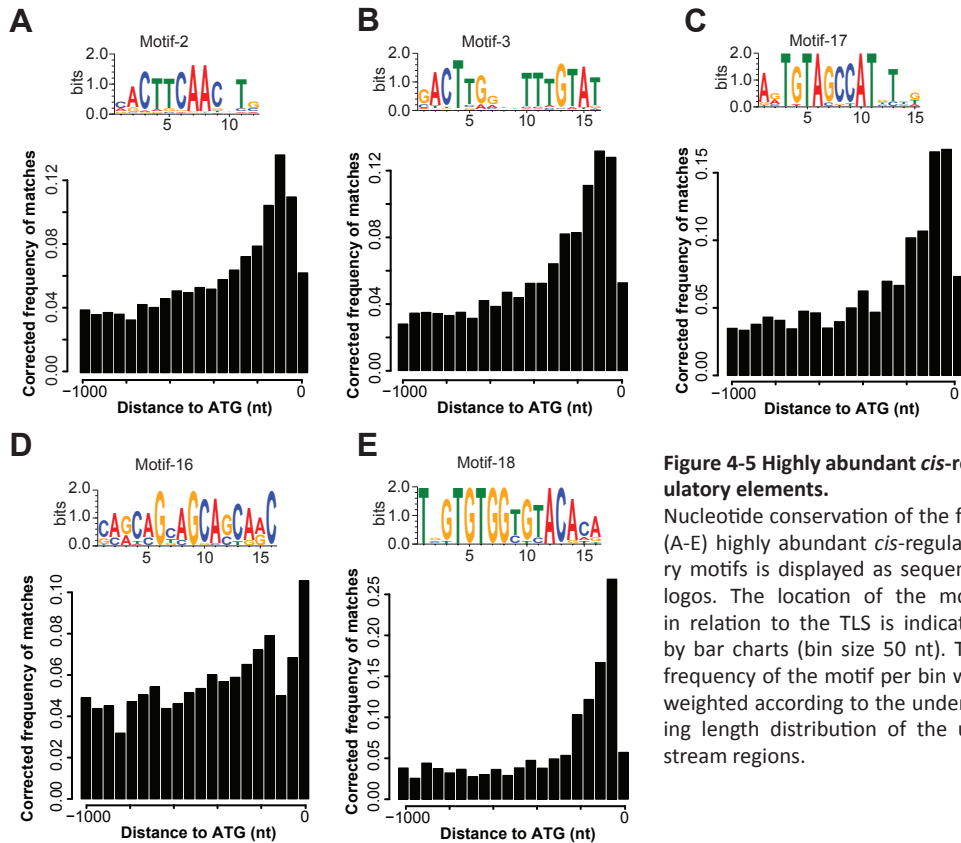
significant positive correlation between the level of motif occurrence and expression levels at one or more of the three time points post inoculation (2-4 dpi;  $q < 0.01$ ). Hence, these four motifs are likely functional binding sites for transcription factors and involved in the regulation of expression of the upstream genes during infection.

We identified a novel DNA motif (motif-1) that does not show any significant similarity to known motifs as determined by a Tomtom search against the JASPAR database (Supplementary Table S4-2). Motif-1 is a highly abundant and conserved motif that is characterized by the consensus inverted repeat sequence TACATGTA and is identified in total in the upstream regions of 12,070 *Phytophthora* genes, 44% of which are from *P. infestans* (Figure 4-4A). The inverted repeat structure is suggestive of a binding site for a homodimeric transcription factor. The presence of motif-1 is significantly correlated with the up-regulation of *P. infestans* genes at 2-4 dpi. Interestingly, the set of genes

that contain this motif in their upstream region is enriched in genes encoding RXLR effectors and genes involved in cell wall organization, carbohydrate metabolism as well as for genes that encode catalytically active proteins, e.g. glycosyl-hydrolases and oxidoreductases (Figure 4-4A & Table 4-1 & Supplementary Table S4-3). The differentially expressed mannitol-dehydrogenase gene (PITG\_00972) is an example of an oxidoreductase within the enriched class of catalytic enzymes that is up-regulated early during infection (Table 4-1). Mannitol can suppress ROS-related plant responses upon secretion in the apoplast and could act as a carbohydrate reservoir (Lewis & Smith 1967; Chaturvedi, Wong, et al. 1996b; Chaturvedi, Flynn, et al. 1996a; Voegelé et al. 2005). It has been suggested that mannitol-dehydrogenases (e.g. MAD1) in the biotrophic fungal plant pathogen *Uromyces fabae* are responsible for the production of mannitol in haustoria (Voegelé et al. 2005), an activity that could also occur in oomycete pathogens. Another example of a stress response gene is a highly expressed (11-fold increased expression at 2 dpi) secreted catalase-peroxidase (PITG\_07143) that could act in counteracting the burst of reactive oxygen species (ROS) by the plant as a defense mechanism upon pathogen infection (Mittler et al. 2004).

We also identified an inverted repeat, AT-rich motif (motif-7) in 1,388 *Phytophthora* genes, 940 of which are from *P. infestans* (Figure 4-4B). This motif shows remote similarity to the eukaryotic TATA-box; a eukaryotic core promoter element that is found in the upstream regions of a quarter of all genes in yeast and human (Yang et al. 2007). Previous analyses of the transcriptional regulation of oomycetes have indicated that oomycete promoters do not contain a canonical TATA-box (Judelson et al. 1992), however non-canonical TATA-box elements that resemble functional TATA-box elements have been discovered in oomycetes before (Judelson & Michelmore 1989; Škalamera & Hardham 2006). Unlike the Inr/FPR element and CCAAT-box, the TATA-like motif does not have a strong positional preference compared to the canonical TATA-box observed in other eukaryotes or the Inr/FPR-element and CCAAT-box in oomycetes. The set of genes with the TATA-like motif in their upstream regions is enriched for genes encoding RXLRs and ELIs and otherwise do not show any significant enrichment for Gene Ontology categories.

Another novel and abundant conserved DNA motif that shows correlation with gene expression during the infection is motif-9. This inverted repeat motif occurs upstream of 1,284 genes and the set of genes is enriched for RXLR effectors (3 fold). These conserved DNA motifs (motif-1, motif-7 and motif-9) are highly abundant in *Phytophthora* genomes, are correlated with the infection-related gene expression levels and are enriched in specific functional categories. Moreover, one of the four positively correlated motifs is the Inr/FPR-element, further emphasizing the adaptation of basic cellular machinery towards pathogenicity (Supplementary Figure S4-1). Hence, the four DNA motifs are relevant candidates for *cis*-acting transcriptional regulatory DNA motifs in pathogenic oomycetes.



**Figure 4-5 Highly abundant *cis*-regulatory elements.**

Nucleotide conservation of the five (A-E) highly abundant *cis*-regulatory motifs is displayed as sequence logos. The location of the motif in relation to the TLS is indicated by bar charts (bin size 50 nt). The frequency of the motif per bin was weighted according to the underlying length distribution of the upstream regions.

### Highly Abundant Motifs in the Genomes of *Phytophthora* spp. are Candidate Binding Sites for Transcriptional Regulators

We expanded the number of candidate motifs by focusing on the ten most abundant motifs within the set of the 19 automatically derived conserved DNA motifs in the upstream regions of the three *Phytophthora* species. These ten motifs include the two common promoter elements (Inr/FPR and CCAAT-box) earlier described, three motifs whose presence is correlated with gene expression levels during infection (motif-1, motif-7 and motif-9) and five additional candidate motifs (Figure 4-5). Whereas the remaining nine motifs occur in less than 100 different upstream regions, these five motifs occur in high abundance in the upstream regions of *Phytophthora* spp., ranging from 12,034 for motif-2 down to 1,397 occurrences for motif-18.

The most abundant of the five motifs is motif-2 which occurs upstream of 12,034 genes. It is a highly conserved CTTCAAC nucleotide motif that shows localization preference at 260 nt upstream of the translation start site (Figure 4-5A). The set of genes with motif-2 in their upstream region is significantly enriched in proteins with acyl-CoA

dehydrogenase and transporter activity (Supplementary Table S4-3C). In total, 606 of the 12,034 genes encode proteins involved in transporter activity, including e.g. a MOP flippase (PITG\_00021) as well as a potential sugar transporter (PITG\_00917).

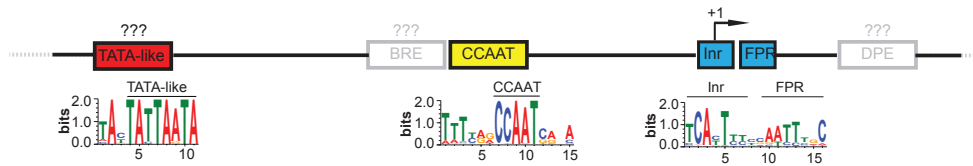
Motif-3, the second most abundant motif, is identified upstream of 5,249 genes, 1,767 of which belong to *P. infestans* (Figure 4-5B). Twelve of the 105 differentially expressed RXLR effector genes in *P. infestans* have this motif in their upstream region, including a member of the *Avrblb2* family (Table 4-1). In contrast to motif-2, the set of genes containing this motif in their promoter is enriched for genes encoding proteins with DNA binding functions (GO: 0003677), many of which are targeted to intracellular organelles (GO: 0043229) such as the nucleus. Of the total of 5,249 genes, 727 genes have either of the two functional annotations and 119 genes share both. These 119 genes include members of the Myb-like transcription factor family (e.g. PITG\_00513 that is significantly up-regulated during infection; see Table 4-1), genes encoding for transcription factors with basic leucine zipper domains (e.g. PITG\_00964), but also genes encoding chromatin remodelers such as histone deacetylases (e.g. PITG\_01897). However, the majority of these genes in *P. infestans* does not show differential expression during the infection process. The high abundance of motif-3 in the genomes of *Phytophthora* spp. and the highly significant enrichment of genes with predicted functions as transcriptional regulators highlight the prominent role of motif-3 as a protein-binding site. Hence, the transcription factor binding motif-3 is a central regulator and an important target for identification of the binding transcription factors and further experimental studies.

Moreover, we identified three additional highly abundant motifs, motif-17, motif-16 and motif-18 (Figure 4-5C-E, Supplementary Table S4-3C). These motifs are found upstream of 1,938, 1,724 and 1,397 genes, respectively. The set of genes containing motif-17 in their upstream region is enriched for genes encoding proteins involved in transferase-activity and amino-acid metabolism, whereas the sets containing motif-16 and motif-18 are enriched for functions such as ATPase activity or intracellular transport, respectively (Supplementary Table S4-3C). Even though the presence of these five motifs is not correlated with the expression levels at 2-4 dpi, they are interesting candidates because of their high abundance and the functional enrichment within the set of genes that have these motifs in their promoter.

## DISCUSSION

Infection of host plants by an oomycete is a complex process that requires the precise expression of proteins encoded in the pathogen's genome. Infection-related proteins directly or indirectly facilitate the tight interaction with the host by suppressing immune responses triggered by the pathogen. How the expression of this complex arsenal of genes, but also other genes encoded in their genomes, is precisely regulated is still largely unknown. To identify *cis*-regulatory motifs that characterize the promoter regions in *Phytophthora* genes, we adopted methods that utilize genome and transcrip-





**Figure 4-6 Architecture of *Phytophthora* promoter.**

Structure of the *Phytophthora* promoter defined by identified DNA motifs determined by our analysis or proposed to be present by indirect evidence. Identified consensus motifs are displayed below for the CCAAT-, the AT-rich TATA-like- and the Inr/FPR-element. Gene names of the orthologs of the TFs described to bind BRE and DPE are in Supplementary Table S4-1B.

to me data to systematically predict conserved DNA elements in genes co-expressed during infection. Our approach yielded 19 potentially active *cis*-regulatory elements, a number that is comparable to similar studies conducted in plants that yielded 34 potential DNA binding sites (Vandepoele et al. 2006). Only very few of the 19 motifs show significant similarity to known *cis*-regulatory motifs (Supplementary Table S4-2). The discovery of novel *cis*-regulatory elements is an important first step towards understanding the regulation of gene expression in oomycetes.

We identified a complex promoter structure that expands our view of the central transcriptional regulation machinery of oomycetes (Figure 4-6). Next to the Inr/FPR-element, we identified and quantified other known eukaryotic elements such as the highly abundant CCAAT-box (Figure 4-2). We identified an AT-rich motif (motif-7) that could represent a functional TATA-like element. Whether the observed AT-rich element is in fact functionally equivalent to the TATA-box element observed in other eukaryotes is unknown. We identified an ortholog of the TATA-box binding protein (TBP) that is encoded in the genomes of all analyzed *Phytophthora* spp. (Supplementary Table S4-1B). The presence of the TBP suggests that AT-rich promoter elements can be bound by the TBP, thereby recruiting the RNA Pol II to the TSS to initiate transcription, especially since it has been shown that TBP binds to a huge variety of AT-rich sequences. Hence, the AT-rich TATA-like box together with the TBP suggests that oomycetes contain functional TATA-box-like elements similar to that of other eukaryotes.

Our analysis did not identify motifs with similarity to the eukaryotic BRE-element or downstream promoter element (DPE), two motifs that are frequently observed as core elements in eukaryotic promoters. However, we found orthologs of the transcription factor II B (TFIIB) as well as the binding factors for the DPE-elements in all *Phytophthora* spp. and in the other oomycetes (Supplementary Table S4-1B). Hence, the presence of the necessary molecular factors encoded in the genomes of oomycetes is an indication for the presence of these elements or of non-canonical, functional replacements in the promoter of oomycete genes.

Only 37% of all genes have an Inr/FPR-element; a percentage that is lower than reported for the eukaryotic Inr-element present in various eukaryotes such as human and yeast (46% and 40%, respectively) (Yang et al. 2007). It is possible that our pipeline

did not automatically predict core promoter elements (e.g. BRE/DPE) or that we underestimated the overall abundance of other motifs, since we searched for motifs within the upstream regions of genes co-expressed under a distinct biological condition. The thereby derived motifs might be biased towards certain nucleotide conservation at positions that do not necessary reflect the consensus. Hence, in combination with a stringent significance cutoff, the biased motifs would not be able to identify all occurrences in the genome and consequently underestimate the true abundance. This is indeed the case for the upstream region of *ipiO1* gene in which the Inr/FPR-element was initially described (Pieterse et al. 1994). If we specifically searched for the occurrence of this motif in the upstream region, we could identify its occurrence at 28 nt upstream of the TLS. However, on a genome-wide search, the occurrence is not significant due to multiple-testing corrections. For a more elaborate unbiased quantification of the core promoter and also other DNA motifs, the identification of these elements under biological conditions other than the infection process is necessary. Currently, the number of different microarray experiments that monitor the changes in gene expression genome-wide is limited. Additional experiments probing different biological conditions would help to reduce the number of false negatives as well as false positives and provide a concise set of differentially expressed genes that could be used for the identification of stage-specific regulatory elements.

Interestingly, the set of genes that are regulated by different combinations of common eukaryotic promoter elements is enriched for distinct functional classes ranging from metabolism to effector genes (Figure 4-3). This functional adaptation of common eukaryotic promoter elements has been observed for yeast: TATA-box containing genes are stress-induced and expressed in extremely high or low levels, linking the TATA-box to transcriptional plasticity (Basehoar et al. 2004). Moreover, in plants and humans the CCAAT-box has been reported upstream of genes involved in development, gene expression, translation and general metabolism (Vandepoele et al. 2006; Dolfini et al. 2009; Jiao et al. 2009), corroborating our observed enrichments in *Phytophthora* (Figure 4-3). Many of the studied gene families, but especially RXLR as well as Crinklers effectors, underwent recent expansions in *Phytophthora* (Jiang et al. 2006; Haas et al. 2009; Schornack et al. 2010; Seidl et al. 2012). Identical upstream regions due to very recent duplications could influence the observed opposing enrichment of these classes in the set of genes containing either the Inr/FPR or the CCAAT-box. To test this hypothesis, we removed upstream regions that exceed similarity that could be expected due to functional DNA elements before assessing the enrichment (95% identity over 50 percent of the sequence). Even though quantitatively the results vary slightly, we overall still observed the opposing patterns of enrichment in GO categories and RXLRs as before, indicating the independence of our observation from bias due to very recent duplications.

We identified 17 additional conserved DNA motifs next to the two common eukaryotic promoter elements (Data S1). Several of these motifs are candidates for functionally active *cis*-regulatory elements because: (i) they are highly abundant in the *Phytophthora* species analyzed, (ii) their presence in the promoter of genes significantly correlates with the gene expression level during infection and (iii) the set of corre-

sponding proteins is enriched for interesting functions. Within the four motifs whose presence significantly correlates with up-regulation during infection, we revealed, next to the Inr/FPR and the putative AT-rich TATA-like element, two novel abundant motifs. This number is slightly lower, most likely due to limitations in the gene expression data, but still comparable to a study in *Plasmodium falciparum* that identified twelve motifs which are significantly correlated with gene expression levels (van Noort & Huynen 2006). Notably, motifs that are positively correlated occur in a high number of different upstream regions and are inverted repeats, suggesting binding by a homodimeric transcription factor.

Motif-1 is highly abundant and correlates with up-regulation of gene expression levels during infection. The set of genes containing this motif in their upstream region is enriched for RXLR effector genes as well as genes with catalytic activity such as glycosyl-hydrolases. In *Caenorhabditis elegans*, taCATGta motifs are rare footprints of Tc1 transposable elements excision (Eide & Anderson 1988). Given the strong conservation of this motif and the high abundance in the analyzed of *Phytophthora* genomes, we do not expect motif-1 to be a transposon footprint. Moreover, a recent analysis of the binding preference of homeodomain DNA-binding domains has identified TACATGTA as the preferred binding site for Irx family transcription factors (Berger et al. 2008); a group of transcription factors that is observed in *Drosophila* as well as in vertebrates and containing the Homeobox KN domain (PF05920). Interestingly, this domain is also present in several predicted transcription factors in the analyzed *Phytophthora* (Supplementary Table S4-1A); hence, these might be interesting candidate transcription factors for motif-1 binding.

Like motif-1, motif-3 is highly abundant. Notably, it is also present upstream of ~700 genes that encode proteins with predicted organelle localization such as the nucleus or DNA binding activity. Of these, 119 have both predicted functional annotations and include several transcription factors of the Myb-like family. The majority of *P. infestans* genes in this set is not differentially expressed during infection. Nevertheless, the high abundance of this motif in the *Phytophthora* genomes and its enrichment in genes encoding nuclear and DNA binding proteins suggests that motif-3 is a functional binding site for an unknown transcription factor that in turn regulates many other transcription factors.

The identification of potential biologically relevant motifs solely by correlating their presence with the gene expression levels is a simplified approach. This is especially apparent in the high variability of expression levels between genes that have one of the correlated motifs in their upstream region (Figure 4-4). *In vivo* there are many factors that influence the transcription of genes such as the chromatin state, the availability of the binding transcription factors and also the presence of other motifs in the proximity that may act together or antagonistic in a regulatory module. Nevertheless, the combination of different criteria, including significant correlation of motif presence with the gene expression level, allows us to generate a concise list of interesting candidates for pending experimental validation; both of the motif itself as well as of the binding

transcription factor.

This analysis provides the first systematic insights in the transcriptional regulation of the late blight pathogen *P. infestans* and two closely related *Phytophthora* species. The identified *cis*-regulatory elements are promising candidates for further experimental validation and identification of the binding transcription factors. In general, biochemical and genetic approaches such as ChIP-Seq are lagging in oomycetes and pathogenic fungi. However, whole genome transcriptomics and thereby derived gene expression data as well as genomic sequences of close relatives will be available in the close future. *In silico* methods such as the one outlined in this study are in an exceptional position to take advantage of these data to gradually close the knowledge gap between well-established model organisms and these important and intriguing groups of pathogens.

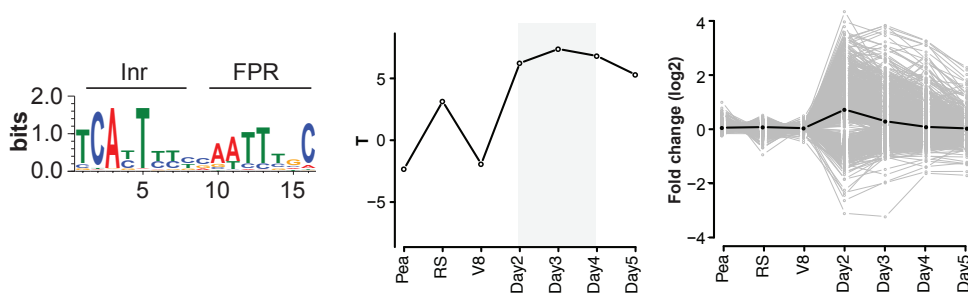
## ACKNOWLEDGMENTS

We would like to thank Lidija Berke, Adrian Schneider and Like Fokkens for fruitful discussions and comments on the manuscript. This work was financed by the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative / Netherlands Organisation for Scientific Research.

## SUPPLEMENTARY MATERIAL

Due to the amount of data, some of the Supplementary additional files, Supplementary tables S4-1 to S4-3 and Supplementary data S4-1, are only accessible online at *PLoS ONE* (<http://dx.plos.org/10.1371/journal.pone.0051295>).

### Supplementary Figures



**Figure S4-1 Positively correlated core promoter motif with gene expression levels.**

The Inr/FPR-element is positively correlated with the gene expression levels during infection. The motif logo, a graph displaying the T-value per data point and the gene expression of all differentially expressed genes that contributed to the correlation are displayed.

## REFERENCES

- Ah-Fong AMV, Xiang Q, Judelson HS. 2007. Architecture of the Sporulation-Specific Cdc14 Promoter from the Oomycete *Phytophthora infestans*. *Eukaryotic Cell*. 6:2222–2230.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.
- Bailey TL, Elkan C. 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol*. 2:28–36.
- Barrett T, Edgar R. 2006. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Meth. Enzymol.* 411:352–369.
- Basehoar AD, Zanton SJ, Pugh BF. 2004. Identification and distinct regulation of yeast TATA box-containing genes. *Cell*. 116:699–709.
- Berger MF et al. 2008. Variation in homeodomain DNA binding revealed by high-resolution analysis of sequence preferences. *Cell*. 133:1266–1276.
- Bussemaker HJ, Li H, Siggia ED. 2001. Regulatory element detection using correlation with expression. *Nat. Genet.* 27:167–171.
- Chalkley G, CP V. 1999. DNA binding site selection by RNA polymerase II TAFs: a TAFII250-TAFII150 complex recognizes the Initiator. *The EMBO Journal*. 18:4835–4845.
- Chaturvedi V, Flynn T, Niehaus WG, Wong B. 1996a. Stress tolerance and pathogenic potential of a mannitol mutant of *Cryptococcus neoformans*. *Microbiology*. 142:937–943.
- Chaturvedi V, Wong B, Newman SL. 1996b. Oxidative killing of *Cryptococcus neoformans* by human neutrophils. Evidence that fungal mannitol protects by scavenging reactive oxygen intermediates. *J. Immunol.* 156:3836–3840.
- Conesa A et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics*. 21:3674–3676.
- Crooks GE, Hon G, Chandonia J-M, Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome Res.* 14:1188–1190.
- Dolfini D, Zambelli F, Pavesi G, Mantovani R. 2009. A perspective of promoter architecture from the CCAAT box. *Cell Cycle*. 8:4127–4137.
- Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics*. 14:755–763.
- Eide D, Anderson P. 1988. Insertion and excision of *Caenorhabditis elegans* transposable element Tc1. *Mol. Cell. Biol.* 8:737–746.
- Elemento O, Tavazoie S. 2005. Fast and systematic genome-wide discovery of conserved regulatory elements using a non-alignment based approach. *Genome Biol.* 6:R18.
- Enright AJ, Van Dongen S, Ouzounis CA. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res.* 30:1575–1584.
- Govers F, Gijzen M. 2006. *Phytophthora* genomics: the plant destroyers' genome decoded. *Mol. Plant Microbe Interact.* 19:1295–1301.
- Grant CE, Bailey TL, Noble WS. 2011. FIMO: scanning for occurrences of a given motif. *Bioinformatics*. 27:1017–1018.
- Grunwald NJ, Garbelotto M, Goss EM, Heungens K, Prospero S. 2012. Emergence of the sudden oak death pathogen *Phytophthora ramorum*. *Trends Microbiol.* 20:131–138.
- Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS. 2007. Quantifying similarity between motifs. *Genome Biol.* 8:R24.
- Haas BJ et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora*

- infestans*. *Nature*. 461:393–398.
- Hahn S, Young ET. 2011. Transcriptional regulation in *Saccharomyces cerevisiae*: transcription factor regulation and function, mechanisms of initiation, and roles of activators and coactivators. *Genetics*. 189:705–736.
- Hoskins RA et al. 2011. Genome-wide analysis of promoter architecture in *Drosophila melanogaster*. *Genome Res*. 21:182–192.
- Javahery R, Khachi A, Lo K, Zenzie-Gregory B, Smale ST. 1994. DNA sequence requirements for transcriptional initiator activity in mammalian cells. *Mol. Cell. Biol.* 14:116–127.
- Jiang RHY, Tripathy S, Govers F, Tyler BM. 2008. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. *Proc. Natl. Acad. Sci. U.S.A.* 105:4874–4879.
- Jiang RHY, Tyler BM, Whisson SC, Hardham AR, Govers F. 2006. Ancient origin of elicitor gene clusters in *Phytophthora* genomes. *Mol. Biol. Evol.* 23:338–351.
- Jiao Y et al. 2009. A transcriptome atlas of rice cell types uncovers cellular, functional and developmental hierarchies. *Nat. Genet.* 41:258–263.
- Judelson HS. 2012. Dynamics and innovations within oomycete genomes: insights into biology, pathology, and evolution. *Eukaryotic Cell*. 11(11):1304–1312.
- Judelson HS. 2007. Genomics of the plant pathogenic oomycete *Phytophthora*: insights into biology and evolution. *Adv. Genet.* 57:97–141.
- Judelson HS, Michelmore RW. 1989. Structure and expression of a gene encoding heat-shock protein Hsp70 from the Oomycete fungus *Bremia lactucae*. *Gene*. 79:207–217.
- Judelson HS, Tyler BM, Michelmore RW. 1992. Regulatory sequences for expressing genes in oomycete fungi. *Mol. Gen. Genet.* 234:138–146.
- Jurka J et al. 2005. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* 110:462–467.
- Kim I, Sinha S, De Crombrughe B, Maity S. 1996. Determination of functional domains in the C subunit of the CCAAT-binding factor (CBF) necessary for formation of a CBF-DNA complex: CBF-B interacts simultaneously with both the CBF-A and CBF-C subunits to form a heterotrimeric CBF molecule. *Mol. Cell. Biol.* 16:4003–4013.
- Kutach AK, Kadonaga JT. 2000. The downstream promoter element DPE appears to be as widely used as the TATA box in *Drosophila* core promoters. *Mol. Cell. Biol.* 20:4754–4764.
- Lewis D, Smith D. 1967. Sugar alcohols (polyols) in fungi and green plants. I. Distribution, physiology and metabolism. *New Phytologist*. 66:143–184.
- Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*. 13:2178–2189.
- Maere S, Heymans K, Kuiper M. 2005. BiNGO: a Cytoscape plugin to assess overrepresentation of gene ontology categories in biological networks. *Bioinformatics*. 21:3448–3449.
- Maity SN, De Crombrughe B. 1998. Role of the CCAAT-binding protein CBF/NF-Y in transcription. *Trends Biochem. Sci.* 23:174–178.
- Majewski J, Ott J. 2002. Distribution and characterization of regulatory elements in the human genome. *Genome Res*. 12:1827–1836.
- Mantovani R. 1998. A survey of 178 NF-Y binding CCAAT boxes. *Nucleic Acids Res.* 26:1135–1143.
- McLeod A, Smart CD, Fry WE. 2004. Core promoter structure in the oomycete *Phytophthora infestans*. *Eukaryotic Cell*. 3:91–99.
- Mittler R, Vanderauwera S, Gollery M, Van Breusegem F. 2004. Reactive oxygen gene network of plants.

- Trends Plant Sci. 9:490–498.
- Müller F, Demény MA, Tora L. 2007. New problems in RNA polymerase II transcription initiation: matching the diversity of core promoters with a variety of promoter recognition factors. *J. Biol. Chem.* 282:14685–14689.
- Pieterse CM et al. 1994. Structure and genomic organization of the *ipiB* and *ipiO* gene clusters of *Phytophthora infestans*. *Gene*. 138:67–77.
- Pritsker M, Liu Y-C, Beer MA, Tavazoie S. 2004. Whole-genome discovery of transcription factor binding sites by network-level conservation. *Genome Res.* 14:99–108.
- Purnell BA, Emanuel PA, Gilmour DS. 1994. TFIID sequence recognition of the initiator and sequences farther downstream in *Drosophila* class II genes. *Genes Dev.* 8:830–842.
- Rayko E, Maumus F, Maheswari U, Jabbari K, Bowler C. 2010. Transcription factor families inferred from genome sequences of photosynthetic stramenopiles. *New Phytol.* 188:52–66.
- Roeder RG. 1998. Role of general and gene-specific cofactors in the regulation of eukaryotic transcription. *Cold Spring Harb. Symp. Quant. Biol.* 63:201–218.
- Roth FP, Hughes JD, Estep PW, Church GM. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* 16:939–945.
- Schorneck S et al. 2010. Ancient class of translocated oomycete effectors targets the host nucleus. *Proc. Natl. Acad. Sci. U.S.A.* 107:17421–17426.
- Seidl MF, Van den Ackerveken G, Govers F, Snel B. 2012. Reconstruction of oomycete genome evolution identifies differences in evolutionary trajectories leading to present-day large gene families. *Genome Biol Evol.* 4:199–211.
- Singer VL, Wobbe CR, Struhl K. 1990. A wide variety of DNA sequences can functionally replace a yeast TATA element for transcriptional activation. *Genes Dev.* 4:636–645.
- Smoot ME, Ono K, Ruscheinski J, Wang P-L, Ideker T. 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics.* 27:431–432.
- Stassen JH, Van den Ackerveken G. 2011. How do oomycete effectors interfere with plant life? *Curr Opin Plant Biol.* 14:1–8.
- Supek F, Bošnjak M, Škunca N, Šmuc T. 2011. REVIGO summarizes and visualizes long lists of gene ontology terms. *PLoS ONE.* 6:e21800.
- Škalamera D, Hardham AR. 2006. PnCcp, a *Phytophthora nicotianae* protein containing a single complement control protein module, is sorted into large peripheral vesicles in zoospores. *Austral. Plant Pathol.* 35:593.
- Tani S, Judelson H. 2006. Activation of zoosporogenesis-specific genes in *Phytophthora infestans* involves a 7-nucleotide promoter motif and cold-induced membrane rigidity. *Eukaryotic Cell.* 5:745–752.
- Tyler BM. 2007. *Phytophthora sojae*: root rot pathogen of soybean and model oomycete. *Mol Plant Pathol.* 8:1–8.
- Tyler BM et al. 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science.* 313:1261–1266.
- Van Dongen S. 2000. A cluster algorithm for graphs. Report INS-R0010, National Research Institute for Mathematics and Computer Science in the Netherlands, Amsterdam.
- van Noort V, Huynen MA. 2006. Combinatorial gene regulation in *Plasmodium falciparum*. *Trends Genet.* 22:73–78.
- Vandepoele K, Casneuf T, Van de Peer Y. 2006. Identification of novel regulatory modules in dicotyledonous plants using expression data and comparative genomics. *Genome Biol.* 7:R103.
- Verrijzer CP, Tjian R. 1996. TAFs mediate transcriptional activation and promoter selectivity. *Trends Biochem.*

- Sci. 21:338–342.
- Voegele RT et al. 2005. Possible roles for mannitol and mannitol dehydrogenase in the biotrophic plant pathogen *Uromyces fabae*. *Plant Physiol.* 137:190–198.
- Whisson SC et al. 2007. A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature.* 450:115–118.
- Wilson D, Charoensawan V, Kummerfeld SK, Teichmann SA. 2008. DBD-taxonomically broad transcription factor predictions: new content and functionality. *Nucleic Acids Res.* 36:D88–92.
- Woychik NA, Hampsey M. 2002. The RNA polymerase II machinery: structure illuminates function. *Cell.* 108:453–463.
- Xiang Q, Kim KS, Roy S, Judelson HS. 2009. A motif within a complex promoter from the oomycete *Phytophthora infestans* determines transcription during an intermediate stage of sporulation. *Fungal Genet. Biol.* 46:400–409.
- Yang C, Bolotin E, Jiang T, Sladek FM, Martinez E. 2007. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene.* 389:52–65.





# 5

# **A Predicted Functional Gene Network for the Plant Pathogen *Phytophthora infestans* as a Framework for Genomic Biology**

Michael F Seidl<sup>1,2</sup>, Adrian Schneider<sup>1</sup>,  
Francine Govers<sup>2,3</sup>, and Berend Snel<sup>1,2</sup>

<sup>1</sup>Theoretical Biology and Bioinformatics, Department of Biology, Utrecht University, Padualaan 8, 3584 CH Utrecht, The Netherlands

<sup>2</sup>Centre for BioSystems Genomics, P.O. Box 98, 6700 AB Wageningen, The Netherlands

<sup>3</sup>Laboratory of Phytopathology, Wageningen University, P.O. Box 8123, 6700 EE Wageningen, The Netherlands

*Submitted*



## ABSTRACT

Associations between proteins are essential to understand cell biology. While this complex interplay between proteins has been studied in model organisms, it has not yet been described for the oomycete late blight pathogen *Phytophthora infestans*.

We present an integrative probabilistic functional gene network that provides associations for 33 percent of the predicted *P. infestans* proteome. Our method unifies available genomic, transcriptomic and comparative genomic data into a single comprehensive network using a Bayesian approach. Enrichment of proteins residing in the same or related subcellular localization validates the biological coherence of our predictions. The network serves as a framework to query existing genomic data using network-based methods, which thus far was not possible in *Phytophthora*. We used the network to study the set of interacting proteins that are encoded by genes co-expressed during sporulation. This identified potential novel roles for proteins in spore formation through their links to proteins known to be involved in this process such as the phosphatase Cdc14.

The functional association network represents a novel genome-wide data source for *P. infestans* that also acts as a framework to interrogate other system-wide data. In both capacities it will improve our understanding of the complex biology of *P. infestans* and related oomycete pathogens.

## INTRODUCTION

The late blight pathogen *Phytophthora infestans* is one of the most destructive pathogens of tomato and potato, and a continuous threat to global food production (Haas et al. 2009). *P. infestans* belongs to the lineage of oomycetes which assembles diverse saprophytic and pathogenic species that share morphological similarities to true fungi (Latijnhouwers et al. 2003), yet are closely related to non-pathogenic diatoms and brown algae. Over the last two decades, *P. infestans* gradually developed into a model organism not only for oomycetes, but also for filamentous plant pathogens. The releases of its genome sequence and that of other closely related oomycetes (Tyler et al. 2006; Haas et al. 2009) have greatly increased our understanding of their complex biology, pathology and evolution (e.g. Seidl et al. 2012; Judelson 2012). So far, however, only individual gene products, mostly in the context of pathogenicity, have been intensively studied (Stassen & Van den Ackerveken 2011). Genome-wide experiments elucidating functional associations among proteins have not yet been performed and as a result, the complex interplay of proteins within a cell and its contribution to fundamental cellular processes is poorly understood.

Even though some proteins operate solitarily, the majority is associated with other proteins. They are embedded in a complex network in which assemblies of proteins synergistically mediate a biological function (Gavin et al. 2002; Krause et al. 2004). Proteins can associate directly by physical interaction, e.g. in protein complexes, or indirectly, e.g. in the same pathway or cellular process. Functional association networks represent the compendium of all possible associations in a cell. *In vivo*, however, these associations are dynamic and depend on physiological conditions such as external stimuli or changes during the life cycle.

A considerable number of functional association networks in many species have been described. These networks are not only derived from large-scale experimentally determined physical associations (Gavin et al. 2002; 2006) but also from integrative approaches combining diverse functional and comparative genomics data. Such integrative networks made a substantial contribution in system-wide understanding of the biology of well-studied model organisms such as *Saccharomyces cerevisiae* (budding yeast) and *Arabidopsis thaliana* (thale cress) (Jansen et al. 2003; Lee et al. 2004; 2010). Many of these studies used a Bayesian framework to integrate heterogeneous data into a single unified network (Jansen et al. 2003; Lee et al. 2004): every data source adds a certain level of evidence to the combined evidence of functional linkage between two proteins. At the same time, this approach accounts for differences in the quality of the individual data sources. The resulting network maximizes the coverage of the proteome while ensuring an acceptable level of confidence (Lee et al. 2004). The reliability of these integrative approaches has been benchmarked using experimental data that are available in these model organisms. While very few protein-protein interactions or functional associations have been reported in *P. infestans* (Blanco & Judelson 2005), a considerable amount of transcriptomic and comparative genomic data for *P. infestans* and other oomycetes is available (Randall et al. 2005; Tyler et al. 2006; Judelson et al.

2008; Haas et al. 2009).

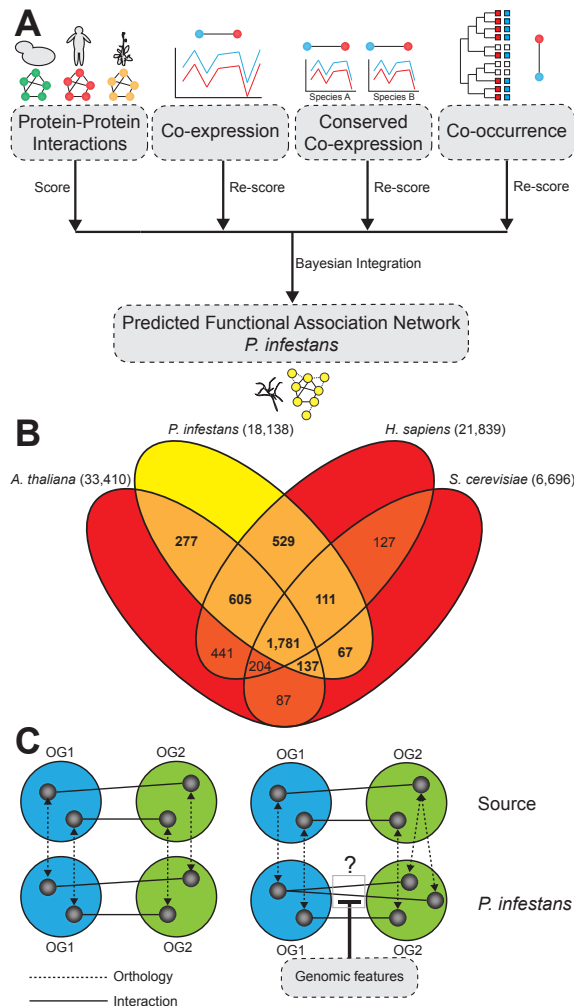
In this study, we present the first functional association network in the oomycete model organism *P. infestans*. Our method integrates diverse functional and comparative genomics data sets into a unified network. The first data set is composed of projected interactions based on interolog mapping, the transfer of protein-protein interactions from one organism to another: proteins in the species of interest interact if their orthologs in another species have been shown to interact (Walhout et al. 2000). The second data set adds predicted associations between proteins encoded by co-expressed genes (Hughes et al. 2000; Gollub et al. 2003). Thirdly, we used conserved co-expression, i.e. orthologs of co-expressed genes in one species are also co-expressed in a related species, to increase the moderate predictive power of gene co-expression towards functional association (van Noort et al. 2003). As a fourth line of evidence we predicted interacting proteins by conserved phylogenetic co-occurrence of the two encoding genes across a considerable amount of divergent species (Pellegrini et al. 1999). This approach assumes that interaction partners should either be gained or lost together, as a single interaction partner cannot perform the full function. We adapted a scoring schema that assesses the merit of each individual data set and subsequently integrates the data using a Bayesian approach yielding a comprehensive functional association network, covering over 30% of the predicted proteome of *P. infestans*. Our predicted network enables the in-depth analysis of complex omics data such as microarrays. For example, in the predicted functional association network we identified functional modules of differentially expressed genes during distinct life phases of *P. infestans*, highlighting dynamic features of this network. These functional modules place unknown gene products in a cellular context. The functional association network represents a valuable addition to the growing genomic resources for *P. infestans* serving as an important framework for in-depth analyses of existing and yet to appear omics data. We anticipate that its availability will add significant knowledge to our understanding of the complex biology of this devastating plant pathogen.

## RESULTS & DISCUSSION

### Adaptation of a Bayesian Scoring Schema in *P. infestans*

To integrate four complementary large-scale transcriptomic and comparative genomic data sets of gene-to-gene (protein-to-protein) associations we adopted a unified scoring schema (Figure 5-1A) that has been applied successfully in other eukaryotes (Lee et al. 2004; 2010). This scoring schema is derived from Bayesian statistics and describes the (log) likelihood score (LLS) of association under given evidence and is corrected for the background expectation of association. Therefore, the LLS is proportional to the confidence of the given experiment to successfully recall known associations (Lee et al. 2004); a LLS of 0 corresponds to random association. More importantly, this unified scoring schema allows accounting for the variability in the predictive quality

between both binary data, such as predicted protein-protein interactions, as well as continuous data with an intrinsic scoring schema, such as the similarity between gene expression profiles. The continuous data is transformed into a range of LLS scores for different values of the intrinsic score (Material and Methods). Due to the lack of experimentally defined protein associations (true positives) and consequently also true negatives in *P. infestans*, we approximated such a set using KEGG maps and Gene Ontology (biological process). Based on these approximations, we derived the prior odds (Supplementary Table S5-1A), i.e. the ratio of probability of functional association and its negation without evidence, and the posterior odds, i.e. the ratio of probability of functional association and its negation given the evidence, for each dataset (Supplementary Table S5-1B) and subsequently determined the LLS.



**Figure 5-1 Prediction of functional association network in *P. infestans*.**

(A) Integration of four distinct data sources to predict the functional association network in *P. infestans*. We predicted protein-protein interactions by projection of interactions from three source organisms (yeast, human, thale cress) to *P. infestans*; co-expression, conservation of co-expression between *P. infestans* and *P. sojae* and phylogenetic co-occurrence in 51 species. Before the integration of the four data sources into a single network, these were scored based on their relative confidence using KEGG maps. (B) Number of orthologous groups between *P. infestans* and the three source organisms used for projecting physical interactions. (C) Projection of physical interactions via orthologous groups. In cases where the mapping was unclear, different genomic features such as co-expression and shared functional annotation were considered to disentangle these specific cases.

### Protein-protein Interactions from Three Model Organisms are Projected to *P. infestans*

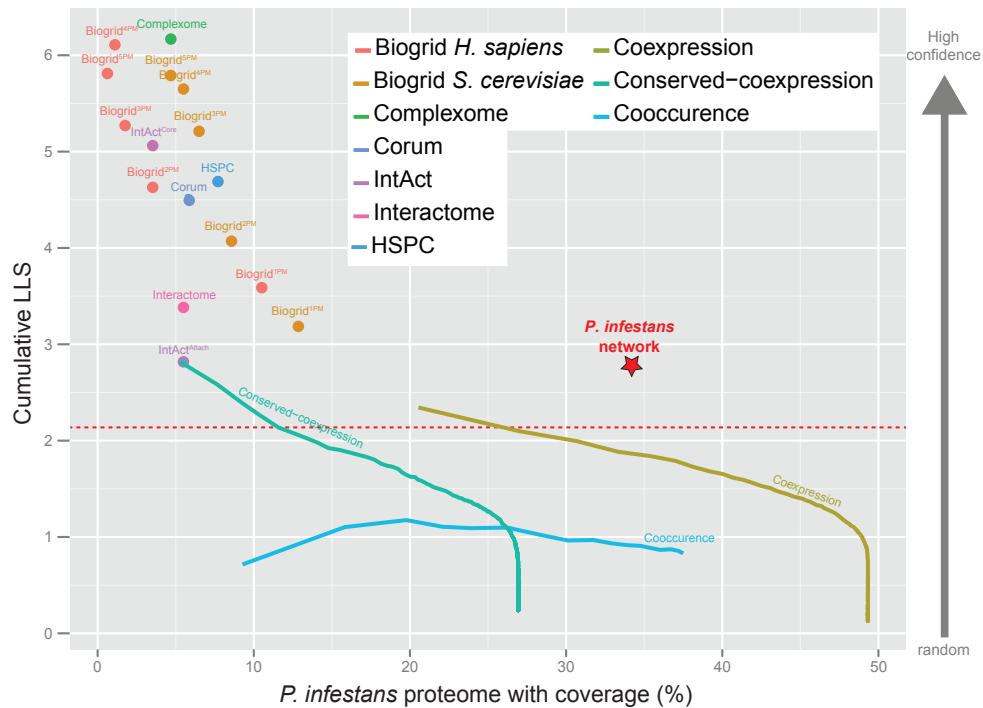
We projected a substantial number of physical interactions between protein pairs based on interolog mapping. To this end, we identified orthologs using an orthology detection algorithm (similar to Orthologous Matrix OMA (Altenhoff et al. 2011)) that we applied to a selection of 51 diverse eukaryotic species. We identified 3,507 orthologous groups (orthologous pairs + inparalogs) between *P. infestans* and at least one of the three genomes (*Homo sapiens* (human), *S. cerevisiae* and *A. thaliana*), of which 1,781 orthologous groups are shared between all four genomes (Figure 5-1B). Using the 3,507 orthologous groups, we projected protein-protein interactions from six different databases that aggregate information from *H. sapiens*, *S. cerevisiae* and *A. thaliana* to *P. infestans* (Supplementary Table S5-1C). The information available from BioGRID and IntAct enabled discrimination between different levels of confidence. Since these interactions are mapped using orthology, some of the orthologous groups also include inparalogs and in some cases it is not directly obvious to which of the possible pairs the functional interaction would be most reliably mapped (Figure 5-1C). These specific cases were disentangled using additional data considering overlapping and complementary functional characteristics, such as gene co-expression and cellular co-localization (Material and Methods).

All sixteen predicted protein-protein interaction networks, derived from the six different databases, have a LLS score that is higher than random linkage (LLS > 0), ranging from 2.8 (IntAct attachments) to 6.1 (Complexome), reflecting their high quality (Figure 5-2 & Supplementary Table S5-1B). The ranking of the LLS for the KEGG and GO benchmark yields similar results (Supplementary Table S5-1B), even though the two benchmark sets are fairly independent: They share only 4,470 pairs of true positives (12.6% of the true positives annotated in KEGG) (Supplementary Figure S5-1), indicating the robustness of the LLS approximation of the mapped datasets.

### Complementary Comparative Genomic and Gene Expression Data are Integrated to Predict Functional Associations in *P. infestans*

To also add complementary data to the mapped physical interactions from distantly related organisms, we used three other large-scale (comparative-) genomic data sets that could be indicative for the association between a pair of proteins (Figure 5-1A); (i) similarity in co-expression patterns, (ii) conservation of co-expression between co-expressed *P. infestans* genes and their orthologs in the soybean pathogen *Phytophthora sojae* and (iii) similarity in phylogenetic co-occurrence profiles measured in 51 eukaryotic species (Material and Methods).

These three genomic datasets score higher than random in our applied LLS scoring schema (Figure 5-2). As expected, their confidence is lower than the predicted protein interaction data, but the coverage of the proteome increases. Gene co-expression on its own has been shown to be a limited predictor of functional association; a correla-



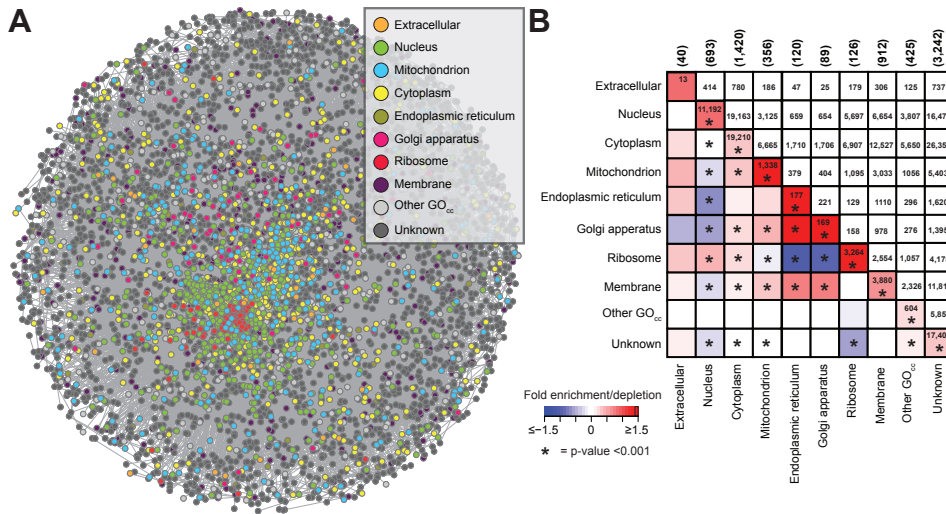
**Figure 5-2 Relationship between the log-likelihood score and the coverage of different data sources.**

The relationship between log-likelihood score and coverage (percentage of *P. infestans* proteome) of the different data sources is displayed. Projected physical interactions are shown by dots and for BioGRID (number of supporting pubmed entries (PM), e.g. 1PM or 3PM) and IntAct (core or attachment) further subdivided based on the reliability. Continuous data sources, e.g. co-expression, are indicated as a line. The dashed red line shows the applied cutoff to include associations between proteins in the predicted *P. infestans* network. The coverage and log-likelihood score for the *P. infestans* association network is denoted with a red star.

tion coefficient of 0.8 corresponds to a LLS of only 0.76. If orthologs of co-expressed genes in related species are also co-expressed, this conserved co-expression is a high quality proxy for functional association (van Noort et al. 2003); a score of 0.8, which is approximated by the average of both correlation coefficients (Material and Methods), corresponds to a LLS of 1.3. This higher quality of conserved co-expression as a proxy for functional association, in return for a smaller coverage, is an observation that is also visible in our scoring schema (Figure 5-2).

### The Prediction and Initial Survey of the Functional Association Network of *P. infestans*

To obtain a comprehensive picture of functional associations, we integrated the four above described large-scale gene association data sets using a naïve Bayesian approach: we additively derived a LLS describing the combined evidence for association among pairs of proteins (Material and Methods). Each individual data source, even



**Figure 5-3 The functional association network and the subcellular localization of its members.**

(A) The predicted functional association network in *P. infestans*. Nodes, representing proteins, are colored according to their subcellular localization approximated by Gene Ontology term. (B) Correlation of subcellular localization with predicted protein associations. The  $\log_2$ -fold enrichment/depletion of protein pairs where both partners are predicted to reside in the same/different subcellular localization compared to the expected numbers is displayed by the heat map (lower half of the symmetrical matrix) (values saturate at  $\pm 1.5$ ); the corresponding raw numbers are shown in upper half. Significant enrichment/depletion (after multiple testing correction) is indicated by asterisk. The total number of proteins predicted to reside in a particular subcellular localization is displayed in brackets above the plot.

though less reliable by itself adds evidence for the functional linkage of two proteins. Thereby, we unified these diverse lines of evidence into a comprehensive functional association network in *P. infestans* while simultaneously controlling quality (expressed by the associated LLS) and coverage of these predictions. We applied a LLS cutoff of 2.1 to each protein pair that corresponds to a very conservative Pearson correlation coefficient for co-expressed gene pairs of 0.97. This cutoff allows the inclusion of associations from genomic data sources if their score is above the LLS cutoff as well as the inclusion of lower scoring associations that require several independent lines of evidence to cumulatively pass the LLS cutoff.

The predicted network in *P. infestans* links 5,942 proteins (~33% of the predicted proteome), with 108,530 functional associations (Supplementary Table S5-2). With a pairwise LLS cutoff of 2.1, the total confidence of the combined network is 2.76 (Figure 5-2; Figure 5-3A). As expected given the applied cutoff, 56% of the functional associations are in part derived from protein interaction data in other species; consequently the *P. infestans* functional association network is mainly a physical interaction network. Moreover, 34,352 of these protein associations (~55%) have additional support based on other large-scale (comparative-) genomic data sets, giving further evidence for the robustness of the predictions. The network comprises 52 connected components (98%



of proteins reside in the largest component; Figure 5-3A). A characteristic path length of 3.4, which is smaller than e.g. the overall protein-protein interaction network of *S. cerevisiae* but similar to the subset of essential proteins (Said et al. 2004), and high clustering coefficient (0.28) are indicative of a dense network that reflects the homology-based projection of complexes and interactions.

Proteins that are part of the network show highly significant enrichment (all p-values  $< 1 \times 10^{-7}$ ) in central cellular processes such as gene expression (GO:0010467), translation (GO:0006412), cellular localization (GO:0051641) and cell cycle (GO:0007049). The majority (58%) of proteins in the network is at least partially projected based on physical interaction which favors evolutionary conserved processes and hence explains the enrichment in core cellular processes. Nevertheless this information is useful as it provides further insights into the wiring of these core processes in *P. infestans*.

The network also includes proteins with putative functions in pathogenicity or proteins that have been shown to induce defense responses in the host (Stassen & Van den Ackerveken 2011); many of which are predicted to be secreted upon infection (Supplementary Table S5-2). The network contains 284 secreted proteins; a 2.6-fold increase to the number we would have obtained if we only considered interactions derived by orthology projection of protein-protein interaction data. The RXLR- and Crinkler-effectors, two classes of host-targeted effectors that most likely promote infection of the host, are highly abundant in the proteome of *P. infestans* (596 RXLR- and 452 Crinkler-effectors) (Haas et al. 2009; Jiang et al. 2008), and also occur 18 and 3 times, respectively, in the predicted network. The associations of these proteins with others are solely based on (conserved-) co-expression data, indicating involvement in the same process, without any evidence for potential physical associations. Two notable classes of highly abundant enzymes that are potentially linked to pathogenicity are glycoside hydrolases and peptidases (Tyler et al. 2006; Haas et al. 2009; Seidl et al. 2011; 2012). We observed 46 glycoside hydrolases and 119 peptidases in the predicted network. This is a considerable increase of  $\sim 2$  fold compared to a network that would only be based on projected physical data.

### The Functional Association Network is Enriched for Co-localized Protein Pairs

Functionally associated proteins that show physical interaction are close together in the same subcellular compartment (Schwikowski et al. 2000; Gandhi et al. 2006). Subcellular localization therefore presents a suitable criterion to assess the biological significance of the predicted associations in *P. infestans* independently of the initial benchmark of (homology-based) KEGG pathways used to derive the LLS for each association. The network displays non-random distribution and local clustering of proteins with the same subcellular localizations, approximated by GO-cellular compartment (Figure 5-3A). To quantify this, we examined the enrichment/depletion in subcellular localization of associated proteins in the predicted *P. infestans* functional association network (Figure 5-3B, Material & Methods). Proteins with the same subcellular localization are significantly enriched amongst associated proteins, in agreement with observa-

tion on directly measured associations in other organisms and confirming the validity of our predicted network. Specifically, proteins residing in the endoplasmic reticulum, the Golgi apparatus and membranes are enriched for interactions, corroborating previous results in human (Gandhi et al. 2006). In accordance with the observations by Gandhi and colleagues (2006), proteins with predicted localization in the nucleus, the ribosome and to a smaller extent the mitochondrion do tend to not interact with proteins present in other sub-compartments.

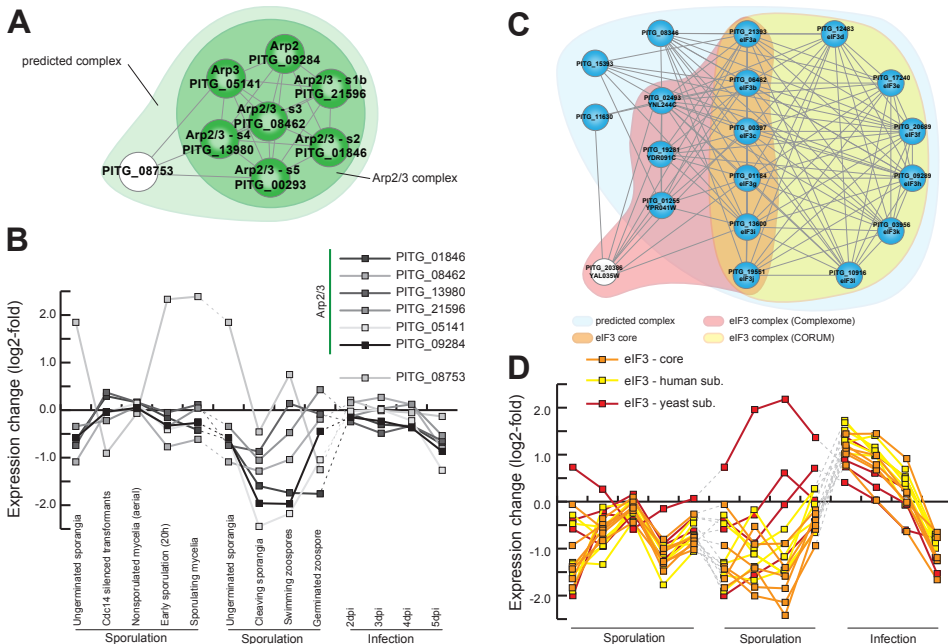
As homology is part of the initial source of the projected physical interactions – i.e. the LLS scoring via the KEGG benchmark as well as the prediction of subcellular localization via GO – we used two additional approaches to assess similarity in subcellular localization of predicted associations independent of homology. In the first approach, we divided the network into two components, one containing associations that are supported by at least one protein-protein interaction dataset (Supplementary Figure S5-2A), and the second, which is merely based on non-physical associations (co-expression, co-occurrence) (Supplementary Figure S5-2B). Both networks yield similar results in the (significant) enrichment of associated proteins predicted to co-localize. In a second independent approach, we used WoLF PSORT that predicts subcellular localization merely on sequence features and not homology (Horton et al. 2007). Again, we found similar patterns of enrichments in proteins with the same subcellular localization. Proteins residing in the nucleus and the mitochondrion showed depletions for associations with proteins predicted to reside elsewhere (Supplementary Figure S5-2C). These patterns are less pronounced, most likely because the prediction algorithm is not optimally trained for oomycete sorting signals. These independently derived similar patterns in enrichment and depletion support our predicted functional associations, even though experimentally verified associations, as present for other species, would provide a superior benchmark set to adjust confidence levels and assess the predicted associations.

### **The Compendium of Protein Complexes Embedded within the Functional Association Network**

One of the major steps in understanding the function of a cell is to identify and determine the composition of its protein complexes. We mined the subset of the predicted functional association network that is supported by at least a single protein-protein interaction to derive protein complexes. We applied the ClusterONE algorithm that detects overlapping protein complexes in weighted networks by searching for sub-graphs that are characterized by many reliable interactions between proteins and separation from the remaining network (Nepusz et al. 2012). In total, we detected 299 protein complexes covering 3,146 proteins (Supplementary Table S5-3).

Due to incomplete proteome annotation, members of a protein complex are unlikely to be identified by functional annotation (GO terms) alone. For example, the Arp2/3 complex, a central organizer of the actin filaments, contains seven subunits in yeast and human (Goley & Welch 2006). While its constitution can be completely retrieved in yeast and human based on its GO term (GO:0005885), the same is impossible in *P. infestans*:

the annotation of the encoding genes is limited (only a single member of the Arp2/3 complex has this term) and higher-level terms such as cytoskeleton are too broad and retrieve too many results. The functional association network is therefore a necessary framework to predict and study the composition of protein complexes in *P. infestans*. Indeed, the Arp2/3 complex is one of the complexes we detected (complex 18). Besides Arp2 and Arp3, which have already been described (Ketelaar et al. 2012), the detected complex contains the remaining five together with an additional subunit (Figure 5-4A). The genes encoding the seven subunits display a high degree of co-expression, whereas the additional protein, a tubulin-tyrosine ligase like protein (TTLL), is not co-expressed (Figure 5-4B), and therefore likely not part of the core Arp2/3 complex. In-depth investigation revealed that the associations to TTLL have been projected via a read-through transcript containing an Arp2/3 subunit and TTLL from human, underscoring the necessity to assess fusion transcripts in future analyses and to include gene expression data



**Figure 5-4 Predicted Arp2/3 and eIF3 complexes in *P. infestans*.**

(A) Automatically predicted Arp2/3 complexes (ClusterONE prediction in light green) include the seven conserved subunits of the eukaryotic Arp2/3 complex (Arp2, Arp3 and the five associated subunits; highlighted in dark green). (B) Gene expression of the predicted Arp2/3 complex (note: there is no gene expression data for Arp2/3 subunit 5 [BROAD:PITG\_00293]). The log<sub>2</sub>-fold change in expression at different time points/developmental stages (averaged replicates) of three different gene expression experiments compared to the gene expression in mycelium/hyphae growth of the respective experiment is displayed in the graph. (C) Automatically predicted eIF3 complex (ClusterONE prediction in light blue), the annotated eIF3 complexes based on CORUM (yellow) and Complexome (red) database as well as the conserved eIF3 core (six subunits; orange). (D) Gene expression of the eIF3 core complex (orange) and the additional subunits predicted by either CORUM (yellow) or Complexome (red). The log<sub>2</sub>-fold change in expression at different time points (averaged replicates; description as in (C)) of three different gene expression experiments compared to the gene expression in mycelium/hyphae growth of the respective experiment is displayed in the graph.

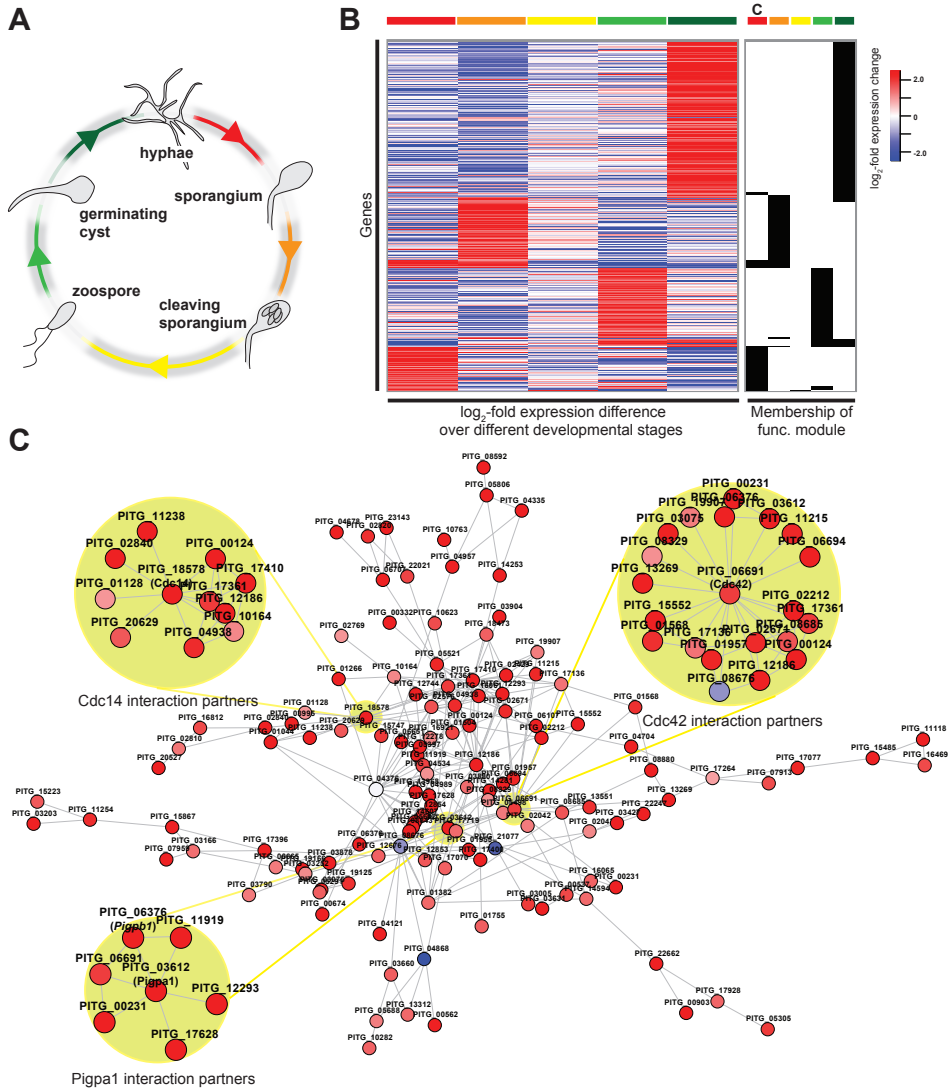
to validate and disentangle predicted protein complexes.

The analysis of another protein complex highlights the necessity of an integrative approach that combines different data sets from diverse organisms: The eukaryotic initiation factor 3 (eIF3) is among the largest translation initiation factors in eukaryotes (Hinnebusch 2006). Its conserved 'core' contains five essential (eIF3a, eIF3b, eIF3c, eIF3g and eIF3i) and one nonessential subunit (eIF3j) (Hinnebusch 2006). Only three of them could have been predicted in *P. infestans* based on GO terms. The eIF3 core is a subset of one of the detected complexes (complex 110; Figure 5-4C) that also contains several other subunits. The *H. sapiens* eIF3 complex described by the CORUM database contains six additional subunits, whereas the eIF3 complex described by the Complexome database contains four additional subunits, all of which have orthologs in *P. infestans*. Our predicted network unifies this information and consequently, the automatically inferred protein complex contains all these subunits, except a single protein from Complexome, and additional three, two of which are also eukaryotic translational initiation factors and hence likely functionally related. The genes encoding the eIF3 core proteins as well as the orthologs of the human complex are highly co-expressed and therefore likely forming a functional complex, whereas the orthologs of the yeast subunits, especially pronounced for *eIF5* [BROAD:PITG\_01255], show a lower level of co-expression (Figure 5-4D). The ATP-binding cassette protein RLI1 (yeast: [SGD:YDR091C]) is a conserved factor that has been implicated in several essential cellular processes such as translational initiation (Dong et al. 2004; Chen et al. 2006) and translational termination and recycling (Shoemaker & Green 2011). According to CORUM database, there is no interaction between human RLI (the ortholog to RLI1 in yeast) and eIF3 core factors, whereas the yeast complex in Complexome and consequently the predicted *P. infestans* network contain this experimentally determined interaction (Dong et al. 2004) (Figure 5-4C).

### Identification of Functional Modules during the Development of *P. infestans*

Microarray technologies are a valuable source for the identification of genes involved in development and pathogenesis in *P. infestans* (Judelson et al. 2008; 2009; Haas et al. 2009). The interpretation of the results is challenging since a direct biological role for differentially expressed genes is not necessarily apparent, especially for uncharacterized gene products. The predicted functional association network provides a convenient framework to enhance the biological interpretation of gene expression data by placing functionally characterized and uncharacterized gene products in their wider cellular context.

We aimed to apply the predicted network to identify functionally related subsets of differentially expressed genes at defined time points in the lifecycle of *P. infestans* (Figure 5-5A). To prevent circularity in the analysis, we excluded associations within the network that were only supported by gene expression data, leaving us with a network of 62,000 associations between 3,500 proteins. We used the algorithm HEINZ (Beisser et al. 2010) that automatically finds the subset of up-regulated genes that are also



**Figure 5-5** Developmental stages of the asexual lifecycle of *P. infestans* and determined functional modules.

(A) Transition between five distinct developmental stages in the asexual lifecycle of *P. infestans*. Gene expression data of these stages are available (Judelson et al. 2008) and were used to assess gene expression changes between the transitions. Transitions between different developmental stages are color coded. (B) The gene expression changes and the membership of all 826 genes predicted in the five functional modules are displayed. The heat map shows the gene expression changes ( $\log_2$ ) for the transitions of two subsequent life stages (same color code as in A; heat map saturated at  $\pm 2.5$ ). Presence (black) or absence (white) of genes in a functional module is highlighted next to the heat map and membership is indicated by color code. C refers to the sporulation module as described in (C). (C) Determined functional module of genes up-regulated during sporulation and their predicted associations. Examples discussed within the text and their directly associated proteins are highlighted with light green. The nodes are colored according to the fold change in expression in sporangia compared to hyphae (same scale as in B).

interconnected by a significant amount of associations. Such an ensemble of genes is referred to as a functional module, and allows identifying and studying proteins likely involved in a defined process and their associations. We studied five different time points during the asexual development of *P. infestans* (Judelson et al. 2008): *in vitro* growing nonsporulating hyphae, sporangia, cleaving sporangia, swimming zoospores and germinating zoospore cysts that contain specialized infection structures called appressoria (Figure 5-5A). We defined differentially expressed genes at each transition and subsequently detected functional modules (Figure 5-5A & Figure 5-5B).

Each developmental transition is represented by a functional module of associated differentially expressed genes. The modules display little overlap and a distinct pattern of gene expression changes (Figure 5-5B). They vary in size, ranging from nearly 400 members at the transition from germinated cyst to hyphal growth, to only two members at the transition from cleaving sporangia to swimming zoospores (Supplementary Table S5-4). The latter transition contains only very few up-regulated genes (FDR 0.05) in the predicted network which is the reason for the small size of the module. Interestingly, the functional module at the former transition is enriched for proteins with a predicted function in proteolysis (GO:0006508;  $p$ -value  $\ll 1 \times 10^{-4}$ ), including twelve, mostly intracellular, peptidases. In contrast, the transition from hyphae to sporangia is significantly ( $p$ -value  $< 1 \times 10^{-4}$ ) enriched for regulation of biological process (GO:0050789) and in particular signal transduction (GO:0007165). Among other proteins involved in regulation we also found ten kinases. Four of these are also found in the functional module of the subsequent transition from sporangium to cleaved sporangium, a module that contains 17 kinases. Kinases have been reported to be among the genes with the highest fold expression change in this transition (Judelson et al. 2008). Oomycetes contain an extensive repertoire of these central regulators (Judelson & Ah-Fong 2010; Seidl et al. 2011). The high abundance of kinases in functional modules points to their prominent role in regulation of sporangium formation. The associations of and amongst these kinases present important novel information that would not have been available using gene expression data alone.

### **The Sporangia Formation Module Contains Genes Encoding Known and Novel Proteins, and Novel Associations**

To highlight the merit of the functional association network as a framework to study gene expression and the predicted associations between co-regulated proteins, we further studied the initial phase of sporulation. In *Phytophthora* this major transition leads to the formation of sporangia, asexual spores that can either germinate directly and infect the host, or develop into a zoosporangium which cleaves and releases multiple zoospores that function as infectious propagules (Figure 5-5A). We identified a module that contains 130 interconnected proteins of which 126 are encoded by up-regulated genes during sporangium formation (Figure 5-5C). This functional module is significantly enriched ( $p$ -value  $< 0.05$ ) for proteins with predicted functions in signal transduction (GO:0007165), cell differentiation (GO:0030154) and developmental processes (GO:0032502). Interestingly, our predicted module contains most proteins known to be

involved in sporangia formation, but also many novel interactors that have not yet been associated with this important process.

In the predicted functional module we observed *Pigpa1* [BROAD:PITG\_03612] and *Pigpb1* [BROAD:PITG\_06376], the alpha and beta subunit of the heterotrimeric G-protein. The transcription of both genes is up-regulated early during spore formation (María Laxalt et al. 2002). Whereas *Pigpb1* silenced mutants have malformed sporangia and very few asexual spores (Latijnhouwers & Govers 2003), *Pigpa1* silenced mutants show altered zoospore mobility, reduction in zoospore release and appressorium formation (Latijnhouwers et al. 2004). *Pigpa1* is predicted to interact with a predicted up-regulated cAMP dependent kinase [BROAD:PITG\_12293] (Figure 5-5C), confirming that the cAMP pathway is indeed involved in processes regulated by *Pigpa/b1* during sporulation as initially suggested by Latijnhouwers and colleagues (2003). One of the predicted interaction partners of both *Pigpa1* and *Pigpb1* is *Rac1* [BROAD:PITG\_06691], a small GTPase of the Ras-like superfamily. Its role as a central regulator is corroborated by several predicted interaction partners: eukaryotic protein kinases such as mitogen-activated kinases [BROAD:PITG\_02212/BROAD:PITG\_17361/BROAD:PITG\_12186] or the phosphatidylinositol-4-phosphate-5-kinase [BROAD:PITG\_15552]. Next to *Rac1*, we observed other signal transduction components related to the Ras superfamily of GTPases such as ARF-like [BROAD:PITG\_08880/BROAD:PITG\_13269] and *Rab* [BROAD:PITG\_19907/BROAD:PITG\_17136], highlighting the importance of these signaling proteins and the associations of these novel candidates for spore formation.

In the functional module we also observed the phosphatase *Cdc14* [BROAD:PITG\_18578]. In eukaryotes, it plays a role in a variety of processes including cell cycle regulation and termination of mitosis. In contrast to its orthologs, *Cdc14* in *P. infestans* is specifically expressed during sporulation, and has a central role in spore formation (Ah-Fong & Judelson 2003). It also seems not to be involved in the regulation of mitosis during normal growth, even though it complements the function of *Cdc14* in yeast (Ah-Fong & Judelson 2003) and therefore might still maintain this regulatory role during sporulation (Ah-Fong & Judelson 2011). Additionally, recent evidence points to a possible role of *P. infestans* *Cdc14* in the development of the flagellum due to its co-localization with the known basal body marker *DIP13* (deflagellation-inducible protein; [BROAD:PITG\_13461]) (Ah-Fong & Judelson 2011). Even though *Cdc14* and *DIP13* show considerable (conserved) co-expression (Pearson correlation coefficient 0.76), this evidence is insufficient to infer association within the framework of our network. Interestingly, *Cdc14* is predicted to interact with a 4.3-fold ( $\log_2$ ) up-regulated tyrosine kinase [BROAD:PITG\_17410] (Figure 5-5C). We observed an association between this kinase and *DIP13* (Supplementary Table S5-2), therefore indirectly linking *DIP13* to *Cdc14* as initially suggested by the co-localization studies by Ah-Fong and colleagues (Ah-Fong & Judelson 2011).

The up-regulated *Cdc14* interaction partners within the functional module include several other kinases such as the 2.5-fold ( $\log_2$ ) up-regulated Ser/Thr kinase [BROAD:PITG\_00124]. Interestingly, we predicted a novel association between *Cdc14*

and NIFC1 [BROAD:PITG\_11238], a protein that contains a nuclear LIM interactor-interacting factors domain and is reported to be involved in transcriptional regulation (Judelson & Tani 2007). *NIFC1* is highly expressed during zoospore-formation (cleavage) (Judelson & Tani 2007; Judelson et al. 2008), whereas *Cdc14* is expressed early during sporangium formation and maintains a high expression level during zoospore-formation. Together with the predicted association between *Cdc14* and the sir2-like histone deacetylase [BROAD:PITG\_10164], these interactions imply a role of *Cdc14* as a transcription regulator to reprogram gene expression during zoospore formation.

We highlighted how the predicted functional association network serves as a valuable framework for the analysis of gene expression data. The delineation of functional modules generates a concise set of candidates and their associations for further studies. The sporangia formation module illustrates this nicely: firstly we identified proteins that have been already experimentally linked to this transition, e.g. *Cdc14* and *Pigpb1*. Subsequently, we were able to place these in their wider cellular context allowing the identification of directly associated proteins. Since many of these have only putative functions (~50%) or are without functional annotation (18%) the functional network approach used in this study revealed interesting novel candidates that may play central roles in sporangia formation.

## CONCLUSIONS

Proteins rarely act alone. They interact either directly or indirectly with other proteins to synergistically mediate biological functions. So far, hardly anything is known about this complex interplay between proteins in oomycetes. The only large-scale experimental study in oomycetes investigated the interactions between effector proteins produced by the downy mildew *Hyaloperonospora arabidopsidis* with known proteins of the *A. thaliana* (thale cress) immune system (Mukhtar et al. 2011). Although this study emphasized the importance of functional association data, it solely addressed complex formation within plant cells and not in the pathogen.

As an initial step on the way to fully expose the ensemble of all functional associations between proteins, we here present the first functional association network in *P. infestans*. We combined available genomic, transcriptomic and comparative genomic data to predict associations (interactions) between proteins resulting in a comprehensive network of gene associations that covers 33 percent of the predicted proteome. As expected, this number is lower than previous studies in *S. cerevisiae* (Lee et al. 2004) or *A. thaliana* (Lee et al. 2010), reflecting the relative paucity of data in *P. infestans* compared to these well studied model organisms. Nevertheless, the availability of these associations is crucial to provide insights into functional genomics. We balanced the coverage with an acceptable level of confidence given all available large-scale data and our *in silico* benchmark. The lack of experimentally confirmed benchmark sets in *P. infestans* limits a completely independent assessment of our prediction. In the future, more complementary gene expression data will most likely be available and consequently, to-



gether with experimentally determined interactions in *P. infestans* and closely related species, the genome-wide prediction of functional associations will be enhanced. We showed that proteins that are predicted to be functionally associated are enriched to reside in the same, or related, cellular sub-compartments, further validating the biological coherence of our predictions and the merit of the applied integrative approach. We exemplified the usability of the predicted functional association network on two examples: We automatically determined protein complexes and subsequently studied their constitution; an analysis that is not possible by just applying functional annotation to the genome. Moreover, we highlighted how the availability of the functional association network together with gene expression data allowed us to predict modules of functionally related genes during distinct phases of development. We exemplified this by analyzing the sporulation module that contained several experimentally characterized proteins such as Cdc14 and Pigpb1. The predicted physical interaction partners to these well-described proteins allowed us to place a concise set of candidates into a prominent role in sporangia formation.

Our study created a so far lacking addition to the growing genomic resources in the plant pathogenic model organism *P. infestans*. We demonstrated that these data are needed to further improve the ability to retrieve biological knowledge from large-scale data such as microarrays, RNA-seq or (phospho-) proteomics. The availability of the predicted functional association network allows a gradual transition from a single gene perspective to a more comprehensive understanding of the complex biology of *P. infestans* and other oomycetes.

## MATERIAL & METHODS

### Prediction of Orthologs between 51 Eukaryotic Species

We defined the groups of orthologs for a set of 51 eukaryotic species that were selected based on the taxonomic diversity. The orthologous groups were computed following an OMA (Orthologous Matrix)-like algorithm (Roth et al. 2008; Altenhoff et al. 2011) which was adjusted to the specific requirements of the analysis. To also identify weaker similarity between sequences we modified the following steps: (i) the minimal alignment score for potential orthologs was reduced to 130, (ii) the minimal alignment coverage was reduced to 40% in the first clustering step (assembling doubly-connected components, as opposed to cliques in the original OMA algorithm) and (iii) alignments with only 25% sequence coverage were added to the best matching cluster. We empirically determined the necessary cutoff values to maximize the inclusion of distant homologs while at the same time avoiding the excessive clustering of paralogs. This approach clustered in total 644,999 proteins into 58,533 orthologous groups. Each group represented all extant descendants from a single gene in the last common ancestor of eukaryotes; or, for a gene invented later, all descendants of that gene.

### Interolog Transfer of Protein-protein Interactions

We retrieved in total sixteen protein-protein interaction networks from six different sources (Supplementary Table S5-1C). Three of these data sets were subsequently subdivided, either to account for different levels of confidence expressed by the number of distinct publications (1PM-5PM) confirming an interaction (BioGRID) or to distinguish between core and attachments (IntAct). BioGRID interactions were mainly based on protein-protein interaction, however if at least a single publication reported the physical associations, also genetic interactions were considered to enhance the support for the specific association (2PM-5PM).

Interactions from the source databases were first mapped to the human Ensembl, yeast and *Arabidopsis* identifiers and subsequently projected from the source species to *P. infestans* using the identified orthologous groups. Since orthologous groups can also contain inparalogs, both in the source genomes (*H. sapiens*, *A. thaliana* and *S. cerevisiae*) and in *P. infestans*, we excluded all genes from the mapping with an alignment score to the source gene of less than 75% of the best matching inparalog, assuming that larger differences might be indicative of neo-functionalization of the paralog. If the mapped pairs still included inparalogs in *P. infestans*, we disentangled these specific cases by applying four different functional criteria to define which of the *P. infestans* proteins most likely retained the interaction. An interaction between two proteins is retained if both proteins (i) are on the same Kyoto Encyclopedia of Genes and Genomes (KEGG) map (Kanehisa et al. 2012), (ii) have protein domains that are known to mediate protein-protein interactions, (iii) share a common Gene Ontology (Ashburner et al. 2000) (GO) term (biological process or cellular component) at a depth of level  $\leq 6$  or  $\leq 5$ , respectively, (iv) share a common GO term (biological process or cellular component) at a depth of level  $\leq 4$  and their expression profiles have a Pearson correlation coefficient  $\geq 0.4$ . If none of these criteria was applicable we chose the protein with the highest similarity to the source protein so that we kept minimally one interaction between a set of orthologous groups.

The details of these four criteria to disentangle inparalogs in *P. infestans* are as follows: (i) To define pairs that are on the same KEGG map, we retrieved 94 predicted KEGG maps for *P. infestans* from the KEGG database (01.05.2012; excluding maps pif01100 and pif01110) that contained in total 1,329 proteins from *P. infestans* (7.5% of the predicted proteome). (ii) Protein domains that are predicted to mediate protein-protein interactions are retrieved from 3did (03.05.2012). Protein domains were predicted for the proteome of *P. infestans* using hmmer3 (Eddy 1998) (gathering cutoff) and a local Pfam-A database (v26) (Finn et al. 2010). (iii) We predicted the GO terms for all predicted proteins in *P. infestans* using the BLAST2GO algorithm (default parameters) (Conesa et al. 2005). Since GO is an acyclic graph, we first searched within each of the two domains (biological process or cellular component) for common GO terms between the two potentially interacting proteins. For all possible combinations of GO terms between the two proteins, we first searched all possible paths for common GO terms that minimize the distance to the initial GO term. If more than one GO term is equally

distant to the initial GO term, we chose the common term that minimized the distance to the root of the ontology. Subsequently, the depth of the common GO term that is shared between the proteins, which can be seen as a measure of functional similarity, is assigned to the pair by calculating the shortest path to the root of the ontology (Supplementary Figure S5-1). (iv) In addition to the approach outlined in (iii), we added gene expression data as a complementary feature (details for the gene expression analysis can be found below). We calculated Pearson correlation coefficients between the expression profiles of two pairs and kept an interacting pair if both the depth cutoff and the correlation cutoff were reached. Suitable cutoffs for the GO depth and the Pearson correlation in (iii) and (iv) were determined by maximizing the positive predicted value and the accuracy while minimizing the false discovery rate for 1,000 randomly picked true positive pairs as defined by KEGG (see above) and 1,000 random gene pairs or 500 pairs for (iii) and (iv), respectively.

### Functional Interactions by Additional Comparative Genomics Data

To define the functional interaction network in *P. infestans*, we added complementary data next to the predicted protein-protein interactions. We used (i) co-expression, (ii) conserved co-expression and (iii) co-occurrence to define these additional functional associations between two genes.

(i) Publicly available gene expression data for *P. infestans* was extracted from NCBI Gene expression omnibus (Barrett & Edgar 2006) with the accessions GSE9623 (Affymetrix), GSE13580 (Affymetrix), and GSE14480 (NimbleGen). The Affymetrix data were normalized using MAS5 and the  $\log_2$  of the expression intensities was computed using Bioconductor (Affy package) (Irizarry et al. 2003). Replicates were averaged and the resulting gene expression vector was normalized calculating the Z-score per unigene. Because the Affymetrix chip was designed prior to the availability of the genome sequence of *P. infestans*, we mapped the unigenes that have been used in the chip design to the transcripts derived from the *P. infestans* genome. We only considered the best hits of each unigene to the transcript set (blastn (Altschul et al. 1990), evalue cutoff  $1 \times 10^{-20}$ ,  $\geq 80$  percent identity). If several independent unigenes have the same transcript as their best hit we assigned the most C-terminal unigene to this transcript, since these unigenes tend to have the highest expression values. Normalized target intensities ( $\log_2$ ) were extracted from the NimbleGene data, replicates were averaged and Z-scores were calculated. The three independent experiments (the union of the genes in the three experiments is 8,947 genes) were combined to compute pairwise Pearson correlation coefficients between all genes.

(ii) To predict pairs of proteins that are encoded by conserved co-expressed gene pairs in *P. infestans*, we used defined orthologs between *P. infestans* and *P. sojae* as outlined above using a confined species selection. Furthermore, we used three publicly available gene expression data sets for *P. sojae* GSE15100 (Affymetrix), GSE22978 (Affymetrix) and GSE735084 (RNA-Seq). The analysis of the two Affymetrix expression sets was conducted as described above, however, before normalization all non-*P. sojae*

probes (the vast majority for this array) were removed. The RNA-seq derived gene expression intensities were  $\log_2$  transformed and otherwise treated similarly to the microarray experiments (see above). Pearson correlation coefficients of the normalized (Z-score) and subsequently combined gene expression values were calculated for all genes (union of the three experiments, i.e. 7,736 genes). A single unified score for each conserved co-expressed gene pair was derived by rescaling (between 0 and 1) the averaged Pearson correlation coefficients of the gene pair in *P. infestans* and the orthologous gene pair in *P. sojae*. The average Pearson correlation was calculated after applying a Fisher's Z-transformation to the individual correlation coefficients.

(iii) We predicted putative pairs of functionally associated protein by comparing the phylogenetic profiles of all genes with at least one gene loss during their evolutionary history. The similarity between profiles was measured by reconstructing the gene gain and loss events within an orthologous group over all 51 eukaryotic species. We ignored duplications, since the presence/absence of a gene within a genome was taken into account. We used 'partial correlations', as described by Cordero et al. (2008) in detail, to compare the gains and losses assigned to the branches of the species tree. The 'partial correlation' is based on the Pearson correlation coefficient of the events, but corrected against genome-wide trends such as whole-genome duplications or genome streamlining. Instead of applying a fixed threshold to indicate which correlation value still corresponds to a potential interaction, we sorted all pairs by their partial correlation and used the top 0.1% pairs.

### Bayesian Integration of Distinct Data Sources

We integrated the different data sources by applying a scoring system that is derived from Bayesian statistics followed by a Bayesian integration approach as outlined by Lee et al. (2004). Briefly, we calculated for each data source the log likelihood score (LLS) that two proteins are functionally linked, defined as  $LSS = \log(O_{\text{Posterior}} / O_{\text{Prior}})$ . The LLS was calculated based on the prior odds ( $O_{\text{Prior}}$ ), describing the ratio of probability of functional linkage and its negation without evidence, and the posterior odds ( $O_{\text{Posterior}}$ ), describing the ratio of probability of functional linkage and its negation under the given evidence. The prior odds can be estimated by the number of protein pairs that share a defined functional annotation, e.g. being on the same KEGG map, and the number of protein pairs that do not share the functional annotation, e.g. residing on two different KEGG maps. Similarly, we derived the posterior odds by estimating the number of protein pairs that share or do not share functional annotation and are supported by the given evidence. We used KEGG or GO – 6th level (biological process) to estimate the prior odds and the posterior odds for each dataset. If the dataset is discrete (e.g. protein-protein interactions) a single LLS is determined (Supplementary Table S5-1B). If the dataset has a continuous scoring schema, e.g. Pearson correlation coefficient for the co-expression data, we first determined a mapping function to re-score the initial score to the corresponding LLS. Therefore, we calculated the LLS for a given bin size and performed non-linear regression to determine the coefficients of the fitted function that is subsequently used to re-score the dataset to the LLS schema (Supplementary Fig-

ure S5-3). The combined LLS of all available evidences for an association between a pair of genes/proteins was calculated using a naïve Bayesian approach:  $LLS_{sum} = \text{SUM}(LSS)_{PPI} + \max(LSS)_{\text{BioGrid Human}} + \max(LSS)_{\text{BioGrid Yeast}} + LLS_{CE} + LLS_{CC} + LLS_{CO}$ . If the summed LLS was smaller than the cutoff, the association was not reported.

### Enrichment/Depletion of Subcellular Localization of Protein Pairs

We assessed the enrichment/depletion of the shared subcellular localization between pairs of associated proteins as outlined by Gandhi et al. (2006). Briefly, we calculated the fold enrichment/depletion of the number of observed edges between proteins of a certain subcellular localization (GO cellular compartment), e.g. number of edges between proteins where one partner is annotated as residing in the nucleus and the other in the endoplasmic reticulum, compared to the expected number of edges based on random networks that maintained the protein annotation, the degree for each protein (number of associations) and the total number of edges. The statistical significance was assessed using a Poisson distribution and the resulting p-value was corrected for multiple testing.

We independently predicted subcellular localization using the WoLF PSORT algorithm that uses sequence features and not homology to assign localization (Horton et al. 2007). We used both animal and fungi presets, assigning subcellular localization to protein upon agreement, otherwise to 'unknown'. Enrichment and depletion was otherwise calculated as described above.

### Functional Modules

Functional modules, i.e. maximally co-regulated sub-networks under a defined condition, were predicted based on differentially expressed genes between two conditions assessed by limma (Smyth 2004). The functional module was identified in a subset of the functional network, excluding associations that were merely supported by gene expression. Moreover, only the proteins whose genes have corresponding expression values and were part of the largest component within the sub-network were considered. The heuristic functional module within each sub-network was identified using BioNet (Beisser et al. 2010), where the p-values obtained from limma were scored using a fitted beta-uniform mixture model and a false discovery rate of 0.01. We were only interested in up-regulated modules during the defined condition, consequently we set all scores of proteins to  $-\text{abs}(S)$  when the gene expression difference expressed as fold ( $\log_2$ ) was smaller than 0, thereby only allowing inclusion of these nodes in the functional module if they connect high scoring nodes.

## ACKNOWLEDGMENTS

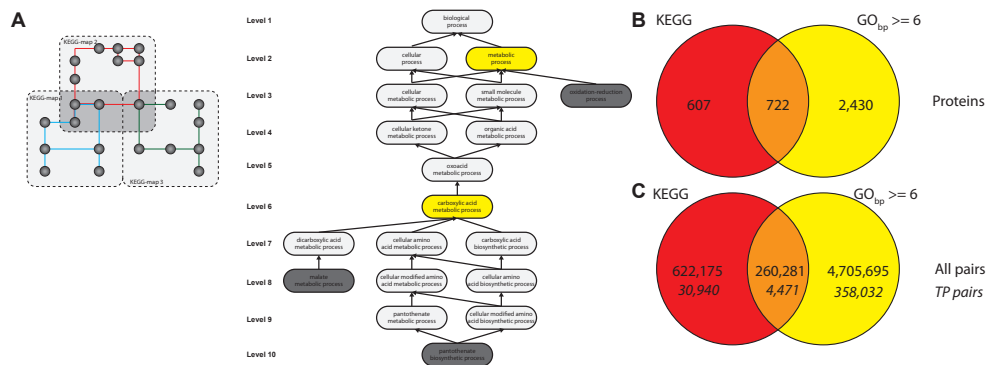
We gratefully acknowledge Lidija Berke for helpful discussions and comments on the manuscript. Moreover, we would thank Daniela Beisser and Tobias Müller for their

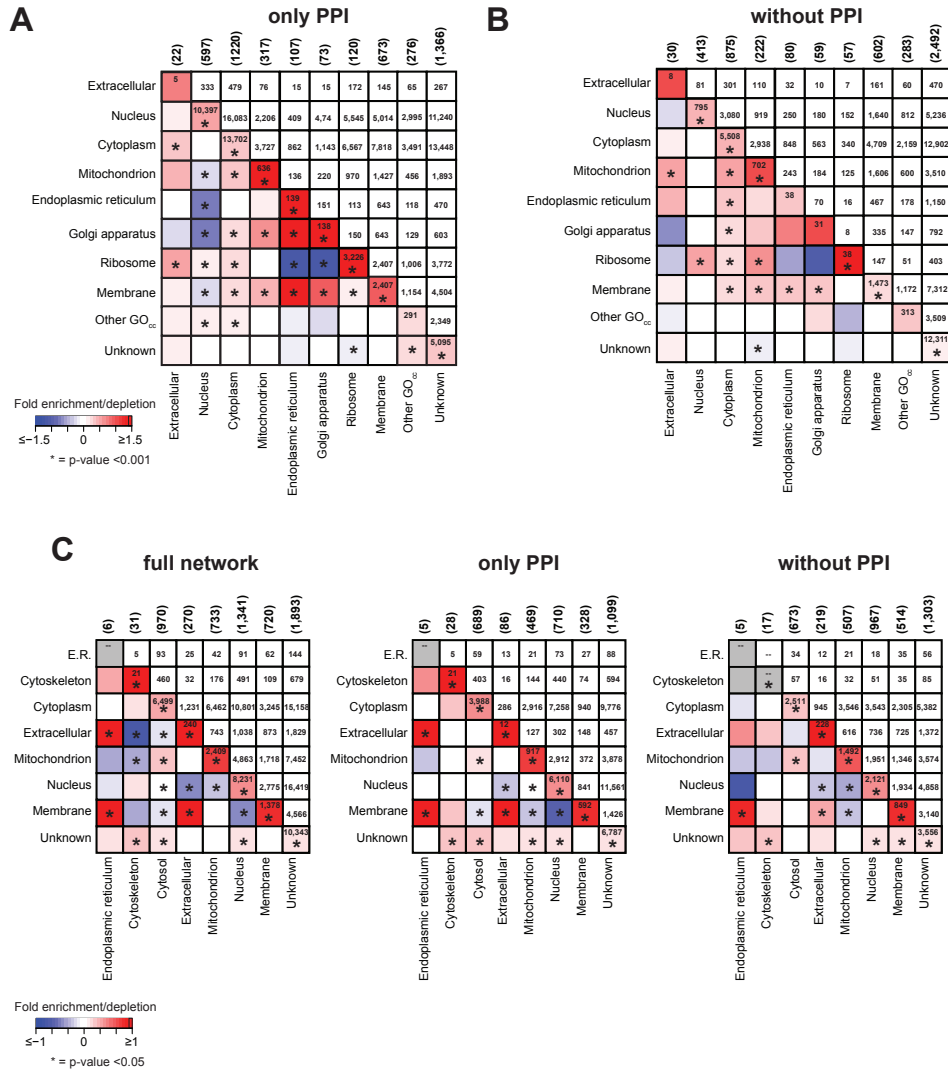
support with BioNET. This project was financed by the Centre for BioSystems Genomics (CBSG) which is part of the Netherlands Genomics Initiative/Netherlands Organisation for Scientific Research.

## SUPPLEMENTARY MATERIAL

Due to the amount of data, some of the Supplementary additional files, Supplementary Tables S5-1 to S5-4, are only accessible online ([http://bioinformatics.bio.uu.nl/michael/index\\_thesis.html](http://bioinformatics.bio.uu.nl/michael/index_thesis.html)).

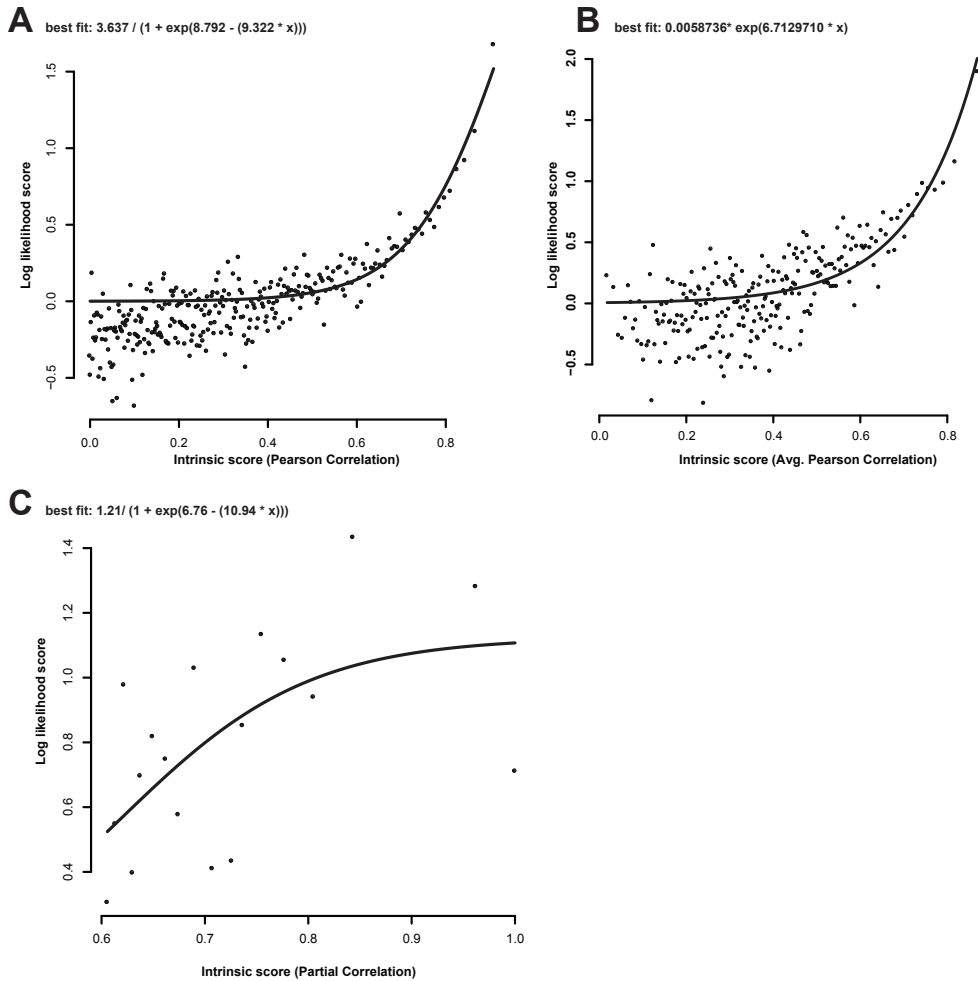
### Supplementary Figures





**Figure S5-2 Correlation of subcellular localization with predicted protein associations.**

The figure displays the log<sub>2</sub>-fold enrichment/depletion of protein pairs where both partners are predicted to reside in the same/different subcellular localization compared to the expected numbers. We discriminated between associations that have predicted protein-protein interactions as a source of evidence (A) and associations that were merely predicted by co-expression, conserved co-expression and co-occurrence (B). Panel (C) shows the same information, however the subcellular localization was predicted using WoLF PSORT (Material and Methods). Enrichment/depletion is displayed by the heatmap (lower half of the symmetrical matrix) (values saturate at ± 1.5 or ±1 for WoLF PSORT); the corresponding raw numbers are shown in upper half. Significant enrichment/depletion (after multiple testing correction) is indicated by an asterisk. The total number of proteins predicted to reside in a particular subcellular localization is displayed in brackets above the plot.



**Figure S5-3 Mapping of continuous scores to the unified log-likelihood schema.**

The figure displays the mapping of intrinsic continuous scores from (A) co-expression, (B) conserved co-expression and (C) phylogenetic co-occurrence to the unified log-likelihood schema. The derived mapping function (based on non-linear regression; Material and Methods) for each mapping is shown.



## REFERENCES

- Ah-Fong AMV, Judelson HS. 2003. Cell cycle regulator *Cdc14* is expressed during sporulation but not hyphal growth in the fungus-like oomycete *Phytophthora infestans*. *Mol. Microbiol.* 50:487–494.
- Ah-Fong AMV, Judelson HS. 2011. New role for *Cdc14* phosphatase: localization to basal bodies in the oomycete *Phytophthora* and its evolutionary coinheritance with eukaryotic flagella. *PLoS ONE.* 6:e16725.
- Altenhoff AM, Schneider A, Gonnet GH, Dessimoz C. 2011. OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Res.* 39:D289–94.
- Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. *J. Mol. Biol.* 215:403–410.
- Ashburner M et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* 25:25–29
- Barrett T, Edgar R. 2006. Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. *Meth. Enzymol.* 411:352–369.
- Beisser D, Klau GW, Dandekar T, Müller T, Dittrich MT. 2010. BioNet: an R-Package for the functional analysis of biological networks. *Bioinformatics.* 26:1129–1130.
- Blanco FA, Judelson HS. 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.
- Chen Z-Q et al. 2006. The essential vertebrate ABCE1 protein interacts with eukaryotic initiation factors. *J. Biol. Chem.* 281:7452–7457.
- Conesa A et al. 2005. Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics.* 21:3674–3676.
- Cordero OX, Snel B, Hogeweg P. 2008. Coevolution of gene families in prokaryotes. *Genome Res.* 18:462–468.
- Dong J et al. 2004. The essential ATP-binding cassette protein RLI1 functions in translation by promoting preinitiation complex assembly. *J. Biol. Chem.* 279:42157–42168.
- Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics.* 14:755–763.
- Finn RD et al. 2010. The Pfam protein families database. *Nucleic Acids Res.* 38:D211–22.
- Gandhi TKB et al. 2006. Analysis of the human protein interactome and comparison with yeast, worm and fly interaction datasets. *Nat. Genet.* 38:285–293.
- Gavin A-C et al. 2002. Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature.* 415:141–147.
- Gavin A-C et al. 2006. Proteome survey reveals modularity of the yeast cell machinery. *Nature.* 440:631–636.
- Goley ED, Welch MD. 2006. The ARP2/3 complex: an actin nucleator comes of age. *Nat. Rev. Mol. Cell Biol.* 7:713–726.
- Gollub J et al. 2003. The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res.* 31:94–96.
- Haas BJ et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature.* 461:393–398.
- Hinnebusch AG. 2006. eIF3: a versatile scaffold for translation initiation complexes. *Trends Biochem. Sci.* 31:553–562.
- Horton P et al. 2007. WoLF PSORT: protein localization predictor. *Nucleic Acids Res.* 35:W585–7.
- Hughes TR et al. 2000. Functional discovery via a compendium of expression profiles. *Cell.* 102:109–126.
- Irizarry RA et al. 2003. Summaries of Affymetrix GeneChip probe level data. *Nucleic Acids Res.* 31:e15.
- Jansen R et al. 2003. A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science.* 302:449–453.

- Jiang RHY, Tripathy S, Govers F, Tyler BM. 2008. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. *Proc. Natl. Acad. Sci. U.S.A.* 105:4874–4879.
- Judelson HS. 2012. Dynamics and innovations within oomycete genomes: insights into biology, pathology, and evolution. *Eukaryotic Cell.* 11(11): 11:1304–1312.
- Judelson HS et al. 2008. Gene expression profiling during asexual development of the late blight pathogen *Phytophthora infestans* reveals a highly dynamic transcriptome. *Mol. Plant Microbe Interact.* 21:433–447.
- Judelson HS, Ah-Fong AMV. 2010. The kinome of *Phytophthora infestans* reveals oomycete-specific innovations and links to other taxonomic groups. *BMC Genomics.* 11:700.
- Judelson HS, Narayan RD, Ah-Fong AMV, Kim KS. 2009. Gene expression changes during asexual sporulation by the late blight agent *Phytophthora infestans* occur in discrete temporal stages. *Mol. Genet. Genomics.* 281:193–206.
- Judelson HS, Tani S. 2007. Transgene-induced silencing of the zoosporegenesis-specific NIFC gene cluster of *Phytophthora infestans* involves chromatin alterations. *Eukaryotic Cell.* 6:1200–1209.
- Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.* 40:D109–14.
- Ketelaar T, Meijer HJG, Spiekerman M, Weide R, Govers F. 2012. Effects of latrunculin B on the actin cytoskeleton and hyphal growth in *Phytophthora infestans*. *Fungal Genet. Biol.* 49:1014–1022.
- Krause R, Mering von C, Bork P, Dandekar T. 2004. Shared components of protein complexes—versatile building blocks or biochemical artefacts? *Bioessays.* 26:1333–1343.
- Latijnhouwers M, de Wit PJGM, Govers F. 2003. Oomycetes and fungi: similar weaponry to attack plants. *Trends Microbiol.* 11:462–469.
- Latijnhouwers M, Govers F. 2003. A *Phytophthora infestans* G-protein beta subunit is involved in sporangium formation. *Eukaryotic Cell.* 2:971–977.
- Latijnhouwers M, Ligterink W, Vleeshouwers VGAA, van West P, Govers F. 2004. A Galpha subunit controls zoospore motility and virulence in the potato late blight pathogen *Phytophthora infestans*. *Mol. Microbiol.* 51:925–936.
- Lee I, Ambaru B, Thakkar P, Marcotte EM, Rhee SY. 2010. Rational association of genes with traits using a genome-scale gene network for *Arabidopsis thaliana*. *Nat. Biotechnol.* 28:149–156.
- Lee I, Date SV, Adai AT, Marcotte EM. 2004. A probabilistic functional network of yeast genes. *Science.* 306:1555–1558.
- María Laxalt A, Latijnhouwers M, van Hulten M, Govers F. 2002. Differential expression of G protein alpha and beta subunit genes during development of *Phytophthora infestans*. *Fungal Genet. Biol.* 36:137–146.
- Mukhtar MS et al. 2011. Independently evolved virulence effectors converge onto hubs in a plant immune system network. *Science.* 333:596–601.
- Nepusz T, Yu H, Paccanaro A. 2012. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods.* 9:471–472.
- Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc. Natl. Acad. Sci. U.S.A.* 96:4285–4288.
- Randall TA et al. 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.
- Roth ACJ, Gonnet GH, Dessimoz C. 2008. Algorithm of OMA for large-scale orthology inference. *BMC Bioinformatics.* 9:518.
- Said MR, Begley TJ, Oppenheim AV, Lauffenburger DA, Samson LD. 2004. Global network analysis of pheno-

- typic effects: protein networks and toxicity modulation in *Saccharomyces cerevisiae*. Proc. Natl. Acad. Sci. U.S.A. 101:18006–18011.
- Schwikowski B, Uetz P, Fields S. 2000. A network of protein-protein interactions in yeast. Nat. Biotechnol. 18(12):1257–1261.
- Seidl MF, Van den Ackerveken G, Govers F, Snel B. 2011. A domain-centric analysis of oomycete plant pathogen genomes reveals unique protein organization. Plant Physiol. 155:628–644.
- Seidl MF, Van den Ackerveken G, Govers F, Snel B. 2012. Reconstruction of oomycete genome evolution identifies differences in evolutionary trajectories leading to present-day large gene families. Genome Biol Evol. 4:199–211.
- Shoemaker CJ, Green R. 2011. Kinetic analysis reveals the ordered coupling of translation termination and ribosome recycling in yeast. Proc. Natl. Acad. Sci. U.S.A. 108:E1392–8.
- Smyth GK. 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Stat Appl Genet Mol Biol. 3:Article3.
- Stassen JH, Van den Ackerveken G. 2011. How do oomycete effectors interfere with plant life? Curr Opin Plant Biol. 14:1–8.
- Tyler BM et al. 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. Science. 313:1261–1266.
- van Noort V, Snel B, Huynen MA. 2003. Predicting gene function by conserved co-expression. Trends Genet. 19:238–242.
- Walhout AJ et al. 2000. Protein interaction mapping in *C. elegans* using proteins involved in vulval development. Science. 287:116–122.





# Summarizing Discussion

# 6



## INTRODUCTION

The work described in this thesis combines four complementary comparative genomic approaches focusing on pathogenic oomycetes. Comparative genomics and integrative bioinformatics are powerful tools to accompany classical experimental work (Bork et al. 1998). This is of special importance for non-model organisms such as oomycetes, but also for many fungi and vertebrates where broad biological insights are lagging behind the possibilities provided by the nowadays more easily accessible large-scale omics data. We have applied highly integrative bioinformatic approaches to the growing amount of omics data revealing different aspects of cellular function and evolution in oomycetes. These different aspects range from the evolution of gene content (chapter 2 and chapter 3) and regulation of gene expression by *cis*-regulatory elements (chapter 4) to functional associations between proteins (chapter 5). Especially in chapter 2 and chapter 5, we were able to derive hypotheses on the function of many uncharacterized gene products, therefore providing significant insight into the biology of oomycetes. Here I reflect on these bioinformatics analyses, especially in the light of related work (see also my list of publications), and discuss recurring themes. This chapter includes an outlook on the paths that this highly dynamic field will most likely take in the near future and emphasizes why comparative genomics is an essential tool to close the gap between available data and testable hypotheses in this interesting group of organisms.

## OOMYCETE GENOME SEQUENCES

Over the last decade we experienced the emergence of large-scale genomics in oomycetes. It started with extensive work on expressed sequence tags (ESTs) (Kamoun et al. 1999; Qutob et al. 2000; Randall et al. 2005; Torto-Alalibo et al. 2005) and the release of the genome sequences of two *Phytophthora* species (Tyler et al. 2006). In recent years a broad range of oomycete genomes has been sequenced (Figure 1-2). The initial decision to sequence two genomes — a result of a compromise between social and scientific interest — also enabled the use of comparative genomics (Govers & Gijzen 2006). The availability of more genomes paved the way to study the evolution and function to an increased depth. Below I will discuss the possible complementary roadmap of sequencing efforts in oomycetes and several open questions that can be answered with the help of additional genome sequences.

Lamour and colleagues depicted already in 2007 three complementary roads to proceed with the sequencing efforts in oomycetes (Lamour et al. 2007): (i) sequencing of a wide spectra of species covering distinct clades of oomycetes, (ii) sequencing of species that represent divergent lifestyles, and (iii) sequencing of strains or sibling species related to those that have already been sequenced. By now all of their predictions have been realized. There are, however, still many gaps where more genome sequences are crucial to answer the many remaining open questions.

Several members of distinct clades of oomycetes and species with divergent life-

styles, e.g. the obligate biotroph *Hyaloperonospora arabidopsidis* (Baxter et al. 2010), have been sequenced (Figure 1-2). However, all oomycetes sequenced so far are pathogens. With the exception of a recently sequenced pathogen of animals, *Saprolegnia parasitica* (Jiang et al. 2013), all sequenced oomycetes parasitize plants; not a single saprophytic species is included. An open question in the evolution of oomycetes is how pathogenicity evolved from likely saprophytic ancestors to (obligate) pathogenic species (chapter 1). There is not a clear black/white difference between the genome content of saprophytes and pathogens: An illustrative example of this continuum is *ECP2*, a gene that encodes a secreted protein causing reduced virulence upon silencing in the plant pathogenic fungus *Cladosporium fulvum* (Laugé et al. 1997). Interestingly, the genomes of the fungal saprophyte *Neurospora crassa*, but also many other non-plant pathogenic as well as pathogenic fungi, encode *ECP2* homologs that represent an ancient superfamily of putative effectors (Galagan et al. 2003; Stergiopoulos et al. 2012). Given this diverse phylogenetic distribution, a distinct function in non-pathogenic species seems likely, e.g. an antagonistic interaction towards other microorganisms (Stergiopoulos et al. 2012). The example of the fungal *ECP2* highlights the necessity to sequence saprophytic species. From an evolutionary perspective, their sequences might hold the key to more precisely understand the emergence, evolution and function of different infection strategies and the genes required to overcome host defenses.

Other questions are connected to the exact constitution of the core genome of oomycetes and the evolution of the families leading to the repertoire observed in extant species. We have shown that genes encoding transcription factors and genes involved in signal transduction have expanded early in the evolution of oomycetes (Seidl et al. 2012; chapter 3). Especially protein kinases are abundant in *Phytophthora* spp. (Judelson & Ah-Fong 2010; Seidl et al. 2012), but also in *S. parasitica* (Jiang et al. 2013), pointing to an early expansion. Full genome sequences of early-branching oomycetes such as *Eurychasma dicksonii* are not yet available. These are crucial to study the oomycete core genome and the evolutionary history of these expanded gene families, providing more resolution at the last common ancestor of the so far sequenced oomycetes (chapter 3).

Especially in the last couple of years, the focus has shifted from the sequencing of more diverse organisms to the sequencing of strains (isolates) and siblings (e.g. Raffaele et al. 2010; Cabral et al. 2011; Cooke et al. 2012). This approach is very powerful due to the availability of next generation sequencing techniques and template genomes that can assist the rapid assembly and subsequent analysis. Whereas studies on more divergent species focus on larger patterns of evolution (see above), studying lineages and sibling species answers pending questions on small scale genomic variations, identification of active effector genes and their contribution to the adaptation to different hosts and evasion of host responses. The studies in emerging *Phytophthora infestans* lineages such as blue *13\_A2* and closely related sibling species such as *Phytophthora mirabilis*, revealed common presence/absence, copy number variations, gene expression and single nucleotide polymorphisms as well as signatures of positive selection. These features occur within the flexible part of the bipartite genome (chapter 1) and are mainly

localized in effector genes thereby likely reflecting genome adaptation following a host jump and elevated virulence (Raffaele et al. 2010; Cooke et al. 2012). Cabral and colleagues (2011) used comparative genomics to analyze the expressed sequence tags of the *H. arabidopsidis* isolate Waco9. They focused on the secretome, especially on RXLR effectors that are thought to function as suppressors of immune responses, and identified *H. arabidopsidis* specific effectors as well as a few Waco9-specific effectors such as *RXLR29*. The authors hypothesize, similar to Cooke and colleagues (2012), that isolate-specific effector repertoires might explain the differences in virulence towards the same hosts, thereby highlighting the merit of sequencing and comparative genomics of closely related species/isolates.

Interestingly, a community-based approach seeks to sequence the whole genus *Phytophthora* thereby providing an unseen wealth of genomic and transcriptomic data (B.M. Tyler, personal communication). These data will allow new comparative studies. The pending questions on the biology and evolution, for example on the regulation of gene expression and the evolution of the genome organization, can be answered. Phylogenetic footprinting to define conserved regulatory elements in the non-coding regions of their genome will most likely considerably enhance previous approaches (chapter 4). Moreover, the exact time and mode of the emergence of the bipartite genome that has so far been observed most prominently in *P. infestans* and closely related sister taxa and to a lesser extent in *Phytophthora sojae* and *Phytophthora ramorum* (chapter 1) is still not fully understood (Haas et al. 2009; Raffaele et al. 2010).

Within oomycetes, complementary paths regarding the sequencing efforts have been rationalized and taken. Even though several questions regarding the fundamental evolution of oomycetes have been successfully answered (Seidl et al. 2012; chapter 3), there are pending questions, such as the origin and evolution of pathogenicity, that would require the genome sequences of additional divergent oomycetes. However given the negative impact of plant pathogens on crop yield, I anticipate that sequencing isolates and close species will have higher priority. Comparative genomics enables us to answer fundamental questions concerning host adaptation as a result of small scale evolution as successfully demonstrated in oomycetes (Raffaele et al. 2010), but also in fungal pathogens (e.g. de Jonge et al. 2012). The perspective to link these variations to host range and lifestyle is tempting and necessary considering the economic and ecological importance of these organisms. Moreover, the application of next generation sequencing together with comparative genomics allows direct and timely surveillance of pathogen populations thereby directly assisting anticipatory breeding efforts by effective deployment of resistance in agriculture (Vleeshouwers et al. 2011; Cooke et al. 2012).

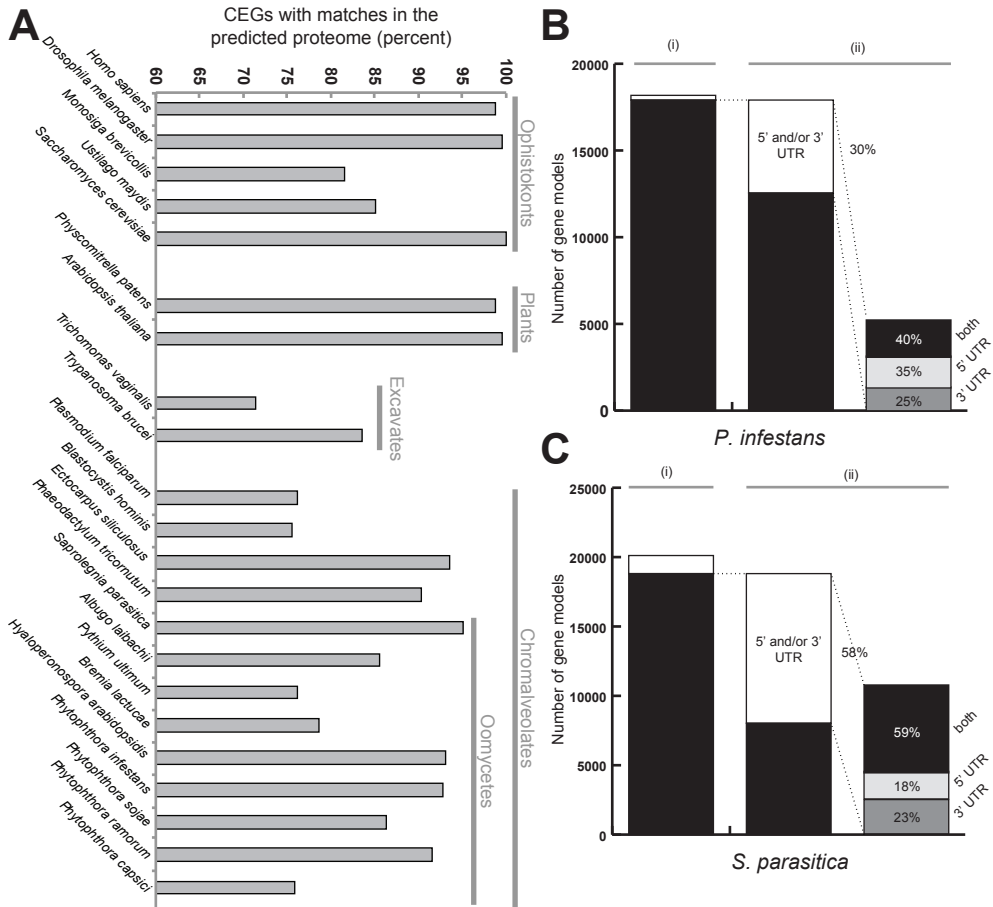


## GENOME ANNOTATION QUALITY AFFECTS COMPARATIVE ANALYSIS

The genome contains all necessary information for the biology of an organism, in particularly the genes and other functional regions such as regulatory DNA (see e.g. the human ENCODE project (Maher 2012)). Therefore, the genome sequence represents a unique resource; its value however depends on the quality of its assembly and subsequent annotation, in first instance of its protein coding genes (Stein 2001). Whereas gene annotation seems trivial in prokaryotic genomes, this endeavor is more challenging in eukaryotes because of the complex gene models containing exons and introns and the presence of pseudogenes. In human, for example, ~12,000 pseudogenes have recently been described, some of which still resemble their 'parents' and are therefore hard to distinguish in an automated annotation pipeline (Pei et al. 2012).

Due to the advances in sequencing technologies, new genome sequences become rapidly available and gene identification and annotation is mostly performed automatically. However, the automated analysis can result in missed gene identifications and introduce misannotations. This is a common problem particularly in organisms that receive little attention or lack the necessary community to provide manual gene annotation. Even in model organisms such as *Arabidopsis thaliana* and human, wrong and changing annotation is apparent (Haas et al. 2005; Harrow et al. 2012). Parra and colleagues (2007;2009) derived a set of low copy eukaryotic genes that are supposed to be present in all eukaryotes and therefore considered highly conserved (core eukaryotic genes (CEG); 248 genes). The number of genes from this set in analyzed genomes correlates reasonably well with the overall genome completeness (Parra et al. 2007; 2009). The percentage of mapped CEGs differs within eukaryotes, in particular between well studied and well annotated model organisms and more recently sequenced organisms and ranges from 70-100 percent (Figure 6-1A; Parra et al. 2009). Overall, oomycetes have a relatively high percentage of matches (Figure 6-1A); even the recently sequenced transcriptome of *Bremia lactucae* covers nearly 80 percent of the CEGs (Stassen et al. 2012). Therefore, the overall completeness of oomycete genome sequences is relatively high and indicates a reasonable annotation quality. However, CEGs provide only rough estimates of the completeness of the genome assembly and, especially due to their conserved nature, annotated gene space.

The apparent absence of genes in a genome can have a technical reason, due to errors in annotation, but can also have biological significance, such as the loss of genes involved in the nitrate assimilation pathway in *H. arabidopsidis* (Baxter et al. 2010; chapter 1). Although erroneous losses had only a minor impact on large-scale evolutionary studies in Chromalveolates (Martens et al. 2008), this issue is of especially important in detailed studies of specific biological pathways or protein complexes. For example a small, but conserved subunit of the anaphase-promoting complex is apparently absent from the annotated gene set in *P. infestans*. Detailed analyses, however, revealed its presence in the genome (B. Snel, personal communication), therefore likely significantly



**Figure 6-1** Completeness of annotated gene sets and frequency of annotated 5' and 3' UTR.

(A) Estimation of the completeness of the annotated gene sets in different sequenced eukaryotic genomes with emphasis on oomycetes. The conserved core genes were derived from Parra and colleagues (2009) and matches were searched in the predicted proteomes. For (B) *P. infestans* and (C) *S. parasitica* the number of genes are displayed: (i) The total number of annotated genes as well as the subset of annotated genes with a corresponding coding region (black), and (ii) the frequency of annotated 5' and/or 3' untranslated regions (UTR) (white) in the annotated coding genes. Both genomes were retrieved from the Broad Institute (09.11.2012).

altering the interpretation of the results.

Next to the possibility of the artificial absence, misannotation affects large-scale comparative genomics research in oomycetes. The identification of novel domain combinations (Seidl et al. 2011; chapter 2) is highly dependent on the quality of the gene models; wrong annotations, especially artificial fusions and fissions, will affect the results even though most of the combinations are conserved in *Phytophthora* (see discussion chapter 2). Moreover, for the majority (70 percent) of protein-coding genes

in *P. infestans* the 5' and/or 3' untranslated regions are not annotated (Figure 6-1B). Without this information, the exact transcription start and stop sites of genes are not known. The lack of knowledge about the precise transcription start site hampers the analysis and evaluation of automatically inferred DNA motifs with respect to their preferential occurrences upstream of coding regions (Seidl et al. 2012; chapter 4) and also identification of possible regulatory elements in the untranslated regions (reviewed in Mignone et al. 2002). By applying bioinformatics, we corroborated in a large-scale study that the oomycete Inr-like element (Inr/FPR; coinciding with the transcription start site Pieterse et al. 1994; McLeod et al. 2004) is indeed the most common eukaryotic promoter element in *P. infestans* occurring in 37% of all genes (Seidl et al. 2012; chapter 4). Identification of this specific DNA motif (see discussion chapter 4) will allow a more precise bioinformatic prediction of the transcription start sites of genes in *Phytophthora* and other oomycetes in the future.

Extensive work on yeast has shown that sequencing of closely related genomes can greatly facilitate genome annotation and gene model validation (Kellis et al. 2003). Kellis and colleagues (2003) applied comparative genomics to four *Saccharomyces* species and identified or refined existing gene models. This led to 15% change in the beforehand published genome annotation. Even though homology has already been used for gene annotation in oomycetes, e.g. in the *P. infestans* genome (Haas et al. 2009), rapid sequencing of oomycetes, especially many more sequences within closely defined taxa such as *Phytophthora* (see above), will allow a more powerful application of synteny and conservation to further refine the initial gene models. It is anticipated that the number of sequenced oomycetes will increase considerably in the next few years. However, due to the relatively large genomes of many members (Figure 1-2) and the high repeat content, assembly and annotation remains challenging, labor intensive and expensive. Therefore, the application of complementary strategies is needed to aid genome annotation and gene model refinement in oomycetes.

The availability of next generation sequencing data such as RNA-Seq will most likely help to solve the outlined limitations and problems, not only in model organisms but, due to its speed and cost efficiency, also in less standard organisms such as oomycetes. Even though RNA-Seq reveals only the transcripts at a defined point of time, pooling different samples will support gene calling and annotation (5'/3' UTR, exons/introns, isoforms) resulting in a significant increase in the quality of the gene sets as well as in the subsequent bioinformatic analyses. Application of RNA-Seq in *de novo* annotations of transcripts in fission yeast, mouse and cucumber significantly enhanced gene annotation (Grabherr et al. 2011; Li et al. 2011). Similar improvements in annotation quality are visible in *S. parasitica* where gene annotation has been supported by RNA-Seq data (Jiang et al. 2013). The number of gene models with 5' and/or 3' UTR is considerably higher compared to *P. infestans* (Figure 6-1B & Figure 6-1C). Moreover, full transcriptome RNA-Seq data will allow a precise evaluation of the observed novel gene combinations in *Phytophthora* spp. and *S. parasitica* (Seidl et al. 2011, chapter 2; Jiang et al. 2013, Morris et al. 2009). Initial studies found rather few introns in *Phytophthora* and other oomycetes, with the exception of *S. parasitica* (Tyler et al. 2006; Jiang et al. 2013).

Analyses in *P. sojae* revealed few genes (122) that likely show alternative splicing (Shen et al. 2011). Therefore, RNA-Seq data, together with the essential analysis tools (Rogers et al. 2012), will shed additional light on possible alternative splicing in oomycetes. So far, RNA-Seq data for oomycetes is sparse as only few experiments in diverse species are available (Lévesque et al. 2010, Jiang et al. 2013, Ye et al. 2011, Links et al. 2011; Kunjeti et al. 2012; Savory et al. 2012), but more data will likely be available in the near future. These data will not only directly enhance genome annotation, but also aid in answering questions regarding novel gene features (Seidl et al. 2011, chapter 2), alternative splice variants and differential gene expression (chapter 5), thereby significantly adding to the biological knowledge in oomycetes.

## COMPLEX OMICS DATA ACCESSIBILITY FOR OOMYCETE BIOLOGISTS

In recent years, huge advances in genomic sequencing and in the availability of large and diverse datasets covering other, often complementary, omics data have been obtained (chapter 1). To make these complex data and the initial analyses easily accessible for all biologists (not only bioinformaticians), integrative platforms that store, link and visualize these data are essential. Whereas a considerable number of platforms exist for model organisms such as yeast (Cherry et al. 1998; 2012), worm (Stein et al. 2001) and fly (Drysdale and FlyBase Consortium 2008), similar platforms for oomycetes are not yet available.

It was realized more than ten years ago that rapid, easy access and availability of the human genome is crucial for scientific progress. The Ensembl database that was developed provides access to human chromosomes, gene models and proteins allowing the study of human genes and their homologs on a genome-wide scale (Flicek et al. 2010). Since then, many genome sequences of a large variety of mainly vertebrates have been integrated in Ensembl, resulting in a unique repository and comparative genomics platform for vertebrate genome sequences (Flicek et al. 2010). Early on, comparative aspects such as genome alignments, genomic variations, automatic (phylogenetically driven) orthology/paralogy determination for gene families (Ensembl compara; Vilella et al. 2009) and comprehensive access with standardized output (IDs, sequence format, query system) have been developed (Kinsella et al. 2011). Other approaches focusing on single model organisms such as the *Saccharomyces* genomics database (SGD), the worm resource and the *Drosophila* database provide centralized databases for (comparative) genomics. Moreover, they also assess and integrate other large-scale complementary datasets such as gene expression, functional annotation and functional associations (chapter 5) that are derived from high-throughput experiments and/or literature.

The outlined databases are the result of a continuous development from simple storage warehouses to comprehensive omics platforms (Flicek et al. 2010). They facilitate experimental design (Cherry et al. 2012) and help experimental biologists to evalu-

ate and compare their results given (all) available genomic data, not only in the specific species but also in their relatives. These complex resources, however, are not yet available for oomycetes. Genomic sequences of oomycetes are mainly distributed over the genome centers that were involved in the initial sequencing such as the Broad Institute (<http://www.broadinstitute.org>) and the Joint Genome Institute (<http://www.jgi.doe.gov>), complicating access and standardization, and hampering subsequent analyses for (non-) bioinformaticians. Moreover, comparative genomics such as phylogenomics analyses and other high throughput data have to be gathered from supplementary material and analyzed independently. Integrative analyses in their wealth are not available to the community. The oomycete research community would benefit largely from a community-driven approach to centralize data storage and integration to access the complex and wealth of large-scale omics data. Recently, some already published oomycete genomes have been deposited in Ensembl protists and in FungiDB (Stajich et al. 2012), thereby providing centralized access. However they lack newly sequenced genomes, accompanying transcriptomics data and integrative analyses provided by the community. Since these data are expected to grow considerably in the close future, it is timely to start a community-driven approach, providing integration of the data discussed in this thesis and even more comparative/integrative data in the future.

## COMPARATIVE GENOMICS IN OOMYCETES AS A BLUEPRINT FOR OTHER TAXA

I highlight in this thesis how the application of integrative computational approaches is feasible given the available — albeit limited — genomic and transcriptomic data. We used these data to predict regulatory motifs in the DNA sequences of three *Phytophthora* genomes, thereby providing the first genome-wide survey of such motifs in oomycetes (Seidl et al. 2012; chapter 4). Moreover, we exploited all available omics data for *P. infestans* to predict the first intra-species functional associations in any oomycete (chapter 5). We used the derived network as a platform to further analyze large-scale transcriptomics data thereby assigning functions and likely associations to many as yet unknown gene products.

The computational approaches described in this thesis provide examples of the crucial importance of comparative genomics to maximize biological insights in oomycetes. However, many more diverse species with economic, ecological and medical importance are already sequenced or will become available in the close future. Our research is applicable to these diverse taxa, too. Therefore, this thesis serves as a ‘blueprint’ to guide and aid further comparative genomics work in other important species with rather limited biological knowledge. For example in *Plasmodium* species, the causal agent of malaria, different authors have conducted similar lines of research primarily applying comparative genomics, which significantly increased insights in the evolution of these pathogens and in gene function (e.g. van Noort & Huynen 2006; Pick et al. 2011). Centralized data sources that integrate functional, genomics, transcriptomics and func-

tional associations data, but especially also comparative genomics (Ensembl compara), are necessary as well. Such data sources are crucial to allow rapid access to newly generated large-scale data and are pivotal for accelerating science in these important taxa.

## CONCLUDING REMARKS

Oomycete research has been a continuous endeavour since the initial description of one of its members over 150 years ago and considerably gained velocity with the availability of their first sequenced genomes. This trend will likely persist because more genomes, also the first sequences from saprophytic oomycetes and especially from siblings and isolates of known pathogens, will become available and focus the research to specific evolutionary and functional questions (chapter 6). This thesis describes a set of comparative genomics papers that integrate the available diverse omics data to answer questions on the function of as yet unknown gene products and the evolution, function and regulation of central cellular processes in oomycetes. Moreover, it highlights the merit of integrative, comparative approaches to close the knowledge gap between the available omics data and the biology of non-model organisms, thereby serving as a 'blueprint' for similar studies in other organisms.

## REFERENCES

- Baxter L et al. 2010. Signatures of adaptation to obligate biotrophy in the *Hyaloperonospora arabidopsidis* genome. *Science*. 330:1549–1551.
- Bork P et al. 1998. Predicting function: from genes to genomes and back. *J. Mol. Biol.* 283:707–725.
- Cabral A et al. 2011. Identification of *Hyaloperonospora arabidopsidis* transcript sequences expressed during infection reveals isolate-specific effectors. *PLoS ONE*. 6:e19328.
- Cherry JM et al. 2012. *Saccharomyces* Genome Database: the genomics resource of budding yeast. *Nucleic Acids Res.* 40:D700–5.
- Cherry JM et al. 1998. SGD: *Saccharomyces* Genome Database. *Nucleic Acids Res.* 26:73–79.
- Cooke DEL et al. 2012. Genome analyses of an aggressive and invasive lineage of the irish potato famine pathogen. *PLoS Pathog.* 8:e1002940.
- de Jonge R et al. 2012. Tomato immune receptor Ve1 recognizes effector of multiple fungal pathogens uncovered by genome and RNA sequencing. *Proc. Natl. Acad. Sci. U.S.A.* 109:5110–5115.
- Drysdale R, FlyBase Consortium. 2008. FlyBase: a database for the *Drosophila* research community. *Methods Mol. Biol.* 420:45–59.
- Flicek P et al. 2010. Ensembl's 10th year. *Nucleic Acids Res.* 38:D557–62.
- Galagan JE et al. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature*. 422:859–868.
- Govers F, Gijzen M. 2006. *Phytophthora* genomics: the plant destroyers' genome decoded. *Mol. Plant Microbe Interact.* 19:1295–1301.
- Grabherr MG et al. 2011. Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 29:644–652.
- Haas BJ et al. 2005. Complete reannotation of the *Arabidopsis* genome: methods, tools, protocols and the

- final release. *BMC Biol.* 3:7.
- Haas BJ et al. 2009. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature.* 461:393–398.
- Harrow J et al. 2012. GENCODE: the reference human genome annotation for The ENCODE Project. *Genome Res.* 22:1760–1774.
- Jiang RHY et al. 2013. Distinctive expansion of potential virulence genes in the genome of the oomycete fish pathogen *Saprolegnia parasitica*. *PLoS Genetics*. In press
- Judelson HS, Ah-Fong AMV. 2010. The kinome of *Phytophthora infestans* reveals oomycete-specific innovations and links to other taxonomic groups. *BMC Genomics.* 11:700.
- Kamoun S, Hrabar 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.
- Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES. 2003. Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature.* 423:241–254.
- Kinsella RJ et al. 2011. Ensembl BioMart: a hub for data retrieval across taxonomic space. *Database (Oxford)*. 2011:bar030.
- Kunjeti SG et al. 2012. RNA-Seq reveals infection-related global gene changes in *Phytophthora phaseoli*, the causal agent of lima bean downy mildew. *Mol Plant Pathol.* 13:454–466.
- Lamour KH, Win J, Kamoun S. 2007. Oomycete genomics: new insights and future directions. *FEMS Microbiol. Lett.* 274:1–8.
- Laugé R, Joosten MHAJ, Van den Ackerveken GFJM, Van den Broek HWJ, de Wit PJGM. 1997. The in planta-produced extracellular proteins ECP1 and ECP2 of *Cladosporium fulvum* are virulence factors. *Mol. Plant Microbe Interact.* 10:725–734.
- Lévesque CA et al. 2010. Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire. *Genome Biol.* 11:R73.
- Li Z et al. 2011. RNA-Seq improves annotation of protein-coding genes in the cucumber genome. *BMC Genomics.* 12:540.
- Links MG et al. 2011. *De novo* sequence assembly of *Albugo candida* reveals a small genome relative to other biotrophic oomycetes. *BMC Genomics.* 12:503.
- Maher B. 2012. ENCODE: The human encyclopaedia. *Nature.* 489(7414):46-8.
- Martens C, Vandepoele K, Van de Peer Y. 2008. Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proc. Natl. Acad. Sci. U.S.A.* 105:3427–3432.
- McLeod A, Smart CD, Fry WE. 2004. Core promoter structure in the oomycete *Phytophthora infestans*. *Eukaryotic Cell.* 3:91–99.
- Mignone F, Gissi C, Liuni S, Pesole G. 2002. Untranslated regions of mRNAs. *Genome Biol.* 3(3)
- Morris PF et al. 2009. Multiple horizontal gene transfer events and domain fusions have created novel regulatory and metabolic networks in the oomycete genome. *PLoS ONE.* 4:e6133.
- Parra G, Bradnam K, Korf I. 2007. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics.* 23:1061–1067.
- Parra G, Bradnam K, Ning Z, Keane T, Korf I. 2009. Assessing the gene space in draft genomes. *Nucleic Acids Res.* 37:289–297.
- Pei B et al. 2012. The GENCODE pseudogene resource. *Genome Biol.* 13:R51.
- Pick C, Ebersberger I, Spielmann T, Bruchhaus I, Burmester T. 2011. Phylogenomic analyses of malaria parasites and evolution of their exported proteins. *BMC Evol. Biol.* 11:167.
- Pieterse CM et al. 1994. Structure and genomic organization of the *ipiB* and *ipiO* gene clusters of *Phytophthora infestans*. *Gene.* 138:67–77.

- Qutob D, Hraber PT, Sobral BW, Gijzen M. 2000. Comparative analysis of expressed sequences in *Phytophthora sojae*. *Plant Physiol.* 123:243–254.
- Raffaele S et al. 2010. Genome evolution following host jumps in the Irish potato famine pathogen lineage. *Science.* 330:1540–1543.
- Randall TA et al. 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.
- Rogers MF, Thomas J, Reddy AS, Ben-Hur A. 2012. SpliceGrapher: detecting patterns of alternative splicing from RNA-Seq data in the context of gene models and EST data. *Genome Biol.* 13:R4.
- Savory EA et al. 2012. mRNA-Seq analysis of the *Pseudoperonospora cubensis* transcriptome during cucumber (*Cucumis sativus* L.) infection. *PLoS ONE.* 7:e35796.
- Seidl MF, Van den Ackerveken G, Govers F, Snel B. 2011. A domain-centric analysis of oomycete plant pathogen genomes reveals unique protein organization. *Plant Physiol.* 155:628–644.
- Seidl MF, Van den Ackerveken G, Govers F, Snel B. 2012. Reconstruction of oomycete genome evolution identifies differences in evolutionary trajectories leading to present-day large gene families. *Genome Biol Evol.* 4:199–211.
- Seidl MF, Wang R-P, Van den Ackerveken G, Govers F, Snel B. 2012. Bioinformatic Inference of Specific and General Transcription Factor Binding Sites in the Plant Pathogen *Phytophthora infestans*. *PLoS ONE.* 7:e51295.
- Shen D, Ye W, Dong S, Wang Y, Dou D. 2011. Characterization of intronic structures and alternative splicing in *Phytophthora sojae* by comparative analysis of expressed sequence tags and genomic sequences. *Can. J. Microbiol.* 57:84–90.
- Stajich JE et al. 2012. FungiDB: an integrated functional genomics database for fungi. *Nucleic Acids Res.* 40:D675–81.
- Stassen JHM et al. 2012. Effector identification in the lettuce downy mildew *Bremia lactucae* by massively parallel transcriptome sequencing. *Mol Plant Pathol.* 13(7):719–731.
- Stein L. 2001. Genome annotation: from sequence to biology. *Nat. Rev. Genet.* 2(7):493–503.
- Stein L, Sternberg P, Durbin R, Thierry-Mieg J, Spieth J. 2001. WormBase: network access to the genome and biology of *Caenorhabditis elegans*. *Nucleic Acids Res.* 29:82–86.
- Stergiopoulos I et al. 2012. *In silico* characterization and molecular evolutionary analysis of a novel super-family of fungal effector proteins. *Mol. Biol. Evol.* 29:3371–3384.
- Torto-Alalibo T et al. 2005. Expressed sequence tags from the oomycete fish pathogen *Saprolegnia parasitica* reveal putative virulence factors. *BMC Microbiol.* 5:46.
- Tyler BM et al. 2006. *Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science.* 313:1261–1266.
- van Noort V, Huynen MA. 2006. Combinatorial gene regulation in *Plasmodium falciparum*. *Trends Genet.* 22:73–78.
- Vilella AJ et al. 2009. EnsemblCompara GeneTrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Res.* 19:327–335.
- Vleeshouwers VGAA et al. 2011. Understanding and exploiting late blight resistance in the age of effectors. *Annu Rev Phytopathol.* 49:507–531.
- Ye W et al. 2011. Digital gene expression profiling of the *Phytophthora sojae* transcriptome. *Mol. Plant Microbe Interact.* 24:1530–1539.




# Summary

Diseases that destroy livestock and crops have had catastrophic effects on human civilization, causing starvation and economic losses and continue to be a constant threat to global food production. Pathogenic species occur throughout the tree of life including many bacteria, eukaryotic microorganisms and metazoans. The fascinating taxonomic class of oomycetes unites several important pathogenic eukaryotes of plants and animals. We have witnessed considerable advances characterizing the (molecular) biology of these species, in particular their interactions with the hosts, in the last decades. Nevertheless, our current knowledge still lags behind the possibilities provided by large-scale (gen)omics data which are nowadays easily accessible. Comparative genomics and integrative bioinformatics provide an important framework to accompany and complement classical experimental work to advance our knowledge of these species. This thesis describes four complementary comparative genomic studies focusing on different aspects of cellular function and evolution of pathogenic oomycetes.

The first two studies (chapter 2 and chapter 3) focused on the evolution of the gene content in pathogenic oomycetes. In chapter 2, we analyzed plant-pathogenic oomycetes using protein domains as a tool to assess the gene content of oomycetes and other eukaryotic pathogens, especially in the light of gene family expansions and novel gene fusions. We find nearly 250 expanded domains in oomycete and fungal plant pathogens of which a substantial part could be linked to a role in pathogenicity. Highly abundant domains with a role in signaling and regulation form a large repertoire of novel combinations. We hypothesize that the analyzed oomycetes might encode proteins that rewire existing signaling networks in a novel way that is distinct from other eukaryotes. In chapter 3, we used a comprehensive phylogenetic approach to study the evolution of the gene content by gene gains, duplications and losses thereby revealing distinct evolutionary paths that shaped the gene content of six extant oomycete pathogens. The branch leading to the plant pathogenic *Phytophthora* is characterized by frequent duplications that represent a major transition point in their evolution. The phylogenetic approach allows the characterization of explicit evolutionary trajectories, e.g. the pattern and timing of duplications and losses, thereby providing additional insights into the complex evolutionary history of oomycete genome evolution.

The following chapter (chapter 4) centered on the regulation of gene expression by identifying *cis*-regulatory DNA motifs in *Phytophthora infestans*, the causal agent of late blight on potato and tomato and one of the most prominent plant pathogens in the group of oomycetes. We conducted the first genome-wide survey for *cis*-regulatory elements in *P. infestans* or any other oomycete. To this end, we applied an *in silico* approach that combines gene co-expression and conservation to infer DNA motifs. We



identified several highly abundant DNA motifs with similarity to common eukaryotic promoter elements. By describing additional elements that occur upstream of a particular group of genes such as transcription factors, we significantly added to the current understanding of transcriptional regulation in oomycetes.

Very limited knowledge on associations or interactions between proteins is available in oomycetes. In chapter 5, we used complementary omics data to predict the first functional association network that links ~33% of the predicted proteome in *P. infestans*. These data allowed us to derive functional modules, i.e. sub-networks of co-expressed genes, involved in sporulation, a fundamental developmental stage in oomycetes. We identify known players such as CDC14, a phosphatase with a critical role in spore formation, and pinpoint novel candidates with as yet unknown roles in this important developmental process. We demonstrate that networks are pivotal frameworks to study other large-scale omics data such as microarrays by creating a concise list of associated candidates for subsequent experimental validation.

The analyses presented in this thesis highlight the merit of integrative and comparative genomics as a pivotal tool to explore the biology and evolution of oomycetes. They provide (testable) hypotheses on the evolution, biology and the function of as yet uncharacterized gene products in oomycetes, thereby significantly advancing our knowledge on this intriguing group of organisms.

# Samenvatting

Ziektes van vee en landbouwgewassen hebben in het verleden een vernietigende uitwerking gehad op menselijke beschavingen door het veroorzaken van hongersnoden en grote economische schade, en zijn nog steeds een bedreiging voor de wereldwijde voedselproductie. Pathogenen komen verspreid voor in de gehele 'tree of life' en zijn te vinden onder de bacteriën, de eukaryote micro-organismen en de metazoa. De oömyceten (waterschimmels) vormen een fascinerende taxonomische groep die belangrijke eukaryote pathogenen van dieren en planten verenigt. De afgelopen decennia is er aanzienlijke vooruitgang geboekt in het onderzoek aan oömyceten en in het bijzonder in het ontrafelen van de interactie van deze pathogenen met hun gastheer op moleculair and cellulair niveau. Desalniettemin is onze kennis nog beperkt. Het is daarom noodzaak om meer en beter gebruik te maken van de mogelijkheden die geboden worden door de grootschalige genomische datasets die tegenwoordig toegankelijk zijn. Vergelijkende genomische analyse en integratieve bioinformatica bieden een raamwerk dat klassieke experimentele methoden versterkt en aanvult, waardoor wij onze kennis van oömyceten kunnen vergroten. Dit proefschrift beschrijft vier complementaire vergelijkende genomische analyse studies, gericht op verschillende aspecten van moleculaire functies in pathogene oömyceten en op de evolutie van oömyceten.

De eerste twee studies (hoofdstukken 2 en 3) richten zich op de evolutie van het genenarsenaal in pathogene oömyceten, welke genfamilies een soort bevat en hoe de genensamestelling van soorten evolueert. In hoofdstuk 2 gebruiken we eiwitdomeinen als instrument om de geninhoud van plantpathogene oömyceten en andere eukaryote plantpathogenen te bestuderen, met nadruk op de rol van de expansie van genfamilies en nieuwe genfusies. We vinden in plantpathogene oömyceten en schimmels 250 domeinen die geëxpandeerd zijn, waarvan een aanzienlijk deel gekoppeld kan worden aan een rol in pathogeniciteit. Veelvoorkomende domeinen met een rol in signaaltransductie netwerken en regulatie vormen een groot repertoire aan nieuwe combinaties waardoor eiwitten ontstaan die niet eerder beschreven zijn. Onze hypothese is dat door deze nieuwe eiwitten de signaaltransductie in oömyceten deels via andere routes verloopt. In hoofdstuk 3 gebruiken we een alomvattende fylogenetische benadering voor het bestuderen van de evolutie van geninhoud door het verschijnen, dupliceren en verdwijnen van genen. Hiermee onthullen wij verschillende evolutionaire paden die vorm gegeven hebben aan de geninhoud van zes hedendaagse pathogene oömyceten. De tak die leidt tot het genus *Phytophthora* met enkel plantpathogene soorten, wordt gekarakteriseerd door veelvuldige duplicaties, symbool voor een belangrijke transitie in de evolutie van *Phytophthora*. Deze fylogenetische benadering maakt het mogelijk om expliciete evolutionaire paden in kaart te brengen, bijvoorbeeld het patroon van duplicaties en verlies van genen, waardoor meer inzicht verkregen wordt in de complexe

evolutionaire geschiedenis van de genoomevolutie van oömyceten.

Het volgende hoofdstuk (hoofdstuk 4) behandelt de regulatie van genexpressie door het identificeren van cis-regulatorische DNA motieven in *Phytophthora infestans*, de veroorzaker van aardappelziekte en één van de meest prominente plantpathogenen binnen de oömyceten. Dit is de eerste genoom-omvattende studie naar cis-regulatorische elementen in *P. infestans* en oömyceten in het algemeen. Hiervoor hebben we een in-silico benadering toegepast waarbij co-expressie en conservering van genen gebruikt worden voor het afleiden van DNA motieven. We hebben verschillende veelvuldig-aanwezige DNA motieven met gelijkenis met bekende eukaryote promotor elementen geïdentificeerd. Door het beschrijven van sequentie elementen die zich 'upstream' bevinden van genen die voor een bepaalde groep eiwitten coderen, zoals bijvoorbeeld transcriptiefactoren of elicitors van afweer in planten, hebben we een substantiële bijdrage geleverd aan ons begrip van transcriptie-regulatie in oömyceten.

Er is erg weinig bekend over de associaties en interacties tussen eiwitten in oömyceten. In hoofdstuk 5 gebruiken we complementaire 'omics' data voor de eerste voorspelling van functionele associaties voor ~33% van het voorspelde proteoom van *P. infestans*. Deze data stellen ons in staat om functionele modules af te leiden (sub-netwerken van genen die gezamenlijk tot expressie komen) betrokken bij sporulatie, een fundamentele ontwikkelingsfase in oömyceten. Hierbij vinden we bekende spelers zoals CDC14, een fosfatase met een cruciale rol in sporevorming, en identificeren we nieuwe kandidaten met tot nu toe onbekende functies in dit belangrijke ontwikkelingsproces. We laten zien dat netwerken een centraal raamwerk zijn voor het bestuderen van andere soorten grootschalige 'omics' data zoals microarrays, door het genereren van een beknopte lijst geassocieerde kandidaten voor experimentele validatie.

De analyses beschreven in dit proefschrift onderstrepen de bijdrage van integratieve bioinformatica en vergelijkende genoomanalyse als cruciale hulpmiddelen voor het verkennen van de biologie en evolutie van oömyceten. Zij leveren (testbare) hypothesen over de evolutie, biologie en functie van tot op heden ongekaracteriseerde genproducten in oömyceten, waardoor onze kennis van deze intrigerende groep micro-organismen aanmerkelijk vergroot wordt.

# Zusammenfassung

Von Schädlingen verursachte Krankheiten an Pflanzen und Tieren hatten in der Vergangenheit katastrophalen Einfluss auf die menschliche Zivilisation. Sie verursachten Hungersnöte und sind auch heute und bleiben auch in der Zukunft eine konstante Bedrohung für die globale Lebensmittelproduktion. Diese Krankheitserreger (pathogene Organismen) umfassen äußerst unterschiedliche Spezies, von einzelligen Bakterien und Eukaryoten bis zu mehrzelligen Eukaryoten wie z.B. Fadenwürmern. Die faszinierende phylogenetische Klasse der Eipilze (oder Oömyceten) beinhaltet einige der wichtigsten Krankheitserreger an Pflanzen und Tieren. Unser Wissen über ihre molekulare Biologie und insbesondere die Interaktion mit ihren Wirt hat sich in den letzten Jahrzehnten konstant weiterentwickelt. Dennoch ist unser jetziges Wissen begrenzt, doch neuartigen Technologien ermöglichen die einfache und schnelle Generierung von einer Vielzahl an biologischen Daten und Informationen. Daher können klassische molekulare Experimente mittels vergleichender Genomanalyse und der Integration dieser verschiedenartigen Daten, zwei wichtige Konzepte in der modernen Biologie, ergänzt und unterstützt werden. Die vorliegende Arbeit beschreibt vier komplementäre Studien, die sich diese Konzepte zu Eigen machen, um diverse Aspekte der molekularen Biologie und der Evolution von Oömyceten im Detail zu untersuchen.

In Kapitel 2 und Kapitel 3 wurden verschiedene Aspekte der Genomevolution von Oömyceten untersucht. In Kapitel 2 benutzten wir Proteindomänen als Hilfsmittel, um die Expansion von Genfamilien und spezifische Domänenkombinationen in pflanzenpathogenen Oömyceten zu erforschen. Wir fanden fast 250 Proteindomänen, die in Oömyceten, aber auch anderen pathogenen Pilzen, expandiert sind. Vielen dieser Domänen konnte eine mögliche Rolle in der Pathogenität dieser Organismen zugewiesen werden. Häufig vorkommende Domänen mit Funktionen in Signaltransduktionswegen oder bei der Regulation von zellulären Prozessen bilden eine Vielzahl an neuartigen und spezifischen Kombinationen in Oömyceten. Daher scheinen die Genome der Oömyceten eine Reihe von Proteinen zu kodieren, die bekannte Signalwege in einer neuartigen Art und Weise miteinander verbinden. Kapitel 3 befasst sich mit der Evolution aller Genfamilien in sechs verschiedenen Oömyceten. Wir benutzten einen phylogenetischen Ansatz, um die einzelnen evolutionären Pfade, das heißt die Abfolge von Geninnovationen, Duplikationen und Verlusten in diesen Familien, nachzuvollziehen. Der Zweig, der zu den pflanzenpathogenen *Phytophthora* führt, ist gekennzeichnet durch eine hohe Anzahl an Duplikationen und ist daher ein wichtiger Schritt in der Evolution dieser Spezies. Da unsere phylogenetische Analyse die genaue zeitliche Abfolge der evolutionären Abläufe dokumentiert, bietet sie daher eine dynamische Übersicht der evolutionären Geschichte der Oömyceten, die mit alternativen Methoden so nicht möglich gewesen wäre.

Das nachfolgende Kapitel (Kapitel 4) beschäftigt sich mit der Regulation von Gene-expression in *Phytophthora infestans*, einem der bedeutendsten Krankheitserreger in der Gruppe der Oömyceten und verantwortlich für die Kraut und Knollenfäule bei Kartoffeln. Wir beschreiben die erste genomweite Identifizierung und Untersuchung von *cis*-regulierenden DNA Elementen in *P. infestans*. Wir benutzten *in silico* Methoden, die auf Geneexpressionsdaten und Sequenzkonservierung zwischen verwandten Spezies basieren. Wir ermittelten mehrere DNA Elemente die eine hohe Ähnlichkeit zu bekannten eukaryotischen Elementen aufweisen. Durch die Identifikation von weiteren DNA Elementen, insbesondere von solchen, die sich in der Nähe von spezifischen Gruppen von Genen, wie zum Beispiel Transkriptionsfaktoren, befinden, haben wir entscheidend zum aktuellen Verständnis der Generegulation in Oömyceten beigetragen.

Das Wissen über die Interaktionen zwischen Proteinen in Oömyceten ist sehr eingeschränkt. In Kapitel 5 verwendeten wir verschiedene sich komplementierende omics-Daten, um das erste Protein-Protein Interaktion Netzwerk in *P. infestans* vorherzusagen. Dieses Netzwerk beschreibt die Interaktionen und funktionelle Verbindungen zwischen 33% aller vorhergesagten Proteine. Mit Hilfe dieser Daten können wir funktionelle Module, eine Gruppe interagierender Proteine deren kodierenden Gene co-exprimiert sind, vorhersagen. Wir untersuchten im Speziellen die Sporulation, einen wichtigen Entwicklungsprozess im Lebenszyklus von *P. infestans*. Dieses funktionelle Modul beinhaltet neben bekannten Proteinen wie CDC14 auch uncharakterisierte Proteine die nun diesem wichtigen Entwicklungsprozess zugeordnet werden konnten. Protein-Protein Netzwerke beschreiben daher essentielle Informationen, die die Interpretation von anderen, oft komplementären, Daten zugänglich machen können, da sie Proteine in ihren zellulären Kontext stellen.

Diese Arbeit hebt die Vorteile von vergleichender Genomanalyse als bedeutendes Werkzeug zum Verständnis der Biologie und Evolution hervor: Wir stellten experimentell verifizierbare Hypothesen zur Evolution, Biologie und Funktion von bisher unbeschriebenen Genen und ihren Produkten in Oömyceten auf. Damit haben wir signifikant zum Verständnis dieser interessanten Gruppe von Spezies beigetragen.

# Curriculum Vitae

Michael Franziskus Seidl was born on the 13th of October 1982 in Darmstadt, Germany. He finished secondary school at the Ulrich-von-Hutten-Gymnasium, Schlüchtern, in 2002. After a year of military service he continued his education at the Technische Universität Darmstadt, Germany, where he obtained his pre-degree in biology. He continued his biology study in Würzburg, Germany, where he focused on bioinformatics, genetics and biochemistry. As part of his studies, he conducted a research internship studying the distribution of single nucleotide polymorphisms in human transcription factors at the Wellcome Trust Centre for Human Genetics, a research institute of the University of Oxford, United Kingdom, under the supervision of dr. Richard R. Copley. After returning to Würzburg he wrote his diploma thesis in the group of prof. dr. Jörg Schultz with the title 'Simplification of molecular machines in model organisms'. In September 2008 he received his diploma (cum laude) in biology from the Julius-Maximilians-Universität Würzburg. Subsequently, he started his PhD research project in the Theoretical Biology & Bioinformatics group at Utrecht University, the Netherlands, in the lab of dr. Berend Snel. His research project was part of a research programme coordinated by the Centre for BioSystems Genomics (CBSG)/ Netherlands Genomics Initiative (NGI) and focused on comparative genomics of oomycete pathogens. He was supervised by dr. Berend Snel, prof. dr. Paulien Hogeweg, prof. dr. Francine Govers, Laboratory of Phytopathology, Wageningen University, The Netherlands, and dr. Guido van den Ackerveken, Plant-Microbe Interactions Group, Utrecht University. In June 2013 he will join the group of dr. Bart Thomma, Laboratory of Phytopathology, Wageningen University, as a research fellow (PostDoc). The results of his PhD research are described in this thesis.





# List of Publications

Michael F Seidl, Adrian Schneider, Francine Govers and Berend Snel  
**A Predicted Functional Gene Network for the Plant Pathogen *Phytophthora infestans* as a Framework for Genomic Biology**  
Submitted

Adrian Schneider, Michael F Seidl and Berend Snel  
**Shared Protein Complex Subunits Contribute to Explaining Disrupted Co-occurrence**  
Submitted

Rays HY Jiang, Irene de Bruijn, Brian J Haas, Rodrigo Belmonte, Lars Löbach, James Christie, Guido Van den Ackerveken, Arnaud Bottin, Bernard Dumas, Lin Fan, Elodie Gaulin, Francine Govers, Laura J Grenville-Briggs, Neil R Horner, Joshua Z Levin, Marco Mammella, Harold JG Meijer, Paul Morris, Chad Nusbaum, Stan Oome, David van Rooyen, Marcia Saraiva, Chris J Secombes, Michael F Seidl, Berend Snel, Joost HM Stassen, Sean Sykes, Sucheta Tripathy, Herbert van den Berg, Julio C Vega-Arreguin, Stephan Wawra, Sarah K Young, Qiandong Zeng, Javier Dieguez-Uribeondo, Carsten Russ, Brett M Tyler and Pieter van West  
**Distinctive Expansion of Potential Virulence Genes in the Genome of the Oomycete Fish Pathogen *Saprolegnia parasitica***  
*PLoS Genetics* (2013; in press)

Michael F Seidl, Rui-Peng Wang, Guido Van den Ackerveken, Francine Govers and Berend Snel  
**Bioinformatic Inference of Specific and General Transcription Factor Binding Sites in the Plant Pathogen *Phytophthora infestans***  
*PLoS ONE* 7(12):e51295 (2012)

Michael F Seidl, Guido Van den Ackerveken, Francine Govers and Berend Snel  
**Reconstruction of Oomycete Genome Evolution Identifies Differences in Evolutionary Trajectories Leading to Present-day Large Gene Families**  
*Genome Biology and Evolution* 4(3):199-211 (2012)

Joost HM Stassen, Michael F Seidl, Pim WJ Vergeer, Isaac J Nijman, Berend Snel, Edwin Cuppen and Guido Van den Ackerveken  
**Effector Identification in the Lettuce Downy Mildew *Bremia lactucae* by Massively Parallel Transcriptome Sequencing**  
*Molecular Plant Pathology* 13(7):719-31 (2011)

Adriana Cabral, Joost HM Stassen, [Michael F Seidl](#), Jaqueline Bautor, Jane E Parker and Guido Van den Ackerveken

**Identification of *Hyaloperonospora arabidopsidis* Transcript Sequences Expressed during Infection Reveals Isolate-specific Effectors**

*PLoS ONE* **6**(5):e19328 (2011)

[Michael F Seidl](#), Guido Van den Ackerveken, Francine Govers and Berend Snel

**A Domain-Centric Analysis of Oomycete Plant Pathogen Genomes Reveals Unique Protein Organization**

*Plant Physiology* **155**:628-644 (2011)

[Michael F Seidl](#) and Jörg Schultz

**Evolutionary Flexibility of Protein Complexes**

*BMC Evolutionary Biology* **9**:155 (2009)

# Acknowledgments

After spending a considerable period of time at the Theoretical Biology and Bioinformatics group I would like to humbly express my thankfulness to several people. They did not only provide guidance and support with the work partially described in this thesis; some have transformed from colleagues to friends. Even though I cannot address each and everyone, I will always remember you and I am thankful for our shared time!


First of all I would like to thank my supervisor Berend Snel. Your knowledge of science is staggering. I can remember several meetings when I left your office astonished by the wealth of input. I think you are one of the rare people that manage to have a brilliant work-life balance, being both a successful group leader and a great father. I found our discussions and thoughts about work, life and future insightful and enjoyable. I appreciated your encouraging words to my regular doubts; I am truly sorry for my occasional bluntness. I believe that you are one of the best supervisors I could imagine and your guidance made me to the scientist I am now.

Next, I would like to thank my promoters Francine Govers and Paulien Hogeweg for their guidance. Paulien, it has been an unimaginable privilege to work in the office next to yours and experience your knowledge and view of science during lunch and coffee breaks or group talks. Francine, working and discussing science with you has been a great pleasure. Your dedication (emails at one o'clock at night), your support and your insightful advice on my manuscripts made a difference in so many cases and were highly appreciated.

Guido van den Ackerveken was the fourth pillar of support during my PhD. Your positive attitude towards science and your clear view on problems and their solution were always an important guidance. I enjoyed our successful collaborations and the nice chats with you, including the discussions over glasses of wine in Toulouse.

Every group reflects the dedication of its leaders. Rob, Paulien, Kirsten, Can and Berend - I am very thankful for the time that I could share in this wonderful group. I hope that you together will make the right decisions to guide this group in a successful future!

Our daily work would be unthinkable without the amazing IT support of Jan Kees. You give us the impression that our computers always 'just' work. Moreover, you always had time and patience for any of my (stupid) questions - you even managed to answer them and provide me with valuable solutions. Without your work, my thesis, and most likely the work of most, if not even all of us, would not have been possible!



Already when I visited the group the first time to give my job interview talk I met two people that immediately made a huge impression on me. They welcomed me with remarkable friendliness and warmth. From colleagues they became very good friends and my paranymphs: Like and Jos. Like, your unconventional view of the world, your entertaining stories on daily life situations (the dog and the soup spoon) made my days so many times. During your PhD you gave birth to two wonderful kids and only because of this I can say that I will be first. Jos, you are one of the smartest people I have ever met. Your view of science changed my way of solving problems. Even though you are not a colleague anymore, I enjoyed our regular dinner/beer appointments; sitting in a pub, discussing work, Dutch and German politics, life and the waitresses. Thanks to both of you for making my stay in Utrecht unforgettable!

John, you were the first bird to leave the warm and safe nest to conquer the scientific world in Nijmegen. You were an important member of the 'Snel trein' to the coffee machine and an entertaining roommate with a profound knowledge of Perl. Gabino – the self-titled super postdoc – even after moving your household twice, I still enjoyed our squash and running competitions. Adrian, watching soccer with you and listening to your phone conversations in Schwyzerdütsch were very amusing experiences. Eelco, I appreciate our squash competitions and your 'bad hair days' still make me smile whenever I see a guy with a hat. Alessia, gaming nights and dinners at your place were a lot of fun.

Hanneke, you are without a doubt the soul of TBB's social life. Your initiatives for presents and group activities and especially the awesome outcomes are unforgettable. You have always been a friend and I hope you will dance into a bright future, whether it is in Hawaii or in the rainy Netherlands! Paola, you always have a positive attitude towards life and a kind word for me whenever things did not go as planned. I will always remember your 'be there' parties, the dinners at your place and your nice company. Ilka, unfortunately you left us quite a while ago. I really miss the concerts we visited, our dinners and just having you around. I would also like to thank Chris, Folkert, Ioana, Johannes, Jorg, Julia, Klaartje, Sai, Sandro, Renske and Thomas for the regular coffee excursions to Gutenberg and the fun time at the 'be there' parties.

Science rarely represents the work of a single person but it is instead a combined endeavor. I had the pleasure to work with a few friendly, smart and driven colleagues in Wageningen and in Utrecht. Harold, I appreciated your insights into *Phytophthora* genomics and I apologize for keeping you awake in Asilomar the night before my first talk at the OMGN. Joost, you were a great travel company and a knowledgeable colleague; our trips to Toulouse, Asilomar and Virginia Tech were memorable and I enjoyed chatting and complaining about our fellow colleagues (you know who I am talking about). I also would like to thank Chiel and Stan for our collaborations. Nijuscha, you made our office to a bright and friendly place. Thank you for highlighting the hidden surprises of Berlin to me.

In the last four years I had the pleasure to supervise five students (Rui, Brand, Esther,



Ethel and Jolien). I hope that I was able to show to you how exciting and sometimes frustrating science can be. I wish you good luck and all the best in your future.

I want to especially thank my parents. Your affection and unconditional support for whatever decision I made in life is beyond words. During my day to day life, my university studies and ultimately during my PhD studies, you always had an open ear for my problems and questions, gave guidance and supported me in every imaginable way: Ich bin Euch so dankbar!

In acknowledgments, the most important people are mentioned at the end, even though they are first in heart and memories. Lidija, you are without any doubt the most important person in my life. Upon your arrival in Berend's group you changed my life and transformed from an incredibly smart, dedicated colleague, to an important friend and ultimately to my wonderful girlfriend. I wish I could be as supportive and forgiving as you were during the last months of my PhD. I enjoy our shared life so much: Lidija, ljubim te!



