UC Riverside

UC Riverside Electronic Theses and Dissertations

Title

Gene Structure and Variation in Phytophthora infestans: Resources for Understanding and Managing a Global Threat to Food Security

Permalink

https://escholarship.org/uc/item/72w4g583

Author

Shrivastava, Jolly

Publication Date

2017

Supplemental Material

https://escholarship.org/uc/item/72w4g583#supplemental

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, availalbe at https://creativecommons.org/licenses/by/4.0/

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA
RIVERSIDE

Gene Structure and Variation in *Phytophthora infestans*: Resources for Understanding and Managing a Global Threat to Food Security

A Dissertation submitted in partial satisfaction
of the requirements for the degree of

Doctor of Philosophy

in

Genetics, Genomics and Bioinformatics

by

Jolly Shrivastava

September 2017

Dissertation committee:
Howard S. Judelson, Chairperson
Thomas Girke
Wenxiu Ma

Copyright by
Jolly Shrivastava
2017

The Dissertation of Jolly Shrivastava is approved:

_____

_____

_____

Committee Chairperson

University of California, Riverside

ABSTRACT OF THE DISSERTATION

Gene Structure and Variation in *Phytophthora Infestans*: Resources for Understanding and Managing a Global Threat to Food Security

by

Jolly Shrivastava

Doctor of Philosophy, Graduate Program in Genetics, Genomics and Bioinformatics
University of California, Riverside, September 2017
Dr. Howard Judelson, Chairperson

Genomic studies on pathogens have given insights to multiple horizons in research that include but are not limited to transcriptional landscaping, host-pathogen interactions, and evolution. These studies have elucidated the role of variation in pathogenic and non-pathogenic strains or species-specific adaptations. We have used genomic and RNA-Seq data to give insight to the genome of *Phytophthora infestans*, that causes late-blight disease in potato and tomato plants. This study has tried to correctly diagnose *P. infestans* in the fields, controlling it by elucidating the variation present in different isolates of *P. infestans* and understanding the mode of evolution among different oomycetes and finally understanding the transcriptional landscape of this pathogen. We have corrected the genome annotation of *P. infestans* helping us to identify pathogenicity related genes and better functional and structural annotation of the genome. We have also identified regions within the genome of different isolates of *P. infestans* that show variation and might help explain the differences that exist in terms of pathogenicity and host

preferences. Apart from this we have also identified differences and similarities in gene expression levels in different oomycetes in different developmental stages.

TABLE OF CONTENTS

LIST OF FIGURES

INTRODUCTION

CHAPTER I

CHAPTER II

CHAPTER III

CHAPTER IV

CHAPTER V

LIST OF TABLES

CHAPTER I

CHAPTER II

CHAPTER III

Introduction

Bioinformatics has been used in the study of microbes in many ways. This includes for sequencing technologies, *in silico* predictions for protein-protein or protein-DNA interactions, performing comparative genomics for gene predictions or functional predictions, and generating databases for genomic and proteomic analysis. Genomic studies on pathogens have given insights to multiple horizons in research that include but are not limited to transcriptional landscaping, genome wide association-studies, host-pathogen interactions, and evolution. These studies have elucidated the role of variation in pathogenic and non-pathogenic strains or species-specific adaptations. Different studies have also identified cellular processes involved in virulence, or role of non-coding RNA and metabolites in regulating the gene expression or host-pathogen interactions.

Pathogen evolution resulting in resistant strains need novel therapeutic strategies to fight the biological threats posed by them. Reduction in sequencing costs have made it easier to gain insight into the genome of these pathogens as more and more strains are being sequenced.

In this study, we have used genomic and RNA-Seq data to give insight to the genome of some of the oomycetes, specifically *Phytophthora infestans*, which is the causal agent of the infamous potato famine in Ireland in 1840. Very few studies have been done at the genome level in *P. infestans,* most of which are directed towards the effector biology. We have used comparative genomics, custom perl scripts and bioinformatic tools such search and alignment algorithms, ortholog identification, variant

annotation and expression calling to help us understand the biology of *P. infestans* in terms of variation, transcriptional landscape, and evolution of its life-style*s*.

Oomycetes: Oomycetes, also known as water molds due to their preference for wet or moist environments, consist of a large number of microorganisms that include saprophytes, plant and animal pathogens and also microorganisms. Oomycetes got their name due to the shape of round oogonia, which contain female gametes.

Oomycetes and fungi have emerged to be some of the biggest threats to food security in current times (Fisher et al., 2012). Diseases caused by oomycetes include late blight of potato caused by *Phytophthora* infestans, sudden oak death caused by *P. ramorum*, root and stem rot of soybean caused by *P. sojae* and downy mildew of grape vine caused by *Plasmopara viticola*. Top 10 oomycetes that infect plants were ranked based on scientific importance and economic losses and included *P. infestans, P. ramorum, P. sojae, P. capsici, P. parasitica* and *Pythium ultimum* (Kamoun et al., 2014).

Classification of oomycetes: The class Oomycetes along with brown algae and diatoms have been classified in the kingdom Stramenopiles within the superkingdom Chromalveolate (Baldauf *et al*., 2000). The Stramenopiles include both autotrophic and heterotrophic organisms. Oomycetes can be found in a wide variety of habitats, from deserts of Iran to artic regions in Antartica (Mirzaee et al., 2009; Bridge et al., 2008). An estimated sixty percent of oomycetes are plant pathogens (Thines and Kamoun, 2010). Oomycetes emerged as plant endophytes approximately around 300-350 million years

ago and phylogenetic analysis of the taxa has shown that parasitism evolved in different lineages of oomycetes at 3 different times – once in Saprolegniales and twice in Peronosporales (Thines and Kamoun, 2010).

Within the order Peronosporales and family Pythiaceae are the genus *Phytophthora* and *Pythium,* both of which are widely known as plant pathogens (Hawksworth et al., 1995). The genus *Phytophthora* is one of the most important plant pathogens and has derived its name from Greek words meaning "plant and destroyer". It contains more than 100 species. Some species have a narrow host range; *P. infestans* is pathogenic to selected members of the *Solanaceae* family and *P. sojae* is also pathogenic to soybean and a few related plants. On the other hand, some species have a broad host range; *P. cinnamomi* can infect over 900 hosts while *P. parasitica* can infect over 100 plant species.

Even though both *Phytophthora* and *Pythium* have arisen from the same ancestor, they have evolved away from each other in terms of life-style, morphology, biochemistry and host range. *Py. ultimum* is a homothallic oomycete and does not need another mating type for sexual reproduction while *P. infestans* is heterothallic and sexual reproduction occurs in presence of correct mating type (Francis and Clair, 1993; Schumann and D'Arcy, 2000). *Py. ultimum* is a necrotroph while *P. infestans* is a hemi-biotroph and while *Py. ultimum* has a broad host range, *P. infestans* has a narrow host range.

Due to similarities in filamentous growth, absorptive mode of nutrition and reproduction via spores, oomycetes were once classified as a class of fungi. However, there are many differences between true fungi and oomycetes (Latijnhouwers *et al*.,

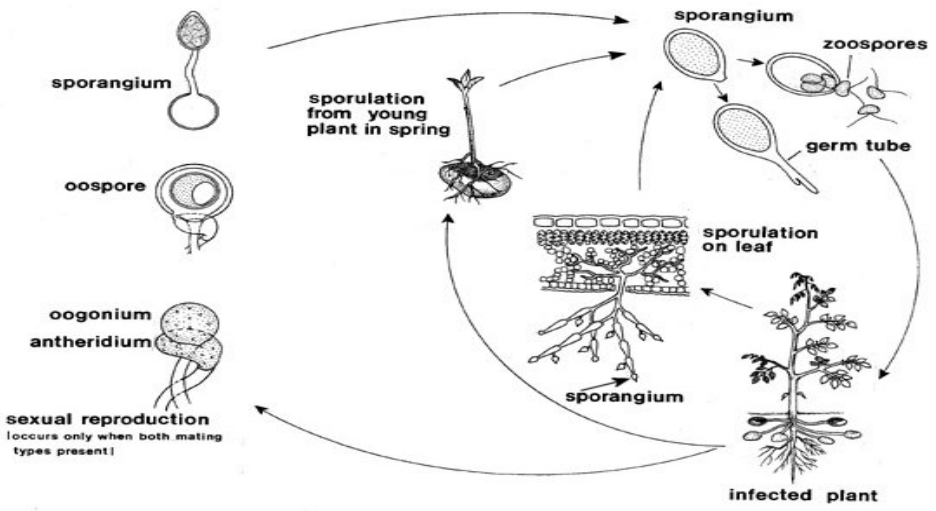
2003); the cell wall of oomycetes is made of beta-glucans and cellulose whereas chitin is the main constituent of fungal cell wall, sexual reproduction in oomycetes leads to the formation of oospores whereas fungal end product is zygospore or ascospore and mycelium is diploid in oomycetes and haploid in most fungi.

***Phytophthora* infestans:**

Sexual and asexual life cycle of P. infestans: P. infestans is a heterothallic oomycete and can undergo sexual reproduction in the presence of both mating types, A1 and A2 (Gallegly and Galindo, 1958). When both mating types are present antheridia and oogonia are formed, which may fuse to form an oospore. Oospore serve two functions – as a survival structure and also as a source of variation through sexual recombination.

The asexual life cycle is primarily responsible for the spread of the pathogen and can be broadly divided into hyphal, sporulation, sporangia cleavage (zoosporogenesis), free-living zoospores, cyst and germinating cyst stages (Fry et al., 2008). The hyphae develop into sporangiophores which bear the sporangia. These sporangia are readily dehiscent and may either germinate through a germ tube at high temperatures (20-25°C) or release wall-less and bi-flagellated zoospores at lower temperatures (10-15°C) (Melhus, 1915). Zoospores can swim for some time and then encyst. A germ tube comes out of the cyst that will then penetrate the leaf or stem tissue, often with the aid of an appressorium (Fry et al., 2008).

Figure 1: Simplified life cycle of *P. infestans.* Adapted from Schuman and D'Arcy, 2000.



Controlling P. infestans: P. infestans is a foliar pathogen, infecting the leaves, stem and fruits of potato and tomato plants but cannot colonize in roots. It is the causal agent of late blight disease in potato and is infamous for triggering the Irish potato famine in 1845 (Bourke, 1964). Even now, the estimated annual damage in potato crops caused by *P. infestans* are over 6 billion dollars (Haverkort *et al*., 2008). Common strategies to control *P. infestans* include growing resistant cultivars (Wastie, 1991; Ross, 1986; Niederhauser and Mills, 1953) or chemical spraying. The use of cultivars with R genes has proved so far to not be very durable in the field as the pathogen evolves to skip identification by R genes. Some of the pesticides used against *Phytophthora* include the phenylamide metalaxyl and a similar isomer mefenoxam, which inhibits RNA-polymerase I (Sukul and Spiteller, 2000). However, many strains of *P. infestans* are resistant to these fungicides (Lee *et al*., 1999).

Mode of nutrition: P. infestans is a hemibiotroph and requires a living host to complete its life cycle. In the early stages of its life cycle, the pathogen propagates asymptomatically by suppressing the host's programmed cell death but in later stages of the life cycle causes necrosis (Lee and Rose, 2010).

Biotrophic phase of *P. infestans* is during 2-4 dpi; starting about day 5, the necrophic phase starts (Haas *et al*., 2009). During the biotrophic phase, branching hyphae with haustoria expand from the site of infection in the intercellular spaces and then into the plant mesophyll cells. During the necrophic phase, the mycelium develops sporangiophores that emerge through stomata and produce numerous sporangia (Aylor *et*

al., 2001). The production of large numbers of spores (up to 300,000 per lesion) that can be dispersed long distances in a short time results in the wide spread of disease (Fry et al., 2008).

Aims and key questions answered in the present study: Global food security has become a major issue in terms of feeding today's 9.2 billion people (Bebber and Gurr, 2015). Apart from abiotic threats that include water crises, higher energy costs and climate change, biotic threats to food security have attracted much research. On average, these biotic threats include viruses, bacteria, fungi, oomycetes and insects, which can cause losses of up to 20% of the total harvest (Oerke 2006; Flood, 2010; Savary et al., 2012). Of the major biotic threats that are present, fungi and oomycetes are seen as the biggest threat to food security (Fisher et al., 2012). The number of diseases caused by fungal and oomycetes pathogens have increased over a period of 50 years (Bebber and Gurr, 2015) and are moving to new territories at a rate of 8 km per annum (Bebber et al., 2013). This scenario calls for more research that will help us in better detection and monitoring of pathogens, better understanding of the biology of these organisms, host-pathogen interactions and differences in life styles of organisms.

In the present study, we have used bioinformatic approaches to address some of the above issues. Some of the research in this study was directly connected to controlling *P. infestans* while some will help in understanding the transcriptional regulation and evolution in different oomycetes which in long term will give us novel mechanisms to control *P. infestans*.

Chapter I: Building a bioinformatic pipeline to identify antigenic peptides that are specific to P. infestans

Timely identification of *P. infestans* is really important because under favorable conditions the entire crop or entire storage facility can be compromised within weeks (Judelson, 2014). Currently PCR-based tests are the most accurate method for correctly identifying *P. infestans* (Cooke et al., 2007). However, PCR requires a lab setting, and thus this type of DNA-based assay may take more time to get results during which the crop might be destroyed. In the present study, we have used a slightly modified approach of "reverse vaccinology" that has been successfully used in identifying antigens in pathogenic strains of *Niserria meningitides, E. coli, Streptococcus* and other pathogens (Pizza et al., 2000; Moriel et al., 2009; Maione et al., 2005). We first identified regions within proteins that are unique to *P. infestans* based on an absence of homology to other *Phytophthora* species at protein and nucleotide level. The regions were then checked for multiple criteria as used in "reverse vaccinology" approach that includes solubility, number of hydrophilic residues, conservation in different strains and antigenicity (Pizza et al., 2000). Apart from the above-mentioned points, we also checked expression levels of transcripts in different developmental stages and confirmed that gene model is correct. All the potential candidates that came out of this analysis and the complete pipeline are described in Chapter I.

Chapter II: RNA-Seq for identifying novel transcripts, alternate splicing and improving gene annotations in plant pathogen Phytophthora infestans

Correct gene annotations were a limiting factor in identifying potential antigenic candidates in Chapter I; RNA-Seq alignments to the reference genome showed that the predicted gene model and as such the predicted protein were often incorrect. Incorrect gene models also hamper other structural and functional analyses; this includes but is not limited to ortholog identification, transcription factor binding site predictions, localization studies, etc. We obtained RNA-Seq reads from two different strains of *P. infestans,* 1306 and 88069, and aligned them to the reference genome of another strain (T30-4). Using these alignments along with multiple tools that included *PASA* (program to align spliced alignments), genome guided *Trinity*, *transdecoder* and *IGV*, we modified the gene models of 11,938 genes (Haas et al., 2003; Grabherr et al., 2011; http://transdecoder.github.io/; Thorvaldsdottir et al., 2013). We then also did a complete functional annotation and secretome identification of the new proteome. We also identified non-coding RNAs using the transcript assemblies. The process of modifying the genes, finding new genes and ncRNAs along with their annotation is explained in Chapter II.

Chapter III: Structural variation identification and annotation in different lineages of P. infestans

Analysis from chapter II where we had reads from two different strains aligned to the assembly of a third isolate showed that many genes had mismatches between different

isolates used in the analysis. This made us wonder if there are big variations between different lineages. Previous studies in different lineages of *P. infestans* have shown variations in terms of mating type, host preference and resistance to fungicides (Danies et al., 2013; Kato et al., 1997; Legard et al., 1995). This variation can be explained if there are differences at DNA levels or due to epigenetics. In this study, we have used 10 different isolates to identify large and small structural variation present and identified the genes that were affected. Since we have used a new assembly that came from 1306 strain, we have annotated the variation based on that assembly. The pipeline to first annotate the assembly and then identify and annotate structural variants present in these lineages is given in Chapter III.

Chapter IV: Expression-profiling of Phytophthora infestans and Pythium ultimum on potato tuber reveals mRNA differences that reflect lifestyle, gene gain and loss, and transcriptional rewiring

Both *P. infestans* and *Py. ultimum* can infect potato tubers; however, both have a different life style. *P. infestans* is a hemi-biotroph while *Py. ultimum* is a necrotroph. This study, performed in collaboration with other members of the laboratory, involved finding orthologs of *P. infestans* and *Py. ultimum*, functional annotation and secretome identification of both *P. infestans* and *Py. ultimum,* and GO term enrichment. This study gives insights into transcriptional differences between *P. infestans* and *Py. ultimum* in terms of different families that include pathogenicity genes, cell-wall degrading enzymes and nutrient uptake. The study also showed that *P. infestans* has a more dynamic

transcriptome, with more genes turning on and off between different developmental stages; *Py*. *ultimum* on the other hand keeps more of its genes turned on. Specific gene families found uniquely in *P. infestans* and *Py*. *ultimum*, orthog identification and common GO annotations are explained in this chapter.

Chapter V: Decay of genes encoding the oomycete flagellar proteome in the downy mildew Hyaloperonospora Arabidopsis.

This chapter is a published paper performed in collaboration with other members of the laboratory. It uses comparative genomics to find orthologs of *P. infestans* flagellar proteins in genomes some of which have retained the ability to form flagella while others who have lost this ability. Flagella play an important role in the life cycle of *P*. *infestans,* as zoospores use their two flagella to swim to the new host. This study investigated the diversity of flagellar proteins in different species and their loss in some. *Hyaloperonospora arabidopsidis* has lost the ability to form flagella and does not have a zoospore stage; however, we found that the genome contains some remnants of the genes that might once be involved in forming flagella. The study also shows expression patterns of flagellar proteins in different developmental stages and found that most flagellar proteins are expressed at higher levels in spores and zoospore stage as compared to non-sporulating mycelia.

References:

Bebber, D. P., & Gurr, S. J. (2015). Crop-destroying fungal and oomycete pathogens challenge food security. Fungal Genetics and Biology, 74, 62-64.

Cooke, D., Schena, L., & Cacciola, S. (2007). Tools to detect, identify and monitor Phytophthora species in natural ecosystems. Journal of Plant Pathology, 89, 13-28.

Danies, G., Small, I. M., Myers, K., Childers, R., & Fry, W. E. (2013). Phenotypic characterization of recent clonal lineages of *Phytophthora infestans* in the United States. Plant Disease, 97, 873-881.

Flood, J. (2010). The importance of plant health to food security. Food Security, 2, 215-231.

Gosting, L. H., Cabrian, K. A. T. H. Y., Sturge, J. C., & Goldstein, L. C. (1984). Identification of a species-specific antigen in *Legionella pneumophila* by a monoclonal antibody. Journal of Clinical Microbiology, 20, 1031-1035.

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., … Regev, A. (2011). Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nature Biotechnology, 29, 644–652.

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., ... & Salzberg, S. L. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research, 31, 5654-5666.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... & MacManes, M. D. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols, 8, 1494-1512.

Kato, M., Mizubuti, E. S., Goodwin, S. B., & Fry, W. E. (1997). Sensitivity to protectant fungicides and pathogenic fitness of clonal lineages of *Phytophthora infestans* in the United States. Phytopathology, 87, 973-978.

Lee, S. J., & Rose, J. K. (2010). Mediation of the transition from biotrophy to necrotrophy in hemibiotrophic plant pathogens by secreted effector proteins. Plant Signaling & Behavior, 5, 769-772.

Le Moigne, V., Gaillard, J. L., Herrmann, J. L., & Roux, A. L. (2017). Virulence mycobacterial determinants represent useful antigens for the serological diagnostic of ntm infection in cystic fibrosis patients. B62 non-tuberculous mycobacteria: clinical aspects and cases (pp. A3957-A3957). American Thoracic Society.

Nakayama, E., Yokoyama, A., Miyamoto, H., Igarashi, M., Kishida, N., Matsuno, K., ... & Takada, A. (2010). Enzyme-linked immunosorbent assay for detection of filovirus species-specific antibodies. Clinical and Vaccine Immunology, 17, 1723-1728.

Niederhauser, J. S. (1991). Phytophthora infestans: the Mexican connection (pp. 25-45). Cambridge University Press, Cambridge.

Oerke, E. C. (2006). Crop losses to pests. The Journal of Agricultural Science, 144, 31-43.

Ross, H. (1986). Potato breeding-problems and perspectives. Advance Plant Breeding Supplement, 13, 82–86.

Savary, S., Ficke, A., Aubertot, J. N., & Hollier, C. (2012). Crop losses due to diseases and their implications for global food production losses and food security. Food Security, 4, 519-537.

Schumann, G.L. and D'Arcy, C. J. (2000). Late blight of potato and tomato. The Plant Health Instructor. DOI: 10.1094/PHI-I-2000-0724-01

Thorvaldsdóttir, H., Robinson, J. T., & Mesirov, J. P. (2013). Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Briefings in Bioinformatics, 14, 178-192.

Wastie, R. L. (1991). Breeding for resistance. Advances in Plant Pathology, 7, 193-224.

Chapter I

Building a pipeline to identify antigenic peptides specific to *P. infestans.*

Abstract: *Phytophthora infestans* is responsible for the late blight disease of potato and tomato plants. Several other *Phytophthora* species infect potato and tomato. Disease management strategies require correct identification of the pathogen as different species call for different controlling strategies. Our study uses a pipeline to identify antigenic regions that can be used to develop an immunodetection assay for a specific Phytophthora species. Using blastp, proteins of *P. infestans* were checked for similarity against four other species of *Phytophthora*. Regions showing low similarity are putative antigenic peptides. Peptides were checked for absence of signal peptide and high antigenicity. Gene models of the candidates were confirmed using RNA-Seq data to ensure that differences were real and not due to bad gene models. Selected peptides were also checked against four other *Phytophthora* and *Solanum* species to confirm that the peptides are unique for *P. infestans*. The putative peptides were checked for conservation in ten isolates of P. infestans to make sure that all the isolates are identified. As of now we have 7 candidates that fulfilled all our above criteria and are now in the process of developing antibody. This pipeline can be used for any species-specific antibody development; however, one has to be cautious about incorrect gene prediction that might give false protein predictions.

Background: Global food security is a major issue as over 800 million people do not have access to sufficient food. Studies have projected that the demand for food will keep on rising and we will need an additional 70% by the year 2050 (Godfray et al., 2010). Approximately over 1 billion people are suffering from malnutrition and almost double the number do not have access to proper vitamin resources (Conway, 2012). One challenge to agricultural production is reduced land area for agriculture and the increase in population, but plant diseases also play an important role in reducing crop production (Nellemann, 2009; Foresight, 2011). It is estimated that crop losses due to plant pathogens can vary from 20%-40% (Savary et al., 2012); thus, the damage caused by these plant pathogens cannot be ignored. Some of the important food crops that are affected by plant pathogens are maize, rice and soybean (crop loss 12%), potatoes with crop losses up to 24%, and wheat and cotton with losses ranging from 50% to 80% (Savary et al., 2012). Among the most-deadly pathogens that have emerged as a big threat to food security are fungi and oomycetes (Fisher et al., 2012).

Some of the important food crops that are affected by plant pathogens are potato and tomato, which are susceptible to diseases caused by *Phytophthora infestans* causing annual losses of ~7-15 billion dollars. *P. sojae* is responsible for the stem and root rot of soybean, resulting in losses of ~6.29 billion dollars in the year 1998 (Wrather et al., 2000). Apart from food crops, many *Phytophthora* species are a big threat to forests, natural ecosystems and ornamental crops across the world; *P. ramorum* has emerged as a deadly pathogen to oak trees causing large-scale decline in tanoak populations in California forests. In the year 2003 plants infected with *P. ramorum* but asymptomatic

were received by large number of nurseries from suppliers on the west coast and threatened an industry worth $13 billion per annum (Jones et al., 2005). The identification and control of *Phytophthora* species is not only important to save the crops but is equally important to save the natural ecosystems.

Phytophthora diagnosis: Traditional methods to identify *Phytophthora* mostly involve visual examination of the morphological characteristics. These include but are not limited to sporangia shape, pedicel shape, papilla shape and hyphal swellings. Some other methods that are used to identify *Phytophthora* in natural environments include isolation, baiting, PCR amplification and immunodetection (Lacey, 1965; Tumwine et al., 2000; Tooley et al., 2000).

Isolation works by isolating *Phytophthora* from infected plant tissue and growing the species on a selective media with antibacterial and antifungal compounds that will promote the growth of *Phytophthora*. This can be then identified using morphological characters (Tsao, 1983). Isolation can give false negative results due to the failure of pathogen growth from the tissue segment (Kox et al., 2007). The reason for failure of growth on selective agar can be due to antagonism from other organisms, inhibition by plant phenols and antibiotics used in the media (Huberli et al., 2000). Hymexazol antibiotic is usually used to control *Pythium* but certain strains of *Phytophthora* are also sensitive as such may die on exposure (Ali-Shtayeh et al., 2003).

Baiting involves floating the susceptible plant tissue in soil water with a high water/soil ratio, and is commonly used to recover *Phytopthora* species from soil.

Zoospores, which will infect the plant tissue, are then collected and placed on selective agar. Species can then be identified on the basis of morphological identification or DNA sequence analysis (Eden et al., 2000). Usually baiting is used to identify the species present in the area but this is a slow and time-consuming technique. Apart from time, zoospores may always not infect the bait, which can lead to false negatives. Different *Phytophthora* species have varying attraction to certain baits and may not infect those baits at all (Tsao, 1983). Studies have shown that even though considerable numbers of zoospores may be present in the water, the results of baiting can be negative (Wilson et al., 2000).

PCR amplification offers many advantages compared to isolation or baiting; it is highly specific, not affected by environmental conditions or stage of development and are easy to set up. The most common region that has been historically used to differentiate species is the ITS region (Ristaino et al., 1998; Hayden et al., 2006; Hughes et al., 2006). Apart from the ITS region, genes encoding 60S ribosomal subunits, β-tubulin, enolase and mitochondrial *cox* regions have also been used (Blair et al., 2008; Kroon et al., 2012; Martin and Tooley, 2003). However, selection of primers, which are the key ingredients of PCR, can affect the sensitivity to detect a species; a single base change in a primer needed to reduce cross reactivity led to a 50-fold change in the sensitivity of PCR to detect *P. ramorum* (Hughes et al., 2006b). Correct primer selection is absolutely important for these studies; PCR amplification followed by restriction digest with mitochondrial *cox1* and *cox2* genes could not be used for infected plant material because

the primers were binding to plant genes and genes from the closely related genus *Pythium* (Martin et al., 2004).

ELISA (enzyme-linked immunosorbent assay) uses antibodies to identify the kind of organism that is affecting the crop. Antibodies are developed by injecting animals (often a mouse or rabbit) with culture extracts that include proteins and carbohydrates produced by the pathogen. Some of these proteins are antigenic and cause an immunogenic reaction, resulting in the production of antibodies, which then bind to antigens and try to clear them out of the circulatory system. In the case of monoclonal antibodies, the cells that produced the immune response are harvested, fused to immortalized cells, and the resulting hybridomas grown in cultures. The antibodies identified may or may not be specific to species because many antigenic peptides are conserved among different species (Rittenber et al., 1997). Many hybridomas may need to be screened to identify those (if any exist) that are species-specific. Nevertheless, antibody-based assays offered by different companies are used for various detection purposes ranging from pathogen testing in plants (Agdia, Agrisera) to testing for cancer and other infectious diseases (biogenex, Bioworld Antibodies). ELISA kits are available to detect *Phytophthora* at the genus level but not at the species level; also, currently available antibodies also show cross reactivity with some species of *Pythium* (Ali-Shatayeh et al., 1991; MacDonald et al., 1990). The commercially available ELISA tests use many different formats that include multi-well plates for high-throughput assays, dipsticks, and lateral flow devices. In order to identify species using ELISA, DNA on the immunostrip can be used for an amplification reaction such as loop mediated isothermal

amplification (LAMP). The LAMP technique works on the same principle as PCR but does not need a thermal cycling equipment and can be done on-site. The above technique has been used to successfully identify *P. palmivora* (Torres et al., 2010)*, P. ramorum* and *P. kernoviae* (Tomlinson et al., 2010). Currently most of the ELISA assays are further confirmed using PCR-based techniques.

Bioinformatic approaches have been used to identify antigenic peptides to generate immune responses in animal models and then used as vaccines (Iurescia et al., 2012; Purcell et al., 2003). Advancements in sequencing technologies have led to the availability of a lot of DNA sequences for a lot of pathogens; these include new bacterial, viral and fungal genomes (Pizza et al., 2000; Tettelin et al., 2000; Giuliani et al., 2006; Pauza et al., 2000). This had resulted in development of tools that can predict antigenicity in proteins (Hopp and Woods, 1989; Kolaskar and Tongaonkar, 1990). Prediction of antigens is also commonly known as "*reverse vaccinology*", which has used both alignment based methods and alignment free methods to predict antigens (Pizza et al., 2000).

Alignment-based methods depend on the alignment of the selected candidates to identify all the proteins that might be virulent based on homology to known virulent proteins, secreted or have putative surface-exposure using prediction software. This approach has been successfully used in identifying potential vaccine candidates in sequenced genomes and was first used in developing vaccines for *Neisseria meningitides B* and has since been a milestone in developing antigens for newly sequenced genomes (Pizza et al., 2000). This strategy includes finding ORFs in newly sequenced species or

the proteome that is then checked for features typical of surface-associated proteins using programs like PSORT (Nakai and Horton, 1999) or SignalP (Bendtsen et al., 2004). The sequences of putative antigens sometimes are also compared to previously known antigens for sequence similarity (Schlessinger et al., 2006). These antigens might then be checked for variation in different strains so that they can identify different strains of the same pathogen (Pizza et al., 2000; Muzzi et al., 2013).

Alignment-free methods to find new antigens depend on the principal chemical amino acid properties and not the homology of the peptide to known peptides. This approach is dependent on training the datasets and then predicting the new antigens based on covariance between the two sets (Doytchinova and Flower, 2007).

Antigen development has been a long-studied topic in vaccine development in animals and designing diagnostic assays for detecting plant pathogens. Various epitope prediction software's have been developed dating back to 1981 when the first epitope prediction method was developed by Hopp and Woods. Most of these methods are for prediction of continuous (linear) epitopes and utilize sequence properties such as surface accessibility, flexibility, hydrophilicity, charge, proximity of the peptide towards N or C-termini of the protein, and secondary structure. Sequence length also plays an important role in identification of the native protein. Large sequence lengths (up to 40 amino acids) are usually better as they can have multiple epitopes and hence can identify the native protein better but are expensive to synthesize. Most researchers prefer an epitope length of 10-20 amino acids that have a more reasonable synthesis cost and still have a decent chance to identify the native protein. Epitopes should have a higher number of

hydrophilic residues as compared to hydrophobic residues that will make it surface accessible and flexible, hence amino and carboxyl terminus of the proteins are preferred as they generally have a higher number of hydrophilic residues. The epitope can be linear stretches of amino acids or discontinuous where the protein folding causes parts of protein to come together.

Even though the peptide-based strategy has greatly enhanced the discovery of new epitopes based on the entire genome of a pathogen, it is still met with challenges that include the synthesis of peptides, coupling of peptide with a carrier protein to generate immune response, and most importantlyl whether the antibody will bind not only the peptide but also the native protein. Synthesis of peptides can be a constraint as peptides rich in hydrophobic residues such as leucine, methionine and valine are usually insoluble in aqueous solutions; on the other hand, peptides that are rich in cysteine or tryptophan can be targets of reactions like oxidation and hence will identify the peptide but not the whole protein (Milton et al., 1990; Fauchere and Pliska, 1983). Small peptides need to be coupled to a carrier proteins to generate an immune response; also, cross-linking peptides to the carrier proteins needed to stimulate the immune response may cause precipitation of the peptide, which will reduce the effectiveness of the candidate antigen (Stawikowski and Fields, 2002).

Controlling *P. infestans:* *P. infestans* can spread by means of infested soil, water, infected plants and aerial dispersion of its spores (Goodwin, 1997). Measures for controlling late blight involve the regular checking of plant tissue for symptoms of a *P. infestans*

infection. These include checking the potato tubers before storing or planting even when they do not show any visible symptoms, and evaluating lesions on leaves and tomato fruits and seeds from tomato fruit. Studies with tomato fruits have shown that *P. infestans* is commonly found in tomato fruit seeds which harbor pathogen propagules; in one study, up to 93% seeds contained viable pathogen propagules when planted on selective medium (Vartanian and Endo, 1985a). Infected plants that are currently showing no symptoms can easily cause large-scale destruction of healthy plants once sporulation occurs. A single potato tuber infected with *P. infestans* can compromise an entire storage facility (Judelson, 2014).

P. infestans can spread really fast and cause defoliation of the plants within 3 weeks after the initial symptoms show up. Also, the pathogen can infect tubers at any stage, even after harvesting of potatoes. One complication in disease management is knowing whether a suspicious-looking lesion contains *P. infestans* or another pathogen, including other *Phytophthora* spp. which spread more slowly. Since the spread of *P. infestans* can be so fast, it is very important to know if the crop is infested with *P. infestans* or some other *Phytophthora* species; usually the farmer would apply fungicides at more frequent intervals if *P. infestans* was present. PCR analysis can give the details of the species but it requires a lab setting, which may take more time, during which the crop might be destroyed.

Genome of *P. infestans* and similarity to other genomes: We wanted to identify proteins/peptides that are uniquely present in *P. infestans* and absent in other species (*P.*

sojae, P. ramorum, P. parasitica and *P. capsici)*. Previous studies have shown that *P. infestans* genome is the largest (240Mb) of all the above-mentioned genomes (*P. sojae* – 95Mb, *P. ramorum* – 65Mb, *P. parasitica* – 95.5Mb and *P. capsici* – 64Mb) and that *P. infestans, P. sojae* and *P. ramorum* share 8,492 ortholog clusters that include 9,583 *P. infestans* genes (Haas et al., 2009). This analysis was based on the annotations from Broad Institute, which predicted 17,797 protein coding genes for *P. infestans.*

Based on differences present between the genome of *P. infestans* and other *Phytophthora* species, it was logical to think that these differences can be exploited to identify antigenic peptides that will specifically bind to *P. infestans* proteins only and not those of other species. These peptides can then be used in a field or lab-based diagnostic antibody assay, without requiring any amplification step or PCR confirmation. Since immunoassays can be performed in the field and since potato growers are used to performing similar assays for viral pathogens, they probably will be adopted by growers more rapidly.

The current strategy to identify species-specific antigen utilizes the Blast technique (Altschul et al., 1990) to identify unique regions or whole proteins that are specific to *P. infestans*.

We first predicted which proteins showed variation at the 5' end, 3' end and regions in the middle of proteins and then using the "reverse vaccinology" approach identified which peptides can be used to generate antibodies.

Consequently, the goal of this study was to identify peptides/proteins that can be used to correctly diagnose *P. infestans* in the field by building a bioinformatics pipeline

and then confirming the output of the pipeline by generating antibodies to the selected candidates and doing ELISA and western blots (Figure 1). We found 11 N-terminus and 13 C-terminus proteins that can be used for generating *P. infestans* specific antibodies and generated monoclonal and polyclonal antibodies against peptides and whole protein for 2 proteins – PITG_02706 and PITG_04766; polyclonal antibodies for peptides and native protein for 3 proteins – PITG_16620, PITG_09393, PITG_16922 (Table 1).

Figure 1: Complete pipeline to identify antigenic peptides specific to P. infestans. Regions specific to *P. infestans* were identified based on similarity searches in other *Phytophthora* species. These were categorized as N-terminus, middle region and C-terminus candidates. The expression levels of the selected candidates were checked in early and late leaf and tuber infections and media cultures. The candidates with higher than average expression levels were checked for correct gene model followed by multiple alignments of orthologs and conservation in different isolates. The selected candidates were then checked for antigenicity. Final candidates were checked using western blots.

Figure 1

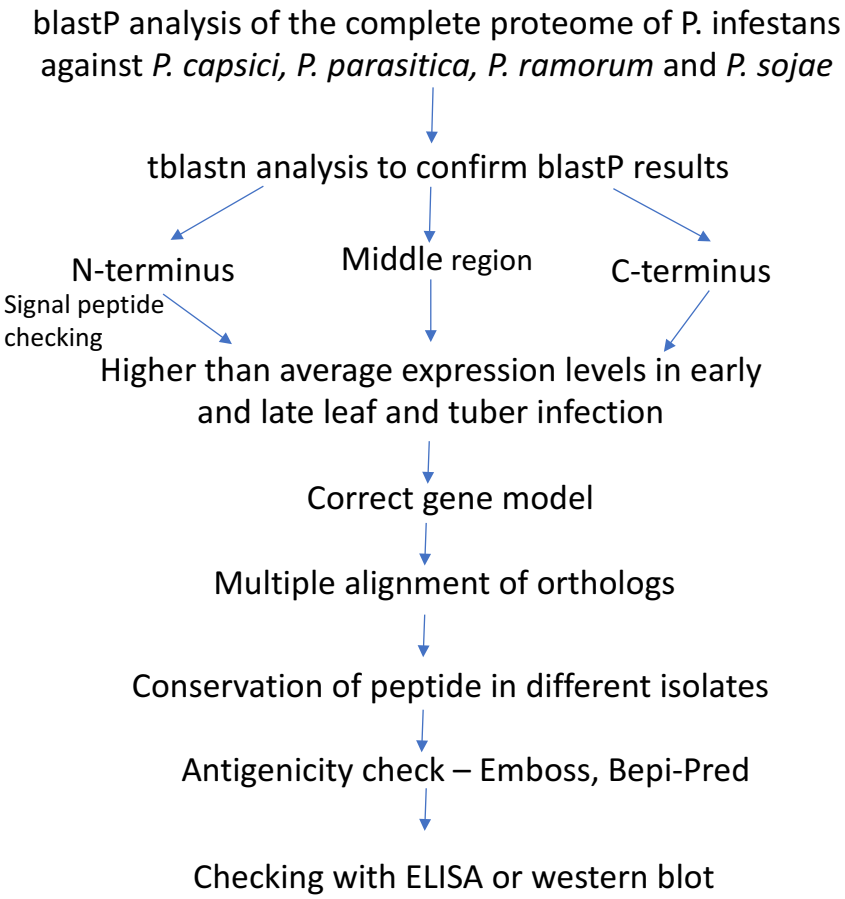

blastP analysis of the complete proteome of P. infestans against *P. capsici, P. parasitica, P. ramorum* and *P. sojae*
tblastn analysis to confirm blastP results
N-terminus
Signal peptide checking
Middle region
C-terminus
Higher than average expression levels in early and late leaf and tuber infection
Correct gene model
Multiple alignment of orthologs
Conservation of peptide in different isolates
Antigenicity check – Emboss, Bepi-Pred
Checking with ELISA or western blot

Table 1: Progress of developing species-specific antibodies for *P. infestans*.

Protein	Peptide monoclonal	Peptide polyclonal	whole protein
PITG_02706	Failed	2 failed	1 in progress
PITG_16620(new)	-	2 in progress	1 in progress
PITG_09393 (old)	-	1 in progress	1 in progress
PITG_16922(new)	-	1 in progress	soon in rabbit
PITG_04766	Failed	2 failed	-

Results:

This study looked at the proteins that share some or no homology to the predicted proteins in other species based on alignment results from Blast hits and then used this information to identify peptides that can be used to develop antibodies to identify *P. infestans* in plants.

We designed a pipeline to predict peptides that are specific to *P. infestans*. The first step was to find peptides that did not show any homology to the proteomes of *P. capsici, P. parasitica, P. sojae* and *P. ramorum*. These regions and proteins are then checked for a variety of criteria that include: (a) checking if the variation is present outside the signal peptide cleavage site, (b) selecting genes with higher than average FPKM values in all developmental stages or highly expressed in early infection stages, (c) correct gene model based on RNA-Seq data, (d) variation confirmation with multiple alignment of orthologs, (e) conservation in different isolates of *P. infestans* and (f) antigenicity and secondary structure predictions using the IEDB analysis website (http://tools.immuneepitope.org/bcell/).

The above criteria were selected due to following reasons. The species-specific region must be outside the signal peptide because if the variation is within the cleavage region then it will not be present in the mature protein and thus not detected by an antibody. It is also preferable that the gene will be expressed at higher than average levels in all developmental stages so that it can be detected at any stage of the disease cycle; higher expression in early infection stages will help identify *P. infestans* even when the plant is asymptomatic. It is also important to ensure that a correct gene model is being

used, to make sure that the peptide is actually a part of the expressed protein and not a false positive that predicts an incorrect protein sequence due to incorrect exon boundaries.

Previous studies have reported variation within single clonal lineages that are mostly attributed to mitotic recombination or mutations (Goodwin et al., 1994; Goodwin et al., 1995). Our lab had sequenced 10 different isolates mentioned above. These isolates are diverse in terms of mating type, resistance to mefenoxam and host susceptibility (Table 2). Apart from the above isolates, we also checked the final candidates against two different species *P. mirabilis* and *P. ipomoeae* that are closely related to *P. infestans* (all of them belong to clade 1c) but are unable to infect potato and tomato (Blair et al., 2008). *P. mirabilis* infects the leafs of *Mirabilis jalapa* and *P. ipomoea* infects *Ipomoea longipedunculata* and are absent in potato fields (Galindo and Hohl, 1985; Flier et al., 2002). The idea behind checking the conservation in these species was a way to confirm the conservation of the peptide; if the peptide is conserved in a different species of *Phytophthora*, the chances that it will be conserved in different isolates of *P. infestans* are higher. We wanted to make sure that the region is highly conserved in different isolates so that we can detect all isolates of *P. infestans* and not just some isolates.

Table 2: Phenotypic characteristics of the strains used in the study.

Strain	Metalaxyl sensitivity	Host preference	Mating type	Source
1306	sensitive	tomato	A1	USA
550	sensitive	potato	A2	Mexico
618	insensitive	not tested	A2	Mexico
1114	insensitive	not tested	A1	Netherlands
6629	sensitive	not tested	A1	Mexico
US-8	insensitive	potato	A2	USA
US-11	insensitive	potato/tomato	A1	USA
US-22	sensitive	tomato	A2	USA
US-23	sensitive	tomato	A1	USA
US-24	sensitive	potato	A1	USA

This pipeline was first implemented by chopping the full protein into 20 amino acid windows which were then checked for similarity to proteins in other species; however, when the putative candidates were checked for variation against the orthologs using multiple alignment of proteins, a high degree of conservation among orthologs was detected. We realized that we should either lower the e-value of blast hits of selected regions to remove false positives or try the whole protein blast. We then checked for regions in the protein based on a ten amino acid window length that showed <60% identity. We used the whole protein blast approach to identify peptides that showed variation. This pipeline was run twice, initially on proteins predicted by Broad institute (Haas et al., 2009) and then after curating the gene models using RNA-Seq data (Chapter II).

Initially the pipeline was run on the gene models from Broad Institute. 17,797 proteins were checked against the proteome of four *Phytophthora* species - *P. sojae, P. parasitica. P. capsici* and *P. ramorum.* We focused mostly on amino and carboxyl termini of proteins as they are more likely to be accessible and flexible, and thus antigenic. The initial parsing to check for peptides that showed variation at the amino and carboxyl terminus gave 1,011 and 1,158 candidates, respectively. The N-terminus candidates were then checked for the presence of signal peptide, which resulted in the removal of 210 candidates in which the unique region was in the signal peptide. Expression data from different developmental stages in leaves, tubers and media (leaves and tubers infected with *P. infestans* and RNA collected at 3 dpi and 6 dpi, *P. infestans* cultures in rye media and the defined media of Xu) for 801 N-terminus and 1,158 C-

terminus candidates was next examined; 27 N-terminus and 19 C-terminus candidates showed higher than average expression levels in all stages. Next, the candidates fulfilling the above criteria were confirmed for correct gene models, multiple alignments of whole protein with best hits from 4 other species, conservation in different isolates of *P. infestans,* and antigenicity. Only 5 of 19 C-terminus candidates and 7 of 27 N-terminus candidates had dependable gene models and were further analyzed with multiple alignments (MLA) of protein with best blast hits. All of the above candidates fulfilled our pipeline requirements. Two proteins, PITG_02706 and PITG_04766, both of them are crinklers and might be involved in inducing necrosis in plants (Van Damme et al., 2012; Stam et al., 2013), had their expression confirmed using the proteomic data from our lab. These were therefore selected for generating monoclonal and polyclonal antibodies against peptides and whole protein.

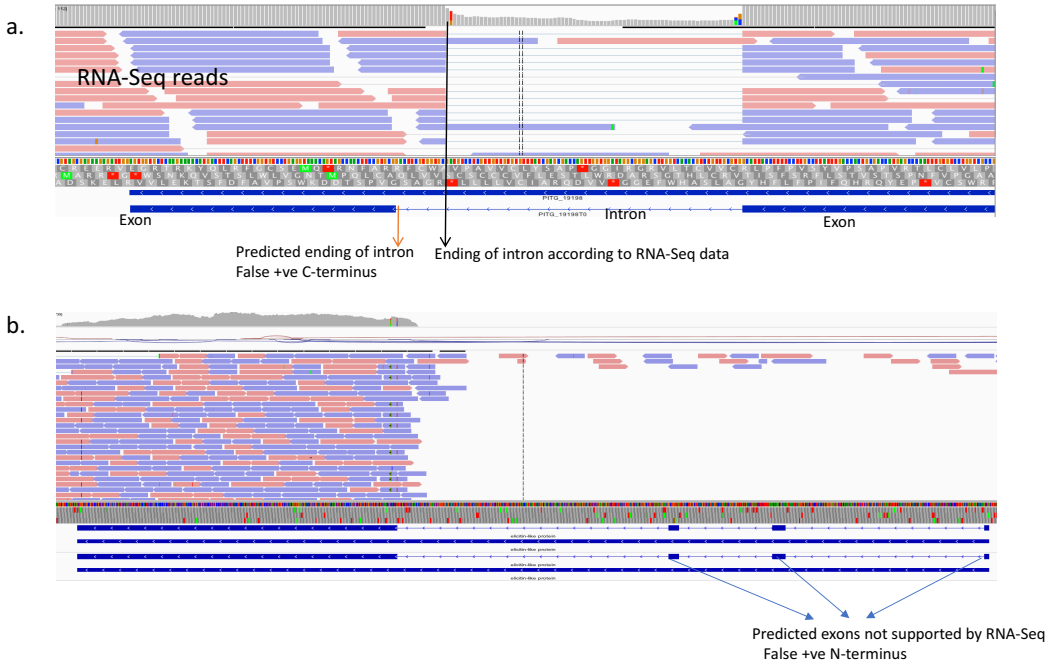
As of now the monoclonal and polyclonal antibodies against peptides of PITG_02706 and PITG_04766 have failed. Peptide monoclonal antibody to PITG_02706 was able to identify the peptide but failed to identify the native protein which is a very common problem in using peptide antibodies because they cannot sufficiently mimic the native protein. We are still in process of developing whole protein polyclonal antibody to PITG_02706.

The biggest challenge or drawback of the pipeline was bad gene models that increased the number of false positives in the dataset. Figure 2 shows some of the scenarios that resulted in candidates that were false positives due to bad gene models.

Figure 2a shows IGV snapshot of a false positive of C-terminus protein where the predicted intron is incorrectly placed leading to a change in frame of translation and hence the predicted protein sequence. When the correct ending of the intron was used to translate the protein, the protein was found to be conserved in different *Phytophthora* species. In Figure 2b there are no RNA-Seq reads to confirm the presence of presence of the predicted exons and introns at 5' end, which showed variation in comparison to other species. The RNA-Seq reads show that the gene has coverage only starting from exon 4, which then shows high conservation among species.

Figure 2: Examples of gene models yielding false positive N-terminus and C-terminus antigen candidates. **a.** Incorrect placement of intron leading to false C-terminus peptide that does not exist (PITG_19198).

b. Incorrect predictions of exons at N-terminus (PITG_12599).



To remove the problem of incorrect gene models we have used RNA-Seq data along with *trinity* and *PASA* pipelines to modify the structures of 11,938 genes (Chapter II). After modification, the complete proteome of *P. infestans* with 22,335 transcripts that are currently present in the reference genome (Chapter II) was checked against the proteome of four *Phytopthora* species - *P. sojae, P. parasitica. P. capsici* and *P. ramorum.* Parsing of the BlastP results gave us 1762 proteins showing variation compared to other species at the start of protein, 1942 proteins showing variation at the end of protein and 516 proteins showing variation in the middle of protein. These were further analyzed using tblastn to circumvent any discrepancies in gene models in other species. Considering that we changed the gene structures of more than half of the originally annotated *P. infestans* genes using RNA-Seq, we could not reliably trust the predicted gene models in other species and hence we did tblastn to find as to how many proteins showed variation due to incorrect gene models in other species. Tblastn analysis removed almost 60% of the proteins that showed variation in the specific regions from the analysis giving us a more reliable set of 687 N-terminus, 773 C-terminus and 333 middle regions candidates that removed much variation due to incorrect gene predictions in other species. This however, will still not correct the assembly problems associated with other species genomes.

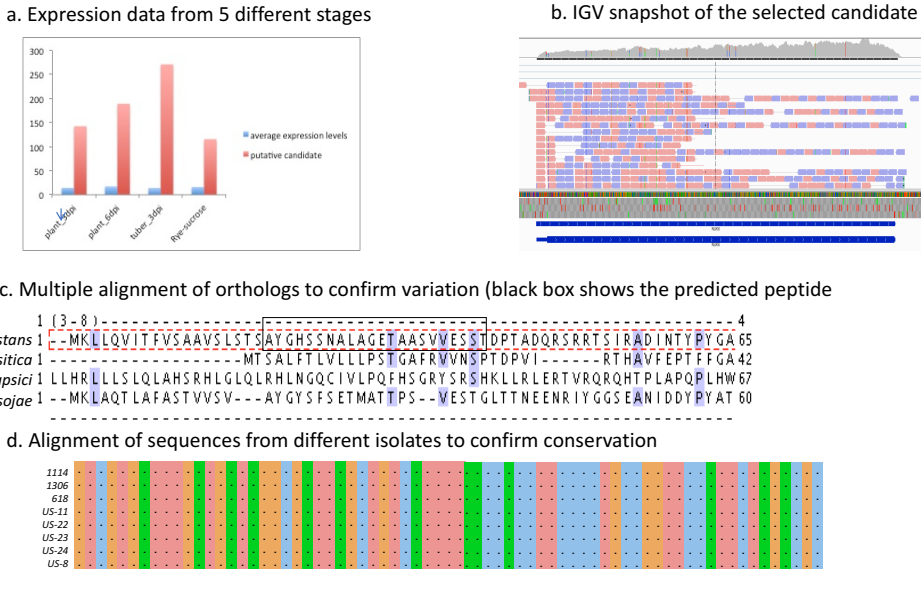
687 proteins that showed variation at N-terminus were then checked for presence of signal peptide to remove candidates which will be absent in mature protein. 157 proteins that showed variation only in the signal peptide region were removed from this analysis.

The resulting 530 N-terminus, 773 C-terminus and 333 middle regions proteins were then checked for expression levels in five different conditions: early leaf infection, late leaf infection, early tuber and late tuber infection, and media cultures. This resulted in 22 N-terminus, 21 C-terminus and 10 middle region proteins that had higher than average expression levels for each life stage. The expression check was done to confirm that an antibody against such proteins could detect *P. infestans* at all stages of development and in all kinds of tissues – leaf or tubers. The proteins showing higher than average expression levels in all developmental stages were then checked for correct gene model, multiple alignment of orthologs and conservation of peptide in different lineages. Eleven out of 22 potential N-terminus candidates were discarded as the amino terminus of the protein looked suspicious (Figure 3); these proteins have two possible methionines, and the second methionine looked more likely to a part of protein. Multiple alignment of 4 proteins showed conservation and 3 showed variation in the isolates of P. infestans. Four amino terminus candidates (PITG_00422, PITG_02977, PITG_19380 and PITG_06708) fulfilled all our criteria were selected for future antibody development. Out of the 21 carboxyl terminal candidates, 4 had a suspicious gene model, 5 showed higher conservation than our criteria for variation between species and 4 had variation in different isolates. Eight C-terminus candidates fulfilled all the above criteria. None of the middle region candidates were confirmed due to small region size that might not be exposed to bind the antibody. All of the above candidates also showed antigenicity in IEDB analysis.

As of now we are in the process of generating polyclonal antibodies against peptides and whole proteins for PITG_05010, PITG_09387, PITG_16620, and PITG_16922 that encode for a hypothetical protein (predicted ORF with no functional annotation) and conserved hypothetical protein (hypothetical protein which is conserved in different species) respectively. PITG_02706 protein has proved to be difficult to express, but efforts are continuing.

A representative example of a good candidate in terms of expression data, correct gene model, multiple alignment of orthologs and conservation in isolates is shown in Figure 3. Figure 3a represents the FPKM values of the putative candidates and average expression levels of all the genes in five developmental stages; 3b shows the alignment of reads to the gene model in genome browser IGV to confirm that gene structure is supported by RNA-Seq; 3c shows the multiple alignments of orthologs to confirm uniqueness of peptide in *P. infestans* (with black box showing the predicted antigenic peptide); 3d shows the conservation of the selected peptide in 8 isolates of *P. infestans*.

Figure 3: Representation of one of the selected candidate (PITG_02706) in terms of **a.** Expression levels **b.** Correct gene model **c.** Multiple alignment of orthologs **d.** Conservation in different *Phytophthora* isolates



Discussion:

Species specific antibodies have been used in the past for both pathogen identification and vaccine generation. Many diagnostic tests have been generated to identify the pathogen present in animal studies (Rusell et al., 1990; Gottstein et al., 1986; Del Brutto et al., 2001). Monoclonal antibody to *Legionella pneumophila* (gram-negative bacterium that causes Legionnaires disease in human) was able to identify the all the serovars of the pathogen and showed no cross-reactivity to non-pneumophilic forms of pathogen (Gosting et al., 1984). Species specific antibody based on a virulence factor was able to differentiate pathogenic *Mycobacterium abscessus* from non-pathogenic *Mycobacterium avium* (Moigne et al., 2017). Antibody to a glycoprotein present in Ebola virus was able to identify a specific species (in this case Zaire Ebola virus) but showed some cross-reactivity to other viruses (Nakayama et al., 2010).

Monoclonal antibodies have been widely used to identify plant pathogens. Monoclonal antibodies for *P. cinnamomi* unambiguously identified the six isolates of *P. cinnamomi*; however, these antibodies worked only for zoospores and cysts (Hardham et al., 1986). In another study on *P. ramorum*, hyper-variable antigenic domains were targeted to generate *P. ramorum* specific polyclonal antibodies that were reactive against zoospores and sporangia (Frankel and Harell, 2017).

This study utilizes a novel approach to identify antigens based on sequence uniqueness and the above pipeline can be extended to any pathogen with reliable gene models.

One of the biggest drawbacks of the above pipeline is dependence on correct gene models of all the species used in the study. Use of tblastn was a good choice to remove incorrect gene models in other species; more than half of the putative candidates were removed from our analysis based on the results of tblastn which looked at translated nucleotide sequences and matched them to the putative candidates. There are 8,492 core orthologs shared between *P. infestans, P. sojae* and *P. ramorum* (Haas et al., 2009). However, based on our tblastn analysis, it is evident that we are missing many more orthologs due to incorrect gene predictions. Correction of gene models will help identify many more proteins that are shared between the *Phytopthora* species giving us a more reliable way to do phylogenetic and evolutionary analysis on these genomes.

This pipeline is an easier way to do a complete proteome blast, analyze the results of both BlastP and Tblastn, checking antigenicity with minimal user input, generate multiple alignments both at protein level and nucleotide levels which give the results in pdf format that can then be loaded in any multiple alignment viewer and manipulated. The only part where user discretion is required is confirming the gene model before finalizing the candidate that will then be used for antibody generation.

Methods:

Identification of peptides and proteins specific to *P. infestans*: Genomic and proteomic data for *P. capsici, P. ramorum* and *P. sojae* genomes were downloaded from JGI website (http://genome.jgi.doe.gov/), while *P. parasitica* genome data were downloaded from Ensembl (http://ensemblgenomes.org/). Complete proteome of *P. infestans* was checked for sequence similarity against four other genomes - *P. capsici* (ver 11.0), *P. parasitica* (PP INRA-310 v2), *P. ramorum* (ver 1.1) and *P. sojae* (ver 3.0) using blastp. N-terminal, C-terminal and middle regions of proteins that showed little or no similarity (<60% similarity) for at least a length of 10 amino acids to any of the four genomes were extracted using a perl script. Proteins that showed no hits with >60% identity throughout their length or no hits at all (orphan proteins) were also extracted using custom perl scripts.

The resulting candidates were then checked for similarity between *P. infestans* proteins and genomic sequences of the above-mentioned species using tblastn. The results of tblastn were then checked to see if the blastp and tblastn results overlap. In cases where the results of blastp and tblastn did not overlap, the intersection between both blastp and tblastn was taken as the region of interest. Orphan proteins with tblastn hits were screened in a similar fashion for peptides with a minimum length of ten amino acids.

Checking for signal peptides: All the proteins that showed variation at the N-terminus were then checked for signal sequence using signalP 4.0 with a D-cutoff of 0.450. The

cleavage site predicted by signalP was then compared to the region coordinates using perl script and the region outside the cleavage site was used for further analysis.

Expression level checking: RNA-Seq reads obtained from tomato leaves and potato tubers infected with *P. infestans* at 3dpi and 6dp and *P. infestans* cultures in rye-sucrose were aligned to the reference genome of T30-4 resulting in five bam files – tomato leaves at 3dpi and 6dpi, potato tubers at 3 dpi and 6 dpi and media. Cufflinks was run on the bam files generated above and a perl script was written to incorporate the expression data from RNA-Seq analysis.

Checking for correct gene model, multiple alignment of orthologs and conservation in strains: The gene models were checked using visualization of bam files in IGV 2.3.9.

After the gene models were confirmed, the variation between different species in the antigenic region was confirmed using multiple alignments. The selected candidates top 10 blast hits in other species along with their sequences were extracted and supplied to MUSCLE (Edgar, 2004) to generate the multiple alignments of the best orthologs. Multiple aligments viewing and manipulation was done using JalView (Clamp et al., 2004).

Genomic reads from 10 different strains (1114, 1306, 6629, 618, 550, US-8, US-11, US-22, US-23 and US-24) were aligned to the reference genome of T30-4 using bowtie2 (Langmead and Salzberg, 2012). Samtools mpileup (Li et al., 2009) and bcftools were then used to extract the sequences from these strains to confirm the conservation.

Checking for antigenicity: A perl script was written to check all the candidate proteins for antigenicity using EMBOSS (Kolaskar and Tongaonkar, 1994). The antigenic coordinates predicted by EMBOSS were compared with the coordinates of the protein (coordinates of region showing variation) and only those candidates where the antigenicity coordinates overlap with protein coordinates overlap were kept for further analysis.

Also after the final selection of candidates the protein antigenic regions were confirmed using IEDB analysis resource (http://tools.immuneepitope.org/bcell/).

References:

Ali-Shatayeh, M., MacDonald, J. D., & Kabashima, J. (1991). A method for using commercial ELISA tests to detect zoospores of *Phytophthora* and *Pythium* species in irrigation water. Plant Disease, 75, 305-311.

Ali-Shtayeh, M., Salah, A. A., & Jamous, R. M. (2003). Ecology of hymexazol-insensitive *Pythium* species in field soils. Mycopathologia, 156, 333-342.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215, 403-410.

Bendtsen, J. D., Nielsen, H., von Heijne, G., & Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. Journal of Molecular Biology, 340, 783-795.

Blair, J. E., Coffey, M. D., Park, S. Y., Geiser, D. M., & Kang, S. (2008). A multi-locus phylogeny for *Phytophthora* utilizing markers derived from complete genome sequences. Fungal Genetics Biology, 45, 266-277.

Clamp, M., Cuff, J., Searle, S. M., & Barton, G. J. (2004). The jalview java alignment editor. Bioinformatics, 20, 426-427.

Conway, G. (2012). One billion hungry: can we feed the world? Cornell University Press.
Van Damme, M., Bozkurt, T. O., Cakir, C., Schornack, S., Sklenar, J., Jones, A. M., & Kamoun, S. (2012). The Irish potato famine pathogen *Phytophthora infestans* translocates the CRN8 kinase into host plant cells. PLoS Pathogen, 8, e1002875.

Del Brutto, O. H., Rajshekhar, V., White Jr, A. C., Tsang, V. C. W., Nash, T. E., Takayanagui, O. M., ... & Botero, D. (2001). Proposed diagnostic criteria for neurocysticercosis. Neurology, 57, 177-183.

Doytchinova, I. A., & Flower, D. R. (2007). Identifying candidate subunit vaccines using an alignment-independent method based on principal amino acid properties. Vaccine, 25, 856-866.

Eden, M. A., Hill, R. A., & Galpoththage, M. (2000). An efficient baiting assay for quantification of *Phytophthora cinnamomi* in soil. Plant Pathology, 49, 515-522.

Edgar, R. C. (2004). MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research, 32, 1792–1797.

Fauchere, J. L., & Pliska, V. (1983). Hydrophobic parameters pi of amino-acid side chains from the partitioning of N-acetyl-amino-acid amides. European Journal of Medicinal Chemistry, 18, 369-375.

Fisher, M. C., Henk, D. A., Briggs, C. J., Brownstein, J. S., Madoff, L. C., McCraw, S. L., & Gurr, S. J. (2012). Emerging fungal threats to animal, plant and ecosystem health. Nature, 484, 186-194.

Flier, W. G., Grünwald, N. J., Kroon, L. P., Van Den Bosch, T. B., Garay-Serrano, E., Lozoya-Saldaña, H., ... & Turkensteen, L. J. (2002). *Phytophthora ipomoeae* sp. nov., a new homothallic species causing leaf blight on *Ipomoea longipedunculata* in the Toluca Valley of central Mexico. Mycological Research, 106, 848-856.

Foresight Commission. (2011). The future of food and farming: Final project report. Office for Science, Government.

Galindo, J., & Hohl, H. R. (1985). *Phytophthora mirabilis*, a new species of *Phytophthora*. Sydowia, 38, 87-96.

Giuliani, M. M., Adu-Bobie, J., Comanducci, M., Aricò, B., Savino, S., Santini, L., ... & Cartocci, E. (2006). A universal vaccine for serogroup *B meningococcus*. Proceedings of the National Academy of Sciences of the United States of America, 103, 10834-10839.

Godfray, H. C. J., Beddington, J. R., Crute, I. R., Haddad, L., Lawrence, D., Muir, J. F., ... & Toulmin, C. (2010). Food security: the challenge of feeding 9 billion people. Science, 327, 812-818.

Goodwin, S. B., Cohen, B. A., & Fry, W. E. (1994). Panglobal distribution of a single clonal lineage of the Irish potato famine fungus. Proceedings of the National Academy of Sciences of the United States of America, 91, 11591-11595.

Gosting, L. H., Cabrian, K., Sturge, J. C., & Goldstein, L. C. (1984). Identification of a species-specific antigen in *Legionella pneumophila* by a monoclonal antibody. Journal of Clinical Microbiology, 20, 1031–1035.

Gottstein, B., Tsang, V. C., & Schantz, P. M. (1986). Demonstration of species-specific and cross-reactive components of *Taenia solium metacestode* antigens. The American Journal of Tropical Medicine and Hygiene, 35, 308-313.

Haas, B. J., Kamoun, S., Zody, M. C., Jiang, R. H., Handsaker, R. E., Cano, L. M., ... & Bozkurt, T. O. (2009). Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature, 461, 393-398.

Hopp, T.P., & Woods, K.R. (1981). Prediction of protein antigenic determinants from amino acid sequences. Proceedings of the National Academy of Sciences of the United States of America, 78, 3824–3828.

Hüberli, D., Tommerup, I.C. & Hardy, G.E.S. (2000) False-negative isolations or absence of lesions may cause mis-diagnosis of diseased plants infected with *Phytophthora cinnamomi*. Australasian Plant Pathology, 29, 164–169.

Hughes, K. J., Tomlinson, J. A., Griffin, R. L., Boonham, N., Inman, A. J., & Lane, C. R. (2006). Development of a one-step real-time polymerase chain reaction assay for diagnosis of *Phytophthora ramorum*. Phytopathology, 96, 975-981.

Iurescia, S., Fioretti, D., Fazio, V. M., & Rinaldi, M. (2012). Epitope-driven DNA vaccine design employing immunoinformatics against B-cell lymphoma: a biotech's challenge. Biotechnology Advances, 30, 372-383.

Judelson, H. S. (2014). Phytophthora infestans. In Genomics of Plant-Associated Fungi and Oomycetes: Dicot Pathogens (pp. 175-208). Springer Berlin Heidelberg.

Dean, R. A., Lichens-Park, A., & Kole, C. (Eds.). (2014). Genomics of Plant-Associated Fungi and Oomycetes: Dicot Pathogens. Springer.

Kolaskar, A. S., & Tongaonkar, P. C. (1990). A semi-empirical method for prediction of antigenic determinants on protein antigens. FEBS Letters, 276, 172-174.

Kox, L. F. F., Brouwershaven, I. V., Vossenberg, B. V. D., Beld, H. V. D., Bonants, P. J. M., & Gruyter, J. D. (2007). Diagnostic values and utility of immunological, morphological, and molecular methods for in planta detection of *Phytophthora ramorum*. Phytopathology, 97, 1119-1129.

Kroon, L. P. N. M., Bakker, F. T., Van Den Bosch, G. B. M., Bonants, P. J. M., & Flier, W. G. (2004). Phylogenetic analysis of *Phytophthora* species based on mitochondrial and nuclear DNA sequences. Fungal Genetics Biology, 41, 766-782.

Langmead, B., & Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. Nature Methods, 9, 357–359.

Le Moigne, V., Gaillard, J. L., Herrmann, J. L., & Roux, A. L. (2017). Virulence Mycobacterial Determinants Represent Useful Antigens for The Serological Diagnostic of NTM Infection in Cystic Fibrosis Patients. B62. Non-tuberculous mycobacteria: clinical aspects and cases (pp. A3957-A3957). American Thoracic Society.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., … 1000 Genome Project Data Processing Subgroup. (2009). The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2078–2079.

MacDonald, J. D., Stites, J., & Kabashima, J. (1990). Comparison of serological and culture plate methods for detecting species of *Phytophthora*, *Pythium*, and *Rhizoctonia* in ornamental plants. Plant Disease, 74, 655-659.

Martin, F. N., & Tooley, P. W. (2003). Phylogenetic relationships among *Phytophthora* species inferred from sequence analysis of mitochondrially encoded cytochrome oxidase I and II genes. Mycologia, 95, 269-284.

Martin, F. N., Tooley, P. W., & Blomquist, C. (2004). Molecular detection of *Phytophthora ramorum*, the causal agent of sudden oak death in California, and two additional species commonly recovered from diseased plant material. Phytopathology, 94, 621-631.

Milton, R. D. L., Milton, S. C., & Adams, P. A. (1990). Prediction of difficult sequences in solid-phase peptide synthesis. Journal of the American Chemical Society, 112, 6039-6046.

Muzzi, A., Mora, M., Pizza, M., Rappuoli, R., & Donati, C. (2013). Conservation of *meningococcal* antigens in the genus *neisseria*. mBio, 4, e00163–13.

Horton, P., & Nakai, K. (1999). A program for detecting sorting signals in proteins and predicting their subcellular localization. Trends in Biochemical Sciences, 24, 34-35.

Nakayama, E., Yokoyama, A., Miyamoto, H., Igarashi, M., Kishida, N., Matsuno, K., ... & Takada, A. (2010). Enzyme-linked immunosorbent assay for detection of *filovirus* species-specific antibodies. Clinical and Vaccine Immunology, 17, 1723-1728.

Nelleman, C. (2009). The environmental food crisis. The Environment's Role in Averting Future Food Crises. A UNEP Rapid Response Assessment (United Nations Environment Program, GRID-Arendal, Arendal, Norway).

Pauza, C. D., Trivedi, P., Wallace, M., Ruckwardt, T. J., Le Buanec, H., Lu, W., ... & Gallo, R. C. (2000). Vaccination with tat toxoid attenuates disease in simian/HIV-challenged macaques. Proceedings of the National Academy of Sciences of the United States of America, 97, 3515-3519.

Pizza, M., Scarlato, V., Masignani, V., Giuliani, M. M., Arico, B., Comanducci, M., ... & Galeotti, C. L. (2000). Identification of vaccine candidates against serogroup *B meningococcus* by whole-genome sequencing. Science, 287, 1816-1820.

Purcell, A. W., Zeng, W., Mifsud, N. A., Ely, L. K., Macdonald, W. A., & Jackson, D. C. (2003). Dissecting the role of peptides in the immune response: theory, practice and the application to vaccine design. Journal of Peptide Science, 9, 255-281.

Ristaino, J. B., Madritch, M., Trout, C. L., & Parra, G. (1998). PCR amplification of ribosomal DNA for species identification in the plant pathogen genus *Phytophthora*. Applied and Environmental Microbiology, 64, 948-954.

Rittenburg, J., Miller, S., & Petersen, F. (1997). Monoclonal antibodies and methods for diagnosis of *Phytophthora* infection. Google Patents. https://google.com/patents/CA1339582C?cl=en.

Russell, H., Tharpe, J. A., Wells, D. E., White, E. H., & Johnson, J. E. (1990). Monoclonal antibody recognizing a species-specific protein from *Streptococcus pneumoniae*. Journal of Clinical Microbiology, 28, 2191-2195.

Savary, S., Ficke, A., Aubertot, J. N., & Hollier, C. (2012). Crop losses due to diseases and their implications for global food production losses and food security. Food Security, 4, 519-537.

Schlessinger, A., Ofran, Y., Yachdav, G., & Rost, B. (2006). Epitome: database of structure-inferred antigenic epitopes. Nucleic Acids Research, 34, D777-D780.

Stam, R., Jupe, J., Howden, A. J., Morris, J. A., Boevink, P. C., Hedley, P. E., & Huitema, E. (2013). Identification and characterisation CRN effectors in *Phytophthora capsici* shows modularity and functional diversity. PloS One, 8, e59517.

Stawikowski, M., & Fields, G. B. (2012). Introduction to peptide synthesis. Current Protocols in Protein Science, 18-1.

Tettelin, H., Saunders, N. J., Heidelberg, J., Jeffries, A. C., Nelson, K. E., Eisen, J. A., ... & Nelson, W. C. (2000). Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. Science, 287, 1809-1815.

Tomlinson, J. A., Dickinson, M. J., & Boonham, N. (2010). Rapid detection of *Phytophthora ramorum* and *P. kernoviae* by two-minute DNA extraction followed by isothermal amplification and amplicon detection by generic lateral flow device. Phytopathology, 100, 143-149.

Tooley, P. W., Bunyard, B. A., Carras, M. M., & Hatziloukas, E. (1997). Development of PCR primers from internal transcribed spacer region 2 for detection of *Phytophthora* species infecting potatoes. Applied and Environmental Microbiology, 63, 1467-1475.

Torres, G. A., Sarria, G. A., Varon, F., Coffey, M. D., Elliott, M. L., & Martinez, G. (2010). First Report of Bud Rot Caused by *Phytophthora palmivora* on African Oil Palm in Colombia. Plant Disease, 94, 1163-1163.

Tsao, P. H. (1983). Factors affecting isolation and quantitation of *Phytophthora* from soil. *Phytophthora* its Biology, Taxonomy, Ecology, and Pathology, 219-236.

Vartanian V.G., & Endo R.M. (1985). Survival of *Phytophthora* infestans in seeds extracted from infected tomato fruits. Phytopathology, 75, 375-378.

Wilson, B. A., Aberton, J., & Cahill, D. M. (2000). Relationships between site factors and distribution of *Phytophthora cinnamomi* in the Eastern Otway Ranges, Victoria. Australian Journal of Botany, 48, 247-260.

Chapter II

Identifying novel transcripts, alternate splicing and improving gene annotations in plant pathogen *Phytophthora infestans* using RNA-seq data

Abstract: Correct gene models and annotations are important for studying this pathogen. In this study, we have used RNA-Seq data to modify the current annotations of the *P. infestans* genome. Trinity and PASA were used to update the existing gene models, identify new genes, non-coding RNAs and alternative splicing events. We used genome-guided Trinity to first align the reads to the T30-4 reference genome, and then PASA to align the transcripts obtained by Trinity back to the genome. Out of 17,797 protein coding genes currently annotated in the latest public release of the genome generated by Broad Institute, our analysis resulted in the modification of 11,938 genes with additions of untranslated regions, changes in CDS boundaries, gene splitting and gene merging. 1603 genes exhibited alternative splicing, which mostly involved intron retention. 3,218 transcripts with open reading frames did not map to any genes in public annotation were screened for orthologs in other *Phytophthora* species resulting in the addition of 482 new protein-coding genes in the current annotation. We have also identified 8,115 non-coding RNAs that were previously absent in Broad Institute annotation. Apart from structural annotation, functional annotation of the current genome has annotated 11,628 transcripts with biological functions and has identified 1,789 secreted proteins. This study will help researchers in terms of accurate prediction of transcription and translational start sites, domain analysis, expression analysis and molecular characterization to name a few.

Background: Correct gene annotations play an important role in understanding the biology of any organism. Genome annotation in terms of identifying gene models is a complex process and relies on multiple factors including gene prediction algorithms, expressed sequence tags (EST) data and homology with proteins from closely related species. The combination of gene predictions, EST alignments and protein sequences from closely related species can lead to spurious gene predictions; EST alignments provide an accurate representation of the transcript but it only gives partial or limited information on the complete gene structure; gene prediction algorithms predict genes based on the training sets and genes that do not have the same coding/non-coding pattern as the training sets might be wrongly predicted (Korf et al., 2004; Mathé et al., 2002; Elsik et al., 2014). Some of the important aspects of genome biology that are affected by incorrect gene models include functional and structural annotation, identification of transcription factor binding sites, protein domain analysis and ortholog identification. Identification of a full transcript set that includes protein coding genes, small and long non-coding RNAs, and splicing isoforms can give us a solid ground for downstream analysis.

RNA-Seq has opened up new doors to identify and correct both protein coding genes and non-coding RNAs in non-model organisms in a high-throughput manner. Different packages like *SOAPdenovo, ABySS* or *trinity* can do a *de-novo* assembly when a reference genome is not present, while packages like *gsnap* or *tophat* can align the reads to reference genome followed by assembly of alignments with programs like *PASA* or *cufflinks* (Li et al., 2010a, b; Birol et al., 2009; Haas et al., 2013; Wu and Nacu, 2010;

Trapnell et al., 2009; Haas et al., 2008). Many model and non-model organism genome annotations have been improved with the use of RNA-Seq data. These include well-known organisms such *Arabidopsis thaliana*, *Aspergillus* and *Saccharomyces* (Roberts et al., 2011; Haas et al., 2003; Cerqueira et al., 2014; Nagalakshmi et al., 2008) and many newly sequenced genome like *Dictyostelium*, *Ailuropoda melanoleuca*, *Tuber melanosporum* (Singh et al., 2017; Chen et al., 2015; Tisserant et al., 2011).

Previous studies have reveals problems related to inaccurate gene models in different species of *Phytopthora* (Gan et al., 2009; Ospina-Giraldo et al., 2010; Roy et al., 2013; Tripathy et al., 2006; Meijer et al., 2014). *De novo* proteome assembly identified 150 proteins that were initially predicted to be non-secretory but after curation had a signal peptide and were added to the secretome of *P. infestans* (Meijer et al., 2014). Gan et al. had to manually curate many incompletely predicted gene models of cyclophilin-encoding genes in *Phytophthora* species, which resulted in a more accurate analysis of homology between orthologs in *P. sojae*, *P. ramorum* and *P. infestans* (Gan et al., 2009). Another study on RxLR motif-containing proteins in *Phytophthora* (a motif that is present in pathogenicity related proteins and is possibly involved in translocating the proteins inside the host cells) found 109 genes lost their qualification as RxLR effectors after gene model corrections (Jiang et al., 2007). In the present study, we address the problems associated with current gene models in the plant pathogen *Phytophthora infestans* using RNA-Seq data.

The assembly of *P. infestans* strain T30-4 was generated using *shot-gun* sequencing and genes were identified based on the gene-finding algorithms *Orthosearch*

and *GeneID* (Haas et al., 2009). *Orthosearch* defines ORFs based on both hidden markov models and by comparison to orthologs from closely related species. *GeneID* is an *ab initio* gene caller that was trained on limited EST evidence from 520 genes; many genes which did not follow the same coding vs non-coding sequences were thus misrepresented in the Broad Institute's annotation. The gene models from *P. sojae* and *P. ramorum* on which the Orthosearch was run are not so reliable, which led to the propagation of many incorrect annotations (Gan et al., 2009; Ye et al., 2011). Also, manual annotation and curation of genes in the prior studies mostly emphasized effector proteins, which play an important role in virulence and host colonization (Haas et al., 2009).

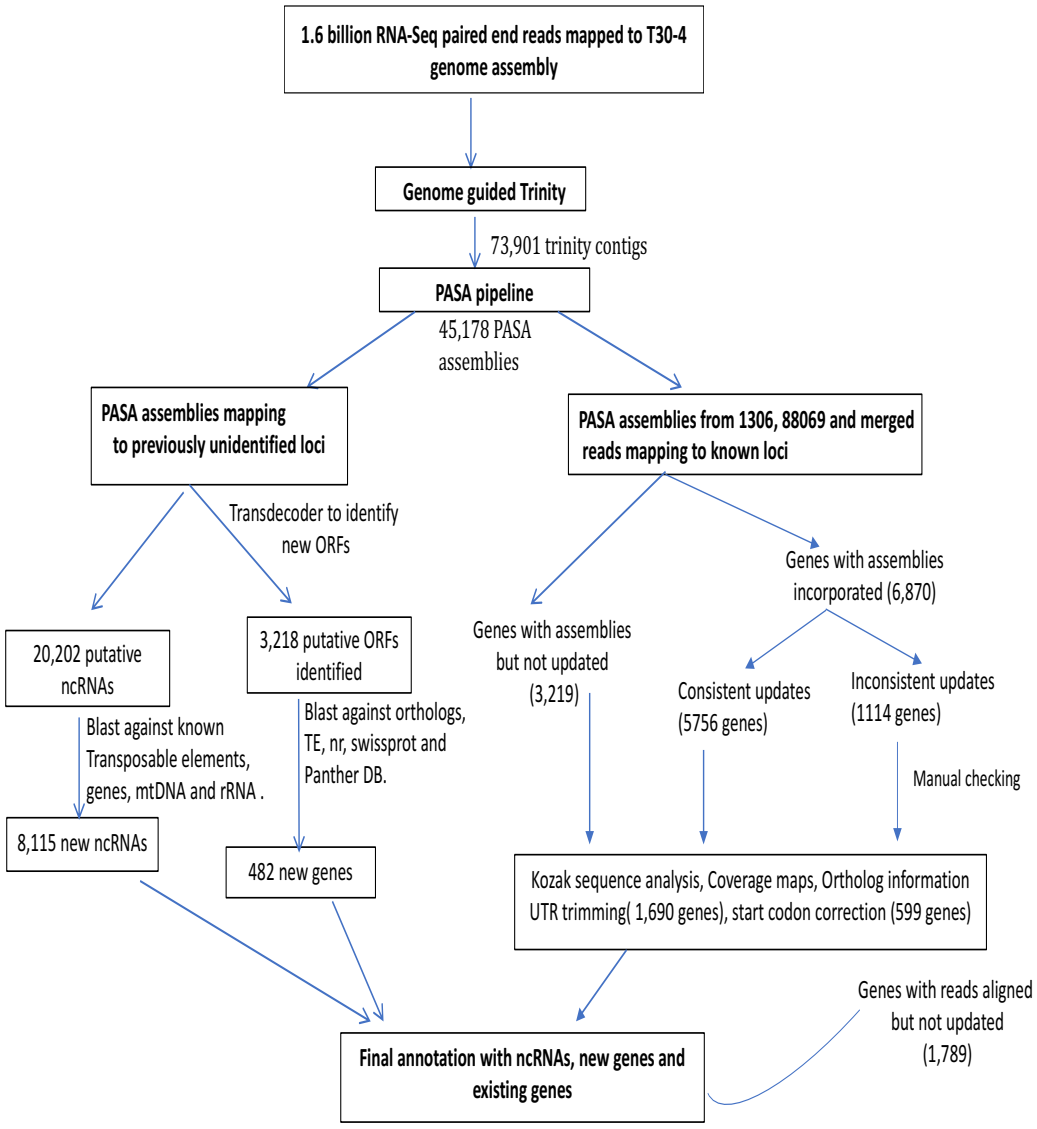
Good quality genomic annotations are required to get a better understanding of various aspects of *Phytophthora* biology; some of these include bioinformatics analysis, while some might directly affect the molecular characterization of the gene of interest. The present study used RNA-Seq data to correct these gene annotations and identify discrepancies that are normally associated with *ab initio* gene callers, although this can only be done for expressed genes. This used RNA-Seq data from different conditions and developmental stages – mycelia, sporangia, cleaving sporangia, germinating cyst, potato tubers and tomato leaves infected with *P. infestans* strains 1306 and 88069. Changes in protein coding genes include exon boundaries corrections, gene merging events, gene splitting events, alternative splicing and re-annotation of protein coding genes to non-coding RNAs or pseudogenes. Also, after incorporating *PASA* assemblies, the final changes in protein coding genes were checked for correct UTR lengths and start codon using strand-specific RNA-Seq data, Kozak sequence and ortholog information to get the

best quality genome annotation. We have not only made changes in the protein coding genes, but also identified new protein coding genes and noncoding RNAs that are present in *P. infestans*.

Results:

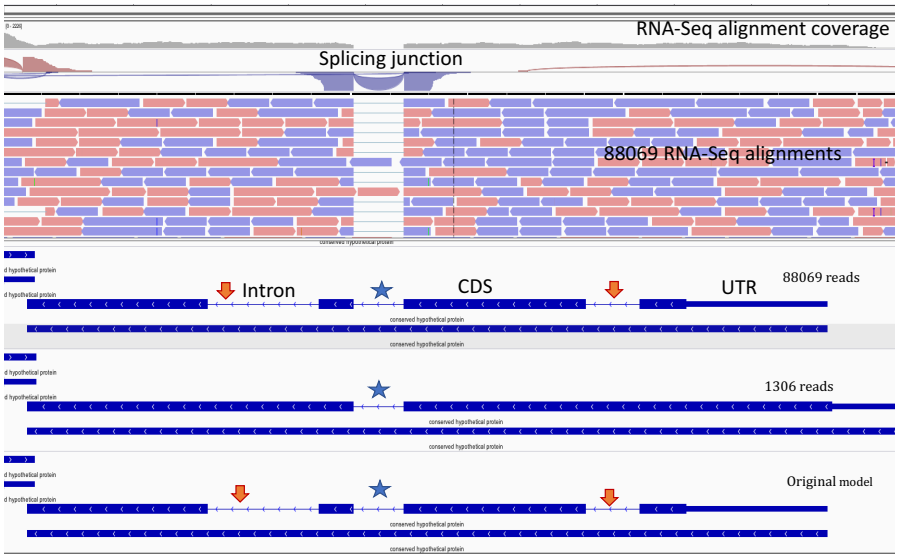
I. Gene structure modification pipeline with RNA-Seq reads: Approximately 1.6 billion RNA-Seq reads obtained from 1306 and 88069-infected potato tubers, tomato leaves and media cultures were used in *trinity* and *PASA* pipelines to update the existing annotation of *P. infestans* and leading to the modification of 5,428 gene structures (Haas et al., 2013; Haas et al., 2003). The complete annotation pipeline to identify and correct genome annotations is given in figure 1.

Figure 1: Complete annotation pipeline that was used to modify the existing gene structures along with addition of novel genes and non-coding RNAs that were missing in the previous annotation.



This study was complicated by the fact that RNA-Seq did not come from the same strain (T30-4) but from two different strains. When the reads from both the strains were merged to run *trinity* and *PASA*, many *PASA* assemblies did not update the gene structure due to mismatches between the reference genome and *PASA* assembly. In order to overcome this difference between strains, we ran 2 separate pipelines for *trinity* and *PASA* – one for 1306 and another for 88069. This helped us in identifying an additional 4,417 genes that were previously left unchanged; however, it also resulted in scenarios where the same gene was differently modified by each strain. 1,114 gene models changed differently between 1306 and 88069 based on their RNA-seq data were visually inspected in IGV to confirm which predicted gene structure is supported by RNA-Seq read alignments. The dataset that correctly represented the gene model was chosen as the final structure. A representative example of one genes that was modified differently by 1306 and 88069 and resulted in different gene models is shown in Figure 2. The original model predicted is at the bottom panel and predicted 3 introns. In contrast, reads from 88069 did not make any changes to the original gene model but it is clear that the 2 of the 3 predicted introns (shown with orange arrows) are not supported by 88069 RNA-Seq alignments, similar alignments were obtained from 1306 reads (data not shown) so in this case only the model predicted by 1306 reads is in sync with RNA-Seq reads and was used to update the gene model.

Figure 2: Comparison of genes modified differently by different strains (1306 reads and 88069 reads). Selection of the correct model based on visualization of RNA-Seq reads in IGV. Arrows point to predicted introns that are not supported by RNA-Seq while the star point to the intron supported by RNA-Seq.

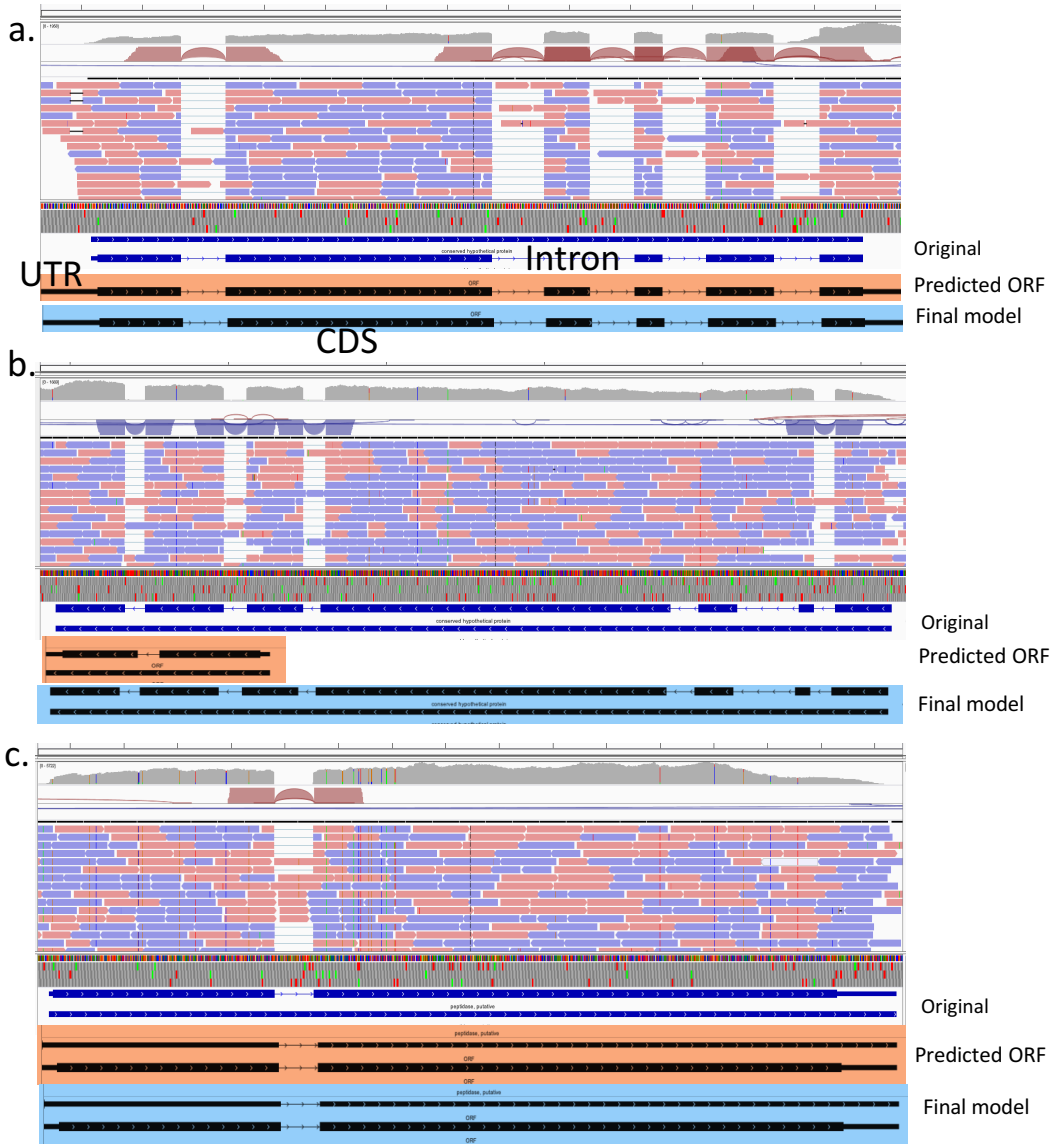


We wanted to make sure that all the genes with assemblies were updated so we checked which of the existing genes have *pasa* assemblies. We used *transdecoder* utility of *PASA* pipeline to predict open reading frames in the assemblies and compared the *transdecoder* predicted models to existing gene models. 3,219 genes had *pasa* assemblies but were not updated and the gene structures were in discordance with *transdecoder* predicted ORFs. We confirmed the results from *pasa* assemblies of these 3219 genes by visualizing the alignments in IGV and modified 1,945 genes that were supported by RNA-Seq data (Figure 3a). The remaining 1,274 predicted ORFs did not match the RNA-Seq alignments and were not incorporated in the respective gene models (Figure 3b). The most common reason that *PASA* did not incorporate those assemblies was due to mismatches present at the exon-intron boundaries to which *PASA* is very sensitive and will not incorporate the *pasa* assemblies to update the model.

Three different scenarios are represented in Figure 3; Figure 3a represents the scenario where *transdecoder*-predicted ORF is supported by RNA-Seq alignments and the gene model is modified as per the *transdecoder*-predicted ORF. Figure 3b represents the scenario where the predicted ORF was not incorporated to modify the gene model because the predicted ORF does not comply with RNA-Seq alignments; RNA-Seq reads show that the gene has 5 exons while the predicted ORF has only 2 exons. 2c represent the scenario where both the predicted and existing ORF are in-sync with each other and thus no changes were required.

Figure 3: Incorporation of ORFs predicted by *transdecoder* to update gene model. a. Correctly predicted ORF incorporated into gene model based on IGV visualization b. Incorrectly predicted ORF does not comply with RNA-Seq data c. Original gene model and predicted ORF are in sync and supported by RNA-Seq. Original model from Broad Institute is shown in white, *transdecoder*-predicted ORF is shown in orange and final gene model present in current gff files is shown in blue.

Figure 3


a.
UTR
Intron
CDS
Original
Predicted ORF
Final model
b.
Original
Predicted ORF
Final model
c.
Original
Predicted ORF
Final model

Confirmation of gene models for antibody development (Chapter I) showed that many genes had either uneven RNA-Seq coverage or no coverage at all. We calculated the coverage of all the genes that were present in the original annotation from the Broad Institute and identified 2,480 genes had <1x coverage (Supplementary table 1). We also identified 1,789 existing genes that had ≥5x read coverage but were unchanged due to lack of *PASA* assemblies. We manually curated those genes using custom perl scripts and visualization of alignments in IGV and were able to modify 1,092 genes using the above strategy. We were not able to confirm the gene models of the remaining 692 genes due to many reasons that include large gene families, not enough coverage to distinguish exon from introns, or gaps in the genome sequence. Genes that were part of families had RNA-Seq reads mapping to multiple locations; hence, we couldn't confirm if the read originated from that particular gene.

II. Kind of annotation changes made to the genome: Using RNA-Seq data we have made changes to gene models that include gene merging, gene splitting, alternative splicing identification, UTR addition and CDS boundary changes (Table 1; Figure 4). Gene models of 11,938 genes, which account for 67% of the total number of protein coding gene have been modified. Apart from modifying the gene structures we also found that 66 of the previously protein coding genes do not have a proper ORF and are now annotated as pseudogenes/ncRNAs in current annotation (Suppl table 2). Comparison of old and new gff files show isoform frequencies that were missing in old annotation, many

more genes have been given a functional annotation and more proteins are predicted to be secreted (Table 2).

We have also identified 200 genes that could not be validated or changed due to assembly problems in T30-4. These genes have N's in the start, end or middle of the mRNA and hence no reliability can be placed on their gene models (Supplementary table 3).

Figure 4: Kinds of annotation changes made to the genome a. UTR additions b. Addition of new exons or removal of introns based on RNA-Seq data c. Extension of proteins at 5' or 3' end d. Merging of two different loci into one gene e. Splitting of genes into 2 or more genes.

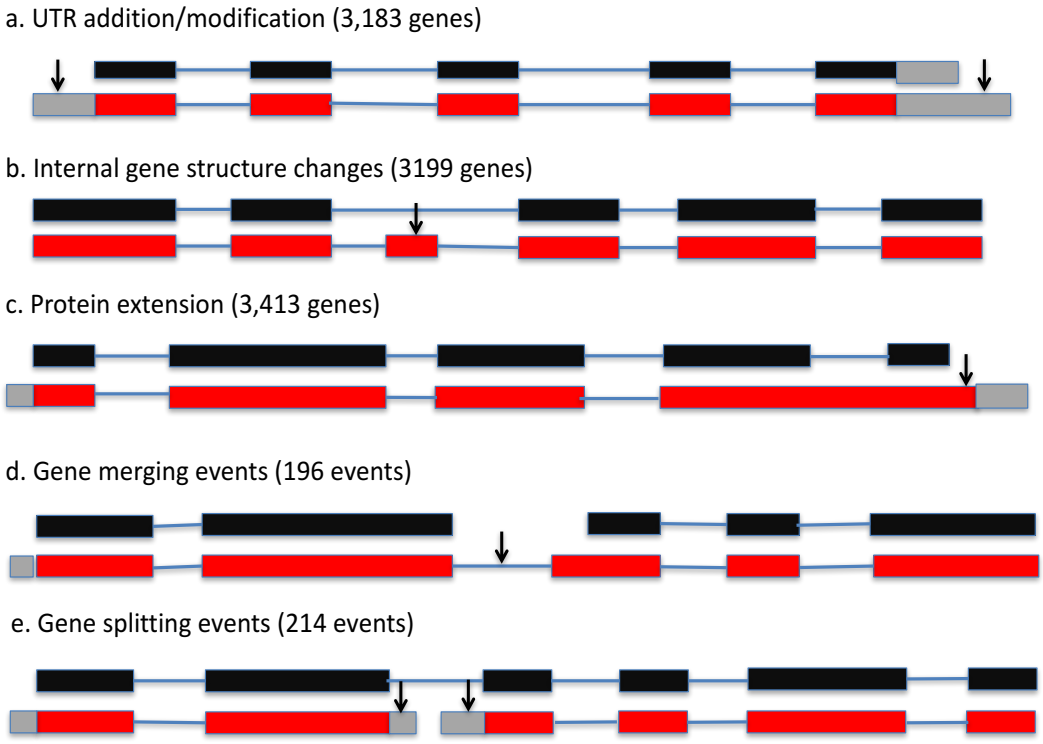


Table 1: Number of genes affected by re-annotation and their categories

Category	# of genes
Total # of genes (Broad)	17,797
Genes Modified	11,938
Genes with unchanged models	2,767
Ambiguous (lack of reads, N's in middle)	4,967
Gene splitting events*	214
ncRNAs (previously protein coding genes)	66
Genes merging events**	195
Genes with alternative splicing	1,603
CDS adjustment*	6,612
UTR addition	3,183

* may include genes with UTR addition also
**details in gene merging event in results section

Table 2: Comparison of old and new gene annotations

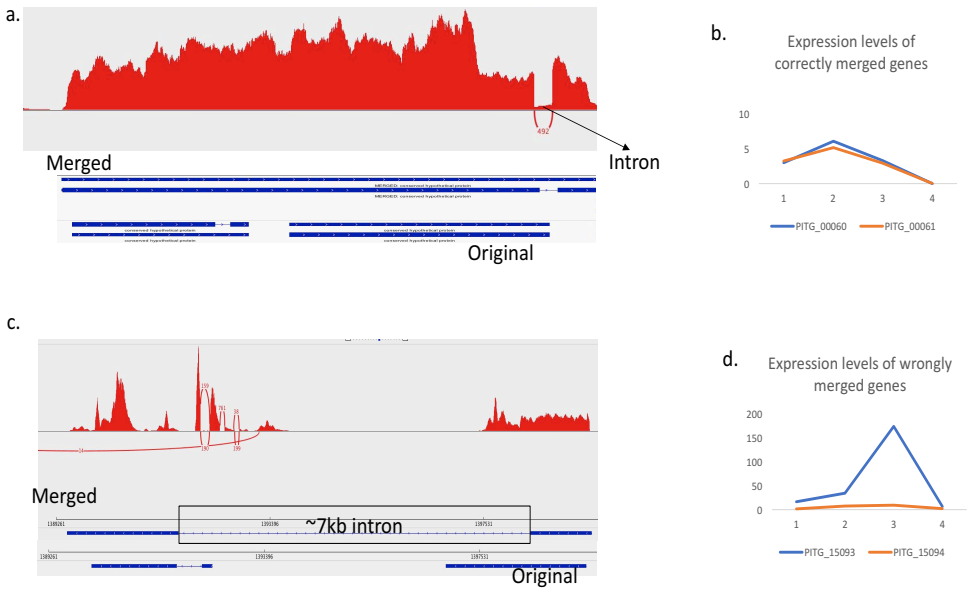
	Old	New
Total # of genes	17,797	18,672
Exon Length*	373.03	540.55
CDS length*	1261.77	1562.53
Intron length*	218.51	104.84
Exons per gene	2.46	2.99
# of genes with isoforms	-	1603
Introns per gene	1.78	2.22
Functional annotation**	6440	10064
Secreted proteins	1415	1789

*Length in base pairs
** Genes with assigned functions as defined by best blast hits in Panther and Swiss-prot with e-value cut-off 1e-10. These also include GO terms.

a. Gene merging events: These include genes that were previously thought to be separate but *PASA* assemblies showed that they were in fact one single gene. A total of 221 merging events were reported by *PASA*. These gene merges were checked for both expression values based on old models and new models and *sashimi* plots to confirm the event (Figure 5). 195 gene merging events were confirmed by visualization using *sashimi* plots in IGV and expression analysis. If the separate genes are part of the same transcript then their expression patterns in different developmental stages should be equal. 195 gene merging events included merging of 2 genes (179 events), 3 gene merging (10 events), 4 genes merged (5 events) and one event of 5 genes merged.

Figure 5: Gene merging events checked for validity using *sashimi* plots. Top model named 'Merged' of 4a and 4b shows the merged gene predicted by *PASA* while the bottom model named 'Original' showed the annotations of two separate genes. 5a represents a correct gene merging predicted by *PASA* which shows equal number of reads in both genes and correct placement of intron. 5b shows expression levels of originally separate genes (now merged) in mycelia (1), sporangia (2), cyst (3) and germinating cyst (4) with FPKM values on y-axis. 5c shows wrong gene merging predicted by *PASA* with 7kb long intron that is not supported by RNA-Seq reads. 5d shows the expression levels of incorrectly merged genes with same x and y-axis as in 5b.

Figure 5


a.
Merged
Intron
Original
b.
Expression levels of correctly merged genes
10
5
0
1
2
3
4
PITG_00060
PITG_00061
c.
Merged
~7kb intron
Original
d.
Expression levels of wrongly merged genes
200
150
100
50
0
1
2
3
4
PITG_15093
PITG_15094

b. Gene splitting events: These included genes that were initially thought to be one single gene but based on ORFs predicted by *transdecoder* on *PASA* assemblies and visualization of RNA-Seq alignments in IGV and were subjected to splitting. In most of the cases the gene was split because *PASA* assembly did not support the predicted intron. A total of 216 genes were split into 2 or more genes.

c. New genes: PASA assemblies with no matches to existing annotation were checked for the presence/absence of ORFs by *transdecoder*. After checking the putative ORFs for proper start/termination codon, putative novel genes were checked for homology in other *Phytophthora* species to confirm if these are real genes. 479 genes have orthologs with 60% similarity along the whole length of the protein in *P. capsici* (125 genes) and *P. parasitica* (225 genes). We have added 482 new genes, with a total of 215 genes shared between *P. infestans, P. capsici* and *P. parasitica* where both *P. capsici* and *P. parasitica* had the almost the same percent identity (within 2%) and e-value.

d. UTR annotations: UTRs play an important role in translation and were annotated for 1,756 genes in the original annotation from Broad Institute which used EST alignments to predict UTRs. Few genes from Broad annotation had EST data and since EST alignments cannot give full length transcripts many genes with UTRs were modified by extending UTRs in current analysis. In the present study, we have added or modified 5' and 3' UTRs in 7567 genes. Previous studies have also shown that 'unstranded' RNA-Seq in gene-dense regions can be erroneously joined to the 5' and 3' end of adjacent genes,

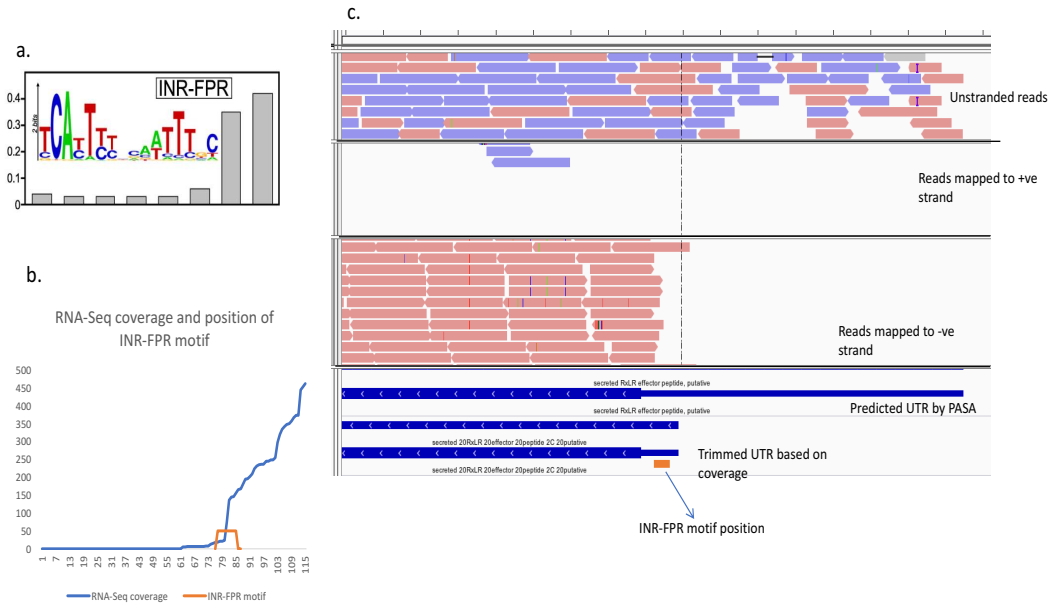
leading to long but incorrect UTR predictions (Testa et al., 2015; Hassan et al., 2012). Since we used 'unstranded' RNA-Seq data, we wanted to make sure that the UTRs predicted by our pipeline are not affected by this feature of gene-dense regions. Studies have shown that the length of UTRs is in general inversely proportional to the translational capability of transcript (Velden et al., 1999); thus, it was very important to identify correct transcription start site.

We used strand-specific RNA-Seq data along with INR-FPR motif analysis (Roy et al., 2013) to confirm the transcription start site. Promoter analysis of *P. infestans* genome identified 3 core promoter elements – INR, FPR and DPEP. INR-FPR motif is linked to promoters with higher than average mRNA expression levels and spans the transcription start site (Roy et al., 2013). All the 7567 genes with UTRs added by *PASA* were checked for the length of 5' and 3' UTRs to find genes with >100 bp long UTR. 4,854 genes had >100 bp 5' or 3' UTRs. We screened for the coverage of these genes using strand-specific data that our lab recently obtained from selected developmental stages. Out of 4854 genes, 2134 genes had >10x coverage and were screened for coverage of 5' and 3' UTR. UTRs were trimmed to the position where the RNA-Seq read coverage dropped to zero based on the strand-specific reads. Using the above strategy, we were able to trim the UTRs of 1,690 genes (Figure 6). 3,965 genes with >100bp 5' UTR were screened for INR-FPR motif within the UTR sequence using the position-weighted matrix for INR-FPR motif and 179 genes were found to have the INR-FPR motif within the UTR. We trimmed the UTR for these 179 genes to the position where the INR-FPR

motif was present. The average 5' UTR size of the *PASA*-predicted UTRs after trimming went from 349 bp to 340 bp while 3' UTR dropped from 780 bp to 640 bp.

Figure 6: INR-FPR motif analysis and coverage data used to trim *PASA*-predicted UTRs. Figure 6a shows the INR-FPR motif that was used in trimming 5' UTRs. Figure 6b shows strand-specific RNA-Seq coverage with x-axis showing the position of the complete predicted UTR along with first 3 bases of coding sequence, y-axis is displaying RNA-Seq coverage for each position in UTR and is displayed in blue. The position of the INR-FPR motif is shown in orange. Figure 6c shows the read alignments in IGV based on 'unstranded' RNA-Seq data, and reads mapping to sense and antisense strands. Bottom panel shows the *PASA*-predicted 5' UTR and trimming of the UTR based on coverage, along with the INR-FPR motif location.

Figure 6


a.
INR-FPR
b.
RNA-Seq coverage and position of INR-FPR motif
RNA-Seq coverage
INR-FPR motif
c.
Unstranded reads
Reads mapped to +ve strand
Reads mapped to -ve strand
Predicted UTR by PASA
Trimmed UTR based on coverage
INR-FPR motif position

e. *Alternative Splicing of genes:* Cufflinks was run on all the genes that were reported by *PASA* as alternatively spliced in all developmental stages. Most of the genes had only one major isoform and only those minor isoforms were reported which had at least 10% expression value as the major isoform. This led to the removal of initially reported 478 isoforms belonging to 174 genes. We have identified 1,603 genes with alternatively spliced variants after filtering of isoforms using the above criteria (Figure 7).

Alternative splicing (AS) has been noted for some gene families in *P. infestans* and *P. sojae* where the majority of AS variants were due to intron retention (García-Bayona et al., 2013; Costanzo et al, 2007). Previous study in *Pseudoperonospora cubensis* validated the alternative splicing of a multi-drug transporter which resulted in two different transcripts; one in which the intron is retained leading to RxLR effector protein with a smaller transcript and another where intron splicing leads to a larger transcript with functional domain (Savory et al., 2012).

This is the first time in *P. infestans* that a genome-wide analysis to find alternative spliced genes have been done. Most of the genes have 2 isoforms (1,004 genes), 221 genes have 3 isoforms and 378 genes have more than 3 isoforms with a maximum of 36 isoforms for PITG_18473 that encodes for 3-phosphoinositide-dependent protein kinase. We also identified genes such as PITG_03751 and PITG_05888 that showed variation at their N-termini, having one isoform with and the other without a signal peptide (Table 3).

Figure 7: Examples of Alternative Splicing events observed in *Phytophthora infestans* and the numbers of each event.

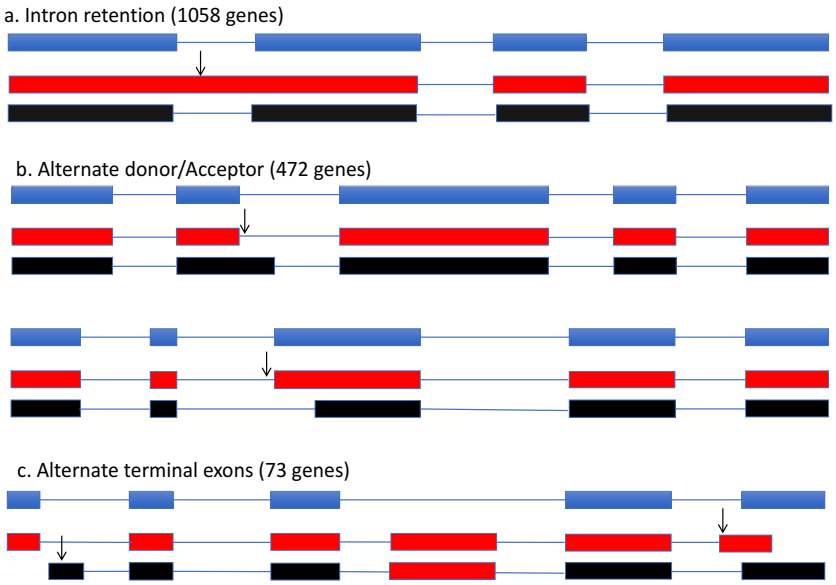


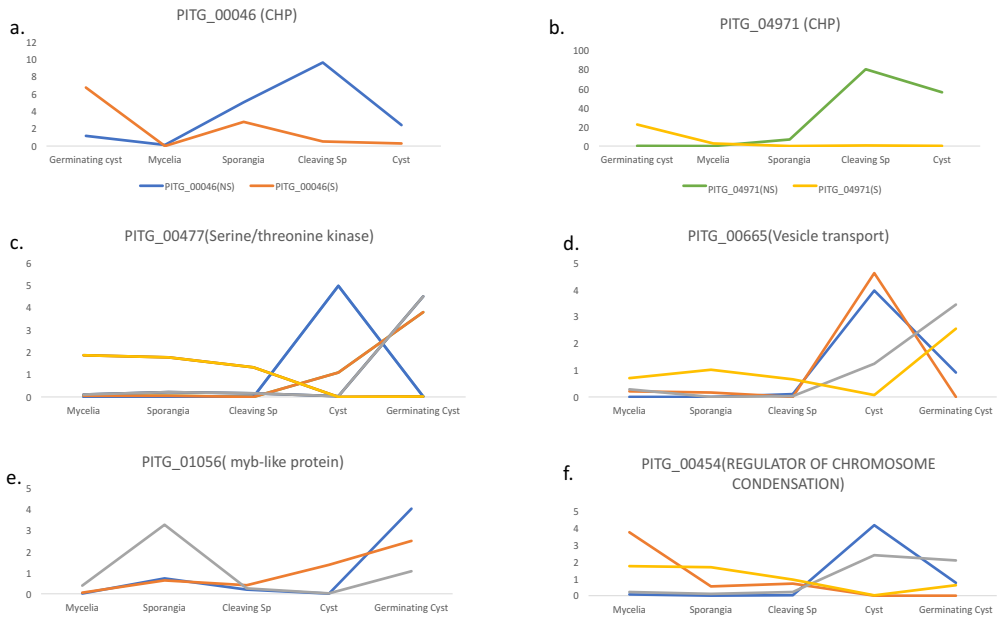
Table 3: Genes with alternative-splicing event where one isoform has and the other does not have signal peptide

Protein Name	Functional Annotation
PITG_00046	Fibrinogen/tenascin/angiopoeitin
PITG_00189	Acid phosphatase-related
PITG_00468	Conserved Hypothetical protein
PITG_01143	N-alpha-acetyltransferase 30
PITG_01509	Conserved Hypothetical protein
PITG_01598	Vesicular mannose-binding lectin
PITG_02075	1-acylglycerol-3-phosphate acyltransferase-related
PITG_02636	Peroxisomal farnesylated protein
PITG_03425	Glycolipid-anchored surface protein 2
PITG_03496	Methyltransferase-like protein 22
PITG_03751	Nipped-b-like protein
PITG_04971	Serine/threonine-protein kinase
PITG_05888	Retinol dehydrogenase 10
PITG_06384	Teneurin and n-acetylglucosamine-1-phosphodiester alpha-n-acetylglucosaminidase
PITG_07307	Filamin-C
PITG_08073	Glucose-6-phosphate 1-dehydrogenase
PITG_08655	Transmembrane emp24 domain protein
PITG_10852	Conserved Hypothetical protein
PITG_12155	Peptidyl-prolyl cis-trans isomerase
PITG_13844	Conserved Hypothetical protein
PITG_13939	Upf0577 protein kiaa1324
PITG_14176	Sphingomyelin phosphodiesterase
PITG_15039	Secreted rxlr effector peptide
PITG_15918	Conserved Hypothetical protein
PITG_17841	Conserved Hypothetical protein
PITG_20094	Tyrosinase
PITG_20762	HD domain-containing protein
PITG_21321	Udp-n-acetylglucosamine--peptide n-acetylglucosaminyltransferase

Intron retention accounted for ~66% of alternatively spliced variants and was the most common form of alternative splicing. Apart from intron retention, other splicing events included alternate donor-acceptor site, alternate start or end exon (Figure 6). Out of 293 genes that exhibited alternative splicing event in 5' or 3' UTRs, 194 genes have showed isoforms due to intron retention/skipping in 5' and 3' UTR.

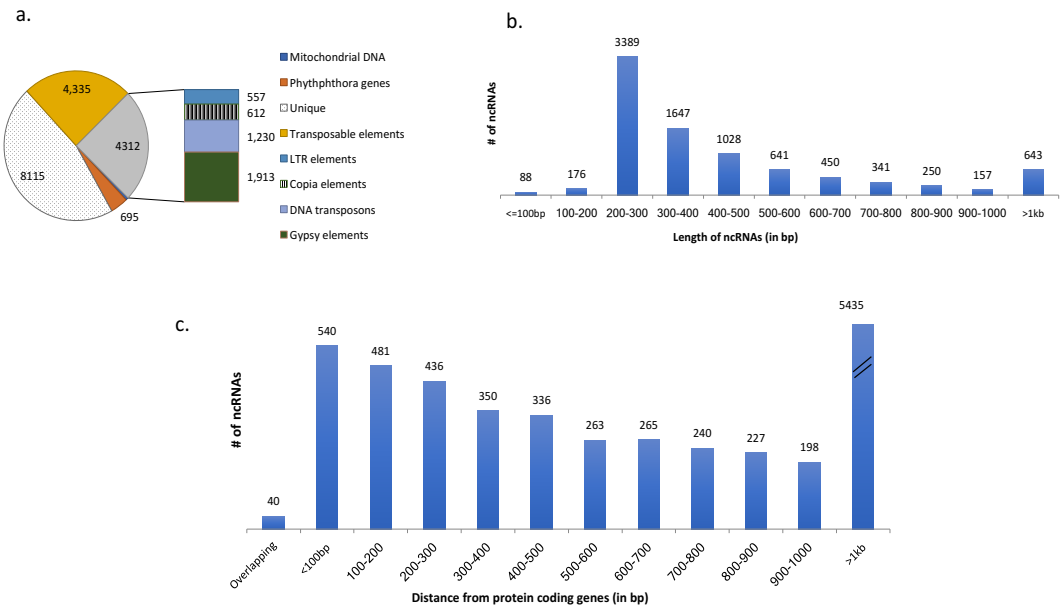
Previous studies have shown that *P. infestans* has a dynamic transcriptome with many genes expressed only in different developmental stages (Judelson et al., 2008; Abrahamian et al., 2016). We wanted to see if any of these isoforms are stage-specific or if the major isoform is dominant in all stages of development. Expression levels of all the isoforms from different developmental stages — mycelia, sporangia, cysts, germinating cysts and zoospores — showed that many isoforms were stage-specific (Figure 8). Also, certain isoforms with signal peptide have shown higher expression in early stages of infection, for example both the secreted forms of PITG_00046 and PITG_04971 (encoding a conserved hypothetical protein) are expressed at higher levels in germinating cysts which indicates early infection stages suggesting that they might play a role in infection while the non-secreted form is expressed in later developmental stages (Figure 8a, b).

Figure 8: Expression levels of genes with differential expression in different developmental stages. Conserved hypothetical protein is CHP. x-axis shows the developmental stages for which expression data for the gene was collected; y-axis shows the FPKM values of the gene in each stage. Gene with one secreted and one non-secreted form are presented in (a) and (b) where (S) represents secreted while (NS) stands for non-secreted. Genes with differentially expressed isoforms but all of them are non-secretory are presented in c-f.



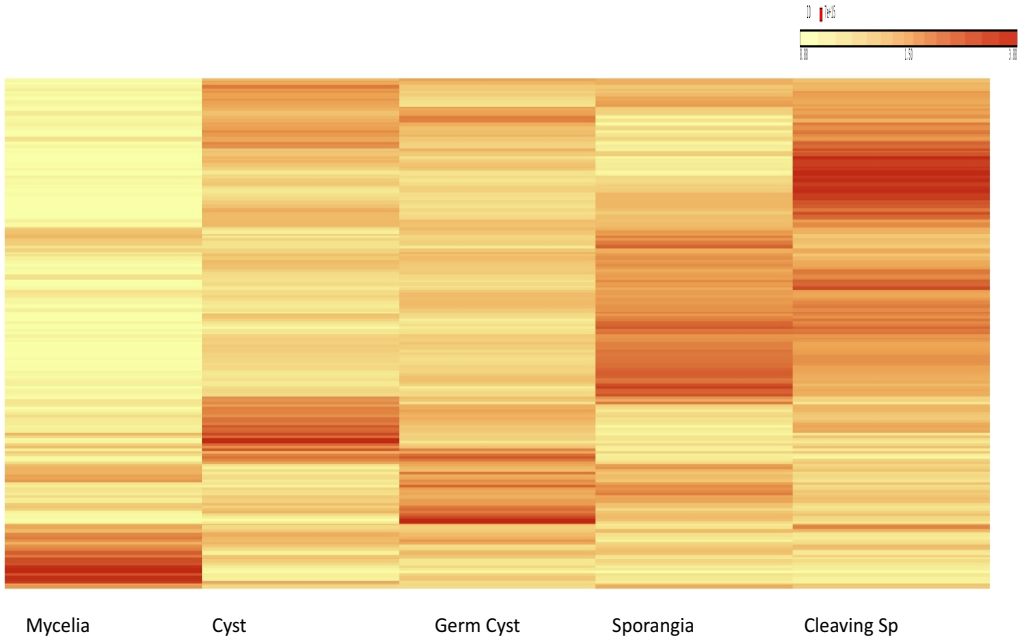
f. P*utative ncRNAs:* 20,061 *PASA* assemblies with >10x RNA-Seq read coverage had no overlap with existing genes and had no predicted ORFs of length ≥100 amino acids. These assemblies are annotated as putative unique ncRNAs after excluding the ones that were having homology to known transposable elements (TE), mtDNA, rRNA and existing protein coding genes with an e-value cut-off of 1e-10 (Figure 9). Figure 9a shows the number of putative ncRNAs that matched TE (LTR elements, Copia elements, DNA transposons and Gypsy elements), mtDNA, rRNA and protein coding genes using the above threshold; 8,115 ncRNA with no homology to any of the above were classified as unique and added to the genome annotation. These putative ncRNAs were checked for their length distribution and their distance from protein coding genes; ncRNAs of size 200-300 bp long were more common than any other type (Figure 9b) and most of the ncRNAs (5,435) were more than 1 kb away from protein-coding genes (Figure 9c). The 40 ncRNAs that overlap with genes were checked for antisense transcripts to see if they might play roles in gene expression; 20 were found to be antisense.

Figure 9: Classification, length distribution and vicinity of ncRNAs to protein-coding genes in *P. infestans.*



Expression patterns of ncRNAs were compared to protein-coding genes and the overall expression levels of ncRNAS (average FPKM 2.5; standard deviation 24.88) were very low as compared to protein-coding genes (average FPKM 30.5; standard deviation 304.78). We also checked if the putative ncRNAs have stage-specific expression patterns and found that 882 ncRNAs have a >2-fold difference and 453 having a >5-fold expression difference (Figure 10) between different developmental stages, using a *p*-value cutoff of 0.0001.

Figure 10: Expression profile of ncRNAs in different developmental stages.



III. Defective alleles of protein-coding genes: Populations reproducing asexually have a tendency to collect defective alleles over time (Goodwin, 1997; Sniegowski et al., 1997). Many of these cause the shortening of the proteins to the extent that either the protein is non-functional or the two forms have different functions. We checked for defective alleles using 1306 reads that might cause a change in protein length.

GATK identified 14,789 SNPs/INDELs present in 1306 when compared to T30-4 reference genome. Out of 14,789 SNPs/INDELs predicted to be present in the entire genome 6,637 SNPs/INDELs were present in genes. These SNPs/INDELs were then annotated using Ensembl Variant annotation package. 477 variants were found to cause premature ending of proteins. These 477 variants were checked for correct gene annotations because the Ensembl annotation database used for annotating variants used old gene models; this removed 284 variants from our analysis because the old model did not match the newer gene models. The remaining 193 genes with potential defects were checked for functional annotations, presence of SNPs close to INDELs and presence of 3 or more SNPs within 10 bp region. Functional annotation of genes showed that 18 variants were present in RxLR genes, which we know are evolving rapidly (Haas et al., 2009; Raffaele et al., 2010). SNP calling is affected by presence INDELs (Van der Auwera et al., 2013) and 24 variants were discarded as they were within 15bp of INDELs. 99 variant genes were discarded due to presence of 3 or more variants within the same 10 bp region based on previous studies that suggested removal of SNP clusters that are within 10 bp of each other (Wu et al., 2010; Nielsen et al., 2011). After filtering of bad gene models, RxLR genes, SNP clusters and SNPs present in vicinity of INDELs

we have identified 52 putative defective alleles, 51 of them were SNPs causing premature termination of proteins (Supplementary table 4).

To confirm the variants using Sanger sequencing, we selected 10 genes that had a functional annotation, were not part of a gene family, and for which the mutation was within the functional domain (as defined by conserved domain database of NCBI) search of the protein. 9 of these were SNPs that caused premature stop and one of them was an INDEL. Sanger sequencing of the region with SNPs or INDELs showed that 7 of the 9 genes were indeed heterozygous as predicted. The SNP in gene PITG_09671 was confirmed by Sanger sequencing of PCR products. Sanger sequencing showed that the SNP in gene PITG_17586, which was predicted to cause premature termination of the protein in 1306, appeared to be homozygous for the apparent defect in 1306 (Table 4).

Table 4: Summary of defective alleles present in 1306 when compared to T30-4 confirmed by Sanger sequencing.

PITG#	Functional annotation	Type	Confirm
PITG_02119	Serine-threonine protein kinase	SNP	Yes
PITG_03143	Pseudouridine-metabolizing bifunctional protein	SNP	Yes
PITG_03289	Protein NLRC5	SNP	Yes
PITG_08734	Chloride chanel	SNP	Yes
PITG_09671	Protein-Tyrosine Phosphatase	SNP	No
PITG_10401	serine-threonine protein Phosphatase	SNP	Yes
PITG_12115	Calmodulin	SNP	Yes
PITG_17094	HTH-type transcriptional regulator	SNP	Yes
PITG_17586	Serine-threonine protein kinase	SNP	Yes
PITG_20241	No conserved domain	INDEL	Yes

IV. Functional annotation of the genome: The 22,246 predicted transcripts currently annotated in the genome were blasted against PANTHER and Swiss-prot database to get annotations and corresponding GO terms (Supplementary table 5). Out of 10,828 conserved hypothetical proteins and hypothetical proteins present in the original annotation, functional annotations were assigned for 3,369 and 112 proteins, respectively. Out of 22,246 proteins present in the current version, we were able to assign GO annotations for 11,630 proteins.

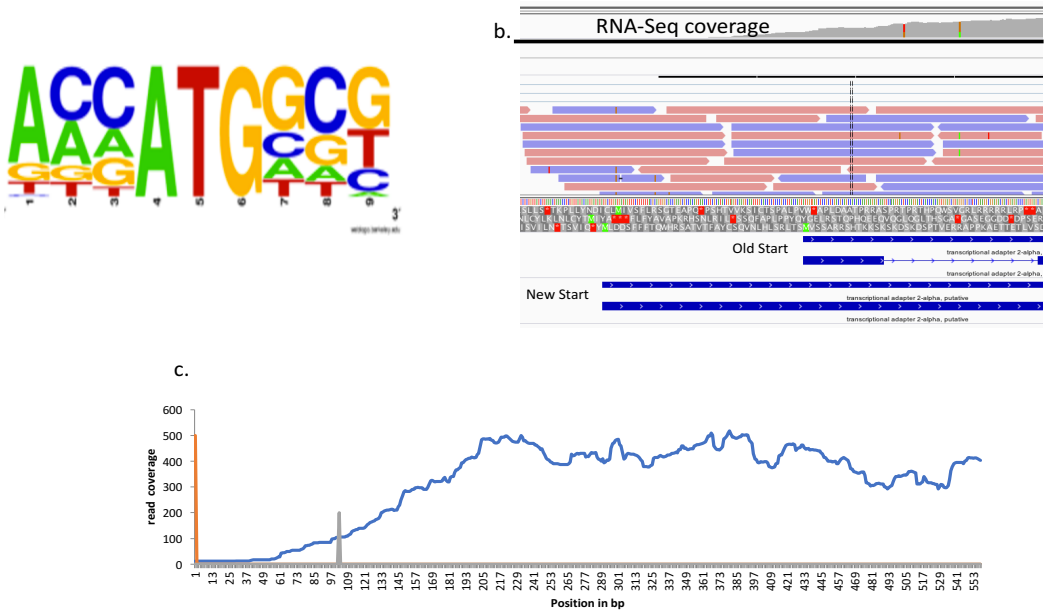
V. Secretome identification: Secreted protein identification is an important part of genome annotation as *Phytophthora* uses these proteins also known as effector proteins to support its colonization in the host (Morgan and Kamoun, 2007; Bozkurt et al., 2012). We have identified 1,789 proteins that are secreted based on SignalP4.0, TargetP1.1 and tmhmm predictions (Supplementary table 6). Proteins predicted to be secreted using SignalP (D-score cut-off of 0.450) and TargetP1.1 (predicted location S) were checked against proteins with transmembrane helix prediction using tmhmm (these proteins have a transmembrane helix and are targeted for membrane) and only those proteins which were predicted to be secreted and lacked transmembrane helices were added to the secretome. Out of these 1789 proteins, 31 are novel genes that were previously not annotated in the genome and are now added in the current annotation. We were also able to compare the current secretome to a previously published study of *P. infestans* which predicted 1415 secreted proteins (Raffaele et al., 2010). We found that 1,246 out of 1,415 proteins were common to our analysis and the previous analysis (Raffaele et al., 2010).

155 proteins that were missing in current secretome compared to the Raffaele et al study were further analyzed for the reason as to why they are not predicted to be secreted (Supplementary table 7). These were checked for changes in their gene models and especially at their N-termini. 56 genes with original gene models were not identified as secreted by SignalP4.0 (D-cutoff -.450). We found 169 genes that have been modified by our pipeline. Out of 169 genes 111 had been modified at their N-termini, resulting in loss of signal peptide, 7 were modified at their C-termini and now have transmembrane helices and thus are not predicted to be secreted, 8 genes were split and only the 5' end of the gene showed signal peptide, 5 genes have isoforms, 7 are merged, 6 have same N-terminus but are not predicted to be secreted by SignalP using the D-cutoff of 0.450, and 20 genes did not have enough reads to confirm the model.

These results caused us to scrutinize the models of all 3,413 genes where the N-terminus of the protein had been changed. The goal was to find if there was enough evidence in terms of Kozak sequence, RNA-Seq read coverage and ortholog search data to confirm or reject the new N-terminus (Figure 11). The translational start codon position of the gene with both old and the new start was weighed against the position weighted matrix of Kozak consensus sequence obtained from 50 ribosomal proteins, which are presumably efficiently translated. Genes that showed higher scoring of translational start site with old gene model were then checked for coverage using RNA-Seq data and correctness of the gene model by visualizing the alignments in IGV. Using the above criteria, we changed the position of the start codon for 599 genes, which identified an additional 61 as encoding secreted proteins.

Figure 11: Correction of start codon based on Kozak consensus sequence scoring of new and old translational start, IGV to confirm the gene model and read coverage. a. Kozak consensus sequence that was used to score the start codon based on old and new translational start b. IGV snapshot of the gene with RNA-Seq coverage at the top. The original model is called 'Old' while the *PASA*-predicted model is called 'New'. c. Coverage plot of the same gene with orange predicting the new start and gray representing the old start.

Figure 11


b.
RNA-Seq coverage
Old Start
New Start
transcriptional adapter 2-alpha,
transcriptional adapter 2-alpha, putative
c.
read coverage
Position in bp

We have also identified 487 proteins that were previously not identified as being secreted. We checked these proteins for the presence of the RxLR motif that is present after the signal peptide cleavage site and a hmm for identifying the RxLR motif (Hass et al., 2009). We found an additional 6 putative RxLRs that satisfy the above criteria; 5 of these genes were previously predicted as not secreted and one gene (PITG_25242) is the novel gene that was previously not annotated in the genome. 5/6 of these RxLRs are single exon genes with their length varying from 300 to 700 amino acids.

VI. Conservation of RxLRs in different isolates of P. infestans: We have seen that there are differences between two different strains 1306 and 88069, which complicated the incorporation of *PASA* assemblies to update gene structures. Differences in strains (1306 and 88069) was seen even with genes that were present in gene dense region, which is evolving at a much slower rate than loci in gene-sparse regions where most of the RxLRs are present (Haas et al., 2009). We wanted to see as to how many of the RxLRs are conserved in different isolates in *P. infestans* or how many RxLRs are expressed in different strains. Genomic sequences from 10 different strains were checked for conservation; these included 550, 618, 1114, 1306, 6629, US8, US11, US22, US23 and US24. We found that out of five hundred fifty-eight RxLRs currently annotated, thirty-four genes were completely absent in some strains (Suppl table 9), while one hundred and twenty-nine genes had missing reads in certain regions of the gene so no reliability can be placed in their conservation. Three hundred and ninety-four genes that had genomic reads

spanning the entire gene length were checked for conservation and fourty-five genes were conserved at protein levels (Supplementary table 8).

We also checked if the conserved RxLRs were expressed in three isolates, using RNA-Seq data from isolates 618, 1306 and 88069 (Supplementary table 9). Out of the 45 genes that were conserved at protein levels in different isolates, 20 genes had <1x coverage, 2 genes (PITG_04164 and PITG_04099) had good coverage in 1306 but <1x coverage in 88069. Only 8 genes had enough coverage in 618; the low coverage in 618 may be attributed to the fact that we had very few samples of 618 and so the data may be biased.

Discussion: Correct gene annotations play a major role in both bioinformatics and molecular biology. Incorrect gene models can affect functional annotation, ortholog identification, promoter analysis, predictions for secreted proteins, tertiary protein structure, expression calling and so forth. In the present study, we have rigorously re-annotated the genome of *P. infestans*.

We checked the if the percent of reads mapped to transcriptome are better with new or old gene models. Tophat (v2.1.1) was run on both the new and old gff files on protein coding genes with the option of exome only to find the percent of reads explained by the exons. Tophat alignment results were 67.8% overall mapping rate to the transcriptome with the new GFF file as compared to 57.2% alignment rate with old GFF file, showing that more reads matched the new gene models as compared to the old one.

We have added UTRs to 7567 genes. UTRs are important in many aspects – translational efficiency, mRNA stability and mRNA expression (Hinnebusch et al., 2017). However, correct identification of UTR boundaries was a challenge in gene-dense regions as RNA-Seq reads from adjacent genes often overlapped, giving unusually long UTRs. We used strand-specific data to trim some of these UTRs but since we had limited strand-specific data, many of these UTRs may still be incorrectly predicted.

Correct identification of UTRs in *P. infestans* opens many possibilities for studying aspects of transcriptional and translational machinery such as uORFs, introns in UTRs, and motif identification, to name a few. Translational ability is affected by many different aspects of UTR that include the presence/absence of uORFs in the 5' region (Hinnebusch, 2011), length of UTR (Toma et al., 2015; Kervestin and Jacobson, 2012),

regulatory motifs and the position of kozak sequence (Kozak, 1986). Presence of uORFs may result in halting of the translation for the main ORF and reduce the translational efficiency. Apart from the presence of uORFs, motifs present in 5' and 3' regions effect mRNA stability and translational efficiency; in animals, AU-rich elements are present in 3'UTRs of genes encoding for transcription factors and cytokines and play an important role in mRNA stability (Chen and Shyu, 1995). TOP (Terminal oligopyrimidine tracts) is a common motif found in 5'UTR of mammalian cells and is involved in both transcription and translation of mRNA (Parry et al., 2010; Pichon et al., 2012). Long 3' UTRs are resistant to nonsense-mediated mRNA decay (NMD) and cis-regulatory element was found in the first 200 nt from the stop codon that suppressed NMD in human mRNAs (Toma et al., 2015). On the other hand, long 5' UTRs have been associated with inhibition of translation especially in cases where uORFs or secondary structures are present (Davuluri et al., 2000).

We also have modified the CDS boundaries of 6612 genes. This included changing the 5' and 3' of proteins, removal or addition of introns, gene merging and alternate splicing. These modifications are of great significance as correct N-termini prediction is required for accurate analyses of promoters including transcription factor binding sites (TFBS), as well as signal peptides used to determine if the protein is targeted for secretion or mitochondria. Thus, the refined gene models might allow the identification of more cis-regulatory elements, or the identification of secreted effector proteins. The latter play an important role in *P. infestans*, as they work to block plant immunity by targeting proteases, glucanases and genes related to cell death in cell

apoplast (Tian et al., 2005; Kamoun, 2003). Cytoplasmic effectors like RxLRs and crinklers have conserved RxLR motif and LFLAK domain respectively (Morgan and Kamoun, 2007; Torto et al., 2003). Both the RxLR motif and the LFLAK domain are present within the first 60 amino acids on the N-terminus. The C-terminus is important for the activity of many effectors inside the plant cells (Bos et al., 2006; Dou et al., 2008); transmembrane helices, which are present at the C-terminal can change the localization of proteins from being secreted to membrane proteins. In the present study, we found that many of the effector proteins that were previously predicted to be secretory (Raffaele et al., 2010) are not secreted since the extended C-terminus included transmembrane helices.

Presence of introns in humans is positively correlated to mRNA stability, with a strong relationship between functional annotation and mRNA turnover (Wang et al., 2002). Apart from functional annotation, when genes with similar functional annotation were clustered and half-life of mRNAs was assessed, mRNA from genes with no introns was less stable when compared to genes with introns (Wang et al., 2007). Further studies with mRNA stability and half-life can be correlated to see if higher intron number leads to longer half-life of mRNAs in *P. infestans*.

Alternative splicing is an important mechanism through which the organisms can increase the complexity of its genome. Introns play an important role in regulating genes at various stages including but not prohibited to mRNA stability, accumulation, and transcript processing and expression (Lu et al., 2008; Samadder et al., 2008). We have identified 1603 genes with alternative splicing events. Many of these alternatively spliced

variants have stage-specific expression patterns and many have variation at the N-terminus leading to the formation of both secreted and non-secreted proteins. Alternative splicing of a multi-drug transporter in *Pseudoperonospora cubensis* results in two isoforms – one which encodes a RxLR protein while the other encodes for a transporter (Savory et al., 2012). Future studies using RT-PCR can confirm if the secreted forms of the P. infestans genes are expressed earlier in the infection cycle (2-4dpi) whereas the non-secreted form is expressed later during sporulation or zoo-sporogenesis.

Methods:

RNA-Seq data: RNA-Seq reads were generated from two strains, 1306 and 88069. 1306 reads came from the following conditions: potato tubers and plant leaves infected by 1306 and collected at 3 and 6 days post infection, 1306 cultures in different media conditions (rye, Xu and casein supplemented Xu media) and different developmental stages (mycelia, sporangia, cleaving sporangia, cysts and germinating cysts). 88069 reads also came from all the developmental stages specified above. All these reads were paired-end without strand specificity.

Apart from the majority of reads that were used to modify the gene structures, at the end of project a small number 1306 strand-specific reads were also used under high and low humidity conditions.

Gene structure improvement: 88069 and 1306 reads were combined, genome-guided *trinity* pipeline was run with default parameters, and the transcripts were supplied to *PASA*. *PASA* was run with settings of 95% identity along 90% transcript length to incorporate the *PASA* assemblies in updating the gene models. This stringent cut-off was used to ensure correct mapping of reads in multi-gene families.

However, since RNA-Seq reads were generated from two different strains, 1306 and 88069, and were aligned to the reference genome of T30-4, many of the *PASA* assemblies were not incorporated to modify gene structures due to mismatches present between *PASA* assemblies and reference genome. To overcome the problem of strain differences that resulted in mismatches between *PASA* assemblies and reference genome,

trinity and *PASA* pipelines were run separately with the 88069 and 1306 reads. Gene structure modifications resulting in different protein sequence and modified commonly by 88069, 1306 and combined reads were manually confirmed using IGV 2.3.3 (Figure 2).

Transdecoder was run on *PASA* assemblies generated by combined (1306+88069) reads to find open reading frames. These ORFs were then compared to existing mRNAs to find if they are in sync with existing gene models. Gene structures that did not match to predictions made by *transdecoder* were checked in IGV to confirm if the *transdecoder* predictions matched with alignments (Figure 3).

Samtools were used to find exons that have <1x coverage based on uniquely mapped reads. A total of 1,789 genes <1x coverage at the 5' and 3' end of gene. These genes were then checked in IGV to find if the 5' or the 3' end is supported and also if the predicted intron is supported by RNA-Seq. The presence/absence of introns was confirmed by comparing the ratio of uniquely mapped in introns vs exons; if the ratio was >=1 then the intron was predicted to be absent.

Identifying problem genes: The existing genome annotation was checked for genes that might still be wrong due to lack of coverage, genome assembly and multi-gene families. *Samtools* (v1.3) was used to calculate the number of reads aligned for each gene, the genes which have <1x coverage were reported (Supplementary table 1).

Existing gene mRNAs were extracted using *bedtools getfasta* (v2.26.0) and checked if there is missing assembly information specified by Ns in the genome. Genes with Ns in middle, end and start are listed in Supplementary table 3.

Defective alleles: 1306 RNA-Seq reads were aligned to the reference genome of T30-4 using *gsnap*. *GATK* (v3.6) and *picard* tools (v2.6.0) were run on the 1306 bam file with option of *–Allow_N_Cigar_reads* that allows RNA-Seq reads present at exon-intron junction. SNPs/INDELs were then annotated using Ensembl Variant annotation package.

To confirm the variants, 500bp upstream and downstream regions of the selected variants were extracted to design primers and these were confirmed using Sanger sequencing.

New genes and ncRNA predictions: Transcriptionally active regions were compared to exiting genome predictions to find new genes that were currently not annotated. *Transdecoder* was run on *PASA* assemblies that did not have annotations in the current genome. The predicted ORFs were then checked for orthologs against *P. capsici, P. parasitica, P. sojae* and *P. ramorum* with e-value cutoff of 1e-20. The ORFs were then checked for a proper start/end codon and blast hits against known transposable elements in *P. infestans*.

PASA assemblies that did not map to existing genome annotation and lacked an open reading frame were deemed to be the ncRNA. These ncRNAs were checked for coverage using *samtools*. *Pasa* assemblies with <5x coverage or missing genomic

assembly were discarded. Blastable database of mitochondrial DNA, transposable elements and currently annotated genes was generated. The refined set was blasted against this database with e-value cutoff 1e-10. Putative ncRNAs with hits in transposable elements, mtDNA and known genes were discarded.

Correct identification of N-terminus: N-terminus was corrected based on three different conditions – ortholog matching, kozak sequence analysis and coverage plots.

A gff file with the new and old start position was generated and fasta sequences based on gff file were extracted. These sequences were then scored against the PWM of Kozak sequence generated based on 50 ribosomal transcripts. Coverage plots of genes that showed a better score with old translational start were generated based on both old and new gene structures. The correctness of gene structure based on old gene model and kozak score was also confirmed by visualization in IGV (Figure 11).

Proteins were checked for similarity against other *Phytophthora* species both at the protein level and genome level and the position of the protein which has the blast hit was recorded. Proteins showing variation of >=20 amino acids outside the signal peptide cleavage site were further analyzed using coverage plots and IGV.

Trimming of UTRs: Genes with 5' or 3' UTR length greater than 100bp were checked for correct transcriptional start. Strand-specific RNA-Seq reads were mapped to T30-4 genome using *gsnap*. The resulting bam file was split into forward and reverse bam files based on strand specificity. The predicted UTRs were checked with two different

criteria: the position of INR-FPR motif and coverage within the UTR based on strand-specific bam files (Figure 6).

Genomic sequences of the 5' UTRs and 100 bp upstream were generated based on the current GFF file. These sequences were checked for the presence of INR-FPR motif within the UTR sequence and upstream UTR sequences. Genes that were found to have INR-FPR motif within the UTR sequences were subsequently trimmed at that position.

Genes with no INR-FPR motif were checked against the positive and negative bam files for coverage based on unique reads (Figure 6) and were trimmed where the unique number of reads was zero.

Functional annotation: Fasta file of the existing proteome was checked for homology against Swiss-prot and PANTHER databases with an e-value cut-off of 1e-10. The best blast hit, along with associated function and GO term of that protein and e-value of Blast hit from both Swiss-prot and PANTHER were then used to give the final functional annotation to the new proteome (Supplementary table 5). No entry was given to proteins that had no hits in either Swiss-prot or PANTHER. The final functional annotation was given based on best blast hit from PANTHER. In cases where the protein name from PANTHER was not available or the name was not informative, Swiss-Prot annotation was given as final annotation.

Secretome identification: The predicted proteins were submitted to 3 different analysis to identify the secretome of *P. infestans*. *SignalP* (v4.1) was used to identify the proteins

with a signal peptide. *TargetP*(v1.1) was used to identify the location where the protein was targeted (mitochondria, secreted or other). Finally, *tmhmm* (v2.0) was used to predict regions with transmembrane helices. Commonly predicted secreted transcripts from *signalP* and *targetP* were further analyzed for transmembrane helices. Proteins with transmembrane helices only in the signal peptide cleavage region were kept in the final secretome (Supplementary table 6).

The proteins that were previously not identified as secreted were also checked for the presence of RxLR motif that is present in many effectors after the signal peptide cleavage site and between position 30-60 amino acids (Haas et al. ,2009).

Supplementary tables information: Attached as supplementary files to thesis.

Supplementary table 1: Genes which have less than 1x coverage in different developmental stages (hyphae, mating, sporangia, zoospores, cyst and germinating cyst) of 1306 and 88069.

Supplementary table 2: Previously protein coding genes that based on RNA-Seq data are now annotated as non-coding RNAs.

Supplementary table 3: Genes that might have incorrect gene models due to gaps in T30-4 assembly. Gaps can be in the start or end or in the middle of the gene and their location in terms of start/end or middle is given.

Supplementary table 4: Defective alleles found in 1306 in comparison to T30-4. The table incorporates the gene name, supercontig#, SNP position, reduction in size of the protein, functional annotation of the gene and the confidence score of variant called.

Supplementary table 5: Functional annotation of all the transcripts present in the current genome annotation based on best blast hits in Panther and Swiss-Prot. The table include the gene name, gene model history in terms of revised or original model, final annotation given, best blast hits in Swiss-Prot and Panther along with GO terms associated with best-blast hits.

Supplementary table 6: Secretome identification of current genome based on *SignalP*, *TargetP* and *tmhmm* predictions.

Supplementary table 7: Analysis of the genes that were predicted to be secreted in previous analysis by Rafaele et al., 2010 but are absent in the newly predicted secretome.

Supplementary table 8: RxLR conserved in 10 different isolates of *P. infestans*

Supplementary table 9: RNA-Seq coverage analysis of conserved RxLRs in 1306, 618 and 88069.

References:

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., ... & Horsman, D. E. (2009). De novo transcriptome assembly with ABySS. Bioinformatics, 25, 2872-2877.

Bos, J. I., Kanneganti, T. D., Young, C., Cakir, C., Huitema, E., Win, J., ... & Kamoun, S. (2006). The C-terminal half of *Phytophthora infestans* RXLR effector *AVR3a* is sufficient to trigger R3a-mediated hypersensitivity and suppress INF1-induced cell death in *Nicotiana benthamiana*. The Plant Journal, 48, 165-176.

Bozkurt, T. O., Schornack, S., Banfield, M. J., & Kamoun, S. (2012). Oomycetes, effectors, and all that jazz. Current Opinion in Plant Biology, 15, 483-492.

Cerqueira, G. C., Arnaud, M. B., Inglis, D. O., Skrzypek, M. S., Binkley, G., Simison, M., ... & Wymore, F. (2014). The *Aspergillus* Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. Nucleic Acids Research, 42, D705-D710.

Chen, M., Hu, Y., Liu, J., Wu, Q., Zhang, C., Yu, J., ... & Wu, J. (2015). Improvement of genome assembly completeness and identification of novel full-length protein-coding genes by RNA-seq in the giant panda genome. Scientific Reports, 5, 18019.

Chen, C. Y. A., & Shyu, A. B. (1995). AU-rich elements: characterization and importance in mRNA degradation. Trends in Biochemical Sciences, 20, 465-470.

Costanzo, S., Ospina-Giraldo, M. D., Deahl, K. L., Baker, C. J., & Jones, R. W. (2007). Alternate intron processing of family 5 endoglucanase transcripts from the genus *Phytophthora*. Current Genetics, 52, 115-123.

Davuluri, R. V., Suzuki, Y., Sugano, S., & Zhang, M. Q. (2000). CART classification of human 5'UTR sequences. Genome Research, 10, 1807-1816.

Dou, D., Kale, S. D., Wang, X., Chen, Y., Wang, Q., Wang, X., ... & McDowell, J. M. (2008). Conserved C-terminal motifs required for avirulence and suppression of cell death by *Phytophthora sojae* effector *Avr1b*. The Plant Cell, 20, 1118-1133.

Haas, B.J., Delcher, A.L., Mount, S.M., Wortman, J.R., Smith Jr, R.K., Jr., Hannick, L.I., Maiti, R., Ronning, C.M., Rusch, D.B., Town, C.D. et al. (2003). Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Research, 31, 5654-5666.

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., ... & MacManes, M. D. (2013). De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. Nature Protocols, 8, 1494-1512.

Jiang, R. H., Tripathy, S., Govers, F., & Tyler, B. M. (2008). RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. Proceedings of the National Academy of Sciences of United States of America, 105, 4874-4879.

Kamoun, S. (2003). Molecular genetics of pathogenic oomycetes. Eukaryotic Cell, *2*, 191-199.

Kervestin, S., & Jacobson, A. (2012). NMD: a multifaceted response to premature translational termination. Nature Reviews Molecular Cell Biology, 13, 700-712.

Korf, I. (2004). Gene finding in novel genomes. BMC Bioinformatics, 5, 59.

García-Bayona, L., Garavito, M. F., Lozano, G. L., Vasquez, J. J., Myers, K., Fry, W. E., ... & Restrepo, S. (2014). De novo pyrimidine biosynthesis in the oomycete plant pathogen *Phytophthora infestans*. Gene, 537, 312-321.

Goodwin, S. B. (1997). The population genetics of *Phytophthora*. Phytopathology, 87, 462-473.

Hassan, M. A., Melo, M. B., Haas, B., Jensen, K. D. C., & Saeij, J. P. J. (2012). De novo reconstruction of the *Toxoplasma gondii* transcriptome improves on the current genome annotation and reveals alternatively spliced transcripts and putative long non-coding RNAs. BMC Genomics, 13, 696.

Haverkort, A. J., Struik, P. C., Visser, R. G. F., & Jacobsen, E. (2009). Applied biotechnology to combat late blight in potato caused by *Phytophthora infestans*. Potato Research, 52, 249-264.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., ... & Zhang, Z. (2010). The sequence and de novo assembly of the giant panda genome. Nature, 463, 311-317.

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., ... & Li, S. (2010). De novo assembly of human genomes with massively parallel short read sequencing. Genome Research, 20, 265-272.

Lu, J., Sivamani, E., Azhakanandam, K., Samadder, P., Li, X., & Qu, R. (2008). Gene expression enhancement mediated by the 5′ UTR intron of the rice *rubi3* gene varied

remarkably among tissues in transgenic rice plants. Molecular Genetics and Genomics, 279, 563-572.

Mathé, C., Sagot, M. F., Schiex, T., & Rouzé, P. (2002). Current methods of gene prediction, their strengths and weaknesses. Nucleic Acids Research, 30, 4103-4117.

Morgan, W., & Kamoun, S. (2007). RXLR effectors of plant pathogenic oomycetes. Current Opinion in Microbiology, 10, 332-338.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., & Snyder, M. (2008). The transcriptional landscape of the yeast genome defined by RNA sequencing. Science, 320, 1344-1349.

Nielsen, R., Paul, J. S., Albrechtsen, A., & Song, Y. S. (2011). Genotype and SNP calling from next-generation sequencing data. Nature Reviews Genetics, 12, 443–451.

Ospina-Giraldo, M. D., McWalters, J., & Seyer, L. (2010). Structural and functional profile of the carbohydrate esterase gene complement in *Phytophthora infestans*. Current Genetics, 56, 495-506.

Parry T. J., Theisen J. W., Hsu J. Y., Wang Y. L., Corcoran D. L., Eustice M., Ohler U., & Kadonaga J. T. (2010). The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. Genes and Development, 24, 2013-2018.

Pichon, X., A Wilson, L., Stoneley, M., Bastide, A., A King, H., Somers, J., & E Willis, A. (2012). RNA binding protein/RNA element interactions and the control of translation. Current Protein and Peptide Science, 13, 294-304.

Roberts, A., Pimentel, H., Trapnell, C., & Pachter, L. (2011). Identification of novel transcripts in annotated genomes using RNA-Seq. Bioinformatics, 27, 2325-2329.

Rose A.B., Emani S., Bradnam K., & Korf I. (2011). Evidence for a DNA-based mechanism of intron-mediated enhancement. Frontiers in Plant Science, 2, 98.

Samadder, P., Sivamani, E., Lu, J., Li, X., & Qu, R. (2008). Transcriptional and post-transcriptional enhancement of gene expression by the 5′ UTR intron of rice *rubi3* gene in transgenic rice cells. Molecular Genetics and Genomics, 279, 429-439.

Savory, E. A., Zou, C., Adhikari, B. N., Hamilton, J. P., Buell, C. R., Shiu, S. H., & Day, B. (2012). Alternative splicing of a multi-drug transporter from *Pseudoperonospora cubensis* generates an RXLR effector protein that elicits a rapid cell death. PLoS One, 7, e34701.

Singh, R., Lawal, H. M., Schilde, C., Glöckner, G., Barton, G. J., Schaap, P., & Cole, C. (2017). Improved annotation with de novo transcriptome assembly in four social amoeba species. BMC Genomics, 18, 120.

Sniegowski, P. D., Gerrish, P. J., & Lenski, R. E. (1997). Evolution of high mutation rates in experimental populations of *E. coli*. Nature, 387, 703.

Testa, A. C., Hane, J. K., Ellwood, S. R., & Oliver, R. P. (2015). CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics, 16, 170.

Tian M., Benedetti B., & Kamoun S. (2005). A second Kazal-like protease inhibitor from *Phytophthora infestans* inhibits and interacts with the apoplastic pathogenesis-related protease *P69B* of tomato. Plant Physiology, 138, 1785-1793.

Tisserant, E., Da Silva, C., Kohler, A., Morin, E., Wincker, P., & Martin, F. (2011). Deep RNA sequencing improved the structural annotation of the *Tuber melanosporum* transcriptome. New Phytologist, 189, 883-891.

Torto, T. A., Li, S., Styer, A., Huitema, E., Testa, A., Gow, N. A. R., & Kamoun, S. (2003). Est mining and functional expression assays identify extracellular effector proteins from the plant pathogen *Phytophthora*. Genome Research, 13, 1675-1685.

Toma, K. G., Rebbapragada, I., Durand, S., & Lykke-Andersen, J. (2015). Identification of elements in human long 3'UTRs that inhibit nonsense-mediated decay. RNA, 21, 887-897.

Trapnell, C., Pachter, L., & Salzberg, S. L. (2009). TopHat: discovering splice junctions with RNA-Seq. Bioinformatics, 25, 1105-1111.

Van der Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.], 11, 11.10.1–11.10.33.

Van Der Velden, A. W., & Thomas, A. A. (1999). The role of the 5' untranslated region of an mRNA in translation regulation during development. The International Journal of Biochemistry & Cell Biology, 31, 87-106.

Wang, Y., Liu, C. L., Storey, J. D., Tibshirani, R. J., Herschlag, D., & Brown, P. O. (2002). Precision and functional specificity in mRNA decay. Proceedings of the National Academy of Sciences of United States of America, 99, 5860-5865.

Wu, T. D., & Nacu, S. (2010). Fast and SNP-tolerant detection of complex variants and splicing in short reads. Bioinformatics, 26, 873-881.

Wu, X., Ren, C., Joshi, T., Vuong, T., Xu, D., & Nguyen, H. T. (2010). SNP discovery by high-throughput sequencing in soybean. BMC Genomics, 11, 469.

Ye, W., Wang, X., Tao, K., Lu, Y., Dai, T., Dong, S., ... & Wang, Y. (2011). Digital gene expression profiling of the *Phytophthora sojae* transcriptome. Molecular Plant-Microbe Interactions, 24, 1530-1539.

Chapter III

Identifying and annotating structural variation in different lineages of *Phytopthora infestans*

Abstract: In this study, we present a systematic approach to identify structural variation present in different lineages of *P. infestans* and which might play a role in phenotypic differences seen in these lineages. Illumina Hi-Seq paired end reads from 10 geographically and phenotypically diverse strains were aligned to the reference genome of 1306 and were used to identify structural variants. We have integrated the results from different structural variant callers and custom filtering to identify a robust set of structural variants. We have identified a total 782 deletion and 508 inversion events that are ≥1kb in 9 different lineages when compared to 1306 strain. Of the 782 deletion and 508 inversion events 343 deletion and 345 inversion events have genes associated with them. Apart from large structural variants, we have also identified small INDELs that are defective alleles in certain lineages. Further characterization of these variants in different lineages may throw light at differences observed between the lineages in terms of mating type, fungicide sensitivity or host preference.

Background: Structural variation including insertions, deletions, translocations and inversions play an important role in genetic variation and have a major influence on phenotypic diversity (Feuk et al., 2006; Mills et al., 2011). Structural variation has been studied both in animal, plants and pathogens; however, the number of studies in humans outweigh other model systems. Structural variation has been widely studied in humans showing that many more base pairs are altered with structural variations as compared to SNPs (Zhang et al., 2009). Many diseases are now known to be linked to structural variation include cancer, autism, Rett syndrome, Alzheimers disease and Parkinson disease (Futreal et al., 2004; Mitelman et al., 2007; Marshall et al., 2008; Szatmari et al., 2007; Rovelet-Lecrux and Campion, 2012; Singleton et al., 2003). Since the reduction in costs in DNA sequencing, many other studies have looked in structural variations in human pathogens. These include but not limited to *Mycobacterium tuberculosis*, *E. coli, Bacillus anthracis, Yesenia pestis* and *Clostridium butyricum* (Coll et al., 2014; Rasko et al., 2008; Read et al., 2002; Welch et al., 2007; Hill et al., 2009). *Mycobacterium tuberculosis* has a SNP/INDEL-based and large structural variant database comparing different isolates, which includes 800 large structural variations (Coll et al., 2014; Coscolla and Gagneux, 2014). Study on pangenome of *E. coli* commensals and pathogenic isolates showed lot of genetic variation linked to virulence of strains (Rasko et al., 2008). Structural variation studies have been done in inbred plant species such as maize, wheat and soybean. Structural variation studies in maize showed that 1,783 SVs affecting 1,270 genes were present in one inbred line and were involved in domestication (Springer et al., 2009). In opium, a cluster of 10 genes was tightly linked to the noscapine

phenotype and was absent in all the non-noscapine lines (Winzer et al., 2012). In wheat, an indel present 50 bp upstream of *Ppd-1* gene (Photoperiod sensitivity gene) influenced flowering time (Nishida et al., 2013), while in soybean structural variants involved gene-families associated with nucleotide-binding and receptor-like protein classes that are important for biotic stress and plant immunity (McHale et al., 2012). Several studies on plant pathogens have shown deletions in avirulence (Avr) genes help pathogen in escaping the identification from corresponding resistance genes in plants (Hartmann et al., 2017; Wollenberg and Schirawski, 2014; González et al., 2011; Gout et al., 2007). In plant pathogen *Leptosphaeria maculans,* which causes stem canker on oilseed rape, deletions of Avr loci were found in >90% of virulent strains and the breakpoints of deletion were conserved among different virulent strains (Gout et al., 2007). Deletion of small secreted gene in populations of *Zymoseptoria tritici* (plant pathogen causing major losses to wheat in Europe) led to a gain of virulence (Hartmann et al., 2017).

In the present study, we sought to find structural variations present in a very important plant pathogen *Phytophthora infestans*, which causes late blight in *Solanaceous* species and results in losses in billions of dollars (Haverkort et al., 2008). *P. infestans* first got attention for the infamous Irish potato famine when the HERB-1 strain from the Americas was introduced in Ireland (Yoshida et al., 2013). US-1 strain was the cause of major epidemic on potato and tomato crops in early 1900s (Goodwin et al., 1994). Even after years of investigations, *P. infestans* still is a devastating pathogen as it has the capability to rapidly adapt to host resistance and can flourish really fast in humid conditions. *P. infestans* can infect potato tubers, tomato fruits and leaves of both potato

and tomato. *P. infestans* is a heterothallic oomycete with mostly diploid chromosomes and can reproduce both sexually and asexually given the mating types. Initially *P. infestans* was thought to be asexual; however, when the two mating types (A1 and A2) were discovered in Toluca valley in central Mexico, its status changed from an asexual to both a sexual and asexually reproducing organism. Sexual reproduction is responsible for genetic variation and recombination and asexual reproduction follows when many sporangia are produced. The most fit single individual usually dominates an entire area by producing thousands of sporangia that are dispersed aerially and causes widespread disease. Most of the sexual reproduction is thought to occur in central Mexico where both mating types A1 and A2 are found in 1:1 ratio and the populations are in Hardy-Weinberg equilibrium (Grünwald et al., 2001). Most of the older lineages in U.S. were thought to be of the A1 mating type, while the A2 mating type was introduced in 1970 (Fry et al., 1993; Fry et al., 1992). Even after the introduction of the A2 mating type in the USA, only a few instances of sexual recombination were reported (Danies et al., 2014). Asexual reproduction, which results in clonal lineages, are usually dominant in a certain region for some time until they are replaced by a more fit lineage. The interest in studying these lineages arose from the late blight epidemics in 2009 and 2010 when US-22 was found to be widespread on tomato plants. This led to many questions about the sensitivity of lineages to fungicides, preferences for certain hosts, mating type, and differences in their rates of germination (Fry et al., 2013; Danies et al., 2013). Many studies have shown ways to identify which clonal lineages are present in an area; some of these include genotyping-by-sequencing (Hansen et al., 2016), SNP-based analysis

(Hansen et al., 2016), restriction fragment length polymorphism (Goodwin et al., 1998), allozyme identification (Goodwin et al., 1995) and microsatellites (Danies et al., 2013; Cooke et al., 2012). The microsatellite markers have been used to distinguish different clonal lineages and to identify diversity within these lineages (Cooke et al., 2012; Li et al., 2012). Even though many studies have looked at the phenotypic differences present in different lineages during a certain period of time and in different areas (Ogoshi et al., 1994; Kato et al., 1997; Mizubuti and Fry, 1998; Perez et al., 2001), not many studies have looked at nucleotide level variation or large structural variations that exist between these lineages. One way to understand the phenotypic variations present in different isolates is to look at the variations which cause large insertions and deletions, since such studies might lead to manipulations of genes involved in immunity to host resistance genes, or identify genes associated with fungicide resistance, mating type, etc. Also important are small INDELs that can cause premature ending or frameshift mutations leading to different secondary and tertiary structure of proteins. The insights to the genome of different lineages can help us to make better decisions in terms of disease management.

We have aimed to identify large and small structural variations that are present in 10 different isolates. The assembly used in this study was recently assembled using PacBio and Mi-Seq reads from 1306 strain and has not been annotated. In the present study, we have used genetically and geographically diverse strains. This is the first study that has tried to identify these differences and then annotate the variants identified. We identified 782 deletions and 508 inversions that are larger than 1 kb, 194 deletions, 5

insertions and 74 inversions that are between 30 bp to 1 kb and 58,880 deletions and 22,348 insertions that are less than 30 bp. We have identified the variants and associated genes in different strains of *P. infestans*.

Results:

Strains used in the present study: We have analyzed 10 different strains that are both sensitive and resistant to metalaxyl and mefenoxam, have different mating types and host preference (Table 1).

Table 1: Phenotypic characteristics of the strains used in the study.

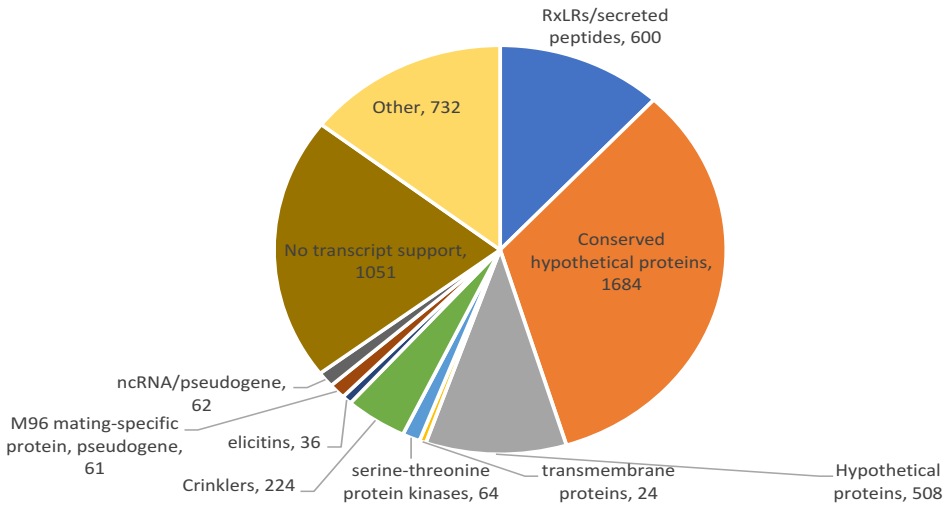
Strain	Metalaxyl sensitivity	Host preference	Mating type	Source
1306	sensitive	tomato	A1	USA
550	sensitive	potato	A2	Mexico
618	insensitive	not tested	A2	Mexico
1114	insensitive	not tested	A1	Netherlands
6629	sensitive	not tested	A1	Mexico
US-8	insensitive	potato	A2	USA
US-11	insensitive	potato/tomato	A1	USA
US-22	sensitive	tomato	A2	USA
US-23	sensitive	tomato	A1	USA
US-24	sensitive	potato	A1	USA

New assembly annotation: Our lab recently sequenced 1306 strain using PacBio Sequencing and generated a new assembly for 1306. The PacBio assembly was 220 Mb in size and had 1439 contigs with N50 size of 540 kb. The assembly was further modified using Illumina Mi-Seq data resulting in 1318 contigs and adding 800,000 bases to the assembly. Previously, another strain of *P. infestans*, T30-4, showed a complicated genome with estimated 74% repeat-rich region where many effector proteins are present (Haas et al., 2009). Annotation of the assembly was important to identify genes that might be involved in different structural variation. We ran *Maker* to identify genes in the 1306 assembly with repeats masked with repeat masker facility, 1306 transcript assemblies from *Trinity* and *Cufflinks*, and the protein fasta file and gff files generated for T30-4 genome assembly (Chapter II). Proteins identified by *Maker* do not always start with methionine, so we checked the position of methionine in the sequence and if the length of protein from first methionine to last stop was ≥100 amino acids, we kept it as a protein-coding gene. After checking for the predicted protein length, we identified 19,040 transcripts belonging to 16,874 genes. These were checked for similarity to existing genes in other strain of *P. infestans* (T30-4). The genes from T30-4 assembly were named *Phytophthora infestans* theoretical genes (PITG) followed by five-digit number. We used both blastP and blastn to find homology between the genes predicted by *Maker* from 1306 and proteins from T30-4. BlastN results were much more reliable way to identify matches between the new genes and prior annotations. This allowed us to give the PITG# to the new *Maker* proteins in cases where the same PITG# was the top blast hit in both blastP and blastn results. Using the above strategy, we were able to assign PITG# to

8,859 proteins. We also found based on best matches (E-value <1e-20) that 68 genes in the new assembly had no match to any exiting genes in T30-4. These might be the putative genes that were specific to 1306.

We also wanted to know which of the T30-4 genes are missing in the new assembly. Analysis of blast results of the new *Maker* gene sequences as query and T30-4 genes as subject showed that 5,046 of the predicted genes in T30-4 have no matching genes in the new assembly. Out of 5046 genes that lacked matching genes in the new assembly, 1051 were found to have no transcript support from the previous study (Chapter 2; Supplementary table 1). We then did a functional classification of the remaining genes that were missing to find their functions. Some of the missing genes belong to the effector families that includes RxLRs, elicitins and crinklers. Figure 1 shows the distribution of the genes missing in current assembly. One logical explanation for the missing genes in 1306 is that by masking the repetitive DNA present in the genome, members of multigene families would have been missed; another possible explanation is that there are big structural differences between 1306 and T30-4, involving the deletion of regions harboring the genes. To overcome genes missing due to masking of the genome is to run *Maker* with no repeat-masking and that should give us all the proteins that are currently missing; structural variations between T30-4 and 1306 that led to missing genes can be addressed by running structural variant analysis with bam files from T30-4 and 1306.

Figure 1: Functional categorization of proteins from the T30-4 assembly that had no blast hits in 1306 assembly based on e-value 1e-20.

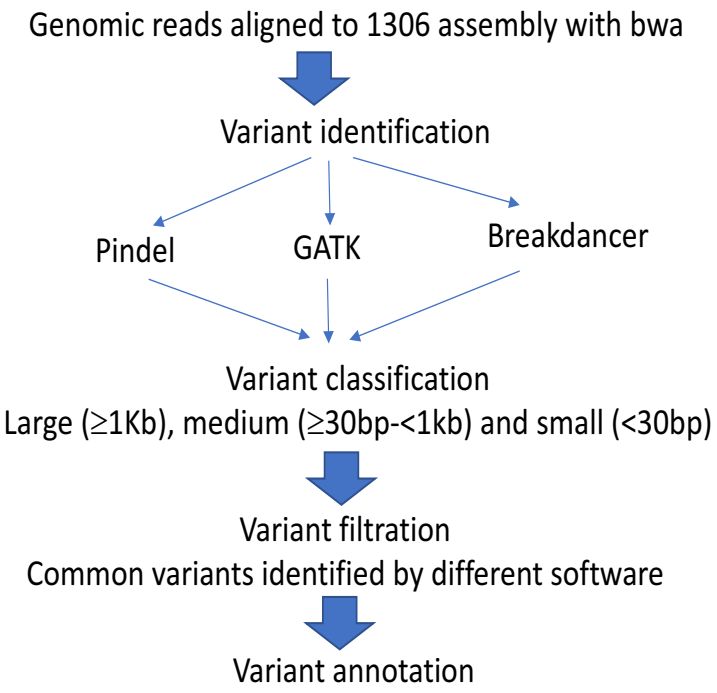


Variant identification: In the present study, we wanted to find structural variations present in different strains when compared to 1306. Illumina Hi-Seq paired-end 50 base-pair reads from 10 different isolates were aligned to the reference genome of 1306. We used three different software – *Pindel*, *BreakDancer* and *GATK* to identify SVs in these lineages that gave us a more robust dataset to work with as they all use different approaches to identify SVs (Ye et al., 2009; Chen et al., 2009; Auwera et al., 2013). Complete pipeline to identify and annotate variants is given in Figure 2.

Figure 2: Complete pipeline to identify and annotate the structural variants in lineages of *P. infestans*.

Reads were aligned to the 1306 assembly using *bwa*. Variant identification included identifying variants with *Pindel*, *BreakDancer* and *GATK* using bam files generated from the bwa alignment. Variant classification included classification of variants as either Large, medium or small. Variant filtration included checking for common predictions from both *Pindel*, *BreakDancer* and *GATK* and removing variants predicted for 1306. Finally, variant annotation involved finding genes present in regions with large deletions and inversions using custom perl scripts and annotating small variants as intergenic, exonic, intronic or splicing variants using *Annovar* variant package.

Figure 2


Genomic reads aligned to 1306 assembly with bwa
Variant identification
Pindel
GATK
Breakdancer
Variant classification
Large (≥1Kb), medium (≥30bp-<1kb) and small (<30bp)
Variant filtration
Common variants identified by different software
Variant annotation

Structural variant callers work on a greedy approach to call variants. For example, even if only 1 out of 100 reads shows the variation, it will be identified and reported as a variant. The best way to filter these kinds of structural variants identified by variant prediction software present in the dataset is to write custom scripts based on the coverage present for each isolate (Ye et al., 2009). We calculated the read coverage for each isolate (Table 2). Since the lowest coverage present was 20x, we decided to use a minimum number of 5x to confirm the variant. Apart from the coverage we intersected the results obtained by *Pindel* and *BreakDancer* to find common large and medium sized variants and *Pindel*, *BreakDancer* and *GATK* to identify small (<30bp) insertions and deletions. *Pindel* works on split-read and pattern-finding approaches to find large sized deletions and medium sized insertions, while *BreakDancer* uses a read-pair approach to find differences between the estimated mapping distance and actual mapping distance between the read-pair. *GATK* UnifiedGenotyper does not work on either split-read or read-pair algorithms but is based on a Bayesian genotype likelihood method to estimate the most likely genotype and allelic frequency.

Table 2: Total # of reads and fold coverage for each lineage used in the study

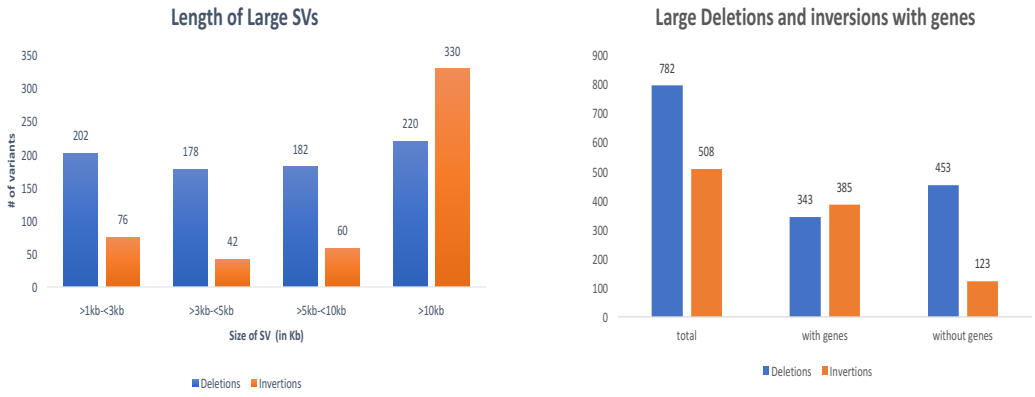
Lineage name	Total # of reads	fold coverage
550	49,786,417	22.63
618	168,897,317	70.37
1114	51,681,957	23.49
1306	119,184,479	49.66
6629	46,052,882	20.93
US-8	57,266,275	23.86
US-11	171,091,836	71.28
US-22	47,235,386	19.68
US-23	51,534,710	21.47
US-24	47,528,203	19.8

Pindel can identify the following kinds of SVs: insertions, deletions, inversions and tandem duplications. By comparison, *BreakDancer* can identify insertions, deletions, inversions and translocation and *GATK* can identify only small insertions and deletions. We focused only on insertions, deletions and inversions because they are identified by both *Pindel* and *BreakDancer*. We first identified all the possible deletions, insertions, and inversions by running *Pindel* and *BreakDancer* individually. Then, we filtered the data to focus on SVs present in both. Finally, the SVs that were detected by both *Pindel* and *BreakDancer* were checked against SVs detected in 1306 and only those SVs that were absent in 1306 were reported. A total of 782 large (≥1kb) deletion and 508 large inversion events were found commonly by *Pindel* and *BreakDancer*; of these, genes were associated with 343 and 385 deletion and inversion events respectively. Large SVs detected by both *Pindel* and *BreakDancer* for all the strains are summarized in Figure 3 and Table 3. Additional files 1 and 2 have details of deletion and inversion SVs found in each strain that includes breakpoint predictions by *Pindel* and *BreakDancer*, # of reads in the strain that supported the prediction, genes associated with SV (specifying the best blast hit of the protein in in the T30-4 assembly, functional annotation, and e-value and bit score of blast hit).

Table 3: Large structural variants identified by *Pindel* and *BreakDancer* in 9 isolates compared to 1306. INS stands for insertions, INV for inversions, and DEL for deletions. (BD) refers to SVs predicted by *BreakDancer*, (P) identified by *Pindel*, and (C) identified by both *Pindel* and *BreakDancer* which were then checked for the presence of genes.

Strain	INS(BD)	INS(P)	INS(C)	INV(BD)	INV(P)	INV(C)	DEL(BD)	DEL(P)	DEL(C)
1114	0	1075	0	1421	1121	16	499	2625	29
618	44	12443	18	563	2263	104	1318	4057	14
550	0	6170	0	1740	3569	335	2536	5485	504
6629	0	4929	0	1706	2321	136	2871	4972	126
US8	0	955	0	470	1181	23	725	2841	46
US11	9	14449	2	2462	1283	32	567	3654	37
US22	0	1014	0	537	1078	24	903	2850	102
US23	2	626	0	836	925	29	806	2383	56
US24	7	913	1	1137	1100	28	730	2757	60

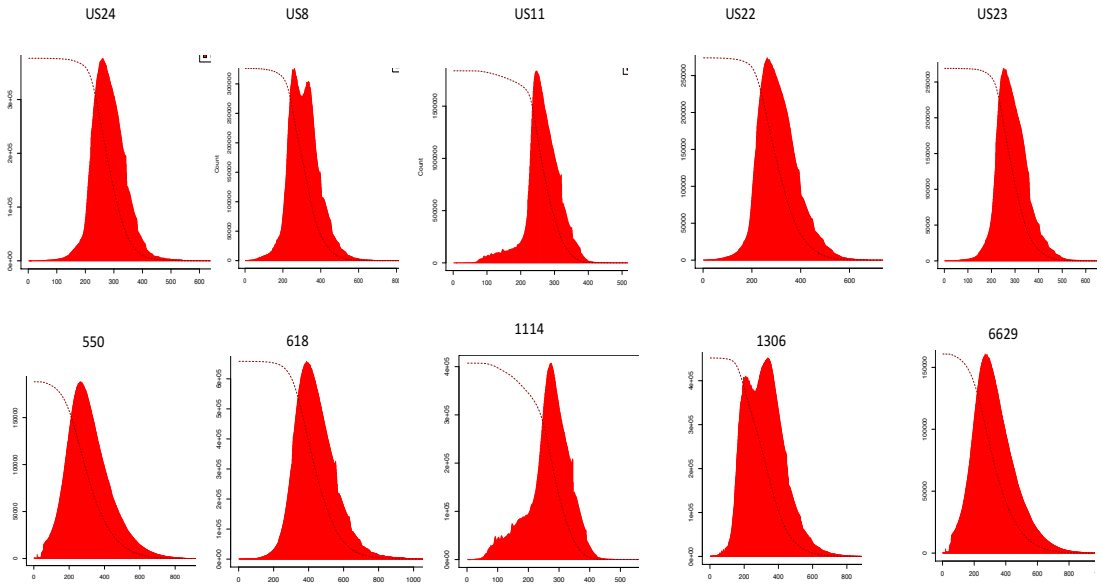
Figure 3: Total # of large (≥1Kb) deletions and inversions commonly predicted by *Pindel* and *BreakDancer*. (a) length of SVs and # of variants falling in each category (b) Deletion and Inversions SVs with genes involved.



One observation based on data from both *Pindel* and *BreakDancer* is that *Pindel* identified more structural variants as compared to *BreakDancer*. The only logical reason that can explain is that with *Pindel* we explicitly specify the insert-size present between read-pairs which can vary within each strain (Figure 4), whereas *BreakDancer* calculates the insert-sizes based on the bam files supplied to it.

As seen in figure 4, most of the read-pairs have an insert length of ~300bp but it clear from the figure that many read-pairs have insert length varying from 100bp to 800bp. While running *Pindel*, the exact insert length of read-pair is given. Since not all base pairs have the same insert size many more structural variants are identified by *Pindel*. However, in case of *BreakDancer* the insert size is calculated based on the bam files of the strains given and thus variants due to bad estimates of insert size are not reported.

Figure 4: Distribution of insert-sizes in 10 different strains used in the present study. The insert sizes of the reads aligned to the reference genome of 1306 for each strain. The x-axis specifies the insert length between read-pairs, y-axis specifies the number of reads falling in the category. Double peak in US-8 and 1306 indicates that there are two common insert-sizes present in the read-pairs of these strains – one is close to 200bp insert and one close to 400bp insert length.



The analysis was repeated with for mid-size variants, which we defined as belonging in the range of ≥30bp to <1kb (Table 4 and Additional files 3,4 and 5). Finally, for short

SVs we focused on insertions and deletion and only the ones that were predicted in

2 out of 3 software (Additional file 6).

We had to confirm large insertion SVs using genome browsers Savant and IGV because the breakpoints predicted by *Pindel* and *BreakDancer* were very different and so were hard to confirm. Even when the common large insertion SVs appear to be true, we could not confirm either the length or annotate the variants because we did not know how long is the insertion. One way around to find the length of long insertions is to take reads that are 200 bp upstream and downstream of the reads supporting the insertion and do a de-novo assembly of the region to find the actual length of the insertion.

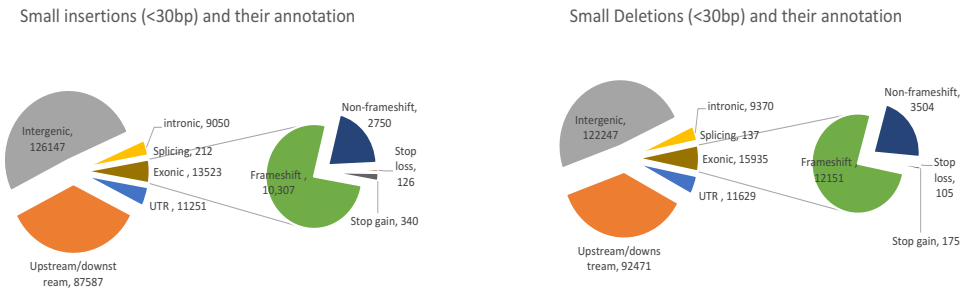
Table 4: Medium structural variants (30bp to 1kb) identified by *Pindel* and *BreakDancer* in 9 isolates compared to 1306. INS stand for insertions, INV for inversions, and DEL for Deletions. (BD) refers to SVs predicted by *BreakDancer*, (P) identified by *Pindel*, and (C) variants identified by both *Pindel* and *BreakDancer* and which were then annotated.

Strain	INS(BD)	INS(P)	INS(C)	INV(BD)	INV(P)	INV(C)	DEL(BD)	DEL(P)	DEL(C)
1114	0	11	0	3590	310	158	499	2049	288
618	181	71	0	213	722	115	1318	3129	177
550	0	682	0	1098	380	187	2536	2512	317
6629	0	657	0	1038	392	201	2871	10622	440
US8	4	14	0	245	348	62	722	2159	259
US11	2464	57	9	3600	475	219	820	2931	368
US22	11	12	0	217	312	80	811	2190	274
US23	359	17	1	457	264	102	879	1928	315
US24	952	16	0	1133	290	143	818	2077	319

Variant annotation: We wanted to identify if these variants have any genes that will help us to understand the differences that exist in different lineages. Perl scripts were used to identify if any large SVs had genes involved; the deletions and inversions SVs that affected genes are given in Figure 3. We were unable to confirm if any of the large insertions had genes for reasons mentioned above.

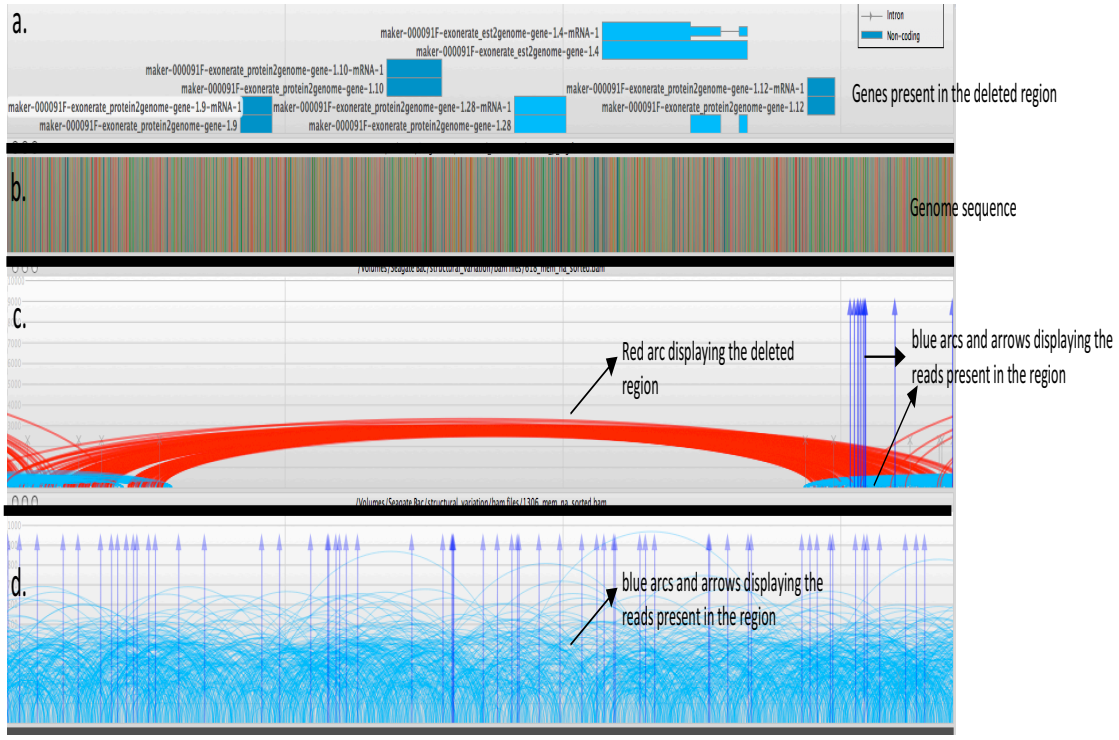
We used *Annovar* variant annotation package to annotate small indels found and were classified as intergenic, upstream/downstream, UTR, intronic, splicing and exonic variants. Variants were annotated based on their location within and outside the genes. Variants present outside the genes but within 1kb of genes were annotated as upstream/downstream variants while variants outside the 1Kb range of genes were annotated as intergenic. Based on gff file, *Annovar* can find protein coding sequences, UTRs, introns and exon-intron junctions. Exonic variants were further classified as stop gain/loss, frameshift and non-frameshift mutations. As expected half of the small variants identified were intergenic; however, even though we identified a lot of frameshift mutations in both insertions and deletions very few caused a stop gain or stop loss (Figure 5).

Figure 5: Small insertions and deletions identified in 9 strains when compared to 1306 and their annotation. Variants were annotated based on their location within and outside the genes. Variants present outside the genes but within 1kb of genes were annotated as upstream/downstream variants while variants outside the 1kb range of genes were annotated as intergenic. Based on gff file, *Annovar* can find protein coding sequence limits, UTRs, introns and exon-intron junctions. Exonic variants were further classified as stop gain/loss, frameshift and non-frameshift mutations.



Variant confirmation: For many of the large variants that have genes we used Savant and IGV genome browser to confirm the variant. Savant use different color arcs to identify different SVs and can display the differences with multiple genome very effectively. A snapshot of the deletion SV present in 618 strain while absent in 1306 is shown in Figure 6. The deletion event is ~3.5 kb in size and is affecting 12 genes.

Figure 6: Deletion structural variant identified in isolate 618 when compared to 1306. Visualization was done using genome browser Savant. a. shows the genes that were predicted to be present in the region. b. represents the genome sequence c. the big red arc shows the reads that span the deletion in 618 with blue arcs and arrows showing the reads present in the region. d. the blue arcs and arrows display the normal read alignments of the reads in 1306.



Discussion:

Chromosomal rearrangements are major drivers of evolution and are also called as the "raw material for evolution" (Dobzhansky, 1937). A large number of studies have been done in humans and other organism where these structural variants have played a role in sex chromosome differentiation, reproductive isolation, speciation, adaptation and pathogenicity (Ranz et al., 2007; Noor et al., 2007; Faria and Navarro, 2010; Conrad and Hurles, 2007; Al-Hasani et al., 2001).

Some studies in other *Phytopthora* species have looked into structural variation within gene families. These studies mostly looked at present/absent polymorphism or copy number variation in effector proteins (Qutob et al., 2009; Raffaele et al., 2010; Cooke et al., 2012). Qutob et al. studied copy number variation in RxLR effector proteins Avr1a and Avr3a in *P. sojae* strains and found that deletion of 2 copies of Avr1a in some strains lead to a change in virulence (Qutob et al., 2009). Another study on clade 1c species (that include *P. infestans, P. ipomoeae, P. mirabilis* and *P. phaseoli*) detected 3,975 copy number variation events and found that most of these events are significantly higher in gene sparse region (Raffaele et al., 2010). Jiang et al. showed copy number variation in *P. infestans* Avr3b-Avr10-Avr11 locus that determines avirulence to potato resistance genes (Jiang et al., 2006).

Even though some studies have looked into variation in different lineages, none of the studies have taken a systematic approach to identify and classify the differences that exist in genome. This is the first study in *P. infestans* that has used a methodical way to first identify and then confirm and annotate these differences. We have identified large

and small differences in different lineages of *P. infestans*. We have used a very conservative approach to identify the present/absent variations in terms of genes and have generated a robust set of the variants by using multiple programs that identify structural variants and then filtering them with different criteria. By using two different programs, each of which uses a different approach to find large SVs, we were able to remove many false positives associated with each approach.

Deletion SVs predicted for three lineages 550, 6629 and US-22 are higher as compared to other strains. All three strains are sensitive to metalaxyl; further characterization of the deletion SVs identified in them can help us find genes that were found in commonly deleted in all the strains. These genes might be involved in metalaxyl resistance in the strains with these genes.

This study has opened up many new aspects in terms of variation present among different strains that will help us in not only identifying regions that might be involved in defining mating type, pathogenicity or fungicide resistance in different strains but also understand as to how these populations are evolving over time. Clustering of these lineages in different ways; for example, clustering all the lineages with same mating type can give us insights to identify the regions that might be involved in determining mating type.

Various studies in plants and animals have shown the strong correlation between structural variants and repeat-rich region (Bzymec and Lovett, 2001; Raskina et al., 2008). Repetitive DNA constituted 74% of the genome in the previous assembly of *P. infestans* strain T30-4. *Maker* has identified repeat-rich regions in the assembly of 1306.

It will be interesting to see as to how many large SVs that were identified in this study are present in this repeat-rich region.

Studies on fungal pathogens have identified set of chromosomes that are not shared among many strains of the same species and are called as "accessory chromosomes" (Cover, 1998; Croll and McDonald, 2012; Quaedvlieg et al., 2011). These chromosomes play an adaptive role in the evolution of pathogen and might be involved in virulence (Goodwin et al., 2011; Stukenbroch et al., 2010). Fifteen chromosomes have been identified in the current assembly of 1306 (*Michael Matson, unpublished data*). We can speculate that some of these chromosomes might be linked to pathogenicity and further characterization of the genes present on these chromosomes can help understand the differences in pathogenicity present among different strains.

Apart from structural variation, we can also use the new assembly to identify genes that might be specific to 1306 strain. Previous studies in fungal species have shown strain specific genes related to pathogen-host interaction and secretory proteins (Hittalmani et al., 2016; Tisserant et al., 2013). We ran MAKER with repeated content of the genome was masked and no genes were predicted in the repeat-rich regions. However, for future studies MAKER can be run without masking of the genome to find effector proteins that differ between the strains. Then using the genomic reads from all the strains used in the study we can determine as to which of the strains have those effector proteins and which have lost the genes.

Methods:

Alignment of reads: Paired end genomic reads from 9 different isolates (550, 618, 1114, 6629, US-8, US-11, US-22, US-23 and US-24) were aligned to the new PacBio assembly of 1306 strain using bwa aligner (Li and Durbin, 2009). 1306 reads were also aligned to the reference genome of 1306 and used as control bam file. The resulting sam files were then converted to bam files and sorted and indexed using samtools (ver 1.3). The sorted bam files were then given to Picard CollectInsertSizeMetrics tool to get the insert sizes between the adapters.

Identification, filtering and classification of structural variants (SVs): Three different softwares *Pindel*, *BreakDancer* and *GATK* were used to identify variants present in 9 different lineages specified above when compared to 1306.

The input to *Pindel* was sorted bam file obtained by samtools of each strain along with bam files from 1306 and insert sizes collected from each of the above bam files. All the structural variants obtained by *Pindel* were then converted to vcf files which in turn were filtered for bad quality reads and mapping quality. The header of each SV reported in original files from *Pindel* were extracted using grep command and a perl script was written to remove all the variants that had less than 5 reads supporting them. The results from the above two steps – filtering using vcf files and removal of SVs based on read coverage – were merged using a perl script giving us a more reliable dataset. This dataset was then checked against SVs called with 1306 strain. Many SVs identified in different lineages were also identified in 1306 and were heterozygous in 1306. In order to remove

SVs that were commonly present in the strain of interest and 1306, only those variants were kept where the number of reads supporting the variant in 1306 were less than 3.

BreakDancer was run with bam files from strains mentioned above along with 1306 bam files for each strain. Custom perl script was written to split the output file from *BreakDancer* into 3 different output files – insertion, deletions and inversions.

Picard tools were first used to sort the bam file obtained by samtools coordinate-wise, and the resulting file was then marked for duplicates followed by addition of readgroups. The sorted bam file with duplicates marked and read-group added was then realigned around INDELs and *GATK* UnifiedGenotyper was run with 1306 assembly as reference genome and bam files from above strains. 1306 bam files were also subjected to variant calling and were used as control files in filtering the variants. The Indels were then also filtered for mapping quality, read quality, and variants present at the end of reads or present on only one strand. The filtered Indels obtained above were then split into insertion and deletion vcf files.

The four output files of *Pindel* for each strain (insertions, deletions, inversions and translocations) were classified as long (≥1 Kb), medium (≥30bp-<1 kb) and small (< 30bp) using custom perl scripts. The same was done for separated insertion, deletion and invertion files obtained by *BreakDancer*.

For small insertions and deletions, only those variants that were present in 2 out of the 3 methods used to identify variants were then checked for annotation. For medium and large sized variants, only those variants were kept that were commonly identified by both *Pindel* and *BreakDancer*.

Identification of genes in the new genome: Genes were predicted in the new assembly using *Maker* annotation transfer tool, which was run with fasta files of proteins predicted in previous assembly of *P. infestans* strain T30-4, transcript evidence from cufflinks, trinity and PASA, and gff3 file from T30-4.

Proteins predicted by *Maker* always do not start with methionine, or may not be proteins at all. Perl scripts were written to check the following (i) proteins start with methionine (ii) protein has a length of 100 amino acids (iii) protein is not a putative ncRNA.

Custom perl scripts were written to generate final gff files and fasta files after filtering the proteins for the above criteria.

Annotation of variants: Custom perl scripts using bedtools were written to find if any of the large and medium SVs have genes involved. Small SVs (<30bp) were annotated using *ANNOVAR* variant annotation package.

References:

Al-Hasani, K., Adler, B., Rajakumar, K., & Sakellaris, H. (2001). Distribution and structural variation of the she pathogenicity island in enteric bacterial pathogens. Journal of Medical Microbiology, 50, 780-786.

Auwera, G. A., Carneiro, M. O., Hartl, C., Poplin, R., del Angel, G., Levy-Moonshine, A., … & DePristo, M. A. (2013). From FastQ data to high confidence variant calls: The Genome Analysis Toolkit best practices pipeline. Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.], 11, 11.10.1–11.10.33.

Chen, K., Wallis, J. W., McLellan, M. D., Larson, D. E., Kalicki, J. M., Pohl, C. S., ... & Shi, X. (2009). BreakDancer: an algorithm for high-resolution mapping of genomic structural variation. Nature Methods, 6, 677-681.

Coll, F., Preston, M., Guerra-Assunção, J. A., Hill-Cawthorn, G., Harris, D., Perdigão, J., ... & Glynn, J. R. (2014). PolyTB: a genomic variation map for *Mycobacterium tuberculosis*. Tuberculosis, 94, 346-354.

Conrad, D. F., & Hurles, M. E. (2007). The population genetics of structural variation. Nature Genetics, 39, S30-S36.

Coscolla, M., & Gagneux, S. (2014). Consequences of genomic diversity in *Mycobacterium tuberculosis*. Seminars in Immunology, 26, 431-444.

Danies, G., Small, I. M., Myers, K., Childers, R., & Fry, W. E. (2013). Phenotypic characterization of recent clonal lineages of *Phytophthora infestans* in the United States. Plant Disease, 97, 873-881.

Faria, R., & Navarro, A. (2010). Chromosomal speciation revisited: rearranging theory with pieces of evidence. Trends in Ecology & Evolution, 25, 660-669.

Fry, W. E., Goodwin, S. B., Dyer, A. T., Matuszak, J. M., Drenth, A., Tooley, P. W., Sujkowski, L. S., Koh, Y. J., Cohen, B. A., Spielman, L. J., Deahl, K. L., Inglis, D. A., & Sandlan, K. P. (1993). Historical and recent migration of *Phytophthora infestans*: Chronology, pathways, and implications. Plant Disease, 77, 653-661.

Fry, W. E., Goodwin, S. B., Matuszak, J. M., Spielman, L. J., Milgroom, M. G., & Drenth, A. (1992). Population genetics and intercontinental migrations of *Phytophthora infestans*. Annual Review of Phytopathology, 30, 107-130.

Fry, W. E., McGrath, M. T., Seaman, A., Zitter, T. A., McLeod, A., Danies, G., ... & Gugino, B. K. (2013). The 2009 late blight pandemic in the eastern United States–causes and results. Plant Disease, 97, 296-306.

Feuk, L., Carson, A. R., & Scherer, S. W. (2006). Structural variation in the human genome. Nature Reviews Genetics, 7, 85-97.

Futreal, P. A., Coin, L., Marshall, M., Down, T., Hubbard, T., Wooster, R., ... & Stratton, M. R. (2004). A census of human cancer genes. Nature Reviews Cancer, 4, 177-183.

González, A., Plener, L., Restrepo, S., Boucher, C., & Genin, S. (2011). Detection and functional characterization of a large genomic deletion resulting in decreased pathogenicity in *Ralstonia solanacearum* race 3 biovar 2 strains. Environmental Microbiology, 13, 3172-3185.

Gout, L., Kuhn, M. L., Vincenot, L., Bernard-Samain, S., Cattolico, L., Barbetti, M., ... & Rouxel, T. (2007). Genome structure impacts molecular evolution at the *AvrLm1* avirulence locus of the plant pathogen *Leptosphaeria maculans*. Environmental Microbiology, 9, 2978-2992.

Hartmann, F. E., Sánchez-Vallet, A., McDonald, B. A., & Croll, D. (2017). A fungal wheat pathogen evolved host specialization by extensive chromosomal rearrangements. The ISME Journal, 11, 1189-1204.

Hill, K. K., Xie, G., Foley, B. T., Smith, T. J., Munk, A. C., Bruce, D., ... & Detter, J. C. (2009). Recombination and insertion events involving the botulinum neurotoxin complex genes in *Clostridium botulinum* types A, B, E and F and *Clostridium butyricum* type E strains. BMC Biology, 7, 66.

Hittalmani, S., Mahesh, H. B., Mahadevaiah, C., & Prasannakumar, M. K. (2016). De novo genome assembly and annotation of rice sheath rot fungus *Sarocladium oryzae* reveals genes involved in Helvolic acid and Cerulenin biosynthesis pathways. BMC Genomics, 17, 271.

Jiang, R. H., Weide, R., van de Vondervoort, P. J., & Govers, F. (2006). Amplification generates modular diversity at an avirulence locus in the pathogen *Phytophthora*. Genome Research, 16, 827-840.

Kato, M., Mizubuti, E. S., Goodwin, S. B., & Fry, W. E. (1997). Sensitivity to protectant fungicides and pathogenic fitness of clonal lineages of *Phytophthora infestans* in the United States. Phytopathology, 87, 973-978.

Mitelman, F., Johansson, B., & Mertens, F. (2007). The impact of translocations and gene fusions on cancer causation. Nature Reviews Cancer, 7, 233-245.

Ogoshi, A., Sato, N., & Fry, W. E. (1994). Migrations and displacements of *Phytophthora infestans* populations in east Asian countries. Phytopathology, 84, 922-927.

Marshall, C. R., Noor, A., Vincent, J. B., Lionel, A. C., Feuk, L., Skaug, J., ... & Thiruvahindrapduram, B. (2008). Structural variation of chromosomes in autism spectrum disorder. The American Journal of Human Genetics, 82, 477-488.

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., ... & Chinwalla, A. (2011). Mapping copy number variation by population-scale genome sequencing. Nature, 470, 59-65.

Mizubuti, E. S., & Fry, W. E. (1998). Temperature effects on developmental stages of isolates from three clonal lineages of *Phytophthora infestans*. Phytopathology, 88, 837-843.

Noor, M. A., Grams, K. L., Bertucci, L. A., & Reiland, J. (2001). Chromosomal inversions and the reproductive isolation of species. Proceedings of the National Academy of Sciences of the United States of America, 98, 12084-12088.

Perez, W. G., Gamboa, J. S., Falcon, Y. V., Coca, M., Raymundo, R. M., & Nelson, R. J. (2001). Genetic structure of Peruvian populations of *Phytophthora infestans*. Phytopathology, 91, 956-965.

Qutob, D., Tedman-Jones, J., Dong, S., Kuflu, K., Pham, H., Wang, Y., ... & Gijzen, M. (2009). Copy number variation and transcriptional polymorphisms of *Phytophthora sojae* RXLR effector genes *Avr1a* and *Avr3a*. PloS One, 4, e5066.

Ranz, J. M., Maurin, D., Chan, Y. S., Von Grotthuss, M., Hillier, L. W., Roote, J., ... & Bergman, C. M. (2007). Principles of genome evolution in the *Drosophila melanogaster* species group. PLoS Biology, 5, e152.

Rasko, D. A., Rosovitz, M. J., Myers, G. S., Mongodin, E. F., Fricke, W. F., Gajer, P., ... & Henderson, I. R. (2008). The pangenome structure of *Escherichia coli*: comparative genomic analysis of *E. coli* commensal and pathogenic isolates. Journal of Bacteriology, 190, 6881-6893.

Read, T. D., Salzberg, S. L., Pop, M., Shumway, M., Umayam, L., Jiang, L., ... & Solomon, D. (2002). Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis*. Science, 296, 2028-2033.

Rovelet-Lecrux, A., & Campion, D. (2012). Copy number variations involving the microtubule-associated protein tau in human diseases. Biochemical Society Transactions, 40, 672-676.

Saville, A., Graham, K., Grünwald, N. J., Myers, K., Fry, W. E., & Ristaino, J. B. (2015). Fungicide sensitivity of US genotypes of *Phytophthora infestans* to six oomycete-targeted compounds. Plant Disease, 99, 659-666.

Singleton, A. B., Farrer, M., Johnson, J., Singleton, A., Hague, S., Kachergus, J., ... & Lincoln, S. (2003). α-Synuclein locus triplication causes Parkinson's disease. Science, 302, 841-841.

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., … Schnable, P. S. (2009). Maize inbreds exhibit high levels of copy number variation (cnv) and presence/absence variation (pav) in genome content. PLoS Genetics, 5, e1000734.

Szatmari, P., Paterson, A. D., Zwaigenbaum, L., Roberts, W., Brian, J., Liu, X. Q., ... & Feuk, L. (2007). Mapping autism risk loci using genetic linkage and chromosomal rearrangements. Nature Genetics, 39, 319-328.

Tisserant, E., Malbreil, M., Kuo, A., Kohler, A., Symeonidi, A., Balestrini, R., ... & Gilbert, L. B. (2013). Genome of an arbuscular mycorrhizal fungus provides insight into the oldest plant symbiosis. Proceedings of the National Academy of Sciences of the United States of America, 110, 20117-20122.

Welch, T. J., Fricke, W. F., McDermott, P. F., White, D. G., Rosso, M. L., Rasko, D. A., ... & Rahalison, L. (2007). Multiple antimicrobial resistance in plague: an emerging public health risk. PloS One, 2, e309.

Winzer, T., Gazda, V., He, Z., Kaminski, F., Kern, M., Larson, T. R., ... & Walker, C. (2012). A *Papaver somniferum* 10-gene cluster for synthesis of the anticancer alkaloid noscapine. Science, 336, 1704-1708.

Wollenberg, T., & Schirawski, J. (2014). Comparative genomics of plant fungal pathogens: The *Ustilago-Sporisorium* paradigm. PLoS Pathogen, 10, e1004218.

Ye, K., Schulz, M. H., Long, Q., Apweiler, R., & Ning, Z. (2009). Pindel: a pattern growth approach to detect break points of large deletions and medium sized insertions from paired-end short reads. Bioinformatics, 25, 2865-2871.

Zhang, F., Gu, W., Hurles, M. E., & Lupski, J. R. (2009). Copy Number Variation in Human Health, Disease, and Evolution. Annual Review of Genomics and Human Genetics, 10, 451–481.

Chapter IV

Lifestyle, gene gain and loss, and transcriptional remodeling cause divergence in the transcriptomes of *Phytophthora infestans* and *Pythium ultimum* during potato tuber colonization

Abstract:

Background: How pathogen genomes evolve to support distinct lifestyles is not well-understood. The oomycete *Phytophthora infestans,* the potato blight agent, is a largely biotrophic pathogen that feeds from living host cells, which become necrotic only late in infection. The related oomycete *Pythium ultimum* grows saprophytically in soil and as a necrotroph in plants, causing massive tissue destruction. To learn what distinguishes their lifestyles, we compared their gene contents and expression patterns in media and a shared host, potato tuber.

Results: Genes related to pathogenesis varied in temporal expression pattern, mRNA level, and family size between the species. A family's aggregate expression during infection was not proportional to size due to transcriptional remodeling and pseudogenization. *Ph. infestans* had more stage-specific genes, while *Py. ultimum* tended towards more constitutive expression. *Ph. infestans* expressed more genes encoding secreted cell wall-degrading enzymes, but other categories such as secreted proteases and ABC transporters had higher transcripts in *Py. ultimum*. Species-specific genes were identified including new *Pythium* genes, perforins, which may disrupt plant membranes. Genome-wide ortholog analyses identified substantial diversified expression, which

correlated with sequence divergence. Pseudogenization was associated with gene family expansion, especially in gene clusters.

Conclusion: This first large-scale analysis of transcriptional divergence within oomycetes revealed major shifts in genome composition and expression, including subfunctionalization within gene families. Biotrophy and necrotrophy seem determined by species-specific genes and varied expression of shared pathogenicity factors, which may be useful targets for crop protection.

Background:

A key issue in plant-microbe interactions is understanding what distinguishes different pathogenic lifestyles. These range from biotrophic relationships in which the pathogen feeds from living host cells, to necrotrophic associations in which the microbe feeds on nutrients released from killed cells (Lewis, 1973). Some pathogens can switch between biotrophic and necrotrophic growth, or grow as saprophytes on organic debris. Events such as gene duplication, loss, regulatory subfunctionalization, neofunctionalization, and horizontal gene transfer are contributors to the diversification of such behaviors by microbes (Skamnioti et al., 2008; Richard et al., 2011).

Genomic, transcriptomic, and other analyses have identified many factors that underlie plant pathogen lifestyles. These include effectors that suppress host defenses during biotrophic growth, enzymes that defeat antimicrobial compounds produced by the host, cell wall-degrading enzymes (CWDEs), and transporters for acquiring nutrients. Attributes specific to each trophic type have remained elusive, however. While defense-suppressing effectors are commonly thought to be specific to biotrophs, some necrotrophs also suppress plant immunity (Weiberg et al., 2013). There is also no general correlation between lifestyle and the number of CWDEs encoded by a pathogen (Kabbage et al., 2015; Zhao et al., 2013; Verma et al., 2016). Instead, biotrophs and necrotrophs may be distinguished by the timing or level of expression of shared factors (Doehlemann et al., 2017; Meinhardt et al., 2014; Lyu et al., 2016).

. One limitation of many biotroph-necrotroph comparisons is that they examine pathogens on different hosts, which may present the microbe with distinct physical or chemical signals (O'connell et al., 2012; Lyu et al., 2015). In the current study, we reduce

such complications by studying *Phytophthora infestans* and *Pythium ultimum* on a common host, potato tubers. *Phytophthora* and *Pythium* are sister taxa in the peronosporalean lineage of oomycetes, and a responsible for blights, rots, and damping-off diseases of thousands of important plant species (Beakes et al., 2012). *Ph. infestans* is the notorious Irish Famine pathogen. Its disease, late blight, occurs on tubers when the pathogen enters natural openings or wounds. *Ph. infestans* is described as a hemibiotroph, since host necrosis occurs only near the end of the life cycle. *Py. ultimum* also infects tubers through wounds but is a necrotroph. Its disease, potato leak, is characterized by dark lesions that culminate in watery rotted tissue. Other distinctions are that only *Py. ultimum* can persist in nature as a saprophyte, only *Py. ultimum* has a broad host range, and only *Ph. infestans* forms abundant sporangia, which appear late in infection. Sporangia release zoospores, which initiate most infections in late blight. While *Py. ultimum* sometimes makes terminal hyphal swellings, it is unclear if they release zoospores (Benhamou et al., 1998; Kamoun et al., 2015).

Oomycete genomes contain many fast-evolving regions and gene families (Dong et al., 2015). Studies in other kingdoms have shown that family expansions are often associated with changes in protein function, pseudogenization, or regulatory subfunctionalization *i.e.* altered transcription (Skamnioti et al., 2008). While there is moderate synteny between the genomes of *Ph. infestans* and *Py. ultimum*, the former is larger (237 vs. 43 Mb) with more predicted genes (17,797 vs. 15,290) (Levesque et al., 2010; Haas et al., 2009). Examples of expanded families in *Ph. infestans* and *Py. ultimum* are RxLR effectors and secreted proteases, respectively. Little is known of how changes in gene families are manifested in their temporal patterns or levels of expression in

oomycetes (Adhikari et al., 2013). Genome-wide expression studies of *Ph. infestans* and *Py. ultimum* during plant infection are also limited. These include a study of *Py. ultimum* growing on *Arabidopsis* seeds in dilute V8 juice (Levesque et al., 2010) and *Ph. infestans* on tomato and potato leaves (Haas et al., 2009; Zuluaga et al., 2016), although the latter lacked sufficient sequence depth at most timepoints to measure most genes.

Here, we use deep sequencing to compare the transcriptomes of *Ph. infestans* and *Py. ultimum* on potato tubers and artificial media. Infection-induced genes were identified in both species, but global analyses and those focused on specific functional groups revealed divergence in stage-specific expression patterns and mRNA levels. Differences between the hemibiotroph and necrotroph could be attributed to at least five phenomena: transcriptional remodeling that affects the level of expression; changes in the timing of expression; expansion and contraction of gene families; species-specific genes; and selective pressure on gene families to support biotrophy or necrotrophy.

RESULTS AND DISCUSSION

Comparisons of disease development

Conditions for examining gene expression were established by testing strategies for inoculating tuber slices with zoospores of *Ph. infestans* and hyphae of *Py. ultimum*. These are the main propagules that infect tubers in soil or storage (Kamoun et al., 2015; Green et al., 2000). With *Ph. infestans,* fairly synchronous infections were obtained by spreading zoospores over each tuber slice. Consistent with the biotrophic nature of the early to middle stages of the disease cycle, the first macroscopic evidence of infection were a few surface hyphae on the non-inoculated side of the slice at 2.5 dpi (days post-infection). Moreover, microscopic analyses performed between 1.5 and 2.5 dpi revealed haustoria. By 3.5 dpi, hyphae and a few sporangia were observed on both sides of the tuber. By 4 dpi, more surface hyphae and sporangia were evident as was slight darkening of host tissue (Fig. 1a). An alternative strategy for infecting tubers by applying a single drop of zoospores proved unsatisfactory. This resulted in lesions that expanded nonuniformly, and often sporulated first at patches distant from the inoculation site.

Disease progression was more rapid with *Py. ultimum.* When tubers were inoculated with a plug of hyphae, an expanding zone of darkened host tissue appeared by 0.5 dpi. At 1.5 dpi, host tissues were reddish brown on the outer periphery of the lesion, brown in the middle-aged part, and black in the oldest part (Fig. 1a). Sparse hyphae were seen on the surface of the brown and black zones. Hyphae were found within the tuber in the discolored region, but few in the non-discolored region. An alternative strategy for infecting tubers by applying hyphae to the entire slice proved unsatisfactory, since movement of the pathogen through the tuber was not uniform.

Based on the above, we collected material for RNA-seq by spreading *Ph. infestans* zoospores on tuber slices and harvesting 1.5, 2.5, and 4 dpi samples as early, middle, and late infection samples. For *Py. ultimum,* infections were initiated using a hyphal plug, and at 1.5 dpi the tuber was dissected into concentric zones representing early, middle, and late stages of colonization. The outer (early) zone included a 3-mm region comprised of one-quarter of nondiscolored tuber tissue and three-quarters reddish brown tissue. The middle region included 3-mm of brown tissue. The late sample included 3-mm of black tissue.

Initial analysis of pathogen transcriptomes

RNA-seq was performed using the three infection stages described above, with three biological replicates. Early and late stages of growth of *Ph. infestans* and *Py. ultimum* on rye and pea broth were also analyzed to allow the identification of infection-induced genes. For *Ph. infestans*, the early and late media cultures represented non-sporulating and sporulating timepoints, respectively. With *Py. ultimum,* sporangia-like hyphal swellings were detected in the late cultures, but their density was <5% of that of *Ph. infestans* sporangia.

The number of reads from the tuber and media samples are shown in Table S1. The fraction coming from the pathogens increased over the course of infection. In early, middle, and late stages of tubers infected with *Ph. infestans,* 3.9, 30, and 76% of reads mapped to the pathogen, respectively. This increased from an average of 13 million (MM) to 234 MM reads for each early and late tuber replicate, respectively. In tubers infected with *Py. ultimum* the early, middle, and late samples resulted in 24, 88, and 90%

of reads mapping to the pathogen, increasing from an average of 38 MM to 74 MM per replicate. The mapping percentage from late *Py. ultimum*-infected tubers is close to that of pure cultures (90 vs. 93%), indicating that little potato RNA persists late in infection. In contrast, host transcripts were eliminated more slowly with *Ph. infestans*. Using a minimum CPM (counts per million mapped reads) cut-off of 1.0, 14,081 *Ph. infestans* and 12,164 *Py. ultimum* genes were expressed in at least one media or tuber condition. CPM calls for both species are shown in Table S2, along with annotations for each gene.

To confirm that *Ph. infestans* went through a normal disease cycle, four stage-specific genes were examined (Fig. 1c). Effector Avr3a and haustorial protein Hmp1 mark the biotrophic stage of infection (Jupe et al., 2013). Both of their genes were expressed at much higher levels in the early and middle stages of tuber colonization than the late stage. Induced more than 100-fold in the late sample was the gene encoding a marker for necrotrophy, NPP1(Jupe et al., 2013). The same was observed for the gene encoding centrin, which is a marker for sporulation, which occurs at the end of the disease cycle (Judelson et al., 2012).

To test further the quality of the data and start to obtain insight into pathogen biology, we performed principal component analysis (PCA). Tight clustering was seen for each biological replicate (Fig. 1d, e). Early rye and pea samples clustered in both species, but not with early tubers, indicating major divergence in the transcriptomes of "young" mycelia in tubers and artificial media.

Heatmaps based on hierarchical clustering also indicated that major changes occurred in both pathogens during *ex planta* and *in planta* growth (Fig. 1b). Based on a fold-change ratio threshold of 4.0 and a false discovery rate (FDR) cut-off of 0.05, 15.5% and

10.0% of *Ph. infestans* and *Py. ultimum* genes were expressed differentially, respectively, in early tubers compared to early media (Fig. 1f). This indicated that while transcriptomic reprogramming occurred in both pathogens, a greater fraction of genes in *Py. ultimum* were expressed constitutively. One type of gene that underlies this difference are RxLR effectors, which are numerous and often infection-induced in *Ph. infestans* but absent from *Py. ultimum*.

Another difference between the species was observed when comparing early and late tubers. While only 9.4% of *Py. ultimum* genes changed by ≥4-fold (FDR<0.05) between early and late infection, 45% of *Ph. infestans* genes changed to the same degree (Fig. 1f). One explanation for this difference is that during late infection *Ph. infestans* switches from biotrophic to necrotrophic growth, and sporulates. This further supports the conclusion that *Py. ultimum* displays a more constitutive pattern of expression of its genes.

Gene ontology (GO) analysis

Infection-induced GO categories (Fig. 2) that were over-represented in both species included carbon oxygen lyase, serine endopeptidase, polysaccharide metabolism, and pectin catabolism; the latter categories represent cell-wall degrading enzymes (CWDEs). This likely reflects the fact that *Ph. infestans* grows biotrophically and causes minimal damage to the host, while *Py. ultimum* is a necrotroph. Many changes observed for *Ph. infestans* resembled those described in a microarray study of *Ph. capsici*, such as the up-regulation early in infection of RxLR effectors and genes involved in transcription (Jupe et al., 2013).

In comparisons of early tubers with media, differences between the pathogens included the RNA metabolism, RNA binding, and ribosome biogenesis categories. These GO terms were over-represented in *Ph. infestans* but under-represented in *Py. ultimum* (Fig. 2a). This suggests that *Ph. infestans* retools its cellular machinery during infection, perhaps reflecting the need to produce haustoria, secrete effectors, and shift from acquiring nutrients from media to the plant. GO categories overrepresented in *Py. ultimum* but not *Ph. infestans* included those associated with transmembrane transporters. As will be described later, many ABC transporters are expressed highly during tuber colonization by *Py. ultimum*. A similar pattern was reported for necrotrophic fungal plant pathogens (Lewis et al., 1973).

A few over-represented GO terms were shared by the two species when comparing late to early tubers (Fig. 2b). These included glycosyl hydrolases, glucosidases, and enzymes associated with lipid metabolism. This may reflect a subtle change from earlier stages, when damage to host structures may be used more to allow hyphae to penetrate host cells than to liberate metabolizable carbon. Most over-represented GO terms were however unique to *Ph. infestans.* These included categories associated with structural components of sporangia (*e.g.* centrosome, cilium) and signal transduction (serine/threonine kinase). The latter could participate in spore-associated functions or a transition to necrotrophic growth. A prior study of *Ph. infestans* observed changes in many protein kinases during sporulation (Ah-Fong et al., 2017). While older cultures of *Py. ultimum* form hyphal swellings and sexual spores, their concentrations were low and likely insufficient to cause obvious gene expression changes.

Several genomic processes may explain the evolution of the differences between *Ph. infestans* and *Py. ultimum* described above and later in this paper. Expression patterns may have changed due to mutations in promoters or transcription factors. Other possibilities are gene gain or loss, including the expansion or contraction of families. In the following sections, we address these processes as well as factors that may underlie biotrophic and necrotrophic growth. First, we will focus on functional classes of genes implicated in pathogenesis. Then, we will study ortholog expression patterns in order to address regulatory subfunctionalization genome-wide.

Expression of putative pathogenicity genes

Preliminary explorations indicated that transcription should be addressed both as the expression patterns of individual genes and aggregate (summed) CPM values within functional categories. To explain why, Fig. 3 illustrates a family of polyphenol oxidases, which may help protect pathogens against toxins or stress (Mayor, 2006). These have 17 paralogs in *Ph. infestans* and 14 in *Py. ultimum*. Conventional heatmap analysis using per-gene normalized data (black and white checkerboards in Fig. 3) indicates that most of the genes are expressed in both species at their highest levels in late tuber and late rye, respectively. However, this is a distorted view since some genes are expressed more highly than others. When the expression of all genes are considered by adding together their individual CPMs, we observe that transcript levels in *Py. ultimum* are highest in early tuber, and fairly constant in *Ph. infestans*.

Similar patterns were seen within many gene families, which based on OrthoMCL analysis encompass about half of the genes in the two species. Therefore, expression will

be described in the following sections both as aggregate CPM and in heatmaps as per-gene normalized values. The latter is the most common way of illustrating expression data, but masks the divergence in CPM of individual genes. It is acknowledged that weakly-expressed genes may have important functions, and that mRNA and protein levels may not increase proportionally.

Secreted proteases. These may degrade structural components of the host to facilitate pathogen ingress, liberate nutrients, or defeat host defense proteins (Van der Hoorn, 2006). We predict that 29 of 101 *Ph. infestans* and 72 of 171 *Py. ultimum* proteases, respectively, are secreted. These numbers are larger than those reported previously (Levesque et al., 2010) since our study included more families of proteases. Of the genes encoding secreted proteases, 25 and 64 were expressed in at least one tissue of each species, respectively, based on a CPM cut-off of 1.0.

Distinct trends were observed for each family (Fig. 4a). In this and subsequent figures, the left column is a heatmap based on per-gene normalized values; only genes with CPM ≥1 in at least one growth condition are included. The pie charts in the center denote the fraction of genes that are up- and down-regulated by ≥3-fold in early tuber versus early media (average of rye and pea), and late tuber versus early tuber. The right column displays the aggregate CPM of the genes in early media, early tuber, and late tuber.

Several classes of proteases showed consistently higher transcription, and more infection-induced genes, in *Py. ultimum* compared to *Ph. infestans*. Eight of 23 expressed (*e.g.* having CPM ≥1) subtilisin proteases in *Py. ultimum* were induced in early tubers

compared to none of the six expressed *Ph. infestans* genes. Combined with expansion of the family in *Py. ultimum*, this resulted in 50-fold higher CPM in that species compared to *Ph. infestans*. mRNA levels were also higher in *Py. ultimum* for aspartyl proteases, which were expressed from one and ten genes in *Ph. infestans* and *Py. ultimum*, respectively. Metalloproteases also had higher total CPM in *Py. ultimum* at each stage. Although more metalloproteases were tuber-induced in *Ph. infestans,* due to low expression this did not translate into higher overall CPM.

Unlike the prior enzymes, trypsin proteases in *Ph. infestans* and *Py. ultimum* had similar numbers of expressed genes (seven and six, respectively) and aggregate CPM in early tubers. More genes were induced in early tubers compared to media in *Ph. infestans* than *Py. ultimum* (four and one, respectively). CPM in both species dropped by >10-fold in late tubers. While half of the *Py. ultimum* genes were late-induced*,* these were expressed at low levels and did not prevent the decline in total CPM.

While levels of subtilisin, trypsin, and metalloproteases all fell in late versus early tubers, the opposite pattern was seen for the eight and 12 cysteine proteases expressed by *Ph. infestans* and *Py. ultimum*, respectively. Both species had genes induced in late tuber, which caused aggregate CPM to rise. In late tubers, cysteine proteases were expressed at about 4000 CPM in both species, which was three to seven-fold more than the other proteases combined.

Overall, the data suggest that proteases act in two phases during infection. The initial phase may involve trypsin proteases in *Ph. infestans* and subtilisin, metallo-, and trypsin proteases in *Py. ultimum*. The subtilisin and metalloproteases are known for their abilities to function in inhospitable environments, which may include the apoplast (Baroncelli et

al., 2016). A second wave in both pathogens may involve cysteine and aspartyl proteases. The broader arsenal of proteases secreted by *Py. ultimum* may contribute to its ability to infect multiple hosts.

Secreted phosphatases. Most secreted phosphatases act on organic compounds to release soluble phosphate, which can be assimilated by the pathogens (Treseder and Lennonb, 2015). *Ph. infestans* and *Py. ultimum* are predicted to encode 26 and 21 of these enzymes, of which 21 and 20 were expressed at CPM ≥1, respectively (Fig. 4b). Of the expressed genes, two per species encoded acid phosphatases with the remainder encoding alkaline-favoring enzymes.

In early tubers compared to media, six acid and zero alkaline phosphatases were induced in *Ph. infestans,* compared to two and two, respectively, in *Py. ultimum.* Both species had a few late infection-induced phosphatases, and a similar fraction of late-repressed genes. As a consequence, aggregate CPM values of the *Ph. infestans* enzymes stayed fairly constant during tuber infection at about 1000 CPM, with an approximate 4:1 ratio of acid to alkaline phosphatases. The predominance of acid-favoring enzymes may be explained by the fact that leaf apoplasts and tubers are mildly acidic (Grignon and Sentenac, 1991).

The situation in *Py. ultimum* was quite different. Levels of acid phosphatase in *Py. ultimum* showed only minor changes between growth conditions, staying near 850 CPM. However, its alkaline phosphatases went from <1 CPM in media to 358 in early tuber and 1648 in late tuber. As a consequence, there was a shift from an excess of acid phosphatases in media and early tubers to an excess of alkaline activity in late tubers.

This was due largely to PYU1_G001049, which accounted for 96% of total CPM in late tubers. The shift might be explained by a need for *Py. ultimum* to prepare for growth in soil after the nutrients from its plant host become depleted. In alkaline soils, inorganic phosphate tends to form insoluble compounds, so alkaline phosphatase may help liberating phosphate from organics to support growth.

Secreted carbonic anhydrase. This converts carbon dioxide to bicarbonate, generating a proton which may contribute to pH homeostasis. The reaction also forms bicarbonate, which may be a co-factor or co-substrate in fatty acid and cAMP pathways [30]. Prior studies indicated that some carbonic anhydrases are infection-induced in *Ph. infestans* (Zuluaga et al., 2016; Raffaele et al., 2010).

Ph. infestans and *Py. ultimum* are predicted to encode 16 and eight carbonic anhydrases, of which seven and eight were predicted to be secreted and six and four expressed at CPM ≥1, respectively (Fig. 4b). A majority were up-regulated in early tubers compared to media. There is a corresponding increase in aggregate CPM in early tubers, but more in *Ph. infestans* (64-fold) than *Py. ultimum* (three-fold). The aggregate CPM in also much higher in *Ph. infestans* in early infection. This may reflect a role in maintaining pH levels optimal for other secreted enzymes during biotrophy. In both species, a role in balancing pH may also be indicated by the induction of several genes during late growth in media, and in late tubers in *Ph. infestans*. Interestingly, this involved genes other than those induced during early infection.

Secreted phospholipase D. A novel feature of certain oomycetes are secreted forms of phospholipase D, which in addition to degrading structural phospholipids can generate the signaling molecule phosphatidic acid [31, 32]. Whether the extracellular generation of this signaling compound affects the pathogen or host is unknown. *Ph. infestans and Py. ultimum* encode 11 and three of the secreted forms of these enzymes, of which 11 and two were expressed at CPM ≥1, respectively (Fig. 4b).

More than three-quarters of the *Ph. infestans* genes, but no *Py. ultimum* genes, were induced in early tubers. This is consistent with a prior report that some of the former were induced during leaf infection (Zuluaga et al., 2016). In our study, expression of nearly all *Ph. infestans* genes declined towards the end of the biotrophic stage. However, secreted phospholipases do not seem to define biotrophic growth since both species had similar CPM levels, and the highest levels, in early tubers.

Secreted protease inhibitors. Members of the kazal and cysteine protease inhibitor families help protect *Ph. infestans* against host enzymes (Song et al., 2009). The transcription patterns and functions of such proteins in necrotrophic oomycetes are not described. *Ph. infestans and Py. ultimum* encode 36 and 19 such proteins, respectively, of which 28 and 18 are predicted to be secreted and 18 and 11 expressed at CPM ≥1, respectively (Fig. 4b). About two-thirds of the *Ph. infestans* genes were induced in early tubers compared to media, declining to low levels late in infection. In contrast, nearly all of the *Py. ultimum* genes were expressed constitutively. Two were induced ≥3-fold in late tubers, but contributed little to total CPM. One may conclude that the existence of protease inhibitors *per se* is not a defining feature of a biotrophic lifestyle. Instead, their

expression level (CPM) or pattern may be a better indicator of lifestyle.

Secreted ribonuclease T2. *Ph. infestans* and *Py. ultimum* encode two and six secreted forms, respectively, of this non-specific ribonuclease, with two and four expressed at CPM ≥1 (Fig. 4b). In *Ph. infestans*, one of the genes was up-regulated during late infection, as well as in late rye and pea cultures. In *Py. ultimum,* three genes were expressed at the highest level in early infection, resulting in peak aggregate CPM at that stage. Of the putative pathogenicity factors addressed by this study, this was one of the few that appeared to be transcribed at a higher level during late infection by *Ph. infestans,* and could thus be linked to necrotrophic growth in both species.

In animals, a portion of secreted T2 enzymes are retained within the cell. There are also examples of secreted T2 enzymes being taken up by animal or plant cells (Luhtala and Parker, 2010). Therefore, the oomycete enzymes might be used to scavenge extracellular nucleic acids for nutrients, turn over pathogen mRNA, or modulate host defenses.

NPP family. Certain oomycetes, fungi, and bacteria express this family of proteins, which cause necrosis in plants. *NPP1,* encoded by gene PITG_16866 of *Ph. infestans*, is often used to mark late infection (Fig. 1c). It is known, however, that *Phytophthora* spp. encode multiple NPP proteins. Many are neither expressed late in infection nor cause necrosis and have unknown cellular roles (Gijzen and Nurnberger, 2006; Feng et al., 2014).

Ph. infestans encodes 55 NPP proteins, of which 40 are predicted to be secreted. Twenty-two were expressed in at least one growth condition at CPM ≥1. Of seven *NPP*

genes in *Py. ultimum,* six were expressed and all encoded secreted proteins (Fig. 5). Six *Py. ultimum* genes, but only nine *Ph. infestans* genes, were conserved at residues shown to be essential for inducing necrosis (Feng et al., 2014).

Intriguingly, transcription of the family varied substantially between the species. All of the *Py. ultimum* genes were expressed much more in early than late tubers. In *Ph. infestans,* three genes including PITG_16866 were expressed ≥3-fold higher in late compared to early tubers, while three other genes were ≥3-fold higher in early tubers. Notably, the *Ph. infestans* genes expressed in early tubers all lacked the residues needed for inducing necrosis. PITG_16866 was expressed at high levels in media and late tubers, but not early tubers, suggesting that its expression is suppressed during early infection to help maintain biotrophy. The proteins predicted as necrosis-causing had an aggregate CPM in early and late tubers of 1.9 and 16 in *Ph. infestans,* and 213 and 6.5 in *Py. ultimum*, respectively, consistent with their association with the necrotrophic stage.

Phylogenetic analyses indicated that the *Py. ultimum* genes form a well-supported clade with PITG_16866*,* even though they have opposite stage-specific expression patterns (Fig. 5). Despite extensive expansion of the family in *Ph. infestans,* the advantages of biotrophy to that species in early infection appears to have selected for mutations of the residues required for plant necrosis, in addition to remodeling its transcription pattern relative to that of *Py. ultimum.*

CRN family. This includes effectors that are secreted and translocated into plant nuclei [37]. However, most members of the family are not secreted and play unknown roles. Like *NPP1,* the *CRN* family is expanded in *Ph. infestans* relative to *Py. ultimum* with 196

and 26 members, respectively (Levesque et al., 2010). Of these, expression was detected at CPM ≥1 for 132 and 14 genes, respectively. No *Py. ultimum* genes were infection-induced. Ten *Ph. infestans* genes (PITG_04767, 04769, 05039, 16575, 16627, 17540, 18835, 19340, 21058, 22536) had ≥3-fold higher expression in tubers compared to media, although only four of those encode secreted proteins. That most CRNs are not induced *in planta* has been described (Haas et al., 2009).

The expression pattern of CRNs differs from that of RxLR effectors (Stassen et al., 2011), which occur in *Ph. infestans* but not *Py. ultimum.* Of 163 expressed *Ph. infestans* RxLRs, 129 were induced ≥3-fold in tubers compared to media. The expression of 121 genes peaked in early tubers.

Plant cell wall degrading enzymes (CWDEs)

Oomycetes and fungi secrete diverse enzymes to depolymerize the cell walls of their hosts. Since both oomycete and plant walls are comprised mostly of β-1,3 and β-1,4 glucans, they are affected by similar enzymes. The presence of transmembrane domains and glycosylphosphatidylinositol (GPI) anchors would be likely features of enzymes that remodel oomycete walls. Therefore, to focus on plant-targeted CWDEs, the following analyses are restricted to secreted proteins that lack transmembrane domains and GPI anchors.

β-1,3-glucanases. These are targeted by glycoside hydrolase families GH17 (β-1,3-glucosidase) and GH81 (β-1,3-glucanase). *Ph. infestans* and *Py. ultimum* are predicted to encode three and five extracellular GH17 enzymes, and five and three GH81 enzymes, respectively. All were expressed based on the CPM threshold of 1.0 (Fig. 6a).

About 80 and 42% of the genes were infection-induced in *Py. ultimum* and *Ph. infestans*, respectively, with the highest levels observed for both species in early tubers. Total CPM was three and ten-fold higher in *Py. ultimum* compared to *Ph. infestans* in early and late tubers, respectively. One *Py. ultimum* gene, PYU1_G013792, was expressed at very high levels, peaking in tubers at 1776 CPM or 79% of the total. In contrast, no single *Ph. infestans* gene accounted for more than 20% of total CPM.

Infection-induced enzymes from *Ph. parasitica* that target β-1,3-glucans have been identified in its interaction with lupin (Blackman et al., 2015). Such enzymes may be used to degrade plant cell walls or callose, which plants deposit in response to pathogens (Ellinger et al., 2013). Both roles are consistent with the higher expression of the enzymes in *Py. ultimum,* in which the need to disrupt cell walls may be more important and the ability to repress host defenses more limited than in *Ph. infestans*.

Cellulases. This activity is conferred primarily by β-glucosidase (GH1, GH3), endo-β-1,4-glucanase (GH5, GH6), and exo-β-1,4-glucanase (GH7) (Blackman et al., 2015). *Ph. infestans* and *Py. ultimum* are predicted to encode one and zero GH1 proteins, respectively, of which the former was expressed at CPM ≥1 in at least one growth condition; eight and three GH3 proteins, of which seven and three were expressed; zero and four GH5 enzymes, with three of the latter expressed; four and zero GH6, all of the former expressed; and one and one GH7, both expressed (Fig. 6a).

Patterns of expression between the species were very different. In early infection, total CPM was more than ten-fold higher in *Ph. infestans* than *Py. ultimum,* due largely to increased expression of enzymes in categories GH3 (PITG_17546, PITG_22095), GH6

(PITG_18388), and GH7 (PITG_06788). In contrast, in late tubers, CPM was three-fold higher in *Py. ultimum* than *Ph. infestans*. The increase in *Py. ultimum* was due mostly to two highly-expressed GH3 (PYU1_G000955, G000957) and one GH5 genes (PYU1_G012486). Although total *Ph. infestans* CPM fell nine-fold in late tubers, one GH1 (PITG_01399) and two GH3 genes (PITG_15905, 03140) were up-regulated.

Since cellulose is the primary component of the plant cell wall, the results suggest that large-scale wall saccharification is not a major feature of pathogenesis by *Py. ultimum* prior to late infection. That most cellulases are late-induced in *Py. ultimum* suggests that the wall becomes a significant carbon source only when other substrates, such as simple sugars, become limiting. The staged expression of cellulases was also reported for *Ph. parasitica* on lupin (Blackman et al., 2015). However, the higher expression of cellulases by *Ph. infestans* in its biotrophic early tuber stage compared to *Py. ultimum* was surprising, since *Phytophthora* spp. are thought to employ stealthy infection strategies that limit host damage during much of the disease cycle (Stassen et al., 2011). The early-induced cellulases in *Ph. infestans* may make focused digests in plant walls to allow haustorial development or make the wall susceptible to later digestion, rather than to immediately macerate the wall.

Hemicellulases. To fully degrade the cell wall, pathogens must digest the heteropolymeric matrix that coexists with cellulose. Enzymes participating in this include β-mannosidase (GH2), endo-1,4-β-xylosidase (GH3), endo-1,4-mannosidase (GH5), β-1,4-xylanase (GH10), and xyloglucan specific-endo-β-1,4-glucanase (GH12).

Ph. infestans encodes a plethora of these enzymes. This includes three GH3 proteins, of which two were expressed at CPM ≥1; one GH5 enzyme, which was expressed; two GH10s, both expressed; and six GH12s, of which four were expressed. Nearly half were induced strongly in early tubers compared to media, paralleling the rise in cellulases. As observed in many other functional categories, a majority (61%) of hemicellulase CPM came from a single GH12 gene, PITG_08944.

In contrast to *Ph. infestans, Py. ultimum* encodes only one GH2 and one GH3, both of which were expressed. Unlike the *Ph. infestans* enzymes, none were infection-induced, and their CPM in early tuber was only one-tenth that of *Ph. infestans.* Expression increased four-fold in late tubers, achieving levels similar to those of *Ph. infestans*.

Our results support the suggestion that *Pythium* spp. have a limited ability to degrade hemicellulase, and that host walls are a major carbon source only in late infection (Zerillo et al., 2013). The timing of expression in *Ph. infestans,* in contrast, suggests that their role is more important in early infection. This matches the conclusion of Blackman *et al*. (Blackman et al., 2015) that hemicellulose degradation by *Ph. parasitica* starts early during infection.

Pectic enzymes. Homogalacturonan, rhamnogalacturonan-I, and substituted galacturonans comprise the pectin matrix of plant cell walls (Zerillo et al., 2013). Enzymes degrading these include pectin lyase (PL1), pectate lyase (PL3), rhamnogalacturonan lyase, polygalacturonase, α-rhamnosidase, rhamnogalacturonyl hydrolase, and α-L-arabinofuranosidase. Their different specificities may control the overall rate of pectin digestion, by affecting the hydrolysis of its side chains.

Each pectin-degrading activity is encoded by multigene families in *Ph. infestans.* Seven of 12 predicted PL1 genes were expressed, as were 11 of the 16 PL3, one of two PL4, all 14 GH28, one of two GH78, both GH105, and none of the GH43 genes. Two-thirds were induced during early tubers, causing a six-fold increase in total CPM compared to media. All genes declined in late tubers, reducing aggregate CPM by 32-fold. The pectin-degrading activities of *Py. ultimum* include twelve PL1s, all of which were expressed; 14 PL3s, of which 10 were expressed; five GH28s, with three expressed; and two PL4s, both expressed. Unlike *Ph. infestans,* there were no GH78 or GH105 genes (Zerillo et al., 2013) but two GH43 genes, both of which were expressed. All of the *Py. ultimum* genes were up-regulated significantly in early tubers, and declined in late tubers.

In most cases, there was a good correlation between gene number and aggregate CPM. For example, the three and 14 GH28 enzymes expressed in *Py. ultimum* and *Ph. infestans,* respectively, yielded 177 and 1375 CPM in early tubers. Similarly, the expansion of expressed, secreted PL1 genes in *Py. ultimum* versus *Ph. infestans* to twelve from seven resulted in CPM totals of 1613 and 633, respectively. However, this proportionality was not observed for PL3, where 10 and 11 genes in *Py. ultimum* and *Ph. infestans* resulted in CPM values of 164 and 992, respectively. Contributing to this disparity was one highly expressed and tuber-induced *Ph. infestans* gene, PITG_14168, which yielded 810 CPM in early tubers.

Ph. infestans also encodes six secreted pectinesterases (CE8), all of which were expressed. Their expression patterns are not included in Fig. 6 since *Py. ultimum* is predicted to lack this activity (Zerillo et al., 2013). The *Ph. infestans* genes were each

induced in early tubers, with aggregate CPM values of 139, 1277, and 40 in media, early tubers, and late tubers, respectively. Most were shown to be up-regulated in zoospores and germinated cysts compared to media, suggesting that they anticipate plant infection (Ah-Fong et al., 2017).

Pectin-degrading enzymes are of particular interest when comparing the outcomes of infection, since while potatoes remain quite firm at the conclusion of the *Ph. infestans* disease cycle, the tubers are liquefied by *Py. ultimum*. Since pectin and pectate lyases (PL1) and polygalacturonase (GH28) were shown to be the most effective in macerating tubers (van den Broek et al., 1997), it was surprising to find that the aggregate CPM of PL1 was about three-fold higher in *Ph. infestans* in early tubers, and GH28 CPM was eight-fold higher in *Ph. infestans.* A related enigma is why *Py. ultimum* does not express any pectin methylesterases; these remove methyl groups as a first step towards exposing the polymer to other enzymes, which would facilitate liquefaction of the tuber.

Besides the enzymes described above, *Ph. infestans* and *Py. ultimum* are predicted to encode one and two GH43 arabinofuranosidases, respectively, of which only the *Py. ultimum* enzymes were expressed. These also target pectin, although hemicellulose and arabinogalactans may also be substrates. Their aggregate CPM in *Py. ultimum* rose from 1.1 to 59 in media and early tubers, respectively, before falling to 6.0 in late tubers.

Genes involved in secretion

Because most pathogenicity factors are extracellular, we examined three classes of proteins involved in the classic secretory pathway. These were SNARE proteins which mediate vesicle fusion, Rab GTPase proteins which help regulate vesicle fusion, and

orthologs of other yeast genes involved in trafficking vesicles for secretion (Schekman, 2002).

Only a few genes showed strong differential expression in *Ph. infestans* or *Py. ultimum*, and there was only minor variation between the species in aggregate CPM (Fig. S1). One subtle difference was that the total CPM of genes encoding SNARE proteins was two-fold higher during tuber infection by *Py. ultimum* compared to *Ph. infestans.* While some genes encoding SNARE or Rab GTPases in *Ph. infestans* were up-regulated during infection (*e.g.* PITG_01853, 10870, 18718), their expression was not high enough to cause a major change in aggregate CPM. A difference was observed during early infection in the aggregate CPM of orthologs of yeast secretory genes, as this was 2.5 times higher in *Ph. infestans* than *Py. ultimum*. This was, however, attributable to just two highly expressed genes encoding chaperones.

It thus appears that most genes involved in secretion are expressed constitutively. It is however possible that some proteins bypass the classical trafficking pathway. This might involve a Golgi bypass, as suggested for animal cells (Grieve and Rabouille, 2011).

.

Detoxification pathways

ABC transporters. These efflux proteins may help defend the pathogens from environmental or plant-generated toxins. *Ph. infestans* and *Py. ultimum* are predicted to encode 198 and 163 of these transporters, respectively, of which 153 and 148 were expressed at CPM ≥1 in least one tissue (Fig. 6b). More genes were induced in early tubers compared to media in *Py. ultimum* than *Ph. infestans* (34 and nine, respectively). In contrast, more were up-regulated in early versus late infection in *Ph. infestans* than *Py.*

ultimum (59 and 12, respectively). Based on aggregate CPM, the *Py. ultimum* mRNAs were more abundant in early versus late infection, while the opposite trend was seen in *Ph. infestans.* Despite having 18% fewer ABC transporter genes, total CPM in *Py. ultimum* was from 25 to 220% higher than *Ph. infestans* in the tissues, with the largest difference in early tubers.

These results suggest that ABC transporters may play more significant roles in necrotrophic growth. This may be due to a need to eliminate host toxins, which as illustrated in Fig. 1a is likely to be more severe and occur earlier in potato leak than late blight. The up-regulation of ABC transporters in *Ph. infestans* in late tubers is consistent with a shift from biotrophic to necrotrophic growth, but could also be associated with sporulation. An earlier study indicated that about 10% of *Ph. infestans* ABC transporters were up-regulated in sporangia produced on artificial media (Ah-Fong et al., 2017).

Catalases. These may aid pathogens by eliminating H_2O_2 made by the host, or by delivering peroxide to the plant (Rolke et al., 2004; Lehmann et al., 2015). *Ph. infestans* and *Py. ultimum* are predicted to encode five and six catalases, respectively, of which five and four were expressed in least one tissue with CPM ≥1.0 (Fig. 6b). Their patterns of expression were opposite of those of ABC transporters, as only *Ph. infestans* genes were significantly up-regulated in early tubers versus media. In addition, total catalase CPM was about ten-fold higher in *Ph. infestans* compared to *Py. ultimum* at all tuber stages. This may suggest that peroxide elimination is more critical for *Ph. infestans.*

Interestingly, 89% of catalase mRNA in *Ph. infestans* in tubers was from a single gene, PITG_07143, which along with PITG_05579 and PYU1_G013215 from *Py.*

ultimum have N-terminal secretion signals. Secreted catalases have also been described in fungal pathogens (Tanabe et al., 2011). While the two *Ph. infestans* genes were up-regulated 60-fold in early tubers compared to media (1012 vs. 8.7 aggregate CPM), PYU1_G013215 was down-regulated six-fold (178 vs. 32 CPM). All were expressed at low levels in late tubers.

The induction of *Ph. infestans* catalases during early infection is consistent with an observation that a catalase was induced early during infection of tobacco by *Ph. Nicotianae* (Blackman and Hardham, 2008). That the most infection-induced and most highly-expressed *Ph. infestans* catalase is secreted may suggest that its role is to eliminate H_2O_2 before it reaches the pathogen, or limit peroxide signaling within the plant (Gechev and Hille, 2005).

While catalase CPM in early tuber was low in *Py. ultimum*, that species may be able to eliminate intracellular peroxides using other enzymes such as glutathione peroxidase.

NADPH oxidase. This generates the toxic compound superoxide. *Ph. infestans* and *Py. ultimum* are predicted to encode 12 and 21 NADPH oxidases, respectively, of which 10 and 19 were expressed at CPM ≥1.0 (Fig. 6b). As in animals, all of the predicted expressed proteins contain transmembrane domains and therefore this is not a secreted activity. Their aggregate CPM was higher in early tubers compared to media in both *Ph. infestans* and *Py. ultimum*, although more *Py. ultimum* genes were infection-induced. Total CPM was four times higher in *Py. ultimum* than *Ph. infestans* in early tuber.

These results are consistent with a model in which the pathogens produce superoxide to damage host enzymes, cell walls, or trigger programmed cell death (Barna et al.,

2012). Similar activities have been proposed for necrotrophic fungi such as *Sclerotinia sclerotiorum* (Kim et al., 2011). This is consistent with the much higher CPM of the enzymes in *Py. ultimum. With Ph. infestans,* the effect of superoxide may be localized or its ability to trigger host cell death may be delayed by effectors during the biotrophic phase.

Glutathione S-transferase (GST). This inactivates toxins by conjugating them to glutathione. *Ph. infestans* and *Py. ultimum* are predicted to encode 17 and 19 GSTs, respectively, of which 12 and 15 were expressed at CPM ≥1.0 (Fig. 6b). A similar fraction of genes were induced in both species during early infection compared to media, however the aggregate CPM was two times higher in *Py. ultimum* than *Ph. infestans.* This was because several *Ph. infestans* were down-regulated in early infection compared to media, and since one infection-induced *Py. ultimum* gene (PYU1_G000253, predicted to be cytoplasmic or peroxisomal) was expressed highly, accounting for 40% of total CPM. A reversal of this expression pattern occurred during late infection, when total CPM was two times higher in *Ph. infestans.*

These results are consistent with a role of GST in necrotrophic growth, *i.e.* all timepoints in *Py. ultimum* and the late timepoint in *Ph. infestans*. GST may protect the pathogens by reducing endogenous or exogenous H_2O_2, other toxins of plant origin, and fatty acid peroxides (Calmes et al., 2015).

Nutrient uptake

Pathogens acquire most nutrients from the host using transporter proteins. The two pathogens allocate roughly the same proportion of their genes to these transporters, 2.6% for *Ph. infestans* and 3.0% for *Py. ultimum*. Overall, the number induced by at least three-fold in early tubers compared to media was slightly higher in *Py. ultimum* than *Ph. infestan*s (11.4% vs. 9.7%), while more were induced in late compared to early tubers by *Ph. infestans* (26.8% vs. 9.6%). Families showing the most changes are discussed in the next sections.

Amino acid/auxin permeases (AAAP). *Ph. infestans* is predicted to encode 57 AAAPs compared to 49 in *Py. ultimum* (Fig. 7). Based on the CPM cut-off of 1.0, 54 *Ph. infestans* and 40 *Py. ultimum* AAAP genes were expressed in at least one of the conditions. While a similar number of genes in each species were up-regulated in early tubers compared to media (about 10%), the aggregate CPM in *Ph. infestans* was twice that of *Py. ultimum.* This was largely due to two highly expressed genes, PITG_20230 and PITG_12808. During the transition from early to late tubers, 28% and 8% of the *Ph. infestans* and *Py. ultimum* genes were up-regulated, respectively. Nevertheless, the aggregate CPM in late tuber was slightly higher in *Py. ultimum* than *Ph. infestans*, 1974 vs. 1470.

Amino acid/polyamine/organocation family (APC). Both species encode similar numbers of these transporters, 27 for *Ph. infestans* and 30 for *Py. ultimum*, with 24 and 29 expressed at CPM ≥1 in at least one tissue, respectively (Fig. 7). Only one and two

genes were up-regulated in *Ph. infestans* and *Py. ultimum* in early tubers compared to media, while ten and three were induced in late versus early tubers. Nevertheless, aggregate CPM levels in tubers were 44-250% higher in *Py. ultimum.* A large contributor was a single highly expressed gene, PYU1_G005219, which accounted for 56% of total transcripts in late tubers.

The APC and AAAP families are believed to represent the major plasma membrane transporters for amino acids in most organisms. Changes in expression during infection could be related to the levels of extracellular amino acids caused by effector action in the case of *Ph. infestans*, or leakage from cells during *Py. ultimum* necrotrophy (Solomon and Oliver, 2001).

Dicarboxylate/ amino acid: cation symporter (DAACS). Substrates of these also include amino acids, although these proteins often accumulate in both intracellular organelles and plasma membrane. *Ph. infestans* and *Py. ultimum* are predicted to encode 11 and 13 DAACS proteins, respectively, of which 11 and nine were expressed with CPM ≥1 in at least one tissue (Fig. 7). Only one gene was induced in early tubers versus media per species. While tissue-specific expression patterns in the two species were similar, this was not true for aggregate CPM which was five to 14 times higher across all tissues in *Ph. infestans.* The peak expression in early tuber was due to a single gene, PITG_09295, which accounted for 83% of total CPM.

Nitrate and ammonium transporters. These present alternative routes for acquiring nitrogen. *Ph. infestans* encodes four predicted nitrate transporters, of which two had a

CPM ≥1. The heatmaps in Fig. 7 indicate that these are expressed more in late than early tubers. However, based on CPM these represent very low transcript levels, two orders of magnitude lower than what occurs in potato and tomato leaves (Abrahamian et al., 2016). Nitrate levels are generally high in leaves and low in tubers.

The expression of nitrate transporters was very different in *Py. ultimum.* Its four predicted transporters were all expressed, with aggregate CPM 100-times that of *Ph. infestans* in early tuber. This was largely due to PYU1_G001247, which at 695 CPM accounted for 95% of total CPM. In media and late tubers, the aggregate CPM of *Ph. infestans* and *Py. ultimum* were similar at 46 and 111 CPM, respectively.

The expression of ammonium transporters (AMTs) was also strikingly different between the species. However, in this case the aggregate CPM in *Ph. infestans* was 100-times higher than *Py. ultimum* in early tubers. Of five predicted AMTs in *Ph. infestans,* two were expressed only in tubers and three in artificial media. Similarly, one of the two predicted *Py. ultimum* AMTs was largely tuber-specific and the other media-specific.

The opposing patterns of expression of these transporters may reflect a greater reliance of *Py. ultimum* on nitrate, and of *Ph. infestans* on ammonium. In soil, nitrate is more abundant and thus more useful to *Py. ultimum* during its saprophytic stage. In contrast, ammonium occurs throughout plant tissues and should be more useful to the more host-dependent species, *Ph. Infestans* (Abrahamian et al., 2016). Alternatively, the main role of the transporters may be for efflux or to alkalinize and weaken host tissues, not to acquire nitrogen for new biomass (Alkan et al., 2008).

SWEET transporters. *Ph. infestans* and *Py. ultimum* are predicted to encode 32 and 28 SWEET transporters, of which 28 and 20 were expressed at CPM ≥1, respectively (Fig. 7). In plants, these participate in the uptake and efflux of mono- and disaccharides (Chen, 2014). Presumably these are used for sugar uptake by the pathogens.

In *Ph. infestans,* the aggregate CPM was similar in all tissues, although a few genes were induced during late infection. In contrast, the aggregate CPM in *Py. ultimum* was three-fold higher in early infection than the other stages. Despite the fairly large size of the SWEET family in *Py. ultimum,* 95% of total CPM in early tubers came from a single gene, PYU1_G003519. This belongs to a cluster of 10 genes, of which eight were not expressed in any tissue. *Ph. infestans* SWEET genes also occurred in clusters, with a majority also lacking expression in the tissues examined.

MFS transporters. This superfamily of secondary transporters was originally thought to only participate in sugar uptake, but was shown later to carry a broader spectrum of substrates (Rolke et al., 2004). MFSs represent nearly a quarter of the nutrient transporters encoded by *Ph. infestans* and *Py. ultimum*, with 112 and 118 genes in the two species, respectively. Of these, 90 and 88% were expressed at CPM ≥1, respectively (Fig. 7).

A similar number of *Ph. infestans* and *Py. ultimum* MFS genes were induced during early infection compared to media, although the aggregate CPM was fairly constant in all tissues and similar between species. Three genes in each species were expressed specifically in tubers compared to media, with half predicted to be hexose transporters (PYU1_G009043 and G009044, PITG_12998).

Glycoside-pentoside-hexuronide family (GPH). This family is believed to be distantly related to the MFS group. In symport with a monovalent cation, they transport a range of carbohydrates that are mostly but not exclusively glycosides. *Ph. infestans* and *Py. ultimum* are predicted to have eight and nine GPH genes, of which four and seven were expressed at CPM ≥1 (Fig. 7). Apart from one *Ph*. *infestans* gene induced in early tuber (PITG_06336) and two *Py. ultimum* genes induced in late tuber (PYU1_G002035, G002036), most were expressed constitutively.

The aggregate CPM values of GPH proteins were 10 to 100 times lower than MFS and SWEET transporters. It thus appears that the MFS and SWEET families bear the primary responsibility for sugar uptake in both pathogens. For *Ph. infestans,* this may entail competition with plant apoplastic transporters, or help reduce leakage of sugars from the pathogen to the apoplast, which may trigger plant defenses (Lingner et al., 2011). Some of the proteins may also serve as sugar sensors and not transporters, as observed in *Colletotrichum* (Lingner et al., 2011).

Global analysis of variation in ortholog expression levels

Above, we described interspecific differences in the level and stage-specific patterns of transcription of genes in functional categories relevant to plant colonization. Only in select cases did family expansions explain the variation in mRNA levels. When all *Ph. infestans* and *Py. ultimum* genes were examined, 21% of families were found to vary in size by two-fold or more; 903 families had >2-fold more genes in *Ph. infestans* and 553 had >2-fold more in *Py. ultimum* (Fig. S2). However, as shown in Fig. S3, the correlation between family size and expression level was weak in *Py. ultimum* (*e.g.* R=0.42 in rye media) and nonexistent in *Ph. infestans* (R=0.03). This is because many family members were often expressed at low levels, as illustrated for polyphenol oxidases in Fig. 3 and NPP1 proteins in Fig. S1.

To better address the extent of transcriptional remodeling between *Ph. infestans* and *Py. ultimum,* we compared orthologs genome-wide. To minimize error, genes were defined as orthologs only if consistent assignments of orthology were obtained using the OrthoMCL and InParanoid pipelines, and the reciprocal best hit approach. To avoid assigning Illumina reads to the wrong gene, genes were excluded if they contained regions >98% identical to another family member. This resulted in the identification of 5559 "high-confidence" ortholog pairs, which included 3945 single-copy genes and 1614 belonging to families.

Frequent divergence in expression level, measured as FPKM (fragments per kilobase per million mapped reads), was observed between *Ph. infestans* and *Py. ultimum* orthologs in all tissues (Fig. 8). For example, more than three-fold divergence was observed in early tuber between 1732 orthologs, or 34% of genes expressed at that stage.

Slightly less divergence occurred in artificial media, and slightly more in late samples, the latter due possibly to the effects of sporulation. The differences between species were not attributable to experimental variation, because only minor changes were observed between biological replicates. The degree of divergence of single-copy genes and those belonging to families was equivalent. Despite the significant divergence of individual genes, the expression levels of most genes were fairly similar, with Pearson's correlation coefficient calculated at 0.88, 0.87, 0.91, 0.71 and 0.80 in early rye, early pea, early tuber, late rye, and late tuber, respectively. There was no significant correlation in randomly shuffled data (R=0.01 to 0.05).

Global analysis of variation in stage-specific patterns of orthologs

Besides divergence in mRNA level, divergence was observed in tissue-specific patterns of expression. As illustrated in Fig. S4a for a comparison between early tuber and media, 833 ortholog pairs were differentially expressed in only one species and 52 in opposite directions, *i.e.* up-regulated in one species and down-regulated in the other. The definition of "differentially expressed" used here is a three-fold change. One interesting example involved secreted catalases. While *Ph. infestans* gene PITG_07143 was induced 164-fold in early tuber compared to media, its *Py. ultimum* ortholog, PYU1_G013215, was down-regulated six-fold. Another example involved acid phosphatases. While *Py. ultimum* gene PYU1_G000880 was induced 16-fold in early tuber, its *Ph. infestans* ortholog, PITG_17872, was down-regulated eight-fold. The other divergently expressed genes are listed in Table S5.

A comparison of ortholog pairs in late and early tubers showed more variation, as 2012 were up or down-regulated only in one species, and 69 were up-regulated in one species but down-regulated in the other (Fig. S4b, Table S6). Orthologs that were up-regulated in late tubers only in *Py. ultimum* had diverse roles that included amino acid catabolism, gluconeogenesis, and dehydrogenase reactions. Genes that were up-regulated only in *Ph. infestans* often encoded flagella-associated proteins. Genes that were up-regulated in one species but down-regulated in the other included many involved in sulfate metabolism, including sulfate transporters and sulfite reductase.

Some of these genes, such as the catalases, fall into categories shown in earlier sections to vary between the two oomycetes. However, this analysis draws a distinct conclusion. Patterns of gene expression have changed not only due to gene gain or loss, or family expansions, but due to the transcriptional remodeling of orthologs. While this may be due to changes in promoters or their cognate transcription factors, it is possible that each species is responding to environmental changes attributable to their biotrophic and necrotrophic lifestyles, such as levels of toxic plant compounds.

Conservation of expression and sequence are positively correlated

Most studies have found that the divergence of expression between orthologs is largely independent of protein similarity (Khaitovich et al., 2005; Dutilh et al., 2006; Tirosh and Barkai, 2008), although studies of mouse and human have yielded contradictory results (Yanai et al., 2004; Jordan et al., 2005). To test this in oomycetes, a genome-wide analysis was performed by graphing the FPKM ratio of *Ph. infestans* and *Py. ultimum* orthologs versus amino acid identity (Fig. 9). This showed that variation in

expression level increased with reduced amino acid identity in all growth conditions, for both single-copy genes and members of families. Our strict rules for assigning orthologs may have, however, eliminated many genes having more divergent expression.

Variation in orthologs based on functional classification

We also tested for correlations in the expression of orthologs categorized by protein function (Fig. 10). The results indicated that orthologs associated with most core cell functions had well-correlated expression values in all conditions. These included genes with roles in DNA replication, ion channels, signal transduction, and amino acid metabolism with the exception of tryptophan and histidine synthesis. Negative correlations were observed for NADPH oxidase, secreted phosphatases, and others. In categories such as CWDEs, secreted proteases (but not all proteases), cytochrome P450s, and glutaredoxin, the divergence was most pronounced during tuber infection. As in prior analyses, the differences may be due to transcriptional remodeling of orthologs or specific responses of each species to environmental changes attributable to their lifestyles.

Family-wide analysis of transcriptional remodeling

The frequency of shifts in the expression of orthologs was addressed by studying 55 OrthoMCL groups in which both species had equal numbers of paralogs and phylogenetic analysis supported the ortholog assignment. In 76% of the families, expression levels and patterns were well-conserved within each ortholog pair. An example is in Fig. 11a, which portrays a small family of calcineurin subunit genes. One ortholog pair was expressed at

relatively high levels in all tissues (PITG_18712, PYU1_G009008), while the other two pairs shared low expression. Fig. 11b illustrates a second, larger family encoding P-ATPases, where expression levels and patterns within each ortholog pair are also generally well-conserved.

In contrast to these examples, expression patterns or levels were conserved poorly in 24% of gene families. An example is in Fig. 11c, which shows a family of metal ion transporters. While one ortholog pair (PITG_00746, PYU1_G0144338) shows high expression in each tissue, expression of the other orthologs has been shuffled. In one ortholog pair, PITG_02263 shows moderate expression while zero expression was seen for PYU1_G014841. In a second ortholog pair, PYU1_G000638 and PITG_03105 show moderate and zero expression, respectively. The ortholog assignments in this family are supported by synteny. The transcriptional network that regulates the family thus appears to have become rewired during evolution.

Genome-wide analysis of pseudogenization

Pseudogenes in families shared by Ph. infestans and Py. ultimum. Many genes in both oomycetes lacked a CPM >1.0 in any of the six growth conditions, and were thus possible nontranscribed pseudogenes. This threshold was selected to avoid counting spurious transcripts, and corresponded to about 20 reads per average gene in our study. In mammals, a CPM of 1.0 equals about 1 transcript per nucleus (Kellis et al., 2014). Excluding RxLR and CRN genes, 25 and 15% of *Ph. infestans* and *Py. ultimum* genes in the functional groups described earlier in this paper were unexpressed. The difference

between species was significant at $p=10^{-6}$ by Fisher's exact test. Possibly contributing to the greater fraction of unexpressed genes in *Ph. infestans* was its increased propensity towards expansion of gene families (Fig. S2). This is commonly associated with pseudogenization or regulatory subfunctionalization, in which genes acquire new expression patterns (Skamnioti et al., 2008).

As an initial test of whether genes with low CPM had undergone pseudogenization, regulatory subfunctionalization, or were simply expressed under growth conditions not represented so far in this study, we examined the *NPP1* family of *Ph. infestans.* Earlier we described how the family had expanded from seven genes in *Py. ultimum,* all expressed, to 55 in *Ph. infestans,* of which 33 had CPM <1. By examining RNA-seq data from five additional developmental stages (purified asexual sporangia, sporangia chilled to differentiate zoospores, swimming zoospores, germinating cysts making appressoria, sexual spores) and a fungicide treatment (0.5 µg/ml metalaxyl), 12 additional genes were deemed to be expressed. The remaining 20 *NPP1* genes are thus candidate nontranscribed pseudogenes. All but one had matching sequences in an assembly of genomic reads from strain 1306.

Next, we performed a genome-wide assessment of whether pseudogenization was associated with the size of a gene family. This focused on genes that contained orthologs in both species, a strategy that would exclude false gene models. By checking for expression in all 12 growth conditions or life stages, we identified 938 candidate pseudogenes. A gene was more likely to be a pseudogene if it belonged to a large family or one that was expanded in *Ph. infestans* compared to *Py. ultimum* (Fig. 12a). A gene was also more likely to be a pseudogene if it was oomycete-specific (Fig. 12b). The

difference was significant at $p=10^{-96}$ by Fisher's exact test. Of the 938 candidates defined by mapping RNA-seq reads to the T30-4 reference genome, nine lacked matching sequences in the genome assembly of strain 1306, and 32 showed enough divergence to impede the mapping of the reads. Therefore, the number of candidate nontranscribed pseudogenes in families shared with *Py. ultimum* is 897.

We also checked for evidence of regulatory subfunctionalization in families common to *Ph. infestans* and *Py. ultimum* by checking for genes that were expressed only during asexual sporulation, zoosporogenesis, or spore germination in *Ph. infestans*. Fitting this criteria were 1574 genes, although it is difficult to claim definitively that these are not sporulation-induced in *Py. ultimum* since whether the latter makes spores is controversial (Benhamou et al., 1998; Kamoun et al., 2015)

. Genes in *Ph. infestans* were more likely to be associated with spores if they had more than two genes per family ($p=10^{-47}$ by Fisher's exact test), or were in a family that was expanded in *Ph. infestans* relative to *Py. ultimum* ($p=10^{-69}$). This is consistent with the occurrence of subfunctionalization and neofunctionalization in other taxa (Skamnioti et al., 2008; Baroncelli et al., 2016).

Other pseudogenes. Also lacking expression in the 12 growth conditions and life stages were 738 *Ph. infestans* genes that lacked orthologs in *Py. ultimum.* Of these, 89 were predicted to encode RxLR proteins and 243 had matches to proteins from other *Phytophthora* spp. in BlastP with $E<10^{-20}$. The remainder, 340, may be false gene models. Eliminating those, there are 1295 candidate pseudogenes in *Ph. infestans* strain 1306. These are listed in Table S7.

By combining *Py. ultimum* RNA-seq data from this study with samples from a prior report that focused on heat and chemical stress (Levesque et al., 2010), we identified 1565 gene models that lacked CPM ≥1.0. By filtering for genes containing matches in other Pythium species, a tentative conclusion is that 1203 are candidate pseudogenes and the rest may be false gene models.

Genomic architecture and pseudogenes. Oomycete genomes contain gene-dense and gene-sparse regions. Since gene evolution is reportedly accelerated in the latter (Dong et al., 2015), the genomic distribution of the pseudogene candidates of *Ph. infestans* was analyzed. The average distance between pseudogenes and neighboring genes was 18% more than for expressed genes. Pseudogenization of sequences in the gene-sparse regions was enhanced, but so was the generation of pseudogenes from tandemly repeated families in gene-dense regions. Twenty percent of pseudogene candidates were adjacent to an expressed member of the same family, and 30% were within four genes of an expressed member of the family.

Species-specific genes

A sign of genome evolution other than divergence of orthologs or gene family expansions are species-specific genes, which may originate through gene loss or gain, including horizontal gene transfer. The RxLR and Crinkler families are one well-studied example in oomycetes, which are abundant in *Phytophthora* but not *Pythium* genomes (Levesque et al., 2010). Little data are available about other species-specific genes. We therefore identified and studied the expression patterns of genes that lacked relatives in

the other species, using a maximum *E*-value threshold of 10^{-25} in BlastP and a minimum protein size of 50 amino acids.

Genes specific to Ph. infestans. These numbered 2801, excluding RxLRs and Crinklers (Table S8). Expressed based on the CPM threshold of 1.0 were only 1209 genes, which suggests that many may be false gene models. Of the expressed genes, 136 encoded secreted proteins. Up-regulated in early and late tuber stages, respectively, were 9% and 38% of genes encoding nonsecreted proteins, and 35% and 30% of those encoding secreted proteins.

Most of the predicted secreted proteins represent categories described previously including protease and glucanase inhibitors, small cysteine-rich proteins, cutinase, and berberine bridge enzymes (Raffaele et al., 2010; Blackman et al., 2015; Damasceno et al. 2008). Contrary to the findings of Ospina-Geraldo *et al*. who examined leaf infections (Ospina-Giraldo

(Ospina-Geraldo et al., 2010), we found that cutinases were induced only in late tubers. The discrepancy could be due to the absence of a cuticle in tubers, or the possibility of suberization in our tuber slices (Kolattukudy and Dean, 1974). Like some fungal cutinases, the *Ph. infestans* enzymes might be able to degrade both cutin and suberin (Kontkanen et al., 2009).

Several intracellular metabolic enzymes were also specific to *Ph. infestans,* including gluconolactonase, sorbitol dehydrogenase, and pyruvate phosphate dikinase. All were highly expressed *in planta*. The latter, which acts at the final step of glycolysis, has a lower free energy than the alternate pyruvate kinase enzyme, and may thus facilitate

gluconeogenesis by being more reversible (Mertens, 1993). The sorbitol dehydrogenases might contribute to osmotic balance or quench reactive oxygen. The glucolactonase might contribute indirectly to the saccharification of cellulose to glucose by eliminating a known inhibitor of β-glucosidase (Ng et al., 2011).

Prior studies identified an early infection-specific gene, *Hmp1,* which encodes a protein of unknown function bearing a signal peptide and transmembrane domain targeted to haustoria (Avrova et al., 2008). A search of the *Ph. infestans*-specific genes for proteins with the same expression pattern and structural features identified three: PITG_06212, PITG_19140, and PITG_16084. Whether these might also be haustorial remains to be determined, although each lacked sequence similarity to Hmp1.

Genes specific to Py. ultimum. Identified were 452 genes, of which 322 were transcribed at CPM ≥1 (Table S8). Sixty-one of the total are predicted to encode secreted proteins, of which 43 were expressed. Up-regulated in early and late tuber stages were 12% and 8.4% of the 322 genes, and 30% and 14% of expressed secreted genes, respectively.

One gene, PYU1_G000106, encodes a protein not described previously in an oomycete (Fig. S5). The 588-amino acid protein exhibits a strong match (E=10^{-42}) to the 394-amino acid membrane-attack complex/perforin (MACPF) domain. In plants and animals, MACPF proteins form pores in cell membranes, leading to programmed cell death in development or defense against pathogens (Osinska et al., 2014; Morita-Yamamuro et al., 2005). In apicomplexans, the proteins are secreted virulence factors (Wirth et al., 2014). We find MACPF orthologs in most *Pythium* spp., but no other oomycete. The *Py. ultimum* gene contains a predicted signal peptide (SignalP score of

0.94), consistent with its affinity in phylogenetic analyses to apicomplexan proteins. The *Py. ultimum* gene is barely expressed in media (averaging 0.6 CPM), but rises to 1.9 in early tuber and 10 CPM in late tuber. We propose that its role is to lyse host cells to liberate nutrients. Its absence from *Ph. infestans* suggests that it is a hallmark of the necrotrophic lifestyle in peronosporalean oomycetes.

One metabolic enzyme specific to *Py. ultimum* was the starch-degrading protein glucoamylase. Full utilization of this plant (but not oomycete) storage molecule requires α-amylase, which cleaves α-1,4 bonds in starch, and glucoamylase, which breaks its α-1,6 linkages. *Ph. infestans* and *Py. ultimum* encode one and three α-amylases, respectively, but only *Py. ultimum* encodes glucoamylase. This explains why starch is an inferior carbon source for *Phytophthora* (Davies, 1959), but a good substrate for *Pythium* (Zerillo et al., 2013). Glucoamylase is encoded by a cluster of four genes in *Py. ultimum*, all expressed and encoding secreted proteins. Their aggregate CPM averaged 124 CPM in media, and rose from 65 in early tuber to 331 in late tuber. The α-amylases of *Py. ultimum* showed a much greater increase during infection, having 1.5 CPM in media, 8.0 in early tuber, and 99 in late tuber. In contrast, the *Ph. infestans* α-amylase was 0.8 CPM in media, 0.2 in early tubers, and 4.8 CPM in late tubers. Thus, it seems that the superior ability of *Py. ultimum* to use starch is due to its gain of glucoamylases (or loss from *Ph. infestans*), three-fold expansion of its α-amylase family relative to *Ph. infestans,* and transcriptional remodeling that increased the CPM of each of its α-amylase genes by about five-fold compared to single *Ph. infestans* gene.

Another feature of metabolism unique to *Py. ultimum* is the presence of N-acetylglucosamine-6-phosphate deacetylase (PYU1_G006885) and glucosamine 6-

phosphate isomerase (PYU1_G006886). These convert chitin into glucose-6-phosphate, which can enter glycolysis. Both are expressed in media and tubers, with the latter doubling in CPM during late infection. Interestingly, the two genes are neighbors in *Py. ultimum* and five other members of the genus. They are not linked in ortholog-containing eukaryotes (fungi and animals), but are often adjacent in bacteria (Dereuse et al., 2013). There is however poor support for horizontal gene transfer from bacteria. Orthologs are found in *Blastocystis*, which like oomycetes belongs to the stramenopile kingdom. However, the genes are absent in other stramenopiles including diatoms and saprolegnian oomycetes, and the *Blastocystis* genes do not form a strong cluster with the *Pythium* genes in phylogenetic analysis. While chitin is not made by potato and is but a minor component (<1%) of *Pythium* cell walls, it comprises a major fraction of soil organic matter (Melida et al., 2013; Clemmensen et al., 2013). The presence of these enzymes may reflect the adaptation of *Pythium* to growth as a saprophyte as well as plant pathogen.

CONCLUSIONS

This study had dual goals: to raise our understanding of differences between biotrophic and necrotrophic growth, and of how the transcriptomes of closely related oomycetes with divergent lifestyles evolved. Differences between *Ph. infestans* and *Py. ultimum* included species-specific genes plus variation in gene family sizes and expression profiles of orthologs. Expansions in families were generally not reflected in aggregate transcript levels, due to divergence in promoter activity or epigenetics. At its extreme, changes within promoters pseudogenized many genes, as reported within gene

families in non-oomycetes. In some cases, changes in promoters may have shifted activity to life stages not shared between *Ph. infestans* and *Py. ultimum.* This phenomenon, known as regulatory subfunctionalization (Skamnioti et al., 2008), could have altered the binding specificity of the cognate transcription factors or the regulation of those factors, in addition to the promoters themselves. Lévesque *et al*. noted that gene orders in *Py. ultimum* and *Ph. infestans* were often inverted*,* which could affect their 5' regulatory regions (Lévesque *et al*., 2010).

Several trends relevant to the issue of biotrophic versus necrotrophic growth were seen in our data. One is that more genes were expressed constitutively in *Py. ultimum* than *Ph. infestans.* This suggests that as a soil-borne necrotroph, *Py. ultimum* goes "full speed ahead" with its pathogenic program, which may help it compete with other microbes. In contrast, *Ph. infestans* deploys its genes in stages. This balances its expression of effectors that suppress plant defenses with proteins such as CWDEs that reduce plant cell integrity. While our results indicate that variation in the timing and level of expression of orthologs may underlie much of the distinctions between biotrophy and necrotrophy, some of this could be secondary responses. For example, the higher expression of ABC and SWEET transporters in *Py. ultimum* may be responses to phytotoxins or nutrients released during necrotrophy, respectively. Whether *Ph. infestans* truly becomes necrotrophic late in infection is controversial; hyphae within necrosing host lesions are usually vacuolated, and sometimes the first macroscopic sign of infection is sporulation (Fry et al., 2013). Nevertheless, late in infection *Ph. infestans* does up-regulate *NPP1*, a necrosis-inducing protein, which in striking contrast is expressed at its highest levels in *Py. ultimum* in early tubers.

The NPP family also provided evidence of selection on gene families, influenced by the biotrophic or necrotrophic lifestyles. In *Py. ultimum,* nearly all members of its expanded family maintained its expression pattern and predicted ability to cause host necrosis. In contrast, in *Ph. infestans* the expansion was followed by pseudogenization or diversification of most proteins to non-necrosing forms, which would support its biotrophy.

Species-specific genes and expanded families also appear to contribute to lifestyle. The presence of glucoamylases only in *Py. ultimum,* for example, may reflect its preference for degrading simple carbohydrates. In contrast, *Phytophthora* is better-equipped to digest complex polysaccharides due to its larger complement of cellulases and hemicellulases. Although not the focus of this study, the RXLR genes of *Ph. infestans* help suppress the host necrosis that is a feature of the *Py. ultimum* interaction (Stassen and Van den Ackerveken, 2011). Conversely, proteins specific to *Py. ultimum* such as MACPF/perforin may accelerate host death. The loss of this ancient cell-lysing protein from *Ph. infestans* is consistent with the theory that biotrophy is more advanced than necrotrophy (Lewis, 1957).

METHODS

Growth of pathogens

Py. ultimum var. *ultimum* was isolated from a potato farm in San Jacinto, CA and maintained at 21ºC in the dark on half-strength V8 medium containing 1.5% agar. The species assignment was verified by ITS sequencing. *Ph. infestans* strain 1306, isolated from tomato in San Diego, CA, and later shown to be also pathogenic on potato, was maintained at 18ºC on rye-sucrose agar.

For expression studies on artificial media, rye-sucrose and pea broth (clarified by centrifugation) were inoculated with 10^4/ml sporangia of *Ph. infestans* or 2-mm plugs from the edges of *Py. ultimum* cultures. Cultures were harvested at "early" and "late" growth stages, which corresponded to roughly 33% and 100% coverage of the liquid surface, respectively. These were 3-day (nonsporulating) and 5-day (sporulating) cultures of *Ph. infestans* and 1.5 and 3-day cultures of *Py. ultimum.* While hyphal swellings considered to resemble sporangia were detected in *Py. ultimum,* their density was low (<5%) compared to that of *Ph. infestans* sporangia and did not increase with further incubation.

Plant infections

Tubers (cv. Russet Burbank) were washed in tap water, immersed in 10% (v/v) household bleach for 15 min, rinsed in distilled water, cut into 2 mm slices using a mandoline, rinsed in water, and blotted dry. The slices were then placed on a metal rack 8-mm above moist paper towels in a box with a tight-fitting lid.

For inoculations with *Ph. infestans,* zoospores were released from 8-day cultures, adjusted to 5 x10^5/ml, and about 0.2 ml was spread on the upper surface of each tuber slice using a rubber policeman. Slices were kept at 18ºC in the dark and frozen in liquid nitrogen after 1.5, 2.5, and 4 days. Since *Py. ultimum* does not readily produce zoospores, infections were performed by placing a 2-mm plug from the growing edge of a 2-day V8 agar culture in the center of the slice. After 1.5 days, tuber tissue was excised as described in Results. For each sample, multiple tuber slices were pooled, with biological replicates (triplicates) performed in separate weeks.

Gene annotation and ortholog analysis

Although annotated gene sets for the two pathogens were available, to allow for valid comparisons these were reanalyzed as described below. The number of genes reported for each functional category may therefore vary slightly from other published studies. Functional annotations including GO terms were obtained by searching the SwissProt and PANTHER databases, using an *E*-value cut-off of 10^{-10}. Proteins predicted by both SignalP v4.1 and TargetP v1.1 (Emanuelsson et al., 2007) to contain signal peptides were classified as secreted, excluding those predicted by TMHMM (Krogh et al., 2001) to contain a transmembrane domain after the signal peptide cleavage site or, in the case of CWDEs, predicted by PredGPI (Pierleoni et al., 2008) to have GPI anchors.

Orthologs were identified by using results from OrthoMCL (Li et al., 2003), InParanoid (Ostlund et al., 2010) using default settings, and reciprocal best hit using a minimum *E*-value in BlastP of 10^{-20}; an assignment of orthology was made only when all three methods gave consistent results. To reduce errors resulting from erroneous gene

models, ortholog pairs that varied in size by more than 50% were excluded from the expression analyses presented in Results. Also excluded from the analyses were genes that had ≥98% DNA identity with another gene, to reduce errors in assigning RNA-seq reads to the correct paralog. Phylogenetic analyses were performed using alignments generated by Muscle, followed by tree-building with PhyML or BioNJ.

RNA-seq analysis

RNA was obtained by grinding tissue to a powder under liquid nitrogen, followed by extraction using Sigma and Agilent Plant RNA kits for the mycelia and tubers, respectively. A preliminary assessment of disease progression was performed by separating RNA on agarose gel and comparing the ratio of plant to pathogen RNA, taking advantage of the fact that oomycete rRNA is larger than plant rRNA. RNA quality was then assessed using an Agilent 2100 Bioanalyzer.

Indexed libraries for sequencing were then prepared using the Illumina Truseq kit v2. Paired-end libraries were quantitated by Qubit analysis, multiplexed and sequenced on an Illumina HiSeq2500, except for the 1.5-day tuber sample which was sequenced on an Illumina NextSeq500. Reads passing the quality filter were aligned to each pathogen's genome using Bowtie 2.2.5 and Tophat 2.0.14, allowing for one mismatch. This used the reference genomes of *Ph. infestans* isolate T30-4 and *Py. ultimum* isolate DAOM BR144 (Levesque et al., 2010; Haas et al., 2009).

Expression and differential expression calls were made with edgeR using TMM normalization, a generalized linear model, and false discovery rate (FDR) calculations based on the Benjamini-Hochberg method (Robinson et al., 2010). Hierarchical

clustering, heatmap generation, and PCA analysis were performed using Partek Genomics Suite. GO term enrichment analysis was performed using the GOHyperGAll script (Horan et al., 2008) with a FDR cut-off of 10^{-3}.

References:

Abrahamian M, Ah-Fong AM, Davis C, Andreeva K, Judelson HS. Gene expression and silencing studies in *Phytophthora infestans* reveal infection-specific nutrient transporters and a role for the nitrate reductase pathway in plant pathogenesis. PLoS Pathog. 2016;12:e1006097.

Adhikari BN, Hamilton JP, Zerillo MM, Tisserat N, Levesque CA, Buell CR. Comparative genomics reveals insight into virulence strategies of plant pathogenic oomycetes. PLoS One. 2013;8:e75072.

Ah-Fong AM, Kim KS, Judelson HS. RNA-seq of life stages of the oomycete *Phytophthora infestans* reveals dynamic changes in metabolic, signal transduction, and pathogenesis genes and a major role for calcium signaling in development. BMC Genomics. 2017;18:198.

Alkan N, Ruhr R, Sherman A, Prusky D. Role of ammonia secretion and pH modulation on pathogenicity of *Colletotrichum coccodes* on tomato fruit. Mol Plant Microbe Interact. 2008;21:1058-1066.

Avrova AO, Boevink PC, Young V, Grenville-Briggs LJ, van West P, Birch PR, Whisson SC. A novel *Phytophthora infestans* haustorium-specific membrane protein is required for infection of potato. Cell Microbiol. 2008;10:2271-2284.

Barna B, Fodor J, Harrach BD, Pogany M, Kiraly Z. The Janus face of reactive oxygen species in resistance and susceptibility of plants to necrotrophic and biotrophic pathogens. Plant Physiol Biochem. 2012;59:37-43.

Baroncelli R, Amby DB, Zapparata A, Sarrocco S, Vannacci G, Le Floch G, Harrison RJ, Holub E, Sukno SA, Sreenivasaprasad S, Thon MR. Gene family expansions and contractions are associated with host range in plant pathogens of the genus *Colletotrichum*. BMC Genomics. 2016;17:555.

Beakes GW, Glockling SL, Sekimoto S. The evolutionary phylogeny of the oomycete "fungi". Protoplasma. 2012;249:3-19.

Benhamou N, Belanger RR. Induction of systemic resistance to Pythium damping-off in cucumber plants by benzothiadiazole: ultrastructure and cytochemistry of the host response. Plant J. 1998;14:13-21.

Blackman LM, Cullerne DP, Torrena P, Taylor J, Hardham AR. RNA-Seq analysis of the expression of genes encoding cell wall degrading enzymes during infection of Lupin (*Lupinus angustifolius*) by *Phytophthora parasitica*. PLoS One. 2015;10:e0136899.

Blackman LM, Hardham AR. Regulation of catalase activity and gene expression during *Phytophthora nicotianae* development and infection of tobacco. Molec Plant Pathol. 2008;9:495-510.

Brouwer H, Coutinho PM, Henrissat B, de Vries RP. Carbohydrate-related enzymes of important *Phytophthora* plant pathogens. Fungal Genet Biol. 2014;72:192-200.

Calmes B, Morel-Rouhier M, Bataille-Simoneau N, Gelhaye E, Guillemette T, Simoneau P. Characterization of glutathione transferases involved in the pathogenicity of *Alternaria brassicicola*. BMC Microbiol. 2015;15:123.

Chen LQ. SWEET sugar transporters for phloem transport and pathogen nutrition. New Phytol. 2014;201:1150-1155.

Clemmensen KE, Bahr A, Ovaskainen O, Dahlberg A, Ekblad A, Wallander H, Stenlid J, Finlay RD, Wardle DA, Lindahl BD. Roots and associated fungi drive long-term carbon sequestration in boreal forest. Science. 2013;339:1615-1618.

Damasceno CM, Bishop JG, Ripoll DR, Win J, Kamoun S, Rose JK. Structure of the glucanase inhibitor protein (GIP) family from *Phytophthora* species suggests coevolution with plant endo-beta-1,3-glucanases. Mol Plant Microbe Interact. 2008;21:820-830.

Davies ME. The nutrition of *Phytophthora fragariae*. Trans Br Mycol Soc. 1959;42:193-200.

Dereuse H, Levy S, Zeng G, Danchin A. Genetics of the PTS components in *Escherichia coli* K-12. FEMS Microbiol Lett. 1989;63:61-67.

Doehlemann G, Okmen B, Zhu W, Sharon A. Plant pathogenic fungi. Microbiol Spectr. 2017;5.

Dong SM, Raffaele S, Kamoun S. The two-speed genomes of filamentous pathogens: waltz with plants. Curr Opin Genet Dev. 2015;35:57-65.

Dutilh BE, Huynen MA, Snel B. A global definition of expression context is conserved between orthologs, but does not correlate with sequence conservation. BMC Genomics. 2006;7.

Elleuche S, Pöggeler S. Carbonic anhydrases in fungi. Microbiology. 2010;156:23-29.

Ellinger D, Naumann M, Falter C, Zwikowics C, Jamrow T, Manisseri C, Somerville SC, Voigt CA. Elevated early callose deposition results in complete penetration resistance to powdery mildew in Arabidopsis. Plant Physiol. 2013;161:1433-1444.

Emanuelsson O, Brunak S, von Heijne G, Nielsen H. Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc. 2007;2:953-971.

Feng BZ, Zhu XP, Fu L, Lv RF, Storey D, Tooley P, Zhang XG. Characterization of necrosis-inducing NLP proteins in *Phytophthora capsici*. BMC Plant Biol. 2014;14.

Fry WE, McGrath MT, Seaman A, Zitter TA, McLeod A, Danies G, Small IM, Myers K, Everts K, Gevens AJ, et al. The 2009 late blight pandemic in the Eastern United States - causes and results. Plant Dis. 2013;97:296-306.

Gechev TS, Hille J. Hydrogen peroxide as a signal controlling plant programmed cell death. J Cell Biol. 2005;168:17-20.

Gijzen M, Nurnberger T. Nep1-like proteins from plant pathogens: recruitment and diversification of the NPP1 domain across taxa. Phytochemistry. 2006;67:1800-1807

Green H, Jensen DF. Disease progression by active mycelial growth and biocontrol of *Pythium ultimum var. ultimum* studied using a rhizobox system. Phytopathology. 2000;90:1049-1055.

Grieve AG, Rabouille C. Golgi bypass: skirting around the heart of classical secretion. Cold Spring Harb Perspect Biol. 2011;3:a005298.

Grignon C, Sentenac H. pH and ionic conditions in the apoplast. Ann Rev Plant Physiol Plant Mol Biol. 1991;42:103-128.

Haas BJ, Kamoun S, Zody MC, Jiang RH, Handsaker RE, Cano LM, Grabherr M, Kodira CD, Raffaele S, Torto-Alalibo T, et al. Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature. 2009;461:393-398.

Horan K, Jang C, Bailey-Serres J, Mittler R, Shelton C, Harper JF, Zhu JK, Cushman JC, Gollery M, Girke T. Annotating genes of known and unknown function by large-scale coexpression analysis. Plant Physiol. 2008;147:41-57.

Jordan IK, Marino-Ramirez L, Koonin EV. Evolutionary significance of gene expression divergence. Gene. 2005;345:119-126.

Judelson HS, Ah-Fong AM, Aux G, Avrova AO, Bruce C, Cakir C, da Cunha L, Grenville-Briggs L, Latijnhouwers M, Ligterink W, et al. Gene expression profiling during asexual development of the late blight pathogen *Phytophthora infestans* reveals a highly dynamic transcriptome. Molec Plant Microbe Interact. 2008;21:433-447.

Judelson HS, Shrivastava J, Manson J. Decay of genes encoding the oomycete flagellar proteome in the downy mildew *Hyaloperonospora arabidopsidis*. PLoS One. 2012;7:e47624.

Jupe J, Stam R, Howden AJM, Morris JA, Zhang RX, Hedley PE, Huitema E. *Phytophthora capsici*-tomato interaction features dramatic shifts in gene expression associated with a hemi-biotrophic lifestyle. Genome Biol. 2013;14:R63.

Kabbage M, Yarden O, Dickman MB. Pathogenic attributes of *Sclerotinia sclerotiorum*: switching from a biotrophic to necrotrophic lifestyle. Plant Sci. 2015;233:53-60.

Kamoun S, Furzer O, Jones JD, Judelson HS, Ali GS, Dalio RJ, Roy SG, Schena L, Zambounis A, Panabieres F, et al. The top 10 oomycete pathogens in molecular plant pathology. Molec Plant Pathol. 2015;16:413-434

Kellis M, Wold B, Snyder MP, Bernstein BE, Kundaje A, Marinov GK, Ward LD, Birney E, Crawford GE, Dekker J, et al. Defining functional DNA elements in the human genome. Proc Natl Acad Sci USA. 2014;111:6131-6138.

Khaitovich P, Hellmann I, Enard W, Nowick K, Leinweber M, Franz H, Weiss G, Lachmann M, Paabo S. Parallel patterns of evolution in the genomes and transcriptomes of humans and chimpanzees. Science. 2005;309:1850-1854.

Kim HJ, Chen CB, Kabbage M, Dickman MB. Identification and characterization of *Sclerotinia sclerotiorum* NADPH oxidases. Appl Environ Microbiol. 2011; 77:7721-7729.

Kolattukudy PE, Dean BB. Structure, gas-chromatographic measurement, and function of suberin synthesized by potato-tuber tissue-slices. Plant Physiol. 1974; 54:116-121.

Kontkanen H, Westerholm-Parvinen A, Saloheimo M, Bailey M, Ratto M, Mattila I, Mohsina M, Kalkkinen N, Nakari-Setala T, Buchert J. Novel *Coprinopsis cinerea* polyesterase that hydrolyzes cutin and suberin. Appl Environ Microbiol. 2009; 75:2148-2157

Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J Molec Biol. 2001;305:567-580.

Kubicek CP, Starr TL, Glass NL. Plant cell wall-degrading enzymes and their secretion in plant-pathogenic fungi. Annu Rev Phytopathol. 2014;52:427-451.

Lehmann S, Serrano M, L'Haridon F, Tjamos SE, Metraux JP. Reactive oxygen species and plant resistance to fungal pathogens. Phytochemistry. 2015;112:54-62.

Levesque CA, Brouwer H, Cano L, Hamilton JP, Holt C, Huitema E, Raffaele S, Robideau GP, Thines M, Win J, et al. Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire. Genome Biol. 2010;11:R73.

Lewis DH. Concepts in fungal nutrition and origin of biotrophy. Biol Rev Camb Philos Soc. 1973;48:261-278.

Li L, Stoeckert CJ, Jr., Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003;13:2178-2189.

Lingner U, Munch S, Deising HB, Sauer N. Hexose transporters of a hemibiotrophic plant pathogen functional variations and regulatory differences at different stages of infection. J Biol Chem. 2011;286:20913-20922.

Luhtala N, Parker R. T2 family ribonucleases: ancient enzymes with diverse roles. Trends Biochem Sci. 2010;35:253-259.

Lyu XL, Shen CC, Fu YP, Xie JT, Jiang DH, Li GQ, Cheng JS: A small secreted virulence-related protein is essential for the necrotrophic interactions of *Sclerotinia sclerotiorum* with its host plants. PLoS Pathog. 2016;12:e1005435.

Lyu XL, Shen CC, Fu YP, Xie JT, Jiang DH, Li GQ, Cheng JS. Comparative genomic and transcriptional analyses of the carbohydrate-active enzymes and secretomes of phytopathogenic fungi reveal their significant roles during infection and development. Sci Rep. 2015;5.

Mayer AM. Polyphenol oxidases in plants and fungi: going places? A review. Phytochemistry. 2006;67:2318-2331

Meijer HJ, Hassen HH, Govers F. *Phytophthora infestans* has a plethora of phospholipase D enzymes including a subclass that has extracellular activity. PLoS One. 2011;6:e17767.

Meinhardt LW, Costa GGL, Thomazella DPT, Teixeira PJPL, Carazzolle MF, Schuster SC, Carlson JE, Guiltinan MJ, Mieczkowski P, Farmer A, et al. Genome and secretome analysis of the hemibiotrophic fungal pathogen, *Moniliophthora roreri*, which causes frosty pod rot disease of cacao: mechanisms of the biotrophic and necrotrophic phases. BMC Genomics. 2014;15:164.

Melida H, Sandoval-Sierra JV, Dieguez-Uribeondo J, Bulone V. Analyses of extracellular carbohydrates in oomycetes unveil the existence of three different cell wall types. Eukaryot Cell. 2013;12:194-203.

Mertens E. ATP versus pyrophosphate: glycolysis revisited in parasitic protists. Parasitol Today. 1993;9:122-126.

Moreno-Hagelsieb G, Latimer K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. Bioinformatics. 2008;24:319-324.

Morita-Yamamuro C, Tsutsui T, Sato M, Yoshioka H, Tamaoki M, Ogawa D, Matsuura H, Yoshihara T, Ikeda A, Uyeda I, Yamaguchi J. The Arabidopsis gene CAD1 controls programmed cell death in the plant immune system and encodes a protein containing a MACPF domain. Plant Cell Physiol. 2005;46:902-912.

Ng IS, Tsai SW, Ju YM, Yu SM, Ho TH. Dynamic synergistic effect on *Trichoderma reesei* cellulases by novel beta-glucosidases from Taiwanese fungi. Bioresour Technol. 2011;102:6073-6081.

O'connell RJ, Thon MR, Hacquard S, Amyotte SG, Kleemann J, Torres MF, Damm U, Buiate EA, Epstein L, Alkan N, et al. Lifestyle transitions in plant pathogenic *Colletotrichum* fungi deciphered by genome and transcriptome analyses. Nat Genet. 2012;44:1060-1065.

Ospina-Giraldo MD, Griffith JG, Laird EW, Mingora C. The CAZyome of *Phytophthora* spp.: a comprehensive analysis of the gene complement coding for carbohydrate-active enzymes in species of the genus *Phytophthora*. BMC Genomics. 2010;11.

Ospina-Giraldo MD, McWalters J, Seyer L. Structural and functional profile of the carbohydrate esterase gene complement in *Phytophthora infestans*. Curr Genet. 2010;56:495-506.

Osinska I, Popko K, Demkow U. Perforin: an important player in immune response. Cent Eur J Immunol. 2014;39:109-115.

Ostlund G, Schmitt T, Forslund K, Kostler T, Messina DN, Roopra S, Frings O, Sonnhammer ELL. InParanoid 7: new algorithms and tools for eukaryotic orthology analysis. Nucl. Acids Res. 2010;38:D196-D203.

Perlin MH, Andrews J, Toh SS. Essential letters in the fungal alphabet: ABC and MFS transporters and their roles in survival and pathogenicity. Adv Genet. 2014;85:201-253.

Pierleoni A, Martelli PL, Casadio R: PredGPI. a GPI-anchor predictor. BMC Bioinformatics. 2008;9.

Raffaele S, Win J, Cano LM, Kamoun S. Analyses of genome architecture and gene expression reveal novel candidate virulence factors in the secretome of *Phytophthora infestans*. BMC Genomics. 2010;11:637.

Richards TA, Soanes DM, Jones MD, Vasieva O, Leonard G, Paszkiewicz K, Foster PG, Hall N, Talbot NJ. Horizontal gene transfer facilitated the evolution of plant parasitic mechanisms in the oomycetes. Proc Natl Acad Sci USA. 2011;108:15258-15263.

Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26:139-140.

Rolke Y, Liu SJ, Quidde T, Williamson B, Schouten A, Weltring KM, Siewers V, Tenberge KB, Tudzynski B, Tudzynski P. Functional analysis of H_2O_2-generating systems in *Botrytis cinerea*: the major Cu-Zn-superoxide dismutase (BCSOD1) contributes to virulence on French bean, whereas a glucose oxidase (BCGOD1) is dispensable. Molec Plant Pathol. 2004;5:17-27.

Saier MH, Jr. Families of transmembrane transporters selective for amino acids and their derivatives. Microbiology. 2000;146:1775-1795.

Schekman R. SEC mutants and the secretory apparatus. Nat Med. 2002;8:1055-1058.

Solomon PS, Oliver RP. The nitrogen content of the tomato leaf apoplast increases during infection by *Cladosporium fulvum*. Planta. 2001;213:241-249.

Song J, Win J, Tian MY, Schornack S, Kaschani F, Ilyas M, van der Hoorn RAL, Kamoun S. Apoplastic effectors secreted by two unrelated eukaryotic plant pathogens target the tomato defense protease Rcr3. Proc Natl Acad Sci USA. 2009;106:1654-1659.

Skamnioti P, Furlong RF, Gurr SJ. The fate of gene duplicates in the genomes of fungal pathogens. Commun Integr Biol. 2008;1:196-198.

Stam R, Jupe J, Howden AJM, Morris JA, Boevink PC, Hedley PE, Huitema E. Identification and characterisation CRN effectors in *Phytophthora capsici* shows modularity and functional diversity. PLoS One. 2013;8:e59517.

Stassen J, Van den Ackerveken G. How do oomycete effectors interfere with plant life? Curr Opin Plant Biol. 2011;14:407-414.

Tanabe S, Ishii-Minami N, Saitoh KI, Otake Y, Kaku H, Shibuya N, Nishizawa Y, Minami E. The role of catalase-peroxidase secreted by *Magnaporthe oryzae* during early infection of rice cells. Mol Plant Microbe Interact. 2011;24:163-171.

Tirosh I, Barkai N. Evolution of gene sequence and gene expression are not correlated in yeast. Trends Genet. 2008;24:109-113.

Treseder KK, Lennonb JT. Fungal traits that drive ecosystem dynamics on land. Microbiol Mol Biol Rev. 2015;79:243-262.

van den Broek LAM, den Aantrekker ED, Voragen AGJ, Beldman G, Vincken JP. Pectin lyase is a key enzyme in the maceration of potato tuber. J Sci Food Agric. 1997;75:167-172.

van der Burgt A, Jashni MK, Bahkali AH, de Wit PJGM. Pseudogenization in pathogenic fungi with different host plants and lifestyles might reflect their evolutionary past. Molec Plant Pathol. 2014;15:133-144.

van der Hoorn RAL. Plant proteases: from phenotypes to molecular mechanisms. Ann Rev Plant Biol. 2008;59:191-223.

Verma S, Gazara RK, Nizam S, Parween S, Chattopadhyay D, Verma PK. Draft genome sequencing and secretome analysis of fungal phytopathogen *Ascochyta rabiei* provides insight into the necrotrophic effector repertoire. Sci Rep. 2016;6.

Weiberg A, Wang M, Lin FM, Zhao HW, Zhang ZH, Kaloshian I, Huang HD, Jin HL. Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. Science. 2013;342:118-123.

Wirth CC, Glushakova S, Scheuermayer M, Repnik U, Garg S, Schaack D, Kachman MM, Weissbach T, Zimmerberg J, Dandekar T, et al. Perforin-like protein PPLP2 permeabilizes the red blood cell membrane during egress of *Plasmodium falciparum* gametocytes. Cell Microbiol. 2014;16:709-733.

Yanai I, Graur D, Ophir R. Incongruent expression profiles between human and mouse orthologous genes suggest widespread neutral evolution of transcription control. OMICS. 2004;8:15-24.

Zerillo MM, Adhikari BN, Hamilton JP, Buell CR, Levesque CA, Tisserat N. Carbohydrate-active enzymes in *Pythium* and their role in plant cell wall and storage polysaccharide degradation. PLoS One. 2013;8:e72572.

Zhao ZT, Liu HQ, Wang CF, Xu JR. Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics. 2013;14:274.

Zuluaga AP, Vega-Arreguin JC, Fei ZJ, Ponnala L, Lee SJ, Matas AJ, Patev S, Fry WE, Rose JKC. Transcriptional dynamics of *Phytophthora infestans* during sequential stages of hemibiotrophic infection of tomato. Molec Plant Pathol. 2016;17:29-41.

Figure 1. Overview of expression data. a, images of uninfected tuber, tuber infected with *Ph. infestans* at 4 dpi, and tuber infected with *Py. ultimum* at 1.5 dpi. *Ph. infestans*-infected tubers are asymptomatic at 1.5 dpi. **b,** Heatmaps of hierarchical clustered TMM-normalized data from artificial media and plant samples for *Ph. infestans* (top) and *Py. ultimum* (bottom). Only genes with CPM ≥1 in at least one condition are shown. **c,** Expression of stage-specific markers in early, middle and late tuber samples of *Ph. infestans*. **d and e,** Principal component analysis (PCA) displaying intrinsic biological variation between replicates of *Ph. infestans* and *Py. ultimum* samples. **f,** $\log_2$ fold-change ratios in comparisons of early tuber versus early media, and late versus early tuber.

Figure 1

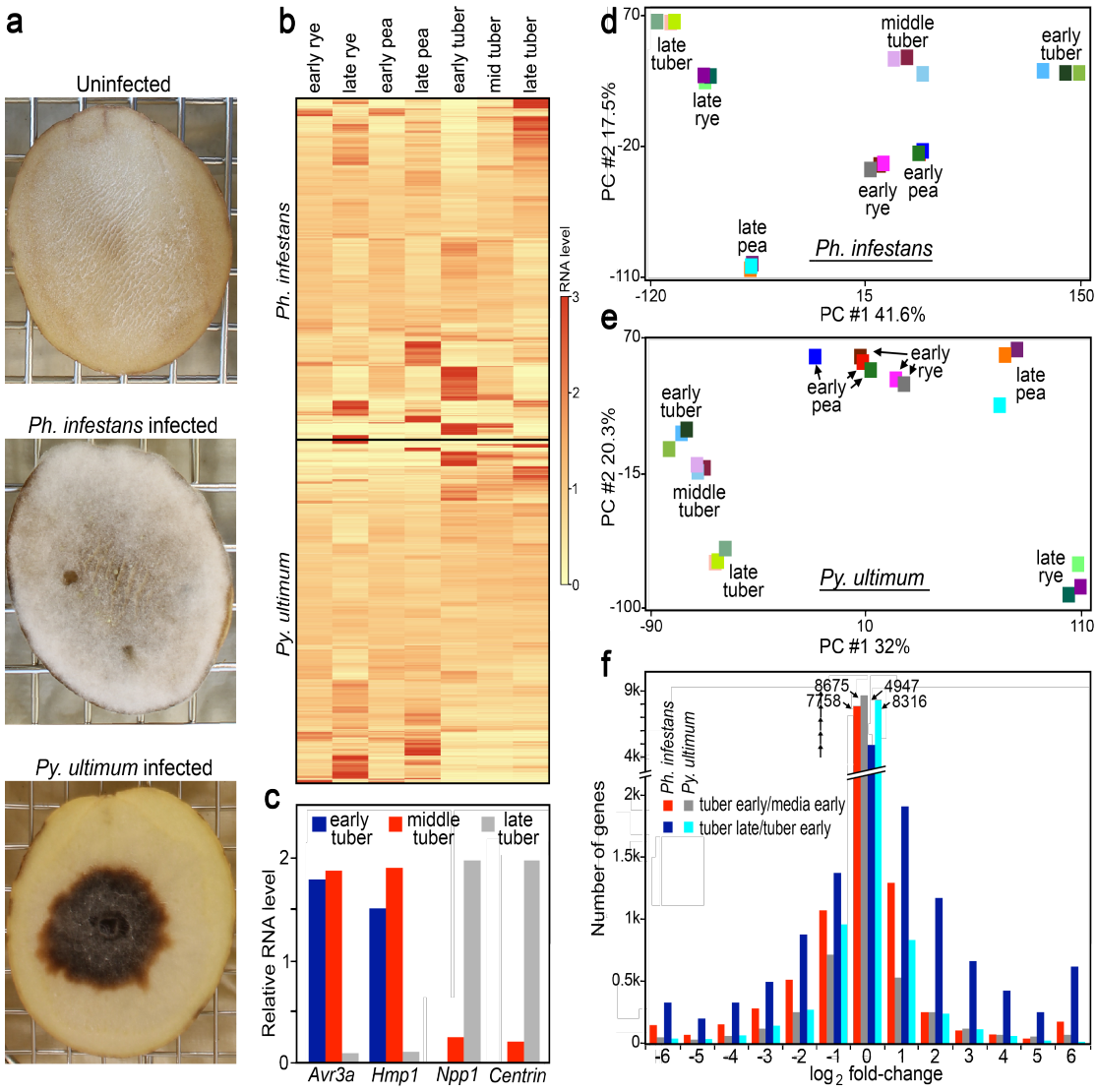

a
Uninfected
Ph. infestans infected
Py. ultimum infected
b
early rye
late rye
early pea
late pea
early tuber
mid tuber
late tuber
Ph. infestans
Py. ultimum
RNA level
c
early tuber
middle tuber
late tuber
Relative RNA level
Avr3a
Hmp1
Npp1
Centrin
d
late tuber
middle tuber
early tuber
late rye
early pea
early rye
late pea
Ph. infestans
PC #2 17.5%
PC #1 41.6%
e
early rye
early pea
late pea
early tuber
middle tuber
late tuber
late rye
Py. ultimum
PC #2 20.3%
PC #1 32%
f
8675
7758
4947
8316
Ph. infestans
Py. ultimum
tuber early/media early
tuber late/tuber early
Number of genes
log2 fold-change

Figure 2. Gene ontology (GO) enrichment analysis. Indicated are terms that are over- or under-represented in genes of *Ph. infestans* (white bars) or *Py. ultimum* (black bars) that are up-regulated in **a,** early tuber compared to the mean of early rye and pea media, and **b,** late tuber versus early tuber.

Figure 2

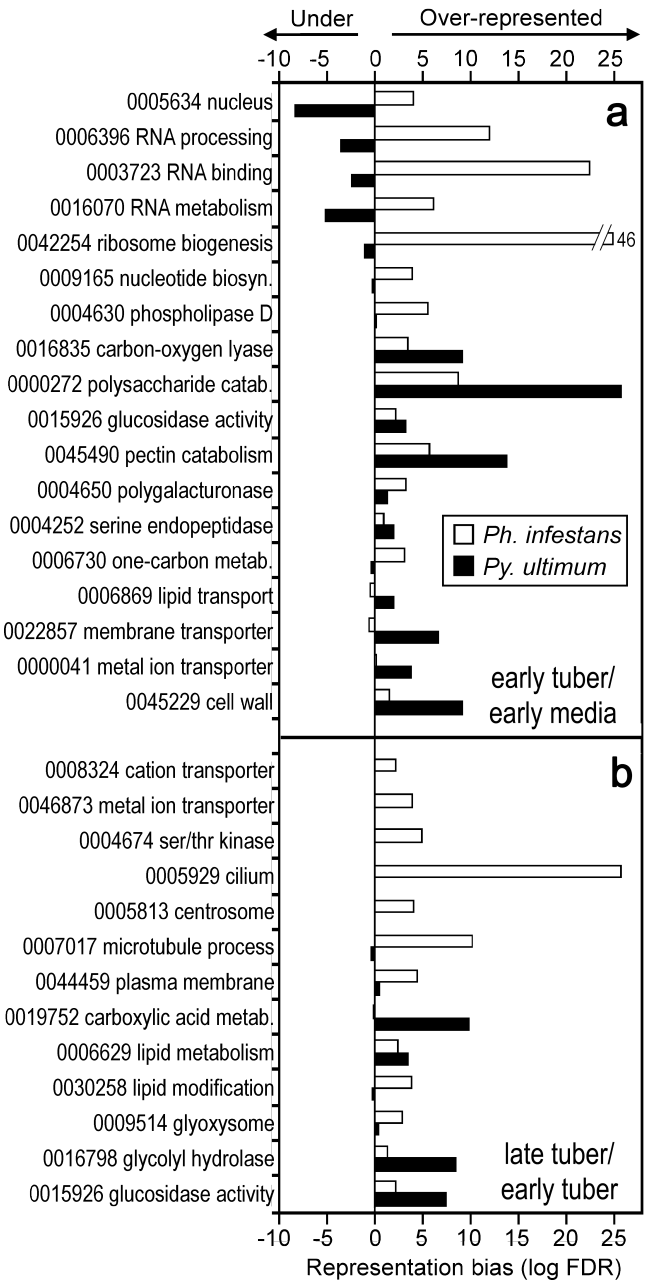

Under
Over-represented
-10 -5 0 5 10 15 20 25
0005634 nucleus
0006396 RNA processing
0003723 RNA binding
0016070 RNA metabolism
0042254 ribosome biogenesis
46
0009165 nucleotide biosyn.
0004630 phospholipase D
0016835 carbon-oxygen lyase
0000272 polysaccharide catab.
0015926 glucosidase activity
0045490 pectin catabolism
0004650 polygalacturonase
0004252 serine endopeptidase
0006730 one-carbon metab.
0006869 lipid transport
0022857 membrane transporter
0000041 metal ion transporter
0045229 cell wall
a
Ph. infestans
Py. ultimum
early tuber/
early media
0008324 cation transporter
0046873 metal ion transporter
0004674 ser/thr kinase
0005929 cilium
0005813 centrosome
0007017 microtubule process
0044459 plasma membrane
0019752 carboxylic acid metab.
0006629 lipid metabolism
0030258 lipid modification
0009514 glyoxysome
0016798 glycolyl hydrolase
0015926 glucosidase activity
b
late tuber/
early tuber
-10 -5 0 5 10 15 20 25
Representation bias (log FDR)

Figure 3. Expression of polyphenol oxidases. The colored bar graphs in the bottom plane show CPM values for individual genes in the six growth conditions in *Ph. infestans* (top) and *Py. ultimum* (bottom). Using the same color scheme, the bar graphs on the back wall portray aggregate CPM for all genes; those with the highest levels (pea late and tuber early in the upper and lower graphs, respectively) are indicated. Heatmaps based on per gene-normalized values are shown in grey scale on the left wall. Genes with CPM <1 are shown as "no reads".

Figure 3

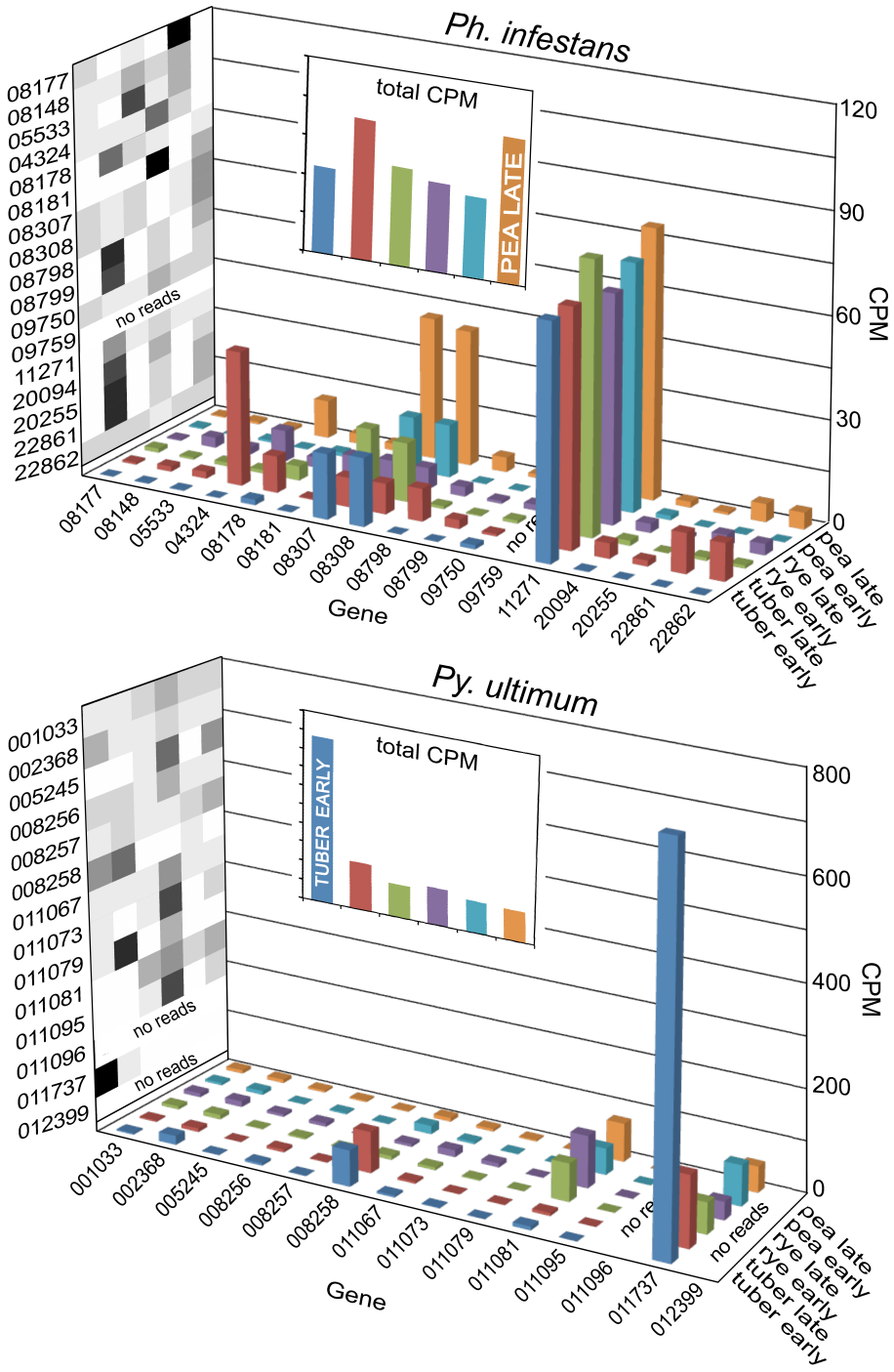

Ph. infestans
total CPM
PEA LATE
08177
08148
05533
04324
08178
08181
08307
08308
08798
08799
09750
09759
11271
20094
20255
22861
22862
no reads
Gene
CPM
120
90
60
30
0
pea late
pea early
rye late
rye early
tuber late
tuber early
Py. ultimum
total CPM
TUBER EARLY
001033
002368
005245
008256
008257
008258
011067
011073
011079
011081
011095
011096
011737
012399
no reads
no reads
Gene
CPM
800
600
400
200
0
pea late
pea early
rye late
rye early
tuber late
tuber early

Figure 4. Expression of selected pathogenesis-related genes. Illustrated are **a,** secreted proteases and **b,** five other classes of secreted pathogenicity factors. Shown on the left side of each panel are heatmaps of genes from *Ph. infestans* (left) and *Py. ultimum* (right) that have CPM ≥1 in at least one growth condition. The pie charts in the middle of each panel represent the fraction of genes from *Ph. infestans* (Pi) and *Py. ultimum* (Pu) that are up-regulated by ≥3-fold (red) or down-regulated (yellow) in early tuber compared to media, or late tuber compared to early tuber. Genes that show smaller changes are represented by the teal (blue-green) slice. Shown on the right side of each panel are bar charts indicating the aggregate CPM of all genes in each functional group in early rye and pea media (averaged; ME), early tuber (TE), and late tuber (TL). Note that Ph. infestans expresses a single aspartyl protease.

Figure 4

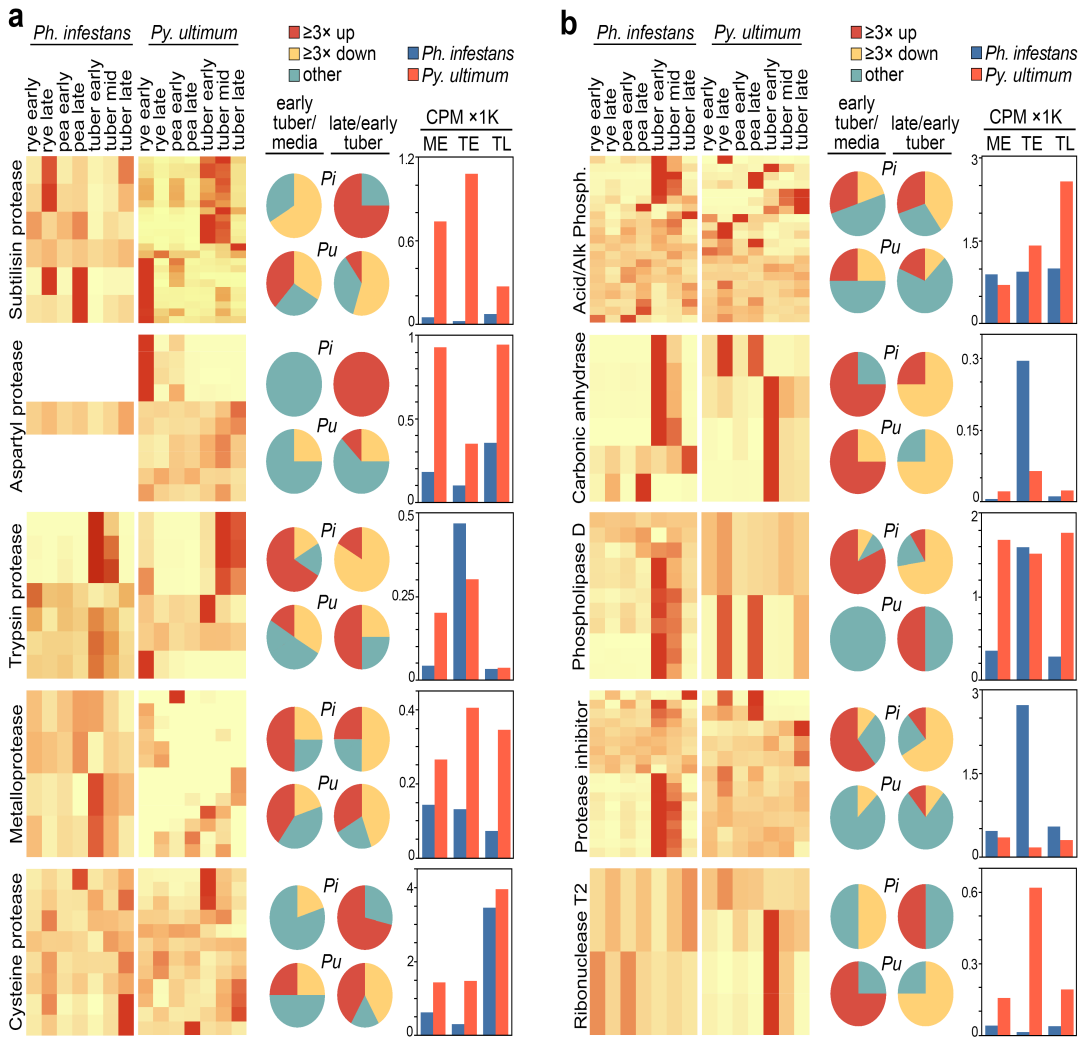

a
b
Ph. infestans
Py. ultimum
rye early
rye late
pea early
pea late
tuber early
tuber mid
tuber late
≥3× up
≥3× down
other
Ph. infestans
Py. ultimum
early tuber/ media
late/early tuber
CPM ×1K
ME TE TL
Subtilisin protease
Aspartyl protease
Trypsin protease
Metalloprotease
Cysteine protease
Acid/Alk Phosph.
Carbonic anhydrase
Phospholipase D
Protease inhibitor
Ribonuclease T2
Pi
Pu

Figure 5. Structure and expression of NPP (NLP) family. The left side of the figure is a cladogram of the genes from *Ph. infestans* (PITG prefix, black lettering) and *Py. ultimum* (PU prefix, green lettering). Bootstrap numbers from PhyML are shown at nodes. The right side of the figure indicates the CPM values in early media (average of early rye and pea, green), early tubers (blue) and late tubers (red bars). An absence of bars indicates CPM ≤1.0. Proteins predicted to be secreted are marked by black circle on the terminal branches of the tree, and those bearing the residues required for plant necrosis are marked by a diamond. These correspond to D112, H120, D123, and E125 of *Phytophthora capsici* NLP1. While PITG_22668 contains those residues, it also contains a S113E substitution which was associated with low necrosis in *Ph. capsici.*

Figure 5

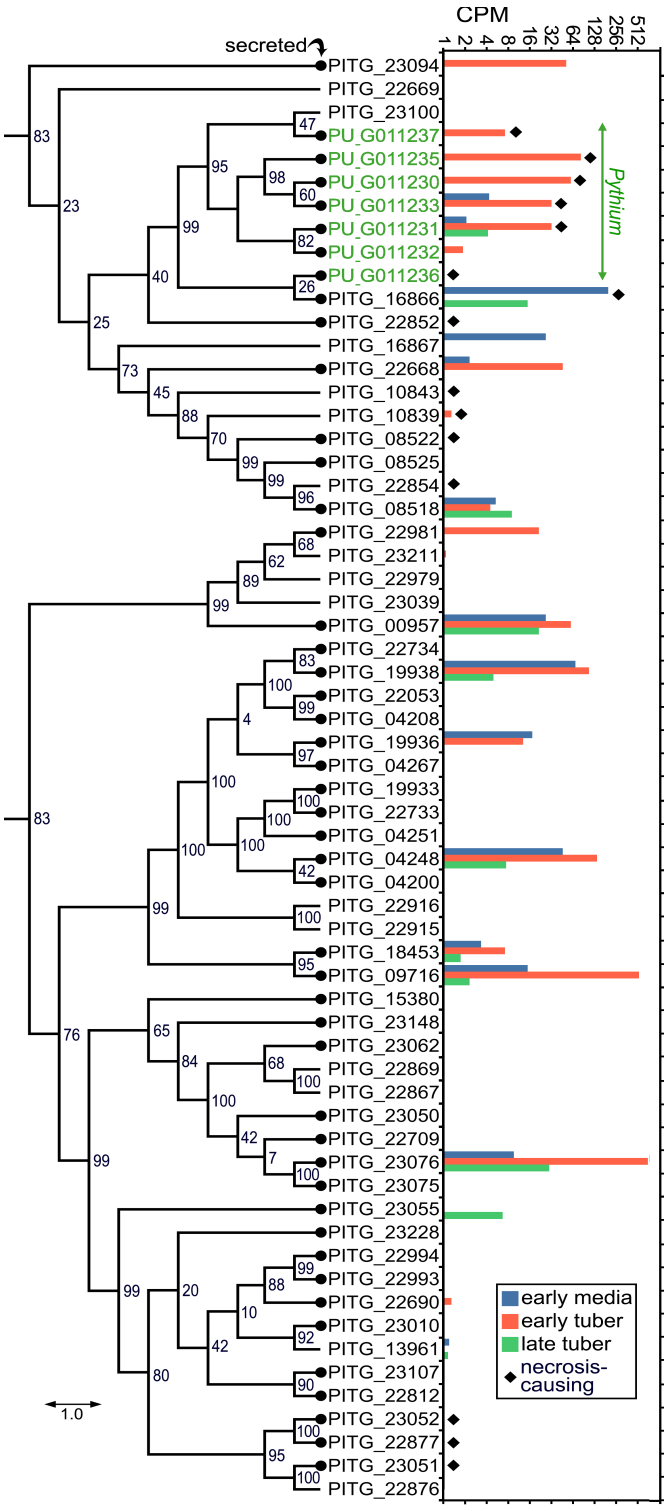

CPM
1
2
4
8
16
32
64
128
256
512
secreted
PITG_23094
PITG_22669
PITG_23100
PU_G011237
PU_G011235
PU_G011230
PU_G011233
PU_G011231
PU_G011232
PU_G011236
PITG_16866
PITG_22852
PITG_16867
PITG_22668
PITG_10843
PITG_10839
PITG_08522
PITG_08525
PITG_22854
PITG_08518
PITG_22981
PITG_23211
PITG_22979
PITG_23039
PITG_00957
PITG_22734
PITG_19938
PITG_22053
PITG_04208
PITG_19936
PITG_04267
PITG_19933
PITG_22733
PITG_04251
PITG_04248
PITG_04200
PITG_22916
PITG_22915
PITG_18453
PITG_09716
PITG_15380
PITG_23148
PITG_23062
PITG_22869
PITG_22867
PITG_23050
PITG_22709
PITG_23076
PITG_23075
PITG_23055
PITG_23228
PITG_22994
PITG_22993
PITG_22690
PITG_23010
PITG_13961
PITG_23107
PITG_22812
PITG_23052
PITG_22877
PITG_23051
PITG_22876
Pythium
early media
early tuber
late tuber
necrosis-causing
1.0

Figure 6. Expression of selected pathogenesis-related genes. Illustrated are **a,** cell wall-degrading enzymes (CWDEs) grouped by activity, and **b,** genes potentially involved in detoxification. The format of the figure is the same as in Fig. 4.

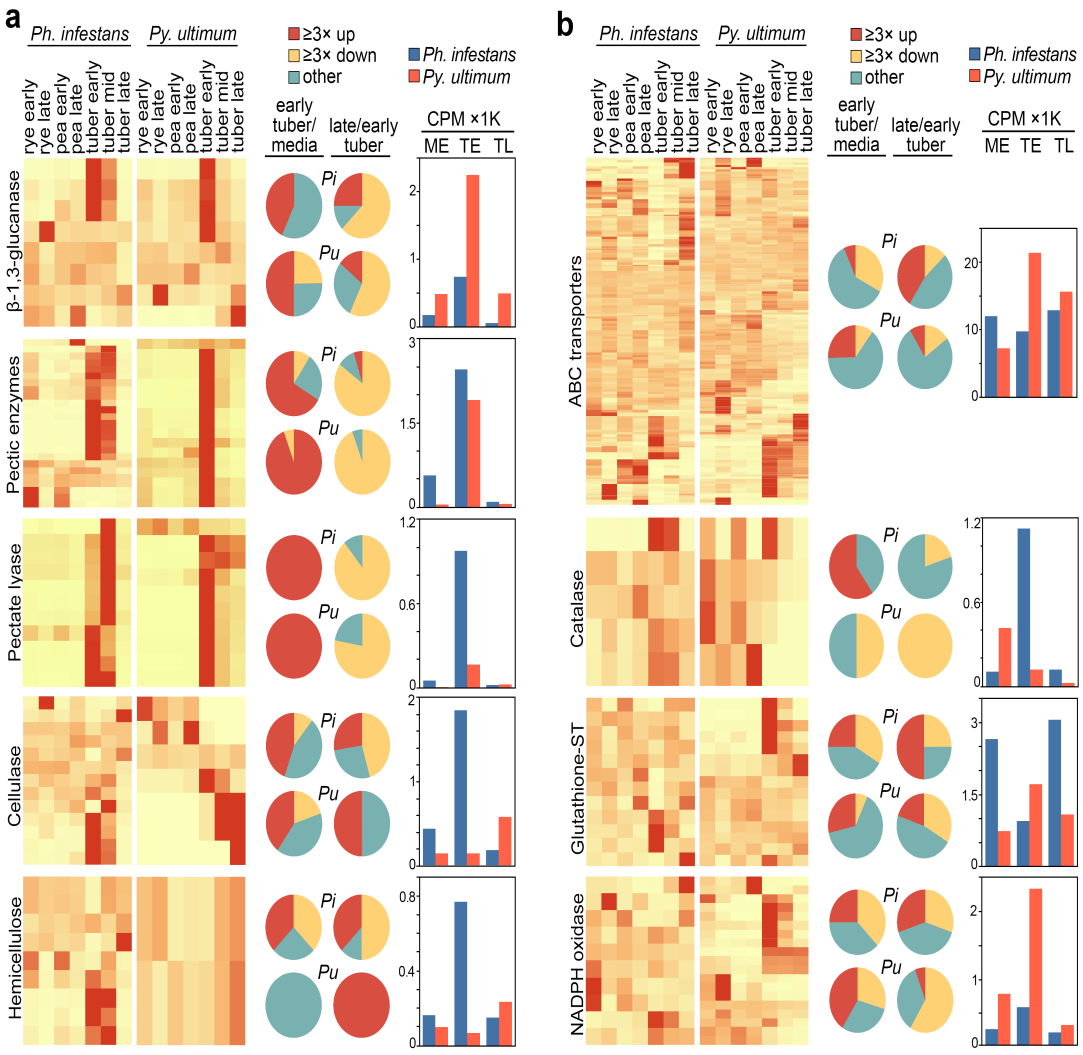


Figure 7. Expression of genes involved in nutrient transport. The format of the figure is the same as in Fig. 4, and functional categories are defined in the main text.

Figure 7

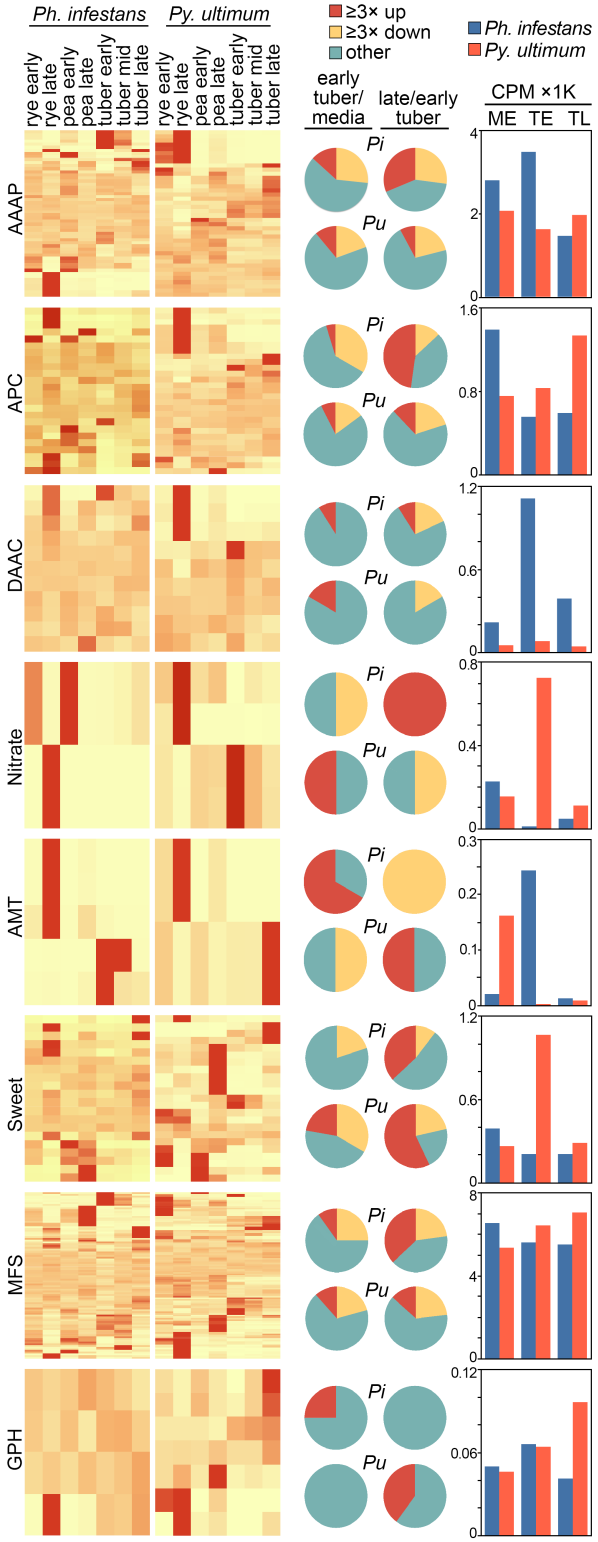

Ph. infestans
Py. ultimum
rye early
rye late
pea early
pea late
tuber early
tuber mid
tuber late
≥3× up
≥3× down
other
early tuber/ media
late/early tuber
Pi
Pu
Ph. infestans
Py. ultimum
CPM ×1K
ME TE TL
AAAP
APC
DAAC
Nitrate
AMT
Sweet
MFS
GPH

Fig. 8. MA plots of ortholog pairs. The y-axis indicates the $\log_2$ ratio of the FPKM values of the two samples being compared (M) while the x-axis shows the $\log_{10}$ product of their FPKM values (A). The top six panels compare *Ph. infestans* (Pi) and *Py. ultimum* (Pu) in early and late rye, pea, and tuber. The bottom two panels compare two replicates of *Ph. infestans* and *Py. ultimum* in early rye. The pie charts indicate the fraction of genes that are up- (+) or down-regulated (-) in *Ph. infestans* compared to *Py. ultimum* based on a fold-change threshold of 3.0. Only genes with FPKM≥1 in at least one species are shown.

Figure 8

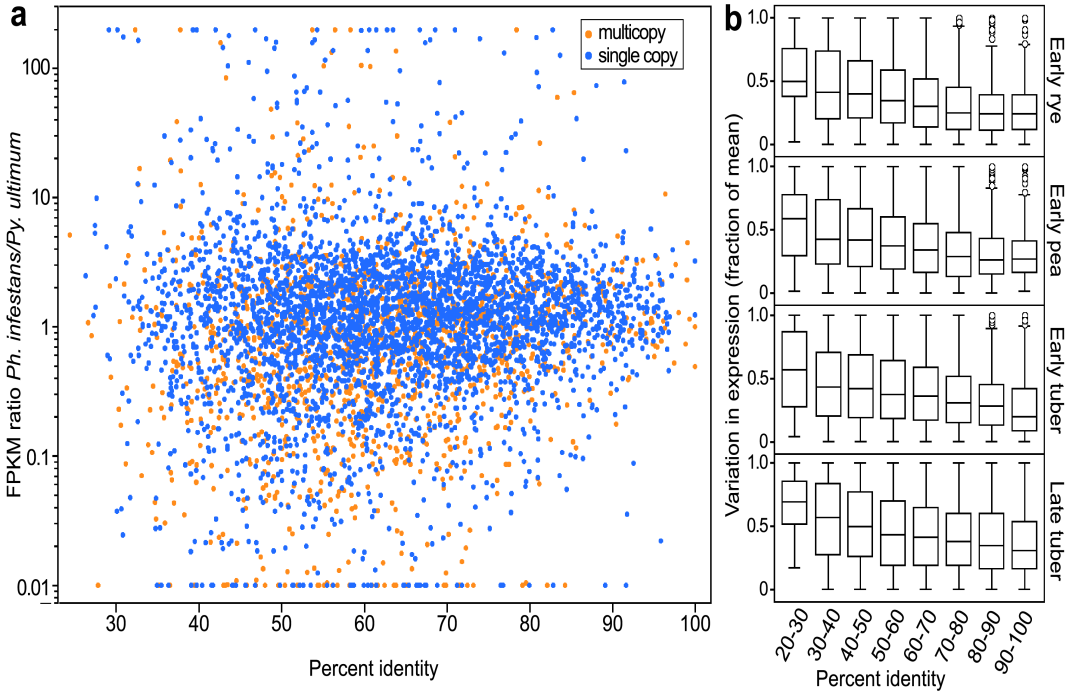

a
multicopy
single copy
FPKM ratio Ph. infestans/Py. ultimum
100
10
1
0.1
0.01
30
40
50
60
70
80
90
100
Percent identity
b
Variation in expression (fraction of mean)
1.0
0.5
0
Early rye
Early pea
Early tuber
Late tuber
20-30
30-40
40-50
50-60
60-70
70-80
80-90
90-100
Percent identity

Figure 9. Relationship of amino acid similarity and expression conservation. a, Plot of FPKM ratio of orthologs (*Ph. infestans* divided by *Py. ultimum*) in early tuber versus amino acid identity. Orange and blue symbols represent members of multigene families and single-copy genes, respectively. **b,** Box-plots indicating variation in expression level in early rye (based on 5273 genes with FPKM≥1), early pea (5211 genes), early tuber (5123 genes), and late tuber (5461 genes) as a function of amino acid identity. Similar trends were seen in late rye and pea. Variation is defined as the difference between the FPKM of orthologs divided by their summed FPKM values. Ratios above 200 or below 0.01 are graphed as 200 and 0.01, respectively.

Figure 9

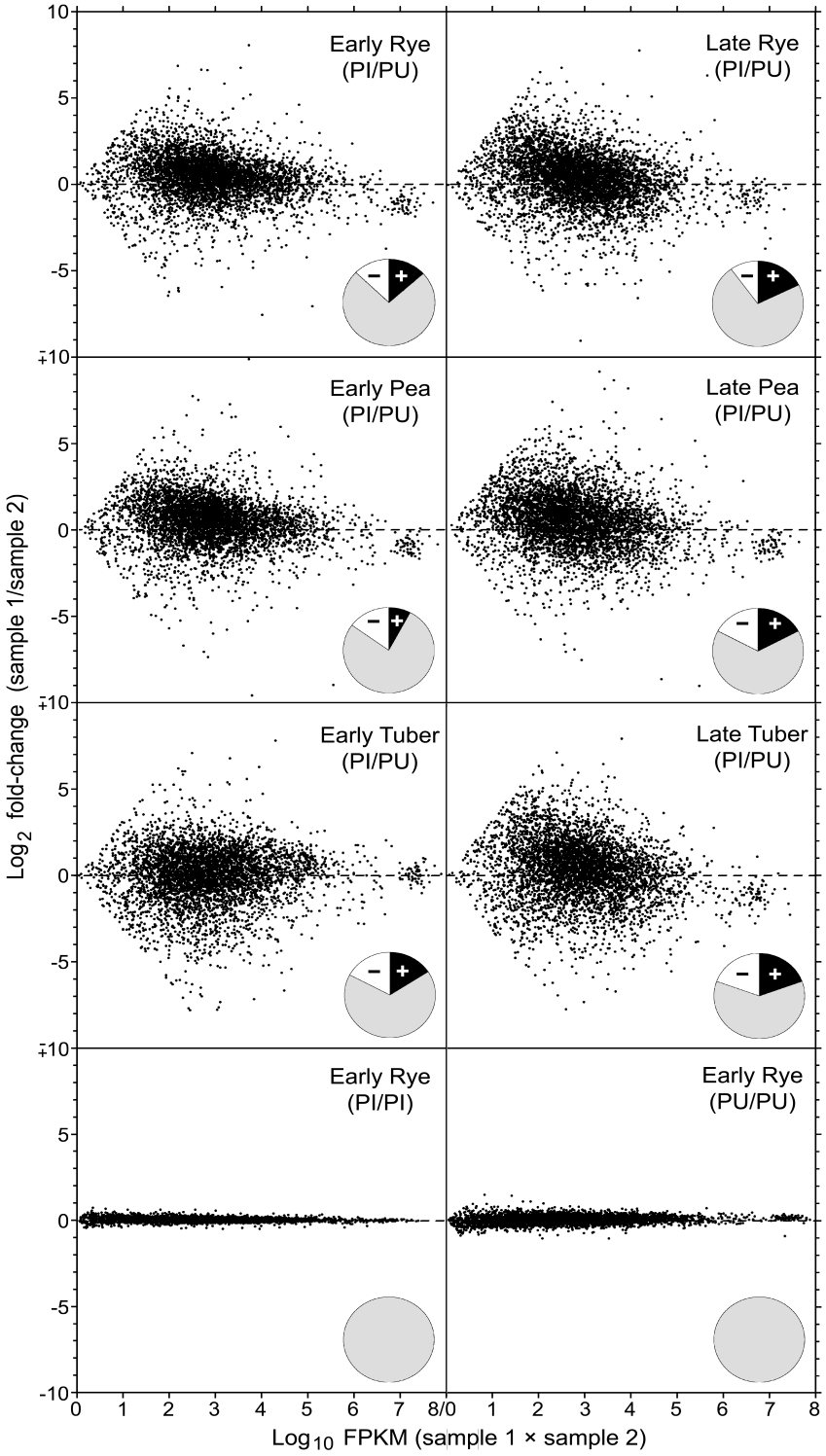

Early Rye
(PI/PU)
Late Rye
(PI/PU)
Early Pea
(PI/PU)
Late Pea
(PI/PU)
Early Tuber
(PI/PU)
Late Tuber
(PI/PU)
Early Rye
(PI/PI)
Early Rye
(PU/PU)
Log2 fold-change (sample 1/sample 2)
Log10 FPKM (sample 1 × sample 2)

Figure 10. Correlation between FPKM of orthologs in selected functional classes. Ortholog pairs were classified into functional groups and filtered to include only those with FPKM≥1 in both species. Pearson's correlation coefficients are graphed as positive (blue) or negative (red) in early rye and pea, and early and late tuber. The area of each circle is proportional to the degree of correlation.

Fig. 10

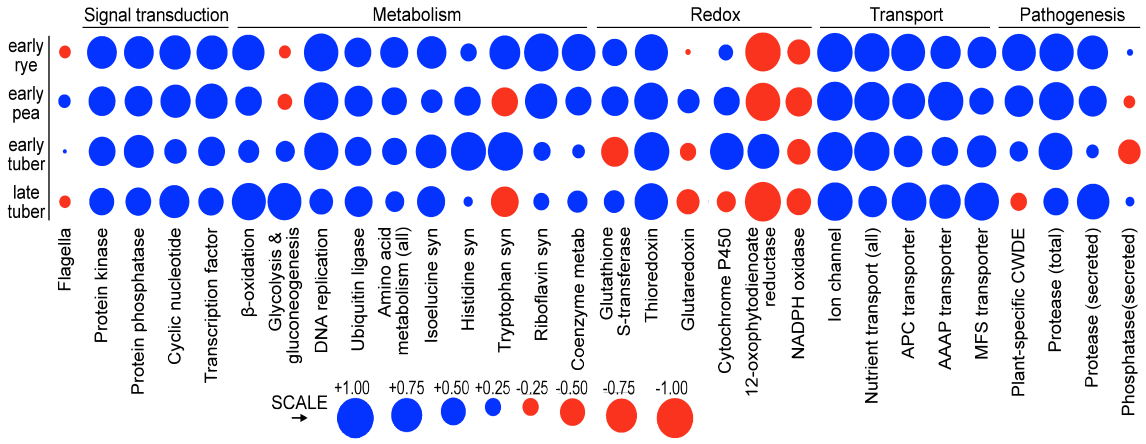


Figure 11. Examples of conserved and discordant expression patterns within orthologous families. Illustrated are three families of orthologs encoding **b,** a regulatory subunit of calcineurin; **b,** P-ATPases; and **c**, a metal ion transporter. Indicated on the left side of each panel are cladogram trees with bootstrap values from PhyML (top) and neighbor-joining. The *Ph. infestans* and *Py. ultimum* gene numbers have PI and PU prefixes, respectively. The bar charts above the trees indicate FPKM of each gene in early tuber, late tuber, and early rye.

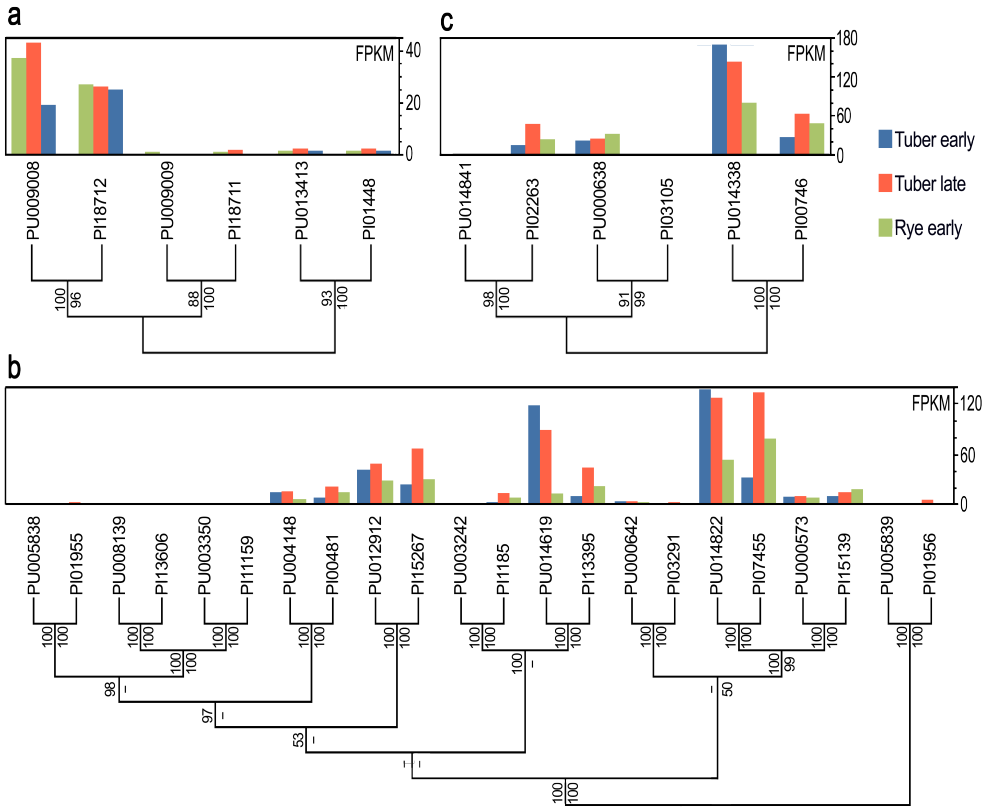
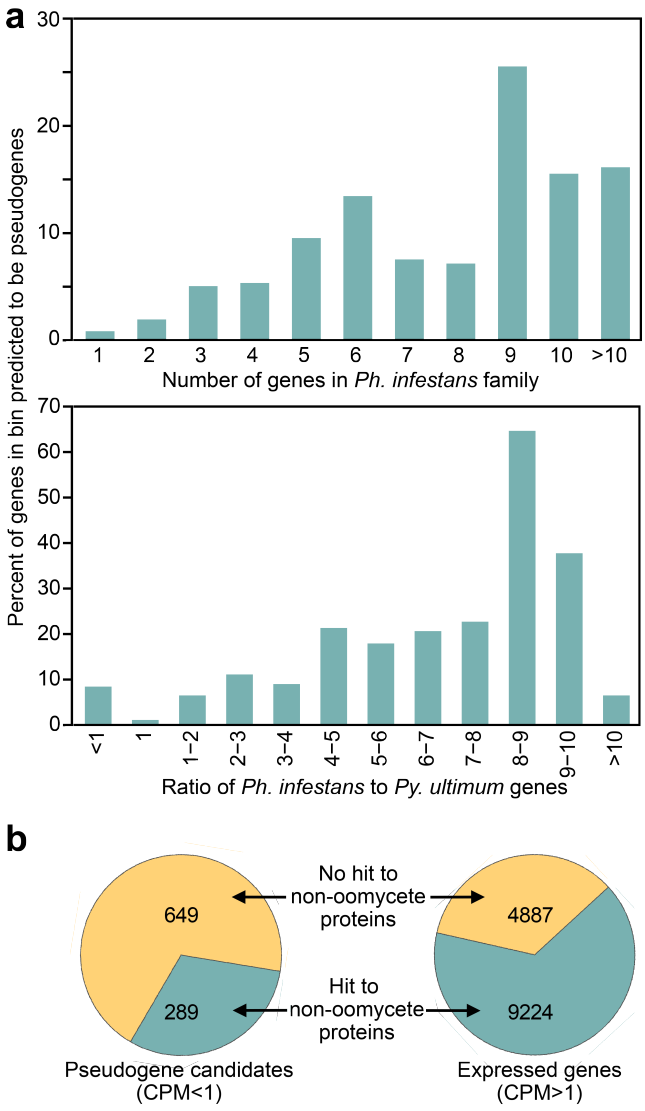


Figure 12. Pseudogenization in *Ph. infestans*. a, relationship between the occurrence of a candidate pseudogene and the size of the gene family (top) or ratio of paralogs in *Ph. infestans* and *Pu. ultimum* (bottom). **b,** comparison of the fraction of pseudogenes and expressed genes having hits in GenBank against non-oomycete proteins.



Chapter V

Decay of Genes Encoding the Oomycete Flagellar Proteome in the Downy Mildew *Hyaloperonospora arabidopsidis*

Abstract

Zoospores are central to the life cycles of most of the eukaryotic microbes known as oomycetes, but some genera have lost the ability to form these flagellated cells. In the plant pathogen *Phytophthora infestans,* genes encoding 257 proteins associated with flagella were identified by comparative genomics. These included the main structural components of the axoneme and basal body, proteins involved in intraflagellar transport, regulatory proteins, enzymes for maintaining ATP levels, and others. Transcripts for over three-quarters of the genes were up-regulated during sporulation, and persisted to varying degrees in the pre-zoospore stage (sporangia) and motile zoospores. Nearly all of these genes had orthologs in other eukaryotes that form flagella or cilia, but not species that lack the organelle. Orthologs of 211 of the genes were also absent from a sister taxon to *P. infestans* that lost the ability to form flagella, the downy mildew *Hyaloperonospora arabidopsidis*. Many of the genes retained in *H. arabidopsidis* were also present in other non-flagellates, suggesting that they play roles both in flagella and other cellular processes. Remnants of the missing genes were often detected in the *H. arabidopsidis* genome. Degradation of the genes was associated with local compaction of the chromosome and a heightened propensity towards genome rearrangements, as such regions were less likely to share synteny with *P. infestans.*

Background:

Flagella and their shorter relatives, cilia, are remarkably conserved in structure in all major eukaryotic groups (Carvalho-Santos et al., 2011). This suggests that these motility and sensory organelles existed in the last eukaryotic common ancestor, LECA. The appearance of microtubular flagella, along with other cytoskeletal components, are hallmarks of the transition from prokaryote to eukaryote (Erickson, 2009). Flagella have been influenced during evolution by the acquisition, loss, or modification of genes. For example, the major component of eukaryotic flagella, tubulin, is thought to have evolved from the prokaryotic cell division protein ftsZ (Erickson, 2009). In addition, evolution of a cytoplasmic pathway for assembling flagella in protozoa such as *Plasmodium* has been linked to the loss of intraflagellar transport proteins, which in other species deliver building blocks from basal bodies to flagellar tips (Jekely and Arendt, 2006).

The most dramatic example of evolutionary change involving flagella is the loss of the organelle. This occurred during the evolution of plants, slime molds, most fungi, and some oomycetes (Carvalho-Santos et al., 2011; Liu et al., 2006; Dick, 2001). It is reasonable to propose that in such cases an initial mutation arose that impaired the formation of flagella, followed by relaxed selection against defects in other genes. Similar events are well-described in bacterial symbionts, where pathway losses were linked to chromosomal rearrangements, deletions, pseudogene and gene remnant accumulation, and reduction in genome size (Keeling and Slamovits, 2005; Moran, 2003). Related events in eukaryotes have been associated with truncation, mutation, or deletion along with transposable element activity (Lai et al., 2004; Charles et al., 2009).

The existence of species that lost the ability to form flagella can help define proteins required for the organelle. By searching for genes shared only by flagellates, prior studies identified between about 75 and 250 candidates depending on the species examined (Merchant et al., 2007; Baron et al., 2007; Fritz-Laylin et al., 2010; Li et al., 2004). In addition, proteomic studies of the organelle identified up to about 300 proteins (Pazour et al., 2005; Yang et al., 2006; Portman et al., 2009; Ostrowski et al., 2002). Not all proteins found by proteomics are necessarily specific to flagella, and some could be contaminants. Many were nevertheless shown to have a flagellar role by RNAi, to be expressed during flagella formation, or to localize to the organelle (Portman et al., 2009; Stolc et al., 2005; Lechtreck et al., 2009). Central features of flagella include microtubules that usually form a 9+2 configuration of outer and inner doublets, radial spokes that join the outer and inner doublets, dynein arms between the doublets, and the basal body. Flagellar structure can vary between species, for example mammalian sperm contain unique fibers peripheral to the 9+2 axoneme, and trypanosomes contain a novel paraflagellar rod (Carvalho-Santos et al., 2011). Combined with the fact that some flagellar proteins may be retained in nonmotile or aciliated species if they serve other roles, there is value to integrating comparative genomics and proteomics across multiple species.

To understand further the diversity of eukaryotic flagella and mechanisms of their elimination from some lineages, here we focus on the oomycetes *Phytophthora infestans*, the potato blight agent, and *Hyaloperonospora arabidopsidis,* a downy mildew pathogen of *Arabidopsis thaliana* (Slusarenko and Schlaich, 2003). Like most oomycetes, *P.*

infestans produces a biflagellated life-stage called the zoospore that helps propagules reach optimal infection or colonization sites (Jupe et al., 2013). Despite its fairly close taxonomic affinity to *P. infestans, H. arabidopsidis* fails to form zoospores, and in a preliminary study we reported that some genes for flagella proteins were absent from the latter's genome (Judelson et al., 2012). To explore this further, in this paper we report the identification of 257 candidate flagella genes from *P. infestans* based on comparative genomics and RNA expression studies, and show that 81% lack orthologs in *H. arabidopsidis.* The loss of flagella from the downy mildew does not seem to be a very ancient event since gene remnants were found for about one-fifth of genes.

RESULTS

Candidate *P. infestans* flagellar proteins from comparative genomics

Two recent studies reported finding about 100 orthologs of known flagella genes in oomycetes (Fritz-Laylin et al., 2010; Levesque et al., 2010). To search more thoroughly for candidates from *P. infestans,* sequences were collected from prior studies of diverse species. These proteins included 95 found by proteomic analysis of flagella from *Chlamydomonas reinhardtii* (Pazour et al., 2005), 195 identified by comparative genomics as being in flagellates such as *C. reinhardtii* but not *H. arabidopsidis* (Merchant et al., 2007), 106 from genomics and proteomics analysis of *Trypanosoma brucei* (Baron et al., 2007; Portman et al., 2009), 182 from a study of *Naegleria gruberi* (Fritz-Laylin et al., 2010), 148 predicted to be in metazoan cilia (Mayor et al., 2006), 51 downregulated in a cilia-lacking mutant of *Caenorhabditis elegans* (van der Hoorn, 2006)], 102 from a proteomics study of *Tetrahymena thermophila* basal bodies (Baroncelli et al., 2016), and other proteins reported to reside in flagella or cilia (van der Hoorn, 2006; Ah-Fong and Judelson, 2011; Yamagishi et al., 2009; Wang et al., 2009; Lauwaet et al., 2011). These candidate flagella components represented 460 nonredundant proteins.

P. infestans was searched for orthologs to the 460 sequences employing the reciprocal best hit method using Blastp. Ortholog assignment was supplemented by phylogenetic analysis when multiple closely-matching proteins were detected. To reduce false negatives due to erroneous gene models, candidates lacking orthologs in *P. infestans* were searched against its assembly as well as the related species *Phytophthora sojae.* A

total of 257 orthologs were thus identified, including one from an uncalled gene in the original genome study (Haas et al., 2009). The common names, *P. infestans* gene model numbers in the Broad Institute database, and predicted functions of each gene are listed in Table S1.

Distribution of orthologs in other eukaryotes

The association of the genes having orthologs in *P. infestans* with flagella or cilia was explored by examining eleven other eukaryotes (Fig. 1). The species examined include six that form motile flagella or cilia (*C. reinhardtii, H. sapiens, N. gruberi, T. thermophila, T. brucei, Giardia lamblia*), *C. elegans* which forms only non-motile sensory cilia, and *Dictyostelium discoideum*, *Schizosaccharomyces pombe, Ostreococcus tauri,* and *H. arabidopsidis* which produce neither cilia or flagella. Phylogenetic analyses indicate that the last species is related closely to *P. infestans* (Runge et al., 2011).

Most of the predicted *P. infestans* flagella-associated proteins, about 200, were conserved in *C. reinhardtii,* humans, *N. gruberi, T. thermophila, and T. brucei.* This number was reduced in *C. elegans* and *G. lamblia* to 117 and 124, respectively, which reflects the lack of motile cilia in the nematode and the compaction of the *G. lamblia* genome (Morrison et al., 2007). About 78% of the proteins lack orthologs in the non-flagellates.

As shown in Fig. 1, *P. infestans* encodes members of all major flagellar protein functional categories as defined by studies of models such as *C. reinhardtii.* The proteins in *P. infestans* include 27 that reside within or attach to the basal body, seven central pair

proteins, all five expected tubulins (α, β, γ, δ, ε), 24 proteins representing heavy, intermediate, and light chain dyneins, and other proteins with roles in cellular regulation, metabolism, protein-protein interactions, and other functions. Also present are 28 proteins involved in intraflagellar transport. These include each IFT protein described for *C. reinhardtii,* along with the kinesins that work with IFT factors to move proteins along the axoneme. *P. infestans* also encodes orthologs of 13 of the identified radial spoke proteins of *C. reinhardtii.* Only one RSP4-like protein is encoded by *P. infestans,* like most eukaryotes except *C. reinhardtii* where the gene is duplicated (Wei et al., 2010). Of the proteins expressed from single genes in the green alga, six come from duplicated genes in *P. infestans*, namely DAU1, DHC10, DHC14, DYF13, FAB57, and FBB15, and RIB72. Four *P. infestans* genes correspond to centrin, which is single-copy in *C. reinhardtii* (DLE2) but also multicopy in humans (LeDizet et al., 1998).

Of orthologs found in non-flagellates, most common are metabolic proteins. Examples include the FM165 acyl transferase, which was detected in *S. pombe*, *D. discoideum, O. tauri,* and *H. arabidopsidis,* and the FM177 oxidoreductase, which was found in the first two species. Such proteins presumably play multiple roles, some of which persist in non-flagellates. A second class of retained proteins include those that mediate protein-protein interactions or protein maturation, such as RSP16, which encodes a Hsp40 chaperone, and RSP12, which encodes a peptidyl-prolyl isomerase.

Forty-one sequences from the *C. reinhardtii* flagellar proteome were not detected in *P. infestans;* most flagellates also lack these proteins (Fig. S1). For example, RSP5 was not detected in humans, *N. gruberi, T. thermophila, T. brucei, G. lamblia*, or *C.*

elegans. Also absent from *P. infestans* were many flagella-associated proteins that had been identified in *N. gruberi* and *T. brucei,* but shown before to be absent from most other flagellates. For example, *P. infestans* lacks orthologs of 26 of the 30 proteins from the paraflagellar rod of *T. brucei,* a structure that is unique to trypanosomes. In contrast, the genes found in *P. infestans* have functions not specific to flagella, such as a calmodulin (*P. infestans* gene model PITG_06514), adenylate kinase (PITG_18377), and phosphatidylinositol 3-related kinase (PITG_02495).

Since alveolates and oomycetes are proposed to have a common ancestry as chromalveolates (Keeling, 2009), it was interesting to note that several proteins present in *P. infestans* and *T. thermophila* were absent from most of the other species. These included three found also in *N. gruberi* (FM145, FM154, BBC52), and two basal body proteins shared only with *T. thermophila.*

Only 44 of the *P. infestans* proteins had detectable orthologs in *H. arabidopsidis.* The fate of the missing genes in this flagella-lacking relative of *P. infestans* is the focus of the last half of this paper.

Most *P. infestans* flagellar protein candidates are induced in spores

Transcript levels for the genes during the *P. infestans* life cycle were measured to test whether their expression patterns were correlated with the timing of zoospore development. This involved microarray and quantitative RT-PCR (qRT-PCR) analysis of nonsporulating hyphae; hyphae early in sporulation, when immature sporangia are just starting to appear; sporangia purified from the hyphae; and motile zoospores released

from sporangia, 2.5 hr after cold-treating the sporangia to induce germination. Data were obtained for 224 genes. These included 116 from microarrays and 108 from qRT-PCR (the microarrays only represented part of the transcriptome; (Judelson et al., 2008; Judelson et al., 2009)). Both qRT-PCR and microarray data were obtained for several genes, with both methods yielding comparable results.

Of the 224 genes thus measured, 172 and 149 were up-regulated by >2.5-fold and >5-fold, respectively, in the early sporulation, sporangia, or zoospore stages compared to nonsporulating hyphae. This is illustrated in Fig. 2, where the data are separated into panels based on whether orthologs were detected in *H. arabidopsidis*. The data are also summarized in Fig. 1, where upwards-pointing arrows denote up-regulated genes.

Most up-regulated genes (induced by >2.5-fold) were induced early in sporulation, reached highest levels in ungerminated sporangia, and declined in zoospores. Their induction during sporulation is consistent with data from inhibitor studies that suggest that most flagella proteins are preformed in sporangia (Clark et al., 1978). This was the case, for example, for most genes encoding the principal components of the axoneme and its assembly machinery including α- and β-tubulin, all 15 genes encoding IFT proteins, all but one of the 22 outer and inner dynein arm protein genes, all flagella-associated kinesins, and all but two radial spoke proteins. Another induced class of genes encoded tubulin-tyrosine ligases, which regulate the recruitment of microtubule-interacting proteins to axonemes (Pathak et al., 2011). Three tubulin-tyrosine ligases (PITG_2721, PITG_03077, PITG_08756, corresponding to FAP67, SSA11, and MOT11, respectively) showed nearly identical induction patterns, rising about 3-fold in early

sporulation compared to nonsporulating hyphae and 10-fold in sporangia compared to hyphae.

A notable difference between induced genes concerned whether their transcripts stayed abundant in swimming zoospores. RNA for about three-quarters declined in that stage compared to sporangia, while one-quarter stayed high. This might reflect differences in the need for *de novo* synthesis after zoosporogenesis to maintain flagella function, or might be a less biologically relevant reflection on their mRNA stability. Examples of genes that had reduced RNA levels after zoospore release included many involved in forming structural components of the axoneme, including all but two of the dyneins, all radial spoke proteins, the tubulin-tyrosine ligases, and most IFT proteins. While RNAs for most genes were still at higher levels in zoospores than hyphae, PITG_20745 (IFT52) and PITG_10096 (IFT139) quickly declined to levels seen in hyphae.

Several genes involved in energy homeostasis were also induced in one or more spore-associated stages, and many maintained high mRNA levels in zoospores. One example is PITG_03419 (CPC1), which encodes a central pair adenylate kinase (reaction: ADP + ADP $\rightarrow$ AMP + ATP). Its RNA levels were three times higher in zoospores than sporangia, which may reflect a role in preserving ATP levels, or ATP/ADP ratios which affect dynein activity (Yagi, 2000). A second adenylate kinase, encoded by PITG_18379, also had the highest levels in zoospores. A gene encoding a nucleoside diphosphate kinase regulatory subunit, PITG_04513 (NDK), also had the most RNA in zoospores (reaction: GTP + ADP $\rightarrow$ GDP + ATP). The NDK catalytic subunit (genes

PITG_03634 and PITG_07886; RIB72) was also elevated in zoospores compared to hyphae, although their highest levels were in sporangia. A phosphagen kinase encoded by PITG_13551 (FCK) was induced >10-fold during early sporulation, but declined in subsequent stages.

Diversity in expression patterns was also observed for genes encoding the mastigoneme (flagella tinsel) proteins, represented by PITG_09306, PITG_10934, and PITG_12878. All had very low RNA levels in nonsporulating hyphae. PITG_10934 and PITG_12878 were induced >10-fold during early sporulation, however PITG_09306 transcripts were not induced strongly until the sporangia stage. The differences in expression patterns may reflect the fact that the proteins form distinct parts of the mastigoneme, *i.e.* shaft versus basal region, which possibly assemble at different times (Yamagishi et al., 2009).

Transcript patterns correlate with gene absence from aflagellates

The *P. infestans* orthologs of genes missing from species lacking flagella were almost always induced during sporulation, which is consistent with the genes' role in the organelle. This can be seen in Fig. 2 by comparing the panels of genes retained and lost in *H. arabidopsidis,* and Fig. 1 where spore-induced genes are marked by arrows. Of the genes lacking orthologs in *H. arabidopsidis,* 92% were up-regulated during sporulation or in spores, while only about 20% of genes having orthologs in *H. arabidopsidis* were induced. Most *P. infestans* orthologs of genes absent from *D. discoideum, O. tauri,* and *S. pombe* were also up-regulated.

About half of proteins retained in the non-flagellates had up-regulated orthologs in *P. infestans*. These include α- and β-tubulin (PITG_7996 and PITG_00156), which comprise a major part of flagella but have many other roles in growth. Another example is the *P. infestans* ortholog of *T. brucei* protein TB10.61.0160 (PITG_17354, named TB-WD in Fig. 1), which is a WD-domain protein. The human ortholog has the highest expression in flagella or cilia-forming cells, but is also expressed in other cell types, which suggests that it has multiple roles (Saeki et al., 2006).

Orthologs of the about half of the genes retained in the non-flagellates were not induced in spores. Two examples are γ-tubulin (PITG_14807, TUBG in Fig. 1) and γ-tubulin associated protein GCP2 (PITG_1154), which acts at microtubule organizing centers maintained in non-flagellates, such as spindle poles or the centrosome. An additional example is RSP12 ortholog PITG_12155, which is predicted to be a peptidyl-prolyl isomerase.

About 8% of the *P. infestans* genes lacking orthologs in non-flagellates were not up-regulated in spores, since some probably act throughout the life cycle. This was also observed in *C. reinhardtii*, where many flagella-associated genes are not induced during flagellar regeneration (Li et al., 2004). One example is PITG_20749, which encodes Rab GTPase IFT27. This protein is unique among IFT factors in that it appears to be needed for general growth based on knockdown studies (Qin et al., 2007). Many of the other non-induced, aflagellate-retained genes have roles in centrioles or basal bodies, which are structurally similar and in many (but not all) species interconvert during the cell cycle (Hoyer-Fender, 2010). Examples include PITG_1877 and PITG_19212, which encode δ-

and ε-tubulin, respectively, and the VFL3 ortholog PITG_13031, which have demonstrated roles in both organelles (Marshall et al., 2001; Wright et al., 1985). Orthologs of others such as PITG_00827, PITG_09512, and PITG_19017 (TT16871, BBC30, FTT18) have been localized to basal bodies and might also have a centriolar role (Baroncelli et al., 2016).

Most of the few remaining proteins that are expressed in all life-stages and retained in non-flagellates have no characterized function. A few have activities for which a flagella-specific role is challenging to conceptualize, such as PITG_05629, which encodes a non-canonical polyA polymerase, and PITG_02853, which encodes enoyl-CoA hydratase.

Fate of flagellar genes in *H. arabidopsidis*

To initiate an investigation into how the flagellar genes were lost from *H. arabidopsidis*, their locations in *P. infestans* were mapped on the latter's major supercontigs (Fig. 3). The flagellar genes were not clustered in *P. infestans*, with most residing within gene-rich islands. Previous studies showed that about 70% of all *P. infestans* genes and 90% of those with orthologs in other oomycetes reside in such islands, in which gene order is often conserved between *Phytophthora* spp. and *H. arabidopsidis* (Baxter et al., 2010; Haas et al., 2009; Armstrong et al., 2005). In our analysis of 30 gene-rich clusters containing a total of 696 and 920 genes from *H. arabidopsidis* and *P. infestans,* respectively, 46% of *H. arabidopsidis* genes shared some degree of synteny with their *P. infestans* counterparts, as they were found in similar

locations within each gene-rich cluster.

The lower amount of downy mildew genes in these clusters compared to *P. infestans* is consistent with the former's smaller total gene content, 14,543 versus 17,797, of which only about 7500 are orthologs (Judelson et al., 2012). Based on the positions of the flagellar genes in *P. infestans* and the commonly conserved gene order between the two species, it seems unlikely that elimination of the flagellar genes from the downy mildew involved a major catastrophic rearrangement, such as the loss of a chromosome arm.

Shared synteny between *P. infestans* and *H. arabidopsidis* was next exploited to search for remnants of the flagellar genes in the latter. Matches with significant *E* values had not been detected in our genome-wide searches, so a more sensitive exploration was pursued by studying the intergenic regions of *H. arabidopsidis* that matched locations of the genes in *P. infestans*. Shared synteny was detected for 179 regions containing the 257 flagella genes; synteny was declared if at least one of the four genes upstream of a targeted region, plus at least one of the four downstream genes, were near each other in the two species. The value should be taken as a rough estimate of the extent of shared synteny, since only draft genomes are available for analysis and gaps would cause false negatives. For comparison, shared synteny between *P. infestans* and *P. sojae* was observed for 83% of the regions containing flagellar genes.

Within the 179 regions that showed shared synteny, gene remnants were detected in 37 cases using LALIGN and TBLASTN as described in Materials and Methods. The authenticity of each match was supported by calculations of false discovery rates, and by assessing the statistical significance of each alignment using randomly shuffled

sequences (Table S2). Compared to the *P. infestans* genes, the *H. arabidopsidis* sequences typically contained multiple base changes including indels that changed the reading frame, as well as rearrangements such as larger deletions and inversions. This is illustrated in Fig. 4 for four selected genes; the total results are tallied in Table S2, and the 37 alignments are shown in Fig. S2. The regions of alignment between the *P. infestans* gene and *H. arabidopsidis* remnant usually spanned only part of the intact gene (42% on average), ranged in length from 83 to 2713 nt, and averaged 57% nucleotide identity. In comparison, flagella candidate genes having orthologs in both species averaged 83% nucleotide identity in alignments.

Interestingly, the elimination of a gene from *H. arabidopsidis* was more prone to be associated with a loss of shared synteny with *P. infestans*. Of the genes retained in *H. arabidopsidis*, about 82% showed shared synteny between the two species (based on the genes flanking the flagella gene) compared to only 66% for regions that lost the gene. The difference is significant by Fisher's Exact Test, with *P* equaling 0.02 (0.01 when recalculated for synteny with *P. sojae*).

The value of focusing searches for remnants to the region of *H. arabidopsidis* predicted to formerly bear the flagella gene is demonstrated by comparing the success here with a prior study (Judelson et al., 2012). That earlier approach checked *H. arabidopsidis* for sequences resembling 84 *C. reinhardtii* flagella genes having orthologs in *Phytophthora.* Due to weaknesses in the genome assembly available at that time, linkages between genes predicted to flank the former flagella gene in *H. arabidopsidis* were detected in only eight cases, and none of the intervening regions contained

remnants. Here we used a dramatically improved assembly in which linkages between flanking genes were detected 69% of the time. Combined with expanding the total analysis from 84 to 257 genes, this enabled the 37 remnants to be detected.

Gene loss is associated with chromosomal compaction

Gene loss in *H. arabidopsidis* was most often associated with a reduction in size of the affected region of the chromosome. This conclusion was drawn from cases where the two genes flanking the flagellar gene had the same order in *P. infestans, P. sojae,* and *H. arabidopsidis*. This is illustrated in Fig. 5A for nine representative genes. For example, orthologs of PITG_08875 had genic and intergenic regions, respectively, of 1.0 and 2.8 kb in *P. infestans* and *P. sojae*, which was reduced to 359 nt in *H. arabidopsidis.* Similarly, the 9 kb interval containing PITG_09288 was reduced to 1.7 kb in *H. arabidopsidis*. Interestingly, gene remnants were always found when the intergenic region in *H. arabidopsidis* was not very reduced compared to *Phytophthora* (PITG_04600, PITG_11526, PITG_12891). This suggests that the rate of gene loss through chromosomal deletion was higher than by base-by-base degeneration.

The distance between genes flanking the flagellar locus was smaller in *H. arabidopsidis* than *P. infestans* in 78% of cases; a similar conclusion was drawn from the prior study of eight loci (Judelson et al., 2012). Interestingly, this trend in intergenic distances is opposite to that observed for genes that share synteny but have functions unrelated to flagella. Of 120 gene pairs selected randomly from 40 different regions of their genomes, intergenic regions were smaller in *H. arabidopsidis* than *P. infestans* only

31% of the time, with median sizes of 310 and 510 nt, respectively. The loss of the flagella gene therefore appeared to facilitate the compaction of the *H. arabidopsidis* genome beyond what would be expected if only the flagella gene and its transcriptional regulatory sequences had been lost.

Oomycete genomes contain a large amount of transposon and retroelement-like sequences, but such elements did not appear to have the predominant role in eliminating flagellar genes. In a search of 32 intergenic regions that formerly contained a flagella gene, repetitive sequences were detected only in the case of IFT80-encoding gene PITG_18714. As illustrated in Fig. 5B, the 3.5 kb region containing the IFT80 gene and flanking noncoding DNA was reduced to 1.5 kb in *H. arabidopsidis.* This included about 400 nt of a Copia-like sequence having 46 relatives in the *H. arabidopsidis* genome, based on BLAST with a $E=10^{-9}$ threshold. It is possible that transposable elements had a greater role than predicted by this analysis, which was biased by its focus on cases where gene order was conserved.

DISCUSSION

Of the 257 *P. infestans* proteins identified as resembling flagellar components of other eukaryotes, most are connected firmly to the biology of that organelle since 77% of the corresponding genes are up-regulated during the life-stages when zoospore components are synthesized, and 82% are absent from the non-flagellated sister taxon *H. arabidopsidis*. While it was not our intention to identify the complete flagellar proteome of *P. infestans,* the 257 proteins are close in number to the roughly 250 estimated by 2-dimensional gel analysis to comprise the *C. reinhardtii* flagellar axoneme (Luck, 1984). Reflecting diversification of flagella during the eukaryotic radiation, about 45 proteins assigned to *C. reinhardtii* flagella lacked orthologs in *P. infestans,* while 7 *P. infestans* sequences lacked orthologs in *C. reinhardtii*. The latter were likely part of the ancestral organelle, since they match proteins in other lower eukaryotes such as *N. gruberi*. Oomycetes may also have evolved novel flagellar proteins, but identifying them is problematic due to an inability to purify enough flagella for analysis.

The highest confidence flagella-focused dataset from *P. infestans* includes about 185 proteins, based on the fraction that were up-regulated during spore formation and lacked orthologs in aflagellates. Combining both criteria reduces false positives, but at the expense of excluding proteins with multiple functions. This is especially true for proteins that may reside in both basal bodies and centrioles, which in many but not all species interconvert during the life cycle (Carvalho-Santos et al., 2011; Hoyer-Fender, 2010). While most basal body proteins have unknown functions, hints may come from their patterns of expression. Constitutive genes may encode proteins present in both basal

bodies and centrioles, such as PITG_09512, which encodes the ortholog of a *T. thermophila* basal body protein, BBC30 (Baroncelli et al., 2016). In contrast, up-regulated genes might influence the disappearance of centrioles and genesis of basal bodies during zoospore development. An example is sporulation-induced gene PITG_14588 (MKS1), which encodes the ortholog of a human protein needed for the migration of centrosomes to the sites of basal body formation (Heath, 1974). Although most oomycetes are known to use centrioles to nucleate the spindle at mitosis, their fate during the cell cycle and flagella-forming stages is not well-understood (Heath, 1974). Strikingly, *H. arabidopsidis* lacks orthologs of all centriole-associated proteins covered by this study, including POC1 which is otherwise widely distributed in eukaryotes. It is unknown whether *H. arabidopsidis* has centrioles.

The up-regulated *P. infestans* genes exhibited diverse kinetics of mRNA induction and persistence, which likely reflects the role of each gene and biology of the spores. Sporangia are an intermediate between hyphae and zoospores. While sporangia can make zoospores soon after maturation, sporangia may remain quiescent for days until zoosporogenesis is stimulated by environment (cold temperatures and free water). Zoospore release can take less than an hour after stimulation, which is consistent with studies using actinomycin-D and cycloheximide that suggested that sporangia already contain all necessary proteins (Clark et al., 1978; Penington et al., 1989). Matching that earlier finding is our observation that virtually all genes for axonemal proteins (dynein arms, radial spokes, etc.) were induced early in sporulation. Why most RNAs persist through later stages may be explained by the need to replenish proteins broken down

during sporangial quiescence, and to preserve motility, which can last a day under ideal conditions. Many late-induced genes may not be needed directly to form zoospores but may instead enable prolonged swimming and chemotaxis. Adenylate kinase PITG_03419 (CPC1), for example, may help maintain ATP levels during the motile period. The need for such functions may extend into subsequent life-stages, such as during encystment of the zoospore, extension of a germ tube from the cyst, or appressorium formation (Jupe et al., 2013).

Why and how the zoospore stage was lost from *H. arabidopsidis* is unknown. The ability of most oomycetes to form zoospores is considered to aid survival by helping them reach nutrients or optimal host infection sites. However, zoospores were lost several times during oomycete speciation, particularly in obligately pathogenic clades such as *Hyaloperonospora, Bremia,* and most *Peronospora* spp. (Beakes et al., 2012). Aflagellates germinate by extending a hyphal tube directly from the asexual spore (conidium), and this also occurs in *Phytophthora* when conditions are too warm to favor zoospore release (Jupe et al., 2013). Presumably the need to maintain genes for both germination pathways represented a heavy load for some obligately pathogenic oomycetes, providing positive selection for their loss. In this regard, the elimination of zoospores parallels the trend of metabolic pathway loss seen in many obligately pathogenic bacteria and protists (Keeling, 2004). It should be noted that zoospores are maintained in some obligately pathogenic oomycetes, including the *Albugo* white rusts and certain downy mildews within *Peronospora*.

One may speculate that destruction of the pathway began in one allele of a flagellar gene that experienced a spontaneous mutation or insertion of a transposable element. The latter are common in most oomycete genomes, and have influenced the evolution of several gene families; *H. arabidopsidis* has a 100 Mb genome, with 42% repetitive DNA content (Baxter et al., 2010; Haas et al., 2009; Raffaele et al., 2010). Considering that oomycetes are diploid, how the remaining functional allele was lost is intriguing. Loss of heterozygosity due to mitotic crossing-over appears frequent in some oomycetes (Lamour et al., 2012) but this might have been ecologically deleterious if the elimination of the remaining functional allele was the result. Alternatively, the first step in zoospore loss could have been a dominant-negative regulatory mutation.

Regardless of the nature of the event that initiated zoospore loss from *H. arabidopsidis*, examining the fate of its remaining flagellar genes was of interest in light of the complexity of oomycete genomes, and the dearth of information about their evolution outside of effector families, retroelements, and occasional gene fusions (Baxter et al., 2010; Morris et al., 2009; Judelson, 2002; Judelson and Ah-Fong, 2010). A fair degree of shared synteny between *Phytophthora* and *H. arabidopsidis* allowed us to identify remnants of many flagella genes and observe a higher propensity for loss of synteny at the affected regions. One factor maintaining gene order in oomycetes is probably their small intergenic distances, most commonly 400-500 nt; this likely restricts the viability of illegitimate recombination events such as unequal crossovers or transposon insertions. Degradation of a flagella gene presumably results in a larger target for viable recombination.

We also observed compaction of the *H. arabidopsidis* genome at most affected loci. Such changes are estimated to account for a reduction of about 1 Mb; overall shrinkage is likely greater since some metabolic genes, protein kinases, and effectors are also reported to be absent in *H. arabidopsidis* (Baxter et al., 2010; Judelson and Ah-Fong, 2010). Most other obligately parasitic eukaryotes such as *Cryptosporidium* and *Entamoeba* have more dramatically reduced genomes as a consequence of gene loss or fusion, intron reduction, and a diminution of intergenic distances (Keeling, 2004). Compaction, however, is not restricted to obligates. For example, the *O. tauri* genome is much smaller than its relatives (Derelle et al., 2006).

It is interesting to note that the loss from *H. arabidopsidis* of a gene seemed to be a better predictor of its flagellar role than whether *O. tauri* lacked an ortholog of a *C. reinhardtii* gene; the fraction of retained flagellar genes was 35% greater in *O. tauri* than in the downy mildew. A recent survey showed that other plants that lack a flagella stage have also retained multiple flagella-associated proteins (Hodges et al., 2011). Retention of a flagella gene is likely explained by its acquisition of additional roles prior to the loss of the organelle. The fraction of retained genes also may reflect the time of divergence of *H. arabidopsidis* and *O. tauri* from their flagellated sister taxa, as well as from plants and the Stramenopile kingdom, which includes oomycetes, from their shared ancestor. Plants and Stramenopiles diverged more than one billion years ago during the Mesoproterozoic era, the two algae from each other 0.5 to 1 billion years ago during the Neoproterozoic (although flagellar loss is probably more recent), and downy mildews from *Phytophthora* less than 65 million years ago with oomycetes themselves having a Neoproterozoic origin

(Dick, 2001; Strullu-Derrien et al., 2011; Omoto and Witman, 1981). Our understanding of oomycete evolution is still developing, and current phylogenetic schemes nest many of the non-flagellated plant pathogens within *Phytophthora* (Runge et al., 2011). Dating the oomycete radiation is challenging in the absence of a clear fossil record, and it is interesting to consider that gene remnants as reported here for *H. arabidopsidis* may inform us about the timing of key events. The fact that gene remnants can still be detected suggests that flagella loss in *Hyaloperonospora* was relatively recent.

MATERIALS AND METHODS

Identification of flagellar protein genes

Candidate proteins from the sources cited in Results were used to search a *P. infestans* protein database using BLASTP. This was performed on a local server or with tools provided for *P. infestans* at the Broad Institute of Harvard and MIT (www.broadinstitute.org, assembly v. 2). The veracity of ortholog assignments was tested using the reciprocal best hit method, by comparing *P. infestans* hits to databases of proteins from *C. reinhardtii* (http://genome.jgi-psf.org, version 3), *Tetrahymena* (ciliate.org, TIGR 2006 version), *T. brucei* database (tritrypdb.org, versions 2.5 and 3), or other species at Genbank. In the summary in Table S1, the common name of the *C. reinhardtii* ortholog as informed by *ChlamyCyc* (chlamyto.mpimp-golm.mpg.de) is used, except for dyneins that were recently renamed (Hom et al., 2011).

If several strong hits were obtained in *P. infestans,* then the sequences were compared phylogenetically to probe orthology further. This was also done whenever the BLASTP *E* value was below 10^{-15}, which was common for some short proteins. This involved performing alignments using the SEAVIEW implementation of MUSCLE followed by PhyML to generate maximum likelihood trees (Gouy et al., 2010). In a few cases the *P. infestans* gene appeared to have undergone a duplication; after excluding the possibility that the duplication was an assembly artifact, both hits were placed on the list of candidate flagellar proteins, annotated with suffixes a and b. If no hits were detected against *P. infestans* proteins, then before concluding that the gene was absent, the test sequences were compared to the *P. infestans* assembly to check for unannotated genes

and to gene models from *P. sojae* using the database at Virginia Bioinformatics Institute (http://vmd.vbi.vt.edu/toolkit). This database was also used to identify *H. arabidopsidis* orthologs of the *P. infestans* genes as described above, using both releases 6.0 and 8.3. This included searching annotated genes using TBLASTN, followed by checking the assembly when hits were not detected.

Ortholog mapping for Fig. 1 was performed using OrthoMCL (version 2.0.3) with the databases listed above as well as those for *C. elegans* (http://wormbase.org, release WS230), *D. discoideum* (http://dictybase.org), *G. lamblia* (http://giardiadb.org, v. 2.5), *H. sapiens* (http://uniprot.org, release 2011-05), *N. gruberi* (http://genome.jgi-psf.org, v. 1), *O. tauri* (http://genome.jgi-psf.org, v. 2), *S. pombe* (http://www.pombase.org, v. 12), and *T. thermophila* (http://ciliate.org, v. 2008), and *Trypanosoma brucei* (http://tritrypdb.org/, v. 4.1). *P. infestans, C. reinhardtii, and N. gruberi* proteins were used as query sequences in OrthoMCL, which was executed using an *E*-value of 10^{-5} and 30% identity as a cut-off, and an inflation parameter of 1.5. Matches with *E*-values below 10^{-10}, and cases where OrthoMCL detected close sequences but could not easily differentiate co-orthologs, were double-checked using the reciprocal best hit method.

Analysis of shared synteny and identification of gene remnants

P. infestans flagellar protein candidates plus the four preceding and following genes on the chromosome were searched against *H. arabidopsidis* assembly v. 6, and match coordinates were compared to assess shared synteny. This was claimed if linkage was observed between at least one upstream and one downstream gene, recognizing that

some false negatives likely occurred due to gaps or assembly errors; N50 values in the *P. infestans* and *H. arabidopsidis* assemblies are 1590 and 116 kb, respectively. Similar criteria were used to compare *P. infestans* and *P. sojae*, using colinearity maps generated by Genome Project Solutions (Hercules, CA). When synteny was detected between *P. infestans* and *H. arabidopsidis,* explorations for gene remnants were performed initially by a DNA-DNA search using LALIGN, in which the *P. infestans* gene trimmed of introns was checked against the relevant region of *H. arabidopsidis*. If matches were not detected, then TBLASTN was used to compare the *P. infestans* protein against *H. arabidopsidis* DNA. False discovery rates were calculated as described in Table S2, and the statistical significance of each LALIGN alignment was calculated using 200 random shuffles of the *H. arabidopsidis* sequence using PRSS (http://fasta.bioch.virginia.edu).

Growth conditions and RNA extractions

Isolate 1306 (A1 mating type, from tomato) was cultured in the dark on rye-sucrose media containing 40 units/ml of nystatin at 18°C. Nonsporulating mycelia were obtained by inoculating 25 ml of rye-sucrose broth in 100 mm plastic dishes with 10^3 sporangia. Tissue was harvested after 72 hr, about one day before sporulation would normally begin. To induce sporulation in a semi-synchronous manner, hyphal mats were removed from broth and placed on 1.5% agar. The sample was then harvested after 20 hr, which is when new sporangia start to be observed. Sporangia were purified from cultures by adding 10 ml of water to each plate, rubbing with a bent glass rod, and passing the resulting fluid through 50 μm mesh to remove hyphal fragments. The

sporangia were then pelleted by centrifugation for 5 min. To obtain zoospores, sporangia were placed in water at 10°C for 4 hr. Most sporangia released their zoospores by 90 min, and the zoospores were pelleted after an additional 60 min. RNA was prepared by grinding tissue in liquid nitrogen, followed by the use of the RNeasy Plant Mini kit (Qiagen, Valencia, CA, USA). RNA quality was assessed spectrophotometrically and by electrophoresis.

Expression analysis

Transcript levels were calculated using Affymetrix microarrays and qRT-PCR. Expression values obtained as described below were combined, and then subjected to per-gene normalization and hierarchical clustering using GeneSpring software (Agilent Technologies, Foster City, CA USA). The microarray data are archived in NCBI GEO as series GSE9623 and GSE13580 (Judelson et al., 2008; Judelson et al., 2009); the arrays themselves are no longer available due to a limited production run. They were designed to detect 15,646 unigenes (mostly EST-based), which appear to represent about two-thirds of the 17,797 genes predicted for *P. infestans* (Haas et al., 2009). Array data were preprocessed using Affymetrix MAS 5.1 software using two biological replicates, genes lacking "present calls" were discarded, and then the remaining data were subjected to median normalization using GeneSpring software (Agilent Technologies, Foster City, California USA).

qRT-PCR employed DNAse-treated RNA, pooled from two biological replicates, which was reverse-transcribed using oligo-dT with a first-strand synthesis kit from

Invitrogen (Carlsbad, CA, USA). Amplifications employed hot-start *Taq* polymerase with primers targeted to the 3' regions of genes, typically yielding amplicons of 150-225, using SYBR Green as a reporter. Reactions were performed in duplicate using the following conditions: one cycle of 95°C for 8 min, and 35 cycles of 95°C for 20 s, 55°C for 20 s, and 72°C for 30 s. Controls lacking reverse transcriptase and melting curves were used to test the data. Results were normalized based on primers for a constitutively expressed gene encoding ribosomal protein S3a, and expression was determined by the ΔΔCT method.

Figure 1. Phylogenetic distribution of flagella-associated components detected in *P. infestans.* Orthologs (black circles) were identified from seven eukaryotes capable of forming flagella or cilia *(C. reinhardtii, H. sapiens, N. gruberi, T. thermophila, C. elegans, G. lamblia)* and four that do not make the organelle *(S. pombe, D. discoideum, O. tauri, H. arabidopsidis)* using *P. infestans, C. reinhardtii, and N. gruberi* proteins as query sequences. Searches were performed using OrthoMCL and supplemented with the reciprocal best hit approach. Genes are grouped by functional categories, although some categories may overlap.

Chlamydomonas
Homo sapiens
Naegleria
Tetrahymena
Trypanosoma
Caenorhabditis
Giardia
Schizosaccharomyces
Dictyostelium
Ostreococcus
Hyaloperonospora
Expression (Pi)
Intraflagellar Transport
Inner Dynein Arm
Centriole
Central Pair
Basal Body
Radial Spoke
Protein Interaction
Outer Dynein Arm
Metabolism
Tubulin-related
Regulatory
Uncharacterized

Figure 2. Expression of predicted flagella-associated genes from *P. infestans*. RNA levels were measured in nonsporulating mycelia (NSM), cultures 20 hours after being induced to sporulate (20H), purified ungerminated sporangia (SPO), and swimming zoospores (ZOO). Panels are split based on the presence or absence of orthologs in *H. arabidopsidis*. Numbers to the left of each image represent the accession number from the Broad Institute database (trimmed of the PITG prefix), and lettering to the right of each image represents the common names of the gene products as listed in Table S1.

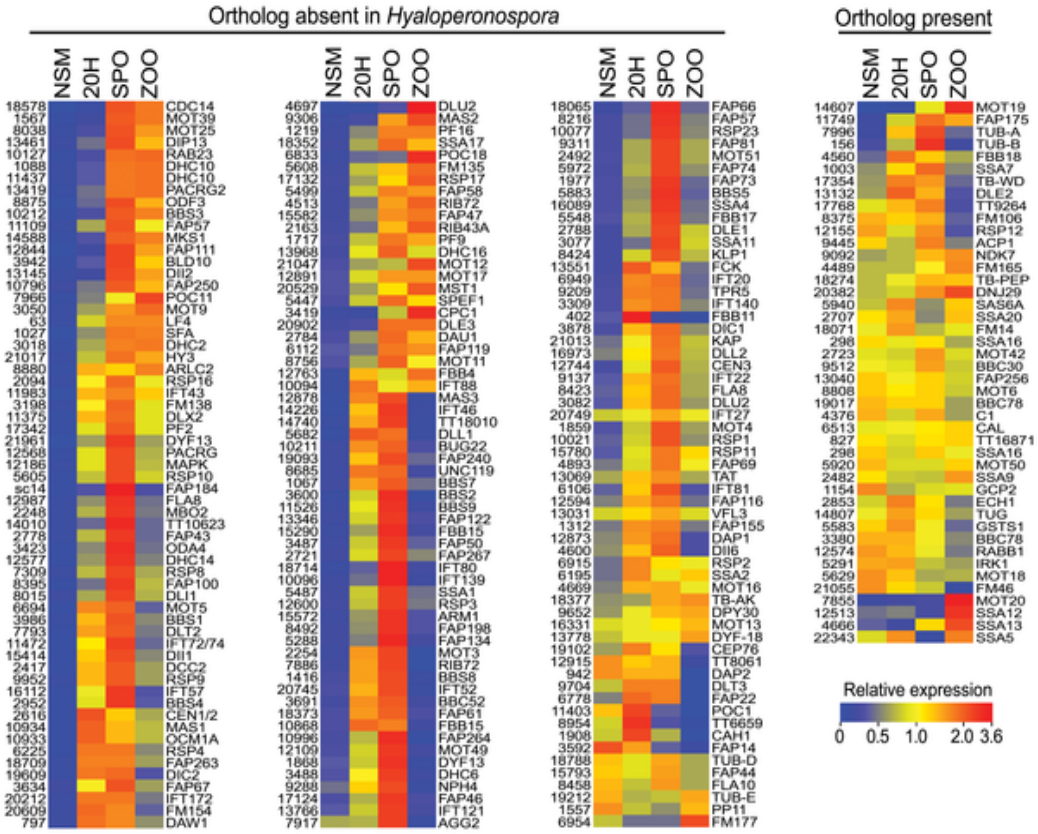


Figure 3. Location of predicted flagella-associated genes from *P. infestans*. Shown are maps of the 40 largest supercontigs from the assembly, with positions of genes marked by vertical bars. An asterisk above the bar indicates that the gene is intact in *H. arabidopsidis,* and an arrow denotes the presence of a gene remnant in *H. arabidopsidis*.

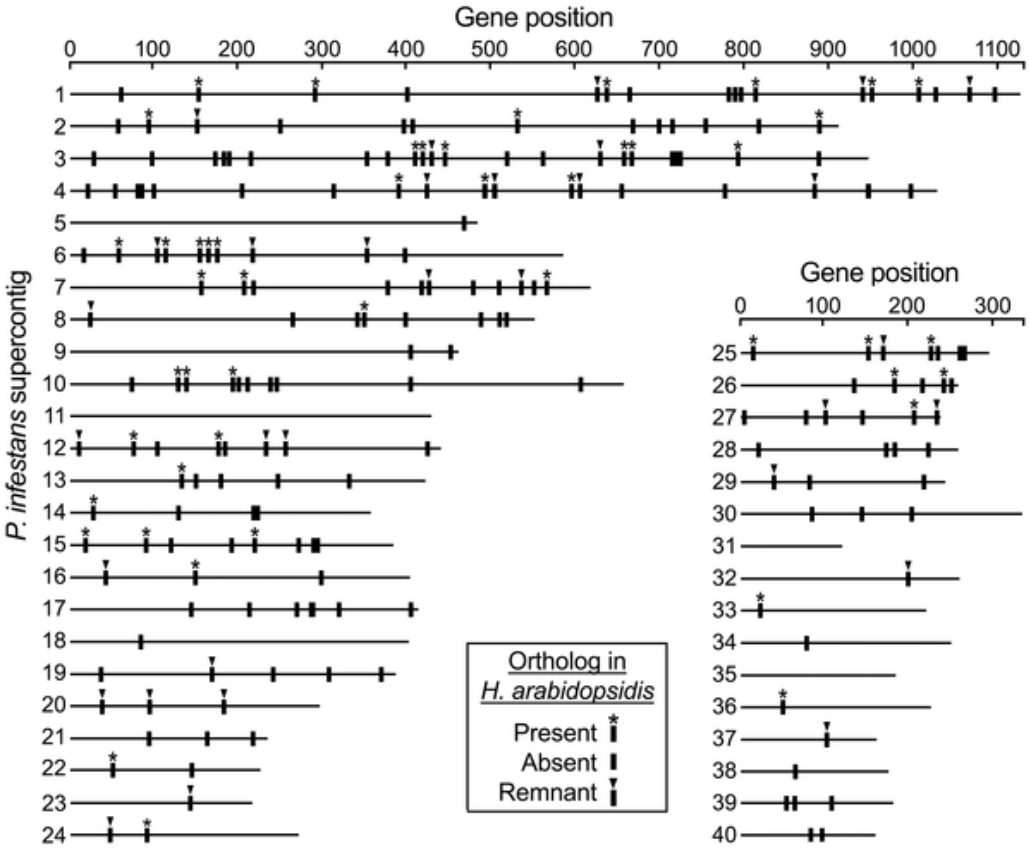


Figure 4. Examples of gene remnants in *H. arabidopsidis.* **A,** Comparisons of three *P. infestans* genes representing orthologs of *C. reinhardtii* BBS2, KAP, and MOT28 and the three corresponding intergenic regions from *H. arabidopsidis*. The images are redrawn from the dot matrix output of NCBI Blast 2.2.26, with diagonals representing regions of similarity. **B,** Alignment of portion of remnant of IFT172 gene from *H. arabidopsidis* (HA) with *P. infestans* gene PITG_20212 (PI). Numbers at right indicate the position within the *P. infestans gene.* The *H. arabidopsidis* sequences correspond to Scaffold 1, positions 1119536 to 1119160, of assembly version 8.3. The total aligned segment spanned 1597 nt, but only 373 nt is shown.

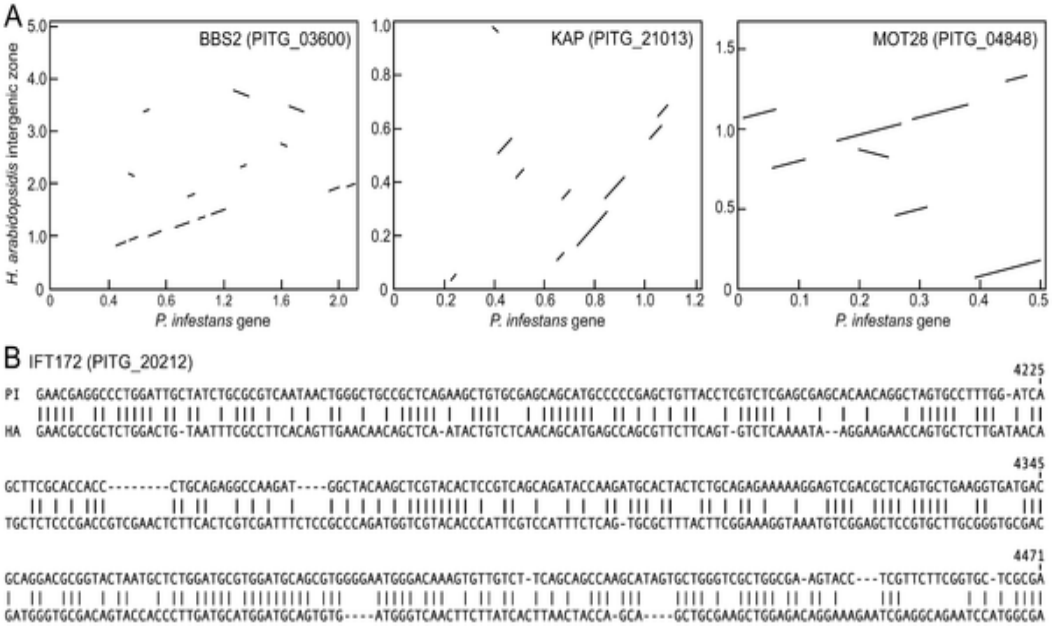
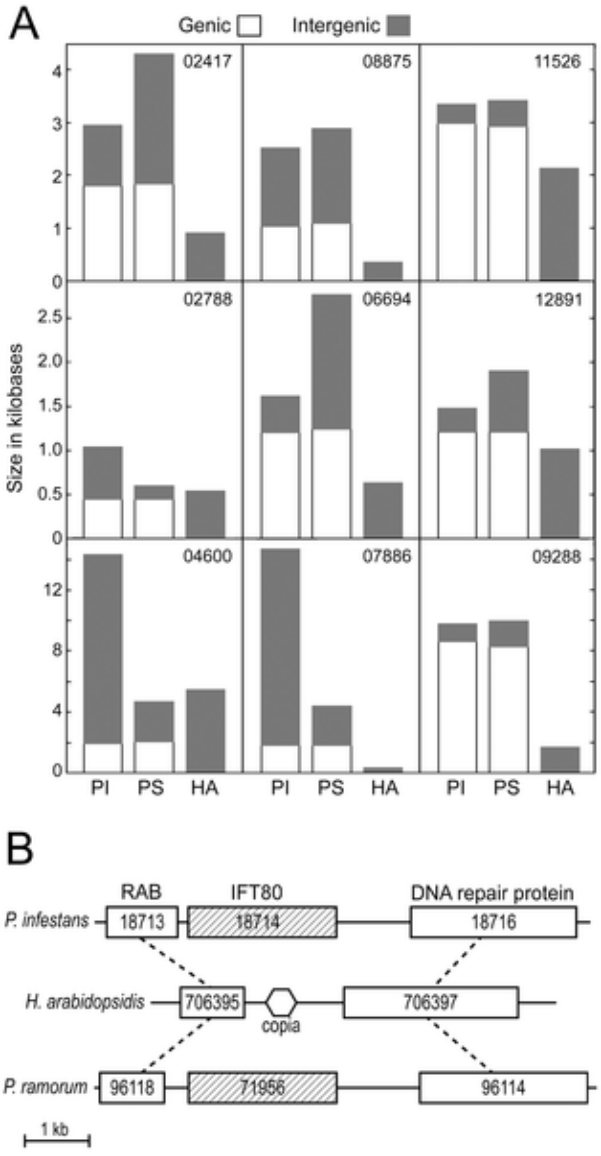


Figure 5. Examples of loci experiencing gene loss in the downy mildew. A, Sizes of genic (white bars) and intergenic (dark bars) regions in *H. arabidopsidis* (HA), *P. infestans* (PI), and *P. sojae* (PS). Genic regions include the entire predicted primary transcript, and the number of the corresponding *P. infestans* gene is marked in the upper right-hand corner of each panel. **B,** *H. arabidopsidis* locus experiencing loss of IFT80 gene, compaction, and transposable element insertion. Dashed lines indicate flanking orthologs in *P. infestans* and *P. ramorum,* boxes represent genes (with database number indicated), and the hexagon in the *H. arabidopsidis* map represents the location of a Copia-like sequence. Maps are drawn to scale using corrected gene models, and illustrate how the region between the RAB and DNA repair protein genes is reduced from 3.5 kb in *Phytophthora* to 1.5 kb in *H. arabidopsidis.*


A
Genic
Intergenic
02417
08875
11526
02788
06694
12891
04600
07886
09288
Size in kilobases
PI
PS
HA
B
RAB
IFT80
DNA repair protein
P. infestans
18713
18714
18716
H. arabidopsidis
706395
copia
706397
P. ramorum
96118
71956
96114
1 kb

References:

Ah-Fong AM, Judelson HS (2011) New role for Cdc14 phosphatase: localization to basal bodies in the oomycete Phytophthora and its evolutionary coinheritance with eukaryotic flagella. PLoS One 6: e16725.

Armstrong MR, Whisson SC, Pritchard L, Bos JI, Venter E, et al. (2005) An ancestral oomycete locus contains late blight avirulence gene Avr3a, encoding a protein that is recognized in the host cytoplasm. Proceeds of the National Academy of Sciences USA 102: 7766-7771.

Baron DM, Ralston KS, Kabututu ZP, Hill KL (2007) Functional genomics in Trypanosoma brucei identifies evolutionarily conserved components of motile flagella. Journal of Cell Science 120: 478-491.

Baxter L, Tripathy S, Ishaque N, Boot N, Cabral A, et al. (2010) Signatures of adaptation to obligate biotrophy in the Hyaloperonospora arabidopsidis genome. Science 330: 1549-1551.

Beakes GW, Glockling SL, Sekimoto S (2012) The evolutionary phylogeny of the oomycete "fungi". Protoplasma 249: 3-19.

Carvalho-Santos Z, Azimzadeh J, Pereira-Leal JB, Bettencourt-Dias M (2011) Evolution: Tracing the origins of centrioles, cilia, and flagella. Journal of Cell Biology 194: 165-175.

Charles M, Tang H, Belcram H, Paterson A, Gornicki P, et al. (2009) Sixty million years in evolution of soft grain trait in grasses: emergence of the softness locus in the common ancestor of Pooideae and Ehrhartoideae, after their divergence from Panicoideae. Molecular Biology and Evolution 26: 1651-1661.

Clark MC, Melanson DL, Page OT (1978) Purine metabolism and differential inhibition of spore germination in Phytophthora infestans. Canadian Journal of Microbiology 24: 1032-1038.

Dawe HR, Smith UM, Cullinane AR, Gerrelli D, Cox P, et al. (2007) The Meckel-Gruber Syndrome proteins MKS1 and meckelin interact and are required for primary cilium formation. Human Molecular Genetics 16: 173-186.

Derelle E, Ferraz C, Rombauts S, Rouze P, Worden AZ, et al. (2006) Genome analysis of the smallest free-living eukaryote Ostreococcus tauri unveils many unique features. Proceedings of the National Academy of Sciences USA 103: 11647-11652.

Dick MW (2001) Straminipilous Fungi. Dordrecht: Kluwer Academic Publishers. 670 p.

Eckhart L, Uthman A, Sipos W, Tschachler E (2006) Genome sequence comparison reveals independent inactivation of the caspase-15 gene in different evolutionary lineages of mammals. Molecular Biology and Evolution 23: 2081-2089.

Erickson HP (2007) Evolution of the cytoskeleton. Bioessays 29: 668-677.

Fritz-Laylin LK, Prochnik SE, Ginger ML, Dacks JB, Carpenter ML, et al. (2010) The genome of Naegleria gruberi illuminates early eukaryotic versatility. Cell 140: 631-642.

Gouy M, Guindon S, Gascuel O (2010) SeaView version 4: a multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Molecular Biology and Evolution 27: 221-224.

Haas BJ, Kamoun S, Zody MC, Jiang RH, Handsaker RE, et al. (2009) Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. Nature 461: 393-398.

Heath IB (1974) Centrioles and mitosis in some oomycetes. Mycologia 66: 354-359.

Hodges ME, Wickstead B, Gull K, Langdale JA (2011) Conservation of ciliary proteins in plants with no cilia. BMC Plant Biology 11: 185.

Hom EF, Witman GB, Harris EH, Dutcher SK, Kamiya R, et al. (2011) A unified taxonomy for ciliary dyneins. Cytoskeleton 68: 555-565.

Hoyer-Fender S (2010) Centriole maturation and transformation to basal body. Seminars in Cell and Developmental Biology 21: 142-147.

Huang X, Miller W (1991) A time-efficient, linear-space local similarity algorithm. Advances in Applied Mathematics 12: 373-381.

Inglis PN, Boroevich KA, Leroux MR (2006) Piecing together a ciliome. Trends Genet 22: 491-500.

Jekely G, Arendt D (2006) Evolution of intraflagellar transport from coated vesicles and autogenous origin of the eukaryotic cilium. Bioessays 28: 191-198.

Judelson HS (2002) Sequence variation and genomic amplification of a family of Gypsy-like elements in the oomycete genus Phytophthora. Molecular Biology and Evolution 19: 1313-1322.

Judelson HS, Ah-Fong AM (2010) The kinome of Phytophthora infestans reveals oomycete-specific innovations and links to other taxonomic groups. BMC Genomics 11: 700.

Judelson HS, Ah-Fong AM, Aux G, Avrova AO, Bruce C, et al. (2008) Gene expression profiling during asexual development of the late blight pathogen Phytophthora infestans reveals a highly dynamic transcriptome. Molecular Plant-Microbe Interactions 21: 433-447.

Judelson HS, Blanco FA (2005) The spores of Phytophthora: weapons of the plant destroyer. Nature Microbiology Reviews 3: 47-58.

Judelson HS, Narayan RD, Ah-Fong AM, Kim KS (2009) Gene expression changes during asexual sporulation by the late blight agent Phytophthora infestans occur in discrete temporal stages. Molecular Genetics and Genomics 281: 193-206.

Judelson HS, Randall TA (1998) Families of repeated DNA in the oomycete Phytophthora infestans and their distribution within the genus. Genome 41: 605-615.

Keeling PJ (2004) Reduction and compaction in the genome of the apicomplexan parasite Cryptosporidium parvum. Developmental Cell 6: 614-616.

Keeling PJ (2009) Chromalveolates and the evolution of plastids by secondary endosymbiosis. Journal of Eukaryotic Microbiology 56: 1-8.

Keeling PJ, Slamovits CH (2005) Causes and effects of nuclear genome reduction. Current Opinion in Genetics and Devevelopment 15: 601-608.

Kilburn CL, Pearson CG, Romijn EP, Meehl JB, Giddings TH, Jr., et al. (2007) New Tetrahymena basal body protein components identify basal body domain structure. Journal of Cell Biology 178: 905-912.

Lai J, Ma J, Swigonova Z, Ramakrishna W, Linton E, et al. (2004) Gene loss and movement in the maize genome. Genome Research 14: 1924-1931.

Lamour K, Mudge J, Gobena D, Hurtado-Gonzales OP, Schmutz J, et al. (2012) Genome sequencing and mapping reveal loss of heterozygosity as a mechanism for rapid adaptation in the vegetable pathogen Phytophthora capsici. Molecular Plant-Microbe Interactions Jun 20.

Lauwaet T, Smith AJ, Reiner DS, Romijn EP, Wong CC, et al. (2011) Mining the Giardia genome and proteome for conserved and unique basal body proteins. International Journal of Parasitology 41: 1079-1092.

Lechtreck KF, Luro S, Awata J, Witman GB (2009) HA-tagging of putative flagellar proteins in Chlamydomonas reinhardtii identifies a novel protein of intraflagellar transport complex B. Cell Motility and Cytoskeleton 66: 469-482.

LeDizet M, Beck JC, Finkbeiner WE (1998) Differential regulation of centrin genes during ciliogenesis in human tracheal epithelial cells. American Journal of Physiology 275: L1145-1156.

Levesque CA, Brouwer H, Cano L, Hamilton JP, Holt C, et al. (2010) Genome sequence of the necrotrophic plant pathogen Pythium ultimum reveals original pathogenicity mechanisms and effector repertoire. Genome Biology 11: R73.

Li JB, Gerdes JM, Haycraft CJ, Fan Y, Teslovich TM, et al. (2004) Comparative genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. Cell 117: 541-552.

Li L, Stoeckert CJ, Jr., Roos DS (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Research 13: 2178-2189.

Liu YJ, Hodson MC, Hall BD (2006) Loss of the flagellum happened only once in the fungal lineage: phylogenetic structure of kingdom Fungi inferred from RNA polymerase II subunit genes. BMC Evolutionary Biology 6: 74.

Luck DJ (1984) Genetic and biochemical dissection of the eucaryotic flagellum. Journal of Cell Biology 98: 789-794.

Marshall WF, Vucica Y, Rosenbaum JL (2001) Kinetics and regulation of de novo centriole assembly. Implications for the mechanism of centriole duplication. Current Biology 11: 308-317.

Merchant SS, Prochnik SE, Vallon O, Harris EH, Karpowicz SJ, et al. (2007) The Chlamydomonas genome reveals the evolution of key animal and plant functions. Science 318: 245-250.

Moran NA (2003) Tracing the evolution of gene loss in obligate bacterial symbionts. Current Opinion in Microbiology 6: 512-518.

Morrison HG, McArthur AG, Gillin FD, Aley SB, Adam RD, et al. (2007) Genomic minimalism in the early diverging intestinal parasite Giardia lamblia. Science 317: 1921-1926.

Morris PF, Schlosser LR, Onasch KD, Wittenschlaeger T, Austin R, et al. (2009) Multiple horizontal gene transfer events and domain fusions have created novel regulatory and metabolic networks in the oomycete genome. PLoS One 4: e6133.

Omoto CK, Witman GB (1981) Functionally significant central-pair rotation in a primitive eukaryotic flagellum. Nature 290: 708-710.

Ostrowski LE, Blackburn K, Radde KM, Moyer MB, Schlatzer DM, et al. (2002) A proteomic analysis of human cilia: identification of novel components. Molecular Cell Proteomics 1: 451-465.

Pathak N, Austin CA, Drummond IA (2011) Tubulin tyrosine ligase-like genes ttll3 and ttll6 maintain zebrafish cilia structure and motility. Journal of Biological Chemistry 286: 11685-11695.

Pazour GJ, Agrin N, Leszyk J, Witman GB (2005) Proteomic analysis of a eukaryotic cilium. Journal of Cell Biology 170: 103-113.

Penington CJ, Iser JR, Grant BR, Gayler KR (1989) Role of RNA and protein synthesis in stimulated germination of zoospores of the pathogenic fungus Phytophthora palmivora. Experimental Mycology 13: 158-168.

Phirke P, Efimenko E, Mohan S, Burghoorn J, Crona F, et al. (2011) Transcriptional profiling of C. elegans DAF-19 uncovers a ciliary base-associated protein and a CDK/CCRK/LF2p-related kinase required for intraflagellar transport. Developmental Biology 357: 235-247.

Portman N, Lacomble S, Thomas B, McKean PG, Gull K (2009) Combining RNA interference mutants and comparative proteomics to identify protein components and dependences in a eukaryotic flagellum. Journal of Biological Chemistry 284: 5610-5619.

Qin H, Wang Z, Diener D, Rosenbaum J (2007) Intraflagellar transport protein 27 is a small G protein involved in cell-cycle control. Current Biology 17: 193-202.

Raffaele S, Farrer RA, Cano LM, Studholme DJ, MacLean D, et al. (2010) Genome evolution following host jumps in the Irish potato famine pathogen lineage. Science 330: 1540-1543.

Runge F, Telle S, Ploch S, Savory E, Day B, et al. (2011) The inclusion of downy mildews in a multi-locus-dataset and its reanalysis reveals a high degree of paraphyly in Phytophthora. IMA Fungus 2: 163-171.

Saeki M, Irie Y, Ni L, Yoshida M, Itsuki Y, et al. (2006) Monad, a WD40 repeat protein, promotes apoptosis induced by TNF-alpha. Biochemistry Biophysics Research Communications 342: 568-572.

Slusarenko AJ, Schlaich NL (2003) Downy mildew of Arabidopsis thaliana caused by Hyaloperonospora parasitica (formerly Peronospora parasitica). Molecular Plant Pathology 4: 159-170.

Stolc V, Samanta MP, Tongprasit W, Marshall WF (2005) Genome-wide transcriptional analysis of flagellar regeneration in Chlamydomonas reinhardtii identifies orthologs of ciliary disease genes. Proceedings of the National Academy of Sciences USA 102: 3703-3707.

Strullu-Derrien C, Kenrick P, Rioult JP, Strullu DG (2011) Evidence of parasitic Oomycetes (Peronosporomycetes) infecting the stem cortex of the Carboniferous seed fern Lyginopteris oldhamia. Proceedings of the Royal Society. Biological Sciences 278: 675-680.

Wang Z, Fan ZC, Williamson SM, Qin H (2009) Intraflagellar transport (IFT) protein IFT25 is a phosphoprotein component of IFT complex B and physically interacts with IFT27 in Chlamydomonas. PLoS One 4: e5384.

Wei M, Sivadas P, Owen HA, Mitchell DR, Yang P (2010) Chlamydomonas mutants display reversible deficiencies in flagellar beating and axonemal assembly. Cytoskeleton 67: 71-80.

Wright RL, Salisbury J, Jarvik JW (1985) A nucleus-basal body connector in Chlamydomonas reinhardtii that may function in basal body localization or segregation. Journal of Cell Biology 101: 1903-1912.

Yagi T (2000) ADP-dependent microtubule translocation by flagellar inner-arm dyneins. Cell Struct Funct 25: 263-267.

Yamagishi T, Motomura T, Nagasato C, Kawai H (2009) Novel proteins comprising the stramenopile tripartite mastigoneme in Ochromonas danica (chrysophyceae). Journal of Phycology 45: 1100-1105.

Yang P, Diener DR, Yang C, Kohno T, Pazour GJ, et al. (2006) Radial spoke proteins of Chlamydomonas flagella. Journal of Cell Science 119: 1165-1174.

Conclusions

Oomycetes have been named as one of the notorious plant pathogens that are causing major destructions to the crops (Fisher et al., 2012). Limited genome-wide studies of oomycetes have been done, most of which were focused on effector biology (Haas et al., 2009; Raffaele et al., 2010; Lévesque et al., 2010; Jiang et al., 2008). These studies have shown the roles and evolution of virulence genes and how they have expanded in certain genomes (Haas et al., 2009). However, in order to control oomycetes, knowing about effector biology alone is not sufficient. Other important aspects to control oomycetes include timely diagnosis, understanding the biology of these organisms in terms of evolution, population structure, adaptation and life style.

In this thesis, we have used bioinformatics approaches to identify and resolve some of the issues that might help in controlling oomycetes, specifically *P. infestans*. Some work in this thesis is directly related to controlling *P. infestans,* as in Chapter I and Chapter III, while Chapters II, IV and V will help in better understanding of the biology of oomycetes and in the long term may provide novel mechanisms to control them. We started with building a pipeline for correct diagnosis of *P. infestans* in the fields (Chapter I), which first involved re-annotating the *P. infestans* genome with RNA-Seq data; this identified and corrected problems associated with gene models that in turn will improve downstream analyses of genes and regulatory pathways (Chapter II), identify differences in lineages of *P. infestans* that will in turn help us identify the molecular basis of phenotypic differences (Chapter III), identify the differences and similarities in two oomycetes that have different life-styles, *P. infestans* and *Py. ultimum* (Chapter IV), and

compare the evolution of eukaryotic species that have lost or retained flagella (Chapter V).

In Chapter I, we have used both genomic and proteomics data in a systematic way to identify species-specific antigens to correctly identify *P. infestans*. The pipeline used protein and nucleotide similarity to find regions of proteins specific to *P. infestans,* which were then assessed to see if they fulfill our criteria for usage as antigens for immunoassay development. This pipeline can be applied to any oomycete with minimal input from the user and is fast in terms of prediction of antigenic peptides. This study has shown a good way of using bioinformatics approaches to filter a large number of proteins based on several criteria that we had set up; without the pipeline, the analysis would have taken an inordinate amount of time to come up with a list of putative candidates. The only drawback is the incorrect predictions of gene models in the species that are used. We were able to filter out any problematic gene models in *P. infestans* by visualization of RNA-Seq data in IGV. For other species, we used tblastn to check similarity that might be missed due to incorrect gene models. The resulting candidates from this pipeline were then used to generate both monoclonal and polyclonal antibodies. As of now two of the proteins (PITG_16620 and PITG_09393) generated from this pipeline look promising where we are getting the expected sized bands for the respective proteins in the western blot using the peptide. One way to make this pipeline more efficient in terms of removing the incorrect gene predictions is to start with Blastn of the entire gene and not the protein blast. This will eliminate the incorrect protein predictions due to erroneous intron

boundaries but it will not eliminate genes which have a different N-terminus or an extended C-terminus.

Correct genome annotation is the first step to understand the biology of any organism. We have utilized RNA-Seq data from many developmental stages to correct existing gene models and identify new transcripts that were missing in the *P. infestans* genome annotation (Chapter II). This is the first study that has used RNA-Seq data to correctly identify the exon-intron junctions, UTR additions/modifications, non-coding RNAs and alternative splicing in the *P. infestans* genome. This study has opened up many new exciting aspects in understanding the expression patterns, functional roles and cis and trans-regulating ncRNAs. We have modified gene models of 67% of the genes that were present in the genome and added 482 new genes that were missing in previous annotation.

Only one study in *P. infestans* has found a long non-coding RNA that was upregulated in the biotrophic phase, which was encoded by multiple loci in the genome (Avrora et al., 2007). We have identified 8,115 unique ncRNAs that are present in *P. infestans* (Chapter II). This is the first study to find ncRNAs at the genome-wide level in any oomycete. Most of these ncRNAs are expressed at very low levels but we found 453 ncRNAs that have >5x differential expression in different developmental stages. We also found that 40 ncRNAs overlap the UTRs of protein-coding genes. The addition of ncRNA in the genome annotation will help us understand the role of ncRNAs in *P. infestans*. Non-coding RNAs are well-studied in fungal genomes where they play important roles in silencing foreign DNA, development, dimorphism and stress responses

(Dang et al., 2011; Nunes et al., 2011; Morrison et al., 2012; Amaral et al., 2013). Further characterization of ncRNAs discovered in this study can help identify targets of ncRNAs in *P. infestans* and find if any of these ncRNAs are involved in pathogenicity or silencing of certain genes.

This is the also the first study that has looked at alternative splicing events in *P. infestans* at a genome-wide level (Chapter II). We found that some of the alternatively-spliced genes have one form that is secreted and another that is not secreted; their expression patterns indicated that the secreted isoform was expressed at higher levels in the biotrophic phase of infection. We also found that 437 genes with isoforms have >5x differential expression of these isoforms in different developmental stages. Studies on fungal genomes have shown different targeting of isoforms to different organelles (Freitag et al., 2012). Since we have seen that some of these isoforms are differentially regulated; it will be interesting to find what role do they play in the development of specific stages.

The correction of gene models, functional annotation of genome and secretome analysis along with identification of ncRNAs and alternative splicing events (Chapter II) will help in further characterization studies of genes and overall transcriptional landscape of *P. infestans*. Re-analyzing some of the previous data in terms of ortholog identification in Chapter V might give us some new genes that were previously missed. Also, many genes have been separated in our analysis; for example, when PITG_10979 was split into 2 genes – the expression levels of PITG_10979b were higher than the parent gene whose expression was predicted low due to PITG_10979a that had very low coverage. This

shows that many earlier expression analysis may be deemed incorrect after changes to gene models.

The only drawback that we had with this study (Chapter II) was that the project was done with 'unstranded' RNA-Seq reads. This led to problems with exceptionally long UTRs and incorrect start codon predictions. We got some stranded RNA-Seq at the end of the project that helped us to correct UTRs and start codons for some genes. Stranded data from more developmental stages can get us better predictions on gene models. We can also use it to get better expression estimates of both protein and non-protein coding genes. Expression data based on stranded RNA-Seq reads can also help us answer the questions how the anti-sense RNAs are regulating the expression levels of protein coding genes that are present in vicinity.

Differences between *P. infestans* lineages have been a major problem in controlling the pathogen in fields; some of the lineages are resistant to commonly used fungicides while some are not, and some lineages are more aggressive pathogens. Some studies have tried to identify SNPs/INDELs in strains that showed higher virulence as compared to others (Cooke et al., 2012; Li et al., 2012); in Chapter III we have looked at large structural variations (SVs) in ten different lineages, which may help us understand the phenotypic variation in different strains. We have identified many large (≥ 1kb) deletions and inversions present in different lineages of *P. infestans*. Many of these large deletions and inversions have genes associated with them. Apart from large structural variants, we have also identified many medium (>30bp - <1kb) and small INDELs (≤30 bp). This is the first systematic study to identify large structural variation in different

lineages of *P. infestans*. The functional annotation of the genes exhibiting variants can help us to understand the phenotypic differences between strains in terms of traits that include mating type, resistance to fungicides or host preference.

One more experiment that will add value to this study (Chapter III) is SNP identification and annotation. SNP annotation can also help us identify defective alleles that easily accumulate in asexual populations (Muller, 1964; Gordo and Charlesworth, 2001; Paland and Lynch, 2006). Also, we were not able to characterize the long insertions – neither their length or if they have genes associated with them. Future studies or collaboration with someone who is good at *de novo* assembly of genomic reads can help find the length of long insertions in different lineages. Another useful addition to the present study (Chapter III) would be clustering lineages by different criteria such as mating type and see which SVs are present in one mating type but not the other. Also, we need to confirm the predicted SVs using some other sequencing method like Sanger sequencing to check if these SVs are real or a mere artifact of mapping or assembly problems.

P. infestans and *Py. ultimum* are both members of oomycetes but have different life styles; *P. infestans* is a hemibiotroph while *Py. ultimum* is a necrotroph. Apart from their feeding habits, *P. infestans* and *Py. ultimum* also differ in many gene families. Some are abundant in *P. infestans* but are missing or less abundant in *Py. ultimum*; examples include RxLRs, which are completely absent in *Py. ultimum*, glycoside hydrolases, and crinklers. *Py. ultimum* has a novel family of putative effector proteins which instead of a RxLR-dEER motif present in *Phytophthora* sp has a YxSL[RK] motif (Lévesque et al.,

2010). In this study which was done in collaboration with Dr. Ah-Fong, (Chapter IV) we have analyzed genes that are present in both *P*. *infestans* and *Py*. *ultimum* and also genes that are specific to each species. This is first study to compare the transcriptional profiles of both pathogens, which among other findings discovered that *P. infestans* has a more dynamic transcriptome with more genes that are turned on and off in different developmental stages compared to *Py. ultimum,* in which more genes are constitutively expressed. The presence of species-specific genes which included RxLRs and Crinkler proteins might give us targets to control these oomycetes in the fields. One of the drawback that we were not able to resolve in this study were the gene models in *Py*. *ultimum* that might have led to both false positives and negatives in terms of ortholog identification, expression calling or secreted proteins to name a few. Another drawback is the gene family size in *Py*. *ultimum*; even though only those families were considered in this study that have a 1:1 ratio of genes; it is possible that some *Py*. *ultimum* genes were not annotated and thus giving wrong number of genes in a gene family.

Evolution plays an important role in shaping the fitness and virulence of pathogenic organisms. While study of evolution may not give direct ideas to control these pathogens, it might give clues as to how these genomes are evolving that in the long run may help in devising strategies to control them. Sporulation in *P*. *infestans* plays an important role in the disease (Fry et al., 2008). Sporulation leads to the formation of flagellated zoospores that can swim for many hours, helping the pathogen find and infect a new host (Judelson and Blanco, 2005; Hardham, 2007); this motile stage is entirely dependent on the flagella. In this study (Chapter V), we have used comparative genomics

to identify orthologs of *P. infestans* flagellar proteins and compared it to another oomycete *Hyaloperonospora arabidopsis* which does not form zoospores. Using comparative genomics from species that have flagella, we have identified 257 proteins in *P. infestans*. Most of these genes were upregulated in sporulation and the swimming zoospore stage. Some of these proteins are structural components of axoneme and basal bodies while others are involved with energy production and utilization. We have identified 211 genes that have no orthologs in the azoosporic species *H. arabidopsis*. Synteny analysis between *P. infestans* and *H. arabidopsis* found remnants of one fifth of these genes in *H. arabidopsidis* showing that these deletions are recent.

In summary, in this study we have used bioinformatics approaches to help us progress towards better strategies to identify and control *P. infestans*. Development of different *in silico* prediction algorithms, sequencing algorithms and reductions in sequencing costs have speeded the process of acquiring and analyzing the data at a much faster pace than before. In this study, we have used DNA and RNA sequencing to answer questions that have given us direct methods to control *P. infestans* (Chapter I and III) or expand information on the nature of transcriptionally active regions, genomic differences between strains, and species-specific genes that can help us find novel targets to control *P. infestans* (Chapter II, IV and V).

References:

Amaral, P. P., Dinger, M. E., & Mattick, J. S. (2013). Non-coding RNAs in homeostasis, disease and stress responses: an evolutionary perspective. Briefings in Functional Genomics, 12, 254-278.

Avrova, A. O., Whisson, S. C., Pritchard, L., Venter, E., De Luca, S., Hein, I., & Birch, P. R. (2007). A novel non-protein-coding infection-specific gene family is clustered throughout the genome of *Phytophthora infestans*. Microbiology, 153, 747-759.

Cooke, D. E., Cano, L. M., Raffaele, S., Bain, R. A., Cooke, L. R., Etherington, G. J., ... & Grünwald, N. J. (2012). Genome analyses of an aggressive and invasive lineage of the Irish potato famine pathogen. PLoS Pathogen, 8, e1002940.

Dang, Y., Yang, Q., Xue, Z., & Liu, Y. (2011). RNA interference in fungi: pathways, functions, and applications. Eukaryotic Cell, 10, 1148-1155.

Freitag, J., Ast, J., & Bölker, M. (2012). Cryptic peroxisomal targeting via alternative splicing and stop codon read-through in fungi. Nature, 485, 522-525.

Gordo, I., & Charlesworth, B. (2001). The speed of Muller's ratchet with background selection, and the degeneration of Y chromosomes. Genetical Research, 78, 149-161.

Lévesque, C. A., Brouwer, H., Cano, L., Hamilton, J. P., Holt, C., Huitema, E., ... & Zerillo, M. M. (2010). Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire. Genome Biology, 11, R73.

Li, Y., van der Lee, T. A. J., Evenhuis, A., van den Bosch, G. B. M., van Bekkum, P. J., Förch, M. G., … & Kessel, G. J. T. (2012). Population dynamics of *Phytophthora infestans* in the Netherlands reveals expansion and spread of dominant clonal lineages and virulence in sexual offspring. G3: Genes|Genomes|Genetics, 2, 1529–1540.

Morrison, E. N., Donaldson, M. E., & Saville, B. J. (2012). Identification and analysis of genes expressed in the *Ustilago maydis* dikaryon: uncovering a novel class of pathogenesis genes. Canadian Journal of Plant Pathology, 34, 417-435.

Muller, H. J. (1964). The relation of recombination to mutational advance. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis, 1, 2-9.

Nunes, C. C., Gowda, M., Sailsbery, J., Xue, M., Chen, F., Brown, D. E., ... & Dean, R. A. (2011). Diverse and tissue-enriched small RNAs in the plant pathogenic fungus, *Magnaporthe oryzae*. BMC Genomics, 12, 288.

Paland, S., & Lynch, M. (2006). Transitions to asexuality result in excess amino acid substitutions. Science, 311, 990-992.