A genomic survey and prediction of the infection strategies employed by *Phoma medicaginis* var. *medicaginis*, a fungal pathogen of alfalfa and the model legume *Medicago truncatula*.

This thesis is presented for the degree of

Doctor of Philosophy

at Murdoch University, 2013

by

Angela. H. Williams, B. App. Sci. (Hons)

 Supervised by:
 Professor Richard P. Oliver
 Dr. Judith Lichtenzveig
 Professor Karam B. Singh
 Assoc. Prof Robert D. Trengove
Cover images

1) *M. truncatula* accession 3054, 10 d after spray inoculation with *P. medicaginis* var. *medicaginis* isolate PmedOMT5 in the glasshouse

2) GFP-expressing PmedOMT5 growing within the leaf of *M. truncatula* accession 3054, 5 dpi

3) Differential interference contrast microscopy of PmedOMT5 spores produced after 9 d growth on PDA

4) *M. truncatula* accession 3054 infiltrated with control media and PmedOMT5 5 week old culture filtrate.
Declaration

I declare that this thesis is my own account of my research and contains as its main content work which has not been previously submitted for a degree at any tertiary education institute.

________________________________________
Angela Helen Williams
Abstract

*Phoma medicaginis* is a fungal pathogen of the order Pleosporales that causes black spot of *Medicago sativa* (alfalfa, lucerne) and the related legume *Medicago truncatula*. This study presents a genomic analysis of the isolate OMT5 (PmedOMT5) and the pathogenicity mechanisms it employs during infection of *M. truncatula*. The genome assembly of ~31Mb covers an estimated 99% of the gene space and analysis outlines the location of repetitive and non-coding RNA, as well as the separately assembled and annotated complete mitochondrial genome. This is the first genome assembly from the genus *Phoma*, which encodes 11,879 manually-annotated protein-coding genes with biological support from proteogenomic and transcriptomic analysis.

Culture filtrate produced by PmedOMT5 grown in a variety of media, produced a chlorotic and necrotic response when infiltrated into plants from *Medicago* species and selected other plant species. The phytotoxic component was localised to a fraction containing molecules greater than 10 kDa, indicating that the active molecule is likely to have a proteinaceous component. The active compound or compounds have high thermal and chemical stability, are able to resist heat and pronase degradation and still produce a consistent response *in planta* after storage for 6 weeks at 4 °C on ice. Unexpectedly, the activity of compounds in the culture filtrate showed little correlation between sensitivity and susceptibility to foliar infection with *P. medicaginis*. Proteogenomic analysis of the PmedOMT5 culture filtrate yielded a list of bioinformatically predicted effector candidates for future testing.

Transcripts were sequenced via RNA-seq from four life-cycle stages of the pathogen: during *in vitro* growth, at vegetative (4d) and sporulating (16d) phases, during production of phytotoxic culture filtrate (4w) and during the early stages of plant invasion (1-5 dpi) on *M. truncatula* leaves. The variation in expression between the conditions was used in a preliminary investigation to identify genes that may play a role in pathogenicity, including prediction of effector candidates. Bioinformatic analysis suggests that PmedOMT5 is able to both detoxify and expel host phytoalexin compounds, which may contribute to its virulence. PmedOMT5 also produces a wide range of cell-wall-degrading enzymes with species-specific expansion of several pectin degrading families when compared to other Pleosporales pathogens.
Manuscripts in preparation incorporating research described in this thesis:

A genomic and transcriptomic analysis of the infection strategies employed by *Phoma medicaginis* var. *medicaginis*, a fungal pathogen of alfalfa and the model legume *Medicago truncatula*.

Angela H. Williams, James K. Hane, Karam B. Singh, Richard P. Oliver, Judith Lichtenzveig

CDSmapper: A pipeline for direct-to-genome mapping of peptide data for the validation of gene annotations and gene discovery

James K. Hane, Angela H. Williams, Francis Kessie, Ramisah Mohd Shah, Karam B. Singh, Judith Lichtenzveig, Richard P. Oliver
Acknowledgements

I would like to thank all my supervisors for their expert advice and assistance. Professor Oliver and A. Prof. Robert Trengove for getting me started, Professor Singh for helping me finish and Dr. Lichtenzveig for being there all the way through.

With special thanks to Dr James Hane for his patient training and never-ending willingness to discuss the possibilities of anything and everything.

I would also like to thank all the members of the Australian Centre for Necrotrophic Fungal Pathogens (ACNFP), the Murdoch Separation Science and Metabolomics Laboratory and the Molecular Plant Pathology and Crop Genomics Laboratory at CSIRO for all their help, encouragement and entertainment, both in and out of the lab.

I am indebted to the Grains Research and Development Corporation (GRDC) for providing me with a scholarship to undertake this research and offer thanks also to Gordon Thompson for his advice and assistance with microscopy, Rob Syme for his enthusiastic computational advice, Julie Lawrence for some fabulous DNA extraction and generally brightening my day and to my fellow genome explorers Francis and Ramisah, we got there in the end although at first the mountain appeared insurmountable.

Thanks also to Douglas Adams, Richard Feynman and Richard Dawkins for inspiring in me a thirst for knowledge about Life, the Universe and Everything and a fundamental scepticism of universally held beliefs. To M, S and C for the company and irreverence. To MS, Margaret and Kate for sharing many cups of tea, encouraging me to start and always believing that I would finish it one day, all is now good to go...

And finally to James without whose love, support and encouragement it is unlikely I would have actually finished and to Dana for inspiring me to be someone you could look up to.
# Table of Contents

Declaration ........................................................................................................................................... 1

Chapter 1 General introduction and literature review ................................................................. 1

1.1 The disease: Spring black stem and leaf spot and its causal agent *P. medicaginis* (Pm) ....................................................................................................................................... 3

1.1.1 Host range and plant organ specificity .............................................................................. 4
1.1.2 The Pathogen: *P. medicaginis* disease cycle ................................................................. 5
1.1.3 Culture morphology and morphological characteristics .............................................. 7
1.1.4 The genus *Phoma* ........................................................................................................ 8
1.1.5 Phylogenetic placement of *P. medicaginis* var. *medicaginis* Malbr. & Roum. 1886 ................................................................................................................................. 9
1.1.6 Plant-pathogenic fungi .................................................................................................... 10
1.1.7 Phylogenetic context: the order Pleosporales and its impact on agriculture ............ 11

1.2 The hosts: Legumes in Australian agriculture ........................................................................ 13

1.2.1 Australian *Medicago* pasture crops .......................................................................... 14
1.2.2 *Medicago sativa* ........................................................................................................ 16
1.2.3 *Medicago truncatula* .................................................................................................. 17
1.2.4 Methods of disease control in the field ......................................................................... 17

1.3 Plant defence against fungal pathogens ................................................................................. 18

1.3.1 Programmed cell death (PCD) ...................................................................................... 20
1.3.2 Reactive oxygen species (ROS) ................................................................................... 22
1.3.3 Phytoalexins and phytoanticipins - secondary metabolite plant defence compounds ................................................................................................................................. 23
1.3.4 Legume secondary metabolites - phenylproponoid derived isoflavonoids... .............. 24
1.3.5 Medicarpin ..................................................................................................................... 25
1.3.6 Response and resistance to *P. medicaginis* in *Medicago* species ......................... 27
1.3.7 Previous work on *P. medicaginis* and the involvement of isoflavonoids in resistance ................................................................................................................................. 29

1.4 Fungal genomics ..................................................................................................................... 30

1.4.1 Currently available fungal whole-genome sequences ............................................... 32

1.5 Aims and overview of this thesis ......................................................................................... 32

1.5.1 Thesis outline ................................................................................................................. 32

Chapter 2 General Materials and Methods .............................................................................. 35

2.1 General methods and procedures ......................................................................................... 36

2.1.1 General media used ....................................................................................................... 36
2.1.2 Seed sources .................................................................................................................. 36
<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1.3</td>
<td>Seed germination</td>
<td>37</td>
</tr>
<tr>
<td>2.1.4</td>
<td>Seed production and storage</td>
<td>37</td>
</tr>
<tr>
<td>2.1.5</td>
<td>Experimental plant conditions – Glasshouse</td>
<td>38</td>
</tr>
<tr>
<td>2.1.6</td>
<td>Experimental plant conditions – Growth chamber</td>
<td>38</td>
</tr>
<tr>
<td>2.1.7</td>
<td>Fungal isolates</td>
<td>38</td>
</tr>
<tr>
<td>2.1.8</td>
<td>Cultivation of fungal isolates for plant inoculation</td>
<td>39</td>
</tr>
<tr>
<td>2.1.9</td>
<td>Harvesting of conidiospores</td>
<td>39</td>
</tr>
<tr>
<td>2.1.10</td>
<td>Re-isolation of pathogenic fungi from infected plants</td>
<td>39</td>
</tr>
<tr>
<td>2.2</td>
<td>Infection methods</td>
<td>40</td>
</tr>
<tr>
<td>2.2.1</td>
<td><em>P. medicaginis</em> inoculation</td>
<td>40</td>
</tr>
<tr>
<td>2.2.2</td>
<td>Spray inoculation modified from (Kamphuis <em>et al.</em> 2008)</td>
<td>40</td>
</tr>
<tr>
<td>2.2.3</td>
<td>Spot inoculation as described in (Kamphuis <em>et al.</em> 2008)</td>
<td>41</td>
</tr>
<tr>
<td>2.2.4</td>
<td>Detached leaf assays</td>
<td>41</td>
</tr>
<tr>
<td>2.2.5</td>
<td>Scoring</td>
<td>41</td>
</tr>
<tr>
<td>2.2.6</td>
<td>Harvesting of plant tissue</td>
<td>41</td>
</tr>
<tr>
<td>2.3</td>
<td>Microscopy and staining for visualisation of fungal cells</td>
<td>42</td>
</tr>
<tr>
<td>2.3.1</td>
<td>Imaging</td>
<td>42</td>
</tr>
<tr>
<td>2.3.2</td>
<td>Staining for microscopy</td>
<td>42</td>
</tr>
<tr>
<td>2.4</td>
<td>DNA analysis</td>
<td>43</td>
</tr>
<tr>
<td>2.4.1</td>
<td>DNA extraction</td>
<td>43</td>
</tr>
<tr>
<td>2.4.2</td>
<td>PCR</td>
<td>44</td>
</tr>
<tr>
<td>2.4.3</td>
<td>PCR for products greater than 1 Kb</td>
<td>44</td>
</tr>
<tr>
<td>2.5</td>
<td>Bionformatics: general methods</td>
<td>44</td>
</tr>
<tr>
<td>2.5.1</td>
<td>Basic Local Alignment Search Tool</td>
<td>44</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Visualisation</td>
<td>45</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Statistical analysis</td>
<td>45</td>
</tr>
</tbody>
</table>

Chapter 3 Genome assembly and analysis of the genomic landscape ............................................. 47

<table>
<thead>
<tr>
<th>Subsection</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1.1</td>
<td>Short read sequencing and data analysis strategies</td>
<td>49</td>
</tr>
<tr>
<td>3.1.2</td>
<td>Features of fungal genome structure</td>
<td>52</td>
</tr>
<tr>
<td>3.1.3</td>
<td>Genome sequences available from the class Dothideomycetes order Pleosporales</td>
<td>54</td>
</tr>
<tr>
<td>3.1.4</td>
<td>Previously existing genomic resources for <em>P. medicaginis</em></td>
<td>54</td>
</tr>
<tr>
<td>3.1.5</td>
<td>Aims for this chapter: building a genomic resource for <em>P. medicaginis</em></td>
<td>55</td>
</tr>
<tr>
<td>3.2</td>
<td>Methods</td>
<td>55</td>
</tr>
<tr>
<td>3.2.1</td>
<td>DNA isolation</td>
<td>55</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Sequencing and assessment of read quality</td>
<td>55</td>
</tr>
</tbody>
</table>
3.2.3 Genome assembly ...................................................................................... 56
3.2.4 Gene prediction and genome characterisation ...................................... 56
3.2.5 Non-coding RNA analysis ....................................................................... 57
3.2.6 Analysis of repetitive DNA content of PmedOMT5 .............................. 57
3.2.7 RIP analysis of repetitive regions ............................................................. 58
3.3 Results ........................................................................................................... 58
3.3.1 Read data and quality ............................................................................. 58
3.3.2 Assembly and read coverage ................................................................. 59
3.3.3 Assessment of gene space coverage within the final assembly .............. 63
3.3.4 Characterisation of assembly and genomic landscape .......................... 63
3.3.5 De novo gene prediction of PmedOMT5 protein set version 1 ............. 63
3.3.6 Assessment of automated gene annotation via comparison of length to
best hit in the NCBI non-redundant protein database (nr) by BLASTP .......... 66
3.3.7 Identification of non-coding RNA (ncRNA) sequences ......................... 67
3.3.8 Repetitive DNA content of PmedOMT5 ................................................ 69
3.3.9 Analysis of Repeat Induced Point Mutation (RIP) in de novo predicted
repeats ............................................................................................................. 77
3.4 Discussion ..................................................................................................... 79
3.4.1 Genome assembly .................................................................................. 79
3.4.2 Assembly characteristics and genome structure .................................. 82
3.4.3 G:C content ........................................................................................... 82
3.4.4 Estimation of PmedOMT5 genome size ................................................ 82
3.4.5 PmedOMT5 gene space ......................................................................... 83
3.4.6 Gene prediction ..................................................................................... 84
3.4.7 Gene distribution ................................................................................... 86
3.4.8 Non coding RNA ................................................................................... 87
3.4.9 Transfer RNAs ....................................................................................... 88
3.4.10 Analysis of repetitive DNA content ..................................................... 88
3.4.11 Types of repetitive DNA in PmedOMT5 ............................................. 89
3.4.12 Evidence for RIP in PmedOMT5 .......................................................... 90
3.4.13 Future strategies for genome finishing ............................................... 91
3.5 Summary ...................................................................................................... 92
Chapter 4 The mitochondrial genome of PmedOMT5 ................................... 93
4.1 Introduction ................................................................................................. 94
4.1.1 Mitochondrial genome size and structure .......................................... 94
4.1.2 Aims for this chapter ........................................................................... 94
4.2 Methods ....................................................................................................... 95
4.3 Results .......................................................................................................................... 96
  4.3.1 Identification and analysis of PmedOMT5 mitochondrial scaffolds .......... 96
  4.3.2 Scaffold joining ........................................................................................................ 96
  4.3.3 Annotation of mitochondrial features ................................................................. 97
  4.3.4 Mapping of genomic and transcriptomic Illumina reads ......................... 100
4.4 Discussion .................................................................................................................. 103
  4.4.1 mtDNA size and structure ................................................................................... 103
  4.4.2 Mitochondrial gene content and translation .................................................. 104
  4.4.3 Mitochondrial introns, intronic endonucleases and rearrangement of gene
       order ......................................................................................................................... 107
  4.4.4 QoI fungicide resistance site ............................................................................ 108
4.5 Summary .................................................................................................................... 109
Chapter 5 Biological support for predicted gene models .............................................. 111
  5.1 Introduction .............................................................................................................. 112
     5.1.1 Gene prediction ................................................................................................. 112
     5.1.2 Proteogenomics ................................................................................................. 113
     5.1.3 RNA-seq ........................................................................................................... 114
     5.1.4 Annotation pipeline ........................................................................................ 115
     5.1.5 Aims for this chapter ....................................................................................... 116
5.2 Materials and methods .............................................................................................. 116
  5.2.1 Proteogenomic analysis ....................................................................................... 116
  5.2.2 Intracellular protein extraction .......................................................................... 116
  5.2.3 Culture filtrate/ extracellular protein extraction ............................................... 117
  5.2.4 Mannitol dehydrogenase (MDH) assay, to check activity in intracellular
       and extracellular fractions .................................................................................. 117
  5.2.5 MS-MS sequencing of intra- and extra-cellular fractions .............................. 117
  5.2.6 Assessment of support for predicted genes via mapping of peptide
       fragments to the annotated gene models and 6-frame translation of PmedOMT5. ...
       ................................................................................................................................. 118
  5.2.7 Transcriptome analysis ....................................................................................... 119
  5.2.8 In vitro fungal tissue used for RNA production .............................................. 119
  5.2.9 Preparation of infected plant material for RNA-seq ....................................... 119
  5.2.10 RNA isolation ..................................................................................................... 120
  5.2.11 Sequencing ......................................................................................................... 121
  5.2.12 Sequence analysis and trimming ...................................................................... 121
  5.2.13 Transcript read alignment ................................................................................ 122
  5.2.14 De novo assembly of transcripts ....................................................................... 122
  5.2.15 Phoma ESTs ........................................................................................................ 123
5.2.16 Combination of evidence for manual gene annotation ........................................... 123
5.2.17 Re-alignment of reads to new gene models and calculation of expression differences .................................................................................................................. 124
5.3 Results .......................................................................................................................... 125
  5.3.1 ESTs ....................................................................................................................... 125
  5.3.2 Proteogenomic gene model support .......................................................................... 126
  5.3.3 Transcriptome libraries .......................................................................................... 128
  5.3.4 Sequence analysis and trimming ............................................................................. 129
  5.3.5 RNAseq read alignment to the genome assembly ................................................... 130
  5.3.6 Differential expression ......................................................................................... 131
  5.3.7 De novo assembled transcriptome .......................................................................... 132
5.4 Discussion .................................................................................................................... 133
  5.4.1 Manual annotation ............................................................................................... 133
  5.4.2 Homology ............................................................................................................. 136
  5.4.3 Proteogenomic analysis ......................................................................................... 136
  5.4.4 Transcriptomic analysis ......................................................................................... 139
  5.4.5 De novo assembled transcriptome .......................................................................... 140
  5.4.6 In planta (IP) expression ....................................................................................... 141
  5.4.7 Summary ............................................................................................................... 142

Chapter 6 Predicted functional annotation of PmedOMT5 proteins, focusing on those potentially involved in pathogenicity .................................................................................. 143
  6.1 Introduction: .............................................................................................................. 144
  6.2 Materials and methods ............................................................................................. 148
    6.2.1 Analysis of mating type genes ............................................................................. 148
    6.2.2 Functional annotation ......................................................................................... 148
    6.2.3 Conserved domain analysis ............................................................................... 149
    6.2.4 Gene Ontology ................................................................................................... 149
    6.2.5 Prediction of cellular localisation and secretion ................................................ 149
    6.2.6 Carbohydrate active enzyme (CAZY) analysis ............................................... 149
    6.2.7 Secondary metabolite synthesis gene prediction ................................................. 149
    6.2.8 Statistical examination for over- or under-representation of protein functional attributes .................................................................................................................. 150
    6.2.9 Comparison to PHIbase proteins with experimentally defined pathogenicity profiles .................................................................................................................. 150
    6.2.10 Analysis of known effector motifs ..................................................................... 150
    6.2.11 Orthology ........................................................................................................... 151
    6.2.12 Score based assessment of potential proteinaceous host-specific effectors .. .............................................................................................................................. 152
6.3 Results ................................................................................................................................. 152
  6.3.1 PmedOMT5 mating type ............................................................................................... 152
  6.3.2 Fungicide target genes in PmedOMT5 ........................................................................ 153
  6.3.3 Putative Functional annotation .................................................................................. 154
  6.3.4 Secreted proteins ......................................................................................................... 154
  6.3.5 In planta (IP) expression .............................................................................................. 156
  6.3.6 Genes that show increased relative expression IP ....................................................... 157
  6.3.7 Carbohydrate-active enzymes (CAZymes) ................................................................. 157
  6.3.8 ESTs ............................................................................................................................ 161
  6.3.9 PmedOMT5 genes with homology in the Plant Host Interaction database (PHIbase) .......................................................................................................................... 162
  6.3.10 Phytoalexin detoxification ......................................................................................... 164
  6.3.11 Expression of genes with Pfam domains known to involved in virulence in other pathogens ............................................................................................................. 167
  6.3.12 NPP1 domain proteins - may induce host cell death ................................................. 167
  6.3.13 LysM domain proteins - may play a role in binding extracellular chitin 167
  6.3.14 CFEM domain proteins ............................................................................................ 168
  6.3.15 CAP domain proteins ............................................................................................... 169
  6.3.16 Secondary metabolite production genes ................................................................. 169
  6.3.17 Over representation of functional annotations in genes detected in planta (IP) ........ 173
  6.3.18 Bioinformatic prediction of candidate effector genes ............................................... 181
  6.3.19 Functional annotations associated with the potential effector candidate set. .......... 185

6.4 Discussion .......................................................................................................................... 187
  6.4.1 Mating type of PmedOMT5 ......................................................................................... 187
  6.4.2 Fungicide resistance .................................................................................................... 189
  6.4.3 Predicted functional annotation of PmedOMT5 genes ............................................. 190
  6.4.4 Secreted proteins ........................................................................................................ 190
  6.4.5 Orthology of PmedOMT5 genes ................................................................................ 192
  6.4.6 Prediction of PmedOMT5 genes with potential roles in pathogenicity ....................... 192
  6.4.7 Membrane transport proteins ..................................................................................... 193
  6.4.8 Fungal growth IP ........................................................................................................ 193
  6.4.9 Secondary metabolite biosynthesis genes ................................................................... 194
  6.4.10 Carbohydrate-active enzymes devoted to plant cell wall depolymerization. ............ 196
  6.4.11 In planta co-expressed gene clusters ....................................................................... 199
  6.4.12 Phytoalexin detoxification genes .............................................................................. 200
6.4.13 NEP1-like (NLP) family genes ................................................................. 204
6.4.14 Effector associated motifs ...................................................................... 205
6.4.15 Prediction of proteinaceous effector candidates ...................................... 206
6.5 Summary ....................................................................................................... 208
Chapter 7 Examining the secretome of PmedOMT5 ......................................... 211
7.1 Introduction: ............................................................................................... 212
7.1.1 Aims for this chapter .............................................................................. 214
7.2 Materials and Methods ............................................................................. 214
7.2.1 Optimisation of phytotoxic culture filtrate production .............................. 214
7.2.2 Bioassay for activity of *P. medicaginis* culture filtrate on *M. truncatula*.. 215
7.2.3 Heat and pronase treatment of culture filtrate ........................................... 216
7.2.4 Infiltration of buffers and controls to test for response in *M. truncatula* 217
7.2.5 PmedOMT5 culture-filtrate host specificity ........................................... 217
7.2.6 Fractionation of the culture filtrate ......................................................... 217
7.2.7 Ion exchange (IEX) fractionation of culture filtrate ............................... 217
7.2.8 Separation of extracellular proteins via SDS-PAGE .............................. 218
7.2.9 Statistical examination for over and under-representation of protein functional attributes in the extracellular fraction .......................................... 218
7.3 Results ......................................................................................................... 218
7.3.1 Optimisation of media for active culture filtrate production .................... 218
7.3.2 Response to brefeldin A infiltration ......................................................... 220
7.3.3 Reaction of *M. truncatula* to infiltrated *P. medicaginis* culture filtrate.. 220
7.3.4 Activity of culture filtrate from three *P. medicaginis* isolates................... 221
7.3.5 Investigating the stability and host-specificity of compounds produced by PmedOMT5 cultured in F2NY media ....................................................... 222
7.3.6 Response of *M. truncatula* accessions to infiltration with PmedOMT5 culture filtrate ................................................................. 223
7.3.7 Evaluating the host specificity of PmedOMT5 culture filtrate ............... 224
7.3.8 Fractionation of culture filtrate by size and ion exchange chromatography. ........................................................................................................ 226
7.3.9 Proteins detected in the culture filtrate via IEX-LC MALDI TOF .......... 226
7.3.10 Analysis of over-representation of functional domains and attributes in the extracellular fraction ................................................................. 231
7.3.11 PHIbase matches to proteins identified in the extracellular fraction ....... 234
7.4 Discussion .................................................................................................... 235
7.4.1 Optimisation of culture filtrate production .............................................. 235
7.4.2 Activity of culture filtrate on *M. truncatula* ......................................... 236
7.4.3 Host specificity of culture filtrate ..................................................237
7.4.4 Fractionation of culture filtrate and secondary metabolites .............238
7.4.5 Thermal stability and proteinase resistance of culture filtrate ..........240
7.4.6 Autofluorescence in culture filtrate infiltrated leaves .....................241
7.4.7 Cerato-platanin ..............................................................................242
7.4.8 Functional annotations over-represented in the extracellular fraction ....242
7.4.9 Determining the nature of potential effectors in PmedOMT5 ............243
7.5 Summary ..........................................................................................245

Chapter 8 General Discussion ..................................................................247
8.1 The pathogen P. medicaginis ..........................................................248
8.2 Genome and assembly........................................................................248
8.2.1 Genome size ................................................................................249
8.3 Gene content of PmedOMT5 .........................................................251
8.4 Mating and sexuality .........................................................................252
8.5 Repetitive DNA and RIP in PmedOMT5 ..........................................253
8.6 The role of overcoming plant secondary metabolites in pathogenicity of Phoma on legumes and legume defence ...........................................254
8.7 Phoma infection mechanisms ............................................................255
8.8 Secreted proteins and effectors .........................................................256
8.9 Summary and future work .................................................................257

Chapter 9 References .............................................................................259
List of figures

Figure 1.1: Growth stages and infection cycle of *P. medicaginis* var. *medicaginis* isolate PmedOMT5................................................................................................................................. 6
Figure 1.2: Crystals produced by *P. medicaginis*. ......................................................... 7
Figure 1.3: Phylogenetic placement of *P. medicaginis*. ................................................ 12
Figure 1.4: Taxonomic relationships within the two major clades of crop legumes (galegoid and phaseoloid). ........................................................................................................... 15
Figure 1.5: *Medicago* pastures, crops and growth regions in Australia......................... 16
Figure 3.1: Variation in N50 and L50 (N50 length) with changing k-mer length of Velvet assemblies .......................................................................................................................... 59
Figure 3.2: Number of predicted genes per scaffold relative to the length of the scaffold. .................................................................................................................................................. 65
Figure 3.3: Percentage completeness of coverage of the top hit by bit score of BLASTP (2.2.25) comparison to NCBI non-redundant protein database .............................................. 66
Figure 3.4: Schematic of the pipeline used for description and annotation of the repetitive DNA content of the PmedOMT5 assembly .................................................................................. 69
Figure 3.5: Proximity by class of repetitive DNA of PmedOMT5 genome assembly to ends of scaffolds ............................................................................................................................. 71
Figure 3.6: RIPCAL analysis of R62 (350 bp) showing high CpA ↔ TpA changes (red) relative to other CpN mutations in a short repeated fragment with similarity to the Molly transposon ........................................................................................................................................ 77
Figure 3.7: Consensus alignment of top ten matches from NCBI protein database to PmedOMT5_02376 ....................................................................................................................................... 79
Figure 3.8: Number of genes with varying intron numbers across the published Dothideomycete species .......................................................................................................................... 85
Figure 4.1: Gel showing amplification between predicted adjacent SSPACE scaffolds but not others ............................................................................................................................................... 99
Figure 4.2: Representative figure of the complete mitochondrial genome of PmedOMT5 and comparison to that of *P. nodorum* SN15 .................................................................................................................. 102
Figure 5.1: Proteogenomics samples (a) SDS PAGE separation of proteins in samples sent for peptide sequencing .......................................................................................................................... 126
Figure 5.2: Summary of overlap between genes with peptide support in intracellular and extracellular proteomic fractions ............................................................................................................ 127
Figure 5.3: *In vitro* RNA-seq growth conditions .................................................. 128
Figure 5.4: *In planta* RNA-seq experimental progress ............................................. 129
Figure 5.5: FastQC analysis of quality scores across the read length 16d lane 1, read 2, raw reads. .............................................................................................................. 129
Figure 5.6: Read coverage over gene length for all genes (a) *in vitro* (b) *in planta*. .... 131
Figure 5.7: Apollo Genome Annotation Curation Tool view depicting a GeneMark predicted v1 gene model that was corrected in version 2. ........................................... 135
Figure 6.1: Amplification of MAT1-2-1 fragment in reference cultures and West Australian *Phoma* spp. isolates. ........................................................................................................... 153
Figure 6.2: Protein localisation ...................................................................................... 155
Figure 6.3: Proportional Venn diagram summarising the overlap between characteristics of the total 5,778 *in planta* expressed genes. ............................................................. 157
Figure 7.1: Scoring range used to assess response to infiltrated culture filtrate. ........ 216
Figure 7.2: *M. truncatula* SA3054 infiltrated with culture filtrate from 6 week PmedOMT5 in Fries 2 .................................................................................................................. 219
Figure 7.3: *M. truncatula* leaves 12 days post infiltration (dpi) ................................. 221
Figure 7.4: Infiltrated leaves produce an autofluorescence reaction similar to that seen when leaves are inoculated with PmedOMT5 spores. ........................................... 222
Figure 7.5: Responses of non-*Medicago* species to PmedOMT5 culture filtrate. ...... 226
Figure 7.6: SDS-PAGE separation of PmedOMT5 culture filtrate............................... 227
Figure 7.7: Functional attributes associated with the 170 proteins identified by peptide matches from the extracellular protein fraction. ......................................................... 227
List of tables

Table 2.1: Non-Medicago seed sources ........................................................................................................... 37
Table 2.2: Original identification, host and location isolated for fungal pathogens used in this study ..................................................................................................................................................... 40
Table 3.1: First stage of genome assembly-optimisation of k-mer length for assembly with Velvet 1.0.14 .......................................................................................................................................................... 60
Table 3.2: Optimisation of Velvet assembly parameters “coverage cut-off” and “expected coverage”, for assemblies created using k-mer length 39 .................................................................................................. 61
Table 3.3: PmedOMT5 final assembly characteristics ....................................................................................... 62
Table 3.4: Number and completeness of 248 ultra-conserved core eukaryote genes (CEGs) identified within the PmedOMT5 final assembly .............................................................................................................. 63
Table 3.5: Predicted gene model statistics (version 1) ........................................................................................... 64
Table 3.6: Number of members in ncRNA families identified on the PmedOMT5 nuclear scaffolds determined by Infernal vs. RFAM .......................................................................................................... 68
Table 3.7: Repetitive DNA classes identified by RepeatMasker in the PmedOMT5 assembly using the de novo repeat library ......................................................................................................................................... 70
Table 3.8: RIP and homology analysis of repetitive DNA classes identified by RepeatMasker in the PmedOMT5 assembly using the de novo repeat library ............................................................................ 72
Table 3.9: Identification of repetitive regions in the PmedOMT5 assembly by similarity to known fungal repeats (Repbase) ....................................................................................................................................................... 75
Table 3.10: RepeatMasker analysis of PmedOMT5 assembly regions with similarity to repeats identified in the genome of P. nodorum SN15 .................................................................................................... 76
Table 3.11: BLAST alignments to N. crassa rid protein (845 aa) [GenBank: AAM27408.1] in Dothideomycete genome assemblies ................................................................................................................................. 78
Table 3.12: Assembly statistics for final PmedOMT5 assembly ........................................................................ 81
Table 3.13: Assembly statistics for selected published Dothideomycete genomes ............................................ 87
Table 3.14: Number and classification of non-coding RNAs (excluding tRNAs) identified in the PmedOMT5 genome .............................................................................................................................................. 88
Table 4.1: Primer sequences used to amplify mitochondrial scaffolds .......................................................... 96
Table 4.2: PmedOMT5 scaffolds predicted to form part of the mitochondrial genome prior to SSPACE scaffolding and PCR joining ................................................................................................................ 98
Table 4.3: Mitochondrial scaffolds joined by SSPACE ....................................................................................... 99
Table 4.4: Amplifying primer pairs and their distance from the scaffold ends ................................................ 99
Table 4.5: Features of the final PmedOMT5 mitochondrial genome sequence........100
Table 4.6: Number of reads aligned to the P. nodorum SN15 mtDNA and PmedOMT5 DNA.................................................................101
Table 4.7: Comparison of P. medicaginis mitochondrial scaffold with published Ascomycete mitochondrial genomes. ..........................................................104
Table 4.8: Comparison of annotated fungal mitochondrial gene content with that of the final assembled PmedOMT5 mitochondrial genome.................................106
Table 4.9: Order of tRNAs in cluster flanking the large ribosomal subunit (rnl) in sequenced Dothideomycete mtDNA ...............................................................108
Table 5.1: Summary of PmedOMT5 version 2 protein-coding gene annotations. ......125
Table 5.2: Comparison of intra-cellular and extra-cellular protein extracts sent for sequencing ........................................................................................................126
Table 5.3: Number of proteogenomic peptide matches supporting the PmedOMT5 version 2 protein set. ......................................................................................127
Table 5.4: Number of RNA-seq reads obtained, the percentage remaining after adapter trimming and quality filtering and number that aligned to the PmedOMT5 genome..130
Table 5.5: Number of genes expressed in each sample by Cufflinks normalised FPKM. ......................................................................................................................131
Table 5.6: Number of de novo assembled Trinity transcripts. ................................132
Table 5.7: Summary of gene support for OMT5 predicted genes..........................133
Table 6.1: List of effector associated motifs searched for and relevant references....151
Table 6.2: Summary of orthology and functional annotations of PmedOMT5 proteins. ......................................................................................................................154
Table 6.3: Total number of each CAZyme family type identified in PmedOMT5......159
Table 6.4: Number of proteins with CAZy domains assigned by PFAM analysis in all PmedOMT5 version 2 proteins and those for which the encoding gene was detected as expressed in planta by alignment of five or more reads. ...........................................160
Table 6.5: Selected expanded CAZyme families in PmedOMT5 potentially relevant to pathogenicity. ..........................................................................................161
Table 6.6: Orthology and expression of genes encoding proteins with a match to PHI:112 maackiain detoxification gene in PHIbase that are expressed in planta .......164
Table 6.7: PmedOMT5 proteins with BLASTP matches to pisatin demethylase proteins in the NCBI protein database. .................................................................166
Table 6.8: Count of genes in PmedOMT5 assembly associated with secondary metabolite production. ..................................................................................................................................................170
Table 6.9: All secondary metabolite production genes identified in PmedOMT5 and their top match in PHIbase .............................................................................................................................................................................171
Table 6.10: Pfam domains over-represented in genes detected as expressed in planta relative to all genes. .............................................................................................................................................................................174
Table 6.11: Pfam domains in genes detected as showing more than a 4-fold increase in expression in planta relative to the other conditions tested genes versus all other genes. .............................................................................................................................................................................175
Table 6.12: Gene Ontologies for all proteins over-represented in the expressed in planta (by five or more reads) set. .............................................................................................................................................................................177
Table 6.13: Gene Ontologies over-represented in proteins encoded by genes that show greater than 4-fold increased relative expression in planta. ........................................................................179
Table 6.14: Scores assigned to various features potentially associated with necrotrophic effectors and counts of genes / proteins with these features. .............................................................................................................................................................................182
Table 6.15: The top effector candidates of P. medicaginis OMT5, based on cumulative scoring of associated gene/protein properties and functional annotations. .................................183
Table 6.16: Top unique (species-specific by orthology) PmedOMT5 effector candidates. .............................................................................................................................................................................184
Table 6.17: Number of over-represented GO terms in the top 111 candidates. ........186
Table 6.18: Number of over-represented Pfam domains in the top 111 candidates. ....187
Table 6.19: Number of PKS-, NRPS-, and Cytochrome P450–encoding genes in the genomes of selected Fungi and Oomycetes compared to PmedOMT5. .................................196
Table 6.20: Comparison of pectate lyase families amongst the Pleosporales. ........198
Table 6.21: Comparison between the total CAZyme family complement of PmedOMT5 with other Pleosporales plant pathogens. .............................................................................................................199
Table 7.1: Mean sensitivity score of infiltrated PmedOMT5 culture filtrate. .............220
Table 7.2: Response to infiltration in M. truncatula accessions SA3054 (S) and SA27063 (R) across two experiments. .............................................................................................................................................................................223
Table 7.3: Mean response score of M. truncatula accessions infiltrated with whole culture filtrate. .............................................................................................................................................................................224
Table 7.4: Sensitivity of various plant species to infiltration with PmedOMT5 culture filtrate. .............................................................................................................................................................................225
Table 7.5: Summary of functional attributes of proteins detected in the extracellular protein fraction.

.................................

.................................

.................................

.................................

.................................................................................................................228

Table 7.6: Top effector candidates identified in the extracellular fraction.

.................230

Table 7.7: Categories of genes over-represented in the extracellular fraction.

.................231

Table 7.8: Pfam domains significantly over-or under-represented in the extracellular protein fraction versus the total protein set.

.................................................................................................................232

Table 7.9: Gene Ontologies (GOs) significantly over- or under-represented in the extracellular fraction versus the total protein set.

.................................................................................................................233
<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>aa</td>
<td>amino acid/s</td>
</tr>
<tr>
<td>AS</td>
<td>acetosyringone</td>
</tr>
<tr>
<td>BAC</td>
<td>bacterial artificial chromosome(s)</td>
</tr>
<tr>
<td>bp</td>
<td>base pairs</td>
</tr>
<tr>
<td>CAZymes</td>
<td>carbohydrate active enzymes</td>
</tr>
<tr>
<td>CDC</td>
<td>conditionally dispensable chromosome</td>
</tr>
<tr>
<td>CDS</td>
<td>coding sequence (region of nucleotides that corresponds to sequence of aa’s in a protein –location includes start and stop codon)</td>
</tr>
<tr>
<td>cDNA</td>
<td>complementary DNA</td>
</tr>
<tr>
<td>CTAB</td>
<td>cetyltrimethylammonium bromide/hexadecyltrimethylammonium bromide</td>
</tr>
<tr>
<td>d</td>
<td>days</td>
</tr>
<tr>
<td>DEPC</td>
<td>diethylpyrocarbonate</td>
</tr>
<tr>
<td>DLA</td>
<td>detached lead assay</td>
</tr>
<tr>
<td>dpi</td>
<td>days post inoculation/infiltration</td>
</tr>
<tr>
<td>EST</td>
<td>expressed sequence tag(s)</td>
</tr>
<tr>
<td>FPKM</td>
<td>Fragments Per Kilobase of transcript per Million mapped reads</td>
</tr>
<tr>
<td>Gb</td>
<td>gigabase pairs</td>
</tr>
<tr>
<td>h</td>
<td>hours</td>
</tr>
<tr>
<td>HGT/LGT</td>
<td>horizontal gene transfer/ lateral gene transfer</td>
</tr>
<tr>
<td>HPLC-DAD</td>
<td>high performance liquid chromatography- diode array detection (UV)</td>
</tr>
<tr>
<td>HSP</td>
<td>high scoring segment pair</td>
</tr>
<tr>
<td>IEX</td>
<td>ion exchange chromatography</td>
</tr>
<tr>
<td>IP</td>
<td>in planta</td>
</tr>
<tr>
<td>ITS</td>
<td>internal transcribed spacer</td>
</tr>
<tr>
<td>kb</td>
<td>kilobase pairs</td>
</tr>
<tr>
<td>LG</td>
<td>linkage group</td>
</tr>
<tr>
<td>LOD</td>
<td>logarithm of the odds</td>
</tr>
<tr>
<td>Mb</td>
<td>mega base pairs</td>
</tr>
<tr>
<td>MFS</td>
<td>Major Facilitator Superfamily</td>
</tr>
<tr>
<td>min</td>
<td>minutes</td>
</tr>
<tr>
<td>MDH</td>
<td>Mannitol dehydrogenase</td>
</tr>
<tr>
<td>MS/MS</td>
<td>multidimensional (2D) mass spectrometry</td>
</tr>
<tr>
<td>MWCO</td>
<td>molecular weight cut-off</td>
</tr>
<tr>
<td>NB-LRR</td>
<td>nucleotide binding site leucine rich repeat</td>
</tr>
</tbody>
</table>
ncRNA  non-coding RNA
NGS     next generation sequencing
NRPS    non-ribosomal peptide synthase
O/N     overnight
OD      optical density
ORF     open reading-frame
PAMP    pathogen-associated molecular pattern
PCD     programmed cell death
PDA     potato dextrose agar
PCR     polymerase chain reaction
PE      paired end
PKS     polyketide synthase
qRT-PCR quantitative real-time polymerase chain reaction
RIP     repeat-induced point mutation
RNA-seq RNA/cDNA sequencing
ROS     reactive oxygen species
rpm     revolutions per minute
rRNA    ribosomal RNA
RT      room temperature
s       seconds
SBSLS   Spring black stem and leaf spot
SDS-PAGE sodium dodecyl sulfate polyacrylamide gel electrophoresis
SNP     single nucleotide polymorphism
sp/mL   spores per millilitre
Syn     synonym
TE      transposable element(s)
TE (buffer) Tris-Cl/EDTA
tRNA    transfer RNA
U       enzyme unit(s)
UTR     untranslated region
w       week(s)
WG(S)S  whole genome (shotgun) sequencing
Supplementary data

Appendices to Chapter 2:
2.1 Media recipes
2.2 GFP-transformation protocol

Appendices to Chapter 3:
3.1 List of nuclear scaffolds their length, GC content, coverage and predicted gene number
3.2 GeneMark predicted version 1 genes
3.3 List of nuclear tRNAs and their co-ordinates
3.4 De novo predicted repeats and the location of matches within the genome
3.5 Repeat matches to SN15
3.6 Repeat matches from REPbase fungal repeats 15.6

Appendices to Chapter 4:
4.1 Mitochondrial gff
4.2 Mitochondrial tRNAs and anti-codons

Appendices to Chapter 5:
5.1 List of all data sets used for comparative analysis
5.2 Primer sets used
5.3 All version 2 PmedOMT5 genes, length, annotation support and functional annotation

Appendices to Chapter 6:
6.1 Relative CAZy domain numbers for each family compared to those identified in other Dothideomycetes analysed by Ohm et al., 2012
6.2 Pfam domains over- or under- represented in the IP expressed gene set
6.3 Pfam domains over- or under- represented in proteins encoded by genes with a relative increase in expression IP
6.4 Over- or under-represented GO terms in genes expressed IP
6.5 Over or under- represented GO terms in proteins encoded by genes with a relative increase in expression IP
6.6 Over- or under- represented GO terms in the top 111 effector candidates
6.7 Over- or under- represented Pfams in the top 111 effector candidates
6.8 CAZymes detected as expressed in planta
6.9 FPKMs of house-keeping genes
Chapter 1
General introduction
and literature review
Given the myriad of potential fungal and bacterial pathogens that plants are continuously exposed to, from the air and below the ground, scientists have asked the question “why the world is still green?” (Boller and He 2009). The answer lies in the evolutionary arms race between plant pathogens and their hosts (Jones and Dangl 2006), where the majority of battles are being won by the plant, although there are many skirmishes that are lost, and the casualties and death toll are large. As there is no evolutionary benefit to a pathogen in extinguishing its sole food source, in nature, plant species are rarely completely killed by their pathogens. However domesticated crops, such as chickpea, can have 100% yield loss under environmental conditions that favour the pathogen (Ozkilinc et al. 2011 and references therein). In this current age of agricultural monocultures there are substantial losses in both yield and nutritional value in grain and forage crops, with plant diseases accounting for the loss of 10-30% of the global harvest each year (Strange and Scott 2005). What decides the outcome of each battle occurs mostly at the molecular level, via the interaction between proteins and metabolites produced by both plants and pathogenic fungi, that has been eloquently described as ‘secretory warfare’ (Huckelhoven and Panstruga 2011).

In recent years, due to declining sequencing costs and increases in throughput, many studies have used whole-genome sequencing to examine the language of the interactions between fungal and oomycete pathogens and their plant hosts. Following sequencing, genome assembly and the prediction of genes and their products, it is possible to identify conserved domains in genes that can be broadly classified via sequence homology to other sequenced species. Comparison of functional annotations between species allows the prediction of common functions and identification of gene families that may play roles relevant to plant pathogenicity. On the fungal side this includes genes encoding enzymes that target and degrade the plant cell walls, as well as small secreted molecules collectively termed “effectors”. Effectors interact with the host cell, altering its structure or function which results in either infection by the fungus or triggering of the host plant’s defences (Kamoun 2006). Whilst on the plant side, the defensive armoury typically contains innate anti-fungal proteins and metabolites, as well as molecules produced in response to detection of pathogen invasion. Genetically-encoded factors such as, the timing of the deployment and the precise location within the cell of these attack and defence responses, play important roles influencing the
outcome of infection, although sometimes this delicate balance is tipped in favour of one or the other by external environmental factors such as temperature and humidity.

This thesis aims to examine some of the mechanisms involved in the pathogenicity of the necrotroph *Phoma medicaginis*, a pathogen of legumes, in particular *Medicago* species. Specifically, this thesis focuses on the genomic analysis of *P. medicaginis* var. *medicaginis* isolate OMT5 (PmedOMT5) and its interaction with the host from which it was isolated, the model legume species *Medicago truncatula*. This introductory chapter is divided into three sections. The first section describes the disease spring black stem and leaf spot and its causal agent *P. medicaginis*, which is the focus of this genomic study. The significance of diseases caused by necrotrophic fungal pathogens is also discussed. The second section outlines its preferred hosts *Medicago* spp. and their importance as pastures and forage crops. This is followed by a brief outline of the mechanisms utilised by plants to defend themselves against fungal pathogens as well as the strategies used by fungi to overcome these defences in order to derive nutrition from the plant. This chapter is concluded with a brief introduction to whole-genome sequencing using next-generation sequencing technologies and its application to fungal pathogenomics.

### 1.1 The disease: Spring black stem and leaf spot and its causal agent *P. medicaginis* (Pm)

*P. medicaginis* var. *medicaginis* and the variant *P. medicaginis* var. *macrospora*, which differ in conidial size and septation, both contribute to spring black stem and leaf spot (SBSLS) of *Medicago sativa* L. (syn: alfalfa, lucerne), and other related legume species (Boerema *et al.* 1993). In ungrazed pastures, loss of herbage and seed of up to 32% and 53% respectively have been reported due to *P. medicaginis* infection (Barbetti 1992, Barbetti 1993). The disease has been observed on all continents (Kinsey 2002) and is the most common disease of alfalfa in many locales including western Canada (Hwang *et al.* 2006), Iran (Naseri and Marefat 2008) and Western Australia (Wright and Burges 2002).

Spring black stem and leaf spot is common in cool, wet springs and results in serious reductions in yield and quality of the first and second hay cuttings (Graham *et al.* 1979, Lamprecht and Knox-Davies 1984, Barbetti 1987, Smith *et al.* 1988). Disease begins when spores, deposited on leaves by wind, insects or rain-splash (Mead 1964b), germinate and invade the plant. Mycelia proliferate throughout the plant tissues
producing symptoms that include dark brown or black spots on leaves, petioles and stems, that can coalesce and form irregularly shaped lesions that can enlarge to girdle the stem (Leath 1990). Lesions that girdle the taproot at or near the crown lead to plant mortality (Kinsey 2002). Leaves turn yellow and drop from the plant prematurely and infected seed pods discoulour and shrivel. The most severe symptoms occur on leaves when the day temperature is 15-21°C (Barbetti 1987). The disease spreads mainly via rain-splash and increases in severity and rate of development, during extended periods of post-inoculation humidity. When there is no increase in humidity post-inoculation, there is little to no disease and the rate of symptom development is very slow (Barbetti 1987). In the field, pycnidia are rarely formed in lesions on living tissue during the growing season but are abundantly observed on fallen leaves and over-wintered stems (Leath 1990). The fungus is found as mycelia and pycnidia on crop refuse and in soil (Mead 1964a). It has also been reported to spread via infected seed where it can survive for up to 8 years (Cormack 1945), but has been reported not to survive in soils for more than 2 years without a host.

*P. medicaginis* var. *medicaginis* infection causes defoliation and premature death of susceptible *Medicago* species including *M. truncatula* (Barbetti 1987), although *P. medicaginis* var. *macrospora* is reported to selectively infect and produce more severe disease on alfalfa (Boerema et al. 1993). There is no complete resistance in *Medicago* species. Following inoculation all species and cultivars become diseased; however there are large differences in the rate and severity of symptom development (Barbetti 1987, Gray et al. 1990, O'Neill et al. 2003). Stem and foliar infection with *P. medicaginis* not only reduces yield but can affect forage quality, reducing the carbohydrate and protein content (Mainer and Leath 1978) and inducing the production of phytoestrogens in the plant including coumesterol (Barbetti 2007), leading to reduced fertility in grazing sheep.

### 1.1.1 Host range and plant organ specificity

The main hosts of *P. medicaginis* are *Medicago* species, although other legumes in the subfamily Faboideae (syn: Papilionoideae, or Papilionaceae) have also been reported as hosts in the literature, including: *Medicago lupinina*, sweet clover, yellow trefoil (*Lotus corniculatus*) and *Melilotus* spp. (Boerema et al. 2004), as well as *Cicer*, *Glycine*, *Lathyrus*, *Lens*, *Phaseolus*, *Pisum*, *Trifolium*, *Trigonella*, *Vicia* and *Vigna* (Kinsey 2002,
Djebali 2013). *P. medicaginis* has also been isolated from non-legume species including *Anacardium, Annona, Beta, Brassica, Chrysanthemum, Curcuma, Cyperus, Fragaria, Juniperus, Linum, Lycopersicon, Madhuca, Nicotiana, Phlox, Saccharum, Solanum, Strigea, Thermeda, Zea* and *Zinnia* (Lakshmanan and Vanterpool 1967, Kinsey 2002). In addition, *P. medicaginis* has been reported under the names *Phoma cuscutae* and *Phoma jatropae* (Boerema et al. 2004). Isolates show significant differences in pathogenicity (Mead and Cormack 1961, Ellwood et al. 2006b, Djebali 2013) and are generally more pathogenic on the plant species they were isolated from than other species tested (Boerema et al. 1965, Djebali 2013).

Although considered to be a foliar pathogen, *P. medicaginis* has occasionally been reported as a primary root pathogen (Rodriguez et al. 1990) and is sometimes observed as a secondary root pathogen in a complex with other species. On its own it has been reported as causing only mild root symptoms when inoculated onto alfalfa roots in the field (Cormack, 1945), but has been shown recently to infect the roots of *M. truncatula*, chickpea and bean under growth chamber conditions (Djebali 2013). *P. medicaginis* has also, less frequently, been described as an opportunistic pathogen, a saprobe (Kinsey 2002) and an endophyte (Weber et al. 2004).

### 1.1.2 The Pathogen: *P. medicaginis* disease cycle

*P. medicaginis* spores usually germinate on the host plant within 12-48 h (Castell-Miller et al. 2007), however germination rates can be affected by the host disease phenotype. Germination rates of 47.2% were observed on susceptible *M. sativa* DLAs and 27.1% on resistant compared with 95% on water agar plates. Germinating spores in contact with leaf surfaces are covered with an extracellular mucilaginous matrix (Castell-Miller et al. 2007) that is not produced during germination on artificial surfaces. Suggesting its production is triggered by detection of plant-related compounds by the pathogen. *Phoma* does not produce separated appressoria however hyphal germ tubes were observed to swell at the terminus (Castell-Miller et al. 2007). Hyphae invade the plant by direct penetration of, or between, epidermal cells and occasionally via stomata (Kamphuis et al. 2008). Growth is initially intercellular then intracellular in dead and dying leaf cells (Mead 1964b). The death of the initially infected cells corresponds with the first visible symptoms on the leaf (Peterson and Melchers 1942).
Necrotic lesions form faster on older leaves than young fresh leaves, as these are often more resistant purportedly due to higher levels of phytoalexins (He and Dixon 2000). Infection rarely spreads to new leaves produced by the plant after spray inoculation (Paiva et al., 1994 and this study) and in the field lesions are often found only on older crown tissues. However when leaf lesions or parts of lesioned stems are placed under high humidity or on water agar, pycnidia are formed within 48h (Mead 1964b and this study).

Figure 1.1: Growth stages and infection cycle of *P. medicaginis* var. *medicaginis* isolate PmedOMT5

(a) Diffraction interference micrograph of PmedOMT5 spores produced after 9 d growth on PDA some with visible septa (indicated by arrow)  (b) Confocal micrograph of GFP-transformed PmedOMT5 spores germinating and penetrating the surface of resistant *M. truncatula* accession SA27063 3 dpi, penetrated cells autofluoresce (arrows)  (c) Transverse cross-section confocal micrograph showing invasive hyphae growing between plant cells (arrow) 4 dpi SA27063  (d) Transverse cross-section confocal micrograph showing invasive hyphae growing through necrotic plant cells *M. truncatula* accession SA3054 5dpi detached leaf assay (DLA)  (e) Spray-inoculated (10⁶ sp/mL) leaf of *M. truncatula* susceptible cultivar A17, showing spreading chlorosis around invasion sites and leaf defoliation 13 dpi  (f) Petiole lesion on spray-inoculated A17 19 dpi, showing browning at cell edges (arrow)  (g) Top-down view, confocal micrograph of GFP-PmedOMT5 pycnidia emerging from the surface of 6 dpi DLA accession SA3054, arrow indicates open ostiole  (h) Methylene-blue stained transverse cross-section of PmedOMT5 pycnidia erupting through surface of *M. truncatula* accession A17 6 dpi, DLA. Microscopic methods are described in chapter 2.3 and transformation methods used to generate GFP-expressing PmedOMT5 in figures b, c, f are described in appendix 2.2.
1.1.3  Culture morphology and morphological characteristics

*P. medicaginis* colony growth is irregular and hyphae are olivaceous or dark olivaceous grey (Noordeloos *et al.* 1993). Pycnidia are black, globose (spherical or approximately spherical), ostiolate, glabrous (devoid of hair) and produced in abundant concentric zones (Boerema *et al.* 2004) with some formed inside the agar. Pycnidial production is more abundant at 21-27 °C and with exposure to light (Morgan-Jones and Burch 1987). Conidia are cylindrical, hyaline (translucent), unicellular and occasionally seen with one septum (Figure 1.1a) (Noordeloos *et al.* 1993). Chlamydospores are usually only formed in older cultures (Boerema *et al.* 1965).

*P. medicaginis* has been observed to produce brefeldin A on dead colonized plant material from infected *Medicago* species (Weber *et al.* 2004), which is thought to inhibit the growth of saprotrophic fungi competing with *P. medicaginis*. Bryoid, dendritic crystals of brefeldin A were also observed to be produced by the reference culture CBS 316.90 when grown for 2-3 weeks on malt agar (Noordeloos *et al.* 1993). *P. medicaginis* isolates from Tunisia were also observed to produce crystals when grown on Sanderson & Srb medium for 3-4 months (Djebali 2013) as were PmedOMT5 isolates grown on PDA for 8 weeks (Figure 1.2).

![Image has been removed due to copyright restrictions](image)

**Figure 1.2:** Crystals produced by *P. medicaginis*.

(a) Reproduced from Djebali 2013-*P. medicaginis* crystals produced after 3-4 months growth on Sanderson & Srb medium at 25 °C with 16h light (b) crystals produced by *P. medicaginis* isolate OMT5 after 8 weeks growth on PDA at 21 °C with 12 h light (Williams *et al.*, unpublished data).

Isolates have been observed to undergo conidial anastamosis both on leaves and *in vitro* (Mead 1964b, Castell-Miller *et al.* 2007) and to produce heterokaryons (Sanderson and Srb 1965).
1.1.4 The genus *Phoma*

The genus *Phoma* has long posed a challenge for plant pathologists and fungal taxonomists due to the asexual nature of the majority of its species, frequent changes in taxonomic placement and its use as a “waste-bin” for unclassified fungi, fitting broad morphological similarity but with different hosts (Aveskamp *et al.* 2010). This means that the genus is both polyphyletic (de Gruyter *et al.* 2009) and contains enormous ecological diversity (Aveskamp *et al.* 2010). Furthermore, sequences deposited in online sequence databases have occasionally been erroneously named, resulting in self-perpetuating errors stemming from mis-identifications following online BLAST searches (Bridge *et al.* 2003, Weber *et al.* 2004).

In 1991, more than 2,000 species were classified as *Phoma* (Montel *et al.* 1991), although this number has since been revised downwards. The genus is currently subdivided into 9 sections that were initially based on morphology (Boerema *et al.* 1997, Boerema *et al.* 2004) but are in the process of being re-classified based on molecular DNA analysis of reference cultures in the Netherlands at the Centraalbureau voor Schimmelcultures (CBS) and Plant Protection Service (Plantenziektenkundige Dienst, PD) (Aveskamp *et al.* 2008, de Gruyter *et al.* 2009, Aveskamp *et al.* 2010, de Gruyter *et al.* 2013). There are presently over 220 recognised *Phoma* species (Aveskamp *et al.* 2008) that occur ubiquitously in the environment and occupy many ecological niches. Approximately half of the ~220 species described in the *Phoma* identification manual (Boerema *et al.* 2004) are pathogens of land plants causing leaf and stem spots (Aveskamp *et al.*, 2010 and references therein), although most are also continuously present in the soil as saprobes and switch to pathogenic lifestyles only when they encounter a suitable host (Aveskamp *et al.* 2008). The remainder are described as endophytic, fungicolous (living on or associated with other fungi) or lichenicolous (living exclusively on lichens), with 10 species even identified as pathogens of humans and other vertebrates and several also found in marine environments (Aveskamp *et al.* 2010 and references therein).

The type species of the genus is *Phoma herbarum* (Boerema 1964), which has no known sexual state. This is also the case for the majority of species identified as *Phoma*, although several teleomorphs have been identified (Aveskamp *et al.* 2010).
1.1.5 Phylogenetic placement of P. medicaginis var. medicaginis Malbr. & Roum. 1886

Throughout its study, the taxonomic placement of P. medicaginis has been controversial and has been altered several times in regard to both genus and species. The species was first described as Phoma herbarum Westd. f. medicaginum Westd. in 1862 (Fungi europ, exs./Klotzschii Herb. mycol. Cont. (Ed. Rabenhorst) Cent. 5, No. 455b in herbarium 'Nationale Plantentuin van Belgie' Meise), then as Phoma medicaginis Malbr. & Roum. apud Roumeguère in 1886 (Fungi Selecti Galliaei Exs. Cent. 37, No. 3675), with the varietal name var. medicagini in 1886 (Revue mycol., Toluse 8: 91. 1886, (sensu stricto).

The current name P. medicaginis var. medicaginis, based on that assigned by Malbranche & Roumeguère, was re-instated by Boerema in 1965 as it was found to be a distinct species from P. herbarum (Boerema et al. 1965). Morgan-Jones and Burch (1987) stated that the variety name medicaginis became redundant when P. pinodella, previously called P. medicaginis var. pinodella, was elevated to species level, but it is still in current use (Aveskamp et al. 2010). Some of the most common previous incarnations include: Phoma herbarum Westd. var. medicaginis Fuckel, Ascochyta imperfecta Peck, Phoma herbarum var. medicaginis or medicaginum, Phoma cuscutae (Negru and Verona 1966), Phoma jatropae or jatrophae (Shreemali 1978, Boerema et al. 1993, Boerema et al. 2004). Most authors describe P. medicaginis var. medicaginis as an asexual species, and although a teleomorph stage has been reported once, this observation has not been confirmed by other studies (Leath 1990).

P. medicaginis is now classified as a member of the family Didymellaceae in the order Pleosporales of the class Dothideomycetes (Figure 1.3) based on molecular phylogeny (Aveskamp et al. 2010). This species has been divided into several variants including P. medicaginis var. medicaginis and P. medicaginis var. macrospora, which differ in conidial size and septation. The macrospora variant is proposed to have arisen in the cold mountainous regions of south west Asia and is more often observed in North America where predominantly cold-tolerant varieties of alfalfa are cultivated, whilst the medicaginis variant is proposed to have secondarily adapted to regions of warmer climate along with its host (Boerema et al. 1993). The pea-pathogenic species previously described as P. medicaginis var. pinodella, which was originally thought to
be closely related, but has since been reclassified as *Peyronellaea pinodella* (Aveskamp *et al.* 2010) also assigned to the family Didymellaceae.

Recent sequencing and phylogenetic studies conducted by Aveskamp (2010), later corroborated by studies including the isolate that is the focus of this study, PmedOMT5 (Ellwood *et al.* 2006b, Kessie 2013) indicate that *P. medicaginis*, is reliably placed within the Didymellaceae and its closest sequenced relatives are the *Ascochyta* pathogens of lentil, faba bean and chickpea. The phylogeny of the *Ascochyta* pathogens closely mimics the phylogeny of their hosts and it has been proposed that they have co-evolved with their hosts, cool season legumes, which are native to southeastern Turkey, Iran, Iraq, and Syria (Peever 2007).

### 1.1.6 Plant-pathogenic fungi

The majority of fungi live saprotrophically, feeding from dead or decayed organic matter, however a small minority have evolved the ability to survive by deriving nutrition from attack of living hosts including plants (Knogge 1998). Pathogenicity on plants begins with attachment of the fungal spore to the host surface, followed by penetration of the outer cell wall and invasion of the plant tissues often by the production of cutinases, cellulases, pectinases and proteases (Knogge 1998 and references therein).

Traditionally plant pathogenic fungi have been divided into three feeding classes: biotrophs - which feed from living tissue often secreting effectors to minimise the host’s defence response and keep the parasitised cells alive; necrotrophs - which derive their nutrition from dead host cells, and; hemi-biotrophs - which live part of their lifecycle as biotrophs before switching to necrotrophy. Whilst these categories convey a general impression of the fungal/host interaction the distinctions between the classes is not always clear (Stergiopoulos *et al.* 2013) and several pathogens originally described as necrotrophs have been identified to live at least part of their lives biotrophically (Goodwin *et al.* 2011, Williams *et al.* 2011). Necrotrophs have recently been identified as much more sophisticated pathogens than originally thought producing reactive oxygen species (ROS) as well as phytotoxic secondary metabolites and proteins in order to promote cell death in the plant rather than simply chewing their way through plant cell walls with a battery of degradative enzymes.
The successful survival of these osmotrophic microorganisms, is dependent upon their secreted arsenal. A successful necrotrophic plant pathogen needs an armoury of effectors to both attack and (or) evade the host plant immune system as well as enzymes for the digestion of plant tissues which are composed of cellulose, lignins, lipids and proteins and the means for uptake of the digested products including simple sugars, amino acids, and fatty acids for use by the growing fungal cell (Soanes et al. 2007).

1.1.7 Phylogenetic context: the order Pleosporales and its impact on agriculture

The majority of fungal species identified to produce host-specific effectors belong to the order Pleosporales of the class Dothideomycetes (Stergiopoulos et al. 2013). This group, which includes *P. medicaginis*, contains several necrotrophic species that cause diseases of important crops such as wheat, barley, corn and canola. Whole-genome sequencing and effector characterisation has been conducted for a number of agriculturally-important pathogens of the Pleosporales, including: *Phaeosphaeria nodorum*, causal agent of stagonospora nodorum blotch of wheat (Friesen et al. 2006, Hane et al. 2007, Liu et al. 2009, Liu et al. 2012, Syme et al. 2013), *Pyrenophora tritici-repentis* (tan spot of wheat) (Tuori et al. 1995, Strelkov et al. 1999, Manning et al. 2013), *Pyrenophora teres f. teres* and *f. maculata*, (barley net blotch) (Sarpeleh et al. 2007, Ellwood et al. 2010), *Leptosphaeria maculans* (black leg of canola) (Van de Wouw et al. 2010, Rouxel et al. 2011), and *Cochliobolus heterostrophus* (southern corn leaf blight) (Yoder 1980, Condon et al. 2013).
Within the order Pleosporales, *P. medicaginis* belongs to the family Didymellaceae, which includes devastating pathogens of many legume crops including: *Ascochyta rabiei* - the causal agent of Ascochyta blight of chickpea (Kaiser 1973); the causal agents of Ascochyta blight on pea- *Peyronellea pinodes* (previously *Mycosphaerella pinodes*), *Ascochyta pisi* and *Peyronellaea pinodella* (previously *P. medicaginis* var. *pinodella*) which are generally found in the field as a complex (Wallen 1965, Bretag et al. 2006); as well as *Ascochyta lentis* and *Ascochyta fabae* that cause disease on lentil and faba bean respectively. These species cause economically important diseases in all areas of the world where these crops are grown (Peever 2007). Most Ascochyta species have a bipolar, heterothallic mating system and the ascospores formed following sexual crossing can travel long distances on the wind, which is believed to be the source of primary inocula (Peever 2007).

The only plant pathogen in the Didymellaceae family that is known to produce a characterised host-specific toxin is *Peyronellaea zeae-maydis* (Aveskamp et al. 2010). *P. zeae-maydis* produces a polyketide called PM-toxin, an analog of T-toxin from *C.*
heterostrophus, that confers pathogenicity on T-cms maize (Yun et al. 1998). Other pathogens of the Didymellaceae family are known to produce non-specific phytotoxic secondary metabolites (Alam et al. 1989, Chen and Strange 1991, Hohl et al. 1991, Latif et al. 1993, Andolfi et al. 2013) and A. rabiei produces a proteinaceous effector for which its range of host-specificity has not yet been determined (Chen and Strange 1994).

1.2 The hosts: Legumes in Australian agriculture

With more than 18,000 member species, the Fabaceae (syn: Leguminosae) family is second only to grasses in its global economic importance to agriculture. Forage and pasture legumes provide important sources of nutrition for animal and dairy production and seed legumes (pulses) are an important protein source for much of the world’s human population. In Australia, animal products based on legume pastures are valued in the billions. Legumes contribute much of the feed for Australia’s wool, dairy and livestock industries, which are valued at $3 billion, $4 billion (NFF 2012), and ~$22 billion per annum respectively (ABARES 2012). Legume crops also aid the $4.8 billion Australian wheat industry (NFF 2012) through nitrogen fixation and disease breaks between crop rotations.

The legume family is the third largest family of flowering plants and is broadly defined by their unusual flower structure, podded fruit, and the ability of the majority (88%) of its species examined to date to form nodules with symbiotic nitrogen-fixing bacteria (De Faria et al. 1989). It is this last characteristic that elevates their importance in agriculture above mere forage crops as they can take nitrogen, an essential component of all proteins, from the air and convert (or ‘fix’) it to ammonia. Upon the plant’s death, fixed nitrogen is released back into the soil and can be used to fertilise future crops, reducing the need for nitrogenous fertilisers. Legumes have been used as pastures and for soil improvement since Roman times, with Varro (37 BC, cited by Fred et al., (1932) in Graham and Vance (2003)) stating “Legumes should be planted in light soils, not so much for their own crops as for the good they do to subsequent crops”.

The diseases caused by fungi are important factors in the yield and grain quality of agricultural legumes, with severity of disease affected by many factors including: crop and pathogen species, geographical area, environmental conditions and cultural practices (Infantino et al. 2006). There is no complete resistance to necrotrophic fungi
in cultivated legumes and disease severity is dependent on cultivar and environment. In cool wet seasons most plants are susceptible and necrotrophic fungi can cause major losses, such as in 2003 when chickpea crops in NSW were decimated due to Ascochyta blight (caused A. rabiei) and similarly in 1988 when wide ranging Ascochyta blight infections occurred in south eastern Australia (CLIMA, Vic DPI). Effective strategies for managing legume diseases caused by necrotrophs involve the application of fungicides in the field and breeding for improved disease resistance. The combined annual losses from major fungal and viral diseases of legume based pastures across southern Australia is estimated to be up to $800 million (Shovelton et al. 2012). However the size and complexity of legume genomes has meant slow progress in genetic characterisation of disease resistance.

1.2.1 Australian Medicago pasture crops

Scientific broadacre agriculture in Australia began in the late 19th century with the use of ley farming (crop rotation with alternate annual seed crops and legume pastures) to counteract the poor nutritional content of the soil. This practice did not show significant development until after the second world war, with a rapid expansion in improved pastures occurring in the 1950s (Wolfe 2009). Since that time forage crops (including Medicago spp.), which are generally grown from April-May (late autumn) to November (late spring), have often been used in rotation with non-legume crops such as wheat. This can produce an increase in wheat grain yield and protein, as well as providing a disease break that is effective against narrow host-range cereal pathogens (Fedorenko et al. 2009).

The genus Medicago originated in the Mediterranean basin (Lesins and Lesins 1979). It belongs to the Trifolieae tribe that also includes other major forage legumes such as clovers (Figure 1.4) and is phylogenetically close to the Vicieae tribe that includes grain legumes such as pea, lentil and Faba bean (Julier et al. 2007). Annual Medicago spp. are adapted to mildly-acidic, neutral or alkaline soils and since naturalisation in the 1800s have been found frequently in valley floors and the mallee soils of the Western Australian wheatbelt (Wolfe 2009). Their growth is mainly limited to southern and coastal Australia (as illustrated in Figure 1.5) due to appropriate day-lengths, temperatures and rainfall in this region (Wolfe 2009). Approximately 17 million ha of perennial and annual legume pastures are grown across the Australian “wheat-sheep
belt” (Woolfe 2009-cited in Shovelton et al. 2012). Ongoing research, and breeding to create improved pasture and forage crops and to introduce beneficial characteristics such as disease resistance is being conducted by several research groups in Australia including SARDI and AMGRC (Nair and Howie 2006, Nichols et al. 2007). These efforts have led to the release of several commercial cultivars.

Figure 1.4: Taxonomic relationships within the two major clades of crop legumes (galegoid and phaseoloid).

Divergence times estimated based on Penalised Likelihood analysis. MYA = million years ago. Figure reproduced from Choi et al. 2004b.
1.2.2 Medicago sativa

Alfalfa (Medicago sativa, lucerne), the oldest forage crop in the world, is an autotetraploid, perennial crop that can be used for pasture but is most commonly harvested as hay for animal fodder (Bauchan and Hossain 2001). Based on archaeological records alfalfa has been cultivated in the Middle East for over 4,000 years and is currently grown on around 30 million hectares in temperate regions worldwide (Yuegao and Cash 2009). It provides a source of high protein fodder with a high yield per hectare. Alfalfa was first introduced to Australia in 1806 and there were 3.5 Mha of lucerne pastures in 1997 which is projected to increase to 7 Mha (Shovelton et al. 2012). Its growth has been suggested as a solution for increasing salinity and waterlogging in the WA wheatbelt due to its deep rooting depth of 2-3m (Wolfe 2009). Between 2000 and 2007 Australia exported over 12 million metric tons as hay and 76,700 tons as pellets (Yuegao and Cash 2009). Serious yield losses caused by leaf and stem diseases like P. medicaginis can cause millions of dollars worth of economic losses and result in poor quality forage (Mainier and Leath 1978, Hwang et al. 2006).
1.2.3 Medicago truncatula

*M. truncatula* Gaertn (commonly known as barrel medic due to the shape of its seed pods) is a favoured model for studying legume biology as a diploid (2n=16) relative of alfalfa (2n=16, 4n=32), with a short generation time and self-fertilising reproduction meaning it is mainly homozygous unlike alfalfa (Cook 1999). *M. truncatula* has been the subject of many studies ranging from metabolome (Farag et al. 2007, Urbanczyk-Wochniak and Sumner 2007, Farag et al. 2008) and proteome analysis (Watson et al. 2003) to examination of its interaction with rhizobia (Barker et al. 1990, Catoira et al. 2000, Wasson et al. 2006). It has been used as a model plant to study the interactions with necrotrophic fungal pathogens by several research groups (Torregrosa et al. 2004, Ellwood et al. 2006b, Moussart et al. 2006, Tivoli et al. 2006a, Tivoli et al. 2006b, Ellwood et al. 2008, Anderson et al. 2010).

The ~500 Mbp genome of the reference accession A17 of *M. truncatula* has been sequenced (Young et al. 2011) along with those of 26 other accessions (Branca et al. 2011) and there are several microarray analyses (Naoumkina et al. 2007, Kamphuis et al. 2011) and well-characterised EST libraries prepared from important life-cycle stages and examining various stress, pathogen and symbiont interactions (Gamas et al. 2007). Unlike *Arabidopsis*, *Medicago* is a true field crop and any desirable characteristics identified can potentially be utilised directly or readily transferred to other legume crops due to their close synteny, although there is considerable variation in DNA content and ploidy, there appears to be good conservation of macrosynteny between the species (Choi et al. 2004a, Aubert et al. 2006). There are also many available genetic and cytogenetic maps (Thoquet et al. 2002, Choi et al. 2004a, Kamphuis et al. 2008, Pierre et al. 2008, Guo et al. 2009) as well as mutant lines generated using various techniques (Tadege et al. 2005, Le Signor et al. 2009) which can facilitate investigation of genes of interest.

1.2.4 Methods of disease control in the field

Disease in alfalfa requires more management than *M. truncatula* due to the perennial nature of the crop. Fungicide treatment of *P. medicaginis* is generally efficient but not usually cost-effective for use in pastures unless they are high value hay or seed production stands (Barbetti 1992, Shovelton et al. 2012). The commercial fungicides benomyl and propiconazole were found to effectively control *P. medicaginis* infection.
(Hwang et al. 2006). However, there may be undesirable effects on livestock from fungicide residues and most studies recommend growing resistant cultivars followed by rotation with non-legume crops to minimise disease build up in the field.

Several biocontrol agents have been tested including *Streptomyces* sp. (Samac et al. 2003) which whilst effective against *P. medicaginis*, significantly reduced plant dry weight. Interestingly the inoculation of the *M. truncatula* accession JA17 with an antagonist strain of *Sinorhizobium meliloti* 10.16/R6 created increased resistance to *P. medicaginis* in the plant that was not seen in plants inoculated with a non-antagonist strain (Mrabet et al. 2011).

### 1.3 Plant defence against fungal pathogens

Plants avoid invasion and colonisation of their tissues by surrounding them with a physical barrier, the plant cell-wall, which is often protected by a waxy cuticle and secreted anti-microbial molecules. To gain access pathogens must either overcome or evade these barriers by entering via wounds or natural openings, such as stomata. Plant defence is an active process that begins with recognition of attempted invasion followed by the coordinated activation of local and systemic biochemical pathways and the transcription and translation of defence related signals. Where, when and at what concentration the relevant molecules accumulate can determine the outcome in favour of plant or pathogen (Hammerschmidt and Kagan 2001). There are two active zones in the plant defence response. Firstly, the local area in contact with the pathogen in which occurs what is known as the hypersensitive response (HR), a localised biochemical attempt by the plant to contain the invasion. Secondly, a plant-wide response, coordinated by the hormone salicylic acid that leads to the development of systemic acquired resistance (SAR), which provides long-term defence against a broad range of pathogens.

The initial defence response is triggered in plant cells by the detection of Pathogen Associated Molecular Patterns (PAMPs) (Jones and Dangl 2006). PAMPs can be molecules associated directly with the pathogen such as chitin or peptidoglycan from fungal and bacterial cell walls or breakdown products of plant cells such as digested cell wall products which are detected by ‘guard’ receptors in the cell membrane known as pattern recognition receptors (PRRs). This system, which efficiently detects and wards off the majority of potential pathogens is known as PAMP-triggered immunity (PTI). It
is a broad-spectrum defence triggered by recognition of invader or non-self components by the plant and is classically associated with the HR. Pathogens are only successful if they are able to produce molecules (effectors) that circumvent the host’s defences resulting in effector-triggered susceptibility (ETS). Plants have in turn evolved a secondary line of defence which can perceive such effectors either directly or indirectly often via NB-LRR receptors within the cell that then signal the initiation of a second layer of defence known as effector-triggered immunity (ETI) (Jones and Dangl 2006). ETI is a more robust and pathogen specific defence response that also often culminates in HR, although the difference between PTI and ETI may simply be a difference of degrees as several components are shared between the responses (Dickman and Fluhr 2013). Pathogens that are able to avoid ETI via the loss or diversification of the recognised effector or the acquisition of additional effectors that can suppress ETI will be favoured by natural selection (Jones and Dangl 2006).

PRRs can detect PAMPs at sub-nanomolar concentrations (Boller and He 2009). However not all plants are sensitive to all PAMPs. If a plant does not possess the appropriate PRR, the PAMP goes undetected. Some PRRs such as FLS2, which detects bacterial flagellin, appear to be conserved in all higher plant species, while others have evolved to recognize different PAMPs in different plant families (Boller and He 2009). One mechanism by which fungi evade PTI is to dephosphorylate kinases that are involved in PRR signalling (reviewed in Boller and He, 2009).

The HR includes the induction of expression of pathogenesis related (PR genes), many of which encode extracellular products (Soh et al. 2012), as well as localised callose and ROS production and accumulation of phytoalexins surrounding the initial infection site (Yoshikawa et al. 1978). Increases in calcium ions (Ca2+) in the cytosol can occur within seconds of PAMP detection and are thought to be a key primary signal for downstream events (Chandra and Low 1997, Grant et al. 2000) as well as efflux of potassium and chloride ions from the cell (Fellbrich et al. 2000).

Light also appears to play an important role in plant defence responses (Dickman and Fluhr 2013). The induction level of defence related transcripts is lower in the dark in wounded plants and it appears that full wound responsiveness requires energy and/or an unknown chloroplast-derived signal (Morker and Roberts 2011) although this may relate more to SAR than to localised HR responses.
1.3.1 Programmed cell death (PCD)

One of the main components of the HR response is rapid plant cell death surrounding the site of pathogen invasion. This is an effective defence against biotrophic pathogens which require living cells for sustenance but is less effective against necrotrophs which induce a spreading cell death from the inoculation site partly by promoting programmed cell death (PCD) in the host cells (Glazebrook 2005). PCD is a genetically controlled, ordered process that requires active metabolism of host cell, HR is a form of PCD but the two are not equivalent (Dickman and Fluhr 2013). Dickman and Fluhr further describe plant-triggered cell death as autophagy but pathogen-controlled cell death as apoptosis (2013), although the lines of distinction between the two are often blurred as they both regulate cell death (Kabbage et al. 2013). Apoptosis is associated with the activation of nucleases that disrupt chromosomal DNA, causing an observable phenomenon known as DNA-laddering, when genomic DNA is analysed via electrophoresis (Ryerson and Heath 1996). This can be observed during the host cell death triggered by the actions of host-specific necrotrophic fungal effectors such as SnTox1 or AAL toxin in effector triggered susceptibility (Wang et al. 1996, Liu et al. 2012). Apoptosis also includes the production of ROS, cell shrinkage, membrane blebbing, nuclear and chromatin condensation, externalization of phosphatidylserine and DNA fragmentation as well as mitochondrial outer membrane permeabilization and depolarization (Williams and Dickman 2008). Whilst autophagy is a process that leads to the recycling of proteins and cellular organelles which are engulfed in characteristic double membrane vesicles which are delivered to the vacuole for degradation and recycling (Dickman and Fluhr 2013).

PCD can be initiated by chloroplasts or mitochondria, the energy organelles of the plant cell, and these are often the ultimate targets of effector molecules secreted by plant pathogenic fungi (Curtis and Wolpert 2004, Walton 2006, Manning et al. 2007). During the HR response the vacuole either fuses with the outer membrane of the plant cell (the plasma membrane) or is destroyed, releasing its contents that include pathogenesis-related (PR) proteins, chitinases and hydrolases into the extracellular space (Hatsugai and Hara-Nishimura 2010). Pathogen effectors that can suppress this membrane fusion are able to delay plant cell death, allowing biotrophic feeding (Eichmann et al. 2004).

If the ability of the plant to control PCD is altered (as in AtBag6 mutants: a stress-upregulated calmodulin-binding protein involved in Arabidopsis PCD) they become
hypersensitive to necrotrophs such as *Botrytis cinerea*, suggesting that autophagy is one of the mechanisms involved in the basal defence response against necrotrophs by stopping fungus induced cell death from spreading (Kabbage and Dickman 2008). However necrotrophic fungi themselves also promote PCD in the plant via the secretion of effectors such as victorin, oxalic acid, SnTox1 and fumonisin BI (Wolpert *et al.* 2002, Liu *et al.* 2012). In *Arabidopsis*, LOV1, a NB-LRR guard protein stimulates host cell death when the *Cochliobolus victoriae* effector, victorin, binds to a thioredoxin protein (TRXh5) responsible for redox control of NPR1, a transcriptional regulator (Lorang *et al.* 2012). The binding of victorin by THXh5 is monitored by LOV1 which, is localised to the cell membrane and when activated, causes cell death. The NB-LRR resistance (R) protein must be present for susceptibility to occur as this is the trigger for PCD. The same gene is also hypothesised to confer resistance to a biotrophic fungus, the rust pathogen *Puccinia coronata*, in oats. In the case of the biotrophic pathogen, the PCD produces resistance to the pathogen whilst for the necrotrophs the fungal triggered PCD appears to facilitate infection.

As both biotrophic or necrotrophic infection of resistant or susceptible plant tissues results in death of the plant cells with different outcomes for the pathogen, it appears that one of the determining factors in the success of a host-pathogen interaction is whether the host or the pathogen is in control (Kabbage *et al.* 2013). Is it a case of the plant sacrificing a few cells for the greater good, or the necrotrophic fungus turning the defences of the host against itself for its own ends? The precise molecular mechanisms which trigger host-cell death are the focus of many current studies.

PCD can also be triggered in the invading fungus by host-plant metabolites such as phytoanticipins and phytoalexins as a plant defence mechanism (Ito *et al.* 2007, Shlezinger *et al.* 2011b). In the interaction between the necrotroph *B. cinerea* and *Arabidopsis thaliana* massive PCD is triggered in the fungus during the first 48 h of infection as the fungus establishes itself in living tissue (Shlezinger *et al.* 2011b). The fungus defends itself by producing anti-apoptosis genes to block the host-triggered PCD preserving sufficient cells to give rise to a new infection throughout the necrotic tissue. Although once the fungus is established and secreting effectors it enters a different growth phase and little fungal PCD occurs (Shlezinger *et al.* 2011b).
1.3.2 Reactive oxygen species (ROS)

Reactive oxygen species (ROS) which include peroxides, superoxides, nitric oxide and oxygen ions (singlet oxygen), play several roles in plant defence including acting as signalling molecules, facilitating the strengthening of cell walls and as anti-microbrial compounds (De Gara et al. 2003). The reinforcement of cell walls by cross-linking of cell wall proteins catalysed by peroxidases may make them more resistant to cell wall degrading enzymes (Brisson et al. 1994) and possibly restricts the entry of effectors (Djebali et al. 2007).

During pathogenesis or wounding plant cells produce large amounts of ROS, mostly via plasma membrane localised NADPH oxidases (Fluhr 2009), or cell wall localised peroxidases (Bolwell 1999). Lower concentrations of ROS, often H$_2$O$_2$ (Bhattacharjee 2012), are produced by the plant to coordinate PCD which results in restricted cell death and resistance to biotrophic pathogens. The plant protects itself against this mild oxidative stress via antioxidant enzymes, however the higher concentrations present during HR are toxic to the plant cell. Biotrophic and hemi-biotrophic pathogens defend themselves against these measures by suppressing ROS production to facilitate invasion of the plant cells (Sohn et al. 2007). However generation of ROS during cuticle penetration and lesion formation in necrotrophs such as B. cinerea might be beneficial to the pathogen.

Some necrotrophic pathogens also manipulate ROS production in the host. The broad host-range necrotrophic pathogen Sclerotinia sclerotiorum can directly induce host cell death and collapse of the vacuole by the secretion of the non-selective phytotoxic effector oxalic acid (Williams et al. 2011). The secretion of oxalic acid triggers a reducing environment in the cell which suppresses the plant’s production of ROS and through this, other defence mechanisms such as callose deposition. However once the infection is established the fungus triggers ROS production in the plant which leads to PCD of the host cells, although recent studies suggest this mechanism may be associated with a form of hemi-biotrophic growth prior to the transition to necrotrophy (Kabbage et al. 2013)
1.3.3 Phytoalexins and phytoanticipins - secondary metabolite plant defence compounds

The majority of plants produce antimicrobial secondary metabolites as part of their normal growth (phytoanticipins) or in response to stress or pathogen attack (phytoalexins). Phytoanticipins, such as saponins, are constitutively produced whilst phytoalexins are synthesised \textit{de novo} by the plant following microbial attack and stresses including UV light and heavy metals (VanEtten \textit{et al.} 1994, Morrissey and Osbourn 1999, VanEtten \textit{et al.} 2001). Although some compounds, such as the methylated flavanone sakuranetin, are expressed as a phytoanticipin in one species (blackcurrant) (Dixon 2001) and act as phytoalexin in another (rice) (Kodama \textit{et al.} 1988). The term phytoalexin was proposed in 1940 by Müller and Börger (Müller and Börger 1940) and is derived from Greek to mean “warding-off agents in plants” (Jeandet \textit{et al.} 2013). Phytoalexins are low molecular weight secondary metabolites produced by the plant following pathogen infection. They are often as toxic to the plant as they are to fungal pathogens (Skipp \textit{et al.} 1977) and are usually produced \textit{de novo} but may also be stored as glycosides, a more hydrophilic form, in vacuoles and released upon attack (Mackenbrock \textit{et al.} 1993).

Phytoalexin production is an active response promoted within the cell as a response to attempted infection which is controlled by complex regulatory mechanisms, at least in \textit{Arabidopsis} where it has been studied in the most detail. Studies examining the responses of mutants at various stages in the ethylene, jasmonic acid (JA) and salicylic acid (SA) signalling pathways as well as the cytochrome p450 gene, \textit{pad3}, that is directly involved in the production of the phytoalexin camalexin, show that different pathways are involved in responses to different pathogens (Jeandet \textit{et al.}, 2013 and references therein). Based on these studies it appears that camalexin production may limit disease symptoms by triggering PCD in the fungus but it is not responsible for limiting entry of the pathogen in resistant interactions (Dixon 2001, Shlezinger \textit{et al.} 2011b).

The first characterised phytoalexin, pisatin, was demonstrated to be produced by \textit{Pisum sativum} (garden or field pea) in response to biotrophic and necrotrophic fungi (Cruickshank and Perrin 1960). Phytoalexins produced by related plant species tend to share a chemical structure. The legume family produce many phytoalexins with an isoflavanoid skeleton; crucifers produce indole alkaloids; cereals produce mostly cyclic
hydroxamic acids and diterpenoids whilst plants of the Solanaceae family produce sesquiterpenoids and polyacetylenes although several plant families produce stilbenes (Pedras and Ahiahonu 2005). The production of phytoalexins tends to be restricted to the tissue colonized by the fungus and cells surrounding the site of infection (Morrissey and Osbourn 1999 and references therein). Studies have shown that an important factor in the success of phytoalexins in plant defence is the production at the right time, concentration and location to be effective in resistance (Pierce et al. 1996).

The importance of phytoanticipins in defence is demonstrated in mutants of oat that cannot produce the saponin, avenacin, leading to increased susceptibility to the fungal pathogen *Gaeumannomyces graminis* var. *tritici* (Papadopoulou et al. 1999). Contrastingly, some fungi have evolved mechanisms to counteract these host defences. One example of this is the fungal pathogen of oat roots, *G. graminis* which produces an enzyme called avenacinase that detoxifies avenacin. Mutants that are unable to metabolise avenacin are non-pathogenic on oats although they retain full pathogenicity on wheat which does not produce saponins (Bowyer et al. 1995).

### 1.3.4 Legume secondary metabolites - phenylproponoid derived isoflavonoids

*Medicago* species are known to produce secondary metabolites via the phenylpropanoid pathway that are directly involved in the interaction between the plant and its environment, with the multiple branches leading to the production of various compounds including anthocyanins, lignins, and isoflavonoids (Yu et al. 2000) (Winkel-Shirley 2001). Phenylpropanoid products play a role in defence against pathogens and environmental stresses including prolonged cold, UV-B light and heavy metals, as well as, establishment of symbiotic nitrogen fixation and attraction of pollinators and seed dispersing organisms (Paiva et al. 1994, Dixon and Paiva 1995, van Rhijn and Vanderleyden 1995). These compounds also have roles in plant and animal health (Dixon and Sumner 2003) with demonstrated estrogenic, antioxidant and anticancer activities and have been used in the treatment of human cardiovascular disease and post-menopausal disorders (Rice-Evans et al. 1997, Dixon and Sumner 2003).

The major products involved in plant defence are isoflavonoids and pterocarpan. Their production has been reported many times in the literature to show a spatial and temporal correlation with resistance (Dixon et al. 1983, Rice-Evans et al. 1997). The signals that
are involved in the initiation of pterocarpan synthesis are not all known but it is known that genes at different stages in the pathway respond to different signals.

All phenylpropanoids are derived from trans-cinnamic acid, which is formed from phenylalanine by the enzyme phenylalanine ammonia-lyase (PAL). PAL is the branch point enzyme between the primary (shikimate pathway) and secondary (phenylpropanoid) metabolism (Dixon and Paiva 1995). The entry point enzyme into the isoflavonoid pathway is isoflavone synthase (IFS) a membrane bound cytochrome p450 attached to the outer surface of the ER by an N-terminal membrane anchor (Kochs and Grisebach 1986). The genes that occur early in the pathway are also involved in the biosynthesis of many other phenylpropanoid products with a wide range of functions and which are expressed under different environmental and developmental conditions which do not result in pterocarpan biosynthesis.

1.3.5 Medicarpin

The major pterocarpan phytoalexin produced upon interaction with fungal pathogens by M. sativa and M. truncatula is medicarpin (Higgins 1972, Paiva et al. 1994, Jasinski et al. 2009). Pterocarpan and isoflavonoid production can also be induced in M. sativa (Dalkin et al. 1990) and M. truncatula cell cultures following elicitation with jasmonic acid or yeast cell walls, which appear to trigger synthesis of the penultimate molecule medicarpin from different points in the phenylpropanoid pathway (Naoumkina et al. 2007). Before elicitation, cell cultures contain no free medicarpin, but do possess small amounts bound as medicarpin-3-O-β-D-glucoside-malonate (Farag et al. 2007, Farag et al. 2008). Following contact with a fungal elicitor M. sativa suspension cell cultures were found to secrete medicarpin and strengthen cell walls via methylation of pectin and other cell wall bound polysaccharides (Kessmann et al. 1990, Edwards et al. 1997).

The initial release of medicarpin following fungal elicitation (0-2h) was observed to occur from hydrolysis of glycoside conjugates which was then followed by the synthesis of new conjugates, which were exported from the cell (Edwards et al. 1997).

The production of medicarpin in planta occurs following elicitation by biotic or abiotic stresses and appears to be localised to affected tissues. Medicarpin is not produced following incubation of M. sativa roots with the symbiotic arbuscular mycorrhizal fungus Glomus versiforme. In fact, a decrease in free medicarpin is observed as the mycorrhiza becomes established although other isoflavonoids are produced (Paiva et al.
HPLC analysis showed production of medicarpin in leaves of the susceptible alfalfa cultivar Apollo with lesions caused by *P. medicaginis* inoculation, but not in those without (Paiva *et al.* 1994). HPLC-DAD analysis of three resistant and three susceptible cultivars of *M. truncatula* shows significantly higher production of medicarpin 48h post inoculation with *P. medicaginis* isolate PmedOMT5 (*p* < 0.01) in resistant accessions when compared to susceptible accessions (Williams, Ellwood, Oliver *et al.* unpublished data). A similar pattern was observed in chickpea cultivars following elicitation (Daniel *et al.* 1990)

Several pathogens of alfalfa have been observed to degrade medicarpin added to culture medium, including *P. medicaginis* and *Stemphylium botryosum* (Higgins 1972). Similar observations have been made for the degradation of pisatin and maackiain, the major phytoalexins produced in pea and chickpea respectively, by fungal pathogens of these hosts (Reimmann and VanEtten 1994, Covert *et al.* 1996, Enkerli *et al.* 1998). Disruption of the genes in the pathogen responsible for phytoalexin detoxification led to decreased virulence but not loss of pathogenicity, whilst transfer of these functions to isolates that did not possess them resulted in only moderate increases in virulence. This suggests that phytoalexins contribute to resistance, but that no single one is wholly responsible (Ciuffetti and VanEtten 1996, Wasmann and VanEtten 1996, Enkerli *et al.* 1998). It has been suggested that the resistance effect of phytoalexins and phytoanticipins may result from cumulative effects (Dixon 2001), suggesting a possible arms race, *i.e.* as a pathogen evolves the ability to detoxify one type, pressure is placed on the plant to produce new toxic metabolites, however the old ones are still retained as they are effective against pathogens not commonly encountered by the host.

Legume plants produce different stereoisomers of medicarpin, (VanEtten *et al.* 1983, Kessmann *et al.* 1990, Guo *et al.* 1994). Alfalfa produces (-) medicarpin and (6aR, 11aR)-medicarpin (Kessmann *et al.* 1990) and peanut produces (+) medicarpin (Blount *et al.* 1994). Interestingly, alfalfa pathogens were found to be more sensitive *in vitro* to the isomer they would not normally encounter, (+) medicarpin, indicating that the detoxifying enzymes from different pathogens have different substrate stereospecificities (van Etten *et al.* 1989, Blount *et al.* 1994).

A recent study of (−) medicarpin isolated from *Canavalia maritima*, a wild Chinese legume, showed a pro-apoptotic effect on the human/cancer cell line He-La cells (Xu *et al.* 2009), however the mechanism of its toxicity to *P. medicaginis* is unknown.
1.3.6 Response and resistance to *P. medicaginis* in *Medicago* species

O’Neil and colleagues studied resistance in 200 accessions of 33 annual *Medicago* species from 5 different continents and found that most were susceptible to *P. medicaginis*, with *M. truncatula* being one of the most susceptible (O’Neill *et al.* 2003). Nevertheless, several sources of resistance have been observed in *Medicago* species including *M. sativa*, *M. truncatula*, *M. sufruticosa* and *M. dzawkhetica* and others (Renfro and Sprague 1959, Mead 1964b, Ellwood *et al.* 2006b, Barbetti 2007, Castell-Miller *et al.* 2007, Kamphuis *et al.* 2008). Although, due to the tetraploid nature of alfalfa crops, inheritance of resistance to *P. medicaginis* has proved complex with both dominant, recessive and epistatic effects identified (Tamimi 1963, cited by Mead 1964b).

Observations by Kamphuis and colleagues (2008) showed no differences in the timing of spore germination or penetration by *P. medicaginis* between resistant (R) and susceptible (S) *M. truncatula* accessions SA27063 (R), SA3054 (S) and A17 (S). Nor were any differences in these characteristics observed by Djebali and colleagues (2007) following inoculation with *P. medicaginis* of *M. truncatula* accessions F83005.5 (S) or DZA45.5 (less S). However, delays of approximately 24 h were observed in spore germination, penetration, development of mycelia and pycnidia formation on detached leaf assays of more resistant lines of *M. sativa* when compared to susceptible lines (Castell-Miller *et al.* 2007). Visible symptom development was also delayed for approximately eight days, after which similar amounts of chlorosis were observed on all genotypes (Castell-Miller *et al.* 2007).

Differences were observed during the colonisation process, with Kamphuis and colleagues (2008) noting that growth of *P. medicaginis* on the resistant line was restricted to one or a few cells only and Djebali and colleagues noting that larger symptomatic areas and faster sporulation occurred in more susceptible accessions (2007). Production of H$_2$O$_2$ surrounding the penetration site was observed by both groups, however Djebali et al. (2007), noted it was more extensive in the less susceptible accession, autofluorescence, possibly due to accumulation of phenolic compounds was also observed in both resistant and susceptible interactions (Kamphuis *et al.* 2008). Djebali and colleagues propose that earlier and higher H$_2$O$_2$ production combined with higher basal levels of peroxidase (POX EC 1.11.1.7) activity may contribute to the lower level of susceptibility observed in DZA45.5, this accession was
ranked as moderately resistant to the Australian isolate PmedOMT5 with a score of 2.19, on a 1-5 scale, where 1 is resistant (Ellwood et al. 2006a).

Resistance to *P. medicaginis* in *M. truncatula* is complex and is controlled by quantitative trait loci with modest effects (Kamphuis et al. 2008). Kamphuis and colleagues identified different QTLs in two populations with the same resistant parent but different susceptible parents, each named *rpm*X for resistance to the necrotroph *P. medicaginis*. In SA27063 x A17 progeny, the identified QTL *rpm1* on linkage group (LG) 4 explained 33.6% of the resistance (LOD = 7.37; *P* < 0.00001) while in SA27063 x SA3054, a QTL on LG 8, *rpm2* (LOD = 6.77; *P* < 0.00001) explained ~29.6% of the resistance (Kamphuis et al. 2008). In SA27063 x SA3054 a second, less significant locus *rpm3* was also identified, located on LG 1 (LOD 3.37, *P*=0.04) which explained 19.4% of the variance and was proposed to potentially have an epistatic effect on *rpm2*. One of the QTLs, *rpm1*, was identified as tightly linked to a cluster of Toll/Interleukin1 receptor nucleotide-binding-site leucine-rich repeat (TIR NBS-LRR) genes and disease resistance protein-like genes. In both cases, the resistance was recessive and the authors postulate that this may be due to host genotype–specific susceptibility loci that interact with effectors produced by *P. medicaginis*. Recessive resistance or dominant sensitivity to host-specific effectors conferring susceptibility has been observed for other necrotrophic pathogens of the order Pleosporales including *P. nodorum* (Friesen et al. 2007).

Unexpectedly high levels of genetic diversity have been found in AFLP analysis of 121 *P. medicaginis* field isolates collected across Minnesota in 2000 (Castell-Miller et al. 2008). This analysis showed little similarity between isolates apart from several small subgroups isolated in close proximity to one another or from different tissues of the same plant. The authors suggest that this indicates that the *P. medicaginis* population is undergoing genetic change or that the (effective) population covers a large geographic area (Castell-Miller et al. 2008). They go on to propose that the lack of defined clonal lineages may be favoured by the nature of the alfalfa population which contains no major resistance genes for spring black stem and leaf spot and thus places no direct selection pressure on the pathogen population.
1.3.7 Previous work on *P. medicaginis* and the involvement of isoflavonoids in resistance

Much of the previous work on the interaction between *P. medicaginis* and *Medicago* species has centred on the production of phytoalexins produced via the phenylproponoid pathway in response to infection (Kessmann *et al.* 1990, Paiva *et al.* 1994, Naoumkina *et al.* 2007, Jasinski *et al.* 2009). Medicarpin has been demonstrated to accumulate in *P. medicaginis* inoculated tissues of *M. sativa* and *M. truncatula* (Higgins 1972; Jasinski *et al.* 2009; Kamphuis *et al.* 2011) and in droplets of *P. medicaginis* spore suspensions on alfalfa leaves (Higgins 1972).

Transcripts of enzymes in the isoflavonoid pathway that lead to the biosynthesis of medicarpin have been observed within 1-2 h of *P. medicaginis* inoculation of alfalfa leaves, prior to the appearance of symptoms. The genes with increased expression include: PAL, chalcone synthase and chalcone isomerise and isoflavone reductase (Paiva *et al.* 1991, Junghans *et al.* 1993, Paiva *et al.* 1994, He and Dixon 2000). Levels of medicarpin itself were observed to increase in *M. sativa* detached leaf assays 4 h post inoculation (hpi) with *P. medicaginis* with sativan detected at 8 hpi reaching a maximum at 4.5 d (Paiva *et al.* 1994). A similar pattern was also observed in *M. truncatula* (Jasinski *et al.* 2009, Kamphuis *et al.* 2011). Other known phytoalexins, sativan, vestitol and 4-O-methoxymedicarpin, were also present in infected leaves, as was the phytoestrogen coumesterol (Paiva *et al.* 1994).

Increased production of medicarpin by over-expressing a gene in the isoflavonoid pathway, isoflavone-O-methyl transferase (IOMT), conferred increased resistance against *P. medicaginis* in alfalfa (He and Dixon 2000). However decreased production of medicarpin in antisense IOMT-transgenic lines did not show a corresponding drop in resistance to *P. medicaginis* (He and Dixon 2000). Although as these experiments were performed using a susceptible alfalfa cultivar (Regen SY) it is possible that a greater difference would have been observed in a resistant cultivar, which has been shown in *M. truncatula* accessions to have increased levels of production of isoflavonoids relative to susceptible cultivars (Kamphuis *et al.* 2011). Several of the intermediate products of the medicarpin synthesis pathway have also been demonstrated to have anti-fungal activity against alfalfa pathogens, which may have resulted in the lack of observed reduction in symptoms (He and Dixon 2000).
Production of a non-endogenous phytoalexin, the stilbene resveratrol, in transgenic alfalfa plants expressing the peanut resveratrol synthase gene, showed an increase in resistance to *P. medicaginis*, reducing the amount of hyphal invasion, size of lesions and number of pycnidia relative to untransformed plants (Hipskind and Paiva 2000). This phytoalexin is produced in grape vines and is involved in their resistance to fungal pathogens. Providing further evidence that organisms are generally intolerant of non-host plant phytoalexins that they would not normally encounter.

Bioassays on agar plates have demonstrated that hyphal growth of *P. medicaginis* is also inhibited by resveratrol and trans-resveratrol-3-O-β-D-glucopyranoside (Hipskind and Paiva 2000). Hyphal growth of some *P. medicaginis* isolates is also inhibited 42% by 0.5 mmol/L medicarpin *in vitro* (Blount *et al.* 1992). However it is difficult to assess, to what concentration the invading fungus would be exposed *in planta*. In most studies, the mean concentration of medicarpin in whole-tissue extracts is measured, whereas the critical parameter is the concentration in the cells that are in direct contact with the pathogen.

1.4 Fungal genomics

The first eukaryotic genome sequenced was the unicellular fungus, baker’s yeast, (*Saccharomyces cerevisiae*, 12 Mb) which reportedly took ~600 researchers six years to complete (Goffeau *et al.* 1996). Until the late 2000’s most genome projects exclusively utilised automated Sanger sequencing (Metzker 2010), which involved the cloning or PCR purification of large DNA fragments followed by labelling of the fragments with fluorescent dideoxynucleotides and separation using capillary electrophoresis with an ABI 3730 xl (Applied Biosystems). The four bases were read separately to produce reads that were on average 700-1,000 bp in length which were then assembled by overlap into contigs. The prohibitive cost of this technique and the degree of data analysis involved meant that most genome analyses was done by consortiums of laboratories, except for species with relatively small genomes, mostly bacteria.

Genome sequencing techniques have changed significantly since the introduction in the mid-2000’s of what is often described as “next-generation sequencing” (NGS), which includes technologies developed by several companies that enable the large scale automation of fragment preparation and parallel sequencing of large numbers of short-fragments simultaneously. The most prevalent commercially available next-generation
instruments include the Illumina Genome Analyzer (GA) IIx and HiSeq 2000 (http://www.illumina.com) the Roche 454 Life Sciences Genome Sequencer FLX (http://www.454.com) and the Applied Biosystems SOLiD (http://www.appliedbiosystems.com) (Metzker 2010). Read lengths for NGS platforms range from 26 to over 400 base pairs, with the volume of data produced per run in the hundreds of megabases to well over a gigabase range for both single fragment and paired end configurations (Ye et al. 2011). The insert size (fragment of DNA from which ends only are sequenced) can also vary upwards from 200 bp with current protocols allowing up to 20 kb mate pairs (Ye et al. 2011). By 2007, there were 170 short read genomes in NCBI, mostly bacterial, and this number has been steadily increasing (Pop and Salzberg 2008).

These technologies have developed at a rapid rate with continuing increases in yield and obtainable sequence lengths coupled with corresponding decreases in cost and run time, making de novo genome sequencing an affordable prospect for many species as the dollars per megabase cost decreases (Sboner et al. 2011). On the other hand, new challenges have arisen, such as the difficulty of assembling the millions of short reads produced which currently requires considerable computational knowledge and resources. Assembly from multi-nucleate or heterozygous species still remains especially challenging and thus far next generation sequencing methods have mostly been used for de novo assembly of haploid genomes or re-sequencing projects. A potential solution for some of these issues is on the horizon with the advent of single molecule sequencing, that do not require prior amplification of the sequencing template. Commercial developments include Pacbio RS (Pacific Biosciences), Ion torrent (Life Technologies), HeliScope (Helicos Biosciences) and MinION (Oxford Nanopore Technologies). These methods require considerably less template DNA and do not require PCR, which can introduce mutations or clonal biases in amplified templates that can masquerade as sequence variants (Metzker 2010). These third generation sequencing technologies tend to produce vastly longer read lengths, in the kilobase range, some having the potential to eventually sequence telomere to telomere for individual chromosomes. For the span of this study however, sequences of these lengths were not commercially available.
1.4.1 Currently available fungal whole-genome sequences

By the year 2013 many species across the evolutionary spectrum have been sequenced including many fungi, although relatively few have been “finished” i.e. each chromosome sequenced from telomere to telomere. As of September 2013 there were 598 ongoing fungal genome BioProjects registered on NCBI, many of which are re-sequencing projects (i.e. different isolates of the same species) and not all of which contain available data yet. A project called diARK has been set up to track all genome sequencing projects (Hammesfahr et al. 2011) and currently lists 462 completed fungal species genome projects (341 Ascomycetes, 81 basidiomycetes, 26 micosporidia and 14 others) and shows that since 2001 the most commonly utilised sequencing method for fungi is still Sanger sequencing although often technologies are used in parallel to get the depth of Illumina sequencing and the length of 454 or Sanger, which allows the assembly of larger scaffolds.

1.5 Aims and overview of this thesis

The aim of this project is to examine the way necrotrophic fungal invaders attack legume hosts and to build knowledge which will facilitate the development of more resistant crops. These efforts are intended to ultimately result in improved crop-protection strategies leading to reduced losses for farmers, decreased use of chemical fungicides and to aid in the development of resistant legume crops via marker assisted selection or genetic engineering.

The practical aims of this thesis were to sequence and assemble the genome of *P. medicaginis* var. *medicaginis* isolate OMT5 (PmedOMT5) and predict its gene content. The predicted gene content was compared with expressed transcripts and proteins to create a set of biologically supported gene models, from which gene functions could be inferred, and utilised to predict which genes are likely to play a direct role in causing disease on the host *M. truncatula*.

1.5.1 Thesis outline

Chapter 2 describes the general materials and methods used for growing plants and fungi, performing infections, microscopic analysis and standard bioinformatic analyses. Chapter 3 outlines assembly of the PmedOMT5 genome and an analysis of its genomic landscape, including its predicted protein-coding genes and non-coding features such as
repeats and non-coding RNA. Chapter 4 describes the separate assembly and annotation of the mitochondrial genome. Chapter 5 outlines the methods used to validate the bioinformatically-predicted gene annotations using RNA-seq and supporting proteomic data from several different life-cycle stages of the fungus including vegetative growth, sporulation, production of phytotoxic compounds, as well as, at the early stages of infection on *M. truncatula*. In chapter 6, the comparative analysis of the RNA-seq data outlined in chapter 5 is analysed in detail to elucidate potential pathogenicity mechanisms of *P. medicaginis* on *M. truncatula*. Chapter 7 provides preliminary insights into the extracellular proteome of PmedOMT5. The major findings and outcomes of these chapters are summarised in the eighth chapter and literature sources are listed in chapter 9.
Chapter 2
General Materials and Methods
2.1 General methods and procedures
All autoclaving was carried out at 121 °C for 20 min unless otherwise specified. The preparation of agarose and polyacrylamide gels, general buffers and solutions was conducted as described in (Sambrook et al. 1989).

2.1.1 General media used
Media recipes are provides in appendix 2.1.

- PDA: potato dextrose agar
- WMA: wheat meal agar
- MM: minimal media
- F2NY: Fries 2 media without yeast
- Fries media: modified from Liu et al. 2004
- TWA: tap water agar
- TE: Tris-EDTA for solubilisation and storage of DNA
- Tris-HCl: for solubilisation of proteins

2.1.2 Seed sources
All M. truncatula ssp. truncatula seeds tested were at least second generation inbred lines derived from seed obtained from the South Australian Research and Development Institute (SARDI) except R108-IC3 (ssp. tricycla) which was obtained via the Commonwealth Scientific and Industrial Research Organisation (CSIRO). M. truncatula accessions used in this study: SA10481, SA7749, SA18543, SA23859, SA28645, SA27063, SA1489, SA3054, SA8623, A17, Borung, Caliph, DZA045.5, DZA315.16, F83005.5, R108-IC3, SA8604, SA11734, SA3047, SA28375 and Sickle (A17- EMS mutant, Penmesta and Cook 1997). M. sativa seeds from accessions: 36325 (AUS), 35043 (AUS), 38082 (AUS), 36442 (AUS), 10119 (USA) were obtained from AMGRC, South Australia. M. sativa Sceptre seeds were was obtained commercially. Most other legume seeds of commercial cultivars were obtained from local seed suppliers (Table 2.1).
Table 2.1: Non-Medicago seed sources.

<table>
<thead>
<tr>
<th>Plant species</th>
<th>Cultivar/Accession/Ecotype</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pea (P. sativum)</td>
<td>Kaspa, Earlicrop Massey, parafield, Greenfeast</td>
<td>Commercial</td>
</tr>
<tr>
<td>Dwarf bean</td>
<td>Borlotti, Bountiful butter</td>
<td>Commercial</td>
</tr>
<tr>
<td>Sub clover</td>
<td>Dalkeith</td>
<td></td>
</tr>
<tr>
<td>Faba bean</td>
<td>Fiord</td>
<td>Commercial</td>
</tr>
<tr>
<td>Mung bean</td>
<td>unknown</td>
<td>Commercial</td>
</tr>
<tr>
<td>Lupin (Lupinus albus)</td>
<td>P27174, Kiev Mutant</td>
<td>Kiev DAFWA</td>
</tr>
<tr>
<td>Narrow-leaf lupin (L. angustifolius)</td>
<td>Unicrop, Tanjil</td>
<td>Unicrop commercial Sona</td>
</tr>
<tr>
<td>Chickpea (Cicer arietinum)</td>
<td></td>
<td>Commercial</td>
</tr>
<tr>
<td>Lentil (Lens culinaris)</td>
<td>Digger, Northfield</td>
<td>Digger commercial Sona</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>Col-0</td>
<td>CSIRO, Dr. J. Anderson</td>
</tr>
<tr>
<td>Lotus japonicus</td>
<td>Gifu</td>
<td>CSIRO, Dr. J. Anderson</td>
</tr>
<tr>
<td>Wheat (Triticum aestivum)</td>
<td>Zippy</td>
<td>DAFWA</td>
</tr>
<tr>
<td>Barley</td>
<td>Flagship</td>
<td>DAFWA</td>
</tr>
</tbody>
</table>

2.1.3 Seed germination

Seeds of inbred *M. truncatula* accessions from the SARDI collection were de-podded then scarified in a mortar with sand. Seeds were surface sterilised by shaking in a mixture of 1% sodium hypochlorite and 5% ethanol for 5 min and rinsed three times with sterilised water. Seeds were then placed on damp filter paper in a petri dish, wrapped in foil and germinated at room temperature (RT) overnight. This was followed by vernalisation for 2-7 days at 4 ºC then germinated seeds were allowed to elongate at room temperature for 24 h.

Germinated seeds were planted in autoclaved vermiculite (The Perlite and Vermiculite company, Jandakot), for growth chamber experiments or sterilized soil mix for glasshouse experiments, seed replication or fungal re-isolation. Seeds of various legumes (listed in Table 2.1) were germinated in petri dishes on moist filter paper in the dark at RT. When a radical had emerged they were planted in sterilized soil in a temperature controlled glasshouse.

2.1.4 Seed production and storage

Plants for seed production were propagated in pasteurised soil in a temperature-controlled glasshouse with cooling only, set to 24 °C and treated regularly with
fungicide and insecticide when necessary. Halogen lamps were used to extend day length during winter. Seeds were collected in net bags placed over the whole plant. Collected seeds were dried at 37 °C for up to 2 weeks and stored as whole pods at RT, in the dark. Seeds were stored in plastic jars with air holes or in unsealed paper envelopes (Garcia et al. 2006).

2.1.5 Experimental plant conditions – Glasshouse

*M. truncatula* plants for fungal inoculation and infiltration were grown in a temperature-controlled glasshouse with cooling only at 24 °C in sterilised vermiculite. Plants in vermiculite were fertilized once weekly with ¼ strength “Grow A” and “Grow B” fertiliser solutions and “Bloom A” and “Bloom B” during flowering (Growth Technology, O’Connor, WA).

2.1.6 Experimental plant conditions – Growth chamber

Plants were grown under 12-16h light at 22 °C, 16 °C overnight and constant 60% humidity. Humidity was increased to a constant 90% following inoculation. Plants were spray-inoculated as for glasshouse experiments described in section 2.2.2 or spot inoculated with 10 µL droplets of 1x10^6 sp/mL containing 0.05% Tween 20 (section 2.2.3).

2.1.7 Fungal isolates

*P. medicaginis* var. *medicaginis* isolate OMT5 was isolated from naturally infected wild *M. truncatula* collected by Prof. R. P. Oliver in Perth, Australia, 2000 (Ellwood et al. 2006b) as was isolate OMT1. Mono-conidial cultures were stored at -80 °C at a concentration of 1 x10^6 sp/ml, in 25% glycerol. Isolates WAC4736 and WAC4741 were collected in Gnowangerup, WA, Australia in 1986 and obtained from the Department of Agriculture and Food, Western Australia, (DAFWA) 3 Baron Hay Court, Perth, Western Australia, 6151. As fungal cultures can lose their virulence when sub-cultured in vitro, these three isolates were periodically infected onto *M. truncatula* and re-isolated from the plants in order to maintain the pathogenicity of the glycerol stock cultures.

All other *P. medicaginis* isolates used for mating type testing were recovered from infected *M. truncatula* in Western Australia (WAC4738, WAC7977 and WAC7980)
were obtained from DAFWA. *P. medicaginis* var. *medicaginis* CBS316.90 and *P. pinodella* CBS318.90 standard reference cultures were acquired from Centraalbureau voor Schimmelcultures (P.O. Box 85167, 3508 AD Utrecht, The Netherlands). Other species used in this study were obtained from DAFWA (*P. pinodella* WAC7978), or the Australian Centre for Necrotrophic fungal pathogens (ACNFP) (M07-4, Me14). The host, location and original identification for each isolate is presented in Table 2. A representative culture of OMT5 is held by DAFWA (Ellwood et al. 2006b).

2.1.8 Cultivation of fungal isolates for plant inoculation

Fungal cultures were maintained on WMA or ½ PDA plates. Plates were inoculated with ~50 µL of spore suspension and 50 µL of sterile Milli Q water, spread with a glass spreader and incubated at 20 ±2 °C with a 12h UV light cycle for 4-6 weeks, until the production of sufficient pycnidia for harvesting.

2.1.9 Harvesting of conidiospores

Conidia were harvested by flooding the plates with sterile water and incubating at RT for ~10 min. The surface of the plate was scraped gently to release the spores and the resulting liquid filtered through a glass-wool syringe or three layers of sterile milk-filters for larger volumes. Concentration of conidia in the filtrate was estimated using a haemocytometer. The filtrate was then diluted to the appropriate concentration for subsequent experiments, usually 1 x 10⁶ sp/mL.

2.1.10 Re-isolation of pathogenic fungi from infected plants

Infected tissue was surface sterilised for 5 min in a mixture of 5% ethanol and 1% sodium hypochlorite, rinsed three times with sterile Milli-Q water and plated onto TWA. Plates were placed at 24 °C, 12h light in a tissue culture room until fungal hyphae extended from the plant tissue. Hyphal tips were collected using a sterile wire loop and inoculated onto PDA plates, which were grown under the same conditions until sporulation. Spores were collected and diluted to produce single spore colonies which were re-isolated and stored at -80 °C in 25% glycerol.
Table 2.2: Original identification, host and location isolated for fungal pathogens used in this study (adapted from Kamphuis et al., 2008).

<table>
<thead>
<tr>
<th>Isolate</th>
<th>Host isolate was found on</th>
<th>Date; Collector; Location</th>
<th>Original identification</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBS316.90</td>
<td><em>M. sativa</em></td>
<td>1990; Noordeloos, M; Czechoslovakia</td>
<td><em>P. medicaginis var. medicaginis</em></td>
</tr>
<tr>
<td>CBS318.90</td>
<td><em>P. sativum</em></td>
<td>1990; Noordeloos, M; Netherlands</td>
<td><em>Phoma</em> (<em>Peyronellaea</em>) <em>pinodella</em></td>
</tr>
<tr>
<td>OMT1</td>
<td><em>M. truncatula</em></td>
<td>2000; Oliver, R; Perth WA</td>
<td><em>P. medicaginis</em></td>
</tr>
<tr>
<td>OMT5</td>
<td><em>M. truncatula</em></td>
<td>2000; Oliver, R; Perth WA</td>
<td><em>P. medicaginis</em></td>
</tr>
<tr>
<td>WAC 4736</td>
<td><em>M. truncatula</em></td>
<td>1986; Barbetti, M; Gnowangerup, WA</td>
<td><em>P. medicaginis</em></td>
</tr>
<tr>
<td>WAC 4738</td>
<td><em>M. truncatula</em></td>
<td>1986; Barbetti, M; Gnowangerup, WA</td>
<td><em>P. medicaginis</em></td>
</tr>
<tr>
<td>WAC 4741</td>
<td><em>M. truncatula</em></td>
<td>1986; Barbetti, M; Gnowangerup, WA</td>
<td><em>P. medicaginis</em></td>
</tr>
<tr>
<td>WAC 7977</td>
<td><em>M. truncatula</em></td>
<td>1981; Barbetti, M; Perth, WA</td>
<td><em>P. medicaginis</em></td>
</tr>
<tr>
<td>WAC 7978</td>
<td><em>M. truncatula</em></td>
<td>1981; Barbetti, M; Perth, WA</td>
<td><em>P. medicaginis</em></td>
</tr>
<tr>
<td>WAC 7980</td>
<td><em>M. truncatula</em></td>
<td>1981; Barbetti, M; Perth, WA</td>
<td><em>P. medicaginis</em></td>
</tr>
<tr>
<td>M07</td>
<td><em>P. sativum</em></td>
<td>2007, J. Lichtenzveig, Merredin, WA</td>
<td><em>Peyronellaea</em> <em>pinodes</em></td>
</tr>
<tr>
<td>Me14</td>
<td><em>Cicer arietinum</em></td>
<td>T. Pfaff, Merredin, WA</td>
<td><em>Didymella rabiei</em></td>
</tr>
</tbody>
</table>

2.2 Infection methods

2.2.1 *P. medicaginis* inoculation

Plants were infected when the third (growth chamber experiments) or fourth (glasshouse experiments) trifoliate leaflet was fully emerged. Conidia were harvested as described in above and diluted to $1.0 \pm 0.5 \times 10^6$ spores/mL.

2.2.2 Spray inoculation modified from (Kamphuis et al. 2008)

Plants were sprayed with a $1.0 \pm 0.5 \times 106$ sp/mL solution containing 0.05% Tween 20 as a surfactant, using an artists airbrush (Paasche Airbrush Co., Harwood Heights, IL,
USA) to just before runoff (approximately 1.5 mL per plant) using a rotating platform to allow even distribution of spores. Control plants were mock-inoculated with 0.05% Tween 20. Plants were incubated under plastic covers in the glasshouse for 72 h to maintain high humidity.

2.2.3 Spot inoculation as described in (Kamphuis et al. 2008).

Plants were spot inoculated with 10 µL of spore suspension (1x10^6 sp/mL concentration 0.05% Tween 20) onto the monofoil, middle and one lateral leaflet of the first and second trifoliate leaves. Control plants were mock-inoculated with 0.05% Tween 20. Plants were incubated under plastic covers in the glasshouse for 72 h to maintain high humidity.

2.2.4 Detached leaf assays

Freshly sectioned trifoliate *M. truncatula* leaves were placed on moist filter paper in petri dishes in a growth room at 22 °C, under 12 h light.

Detached leaf assays were performed by inoculating each trefoil with a 10 µl droplet of 1 x 10^6 sp/mL in 0.05% Tween20.

2.2.5 Scoring

Following spot inoculation infection was scored on a 0-5 scale at 7 dpi and re-scored at 10-14 dpi to confirm as described previously by (Salter and Leath 1992, Kamphuis et al. 2008). Individual accession’s disease scores were averaged for data comparisons. The statistical software package JMP (version 5.1 or 7; SAS Institute, Cary, NC) was used to compare means using the Tukey-Kramer honestly significant difference test at \( p \leq 0.05 \).

Following spray inoculation, infection was scored at 7 and 10 days post inoculation using a 5 point scale (0-5).

2.2.6 Harvesting of plant tissue

At the desired time-points seedlings were cleaned of the growth matrix and where necessary were sectioned for tissue-specific studies. Extracted tissues were frozen immediately in liquid nitrogen and stored at – 80 °C until analysis.
2.3 Microscopy and staining for visualisation of fungal cells

2.3.1 Imaging

All light microscopy was performed using an Olympus BH-2 compound microscope fitted with Olympus DP12 image acquisition hardware. Samples were illuminated using white light or UV with the appropriate filters: blue for eGFP (WB filter U-MWB2, 460-490 nm excitation filter and a 520 nm barrier filter) or green for dTomato (WG filter U-MWG2, excitation 510-550nm, emission 590nm DM 570).

Confocal microscopy was performed using a LaserSharp Confocal System (BioRad) combined with a Nikon inverted microscope. GFP images were collected by exciting at 488 nm and collecting at both 522 and 605 nm. Optical sections were typically collected over a depth of 40 µm using 1-5 µm sections.

All photographs were taken with a Nikon Coolpix 995.

2.3.2 Staining for microscopy

Hyphae on the surface of the leaf were stained with either Calcofluor white (stains cellulose, chitin and callose) or 3,3′-dihexyloxacarbocyanine iodide (DiOC6(3)) (shows cytoplasm by staining the endoplasmic reticulum, vesicle membranes and mitochondria) to improve visualisation under UV light. Hyphae within the leaf were visualised using trypan blue and bright field microscopy or confocal microscopy of a GFP-transformant (appendix 2.2)

Calcofluor white - Leaves were placed in a solution of 0.1% Calcofluor white (Fluorescent Brightner 28 Sigma-Aldrich, Australia) dissolved in 0.1 M Tris-HCl pH 8.3, for 5 mins on an orbital shaker in the dark. Leaves were then rinsed in water for 15 mins in the same manner and examined under UV light where hyphae appear blue.

DiOC6(3) - Samples were stained by immersion in 50 µg/mL solution in water (prepared from a 0.5 mg/mL stock solution in 100% ethanol) for 5 min on an orbital shaker in the dark. Leaves were then rinsed in water for 5-15 mins in the same manner. Fungal cells fluoresced green when excited with blue light and appeared yellow under UV light.

Trypan blue staining - Leaves were cleared overnight in Farmers fluid (100mL acetic acid, 30 mL chloroform, 60ml ethanol) then stained in 1:4 v/v trypan blue solution
(Fluka) by boiling for 5 min and cooling to RT. Leaves were then de-stained in 100% ethanol or 2:1 ethanol: lactophenol and visualised via bright field microscopy.

2.4 DNA analysis

2.4.1 DNA extraction

All material harvested for DNA extraction was immediately frozen in liquid nitrogen and stored at -80 °C until processing. Tissue was ground by hand in a cooled, sterile mortar and pestle.

Unless otherwise specified, fungal genomic DNA was extracted using a QIAGEN BioSprint 15 (Thermo Electron Corporation, Finland) according to the manufacturer’s directions (Purification of DNA from Plant Tissue- BioSprint DNA Plant Handbook 03/2005 p.15-19).

For more difficult extractions or where larger amounts of DNA were required a CTAB extraction protocol was used, modified from Cook et al. (1990). Frozen plant/fungal material was ground in liquid nitrogen and mixed with 700 µl pre-heated CTAB extraction buffer (2% w/v CTAB, 1.4 M sodium chloride (NaCl), 0.2% w/v β-mercaptoethanol, 20 mM EDTA, 100 mM tris-HCl) at 60°C and incubated for 30 min. An equal volume of chloroform/isoamyl-alcohol (24:1) was added and mixed gently for 5 min before centrifugation at 12,000 rpm for 20 min. The supernatant was removed to a new eppendorf tube and treated with RNase (Sigma –Aldrich, Australia) 20 U/mL at 37°C for 30 min. Chloroform/isoamyl extraction was repeated until the aqueous phase was clear.

Supernatant was transferred to a new tube and the DNA precipitated with 0.1 volume sodium acetate (3M, pH 5.2) and 0.6 volumes isopropanol at -20 °C for 30 min before centrifugation at 12,000 rpm for 10 min. The supernatant was discarded and the DNA pellet washed with 500 µL 70 % ethanol before air drying and re-suspension in 100 µL sterile TE (pH 8.0).

After extraction the concentration of all DNA was determined spectrophotometrically and tested for integrity by running on a 1% agarose gel.
2.4.2 PCR

The general polymerase chain reaction (PCR) protocol used throughout this thesis involved standard reagents of 50-100 ng of genomic DNA template, 1 x PCR reaction buffer (50 mM KCl, 10 mM Tris-HCl, 1.5 mM MgCl₂, 0.1% Triton-X), 0.25 mM of each dNTP, 10 pM of each primer and 1 U of Taq polymerase. Thermocycling conditions of 2-5 min denaturation at 94 °C, followed by 35 cycles of: 94 °C for 30 s, 50-60 °C for 30 s annealing (specific to primer-pairs), 72 °C for 30-120 s extension (dependent upon the length of the desired amplicon) and a final extension step of 7 min at 72 °C. All PCRs were carried out on an Eppendorf thermocycler.

2.4.3 PCR for products greater than 1 Kb

For longer products iProof taq (BIO-RAD, USA) was substituted for standard Taq polymerase.

1x PCR reaction: 0.5 reaction volume of 2x iProof HF mastermix containing 0.04 U/µL Pyrococcus-like proofreading enzyme fused to a dsDNA binding protein, Sso7d. (BIO-RAD Laboratories, Hercules, Ca.,USA), 5pM of each primer, 100 ng of DNA.

Thermocycling conditions of 30 s denaturation at 98 °C, followed by 35 cycles of: 98 °C for 5 s, 50-60 °C for 30 s annealing (specific to primer-pairs), 72 °C for 30 s – 2 min extension (dependent upon the length of the desired amplicon) and a final extension step of 10 min at 72 °C.

2.5 Bioinformatics: general methods

Bioinformatic analysis was carried out via a local Linux computer (running Ubuntu) or with the assistance of resources provided by the High Performance Computing Cluster XE at Murdoch University and the National Compute Infrastructure specialised facility in Bioinformatics (NCI-SF Bioinformatics) High Performance Computing Cluster (Barrine) at the University of Queensland, provided through the National Computational Merit Allocation Scheme supported by the Australian Government. All data was stored on the iVEC Petascale Data Store at Bentley, WA.

2.5.1 Basic Local Alignment Search Tool

BLAST (Altschul et al. 1990) searches were conducted via the NCBI website, the NCI-SF Bioinformatics facility or a local copy (version 2.2.25). Unless otherwise specified
BLAST searches were done at a significance score of $1e^{-5}$, which implies a 1 in 10,000 chance of a random match. Routine analysis included BLASTP (protein query to protein database search), BLASTN (nucleotide to nucleotide), TBLASTN (protein to nucleotide), TBLASTX (translated nucleotide to translated nucleotide) and BLASTX (translated nucleotide to protein).

### 2.5.2 Visualisation

Sequence alignments were visualised in GenomeView (Abeel et al. 2011), IGV (Thorvaldsdóttir et al. 2013) or Apolllo (Lewis et al. 2002).

Proportional Venn diagrams were drawn with the Applet For Drawing 3 Set Area-Proportional Venn Diagrams (Chow and Rodgers 2005).

### 2.5.3 Statistical analysis

All statistical analysis was performed using JMP v7 from SAS Institute, North Carolina.
Chapter 3
Genome assembly and analysis of the genomic landscape
3.1 Introduction

Following its advent in 1977 (Sanger et al. 1977) nucleotide sequencing has become a standard tool in the biologist’s toolkit. It is routinely employed to investigate diverse biological questions about the evolution and organisation of the genomic landscape. Sequencing coupled with comparative genomics allows investigation into the conservation, alteration and transfer of genes between species providing a window to their evolutionary history.

Whole genome sequencing has rapidly advanced the body of knowledge surrounding phytopathogenic fungal species and opened new avenues for the discovery of pathogenicity effectors. The increasing speed, availability and decreasing cost of new sequencing technologies coupled with an expansion in analytical bioinformatic techniques, has led to rapid advances in our understanding of fungal pathogenicity arsenals. Previous studies have identified novel pathogenicity factors, host specificity factors and identified candidates of horizontal transfer that can either confer or enhance the virulence of the receiving isolates (Friesen et al. 2006, Hane et al. 2007, Ma et al. 2010).

The first step in these investigations is assembling an approximation of the genomic organisation from the fragmented sequences obtained. Whilst a genome assembly in and of itself tells little about an organism’s pathogenicity strategies, coupling this data with biological studies can aid in the rapid and accurate identification of the transcribed genes and translated proteins of the species of interest and offer insights into gene product function and regulation under different conditions.

To advance the knowledge of pathogenicity mechanisms employed by P. medicaginis, sequencing of the genome of isolate OMT5 (PmedOMT5) (Ellwood et al. 2006b), was commissioned utilising the next-generation sequencing (NGS) technologies developed by Solexa/Illumina. As a pathogen of the model legume species M. truncatula and belonging to a class of fungi known to produce host-specific effectors, the main goal of genomic sequencing, of PmedOMT5, apart from creating a reference assembly, was to examine the pathogenicity mechanisms employed by fungi against legume hosts.

P. medicaginis is thought to reproduce asexually, (Boerema et al. 2004, Aveskamp et al. 2010), thus information about its genomic structure cannot be elucidated by common population mapping techniques and as the PmedOMT5 genome is haploid, assembly
was anticipated to be relatively simple using currently available techniques. It has been previously demonstrated that the majority of the genes encoded by a fungal pathogen, with a simple genome structure, can be identified even from an unfinished, fragmented genome assembly (Ellwood et al. 2010).

3.1.1 Short read sequencing and data analysis strategies

3.1.1.1 The Illumina/Solexa sequencing method and introduced errors

The Illumina Genome Analyser was introduced in 2006 and is based on the concept of “sequencing by synthesis” producing sequence reads of up to 150 bp from tens of millions of simultaneously surface amplified DNA fragments. The Illumina platform is currently the most popular choice for genomic sequencing (Metzker 2010, Hammesfahr et al. 2011) due to its relatively low cost and high yield. A run of 75 cycles can take 3-9 days (for a GAII paired end run) and typically yields 40-50 million sequences of paired 75 bp reads after quality filtering (Mardis 2008). For paired-end reads, after the first read is sequenced the forward product strand is removed and the complementary strand re-synthesised and sequenced in the same manner.

All sequencing technologies have the potential to introduce errors into the sequence data. There are several types of errors observed in Illumina data. Substitutions are the most common, with a higher proportion following a G base (Metzker 2010). Other errors occur randomly caused by ‘phasing’ or ‘pre-phasing’ of individual sequence clusters (Metzker 2010) resulting in skipped bases, although there are typically a low number of indel errors (<0.1%) (Dohm et al. 2008) especially in comparison to 454 sequencing where indels are prominent due to the pyrosequencing method used (Nagarajan and Pop 2010). These errors can often be corrected by other reads overlapping the same sequence due to the large number of reads Illumina sequencers can produce (up to 3 billion paired-end reads per HiSeq 1000 flow cell, 270-300 Gb (Illumina 2011). Dohm et al., (2008) found that 20-fold coverage is sufficient to compensate for sequencer induced errors with correct reads.

An under-representation of sequences from A:T and G:C rich regions has been observed in Illumina data, putatively caused by an amplification bias during template preparation (Dohm et al. 2008, Hillier et al. 2008, Metzker 2010). There is also a marked tendency for the quality score to decline at the end of longer reads (over 50bp). This can be
ameliorated post-sequencing by bioinformatically trimming bases below a desired quality threshold. Dohm et al., observed error rates ranging from 0.3% incorrect calling of the first base at the beginning of a read increasing to 3.8% at the end (Dohm 2008).

3.1.1.2 Genome assembly from NGS data via manipulation of de Bruijn graphs

From the millions of reads generated via NGS, researchers must create an assembly that forms a set of contiguous sequences (contigs) that represent the alignment of nucleotides on the DNA strands (Zerbino and Birney 2008). Previous technologies developed for the alignment of Sanger reads involved creating an overlap consensus however the short read lengths coupled with the enormous volume of data generated by Illumina sequencing are not suitable for this approach. In the past few years a plethora of new techniques has been developed to tackle the issue. As of March 2011, there were 24 de novo short read genome assembly algorithms available under academic licence, most using either string or graph based models (Zhang et al. 2011). Many of these new assemblers use a technique for reconstructing a path through the short reads known as a deBruijn graph (Nagarajan and Pop 2010). The main rationale behind the application of deBruijn graphs is an efficient use of computational memory to handle the large number of partially repetitive sequence reads (as each part of the genome is sequenced many times from many different fragments).

Briefly, a deBruijn graph assembler starts by decomposing the set of reads into a set of shorter segments of length $k$ nucleotides ($k$-mers). A graph is constructed which contains the $k$-mers segments as nodes and represents overlaps as edges. Two nodes are connected if they overlap in one of the original reads (Pop and Salzberg 2008). In an ideal world a correct reconstruction of the genome is represented as an Eulerian path through this graph that traverses all the edges only once, although in practice the nodes become highly interconnected when repetitive regions of DNA or reads errors are encountered. Fragmenting the original reads into smaller segments, allows an easier solution for calculation of this path, allows reads of different lengths to be processed together and also allows the high redundancy in the sequence obtained to be handled by the graph without affecting the number of nodes (Pop and Salzberg 2008).

Irrespective of genome size, de novo assemblies constructed from short-read data are highly fragmented (Pop 2009, Simpson et al. 2009). One of the major problems for NGS assemblies using deBruijn graphs is assembling repetitive regions of the genome
(Treangen and Salzberg 2011). Any overlaps between sequences are correspondingly short and many reads in repeated genomic regions will have only single or no base differences (Zerbino and Birney 2008). Repeats may be represented by a single or a reduced number of nodes in the deBrujin graph with multiple links to other non-repetitive sequences. A high number of connections between nodes in the graph caused by the repeats prevents them being resolved into a single Eulerian path and leads to a fragmented assembly (Zerbino and Birney 2008).

At the time of assembly of the PmedOMT5 genome the only non-bacterial published genomes assembled purely from Illumina short reads were the fungus *Pyrenophora teres f. teres* (Ellwood et al. 2010, Li et al. 2010) and Panda (*Ailuropoda melanoleuca*) (Li et al. 2010), although several genomes have since been assembled from the combination of 454 and Illumina data *e.g.* chicken (Ye et al. 2011) and the fungus *Sordaria macrospora* (Nowrousian et al. 2010). Both of these studies demonstrate high accuracy for the Illumina based assemblies when compared to a reference Sanger sequenced genome assembly or Sanger sequenced BACs. There is currently a multitude of Illumina only re-sequencing projects in which new isolates of species with a reference genome assembled using other methods are sequenced for comparative analysis.

### 3.1.1.3 Measuring completeness of a genome assembly

As there are very few finished (telomere to telomere) genome sequences available, several metrics have been identified to enable comparison between assemblies including N50 and L50 (also known as N50 length). If the assembled contigs are ranked in size order from largest to smallest, N50 describes the number of contigs which contain more than 50% of the assembled nucleotides and L50 the length for which 50% of all bases in the assembly are in a contigs or scaffold of at least that length (Parra et al. 2007, Miller et al. 2010). There is still discussion amongst the bioinformatics community about whether these are the best measures of an assembly. Others factors must also be taken into account such as average contig length, maximum contig length and mate constraint satisfaction *i.e.* do mate-paired reads map to the assembly the correct distance apart and in the correct orientation? (Phillippy et al. 2008, Earl et al. 2011). Although metrics such as N50 are useful for describing the nucleotide content of a genome project, they
do not describe the state of the gene space but this can be assessed by comparison to a set of core eukaryotic genes (Parra et al. 2007).

3.1.2 Features of fungal genome structure

3.1.2.1 Number and size of chromosomes

The number of chromosomes in a fungal species varies widely even between strains and can also alter during different life cycle stages, with observations ranging from 3 in haploid *S. pombe* (Wood et al. 2002) to greater than 50 in the polyploid Blastocladiales mycete *Alomyces macrognus* (Emerson and Wilson 1954). Some fungi can show extreme genome plasticity with chromosomes gained, lost, split or joined during meiosis as observed in the Dothideomycete *M. graminicola* (Goodwin et al. 2011). Fungi may also possess dispensable, supernumerary or conditionally-dispensable chromosomes (CDCs), which are not essential for growth or sex (Coleman et al. 2009, Ma et al. 2010) but may confer an evolutionary advantage. CDCs have been observed in three Dothideomycetes: *Alternaria alternata*, *L. maculans* and *M. graminicola* and can play important roles in pathogenicity (Akagi et al. 2009, Goodwin et al. 2011, Rouxel et al. 2011). They tend to be highly repetitive and gene sparse with a high A:T content.

The number of chromosomes that make up the PmedOMT5 genome is unknown, but other Dothideomycetes range from 9-11 in *Altenaria brassicicola* to 21 in *M. graminicola*, with 14-19 observed in *P. nodorum* by electrophoretic karyotyping (Cooley and Caten 1991) and 12-16 in its close relative *D. rabiei* (Akamatsu et al. 2012). Chromosome size in fungi can range from 0.27 Mb in *Colletotrichum gloeosporioides* to 12.6 Mb in *N. crassa* (Zolan 1995). Chromosome size in the Dothideomycete fungus *L. maculans* ranges from 0.7 to 3.7 Mb and other Dothideomycetes appear to follow a similar pattern (Zolan 1995).

All fungi contain mitochondrial organelles present in the cytoplasm which contain the mitochondrial genome. Scaffolds originating from the mitochondrial genome were identified, assembled and analysed separately from the nuclear scaffolds and are discussed in chapter 4.
3.1.2.2 Repetitive DNA in fungi

Repetitive DNA elements are found ubiquitously in all fungal genomes and have been classified as: microsatellites or simple repeats (20-60 bp with tandem repetitions of 1-5 nucleotides), tandem repeats (more complex, 100-200 bp sequences typically found around centromeres and telomeres of chromosomes), minisatellites (10-60 bp motifs amplified to lengths of 0.1-3 kbp varying in length but often with conserved positions), transposons (mobile elements) and retrotransposons (Wöstemeyer and Kreibich 2002, Jurka et al. 2005). Transposons are classified into class I (which possess their own reverse transcriptase and transpose via RNA intermediates) and class II (which transpose directly at the DNA level) (Wöstemeyer and Kreibich 2002).

Transposable elements (TEs) have been identified in all prokaryotic and eukaryotic life-forms sequenced to date (Kempken and Kück 1998). Unlike normal genes they are able to change their position within their host genome. The subsequent transposition into coding sequences and their initiation of chromosome rearrangements can have a tremendous impact on gene expression and genome evolution (Kempken and Kück 1998).

3.1.2.3 Repeat-induced Point Mutation (RIP) in repetitive DNA

First identified in N. crassa (Selker et al. 1987, Selker 1990), repeat-induced point mutation (RIP) is a mechanism employed by some fungi which alters the sequence of repeated regions in the genome prior to the sexual cycle. It often results in the introduction of stop codons into repetitive sequences due to the introduction of C:G to T:A transition mutations in both strands of a repeat. RIP is noted to occur in sequences over 400 bp sharing greater than 80% identity in N. crassa (Cambareri et al. 1991, Watters et al. 1999). In practice this often results in the inactivation of transposon genes (Watters et al. 1999) as they occur in multiple copies and as RIP tends to introduce stop codons which inactivates transposases, littering the genome sequence with non-functional repeat relics. In M. graminicola all transposons with more than 10 copies contained stop codons in their coding regions indicating they had been inactivated (Goodwin et al. 2011).

The presence of transposons can significantly impact the evolution of genes neighbouring the insertion point. In L. maculans (a sexual Dothideomycete) RIP like
mutations were identified in some single copy genes (notably in two pathogenicity effector genes \textit{AvrLm1} and \textit{AvrLm6}), in proportion to their proximity to RIP degenerated transposable elements (Van de Wouw \textit{et al.} 2010). This phenomenon of “RIP leakage” where the mechanism responsible for creating RIP in repetitive sequence appears to overrun into non-repetitive genomic regions creating the potential for increased mutation, has also been observed in \textit{N. crassa} up to 930 bp from the border of a nearby duplicated sequence (Irelan \textit{et al.} 1994).

3.1.3 Genome sequences available from the class Dothideomycetes order Pleosporales

At the time of initiating the genome analysis described herein whole genome sequences had been published for several species in the order Pleosporales, including \textit{P. nodorum} (Hane \textit{et al.} 2007), \textit{L. maculans} (Rouxel \textit{et al.} 2011) and \textit{P. teres f. teres} (Ellwood \textit{et al.} 2010). The genome of \textit{P. nodorum} (isolate SN15), which was sequenced by the BROAD institute and analysed by the ACNFP in 2005 is the most well characterised closely related genome to \textit{P. medicaginis} with protein and transcript expression support (microarray and EST) for the majority of the currently annotated gene set (version 3) (Bringans \textit{et al.} 2009, Ipcho \textit{et al.} 2011). For these reasons SN15 has been used as a reference, along with the genome assemblies of other sequenced members of the class Dothideomycetes and division Ascomycota, throughout the assembly and analysis phase for comparative genomic studies.

3.1.4 Previously existing genomic resources for \textit{P. medicaginis}

Prior to the sequencing of the genome of isolate OMT5 there were only 11 publically available \textit{P. medicaginis} var. \textit{medicaginis} DNA sequences in NCBI. There were 144 \textit{P. medicaginis} sequences from various other variants and 4,510 from different \textit{Phoma} species, although as \textit{Phoma} is an amorphous genus some of these species have since been re-assigned to other taxa (Aveskamp \textit{et al.} 2010). These sequences consisted mainly of the ribosomal internal transcribed spacer region (ITS) regions from several isolates in addition to regions of the beta tubulin, actin, elongation factor 1 alpha and calmodulin genes employed in phylogenetic studies as well as a mitochondrial rDNA region and three partial polyketide synthase genes.
3.1.5 Aims for this chapter: building a genomic resource for \textit{P. medicaginis}

This chapter describes the assembly of the PmedOMT5 genome from Illumina sequencing data and outlines its genomic landscape, including \textit{ab initio} annotation of genes and identification of repetitive regions and non-coding RNA. The intent was to produce the best assembly possible with the available data which would serve as a genomic reference for subsequent \textit{Phoma} studies, with an emphasis on predicting a set of genes comprising the pathogenicity arsenal of \textit{P. medicaginis} var. \textit{medicaginis}. Therefore features of the genomic landscape not relevant to nuclear protein-encoding genes, such as mitochondrial DNA sequences, repetitive DNA regions and non-coding RNA were identified to facilitate separation of these features from the dataset of nuclear-encoded genes. This enabled subsequent efforts to focus on nuclear genic regions for pathogenicity analysis and effector prediction.

3.2 Methods

3.2.1 DNA isolation

\textit{P. medicaginis} var. \textit{medicaginis} isolate OMT5 was grown in 200 mL minimal media with 30mM glutamate for 6 days in a static 500 mL flask at RT. DNA was extracted using the CTAB protocol as described in section 2.4 with minor modifications that included 45 min RNase treatment followed by pooling of several extractions and subsequent RNase treatment. The genomic DNA sent for sequence analysis had a final concentration of 547 ng/µl in a 40 µl volume with A260/A280 ratio of 1.83.

3.2.2 Sequencing and assessment of read quality

Sequencing of 75 nucleotide, paired-end reads was performed on randomly sheared DNA fragments (~200 bp) in a single lane at the Allan Wilson Centre Genome Service (AWCGS), Massey University, Palmerston North, New Zealand using the Ilumina/Solexa Genome Analyzer (GAIIX) platform. DNA preparation, cluster formation, primer hybridisation and DNA amplification reactions were according to the manufacturer's recommended protocol (Bentley \textit{et al.} 2008). Read data was pre-filtered via Illumina's Pipeline v.1.4 and IPARv.1.3, reads failing a 'chastity' filter of 0.6 were discarded by the AWCGS.
Once the data was received, read quality was assessed using FastQCv 0.7.0 (FASTQC) and SolexaQA (Cox et al. 2010).

### 3.2.3 Genome assembly

Draft assemblies were created using ABBySS v1.2.6 (Simpson et al. 2009) and Velvet version 1.0.14 (Zerbino and Birney 2008) testing k-mers between 21 and 71, in increments of 10 then optimising between 31 and 41 in increments of two nucleotides.

After optimising for k-mer length, a Velvet assembly (k=39) was selected and further optimised by adjusting values for the parameters “expected coverage” and “coverage cutoff”. The expected coverage parameter reportedly drives scaffolding and homolog splitting, distinguishing between members of a gene family and outright repeats by allowing traversal of repetitive nodes multiple times based on coverage. The coverage cutoff parameter was used to exclude short low-coverage nodes from the final assembly (Zerbino 2008).

The average expected k-mer coverage of the assembly was determined automatically by the velvetg program as 19.12 by determining the length weighted median contig coverage (Zerbino 2008). Coverage cut-off values of 2, 5 and 10 were tested and a minimum of 10 paired reads were required to create a join between scaffolds.

The expected insert size was set to 220 bp based on the average distance between mapped reads (including sequenced ends) as determined by an initial ABBySS v 1.0.14 assembly (data not presented). For the final assembly, minimum contig size was set to 200 bp as this is the smallest contig size submissible to NCBI.

### 3.2.4 Gene prediction and genome characterisation

CEGMA v2.4 was used to assess the number of core eukaryotic genes (CEGs) present in the final assembly (Parra et al. 2007, Parra et al. 2009), as a means of estimating the completeness of the gene content. GeneMark.hmm ES version 2.3a (beta) (Ter-Hovhannisyan et al. 2008), an unsupervised gene prediction algorithm, was used to predict gene models. Six-frame open reading frame translation of the scaffolds was obtained using getorf (mEMBOSS 6.1.0.1 available from http://emboss.sourceforge.net/). Getorf produced a 6-frame translation of open reading frames (ORFs) greater than 30 nucleotides (10 amino acids) between stop codons but not including the stop codon.
3.2.5 Non-coding RNA analysis

Nuclear tRNAs were identified using tRNAscan-SE v1.3, with search parameters eukaryotic and cove only (Lowe and Eddy 1997). Non-coding RNA (ncRNA) regions within the genome assembly were identified using Infernal 1.1 rc2 (Nawrocki et al. 2009), by searching against the Rfam database 11.0 (Griffiths-Jones et al. 2003, Griffiths-Jones et al. 2005) using the gathering cut-off parameter. Overlap between predicted ncRNA and manually annotated gene models described in chapter 5 was assessed using BedTools (intersectBED) (Quinlan and Hall 2010). Predicted proteins were compared via BLASTP to the NCBI protein database (nr) at a significance threshold (e value threshold) of 1x10⁻⁵ (Altschul et al. 1990).

3.2.6 Analysis of repetitive DNA content of PmedOMT5

In order to analyse the occurrence of repetitive DNA in the PmedOMT5 assembly, scaffold sequences were analysed using RepeatScout (1.0.5) (Price et al. 2005) to create a list of 14 bp oligomers that occurred multiple times within the genome (max. 500bp apart, min. three copies). These were then filtered to remove low complexity regions (calculated to be > 50% low complexity using TRF (Benson 1999) or NSEG (Wootton and Federhen 1993) and regions less than 50 bp in length. Remaining repeats were aligned using CAP3 (Huang and Madan 1999) to identify and combine any regions that overlapped (parameters -h70 –z 1 –p 70). The output of the assembled repeats and remaining singletons were analysed using RepeatMasker v 3.3.0 (sensitive mode) (http://www.repeatmasker.org, Smit et al., 1996-2004). The output of RepeatMasker was filtered for repeats that occurred nine or more times within the genome to create a set of de novo repeat family consensus sequences.

De novo predicted repeats were analysed by BLASTN and TBLASTX against the NCBI nucleotide and protein databases (nt, nr), the P. nodorum SN15 repeats (Hane and Oliver 2008, Hane and Oliver 2010) and the fungal database at REPBASE (Repbase15.0.6-fngrep) (Jurka et al. 2005, Kapitonov and Jurka 2008). The top hit was identified by bit score, to assign a putative identity to repeat families. To determine repeat locations in relation to scaffold co-ordinates, repeats identified by RepeatMasker with similarity to repeats in the fungal database at REPBASE were plotted based on the calculated distance from the end of the repeat match to the closest end of the scaffold using JMP (SAS Institute Inc version 7).
RepeatMasker was also used to analyse the entire assembly for regions similar to repeats in two repeat libraries: SN15 repeats and Repbase fungal repeats. These libraries were generated from the combined consensus and deRIPed consensus *P. nodorum* SN15 repeat sequences (sequences with RIP like changes reverted to a potential ancestral form) (Hane and Oliver 2008, Hane and Oliver 2010) and the fungal database at REPBASE (Repbase15.0.6-fngrep) (Jurka *et al.* 2005).

One of the major difficulties of NGS assemblies is assembling repetitive regions of the genome. To examine the possibility that reads not used in the assembly were derived from repetitive sequences, unassembled reads were checked for alignment to repetitive regions in *P. nodorum* and Repbase via BWA0.5.9 using default parameters (Li and Durbin 2009) and Bowtie2 (Langmead and Salzberg 2012) (Parameters -N 1 -L 8 -f -p 8 –local).

### 3.2.7 RIP analysis of repetitive regions

In order to determine if RIP occurs in PmedOMT5, *de novo* predicted repeats were analysed using RIPCAL (Hane and Oliver 2008, Hane 2011) to generate alignment based repeat models (Parameters: alignment, gff, fasta, degenerate). The extent of RIP was quantified using the ‘RIP dominance’ score which calculates the most dominant forms of RIP mutation in a repeated sequence (Hane and Oliver 2008). In order to identify the possible ancestral sequence from which PmedOMT5 repeats may have been derived; *de novo* repeat models were also subjected to deRIP analysis (Hane and Oliver 2010).

To identify potential homologs of the *rid* gene required for RIP to occur in *N. crassa* [NCBI accession: AAM27408], BLASTP and TBLASTN analysis of the predicted PmedOMT5 proteins and scaffolds and those of related Dothideomycetes was conducted.

### 3.3 Results

#### 3.3.1 Read data and quality

A total of 11,233,131 paired-end reads (1.68 Gb), 75 nucleotides in length passed quality filtering and were used for subsequent *de novo* genome assembly without any further filtering. The average quality score of each base position of all reads was above
a Phred quality score of 20, which equates to a base calling accuracy of greater than 99%.

3.3.2 Assembly and read coverage

Test assemblies were run using Velvet (v1.0.14) at various k-mer lengths, (without optimising for coverage) to select the k-mer parameter that produced the largest N50 length and lowest N50 contig number. Assembly data is summarised in Figure 3.1, highlighting the selected optimum k-mer, with the largest L50 and showing the small changes in N50 as the k-mer length varies. Detailed comparisons are presented in Table 3.1 illustrating the enormous variations of maximum scaffold length with varying k-mer length.

For all assemblies a minimum contig size of 200 bp was applied. For optimisation of coverage parameters k-mer 39 was selected as it produced the lowest N50, longest N50 length, smallest number of scaffolds (consistent with the predicted genome assembly size) and the longest average scaffold length.

![Figure 3.1: Variation in N50 and L50 (N50 length) with changing k-mer length of Velvet assemblies (see Table 3.1 for detailed statistics). Red box indicates selected optimum k-mer assembly (k39) with largest L50 and smallest N50.](image-url)
Table 3.1: First stage of genome assembly optimisation of k-mer length for assembly with Velvet 1.0.14.

<table>
<thead>
<tr>
<th>k-mer length</th>
<th>21</th>
<th>31</th>
<th>33</th>
<th>35</th>
<th>37</th>
<th>39*</th>
<th>41</th>
<th>51</th>
<th>61</th>
<th>71</th>
</tr>
</thead>
<tbody>
<tr>
<td>N50</td>
<td>1,745</td>
<td>351</td>
<td>328</td>
<td>317</td>
<td>310</td>
<td>302</td>
<td>307</td>
<td>694</td>
<td>7,136</td>
<td>9</td>
</tr>
<tr>
<td>L50 (N50 length) (bp)</td>
<td>4,977</td>
<td>27,224</td>
<td>28,895</td>
<td>30,003</td>
<td>30,983</td>
<td>31,106</td>
<td>30,382</td>
<td>13,460</td>
<td>1,303</td>
<td>1,630</td>
</tr>
<tr>
<td>Average scaffold length (bp)</td>
<td>2,432</td>
<td>9,711</td>
<td>10,471</td>
<td>11,105</td>
<td>11,396</td>
<td>11,522</td>
<td>11,424</td>
<td>6,196</td>
<td>939</td>
<td>810</td>
</tr>
<tr>
<td>Total assembly length (Mbp)</td>
<td>30.24</td>
<td>30.57</td>
<td>30.59</td>
<td>30.6</td>
<td>30.6</td>
<td>30.6</td>
<td>30.7</td>
<td>29.8</td>
<td>0.72</td>
<td></td>
</tr>
<tr>
<td>Number of scaffolds</td>
<td>12,431</td>
<td>3,147</td>
<td>2,921</td>
<td>2,756</td>
<td>2,688</td>
<td>2,660</td>
<td>2,684</td>
<td>4,956</td>
<td>31,770</td>
<td>89</td>
</tr>
</tbody>
</table>

*Selected optimum k-mer (k39)
Table 3.2: Optimisation of Velvet assembly parameters “coverage cut-off” and “expected coverage”, for assemblies created using k-mer length 39.

<table>
<thead>
<tr>
<th>Test assembly number:</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly conditions tested:</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Coverage cut off</td>
<td>2</td>
<td>5</td>
<td>5</td>
<td>5</td>
<td>auto (9.561)</td>
<td>10</td>
</tr>
<tr>
<td>Minimum contig size set</td>
<td>200</td>
<td>200</td>
<td>not set</td>
<td>200</td>
<td>200</td>
<td>200</td>
</tr>
</tbody>
</table>

| Assembly statistics: |
|----------------------|----|----|----|----|----|----|
| Percentage of reads used in assembly | 84.2 | 85.99 | 87.8 | 86.4 | 86.8 | 86.8 |
| Number of scaffolds  | 1,155 | 2,660 | 2023 | 965 | 911 | 911 |
| N50                  | 90  | 302 | 91  | 91  | 101 | 101 |
| L50 (bp)             | 104,675 | 31,106 | 105,011 | 105,011 | 92,373 | 92,373 |
| Maximum contig size (bp) | 375,306 | 160,364 | 374,744 | 374,744 | 371,867 | 371,867 |
| Total assembly size (bp) | 30,907,835 | 30,662,184 | 30,885,734 | 30,777,866 | 30,618,426 | 30,618,426 |
| Average scaffold length (bp) | 26,736 | 11,522 | 15,260 | 31,861 | 33,572 | 33,572 |
| Minimum contig length | 200 | 200 | 77  | 200 | 200 | 200 |

For all assemblies an estimated insert length of 220 bp was used as had been determined by observation from preliminary ABYSS draft assemblies. Final assembly (test4) used for all subsequent analysis is highlighted in bold.
The optimal PmedOMT5 assembly resulted from the following parameters: covaerage cut off-5, minimum contig size- 200, expected coverage- 19.2 (Table 3.2) and this assembly was used for all further analysis. Although a coverage cut-off of 2 produced a slightly lower N50 of 90 rather than 91, the average scaffold length was much shorter, whilst using a coverage cut-off of 10 gave a lower number of contigs with a slightly longer average length but a slightly larger N50 and shorter L50, the selected conditions resulted in more large contigs and a small N50. Gene prediction using GeneMark and subsequent BLAST analysis of the predicted gene models showed an improvement in accuracy compared to previous other draft assemblies created using the ABysS algorithm (Simpson et al. 2009) without obvious incorrect joins. There was also an improvement in assembly of some repetitive regions such as the ribosomal internal transcribed spacer region (data not shown).

The average scaffold coverage in the final assembly of ~31 Mb was 150-fold and ranged from 9 to 3,687-fold for individual scaffolds. The higher coverage was observed on mitochondrial and repetitive scaffolds, with the highest coverage seen on scaffolds containing rDNA repeats. Scaffolds containing predicted genes had an average of 38-fold nucleotide coverage. The overall median coverage was 39.5-fold, consistent with the average coverage of gene-encoding nuclear scaffolds. The statistics of the final assembly (test 4 in Table 3.2) are summarised in Table 3.3.

Table 3.3: PmedOMT5 final assembly characteristics.

<table>
<thead>
<tr>
<th>Assembly statistics</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of scaffolds</td>
<td>965</td>
</tr>
<tr>
<td>Number of nuclear scaffolds</td>
<td>952</td>
</tr>
<tr>
<td>Number of mitochondrial scaffolds</td>
<td>13</td>
</tr>
<tr>
<td>Number of nuclear scaffolds &gt; 1 kbp in length</td>
<td>643</td>
</tr>
<tr>
<td>Number of nuclear scaffolds &gt; 10 kbp in length</td>
<td>419</td>
</tr>
<tr>
<td>Total assembly size (bp)</td>
<td>30,777,866</td>
</tr>
<tr>
<td>Number of scaffolds with gaps (stretches of Ns)</td>
<td>365</td>
</tr>
<tr>
<td>Number of unknown bases in total assembly (Ns)</td>
<td>36,608</td>
</tr>
<tr>
<td>Average nucleotide coverage of scaffolds containing genes</td>
<td>38.3</td>
</tr>
<tr>
<td>Number of reads used in the assembly</td>
<td>19,318,780</td>
</tr>
</tbody>
</table>
3.3.3 Assessment of gene space coverage within the final assembly

Genis Parra and colleagues designed a method for assessing the completeness of an eukaryote genome assembly by identifying a set of 248 core eukaryotic genes (CEGs), that are extremely highly conserved and are generally present in low copy numbers in higher eukaryotes. They demonstrated that the proportion of these genes that can be identified in a genome assembly provides a rough approximation of the proportion of all genes that may be present (Parra et al. 2009). CEGMA analysis of the PmedOMT5 assembly shows that 245 of the 248 CEGs (98.79%) were detected in their complete form. A further two genes were detected in a partial form making a total of 247 genes (99.6%) that have at least partial matches (CEGMA defines a partial match as less than 70% of the predicted gene aligning to the CEG) (Table 3.4). CEGMA analysis divides CEGs into groups with varying levels of conservation, of which group 1 is the least conserved, and analyses the number of orthologs for each CEG present in the genome as an indicator of expansion of a gene family in the genome under study. Based on the CEGMA analysis the PmedOMT5 assembly covers the majority of the gene space, that is, the majority of nuclear encoded genes are likely to be present in the assembly.

Table 3.4: Number and completeness of 248 ultra-conserved core eukaryote genes (CEGs) identified within the PmedOMT5 final assembly.

<table>
<thead>
<tr>
<th></th>
<th>Number of conserved CEGs present in genome</th>
<th>Percentage conserved CEGs present</th>
<th>Total number of CEGs present including putative orthologs</th>
<th>Average number of orthologs per CEG</th>
<th>Percentage of detected CEGs that have more than one ortholog</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complete</td>
<td>245</td>
<td>98.70</td>
<td>283</td>
<td>1.16</td>
<td>12.24</td>
</tr>
<tr>
<td>Group 1</td>
<td>64</td>
<td>96.97</td>
<td>71</td>
<td>1.11</td>
<td>6.25</td>
</tr>
<tr>
<td>Group 2</td>
<td>56</td>
<td>100</td>
<td>64</td>
<td>1.14</td>
<td>12.50</td>
</tr>
<tr>
<td>Group 3</td>
<td>60</td>
<td>98.36</td>
<td>76</td>
<td>1.27</td>
<td>21.67</td>
</tr>
<tr>
<td>Group 4</td>
<td>65</td>
<td>100</td>
<td>72</td>
<td>1.11</td>
<td>9.23</td>
</tr>
<tr>
<td>Partial</td>
<td>247</td>
<td>99.6</td>
<td>299</td>
<td>1.21</td>
<td>15.79</td>
</tr>
<tr>
<td>Group 1</td>
<td>65</td>
<td>98.48</td>
<td>72</td>
<td>1.11</td>
<td>6.15</td>
</tr>
<tr>
<td>Group 2</td>
<td>56</td>
<td>100</td>
<td>70</td>
<td>1.25</td>
<td>19.64</td>
</tr>
<tr>
<td>Group 3</td>
<td>61</td>
<td>100</td>
<td>82</td>
<td>1.34</td>
<td>26.23</td>
</tr>
<tr>
<td>Group 4</td>
<td>65</td>
<td>100</td>
<td>75</td>
<td>1.15</td>
<td>12.31</td>
</tr>
</tbody>
</table>

3.3.4 Characterisation of assembly and genomic landscape

3.3.5 De novo gene prediction of PmedOMT5 protein set version 1

*De novo* gene predictions from the 965 assembled scaffolds were created via GeneMark (Ter-Hovhannisyan et al. 2008). This program is optimised for fungal *de novo* gene prediction and assigns gene numbers to regions of the genome based on an Eukaryotic
Self-training method. The initial GeneMark-predicted protein set consisted of 10,843 predicted proteins ranging from 2 to 9,592 amino acids (aa) in length, 10,377 (95.7%) of which contained both a start and a stop codon. These predictions were filtered to create a set of 10,478 proteins greater than or equal to 50 aa (i.e. 150 nucleotides). Many of the short (≤50 aa) predicted proteins were found on scaffolds of less than 1 kb (64%) or were incomplete i.e. did not contain either a start and (or) stop codon (49%). The short predicted proteins (≤50 aa) were retained for initial analysis but unless validated by experimental evidence, homology to other fungal genes or re-annotated to longer models (see chapter 5), they were excluded from the final protein set (version 2). Statistics for the predicted version 1 gene models are outlined in Table 3.5.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>All genes</th>
<th>Proteins ≥ 50 aa</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of proteins predicted by GeneMark ES</td>
<td>10,843</td>
<td>10,478</td>
</tr>
<tr>
<td>Percentage of complete gene models (start and stop)</td>
<td>95.7</td>
<td>97.2</td>
</tr>
<tr>
<td>Incomplete genes (missing start or stop codon)</td>
<td>466</td>
<td>286</td>
</tr>
<tr>
<td>Mean length of proteins (aa)</td>
<td>473</td>
<td>489</td>
</tr>
<tr>
<td>Mean gene length (nucleotides)</td>
<td>1,566</td>
<td>1,617</td>
</tr>
<tr>
<td>Mean CDS length (nucleotides)</td>
<td>1,422</td>
<td>1,470</td>
</tr>
<tr>
<td>Mean number of exons per gene</td>
<td>2.54</td>
<td>2.58</td>
</tr>
<tr>
<td>Mean exon size (nucleotides)</td>
<td>695</td>
<td>718</td>
</tr>
<tr>
<td>Percentage of fungal protein homologs*</td>
<td>91.4</td>
<td>94.5</td>
</tr>
</tbody>
</table>

* BLASTP match to nr fungal gene, significant at an e value cut-off of ≤ 1x 10^-5

The number of predicted genes identified on each scaffold was plotted relative to the scaffold length to aid in the identification of scaffolds with unusually low gene content. A trend line was included to enable better visualisation of outliers which may contain less genes than others and thus form part of dispensable chromosomes (Figure 3.2). Several outliers were identified but further analysis showed that they were unlikely to form part of conditionally dispensable chromosomes, which are known to show a pattern of low gene content relative to core chromosomes in other Dothideomycetes.
Figure 3.2: Number of predicted genes per scaffold relative to the length of the scaffold. Horizontal line indicates the L50 (N50 length) in bp, diagonal trend line calculated to facilitate distinction of outliers.
3.3.6 Assessment of automated gene annotation via comparison of length to best hit in the NCBI non-redundant protein database (nr) by BLASTP

The majority (9,914) of predicted proteins showed homology to proteins in the non-redundant protein database at NCBI below a significance threshold of $1 \times 10^{-5}$. Of these 7,527 had protein matches in NCBI that covered over 80% of the top hit, illustrated in Figure 3.3.

None of the predicted proteins with lengths under 30 amino acids showed significant alignments. The longest protein with significant homology is 9,592 aa (Pmed_00341v1). This gene has homologs in other Dothideomycetes of similar lengths: *P. teres* f. *teres* (PTT_10986, [GenBank: EFQ92021.1], 9,819 aa- match bit score 1292, e=0.0) and *P. tritici-repentis* (PTRG_02062, 9,643aa, [NCBI: XP_001793282.1]) suggesting that this gene model is accurately annotated. All these proteins contain the conserved domains: chromosome segregation protein SMC [pfam: 2463] and ribonuclease E [CDD: PRK10811] and show partial matches across their entire lengths,
although they exhibit high variability outside these domains, resulting in only 36.8% hit completeness which was determined based on cumulative HSP matches as depicted in Figure 3.3.

3.3.7 Identification of non-coding RNA (ncRNA) sequences

In total, 137 tRNA genes were identified including those on several scaffolds predicted to be mitochondrial in origin (see Chapter 4). Of these, 116 were present on nuclear scaffolds and encode anti-codons for all of the 20 common amino acids except histidine and cysteine. Twelve predicted tRNA pseudogenes were also identified.

Nuclear tRNA’s were found on 85 scaffolds. Most occurred without neighbouring tRNAs, although several small clusters of up to four per scaffold were identified (Appendix 3.3: Nuclear tRNA co-ordinates).

Infernal analysis (Nawrocki et al. 2009) predicted 422 genomic regions containing ncRNA. This list was manually curated to capture only those with matches showing a significance value below a threshold of 1x10^{-2} and that did not show a strong bias score which can indicate potential false positive matches. Of the remaining 208, 38 were predicted to overlap gene regions. From those 38, 18 were predicted to occur within introns, 4 on the opposite strand to the coding region and the rest to overlap untranslated region (UTRs). Of the 38 proteins overlapped, 10 have no BLASTP hit to the GenBank protein data set (NR) below a significance threshold of 10^{-5}. Genes with predicted overlaps were manually examined and re-annotated if necessary (chapter 5).

The number of members of each family of ncRNAs identified in the assembly, the majority of which were ribosomal or small nucleolar RNAs is summarised in Table 3.6 and Table 3.14.
Table 3.6: Number of members in ncRNA families identified on the PmedOMT5 nuclear scaffolds determined by Infernal vs. RFAM (e-value ≤ 0.02).

<table>
<thead>
<tr>
<th>Non-coding RNA family</th>
<th>Rfam accession</th>
<th>Number in OMT5 genome</th>
<th>Rfam description of function</th>
</tr>
</thead>
<tbody>
<tr>
<td>5S_rRNA</td>
<td>RF00001</td>
<td>34</td>
<td>5S ribosomal RNA</td>
</tr>
<tr>
<td>Afu_182</td>
<td>RF01496</td>
<td>1</td>
<td>A. fumigatus snoRNA Afu_182</td>
</tr>
<tr>
<td>Afu_190</td>
<td>RF01498</td>
<td>1</td>
<td>A. fumigatus snoRNA Afu_190</td>
</tr>
<tr>
<td>Afu_198</td>
<td>RF01500</td>
<td>1</td>
<td>A. fumigatus snoRNA Afu_198</td>
</tr>
<tr>
<td>Afu_300</td>
<td>RF01509</td>
<td>1</td>
<td>A. fumigatus snoRNA Afu_300</td>
</tr>
<tr>
<td>Afu_309</td>
<td>RF01512</td>
<td>1</td>
<td>A. fumigatus snoRNA Afu_309</td>
</tr>
<tr>
<td>Afu_335</td>
<td>RF01513</td>
<td>1</td>
<td>A. fumigatus snoRNA Afu_335</td>
</tr>
<tr>
<td>Afu_455</td>
<td>RF01860</td>
<td>1</td>
<td>A. fumigatus snoRNA Afu_455</td>
</tr>
<tr>
<td>Afu_513</td>
<td>RF01514</td>
<td>1</td>
<td>A. fumigatus snoRNA Afu_513</td>
</tr>
<tr>
<td>Afu_514</td>
<td>RF01515</td>
<td>1</td>
<td>A. fumigatus snoRNA Afu_514</td>
</tr>
<tr>
<td>EAV_LTH</td>
<td>RF00498</td>
<td>1</td>
<td>Equine arteritis virus leader TRS hairpin (LTH)</td>
</tr>
<tr>
<td>Fungi_SRP</td>
<td>RF01502</td>
<td>1</td>
<td>Fungal signal recognition particle RNA</td>
</tr>
<tr>
<td>Fungi_U3</td>
<td>RF01846</td>
<td>2</td>
<td>Fungal small nucleolar RNA U3</td>
</tr>
<tr>
<td>HgcC</td>
<td>RF00062</td>
<td>1</td>
<td>HgcC family RNA</td>
</tr>
<tr>
<td>mir-276</td>
<td>RF00637</td>
<td>1</td>
<td>microRNA mir-276</td>
</tr>
<tr>
<td>mir-331</td>
<td>RF00769</td>
<td>7</td>
<td>microRNA mir-331</td>
</tr>
<tr>
<td>mir-598</td>
<td>RF01059</td>
<td>1</td>
<td>microRNA mir-598</td>
</tr>
<tr>
<td>P10</td>
<td>RF01668</td>
<td>1</td>
<td>Pseudomonas sRNA P10</td>
</tr>
<tr>
<td>RNaseP_nuc</td>
<td>RF00009</td>
<td>1</td>
<td>Nuclear RNase P</td>
</tr>
<tr>
<td>snoR38</td>
<td>RF00213</td>
<td>2</td>
<td>Small nucleolar RNA R38</td>
</tr>
<tr>
<td>SNORD14</td>
<td>RF00016</td>
<td>2</td>
<td>Small nucleolar RNA SNORD14</td>
</tr>
<tr>
<td>SNORD18</td>
<td>RF00093</td>
<td>1</td>
<td>Small nucleolar RNA SNORD18</td>
</tr>
<tr>
<td>snosnR60_Z15</td>
<td>RF00309</td>
<td>1</td>
<td>Small nucleolar RNA</td>
</tr>
<tr>
<td>snosnR61</td>
<td>RF00476</td>
<td>1</td>
<td>Small nucleolar RNA snR61/Z1/Z11</td>
</tr>
<tr>
<td>snoZ13_snr52</td>
<td>RF00335</td>
<td>1</td>
<td>Small nucleolar RNA Z13/snr52</td>
</tr>
<tr>
<td>snR191</td>
<td>RF01263</td>
<td>1</td>
<td>Small nucleolar RNA snR191</td>
</tr>
<tr>
<td>snR32</td>
<td>RF01247</td>
<td>1</td>
<td>Small nucleolar RNA snR32</td>
</tr>
<tr>
<td>snR36</td>
<td>RF01242</td>
<td>1</td>
<td>Small nucleolar RNA snR36</td>
</tr>
<tr>
<td>snR44</td>
<td>RF01262</td>
<td>1</td>
<td>Small nucleolar RNA snR44</td>
</tr>
<tr>
<td>snR51</td>
<td>RF01214</td>
<td>1</td>
<td>Small nucleolar RNA snR51</td>
</tr>
<tr>
<td>snR73</td>
<td>RF01207</td>
<td>1</td>
<td>Small nucleolar RNA snR73</td>
</tr>
<tr>
<td>snR75</td>
<td>RF01185</td>
<td>1</td>
<td>Small nucleolar RNA snR75</td>
</tr>
<tr>
<td>SPRY4-IT1_1</td>
<td>RF02038</td>
<td>1</td>
<td>SPRY4-IT1 conserved region 1</td>
</tr>
<tr>
<td>tmRNA</td>
<td>RF00023</td>
<td>2</td>
<td>transfer-messenger RNA</td>
</tr>
<tr>
<td>TPP</td>
<td>RF00059</td>
<td>5</td>
<td>TPP riboswitch (THI element)</td>
</tr>
<tr>
<td>tRNA</td>
<td>RF00005</td>
<td>119</td>
<td>tRNA</td>
</tr>
<tr>
<td>U1</td>
<td>RF00003</td>
<td>3</td>
<td>U1 spliceosomal RNA</td>
</tr>
<tr>
<td>U4</td>
<td>RF00015</td>
<td>1</td>
<td>U4 spliceosomal RNA</td>
</tr>
<tr>
<td>U5</td>
<td>RF00020</td>
<td>1</td>
<td>U5 spliceosomal RNA</td>
</tr>
<tr>
<td>U6</td>
<td>RF00026</td>
<td>2</td>
<td>U6 spliceosomal RNA</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>208</td>
<td></td>
</tr>
</tbody>
</table>
3.3.8 Repetitive DNA content of PmedOMT5

A schematic outline of the approach used to investigate the repetitive content of PmedOMT5 is presented in Figure 3.4. Of the 965 scaffolds in the PmedOMT5 assembly, 612 contained repetitive regions, including low complexity repeats, 151 contained complex repeats including TEs and six are composed entirely of repeats, all rDNA fragments.

De novo analysis of repetitive regions identified 109 repeated sequences in the PmedOMT5 genome. A total of 101 remained after filtering low complexity loci to remove sequences less than 50 bp or greater than 50% low complexity judged by NSEG or TRF. Following CAP3 contig assembly there were 83 de novo repeat families. Forty of these occurred nine or more times in the assembly and were not low complexity or simple repeats (Table 3.7).

The majority of repetitive regions identified by the de novo analysis were simple repeats, short regions of repeated nucleotide patterns that usually do not occur in genic regions but are dispersed throughout the genome. Where possible the role and origin of the 40 unknown repeat classes was identified by homology to previously identified
repeats using RepeatMasker and BLAST versus NCBI, SN15 characterised repeats or the REPBASE database (15.0.6). Low complexity (G:C and A:T rich regions) and simple repeats (mostly SSRs) (Table 3.7) were distributed throughout the assembly (Figure 3.5) but transposons were more often identified close to the end of scaffolds.

RIP dominance scores were calculated to examine the type of RIP most prevalent in PmedOMT5. A higher RIP dominance score indicates a greater intensity of RIP-like mutations. In L. maculans, repeat classes that showed a RIP dominance score of greater than 1 were classified as heavily RIPped (Van de Wouw 2010). In the majority of cases in PmedOMT5 CpA↔TpA was the dominant form followed by CpT↔TpT. Of the de novo repeats derived from transposons, the longer the consensus sequence the more RIP-like changes were observed. RIP dominance scores for de novo repeats are listed in Table 3.8.

Repeated regions identified by homology to those in SN15 and REPBASE were typically only short matches (Table 3.9, Table 3.10), and as can be seen in table 3.10 only low percentages of the complete repeats seen in SN15 were identifiable in PmedOMT5. Bowtie2 analysis aligned 11,904 of the unassembled reads one or more times to the SN15 consensus repeats with the majority of these (99.85%), aligning to ribosomal repeats.

Table 3.7: Repetitive DNA classes identified by RepeatMasker in the PmedOMT5 assembly using the de novo repeat library.

<table>
<thead>
<tr>
<th>Repeat class/family</th>
<th>Occurrence of repeat type in assembly</th>
<th>Number of repeat types in class that occur in PmedOMT5</th>
<th>Total length of repeat matches in assembly (bp)</th>
<th>Percentage of total assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low complexity</td>
<td>234</td>
<td>2</td>
<td>8,194</td>
<td>0.03</td>
</tr>
<tr>
<td>Simple repeat</td>
<td>3,477</td>
<td>148</td>
<td>216,793</td>
<td>0.70</td>
</tr>
<tr>
<td>Unknown</td>
<td>988</td>
<td>40</td>
<td>192,006</td>
<td>0.62</td>
</tr>
<tr>
<td>Total</td>
<td>4,699</td>
<td>190</td>
<td>416,993</td>
<td>1.35</td>
</tr>
</tbody>
</table>
Figure 3.5: Proximity by class of repetitive DNA of PmedOMT5 genome assembly to ends of scaffolds. (Marker colours: green- DNA transposons, purple- LTR retrotransposons aqua- ribosomal RNA repeats, red- non LTR retrotransposons, orange- centromeric repeats, box plots indicate median and 25th to 75th quantile). Classes are presented as reported by matches to Repbase.
Table 3.8: RIP and homology analysis of repetitive DNA classes identified by RepeatMasker in the PmedOMT5 assembly using the de novo repeat library. (RIP dominance scores > 1 highlighted in bold).

<table>
<thead>
<tr>
<th>Repeat name*</th>
<th>Length of consensus</th>
<th>Shortest</th>
<th>Longest</th>
<th>Average</th>
<th>Total sequence matches in assembly (bp)</th>
<th>Description based on top BLAST hit</th>
<th>CpA←→TpA</th>
<th>CpC←→TpC</th>
<th>CpG←→TpG</th>
<th>CpT←→TpT</th>
</tr>
</thead>
<tbody>
<tr>
<td>RepeatContig1</td>
<td>276</td>
<td>50</td>
<td>276</td>
<td>173</td>
<td>9,889</td>
<td>Similar to region of emb</td>
<td>FP929139.1 Leptosphaeria maculans v23.1.3 ln_SuperContig_0_v2</td>
<td>0.57</td>
<td>0.46</td>
<td>0.25</td>
</tr>
<tr>
<td>RepeatContig2</td>
<td>1640</td>
<td>54</td>
<td>1640</td>
<td>615</td>
<td>41,845</td>
<td>Mariner/Tc1; DNA transposon</td>
<td>1.06</td>
<td>0.15</td>
<td>0.13</td>
<td>0.53</td>
</tr>
<tr>
<td>RepeatContig3</td>
<td>117</td>
<td>77</td>
<td>117</td>
<td>100</td>
<td>1,095</td>
<td>Unknown</td>
<td>1.30</td>
<td>0.04</td>
<td>0.44</td>
<td>0.00</td>
</tr>
<tr>
<td>RepeatContig4</td>
<td>1051</td>
<td>37</td>
<td>914</td>
<td>208</td>
<td>9,547</td>
<td>Mariner/Tc1; DNA transposon</td>
<td>1.19</td>
<td>0.09</td>
<td>0.15</td>
<td>0.45</td>
</tr>
<tr>
<td>RepeatContig5</td>
<td>597</td>
<td>46</td>
<td>597</td>
<td>244</td>
<td>4,872</td>
<td>LTR retrotransposon</td>
<td>1.97</td>
<td>0.21</td>
<td>0.05</td>
<td>0.35</td>
</tr>
<tr>
<td>RepeatContig6</td>
<td>263</td>
<td>35</td>
<td>263</td>
<td>156</td>
<td>2,188</td>
<td>Unknown</td>
<td>1.51</td>
<td>0.09</td>
<td>0.24</td>
<td>0.22</td>
</tr>
<tr>
<td>RepeatContig7</td>
<td>390</td>
<td>36</td>
<td>390</td>
<td>164</td>
<td>7,874</td>
<td>Mariner/Tc1; DNA transposon</td>
<td>1.33</td>
<td>0.08</td>
<td>0.55</td>
<td>0.25</td>
</tr>
<tr>
<td>RepeatContig8</td>
<td>3593</td>
<td>77</td>
<td>3309</td>
<td>544</td>
<td>35,920</td>
<td>LTR retrotransposon</td>
<td>0.96</td>
<td>0.10</td>
<td>0.05</td>
<td>2.07</td>
</tr>
<tr>
<td>RepeatContig9</td>
<td>496</td>
<td>159</td>
<td>496</td>
<td>356</td>
<td>3,205</td>
<td>Unknown</td>
<td>0.69</td>
<td>0.22</td>
<td>0.11</td>
<td>0.26</td>
</tr>
<tr>
<td>R107</td>
<td>59</td>
<td>30</td>
<td>59</td>
<td>53</td>
<td>1,111</td>
<td>Unknown</td>
<td>0.10</td>
<td>0.23</td>
<td>0.64</td>
<td>0.00</td>
</tr>
<tr>
<td>R13</td>
<td>217</td>
<td>53</td>
<td>217</td>
<td>164</td>
<td>3,930</td>
<td>Mariner/Tc1; DNA transposon</td>
<td>0.72</td>
<td>0.13</td>
<td>0.49</td>
<td>0.23</td>
</tr>
</tbody>
</table>

*Repeat Contigs are repeats identified by RepeatMasker that could be concatenated via CAP3.

Table continued overleaf
<table>
<thead>
<tr>
<th>Repeat name</th>
<th>Length of consensus</th>
<th>Shortest</th>
<th>Longest</th>
<th>Average</th>
<th>Total sequence matches in assembly (bp)</th>
<th>Description based on top BLASThit</th>
<th>CpA↔TpA</th>
<th>CpC↔TpC</th>
<th>CpG↔TpG</th>
<th>CpT↔TpT</th>
</tr>
</thead>
<tbody>
<tr>
<td>R15</td>
<td>67</td>
<td>38</td>
<td>67</td>
<td>61</td>
<td>788</td>
<td>mariner pogo clade</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>R18</td>
<td>83</td>
<td>80</td>
<td>83</td>
<td>82</td>
<td>742</td>
<td>tRNA</td>
<td>0.00</td>
<td>0.29</td>
<td>0.05</td>
<td>0.81</td>
</tr>
<tr>
<td>R19</td>
<td>83</td>
<td>38</td>
<td>83</td>
<td>74</td>
<td>1,116</td>
<td>Unknown</td>
<td>0.08</td>
<td>0.25</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>R21</td>
<td>387</td>
<td>38</td>
<td>387</td>
<td>274</td>
<td>7,939</td>
<td>MOLLY Mariner Tc1</td>
<td>0.56</td>
<td>0.13</td>
<td>0.34</td>
<td>0.42</td>
</tr>
<tr>
<td>R24</td>
<td>84</td>
<td>38</td>
<td>84</td>
<td>68</td>
<td>1,705</td>
<td>DNA mariner</td>
<td>2.02</td>
<td>0.34</td>
<td>0.10</td>
<td>0.12</td>
</tr>
<tr>
<td>R29</td>
<td>450</td>
<td>31</td>
<td>450</td>
<td>286</td>
<td>4,007</td>
<td>marinerTc1</td>
<td>1.09</td>
<td>0.13</td>
<td>0.33</td>
<td>0.33</td>
</tr>
<tr>
<td>R3</td>
<td>130</td>
<td>30</td>
<td>130</td>
<td>89</td>
<td>7,612</td>
<td>5sRNA interspersed repeat</td>
<td>0.37</td>
<td>0.34</td>
<td>0.35</td>
<td>0.35</td>
</tr>
<tr>
<td>R30</td>
<td>82</td>
<td>36</td>
<td>82</td>
<td>67</td>
<td>2,468</td>
<td>Unknown</td>
<td>0.38</td>
<td>0.20</td>
<td>0.43</td>
<td>0.00</td>
</tr>
<tr>
<td>R32</td>
<td>64</td>
<td>60</td>
<td>64</td>
<td>62</td>
<td>1,437</td>
<td>Unknown</td>
<td>0.63</td>
<td>0.06</td>
<td>0.09</td>
<td>0.75</td>
</tr>
<tr>
<td>R36</td>
<td>226</td>
<td>37</td>
<td>226</td>
<td>137</td>
<td>1,781</td>
<td>Unknown</td>
<td>1.70</td>
<td>0.02</td>
<td>0.18</td>
<td>0.03</td>
</tr>
<tr>
<td>R39</td>
<td>195</td>
<td>118</td>
<td>195</td>
<td>175</td>
<td>2,094</td>
<td>Unknown</td>
<td>1.43</td>
<td>0.23</td>
<td>0.19</td>
<td>0.25</td>
</tr>
<tr>
<td>R40</td>
<td>340</td>
<td>38</td>
<td>340</td>
<td>239</td>
<td>3,346</td>
<td>Mariner</td>
<td>1.04</td>
<td>0.07</td>
<td>0.20</td>
<td>1.24</td>
</tr>
<tr>
<td>R41</td>
<td>61</td>
<td>46</td>
<td>61</td>
<td>55</td>
<td>491</td>
<td>Unknown</td>
<td>0.12</td>
<td>0.75</td>
<td>0.04</td>
<td>0.67</td>
</tr>
<tr>
<td>R43</td>
<td>172</td>
<td>32</td>
<td>172</td>
<td>90</td>
<td>2,882</td>
<td>Unknown</td>
<td>1.78</td>
<td>0.25</td>
<td>0.18</td>
<td>0.15</td>
</tr>
<tr>
<td>R44</td>
<td>51</td>
<td>38</td>
<td>51</td>
<td>49</td>
<td>938</td>
<td>Unknown</td>
<td>0.69</td>
<td>0.18</td>
<td>0.07</td>
<td>0.09</td>
</tr>
</tbody>
</table>

Table continued overleaf
<table>
<thead>
<tr>
<th>Repeat name</th>
<th>Length of consensus</th>
<th>Shortest</th>
<th>Longest</th>
<th>Average</th>
<th>Total sequence matches in assembly (bp)</th>
<th>Description based on top BLAST hit</th>
<th>CpA&lt;&gt;TpA</th>
<th>CpC&lt;&gt;TpC</th>
<th>CpG&lt;&gt;TpG</th>
<th>CpT&lt;&gt;TpT</th>
</tr>
</thead>
<tbody>
<tr>
<td>R45</td>
<td>248</td>
<td>77</td>
<td>248</td>
<td>170</td>
<td>2,041</td>
<td>DNA transposon</td>
<td>1.37</td>
<td>0.04</td>
<td>0.13</td>
<td>0.50</td>
</tr>
<tr>
<td>R53</td>
<td>92</td>
<td>80</td>
<td>92</td>
<td>91</td>
<td>1,449</td>
<td>Unknown</td>
<td>0.37</td>
<td>0.21</td>
<td>0.25</td>
<td>1.32</td>
</tr>
<tr>
<td>R57</td>
<td>138</td>
<td>64</td>
<td>138</td>
<td>128</td>
<td>2,041</td>
<td>Unknown</td>
<td>0.76</td>
<td>0.16</td>
<td>0.18</td>
<td>0.95</td>
</tr>
<tr>
<td>R58</td>
<td>159</td>
<td>68</td>
<td>159</td>
<td>130</td>
<td>2,865</td>
<td>Unknown</td>
<td>0.27</td>
<td>0.29</td>
<td>0.30</td>
<td>0.73</td>
</tr>
<tr>
<td>R62</td>
<td>350</td>
<td>64</td>
<td>350</td>
<td>244</td>
<td>4,396</td>
<td>MOLLY Mariner Tc1</td>
<td>1.82</td>
<td>0.02</td>
<td>0.18</td>
<td>0.29</td>
</tr>
<tr>
<td>R66</td>
<td>261</td>
<td>139</td>
<td>261</td>
<td>237</td>
<td>3,086</td>
<td><em>Phaeosphaeria nodorum</em> Pyrenophora teres f. teres* 0-1 hypothetical protein</td>
<td>1.68</td>
<td>0.24</td>
<td>0.20</td>
<td>0.19</td>
</tr>
<tr>
<td>R68</td>
<td>52</td>
<td>33</td>
<td>52</td>
<td>49</td>
<td>989</td>
<td>Unknown</td>
<td>0.60</td>
<td>0.00</td>
<td>0.05</td>
<td>0.11</td>
</tr>
<tr>
<td>R71</td>
<td>60</td>
<td>35</td>
<td>60</td>
<td>54</td>
<td>590</td>
<td>LTR retrotransposon,Gypsy-2-I_AF_Gypsy_Aspergillus_fumigatus</td>
<td>0.60</td>
<td>0.17</td>
<td>0.00</td>
<td>0.23</td>
</tr>
<tr>
<td>R76</td>
<td>63</td>
<td>59</td>
<td>63</td>
<td>62</td>
<td>806</td>
<td>Unknown</td>
<td>0.44</td>
<td>0.03</td>
<td>0.31</td>
<td>0.40</td>
</tr>
<tr>
<td>R77</td>
<td>74</td>
<td>33</td>
<td>74</td>
<td>67</td>
<td>875</td>
<td>Unknown</td>
<td>1.12</td>
<td>0.46</td>
<td>0.02</td>
<td>0.07</td>
</tr>
<tr>
<td>R87</td>
<td>73</td>
<td>42</td>
<td>73</td>
<td>66</td>
<td>661</td>
<td>Unknown</td>
<td>1.40</td>
<td>0.47</td>
<td>0.00</td>
<td>0.24</td>
</tr>
<tr>
<td>R9</td>
<td>141</td>
<td>63</td>
<td>141</td>
<td>119</td>
<td>4,756</td>
<td>Unknown</td>
<td>0.90</td>
<td>0.27</td>
<td>0.33</td>
<td>0.23</td>
</tr>
<tr>
<td>R91</td>
<td>101</td>
<td>38</td>
<td>101</td>
<td>85</td>
<td>2,878</td>
<td>Unknown</td>
<td>1.17</td>
<td>0.01</td>
<td>0.12</td>
<td>0.19</td>
</tr>
<tr>
<td>R95</td>
<td>310</td>
<td>95</td>
<td>310</td>
<td>234</td>
<td>3,751</td>
<td>Mariner/Tc1; DNA transposon</td>
<td>1.16</td>
<td>0.23</td>
<td>0.20</td>
<td>0.13</td>
</tr>
</tbody>
</table>
Table 3.9: Identification of repetitive regions in the PmedOMT5 assembly by similarity to known fungal repeats (Repbase).

<table>
<thead>
<tr>
<th>Repeat type</th>
<th>Repbase Repeat class/family</th>
<th>Av. repeat match length</th>
<th>Occurrence of repeat within assembly</th>
<th>Total length of repeat matches</th>
<th>Percentage of total assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type I:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Retrotransposons</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>non-LTR</td>
<td>LINE/I</td>
<td>54</td>
<td>1</td>
<td>54</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LINE/Penelope</td>
<td>69</td>
<td>1</td>
<td>69</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>LINE/Tad1</td>
<td>651</td>
<td>2</td>
<td>1302</td>
<td>0.004</td>
</tr>
<tr>
<td></td>
<td>LINE/Tad1 (possible)</td>
<td>185</td>
<td>2</td>
<td>370</td>
<td>0.001</td>
</tr>
<tr>
<td>LTR</td>
<td>LTR/Copia</td>
<td>163</td>
<td>11</td>
<td>1789</td>
<td>0.006</td>
</tr>
<tr>
<td></td>
<td>LTR/Gypsy</td>
<td>375</td>
<td>49</td>
<td>18363</td>
<td>0.06</td>
</tr>
<tr>
<td><strong>Type II:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DNA transposons</td>
<td>DNA/Maverick</td>
<td>44</td>
<td>1</td>
<td>44</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>DNA/TcMar-Ant1</td>
<td>49</td>
<td>2</td>
<td>98</td>
<td>0</td>
</tr>
<tr>
<td></td>
<td>DNA/TcMar-Fot1</td>
<td>485</td>
<td>90</td>
<td>43662</td>
<td>0.142</td>
</tr>
<tr>
<td></td>
<td>DNA/TcMar-Mariner</td>
<td>328</td>
<td>5</td>
<td>1638</td>
<td>0.005</td>
</tr>
<tr>
<td></td>
<td>DNA/TcMar-Pogo</td>
<td>348</td>
<td>2</td>
<td>696</td>
<td>0.002</td>
</tr>
<tr>
<td></td>
<td>DNA(possible transposon)</td>
<td>13</td>
<td>1</td>
<td>13</td>
<td>0</td>
</tr>
<tr>
<td><strong>Other:</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rRNA</td>
<td>143</td>
<td>83</td>
<td></td>
<td>11886</td>
<td>0.039</td>
</tr>
<tr>
<td>Unknown/centromeric</td>
<td>106</td>
<td>4</td>
<td></td>
<td>422</td>
<td>0.001</td>
</tr>
<tr>
<td>Satellite</td>
<td>73</td>
<td>1</td>
<td></td>
<td>73</td>
<td>0</td>
</tr>
<tr>
<td>Low_complexity</td>
<td>71</td>
<td>632</td>
<td></td>
<td>44942</td>
<td>0.146</td>
</tr>
<tr>
<td>Simple_repeat</td>
<td>61</td>
<td>2961</td>
<td></td>
<td>181861</td>
<td>0.591</td>
</tr>
</tbody>
</table>
Table 3.10: RepeatMasker analysis of PmedOMT5 assembly regions with similarity to repeats identified in the genome of *P. nodorum* SN15.

<table>
<thead>
<tr>
<th>Repeat type</th>
<th>Predicted origin of Repeat (Hane and Oliver 2010)</th>
<th>Total length of repeat matches in the assembly</th>
<th>Repeat Match Length</th>
<th>Percentage of average repeat match length in PmedOMT5 to repeat length in SN15</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Average</td>
<td>Shortest</td>
<td>Longest</td>
</tr>
<tr>
<td>molly</td>
<td>Mariner/Tc1-like DNA transposon</td>
<td>19,618</td>
<td>363</td>
<td>59</td>
</tr>
<tr>
<td>molly_derip</td>
<td></td>
<td>19,994</td>
<td>425</td>
<td>89</td>
</tr>
<tr>
<td>consensus pixie</td>
<td></td>
<td>11,316</td>
<td>377</td>
<td>56</td>
</tr>
<tr>
<td>pixie_derip</td>
<td>Mariner/Tc1-like DNA transposon</td>
<td>6,173</td>
<td>257</td>
<td>58</td>
</tr>
<tr>
<td>consensus r25</td>
<td></td>
<td>861</td>
<td>96</td>
<td>59</td>
</tr>
<tr>
<td>r25_derip</td>
<td>histone H3 DNA transposon</td>
<td>1,009</td>
<td>84</td>
<td>52</td>
</tr>
<tr>
<td>r25_consensus r31</td>
<td>DNA transposon</td>
<td>199</td>
<td>199</td>
<td>199</td>
</tr>
<tr>
<td>r31_derip</td>
<td>Mariner/Tc1-like DNA transposon</td>
<td>1,803</td>
<td>361</td>
<td>124</td>
</tr>
<tr>
<td>r31_consensus r39</td>
<td></td>
<td>7,840</td>
<td>302</td>
<td>75</td>
</tr>
<tr>
<td>r39_derip</td>
<td></td>
<td>5,841</td>
<td>307</td>
<td>97</td>
</tr>
<tr>
<td>r51_derip</td>
<td>Mariner/Tc1-like DNA transposon</td>
<td>1,818</td>
<td>182</td>
<td>57</td>
</tr>
<tr>
<td>r51_consensus x0</td>
<td></td>
<td>1,709</td>
<td>570</td>
<td>181</td>
</tr>
<tr>
<td>x0_derip</td>
<td>Non-LTR retrotransposon</td>
<td>1,427</td>
<td>285</td>
<td>139</td>
</tr>
<tr>
<td>consensus x11</td>
<td>Gypsy-like retrotransposon</td>
<td>698</td>
<td>233</td>
<td>90</td>
</tr>
<tr>
<td>x11_derip</td>
<td></td>
<td>678</td>
<td>170</td>
<td>104</td>
</tr>
<tr>
<td>x5</td>
<td>Gypsy-like retrotransposon</td>
<td>115</td>
<td>115</td>
<td>115</td>
</tr>
<tr>
<td>x5_derip</td>
<td></td>
<td>324</td>
<td>324</td>
<td>324</td>
</tr>
<tr>
<td>consensus x96</td>
<td>uncharacterised</td>
<td>1,737</td>
<td>434</td>
<td>87</td>
</tr>
<tr>
<td>x96</td>
<td></td>
<td>276</td>
<td>138</td>
<td>106</td>
</tr>
<tr>
<td>x1</td>
<td>rDNA repeat (encoding large and small rRNA subunits)</td>
<td>5,940</td>
<td>297</td>
<td>50</td>
</tr>
</tbody>
</table>
3.3.9 Analysis of Repeat Induced Point Mutation (RIP) in *de novo* predicted repeats

RIPCAL analysis showed some evidence of RIP having occurred in PmedOMT5. An example is shown in Figure 3.6 which depicts the prevalence of CpA to TpA changes observed in a Molly-like transposon fragment. DeRIP analysis was uninformative, most likely due to the short nature of the repetitive regions identified (data not presented).

![Figure 3.6](image)

**Figure 3.6:** RIPCAL analysis of R62 (350 bp) showing high CpA ↔ TpA changes (red) relative to other CpN mutations in a short repeated fragment with similarity to the Molly transposon (CpA↔TpA mutations are shown in red).

The *rid* gene (*RIP* defective) [GenBank: AAM27408.1] a cytosine demethylase was found to be essential for RIP in *N. crassa* (Freitag *et al.* 2002). PmedOMT5 predicted version 1 protein PmedOMT5_2376 shows similarity to this protein at a similar level to that seen in other Dothideomycetes in which RIP is known to occur (BLASTP match bit score 135, e value 4x10^{-33}), details in Table 3.11.
Table 3.11: BLAST alignments to *N. crassa* rid protein (845 aa) [GenBank: AAM27408.1] in Dothideomycete genome assemblies (below significance threshold e= 1x10\(^{-5}\)).

<table>
<thead>
<tr>
<th>Species/isolate</th>
<th>BLAST search method</th>
<th>Bit score</th>
<th>E value</th>
<th>Database Searched</th>
<th>Region match in <em>N. crassa</em> RID protein</th>
<th>Region match in subject to <em>N. crassa</em> RID protein query</th>
<th>Match</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P. medicaginis</em> OMT5</td>
<td>TBLASTN</td>
<td>75.1</td>
<td>6x10(^{-33})</td>
<td>PmedOMT5scaffolds</td>
<td>94-365, 361-453, 453-592</td>
<td>157881-157072, 157037-156750, 156703-156263</td>
<td>Scaffold 731</td>
</tr>
<tr>
<td><em>P. medicaginis</em> OMT5</td>
<td>BLASTP</td>
<td>135</td>
<td>4x10(^{-33})</td>
<td>Pmed OMT5 predicted proteins v1</td>
<td>56-170, 367-581</td>
<td>64-183, 187-411</td>
<td>PmedOMT5_2376</td>
</tr>
<tr>
<td><em>P. nodorum</em> SN15</td>
<td>TBLASTN</td>
<td>89</td>
<td>4x10(^{-38})</td>
<td>StanoAssembly JGI</td>
<td>297-366, 364-450, 453-580</td>
<td>629076-628867, 628825-628562, 628506-628099</td>
<td>Scaffold 11</td>
</tr>
<tr>
<td><em>P. nodorum</em> SN15</td>
<td>BLASTP</td>
<td>45</td>
<td>8x10(^{-5})</td>
<td>SN15 proteins v 123</td>
<td>54-195</td>
<td>74-228</td>
<td>SNOT_07571.2</td>
</tr>
<tr>
<td><em>L. maculans</em> v23.1.3</td>
<td>BLASTP</td>
<td>198</td>
<td>1x10(^{-50})</td>
<td>NCBI non-redundat protein database (nr)</td>
<td>30-203, 259-608</td>
<td>35-234, 307-661</td>
<td>LEMA_P040230.1</td>
</tr>
<tr>
<td><em>L. maculans</em> v23.1.3</td>
<td>TBLASTN</td>
<td>98.2</td>
<td>6x10(^{-35})</td>
<td>NCBI non-redundat nucleotide database (nt)</td>
<td>66-459, 453-608</td>
<td>625308-626672, 626708-627196</td>
<td>Lm_SuperContig_A_v2</td>
</tr>
<tr>
<td><em>Pyrenophora tritici-repentis</em> Pt-1C-BFP</td>
<td>BLASTP</td>
<td>249</td>
<td>1x10(^{-47})</td>
<td>NCBI non-redundat protein database (nr)</td>
<td>7-587</td>
<td>33-624</td>
<td>PTRG_05633C-5 cytosine methyltransferase DmtA</td>
</tr>
</tbody>
</table>
Figure 3.7: Consensus alignment of top ten matches from NCBI protein database to PmedOMT5_02376. Intensity of colour indicates strength of conservation, bottom line shows consensus sequence.

The protein PmedOMT5T_02376 (v1) contains a conserved domain (Cyt_C5_DNA_methylase, cd00315, PFAM: PF00145) for C-5-cytosine-specific DNA methylation (from aa 156-404). A consensus alignment of the top ten hits in NCBI shows conserved amino acids around the domain region but much less across the length of the protein. This is illustrated in Figure 3.7 with a clustalw alignment (Larkin et al. 2007, Goujon et al. 2010) coloured by consensus.

3.4 Discussion

3.4.1 Genome assembly

The purpose of this genome assembly was to capture as much reliable information as possible about the gene content of the organism on a modest budget. Initial analysis shows a gene catalog not dissimilar from other Dothideomycete fungi, assembled into a reasonable number of scaffolds (965) for an NGS paired-end only assembly with a total length of ~31 Mb.

Scaffolds are made up of multiple contigs that are joined by read-pairs, whilst contigs are contiguous stretches of overlapping reads. Assembly algorithms are not always able to assemble across the gap between the connecting read-pairs and thus scaffolds may contain stretches of unknown nucleotides represented by N’s. Of the 952 nuclear scaffolds in the PmedOMT5 assembly, 365 contain gaps represented by stretches of Ns ranging from 10 to 171 bp. The length of gaps is created by predicting the amount of missing data between the contigs based on the expected insert size.

Scaffolds shorter than 200 bp were excluded from this assembly as these are not accepted for submission to the NCBI Whole Genome Shotgun (WGS) database,
however their constituent reads were retained and could be included in future rounds of assembly curation. These scaffolds are unlikely to contain complete gene models and are likely to have resulted from repetition in the genome or be artefacts due to sequencing errors.

Although there was a tailing off of quality scores at the ends of the reads as has been observed previously by many authors for Illumina-generated sequence data, at the time of initial assembly reads were not trimmed, as the quality scoring for this particular sequencing run was found to be not reliable. Because of this, adapter sequences were not trimmed from reads used for the assembly which may account for the inability of some of the reads to be incorporated into the final assembly. However the difference in the number of reads that could be mapped back to the final assembly via Bowtie2 was small (96.37% for reads trimmed of adapters using Scythe versus 95.04% of raw reads).

The PmedOMT5 assembly was completed using Velvet, a de novo genome assembly program that is known and trusted for its stringency and accuracy. In a comparison of de novo genome assembly algorithms, Velvet performed well, assembling over 95% of the yeast genome with 75 bp paired end (PE) data, with 100% of contigs correctly mapped based on simulated data compared to the reference genome (Zhang et al. 2011). Velvet was shown to produce the best assembly of S. cerevisiae (12Mb) from simulated reads in a comparison of five assemblers, in which only ABySS and Velvet were able to complete the task (Haiminen et al. 2011).

Previous authors have shown that a de novo assembly based on Illumina reads is sufficient to capture the majority of genes present in a small genome such as P. syringae 6 Mb (Farrer et al. 2009) and the fungus P. teres f. teres of ~42 Mb (Ellwood et al. 2010). The genome size of P. teres is in a similar range to that of PmedOMT5, estimated to be between 30-40 Mb, based on the known size of closely related species. Nowroussian et al., (2010) demonstrated that even an Illumina-only Velvet assembly of Sordaria macrospora (~40 Mb) with a large number of gaps, contained the majority of genes identified in a later assembly that included 454 data.

Based on comparisons to other sequenced fungi including closely related Dothideomycetes, no obviously mis-assembled regions have been identified in the Velvet assembly to date. Assembly statistics are summarised in Table 3.12. Although neither the rDNA repeat units nor complex repeats were represented well in the
assembly as has been previously observed for Solexa/Illumina assemblies (Nowrousian et al. 2010).

Table 3.12: Assembly statistics for final PmedOMT5 assembly.

<table>
<thead>
<tr>
<th>Assembly metrics</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of scaffolds</td>
<td>965</td>
</tr>
<tr>
<td>Number of nuclear scaffolds</td>
<td>952</td>
</tr>
<tr>
<td>Number of mitochondrial scaffolds</td>
<td>13</td>
</tr>
<tr>
<td>N50</td>
<td>91</td>
</tr>
<tr>
<td>L50/N50 length (bp)</td>
<td>105,011</td>
</tr>
<tr>
<td>Maximum scaffold length (bp)</td>
<td>374,744</td>
</tr>
<tr>
<td>Mean scaffold length (bp)</td>
<td>31,861</td>
</tr>
<tr>
<td>Minimum scaffold length (bp)</td>
<td>201</td>
</tr>
<tr>
<td>Number of scaffolds &gt; 1kbp in length</td>
<td>654</td>
</tr>
<tr>
<td>Total assembly size (bp)</td>
<td>30,777,866</td>
</tr>
<tr>
<td>Number of scaffolds with gaps</td>
<td>365</td>
</tr>
<tr>
<td>Number of unknown bases in assembly</td>
<td>36,608</td>
</tr>
<tr>
<td>Average nucleotide coverage of scaffolds containing genes</td>
<td>38.3</td>
</tr>
<tr>
<td>Percentage of repetitive DNA in assembly</td>
<td>1.35</td>
</tr>
<tr>
<td>Percentage GC content of assembly</td>
<td>53.07</td>
</tr>
<tr>
<td>Open reading frames predicted</td>
<td>1,324,085</td>
</tr>
<tr>
<td>Number of nuclear tRNAs identified</td>
<td>116</td>
</tr>
<tr>
<td>Number of scaffolds containing predicted ver. 1 genes</td>
<td>919</td>
</tr>
<tr>
<td>Scaffolds with one gene only</td>
<td>393</td>
</tr>
<tr>
<td>Number of scaffolds containing predicted genes ≥ 50aa</td>
<td>637</td>
</tr>
</tbody>
</table>
3.4.2 Assembly characteristics and genome structure

3.4.3 G:C content

The G:C content of fungal genomes can vary from as low as 32.52% in *Pneumocystis carinii* to 56.97% in *Phanerochaete chrysosporium* (Jung et al. 2008). The average G:C% of PmedOMT5 scaffolds was found to be 53.07%, which is slightly higher than that of other Dothideomycetes (see Table 3.13). This may be due to lower repetitive content present in the genome (and thus assembly) or an Illumina sequencing bias for higher coverage of G:C-equilibrated (~50%) DNA regions (Dohm et al. 2008, Hillier et al. 2008, Metzker 2010). The local G:C content of individual PmedOMT5 scaffolds ranged from 23.1-73.1%. Low average G:C content can indicate regions of A:T rich repeats. In *L. maculans* the average genomic G:C content was lower than average for Dothideomycetes at 44.1%, due to the presence of A:T-rich isochores, however coding regions displayed a normal G:C content (50.5%) (Rouxel et al. 2011). Jung et al. (2008) showed that there is no correlation between the proportion of ORFs in the genome and the G:C content. A possible explanation for this is that fungal species have been invaded by transposons at different times in their evolution and to different extents. These repeats may have subsequently been altered by repeat-induced point mutations (Hane and Oliver 2008, Clutterbuck 2010), which induces cytosine to thymine, thus lowering their overall G:C content but not significantly altering the gene content.

3.4.4 Estimation of PmedOMT5 genome size

Genome size varies widely in fungi mostly dependent upon the number of transposon invasions undergone. The sequenced Ascomycete fungal genome assemblies range in size from 12 Mb for *S. cerevisiae* to 125 Mb for the black truffle *Tuber melanosporum* (Martin et al. 2010), which contains 58% TEs.

A rough estimate of the PmedOMT5 genome size can be calculated based on read data, median coverage (39.5-fold) and the fact that 86.4% of reads were used to create the assembly. If it is assumed that the rest of the sequences that could not be assembled (3,052,124 reads) make up genomic regions with roughly the same coverage of 75 bp reads, this could be assembled into a complete genome of 36.6 Mbp ((3,052,124 x75)/39.5) + 30,777,866 = 36,573,038 nucleotides). This is close to the median euascomycete genome size calculated by Spanu et al., of 36.7 Mb (Spanu et al. 2010).
But this is simply a rough estimate, as the true final number will depend on how many of the sequences did not assemble due to errors or adapter contamination. This could also depend on the proportion of A:T rich regions that were not efficiently amplified by the Illumina chemistry and thus may have lower amounts of sequenced coverage that result in them not being included in the Velvet assembly due to low coverage. Thus it is predicted that the actual genome size of OMT5 is in the range of 30-40 Mb, which is similar to that predicted for the closely related *D. rabiei* based on cytological data of 23-34 Mb (Akamatsu *et al.* 2012). Quoted genome sizes for published species are likely to be underestimates as not all regions of fungal genomes can be completely assembled even from longer read Sanger sequenced data, notably centromeric and telomeric regions.

3.4.5 PmedOMT5 gene space

The majority of conserved core eukaryotic genes (99.6%) were identified in the PmedOMT5 genome assembly. There were only three genes that were listed as partial or absent: two genes (KOG3493- 73 aa, ubiquitin-like protein, KOG1555- 306 aa, 26S proteasome regulatory complex, subunit RPN11) were present but in regions of incomplete assembly, *i.e.* N’s present, and thus did not meet the strict CEGMA alignment requirements (Parra *et al.* 2009), that the proportion of the predicted gene in PmedOMT5 that aligned to the CEG profile was not greater than 70%.

The third gene (KOG1185- Thiamine pyrophosphate-requiring enzyme (a putative 2-hydroxyacyl-CoA lyase which catalyzes a carbon-carbon cleavage reaction, cleaving a 2-hydroxy-3-methylacyl-CoA into formyl-CoA and a 2-methyl-branched fatty aldehyde) appears to have been duplicated in PmedOMT5 and neither gene shows very strong remaining homology to the representative KOGs. The gene may also have been excluded during the CEGMA assessment of orthology which requires that a gene prediction must both align to the profile CEG and produce a score that exceeds any that can be produced from the alignment of any non-core gene (Parra *et al.* 2009). The missing CEG belongs to group 1, the least conserved group of CEGs.

Overall the CEGMA evaluation indicates a near complete capture of the gene space in PmedOMT5. This is on par with CEGMA analysis of the Sanger sequenced fungal genome assemblies of *N. crassa* and *M. grisea* which showed the presence of 245 genes (98.8%) and 246 complete genes (99.6%), respectively (Parra *et al.* 2009).
3.4.6 Gene prediction

To identify regions of the genome encoding genes, GeneMark-ES version 2.3 (Ter-Hovhannisyan et al. 2008), an ab initio gene prediction algorithm designed to work with fungal genomes was used. GeneMark uses Hidden Markov Models to model intron sites with and without branch points and a heuristic method for analysis of local G:C content to predict protein coding and non-coding regions. The GeneMark-E program determines the protein-coding potential of a DNA sequence (within a sliding window) by using species specific parameters of the Markov models of coding and non-coding regions (GeneMark). GeneMark was used to predict protein-encoding regions of the *P. teres* f. *teres* assembly (Ellwood et al. 2010) and the algorithm has been demonstrated to compare favourably to the accuracy of gene finders that employ supervised training based on EST data sets (Ter-Hovhannisyan et al. 2008). The number of predicted genes in PmedOMT5 greater than 50 aa (10,478) is comparable to that seen in other Dothideomycetes (Table 3.13).

The majority of the gene models predicted by GeneMark appear to be accurate (examined in detail in chapter 5) as indicated by homology to predicted genes in other fungi which were predicted using different algorithms and some of which have been biologically validated. This can be seen by the percentage (94.5% over 50 aa) of predicted genes with homology to the non redundant protein database at NCBI (See Figure 3.3).

Whilst most PmedOMT5 genes have BLASTP matches to known fungal proteins, not all have hits across the entire length of the GenBank sequence (Figure 3.3). It is impossible to know without further analysis if the apparent incompleteness of the match is due to differing gene structures between species or incomplete annotation in either genome. There are very few fully curated gene sets available amongst closely related species. Only one conserved gene locus appears to have been missed entirely by the GeneMark algorithm (3-dehydroquinate dehydratase, *Qa2*, *P. nodorum* NCBI accession: XM_001801631) which was identified in PmedOMT5 based on homology to other Dothideomycete genomes. This gene was also not predicted by FGENESH or GENESCAN analysis of the same genomic region (data not presented). Re-annotation and validation of the predicted gene models based on biological data and protein homology in closely related species is discussed in chapter 5.
Most fungal introns are short and have canonical splice sites, in which the first two nucleotides at the 5’ end of the intron are GU and the last before the 3’ exon are AG. Fungal introns have been observed with lengths ranging from 27-1,978 nucleotides (Kupfer et al. 2004) but more typically are between 50-100 bp. Predicted PmedOMT5 genes contained 0-13 introns. In S. pombe it was observed that as the number of introns in a gene increases by one the number of genes with that number of introns halves (Wood 2002). A similar pattern is observed in other fungi, including PmedOMT5, as illustrated in Figure 3.8.

![Figure 3.8: Number of genes with varying intron numbers across the published Dothideomycete species.](image)

The average gene size including introns is 1,617 bp which is consistent with other fungal genomes as illustrated in Table 3.13. The average CDS length of 1,422 bp was similar to that previously reported by Nowruussian et al. in Sordaria macrospora (1,432 bp) and the mean number of exons per gene (2.58) similar to that observed in P. teres f. teres (2.53) (Ellwood et al. 2010) and L. maculans (2.8) (Rouxel et al. 2011- supp. data).
3.4.7 Gene distribution

An important area of investigation in phytopathogenic Dothideomycete genome structure involves the identification of potential dispensable chromosomes that may contain transferable pathogenicity genes (Goodwin et al. 2011, Rouxel et al. 2011). Supernumerary or conditionally dispensable chromosomes (CDCs) are found in several fungal species and were first identified in *N. haematococca* in 1991 (Miao et al. 1991). By definition these chromosomes are not required for growth in all conditions but confer an adaptive advantage in certain habitats (Covert 1998), and they have often been found to contain genes essential for pathogenicity, such as the *SIX* (secreted in xylem) genes from the tomato pathogen, *F. oxysporum* f. sp. *lycoperscici* (Ma et al. 2010). Transfer of the CDC that the *SIX* genes are located on to an isolate of *F. oxysporum* that was non-pathogenic on tomato was sufficient to convert the isolate to a tomato pathogen (Ma et al. 2010). CDCs are often highly repetitive, gene sparse, with a lower G:C content than the rest of the genome and may contain genes unique to their species. They are often found only in certain members of a species. Regions of the assembly that are low in gene density may form parts of CDCs or contain potential effector genes in proximity to repetitive regions. To investigate this, the number of genes per scaffold was plotted against the length of the scaffold (Figure 3.2). Scaffolds that have a lower number of genes on average for their length were observed to contain more repetitive DNA although gene density appears relatively constant across the scaffolds. The average gene density in PmedOMT5 of 3.5 genes per 10 kb is consistent with that seen in other Dothideomycete assemblies, indicating that the PmedOMT5 assembly and its gene predictions are comparable to these species (Table 3.13). In conclusion there is no strong evidence for dispensable chromosomes in PmedOMT5.
Table 3.13: Assembly statistics for selected published Dothideomycete genomes.

<table>
<thead>
<tr>
<th>Genome features</th>
<th>PmedOMT5</th>
<th>Pn&lt;sup&gt;a&lt;/sup&gt;</th>
<th>Lm&lt;sup&gt;b&lt;/sup&gt;</th>
<th>Ptr&lt;sup&gt;c&lt;/sup&gt;</th>
<th>Mg&lt;sup&gt;d&lt;/sup&gt;</th>
</tr>
</thead>
<tbody>
<tr>
<td>Assembly size (Mb)</td>
<td>30.78</td>
<td>37.21</td>
<td>45.12</td>
<td>37.84</td>
<td>39.7</td>
</tr>
<tr>
<td>GC%</td>
<td>53.2</td>
<td>50.3</td>
<td>44.1</td>
<td>50.9</td>
<td>51.7</td>
</tr>
<tr>
<td>Predicted protein coding sequences</td>
<td>10,478</td>
<td>12,194</td>
<td>12,469</td>
<td>12,171</td>
<td>10,933</td>
</tr>
<tr>
<td>Gene density per 10 kb</td>
<td>3.5</td>
<td>3.1</td>
<td>2.8*</td>
<td>3.8</td>
<td>3.4</td>
</tr>
<tr>
<td>Average gene length (bp)</td>
<td>1,617</td>
<td>1,326</td>
<td>1,323</td>
<td>1,618</td>
<td>1,144</td>
</tr>
</tbody>
</table>

Data obtained from: a (Hane et al. 2011b), b (Rouxel et al. 2011), c (Ciufetti 2008), d (Goodwin et al. 2011). Species names: Pn- P. nodorum, Lm- L. maculans, Ptr- P. tritici-repentis, Mg- M. graminicola
*gene density in core genome excluding repeated elements but including gaps in genome sequence only = 4.2 (Rouzel et al, 2011)

3.4.8 Non coding RNA

An important part of determining gene location and structure in the genome is also to determine where they are not, i.e. excluding repeats, rRNA and other non-coding RNA (ncRNA) regions from consideration. The main incentive behind the ncRNA analysis was to identify short gene models that had been mistakenly identified as novel PmedOMT5 proteins during annotation (chapter 5) and thus remove them from subsequent effector analysis. For this purpose the software package Infernal (Nawrocki et al. 2009) was used to search for matches in the genome to sequences from the Rfam database, including transfer RNAs and ribosomal RNA as well as experimentally identified and bioinformatically predicted small, micro and small nucleolar (sno) RNAs. Using very strict criteria, 208 conserved ncRNAs were identified in the PmedOMT5 genome (Table 3.14).

No ncRNA overlap was detected with any of the effector candidate genes described in chapters 6 and 7. The most abundant ncRNAs in the nuclear genome assembly are the tRNAs, followed by replicates of the 5S rRNA, copies of which are found scattered throughout the genome, as is typical in fungi (Rooney and Ward 2005). The next most
abundant type was small nucleolar RNAs (snoRNAs) which play a role in guiding the chemical modification of other RNAs. The ncRNA analysis revealed a similar profile to that observed in another Ascomycete fungus *Aspergillus fumigatus* (Jöchl *et al.* 2008).

**Table 3.14:** Number and classification of non-coding RNAs (excluding tRNAs) identified in the PmedOMT5 genome.

<table>
<thead>
<tr>
<th>ncRNA type</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>5S rRNA (ribosomal)</td>
<td>34</td>
</tr>
<tr>
<td>snoRNA (small nucleolar)</td>
<td>26</td>
</tr>
<tr>
<td>microRNA</td>
<td>9</td>
</tr>
<tr>
<td>snRNA (small nuclear)</td>
<td>7</td>
</tr>
<tr>
<td>other</td>
<td>4</td>
</tr>
<tr>
<td>tmRNA (transfer-messenger)</td>
<td>2</td>
</tr>
<tr>
<td>lncRNA (long non-coding)</td>
<td>1</td>
</tr>
<tr>
<td>Riboswitch regulating gene expression</td>
<td>1</td>
</tr>
<tr>
<td>RNase</td>
<td>1</td>
</tr>
</tbody>
</table>

### 3.4.9 Transfer RNAs

From the 116 tRNA genes predicted by tRNAscan, the majority of tRNAs (18) with anti-codons for the twenty standard amino acids were identified (excluding His and Cys). These two amino acids were present in translated proteins indicating that the tRNAs are simply missing from the assembly or in mis-assembled regions. In the *S. pombe* genome assembly tRNAs were mostly identified near centromeres (Wood *et al.* 2002) which may account for the low number seen in PmedOMT5, due to the difficulty in assembling the repetitive regions near the centromeres. The number of tRNAs in other fungi varies considerably, with 275 in *S. cerevisiae* (Goffeau *et al.* 1996) and 178 in *A. fumigatus* (Chan and Lowe 2009). In general a lower number is seen in the Dothideomycetes: *L. maculans* (95) (Rouxel *et al.* 2011-supp. data) and *P. nodorum* (117 including pseudo tRNAs) (http://www.broadinstitute.org/annotation/genome/stagonospora_nodorum/GenomeStats.html- accessed 130704). As the majority of tRNAs for all common amino acids were identified in PmedOMT5, this further indicates that the assembly is likely to accurately represent the majority of functionally important regions of the genome.

### 3.4.10 Analysis of repetitive DNA content

Repeats are known to be drivers of genome evolution (Kazazian 2004) which can sometimes play a beneficial role, facilitating the alteration of gene expression and sequence. In fungi even simple repeats can lead to genomic rearrangement as evidenced
by Ellison et al., (2011) where a low complexity sequence near the MAT region of Neurospora tetrasperma appears to have facilitated the inversion of a 68 kb region.

Current assembly algorithms are unable to correctly re-assemble regions of high sequence similarity which occur multiple times in the genome from short reads (75 bp, for PmedOMT5) because of the difficulty of resolving repetitive regions longer than the insert length between paired sequenced reads. In the case of PmedOMT5 this distance was ~220 bp, whilst the average length of known fungal transposons in Repbase is 2,719 bp (Repbase 15.6). It was possible to identify fractions of 40 repetitive element families with lengths ranging from 30-3,309 bp.

The main aim of the repeat analysis was to identify the location of repetitive regions within the genome that may be proximal to potential effectors, which have the potential to be affected by RIP leakage (Van de Wouw et al. 2010) and to gain insights into possible regions of genome rearrangement which may also include regions transferred from other species via lateral gene transfer (LGT). Several methods were used to identify PmedOMT5 repeats and the position they might occupy in the genome (summarised in Figure 3.4), which can be postulated from their occurrence at the ends of scaffolds. However the comprehensive number and length of all repeats cannot be ascertained from an NGS assembly, as each break in the assembly may contain multiple repeats.

3.4.11 Types of repetitive DNA in PmedOMT5

Only 16 out of the 40 repeat families identified de novo in PmedOMT5 had BLAST matches below a significance threshold of 1x 10^{-5} (or even 1x10^{-2}) that could be used for identification. Based on similarity to fungal repeats found in REPBASE (Kapitonov and Jurka 2008), the most predominant repeat family is Mariner class transposons which include the Molly transposon [GenBank: AJ488502.1] identified in P. nodorum, that are represented by over 90 regions within the PmedOMT5 assembly totalling over 43 kb.

Most of the repeat regions identified appear to be small parts of whole transposon copies that could not be assembled in their entirety. It is difficult to estimate what proportion of each repeat family has been assembled in PmedOMT5 as the lengths of repeats catalogued in REPBASE vary in length depending on the species they were
identified in. The length of Mariner class transposons can range from 484 bp in *A. oryzae* to 2,025 bp in *E. nidulans* (Repbase15.0) whilst the Molly consensus sequence is 1,897 bp in *P. nodorum* SN15.

The total predicted repeat content of PmedOMT5 by RepeatMasker is only 1.35%, in contrast with the higher numbers seen in other published Dothideomycete assemblies: 4.5% in *P. nodorum* SN15 (Hane *et al.* 2007), 21% in *M. graminicola* (Goodwin *et al.* 2011) and 34% in *L. maculans* (Rouxel *et al.* 2011).

A very small percentage of the ~three million reads not used in the assembly (3,052,124 reads) can be aligned to known fungal repetitive DNA (0.3 % to SN15 repeats and 0.04% to Repbase fungal repeats) when using Bowtie2 in a relaxed alignment mode. This is a further indication that the majority of the unassembled reads do not consist of known repeats.

### 3.4.12 Evidence for RIP in PmedOMT5

RIP analysis was performed in order to identify the potential for RIP leakage from transposon insertions in the genome to putative effector candidates. *De novo* repeats were analysed for the presence of RIP mutations using RIPCAL (Hane *et al.* 2007). Most of the transposon fragments identified show RIP like changes between different copies of the repeat in the assembly. RIP dominance scores were calculated to examine the type of RIP most prevalent in PmedOMT5. In the majority of cases, CpA\(\leftrightarrow\)TpA was the dominant form followed by CpT\(\leftrightarrow\)TpT. This is consistent with the mutation biases observed in the majority of filamentous Ascomycetes (Clutterbuck 2010).

In *L. maculans* repeat classes that showed a RIP dominance score of greater than one were classified as heavily ‘RIPed’ (Van de Wouw *et al.* 2010). Figure 3.6 illustrates a short repetitive sequence (R62-consensus length 350 bp) that has similarity to the Molly repeat identified in *P. nodorum*, with a CpA:TpA dominance score of 1.82. This indicates that the original transposon insertion has undergone RIP. The majority of other *de novo* identified repeats were too short to comprehensively analyse the extent of RIP, although in general more RIP-like changes were observed in longer identified repeats (*e.g.* Molly fragment (959 bp) on scaffold 2,433 contains multiple stop codons in the protein sequence for all 6 frames). This pattern suggests that RIP may be occurring or have occurred in the past, but this is difficult to assess due to the inability to assemble
long repetitive regions. The presence of RIP may serve as evidence for the hypothesis that *P. medicaginis* does exist as a sexual species but the teleomorph has not yet been identified.

RIP is not active in all fungal species (Clutterbuck 2010, Nowrousian *et al.* 2010) and the methods by which it occurs are yet to be fully elucidated (Clutterbuck 2010). The only gene known to be essential for RIP is the *N. crassa* *rid* gene [GenBank: AAM27408.1]. All members of the subphylum Pezizomycotina analysed by Clutterbuck, which showed evidence of RIP, contained genes with similarity to the *rid* gene (Clutterbuck 2010). The protein PmedOMT5T_02376 (v1) shows some similarity to this protein, illustrated in Figure 3.7 (bit score 135) and is predicted to have cytosine methyltransferase activity. The presence of this gene is further evidence that RIP may be actively occurring in PmedOMT5. However, *S. macrospora* also possesses a homolog of the *N. crassa* *rid* gene but there is no evidence of active RIP (Nowrousian *et al.* 2010). Transcript analysis of RNA-seq data (chapter 5) suggests that PmedOMT5 still contains several actively transcribed transposons. Illumina based RNA-seq analysis of PmedOMT5 performed during several growth phases (outlined in detail in chapter 5): including vegetative growth, *in vitro* asexual sporulation (16d on minimal media) and pathogen growth *in planta* showed very little expression of the *rid* homolog (PmedOMT5_02870 v2 protein) with only 20 mapped reads across all conditions. However this gene would not be expected to be highly expressed under the non-meiotic conditions tested.

### 3.4.13 Future strategies for genome finishing

Almost all fungal genome assemblies remain ‘unfinished’ due to the difficulty of assembling around repetitive regions. Whilst it is evident that the majority of the gene content can still be examined in these assemblies, for more complete evolutionary analysis it may be necessary to improve the length of the contiguous sequences to ensure no important data is missing. There are several techniques that can be used to refine assemblies, collectively known as “genome finishing”.

Finishing the genome of PmedOMT5 would first require the determination of the number and size of the chromosomes. One method which is commonly used to assess this in fungal genome analysis is electrophoretic karyotyping (Carle and Olson 1985, Chu *et al.* 1986) which separates chromosomes based on their different mobilities in an
electric field of altering polarities. The order of the existing scaffolds must then be mapped to the chromosomes using techniques such as genetic linkage mapping (not currently possible for PmedOMT5 unless a second mating type is identified), optical mapping: which produces an ordered restriction map from a single DNA molecule (Schwartz et al. 1993) or HAPPY mapping: creation of a map from fragmented DNA based on co-amplification of fragments (Dear and Cook 1993). These techniques combined with sequencing of large insert mate-pair libraries or Sanger sequenced BACs which can straddle repeats and thus join contigs, third generation sequencing technologies such as SMRT from Pacific Biosciences (Chin et al. 2013) or the single molecule nanopore sequencer under development by Oxford Nanopore Technologies (Check-Hayden 2012-02-17), would enable the creation of larger scaffolds and potentially complete chromosomal sequences.

Many filamentous Ascomycete fungi, especially in the class Dothideomycetes, display a pattern of chromosome conservation known as mesosynteny, where genes show a tendency to be retained on equivalent chromosomes, although the order and orientation of genes are rearranged (Hane et al. 2011a). This pattern has been exploited in the genome finishing strategy of M. graminicola, the first fully “finished” telomere-to-telomere sequenced Dothideomycete (Goodwin et al. 2011). This strategy could also potentially be used to improve the PmedOMT5 assembly, by comparison to a more completely assembled closely related species.

3.5 Summary

The work presented in this chapter describes the strategies used to create the first reference assembly of a P. medicaginis genome using only short-sequence reads. This draft assembly compares well to other next-generation short read assemblies and facilitated the analysis of its genomic landscape, including gene content, non-coding RNA and repetitive regions. Comparison to other Dothideomycete genome assemblies created using first generation sequencing techniques shows a similarity in terms of gene content and structure, which demonstrates the validity of the strategies used. The PmedOMT5 genome will serve as an important resource for Phoma research and studies of fungal genome evolution.
Chapter 4
The mitochondrial genome of PmedOMT5
4.1 Introduction

Mitochondria are essential organelles present in virtually all eukaryotic cells, which originated from an ancestral alpha-proteobacterial endosymbiont (Gray et al. 1999). All fungi contain mitochondria in varying numbers per cell. Their primary function is the production of adenosine triphosphate (ATP) which serves as an energy source for the fungal cell. Mitochondria contain their own separate genome with prokaryote features and many use a non-standard genetic code. The mitochondrial genome (mtDNA) encodes some of the genes required for its own expression including rRNAs and tRNAs as well as genes for ATP production and is present in high copy number relative to the nuclear genome. Some mitochondrial genes are also encoded by the nuclear genome, mostly those involved in transport and regulatory functions (Karlberg et al. 2000). Whilst mitochondrial genes are not presently known to play a direct role in pathogenicity they do have a role in fungicide resistance, with mutations in the cytochrome b (cytb) gene conferring resistance to the strobilurin group of fungicides (Gisi et al. 2000).

4.1.1 Mitochondrial genome size and structure

Fungal mitochondrial genomes are most commonly found as single circular DNA molecules, although they can be linear, or segmented with linear and circular components (Hausner et al. 2003). Genes are often encoded on only one strand (Kennell and Cohen 2004), although in the mitochondria of Dothideomycetes M. graminicola and P. nodorum both strands encode genes (Hane et al. 2007, Torriani et al. 2008).

Fungal mitochondrial genomes can vary significantly in size, even within a genus, but are generally in the 30-80 kb range (Clarke-Walker 1992). The size variation is most often due to optional introns and to variations in the size of non-coding intergenic sequences that are often repetitive. Smaller size differences may also be due to varying gene content i.e. differing copy number of tRNAs and ORFs with unknown functions (Clarke-Walker 1992).

4.1.2 Aims for this chapter

The goal of this chapter was to identify and assemble the PmedOMT5 mtDNA and to annotate the mitochondrial genes for future comparative analysis and as potential targets for host defences or fungicides.
4.2 Methods
Mitochondrial scaffolds were identified by BLASTN (Altschul et al. 1990) of PmedOMT5 scaffolds to *P. nodorum* (SN15) *L. maculans* (v23.1.3), *M. graminicola* and *Sordaria macrospora* mitochondrial genomes with an e value threshold of $1 \times 10^{-5}$ and TBLASTN with *P. nodorum*, *N. crassa* and *S. macrospora* mitochondrial proteins. BLASTN analysis of SN15 mitochondrial ribosomal DNA sequences were used to identify the small and large rDNA subunits.

Scaffolds that were predicted to be mitochondrial based on BLAST hits were examined for the presence of tRNAs using tRNAscanSE using the following parameters: search mode, Cove only, source- Mito/chloroplast, genetic code for tRNA isotype prediction- Mold and Protozoan mito (Lowe and Eddy 1997).

Predicted mitochondrial scaffolds were further examined for the presence of predicted nuclear encoded genes and G:C content consistent with mitochondrial DNA using Geneious Pro 5.3.6 (Drummond et al. 2012).

MetaGeneMark (v 2.7d), a gene prediction program designed for use with microbial genomes was used to try to predict gene borders (Zhu et al. 2010) using the parameters: Kingdom-Archaea or bacteria, from scaffolds identified to be mitochondrial.

Bowtie2 (v2.0.0 beta 5) (Langmead and Salzberg 2012) was used to identify reads that were mitochondrial in origin but had not formed part of the original Velvet assembly (chapter 3.3.3) by mapping unassembled reads to the *P. nodorum* SN15 mitochondrial genome (Hane et al. 2007) (parameters: -N 1, –local). Bowtie2 mapping of all PmedOMT5 genomic reads was also conducted (parameter: sensitive). Mitochondrial scaffolds were concatenated using SSPACE v.1.2 (Boetzer et al. 2010) SSPACE parameters used: were default: –k 3 –x 1).

Primers were designed to amplify out from the ends of all three scaffolds using Primer 3 v 0.4.0 (Rozen and Skaletsky 2000). Primers were tested in all possible combinations using the iProof PCR protocol described in section 2.4.3 with an amplification temperature of 54 °C and an extension time of 30 sec. PCR products were subjected to electrophoresis in 1% agarose for 40 min at 70 volts.
Table 4.1: Primer sequences used to amplify mitochondrial scaffolds.

<table>
<thead>
<tr>
<th>Primer name</th>
<th>length</th>
<th>Tm</th>
<th>GC%</th>
<th>5'-3' sequence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pm428_F2</td>
<td>22</td>
<td>60.06</td>
<td>50</td>
<td>GCTAAAGGAGGAGGAGGATAA</td>
</tr>
<tr>
<td>Pm428_R3</td>
<td>22</td>
<td>59.89</td>
<td>45.45</td>
<td>ATACTAACCTCCAGGAGCAAA</td>
</tr>
<tr>
<td>Pm524_F4</td>
<td>22</td>
<td>60.1</td>
<td>40.91</td>
<td>AGATCGGATGGCTCTTTTAA</td>
</tr>
<tr>
<td>Pm524_R3</td>
<td>23</td>
<td>60.15</td>
<td>43.48</td>
<td>GGTGTATGACAGACATTCTGC</td>
</tr>
<tr>
<td>Pm463_R2</td>
<td>22</td>
<td>59.66</td>
<td>45.45</td>
<td>GATAAGTCCGCAAACTCTGCT</td>
</tr>
<tr>
<td>Pm463_F1</td>
<td>22</td>
<td>60.13</td>
<td>45.45</td>
<td>AGGAAAGCTTCACAGGACAATA</td>
</tr>
</tbody>
</table>

Following PCR confirmation of predicted joins an entire mitochondrial scaffold was created linked by N’s where sequence could not be accurately determined. This final scaffold was again examined by MetaGeneMark, tRNAscanSE and BLAST in order to create a final manually annotated mitochondrial gene set (Appendix 4.1: mitochondrion gff3). CGView Server (Grant and Stothard 2008) was used to generate a figure illustrating the layout of the genome. For comparative purposes the SN15 mitochondrion was illustrated in the same manner.

Transcriptome reads (details provided in chapter 5.1.3) were aligned to the assembled mitochondrial scaffold to examine expression of predicted genes using TopHat v2.0.4 (Trapnell et al. 2009) (parameters: -very-sensitive -r 50 --mate-std-dev 100 -i 20 -I 4000 -g 20 -m 0 --min-coverage-intron 20 --coverage-search --microexon-search).

4.3 Results

4.3.1 Identification and analysis of PmedOMT5 mitochondrial scaffolds

Thirteen scaffolds assembled by Velvet, totalling 38,051 bp in length, were predicted to originate from the mitochondrial genome based on BLAST data and were separated from the nuclear assembly. These scaffolds contained portions of all genes usually encoded by the mitochondrion in fungi. None of the predicted mitochondrial contigs contained gaps. These scaffolds and the genes contained within them are listed in Table 4.2.

4.3.2 Scaffold joining

In order to create an assembly of the entire mtDNA the orientation of the Velvet scaffolds was examined using Bowtie2. Bowtie2, is an alignment tool which aligns reads to a reference sequence allowing for small differences in sequence. Bowtie2 analysis using all PmedOMT5 reads showed that many of the paired-end reads mapped
with one end of a pair to one mitochondrial scaffold, whilst the other mapped to a different mitochondrial scaffold. Therefore an attempt to join scaffolds using SSPACE v.1.2 (Boetzer et al. 2010) was made which further concatenated the mitochondrial scaffolds. SSPACE is a scaffolding tool that analyses paired-end sequence data utilising Bowtie to assess the order and orientation of contigs produced by a de novo assembler. Where possible it extends and joins the assembled contigs based on the read pair information. The 13 mitochondrial scaffolds were concatenated by SSPACE to create three scaffolds totalling 37,917 bp in length with five unknown bases. Joins were based on reads which had pairs anchored in different scaffolds and some small overlaps between extended scaffold ends. The sizes of these new scaffolds are presented in Table 4.3. To determine the order of these 3 scaffolds around the mtDNA they were tentatively assembled into a contiguous sequence joined by N’s to represent unknown sequence based on alignment to other mitochondrial genomes from the class Dothideomycetes. As mitochondrial genomes can be extensively re-shuffled even between closely related species, primers were designed to amplify out from the ends of all three scaffolds to confirm the predicted orientation.

PCR amplification from the ends of the SSPACE concatenated scaffolds produced three bands indicating the order in which the scaffolds were aligned around the mitochondrial genome was as predicted (Figure 4.1). Distance from the designed primers to the end of the scaffolds is listed in Table 4.4. Incorporation of the predicted sized gaps between the SSPACE joined scaffolds resulted in assembly of the complete PmedOMT5 mitochondrial genome of approximately 38,538 bp. A single pseudo-molecule was created containing a total of 626 (621+5) unknown bases.

4.3.3 Annotation of mitochondrial features

Following the analysis described above, the gene features of the assembled mitochondrial genome were manually annotated to create the gene data presented in Table 4.5 (Appendix 4.1: mitochondrion gff3). Analysis with tRNAscanSE (Lowe and Eddy 1997) showed the presence of single copies of all essential tRNAs, with duplicate copies of Arginine, Leucine, Methionine and Serine. Co-ordinates and tRNA anticodons are listed in appendix 4.2.
Table 4.2: PmedOMT5 scaffolds predicted to form part of the mitochondrial genome prior to SSPACE scaffolding and PCR joining.

<table>
<thead>
<tr>
<th>Scaffold</th>
<th>%G:C</th>
<th>Scaffold length</th>
<th>Mitochondrial features present (includes partial)</th>
<th>Scaffold nucleotide coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>24</td>
<td>35.98</td>
<td>1,804</td>
<td>ss rRNA</td>
<td>913</td>
</tr>
<tr>
<td>35</td>
<td>34.39</td>
<td>2,172</td>
<td>ls rRNA, Thr1, Met1, Glu1, Ala1, Phe1, Leu1, Gln1, His1, Met2, <em>atp6</em>, Cys1, <em>cox1</em>-2</td>
<td>802</td>
</tr>
<tr>
<td>66</td>
<td>27.81</td>
<td>5,962</td>
<td><em>ls rRNA, Thr1, Met1, Glu1, Ala1, Phe1, Leu1, Gln1, His1, Met2, atp6, Cys1, cox1-2</em></td>
<td>154</td>
</tr>
<tr>
<td>215</td>
<td>29.68</td>
<td>8,473</td>
<td><em>nad4, nad3, nad2, cox3, Arg1</em></td>
<td>233</td>
</tr>
<tr>
<td>224</td>
<td>28.3</td>
<td>2,572</td>
<td>Leu2, Tyr1, Asn1, <em>nad6</em>, Val1, Lys1, Gly1, Asp1</td>
<td>236</td>
</tr>
<tr>
<td>243</td>
<td>30.44</td>
<td>2,710</td>
<td><em>cytb</em></td>
<td>247</td>
</tr>
<tr>
<td>294</td>
<td>26.99</td>
<td>3,224</td>
<td><em>rps5, nad1.1</em></td>
<td>110</td>
</tr>
<tr>
<td>299</td>
<td>30.77</td>
<td>3,422</td>
<td>Trp1, Ile1, Ile2, Arg2, Ser1, Pro1, ls RNA</td>
<td>442</td>
</tr>
<tr>
<td>373</td>
<td>32.43</td>
<td>3,207</td>
<td><em>nad4L, nad5.1</em></td>
<td>338</td>
</tr>
<tr>
<td>628</td>
<td>29.34</td>
<td>1,145</td>
<td><em>cox1</em>-2, LAGLIDADG endonuclease</td>
<td>201</td>
</tr>
<tr>
<td>1,039</td>
<td>27.27</td>
<td>407</td>
<td>Ser2</td>
<td>119</td>
</tr>
<tr>
<td>1,237</td>
<td>30.74</td>
<td>2,619</td>
<td><em>cox1</em>-2</td>
<td>261</td>
</tr>
<tr>
<td>1,804</td>
<td>32.93</td>
<td>334</td>
<td><em>cox1</em>-2</td>
<td>327</td>
</tr>
<tr>
<td>Total/Average</td>
<td>30.54</td>
<td>38,051</td>
<td></td>
<td>337</td>
</tr>
</tbody>
</table>
Table 4.3: Mitochondrial scaffolds joined by SSPACE.

<table>
<thead>
<tr>
<th>SSPACE mitochondrial scaffolds</th>
<th>Length (bp)</th>
<th>Order in which velvet contigs are joined to create SSPACE scaffolds*</th>
</tr>
</thead>
<tbody>
<tr>
<td>428.1</td>
<td>18,801</td>
<td>215F_24R_224F_1039R_299F_35F</td>
</tr>
<tr>
<td>463.1</td>
<td>13,156</td>
<td>66F_1804F_628F_1237F_373F</td>
</tr>
<tr>
<td>524.1</td>
<td>5,960</td>
<td>294F_243R</td>
</tr>
<tr>
<td>Total length</td>
<td>37,917</td>
<td></td>
</tr>
</tbody>
</table>

*F and R indicate the orientation of the original Velvet scaffold in the SSPACE scaffold.

Figure 4.1: Gel showing amplification between predicted adjacent SSPACE scaffolds but not others. Lane 1 primer combination: 428F-463R, lane 3: 524F-428R, lane 5: 463F 524R.

Table 4.4: Amplifying primer pairs and their distance from the scaffold ends.

<table>
<thead>
<tr>
<th>Primer pair</th>
<th>~amplicon size(bp)</th>
<th>Distance of primers from scaffold end</th>
<th>~ gap length</th>
</tr>
</thead>
<tbody>
<tr>
<td>428F-463R</td>
<td>900</td>
<td>279+392</td>
<td>229</td>
</tr>
<tr>
<td>524F-428R</td>
<td>950</td>
<td>418+297</td>
<td>235</td>
</tr>
<tr>
<td>463F 524R</td>
<td>1200</td>
<td>648+395</td>
<td>157</td>
</tr>
<tr>
<td>Gap length</td>
<td></td>
<td></td>
<td>621</td>
</tr>
<tr>
<td>Total SSPACE joined scaffold length</td>
<td></td>
<td></td>
<td>37,917</td>
</tr>
<tr>
<td>Total size: SSPACE scaffolds plus gaps</td>
<td></td>
<td></td>
<td>38,538</td>
</tr>
</tbody>
</table>
Table 4.5: Features of the final PmedOMT5 mitochondrial genome sequence.

<table>
<thead>
<tr>
<th>Name of feature</th>
<th>Feature type</th>
<th>Start</th>
<th>End</th>
<th>Strand</th>
</tr>
</thead>
<tbody>
<tr>
<td>nad4</td>
<td>gene</td>
<td>50</td>
<td>1954</td>
<td>-</td>
</tr>
<tr>
<td>nad3</td>
<td>gene</td>
<td>2473</td>
<td>3285</td>
<td>-</td>
</tr>
<tr>
<td>nad2</td>
<td>gene</td>
<td>3286</td>
<td>5013</td>
<td>-</td>
</tr>
<tr>
<td>cox3(exon 2)</td>
<td>exon</td>
<td>5197</td>
<td>5374</td>
<td>-</td>
</tr>
<tr>
<td>cox3 (exon 1)</td>
<td>exon</td>
<td>6015</td>
<td>6643</td>
<td>-</td>
</tr>
<tr>
<td>Unknown open reading frame</td>
<td>ORF</td>
<td>7238</td>
<td>7867</td>
<td>+</td>
</tr>
<tr>
<td>Arg 1</td>
<td>tRNA</td>
<td>8149</td>
<td>8219</td>
<td>+</td>
</tr>
<tr>
<td>Small ribosomal subunit</td>
<td>rRNA</td>
<td>8761</td>
<td>9881</td>
<td>+</td>
</tr>
<tr>
<td>Leu 1</td>
<td>tRNA</td>
<td>10555</td>
<td>10617</td>
<td>+</td>
</tr>
<tr>
<td>Tyr</td>
<td>tRNA</td>
<td>11046</td>
<td>11130</td>
<td>+</td>
</tr>
<tr>
<td>Asn</td>
<td>tRNA</td>
<td>11276</td>
<td>11346</td>
<td>+</td>
</tr>
<tr>
<td>nad6</td>
<td>gene</td>
<td>11695</td>
<td>12291</td>
<td>+</td>
</tr>
<tr>
<td>Val</td>
<td>tRNA</td>
<td>12380</td>
<td>12452</td>
<td>+</td>
</tr>
<tr>
<td>Lys</td>
<td>tRNA</td>
<td>12510</td>
<td>12581</td>
<td>+</td>
</tr>
<tr>
<td>Gly</td>
<td>tRNA</td>
<td>12597</td>
<td>12667</td>
<td>+</td>
</tr>
<tr>
<td>Asp</td>
<td>tRNA</td>
<td>12670</td>
<td>12741</td>
<td>+</td>
</tr>
<tr>
<td>Ser 1</td>
<td>tRNA</td>
<td>12942</td>
<td>13021</td>
<td>+</td>
</tr>
<tr>
<td>Trp</td>
<td>tRNA</td>
<td>13343</td>
<td>13414</td>
<td>+</td>
</tr>
<tr>
<td>Ile</td>
<td>tRNA</td>
<td>13523</td>
<td>13594</td>
<td>+</td>
</tr>
<tr>
<td>Arg 2</td>
<td>tRNA</td>
<td>13667</td>
<td>13738</td>
<td>+</td>
</tr>
<tr>
<td>Ser 2</td>
<td>tRNA</td>
<td>14162</td>
<td>14246</td>
<td>+</td>
</tr>
<tr>
<td>Pro</td>
<td>tRNA</td>
<td>14770</td>
<td>14842</td>
<td>+</td>
</tr>
<tr>
<td>Large ribosomal subunit</td>
<td>rRNA</td>
<td>14937</td>
<td>17604</td>
<td>+</td>
</tr>
<tr>
<td>Thr</td>
<td>tRNA</td>
<td>18536</td>
<td>18606</td>
<td>+</td>
</tr>
<tr>
<td>Met 1</td>
<td>tRNA</td>
<td>18631</td>
<td>18701</td>
<td>+</td>
</tr>
<tr>
<td>Gln</td>
<td>tRNA</td>
<td>19659</td>
<td>19731</td>
<td>+</td>
</tr>
<tr>
<td>Ala</td>
<td>tRNA</td>
<td>19755</td>
<td>19826</td>
<td>+</td>
</tr>
<tr>
<td>Phe</td>
<td>tRNA</td>
<td>20120</td>
<td>20192</td>
<td>+</td>
</tr>
<tr>
<td>Leu 2</td>
<td>tRNA</td>
<td>20493</td>
<td>20574</td>
<td>+</td>
</tr>
<tr>
<td>Gln</td>
<td>tRNA</td>
<td>20652</td>
<td>20723</td>
<td>+</td>
</tr>
<tr>
<td>His</td>
<td>tRNA</td>
<td>20729</td>
<td>20801</td>
<td>+</td>
</tr>
<tr>
<td>Met 2</td>
<td>tRNA</td>
<td>21189</td>
<td>21260</td>
<td>+</td>
</tr>
<tr>
<td>atp6</td>
<td>gene</td>
<td>22603</td>
<td>23379</td>
<td>+</td>
</tr>
<tr>
<td>Cys</td>
<td>tRNA</td>
<td>23526</td>
<td>23595</td>
<td>+</td>
</tr>
<tr>
<td>cox1,2 polyprotein (exon1)</td>
<td>exon</td>
<td>24409</td>
<td>25532</td>
<td>+</td>
</tr>
<tr>
<td>LAGLIDADG</td>
<td>Intrinsic endonuclease</td>
<td>25569</td>
<td>26567</td>
<td>+</td>
</tr>
<tr>
<td>cox1,2 polyprotein (exon2)</td>
<td>exon</td>
<td>26626</td>
<td>27870</td>
<td>+</td>
</tr>
<tr>
<td>nad4L</td>
<td>gene</td>
<td>29163</td>
<td>29432</td>
<td>+</td>
</tr>
<tr>
<td>nad5</td>
<td>gene</td>
<td>30134</td>
<td>31972</td>
<td>+</td>
</tr>
<tr>
<td>nad1</td>
<td>gene</td>
<td>32344</td>
<td>33414</td>
<td>-</td>
</tr>
<tr>
<td>rps5</td>
<td>gene</td>
<td>34012</td>
<td>35319</td>
<td>-</td>
</tr>
<tr>
<td>cyt b</td>
<td>gene</td>
<td>36722</td>
<td>37891</td>
<td>-</td>
</tr>
</tbody>
</table>

4.3.4 Mapping of genomic and transcriptomic Illumina reads

Unassembled PmedOMT5 reads were mapped to the *P. nodorum* SN15 mtDNA to estimate the amount of the PmedOMT5 mtDNA that had not been able to be assembled into scaffolds (Table 4.6). The relatively low number indicated that although the
predicted PmedOMT5 mitochondrial scaffolds had a total length ~ 10kb shorter than the SN15 mtDNA that no large regions were unassembled.

All genomic reads were mapped to the final Velvet assembly including the completed PmedOMT5 mtDNA to ascertain the read coverage for each scaffold and confirm to the mtDNA assembly. Transcriptome reads (details outlined in section 5.1.3) showed that all predicted protein-coding genes were expressed.

Table 4.6: Number of reads aligned to the P. nodorum SN15 mtDNA and PmedOMT5 DNA.

<table>
<thead>
<tr>
<th>Reads</th>
<th>Reference</th>
<th>Number of reads mapped</th>
<th>% of reads mapped</th>
<th>Mapping program</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unassembled genomic reads</td>
<td>SN15 mtDNA</td>
<td>5,102</td>
<td>0.02</td>
<td>Bowtie2</td>
</tr>
<tr>
<td>All genomic reads</td>
<td>SN15 mtDNA</td>
<td>81,223</td>
<td>0.36</td>
<td>Bowtie2</td>
</tr>
<tr>
<td>All genomic reads</td>
<td>PmedOMT5 mtDNA</td>
<td>162,869</td>
<td>0.73</td>
<td>Bowtie2</td>
</tr>
<tr>
<td>All in vitro transcriptome reads</td>
<td>PmedOMT5 mtDNA</td>
<td>354,007</td>
<td>0.65</td>
<td>TopHat2</td>
</tr>
</tbody>
</table>
Figure 4.2: Representative figure of the complete mitochondrial genome of PmedOMT5 and comparison to that of *P. nodorum* SN15.
4.4 Discussion

4.4.1 mtDNA size and structure

The predicted PmedOMT5 mitochondrial scaffold of 38.5 kb is smaller than the mtDNA of other sequenced Dothideomycetes yet contains a similar number and type of genes (summarised in Table 4.8). As a contiguous scaffold was able to be assembled and no telomeric sequences were identified it is clear the mtDNA has a circular structure. There was no evidence suggesting PmedOMT5 contains mitochondrial plasmids. The mtDNA has an average G:C content of 30.5%, which is consistent with that of most filamentous fungal mitochondria (Torriani et al. 2008) and is lower than the nuclear scaffold average of 53%.

The PmedOMT5 mitochondrial genome has a similar structure to Aspergillus nidulans (33 kb), both are smaller than the other Ascomycete fungi due to shorter intergenic spaces and fewer introns whilst possessing a similar gene set (Brown et al. 1985). As can be seen from Table 4.7 the two most closely related sequenced species have smaller mtDNA sizes than those of the more distant Sordariomycetes, with the exception of the expanded L. maculans mtDNA which has been invaded by intronic endonucleases (J.Hane- pers. comm.) which are also seen in the Sordariomycetes. The L. maculans mtDNA also contains four additional linear plasmids (Rouxel et al. 2011).

Mapping of the unassembled reads to the P.nodorum SN15 mtDNA suggests that there is not a large section of the PmedOMT5 mtDNA that was not assembled and the ability to amplify across the known gaps to produce fragments of small sizes confirms the mitochondrial genome is considerably smaller than that of other Dothideomycetes.
4.4.2 Mitochondrial gene content and translation

Fungal mitochondrial genomes usually contain the genes which encode the hydrophobic subunits of respiratory chain complexes, as well as genes for the organellar translation system: large and small ribosomal RNAs and a full set of tRNAs, sufficient to read all codons based on the extended wobble hypothesis (Bonitz et al. 1980, Gray et al. 1999). The standard repertoire encodes apocytochrome b; cytochrome oxidase subunits 1, 2, and 3; NADH dehydrogenase subunits 1, 2, 3, 4, 4L, 5 and 6; and ATPase subunits 6, 8 and 9 (http://pages.slu.edu/faculty/kennellj/table_mt_genomes.html accessed 4.3.2011).

One of the distinguishing features between mtDNA of fungi of the order Capnodiales and order Pleosporales, to which Pm-OMT5 belongs, is the absence of the genes for *atp8* and *atp9*. The gene for *atp9* is encoded by both the nuclear and mitochondrial genomes in other Ascomycetes, *Neurospora crassa* and *Aspergillus nidulans* (van den Boogaart et al. 1982, Brown et al. 1985, Kennell and Cohen 2004) but is absent from *Podospora anserina* and the Pleosporales *P. nodorum* SN15 and PmedOMT5. The gene for *atp8* is also absent from the mtDNA of *P. nodorum* and PmedOMT5 (order Pleosporales) but not *M. graminicola* (order Capnodiales) (Torriani et al. 2008). All other commonly found genes that were identified in the two other published Dothideomycete mitochondrial genomes were identified (Table 4.8) in PmedOMT5, although the *nad1* gene contains no stop codon and is missing the last ~10 aa as it extends into the gap between scaffolds.

### Table 4.7: Comparison of *P. medicaginis* mitochondrial scaffold with published Ascomycete mitochondrial genomes.

<table>
<thead>
<tr>
<th>Strain/isolate</th>
<th><em>Pm</em></th>
<th><em>Pn</em></th>
<th><em>Mg</em></th>
<th><em>Lm</em></th>
<th><em>Sm</em></th>
<th><em>Nc</em></th>
<th><em>Pu</em></th>
<th><em>An</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Order</td>
<td>Dothid</td>
<td>Dothid</td>
<td>Dothid</td>
<td>Dothid</td>
<td>Sord</td>
<td>Sord</td>
<td>Sord</td>
<td>Euro</td>
</tr>
<tr>
<td>Size (kb)</td>
<td>38.5</td>
<td>49.8</td>
<td>43.9</td>
<td>154.8</td>
<td>88.4</td>
<td>64.8</td>
<td>94.2</td>
<td>33</td>
</tr>
<tr>
<td>G:C content</td>
<td>30.5</td>
<td>29.4</td>
<td>32</td>
<td>30</td>
<td>33.6</td>
<td>36.1</td>
<td>29.9</td>
<td>-</td>
</tr>
<tr>
<td>Reference</td>
<td>Hane et al., 2007</td>
<td>Torriani et al., 2008</td>
<td>Rouxel et al., 2010</td>
<td>Nowrousian et al., 2010</td>
<td>Griffiths et al., 1995</td>
<td>Cummings et al., 1990</td>
<td>Brown et al., 1985</td>
<td></td>
</tr>
</tbody>
</table>

The full complement of twenty essential amino acid transporter RNAs and redundant copies of Arg, Leu, Met and Ser are present in the PmedOMT5 mtDNA. None of the predicted tRNA genes contain introns. Several of the duplicated tRNAs in these species were not seen e.g. unlike the closely related *P. nodorum*, PmedOMT5 has only two detectable methionine tRNAs rather than three. The order of the genes in the tRNA cluster flanking the large ribosomal subunit was similar to that in the Pleosporales SN15, with the exception of the third methionine (Table 4.9).

A third-position codon bias of A and T is commonly observed among fungal mitochondrial genes (Gray and Lang 1998) and there are many species that deviate from the universal code (Bonitz *et al.* 1980, Knight *et al.* 2001). Most Ascomycete and all analysed Dothideomycete mitochondrial genomes sequenced to date utilise genetic code 4 (The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code - http://www.ncbi.nlm.nih.gov/Taxonomy/Utils/wprintgc.cgi) in which the UGA codon, that encodes ‘stop’ in the standard amino acid translation table 1, is translated as tryptophan (Paquin *et al.* 1997, Hane *et al.* 2011b). The PmedOMT5 mitochondrial genome appears to share this codon usage.

Fungal mitochondrial genomes may contain intron-encoded ORFs as well as varying numbers of unique ORFs of unknown function and several also encode a ribosomal protein associated with the small rRNA (Kennell and Cohen 2004). The ribosomal protein (rps5) is observed in PmedOMT5 as in SN15. There was only one PmedOMT5 ORF of significant size (210 codons) which bears no similarity to other known mitochondrial genes (Tables 4.5 and 4.8, Figure 4.2). Transcriptome read alignments show that this ORF is expressed as are all other protein-coding genes.
Table 4.8: Comparison of annotated fungal mitochondrial gene content with that of the final assembled PmedOMT5 mitochondrial genome.

<table>
<thead>
<tr>
<th></th>
<th>P. medicaginis</th>
<th>P. nodorum</th>
<th>M. graminicola</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>tRNAs</strong></td>
<td>24</td>
<td>27</td>
<td>27</td>
</tr>
<tr>
<td><strong>rRNAs</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>large ribosomal RNA subunit (ml)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>small ribosomal RNA subunit (rns)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td><strong>Mitochondrial protein-encoding genes</strong></td>
<td>12</td>
<td>12</td>
<td>14</td>
</tr>
<tr>
<td>ATP synthase subunits</td>
<td>1</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>atp6</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>atp8</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>atp9</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>cytochrome oxidase subunits</td>
<td>3*</td>
<td>3*</td>
<td>3</td>
</tr>
<tr>
<td>cox1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>cox2</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>cox3</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>cytochrome b (cytb)</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>nicotinamide adenine dinucleotide ubiquinone oxidoreductase subunits</td>
<td>7</td>
<td>7</td>
<td>7</td>
</tr>
<tr>
<td>nad1</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>nad2</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>nad3</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>nad4</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>nad4L</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>nad5</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>nad6</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>5S ribosomal protein (rps5)</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td><strong>Unknown ORFs</strong></td>
<td>1</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td><strong>Intronic endonucleases</strong></td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
</tbody>
</table>

*CoxI and II are joined in P. nodorum SN15 and most likely also in PmedOMT5. Adapted from (Hane et al. 2011b) with data from (Hane et al. 2007, Torriani et al. 2008).
4.4.3 Mitochondrial introns, intronic endonucleases and rearrangement of gene order

Ascomycete mitochondria often contain large numbers of introns (Paquin et al. 1997), the highest number was observed in *P. anserina* where 36 introns occupy ~ 60% of the mtDNA (Cummings et al. 1990). There are two major groups of mitochondrial introns, I and II, classified according to conserved secondary structure and the occurrence of short sequence motifs (Michel et al. 1982), reviewed by Lang in 2007. Most fungal mitochondrial introns belong to group I, e.g. 33 out of 36 in *P. anserina* (Cummings et al. 1990, Paquin et al. 1997). Many group I introns contain a homing (DNA) endonuclease gene that acts in the transfer and site-specific integration or “homing” of introns. Group II introns typically contain ORFs that code for reverse-transcriptase-like proteins (Hausner 2003). Both groups may play a role in genomic rearrangements (Dujon 1989) and transfer via HGT between species and possibly even phyla (Vaughn et al. 1995). Introns can be inserted into many different mitochondrial genes, with a strong preference for sequences from the most conserved protein-coding genes, *cox1* and *cytb*. No introns have been found in fungal mitochondrial tRNA genes and relatively few are located in rRNA genes. Genes such as *nad6*, *cox2* or *atp9* rarely contain introns (Paquin et al. 1997). Mitochondrial introns are often located in identical gene positions in different fungal species (Paquin et al. 1997).

The majority of group I ORFs fall into two families of homing endonucleases (Galburt and Stoddard 2002, Lang et al. 2007) encoding either LAGLIDADG or GIY-YIG as their conserved sequence motifs (Belfort and Perlman 1995). Only two PmedOMT5 genes (*cox3* and *cox1-2*) contain introns, compared to six in *P. nodorum* mtDNA and none in *M. graminicola*. The largest intron seen in PmedOMT5 belongs to group I, with a conserved LAGLIDAG motif [pfam: LAGLIDAG_1 PF00961] which appears in the *cox1* gene. Its equivalent in SN15 resides instead within the *cytb* gene intron, indicating it has moved since the species diverged or that the species were both invaded following divergence. The genes for *cox1* and *cox2* also appear to belong to a single transcriptional unit in OMT5 as in SN15, as the *cox1* gene contains no stop codon.

Extensive rearrangements of mtDNA are often seen in the Fungi, although some genes commonly remain linked. One example of this is the tRNA clusters surrounding the large ribosomal subunit gene (Clarke-Walker 1992) which are also retained in
PmedOMT5 (Table 4.9). The gene locations and/or orientations of *nad1*, *nad4*, *cytb*, *rps5*, *atp6* and the LAGLIDADG endonuclease were altered between PmedOMT5 and its closest sequenced relative SN15.

**Table 4.9: Order of tRNAs in cluster flanking the large ribosomal subunit (rnl) in sequenced Dothideomycete mtDNA (Adapted from Hane et al, 2011).**

<table>
<thead>
<tr>
<th>Species/Isolate</th>
<th>Order</th>
<th>5’ Upstream of rnl a,b</th>
<th>3’ Downstream of rnl b</th>
<th>GenBank accession number</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>P. medicaginis var. medicaginis</em> OMT5</td>
<td>Pleosporales</td>
<td>VKGDS₁WIRS₂P</td>
<td>TM₁EAFL₂QHM₂</td>
<td>N/A</td>
</tr>
<tr>
<td><em>Phaeosphaeria nodorum</em> SN15</td>
<td>Pleosporales</td>
<td>VKGDS₁WIRS₂P</td>
<td>TM₁M₂EAFLQHM₃</td>
<td>EU053989</td>
</tr>
<tr>
<td><em>Mycosphaerella graminicola</em> IPO323</td>
<td>Capnoidiales</td>
<td>GDS₁WIS₂P</td>
<td>M₁L₁EAFL₂YQM₂HRM₃</td>
<td>EU090238</td>
</tr>
</tbody>
</table>

- The numbers 1, 2, 3 indicate the presence of more tRNA genes for the same amino acid in the consensus sequences.
- Capital letters refer to tRNA genes with anti-codon for: R, arginine; K, lysine; G, glycine; D, aspartic acid; S, serine; W, tryptophan; I, isoleucine; P, proline; T, threonine; E, glutamic acid; V, valine; L, leucine; A, alanine; F, phenylalanine; Q, glutamine; H, histidine; Y, tyrosine; N, asparagine.

### 4.4.4 QoI fungicide resistance site

Following the introduction of strobilurin (QoI) fungicides in 1996, many fungal crop pathogens have independently acquired resistance or insensitivity to this group of fungicides. The QoI (Quinone outside inhibitor) fungicides act by blocking electron transport in the membrane-embedded cytochrome bc₁ complex and thus inhibiting respiration. The primary molecular mechanism of resistance in the majority of fungi is a point mutation at codon 143 in the mitochondrial encoded cytochrome b gene (*CYTB*), which results in the substitution of glycine (G) with alanine (A) (G143A) (Gisi *et al.* 2000). A second mutation F129L has been observed in *A. solani* (Pasche *et al.* 2005) which confers a more moderate resistance.

None of the known mutations that confer fungicide resistance were observed in the PmedOMT5 *CYTB* gene. The G143A mutation is not known to have occurred in fungi that contain an intron immediately following the highly conserved glycine 143 (Grasso *et al.* 2006). Alteration of this codon hinders splicing of the intron bi2 (Vallieres *et al.* 2006).
2011), thus interrupting the CYTB open-reading frame. As the PmedOMT5 CYTB gene contains no intron it has potential to develop QoI resistance in the field.

4.5 Summary
Using a combination of strategies the assembly of the complete PmedOMT5 mitochondrial genome sequence was completed and its gene content analysed. The gene content was found to be similar to other Dothideomycete fungal mitochondria however a lower presence of introns and endonucleases than the currently sequenced Dothideomycetes was observed, as were several rearrangements. The PmedOMT5 mitochondrial sequence provides a valuable resource for phylogenetic and fungicide resistance studies in Phoma and related species.
Chapter 5
Biological support for predicted gene models
5.1 Introduction

An accurate and comprehensive overview of the gene content of an organism is essential for all downstream analysis and requires the combination of multiple and independent layers of evidence. Inaccurate annotation of gene structure can complicate analyses including: orthology, alignment of conserved domains and detection of gene presence, making subsequent biological analysis such as knockouts and cloning much more difficult to perform and interpret accurately.

A critical factor in comparative genomics is the ability to identify conserved domains with known functions. The conservation of these domains in a newly sequenced genome allows the transfer of functional annotations from experimentally validated genes to its predicted gene models. This forms the basis for subsequent comparative genomics analyses using Pfam and GO functional annotations, which will be discussed further in chapter 6. If gene structure is not correctly annotated, important functional annotations can be missed or duplicated. The major component of comparative genomics is comparisons between species of the numbers of genes with specific functional annotations, thus inaccuracies stemming from mis-annotation can lead to misguided interpretations of the data in a biological context.

A variety of methods can be employed to gauge the accuracy of gene calling including: transcriptomics (Li et al. 2011), proteogenomics (Dybas et al. 2008, Bringans et al. 2009) and the examination of conserved genes with regard to phylogeny or the co-localization of functionally linked genes (Korbel et al. 2004).

5.1.1 Gene prediction

Most currently available ab initio gene prediction software relies on hidden Markov models (HMM) which are effective for the locus-level prediction of highly conserved, or “core” genes. However these methods have lower accuracy for genes which differ from the norm in terms of G:C content, size and distribution of exon/intron borders or even potentially codon usage if their presence in the genome is the result of recent lateral gene transfer (Richards et al. 2011). These non-standard genes can be improperly predicted or in some cases missed altogether. HMM gene prediction algorithms look for patterns that are known to be conserved in eukaryotic genes such as G:C content variation, start/stop codons and donor and acceptor splice sites. Some extrinsic gene
finding algorithms can be trained using EST data or gene annotations from another species, however even closely related organisms can differ significantly with regard to these characteristics (Korf 2004). In the majority of cases, *ab initio* gene predictors correctly identify a gene locus (Coghlan *et al.* 2008, Yandell and Ence 2012), although they are less accurate at determining the complete gene structure and can differ in the correct number of exons called, start codons and splice positions. Analysis of gene prediction in rice has shown that most gene prediction software, including GeneMark, is only 70-80% accurate in terms of exon prediction accuracy (Haas *et al.* 2008). Most software are more accurate at predicting internal exon boundaries (~85%) with lower (44-68%) accuracy of prediction of initial or terminal exons (Haas *et al.* 2008).

Because host-specific effectors often only exist in the pathogenic species with little or no homology to other genes they will not always be well predicted by *ab initio* prediction software. Short and/or single exon genes are predicted poorly by most gene finding algorithms (Goodswen *et al.* 2012). Example of this are the characterised pathogenicity genes *ToxA* and *Tox3* of *S. nodorum*, for which the *ab initio* gene finder GeneMark produced poor predictions missing the terminal exon in *ToxA* and completely missing *Tox3* (Pers. Comm.- K-C. Tan, R.P. Oliver, R. Syme).

### 5.1.2 Proteogenomics

Proteogenomics is a technique used to match sequenced peptide fragments derived from proteins produced by the organism, to a reference genome or transcriptome assembly in order to identify proteins translated under the conditions examined. In recent years, proteogenomic analysis has been used to validate and correct the annotation of the fungal genome assemblies of *Phaeosphaeria nodorum* (Bringans *et al.* 2009) and *Blumeria graminis* (Spanu *et al.* 2010). Unlike transcriptomics which can capture a wide range of gene expression levels in one sweep, proteomics is much more sensitive to the large variation in protein abundance within a cell (Brunner *et al.* 2007) and can have lower detection rates for low molecular weight proteins (Armengaud 2009). However it can provide a much more targeted analysis, depending on the protein isolation techniques used, to provide a snapshot of the proteins present in the cell at specific locations or translated under different conditions, although for genome-wide gene validation several growth conditions should be used. The protein content of a cell
is not always well represented by the transcripts present, as proteins have different rates of degradation and may exist in a cell long after the transcript that encoded them has been degraded. Many authors have shown there is not a direct correlation between transcription and translation, reviewed by Greenbaum et al., (2003). Proteomic analysis allows researchers to investigate which proteins in the cell may be involved in specific molecular reactions such as the interaction between powdery mildew haustoria and their host plant cell (Godfrey et al. 2009). Another advantage of proteomics is that it allows selection of proteins with particular physical properties (e.g. hydrophobic or positively/negatively charged) or which localise to particular regions of the cell, which is not possible with transcriptomics e.g. selection of secreted (extra-cellular) proteins which may offer insights into which proteins the pathogen is secreting into the plant cell and thus facilitate identification of effector molecules.

Proteogenomic validation of predicted gene models begins with chromatographic separation of trypsin digested proteins followed by analysis with a mass spectrometer. The sequence of the peptide fragments is predicted via analysis of MS/MS fragmentation patterns. This peptide sequence is then mapped to a database of predicted proteins and to a six-frame translation of the genome assembly (both of which have been bioinformatically trypsin-digested in silico) to determine if the peptide data supports the existing gene models or predicts that the gene or specific CDS features should be annotated in a different reading frame.

5.1.3 RNA-seq

RNA-seq provides the most broadly informative support for whole gene annotation, defining exon boundaries, untranslated regions (UTRs), alternate exon splicing and can aid in the detection of previously unidentified genes that were not predicted in silico. RNA-seq can also provide additional data for correction of small indels in the original genome assembly (Uyar et al. 2012) but the potential for post-transcriptional mRNA modifications must also be taken in to account. RNA-seq provides a greater depth of data than expressed sequence tags (ESTs) (Wang et al. 2009) for considerably less cost although analysis of these data can be challenging due to size and complexity (Yandell and Ence 2012).

There are two main methods of RNA-seq data analysis: mapping sequenced reads to a genome assembly then assembling transcripts from the mapped reads (e.g. using the
The Tuxedo suite of tools (Trapnell et al. 2009, Trapnell et al. 2010) or creating a de novo assembly of the transcriptome directly from the RNA data utilising tools such as Trinity (Grabherr et al. 2011) or Oases (Schulz et al. 2012). An essential requirement for the first is the existence or availability of a relatively complete reference genome that will allow for alignment of the majority of mRNA-derived reads. The second technique is often used by researchers working with non-model organisms as the de novo transcriptome assembly can be used to investigate the gene content in the absence of a reference genome. Assembly of genes in a de novo transcriptome assembly is less complex than de novo genome assembly due to the lower amount of interspersed repetitive DNA. Thus the genes may be more accurately and completely assembled than those from transcript reads mapped to an Illumina sequencing based assembly of a highly repetitive genome that can be highly fragmented (Grabherr et al. 2011, Yandell and Ence 2012). Additionally this technique can also be used to identify expressed regions that were not assembled well in the reference genome sequence.

5.1.4 Annotation pipeline

A more holistic approach for gene annotation is the use of automated pipelines to combine different layers of evidence. To facilitate gene discovery and accurate annotation many automated pipelines (such as MAKER2 (Cantarel et al. 2008), or PASA/EVidenceModeler (Haas et al. 2003, Haas et al. 2008)) build a complete picture of gene structure by combining homology, proteomic and/or transcriptomic data where available. Utilizing transcript data has the advantage of helping to discover the gene UTR borders as opposed to simply coding exons (CDS), whilst proteomics can provide information as to which isoforms are expressed in the analysed sample and which transcripts are leading to translation. This is especially useful during this era of high-throughput, short-read sequencing, as short RNA-seq reads may potentially represent full-length mRNAs or degraded, non-functional fragments. Furthermore, while the use of genome homology to proteins from other species greatly improves annotation accuracy compared to in silico predictions, homology data should be treated with care. A newly annotated gene’s functional predictions depend upon how well related gene models in the other species have been annotated. There are many mis-annotated models in public databases, although less are found in manually curated data sets (Schnoes et al. 2009). Manual annotation - which involves examination by a researcher of all
available data for a gene’s structure in a viewer and if necessary, curation in an annotation editor - is still considered ‘the gold standard’ approach for producing high quality and accurate genome annotations (Haas et al. 2008).

5.1.5 Aims for this chapter

In this chapter the aim was to create the most accurately annotated gene models for PmedOMT5 based on the currently available evidence: including homology, RNA-seq of transcripts at several life-cycle phases of the fungus and intra- and extra-cellular peptide sequences whilst excluding known non-coding and repetitive regions of the DNA. The work described in this chapter entails the biological experiments conducted to provide evidence for the genome annotation and the strategies used to validate the predicted gene set.

5.2 Materials and methods

5.2.1 Proteogenomic analysis

5.2.2 Intracellular protein extraction

PmedOMT5 spores (1 ml at 1x10^6 sp/mL) were incubated in 500 mL flasks containing 100 mL minimal media pH 6.0 with 25 mM glucose as a carbon source for 7 d at 20 °C, 150 rpm. Mycelia were harvested by straining through three layers of sterile milk filters and rinsed with sterile water. Washed mycelia were lyophilised for 48 h then ground in 10 mM Tris-Cl pH 7.5 with glass beads and centrifuged at 16,000 g for 15 min at 4 °C. The supernatant was carefully removed with an 18 gauge needle. This was repeated with the remaining sample and the supernatants combined. The extracted proteins were treated with 10 U DNase (Sigma, Castle Hill, Australia) and 10 U RNase (Sigma, St. Louis, USA) to remove contaminating nucleic acids. Protein concentration was determined via the Cu^2+ reduction/bicinchoninic acid (BCA) method (Smith et al. 1985) (Sigma, St. Louis, USA). All protein samples were checked via SDS-PAGE (12 % glycine gel) and stained with Coomassie blue (Sambrook et al. 1989) to ensure that proteolysis during sample preparation was minimal (Figure 5.1). Following extraction, samples were aliquoted and stored at -80 °C until analysis.
5.2.3 Culture filtrate/ extracellular protein extraction

Extracellular samples were cultured under conditions previously determined to produce a culture filtrate that caused necrosis on the host species *M. truncatula* as described in chapter 7. Fries 2 media without yeast extract (100mL) was inoculated with 1 mL 1x10^6 sp/mL and cultured for 3 d at 27°C at 100 rpm, followed by 33 d static growth in the dark. Extracellular proteins were harvested by filtration (0.2µm) to remove mycelia and spores. In order to concentrate proteins and remove salts etc. the buffer was exchanged using Millipore 3 kDa molecular weight cut-off (MWCO) filters and proteins were further purified via PD-10 columns with elution in Fluka MS grade water (Sigma, St.Louis, USA).

Activity of culture filtrate (CF) was checked *in planta* as described in chapter 7 (illustrated in Figure 5.1b) and mannitol dehydrogenase activity was tested to check for contamination with intracellular fraction.

5.2.4 Mannitol dehydrogenase (MDH) assay, to check activity in intracellular and extracellular fractions

NADPH-dependent fructose reduction was measured by proxy, via oxidation of NADPH as detected by a change in absorbance at 340 nm over an incubation period of 30 min (Joosten *et al*. 1990, Noeldner *et al*. 1994).

333 µL of 2.4 M fructose solution was incubated with 10 µL of 25 mM NADPH, 107 µL of sterile water and 50 uL of cell free extract and absorbance measured using a spectrophotometer.

5.2.5 MS-MS sequencing of intra- and extra-cellular fractions

Protein samples were digested with trypsin for comparison to an in silico trypsin digest of the predicted proteins. Protein separation and peptide mass fingerprinting were carried out essentially as described in Bringans *et al*. (2009) by Proteomics International, Perth, Australia.

Briefly, 200 µg protein sample was reduced, alkylated and trypsin digested according to the iTRAQ protocol (Applied Biosystems). Peptides were desalted on a Strata-X 33 µm polymeric reversed phase column (Phenomenex) and dissolved in 70 µl of buffer (2% acetonitrile and 0.05% trifluoroacetic acid) before separation by strong cation exchange
liquid chromatography on an Agilent 1100 HPLC system using a PolySulfoethyl column (4.6 x 100 mm, 5 μm, 300 Å). Peptides were eluted with a linear gradient of 0-400 mM KCl. Strong-cation exchange fractions were desalted and loaded onto an Ultimate 3000 nano-HPLC system (Dionex, C18 PepMap100, 3 μm) and separated using a gradient of 10-40% acetonitrile (0.1% trifluoroacetic acid). Separated peptides were spotted using a ProBot (LC Packings) robotic spotter and analysed via MALDI-TOF-TOF. Spectra were analysed to identify proteins of interest using Mascot sequence matching software version 2.2.04 (Matrix Science) within databases of in silico digests of the 6-frame translation of the open reading frames of the PmedOMT5 assembly and the PmedOMT5 version 1 predicted proteins. Protein hits were considered based on MOWSE (MudPIT) scoring (which removes proteins that have high protein scores purely because they have a large number of low-scoring peptide matches) and only from rank 1 peptide matches to the database above the homology threshold with a significance value of $p < 0.05$. Search parameters included a peptide and fragment mass tolerance of ± 4kDa allowing for the modification of methionine by oxidation.

5.2.6 Assessment of support for predicted genes via mapping of peptide fragments to the annotated gene models and 6-frame translation of PmedOMT5.

Peptide data from MASCOT analysis was filtered to identify only unique peptides over the identity threshold in comparison to the automatically-generated decoy database. Six-frame ORF translations between stop codons of the scaffolds were obtained as described in section 3.2.4. The unique peptides were then mapped to the PmedOMT5 scaffolds using the CDSmapper suite (available from https://sourceforge.net/projects/cdsmapper/). These scripts enable the mapping of peptide matches to their genomic locations and comparison with a previously annotated gene set. Briefly, CDSmapper.pl creates a GFF3 file detailing the genomic location of peptide matches to the proteins or ORFs. The script testpeptideframes.pl tests whether or not the peptide sequences overlap with the previously predicted genes based on GFF co-ordinates and summarise_frametest.pl summarises the results of testpeptideframes.pl. Compare2gff.pl compares the peptide GFF file to the predicted protein GFF so peptides can be matched back to the predicted proteins and scaffolds.

Where peptide alignments to the scaffolds conflicted with the currently annotated gene set, each match was assessed individually. Initial analysis was by BLASTP comparison
of the predicted gene model and the alternate model suggested by the peptide. This was followed by TBLASTN comparisons with the predicted proteomes of 10 other Ascomycete fungi (Aspergillus nidulans, Pyrenophora teres f. teres, Pyrenophora tritici-repentis, Leptosphaeria maculans, Zymoseptoria tritici (formerly Mycosphaerella graminicola), Magnaporthe oryzae (formerly grisea), Ascochyta rabiei, Peyronellaea pinodes, Phaeosphaeria nodorum and Neurospora crassa), details of data sets in appendix 5.1. Finally, the curated peptide data was manually assessed in conjunction with other supporting data to determine whether genes required re-annotation.

5.2.7 Transcriptome analysis

5.2.8 In vitro fungal tissue used for RNA production

PmedOMT5 was grown in 100 mL Fries 2 media without yeast as described in chapter 7.2.1. Culture filtrate from the media was infiltrated into M. truncatula plants and produced typical symptoms including chlorosis as shown in Figure 5.1b. Fungal mycelial tissue was harvested by filtration from three separate flasks, snap frozen and stored at -80 °C. Three biological replicates were harvested and extracted separately.

Droplets of P. medicaginis var. medicaginis isolate PmedOMT5 (5 x 50 µL at a concentration of 1x10⁶ spores/mL from glycerol stocks) were spotted onto minimal media plates (25mM glucose as a carbon source). Mycelia were harvested at 4 d (vegetative mycelial growth only) and 16 d (Mycelia and pycnidia containing pycnidiospores) by scraping the plates with a sterile spatula and tissue stored at -80 °C until RNA extraction. Three biological replicates were harvested and extracted separately.

5.2.9 Preparation of infected plant material for RNA-seq

M. truncatula plants (cultivar A17) were grown in a growth chamber with 12 h light at 20 ±2 °C. Four week old plants were spot inoculated on the monofoil and central leaflet of all mature leaves with 10 µL droplets containing 2 x 10⁶ PmedOMT5 spores/mL with 0.05% Tween20 or only 0.05% Tween20 for mock inoculated plants. There were 15 plants per treatment block of a split plot design. Throughout the experiment leaves were examined via microscopy using various staining techniques (Figure 5.3), described in
chapter 2.3.2. The area of leaf tissue under the droplet was excised using sterile scissors at 1, 2, 3, 4, 5, 7 and 10 dpi from 20 individual infected leaflets per experimental unit and immediately frozen in liquid nitrogen. Three pooled replicates were harvested at each time-point.

5.2.10 RNA isolation

RNA was extracted from infected plants at 1-5 dpi using Trizol (Invitrogen Corp., Carlsbad, CA) essentially according to the manufacturer’s instructions with a high salt precipitation step added for in planta samples.

Briefly, approximately 150 mg of fresh tissue was ground in liquid nitrogen and combined with 1 mL of Trizol reagent (Invitrogen Corp., Carlsbad CA) in a 1.5 mL eppendorf tube then incubated at RT for 15 min, before storing on ice. Samples were centrifuged at 12,000 g, 4 °C for 10 min. The supernatant was removed to a new tube and combined with 200 µL chloroform. Samples were then incubated at RT for approximately 3 mins until separation of the phases began and centrifuged at 12,000 g, 4 °C for 15 min. The upper phase was combined with 300 µL high-salt precipitation solution (0.8 M sodium citrate/1.2 M NaCl) and 300 µL isopropanol before incubating at RT for 10 min to selectively precipitate total RNA. Samples were then centrifuged for 10 min at 12,000 g, 4°C for 10 min. Precipitated RNA was washed twice in 75% ethanol and resuspended in diethylpyrocarbonate (DEPC)-treated water. To facilitate re-suspension of the RNA pellet, samples were heated to 60 °C for 10 min.

The three biological replicates were extracted separately and 1 µL was run on a 1% agarose gel to analyse the integrity of the RNA. RNA was also analysed spectrophotometrically to ascertain the quantity and purity before proceeding. RNA was diluted to 200 ng/µL and treated with DNase 1 (Ambion DNA-Free, Applied Biosytems, CA, USA) according to the manufacturer’s instructions, before pooling for sequencing. Total RNA (400 ng from each time-point and replicate) was pooled for samples from 1-3 dpi, with 200 ng from each replicate of the 4 and 5 dpi samples, producing a total of 4,800 ng of RNA from which the sequencing library was generated.

For the fungal (non- infection) samples, RNA was extracted separately from each of three biological replicates using the Trizol protocol exactly as described by the manufacturer (Invitrogen Corp., Carlsbad, CA). Samples were DNase treated (Ambion,
Applied Biosystems, CA, USA) pooled and concentrated/cleaned using Qiagen RNAeasy columns as per the manufacturer’s instructions (QIAGEN GmbH, Germany).

Before submitting for sequencing all RNA was tested via PCR to ensure there was no contamination with DNA (Primers HMG-236R_AW and HMG-8L_AW, amplification conditions described in appendix 5.2). One microlitre of each pooled RNA sample was analysed using a Nanodrop and a further 1 µL of the pooled sample was run on a 1% agarose gel to visually analyse the RNA quantity and integrity.

Upon receipt at The Australian Genome Research Facility (AGRF) total RNA was analysed using a BioAnalyzer and all samples were found to have a RIN >8 and 28S:18S ratio ranging from 1.3-1.8.

5.2.11 Sequencing

RNA-library preparation and sequencing were conducted at AGRF, Melbourne, Australia using an Illumina GAIIx. Four libraries were sequenced: three in vitro (4d, 16d and F2NY) and one from the combined in planta libraries (1-5 dpi).

The in vitro samples were bar-coded, pooled and sequenced over two lanes. The pooled in planta samples were run in two full lanes. All sequencing was run for 100 paired-end cycles. Sequence data files were generated via the Illumina 1.9.0 pipeline.

5.2.12 Sequence analysis and trimming

Read quality was inspected with SolexaQA (Cox et al. 2010) and FastQC v9.2 (Figure 5.5) (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Reads from each lane were trimmed of adapters and low quality bases using scripts written by the USC Bioinformatics group and publically available on github (https://github.com/vsbuffalo/scythe and https://github.com/najoshi/sickle). Scythe (vsbuffalo-scythe-8472bb2C) was used to trim Illumina adapter sequence from 5’ end of reads and Sickle (najoshi-sickle-79cb240) was used to remove low quality bases at the 3’ end. Only bases with a Phred quality score ≥ 30 (i.e. a 99.9% probability that the base was correctly assigned) were retained. Following trimming, read sets were filtered to retain only paired reads ≥ 25 bp in length. Single reads whose pair had been discarded based on filtering results were retained for downstream analysis. Post processing, reads from the same sample (sequenced in different lanes) were combined for analysis.
5.2.13 Transcript read alignment

To determine splice-junction sites and read coverage depth, RNA-seq libraries were aligned individually to the genome assembly via TopHat 2.0 beta with parameters: minimum intron size 20 bp, maximum intron 4,000 bp, coverage search, microexon search enabled, very sensitive mapping, 20 read matches maximum, reporting discordant and secondary alignments (Trapnell et al. 2009). Intron boundaries were determined based on analysis of intron sizes in five fungi by Kupfer and colleagues (Kupfer et al. 2004).

Cufflinks 1.1.0 (Trapnell et al. 2010) was used to create transcripts from the in vitro reads aligned to the genome and the GeneMark predicted v1 gene set. This analysis was later used as an input for EVidence Modeler (EVM), described in section 5.2.17.

5.2.14 De novo assembly of transcripts

Trinity (Grabherr et al. 2011) release 2011-11-26, was used for de novo transcriptome assembly of each individual library using the following parameters: jaccard clipping, minimum contig length 99 bp, min kmer coverage 5. A combined in vitro assembly, incorporating reads from the 4 d, 16 d and Fries media libraries, was also assembled to increase the chances of assembling complete transcripts of genes with low expression. This resulted in five assembled transcript libraries: 4d, 16d, F2NY (35d), the combined in vitro library and the in planta library.

All assembled transcripts were aligned to the PmedOMT5 genome assembly using BLAT (Kent 2002) and the non-redundant NCBI protein database using BLASTX. Those that did not align to the PmedOMT5 assembly were analysed via the CENSOR database (Kohany et al. 2006).

In planta derived transcripts were also aligned to the M. truncatula A17 reference genome assembly version 3.5 (available at www.medicagohapmap.org/?genome) (Young et al. 2011) for the purpose of distinguishing plant sequences from fungal sequences. Transcripts were selected for subsequent analysis if they aligned exclusively to the PmedOMT5 scaffolds. Those that aligned to both plant and fungal scaffolds but had a stronger match by bit score to the PmedOMT5 assembly were also selected. These two sets of transcripts made up the filtered in planta data set. The combined in vitro
transcripts and the filtered PmedOMT5 *in planta* transcripts were aligned to the genome using AAT (Huang *et al.* 1997).

5.2.15 Phoma ESTs

Expressed Sequence Tags (ESTs) isolated from excised *M. truncatula* young trifoliate leaves that were dip-inoculated in a spore suspension of an unspecified *P. medicaginis* isolate then harvested at 0, 15, and 30 minutes and 1, 2, 3, 6, 14, 24, 48, 72, and 96 h, were downloaded from NCBI [NCBI_EST:LIBEST_010527_Phoma-infected], (accession numbers BQ137938-BQ141532, (Gamas *et al.* 2007), Watson, Paiva *et al.*, unpublished). These ESTs were aligned to the PmedOMT5 genome and the *M. truncatula* A17 genome assembly v3.0 (Young *et al.* 2011) using BLAT (Kent 2002) with default settings. Sequences that matched PmedOMT5 with a higher bit score than the match of the same sequence to the *Medicago* genome assembly were selected and used as further evidence for annotation.

5.2.16 Combination of evidence for manual gene annotation

AAT (Huang *et al.* 1997) was used to align protein data from 14 fungal species (*P. nodorum*, *N. crassa*, *M. oryzae*, *Dothistroma septosporum*, *L. maculans*, *P. teres* f. *teres*, *F. oxysporum* f. sp. *lycoperscici*, *Aspergillus nidulans*, *Alternaria brassicicola*, *Zymoseptoria tritici*, *Ascochyta rabiei*, *Peyronellaea pinodes*, *Pyrenophora tritici-repentis*, *Cochliobolus heterostrophus*), as well as, all proteins from the Pathogen Host Interactions database (PHIbase version 3.2 (http://www.phi-base.org/) (Baldwin *et al.* 2006), and publically available EST data from three species (*P. medicaginis*, *Ascochyta rabiei* and *Colletotrichum trifolii*), to the PmedOMT5 genome. A complete list of species, isolates, strains and annotation version numbers related to the datasets used in this thesis are listed in appendix 5.1.

PASA (r2011-05-20) was used to determine the exon structure based on the alignment of *de novo* Trinity assembled transcripts to the genome (minimum percentage aligned 75, maximum intron length 5000 bp) (Haas *et al.* 2003, Haas *et al.* 2008). The PASA output was combined with the following supporting data (weighted as indicated in parentheses) as inputs for the automated annotation pipeline EVidenceModeler (EVM) r2012-06-25 (Haas *et al.* 2008): TopHat junctions (15), PASA predictions (10), GeneMark predictions (8), other fungal protein alignments (5), EST alignments (7),
Trinity alignments (8), Pmed EST alignments (8) and Cufflinks (4). The Apollo Genome Annotation Curation Tool v1.11.6 (Lewis et al. 2002) and IGV v2.0.7 (Thorvalsdóttir et al. 2013) were used to evaluate the EVidenceModeler predictions in combination with their supporting evidence via manual inspection and BLAST analysis to confirm exon structure. These data were used in conjunction with tRNA, ncRNA and repeat analysis to exclude probable non-coding regions from the protein-coding gene prediction.

All genes were manually annotated using Apollo according to the following criteria: GeneMark gene models were retained unless there was sufficient conflicting evidence such as expression data (PASA and TopHat alignments), peptide data or reasonable alignments from other fungal sequences. Annotation was amended if there was overlap with predicted repetitive DNA or non-coding RNA. Where there was conflicting evidence, gene annotations were adjusted to the most likely structure based on available evidence. Where multiple isoforms from a single locus were present the reference model was annotated as the most abundant model based on the RNA-seq data from the conditions analysed. Predicted proteins less than 50 aa were discarded if none of the supporting evidence described above was found to support their retention.

5.2.17 Re-alignment of reads to new gene models and calculation of expression differences

Following manual inspection and adjustment of the annotation, reads were re-aligned to the scaffolds (re-ordered by size) using TopHat 2.0.4 with the same parameters described in section 5.2.13. Differences in expression were calculated using CuffDiff (Cufflinks 2.0.2) (Trapnell et al. 2010, Trapnell et al. 2012) with the following parameters: frag-bias-correct, multi-read-correct, min-alignment-count 5, upper-quartile-norm and compatible-hits-norm. Transcripts with more than a four-fold change ($\log_2 \geq 2$) in FPKM were considered as differentially expressed between conditions.

In planta reads were also aligned to the A17 M. truncatula genome assembly v3.5 to determine the number of plant expressed genes, using TopHat 2.0.4 with the same parameters except that micro-exon and coverage search were disabled and the maximum intron size was set to 15 kb.
5.3 Results

Manual curation of the gene set predicted by EvidenceModeler resulted in 11,879 reliable gene models (designated “PmedOMT5 version 2 protein set”) that were supported by peptide, transcript and/or homology-based evidence. A summary of the version 2 gene set is presented in Table 5.1. A complete description of all version 2 gene models, their level of annotation support and functional annotation is provided in appendix 5.3. Gene models with functional annotations that are potentially relevant to pathogenicity are discussed in further detail in chapter 6.

Table 5.1: Summary of PmedOMT5 version 2 protein-coding gene annotations.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total length (aa)</td>
<td>5,451,839</td>
</tr>
<tr>
<td>Average length (aa)</td>
<td>459</td>
</tr>
<tr>
<td>Maximum length (aa)</td>
<td>9,607</td>
</tr>
<tr>
<td>Minimum length (aa)</td>
<td>50</td>
</tr>
<tr>
<td>N50</td>
<td>2,997</td>
</tr>
<tr>
<td>L50 (aa)</td>
<td>576</td>
</tr>
<tr>
<td>Total number of genes</td>
<td>11,879</td>
</tr>
<tr>
<td>Incomplete genes (lacking a start and/or stop codon)</td>
<td>230</td>
</tr>
<tr>
<td>Mean intron length (nucleotides)</td>
<td>53</td>
</tr>
<tr>
<td>Mean intergenic distance</td>
<td>1,052</td>
</tr>
</tbody>
</table>

5.3.1 ESTs

Of the 3,281 ESTs available from the Phoma-infected cotyledon library at GenBank, [NCBI_EST:LIBEST_010527_Phoma-infected] 93 (2.8%) mapped with a higher bit score to the PmedOMT5 assembly providing additional support for 92 genes. The majority of the remainder, 2,499 ESTs (76.2%) mapped uniquely to the M. truncatula A17 genome v3.0 (Cannon et al. 2006, Young et al. 2011) via BLAT at a significance threshold below 1x10^-5. These data were used for annotation and in planta expression analysis discussed in chapter 6.3.8.
5.3.2 Proteogenomic gene model support

In total 570 (4.8%) PmedOMT5 v2 proteins were supported by peptide data, 454 (3.8%), had 6-frame only support. This indicates that these genes or those with a similar domain have peptide support.

Table 5.2: Comparison of intra-cellular and extra-cellular protein extracts sent for sequencing.

<table>
<thead>
<tr>
<th>Sample type</th>
<th>Protein conc (mg/mL)</th>
<th>Volume (mL)</th>
<th>Total protein (mg)</th>
<th>Solvent</th>
<th>MDH in planta activity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intracellular</td>
<td>10.35</td>
<td>4.5</td>
<td>46.58</td>
<td>Tris-Cl, pH 7.5</td>
<td>yes</td>
</tr>
<tr>
<td>Extracellular</td>
<td>1.88</td>
<td>0.2</td>
<td>0.38</td>
<td>water</td>
<td>no</td>
</tr>
</tbody>
</table>

*mannitol dehydrogenase

Figure 5.1: Proteogenomics samples (a) SDS PAGE separation of proteins in samples sent for peptide sequencing, amount loaded is displayed above the figure. (b) Effect of infiltrated PmedOMT5 culture filtrate (CF) grown for 35d in F2NY media sent for peptide sequencing, contrasted with infiltrated media (Fries2 no yeast) on leaflets of *M. truncatula* accession SA3054, 7 dpi.
Table 5.3: Number of proteogenomic peptide matches supporting the PmedOMT5 version 2 protein set.

<table>
<thead>
<tr>
<th></th>
<th>Intracellular proteins</th>
<th>Intracellular 6 frame</th>
<th>Extracellular proteins</th>
<th>Extracellular 6 frame</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of proteins identified by peptide matches</td>
<td>433</td>
<td>380</td>
<td>159</td>
<td>117</td>
</tr>
<tr>
<td>Number of unique peptides matched</td>
<td>1,257</td>
<td>970</td>
<td>357</td>
<td>236</td>
</tr>
<tr>
<td>Number of proteins identified by multiple unique peptides (range of peptide count per protein)</td>
<td>244</td>
<td>199</td>
<td>77</td>
<td>53</td>
</tr>
<tr>
<td>Number of proteins identified by single peptides</td>
<td>188</td>
<td>181</td>
<td>82</td>
<td>64</td>
</tr>
</tbody>
</table>

Figure 5.2: Summary of overlap between genes with peptide support in intracellular and extracellular proteomic fractions.

Of the 170 proteins identified in the extracellular fraction 58 (34%) were also observed in the intra-cellular fraction. Of these 37 (64%) were predicted to be secreted by all secretion prediction programs (WolfPSORT, SignalP, Phobius- chapter 6.2.5). Of the proteins detected only in the extracellular fraction 59 out of 112 (53%) were predicted to be secreted by all three methods and 73 (65%) were predicted to be secreted by one or more methods. Of those with intracellular peptide support that did not overlap with the extracellular fraction (400), only 61 (15%) were predicted to be secreted by all three methods and 97 (24%) were predicted to be secreted by one or more.
5.3.3 Transcriptome libraries

Transcript data was produced from three in vitro cultured libraries corresponding to the vegetative growth, sporulation and secretion phases of the fungal growth cycle, as illustrated in Figure 5.3.

Transcripts were also isolated from 1-5 dpi M. truncatula plants spot inoculated with PmedOMT5 spores. The course of the experiment is illustrated in Figure 5.4 which shows the plants first visible response to inoculation at 2 dpi and the continuing response in the plant as the hyphal growth progressed. This time-course covers the important establishment phase of the disease cycle on the host, from attachment and germination of the pycnidiospores on the plant surface to host penetration and cell death followed by fungal proliferation within the host.

Figure 5.3: In vitro RNA-seq growth conditions: a) 4d MM 30mM glucose, vegetative growth b) 16d MM 30 mM glucose, sporulating c) 4 weeks growth in Fries media - chlorosis/necrosis-inducing (active in planta culture filtrate).
Figure 5.4: *In planta* RNA-seq experimental progress: (a) Inoculated A17 2 dpi plant reaction developing under spore droplet inoculation site (Bright field micrograph); (b) 3 dpi inoculated leaflets showing developing lesions ‘l’, anthocyanin mark indicated by ‘a’ in centre of all leaflets is characteristic of *M. truncatula* accession A17; (c) 3 dpi spores stained with DIOC6 under UV light; (d) 5 dpi Trypan blue-stained hyphae showing leaf penetration via darkly pigmented stomata (arrow) (note: stomata is not clearly visible in this plane); (e) 5 dpi inoculated plants showing spreading lesions and chlorosis spreading out from the inoculation site ahead of hyphal growth (arrow); (f) 11 dpi Trypan blue-stained hyphal growth and plant pigmentation spreading out from the inoculation site.

5.3.4 Sequence analysis and trimming

FastQC analysis showed a median quality across all bases in a read with Phred quality score greater than 34. However there was a marked tail showing wide variation in quality at the ends of reads beginning around 75 bp becoming highly variable after 90 bp (Figure 5.5). In order for reads to be accurately re-aligned to the genome low quality bases and adapter sequences were removed (refer to Methods section 5.2.12).

Figure 5.5: FastQC analysis of quality scores across the read length 16d lane 1, read 2, raw reads.
5.3.5 RNAseq read alignment to the genome assembly

The majority (88%) of all in vitro derived reads could be re-aligned to the PmedOMT5 genome assembly (Table 5.4). However only a very small proportion (0.35%) of the total reads obtained from the early stages of plant infection could be aligned to the PmedOMT5 assembly, with a much larger proportion of 64% (37,192,018 reads) aligning to the plant genome assembly as expected. The reads that aligned to the Medicago assembly supported the expression of 45,108 M. truncatula genes or 95% of the available gene set of its 47,845 experimentally validated genes (Young et al. 2011). A large number of PmedOMT5 genes 8,048 (68%) were covered across their entire length by in vitro-reads but, of the genes detected as expressed in planta, many had incomplete coverage, illustrated in Figure 5.6. The total number of version 2 genes with one or more aligned reads in any sample was 11,643 (98%).

Table 5.4: Number of RNA-seq reads obtained, the percentage remaining after adapter trimming and quality filtering and number that aligned to the PmedOMT5 genome.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Total number of reads obtained</th>
<th>Total number of reads after trimming/filtering</th>
<th>Percentage of reads retained following trimming</th>
<th>Total number of mapped trimmed reads</th>
<th>Percentage of trimmed reads aligned to assembly</th>
<th>Coverage of total CDS length</th>
</tr>
</thead>
<tbody>
<tr>
<td>4d</td>
<td>14,100,702</td>
<td>13,095,640</td>
<td>92.87%</td>
<td>12,381,239</td>
<td>94.54%</td>
<td>64x</td>
</tr>
<tr>
<td>16d</td>
<td>17,072,320</td>
<td>15,808,138</td>
<td>92.60%</td>
<td>14,392,405</td>
<td>91.04%</td>
<td>74x</td>
</tr>
<tr>
<td>F2NY</td>
<td>27,845,674</td>
<td>25,963,820</td>
<td>93.24%</td>
<td>21,616,171</td>
<td>83.25%</td>
<td>112x</td>
</tr>
<tr>
<td>Total in vitro reads</td>
<td>59,018,696</td>
<td>54,867,598</td>
<td>92.97%</td>
<td>48,389,815</td>
<td>88.19%</td>
<td>251x</td>
</tr>
<tr>
<td>Total in planta (1-5d) reads</td>
<td>61,725,890</td>
<td>58,376,942</td>
<td>94.57%</td>
<td>204,613</td>
<td>0.35%</td>
<td>1x</td>
</tr>
</tbody>
</table>
5.3.6 Differential expression

The number of genes with normalised expression data from each sample was calculated using Cufflinks/Cuffdiff (Trapnell et al. 2010, Trapnell et al. 2012) and is presented in Table 5.5. Cuffdiff calculates transcript abundance as a normalised “expected number of fragments per kilobase of transcript sequence per millions base pairs sequenced” (FPKM). This effectively normalises transcript abundance for variation in sequence library yield and gene length, thus enabling meaningful comparison between genes or experimental conditions (Trapnell et al. 2012). The use of FPKM rather than direct read counts corrects for two potential differences when comparing RNA-seq libraries: 1) the fact that longer mRNAs produce more fragments in an Illumina sequencing library preparation, where the cDNA is sheared and size selected to optimise the output from the sequencer (Trapnell et al. 2012) and 2) two sequencing runs of the same library can produce different volumes of sequenced reads and can only be accurately compared when normalised to the total yield of the machine (Trapnell et al. 2012).

Table 5.5: Number of genes expressed in each sample by Cufflinks normalised FPKM.

<table>
<thead>
<tr>
<th>Sample</th>
<th>Genes expressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>4d</td>
<td>10,880</td>
</tr>
<tr>
<td>16d</td>
<td>11,324</td>
</tr>
<tr>
<td>F2NY media</td>
<td>9,950</td>
</tr>
<tr>
<td><em>In planta</em></td>
<td>8,018</td>
</tr>
</tbody>
</table>
5.3.7  *De novo* assembled transcriptome

From the combined *in vitro* library the majority of assembled transcripts (96%) aligned to the assembly, details in Table 5.4. From the *in planta* library only 1,985 (2.1%) assembled transcripts aligned to the PmedOMT5 assembly with a higher bit score than to *M. truncatula*.

There were 1,013 assembled *in vitro* transcripts that did not align to the PmedOMT5 genome assembly (ranging in length from 100-5,403 nucleotides with a total length of 72,728 nucleotides). CENSOR analysis (Kohany et al. 2006) showed that 256 of these were repetitive fragments, with 237 of these matching to transposable elements, including the fungal retrotransposons Molly [GeneBank: AJ488502.1] and Pixie [GenBank: AJ488503.1]. Another 71 were longer, with an average length of 1,895 nucleotides, and had strong hits to other fungal proteins or genomes, three of which matched polyketide synthase gene fragments.

Table 5.6: Number of *de novo* assembled Trinity transcripts.

<table>
<thead>
<tr>
<th>Statistics</th>
<th><em>in vitro</em></th>
<th><em>in planta</em></th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Trinity transcripts assembled</td>
<td>26,101</td>
<td>96,478</td>
</tr>
<tr>
<td>Number of Trinity transcripts that align to the PmedOMT5 assembly</td>
<td>25,110</td>
<td>1,985</td>
</tr>
<tr>
<td>Percentage of transcripts that align to PmedOMT5 assembly</td>
<td>96.1%</td>
<td>2.1%</td>
</tr>
<tr>
<td>Number of Trinity transcripts that align to the <em>M. truncatula</em> assembly v 3.5</td>
<td>-</td>
<td>73,375</td>
</tr>
<tr>
<td>Percentage of transcripts that align to the <em>M. truncatula</em> assembly v 3.5</td>
<td>-</td>
<td>76.1%</td>
</tr>
<tr>
<td>Number of v2 genes overlapped by Trinity transcripts</td>
<td>10,162</td>
<td>970</td>
</tr>
<tr>
<td>Percentage of all v2 genes overlapped by Trinity transcripts</td>
<td>85.6%</td>
<td>8.2%</td>
</tr>
</tbody>
</table>
Table 5.7: Summary of gene support for OMT5 predicted genes.

<table>
<thead>
<tr>
<th>Support for gene models</th>
<th>Number of supported v2 gene models</th>
<th>Percentage of supported gene models</th>
</tr>
</thead>
<tbody>
<tr>
<td>Transcript read support</td>
<td>11,427</td>
<td>96.2%</td>
</tr>
<tr>
<td>(min 5 reads aligned)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Transcript support</td>
<td>10,220</td>
<td>86.0%</td>
</tr>
<tr>
<td>(by alignment from Trinity assembly)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Homology</td>
<td>10,352</td>
<td>87.1%</td>
</tr>
<tr>
<td>(GenBank fungal protein match)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Proteogenomic</td>
<td>570</td>
<td>4.8%</td>
</tr>
</tbody>
</table>

5.4 Discussion

The analysis presented in this chapter outlines the work conducted to validate the predicted gene content of PmedOMT5.

5.4.1 Manual annotation

One of the main difficulties in gene prediction in fungi is caused by the close proximity of their genes. Fungal genomes are known to have much shorter intergenic distances than other eukaryotes such as plants and mammals, the average intergenic distance in the Sordariamycete *N. crassa* is ~ 2kb (Galagan et al. 2003). During RNA-seq analysis of the yeast *Candida parapsolosis* the UTRs of ~50% of adjacent genes were found to overlap. These overlaps occurred both on opposite strands and in the same strand (Guida et al. 2011). In PmedOMT5, transcript reads aligning to neighbouring gene regions were observed to map as little as 20 bp apart and in some cases could not be separated at all. This gene density is known to complicate *ab initio* gene prediction in fungi and cannot be resolved by a combinatorial approach where transcript data is used to delineate exon structure, unless stranded RNA-seq data is used.

Incorrect merging of adjacent gene loci by GeneMark was the major problem identified with the PmedOMT5 version 1 gene models. This made assessment of transcript data very difficult as reads aligned to one half of a merged gene were counted by Cufflinks as being across the whole gene. This meant that genes that actually had increased expression *in planta* were not able to be accurately detected, as the different expression levels for two adjacent genes would be averaged across their combined lengths. This
joining of genes also complicated orthology and functional analysis due to the presence of two domains of dissimilar functions occurring within one gene model. It was also observed that there were several loci that were not predicted in the version 1 gene set but had aligned reads from the in planta sample. These are potential effector gene candidates and would not have been discovered if not manually annotated. Due to the difficulty in processing the sheer number of genes with supporting data and in order to utilise all the data concerning junctions and UTRs from the RNA-seq, the decision was made to use the EVM automated pipeline to combine all available data and then manually curate its predicted gene models.

An example is shown in Figure 5.7 of a GeneMark predicted version 1 gene (aqua) that required manual annotation to split into two correct gene models, based on transcript and homologous protein alignments. The majority of protein alignments (magenta) showed two separate genes, as did the Trinity/PASA aligned transcripts (blue). Note that no TopHat junctions (olive) aligned to the predicted intron in the GeneMark model, which is actually comprised of the UTR from one gene and the intergenic space. The EVM annotation prediction (white) was also unable to separate the two gene models most likely due to the presence of a match to another incorrectly joined protein (from another fungus also annotated solely from GeneMark predictions). It is worth noting that Cufflinks (yellow) did predict the separated gene models but as alternate spliced isoforms of the original gene prediction.

Apollo, the software used to visualise the gene annotations and supporting data displayed in Figure 5.7, was configured to display on either strand some non-strand-specific features such as Trinity and Cufflinks transcript alignments (which can potentially map on either strand), whilst strand-specific features such as protein matches and splice junctions are always displayed on their aligned strand. BLASTP analysis of the two new protein models from the example shown in Figure 5.7, revealed their identity as genes encoding an ATP-binding cassette protein and dynactin.
Although 98% of genes had alignments of at least one RNA-seq read, this number was too low to distinguish from background mis-matches, therefore no gene was annotated on the basis of having only a single aligned read, as alignment of short reads of 100 bp especially where the pair does not map within range are potentially false positives. Gene models that had only low levels of RNA-seq support were instead annotated based on their homology with other fungal proteins, following manual inspection. Eighty-six percent of genes were supported by alignments from the Trinity assembled transcripts, providing a much more reliable level of support as the transcripts were first assembled with a minimum kmer coverage of five at each nucleotide, followed by alignment of the whole assembled transcript to the genome assembly. This demonstrates that the gene set remains well supported when compared to a more conservative transcript data set.

The version 2 gene set contained 1,036 genes not identified by GeneMark or whose models had been incorrectly joined together. A total of 6,764 models (57%) remained unchanged at the amino acid sequence level in version 2 from the GeneMark predicted v1 gene set. However GeneMark does not predict UTRs which were added to many of these genes where the borders could be accurately determined. Following manual annotation, there still remained 230 incomplete gene models, that were missing either a start or a stop codon, due to incomplete assembly of the genomic region surrounding them.
5.4.2 Homology

Genes predicted *ab initio* will most likely have some form of BLASTP match to the non-redundant protein database at NCBI as long as at least one of the exons is annotated in the correct reading frame. But in order to check if low scoring matches are due to real sequence differences or simply mis-annotation of the gene model in one of the compared species, the results must often still be manually inspected.

Manual annotation and BLAST analysis revealed that in many publically available fungal genomes the 5SrRNA has been annotated as a small protein-coding gene. This was confirmed by Infernal analysis of PmedOMT5 (chapter 3.4.8). Because of this mis-annotation the genomic sequence would show a homologous protein match in all compared species even though no translated protein exists. Several researchers are now looking at errors in public data bases including Pfam (Eberhardt *et al.* 2012), with the aim of minimising perpetuation of these types of annotation errors by automated pipelines.

5.4.3 Proteogenomic analysis

Shotgun proteomics, whereby proteins are digested into peptides and then sequenced via tandem mass spectrometry (MS/MS) has been routinely used to validate gene annotation and identify proteins translated under specific experimental conditions. It provides the largest amount of peptide data, although there is a known bias towards larger and more abundant proteins (Brunner *et al.* 2007). Because not all proteins are expressed at all times or can be isolated and detected using all methods (especially those of fleeting expression or low abundance), it would be much more difficult and expensive to validate the entire gene set based on proteogenomics alone. Nevertheless, when used to identify proteins with specific properties, such as secreted proteins in pathogenically-active culture-filtrate fractions, it is of extremely high value for the purpose of gene annotation.

In this study a total of 570 predicted version 2 proteins were found to have support from at least one peptide in either of the samples tested. As illustrated in Table 5.2, much less protein was obtained from the extracellular fraction. This may account in part for the lower number of proteins identified from this fraction. Examination of mannitol dehydrogenase (MDH) activity in the fractions sent for peptide sequencing (Table 5.2)
indicates that there was unlikely to be substantial contamination of the extracellular fraction with intracellular proteins and that the proteins extracted from the intracellular fraction were sufficiently intact to retain their enzymatic activity. Although it is possible that degradation of the proteins in the extracellular fraction may have affected detection of MDH activity as there were no stabilising factors included in the extraction media.

As 58 proteins were detected in both samples (section 5.3.2), and not all proteins detected in the extracellular fraction were predicted to be secreted, it is possible that there was some cross-contamination of the extracellular fraction. However, fungi are known to have non-classical secretion pathways than cannot be well computationally predicted so the fact that these proteins were not predicted to be secreted by the methods used does not mean they are not secreted by the fungus. Thus the source of these proteins is not known. Detection of intracellular contamination of the extracellular fraction could be assessed in future experiments via methods such as western blotting against an intra-cellular protein e.g. actin or elongation factor 1-alpha, to identify the presence of contaminating intracellular proteins.

There was support for a smaller number of gene models based on the digested 6-frame matches compared to the digested protein matches due to the larger number of in silico peptide fragments in that data set (1,324,085 open reading frames prior to trypsin digestion) as opposed to the protein set (10,843 v1 predicted gene models prior to digestion). Whether a peptide match is reported by MASCOT is dependent upon its expectation value (e-value) however as the e-value is dependent upon the size of the database being searched some of the matches that were above the significance threshold in the 6-frame data set are below the significance threshold in the protein database due to its smaller size.

For the purposes of this study genes with only one peptide hit were considered as having support for translation, as long as they also had supporting homology or transcript data. Only peptides that had been filtered for unique sequences with matches to the database above the homology threshold with a significance value of \( p < 0.05 \) were considered. The homology threshold is designed to identify outliers in the MS/MS spectrum that are unlikely to be the result of background noise.

A number of issues need to be considered when dealing with proteogenomic data. Importantly, a peptide data match does not validate the entire gene model, simply adds
another level of support to the prediction. Providing a further indication that the annotation in the region that is supported by the peptide is likely to be in the correct strand and frame. A single peptide also cannot be considered as strong experimental support for gene models derived from a fragmented genome as the peptides can only be reliably matched to the input data. If the assembly is incomplete a peptide cannot be matched to a missing sequence. Having issued these caveats, based on analysis in the previous chapter it is likely that the majority of the gene coding regions of the PmedOMT5 genome were captured in the assembly (CEGMA analysis- chapter 3.3.3), but there is still a small amount of missing data as evidenced by the gaps in the assembly and unmapped de novo assembled transcript data, discussed in section 5.4.5.

In previous studies using peptide data to validate gene annotation (Spanu et al. 2010, supplementary data), genes were required to have matches from a minimum of two peptides. In this study, as the peptide data was not used as the primary means of validating gene annotation but merely to support the transcript data, one peptide match was considered sufficient if it did not conflict with other supporting data.

A small number of peptides (26) matched within a gene locus but on the opposing strand, placing them in conflict with previous gene models (version 1). However their locations were not supported by BLAST data. i.e. there were gene models in these genomic regions and their current structures were supported both by transcript data and homology to other fungal proteins in the NCBI database or the local Dothideomycete database. These peptides may have been derived from real Phoma proteins that reside within unassembled regions of the genome, from contaminants or misinterpretation of the spectral data. Another possibility is that the mass spectra were derived from post-translationally modified proteins and therefore their masses were not accurately assigned to the peptide database. Many of these spurious peptides had scores just above the identity threshold, whereas validated peptides had scores of 15 or more above the threshold.

When comparing the transcriptomics and proteomics data, only one mismatch was observed between these datasets. One peptide mapped to an intron that was well supported by transcript data but may simply have been the product of an alternately spliced isoform whose transcript was not captured by RNA-seq. A single gene (PmedOMT5_01555) has a peptide match but no overlapping transcript reads. This gene is highly conserved across Ascomycete fungi and contains a conserved HET domain.
The peptide was derived from the intracellular 6-frame dataset, a sample for which there is no corresponding transcript sequencing library (7 days growth in liquid minimal media) which suggests that this protein may not be expressed in the other conditions examined.

On its own, the proteogenomic evidence provides relatively weak support for v2 genes. Many genes have only one peptide match and peptides mapped to version 1 proteins were simply re-mapped to the version 2 proteins due to a lack of resources for a repeat run of the Mascot analysis with the new protein set. Therefore, this data is not as reliable as it could have been. The gene models most confidently supported by peptide evidence are those with support from multiple peptides matching to the 6 frame translation (199 intracellular and 53 extracellular –total 239, 2% of v2 proteins). Nevertheless the remaining peptides also concur with the gene models predicted by both RNA-seq and homology and thus provide another layer of supporting evidence for these genes.

5.4.4 Transcriptomic analysis

Several studies in recent years have examined the fungal transcriptome both in vitro (Guida et al. 2011) and in planta (O’Connell et al. 2012, Soanes et al. 2012). These studies have demonstrated that RNA-seq can be used to validate predicted gene structure, identify novel protein-coding genes and elucidate pathogen plant interactions. RNA-seq is noted to have substantially less technical variability than other gene expression assays (Trapnell et al. 2012) however sensitivity is lower if transcripts have less than 5-fold read coverage (Trapnell et al. 2009).

Both methods of RNA-seq analysis, read-mapping and de novo assembly were utilised in this study in order to optimise the chances of discovering novel PmedOMT5 transcripts. RNA-seq analysis was able to confirm the annotated exon structure and mRNA expression of the vast majority of PmedOMT5 genes, with 96% having more than 5 reads aligned and with 86 % supported by Trinity de novo assembled transcripts.

The RNA-seq experiments were designed in order to capture information about the largest number of genes possible. In order to do this several different life cycle stages were sequenced in order to provide sufficient data to validate the gene annotations, as well as allow for the possibility of detecting potential effector genes expressed either in
*planta* or in mycelia grown in Fries media which secretes a phytotoxic component. At 4 days growth on minimal media *P. medicaginis* was still in a vegetative growth phase but at 16d was producing mature pycnidia releasing pycnidiospores. In Fries media after 4 weeks growth many of the small colonies arising from individual spores had coalesced to form a pink coloured mat across the surface of the media with some darker coloured colonies that have fixed to the edge of the flask (Figure 5.3). The *in planta* time-points sequenced covered the germination of the fungal spores on the host, followed by penetration of the cuticle and fungal growth within the host cells. In combination these data cover many significant stages in the fungal life cycle and in the establishment of the pathogen on the host.

Whilst the main purpose of the study was to identify genes for annotation, differential expression between the infection and non-infection samples was evaluated for the purpose of identifying genes that may play a role in pathogenicity. These will be discussed in chapter 6.

While calculation of differential expression by Cuffdiff was generated, these data were not used to compare the differences in expression between all samples only to highlight genes expressed more highly *in planta* than *in vitro*. Although the sequences obtained were the results of pooled biological samples of at least three replicates (and more individual plants for the *in planta* analysis) and the libraries were each sequenced in two separate sequencing runs, there was only one library preparation from each of the pooled samples, thus there is no replication from which reliable conclusions could be drawn. Differences can be inferred from detection of expression in one condition versus another but will require further confirmation. The most robust way to do this would be via qRT-PCR analysis of individual genes from an independent experiment or further RNA-seq of independent experiments with a larger number of biological replicates sequenced separately, preferably using stranded-RNAseq.

### 5.4.5 *De novo* assembled transcriptome

The number of *de novo* assembled transcripts from the *in vitro* library that did not align to the PmedOMT5 genome was quite low (3.9%) and accounted for only 70 kb of sequence. This indicates that the unassembled parts of the *Phoma* genome are unlikely to be large or to contain large numbers of genes. Analysis demonstrates that they do contain some repetitive elements including transposons. The fact that the transposons
were not able to be aligned to the assembly was not surprising as a known issue with assemblies from Illumina-sequenced genomes is the difficulty in assembling repetitive regions of the genome accurately. Most importantly what this shows is that there are transposons in the PmedOMT5 genome that are being actively transcribed and not all are RIP-inactivated.

5.4.6 In planta (IP) expression

In planta expression of PmedOMT5 genes was gauged by inoculating *M. truncatula* accession A17 known to be susceptible to PmedOMT5 (Ellwood *et al.* 2006b, Kamphuis *et al.* 2008). Whilst not the most susceptible *Medicago* accession to PmedOMT5, A17 was selected as it is the reference cultivar for the *Medicago* genome sequencing project (Young *et al.* 2011).

Although a much lower number of fungal reads were detected in planta, known fungal house-keeping genes (*actin, eIFα* and *beta-tubulin*) were detected and Cuffdiff analysis of normalised FPKMs did not indicate changes in expression levels greater than the 4-fold threshold (Appendix 6.9). Additionally a wide variation in expression is observed across in planta expressed genes despite the low sequence coverage, although very lowly expressed genes are likely not to have been detected or not counted due to the presence of less than five mapped reads.

The EST library examined [NCBI_EST:LIBEST_010527_Phoma-infected] was isolated from early time-points 1-4 d in an infection-mimicking library, therefore a similar profile was expected to that observed in the 1-5d IP RNA-seq experiment performed in this study. As observed in the IP RNA-seq library the majority of the transcripts were derived from plant mRNA and only a small amount of fungal transcripts were observed, 2.1% of the *de novo* assembled IP RNA-seq transcripts and 2.8% of the EST library. However these time-points are significant in pathogenicity studies, as this is the time when currently known effector genes have been observed to be expressed in the fungus (Ipcho *et al.* 2011, O'Connell *et al.* 2012) and when the reaction is first observed in the host (Pandelova *et al.* 2009).

In summary the transcriptomics analysis provided a vast amount of supporting data for gene annotation and provides insights into the gene expression of PmedOMT5 during several life cycle stages including a pathogenic interaction. The potential roles in
pathogenicity for PmedOMT genes detected during the preliminary IP analysis are discussed in the following chapter.

5.4.7 Summary

The *P. medicaginis* isolate OMT5 genome is predicted to encode 11,879 genes (98%) of which have some level of biological support. This number is in keeping with that seen in other fungi of the order Pleosporales (Ohm *et al.* 2012) and close to the 12,394 seen in *Didymella exigua* CBS.183 55, the only publically available sequenced member of the Didymellaceae family.

The manually annotated gene set provides a valuable resource for researchers working with *Phoma* as well as other closely related pathogens. It also provides a valuable resource for research examining the gene repertoire shared by plant pathogenic fungi.
Chapter 6
Predicted functional annotation of PmedOMT5 proteins, focusing on those potentially involved in pathogenicity
6.1 Introduction:

One of the major goals of whole-genome sequencing in the context of pathogenicity is to shed light on the gene content of an organism. This information can be used to identify similarities and differences between close and distantly related species in order to infer what makes one species or isolate a pathogen of a certain host and not another. In the previous chapter, the gene content of PmedOMT5 was determined. This chapter uses comparative bioinformatics analysis with other pathogen species to highlight genes and biological processes of importance in the PmedOMT5-Medicago interaction, including identification of mating type genes and potential fungicide targets.

In multicellular filamentous fungi, sexual reproduction requires differentiation of reproductive structures and complex developmental processes. Heterothallic (obligate outcrossing) Dothideomycetes have a single MAT locus with two alternate forms (idiomorphs) that must be different for two isolates to mate. These ideomorphs encode single proteins with differing DNA-binding domains, an α domain for MAT1-1 isolates and a high mobility group (HMG) DNA-binding domain for MAT1-2 isolates (Turgeon 1998). The identification of the structure of the MAT locus can provide important information about the lifestyle and evolution of the fungus.

The prevailing wisdom about the interactions between plants and their pathogens describes two main defence strategies employed by plants in their arms race with their invaders. The first level of defence works on the basis of the recognition of conserved, or slowly evolving, PAMPs (Jones and Dangl 2006). These are recognised by membrane-bound receptor molecules in the host cell and trigger a basal defence response known as PAMP-triggered immunity (PTI). When successful, this mode of defence halts further colonisation by biotrophic pathogens (Jones and Dangl 2006). PTI can be overcome by biotrophic pathogens that secrete “effector” molecules, which modify the plant’s response, resulting in effector-triggered susceptibility (ETS) which allows the pathogen to exploit the host cell without killing it (Stergiopoulos and de Wit 2009).

The second bastion of plant defence, known as effector-triggered immunity (ETI), occurs inside the cell when effectors (some also known as avirulence proteins) secreted by the fungus interact either directly or indirectly with host receptors, which are often NBS-LRR proteins (Jones and Dangl 2006). If the effector is ‘recognised’ by its
cognate receptor it triggers a defensive response known as the ‘hypersensitive response’ at the site of infection. This is usually accompanied by an oxidative burst leading to cell death of the infected cells, resulting in cessation of the invasion by biotrophic pathogens and thus effector-triggered immunity. For necrotrophic pathogens this reaction between an effector and its cognate receptor or sensitivity gene has been shown to trigger programmed cell death (PCD) allowing for successful invasion and disease (Stergiopoulos et al. 2013). The interaction of a necrotrophic host-specific effector and the receptor gene required to trigger disease susceptibility is considered an inverse of the typical ‘gene-for-gene’ reaction (Flor 1942) seen in plant-biotroph interactions usually resulting in ETI and there is tantalising evidence that some may interact directly or indirectly with the same NBS-LRR protein with different results. This raises the possibility that in some cases the same NBS-LRR protein in a host cell may confer both resistance to a biotrophic pathogen and susceptibility to a necrotroph (Lorang et al. 2012, Stergiopoulos et al. 2013).

Pathogenicity in fungi has evolved multiple times (van der Does and Rep 2007), potentially due to the acquisition of effector genes by species that previously had led saprophytic or endophytic lifestyles. Effector genes of a pathogen encode products that have an effect on host plant metabolism or defences. Effector emergence in a pathogen can arise via lateral gene transfer or duplication, expansion and subsequent divergence of paralogs (Friesen et al. 2006, Godfrey et al. 2010). Genes that encode effectors (previously described as host-specific toxins or avirulence determinants) are often unique to a pathogen, although there is mounting evidence that these genes can be transferred between fungal species (Friesen et al. 2006) and even across phyla (de Jonge et al. 2012). Many effector genes have been found in proximity to regions of repetitive DNA including *P. nodorum* SnToxA and SnTox3, which are flanked by A:T-rich sequences containing long terminal repeat (LTR) retrotransposons (Oliver and Solomon 2010). It has been hypothesised that proximity to transposable elements may confer the potential to relocate to new genomes. In some cases this may even facilitate gene transfer across kingdoms, as is the case for Ave1, a *Verticillium* effector suggested to increase xylem sap flow leading to accelerated host colonisation, which is proposed to have been transferred from a plant to a fungus (de Jonge et al. 2012). Proximity to repetitive DNA, regardless of transposon activity, may also allow for effector adaptation via increased mutation rates, from the leakage of repeat-induced point mutations that
have been targeted to a nearby repeat (*e.g.* Anrml4 and Avrlm6 in *L. maculans* (Van de Wouw *et al.* 2010)) or increased potential for recombination.

Effectors can be host specific or non-specific. Host-specific effectors confer pathogenicity on a particular host species to an isolate, whilst non-specific effectors are effective on a broad range of hosts and tend to play roles in virulence of the pathogen, contributing to the pathogen’s ability for growth within the plant (Tsuge *et al.* 2013). Host-specific effectors identified to date are limited to less than 20 pathogenic fungi and most of these are low-molecular-weight secondary metabolites (Tsuge *et al.* 2013), 11 of which are produced by fungi in the genera *Alternaria* and *Cochliobolus*. One of the first host-specific effectors identified was AK-toxin identified in the culture filtrate of *Alternaria kikuchiana* (Tanaka 1933).

The mode of uptake of effectors is not known for all models but has been proposed for Oomycete effectors containing a conserved RXLR-dEER amino-acid motif to involve binding to phosphatidylinositol-3-phosphate (PI3P) lipids in the host cell membrane (Kale *et al.* 2010), which enables uptake into the plant cell directly. In this and several other cases, host-cell import is also dependent upon the presence of certain motifs in the effector protein sequence, usually located at the N terminal (outlined in Table 6.1 and section 6.4.14 and reviewed in Kale *et al.*, 2012). A similar mechanism is thought to operate for some fungal effectors with related motifs. The RGD motif identified in the PtrToxA protein has also been demonstrated to be essential for its uptake into the cell and induction of necrosis (Meinhardt *et al.* 2002, Manning *et al.* 2004).

Not all genes involved in pathogenicity and virulence are effectors. Genes found to be important in pathogenicity play a vast range of roles in fungi such as penetration, sporulation, signalling, detoxification of defence compounds and carbohydrate degradation (Baldwin *et al.* 2006), or simply confer increased growth rate or the ability to sporulate faster allowing the pathogen to better evade the host’s defences.

Plant pathogens with different feeding strategies (biotrophs, hemi-biotrophs and necrotrophs) share some similarities in the types of proteins involved in causing disease on their hosts (Baldwin *et al.* 2006). Two categories of proteins important for virulence in fungal pathogens are: 1) carbohydrate active enzymes (CAZymes) that can break down plant cell wall components during the pathogen’s attempt to gain access to the nutrients behind the wall and generate a nutrient source for the fungus from the
breakdown of the cell-wall components (Favaron et al. 2004, Sexton et al. 2006, Andrew et al. 2009) and 2) proteins that metabolise, i.e. neutralise, defence molecules produced by the plant in response to infection, such as phytoalexins (Pedras and Ahiahonu 2005).

One of the goals of this study was to identify genes involved in pathogenicity, including prediction of effector candidates as well as other genes that may be important for the characterisation of the species. This also extended to the examination of factors that can have an effect on disease in the field, such as mating types (MAT idiomorphs) and fungicide resistance genes. Two types of effectors were sought: 1) host-specific necrotrophic effectors that occur only in this pathogen (species-specific) or may have potentially been introduced to PmedOMT5 via lateral gene transfer (LGT) from other pathogens, and 2) conserved, non-specific effectors that may be involved in a wide range of plant-fungal interactions e.g. the extracellular chitin-binding protein Ecp6 which protects chitin molecules of invading fungi from being detected by host defences (de Jonge et al. 2010). To further the search for the latter, PmedOMT5 genes that matched genes that have been investigated experimentally for pathogenicity phenotypes in other pathosystems listed in the Pathogen-Host Interaction database (PHIbase, http://www.phi-base.org/) were examined. The genes described in PHIbase are characterised in terms of their disease phenotype upon disruption of the gene of interest. Studies have shown that disruption of homologous genes in other species often also causes reduced virulence or loss of pathogenicity (Baldwin et al. 2006).

To identify genes with possible roles in pathogenicity in P. medicaginis, mRNA transcripts from four life-cycle stages of the pathogen were sequenced. These included three in vitro growth stages: at vegetative (4d) and sporulating (16d) phases, during production of active culture filtrate (5w) and one during the early stages of infection of M. truncatula leaves (1-5 dpi, when effector genes in other species are known to be expressed) (Chapter 5). The quantitative variations in expression between these in planta and in vitro conditions were used to identify genes that may play a role in pathogenicity. This information was combined with proteomic analyses whereby proteins were identified as produced in an “infection-mimicking” growth stage of liquid culture, during which culture filtrate produced a visible phytotoxic effect when infiltrated into plants presumably due to secreted proteins (chapter 7). The cumulative data from these two sets: 1) genes that were detected as expressed during early in planta
infection (1-5 dpi) and 2) proteins secreted extracellularly by the fungus, were used to predict a set of potential pathogenicity determinants that would be prioritised for future experimental validation.

**Aims for this chapter**

In this chapter the orthology and transcript analyses used to annotate the genome were built upon and functional classifications were added to the PmedOMT5 version 2 protein set based on orthology to characterised genes in public databases (Pfam, InterPro, GO). This information was used to classify genes into: 1) genes potentially involved in pathogenicity as novel effector genes that were expressed *in planta* or; 2) gene families that may play a more conserved role in pathogenicity of *P. medicaginis* on *M. truncatula*, based on their functional classification, including predicted secretion, activity against carbohydrates or matches to conserved amino acid motifs of known pathogenicity genes.

### 6.2 Materials and methods

#### 6.2.1 Analysis of mating type genes

The MAT region from *Ascochyta lentis* [Genbank: DQ341315.2] and *Ascochyta rabiei* [Genbank: DQ341312.1] were aligned using clustalW (Chenna *et al.* 2003, Larkin *et al.* 2007) and primers designed to amplify the conserved region of the HMG domain using Primer3plus (Rozen and Skaletsky 2000). These primers were tested on several *P. medicaginis* isolates including the reference strain and other related *Phoma* species (WAC7988, WAC7978, WAC7980, WAC7977, WAC4738, WAC, 4736, WAC4741, OMT1, OMT5, CBS316.90, CBS318.90). Primers were also designed to try to amplify the entire MAT idiomorph region and potentially from an area external to the MAT region into the conserved alpha domain of MAT1-1-1 if present.

#### 6.2.2 Functional annotation

PmedOMT5 version2 predicted proteins were assigned putative functional roles via bioinformatic analysis of conserved domains and predicted protein properties. As functional annotation methods were numerous and diverse, their descriptions have been divided into the following sub-sections.
6.2.3  Conserved domain analysis

Conserved domains were identified using HMMER v 3.0 (Finn et al. 2010) against the PFAM-A database (v 27.0) with gathering cut-offs defined individually for each HMM. Protein predictions were also examined using InterProScan (Zdobnov and Apweiler 2001, Quevillon et al. 2005) and the conserved domain database (CDD) (Marchler-Bauer et al. 2011).

6.2.4  Gene Ontology

Predicted proteins were analysed using BLASTP (Altschul et al. 1997) to the non-redundant protein database downloaded from NCBI (24/03/2013) and the results analysed using Blast2GO v 2.3.6 (Conesa et al. 2005), with the following parameters: annotation, E-value hit filter 1E^-6, annotation cut-off 55, GO weight 5, high scoring segment pairs (HSP) hit coverage cut off 0. Only validated GO terms were used for functional annotation.

6.2.5  Prediction of cellular localisation and secretion

Putative secreted proteins were identified via combined analysis with SignalP 4.1b (Petersen et al. 2012) parameter: eukaryote, WolfPSort (Horton et al. 2007) parameter: selected organism fungi and Phobius 1.01 (Kall et al. 2004). These analyses were combined to create a list of proteins predicted to be secreted by all three algorithms or those predicted to be secreted by one or more. The presence of transmembrane domains was analysed using Phobius 1.01 (Kall et al. 2004) and TMHMM (Krogh et al. 2001).

6.2.6  Carbohydrate active enzyme (CAZY) analysis

CAZyme activities (www.cazy.org), (Cantarel et al. 2009) were assigned to protein sequences via the web interface of the CAZymes Analysis Toolkit (CAT) 1.1 beta (http://mothra.ornl.gov/cgi-bin/cat.cgi), (Park et al. 2010) using its default settings for PFAM based analysis (E value- 0.01, bit score- 55, level of support- 40 and using rules).

6.2.7  Secondary metabolite synthesis gene prediction

PKS genes were identified by PFAM domain and further analysed for domain structure via the PKS/NRPS analysis website at http://nrps.igs.umaryland.edu/nrps/ (Bachmann and Ravel 2009). All genes were analysed via the Secondary Metabolite Unique
Regions Finder (SMURF) (http://jcvi.org/smurf/index.php) for prediction of secondary metabolite biosynthesis gene clusters (Khaldi et al. 2010).

6.2.8 Statistical examination for over- or under-representation of protein functional attributes

The number of genes with specific functional attributes (secreted proteins, GOs, Pfam domains and CAZymes) that were found in genes detected as expressed in planta or showing a more than 4-fold increase in expression in planta compared to the other conditions tested versus those in the version 2 protein set as a whole, were examined using Fisher’s exact test at a significance threshold of $p \leq 0.01$.

6.2.9 Comparison to PHIbase proteins with experimentally defined pathogenicity profiles

All PmedOMT5v2 and scaffolds were compared to experimentally tested pathogenicity proteins identified in the Pathogen-Host Interaction database (PHIbase) (Baldwin et al. 2006, Winnenburg et al. 2006, Winnenburg et al. 2008) by BLAST versus PHIbase v 3.2 (924 sequences) and all 3.4 (Uniprot- 1,245, NCBI- 806 proteins and NCBI- 67 nucleotide sequences) for which data was readily available. BLASTN, BLASTP and TBLASTN with a significance threshold of $10^{-5}$ to identify any genes with similarity to experimentally validated plant pathogenicity genes. For genes with multiple matches the best match was selected by highest bit score from the closest related pathogen.

6.2.10 Analysis of known effector motifs

Proteins were assessed for the presence of motifs known to be associated with effectors in other fungal and oomycete pathogens, in both the initial 70 aa, and at any location within the encoded protein using EMBOSSpreg v 6.5.7 (http://emboss.sourceforge.net/apps/release/6.3/emboss/apps/preg.html). Motifs and references listed in Table 6.1.
Table 6.1: List of effector associated motifs searched for and relevant references.

<table>
<thead>
<tr>
<th>Motif</th>
<th>Reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>RGD</td>
<td>Sarma et al. (2005), Manning et al. (2004), Meinhardt et al. (2002)</td>
</tr>
<tr>
<td>RXLR(X[5,30])EER</td>
<td>Whisson et al. (2007), Dou et al. (2008), Grouffaud et al. (2008)</td>
</tr>
<tr>
<td>RXLR</td>
<td>Whisson et al. (2007), Dou et al. (2008), Grouffaud et al. (2008)</td>
</tr>
<tr>
<td>RFYR</td>
<td>Kale et al. (2010)</td>
</tr>
<tr>
<td>RTLK</td>
<td>Kale et al. (2010)</td>
</tr>
<tr>
<td>RYWT</td>
<td>Kale et al. (2010)</td>
</tr>
<tr>
<td>RMLH</td>
<td>Kale et al. (2010)</td>
</tr>
<tr>
<td>RIER</td>
<td>Kale et al. (2010)</td>
</tr>
<tr>
<td>CHXC</td>
<td>Kemen et al. (2011)</td>
</tr>
<tr>
<td>LXLFLAK</td>
<td>Win et al. (2007), Haas et al. (2009)</td>
</tr>
<tr>
<td>YXSL[RK]</td>
<td>Levesque et al. (2010)</td>
</tr>
<tr>
<td>[YFW]XC</td>
<td>Godfrey et al. (2010)</td>
</tr>
<tr>
<td>HVLVXXP</td>
<td>Haas et al. (2009)</td>
</tr>
<tr>
<td>[FL]XLYLALK</td>
<td>Gaulin et al. (2008)</td>
</tr>
<tr>
<td>LXLYLAXR</td>
<td>Chueng et al. (2008)</td>
</tr>
<tr>
<td>KECD</td>
<td>Nicastro et al. (2009)</td>
</tr>
<tr>
<td>RQHHRK(X[5,15])HRRHK</td>
<td>Kemen et al. (2005)</td>
</tr>
<tr>
<td>[L]XR</td>
<td>Yoshida et al. (2009), Li et al. (2009)</td>
</tr>
<tr>
<td>[RK]CXXC(X[6,18])H</td>
<td>Yoshida et al. (2009)</td>
</tr>
<tr>
<td>[RK][YY][L]R</td>
<td>Ridout et al. (2006)</td>
</tr>
</tbody>
</table>

Note: RXLR and RXLR...dEER were counted as separate motifs. Parentheses “()” indicate a subset of the pattern. Brackets “[]” indicate multiple possible amino acids at that position. The letter ‘X’ indicates any amino acid can be substituted at that position. Braces “{}” indicate the minimum and maximum allowable range of amino-acids at a given position.

6.2.11 Orthology

Proteinortho v4.1 (Lechner et al. 2011) was used to detect orthologs of PmedOMT5v2 proteins compared with 41 other fungi and Oomycetes: Didymella rabiei, Peyronellaea pinodes, Didymella exigua, Phaeosphaeria nodorum, Pyrenophora teres f. teres, P. tritici-repentis, Leptosphaeria maculans, Zymoseptoria tritici (formerly Mycosphaerella graminicola), Magnaporthe oryzae (formerly grisea), Neurospora crassa, Dothistroma septosporum, Alternaria brassicicola, Cochliobolus heterostrophus, Macrophomina phaseolina, Mycosphaerella fijiensis, Grosmania clavigeria, Sordaria macrospora, Podospora anserina, Trichoderma reesei, Fusarium graminearum, F. oxysporum f. sp. lycopersici, F. verticilliodes, Nectria haematococca, Uncinocarpus reesii, Histoplasma capsulatum, Aspergillus clavatus, A. flavus, A. fumigatus, A. nidulans, A. niger, A. oryzae, Neosartorya fischeri, Coccidioides immitis, Blumeria graminis, Tuber melanosporum, Saccharomyces cervisiae, Schizosaccharomyces pombe, Ustilago maydis, Phytophthora infestans, P. ramorum and Phythium ultimum.
Details of the specific isolates and data sets used are provided in appendix 5.1. Orthologs were determined via reciprocal BLASTP using parameters: e=1e⁻⁵ id=25, conn=0.1, m=0.75, retaining both pairs and singletons.

6.2.12 Score based assessment of potential proteinaceous host-specific effectors

Proteins were assigned ‘effector scores’ based on the sum of weighted scores for various attributes according to criteria deemed to have potential for association with effector genes based on the literature, with a higher score indicating a higher likelihood of pathogenic activity. Scores were assigned as follows: detected as expressed in planta by Trinity alignment (8), protein detected in the extracellular fraction (7), increased expression in planta (4), greater than 50% of gene covered by reads IP (2), EST support (3), no BLASTP hit in GenBank (4), greater than two cysteines in the mature protein (3), presence of known effector associated domain (1), predicted molecular weight of less than 40 kDa (5), predicted to be secreted by all in silico analyses (3), predicted to be secreted by one or more methods (2), within 2 kb of a repetitive region (2) and greater than 5% cysteines (3), five or more reads detected IP (2). Proteins with scores greater than 25, were deemed to have greater potential as effector candidates.

6.3 Results

6.3.1 PmedOMT5 mating type

The PmedOMT5 assembly was found to contain a predicted gene encoding a 351 aa protein (PmedOMT5_08081) with high homology to the A. rabiei MAT1-2-1 protein [GenBank: ABC70319]. The gene is located on a 69,938 kb scaffold, which contains 32 genes, PmedOMT5_8069- PmedOMT5_8100. Genes surrounding the MAT idiomorph were conserved in PmedOMT5 and several related species (data not shown). A combination of different primer sets was used to test the amplification of the MAT region in the PmedOMT5 and other isolates from Phoma spp. described in Ellwood et al. (2006b) as well as the reference isolates for P. medicaginis var. medicaginis (CBS316.90) and P. pinodella (CBS 318.90) from Centraalbureau voor Schimmelcultures (CBS-Utrecht, The Netherlands).

All seven West Australian isolates identified as P. medicaginis contain the MAT1-2-1 idiomorph based on amplification with primers specific to the conserved HMG region (Figure 6.1), as does the CBS reference culture for this species (product not shown for
Expression of mRNA for this gene was very low under the conditions studied but supported the intron location within the gene model.

Figure 6.1: Amplification of MAT1-2-1 fragment in reference cultures and West Australian *Phoma* spp. isolates (described in Ellwood et al. 2006).

(a) PCR products following amplification with (A) HMG primer set and (B) mat1-2 primer sets from genomic DNA from 1) PmedOMT5 (arrow indicates band of interest 228 bp), 2) CBS316.90 reference isolate *P. medicaginis* var. *medicaginis*, 3) CBS 318.90 reference isolate *Phoma pinodella*, 4) WAC7980 *P. medicaginis*, arrow indicates band of interest 216 bp, NTC) ‘no template control’, unmarked lanes contain DNA ladder (HyperladderI- Bioline, Australia). (b) PCR products following amplification with HMG primers from isolates characterised in Ellwood et al 2006. 1) isolate WAC7988 *Phoma exigua*, 2) WAC7978 *Phoma pinodella*, 3) WAC7980 *P. medicaginis*, 4) WAC7977 *P. medicaginis*, 5) WAC4738 *P. medicaginis*, 6) WAC4736 *P. medicaginis*, 7) OMT1 *P. medicaginis* 8) OMT5 *P. medicaginis* (PmedOMT5), arrow indicates band of interest, 228 bp, unmarked lane contains DNA ladder (HyperladderIV-Bioline, Australia).

### 6.3.2 Fungicide target genes in PmedOMT5

The fungal proteins \(\beta\)-tubulin, cytochrome P450 51- sterol 14 alpha-demethylase (CYP51) and mitochondrial cytochrome b (CYTB) are the target sites of fungicide classes in current use (benzimidazoles, triazoles and strobilurins, respectively). Potential mutation sites in the PmedOMT5 \(\beta\)-tubulin, CYP51- and CYTB-homologous genes, which determine fungicide resistance or susceptibility, were examined to assess the state of fungicide applicability.

The cytochrome \(b\) gene is present on the PmedOMT5 mitochondrial scaffold (chapter 4) and does not contain an intron based on assembly and transcript data. There is no G143A mutation in this isolate but the potential for mutation is there. Only one copy of CYP51 (*PmedOMT5_04693*) was identified, which shows moderate expression levels
and no mutations that are known to cause resistance. The $\beta$-tubulin gene ($PmedOMT5_00352$) shows no E198A mutation, the major resistance conferring mutation observed in other species.

### 6.3.3 Putative Functional annotation

The manually-annotated PmedOMT5 version 2 predicted proteins were functionally characterised via bioinformatic analysis of conserved domains and comparisons to annotated proteins in relevant databases. The functional characterisations for each gene are listed per gene in appendix 5.3 and summarized in Table 6.2 - Table 6.14 focusing on potential pathogenicity related gene groups. More detailed analyses are presented for those characterisations that may have an effect on pathogenicity based on their detection in genes expressed in planta.

The majority of predicted PmedOMT5 proteins (85\%) have orthologs in other fungal species, with 87/11,879 predicted to have orthologs only within the family Didymellaceae. A summary of the number of proteins with orthologs and putative functional annotations based on the analyses performed is presented in Table 6.2.

**Table 6.2: Summary of orthology and functional annotations of PmedOMT5 proteins.**

<table>
<thead>
<tr>
<th>Description</th>
<th>Number</th>
</tr>
</thead>
<tbody>
<tr>
<td>PmedOMT5 proteins with orthologs in at least one of the 41 species compared</td>
<td>10,041</td>
</tr>
<tr>
<td>Ortholog groups containing PmedOMT5 proteins with orthologs in at least one of the 41 species compared</td>
<td>9,001</td>
</tr>
<tr>
<td>PmedOMT5 proteins with no orthologs in other species or paralogs in PmedOMT5 (unique, species-specific)</td>
<td>1,812</td>
</tr>
<tr>
<td>PmedOMT5 proteins with paralogs (2,080 with orthologs in other species)</td>
<td>2,106</td>
</tr>
<tr>
<td>PmedOMT5 proteins with paralogs and no orthology in other species (species-specific)</td>
<td>26</td>
</tr>
<tr>
<td>Proteins with match to PHIbase sequences</td>
<td>2,776</td>
</tr>
<tr>
<td>Proteins containing Pfam domains</td>
<td>7,826</td>
</tr>
<tr>
<td>Proteins with InterProscan match</td>
<td>5,022</td>
</tr>
<tr>
<td>Proteins with validated GO annotations</td>
<td>6,356</td>
</tr>
<tr>
<td>Proteins that have KEGG enzyme codes (ECs)</td>
<td>1,820</td>
</tr>
<tr>
<td>Proteins with CAZyme annotations</td>
<td>472</td>
</tr>
<tr>
<td>Proteins with amino acid motifs found to be associated with effectors in other species (at any position)</td>
<td>6,748</td>
</tr>
</tbody>
</table>

### 6.3.4 Secreted proteins

A total of 1,020 proteins were predicted to be secreted by all of three prediction methods (SignalP, Wolf PSORT and Phobius), with 2,114 proteins predicted to be secreted by one or more methods. The intersection between the three prediction methods is illustrated in Figure 6.2a. Of those proteins predicted to be secreted by all
methods, 20 were predicted to contain more than one transmembrane domain by both TMHMM and Phobius. Of those 1,020 proteins predicted to be secreted by all methods, transcripts of 859 of the genes encoding them were detected as more highly expressed in the induced phytotoxin production Fries media (F2NY) sample and 96 of those proteins were detected in the extracellular protein fraction via proteomic analysis. Of the total 1,020 proteins predicted to be secreted by all 3 methods, 20% (203) had CAZyme annotations.

A total of 420 proteins that were predicted to be secreted by all three methods were expressed in planta with 119 showing a more than 4-fold increase in relative expression compared to the other three in vitro samples derived from vegetative and sporulating cultures and during production of secreted phytotoxic metabolites in liquid culture. Of those 119, only 24 were CAZymes with cell-wall degrading functions such as cutinases, pectate lyases, pectin esterases and glycosyl hydrolases. Other proteins predicted to be secreted include subtilisin-like serine proteases and 10 proteins with no BLAST hit to the NCBI protein database. The total number of Wolf PSORT predicted, secreted proteins, is similar across a range of plant pathogens (mean 1,324, standard deviation 140) Figure 6.2b.

![Figure 6.2: Protein localisation](image)

a) The number of the 2,114 total predicted secreted proteins predicted to be secreted by each program and their intersect b) Comparison of the localisation of fungal proteins predicted by Wolf PSORT in fungal plant pathogens, the model organism *Aspergillus nidulans* and the orange bread mold *Neurospora crassa*.
6.3.5  *In planta* (IP) expression

The majority of the analyses presented in this chapter focuses on the 5,778 genes that were detected as expressed *in planta* (IP) calculated by the mapping of five or more RNA-seq reads to the gene loci. This low cut-off was selected as the study focused on early time-points during infection, meaning the number of fungal transcripts sequenced was low, reflecting the low fungal biomass at this early stage of the infection. The decision was made to accept the risk of false positives rather than miss anything that might be important in pathogenicity but without sufficient sequence depth. However there will also be false negatives, as reads from some expressed genes may fall below the threshold set. Thus the emphasis for analysis of this preliminary experiment was placed on what was detected as expressed *in planta* and might thus be contributing to pathogenicity. Although Cuffdiff calculates an FPKM value for all genes that have one or more mapped reads, these were only reported as greater than zero in this chapter if there were more than five reads mapped to the locus.

Initial analyses were targeted to genes that may have roles in the pathogenicity or virulence of PmedOMT5 based on their similarity to characterised genes in the other fungi. Of the genes that were detected IP, 23% (1,348) had characteristics associated with pathogenicity in the form of matches to known pathogenicity or virulence genes in PHIbase, CAZyme annotations or they were detected in the extracellular protein fraction (Chapter 7) and thus potentially secreted *in planta* (summarised in Figure 6.3). Of those 92 genes that had matches in the *P. medicaginis-M. truncatula* EST library, 76 (83%) were also detected as expressed IP by five or more mapped reads in this study, a further nine had 1-4 mapped reads.
6.3.6 Genes that show increased relative expression IP

In total, 461 genes show increased expression IP relative to the other conditions tested based on CuffDiff analysis: 16 of these were also detected in the extracellular peptide fraction, 42 were annotated as CAZymes and 106 have matches to PHIbase genes involved in virulence and pathogenicity. Some of the main roles found in this gene set include: proteins involved in plant cell wall degradation and fungal cell wall reorganization, transporters, cytochrome p450s, oxidation-reduction processes and proteolysis. Important genes identified in this group are discussed below.

6.3.7 Carbohydrate-active enzymes (CAZymes)

Carbohydrate-active enzymes (CAZymes) are responsible for the breakdown, biosynthesis or modification of carbohydrate moieties of glucoconjugates or oligo- and poly-saccharides (Zhao et al. 2013). They are important in plant pathogenic fungi for the role they play in degrading the plant cell walls to allow fungal penetration and nutrient access, as well as in the biosynthesis and modification of fungal cell walls. There are four main functional enzyme classes: glycoside hydrolases (GHs),

Figure 6.3: Proportional Venn diagram summarising the overlap between characteristics of the total 5,778 in planta expressed genes.
polysaccharide lyases (PLs) and carbohydrate esterases (CEs) known collectively as cell-wall degrading enzymes (CWDEs), as well as glycosyltransferases (GTs) with an additional category for the non-enzymatic carbohydrate-binding modules (CBMs) (Cantarel et al. 2009). The presence of carbohydrate active enzymes in PmedOMT5 was identified by comparing conserved Pfam domains to the CAZy database which contains families that are based on experimentally characterised proteins combined with sequences with significant similarity from public databases (Cantarel et al. 2009). Most CAZymes have a complex modular architecture that may involve several functional domains as well as several non-catalytic domains (CBMs) (Park et al. 2010). CAZy analysis via the database analyses each module separately and thus the same protein may be assigned to several families, as was the case for 11 PmedOMT5 proteins. The number of occurrences of members of each CAZyme class in PmedOMT5 proteins is listed in Table 6.3.

Initial CAZyme analysis showed a similar profile to other fungi of the order Pleosporales (Ohm et al. 2012) with 472 proteins (4% of the total PmedOMT5 proteins) containing 95 individual types of CAZy domains (CAZyme domains by protein are listed in appendix 5.3 and their relative numbers in each family compared to other Dothideomycetes are listed in appendix 6.1). Of the 472 genes encoding proteins with CAZyme domains, 257 were detected as expressed in planta (overlapped by 5 or more reads-Appendix 6.8) with 42 of those showing more than 4-fold increase in relative expression. The number of proteins with CAZy domains assigned by PFAM analysis in all proteins and those for which the encoding gene was detected as expressed in planta are outlined in Table 6.4. Of those detected in planta, 81 had matches to PHIbase genes annotated as playing a role in plant pathogenicity (with 16 of these showing increased expression IP). There were 26 CAZymes detected in the extracellular protein fraction and genes for 21 of those were also detected as expressed in planta (discussed in chapter 7). Amongst those genes that showed increased expression IP versus the other conditions tested or were simply detected as expressed IP, no individual CAZymes were identified as over- or under-represented using Fishers exact test at $p \leq 0.01$. 
Table 6.3: Total number of each CAZyme family type identified in PmedOMT5.

<table>
<thead>
<tr>
<th>CAZyme family type</th>
<th>Number of proteins containing CAZyme domains per family in PmedOMT5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycoside hydrolases (GH)</td>
<td>243</td>
</tr>
<tr>
<td>Glycosyltransferases (GT)</td>
<td>113</td>
</tr>
<tr>
<td>Carbohydrate esterases (CE)</td>
<td>91</td>
</tr>
<tr>
<td>Polysaccharide lyases (PL)</td>
<td>20</td>
</tr>
<tr>
<td>Carbohydrate-binding modules (CBMs)</td>
<td>15</td>
</tr>
</tbody>
</table>

*11 proteins contain domains from two CAZy families

The CAZyme profile of PmedOMT5 is in the most part, similar to other pathogens of dicotyledonous plants, with larger numbers of enzymes devoted to pectin degradation (Zhao *et al.* 2013) such as PL1, PL3, GH28 and GH88 (Table 6.4). Some expansions of gene families with CAZy domains were observed (Table 6.5), relative to the numbers in other Pleosporales plant pathogens analysed by Ohm and colleagues (2012). These included glycoside hydrolases involved in pectin degradation and chitin synthases.
Table 6.4: Number of proteins with CAZy domains assigned by PFAM analysis in all PmedOMT5 version 2 proteins and those for which the encoding gene was detected as expressed in planta by alignment of five or more reads.

<table>
<thead>
<tr>
<th>CAZy family</th>
<th>Number of genes with CAZy family domain</th>
<th>Number of genes with CAZy family domain expressed</th>
<th>CAZy family</th>
<th>Number of genes with CAZy family domain</th>
<th>Number of genes with CAZy family domain expressed</th>
</tr>
</thead>
<tbody>
<tr>
<td>CBM1</td>
<td>5</td>
<td>3</td>
<td>GH62</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>CBM13</td>
<td>1</td>
<td>-</td>
<td>GH63</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>CBM20</td>
<td>3</td>
<td>2</td>
<td>GH65</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CBM21</td>
<td>1</td>
<td>1</td>
<td>GH71</td>
<td>2</td>
<td>-</td>
</tr>
<tr>
<td>CBM42</td>
<td>1</td>
<td>-</td>
<td>GH72</td>
<td>7</td>
<td>6</td>
</tr>
<tr>
<td>CBM43</td>
<td>2</td>
<td>2</td>
<td>GH76</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>CBM48</td>
<td>1</td>
<td>1</td>
<td>GH78</td>
<td>6</td>
<td>3</td>
</tr>
<tr>
<td>CBM50</td>
<td>1</td>
<td>1</td>
<td>GH81</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>CE1</td>
<td>11</td>
<td>7</td>
<td>GH85</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CE4</td>
<td>10</td>
<td>9</td>
<td>GH88</td>
<td>6</td>
<td>4</td>
</tr>
<tr>
<td>CE5</td>
<td>11</td>
<td>6</td>
<td>GH89</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>CE8</td>
<td>4</td>
<td>3</td>
<td>GH92</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>CE9</td>
<td>6</td>
<td>4</td>
<td>GH94</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>CE10</td>
<td>48</td>
<td>39</td>
<td>GH109</td>
<td>8</td>
<td>7</td>
</tr>
<tr>
<td>CE14</td>
<td>1</td>
<td>1</td>
<td>GH125</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GH1</td>
<td>5</td>
<td>2</td>
<td>GH127</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>GH2</td>
<td>6</td>
<td>6</td>
<td>GH128</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>GH3</td>
<td>21</td>
<td>17</td>
<td>GT1</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>GH5</td>
<td>7</td>
<td>4</td>
<td>GT2</td>
<td>59</td>
<td>44</td>
</tr>
<tr>
<td>GH6</td>
<td>3</td>
<td>3</td>
<td>GT3</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH7</td>
<td>3</td>
<td>2</td>
<td>GT4</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>GH10</td>
<td>4</td>
<td>3</td>
<td>GT8</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>GH11</td>
<td>4</td>
<td>1</td>
<td>GT15</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>GH12</td>
<td>3</td>
<td>-</td>
<td>GT20</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>GH13</td>
<td>6</td>
<td>5</td>
<td>GT22</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>GH15</td>
<td>3</td>
<td>3</td>
<td>GT24</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH16</td>
<td>10</td>
<td>9</td>
<td>GT28</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>GH17</td>
<td>1</td>
<td>1</td>
<td>GT32</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>GH18</td>
<td>11</td>
<td>8</td>
<td>GT34</td>
<td>4</td>
<td>4</td>
</tr>
<tr>
<td>GH20</td>
<td>3</td>
<td>3</td>
<td>GT35</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH26</td>
<td>2</td>
<td>-</td>
<td>GT39</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>GH27</td>
<td>4</td>
<td>2</td>
<td>GT41</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH28</td>
<td>10</td>
<td>4</td>
<td>GT48</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH29</td>
<td>1</td>
<td>-</td>
<td>GT50</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH31</td>
<td>10</td>
<td>8</td>
<td>GT54</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH32</td>
<td>2</td>
<td>2</td>
<td>GT55</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>GH35</td>
<td>5</td>
<td>1</td>
<td>GT57</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>GH36</td>
<td>1</td>
<td>1</td>
<td>GT58</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH37</td>
<td>2</td>
<td>2</td>
<td>GT59</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH38</td>
<td>1</td>
<td>1</td>
<td>GT62</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>GH42</td>
<td>1</td>
<td>1</td>
<td>GT66</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH43</td>
<td>15</td>
<td>8</td>
<td>GT69</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>GH45</td>
<td>2</td>
<td>1</td>
<td>GT71</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>GH47</td>
<td>11</td>
<td>10</td>
<td>GT76</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>GH51</td>
<td>2</td>
<td>2</td>
<td>PL1</td>
<td>9</td>
<td>6</td>
</tr>
<tr>
<td>GH53</td>
<td>1</td>
<td>1</td>
<td>PL3</td>
<td>10</td>
<td>3</td>
</tr>
<tr>
<td>GH54</td>
<td>1</td>
<td>-</td>
<td>PL7</td>
<td>1</td>
<td>-</td>
</tr>
<tr>
<td>GH61</td>
<td>21</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6.5: Selected expanded CAZyme families in PmedOMT5 potentially relevant to pathogenicity.

<table>
<thead>
<tr>
<th>CAZy family</th>
<th>Typical substrate</th>
<th>Number in PmedOMT5</th>
<th>Average number in Pleosporales pathogens of dicots*</th>
<th>Average number in Pleosporales pathogens of monocots*</th>
</tr>
</thead>
<tbody>
<tr>
<td>GH28</td>
<td>Pectin</td>
<td>10</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>GH88</td>
<td>Pectin</td>
<td>6</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>PL7</td>
<td>Alginate</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>GT2</td>
<td>chitin</td>
<td>59</td>
<td>15</td>
<td>17</td>
</tr>
</tbody>
</table>

*Data from Ohm 2012, Dicot pathogen species: L. maculans and Alternaria brassicicola Monocot pathogen species: Cochliobolus sativa, Cochliobolus heterostrophus, P. teres f. teres, P. tritici repentinis, P. nodorum and Setosphaeria turcica.

6.3.8 ESTs

Of the 92 genes that had matches to the 93 ESTs from infected Medicago leaves described in section 5.2.15, 76 (83%) were detected as expressed IP, with four showing a relative increase in expression. These included an ABC transporter, PmedOMT5_09990 with PHIbase match (PHI:2309) to a transporter of the phytoalexin camalexin (Stefanato et al. 2009), an FAD-linked oxidase (PmedOMT5_10178) and a phenol 2 monoxygenase (PmedOMT5_08261). In total 22 genes with EST matches also had PHIbase matches, 18 of which were detected as expressed IP in this study. This included a two-component histidine kinase (PHI:550) and several transcription factors and CAZymes.

A separate four genes with EST matches also had peptide matches in the extracellular protein fraction, three of which were detected as expressed IP including a thioredoxin, a GMC-oxidoreductase (PHI:199-Alcohol oxidase from P. fulvum (Segers et al. 2001) and a serine carboxypeptidase. Only 13 proteins (14%) predicted to be secreted by one or more methods (six of which were predicted secreted by all methods) were detected amongst proteins with EST matches (appendix 5.3). There were three proteins with EST matches that had no BLASTP matches to the NCBI protein database, all were predicted to be less than 40 kDa but only one was predicted to be secreted (PmedOMT5_2693, effector score 22). This was also the only gene with an EST match that was predicted to be secreted and was identified as expressed in planta in this experiment.
PmedOMT5 genes with homology in the Plant Host Interaction database (PHIbase)

The Plant Host Interaction database (PHIbase) contains details of 2,630 entries with 2,442 unique PHI numbers whose role in pathogenicity has been experimentally examined (Baldwin et al. 2006, Winnenburg et al. 2006, Winnenburg et al. 2008). The PHIbase entries correspond to bacterial and fungal gene products that play a role in pathogen specificity, virulence and chemical sensitivity (version 3.4 released Feb 2013).

A total of 2,776 PmedOMT5 sequences had a BLASTP or BLASTN match to a corresponding sequence in the PHI database below a significance threshold of 10^{-5}. Of these 1,726 matches were to genes annotated as causing a loss of pathogenicity (358) or reduction in virulence (1,368) on their host when disrupted experimentally and a further 26 had matches to genes classified as effectors or plant avirulence determinants. Of the total 1,748 with characterised roles in pathogenicity, 1,099 were detected as expressed by PmedOMT5 during the first five days of infection of *M. truncatula* accession A17 by more than five mapped RNA-seq reads. These 1,099 genes had matches to 558 PHIbase accessions. The top PHIbase hit by bit score for all genes with a confirmed role in pathogenicity is listed in appendix 5.3. The number of gene matches per PHIbase accession ranged from 73, to PHI:438, a *Botrytis cinerea* gene that encodes a cytochrome p450 monoxygenase to only one, for 287 separate PHIbase entries. Of the 461 genes that showed increased relative expression IP, 106 had PHIbase matches, 18 of these to cytochrome p450 proteins.

Many of the BLASTP matches were not strong (bit scores of less than 200), only 411 genes had bit scores of greater than or equal to 200, and not all of the genes are orthologous but may only contain similar domains. Reciprocal matches were not used, as the aim was to use the widest possible net to identify potential pathogenicity genes.

Of those genes with PHI base matches that were expressed IP, there were 10 with functional annotations suggesting a role in the synthesis of secondary metabolites including PKSs, NRPSs, as well as NRPS-like and PKS-like proteins. These are summarised in Table 6.9 along with other PmedOMT5 genes predicted to be involved in secondary metabolite production.

Two genes (*PmedOMT5_09049* and *PmedOMT5_00861*) were identified that had matches to genes involved in the detoxification of saponins in other species, PHI:24
avenecinase. *M. truncatula* and alfalfa are known to produce saponins (Huhman and Sumner 2002, Kapusta *et al.* 2005), a class of phytoanticipins, many of which have antifungal activity (reviewed by Morrissey and Osbourn 1999). These compounds are thought to act by interacting with sterols in the fungal membrane causing a loss of integrity, but have also been proposed to trigger PCD in fungi (Ito *et al.* 2007). A number of fungi can degrade their host’s saponins often by hydrolysis of sugar molecules from the saponin backbone. These PmedOMT5 genes contain GH3 domains, which remove single glycosyl residues from their substrates and show increased relative expression IP. PmedOMT5_09049 is predicted to be secreted by all three methods used and has orthologs in 31 species, PmedOMT5_00861 is only predicted to be secreted by one algorithm. Two other genes with GH3 domains that show increased relative expression IP, *PmedOMT5_00259* and *PmedOMT5_00259* had matches to tomatinase (*tom1*), a saponin detoxification enzyme from *F. oxysporum* f. sp *lycopersici*, which is essential for full virulence on tomato (Pareja-Jaime *et al.* 2008). These genes may play a role in virulence of PmedOMT5 by reducing the activity of plant saponins against the fungus.

*PmedOMT5_09990* has a strong match (bit score 1,889) to PHI:2309, *BcatrB*, an ABC transporter important for pathogenicity of *Botryotinia fuckeliana* on *Arabidopsis* (Stefanato *et al.* 2009). This gene shows high relative expression IP and had a match in the *P. medicaginis/Medicago* EST library. The role of *BcatrB* is efflux of the phytoalexin camalexin from fungal cells (Stefanato *et al.* 2009). *PmedOMT5_03325* also had a PHIbase match to *BcatrB*, PHI:2309, (bit score 1,081), with several BLASTP matches in NCBI characterised as a brefeldin A resistance proteins or efflux transporters in *Fusarium oxysporum* f. sp *cubense* race 1 (bit score 1947).

Other genes of interest with strong PHIbase matches expressed IP include isocitrate lyase, a protein essential for pathogenicity in *L. maculans* (Idnurm and Howlett 2002). This gene shows increased relative expression IP suggesting the importance of the glyoxylate pathway in *Phoma* pathogenicity. Other potentially important for pathogenicity predicted roles in genes with PHIbase matches expressed IP include: xylanases, pectinases and various other CAZyme activities, transportation in and out of the cell, transcription factors and the regulatory proteins *FgVELB* velvet *PmedOMT5_04941* (PHI:2427), and *LaeA* *PmedOMT5_02796* (PHI:2315).
Phytoalexins are low molecular weight antimicrobial compounds produced by plants in response to pathogen attack or wounding. A total of seven PmedOMT5 proteins had matches to a protein in PHIbase, MAK1 (PHI:112), which is able to detoxify the chickpea phytoalexins maackiain and medicarpin (Denny and VanEtten 1982). These are pterocarpan molecules produced by clover, alfalfa and chickpea which are structurally similar and are toxic to several genera of fungi that are pathogens and non-pathogens of legumes (Higgins 1972, Duczek and Higgins 1976, Denny and VanEtten 1981, Delserone et al. 1992). The ability to degrade medicarpin has been proposed as an important factor in determining pathogenicity on alfalfa (Higgins 1972). MAK1 encodes an FAD-binding-monoxygenase that converts maackiain to the non-fungitoxic 1a-hydroxy-maackiain and is encoded on a 1.6 Mb dispensable chromosome in Nectria haematococca (Covert et al. 1996, Enkerli et al. 1998). The gene MAK1, conferred reduced virulence on chickpea following targeted disruption and increased virulence following the addition of multiple copies to a less virulent N. haematococca isolate (Enkerli et al. 1998). Of the seven PmedOMT5 genes with matches to MAK1, four were expressed IP and two of those showed increased relative expression IP (Table 6.6). Several of these genes also showed expression of adjacent genes IP.

Table 6.6: Orthology and expression of genes encoding proteins with a match to PHI:112 maackiain detoxification gene in PHIbase that are expressed in planta.

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Expressed IP</th>
<th>No. of species with orthologs</th>
<th>Ortholog in Nh</th>
<th>Number of orthologs</th>
<th>Bit score of BLASTP match to PHI:112</th>
<th>Cluster of (3 or more consecutive) genes expressed IP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PmedOMT5_10173</td>
<td>E/I</td>
<td>0</td>
<td>no</td>
<td>0 53</td>
<td></td>
<td>Expressed IP in cluster on scaffold 260</td>
</tr>
<tr>
<td>PmedOMT5_10106</td>
<td>E/I</td>
<td>29</td>
<td>yes</td>
<td>47 450</td>
<td></td>
<td>Expressed IP in cluster on scaffold 256</td>
</tr>
<tr>
<td>PmedOMT5_06000</td>
<td>E</td>
<td>32</td>
<td>yes</td>
<td>68 inc. PmedOMT5_1040</td>
<td>106</td>
<td>Adjacent gene expressed IP scaffold 91, very weakly expressed cluster IP includes cutinase</td>
</tr>
<tr>
<td>PmedOMT5_05534</td>
<td>E</td>
<td>25</td>
<td>yes</td>
<td>35 291</td>
<td></td>
<td>Expressed in cluster on scaffold 81</td>
</tr>
</tbody>
</table>

E= expressed in planta, I= increased relative expression in planta, Nh=N. haematococca
Pisatin is the major phytoalexin produced by garden pea *Pisum sativum* L., (Wu and VanEtten 2004). Pisatin can be detoxified by *N. haematococca* MP VI and related species that are also pathogenic on pea, via pisatin demethylase (PDA), a cytochrome P450 monooxygenase (CYP57 family) (Reimmann and VanEtten 1994, Wasmann and VanEtten 1996, Liu et al. 2003). In the PmedOMT5 genome, 13 proteins had BLASTP matches to pisatin demethylases from various fungal species of the Dothideomycetes, eight of which were expressed IP, with six showing increased relative expression (Table 6.7).

Two of these potential phytoalexin detoxification genes *PmedOMT5_10173* (similar to *MAK1*) and *PmedOMT5_10174* (similar to *PDA*) are beside each other in a cluster of genes expressed *in planta*, yet while they both have potential homologs neither has orthologs in other species examined in this study.
Table 6.7: PmedOMT5 proteins with BLASTP matches to pisatin demethylase proteins in the NCBI protein database.

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Paralog group in PmedOMT5 or orthologs from other species</th>
<th>Expressed IP</th>
<th>Cluster of (3 or more consecutive) genes expressed IP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PmedOMT5_02903</td>
<td>PmedOMT_02903, PmedOMT5_08259, PmedOMT_08157</td>
<td>E/I</td>
<td>Y</td>
</tr>
<tr>
<td>PmedOMT5_03185</td>
<td>PmedOMT5_03185, PmedOMT5_08423, PmedOMT5_11322</td>
<td>E</td>
<td>Y</td>
</tr>
<tr>
<td>PmedOMT5_03385</td>
<td>PmedOMT5_03385, PmedOMT5_09423, PmedOMT5_05966</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_05737</td>
<td>unique</td>
<td>E/I</td>
<td>N nearby genes not consecutive show low expression IP</td>
</tr>
<tr>
<td>PmedOMT5_05966</td>
<td>PmedOMT5_03385, PmedOMT5_09423, PmedOMT5_05966</td>
<td>E</td>
<td>N nearby genes not consecutive show low expression IP</td>
</tr>
<tr>
<td>PmedOMT5_07265</td>
<td>unique</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PmedOMT5_07751</td>
<td>Orthologs in 9 species, none from pea pathogens</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>PmedOMT5_08157</td>
<td>PmedOMT_02903, PmedOMT5_08259, PmedOMT_08157</td>
<td>E/I</td>
<td>Y</td>
</tr>
<tr>
<td>PmedOMT5_08259</td>
<td>PmedOMT_02903, PmedOMT5_08259, PmedOMT_08157</td>
<td>E/I</td>
<td>Y, 8621 EST support</td>
</tr>
<tr>
<td>PmedOMT5_08423</td>
<td>PmedOMT5_03185, PmedOMT5_08423, PmedOMT5_11322</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_09423</td>
<td>PmedOMT5_03385, PmedOMT5_09423, PmedOMT5_05966</td>
<td>E/I</td>
<td>N nearby genes not consecutive low expression IP</td>
</tr>
<tr>
<td>PmedOMT5_10174</td>
<td>unique</td>
<td>E/I</td>
<td>Y highly expressed cluster with four PHI matches and EST support</td>
</tr>
<tr>
<td>PmedOMT5_11322</td>
<td>PmedOMT5_03185, PmedOMT5_08423, PmedOMT5_11322</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

E= expressed in planta, I= increased relative expression in planta
6.3.11 Expression of genes with Pfam domains known to involved in virulence in other pathogens

Homologs of genes in several families or with domains known to be associated with pathogenicity in other fungi were identified amongst the genes expressed IP. Relevant gene families which have previously been well-characterised in other pathogens are outlined below.

6.3.12 NPP1 domain proteins - may induce host cell death

NEP1, a member of what is now classified as the necrosis- and ethylene-inducing-like protein (NLP) gene family, was first identified in the culture filtrates of *F. oxysporum* that caused necrosis and ethylene production in the coca plant, *Erythoxylum coca* (Baily 1995). This family includes proteins from bacteria and fungi containing the motif ‘GHRHDWE’ (Fellbrich *et al.* 2002) and shows significant expansion in the oomycetes (up to 60 copies, Gijzen and Nurnberger 2006), where they are expressed during the transition to necrotrophy (Qutob *et al.* 2002). NLPs generally induce defence responses, such as callose apposition and accumulation of ROS and ethylene at low concentrations (Pemberton and Salmond 2004, Qutob *et al.* 2006) and at higher concentrations cause cell death in dicotyledonous plants but not monocots (Bailey 1995, Pemberton and Salmond 2004, Gijzen and Nurnberger 2006, Kamoun 2006).

Genes in this family share a conserved NPP1 domain, which was identified in three PmedOMT5 genes (*PmedOMT5_11472*, *PmedOMT5_00836* and *PmedOMT5_08503*), all of which were predicted to be secreted by all methods. *PmedOMT5_11472* and *PmedOMT5_00836* both showed increased relative expression IP, but neither was detected in the extracellular protein fraction. *PmedOMT5_11472* (effector score 30) has strong similarity to the NEP1 precursor, whilst *PmedOMT5_00836* (effector score 21) is more similar to NEP2 [PHI:658].

6.3.13 LysM domain proteins - may play a role in binding extracellular chitin

The Lysine motif (LysM) is a globular domain of approximately 40 amino acids [CDD:cd00118] found in bacteria and eukaryotes. It plays a role in PAMP-triggered immunity (PTI) in plants, triggering a defence response that includes an oxidative burst after binding pathogen cell wall components, chitin from fungi and peptidoglycans from
bacteria (Knogge and Scheel 2006). Several LysM receptor kinases that bind chitin and trigger host immunity have been identified in plants including CERK1 in Arabidopsis (Wan et al. 2008) and CEBiP in rice (Kaku et al. 2006). They also play a role in nod factor binding and recognition of rhizobial legume symbionts (Radutoiu et al. 2003).

Ecp6, an extracellular fungal protein with three LysM domains, first identified in the apoplastic space of tomatoes infected with Passalora fulvum (Bolton et al. 2008) competes with the plant chitin binding receptors for chitin oligosaccharides preventing the activation of PTI and is required for complete virulence of P. fulvum (Bolton et al. 2008, de Jonge et al. 2010). In the rice pathogen M. oryzae, secretion of an Ecp6 ortholog called secreted LysM protein 1 (Slp1) has been shown to accumulate between the fungal cell wall and the rice plasma membrane (Mentlak et al. 2012).

Seven PmedOMT5 genes contain LysM (chitin binding) domains. Two of these were predicted to be secreted by all of the algorithms used, but none were detected in the in vitro extracellular protein fraction. Three were expressed in planta, one of which (PmedOMT5_11730) showed a more than 4-fold relative increase in expression IP (FPKM: 4d-1, 16d-2, F2NY-40, IP-1,358). The two LysM-domain-encoding genes predicted to be secreted (PmedOMT5_11730-effector score 29 and PmedOMT5_03744-effector score 15), were both detected as expressed IP and have BLASTP matches to the Ecp6 protein from Passalora fulva although PmedOMT5_03744 has a lower match and contains only 1 LysM domain. PmedOMT5_11730, encodes a 202 aa mature protein (after cleavage of the signal peptide from the 216 aa pre-protein), containing nine cysteines and three LysM domains, and has a BLASTP match to PHI:2404 (SLP1 from M. grisea, bit score 89.4). There were 16 orthologs to PmedOMT5_11730, 15 in the Ascomycetes and 1 in Phytophthora ramorum. Passalora fulvum was not included in the orthology study but PmedOMT5_11730 was the reciprocal best hit for the Ecp6 protein [GeneBank: ACF19427.1] by BLASTP.

6.3.14 CFEM domain proteins

PmedOMT5_10302 (effector score 28) and PmedOMT5_01824 (effector score 25) are expressed IP and encode small and predicted-to-be-secreted proteins that contain 10 and 8 cysteine residues respectively. Both are predicted to contain a fungal-specific, cysteine-rich domain known as ‘common in several fungal extracellular membrane proteins’ or CFEM [Pfam:PF05730], that is predicted to be localised to extracellular
membrane proteins and proposed to be associated with pathogenesis (Kulkarni et al. 2003). Fourteen PmedOMT5 genes contain this domain, 8 of which are expressed IP with three showing increased relative expression. Of those 14, eight have matches to PHI:404, a M. oryzae gene, Pth11, disruption of which reduces the ability to form appressoria in mutants (DeZwaan et al. 1999).

6.3.15 CAP domain proteins

Five small, predicted secreted PmedOMT5 proteins contain CAP domains [Pfam:PF00188 named for ‘cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins’ (CAP)]. The gene encoding one of these PmedOMT5_5918 (effector score 26), is very highly expressed IP (FPKM: 3,072). CAP domains are found in a large family of secreted cysteine-rich proteins (SCP superfamily) from eukaryotes and prokaryotes including plant pathogenesis-related proteins (e.g. PR1). They have proposed roles as as proteases or protease inhibitors, with potential effects on signalling processes including the regulation of the extracellular matrix, sterol binding and branching morphogenesis. Four of these genes have matches to a PHIbase entry PHI:184, a Candida albicans gene, TUP1. TUP1, encodes a protein which regulates transcription of genes involved in filamentous fungal growth whose disruption conferred reduced virulence on animal hosts (Braun et al. 2000). These proteins have been proposed as fungal virulence factors with an expansion seen in pathogenic fungi that may have arisen via an HGT event from plants (reviewed in Schneiter and Di Pietro 2013). The encoded proteins have a small, dense structure stabilised by hydrophobic interactions, multiple hydrogen bonds and two conserved disulphide bonds proposed to provide high thermal, pH, and proteolytic stability in the extracellular environment (Schneiter and Di Pietro 2013).

6.3.16 Secondary metabolite production genes

Many fungi produce secondary metabolites of varying functions, some of which act as host-specific effectors, such as T-toxin, produced by C. heterostrophus race T (Baker et al. 2006). These are often synthesised by small clusters of genes that contain a large multi-domain enzyme or ‘backbone’ gene. The most common classes of these are polyketide synthases (PKS), nonribosomal peptide synthases (NRPS), hybrid forms of
these two (PKS-NRPS), dimethylallyltranstransferases (DMAT) and terpene cyclases (terpene synthases) (Khaldi et al. 2010).

Analysis with SMURF identified 35 ‘backbone’ genes involved in the production of secondary metabolites (summarised in Table 6.8). Of these 35, 14 were expressed IP, with four PKS genes and one NRPS showing a more than a 4–fold increase in expression IP relative to the other conditions tested (Table 6.9). Twenty-six of the secondary metabolite synthesis-related proteins had matches to proteins, the mutation of which resulted in reduced virulence or loss of pathogenicity phenotypes, listed in PHIbase and 10 of the 26 were expressed IP. Two secondary metabolite production proteins were detected by peptide matches in the intracellular fraction. These genes, their annotated function, expression IP and PHIbase match are listed in Table 6.9. For most of the secondary metabolite production genes that were expressed IP (all except \emph{PmedOMT5_09890}), the surrounding genes were also expressed, indicating small gene clusters that may be co-regulated during infection.

Table 6.8: Count of genes in PmedOMT5 assembly associated with secondary metabolite production.

<table>
<thead>
<tr>
<th>Secondary metabolite production genes</th>
<th>Number in PmedOMT5 assembly</th>
</tr>
</thead>
<tbody>
<tr>
<td>DMAT</td>
<td>1</td>
</tr>
<tr>
<td>HYBRID PKS/NRPS</td>
<td>1</td>
</tr>
<tr>
<td>NRPS</td>
<td>7</td>
</tr>
<tr>
<td>NRPS-Like^*</td>
<td>5</td>
</tr>
<tr>
<td>PKS</td>
<td>19</td>
</tr>
<tr>
<td>PKS-Like^*</td>
<td>2</td>
</tr>
</tbody>
</table>

^*NRPS-like and PKS-like genes contain at least two domains of the specified enzyme module but score below a trusted HMM cut-off (Khaldi, 2004).
Table 6.9: All secondary metabolite production genes identified in PmedOMT5 and their top match in PHIbase (table continues overleaf).

<table>
<thead>
<tr>
<th>Gene ID</th>
<th>Annotated gene function</th>
<th>SMURF backbone gene</th>
<th>PHI ID</th>
<th>Bit score</th>
<th>Pathogen Species (PHIbase)</th>
<th>Phenotype of mutant</th>
<th>PHI accession</th>
<th>Expressed IP</th>
<th>Increased expression IP</th>
</tr>
</thead>
<tbody>
<tr>
<td>PmedOMT5_01171</td>
<td>melanin biosynthesis</td>
<td>PKS</td>
<td>PHI:40</td>
<td>1896</td>
<td>Colletotrichum lagenarium</td>
<td>RV</td>
<td>BAA18956</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_01242</td>
<td>beta-ketoacyl synthase</td>
<td>PKS-Like</td>
<td>PHI:101</td>
<td>79</td>
<td>Aspergillus fumigatus</td>
<td>RV</td>
<td>AAC39471</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_01372</td>
<td>nrps-like enzyme</td>
<td>NRPS-Like</td>
<td>-</td>
<td>-</td>
<td>Candida albicans</td>
<td>LP</td>
<td>AAA34345</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_01385</td>
<td>fatty acid synthase subunit alpha reductase</td>
<td>PKS-Like</td>
<td>PHI:96</td>
<td>1061</td>
<td>Magnaporthe grisea</td>
<td>EF</td>
<td>B1GVX7</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_05119</td>
<td>nonribosomal peptide l-aminoadipate-semialdehyde dehydrogenase</td>
<td>NRPS</td>
<td>PHI:12</td>
<td>1547</td>
<td>Cochliobolus carbonum</td>
<td>LP</td>
<td>AAA33023</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_05427</td>
<td>dimethylallyl tryptophan synthase</td>
<td>DMAT</td>
<td>-</td>
<td>-</td>
<td>Cochliobolus sativus</td>
<td>RV</td>
<td>G8DNS9</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_06320</td>
<td>nonribosomal peptide synthetase 6</td>
<td>NRPS</td>
<td>PHI:2261</td>
<td>2799</td>
<td>Cochliobolus sativus</td>
<td>E/I</td>
<td>G8DNS9</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_06401</td>
<td>polyketide synthase</td>
<td>PKS</td>
<td>PHI:2425</td>
<td>956</td>
<td>Cochliobolus heterostrophus</td>
<td>RV</td>
<td>Q6RKG2</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_07598</td>
<td>6-methylsalicylic acid synthase</td>
<td>PKS</td>
<td>PHI:55</td>
<td>396</td>
<td>Cochliobolus heterostrophus</td>
<td>RV</td>
<td>AAB08104</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_07774</td>
<td>destruxin synthetase</td>
<td>NRPS</td>
<td>PHI:12</td>
<td>1529</td>
<td>Cochliobolus carbonum</td>
<td>LP</td>
<td>AAA33023</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_08384</td>
<td>nonribosomal siderophore peptide synthetase</td>
<td>NRPS</td>
<td>PHI:160</td>
<td>1816</td>
<td>Alternaria alternata</td>
<td>LP</td>
<td>AAF01762</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_08495</td>
<td>polyketide synthase</td>
<td>PKS</td>
<td>PHI:416</td>
<td>62.4</td>
<td>Cochliobolus heterostrophus</td>
<td>RV</td>
<td>AAX09988</td>
<td>-</td>
<td></td>
</tr>
<tr>
<td>Gene ID</td>
<td>Annotated gene function</td>
<td>SMURF backbone gene</td>
<td>PHI ID</td>
<td>Bit score</td>
<td>Pathogen Species (PHIbase)</td>
<td>Phenotype of mutant*</td>
<td>PHI accession</td>
<td>Expressed IP/ Increased expression IP</td>
<td></td>
</tr>
<tr>
<td>------------------</td>
<td>-------------------------------------------</td>
<td>---------------------</td>
<td>--------------</td>
<td>-----------</td>
<td>--------------------------------------------</td>
<td>----------------------</td>
<td>--------------</td>
<td>-------------------------------------</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_09358</td>
<td>polyketide synthase</td>
<td>PKS</td>
<td>PHI:2379</td>
<td>841</td>
<td>* Alternaria brassicicola</td>
<td>RV</td>
<td>D2E9X0</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_09406</td>
<td>polyketide synthase</td>
<td>PKS</td>
<td>PHI:55</td>
<td>768</td>
<td>* Cochliobolus heterostrophus</td>
<td>RV</td>
<td>AAB08104</td>
<td>E/I</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_09636</td>
<td>polyketide synthase</td>
<td>PKS</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>RV</td>
<td>AAB08104</td>
<td>E/I</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_09639</td>
<td>l-aminoadipate-semialdehyde dehydrogenase</td>
<td>NRPS-Like</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_09808</td>
<td>polyketide</td>
<td>PKS</td>
<td>PHI:2379</td>
<td>875</td>
<td>* Alternaria brassicicola</td>
<td>RV</td>
<td>D2E9X0</td>
<td>E/I</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_09812</td>
<td>polyketide synthase</td>
<td>PKS</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_09896</td>
<td>nonribosomal peptide synthetase 10</td>
<td>PKS</td>
<td>PHI:2290</td>
<td>1092</td>
<td>* Botrytis cinerea</td>
<td>RV</td>
<td>B1GVX7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_10258</td>
<td>nonribosomal peptide synthetase 10</td>
<td>PKS</td>
<td>PHI:2511</td>
<td>1355</td>
<td>* Aspergillus fumigatus</td>
<td>RV</td>
<td>Q4WT66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_10406</td>
<td>polyketide synthase</td>
<td>PKS</td>
<td>PHI:55</td>
<td>1278</td>
<td>* Cochliobolus heterostrophus</td>
<td>RV</td>
<td>AA08104</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_10407</td>
<td>polyketide synthase</td>
<td>PKS</td>
<td>PHI:40</td>
<td>399</td>
<td>* Colletotrichum lagenarium</td>
<td>RV</td>
<td>BAA18956</td>
<td>E/I</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_10680</td>
<td>nonribosomal peptide synthetase 10</td>
<td>NRPS-Like</td>
<td>PHI:2259</td>
<td>282</td>
<td>* Cochliobolus sativus</td>
<td>RV</td>
<td>G8DNT1</td>
<td>E</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_10816</td>
<td>non-ribosomal peptide synthetase</td>
<td>PKS</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_11227</td>
<td>non-ribosomal peptide synthetase</td>
<td>NRPS-Like</td>
<td>PHI:2511</td>
<td>1155</td>
<td>* Aspergillus fumigatus</td>
<td>RV</td>
<td>Q4WT66</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_11261</td>
<td>polyketide synthase</td>
<td>PKS</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_11296</td>
<td>polyketide synthase</td>
<td>PKS</td>
<td>PHI:2290</td>
<td>1595</td>
<td>* Botrytis cinerea</td>
<td>RV</td>
<td>B1GVX7</td>
<td>E/I</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_11355</td>
<td>polyketide synthase</td>
<td>PKS</td>
<td>-</td>
<td>-</td>
<td>*</td>
<td>-</td>
<td>-</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_11384</td>
<td>polyketide</td>
<td>PKS</td>
<td>PHI:2379</td>
<td>1278</td>
<td>* Alternaria brassicicola</td>
<td>RV</td>
<td>D2E9X0</td>
<td>E/I</td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_11453</td>
<td>mycocerosic acid synthase</td>
<td>PKS</td>
<td>PHI:2379</td>
<td>881</td>
<td>* Alternaria brassicicola</td>
<td>RV</td>
<td>D2E9X0</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* LP= loss of pathogenicity, RV = reduced virulence, EF = effector (avirulence), E= expressed in planta, I= increased relative expression in planta
6.3.17 Over representation of functional annotations in genes detected \textit{in planta} (IP)

In order to provide an insight into the types of genes that were expressed IP, domains that were over- or under-represented in the genes detected as expressed IP or with a more than 4-fold increase in expression versus the other conditions tested, were analysed relative to all genes and compared to randomly sampled pools of the same size. Amongst those genes that were expressed $\geq$ 4-fold IP versus the other conditions tested, no CAZymes were identified as over-represented at $p \leq 0.01$.

Only five Pfam domains were significantly over represented in the IP expressed genes (Table 6.10, under-represented genes are listed in appendix 6.2). These include fungal specific transcription factors, zinc finger domains (Zn(2)-Cys(6)), tyrosine kinases, mitochondrial Rho proteins and GTP-binding proteins with a range of roles including control of vesicle transport from the cell. Twelve of the genes with GTP-binding domains have PHIbase matches, with nine of these to PHI:339 a Rab/GTPase from \textit{Colletotrichum lindemuthianum} important for virulence on bean (Siriputthaiwan \textit{et al.} 2005) and the remaining three (two to PHI:281, one to PHI:182), to genes with important roles in the virulence of fungal pathogens of animals related to growth within their hosts.

Of the proteins with tyrosine kinase domains, 72 have matches to genes in PHIbase that cause either loss of pathogenicity or reduced virulence when mutated. One of these \textit{PmedOMT5_00695} hast a strong match to PHI:448, a MAP kinase from \textit{P. nodorum} that caused a loss of pathogenicity following transposon mutagenesis (Solomon \textit{et al.} 2005).
Table 6.10: Pfam domains over-represented in genes detected as expressed *in planta* relative to all genes.

<table>
<thead>
<tr>
<th>Pfam domain name</th>
<th>Pfam</th>
<th>Role</th>
<th>Number of Pfam domains in proteins with genes expressed IP by 5 or more reads (5,778)</th>
<th>Number of Pfam domains in all proteins (11,879)</th>
<th>p value</th>
<th>Number that would be likely to be detected by chance in a random sample of this size (5,778)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Fungal_trans</strong></td>
<td>PF04082.13</td>
<td>Fungal specific transcription factor</td>
<td>89</td>
<td>126</td>
<td>1.55E-03</td>
<td>61</td>
</tr>
<tr>
<td><strong>GTP_EFTU</strong></td>
<td>PF00009.22</td>
<td>GTP binding and signalling</td>
<td>28</td>
<td>32</td>
<td>8.33E-03</td>
<td>16</td>
</tr>
<tr>
<td><strong>Miro</strong></td>
<td>PF08477.8</td>
<td>Mitochondrial Rho proteins</td>
<td>32</td>
<td>39</td>
<td>9.10E-03</td>
<td>19</td>
</tr>
<tr>
<td><strong>Pkinase_Tyr</strong></td>
<td>PF07714.12</td>
<td>Tyrosine kinase-transfers phosphate from ATP to proteins</td>
<td>80</td>
<td>2.61E-03</td>
<td>56</td>
<td></td>
</tr>
<tr>
<td><strong>Zn_clus</strong></td>
<td>PF00172.13</td>
<td>Zinc finger proteins</td>
<td>114</td>
<td>187</td>
<td>7.89E-03</td>
<td>91</td>
</tr>
</tbody>
</table>
In the examination of genes which had a more than 4-fold increase in relative expression IP (Table 6.11, genes under-represented in appendix 6.3), a range of Pfam domains with roles related to oxidoreduction, hydroxylation, and FAD and NAD binding were identified. These domains were observed in a range of proteins with putative roles including maackiain detoxification, cytochrome p450s and mannitol dehydrogenase.

There are seven genes in the PmedOMT5 genome containing Pfam domain PF00296.15, which is listed as a Bac-luciferase-like flavin monooxygenase that catalyses the oxidation of long-chain aldehydes and releases energy in the form of visible light, and which uses flavin as a substrate rather than a cofactor, although the Interpro description of the domain IPR011251 (found in four of these genes) notes that
there are structural similarities between the bacterial luciferase and non-fluorescent flavoproteins such as alkane sulfonate monoxygenase SsuD (Eichhorn et al. 1999). Four of these genes show increased expression in planta and all have strong homology to fungal genes in the NCBI protein database characterised as methanesulfonate monoxygenases.

There was also a considerable over-representation of proteases and membrane transporter domains, including MFS (Major Facilitator Superfamily) and sugar transporters, which are presumably involved in break-down and subsequent uptake of plant components. Although ABC transporters may also play roles in efflux of plant defence compounds or fungal secondary metabolites from the cell, MFS transporters, have been proposed to only be capable of transporting small solutes in response to chemiosmotic ion gradients (Pao et al. 1998)

A similar analysis was carried out examining over-represented GO terms in genes detected as expressed IP versus the other conditions tested (Table 6.12, appendix 6.4). The majority of IP-over–represented GO terms appear to be related to breakdown of proteins and nucleotides, transcription and growth, suggesting substrate degradation by the pathogen and advancement within the host. There was also an increase in genes expressed IP with roles in vesicle-mediated transport and ubiquitin-mediated catabolism. Furthermore, for IP-expressed genes that also had more than a 4-fold relative increase in expression IP (Table 6.13, appendix 6.5), there was an increase in expression of genes encoding proteins with oxidoreductase activities and those involved in transmembrane transport.
Table 6.12: Gene Ontologies for all proteins over-represented in the expressed *in planta* (by five or more reads) set. (Table continues overleaf).

<table>
<thead>
<tr>
<th>GO number</th>
<th>GO term (expressed IP 5 or more reads)</th>
<th>GO Process*</th>
<th>GO count IP expression (5,778)</th>
<th>GOs in all genes (11,979)</th>
<th>p value</th>
<th>Chance #</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0000166</td>
<td>nucleotide binding</td>
<td>F</td>
<td>253</td>
<td>412</td>
<td>4.25E-04</td>
<td>200.40</td>
</tr>
<tr>
<td>GO:0000981</td>
<td>sequence-specific DNA binding RNA polymerase II transcription factor activity</td>
<td>F</td>
<td>104</td>
<td>167</td>
<td>7.17E-03</td>
<td>81.23</td>
</tr>
<tr>
<td>GO:0003677</td>
<td>DNA binding</td>
<td>F</td>
<td>196</td>
<td>309</td>
<td>5.23E-04</td>
<td>150.30</td>
</tr>
<tr>
<td>GO:0003700</td>
<td>sequence-specific DNA binding transcription factor activity</td>
<td>F</td>
<td>43</td>
<td>55</td>
<td>5.70E-03</td>
<td>26.75</td>
</tr>
<tr>
<td>GO:0003723</td>
<td>RNA binding</td>
<td>F</td>
<td>88</td>
<td>115</td>
<td>3.69E-04</td>
<td>55.94</td>
</tr>
<tr>
<td>GO:0003735</td>
<td>structural constituent of ribosome</td>
<td>F</td>
<td>104</td>
<td>126</td>
<td>2.06E-05</td>
<td>61.29</td>
</tr>
<tr>
<td>GO:0003743</td>
<td>translation initiation factor activity</td>
<td>F</td>
<td>42</td>
<td>49</td>
<td>2.41E-03</td>
<td>23.83</td>
</tr>
<tr>
<td>GO:0003924</td>
<td>GTPase activity</td>
<td>F</td>
<td>60</td>
<td>75</td>
<td>1.21E-03</td>
<td>36.48</td>
</tr>
<tr>
<td>GO:0004674</td>
<td>protein serine/threonine kinase activity</td>
<td>F</td>
<td>67</td>
<td>90</td>
<td>2.12E-03</td>
<td>43.78</td>
</tr>
<tr>
<td>GO:0004713</td>
<td>protein tyrosine kinase activity</td>
<td>F</td>
<td>51</td>
<td>64</td>
<td>2.54E-03</td>
<td>31.13</td>
</tr>
<tr>
<td>GO:0005488</td>
<td>Binding</td>
<td>F</td>
<td>87</td>
<td>134</td>
<td>6.27E-03</td>
<td>65.18</td>
</tr>
<tr>
<td>GO:0005515</td>
<td>protein binding</td>
<td>F</td>
<td>391</td>
<td>583</td>
<td>1.01E-07</td>
<td>283.57</td>
</tr>
<tr>
<td>GO:0005524</td>
<td>ATP binding</td>
<td>F</td>
<td>409</td>
<td>574</td>
<td>4.12E-10</td>
<td>279.20</td>
</tr>
<tr>
<td>GO:0005525</td>
<td>GTP binding</td>
<td>F</td>
<td>88</td>
<td>110</td>
<td>1.49E-04</td>
<td>53.50</td>
</tr>
<tr>
<td>GO:0005622</td>
<td>Intracellular</td>
<td>C</td>
<td>128</td>
<td>187</td>
<td>5.65E-04</td>
<td>90.96</td>
</tr>
<tr>
<td>GO:0005634</td>
<td>Nucleus</td>
<td>C</td>
<td>297</td>
<td>446</td>
<td>3.88E-06</td>
<td>216.94</td>
</tr>
<tr>
<td>GO:0005667</td>
<td>transcription factor complex</td>
<td>C</td>
<td>50</td>
<td>64</td>
<td>3.45E-03</td>
<td>31.13</td>
</tr>
<tr>
<td>GO:0005737</td>
<td>Cytoplasm</td>
<td>C</td>
<td>193</td>
<td>254</td>
<td>7.15E-07</td>
<td>123.55</td>
</tr>
<tr>
<td>GO:0005739</td>
<td>Mitochondrion</td>
<td>C</td>
<td>89</td>
<td>99</td>
<td>9.64E-06</td>
<td>48.15</td>
</tr>
<tr>
<td>GO:0005840</td>
<td>Ribosome</td>
<td>C</td>
<td>150</td>
<td>173</td>
<td>7.37E-08</td>
<td>84.15</td>
</tr>
<tr>
<td>GO:0006094</td>
<td>Gluconeogenesis</td>
<td>P</td>
<td>29</td>
<td>33</td>
<td>7.24E-03</td>
<td>16.05</td>
</tr>
<tr>
<td>GO:0006119</td>
<td>oxidative phosphorylation</td>
<td>P</td>
<td>26</td>
<td>27</td>
<td>5.29E-03</td>
<td>13.13</td>
</tr>
<tr>
<td>GO:0006144</td>
<td>purine nucleobase metabolic process</td>
<td>P</td>
<td>53</td>
<td>72</td>
<td>5.55E-03</td>
<td>35.02</td>
</tr>
<tr>
<td>GO number</td>
<td>GO term (expressed IP 5 or more reads)</td>
<td>GO Process*</td>
<td>GO count IP expression (5,778)</td>
<td>GOs in all genes (11,979)</td>
<td>p value</td>
<td>Chance*</td>
</tr>
<tr>
<td>---------------</td>
<td>-------------------------------------------------------------------------------------------------------</td>
<td>-------------</td>
<td>--------------------------------</td>
<td>----------------------------</td>
<td>-------------</td>
<td>-----------</td>
</tr>
<tr>
<td>GO:0006184</td>
<td>GTP catabolic process</td>
<td>P</td>
<td>56</td>
<td>66</td>
<td>7.33E-04</td>
<td>32.10</td>
</tr>
<tr>
<td>GO:0006200</td>
<td>ATP catabolic process</td>
<td>P</td>
<td>47</td>
<td>56</td>
<td>1.91E-03</td>
<td>27.24</td>
</tr>
<tr>
<td>GO:0006355</td>
<td>regulation of transcription, DNA-dependent</td>
<td>P</td>
<td>104</td>
<td>144</td>
<td>4.90E-04</td>
<td>70.04</td>
</tr>
<tr>
<td>GO:0006357</td>
<td>regulation of transcription from RNA polymerase II promoter</td>
<td>P</td>
<td>57</td>
<td>69</td>
<td>9.81E-04</td>
<td>33.56</td>
</tr>
<tr>
<td>GO:0006412</td>
<td>Translation</td>
<td>P</td>
<td>98</td>
<td>122</td>
<td>6.46E-05</td>
<td>59.34</td>
</tr>
<tr>
<td>GO:0006446</td>
<td>regulation of translational initiation</td>
<td>P</td>
<td>44</td>
<td>51</td>
<td>1.84E-03</td>
<td>24.81</td>
</tr>
<tr>
<td>GO:0006457</td>
<td>protein phosphorylation</td>
<td>P</td>
<td>42</td>
<td>56</td>
<td>9.02E-03</td>
<td>27.24</td>
</tr>
<tr>
<td>GO:0006511</td>
<td>ubiquitin-dependent protein catabolic process</td>
<td>P</td>
<td>30</td>
<td>35</td>
<td>7.87E-03</td>
<td>17.02</td>
</tr>
<tr>
<td>GO:0006886</td>
<td>intracellular protein transport</td>
<td>P</td>
<td>70</td>
<td>74</td>
<td>2.64E-05</td>
<td>35.99</td>
</tr>
<tr>
<td>GO:0007264</td>
<td>small GTPase mediated signal transduction</td>
<td>P</td>
<td>36</td>
<td>43</td>
<td>5.35E-03</td>
<td>20.92</td>
</tr>
<tr>
<td>GO:0008270</td>
<td>zinc ion binding</td>
<td>F</td>
<td>358</td>
<td>568</td>
<td>1.36E-05</td>
<td>276.28</td>
</tr>
<tr>
<td>GO:0008565</td>
<td>protein transporter activity</td>
<td>F</td>
<td>23</td>
<td>23</td>
<td>6.24E-03</td>
<td>11.19</td>
</tr>
<tr>
<td>GO:0009069</td>
<td>serine family amino acid metabolic process</td>
<td>P</td>
<td>75</td>
<td>98</td>
<td>8.45E-04</td>
<td>47.67</td>
</tr>
<tr>
<td>GO:0015031</td>
<td>protein transport</td>
<td>P</td>
<td>42</td>
<td>51</td>
<td>3.66E-03</td>
<td>24.81</td>
</tr>
<tr>
<td>GO:0016020</td>
<td>Membrane</td>
<td>C</td>
<td>187</td>
<td>288</td>
<td>3.28E-04</td>
<td>140.08</td>
</tr>
<tr>
<td>GO:0016021</td>
<td>integral to membrane</td>
<td>C</td>
<td>405</td>
<td>735</td>
<td>2.99E-03</td>
<td>357.51</td>
</tr>
<tr>
<td>GO:0016192</td>
<td>vesicle-mediated transport</td>
<td>P</td>
<td>36</td>
<td>40</td>
<td>2.73E-03</td>
<td>19.46</td>
</tr>
<tr>
<td>GO:0016310</td>
<td>Phosphorylation</td>
<td>P</td>
<td>60</td>
<td>80</td>
<td>2.92E-03</td>
<td>38.91</td>
</tr>
<tr>
<td>GO:0042254</td>
<td>ribosome biogenesis</td>
<td>P</td>
<td>116</td>
<td>144</td>
<td>1.56E-05</td>
<td>70.04</td>
</tr>
<tr>
<td>GO:0043565</td>
<td>sequence-specific DNA binding</td>
<td>F</td>
<td>43</td>
<td>57</td>
<td>8.01E-03</td>
<td>27.73</td>
</tr>
<tr>
<td>GO:0045449</td>
<td>Regulation of gene-specific transcription</td>
<td>P</td>
<td>65</td>
<td>87</td>
<td>2.27E-03</td>
<td>42.32</td>
</tr>
<tr>
<td>GO:0046486</td>
<td>glycerolipid metabolic process</td>
<td>P</td>
<td>36</td>
<td>44</td>
<td>6.55E-03</td>
<td>21.40</td>
</tr>
<tr>
<td>GO:0046872</td>
<td>metal ion binding</td>
<td>F</td>
<td>130</td>
<td>184</td>
<td>2.38E-04</td>
<td>89.50</td>
</tr>
<tr>
<td>GO:0051082</td>
<td>unfolded protein binding</td>
<td>F</td>
<td>35</td>
<td>42</td>
<td>6.11E-03</td>
<td>20.43</td>
</tr>
<tr>
<td>GO:0055114</td>
<td>oxidation-reduction process</td>
<td>P</td>
<td>468</td>
<td>778</td>
<td>2.27E-05</td>
<td>378.42</td>
</tr>
</tbody>
</table>

* F=Molecular Function; P=Biological Process; C=Cellular Component # Number that would be likely to be detected by chance in a random sample of this size (5,778)
Table 6.13: Gene Ontologies over-represented in proteins encoded by genes that show greater than 4-fold increased relative expression in planta (Table continues overleaf).

<table>
<thead>
<tr>
<th>GO number</th>
<th>GO term</th>
<th>GO Process</th>
<th>Number of proteins with GO term encoded by gene with more than 4-fold increase in relative expression IP (461)</th>
<th>Number of GO terms in all proteins (11,879)</th>
<th>p value</th>
<th>Number that would be likely to be detected by chance in a random sample of this size (461)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0004181</td>
<td>metallocarboxypeptidase activity</td>
<td>F</td>
<td>4</td>
<td>10</td>
<td>1.32E-03</td>
<td>0.39</td>
</tr>
<tr>
<td>GO:0004497</td>
<td>monooxygenase activity</td>
<td>F</td>
<td>14</td>
<td>74</td>
<td>3.53E-06</td>
<td>2.87</td>
</tr>
<tr>
<td>GO:0004553</td>
<td>hydrolase activity, hydrolyzing O-glycosyl compounds</td>
<td>F</td>
<td>10</td>
<td>92</td>
<td>3.29E-03</td>
<td>3.57</td>
</tr>
<tr>
<td>GO:0005215</td>
<td>transporter activity</td>
<td>F</td>
<td>9</td>
<td>53</td>
<td>3.67E-04</td>
<td>2.06</td>
</tr>
<tr>
<td>GO:0005315</td>
<td>inorganic phosphate transmembrane transporter activity</td>
<td>F</td>
<td>4</td>
<td>8</td>
<td>7.04E-04</td>
<td>0.31</td>
</tr>
<tr>
<td>GO:0005506</td>
<td>iron ion binding</td>
<td>F</td>
<td>27</td>
<td>161</td>
<td>1.31E-09</td>
<td>6.25</td>
</tr>
<tr>
<td>GO:0005618</td>
<td>cell wall</td>
<td>C</td>
<td>4</td>
<td>11</td>
<td>1.73E-03</td>
<td>0.43</td>
</tr>
<tr>
<td>GO:0006040</td>
<td>amino sugar metabolic process</td>
<td>P</td>
<td>7</td>
<td>51</td>
<td>4.30E-03</td>
<td>1.98</td>
</tr>
<tr>
<td>GO:0006118</td>
<td>electron transport</td>
<td>P</td>
<td>22</td>
<td>196</td>
<td>1.53E-05</td>
<td>7.61</td>
</tr>
<tr>
<td>GO:0006508</td>
<td>proteolysis</td>
<td>P</td>
<td>15</td>
<td>175</td>
<td>3.06E-03</td>
<td>6.79</td>
</tr>
<tr>
<td>GO:0006807</td>
<td>nitrogen compound metabolic process</td>
<td>P</td>
<td>6</td>
<td>46</td>
<td>9.49E-03</td>
<td>1.79</td>
</tr>
<tr>
<td>GO:0006810</td>
<td>transport</td>
<td>P</td>
<td>9</td>
<td>92</td>
<td>8.72E-03</td>
<td>3.57</td>
</tr>
<tr>
<td>GO:0006817</td>
<td>phosphate ion transport</td>
<td>P</td>
<td>4</td>
<td>7</td>
<td>4.87E-04</td>
<td>0.27</td>
</tr>
<tr>
<td>GO:0008152</td>
<td>metabolic process</td>
<td>P</td>
<td>25</td>
<td>335</td>
<td>1.15E-03</td>
<td>13</td>
</tr>
<tr>
<td>GO:0008762</td>
<td>UDP-N-acetylmuramate dehydrogenase activity</td>
<td>F</td>
<td>7</td>
<td>41</td>
<td>1.53E-03</td>
<td>1.59</td>
</tr>
<tr>
<td>GO:0009055</td>
<td>electron carrier activity</td>
<td>F</td>
<td>21</td>
<td>147</td>
<td>8.86E-07</td>
<td>5.7</td>
</tr>
<tr>
<td>GO:0015205</td>
<td>nucleobase transmembrane transporter activity</td>
<td>F</td>
<td>3</td>
<td>9</td>
<td>8.11E-03</td>
<td>0.35</td>
</tr>
<tr>
<td>GO:0015851</td>
<td>nucleobase transport</td>
<td>P</td>
<td>3</td>
<td>9</td>
<td>8.11E-03</td>
<td>0.35</td>
</tr>
<tr>
<td>GO:0016021</td>
<td>integral to membrane</td>
<td>C</td>
<td>44</td>
<td>735</td>
<td>1.59E-03</td>
<td>28.52</td>
</tr>
<tr>
<td>GO:0016491</td>
<td>oxidoreductase activity</td>
<td>F</td>
<td>40</td>
<td>428</td>
<td>5.35E-07</td>
<td>16.61</td>
</tr>
<tr>
<td>GO:0016702</td>
<td>oxidoreductase activity, acting on single donors with incorporation of molecular oxygen, incorporation of two atoms of oxygen</td>
<td>F</td>
<td>7</td>
<td>31</td>
<td>3.83E-04</td>
<td>1.2</td>
</tr>
<tr>
<td>GO number</td>
<td>GO term</td>
<td>GO Process</td>
<td>Number of proteins with GO term encoded by gene with more than 2-fold increase in relative expression IP (461)</td>
<td>Number of GO terms in all proteins (11,879)</td>
<td>p value</td>
<td>Number that would be likely to be detected by chance in a random sample of this size (461)</td>
</tr>
<tr>
<td>-----------</td>
<td>-------------------------------------------------------------------------</td>
<td>------------</td>
<td>----------------------------------------------------------------------------------------------------------------</td>
<td>----------------------------------------------</td>
<td>-----------</td>
<td>-------------------------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>GO:0016705 oxidoreductase activity, acting on paired donors, with incorporation or reduction of molecular oxygen</td>
<td>F</td>
<td>14</td>
<td>53</td>
<td>1.17E-07</td>
<td>2.06</td>
<td></td>
</tr>
<tr>
<td>GO:0016810 hydrolase activity, acting on carbon-nitrogen (but not peptide) bonds</td>
<td>F</td>
<td>4</td>
<td>15</td>
<td>4.23E-03</td>
<td>0.58</td>
<td></td>
</tr>
<tr>
<td>GO:0019439 aromatic compound catabolic process</td>
<td>P</td>
<td>3</td>
<td>5</td>
<td>2.40E-03</td>
<td>0.19</td>
<td></td>
</tr>
<tr>
<td>GO:0020037 heme binding</td>
<td>F</td>
<td>21</td>
<td>151</td>
<td>1.29E-06</td>
<td>5.86</td>
<td></td>
</tr>
<tr>
<td>GO:0031177 phosphopantetheine binding</td>
<td>F</td>
<td>4</td>
<td>16</td>
<td>5.09E-03</td>
<td>0.62</td>
<td></td>
</tr>
<tr>
<td>GO:0050660 flavin adenine dinucleotide binding</td>
<td>F</td>
<td>15</td>
<td>128</td>
<td>2.06E-04</td>
<td>4.97</td>
<td></td>
</tr>
<tr>
<td>GO:0055085 transmembrane transport</td>
<td>P</td>
<td>31</td>
<td>448</td>
<td>1.02E-03</td>
<td>17.39</td>
<td></td>
</tr>
<tr>
<td>GO:0055114 oxidation-reduction process</td>
<td>P</td>
<td>83</td>
<td>778</td>
<td>3.21E-16</td>
<td>30.19</td>
<td></td>
</tr>
</tbody>
</table>

* F=Molecular Function; P=Biological Process; C=Cellular Component
6.3.18 Bioinformatic prediction of candidate effector genes

Proteins were assigned a score based on various characteristics known to be associated with effector genes in the literature. Each characteristic was given a score from 1 - 8, weighted according to their perceived relevance to pathogenicity effectors. Cumulative scores, derived from the sum of matching characteristics, were used in several categories to give increased weight to proteins that were highly likely to be effectors, *i.e.* highly expressed *in planta*, had high relative cysteine counts for their size or were predicted to be secreted by more than one algorithm. Conversely all genes detected by only a few reads IP or predicted to be secreted by only one method, received a more modest number of points towards the final score. Using the cumulative scoring regime outlined in section 6.2.12 and Table 6.14, 33 candidates were identified with a score of 30 or greater, 111 scored greater than or equal to 25 and 324 scored greater than or equal to 20 (scores for all proteins are listed in Appendix 5.3).

Of those 111 genes that scored greater than 25, 107 were detected as expressed in IP (46 with increased relative expression) by mapped reads or Trinity alignments and 32 of those were also detected in the extracellular fraction. Two other genes were detected in the extracellular fraction only. The remaining two were uncharacterised (no BLAST, PFAM, PHIbase or GO matches), small cysteine-rich proteins that were also predicted to be secreted.

Eleven of the proteins in the top 111 have PHIbase matches to proteins with previously described roles in pathogenicity, including: an oxidoreductase (PmedOMT5_06753), an alcohol oxidase (PmedOMT5_06610), a cutinase (PmedOMT5_08831) and an endopolygalacturonase (PmedOMT5_03759), as well as a maackian detoxification protein (PmedOMT5_10173). There were also proteins with matches to two known non-specific effectors PmedOMT5_11730- Ecp6 (effector score of 29) and PmedOMT5_11472 - NPP1 (effector score of 30). There were 34 proteins in the top 111 that had no hit by BLASTP to the NCBI protein database, but only 15 of these had no match by TBLASTN to the genome sequence of the closely related species *Ascochyta rabiei* (R. Mohd Shah, A.H. Williams, J. K. Hane, R.P. Oliver, J. Lichtenzveig-unpublished data). There were five genes in the top 111 that were also supported by mapped Phoma-Medicago infection ESTs and 53 proteins with characterised domains, while the functions of the remainder are unknown. There were nine genes with
CAZyme annotations indicating possible roles in plant cell-wall degradation or fungal cell-wall restructuring.

The top 18 candidates with scores of 32 or more are listed in Table 6.15 with those unique to PmedOMT5 based on orthology listed in Table 6.16. In the top 18, only two genes have assigned GO or Pfam domains, the others had no evidence for characterised function. One gene *PmedOMT5_05568* has EST support. The highest score was 42 for PmedOMT5 _08005, a 79 aa protein with a [YFW]XC domain at the C terminal (73-75). This gene is most highly expressed *in planta* (FPKM- 4d: 3, 16d: 2, F2NY: 16, IP: 827) and had no orthologs in the species tested or BLASTP matches to the NCBI protein database.

Table 6.14: Scores assigned to various features potentially associated with necrotrophic effectors and counts of genes / proteins with these features.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Weighted Score</th>
<th>Number of genes that matched criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>De novo</em> assembled IP transcript</td>
<td>8</td>
<td>970</td>
</tr>
<tr>
<td>Predicted molecular weight of less than 40 kDa</td>
<td>5</td>
<td>5,510</td>
</tr>
<tr>
<td>Detected in extracellular fraction</td>
<td>7</td>
<td>170</td>
</tr>
<tr>
<td>No BLASTP match in GenBank (below e-5)</td>
<td>7</td>
<td>1,527</td>
</tr>
<tr>
<td>Greater than 4-fold increase in relative expression IP</td>
<td>4</td>
<td>461</td>
</tr>
<tr>
<td>EST match in <em>Phoma</em> IP library</td>
<td>3</td>
<td>92</td>
</tr>
<tr>
<td>Predicted to be secreted by all methods</td>
<td>3</td>
<td>1,020</td>
</tr>
<tr>
<td>Mature protein is &gt; 5% of cystines</td>
<td>3</td>
<td>91</td>
</tr>
<tr>
<td>Two or more cysteines in the mature protein</td>
<td>3</td>
<td>1,024</td>
</tr>
<tr>
<td>Greater than 50% of CDS covered by IP RNAseq reads</td>
<td>2</td>
<td>1,775</td>
</tr>
<tr>
<td>Predicted to be secreted by one or more methods</td>
<td>2</td>
<td>2,114</td>
</tr>
<tr>
<td>Known effector associated domain</td>
<td>1</td>
<td>7,013</td>
</tr>
<tr>
<td>Gene is within 2kb of a repetitive region</td>
<td>2</td>
<td>2,142</td>
</tr>
<tr>
<td>Genes with 5 or more mapped RNAseq reads <em>in planta</em></td>
<td>2</td>
<td>5,778</td>
</tr>
</tbody>
</table>
Table 6.15: The top effector candidates of *P. medicaginis* OMT5, based on cumulative scoring of associated gene/protein properties and functional annotations.

<table>
<thead>
<tr>
<th>Locus ID</th>
<th>Effector score</th>
<th>Expression</th>
<th>Peptide match from the extracellular protein fraction</th>
<th>Predicted molecular weight (kDa)</th>
<th>Mature protein length</th>
<th>Cysteine count mature protein</th>
<th>Cysteine percentage of mature protein</th>
<th>Predicted secreted by all 3 methods</th>
<th>Within 2kb of repeat</th>
<th>Motif associated with known effectors</th>
<th>EST match</th>
<th>PHIbase or Pfam match</th>
<th>TBLASTX match to A. mali scaffolds</th>
<th>BLASTP match to NCBI database bit score</th>
</tr>
</thead>
<tbody>
<tr>
<td>PmedOMT5_08005</td>
<td>42</td>
<td>E/I</td>
<td></td>
<td>8.3</td>
<td>63</td>
<td>6</td>
<td>9.5</td>
<td>y</td>
<td>y</td>
<td>[YFW]XC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_11505</td>
<td>40</td>
<td>E/I</td>
<td></td>
<td>9.8</td>
<td>74</td>
<td>6</td>
<td>8.1</td>
<td>y</td>
<td>y</td>
<td>[YFW]XC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_03386</td>
<td>39</td>
<td>E/I</td>
<td></td>
<td>9.1</td>
<td>67</td>
<td>6</td>
<td>9</td>
<td>y</td>
<td>y</td>
<td>[YFW]XC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_11465</td>
<td>39</td>
<td>E/I</td>
<td></td>
<td>15.4</td>
<td>134</td>
<td>2</td>
<td>1.5</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_10797</td>
<td>37</td>
<td>E/I</td>
<td></td>
<td>16.9</td>
<td>141</td>
<td>5</td>
<td>3.5</td>
<td>y</td>
<td>[YFW]XC</td>
<td>y</td>
<td>186</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_00021</td>
<td>36</td>
<td>E/I</td>
<td></td>
<td>22.6</td>
<td>195</td>
<td>3</td>
<td>1.5</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_03659</td>
<td>36</td>
<td>E/I</td>
<td></td>
<td>14.4</td>
<td>126</td>
<td>4</td>
<td>3.2</td>
<td>y</td>
<td>Ribonuclease</td>
<td>y</td>
<td>186</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_05568</td>
<td>36</td>
<td>E</td>
<td></td>
<td>32.8</td>
<td>291</td>
<td>11</td>
<td>3.8</td>
<td>y</td>
<td>[YFW]XC</td>
<td>y</td>
<td>186</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_10389</td>
<td>36</td>
<td>E/I</td>
<td></td>
<td>31.8</td>
<td>280</td>
<td>2</td>
<td>0.7</td>
<td>y</td>
<td>y</td>
<td>399</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_00560</td>
<td>35</td>
<td>E</td>
<td></td>
<td>20.8</td>
<td>196</td>
<td>5</td>
<td>2.6</td>
<td>y</td>
<td>y</td>
<td>[LJ]XAR</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_05310</td>
<td>34</td>
<td>E(trinity)</td>
<td></td>
<td>6.9</td>
<td>43</td>
<td>4</td>
<td>9.3</td>
<td>y</td>
<td>y</td>
<td>[YFW]XC</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_02275</td>
<td>33</td>
<td>E/I</td>
<td></td>
<td>24.4</td>
<td>222</td>
<td>0</td>
<td>0</td>
<td>y</td>
<td>y</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_10901</td>
<td>33</td>
<td>E/I</td>
<td></td>
<td>19.2</td>
<td>167</td>
<td>12</td>
<td>7.2</td>
<td>y</td>
<td>[YFW]XC</td>
<td>y</td>
<td>120</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_00533</td>
<td>32</td>
<td>E/I</td>
<td></td>
<td>18.7</td>
<td>162</td>
<td>10</td>
<td>6.2</td>
<td>y</td>
<td>[YFW]XC</td>
<td>y</td>
<td>150</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_01232</td>
<td>32</td>
<td>E/I</td>
<td></td>
<td>11.1</td>
<td>87</td>
<td>6</td>
<td>6.9</td>
<td>y</td>
<td>y</td>
<td>137</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_06753</td>
<td>32</td>
<td>E/I</td>
<td></td>
<td>57.8</td>
<td>527</td>
<td>5</td>
<td>0.9</td>
<td>y</td>
<td>[LJ]XAR, [YFW]XC</td>
<td>PHI:1046, bit score 155, Pfam:FAD_binding_4</td>
<td>y</td>
<td>576</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_07632</td>
<td>32</td>
<td>E/I</td>
<td></td>
<td>19.5</td>
<td>163</td>
<td>16</td>
<td>9.8</td>
<td>y</td>
<td>[YFW]XC</td>
<td>y</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 6.16: Top unique (species-specific by orthology) PmedOMT5 effector candidates.

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>Effector score</th>
<th>Expression</th>
<th>4d FPKM</th>
<th>16d FPKM</th>
<th>F2NY FPKM</th>
<th>IP FPKM</th>
<th>Known effector associated motif</th>
<th>aa length</th>
<th>Predicted molecular weight (kDa)</th>
<th>Predicted secreted one or more algorithms</th>
<th>gene is within 2kb of a repeat</th>
<th>Cysteines in mature protein</th>
<th>Number of cysteines in whole protein</th>
<th>greater than 5% cysteines in mature protein</th>
</tr>
</thead>
<tbody>
<tr>
<td>PmedOMT5_08005</td>
<td>42</td>
<td>E/I</td>
<td>3</td>
<td>2</td>
<td>16</td>
<td>827</td>
<td>[YFW]XC</td>
<td>79</td>
<td>8.3</td>
<td>y</td>
<td>y</td>
<td>6</td>
<td>6</td>
<td>y</td>
</tr>
<tr>
<td>PmedOMT5_11465</td>
<td>39</td>
<td>E/I</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1029</td>
<td></td>
<td>87</td>
<td>9.1</td>
<td>y</td>
<td></td>
<td>6</td>
<td>6</td>
<td>y</td>
</tr>
<tr>
<td>PmedOMT5_03386</td>
<td>39</td>
<td>E/I</td>
<td>2</td>
<td>0</td>
<td>22</td>
<td>464</td>
<td></td>
<td>56</td>
<td>6.0</td>
<td>y</td>
<td>y</td>
<td>6</td>
<td>6</td>
<td>y</td>
</tr>
<tr>
<td>PmedOMT5_05310</td>
<td>34</td>
<td>E/I</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>[YFW]XC</td>
<td>63</td>
<td>6.9</td>
<td>y</td>
<td>y</td>
<td>4</td>
<td>5</td>
<td>y</td>
</tr>
<tr>
<td>PmedOMT5_10596</td>
<td>31</td>
<td>E/I</td>
<td>0</td>
<td>2</td>
<td>5</td>
<td>283</td>
<td></td>
<td>63</td>
<td>6.3</td>
<td>y</td>
<td></td>
<td>8</td>
<td>8</td>
<td>y</td>
</tr>
<tr>
<td>PmedOMT5_09579</td>
<td>30</td>
<td>E/I</td>
<td>9</td>
<td>16</td>
<td>43</td>
<td>136</td>
<td></td>
<td>53</td>
<td>5.5</td>
<td>y</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_01684</td>
<td>29</td>
<td>E/I</td>
<td>17</td>
<td>46</td>
<td>48</td>
<td>467</td>
<td>RXLR</td>
<td>80</td>
<td>8.9</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_05659</td>
<td>28</td>
<td>E</td>
<td>16</td>
<td>16</td>
<td>20</td>
<td>35</td>
<td>[LI]XAR</td>
<td>270</td>
<td>28.2</td>
<td>y</td>
<td>y</td>
<td>18</td>
<td>18</td>
<td>y</td>
</tr>
<tr>
<td>PmedOMT5_08748</td>
<td>27</td>
<td>E</td>
<td>322</td>
<td>161</td>
<td>608</td>
<td>254</td>
<td>[LI]XAR</td>
<td>82</td>
<td>8.5</td>
<td>y</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_05864</td>
<td>27</td>
<td>E</td>
<td>37</td>
<td>194</td>
<td>443</td>
<td>54</td>
<td>CHXC,RXLR</td>
<td>113</td>
<td>12.7</td>
<td>y</td>
<td></td>
<td>8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_10533</td>
<td>27</td>
<td>E</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>422</td>
<td>[YFW]XC</td>
<td>79</td>
<td>9.5</td>
<td>y</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_10443</td>
<td>27</td>
<td>E</td>
<td>2869</td>
<td>912</td>
<td>567</td>
<td>1962</td>
<td>[LI]XAR</td>
<td>139</td>
<td>14.3</td>
<td>y</td>
<td></td>
<td>2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PmedOMT5_10027</td>
<td>25</td>
<td>E</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>0</td>
<td></td>
<td>133</td>
<td>14.8</td>
<td>y</td>
<td>y</td>
<td>6</td>
<td>6</td>
<td>y</td>
</tr>
<tr>
<td>PmedOMT5_08696</td>
<td>25</td>
<td>E</td>
<td>200</td>
<td>174</td>
<td>45</td>
<td>80</td>
<td>[YFW]XC</td>
<td>102</td>
<td>11.6</td>
<td></td>
<td></td>
<td>3</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
6.3.19 Functional annotations associated with the potential effector candidate set

The number of over and under-represented functional annotations was also calculated for the top 111 effector candidates, relative to the complete set of PmedOMT5 proteins. None of the known effector-associated domains were over represented in the top 111 candidates. The CAZyme GH72 was found to be over-represented. This family consists of β-1,3-glucanosyltransglycosylases. β-1,3-glucan is one of the major structural components of fungal cell walls, to which chitin and other polysaccharides are attached (Fontaine et al. 1997). One of the genes with a GH72 domain (PmedOMT5_01156) has several PHI matches (PHI:256, 522, 526, 1071) to a gene encoding a glycosylphosphatidylinositol-anchored surface protein (GAS1) identified in various fungal species as an important protein for the correct assembly and cross-linking of fungal cell-walls, mutation of which can affect virulence and pathogenicity in planta (Nuoffer et al. 1991, Saporito-Irwin et al. 1995, Caracuel et al. 2005). This gene has several orthologs in PmedOMT5 that were also detected as expressed IP and identified in the extracellular protein fraction, including PmedOMT5_03404.

Six GO terms were over-represented in the top effector candidates (Table 6.17, under represented candidates are listed in appendix 6.6), including those with roles in carbohydrate metabolism and fungal cell wall biosynthesis.
Table 6.17: Number of over-represented GO terms in the top 111 candidates.

<table>
<thead>
<tr>
<th>GO number</th>
<th>GO description</th>
<th>GO Ontology</th>
<th>Number of GOs in top ranked effector candidates (111)</th>
<th>Number of GOs in all proteins (11,879)</th>
<th>p value</th>
<th>Number expected to occur by chance in a sample of this size (111)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0005975</td>
<td>carbohydrate metabolic process</td>
<td>P</td>
<td>7</td>
<td>188</td>
<td>1.76E-03</td>
<td>1.76</td>
</tr>
<tr>
<td>GO:0043169</td>
<td>cation binding aspartic-type endopeptidase activity</td>
<td>F</td>
<td>6</td>
<td>91</td>
<td>2.43E-04</td>
<td>0.85</td>
</tr>
<tr>
<td>GO:0004190</td>
<td></td>
<td>F</td>
<td>3</td>
<td>24</td>
<td>1.82E-03</td>
<td>0.22</td>
</tr>
<tr>
<td>GO:0008762</td>
<td>UDP-N-acetylmuramate dehydrogenase activity</td>
<td>F</td>
<td>3</td>
<td>41</td>
<td>7.05E-03</td>
<td>0.38</td>
</tr>
<tr>
<td>GO:0006662</td>
<td>glycerol ether metabolic process</td>
<td>P</td>
<td>2</td>
<td>9</td>
<td>4.30E-03</td>
<td>0.08</td>
</tr>
<tr>
<td>GO:0009277</td>
<td>fungal-type cell wall</td>
<td>C</td>
<td>2</td>
<td>2</td>
<td>5.00E-04</td>
<td>0.02</td>
</tr>
</tbody>
</table>

There were six Pfam domains over-represented in the top 111 candidates (Table 6.17, appendix 6.7) including the GH72 and GH2 CAZyme domains discussed previously (section 6.3.19), that were found in proteins with putative roles in assembly and cross-linking of fungal cell walls, as well as the CFEM extracellular fungal membrane domain proposed to be associated with pathogenesis (Kulkarni et al. 2003). Two proteins contain thioredoxin domains, these are found in redox proteins that can act as antioxidants, one of these, PmedOMT5_04414 was also identified in the extracellular protein fraction and its encoding gene had a match in the Phoma/Medicago EST library.
Table 6.18: Number of over-represented Pfam domains in the top 111 candidates.

<table>
<thead>
<tr>
<th>Pfam</th>
<th>Pfam domain name</th>
<th>Role</th>
<th>Number of genes with Pfam in top candidates (111)</th>
<th>Number of occurrences of Pfam domain in all proteins (11,879)</th>
<th>p value</th>
<th>Number expected to occur by chance in a sample of this size (111)</th>
</tr>
</thead>
<tbody>
<tr>
<td>PF00085.15</td>
<td>Thioredoxin</td>
<td>Redox, antioxidant, cell signalling</td>
<td>2</td>
<td>14</td>
<td>8.97E-03</td>
<td>0.13</td>
</tr>
<tr>
<td>PF01565.18</td>
<td>FAD_binding_4</td>
<td>FAD binding domain, most oxidoreductases</td>
<td>3</td>
<td>40</td>
<td>6.63E-03</td>
<td>0.37</td>
</tr>
<tr>
<td>PF02836.12</td>
<td>Glyco_hydro_2_C</td>
<td>Glycoside hydrolase family 2, includes β-galactosidases, β-glucuronidases</td>
<td>2</td>
<td>8</td>
<td>3.55E-03</td>
<td>0.07</td>
</tr>
<tr>
<td>PF03198.9</td>
<td>Glyco_hydro_72</td>
<td>Glycoside hydrolase family 72, cleave bonds between carbohydrates</td>
<td>3</td>
<td>7</td>
<td>8.70E-05</td>
<td>0.07</td>
</tr>
<tr>
<td>PF03372.18</td>
<td>Exo_endo_phos</td>
<td>Endonuclease, phosphatise, intra cellular signalling</td>
<td>2</td>
<td>12</td>
<td>6.93E-03</td>
<td>0.11</td>
</tr>
<tr>
<td>PF05730.6</td>
<td>CFEM</td>
<td>Fungal specific cysteine rich domain</td>
<td>2</td>
<td>14</td>
<td>8.97E-03</td>
<td>0.13</td>
</tr>
</tbody>
</table>

6.4 Discussion

6.4.1 Mating type of PmedOMT5

Sexual reproduction in fungi is an efficient mechanism for pathogens to rapidly generate genetic diversity by meiotic recombination (Chen and McDonald 1996), facilitating adaptation to new environments. In a fungus undergoing sexual reproduction there is also potential for meiotically-associated repeat-induced point (RIP) mutation (Hane and Oliver 2008, Van de Wouw et al. 2010), which can lead to further variation in duplicated genes or genes proximal to repetitive DNA. In light of this knowledge, consideration of the reproductive strategies of a pathogen is important in assessing its potential to adapt to and overcome disease management strategies both of its host and in the field.
Sexual compatibility in the Ascomycetes is governed by mating type genes. There are two main forms, heterothallic (self-sterile) and homothallic (self-fertile) both of which exist in the Didymellaceae family (Woudenberg et al. 2011). In all heterothallic (outcrossing) filamentous Ascomycetes studied to date, sexual reproduction is controlled by a single regulatory locus, the mating-type or MAT locus (Woudenberg et al. 2011). This locus contains one or another distinct, complementary idiomorph named either MAT1-1 or MAT1-2 (Turgeon and Yoder 2000). Homothallic fungi do not require a complementary isolate to complete their sexual cycle but possess both mating-types, either linked or unlinked within their genome. Heterothallicism in the Dothideomycetes is proposed to have evolved via deletion from a homothallic (self-fertile) ancestor (discussed in Amselem et al. 2011). MAT regions often show evidence of rearrangements from deletions of one or other mat loci leaving different remnants in different genomes resulting from insertions and rearrangements around repeated regions as seen in Neurospora tetrasperma (Ellison et al. 2011). Although homothallism in Didymella clematis is proposed to have arisen via a single crossover between MAT1-1 and MAT1-2 sequences from heterothallic ancestors (Woudenberg et al. 2011), indicating that MAT loci can be potentially both gained and lost.

*P. medicaginis* is believed to exist as an asexual species (Boerema et al. 2004) like many plant pathogens (Kronstad and Staben 1997), however its MAT locus has never previously been studied. In light of this, potential mating type genes were searched for in the PmedOMT5 genome as their discovery may shed light on both the pathogen’s evolution and its potential for mating. Following the identification of the MAT1-2-1 locus in the PmedOMT5 genome other isolates were examined since if a second mating type locus were to be discovered this would mean that *P. medicaginis* may be capable of sexual reproduction. *D. rabiei*, a closely related heterothallic species, has two mating types, however geographical separation has been reported in several populations (Kaiser et al., 1997, Barve et al. 2003 and references therein and references therein) and there is currently believed to be only one mating type (MAT1-2-1) present in Australia based on molecular evidence (Phan et al. 2003), rendering the population in this country only capable of asexual reproduction.

All isolates analysed by PCR carried a MAT1-2-1 gene including the CBS reference for *P. medicaginis var. medicaginis* and *P. pinodella*, as well as WAC47988 *D. exigua*. Primers were also designed to amplify the entire MAT idiomorph region and from an
area external to the MAT region into the conserved alpha domain of MAT1-1-1 if present. However none of these amplified a single product in the isolates tested.

The taxonomy of the genus *Phoma* is still under revision. Since 2006, scientists have been working on the molecular classification of this polyphyletic genus whose previous classification based on morphological characteristics had led to numerous artificial groupings (de Gruyter et al. 2013). In 2009, *Phoma* was divided into nine sections all of which are grouped within the order Pleosporales. Five of these sections form a distinct clade within the family Didymellaceae including those (mainly asexual) species that retained the name *Phoma* (de Gruyter et al. 2009). However the genus *Phoma* within the family Didymellaceae is still polyphyletic and contains both sexual species and those for which a teleomorph has not yet been identified (de Gruyter et al. 2009, Aveskamp et al. 2010), and it is still unknown whether or not these species have lost the ability to reproduce sexually. Aveskamp and colleagues (2010) postulate that *Phoma* spp. may simply have a “cryptic” sexual state that is only induced under specific conditions. The identification of the mating type of *P. medicaginis*, presented here may help to shed some light on the evolution of the *Phoma* genus in future studies.

### 6.4.2 Fungicide resistance

Whilst fungicide resistance genes are not usually involved in pathogenicity they determine the effectiveness of disease control in the field once plants are infected. Thus, potential mutation sites in PmedOMT5 genes were examined to assess the state of fungicide applicability.

Plant disease control with chemicals began in the 1850’s with the use of Bordeaux mixture to control downy mildew (Knight et al. 1997). In the 1960s, the use of site-specific systemic fungicides, which are absorbed into the sap stream of the plant, mostly replaced the previous multi-site inhibitors (reviewed in Ma and Michailides, 2005). Since that time there has been an increasing number of reports of resistance emerging mostly conferred by point mutations in the target genes (Gisi et al. 2000). The benzimidazoles, which bind to fungal β-tubulin, were one of the earliest classes of systemic fungicides and a large number of agricultural pathogens have since evolved resistance. Fortunately in PmedOMT5, the β-tubulin gene (*PmedOMT5_00352*) lacks the E198A mutation, which confers high benzimidazole resistance (Ma 2005). One of the next major fungicide classes in use is the strobilurins or QoI inhibitors, which inhibit
the cytochrome b (Cytb) protein, a major component of mitochondrial complex III. Resistance was first seen in early 2000 and is conferred by a G143A, glycine to alanine mutation in the Cytb gene (Gisi et al. 2002). The Cytb gene in PmedOMT5 (chapter 4.4.4) does not contain an intron and has no G143A mutation, but the potential for mutation is high based on the frequency of strobilurin resistance arising in other Dothideomycetes (Lesniak et al. 2011, Bolton et al. 2013).

Azole fungicides inhibit the cytochrome P450 14-alpha-demethylase, interrupting the conversion of lanosterol to ergosterol a component of the fungal cell membrane. Resistance to this class can come from point mutations in cyp51, over-expression of cyp51, or up regulation of efflux proteins. There appears to be only one copy of cyp51 (PmedOMT5_04693) which shows moderate expression levels and no mutations that are known to cause resistance. Thus control of P. medicaginis appears to be feasible with current fungicides. Nevertheless the potential for the emergence of these mutations in P. medicaginis, as with any other fungal pathogen, remains an ongoing concern.

6.4.3 Predicted functional annotation of PmedOMT5 genes

Before genes could be assigned a putative role in pathogenicity they first needed to be characterised based on their orthology and similarity to characterised genes. However, the function of the vast majority of genes identified across all species remains unknown, even in meticulously studied species. The characterised domain content of the human, Drosophila melanogaster, C. elegans, A. thaliana and S. cerevisiae proteomes varies from 57-75% (Yandell 2012). In PmedOMT5, only 66% (7,826) of the proteins contained conserved Pfam domains that could be used to infer their function in the cell although this number reflects well on the accuracy of the annotation.

6.4.4 Secreted proteins

One of the first criteria for identifying effectors is that they are secreted into the extracellular environment of the fungus. Nearly all secreted fungal proteins contain a signal peptide at the N-terminus that targets them for secretion to the rough ER or the Golgi complex. The signal peptide is typically 15-30 amino acids long and is cleaved from the mature protein during transport across the membrane (Min 2010). Accurate computational prediction of protein localisation is not a simple task and no method known to date has a 100% success rate, with the best results achieved by using a
combination of methods (Min 2010). Thus, in order to create an accurate list of secreted proteins that would serve as the initial short list of predicted effectors, the PmedOMT5 version 2 proteins were analysed by a variety of secretion prediction programs.

WoLFPsSORT predicts the sites of sub-cellular localisation of proteins based on their amino acid sequences and features such as the presence of sorting signals and functional motifs (e.g. DNA-binding) (Horton et al. 2006, Horton et al. 2007). Proteins are classified into more than 10 localisation sites, which can include dual localisations for proteins that shuttle between two locations, such as the cytosol and the nucleus. WoLFPsSORT predictions have been demonstrated to be more sensitive for certain locations such as the nucleus and mitochondria and less so for sites such as the peroxisome and Golgi bodies (Horton et al. 2007). Genes that are predicted to have an extracellular location include those with GO ID numbers GO:0005576 and GO:0005618, which account for extracellular and cell wall respectively. SignalP v 4 (Petersen et al. 2012) and Phobius (Kall et al. 2004) both predict proteins to be secreted by looking for conserved residues within the N-terminal amino acids and separating out those that contain multiple TM domains. Secretion prediction programs are known to have difficulty distinguishing the alpha helix contained in the secretion signal before the cleavage site, which is used for transport across the membrane, from actual TM domains (Petersen et al. 2012). For this reason, the 78 proteins, predicted to be secreted by all three methods, which had only one TM domain, were also retained in the predicted secreted set for further analysis. Those with predicted functions as transporters inferred from BLAST and BLAST2GO analyses were excluded from the potential effector set.

The intersection of all three predictive methods was used as the main selection criteria for determining the potential extracellular secretion of proteins (Figure 6.2). Information relating to those proteins predicted to be secreted by only one program was retained and used in combination with other analyses. Figure 6.2b illustrates that the majority of Ascomycetes examined had similarly sized extracellular profiles ~10% of the total gene count per organism.

There were 213 small secreted proteins that met the criteria used by Ohm and colleagues (2012): less than 200 aa, predicted to be secreted by SignalP and with one or less N terminal transmembrane domains predicted by TMHMM. This is a similar
number to that observed by Ohm in other Dothideomycetes: *P. nodorum* 209, *L. maculans* 188, *P. tritici-repentis* 223. Of these 213, 33 were found in the top 111 ranked candidates and effector scoring across all ranged from 42 (the top score) to 7.

### 6.4.5 Orthology of PmedOMT5 genes

Most of the PmedOMT5 proteins (85%) genes have orthologs, genes in another species that are estimated to have evolved from a common ancestral gene by speciation based on reciprocal BLASTP matches. Orthologs tend to retain the same function in the course of evolution although paralogs, duplicated genes within a genome can evolve different functions. Ortholog analysis of 42 fungal and oomycete species identified 146,962 ortholog groups with only 320 genes common to all species. Of those genes unique to PmedOMT5 or with paralogs only in PmedOMT5, 31 were found in the top 111 effector candidates.

### 6.4.6 Prediction of PmedOMT5 genes with potential roles in pathogenicity

From the pool of 11,879 PmedOMT5 version 2 genes annotated in chapter 6, a method was needed to identify those that may play a role in pathogenicity on *Medicago* spp. An obvious choice was to first identify those genes which were similar to genes known to be important for pathogenicity in other fungal species. However this approach can only identify conserved proteins and would miss any novel proteins produced by PmedOMT5. Thus a strategy combining multiple analyses was used. The first step was to identify all genes expressed in planta, then identify those that had homologs (or conserved domains that might indicate function) with genes in other species and thirdly to rank PmedOMT5 genes for their potential as effectors.

Plant pathogens have been demonstrated to produce a variety of compounds in vitro, some of these are phytotoxic but not all are directly involved in pathogenicity (Turner 1971). Close to half of the PmedOMT5 genes (49%) were detected as expressed IP by five or more mapped reads. In order to identify those that may play crucial roles in pathogenicity and virulence these genes were functionally categorised and examined by comparison to known pathogenicity genes in other species. Forty-eight genes were detected as expressed IP that showed no expression in any in vitro samples and thus have potential roles in plant pathogenicity. These included several cytochrome p 450s, pisatin demethylase orthologs, MFS transporters and pectate lyases.
6.4.7 Membrane transport proteins

The PmedOMMT5 genome contains 43 genes encoding proteins with ATP-binding cassette (ABC) superfamily transporter domains, 30 of which are expressed IP. ABC transporters are responsible for the transport of fungitoxic compounds from the cell, thus conferring decreased sensitivity to fungicides and plant defence compounds, such as phytoalexins and phytoanticipins. Most have a broad substrate specificity, but some are specific (Stergiopoulos et al. 2002). One protein, (PmedOMT5_03325) has a proposed role as a brefeldin A exporter, which may protect PmedOMT5 from the antifungal effects of its own secondary metabolite brefeldin A (Weber et al. 2004).

Another (PmedOMT5_09990), which has increased relative expression IP, has a strong match to a known phytoalexin (camalexin) exporter (Stefanato et al. 2009). The roles of the others are unknown, but in addition to efflux of fungicides and plant metabolites, ABC transporters can also be responsible for the secretion of mating type factors and mediating the secretion of host-specific effectors (Stergiopoulos et al. 2002), suggesting potentially important roles for these genes during fungal life and infection cycles.

6.4.8 Fungal growth IP

Many of the genes with higher expression detected in planta relative to the in vitro conditions tested suggest that significant breakdown and re-structuring of the fungal cell-walls is occurring. It cannot be determined at this stage if this is due to fungal cell wall degradation by plant enzymes followed by recovery of the fungus or simply growth of the fungus through the plant as infection progresses. In the analyses of other necrotrophic fungi/plant interactions, it has been observed that significant cell death in the pathogen, at early stages of infection (< 48h), is triggered by the host plant defences activating fungal PCD (Shlezinger et al. 2011a, Shlezinger et al. 2011b). Anti-apoptotic genes have been demonstrated to play an important role in protection of necrotrophic pathogens B. cinerea and C. heterostrophus against plant defences (Shlezinger et al. 2011b), by allowing a few fungal cells to survive the initial host onslaught and then grow through the necrotic plant cells. Whilst orthologs of two of these anti-apoptotic genes BcBIR and BcNma (PHI:2334) were identified as expressed IP they were not detected as strongly expressed in the P. medicaginis/M. truncatula interaction, with only 2 RNAseq reads mapping to the BcBIR homolog and 8 to BcNma. This is possibly due to the fact that the IP RNA-seq reads were derived from a single pool of 1-5 dpi
tissues and these genes are only expressed at the very early stage of plant-pathogen interaction. The role of these genes in the establishment of PmedOMT5 infection warrants further investigation.

6.4.9 Secondary metabolite biosynthesis genes

Fungi produce many secondary metabolite products some of which play a role in pathogenicity such as T-toxin (Yang et al. 1996), others can cause post-harvest contamination of seed which can be toxic to humans and animals, such as trichothecenes which may also play roles in virulence (Proctor et al. 2002) and aflatoxins (Bhatnagar et al. 2003). Others are produced in the plant and have a hormonal effect, such as such as zearalenones which causes premature labour in stock (Gaffoor et al. 2005) or have a proposed role in establishing an anti-microbial exclusion zone in an infected plant for the producing pathogen, such as brefeldin A(Weber et al. 2004). Several fungal host-specific effectors are non-ribosomal peptide products of NRPSs, including victorin (Navarre and Wolpert 1999), HC-toxin (Walton 2006) and AM-toxin (Johnson et al. 2000). Additionally, the product of a PKS/NRPS hybrid enzyme (Ace1) in M. grisea can act as an effector (avirulence factor), triggering resistance in rice cultivars harbouring the corresponding resistance gene (Böhnert et al. 2004).

Secondary metabolites in fungi are the products of large multi-modular enzymes that consist of conserved functional domains characteristic to each enzyme type, with the three main classes being PKSs, NRPSs or DMATs. PKSs are enzyme complexes that contain a coordinated group of active sites which synthesise their products via the stepwise assembly of small 2-4-carbon building blocks including acetyl-CoA, propionyl CoA, butyryl-CoA and their activated derivatives. PKS genes are known to be more abundant in the phytopathogenic Ascomycetes compared to their saprophytic counterparts (Kroken et al. 2003). Identifying the roles of all secondary metabolite production genes in PmedOMT5 will require experimental validation, however putative roles can be assigned to several based on their high sequence homology with other characterised PKSs. Without near-perfect sequence homology, however, this approach is not feasible as the products of PKS genes are also dependent upon the order of their functional domains, i.e. two PKSs that produce the same or similar products must have similar domains in the same order. Even for slightly divergent homologs with identical domain order, there is no guarantee they would produce the same secondary metabolite.
Some of the few sequences available for *P. medicaginis* before the genome was sequenced were three partial PKS gene sequences from isolate AS4 isolated from *M. sativa* (Akamatsu et al. 2010). The protein sequence for one of these AsPKS1 [GenBank: GQ150545, 229 aa] had 100% amino acid conservation to part of PmedOMT5_01171. AsPKS1 grouped with sequences characterised as non–reduced type polyketide synthases responsible for the production of the 1,8-dihydroxynaphthalene-melanin (DHN-melanin) pigment by *A. rabiei* (Akamatsu et al. 2010). Although DHN-melanin is required for pathogenicity in fungi that form melanised (structurally reinforced)-appressoria (Kubo et al. 1982, Wolkow et al. 1983), this gene was found not to be a requirement for pathogenicity in other Dothideomycetes including *Alternaria alternata* and *Bipolaris oryzae* (Kubo et al. 1989, Moriwaki et al. 2004). In *A. rabiei* site-directed mutagenesis of this gene resulted in mutants that, although they produced melanin-deficient pycnidia, did not show reduced pathogenicity on chickpea (Akamatsu et al. 2010). Akamatsu and colleagues (2010) tested the hypothesis that melanisation may play a role in protecting asexual pycnidia from UV radiation in the field, showing that *ArPKS1* mutants were more sensitive than wild-type *A. rabiei* pycnidia to UV light. The PmedOMT5 ortholog showed the highest FPKM value in the 16 d (pycnidia production on minimal media) sample.

Two other *P. medicaginis* PKS genes had been previously reported by Akamatsu and colleagues (Akamatsu et al. 2010). AsKS3 and AsKS2 grouped in different subclades of a separate clade from As PKS1. Analysis performed in this study indicates that the orthologs of these two genes in PmedOMT5 are paralogs. PmedOMT5_5798 has a BLASTP match to AsPKS2 (GQ150549.1) with 92% aa homology for the partial match. PmedOMT5_5798 has 2 paralogs in PmedOMT5, PmedOMT5_10407 (BLASTP match to AsPKS3_GQ150550.1, 235 aa, 98% aa identity) and PmedOMT5_10816, as well as orthologs in 25 fungal species. According to Akamatsu and colleagues (2010) the PKSs in clade 2 are thought to build non-reduced and methyl group-attached polyketides, which may be non-essential for fungal life, such as phytotoxins or mycotoxins. There is only one characterised gene in this clade, a PKS responsible for the synthesis of the nephrotoxic mycotoxin citrinin in *Monascus purpureus* (Shimizu et al. 2005). All three paralogs were detected as expressed in vitro but only PmedOMT5_10407 was detected as expressed IP and also showed increased relative expression IP compared to the other conditions tested.
Some PmedOMT5 secondary metabolism genes that were expressed in planta appear to have roles in response to oxidative stress. PmedOMT5_06337, characterised as an NRPS, also showed increased relative expression IP. This gene has a strong match (bitscore: 2,799) to NPS6 (PHI:2261), the deletion of which slightly reduced the virulence of Cochliobolus sativus on barley. A homolog of NPS6 in the closely related C. heterostrophus, is involved in virulence and tolerance to oxidative stress (Lee et al. 2005) and has a conserved role in iron uptake and extracellular siderophore biosynthesis in filamentous Ascomycetes (Oide et al. 2006). PmedOMT5_10680 is an ortholog of PKS10 from C. heterostrophus (Lee et al. 2005), an ancestral PKS conserved across the sequenced Dothideomycetes with a role in oxidative stress response (Ohm et al. 2012).

When comparing the numbers of secondary metabolite biosynthesis genes between sequenced filamentous Ascomycetes and oomycetes, PmedOMT5 shows a profile closest to that of the Dothideomycete P. nodorum (Table 6.19). However, comparisons of secondary metabolite production genes in the class Dothideomycetes by Ohm and colleagues (2012) indicate that there is little conservation in the roles of secondary metabolite biosynthesis genes, excluding the relatively well-conserved NPS6 and PKS10.

### Table 6.19: Number of PKS-, NRPS-, and Cytochrome P450–encoding genes in the genomes of selected Fungi and Oomycetes compared to PmedOMT5.

<table>
<thead>
<tr>
<th>Species*</th>
<th>PKS</th>
<th>NRPS</th>
<th>PKS-NRPS</th>
<th>P450</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. nidulans</td>
<td>27</td>
<td>13</td>
<td>1</td>
<td>102</td>
</tr>
<tr>
<td>M. grisea</td>
<td>23</td>
<td>6</td>
<td>8</td>
<td>115</td>
</tr>
<tr>
<td>N. crassa</td>
<td>7</td>
<td>3</td>
<td>0</td>
<td>39</td>
</tr>
<tr>
<td>P. chrysosporium</td>
<td>9</td>
<td>7</td>
<td>0</td>
<td>148</td>
</tr>
<tr>
<td>P. sojae</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>30</td>
</tr>
<tr>
<td>P. ramorum</td>
<td>0</td>
<td>4</td>
<td>0</td>
<td>24</td>
</tr>
<tr>
<td>S. nodorum</td>
<td>19</td>
<td>8</td>
<td>1</td>
<td>103</td>
</tr>
<tr>
<td>PmedOMT5</td>
<td>19</td>
<td>7</td>
<td>1</td>
<td>101</td>
</tr>
</tbody>
</table>

*Data adapted from Soanes 2007

#### 6.4.10 Carbohydrate-active enzymes devoted to plant cell wall depolymerization.

Plant cell-walls, composed mainly of pectins, celluloses, hemi-celluloses and lignins, act as a physical barrier to keep pathogens from entering. From the perspective of a pathogen, plant cell-walls are a barrier that must be broken down to obtain nutrition from within and in the process plant cell walls can also serve as a carbon source for successful pathogens (Ohm et al. 2012, Zhao et al. 2013). A recent study by Zhao and colleagues (2013) compared the CAZyme complements of 103 fungi, including a range
of pathogens, saprophytes and fungal phyla, which demonstrated that the CAZyme profile of an organism reflects its life-style. Plant pathogenic fungi tend to produce a larger total number of CAZymes than saprophytic fungi, with pathogens of different groups of plant hosts or different trophisms, also differing in the contents of their cell wall degrading arsenal (King et al. 2011, Zhao et al. 2013). CAZymes devoted to plant cell-wall depolymerisation are greatly reduced in biotrophs such as *Blumeria graminis* f. sp. *hordei* (Spanu et al. 2010) and pathogens of dicots tend to have more pectin degrading enzymes than pathogens of monocots, due to the higher amounts of pectin present in dicot cell-walls (Zhao et al. 2013). Some CWDEs such as pectinases and xylanases have demonstrated roles in pathogenicity (Douaiher et al. 2007, Kikot et al. 2009). Some of these enzymes, or the by-products of their degradation of host-plant molecules, have also become the target of PTI-based plant defences (Matzinger 2007). For most necrotrophic fungi, pectinases, which are CWDEs capable of macerating plant tissues and killing plant cells, are the first enzymes to be produced (Cooper 1983).

When compared to other Pleosporales pathogens (Table 6.20) PmedOMT5 has a similar CAZyme profile to the other dicot pathogens, with larger numbers of PL1 and 3 families as well as, one PL7 (alginate lyase), which was not observed in other Pleosporales (Ohm 2012-supp. data) but has been observed in other species of filamentous Ascomycetes, including *P. fulvum* and *Botrytinia fuckeliana*. Genes with this domain (Pfam: PF08787) have been characterised as alginate lyases, enzymes that degrade the linear polysaccharide alginates found in the cell walls of brown algae and are involved in the cleavage of other polysaccharides containing beta-D-mannuronate residues. PL1 family members are important in plant pathogenicity with many more members found in plant pathogens than saprotrophic fungi (Zhao et al. 2013).
Table 6.20: Comparison of pectate lyase families amongst the Pleosporales.

<table>
<thead>
<tr>
<th>Species*</th>
<th>Trophism and host clade</th>
<th>1</th>
<th>3</th>
<th>4</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>11</th>
<th>14</th>
<th>20</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>PmedOMT5</td>
<td>N, D</td>
<td>9</td>
<td>10</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>20</td>
</tr>
<tr>
<td>Leptosphaeria maculans v23.1.3</td>
<td>H, D</td>
<td>9</td>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>19</td>
</tr>
<tr>
<td>Alternaria brassicicola ATCC 96866</td>
<td>N, D</td>
<td>8</td>
<td>11</td>
<td>3</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>23</td>
</tr>
<tr>
<td>Cochliobolus heterostrophus C5</td>
<td>N, M</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>Cochliobolus heterostrophus C4</td>
<td>N, M</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>Cochliobolus sativus ND90Pr</td>
<td>H, M</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>Setosphaeria turcica Et28A</td>
<td>H, M</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>14</td>
</tr>
<tr>
<td>Pyrenophora teres f. teres 0-1</td>
<td>N, M</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Pyrenophora tritici-repentis Pt-1C-BFP</td>
<td>N, M</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>P. nodorum SN15</td>
<td>N, M</td>
<td>4</td>
<td>2</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>Pleosporales average</td>
<td></td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>15</td>
</tr>
<tr>
<td>Monocot average</td>
<td></td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>13</td>
</tr>
<tr>
<td>Dicot average</td>
<td></td>
<td>9</td>
<td>9</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>21</td>
</tr>
</tbody>
</table>

N= necrotroph, H=hemi-biotroph, D= dicot, M= monocot  * Data for other species from Ohm et al. 2012.

Glycosyl transferases are enzymes responsible for the biosynthesis of glycosidic bonds in fungal cell wall assembly and reorganisation. PmedOMT5 has a higher number (113) compared to other Pleosporales pathogens analysed by Ohm and colleagues (2012), which had an average of 94 (Table 6.21). Highly expanded families include GT2 (known roles in chitin synthesis, and methyl-group transfer among others) which has 59 members in PmedOMT5, 44 of which are expressed IP, with an average for monocot Pleosporales pathogens of 17 and for dicot pathogens, 15. Not all of the 59 proteins predicted to contain domains belonging to the GT2 family also have EC numbers assigned and they have a range of predicted functions based on orthology including histidine kinases, metal ion binding and involvement in fatty acid biosynthesis in addition to chitin synthases. With representatives from all these groups expressed IP as well as an ortholog of LaeA, a conserved fungal transcriptional regulator, shown to be important for virulence (Bok et al. 2005), PmedOMT5_02796 (PHI:2315).

In comparison with CAZyme analysis of other Pleosporales species presented by Ohm and colleagues (2012) (Table 6.21), PmedOMT5 appeared to have an unusually enlarged number of CEs and fewer CBMs. CAZyme analysis for PmedOMT5 was performed using Pfam analysis only, which is more conservative than analyses which incorporate orthology, as was used by Ohm and colleagues, followed by manual curation. The CAZyme orthology analysis of PmedOMT5 proteins predicted a total of
267 proteins with CBMs, although only 29 show consistent domains and length with their matches. To ensure that only correctly assigned CAZyme activities were compared, only the more conservative Pfam data was used, which may have resulted in a somewhat conservative underestimation of some CAZymes, particularly CBMs.

The inflated number of CE genes in PmedOMT5 results from the inclusion of the CE10 family, which has 48 members in PmedOMT5. The CE10 family was omitted from the data presented by Ohm and colleagues (2012) as the CAZY database states that it no longer supports this class as a CAZyme, as most if not all of the characterised esterases in this family act on non-carbohydrate substrates. When proteins with CE10 domains are removed from this analysis PmedOMT5 has only 43 CEs, which is consistent with other fungi of the order Pleosporales.

<table>
<thead>
<tr>
<th>Species</th>
<th>GH</th>
<th>PL</th>
<th>CE</th>
<th>CBM</th>
<th>GT</th>
</tr>
</thead>
<tbody>
<tr>
<td>PmedOMT5</td>
<td>243</td>
<td>20</td>
<td>91 (43 (^a))</td>
<td>15</td>
<td>113</td>
</tr>
<tr>
<td>Cochliobolus heterostrophus C5</td>
<td>292</td>
<td>15</td>
<td>49</td>
<td>101</td>
<td>103</td>
</tr>
<tr>
<td>Cochliobolus heterostrophus C4</td>
<td>276</td>
<td>15</td>
<td>46</td>
<td>63</td>
<td>96</td>
</tr>
<tr>
<td>Cochliobolus sativus ND90Pr</td>
<td>272</td>
<td>15</td>
<td>47</td>
<td>97</td>
<td>99</td>
</tr>
<tr>
<td>Setosphaeria turcica Et28A</td>
<td>254</td>
<td>14</td>
<td>41</td>
<td>68</td>
<td>94</td>
</tr>
<tr>
<td>Alternaria brassicicola ATCC 96866</td>
<td>248</td>
<td>23</td>
<td>40</td>
<td>53</td>
<td>83</td>
</tr>
<tr>
<td>Pyrenophora teres f. teres 0-1</td>
<td>254</td>
<td>10</td>
<td>39</td>
<td>60</td>
<td>98</td>
</tr>
<tr>
<td>Pyrenophora tritici-repentis Pt-1C-BFP</td>
<td>245</td>
<td>10</td>
<td>38</td>
<td>45</td>
<td>91</td>
</tr>
<tr>
<td>Leptosphaeria maculans v23.1.3</td>
<td>231</td>
<td>19</td>
<td>33</td>
<td>52</td>
<td>94</td>
</tr>
<tr>
<td>Stagonospora nodorum SN15</td>
<td>273</td>
<td>10</td>
<td>49</td>
<td>61</td>
<td>89</td>
</tr>
<tr>
<td>Average for all Pleosporales</td>
<td>261</td>
<td>15</td>
<td>42</td>
<td>67</td>
<td>94</td>
</tr>
<tr>
<td>Average for Pleosporales monocot pathogens</td>
<td>267</td>
<td>13</td>
<td>44</td>
<td>71</td>
<td>96</td>
</tr>
<tr>
<td>Average for Pleosporales dicot pathogens</td>
<td>240</td>
<td>21</td>
<td>37</td>
<td>53</td>
<td>89</td>
</tr>
</tbody>
</table>

\(^a\) excluding CE 10 family GH glycoside hydrolases, PL polysaccharide lyases, CE carbohydrate esterases, CBM carbohydrate-binding modules (non-enzymatic), GT glycosyltransferases

**6.4.11 In planta co-expressed gene clusters**

Clusters of genes that were detected as expressed IP were identified by visually inspecting IP expressed genes for those with three or more consecutive locus numbers or in close proximity on the same scaffold, as some fungi, notably the basidiomycete *Ustilago maydis*, a pathogen of corn, are known to express effector genes in clusters (Kamper et al. 2006). Many of the clusters identified were related to basic cellular processes or small groups of 3-5 genes frequently containing cytochrome p450s or various kinases.
A cluster of genes was identified as strongly expressed IP on scaffold_260 (~37 kbp). This scaffold has a lower average G:C content (48%) in comparison to the genome wide average of 53%. Scaffold_260 encodes 19 genes, 14 of which were expressed IP, 10 showing a more than 4-fold increase in relative expression, nine of which are consecutive \textit{PmedOMT5\_10172-10180}. Of those 14 genes expressed IP, five have PHIbase hits (four show a more than 4-fold increase in relative expression up IP). Two genes in this cluster have EST support (one of which, \textit{PmedOMT5\_10178} shows increased relative expression IP), which provides a further independent validation of the importance of their expression in the \textit{P. medicaginis/M. truncatula} interaction as it is observed across two \textit{Phoma} isolates in different infections. This cluster includes two genes showing increased relative expression IP with matches to the phytoalexin detoxification genes \textit{MAK1} and \textit{PDA}. Proposed roles for other genes in this cluster include: maleylacetate reductase, vanillin dehydrogenase, catechol dioxygenase, isoflavone reductase family protein, a short chain dehydrdogenase and two fungal specific transcription factors.

Several other small clusters were identified surrounding putative phytoalexin detoxification genes, as well as genes involved in the production of secondary metabolites. While most genes in these clusters had orthologs, few of these clusters shared the same gene order with any of the other fungal species analysed.

\textbf{6.4.12 Phytoalexin detoxification genes}

Plant pathogenic microorganisms, are in general, more resistant to the phytoalexins produced by their host plants than to those of non-host species (VanEtten \textit{et al.} 2001). Plants produce phytoalexins during both susceptible and resistant interactions, although in some resistant interactions the response is faster and compounds may accumulate to higher levels in resistant plants (Morrissey and Osbourn 1999, Kamphuis \textit{et al.} 2011). Thus a pathogenic fungus would be continuously exposed to the phytoantycipins and phytoalexins produced by its host and evolution of tolerance to the host phytoalexins may be expected to occur and contribute to virulence (VanEtten \textit{et al.} 2001).

Fungal tolerance of phytoalexins is generally correlated with the ability of to detoxify them (Denny and VanEtten 1982), although tolerance can also be conferred by non-degradative mechanisms (Denny and VanEtten 1981, Denny \textit{et al.} 1987). For example, an ortholog of an ABC transporter from \textit{B. cinerea}, \textit{BcatrB} (PHI:2309), was found to
show increased relative expression in PmedOMT5 during the early stage of infection. This transporter exports the brassica phytoalexin, camalexin from the fungal cell. This increases the tolerance of the pathogen to the phytoalexin and a similar phytoalexin-export process may potentially occur in PmedOMT5, mediated by this ortholog or other uncharacterised transporters identified as expressed IP.

*M. sativa* and *M. truncatula* produce medicarpin, sativan, daidzein and formononetin via the phenylpropanoid pathway (Martin and Dewick 1980, Deavours and Dixon 2005, Farag et al. 2007, Farag et al. 2008, Jasinski et al. 2009). These compounds have been classified as phytoalexins and can be degraded by some fungi, e.g. daidzein by *Aspergillus saitoi* (Esaki et al. 1998) and formononetin by *Fusarium* sp. (Weltring et al. 1982). The production of some of these compounds is increased following inoculation of alfalfa and *M. truncatula* with *P. medicaginis* (Deavours and Dixon 2005, Kamphuis et al. 2011). Previous studies have indicated that *P. medicaginis* isolates can detoxify or are insensitive to medicarpin or its precursors daidzein, formononetin and 2'-hydroxyformononetin (Higgins 1972, Blount et al. 1992).

Growth of *P. medicaginis* (syn. *P. herbarum var. medicaginis*) was not inhibited by medicarpin concentrations up to 75 µg/g/mL on agar plates, although this concentration was sufficient to inhibit growth of a non-alfalfa pathogen *Helminthosporium turcicum* (syn: *Setosphaeria turcica*) (Higgins 1972). The medicarpin concentration in alfalfa leaves infected with *P. medicaginis* was estimated to be 27µg/g fresh weight, although this is likely to be an underestimate as the extraction and purification process was only 70% efficient and only 25% of the leaf was infected (Higgins 1972). Medicarpin levels in alfalfa suspension cultures have been reported as high at 800 nmol/g (~216 µg/g) of cell fresh weight, following elicitation with a PAMP-rich fungal cell-wall preparation from *Colletotrichum lindemuthianum* (Kessmann et al. 1990). *P. medicaginis* has been demonstrated to degrade medicarpin that was added to media to form two new unidentified compounds and spores were also found to degrade medicarpin on the leaf surface. This degradation was proposed to occur extracellularly (Higgins 1972). *PmedOMT5_10173*, a gene homologous to the medicarpin detoxification gene *MAKI*, showed high expression *in planta*, was predicted to be secreted by all three methods and is therefore a strong candidate for a gene that may degrade medicarpin in PmedOMT5.
Conversely, studies by Blount and colleagues (1992) showed that growth of *P. medicaginis* may be inhibited by medicarpin (42%) and its precursor vestitone (44%) at 0.5 M concentration, although this may be much a higher concentration than normally encountered in plants. However the results of this study were difficult to interpret but seemed to indicate a trend of increased susceptibility as medicarpin concentration increased based on a series of plate assays. As *P. medicaginis* grows very slowly (up to 50% slower than the other isolates tested) and showed lobulated rather than radial growth, there was large variation within the experiment making it difficult to accurately determine growth rates (Blount *et al.* 1992). *N. haematococca* isolates previously demonstrated to detoxify maackiain used in the same study were not inhibited by medicarpin (Blount *et al.* 1992). These contrasting results may indicate a difference in the presence, activity or number of detoxification genes in each isolate. Increased virulence by the fungal pathogen has been associated with the speed at which it is able to detoxify the host’s phytoalexins (Delserone *et al.* 1999).

A strong correlation between higher or earlier phytoalexin production and greater disease resistance has been reported for a number of plant species including chickpea (Weigand *et al.* 1986) and others discussed by Mudodi and colleagues (2001). In alfalfa, following pathogen attack, a high level of medicarpin accumulates and is believed to increase the resistance of alfalfa to certain pathogens (Latunde-Dada *et al.* 1987, Paiva *et al.* 1991, Paiva *et al.* 1994). The mechanism by which phytoalexins act against fungi is not clear although the *Arabidopsis thaliana* phytoalexin, camalexin, has been demonstrated to induce programmed cell death (PCD) in *B. cinerea in vitro* (Shlezinger *et al.* 2011a). Additionally, pad3 *Arabidopsis* mutants which are deficient in camalexin production (Glawischnig 2007) showed increased infection, relative to the wildtype Col-0, when infected with *B. cinerea* mutants with attenuated expression of the anti-apoptotic gene *BcBIR1*. Attenuated expression of *BcBIR1* resulted in a decrease in anti-apoptotic activity with corresponding increased PCD and increased sensitivity to camalexin, whilst over-expression of *BcBIR1* showed reduced PCD and reduced sensitivity to camalexin (Shlezinger *et al.* 2011b). These data suggest that exposure to phytoalexins produced by living plant cells, early in infection, may be triggering host-induced cell death in the fungus, that is later overcome by pathogens in compatible interactions as they grow into necrotic tissue (Shlezinger *et al.* 2011b).
Several genes were identified in PmedOMT5 that may facilitate the degradation of phytoalexins produced by its host plants. These genes had similarities to genes shown to degrade medicarpin and maackiain (MAK1) and pisatin (PDA) in N. haematococca. *P. medicaginis* isolates have previously been shown to degrade both medicarpin (Higgins 1972) and pisatin (Delserone *et al.* 1992). *P. medicaginis* is known to causes symptoms on pea (Kinsey 2002), but neither *M. truncatula* or *M. sativa* is known to produce pisatin although it can be produced in *M. truncatula* cell cultures if fed 6a-hydroxymaackiain (Liu *et al.* 2006). Possibly some of the genes characterised as pisatin demethylases are detoxifying other pterocarps such as sativan which is produced by both alfalfa and *M. truncatula*. A gene involved in the conversion of vestitol to sativan is known to be transcriptionally up-regulated following *P. medicaginis* infection of *M. truncatula* (Deavours *et al.* 2006), so it is likely to be found in infected tissue but its toxicity to *P. medicaginis* is unknown. These genes may simply encode non-specific cytochrome p450s in PmedOMT5 as they share only ~49% identity with the *N. haematococca* PDA genes which all share ~90% identity with each other (Maloney and VanEtten 1994), although pisatin demethylase genes identified in other species pathogenic on pea *A. pisi*, *M. pinodes* (syn: *P. pinodes*) or *D. pinodella* (syn: *P. pinodella*) also have low homology to PDA genes in *N. haematococca* (Delserone *et al.* 1999).

Phytoalexin detoxification genes are presumed to have arisen in different species through convergent evolution based on the varying mechanisms used to detoxify the same product although there is some evidence that they may have arisen in some *Fusarium* species via HGT (Liu *et al.* 2003), this does not appear to be the common method of acquiring them (VanEtten *et al.* 2001). MAK1 and several PDA genes exist on a dispensable chromosome in *N. haematococca* (Wasmann and VanEtten 1996, Enkerli *et al.* 1998) and PmedOMT5 scaffold 260 contains both MAK1 and PDA, as well as some genes that are also observed on the same chromosome in *N. haematococca*. However, the available genome sequence evidence suggests that the order of homologous genes have been substantially rearranged between these species, which is more consistent with vertical inheritance from a common ancestor or ancient LGT, rather than recent acquisition.

These data suggest that phytoalexin tolerance or detoxification may be some of the mechanisms *P. medicaginis* uses to evade host defences. Both PDA and MAK1 have
been demonstrated to increase virulence on pea and chickpea respectively, when inserted into isolates that lack these genes (Ciuffetti and VanEtten 1996, Enkerli et al. 1998). Additionally, disruption or deletion of these genes leads to a decrease in virulence, however neither has been demonstrated as essential for pathogenicity. Further experimentation would be required to confirm if this is also the case for PmedOMT5.

6.4.13 NEP1-like (NLP) family genes.

Based on crystal structure analysis and mutagenesis, it has been proposed that NLPs can function as cytolytic toxins that induce plasma membrane leakage, thus causing cell death (Ottmann et al. 2009; Qutob et al. 2006) via a mechanism that is genetically distinct from immunity-associated PCD (Qutob et al. 2006). Not all NLPs from all species induce host-cell death (Santhanam et al. 2013). Most fungal genomes contain up to three NLP genes (Santhaman 2013 and references therein) and three were identified in the PmedOMT5 genome. The predicted proteins of all three genes contained the heptapeptide motif ‘GHRHDWE’, conserved in NLPs.

Expansion of NLP genes in the broad host-range vascular wilt pathogen *Verticillium dahliae*, has been proposed to contribute to its virulence. However only two out of seven members (NLP1 and NLP2) induced plant cell death and had an effect on virulence following targeted deletion (Santhanam et al. 2013). Interestingly the two active genes were expressed differently during infection of different hosts, with NLP1 expressed in both tomato and *Nicotiana bethamiana*, whilst NLP2 was expressed only during infection of tomato. Disruption of NLP1 caused a reduction in virulence on tomato, *Arabidopsis thaliana* and *N. bethamiana*, whilst NLP2 mutants showed less virulence on tomato and *A. thaliana*, but not *N. bethamiana*. Disruption of NLP1 also resulted in reduced growth and conidiospore production in vitro (Santhanam et al. 2013).

Further support for a somewhat host-specific effector activity of NLPs comes from a single NPP1-domain-containing protein identified in *M. graminicola*, MgNLP. This NLP homolog induced necrotic cell death and activation of defence related genes in *Arabidopsis*, but not in a susceptible wheat line following heterologous expression and infiltration (Motteram et al. 2009). Targeted deletion of *MgNLP* had no effect on virulence or pathogenicity on wheat, although the gene showed high expression in planta prior to observation of symptoms (Motteram et al. 2009). However, the
disruption of an \textit{NLP1} gene in the bacterial soft rot potato pathogen \textit{Erwinia carotovora} (Pemberton \textit{et al.} 2005) did result in reduced virulence. Deletion of two NLP-domain genes identified in \textit{B. cinerea}, both of which were expressed during infection of tomato leaves and whose encoded proteins caused necrosis when infiltrated into \textit{N. benthamiana}, showed no reduction in virulence (Arenas \textit{et al.} 2010). Although possibly their effects were overshadowed by other effector genes produced by this species (Arenas \textit{et al.} 2010). Thus, the role of NLPs in conferring virulence is ambiguous and does not always conform to a similar pattern in different pathosystems indicating that these genes have evolved to have different roles in different pathogens and on different hosts. Furthermore, Motteram and colleagues (2009) proposed that \textit{MgNLP} may have a similar role to that proposed for brefeldin, \textit{i.e.} inhibitory activity against filamentous fungi and other microbes (Weber \textit{et al.} 2004), which may explain why the gene is also maintained in non-plant pathogens and why its presence does not always correlate with disease phenotype. For these reasons, the possible role of NLP proteins in PmedOMT5 pathogenicity should be further examined, beginning with the two that showed increased expression IP, PmedOMT5_11472 which had the strongest similarity to NEP1 and \textit{PmedOMT5}_00836 which was most similar to NEP2.

### 6.4.14 Effector associated motifs

Effector associated motifs that had been identified in other plant-pathogenic species were examined for over-representation in the top 111 effector candidates, the 304 genes unique to PmedOMT5 that were expressed IP and the 213 SSPs, none were found to be significantly over-represented. The \texttt{[LI]XAR}, RXLR and RGD motifs were however found to be significantly over-represented in genes expressed IP by Fisher’s exact test \( (p \leq 0.01) \). These domains are proposed to be associated with the entry of fungal effectors into the host cell, with RXLR domains suggested to mediate direct transport of \textit{Phytophthora} proteins into the plant cell via interaction with host cell-membrane phospholipids (Kale \textit{et al.} 2010). The significance of these motifs in PmedOMT5 is unclear but initial analysis shows few are associated with proteins predicted to be secreted or occurring in the N terminal portion of the proteins where they would be most likely to occur if involved in translocation. In the top 111 effector candidates, 41 proteins contained \texttt{[Y/F/W]XC} domains, although this domain is not significantly over-represented in this set. The \texttt{[Y/F/W]XC} domain was previously identified in haustoria-
secreted effector-like proteins in powdery mildew and rust fungi (Godfrey \textit{et al.} 2010). In these fungi the motif occurred shortly after the signal peptide (within 17 aa), although in many of the PmedOMT5 effector candidates the motifs appeared closer to the C terminus. In \textit{Blumeria graminis} these genes, although highly divergent, are thought to have originated from a single ancestor where only the one specific codon is maintained, as well as exon structure and intron phase. It is unclear if their appearance in PmedOMT5 effector candidates is significant, but may warrant further investigation. A search for novel conserved motifs in PmedOMT5 effector candidates revealed no obviously associated highly represented motifs.

\textbf{6.4.15 Prediction of proteinaceous effector candidates}

There are two main bioinformatic strategies to look for novel effector genes that may be involved in pathogenicity. One can either a) look \textit{de novo} for patterns or domains associated with genes expressed during infection or at specific locations during infection such as the plant/haustorial interface (Pedersen \textit{et al.} 2012, Saunders \textit{et al.} 2012) or b) search for proteins that share characteristics with proteins demonstrated to play a role in pathogenicity and virulence in previous studies. In recent years many genome studies have gone ‘effector hunting’, looking for small secreted molecules that may be tripping the switch in plant cells to allow necrotrophic invasion (Rouzel \textit{et al.} 2011, O’Connell \textit{et al.} 2012, Condon \textit{et al.} 2013). By making a short list of bioinformatically-predicted, small, secreted proteins with other characteristics associated with experimentally verified effector proteins researchers can lower the number of candidates that need to be tested in the laboratory for their species of interest. Effectors identified to date have tended to be small, secreted proteins, expressed \textit{in planta}, though not necessarily more highly, with a high cysteine count (which confers stability on proteins secreted outside the cell) and some have been identified proximal to repetitive DNA regions. These criteria were used to rank potential effector candidates in PmedOMT5.

Amongst the top 111 ranked effector candidates (those with an effector score of greater than 25) are several genes with known roles in pathogenicity or that contain domains predicted to be involved in pathogenicity including CAP and CFEM domain-containing proteins which were also observed to be over-abundant in secreted rust haustorial proteins (Saunders \textit{et al.} 2012). Proteins with known roles include homologues of the
non-specific effectors Ecp6 and NEP1. Many other genes (77 in the top 111) have
BLAST matches of varying strengths to uncharacterised genes from other species, but
this does not exclude them as host-specific effectors. One of the top ranked predicted
effectors (effector score of 38) \textit{PmedOMT5\_11737}, has a paralog \textit{Pmed\_OMT5\_03381}
(effector score of 5) and orthologs of both have been found only in other Didymellaceae
species. This may suggest the recent evolution and expansion of an effector family in
this closely related group of legume pathogens.

One other potential effector candidate \textit{PmedOMT5\_03659} was highly expressed IP but
had little expression under other conditions and was also detected in the extracellular
protein fraction. This gene contains a conserved ribonuclease domain. The conserved
domain database match for this gene [CDD:CD00606, PFAM:pfam00545] describes the
related molecules as mostly guanyl-RNAs or ribotoxins, a group of fungal cytotoxins,
specifically cleaving the sarcin/ricin loop (SRL) structure of the 23-28S rRNA and
therefore being very potent inhibitors of protein synthesis. In a recent study of powdery
mildew, a number of candidate secretory effector proteins were identified that had
similarities to ribonucleases (of a different type than \textit{PmedOMT5\_03659}), including
conserved structure (Pliego et al. 2013). The authors note that ribonucleaseses contain a
conserved disulphide bond that increases extracellular stability and are highly resistant
to proteolytic degradation. They also propose that these proteins may modify host
defences by interaction with plant RNA (Pedersen et al. 2012). \textit{PmedOMT5\_03659} has
orthologs in 31 of the other fungal species analysed and is therefore unlikely to be
involved in host specific pathogenicity unless these RNAses have diverged to target
different host RNAs, but may be an important virulence factor. The role of this gene in
\textit{PmedOMT5} pathogenicity clearly warrants further investigation.

Experimental analysis is required to investigate the activity and potential host-
specificity of these predicted effectors. Initial characterisation would begin with the
confirmation of their early expression \textit{in planta}, via qRT-PCR from a time-course
experiment with resistant and susceptible hosts. This could be followed by gene knock-
out/disruption to assess if they are essential for pathogenicity or simply virulence
factors. Their activity could also be assessed via synthesis of a cloned version of these
proteins in an expression vector such as \textit{Pichia pastoris} (Liu et al. 2009) followed by
infiltration into the host plant or potentially transient expression in \textit{Nicotiana bethamiana}
or \textit{M. truncatula} leaves via agroinfiltration (Pais et al. 2013).
6.5 Summary

Based on a comprehensive series of bioinformatic analyses, it appears that upon invasion of the host cell PmedOMT5 secretes an Ecp6 ortholog to bind its extracellular chitin molecules to evade host PAMP-recognition and PTI. *P. medicaginis* may defend itself from the phytoanticipins, and once detected, phytoalexins produced by its host, by both actively degrading and expelling them from the fungal cell. During fungal invasion of *M. truncatula*, PmedOMT5 also produces a battery of proteins involved in degrading and digesting plant cell components including cell-wall-degrading enzymes and peptidases. The degraded products are then taken up by the fungus to create its own cell walls and fuel its growth as it ramifies throughout the host tissue. As the pathogen growth proceeds through the host tissue it is preceded by a zone of chlorotic cells. Cell death in the host within this zone is likely to be caused by secreted molecules, which may include NEP1 and ceratoplatanin-orthologs (Chapter 7.4.6), as well as a number of novel host-specific effectors. A number of potential candidates for host-specific effectors were identified in this study however their potential roles in the pathogenicity of PmedOMT5 will require experimental validation. It is not clear at this stage if pathogenicity in *P. medicaginis* is truly host-specific or largely due to a combination of non-specific effectors that allow it to overcome plant defences through the sum of their phytotoxic effects, rather than host-specific ones. Observations of its host-specificity may in fact be due not to a host-specific effector, but in part to differences in concentrations and timing of host production of phytoalexins and the effectiveness with which *P. medicaginis* is able to successfully degrade or expel these compounds.

This chapter represents the culmination of much of the bioinformatic work presented in this thesis, which has progressed from building basic genomic resources, prediction of gene content and function to interpretation of that data in the context of plant pathogenicity. Many questions concerning the molecular mechanisms of PmedOMT5 infection remain unanswered, for which further bioinformatic analysis alone will likely not suffice and traditional experimental approaches will be required. Nevertheless, prior to undertaking this genome project we had little knowledge of which questions to ask. The insights gained from this study, in particular the prediction of candidate pathogenicity effector genes, will facilitate further bioinformatic and experimental investigations into the pathogenicity mechanisms of *P. medicaginis* and comparisons with other necrotrophic fungal pathogens of legumes.
Despite the low coverage of *P. medicaginis* transcripts *in planta* at the early-infection timepoints investigated, it was felt that sufficient data was available to make predictions of effector candidates however their expression patterns should be further examined using qRT-PCR or further RNAseq prior to experimental analysis.
Chapter 7
Examining the secretome of PmedOMT5
7.1 Introduction:

The current focus of many fungal whole-genome sequencing studies is the elucidation of the effector arsenal employed by the pathogen of interest against its hosts. Plant-pathogenic fungi produce several classes of what have come to be termed ‘effector’ molecules. Effectors are molecules that can manipulate the host-cell’s structure or function, which enables either infection of the host (collectively termed virulence factors or toxins) or triggers defence responses in the host (avirulence factors or elicitors) (Kamoun 2007). This study focuses on the former, with subsequent use of the term ‘effector’ referring to virulence factors unless otherwise indicated.

Effectors can be divided into two major classes: non-selective (non-host-specific) and host-specific and the molecules in either class can be either proteinaceous or secondary metabolites (Stergiopoulos et al. 2013). Non-selective effectors act on a broad range of host plants contributing to virulence, but not acting as determinants of host specificity. Some non-selective effectors may exhibit differential toxicity amongst plant species, but their activity is not highly correlated with the host range of their producer (Knoche and Duvick 1987). Host-specific effectors (previously known as host-specific toxins or HSTs) are toxic only to host plants and play an important role in pathogenicity as determinants of host-range (Wolpert et al. 2002). When a host-specific effector interacts with its cognate receptor it creates a response in the plant that mimics the effect of inoculation with the pathogen. These host-specific effectors may not be present in all strains of a fungal species (Jones and Dangl 2006, Liu et al. 2012), nor is the cognate receptor found in all cultivars of the host species.

Necrotrophic and biotrophic fungi are characterised by their different feeding strategies. Biotrophic fungi secrete effectors that suppress the host defence response, keeping the host cells alive so the fungus can feed on living cells undeterred. Conversely, necrotrophs secrete effectors that promote cell death so they can feed on the remains (Glazebrook 2005). The term ‘necrotrophic effectors’ has been used in some studies to describe effectors whose presence results in disease susceptibility rather than immunity, when they interact with a corresponding sensitivity gene in the host plant (Liu et al. 2012).

Detection of host-specific effectors may be challenging depending on the pathosystem and effector in question. Some pathogens are known to selectively produce host-specific
effectors only when in contact with the host plant surface. For example, *Alternaria brassicicola* and *A. panax* which produce AB-toxin and AP-toxin, respectively, only on host plants (Parada *et al.* 2008), but not in culture or on non-host plants. Other host-specific effectors such as the proteinaceous necrotrophic effector SnTox3 of *P. nodorum*, are highly expressed both in culture, as well as during the early stages of infection, with maximum expression *in planta* observed early (3 dpi for SnTox3) but decreasing once the infection is established (Liu *et al.* 2009). Furthermore, most known host-specific effectors, whether proteinaceous or secondary metabolites, exhibit slight variation in structure between isolates (Knoche and Duvick 1987, Friesen *et al.* 2006, Liu *et al.* 2012) which can affect their activity (Tan *et al.* 2012). Many of the secondary metabolite toxins exist as families of compounds, each member of which is produced in different amounts and with different potencies (Walton 1996).

The majority of known host-specific effectors are low molecular weight secondary metabolites belonging to several classes of chemical compounds. Those identified in liquid fungal culture extracts include polyketides, non-ribosomal peptides, alkaloids and terpenes (Parada *et al.* 2008), with only a small number of proteinaceous host-specific effectors having been characterised thus far (Stergiopoulos *et al.* 2013). Previous reports of characterised proteinaceous fungal effectors include proteins that are small (less than 30 kDa mature protein), soluble (lack transmembrane domains) and are cysteine-rich (provides stability in the extracellular environment) with a putative signal peptide (Bhadauria *et al.* 2011). They may also be of novel sequence (not homologous to known sequences) or highly similar to an otherwise novel sequence in an unrelated pathogen, which may be indicative of a lateral gene transfer (LGT) event or orthology. Most proteinaceous effectors also show a pattern of early expression (within 12 hpi and decreasing after 3 dpi) during infection (Liu *et al.* 2012). The first host-specific proteinaceous effector to be identified was ToxA of *P. tritici-repentis*, which induces necrosis on ToxA-sensitive wheat cultivars (Ballance *et al.* 1989, Tomas *et al.* 1990, Tuori *et al.* 1995) in the presence of the corresponding sensitivity gene *Tsn1* on chromosome 5b of wheat (Faris *et al.* 2010). The ToxA gene was also later identified in *P. nodorum* and is proposed to have been recently transferred into the *P. tritici-repentis* genome via lateral gene transfer (LGT) from *P. nodorum* (Friesen *et al.* 2006).

The mechanisms by which necrotrophic host-specific effectors cause programmed cell death (PCD) in sensitive plants are currently not well understood (Stergiopoulos *et al.*
The proteins PtrToxA and SnToxA are believed to target components of photosystem II in the chloroplast (Manning et al. 2009), although their precise binding targets are unknown. While the secondary metabolite T toxin of *C. heterostrophus* targets a mitochondrial protein in sensitive hosts (Levings et al. 1995).

Little is currently known about the molecules that mediate the interaction of *P. medicaginis* var. *medicaginis* with *Medicago* species, although the work of Kamphuis and colleagues paints a tantalising picture of the potential involvement of host-specific effectors (Kamphuis et al. 2008). This work identified two dominant susceptibility loci, in crosses of two different susceptible *M. truncatula* accessions (A17 or 3054) with the same resistant accession (SA27063) which conferred susceptibility to foliar infection with *P. medicaginis*. Research elucidating host-pathogen interactions using functional genomics in other fungal species has provided the precedent for use of a similar approach in examining the interaction between *P. medicaginis* and *M. truncatula*.

### 7.1.1 Aims for this chapter

The objectives for this chapter were to determine the optimal conditions to induce secretion in culture of active molecules by *P. medicaginis* and perform a preliminary characterisation of their attributes and host-specificity. This was followed by a bioinformatic survey of the proteins identified by mass spectrometry in the extracellular fraction that showed activity when infiltrated into *M. truncatula* plants. Proteomic, transcriptomic and bioinformatic data were combined to predict host specific effector candidates for further investigation.

### 7.2 Materials and Methods

#### 7.2.1 Optimisation of phytotoxic culture filtrate production

Various media were tested for the production of culture filtrate based on those that had been successfully used for the culture of other fungal species: Fries 2, Fries 3, minimal media or Fries 2 without yeast (F2NY) (media recipes in appendix 2.1), Fries 2 and 3 media have the same basic recipe with different micronutrient components.

Spores were either harvested fresh on the day of inoculation from PDA plates as described in chapter 2.1.9 or from previously harvested spores stored at the same concentration (1x10^6 sp/mL) in 25% glycerol at -80 °C. No significant difference was
found in the plant response to culture filtrates produced from media inoculated with spores stored in glycerols or harvested from PDA plates on the day of inoculation (data not presented). Thus stored spores were used to ensure flasks were inoculated with spores from identical sources.

Flasks were inoculated and incubated based on the method described by (Liu et al. 2004) for the production of culture filtrate from *P. nodorum*. Three 250 mL Erlemyer flasks containing 100 mL growth media (either minimal media, Fries 2, Fries 3 or F2NY) were inoculated with 1 mL of 1 x 10^6 spores/mL of *P. medicaginis* strains (PmedOMT5, WAC4741 or WAC4736) and incubated on an orbital shaker at 27 °C, 100 rpm for 3 days in the dark. Flasks were then stored at 20 °C ±2 in the dark for 2-6 weeks before harvesting of culture filtrate.

Small volumes of culture filtrate were collected by passing the liquid through a sterile disposable syringe containing glass wool to remove hyphae and then through a 0.2 µm filter to remove all spores. Larger quantities (3 pooled flasks per experiment) were strained through a Büchner funnel lined with three layers of filter paper (Whatman No.1) and two layers of Miracloth (Calbiochem), followed by a 0.2 µm vacuum filter (PALL, USA).

Culture filtrate from each experiment was plated on to half strength PDA to determine if there was any fungal growth or bacterial contamination that may account for potential false positive reactions *in planta*.

**7.2.2 Bioassay for activity of *P. medicaginis* culture filtrate on *M. truncatula***

Leaves of 3-4 week old *M. truncatula* accession SA3054 plants were infiltrated with 20-50 µL of culture filtrate or the prepared culture filtrate fraction using a needleless 1 mL syringe. The edge of the zone of infiltration was marked with a Staedtler Lumocolor permanent marker, confirmed on *M. truncatula* not to independently cause a reaction in the leaf. Infiltrated leaves were examined microspically as described in section 2.3, and compared to results observed when leaves were inoculated with GFP-transformed spores of PmedOMT5 as described in appendix 2.2.

Following media optimisation, four week old *Medicago* plants of 21 accessions that had previously been shown to have different responses to spray or spot inoculation with *P. medicaginis* spores (Ellwood et al. 2006b): SA10481, SA7749, SA18543, SA23859,
SA28645, SA27063, SA1489, SA3054, SA8623, A17, Borung, Caliph, DZA045.5, DZA315.16, F83005.5, R108-IC3, SA8604, SA11734, SA3047, Sickle and SA28375 were infiltrated with 5 week-old PmedOMT5 culture filtrate.

Infiltrations in the growth chamber were completed approximately 1 h after the start of the 12h light cycle. In the glasshouse, times varied but were recorded so they could be analysed for any differences that may be due to diurnal variations in light-levels between experiments.

Response to infiltration of culture filtrate was scored according to the following scale at 6-10 dpi: 0) no visible reaction; 1) slight chlorosis, 2) chlorosis no necrotic spotting, 3) chlorosis and/or brown spotting, and 4) strong chlorosis and brown spotting (Figure 7.1).

Figure 7.1: Scoring range used to assess response to infiltrated culture filtrate.

For each treatment in an experiment a minimum of four replicates was used and the mean and variance of the response scores were calculated. Where variances were equal, these were compared using the Tukey-Kramer HSD test at α 0.05 with JMP version 7, SAS Institute Inc.

7.2.3 Heat and pronase treatment of culture filtrate

Active culture filtrates were tested for stability of active components via: heat treatment for 2 h at 80 °C and protein degradation by treatment with 2 mg/mL pronase (Calbiochem, EMDMillipore, USA) at 37 °C for 2 h, then heat-deactivated for 20 min at 80 °C. Samples at RT were infiltrated into M. truncatula accession SA3054 along with untreated culture filtrate and a control containing 2 mg/mL heat deactivated pronase in water.
7.2.4 Infiltration of buffers and controls to test for response in *M. truncatula*

*M. truncatula* accessions were also infiltrated with other substances as controls for the various fractionation methods and to test their response. These included: water, Fries 2 media, F2NY media, minimal media, Fries 3 media, 10 mM NaAc buffer pH 5.0, 10mM sodium phosphate buffer pH 7, 10% ethanol, 1, 5, 25, 50, 100, 250, 500 and 1000 µg/mL brefeldin A (Sigma, Missouri, USA) in 0.1% ethanol, 50, 100, 250, 500 and 1000 mM NaCl, pronase-treated culture filtrates and deactivated pronase (Calbiochem, EMD Millipore, USA) in water.

7.2.5 PmedOMT5 culture-filtrate host specificity

The response of *Arabidopsis*, barley, wheat and several legume species (listed in table 2.1) to infiltration of PmedOMT5 culture filtrate, was evaluated using the bioassay described in 7.2.2 on fully expanded leaves of plants 2 - 6 weeks of age, depending on the plant species.

A minimum of four individual plants were infiltrated per species as soon as there were appropriately sized fully expanded leaves. Plants were scored 0 for no visible reaction and 1 for a visible reaction, which included chlorosis and necrosis.

7.2.6 Fractionation of the culture filtrate

Culture filtrate was size-fractionated using Amicon Ultra-15 molecular weight cut-off filters at 3, 10, 30 and 50 kDa (Millipore, Ireland), pre-rinsed with 10 mM NaPO₄ buffer pH 7. Following filtration, the concentrated protein was resuspended in a total volume one tenth of the original volume of media (10x concentration). PD-10 desalting columns (GE Healthcare, Sweden) with a bead pore exclusion limit of 5 kDa were used for desalting and buffer exchange of culture filtrate prior to ion exchange separation.

7.2.7 Ion exchange (IEX) fractionation of culture filtrate

Prior to IEX culture filtrate was dialysed in 3,500 MWCO dialysis tubing (Snakeskin Thermo Scientific) in 10 mM NaAc buffer pH 5 for 48 h with a change of buffer after 24 h.

Approximately 50 mL of dialysed active culture filtrate was loaded on to a 1 mL cation exchange column (High Trap SP FF- GE Healthcare, Sweden) at 1 mL per minute via a BIORAD Econopump system. The column was washed with start buffer (10 mM
sodium acetate (NaAc) buffer pH5.0) for 15 min then eluted with a gradient of 0-1 M NaCl over 20 minutes, collecting 1 mL fractions every minute as they were eluted from the column, followed by immediate storage on ice. UV absorbance at 280 nm was used to detect the presence of proteins. Eluted 1 mL fractions were dialysed for 48 h in 1.5 mL tubes capped with 3,500 MWCO dialysis tubing (Snakeskin Thermo Scientific) in 10 mM NaAc buffer pH 5.0 before infiltration into the plant.

Each fraction was infiltrated into a minimum of four leaves, along with control solutions of 50, 100, 250, 500 and 1000 mM NaCl and 10 mM sodium acetate buffer pH 5.0.

7.2.8  **Separation of extracellular proteins via SDS-PAGE**

The contents of dialysed fractions and whole culture filtrates from various media were visualised using SDS-PAGE. Samples (1 mL) were lyophilised O/N and resuspended in 30 µL deionised water, then heated for 10 min at 40 °C in 15 µL SDS-Bromophenol blue loading buffer to resuspend the proteins. This was followed by denaturation at 85 °C for 5 min. The resulting mixture was clarified by centrifugation at full speed for 5 min to remove precipitates and loaded on to 12% glycine or 16.5% Tris-tricine acrylamide gels prepared according to Sambrook *et al.* (1989).

7.2.9  **Statistical examination for over and under-representation of protein functional attributes in the extracellular fraction**

The number of genes with specific functional attributes (secreted proteins, GOs, Pfam domains and CAZymes) in the extracellular fraction, versus those in the version 2 protein set as a whole, were examined using Fisher’s exact test at a significance threshold of $p \leq 0.01$.

7.3  **Results**

7.3.1  **Optimisation of media for active culture filtrate production**

Initial infiltrations of PmedOMT5 culture filtrate produced in Fries2 media showed a chlorotic response to the culture filtrate. However Fries 2 media alone, also produced a chlorotic response in more than 50% of infiltrated leaves of accession SA3054 (Figure 7.2), this response was also observed in other legumes such as *Lupinus angustifolius* but was not seen in pea, chickpea or lentil (data not presented).
Further testing with culture filtrate derived from PmedOMT5 grown in a variety of media (Fries2 media without glucose, minimal media, Fries 3 media and Fries 2 media without yeast [F2NY]) for 2-5 weeks, showed that culture filtrate from PmedOMT5 grown in F2NY for 5 weeks produced the highest mean response score with the most consistent results and there was no consistent reaction to the media alone (Table 7.1). All further testing was conducted with F2NY. After 4-6 weeks growth in F2NY the pH range of the culture filtrate was typically 5 (ranging from 4.5-6.5) and the mean protein concentration was 1.1 mg/mL (ranging from 0.8-2.5 mg/ml).
Table 7.1: Mean sensitivity score of infiltrated PmedOMT5 culture filtrate at different growth times in various media.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>Growth time</th>
<th>Mean score</th>
<th>Std. deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Media alone</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>water</td>
<td>-</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>MM</td>
<td>-</td>
<td>0.3</td>
<td>0.6</td>
</tr>
<tr>
<td>Fries 2 no yeast (F2NY)</td>
<td>-</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>Fries 2</td>
<td>-</td>
<td>3.4</td>
<td>0.5</td>
</tr>
<tr>
<td>Fries 2 no sucrose</td>
<td>-</td>
<td>3.1</td>
<td>0.9</td>
</tr>
<tr>
<td>Fries 3</td>
<td>-</td>
<td>2.0</td>
<td>2.1</td>
</tr>
<tr>
<td>Minimal media</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>14 d</td>
<td>1.2</td>
<td>1.3</td>
</tr>
<tr>
<td></td>
<td>37 d</td>
<td>2.5</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>54 d</td>
<td>2.2</td>
<td>0.8</td>
</tr>
<tr>
<td>Fries 2 no yeast (F2NY)</td>
<td>14d</td>
<td>2.5</td>
<td>0.7</td>
</tr>
<tr>
<td></td>
<td>21d</td>
<td>2.8</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>28 d</td>
<td>2.1</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>35 d</td>
<td>3.0</td>
<td>1.1</td>
</tr>
<tr>
<td>Fries 2</td>
<td>35d</td>
<td>3.5</td>
<td>0.6</td>
</tr>
<tr>
<td>Fries 3</td>
<td>33d</td>
<td>3.2</td>
<td>1.5</td>
</tr>
</tbody>
</table>

*Each mean is from a minimum of 12 replicates

7.3.2 Response to brefeldin A infiltration

*P. medicaginis* has been reported to produce brefeldin A (Weber et al. 2004) and it was hypothesized that this secondary metabolite may be responsible for the reaction observed in *M. truncatula* following infiltration with *P. medicaginis* culture filtrate. Brefeldin A has been previously observed to have an effect on the growth of *M. sativa* roots and shoots (Weber et al. 2004).

Infiltration with brefeldin A produced a slight but visible change in colour in the infiltrated zone at concentrations above 500 µg/mL. The reaction did not progress to chlorosis or necrosis and did not resemble the response observed following culture filtrate infiltration (Figure 7.3c).

7.3.3 Reaction of *M. truncatula* to infiltrated *P. medicaginis* culture filtrate

The infiltration of the *P. medicaginis* culture filtrate into *M. truncatula* plants produced a visible reaction within the infiltrated zone of the leaf. This reaction ranged from a very slight chlorosis to large brown necrotic spots. Visible symptoms first appeared 2 days post infiltration (dpi) and progressed to a maximum at 8-10 dpi (Figure 7.3).
Figure 7.3: *M. truncatula* leaves 12 days post infiltration (dpi)  

a) Left leaflet: water, centre leaflet: PmedOMT5 greater than 3 kDa fraction from 37d culture, Right leaflet: PmedOMT5 culture filtrate 37d (marked with •)  
b) PmedOMT5 less than 3 kDa fraction from 37d culture  
c) 500µg brefeldin A (•), white arrow indicates marker outline of infiltrated zone.

7.3.4 Activity of culture filtrate from three *P. medicaginis* isolates

Infiltrated culture filtrate from all three *P. medicaginis* isolates (PmedOMT5, WAC4736 and WAC4741) grown in Fries 2 media for 23-42 d, produced a reaction when infiltrated into *M. truncatula* SA3054. This *Medicago* accession is susceptible to foliar inoculation with all three isolates (Ellwood *et al.* 2006b). Activity in all isolates was limited to the greater than 3 kDa fraction (Figure 7.3 a and b), when separated via molecular weight cut-off filters (data for other isolates not presented). All further testing was conducted with isolate PmedOMT5 only.

Infiltration of PmedOMT5 culture filtrate caused necrosis and the production of autofluorescent compounds in cells within the infiltrated zone of the leaf. When infiltrated leaves were examined microscopically following excitation with blue light (460-490nm), they showed a similar appearance to cells that were inoculated with PmedOMT5 spores (Figure 7.4), consisting of necrotic cells surrounded by cells producing an autofluorescent compound. Autofluorescence was not observed in cells infiltrated with water or sodium phosphate (data not shown).
Figure 7.4: Infiltrated leaves produce an autofluorescence reaction similar to that seen when leaves are inoculated with PmedOMT5 spores.

a) Micrograph of *M. truncatula* SA3054 leaf illuminated with blue light (460-490 nm) showing necrosis (brown) and autofluorescence (yellow) within the zone of infiltration (to the right of arrow), 12 dpi with PmedOMT5 culture filtrate grown in Fries2 media without yeast (F2NY), white arrow indicates marker outline surrounding infiltrated zone. b) A17 detached leaf assay (DLA) 6dpi spot inoculated with 10µL 1x10^6 PmedOMT5-GFP transformed spores, surrounded by autofluorescing cells (yellow) and healthy cells (red) c) Greater magnification of A17 DLA 6dpi spot inoculated with PmedOMT5-GFP transformed spores (green), invaded necrotic cells (brown) surrounded by autofluorescent cells (yellow) d) Confocal micrograph (200x magnification) showing *M. truncatula* SA27063 5 d after spot inoculation with GFP transformed PmedOMT5 spores (green) showing auto fluorescence (red) of penetrated cells e) Figure 2E reproduced from Kamphuis *et al.*, 2008, showing autofluorescence in SA3054 leaf 48 h after spot inoculation with PmedOMT5 stained with DiOC6 and illuminated with blue light: sp- spores, ifh- infection hyphae, p- point of penetration.

7.3.5 Investigating the stability and host-specificity of compounds produced by PmedOMT5 cultured in F2NY media

To investigate the stability of the effector molecules causing the toxicity reaction *in planta* and potential differences between *Medicago* accessions that are observed to be resistant (R) or susceptible (S) to PmedOMT5 shoot inoculation, the culture filtrate was treated with heat and/or pronase and infiltrated into accessions SA3054 (S) and SA27063 (R). Whilst heat and pronase treatment lowered the mean score in some cases, the difference was not significant via Tukey-Kramer HSD analysis *p* < 0.05. Nor was there a significant difference between the response in SA3054 (S) and SA27063 (R). In both accessions there was a significant difference in the mean response score for the
greater than 10 kDa fraction versus the whole culture filtrate and all other treatments (Table 7.2).

Early tests showed that the 10-30 kDa, greater than 30 kDa and greater than 50 kDa fractions of culture filtrate, produced a reaction in planta that was not significantly reduced by heat treatment at 80°C (data not shown). However SDS-PAGE analysis of the fractionated proteins showed that fractionation was not complete and therefore the activity can only reliably be assigned to the greater than 10 kDa fraction (data not presented).

Table 7.2: Response to infiltration in *M. truncatula* accessions SA3054 (S) and SA27063 (R) across two experiments.

<table>
<thead>
<tr>
<th>Treatment</th>
<th>3054 mean response score 8 dpi</th>
<th>3054 Std Dev</th>
<th>3054 level*</th>
<th>27063 mean response score 8 dpi</th>
<th>27063 Std Dev</th>
<th>27063 level*</th>
</tr>
</thead>
<tbody>
<tr>
<td>PmedOMT5 culture filtrate F2NY</td>
<td>1.8</td>
<td>0.6</td>
<td>B</td>
<td>2.2</td>
<td>0.7</td>
<td>B</td>
</tr>
<tr>
<td>Pronase treated CF</td>
<td>1.2</td>
<td>0.4</td>
<td>BC</td>
<td>2.0</td>
<td>0.8</td>
<td>B</td>
</tr>
<tr>
<td>Heat treated CF</td>
<td>1.9</td>
<td>0.9</td>
<td>B</td>
<td>2.0</td>
<td>0.8</td>
<td>B</td>
</tr>
<tr>
<td>&gt;10 kDa fraction F2NY</td>
<td>3.9</td>
<td>0.4</td>
<td>A</td>
<td>4.0</td>
<td>0.0</td>
<td>A</td>
</tr>
<tr>
<td>&lt;10 kDa fraction F2NY</td>
<td>0.9</td>
<td>0.3</td>
<td>CD</td>
<td>0.5</td>
<td>1.0</td>
<td>C</td>
</tr>
<tr>
<td>10 mM sodium phosphate buffer</td>
<td>0.6</td>
<td>0.5</td>
<td>CDE</td>
<td>0.3</td>
<td>0.5</td>
<td>C</td>
</tr>
<tr>
<td>Water</td>
<td>0.4</td>
<td>0.8</td>
<td>DE</td>
<td>0.0</td>
<td>0.0</td>
<td>C</td>
</tr>
<tr>
<td>Pronase treated water control</td>
<td>0.0</td>
<td>0.0</td>
<td>E</td>
<td>0.0</td>
<td>0.0</td>
<td>C</td>
</tr>
</tbody>
</table>

Experiment 2:

<table>
<thead>
<tr>
<th>Treatment</th>
<th>3054 mean response score 8 dpi</th>
<th>3054 Std Dev</th>
<th>3054 level*</th>
<th>27063 mean response score 8 dpi</th>
<th>27063 Std Dev</th>
<th>27063 level*</th>
</tr>
</thead>
<tbody>
<tr>
<td>PmedOMT5 CF F2NY</td>
<td>2.5</td>
<td>0.5</td>
<td>-</td>
<td>2.4</td>
<td>0.8</td>
<td>-</td>
</tr>
<tr>
<td>Pmed OMT5CF F2NY after storage at 4°C on ice for 6 weeks</td>
<td>2.4</td>
<td>0.5</td>
<td>-</td>
<td>2.4</td>
<td>0.8</td>
<td>-</td>
</tr>
</tbody>
</table>

*Tukey-Kramer HSD levels not connected by the same letter are significantly different from one another at p < 0.05.

The mean SA3054 response to infiltrated culture filtrate from PmedOMT5 cultured in F2NY was 2.4 over 14 independent experiments. No significant difference was seen in the mean response score after the culture filtrate was stored at 4 °C on ice for 6 weeks (Table 7.2).

7.3.6 Response of *M. truncatula* accessions to infiltration with PmedOMT5 culture filtrate

Sensitivity to PmedOMT5 culture filtrate was examined in a range of *M. truncatula* accessions with known responses to PmedOMT5 foliar inoculation (Table 7.3).
Sensitivity to culture filtrate and susceptibility to PmedOMT5 inoculation showed no correlation.

Table 7.3: Mean response score of *M. truncatula* accessions infiltrated with whole culture filtrate.

<table>
<thead>
<tr>
<th>Accession/cultivar</th>
<th>Mean score</th>
<th>Std dev</th>
<th>Level</th>
<th>Response to PmedOMT5 foliar inoculation</th>
</tr>
</thead>
<tbody>
<tr>
<td>SA23859</td>
<td>4.0</td>
<td>0.0</td>
<td>A</td>
<td>R</td>
</tr>
<tr>
<td>SA7749</td>
<td>4.0</td>
<td>0.0</td>
<td>AB</td>
<td>S</td>
</tr>
<tr>
<td>SA1489</td>
<td>3.7</td>
<td>0.4</td>
<td>ABC</td>
<td>R</td>
</tr>
<tr>
<td>SA27063</td>
<td>3.5</td>
<td>0.8</td>
<td>ABCD</td>
<td>R</td>
</tr>
<tr>
<td>SA28645</td>
<td>3.2</td>
<td>1.2</td>
<td>ABCDE</td>
<td>S</td>
</tr>
<tr>
<td>SA18543</td>
<td>3.2</td>
<td>0.9</td>
<td>ABCDE</td>
<td>S</td>
</tr>
<tr>
<td>R108-1C3</td>
<td>3.0</td>
<td>0.9</td>
<td>ABCDEF</td>
<td>S*</td>
</tr>
<tr>
<td>DZA315.16</td>
<td>2.9</td>
<td>0.9</td>
<td>ABCDEF</td>
<td>MR</td>
</tr>
<tr>
<td>SA8604</td>
<td>2.9</td>
<td>0.8</td>
<td>ABCDEF</td>
<td>S</td>
</tr>
<tr>
<td>SA11734</td>
<td>2.6</td>
<td>1.2</td>
<td>BCDEF</td>
<td>S</td>
</tr>
<tr>
<td>Caliph</td>
<td>2.6</td>
<td>1.1</td>
<td>BCDEF</td>
<td>-</td>
</tr>
<tr>
<td>DZA045.5</td>
<td>2.4</td>
<td>0.8</td>
<td>CDEF</td>
<td>MR</td>
</tr>
<tr>
<td>SA3047</td>
<td>2.4</td>
<td>1.2</td>
<td>CDEF</td>
<td>R</td>
</tr>
<tr>
<td>A1</td>
<td>2.4</td>
<td>1.0</td>
<td>CDEF</td>
<td>S</td>
</tr>
<tr>
<td>F803005.5</td>
<td>2.3</td>
<td>0.8</td>
<td>DEF</td>
<td>S*</td>
</tr>
<tr>
<td>SA10481</td>
<td>2.2</td>
<td>1.0</td>
<td>DEF</td>
<td>R</td>
</tr>
<tr>
<td>SA3054</td>
<td>2.1</td>
<td>1.0</td>
<td>DEF</td>
<td>S</td>
</tr>
<tr>
<td>Sickle</td>
<td>2.1</td>
<td>1.0</td>
<td>CDEF</td>
<td>S</td>
</tr>
<tr>
<td>SA28375</td>
<td>2.0</td>
<td>0.9</td>
<td>EF</td>
<td>R</td>
</tr>
<tr>
<td>Borung</td>
<td>1.9</td>
<td>1.2</td>
<td>EF</td>
<td>S</td>
</tr>
<tr>
<td>SA8623</td>
<td>1.6</td>
<td>0.5</td>
<td>F</td>
<td>R</td>
</tr>
</tbody>
</table>

Levels not connected by the same letter are significantly different from one another *p* < 0.05
Response to foliar inoculation based on Ellwood *et al.*, 2006, *Djebali et al.*, 2013 or *this study,* S=sensitive, R= resistant, MR=moderately resistant

7.3.7 Evaluating the host specificity of PmedOMT5 culture filtrate

A range of plant species were infiltrated with whole PmedOMT5 culture filtrate to assess their sensitivity. Responses in non-*Medicago* plants were scored as sensitive or insensitive. Sensitivity was observed as chlorosis with some brown spotting, which occasionally progressed to necrosis over time (Figure 7.5a). Several of the plants with sensitive responses are not known to be hosts to *P. medicaginis.*

*M. sativa* produced responses similar to those seen in *M. truncatula* (Figure 7.5b). All *M. sativa* accessions except Sceptre were sensitive, with no significant difference in mean scores (data not shown). Sceptre is listed as moderately resistant to *Phoma* spp. in studies by the Department of Food and Agriculture WA (Loo *et al.* undated).
Table 7.4: Sensitivity of various plant species to infiltration with PmedOMT5 culture filtrate.

<table>
<thead>
<tr>
<th>Plant species</th>
<th>Cultivar/Accession/Ecotype</th>
<th>Response to infiltration</th>
<th>Response to <em>P. medicaginis</em> foliar inoculation</th>
<th>Literature reports of <em>P. medicaginis</em> pathogenicity on genus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pea (<em>Pisum sativum</em>)</td>
<td>Kaspa, Earlicrop Massey, Greenfeast, Parafield</td>
<td>y</td>
<td>S</td>
<td>y&lt;sup&gt;a,c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Dwarf bean (<em>Phaseolus vulgaris</em>)</td>
<td>Borlotti, Bountiful butter</td>
<td>n</td>
<td>-</td>
<td>y&lt;sup&gt;a,c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Sub clover (<em>Trifolium subterraneum</em>)</td>
<td>Dalkeith</td>
<td>y</td>
<td>-</td>
<td>y&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Faba bean (<em>Vicia faba</em>)</td>
<td>Fiord</td>
<td>n</td>
<td>-</td>
<td>y&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>Mung bean (<em>Vigna radiata</em>)</td>
<td>unknown (from health food store)</td>
<td>n</td>
<td>-</td>
<td>y&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td>White lupin (<em>Lupinus albus</em>)</td>
<td>P27174, Kiev</td>
<td>y</td>
<td>S</td>
<td>y&lt;sup&gt;b&lt;/sup&gt;</td>
</tr>
<tr>
<td>Narrow-leaf lupin (<em>L. angustifolius</em>)</td>
<td>Uncrop, Tanjil</td>
<td>y</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Alfalfa (<em>Medicago sativa</em>)</td>
<td>36325, 35043, 38082, 36442, 10119</td>
<td>y</td>
<td>-</td>
<td>y&lt;sup&gt;a,c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Alfalfa <em>Medicago sativa</em></td>
<td>Sceptre</td>
<td>n</td>
<td>-</td>
<td>y&lt;sup&gt;a,b,c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Chickpea (<em>Cicer arietinum</em>)</td>
<td>Sona, Genesis 509, Genesis 090</td>
<td>y (weak)</td>
<td>S</td>
<td>y&lt;sup&gt;a,c&lt;/sup&gt;</td>
</tr>
<tr>
<td>Lentil (<em>Lens culinaris</em>)</td>
<td>Digger, Northfield</td>
<td>n</td>
<td>S</td>
<td>y&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td><em>Arabidopsis thaliana</em></td>
<td>Col-0</td>
<td>n</td>
<td>S</td>
<td>y&lt;sup&gt;a&lt;/sup&gt;</td>
</tr>
<tr>
<td><em>Lotus japonicus</em></td>
<td>Gifu</td>
<td>y (weak)</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Wheat (<em>Triticum aestivum</em>)</td>
<td>Zippy</td>
<td>n</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Barley (<em>Hordeum vulgare</em>)</td>
<td>Flagship</td>
<td>y</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

Response to foliar inoculation based on this study: R = resistant, S = susceptible. a- (Kinsey 2002) b- (Ellwood *et al.* 2006a) c- (Djebari 2013)
In all bean varieties infiltrated there was a small dark region directly beneath the infiltration site that became brown and appeared necrotic, however this reaction was never seen to spread to the boundaries of the infiltrated zone as in other sensitive plant species (figure 7.5 c). This reaction was not observed following infiltration of water or F2NY media.

Figure 7.5: Responses of non-Medicago species to PmedOMT5 culture filtrate.

- a) Pea ‘Earlicrop Massey’ 8dpi
- b) M. sativa 36325 6 dpi
- c) Dwarf bean Borlotti 10 dpi
- d) Barley ‘Flagship’

‘Flagship’ 1 • indicates culture filtrate, •• indicates F2NY media, unlabeled circle on leaf indicates water infiltration.

7.3.8 Fractionation of culture filtrate by size and ion exchange chromatography

Cation exchange separation of dialysed 35 d PmedOMT5 culture filtrate into 20 fractions over a gradient of 0 - 1M NaCl showed some separation of proteins. The mean score from whole culture filtrate infiltration was 3.5 (±0.6). No reactions were seen in response to 10 mM sodium acetate buffer infiltration. Mean scores for fractions ranged between 0.5 and 4, however response by the plant appeared to coincide with protein concentration rather than specific protein bands. The unbound fraction also showed strong activity. Several other separations were performed using different salt gradients for both cation and anion exchange columns but significant separation was not achieved (data not presented).

SDS-PAGE analysis of the PmedOMT5 culture filtrate showed a high concentration of proteins with a molecular weight of greater than 40 kDa (Figure 7.6) as had been observed with separation using molecular weight cut-off filters (Section 7.3.5).

7.3.9 Proteins detected in the culture filtrate via IEX-LC MALDI TOF

As described in section 5.2.3 the >3 kDa fraction of 35 day old PmedOMT5 culture filtrate with a mean score of 3.5 on accession SA27063 and 3.1 on cultivar A17 was submitted to peptide sequencing. The resulting peptide fragments were mapped to the 6-frame translation of the genome and version 2 proteins as described in chapter 5.2.6.
This resulted in the identification of 170 proteins that were supported by mapped peptide data with MUDPIT scores above the confidence limit (p < 0.05). The majority of the detected extracellular proteins were predicted to be secreted, expressed in planta or contain CAZyme domains (summarised in Figure 7.7 and Table 7.5).

![Figure 7.6: SDS-PAGE separation of PmedOMT5 culture filtrate.](image)

From 33 d culture in F2NY (CF) and greater than 3.5 kDa fraction of culture filtrate (>3.5 kDa), low molecular weight marker (LM), high molecular weight marker (HM).

![Figure 7.7: Functional attributes associated with the 170 proteins identified by peptide matches from the extracellular protein fraction.](image)
Table 7.5: Summary of functional attributes of proteins detected in the extracellular protein fraction.

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Count in extracellular fraction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identified by at least one peptide match</td>
<td>170</td>
</tr>
<tr>
<td>6-frame genome translation support by one or more peptides</td>
<td>117</td>
</tr>
<tr>
<td>Identified by two or more peptides from the 6-frame fraction</td>
<td>53</td>
</tr>
<tr>
<td>Predicted to be secreted by all three methods</td>
<td>96</td>
</tr>
<tr>
<td>Predicted to be secreted by one or more methods</td>
<td>115</td>
</tr>
<tr>
<td>CAZyme annotation</td>
<td>26</td>
</tr>
<tr>
<td>Expressed in planta (five or more mapped reads)</td>
<td>113</td>
</tr>
<tr>
<td>More highly expressed in planta</td>
<td>20</td>
</tr>
<tr>
<td>EST match</td>
<td>4</td>
</tr>
<tr>
<td>PHIbase pathogenicity-related match</td>
<td>33</td>
</tr>
<tr>
<td>Predicted molecular weight of less than 40 kDa</td>
<td>66</td>
</tr>
<tr>
<td>Number with pathogenicity score of 25 or over</td>
<td>34</td>
</tr>
</tbody>
</table>

Selection based on pathogenicity attributes as described in chapter 6.2.12 that includes only those proteins with peptides identified in the extracellular fraction resulted in a short-list of five effector candidates: PmedOMT5_11737, PmedOMT5_10797, PmedOMT5_10389, PmedOMT5_00021 and PmedOMT5_05568 (Table 7.6). All of these proteins have BLASTP matches to fungal proteins in the NCBI protein database, with bit scores ranging from 55.8 to 399. None have conserved Pfam domains or matches to proteins with characterised roles in pathogenicity in other species in PHIbase.

Four of the proteins identified in the extracellular fraction have no BLASTP match to the NCBI protein database making them Phoma-specific. But none of their encoding genes had detectable levels of expression in planta, making them less likely to be pathogenicity effector candidates, although because of the low levels of fungal reads obtained from the IP RNAseq they were still considered as potential candidates. Eight genes have no orthologs in the species tested. Only one of these, PmedOMT5_05765 was detected as expressed IP, but this gene also had a BLASTP match to a P. nodorum protein [NCBI:AT85189.2] hypothetical protein SNOG_07723 ] bit score 283, and contained a conserved ubiquitin-3-binding protein domain [Pfam: PF09792]. Another of these PmedOMT5 unique genes belongs to the pectate lyase family that is expanded in PmedOMT5.
One other effector candidate, identified in the extracellular fraction, as well as more highly expressed \textit{in planta} than under the other conditions tested, was \textit{PmedOMT5}_03659, which had a pathogenicity score of 36 and contains a Pfam domain associated with fungal cytotoxins (discussed in 6.4.15).
Table 7.6: Top effector candidates identified in the extracellular fraction.

<table>
<thead>
<tr>
<th>Protein ID</th>
<th>4d MM FPKM</th>
<th>16d MM FPKM</th>
<th>35d MM FPKM</th>
<th>IP FPKM</th>
<th>NCBI top hit accession and species</th>
<th>BLASTP to nr significance and bit score</th>
<th>aa length</th>
<th>Predicted molecular weight (kDa)</th>
<th>Percent cysteine in mature protein</th>
<th>Cysteine count in mature protein</th>
<th>Predicted mature protein length</th>
<th>Effector score</th>
<th>Number of species with orthologs</th>
</tr>
</thead>
<tbody>
<tr>
<td>PmedOMT5_11737</td>
<td>15.8</td>
<td>6.1</td>
<td>126.5</td>
<td>568.7</td>
<td>ELA34145 Colletotrichum gloeosporioides Nara gc5</td>
<td>8E^-08, 55.8</td>
<td>148</td>
<td>15.4</td>
<td>1.5</td>
<td>2</td>
<td>134</td>
<td>38</td>
<td>3</td>
</tr>
<tr>
<td>PmedOMT5_10797</td>
<td>19.3</td>
<td>34.0</td>
<td>445.6</td>
<td>331.4</td>
<td>XP_003836972 Leptosphaeria maculans JN3</td>
<td>2E^-58, 186</td>
<td>157</td>
<td>16.9</td>
<td>3.5</td>
<td>5</td>
<td>141</td>
<td>37</td>
<td>12</td>
</tr>
<tr>
<td>PmedOMT5_05568</td>
<td>619.6</td>
<td>980.3</td>
<td>2656.9</td>
<td>2397.0</td>
<td>EMD64635.1 Bipolaris sorokiniana</td>
<td>3E^-82, 399</td>
<td>305</td>
<td>32.8</td>
<td>3.8</td>
<td>11</td>
<td>291</td>
<td>36</td>
<td>11</td>
</tr>
<tr>
<td>PmedOMT5_10389</td>
<td>8.3</td>
<td>46.1</td>
<td>214.7</td>
<td>279.5</td>
<td>XP_001800655 P. nodorum SN15</td>
<td>1E^-138, 399</td>
<td>293</td>
<td>31.8</td>
<td>0.7</td>
<td>2</td>
<td>280</td>
<td>36</td>
<td>20</td>
</tr>
<tr>
<td>PmedOMT5_00021</td>
<td>1.1</td>
<td>123.6</td>
<td>2.5</td>
<td>294.5</td>
<td>XP_001792580 P. nodorum SN15</td>
<td>2E^-91, 279</td>
<td>210</td>
<td>22.6</td>
<td>1.5</td>
<td>3</td>
<td>195</td>
<td>36</td>
<td>18</td>
</tr>
</tbody>
</table>

FPKM - expected number of fragments per kilobase of transcript sequence, per millions base pairs sequenced. nr- NCBI GenBank non-redundant protein database. Mature protein – predicted product after cleavage of predicted signal peptide. *Effector score calculated as described in 6.2.12
7.3.10 Analysis of over-representation of functional domains and attributes in the extracellular fraction.

Predicted secreted proteins, genes expressed *in planta*, and genes more highly expressed *in planta*, were found to be significantly more abundant in the extracellular fraction when compared to the proteome as a whole (Fisher’s exact test, significance threshold of $p < 0.01$).

Table 7.7: Categories of genes over-represented in the extracellular fraction.

<table>
<thead>
<tr>
<th>Category</th>
<th>Number in extracellular (total 170)</th>
<th>Number in total protein set (total 11,879)</th>
<th>$p$ value</th>
</tr>
</thead>
<tbody>
<tr>
<td>More highly expressed <em>in planta</em></td>
<td>20</td>
<td>461</td>
<td>1.03E-05</td>
</tr>
<tr>
<td>Expressed <em>in planta</em></td>
<td>113</td>
<td>5,778</td>
<td>2.00E-06</td>
</tr>
<tr>
<td>Predicted to be secreted by all three methods</td>
<td>97</td>
<td>1,020</td>
<td>3.18E-56</td>
</tr>
</tbody>
</table>

Two CAZy domains were also found to be significantly over-represented in the extracellular fraction. The first of these was CBM43, members of which have been demonstrated to bind β-1,3-glucans. The second was GH16, which cleaves glycosidic bonds in a range of glucans and galactans.

Over-represented Pfam domains included PF12708:Pectate_lyase_3 found in pectate lyases, which act as virulence factors that degrade the pectin components of plant cell walls (Mayans *et al.* 1997). There was also an increase in the number of oxidoreductase enzymes, as well as, proteins with an X8 domain, which is synonymous with the CBM43 CAZy domain and corroborated as significantly over-represented by the previously described CAZy analysis. Table 7.8 and Table 7.9 outline the Pfam domains and GOs that were more highly represented in the extracellular fraction when compared to the entire protein set. There were significantly more proteins with GO terms related to serine and glycine metabolism, hydrolase activity and extracellular localisation. There were also significantly lower numbers of genes identified in the extracellular fraction compared to the gene set as a whole with GOs characterised as zinc ion binding or as integral to membranes.
Table 7.8: Pfam domains significantly over-or under-represented in the extracellular protein fraction versus the total protein set.

<table>
<thead>
<tr>
<th>Pfam domain name</th>
<th>Pfam</th>
<th>Description</th>
<th>Number in extracellular (total 170)</th>
<th>No. in total protein set (total 11,879)</th>
<th>p value</th>
<th>Number expected to occur by chance in a sample size of 170</th>
</tr>
</thead>
<tbody>
<tr>
<td>BBE</td>
<td>PF08031.7</td>
<td>Berberine bridge enzyme</td>
<td>4</td>
<td>17</td>
<td>0.00019</td>
<td>0.2</td>
</tr>
<tr>
<td>But2</td>
<td>PF09792.4</td>
<td>Ubiquitin 3 binding protein But2 C-terminal domain</td>
<td>2</td>
<td>4</td>
<td>0.00285</td>
<td>0.1</td>
</tr>
<tr>
<td>Copper-bind</td>
<td>PF00127.15</td>
<td>Plastocyanin family of copper binding proteins protease</td>
<td>2</td>
<td>3</td>
<td>0.00193</td>
<td>0.1</td>
</tr>
<tr>
<td>DJ-1_PfpI</td>
<td>PF01965.19</td>
<td>Plastocyanin family of copper binding proteins protease</td>
<td>2</td>
<td>8</td>
<td>0.00807</td>
<td>0.1</td>
</tr>
<tr>
<td>FAD_binding_4</td>
<td>PF01565</td>
<td>Fad binding domain</td>
<td>5</td>
<td>40</td>
<td>0.00038</td>
<td>0.1</td>
</tr>
<tr>
<td>Ferritin_2</td>
<td>PF13668.1</td>
<td>Ferritin-like domain</td>
<td>2</td>
<td>4</td>
<td>0.00285</td>
<td></td>
</tr>
<tr>
<td>Glyco_hydro_2_C</td>
<td>PF02836.12</td>
<td>GH2 glycohydrolase glucose-methanol-choline oxidoreductase family-C-terminal steroid binding domain</td>
<td>2</td>
<td>8</td>
<td>0.00807</td>
<td>0.6</td>
</tr>
<tr>
<td>GMC_oxred_C</td>
<td>PF05199.8</td>
<td>glucose-methanol-choline oxidoreductase family-C-terminal steroid binding domain</td>
<td>4</td>
<td>30</td>
<td>0.0012</td>
<td>0.1</td>
</tr>
<tr>
<td>GMC_oxred_N</td>
<td>PF00732.14</td>
<td>glucose-methanol-choline oxidoreductase family-FAD ADP binding N terminal domain</td>
<td>4</td>
<td>30</td>
<td>0.0012</td>
<td>0.4</td>
</tr>
<tr>
<td>Pectate_lyase_3</td>
<td>PF12708.2</td>
<td>Pectate lyase family related to CAZy GH28</td>
<td>3</td>
<td>7</td>
<td>0.0031</td>
<td>5.4</td>
</tr>
<tr>
<td>WD40</td>
<td>PF00400.27</td>
<td>Domain involved in signal transduction</td>
<td>0</td>
<td>170</td>
<td>0.00164</td>
<td>5.4</td>
</tr>
<tr>
<td>X8</td>
<td>PF07983.8</td>
<td>X8 domain/CBM43</td>
<td>2</td>
<td>2</td>
<td>0.00117</td>
<td>0.1</td>
</tr>
</tbody>
</table>
Table 7.9: Gene Ontologies (GOs) significantly over- or under-represented in the extracellular fraction versus the total protein set.

<table>
<thead>
<tr>
<th>GO number</th>
<th>GO Process</th>
<th>GO description</th>
<th>Number in extracellular (total 170)</th>
<th>Number in total protein set (total 11,879)</th>
<th>p value</th>
<th>Number expected by chance in a sample the size of the extracellular fraction (170)</th>
</tr>
</thead>
<tbody>
<tr>
<td>GO:0006066 P</td>
<td>alcohol metabolic process</td>
<td>4</td>
<td>24</td>
<td>0.000561</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>GO:0006040 P</td>
<td>amino sugar metabolic process</td>
<td>4</td>
<td>51</td>
<td>0.006421</td>
<td>0.73</td>
<td></td>
</tr>
<tr>
<td>GO:0004190 F</td>
<td>aspartic-type endopeptidase activity</td>
<td>3</td>
<td>24</td>
<td>0.005773</td>
<td>0.34</td>
<td></td>
</tr>
<tr>
<td>GO:0007047 P</td>
<td>cellular cell wall organization</td>
<td>2</td>
<td>2</td>
<td>0.001155</td>
<td>0.03</td>
<td></td>
</tr>
<tr>
<td>GO:0008812 F</td>
<td>choline dehydrogenase activity</td>
<td>4</td>
<td>23</td>
<td>0.000488</td>
<td>0.33</td>
<td></td>
</tr>
<tr>
<td>GO:0005576 C</td>
<td>extracellular region</td>
<td>7</td>
<td>74</td>
<td>0.000124</td>
<td>1.06</td>
<td></td>
</tr>
<tr>
<td>GO:0050660 F</td>
<td>flavin adenine dinucleotide binding</td>
<td>8</td>
<td>128</td>
<td>0.000548</td>
<td>1.83</td>
<td></td>
</tr>
<tr>
<td>GO:0006544 P</td>
<td>glycine metabolic process</td>
<td>4</td>
<td>56</td>
<td>0.008562</td>
<td>0.80</td>
<td></td>
</tr>
<tr>
<td>GO:0016787 F</td>
<td>hydrolase activity</td>
<td>10</td>
<td>216</td>
<td>0.001015</td>
<td>3.09</td>
<td></td>
</tr>
<tr>
<td>GO:0016021 C</td>
<td>integral to membrane</td>
<td>3</td>
<td>735</td>
<td>0.004637</td>
<td>10.52</td>
<td></td>
</tr>
<tr>
<td>GO:0006563 P</td>
<td>L-serine metabolic process</td>
<td>4</td>
<td>55</td>
<td>0.008104</td>
<td>0.79</td>
<td></td>
</tr>
<tr>
<td>GO:0004650 F</td>
<td>polygalacturonase activity</td>
<td>2</td>
<td>9</td>
<td>0.009592</td>
<td>0.13</td>
<td></td>
</tr>
<tr>
<td>GO:0004252 F</td>
<td>serine-type endopeptidase activity</td>
<td>5</td>
<td>31</td>
<td>0.00013</td>
<td>0.44</td>
<td></td>
</tr>
<tr>
<td>GO:0006566 P</td>
<td>threonine metabolic process</td>
<td>4</td>
<td>54</td>
<td>0.00766</td>
<td>0.77</td>
<td></td>
</tr>
<tr>
<td>GO:0008762 F</td>
<td>UDP-N-acetylMuramuramase dehydrogenase activity</td>
<td>4</td>
<td>41</td>
<td>0.003224</td>
<td>0.59</td>
<td></td>
</tr>
<tr>
<td>GO:0008270 F</td>
<td>zinc ion binding</td>
<td>1</td>
<td>568</td>
<td>0.002161</td>
<td>8.13</td>
<td></td>
</tr>
</tbody>
</table>
7.3.11 PHIbase matches to proteins identified in the extracellular fraction

Thirty-three PmedOMT5 proteins identified in the extracellular fraction were found to have BLAST matches (below a significance threshold of $10^{-5}$) to genes in the Pathogen-Host Interaction database (PHIbase) (Winnenburg et al. 2006). Only those proteins with matches to genes whose experimental disruption produced a reduction in virulence or loss of pathogenicity in planta were considered (data on all PHI matches is contained in appendix 5.3). Of the 33 genes with PHIbase matches 28 are expressed in planta and 13 of these have KEGG enzyme codes (ECs) that were assigned by Blast2GO. The putative roles of these ECs included: cell wall degrading enzymes (10 glycohydrolases, including: GH37 trehalase, GH 28, GH16, GH88, and GH72), serine kinases (2), MFS transporters (2), glycosylases, chitin binding, a 1,3-beta-glucanosyltransferase and a MAP kinase.

Of those genes with PHIbase matches that were also expressed in planta PmedOMT5_01156, a beta-1,3-glucanosyltransferase, had a match to a Fusarium oxysporum gene (PHI:522) that is required for virulence on tomato fruits and roots (Caracuel 2005). This protein has a proposed role in fungal cell wall biosynthesis within the plant host. Another, PmedOMT5_02144, has a strong match to a gene encoding a serine/threonine-protein kinase protein from F. graminearum ([PHI:1193], [UniProt:I1RGD6], Fg02795.1, a homolog of S. cerevisiae SKY1), disruption of which caused a reduction in virulence on wheat, reduced vegetative spore counts and defective sexual spore production (Wang et al. 2011).

PmedOMT5_10285, matches to a ribonuclease T2 gene (MGG_10510 PHI:811) identified as playing a role in pathogenicity in M. oryzae (Jeon et al. 2007). Disruption of this gene by T-DNA insertion had the effect of reducing virulence. This type of secreted RNase is expressed by plants in response to wounding or pathogen invasion and has a role scavenging phosphate [CDD: cd01061]. It was expressed by PmedOMT5 in planta but not increased in comparison to the other conditions examined.

Another five genes had matches to genes categorised in PHIbase as loss of pathogenicity mutation. These included PmedOMT5_01469, which matches a putative glyoxaloxidase molecule with a chitin-binding domain (PHI: 352 bit score 201), whose disruption causes loss of pathogenicity in Ustilago maydis (Leuthner et al. 2005). This protein also contains a hevein chitin binding domain. Glyoxaloxidase enzymes catalyse
the oxidation of aldehydes to carboxylic acids, coupled with the reduction of dioxygen to hydrogen peroxide and have an essential role in the degradation of extracellular lignin in the wood-rot fungus *Phaenerochaete chrysosporium* (Whittaker *et al.* 1999).

One gene (*PmedOMT5_06610*) with a high number of proteogenomic peptide matches (7) and containing a GMC oxidoreductase domain [Pfam: PF05199.8], was also identified in the EST set from *Phoma* inoculated leaves. This gene also has a match to an alcohol oxidase gene in *Passalora fulva* (*AOX1*, PHI:199, bit score 144), the deletion of which resulted in greatly reduced pathogenicity on tomato (*Segers et al.* 2001).

*PmedOMT5_01552* was expressed in planta and has a match to a cerato-platanin domain [Pfam:PF07249] containing protein (*L. maculans* gene *sp1*, PHI:695, bit score- 145). The protein products of orthologous genes from several fungal species were found to induce both necrosis and phytoalexin synthesis in many plant species. However, *L. maculans sp1* mutants were characterised as unaffected in pathogenicity by gene disruption with a hygromycin resistance gene insert (*Wilson et al.* 2002).

### 7.4 Discussion

#### 7.4.1 Optimisation of culture filtrate production

Various media previously used for the successful production of effectors in closely related fungal species were tested to determine if they also induced the production of compounds in the culture filtrate of PmedOMT5 that were active in planta. These included: minimal media, Fries 2 and Fries 3 and a combination of Fries 2/3 micronutrients (Fries media are the same basic recipe with different combination of micronutrients-listed in appendix 2.1). Related Dothideomycete crop pathogens, *P. nodorum* (*Liu et al.* 2004), *P. teres f. teres* and *P. teres f. maculata* (*Sarpeleh et al.* 2007) and *P. tritici-repentis* (*Strelkov et al.* 1999) all produce active phytotoxic culture filtrates in Fries 2 media. However, initial tests showed that some *M. truncatula* accessions were sensitive to a component of the Fries 2 media. Other researchers have observed a similar response to un-inoculated media infiltration which they attributed to osmotic damage from the high concentration of unconsumed sugars in the media (*Tomas and Bockus 1987*). But in the case of *M. truncatula*, the response was determined to be coming from the yeast extract. This conclusion was based on the fact that fractionated media showed a response to the greater than 50 kDa fraction of the
media only and elimination of sugar from the media did not reduce the effect (data not shown), nor was the reaction observed following infiltration with minimal media which contains a similar amount of sucrose but no yeast extract. *M. truncatula* root-cell cultures are known to produce a defence response, including up-regulation of isoflavonoid production, in the presence of yeast elicitors based on metabolomic and microarray analysis (Suzuki *et al.* 2005, Farag *et al.* 2008).

This background response to yeast media was more pronounced in accession SA3054, the susceptible line used for testing the activity of the culture filtrate, than other *M. truncatula* lines. It was also observed in *Lupinus albus* Kiev mutant (data not shown). Tomas and Bockus (1987) noted a decrease in media response in older leaves, but variation in response to leaf age was not tested in this study. Removing yeast extract from the media greatly reduced the response to media but also slightly lowered the mean score from the culture filtrate. Nevertheless, response scores to the culture filtrates from F2NY cultures were still the highest and most consistent of all samples tested.

### 7.4.2 Activity of culture filtrate on *M. truncatula*

All of the three isolates tested (PmedOMT5, WAC4741 and WAC4736) produced similar phytotoxic responses upon culture filtrate infiltration into SA3054, an accession susceptible to foliar inoculation with these isolates (Ellwood *et al.* 2006b). PmedOMT5 grown in F2NY for 4-6 weeks consistently produced the strongest response when infiltrated into SA3054, with no consistent response to the control media alone. Infiltration of 5 week old culture filtrate into other *M. truncatula* accessions produced a range of responses that had no correlation to previously observed responses to foliar inoculation.

Insect damage from thrips, which appeared as white flecks on the leaves, was occasionally observed in glasshouse experiments. Interestingly, this was most often observed within the infiltrated zone. It is possible that the thrips may have been attracted by unconsumed sugars in the infiltrated media or by a volatile the plant may have been emitting (or had ceased to produce) in response to the infiltration.
7.4.3 Host specificity of culture filtrate

Infiltration of the culture filtrate caused a response in many plant species that did not correspond with the known host range of *P. medicaginis*, either observed in this study or described in the literature. Responses were detected in plants, which *P. medicaginis* is not known to infect, such as wheat and barley. In alfalfa, the cultivar Sceptre was the only one which appeared insensitive to infiltration with the culture filtrate, although, this cultivar was previously determined to be highly susceptible to PmedOMT5 foliar infection.

Previous studies have described *P. medicaginis* var. *medicaginis* (Ellwood et al., 2006) and another variant *P. medicaginis* var. *macrospora* (Boerema et al. 2004) as host-specific pathogens limited to the infection of *Medicago* species and lupin. Other authors, however, have described *P. medicaginis* variants as having broad host ranges, though possibly as secondary or opportunistic pathogens (Kinsey 2002), and have observed these variants on many legumes including clover, pea and lentil, as well as, non-legumes including flax and canola (Lakshmanan and Vanterpool 1967, Boerema et al. 2004). *P. medicaginis* has also been observed to grow on wheat and barley seeds (Renfro and Wilcoxson 1963). The precise phylogenetic placement of some of the species described as *P. medicaginis* is not always clear, as classification of the genus *Phoma* was originally based on morphology in culture, but has recently been restructured based on molecular data (Aveskamp et al. 2010, de Gruyter et al. 2013). Few of the previously described studies describing *P. medicaginis* as a broad host-range pathogen included any molecular data, and based fungal identification on morphological features only.

A possible explanation for the lack of host specificity of the culture filtrate is that PmedOMT5 produces non-selective effectors, which can cause a response on multiple hosts. There may also be additional factors, such as variation in the ability of different species or cultivars to recognise *P. medicaginis* and initiate host defences. Indeed, it could even be speculated that host range in *Phoma* is largely determined by its susceptibility to plant defences, as opposed to susceptibility of plant hosts to its effector arsenal. Another explanation for the strong responses on non-host plants may be that some secreted proteins, particularly cell wall degrading enzymes, may not normally
penetrate below the cuticle of non-host species. Once inside the plant, however, they could act as they would on a susceptible host.

Other non-host specific proteins may also have contributed to the phytotoxic effect of the culture filtrate. One of the proteins identified in the culture filtrate was an ortholog of proteins containing a cerato-platanin domain, which are known to produce a response in a broad range of hosts (discussed in section 0). Dothideomycetes and other fungi also produce non-selective proteinaceous effectors such as Ecp6 (de Jonge et al. 2010) or those that are partially selective such as NEP1 (NLP1) (Santhanam et al. 2013). Genes for both of these small secreted proteins are present in the PmedOMT5 genome. Transcripts were also detected as expressed in planta and in mycelia grown in F2NY for 35d (Pmed_117370 homologous to Ecp6, FPKM: F2NY- 40, IP- 1358, PmedOMT5_11472 homologous to NLP1, FPKM: F2NY- 3, IP-432), but their corresponding proteins were not detected in the extracellular fraction by proteomic analysis. Expression of Ecp6 homologs is known to be increased in planta in other species and it may not be produced in detectable quantities in the PmedOMT5 culture filtrate although other proteins were identified in the extracellular fraction, which had lower FPKMs in the F2NY sample. NEP1 however was originally isolated from the 7 d old culture filtrate of F. oxysporum (Bailey 1995) and showed decreased production after 22 d. Why it was not detected in the PmedOMT5 culture filtrate is unknown. Possibly it is only produced earlier in culture and there was insufficient for detection at 35 d. It should be noted that the role of Ecp6 is that of an effector in the more traditional sense, in that, it does not interact directly with the plant cell to cause necrosis, as necrotrophic effectors do. Its role is defensive rather than offensive, ‘shielding’ the fungal cell from recognition by the plant host defences (PTI) by binding to chitin outside the fungal cell membrane and thereby indirectly promoting pathogenicity (de Jonge et al. 2010).

7.4.4 Fractionation of culture filtrate and secondary metabolites

Fractionation of culture filtrates into fractions above and below 3 or 10 kDa consistently showed that the higher molecular weight fractions caused a stronger reaction in the plant. This may simply be due to a higher protein concentration (10-fold) than the lower molecular weight fraction. As shown in Figure 7.6, the culture filtrate of PmedOMT5 contains a large number of proteins greater than 50 kDa. In order to fully investigate the
activity and potential host specificity of smaller compounds these will need to be further separated in future studies.

Several fungi previously identified as belonging to the genus *Phoma* have been identified as producing phytotoxic secondary metabolites (Rai et al. 2009). Two secondary metabolites have previously been isolated from fungi identified as *P. medicaginis*, which cause inhibition of plant germination or reduced growth, phomic acid and brefeldin A (Lakshmanan and Vanterpool 1967, Weber et al. 2004). Phomic acid, a water soluble acidic compound with a high melting point of 260 °C, was isolated from a fungus described as *P. medicaginis* infecting flax seed (*Linum usitatissimum*) (Lakshmanan and Vanterpool 1967). Brefeldin A is a macrolide polyketide product of an unknown PKS and is produced by several taxonomically unrelated fungi. It was observed to be produced by *P. medicaginis* isolates growing on *Medicago lupulina* (Weber et al. 2004). Brefeldin A has been reported to cause necrosis on punctured leaves of *Xanthium occidentale* (Vurro et al. 1998), but no necrosis was observed when brefeldin A was infiltrated into *M. truncatula* leaves (Figure 7.3c). Dothideomycete fungi are known to produce many non-selective secondary metabolite effectors that are involved in virulence, mainly via effects on common plant physiological processes, such as, through production of ROS (cercosporins: *Cercospora* spp.) or inhibition of actin (cytochalasins) (Stergiopoulos et al. 2013). However production of water soluble phytotoxic secondary metabolites (<3kDa) by PmedOMT5 was not observed under the culture conditions tested.

Other closely related fungi within the family *Didymellaceae* are also known to produce small non-proteinaceous secondary metabolites, including the recently isolated, light activated phytotoxic anthraquinone lentsione from *Ascochyta lentis* (Andolfi et al. 2013), which is also toxic to lentil, chickpea, pea and faba bean. Also the recently renamed and re-classified *Peyronellaea zeae-maydis* (Aveskamp et al. 2010) produces a polyketide called PM-toxin, an analog of T-toxin from *C. heterostrophus*, that confers pathogenicity on *T-cms* maize (Yun et al. 1998) and the phytotoxic solanopyrones A, B, C and cytochalasin D are produced by very close relative *Didymella rabiei* (Alam et al. 1989, Chen and Strange 1991, Hohl et al. 1991, Latif et al. 1993). Additionally, species that were previously classified as *Phoma* have also been identified as producing secondary metabolites with a range of activity classes including phytotoxins, antimicrobials, herbicides and gluconeogeneisis inhibition (Ohtsu et al. 2005, Rai et al. 2009).
2009). However taxonomic re-classification has meant that many of these species have been moved to different genera and families including: *Phoma macdonaldii* to *Leptosphaeria lindquistii* and *Phoma lingam* to *Leptosphaeria maculans*. While most former *Phoma* spp. remain within the Pleosporales, most do not group within the Didymellaceae family to which *P. medicaginis* belongs (de Gruyter et al. 2013). Similar molecules may be produced by *P. medicaginis*, but do not appear to be optimally produced in the media used for this study. The volume of liquid cultures used may dilute any metabolites (cultures were grown in 100 mL media) requiring further concentration to have reached *in planta* active concentrations or may not have contained the correct micronutrients or plant stimuli to induce production. It has previously been stated that the closely related *D. rabiei* only produces the secondary metabolite solanopyrones in the presence of chickpea plant extracts in media (Alam et al. 1989). Although it was later shown that these plant extracts could be replaced by salts of particular divalent metal cations (Chen and Strange 1991), which may require optimisation for each fungal species.

### 7.4.5 Thermal stability and proteinase resistance of culture filtrate

The activity of the components of the culture filtrate larger than 10 kDa were not significantly diminished by heat treatment for 2 h at 80 °C or pronase digestion for 2 h at 37 °C. These conditions should decrease the activity of most proteinaceous effectors although some other compounds have been identified that show heat resistance. For example, PtToxB is stable for 1 h at 55 °C (Strelkov et al. 1999). An extracellular chitin deacetylase produced by *Aspergillus nidulans* (Alfonso et al. 1995) was also heat stable after 1 h incubation at 100 °C. Finally, SnTox1, an effector identified in *P. nodorum* containing 16 cysteines that are predicted to form multiple disulphide bridges, is only partially inactivated by boiling for 1 h at 100 °C. This lends weight to the hypothesis that secreted effector proteins of PmedOMT5, may also be heat and proteinase resistant, for those that are cysteine rich.

Glycoproteins resistant to proteinase digestion have also been identified in another fungal species of the order Pleosporales. *Phoma tracheiphila*, currently re-named *Plenodomus tracheiphilus* (de Gruyter et al. 2013), causes Mal secco disease of citrus trees in the Mediterranean. Compounds identified from partially purified culture filtrate include non-selective glycoprotein toxins of 60 and 93 kDa, which induced light-
dependent toxin symptoms on citrus leaves within 4-7 d. The 60 kDa protein is proposed to be completely encapsulated by the carbohydrate moiety rendering it resistant to protease degradation (Fogliano et al. 1998). A 7.6 kDa proteinaceous and possibly glycosidic compound has also been isolated from the closely related D. rabiei that was phytotoxic to chickpea cells in culture (Chen and Strange 1994). However this compound has not yet been fully characterised or its host-specificity tested. The fact that glycoproteins have been identified from closely related fungi and their high stability suggests that P. medicaginis may also produce glycoproteins, which may also account for the stability of its phytotoxic compounds.

7.4.6 Autofluorescence in culture filtrate infiltrated leaves

Autofluorescence in plant cells is often due to the accumulation of flavonoid phytoalexins and lignin (Nicholson and Hammerschmidt 1992). The presence of autofluorescence can be used as an indication of cell damage due to the hypersensitive response (Heath 1998). Recombinantly produced polyhistidine-tagged L. maculans SP-1 protein (an ortholog of the predicted cerato-platanin protein in PmedOMT5, PmedOMT5_01552) infiltrated into canola leaves at a concentration of 30 ng or infiltrated crude culture filtrate both induced autofluorescence. Neither a His-tagged control protein nor chicken lysozyme protein of similar size to SP-1 did (Wilson et al. 2002), indicating that autofluorescence is generated only in response to specific proteins that invoke a defence response in the plant. Autofluorescence has also been observed in chickpea leaves incubated with a mixture of solanopyrones A, B and C (Hohl et al. 1991) and following the infiltration a cerato-platanin ortholog from Ceratocystis fimbriata f. sp. platani to tobacco leaves or its application to plane tree leaves (Pazzagli et al. 1999). Cells penetrated by PmedOMT5 hyphae were observed to autofluoresce in this study and previously by Kamphuis and colleagues (2008) (Figure 7.4). The observation that plant cells within the PmedOMT5 culture filtrate infiltration zone produce autofluorescent compounds suggests that infiltration of PmedOMT5 culture filtrate is causing a response within the leaf similar to that observed after foliar inoculation with PmedOMT5 spores.
7.4.7 Cerato-platanin

The cerato-platanins are a family of fungal-secreted, PAMP proteins, originally identified in the plane tree pathogen *Ceratocystis platani* (Pazzagli *et al.* 1999). They can induce plant defence responses including: plasmolysis, necrosis, accumulation of phenolic compounds and phytoalexins, production of ROS and up-regulation of defence response genes in host and non-host plants, summarised in de Oliveira *et al.* (2011). Proteins with a cerato-platanin domain are small (~150 aa), hydrophobin-like and contain four cysteine residues that form two disulphide bonds (Pazzagli *et al.* 1999). They are found in the cell walls of hyphae and conidia and can be secreted in culture (Boddi *et al.* 2004). They self-aggregate (Pazzagli *et al.* 2009), have the ability to bind oligosaccharides and are hypothesised to play roles in fungal growth, cell wall morphogenesis and development (de Oliveira *et al.* 2011). Orthologs from *B. cinerea* (Frias *et al.* 2012) and *Magnaporthe oryzae* (Yang *et al.* 2009) have been found to trigger PTI and SAR in many host and non-host species conferring resistance to other pathogenic species. Targeted disruption of the cerato-platanin gene (*sp1*) in *L. maculans* showed no reduction in virulence (Wilson *et al.* 2002).

Two 6-frame mapped peptides support the presence of an ortholog of this protein in the extracellular fraction of PmedOMT5 and it has a predicted effector score of 24. The gene *PmedOMT5_01552*, encoding a protein product that contains a cerato-platanin domain [Pfam: Pf07249.7], was found to be most highly expressed at 35d in F2NY culture medium (FPKM=528) and was also expressed *in planta* (FPKM=24.1).

There are cerato-platanin orthologs in all of the 36 Ascomycetes examined in this study including non-pathogens. The role in virulence of cerato-platanin is unknown although it has been hypothesised to bind oligosaccharides in a similar manner to Ecp6 although with a lower affinity (de Oliveira *et al.* 2011).

7.4.8 Functional annotations over-represented in the extracellular fraction

Common themes observed amongst domains over-represented in the extracellular fraction appear to be associated with degradation of pectin and reorganization of chitin in fungal cell walls for growth. Pectin is more prevalent in dicots than monocots and these domains are commonly found in larger numbers in dicot pathogens relative to pathogens of monocot plants. Two CAZy domains, CBM43, many of which are
reported to bind β-1,3-glucans in the CAZydatabase (Cantarel et al. 2009), and GH16 (which act on a range of glucans) were also found to be significantly over-represented in the extracellular fraction. Both of these CAZy domains were identified in all of the 103 fungi examined by Zhau and colleagues (2013) which included pathogenic and non-pathogenic species of Ascomycota, Basidiomycota, Chytridomycota and Zygomycota (Zhao et al. 2013). Zhao et al., (2013), state that the known substrate for the most common form of GH16 is fungal cell walls, thus these genes potentially play a role in fungal cell wall re-organization or morphogenesis within the plant. Many proteins greater than 50 kDa were observed in the culture filtrate (Figure 7.6 and Figure 5.1) and include several CAZymes, 26 of which were identified in the culture filtrate with predicted molecular weights ranging from 28-100 kDa, with 22 out of 26 greater than 40 kDa. Attempts were made to remove these from the fractionated culture filtrate using molecular weight cut-off filters, but the filters easily became blocked and good separation was not achieved. The presence of these may be partially responsible for the activity of the culture filtrate.

7.4.9 Determining the nature of potential effectors in PmedOMT5

The assays described in this thesis isolated the active component of the PmedOMT5 culture filtrate to the greater than 10 kDa fraction. This means the active compound(s) are highly likely to be proteinaceous, as secondary metabolites are not reported to occupy this weight range. The active compound or compounds are stable, able to resist heat and pronase degradation, and still produce a similar response in planta after storage for 6 weeks at 4 °C on ice.

Further purification of the proteins present in PmedOMT5 culture filtrate is required to assess the activity of individual components. One possible strategy for better separation may include: ammonium sulphate precipitation followed by sepharose size selection prior to stepwise ion exchange chromatography followed by hydrophobic interaction chromatography (HIC).

Many effectors produced by Dothideomycete fungi are known to be active only in the presence of light including: Sn/PtrToxA, the unkown proteinaceous toxin from P. teres (Sarpeleh et al. 2008), the A. lentis secondary metabolite toxin lentsone (Andolfi et al. 2013) and cercosporin produced by Cercospora spp. (Daub and Hangarter 1983). Sarpeleh and colleagues (2008) also observed a reduction in the activity of the
infiltrated semi-purified toxins at temperatures of 4 °C. Thus future experiments may benefit from examining the effects of light and heat variation on the activity of the culture filtrate.

A list of the top five effector candidates identified in the culture filtrate is presented in Table 7.6. The next step in analysis of *P. medicaginisc* effectors would be to sequence these regions (and those identified in the previous chapter as showing increased expression IP) from other *P. medicaginis* isolates to confirm presence of the same genes, conserved expression patterns and if their presence correlates with activity in *Medicago* and if so examine whether or not these genes may be undergoing positive selection. It is possible that different effector complements may account for the differences in aggressiveness on different hosts of the *P. medicaginis* variants *medicaginis* and *macrospora*.

The fact that the PmedOMT5 culture filtrate did not produce a clear host-selective effect may be due to it producing multiple effectors with a range of activities on different hosts. In *P. nodorum*, a range of effectors have been identified with varying effects on disease development, each creating a response in a host wheat plant with a corresponding sensitivity gene: ToxA-Tsn1 95%, SnTox1-Snn1 58%, SnTox2-Snn2 47%, SnTox3-Snn3 17% and SnTox4-Snn4 41% (Friesen et al. 2007, Friesen et al. 2008a, Friesen et al. 2008b, Stergiopoulos et al. 2013). The resistance identified by Kamphuis and colleagues (2008) in *M. truncatula* to *P. medicaginis* var. *medicaginis* was recessive, with different dominant susceptibilities observed in each line crossed with a resistant parent. A highly significant QTL was identified at a different locus in each cross (SA27063 x A17 and SA27063 x SA3054), each accounting for approximately 30% of the resistance suggesting the presence of at least two susceptibility loci. It is possible that the susceptibility of these lines is governed by sensitivity to different components of the PmedOMT5 culture filtrate, but disease may also be affected by the presence of non-selective effectors that contribute to the symptoms observed *in planta*.

Ellwood and colleagues (2006) found that the *M. truncatula* lines tested differed in their responses to different *P. medicaginis* isolates, indicating that the different isolates may also produce different effectors that interact with different sensitivity genes in *M. truncatula*. It is also possible that PmedOMT5 does not produce host specific effectors.
but instead relies on a collection of effectors that have activity in a range of hosts such as Ecp6, cerato-platain and NEP1, all of which are expressed in planta.

7.5 Summary
Experimental evidence indicates that the three isolates of *P. medicaginis* tested all secrete proteinaceous molecules that cause a response *in planta*. The active component is greater than 10 kDa in size. However, to date no host-specific effector has been identified in *P. medicaginis var. medicaginis*. None of the isolates were found to secrete secondary metabolites that cause a visible reaction in the plant under the conditions tested.

A significant output of this study has been the identification of strong candidates for further biological testing using a data driven approach combining whole genome sequencing with transcriptomic and proteogenomic data. Some insights were gained into the secretome of *P. medicaginis* but further investigation is needed to characterise the effector complement of PmedOMT5. Current data indicates PmedOMT5 also produces non-selective proteinaceous effectors.
Chapter 8
General Discussion
8.1 The pathogen *P. medicaginis*

*P. medicaginis* is a necrotrophic fungal pathogen of legumes that causes spring black stem and leaf spot on its main hosts *Medicago* species. *P. medicaginis* has been described as a weak pathogen (Boerema *et al.* 1965) and the success of infection appears heavily dependent upon environmental factors including temperature and humidity (Barbetti 1987). Although under conducive conditions the fungus can cause disease on all parts of *Medicago* plants above and below ground.

Prior to this study there were very few genomic or transcriptomic sequences available for *P. medicaginis*. The majority of resources available were sequences from conserved genes or genomic regions used for phylogenetic studies and a small number of ESTs from infected *M. truncatula* cotyledons (Gamas *et al.* 2007). Although considerable research had been done on the response of *Medicago* species inoculated with *P. medicaginis* (Paiva *et al.* 1994, Jasinski *et al.* 2009, Kamphuis *et al.* 2011), elucidating the involvement of the phenylpropanoid and octadecanoid pathways in mediating and signalling the plant defence response, little was known about the pathogen’s virulence arsenal at the genomic level. The major outcomes of this study include the first draft assembly of a fungal genome from the genus *Phoma*, its biologically supported gene prediction and inferred annotations of gene functions and biological roles.

8.2 Genome and assembly

A genome assembly is a representation of the actual deoxyribonucleic acid content of an organism. In most cases, it is almost impossible to assemble genomic sequences completely into chromosomes and an assembly alone usually contains no information about the three dimensional organisation of the chromosomes around histones that may influence the co-ordination of transcription of distantly spaced genomic loci (Tanizawa *et al.* 2010). In a recent comparison of assembly algorithms for short-read sequences, it was demonstrated by Salzberg and colleagues that a short read assembly will never provide a complete picture of a complex genome (Salzberg *et al.* 2012). Their analysis showed that the best currently available assembler in terms of contiguity and correctness was Allpaths-LG. However, this could not be used with PmedOMT5 as it requires at least one overlapping mate pair library which was not available for this organism at the time of assembly. It is possible that small minor errors have been introduced in to the genome by the chosen assembler, Velvet. Only further sequencing of PmedOMT5 or a
similar isolate, preferably with longer read lengths and a larger insert size will be able to identify these. Nevertheless it is estimated that the assembly covers 98% of the protein-coding gene space based on CEGMA analysis (Parra et al. 2007).

**8.2.1 Genome size**

The PmedOMT5 genome assembly consists of 952 nuclear scaffolds with a total length of ~ 31 Mb plus one scaffold of 38.5 kb that represents the mitochondrial genome. These scaffolds contain an estimated 1.35% of repetitive DNA and 36 kb of identified non-coding RNA.

Fungal genomes show remarkable diversity in genome size and architecture (Raffaele and Kamoun 2012). PmedOMT5 has a relatively small genome, even allowing for the fact that the assembly was constructed from short-read data only, which appears to be due to a lower number of repeats than seen in related species, as the gene count is similar to that seen in other Dothideomycete pathogens. As outlined in chapter section 3.4.4, it is estimated that ~5 Mb of the PmedOMT5 genome is not represented in the assembly based on the predicted genome size of ~36 Mb derived from the median read coverage and the actual genome size is likely to lie within this range (31-36 Mb).

Genome sizes for the closely related Didymellaceae fungi *A. rabiei* and *A. fabae* range from 23.9-33.5 Mb and 33.3-42.9 Mb respectively based on electrophoretic karyotyping (Akamatsu et al. 2012). As only a small proportion of the RNA-seq derived reads were unable to be aligned to the assembly (12% *in vitro* derived), it is likely that the unassembled sequence does not contain a large number of genes. Furthermore, only a small proportion of the unassembled reads could be aligned to known repetitive sequences from other fungi (0.3%) thus they are unlikely to originate from conserved TEs, rather they may be derived from telomeres, centromeres and rDNA which were also not well represented in the current assembly. These regions of a genome are notoriously difficult to sequence and assemble even using Sanger technology (Cambareri et al. 1998). Alternatively these sequences may derive from novel TEs found in members of the Didymellaceae that are uncharacterised in other species.

The PmedOMT5 genome assembly is several megabases smaller than those of other Pleosporales pathogens. *P. nodorum* and *P. tritici-repentis* are both ~ 37 Mb and *L. maculans* is ~45 Mb (Hane et al. 2007, Rouxel et al. 2011, Manning et al. 2013). The 38.5 kb PmedOMT5 mitochondrial genome is also small compared to closely related
species and shows comparatively much less endonuclease invasion than is seen in other Dothideomycete species. The mitochondrial genome of *P. nodorum* is 50kb and contains several endonucleases and the *L. maculans* mtDNA is 155 kb and shows extensive endonuclease invasion. In *P. nodorum* pathogenic strains were observed to have larger mitochondrial genomes by ~7kb (Syme *et al.* 2013), including additional intronic endonucleases and ORFs. Thus, not only does the nuclear genome of *PmedOMT5* appear repeat sparse, so does the mitochondrial genome. The correlation between larger genome size due to increased repeat content and increased genome plasticity has been noted and proposed to facilitate an increased propensity for recombination in repeat rich areas (Raffaele and Kamoun 2012).

Unlike bacterial genomes, which have been demonstrated to ‘slim-down’ over the course of their evolution due to repeated loss of non-essential genes, often leaving them with very specific host-ranges or ecological niches, fungal genomes show a greater variation in size and tend to retain genes, e.g. most saprotrophs have CAZYs for pectin degradation and fungal pathogens of animals retain the ability to degrade cellulose (Zhao *et al.* 2013). The mechanisms by which fungal genomes can decrease in size include unequal crossing-over that induces localized deletions of one or more genes, or other types of chromosomal rearrangement that could potentially induce large-scale deletions. Whilst the mechanisms responsible for increased genome size can include duplication of single or groups of genes, duplication of all or selected chromosomes, hybridization between species and horizontal transfer of genes or chromosomes (Ma *et al.* and references therein 2013). In the genus *Fusarium*, genome size ranges from 36 to 61Mb in the five sequenced species and formae speciales (Ma *et al.* 2013). There does not seem to be a clear evolutionary trend in the genome sizes of fungi suggesting either a continuous loss or gain of material. Instead analysis shows that genetic material is continually lost and gained via HGT and transposon-mediated transfer over the evolutionary history of fungi. Why one species is more susceptible to transposon invasion than another is unknown, but possibly has to do with the number of other species it encounters in close proximity *i.e.* that share the same host or environment. *P. medicaginis* is known to produce brefeldin A, a fungitoxic compound that may reduce the number of interactions it encounters with other fungi and thus the potential for transfer of transposons and their potential “hitchhiker” HGT genes. No evidence was observed that suggested the presence of conditionally dispensable chromosomes in
PmedOMT5 nor have any been observed in analysis of related A. rabiei isolates (Akamatsu et al. 2012).

8.3 Gene content of PmedOMT5

Prediction of the location and structure of genes that determine the molecules encoded by a genome is not a simple task. Annotation of the Sanger sequenced mouse genome (Waterston et al., 2002) was supported by millions of ESTs and thousands of full length cDNAs and the result of considerable time, effort and finances poured into it by a large number of groups. However, analysis of the mouse RNAseq data by two new algorithms (Cufflinks and Scripture) (Guttman et al. 2010, Trapnell et al. 2010), validated by laboratory studies, identified hundreds-thousands of novel protein coding genes, novel isoforms of known genes and non-coding RNAs. The authors of these studies state that the explanation for these newly discovered genes is that some genes/isoforms are only expressed under certain conditions. In order to validate as many of the predicted PmedOMT5 genes as possible, RNA-sequencing was conducted for a number of samples that covered important stages in the lifecycle of P. medicaginis including vegetative growth, sporulation, phytotoxic metabolite production and during the initial stages of plant infection.

The assembled nuclear scaffolds were predicted to encode 11,879 proteins and the majority of these (96%) were supported by transcriptomic reads and/or proteogenomic mapped peptides and the remaining 4% had homology in other fungal species. Overall, the gene content of PmedOMT5 appears similar to other Pleosporales plant pathogens with only 1,812 Phoma-specific genes identified, 298 of which were detected as expressed during the early stage of infection examined. Expansions were seen in several CAZy families related to pectin degradation, potentially due to the higher pectin content of legumes in comparison to the cereal hosts of the majority of other fungal species examined.

The interaction between P. medicaginis and M. truncatula forms a good model pathosystem for examining the interaction between necrotrophic fungi and legumes, as both host and pathogen genomes are sequenced and genetically tractable. There are few other pathosystems where both host and pathogen are sequenced and well-characterised apart from rice and the causal agent of rice blast, the sordariomycete, Magnaporthe oryzae. However, the pathogenicity mechanisms employed by the hemi-biotrophic M.
oryzae appear significantly different to those of necrotrophic pathogens of the order Pleosporales. Recent studies which confirm that *P. medicaginis* isolates isolated from *M. truncatula* are pathogenic on a range of legumes (Djebali 2013) suggest that findings from this study can be applied to a range of interactions between legumes and necrotrophic fungal pathogens.

*M. truncatula* has a much smaller and less complex genome than most other legumes, including the related tetraploid alfalfa (*M. sativa*), which makes this model plant amenable to molecular and genetic analysis. There are also a large number of TILLING and TNT gene disruption lines available and protocols for RNAi and *Agrobacterium rhizogenes* mediated hairy root transformations as well as stable *Agrobacterium tumefaciens* mediated transformation (Boisson-Dernier et al. 2001, Chabaud et al. 2003, Tadege et al. 2005, Le Signor et al. 2009) in *M. truncatula*. PmedOMT5 is also readily transformable via *Agrobacterium tumefaciens* mediated transformation (Williams, Lichtenzveig, Oliver unpublished data) although homologous recombination has not yet been tested. Investigation of the ability to transform *P. medicaginis* was undertaken with the two-fold aim of establishing a transformation technique that could potentially be later used to knock-out genes of interest and to create a GFP-expressing strain for further investigation of the *in planta* behaviour of PmedOMT5 (appendix 2). These transformants could potentially be used in future studies to enable easier observation of infection correlated with production of metabolites of interest in the plant.

### 8.4 Mating and sexuality

No conclusive evidence of a sexual stage has been reported for *P. medicaginis*, like the majority of species currently classified in the genus *Phoma*. All of the *P. medicaginis* isolates examined in this study contain the *mat 1-2-1* gene. Thus, while it has potential for sexual recombination there is no evidence that it is occurring in the field. However, heterokaryosis (two or more genetically different nuclei in a cell) and parasexuality (genetic recombination without meiosis) have been observed in *P. medicaginis* (Sanderson and Srb 1965) between auxotrophic mutants. Anastamosis has also been observed between mutants from single spores of different isolates (Mead 1964b), as well as, between conidia and/or hyphae of the same isolate 72 hpi on the surface of alfalfa leaves (Castell-Miller et al. 2007) and between different isolates on agar plates (Ellingboe 1959). Nuclei were observed in the bridging hyphae (Sanderson and Srb
1965) suggesting the transfer of genetic material and cells with one or more nuclei were observed in cells following anastomosis.

This mechanism has the potential to account for the variation observed in Phoma populations (Castell-Miller et al. 2008), but further study is required to determine how frequently this occurs in the field. While this demonstrates a mechanism by which HGT may be occurring, no obvious candidates of HGT either of individual genes or chromosomal segments, were identified in the assembly, based on comparison of gene content and orthology to both closely related and more distantly related sequenced species.

8.5 Repetitive DNA and RIP in PmedOMT5

RIP is a mechanism in fungi that affects duplicated sequences by inducing transition mutations. RIP is proposed to defend fungal genomes against transposon invasion by creating stop codons in the genes encoding transposases, which would limit transposons’ ability to move around the genome. A byproduct of this mechanism is reduced occurrence of gene duplication in species with high RIP occurrence. RIP has been demonstrated in N. crassa to occur in haploid nuclei during the dikaryotic stage following fertilisation but prior to meiotic DNA replication, which in filamentous Ascomycetes occurs prior to karyogamy (nuclear fusion) (Selker 1990). The only gene identified to date as essential for RIP in N. crassa is the rid gene (RIP defective), which has structural similarities to cytosine demethylase genes (Freitag et al. 2002). Following the occurrence of RIP, affected sequences are found to be methylated in vegetative mycelium. In fungi where RIP can occur it is not always observed to have occurred in all copies of a transposon (Clutterbuck 2010, and references therein).

Repetitive elements in a genome can cause ectopic recombination which may result in large-scale sequence deletion or duplications. Although the total predicted repeat content of PmedOMT5 was low, only 1.35% in contrast with the higher numbers seen in other published Dothideomycete assemblies, 4.5% in P. nodorum SN15 (Hane et al. 2007), 21% in M. graminicola (Goodwin et al. 2011) and 34% in L. maculans (Rouxel et al. 2011), several copies of transposons were identified as actively transcribed based on the alignment of mapped RNA-seq reads. No intact mobile elements were detected in N. crassa, the species where RIP was first identified (Galagan and Selker 2004). Although the genome has a repeat content of ~10%, all duplication is proposed to have
occurred prior to the evolution of RIP in this species. The low number of repeats in the PmedOMT5 genome is possibly due to RIP inactivation of invading transposons, which would prevent them from propagating throughout the genome, although the mechanism by which this may occur is unknown. The PmedOMT5 genome does contain a potential homolog of the *N. crassa* *rid* gene which shows very low expression under the conditions tested.

If RIP is occurring in PmedOMT5, an important question is when? It may have occurred in the past in a sexual ancestor species, or only rarely when *Phoma* encounters an isolate of the opposite mating type. Possibly, it could be occurring following anastomosis when two haploid nuclei from different isolates, or potentially the same isolate exist in the same cell. It has been proposed that the cost paid by genomes to protect themselves from transposon invasion is at the expense of gene variability and the evolutionary benefits that transposons can bring (Galagan and Selker 2004). However RIP itself has the potential to introduce variability into genes by ‘RIP leakage’, whereby single copy genes adjacent to replicated sequences are also mutated introducing new variation (Van de Wouw *et al.* 2010). Finding the answer to this question will have a significant impact on our understanding of pathogen adaptability in *P. medicaginis* and other asexual species.

### 8.6 The role of overcoming plant secondary metabolites in pathogenicity of *Phoma* on legumes and legume defence

As previously proposed by several authors (Higgins 1972, Castell-Miller *et al.* 2007), *P. medicaginis* was found to encode the requisite genes for degrading or excreting phytoalexins produced by legume species. These genes were expressed *in planta* during the early stages of infection when the pathogen is establishing itself in the plant. Several studies have shown that increased production of isoflavonoids or other phenolic compounds in the plant, whether genetically engineered or naturally occurring, provides increased resistance to *P. medicaginis* in *Medicago* spp. (He and Dixon 2000, Hipskind and Paiva 2000, Deavours and Dixon 2005, Kamphuis *et al.* 2011). One potential mechanism for the differences in virulence between *P. medicaginis* isolates is variation in the number of these genes in the genome and the rate and timing at which they are expressed. It appears that the ability to degrade or expel from the fungal cell the phytoalexins produced by legume plants is one of the key factors shared by legume
pathogens found within the families Didymellaceae and Nectriaceae. Potentially, acquisition of this trait is what allows these pathogens to establish on legume hosts when other fungal species cannot. It is clear from the results of many studies (Castell-Miller et al. 2007, Kamphuis et al. 2008, Djebali 2013) that isoflavonoid production has little to no effect on the initial invasion of the host cell, as in all investigations spores both germinated on and penetrated, both susceptible and more resistant plants, including those with measured higher isoflavonoid levels. However, it may be that the production of phytoalexins at greater levels than the pathogen can cope with is a limiting factor for the growth of *P. medicaginis* within the host tissues, and this is one of the factors preventing further proliferation in more resistant *Medicago* spp. and cultivars.

### 8.7 Phoma infection mechanisms

Although classed as a necrotroph *P. medicaginis* does not appear to use very aggressive strategies to kill its host’s tissues. While the initially invaded cells quickly become necrotic, the lesions do not spread very rapidly or widely unless conditions are very favourable, although it can cause large losses under high inoculum loads and favourable conditions in field inoculated plants (Barbetti 1987). It often does not sporulate on living host tissues presumably as it is unable to sufficiently overcome host defences but it can persist within the host tissue, and has even been described as growing endophytically (Weber et al. 2004). It is frequently found to infect the older tissues of plants such as the crown of alfalfa, where more susceptible older leaves with lower medicarpin content are found. One of the major problems of *P. medicaginis* infection in the field is loss of yield, due to the defensive mechanisms employed by *Medicago* species that ultimately lead to dropping of infected leaves from the base of the petiole. *P. medicaginis* appears to be able to persist in the plant asymptotically following the initial infection and therefore must have a mechanism for evading or suppressing host defence, at least locally. This may also be a mechanism that allows the plant and pathogen to coexist, until the fungus gets senescent tissue to feed on, but the whole host plant doesn’t die. In the field, a certain amount of disease is tolerated and only causes a problem when, under conducive conditions, too much defoliation occurs or oestrogenic compounds build up in the plant (Barbetti 1987, Barbetti 2007).

*P. medicaginis* has been reported to penetrate the plant cell via secretion of a battery of secreted enzymes, which is corroborated by the detection of 257 genes encoding
CAZymes expressed *in planta*, many of which may be involved in cell wall degradation during the early stages of infection.

### 8.8 Secreted proteins and effectors

In the last decade there has been a surge in plant pathogenomics projects, ultimately aimed at elucidating the effector complements of a diverse range of fungal pathogens. Plant pathogenic fungi produce several classes of effector molecules, which have an effect on the host plant’s cells leading to pathogenicity in compatible reactions or resistance in incompatible reactions. Effectors are thought to evolve in concert with changes in plant defence mechanisms (Godfrey *et al.* 2010). As a plant evolves a method of detecting a molecule, pathogens that have evolved novel undetectable effectors remain pathogenic on the plant and have a selective advantage over previously successful pathogens which produce detectable effectors that become less numerous over time.

The majority of secreted proteins amongst filamentous Ascomycetes are conserved among necrotrophs, hemi-biotrophs and saprotrophs, but there is a greater variation when compared to biotrophs (Brown *et al.* 2012). The two major classes of effectors are host-specific and non-selective. The molecules in either class can be either proteinaceous or secondary metabolites, such as polyketides. Host-specific effectors are active only against particular plant species and effectively determine the host range of the pathogen (Wolpert *et al.* 2002).

PmedOMT5 was predicted to secrete 1,020 proteins. The genes encoding 420 of these were detected as expressed IP based on a preliminary low-coverage RNA-seq study of early infection time-points. A total of 170 proteins were identified via proteogenomic analysis of the *in vitro* extracellular proteome and 115 (68 %) of these were predicted to be secreted by at least one method, the same percentage (68%) as was identified in a similar study in *F. graminearum* (Paper *et al.* 2007). Although, as fungi are known to have non-classical secretion pathways, which are utilised by some effectors (Ridout *et al.* 2006, Liu *et al.* 2014), and which cannot be well computationally predicted, the lack of predicted secretion does not mean a protein may not function as an effector.

In *P. medicaginis*, predicted effectors were identified with the support of preliminary RNA-seq data from the early stages (1-5 dpi) of infection on the host *M. truncatula* and sequenced peptide fragments isolated from phytotoxic culture filtrate. Examination of
these two data sets combined with examination of protein characteristics known to be associated with fungal host–specific effectors resulted in prediction of a short-list of 111 effector candidates. These candidates were expressed in the early stages of the P. medicaginis – M. truncatula interaction or were secreted into phytotoxic culture filtrate, although confirmation of their expression patterns, activity and host-specificity will require further experimental validation.

From the top 111 candidates, three subsets of candidates were identified: those that scored highest in attributes previously associated with characterised effectors in other species, those detected in the phytotoxic extracellular fraction, and those that had effector-associated characteristics and were unique to PmedOMT5 amongst sequenced species.

Effectors identified in other filamentous plant pathogens often contain conserved motifs or intragenic repeats that are specific to that species or genus. However, examination of PmedOMT5 effector candidates did not reveal any obvious conserved motifs. While RXLR motifs were identified in many PmedOMT5 proteins, few were seen in the N-terminal region. This particular motif is not known to be associated with translocation into host cells in fungi although several “RXLR-like” motifs have been identified in fungal intracellular avirulence proteins including AvrLm6 which contains an ‘RYWT’ motif (Kale et al. 2010, Kale 2012) proposed to be associated with membrane transport. These genes also contain secondary non-functional domains, which have been proposed to play a role in stability of the protein.

8.9 Summary and future work

The significant outcomes of this thesis are the provision of a solid foundation of genomic resources, upon which to build future studies investigating the pathogenicity of necrotrophic fungal pathogens of legumes. This thesis has also revealed many tantalising insights into the gene content of PmedOMT5 and their potential roles in plant pathogenicity. The genome assembly of P. medicaginis var. medicaginis isolate OMT5 and its supported gene models will facilitate further comparative genomic studies. However the predicted effector candidates will require further experimental validation. Initially studies to confirm their expression during the course of infection in planta either via qRT-PCR or for greater detail and insight possibly further RNA-seq data from a time-course experiment including three or more replicates and preferentially
stranded RNA-seq. This analysis should support the further investigation of the currently predicted effectors and if additionally coupled with a new assembly utilising long reads could potentially identify new candidates from currently unassembled or mis-assembled genomic regions. Once further confirmation is obtained, future work may involve the purification or heterologous expression of predicted effector candidates to characterise their phytotoxicity and host-specificity.

The various insights into the nature of pathogenicity in Phoma presented in this thesis can be of assistance to legume breeding programs. In particular, advances in effector validation can be used to aid in disease resistance breeding efforts as has been the case for ToxA and wheat. Although SBSLS caused by *P. medicaginis* can be treated with fungicides, effective treatment is not always feasible due to high costs and there are also concerns regarding adverse effects on humans, livestock and surrounding ecosystems from high levels of fungicide residues remaining on pasture and fodder crops. Significant benefits to farmers could come from the integration of higher levels of resistance in plants, through genetic screening against a panel of Phoma effectors with validated phytotoxic activities.

In conclusion, the sequencing and assembly of the genome of *P. medicaginis*, in combination with the availability of the genome sequence of *M. truncatula* and the array of available genomic, transcriptomic and proteomic data for both organisms, provides a model pathosystem for investigating the interaction of necrotrophic fungal pathogens and legumes.
Chapter 9
References


large sized conidial dimorphs, in vivo sometimes as *Stagonosporopsis*

Boerema, G. H. (1964). *Phoma herbarum* Westend., the type-species of the form genus

the black stem fungi on lucerne and red clover and the footrot fungus of pea.

Identification Manual: Differentiation of Specific and Infra-Specific Taxa in
Culture. Wallingford, Oxfordshire, UK, CABI Publishing.

Boerema, G. H., R. Pieters and M. C. Hamers (1993). Check-list for scientific names of
common parasitic fungi. Supplement Series 2c, d (additions and corrections):
Fungi on field crops: pulse (legumes), forage crops (herbage legumes),
vegetables and cruciferous crops. Netherlands Journal of Plant Pathology 99
(Supplement 1): 1-29.


Böhnert, H. U., I. Fudal, W. Dioh, D. Tharreau, J. L. Notteghem and M. H. Lebrun
(2004). A putative polyketide synthase/peptide synthetase from *Magnaporthe
grisea* signals pathogen attack to resistant rice. The Plant Cell 16 (9): 2499-513.

Boisson-Dernier, A., M. Chabaud, F. Garcia, G. Bécard, C. Rosenberg and D. G. Barker
(2001). *Agrobacterium rhizogenes*-Transformed Roots of *Medicago truncatula*
for the Study of Nitrogen-Fixing and Endomycorrhizal Symbiotic Associations.

Bok, J. W., S. A. Balajee, K. A. Marr, D. Andes, K. F. Nielsen, J. C. Frisvd and N. P.
Eukaryotic Cell 4 (9): 1574-82.

recognition receptors in plants and effectors in microbial pathogens. Science
324: 742 - 44.

associated with QoI resistance in *Cercospora beticola* field isolates from

Bolton, M. D., H. P. van Esse, J. H. Vossen, R. de Jonge, I. Stergiopoulos, I. J. E.
Stulemeijer, G. C. M. van den Berg, O. Borrás-Hidalgo, H. L. Dekker, C. G. de
The novel *Cladosporium fulvum* lysin motif effector *Ecp6* is a virulence factor
with orthologues in other fungal species. Molecular Microbiology 69: 119-36.


Kessie, F. (2013). Inisghts into the genome and secretome of Didymella pinodes, the causal agent of Ascochyta blight of pea. Ph. D., Murdoch University.


isoflavonoid O-methyltransferases in the evolution of plant defense responses.
The Plant Cell 18 (12): 3656-69.

Zhang, L. Gu, B. Zhang and D. Dou (2014). Unconventionally secreted effectors
of two filamentous pathogens target plant salicylate biosynthesis. Nature
Communications 5.

genes in vivo and in vitro, characterization of the flanking regions of the
PEP cluster and evidence that the PEP cluster region resulted from horizontal
gene transfer in the fungal pathogen Nectria haematococca. Current Genetics 44
(2): 95-103.

Liu, Z., J. D. Faris, R. P. Oliver, K. C. Tan, P. S. Solomon, M. C. McDonald, B. A.
McDonald, A. Nunez, S. Lu, J. B. Rasmussen and T. L. Friesen (2009). SnTox3
acts in effector triggered susceptibility to induce disease on wheat carrying the

Liu, Z., Z. Zhang, J. D. Faris, R. P. Oliver, R. Syme, M. C. McDonald, B. A.
The cysteine rich necrotrophic effector SnTox1 produced by Stagonospora
nodorum triggers susceptibility of wheat lines harboring Sun1. PLoS Pathogens
8 (1): e1002467.

Liu, Z. H., J. D. Faris, S. W. Meinhardt, S. Ali, J. B. Rasmussen and T. L. Friesen
Wheat to a Partially Purified Host-Selective Toxin Produced by Stagonospora
nodorum. Phytopathology 94 (10): 1056-60.

Loo, C., P. Dolling and S. Mokhtari (undated). Lucerne (Medicago sativa). Herbaceous
perennial legumes, Department of Agriculture and Food, Government of Western
Australia.

and T. J. Wolpert (2012). Tricking the Guard: Exploiting Plant Defense for

of transfer RNA genes in genomic sequence. Nucleic Acids Research 25 (5):
955-64.


Ma, L.-J., H. C. van der Does, K. A. Borkovich, J. J. Coleman, M.-J. Daboussi, A. Di
Pietro, M. Dufresne, M. Freitag, M. Grabherr, B. Henrisat, P. M. Houterman, S.
H. Bluhm, A. Breakspear, D. W. Brown, R. A. E. Butchko, S. Chapman, R.
Coulson, P. M. Coutinho, E. G. J. Danchin, A. Diener, L. R. Gale, D. M.


Wright, D. and N. Burges (2002). Fungal diseases present in Western Australian lucerne crops. Crop updates 2002 - farming systems, Department of Agriculture, Western Australia.


