The secretome of the tomato pathogen *Cladosporium fulvum*

a comparative analysis of fungal pathogens

MSc thesis
Mattias de Hollander

The secretome of the tomato pathogen *Cladosporium fulvum*

a comparative analysis of fungal pathogens

THESIS

submitted in partial fulfillment of the
requirements for the degree of

MASTER OF SCIENCE

in

BIOINFORMATICS

by

Mattias de Hollander

born in Amstelveen, the Netherlands

Cover picture: Microscopic view of the spherical spores produced for dispersal by the fungus *Emericella nidulans* which are coated in a thin layer of the protein hydrophobin (Source: BASF - The Chemical Company 2008, `http://idw-online.de/pages/de/image68235`)

The secretome of the tomato pathogen *Cladosporium fulvum*

a comparative analysis of fungal pathogens

Student name:

Mattias de Hollander

Student registration number:

850121-354-030

Email:

mdehollander@gmail.com

Course code:

BIF-80333

Supervisor(s):

Erwin Datema, chairgroup Bioinformatics

Dr. Ir. Harrold van den Burg, chairgroup Phytopathology

Dr. Ioannis Stergiopoulos, chairgroup Phytopathology

Examinator(s):

Prof. Dr. Ir. Pierre de Wit, chairgroup Phytopathology

Dr. Roeland van Ham, chairgroup Bioinformatics

Contents

Abstract

Cladosporium fulvum (syn. *Passalora fulva*) is a biotrophic fungal pathogen that causes leaf mold on tomato. It belongs to the family of *Mycosphaerellaceae* within the Dothideomycete class of fungi, a major class of plant pathogenic fungi. During infection *C. fulvum* secretes a number of small proteins, which are collectively called effectors, into the apoplast of infected leafs. These effector proteins play an important role during pathogenesis by assisting the fungus in establishing disease. So far, ten effector proteins have been characterized from *C. fulvum*. In this study we aim to identify and functionally annotate the secretome of *C. fulvum*.

Currently, the genome sequences of at least seven *Dothideomycete* species are available, including those of *Mycosphaerella fijiensis,* a pathogen of banana, and *M. graminicola*, a pathogen of wheat that are both phylogenetically closely related to *C. fulvum*. Recently, the genome of *C. fulvum* has been sequenced. Here, using a combination of computational methods the secretomes for these seven *Dothideomycete* species were determined. Protein-coding genes in *C. fulvum* were predicted using GeneMark-ES. The proteomes of *M. fijiensis* and *M. graminicola* were obtained from the Joint Genome Institute (JGI) genome portal. In order to identify protein families an all-against-all similarity search was performed for all encoded proteins. Groups of functionally related proteins were identified with the MCL algorithm. We then defined a secreted protein as a protein containing a predicted signal peptide, without putative transmembrane domains outside the secretion signal or GPI-anchored regions. All secreted proteins were functionally annotated by integrating BLAST similarities and predicted InterPro domains using Blast2GO.

About half of the 1,275 predicted secreted proteins in *C. fulvum* have orthologues in *M. graminicola* and *Mycosphaerella fijiensis*, 289 have an ortholog in one of the two species, whereas 216 appear to be specific to *C. fulvum,* including chitin-binding proteins and hydrophobins. Furthermore, the secretome of *C. fulvum* contains several groups of proteins which occur more frequently compared to other Dothideomycetes. Four fungal hydrophobins are unique in *C. fulvum* and other classes of proteins like Cytochrome P450s, chloroperoxidases, and glycoside hydrolases are also more prevalent in the *C. fulvum* secretome.

Acknowledgements

This thesis has been a collaboration between the Bioinformatics chairgroup and the Phytopathology chairgoup of the Wageningen University. I would like to thank Prof. Dr. Jack Leunissen and Dr. Roeland van Ham of the bioinformatics department and Prof. Dr. Ir. Pierre de Wit of the phytopathology department for bringing the two research fields together. For me it was really a honor to form the connection between the two departments.

I have always said that I would like to work together with wetlab-biologists as a bioinformatician, because computational models and techniques are most of the time only able to give predictions of what happens in cells. Being close to the wetlab means that your results in the future can be validated with *in vivo* or *in vitro* experiments. This again can generate new work for bioinformaticians, making it a continuous collaboration.

My supervisors of the phytopathology group, Dr. Ir. Harrold van den Burg and Dr. Ioannis Stergiopoulos have been a great help in interpreting the results. Especially Harrold gave me useful explanations for the observations I had made and explained how we could link the different functions we predicted into a more broader biological perspective. My other supervisor from the bioinformatics group, Erwin Datama, assisted me more with the setup of the project and form bioinformatics related question he always had time for me.

Having multiple supervisors with different backgrounds can give you a lot of input and motivation. Sometimes it was difficult to get everyone pleased and on the same line, but in the end I think I am pleased with the result.

I also want to mention the people from the phytopathology group for their input they gave during the work discussions and the people of the Greenomics room for more computer related topics. In particular Jan van Haarst who teached me some useful awk and sed commands.

In a personal context, Emily Roeder gave me a lot of motivation to pursue my research during the nice breaks at the small lake behind the Lumen building and the time after office hours. Unfortunately, she went back to the United States at the end of the year to continue her own journey of life, but I want to thank her for some really wonderful months.

Mattias de Hollander

Wageningen, the Netherlands

February 22th, 2010

Chapter 1

Introduction

1.1 Pathogen host interaction

Cladosporium fulvum is a biotrophic fungal pathogen that causes leaf mold on tomato (Figure 1.1). This means that the fungus grows and obtains its nutrients from living plant material rather than first killing the host cells. *C. fulvum* belongs to the family of *Mycosphaerellaceae* within the Dothideomycete class of fungi. Dothideomycete is one of the largest and most important classes of fungi that collectively infect economically important plants like maize, barley, wheat, banana and tomato (Galagan *et al.*, 2007). Currently, the genome sequences of at least six Dothideomycete species are available, including those of *Mycosphaerella fijiensis,* a pathogen of banana, and *M. graminicola*, a pathogen of wheat. Both the species are phylogenetically closely related to *C. fulvum.*

Figure 1.1: Symptoms of infection by *C. fulvum* on a tomato leaf

Defense responses of plants against pathogens consist of two steps: First, plants can activate an immune response by recognition of pathogen-associated molecular patterns (PAMPs) by PAMP-recognition receptors (PRRs). This activates a PAMP-triggered immunity (PTI) response that prevents further colonisation of the host (Birch *et al.*, 2008). Pathogens overcome this first line of defense by secreting effectors that can suppress PTI. In response plants, have developed the tools to detect these effectors by Resistance (R) proteins and this activates the second step in defense responses, meaning effector-triggered-immunity (ETI). Most typically, ETI culminates in a Hyper-

sensitive Response (HR) that results in growth arrest of the fungus and this response is more rapid than PTI (Birch *et al.*, 2008). When no corresponding R proteins are present in the plant, the pathogen can further infect the plant and cause disease (Stergiopoulos and de Wit, 2009).

During the colonization of tomato plants, *C. fulvum* remains inside the apoplast and does neither penetrate host cells nor develop haustoria and secretes a number of effectors into the apoplast of tomato leafs. These effector proteins play a vital role during pathogenesis as they assist the fungus in establishing disease (Stergiopoulos and de Wit, 2009). Until know, four effector or otherwise known as avirulence (*Avr*) genes, namely *Avr2, Avr4, Avr4E* and *Avr9,* have been identified from *C. fulvum*, which are recognized in tomato by the cognate Cf resistance proteins Cf-2, Cf-4, Cf-4E and Cf-9, respectively. In Figure 1.2 the recognition of Avr2 by the plant resistance protein Cf-2 can be seen. In addition, six extracellular proteins (Ecps) have been characterized, namely Ecp1, Ecp2, Ecp4, Ecp5, Ecp6 and Ecp7, and for at least four of them (Ecp1, Ecp2, Ecp4, and Ecp5), resistance traits have been characterized in wild *Solanum* species. All secreted Avr and Ecp proteins are relatively small and contain an even number of cysteine residues, which are involved in the formation of disulphide bonds that are necessary for protein stability in the harsh protease-rich environment of the tomato apoplast (Joosten *et al.*, 1997).

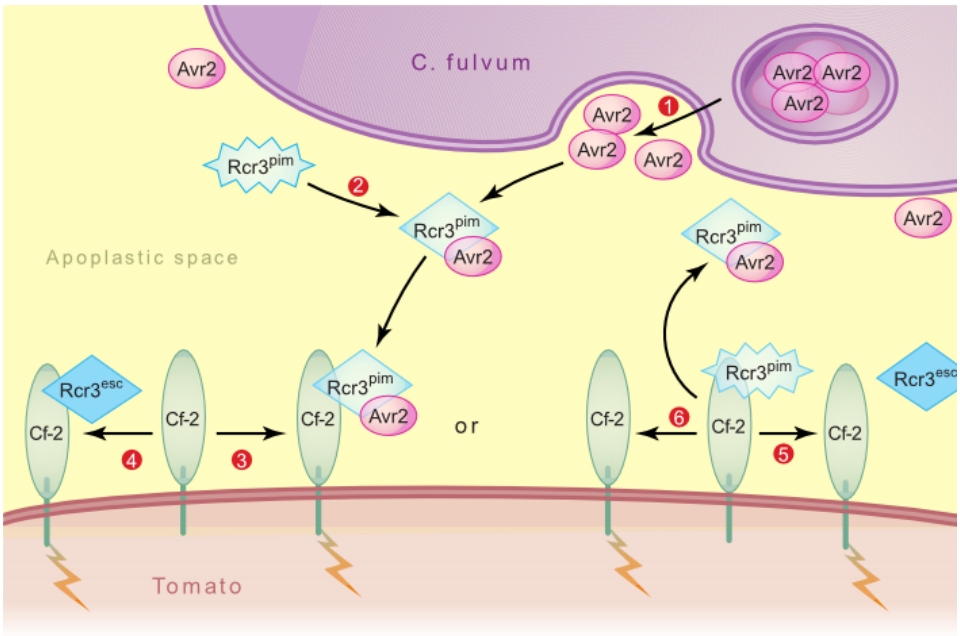

Figure 1.2: Recognition of pathogen effector proteins (Avr2) by plant resistance proteins (Cf-2). (Schulze-Lefert and Bieri, 2005)

Plant pathogenic bacteria can translocate effector proteins into the cytoplasm of host cells by

using the Type III Secretion System (TTSS). Many of these injected bacterial effectors were shown to suppress PTI (Birch *et al.*, 2008). Within oomycetes a similar mechanism has been detected, in which effector proteins containing a RXLR-dEER motif can enter host cells (Birch *et al.*, 2008). For some fungal effectors it has been shown that they are putatively translocated into the host cell, where they interact with cytoplasmic or nuclear R proteins (Catanzariti *et al.*, 2007; Mueller *et al.*, 2008). However, these effector proteins do not have a clear conserved motif that can be implicated in translocation of the proteins inside the host cell.

Previously, secreted proteins have been identified from cDNA libraries and complete genome sequences. Analysis of EST sequences from the apple scab pathogen *Venturia inaequalis* identified new candidate effector genes, which are coding for small, cysteine-rich proteins containing a putative signal peptide (Bowen *et al.*, 2009). The repertoire of putatively secreted proteins has been determined for the wheat pathogen *Stagonospora nodorum*, which also belongs to the Dothideomycete class of fungi (Galagan *et al.*, 2007). One third of the 1782 proteins that were predicted to be extracellular could not be assigned with an GO annotation, leaving the roles of the remaining proteins undetermined. The secretome of *Ustilago maydis*, a pathogen of maize, revealed that many effector genes are clustered in the genome resulting in effector rich sequence islands (Kämper *et al.*, 2006; Mueller *et al.*, 2008). Comparative analysis of two *Phytopthora* genomes has shown that some gene families that encode secreted proteins have been expanded (Jiang *et al.*, 2006).

Currently, the genome of *C. fulvum* is being sequenced. Here we studied the genome annotation of secreted proteins and tried to identify novel effector proteins. In addition, the genome sequence of *C. fulvum* was used to perform a comparative analysis between *C. fulvum* and other sequenced Dothideomycetes like *M. graminicola* and *M. fijiensis*. The number of secreted proteins and gene families in *C. fulvum* was compared with the secretomes of *M. fijiensis*, *M. graminicola* and other Dothideomycetes.

Most studies on secreted proteins predict signal peptides using SignalP (Bendtsen *et al.*, 2004), together with programs to predict transmembrane domains, subcellular localisation and glycosylphosphatidylinisotol (GPI) lipid modification (anchor) sites (Mueller *et al.*, 2008). A recent evaluation of signal peptide prediction programs showed that SignalP was the most accurate program in signal peptide and cleavage site prediction. The prediction performance could further be improved by combining multiple methods into consensus prediction (Zhang *et al.*, 2009). In this study, the selection of secreted proteins was performed in an analogous way to studies on other biotrophic pathogens, like the maize pathogen *Ustilago maydis* (Mueller *et al.*, 2008).

1.2 Functions of secreted proteins

In order to survive within the plant *C. fulvum* needs to degrade parts of the plant cell wall. Therefore it is expected that the secretome of *C. fulvum* contains a number of proteins that degrade plant tissue. Also proteins without an enzymatic function, like the already isolated and characterized cysteine-rich hydrophobins (Segers *et al.*, 1999), should occur in the secretome.

1.2.1 Enzymes involved in plant cell wall degradation

Leaves and other aerial parts of plants are covered by the cuticle made of cutin (Figure 1.3). The plant cuticle is attached to the cell wall by a layer of pectin. The cuticle protects plants against fungal pathogens by forming a physical barrier. All pathogens need an entry-point, like a wound, or they have to (partially) break down the cuticle before they can cause infection (Chassot *et al.*, 2007; Kolattukudy, 1985). Cutinases are enzymes that can catalyse the hydrolysis of cutin, and allow the fungus to penetrate through the cuticle. However, *C. fulvum* can enter the host through the stomata if it settles on the abaxial (lower) side of a leaf and germinates (Thomma *et al.*, 2005). Therefore it does not need to mechanically or enzymatically break down the cuticle in order to cause infection.

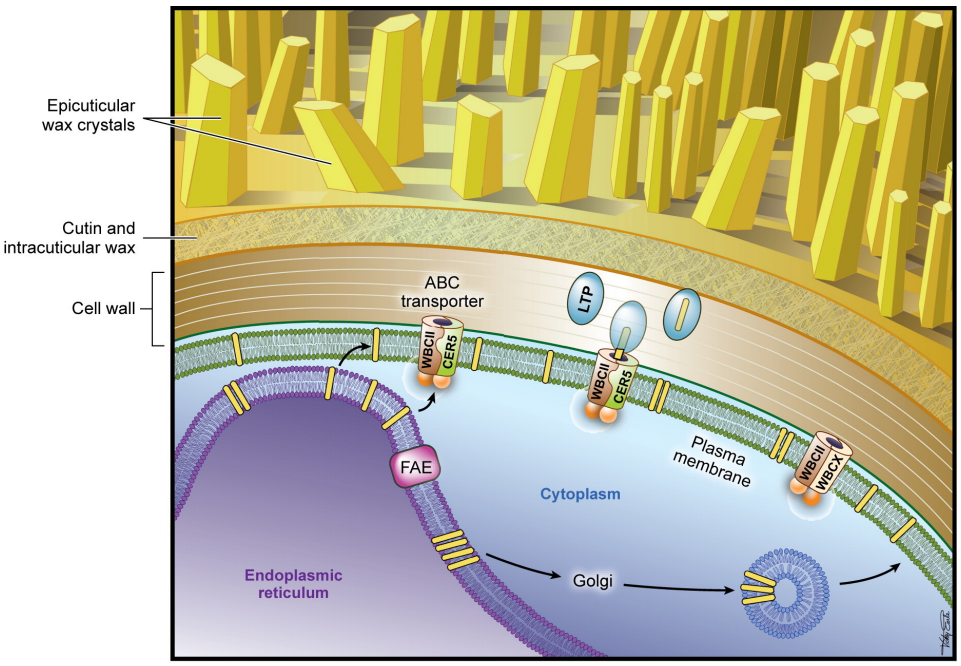

Figure 1.3: Schematic impression of the plant cell wall and the cuticle (Samuels *et al.*, 2008).

Plants form two different kind of cell walls that have a different composition and function. The

primary cell wall surrounds growing and dividing cells and is part of the apoplast (Figure 1.4a). They are mainly composed of carbohydrates like celluloses, hemicelluloses and pectins (Cosgrove, 2005). The secondary cell wall is much thicker and is only present in growing cells (Figure 1.4b). It consists of cellulose, lignin, and xylan (Persson *et al.*, 2007). Lignin provides strength and protection against biodegradation to the plant cell wall, by forming a matrix surrounding the cellulose and hemicelluloses (Schoemaker and Piontek, 1996). Lignin peroxidases are a family of extracellular heme proteins and play a main role in lignin biodegradation (Reddy and D'Souza, 1994).
Therefore, it is expected that the genome of *C. fulvum* encodes a number of plant cell wall degrading enzymes such as pectin lyases, endoglucanases and glucosidases similar to the genomes of other plant pathogens like *Ustilago maydis* (Mueller *et al.*, 2008).

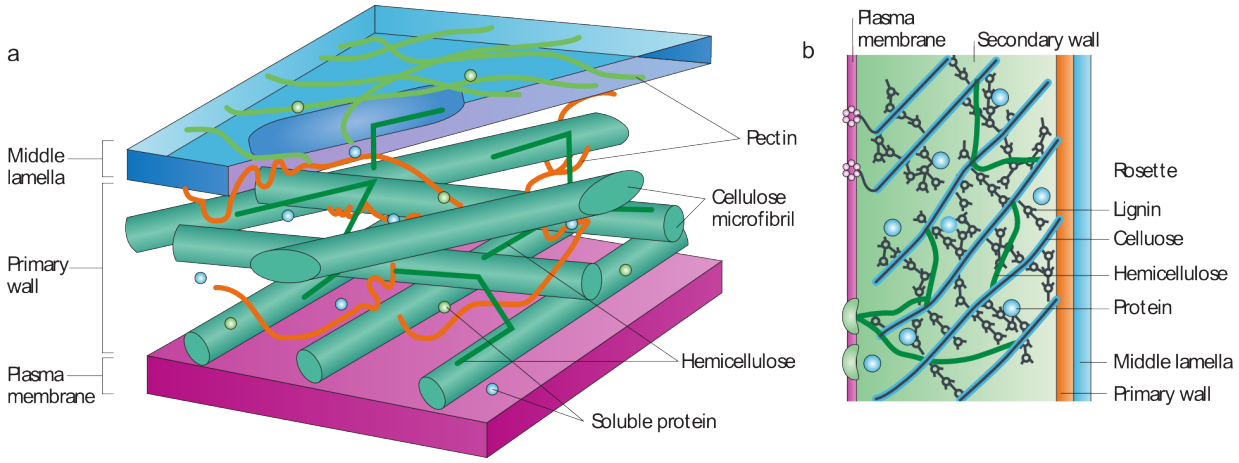

Figure 1.4: Molecular structure of the primary (a) and secondary (b) cell wall in plants.

1.2.2 Characterized proteins

Before the release of the genome sequence, there are already 41 entries in the UniProt database for *C. fulvum*. Five of them have been manually annotated in the UniProtKB/Swiss-Prot database, including two Avrs and two beta-glucosidases. The remaining UniProtKB/TrEMBL entries comprehend five Ecps and eight hydrophobins.

Hydrophobins are small proteins that contain eight cysteine residues with a conserved spacing (Kershaw and Talbot, 1998). Hydrophobins are found in many filamentous fungi on the outer surface of the spores (conidia) where they are involved in the communication between the fungus and its environment (Staples, 2001). Two distinct classes of hydrophobins have been characterized based on their solubility and hydropathic properties, but they nevertheless have a similar size and cysteine

spacing is conserved (Nielsen *et al.*, 2001; Whiteford and Spanu, 2001). In *C. fulvum* six hydrophobin genes (HCf-1 to -6) have been identified and characterized so far (Whiteford and Spanu, 2001). Four of these hydrophobins (HCf-1 to HCf-4) are related and belong to class I hydrophobins while HCf-5 and HCf-6 are different and they belong to class II hydrophobins. Experiments showed that HCf-1 is not required for pathogenicity on the tomato, but could play a role in the dispersal mediated by water droplets. The six hydrophobins in *C. fulvum* are expressed under different nutritional conditions and in a different stage of development. This suggest at most that each of these hydrophobins have a seperate function (Whiteford and Spanu, 2001).

1.3 Aim and Objectives

The aim of this study is to identify and functionally annotate the secretome of *C. fulvum*. A first objective is to create a computational pipeline that predicts secreted proteins using the encoded proteins of the *C. fulvum* genome.

Next, the secretome will be functionaly annotated using sequence similarity methods that predict Interpro domain and Gene Ontology terms. Using the annotated secretome it is possible to focus on the second objective, describing how *C. fulvum* establishes an infection, survives and reproduces within the tomato plant.

Futhermore, the characteristics of known effector proteins can help in selecting a list of small secreted proteins, which are likely to be candidate new effectors from *C. fulvum*.

Additionaly, the secretome of *C. fulvum* will be compared with the secretomes of *M. fijiensis, M. graminicola* and other species of Dothideomycetes to find homologous and species-specific secreted proteins. Within this light, an attempt will be made to find out what makes *C. fulvum* different in terms of secreted proteins compared to the other fungal pathogens. Are there any classes of proteins, which are only present in the genome of *C. fulvum*? Or have certain groups of proteins expanded in *C. fulvum* after speciation from other Dothideomycetes? These questions form the final objective of this study: analyze the difference in functions of the proteins encoded by the genomes of seven Dothideomycetes.

Chapter 2

Material and Methods

2.1 Genome data

The genome of *C. fulvum* has been sequenced and assembled by Plant Research International (PRI) and annotated using an automated pipeline (Fiers *et al.*, 2008). Protein-coding genes in *C. fulvum* were predicted using GeneMark-ES (Ter-Hovhannisyan *et al.*, 2008).

Protein sequences for other Dothideomycetes were obtained from the Joint Genome Institute (JGI) portal or the Broad Institute portal. For a complete overview of all organisms and the number of encoded proteins see table 2.1.

2.2 Secretome pipeline

Several studies have shown that secreted proteins play an important role in establishing plant disease (Kämper *et al.*, 2006; Talbot *et al.*, 2008). There exist a number of tools to predict whether a protein is likely to be secreted (reviewed in (Klee and Ellis, 2005; Zhang *et al.*, 2009). In this study several filtering steps are applied on all protein sequences of a genome to select a final set of putative secreted proteins. The applied strategy is similar to the method used by Mueller *et. al.* on the *Ustilago maydis* genome (Mueller *et al.*, 2008). For the prediction of the signal peptide SignalP 3.0 (Bendtsen *et al.*, 2004) is used, because is has been shown that this program is the most accurate in signal peptide prediction in bacteria (96%) (Zhang *et al.*, 2009) and eukaryotes (90%) (Klee and Ellis, 2005)

First, the N-terminal sequence of each protein is analysed for the presence of a signal peptide using SignalP 3.0. This tool incorporates a method that combines several artificial neural networks and hidden Markov models to give a signal peptide prediction. Only proteins which have at least 2

Table 2.1: Predicted number of proteins in seven *Dothideomycetes*

Organisms	**Species abbreviation**	**Number of predicted proteins**	**Website**
Cladosporium fulvum	Cf	16,672	In-house sequencing project
Mycosphaerella fijiensis	Mf	10,313	http://genomeportal.jgi-psf.org/Mycfi1
Mycosphaerella graminicola	Mg	10,933	http://genomeportal.jgi-psf.org/Mycgr3
Stagonospora nodorum	Sn	16,597	http://www.broadinstitute.org/annotation/genome/stagonospora_nodorum/
Pyrenophora tritici-repentis	Pt	12,169	http://www.broadinstitute.org/annotation/genome/pyrenophora_tritici_repentis.3/MultiDownloads.html
Cochliobolus heterostrophus C5	Ch	9,633	http://genomeportal.jgi-psf.org/CocheC5_1/CocheC5_1.download.ftp.html
Alternaria brassicicola	Ab	10,688	http://genomeportal.jgi-psf.org/Altbr1/Altbr1.download.ftp.html

positive predictions for either D-score, S-mean score predicted by the neural network of SignalP or HMM prediction were kept for further analysis. Next, proteins that contain a putative mitochondrial targeting signal as predicted by TargetP (Emanuelsson *et al.*, 2000) were removed. The remaining proteins were screened for the presence of transmembrane spanning regions (TMs) using the TMHMM program (Krogh *et al.*, 2001). Only proteins containing no TM or proteins containing a single TM that overlaps with the secretion signal were kept in the dataset. Finally, proteins containing a potential GPI-anchor signal as predicted by the PredGPI web-service were discarded (Pierleoni *et al.*, 2008).

2.3 Functional annotation

Protein sequences were functionally annotated using a standardized approach. For each protein multiple associated Gene Ontology categories are predicted using the Blast2GO tool (Conesa and Götz, 2008). Blast2GO was run with default parameter settings and given BLAST results against the non-redundant database of NCBI (January 2009) as input. Additionaly, each protein sequence was scanned for the presence of protein domains using InterproScan v4.4 including domain profiles from the Pfam, SMART, superfamily and Gene3D database (February 2009).

The number of Interpro domain detected in the proteins of all analyzed Dothideomycetes were compared to each other using a strategy applied by Martens *et. al* (Martens *et al.*, 2008). For each Interpro term the number of proteins predicted to have this domain per species were counted. A z-score is calculated by substracting the average number of proteins per domain from the total

number of proteins per species and dividing it by the standard deviation. Note that this score uses the absolute differences in protein numbers and does not take into account the size of the secretome. A high z-score for a Interpro domain denotes that for a specific species more proteins are predicted with that domain compared to the average number of proteins with the same domain for all analyzed species.

2.4 Protein clustering

The grouping of proteins of different species in protein families is a good guide in functional genomics and evolutionary analysis (Enright *et al.*, 2002). A protein family can be defined as a group of evolutionary related proteins that share significant sequence similarity (Dayhoff, 1976). Protein families can aid the functional characterization of uncharacterized members using the functions of their family members (Liu and Teow, 2005). Furthermore, the expansion or contraction of protein families in related species can be used to foresee the importance that they have acquired in the species after speciation events.

Most methods that are currently being used for the clustering of protein sequences rely on protein sequence similarity. Often these tools incorrectly group proteins in a single family due to a number of problems (Enright *et al.*, 2002; Liu and Teow, 2005). The main problem is that proteins can contain multiple domains. The presence of a shared domain does not mean that these proteins are also involved in the same biological process (Chen *et al.*, 2007). Therefore, only if they have a highly similar domain architecture they should be grouped together. Moreover, some protein domains play a role in a whole range of cellular functions. Clustering of proteins with such promisicuous domains will not reflect a group of functionaly related proteins. Finally, automated gene calling can result in fragmented proteins. This can lead to incorrect assignment of a protein to a family.

All protein clustering methods require sequence similarity relationships as input, such as the ones produced by BLAST, but they differ in the way that they handle the data. Some tools require additional information like the presence of protein domains, such as the the GeneRAGE algorithm (Enright and Ouzounis, 2000). These have been proven to be successful in creating databases of protein families such as PRODOM (Corpet *et al.*, 2000), but they are also hampered by the problems addressed above (Liu and Teow, 2005) and thus can not group the proteins correctly.

The Markov Cluster Algorithm (MCL) is able to overcome most of the above mentioned problems (Enright *et al.*, 2002), and will therefore be used in this study. It makes use of a graph, were nodes represent proteins and edges represent sequence similarity. The MCL algorithm applies random

walks through the sequence similarity graph and iteratively eliminates the inter-families similarities caused by multi-domain proteins (Liu and Teow, 2005). The tightness or granularity of the clusters is controlled by a inflation factor.

In order to identify protein families in *C. fulvum* and six other Dothideomycetes we performed an all-against-all similarity search (BLASTP, E-value cut-off e-3). First, the protein sequences from the different genomes are grouped into protein families using the Markov Clustering (MCL) algorithm (Enright *et al.*, 2002) with an inflation factor of 2.5 after a performance benchmark test with various threshold settings. The workflow for the clustering of similarity graphs was used as described on the MCL website (http://www.micans.org/mcl/man/clmprotocols.html, version 1.008, 09-308). A description was added to families based on the most frequently occuring InterPro motifs in all members of a family.

2.5 Orthology detection

Proteins from all genomes are clustered into orthologous and paralogous groups using OrthoMCL v2.0 (Li *et al.*, 2003) with default settings using an all-against-all BLASTP dataset as input.

Chapter 3

Results

3.1 Identification of secreted proteins

Cladosporium fulvum infections are assisted by the use of proteins that are secreted by the fungus in the apoplast of the leaves. To gain more insight in the role of these proteins in pathogenesis the secreted proteins are identified in the recently sequenced genome using computational methods. All genes encoded in the genome of *C. fulvum* are screened for the presence of a predicted signal peptide, putative transmembrane domains outside the secretion signal or GPI-anchored regions in their translated protein sequence (Figure 3.1).

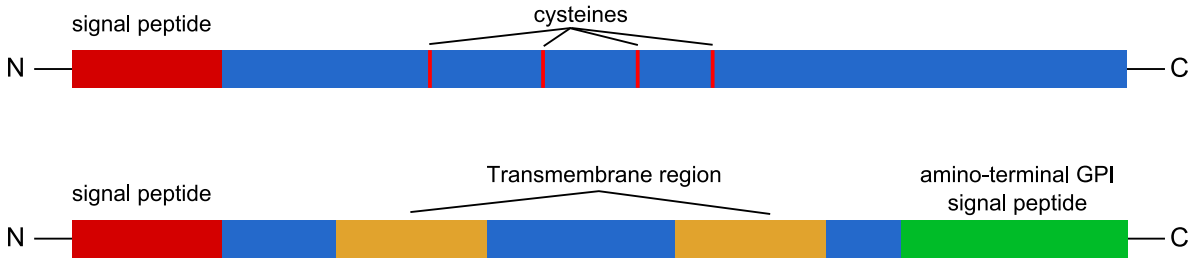

Figure 3.1: Effector proteins contain a signal peptide with positive number of conserved cysteines (top). A secreted protein is defined as a protein containing a predicted signal peptide, without putative transmembrane domains outside the secretion signal or GPI-anchored regions (bottom).

In total 1,275 out of the 16,672 predicted proteins in *C. fulvum* are identified as being putatively secreted proteins (Figure 3.2). These were selected from a large pool of 1,877 proteins that contained a predicted signal peptide at their N-terminal sequence. However, during following-up filtering steps 116 mitochondrial targeted proteins, 410 proteins containing transmembrane regions and 113 proteins having a potential GPI-anchor peptide were discarded. Based on this analysis, the set of predicted secreted proteins corresponds to 7.6% of all proteins encoded in the genome of *C. fulvum*. For

further functional analysis of the secretome of *C. fulvum* the secreted proteins are also predicted for six other Dothideomycete species, including *Mycosphaerella graminicola, Mycosphaerella fijiensis, Stagonospora nodorum, Pyrenophora tritici-repentis, Cochliobolus heterostrophus C5* and *Alternaria brassicicola.* The secretomes of these species contain on average 8.9% of all encoded proteins (Table 3.1). In absolute numbers most secreted proteins are predicted for *S. nodorum* with a 1,298 putatively secreted proteins. With a secretome size of 10.7 % of all predicted proteins, the secretome of *P. tritici-repentis* is relative terms the largest one. The proportion of secreted proteins in the *C. fulvum* genome is with 7.6% of all encoded proteins the smallest.

Furthermore, the set of all putatively secreted proteins is screened for putative effectors that could be play a role in pathogenesis. The main characteristics of known effector proteins is that they are relatively small and contain an even number of cysteines. Using these two criteria we selected a set of 135 proteins from the secretome of *C. fulvum* that are smaller than 255 amino acids,and contain at least 5 cysteine residues are selected. The full list of small secreted proteins (SSPs) can be found in Appendix table A.1.

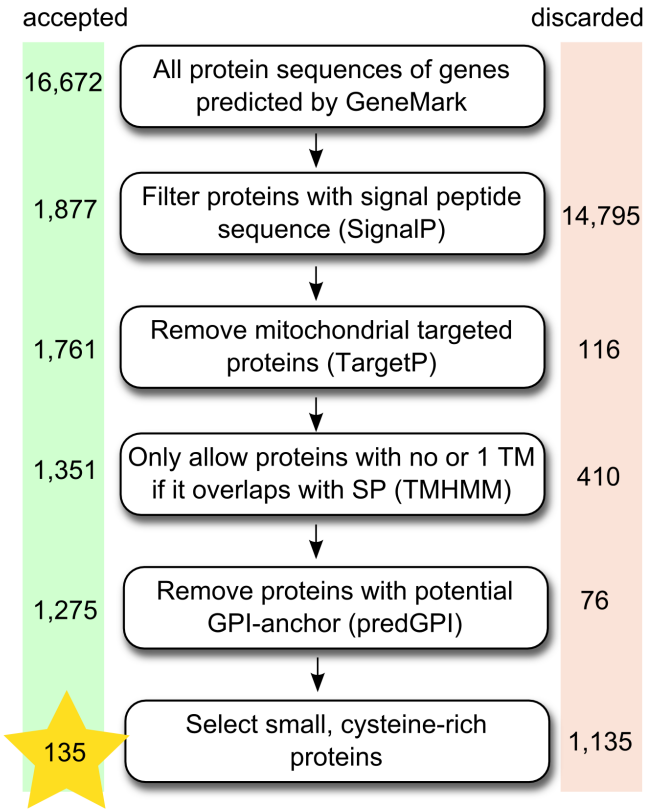

Figure 3.2: Strategy and filtering steps used to select putatively secreted proteins. For *Cladosporium fulvum* 1,275 proteins revealed to be putatively secreted (7.6% of the total proteome).

Table 3.1: Number of secreted proteins for three currently sequenced *Dothideomycete* species. On average 8% of all encoded proteins are predicted to be secreted.

Species	**Number of secreted proteins**	**Percentage of predicted proteins**
Cladosporium fulvum	1,275	7.6
Mycosphaerella graminicola	1,001	9.2
Mycosphaerella fijiensis	809	7.8
Stagonospora nodorum	1,713	10.3
Pyrenophora tritici-repentis	1,298	10.7
Cochliobolus heterostrophus C5	849	8.8
Alternaria brassicicola	871	8.1

3.2 Annotation of secreted proteins

In order to gain a better understanding on the roles that secreted proteins might play during pathogenesis, they are computationally annotated using Interpro motifs and Gene Ontology (GO) terms.

3.2.1 Interpro motifs and Gene Ontology

From the total number of 1,275 putatively secreted proteins in *C. fulvum*, 559 proteins were found containing an Interpro motif and/or having a GO term assigned to them. Some proteins are annotated based only on Interpro domain (145 proteins), some based on GO assignment (27 proteins).
As can be seen in figure 3.3, the majority of all putatively secreted proteins are not assigned with neither an Interpro domain (44.8%) or a GO term (54.0%). However, this is not a secreted protein-related feature, as 51.3% and 59.7% of all encoded proteins in the *C. fulvum* genome have no Interpro or GO term assigned to them, respectively. Even less annotations can be given for the small secreted proteins (SSPs) in *C. fulvum*. Out of all 135 putative SSPs, 18.5% (25) are assigned with an Interpro domain, including eight putative hydrophobins, three putative cutinases, and a putative chitin-binding domain (Table 3.2)

3.2.2 Clustering

About half of all putatively secreted proteins in *C. fulvum* can be assigned with an Interpo motif or GO term. In order to be able to describe the functions of secreted proteins in more detail, all encoded proteins in the *C. fulvum* genome are clustered based on sequence similarity using the Markov Clustering (MCL) algorithm into families with proteins from the six Dothideomycetes. All members of a cluster are grouped together because they share significant sequence similarity and is

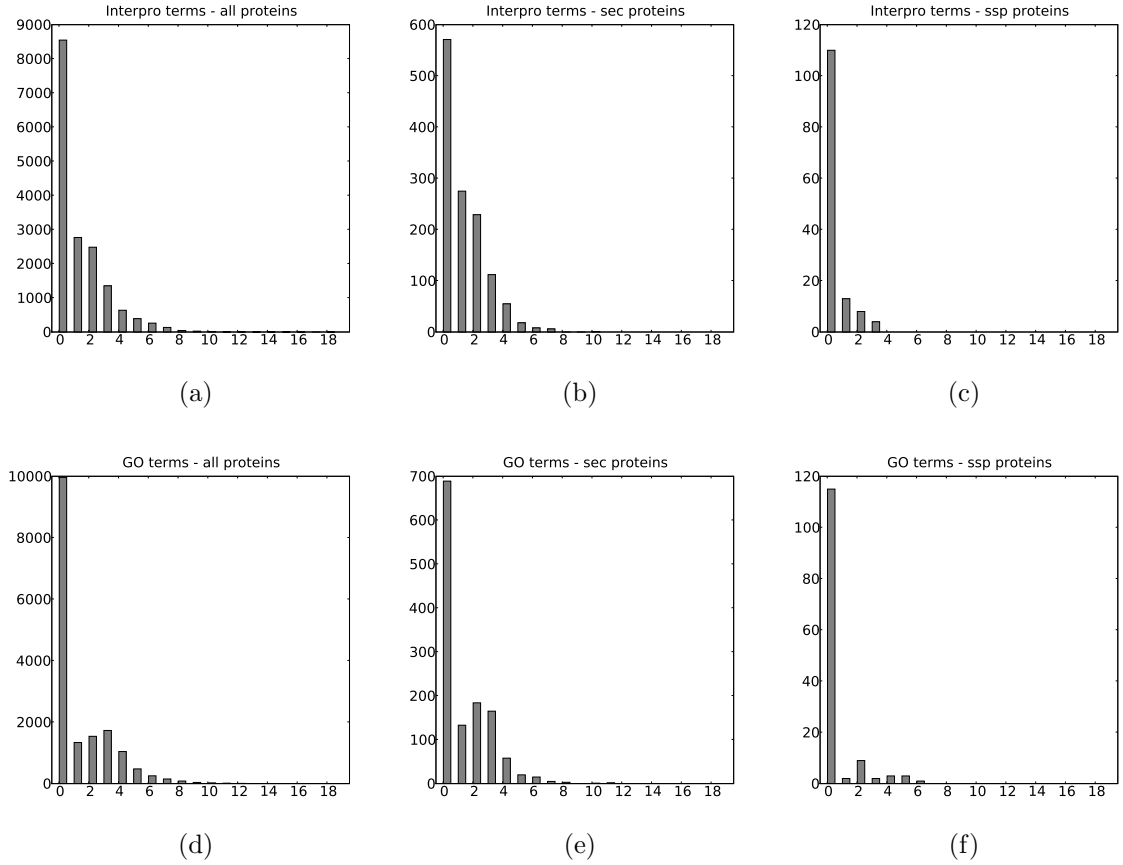

Figure 3.3: Distribution of the number of Interpro terms (a-c) and Gene Ontology terms (d-f) per protein. On the x-axes the total number of annotation terms per protein is plotted against the total number of proteins containing that amount of annotation terms (y-axes) for all proteins in *C. fulvum* (a,d), putatively secreted proteins (b,e) and small secreted proteins (c,f)

Table 3.2: Predicted small secreted proteins in *C. fulvum*: 135 proteins are smaller than 255 amino acids and contain at least 5 cysteine residues. Shown are the 25 proteins that are predicted to contain a Interpro domain.

Protein ID	Size	Number of cysteines	Interpro domain	Interpro description
Cf|2592304	228	8	IPR000675; IPR001412;	Cutinase; Aminoacyl-tRNA synthetase, class I, conserved site;
Cf|2599123	223	7	IPR000675; IPR011150;	Cutinase; Cutinase, monofunctional;
Cf|2592595	223	6	IPR000675;	Cutinase
Cf|2594577	222	6	IPR000873;	AMP-dependent synthetase and ligase
Cf|2601278	244	8	IPR001223; IPR013781; IPR017853	Glycoside hydrolase, family 18, catalytic domain; Glycoside hydrolase, subgroup, catalytic core; Glycoside hydrolase, catalytic core
Cf|2595228	121	8	IPR001338	Fungal hydrophobin
Cf|2597786	228	8	IPR001338	Fungal hydrophobin
Cf|2588266	103	8	IPR001338	Fungal hydrophobin
Cf|2591325	110	9	IPR001338	Fungal hydrophobin
Cf|2592466	105	8	IPR001338	Fungal hydrophobin
Cf|2594135	225	6	IPR001602	Protein of unknown function UPF0047
Cf|2592097	171	10	IPR001969	Peptidase aspartic, active site
Cf|2601254	228	9	IPR002482	Peptidoglycan-binding LysM
Cf|2589602	223	6	IPR002482	Peptidoglycan-binding LysM
Cf|2597781	135	8	IPR002557	Chitin binding protein, peritrophin-A
Cf|2596116	229	6	IPR002818	ThiJ/PfpI;
Cf|2591917	158	8	IPR002889; IPR013994	Carbohydrate-binding WSC; Carbohydrate-binding WSC, subgroup
Cf|2590916	220	7	IPR005132; IPR009009; IPR014734	Rare lipoprotein A; Barwin-related endoglucanase; Pollen allergen, N-terminal
Cf|2601014	245	6	IPR006094; IPR016166	FAD linked oxidase, N-terminal; FAD-binding, type 2
Cf|2600764	220	6	IPR006863	Erv1/Alr
Cf|2589608	87	8	IPR008427; IPR014005	Extracellular membrane protein, 8-cysteine region, CFEM; Extracellular membrane protein, 8-cysteine region, fungi
Cf|2602579	165	8	IPR010636	Hydrophobin 2
Cf|2588532	112	8	IPR010636	Hydrophobin 2
Cf|2586674	184	8	IPR010636	Hydrophobin 2
Cf|2597260	121	6	IPR013032	EGF-like region, conserved site

therefore called a protein (or gene) family.

In total 18,865 protein families are found in the six analysed Dothideomycete species, with 2,163 families containing more than 40% secreted proteins. In figure 3.4 the size of the protein families is plotted against the number of secreted (grey) and non-secreted (black) protein families is plotted. Most proteins of the putatively secreted proteins (54.7%) are not clustered with any other protein and thus form so-called singletons. These families have only one member and form the first bars on the left in the figure. Small families occur more frequently compared to larger families.

There is a small peak for proteins with seven members. In figure 3.5, the distribution of the number of species in cluster with 7 members can be seen. This figure shows that for each family approximately one protein per species is present. Presumably, the functions carried out by these proteins are well conserved across all Dothideomycetes and the proteins have almost no paralogs.

More than 50% of the protein families that consist of more than 40% of secreted proteins, are composed of only one protein (1,184 out of 2,163). A similar observation can be made for non-secreted proteins. Out of the set of families with secreted proteins, 548 can be assigned with an Interpro term.

3.2.3 Orthology

Besides sorting all proteins from the 7 species of Dothideomycetes in groups of homologues clustering based on orthology can also provide a wealth of information. Therefore, all proteins encoded in the genomes of *C. fulvum*, *M. graminicola* and *M. Fijiensis,* are clustered in orthologous pairs using OrthoMCL. About half of the 1,275 predicted secreted proteins in *C. fulvum* have orthologues in the two other species, 289 have an ortholog in one of the two species, whereas 216 appear to be specific to *C. fulvum*. Within the list of putatively secreted proteins with no orthologs in the other *Mycosphaerellaceae* are proteins that are annotated with functions like chitin-binding and hydrophobins.

3.3 Functional analysis of secreted proteins

The number of proteins per Interpro term are compared among the seven Dothideomycetes in order to find terms that are present in all species or abundant or unique in *C. fulvum*.
For each Interpro term the total number of proteins containing this term in the secretome is counted per species. To compare the different profiles, the protein counts are converted into z-scores by substracting the average number of proteins per domain from the total number of proteins with that

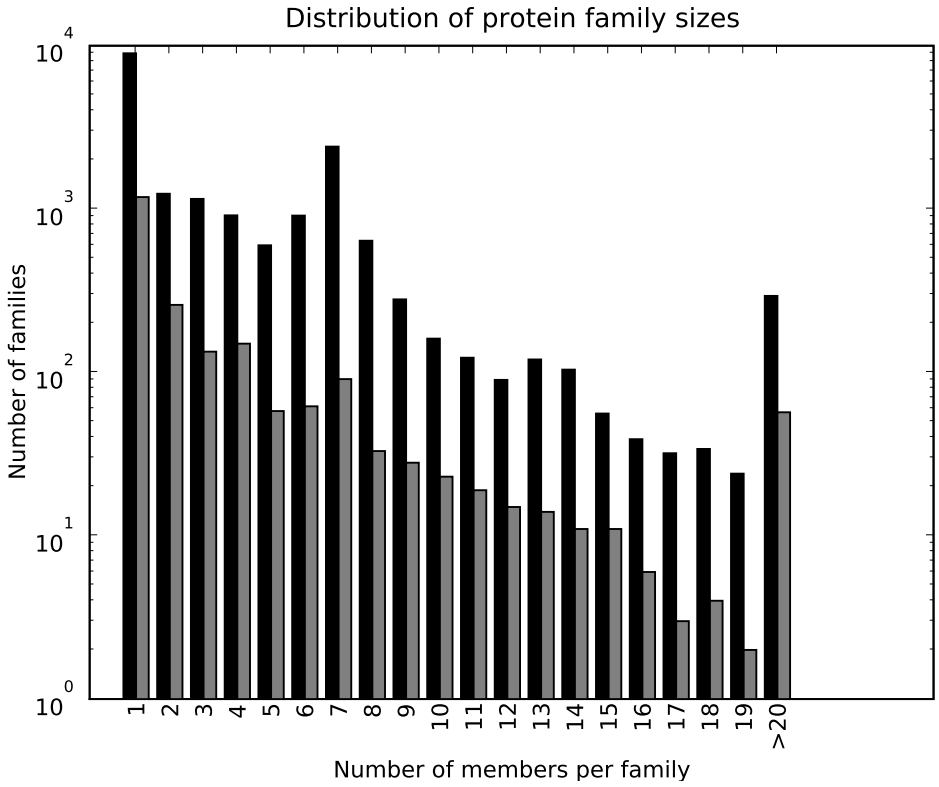

Figure 3.4: Distribution of protein families in six *Dothideomycete* species consisting of more than 40% secreted proteins (grey) and less then 40% secreted proteins (black). On the x-axis the size of the protein families is plotted against the number of families on the y-axis (log-scale). The last bin contains all protein families containing more than 20 proteins.

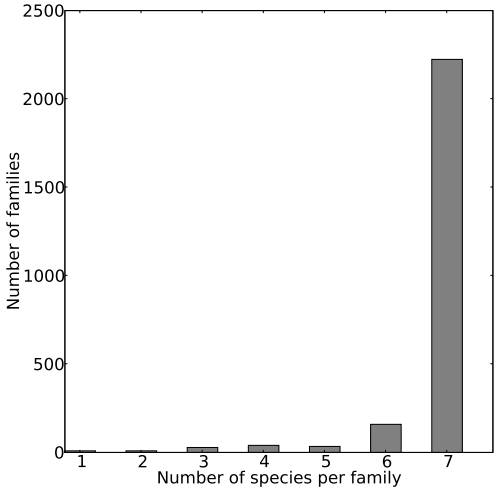

Figure 3.5: Protein families with seven members. The number of species per family is plotted against the total number of families with that amount of species. Most of the protein families with seven members contain a protein from each of the seven analysed *Dothideomycetes*.

domain per species and dividing it by the standard deviation. Per species, the z-score denotes if for a certain Interpro term, more or less proteins are present in respect to the mean protein count for all species.

In figure 3.6A the 25 most frequently occurring Interpro term in the secretome of *C. fulvum* can be seen together with the number of proteins with this domain in the other Dothideomycetes. The 61 proteins with a putative cytochrome P450 (CYP) domain (IPR001128) form the largest group of proteins in the secretome of *C. fulvum*. This number is significantly higher as compared to the average number of 29 CYPs found in the other Dothideomycetes. For example, in the secretome of *M. fijiensis* there are 36 CYP proteins predicted and for *S. nodorum* 28 CYP proteins. The z-score for *C. fulvum* (1.89) indicates that this group of proteins is more abundant in the secretome of this species. In figure 3.6A this can be seen as a green box (high z-score) with respect to a black or red box (low-zscore). Also, the number of predicted CYP proteins in *Mycosphaerellaceae* is higher than for *Pleosporaceae* (*S. nodorum, P. tritici-repentis, C. heterostrophus and A. brassicicola*).

The second largest group of proteins in the secretome of *C. fulvum* are the predicted glycoside hydrolases. In *C. fulvum* 57 proteins are predicted to contain the catalytic core domain of glycoside hydrolases (IPR017853). This domain is common to many different families of glycosyl hydrolases and are involved in plant cell wall degradation (see section below). The absolute number of proteins with this domain is almost the same for *C. fulvum* and *S. nodorum*, but for *M. graminicola* there are only 33 proteins predicted.

There are also domains present in specific families of glycoside hydrolases. In the *C. fulvum* genome for example 12 putative glycoside hydrolases of family 3 (GH3) (IPR002772) are predicted to be encoded, where the other Dothideomycetes have between 5 and 8 proteins with the same domain. The GH3 family consist of a number of glucosidases and xylanases, including -glucosidase, glucan 1,4–glucosidase and xylan 1,4–xylosidase. This suggests that the cell wall of the tomato plant is composed of several components, which are hydrolised by members of the GH3 family, and are not present in the hosts of the other Dothideomycetes. The GH3 family is also more abundant in all proteins encoded in the genome as compare as compared to the other species (data not shown). This means the GH3 family is not only larger in the secretome, but there are more members of this family encoded in the genome of *C. fulvum*.

In figure 3.6B the 25 Interpro domains with the highest z-score for *C. fulvum* are shown. There are more proteins predicted to contain these Interpro domains in *C. fulvum* than for the other species. In the figure, this can be seen as a green box versus a black or a red box in the other species.

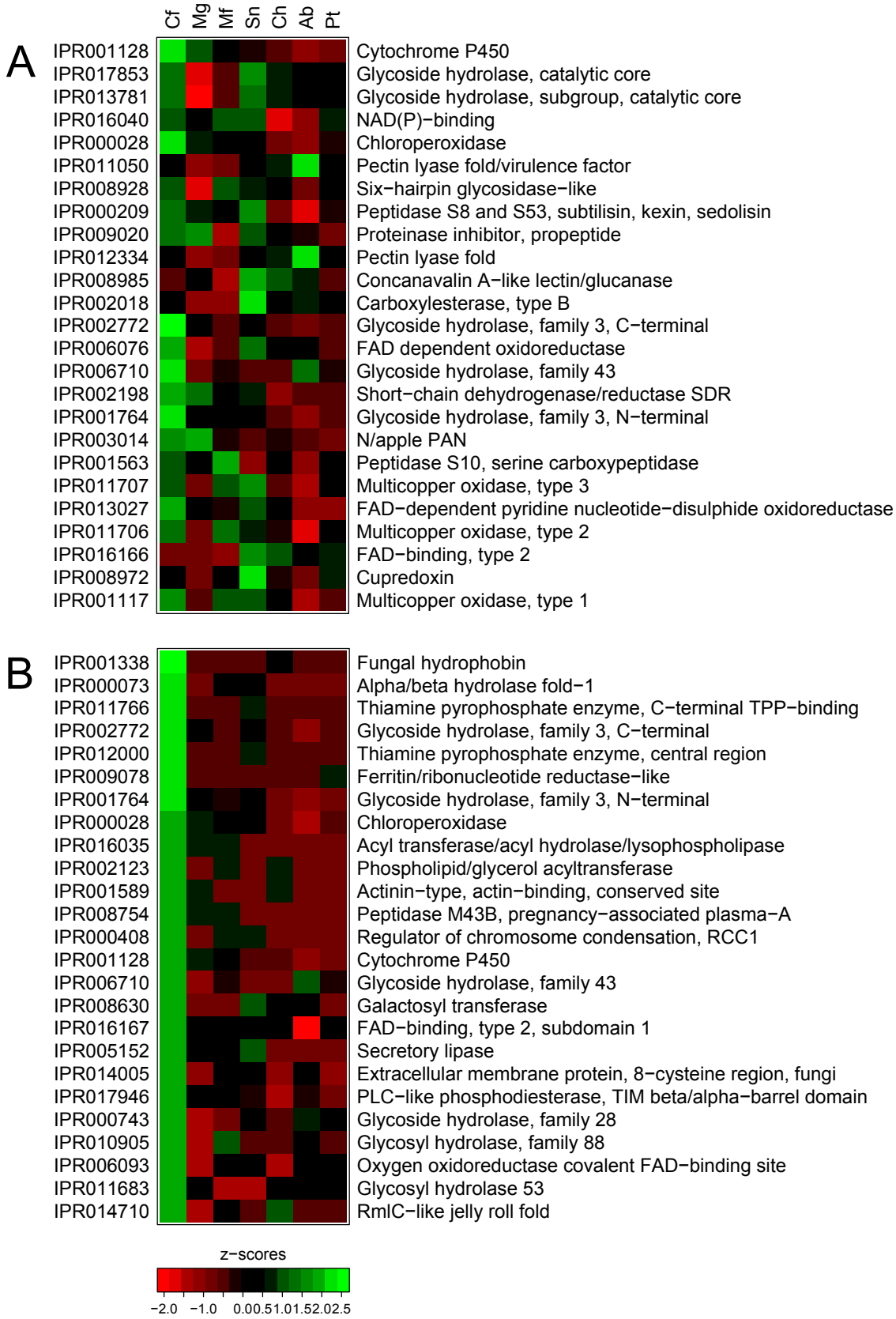

Figure 3.6: Levelplot of the 25 most frequently occuring Interpro terms in the secretome of *C. fulvum* seen together with the number of proteins with this domain in the other *Dothideomycetes* (A) and the 25 interpro domains with the highest z-score for *C. fulvum* (B). Coloring is according to z-scores. A high z-score (green) denotes that a specific Interpro motif is present in more proteins compared to other species with a lower z-score (red).

For example, there are five proteins in *C. fulvum* predicted to have a domain specific for fungal hydrophobins, whereas *C. heterostrophus* is the only other species containing one protein with this domain. Because hydrophobins have been studied before in *C. fulvum* and now they appear to be almost unique they will be discussed in more detail in the next section.

3.3.1 Hydrophobins

All six known hydrophobins genes were found back in the automated gene prediction performed on the genome of *C. fulvum*, except HCf-5 (Table 3.3). However, HCf-5 is present in the genome (TBLASTN, data not shown). Four of the proteins (HCf-1 to HCf-5) are predicted to contain a domain common in fungal hydrophobins (IPR001338) while HCf-6 has a domain found in hydrophobins, which is restricted only to *Ascomycetes* (IPR010636). In the genome of *C. fulvum* we identified four additional genes, which code for a protein containing one of the hydrophobin specific domains. One protein is more similar to Class I hydropobins (Cf|2591325) and the other proteins have a domain similar to Hcf-6 and thus belong to Class II hydrophobins. All the new putative hydrophobins contain 8 conserved cysteines residues, except for Cf|2591325, which has no cysteine residue at the fourth conserved position (Figure 3.7 and 3.8). All these proteins are also relatively small. The class I hydrophobins are approximately 100 amino acids long, although HCf-4 is an exception to this (228 aa). The class II hydrophobins are larger, with an approximate size of 150 aa. One of the new putative class II hydrophobins is significantly larger with a size of 358 aa.

An additional putative class I hydrophobin is detected by using a HMM profile search. A HMM profile was build based on the conserved cysteine motif in class I hydrophobins. A search against the *C. fulvum* genome yielded a hydrophobin candidate (Cf|2594738), containing a similar cysteine spacing pattern like class I hydrophobins. This protein was missed by standard similarity searches because it has several low complexity regions.

The class I and class II hydrophobins form distinct protein families. All known class I hydrophobins (HCf-1 to -6) group together in a single *C. fulvum* specific family, while the putative class I hydrophobin (Cf|2591325) ends up in a singleton, indicating limited sequence similarity. The hydrophobin domain in this protein is probably the most conserved part of the sequence, while the rest of the protein has changed as compared to the other hydrophobins. The class II hydrophobins form a protein family with proteins from other Dothideomycetes like *M. graminicola* (4), *M. fijiensis* (1), *S. nodorum* (2), *P. tritici-repentis* (3), *C. heterostrophus* (3) and *A. brassicicola* (2).

Figure 3.7: Alignment of Class I hydrophobins in *C. fulvum*. The conserved cysteine motif is clearly visible. The putative hydrophobin Cf|2591325 is the only protein without a cysteine at position 189.

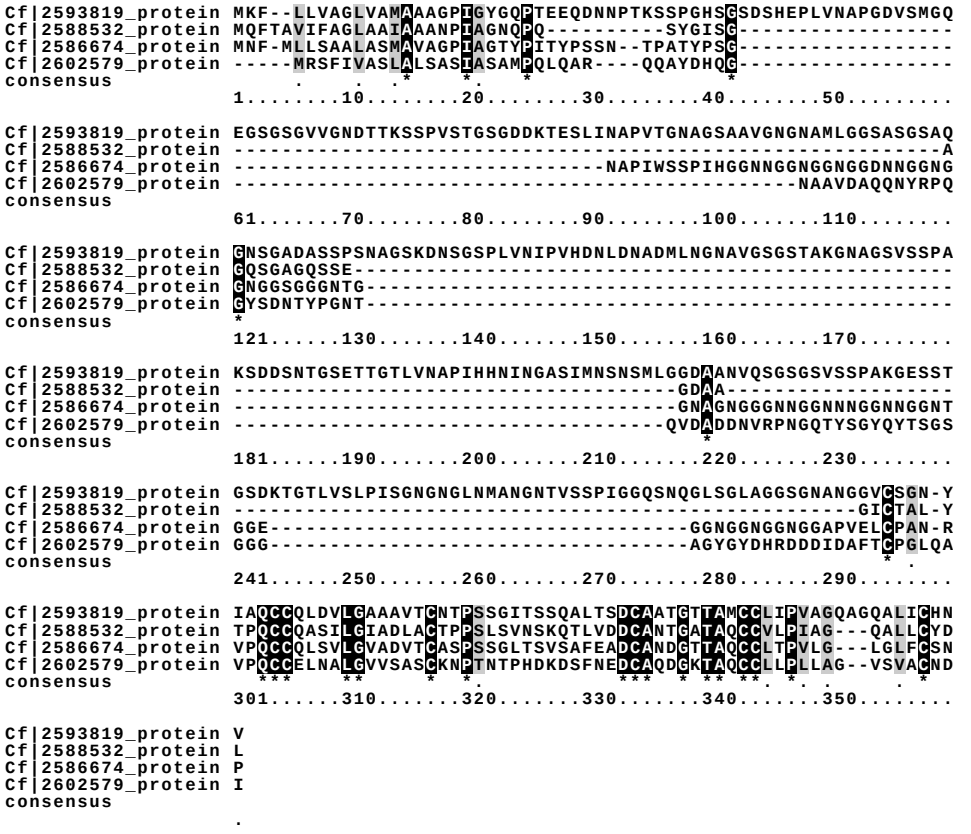

Figure 3.8: Alignment of Class II hydrophobins in *C. fulvum*. The motif of eight cysteine residues is conserved in all protein sequences.

Table 3.3: Known and novel class I and class II hydrophobins in *C. fulvum*

Class	**Gene name**	**Protein**	**Size (aa)**	**Interpro**
I	hcf-1	Cf|2592466	105	IPR001338[1]
I	hcf-2	Cf|2595228	121	IPR001338
I	hcf-3	Cf|2588266	103	IPR001338
I	hcf-4	Cf|2597786	228	IPR001338
I	hcf-5	-	88	-
I	-	Cf|2591325	110	IPR001338
I	-	Cf|2594738[3]	268	-
II	hcf-6	Cf|2602579	165	IPR010636[2]
II	-	Cf|2593819	358	IPR010636
II	-	Cf|2586674	184	IPR010636
II	-	Cf|2588532	112	IPR010636

[a]Hydrophobin domain
[b]Hydrophobin 2 domain restricted to Ascomycetes
[c]Detected with HMM profile search

3.4 More remarkable patterns in domain counts

Another group of proteins are the ones that are predicted to contain a domain that is not present in the proteins of other Dothideomycetes (Figure 3.9A). Most of the times one protein is predicted in *C. fulvum* that has an Interpro motif that is not present in any other proteins. This can indicate false annotations in the secretome or even not correctly predicted genes. However, it can also mean that these proteins are unique for *C. fulvum*, because it grows in a different host as compared to the other species. For example, the *C. fulvum* secretome contains 2 ubiquitins, a protein with lysozyme activity belonging to the glycoside hydrolase family 25 and several other proteins, which are predicted to be secreted. These proteins are unique to *C. fulvum* and were not found in any of the other six Dothideomycete species examined.

There are also several proteins predicted to be secreted in other Dothideomycetes but were absent in the secretome of *C. fulvum* (Figure 3.9B). This can be an indication of the number of proteins, which are false negatives regarding signal peptide predictions. This is the case for proteins of the peptidase S41 family. In each of the Dothideomycete species there is at least one protein present containing the functional domain for this family except for *C. fulvum*. The genome of *C. fulvum*

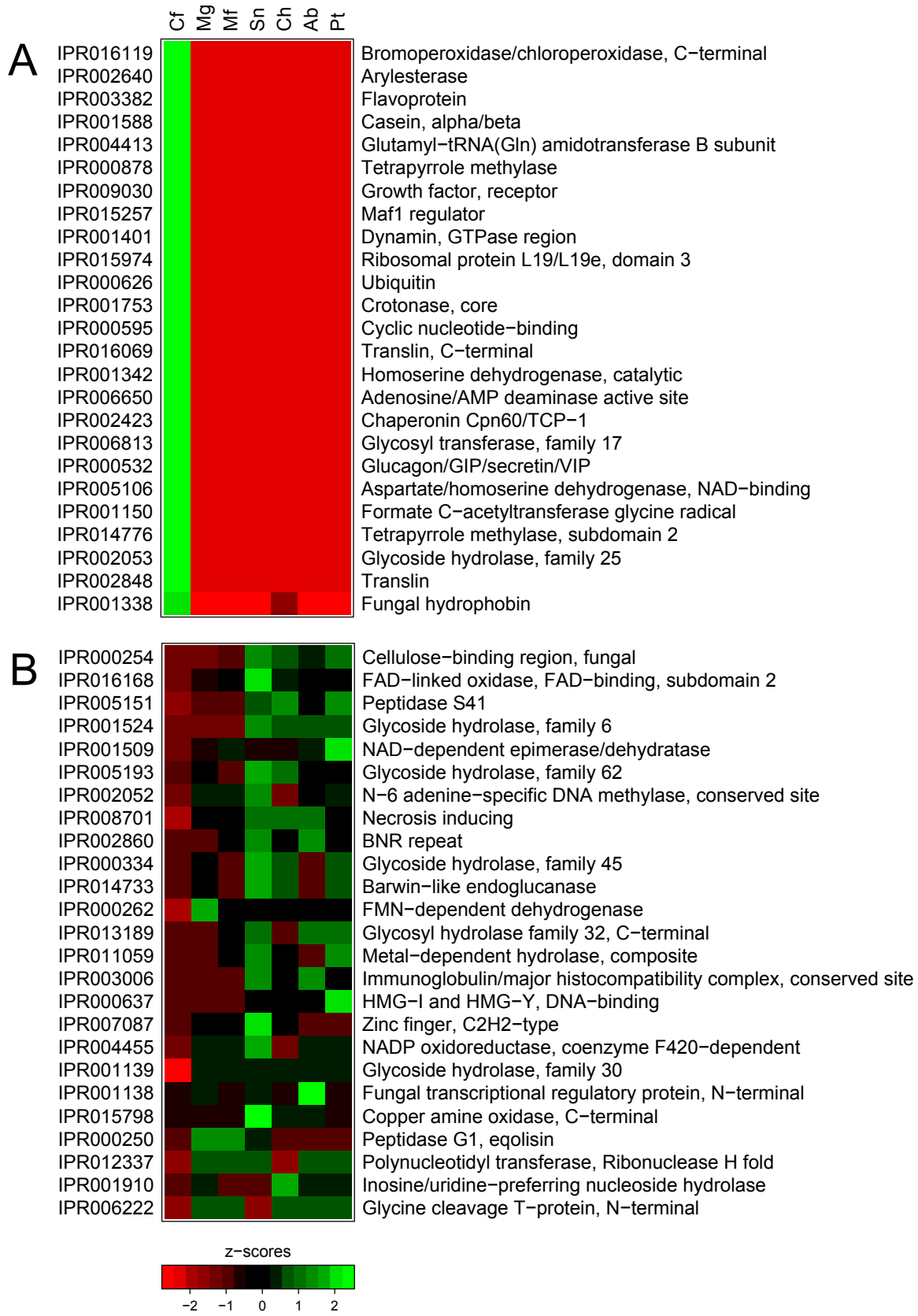

Figure 3.9: Levelplot of proteins in *C. fulvum* containing a domain that is not present in the proteins of other *Dothideomycetes* or domains that are not present in secreted *C. fulvum* proteins (B). Coloring is according to z-scores. A high z-score (green) denotes that a specific Interpro motif is present in more proteins compared to other species with a lower z-score (red).

encodes three proteins of the peptidase S41 family, but none of them is predicted to be secreted. Another possibilities are that these genes are missed by the gene prediction models, are not included in the assembly or are not present in the genome of *C. fulvum*.

A final group of proteins that will be discussed over here are the proteins with a domain present in *Mycosphaerellaceae* but absent in *Pleosporaceae* or *vice versa* (Figure 3.10). For example, there are proteins in the Pleosporaceae family containing several domains related to copper amine oxidase, which are not present in the *Mycosphaerellaceae*. Also, the glycoside hydrolase family 6 domain is present in at least 3 proteins in the Pleosporaceae, but neither *C. fulvum* or any of the other Mycosphaerellaceae is predicted to contain a protein with this domain. This can mean that these proteins are actually not encoded in the genome of these species, or they or not predicted to be secreted.

3.4.1 Plant cell wall degradation

Until know the focus was mainly on the comparison of secreted proteins within the class of Dothideomycetes. It is also interesting to analyze how *C. fulvum* can degrade parts of the plant cell in order to survive in the apoplast. Therefore, in this section proteins containing a domain associated with plant cell wall degradation are discussed.

A fungus needs to weaken a plant cell wall in order to survive in the host. *C. fulvum* enters the plant cell through the lower side of the leaves and via the stomata (Thomma *et al.*, 2005). It has not been shown that *C. fulvum* can forcibly penetrate plants by degrading the cuticle layer present at the leaf surface, as is common for other pathogens. It is therefore interesting to observe that the genome of *C. fulvum* is predicted to encode for 8 putative cutinases (IPR000675), which can catalyse the hydrolysis of cutin. If *C. fulvum* is really able to penetrate through the cuticle or that the cutinase have a different function depends the stage of infection in which they are expressed. Three of these putative cutinases occur in the list of small secreted proteins.

A number of putative endoglucanases and endoxylanases are predicted to be secreted from *C. fulvum* in the tomato apoplast. This suggests that cellulose and xylan from the plant can be hydrolysed by the fungus upon infection. The Carbohydrate-Active Enzyme (CAZy) database contains an overview of enzymes that breakdown carbohydrates (Cantarel *et al.*, 2009). These glycosyl hydrolases are grouped in families based on sequence similarity.

The genome of *C. fulvum* contains two predicted proteins of the glycoside hydrolase family 11 (GH11) and another two of the glycoside hydrolase family 12 (GH12). All these proteins contain the

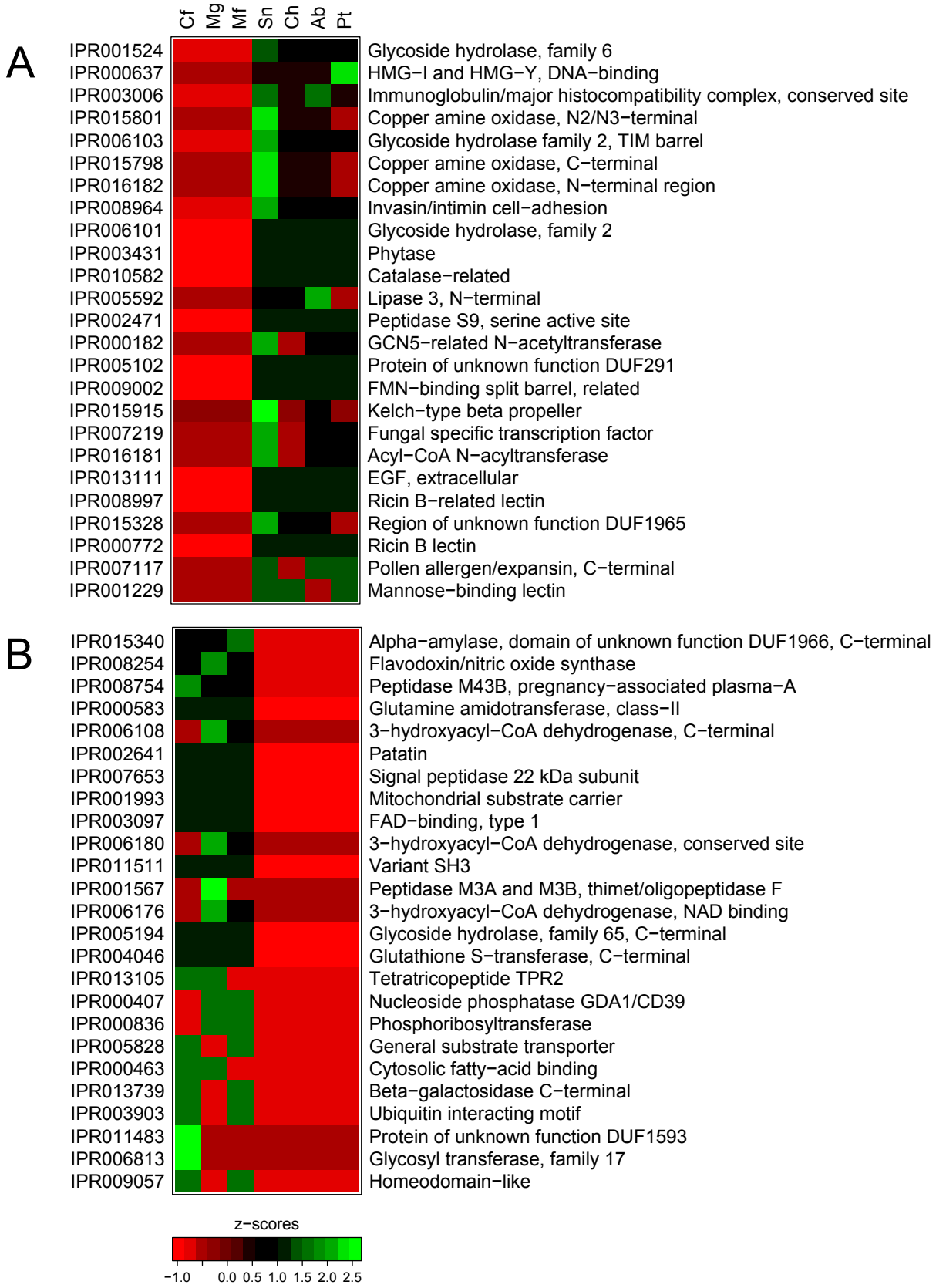

Figure 3.10: Levelplot of proteins with a domain present in *Mycosphaerellaceae* but absent in *Pleosporaceae* or *vice versa*.Coloring is according to z-scores. A high z-score (green) denotes that a specific Interpro motif is present in more proteins compared to other species with a lower z-score (red).

catalytic domain for glycoside hydrolases (IPR013319). Proteins of the GH11 family are only able to breakdown xylan (Cf|2588225, Cf|2600360) while the other family contains proteins (Cf|2598575, Cf|2587716) with a broader range of activities (endoglucanases, xyloglucanases). The two proteins belonging to the GH12 family were also annotated with the Gene Ontology term for cellulase activity using the Blast2GO tool.

Besides cellulose and xylan, *C. fulvum* is predicted to be able to break down pectin in the plant cell wall. Pectate lyases are fungal virulence factors that degrade the pectic components of the plant cell wall (Mayans *et al.*, 1997). There are 7 proteins of the GH28 family predicted in the genome of *C. fulvum*. This family contains a group of polygalacturonases, which are able to hydrolyse pectate and other galacturonans. All these proteins contain a pectin lyase fold domain (IPR011050), a single-stranded, right-handed parallel beta-helix (IPR006626) and have the Gene Ontology classification for carbohydrate metabolic process. The same domains are present in a pectin lyase of *Aspergillus niger* and several other virulence factors including bacterial pectate lyases, fungal and bacterial galacturonases (Jenkins *et al.*, 1998). Another group of 8 proteins was predicted in *C.fulvum,* which all contain a pectin lyase domain but are missing the parallel beta strands common to pectate lyases. It is possible that the beta strands in these proteins are not correctly predicted or they form a class of proteins that fold in a different way as known pectate lyases.
The pectate lyases in *C. fulvum* form multiple protein families with related proteins in other Dothideomycetes. For example, three pectate degrading proteins (Cf|2589803, Cf|2592083, Cf|2594072) form a family with 13 pectate lyases from *A. brassicicola, C. heterostrophus M. fijiensis, P. triticirepentis* and *S. nodorum*. This indicates that the proteins involved in the breakdown of pectin in the cell wall are conserved among Dothideomycetes and relative species.

C. fulvum might also be able to degrade lignin from the plant cell wall. The first lignin peroxidase has been found in the white-rot basidiomycete *Phanerochaete chrysosporium* (AAA33739, Ligninase LG5). This protein shows some sequence similarity with a protein in *C. fulvum* (Cf|2600264). Lignin peroxidases can contain several functional domains, including a haem peroxidase domain (IPR002016) and a ligninase domain (IPR001621). The putative lignin peroxidase in *C. fulvum* contains a haem peroxidase domain but there is no ligninase domain predicted. The same is true for two other proteins (Cf|2588060; Cf|2595350).

The heam peroxidase domain is a signature of lignin peroxidases. The proteins in *C. fulvum* might still be able to degrade lignin, but they can contain a lignin peroxidase domain that is slightly different from other ligninases. This can explain why it is not predicted in the genome of *C. fulvum*.

Table 3.4: Predicted enzymes involved in degradation of the plant cell wall

Proteins	**Interpro domains**	**Enzymatic function**	**Plant cell wall component**
Cf|2588225; Cf|2600360	IPR001137	GH11, xylan degradation	xylan
Cf|2587716; Cf|2598575	IPR002594; IPR008985; IPR013319	endoglucanase	cellulose
Cf|2587312; Cf|2589803; Cf|2592083; Cf|2599059; Cf|2588509; Cf|2593889; Cf|2600003	IPR000743; IPR011050; IPR006626	Pectin lyase	pectin
Cf|2592674; Cf|2588787; Cf|2601629; Cf|2599841; Cf|2598462 ;Cf|2590966; Cf|2588439; Cf|2601539; Cf|2594885	IPR011050	Pectate lyase	pectin
Cf—2600264; Cf—2588060; Cf—2595350	IPR00216; IPR001621	Lignin peroxidase	lignin

3.4.2 Peptidases

In agreement with the secretome of *U. maydis*, a pathogen of maize, the genome of *C. fulvum* encodes a number of putative secreted proteins that degrade other components of the plant cell besides cell wall degradation. These enzymes probably help the fungus to survive within the plant and suppress defense mechanisms of the plant (Mueller *et al.*, 2008).

The set of secreted proteases includes a number of endopeptidases and carboxypeptidases. The group of serine-type endopeptidases (GO:0004252; IPR000209; IPR015366) (17 proteins) is clearly overrepresented as compared to the other endopeptidases like aspartic- and metallo-endopeptidases (both 3 proteins). The majority of the serine-type endopeptidases belong to the S8 and S53 family according to the MEROPS peptidase database (Rawlings *et al.*, 2008).

3.5 Small secreted proteins

The effector proteins are a special class of proteins in the secretome. They are mostly small and contain a large number of cysteines. Currently ten effector proteins are known for *C. fulvum*. Four Avirulence (Avr) genes, namely Avr2, Avr4, Avr4E and Avr9 and six Extracellular proteins (Ecps), Ecp1, Ecp2, Ecp4, Ecp5, Ecp6 and Ecp7, have been identified. All these secreted effector proteins are relatively small (between 60 and 230 amino acids) and contain an even number of cysteine residues. Five of them are picked up by automated gene prediction (GeneMarkES) (Table 3.5). Most gene

Table 3.5: Characteristics of the effector proteins in *C. fulvum*. Some known effector genes are not present in the automated genome annotation, and are therefore not used in this analysis (data shown as not applicable (n.a.))

Gene name	**Protein ID**	**Size (aa)**	**Secreted**	**Interpro**	**Orthologs (only in Mf, Mg)**	**Family size (in Cf)**
Avr2	-	78	n.a	n.a.	n.a.	n.a.
Avr4	Cf|2597781	135	yes	IPR002557[1]	1	-
Avr4e	Cf|2597260	121	yes	IPR013032[2]	0	-
Avr9	-	63	n.a.	n.a.	n.a.	n.a.
Ecp1	-	96	n.a.	n.a.	n.a.	n.a.
Ecp2	Cf|2589111	165	yes	-	2	3
-	Cf|2591819	350	yes	-	1	3
-	Cf|2587574	147	yes	-	0	3
Ecp4	Cf|2589430	138	yes	-	0	-
Ecp5	-	115	n.a.	n.a.	n.a.	n.a.
Ecp6	Cf|2601254	228	yes	IPR002482[3]	2	1
Ecp7	-	84	n.a.	n.a.	n.a.	n.a.

[a]Chitin binding protein, peritrophin-A
[b]EGF-like region, conserved site
[c]Peptidoglycan-binding LysM

prediction programs have difficulties predicting small open reading frames (Hanada *et al.*, 2007). This explains why only the five largest effector proteins are detected.

Avr4 encodes a 135 amino acid protein, which binds to chitin and protects *C. fulvum* against chitinases during infection (Stergiopoulos and de Wit, 2009; van den Burg *et al.*, 2006). Avr4 contains a chitin binding domain (IPR002557), and a homologue is found inincluding those of *M. fijiensis*.

The only two effector proteins, which have many orthologs in other *Ascomycetes* and Dothideomycetes are Ecp2 and Ecp6. Ecp2 forms a family with two other proteins from *C. fulvum*, one from *C. heterostrophus*, three from *M. fijiensis* and one from *M. graminicola*. The alignment of Ecp2 homologs in *C. fulvum* (Figure 3.11) and the other Dothideomycetes (Figure 3.12) reveals a conserved pattern of amino acids, including four cysteine residues. Based on this conserved motif in the alignment a HMM model is build to find additional proteins. Five other proteins contain a similar pattern of amino acids (Table 3.6). The Ecp2 homologs in *P. tritici-repentis* and *S. nodorum* were not detected using standard similarity searches. Moreover, two additional homologs in *M.*

Table 3.6: Homologs of the Ecp2 protein in *C. fulvum*

Organism	**Protein ID**
C. fulvum	Cf|2589111, Cf|2591819, Cf|2587574
M. fijiensis	Mf|52972; Mf|23545; Mf|60658
M. graminicola	Mg|104404; Mg|107904; Mg|111636
C. heterostrophus C5	Ch|100623
P. tritici-repenti	Pt|PTRT_07799; Pt|PTRT_00203
S. nodorum	Sn|SNOT_04278

graminicola are found, making the total of Ecp2 homologs per species for *C. fulvum* , *M. graminicola* and *M. fijiensis* to a total of three.

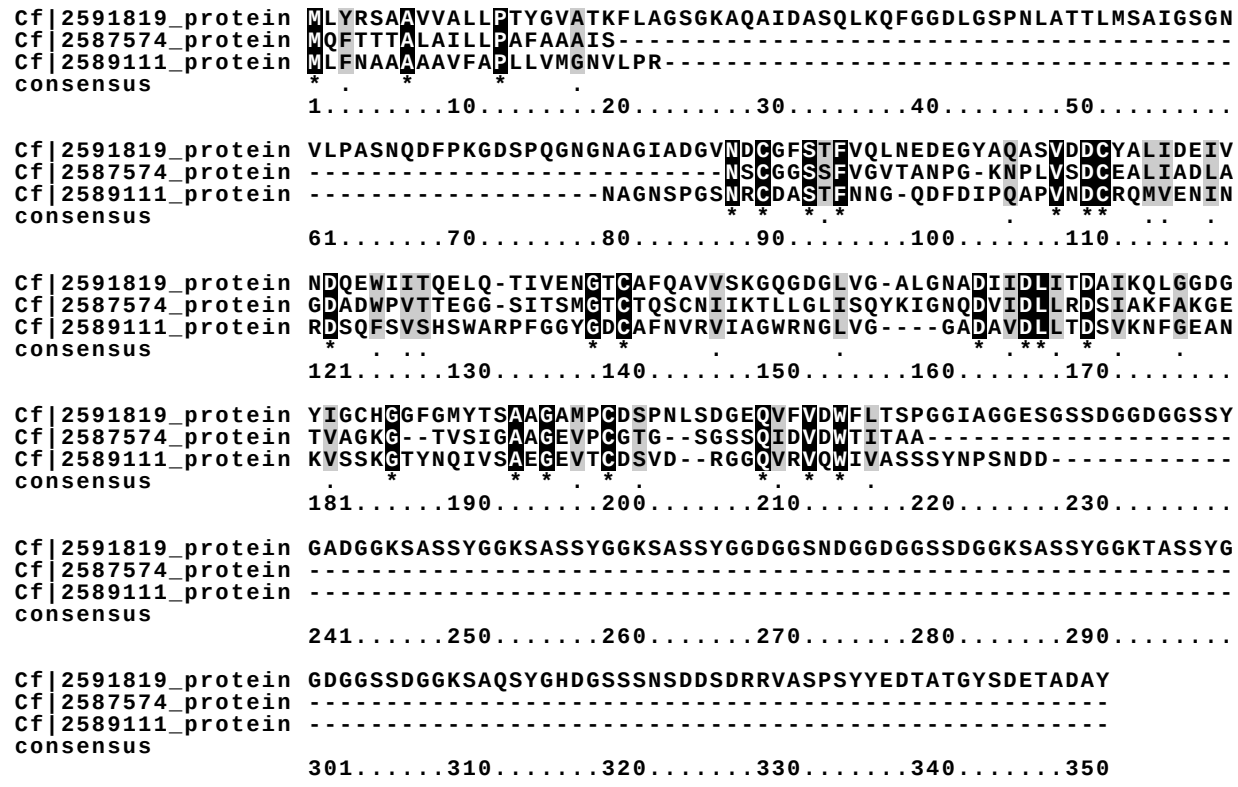

Figure 3.11: Alignment of the Ecp2 protein of *C.fulvum* (Cf|2589111) with potential paralogs.

Ecp6 contains a LysM domain (IPR002482), which is common in carbohydrate binding proteins and may be involved in chitin binding (de Koster *et al.*, 2008). In this study, Ecp6 orthologous have been found in *M. graminicola* and *M. fijiensis*, which also contain the LysM domain. However, the Ecp6 protein in *C. fulvum* does not group with other proteins in a family based on sequence

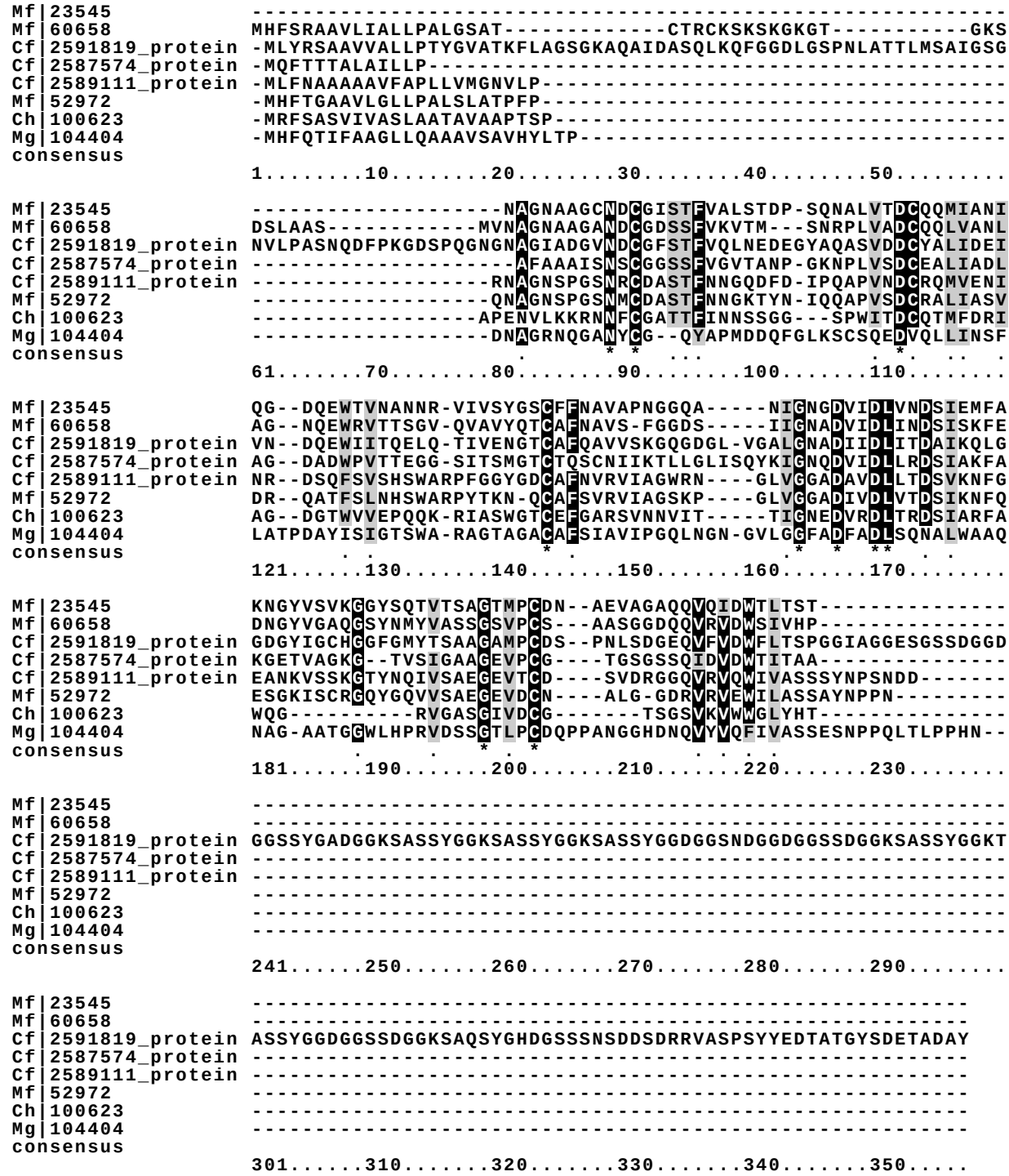

Figure 3.12: Alignment of the *C. fulvum* Ecp2 proteins with homologs in *M. fijiensis*, *M. graminicola* and *C. heterostrophus C5*

similarity. Among the secreted proteins in *C. fulvum* there are two proteins containing a LysM domain (Cf|2596443 and Cf|2589602) with a similar size as Ecp6 (276 and 223 aa respectively). These proteins could be potential paralogs of Ecp6.

3.6 Additional findings regarding the genome of C. fulvum

3.6.1 Ubiquination

Ubiquination is one of the most common post-translationd modifications after phosphorylation and it is involved in a number of cellular processes, including meiosis, cellular proliferation and development (Semple, 2003). The most well-known role of ubuiqitin is labeling proteins for degradation, but it is also shown to play a role in transcriptional regulation (Conaway *et al.*, 2002).

The transfer of ubiquitin requires at least three enzyme complexes to be active in order to activate (E1, IPR000011), conjugate (E2, IPR000608) and ligate (E3, IPR000569, IPR003613) ubiquitin covalently to a substrate (Furukawa *et al.*, 2003). The familly of E3 ligases is formed by proteins containing either a HECT (Homologous to the E6-AP Carboxyl Terminus) domain or a RING (Really Interesting New Gene)domain (Geyer *et al.*, 2003). Among the best studied E3 RING ligases are the SCF (Skp1-Cullin-F-box protein) complexes. They are composed of proteins containing a SKP1 (S-phase kinase-associated protein 1) component (IPR001232), cyclin-like F-box domain (IPR001810), cullin (IPR001373) and a zinc-finger RING (IPR001841). The F-box adaptors mediate substrate binding and specificity (Geyer *et al.*, 2003). The SCF complexes contain separate adapter and substrate proteins, although for CUL3 ligases, the adaptor and substrate-receptor functions are merged in a single 'Broad complex, Tramtrack, Bric-a-brac' (BTB)-domain-containing polypeptide (Petroski and Deshaies, 2005).

A key feature of E3 ligases is that each cullin can assemble with numerous substrate receptors to form ubiquitin ligases that share a common catalytic core yet recruit different substrates (Petroski and Deshaies, 2005). For CUL1, the substrate receptors are recruited by the adapter protein SKP1. The N-terminal domain of SKP1 binds to CUL1 and the C-terminal region binds to a F-box motif of a substrate receptor (Figure 3.13a). The other cullins bind to the substrate receptors containing suppressor of cytokine signalling/elongin-BC (SOCS/BC) boxes (Figure 3.13b) or a BTB/POZ domain (Figure 3.13c).

The genome of *C. fulvum* encodes several proteins related to ubiquination, including cullins, zinc-fingers, F-box proteins and proteins involved in the three ubiquitin complexes. The number of proteins in *C. fulvum* containing a BTB/POZ domain, a cyclin-like F-box motif or a RING-type zinc

finger domain are increased as compared to species like *M. graminicola* and *M. fijiensis* (Figure 3.14). Interestingly, all these proteins are part of substrate receptors of E3 ligases. This means that a large number of different E3 ligases can be formed, targeting a broad range of substrates. As compared to the other Dothideomycetes *C. fulvum* looks like to exhibit a broader substrate-specifity for E3 ligases.

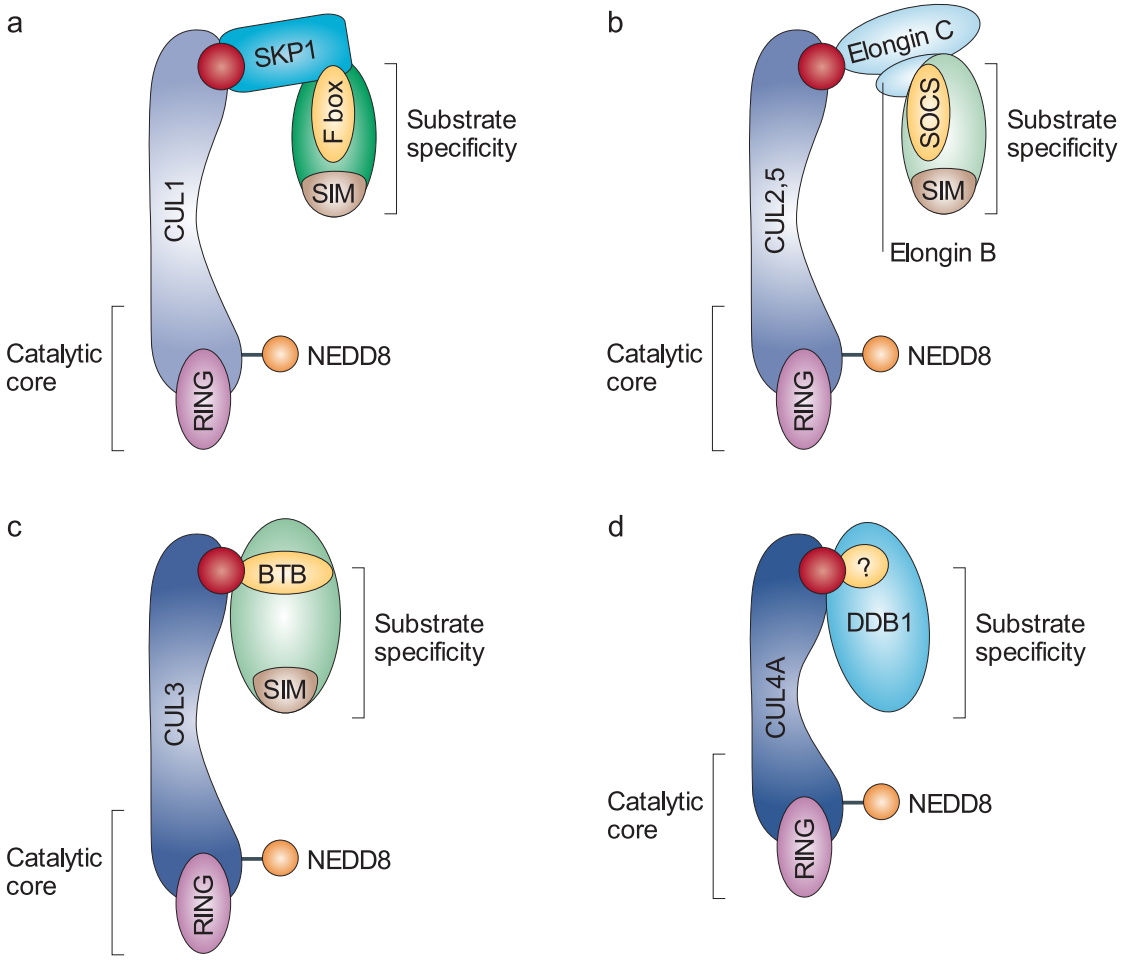

Figure 3.13: Composition of different cullin RING ligases. The motifs that link substrate receptors to their cognate adaptors are shown in yellow. (Petroski and Deshaies, 2005)

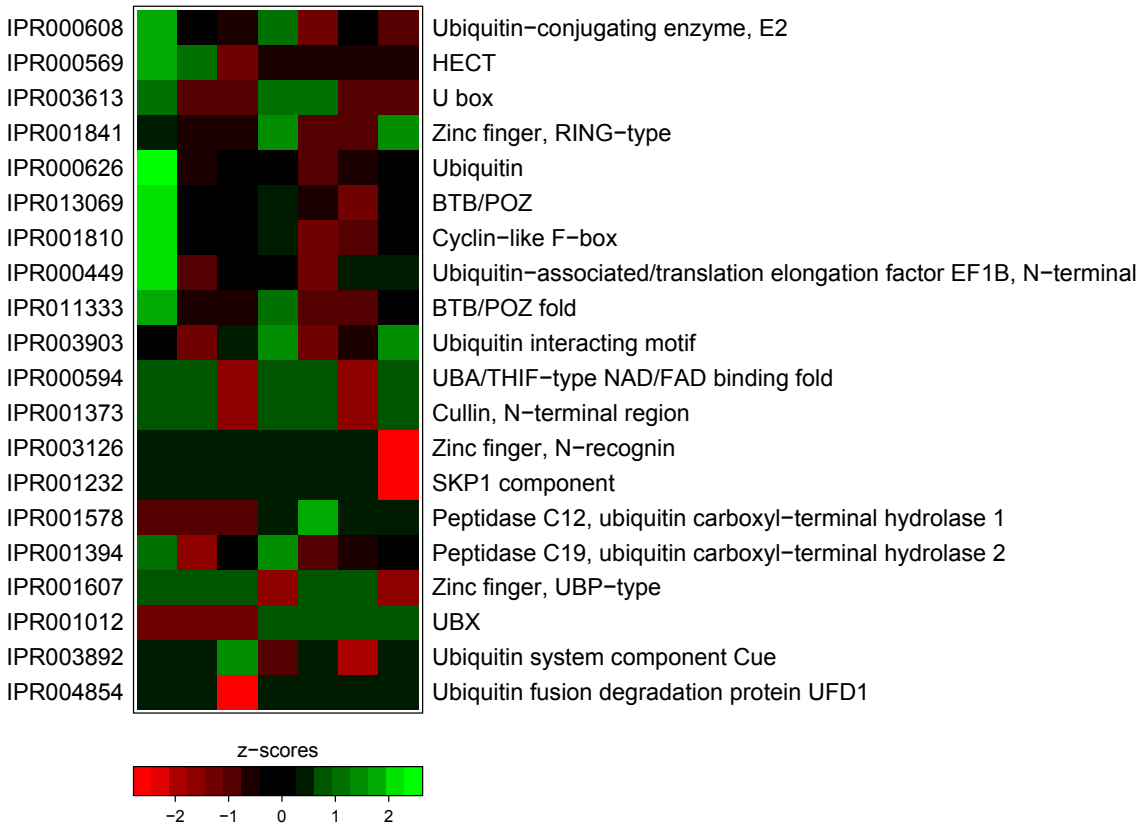

Figure 3.14: Levelplot of proteins in *C. fulvum* associated with ubiquination. Coloring is according to z-scores. A high z-score (green) denotes that a specific Interpro motif is present in more proteins compared to other species with a lower z-score (red).

Chapter 4

Discussion

The predicted secretome in *C. fulvum* consists of 7.6% of all predicted proteins encoded in the genome. For the other analysed Dothideomycetes the secretomes range from 7.8% to 10.7% of all predicted proteins per species, with an average secretome size of 8.9%. The secretomes of other fungal pathogens have been predicted previously. For example, the secretome of *U. maydis* compromises 6.1% of the predicted protein-encoding genes (Kämper *et al.*, 2006) while the rice blast fungus *Magnaporte grisea* has a predicted secretome of 6.7% (Pan *et al.*, 2005). The secretome size can vary greatly depending on the programs that are being used. In a comparative secretome analysis of phytopathogenic ascomycetes, 12% of all predicted proteins in *M. grisea* were predicted to be secreted (Talbot *et al.*, 2008) in contrast to the previously reported 6.7%. The predicted secretome of the other fungi varied from 5%-12% of all predicted proteins. This indicates that it is difficult to compare the number of predicted secreted proteins if different programs have been used.
It needs to be stated that all conclusions in this study are based upon predictions. First, it will be necessary to prove that the proteins for which we predicted a secretion signal are actually secreted. The SignalP 3.0 program, which has been used for signal peptide prediction, has a accuracy of 96% in bacteria (Zhang *et al.*, 2009) and 90% in eukaryotes (Klee and Ellis, 2005). This shows that the prediction of secretion is accurate, but proteins of interest still need to be experimentally validated.

Of all 1,275 *C. fulvum* proteins predicted to be secreted, 716 (56%) can not be assigned a function. In *U. maydis* it is also shown that most of the secreted proteins cannot be functionally annotated (70%) (Kämper *et al.*, 2006). This indicates that it is common for secreted proteins in fungi that only one out of two proteins can be annotated. The number of small secreted proteins with annotation is even lower.

Furthermore, the quality of the annotation is directly related to the quality of the predicted gene structures. The contigs in the draft genome of C. fulvum have a median coverage of 10.4x, resulting

in 7,877 contigs divided over 4,755 scaffolds. The relatively high number of contigs makes it probable that some genes are divided over multiple contigs. Since the genes were predicted exclusively using a de novo prediction algorithm, some genes may only be partially predicted or may miss introns/exons or have incorrectly predicted introns/exons. Expressed sequence tags (ESTs) or proteomics data is therefore useful to validate gene structures in the genome of *C. fulvum*. For example, ESTs obtained from cDNA libraries indicate which genes are transcibed into mRNA and have been used to validate predicted protein-coding sequences in *M. grisea* (Numa *et al.*, 2009). Futhermore, proteomics approaches can confirm which transcripts are translated into proteins. This has been demonstrated to be advantageous for several species including *Aspergillus niger* (Wright *et al.*, 2009) and *Arabidopsis thaliana* (Baginsky *et al.*, 2008). Both EST and proteomics data can help to validate the gene structures for *C. fulvum*.

Finally, to have certainty about protein functions, the location in the host can indicate whether the predicted secreted protein is indeed associated with the fungal cell wall, the apoplast or that proteins can even be translocated across the plasma membrane.

4.1 Protein family clustering and annotation

The protein clustering resulted in multiple protein families, which were given the same annotation. This means that 40% of the proteins in these families contain the same Interpro domain(s), but are nevertheless grouped in a single cluster. Detailed inspection of some of these families demonstrates that there is not enough local sequence similarity between the members of the different families.

For example, the cutinase family, which the proteins of these family can break down the cuticle of the plant and forms the first barrier for infection. As described in the Results section, the *C. fulvum* genome encodes 8 putative cutinases based on the presence of the Interpro domain for cutinases (IPR000675), all of which are also predicted to be secreted. These proteins are divided over four MCL clusters. Two *C. fulvum* proteins (Cf|2588886, Cf|2599123) form a cluster with 31 proteins from all other analysed Dothideomycetes. They have moderate BLAST hits (percentage identity up to 40% and e-value lower than 0.001) to all proteins in the same cluster, but not to proteins of the other cutinase clusters. This explains why the MCL algorithm was not able to group them together, even with lower inflation factors. Apparently, the proteins containing the cutinase domain do not share enough sequence similarity to have significant BLAST hits, but are nevertheless similar enough to match the same Hidden Markov Model (HMM) describing the cutinase domain.

A comparable observation can be made for more protein families, including the glycoside hydrolase

family 43. This suggests that there are more clusters with the same Interpro annotation. The limiting factor for the MCL clustering are the sequence similarities provided by BLAST (data not shown). The question arises whether all proteins that contain the same Interpro motif should be grouped into a single protein family. Maybe the sequences of the proteins with the same domain composition have diverged so much that they form subfamilies or maybe not all protein domains have been predicted by InterproScan.

If it is required that proteins containing the same domain are clustered together in a single family in order to have one family per function, three alternative methods can be applied. First, an algorithm that also takes additional information on domains as input can be used. But the disadvantage is that those tools are hampered by multi-domain proteins and promiscuous domains as discussed in the introduction. A second option is to lower the BLAST threshold so that more distant proteins can be grouped together. Clustering proteins with a lower sequence similarity threshold brings the risk that unrelated proteins are clustered and it does not guarantee that proteins containing the same domain are included. A third and probably the best alternative method is to have another step after the MCL clustering. Clusters containing proteins with a similar domain structure could be merged together.

An interesting analysis to follow the protein clustering would be to have a closer look at protein family evolution. Differences in family size due to gene duplication and gene loss can provide clues to evolutionary forces that caused divergence within the Dothideomycetes. For example, the Computational Analysis of gene Family Evolution (CAFE) can be used for statistical analysis of the evolution of the size of protein families (Bie *et al.*, 2006). Given a phylogenetic tree and a matrix of protein family sizes it uses a stochastic birth-and-death process to model the evolutionary force and can identify protein families that have accelerated rates of gain and loss. Previously this tool was used to study the evolution of mammalian genomes in combination with the MCL algorithm (Demuth *et al.*, 2006). The CYP protein families and the hydrophobin family are predicted to be larger for *C. fulvum* than for the other Dothideomycetes, and are therefore good candidates for a gene gain/loss analysis.

Instead of analyzing the evolution of all protein families, the evolutionary history of a specific protein family can be studied. Divergent rates of gene gain and loss have been reported for the cutinase family in five Ascomycetes (Skamnioti *et al.*, 2008). They reported no positive selection on any of the cutinases in *M. grisea* is found, which indicates that no evolutionary pressure has acted on the cutinases after gene duplication. This means that the emergence of new functions is more likely

driven by changes in the promoter composition or in the interaction network of these cutinases. An analogous study can be applied on a *C. fulvum* protein family of interest.

Chapter 5

Conclusions and Future Work

5.1 Conclusions

In total 1,275 proteins from *C. fulvum* are predicted to be secreted including enzymes, which aid the fungus to establish a biotrophic relationship with its host. The presence of cell-wall degrading enzymes, which can break down components such as cellulose, xylan, pectin and lignin shows that *C. fulvum* is at least able to partially hydrolyse the plant cell walls. Also the predicted number of chloroperoxidases and glycoside hydrolases is higher in the *C. fulvum* secretome as compared to the secretomes of other Dothideomycetes. *C. fulvum* stays inside the apoplast during infection and until know it has not been shown that proteins can pass the plant cell wall. A role of the cell wall degrading enzymes can be the attachement to the surface of the cell wall and the partially degradation of cell wall components to obtain nutrients.

Besides degrading parts of the plant cell wall, proteases encoded in the *genome C. fulvum* indicate that other components of the plant cell wall are broken down. These enzymes probably help the fungus to survive within the plant and suppress defence mechanisms of the plant. The observed number of serine-type endopeptidases (17) is higher than the average number found for other types of endopeptidases (average 3). This may indicate that serine-type endopeptidases are more encoded in the *C. fulvum* genome , or that the other types of endopeptidases are not captured in the draft genome of *C. fulvum.*

Within the predicted secretome of *C. fulvum* there are 135 proteins selected, which are smaller than 255 amino acids and contain more than 5 cysteines residues, because this are the main characteristics of known effector proteins. These 135 small secreted proteins (SSP) could therefore be play a role in pathogenesis and should be analyzed more intensively using wet-lab experiments. Out of all 135 putative SSPs, 18.5% are assigned with an Interpro domain, including eight putative hy-

drophobins, three putative cutinases and two putative chitin-binding proteins.
Five of the hydrophobins are unique for *C. fulvum* and have not been reported before. The absence of these hydrophobins and cutinases in the other Dothideomycete genomes could be the effect of not predicted genes. The nucleotide sequences of the genes that encode these proteins should therefore be searched against the translated nucleotide sequences in all reading frames of those species, to detect missing genes. If the absence of these genes is confirmed the question why these proteins are more present in the *C. fulvum* genome should be addressed.
The secretome of *C. fulvum* encodes for 8 putative cutinases, with three of these proteins also occuring in the list of SSPs. This observation is quite remakable because cutinases are able to break down the cuticle layer of plant leaves. It has not been shown that *C. fulvum* can forcibly penetrate plants by degrading the cuticle layer. The presence of the putative cutinases in the genome of *C. fulvum* does not mean that the cuticle is broken down. The proteins therefore need to be studied in more detail. Maybe *C. fulvum* partially degrades the cuticle when it is growing at the upper side of plant leaves in order to obtain some nutrients.

The secretome of *C. fulvum* contains several groups of proteins, which occur more frequently as compared to other Dothideomycete genomes. For example, there are 61 putative cytochrome P450 (CYP) proteins encoded in the *C. fulvum* genome, as compared to an average value of 29 proteins in the other Dothideomycete genomes. The CYP proteins have a broad range of functions, including the biosynthesis of secondary metabolite toxins (Howlett, 2006). In biotrophic fungi these toxins can trigger plant cell death to provide nutrition for fungal growth and colonisation of plant tissue (Howlett, 2006). *C. fulvum* is a hemi-biotroph and in general this class of pathogens have an initial biotrophic phase, then become necrotrophic (Howlett, 2006). Necrotrophs often express cell-wall-degrading enzymes and toxins. The fact that the genome of *C. fulvum* is predicted to encode for proteins that break down parts of the cell wall and CYP proteins that produce toxins, makes it interesting to know when these proteins are expressed. Probably they are expressed during the necrotrophic phase.
Five out of the ten known effector proteins are predicted using automated gene finders. Effector proteins are usually very small, which explains why they are not picked up in this way. This indicates that the number of small secreted proteins in fact could be twice as large as now has been found. Two new putative homologs of the Ecp2 form a protein family with proteins form other Dothideomycetes. The alignment of these proteins reveals a conserved pattern of amino acids, including four cysteine residues. This pattern can be used to build a Hidden Markov Model to search for similar patterns

in other fungal species.

5.2 Future work

Some known effector genes of *C. fulvum* are not present in the automated genome annotation. For future research on small secreted proteins it is therefore useful to include those proteins also in the analysis. Most gene prediction programs have difficulties predicting small open reading frames (Hanada *et al.*, 2007). Therefore an alternative approach should be applied in order to find more small secreted proteins in the genome. This study showed that standard sequence similarity searches are not sufficient for locating effector proteins in related species. In bacteria, effector proteins are shown to have altered nucleotide statistics as compared to all other genes due to horizontal gene transfer (Genin and Boucher, 2004). For example, several avirulence and virulence genes in *Pseudomonas syringae* are found to have a GC content significantly lower than that of other protein-coding genes in the genome and are located in proximity of each other and mobile elements in so called pathogenicity islands (Mansfield, 2009; Vivian *et al.*, 1999). The same observation is made for Type III Secretion System (TTSS) effectors in bacteria (Genin and Boucher, 2004) and in *U. maydis* 18.6% of all genes encoded by putatively secreted proteins are arranged in 12 gene clusters. The specific upregulation of seven of these gene clusters in tumour tissues indicated a possible function during pathogenic development (Kämper *et al.*, 2006). For the predicted secreted proteins in *C. fulvum* an expression analysis can also help to verify the predicted functions. An expression study in *M. grisea* showed that several putatively secreted cutinases are significantly up-regulated during infection-related development (Pan *et al.*, 2005).

In pathogenic fungi the effector genes evolve generally faster than average genes or are acquired more recently (Jiang, 2006), which suggests deviating nucleotide characteristics like in bacteria. Therefore, putative fungal effector proteins can be identified by analyzing all open reading frames (ORFs) in a genome for atypical GC content, codon usage and other relevant nucleotide statistics (Alfano, 2009; Juhas *et al.*, 2009). Also the grouping of genes in pathogenicity islands and the presence of nearby mobile elements or tRNAs can be taken into account.

Finally, wet-lab experiments can confirm the predictions made in this study. The set of 135 small secreted proteins forms a good starting point for studies focussed on novel effector proteins involved in pathogenesis. Within this set of proteins, the five predicted hydrophobins and three cutinases are of high interest.

Appendix A

Appendix

Table A.1: List of 135 putative effector proteins. The criteria are selected based on the characteristics of known effector proteins. Only putative secreted proteins smaller than 255aa and which contain more than 5 cysteines are selected.

Protein ID	Size	Number of cysteines	Interpro domain	Interpro description
Cf|2591019	59	6		
Cf|2602557	69	6		
Cf|2595246	77	8		
Cf|2594330	79	7		
Cf|2588287	82	9		
Cf|2589608	87	8	IPR008427; IPR014005	Extracellular membrane protein, 8-cysteine region, CFEM; Extracellular membrane protein, 8-cysteine region, fungi
Cf|2588228	89	10		
Cf|2592635	91	10		
Cf|2591520	91	6		
Cf|2587761	92	10		
Cf|2586666	93	6		
Cf|2590773	93	10		
Cf|2598071	95	7		
Cf|2589524	96	8		
Cf|2601692	98	10		
Cf|2601691	99	6		
Cf|2598649	99	7		
Cf|2597882	101	6		
Cf|2592380	102	8		

-continued on next page-

Table A.1 – continued from previous page

Protein ID	Size	Number of cysteines	Interpro domain	Interpro description
Cf|2588266	103	8	IPR001338	Fungal hydrophobin
Cf|2587582	104	6		
Cf|2596399	104	7		
Cf|2593666	105	8		
Cf|2590602	105	8		
Cf|2592466	105	8	IPR001338; IPR001338	Fungal hydrophobin; Fungal hydrophobin
Cf|2598764	105	7		
Cf|2587137	107	6		
Cf|2591325	110	9	IPR001338	Fungal hydrophobin
Cf|2591448	111	6		
Cf|2588532	112	8	IPR010636; IPR010636; IPR010636	Hydrophobin 2; Hydrophobin 2; Hydrophobin 2
Cf|2593992	114	6		
Cf|2595914	115	7		
Cf|2590484	119	8		
Cf|2595228	121	8	IPR001338	Fungal hydrophobin
Cf|2597260	121	6	IPR013032	EGF-like region, conserved site
Cf|2599225	122	8		
Cf|2593484	124	6		
Cf|2602036	126	6		
Cf|2602471	127	8		
Cf|2591576	127	8		
Cf|2591288	128	6		
Cf|2593270	130	8		
Cf|2590621	132	8		
Cf|2602786	132	12		
Cf|2591142	134	8		
Cf|2595141	135	11		
Cf|2589959	135	6		
Cf|2597781	135	8	IPR002557; IPR002557; IPR002557	Chitin binding protein, peritrophin-A; Chitin binding protein, peritrophin-A; Chitin binding protein, peritrophin-A
Cf|2589430	138	6		
Cf|2599617	139	6		

-continued on next page-

Table A.1 – continued from previous page

Protein ID	Size	Number of cysteines	Interpro domain	Interpro description
Cf|2602469	140	8		
Cf|2598868	140	6		
Cf|2588443	141	6		
Cf|2594470	146	7		
Cf|2592387	148	10		
Cf|2601039	149	6		
Cf|2599334	153	6		
Cf|2596932	154	7		
Cf|2601269	155	8		
Cf|2602667	155	7		
Cf|2596315	157	6		
Cf|2591917	158	8	IPR002889; IPR013994	Carbohydrate-binding WSC; Carbohydrate-binding WSC, subgroup
Cf|2593173	160	11		
Cf|2592005	162	8		
Cf|2591840	163	6		
Cf|2591090	163	8		
Cf|2594970	165	14		
Cf|2602579	165	8	IPR010636; IPR010636; IPR010636	Hydrophobin 2; Hydrophobin 2; Hydrophobin 2
Cf|2596972	165	8		
Cf|2596263	166	6		
Cf|2598661	167	9		
Cf|2590449	170	6		
Cf|2592097	171	10	IPR001969	Peptidase aspartic, active site
Cf|2598763	173	6		
Cf|2588513	175	10		
Cf|2594980	175	8		
Cf|2598660	176	8		
Cf|2594541	176	16		
Cf|2595946	177	13		
Cf|2598724	178	6		
Cf|2601847	179	6		

-continued on next page-

Table A.1 – continued from previous page

Protein ID	Size	Number of cysteines	Interpro domain	Interpro description
Cf|2589214	180	12		
Cf|2586674	184	8	IPR010636; IPR010636; IPR010636	Hydrophobin 2; Hydrophobin 2; Hydrophobin 2
Cf|2587023	184	8		
Cf|2587671	187	8		
Cf|2598092	187	7		
Cf|2595265	189	8		
Cf|2590833	190	6		
Cf|2597147	191	6		
Cf|2596658	195	8		
Cf|2601903	196	8		
Cf|2593024	197	7		
Cf|2592575	200	10		
Cf|2591384	201	6		
Cf|2599485	206	13		
Cf|2587567	206	12		
Cf|2596973	208	8		
Cf|2592947	208	6		
Cf|2597291	209	10		
Cf|2594195	209	7		
Cf|2601246	210	7		
Cf|2590753	210	9		
Cf|2602367	211	9		
Cf|2589888	216	6		
Cf|2592690	217	11		
Cf|2590916	220	7	IPR005132; IPR009009; IPR014734	Rare lipoprotein A; Barwin-related endoglucanase; Pollen allergen, N-terminal
Cf|2600764	220	6	IPR006863; IPR006863; IPR006863	Erv1/Alr; Erv1/Alr; Erv1/Alr
Cf|2599724	221	7		
Cf|2594577	222	6	IPR000873; IPR000873; noIPR; noIPR; noIPR	AMP-dependent synthetase and ligase; AMP-dependent synthetase and ligase; unintegrated; unintegrated; unintegrated
Cf|2592595	223	6	IPR000675; noIPR; noIPR	Cutinase; unintegrated; unintegrated

-continued on next page-

Table A.1 – continued from previous page

Protein ID	Size	Number of cysteines	Interpro domain	Interpro description
Cf|2599123	223	7	IPR000675; IPR011150; IPR011150; noIPR; noIPR	Cutinase; Cutinase, monofunctional; Cutinase, monofunctional; unintegrated; unintegrated
Cf|2589602	223	6	IPR002482; noIPR	Peptidoglycan-binding LysM; unintegrated
Cf|2594135	225	6	IPR001602; IPR001602; IPR001602; IPR001602	Protein of unknown function UPF0047; Protein of unknown function UPF0047; Protein of unknown function UPF0047; Protein of unknown function UPF0047
Cf|2598582	226	6		
Cf|2592304	228	8	IPR000675; IPR001412; noIPR; noIPR	Cutinase; Aminoacyl-tRNA synthetase, class I, conserved site; unintegrated; unintegrated
Cf|2592728	228	11		
Cf|2597786	228	8	IPR001338	Fungal hydrophobin
Cf|2601254	228	9	IPR002482; IPR002482; IPR002482; IPR002482; IPR002482; IPR002482; noIPR; noIPR; noIPR	Peptidoglycan-binding LysM; Peptidoglycan-binding LysM; Peptidoglycan-binding LysM; Peptidoglycan-binding LysM; Peptidoglycan-binding LysM; Peptidoglycan-binding LysM; unintegrated; unintegrated; unintegrated
Cf|2596116	229	6	IPR002818; noIPR; noIPR	ThiJ/PfpI; unintegrated; unintegrated
Cf|2593491	229	6		
Cf|2601732	233	6		
Cf|2601519	236	10		
Cf|2596063	238	8		
Cf|2601358	240	8		
Cf|2599383	241	9		
Cf|2598054	243	12		
Cf|2601278	244	8	IPR001223; IPR013781; IPR017853	Glycoside hydrolase, family 18, catalytic domain; Glycoside hydrolase, subgroup, catalytic core; Glycoside hydrolase, catalytic core
Cf|2593553	244	8		
Cf|2601014	245	6	IPR006094; IPR016166	FAD linked oxidase, N-terminal; FAD-binding, type 2
Cf|2592210	247	8		
Cf|2588947	248	6		
Cf|2600419	248	11		
Cf|2595104	249	6		

-continued on next page-

Table A.1 – continued from previous page

Protein ID	Size	Number of cysteines	Interpro domain	Interpro description
Cf|2591403	251	8		
Cf|2592533	253	9		

References

Alfano,J.R. (2009) Roadmap for future research on plant pathogen effectors. *Molecular Plant Pathology,* **10**, 805–813.

Baginsky,S., Baerenfaller,K., Grossmann,J., Grobei,M.A., Hull,R., Hirsch-Hoffmann,M., Yalovsky,S., Zimmermann,P., Grossniklaus,U. and Gruissem,W. (2008) Genome-scale proteomics reveals arabidopsis thaliana gene models and proteome dynamics. *Science (New York, N.Y.),* **320**, 938–941.

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

Bie,T.D., Cristianini,N., Demuth,J.P. and Hahn,M.W. (2006) Cafe: a computational tool for the study of gene family evolution. *Bioinformatics (Oxford, England),* **22**, 1269–1271.

Birch,P.R.J., Boevink,P.C., Gilroy,E.M., Hein,I., Pritchard,L. and Whisson,S.C. (2008) Oomycete rxlr effectors: delivery, functional redundancy and durable disease resistance. *Current Opinion in Plant Biology,* **11**, 373–379.

Bowen,J.K., Mesarich,C.H., Rees-George,J., Cui,W., Fitzgerald,A., Win,J., Plummer,K.M. and Templeton,M.D. (2009) Candidate effector gene identification in the ascomycete fungal phytopathogen venturia inaequalis by expressed sequence tag analysis. *Molecular Plant Pathology,* **10**, 431–448.

Cantarel,B.L., Coutinho,P.M., Rancurel,C., Bernard,T., Lombard,V. and Henrissat,B. (2009) The carbohydrate-active enzymes database (cazy): an expert resource for glycogenomics. *Nucleic Acids Research,* **37**, D233–238.

Catanzariti,A.M., Dodds,P.N. and Ellis,J.G. (2007) Avirulence proteins from haustoria-forming pathogens. *FEMS Microbiology Letters,* **269**, 181–188.

Chassot,C., Nawrath,C. and Métraux,J.P. (2007) Cuticular defects lead to full immunity to a major plant pathogen. *The Plant Journal: For Cell and Molecular Biology,* **49**, 972–980.

Chen,F., Mackey,A.J., Vermunt,J.K. and Roos,D.S. (2007) Assessing performance of orthology detection strategies applied to eukaryotic genomes. *PLoS ONE,* **2**, e383.

Conaway,R.C., Brower,C.S. and Conaway,J.W. (2002) Emerging roles of ubiquitin in transcription regulation. *Science (New York, N.Y.),* **296**, 1254–1258.

Conesa,A. and Götz,S. (2008) Blast2go: a comprehensive suite for functional analysis in plant genomics. *International Journal of Plant Genomics,* **2008**, 619832.

Corpet,F., Servant,F., Gouzy,J. and Kahn,D. (2000) Prodom and prodom-cg: tools for protein domain analysis and whole genome comparisons. *Nucleic Acids Research,* **28**, 267–269.

Cosgrove,D.J. (2005) Growth of the plant cell wall. *Nature Reviews. Molecular Cell Biology,* **6**, 850–861.

Dayhoff,M.O. (1976) The origin and evolution of protein superfamilies. *Federation Proceedings,* **35**, 2132–2138.

de Koster,C.G., de Wit,P.J.G.M., Joosten,M.H.A.J., Thomma,B.P.H.J., Bolton,M.D., van Esse,H.P., Vossen,J.H., de Jonge,R., Stergiopoulos,I., Stulemeijer,I.J.E., van den Berg,G.C.M., Borrás-Hidalgo,O. and Dekker,H.L. (2008) The novel cladosporium fulvum lysin motif effector ecp6 is a virulence factor with orthologues in other fungal species. *Molecular Microbiology,* **69**, 119–136.

Demuth,J.P., Bie,T.D., Stajich,J.E., Cristianini,N. and Hahn,M.W. (2006) The evolution of mammalian gene families. *PloS One,* **1**, e85.

Emanuelsson,O., Nielsen,H., Brunak,S. and von Heijne,G. (2000) Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. *Journal of Molecular Biology,* **300**, 1005–1016.

Enright,A.J., Dongen,S.V. and Ouzounis,C.A. (2002) An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research,* **30**, 1575–1584.

Enright,A.J. and Ouzounis,C.A. (2000) Generage: a robust algorithm for sequence clustering and domain detection. *Bioinformatics (Oxford, England),* **16**, 451–457.

Fiers,M.W.E.J., van der Burgt,A., Datema,E., de Groot,J.C.W. and van Ham,R.C.H.J. (2008) High-throughput bioinformatics with the cyrille2 pipeline system. *BMC Bioinformatics,* **9**, 96.

Furukawa,M., He,Y.J., Borchers,C. and Xiong,Y. (2003) Targeting of protein ubiquitination by btb-cullin 3-roc1 ubiquitin ligases. *Nature Cell Biology,* **5**, 1001–1007.

Galagan,J.E., Torriani,S.F.F., McDonald,B.A., Oliver,R.P., Hane,J.K., Lowe,R.G.T., Solomon,P.S., Tan,K.C., Schoch,C.L., Spatafora,J.W., Crous,P.W., Kodira,C. and Birren,B.W. (2007) Dothideomycete plant interactions illuminated by genome sequencing and est analysis of the wheat pathogen stagonospora nodorum. *The Plant Cell,* **19**, 3347–3368.

Genin,S. and Boucher,C. (2004) Lessons learned from the genome analysis of ralstonia solanacearum. *Annual Review of Phytopathology,* **42**, 107–134.

Geyer,R., Wee,S., Anderson,S., Yates,J. and Wolf,D.A. (2003) Btb/poz domain proteins are putative substrate adaptors for cullin 3 ubiquitin ligases. *Molecular Cell,* **12**, 783–790.

Hanada,K., Zhang,X., Borevitz,J.O., Li,W.H. and Shiu,S.H. (2007) A large number of novel coding small open reading frames in the intergenic regions of the arabidopsis thaliana genome are transcribed and/or under purifying selection. *Genome Research,* **17**, 632–640.

Howlett,B.J. (2006) Secondary metabolite toxins and nutrition of plant pathogenic fungi. *Current Opinion in Plant Biology,* **9**, 371–375.

Jenkins,J., Mayans,O. and Pickersgill,R. (1998) Structure and evolution of parallel beta-helix proteins. *Journal of Structural Biology,* **122**, 236–246.

Jiang,R. (2006). Footprints of evolution: the dynamics of effector genes in the phytophthora genome.

Jiang,R.H.Y., Tyler,B.M. and Govers,F. (2006) Comparative analysis of phytophthora genes encoding secreted proteins reveals conserved synteny and lineage-specific gene duplications and deletions. *Molecular Plant-Microbe Interactions: MPMI,* **19**, 1311–1321.

Joosten,M.H., Vogelsang,R., Cozijnsen,T.J., Verberne,M.C. and Wit,P.J.D. (1997) The biotrophic fungus cladosporium fulvum circumvents cf-4-mediated resistance by producing unstable avr4 elicitors. *The Plant Cell,* **9**, 367–379.

Juhas,M., van der Meer,J.R., Gaillard,M., Harding,R.M., Hood,D.W. and Crook,D.W. (2009) Genomic islands: tools of bacterial horizontal gene transfer and evolution. *FEMS Microbiology Reviews,* **33**, 376–393.

Kämper,J., Kahmann,R., olker,M.B., Ma,L.J., Brefort,T., Saville,B.J., Banuett,F., Kronstad,J.W., Gold,S.E., uller,O.M., Perlin,M.H., osten,H.A.B.W., de Vries,R., Ruiz-Herrera,J., na,C.G.R.P., Snetselaar,K., McCann,M., Pérez-Martín,J., ugge,M.F., Basse,C.W., Steinberg,G., Ibeas,J.I., Holloman,W., Guzman,P., Farman,M., Stajich,J.E., Sentandreu,R., González-Prieto,J.M., Kennell,J.C., Molina,L., Schirawski,J., Mendoza-Mendoza,A., Greilinger,D., unch,K.M., ossel,N.R., Scherer,M., Vranes,M., Ladendorf,O., Vincon,V., Fuchs,U., orn Sandrock,B., Meng,S., Ho,E.C.H., Cahill,M.J., Boyce,K.J., Klose,J., Klosterman,S.J., Deelstra,H.J., Ortiz-Castellanos,L., Li,W., Sanchez-Alonso,P., Schreier,P.H., auser Hahn,I.H., Vaupel,M., Koopmann,E., Friedrich,G., Voss,H., uter,T.S., Margolis,J., Platt,D., Swimmer,C., Gnirke,A., Chen,F., Vysotskaia,V., Mannhaupt,G., uldener,U.G., unsterk otter,M.M., Haase,D., Oesterheld,M., Mewes,H.W., Mauceli,E.W., DeCaprio,D., Wade,C.M., Butler,J., Young,S., Jaffe,D.B., Calvo,S., Nusbaum,C., Galagan,J. and Birren,B.W. (2006) Insights from the genome of the biotrophic fungal plant pathogen ustilago maydis. *Nature,* **444**, 97–101.

Kershaw,M.J. and Talbot,N.J. (1998) Hydrophobins and repellents: proteins with fundamental roles in fungal morphogenesis. *Fungal Genetics and Biology: FG & B,* **23**, 18–33.

Klee,E. and Ellis,L. (2005) Evaluating eukaryotic secreted protein prediction. *BMC Bioinformatics,* **6**, 256.

Kolattukudy,P.E. (1985) Enzymatic penetration of the plant cuticle by fungal pathogens. *Annual Review of Phytopathology,* **23**, 223–250.

Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *Journal of Molecular Biology,* **305**, 567–580.

Li,L., Stoeckert,C.J. and Roos,D.S. (2003) Orthomcl: identification of ortholog groups for eukaryotic genomes. *Genome Research,* **13**, 2178–2189.

Liu,H. and Teow,L.N. (2005) *Performance Evaluation of Protein Sequence Clustering Tools.*

Mansfield,J.W. (2009) From bacterial avirulence genes to effector functions via the hrp delivery system: an overview of 25 years of progress in our understanding of plant innate immunity. *Molecular Plant Pathology,* **10**, 721–734.

Martens,C., Vandepoele,K. and de Peer,Y.V. (2008) Whole-genome analysis reveals molecular innovations and evolutionary transitions in chromalveolate species. *Proceedings of the National Academy of Sciences,* **105**, 3427–3432.

Mayans,O., Scott,M., Connerton,I., Gravesen,T., Benen,J., Visser,J., Pickersgill,R. and Jenkins,J. (1997) Two crystal structures of pectin lyase a from aspergillus reveal a ph driven conformational change and striking divergence in the substrate-binding clefts of pectin and pectate lyases. *Structure (London, England: 1993),* **5**, 677–689.

Mueller,O., Kahmann,R., Aguilar,G., Trejo-Aguilar,B., Wu,A. and de Vries,R.P. (2008) The secretome of the maize pathogen ustilago maydis. *Fungal Genetics and Biology: FG & B,* **45 Suppl 1**, S63–70.

Nielsen,P.S., Clark,A.J., Oliver,R.P., Huber,M. and Spanu,P.D. (2001) Hcf-6, a novel class ii hydrophobin from cladosporium fulvum. *Microbiological Research,* **156**, 59–63.

Numa,H., Nishimura,M., Tanaka,T., Kanamori,H., Yang,C.C., Matsumoto,T., Nagamura,Y. and Itoh,T. (2009) Genome-wide validation of magnaporthe grisea gene structures based on transcription evidence. *FEBS Letters,* **583**, 797–800.

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

Persson,S., Caffall,K.H., Freshour,G., Hilley,M.T., Bauer,S., Poindexter,P., Hahn,M.G., Mohnen,D. and Somerville,C. (2007) The arabidopsis irregular xylem8 mutant is deficient in glucuronoxylan and homogalacturonan, which are essential for secondary cell wall integrity. *The Plant Cell,* **19**, 237–255.

Petroski,M.D. and Deshaies,R.J. (2005) Function and regulation of cullin-ring ubiquitin ligases. *Nature Reviews. Molecular Cell Biology,* **6**, 9–20.

Pierleoni,A., Martelli,P.L. and Casadio,R. (2008) Predgpi: a gpi-anchor predictor. *BMC Bioinformatics,* **9**, 392.

Rawlings,N.D., Morton,F.R., Kok,C.Y., Kong,J. and Barrett,A.J. (2008) Merops: the peptidase database. *Nucleic Acids Research,* **36**, D320–325.

Reddy,C.A. and D'Souza,T.M. (1994) Physiology and molecular biology of the lignin peroxidases of phanerochaete chrysosporium. *FEMS Microbiology Reviews,* **13**, 137–152.

Samuels,L., Kunst,L. and Jetter,R. (2008) Sealing plant surfaces: cuticular wax formation by epidermal cells. *Annual Review of Plant Biology,* **59**, 683–707.

Schoemaker,H.E. and Piontek,K. (1996) On the interaction of lignin peroxidase with lignin. *Pure and Applied Chemistry,* **68**, 2089–2096.

Schulze-Lefert,P. and Bieri,S. (2005) Plant sciences. recognition at a distance. *Science (New York, N.Y.),* **308**, 506–508.

Segers,G.C., Hamada,W., Oliver,R.P. and Spanu,P.D. (1999) Isolation and characterisation of five different hydrophobin-encoding cdnas from the fungal tomato pathogen cladosporium fulvum. *Molecular & General Genetics: MGG,* **261**, 644–652.

Semple,C.A.M. (2003) The comparative proteomics of ubiquitination in mouse. *Genome Research,* **13**, 1389–1394.

Skamnioti,P., Furlong,R.F. and Gurr,S.J. (2008) Evolutionary history of the ancient cutinase family in five filamentous ascomycetes reveals differential gene duplications and losses and in magnaporthe grisea shows evidence of sub- and neo-functionalization. *The New Phytologist,* **180**, 711–721.

Staples,R.C. (2001) A hydrophobin aids water-mediated dispersal of cladosporium conidia. *Trends in Plant Science,* **6**, 343–344.

Stergiopoulos,I. and de Wit,P.J.G.M. (2009) Fungal effector proteins. *Annual Review of Phytopathology,* **47**, 233–263.

Talbot,N.J., Soanes,D.M., Alam,I., Cornell,M., Wong,H.M., Hedeler,C., Paton,N.W., Rattray,M., Hubbard,S.J. and Oliver,S.G. (2008) Comparative genome analysis of filamentous fungi reveals gene family expansions associated with fungal pathogenesis. *PloS One,* **3**, e2300.

Ter-Hovhannisyan,V., Lomsadze,A., Chernoff,Y.O. and Borodovsky,M. (2008) Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Research,* **18**, 1979–1990.

Thomma,B.P.H.J., Esse,H.P.V., Crous,P.W. and Wit,P.J.G.M.D. (2005) Cladosporium fulvum (syn. passalora fulva), a highly specialized plant pathogen as a model for functional studies on plant pathogenic mycosphaerellaceae. *Molecular Plant Pathology,* **6**, 379–393.

van den Burg,H.A., Harrison,S.J., Joosten,M.H.A.J., Vervoort,J. and de Wit,P.J.G.M. (2006) Cladosporium fulvum avr4 protects fungal cell walls against hydrolysis by plant chitinases accumulating during infection. *Molecular Plant-Microbe Interactions: MPMI,* **19**, 1420–1430.

Vivian,A., Jackson,R.W., Athanassopoulos,E., Tsiamis,G., Mansfield,J.W., Sesma,A., Arnold,D.L., Gibbon,M.J., Murillo,J. and Taylor,J.D. (1999) Identification of a pathogenicity island, which contains genes for virulence and avirulence, on a large native plasmid in the bean pathogen pseudomonas syringae pathovar phaseolicola. *Proceedings of the National Academy of Sciences of the United States of America,* **96**, 10875–10880.

Whiteford,J.R. and Spanu,P.D. (2001) The hydrophobin hcf-1 of cladosporium fulvum is required for efficient water-mediated dispersal of conidia. *Fungal Genetics and Biology: FG & B,* **32**, 159–168.

Wright,J.C., Sugden,D., Francis-McIntyre,S., Riba-Garcia,I., Gaskell,S.J., Grigoriev,I.V., Baker,S.E., Beynon,R.J. and Hubbard,S.J. (2009) Exploiting proteomic data for genome annotation and gene model validation in aspergillus niger. *BMC Genomics,* **10**, 61.

Zhang,X., Li,Y. and Li,Y. (2009) Evaluating signal peptide prediction methods for gram-positive bacteria. *Biologia,* **64**, 655–659.