8-2008

A proteomic investigation of *Phytophthora* species using mass spectrometry and reverse genetics

Alon Savidor
University of Tennessee - Knoxville

To the Graduate Council:

I am submitting herewith a dissertation written by Alon Savidor entitled "A proteomic investigation of *Phytophthora* species using mass spectrometry and reverse genetics." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

Kurt H. Lamour, W. Hayes McDonald, Major Professor

We have read this dissertation and recommend its acceptance:

Jeffrey Becker, Loren Hauser, Neal Stewart

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

To the Graduate Council:

I am submitting herewith a dissertation written by Alon Savidor entitled “A proteomic investigation of *Phytophthora* species using mass spectrometry and reverse genetics.” I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Life Sciences.

_____
Kurt H. Lamour, Major Professor

_____
W. Hayes McDonald, Major Professor

We have read this dissertation
and recommended its acceptance:

_____
Jeffrey Becker

_____
Loren Hauser

_____
Neal Stewart

Accepted for the Council:

_____
Carolyn R. Hodges, Vice Provost and
Dean of the Graduate School

Original signatures are on file with official student records.

A proteomic investigation of *Phytophthora* species using mass spectrometry and reverse genetics

A Dissertation Presented for the Doctor of Philosophy Degree
The University of Tennessee, Knoxville

Alon Savidor
Aug. 2008

Acknowledgements

I would like to thank all the people who assisted me in the completion of the research presented in this dissertation. First, I would like to thank my graduate advisors Dr. Kurt H. Lamour and Dr. W. Hayes McDonald for their guidance in my doctoral studies and for their efforts in facilitating this collaborative project. I would like to thank my committee members Dr. Jeffrey Becker, Dr. Loren Hauser, and Dr. Neal Stewart for their support, guidance, and involvement in my research project.

I would like to thank my fellow graduate students Ryan Donahoo and Oscar Hurtado-Gonzales for teaching me laboratory work with *Phytophthora* as well as other molecular techniques. I would like to thank Ledare Finley for her guidance in the lab and for sharing her extensive technical and molecular biology knowledge.

I would like to thank everyone in the Organic and Biological Mass Spectrometry group at the Oak Ridge National Laboratory (ORNL), including Dr. Gregory B. Hurst for his technical assistance, and Becky R. Maggard for her secretarial assistance. I would like to thank Miriam Land for bioinformatics resources. I would like to thank Dr. David L. Tabb for bioinformatic software used in this study including DBDigger, DTASelect, and SQTRevPuller.

I would like to thank all of our collaborators for their contribution to this study. I would especially like to thank the research group at the Virginia Bioinformatics Institute, including Dr. Brett M. Tyler for facilitating training and collaborative research, and Dr. Sucheta Tripathi for bioinformatic resources. I would like to thank Dr. Sophien Kamoun and Liliana Cano (The Sainsbury Laboratory, UK) for *P. infestans* samples.

Finally, I would like to acknowledge support from the University of Tennessee (Knoxville)-ORNL Graduate School of Genome Science and Technology, and from the Chemical Division at ORNL. Much of the funding for this research came from the United States Department of Energy's Office of Biological and Environmental Sciences through the Genomics: GTL Center for Molecular and Cellular Systems and the Genomics:GTL grant. Funding was also provided by the ORNL Laboratory Director's Research and Development Program. Without these funding sources, the research presented here would not have been possible.

Abstract

Organisms in the genus *Phytophthora* are important plant pathogens, although understudied. *Phytophthora* was first brought into human awareness with the identification of *P. infestans* as the culprit for the Irish potato famine in the mid 1800s. Since then, over 80 *Phytophthora* species have been identified, many of which infect a wide variety of crops worldwide with devastating results.

Traditionally, much of the work aimed at controlling *Phytophthora* diseases involved applied research. In recent years there has been a marked increase in molecular work on *Phytophthora*. This increase is evident not only from increased funding by agencies such as the National Science Foundation (NSF), but also from the type of research applied to *Phytophthora* for the first time. The first *Phytophthora* species to have their genomes sequenced were *P. sojae* and *P. ramorum* at 2004. Since then the genomes of two more *Phytophthora* species- *P. capsici* and *P. infestans*, were also sequenced.

Availability of *Phytophthora* genome sequences provided us with the basis necessary for a proteomic investigation of these organisms. The study presented here represents the first large scale proteomic study of any *Phytophthora* species. Using mass spectrometry and available or newly developed bioinformatic tools we measured the proteomes of different asexual *Phytophthora* life stages. We also measured the protein complement of *P. capsici* infected tomato plants, the so called "interactome", in order to gain an insight into the biological processes occurring in the pathogen during infection, and in the plant in response to the pathogen. We also used data from these proteomic experiments as a part of a novel approach aimed at improving the genome annotation of

those *Phytophthora* species. Finally, we used different molecular techniques, including a reverse genetic technique called Targeted Induced Local Lesions in Genome (TILLING), to begin characterization of a few protein targets identified in those experiments.

The accumulated data from all our experiments identified certain molecular processes, metabolic and others, that may explain the success of *Phytophthora* as a plant pathogen. The data from these experiments provides a platform on which future experiments can be based on to further characterize these interesting organisms.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Phytophthora

The Latin name "*Phytophthora*" directly translates to "plant destroyer". As their name suggests, *Phytophthora* species are a group of destructive plant pathogens capable of impacting a large variety of plants. The first *Phytophthora* species to be described was *P. infestans*, the causal agent of the late blight disease of potato and tomato. During 1845 and 1846 *P. infestans* totally destroyed Ireland's staple potato crop on which the lower socio-economic class had come to depend[1]. The later known "Irish potato famine" led to mass starvation and poverty, which subsequently lead to massive immigration of people to the United States and elsewhere[1]. This was arguably the single most devastating plant disease responsible for widespread human suffering and sociological impact[1].

While *P. infestans* continues to cause damages to potato and tomato crops, other *Phytophthora* species are no less important. Since the late 1800s over 80 species of *Phytophthora* have been described, many of which cause heavy damage to agriculturally and ornamentally important crops, and can even present an ecological threat. *Phytophthora cinnamomi* devastated 250,000 acres of jarrah (*Eucalyptus marginata*) tree forests in Western Australia and contributed to the demise of the American chestnut (*Castanea dentata*) in the southeastern U.S. in the previous century[2, 3]. *Phytophthora ramorum*, one of the major study subjects of this work, is the causal agent of the sudden oak death disease—a disease that is currently devastating oak forests in California and Oregon, and is feared to have destructive effects if spread to the Smoky Mountain region

here in East Tennessee. Other *Phytophthora* species cause agricultural damages estimated in the billions of dollars annually world wide, impacting diverse cash crops such as soybean, peppers, cucumbers, apples and cacao. Different *Phytophthora* species have very different hosts, both in terms of the plant species they are able to infect, and in the range of hosts. Some *Phytophthora* species such as *P. cinnamomi*, *P. ramorum*, and *P. cactorum*, are able to infect hundreds of different plant species, while others, such as *P. sojae* and *P. fragariae* var. *oryzobladis* are specific to a single host. The molecular mechanisms responsible for the different host ranges of different *Phytophthora* species are not understood, and this was one of the questions this study aimed at probing.

Phytophthora species are capable of reproducing sexually and asexually. Some *Phytophthora* species are heterothallic, meaning that in order for sexual spores (oospores) to be produced, both mating types, A1 and A2, of the same species must grow together. Other *Phytophthora* species are homothallic, where oospores can be produced when only a single mating type is present in the culture. Oospores have thick cell walls and can withstand dry conditions much better than asexual spores. Oospores can survive in the soil for long periods of time and re-infect their host plant in subsequent growing seasons. While the oospores are important for long term survival and generation of diversity through sexual recombination, it is the asexual life cycle that is responsible for rapid propagation and spread of the disease in the field or forest due to its relatively short completion time and the high number of propagules produced. Through the asexual life cycle, the organism is able to differentiate into different life stages including mycelium, sporangium, zoospore, cyst, and germinating cyst (Figure. 1.1).

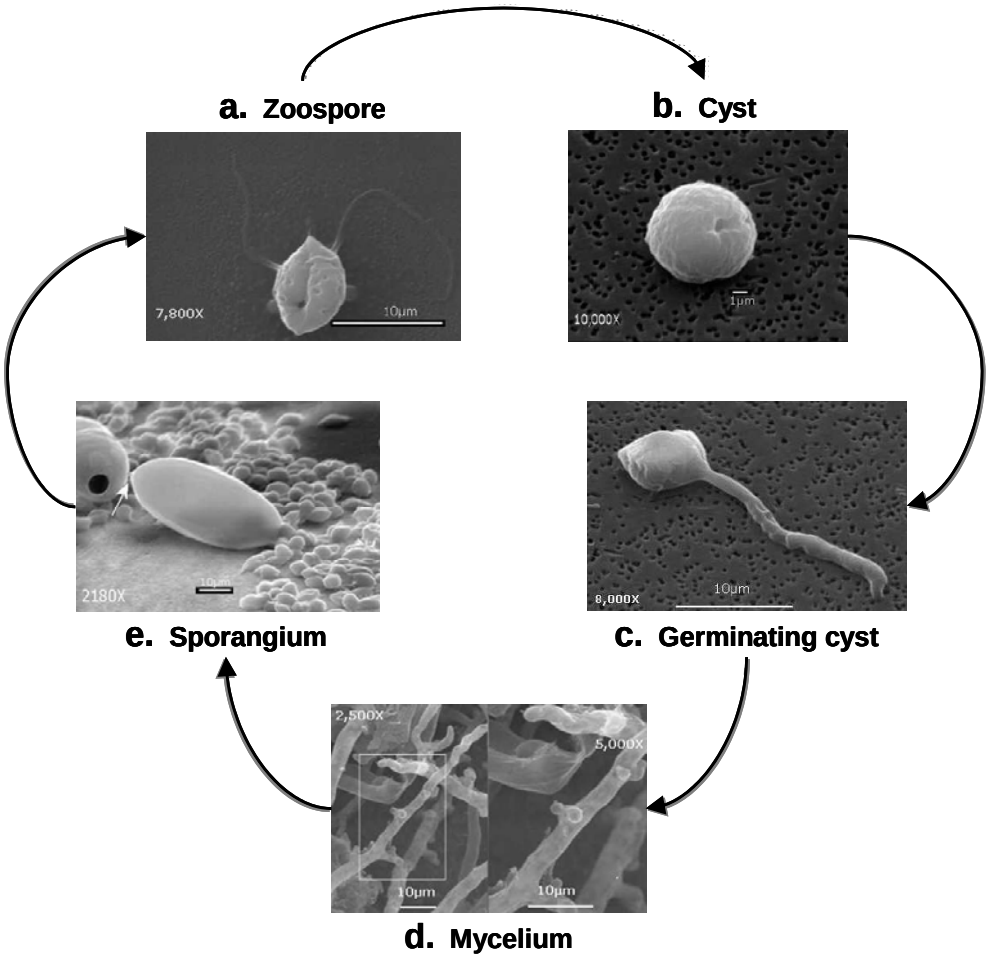


Figure 1.1- The asexual life cycle of *Phytophthora*. Scanning electron micrographs of asexual life stages of *P. ramorum*. Motile bi-flagellated zoospore **(a)** swim chemotactically towards its plant host. Upon contact with the plant (or shaking in the lab) the zoospore sheds its flagella and encysts **(b)**. The cyst germinates **(c)**, and the germ tube penetrates into the plant tissue where it starts growing vegetatively through the plant tissue as mycelium **(d)**. The mycelium can grow out of the plant tissue where, under appropriate stimulation, its terminal ends can differentiate into sporangia **(e)**, which are structures containing multiple nuclei which can differentiate into zoospores. Zoospores can then be released from the sporangia and repeat the cycle. Micrographs were supplied courtesy of Dr. Edwin R. Florance, Lewis & Clark College, Portland, OR.

Infection of a plant can be initiated when a zoospore, a motile kidney shaped bi-flagellated cell, interacts with a compatible plant host. Upon contact with the plant host the zoospore sheds its flagella, encysts, and adheres. Shortly after encystment the cyst germinates by producing a germ tube that penetrates the host tissue directly or through wounds or natural openings. Once inside the plant, the pathogen grows and ramifies through the plant tissue as mycelium, the vegetative growth life stage of *Phytophthora*. In a compatible interaction the initial growth of the mycelium in the plant tissue is biotrophic, where the pathogen evades the plant defense responses. Later, growth is switched to necrotrophic and the plant tissue is destroyed. The mycelium can also emerge and grow outside of the plant tissue. When exposed to the new environment outside of the plant, the terminal ends of the mycelium can differentiate into sporangia, vessel like structures that contain multiple nuclei in their cytoplasm. The nuclei in the sporangia can differentiate into individual zoospores which can then be released from the sporangium and continue the asexual cycle.

The genus *Phytophthora* is included in the Oomycete phylum. Historically the Oomycetes were classified as fungi, as their infection typically resembles fungal infection morphologically. In addition, the Oomycetes produce similar life stages to those of true fungi. However, major differences between the Oomycetes and true fungi have been identified. Morphologically, in contrast to the mycelia of fungi, the Oomycetes lack septa in the mycelium. Additionally, unlike fungi, the motile *Phytophthora* zoospores are bi-flagellated with one whiplash and one tinsel flagellum. Beyond the morphological differences between Oomycetes and fungi, many biochemical and molecular differences have been identified over the years. While most fungi are haploid through most of their

life cycle, the Oomycetes are diploid for most of their life cycle. The major component of true fungi cell wall is chitin, whereas the oomycete cell was is mainly composed of cellulose and β-glucans[4]. In addition, unlike fungi, *Phytophthora* do not synthesize their own sterols, but require an exogenous sterol source for sporulation[5]. With the advance in sequencing techniques and phylogenetic analysis tools in the last two decades, a clearer picture of the evolutionary origin of *Phytophthora* began to emerge. Today, it is widely accepted that the Oomycetes are completely distinct from true fungi. They form a unique branch on the evolutionary tree, and are actually more closely related to brown algae and diatoms than to fungi[6, 7](Figure 1.2).

The phylogenetic uniqueness of the Oomycetes has both practical and applied implications. One such implication is control of the disease in the field or forest. Since the Oomycetes and true fungi are so different, most fungicides and antibiotics are ineffective against *Phytophthora* infections. This is particularly true for polygene antibiotics such as pimaricin that target the sterol synthesis pathway. Fungal growth is inhibited by such antibiotics, while the Oomycetes are uniquely resistant due to the lack of dependence on endogenous sterol synthesis[8]. As a consequence, only few chemical treatments for the control of *Phytophthora* diseases are effective and economically feasible.

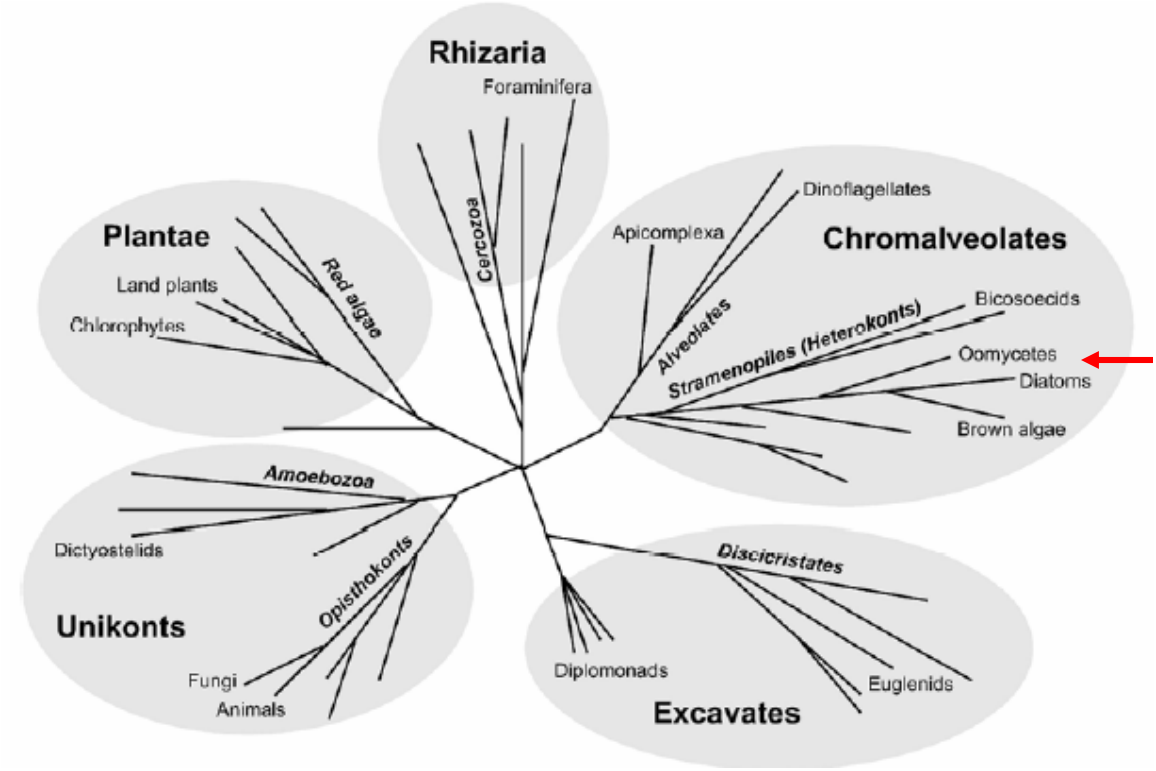


Figure 1.2- A schematic phylogeny of eukaryotes. This hypothetical tree was drawn according to recent reviews by Baldauf (2003) and Keeling and associates (2005). The five proposed eukaryotic supergroups are shaded and labeled in bold font. For simplicity, other labels correspond to selected examples; major subbranches are shown in italics, and representative groups of organisms are provided at the branch termini. Plantae, Rhizaria, Chromalveolates, and Excavates represent the bikont lineages. Red arrow points to the Oomycetes. Figure taken from Govers et al. (2006)[6].

Another practical challenge attributed to the evolutionary uniqueness of the Oomycetes is that their genomes contain many genes with non-typical sequences (e.g. infrequent hexamer repeats, splice sites, etc) and unknown function. Homology searches to other species are often ineffective since homologous genes are not found in other organisms. Thus, there is great room for improvement in gene calling and functional annotations. Another goal of this study was to assist in the genome annotation of *Phytophthora* species.

While a fair amount of genetic work has accumulated in the literature, only a limited number of large scale studies have been done on *Phytophthora*. Several transcriptomics studies have been carried out, typically involving measurements of expressed sequence tags (ESTs) and construction of cDNA libraries[9-13]. Moy et al. constructed a microarray chip containing cDNA transcripts from both *P. sojae* and its soybean host, and measured transcript level after infection[14]. Proteomic efforts were limited to a few studies involving 2-Dimentional polyacrylamide electrophoretic gel separations (2D-PAGE), typically followed by Matrix assisted laser disorption ionization (MALDI) and Time of Flight (TOF) mass spectrometry[15-18]. In 2004 the first genomes of the two *Phytophthora* species, *P. sojae* and *P. ramorum*, were sequenced and annotated [7]. In 2007, the draft genomes of *P. capsici* and *P. infestans* became available, although assembly and annotation of these genomes are still in progress. Availability of the complete genome sequences of these organisms paved the way to use these organisms in shotgun proteomics experiments that were previously impossible to perform. This study represents the first large scale proteomics study on these sequenced *Phytophthora* species.

Proteomics

The "proteome" refers to the entire expressed set of proteins in a given biological sample such as a cell culture or a tissue. The objective of the field of proteomics is to sample as completely as possible the proteome to be studied. This is no easy task considering that the complexity of even the simplest of bacteria, the simplest autonomous life form, is enormous, containing thousands of different species of proteins which range widely in abundance, structure, and chemical properties. Over the last two decades, mass spectrometry has become the main tool for proteomics investigation. To allow efficient proteomic investigation, improvement in all parts of the process was required: separation, MS analysis, and bioinformatic analysis of the results.

For a quality measurement of individual proteins, efficient separation of the proteins or their peptides in more than one dimension is required. The first step allowing the feasibility of analyzing an entire proteome was the development of two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) in the mid 1970s[19]. In 2D-PAGE, proteins are separated based on their isoelectric point in the first dimension, and then based on their molecular weight in the second dimension. This technique allows separating and visualizing hundreds and even thousands of proteins on a single gel. Gel spots can then be excised, treated, and analyzed to reveal the identity of the protein in the spot. While 2D-PAGE is widely used, it has several drawbacks. While many spots can be visualized on a gel, only a fraction of them can be identified using mass spectrometry or other techniques due to sample loss during treatment of the excised gel spot and co-migration of different proteins complicating interpretation and also interfering with

quantitation. As a result, only tens to low hundreds of proteins are typically identified in a single 2D-PAGE experiment. Also, due to the dimensions of separations, 2D-PAGE is biased against proteins with extreme pH, molecular weight, or low abundance. Finally, excision, treatment, and analysis of each individual gel spot is time consuming and resource intensive.

Another milestone improvement in the analysis of whole proteomes was the development of 2-dimensional liquid chromatography (LC/LC) that could be directly interfaced with a mass spectrometer. Multidimensional protein identification technology (MudPIT), published in 2001[20], is one strategy that utilizes automated LC/LC, and became a popular tool for proteomic investigation. MudPIT falls under the category of "shotgun" or "bottom-up" proteomics. Instead of separating intact proteins as in 2D-PAGE, shotgun proteomic approaches rely on peptides generated by protein digestion with proteases or chemicals. A typical proteome contains thousands of different proteins. Digestion of a proteome results in a peptide mixture which is at least an order of magnitude more complex than the original protein mixture. This seems counter-intuitive, but in fact, peptides are much easier to separate by LC/LC and analyze using mass spectrometry than intact proteins. A MudPIT experiment starts with digestion of a proteome with a protease, typically trypsin, followed by an offline loading of the peptide mixture onto a bi-phasic fused silica column containing a reverse phase (RP) resin followed by a strong cation exchange resin (SCX)[20]. After loading, the column is placed on the back end of a front column loaded with RP, online with the mass spectrometer. Separation of the peptides occurs on the SCX in the first dimension and the front column's RP in the second dimension. The additional RP on the back column is used for

cleaning and desalting purposes only. As peptides elute off the front column, they enter directly into the mass spectrometer. They are then subjected to tandem mass spectrometry where they are isolated, fragmented, and the resulting peptide fragments measured. Finally, the accumulated spectra of fragmented peptides are searched against a given protein database to identify the peptides giving rise to them, and therefore the proteins those peptides originated from. A diagram of the entire MudPIT process is shown in figure 1.3. The use of MudPIT has several advantages over 2D-PAGE, including better sensitivity and dynamic range, less bias, higher throughput, and better quantitation capabilities. All of the mass spectrometry experiments performed in this study utilized the MudPIT approach.

Specific aims

This study was initiated with 3 specific aims:

1) Contribute to the annotation of the newly sequenced *Phytophthora* genomes.
2) Identify candidate *Phytophthora* proteins that are involved in early infection.
3) Identify candidate *Phytophthora* proteins that are involved in host range.

Chapter 2 describes the development of the "Expressed Peptide Tags (EPTs)"—a strategy that uses mass spectrometry to increase confidence in gene calling of newly sequenced genomes. This effort was directed at specific aim 1, and can be applied to any other eukaryotic or prokaryotic organism. Indeed, EPT mapping was applied to *P. capsici* and *P. infestans* as their genome sequences became available, and might be applied to the bacterium *Azospirillum brasilense* in the near future. Chapters 3-5 are aimed at probing

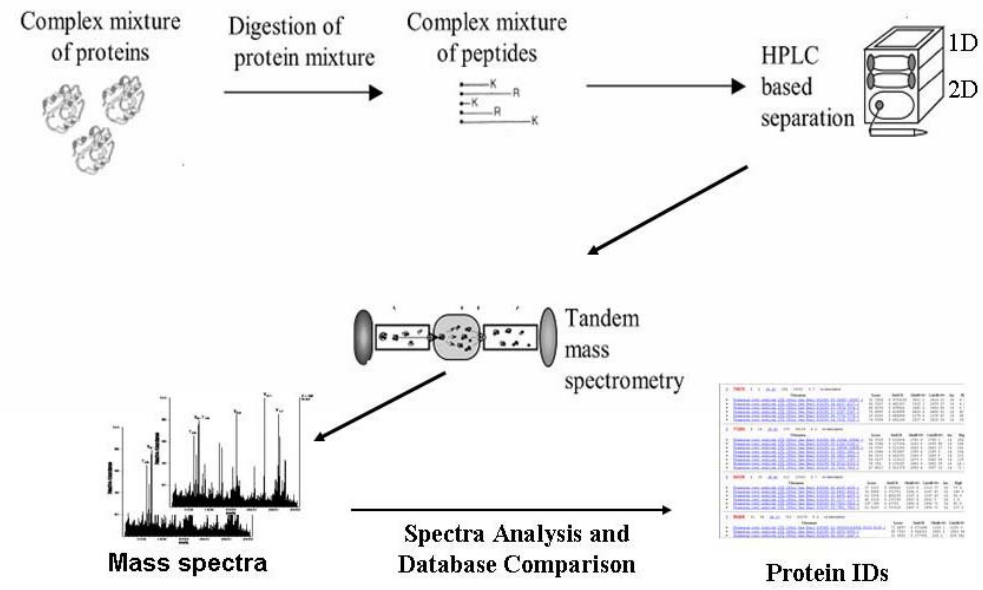


Figure 1.3- Schematic representation of Multidimensional Protein Identification Technology (MudPIT). The objective of MudPIT is to sample and identify as many proteins as possible from the original sample. A complex starting mixture of proteins, typically a proteome, is digested with trypsin which cleaves the proteins into peptides after cysteines and arginines. The resulting peptide mixture is then separated in 2 dimensions by high performance liquid chromatography (HPLC). Eluting peptides enter directly into the mass spectrometer, where they are isolated and fragmented. The peptide fragments are then measured producing spectra that are then computationally searched against a given protein database. Identified peptides lead to identification of the proteins they originated from.

both the host range and temporal expression pattern questions posed by specific aims 2 and 3. In chapter 3 we took advantage of the unique opportunity of having the sequenced genomes of two closely related eukaryotic organisms to develop a strategy for identifying targets for either host range or early/late infection stages of *P. sojae* and *P. ramorum*. Chapter 4 describes the proteomic investigation of *P. capsici*, sequenced in 2007, using some of the tools developed in chapter 3. While chapters 3 and 4 describe the study of different life stages of *Phytophthora* in pure culture, chapter 5 describes a proteomics investigation of actual infection of tomato with *P. capsici*.

References

1. Erwin, D. C., and Ribeiro, O. K., *Phytophthora Diseases Worldwide*. The American Phytopathological Society: St. Paul, MN, 1996; p 562.
2. Podger, F. D., Phytophthora cinnamomi, a cause of lethal disease in indigenous plant communities in western Australia. *Phytopathology* **1972,** 62, 972-81.
3. Crandall, B. S.; Gravatt, G. F.; Ryan, M. M., Root disease of Castanea species and some coniferous and broadleaf nursery stocks, caused by Phytophthora cinnamomi. *Phytopathology* **1945,** 35, 162-180.
4. Bartnicki-Garcia, S.; Wang, M. C., *Biochemical aspects of morphogenesis in Phytophthora*. American Phytopathological Society: St. Paul, Minn, 1983; p 392.
5. Hendrix, J. W., Sterols in growth and reproduction of fungi. *Annu Rev Phytopathol* **1970,** 8, 111-130.
6. Govers, F.; Gijzen, M., Phytophthora genomics: the plant destroyers' genome decoded. *Mol Plant Microbe Interact* **2006,** 19, (12), 1295-301.
7. Tyler, B. M.; Tripathy, S.; Zhang, X.; Dehal, P.; Jiang, R. H.; Aerts, A.; Arredondo, F. D.; Baxter, L.; Bensasson, D.; Beynon, J. L.; Chapman, J.; Damasceno, C. M.; Dorrance, A. E.; Dou, D.; Dickerman, A. W.; Dubchak, I. L.; Garbelotto, M.; Gijzen, M.; Gordon, S. G.; Govers, F.; Grunwald, N. J.; Huang, W.; Ivors, K. L.; Jones, R. W.; Kamoun, S.; Krampis, K.; Lamour, K. H.; Lee, M. K.; McDonald, W. H.; Medina, M.; Meijer, H. J.; Nordberg, E. K.; Maclean, D. J.; Ospina-Giraldo, M. D.; Morris, P. F.; Phuntumart, V.; Putnam, N. H.; Rash, S.; Rose, J. K.; Sakihama, Y.; Salamov, A. A.; Savidor, A.; Scheuring, C. F.; Smith, B. M.; Sobral, B. W.; Terry, A.; Torto-Alalibo, T. A.; Win, J.; Xu, Z.; Zhang, H.; Grigoriev, I. V.; Rokhsar, D. S.; Boore, J. L., Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **2006,** 313, (5791), 1261-6.
8. Eckert, J. W.; Tsao, P. H., A selective antibiotic medium for isolation of Phytophthora and Pythium from plant roots. *Phytopathology* **1962,** 52, 771-777.
9. Kamoun, S.; Hraber, P.; Sobral, B.; Nuss, D.; Govers, F., Initial assessment of gene diversity for the oomycete pathogen Phytophthora infestans based on expressed sequences. *Fungal Genet Biol* **1999,** 28, (2), 94-106.
10. Qutob, D.; Hraber, P. T.; Sobral, B. W.; Gijzen, M., Comparative Analysis of Expressed Sequences in Phytophthora sojae. *Plant Physiol* **2000,** 123, (1), 243-54.
11. Shan, W.; Marshall, J. S.; Hardham, A. R., Gene expression in germinated cysts of Phytophthora nicotianae. *Mol Plant Pathol* **2004,** 5, (4), 317-30.
12. Chen, X.; Shen, G.; Wang, Y.; Zheng, X.; Wang, Y., Identification of Phytophthora sojae genes upregulated during the early stage of soybean infection. *FEMS Microbiol Lett* **2007,** 269, (2), 280-8.
13. Wang, Z.; Wang, Y.; Chen, X.; Shen, G.; Zhang, Z.; Zheng, X., Differential screening reveals genes differentially expressed in low- and high-virulence near-isogenic Phytophthora sojae lines. *Fungal Genet Biol* **2006,** 43, (12), 826-39.
14. Moy, P.; Qutob, D.; Chapman, B. P.; Atkinson, I.; Gijzen, M., Patterns of gene expression upon infection of soybean plants by Phytophthora sojae. *Mol Plant Microbe Interact* **2004,** 17, (10), 1051-62.

15. Shepherd, S. J.; van West, P.; Gow, N. A., Proteomic analysis of asexual development of Phytophthora palmivora. *Mycol Res* **2003,** 107, (Pt 4), 395-400.
16. Mitchell, H. J.; Kovac, K. A.; Hardham, A. R., Characterisation of Phytophthora nicotianae zoospore and cyst membrane proteins. *Mycol Res* **2002,** 106, (10), 1211-1223.
17. Torto, T. A.; Li, S.; Styer, A.; Huitema, E.; Testa, A.; Gow, N. A.; van West, P.; Kamoun, S., EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen Phytophthora. *Genome Res* **2003,** 13, (7), 1675-85.
18. Ebstrup, T.; Saalbach, G.; Egsgaard, H., A proteomics study of in vitro cyst germination and appressoria formation in Phytophthora infestans. *Proteomics* **2005,** 5, (11), 2839-48.
19. O'Farrell, P. H., High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **1975,** 250, (10), 4007-21.
20. Washburn, M. P., Wolters, D., and Yates III, J.R., Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology* **2001,** 19, 242-247.

Chapter 2

Expressed Peptide Tags: An additional layer of data for genome annotation

Published as Savidor, A., Donahoo, R. S., Hurtado-Gonzales, O., Verberkmoes, N. C., Shah, M. B., Lamour, K. H., and McDonald, W. H. (2006) Expressed Peptide Tags: An Additional Layer of Data for Genome Annotation. *J Proteome Res* 5, 3048-3058.

Abstract

While genome sequencing is becoming ever more routine, genome annotation remains a challenging process. Identification of the coding sequences within the genomic milieu presents a tremendous challenge, especially for eukaryotes with their complex gene architectures. Here we present a method to assist the annotation process through the use of proteomic data and bioinformatics. Mass spectra of digested protein preparations of the organism of interest were acquired and searched against a protein database created by a six frame translation of the genome. The identified peptides were mapped back to the genome, compared to the current annotation, and then categorized as supporting or extending the current genome annotation. We named the classified peptides Expressed Peptide Tags (EPTs). The well annotated bacterium *Rhodopseudomonas palustris* was used as a control for the method and showed high degree of correlation between EPT mapping and the current annotation, with 86% of the EPTs confirming existing gene calls and less than 1% of the EPTs expanding on the current annotation. The eukaryotic plant

pathogens *Phytophthora ramorum* and *Phytophthora sojae*, whose genomes have been recently sequenced and are much less well annotated, were also subjected to this method. A series of algorithmic steps were taken to increase the confidence of EPT identification for these organisms, including generation of smaller sub-databases to be searched against, and definition of EPT criteria that accommodates the more complex eukaryotic gene architecture. As expected, the analysis of the *Phytophthora* species showed less correlation between EPT mapping and their current annotation. While ~76% of *Phytophthora* EPTs supported the current annotation, a portion of them (7.7% and 12.9% for *P. ramorum* and *P. sojae*, respectively) suggested modification to current gene calls or identified novel genes that were missed by the current genome annotation of these organisms.

Introduction

Since the development of automated high throughput sequencing techniques, complete genome sequences of many organisms have been determined. To date, the genomes of 655 species have been completely sequenced (http://igweb.integratedgenomics.com), and many more species have genomic data in the form of expressed sequence tags (ESTs) and cDNA sequences. The advent of higher throughput sequencing techniques[1-3] is expected to allow rapid sequencing of many more species at a dramatically reduced cost. Determination of the DNA sequence is, however, only the initial step in understanding the biology of the organism. The next task is genome annotation, which is achieved by determining which regions of the DNA sequence are coding sequences (genes) and the likely function of each. With the ever

increasing amount of genomic DNA sequence, annotation is primarily done computationally by employing gene-prediction programs. However, computational gene prediction remains a significant scientific challenge[4], especially when applied to higher organisms with complex gene architectures[5]. This fact is evident in a study of the *Caenorhabditis elegans* genome annotation, which suggested that 50% of the predicted genes had introns/exons that were incorrectly annotated[6]. Over the last two decades, many gene prediction programs have been developed (http://www.nslij-genetics.org/gene/programs.html) to address this problem from different angles (reviewed in Mathe *et al*.[7]). One approach involves gene prediction based on classification of coding versus non coding sequences. This prediction is based on extrinsic factors such as DNA sequence similarity to known coding sequences, as well as intrinsic factors such as nucleotide composition (i.e. G+C content) and hexamer usage. A second gene prediction approach involves detection of functional sites specific to a gene, such as splice sites, and start and stop codons. Detection of these functional sites is based on matching to a consensus sequence, where some variation may be permitted. The consensus sequence is determined from a multiple alignment of functionally related documented sequences. Combinations of these approaches are often employed.

Regardless of their underlying approach, all gene prediction programs adopt certain assumptions about the DNA sequence of genes and their features. Thus, only genes that comply with these assumptions will be identified; other genes, such as those containing non-typical sequences will be missed (false negatives). In addition, non-

coding sequences that comply with these assumptions to a sufficient degree will be identified as coding regions (false positives).

Mass spectrometry (MS) has long been used as a sensitive technique for analysis of complex mixtures of proteins[8-12]. When dealing with very complex protein mixtures, such as proteomes, a protein identification approach known as "shotgun proteomics" is often employed[13-15]. Briefly, following proteolytic digestion of the protein mixture, the resulting peptides are separated by chromatography and analyzed by tandem mass spectrometry (MS/MS) in which individual peptides are isolated, fragmented, and the resulting fragment ions are measured. Identification of peptides from tandem mass spectra is typically done computationally using a database search strategy. In database searching, a computer program attempts to find the best match between an experimental spectrum and the predicted spectrum of a peptide from a given protein database. Candidate peptides from the database are scored according to how well their theoretical fragmentation pattern corresponds to the experimental MS/MS spectrum. If it passes appropriate filtering criteria, the best scoring peptide is identified as the peptide giving rise to the spectrum. Identification of peptides results in the identification of the proteins from which they were derived. Searching MS/MS spectra against protein databases using the Sequest[16, 17] and Mascot[18] algorithms has been the gold standard for database searching for several years. Recently, Tabb et al. introduced DBDigger[19, 20], a more efficient and sensitive database searching algorithm that has facilitated much of the analysis presented here. All these programs allow robust, automated, and high-throughput protein identification. However, correct identification by these programs relies on an accurate and complete protein database to be searched against. Often, the protein

database to be searched against is at least partially predicted by gene prediction algorithms. Thus, proteins that are products of genes that were not identified by the gene prediction programs will not be included in the theoretical protein database and cannot be identified by the process described above, even though their peptides were measured.

To overcome the possibility of excluding unpredicted proteins from the protein-database, several groups used a protein database derived from a theoretical translation of genome sequences in all six frames against which to search their mass spectrometry derived data[17, 21-28] Arthur and Wilkins searched peptide mass fingerprints of 20 2D-gel spots from extracellular matrix of *Mycobacterium tuberculosis* against a six-frame translation of the *M. tuberculosis* genome, mapped the peptides back to the genome, and identified regions with high concentration of peptide hits as coding regions[22]. They also demonstrated this strategy in limited application for a single human chromosome[22]. Jaffe et al. mapped *Mycoplasma pneumoniae* peptide hits, identified by searching MS/MS data against a six-frame translation database, onto genomic scaffolds. Open reading frames (ORFs) with peptide hits were then compared to the current genome annotation[25]. Smith et al. searched MS/MS proteomic data against the six-frame-translated genome of *Tetrahymena thermophila* and identified putative coding regions through a BLAST search of areas of the genome that contained peptide hits[27].

Searching against a six-frame-translation protein database has the advantage of not relying on any assumptions made to generate the predicted protein database. However, the inclusion of the translated product of every theoretical ORF in the protein database results in a protein database containing many more entries than the actual number of real proteins, especially for eukaryotes where most of the genome sequence is

non-coding. Searching such larger database may result in an increased false positive rate, because there is a greater chance that a given spectrum would correlate better to a random entry in the database than to the true peptide it was derived from[29]. A search against a very large database may also create another scenario in which a random entry in the database would indeed score less than the true peptide from which the mass spectrum originated, but only slightly less. Since typically search results are filtered to include only those peptides that scored well AND scored significantly higher than the next best scoring peptide for that spectrum (e.g. the DCN filter, which filters out best scoring peptides that scored less than a specified percentage difference than the second best scoring candidate peptide), such a scenario would lead to a reduction of true peptide identifications. Thus, an increase in the false negative rate is another outcome of searching against very large databases. A final practical downfall of this approach is the increased computational time needed to search the much larger databases.

While previously published studies utilized the six-frame translation approach, they were performed on relatively simple or well studied systems. Most published studies employed bacteria with small genomes and simple gene architectures. A few studies on eukaryotes tried to decrease the false positive identification rate by simply increasing the filter thresholds required for peptide identification[30], integrating data from multiple experiments[31] or integrating peptide identification results from different algorithms[32]. Nesvizhskii et al. published a method utilizing iterative searching of spectra that were unassigned in conventional database searching, and may be used for reduction in false negative identification rate[33]. However, none of these studies addressed the need to simultaneously decrease both false positive and false negative identification rates. To the

best of our knowledge, no systematic approach that addresses those issues has been published for the analysis of a eukaryotic proteome in conjunction with the organism's current genome annotation. Here we present a method to incorporate MS proteomic data into the annotation process of both prokaryotic and eukaryotic organisms. This method can be particularly useful for newly sequenced but less well studied organisms, regardless of pre-existing primary data. We show that this method can be applied to eukaryotic organisms with exons and a genome size of at least 100 Mbp. Peptides identified from a search against six-frame translated genomes were mapped onto the genomic sequence and correlated to the current annotation. Peptides that were mapped to a predicted gene call were considered as supporting the current annotation. Clusters of peptide hits that hit outside a predicted gene call were considered either as hitting novel genes or expanding current gene annotations, depending on their distance from a predicted gene call. We named these mapped peptides "Expressed Peptide Tags" (EPTs). Bioinformatic tools were developed to reduce the false positive and false negative identification rates and increase confidence in identification of existing or novel gene predictions in eukaryotes with large genomes and complex gene architecture.

Much like ESTs, EPTs can be used in conjunction with *in silico* gene predictions to increase confidence in the annotation, especially for organisms with limited pre-existing genomic or proteomic data. In this study, three microbial organisms were subjected to the procedure described above: *Rhodopseudomonas palustris*, *Phytophthora sojae*, and *Phytophthora ramorum*. The purple non-sulfur phototrophic bacterium *R. palustris* is a metabolically versatile organism that has been sequenced, extensively studied, and well annotated[34]. *R. palustris* was therefore chosen to test the utility of using

EPTs against a well annotated genome. Organisms in the genus *Phytophthora* are eukaryotic plant pathogens within the phylum Oomycetes. The genomes of both *P. sojae* and *P. ramorum* have recently been sequenced by the Joint Genome Institute (http://www.jgi.doe.gov/) and are currently being annotated. Although organisms of the genus *Phytophthora* share many morphologic similarities to fungi (Kingdom Myceteae), they are phylogenetically distinct. Ribosomal RNA sequencing suggests that *Phytophthora* is more closely related to brown algae and diatoms than to fungi, and is classified into Kingdom Chromista[35]. Hence, it is possible that the genomes of these organisms contain many genes with non-typical sequences that are likely to be missed using conventional annotation tools. As expected, correlation between EPT data and the current annotation was found to be better for *R. palustris* than for *P. sojae* and *P. ramorum*. This was evident from the higher number of expanded or novel genes identified for the *Phytophthora* species as compared to those of *R. palustris*. We anticipate that proteomic studies and the resulting EPTs will play an increasingly important role in aiding the annotation of newly sequenced organisms, especially those of poorly studied genera.

Materials and methods

R. palustris cell culture, protein preparation and LC-MS/MS analysis

Wild type *R. palustris* CGA0010 strain was grown to mid-log phase under photoheterotrophic conditions and proteome fractions were prepared as described previously[36]. For this study only crude soluble and membrane fractions were used.

Proteomes were denatured, digested with sequencing grade trypsin and concentrated as described previously[36]. Digested proteome fractions were analyzed by Multidimensional protein identification technology (MudPIT) as described below except 12-step 24-hour analyses were used instead of 11-step, 22-hour.

Phytophthora cell culture and life stage isolation

P. sojae and *P. ramorum* were grown in clarified antibiotic amended V8 juice broth to generate mycelium and on antibiotic amended V8 agar plates to generate asexual sporangia and zoospores according to standard protocols[37]. Zoospores were stimulated to encyst by shaking them vigorously for 1-2 minute and then waiting approximately 1 hour before harvesting the germinating cysts. The mycelium and germinating cysts were collected into 2 ml screw top microfuge tubes pre-loaded with four 3 mm glass balls and freeze dried for 48 hours in a LabConco stoppering tray freeze dry apparatus (Labconco Corp.; Kansas City, MO). The dried material (~15 mg dried mycelium, ~10 mg dried germinating cyst) was then ground into a fine powder using a Mixer Mill 300 disruption device (Qiagen, Valencia, CA) set to the highest setting for 1 minute.

Phytophthora lysate preparation

Homogenate was resuspended in 300 ul 8 M urea 100 mM Tris buffer pH 8.5 and vortexed vigorously. Suspended lysate was then centrifuged at 14000Xg for 10 minutes at 4˚C to pellet cell debris. Supernatant was transferred to a new Eppendorf tube and centrifuged for 1 hour at 4˚C for fractionation. The top 150 ul was transferred to a new eppendorf tube and considered as “crude soluble fraction” while the remaining

supernatant was considered as “membrane fraction”. For this study, only crude soluble fraction was used for *Phytophthora* analysis. The sample was reduced with tris(2-carboxyethyl)phosphine (TCEP) to final concentration of ~6.5 mM for 20 minutes at room temperature, and alkylated with iodoacetamide (IAA) to final concentration of ~2.5 mM for 20 minutes at room temperature in the dark. 450 ul 100 mM tris buffer pH 8.5 was then added to reduce urea concentration to 2 M, and $CaCl_2$ added to a final concentration of 0.8 mM. Proteins were then digested overnight with 20 μg sequencing grade trypsin (Promega, WI) at 37˚C. The following day, another 20 μg sequencing grade trypsin was added and again incubated overnight at 37˚C. Before loading onto a column samples were acidified to 4% with formic acid and filtered through Ultrafree-MC (0.45 μm filter unit) Microcon tubes (Millipore, MA).

Phytophthora LC/LC-MS/MS analysis

Filtered and acidified samples originating from approximately 5mg dry *Phytophthora* mycelium before grinding or 3mg dry *Phytophthora* cyst before grinding were loaded onto a biphasic fused silica column offline[38, 39]. Column inner diameter was 150 μm. The first phase consisted of 3-3.5 cm reverse-phase (RP) material (Aqua, C18 5μm 200A, Phenomenex, Torrance, CA), while the second phase consisted of 3.5-4 cm Luna strong cation exchange (SCX) material (Luna SCX 5μm 100 A, Phenomenex, Torrance, CA). The column was packed with the two phases sequentially via a pressure cell (New Objective, Woburn, MA). After the entire sample was loaded, buffer B (70% acetonitrile, 30% H_2O, 0.1% Formic acid) was run through the column for 30 minutes to clear hydrophobic molecules that were leading to column clogging, and the column was

then re-equilibrated with buffer A (95% H_2O, 5% acetonitrile, 0.1% formic acid) for 20 minutes. Although highly negatively charged peptides may have eluted and been lost, this procedure prevented column clogging while allowing direct sample loading and desalting prior to eluting peptides onto the SCX. The column was then used in a two-dimensional (2D) nano-LC-MS/MS system with a split-phase column[39], by mounting the column directly upstream of a 100 m inner-diameter front column packed with 15 cm Jupiter reverse-phase material (Jupiter C18 5µm 300 A, Phenomenex, Torrance, CA). The front column (Pico Tip, 100µm with 15µm tip from New Objective, Woburn, MA) was also packed via a pressure cell. For LC/LC-MS/MS analysis, a Surveyor HPLC (ThermoElectron, San Jose, CA) was interfaced with a linear ion trap mass spectrometer (LTQ, ThermoElectron, San Jose, CA). Flow rate off the front column was 200-300 nl/min. The crude soluble proteome was analyzed via a 22-hour, 11-step MudPIT analysis, as previously described[38, 40, 41], with the only exception of omitting the first MudPIT step in the analysis of *Phytophthora* samples. Only the first MudPIT experiment that was performed (*P. ramorum* germinating-cyst 1) was done with the full 24-hour, 12 step analysis (column clogging in subsequent experiments required the adjustment). Experiments were performed in duplicates on soluble fractions from two life stages (mycelium and germinating-cyst) from *P. sojae* and *P. ramorum*.

Bioinformatics

Predicted protein databases as well as genome sequences of *R. palustris*, *P. sojae*, and *P. ramorum* were downloaded from the JGI website (http://www.jgi.doe.gov/). A series of Perl scripts was created for translating nucleotide databases and for other

downstream data analysis applications. The predicted protein database of each organism was reversed and a concatenated database containing the forward and reversed database was created. MS/MS spectra from MudPIT experiments were searched against the original predicted protein databases, as well as the concatenated forward and reversed database, using the DBDigger algorithm in fully tryptic mode with the MASPIC scorer. MASPIC score filter values required to allow 1 hit to the reversed database for every 19 to the forward database (5% false discovery rate) were determined using the SQTRevPuller algorithm courtesy of David Tabb. Based on the score distribution of reverse hits, SQTRevPuller seeks a threshold for score and DCN that will remove a user-specified percentage of these reverse hits[42]. The determined values were used to filter the list of peptides identified in the search against the original predicted protein database. Protein identification required 2 peptides per protein for identification.

The genome sequence of each organism was translated in all six frames, and translation products of all theoretical open reading frames (ORFs) that contained at least 35 amino acid residues between stop codons were included in a theoretical protein database in a FASTA format. Each entry in the database contained a header indicating the start and stop nucleotide numbers in the genomic sequence, as well as the frame and the scaffold to which it mapped. Spectra from MudPIT experiments were searched against this protein database using the DBDigger algorithm in fully tryptic mode. Sub-databases were generated for *P. sojae* and *P. ramorum* based on the search results as follows: the 20% false discovery rate filter values were determined as described above. The first and second best scoring peptides for each spectrum were included in a list if their scores passed these filters. Entries in the original database that contained two or more peptides

in the list were included in the sub-database. Entries in the original database that contained only one peptide in the list, but were within 200 bp from an entry containing 2 or more peptides in the list, were also included in the sub-database. The MS/MS spectra were searched against the sub-databases using the DBDigger algorithm in the semi-tryptic mode. A representative fully tryptic search using the Sequest algorithm was also performed on one MudPIT experiment from each of the *Phytophthora* species against the complete six-frame-translation protein database. All experimental RAW files, databases, unfiltered (DTASelect.txt) and filtered search results (DTASelect-filter.txt and DTASelect.html) used in this study are available for download at http://compbio.ornl.gov/mspipeline/ept. The Sequest search results were similar to the DBDigger search results (data not shown), and suggested the previously reported consistency between the algorithms[19] when applied even to large databases.

For *P. sojae* and *P. ramorum*, each peptide was then mapped back to the genome and was classified into one of four categories (Figure 2.1): (i) *Confirming* current gene call- if the peptide was mapped to a predicted coding region (ii) *Expanding* current gene call- if the peptide was located outside of a predicted coding region (even if only on a different reading frame) but within 200 base pairs from the expanded gene AND there was an additional peptide hit in the expanded gene within 1000 bp from it. A peptide was also considered as *expanding* a current gene call if it was located within 500 bp from the gene, had an additional peptide hit in the expanded gene within 1000 bp from it, AND was in the same ORF as the expanded gene (or exon it expanded) (iii) Suggesting *novel* gene- if the peptide was located further than 200 bp away from any predicted gene call, and was a part of a cluster of at least three different peptide hits (different amino acid

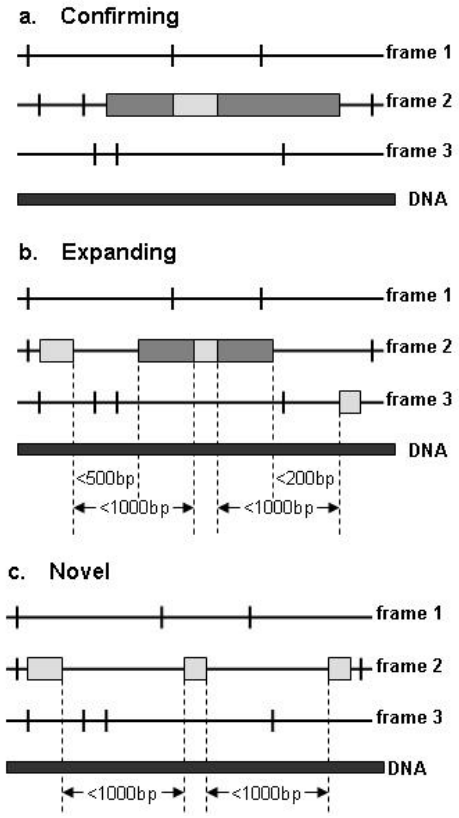


Figure 2.1- EPT rules. (a) *Confirming* EPT- The peptide is located within a predicted gene. (b) *Expanding* EPTs- peptide is located less than 200 bp away from a predicted gene (right peptide) which was confirmed with a *confirming* EPT (middle peptide) within 1000 bp. The *expanding* EPT may suggest an additional exon not included in the current gene prediction. If the peptide is on the same ORF as the gene (left peptide), it may be located within 500 bp away from the gene. (c) *Novel* EPTs- a cluster of at least 3 peptides separated by less than 1000 bp from one another, and located more than 200 bp away from a confirmed gene. For *R. palustris*, *expanding* EPTs had to be on the same ORF and within 500 bp from the expanded gene, and *novel* EPT clusters also had to be on the same ORF. Legend: bottom thick black line- One DNA strand. Horizontal Black lines- Amino acid sequence of the translated DNA sequence in all 3 frames. Vertical black lines- Stop codons. Dark grey rectangle- predicted protein. Light grey boxes- peptides hits (EPTs).

sequences) separated by intervals of 1000 bp or less. (iv) None of the above. For the exon-less bacterium *R. palustris*, classification definitions were similar, with the exception that *expanding* peptides always had to be in the same ORF as the expanded gene, and all peptides suggesting a *novel* gene had to be in the same ORF. This classification was done automatically and results formatted to allow visualization in the Artemis genome browser (The Sanger Institute, http://sanger.ac.uk/Software/Artemis/index.shtml).

cDNA preparation and analysis

RNA was isolated from ~0.7 g *P. ramorum* mycelium. Mycelium was grown as described elsewhere[43]. Mycelium was frozen in liquid nitrogen and was immediately ground with mortar and pestle. Total RNA isolation was performed using the TRIzol reagent (Invitrogen, CA) following manufacturer instructions. RNA preparation was treated with DNase I (Ambion, TX) according to manufacturer instructions, and ~0.4 ug treated RNA was used to synthesize cDNA using the REVERSE-IT 1st Strand Synthesis Kit (Abgene, UK) according to manufacturer instructions. Primer sequences for the novel_39 gene were: 5'-CAACTGGGTGGTGCTCTACA-3', and 5'-AGCTCCTCCTTCAGGTCACA-3' (amplifying 336 bp). Primers for the expanded_93339 gene were: 5'- AACTCGCATTGTGACTCCGTTTG-3' and 5'-CACAATGTCGCGGTCGTCGAAG- 3' (377 bp). Primers for the □-tubulin control were: 5'- TACGGTGACCTGAACCACCT- 3' and 5'- GGGATCCACTCGACGAAGTA- 3' (373 bp). PCR reaction was set to cycle 35 times with annealing temperature of 60°C and elongation time of 30 seconds at 72°C. PCR products were resolved on 1% agarose gel.

PCR products were cleaned using the Qiagen PCR clean up kit (Qiagen, CA) following the manufacturer instructions. Cleaned PCR products were sequenced at the Molecular Biology Resource Facility (University of Tennessee). Sequences were analyzed using the Sequencher 4.5 software (Gene Codes, MI). The sequences were also BLASTed (http://genome.jgi-psf.org/cgi-bin/runAlignment?db=ramorum1&advanced=1) against the *P. ramorum* genome to ensure that the sequences were unique, and that the mRNA did not originate from elsewhere in the genome.

Results and Discussion

The organisms *Rhodopseudomonas palustris*, *Phytophthora sojae,* and *Phytophthora ramorum* were subjected to MS/MS proteomic analysis. In order to express as many proteins as possible, proteolytic preparations of soluble fractions of lysates from two life stages (mycelium and germinating cyst) from *P. ramorum* and *P. sojae*, as well as soluble and membrane fractions from *R. palustris* were made. The proteolytic preparations were subjected to duplicate MudPIT experiments. First, MS/MS spectra of each organism were searched against their respective predicted protein database. Results are presented in table 2.1. These searches were done in order to get a qualitative assessment of the data and were not used for further analysis in this study. Consistent numbers of proteins were identified in each MudPIT run from identical samples, indicating a quality dataset. Next, the same MS/MS spectra were searched against a six-frame translation protein database containing translation products of all theoretical open-

Table 2.1- Proteins identified from searches against predicted protein databases

Organism	**Life stage/ growth condition**	**Proteins identified in MudPIT set 1**	**Proteins identified in MudPIT set 2**
R. palustris	Soluble fraction	1896 (19,230)	1727 (16,532)
	Membrane fraction	1509 (11,977)	1456 (11,226)
P. sojae	Mycelium	1034 (4,715)	909 (4,175)
	Germinating cyst	1462 (8,102)	1339 (7,968)
P. ramorum	Mycelium	1379 (6,700)	1592 (7,303)
	Germinating cyst	1361 (7,104)	1364 (6,364)

* Database searches were done using DBDigger. At least two peptides at 5% false discovery rate were required for protein identification. Peptide numbers are in parenthesis.

reading-frames (ORFs) with at least 35 amino acid residues between stop codons. The genomes of the *Phytophthora* species are fairly large (95 Mbp for *P. sojae* and 62 Mbp for *P. ramorum*). The six-frame translation protein databases for the *Phytophthora* species contain hundreds of thousands of entries, corresponding to millions of possible peptides. Thus, when trying to match a spectrum to a peptide from the database, the search algorithm attempts to find the best possible match among millions of candidates in the database and is likely to produce many false identifications[44]. Estimation of the false positive identification rate can be done by searching MS/MS spectra against a concatenated database including a forward protein database as well as its reversed version (where the amino-acid sequence of each protein is written from C-terminal to N-terminal)[8, 45]. A search against such a concatenated database was done for one representative MudPIT experiment for each organism. Filter thresholds that provided 19 identifications from the forward database for every 1 identification from the reverse database (~5% false discovery rate) were determined. The false discovery rate may even be less than 5% since peptide sequences that were identified from the reverse database were considered as false positives even if they also occurred and identified from the forward database. The results of the DBDigger search after filtering for all 3 organisms are presented in table 2.2. Each peptide was then mapped back to the genome and was classified into one of four categories: *Confirming* an existing gene call, *expanding* a current gene call, suggesting a *novel* gene, or none of the above (see bioinformatics section in materials and methods, figure 2.1). We have termed these classified peptides Expressed Peptide Tags (EPTs). The results from this analysis are presented in table 2.3

and table 2.4. This classification was automated and output in a format for visualization in Artemis (Sanger Institute), allowing manual examination of the results.

As can be seen from table 2.3, a larger proportion of the peptides could not be classified into any EPT category for the *Phytophthora* species than for *R. palustris*. This could be for several reasons, one of which is their much larger genomes. In order to reduce the size of the protein databases of *P. sojae* and *P. ramorum*, and increase confidence in peptide identification, a smaller protein database, containing only products of those ORFs from which at least two different peptides were identified, was constructed. The list of peptides used for this application was generated by including both the first and second best scoring peptides for each spectrum in the searches against the original six-frame translation protein databases (4 MudPIT experiments for each organism). The filter thresholds used to filter the peptide list were more liberal, excluding only 80% of peptide hits from the reverse database in a representative concatenated-database search (estimated 20% false positive rate). This was done in order to include as many potential true coding ORFs as possible, without dramatically increasing the size of the resulting sub-database. ORFs with only a single peptide hit from the list were also included in the sub-databases if they were within 200 bp from an ORF with multiple peptide hits on the same DNA strand. This was done in order to include short exons that may only give rise to a single tryptic peptide. The constructed sub-databases were significantly smaller than the original six-frame translation databases: 40,249 versus ~1.3 million for *P. sojae*, and 100,521 versus 941,453 for *P. ramorum*. The *P. sojae* sub-database was smaller than the *P. ramorum* database, even though its original six-frame translation protein database was larger. This may be the result of the higher filter

Table 2.2- Peptides identified from searches against six-frame translation protein databases

Organism	Life stage/ fraction	Peptides identified in MudPIT set 1	Peptides identified in MudPIT set 2	Total number of peptides identified
R. palustris	Soluble	18,484	15,873	56,880
	Membrane	11,612	10,911	
P. sojae	Mycelium	8,243	7,716	42,177
	Germinating cyst	13,392	12,826	
P. ramorum	Mycelium	14,193	14,397	58,016
	Germinating cyst	15,122	14,304	

Table 2.3- EPTs identified by searching against six-frame translation databases

Organism	Confirming	Expanding	Novel	None	Total
R. palustris	48,944 (86%)	162 (0.3%)	360 (0.6%)	7,414 (13%)	56,880
P. sojae	29,460 (69.8%)	1,253 (3%)	2,602 (6.2%)	8,862 (21%)	42,177
P. ramorum	34,739 (59.9%)	2,553 (4.4%)	2,215 (3.8%)	18,509 (31.9%)	58,016

Table 2.4- Proteins corresponding to EPTs

Organism	Confirmed	Expanded	Novel	Total genes predicted in genome
R. palustris	2,814 (58.2%)	111 (2.3%)	85 (1.8%)	4,836
P. sojae	3,797 (20%)	568 (3%)	190 (1%)	19,027
P. ramorum	4,803 (29.9%)	1,538 (9.6%)	548 (3.4%)	16,066

thresholds required to provide the same false discovery rate. In other words, a larger original database leads to more false positives, resulting in a requirement for increased filter thresholds and thus fewer peptides are used in creating the subsequently smaller sub-database.

Generation of and search against the sub-databases reduces both false positive and false negative peptide identifications. The false negative rate is reduced because best scoring true peptides that scored well, but not significantly higher than the next best scoring peptide, would not be filtered out due to omission of the DCN filter (the two best scoring peptides are included in the peptide list used to generate the sub-database, regardless of the difference in their score).

As mentioned earlier, false positive identifications result from a better correlation between the experimental spectrum and a random entry in the database than the actual peptide generating the spectrum. The probability that two false positive occurrences would originate from the same non-coding theoretical ORF is very low. For example, assuming a 20% false positive rate for the peptides in the peptide list used to generate the *P. sojae* sub-database, a random ORF would have a probability of < 1 % to be included in the sub-database. Thus, the generation of a sub-database results in exclusion of the vast majority of these theoretical non-coding ORFs from the sub-database. When the sub-database is searched against with the same MS/MS data, the second best scoring peptide from the original six-frame search, which is likely the true peptide, would now be identified as the best scoring one. Searching against the smaller sub-database also reduces searching time dramatically, as has been previously described[46].

The smaller sub-databases were searched with the same proteomic data. A representative concatenated forward/reverse protein sub-database was also generated and searched against to determine appropriate filter threshold values. Search results against the forward sub-databases were filtered with threshold values estimated to exclude 95% false positive identifications. The entire process is outlined in figure 2.2. The results from the searches against the sub-databases for *P. sojae* and *P. ramorum* are presented in table 2.5. EPT classifications are presented in tables 2.6 and 2.7. This process was also applied to the *R. palustris* data, and no significant change in EPT proportions was observed when the sub-database was used (data not shown). This could be explained by the relatively small original six-frame translation database (82,096 entries). Construction of the *R. palustris* sub-database (>40% of original six-frame translation database size) does not decrease the search space enough to reduce false positive identification by a significant amount.

As can be seen from these tables, searching against the smaller database increased the number of classified EPTs, and reduced the number of peptides that did not fit any of the categories by ~50%. While the percentage of *expanding* and *novel* EPTs of the total did not change substantially, the percentage of *confirming* EPTs increased markedly, thereby increasing confidence in their corresponding protein identifications. If the search against the sub-database had yielded more *confirming* EPTs simply due to the increased probability of a peptide to hit an existing entry in such smaller database, a proportional increase in the *expanded* and *novel* categories would also be expected.

Figure 2.2- Annotation pipeline for *P. ramorum* and *P. sojae*. Two MudPIT experiments were performed on each of the mycelium and germinating-cyst life stages of each organism. Searches against both six-frame translation protein database and protein sub-database used the same MudPIT generated spectra for each organism.

* For sub-database creation, both the first and the second best scoring peptides for each spectrum were included in the list if they passed the filter threshold.

** For organisms with small genomes, generation of a sub-database may not be necessary. In such case, EPT generation can start from this point of the pipeline using the entire six-frame translation database.

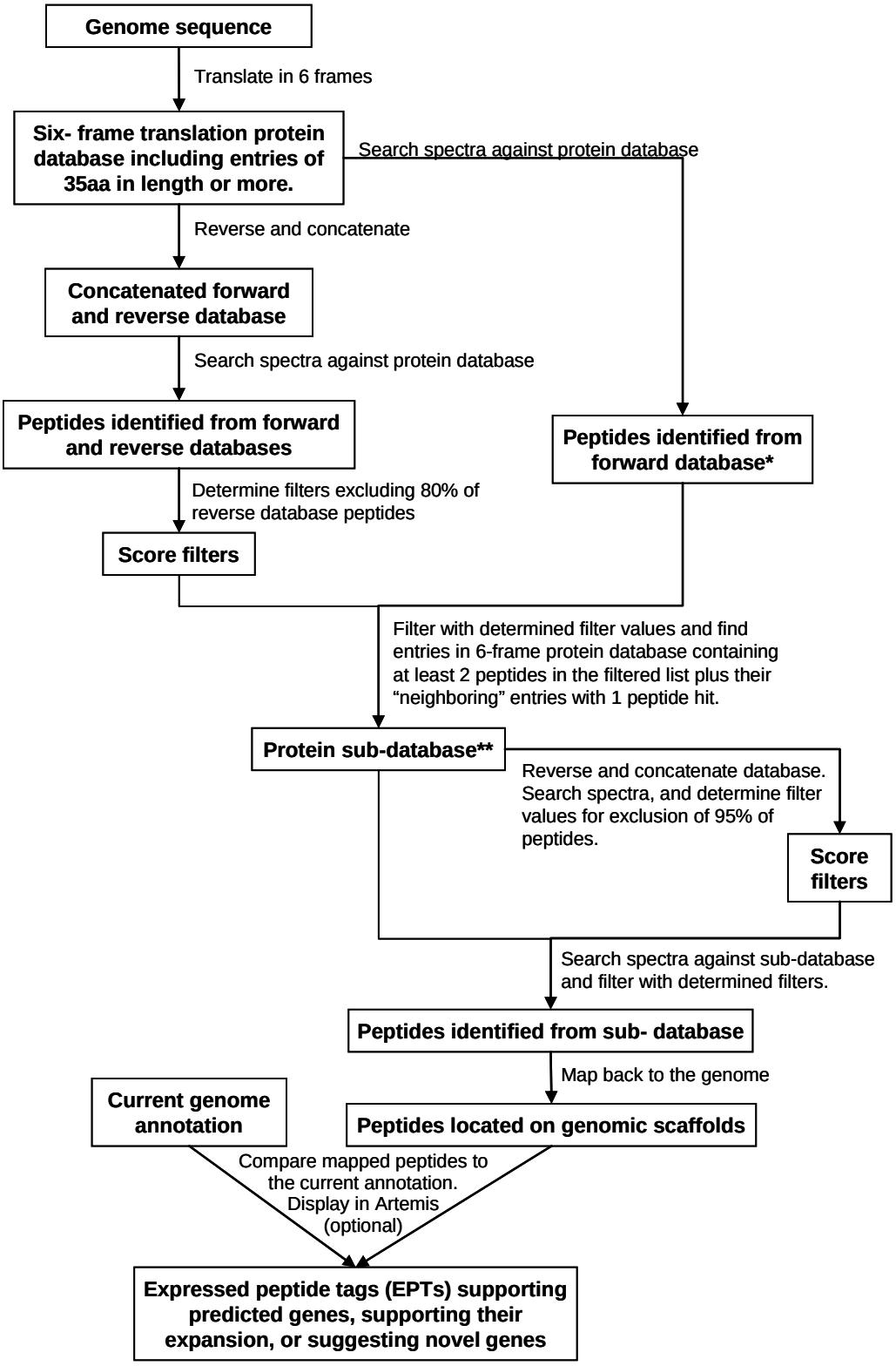


Figure 2.2- Continued.

Table 2.5- Peptides identified from searches against the generated sub-databases

Organism	Life stage/ fraction	Peptides identified in MudPIT set 1	Peptides identified in MudPIT set 2	Total number of peptides identified
P. sojae	Mycelium	11,850	9,750	50,407
	Germinating cyst	14,242	14,565	
P. ramorum	Mycelium	13,357	13,539	53,170
	Germinating cyst	13,338	12,936	

Table 2.6- EPTs identified by searching against protein sub-databases

Organism	Confirming	Expanding	Novel	None	Total
P. sojae	38,853 (77.1 %)	2,601 (5.1 %)	3,945 (7.8 %)	5,008 (9.9 %)	50,407
P. ramorum	40,122 (75.5 %)	2,575 (4.8 %)	1,541 (2.9 %)	8,932 (16.8 %)	53,170

Table 2.7- Proteins corresponding to EPTs in sub-database analysis

Organism	Confirmed	Expanded	Novel	Total genes predicted in genome
P. sojae	3,967 (20.8%)	927 (4.9%)	329 (1.7%)	19,027
P. ramorum	4,958 (30.9%)	1,327 (8.3%)	331 (2.1%)	16,066

The bacterium *R. palustris* was chosen as a case study to demonstrate the correlation between our proteomic data and the genome annotation due to its well annotated genome and relatively simple gene architecture. Eighty six percent of the total number of EPTs identified for *R. palustris* were classified as *confirming* EPTs, indicating a tight correlation between the EPT mapping and the current annotation. Most of the *confirming* EPTs hit an ORF with at least one additional peptide hit, providing strong evidence for the presence of their corresponding protein in the sample, and reducing the likelihood that they are spurious false positives. Figure 2.3a shows the distribution of EPTs per confirmed gene for *R. palustris*. 162 *expanding* EPTs suggested extension of the N-termini of 111 gene predictions (some N-termini were expanded with more than one EPT). 360 peptides, corresponding to only 85 ORFs, suggested novel loci. These observations are consistent with a well annotated genome where most genes are predicted correctly. Figure 2.4 shows examples of categorized *R. palustris* genes.

The analysis of the *Phytophthora* species showed looser correlation between the EPTs and current genome annotation than was observed for *R. palustris*. On average, 76.3% of all identified EPTs for each organism fell into the *confirming* category, 4.9% fell into the *expanding* category, 5.3% fell into the *novel* category, and 13.3% did not fit any of the categories (examples shown in figure 2.5). Thus, in comparison to *R. palustris*, EPT mapping showed a reduced rate of *confirming* EPTs for *Phytophthora*, as well as an increased rate of *novel* and *expanding* EPTs. This looser correlation may be due to: (i) Less complete and accurate genome annotation for *P. sojae* and *P. ramorum* (ii) While *R. palustris* is a haploid bacterium, *Phytophthora* is a diploid eukaryote which can support allelic polymorphism. Thus, a peptide different in a single amino-acid than the predicted

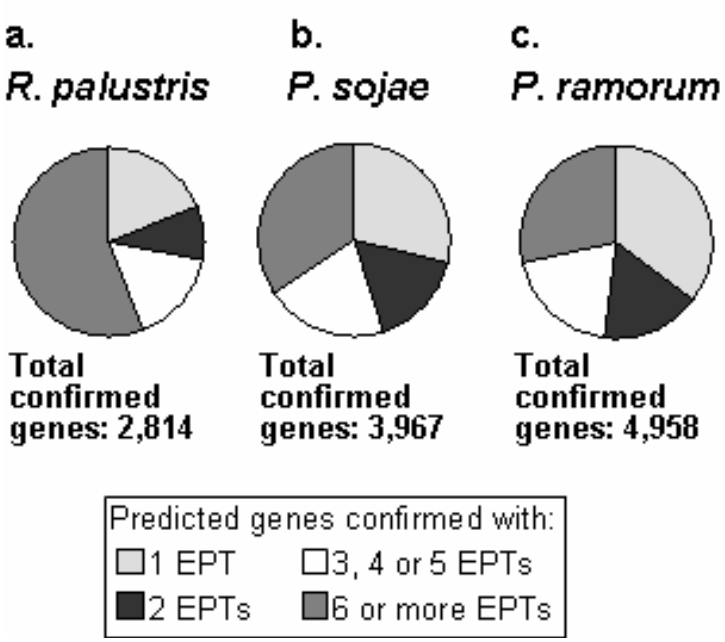


Figure 2.3- Predicted genes with *confirming* EPTs. For each organism, spectra from 4 MudPIT experiments were searched against its protein database (sub-databases in the case of *P. sojae* and *P. ramorum*). Peptides that were mapped back to predicted genes were considered *confirming* EPTs. The pie charts show the distribution of *confirming* EPTs per confirmed predicted gene for each organism.

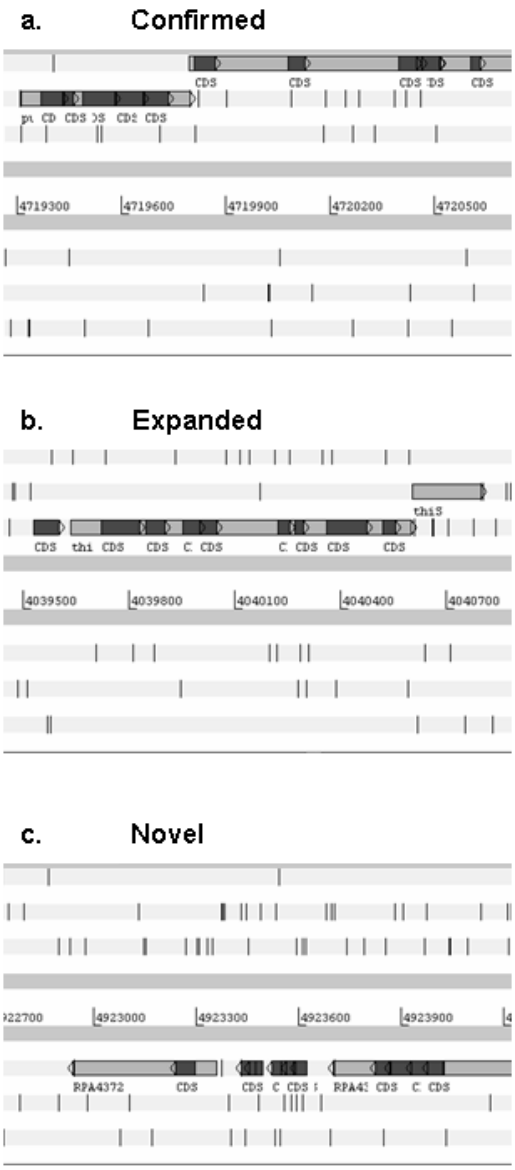


Figure 2.4- Artemis screen-shots of *R. palustris* EPTs. Central dark grey lines- Both DNA strands. Light grey lines- Amino acid sequences of translated DNA in all 6 frames. Vertical black lines- Stop codons. Dark grey rectangles on the amino acid sequences- predicted protein. Black boxes on the amino acid sequence - peptides hits (EPTs). (a) Two predicted genes (on frame 2 and frame 1) confirmed with multiple EPTs. (b) An *expanding* EPT (left) suggesting the N-terminal expansion of a predicted gene confirmed with multiple *confirming* EPTs. (c) Multiple EPTs located in between two predicted genes on frame 6, suggesting a novel gene.

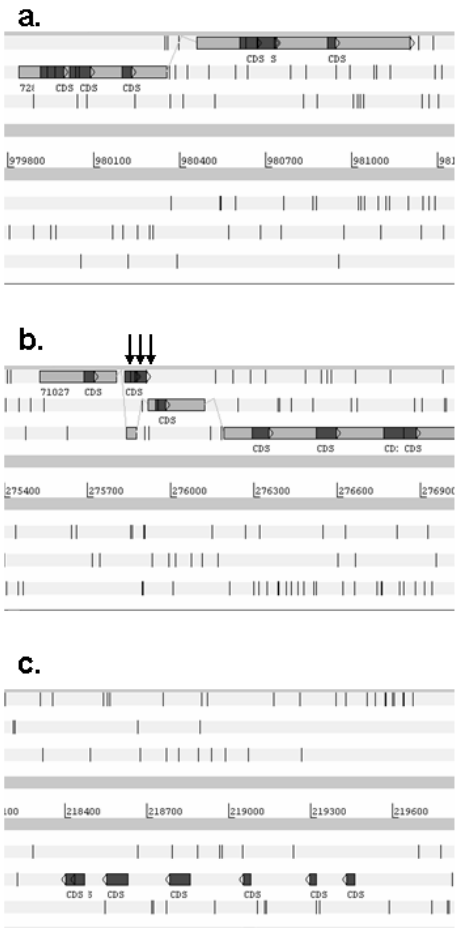


Figure 2.5- Artemis screen-shots of *P. ramorum* EPTs. Central dark grey lines- Both DNA strands. Light grey lines- Amino acid sequences of translated DNA in all 6 frames. Vertical black lines- Stop codons. Dark grey rectangles on the amino acid sequences- predicted protein. Black boxes on the amino acid sequence - peptides hits (EPTs). (a) A single predicted gene with two exons (exon 1 on frame 2 splices to exon 2 on frame 1) confirmed with multiple EPTs. (b) Gene Pr_71027 is predicted to have 4 exons. Exon 1 (leftmost), exon 3 (second from the right) and exon 4 (rightmost) were confirmed with multiple EPTs. Three *expanding* EPTs (pointed to by the black arrows) were mapped between exon 1 and 3, suggesting the C-terminal extension of exon 1 or the location of exon 2 on frame 1. (c) Seven different *novel* EPTs on frame 6, suggesting a novel gene.

translation sequence of the published *P. sojae* genome will not be identified and mapped back to its correct location on the genome, resulting in a false negative and a reduced *confirming* EPT rate. If the incorrectly identified peptide maps somewhere else in the genome, the result will also be a false positive if it passes the filter criteria.

As was the case for *R. palustris*, the majority of confirmed *Phytophthora* genes contained more than one EPT (figure 2.3b and 2.3c). Due to the size of the database, the likelihood that multiple false positive EPTs will map to a single non-coding translated sequence is low. Therefore, multiple EPTs that were located to a single confirmed gene increased the confidence that this was indeed an expressed protein and not a non-coding region with a single miss-identified peptide hit. A single EPT located to an ORF may be a false positive or may have been derived from a low abundance protein from which no other peptides were detected. Additional experiments on different expressed proteomes of these organisms (achieved by different culture conditions or isolation of different life stages) may resolve some of these ambiguities. Under these alternate conditions, some proteins will be more highly expressed, and additional peptides derived from them will be identified. If the ORF in question is non-coding, it is unlikely to obtain additional EPTs and the confidence in its identification will not increase.

The remaining EPTs which were not classified as any of the previous three categories may be either false positives or false negatives. Incorrect classifications may be the result of one or more reasons: (i) The peptide was not included in any of the categories due to the definitions of their nucleotide distance cut-offs. For example, an *expanding* peptide that was derived from a *Phytophthora* gene exon more than 200 bp

away from the predicted end of the gene would not be considered as *expanding*. (ii) The peptide was derived from a low abundance unpredicted protein, and the absence of additional peptide identification from the same protein prevented it from passing the two/three hit filter (iii) Low quality spectra identified incorrectly (iv) A peptide sequence that occurs in a protein can also be found in an additional non-coding ORF in the six-frame protein database (redundant). Although the identification is derived from the true peptide, a peptide hit would also be displayed for the rest of these sequence occurrences. (v) Imperfect genome sequence/ polymorphism which resulted in a wrong amino acid in an entry of the translated protein database. A peptide containing a different sequence will most likely not be identified correctly. (vi) Peptides of contaminating proteins in the protein preparation may also result in false identifications.

While confidence in individual peptide identifications was approximated at 95% based on a reverse database search, the confidence level in identification of a protein and more importantly its classification into a certain category is more problematic. Since the clustering rules required that multiple peptides be present in close proximity to each other, it was very unlikely that under those circumstances (i.e. number of spectra searched and the database size) that two or more incorrect hits clustered together and classified into an individual EPT category. Deciding which peptides to use for classifying a protein depends on the inclusion of these peptides in an EPT category, which in turn depends on the presence of other peptides, as well as predicted genes localized in its vicinity. Together, these factors make confidence at the protein and protein classification levels difficult to infer accurately.

In order to demonstrate the utility of EPT mapping, expression of candidate genes suggested by *expanding* or *novel* EPTs was monitored. RNA was isolated from *P. ramorum* mycelium, and cDNA was synthesized from it. Candidate genes were then PCR amplified with targeted primers, and products analyzed by agarose gel electrophoresis. The results are presented in figure 2.6. Gene Pr_93339 was predicted to span 532 bp and 2 exons (nucleotides 683452 to 683483 and 683510-684011 in genomic scaffold 1). In addition to a *confirming* EPT at the C-terminal of predicted exon 2, multiple *expanding* EPTs were identified both up and downstream of exon 2 on frame 3 (Figure 2.6b) and were included in the *expanding gene call* category. Primers were designed to amplify cDNA beyond exon 2. Presence of a gel-band of the correct length indicated RNA expression of the expanded version of the protein suggested by the EPT data (Figure 2.6c).

A novel gene, designated Novel_39, was evidenced in scaffold 39 of *P. ramorum*, starting downstream to nucleotide 337778 on the reverse DNA strand. This region of the genome was not predicted to contain a gene by the current annotation. However, our proteomic data, including MudPIT runs from both *P. ramorum* mycelium and *P. ramorum* germinating cyst life stages, identified multiple *novel* EPTs within a range of ~1 Kbp in frames 6, 5 and 4 (figure 2.6a), possibly corresponding to three exons of the gene. PCR amplification was carried out on *P. ramorum* mycelium cDNA with primers designed to include the 5' most EPT as well as the 3' most EPT on frame 6 (Figure 2.6c). The appearance of a band on an agarose gel corresponding to a fragment of ~350 bp confirmed that this ORF is indeed a gene which is expressed in the mycelium life stage of *P. ramorum*. A sequencing reaction confirmed the identity of both sequences (data not

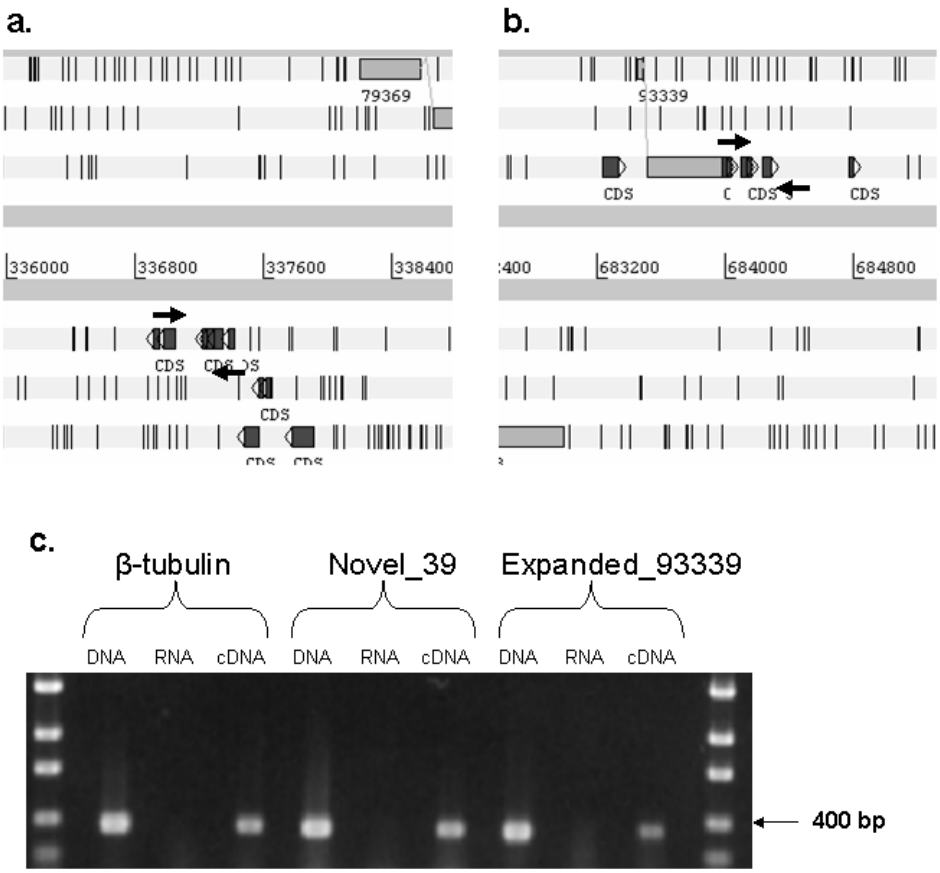


Figure 2.6- Validation of EPTs. Central dark grey lines- Both DNA strands. Light grey lines- Amino acid sequences of translated DNA in all 6 frames. Vertical black lines- Stop codons. Dark grey rectangles on the amino acid sequences- predicted protein. Black boxes on the amino acid sequence - peptides hits (EPTs). Black arrows- primers used for RT-PCR (a) Artemis screen shot of *novel* EPTs. A cluster of 11 different *novel* EPTs was identified in scaffold 39 within a region of ~1Kbp. The position of the EPTs may suggest a novel gene with 3 exons: exon 1 on frame 4, exon 2 on frame 5 and exon 3 on frame 6. The 5'-most EPT on frame 4 (leftmost black box on the lowest amino acid sequence) may be a false positive, or may indicate alternative splicing of the gene. (b) Artemis screen shot of expanding EPTs. Gene Pr_93339 is predicted to have 2 exons (exon 1 on frame 1 and exon 2 on frame 3). In addition to a *confirming* EPT at the C-terminal of predicted exon 2, multiple expanding EPTs were identified both up and downstream of exon 2 on frame 3. (c) cDNA was transcribed from *P. ramorum* mycelium mRNA which was treated with DNase. Genomic DNA preparations, RNA preparations and cDNA preparations were all PCR amplified and resolved by electrophoresis on an agarose gel. Primers for amplification were designed to amplify a ~350 bp region of interest for each of the genes showed in 2.6a and 2.6b (primers indicated by black arrows). A band at ~350bp in each cDNA lane indicated the suggested regions of interest were indeed expressed. The lack of PCR product in all the control RNA lanes indicated that no DNA contamination was present in the RNA preparation, and that it was only the synthesized cDNA that gave rise to the PCR product in the cDNA lanes.

shown). The sequences were also Blasted against the *P. ramorum* genome, and a failure to match any other sequences in the genome indicated that the loci are expressed and not the result of distinct duplicated regions. All of these lines of evidence prove that EPTs mapped to these two loci correctly identified genes that were not predicted correctly in the current annotation. These verification experiments could be performed for any gene identified as *novel* or expanded by EPT mapping.

Conclusions

We have developed a method to integrate proteomic MS data with current genome annotations, in order to improve the quality of gene predictions. In this method, peptides are identified by searching mass spectrometric data against a six-frame translation protein database. These peptides are then mapped back to the genome, compared to the current annotation, and classified into categories. These peptides, which we term Expressed Peptide Tags (EPTs), are used to confirm, expand, or make novel gene annotations. While previous studies have used somewhat similar approaches, most of them employed bacteria with simple gene architectures, and none used a multi-tiered approach to address the consequences of analysis of eukaryotes with large genomes and complex gene architectures. Although comparing proteomic data to current annotations has been previously proposed and implemented[23, 25], no practical tools have been published to efficiently apply this idea to poorly characterized complex eukaryotes. In this study, a series of measures was taken in order to address challenges in eukaryotic genome annotation, including working definitions of criteria for EPT classification

(*confirming*, *expanding,* or *novel*) for eukaryotic organisms with more complex gene architectures. These definitions will obviously be different for organisms with dramatically different gene densities and structures. It should be noted that multiple EPTs mapped to the shared sequences of gene families may result in incorrect identification of ORFs. This is a common problem of shotgun proteomics that could be addressed by identifying unique peptides from such genes. During this process we explored the generation of sub-databases. Such smaller sub-databases improve the confidence in gene identification, and can also reduce the search time dramatically. Once the sub-databases are generated, subsequent searches with data from new MS/MS experiments on similar biological protein preparation are much quicker than performing searches against the entire six-frame translation protein database. Of course, when new growth conditions or life stages are analyzed, a new sub-database will have to be generated to include additional proteins that are uniquely expressed in those samples. The generation of a sub-database reduces the search space, therefore allowing more relaxed filter thresholds for confident peptide identification compared to the entire six-frame translation. Thus, fewer true peptides are filtered out and the number of confident peptide identifications increases. Also due to the reduced search space, semi-tryptic searches become practical in terms of searching time, therefore allowing identification of this class of peptides. When used in conjunction with existing gene prediction algorithms, ESTs and cDNA sequences, EPTs can provide an additional layer in the accumulated data on which gene prediction is based. In the future, EPTs could be integrated into the automated annotation process in the same fashion that ESTs are currently being used. Although EPT data may not be as unequivocal as EST data, with the use of appropriate filters EPTs can provide a

powerful tool for increasing confidence in automated gene predictions and have the added benefit of yielding functional proteomic data in parallel.

References

1. Eisenstein, M., Light-speed genomics. *Nature methods* **2005,** 2, (9), 646 -646.
2. Smailus DE, M. A., Dextras P, Marra MA, Holt RA., Simple, robust methods for high-throughput nanoliter-scale DNA sequencing. *Genome Res.* **2005,** In print.
3. Mitraa, R. D., Shendurea, J., Olejnikb, J., Edyta-Krzymanska-Olejnikb and Church G.M., Fluorescent in situ sequencing on polymerase colonies. *Analytical biochemistry* **2003,** 320, (1), 55-65.
4. Majoros W.H., P. M., Antonescu C., and Salzberg S.L., GlimmerM, Exonomy and Unveil: three ab initio eukaryotic genefinders. *Nucleic Acids Res* **2003,** 31, (13), 3601-4.
5. Claverie, J. M.; Poirot, O.; Lopez, F., The difficulty of identifying genes in anonymous vertebrate sequences. *Comput Chem* **1997,** 21, (4), 203-14.
6. Reboul J, V. P., Rual JF, Lamesch P, Martinez M, Armstrong CM, Li S, Jacotot L, Bertin N, Janky R, Moore T, Hudson JR Jr, Hartley JL, Brasch MA, Vandenhaute J, Boulton S, Endress GA, Jenna S, Chevet E, Papasotiropoulos V, Tolias PP, Ptacek J, Snyder M, Huang R, Chance MR, Lee H, Doucette-Stamm L, Hill DE, Vidal M, C. elegans ORFeome version 1.1: experimental verification of the genome annotation and resource for proteome-scale protein expression. *Nat Genet* **2003,** 34, (1), 35-41.
7. Mathe, C., Sagot, M., Schiex, T., and Rouze, P., Current methods of gene prediction, their strengths and weaknesses. *Nucleic acid research* **2002,** 30, (19), 4103-4117.
8. Peng, J., Elias, J.E., Thoreen, C.C., Licklider, L.J., and Gygi, S.P., Evaluation of mutidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large scale protein analysis: The yeast proteome. *Journal of proteome research* **2003,** 2, 43-50.
9. Florens, L.; Washburn, M. P.; Raine, J. D.; Anthony, R. M.; Grainger, M.; Haynes, J. D.; Moch, J. K.; Muster, N.; Sacci, J. B.; Tabb, D. L.; Witney, A. A.; Wolters, D.; Wu, Y.; Gardner, M. J.; Holder, A. A.; Sinden, R. E.; Yates, J. R.; Carucci, D. J., A proteomic view of the Plasmodium falciparum life cycle. *Nature* **2002,** 419, (6906), 520-6.
10. Zhu, W., Reich, C.I., Olsen, G.J., Giometti, C.S., Yates, J.R. III, Shotgun proteomics of *Methanococcus jannaschii* and insights into methanogenesis. *J Proteome Res.* **2004,** 3, (3), 538-48.
11. Eriksson. J., F., D., Protein identification in complex mixtures. *J Proteome Res.* **2005,** 4, (2), 387-93.
12. Xiang, R., Shi, Y., Dillon, D.A., Negin, B., Horvath, C., Wilkins, J.A., 2D LC/MS analysis of membrane proteins from breast cancer cell lines MCF7 and BT474. *J Proteome Res.* **2004,** 3, (6), 1278-83.
13. VerBerkmoes, N. C., Connelly, H.M., Pan, C., and Hettich, R.L., Mass spectrometric approaches for characterizing bacterial peoteomes. *Expert Rev Proteomics* **2004,** 1, (4), 433-47.
14. Chong, P. K., Wright, P.C., Identification and characterization of the Sulfolobus solfataricus P2 proteome. *J Proteome Res.* **2005,** 4, (5), 1789-98.

15. Delahunty, C., Yates, J.R. III, Protein identification using 2D-LC-MS/MS. *Methods* **2005,** 35, (3), 248-255.
16. Eng, J. K., McCormack, A.L., Yates, J.R., III, An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **1994,** 5, 976-989.
17. Yates, J. R. I., Eng, J.K., McCormack, A.L., Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Analytical chemistry* **1995,** 67, (18), 3202-10.
18. Perkins, D. N., Pappin, J.C., Creasy, D.M., Cottrell, J.S., *Electrophoresis* **1999,** 20, 3551-3567.
19. Tabb, D. L.; Narasimhan, C.; Strader, M. B.; Hettich, R. L., DBDigger: reorganized proteomic database identification that improves flexibility and speed. *Anal Chem* **2005,** 77, (8), 2464-74.
20. Narasimhan, C., Tabb, D.L., Verberkmoes, N.C., Thompson, M.R., Hettich, R.L., Uberbacher, E.C., MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Analytical chemistry* **2005,** 77, (23), 7581-93.
21. Allmer, J., Markert, C., Stauber, E.J., and Hippler, M., A new approach that allows identification of intron-split peptides from mass spectrometric data in genomic databases. *FEBS Letters* **2004**, (562), 202-206.
22. Arthur, J. W., and Wilkins, M.R., Using Proteomics to Mine Genome Sequences. *Journal of proteome research* **2004,** 3, 393-402.
23. Carlton, J. M., Angiuoli, S.V., Suh, B.B., Kooij, T.W., Pertea, M., Silva, J.C., Ermolaeva, M.D., Allen, J.E., Selengut, J.D., Koo, H.L., Peterson, J.D., Pop, M., Kosack, D.S., Shumway, M.F., Bidwell, S.L., Shallom, S.J., van Aken, S.E., Riedmuller, S.B., Feldblyum, T.V., Cho, J.K., Quackenbush, J., Sedegah, M., Shoaibi, A., Cummings, L.M., Florens, L., Yates, J.R., Raine, J.D., Sinden, R.E., Harris, M.A., Cunningham, D.A., Preiser, P.R., Bergman, L.W., Vaidya, A.B., van Lin, L.H., Janse, C.J., Waters, A.P., Smith, H.O., White, O.R., Salzberg, S.L., Venter, J.C., Fraser, C.M., Hoffman, S.L., Gardner, M.J., Carucci, D.J., Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. *Nature* **2002,** 419, (6906), 512-519.
24. Giddings, M. C., Shah, A.A., Gesteland, R., and Moore, B., Genome-based peptide fingerprint scanning. *Proc Natl Acad Sci U S A* **2003,** 100, (1), 20-25.
25. Jaffe, J. D., Berg, H.C., Church, G.M., Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **2004,** 4, 59-77.
26. Küster, B., Mortensen, P., Andersen, J.S., and Mann, M., Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **2001,** 1, 641–650.
27. Smith, J. C., Northey, J.G.B., Garg, J., Pearlman, R.E., and Michael Siu, K.W., Robust method for proteome analysis by MS/MS using an entire translated genome: Demonstration on the ciliome of Tetrahymena thermophila. *Journal of proteome research* **2005**.
28. Kalume DE, P. S., Reddy R, Zhong J, Okulate M, Kumar N, Pandey A, Genome annotation of Anopheles gambiae using mass spectrometry-derived data. *BMC Genomics* **2005,** 6, 128.

29. Resing, K. A., Meyer-Arendt, K., Mendoza, A.M., Aveline-Wolf, L.D., Jonscher, K.R., Pierce, K.G., Old, W.M., Cheung, H.T., Russell, S., Wattawa, J.L., Goehle, G.R., Knight, R.D., Ahn, N.G., Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics. *Analytical chemistry* **2004,** 76, (13), 3556-68.
30. Colinge J, C. I., Reffas S, Mahe E, Niknejad A, Rey PA, Mattou H, Moniatte M, Bougueleret L., Experiments in searching small proteins in unannotated large eukaryotic genomes. *J Proteome Res.* **2005,** 4, (1), 167-174.
31. Desiere F, D. E., Nesvizhskii AI, Mallick P, King NL, Eng JK, Aderem A, Boyle R, Brunner E, Donohoe S, Fausto N, Hafen E, Hood L, Katze MG, Kennedy KA, Kregenow F, Lee H, Lin B, Martin D, Ranish JA, Rawlings DJ, Samelson LE, Shiio Y, Watts JD, Wollscheid B, Wright ME, Yan W, Yang L, Yi EC, Zhang H, Aebersold R., Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry. *Genome Biol.* **2005,** 6, (1), R9.
32. McGowan, Annotation of the Human Genome by High-Throughput Sequence Analysis of Naturally Occurring Proteins. *Current Proteomics* **2004,** 1, 41-48.
33. Nesvizhskii AI, R. F., Grossmann J, Vogelzang M, Eddes JS, Gruissem W, Baginsky S, Aebersold R., Dynamic spectrum quality assessment and iterative computational analysis of shotgun proteomic data: toward more efficient identification of post-translational modifications, sequence polymorphisms, and novel peptides. *Mol Cell Proteomics* **2006,** 5, (4), 652-670.
34. Larimer, F. W., Chain, P., Houser, L., Lamerdin, J., Malfatti, S., Do, L., Land, M.L., Pelletier, D.A., Beatty, J.T., Lang, A.S., Tabita, F.R., Gibson, J.L., Hanson, T.E., Bobst, C., Torres y Torres, J.L., Peres, C., Harrison, F.H., Gibson, J., and Harwood, C.S., Complete genome sequence of the metabolically versatile photosynthetic bacterium Rhodopseudomonas palustris. *Nature Biotechnology* **2004,** 22, 55-61.
35. Tyler, B. M., Molecular Basis of Recognition Between Phytophthora Pathogens and their Hosts. *Annu. Rev. Phytopathol.* **2002,** 40, 137-67.
36. Verberkmoes, N. C., Shah, M.B., Lankford, P.K., Pelletier, D.A., Strader, M.B., Tabb, D.L., McDonald, W.H., Barton, J.W., Hurst, G.B., Hauser, L., Davison, B.H., Beatty, J.T., Harwood, C.S., Tabita, F.R., Hettich, R.L., Larimer, F.W., Determination and Comparison of the Baseline Proteomes of the Versatile Microbe Rhodopseudomonas palustris under Its Major Metabolic States. *J Proteome Res.* **2006,** 5, (2), 287-298.
37. Erwin, D. C., and Ribeiro, O. K., *Phytophthora Diseases Worldwide*. The American Phytopathological Society: St. Paul, MN, 1996; p 562.
38. McDonald, W. H.; Ohi, R.; Miyamoto, D.; Mitchison, T. J.; Yates, J. R., Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimensional LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. *Int J Mass Spectrometry* **2002,** 219, 245-251.
39. Ram, R. J., Verberkmoes, N.C., Thelen, M.P., Tyson, G.W., Baker, B.J., Blake, R.C. 2nd, Shah, M., Hettich, R.L., Banfield, J.F., Community proteomics of a natural microbial biofilm. *Science* **2005,** 308, (5730), 1915-20.
40. Washburn, M. P., Wolters, D., and Yates III, J.R., Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology* **2001,** 19, 242-247.

41. Wolters, D. A., Washburn, M.P., and Yates III, J.R., An automated multidimensional protein identification technology for shotgun proteomics. *Analytical chemistry* **2001,** 73, (23), 5683-90.
42. Tabb, D. L.; Shah, M. B.; Strader, M. B.; Connelly, H. M.; Hettich, R. L.; Hurst, G. B., Determination of Peptide and Protein Ion Charge States by Fourier Transformation of Isotope-Resolved Mass Spectra. *Journal of the American Society for Mass Spectrometry* **2006,** 17, (7), 903-915.
43. Finley, L. a., A strategy for recovering high quality genomic DNA from a large number of Phytophthora isolates. *Mycologia* **2006,** In press.
44. MacCoss, M. J., Computational analysis of shotgun proteomics data. *Current Opinion in Chemical Biology* **2005,** 9, (1), 88-94.
45. Moore, R. E., Young, M.K., Lee, T.D., Qscore: an algorithm for evaluating SEQUEST database search results. *J. Am. Soc. Mass Spectrom.* **2002,** 13, (4), 378-86.
46. Craig R, B. R., A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom* **2003,** 17, (20), 2310-6.

Chapter 3

Cross-species global proteomics reveals conserved and unique processes in *Phytophthora sojae* and *P. ramorum*

Published as Savidor, A.; Donahoo, R. S.; Hurtado-Gonzales, O.; Land, M. L.; Shah, M. B.; Lamour, K. H.; McDonald, W. H., Cross-species global proteomics reveals conserved and unique processes in Phytophthora sojae and P. ramorum. *Mol Cell Proteomics* **2008**.

Abstract

Phytophthora ramorum and *Phytophthora sojae* are destructive plant pathogens. *Phytophthora sojae* has a narrow host range whereas *P. ramorum* has a wide host range. A global proteomic comparison of the vegetative (mycelium) and infective (germinating-cyst) life-stages of *P. sojae* and *P. ramorum* was conducted to identify candidate proteins involved in host range, early infection and vegetative growth. Sixty-two candidates for early infection, 26 candidates for vegetative growth, and numerous proteins that may be involved in defining host specificity were identified. In addition, common life stage proteomic trends between the organisms were observed. In mycelia, proteins involved in transport and metabolism of amino acids, carbohydrates and other small molecules were up-regulated. In the germinating cysts, up-regulated proteins associated with lipid transport and metabolism, cytoskeleton and protein synthesis were observed. It appears that the germinating cyst catabolizes lipid reserves through the β-oxidation pathway to drive the extensive protein synthesis necessary to produce the germ tube and initiate infection. Once inside the host, the pathogen switches to vegetative growth, where energy is derived from glycolysis and utilized for synthesis of amino acids and other molecules that assist survival in the plant tissue.

Introduction

Organisms of the genus *Phytophthora* are destructive plant pathogens capable of infecting many agriculturally and ornamentally important crops[1]. To date, there are over 80 recognized species of *Phytophthora*. While some *Phytophthora* species are able to infect a broad range of host plants, others are limited to a single host. *Phytophthora ramorum* and *P. sojae* are examples of each of these groups, respectively. The recently characterized *P. ramorum* is the causal agent of sudden oak death disease[2, 3]. In addition to its destructive effect on live oak trees, as currently unfolding in forests in California and Oregon, *P. ramorum* is capable of infecting a wide range of trees and shrubs, such as bay laurel and viburnum[4]. On the other hand, *P. sojae*, the causal agent of soybean root and stem rot, has a very narrow host range. Races of *P. sojae* are cultivar specific; with certain *P. sojae* isolates only infecting certain varieties of soybean[1].

The genomes of *P. sojae* and *P. ramorum* have recently been sequenced[5]. The genome of *P. sojae* is 95 Mbp and predicted to encode 19,027 genes while *P. ramorum*'s genome is 65 Mbp, and predicted to encode 15,743 genes. When compared to one another, the two organisms have a high degree of orthology and synteny between their genomes [5, 6]. Thus, despite the similarity between the two organisms, there are obviously specific differences that define the organisms in terms of their unique properties. One aim of this study was to identify candidate proteins that might be involved in host-range capacity. This was achieved by comparing the expressed proteomes of *P. sojae* and *P. ramorum* and identifying differences between them.

Phytophthora species are capable of reproducing sexually and asexually. During the sexual life cycle oospores, the sexual spores, are produced. Oospores have thick cell

walls and can survive in the soil for years, thus allowing re-infection of their host plant in subsequent growing seasons. However, since the oospores require a dormancy period of several weeks before germination, it is the asexual life cycle that is responsible for rapid propagation and spread of the disease in the field or forest. During the asexual life cycle, the organism is able to differentiate into different life stages including mycelium, sporangium, zoospore, cyst and germinating cyst (Figure. 3.1).

Infection of a plant can be initiated when a zoospore, a motile kidney shaped bi-flagellated cell, interacts with a compatible plant host. Upon contact with the plant host the zoospore sheds its flagella, encysts and adheres. Shortly after encystment the cyst germinates by producing a germ tube that penetrates the host tissue directly or through wounds or natural openings. Once inside the plant, the pathogen grows and ramifies through the plant tissue as mycelium, the vegetative growth life stage of *Phytophthora*. In a compatible interaction the initial growth of the mycelium in the plant tissue is biotrophic, where the pathogen evades the plant defense responses. Later, growth is switched to necrotrophic and the plant tissue is destroyed. A second aim of this study was to identify candidate proteins involved in initiation of infection. This was achieved by comparing the proteomes of the germinating cyst and the mycelium life stages for both organisms.

Proteomic investigation of the mycelium and germinating cyst life stages from *P. sojae* and *P. ramorum* was carried out using multidimensional protein identification technology (MudPIT)[7]. As compared to other proteomic techniques such as two-dimensional electrophoresis (2-DE) which typically identifies a few tens of proteins, MudPIT is superior, identifying hundreds and even thousands of proteins in a single

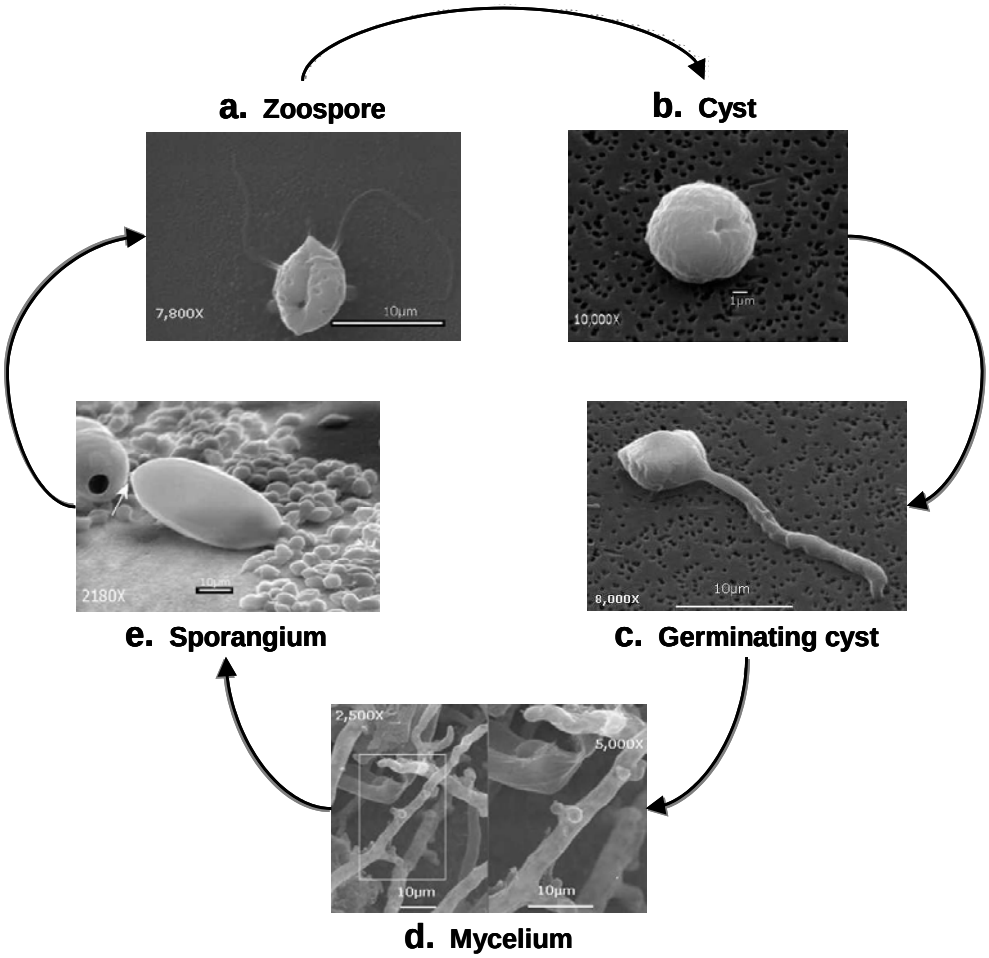


Figure 3.1- The asexual life cycle of *Phytophthora*. Scanning electron micrographs of asexual life stages of *P. ramorum*. A motile bi-flagellated zoospore **(a)** swims chemotactically towards its plant host. Upon contact with the plant (or shaking in the lab) the zoospore sheds its flagella and encysts **(b)**. The cyst germinates **(c)**, and the germ tube penetrates into the plant tissue where it starts growing vegetatively through the plant tissue as mycelium **(d)**. The mycelium can grow out of the plant tissue where, under appropriate stimulation, its terminal ends can differentiate into sporangia **(e)**, which are structures containing multiple nuclei which can differentiate into zoospores. Zoospores can then be released from the sporangia and repeat the cycle. Micrographs were supplied courtesy of Dr. Edwin R. Florance, Lewis & Clark College, Portland, OR.

experiment. Such a strategy for proteomic investigation has been used extensively on many organisms (i.e. [7-10]). Several groups compared the identified proteomes of an organism grown under different treatments (comparative proteomics) in order to elucidate biological responses to these treatments[11, 12]. Comparative proteomics was also applied to different life stages of eukaryotic organisms in order to gain insight to their life cycle[13, 14]. Here we present an approach incorporating comparative proteomics with orthology to investigate cellular processes in closely related organisms. The rationale is that orthologous proteins that are up-regulated in a certain life-stage in both organisms may be specifically involved in that life-stage; whereas proteins with no similarly expressed orthologs, or with no orthologs at all, may contribute specific functions to that organism. In our case, orthologous proteins that were up-regulated in germinating-cyst both in *P. sojae* and *P. ramorum* were identified as candidates involved in cyst germination, and therefore may be involved in early infection. Orthologous protein pairs that were up-regulated in mycelium were identified as candidates important for vegetative growth. Proteins expressed uniquely in *P. ramorum* were identified as candidates allowing its broad host-range capacity, while uniquely expressed *P. sojae* proteins were identified as candidates for defining its narrow host-range. In addition, the methodological approach presented here includes use of orthology and a variety of novel bioinformatic strategies to analyze the data and represents a step forward in the analysis of proteomic data from closely related species.

Materials and Methods

Phytophthora cell culture, life stage isolation, lysate preparation and LC/LC-MS/MS analysis

Cell tissue was prepared as previously described[15]. Briefly, mycelia were generated by growing *P. sojae* strain P6497 (University of California, Riverside, U.S.A.) and *P. ramorum* strain LT191 (University of California, Riverside, U.S.A.) in clarified antibiotic amended V8 juice broth. Asexual sporangia and zoospores were generated by growing *P. sojae* and *P. ramorum* on antibiotic amended V8 agar plates according to standard protocols[1]. Zoospores were stimulated to encyst by vigorous shaking for 1-2 minutes and then waiting approximately 1 hour before harvesting the germinating cysts. The mycelium and germinating cyst were freeze-dried and ground using glass beads.

Lysates were fractionated to membrane and soluble fractions by centrifugation and the soluble fractions were digested with trypsin after denaturation, reduction and alkylation as previously described[15]. Filtered and acidified samples were then loaded offline on a bi-phasic 150 μm inner-diameter fused silica column containing 3-3.5 cm C18 reverse phase material followed by 3.5-4 cm strong cation exchange phase. Loaded columns were placed directly upstream of a 100 μm inner-diameter front column packed with 15cm C18 reverse-phase material. Samples were analyzed by 11-step MudPIT (LC/LC-MS/MS) on a linear ion trap mass spectrometer (LTQ, ThermoElectron, San Jose, CA) as previously described[15]. Flow rate at the tip was ~200-300 nl/min.

Informatics

MS/MS spectra from RAW files of MudPIT experiments were extracted into MS2 files using Raw2MS2.exe[16, 17]. Spectra were then assigned charge using MS2ZAssign[17], and searched against the predicted protein databases using the DBDigger algorithm[18] in fully tryptic mode with the MASPIC scorer[17]. No limit on missed cleavages was specified. The *P. sojae* protein database used was Psojae_proteins_VMD_V-1.0, release date-7/29/2004 and included 19,027 *P. sojae* proteins as well and 52 common contaminants. The *P. ramorum* protein database used was Pramorum_proteins_VMD_V-1.0, release date-7/29/2004 and included 15,743 *P. ramorum* proteins as well and 52 common contaminates. Search results were filtered with the DTASelect algorithm[19] using filter values allowing 5% false discovery rate (Table S3.1) based on a concatenated reversed protein database search, as previously described[15]. Briefly, spectra from one representative MudPIT experiment for each organism was searched against a protein database containing the respective organism's protein sequences, the contaminants, and the reversed sequence of both proteins and contaminants (essentially representing random entries with the same amino acid composition and length). Filter values allowing 1 hit to the reverse database for every 19 to the forward database (5% false discovery rate) were determined using the SQTRevPuller algorithm[20], and applied thereafter to all search results for the same organism. For all searches, a fixed modification of +57 Da on cysteine residues was included as cysteines were considered to be fully carboxiamidomethlyated. No variable modifications were included. Protein identification required 2 peptides per protein for identification. Mass tolerance for precursor ions was 3 Da, and mass tolerance for fragment ions was 0.5 Da. Comparison between protein lists

was carried out using Microsoft Access and PERL scripts. To compare between life stages, the union of the duplicate MudPIT results of the germinating cyst was compared to the union of the duplicate MudPIT results of the mycelium from the same organism.

Spectral count was considered as the total number of MS/MS spectra corresponding to peptides from a given protein. Normalization of spectral counts was done using Normalized Spectral Abundance Factors (NSAF) as previously described[21]. Proteins groups that shared all of their peptides were assigned a NSAF value equal to the NSAF value for the group divided by the number of proteins in the group. A protein was determined to be up-regulated in one life stage versus another if (i) it was identified in both MudPIT duplicates of the up-regulated life stage and (ii) the sum of the normalized spectral counts from both duplicates was at least 5 times higher than the sum of normalized spectral counts in the other life stage duplicates. For statistical significance, a G-test was set up in Microsoft Excel such that for each protein the sums of normalized spectral counts in each life stage were tested against the null hypothesis that they are not different from an expected 1:1 ratio, with a two-tailed P-value<0.05.

Gene Onthologies (GO) annotations as well as Eukaryotic Orthologous Groups (KOG) annotations were downloaded from the JGI website (http://www.jgi.doe.gov/) into a Microsoft Access relational database and associated with their corresponding protein IDs. Signal peptide assignment was done using a local copy of SignalP (http://cbs.dtu.dk/services/SignalP/).

Orthology was determined by a reciprocal BLASTp of the *P. sojae* and *P. ramorum* protein databases. A pair of proteins was assigned orthology if both proteins were the best hit on their partner's BLAST results and had an E value $<10^{-50}$ (Table

S3.2). Differential expression between orthologous proteins was determined as described above for a protein in different life stages.

To compare between the two organisms, the union of the duplicate MudPIT results of each life stage from *P. ramorum* was compared, based on orthology, to the union of duplicate MudPIT results from the same life stage in *P. sojae*. For identification of host-range specific candidates, the raw lists of organism specific proteins (Tables S3.3 and S3.4) were manually examined. Steps that were taken to identify host-range specific candidate from the raw lists included (a) High spectral count for confident protein identification and abundance (b) Observation of KOG and GO annotation (c) BLASTp against the other organism protein database for identification of homologs with similar expression (d) BLASTp against the non-redundant protein database for identification of homologs in other species (e) ClustalW alignment with homologous proteins to confirm homology (f) Pfam analysis for domain identification (g) Literature search.

All raw data and supporting analysis are available at: http://compbio.ornl.gov/mspipeline/phyto_comparison. Supplementary material is available through the Molecular & Cellular Proteomics Journal.

Results and Discussion

Comparison of the global proteomes of germinating cyst and mycelium life stages

Infection of a plant host by *Phytophthora* can be initiated when a *Phytophthora* zoospore adheres and encysts on the plant tissue, produces a germ tube, gains access to

the plant host, and starts growing vegetatively throughout the plant tissue as mycelium. We investigated the proteomic differences between the germinating cyst and mycelium life stages from *P. sojae* and *P. ramorum* by MudPIT in order to identify candidate proteins that might be involved in early infection. Duplicate MudPIT experiments were carried out on each life stages from *P. ramorum* and *P. sojae*. All together, 3897 *P. ramorum* and 2970 *P. sojae* proteins were identified. In the germinating cyst 2065 and 2089 *P. ramorum* proteins were identified respectively in the first and second replicates (Tables S3.5-S3.6), with 1293 of them found in both replicates (63% reproducibility). In the *P. ramorum* mycelium 1966 and 1962 proteins were identified in the first and second replicates (Tables S3.5-S3.6), with 1339 of them found in both replicates (68% reproducibility). For *P. sojae* 1940 and 1779 germinating cyst proteins were identified respectively in the first and second replicates (Tables S3.5, S3.7), with 1248 of them found in both replicates (70% reproducibility). In the *P. sojae* mycelium 1313 and 1216 proteins were identified in the first and second replicates (Tables S3.5, S3.7), with 825 of them found in both replicates (68% reproducibility). It should be noted that the major factor in reducing reproducibility comes from identification of low abundance proteins from which two peptides happened to be identified in one replicate but not the other. For abundant proteins the reproducibility rate is much higher than stated above.

One group of proteins that were identified in both life stages of both organisms included superoxide dismutases, catalases and peroxidases. These enzymes break down reactive oxygen species (ROS)[22-25] that are produced as by-products of aerobic respiration, as well as by plants defense mechanisms in response to pathogens[26-28]. It has been shown that pathogenesis of certain bacteria and fungi can be correlated to their

ROS-inactivating enzymes[29-31]. Thus, constituent expression of these proteins in *Phytophthora* may contribute to their success as plant pathogens.

On average, 14% of all identified proteins in each organism had a predicted signal peptide, and are thus likely secreted or membrane proteins. This number of secreted or membrane proteins is high, especially considering that only an estimated 8% of all predicted genes in the genome are predicted to be secreted[5, 6], and that the protein preparations were not designed specifically to contain such proteins (only crude soluble fractions of total lysates were analyzed). Among others, these signal-peptide-containing-proteins included peptidases and other hydrolases, chaperones, transporters, and structural molecules. Among the secreted proteins, 42 *P. sojae* and 46 *P. ramorum* proteins contained the RxLR motif within the first 30 to 60 residues of the N-terminal, possibly classifying them as members of the RxLR family; a group of effector proteins that is thought to be targeted to the host cell cytosol, where they manipulate host defense responses[32-34].

Relative protein expression

For relative protein expression analysis, protein abundance was determined based on normalized spectral counts (see material and methods). Spectral count has been shown to provide a reliable parameter for estimation of relative protein abundance in proteomes up to at least 3 orders of magnitude[35]. Normalization of spectral count according to published protocols[21] was done in order to meaningfully compare protein abundances across different experiments. Within the same experiment normalization minimizes the bias of higher spectral count number towards longer proteins, and between different

experiments it minimizes differences that may arise due to variation in experimental procedures (initial protein concentration in each sample, protein digestion efficiency, instrument performance, etc.). Each identified protein was associated with its GO and KOG annotations, and proteins were clustered into functional groups based on their KOG annotation. The proportion of each functional category of the total was calculated based on the sum of normalized spectral counts of all proteins in each category. A breakdown of the identified proteomes of the mycelium and germinating cyst life stages of *P. ramorum* and *P. sojae* are shown in figure 3.2. Very few functional categories changed significantly and the proportion of most categories remained fairly constant between germinating-cyst and mycelium (Figure 3.2). These results indicate that while some distinct processes are taking place in each life stage, other common functions are being carried out in both.

In order to focus on cellular processes that are unique to each of the life stages, subset proteomes were compiled that contained only those proteins that were differentially expressed between germinating cyst and mycelium. A protein was considered to be up-regulated in a specific life stage if its abundance, based on normalized spectral counts, was at least 5 fold higher than in the other life stage (see informatics section in material and methods). 686 proteins were determined to be differentially expressed between germinating cyst and mycelium for *P. ramorum* (Table S3.8, Figure 3.3) and 513 for *P. sojae* (Table S3.9, Figure 3.3). A G-test on normalized spectral counts between the life-stages identified a statistical significant difference for all of these suggested differentially expressed targets (with a P-value<0.05) as well as other proteins not passing our stringent criteria (data not shown). When examining only the

Figure 3.2- Distribution of protein functional categories in the global proteomes of the germinating cyst and mycelium of *P. ramorum* and *P. sojae*. Each identified protein was associated with its KOG annotation and protein categories were clustered based on their KOG functional annotation. The proportion of each protein of the total was calculated based on Normalized Spectral Abundance Factors (NSAF) as previously described[21]. **(a)** Each pie chart represents the union of duplicate MudPIT experiments. The numbers in parenthesis represent the number of proteins in the respective functional category. **(b)** and **(c)** For comparison purposes, all functional categories with >5% of the total are presented as column charts.

(a)

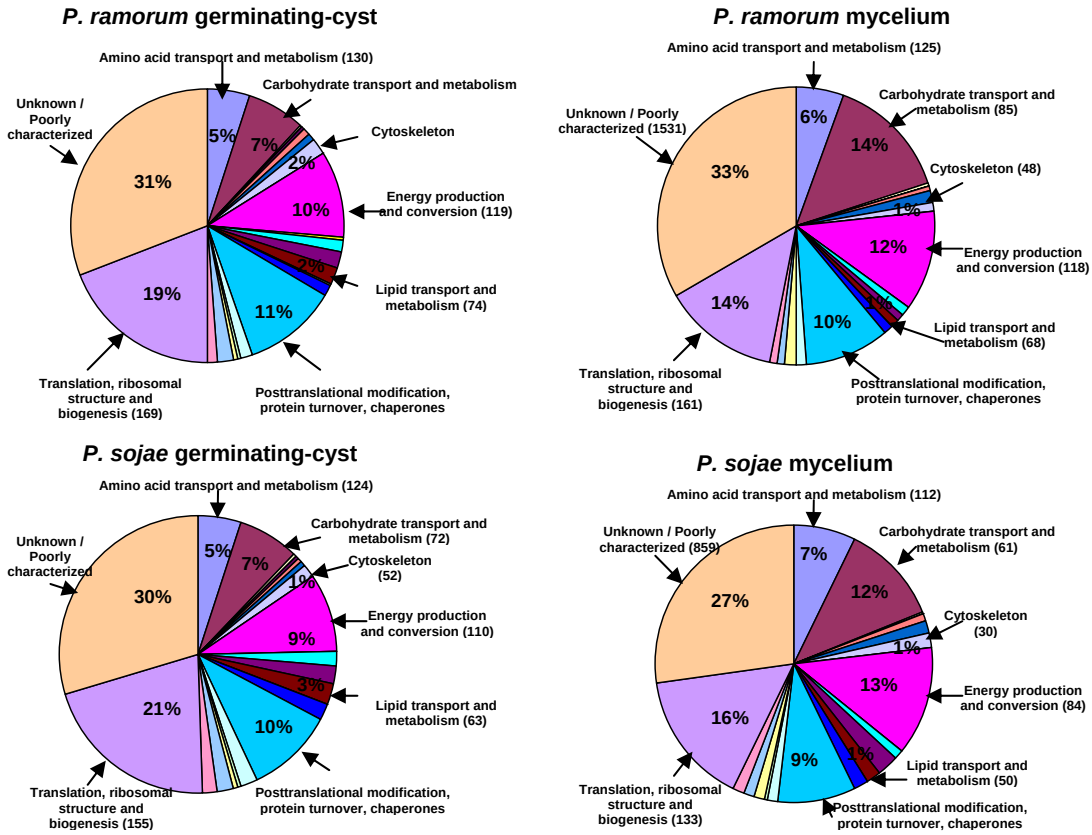
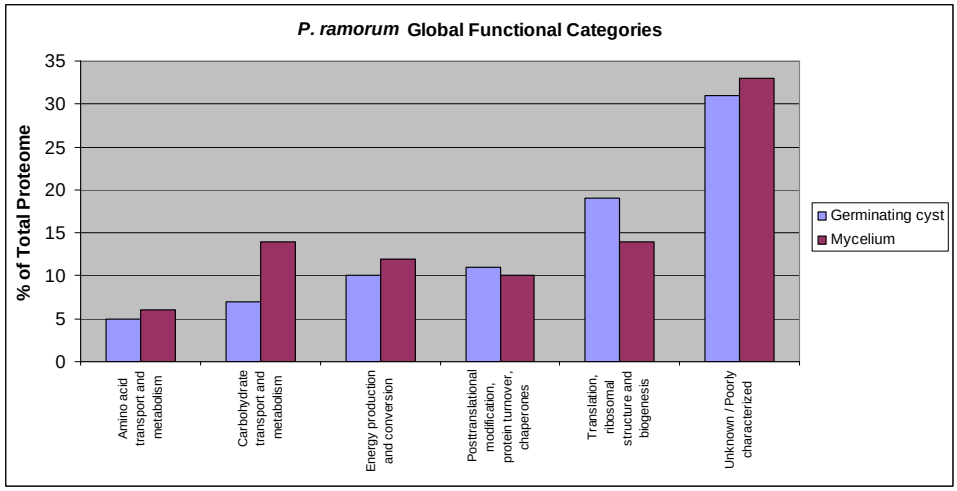


Figure 3.2- continued

(b)



(c)

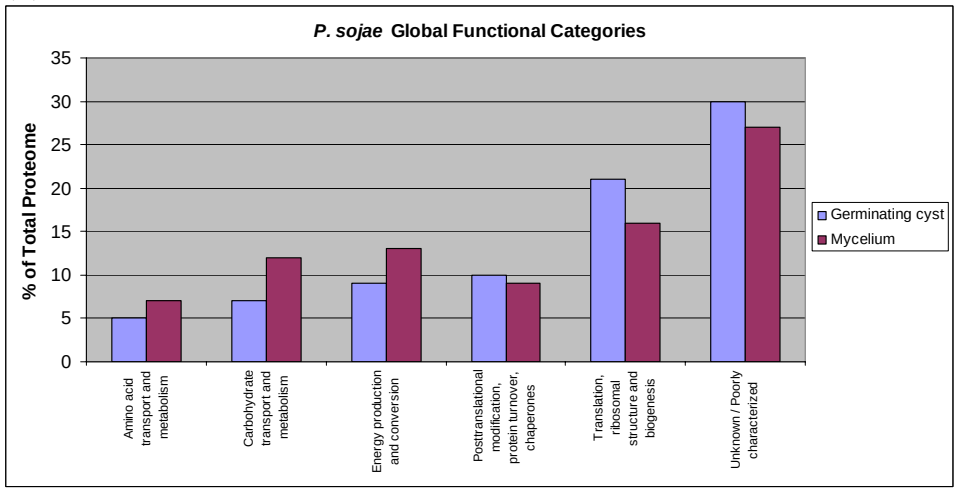


Figure 3.2- continued

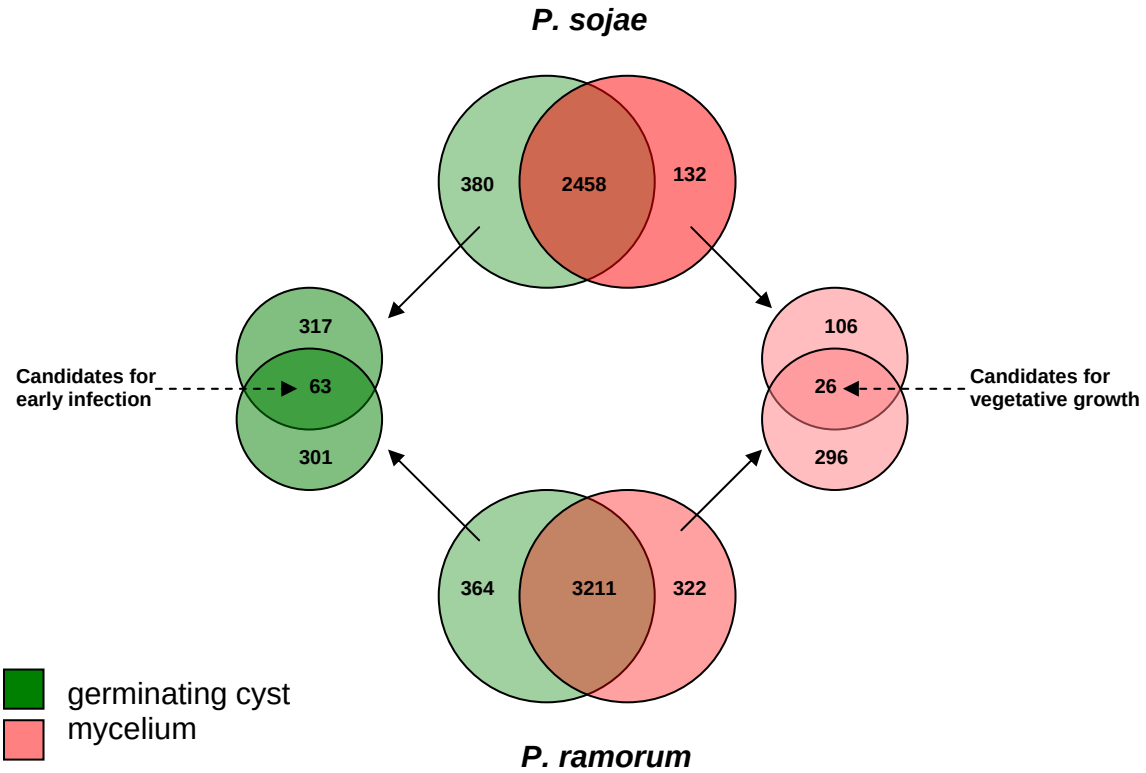


Figure 3.3- Identification of candidate proteins for involvement in early infection or vegetative growth. For each organism, the total identified proteomes of the germinating cyst and mycelium were compared (top and bottom overlapping circles). The non-overlapping regions represent the differentially expressed proteins in each life stage, while the overlapping regions represent the rest of the identified proteins. Orthologous proteins that were up-regulated in the germinating cyst life stage of both organisms (overlapping region, middle left) were identified as candidates for early infection. Orthologous proteins that were up-regulated in the mycelium life stage of both organisms (overlapping region, middle right) were identified as candidates for vegetative growth.

Figure 3.4- Distribution of protein functional categories in the differentially expressed proteomes of the germinating cyst and mycelium of *P. ramorum* and *P. sojae*. Each identified protein was associated with its KOG annotation and protein categories were clustered based on their KOG functional annotation. The proportion of each protein of the total was calculated based on Normalized Spectral Abundance Factors (NSAF) as previously described[21]. Only proteins that were determined to be differentially expressed (see materials and methods) were included in this analysis. **(a)** Each pie chart represents the union of duplicate MudPIT experiments. The numbers in parenthesis represent the number of proteins in the respective functional category. **(b)** and **(c)** For comparison purposes, all functional categories with >5% of the total (with the exception of the "Unknown" category) are presented as column charts.

(a)

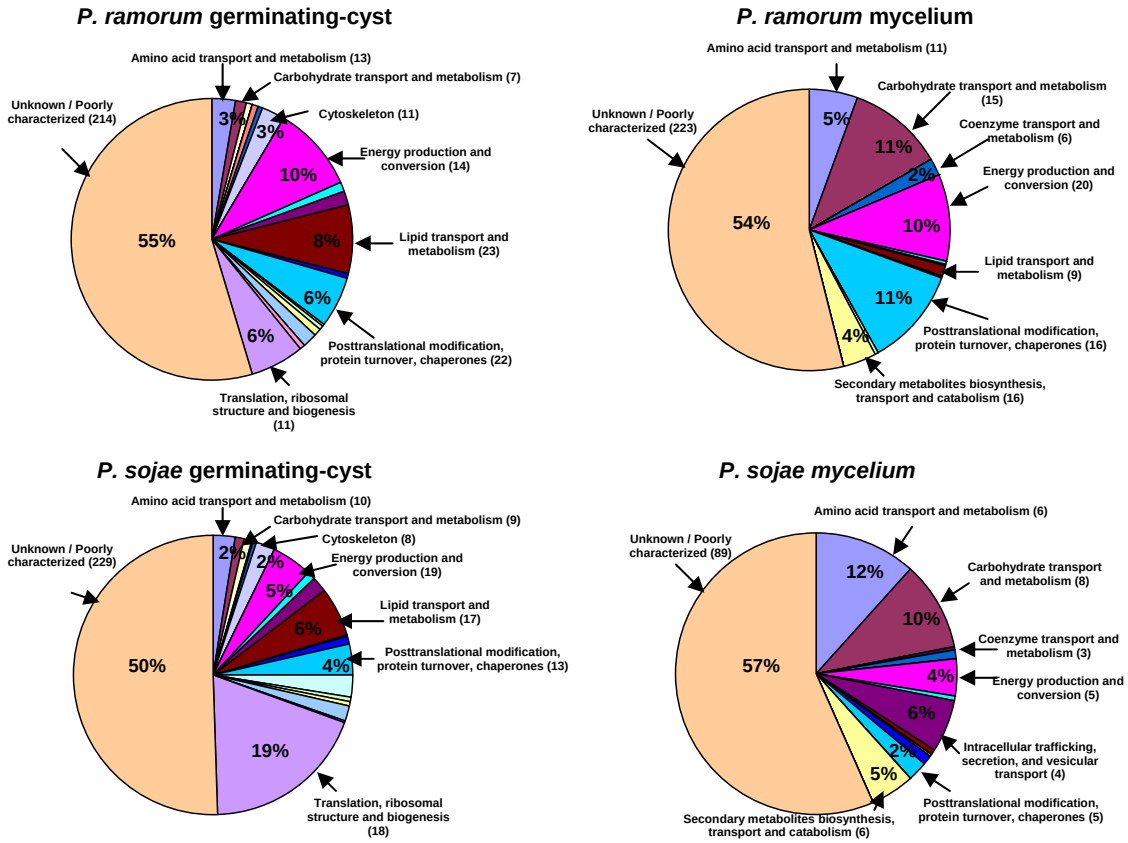
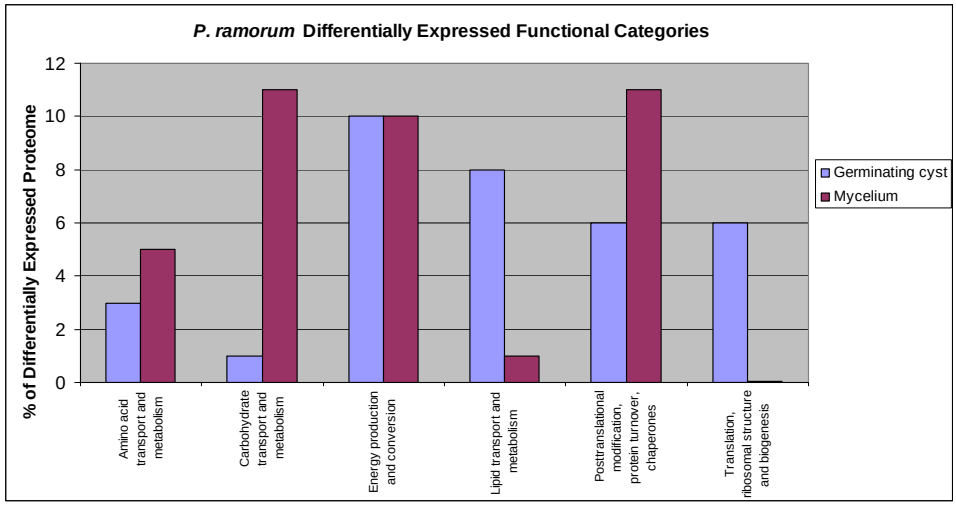


Figure 3.4- continued

(b)



(c)

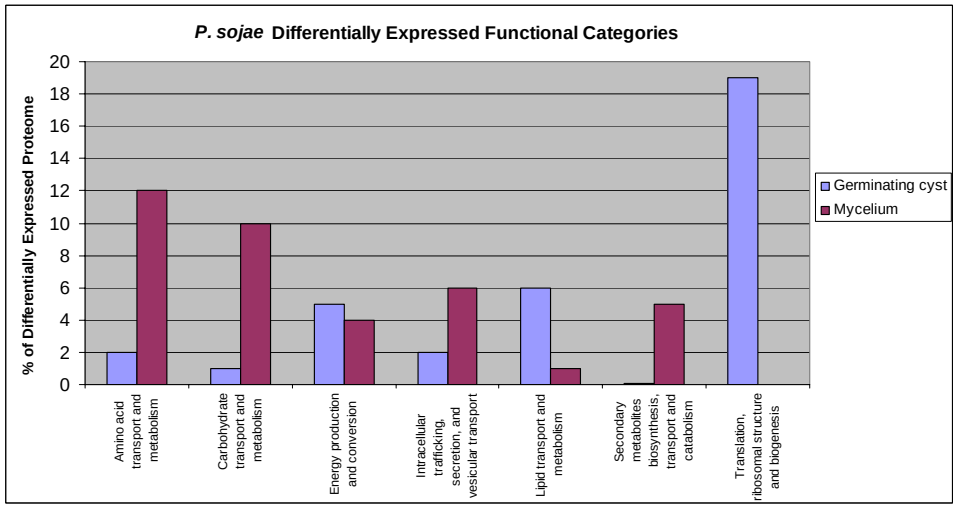


Figure 3.4- continued

differentially expressed proteins, the proteomic differences between germinating cyst and mycelium become more apparent (Figure 3.4). Thus, ignoring the common background (i.e. house keeping proteins with equivalent abundance in both life stages) aids in emphasizing the differences between the two. Interestingly, the proportion of unknown proteins increases dramatically in the differentially expressed protein sets as compared to the global proteome analysis (compare figure 3.2 to figure 3.4), suggesting that many life-stage specific proteins have yet to be characterized in these or other related organisms.

Germinating cyst proteome

In the germinating-cyst life stage from both *P. sojae* and *P. ramorum*, three KOG categories showed a substantial increase in expression as compared to mycelium: *Translation, ribosomal structure and biogenesis* (protein synthesis), *Cytoskeleton*, and *Lipid transport and metabolism* (Figure 3.4). Up regulated proteins in the *Lipid transport and metabolism* category included an array of proteins spanning the entire β-oxidation pathway, which is the process of fatty acid catabolism[36] (Figure 3.5). These include acyl-coA synthases (activates fatty acids), carnitine O-acyl transferase (transports peroxisomal activated fatty acids into the mitochondria), acyl-coA dehydrogenase (catalyzes fatty acid oxidation), and hydroxyacyl-coA dehydrogenase (subsequent oxidation of fatty acids), among others. Based on this, it appears that the source of energy production prior to penetration and harvesting of host resources is via lipid catabolism. A previous study in *Phytophthora infestans* found one component of this pathway, acyl-CoA synthetase, to be up-regulated in germinating-cyst[37], leading the authors to hypothesize lipid metabolism

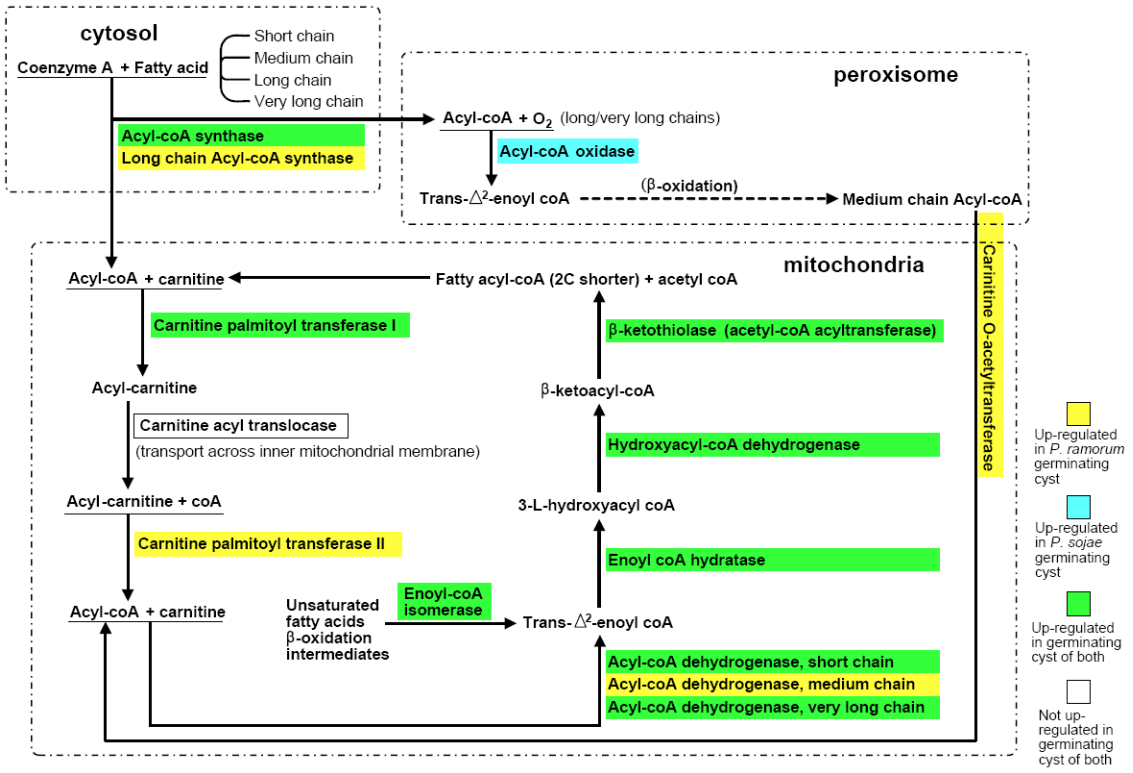


Figure 3.5- Lipid catabolism via the β-oxidation pathway.
Multiple enzymes involved in the catabolism of fatty acids were found to be up-regulated in the germinating cyst of *P. sojae* and *P. ramorum*. For simplicity, the cofactors, reducing agents, water molecules, ATP molecules etc. were excluded from the diagram, leaving only the proteins and substrates. The diagram is based on Kunau et al.[36].

as one of several possible energy production resources. Considering this, our results suggest that germinating-cyst β-oxidation may be applicable across the entire genus. While it occurs in peroxisomes and mitochondria of mammals and some algae, β-oxidation occurs exclusively in peroxisomes of plants and fungi[36, 38]. Both peroxisomal and mitochondrial specific proteins were identified, suggesting β-oxidation in both organelles. Thus, β-oxidation of fatty acids in both organelles suggests a closer evolutionary connection of *Phytophthora* to animals than plants or fungi.

An increase in the *Translation, ribosomal structure and biogenesis* and *Cytoskeleton* categories is consistent with the process of cyst germination. During this process, the cyst produces a long germ-tube that, during infection, penetrates host plant tissue. Presumably, production of the germ tube requires large amounts of cytoskeleton proteins, as well as other proteins, and thus a concomitant increase in the mentioned categories is expected. Supporting this notion is the fact that in both organisms several proteins associated with microtubule-based movement were up-regulated in the germinating cyst life stage, but were absent or down regulated in mycelium. These proteins may be involved in shuttling specific proteins destined for interaction with the host to the extremities of the germ tube. An increase in the *Translation, ribosomal structure and biogenesis* category may also represent ramping of protein production in the transition from a cyst to vegetative growth as mycelium. These results are consistent with previous studies[37, 39-41] which suggested up-regulation of de-novo RNA and protein synthesis in the germinating cyst of other *Phytophthora* species.

Notably, lipid metabolism via the peroxisomal β-oxidation pathway has been shown to be essential for *Arabidopsis* seedling growth[42-45]. Components of the β-

oxidation pathway have also been shown to be involved in the control of seed germination in *Arabidopsis*[46, 47]and have been shown to poses a critical role in appresorium (the pre-infective structure produced by the germinating cyst) function of the fungus *Magnaporthe grisea*[48]. Cytoskeleton proteins such as actin and tubulin have also been shown to play an important part in germination of *Arabidopsis* seed[49] and rice pollen[50], as well as in spore germination of filamentous fungi such as *Aspergillus nidulans*[51]. Finally, increased protein synthesis has also been shown to occur in plant seed, plant pollen and fungal spore germination[50, 52, 53]. All of these may point to conserved mechanisms for germination for plant, fungi and oomycetes.

In Arabidopsis it has been shown that protein synthesis during seed germination is dependent on mRNA that has been pre-synthesized during seed maturation[53]. While it was abolished in the presence of a translational inhibitor, seed germination proceeded in the absence of transcription in the germinating seed. The fact that a large increase is observed for the *Translation, ribosomal structure and biogenesis* category in the germinating cyst (figure 3.4), but no increase is observed for the *Transcription* category (remains less than 1%) may suggest a similar mechanism in which protein synthesis in the germinating cyst is dependent on mRNA pre-synthesized earlier in the life cycle, perhaps in during zoospore encystment. However, this speculation should be tested since transcription factors and other proteins involved in transcription are typically very unabundant and are not likely to be identified in such global proteomic survey.

Sixty-three orthologous protein pairs were up-regulated in the germinating cyst of both organisms (Figure 3.3, Table S3.10). Thus, they were identified as candidates for involvement in early infection. A group of four such orthologous protein pairs was highly

up regulated in the germinating cyst of both organisms, while completely undetected in the mycelium samples (one example in table 3.1). These proteins were annotated as structural molecules, and although they vary in sequence, they all share multiple 'laminin-type EGF-like' domains. The EGF-like domain is a sequence of 30-40 amino acids found in the epidermal growth factor (EGF), as well as in many other proteins from various organisms[54]. This domain is typically found in the extracellular domain of membrane-bound proteins or in secreted proteins[54, 55] that typically mediate cell adhesion, growth migration, and differentiation [56-59]. All *Phytophthora* proteins in this group are predicted to contain a signal peptide, indicating that these proteins may be acting at the interface between the host and the pathogen. These proteins may play a part in adhesion of the cyst to the plant host, and facilitate both germination of the cyst on the plant tissue and penetration of the germ tube into the host tissue. BLAST searches of these *Phytophthora* proteins show high sequence similarity to the products of three Sexually Induced Genes (SIG1, SIG2 and SIG3) in the diatom *Thalassiosira weissflogii*. The SIG proteins also show a striking similarity to other extracellular proteins involved in cell-cell interactions[60], supporting the hypothesis that the corresponding *P. sojae* and *P. ramorum* proteins are involved in interaction with their plant hosts.

Another pair of similarly regulated orthologs includes *P. ramorum* protein Pr_83420 and *P. sojae* protein Ps_142672. Both proteins were highly up regulated in germinating cyst (Table 3.1). While the function of these proteins is unknown, they contain multiple ricin-B lectin domains. Ricin is a lectin found in castor beans, and is toxic to people, animals and insects. The ricin B lectin domain is found in many carbohydrate-recognition proteins including plant and bacterial AB-toxins, glycosidases

Table 3.1: Examples of candidate proteins for life stage specificity or host range

Candidate[1] Number	Organism	Protein ID	Up-regulation life-stage	Fold increase	Spectral count Cyst / Myc	Function	Possible role
1	*P. ramorum*	Pr_74450	Germinating cyst	>5	19 / 0	Structural molecule, Laminin-type EGF-like domain	Early infection
1	*P. sojae*	Ps_140046	Germinating cyst	>5	91 / 0	Structural molecule, Laminin-type EGF-like domain	Early infection
2	*P. ramorum*	Pr_83420	Germinating cyst	>5	248 / 0	Unknown, ricin domain	Early infection
2	*P. sojae*	Ps_142672	Germinating cyst	9	159 / 13	Unknown, ricin domain	Early infection
3	*P. ramorum*	Pr_87864	Germinating cyst	6	44 / 10	Unknown, alpha-mannosidase similarity	Early infection
3	*P. sojae*	Ps_145478	Germinating cyst	17	75 / 4	Unknown, alpha-mannosidase similarity	Early infection
4	*P. ramorum*	Pr_80246	Germinating cyst	8	131 / 27	Alpha-mannosidase	Early infection
4	*P. sojae*	Ps_131780	Germinating cyst	5	184 / 28	Alpha-mannosidase	Early infection
5	*P. ramorum*	Pr_75997	Mycelium	14	6 / 107	Coproporphyrinogen III oxidase	Vegetative growth
5	*P. sojae*	Ps_132631	Mycelium	5	4 / 16	Coproporphyrinogen III oxidase	Vegetative growth
6	*P. ramorum*	Pr_85749	Mycelium	20	4 / 123	Annexin	Vegetative growth
6	*P. sojae*	Ps_144346	Mycelium	>5	0 / 91	Annexin	Vegetative growth
7	*P. ramorum*	Pr_87858	Mycelium	>5	0 / 179	Annexin	Vegetative growth
7	*P. sojae*	Ps_145567	Mycelium	>5	0 / 7	Annexin	Vegetative growth
8	*P. ramorum*	Pr_77742	Mycelium	8	6 / 77	Phospholipase D	Vegetative growth
8	*P. sojae*	Ps_127024	Mycelium	13	7 / 64	Phospholipase D	Vegetative growth
9	*P. sojae*	Ps_127028	Germinating cyst	>5	28 / 0	Glycosyl hydrolase family 81	Soybean infection

10	*P. sojae*	Ps_126853	Germinating cyst	>5	13 / 0	Similar to CRN3 *from P. infestans*	Soybean infection
11	*P. sojae*	Ps_145528	Germinating cyst	>5	17 / 0	Similar to CRN5 *from P. infestans*	Soybean infection
12	*P. sojae*	Ps_140694	Mycelium	11	7 / 57	CBEL – cell surface glycoprotein	Soybean infection
13	*P. ramorum*	Pr_78735	Mycelium	>5	0 / 51	Nucleoside-diphosphate-sugar epimerase	Broad host range
13	*P. sojae*	Ps_133984	-		0 / 0	Nucleoside-diphosphate-sugar epimerase	-
14	*P. ramorum*	Pr_76706	Mycelium	>5	0 / 30	Sugar transporter of the major facilitator superfamily	Broad host range

[1] Proteins with the same candidate number are orthologs. Individual candidates have no orthologs.

or proteases[61-63]. Up-regulation in the germinating cyst life stage may be directed at interaction with host carbohydrate moieties such as sugars or glycoproteins as a part of the chemical warfare that occurs between the pathogen and the host. A final example of proteins potentially involved in early infection includes two orthologous pairs annotated as lysosomal alpha-mannosidases. These proteins were all up-regulated in germinating cyst (Table 3.1). Lysosomal alpha-mannosidase is an enzyme that mediates the catabolism of N-linked carbohydrates released during the glycoprotein degradation process [64]. An increase in expression of these proteins in the germinating cyst may indicate increased internal glycoprotein turnover in the transition from cyst to mycelium, or perhaps even degradation of secreted host glycoprotein taken up by the pathogen.

Mycelium proteome

When compared to germinating cyst, the portion of differentially expressed proteins involved in transport and metabolism of amino acids, carbohydrates, coenzymes and secondary metabolites consistently increased in the mycelium of both organisms (Figure 3.4). The greater representation of proteins involved in carbohydrate transport and metabolism is not surprising considering that the mycelium is the vegetative growth life stage, in which the organism utilizes nutrients from its environment. When *Phytophthora* grows necrotrophically inside the plant tissue, the environment in which the mycelium is present is rich in carbohydrates from the plant host, and utilization of these molecules is advantageous for the pathogen. While most proteins in this category are associated with carbohydrate transport and glycolysis (catabolism of glucose for

energy production), proteins in the other up-regulated mycelium functional categories are associated with biosynthesis. Thus, it appears that *Phytophthora* mycelium utilizes external glucose as an energy source to drive biosynthesis of amino acids and other small molecules necessary for vegetative growth.

Other classes of proteins up-regulated in the mycelium life stage included transporters, permeases, various lyases, kinases, and oxidoreductases (Tables S3.8-S3.9). Twenty six orthologous proteins were up regulated in the mycelia of both organisms and were identified as candidates for involvement specifically in vegetative growth (Figure 3.3, Table S3.10). One such pair includes *P. sojae* Ps_132631 and *P. ramorum* Pr_75997 (Table 3.1). These proteins are coproporphyrinogen III oxidases, an enzyme class participating in the process of heme biosynthesis[65]. Heme and its derivatives are cofactors of different enzymes in different organisms, including cytochromes in the respiratory chains. In yeast, the rate-limiting coproporphyrinogen oxidase is transcriptionally induced under low oxygen conditions[66, 67], thus increasing heme production and possibly enabling improved respiration. Although it is unclear if similar regulation of this enzyme occurs in *Phytophthora*, it is interesting to speculate that up-regulation of this enzyme during growth of the mycelium through the plant tissue, where oxygen levels are low, allows survival of the pathogen inside the plant tissue.

Other orthologous protein pairs up-regulated in mycelium are *P. sojae* Ps_144346 and *P. ramorum* Pr_85749, as well as *P. sojae* Ps_145567 and *P. ramorum* Pr_87858 (table 3.1). A close examination of the genomic sequence suggested that these were miss-annotated gene models encoding a single *P. ramorum* protein and a single *P. sojae* protein. These are annexin-like proteins, a family of eukaryotic Ca^{2+}-dependent

phospholipid binding proteins involved in diverse functions such as exocytosis, membrane fusion, phospholipase inhibition, and Ca^{2+}-channel regulation[68-70]. *P. ramorum* Pr_87858 has been previously identified as a mycelial cell wall associated protein[71]. While the exact function of these *Phytophthora* annexins is unknown, they may be important for mycelial growth through the host tissue.

A final example of candidates that might be important for vegetative mycelium growth includes two orthologous protein pairs: *P. sojae* Ps_127024 and *P. ramorum* Pr_77742 (Table 3.1), and *P. sojae* Ps_127026 and *P. ramorum* Pr_77744. These proteins contain catalytic domains similar to phospholipase D (PLD). PLDs are enzymes that hydrolyze the terminal phosphodiester bond of phospholipids into free choline and phosphatidic acid, a compound heavily involved in signal transduction[72]. *Phytophthora* PLDs have unusual domain organization and composition as compared to other known eukaryotic PLDs and are thought to participate in non-typical biochemical pathways[73]. The up-regulation of these proteins in the *Phytophthora* mycelium may facilitate signal transduction or other biochemical processes important for growth inside the plant tissue.

Identification of candidate proteins involved in host specificity

The host ranges of *P. sojae* and *P. ramorum* are very different. While *P. ramorum* can infect a broad range of plant hosts, including many varieties of trees and shrubs, *P. sojae* can only infect certain varieties of soybean (*Glycine max*). The ability of *P. sojae* and *P. ramorum* to infect different hosts must be represented in their proteomic makeup. Therefore, the proteomes of the same life stage from *P. sojae* and *P. ramorum* were compared in order to identify candidate proteins involved in host-range specificity.

Orthologous proteins that did not follow similar expression pattern and proteins with no orthologs that were expressed in one organism can be considered as such candidates. The list of proteins following these criteria was large (Figute 3.6, Tables S3.3, S3.4, S3.11). In addition to proteins involved in host specificity, this list likely contained proteins that account for other differences between *P. sojae* and *P. ramorum*, such as adaptation to growth environment (below ground vs. foliar), sexuality (self-fertile vs. out-crossing), and different metabolic requirements. We therefore examined the list manually for proteins whose predicted function suggests involvement in host specificity.

Candidate *P. sojae* proteins involved in soybean specificity (narrow host range)

Many of the proteins that were identified as candidates specifically involved in *P. sojae* host range were annotated as protein or carbohydrate hydrolases. One such example is *P. sojae* Ps_127028 (Table 3.1) which was found to be up regulated in the germinating cyst. No *P. ramorum* ortholog was identified for this protein, and *P. ramorum* proteins that shared sequence similarity with it were undetected in both life stages of *P. ramorum*. This protein belongs to the glycosyl hydrolase family 81 (beta-1,3-glucanases), and share sequence similarity with glycosyl hydrolases from different fungi. Glycosyl hydrolases are enzymes that catalyze hydrolysis of the glycosidic bond between carbohydrates or between a carbohydrate and a non-carbohydrate molecule[74] and this protein may be involved in degradation of specific soybean polysaccharides or glycoproteins.

A second interesting group of proteins that may be specific to soybean infection includes 31 *P. sojae* proteins with high sequence similarity to the Crinkling and Necrosis

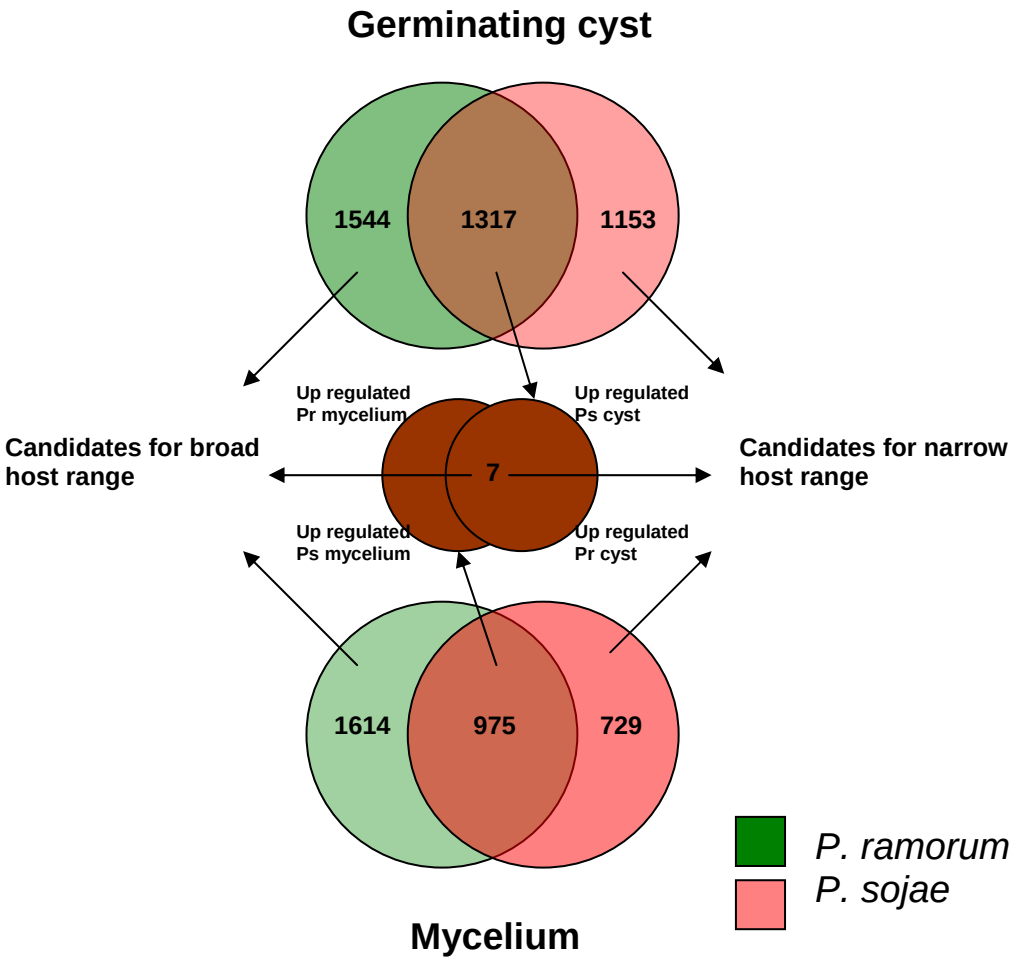


Figure 3.6- Identification of candidate proteins for involvement in narrow vs. broad host range capabilities. The total identified proteome (union of two MudPIT experiments) of each *P. ramorum* life stage was compared to the proteome of the same life stage from *P. sojae* (top and bottom overlapping circles). Overlapping regions represent orthologous proteins found in both, while non-overlapping regions represent proteins with no orthologs, and proteins whose orthologous counterparts were not identified in the other organism. Also, orthologous proteins that were identified in both organisms but followed a different expression pattern with respect to the life stages (middle overlapping circles) were also identified as candidates for host range capabilities. Pr = *P. ramorum*, Ps = *P. sojae*

inducing proteins CRN2, CRN3, CRN5, CRN7 and CRN13 from *P. infestans*. Nearly all of these *P. sojae* proteins were found to be up-regulated in the germinating cyst life stage, although some less than 5 fold (Table S3.12, two examples are presented in Table 3.1), and had no *P. ramorum* orthologs. The CRN family of proteins is unique to *Phytophthora*[75]. While their function is not precisely defined, members of this family are secreted proteins that have been suggested to play an important part in the interaction with the host plant[37]. Both resistant and susceptible host plants showed a necrotic response as well as other defense responses to virally transduced *P. infestans crn* [75]. It is possible that these *P. sojae* proteins are required for successful infection initiation of soybean but not for *P. ramorum* hosts.

A final example of a protein that may contribute to successful infection of soybean is *P. sojae* Ps_140694. This is an extracelluar protein which was highly up-regulated in the mycelium (table 3.1). While multiple homologous proteins are predicted in *P. ramorum*, most of them were undetected, none was up-regulated in the mycelium life stage, and some were slightly up-regulated in the germinating cyst. This *P. sojae* protein shares sequence and domain similarity with the CBEL (cellulose binding elicitor lectin) group of *Phytophthora* proteins. CBELs are glycoproteins that are localized to the *Phytophthora* cell surface, and elicit hypersensitive-response (HR)-like necrosis and defense gene expression in the host plant[76, 77]. Functionally they are implicated in adhesion and cell wall structure, as well as in sensing exogenous cellulose[78]. The unique expression pattern of *P. sojae* Ps_140694 may represent specific adaptation to soybean invasion or evasion from its early defense responses. Another possibility is that the broad

host range *P. ramorum* has a more elaborate regulation of its homologous CBEL proteins, and expresses them only in the presence of a specific host.

Candidate *P. ramorum* proteins involved in broad host range capability

Proteins uniquely expressed in *P. ramorum* may contribute to its ability to infect a broad range of hosts. As was the case for *P. sojae*, the identified candidates included peptidases and other hydrolases, oxidoreductases, and kinases. Another family of *P. ramorum* proteins that may be involved in broad host range infection ability includes proteins Pr_78733, Pr_78734, Pr_78735, Pr_78744, Pr_86077 and Pr_86078. All of these proteins were found to be up-regulated in *P. ramorum* mycelium (example in Table 3.1). While 4 of the 6 proteins had orthologs in *P. sojae*, these orthologs as well as non-orthologous proteins with sequence similarity were not detected in both life stages of *P. sojae*. Although annotated as unknowns, these *P. ramorum* proteins show high similarity to a predicted nucleoside-diphosphate-sugar epimerase from the plant-pathogenic ascomycete fungus *Gibberella zeae* (anamorph *Fusarium graminearum*), the causative agent of head blight of wheat[79]. Epimerases are ubiquitous enzymes that change the stereochemistry of carbohydrate moieties, and are involved in many cellular processes[80]. In bacteria epimerases are involved in production of complex carbohydrate polymers that, when incorporated into the cell wall or envelop, protect the cell from its host's immune responses[80]. While some epimerases are involved in vital cellular processes, such as the Calvin cycle in plants[80], it is possible that since their *P. sojae* counterparts were not detected, these *P. ramorum* proteins are epimerases involved in production of

carbohydrates that protect *P. ramorum* from its different plant hosts, or allow utilization of a broader range of host sugars.

As a final example, protein Pr_76706 was found to be up-regulated in the mycelium of *P. ramorum* and no direct ortholog to it was identified in *P. sojae* (Table 3.1). Although multiple *P. sojae* proteins showed high sequence similarity to it, none were expressed at a significant level. This *P. ramorum* protein was annotated as a sugar transporter of the major facilitator superfamily. Proteins of this family are integral membrane proteins capable of transporting a wide variety of substrates, including simple sugars, oligosaccharides, inositols, drugs, amino acids, nucleosides, organophosphate esters, Krebs cycle metabolites, and many different organic and inorganic ions[81]. One possibility is that uptake of certain molecules present in the protoplast of the host plant by the *P. ramorum* mycelium assists survival and growth of the pathogen inside the plant tissue.

Conclusions

A novel proteomic strategy for identification of candidate proteins involved in different cellular processes was applied to two life stages of two *Phytophthora* species. Ninety-one orthologous protein pairs that followed similar expression patterns in *P. ramorum* and *P. sojae* were identified as candidates for early infection or for vegetative growth. Many species specific proteins were identified, some of which were suggested as candidates for broad or narrow host range capability. These include peptidases and other hydrolases, transporters, and proteins involved in signal transduction mechanisms. Global analysis of the germinating cyst proteome from both species shows increased production

of proteins involved in lipid transport and metabolism, cytoskeleton and protein synthesis. Global analysis of the mycelium life stage shows increased production of proteins involved in transport and metabolism of many molecules, including amino acids, carbohydrates, coenzymes, and secondary metabolites.

Based on these finding we propose a model in which germinating cysts of *Phytophthora* catabolize lipid reserves through the β-oxidation pathway as an energy source (Figure 3.7a). This energy is used to gear-up protein production in the transition from a cyst to mycelium, as well as for the production of the germ tube and initiating infection. During the process of germination, structural cytoskeleton proteins are up-regulated for construction of the germ tube, and proteins involved in microtubule based movement are up-regulated for shuttling of molecules required for successful infection to the terminal end of the germ tube, where interaction with the plant host takes place. The germinating cyst adheres to the plant host through the expression of secreted and membrane structural proteins, including laminin-type EGF-like domain containing proteins. Combating host defenses is achieved through expression of enzymes that manipulate host defenses or degrade host proteins and carbohydrates.

Once inside the host tissue, the pathogen grows vegetatively as mycelium. At this life stage, the pathogen absorbs its nutrients from the degraded tissue of its plant host. Specifically, the host nutrients are utilized as fuel to drive glycolysis, and the energy that is produced by this process is utilized to drive synthesis of amino acids and other small molecules necessary for vegetative growth (Figure 3.7b).

Each *Phytophthora* species expresses a unique set of proteins during infection. These include proteases and other hydrolyses, as well as proteins that are involved in

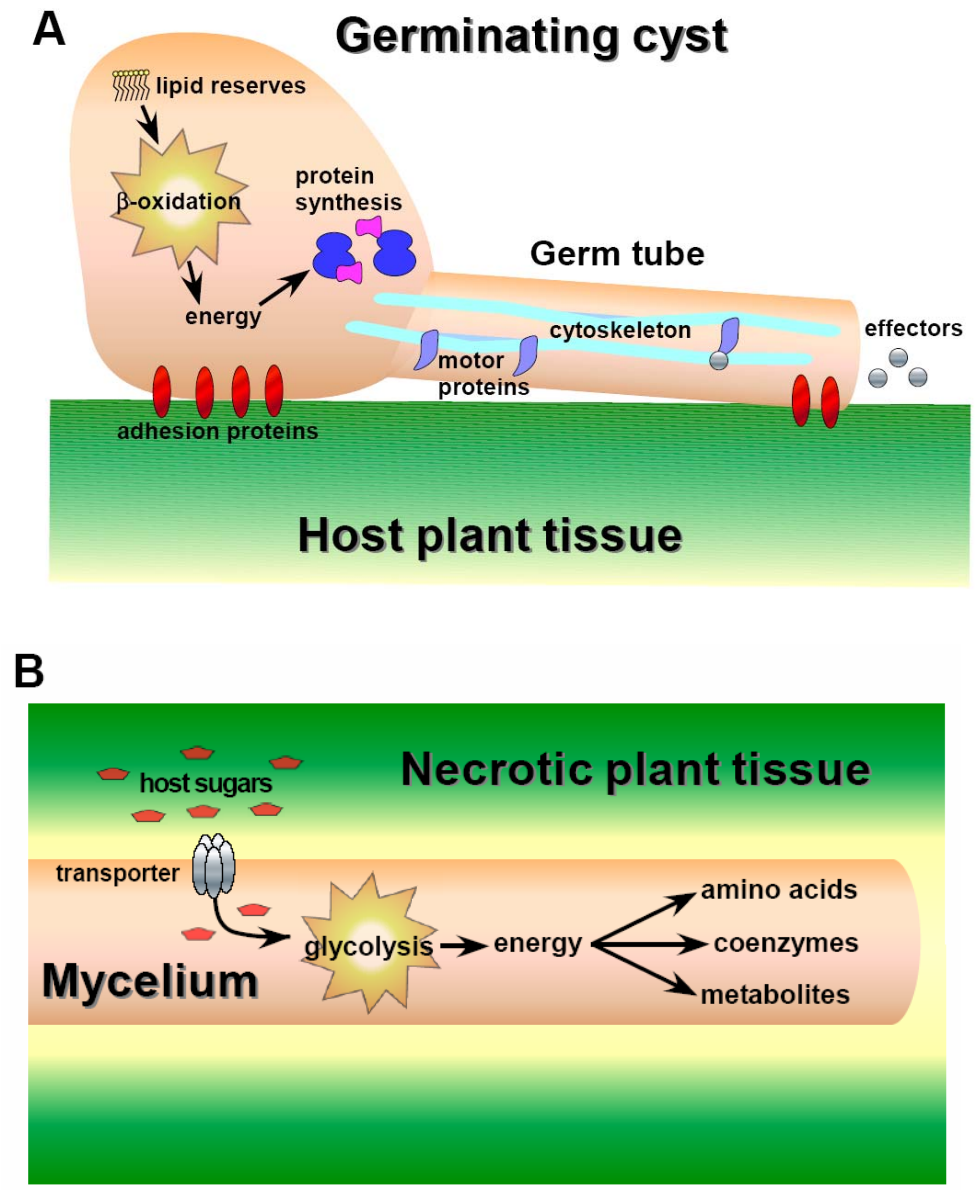


Figure 3.7- Proposed model for *Phytophthora* germinating-cyst and mycelium metabolic regulation. Based on the proteins that were found to be up-regulated in their respective life stages, it is proposed that the germinating-cysts catabolize internal lipid reserves, and use this energy for extensive protein synthesis including cytoskeleton and motor proteins required for generation of the germ tube. Adhesion proteins are also up-regulated and mediate adhesion of the germinating cyst to the plant. Once inside the host, *Phytophthora* grows vegetatively as mycelium where host sugars are transported and utilized as an energy source for glycolysis. Energy derived by this process is used to drive synthesis of many small molecules, including amino acids, coenzymes, and secondary metabolites.

protein-protein interaction and in signal transduction. *Phytophthora sojae*, which has a narrow host range and can only infect soybean uniquely expresses a group of Crinkling and Necrosis inducing proteins during cyst germination which may contribute to its specificity for soybean. *Phytophthora ramorum*, which has a broad host range, uniquely expresses a group of proteins involved in sugar transport and modification during vegetative growth, which may aid in infection of its different hosts. The hypothetical role of targets identified in this study may be examined in the future using molecular tools available for *Phytophthora* manipulation, including RNAi, Targeted Induced Local Lesions in Genomes (TILLING), transformation, and classical genetics. Indeed, future study in our laboratory will focus on further characterization of early infection and species specific targets identified in this study.

With few exceptions, our proteomic study agrees with or complements results from several transcriptomic studies carried out in *Phytophthora*[82-85]. Recent studies of *P. sojae*[83, 85] and *P. parasitica*[82] identified multiple ESTs representing genes involved in glycolysis in mycelium or *in-planta*, further supporting our findings. Contrary to our results, differential hybridization analysis of cDNA libraries from *P. nicotianae* identified glycolytic genes to be up-regulated in the germinating cyst[84]. However, in accordance with our finding, the same study did suggest up-regulation of genes encoding proteins involved in protein biosynthesis, cell wall biogenesis and adhesion in the germinating cyst. Grenville-Briggs et al. identified 5 genes involved in amino acid biosynthesis to be up-regulated in appresorium forming germinating cyst and potato infection in *P. infestans* using 2DE and Real-Time RT-PCR[86]. With one exception (threonine synthase), the homologs of all of these genes in *P. sojae* and *P. ramorum* were found to be somewhat

up-regulated in mycelium in this study. Torto et al. has also identified many ESTs representing *P. sojae* genes potentially involved in pathogenicity, such as hydrolases, elicitins, CRNs and RxLRs[85]. The proteomic study presented here not only corroborates the presence of these gene products, but also suggests their temporal expression. Any discrepancy between our proteomic identification and previous transcriptomic studies might be explained by sampling bias of each technique, biological variation of different species, different growth conditions or sample preparation, or by a true lack of correlation between the mRNA level and protein level in the cell.

Recently, the genomes of *P. capsici* and *P. infestans* were sequenced, and they are currently being assembled and annotated. Both organisms have host ranges that are very different from those of *P. sojae* and *P. ramorum*. Expansion of the cross-species proteomic analysis to include these newly sequenced *Phytophthora* species should prove useful for testing whether the proteins identified in this study are life-stage or host-range specific and will help refine the model described above. Chapter 4 describes expansion of this study to include *P. capsici*. Additionally, once a susceptible host is fully sequenced, a proteomic study of *Phytophthora* as it infects the plant will allow examination of our infection model. Chapter 5 describes such *in planta* experiments on *P. capsici* infected tomato plant, despite availability of only a partial tomato genome sequence. Finally, the strategy for cross-species comparative analysis outlined here provides a powerful strategy for investigating biological processes in closely related organisms.

References

1. Erwin, D. C., and Ribeiro, O. K., *Phytophthora Diseases Worldwide*. The American Phytopathological Society: St. Paul, MN, 1996; p 562.
2. Rizzo, D. M.; Garbelotto, M.; Davidson, J. M.; Slaughter, G. W.; Koike, S. T., Phytophthora ramorum as the cause of extensive mortality of Quercus spp. and Lithocarpus densiflorus in California. *Plant Disease* **2002**, (86), 205–214.
3. Werres, S.; Marwitz, R.; Man in't Veld, W. A.; M., d. C. A. W. A.; Bonants, P. J. M.; de Weerdt, M.; Themann, K.; Ilieva, E.; Baayen, R. P., Phytophthora ramorum sp. nov., a new pathogen on Rhododendron and Viburnum. *Mycological Research* **2001**, (105), 1166–1175.
4. Rizzo, D. M.; Garbelotto, M.; Hansen, E. M., PHYTOPHTHORA RAMORUM: Integrative Research and Management of an Emerging Pathogen in California and Oregon Forests. *Annual Review of Phytopathology* **2005,** 43, (1), 309-C-4.
5. Tyler, B. M.; Tripathy, S.; Zhang, X.; Dehal, P.; Jiang, R. H.; Aerts, A.; Arredondo, F. D.; Baxter, L.; Bensasson, D.; Beynon, J. L.; Chapman, J.; Damasceno, C. M.; Dorrance, A. E.; Dou, D.; Dickerman, A. W.; Dubchak, I. L.; Garbelotto, M.; Gijzen, M.; Gordon, S. G.; Govers, F.; Grunwald, N. J.; Huang, W.; Ivors, K. L.; Jones, R. W.; Kamoun, S.; Krampis, K.; Lamour, K. H.; Lee, M. K.; McDonald, W. H.; Medina, M.; Meijer, H. J.; Nordberg, E. K.; Maclean, D. J.; Ospina-Giraldo, M. D.; Morris, P. F.; Phuntumart, V.; Putnam, N. H.; Rash, S.; Rose, J. K.; Sakihama, Y.; Salamov, A. A.; Savidor, A.; Scheuring, C. F.; Smith, B. M.; Sobral, B. W.; Terry, A.; Torto-Alalibo, T. A.; Win, J.; Xu, Z.; Zhang, H.; Grigoriev, I. V.; Rokhsar, D. S.; Boore, J. L., Phytophthora genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **2006,** 313, (5791), 1261-6.
6. Jiang, R. H.; Tyler, B. M.; Govers, F., Comparative analysis of Phytophthora genes encoding secreted proteins reveals conserved synteny and lineage-specific gene duplications and deletions. *Mol Plant Microbe Interact* **2006,** 19, (12), 1311-21.
7. Washburn, M. P., Wolters, D., and Yates III, J.R., Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology* **2001,** 19, 242-247.
8. Cagney, G.; Park, S.; Chung, C.; Tong, B.; O'Dushlaine, C.; Shields, D. C.; Emili, A., Human tissue profiling with multidimensional protein identification technology. *J Proteome Res* **2005,** 4, (5), 1757-67.
9. Skipp, P.; Robinson, J.; O'Connor, C. D.; Clarke, I. N., Shotgun proteomic analysis of Chlamydia trachomatis. *Proteomics* **2005,** 5, (6), 1558-73.
10. Vitali, B.; Wasinger, V.; Brigidi, P.; Guilhaus, M., A proteomic view of Bifidobacterium infantis generated by multi-dimensional chromatography coupled with tandem mass spectrometry. *Proteomics* **2005,** 5, (7), 1859-67.

11. Brown, S. D.; Thompson, M. R.; Verberkmoes, N. C.; Chourey, K.; Shah, M.; Zhou, J.; Hettich, R. L.; Thompson, D. K., Molecular dynamics of the Shewanella oneidensis response to chromate stress. *Mol Cell Proteomics* **2006,** 5, (6), 1054-71.
12. Chu, D. S.; Liu, H.; Nix, P.; Wu, T. F.; Ralston, E. J.; Yates, J. R., 3rd; Meyer, B. J., Sperm chromatin proteomics identifies evolutionarily conserved fertility factors. *Nature* **2006,** 443, (7107), 101-5.
13. Florens, L.; Washburn, M. P.; Raine, J. D.; Anthony, R. M.; Grainger, M.; Haynes, J. D.; Moch, J. K.; Muster, N.; Sacci, J. B.; Tabb, D. L.; Witney, A. A.; Wolters, D.; Wu, Y.; Gardner, M. J.; Holder, A. A.; Sinden, R. E.; Yates, J. R.; Carucci, D. J., A proteomic view of the Plasmodium falciparum life cycle. *Nature* **2002,** 419, (6906), 520-6.
14. Leifso, K.; Cohen-Freue, G.; Dogra, N.; Murray, A.; McMaster, W. R., Genomic and proteomic expression analysis of Leishmania promastigote and amastigote life stages: The Leishmania genome is constitutively expressed. *Mol Biochem Parasitol* **2007,** 152, (1), 35-46.
15. Savidor, A.; Donahoo, R. S.; Hurtado-Gonzales, O.; Verberkmoes, N. C.; Shah, M. B.; Lamour, K. H.; McDonald, W. H., Expressed Peptide Tags: An Additional Layer of Data for Genome Annotation. *J Proteome Res* **2006,** 5, (11), 3048-3058.
16. McDonald, W. H.; Tabb, D. L.; Sadygov, R. G.; MacCoss, M. J.; Venable, J.; Graumann, J.; Johnson, J. R.; Cociorva, D.; Yates, J. R., MS1, MS2, and SQT-three unified, compact, and easily parsed file formats for the storage of shotgun proteomic spectra and identifications. . *Rapid Commun Mass Spectrom* **2004,** 18, (18), 2162-8.
17. Narasimhan, C., Tabb, D.L., Verberkmoes, N.C., Thompson, M.R., Hettich, R.L., Uberbacher, E.C., MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Analytical chemistry* **2005,** 77, (23), 7581-93.
18. Tabb, D. L.; Narasimhan, C.; Strader, M. B.; Hettich, R. L., DBDigger: reorganized proteomic database identification that improves flexibility and speed. *Anal Chem* **2005,** 77, (8), 2464-74.
19. Tabb, D. L.; McDonald, W. H.; Yates, J. R., 3rd, DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* **2002,** 1, (1), 21-6.
20. Tabb, D. L.; Shah, M. B.; Strader, M. B.; Connelly, H. M.; Hettich, R. L.; Hurst, G. B., Determination of Peptide and Protein Ion Charge States by Fourier Transformation of Isotope-Resolved Mass Spectra. *Journal of the American Society for Mass Spectrometry* **2006,** 17, (7), 903-915.
21. Florens, L.; Carozza, M. J.; Swanson, S. K.; Fournier, M.; Coleman, M. K.; Workman, J. L.; Washburn, M. P., Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **2006,** 40, (4), 303-11.
22. Benov, L.; Fridovich, I., Superoxide dismutase protects against aerobic heat shock in Escherichia coli. *J Bacteriol* **1995,** 177, (11), 3344-6.
23. Fridovich, I., Superoxide radical and superoxide dismutases. *Annu Rev Biochem* **1995,** 64, 97-112.

24. Lanfranco, L.; Novero, M.; Bonfante, P., The mycorrhizal fungus Gigaspora margarita possesses a CuZn superoxide dismutase that is up-regulated during symbiosis with legume hosts. *Plant Physiol* **2005,** 137, (4), 1319-30.
25. Garcia-Brugger, A.; Lamotte, O.; Vandelle, E.; Bourque, S.; Lecourieux, D.; Poinssot, B.; Wendehenne, D.; Pugin, A., Early signaling events induced by elicitors of plant defenses. *Mol Plant Microbe Interact* **2006,** 19, (7), 711-24.
26. Apostol, I.; Heinstein, P. F.; Low, P. S., Rapid Stimulation of an Oxidative Burst during Elicitation of Cultured Plant Cells : Role in Defense and Signal Transduction. *Plant Physiol* **1989,** 90, (1), 109-116.
27. Jennings, D. B.; Ehrenshaft, M.; Pharr, D. M.; Williamson, J. D., Roles for mannitol and mannitol dehydrogenase in active oxygen-mediated plant defense. *Proc Natl Acad Sci U S A* **1998,** 95, (25), 15129-33.
28. Keller, T.; Damude, H. G.; Werner, D.; Doerner, P.; Dixon, R. A.; Lamb, C., A plant homolog of the neutrophil NADPH oxidase gp91phox subunit gene encodes a plasma membrane protein with Ca2+ binding motifs. *Plant Cell* **1998,** 10, (2), 255-66.
29. De Groote, M. A.; Ochsner, U. A.; Shiloh, M. U.; Nathan, C.; McCord, J. M.; Dinauer, M. C.; Libby, S. J.; Vazquez-Torres, A.; Xu, Y.; Fang, F. C., Periplasmic superoxide dismutase protects Salmonella from products of phagocyte NADPH-oxidase and nitric oxide synthase. *Proc Natl Acad Sci U S A* **1997,** 94, (25), 13997-4001.
30. Mandell, G. L., Catalase, superoxide dismutase, and virulence of Staphylococcus aureus. In vitro and in vivo studies with emphasis on staphylococcal--leukocyte interaction. *J Clin Invest* **1975,** 55, (3), 561-6.
31. Narasipura, S. D.; Ault, J. G.; Behr, M. J.; Chaturvedi, V.; Chaturvedi, S., Characterization of Cu,Zn superoxide dismutase (SOD1) gene knock-out mutant of Cryptococcus neoformans var. gattii: role in biology and virulence. *Mol Microbiol* **2003,** 47, (6), 1681-94.
32. Birch, P. R.; Rehmany, A. P.; Pritchard, L.; Kamoun, S.; Beynon, J. L., Trafficking arms: oomycete effectors enter host plant cells. *Trends Microbiol* **2006,** 14, (1), 8-11.
33. Ellis, J.; Catanzariti, A. M.; Dodds, P., The problem of how fungal and oomycete avirulence proteins enter plant cells. *Trends Plant Sci* **2006,** 11, (2), 61-3.
34. Rehmany, A. P.; Gordon, A.; Rose, L. E.; Allen, R. L.; Armstrong, M. R.; Whisson, S. C.; Kamoun, S.; Tyler, B. M.; Birch, P. R.; Beynon, J. L., Differential recognition of highly divergent downy mildew avirulence gene alleles by RPP1 resistance genes from two Arabidopsis lines. *Plant Cell* **2005,** 17, (6), 1839-50.
35. Liu, H.; Sadygov, R. G.; Yates, J. R., 3rd, A model for random sampling and estimation of relative protein abundance in shotgun proteomics. *Anal Chem* **2004,** 76, (14), 4193-201.
36. Kunau, W. H.; Dommes, V.; Schulz, H., beta-oxidation of fatty acids in mitochondria, peroxisomes, and bacteria: a century of continued progress. *Prog Lipid Res* **1995,** 34, (4), 267-342.
37. Ebstrup, T.; Saalbach, G.; Egsgaard, H., A proteomics study of in vitro cyst germination and appressoria formation in Phytophthora infestans. *Proteomics* **2005,** 5, (11), 2839-48.

38. Poirier, Y.; Antonenkov, V. D.; Glumoff, T.; Hiltunen, J. K., Peroxisomal beta-oxidation--a metabolic pathway with multiple functions. *Biochim Biophys Acta* **2006,** 1763, (12), 1413-26.
39. Clark, M. C.; Melanson, D. L.; Page, O. T., Purine metabolism and differential inhibition of spore germination in Phytophthora infestans. *Can J Microbiol* **1978,** 24, (9), 1032-8.
40. Kramer, R.; Freytag, S.; Schmelzer, E., *In vitro* formation of infection structures of *Phytophthora infestans* is associated with synthesis of stage specific polypeptides. *Eur. J. Plant Pathol.* **1997,** 103, 43-53.
41. Penington, C. J.; Iser, J. R.; Grant, B. R.; Gayler, K. R., The role of RNA and protein synthesis in stimulated germination of the zoospores of the pathogenic fungus *Phytophthora palmivora*. *Exp. Mycol.* **1989**, (13), 158-168.
42. Germain, V.; Rylott, E. L.; Larson, T. R.; Sherson, S. M.; Bechtold, N.; Carde, J. P.; Bryce, J. H.; Graham, I. A.; Smith, S. M., Requirement for 3-ketoacyl-CoA thiolase-2 in peroxisome development, fatty acid b-oxidation and breakdown of triacylglycerol in lipid bodies of Arabidopsis seedlings. *The Plant Journal* **2001,** 28, (1), 1-12.
43. Pinfield-Wells, H.; Rylott, E. L.; Gilday, A. D.; Graham, S.; Job, K.; Larson, T. R.; Graham, I. A., Sucrose rescues seedling establishment but not germination of Arabidopsis mutants disrupted in peroxisomal fatty acid catabolism. *The Plant Journal* **2005,** 43, (6), 861-72.
44. Hayashi, M.; Toriyama, K.; Kondo, M.; Nishimura, M., 2,4-Dichlorophenoxybutyric Acid–Resistant Mutants of Arabidopsis Have Defects in Glyoxysomal Fatty Acid b-Oxidation. *The Plant Cell* **1998,** 10, (2), 183-195.
45. Geopfert, S.; Poirier, Y., Beta-oxidation in fatty acid degradation and beyond. *Curr Opin Plant Biol* **2007,** 10, (3), 245-51.
46. Footitt, S.; Marquez, J.; Schmuths, H.; Baker, A.; Theodoulou, F. L.; Holdsworth, M., Analysis of the role of COMATOSE and peroxisomal beta-oxidation in the determination of germination potential in Arabidopsis. *J Exp Bot* **2006,** 57, (11), 2805-14.
47. Penfield, S.; Graham, S.; Graham, I. A., Storage reserve mobilization in germinating oilseeds: Arabidopsis as a model system. *Biochem Soc Trans* **2005,** 33, (Pt 2), 380-3.
48. Chen, X.; Shen, G.; Wang, Y.; Zheng, X.; Wang, Y., Identification of Phytophthora sojae genes upregulated during the early stage of soybean infection. *FEMS Microbiol Lett* **2007,** 269, (2), 280-8.
49. Gallardo, K.; Job, C.; Groot, S. P.; Puype, M.; Demol, H.; Vandekerckhove, J.; Job, D., Proteomic analysis of arabidopsis seed germination and priming. *Plant Physiol* **2001,** 126, (2), 835-48.
50. Dai, S.; Chen, T.; Chong, K.; Xue, Y.; Liu, S.; Wang, T., Proteomics identification of differentially expressed proteins associated with pollen germination and tube growth reveals characteristics of germinated Oryza sativa pollen. *Mol Cell Proteomics* **2007,** 6, (2), 207-30.
51. d'Enfert, C., Fungal Spore Germination: Insights from the Molecular Genetics of Aspergillus nidulans and Neurospora crassa. *Fungal Genet Biol* **1997,** 21, (2), 163-72.

52. Dai, S.; Li, L.; Chen, T.; Chong, K.; Xue, Y.; Wang, T., Proteomic analyses of Oryza sativa mature pollen reveal novel proteins associated with pollen germination and tube growth. *Proteomics* **2006,** 6, (8), 2504-29.
53. Rajjou, L.; Gallardo, K.; Debeaujon, I.; Vandekerckhove, J.; Job, C.; Job, D., The effect of alpha-amanitin on the Arabidopsis seed proteome highlights the distinct roles of stored and neosynthesized mRNAs during germination. *Plant Physiol* **2004,** 134, (4), 1598-613.
54. Campbell, L. D.; Bork, P., Epidermal growth factor-like modules. *Current Opinions in Structural Biology* **1993,** 3, 385-392.
55. Bork, P.; Downing, A. K.; Kieffer, B.; Campbell, I. D., Structure and distribution of modules in extracellular proteins. *Q Rev Biophys* **1996,** 29, (2), 119-67.
56. Beck, K.; Hunter, I.; Engel, J., Structure and function of laminin: anatomy of a multidomain glycoprotein. *Faseb J* **1990,** 4, (2), 148-60.
57. Engel, J., Laminins and other strange proteins. *Biochemistry* **1992,** 31, (44), 10643-51.
58. Hohenester, E.; Engel, J., Domain structure and organisation in extracellular matrix proteins. *Matrix Biol* **2002,** 21, (2), 115-28.
59. Liu, B. F.; Ma, J.; Cui, F. Z., Regulation of charged groups and laminin patterns for selective neuronal adhesion *Colloids and Surfaces B: Biointerfaces* **2006,** 53, 175-178.
60. Armbrust, E. V., Identification of a new gene family expressed during the onset of sexual reproduction in the centric diatom Thalassiosira weissflogii. *Appl Environ Microbiol* **1999,** 65, (7), 3121-8.
61. Dodd, R. B.; Drickamer, K., Lectin-like proteins in model organisms: implications for evolution of carbohydrate-binding activity. *Glycobiology* **2001,** 11, (5), 71R-9R.
62. Mishra, V.; Ethayathulla, A. S.; Sharma, R. S.; Yadav, S.; Krauspenhaar, R.; Betzel, C.; Babu, C. R.; Singh, T. P., Structure of a novel ribosome-inactivating protein from a hemi-parasitic plant inhabiting the northwestern Himalayas. *Acta Crystallogr D Biol Crystallogr* **2004,** 60, (Pt 12 Pt 2), 2295-304.
63. Shimoi, H.; Iimura, Y.; Obata, T.; Tadenuma, M., Molecular structure of Rarobacter faecitabidus protease I. A yeast-lytic serine protease having mannose-binding activity. *J Biol Chem* **1992,** 267, (35), 25189-95.
64. al Daher, S.; de Gasperi, R.; Daniel, P.; Hall, N.; Warren, C. D.; Winchester, B., The substrate-specificity of human lysosomal alpha-D-mannosidase in relation to genetic alpha-mannosidosis. *Biochem J* **1991,** 277 (Pt 3), 743-51.
65. Xu, K.; Elliott, T., An oxygen-dependent coproporphyrinogen oxidase encoded by the hemF gene of Salmonella typhimurium. *J Bacteriol* **1993,** 175, (16), 4990-9.
66. Zagorec, M.; Buhler, J. M.; Treich, I.; Keng, T.; Guarente, L.; Labbe-Bois, R., Isolation, sequence, and regulation by oxygen of the yeast HEM13 gene coding for coproporphyrinogen oxidase. *J Biol Chem* **1988,** 263, (20), 9718-24.
67. Zagorec, M.; Labbe-Bois, R., Negative control of yeast coproporphyrinogen oxidase synthesis by heme and oxygen. *J Biol Chem* **1986,** 261, (6), 2506-9.
68. Gerke, V.; Moss, S. E., Annexins: From structure to function. *Physiol. Rev.* **2002,** 82, (2), 331-371.

69. Raynal, P.; Pollard, H. B., Annexins: the problem of assessing the biological role for a gene family of multifunctional calcium- and phospholipid-binding proteins. *Biochim Biophys Acta* **1994,** 1197, (1), 63-93.
70. Romisch, J.; Paques, E. P., Annexins: calcium-binding proteins of multi-functional importance? *Med Microbiol Immunol* **1991,** 180, (3), 109-26.
71. Meijer, H. J.; van de Vondervoort, P. J.; Yin, Q. Y.; de Koster, C. G.; Klis, F. M.; Govers, F.; de Groot, P. W., Identification of cell wall-associated proteins from Phytophthora ramorum. *Mol Plant Microbe Interact* **2006,** 19, (12), 1348-58.
72. Cazzolli, R.; Shemon, A. N.; Fang, M. Q.; Hughes, W. E., Phospholipid signalling through phospholipase D and phosphatidic acid. *IUBMB Life* **2006,** 58, (8), 457-61.
73. Meijer, H. J.; Govers, F., Genomewide analysis of phospholipid signaling genes in Phytophthora spp.: novelties and a missing link. *Mol Plant Microbe Interact* **2006,** 19, (12), 1337-47.
74. Henrissat, B.; Callebaut, I.; Fabrega, S.; Lehn, P.; Mornon, J. P.; Davies, G., Conserved catalytic machinery and the prediction of a common fold for several families of glycosyl hydrolases. *Proc Natl Acad Sci U S A* **1995,** 92, (15), 7090-4.
75. Torto, T. A.; Li, S.; Styer, A.; Huitema, E.; Testa, A.; Gow, N. A.; van West, P.; Kamoun, S., EST mining and functional expression assays identify extracellular effector proteins from the plant pathogen Phytophthora. *Genome Res* **2003,** 13, (7), 1675-85.
76. Mateos, F. V.; Rickauer, M.; Esquerre-Tugaye, M. T., Cloning and characterization of a cDNA encoding an elicitor of Phytophthora parasitica var. nicotianae that shows cellulose-binding and lectin-like activities. *Mol Plant Microbe Interact* **1997,** 10, (9), 1045-53.
77. Séjalon-Delmas, N.; Villalba, F.; Bottin, A.; Rickauer, M.; Dargent, R.; Esquerré-Tugayé, M. T., Purification, elicitor activity, and cell wall localisation of a glycoprotein from Phytophthora parasitica var. nicotianae, a fungal pathogen of tobacco. . *Phytopathology* **1997**, (87), 899 -909.
78. Gaulin, E.; Jauneau, A.; Villalba, F.; Rickauer, M.; Esquerre-Tugaye, M. T.; Bottin, A., The CBEL glycoprotein of Phytophthora parasitica var-nicotianae is involved in cell wall deposition and adhesion to cellulosic substrates. *J Cell Sci* **2002,** 115, (Pt 23), 4565-75.
79. Bai, G.; Shaner, G., Management and resistance in wheat and barley to fusarium head blight. *Annu Rev Phytopathol* **2004,** 42, 135-61.
80. Allard, S. T.; Giraud, M. F.; Naismith, J. H., Epimerases: structure, function and mechanism. *Cell Mol Life Sci* **2001,** 58, (11), 1650-65.
81. Pao, S. S.; Paulsen, I. T.; Saier, M. H., Jr., Major facilitator superfamily. *Microbiol Mol Biol Rev* **1998,** 62, (1), 1-34.
82. Panabières, F.; Amselem, J.; Galiana, E.; Le Berre, J. Y., Gene identiWcation in the oomycete pathogen Phytophthora parasitica during in vitro vegetative growth through expressed sequence tags. *Fungal Genet Biol* **2005,** 42, (7), 611-23.
83. Qutob, D.; Hraber, P. T.; Sobral, B. W.; Gijzen, M., Comparative Analysis of Expressed Sequences in Phytophthora sojae. *Plant Physiol* **2000,** 123, (1), 243-54.
84. Shan, W.; Marshall, J. S.; Hardham, A. R., Gene expression in germinated cysts of Phytophthora nicotianae. *Mol Plant Pathol* **2004,** 5, (4), 317-30.

85. Torto-Alalibo, T. A.; Tripathy, S.; Smith, B. M.; Arredondo, F. D.; Zhou, L.; Li, H.; Chibucos, M. C.; Qutob, D.; Gijzen, M.; Mao, C.; Sobral, B. W.; Waugh, M. E.; Mitchell, T. K.; Dean, R. A.; Tyler, B. M., Expressed sequence tags from phytophthora sojae reveal genes specific to development and infection. *Mol Plant Microbe Interact* **2007,** 20, (7), 781-93.
86. Grenville-Briggs, L. J.; Avrova, A. O.; Bruce, C. R.; Williams, A.; Whisson, S. C.; Birch, P. R.; van West, P., Elevated amino acid biosynthesis in Phytophthora infestans during appressorium formation and potato infection. *Fungal Genet Biol* **2005,** 42, (3), 244-56.

Chapter 4

A proteomic investigation of *Phytophthora capsici* life stages

Abstract

Phytophthora capsici is a plant pathogen infecting many agricultural crops. Rapid spread of the disease in the field is the result of the exponential growth of *P. capsici* through the asexual life cycle. We therefore carried out a proteomic investigation of three *P. capsici* asexual life stages: germinating cyst, mycelium, and sporangium. The data from these experiments were compared to a similar study of *P. sojae* and *P. ramorum* (Chapter 3). The accumulated data from the study of all 3 species supports the hypothesis posed previously and suggests that the germinating cyst uses β-oxidation to break down internal lipid reserves as the major energy source while utilizing the generated energy for extensive protein synthesis. The mycelium acquires its energy through glycolysis and uses it for synthesis of small molecules. Identification of several life stage and host range specific protein candidates was made and supported the hypothesized role of these proteins in infection. Surprisingly, the sporangium proteome was remarkably similar to the germinating cyst proteome, possibly due to the similar environment in which they are both found.

Introduction

Phytophthora species are a group of fungal like plant pathogens, capable of infecting a vast variety of plants. *Phytophthora capsici* is one *Phytophthora* species that

causes heavy damage to agricultural crops. It was first described in 1922 as the causal agent of blight of *Capsicum annuum* L. (chili pepper) in New Mexico[1], hence its name. However, in addition to almost all varieties of peppers, *P. capsici* is also able to infect many other agriculturally important vegetable crops in the Solanaceae family (e.g. tomato, eggplant), cucurbits (e.g. cucumber, squashes), and others (e.g. cacao, cotton)[1].

Like other *Phytophthora* species, *P. capsici* is capable of reproducing sexually and asexually, with the asexual life cycle facilitating the fast spread of the disease in the field. In the asexual life cycle, a motile zoospore swims chemotactially towards the plant host, and upon contact with the plant it sheds its flagella, encysts, and adheres. Shortly thereafter, the cyst germinates, penetrates into the plant tissue, and starts growing vegetatively through the plant tissue as mycelium. The mycelium can also emerge and grow out of the plant tissue. The terminal ends of the mycelium, having sensed the new environment outside the plant tissue, can be stimulated to differentiate into sporangia, which are multi-nucleated vessel like structures. In the field, the sporangia can be severed from its stem by wind or rain. The sporangium has the ability to germinate directly and initiate infection of a plant on which it landed. A second possibility is for the nuclei in the sporangium to differentiate into zoospores which can then be released from the sporangium and swim to a new plant host, thus continuing the asexual cycle and propagating the disease in the field.

As we were interested in the molecular mechanisms underlying infection of *Phytophthora*, we previously carried out a proteomic investigation of the germinating cyst and mycelium life stages from *P. sojae* and *P. ramorum*. The germinating cyst and the mycelium life stages, although *in vitro*, represent the early and late stages of

infection, respectively. Here we present a similar proteomic study on different life stages of *P. capsici*, including the germinating cyst, mycelium, and sporangium.

The genome sequencing of *P. capsici* has recently been completed. The sequencing of *P. capsici* genome was a collaborative effort involving both the traditional Sanger sequencing technique and the new 454 sequencing platform[2], making *P. capsici* the first eukaryote to be sequence by this integrated technique. While the assembly and annotation of the *P. capsici* genome is still a work in progress, a working draft of the entire list of predicted genes (Forge 5), ~12,000 in total, has been completed and made available to members of the *Phytophthora* community through the Joint Genome Institute. The availability of a protein database allowed us to use *P. capsici* in our global proteomics studies.

Materials and methods

Phytophthora capsici cell culture, life stage isolation, lysate preparation and LC/LC-MS/MS analysis

Experimental procedures were similar to those described in chapter 3. Briefly, mycelia were generated by growing *P. capsici* strain 1534 in clarified antibiotic amended V8 juice broth. Asexual sporangia and zoospores were generated by growing *P. capsici* under constant fluorescent light at room temperature on antibiotic amended V8 agar plates, according to standard protocols[1]. For isolation of sporangia, one week old sporangia rich cultures on V8 agar plates were scraped with a glass rod and the accumulated sporangial mass was scooped with a spatula directly into an Eppendorf tube. Examination under a light microscope confirmed a nearly pure sporangial preparation

with very little contaminating mycelia debris. For germinating cyst isolation, one week old sporangial rich plates were flooded with cold sterile distilled water and kept at 4°C for 45 minutes to stimulate zoospore release. Zoospores were then stimulated to encyst by vigorous shaking for 1-2 minutes and then waiting approximately 2 hours before harvesting the germinating cysts. The sporangia, mycelium, and germinating cyst were freeze-dried and ground using glass beads.

The ground tissues were resuspended in 8M urea and vortexed vigorously. Lysates of all three life stages were fractionated to membrane and soluble fractions by centrifugation. The soluble fractions were then reduced with tris(2-carboxyethyl)phosphine (TCEP), alkylated with iodoacetamide (IAA), and digested with trypsin as previously described[3]. Samples were then filtered, acidified, and loaded offline on a bi-phasic 150 μm inner-diameter fused silica column containing 3-3.5 cm C18 reverse phase material followed by 3.5-4 cm strong cation exchange phase. Loaded columns were placed directly upstream of a 100 μm inner-diameter front column packed with 15 cm reverse-phase material (Jupiter C18 5μm 300 A, Phenomenex, Torrance, CA). Samples were analyzed by 12-step MudPIT (LC/LC-MS/MS) on a linear ion trap mass spectrometer (LTQ, ThermoElectron, San Jose, CA) as previously described[3]. Flow rate at the tip was ~200-300 nl/min. Replicate experiments were carried out on each life stage.

Informatics

The *P. capsici* protein database as well as the GO and KOG functional annotations were downloaded from the JGI website (http://www.jgi.doe.gov/). MS/MS

spectra acquired during MudPIT experiments were extracted, assigned charge, and searched against the predicted protein *P. capsici* databases using the DBDigger algorithm[4] in fully tryptic mode with the MASPIC scorer[5] as described in chapter 3. Search results were filtered with DTASelect[6] using filter values allowing 5% false discovery rate based on a concatenated reversed protein database search, as previously described[3]. Protein identification required 2 peptides per protein for identification. Identified proteins were entered into a relational database and associated with their functional annotations using Microsoft Access. To compare between life stages, the union of the duplicate MudPIT results of the germinating cyst was compared to the union of the duplicate MudPIT results of the mycelium, and the union of the duplicate MudPIT results of the mycelium was compared to the union of the duplicate MudPIT results of the sporangia.

Spectral count was considered as the total number of MS/MS spectra corresponding to peptides from a given protein. Normalization of spectral counts was done using Normalized Spectral Abundance Factors (NSAF) as previously described[7]. Proteins groups that shared all of their peptides were assigned a NSAF value equal to the NSAF value for the group divided by the number of proteins in the group. A protein was determined to be up-regulated in one life stage versus another if (i) it was identified in both MudPIT duplicates of the up-regulated life stage and (ii) the sum of the normalized spectral counts from both duplicates was at least 5 times higher than the sum of normalized spectral counts in the other life stage duplicates.

Results and discussion

The asexual life cycle of *Phytophthora* is responsible for the rapid propagation of disease in the field. In order to study the underlying molecular processes and identify individual proteins that may play a role in infection, we characterized the proteomes of three *P. capsici* asexual life stages: germinating cyst, mycelium, and sporangium. MudPIT experiments were carried out in duplicates on protein preparations from each isolated life stage. Thousands of proteins were identified in each experiment (Table 4.1). While the percentage of proteins identified in both replicates was high for the sporangia experiments, a lower percentage was observed for the germinating cysts and mycelia, which could possibly be explained by a high number of low abundance proteins that happened to be identified in one of the two replicates.

As in *P. sojae* and *P. ramorum* (Chapter 3), catalases, superoxide dismutases, oxidoreductases, and multiple potential RxLR proteins (77 altogether) were identified in the different life stages of *P. capsici*. As previously described for the proteomic investigation of *P. sojae* and *P. ramorum*, protein abundance was normalized according to Normalized Spectral Abundance Factors (NSAF), followed by clustering of proteins according to their functional annotations. Similar to the observations for the other species, while some global proteomic trends differed between the life stages, most remained constant in magnitude (Figure 4.1), suggesting maintenance of constitutive house keeping and other processes during the entire life cycle.

Table 4.1- Total identified *P. capsici* proteins

Life stage	Replicate 1	Replicate 2	Overlap
Germinating cyst	2674	2728	62%
Mycelium	1487	1399	54%
Sporangium	1223	1045	77%

Figure 4.1- Distribution of protein functional categories in the global proteomes of the germinating cyst, mycelium, and sporangium of *P. cpaisici*. Each identified protein was associated with its KOG annotation and protein categories were clustered based on their KOG functional annotation. The proportion of each protein of the total was calculated based on Normalized Spectral Abundance Factors (NSAF) as previously described[7]. Each pie chart represents the union of duplicate MudPIT experiments. The numbers in parenthesis represent the number of proteins in the respective functional category.

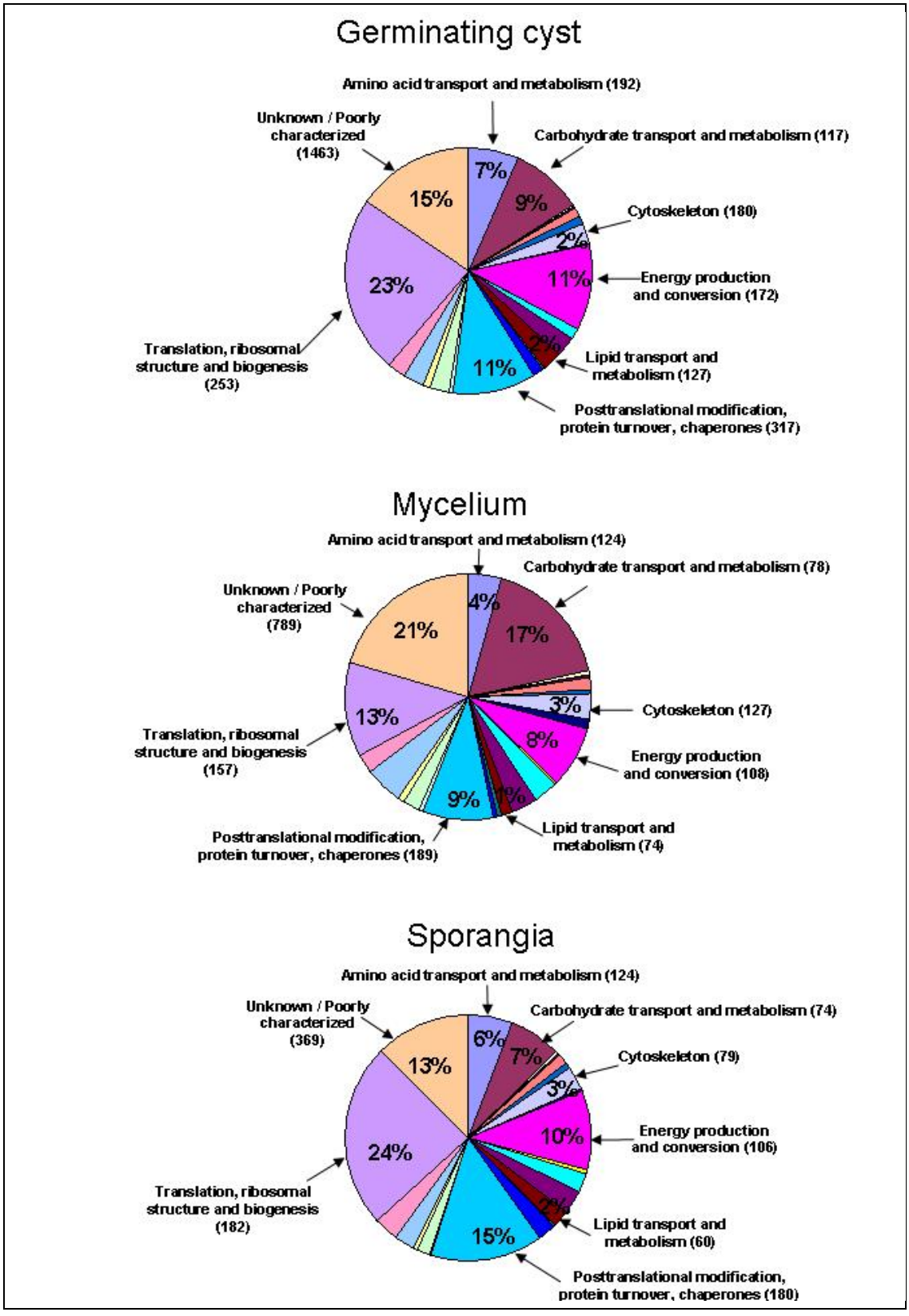


Figure 4.1- Continued

Relative protein expression

As was described for *P. ramorum* and *P. sojae*, the normalized abundance of each protein was used for determination of relative protein abundance. As described before, a protein was considered to be differentially expressed if 2 different criteria were met: (i) The protein was identified in both replicated experiments of the up-regulated life stage. (ii) The protein's NSAF value was at least 5 times greater in the up-regulated life stage. Since 3 different life stages were investigated in this study, they were compared in a binary fashion based on their chronological order in the life cycle. Thus, two binary comparisons were made: germinating cyst to mycelium, and mycelium to sporangium. Figure 4.2 presents the number of differentially expressed proteins from each comparison. Hundreds of life stage specific proteins were identified. When considering the functional category of only those proteins that were differentially expressed in the different life stages, major proteomic trends can be observed (Figure 4.3).

Germinating cyst and sporangium proteomes

Unexpectedly, the proteomic makeup of the germinating cyst and sporangium were remarkably similar. Not only the portion of functional categories of the total were almost identical (Figure 4.3), but also many of the same proteins seemed to be up-regulated to a similar degree in both life stages, as compared to mycelium. The dominating functional category in the germinating cyst and sporangium was *Translation, ribosomal structure and biogenesis*. It included translation elongation factors, ribosomal proteins, tRNA associated proteins, and other proteins involved in protein synthesis. These findings support our previous hypothesis that increased protein synthesis is

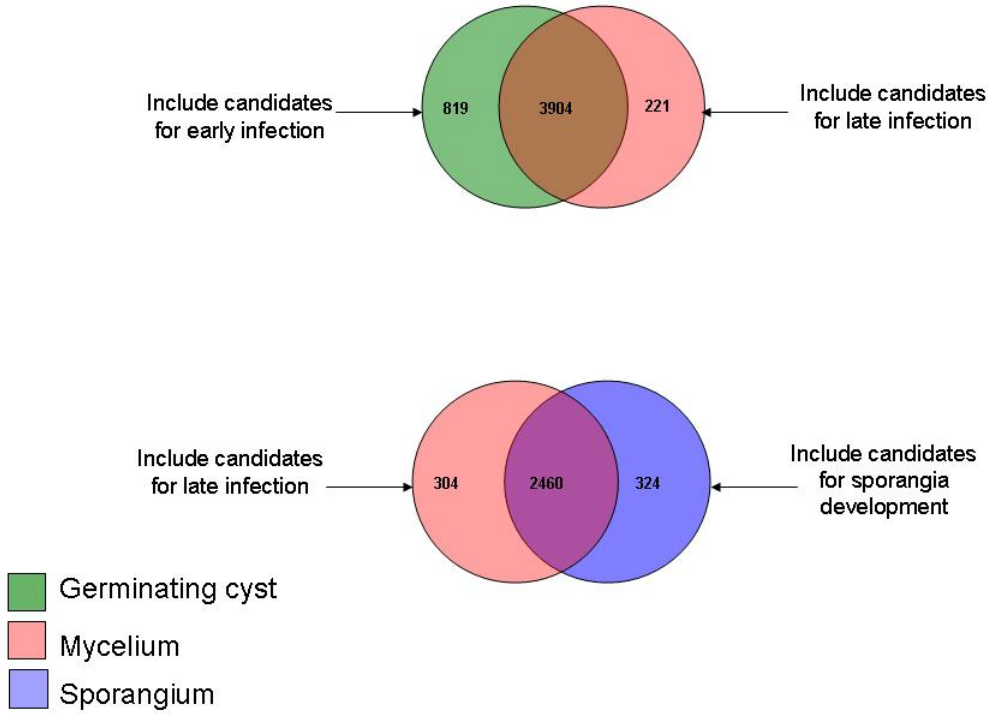


Figure 4.2- Identification of candidate proteins for involvement in early infection, late infection, and sporangium development. The total identified proteomes of the germinating cyst, mycelium, and sporangium were compared. The top Venn diagram represents the comparison between germinating cyst and mycelium. The bottom Venn diagram represents that comparison between mycelium and sporangium. The non-overlapping regions represent the differentially expressed proteins in each life stage, while the overlapping regions represent the rest of the identified proteins.

Figure 4.3- Distribution of protein functional categories in the differentially expressed proteomes of the germinating cyst, mycelium, and sporangium of *P. capsici.* Each identified protein was associated with its KOG annotation and protein categories were clustered based on their KOG functional annotation. The proportion of each protein of the total was calculated based on Normalized Spectral Abundance Factors (NSAF) as previously described[7]. Only proteins that were determined to be differentially expressed (see materials and methods) were included in this analysis. Each pie chart represents the union of duplicate MudPIT experiments. The numbers in parenthesis represent the number of proteins in the respective functional category.

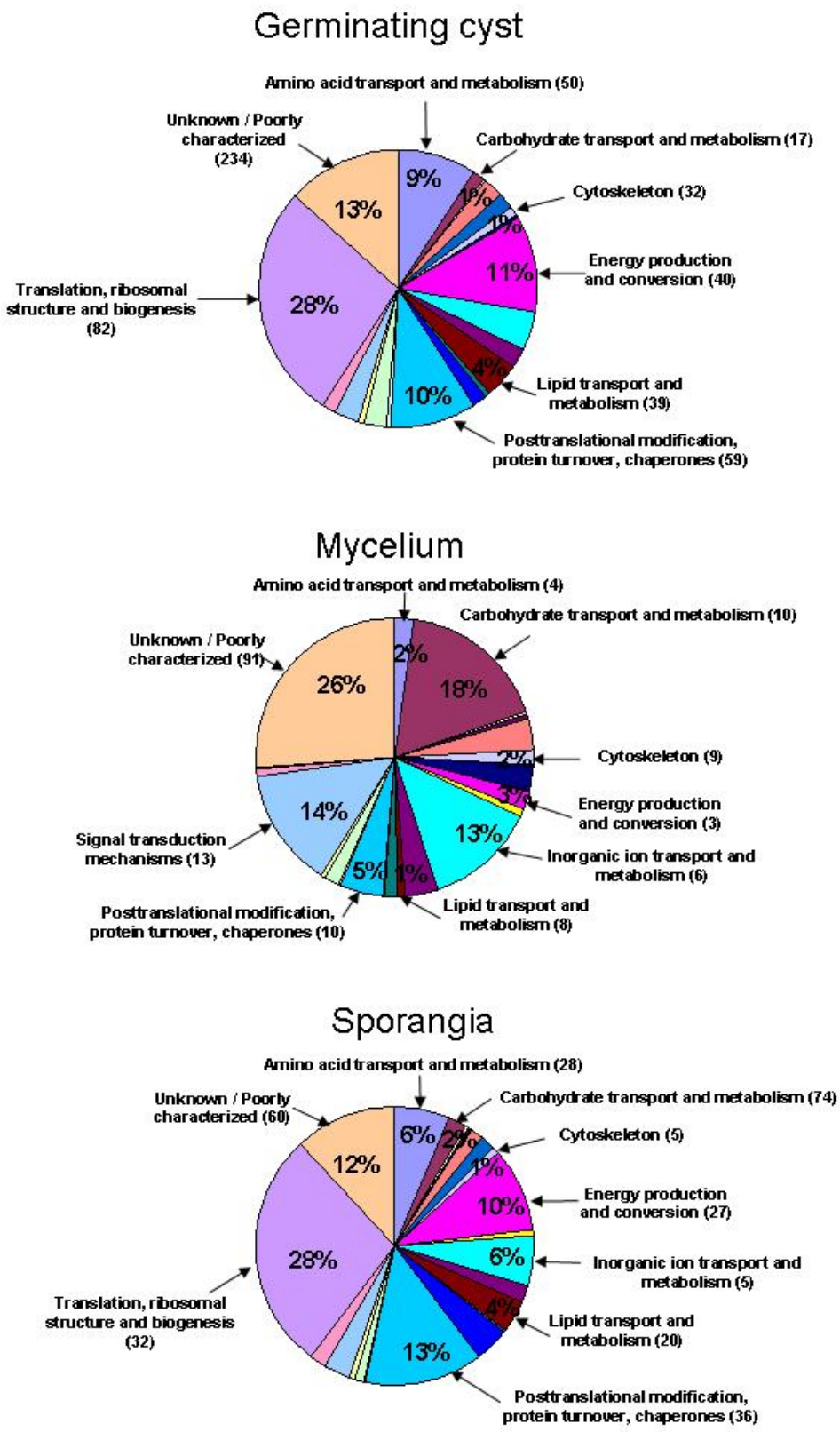


Figure 4.3- Continued

important for cyst germination, as production of new proteins (structural and others) is necessary for construction of the germ tube and new hyphae. The same process might be required for sporangial development as zoospore differentiation inside the sporangium is an active process that might require extensive synthesis of new proteins.

Another functional category including proteins specifically up-regulated in germinating cyst and sporangium was *Lipid transport and metabolism*. As was the case for *P. sojae* and *P. ramorum*, this category included multiple proteins involved in β-oxidation. Up-regulation of these proteins supports the hypothesized catalysis of internal lipid reserves in the germinating cyst and sporangium as an energy source, since both of these life stages are devoid of an external energy source.

A third category that was much more prominent in the germinating cyst and the sporangium than in the mycelium was *Energy production and conversion*. This category included multiple mitochondrial proteins in addition to several nuclear proteins involved in energy production, once again providing evidence for the highly energetic processes occurring in the germinating cyst and sporangium, namely construction of the germ tube and zoospore differentiation, respectively.

Individual protein candidates for early infection were previously identified for *P. sojae* and *P. ramorum* (Chapter 3). *Phytophthora capsici* orthologs of these proteins were also found to be up-regulated in the germinating cyst, further supporting the hypothesis that there are important for early infection. These included secreted proteins with EGF-like domains, a protein containing multiple ricin domains, and alpha-mannosidases.

Mycelium proteome

One of the biggest differences in cellular processes between the mycelium and the other two life stages is in the *Carbohydrate transport and metabolism* category (Figure 4.3). As was previously observed for *P. sojae* and *P. ramorum*, this category is much more abundant in the mycelium than in other life stages. The most abundant protein in this category is a Glyceraldehyde 3-phosphate dehydrogenase, an enzyme involved in glycolysis. Other proteins in this category that were up-regulated in the mycelium included beta-glucocerebrosidases, housekeeping enzymes that are involved in the breakdown of glucocerebroside into a glucose and ceramide. Thus, they represent a catalysis step up-stream of glycolysis. Taken together, these data support the hypothesis that the mycelium obtains its energy through glycolysis of sugars, likely from an external source.

While the *Inorganic ion transport and metabolism* and *Signal transduction mechanisms* categories seemed more abundant in the mycelium than in the other two life stages, a close examination revealed that this increase in abundance is actually due to a family of three proteins (Pc_3604, Pc_8368, and Pc_7393) that were extremely highly up-regulated in this life stage, and whose function is unknown and mis-annotated. This is a group of un-characterized repetitive proteins which will be further described in chapter 6. While several proteins involved in the biosynthesis of amino acids and secondary metabolites were up-regulated in the mycelium, the abundances of their respective categories in *P. capsici* were not as prominent as they were in the *P. sojae* and *P. ramorum* mycelium. This discrepancy may represent a sampling issue or metabolic differences between *P. capsici* and the other two species.

Proteins up-regulated in the mycelium of *P. capsici* also included orthologs of individual proteins that were identified as candidates for late infection in *P. sojae* and *P. ramorum* (Chapter 3). These included the repetitive proteins mentioned above, and a group of 4 annexins.

Candidate proteins involved in host range

In our previous study we tried to identify candidate proteins that might contribute to the ability to infect a broad or a narrow range of hosts. This was done by comparing the proteomes of *P. sojae* that has a narrow host range and *P. ramorum* that has a broad host range. As mentioned earlier, *P. capsici* has a broad range of host plants that it is capable of infecting. Thus, it is expected that the expression patterns of *P. capsici* orthologs of the previously identified *P. ramorum* candidates would be similar, while *P. capsici* orthologs of the previously identified *P. sojae* candidates would either not exist in the *P. capsici* genome or would have a different expression pattern than the *P. sojae* proteins. This indeed was the case for several of the previously identified candidates. For example, Pc_13762 is a nucleoside-diphosphate-sugar epimerase that was up-regulated in *P. capsici* mycelium. Its ortholgous *P. ramorum* protein was also up-regulated in the mycelium and identified as a candidate for broad host range capability due to its predicted function as a carbohydrate modifier and the fact that its *P. sojae* ortholog was not detected in any life stage. Also, many of the crinkling and necrosis inducing *P. sojae* proteins (CRNs) that were identified as candidates specifically for soybean infection (narrow host range) did not have orthologs in the *P. capsici* genome.

Conclusions

A proteomic investigation of three *P. capsici* asexual life stages was performed. The combined results of this study and our previous study of *P. ramorum* and *P. sojae* help support some of the hypothesis made regarding cellular process and metabolic pathways occurring in the different life stage. The data from all 3 organisms suggests that β-oxidation provides the primary energy source for the germinating cyst, while the mycelium uses glycolysis, potentially of host sugars, to obtain its energy. During cyst germination many proteins that are involved in protein synthesis are up-regulated, presumably increasing production of structural and enzymatic proteins that are important for germ tube construction and interaction with the host. The mycelium seems to utilize its derived energy for synthesis or transport of other small molecules such as amino acids and inorganic ions, although the number and abundance of proteins in these categories were lower in *P. capsici* as compared to *P. sojae* and *P. ramorum*.

Surprisingly, the sporangium proteome was strikingly similar to the proteome of the germinating cyst. Not only the magnitude of the functional categories was almost identical between the two life stages, but many of the same proteins were expressed similarly. Like the germinating cyst, proteins involved in β-oxidation and downstream energy production, as well as proteins involved in protein synthesis were up-regulated in the sporangium. This proteomic similarity may reflect similar circumstances in those life stages. While there are obvious differences between the germinating cyst and the sporangium, they are both cellular structures that differentiate outside the plant tissue and therefore in the absence of a surrounding external energy source. Thus, catalysis of internal lipid reserves via the β-oxidation pathway as an energy source may be a common

solution to the same problem. Since cyst germination and zoospore differentiation are rapid processes that result in dramatic morphological changes, both are presumably highly energetic processes requiring large amounts of energy as well as *de novo* synthesis of structural and enzymatic proteins. Thus, the increase in proteins involved in energy production and protein synthesis may satisfy these needs.

In conclusion, the study presented here describes the proteomic profiles of the *P. capsici* germinating cyst, mycelium, and sporangium life stages. When taken in aggregate with previous studies of *P. sojae* and *P. ramorum*, the study presented here adds to the accumulated knowledge of asexual *Phytophthora* life cycle and provides support to previously made hypothesis regarding the *Phytophthora* metabolic pathways, cellular process, and host range involvement of specific proteins.

References

1. Erwin, D. C., and Ribeiro, O. K., *Phytophthora Diseases Worldwide*. The American Phytopathological Society: St. Paul, MN, 1996; p 562.
2. Margulies, M.; Egholm, M.; Altman, W. E.; Attiya, S.; Bader, J. S.; Bemben, L. A.; Berka, J.; Braverman, M. S.; Chen, Y. J.; Chen, Z.; Dewell, S. B.; Du, L.; Fierro, J. M.; Gomes, X. V.; Godwin, B. C.; He, W.; Helgesen, S.; Ho, C. H.; Irzyk, G. P.; Jando, S. C.; Alenquer, M. L.; Jarvie, T. P.; Jirage, K. B.; Kim, J. B.; Knight, J. R.; Lanza, J. R.; Leamon, J. H.; Lefkowitz, S. M.; Lei, M.; Li, J.; Lohman, K. L.; Lu, H.; Makhijani, V. B.; McDade, K. E.; McKenna, M. P.; Myers, E. W.; Nickerson, E.; Nobile, J. R.; Plant, R.; Puc, B. P.; Ronan, M. T.; Roth, G. T.; Sarkis, G. J.; Simons, J. F.; Simpson, J. W.; Srinivasan, M.; Tartaro, K. R.; Tomasz, A.; Vogt, K. A.; Volkmer, G. A.; Wang, S. H.; Wang, Y.; Weiner, M. P.; Yu, P.; Begley, R. F.; Rothberg, J. M., Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **2005,** 437, (7057), 376-80.
3. Savidor, A.; Donahoo, R. S.; Hurtado-Gonzales, O.; Verberkmoes, N. C.; Shah, M. B.; Lamour, K. H.; McDonald, W. H., Expressed Peptide Tags: An Additional Layer of Data for Genome Annotation. *J Proteome Res* **2006,** 5, (11), 3048-3058.
4. Tabb, D. L.; Narasimhan, C.; Strader, M. B.; Hettich, R. L., DBDigger: reorganized proteomic database identification that improves flexibility and speed. *Anal Chem* **2005,** 77, (8), 2464-74.
5. Narasimhan, C., Tabb, D.L., Verberkmoes, N.C., Thompson, M.R., Hettich, R.L., Uberbacher, E.C., MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Analytical chemistry* **2005,** 77, (23), 7581-93.
6. Tabb, D. L.; McDonald, W. H.; Yates, J. R., 3rd, DTASelect and Contrast: tools for assembling and comparing protein identifications from shotgun proteomics. *J Proteome Res* **2002,** 1, (1), 21-6.
7. Florens, L.; Carozza, M. J.; Swanson, S. K.; Fournier, M.; Coleman, M. K.; Workman, J. L.; Washburn, M. P., Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **2006,** 40, (4), 303-11.

Chapter 5

The *Phytophthora capsici*-Tomato Interactome

Abstract

Phytophthora capsici is a vegetable pathogen capable of infecting tomatoes, peppers, and cucumbers among others. Like other *Phytophthora* species *P. capsici* is a hemibiotroph, initiating infection while growing biotrophically before becoming necrotrophic and destroying plant tissue. Multidimensional protein identification technology (MudPIT) was used to measure the "interactome" (the protein complement) of *P. capsici*-infected tomato plants over the course of 3 days of infection. We identified multiple *P. capsici* proteins that may be involved in infection, as well as tomato proteins that may be involved in the response to infection. As infection progresses, tomato metabolism seems to shift from housekeeping mode to defense mode, while *P. capsici* up-regulates proteins that help counteract tomato defenses. Potential specific markers for *P. capsici* nectrotrophy, such as pectate lyase and phenolic acid decarboxylase, were also identified.

Introduction

Phytophthora capsici is a fungal-like destructive plant pathogen with a broad host range. It is capable of infecting many agriculturally important crops including different solanaceae species (e.g. tomatoes, peppers) and cucurbit species (e.g. cucumbers,

squashes). Infection by *Phytophthora capsici* is responsible for heavy economical losses to farmers estimated at many millions of dollars annually worldwide.

The diseases caused by *P. capsici* include foliar blights, fruit rots, and stem and root rots. Like other *Phytophthora* species, *P. capsici* is hemibiotrophic. Thus, diseases caused by *P. capsici* are characterized by an initial biotrophic phase where the pathogen grows intercellularly through the plant tissue but is thought to evade detection by the host plant[1]. The area of infection is typically observed as a water soaked lesion. Later on, pathogen growth is switched to necrotrophic, where the pathogen grows both intercellularly and intracellularly, and plant tissue is necrotized and destroyed[2].

The duration of the biotrophoic phase varies between *Phytophthora* species on different hosts. The biotrophic phase of *P. infestans* on potato lasts about 1-2 days after inoculation[3]. It has been suggested that for *P. sojae* on soybean the transition from biotrophy to necrotrophy occurs between 12 and 24 hours after infection initiation[4]. Histological analysis of *P. parasitica* infected tomato showed that despite the fact that infected plants did not show macroscopic infection symptoms until day 4 after infection, the pathogen was already present in host tissue one day after infection, and kept progressing consistently throughout the plant root until plant death at around day 10[5]. Electron microscopy observation of the progress of invasion of *Phytophthora parasitica* var. *nicotianae* into the root of its tobacco host showed that the epidermis of the plant was colonized 24-48 hours after inoculation, with mainly intracellular mycelial growth of the pathogen. By 72 hours after inoculation the pathogens spread intercellularly through the plant cortex, and by 96 hours after inoculation pathogen growth increased significantly and spread throughout the plant[1]. The biotrophic phase of *P. capsici* is not

well characterized, and one of the goals of this study was to add to the current knowledge of the early infection stage of this pathogen.

Little is known about the molecular mechanisms behind the interplay between *P. capsici* specifically or *Phytophthora* species in general and their plant hosts, although individual genes or proteins involved in such interaction have been identified in different *Phytophthora* species. Several trascriptomic studies have been performed in different *Phytophthora* species (mostly *P. infestans*, *P. nicotianae*, and *P. sojae*) during infection of their hosts in order to identify pathogen genes which are expressed during different stages of the infection[4-10]. During infection of potato, Pieterse et al. identified 9 expressed *P. infestans* genes using differential hybridization [6] and Beyer et al. showed 6 *P. infestans* genes to be up-regulated during infection using suppression subtractive hybridization[11]. *Phytophthora capsici* is under-represented in this group of studies. As for proteomics, a single 2D-PAGE study was carried out on *P. infestans* infected potato, resulting in identification of 5 proteins up-regulated during infection[8]. However, no large-scale proteomic investigation of a *Phytophthora* infected host has been done to date.

Recently, the genome of *P. capsici* has been sequenced, allowing large scale genomic, transcriptomic, and proteomic analysis. No genome of a *Phytophthora* host has been completed to date, but a portion of the genome of tomato (*Lycopersicon esculentum*), a natural host for *P. capsici*, has been sequenced through an international initiative known as the "International Solanaceae Genome Project" (SOL) (http://www.sgn.cornell.edu/index.pl)[12]. While large segments of the tomato genome are yet to be sequenced, the currently available sequence, containing primarily gene dense euchromatin regions, presents a platform on which initial analysis can be based. Indeed,

even at this stage, ~34,000 tomato genes are predicted in the available sequence, based on homology similarity to *Arabidopsis thaliana*. Here we present the first large scale proteomic study of a *Phytophthora* infected host. *Phytophthora capsici* infected tomato was investigated over a time course of 3 days, and both pathogen and host proteins were monitored. Identification and relative abundance determination of both host and pathogen proteins can shed light on the complex molecular interactions that drive pathogen invasion, host defense, and resulting infection.

Materials and methods

Growth of *P. capsici* & tomato plants

Phytophthora capsici strain 1534 (the source for genome sequencing) was grown on UCV8 juice agar plates amended with PARP (100 ppm of pimaricin, 100 ppm of ampicillin, 30 ppm of rifampicin, and 100 ppm of pentachloronitrobezene). Plates were kept under constant fluorescent light at room temperature to induce sporangia formation. One week old sporangia rich cultures were flooded with sterile distilled cold water and kept at 4°C for ~45 min in order to stimulate zoospore release. The water was then collected and filtered through a Kimwipe in order to get rid of mycelium debris. Zoospore suspension was adjusted to ~$1X10^4$ zoospores per ml, and transferred to 200-ml spray bottles for plant inoculation.

Tomato plants (*Lycopersicon esculentum* cultivar Ohio 7814) were grown in pots from seed in a growth chamber at 28 °C, under 14 h-light/10 h-dark cycle. High relative humidity (70%) was maintained using humidity domes. Plants were watered daily or as needed. Growth conditions were maintained prior to and post inoculation.

Plant inoculation

Five week old tomato plants were sprayed evenly with the zoospore mixture to ensure massive and evenly distributed infection throughout the plant. Three different control plants were sprayed with distilled water. At least 10 leaves from each plant were collected from the control (water sprayed) plants (healthy, H), as well as from plants infected with *P. capsici* at one day after inoculation (DAI 1), two days after inoculation (DAI 2), and three days after inoculation (DAI 3). These time points possibly correspond to the different stages of infection: biotrophy (DAI 1), transition phase from biotrophy to necrotrophy (DAI 2), and necrotrophy (DAI 3).

Protein isolation

Two protein fractions were isolated from tomato leaves collected from healthy and infected plants: Intercellular fluids (IF), and total soluble proteins (SP). For each preparation, several leaves from the same plant were pooled together. Intercellular fluids were collected according to the method of de Wit and Spikman[13]. Total soluble proteins fraction was isolated by grinding the leaves left after IF preparation using mortar and pestle with liquid nitrogen. The ground up material was then suspended in lysis buffer and stored in -80°C until further processing.

Protein precipitation

Proteins in both IF and SP fractions were precipitated with addition of 1/3 volume 100% Trichloroacetic acid (Promega, Madison, WI), incubation on ice for 1h, and

centrifugation at 14,000g for 30 min at 4°C. The supernatant was then removed and the pellet was washed twice with ice-cold 100% acetone (Sigma) with 5 min centrifugation at 14,000g at 4°C in between. Finally, the supernatant was removed and the pellet was allowed to air dry.

Protein digestion

Precipitated protein pellets were resuspended in 300 ul 8 M urea and were vigorously vortexed until the pellets dissolved. Not all pellets dissolved completely (in particular SP fractions) and some required additional suspension time before pellets became soluble. Disulphide bonds were reduced by addition of tris(2-carboxyethyl)phosphine (TCEP) to a final concentration of 6.5 mM for 20 min at room temperature, and alkylated by addition of iodoacetamide (IAA) to a final concentration of 2.5 mM for 20 min at room temperature in the dark. A volume of 900 ul of 100 mM Tris buffer, pH 8.5, 1mM $CaCl_2$ was then added to reduce the urea concentration to 2 M. Proteins were then digested overnight with 20 ug of sequencing-grade trypsin (Promega, WI) at 37°C. The following day, another 20 ug of sequencing-grade trypsin was added and again incubated overnight at 37°C. Total soluble protein samples were cleaned and desalted using Sep-Pak light C-18 solid phase extraction (Waters, Milford, MA). Sep-Pak was not used for IF fractions due to their visibly lower protein content. Before loading onto a column, samples were acidified to 4% with formic acid, filtered through Ultrafree-MC (0.45 um filter unit) Microcon tubes (Millipore, MA), and diluted X 2 with buffer A (95% H_2O, 5% ACN, 0.1 FA) to improve loading.

LC/MS/MS

Filtered and acidified samples corresponding to ~0.5 ml of original IF or SP in lysis buffer were loaded offline onto a biphasic 150um inner diameter fused silica column loaded with 3-3.5 cm reverse-phase (RP) material (Aqua, C18, 5um, 200 Å, Phenomenex, Torrance, CA), followed by 3-3.5 cm strong-cation exchange (SCX) material (Luna SCX, 5um 100 Å, Phenomenex, Torrance, CA) as previously described[14, 15]. The columns were packed with the two phases sequentially, and samples then loaded, via a pressure cell (New Objective, Woburn, MA). The loaded column was mounted directly upstream of a 100 um inner-diameter front column packed via pressure cell with 15 cm Jupiter reverse-phase material (Jupiter C18, 3 um, 300 Å, Phenomenex, Torrance, CA). For LC/LC-MS/MS analysis, an Ultimate HPLC (LC Packings, Sunnyvale, CA) was interfaced with a linear ion trap mass spectrometer (LTQ, ThermoElectron, San Jose, CA). The samples were then was analyzed via a 23-h, 12-step MudPIT analysis, as previously described[14, 16, 17], with the exception of a 1 h instead of 2 h first step gradient. Two replicates were performed for each of sample set.

Bioinformatics

Predicted *P. capsici* protein database, as well as KOG and GO functional annotations, were downloaded from the JGI Web site (http://shake.jgi-psf.org/Phyca1/Phyca1.home.html). Predicted tomato protein database, including functional annotations, was supplied by the SOL consortium. The MS/MS data was searched against a concatenated database containing both tomato and *P. capsici* proteins using the DBDigger algorithm[18] in a fully tryptic mode using the MASPIC scorer[19]. For

each protein fraction (IF or SP) a representative experiment data was searched against a concatenated forward and reversed protein database. Based on the search results, filter values corresponding to a 5% false discovery rate were determined using the SQTRevPuller algorithm as previously described[20, 21], and those filter values were applied to the search results of the rest of the experiments of that protein fraction. Identification of two peptides per protein was required for protein identification. Spectral count was considered as the total number of MS/MS spectra corresponding to peptides from a given protein. Spectral counts were normalized based on Normalized Spectral Abundance Factors (NSAF) as previously described[22]. It should be noted that normalization was performed for each species individually (*P. capsici* NSAF value was determined based on the total spectral counts of all *P. capsici* proteins in a single MudPIT experiment, and tomato NSAF value was determined based on the total spectral counts of all tomato proteins in that same MudPIT experiment). Protein groups that shared all of their peptides were assigned a NSAF value equal to the NSAF value for the group divided by the number of proteins in that group. For each protein, an average NSAF value for each day was calculated including all 4 experiments (IF replicate 1, IF replicate 2, SP replicate 1, SP replicate 2) from that day. Using Microsoft Access a relational database was constructed such that each protein was associated with its average NSAF value for each day and its functional annotations.

Results and discussion

Infection of tomato by *P. capsici* results in rapid progress of the disease with complete devastation of the plant by 3 to 4 days after infection (Figure 5.1). The

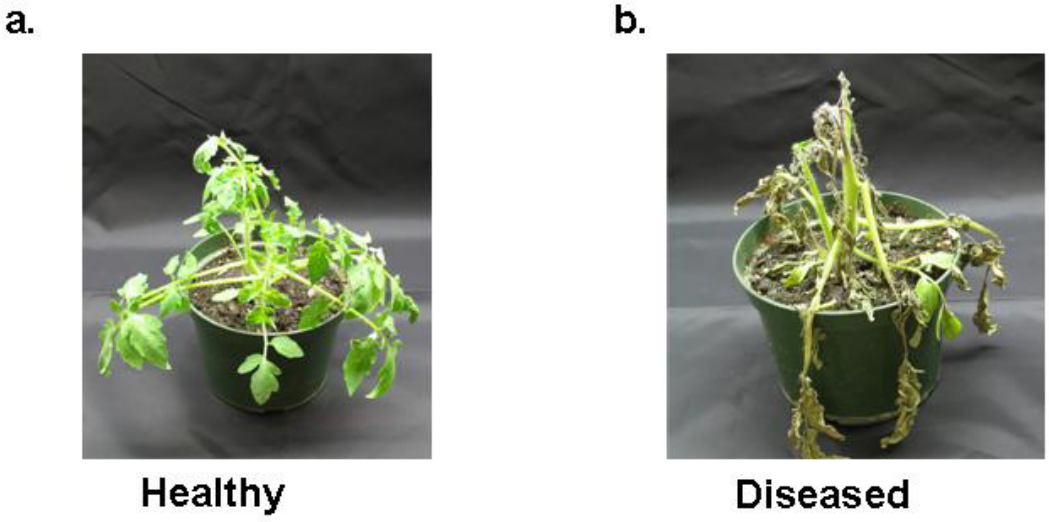


Figure 5.1- *P. capsici* infected tomato plant. A 5 week old healthy tomato plant (a) was sprayed with a concentrated *P. capsici* zoospore solution. Three days later the tomato plant was completely devastated (b).

chemical warfare between the *P. capsici* pathogen and the tomato host is likely represented in their proteomic makeup. We therefore investigated the proteomes of both plant and pathogen during the course of 3 days of infection. Five-week old tomato plants were sprayed evenly with a concentrated solution of *P. capsici* zoospores. A sample of several leaves from a single plant was taken from un-infected plants (control) as well as from infected plants at 1, 2, and 3 days after infection (DAI 1, DAI 2, and DAI 3, respectively). Figure 5.2 shows example of leaves taken from healthy and diseased plants at each sampling point. At day 1 after infection the leaves seemed healthy, with the exception of tiny dark spots at different locations on the leaf (Figure 5.2b). At this stage of the infection *P. capsici* may be growing biotrophically by evading recognition by the plant, therefore avoiding mounting of a defense response by the plant at sites of infection. The tiny dark spots may represent locations where *P. capsici* has invaded, grown, and already turned nectrotrophic, or location where the plant has successfully mounted a defense response locally and produced a hypersensitive response (HR). At day 2 after infection, obvious necrotized lesions appeared on most leaves. The necrotized tissue was brown and desiccated with clear borders separating it from the healthy looking tissue (Figure 5.2c). By day 3 after infection, all leaves were either completely or mostly necrotized (Figure 5.2d).

To probe the plant and pathogen proteomic changes occurring during infection, we analyzed both intercellular fluids (IF) and total soluble proteins (SP) preparations of infected leaves. A separate IF fraction was prepared because much of the interactions between the plant and the pathogen are thought to occur in the intercellular space of the plant (including secreted proteins from both). However, the abundance of these secreted

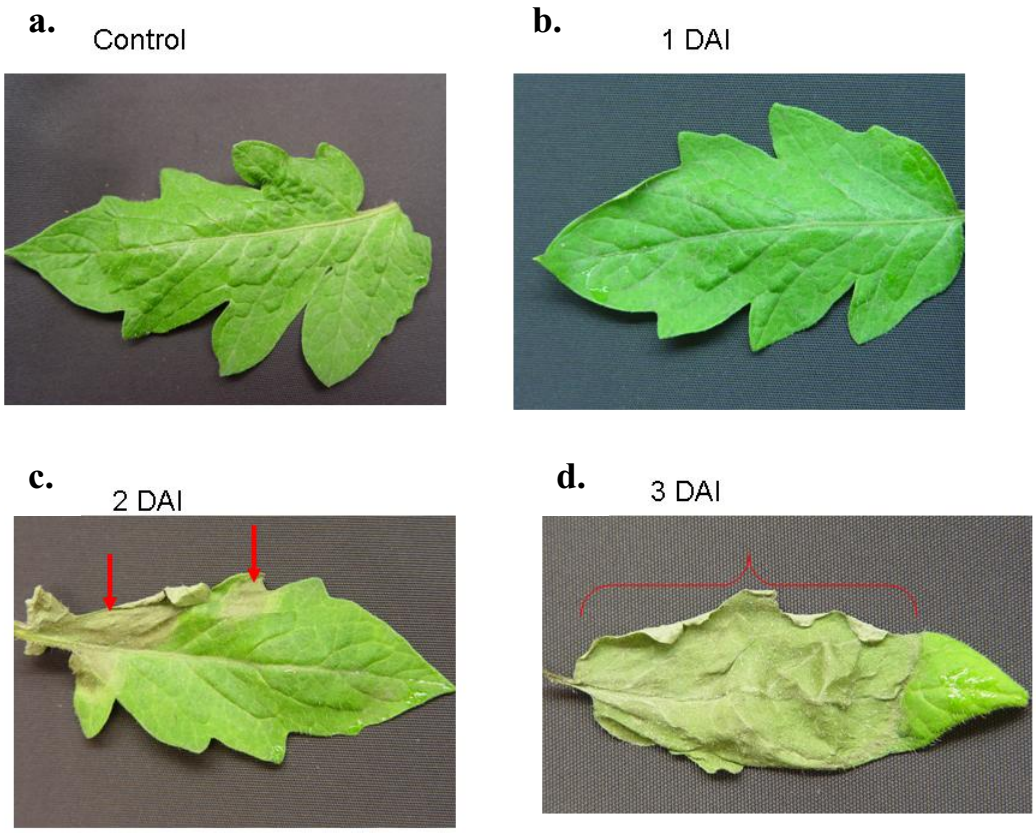


Figure 5.2- *P. capsici* disease progression in tomato. Healthy 5 week old tomato plants were sprayed with a concentrated *P. capsici* zoospore suspension. Leaves were collected from (**a**) un-infected tomato plants (control), as well as from infected plants at (**b**) one day after infection, (**c**) two days after infection, and (**d**) three days after infection. Red arrows and brace indicate the location of necrotized tomato tissue.

proteins is likely very low in comparison to intracellular proteins. Thus, these secreted proteins are likely to be masked by those abundant proteins and elude detection. Extraction of intercellular fluids was therefore aimed at enriching for such secreted proteins. Extraction of intercellular fluids was done on intact infected leaves in order to avoid intracellular proteins as much as possible. After extraction, the same leaves were used for soluble proteins isolation, this time dissecting the leaves around the visible lesions in order to enrich for proteins that are involved in the interaction between the plant and the pathogen, and that might not be expressed similarly at healthy regions of the leaves. Experiments were replicated on each protein preparation from each day. The results of all of the experiments are presented in table 5.1.

Altogether, 5409 proteins were identified from both tomato and *P. capsici*, with the majority of proteins identified (82%) originating from tomato. As expected, the percentage of *P. capsici* proteins of the total number of proteins identified increased consecutively from day 1 to day 3 after infection, as the pathogen grew through the plant tissue and represented a greater mass of the infected tissue (figure 5.3). The abundance of each protein was normalized (see bioinformatics section in materials and methods), and the expression trend was observed.

P. capsici proteins with possible role in early infection

Of the 993 identified *P. capsici* proteins, many did not exhibit any marked change in expression during the 3 day course of infection, while some proteins did. *Phytophthora capsici* proteins that were up-regulated in the early stages of infection may be important for infection initiation and biotrophic phase, while proteins that were up- regulated during

Table 5.1- Total number of proteins identified in MudPIT experiments of *P. capsici* infected tomato leaves

	IF #1	IF #2	SP #1	SP #2
Healthy plant	1398	1250	954	799
DAI 1	829	1093	1571	1344
DAI 2	931	N/A	2046	1897
DAI 3	708	1093	1116	N/A

IF #1- Intercellular fluid, replicate 1
IF #2- Intercellular fluid, replicate 2
SP #1- Soluble protein, replicate 1
SP #2- Soluble protein, replicate 2
DAI- Days after infection
N/A- Not available. Failed experiment

the later stages of infection may be important for colonization and necrotizing of the host tissue. However, as figure 5.3 indicates, the vast majority of peptide and protein identification in day 1 originated from tomato proteins as a result of little pathogen mass in the infected leaf. While normalization of protein abundance corrects relative abundances of detected proteins to some extent, it does not completely solve the problem of proteins below our limits of detection in the early stages of infection. Thus, there are likely *P. capsici* proteins that play a role in early stages of infection but are not abundant enough in the total plant tissue to be detected.

Altogether, 195 *P. capsici* proteins were identified one day after infection, some of which were most abundant at that point as compared to later stages in the infection. Some of these proteins are cytoskeleton proteins (Figure 5.4a) as well as proteins involved in translation (Figure 5.4b). We have previously shown that these classes of proteins are up-regulated in the germinated cyst life stage of *P. sojae*, *P. ramorum* (Chapter 3) and *P. capsici* (Chapter 4). We hypothesized that the processes in which these proteins are involved, namely extensive protein synthesis and generation of cytoskeletal proteins for construction of the germ tube and initial hyphae, play an important part in cyst germination and early infection. The similar pattern observed here during infection further supports this hypothesis.

Another group of proteins that was most abundant during early stages of infection included ubiquitin and ubiquitin-like proteins (Figure 5.4c). Ubiquitin is a highly conserved protein found in all living cells. In a process known as ubiquitination, single or multiple ubiquitin molecules are covalently attached to proteins[23]. Ubiquitination typically targets those modified proteins for degradation by the proteosome[24, 25]. The

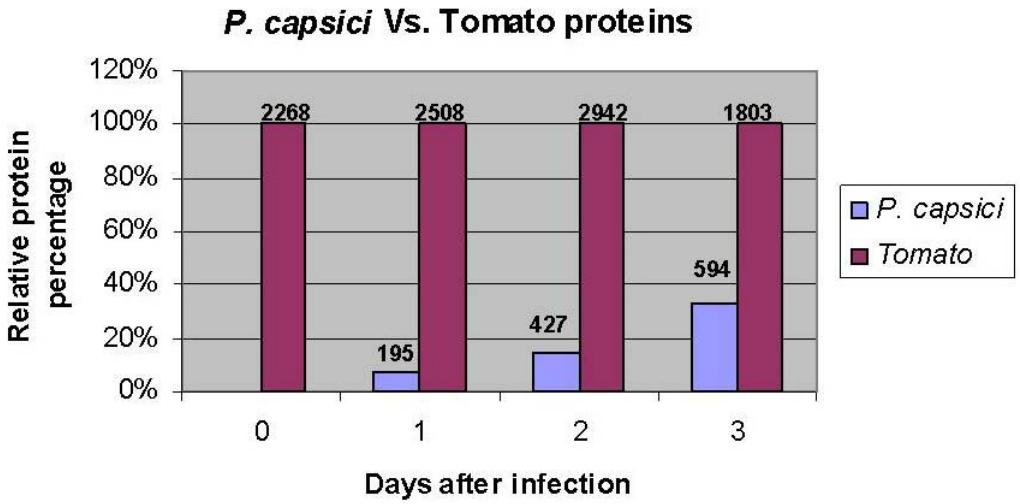


Figure 5.3- Relative abundance of total *P. capsici* and tomato proteins identified. Four MudPIT experiments were carried out on protein preparations (two IF replicates, two SP replicates) from uninfected plants (designated "0 days after infection") as well as from infected plants at day 1, 2, and 3 after infection. The red bars represent the total number of unique tomato proteins identified from all 4 experiments for a single day, normalized to 100%. The blue bars represent the total number of unique *P. capsici* proteins identified from all 4 experiments for a single day, normalized relative to the total tomato proteins. The numbers above the bars represent the total number of unique identified proteins.

increased rate of ubiquitination early in infected suggests degradation of many proteins, some of which may be detected by the host plant and result in HR. Ubiquitin-like proteins are also protein modifiers which are similar in sequence to ubiquitin, but are distinct. Ubiquitin-like proteins function as critical regulators of many cellular processes, including transcription, DNA repair, signal transduction, autophagy, and cell-cycle control[26]. Thus, high expression of ubiquitin and ubiquitin-like proteins in the early stages of infection may be partly responsible for the marked cellular changes occurring in the pathogen cell in the transition from biotrophy to nectrotrophy. It should be noted however that ubiquitin and ubiquitin-like proteins are extremely conserved through evolution. As a result, all identified *P. capsici* peptides derived from this category of proteins could also be found in the tomato proteins of this class. Thus, a change in abundance of these proteins may represent a change in the net abundance of tomato and *P. capsici* ubiquitin proteins, rather than *P. capsici* alone.

While other proteins were identified in day 1 after infection including proteins that are involved in signal transduction, peptide transport, and transcription regulation, as well as proteins of unknown function. However, many were only detected in very low abundance, and conclusions could not be confidently drawn based on them. This is indeed the major challenge for measuring the proteome of a pathogen early in infection within the dominating background of the host.

P. capsici proteins with possible role in late infection

Phytophthora capsici proteins up-regulated during later stages of infection may be responsible for its necrotrophic phase. We showed previously that when grown *in-vitro*,

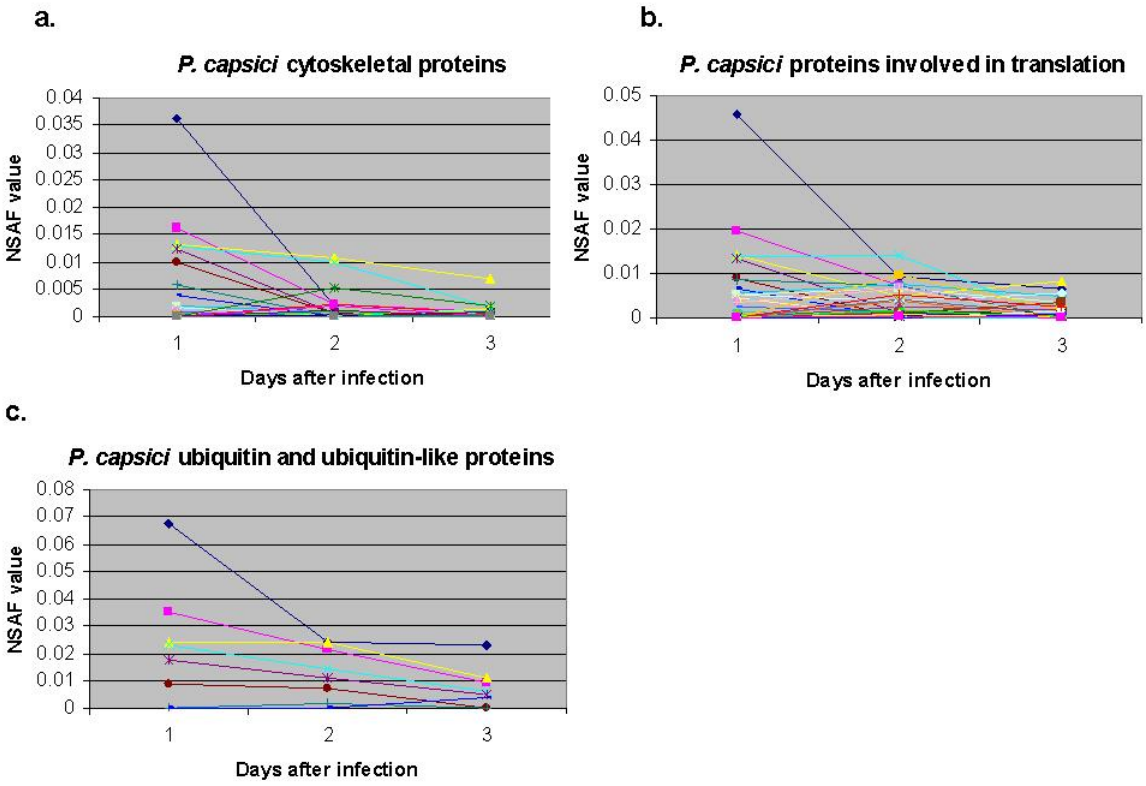


Figure 5.4- *P. capsici* proteins up-regulated in early infection. Four MudPIT experiments were carried out on protein preparations (two IF replicates, two SP replicates) from infected plants at day 1, 2, and 3 after infection. For each experiment the protein abundance was normalized based on Normalized Spectral Abundance Factors (NSAF) as previously described. Each data point in the graph represents the average NSAF value for all 4 experiments for a single *P. capsici* protein in a single day. Expression pattern of proteins of the following functional categories are shown: (a) Cytoskeletal proteins (b) proteins involved in translation (c) Ubiquitin and Ubiquitin-like proteins.

glycolytic enzymes were up-regulated in the mycelium of *P. sojae* and *P. ramorum*. We hypothesized that this is also the case *in planta*, where the pathogen utilized host sugars in the necrotized tissue at later stages of infection. Observation of the glycolytic proteins over the course of *P. capsici* infection of tomato supports this hypothesis, as several glycolytic enzymes were identified and determined to be up-regulated in later stages of infection (Figure 5.5a).

Our previous study has also shown that during cyst germination of *P. sojae*, *P. ramorum* and *P. capsici*, multiple proteins involved in lipid degradation via the β-oxidation pathway were up-regulated. We hypothesized that since zoospore encystment and germination occurs in a nutrient deprived environment outside the plant tissue, the germinating cyst derives its energy from β-oxidation catabolism of its lipid reserves. Surprisingly in this study we observed an increase in β-oxidation proteins in late stages of infection (Figure 5.5b). One possible explanation is that after an initial reduction in β-oxidation activity immediately after cyst germination and lasting for ~1-2 days, β-oxidation resumes, but only in the presence of a host. Perhaps the nutrient rich media in which the mycelium is grown *in vitro* is not conducive for induction of the β-oxidation pathway in *P. capsici*, while the harsher environment (even during nectrotrophy) in the plant tissue is. It is also possible that *P. capsici* imports and catabolizes host lipids that are available in the necrotized tissue, and that the increase in β-oxidation proteins represents lipid catabolism of an external rather than an internal source.

Three *P. capsici* proteins (Pc_12939, Pc_48539, Pc_7649) that exhibited a dramatic increase in abundance constantly during the second and third day of infection included two different manganese superoxide dismutases and one copper/zinc superoxide

Figure 5.5- *P. capsici* proteins up-regulated in late infection. Four MudPIT experiments were carried out on protein preparations (two IF replicates, two SP replicates) from infected plants at day 1, 2, and 3 after infection. For each experiment the protein abundance was normalized based on Normalized Spectral Abundance Factors (NSAF) as previously described. Each data point in the graph represents the average NSAF value for all 4 experiments for a single *P. capsici* protein in a single day. Expression pattern of proteins of the following functional categories are shown: (a) Glycolytic enzymes (b) protein involved in β-oxidation (c) superoxide dismutases (d) catalases (e) phenolic acid decarboxylase (f) phospholipase Ds (g) pectate lyase (h) coproporphyrinogen III oxidases (i) annexins.

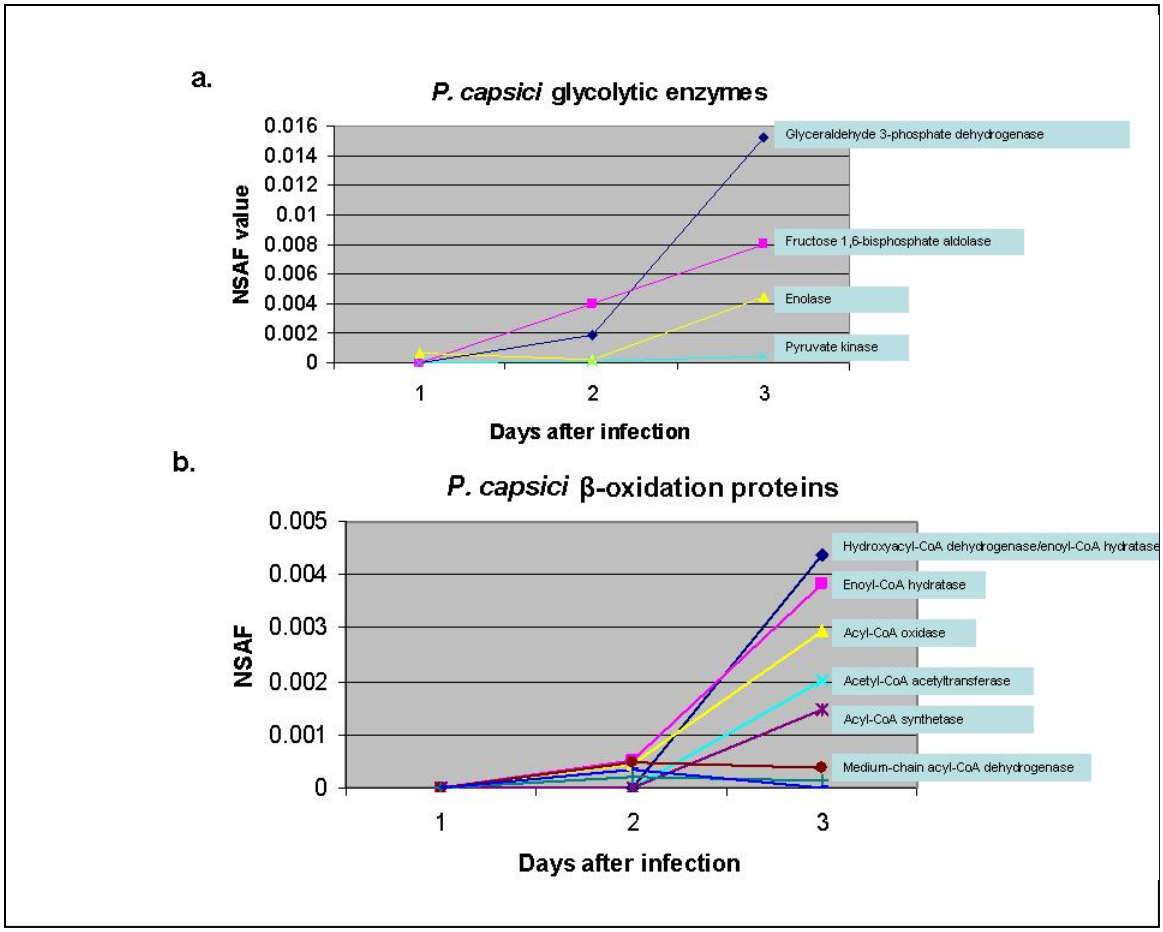


Figure 5.5- Continued

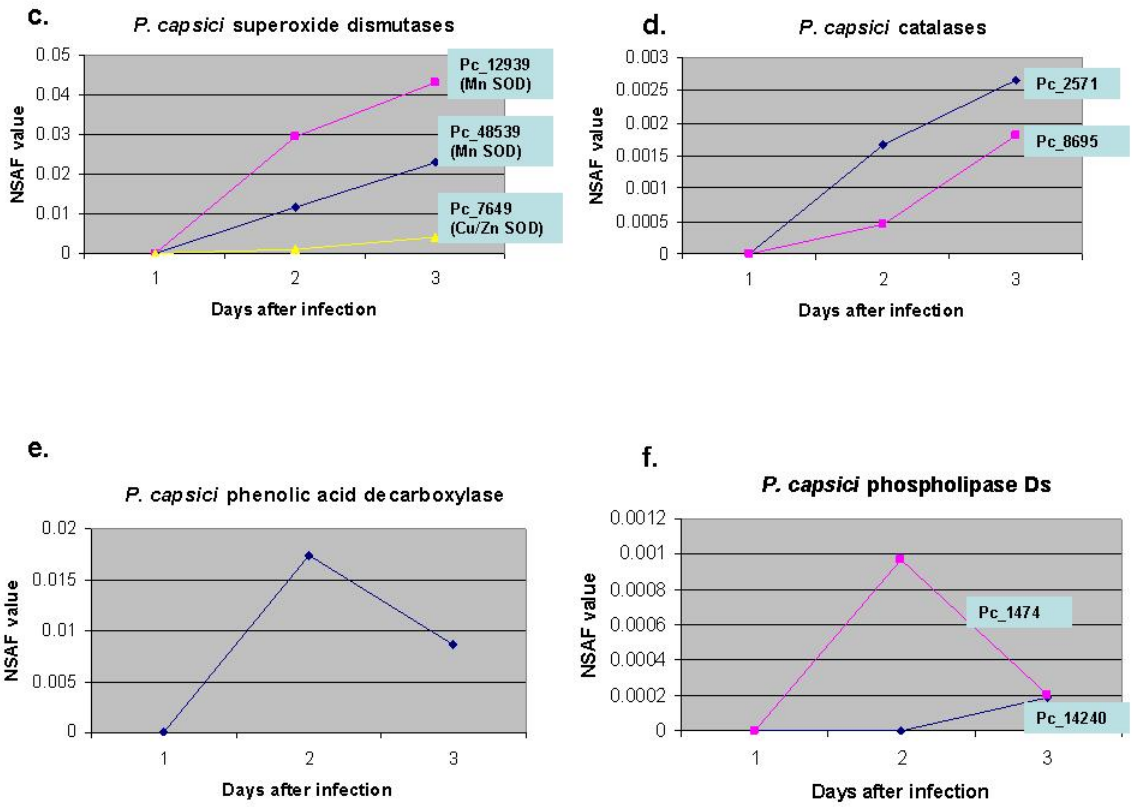


Figure 5.5- Continued

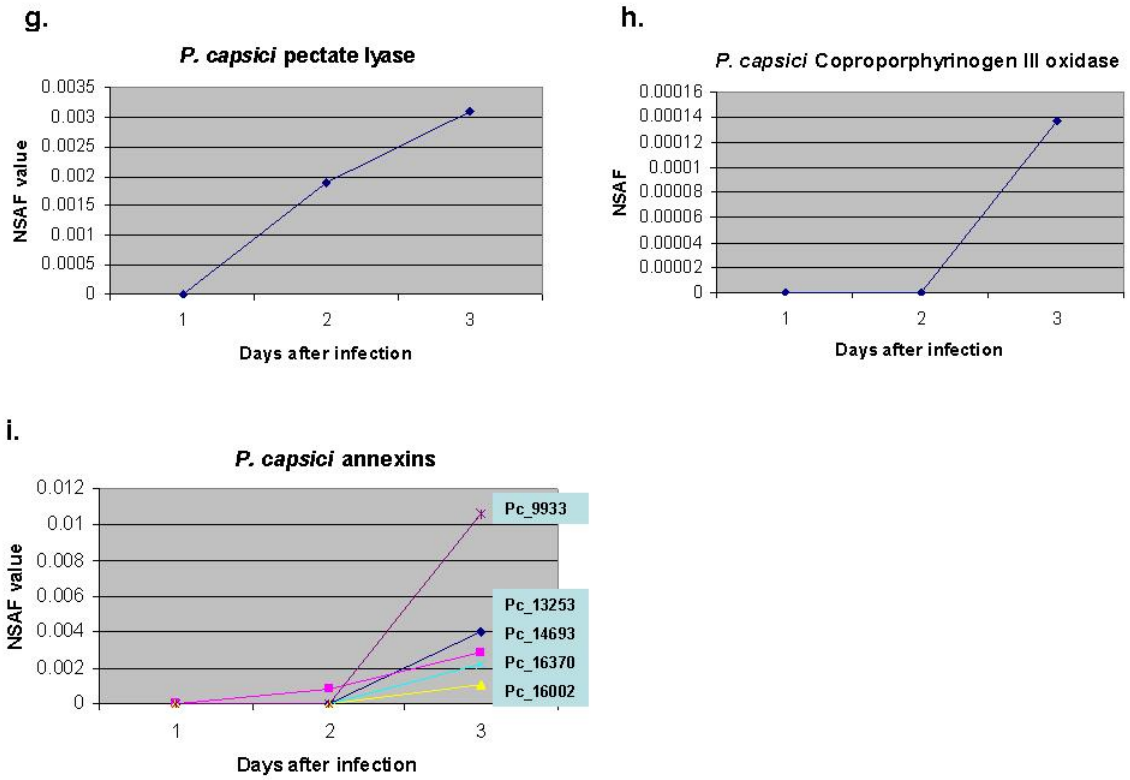


Figure 5.5- Continued

dismutase (Figure 5.5c). Superoxide dismutases are enzymes that catalyze the dismutation of the superoxide anion radical into oxygen and hydrogen peroxide, and therefore serve as major antioxidants countering and neutralizing this toxic compound[27, 28]. Catalases are another group of enzymes scavenging reactive oxygen species, two of which (Pc_2571, Pc_8695) were also found to increase in *P. capsici* during the second and third days of infection (Figure 5.5d). Catalases catalyze the breakdown of hydrogen peroxide into water and oxygen, and are therefore also important antioxidants protecting the living cell[29]. As a part of their defense arsenal, plants use a variety of reactive oxygen species (ROS) to fight invading microorganisms. ROS such as superoxide radicals, hydrogen peroxide, and hydroxyl radicals, can disrupt cellular membranes and cellular metabolism, damage proteins, and induce oxidative changes in DNA[30]. They are therefore damaging and often deleterious to cellular growth and development of living organisms. Upon sensing of a microbial pathogen, plants trigger the so called oxidative burst, in which huge amounts of ROS are produced rapidly and transiently, in an attempt to fend off the invading pathogen[31], or to initiate signal transduction resulting in additional defense responses, including HR[32]. The ability of *P. capsici* to progress with infection despite such potential harsh environment in the tomato tissue may be due to its ability to counter these ROS with the above superoxide dismutases and catalases.

Protein Pc_13566 is another protein that was found to be up-regulated at day 2 and 3 after infection (Figure 5.5e). Highly conserved homologs to this protein were also found in other *Phytophthora* species as well as in other fungal plant pathogens such as *Fusarium graminearum* (a.k.a *Gibberella zeae*), *Phaeosphaeria nodorum*, and *Ustilago*

maydis. While all of these proteins are annotated as conserved hypothetical proteins with an uncertain function, they share domain and sequence similarity to bacterial phenolic acid decarboxylases (PAD). Phenolic acids are aromatic acids that are constituents of plant cell wall, where they bind the complex lignin polymer to the hemicellulose and cellulose[33, 34]. Once released from the cell wall by the activity of hydrolases and esterases, PADs can use these free phenolic acids as their substrate, and convert them into their vinyl phenol derivatives[35, 36]. Thus, up-regulated expression of Pc_13566 during the later stages of infection may be contributing to the necrotrophy of the plant tissue by catalyzing breakdown of phenolic acids and thus promoting plant cell wall degradation.

Another protein that was up-regulated in late infection may contribute to necrotrophy is Pc_32567 (figure 5.5g), an extracellular pectate lyase. This enzyme is responsible for the maceration and soft-rotting of plant tissue by catalyzing the cleavage of de-esterified pectin, a major component of the primary cell walls of many higher plants, and which helps bind plant cells together and regulate water in the plant tissue[37, 38].

Finally, orthologs from three groups of proteins shown to be up-regulated in mycelium of *P. sojae* and *P. ramorum in vitro*[39] were also up-regulated in the late stages of *P. capsici* infection. The first group included 2 phospholipase D (PLDs) proteins (figure 5.5f). PLDs are enzymes that hydrolyze phospholipids, producing choline and phosphatidic acid, a compound heavily involved in signal transduction[40]. Thus, these PLDs may facilitate signal transduction pathways that allow *Phytophthora* to grow through the plant tissue.

Coproporphyrinogen III oxidase is an enzyme that participates in the process of heme biosynthesis[41]. Its up-regulation during the late stages of infection (figure 5.5h), as well as in *P. sojae* and *P. ramorum* mycelia, supports the hypothesis that this protein is important for survival of the pathogen inside the plant tissue, possibly by increasing heme production, and therefore improving respiration in the low oxygen environment in the plant tissue.

The third group of *Phytophthora* proteins that were up-regulated in mycelium *in vitro* and late infection *in planta* are the annexins (figure 5.5i). Annexins are Ca^{2+}-dependent phospholipid binding proteins that are involved in diverse functions such as exocytosis, membrane fusion, phospholipase inhibition, and Ca^{2+}-channel regulation[42-44]. While their function in *Phytophthora* is unknown, they may play an important role in late infection.

Tomato response to *P. capsici* infection

Most of the proteins identified in this set of experiments were tomato proteins (Figure 5.3). Of these, not surprisingly, the most abundant proteins were components of Ribulose-1,5-bisphosphate carboxylase/oxygenase (Rubisco). Rubisco, the most abundant enzyme on earth[45], is an enzyme that catalyzes the first step in carbon fixation in plants, algae, and phototrophic bacteria. It consists of eight large subunits and eight small subunits[46]. The expression trends of both large and small subunits suggest that Rubisco is down-regulated over the course of *P. capsici* infection (Figure 5.6a). Down-regulation of this abundant house-keeping enzyme may suggest a transition from normal

house-keeping operation to defense mode operation, as has been previously observed in tomato response to infection with bacteria[47].

Indeed, other proteins that are associated with defense responses seem to be up-regulated during the course of infection. One such example includes a group of proteins similar to pathogenesis related protein 1 (PR-1) from Arabidopsis (Figure 5.6b). PR-1 exhibits antifungal activity, and is considered an excellent marker of plant disease resistance against pathogens[48]. In fact, over-expression of PR-1a in tobacco increased plant resistance to infection by two *Phytophthora* species[49]. Thus, elevated expression of these proteins may represent an attempt made by the tomato to fend off the invading pathogen.

In contrast to PR-1, another group of pathogenesis related proteins of the Bet v I family showed a different expression pattern (Figure 5.6c). Most of them seemed to be down-regulated at the first day after infection, peak on day 2, and then dramatically decrease in abundance 3 days after infection. Other proteins in this group showed a constant decrease in abundance. While the function of these proteins is unknown, their expression pattern may suggest a response to a specific stage of infection, perhaps the switch from biotrophy to nectrotrophy of *P. capsici* between day 1 and 2 after infection.

Other plant proteins that have previously been shown to be associated with disease resistance include chitinases and β-1,3-glucanases[50]. Both enzymes are digestive enzymes that break down glycosidic bonds. Specifically, chitinases break down chitin, a molecule that is the major component of fungal cell wall. β-1,3-glucanases break down β-1,3-glucans, the major component of Oomycetes cell wall. Tomato proteins including members of both of these categories were up-regulated in later stages of the infection

(Figure 5.6d,e), once again suggesting they are involved in the defense response by the plant.

A group of tomato protease inhibitors was also found to be up-regulated in late stages of the infection (Figure 5.6f). Induction of protease inhibitors appears to be an integral part of plant defense system in response to pathogens, as they were shown to accumulate in plants infected with different pathogens, including tomato infected with *P. infestans*[51], and tobacco plants treated with elicitor preparation obtained from its pathogen *Phytophthora parasitica* var. *nicotianae*[52]. Specifically, protease inhibitors can function in protection against proteolytic attack by an invading pathogen or control endogenous proteases for various purposes[53-55].

While the general trend emerging from these proteomic experiments suggests consistent increase in abundance of groups of plant proteins involved in defense, some plant defense proteins show a different and unexpected pattern of regulation. For example, lipoxygenases (LOXs) are enzymes that catalyse the oxidation of polyunsaturated fatty acids containing a *cis*, *cis*-1,4-pentadiene moiety[56]. In plants, LOXs are induced in response to pathogen attack and upon wounding[57, 58]. LOXs play a key role in the biosynthetic pathway of jasmonic acid (JA), a signal-transduction pathway that underlies induced defenses[59]. The JA pathway is aimed at plant resistance against necrotrophs, and accumulation of jasmonic acid results in activation of proteins with potent antifungal activities[60-62]. Thus, it was expected LOXs would be up-regulated during infection, in particular toward the late stages of infection. In contrast, as could be in figure 6g, tomato LOXs were down-regulated as infection progresses. One possibility

Figure 5.6- Differentially expressed tomato proteins during *P. capsici* infection. Four MudPIT experiments were carried out on protein preparations (two IF replicates, two SP replicates) from uninfected tomato plants (designated "0 days after infection), as well as from infected plants at day 1, 2, and 3 after infection. For each experiment the protein abundance was normalized based on Normalized Spectral Abundance Factors (NSAF) as previously described[7]. Each data point in the graph represents the average NSAF value for all 4 experiments for a single tomato protein in a single day. Expression pattern of proteins of the following functional categories are shown: (a) RUBISCO proteins (b) pathogenesis related PR1 proteins (c) pathogenesis related Bet v I proteins (d) chitinases (e) β-1,3 glucanases (f) protease inhibitors (g) lipoxygenases

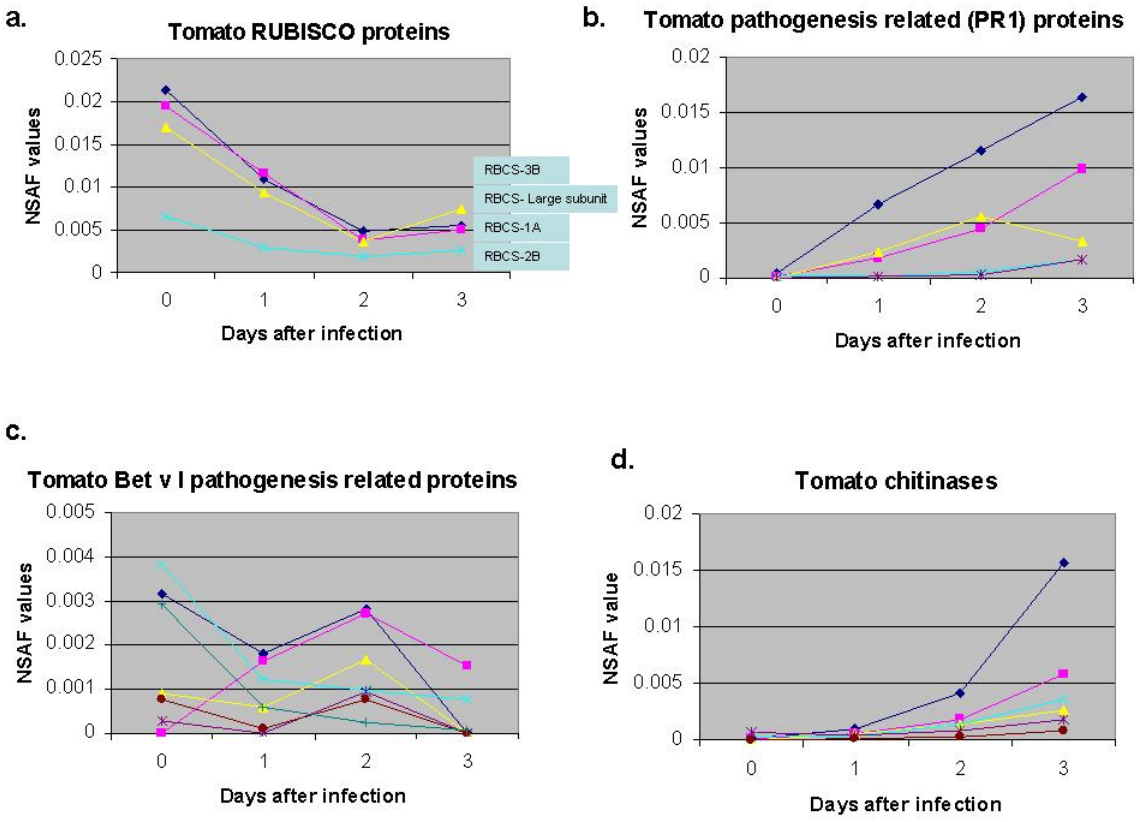


Figure 5.6- Continued

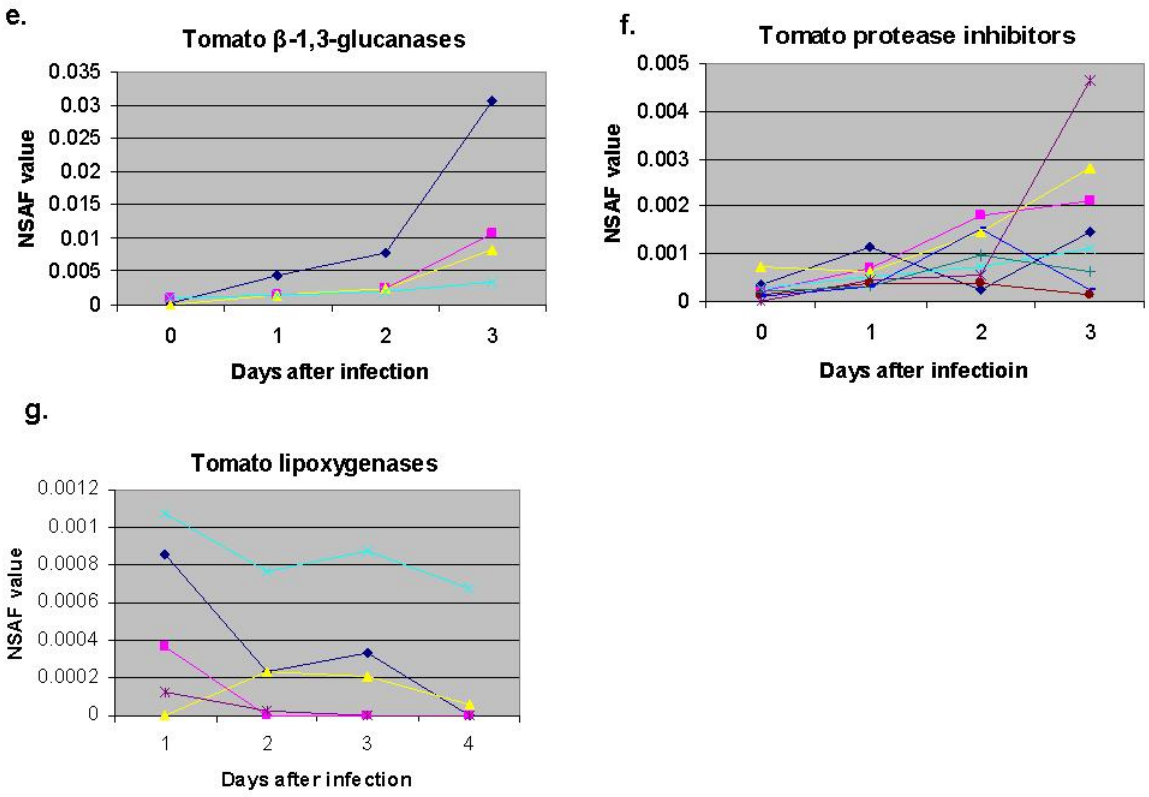


Figure 5.6- Continued

is that the tomato does not recognize *P. capsici* as a nectrotroph, or that *P. capsici* is actively inhibiting the JA pathway in the plant by an unknown mechanism.

Conclusions

The first large scale proteomics study of the interactome of *Phytophthora* and its plant host during infection was performed. Both *P. capsici* and tomato proteins were semi-quantitatively measured over the course of 3 days of infection. During early infection, many *P. capsici* proteins that are involved in protein synthesis and cytoskeleton were up-regulated, in agreement with our previous *in vitro* experiments. Ubiquitin and Ubiquitin-like proteins were also up-regulated in early stages of infection, possibly indicating significant cellular changes occurring in the *Phytophthora* cell in the transition from biotrophy to necrotrophy. Later in infection, *P. capsici* proteins that are associated with countering reactive oxygen species were up-regulated. Such proteins may help *Phytophthora* survival in the infected plant tissue. Phenolic acid decarboxylase and pectate lyase proteins were also up-regulated during late infection, and may represent markers for necrotrophy. Metabolically, catalysis of both sugars and lipids seem to increase in *P. capsici* in late stages of infection by glycolysis and β-oxidation, respectively, possibly indicating utilization of those host molecules present in the necrotized tissue.

During infection, proteomic trends in the tomato are in agreement with previously described shift from house-keeping metabolism to defense mode. As the disease progresses, defense proteins such as pathogenesis related proteins, chitinases, glucanases

and protease inhibitors are up-regulated. Nonetheless, the perception of *Phytophthora* and the mounted response by the tomato is probably too late and ineffective in stopping the already established infection. It has been suggested that even susceptible plants are capable of mounting a defense response, although not quickly enough to halt the pathogen from progressing through infection[1, 63], and our results are in agreement with this hypothesis. It also appears that the plant fails to recognize *P. capsici* as a necrotroph and thus does not trigger the jasmonic acid pathway for combating the pathogen, as evident from down-regulation of tomato lipoxygenases during infection. Another possibility is that despite its recognition, *P. capsici* actively suppresses components of the downstream signaling cascade in the plant, resulting in suppression of the jasmonic acid pathway.

Several challenges are associated with MudPIT experiments carried out on protein preparations from infected plant tissues. The first challenge, as evident from figure 5.3, is dynamic range. While some of the most interesting cellular processes and chemical warfare between the pathogen and the host occur at the very early stages of infection, it is difficult to identify the participating proteins, as they may be in low abundance and localized to a small area of the leaf. When isolating proteins from a leaf early in infection, the majority of proteins are likely collected from a non-infected region of the leaf, and the measured proteomic profile may not accurately represent the true proteome at the localized site of infection. Normalization of the relative abundance of proteins helps mitigate this problem, but it is not effective for proteins that are completely missed due to their relative low abundance.

A second challenge involves protein homology. In a typical proteomic experiment on a single organism, confident unique identification of a protein may be complicated by reliance on identification of redundant peptides that may also exist in other proteins. This is especially true in eukaryotes with extensive gene families and many paralogs. In our experiments, the problem is compounded by the presence of two eukaryotes in the same sample. Not only does each organism contain gene families and paralogs in its genome, but in this case the sample also contains orthologs and other homologs between the host and the pathogen. Thus, identification of peptides from conserved proteins may be ambiguous as for the origin organism of those proteins. To increase confidence in protein identification, identification of unique peptides that are found only in the sequence of one organism, should be done.

A third challenge involves the state of annotation of the genomes of both tomato and *P. capsici*. The tomato genome project is underway and incomplete. While the tomato protein database used for this study includes the majority of proteins estimated to be encoded in the tomato genome, other tomato proteins are likely missing from it. While the genome sequencing of *P. capsici* is complete, the genome annotation is still a work in progress, and many gene models are likely missing or incorrect. Since the method of peptide and protein identification is dependent on correlation of a spectrum to the correct sequence in the protein database, an inaccurate protein database may compromise the quality of protein identification. Thus, increased false positive and false negative rates for peptide and protein identification may be observed, requiring stringent filtering criteria for confident protein identification in order to mitigate some of these effects. As the quality of both protein databases improve, the same spectra from our experiments can be

searched against them to produce better results. Another way to increase confidence in peptide identification in such a complex mixture, and thus identify the true origin of each peptide, would be to perform the same experiments on a higher resolution instrument such as the Orbitrap mass spectrometer.

In conclusion, this study represents the first large scale proteomic investigation of an Oomycete:host interactome over the course of infection. The results from this study provide a basis on which future specific follow-up experiments can be based in order to characterize the role of different pathogen or host proteins in infection as well as serve as a starting point for future large scale proteomics studies.

References

1. Benhamou, N.; Cote, F., Ultrastructure and cytochemistry of pectin and cellulose degradation in tobacco roots infected by Phytophthora parasitica var. nicotianae. *Phytopathology* **1992,** 82, 468-478.
2. Drenth, A.; Goodwin, S. B., Population structure: oomycetes In *In Structure and Dynamics of Fungal Populations*, Worral, J. J., Ed. Kluwer Academic Publishers Dordrech, 1999; pp 195– 224.
3. van West, P.; Vleeshouwers, V. G. A. A., The Phytophthora infestans–host interaction. In *Plant Pathogen Interactions*, Talbot, N. J., Ed. Blackwell Scientific Publishers: Oxford 2004; Vol. 11, pp 219–242.
4. Moy, P.; Qutob, D.; Chapman, B. P.; Atkinson, I.; Gijzen, M., Patterns of gene expression upon infection of soybean plants by Phytophthora sojae. *Mol Plant Microbe Interact* **2004,** 17, (10), 1051-62.
5. Le Berre, J. Y.; Engler, G.; Panabieres, F., Exploration of the late stages of the tomato-Phytophthora parasitica interactions through histological analysis and generation of expressed sequence tags. *New Phytol* **2008,** 177, (2), 480-92.
6. Pieterse, C. M. J.; Riach, M. B. R.; Bleker, T.; van den Berg-Velthuis, G. C. M.; Govers, F., Isolation of putative pathogenicity genes of the potato late blight fungus Phytophthora infestans by differential hybridization of a genomic library *Physiological and Molecular Plant Pathology* **1993,** 43, (1), 69-79.
7. Avrova, A. O.; Venter, E.; Birch, P. R.; Whisson, S. C., Profiling and quantifying differential gene transcription in Phytophthora infestans prior to and during the early stages of potato infection. *Fungal Genet Biol* **2003,** 40, (1), 4-14.
8. Grenville-Briggs, L. J.; Avrova, A. O.; Bruce, C. R.; Williams, A.; Whisson, S. C.; Birch, P. R.; van West, P., Elevated amino acid biosynthesis in Phytophthora infestans during appressorium formation and potato infection. *Fungal Genet Biol* **2005,** 42, (3), 244-56.
9. Qutob, D.; Kamoun, S.; Gijzen, M., Expression of a Phytophthora sojae necrosis-inducing protein occurs during transition from biotrophy to necrotrophy. *Plant J* **2002,** 32, (3), 361-73.
10. Torto-Alalibo, T. A.; Tripathy, S.; Smith, B. M.; Arredondo, F. D.; Zhou, L.; Li, H.; Chibucos, M. C.; Qutob, D.; Gijzen, M.; Mao, C.; Sobral, B. W.; Waugh, M. E.; Mitchell, T. K.; Dean, R. A.; Tyler, B. M., Expressed sequence tags from phytophthora sojae reveal genes specific to development and infection. *Mol Plant Microbe Interact* **2007,** 20, (7), 781-93.
11. Beyer, K.; Jimenez Jimenez, S.; Randall, T. A.; Lam, S.; Binder, A.; Boller, T.; Collinge, M., Characterization of Phytophthora infestans genes regulated during the interaction with potato. *Mol Plant Pathol* **2002,** 3, (6), 473-485.
12. Shibata, D., Genome sequencing and functional genomics approaches in tomato. *J Gen Plant Pathol* **2005,** 71, 1-7.
13. de Wit, P. J. G. M.; Spikman, G., Evidence for the occurrence of race and cultivar-specific elicitors of necrosis in intercellular fluids of compatible interactions of Cladosporium fulvum and tomato. *Physiol. Plant Pathol.* **1982,** 21, 1-11.

14. McDonald, W. H.; Ohi, R.; Miyamoto, D.; Mitchison, T. J.; Yates, J. R., Comparison of three directly coupled HPLC MS/MS strategies for identification of proteins from complex mixtures: single-dimensional LC-MS/MS, 2-phase MudPIT, and 3-phase MudPIT. *Int J Mass Spectrometry* **2002,** 219, 245-251.
15. Ram, R. J., Verberkmoes, N.C., Thelen, M.P., Tyson, G.W., Baker, B.J., Blake, R.C. 2nd, Shah, M., Hettich, R.L., Banfield, J.F., Community proteomics of a natural microbial biofilm. *Science* **2005,** 308, (5730), 1915-20.
16. Washburn, M. P., Wolters, D., and Yates III, J.R., Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nature Biotechnology* **2001,** 19, 242-247.
17. Wolters, D. A., Washburn, M.P., and Yates III, J.R., An automated multidimensional protein identification technology for shotgun proteomics. *Analytical chemistry* **2001,** 73, (23), 5683-90.
18. Tabb, D. L.; Narasimhan, C.; Strader, M. B.; Hettich, R. L., DBDigger: reorganized proteomic database identification that improves flexibility and speed. *Anal Chem* **2005,** 77, (8), 2464-74.
19. Narasimhan, C., Tabb, D.L., Verberkmoes, N.C., Thompson, M.R., Hettich, R.L., Uberbacher, E.C., MASPIC: intensity-based tandem mass spectrometry scoring scheme that improves peptide identification at high confidence. *Analytical chemistry* **2005,** 77, (23), 7581-93.
20. Tabb, D. L.; Shah, M. B.; Strader, M. B.; Connelly, H. M.; Hettich, R. L.; Hurst, G. B., Determination of Peptide and Protein Ion Charge States by Fourier Transformation of Isotope-Resolved Mass Spectra. *Journal of the American Society for Mass Spectrometry* **2006,** 17, (7), 903-915.
21. Savidor, A.; Donahoo, R. S.; Hurtado-Gonzales, O.; Verberkmoes, N. C.; Shah, M. B.; Lamour, K. H.; McDonald, W. H., Expressed Peptide Tags: An Additional Layer of Data for Genome Annotation. *J Proteome Res* **2006,** 5, (11), 3048-3058.
22. Florens, L.; Carozza, M. J.; Swanson, S. K.; Fournier, M.; Coleman, M. K.; Workman, J. L.; Washburn, M. P., Analyzing chromatin remodeling complexes using shotgun proteomics and normalized spectral abundance factors. *Methods* **2006,** 40, (4), 303-11.
23. Ciechanover, A., The ubiquitin-proteasome proteolytic pathway. *Cell* **1994,** 79, (1), 13-21.
24. Hershko, A.; Ciechanover, A., The ubiquitin system for protein degradation. *Annu Rev Biochem* **1992,** 61, 761-807.
25. Hochstrasser, M., Ubiquitin, proteasomes, and the regulation of intracellular protein degradation. *Curr Opin Cell Biol* **1995,** 7, (2), 215-23.
26. Kerscher, O.; Felberbaum, R.; Hochstrasser, M., Modification of proteins by ubiquitin and ubiquitin-like proteins. *Annu Rev Cell Dev Biol* **2006,** 22, 159-80.
27. Bannister, J. V.; Bannister, W. H.; Rotilio, G., Aspects of the structure, function, and applications of superoxide dismutase. *CRC Crit Rev Biochem* **1987,** 22, (2), 111-80.
28. Fridovich, I., Superoxide anion radical (O2-.), superoxide dismutases, and related matters. *J Biol Chem* **1997,** 272, (30), 18515-7.
29. Chelikani, P.; Fita, I.; Loewen, P. C., Diversity of structures and properties among catalases. *Cell Mol Life Sci* **2004,** 61, (2), 192-208.

30. Halliwell, B.; Gutteridge, J. M. C., *Free Radicals in Biology and Medicine*. Oxford University Press: Oxford, 1998.
31. Wojtaszek, P., Oxidative burst: an early plant response to pathogen infection. *Biochem J* **1997,** 322 (Pt 3), 681-92.
32. Tenhaken, R.; Levine, A.; Brisson, L. F.; Dixon, R. A.; Lamb, C., Function of the oxidative burst in hypersensitive disease resistance. *Proc Natl Acad Sci U S A* **1995,** 92, (10), 4158-63.
33. Faulds, C. B.; Williamson, G., The role of hydroxycinnamates in the plant cell wall. *J Agric Food Chem* **1999,** 79, (3), 393-395.
34. Klepacka, J.; Fornal, L., Ferulic acid and its position among the phenolic compounds of wheat. *Crit Rev Food Sci Nutr* **2006,** 46, (8), 639-47.
35. Christov, L. P.; Prior, B. A., Esterases of xylan-degrading microorganisms: production, properties, and significance. *Enzyme Microb Technol* **1993,** 15, (6), 460-75.
36. Barthelmebs, L.; Divies, C.; Cavin, J. F., Expression in Escherichia coli of native and chimeric phenolic acid decarboxylases with modified enzymatic activities and method for screening recombinant E. coli strains expressing these enzymes. *Appl Environ Microbiol* **2001,** 67, (3), 1063-9.
37. Carpita, N. C.; Gibeaut, D. M., Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth. *Plant J* **1993,** 3, (1), 1-30.
38. Marin-Rodriguez, M. C.; Orchard, J.; Seymour, G. B., Pectate lyases, cell wall degradation and fruit softening. *J Exp Bot* **2002,** 53, (377), 2115-9.
39. Savidor, A.; Donahoo, R. S.; Hurtado-Gonzales, O.; Land, M. L.; Shah, M. B.; Lamour, K. H.; McDonald, W. H., Cross-species global proteomics reveals conserved and unique processes in Phytophthora sojae and P. ramorum. *Mol Cell Proteomics* **2008**.
40. Cazzolli, R.; Shemon, A. N.; Fang, M. Q.; Hughes, W. E., Phospholipid signalling through phospholipase D and phosphatidic acid. *IUBMB Life* **2006,** 58, (8), 457-61.
41. Xu, K.; Elliott, T., An oxygen-dependent coproporphyrinogen oxidase encoded by the hemF gene of Salmonella typhimurium. *J Bacteriol* **1993,** 175, (16), 4990-9.
42. Gerke, V.; Moss, S. E., Annexins: From structure to function. *Physiol. Rev.* **2002,** 82, (2), 331-371.
43. Raynal, P.; Pollard, H. B., Annexins: the problem of assessing the biological role for a gene family of multifunctional calcium- and phospholipid-binding proteins. *Biochim Biophys Acta* **1994,** 1197, (1), 63-93.
44. Romisch, J.; Paques, E. P., Annexins: calcium-binding proteins of multi-functional importance? *Med Microbiol Immunol* **1991,** 180, (3), 109-26.
45. Dhingra, A.; Portis, A. R., Jr.; Daniell, H., Enhanced translation of a chloroplast-expressed RbcS gene restores small subunit levels and photosynthesis in nuclear RbcS antisense plants. *Proc Natl Acad Sci U S A* **2004,** 101, (16), 6315-20.
46. Houtz, R. L.; Portis, A. R., Jr., The life of ribulose 1,5-bisphosphate carboxylase/oxygenase--posttranslational facts and mysteries. *Arch Biochem Biophys* **2003,** 414, (2), 150-8.
47. Balaji, V.; Mayrose, M.; Sherf, O.; Jacob-Hirsch, J.; Eichenlaub, R.; Iraki, N.; Manulis-Sasson, S.; Rechavi, G.; Barash, I.; Sessa, G., Tomato transcriptional changes in

response to Clavibacter michiganensis subsp. michiganensis reveal a role for ethylene in disease development. *Plant Physiol* **2008**.
48. Ryals, J. A.; Neuenschwander, U. H.; Willits, M. G.; Molina, A.; Steiner, H. Y.; Hunt, M. D., Systemic Acquired Resistance. *Plant Cell* **1996,** 8, (10), 1809-1819.
49. Alexander, D.; Goodman, R. M.; Gut-Rella, M.; Glascock, C.; Weymann, K.; Friedrich, L.; Maddox, D.; Ahl-Goy, P.; Luntz, T.; Ward, E.; et al., Increased tolerance to two oomycete pathogens in transgenic tobacco expressing pathogenesis-related protein 1a. *Proc Natl Acad Sci U S A* **1993,** 90, (15), 7327-31.
50. Mauch, F.; Mauch-Mani, B.; Boller, T., Antifungal hydrolases in pea tissue. II. lnhibition of funga1 growth by combinations of chitinase and fi-1,3-glucanase. *Plant Physiol* **1998**, (88), 936-942.
51. Peng, J. H.; Black, L. L., Increased proteinase inhibitor activity in response to infection of resistant tomato plants by Phytophthora infestans. *Phytopathology* **1976,** 66, 958-963.
52. Rickauer, M.; Fournier, J.; Esquerre-Tugaye, M. T., Induction of Proteinase Inhibitors in Tobacco Cell Suspension Culture by Elicitors of Phytophthora parasitica var. nicotianae. *Plant Physiol* **1989,** 90, (3), 1065-1070.
53. Richardson, M., The proteinase inhibitors of plants and micro-organisms *Phytochemistry* **1977,** 16, (2), 159-169.
54. van der Hoorn, R. A.; Jones, J. D., The plant proteolytic machinery and its role in defence. *Curr Opin Plant Biol* **2004,** 7, (4), 400-7.
55. Christeller, J.; Laing, W., Plant serine proteinase inhibitors. *Protein Pept Lett* **2005,** 12, (5), 439-47.
56. Brash, A. R., Lipoxygenases: occurrence, functions, catalysis, and acquisition of substrate. *J Biol Chem* **1999,** 274, (34), 23679-82.
57. Rosahl, S.; Feussner, I., Plant Lipids. In Murphy, D., Ed. Blackwell Publishing Ltd.: Oxford, U.K., 2005; pp 329–354.
58. Chen, X.; Shen, G.; Wang, Y.; Zheng, X.; Wang, Y., Identification of Phytophthora sojae genes upregulated during the early stage of soybean infection. *FEMS Microbiol Lett* **2007,** 269, (2), 280-8.
59. Kessler, A.; Baldwin, I. T., Plant responses to insect herbivory: the emerging molecular analysis. *Annu Rev Plant Biol* **2002,** 53, 299-328.
60. Penninckx, I. A.; Eggermont, K.; Terras, F. R.; Thomma, B. P.; De Samblanx, G. W.; Buchala, A.; Metraux, J. P.; Manners, J. M.; Broekaert, W. F., Pathogen-induced systemic activation of a plant defensin gene in Arabidopsis follows a salicylic acid-independent pathway. *Plant Cell* **1996,** 8, (12), 2309-23.
61. Lorenzo, O.; Piqueras, R.; Sanchez-Serrano, J. J.; Solano, R., ETHYLENE RESPONSE FACTOR1 integrates signals from ethylene and jasmonate pathways in plant defense. *Plant Cell* **2003,** 15, (1), 165-78.
62. Spoel, S. H.; Johnson, J. S.; Dong, X., Regulation of tradeoffs between plant defenses against pathogens with different lifestyles. *Proc Natl Acad Sci U S A* **2007,** 104, (47), 18842-7.
63. Kuc, J., *Innovative Approaches to Plant Disease Control*. John Wiley & Sons: New York, 1987; p 255-274.

64. Buddc, A. D.; Helgeson, J. P., Phytoalexins in tobacco callus tissue challanged by zoospores of Phytophthora parasitica var. nicotianae. *Phytopathology* **1981,** 71, 206.

Chapter 6

Conclusions

The study presented here represents the first large scale proteomic study on *Phytophthora*. While much of the available literature on the Oomycetes is focused on *P. infestans* and *P. sojae*, we expanded the range of species to also include *P. ramorum* and *P. capsici*, two under-studied pathogens of high ecological and agricultural importance.

We were interested in the mechanism of *Phytophthora* pathogenesis, and in particular in early infection. To investigate the infection process we carried out both *in vitro* and *in planta* experiments. The *in vitro* experiments included isolation of three asexual life stages-- germinating cyst, mycelium, and sporangium, followed by preparation and analysis of their proteomes. The *in planta* experiments included proteomic investigation of a timeline of *P. capsici* infected tomato plants. The combined data from all of these experiments provided new insights into *Phytophthora* biology and pathogenesis, helped generate new hypothesis and test previously formulated hypothesis.

As means to answering biological questions, we developed different techniques that incorporated biology, mass spectrometry, and bioinformatics. These techniques are not limited to the study of *Phytophthora*, but can be applied essentially to any organism, provided that sufficient genome information is available. Chapter 2 describes the development of one such technique, EPT mapping, for improvement of genome annotation. EPT mapping was developed to address the first specific aim of this study, which involved the use of our proteomics data to improve the genome annotation of

Phytophthora species. Application of this technique to different *Phytophthora* species supported many predicted genes models, expanded many others, and identified multiple new genes that were previously missed by the computational genome annotation. The results of the analysis were biologically validated for some individual genes, and demonstrated the effectiveness of this technique.

The other two specific aims defined for this study intended to identify proteomic trends and individual proteins which are responsible for *Phytophthora* infection and for definition of host range. Chapters 3-5 address these aims. The recent sequencing several *Phytophthora* species provided us with a unique opportunity to incorporate proteomics and orthology of those closely related eukaryotes in order to identify proteins of interest. Chapter 3 describes a cross-species proteomic technique using closely related species for identification of protein trends and protein candidates important for *Phytophthora* pathogenesis. Chapter 4, which includes proteomics experiments on 3 asexual life stages from *P. capsici*, builds on the accumulated data obtained from the experiments on *P. sojae* and *P. ramorum* described in chapter 3. In addition to identification of candidate proteins important for early or late infection, those experiments also identified metabolic trends that are common to all of these species, including β-oxidation of internal lipid reserves during cyst germination, and glycolysis as the main energy producing process during vegetative growth. The *in planta* experiments described in chapter 5 are also aimed at identification of *Phytophthora* proteins involved in pathogenesis, and also provide insight into the molecular response of the tomato plant host to *Phytophthora*. The data suggests that during *P. capsici* infection of tomato, the host plant fails to recognize the invading hemibiotrophic pathogen in a timely fashion, and therefore fails to activate

adequate defense responses, such as the jasmonic acid pathway, to fend off infection. As infection progresses, the plant does sense the invasion and transits from house keeping mode to defense mode. However this response is too little too late, as the pathogen has already established infection and possesses tools, such as ROS scavenging enzymes, to combat plant defenses. Possible specific markers of *Phytophthora* necrotrophy, such as pectate lyase and phenolic acid decarboxylase, were also identified and may contribute to our understanding of hemibiotrophy.

As identification of hundreds and even thousands of proteins in a single experiment has become routine in the last several years, the need for biological meaningful querying of the data is in growing demand. The techniques that were developed and applied to the sequenced *Phytophthora* species in this study can be applied to many other prokaryotic and eukaryotic organisms which are expected to be sequenced in the future. Specifically for *Phytophthora* research, the raw and analyzed data from most of this study is or will be publicly available, providing the research community with a platform on which to build on and expand.

Vita

Alon Savidor was born in Tel-Aviv, Israel on May 17th, 1977. He was raised by his parents, Eliezer and Michal Savidor, in Ramat Hasharon, Israel, where he graduated from Rotberg High School in 1995. He served in the Israeli Defense Force from 1995 to 1998. He enrolled at Southern Illinois University in 1999, and graduated with a B.S. in Microbiology and minors in Computer science and Chemistry in 2003. He enrolled in the University of Tennessee-Oak Ridge National Laboratory Graduate School of Genome Science and Technology in 2003 to pursue his doctorate in Life Sciences. He graduated with a Ph.D. in 2008.