UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Genome assembly, annotation and comparative analysis of the maize fungal pathogen *Cercospora zeina*

by

Nicholas Abraham Olivier

Submitted in partial fulfilment of the requirements for the degree
PhD(Bioinformatics)

in the Faculty of Natural and Agricultural Science
Department of Biochemistry, Genetics and Microbiology
University of Pretoria
Pretoria

Supervisor: Prof Oleg Reva
Co-supervisor: Prof Dave Berger

June 2019

Table of Contents

Selection of *Cercospora*-specific genes for phylogenetic species classification

Declaration

I, Nicholas Abraham Olivier, declare that the thesis, which I hereby submit for the degree PhD(Bioinformatics) at the University of Pretoria, is my own work and has not been submitted by me for a degree at this or any other tertiary institution.

Signature:

Date: 2019/06/24

Acknowledgements

I wish to thank the following persons and institutions:

- My co-supervisor **Prof Dave Berger** for the opportunity for this project, as well as endless encouragement and guidance.
- My supervisor **Prof Oleg Reva** for all of the valuable discussions we shared and great advice.
- **Dr Yao-Cheng Lin** for crucial guidance with genome assembly and annotation during research visits in Belgium.
- **Prof Fourie Joubert** for assistance with analyses and computing infrastructure.
- **Prof Yves Van de Peer** for hosting two research visits to the Bioinformatics and Evolutionary Genomics Group in Belgium.
- **Dr Ronnie de Jonge** for providing unpublished genome annotation data, and for interesting discussions on comparative genomics.
- Current and past members of the MPPI and CFPR research groups in the Department of Plant and Soil Sciences.
- Current and past members of the Centre for Bioinformatics and Computational Biology.
- The National Research Foundation, University of Pretoria postgraduate bursary scheme, as well as the University of Pretoria Genome Research Institute and Department of Plant and Soil Sciences for financial support.
- Colleagues and students in the Department of Plant and Soil Sciences.
- Most importantly my **wife and son** for support, motivation, advice and valuable discussions. This degree was not possible without you!

Preface

Grey leaf spot is one of the most significant yield-limiting diseases of maize, especially in sub-Saharan Africa. The fungal pathogen *Cercospora zeina* is responsible for the disease in Africa and other regions in the world, negatively affecting food security for populations relying on maize as staple food. Studies into the pathogen's mechanism of infection and pathogenesis have traditionally relied on morphological investigations and narrowly focused genetic and biochemical methods. Studying the genome and gene complement of *C. zeina* and identifying the genes and processes important for infection and pathogenicity could allow for the implementation of functional genomics approaches to describe the pathogen-host response and develop novel strategies for resistance breeding. The main aim of this study was the assembly and annotation of the *C. zeina* genome, and the comparison of genomic elements with other *Cercospora* species.

Chapter 1 is a Literature Review that provided a background on the two *Cercospora* species causing grey leaf spot in maize. The methods of host infection and pathogenesis were discussed, as well as major difference between the two species. Lifestyles of fungi were subsequently discussed, focusing on the effectors and immune response modulation affected by biotrophic, necrotrophic and hemibiotrophic phytopathogenic species. The classification of fungal species in the Dothideomycete class and *Cercospora* genus were also described. Shortcomings in the present multi-locus sequencing approach were highlighted, showing the need for improved loci for species classification in the *Cercospora* genus which contains a significant number of economically important plant-pathogenic species.

The first aim of the study was the sequencing and assembly of the *Cercospora zeina* genome using the Illumina next-generation technology. **Chapter 2** describes the isolation and quality control of the *C. zeina* genomic DNA. The quality control and assembly of the sequencing reads using several genome assembly algorithms are discussed, along with the completeness assessment of the assembled genome. The isolation of RNA from *in vitro C. zeina* cultures, sequencing and subsequent transcriptome assembly are shown, including the use of the RNA sequencing reads to assess the transcriptional functionality of the genome assembly. Finally the ITS and TEF1α sequences were extracted from the genome sequence and used in a phylogenetic analysis with related species to confirm the *C. zeina* origin of the sequenced DNA.

The second aim required the annotation of the genome assembly to identify and classify gene regions in the genome assembly. In **Chapter 3** the MAKER genome annotation pipeline was configured using *C. zeina*-specific sequence data and gene prediction models to identify promotor, protein-coding and repeat elements in the genome assembly. The resultant proteome of *C. zeina* was compared with that of the closely related *Cercospora zeae-maydis*, *Cercospora beticola* and *Cercospora berteroae*. Functional annotation of proteins of specific classes were performed to identify differences in secreted proteins, carbohydrate-active enzymes, lipases, proteases and components of secondary metabolite biosynthesis clusters. The synteny of the genes in the cercosporin toxin biosynthesis cluster was also confirmed in all four species.

The third aim involved the search for single copy orthologous proteins specific to *Cercospora* species for use as additional classification loci for the *Cercospora* genus. During the analyses in **Chapter 4** a number of Dothideomycete species' proteomes were analyzed for orthologous relationships. The single-copy orthologs specific to the four *Cercospora* species from Chapter 3 were analyzed for phylogenetic information content, and eight genes selected for the design of PCR primers in regions of protein identity. Primers were synthesized and tested for specificity during amplification of *C. zeina* and *C. zeae-maydis* genomic DNA for four of the genes. Degenerate primer pairs for two genes were selected for further analysis, since PCR amplification was specific for the gene products as confirmed with DNA sequencing. Further analysis falls outside the scope of this study.

The Conclusions and Future Prospects section in **Chapter 5** summarizes the results of the study, while shortcomings are discussed and future directions of interest are highlighted.

Outputs from the study

Publication:

Wingfield, B. D., Berger, D. K., Steenkamp, E. T., Lim, H. J., Duong, T. A., Bluhm, B. H., de Beer, Z. W., De Vos, L., Fourie, G., Naidoo, K., Olivier, N., Lin, Y. C., Van de Peer, Y., Joubert, F., Crampton, B. G., Swart, V., Soal, N., Tatham, C., van der Nest, M. A., van der Merwe, N. A., van Wyk, S., Wilken, P. M., Wingfield, M. J. (2017) IMA Genome-F 8: Draft genome of *Cercospora zeina*, *Fusarium pininemorale*, *Hawksworthiomyces lignivorus*, *Huntiella decipiens* and *Ophiostoma* ips. *IMA Fungus* **8(2)**:385-396

Additional co-authored publications:

Swart, V., Crampton, B. G., Ridenour, J. B., Bluhm, B. H., Olivier, N. A., Meyer, J. J. M., Berger, D. K. (2017) Complementation of CTB7 in the maize pathogen *Cercospora zeina* overcomes the lack of *in vitro* cercosporin production. *Molecular Plant Microbe Interactions* **30(9)**:710-724

Muller, M. F., Barnes, I., Kunene, N. T., Crampton, B. G., Bluhm, B. H., Phillips, S. M., Olivier, N. A., Berger, D. K. (2016) *Cercospora zeina* from maize in South Africa exhibits high genetic diversity and lack of regional population differentiation. *Phytopathology* **106(10)**:1194-1205

Additional Publications using the described genome data:

Berger, D. K., Carstens, M., Korsman, J. N., Middleton, F., Kloppers, F. J., Tongoona, P., Myburg, A. A. (2014) Mapping QTL conferring resistance in maize to grey leaf spot disease caused by *Cercospora zeina*. *BMC Genetics* **15**:60

Nsibo, D. L., Barnes, I., Kunene, N. T., and Berger, D. K. (2019) Influence of farming practices on the population genetics of the maize pathogen *Cercospora zeina* in South Africa. *Fungal Genetics and Biology* **125**:36-44

Conference presentations:

Berger, D. K., Swart, V., Crampton, B. G., Ridenour, J. B., Olivier, N. A., Meyer, J. J. M., Bluhm, B. H. (March 2017) Molecular basis for lack of cercosporin production in *Cercospora zeina*, grey leaf spot pathogen of maize, Dothideomycete workshop, 29th Fungal Genetics Conference, Genetics Society of America, Asilomar, California, USA.

Berger, D. K., Muller, M. F., Kunene, N. T., Crampton, B. G., Bluhm, B., Phillips, S., Olivier, N. A. Barnes, I. 2017. High genetic diversity of the grey leaf spot pathogen, *Cercospora zeina*, observed in commercial maize in South Africa, 50th Congress of the Southern African Society for Plant Pathology, Champagne Castle Sports Resort, Drakensberg, KZN, RSA.

Conference posters:

Olivier, N. A., Lin, Y-C., Van de Peer, Y., Reva, O., Bluhm, B. H., Berger, D. K. (2014) Genome assembly and annotation of the maize fungal pathogen *Cercospora zeina*, SASBi-SAGS Joint Congress, Kwalata Game Lodge, Pretoria.

Berger, D. K., Muller, M. F., Kunene, N. T., Crampton, B. G., Bluhm, B. H., Phillips, S., Olivier, N. A., Barnes, I. (2016) High genetic diversity of the grey leaf spot pathogen, *Cercospora zeina*, observed in commercial maize in South Africa, XVII International Congress on Molecular Plant-Microbe Interactions, Portland, Oregon, USA.

Abbreviations

A	Adenine
ABC	ATP-Binding Cassette
AED	Annotation Edit Distance
antiSMASH	Antibiotics and Secondary Metabolite Analysis Shell
BAM	Binary alignment map
BOL	Barcode of Life
bp	Base pair
BRE	B Recognition Element
BUSCO	Benchmarking Universal Single-Copy Orthologs
C	Cytosine
CAZymes	Carbohydrate-active enzymes
cber	*Cercospora berteroae*
cbet	*Cercospora beticola*
CBM	Carbohydrate-Binding Modules
CBOL	Consortium for the Barcode of Life
CBS	Centraalbureau voor Schimmelcultures
CDD	Conserved Domain Database
CDS	Coding Sequence
CE	Carbohydrate esterases
CEGMA	Core Eukaryotic Genes Mapping Approach
ceze	*Cercospora zeina*
cezm	*Cercospora zeae maydis*
CHPC	Centre for High Performance Computing
CHS	Chitin synthase
COG	Clusters of Orthologous Groups
Contig	Continuous sequence
CPU	Central Processing Unit
CSIR	Council for Scientific and Industrial Research
CTAB	Hexadecyltrimethylammonium bromide
CTB	Cercosporin Toxin Biosynthesis
DDBJ	DNA Data Bank of Japan
dbCAN	Database for automated carbohydrate-active enzyme annotation
dd-H2O	Double-distilled water
DNA	Deoxyribonucleic acid
dNTP	Deoxynucleotide
EDTA	Ethylenediaminetetraacetic acid

EMBL-EBI	European Molecular Biology Laboratory - European Bioinformatics Institute
ENA	European Nucleotide Archive
EST	Expressed Sequence Tag
ETI	Effector Triggered Immunity
ETS	Effector Triggered Susceptibility
FABI	Forestry and Agricultural Biotechnology Institute
FR	Forward-reverse
G	Guanine
GAPDH	Glyceraldehyde-3-phosphate dehydrogenase
Gbp	Gigabase pair
GC	Guanine-Cytosine
GFF	Generic Feature Format
GH	Glycoside hydrolases
GLS	Grey leaf spot
GO	Gene Ontology
GUI	Graphic User Interface
HAMAP	High-quality Automated and Manual Annotation of Proteins
HMM	Hidden Markov Model
HR	Hypersensitive Response
HST	Host-specific toxin
Indel	Insertion/deletion
InParanoid	In-paralog and ortholog identification
ITS	Internal Transcribed Spacer
JA	Jasmonic acid
JGI	Joint Genome Institute
Kbp	Kilobase pair
KEGG	Kyoto Encyclopedia of Genes and Genomes
KOG	Eukaryotic Orthologous Groups
LCA	Last Common Ancestor
LSU	Large Ribosomal Subunit
MAMP	Microbe-Associated Molecular Pattern
MCL	Maximum Composite Likelihood
MCM7	Mini-chromosome Maintenance Complex Component 7
MFS	Major Facilitator Superfamily

miRNA	MicroRNA
MP3KB	3Kb mate-pair sequence library
MP8KB	8Kb mate-pair sequence library
MUCL	Mycothèque de l'Université Catholique de Louvain
N	No-call
NCBI	National Center for Biotechnology Information
ncRNA	Non-coding RNA
ng	Nanogram
NP	Non-deterministic polynomial-time
nrDNA	Nuclear ribosomal DNA
NRPS	Non-ribosomal peptide synthetase
OMA	Orthologous Matrix
PAMP	Pathogen-Associated Molecular Pattern
PANTHER	Protein Analysis Through Evolutionary Relationships
PCD	Programmed cell death
PCR	Polymerase chain reaction
PDA	Potato dextrose agar
PDA-AP	Potato dextrose agar – ammonium phosphate
PDB	Potato dextrose broth
PE	Paired-end sequence library
PGK	Phosphoglycerate kinase
PKS	Polyketide synthase
PL	Polysaccharide lyases
pmol	Picomole
ProDom	Protein Domain Database
PRR	Pattern Recognition Receptor
PTI	Pathogen Triggered Immunity
PVP	Polyvinylpyrrolidone
QC	Quality control
RF	Reverse-forward
RFLP	Restriction fragment length polymorphism
RIP	Repeat-Induced Point
RNA	Ribonucleic acid
RNAseq	RNA sequencing
ROS	Reactive oxygen species
RPB1	DNA-directed RNA polymerase II subunit rpb1
RPB2	DNA-directed RNA polymerase II subunit rpb2
RQI	RNA quality indicator

rRNA	Ribosomal RNA
SA	Salicylic acid
SAM	Sequence alignment map
SFLD	Structure-Function Linkage Database
SIMAP	Similarity Matrix of Proteins
siRNA	Small interfering RNA
SMART	Simple Modular Architecture Research Tool
SMURF	Secondary Metabolite Unique Regions Finder
SNAP	Semi-HMM-based Nucleic Acid Parser
snRNA	Small nuclear RNA
snoRNA	Small nucleolar RNA
SSP	Small secreted proteins
SSU	Small ribosomal subunit
T	Thymine
TE	Transposable element
TEF1	Translation elongation factor 1-alpha
TOPI	DNA topoisomerase 1
TPS	Terpene synthases
TrEMBL	Translated European Molecular Biology Laboratory Nucleotide Sequence Database
Tris	Tris(hydroxymethyl)aminomethane
tRNA	Transfer RNA
TSS	Transcription start site
tub2	Beta-tubulin
UniParc	UniProt Archive
UniProt	Universal Protein Resource
UniProtKB	UniProt Knowledgebase
UniRef	UniProt Reference Clusters
USA	United States of America
UTR	Untranslated region
Vulgar	Verbose Useful Labelled Gapped Alignment Report
WGS	Whole genome sequence
YPD	Yeast peptone dextrose

Index of Figures

Index of Tables

Chapter 1

Cercospora zeina, Grey leaf spot and fungal genomics

N. A. Olivier, O. Reva, D. K. Berger

Department of Plant and Soil Sciences, Forestry and Agricultural Research Institute (FABI), Faculty of Natural and Sciences, University of Pretoria, Pretoria, 0002, South Africa

The manuscript was researched and written by N. A. Olivier. Prof O. Reva and Prof D. K. Berger assisted with critical revision of the manuscript.

1.1 Introduction

There are an estimated 50,000 edible plant species in the world. According to the United Nations Food and Agriculture Organization (FAO) only three of these plant species, i.e. maize, wheat and rice, provide up to 60% of the total energy intake of around 4 billion people, while up to 90% of the total energy intake is provided by only 15 crop species (FAO, 1995). Maize is the most utilized of all the staple crops, providing up to 20% of the global calorie intake, and is especially higher in sub-Saharan Africa where more than 50% of the population depends on maize. Additionally maize is the main global livestock feed crop, and with global dietary compositions shifting to include more animal products, the use of maize for livestock is expected to increase. As competition for maize as food crop, there is also an increasing use of maize for biofuel production, especially in the USA. The availability of maize for food and livestock is therefore decreased to a certain extent, while the price of maize is continually increasing due to the rising demand (Alexandratos & Bruinsma, 2012).

Due to a variety of factors, including climate change, diseases, poor land management and limited resources for subsistence farmers, the outlook for maize yield improvement is currently negative. Limited access to fertilizers, as well as spiraling fertilizer prices, are decreasing the yields possible with these chemical additives. Changing weather patterns and temperatures are also disrupting traditional planting practices and crop selections, while crop diseases are expected to infect areas not previously at risk. One of the main solutions to diseases remain the use of traditional breeding or genetic modification to generate disease tolerant or resistant cultivars (Alexandratos & Bruinsma, 2012 ; OECD/FAO, 2016).

The study of fungal pathogens affecting maize and maize production is crucial for food quality and security. To understand maize fungal pathogens, their infection strategies and host interactions it is increasingly necessary to study their genome sequences and gene complements to evaluate their infection potentials. Grey leaf spot (GLS) of maize is one major pathogen affecting maize production.

1.1.1 Grey leaf spot

Grey leaf spot is a devastating foliar disease in maize (*Zea mays*), and is one of the most significant yield-limiting diseases of maize worldwide (Crous *et al.*, 2006). Maize is a staple food crop, especially in sub-Saharan Africa, and globally supplies 15% of the world's protein and 20% of the world's calories to more than 200 million people. The largest suppliers of maize include the USA, China, Brazil and Mexico. South Africa ranks as the largest producer in Africa, followed by Nigeria (Nuss & Tanumihardjo, 2010). The causal fungus, initially identified as *Cercospora zeae-maydis* (Tehon & Daniels, 1925 ; Chupp, 1953), was originally designated as a minor pathogen, although the practices of no-tillage, corn-on-corn cultivation and susceptible cultivars greatly increased the severity of infection and disease impact in various regions of the world (Latterell &

Rossi, 1983 ; Ward *et al.*, 1999). Estimates for yield losses range from 11% to 70%, depending on the hybrid and climatic conditions (Ward *et al.*, 1999).

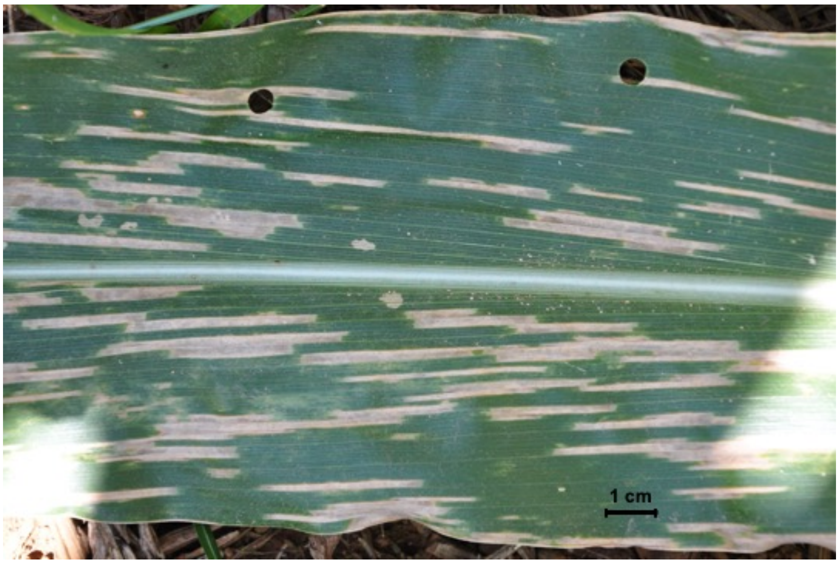


Figure 1.1 **Grey leaf spot symptoms on a maize leaf surface.** *The bar indicates a length of 1 cm.* (*Source: Berger, D. K., Cedara, South Africa)*

Early symptoms of GLS show up in lower leaves as small tan spots with chlorotic borders, and are indistinguishable from symptoms of other foliar pathogens. Mature lesions are readily distinguishable from other pathogens by their tan to grey colour and rectangular shape running parallel to leaf veins. Lesions coalesce during latter infection stages (Figure 1.1), and can blight the entire leave surface, thereby reducing the photosynthetic potential of infected leaves (Ward *et al.*, 1999). Infections can spread to the sheath or husk tissues, and stalk infections have also been reported (Rees & Jackson, 2008). Infection of upper leaves are thought to result from inoculum from infected lower leaves, and since the upper eight or nine leaves contribute 75% to 90% of the photosynthetic yield for grainfill, advanced infections can lead to increased yield losses.

No-tillage strategies usually lead to increased accumulation of inoculum in crop residue which serve to initiate earlier infections and more severe disease development during subsequent planting seasons (Ward *et al.*, 1999). The infection usually starts from wind or water dispersal of conidia from the infected debris to lower leaves, where the conidia can germinate on the leaf surface within three hours (Latterell & Rossi, 1983), with 90 – 100% of conidia germinating within 12 hours (Beckman & Payne, 1982). Following germination, stomatal tropism is observed, followed by appressoria formation over stomata. Stomatal penetration by infection hyphae leads to the colonization of the substomatal cavity, with the subsequent internal colonization confined to the air spaces surrounding the mesophyll or parenchymal tissue (Beckman & Payne, 1982). The release of cercosporin is believed to damage cell walls and cause nutrients to leak into the intercellular spaces (Daub & Ehrenshaft, 2000). This process continues until mycelium colonizes much of the accessible leaf volume, though confined by the major leaf veins. Following host tissue necrosis, hyphal strands grow towards the guard cell area of the substomatal cavity and differentiate to form a stroma that fills the cavity. Conidiophores forming from the stroma erupt from the stomatal opening and form conidia for subsequent dispersal and infection. The stroma can also remain intact in maize debris to form inoculum for subsequent seasonal infections (Beckman & Payne, 1982).

Micro-cycle conidiation has been observed in *C. zeae-maydis*, with primary conidia germinating to form secondary conidia without stomatal penetration or resorting to mycelium formation (Lapaire & Dunkle, 2003). This enables a population of conidia to increase its numbers four-fold in two days when formed on trichomes. Multiple generations of microcycle conidiation prove viable, though after five generations conidial viability is lost without growth on a nutrient-rich medium. During low humidity periods these secondary conidia dry out and are wind-dispersed, though rehydration during spells of high-humidity can return these conidia to viability (Lapaire & Dunkle, 2003).

Management strategies for GLS include resistant/tolerant hybrids, crop rotation, tillage practices and foliar fungicide application (Ward *et al.*, 1999 ; Rees & Jackson, 2008). In traditional tillage regions the fungus is exposed to soil microorganisms which generally outcompete it, leading to decreased infection efficiency (Lapaire & Dunkle, 2003). In addition, the fungus infection efficiency is also severely weakened when crops are rotated and the fungus has to survive for an extended time without host interactions. In general the fungus does not survive for longer than a year on diseased maize debris (Latterell & Rossi, 1983).

Due to the practice of planting high-yield hybrids over resistant cultivars, the application of fungicides has traditionally been important in maintaining yields. Studies have confirmed the efficacy in yield increases based on single fungicide applications (Munkvold *et al.*, 2001 ; Dhami *et al.*, 2015), thought the profitability was still linked to prevalent maize prices. The use of resistant cultivars did not show a similar yield increase following fungicide application, and the use of these cultivars is considered to be the most economical and effective choice for the mitigation of GLS in maize (Munkvold *et al.*, 2001).

Genetic studies on the *C. zeae-maydis* population in the USA identified two taxonomically identical, but genetically distinct subgroups, subsequently labelled as *C. zeae-maydis* Group I and Group II (Wang *et al.*, 1998 ; Dunkle & Levy, 2000). The two groups were characterized and found to be two distinct species, i.e. *C. zeae-maydis* and *Cercospora zeina* (Crous *et al.*, 2006). To date, *C. zeae-maydis* infections have not been identified in Africa (Meisel *et al.*, 2009 ; Nsibo *et al.*, 2019), while both species are prevalent in the USA. Several theories on the origin of *C. zeina* have been postulated

(Ward *et al.*, 1999 ; Dunkle & Levy, 2000), although no definitive conclusions have been reached. The significant genetic diversity in the African *C. zeina* population compared to the USA suggests that the pathogen was introduced to the USA from Africa (Wang *et al.*, 1998 ; Dunkle & Levy, 2000). Phylogenetic analysis of ITS regions showed that the *C. zeina* ITS sequence was more similar to the *C. sorghii* var *sorgii* isolate than *C. zeae-maydis*, though the bootstrap support was weak. This suggested that in Africa *C. zeina* might have originated from sorghum but changed its host-specificity to maize (Crous *et al.*, 2006). During the study the phylogenetic analysis using the combined sequences of the ITS, elongation factor 1-alpha, actin, calmodulin and histone H3 regions excluded the *C. sorghii* var *sorgii* isolate sequences, and therefore the origins of *C. zeina* could not be studied using these data.

1.1.2 *Cercospora zeina*

C. zeina is an Acomycete pathogenic fungus of the Class *Dothideomycetes* in the order *Capnodiales*. It was first identified in the USA as *C. zeae-maydis* Group II, but the first report from Africa was in Kwa-Zulu Natal (South Africa) in 1988 (Ward *et al.*, 1999). Subsequently reports indicate an extensive distribution of *C. zeina*, with confirmations of maize infections in Brazil, China, Kenya, Rwanda, Uganda, Zambia and Zimbabwe (Dunkle & Levy, 2000 ; Goodwin *et al.*, 2001 ; Okori *et al.*, 2003 ; Liu & Xu, 2013 ; Neves *et al.*, 2015).

C. zeina was shown to have a slower growth rate *in vivo* and *in vitro* than *C. zeae-maydis*, and also lacks the ability to synthesize cercosporin *in vitro* (Wang *et al.*, 1998 ; Crous *et al.*, 2006). The absence of cercosporin *in planta* has not been confirmed, though the infection efficiency was not perceived to be significantly different from *C. zeae-maydis* when the slower growth rate was taken into account (Wang *et al.*, 1998). Morphological studies on the conidiophores and conidia of *C. zeina* and *C. zeae-maydis* indicated few apparent differences in Wang *et al.* (1998), since the general morphology of these structures were found to be variable and dependent on environmental conditions. However, Crous *et al.* (2006) did find that *C. zeina* has shorter conidiophores (up to 100μm) compared to *C. zeae-maydis* (180μm).

Due to the high humidity required for successful GLS infection (Paul & Munkvold, 2005), *C. zeina* is not endemic to all regions where maize is cultivated. In South Africa, GLS is endemic in the subtropical regions of KwaZulu-Natal and the Eastern Cape, with outbreaks more common during years with average or above-average rainfall (Berger, D. K., personal communication). As mentioned in Latterell *et al.* (1983) when studying *C. zeae-maydis*, the largest factor for increased disease severity appears to be high relative humidity, while optimal temperature and rainfall did not play as big a role, except in the promotion of the required humidity levels.

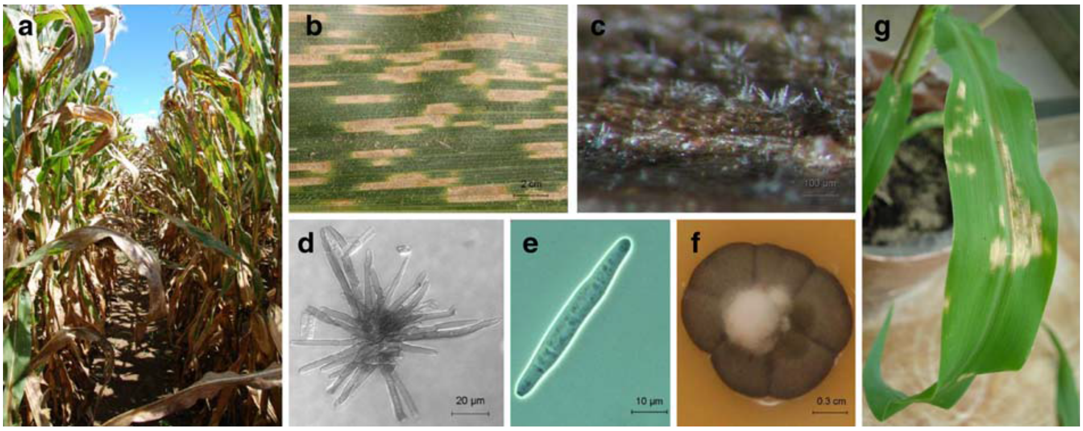


Figure 1.2 ***C. zeina*** **infection and conidiophore and conidial morphology.** *(**a**) GLS symptoms in a maize field, (**b**) GLS lesions on the maize leaf surface, (**c**) conidiophores on leaf surface, (**d**) conidiophores, (**e**) conidia, (**f**)* C. zeina in vitro*, and (**g**) artificially inoculated maize leaf. Source: (Meisel et al., 2009)*

Control of *C. zeina* infections using fungicide application has been successful, and the active ingredients normally include members and combinations from the Quinine outside inhibitor (Strobilurin), Demethylation inhibitor (Triazole) and Multi-site activity classes (Smith, 2013). Although fungicides can greatly increase yield, there is a negative impact on the environment. Detrimental effects on beneficial insects are only one of these, while fungal resistance to generally used fungicides is an increasing concern. In addition, some fungicidal classes have an impact on photosynthesis and stomatal function, thereby decreasing the efficacy and long term usage potential (Petit *et al.*, 2012 ; Smith, 2013)

1.1.3 Cercosporin

Cercosporin is a phytoactive toxin produced by a large number of *Cercospora* species, and is part of the perylenequinone family of molecules. Perylenequinone toxins are produced by at least eight genera of fungal pathogens, and require light activation to produce activated oxygen species which damage plant cell walls and assist pathogenesis (Daub & Ehrenshaft, 2000). The cercosporin toxin, originally isolated from *C. kikuchii* cultures (Kuyama & Tamura, 1957), is produced via a polyketide synthesis pathway, and is distinguishable as a red pigment in culture. The requirement for light for successful *Cercospora* pathogenesis is a strong motivation for the importance of cercosporin during infection. It was shown that the absence of light did not affect stomatal penetration, but did decrease infection efficiencies and symptom severity (Daub & Ehrenshaft, 2000).

The biosynthetic pathway for cercosporin was characterized in *C. nicotianae*, and consists of eight genes in a CTB biosynthetic cluster (Newman & Townsend, 2016). The presence of the conserved cluster has also been confirmed in *C. zeina, C. zeae-maydis* (Swart *et al.*, 2017), *Cercospora beticola*, *Cercospora berteroae* and *Cercospora canescens* (de Jonge *et al.*, 2018). In *C. zeina* the lack of cercosporin production *in vitro* was shown to be due to a mutation in a FAD-dependent monooxygenase gene (CTB7) leading to the creation of a pseudogene (Figure 1.3). Complementation with the *C. zeae-maydis* gene restored cercosporin production *in vitro* (Swart *et al.*, 2017). The study by de Jonge *et al.* (2018) also used a gene knock-out approach to show that four adjacent genes are also required for cercosporin synthesis, and that the cluster is actually composed of 12 genes. The presence and synteny of these four additional genes was confirmed in *C. zeina*, though no gene expression analysis has been published.

The cytotoxic effect of cercosporin on plant cells have been extensively studied, but recently cercosporin was shown to also exhibit cytotoxic behavior against human tumour cells (Mastrangelopoulou *et al.*, 2018). The question of resistance against cercosporin, especially by the producing fungi, becomes relevant in this light. Multiple mechanisms have been proposed for the resistance of *Cercospora* species, though the primary mechanism appears to be the reversible reduction of the perylenequinone moiety to dihydrocercosporin, a phyto-inactive molecule. This molecule is spontaneously oxidized to the phytoactive cercosporin following export from the fungus (Newman & Townsend, 2016). Other mechanisms entail the presence of membrane transporter proteins which cause the efflux of cercosporin from fungal hyphae into the plant. CFP from *C. kikuchii* is a Major Facilitator Superfamily (MFS) transporter required for both cercosporin production and resistance (Callahan *et al.*, 1999), while the CTB biosynthesis cluster MFS transporter (CTB4) is required for the production of, though not resistance to cercosporin. The ATP-Binding Cassette (ABC) transporters ATR1 and *CnATR2* were found to be required for resistance to cercosporin in *C. nicotianae* (Amnuaykanjanasin & Daub, 2009 ; Beseli *et al.*, 2015). The use of transporters have been considered for host resistance engineering, and subsequently the *cpd1* (cercosporin photosensitizer detoxification) gene from *Saccharomyces cerevisiae* was expressed in tobacco, conferring resistance to cercosporin (Panagiotis *et al.*, 2007). This result increased the likelihood that transgenic crops expressing these transporters could enhance resistance to cercosporin-producing pathogens in future.

The study of the infection strategies by the maize-infecting *Cercospora* species have been hampered by the absence of genome sequences for the two species. Although the genome of *C. zeae-maydis* was sequenced in 2011 (JGI, 2011), no subsequent functional analysis of the genome in terms of infection and physiology has been published to date. The use of fungal genome assemblies for the study of infection strategies and the identification of novel effector molecules is a tested approach, and has been successful in various studies (Amselem *et al.*, 2011 ; de Jonge *et al.*, 2018 ; Hane *et al.*, 2007 ; Ohm *et al.*, 2012).

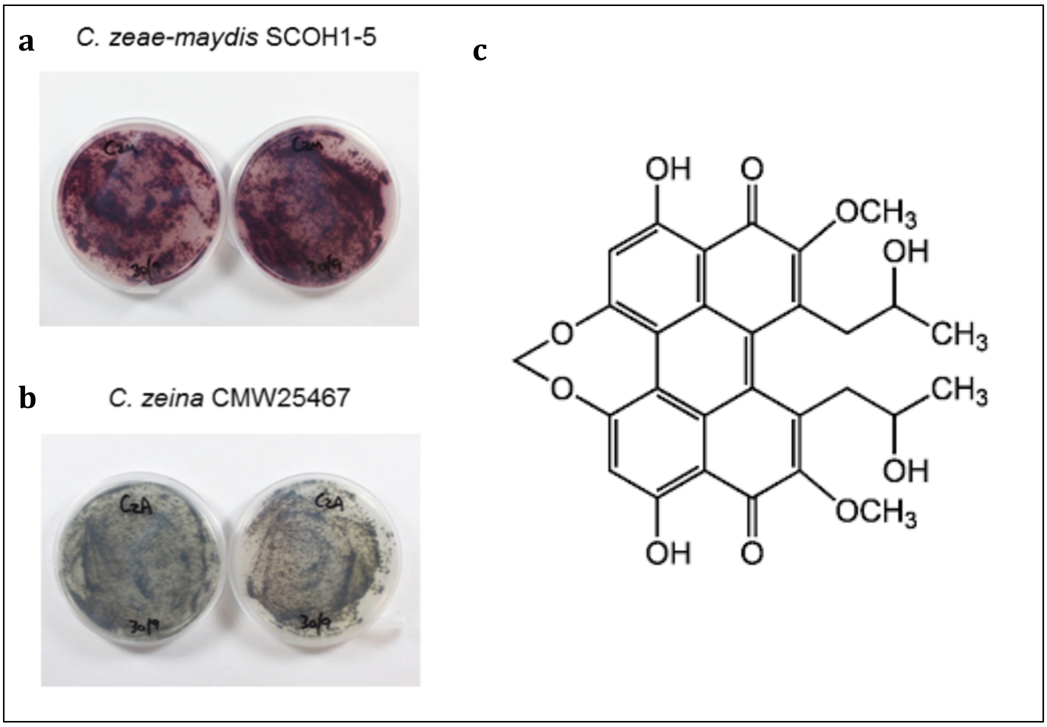


Figure 1.3 ***In vitro* cercosporin production by *C. zeae-maydis and C. zeina.*** *(**a**) Presence of red cercosporin in* C. zeae-maydis *cultures, (**b**) Absence of red cercosporin in* C. zeina *cultures, (**c**) Structure of cercosporin. Sources: a-b: (Swart et al., 2017), c: AdipoGen Life Sciences (https://adipogen.com/ag-cn2-0111-cercosporin.html/).*

1.1.4 Ascomycete pathogenic fungi

The Fungal Kingdom is comprised of one subkingdom, seven phyla and ten sub-phyla. The subkingdom *Dikarya* includes mainly the Ascomycota and Basidiomycota phyla (Ebersberger *et al.*, 2012), which jointly contain a large number of plant pathogenic (phytopathogenic) species (Hibbett *et al.*, 2007 ; Doehlemann *et al.*, 2017). These infect a diverse and large number of plant species, and differ in their lifestyles and infection strategies. Pathogenic fungi colonizing and obtaining nutrients exclusively from living plant tissue are classified as biotrophs, while those infecting and killing host cells and feeding on dead or necrotic tissue are classified as necrotrophs. Hemibiotrophs are pathogenic fungi which infect and initially feed on living tissue, but switch to a necrotrophic stage after killing host cells (Koeck *et al.*, 2011).

Many Ascomycete phytopathogenic species have the ability to reproduce both sexually and asexually, producing spores or similar structures which are spread by wind or water. However, fungi reproducing sexually do not occur as male or female phenotypes, but rather as one of several mating types (bipolar in the case of Ascomycetes) distinguishable only at the molecular level. Haploid cells of compatible (opposite) mating types fuse, followed by nuclear fusion to produce new diploid cells. A final meiosis step results in new haploid cells, with possible nuclear recombination resulting during the nuclear fusion process (Heitman *et al.*, 2013). Although no sexual reproductive structures have been observed for many of these species, the presence of sexual reproduction is often inferred from the presence of bipolar mating type genes. The ratio of these genes in fungal populations can also be an indication of the presence of sexual reproduction (Heitman *et al.*, 2013). Although wild-type populations have been found to retain the required mating type genes, the occurrence of sexual reproduction appears to be uncommon, especially in host-specific pathogens where meiosis and the potential shuffling of alleles could lead to the loss of critical virulence factors. It is hypothesized that sexual reproduction evolved in response to environmental stresses, especially for the repair of DNA damage via homologous recombination (Wallen & Perlin, 2018). Therefore the retention of mating-type genes, in the absence of wide-spread sexual reproduction, possibly confers improved evolutionary fitness during increased environmental stresses, while the removal of these genes disables a major mechanism for DNA repair (Heitman *et al.*, 2013 ; Wallen & Perlin, 2018).

Asexual reproduction in Ascomycetes is a mitosis-driven process where conidia are produced from specialized structures known as conidiophores. These structures which normally show distinctive morphological differences between species, have branched ends where conidia are produced, bud off and are dispersed. Each conidium is genetically identical to the parent cell and allow fungi to rapidly reproduce and multiply (Taylor *et al.*, 1999). The formation of conidial anastomosis tubes fusing adjacent germinating conidia might lead to the exchange of genetic material between different conidial genotypes, even between incompatible genotypes. This mechanism could explain some chromosomal variation and recombination in the absence of recognized sexual recombination (Gabriela Roca *et al.*, 2005).

Following the dispersal of spores or conidia, the initial stage in host leaf penetration is the attachment to the host surface. To prevent the spores or conidia from being washed or blown from the leaf surface the secretion of extracellular matrix is crucial, although a hydrophobic interaction between the conidium and cuticle might constitute the initial binding phase (Mendgen *et al.*, 1996). Generally fungal adhesives are composed of water-insoluble glycoproteins, while lipids and polysaccharides have also been detected in the adhesive material (Tucker & Talbot, 2001). It has been shown in the rice blast fungus *Magnaporthe oryzae* that mucilage, a ubiquitous mixture including high-molecular weight glycoproteins (Qu *et al.*, 2017), is secreted from the periplasmic compartment at the conidial apex upon hydration of the conidium, and aids in attachment on the host surface (Hamer *et al.*, 1988 ; Doehlemann *et al.*, 2017). Mucilage of spores and conidia is required for enhanced pathogenicity in a range of plant and insect infecting fungi (Qu *et al.*, 2017).

The role and importance of cutinases in pathogenesis have been well studied, and the enzyme has been detected in conidial extracellular matrix during adhesion of *Colletotrichum graminicola* (Mendgen *et al.*, 1996). It is hypothesized that cutinases degrade the cuticle to decrease the hydrophobicity of the leaf surface to enhance conidial attachment (Nicholson *et al.*, 1993). Studies on *Fusarium solani* f. sp. *pisi* have shown that the infection rate decreases in the presence of cutinase-specific inhibitors and antibodies, indicating that the enzyme is important for pathogenesis. It was also shown that the expression of cutinase genes increases in the presence of cuticle breakdown products. To date multiple cutinase paralogs have been identified for several species, with each enzyme active during different life stages e.g. saprophytic or biotrophic (Mendgen *et al.*, 1996).

Fungi directly penetrating hosts require cell-wall degrading enzymes to successfully infect. The presence of polygalacturonase and polygalacturonate lysases were confirmed at infection sites, though only antibodies targeting the lyases provided protection to hosts (Mendgen *et al.*, 1996). The role of pectic enzymes have been shown to be important during pathogenesis for *Botrytis cinerea*, while *Glomerella cingulate* does not require these for penetration (Mendgen *et al.*, 1996). The redundancy and variable regulation of cell-wall degrading enzymes make these enzymes difficult to study, since the presence of cell-wall derived polymers can lead to the expression of a range of degrading enzymes in culture, including polygalacturonases, pectin and pectate lyases, and pectin methylesterases. The role of these enzymes in cell-wall degradation during host penetration has not been confirmed, since these enzymes might also play a role during saprophytic growth (Mendgen *et al.*, 1996). Expressed proteases have been detected in several species, though the precise role has not been confirmed. In *Colletotrichum graminicola,* a metalloprotease is required for virulence on maize leaves, with increased maize chitinase activity shown in the absence of the protease. Therefore protease effectors could be important for modulating the plant immune system, especially for biotrophs (Sanz-Martin *et al.*, 2016). Lignin, a phenolic compound, is an important structural component of plant cell walls and lignin deposition is a plant immune response mechanism to strengthen cell walls to prevent further pathogen ingress (Kombrink & Somssich, 1995). Studies have shown that the inhibition of laccases (enzymes acting on phenolic compounds) provide protection against *B. cinerea*, indicating that the breakdown of lignin by these pathogen enzymes is a crucial process in host infection (Mendgen *et al.*, 1996).

Once conidia germinate following successful adhesion, they produce primary hyphae, also known as germinating tubes, which differentiate into hyphae and mycelium. Primary hyphae show polarized cell growth on the host surface in response to physical or chemical stimuli. The mechanism of stimuli recognition by the fungi is not known, although the G-protein-coupled receptor, Pth11, and related G-α- and G-βγ-subunit proteins are required for further infection stages (DeZwaan *et al.*, 1999 ; Wilson & Talbot, 2009 ; Doehlemann *et al.*, 2017). Host penetration is normally facilitated by bulges in the runner hyphae, known as appressoria, and there are three invasion strategies involving their formation. The first type involves the formation of an appressorium with chitin enriched cell walls and a melanized inner layer. The melanin layer assists with the passage of water through the plasma membrane, but not osmotically active compounds such as glycerol. The increase in the concentration of these osmotically active compounds inside the appressorium leads to the increased influx of water, thereby greatly increasing the turgor pressure. The formation of penetration hyphae, assisted by the turgor pressure, leads to the penetration of the host tissue. The presence of fungal proteases and cell wall degrading enzymes could also assist in the weakening of the host cell walls. Once the host tissue has been penetrated, the invasion hyphae can invade inter- and intracellular spaces (Mendgen *et al.*, 1996 ; Tucker & Talbot, 2001 ; Doehlemann *et al.*, 2017). The second type is the formation of a hyphal tip swelling which differentiates into an appresorial-like structure, but lacking the melanin layer. The process of host penetration is through biolytic action rather than physical force, as in the case of *B. cinerea* (Doehlemann *et al.*, 2017). The third type involves the invasion of host tissue directly through open stomata, such as the invasion of *Passalora fulva* to colonize the extracellular space of tomato leaves (Thomma *et al.*, 2005 ; Doehlemann *et al.*, 2017). In some species, e.g. *C. zeae-maydis* the appressoria form directly over stomata, with penetration pegs entering the open stomatal structures without the need for cell wall penetration (Beckman & Payne, 1982).

The plant immune system have the ability to recognize the presence of pathogenic fungi on different levels (Jones & Dangl, 2006). The initial recognition is usually of important molecules which are an integral part of the pathogen cell wall structure and can therefore not easily be altered for immune system evasion. These molecules are known as pathogen or microbe-associated molecular patterns (PAMPs/MAMPs) and are recognized by plant pattern recognition receptors (PRRs) (Jones & Dangl, 2006 ; Koeck *et al.*, 2011). In fungi, the cell wall is primarily composed of chitin polymers of linear β-1,4-linked N-acetyl-glucosamines. Plant chitinases in the apoplast degrade fungal chitin to monomers, leading to the recognition of the monomers by plant receptors on the cell membrane (Zipfel, 2009). These are usually transmembrane proteins containing extracellular lysin motif (LysM) domains which recognize N-acetylglucosamine ligands such as chitin monomers (Gust *et al.*, 2012 ; Tanaka *et al.*, 2013). Although intracellular protein kinase domains have only been detected in some chitin receptors, all have been found to be essential for chitin-triggered innate immune responses (Tanaka *et al.*, 2013).

Recognition of chitin leads to one example of PAMP-triggered immunity (PTI) (Figure 1.4), which is regarded as part of the basal or innate plant immune system targeting a broad range of microbes (Li, B. *et al.*, 2016). The plant response to PTI includes the generation of reactive oxygen species (ROS) (Torres *et al.*, 2006 ; Lehmann *et al.*, 2015), pathogenesis-related proteins such as PR-1 (Breen *et al.*, 2017), ethylene and salicylic acid (SA), and compounds required for strengthening the cell-wall such as callose (Li, B. *et al.*, 2016). In addition, the activation of the phenylalanine ammonia-lyase (PAL) enzyme is important against necrotrophic fungi, since transgenic tobacco plants with suppressed PAL activity showed a marked increase in pathogenicity by *C. nicotianae*

(Maher *et al.*, 1994). PAL is involved in the SA and lignin biosynthesis pathways, both of which are components of the plant immune response (Kachroo *et al.*, 2016).

Effector triggered immunity (ETI) is generally triggered when the inter- or intracellular effectors secreted by a pathogen are detected (Figure 1.4), and is generally a more pathogen/race specific response to host-adapted pathogens (Li, B. *et al.*, 2016). The main response during ETI is the hypersensitive response (HR) which is activated upon binding of a pathogen effector molecule to a resistance (R) protein. This leads to localized cell-death and denies the flow of nutrients to invading biotrophic pathogens. Repeated cycles of effector mutation and R-protein evolution is involved in the successful colonization by pathogens, or plant resistance to pathogenesis (Jones & Dangl, 2006).

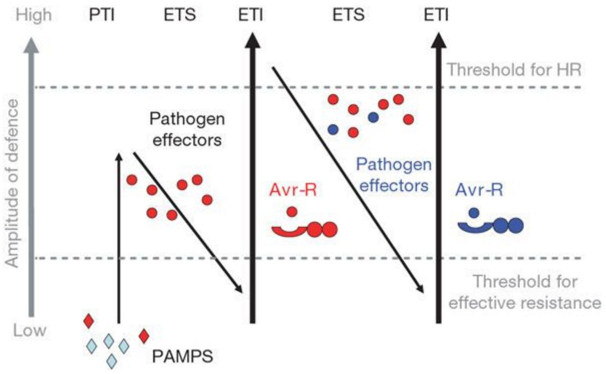


Figure 1.4 The zigzag model of plant-pathogen interaction as a function of the quantitative plant immune response (Jones & Dangl, 2006). *In phase 1 the plant detects the presence of the pathogen associated microbial patterns (PAMPs) to trigger PAMP-triggered immunity (PTI). Effectors secreted by the pathogen suppress PTI during phase 2, resulting in effector triggered susceptibility (ETS). During phase 3 an effector (red) is recognized by a NB-LRR protein, activating effector triggered immunity (ETI), thereby initializing the hypersensitive response (HR) cell death when a threshold is crossed. Pathogens negative for the red effector are selected and through new or altered effectors again suppress ETI in phase 4. In subsequent phases plants select for new NB-LRR recognition proteins for activating PTI, while new effectors from pathogens subsequently aim to activate ETS.*

Processes following host penetration depend on the life stage of the invading pathogen. In biotrophic species, the invading penetration peg usually develops a specialized hypha called haustorium. These structures form behind the plant cell walls without damaging the cell membrane (Figure 1.5). The resultant complex contains the haustorium invaginated by the host plasma membrane, while the haustorium is enveloped by a polysaccharide-rich extra-haustorial matrix. Secretion of a range of fungal effectors modulates the host defense to block the HR, while haustorial transporters take up nutrients from the host cell (Mendgen & Hahn, 2002 ; Doehlemann *et al.*, 2017). The haustorial membrane differs in composition from primary and invasive hyphae, with altered carbohydrate and glycoprotein compositions (Mackie *et al.*, 1993), as well as amino acid transporters specific to the structure (Hahn *et al.*, 1997). The invaginated host plasma membrane is also a highly specialized structure, with the lipid and protein composition altered to suit the fungal nutrient uptake mechanism (Perfect & Green, 2001). In addition, the membrane displays increased thickness and folds, presumably increasing the surface area for fungal contact and nutrition (Manners & Gay, 1983). Alternatively, biotrophic species like *P. fulva* exclusively occupies the intracellular and apoplastic spaces of tomato leaves without intracellular invasion (Bolton *et al.*, 2008).

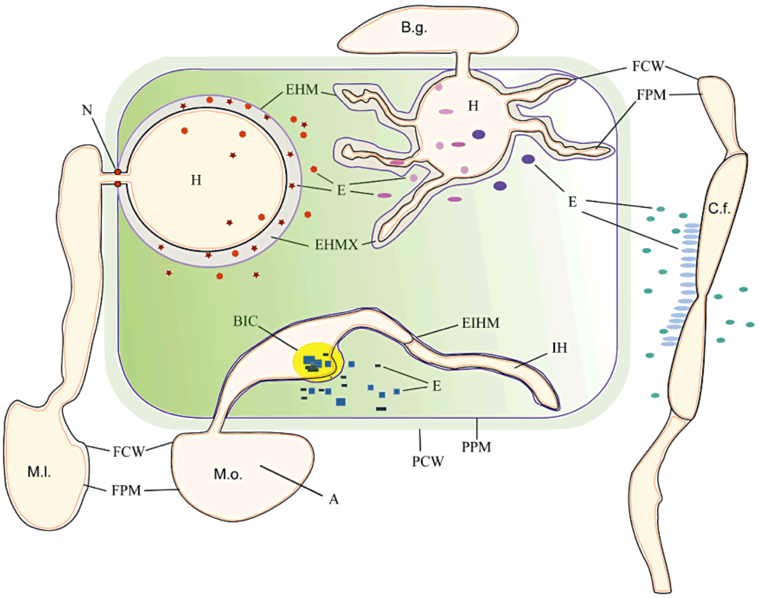


Figure 1.5 Overview of inter and intracellular fungal infection structures (Koeck *et al.*, 2011). *The two biotrophic Ascomycetes employ different strategies, with* Passalora fulva *(**C.f.**) occupying the intercellular space, while* Blummeria graminis *(**B.g.**) penetrates the host membrane and obtains nutrients via a haustorium (**H**). The hemibiotrophic Ascomycete* Magnaporthe oryzae *(**M.o.**) utilizes invasion hyphae during the biotrophic phase, and the Basidiomycete* Melampsora lini *(**M.l.**) also utilizes a haustorial structure. All species secrete effectors (**E**) to modulate the host immune system and increase nutrition access and acquisition.*

***A**, appressorium; **BIC**, biotrophic invasive complex; **E**, effector; **EHM**, extrahaustorial membrane; **EHMX**, extrahaustorial matrix; **EIHM**, extrainvasive hyphal membrane; **FCW**, fungal cell wall; **FPM**, fungal plasma membrane; **H**, haustorium; **IH**, invasive hyphae; **N**, neckband; **PCW**, plant cell wall; **PPM**, plant plasma membrane.*

Biotrophic fungi secrete a large number of diverse protein effectors which typically show little sequence similarity to know proteins and have functions which are largely unknown (Koeck *et al.*, 2011). Effectors such as Avr4 and Ecp6 have been found to be crucial for successful biotrophic fungi colonization, since these effectors contain LysM carbohydrate-binding domains similar to the LysM motifs present in host PRRs. In this case, Avr4 binds chitin in fungal hyphae to protect against host chitinases, while Ecp6 binds chitin monomers liberated from the fungal cell wall by chitinase action (De Jonge *et al.*, 2011). By sequestering these breakdown products of chitin before the PRRs can recognize chitin and launch PTI, fungi can successfully colonize plant cells without activating the HR (Dolfini *et al.*, 2009 ; Stergiopoulos & de Wit, 2009 ; De Jonge *et al.*, 2011). In fact, the inhibition of plant cell death is crucial for biotrophic survival, and *M. oryzae* was found to secrete 11 suppressors of programmed cell death (PCD) during its biotrophic phase (Fernandez & Orth, 2018). An additional strategy of biotrophic fungi is the downregulation of the SA signaling and systemic acquired resistance pathways to evade the plant immune response (Tanaka *et al.*, 2015). An example is the fungal vascular wilt pathogen *Verticillium dahlia* which secretes the isochorismatase enzymes PsIsc1 and VdIsc1, blocking the synthesis of SA (Tanaka *et al.*, 2015).

Necrotrophic pathogens infecting a narrow host range generally rely on the production of host-specific toxins (HST) which interact with specific host genes (Table 1.1) (Greenberg & Yao, 2004 ; Oliver & Solomon, 2010). This process is similar to the interaction of effector proteins with resistance, or R-proteins (Doehlemann *et al.*, 2017). Host proteins interacting with these toxins are also called susceptibility (S) proteins, and only one dominant copy of the specific S-protein gene is required to trigger host susceptibility (Friesen *et al.*, 2006). These interactions are thought to be deliberate by the pathogen to activate plant ETI responses, thereby leading to hypersensitive responses and the localized cell-death required for the pathogen to colonize the host. These toxins have thus also been classified as effector molecules (Hammond-Kosack & Rudd, 2008 ; Faris *et al.*, 2010 ; De Jonge *et al.*, 2011). Therefore, plants lacking the specific R-protein linked to the HST will show resistance to the fungus, hence the narrow host-specificity of these fungi. It has not been shown that all narrow host-range fungi produce these HSTs, although alternative mechanisms for host-specificity have not been confirmed.

Necrotrophic pathogens infecting a broad host range lack any HST, and therefore host resistance to these pathogens is more complex, typically being quantitative in nature (Oliver & Solomon, 2010 ; Doehlemann *et al.*, 2017). These fungi normally target the host cell death (apoptosis) pathways, with initial local necrosis followed by a more general PCD (Greenberg & Yao, 2004 ; Amselem *et al.*, 2011). Though not all the components of the host general PCD pathway(s) have been described to date, HR is preceded by an oxidative burst, ion channel activity, Nitric Oxide (NO), as well as the interaction between some of these different signals (Greenberg & Yao, 2004) while oxalic acid has also been implicated in *Sclerotinia sclerotiorum*-mediated PCD (Doehlemann *et al.*, 2017). The idea that necrotrophic fungi release effectors solely to affect cell death for nutrition is not conclusive, since limiting host ROS secretion is also an important consideration (Oliver & Solomon, 2010). However, there is evidence for *L. maculans* contributing to ROS production and secretion to exploit the host's oxidative burst for pathogenesis (Li, C. *et al.*, 2008).

Table 1.1 Host-specific toxins produced by narrow host-range necrotrophic fungi.

Pathogen	Host Specific Toxin(s)	Reference
Alternaria alternate	AAL-toxin Ta, Tb / AF-toxin I, II, III	(Oliver & Solomon, 2010 ; Tsuge *et al.*, 2013)
Alternaria citri	ACT-toxin I, II / ACR-toxin I	(Tsuge *et al.*, 2013)
Alternaria kikuchiana	AK-toxin I, II	(Tsuge *et al.*, 2013)
Alternaria longipes	AT-toxin	(Tsuge *et al.*, 2013)
Alternaria mali	AM-toxin I, II, III	(Tsuge *et al.*, 2013)
Botrytis cinerea	NEP1-like	(Oliver & Solomon, 2010)
Cochliobolus carbonum	HC-Toxin / Tox1	(Friesen *et al.*, 2008 ; Oliver & Solomon, 2010)
Cochliobolus heterostrophus	T-toxin	(Oliver & Solomon, 2010)
Cochliobolus victoriae	Victorin	(Oliver & Solomon, 2010)
Corynespora cassiicola	Cassiicolin	(Friesen *et al.*, 2008)
Phyllosticta maydis	PM-Toxin	(Friesen *et al.*, 2008)
Pyrenophora tritici-repentis	ToxA	(Oliver & Solomon, 2010)
Rhynchosporium secalis	NIP1	(Oliver & Solomon, 2010)
Stagonospora nodorum	Tox3 / SnTox1 / SnTox2 / ToxA	(Friesen *et al.*, 2008 ; Oliver & Solomon, 2010)

A deluge of secreted biolytic and cell wall degrading enzymes have been characterized in various necrotrophs, while the diversity and redundancy in the range of these enzymes indicate their importance in infection and necrosis (Doehlemann *et al.*, 2017). These enzymes include carbohydrate-active enzymes (Zhao *et al.*, 2013), proteases (Yike, 2011) and lipases (Subramoni *et al.*, 2010). Most of the necrotrophic fungal effectors are classified as small secreted proteins, while many of these are cysteine rich (Stergiopoulos & de Wit, 2009), with disulfide bridges contributing to stability against protease action (Bolton *et al.*, 2008). The mode of action of these effectors are believed to involve direct or indirect interaction with host susceptibility gene products or receptors, or evasion of the host biolytic enzymes (Oliver & Solomon, 2010). The functions of a few effectors have been characterized, and they include protease inhibitors (Sabotič & Kos, 2017) and PAMP-scavenging proteins (De Jonge *et al.*, 2011), while the *S. sclerotiorum* SsSSVP1 effector dislocates the cytochrome b-c1 complex subunit 8 from the mitochondria to the cytoplasm, leading to PCD (Lyu *et al.*, 2016). A number of effectors also target the jasmonic acid (JA) and ethylene signaling pathways to downregulate immune responses during early infection, e.g. by activating the SA pathway instead (Zhu *et al.*, 2013). During its necrotrophic phase, *M. oryzae* secretes a monooxygenase which hydroxylates endogenous free JA to disrupt the JA pathway (Fernandez & Orth, 2018). Finally, other classes of effectors include small RNAs delivered by *B. cinerea* which bind to the *Arabidopsis* and tomato Argonaute 1 proteins to selectively suppress host immune gene expression (Weiberg *et al.*, 2013).

The hemibiotrophic fungal lifestyle combines structural and effector elements of both biotrophic and necrotrophic fungi, since these pathogens need to keep the host alive during the initial biotrophic phase, only transitioning to a necrotrophic phase later in the infection cycle. The classification of true hemibiotrophs is not absolutely defined, since any species not typified as true biotrophs or necrotrophs are potentially hemibiotrophic in nature. Therefore some species belonging to the *Fusarium*, *Verticillium* and *Mycosphaerella* genera are classified as hemibiotrophic due to their initial latency period, but many have been found to lack the typical biotrophic feeding structures. Furthermore, they may remain in the apoplastic or intercellular spaces and refrain from close host cell contact (Doehlemann *et al.*, 2017). In true hemibiotrophs, the biotrophic phase can last from days to several months depending on the species and host (Doehlemann *et al.*, 2017).

While biotrophic fungi generally use haustorial-structures to obtain nutrients from plant cells, hemibiotrophic fungi such as *Colletotrichum lindemuthianum* (Figure 1.6) rather use invasive hyphae which interact with much less specialized invaginated plant plasma membranes than the haustorial structures formed by biotrophic fungi, probably reflecting the shorter biotrophic interaction required by the pathogen (Perfect & Green, 2001). It was confirmed in *Colletotrichum gloeosporioides* that intracellular/invasive hyphae participate in nutrient uptake, and that these structures were analogous to biotrophic haustoria (Wei *et al.*, 2004). Generally invasive hyphae are restricted to a single cell, but these might spread to other cells via plasmodesmata (Fernandez & Orth, 2018) during the advent of host necrosis, while the formation of secondary hyphae which are un-encapsulated by the host membrane are features of the necrotrophic phase (Mendgen & Hahn, 2002). In the specific case of *C. lindemuthianum* the spreading intercellular hyphae can re-differentiate into biotrophic stage feeding structures in adjacent cells, thereby re-initiating the biotrophic phase (Perfect & Green, 2001).

During infection of rice (*Oryza sativa*), the hemibiotroph *M. oryzae* develops an invasive hyphae containing a lobed structure with accumulated effector proteins. This structure is known as the biotrophic interfacial complex (BIC), and new BICs are formed in invasive hyphae infecting adjacent cells (Koeck *et al.*, 2011). Secreted effectors from the BIC not only have a local function, but also move to adjacent cells to prepare these cells for invasive hyphae proliferation (Koeck *et al.*, 2011 ; Fernandez & Orth, 2018).

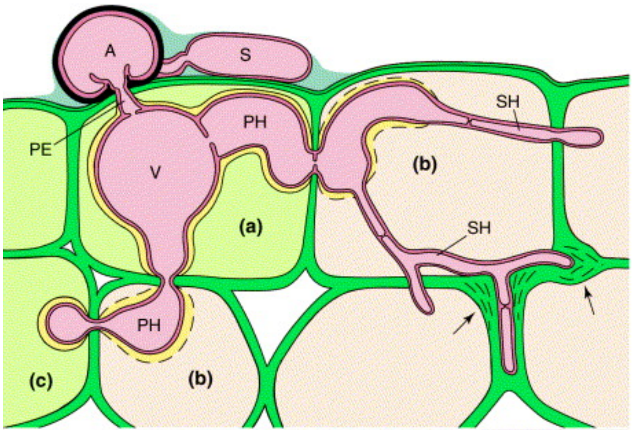


Figure 1.6 Infection strategy of the hemibiotroph *Colletotrichum lindemuthianum* (Mendgen & Hahn, 2002). *Following the germination of a spore (**S**) on the host surface, the formation of an appressorium (**A**) leads to host cell penetration via a penetration hypha (**PE**). During the biotrophic phase (**a**) a vesicle (**V**) and primary hypha (**PH**) form from the penetration hypha, both of which are surrounded by the invaginated plant plasma membrane. Several days after colonization the plant plasma membrane disintegration leads to host cell death during the necrotrophic phase (**b**). New primary hyphae colonize adjacent cells, with the short biotrophic phase again followed by the necrotrophic phase (**c**), a sequence repeated until narrow secondary hyphae (**SH**) are formed. These are not surrounded by a host membrane and the secretion of cell-wall degrading enzymes (indicated by arrows) lead to large scale host cell degradation.*

The switch from biotrophic to necrotrophic stages have not been studied for most hemibiotrophs, though the *CLTA1* gene, encoding a GAL4-like transcriptional activator, was found to be important in *C. lindemuthianum* (Dufresne *et al.*, 2000 ; Oliver & Ipcho, 2004). In addition, expression of the *Colletotrichum* CIH1 gene was found to be important for establishing biotrophy, but expression stopped at the onset of necrotrophy. This suggests an important role for the proline-rich glycoprotein during biotrophy, with the protein localized to the biotrophic interface (Perfect *et al.*, 2000). Necrosis is mediated by, amongst others, fungal effectors recognized by host R-proteins leading to HR, along with the secretion of numerous cell-wall degrading enzymes (Koeck *et al.*, 2011). In *C. gloeosporioides* the release of a pectate lyase was crucial for both cell wall degradation, but also the reduction of host defense reactions. The concurrent release of ammonia counteracts acidic pH levels to optimize pectate lyase activity (Mendgen & Hahn, 2002). In *M. oryzae* the secreted MSP1 protein induces peroxide production and PCD in rice and barley leaves, increasing pathogenicity (Doehlemann *et al.*, 2017).

1.1.5 The *Cercospora* genus

The Dothidiomycetes is the largest class of the Ascomycete fungi, and is also regarded as the most ecologically diverse (Ohm *et al.*, 2012). The class contains 12 orders, 90 families, 1,300 genera and more than 19,000 species have been identified (Schoch *et al.*, 2009 ; Ohm *et al.*, 2012). Plant pathogens have been found to occur in six of these 12 orders, and they cause some of the most economically important diseases in cereals, trees, dicots and tropical fruit (Ohm *et al.*, 2012).

The *Cercospora* genus, in the family *Mycosphaerellaceae*, order Capnodiales and class Dothideomycetes was originally thought to include more than 3,000 species (Pollack, 1987 ; Crous & Braun, 2003), with species normally named for the host they infect (Chupp, 1954). Since *Cercospora* species were considered to be host-specific, a new species name was created for each newly discovered isolate, leading to the large number of species in the genus (Chupp, 1954 ; Pollack, 1987). Subsequent divisions of the genus was made on the basis of morphology, using characteristics such as conidiomatal structure, mycelium, conidiophores, conidiogenous cells and conidia (Groenewald *et al.*, 2013). A refinement by Crous and Braun (2003) in particular used the structures of conidiogenous loci and hila, as well as the absence or presence of pigmentation in conidiophores and conidia to clarify species in the genus. As a result of the refinement, almost 50 separate genera were identified in the *Cercospora* complex (Crous & Braun, 2003), while recognizing 659 true *Cercospora* species (Groenewald *et al.*, 2013). A further 281 species were included in the *Cercospora apii s. lat* complex representing species which are morphologically indistinguishable from *C. apii* (Crous & Braun, 2003 ; Groenewald *et al.*, 2013).

The production of cercosporin is virtually universal in the *Cercospora* genus, with very few exceptions. Phylogenetic analyses using the internal transcribed spacer areas (ITS) suggested that the species shared a recent common ancestor and that it was most probably a cercosporin producer (Goodwin *et al.*, 2001). Absence of cercosporin production in related genera suggests a single evolutionary origin of the toxin production, most probably by the recent ancestor, and that non-producers were most likely in different genera (Fajola, 1978 ; Goodwin *et al.*, 2001). Non-producers, confirmed to be in the genus by ITS sequence data most probably lost the ability to produce the toxin during species radiation, as in the case of *C. zeina* where the biosynthesis cluster is present, with only one of the genes (*ctb7*) producing a non-functional protein due to a mutation (Goodwin *et al.*, 2001 ; Swart *et al.*, 2017).

For most *Cercospora* species a sexual stage have not been identified, and where a teleomorph was found these were invariably in the genus *Mycosphaerella* (Goodwin *et al.*, 2001). An unconfirmed *Mycosphaerella* teleomorph for *C. zeae-maydis* was proposed in 1977 (Latterell & Rossi, 1983 ; Goodwin *et al.*, 2001), though no follow-up studies are evident. Analysis of the mating-type genes of selected *Cercospora* species did show that *C. beticola*, *C. zeae-maydis* and *C. zeina* were all heterothallic (Groenewald *et al.*, 2013), and that the two mating types were equally distributed in the populations for all three of the species (Groenewald *et al.*, 2006 ; Muller *et al.*, 2016). The converse was found for *C. apii* and *C. apiicola*, where only one of the mating types were identified in the sampled populations, suggesting that these species did not have a sexual stage, or that sampling was insufficiently deep (Groenewald *et al.*, 2006).

The proposed host specificity of the *Cercospora* species is also not universally factual, since species of *C. apii*, initially discovered on celery (Groenewald *et al.*, 2013) have been found on plant hosts in 86 genera of plant families (Crous & Braun, 2003). Similarly, *C. beticola*, which is very closely related to *C. apii*, has also been found on members of the *Apium, Chrysanthemum, Limonium, Malva, Spinacia* genera (Groenewald *et al.*, 2006), although it is mainly associated with pathogenicity of sugar beet, *Beta vulgaris* (Weiland & Koch, 2004). The possibility has been postulated that *Cercospora* species are pathogenic mainly on their respective natural hosts, but that some could infect secondary hosts while the natural host is not available (Groenewald *et al.*, 2005), and that during these events the non-host infecting *Cercospora* species could rather be classified as saprobes or secondary invaders (Crous & Groenewald, 2005).

To date the standard loci used for reconstructing the phylogeny of *Cercospora* species have comprised the ITS areas, including the adjacent 5.8S rRNA gene of the nuclear ribosomal DNA operon (Goodwin *et al.*, 2001 ; Meisel *et al.*, 2009 ; Groenewald *et al.*, 2013), actin (Groenewald *et al.*, 2013), translation elongation factor 1-α (TEF1) (Meisel *et al.*, 2009 ; Groenewald *et al.*, 2013), calmodulin (Groenewald *et al.*, 2013) and histone H3 (Meisel *et al.*, 2009 ; Groenewald *et al.*, 2013). Of these loci, the use of ITS was found to be a good indicator of genus identity, but was poor at species-level classification, and that a multi-locus approach was required (Groenewald *et al.*, 2013).

Further phylogenetic analysis of the *Cercospora* genus performed by Groenewald *et. al.* (2013) identified 19 *Cercospora* species groups (named *Cercospora* sp. A-S) which could not reliably be classified as any known *Cercospora* species based on the available loci sequence data (Groenewald *et al.*, 2013). In addition to the standard loci, data from additional loci were unsuccessfully included to attempt the resolution of the cryptic taxa in *Cercospora* sp. Q group. These loci included the genes for glyceraldehyde-3-phosphate dehydrogenase (GAPDH), chitin synthase (CHS), beta-tubulin, a mini-chromosome maintenance complex component 7 (MCM7), as well as part of the mitochondrial small subunit rRNA gene (Groenewald *et al.*, 2013).

An analysis of *Cercospora* species in Iran used the five standard loci, with the addition of the second largest subunit of RNA-polymerase II and new primers for both GAPDH and tubulin to resolve cryptic taxa and develop a DNA barcode for the *Cercospora* genus (Bakhshi *et al.*, 2018). The additional loci was useful in classifying some of the cryptic species, but did not fully resolve these groups, while an additional cryptic group, *Cercospora* sp. T. (Bakhshi *et al.*, 2015) was also unresolved. In addition, none of the genes provided the resolution required to be a barcode for the genus, and therefore the need exists for additional loci to expand or replace the current eight gene multi-locus approach (Bakhshi *et al.*, 2018).

1.2 References

Alexandratos, N., and Bruinsma, J. (2012) World Agriculture Towards 2030/2050: The 2012 Revision. *ESA Working paper No. 12-03. FAO, Rome*

Amnuaykanjanasin, A., and Daub, M. E. (2009) The ABC transporter ATR1 is necessary for efflux of the toxin cercosporin in the fungus *Cercospora nicotianae*. *Fungal Genetics and Biology* **46(2)**:146-158

Amselem, J., Cuomo, C. A., van Kan, J. A., Viaud, M., Benito, E. P., Couloux, A., Coutinho, P. M., de Vries, R. P., Dyer, P. S., Fillinger, S., Fournier, E., Gout, L., Hahn, M., Kohn, L., Lapalu, N., Plummer, K. M., Pradier, J. M., Quevillon, E., Sharon, A., Simon, A., ten Have, A., Tudzynski, B., Tudzynski, P., Wincker, P., Andrew, M., Anthouard, V., Beever, R. E., Beffa, R., Benoit, I., Bouzid, O., Brault, B., Chen, Z., Choquer, M., Collemare, J., Cotton, P., Danchin, E. G., Da Silva, C., Gautier, A., Giraud, C., Giraud, T., Gonzalez, C., Grossetete, S., Guldener, U., Henrissat, B., Howlett, B. J., Kodira, C., Kretschmer, M., Lappartient, A., Leroch, M., Levis, C., Mauceli, E., Neuveglise, C., Oeser, B., Pearson, M., Poulain, J., Poussereau, N., Quesneville, H., Rascle, C., Schumacher, J., Segurens, B., Sexton, A., Silva, E., Sirven, C., Soanes, D. M., Talbot, N. J., Templeton, M., Yandava, C., Yarden, O., Zeng, Q., Rollins, J. A., Lebrun, M. H., and Dickman, M. (2011) Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *PLoS Genetics* **7(8)**:e1002230

Bakhshi, M., Arzanlou, M., Babai-Ahari, A., Groenewald, J. Z., and Crous, P. W. (2018) Novel primers improve species delimitation in *Cercospora*. *IMA Fungus* **9**:299-332

Bakhshi, M., Arzanlou, M., Babai-Ahari, A., Groenewald, J. Z., Braun, U., and Crous, P. W. (2015) Application of the consolidated species concept to *Cercospora* spp. from Iran. *Persoonia* **34**:65-86

Beckman, P. M., and Payne, G. A. (1982) External growth, penetration, and development of *Cercospora-Zeae-Maydis* in corn leaves. *Phytopathology* **72(7)**:810-815

Beseli, A., Amnuaykanjanasin, A., Herrero, S., Thomas, E., and Daub, M. E. (2015) Membrane transporters in self resistance of *Cercospora nicotianae* to the photoactivated toxin cercosporin. *Current Genetics* **61(4)**:601-620

Bolton, M. D., van Esse, H. P., Vossen, J. H., de Jonge, R., Stergiopoulos, I., Stulemeijer, I. J., van den Berg, G. C., Borras-Hidalgo, O., Dekker, H. L., de Koster, C. G., de Wit, P. J., Joosten, M. H., and Thomma, B. P. (2008) The novel *Cladosporium fulvum* lysin motif effector Ecp6 is a virulence factor with orthologues in other fungal species. *Molecular Microbiology* **69(1)**:119-136

Breen, S., Williams, S. J., Outram, M., Kobe, B., and Solomon, P. S. (2017) Emerging insights into the functions of Pathogenesis-Related Protein 1. *Trends in Plant Science* **22(10)**:871-879

Callahan, T. M., Rose, M. S., Meade, M. J., Ehrenshaft, M., and Upchurch, R. G. (1999) CFP, the putative cercosporin transporter of *Cercospora kikuchii*, is required for wild type cercosporin production, resistance, and virulence on soybean. *Molecular Plant-Microbe Interactions* **12**:901–910

Chupp, C. (1953) A monograph of the fungus genus *Cercospora*, Ithaca, N.Y.

Crous, P., Groenewald, J., Groenewald, M., Caldwell, P., Braun, U., and Harrington, T. (2006) Species of *Cercospora* associated with grey leaf spot of maize. *Studies in Mycology* **55**:189-197

Crous, P. W., and Braun, U. (2003) *Mycosphaerella* and its anamorphs 1. Names published in *Cercospora* and *Passalora*. *CBS Biodiversity Series* **1**:1-571

Crous, P. W., and Groenewald, J. Z. (2005) Hosts, species and genotypes: opinions versus data. *Australasian Plant Pathology* **34**:463–470

Daub, M. E., and Ehrenshaft, M. (2000) The photoactivated *Cercospora* toxin cercosporin: contributions to plant disease and fundamental biology. *Annual Reviews in Phytopathology* **38**:461-490

De Jonge, R., Bolton, M. D., and Thomma, B. P. (2011) How filamentous pathogens co-opt plants: the ins and outs of fungal effectors. *Current Opinion in Plant Biology* **14(4)**:400-406

De Jonge, R., Ebert, M. K., Huitt-Roehl, C. R., Pal, P., Suttle, J. C., Spanner, R. E., Neubauer, J. D., Jurick, W. M., Stott, K. A., Secor, G. A., Thomma, B. P. H. J., Van de Peer, Y., Townsend, C. A., and Bolton, M. D. (2018) Gene cluster conservation provides insight into cercosporin biosynthesis and extends production to the genus *Colletotrichum*. *Proceedings of the National Academy of Sciences of the United States of America* **115(24)**:E5459-E5466

DeZwaan, T. M., Carroll, A. M., Valent, B., and Sweigard, J. A. (1999) *Magnaporthe grisea* pth11p is a novel plasma membrane protein that mediates appressorium differentiation in response to inductive substrate cues. *Plant Cell* **11(10)**:2013-30

Dhami, N. B., Kim, S. K., Paudel, A., Shrestha, J., and Rijal, T. R. (2015) A review on threat of grey leaf spot disease of maize in Asia. *Journal of Maize Research and Development* **1(1)**:71-85

Doehlemann, G., Okmen, B., Zhu, W., and Sharon, A. (2017) Plant pathogenic fungi. *Microbiology Spectrum* **5(1)**:FUNK-0023-2016

Dolfini, D., Zambelli, F., Pavesi, G., and Mantovani, R. (2009) A perspective of promoter architecture from the CCAAT box. *Cell Cycle* **8(24)**:4127-4137

Dufresne, M., Perfect, S., Pellier, A. L., Bailey, J. A., and Langin, T. (2000) A GAL4-like protein is involved in the switch between biotrophic and necrotrophic phases of the infection process of *Colletotrichum lindemuthianum* on common bean. *Plant Cell* **12(9)**:1579-1590

Dunkle, L. D., and Levy, M. (2000) Genetic relatedness of African and United States populations of *Cercospora zeae-maydis*. *Phytopathology* **90(5)**:486-90

Ebersberger, I., de Matos Simoes, R., Kupczok, A., Gube, M., Kothe, E., Voigt, K., and von Haeseler, A. (2012) A consistent phylogenetic backbone for the fungi. *Molecular Biology and Evolution* **29(5)**:1319-1334

Fajola, A. O. (1978) Cercosporin, a phytotoxin from *Cercospora* spp. *Physiological Plant Pathology* **13**:157-164

FAO (1995) Staple foods: What do people eat? www.fao.org/3/u8480e/U8480E07.htm.

Faris, J. D., Zhang, Z., Lu, H., Lu, S., Reddy, L., Cloutier, S., Fellers, J. P., Meinhardt, S. W., Rasmussen, J. B., Xu, S. S., Oliver, R. P., Simons, K. J., and Friesen, T. L. (2010) A unique wheat disease resistance-like gene governs effector-triggered susceptibility to necrotrophic pathogens. *Proceedings of the National Academy of Sciences of the United States of America* **107(30)**:13544-13549

Fernandez, J., and Orth, K. (2018) Rise of a cereal killer: the biology of *Magnaporthe oryzae* biotrophic growth. *Trends in Microbiology* **26(7)**:582-597

Friesen, T. L., Faris, J. D., Solomon, P. S., and Oliver, R. P. (2008) Host-specific toxins: effectors of necrotrophic pathogenicity. *Cellular Microbiology* **10(7)**:1421-1428

Friesen, T. L., Stukenbrock, E. H., Liu, Z., Meinhardt, S., Ling, H., Faris, J. D., Rasmussen, J. B., Solomon, P. S., McDonald, B. A., and Oliver, R. P. (2006) Emergence of a new disease as a result of interspecific virulence gene transfer. *Nature Genetics* **38(8)**:953-956

Gabriela Roca, M., Read, N. D., and Wheals, A. E. (2005) Conidial anastomosis tubes in filamentous fungi. *FEMS Microbiology Letters* **249(2)**:191-198

Goodwin, S. B., Dunkle, L. D., and Zismann, V. L. (2001) Phylogenetic analysis of *Cercospora* and *Mycosphaerella* based on the internal transcribed spacer region of ribosomal DNA. *Phytopathology* **91(7)**:648-658

Greenberg, J. T., and Yao, N. (2004) The role and regulation of programmed cell death in plant-pathogen interactions. *Cellular Microbiology* **6(3)**:201-211

Groenewald, J. Z., Nakashima, C., Nishikawa, J., Shin, H. D., Park, J. H., Jama, A. N., Groenewald, M., Braun, U., and Crous, P. W. (2013) Species concepts in *Cercospora*: spotting the weeds among the roses. *Studies in Mycology* **75(1)**:115-170

Groenewald, M., Groenewald, J. Z., and Crous, P. W. (2005) Distinct species exist within the *Cercospora apii* morphotype. *Phytopathology* **95(8)**:951-959

Groenewald, M., Groenewald, J. Z., Braun, U., and Crous, P. W. (2006) Host range of *Cercospora apii* and *C. beticola* and description of *C. apiicola*, a novel species from celery. *Mycologia* **98(2)**:275–285

Groenewald, M., Groenewald, J. Z., Harrington, T. C., Abeln, E. C., and Crous, P. W. (2006) Mating type gene analysis in apparently asexual *Cercospora* species is suggestive of cryptic sex. *Fungal Genetics and Biology* **43(12)**:813-25

Gust, A. A., Willmann, R., Desaki, Y., Grabherr, H. M., and Nurnberger, T. (2012) Plant LysM proteins: modules mediating symbiosis and immunity. *Trends in Plant Science* **17(8)**:495-502

Hahn, M., Neef, U., Struck, C., Gottfert, M., and Mendgen, K. (1997) A putative amino acid transporter is specifically expressed in haustoria of the rust fungus *Uromyces fabae*. *Molecular Plant-Microbe Interactions* **10:**438-445

Hamer, J. E., Howard, R. J., Chumley, F. G., and Valent, B. (1988) A mechanism for surface attachment in spores of a plant pathogenic fungus. *Science* **239(4837)**:288-290

Hammond-Kosack, K. E., and Rudd, J. J. (2008) Plant resistance signalling hijacked by a necrotrophic fungal pathogen. *Plant Signaling & Behavior* **3(11)**:993-995

Heitman, J., Sun, S., and James, T. Y. (2013) Evolution of fungal sexual reproduction. *Mycologia* **105(1)**:1-27

Hibbett, D. S., Binder, M., Bischoff, J. F., Blackwell, M., Cannon, P. F., Eriksson, O. E., Huhndorf, S., James, T., Kirk, P. M., Lucking, R., Thorsten Lumbsch, H., Lutzoni, F., Matheny, P. B., McLaughlin, D. J., Powell, M. J., Redhead, S., Schoch, C. L., Spatafora, J. W., Stalpers, J. A., Vilgalys, R., Aime, M. C., Aptroot, A., Bauer, R., Begerow, D., Benny, G. L., Castlebury, L. A., Crous, P. W., Dai, Y. C., Gams, W., Geiser, D. M., Griffith, G. W., Gueidan, C., Hawksworth, D. L., Hestmark, G., Hosaka, K., Humber, R. A., Hyde, K. D., Ironside, J. E., Koljalg, U., Kurtzman, C. P., Larsson, K. H., Lichtwardt, R., Longcore, J., Miadlikowska, J., Miller, A., Moncalvo, J. M., Mozley-Standridge, S., Oberwinkler, F., Parmasto, E., Reeb, V., Rogers, J. D., Roux, C., Ryvarden, L., Sampaio, J. P., Schussler, A., Sugiyama, J., Thorn, R. G., Tibell, L., Untereiner, W. A., Walker, C., Wang, Z., Weir, A., Weiss, M., White, M. M., Winka, K., Yao, Y. J., and Zhang, N. (2007) A higher-level phylogenetic classification of the Fungi. *Mycology Research* **111(Pt 5)**:509-547

JGI (2011) Home - *Cercospora zeae-maydis* v1.0. https://genome.jgi.doe.gov/Cerzm1/Cerzm1.home.html.

Jones, J. D., and Dangl, J. L. (2006) The plant immune system. *Nature* **444(7117)**:323-329

Kachroo, P., Lim, G., and Kachroo, A. (2016) Nitric oxide-mediated chemical signaling during systemic acquired resistance. Edited by D. Wendehenne. Vol. 77, *Advances in Botanical Research*. Elsevier Science, Cambridge, MA, USA.

Koeck, M., Hardham, A. R., and Dodds, P. N. (2011) The role of effectors of biotrophic and hemibiotrophic fungi in infection. *Cellular Microbiology* **13(12)**:1849-1857

Kombrink, E., and Somssich, I. E. (1995) Defense responses of plants to pathogens. *Advances in Botanical Research, Vol 21* **21**:1-34

Kuyama, S., and Tamura, T. (1957) Cercosporin - a pigment of *Cercosporina kikuchii* Matsumoto Et Tomoyasu. 2. Physical and chemical properties of cercosporin and its derivatives. *Journal of the American Chemical Society* **79(21)**:5726-5729

Lapaire, C. L., and Dunkle, L. D. (2003) Microcycle conidiation in *Cercospora zeae-maydis*. *Phytopathology* **93(2)**:193-199

Latterell, F. M., and Rossi, A. E. (1983) Grey leaf spot of corn: a disease on the move. *Plant Disease* **67(8)**:842-847

Lehmann, S., Serrano, M., L'Haridon, F., Tjamos, S. E., and Metraux, J. P. (2015) Reactive oxygen species and plant resistance to fungal pathogens. *Phytochemistry* **112**:54-62

Li, B., Meng, X., Shan, L., and He, P. (2016) Transcriptional regulation of pattern-triggered immunity in plants. *Cell Host & Microbe* **19(5)**:641-650

Li, C., Barker, S. J., Gilchrist, D. G., Lincoln, J. E., and Cowling, W. A. (2008) *Leptosphaeria maculans* elicits apoptosis coincident with leaf lesion formation and hyphal advance in Brassica napus. *Molecular Plant-Microbe Interactions* **21(9)**:1143-1153

Liu, K. J., and Xu, X. D. (2013) First report of grey leaf spot of maize caused by Cercospora *zeina* in China. *Plant Disease* **97(12)**:1656-1656

Lyu, X., Shen, C., Fu, Y., Xie, J., Jiang, D., Li, G., and Cheng, J. (2016) A small secreted virulence-related protein is essential for the necrotrophic interactions of *Sclerotinia sclerotiorum* with its host plants. *PLoS Pathogens* **12(2)**:e1005435

Mackie, A. J., Roberts, A. M., Callow, J. A., and Green, J. R. (1993) Glycoproteins recognised by monoclonal antibodies UB7, UB8 and UB10 are expressed early in the development of pea powdery mildew haustoria. *Physiological and Molecular Plant Pathology* **43**:135-146

Maher, E. A., Bate, N. J., Ni, W., Elkind, Y., Dixon, R. A., and Lamb, C. J. (1994) Increased disease susceptibility of transgenic tobacco plants with suppressed levels of preformed phenylpropanoid products. *Proceedings of the National Academy of Sciences of the United States of America* **91(16)**:7802-7806

Manners, J. M., and Gay, J. L. (1983) The host–parasite interface and nutrient transfer in biotrophic parasitism. Edited by J. A. Callow, *Biochemical Plant Pathology*. Wiley & Sons Ltd, Chichester, UK.

Mastrangelopoulou, M., Grigalavicius, M., Berg, K., Menard, M., and Theodossiou, T. A. (2018) Cytotoxic and Photocytotoxic Effects of Cercosporin on Human Tumor Cell Lines. *Journal of Photochemistry and Photobiology* **95(1)**:387-396

Meisel, B., Korsman, J., Kloppers, F., and Berger, D. (2009) *Cercospora zeina* is the causal agent of grey leaf spot disease of maize in southern Africa. *European Journal of Plant Pathology* **124**:577-583

Mendgen, K., and Hahn, M. (2002) Plant infection and the establishment of fungal biotrophy. *Trends in Plant Science* **7(8)**:352-356

Mendgen, K., Hahn, M., and Deising, H. (1996) Morphogenesis and mechanisms of penetration by plant pathogenic fungi. *Annual Reviews in Phytopathology* **34**:367-386

Muller, M. F., Barnes, I., Kunene, N. T., Crampton, B. G., Bluhm, B. H., Phillips, S. M., Olivier, N. A., and Berger, D. K. (2016) *Cercospora zeina* from maize in South Africa exhibits high genetic diversity and lack of regional population differentiation. *Phytopathology* **106(10)**:1194-1205

Munkvold, G. P., Martinson, C. A., Shriver, J. M., and Dixon, P. M. (2001) Probabilities for profitable fungicide use against grey leaf spot in hybrid maize. *Phytopathology* **91(5)**:477-484

Neves, D. L., Silva, C. N., Pereira, C. B., Campos, H. D., and Tessmann, D. J. (2015) *Cercospora zeina* is the main species causing grey leaf spot in southern and central Brazilian maize regions. *Tropical Plant Pathology* **40(6)**:368-374

Newman, A. G., and Townsend, C. A. (2016) Molecular characterization of the cercosporin biosynthetic pathway in the fungal plant pathogen *Cercospora nicotianae*. *Journal of the American Chemical Society* **138(12)**:4219-4228

Nicholson, R. L., Kunoh, H., Shiraishi, T., and Yamada, T. (1993) Initiation of the infection process by *Erysiphe graminis* - conversion of the conidial surface from hydrophobicity to hydrophilicity and influence of the conidial exudate on the hydrophobicity of the barley leaf surface. *Physiological and Molecular Plant Pathology* **43(4)**:307-318

Nsibo, D. L., Barnes, I., Kunene, N. T., and Berger, D. K. (2019) Influence of farming practices on the population genetics of the maize pathogen *Cercospora zeina* in South Africa. *Fungal Genetics and Biology* **125**:36-44

Nuss, E. T., and Tanumihardjo, S. A. (2010) Maize: A paramount staple crop in the context of global nutrition. *Comprehensive Reviews in Food Science and Food Safety* **9(4)**:417-436

OECD/FAO. (2016) Agriculture in Sub-Saharan Africa: Prospects and challenges for the next decade. *OECD-FAO Agricultural Outlook 2016-2025, OECD Publishing, Paris.*

Ohm, R. A., Feau, N., Henrissat, B., Schoch, C. L., Horwitz, B. A., Barry, K. W., Condon, B. J., Copeland, A. C., Dhillon, B., Glaser, F., Hesse, C. N., Kosti, I., LaButti, K., Lindquist, E. A., Lucas, S., Salamov, A. A., Bradshaw, R. E., Ciuffetti, L., Hamelin, R. C., Kema, G. H., Lawrence, C., Scott, J. A., Spatafora, J. W., Turgeon, B. G., de Wit, P. J., Zhong, S., Goodwin, S. B., and Grigoriev, I. V. (2012) Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PLoS Pathogens* **8(12)**:e1003037

Okori, P., Fahleson, J., Rubaihayo, P. R., Adipala, E., and Dixelius, C. (2003) Assessment of genetic variation among East African *Cercospora zeae-maydis*. *African Crop Science Journal* **11(2)**:75-86

Oliver, R. P., and Ipcho, S. V. (2004) *Arabidopsis* pathology breathes new life into the necrotrophs-vs.-biotrophs classification of fungal pathogens. *Molecular Plant Pathology* **5(4)**:347-352

Oliver, R. P., and Solomon, P. S. (2010) New developments in pathogenicity and virulence of necrotrophs. *Current Opinions in Plant Biology* **13**:415-419

Panagiotis, M., Kritonas, K., Irini, N. O., Kiriaki, C., Nicolaos, P., and Athanasios, T. (2007) Expression of the yeast cpd1 gene in tobacco confers resistance to the fungal toxin cercosporin. *Biomolecular Engineering* **24(2)**:245-251

Paul, P. A., and Munkvold, G. P. (2005) Influence of temperature and relative humidity on sporulation of *Cercospora zeae-maydis* and expansion of grey leaf spot lesions on maize leaves. *Plant Disease* **89(6)**:624-630

Perfect, S. E., and Green, J. R. (2001) Infection structures of biotrophic and hemibiotrophic fungal plant pathogens. *Molecular Plant Pathology* **2(2)**:101-108

Perfect, S. E., Pixton, K. L., O'Connell, R. J., and Green, J. R. (2000) The distribution and expression of a biotrophy-related gene, CIH1, within the genus *Colletotrichum*. *Molecular Plant Pathology* **1(4)**:213-221

Petit, A. N., Fontaine, F., Vatsa, P., Clement, C., and Vaillant-Gaveau, N. (2012) Fungicide impacts on photosynthesis in crop plants. *Photosynthesis Research* **111(3)**:315-326

Pollack, F. G. (1987) An annotated compilation of *Cercospora* Names. Vol. 12, *Mycologia memoirs*. Schweizerbart Science Publishers, Stuttgart, Germany.

Qu, J., Zou, X., Yu, J., and Zhou, Y. (2017) The conidial mucilage, natural film coatings, is involved in environmental adaptability and pathogenicity of *Hirsutella satumaensis* Aoki. *Scientific Reports* **7(1)**:1301

Rees, J. M., and Jackson, T. A. (2008) Grey Leaf Spot of Corn. *University of Nebraska-Extension, Institute of Agriculture and Natural Resources*

Sabotič, J., and Kos, J. (2017) Fungal Protease Inhibitors. Edited by J. Mérillon, and K. G. Ramawat, *Fungal Metabolites*. Springer Nature, Switzerland.

Sanz-Martin, J. M., Pacheco-Arjona, J. R., Bello-Rico, V., Vargas, W. A., Monod, M., Diaz-Minguez, J. M., Thon, M. R., and Sukno, S. A. (2016) A highly conserved metalloprotease effector enhances virulence in the maize anthracnose fungus *Colletotrichum graminicola*. *Molecular Plant Pathology* **17(7)**:1048-1062

Schoch, C. L., Crous, P. W., Groenewald, J. Z., Boehm, E. W., Burgess, T. I., de Gruyter, J., de Hoog, G. S., Dixon, L. J., Grube, M., Gueidan, C., Harada, Y., Hatakeyama, S., Hirayama, K., Hosoya, T., Huhndorf, S. M., Hyde, K. D., Jones, E. B., Kohlmeyer, J., Kruys, A., Li, Y. M., Lucking, R., Lumbsch, H. T., Marvanova, L., Mbatchou, J. S., McVay, A. H., Miller, A. N., Mugambi, G. K., Muggia, L., Nelsen, M. P., Nelson, P., Owensby, C. A., Phillips, A. J., Phongpaichit, S., Pointing, S. B., Pujade-Renaud, V., Raja, H. A., Plata, E. R., Robbertse, B., Ruibal, C., Sakayaroj, J., Sano, T., Selbmann, L., Shearer, C. A., Shirouzu, T., Slippers, B., Suetrong, S., Tanaka, K., Volkmann-Kohlmeyer, B., Wingfield, M. J., Wood, A. R., Woudenberg, J. H., Yonezawa, H., Zhang, Y., and Spatafora, J. W. (2009) A class-wide phylogenetic assessment of Dothideomycetes. *Studies in Mycology* **64**:1-15S10

Smith, D. (2013) A3878 Fungicide resistance management in corn, soybean, and wheat in Wisconsin *University of Wisconsin-Madison Extension*

Stergiopoulos, I., and de Wit, P. J. (2009) Fungal effector proteins. *Annu Rev Phytopathol* **47**:233-263

Subramoni, S., Suárez-Moreno, Z. R., and Venturi, V. 2010. Lipases as pathogenicity factors of plant pathogens. Pages 3269-3277. in: Handbook of Hydrocarbon and Lipid Microbiology K. N. Timmis, ed. Springer, Berlin, Heidelberg.

Swart, V., Crampton, B. G., Ridenour, J. B., Bluhm, B. H., Olivier, N. A., Meyer, J. J. M., and Berger, D. K. (2017) Complementation of CTB7 in the maize pathogen *Cercospora zeina* overcomes the lack of *in vitro* cercosporin production. *Molecular Plant-Microbe Interactions* **30(9)**:710-724

Tehon, L. R., and Daniels, E. Y. (1925) Notes on the parasitic fungi of Illinois. II. *Mycologia* **17(6)**:240-249

Tanaka, K., Nguyen, C. T., Liang, Y., Cao, Y., and Stacey, G. (2013) Role of LysM receptors in chitin-triggered plant innate immunity. *Plant Signaling & Behavior* **8(1)**:e22598

Tanaka, S., Han, X., and Kahmann, R. (2015) Microbial effectors target multiple steps in the salicylic acid production and signaling pathway. *Frontiers in Plant Science* **6**:349

Taylor, J., Jacobson, D., and Fisher, M. (1999) The evolution of asexual fungi: Reproduction, speciation and classification. *Annual Reviews in Phytopathology* **37**:197-246

Thomma, B. P., HP, V. A. N. E., Crous, P. W., and PJ, D. E. W. (2005) *Cladosporium fulvum* (syn. *Passalora fulva*), a highly specialized plant pathogen as a model for functional studies on plant pathogenic *Mycosphaerellaceae*. *Molecular Plant Pathology* **6(4)**:379-393

Torres, M. A., Jones, J. D., and Dangl, J. L. (2006) Reactive oxygen species signaling in response to pathogens. *Plant Physiology* **141(2)**:373-378

Tsuge, T., Harimoto, Y., Akimitsu, K., Ohtani, K., Kodama, M., Akagi, Y., Egusa, M., Yamamoto, M., and Otani, H. (2013) Host-selective toxins produced by the plant pathogenic fungus *Alternaria alternata*. *FEMS Microbiology Reviews* **37(1)**:44-66

Tucker, S. L., and Talbot, N. J. (2001) Surface attachment and pre-penetration stage development by plant pathogenic fungi. *Annual Reviews in Phytopathology* **39**:385-417

Wallen, R. M., and Perlin, M. H. (2018) An overview of the function and maintenance of sexual reproduction in dikaryotic fungi. *Frontiers in Microbiology* **9**:503

Wang, J., Levy, M., and Dunkle, L. D. (1998) Sibling species of *Cercospora* associated with grey leaf spot of maize. *Phytopathology* **88(12)**:1269-1275

Ward, J. M. J., Stromberg, E. L., Nowell, D. C., and Nutter, F. W. (1999) Grey leaf spot: A disease of global importance in maize production. *Plant Disease* **83**:884-895

Wei, Y., Shen, W., Dauk, M., Wang, F., Selvaraj, G., and Zou, J. (2004) Targeted gene disruption of glycerol-3-phosphate dehydrogenase in *Colletotrichum gloeosporioides* reveals evidence that glycerol is a significant transferred nutrient from host plant to fungal pathogen. *Journal of Biological Chemistry* **279(1)**:429-435

Weiberg, A., Wang, M., Lin, F. M., Zhao, H., Zhang, Z., Kaloshian, I., Huang, H. D., and Jin, H. (2013) Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. *Science* **342(6154)**:118-123

Weiland, J., and Koch, G. (2004) Sugarbeet leaf spot disease (*Cercospora beticola* Sacc.). *Molecular Plant Pathology* **5(3)**:157-166

Wilson, R. A., and Talbot, N. J. (2009) Under pressure: investigating the biology of plant infection by *Magnaporthe oryzae*. *Nature Reviews Microbiology* **7(3)**:185-195

Yike, I. (2011) Fungal proteases and their pathophysiological effects. *Mycopathologia* **171(5)**:299-323

Zhao, Z., Liu, H., Wang, C., and Xu, J. R. (2013) Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. *BMC Genomics* **14**:274

Zhu, W., Wei, W., Fu, Y., Cheng, J., Xie, J., Li, G., Yi, X., Kang, Z., Dickman, M. B., and Jiang, D. (2013) A secretory protein of necrotrophic fungus *Sclerotinia sclerotiorum* that suppresses host resistance. *PLoS ONE* **8(1)**:e53901

Zipfel, C. (2009) Early molecular events in PAMP-triggered immunity. *Current Opinions in Plant Biology* **12(4)**:414-420

Chapter 2

Genome assembly of the maize fungal pathogen, *Cercospora zeina*

N. A. Olivier, Y-C. Lin, F. van Staden, B. G. Crampton, V. Swart, F. Joubert, Y. Van de Peer, O. Reva and D. K. Berger

Department of Plant and Soil Sciences, Faculty of Natural and Soil Sciences, Forestry and Agricultural Research Institute, University of Pretoria, Pretoria, 0002, South Africa

Data from this chapter, together with Chapter 3 was published as a genome announcement in *IMA Fungus*. Analyses performed on SSR content of the genome assembly were published in *Phytopathology*. The second and third genome assemblies, the transcriptome assembly, the quality controls, mapping and phylogenetic analysis were performed by N. Olivier. Dr Y-C. Lin from the laboratory of Prof Y. Van de Peer provided invaluable technical and procedural assistance for all aspects of the genome and transcriptome assembly and analyses. The first genome assembly was performed by a student, F. van Staden, with the aid of Prof F. Joubert. The *C. zeina* culturing and genomic DNA isolation were performed by Dr B. G. Crampton. The *in vitro C. zeina* culturing and RNA isolations were performed by Ms V. Swart. Prof O. Reva provided assistance with critical revision of the manuscript. Prof D. K. Berger was responsible for planning, funding and co-ordinating the sequencing project. He provided direction in interpreting the results and critical revision of the manuscript. He also arranged funding for a research visit to the laboratory of Prof. Y. Van de Peer where a significant amount of the research for the chapter was completed.

2.1 Abstract

Maize (*Zea mays*) is an important staple food crop in many part of the world, especially Sub-Saharan Africa. Grey Leaf Spot (GLS) is a devastating foliar disease in maize plants, leading to severe yield losses, and this presents a threat to food security in many impoverished nations dependent on maize. Two causal agents for GLS have been identified, with *C. zeae-maydis* the dominant pathogen in the USA and *C. zeina* displaying a reduced distribution. To date, *C. zeina* has been shown to be the only causal agent for GLS in Africa. Due to the paucity of information regarding the mechanism of pathogenicity of *C. zeina*, as well as the absence of gene information for functional genomics studies, the genome was sequenced at the Purdue Genomics Core Facility from three different insert-size sequencing libraries. The sequence data was used to construct three separate genome assemblies. The first assembly displayed an abnormal GC-content profile and an insufficient number of transcript sequences mapped for it to be functional, while the second assembly was too fragmented and contained too many no-call nucleotides to be useful. The third assembly contained the expected GC-content and had an N50 of 160,632 bp with an assembly length of 40,773,084 bp. The majority of the transcript sequences mapped to this assembly, rendering it useful for functional genomics. Finally, a phylogenetic analysis of two genes in the assembly, together with those of other *Cercospora* species, indicated that the genome assembly grouped with other C. *zeina* strains as expected. The completed assembly was uploaded to the NCBI WGS database. The assembly was subsequently used for gene prediction and annotation in Chapter 3.

2.2. Introduction

Functional genomics studies utilize genome-wide gene and protein information to investigate processes of interest in a global context. This involves functional, transcriptional and regulation information usually obtained from high-throughput methods. These methods include microarrays and next-generation sequencing (https://www.nature.com/subjects/functional-genomics). To place processes in a genome-wide context it is optimal to have a functionally annotated genome for the organism under investigation. For economically important crop pathogens, such as *Cercospora zeina*, this is very important when infection and pathogenesis strategies are being studied, since these processes are multi-local and gene positions, sequences and global gene expression patterns are required to investigate the pathogenesis process (Zhao, X. *et al.*, 2007). In cases where an annotated genome is not available it is necessary to sequence the genome of interest and create a functional genome assembly which can be annotated for gene content.

In principle, genome assembly can be defined as the reconstruction of a genome sequence from a collection of randomly sampled sequences (Narzisi & Mishra, 2011). This has proven a complex problem since the first generation Sanger sequencing instruments produced multiple short reads (Zhang *et al.*, 2011). The shotgun-sequencing approach used by Celera Genomics during the sequencing of the *Drosophila* (Adams *et al.*, 2000 ; Myers *et al.*, 2000) and *Homo sapiens* (Venter *et al.*, 2001) genomes increased the speed of the sequencing process, but necessitated the assembly of shorter sequences into longer contiguous sequences (contigs), and scaffolding the contigs into long sequences approaching chromosome length. The Celera Genomics sequencing strategy relied on the creation of ~500bp long paired reads obtained from sequencing the ends of 2kbp, 10kbp and 50kbp shotgun clone-libraries. The Celera Assembler (Myers *et al.*, 2000) was developed to combine the data from the smaller insert sequencing libraries to create contigs, and use the insert distance between the read pairs from the longer sequencing libraries to scaffold the contigs.

The advent of second-generation short-read sequencing technologies dramatically increased the throughput and amount of data of sequencing projects, while sequencing complex genomes in a fraction of the time (Florea *et al.*, 2011). Assembly algorithms have to deal with the computational burden from large amounts of data, since this falls into the class of Non-deterministic Polynomial-time (NP) hard problems for which there are no efficient computational solutions. In addition, new platform-specific error-types also have to be taken into account (Pop, 2009).

The presence of repeat regions was the biggest challenge in creating long scaffolds, since the current technology of short reads do not span longer repetitive regions (Pop & Salzberg, 2008). The ambiguity surrounding the exact positional identification for the contig sequences surrounding repeat regions can only be resolved by sequencing the repeats plus the flanking regions. For very long repeats or tandem repeat regions, there are no practical methods for correct assembly into contigs. The repeat regions therefore split contigs and only by using longer insert mate-pair libraries can some positional information be inferred for adjacent contigs (Zhang *et al.*, 2011).

The non-uniform coverage of sequence reads across a genome is another challenge for assembly, since regions with low coverage will not be well-represented in read data. This also introduces gaps between contigs, and since coverage-based statistics for genome assembly diagnostics usually model a uniform coverage, these regions would invalidate the inherent statistical assumptions and undermine the assembly evaluation (Miller *et al.*, 2010).

To generate a genome assembly from a large number of short-read sequences two main approaches can be followed. Firstly, in cases where a high quality reference genome or genome of a closely related species is available, the reads can be mapped to this reference genome to generate an assembly based on the information in the reference genome. Secondly, the reads can be assembled into a genome without the use of a reference genome, also known as the *de novo* approach.

2.2.1 Reference mapping genome assembly

The reference mapping approach is faster and less computationally intensive, and in cases where the species are closely related could yield a very accurate genome. Some popular aligners include Bowtie2 (Langdon, 2015), BWA (Li, H. & Durbin, 2009), SOAP2 (Li, H. & Durbin, 2009), and BFAST (Homer *et al.*, 2009). These aligners differ in the mapping and algorithm approaches, their computational requirements and the level to which they can accommodate mismatches and gaps or insertions/deletions (indels) in sequences. Two popular algorithms include the Burrows-Wheeler transformation (BWA, Bowtie2, SOAP2), and hashing algorithms where the reference genome is converted to a hash table (BFAST).

2.2.1.1 Burrows-Wheeler transformation

Burrows-Wheeler (BW) transformations utilize suffix arrays where suffixes are defined as substrings of sequences that start at any position in the sequence and end at the same position as the end of the sequence. To make the approach functional, circular permutations are created with a spacer added to the end of the suffix followed by the start of the genome. In this way the array is populated and the array dimensions are equal to the length of the genome sequence. The BW transform is then merely the last column of the circular suffix array, thus a single letter for the last nucleotide. By alternate concatenation of and lexicographic sorting of the BW transforms for an array, the sequence of the genome can be reconstructed.

Searching the array to confirm the presence and position of a specific read (length n) involves traversing the array. First the final base (n) in the read is identified in the final BW transform column of the array. Once these are identified (here it should find ~25% of all BW transforms), the subset in the array is sorted and the n-1 base in the read is searched for in the adjoining column. If identified, a new subset is created and sorted, followed by identifying the n-2 base in the adjoining BW transform column. By repeating the process, the presence and position of the read in the reference genome can be confirmed in this way. The algorithm is very sensitive for mismatches, and each algorithm incorporates different methods to deal with repeat regions and indels. Bowtie2, for example, can only accommodate 1-2 errors per 50bp (Schbath *et al.*, 2012).

2.2.1.2 Hashing algorithms

Hashing algorithms create numerical hash values for each sequence region in the genome. These values are then compiled into hash tables which can easily be searched. Hash values are then similarly computed for read sequences and these are then searched against the genome hash table. A seed and extend step attempts to confirm the alignment of locations in the read adjacent to the seed sequence to evaluate if the read is in the correct position. This approach potentially increases the speed and efficiency in computational memory allocation. Mismatches in the read sequence cause the hash values to be split into smaller seeds, which becomes computationally problematic. Since seeds of less than 10 nucleotides are rarely used due to memory constraints, aligners that use this approach have difficulty in mapping reads with a large number of mismatches. (Schbath *et al.*, 2012).

Both these approaches have problems when reads partially or completely map to repeat regions. In some cases the reads are removed from consideration, while in other cases the reads are randomly mapped to one or multiple regions (Trapnell & Salzberg, 2010). Structural variation between the reference and reads could also influence mapping to certain regions. Inversions, transpositions, insertions, deletions and duplication events can lead to mis-alignment of reads, or cause reads to be removed from the assembly. In these cases, it might be prudent to compare *de novo* assemblies to the mapped assemblies to find additional information about these regions (Otto, 2015).

2.2.2 *De novo* genome assembly

Most *de novo* assembler assume that if two sequencing reads share an overlapping sub-sequence they are likely to have originated from the same region of the genome. Subsequent combinations of these similar reads create consensus sequences which can then be combined with more overlapping reads. The algorithms used by *de novo* assemblers can broadly be classified into two categories, i.e Greedy and Graph based (Pop, 2009 ; Narzisi & Mishra, 2011).

2.2.2.1 Greedy algorithms

Greedy algorithms were initially developed for shotgun sequence assemblies. The approach is to attempt the alignment and subsequent combination of read-pairs showing the highest scoring overlap. These reads are merged and reinserted into the read pool. The cycle is repeated until no more mergers are possible. The result is a pool of contigs with breaks between them representing repeat or low-coverage regions that cannot be resolved (Narzisi & Mishra, 2011). The approach optimizes local maximum overlap efficiencies which might not be globally applicable, and might lead to mis-assemblies of repeat regions (Pop, 2009). Well known assemblers based on this algorithm include CAP3 (Huang & Madan, 1999), TIGR (Sutton *et al.*, 1995) and SSAKE (Warren *et al.*, 2007).

2.2.2.2 Graph-based algorithms

Graph-based assemblers normally treat reads and overlaps as nodes and edges in a graph space. Relationships between nodes/edges can be traversed to form a path, which in genome sequencing relates to the genome sequence assembly. Different graph-based methods can be applied in different algorithms, e.g. simple paths where nodes are visited at most once, or Euler paths where edges are traversed only once. Depending on the algorithm requirements and definitions of nodes and edges, graph-based methods can be incorporated into assembly algorithms. Since non-repeat sequences will induce a single, unique path while repeat sequences form cycles, divergences and convergences, graph-based methods are especially useful in finding and cataloguing repeat regions. In addition, some graph-based methods don’t rely on pairwise sequence alignments to find overlaps and is therefore less computationally expensive. Two graph-based approaches dominate this field, i.e. Overlap Layout Consensus (OLC) and Sequencing by Hybridization (SBH) (Miller *et al.*, 2010).

The OLC approach splits the process into three parts to obtain a more globally efficient solution. Firstly a list of pair-wise read overlaps are created. A graph is created from these alignments with each read represented by a node, while the edges are constructed between overlapping read-pairs. Next the graph is analysed to identify segments of the genome for which there are unambiguous paths through the graph, thus traversing each node only once. Finally the genome is constructed by inferring the DNA sequence from the optimal path identified during the layout step. The reads overlapping each position are used to reach a consensus identity for that position based on read coverage and quality scores (Pop, 2009). Arachne (Batzoglou *et al.*, 2002) and Edena (Hernandez *et al.*, 2008) are examples of assemblers that use the OLC approach.

Conversely the SBH approach treats overlaps as nodes in a graph, and the reads involved in the overlap as edges. Reads are partitioned into a collection of k-mers and a de Bruijn graph constructed in which each edge is a k-mer from this collection. An Eulerian path is constructed in which each edge is contained only once. Theoretically the graph is linear in size and computationally relatively quick. Unfortunately there are several complications inhibiting the computation of the Eulerian path. Sequencing errors can create false-positive edges, while a larger k-mer size will greatly inflate the graph size. In addition, a de Bruijn graph may not have a unique Euler path, and finding an ideal Euler path within specified constraints becomes an NP-hard problem. Assemblers apply different transformation approaches to compute optimal Euler paths within computational constraints (Pop, 2009). *De novo* assemblers utilizing the SBH algorithm include Velvet (Zerbino & Birney, 2008), SOAPdenovo (Li, R., Zhu*, et al.*, 2010), Euler (Pevzner *et al.*, 2001) and ABySS (Simpson *et al.*, 2009).

2.2.3 Hybrid algorithms

To incorporate the advantages of each of the assembly algorithms, some hybrid algorithms have been developed. Typical approaches involve separately performing assemblies with reference mapping and *de novo* methods, and then combining the resultant contigs according to some specified criteria. For example, the CD-Hybrid strategy (Ji *et al.*, 2011) uses three *de novo* assembly algorithms, i.e. Velvet, ABySS and SOAPdenovo to create separate genome assemblies. The *de novo* contigs are combined and a contig is selected only if identical to contigs in at least two of the assemblies. In parallel, short read data is assembled to multiple reference genomes by the AMOScmp algorithm. The reference assembly contigs are compared to the *de novo* contigs, and following criteria limiting the numbers of indels and setting a lower limit of the depth of coverage, the contigs that conform are combined with the *de novo* contigs and assembled using Minimo (Treangen *et al.*, 2011).

2.2.4 RNA sequencing read mapping

Mapping RNA sequencing (RNAseq) reads to a reference genome can be considered a special case of genome mapping, as the data also consists of multiple short reads, but usually just from short-insert fragment libraries. Since introns sequences in the reference genome are absent from the sequencing reads, this presents problems to strict genome mapping algorithms as these regions present as errors. In addition, the presence of paralogous gene families, low-complexity sequences and the high sequence similarity between alternatively spliced isoforms of the same gene all contribute to a large proportion of reads mapping to multiple regions of the genome with equal efficiency (Mortazavi *et al.*, 2008 ; Li, B., Ruotti*, et al.*, 2010 ; Zhao, S., 2014). Since the library insert sizes are generally too small for read-pairs to span introns to resolve many of these problems, read mappers have to correct and/or account for reads mapping to multiple sites. The main strategies involve either discarding these reads (Marioni *et al.*, 2008), or allow a portion of these reads to map to genes in proportion to unique read coverage for these genes (Mortazavi *et al.*, 2008). Thirdly some algorithms allow for the truncation of reads ends, thereby yielding an incomplete alignment when an entire sequence cannot be mapped (Engstrom *et al.*, 2013).

The majority of mapping algorithms rely on a two-step process, i.e. an initial alignment of reads to a reference genome, with a subsequent analysis of the alignments to identify splice sites (Zhao, S., 2014). These splice site information is then used to correctly map the reads that were initially unmapped. These aligning algorithms are included in the GSNAP (Wu & Nacu, 2010), Tophat (Kim *et al.*, 2013), and STAR (Dobin *et al.*, 2013) read mappers. This approach does require that exons have a certain expression level, and might thus miss rare splicing events (Zhao, S., 2014). Another approach is to allow the majority of correctly mapped overlapping reads to decide the correct splice site, as in the Subread aligner (Liao *et al.*, 2013). Finally, reads can be mapped to a reference transcriptome, although this will only map to and identify known splice-boundaries, and a representative reference genome is not always available (Zhao, S., 2014).

2.2.5 Transcriptome assembly

The assembly of RNA sequences into a transcriptome is especially useful when a reference genome is not available or the genome of a closely related species is not suitable, or even when a transcriptome is a more functional and relevant research requirement. This is especially true when considering the difference in sequencing cost and analysis expertise required between transcriptome and genome sequencing and assembly.

In theory the transcription assembly problem appears to be similar to the genome assembly problem, although short-read assemblers such as Velvet and ABySS cannot be directly utilized for this application. One limitation involves the basic requirement of these assemblers for a consistent sequencing coverage across all regions of the transcriptome, while the differential gene expression behaviour between transcripts will cause local coverages to vary by orders of magnitude. Additionally, alternative splicing of transcripts will increase the computational burden to represent a true version of the transcriptome (Chang *et al.*, 2015).

Two general categories of transcriptome assembly algorithms have been developed which closely mirror the genome assembly approaches, i.e. reference-based and *de novo* assembly. For reference-based approaches such as Cufflinks (Trapnell *et al.*, 2010) and Scripture (Guttman *et al.*, 2010) the first step involves the alignment of the reads to a high quality reference genome using a splice-aware mapper such as Tophat or GSNAP. The read-overlap information is used to construct a graph which includes all splicing variants, which is then traversed to recover the transcriptome with full-length isoforms (Chang *et al.*, 2015).

When a reference genome is not available, or the reference genome is altered with respect to the transcriptome as in cancerous versus healthy patients, a *de novo* transcriptome must be assembled. Since some *de novo* assemblers rely on principles applicable to genome assembly, they might not be very accurate. Other assemblers, such as Oases (Schulz *et al.*, 2012) and SOAPdenovo-Trans (Xie *et al.*, 2014) were developed using genome assemblers as basis (here Velvet and SOAPdenovo respectively), while adjusting the assembly for transcriptomes using additional algorithms. Trinity (Grabherr *et al.*, 2011 ; Haas *et al.*, 2013) was the first transcriptome-specific assembler, and it functions by first extending reads into longer contigs. These contigs are used to construct a de Bruijn graph which is then used to derive the splice-isoform paths for the final assembly (Chang *et al.*, 2015).

Transcriptome assembly presents novel challenges to assembly algorithms, since only the most efficient assemblers (including Trinity and SOAPdenovo-Trans) can successfully and accurately assembly highly expressed genes. In genome assembly excessive sequencing is beneficial, while in transcriptome sequencing both insufficient and excessive sequencing can thus present a problem. This also relates to the amount of sequence data required to assemble a functional gene-set. In a study on *Arabidopsis*, a sequence library of 4.2Gbp represented a 96.2% total set of predicted genes, while doubling the sequence data only increased the predicted gene-set by 3.7%. To represent a functional dataset it is therefore not practically feasible to generate excessive sequencing data-sets (Gruenheit *et al.*, 2012 ; Honaas *et al.*, 2016).

An additional consideration when sequencing small, gene-dense genomes such as fungi, is overlapping of UTR regions of genes which is quite common. For assemblers, which do not take this into account, the result could be incorrect assembly of end-to-end fused transcripts into chimeric genes. The Trinity assembler takes this into account, and analyses the consistency of read pairings across transcript lengths to correct for chimeras (Grabherr *et al.*, 2011 ; Haas *et al.*, 2013).

2.2.6 Assembly quality evaluation

Due to the large number of assemblers and algorithms available, it is critical to have a relevant set of standard statistical parameters with which to objectively evaluate the quality of a genome assembly. A set of benchmarking approaches was attempted by the dnGASP (*de novo* Genome Assembly Project; http://cnag.bsc.es/), GAGE (Genome Assembly Gold-standard Evaluations; (Salzberg *et al.*, 2012)), and the Assemblathon (Earl *et al.*, 2011) by utilizing simulated and real datasets and evaluating the performance of assembly algorithms and pipelines. The Assemblathon 2 event aimed at using un-simulated sequence data without a known reference genome sequence to firstly evaluate the performance of 21 assembly algorithms, and secondly to assess a set of metrics and their suitability in evaluating the quality of genome assemblies. A set of 105 metrics were analysed, and the results show that few assemblers performed consistently well on a diverse set of metrics. Several of the metrics were not generally applicable, since they rely on the availability of fosmid sequences and optical mapping for improving scaffolding of contigs. To generate consistent evaluation statistics the assemblathon_stats.pl (github.com/ucdavis-bioinformatics/assemblathon2-analysis/blob/master/assemblathon_stats.pl) Perl script was developed for the

Assemblathon event. The script is freely available for download, implementation and analyses of genome and transcriptome assemblies for a number of common, but important quality metrics (Earl *et al.*, 2011;Bradnam *et al.*, 2013)(Table 2.1).

Table 2.1 Genome assembly quality metrics evaluated by the assemblathon.pl script (Bradnam *et al.*, 2013). *Metrics for contig quality, scaffolding quality, and scaffolding content are listed.*

Metrics evaluating scaffolding quality	**Metrics evaluating contig quality**
Number of scaffolds	Number of contigs
Total size of scaffolds	Total size of contigs
Longest scaffold	Longest contig
Shortest scaffold	Shortest contig
Number of scaffolds > 1K nt	Number of contigs > 1K nt
Number of scaffolds > 10K nt	Number of contigs > 10K nt
Number of scaffolds > 100K nt	Number of contigs > 100K nt
Number of scaffolds > 1M nt	Number of contigs > 1M nt
Number of scaffolds > 10M nt	Number of contigs > 10M nt
Mean scaffold size	Mean contig size
Median scaffold size	Median contig size
N50 scaffold length	N50 contig length
L50 scaffold count	L50 contig count
scaffold %A	contig %A
scaffold %C	contig %C
scaffold %G	contig %G
scaffold %T	contig %T
scaffold %N	contig %N
scaffold %non-ACGTN	contig %non-ACGTN
Number of scaffold non-ACGTN nt	Number of contig non-ACGTN nt
Metrics evaluating scaffolding content	
Percentage of assembly in scaffolded contigs	
Percentage of assembly in unscaffolded contigs	
Average number of contigs per scaffold	
Average length of break (>25 Ns) between contigs in scaffold	
Number of contigs in scaffolds	
Number of contigs not in scaffolds	

The consensus for quality assessment metrics are driven by a large number of long contigs/scaffolds, a high completion rate of the expected genome size and a smaller number of contigs when the assembly size approaches the expected genome size. The metric used most often is the N50 value, which is calculated by summing all contig/scaffold lengths, starting with the longest, and observing the length that takes the sum length past 50% of the total assembly length (Bradnam *et al.*, 2013). It is a measure of the completeness of the genome since a large N50 value would indicate that the majority of the bases are in very large contigs or scaffolds, and an ideal approach would be a N50 value approaching the average chromosome size for the organism, with a small number of contigs or scaffolds, again approaching the number of chromosomes. Assembly approaches which emphasize longer contig/scaffold size (thus a high N50 value) above all others might not be biologically accurate, since it is common to find misassembled large contigs/scaffolds which will result in artificially large N50 values. The L50 value is related to the N50 value as it is defined as the smallest number of contigs for which the sum of their lengths produce the N50 value, with a smaller number being ideal.

Due to the occurrence of mis-assembly and false concatenation of contigs and scaffolds, the inclusion of a functional or gene-driven approach to quality evaluation is essential. Two datasets are available to analyse genome assemblies for genes occurring in all species of an order, also known as the core genes. The Core Eukaryotic Genes Mapping Approach (CEGMA) is a pipeline for identifying a subset of 248 conserved core eukaryotic genes in a genome assembly. The results are provided for full length, partial and missing core genes in the assembled genome (Parra *et al.*, 2007). The Benchmarking Universal Single-Copy Orthologs (BUSCO) completeness assessments utilise lineage-specific single-copy orthologous gene-sets in either the gene-set or protein-set for genome or transcriptome assemblies. Results are presented as percentages of single-copy complete, duplicated complete, fragmented and missing genes (Simao *et al.*, 2015). Although these approaches will give a percentage completion of the assembly based on these selected genes, it is advisable to include transcriptome information mapped to the assembly to evaluate the functional relevance of the genome or transcriptome.

In this study we sequenced the genomic DNA of *a C. zeina* strain, and created a functional genome assembly that is transcriptionally representative. The objectives were: i) to obtain useful sequencing data from the *C. zeina* isolate, ii) to create a genome assembly that was functional for subsequent gene prediction, iii) create a quality transcriptome assembly to increase the subsequent gene-prediction accuracy and iv) to show that the genome sequence was from a *C. zeina* strain, and not contamination. This study was the first to sequence the genome of the *C. zeina* species and to obtain quality genome and transcriptome assemblies that could be used for subsequent gene prediction and functional genomics studies.

2.3 Materials and Methods

2.3.1 Chemicals

All chemicals were purchased from Merck Chemicals (Germany), unless otherwise stated.

2.3.2 *Cercospora zeina* strain

The *Cercospora zeina* strain CMW 25467 was collected from infected maize (*Zea mays*) leaves in the Mkushi region of Zambia in March 2007 (Meisel *et al.*, 2009). Cultures of the strain have been deposited in The Forestry and Agricultural Biotechnology Institute (FABI) culture collection (CMW 25467), the Belgian Coordinated Collections of Micro-organisms/Mycothèque de l'Université catholique de Louvain (MUCL 51677), the Centraalbureau voor Schimmelcultures (CBS 142763), and South African National Collection of Fungi (PREM 61898, dried culture).

2.3.3 Culturing of *C. zeina* for DNA isolation

A glycerol stock of the *C. zeina* CMW 25467 strain was grown on V8 agar [20% (v/v) Campbells V8 juice, 2% (w/v) Bacterial Agar, 0.349% (w/v) $CaCO_3$] containing 100µg/litre Cefotaxime. Regions of dense growth producing conidia were cut from the media and transferred to new V8 agar plates using the patting technique. Cultures were incubated at ambient temperature in constant darkness to promote conidiation. Conidia were collected by washing the media surface in Potato Dextrose Broth (PDB) [2.4% (w/v)] followed by agitation of the media surface to increase the conidial yield in the suspension. Subsequently seven $1x10^5$ conidia/ml aliquots were separately cultured in PDB and incubated at ambient temperature with shaking at 50rpm. Cultures were continuously inspected visually and as cultures reached the melanized stage (~3-5 days) the fungal hyphal tissue was collected by centrifugation and washed with [100mM Tris(hydroxymethyl)aminomethane (Tris), pH8]. The quality of the isolated hyphae tissues of seven cultures were evaluated individually.

2.2.4 Culturing of *C. zeina* for RNA isolation

The *C. zeina* CMW 25467 cultures for RNA isolation was maintained by V. Swart (Swart, 2017). A glycerol stock was grown on V8 agar [20% (v/v) Campbells V8 juice, 2% (w/v) Bacterial Agar, 0.349% (w/v) $CaCO_3$] containing 100µg/l Cefotaxime (Aspen Pharmacare, South Africa). Regions of dense growth producing conidia were cut from the media and transferred to new V8 agar plates using the patting technique. Cultures were incubated at ambient temperature in constant darkness to promote conidiation. Conidiating cultures were maintained on V8 agar prior to conidial transfer to the seven different *in vitro* growth conditions selected for the culturing of *C. zeina* for RNA isolation (Table 2.2). For solid media, conidia were transferred onto sterile cellophane sheets overlaid onto the particular media. For liquid media, conidia were transferred to a 1000ml flask containing 400ml of the relevant growth media. Cultures were incubated on the indicated media for seven days prior to RNA isolation, except for the V8 agar where cultures were incubated for three days (Table 2.2).

Media were prepared as follow:
Complete medium: 1% (w/v) Glucose, 0.1% (w/v) Yeast extract, 0.1% (w/v) Casein hydrolysate, 0.1% (w/v) $Ca(NO_3)_2.4H_2O$, 1% (v/v) mineral solution [2% (w/v) KH_2PO_4, 2.5% (w/v) $MgSO_4.7H_2O$, 1.5% (w/v) NaCl];
PDA AP: PDA supplemented with 10mM $NH_4H_2PO_4$ [0.114% (w/v) $NH_4H_2PO_4$];
Cornmeal agar: 1.7% (w/v) cornmeal agar (Merck Chemicals, Germany;
PDA pH8: 3.9% (w/v) PDA, pH adjusted to pH8 using sodium carbonate-sodium bicarbonate buffer [0.29% (w/v) Na_2CO_3 and 0.76% (w/v) $NaHCO_3$];
PDB pH3: 2.4% (w/v) PDB, pH adjusted to pH3 using citric acid-Na_2HPO_4 [1.67% (w/v) Citric acid and 0.58% (w/v) Na_2HPO_4];
V8 media: 20% (v/v) Campbells V8 juice, 2% (w/v) Bacterial Agar, 0.349% (w/v) $CaCO_3$
Yeast Peptone Dextrose (YPD) media: 0.05% (w/v) Peptone, 0.05% (w/v) Yeast extract, 0.5% (w/v) Glucose, 1.8% (w/v) NaCl.

Table 2.2 *In vitro* growth conditions for the culturing of *C. zeina* (Swart, 2017).

Media	Solid/liquid media	Light conditions	Harvesting time	Notes
Complete medium	Solid	Light	7 dpi	Nutrient rich media
Cornmeal agar	Solid	Light	7 dpi	Grain-rich media
PDA-AP	Solid	Light	7 dpi	Represses cercosporin production in *Cercospora spp.*
PDA pH8	Solid	Light	7 dpi	Alkaline
PDB pH3	Liquid	Light	7 dpi	Acidic
V8 media	Solid	Dark	3 dpi	Induces conidiation
YPD media	Liquid	Light	7 dpi	Media used to culture *C. zeina* for DNA isolation

2.3.5 DNA isolation

C. zeina CMW 25467 DNA was isolated by using a modified Hexadecyltrimethylammonium bromide (CTAB) method (Ma *et al.*, 2010). To prepare the CTAB buffer 125µl ß-Mercaptoethanol was added to 25ml CTAB buffer [77 mM Tris(hydroxymethyl)aminomethane hydrochloride, 28% (v/v) 5M NaCl, 4% (v/v) 0.5M Ethylenediaminetetraacetic acid (EDTA), 55mM CTAB], followed by 0.1% (w/v) Polyvinylpyrrolidone (PVP). The solution was incubated at 65°C until the PVP dissolved.

Fungal hyphae tissues of the seven biological replicates were separately ground in a mortar-and-pestle under liquid nitrogen, with care taken to ensure that the tissue remained frozen. The ground tissue was placed in plastic tubes, and 1ml CTAB buffer and 14U RNase A (QIAGEN, Germany) added to each sample. The tubes were pulse vortexed, incubated at 37°C for 30 min and pulse-vortexed. Following incubation at

65°C for 60 min with tube inversion after 30 min, 1ml Chloroform was added and the tubes inverted for 1 min. The tubes were centrifuged for 20 min at 10,000 x *g* and the aqueous phases transferred to new tubes. A total of 0.8 x (the estimated volume of the aqueous phases) 7.5M ammonium acetate at 4°C, and 0.54 x (the estimated volume of the aqueous phases) 100% isopropanol at 4°C were added and the solutions mixed by repeated inversion for 1 min. Following incubation at -20°C for 60 min the solutions were centrifuged for 20 min at 10,000 x *g* and the supernatants discarded. A total of 1ml 70% (v/v) ethanol was added and incubated at 65°C for 60 min followed by pulse-vortexing. The supernatants were discarded after centrifugation for 10 min at 10,000 x *g*, and 1 ml 70% (v/v) ethanol added. The solutions were pulse-vortexed and the supernatant discarded after centrifugation for 10 min at 10,000 rpm. Ethanol was evaporated and the pellets dissolved with the addition of 50µl ddH$_2$O.

Purified DNA was analysed on the Nanodrop™ ND2000 instrument (ThermoScientific, USA). Purity, RNA contamination and concentration of the DNA was evaluated using gel electrophoresis on a 2% (w/v) agarose gel in TAE buffer (40mM Tris, 20mM Glacial acetic acid, 2mM EDTA). Aliquots of the Generuler 1kb ladder (ThermoFisher Scientific, MA, USA) were analysed in duplicate on the gel to provide size standards. Dilutions of lambda DNA aliquots (stock 300ng/µl, ThermoFisher Scientific, MA, USA) were used as standard to visually assess concentration of the DNA.

2.3.6 RNA isolation

RNA was isolated from *C. zeina* CMW 25467 cultured in seven different growth media by V. Swart (Swart, 2017). *C. zeina* mycelial growth was removed from the respective growth media plates along with the cellophane sheets and flash frozen in liquid nitrogen. For the liquid growth media, the *C. zeina* cultures were filtered through Whatmann's nr 1 filter paper and the retained mycelia flash frozen in liquid nitrogen. Mycelia was ground in liquid nitrogen and 3g used for total RNA isolation using QIAzol Lysis Reagent (QIAGEN, Germany) as per the manufacturer's specifications. Total RNA (100µg) was purified using the RNeasy Mini Kit and the RNase-free DNase set for on-column DNA digestion (QIAGEN, Germany) as per the manufacturer's specifications. The quantity and purity of the RNA was analysed using the Thermo Scientific Nanodrop™ 2000 spectrophotometer (ThermoFisher Scientific). RNA quality was assessed with the Experion™ Automated Electrophoresis System (Bio-Rad, California, USA) using the Experion™ RNA StdSens kit.

2.3.7 High-throughput Sequencing

Three genomic DNA sequencing libraries were prepared by the Purdue Genomics Core Facility (Purdue University, IN, USA) according to the manufacturer's instructions. The libraries comprised of a 400bp paired-end (PE) library, a 3kb mate-pair (MP3KB) library and a 8kb mate-pair (MP8KB) library. Sequencing of the libraries was performed on an Illumina HiSeq2000 instrument using 100bp paired-end sequencing chemistry.

Seven RNAseq libraries were prepared by the Beijing Genome Institute (Hong Kong, China) according to the manufacturer's instructions. Ribosomal RNA was depleted prior to the preparation of 200bp insert libraries. Sequencing of the libraries was performed on an Illumina HiSeq2000 instrument using 100bp paired-end sequencing chemistry.

2.3.8 Quality control of sequencing data

The quality of the sequence reads were evaluated by FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/) using the default parameters. The sequencing reads were not quality filtered and trimmed during the first assembly by F van Staden. Prior to the second and third assemblies the sequence reads for all libraries were quality filtered and trimmed with Trimmomatic v. 0.30 (Bolger *et al.*, 2014) using the following respective commands for each library (RNAseq libraries treated separately):

PE:

```
java -jar trimmomatic-0.32.jar PE -threads 2 -phred64 –trimlog
name_of_PE_logfile.log PE_forward_reads.fastq PE_reverse_reads.fastq
PE_forward_paired.fq PE_forward_single.fq PE_reverse_paired.fq
PE_reverse_single.fq LEADING:12 SLIDINGWINDOW:4:15
```

MP3KB:

```
java -jar trimmomatic-0.32.jar PE -threads 2 -phred64 –trimlog
name_of_MP3KB_logfile.log MP3KB_forward_reads.fastq MP3KB_reverse_reads.fastq
MP3KB_forward_paired.fq MP3KB_forward_single.fq MP3KB_reverse_paired.fq
MP3KB_reverse_single.fq LEADING:12 SLIDINGWINDOW:4:15
```

MP8KB:

```
java -jar trimmomatic-0.32.jar PE -threads 2 -phred64 –trimlog
name_of_MP8KB_logfile.log MP8KB_forward_reads.fastq MP8KB_reverse_reads.fastq
MP8KB_forward_paired.fq MP8KB_forward_single.fq MP8KB_reverse_paired.fq
MP8KB_reverse_single.fq LEADING:12 SLIDINGWINDOW:4:15
```

RNAseq libraries:

```
java -jar trimmomatic-0.32.jar PE -threads 2 -phred64 –trimlog
name_of_RNAlib_logfile.log RNAlib_forward_reads.fastq RNAlib_reverse_reads.fastq
RNAlib_forward_paired.fq RNAlib_forward_single.fq RNAlib_reverse_paired.fq
RNAlib_reverse_single.fq LEADING:12 SLIDINGWINDOW:4:15
```

Settings in the commands were as follows:

threads	Number of CPU cores used
phred64	Phred64 quality scores used for read quality evaluation
LEADING	Number of bases to remove from start of each read
SLIDINGWINDOW	Sliding window of 4 bases with reads with minimum average of Q15 across entire read length retained

2.3.9 Genome assembly

Three genome assemblies were prepared from the genome sequence data, using different combinations of data and assembly software for each assembly.

2.3.9.1 First assembly

The first assembly was performed by F. van Staden using the reads of all three libraries combined. VelvetOptimizer (Zerbino & Birney, 2008) was used to estimate the optimal k-mer size. The assembly was constructed using Velvet, with the optimal k-mer size of 77 bp specified. The command used for the assembly was:

```
velveth  name-of-assembly  77  -fastq  -shortPaired  /path/to/MP3KB/MP3KB.fastq  -
shortPaired2 /path/to/PE/PE.fastq -shortPaired2 /path/to/MP8KB/MP8KB.fastq
```

2.3.9.2 Second assembly

The second assembly was performed by only using the MP8KB library reads. The reads were treated as single-ended to negate the large distance between read-pairs. The CLCAssembler script of the CLC Genomics Workbench (https://www.qiagenbioinformatics.com/) was used to assemble the reads. The command used was:

```
clc_assembler    -q    forward_paired.fq    reverse_paired.fastq    -o
name_of_assembly –p no --cpus 10 --no-scaffolding
```

Settings in the command were as follows:

q	Sequence files in fastq-format
o	Name of the assembly output
p	Paired reads (yes/no)
cpus	Number of server/computer cores used during the assembly
no-scaffolding	Contigs are created but no scaffolds are created from the pair-distance data

Following the completion of the CLCAssembler assembly the completeness of the assembly was evaluated with the Assemblathon_stats.pl script (Bradnam *et al.*, 2013).

2.3.9.3 Third assembly

The third assembly was performed using only the MP8KB library reads, while treating the reads as single-end to negate the large distance between read-pairs. VelvetOptimiser (Zerbino & Birney, 2008) was used to estimate the optimal k-mer size. The assembly was constructed using Velvet, with the optimal k-mer size supplied by VelvetOptimiser and the minimum contig size specified as 200bp. The command used was:

```
perl VelvetOptimiser-2.2.5/VelvetOptimiser.pl --s 61 --e 85 --f '-fastq
-short -shortPaired /path/to/MP8KB/8kb-MP_forward_paired.fq -fastq
-short2 /path/to/8kb-MP/8kb-MP_reverse_paired.fq' --t 10
--o='-min_contig_lgth 200' -d /path/to/output/folder/ --v
```

Settings in the command were as follows:
s	The starting (lower) hash value
e	The end (higher) hash value
f	The file section of the velveth command line
t	The maximum number of simultaneous velvet instances to run
o	The file section of the velveth command line
d	The name of the directory to put the final output into
v	Verbose logging, includes all velvet output in the logfile

Following the completion of the Velvet assembly process, the completeness of the assembly was evaluated with the Assemblathon_stats.pl script (Bradnam *et al.*, 2013).

2.3.10 Scaffolding of third assembly contigs

The read-pair distance inherent in the three libraries were utilized to scaffold the second and third assembly contigs using the SSPACE v. 2.0 software (Boetzer *et al.*, 2011). To determine the optimal errors for the respective library insert sizes, the assembly contigs were scaffolded separately by the reads of each library with varying errors specified. The scaffolding completeness for each error was evaluated using the Assemblathon_stats.pl script. The selection criterion for selecting the optimal error for each library was the error percentage giving rise to the highest N50 value for the resultant assembly. Errors of 25% and 50% were evaluated. The settings input-file for SSPACE specified the error for each library, as well as the orientation of the reads based on the read orientation during sequencing. The PE library has a read orientation of Forward-Reverse (FR), while the two mate-pair libraries have Reverse-Forward (RF) orientations. The scaffolding completeness for the assemblies was evaluated using the Assemblathon_stats.pl script (Bradnam *et al.*, 2013).

2.3.11 Filling of sequence gaps

SSPACE generated scaffolds with multiples of no-call (N) bases in scaffold sequences to serve as placeholders for intra-scaffold sequence gaps. To correctly identify some of these no-call bases the Gapfiller v 1.11 software package (Boetzer & Pirovano, 2012) was utilized. The input and configuration setup is identical to the SSPACE package. Gapfiller removes 10bp from each end of each read, then aligns the reads with the assembled scaffolds to replace no-call bases where appropriate. The optimized error in read insert for each library is used to correlate the read-pair alignment to the scaffolds. The Gapfiller package was used to identify nucleotides in no-call gaps in the third assembly following scaffolding with SSPACE. The gapfilled completeness for the assembly was evaluated using the Assemblathon_stats.pl script (Bradnam *et al.*, 2013).

2.3.12 Completeness evaluation of genome assemblies

To evaluate the completeness of the genome assemblies the CEGMA (Parra *et al.*, 2007) and BUSCO (Simao *et al.*, 2015) packages were used. The BUSCO analysis was performed using the Ascomycota gene set which evaluates the presence of 1,315 BUSCO groups.

2.3.13 RNAseq read mapping

The filtered and trimmed RNAseq datasets were combined into two files, i.e. one file with all the forward reads, and one file with all the reverse reads. The combined reads were mapped to the relevant genome assemblies using the Tophat2 package (Kim *et al.*, 2013). The generic command used was:

```
tophat2 -p 8 -r 0 --mate-std-dev 200 -o /path/to/output/folder/
/path/to/assembly-bowtie-index/index-name /path/to/RNAseq/data/forward-
reads.fastq /path/to/RNAseq/data/reverse-reads.fastq
```

Settings in the command were as follows:

p	Number of CPU cores used during the assembly
r	The expected (mean) inner distance between mate pairs
mate-std-dev	The standard deviation for the distribution on inner distances between mate pairs
o	The name of the TopHat output directory

To evaluate the number of RNAseq reads mapping to the respective assemblies, the `flagstat` function in SAMtools (Li, H. *et al.*, 2009) was used. This function provides a breakdown of the number of reads mapping to the assembly, both in total, and in terms of pairs of reads or singletons mapping, while also providing a measure of reads mapping to multiple areas on the assembly.

2.3.14 Mate-pair library insert size QC

To estimate the insert sizes of the mate-pair libraries used for scaffolding the assembled contigs, the CollectInsertSizeMetrics of the Picard tools package (Broad) was used. The generic command used was:

```
java -jar picard.jar CollectInsertSizeMetrics I=input.bam
O=insert_size_metrics.txt H=insert_size_histogram.pdf  M=0.5
```

Settings in the command were as follows:

I	Input file in bam format
O	Output file
H	Histogram pdf output file
M	Minimum percentage option (default 0.5)

The tool outputs the percentages of read pairs in each of the three orientations (FR, RF, and TANDEM) as a histogram as estimated from the scaffolding distances in the assembly. The minimum percentage option sets a threshold for removing read orientation categories that have fewer than this percentage of overall reads.

2.3.15 Genbank submissions

The draft genome was deposited at DDBJ/ENA/GenBank and is available under the accession number MVDW00000000; Biosample SAMN06067857; Bioproject PRJNA355276. The RNAseq data was deposited in the NCBI Gene Expression Omnibus and is available under the accession number GSE90705.

2.3.16 Phylogenenetic analysis

To verify the identity of the genome sequenced strain as *C. zeina* CMW 25467, the phylogenetic relationship of the genome sequence with other *Cercospora* species was determined. The Translation Elongation Factor 1-alpha (TEF1) and Internal Transcribed Spacer (ITS) sequences for selected *Cercospora* species (Tables 2.2 and 2.3) were downloaded from the Genbank nucleotide database. The relevant sequences for the genome assembly were extracted using blastn. The query sequences for the blastn analysis were from Meisel *et al.* (2009). The extracted sequences are labelled as the genome sequence in the phylogenetic tree and related tables. The two sequences for each species were concatenated and aligned with ClustalW (Thompson *et al.*, 1994). The phylogenetic relationship was inferred using the Maximum Likelihood method based on the Tamura 3-parameter model (Tamura, 1992) using MEGA7 (Kumar *et al.*, 2016), with confidence at nodes gained using bootstrap analysis (Felsenstein, 1985) with 100 bootstrap replicates tested. Branches corresponding to partitions reproduced in less than 50% bootstrap replicates were collapsed. The percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates) are shown next to the branches (Felsenstein, 1985). Initial tree(s) for the heuristic search were obtained by applying the Neighbor-Joining method to a matrix of pairwise distances estimated using the Maximum Composite Likelihood (MCL) approach. A discrete Gamma distribution was used to model evolutionary rate differences among sites (5 categories (+G, parameter = 0.2308)). All positions with less than 95% site coverage were eliminated. The final dataset contained a total of 692 positions. Two strains of *Mycosphaerella thailandica* (CBS 116367 and CPC 10548) were used to root the cladogram.

Table 2.3 Genbank accession information for Internal Transcribed Spacer sequences used for phylogenetic analysis

Genbank	Definition
JX143523.1	*Cercospora achyranthis* strain CBS 132613 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143525.1	*Cercospora alchemillicola* strain CPC 5259 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143528.1	*Cercospora althaeina* strain CBS 126.26 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
AY840520.1	*Cercospora apii* strain CBS 116504 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
KF251296.1	*Cercospora apii* strain CBS 118712 culture-collection CBS:118712 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
AY840536.1	*Cercospora apiicola* strain CBS 116457 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
KF251297.1	*Cercospora ariminensis* strain CBS 137.56 culture-collection CBS:137.56 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomalRNA gene, partial sequence
JX143538.1	*Cercospora armoraciae* strain CBS 115060 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
AY840529.1	*Cercospora beticola* strain CBS 116502 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143561.1	*Cercospora campi-silii* strain CBS 132625 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143570.1	*Cercospora celosiae* strain CBS 132600 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143578.1	*Cercospora chinensis* strain CBS 132612 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
KC005779.1	*Cercospora chrysanthemoides* culture-collection CPC:20529 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143583.1	*Cercospora coniogrammes* strain CBS 132634 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
KJ886441.1	*Cercospora convolvulicola* strain CCTU 1083 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence

KJ886445.1	*Cercospora conyzae-canadensis* strain CCTU 1119 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143584.1	*Cercospora corchori* strain MUCC 585 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143587.1	*Cercospora delaireae* strain CBS 132595 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143591.1	*Cercospora dispori* strain CBS 132608 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143593.1	*Cercospora euphorbiae-sieboldianae* strain CBS 113306 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partialsequence
JX143594.1	*Cercospora fagopyri* strain CBS 132623 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
KJ886513.1	*Cercospora iranica* strain CCTU 1137 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143619.1	*Cercospora kikuchii* strain CBS 132633 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143621.1	*Cercospora lactucae-sativae* strain CBS 132604 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143627.1	*Cercospora mercurialis* strain CBS 549.71 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143632.1	*Cercospora olivascens* strain CBS 253.67 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143634.1	*Cercospora pileicola* strain CBS 132607 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143638.1	*Cercospora punctiformis* strain CBS 132626 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143646.1	*Cercospora ricinella* strain CBS 132605 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143649.1	*Cercospora senecionis-walkeri* strain CBS 132636 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
KJ886523.1	*Cercospora solani* strain CCTU 1043 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
KJ886525.1	*Cercospora sorghicola* strain CCTU 1173 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence

DQ185071.1	*Cercospora sp.* F JZG-2013 strain CPC 12062 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143734.1	*Cercospora vignigena* strain CBS 132611 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143737.1	*Cercospora violae* strain CBS 251.67 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
DQ185072.1	*Cercospora zeae-maydis* strain CBS 117755 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
DQ185073.1	*Cercospora zeae-maydis* strain CBS 117756 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
DQ185074.1	*Cercospora zeae-maydis* strain CBS 117757 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
DQ185075.1	*Cercospora zeae-maydis* strain CBS 117758 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
DQ185076.1	*Cercospora zeae-maydis* strain CBS 117759 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
DQ185077.1	*Cercospora zeae-maydis* strain CBS 117760 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
DQ185078.1	*Cercospora zeae-maydis* strain CBS 117761 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
DQ185079.1	*Cercospora zeae-maydis* strain CBS 117762 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
DQ185080.1	*Cercospora zeae-maydis* strain CBS 117763 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
JX143742.1	*Cercospora zeae-maydis* strain CBS 132668 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
EU569228.1	*Cercospora zeina* strain CMW 25442 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
EU569225.1	*Cercospora zeina* strain CMW 25445 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
EU569219.1	*Cercospora zeina* strain CMW 25448 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
EU569229.1	*Cercospora zeina* strain CMW 25452 internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence

EU569226.1	*Cercospora zeina* strain CMW 25459 internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
EU569224.1	*Cercospora zeina* strain CMW 25462 internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
EU569227.1	*Cercospora zeina* strain CMW 25467 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
DQ185081.1	*Cercospora zeina* strain CPC 11995 (ex-type) 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence
KF901776.1	*Pseudocercospora thailandica* culture-collection CBS:116367 internal transcribed spacer 1, partial sequence; 5.8S ribosomal RNA gene, complete sequence; and internal transcribed spacer 2, partial sequence
AY752157.1	*Pseudocercospora thailandica* strain CPC 10548 18S ribosomal RNA gene, partial sequence; internal transcribed spacer 1, 5.8S ribosomal RNA gene, and internal transcribed spacer 2, complete sequence; and 28S ribosomal RNA gene, partial sequence

Table 2.4 **Genbank accession information for Translation Elongation Factor 1-alpha (TEF1) sequences used for phylogenetic analysis**

Genbank	Definition
JX143277.1	*Cercospora achyranthis* strain CBS 132613 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143279.1	*Cercospora alchemillicola* strain CPC 5259 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143282.1	*Cercospora althaeina* strain CBS 126.26 translation elongation factor 1-alpha (tef1) gene, partial cds
AY840487.1	*Cercospora apii* strain CBS 116504 translation elongation factor 1-alpha (tef1) gene, exons 2, 3 and partial cds
KF253244.1	*Cercospora apii* strain CBS 118712 culture-collection CBS:118712 translation elongation factor 1-alpha (tef1) gene, partial cds
AY840503.1	*Cercospora apiicola* strain CBS 116457 translation elongation factor 1-alpha (tef1) gene, exons 2, 3 and partial cds
KF253245.1	*Cercospora ariminensis* strain CBS 137.56 culture-collection CBS:137.56 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143292.1	*Cercospora armoraciae* strain CBS 115060 translation elongation factor 1-alpha (tef1) gene, partial cds
AY840496.1	*Cercospora beticola* strain CBS 116502 translation elongation factor 1-alpha (tef1) gene, exons 2, 3 and partial cds
JX143315.1	*Cercospora campi-silii* strain CBS 132625 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143326.1	*Cercospora celosiae* strain CBS 132600 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143334.1	*Cercospora chinensis* strain CBS 132612 translation elongation factor 1-alpha (tef1) gene, partial cds
KC005813.1	*Cercospora chrysanthemoides* culture-collection CPC:20529 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143341.1	*Cercospora coniogrammes* strain CBS 132634 translation elongation factor 1-alpha (tef1) gene, partial cds
KJ886280.1	*Cercospora convolvulicola* strain CCTU 1083 translation elongation factor 1-alpha (tef1) gene, partial cds
KJ886284.1	*Cercospora conyzae-canadensis* strain CCTU 1119 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143342.1	*Cercospora corchori* strain MUCC 585 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143345.1	*Cercospora delaireae* strain CBS 132595 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143349.1	*Cercospora dispori* strain CBS 132608 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143351.1	*Cercospora euphorbiae-sieboldianae* strain CBS 113306 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143352.1	*Cercospora fagopyri* strain CBS 132623 translation elongation factor 1-alpha (tef1) gene, partial cds
KJ886352.1	*Cercospora iranica* strain CCTU 1137 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143378.1	*Cercospora kikuchii* strain CBS 132633 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143380.1	*Cercospora lactucae-sativae* strain CBS 132604 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143386.1	*Cercospora mercurialis* strain CBS 549.71 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143391.1	*Cercospora olivascens* strain CBS 253.67 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143393.1	*Cercospora pileicola* strain CBS 132607 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143397.1	*Cercospora punctiformis* strain CBS 132626 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143405.1	*Cercospora ricinella* strain CBS 132605 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143408.1	*Cercospora senecionis-walkeri* strain CBS 132636 translation elongation factor 1-alpha (tef1) gene, partial cds
KJ886362.1	*Cercospora solani* strain CCTU 1043 translation elongation factor 1-alpha (tef1) gene, partial cds

KJ886364.1	*Cercospora sorghicola* strain CCTU 1173 translation elongation factor 1-alpha (tef1) gene, partial cds
DQ185083.1	*Cercospora sp.* F JZG-2013 strain CPC 12062 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143493.1	*Cercospora vignigena* strain CBS 132611 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143496.1	*Cercospora violae* strain CBS 251.67 translation elongation factor 1-alpha (tef1) gene, partial cds
DQ185084.1	*Cercospora zeae-maydis* strain CBS 117755 translation elongation factor 1-alpha (tef1) gene, partial cds
DQ185085.1	*Cercospora zeae-maydis* strain CBS 117756 translation elongation factor 1-alpha (tef1) gene, partial cds
DQ185086.1	*Cercospora zeae-maydis* strain CBS 117757 translation elongation factor 1-alpha (tef1) gene, partial cds
DQ185087.1	*Cercospora zeae-maydis* strain CBS 117758 translation elongation factor 1-alpha (tef1) gene, partial cds
DQ185088.1	*Cercospora zeae-maydis* strain CBS 117759 translation elongation factor 1-alpha (tef1) gene, partial cds
DQ185089.1	*Cercospora zeae-maydis* strain CBS 117760 translation elongation factor 1-alpha (tef1) gene, partial cds
DQ185090.1	*Cercospora zeae-maydis* strain CBS 117761 translation elongation factor 1-alpha (tef1) gene, partial cds
DQ185091.1	*Cercospora zeae-maydis* strain CBS 117762 translation elongation factor 1-alpha (tef1) gene, partial cds
DQ185092.1	*Cercospora zeae-maydis* strain CBS 117763 translation elongation factor 1-alpha (tef1) gene, partial cds
JX143501.1	*Cercospora zeae-maydis* strain CBS 132668 translation elongation factor 1-alpha (tef1) gene, partial cds
EU569216.1	*Cercospora zeina* strain CMW 25442 translation elongation factor 1-alpha (tef1) gene, partial cds
EU569217.1	*Cercospora zeina* strain CMW 25445 translation elongation factor 1-alpha (tef1) gene, partial cds
EU569208.1	*Cercospora zeina* strain CMW 25448 translation elongation factor 1-alpha (tef1) gene, partial cds
EU569213.1	*Cercospora zeina* strain CMW 25452 translation elongation factor 1-alpha (tef1) gene, partial cds
EU569215.1	*Cercospora zeina* strain CMW 25459 translation elongation factor 1-alpha (tef1) gene, partial cds
EU569210.1	*Cercospora zeina* strain CMW 25462 translation elongation factor 1-alpha (tef1) gene, partial cds
EU569218.1	*Cercospora zeina* strain CMW 25467 translation elongation factor 1-alpha (tef1) gene, partial cds
DQ185093.1	*Cercospora zeina* strain CPC 11995 (ex-type) translation elongation factor 1-alpha (tef1) gene, partial cds
AY840476.1	*Mycosphaerella thailandica* strain CBS 116367 translation elongation factor 1-alpha (tef1) gene, exons 2, 3 and partial cds
AY840477.2	*Mycosphaerella thailandica* strain CPC 10548 translation elongation factor 1-alpha (tef1) gene, exons 2, 3 and partial cds

2.3.17 Transcriptome assembly

The combined RNAseq reads (Section 2.3.13) were assembled into a transcriptome using the Trinity software package (Grabherr *et al.*, 2011 ; Haas *et al.*, 2013). The generic command used was:

```
perl      /path/to/trinity/Trinity.pl      --seqType      fq      --left
/path/to/reads/forward-reads.fq --right /path/to/reads/reverse-reads.fq
--output /path/to/output/directory --JM 50G --CPU 12
--path_reinforcement_distance 0 --group_pairs_distance 400 --jaccard_clip
```

Settings in the command were as follows:

seqType	Input file-type, here fastq
right	Forward read input file
left	Reverse read input file
output	Output directory
JM	Limit memory usage to 50Gb
CPU	Number of CPU cores used in the assembly
jaccard_clip	Minimize transcript fusion for gene-dense genome
group_pairs_distance	Maximum length expected between fragment pairs
path_reinforcement_distance	Minimum overlap of reads with growing transcript

The completeness of the assembly was evaluated using the Assemblathon_stats.pl script (Bradnam *et al.*, 2013) as well as the BUSCO (Simao *et al.*, 2015) package.

2.4 Results

2.4.1 DNA isolation

DNA was isolated from seven cultures of the *C. zeina* CMW 25467 strain to ensure adequate backup for sequencing purposes. The DNA was evaluated for purity on a Nanodrop™ (Table 2.5), with the A260/A280 ratios falling into the acceptable range for all samples (according to the sequencing provider's requirements). To evaluate the level of RNA contamination in the samples, a 1µl aliquot of each sample was analysed using agarose gel electrophoresis (Figure 2.1). None of the samples showed RNA contamination, while the DNA concentration was estimated by comparison to a serial dilution of lambda phage-DNA standards on the same gel (according to the sequencing provider's specifications). A total of ~200µl purified DNA was obtained for each isolate, and shipped to the sequencing service provider.

Table 2.5 Concentration and purity data for DNA isolations from *C. zeina* CMW 25467 cultures. *DNA concentrations were estimated from the lambda-phage DNA standard serial dilution in the agarose gel.*

Sample ID	A260/A280	A260/A230	Concentration from gel (ng/µl)*
Cz 1	1.97	1.26	15
Cz 2	2.00	2.02	63
Cz 3	1.92	1.45	30
Cz 4	1.93	1.64	30
Cz 5	1.97	1.72	40
Cz 6	1.94	1.49	30
Cz 7	1.95	1.45	30

* Please refer to Figure 2.1

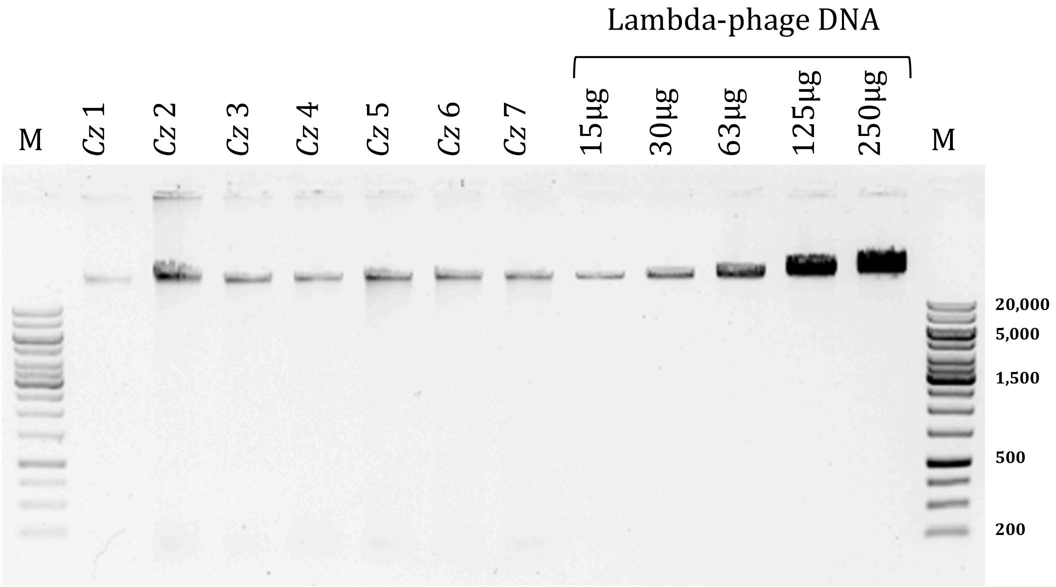


Figure 2.1 Agarose gel image of DNA isolated from seven *C. zeina* CMW25467 cultures (*Cz 1-Cz 7*). *A 1µl aliquot was loaded on the gel for each isolate. A Fermentas Generuler™ 1kb ladder (M) aliquot was loaded to evaluate fragment sizes (base pair sizes for selected bands indicated). A serial dilution of lambda-phage DNA (as indicated) was loaded to evaluate concentrations in the indicated lanes.*

The sequencing provider utilized several of the gDNA samples during optimization of the library building process, but no further specifics were provided on which gDNA samples were sequenced, and whether samples were pooled during library construction.

2.4.2 RNA isolation

The RNA isolation was performed by Ms V. Swart during her PhD study, where she investigated the expression of cercosporin-synthesis genes. The results were included here for completeness, but were discussed in her PhD thesis (Swart, 2017). RNA was isolated and 3µl of each sample submitted to Experion™ RNA analysis. The Experion™ analysis provides a RNA Quality Indicator (RQI) number that provides a numerical indication of the intactness of the RNA. The RQI values can range from 0 for totally degraded RNA, to 10 for completely intact RNA. For RNAseq analysis it is ideal to sequence samples with RQI values >8. The samples submitted for RNAseq conformed to the basic quality requirement for the samples (Table 2.6, Figure 2.2), and also passed the basic QC requirement performed by the sequencing service provider upon sample delivery.

Table 2.6 Purity and quality analysis of RNA isolated from seven growth conditions for sequencing. *Each sample represented a pool of three biological replicates.*

Sample Name	Nanodrop™ Purity (A260/A280)	Nanodrop™ Purity (A260/A230)	Experion™ RNA Concentration (ng/µl)	Experion™ RNA Quality Indicator
Complete Media	2.16	2.27	468.29	9.0
Cornmeal Agar	2.16	2.44	403.96	10
PDA-AP	2.16	2.43	547.22	9.9
PDA-pH8	2.15	2.40	326.17	9.8
PDB-pH3	2.16	2.43	617.11	8.1
V8	2.13	2.47	310.39	10
YPD	2.14	2.41	475.84	8.7

2.4.3 Sequencing data

The sequencing data was downloaded from the relevant sequencing facility ftp-servers. The data was in fastq format, and the number of reads and amount of sequence base-pairs per sample are listed in Table 2.7.

Table 2.7 Illumina sequence data obtained for *C. zeina* genomic and transcriptomic sequencing libraries. *The total number of reads and base pairs for each sample are a combined count for both forward and reverse reads.*

Sample, Library type	Total data / reads	Sequence Facility
Genomic DNA, paired-end (400 bp), 2 x 100bp	2.8Gbp / 28,291,520 total reads	Purdue Genomics Core Facility, Purdue University, IN, USA
Genomic DNA, short mate-pair (3Kbp), 2 x 100bp	3.5Gbp / 34,807,200 total reads	
Genomic DNA, long mate-pair (8Kbp) , 2 x 126 bp	10.1Gbp / 80,335,114 total reads	
Complete media RNA, paired-end, 2 x 100bp	2.8Gbp / 27,615,964 total reads	BGI-Hong Kong, China
Cornmeal agar RNA, paired-end, 2 x 100bp	2.8Gbp / 27,776,658 total reads	
PDA AP RNA, paired-end, 2 x 100bp	2.6Gbp / 26,300,084 total reads	
PDA pH8 RNA, paired-end, 2 x 100bp	2.6Gbp / 26,021,246 total reads	
PDB pH3 RNA, paired-end, 2 x 100bp	2.7Gbp / 27,198,070 total reads	
V8 media RNA, paired-end, 2 x 100bp	2.7Gbp / 26,831,318 total reads	
YPD RNA, paired-end, 2 x 100bp	2.7Gbp / 27,472,074 total reads	

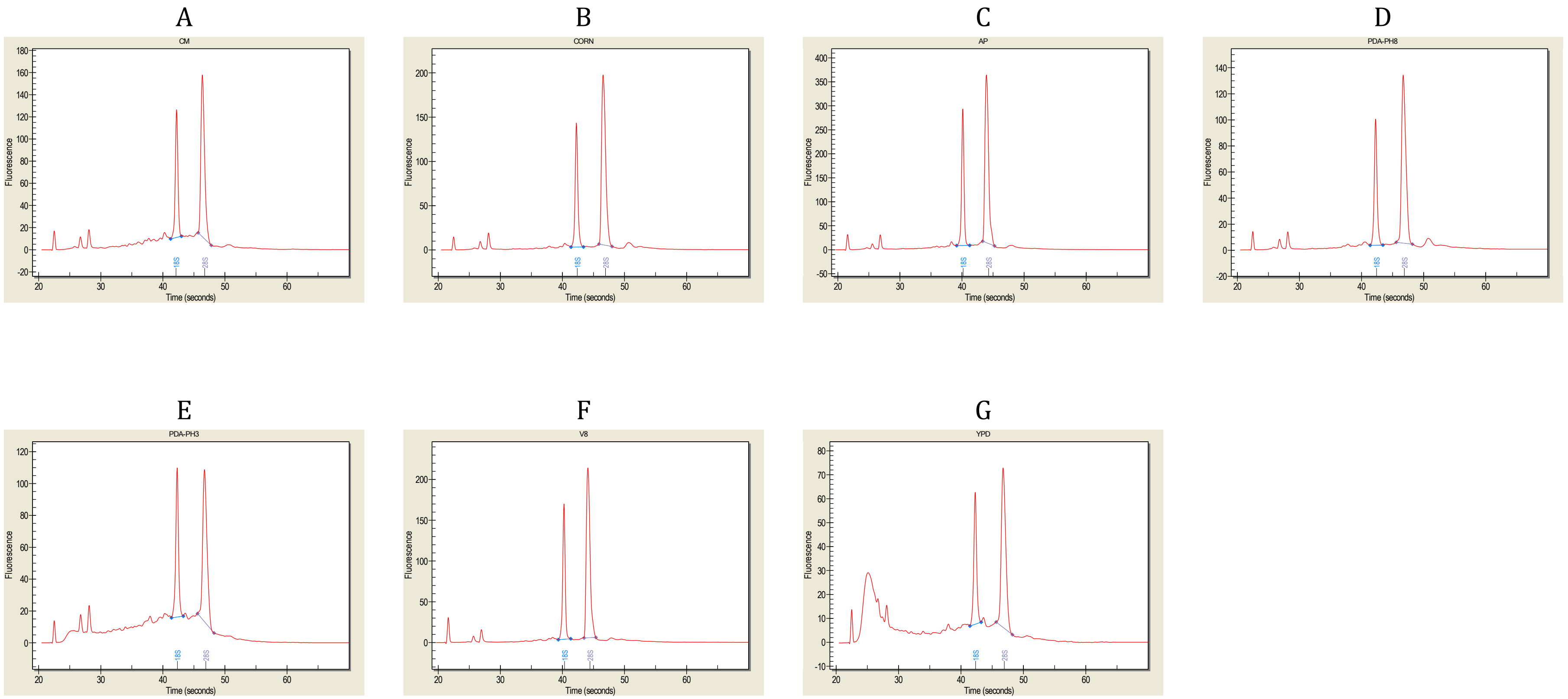


Figure 2.2 Experion™ electrogram profiles for the RNA isolated from seven growth conditions for sequencing. *Each sample represented a pool of three biological replicates. A) Complete Media; B) Cornmeal Agar; C) PDA-AP; D) PDA-pH8; E) PDB-pH3; F) V8 agar; and G) YPD.*

2.4.4 Sequencing data quality control

To evaluate the data quality and content for each sequence file, the FastQC script was used with default parameters, and the output was written to html format. The output report provides quality information for the following metrics:

- Per base sequence quality
- Per sequence quality scores
- Per base sequence content
- Per base GC content
- Per sequence GC content
- Per base N content
- Sequence Length Distribution
- Sequence Duplication Levels
- Overrepresented sequences
- Kmer Content

Along with detailed graphs for the respective metric options per aggregate sequence base position, information is provided whether the data set passes or fails the specific metric, or notes any associated warnings not serious enough for indicating a failed metric.

The sequencing data QC was not performed before the first assembly, and as a result the anomalous Per base GC and Per sequence GC content for both the PE (Figure 2.3 C, D) and MP3KB (Figure 2.4 C, D) libraries were not discovered. The cause of the anomalies is not known, but it is most likely an artefact of the library building chemistry used. Due to these anomalies, it was possible that the assembly from these reads would produce a non-optimal genome assembly. The Per base GC and Per sequence GC content (Figure 2.5 C, D) for the MP8KB library showed no anomalies.

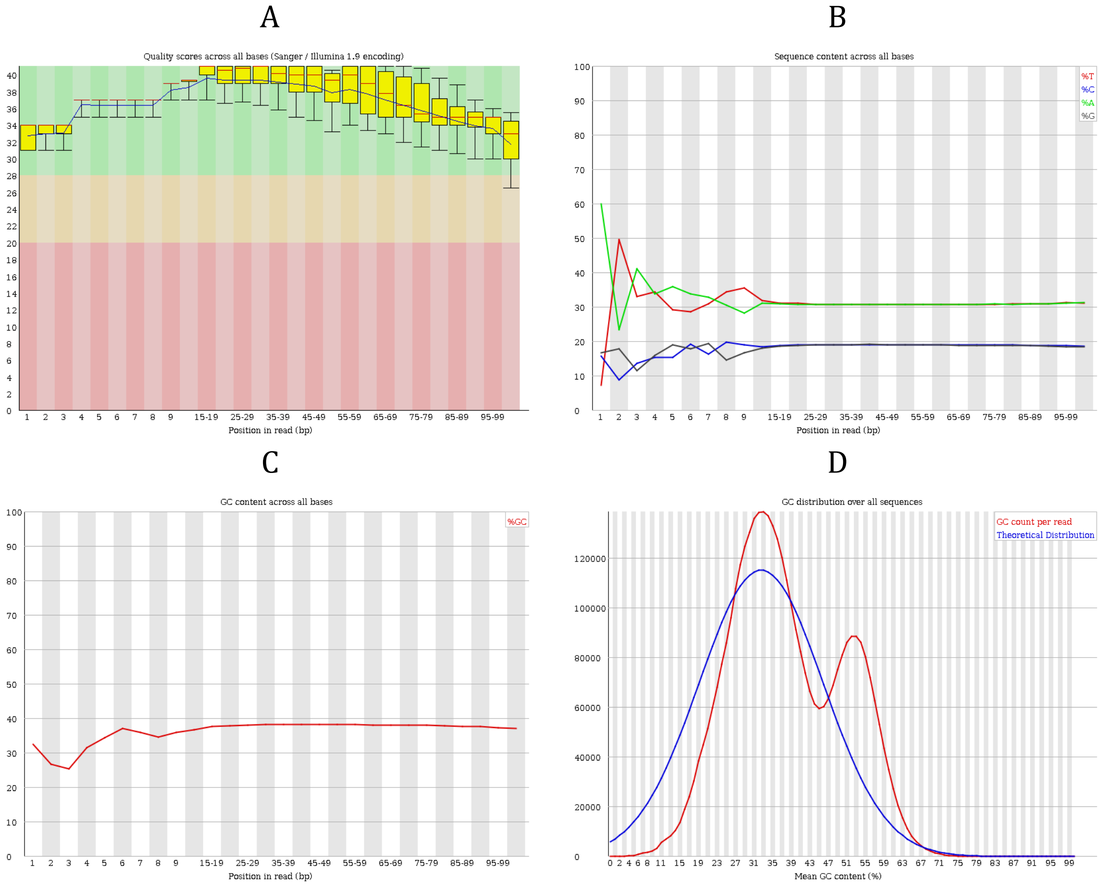


Figure 2.3 FastQC output for selected quality criteria for the PE genome sequencing library. *A) Per base sequence quality; B) Per base sequence content; C) Per base GC content; and D) Per sequence GC content.*

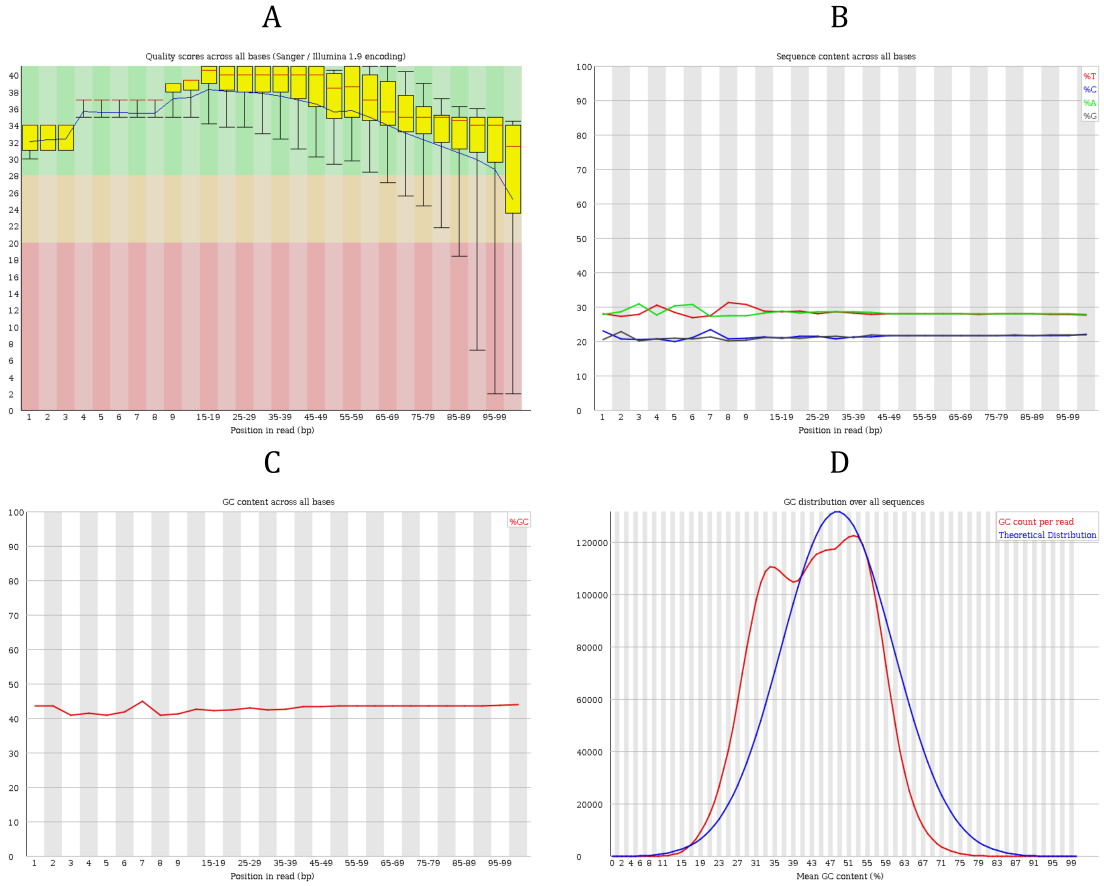


Figure 2.4 FastQC output for selected quality criteria for the MP3KB genome sequencing library. *A) Per base sequence quality; B) Per base sequence content; C) Per base GC content; and D) Per sequence GC content.*

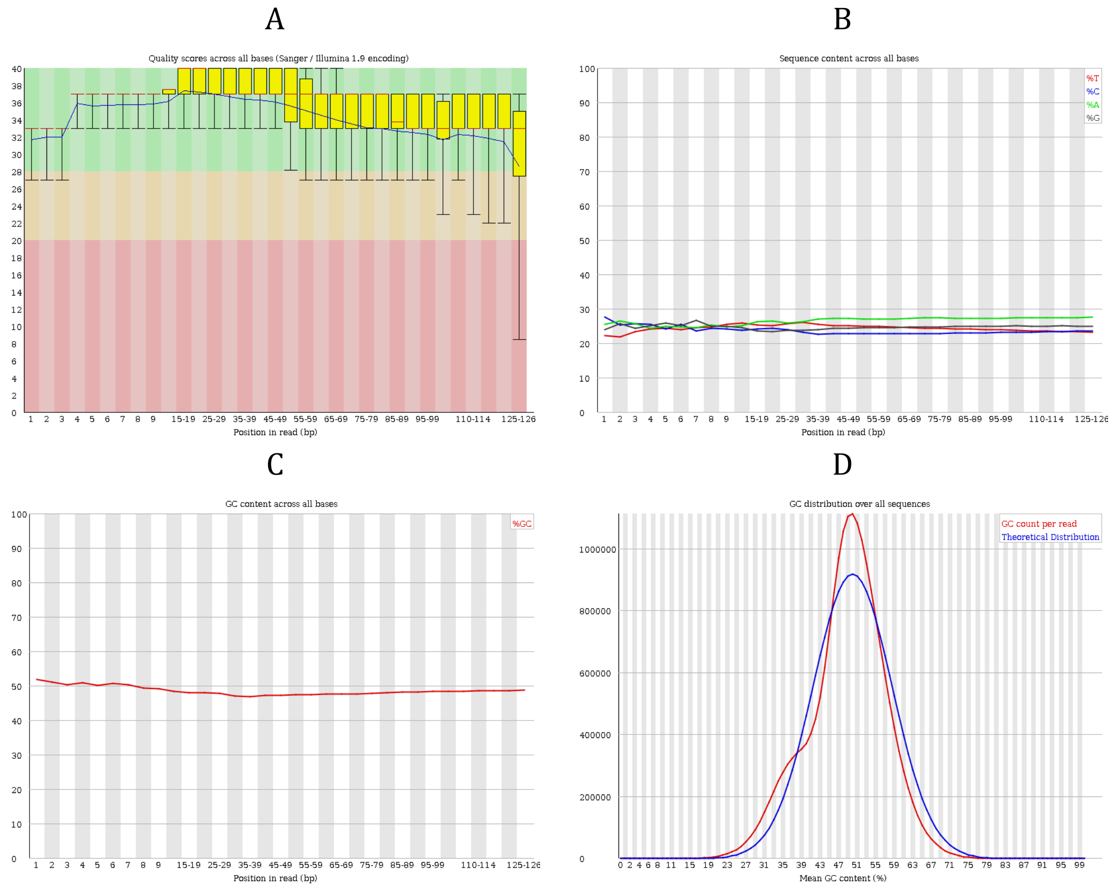


Figure 2.5 FastQC output for selected quality criteria for the MP8KB genome sequencing library. *A) Per base sequence quality; B) Per base sequence content; C) Per base GC content; and D) Per sequence GC content.*

2.4.5 First assembly

The first assembly was performed by F. van Staden, without performing quality filtering and trimming on the raw sequence data. The PE library data was used for the assembly, while the MP3KB and MP8KB libraries were used to scaffold the assembly using SSPACE. The final scaffolded assembly showed a N50 of 720,376 bp, as well as a total size close to the expected genome size (Table 2.8).

Table 2.8 Assembly statistics of the first assembly following contig scaffolding with SSPACE.

Description	Assembly statistics
Number of scaffolds	913
Total size of scaffolds	46,610,549 bp
Longest scaffold	2,282,376 bp
Shortest scaffold	2,001 bp
N50 scaffold length	720,376 bp
N content (%)	10.53%

The first assembly CEGMA completeness evaluation predicted the genome to be 95.16% complete with 236 of 248 complete conserved genes present. The

BUSCO evaluation using the Ascomycota dataset yielded a completeness report of C: 98.5%: (98.3% Complete and single-copy BUSCOs, D: 0.2% duplicated BUSCOs, 0.6% fragmented BUSCOs, M: 0.9% missing BUSCOs, total 1,315 genes evaluated).

The pooled RNAseq reads of the *in vitro* growth conditions were mapped to the first assembly to evaluate to what extent the assembly represented a transcriptionally functional genome. From the 189,215,414 total reads analysed in the combined RNAseq samples only 964,323 (0.5%) reads mapped to the assembly, while 188,251,091 (99.5%) reads did not map. The GC-content abnormalities in the PE and MP3KB libraries did therefore influence the quality of the assembly, and the assembly was not transcriptionally functional.

2.4.6 Second assembly

The absence of GC-content abnormalities, as well as the large amount of sequence data in the MP8KB library suggested that this library could serve as basis for the genome assembly. The sequence reads were treated as single end to negate the possible large insert size bias in the assembly. CLC Genomics Workbench yielded an assembly with a small N50 value of 2,662 bp. Upon scaffolding the assembly contigs in SSPACE using the PE and MP8KB libraries, and allowing for library insert size errors of 50%, the completeness of the assembly improved, with the N50 value increasing to 42,169 bp and an assembled genome size close to the expected genome size (Table 2.9). A concern was the large percentage of N-bases following scaffolding, which indicated the genome was still significantly fragmented. Due to the fragmentation and poor assembly statistics the RNAseq reads were not mapped to this assembly.

Table 2.9 Statistics of the second assembly before and after contig scaffolding with SSPACE.

Description	Before SSPACE	After SSPACE
Number of scaffolds	19,096	8,678
Total size of scaffolds	29,935,079 bp	46,203,281 bp
Longest scaffold	21,164 bp	241,008 bp
Shortest scaffold	200bp	200bp
N50 scaffold length	2,662 bp	42,169 bp
N content (%)	0%	35.21%

2.4.7 Third assembly

The VelvetOptimizer software package was used to estimate the optimum hash value (kmer size) for the third assembly. VelvetOptimizer subsequently passed the optimized values to Velvet to perform the assembly, and the minimum contig size was set to 200 bp. As with the second assembly, the MP8KB library was used while the sequence reads were treated as single end. Following the assembly process the assembly showed a marked improvement in completeness when compared to the second assembly. After scaffolding with SSPACE, using all three libraries and a library insert size error of 25%, the assembly completeness improved with the N50 value increasing to 156,483 bp and a total genome containing a significantly smaller number of N-bases when compared to the second assembly. Gapfilling increased the N50 value to a final value of 160,632 bp with a total genome assembly smaller than the expected genome size, but less fragmented and with a smaller percentage of N-bases than the second assembly (Table 2.10).

Table 2.10 Statistics of the third assembly. *The data before and after scaffolding with SSPACE and after the filling of sequence gaps with Gapfiller are included*.

Description	Before SSPACE	After SSPACE	After Gapfiller
Number of scaffolds	21,609	10,044	10,044
Total size of scaffolds	36,579,835 bp	41,431,997 bp	40,773,084 bp
Longest scaffold	205,526 bp	940,459 bp	938,006 bp
Shortest scaffold	200 bp	200 bp	119 bp
N50 scaffold length	21,078 bp	156,483 bp	160,632 bp
N content (%)	0%	11.72%	9.01%

The CEGMA completeness prediction for the third assembly was 94.7% with 235 of 248 complete conserved genes predicted. The BUSCO evaluation using the Ascomycota dataset yielded a completeness report of C: 95.4%: (95.4% Complete and single-copy BUSCOs, D: 0.0% duplicated BUSCOs, 2.1% fragmented BUSCOs, M: 2.5% missing BUSCOs, total 1,315 genes evaluated).

The pooled RNAseq reads of the *in vitro* growth conditions were mapped to the third assembly to evaluate to what extent the assembly could represent a transcriptionally functional genome. From a total of 189,215,414 analysed reads in the combined RNAseq sample pool 187,971,976 (99.3%) reads mapped to the assembly, while 1,243,438 (0.7%) reads did not map. The third assembly could therefore be considered as a transcriptionally functional assembly and could be used in functional studies.

2.4.9 Sequencing library insert size estimation

The insert sizes of the three genome sequencing libraries (PE, MP3KB and MP8KB) were estimated with the Picard tools CollectInsertSizeMetrics tool. The tool maps the relevant library reads to the assembly and provides statistical data for size distribution of the reads in these libraries, as well as a graphical histogram output that represents the size distribution of the inserts. This data allows the estimation of the size-selection accuracy during library preparation. This would impact scaffolding of contigs since the incorrect size-selection would lead to the incorrect size specified to SSPACE, and therefore the distances between paired reads on different scaffolds would be incorrectly applied during the scaffolding process.

Table 2.11 Insert size estimation data for the genome sequencing libraries. *The number of read-pairs indicate the number of reads involved in the scaffolding process. The read pair orientation for paired-end libraries should be FR, while mate-pair libraries should have a RF orientation.*

Library	Median insert size (bp)	Minimum insert size (bp)	Maximum insert size (bp)	Mean insert size (bp)	Number of read pairs	Read pair orientation
PE	227	47	644134	235	13,271,629	FR
MP3KB	296	57	832,693	301	1,682,063	FR
	3,182	68	558,406	459	2,789,574	RF
MP8KB	5,751	30	822,181	6,172	10,787,711	RF

The PE library (Figure 2.6) showed a median insert size of 227 bp, which is almost half of the expected size of 400 bp as selected during library construction, with a size deviation of 44%. The MP3KB library (Table 2.11, Figure 2.7A), showed a correct median insert size according to the sequencing service provider workflow. The small mean insert size showed that the library size selection was not performed to standard, which is confirmed by the large amount of reads incorrectly predicted with a read pair orientation of FR which is more characteristic of a paired-end library. The MP8KB library (Table 2.11, Figure 2.7B) showed a mean insert size which was just outside the 25% error range specified to SSPACE (28% of 8,000 bp), but was suitable for use in the scaffolding process. The library showed a smaller than expected median insert size (71% of 8,000 bp). The absence of reads predicted in the RF orientation (Table 2.11) supports the acceptable insert size correlation with the experimental protocols.

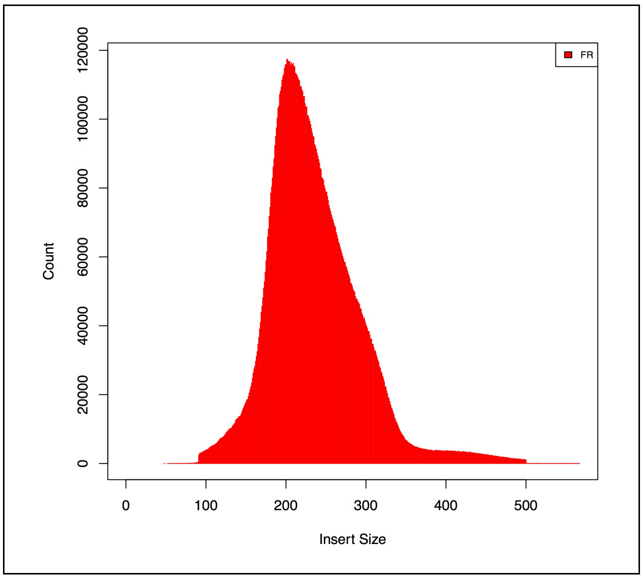


Figure 2.6 Paired-end library insert size estimation by Picard tools. *Insert sizes of the PE library with all data correctly identified in the FR orientation.*

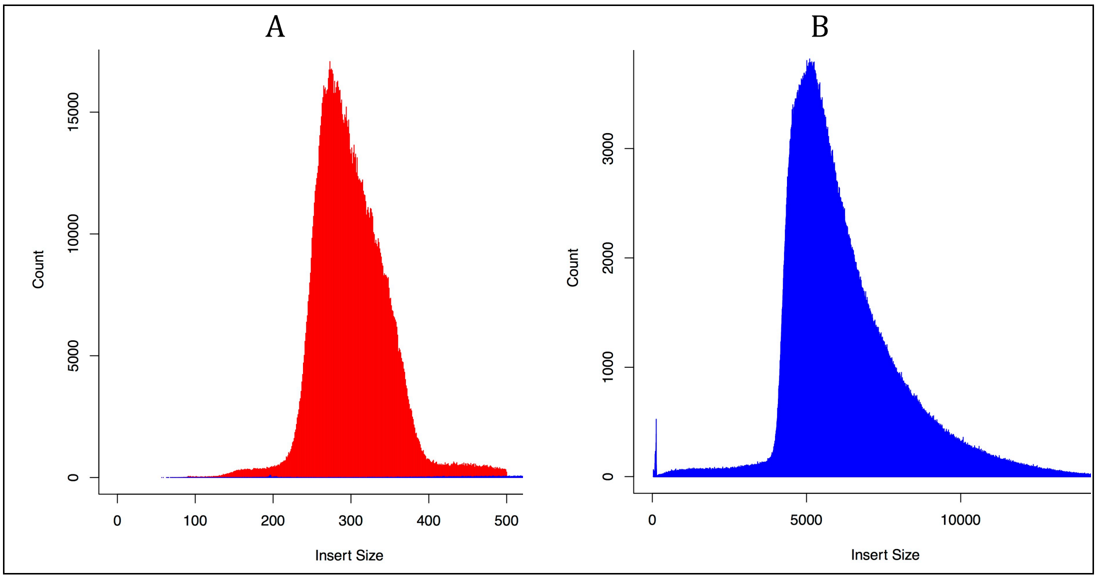


Figure 2.7 Mate-pair library insert size estimation by Picard tools. *A) Insert sizes of the MP3KB library with the majority of data incorrectly identified in the FR orientation and the data for the RF orientation not provided, B) Insert sizes of the MP8KB library with all data correctly identified in the RF orientation.*

2.4.8 Genbank submission

The genome assembly was uploaded to the NCBI Whole Genome Sequence (WGS) database. The Genbank automated quality control software indicated some abnormalities to be corrected in the third assembly before it could be accepted for acceptance and public upload. Firstly, due to the remaining sequence library adapters not being removed during the quality trimming and filtering step, some sequences were still present in the genome assembly and were removed manually. Secondly, the contigs with sizes smaller than 200bp were removed from the assembly due to the strict size-limit cutoff for upload to the NCBI WGS database. Thirdly, two of the contigs were identified as phage sequence, i.e. Salmonella phage iEPS5 and Enterobacteria phage phiX174. Both these contigs were removed from the third assembly. Subsequently the genome assembly completeness (Table 2.12) was evaluated using the Assemblathon_stats.pl script (Bradnam *et al.*, 2013).

Table 2.12 **Statistics of the third assembly following Genbank submission changes.**

Description	**Final statistics**
Number of scaffolds	10,023
Total size of scaffolds	40,755,333 bp
Longest scaffold	937,986 bp
Shortest scaffold	200 bp
N50 scaffold length	160,632 bp
N content (%)	9.02%
GC content in scaffolds (%) – N excluded	49.7%
Percentage of assembly in scaffolded contigs	92.1%
Percentage of assembly in unscaffolded contigs	7.9%

2.4.9 Phylogenetic analysis

Due to the slow growth-rate of *C. zeina* in culture, the possibility of fungal or bacterial contamination with faster growth rates is a concern. The genome assembly resulting from such contaminated reads could yield a hybrid assembly containing sequences from multiple species, and therefore the evolutionary relationship with other *Cercospora* species was evaluated. Following the methodology of Meisel *et. al.* (2009) the ITS and TEF1 gene sequences were obtained for 55 *Cercospora* species and concatenated. The relevant sequences for the genome assembly were obtained as described, and included in the analysis. For both of the query sequences only one blastn hit was obtained from the genome assembly, showing that these regions were not duplicated during the assembly process, or that potential contamination was most likely not present. The genome assembly sequences showed one base difference for each gene when compared to the CMW 25467 strain (Meisel *et al.*, 2009). Since the assembly was generated from DNA from the *C. zeina* CMW 25467 strain these sequences were expected to be identical. The ClustalW alignment of the sequences was trimmed to only include blocks of globally representative sequence regions, since the sequenced regions for some species displayed longer or shorter gene-lengths and thus introduced areas with low phylogenetic information if not removed. The maximum likelihood consensus tree (Figure 2.8) showed a clear distinction between the *C. zeina* isolates and other *Cercospora* species with a significant bootstrap support for the separation. Within the *C. zeina* clade, however, the bootstrap support is not significant, indicating the expected low diversity within the isolates analysed. The genome assembly grouped together with the *C. zeina* CMW 25467 isolate (Meisel *et al.*, 2009) as expected. Most relevant, though, is the unequivocal clustering of the genome assembly sequences within the *C. zeina* clade with very strong bootstrap support, indicating that there was no significant fungal contamination in the fungal DNA isolate submitted for sequencing. Together with the very low number of non-fungal sequence contigs discovered during the Genbank submission process, it is clear that the genome assembly did indeed arise from a *C. zeina* isolate.

2.4.10 Transcriptome assembly

To obtain a transcriptome with as many representative transcripts as possible, the RNAseq reads for all the sequenced conditions (Table 2.1, 2.6) were combined. The Trinity package was used to create the transcriptome assembly. Since this is a fungal genome and the gene density is expected to be higher, the `jaccard_clip` option was used to minimize falsely fusion transcripts. The allowed memory for the assembly was also limited, which negatively impacted the operation time, but made more optimal use of the computing resources. The assembly showed a larger total size than the expected genome size, and with a N50 value that is larger than the expected average gene length for the species (Table 2.13).

Table 2.13 Statistics of the Trinity transcriptome assembly.

Description	Assembly statistics
Number of contigs	20,988
Total size of contigs	62,146,746 bp
Longest contig	26,356 bp
Shortest contig	201 bp
N50 contig length	4,974 bp
A content (%)	22.69%
C content (%)	27.35%
G content (%)	27.41%
T content (%)	22.55%

The BUSCO evaluation using the Ascomycota dataset and the transcriptome setting yielded a completeness report of C: 89.4%: (57.5% Complete and single-copy BUSCOs, D: 31.9% duplicated BUSCOs, 9% fragmented BUSCOs, M: 1.6% missing BUSCOs, total 1,315 genes evaluated).

The pooled RNAseq reads of the *in vitro* growth conditions were mapped to the transcriptome assembly to evaluate to what extent the assembly was representative of the read data. From a total of 189,215,414 analysed reads in the combined RNAseq sample pool, 182,620,495 (96.5%) reads mapped to the assembly, while 6,594,919 (3.5%) reads did not map. Of the mapped reads, 32.9% mapped to multiple regions on the transcriptome assembly.

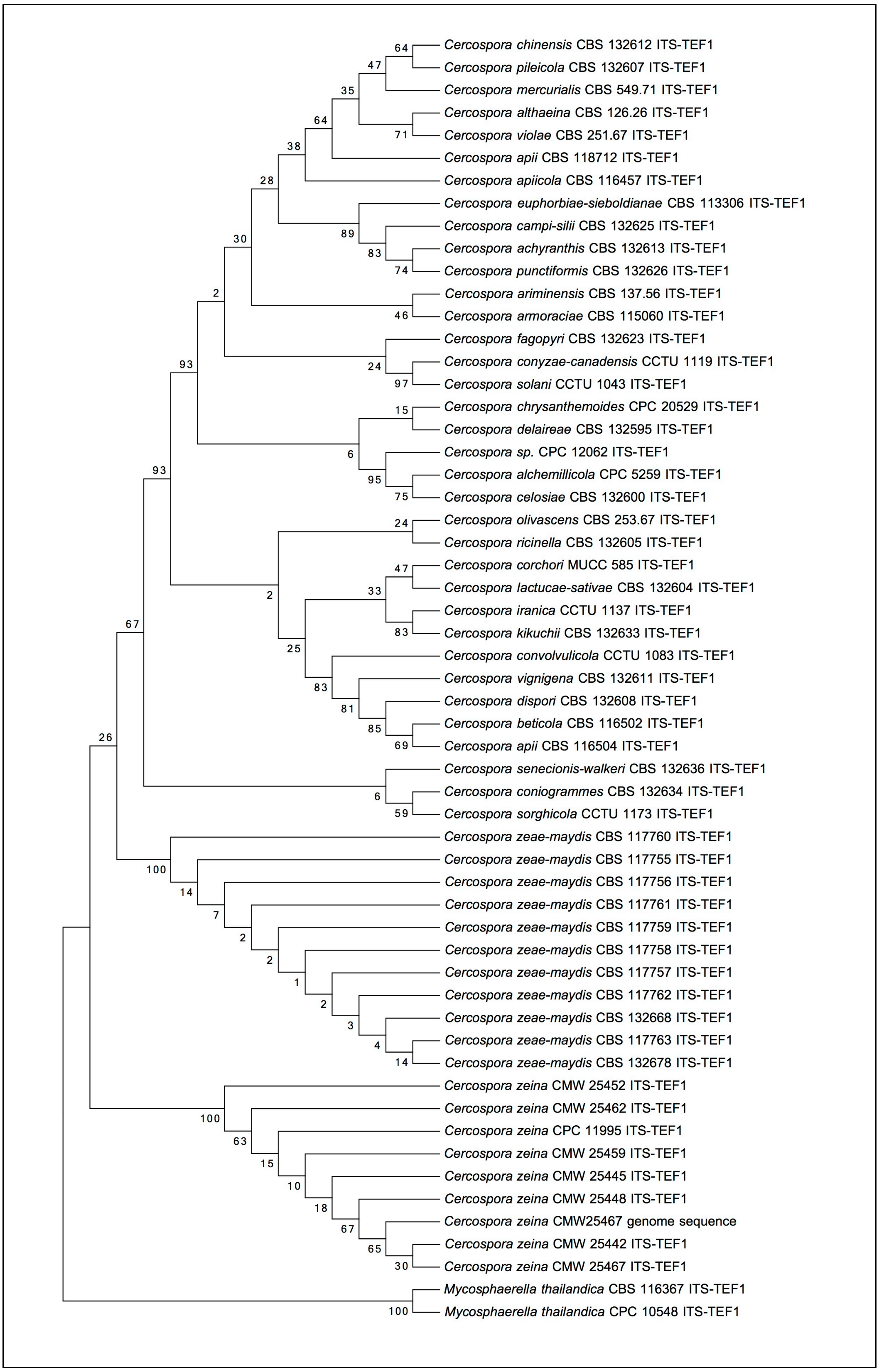

Cercospora chinensis CBS 132612 ITS-TEF1
Cercospora pileicola CBS 132607 ITS-TEF1
Cercospora mercurialis CBS 549.71 ITS-TEF1
Cercospora althaeina CBS 126.26 ITS-TEF1
Cercospora violae CBS 251.67 ITS-TEF1
Cercospora apii CBS 118712 ITS-TEF1
Cercospora apiicola CBS 116457 ITS-TEF1
Cercospora euphorbiae-sieboldianae CBS 113306 ITS-TEF1
Cercospora campi-silii CBS 132625 ITS-TEF1
Cercospora achyranthis CBS 132613 ITS-TEF1
Cercospora punctiformis CBS 132626 ITS-TEF1
Cercospora ariminensis CBS 137.56 ITS-TEF1
Cercospora armoraciae CBS 115060 ITS-TEF1
Cercospora fagopyri CBS 132623 ITS-TEF1
Cercospora conyzae-canadensis CCTU 1119 ITS-TEF1
Cercospora solani CCTU 1043 ITS-TEF1
Cercospora chrysanthemoides CPC 20529 ITS-TEF1
Cercospora delaireae CBS 132595 ITS-TEF1
Cercospora sp. CPC 12062 ITS-TEF1
Cercospora alchemillicola CPC 5259 ITS-TEF1
Cercospora celosiae CBS 132600 ITS-TEF1
Cercospora olivascens CBS 253.67 ITS-TEF1
Cercospora ricinella CBS 132605 ITS-TEF1
Cercospora corchori MUCC 585 ITS-TEF1
Cercospora lactucae-sativae CBS 132604 ITS-TEF1
Cercospora iranica CCTU 1137 ITS-TEF1
Cercospora kikuchii CBS 132633 ITS-TEF1
Cercospora convolvulicola CCTU 1083 ITS-TEF1
Cercospora vignigena CBS 132611 ITS-TEF1
Cercospora dispori CBS 132608 ITS-TEF1
Cercospora beticola CBS 116502 ITS-TEF1
Cercospora apii CBS 116504 ITS-TEF1
Cercospora senecionis-walkeri CBS 132636 ITS-TEF1
Cercospora coniogrammes CBS 132634 ITS-TEF1
Cercospora sorghicola CCTU 1173 ITS-TEF1
Cercospora zeae-maydis CBS 117760 ITS-TEF1
Cercospora zeae-maydis CBS 117755 ITS-TEF1
Cercospora zeae-maydis CBS 117756 ITS-TEF1
Cercospora zeae-maydis CBS 117761 ITS-TEF1
Cercospora zeae-maydis CBS 117759 ITS-TEF1
Cercospora zeae-maydis CBS 117758 ITS-TEF1
Cercospora zeae-maydis CBS 117757 ITS-TEF1
Cercospora zeae-maydis CBS 117762 ITS-TEF1
Cercospora zeae-maydis CBS 132668 ITS-TEF1
Cercospora zeae-maydis CBS 117763 ITS-TEF1
Cercospora zeae-maydis CBS 132678 ITS-TEF1
Cercospora zeina CMW 25452 ITS-TEF1
Cercospora zeina CMW 25462 ITS-TEF1
Cercospora zeina CPC 11995 ITS-TEF1
Cercospora zeina CMW 25459 ITS-TEF1
Cercospora zeina CMW 25445 ITS-TEF1
Cercospora zeina CMW 25448 ITS-TEF1
Cercospora zeina CMW25467 genome sequence
Cercospora zeina CMW 25442 ITS-TEF1
Cercospora zeina CMW 25467 ITS-TEF1
Mycosphaerella thailandica CBS 116367 ITS-TEF1
Mycosphaerella thailandica CPC 10548 ITS-TEF1
64
47
35
64
71
38
28
89
83
74
30
46
2
24
97
93
15
6
95
75
93
24
47
33
83
2
67
25
83
81
85
69
26
6
59
100
14
7
2
2
1
2
3
4
14
100
63
15
10
18
67
65
30
100

Figure 2.8 Maximum likelihood phylogenetic analysis of TEF1 and ITS sequences for selected *Cercospora* species and the assembled genome sequence. *The numbers on branching points indicate the percentage of replicate trees in which the associated taxa clustered together in the bootstrap test (100 replicates). The C. zeina CMW 25467 genome sequence label indicates the position of the sequences from the genome assembly.*

2.5 Discussion

The study yielded quality sequencing data from genomic and transcriptomic sequencing libraries. The genomic sequence data was used to construct a genome assembly for *C. zeina* that is transcriptionally functional and can be used for genome annotation and functional genomics studies. A transcriptome assembly was constructed from the transcriptome sequencing libraries and was shown to be representative of the transcriptome sequencing libraries. The transcriptome assembly is suitable for use in genome annotation.

A representative and transcriptionally functional genome assembly is a valuable tool for studying and understanding the functional interaction between pathogenic organisms and their hosts. However, due to the high genome coverage and resolution power possible with high-throughput sequencing, contaminating sequences will cause assemblies with unrelated sequences that might be falsely attributed to the organism studied. In a study on quality control in assembling human genomes it was found that a genome assembly from a Yoruban cell-line contained a large number of sequences from the Epstein-Barr virus used during immortalising cell-lines. Similarly, another human sample contained contaminating sequences that were subsequently incorrectly classified as novel insertional polymorphisms (Alkan *et al.*, 2011). In cases where the contaminating organism is a related species, it might not always be possible to detect this contamination and the resultant genome assembly would not be a true reflection of the genome of the species of interest. It is thus crucial that strict contamination control is maintained during the culturing and DNA isolation steps. In the case of foliar plant pathogenic fungi, it is critical to ensure that conidia are collected from the lesions associated with the target species.

The contamination of cultures is probable when it is not feasible to follow strict sterile methodology. This is especially troublesome when culturing slow-growing organisms such as *C. zeina*. The possibility exists for viral, bacterial or fungal contamination from surface contaminants or symbionts. The use of antibiotics could reduce the burden of bacterial contamination, but this will not prevent viral or fungal contamination, and here microscopic and phenotypic inspection of cultures is important. Viral contamination is problematic since it is not feasible to screen for all viruses, not least due to the large number of viral nucleic acid types, i.e. double-strand DNA, single-strand DNA, double-strand RNA, single-strand RNA (both positive and negative strand) and DNA/RNA hybrid (Mayo & Pringle, 1998). In this case there were Salmonella phage iEPS5 and Enterobacteria phage phiX174 genomes detected in the assembly, contained in one contig each, and these were subsequently removed from the assembly prior to upload to Genbank.

The phylogenetics analysis using the ITS and TEF1 sequences (Figure 2.8) confirmed that the genome sequence data is consistent with *C. zeina*, and that the correct species was sequenced. Contamination by bacteria and other fungi is common in the absence of strict sterility controls, while viruses and phages are difficult to detect in asymptomatic cultures. and A standard QC step for checking bacterial contamination would be the Polymerase Chain Reaction (PCR) amplification of the 16S region of bacteria from the isolated genomic DNA. A positive result in this amplification step would indicate the presence of bacterial DNA. The result is, of course, dependent on the use of suitable universal primers for the 16S region. In this study there was no evidence that bacterial contamination was present in the genome assembly. The detection of viral contamination of cultures is complicated (Merten, 2002), while no need for detection is evident in asymptomatic cultures. The detection of fungal DNA contamination in a fungal culture is more complicated. Sequencing the ITS region might show if contamination is present, although the similarity in ITS sequences between closely related species might be very high and could result in inconclusive results. In the present study neither PCR amplification of the 16S region, nor ITS sequencing from the isolated fungal genomic DNA were employed. The use of microscopic and phenotypic confirmation were extensive, and the use of antibiotics theoretically inhibited the growth of bacteria in all cultures. The microscopic study of fungal cultures grown on solid media to confirm phenotypic traits is one method that is reliable if there are no closely related species which might have a high physiological similarity. Due to the very close conidiophore and conidia structure morphology between *C. zeina* and *C. zeae-maydis* (Braun *et al.*, 2015), this could be a concern. The possibility existed that the incorrect *Cercospora* species (*C. zeae-maydis*) or a mixed culture was collected from the field samples. When GLS was initially characterized, the causal agents were identified as *C. zeae-maydis* type I and type II , with *C. zeae-maydis* type II subsequently re-classified as *C. zeina*. To date only two *Cercospora* species have been described infecting *Z. mays* worldwide, i.e. *C. zeina* and *C. zeae-maydis* (Crous *et al.*, 2006), with only *C. zeina* endemic to Africa (Meisel *et al.*, 2009). Though the similar lesion morphology might have led to mixed-species collection on original leaf tissue, the ITS sequence data and phylogenetic analysis indicated the genome sequence grouping in the *C. zeina* species clade with high bootstrap support (Figure 2.8).

Quality control of the respective sequencing library data showed that the MP8KB library passed the required metrics, though the PE and MP3KB libraries showed anomalies in the GC-distribution profiles. The respective quality of the read pairs in next generation paired-end sequence data can be highly variable. It is more common for the reverse pairs of paired-end reads to have decreased quality profiles due to an increase in the library template’s reagent and laser exposure, while the Illumina Real-Time Analysis (RTA) software acquisition of cluster information might impact the first 10-15 bases of reads (Guo *et al.*, 2014). Three of the quality metrics highlighted during the FastQC assessment (Figures 2.3, 2.4 and 2.5) are the most important to consider and would indicate anomalies that would affect downstream procedures to the largest extent. One metric not indicated was the level of sequence duplication. Although this metric did not pass the sequence read quality as ideal for any of the libraries, it did not provide disqualifying quality scores, and the sequence duplication levels for the respective libraries were in the same order of magnitude (data not shown). The anomalous quality metrics in the PE and MP3KB sequencing libraries were most pronounced when taking the GC-content distribution into account. This was most probably due to problems experienced during the library preparations step. The quality of the DNA sample could impact the quality to this extent, but it is not known which of the DNA isolates were used for the procedures. Due to the complexity in library construction for the MP8Kb sequencing library it was prepared separately from the others, and was most likely prepared with a different or newer reagent set which did not cause the same anomalies as the first set. Fortuitously this library also produced the largest number of sequencing reads. The sequencing service provider did not inform whether a similar library dilution and molar fraction was loaded on the flow cell. If this was the case then the problems with the library preparation would have caused problems during sequencing and led to the decrease in read numbers for the PE and MP3KB libraries when compared to the more ideal situation with the MP8KB library.

The quality of the RNAseq libraries were as expected for the high quality RNA provided to the supplier, as well as the known level of throughput and expertise from the sequence provider. As with the DNA libraries, the leading 12 bases were removed from the front of each read to negate the RTA software impact on the leading bases. While the unpaired DNA library sequence data from the Trimmomatic filtering step could be used for the assembly process, it was decided to only use the paired RNAseq data. This improved the mapping of the RNAseq data to the genome assemblies due to the known insert sizes of the libraries. The additional genome assembly quality control dimension provided by the RNAseq data mapping to the genome assembly was very important, and is an approach often used (Curtin *et al.*, 2012 ; Xia *et al.*, 2017). In addition, the RNAseq reads could also be useful for providing gene expression information between the different growth media. Since the media were chosen to emphasize or enhance certain characteristics of the fungal growth and development cycles this might shed some light on the genes responsible.

The genome assembly was optimized over three iterations, with the initial quality evaluation of each assembly based on the completeness statistics from the assemblathon script. The first assembly appeared to be very complete based on the assembly statistics. The N50 value showed that the majority of the assembly was in large contigs over 720Kbp (Table 2.8), while the size of the genome (without N) was within 7% of the predicted genome size of ~45Mbp. The RNAseq data was not available immediately following the assembly process, and the mapping of the reads to the assembly was thus not performed at fist. When the quality of the raw sequencing data was evaluated before annotation, it was found to have anomalous GC-profiles. The subsequent mapping of the RNAseq data showed that the assembly did not represent a transcriptionally functional genome. This is not a unique occurrence (Salzberg *et al.*, 2012), since one of the common problems associated with genome assembly from short read sequence data is the artificial concatenation of unrelated reads to form larger contigs. The main quality metrics used are the N50 and assembly size, thus emphasizing assembly size over contig accuracy and quality (Narzisi & Mishra, 2011), while the correctness of an assembly varies widely and is not well correlated with statistics on contiguity (Salzberg *et al.*, 2012). As a case in point a study in *Bos taurus* involved two separate assemblies with identical input data giving rise to two genomes with varying qualities and gene content (Florea *et al.*, 2011). Another study focussed on evaluating the effect of GC-bias in sequencing data, and concluded that, since higher GC-content regions are not amplified as efficiently as lower GC-regions, these regions are then under-represented in the assembly, with a deleterious effect on the assembly quality (Chen, Y. C. *et al.*, 2013). From the GC-content graphs of the PE and MP3KB libraries (Figures 2.1 and 2.1) it can be seen that the main GC-content of these libraries is ~40%, where the actual content should be ~50%, and the lower GC-reads are preferentially used. Therefore some degree of functional evaluation should be performed on genome assemblies in concert with size and completeness assessments to obtain a true measure of assembly quality. It is interesting to note that the GEGMA and BUSCO completeness assessments, involving the prediction of selected core genes, indicated a genome completion of >95%. This indicates that these genes were complete/present to such a degree that the predicted gene models conformed to a certain threshold acceptable for the two prediction tools.

The use of the MP8KB sequencing library reads to build the assembly proved successful though unorthodox, especially as the reads were treated as single reads without a paired partner. Using the paired-end reads as single reads increased the number of reads that could be used, without incorporating the large mate-pair insert-size bias into the assembly. There is a precedent since the current genome assemblers, and specifically Velvet, were initially developed to assemble the very short single-end sequence reads (32 bp) from the first Illumina sequencing runs (Zerbino & Birney, 2008). Normally the strategy would involve a large number of short-insert paired-end reads which would form the basis of the assembly, while the larger-insert mate-pair libraries would largely be used to scaffold the contigs (Li, R., Fan*, et al.*, 2010). In this case it was not possible, and the use of the MP8KB library for the assembly was required. Due to the absence of read-pair distance information it is possible that adjacent paralog genes with close sequence similarity was either mis-assembled or condensed into one gene.

The second assembly with the CLCAssembler was performed in parallel with the third Velvet assembly to evaluate the performance of the assembler with the selected data. The laboratory of Prof Van de Peer (Martin *et al.*, 2008 ; De Schutter *et al.*, 2009 ; Young *et al.*, 2011 ; Nystedt *et al.*, 2013) have assembled some genomes using this approach, but since the data composition was not ideal for the assembler, it was not expected to give a high quality assembly. Evaluation of the quality of the assembly was based solely on N50 and genome size estimations, and since the assembly did not compare well with the third Velvet assembly, optimization of the process was therefore terminated. In addition, the large increase in N50 and total assembly size after scaffolding was largely due to a large increase in the number of Ns added as placeholders to scaffold contigs (Table 2.9). In total a third of the assembly was composed of Ns and this artificially improved the assembly statistics without contributing functional value. The RNAseq data was not mapped to the assembly due to all these factors.

Velvet was the genome assembler of choice due to the extensive fungal genome assembly expertise of the collaborators (Martin *et al.*, 2008 ; De Schutter *et al.*, 2009 ; Duplessis *et al.*, 2011 ; Hacquard *et al.*, 2012 ; Roelants *et al.*, 2013 ; Delhomme *et al.*, 2015). The third assembly was not shown to be as complete as the first assembly, although the GC-composition was at ~50% where it was expected based on comparison data with other Dothideomycete genomes (Ohm *et al.*, 2012). The scaffolding process increased the N50 and total genome size appreciably, and the number of Ns was much lower than for the CLC assembly at ~11%, which was similar to the first assembly (Tables 2.9 and 2.11). The total genome size and largest scaffold size did decrease following gap filling, but this was due to removal of less reliable reads on the ends of contigs before filling gaps. It did have the effect of increasing the N50, and thus increased the size of most of the contigs in the assembly. A concern was the skewed insert-size of the MP3KB reads used to assist in scaffolding. Based on the number of reads contributing to the scaffolding, five times more reads for the MP8KB library were used in the process when compared to the MP3KB library. Therefore it might be assumed, without experimental support, that the MP3KB reads used for the scaffolding are the reads with the correct insert-size distribution. This might have been a concern were it not for the high number of RNAseq reads mapping to the assembly. This indicated that the assembly is transcriptionally functional, and could be used for the annotation process. Although the assembly is less complete based on the CEGMA and BUSCO predictions, it is acceptable for functional purposes.

The RNAseq reads were important for evaluating the quality of the genome assembly in terms of the transcriptional functionality. Subsequently the reads were assembled to obtain a transcriptome assembly. It is difficult to estimate the completeness and quality of a transcriptome assembly, since the genes present are a function of the state of gene transcription activity in the cell at the time of sampling, and as such it is highly unlikely that all genes in the genome will be represented in the transcriptome. Previously it was reported that standard genome assembly parameters, such as N50 value, were not sufficient for assessing transcriptome assembly quality. Functional methods are more informative, including mapping reads back to the assembly and similarity searches for core genes (Garg *et al.*, 2011). The *C. zeina* transcriptome shows a N50 value that is larger than other fungal transcriptome assemblies (Chen, Y. *et al.*, 2017 ; Giosa *et al.*, 2017), and might be indicative of the incorrect concatenation of contigs due to the close proximity of genes in fungal genomes (Galagan *et al.*, 2005). The size of the assembly is larger than the genome for the organism, and this is most probably linked to the presence of splice variants for a large number of genes which are all accounted for in the transcriptome (Honaas *et al.*, 2016). The BUSCO report does shows many duplicated genes just in this dataset (420 of 1315) and therefore we can assume a similar proportion for the rest of the genes. In total 89.4% of the BUSCO genes are predicted to be present. The RNAseq reads also map back to the transcriptome to a very high degree, indicating that the assembly process did not assemble the data into distorted and non-functional genes. Analysis of the transcriptome assembly composition revealed significant differences in GC over AT content in the transcriptome, with the transcriptome containing 54.76% guanine and cytosine (genome GC content 49.7%), and only 45.24% adenine and thymine (genome AT content 50.3%). This is consistent with other studies, including the human genome, where coding exons also show elevated GC-content (Louie *et al.*, 2003). In the following chapter the transcriptome assembly was utilized during the genome annotation process.

The *C. zeina* genome assembly constructed in this study is transcriptionally functional, has a majority of sequence bases in larger contigs, contains the majority of *Ascomycota* core genes and can be considered sufficiently complete for gene prediction and annotation procedures in the following chapter.

2.6 References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., George, R. A., Lewis, S. E., Richards, S., Ashburner, M., Henderson, S. N., Sutton, G. G., Wortman, J. R., Yandell, M. D., Zhang, Q., Chen, L. X., Brandon, R. C., Rogers, Y. H., Blazej, R. G., Champe, M., Pfeiffer, B. D., Wan, K. H., Doyle, C., Baxter, E. G., Helt, G., Nelson, C. R., Gabor, G. L., Abril, J. F., Agbayani, A., An, H. J., Andrews-Pfannkoch, C., Baldwin, D., Ballew, R. M., Basu, A., Baxendale, J., Bayraktaroglu, L., Beasley, E. M., Beeson, K. Y., Benos, P. V., Berman, B. P., Bhandari, D., Bolshakov, S., Borkova, D., Botchan, M. R., Bouck, J., Brokstein, P., Brottier, P., Burtis, K. C., Busam, D. A., Butler, H., Cadieu, E., Center, A., Chandra, I., Cherry, J. M., Cawley, S., Dahlke, C., Davenport, L. B., Davies, P., de Pablos, B., Delcher, A., Deng, Z., Mays, A. D., Dew, I., Dietz, S. M., Dodson, K., Doup, L. E., Downes, M., Dugan-Rocha, S., Dunkov, B. C., Dunn, P., Durbin, K. J., Evangelista, C. C., Ferraz, C., Ferriera, S., Fleischmann, W., Fosler, C., Gabrielian, A. E., Garg, N. S., Gelbart, W. M., Glasser, K., Glodek, A., Gong, F., Gorrell, J. H., Gu, Z., Guan, P., Harris, M., Harris, N. L., Harvey, D., Heiman, T. J., Hernandez, J. R., Houck, J., Hostin, D., Houston, K. A., Howland, T. J., Wei, M. H., Ibegwam, C., Jalali, M., Kalush, F., Karpen, G. H., Ke, Z., Kennison, J. A., Ketchum, K. A., Kimmel, B. E., Kodira, C. D., Kraft, C., Kravitz, S., Kulp, D., Lai, Z., Lasko, P., Lei, Y., Levitsky, A. A., Li, J., Li, Z., Liang, Y., Lin, X., Liu, X., Mattei, B., McIntosh, T. C., McLeod, M. P., McPherson, D., Merkulov, G., Milshina, N. V., Mobarry, C., Morris, J., Moshrefi, A., Mount, S. M., Moy, M., Murphy, B., Murphy, L., Muzny, D. M., Nelson, D. L., Nelson, D. R., Nelson, K. A., Nixon, K., Nusskern, D. R., Pacleb, J. M., Palazzolo, M., Pittman, G. S., Pan, S., Pollard, J., Puri, V., Reese, M. G., Reinert, K., Remington, K., Saunders, R. D., Scheeler, F., Shen, H., Shue, B. C., Siden-Kiamos, I., Simpson, M., Skupski, M. P., Smith, T., Spier, E., Spradling, A. C., Stapleton, M., Strong, R., Sun, E., Svirskas, R., Tector, C., Turner, R., Venter, E., Wang, A. H., Wang, X., Wang, Z. Y., Wassarman, D. A., Weinstock, G. M., Weissenbach, J., Williams, S. M., WoodageT, Worley, K. C., Wu, D., Yang, S., Yao, Q. A., Ye, J., Yeh, R. F., Zaveri, J. S., Zhan, M., Zhang, G., Zhao, Q., Zheng, L., Zheng, X. H., Zhong, F. N., Zhong, W., Zhou, X., Zhu, S., Zhu, X., Smith, H. O., Gibbs, R. A., Myers, E. W., Rubin, G. M., and Venter, J. C. (2000) The genome sequence of *Drosophila melanogaster*. *Science* **287(5461)**:2185-2195

Alkan, C., Sajjadian, S., and Eichler, E. E. (2011) Limitations of next-generation genome sequence assembly. *Nature Methods* **8(1)**:61-65

Batzoglou, S., Jaffe, D. B., Stanley, K., Butler, J., Gnerre, S., Mauceli, E., Berger, B., Mesirov, J. P., and Lander, E. S. (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Research* **12(1)**:177-89

Boetzer, M., and Pirovano, W. (2012) Toward almost closed genomes with GapFiller. *Genome Biology* **13(6)**:R56

Boetzer, M., Henkel, C. V., Jansen, H. J., Butler, D., and Pirovano, W. (2011) Scaffolding pre-assembled contigs using SSPACE. *Bioinformatics* **27(4)**:578-579

Bolger, A. M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30(15)**:2114-2120

Bradnam, K. R., Fass, J. N., Alexandrov, A., Baranay, P., Bechner, M., Birol, I., Boisvert, S., Chapman, J. A., Chapuis, G., Chikhi, R., Chitsaz, H., Chou, W. C., Corbeil, J., Del Fabbro, C., Docking, T. R., Durbin, R., Earl, D., Emrich, S., Fedotov, P., Fonseca, N. A., Ganapathy, G., Gibbs, R. A., Gnerre, S., Godzaridis, E., Goldstein, S., Haimel, M., Hall, G., Haussler, D., Hiatt, J. B., Ho, I. Y., Howard, J., Hunt, M., Jackman, S. D., Jaffe, D. B., Jarvis, E. D., Jiang, H., Kazakov, S., Kersey, P. J., Kitzman, J. O., Knight, J. R., Koren, S., Lam, T. W., Lavenier, D., Laviolette, F., Li, Y., Li, Z., Liu, B., Liu, Y., Luo, R., Maccallum, I., Macmanes, M. D., Maillet, N., Melnikov, S., Naquin, D., Ning, Z., Otto, T. D., Paten, B., Paulo, O. S., Phillippy, A. M., Pina-Martins, F., Place, M., Przybylski, D., Qin, X., Qu, C., Ribeiro, F. J., Richards, S., Rokhsar, D. S., Ruby, J. G., Scalabrin, S., Schatz, M. C., Schwartz, D. C., Sergushichev, A., Sharpe, T., Shaw, T. I., Shendure, J., Shi, Y., Simpson, J. T., Song, H., Tsarev, F., Vezzi, F., Vicedomini, R., Vieira, B. M., Wang, J., Worley, K. C., Yin, S., Yiu, S. M., Yuan, J., Zhang, G., Zhang, H., Zhou, S., and Korf, I. F. (2013) Assemblathon 2: evaluating *de novo* methods of genome assembly in three vertebrate species. *Gigascience* **2(1)**:10

Braun, U., Crous, P. W., and Nakashima, C. (2015) Cercosporoid fungi (*Mycosphaerellaceae*) 3. Species on monocots (*Poaceae*, true grasses). *IMA Fungus* **6(1)**:25-97

Broad. http://broadinstitute.github.io/picard/.

Chang, Z., Li, G., Liu, J., Zhang, Y., Ashby, C., Liu, D., Cramer, C. L., and Huang, X. (2015) Bridger: a new framework for de novo transcriptome assembly using RNA-seq data. *Genome Biology* **16**:30

Chen, Y., Cao, Q., Tao, X., Shao, H., Zhang, K., Zhang, Y., and Tan, X. (2017) Analysis of *de novo* sequencing and transcriptome assembly and lignocellulolytic enzymes gene expression of *Coriolopsis gallica* HTC. *Bioscience, Biotechnology, and Biochemistry* **81(3)**:460-468

Chen, Y. C., Liu, T., Yu, C. H., Chiang, T. Y., and Hwang, C. C. (2013) Effects of GC bias in next-generation-sequencing data on *de novo* genome assembly. *PLoS ONE* **8(4)**:e62856

Crous, P., Groenewald, J., Groenewald, M., Caldwell, P., Braun, U., and Harrington, T. (2006) Species of *Cercospora* associated with grey leaf spot of maize. *Studies in Mycology* **55**:189-197

Curtin, C. D., Borneman, A. R., Chambers, P. J., and Pretorius, I. S. (2012) *De novo* assembly and analysis of the heterozygous triploid genome of the wine spoilage yeast *Dekkera bruxellensis* AWRI1499. *PLoS ONE* **7(3)**:e33840

De Schutter, K., Lin, Y. C., Tiels, P., Van Hecke, A., Glinka, S., Weber-Lehmann, J., Rouze, P., Van de Peer, Y., and Callewaert, N. (2009) Genome sequence of the recombinant protein production host *Pichia pastoris*. *Nature Biotechnology* **27(6)**:561-566

Delhomme, N., Sundstrom, G., Zamani, N., Lantz, H., Lin, Y. C., Hvidsten, T. R., Hoppner, M. P., Jern, P., Van de Peer, Y., Lundeberg, J., Grabherr, M. G., and Street, N. R. (2015) Serendipitous meta-transcriptomics: The fungal community of Norway spruce (*Picea abies*). *PLoS ONE* **10(9)**:e0139080

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., and Gingeras, T. R. (2013) STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29(1)**:15-21

Duplessis, S., Cuomo, C. A., Lin, Y. C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., Joly, D. L., Hacquard, S., Amselem, J., Cantarel, B. L., Chiu, R., Coutinho, P. M., Feau, N., Field, M., Frey, P., Gelhaye, E., Goldberg, J., Grabherr, M. G., Kodira, C. D., Kohler, A., Kues, U., Lindquist, E. A., Lucas, S. M., Mago, R., Mauceli, E., Morin, E., Murat, C., Pangilinan, J. L., Park, R., Pearson, M., Quesneville, H., Rouhier, N., Sakthikumar, S., Salamov, A. A., Schmutz, J., Selles, B., Shapiro, H., Tanguay, P., Tuskan, G. A., Henrissat, B., Van de Peer, Y., Rouze, P., Ellis, J. G., Dodds, P. N., Schein, J. E., Zhong, S., Hamelin, R. C., Grigoriev, I. V., Szabo, L. J., and Martin, F. (2011) Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proceedings of the National Academy of Sciences of the United States of America* **108(22)**:9166-9171

Earl, D., Bradnam, K., St John, J., Darling, A., Lin, D., Fass, J., Yu, H. O., Buffalo, V., Zerbino, D. R., Diekhans, M., Nguyen, N., Ariyaratne, P. N., Sung, W. K., Ning, Z., Haimel, M., Simpson, J. T., Fonseca, N. A., Birol, I., Docking, T. R., Ho, I. Y., Rokhsar, D. S., Chikhi, R., Lavenier, D., Chapuis, G., Naquin, D., Maillet, N., Schatz, M. C., Kelley, D. R., Phillippy, A. M., Koren, S., Yang, S. P., Wu, W., Chou, W. C., Srivastava, A., Shaw, T. I., Ruby, J. G., Skewes-Cox, P., Betegon, M., Dimon, M. T., Solovyev, V., Seledtsov, I., Kosarev, P., Vorobyev, D., Ramirez-Gonzalez, R., Leggett, R., MacLean, D., Xia, F., Luo, R., Li, Z., Xie, Y., Liu, B., Gnerre, S., MacCallum, I., Przybylski, D., Ribeiro, F. J., Yin, S., Sharpe, T., Hall, G., Kersey, P. J., Durbin, R., Jackman, S. D., Chapman, J. A., Huang, X., DeRisi, J. L., Caccamo, M., Li, Y., Jaffe, D. B., Green, R. E., Haussler, D., Korf, I., and Paten, B. (2011) Assemblathon 1: a competitive assessment of de novo short read assembly methods. *Genome Research* **21(12)**:2224-2241

Engstrom, P. G., Steijger, T., Sipos, B., Grant, G. R., Kahles, A., Ratsch, G., Goldman, N., Hubbard, T. J., Harrow, J., Guigo, R., Bertone, P., and Consortium, R.

(2013) Systematic evaluation of spliced alignment programs for RNA-seq data. *Nature Methods* **10(12)**:1185-1191

Felsenstein, J. (1985) Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39(4)**:783-791

Florea, L., Souvorov, A., Kalbfleisch, T. S., and Salzberg, S. L. (2011) Genome assembly has a major impact on gene content: a comparison of annotation in two *Bos taurus* assemblies. *PLoS ONE* **6(6)**:e21400

Galagan, J. E., Henn, M. R., Ma, L. J., Cuomo, C. A., and Birren, B. (2005) Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Research* **15(12)**:1620-1631.

Garg, R., Patel, R. K., Tyagi, A. K., and Jain, M. (2011) *De novo* assembly of chickpea transcriptome using short reads for gene discovery and marker identification. *DNA Res.* **18(1)**:53-63

Giosa, D., Felice, M. R., Lawrence, T. J., Gulati, M., Scordino, F., Giuffre, L., Lo Passo, C., D'Alessandro, E., Criseo, G., Ardell, D. H., Hernday, A. D., Nobile, C. J., and Romeo, O. (2017) Whole RNA-sequencing and transcriptome assembly of *Candida albicans* and *Candida africana* under chlamydospore-inducing conditions. *Genome Biology and Evolution* **9(7)**:1971-1977

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., di Palma, F., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., Friedman, N., and Regev, A. (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology* **29(7)**:644-52

Gruenheit, N., Deusch, O., Esser, C., Becker, M., Voelckel, C., and Lockhart, P. (2012) Cutoffs and k-mers: implications from a transcriptome study in allopolyploid plants. *BMC Genomics* **13**:92

Guo, Y., Ye, F., Sheng, Q., Clark, T., and Samuels, D. C. (2014) Three-stage quality control strategies for DNA re-sequencing data. *Briefings in Bioinformatics* **15(6)**:879-889

Guttman, M., Garber, M., Levin, J. Z., Donaghey, J., Robinson, J., Adiconis, X., Fan, L., Koziol, M. J., Gnirke, A., Nusbaum, C., Rinn, J. L., Lander, E. S., and Regev, A. (2010) *Ab initio* reconstruction of cell type-specific transcriptomes in mouse reveals the conserved multi-exonic structure of lincRNAs. *Nature Biotechnology* **28(5)**:503-510

Haas, B. J., Papanicolaou, A., Yassour, M., Grabherr, M., Blood, P. D., Bowden, J., Couger, M. B., Eccles, D., Li, B., Lieber, M., MacManes, M. D., Ott, M., Orvis, J., Pochet, N., Strozzi, F., Weeks, N., Westerman, R., William, T., Dewey, C. N., Henschel, R., LeDuc, R. D., Friedman, N., and Regev, A. (2013) *De novo* transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nature Protocols* **8(8)**:1494-1512

Hacquard, S., Joly, D. L., Lin, Y. C., Tisserant, E., Feau, N., Delaruelle, C., Legue, V., Kohler, A., Tanguay, P., Petre, B., Frey, P., Van de Peer, Y., Rouze, P., Martin, F., Hamelin, R. C., and Duplessis, S. (2012) A comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (poplar leaf rust). *Molecular Plant-Microbe Interactions* **25(3)**:279-293

Hernandez, D., Francois, P., Farinelli, L., Osteras, M., and Schrenzel, J. (2008) *De novo* bacterial genome sequencing: millions of very short reads assembled on a desktop computer. *Genome Research* **18(5)**:802-809

Homer, N., Merriman, B., and Nelson, S. F. (2009) BFAST: an alignment tool for large scale genome resequencing. *PLoS ONE* **4(11)**:e7767

Honaas, L. A., Wafula, E. K., Wickett, N. J., Der, J. P., Zhang, Y., Edger, P. P., Altman, N. S., Pires, J. C., Leebens-Mack, J. H., and dePamphilis, C. W. (2016) Selecting superior *de novo* transcriptome assemblies: lessons learned by leveraging the best plant genome. *PLoS ONE* **11(1)**:e0146062

Huang, X., and Madan, A. (1999) CAP3: A DNA sequence assembly program. *Genome Research* **9(9)**:868-877.

Ji, Y., Shi, Y., Ding, G., and Li, Y. (2011) A new strategy for better genome assembly from very short reads. *BMC Bioinformatics* **12**:493

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14(4)**:R36

Kumar, S., Stecher, G., and Tamura, K. (2016) MEGA7: Molecular Evolutionary Genetics Analysis version 7.0 for bigger datasets. *Molecular Biology and Evolution* **33(7)**:1870-1874

Langdon, W. B. (2015) Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Mining* **8(1)**:1

Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., and Dewey, C. N. (2010) RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics* **26(4)**:493-500

Li, H., and Durbin, R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25(14)**:1754-1760

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., and Genome Project Data Processing, S. (2009) The Sequence Alignment Map format and SAMtools. *Bioinformatics* **25(16)**:2078-2079

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., Li, Y., Li, S., Shan, G., Kristiansen, K., Li, S., Yang, H., Wang, J., and Wang, J. (2010) *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research* **20(2)**:265-272

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., Huang, Q., Cai, Q., Li, B., Bai, Y., Zhang, Z., Zhang, Y., Wang, W., Li, J., Wei, F., Li, H., Jian, M., Li, J., Zhang, Z., Nielsen, R., Li, D., Gu, W., Yang, Z., Xuan, Z., Ryder, O. A., Leung, F. C., Zhou, Y., Cao, J., Sun, X., Fu, Y., Fang, X., Guo, X., Wang, B., Hou, R., Shen, F., Mu, B., Ni, P., Lin, R., Qian, W., Wang, G., Yu, C., Nie, W., Wang, J., Wu, Z., Liang, H., Min, J., Wu, Q., Cheng, S., Ruan, J., Wang, M., Shi, Z., Wen, M., Liu, B., Ren, X., Zheng, H., Dong, D., Cook, K., Shan, G., Zhang, H., Kosiol, C., Xie, X., Lu, Z., Zheng, H., Li, Y., Steiner, C. C., Lam, T. T., Lin, S., Zhang, Q., Li, G., Tian, J., Gong, T., Liu, H., Zhang, D., Fang, L., Ye, C., Zhang, J., Hu, W., Xu, A., Ren, Y., Zhang, G., Bruford, M. W., Li, Q., Ma, L., Guo, Y., An, N., Hu, Y., Zheng, Y., Shi, Y., Li, Z., Liu, Q., Chen, Y., Zhao, J., Qu, N., Zhao, S., Tian, F., Wang, X., Wang, H., Xu, L., Liu, X., Vinar, T., Wang, Y., Lam, T. W., Yiu, S. M., Liu, S., Zhang, H., Li, D., Huang, Y., Wang, X., Yang, G., Jiang, Z., Wang, J., Qin, N., Li, L., Li, J., Bolund, L., Kristiansen, K., Wong, G. K., Olson, M., Zhang, X., Li, S., Yang, H., Wang, J., and Wang, J. (2010) The sequence and *de novo* assembly of the giant panda genome. *Nature* **463(7279)**:311-317

Liao, Y., Smyth, G. K., and Shi, W. (2013) The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote. *Nucleic Acids Research* **41(10)**:e108

Louie, E., Ott, J., and Majewski, J. (2003) Nucleotide frequency variation across human genes. *Genome Research* **13(12)**:2594-2601

Ma, L.-J., van der Does, H. C., Borkovich, K. A., Coleman, J. J., Daboussi, M.-J., Di Pietro, A., Dufresne, M., Freitag, M., Grabherr, M., Henrissat, B., Houterman, P. M., Kang, S., Shim, W.-B., Woloshuk, C., Xie, X., Xu, J.-R., Antoniw, J., Baker, S. E., Bluhm, B. H., Breakspear, A., Brown, D. W., Butchko, R. A. E., Chapman, S., Coulson, R., Coutinho, P. M., Danchin, E. G. J., Diener, A., Gale, L. R., Gardiner, D. M., Goff, S., Hammond-Kosack, K. E., Hilburn, K., Hua-Van, A., Jonkers, W., Kazan, K., Kodira, C. D., Koehrsen, M., Kumar, L., Lee, Y.-H., Li, L., Manners, J. M., Miranda-Saavedra, D., Mukherjee, M., Park, G., Park, J., Park, S.-Y., Proctor, R. H., Regev, A., Ruiz-Roldan, M. C., Sain, D., Sakthikumar, S., Sykes, S., Schwartz, D. C., Turgeon, B. G., Wapinski, I., Yoder, O., Young, S., Zeng, Q., Zhou, S., Galagan, J., Cuomo, C. A., Kistler, H. C., and Rep, M. (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature* **464**:367-373

Marioni, J. C., Mason, C. E., Mane, S. M., Stephens, M., and Gilad, Y. (2008) RNA-seq: an assessment of technical reproducibility and comparison with gene expression arrays. *Genome Research* **18(9)**:1509-1517

Martin, F., Aerts, A., Ahren, D., Brun, A., Danchin, E. G., Duchaussoy, F., Gibon, J., Kohler, A., Lindquist, E., Pereda, V., Salamov, A., Shapiro, H. J., Wuyts, J., Blaudez, D., Buee, M., Brokstein, P., Canback, B., Cohen, D., Courty, P. E., Coutinho, P. M., Delaruelle, C., Detter, J. C., Deveau, A., DiFazio, S., Duplessis, S., Fraissinet-Tachet, L., Lucic, E., Frey-Klett, P., Fourrey, C., Feussner, I., Gay, G., Grimwood, J., Hoegger, P. J., Jain, P., Kilaru, S., Labbe, J., Lin, Y. C., Legue, V., Le Tacon, F., Marmeisse, R., Melayah, D., Montanini, B., Muratet, M., Nehls, U., Niculita-Hirzel, H., Oudot-Le Secq, M. P., Peter, M., Quesneville, H., Rajashekar, B., Reich, M., Rouhier, N., Schmutz, J., Yin, T., Chalot, M., Henrissat, B., Kues, U., Lucas, S., Van de Peer, Y., Podila, G. K., Polle, A., Pukkila, P. J., Richardson, P. M., Rouze, P., Sanders, I. R., Stajich, J. E., Tunlid, A., Tuskan, G., and Grigoriev, I. V. (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature.* **452(7183)**:88-92

Mayo, M. A., and Pringle, C. R. (1998) Virus taxonomy -1997. *Journal of General Virology* **79 (Pt 4)**:649-657

Meisel, B., Korsman, J., Kloppers, F., and Berger, D. (2009) *Cercospora zeina* is the causal agent of grey leaf spot disease of maize in southern Africa. *European Journal of Plant Pathology* **124**:577-583

Merten, O. W. (2002) Virus contaminations of cell cultures - A biotechnological view. *Cytotechnology* **39(2)**:91-116

Miller, J. R., Koren, S., and Sutton, G. (2010) Assembly algorithms for next-generation sequencing data. *Genomics.* **95(6)**:315-27

Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., and Wold, B. (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. *Nature Methods* **5(7)**:621-628

Myers, E. W., Sutton, G. G., Delcher, A. L., Dew, I. M., Fasulo, D. P., Flanigan, M. J., Kravitz, S. A., Mobarry, C. M., Reinert, K. H., Remington, K. A., Anson, E. L., Bolanos, R. A., Chou, H. H., Jordan, C. M., Halpern, A. L., Lonardi, S., Beasley, E. M., Brandon, R. C., Chen, L., Dunn, P. J., Lai, Z., Liang, Y., Nusskern, D. R., Zhan, M., Zhang, Q., Zheng, X., Rubin, G. M., Adams, M. D., and Venter, J. C. (2000) A whole-genome assembly of *Drosophila*. *Science* **287(5461)**:2196-2204

Narzisi, G., and Mishra, B. (2011) Comparing *de novo* genome assembly: the long and short of it. *PLoS ONE* **6(4)**:e19175

Nystedt, B., Street, N. R., Wetterbom, A., Zuccolo, A., Lin, Y. C., Scofield, D. G., Vezzi, F., Delhomme, N., Giacomello, S., Alexeyenko, A., Vicedomini, R., Sahlin, K., Sherwood, E., Elfstrand, M., Gramzow, L., Holmberg, K., Hallman, J., Keech, O., Klasson, L., Koriabine, M., Kucukoglu, M., Kaller, M., Luthman, J., Lysholm, F., Niittyla, T., Olson, A., Rilakovic, N., Ritland, C., Rossello, J. A., Sena, J., Svensson, T., Talavera-Lopez, C., Theissen, G., Tuominen, H., Vanneste, K., Wu, Z. Q., Zhang, B., Zerbe, P., Arvestad, L., Bhalerao, R., Bohlmann, J., Bousquet, J., Garcia Gil, R., Hvidsten, T. R., de Jong, P., MacKay, J., Morgante, M., Ritland, K., Sundberg, B., Thompson, S. L., Van de Peer, Y., Andersson, B., Nilsson, O., Ingvarsson, P. K., Lundeberg, J., and Jansson, S. (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature* **497(7451)**:579-584

Ohm, R. A., Feau, N., Henrissat, B., Schoch, C. L., Horwitz, B. A., Barry, K. W., Condon, B. J., Copeland, A. C., Dhillon, B., Glaser, F., Hesse, C. N., Kosti, I., LaButti, K., Lindquist, E. A., Lucas, S., Salamov, A. A., Bradshaw, R. E., Ciuffetti, L., Hamelin, R. C., Kema, G. H., Lawrence, C., Scott, J. A., Spatafora, J. W., Turgeon, B. G., de Wit, P. J., Zhong, S., Goodwin, S. B., and Grigoriev, I. V. (2012) Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PLoS Pathogens* **8(12)**:e1003037

Otto, T. D. (2015) From sequence mapping to genome assemblies. *Methods in Molecular Biology* **1201**:19-50

Parra, G., Bradnam, K., and Korf, I. (2007) CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics.* **23(9)**:1061-1067

Pevzner, P. A., Tang, H., and Waterman, M. S. (2001) An Eulerian path approach to DNA fragment assembly. *Proceedings of the National Academy of Sciences of the United States of America* **98(17)**:9748-9753

Pop, M. (2009) Genome assembly reborn: recent computational challenges. *Briefings in Bioinformatics* **10(4)**:354-366

Pop, M., and Salzberg, S. L. (2008) Bioinformatics challenges of new sequencing technology. *Trends in Genetics* **24(3)**:142-149

Roelants, S. L., Saerens, K. M., Derycke, T., Li, B., Lin, Y. C., Van de Peer, Y., De Maeseneire, S. L., Van Bogaert, I. N., and Soetaert, W. (2013) *Candida bombicola* as a platform organism for the production of tailor-made biomolecules. *Biotechnology and Bioengineering* **110(9)**:2494-503

Salzberg, S. L., Phillippy, A. M., Zimin, A., Puiu, D., Magoc, T., Koren, S., Treangen, T. J., Schatz, M. C., Delcher, A. L., Roberts, M., Marcais, G., Pop, M., and Yorke, J. A. (2012) GAGE: A critical evaluation of genome assemblies and assembly algorithms. *Genome Research* **22(3)**:557-567

Schbath, S., Martin, V., Zytnicki, M., Fayolle, J., Loux, V., and Gibrat, J. F. (2012) Mapping reads on a genomic sequence: an algorithmic overview and a practical comparative analysis. *Journal of Computational Biology* **19(6)**:796-813

Schulz, M. H., Zerbino, D. R., Vingron, M., and Birney, E. (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. *Bioinformatics* **28(8)**:1086-1092

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31(19)**:3210-3212

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, I. (2009) ABySS: a parallel assembler for short read sequence data. *Genome Research* **19(6)**:1117-1123

Sutton, G. G., White, O., Adams, M. D., and R., K. A. (1995) TIGR Assembler: A new tool for assembling large shotgun sequencing projects. *Genome Science and Technology* **1**:9-19

Swart, V. (2017) Functional genomics of the cercosporin biosynthetic gene cluster in the maize pathogen *Cercospora zeina*. *PhD Thesis, University of Pretoria, Pretoria, South Africa*

Tamura, K. (1992) Estimation of the number of nucleotide substitutions when there are strong transition-transversion and G+C-content biases. *Molecular Biology and Evolution* **9(4)**:678-687

Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research* **22(22)**:4673-4680

Trapnell, C., and Salzberg, S. L. (2010) How to map billions of short reads onto genomes. *Nature Biotechnology* **27(5)**:455-457

Trapnell, C., Williams, B. A., Pertea, G., Mortazavi, A., Kwan, G., van Baren, M. J., Salzberg, S. L., Wold, B. J., and Pachter, L. (2010) Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nature Biotechnology* **28(5)**:511-515

Treangen, T. J., Sommer, D. D., Angly, F. E., Koren, S., and Pop, M. (2011) Next generation sequence assembly with AMOS. *Current Protocols in Bioinformatics.* **Chapter(11)**:Unit 11.8

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., Gocayne, J. D., Amanatides, P., Ballew, R. M., Huson, D. H., Wortman, J. R., Zhang, Q., Kodira, C. D., Zheng, X. H., Chen, L., Skupski, M., Subramanian, G., Thomas, P. D., Zhang, J., Gabor Miklos, G. L., Nelson, C., Broder, S., Clark, A. G., Nadeau, J., McKusick, V. A., Zinder, N., Levine, A. J., Roberts, R. J., Simon, M., Slayman, C., Hunkapiller, M., Bolanos, R., Delcher, A., Dew, I., Fasulo, D., Flanigan, M., Florea, L., Halpern, A., Hannenhalli, S., Kravitz, S., Levy, S., Mobarry, C., Reinert, K., Remington, K., Abu-Threideh, J., Beasley, E., Biddick, K., Bonazzi, V., Brandon, R., Cargill, M., Chandramouliswaran, I., Charlab, R., Chaturvedi, K., Deng, Z., Di Francesco, V., Dunn, P., Eilbeck, K., Evangelista, C., Gabrielian, A. E., Gan, W., Ge, W., Gong, F., Gu, Z., Guan, P., Heiman, T. J., Higgins, M. E., Ji, R. R., Ke, Z., Ketchum, K. A., Lai, Z., Lei, Y., Li, Z., Li, J., Liang, Y., Lin, X., Lu, F., Merkulov, G. V., Milshina, N., Moore, H. M., Naik, A. K., Narayan, V. A., Neelam, B., Nusskern, D., Rusch, D. B., Salzberg, S., Shao, W., Shue, B., Sun, J., Wang, Z., Wang, A., Wang, X., Wang, J., Wei, M., Wides, R., Xiao, C., Yan, C., Yao, A., Ye, J., Zhan, M., Zhang, W., Zhang, H., Zhao, Q., Zheng, L., Zhong, F., Zhong, W., Zhu, S., Zhao, S., Gilbert, D., Baumhueter, S., Spier, G., Carter, C., Cravchik, A., Woodage, T., Ali, F., An, H., Awe, A., Baldwin, D., Baden, H., Barnstead, M., Barrow, I., Beeson, K., Busam, D., Carver, A., Center, A., Cheng, M. L., Curry, L., Danaher, S., Davenport, L., Desilets, R., Dietz, S., Dodson, K., Doup, L., Ferriera, S., Garg, N., Gluecksmann, A., Hart, B., Haynes, J., Haynes, C., Heiner, C., Hladun, S., Hostin, D., Houck, J., Howland, T., Ibegwam, C., Johnson, J., Kalush, F., Kline, L., Koduru, S., Love, A., Mann, F., May, D., McCawley, S., McIntosh, T., McMullen, I., Moy, M., Moy, L., Murphy, B., Nelson, K., Pfannkoch, C., Pratts, E., Puri, V., Qureshi, H., Reardon, M., Rodriguez, R., Rogers, Y. H., Romblad, D., Ruhfel, B., Scott, R., Sitter, C., Smallwood, M., Stewart, E., Strong, R., Suh, E., Thomas, R., Tint, N. N., Tse, S., Vech, C., Wang, G., Wetter, J., Williams, S., Williams, M., Windsor, S., Winn-Deen, E., Wolfe, K., Zaveri, J., Zaveri, K., Abril, J. F., Guigo, R., Campbell, M. J., Sjolander, K. V., Karlak, B., Kejariwal, A., Mi, H., Lazareva, B., Hatton, T., Narechania, A., Diemer, K., Muruganujan, A., Guo, N., Sato, S., Bafna, V., Istrail, S., Lippert, R., Schwartz, R., Walenz, B., Yooseph, S., Allen, D., Basu, A., Baxendale, J., Blick, L., Caminha, M., Carnes-Stine, J., Caulk, P., Chiang, Y. H., Coyne, M., Dahlke, C., Mays, A., Dombroski, M., Donnelly, M., Ely, D., Esparham, S., Fosler, C., Gire, H., Glanowski, S., Glasser, K., Glodek, A., Gorokhov, M., Graham, K., Gropman, B., Harris, M., Heil, J., Henderson, S., Hoover, J., Jennings, D., Jordan, C., Jordan, J., Kasha, J., Kagan, L., Kraft, C., Levitsky, A., Lewis, M., Liu, X., Lopez, J., Ma, D., Majoros, W., McDaniel, J., Murphy, S., Newman, M., Nguyen, T., Nguyen, N., Nodell, M., Pan, S., Peck, J., Peterson, M., Rowe, W., Sanders, R., Scott, J., Simpson, M., Smith, T., Sprague, A., Stockwell, T., Turner, R., Venter, E., Wang, M., Wen, M., Wu, D., Wu, M., Xia, A., Zandieh, A., and Zhu, X. (2001) The sequence of the human genome. *Science* **291(5507)**:1304-1351

Warren, R. L., Sutton, G. G., Jones, S. J., and Holt, R. A. (2007) Assembling millions of short DNA sequences using SSAKE. *Bioinformatics.* **23(4)**:500-501

Wu, T. D., and Nacu, S. (2010) Fast and SNP-tolerant detection of complex variants and splicing in short reads. *Bioinformatics* **26(7)**:873-881

Xia, E. H., Yang, D. R., Jiang, J. J., Zhang, Q. J., Liu, Y., Liu, Y. L., Zhang, Y., Zhang, H. B., Shi, C., Tong, Y., Kim, C., Chen, H., Peng, Y. Q., Yu, Y., Zhang, W., Eichler, E. E., and Gao, L. Z. (2017) The caterpillar fungus, *Ophiocordyceps sinensis*, genome provides insights into highland adaptation of fungal pathogenicity. *Scientific Reports* **7(1)**:1806

Xie, Y., Wu, G., Tang, J., Luo, R., Patterson, J., Liu, S., Huang, W., He, G., Gu, S., Li, S., Zhou, X., Lam, T. W., Li, Y., Xu, X., Wong, G. K., and Wang, J. (2014) SOAPdenovo-Trans: *de novo* transcriptome assembly with short RNA-Seq reads. *Bioinformatics* **30(12)**:1660-1666

Young, N. D., Debelle, F., Oldroyd, G. E., Geurts, R., Cannon, S. B., Udvardi, M. K., Benedito, V. A., Mayer, K. F., Gouzy, J., Schoof, H., Van de Peer, Y., Proost, S., Cook, D. R., Meyers, B. C., Spannagl, M., Cheung, F., De Mita, S., Krishnakumar, V., Gundlach, H., Zhou, S., Mudge, J., Bharti, A. K., Murray, J. D., Naoumkina, M. A., Rosen, B., Silverstein, K. A., Tang, H., Rombauts, S., Zhao, P. X., Zhou, P., Barbe, V., Bardou, P., Bechner, M., Bellec, A., Berger, A., Berges, H., Bidwell, S., Bisseling, T., Choisne, N., Couloux, A., Denny, R., Deshpande, S., Dai, X., Doyle, J. J., Dudez, A. M., Farmer, A. D., Fouteau, S., Franken, C., Gibelin, C., Gish, J., Goldstein, S., Gonzalez, A. J., Green, P. J., Hallab, A., Hartog, M., Hua, A., Humphray, S. J., Jeong, D. H., Jing, Y., Jocker, A., Kenton, S. M., Kim, D. J., Klee, K., Lai, H., Lang, C., Lin, S., Macmil, S. L., Magdelenat, G., Matthews, L., McCorrison, J., Monaghan, E. L., Mun, J. H., Najar, F. Z., Nicholson, C., Noirot, C., O'Bleness, M., Paule, C. R., Poulain, J., Prion, F., Qin, B., Qu, C., Retzel, E. F., Riddle, C., Sallet, E., Samain, S., Samson, N., Sanders, I., Saurat, O., Scarpelli, C., Schiex, T., Segurens, B., Severin, A. J., Sherrier, D. J., Shi, R., Sims, S., Singer, S. R., Sinharoy, S., Sterck, L., Viollet, A., Wang, B. B., Wang, K., Wang, M., Wang, X., Warfsmann, J., Weissenbach, J., White, D. D., White, J. D., Wiley, G. B., Wincker, P., Xing, Y., Yang, L., Yao, Z., Ying, F., Zhai, J., Zhou, L., Zuber, A., Denarie, J., Dixon, R. A., May, G. D., Schwartz, D. C., Rogers, J., Quetier, F., Town, C. D., and Roe, B. A. (2011) The *Medicago* genome provides insight into the evolution of rhizobial symbioses. *Nature.* **480(7378)**:520-524

Zerbino, D. R., and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research* **18(5)**:821-829

Zhang, W., Chen, J., Yang, Y., Tang, Y., Shang, J., and Shen, B. (2011) A practical comparison of *de novo* genome assembly software tools for next-generation sequencing technologies. *PLoS ONE* **6(3)**:e17915

Zhao, S. (2014) Assessment of the impact of using a reference transcriptome in mapping short RNA-Seq reads. *PLoS One.* **9(7)**:e101374

Zhao, X., Mehrabi, R., and Xu, J. R. (2007) Mitogen-activated protein kinase pathways and fungal pathogenesis. *Eukaryot Cell* **6(10)**:1701-1714

Chapter 3

Genome annotation of the maize fungal pathogen, *Cercospora zeina*

N. A. Olivier, Y-C. Lin, Y. Van de Peer, D. K. Berger

Department of Plant and Soil Sciences, Faculty of Natural and Soil Sciences, Forestry and Agricultural Research Institute, University of Pretoria, Pretoria, 0002, South Africa

Data from this chapter, together with Chapter 2 was published as a genome announcement in *IMA Fungus* (Wingfield *et al.*, 2017). A study on the cercosporin biosynthesis gene cluster was published in *Molecular Plant Microbe Interactions* (Swart *et al.*, 2017). I performed all the annotation, quality control and functional prediction analyses. Dr Y-C. Lin from the laboratory of Prof Y. Van de Peer provided invaluable technical and procedural assistance for all aspects of the gene prediction process. Prof D. K. Berger was responsible for planning, funding and coordinating the sequencing and annotation project. He provided direction in interpreting the results and critical revision of the manuscript. He also arranged funding for a research visit to the laboratory of Prof. Y. Van de Peer where the initial annotation procedures for the chapter were conducted.

3.1 Abstract

The *Cercospora zeina* genome assembly enabled the design of systems biological research projects aimed at understanding the host-pathogen interaction between *C. zeina* and *Zea mays* at genomic level, as well as the fungal population structure in South Africa. Functional biological studies, however, require knowledge of the gene content of an organism. Protein sequence data from related organisms were mapped to the genome assembly of *C. zeina*, together with RNAseq data and Uniprot protein information. The lines of evidence were used to manually predict the structure of 145 genes, and these were subsequently used to train gene prediction algorithms specific for the *C. zeina* genome. Following manual confirmation of the accuracy of these algorithms, and gene model structure editing for Genbank upload, the final number of predicted genes was found to be 10,193. The functional content of the genes were predicted using various analysis tools. The presence of carbohydrate active enzymes, secondary metabolite genes, as well as secreted lipases, secreted small proteins, and secreted proteases in *C. zeina* were reported in comparison to three closely related species, i.e. *Cercospora beticola*, *Cercospora berteroae* and *Cercospora zeae-maydis*.

3.2. Introduction

The explosive rise in the number of sequenced genomes from organisms from all taxonomical domains have revolutionized the study of biology (Hood & Rowen, 2013). This has also increased the number and impact of comparative genome studies. However, a genome sequence without the associated functional content information is not conducive to in-depth study. The process of genome annotation is therefore closely linked to genome sequencing to unlock all aspects of genomic information, and is defined as the process of identifying the position and identity of functional units on genome sequences. These units include the regions of the genome where genes are located, as well as regions for repetitive DNA, non-coding RNA and regulatory regions (International Human Genome Sequencing, 2004).

3.2.1 Repetitive DNA

Repetitive DNA composes the largest part of the eukaryotic genome, with up to 98% of the human genome classified as non-protein coding and containing up to 50% repetitive DNA (Treangen & Salzberg, 2011). There are three main classes of repetitive DNA in eukaryotic DNA, i.e. Dispersed repeats/transposable elements, structural components of chromosomes and tandem repeats.

3.2.1.1 Transposable elements

Transposable elements (also transposons), are classified relative to the type of source molecule, with retro-transposons arising from an RNA intermediate, while DNA transposons are moved via DNA. Transposons are inserted over broad regions of the eukaryotic nuclear genome, and accumulate to different degrees depending on the class of transposon and the host species (Biscotti, Olmo*, et al.*, 2015). Functionally the transposons have not been extensively characterized, though the possible involvement of V-SINE elements with miRNAs forms part of gene-regulatory networks in vertebrates (Scarpato *et al.*, 2015). In plants the amplification of different retroelements is suggested to have an effect on genome behavior and divergence, and as a result could lead to speciation (Santos *et al.*, 2015). Finally, the high GC-content of some transposons have been shown to form structure made of four DNA strands known as G-quadruplexes. These structures have an effect on replication, transcription, translation, chromatin status and recombination, and therefore contributed to the evolution of a cellular regulatory network (Kejnovsky *et al.*, 2015).

3.2.1.2 Centromeric and telomeric repeats

Centromeric and telomeric repeats are considered structural components of eukaryotic chromosomes, since they contribute to the heterochromatic packing of DNA in these regions which contribute to function. The heterochromatin packing in human centromere regions, for example, aids with kinetochore binding and function during mitosis (Mehta *et al.*, 2010). Telomeres are heterochromatic regions at the ends of chromosomes containing six-to-eight base-pair guanine repeats which serve to prevents genetic information loss during DNA replication. By shortening telomeric repeat regions (200-400bp per cell division), and also avoiding chromosome fusion by their presence, these repeats maintain cell health and function, and prevent pre-mature apoptosis or senescence (Biscotti, Canapa*, et al.*, 2015).

3.2.1.3 Tandem repeats

Tandemly repeated DNA comprises of motifs which are arranged adjacently in repeating arrays. Some of these repeats are classified as satellites, and make up a large proportion of heterochromatic DNA. These sequences can also occur in specific chromosome regions (pericentromeric, intercalary and sub-telomeric) and in some organisms these repeats can even be species and chromosome-specific (Heslop-Harrison & Schmidt, 2001).

Microsatellite repeat regions are found in most eukaryotic organisms and are short repeats (1-6bp) widely dispersed through organisms, in both coding and non-coding DNA, and in the nuclear, chloroplast and mitochondrial genomes. There is great diversity in the number of repeats of microsatellite sequences per locus, even within species, and these are used as polymorphic markers to study population structure and diversity (Heslop-Harrison & Schmidt, 2001 ; Kalia *et al.*, 2011). The functions of microsatellites depend on their genomic location. SSRs in coding regions could affect gene activation or lead to pseudogenes, while an intronic or UTR location may impact mRNA splicing, gene silencing and transcription or translation (Lawson & Zhang, 2006). Mini-satellites are another class of tandem repeat, consisting of longer repeating units (10–50 bp), are also distributed across genomic loci, but are otherwise poorly characterized (Heslop-Harrison & Schmidt, 2001).

3.2.2 Non-coding RNA

Non-coding RNA (ncRNA) is comprised of a large number of classes. They include transfer RNAs (tRNA), ribosomal RNAs, small nucleolar RNAs (snoRNAs), small nuclear RNAs (snRNA), microRNAs (miRNA), small interfering RNAs (siRNA), piwi-interacting RNAs, extracellular RNAs, small Cajal body-specific RNAs (International Human Genome Sequencing, 2004), and long non-coding RNA. These molecules are not translated into proteins, but can nevertheless play functional roles in various eukaryotic cellular processes. Conserved ncRNAs present in all cell lineages play important central roles, while more transient ncRNAs are specific to only a few closely related species. Non-coding RNAs have been shown to play important roles specifically in translation (Harish & Caetano-

Anolles, 2012), RNA splicing (Kishore & Stamm, 2006), DNA replication (Zhang *et al.*, 2011), gene regulation (Reiner *et al.*, 2006), genome defense (Aravin *et al.*, 2008) and chromosome structure (Jady *et al.*, 2006).

3.2.3 Regulatory regions

Regulatory regions in eukaryotic organisms are composed of promoters, enhancers, silencers and insulators (Lodish, 2000 ; Kellis *et al.*, 2014). The detection of these elements are more challenging than for protein-coding regions, since they are usually short and can tolerate higher sequence variation than other functional elements. In addition, their distribution and behavior follow few known rules (Kellis *et al.*, 2003 ; Kellis *et al.*, 2014).

3.2.3.1 Promoters

Promoters are DNA sequences which indicate the initiation of transcription for the RNA polymerase enzyme (Lodish, 2000). Generally eukaryotic promoters were classified of either the TATA or CpG classes, although additional classifications have refined the function of specific promotors to certain tissue types or biological processes, with a total of 10 classes evident. Promoters are situated upstream from the transcription start site (TSS), and is recognized by the presence of elements like a TATA box, INR box, GC-box, CCAAT-box and BRE (Gagniuc & Ionescu-Tirgoviste, 2012). These elements generally serve as binding sites for specific transcription factors regulating transcription, especially TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH (Smale & Kadonaga, 2003).

The TATA box was the first promoter element discovered in highly transcribed genes, and is situated ~25-35 bp upstream from the TSS. The TATA box function appears to be as positioning guide for the RNA polymerase to initiate transcription at the correct TSS. Instead of a TATA box, some eukaryotic genes contain an initiator element (INR box) 2 bp upstream from the TSS, though the function of both these elements appears to be identical (Smale & Kadonaga, 2003). In contrast to the well-defined TSS of genes containing TATA boxes or initiators, other genes show transcription initiation over extended regions. These genes have alternative promoter elements comprised of a stretch of GC-rich sequence upstream from the transcription start site region. These CpG islands are common in genes involved with intermediates in secondary metabolism, as well as in housekeeping genes (Lodish, 2000 ; Adachi & Lieber, 2002). The BRE (B recognition element) is located immediately upstream of the TATA box in some genes, and was found to be the binding site of TFIIB during initiation of the pre-initiation complex (Smale & Kadonaga, 2003).

The GC-box and CCAAT-box are promoter elements situated further upstream from the TSS and are both binding sites for multiple transcription factors. The

GC-box is located ~110 bp upstream from the TSS, and is the binding site for Zinc-finger proteins (Lundin *et al.*, 1994). The CCAAT-box is located 60-100 bp upstream from the TSS, and genes containing this element appears to require enhanced gene transcription activity. The element is absent in universally prevalent genes (Dolfini *et al.*, 2009)

A common feature of many organisms is the presence of bidirectional promoters (Koyanagi *et al.*, 2005 ; Wei *et al.*, 2011). Two adjacent functionally related genes on opposite strands, and with their 5' ends oriented towards one another, are often under the control of a shared bidirectional promoter. This allows the genes to be co-regulated and co-expressed (Adachi & Lieber, 2002). The prevalence of bidirectional promoters in the human genome is significant, comprising ~11% of genes (Trinklein *et al.*, 2004). A common feature of bidirectional promoters is the presence of a CpG island between the genes, usually overlapping at least the first exon of both genes (Adachi & Lieber, 2002).

3.2.3.2 Enhancers

Enhancers are control elements which can affect the transcription of genes, even though they can be located tens of thousands of base pairs from the TSS. They can be found up- or downstream from the promoter of the gene of interest, and may even be located in an intron of the gene. The looping of intervening DNA brings the enhancer elements in contact with the core promoter region for activity. The same activating factors that bind to the core promoter area can also bind to enhancer elements (Maston *et al.*, 2006).

3.2.3.3 Silencers

Silencers share many properties ascribed to enhancers, but they confer repressing effects on gene expression. They are positionally independent from the core promoter, and co-operative binding of different silencers can affect specific gene silencing. Silencers are binding sites for negative transcription factors, also known as repressors. Most genes are thought to be functionally repressed until the binding of the silencer is disrupted and transcription can be initiated (Maston *et al.*, 2006).

3.2.3.4 Insulators

Insulator elements function to insulate genes from the effects of transcriptional activity in neighbouring genes. They are typically 0.5-3 kb in size, and function in a position-dependent manner (Maston *et al.*, 2006). Insulators are rare, and might only be found in regions with a high occurrence of coding and regulatory information (Fourel *et al.*, 2004). The precise mechanism of action is unknown, but involves either or both the binding of insulators to transcriptional activators, or the specific arrangement of chromatin structure (Maston *et al.*, 2006).

3.2.4 Protein-coding regions

The coding regions or sequences (CDS) are those components of genes encoding protein sequence information. On genomic DNA the CDS is located downstream from the promoter area, and is located on only one of the DNA strands. The coding region is not continuous, since information is encoded in blocks called exons, connected by regions not present on mature mRNA, the introns. When transcribed, the mRNA product is processed to remove intronic sequences, and the exons concatenated to allow translation of the complete protein-coding information of the gene. Coding sequences start with a start codon, usually ATG, and end with a stop codon, one of TAA, TAG or TGA. On either side of these codons are regions that are transcribed on the mRNA molecule, but don't code for protein information. These are the 5' and 3' UTRs, and often contain regulatory sequences that control translation (Polyak & Meyerson, 2003). Exon-intron boundaries contain highly conserved sequences recognized by the spliceosome complex responsible for splicing out introns from the pre-mRNA molecule, i.e. GT on 5' and AG on the 3' end. During the splicing process the exon composition of mRNA can be adjusted to yield a range of proteins from the same mRNA molecule, a process called alternative splicing. It is estimated that up to 95% of human multi-exonic genes yield products of alternative splicing (Pan *et al.*, 2008).

3.2.5 Structural annotation

During the structural annotation process the presence and position of genomic elements are identified. This *in silico* process can be undertaken by using either empirical or *ab initio* methods (Stein, L., 2001). Three basic approaches exist which require varying degrees of external evidence. The first approach uses a single ab initio gene predictor to yield the most likely CDS for each gene model. The second approach combines the prediction output of several gene predictors, and uses a consensus chooser to obtain the best consensus CDS. The third approach involves elements of the first two approaches, but also relies on external evidence, such as RNAseq data and EST information, as well as manual curation, to yield CDS models most consistent with the evidence (Yandell & Ence, 2012). In addition to CDSs and exon-intron boundaries, the co-ordinates of 5' and 3' UTRs must also be predicted due to their important role in regulation.

3.2.5.1 Empirical annotation methods

Empirical annotation methods function by searching the genome for sequence similarity to extrinsic information in known databases. Usually protein or protein-coding DNA sequence data from publicly available databases are used to identify conserved exons using similarity search algorithms like BLAST or Smith Waterman. This approach can be computationally intensive and relies on the quality and depth of sequences in the extrinsic databases. Since BLAST does not have splice awareness models, it will only approximate the correct exon splice boundaries (Sleator, 2010 ; Yandell & Ence, 2012). Splice-aware aligners like Splign (Kapustin *et al.*, 2008), Spidey (Wheelan *et al.*, 2001), and Exonerate (Slater & Birney, 2005) can be used to accurately align protein, mRNA and EST data to the genome assembly. The availability of RNAseq data from the genome of interest is invaluable to assist in the correct delineation of exons and splice-sites (Yandell & Ence, 2012).

3.2.5.2 *Ab initio* annotation methods

Ab initio (*de novo*) gene predictors rely on mathematical models to predict the presence of genes and the correct exon-intron boundaries. The algorithms underlying these models rely on neural networks, Fourier transformations, HMMs (Sleator, 2010) and machine learning techniques (Ratsch *et al.*, 2007). The fact that *ab initio* gene predictors don't require external evidence to function is a great advantage, especially when studying non-model organisms. Unfortunately the training parameters for many predictors were compiled from model organisms, and might not be suitable for all organism under study. In particular there are great variations in intron lengths, codon usage and GC content to be taken into account. Although the sensitivity of these models can greatly improve with training, the accuracy decreases due to the increase in false-positives (Korf, 2004 ; Yandell & Ence, 2012). Some *ab initio* predictors include Genemark (Ter-Hovhannisyan *et al.*, 2008), GENSCAN (Burge & Karlin, 1997), and Glimmer (Salzberg *et al.*, 1999) amongst others.

3.2.5.3 Combined annotation methods

There is a significant improvement in gene prediction accuracy when using *ab initio* gene predictors which have been trained on genome-specific data, and can also utilize extrinsic data, e.g. ESTs and RNAseq data (Sleator, 2010 ; Yandell & Ence, 2012). The approach is also called evidence-driven gene prediction, and a large number of software tools have been developed to incorporate this functionality. These include SNAP (Korf, 2004), AUGUSTUS (Stanke *et al.*, 2006), FGENESH (Salamov & Solovyev, 2000), Twinscan (Korf *et al.*, 2001) and BRAKER (Hoff *et al.*, 2016). Although these tools have greatly improved the process, they can be more difficult to implement due to the additional computational alignment, mapping and post-processing required. Several genome annotation pipelines have been developed to incorporate all these principles, but to require less user supervision. MAKER (Cantarel *et al.*, 2008), PASA (Haas *et al.*, 2003) and Gnomon (Souvorov, 2010) can utilize a number of gene predictors and feed the extrinsic data to the gene predictors during the analysis. A choosing algorithm selects the most probable and representative annotation, and UTRs are added with reference to the RNAseq or EST data provided to obtain gene models most consistent with the extrinsic evidence. The NCBI (Kitts, 2003) and Ensemble (Curwen *et al.*, 2004) have similar pipelines available (Yandell & Ence, 2012).

3.2.6 Editing and viewing

Manual verification and curation of the genome annotation is an important part of the annotation process. Verification of gene models are performed with reference to information in the relevant public and private databases. Genome annotation editing tools import the genome sequence, genome annotation prediction and extrinsic evidence in a graphic user interface (GUI), where it is possible to change the exon boundaries, create and remove gene models based on evidence, and export the relevant genome annotation files in standard formats (Yandell & Ence, 2012). Some examples of these editing tools are Genomeview (Abeel *et al.*, 2012), Webapollo (Lewis *et al.*, 2002), Artemis (Rutherford *et al.*, 2000) and ORCAE (Sterck *et al.*, 2012). In addition, the genome annotation can be visualized for general release using a range of web genome browsers, such as JBROWSE (Buels *et al.*, 2016) or GBROWSE (Stein, L. D., 2013), while local software like Integrated Genome Viewer (Robinson *et al.*, 2011) can also be used.

3.2.7 Functional annotation

Assigning biological information to the functional elements in the genome annotation is one of the prime goals of a successful genome sequence project, since a well-characterized proteome is required for additional functional or comparative genomics research. Low throughput methods are the most reliable in ascertaining specific protein functions, but these methods are expensive and time-consuming and at best only suitable for small subsets of genes of interest. As a result many research groups focus on developing high-throughput computational tools for predicting biological function based on sequence and structural motifs. The number of tools available for protein annotation is large and diverse, and a subset of these are discussed in the following sections.

3.2.7.1 Gene Ontology

Standardizing the nomenclature of protein function is critical to ensure uniformity and remove unnecessary duplication in protein function descriptions. The Gene Ontology Consortium has the goal of producing and maintaining a controlled vocabulary to describe the roles of genes and gene products in organisms. The gene ontology (GO) classification is constructed around three independent ontologies which serve as reference for protein functional description. The biological process ontology refers to the biological objective to which the protein contributes, e.g. 'carbohydrate metabolism'. The molecular function ontology is defined as the biochemical activity of a protein, e.g. 'deacetylase'. The cellular component refers to the place in the cell where the protein is active, e.g 'cytoplasm'. By combining these ontologies and populating broad or specific descriptions, the function of proteins can be classified in an objective, controlled manner (Ashburner *et al.*, 2000 ; The Gene Ontology, 2017).

3.2.7.2 Universal Protein Resource

The Universal Protein Resource (UniProt) is a resource for protein sequence and annotation data. Three databases, the UniProt Knowledgebase (UniProtKB), the UniProt Reference Clusters (UniRef), and the UniProt Archive (UniParc), house various types of protein data, with varying levels of computational and human curation. An example is the UniProtKB which houses the TrEMBL computationally analyzed records, which is further curator-evaluated with additional literature information, and these reviewed records are added to the SwissProt database. The databases and associated webservers (www.uniprot.org/) can be freely accessed and annotation analyses performed, while the selected databases can also be downloaded for use in local analyses and pipelines (UniProt Consortium, 2018).

3.2.7.3 InterProScan

The InterPro consortium (InterPro), hosted by the European Bioinformatics Institute at the European Molecular Biology Laboratory (EMBL-EBI) is composed of databases and resources providing functional analysis of protein sequences (Finn *et al.*, 2017). This allows the classification of proteins into families, and identifying domains and important functional sites. InterProScan (Jones *et al.*, 2014) is the software allowing the analysis of protein and nucleic acid sequences with the predictive models contained in the InterPro consortium. The output provides a comprehensive analysis of the relevant domains and signatures in proteins for possible functional classification.

A number of additional databases and tools can also be included in the InterProScan workflow, although third-party licensing agreements have to be considered (Table 3.1).

Table 3.1 Databases and tools that can be combined in InterProScan analyses.

Database/Tool	URL	Citation
CATH-Gene3D	http://www.cathdb.info/	(Dawson *et al.*, 2017)
CDD	https://www.ncbi.nlm.nih.gov/cdd/	(Marchler-Bauer *et al.*, 2017)
HAMAP	http://hamap.expasy.org/	(Pedruzzi *et al.*, 2015)
HMMER	http://hmmer.org/	(Eddy, 2011)
MobiDB	http://mobidb.bio.unipd.it/	(Piovesan *et al.*, 2018)
Ncoils	https://bio.tools/ncoils/	(Lupas *et al.*, 1991)
PANTHER	http://www.pantherdb.org/	(Mi *et al.*, 2013)
Phobius	http://phobius.sbc.su.se/	(Kall *et al.*, 2004)
PIRSF	http://pir.georgetown.edu/pirwww/dbinfo/pirsf.shtml	(Wu *et al.*, 2004)
PFam	http://pfam.xfam.org/	(Finn *et al.*, 2016)
PRINTS	http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/	(Attwood *et al.*, 2012)
ProDom	http://prodom.prabi.fr/	(Servant *et al.*, 2002)
Prosite	http://prosite.expasy.org/	(Sigrist *et al.*, 2002)
SFLD	http://sfld.rbvi.ucsf.edu/django/	(Akiva *et al.*, 2014)
SignalP	http://www.cbs.dtu.dk/	(Petersen *et al.*, 2011)
SMART	http://smart.embl-heidelberg.de/	(Letunic & Bork, 2017)
SuperFamily	http://supfam.org/SUPERFAMILY/	(Gough *et al.*, 2001)
TIGRFAM	http://www.jcvi.org/cgi-bin/tigrfams/index.cgi	(Haft *et al.*, 2003)1
TMHMM	http://www.cbs.dtu.dk/	(Krogh *et al.*, 2001)

3.2.7.4 BLAST2GO

BLAST2GO is a computational pipeline for protein annotation. By using the BLAST algorithm to perform similarity searches to existing, characterized proteins, the functional information from the known proteins is transferred to the unknown query sequences. The data is represented by GO categories (Gotz *et al.*, 2008).

3.2.7.5 Clusters of Orthologous Groups

The Clusters of Orthologous Groups (COG) database provides a measure of the phylogenetic relationships of proteins encoded in complete genomes. COGs were determined by comparing protein sequences of complete genomes, and contain proteins or groups of paralogs from at least three lineages. Information for four functional categories are presented, i.e. cellular processes and signaling, information storage and processing, metabolism and poorly characterized proteins, with functional sub-classifications for all categories (Tatusov *et al.*, 2000). The current COG database contains both prokaryotic clusters (COGs) and eukaryotic clusters (KOGs). The KOG categories and functional classifications are provided in Table A6 in the Appendix.

3.2.7.8 KEGG

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a computerized resource for biological interpretation of biological and genomic data. Several databases contain collections of gene and protein information relevant to specific subcategories. The PATHWAY database is collection of pathway maps which can be used to link a gene and protein to a specific pathway and associated function. The maps are organized into seven sections, i.e. metabolism, genetic information processing, environmental information processing, cellular processes, organismal systems, human diseases and drug development. Genomic data is stored in the GENES and GENOMES databases, with links of un-annotated proteins made to existing data via ortholog prediction (Kanehisa & Goto, 2000)

3.2.7.6 SignalP

Signal peptides are sorting signals for shunting newly synthesized proteins towards the secretory pathway. These peptide sequences are short, typically 16-30 amino acids in length, and are present at the protein N-terminus (Blobel & Dobberstein, 1975 ; Martoglio, 2003). SignalP is a software tool to predict the presence and location of signal peptide cleavage sites in amino acid sequences. The prediction is based on a combination of several artificial neural networks, and outputs are provided as presence/absence, with the position of the peptides if present (Nielsen, 2017).

3.2.7.7 TMHMM

TMHMM is a predictive tool for transmembrane helices, and relies on a HMM trained with experimentally confirmed data. The output provides information on the number of predicted transmembrane helices, as well as the expected topologies of these regions (Krogh *et al.*, 2001).

3.2.7.9 Specialized webservers

In addition to free-standing prediction tools, a large number of web-servers are available for the general or more specialized annotation of proteins and protein domains. Some selected examples are discussed in the following sections.

3.2.7.9.1 Carbohydrate active enzymes

The Carbohydrate Active EnZyme (CAZy) classification is a sequence-based classification system of enzymes that are involved in the synthesis or break-down of saccharides. The system comprises several modules or families, i.e. glycosyltransferases, glycoside hydrolases, polysaccharide lyases, carbohydrate esterases, auxiliary activity families and carbohydrate binding modules. The CAZy classification and associated database undergoes continual curation, with new enzyme and module classes added continually. Two webservers, CAZy (CAZy ; Lombard *et al.*, 2014) and dbCAN (dbCAN ; Yin *et al.*, 2012) are available for predicting carbohydrate active enzymes in a protein sequence set. Results are provided as summary of all classes of CAZy classes represented, and also specific classification details for each submitted protein sequence.

3.2.7.9.2 Secondary metabolite biosynthesis gene clusters

Bacterial and fungal secondary metabolites are rich sources of novel bioactive compounds with pharmaceutical value, but a subset of these can also be involved in pathogenesis. Studies on the genes responsible for secondary metabolism synthesis indicated that gene clustering is a hallmark of the great majority of these genes and was observed in almost all secondary metabolite classes. In addition, due to the limited carbon sources for synthesis the backbone biosynthesis genes show significant domain similarity. Once the backbone genes have been predicted, the adjacent genes are evaluated for contribution to the synthesis of the class of secondary metabolite indicated by the backbone genes (Cacho *et al.*, 2014).

Two webservers dedicated to secondary metabolite gene-cluster prediction are antiSMASH (Weber *et al.*, 2015) and SMURF (Secondary Metabolite Unique Regions Finder) ; (Khaldi *et al.*, 2010). antiSMASH utilizes HMMs specific for certain types of gene clusters in nucleotide input sequences, and provides information on the backbone and related genes in each cluster along with information on related organisms in the antiSMASH database. In addition, COGs are provided for all genes involved in the cluster to attempt to assign function to uncharacterized proteins. antiSMASH provides two separate dedicated prediction servers, i.e. a bacterial (antiSMASH-bacterial) and fungal (antiSMASH-fungal) version. There are also stand-alone versions available for local analysis. SMURF (SMURF) relies on PFAM and TIGRFAM domain content, as well as chromosomal position for prediction. Input files are the protein sequences for an organism, as well as a gene co-ordinate file for chromosomal positioning. Output files provide information on genes involved in specific clusters, as well as the backbone genes defining the cluster function.

3.2.7.9.3 The peptidase database

The MEROPS database (MEROPS ; Rawlings *et al.*, 2018) is a collection of sequences for peptidases and inhibitors, and has a hierarchical, structure based organization. Proteins are assigned to families based on function, and families believed to be homologous are grouped into clans. The database can be downloaded and used for BLAST similarity searches to putatively identify proteases and inhibitors in the organism of interest. The classification includes aspartic, cysteine, glutamic, metallo, asparagine, mixed, serine, threonine, unknown and compound proteases. A web-server BLAST option is also available on the database website.

3.2.7.9.4 Lipase Engineering Database

The Lipase Engineering Database (LED ; Fischer & Pleiss, 2003) aims to provide lipase sequence and classification information for facilitating biotransformation and biocatalyst engineering. Lipases share an α/β-hydrolase structural fold and the catalytic triad GxSxG-motif (Mehta *et al.*, 2010). Based on the structural oxyanion hole formed by the backbone amides of two conserved residues, lipases in the database are classified into three classes: GX, GGGX, and Y. In addition, fungal lipases have been grouped into five subclasses, i.e. two in the GX class, two in the GGGX class and one in the Y class. The database can be downloaded for BLAST similarity searches to identify putative lipases in the organism of interest. A web-server BLAST option is also available on the database website.

In this study we predicted and functionally annotated gene models in the *C. zeina* genome assembly in comparison with closely related *Cercospora* species. The objectives were: i) to predict gene models on the genome assembly of *C. zeina*, ii) to predict functional annotations on the genes, and iii) to confirm the functional and biological relevance of the gene annotations by comparison to that of closely related species. *Ab initio* gene prediction with the MAKER pipeline yielded 10,339 gene models which were reduced to a final number of 10,193 following manual curation, with core gene content analysis indicating the gene annotation to be 95.4% complete. The functional gene prediction was performed using InterProScan analysis, web server and local similarity searches to specific protein classes, and by manual gene structure confirmation of the cercosporin biosynthesis cluster proteins. The numbers and classes of carbohydrate active enzymes, secondary metabolite production genes, as well as secreted small proteins, proteases and lipases were extracted and compared with the functional content of related *Cercospora* species.

3.3 Materials and Methods

3.3.1 First manual gene-prediction

Manual gene prediction procedures require evidence for possible gene region positions in the genome assembly. This evidence is obtained from similarity searches of known genes/proteins to the genome assembly. Protein products or Expressed Sequence Tag (EST) information from closely related organisms are generally used.

3.3.1.1 Mapping with *C. zeae-maydis* protein data

The protein annotation and transcript data from the *C. zeae-maydis* genome assembly were downloaded (Strain SCOH1-5, JGI Project ID 401984) and mapped to the *C. zeina* genome assembly using the Exonerate software (Slater & Birney, 2005). The command used was:

```
exonerate --model protein2genome /path/to/czm/data/czm_prot.fasta
/path/to/genome-assembly/genome-assembly.fasta --refine full --score
100 --showvulgar yes --softmaskquery no --minintron 20 --maxintron
15000 --showalignment no --showtargetgff yes > output.gff
```

Settings in the command were as follows:

model — The alignment model used, in this case the alignment of protein sequences to genomic DNA

refine — Refined alignments generated by heuristics using dynamic programming over larger regions. Exhaustive alignments were calculated from the pair of sequences in their entirety

score — Overall score threshold. Alignments were not reported below this threshold

showvulgar — Alignments shown in Verbose Useful Labelled Gapped Alignment Report (vulgar) format

softmaskquery — Indicated whether the query was softmasked

minintron — Minimum intron length limit

maxintron — Maximum intron length limit

showalignment — Alignments presented in human readable form

showtargetgff — GFF output reported for features on the target sequence

3.3.1.2 Mapping with *C. zeae-maydis* transcript data

The transcript annotation data from the *C. zeae-maydis* genome assembly was mapped to the *C. zeina* genome assembly using the Exonerate software (Slater & Birney, 2005). The command used was:

```
exonerate --model est2genome /path/to/czm/data/czm_est.fasta
/path/to/genome-assembly/genome-assembly.fasta --refine full --score
100 --showvulgar yes --softmaskquery no --minintron 20 --maxintron
15000 --showalignment no --showtargetgff yes > output.gff
```

Settings in the command were as follows:

model — The alignment model used, in this case the alignment of transcript sequences to genomic DNA

refine — Refined alignments generated by heuristics using dynamic programming over larger regions. Exhaustive alignments were calculated from the pair of sequences in their entirety

score — Overall score threshold. Alignments were not reported below this threshold

showvulgar — Alignments shown in Verbose Useful Labelled Gapped Alignment Report (vulgar) format

softmaskquery	Indicated whether the query was softmasked
minintron	Minimum intron length limit
maxintron	Maximum intron length limit
showalignment	Alignments presented in human readable form
showtargetgff	GFF output reported for features on the target sequence

3.3.1.3 Mapping with *C. zeina* Trinity assembly

The Trinity assembly (Section 2.4.10) was mapped to the *C. zeina* genome assembly using the Exonerate software (Slater & Birney, 2005). The command used was:

```
exonerate  --model  est2genome  /path/to/trinity-assembly/cz_trinity-
assembly.fasta   /path/to/genome-assembly/genome-assembly.fasta   --
refine  full  --score  100  --showvulgar  yes  --softmaskquery  no  --
minintron 20 --maxintron 15000 --showalignment no --showtargetgff yes
> output.gff
```

Settings in the command were as follows:

model	The alignment model used, in this case the alignment of transcript (EST) sequences to genomic DNA
refine	Refined alignments generated by heuristics using dynamic programming over larger regions. Exhaustive alignments were calculated from the pair of sequences in their entirety
score	Overall score threshold. Alignments were not reported below this threshold
showvulgar	Alignments shown in Verbose Useful Labelled Gapped Alignment Report (vulgar) format
softmaskquery	Indicated whether the query was softmasked
minintron	Minimum intron length limit
maxintron	Maximum intron length limit
showalignment	Alignments presented in human readable form
showtargetgff	GFF output reported for features on the target sequence

3.3.1.4 Creating a local BLAST database

A local BLAST database is required when performing BLAST-analyses using the command line BLAST scripts. Reference sequence files in the fasta-format are required. Databases for either protein or nucleic acid input data can be created using the makeblastdb command (Altschul *et al.*, 1990). The generic command used was:

```
makeblastdb —dbtype nucl —in /path/to/sequences.fasta —out
/path/to/db-name —parse_seqids
```

Settings in the command were as follows:

dbtype Type of molecule for database, i.e. nucleic acid (nucl) or protein (prot)

in Name of input sequence file

out Name of database to be created

parse_seqids Enable the retrieval of sequences based on the fasta-file header identifiers

3.3.1.5 Mapping with UniProt Swiss-prot data

The UniProt Swiss-Prot database of manually curated protein sequence data was downloaded and mapped to the *C. zeina* genome assembly using the BLASTX algorithm (Altschul *et al.*, 1990). The command used was:

```
blastx -db /path/to/database/uniprot_sprot -query /path/to/genome-
assembly/genome-assembly.fasta -out /path/to output-directory -evalue
1e-10 -outfmt 6 -num_threads 8
```

Settings in the command were as follows:

db Name of the BLAST database

query Name of the input sequence file

out Name of the output directory

evalue E-value cutoff for BLAST similarity search

outfmt The format of the output file, in this case tab-delimited text

num_threads Number of CPU cores to use during the analysis

3.3.1.6 Mapping with non-redundant NCBI database

The NCBI (NCBI) non-redundant protein sequence database was downloaded and mapped to the *C. zeina* genome assembly using the BLASTX algorithm (Altschul *et al.*, 1990). The command used was:

```
blastx -db /path/to/database/nr -query /path/to/genome-
assembly/genome-assembly.fasta -out /path/to output-directory -evalue
1e-10 -outfmt 6 -num_threads 8
```

Settings in the command were as follows:

db Name of the BLAST database

query Name of the input sequence file

out Name of the output directory

evalue E-value cutoff for BLAST similarity search

outfmt	The format of the output file, in this case tab-delimited text
num_threads	Number of CPU cores to use during the analysis

3.3.1.7 Mapping with *C. nicotianae* CTB-cluster genes

The protein sequences for the eight genes responsible for (CTB) in *C. nicotianae* were downloaded from the NCBI Genbank database (NCBI) (Table 3.2) and used to create a local BLAST database.

Table 3.2 Genbank sequence information for the *C. nicotianae* CTB genes

Gene name	Gene abbreviation	Genbank accession
Polyketide synthase	CTB1	AAT69682.1
O-methyltransferase	CTB2	ABK64180.1
Cercosporin toxin biosynthesis protein	CTB3	ABC79591.2
MFS transporter	CTB4	ABK64181.1
Oxidoreductase	CTB5	ABK64182.1
Reductase	CTB6	ABK64183.1
Oxidoreductase	CTB7	ABK64184.1
Zinc-finger transcription factor	CTB8	ABK64185.1

The sequences were mapped to the *C. zeina* genome assembly using the BLASTX algorithm (Altschul *et al.*, 1990). The command used was:

```
blastx -db /path/to/database/cnic-ctb -query /path/to/genome-
assembly/genome-assembly.fasta -out /path/to output-directory -evalue
1e-10 -outfmt 6 -num_threads 8
```

Settings in the command were as follows:

db	Name of the BLAST database
query	Name of the input sequence file
out	Name of the output directory
evalue	E-value cutoff for BLAST similarity search
outfmt	The format of the output file, in this case tab-delimited text
num_threads	Number of CPU cores to use during the analysis

3.3.1.8 Bowtie2 index for *C. zeina* genome assembly

A Bowtie2 (Langdon, 2015) index is required when mapping data with the Tophat2 software (Kim *et al.*, 2013). The genome assembly sequence data was used to build the index, and the generic command used was:

```
bowtie2-build -f /path/to/reference/reference.fasta
/path/to/index/index-name
```

Settings in the command were as follows:

f	The reference sequence input file in the fasta-format

3.3.1.9 Mapping *C. zeina* RNAseq data

The RNAseq data from the individual and combined *in vitro C. zeina* cultures (Section 2.3.7) were mapped to the *C. zeina* genome assembly using the Tophat2 software (Kim *et al.*, 2013). The generic command used was:

```
tophat2 -p 8 -r 0 --mate-std-dev 200 -o /path/to/output-dir
/path/to/bowtie-index/index /path/to/sequence-data/forward-reads.fq
/path/to/sequence-data/reverse-reads.fq
```

Settings in the command were as follows:

p Number of CPU cores to use during the mapping

r The expected (mean) inner distance between mate pairs

mate-std-dev The standard deviation for the distribution on inner distances between mate pairs

o Output directory

3.3.2 SNAP gene training set

A manually curated gene-model annotation dataset is required as training data for the SNAP *ab initio* gene-predictor (Korf, 2004). The genome assembly was loaded in the GenomeView genome browser (Abeel *et al.*, 2012). The respective similarity mapping data output files (Sections 3.3.1.1 - 3.3.1.8) were subsequently imported into GenomeView. In addition, the junction information file provided by the Tophat2 mapping was also imported to provide information on the possible intron/exon junction positions. Initially only the genome assembly contigs with *czeina1* and *czeina2* designations were analyzed for gene-content. The mapping information provided the reference positional information to create hypothetical gene models on the reference genome backbone in GenomeView. By using the junction information from Tophat2 as guide for the presence and position of introns, and performing nucleotide and protein similarity searches on each new gene-model using the NCBI BLAST-server (BLAST), the correct co-ordinates for each new gene model could be adjusted. Only gene-models with BLAST-hits to genes/proteins with known function (no hypothetical, unknown or similar descriptions) were included in the initial annotation. A total of 145 gene models with known function and correctly predicted gene co-ordinates (relative to known genes from other organisms) were annotated. The output from GenomeView was an annotation file in the General Feature Format (GFF).

The GFF annotation file was converted to the SNAP-specific ZFF-file format. The ZFF-format combines FASTA and GFF properties, and the created ZFF-file was in the short-format suitable for SNAP optimization. Both the ZFF-file and genome assembly were checked for gene-prediction errors, followed by conversion of all gene-models to plus-stranded. The gene-prediction parameters were estimated, and the final output was a Hidden Markov Model specific for *C. zeina*.

3.3.3 First annotation prediction

MAKER-P (Campbell, Law*, et al.*, 2014) is a parallelization-enabled MAKER application initially developed for the annotation of large, repeat-rich plant genomes. A MAKER-P instance was created on the Atmosphere iPlant initiative (Cyverse). The application requires three project-specific options files that pass the relevant settings to MAKER for each run, i.e. `maker_opts.ctl`, `maker_bopts.ctl`, `maker_exe.ctl`. By altering the parameters in these files, the behavior of the MAKER application can be adjusted. The `maker_opts.ctl` file passes the required general behavioral settings to MAKER (Appendix Table A1).

For the first MAKER run, the only input was the genome assembly, the SNAP HMM file, *C. zeae-maydis* transcript data (Section 3.3.1.2) and Transposable Element (TE) sequence information for masking these regions during the gene prediction process (provided with the MAKER-P instance). The `maker_bopts.ctl` file provides the BLAST and Exonerate statistics thresholds (Appendix Table A2). The applications and algorithms required by MAKER were specified in the `maker_exe.ctl` file, with the relevant paths to the software provided (Appendix Table A3).

Following the completion of the first annotation run, the MAKER-P `fasta_merge` and `gff3_merge` scripts were executed to extract the relevant data from the MAKER output directories. Three final output files were created, i.e. the complete genome annotation GFF file, the complete transcript sequences FASTA file and the complete protein sequences FASTA file.

3.3.4 Validation of SNAP gene prediction accuracy

To verify the accuracy of the SNAP HMM gene-predictor, the complete genome annotation GFF file was loaded in GenomeView together with the genome assembly. The gene-model prediction accuracy was manually checked on the genome assembly contigs with *czeina3* and *czeina4* designations. The co-ordinates of a further 45 gene models with known function and correctly predicted gene coordinates were confirmed. Where the SNAP gene-predictions were not accurate, the gene coordinates were adjusted to conform to the coordinates of known genes. The coordinate information of the additional genes was added to the SNAP ZFF file, and the *C. zeina*-specific SNAP HMM updated.

3.3.5 AUGUSTUS training data

The AUGUSTUS *ab initio* gene predictor requires a training parameter file for gene-prediction. Although AUGUSTUS have pre-trained parameter files available, these might not be suitable for the species under investigation. To create the required input, the `gff2gbSmallDNA.pl` script was used to create a Genbank-format file using the genome assembly and an annotation GFF file as input. The Genbank-file was used as input for the `randomSplit.pl` script, which created the training and testing gene-sets. The training data-sets usually comprised of 100 genes. The `optimize_augustus.pl` script was used to train AUGUSTUS and create the species-specific prediction parameters. Following a subsequent re-training and optimization using the `etraining` script, the prediction accuracy on the training data-set was evaluated by executing the `augustus` script with the species-specific prediction parameters and test gene-set as input.

The first data-set used to train AUGUSTUS consisted of the gene-coordinate information for the 195 manually annotated genes from the *C. zeina* genome assembly (Section 3.3.4). The second training set consisted of the MAKER-derived annotation data from the first annotation (Section 3.3.3) with the highest confidence as indicated by the Annotation Edit Distance (AED). The MAKER gene predictions with AED values between 0 and 0.2 were selected as highly confident and the coordinates compiled into a GFF annotation file.

Following AUGUSTUS training the prediction report provided quality metrics to validate the prediction accuracy of the 100 test gene-set. A gene level sensitivity of >20% indicates that the training dataset is large enough and the prediction accuracy of the prediction HMM is high enough for *ab initio* gene prediction.

3.3.6 Hard masking repeats

Repeats in the genome assembly were hard-masked with RepeatMasker (Smit *et al.*, 2013–2015) using the NCBI repeats database. The script used was:

```
RepeatMasker -engine ncbi -x -excln -species fungi
genome.assembly.fasta
```

Settings in the command were as follows:

engine	Search engine used for analysis
x	Returns repetitive regions masked with X
excln	Calculates repeat densities
species	Specify the species or clade of the input sequence, must occur in the repeat database

3.3.7 Genemark-ES training data

The GeneMark-ES *ab initio* gene predictor utilizes unsupervised training to generate a HMM specific for the species of interest. The *C. zeina* genome assembly hard-masked for repeat content (Section 3.3.6) was used as input and analyzed with the Genemark-ES predictor, with the GeneMark-ES HMM file generated. The script for training the GeneMark-ES predictor was:

```
perl /path/to/software/gm_es.pl --min_contig 200 --max_contig 938006
/path/to/assembly/genome.x-masked.fa > /path/to/output-
dir/genemark_hmm.gff 2> error.logfile
```

Settings in the command were as follows:

min_contig The minimum contig size was set as 200bp

max_contig The maximum contig size was set as 938,006bp, the size of the largest contig in the genome assembly

3.3.8 Second annotation prediction

The second MAKER-P annotation prediction utilized the SNAP, AUGUSTUS and GeneMark-ES *ab initio* gene predictors. The respective species-specific training HMMs for the gene predictors were specified in the MAKER options file. In addition, the *C. zeina* Trinity transcriptome assembly mapped to the genome (Section 3.3.1.3) was included to provide EST-information.

Following the completion of the second annotation run, the MAKER-P `fasta_merge` and `gff3_merge` scripts were executed to extract the relevant data from the MAKER output directories. Three final output files were created, i.e. the complete genome annotation GFF file, the complete transcript sequences FASTA file and the complete protein sequences FASTA file.

3.3.9 Third annotation prediction

The third MAKER-P annotation prediction only utilized the SNAP and AUGUSTUS *ab initio* gene predictors. The respective species-specific training HMMs for the gene predictors were specified in the MAKER options file. In addition, the Trinity transcriptome assembly mapped to the genome (Section 3.3.1.3) was included to provide EST-information.

Following the completion of the third annotation run, the MAKER-P `fasta_merge` and `gff3_merge` scripts were executed to extract the relevant data from the MAKER output directories. Three final output files were created, i.e. the complete genome annotation GFF file, the complete transcript sequences FASTA file and the complete protein sequences FASTA file.

3.3.10 GFF3 Annotation file format

The MAKER derived annotation file (GFF format) did not fully conform to the required ontology in terms of the attribute column content. The annotation file was converted to GFF3 specifications (www.sequenceontology.org) to reflect hierarchical groupings of features and sub-features.

3.3.11 Annotation submission to Genbank

The annotation file was uploaded to the NCBI WGS database to complement the genome assembly accession (Section 2.4.8). The GFF file contents were converted to the tab-delimited TBL format specified by the NCBI, followed by conversion to the required `.sqn` upload-format via the command-line `tbl2asn` program (tbl2asn2). Discrepancies in the gene feature coordinates identified by the `tbl2asn` file were corrected. These included incorrect coordinates for splice-donors/acceptor sites and stop and start codons, as well as truncated introns and exons with lengths less than the minimum required by Genbank. Incorrect gene-structures with no similarity to genes in the NCBI Genbank database were removed from the annotation file. Several genes predicted incorrectly as directly adjacent were manually combined into larger genes with correct similarities to genes in the Genbank database. Several large genes incorrectly merged during the prediction process were split into smaller genes with correct similarities to genes in the Genbank database. The 5' and 3' untranslated regions (UTR) of many genes were incorrectly predicted and often spanned gene-coding regions on the opposite strand. These were subsequently removed from the annotation file. The GFF annotation file was edited and updated with the corrected coordinate information. The protein sequences for *C. zeina* were obtained from the genome assembly accession page at the NCBI following public release of the genome assembly data for subsequent analysis.

3.3.12 Completeness analysis of genome annotation

To evaluate the completeness of the genome annotation, the BUSCO (Simao *et al.*, 2015) package was used. The analysis was performed using the Ascomycota-specific HMM predictor which evaluates the presence of 1,315 BUSCO protein groups specific to the Ascomycota. To compare the completeness of the annotation with other available *Cercospora* species, the total predicted proteomes were also obtained for *C. berteroae* (Strain CBS538.71, NCBI BioProject accession PRJNA270309, NCBI BioSample SAMN08288628) and *C. beticola* (Strain 09-40, NCBI BioProject accession PRJNA270309, NCBI BioSample SAMN03265455). Similar BUSCO analyses were performed on all additional proteomes.

3.3.13 Ortholog inference in *Cercospora* species

Orthologous genes were inferred between *C. zeina* and the additional *Cercospora* species (Section 3.3.1.1 and 3.3.12) with the OrthoFinder 2.2.6 package (Emms & Kelly, 2015), using the BLAST (Altschul *et al.*, 1990) similarity search gene-set. The BLAST e-value cut-off was set at 1x10^{-3}. Proteome-specific protein identifiers and sequences for each orthogroup were extracted from the `Orthogroups.txt` results file using command-line tools, and the number of genes in each orthogroup for the respective species were extracted from the `Orthogroups.GeneCount.csv` output file.

3.3.14 Cercosporin toxin biosynthesis cluster comparison

The cercosporin biosynthesis cluster genes predicted in Section 3.3.1.7 using the *C. nicotianae* genes were manually confirmed using Genomeview and the *C. nicotianae* mapping. The cluster genes in the other *Cercospora* species were also confirmed using GenomeView, using the co-ordinates in the respective genome annotation GFF-files to find the gene models in the respective genomes. Due to the genome annotation of *C. nicotianae* not being available, TBLASTN homology searches of the *C. nicotianae* CTB cluster proteins (Table 3.2) to the genome assembly was used to obtain co-ordinates for the gene positions. The gene models were subsequently refined by changing the open reading frame. The synteny of the gene clusters in all the species were determined and schematically represented.

3.3.15 Functional annotation and comparison of *Cercospora* genes

The protein sequences and genome annotation GFF-files for *C. berteroae*, *C. beticola* and *C. zeae-maydis* were obtained as described in Section 3.3.1.1 and Section 3.3.12. Functional gene prediction was performed simultaneously on all proteomes to standardize comparisons.

3.3.15.1 InterProScan analyses

The prediction of functional domains and motifs in the proteomes were performed using InterProScan 5.29 (Jones *et al.*, 2014). The databases and tools included in the analyses are listed in Table 3.1. Input for the analyses were the multi-fasta protein sequence files for the respective species, while the outputs were in tab-delimited text format for parsing with standard tools.

3.3.15.2 Eukaryotic orthologous groups

The eukaryotic-specific COGs were predicted using the online eggnog-mapper (emapper ; Huerta-Cepas *et al.*, 2017), a tool on the EggNOG v4.5.1 web-server (Huerta-Cepas *et al.*, 2016). The mapping was performed using Diamond, with the recommended settings used for the Taxonomic Scope, Orthologs and Gene

Ontology evidence. Input for the analyses were the multi-fasta protein sequence files for the respective species, while the outputs were in tab-delimited text format for parsing with standard tools.

3.3.15.3 Small secreted proteins

Small secreted proteins (SSP) were predicted from the InterProScan results using the predicted SignalP output for each protein. All proteins with a predicted signal peptide in the first 20 amino acids and a total length less than or equal to 200 amino acids were classified as SSPs.

3.3.15.4 Carbohydrate activating enzymes

Carbohydrate activating enzymes (CAZymes) were predicted separately for each species using the dbCAN v1 (Yin *et al.*, 2012). Input for the analyses were the multi-fasta protein sequence files for the respective species, while the outputs were in tab-delimited text format for parsing with standard tools.

3.3.15.5 Secondary metabolite production genes

Secondary metabolites have been shown to be required for virulence and host-specificity in *Dothideomycetes* (Panaccione *et al.*, 1992 ; Yang *et al.*, 1996). Genes involved in secondary metabolite production, i.e. polyketide synthases (PKS), non-ribosomal peptide synthetases (NRPS) and terpene synthases (TPS) were subsequently predicted. PKSs and NRPSs were predicted using the SMURF webserver (Khaldi *et al.*, 2010), with required input files being the relevant proteome sequence files in multi-fasta format, as well as gene co-ordinate files created by editing the relevant genome annotation GFF files. The gene co-ordinate file contained the protein ID (same as protein IDs in multi-fasta file), chromosome/contig ID, 5' gene start and 3' gene stop co-ordinates and the protein name/function/definition (if available). Outputs were in tab-delimited text format for parsing with standard tools.

TPSs were predicted using the fungal antiSMASH secondary metabolite gene cluster prediction web-server (Weber *et al.*, 2015), with input for the analyses being the respective genome assemblies. Input for the analyses were the multi-fasta protein sequence files for the respective species, while the outputs were graphically represented on the antiSMASH website. Only PKS predictions were taken into account and correlated with the respective genes from the protein set for each species. The orthologs for each PKS protein were determined using the Orthofinder orthology analysis (Section 3.3.13)

3.3.15.6 Secreted lipases

Secreted lipases were predicted with BLASTP similarity searches against sequences from the Lipase Engineering Database (Fischer & Pleiss, 2003). The multi-fasta sequence file for each Lipase superfamily was downloaded, a database created from the sequence file and a BLASTP similarity search performed using the relevant proteome sequence files for each *Cercospora* species as input. The identified lipases were checked for possible secretion by confirming the presence of signal peptide sequences as predicted by the SignalP output of the InterProScan analysis (Section 3.3.15.1). The orthologs for the secreted lipases were determined using the Orthofinder orthology analysis (Section 3.3.13).

3.3.15.7 Secreted proteases

Secreted proteases were predicted for each species by BLAST similarity searches to the full-length peptidase list (protease.lib) from the MEROPS database v12.0 (Rawlings *et al.*, 2018). The full peptidase library was downloaded the BLAST searches performed against each species with e-values of 0.0001. The presence of Pfam identifiers linked to 'protease', 'proteinase' and 'peptidase', but excluding 'inhibitor' were also extracted from the InterProScan analysis. The InterProScan output positive for SignalP regions were used to select for the protein sub-set predicted to be secreted. The MEROPS IDs were extracted only for these proteins using command line tools, followed by the removal of duplicate IDs for each protein.

3.4 Results

3.4.1 Manual gene prediction

The genome assembly and all levels of mapping evidence were loaded in GenomeView (Figure 3.1). By using the mapping evidence as starting point for each predicted gene model, the reading frames were iteratively adjusted to conform to the RNA-sequencing intron/exon structure, the appropriate splice donor-acceptor sequences and BLAST similarity to known proteins on Genbank.

The initial gene prediction on the *czeina1* and *czeina2* scaffold yielded 145 high-confidence gene-models when compared to known genes in the Genbank sequence database. The annotations were saved into an annotation file (GFF) for further predictions.

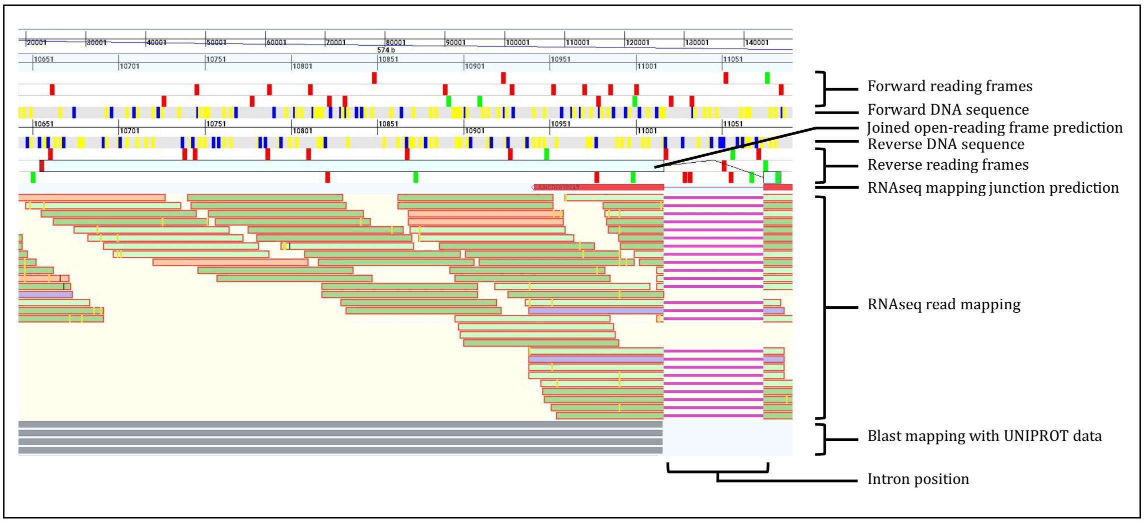


Figure 3.1 GenomeView manual gene structure prediction. *The forward and reverse sequences of the genome are used to evaluate respective forward and reverse open reading frames. The yellow and blue vertical blocks indicate possible splice donor and acceptor sites. Green vertical blocks indicate possible start codons, while red vertical blocks indicate possible stop codons. Evidence from BLAST mapping to UNIPROT data, as well as RNAseq mapped reads are used as guide to create and merge possible open reading frames.*

3.4.2 First MAKER prediction

The first MAKER prediction, using only the trained SNAP gene-predictor yielded a total of 10,407 predicted gene models. The number of gene models was similar in range to related *Dothideomycete* genomes (Table 3.3).

Table 3.3 Number of predicted gene models for *Dothideomycete* species closely related to *C. zeina*.

Organism	Number of predicted gene models
Cercospora berteroae	11,903
Cercospora beticola	12,468
Cercospora zeae-maydis	12,020
Septoria musiva	10,233
Septoria populicola	9,739

3.4.3 AUGUSTUS training

The GFF file with the manually annotated genes (Section 3.4.1) with AED value between 0 and 0.2 was converted to a Genbank-format file (`.gbk`), resulting in a gene-set containing 3,506 genes. Following the splitting of the original dataset, training and testing datasets were obtained containing 100 and 3,406 genes respectively. The optimized AUGUSTUS gene-prediction HMM was evaluated using the testing dataset, yielding a gene-level sensitivity of 0.45 (Table 3.4), which indicated the HMM was sufficiently accurate for predicting genes in the full genome (Stanke *et al.*, 2006).

Table 3.4 **AUGUSTUS prediction training accuracy metrics**

	Sensitivity	Specificity
Nucleotide level	0.952	0.706
Exon level	0.668	0.48
Gene level	0.45	0.274

3.4.4 Second MAKER annotation prediction

The HMM created during the Genemark-ES training predicted gene models without error and passed the prescribed quality indications during the training process, and was deemed acceptable for use in gene prediction. The second MAKER annotation run yielded no predicted gene models, and log-files showed that the performance of the Genemark-ES predictor was incompatible with the other gene predictors. The AUGUSTUS and SNAP predictors utilized *C. zeina*-specific extrinsic data during training, while the GeneMark-ES predictor relied primarily on general fungal information. It was therefore decided to rely on only the *C. zeina*-specific predictors for subsequent gene prediction.

3.4.5 Third MAKER annotation prediction

The third MAKER annotation run, utilizing the SNAP and AUGUSTUS trained predictors yielded 10,339 predicted gene models. Following editing of the genome during the NCBI WGS upload process, the number of gene models were decreased to 10,193. This was the final number of gene models and protein set which were used in all subsequent analyses.

3.4.6 BUSCO completeness analysis

The BUSCO evaluation of the *C. zeina* genome annotation protein set yielded a completeness report of C: 95.4%: (95.4% Complete and single-copy BUSCOs, D: 0% Complete and duplicated BUSCOs, 2.1% fragmented BUSCOs, M: 2.5% missing BUSCOs, total 1,315 proteins evaluated). This completeness figure, on protein level, corresponds very well with the completeness figure for the genome assembly (95.4%) on nucleotide level (Section 2.4.7). The BUSCO completeness of the *C. zeina* proteome was compared with the BUSCO completeness predicted for the other *Cercospora* species (Table 3.5), with results in the similar range.

Table 3.5 **BUSCO completeness prediction comparison between *Cercospora* genome annotations**

Description	*C. zeina*	*C. berteroae*	*C. beticola*	*C. zeae-maydis*
Total BUSCOs	1315	1315	1315	1315
Complete BUSCOs (C)	95.4%	98.3%	97.9%	95.8%
Complete and single-copy BUSCOs (S)	95.4%	98.3%	95.8%	95.7%
Complete and duplicated BUSCOs (D)	0%	0%	2.1%	0.1%
Fragmented BUSCOs (F)	2.1%	1.1%	1.0%	2.6%
Missing BUSCOs (M)	2.5%	0.6%	1.1%	1.6%

3.4.7 Cercosporin toxin biosynthesis cluster

The presence of the cluster of 8 genes responsible for the synthesis of the cercosporin toxin were confirmed for *C. zeina* on scaffold *czeina49* of the genome assembly (Figure 3.2).

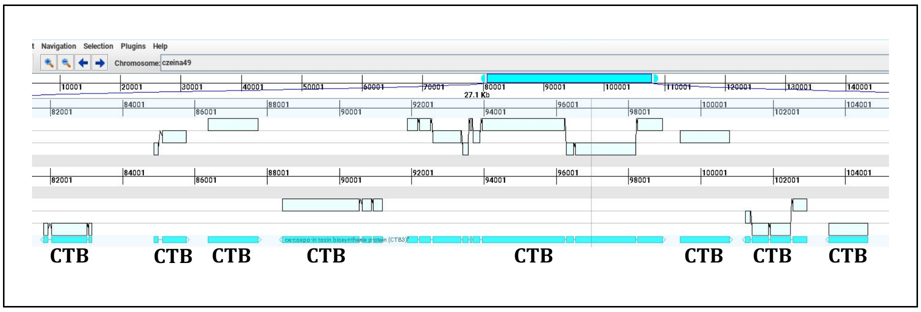


Figure 3.2 GenomeView gene structures of the *C. zeina* cercosporin biosynthesis cluster genes.

The cluster genes were uploaded to Genbank with accession numbers as in Table 3.6. Due to a nonsense mutation in the *C. zeina* Oxidoreductase gene (CTB7), this is classified as a pseudogene and does not have an accession number (Swart *et al.*, 2017), but is still visualized in Genomeview for comparison purposes.

Table 3.6 *C. zeina* cercosporin toxin biosynthesis cluster genes (Swart *et al.*, 2017). *In* C. zeina *CTB7 is a psedogene and does not have an accession number.*

CTB	Gene name / function	Genbank accession
CTB1	Polyketide synthase	ARU80380.1
CTB2	O-methyltransferase	ARU80381.1
CTB3	Cercosporin toxin biosynthesis protein	ARU80379.1
CTB4	MFS transporter	ARU80382.1
CTB5	Oxidoreductase	ARU80378.1
CTB6	Reductase	ARU80383.1
CTB7	Oxidoreductase (pseudogene)	N/A
CTB8	Zinc finger transcription factor	ARU80376.1

3.4.8 *Cercospora* species cercosporin toxin biosynthesis cluster comparison

To evaluate the accuracy and biological applicability of the *C. zeina* genome annotation, the gene models predicted for the cercosporin biosynthesis gene clusters for *C. berteroae*, *C. beticola*, *C. zeae-maydis* and *C. nicotianae* were visualized using GenomeView (Figure 3.3). Using the respective gene co-ordinates from the genome annotation GFF files the gene structures were recreated and confirmed using NCBI BLASTP comparisons to the host protein sequences. For the *C. zeae-maydis* annotation which contains multiple CDS and exon hits per gene locus, the gene-structures were recreated using TBLASTN co-ordinates mapped to the genome assembly, and confirmed with NCBI BLASTP comparisons to the nearest ortholog between the three known *Cercospora* species from the study. The *C. nicotianae* cluster gene co-ordinates are not known due to the genome annotation not being available. The TBLASTN co-ordinates of the *C. nicotianae* cluster gene comparison to the genome assembly was used to recreate the gene-structures, with confirmation using the NCBI BLASTP comparisons to the *C. nicotianae* Genbank sequences.

The cluster genes were predicted for all the species, and the gene order and close proximity of all genes within species were similar for all species. The accessions of the respective CTB cluster genes for all the annotated *Cercospora* species are listed in Table 3.7.

Table 3.7 Cercosporin toxin biosynthesis cluster protein accession numbers of *Cercospora* species.

CTB	*C. zeina* [a]	*C. zeae-maydis* [b]	*C. berteroae* [c]	*C. beticola* [d]
CTB1	ARU80380.1	e_gw1.15.349.1	PPJ53310.1	PIB02405.1
CTB2	ARU80381.1	e_gw1.15.62.1	PPJ53319.1	PIB02404.1
CTB3	ARU80379.1	e_gw1.15.397.1	PPJ53309.1	PIB02398.1
CTB4	ARU80382.1	e_gw1.15.135.1	PPJ53320.1	PIB02403.1
CTB5	ARU80378.1	e_gw1.15.344.1	PPJ53318.1	PIB02399.1
CTB6	ARU80383.1	e_gw1.15.258.1	PPJ53311.1	PIB02402.1
CTB7	-	e_gw1.15.87.1	PPJ53317.1	PIB02400.1
CTB8	ARU80376.1	N/A	PPJ53316.1	PIB02401.1

(a,c,d) Accessions from Genbank; (b) Gene number from C. zeae-maydis annotation GFF file.

The order of the genes in the clusters were inverted for the various organisms, but this was solely due to the respective sequence orientation of the genome assembly scaffolds where the gene clusters were identified. When analyzing the directions of genes in all clusters it is clear that the synteny of all cluster genes in all genomes are similar (Figure 3.4).

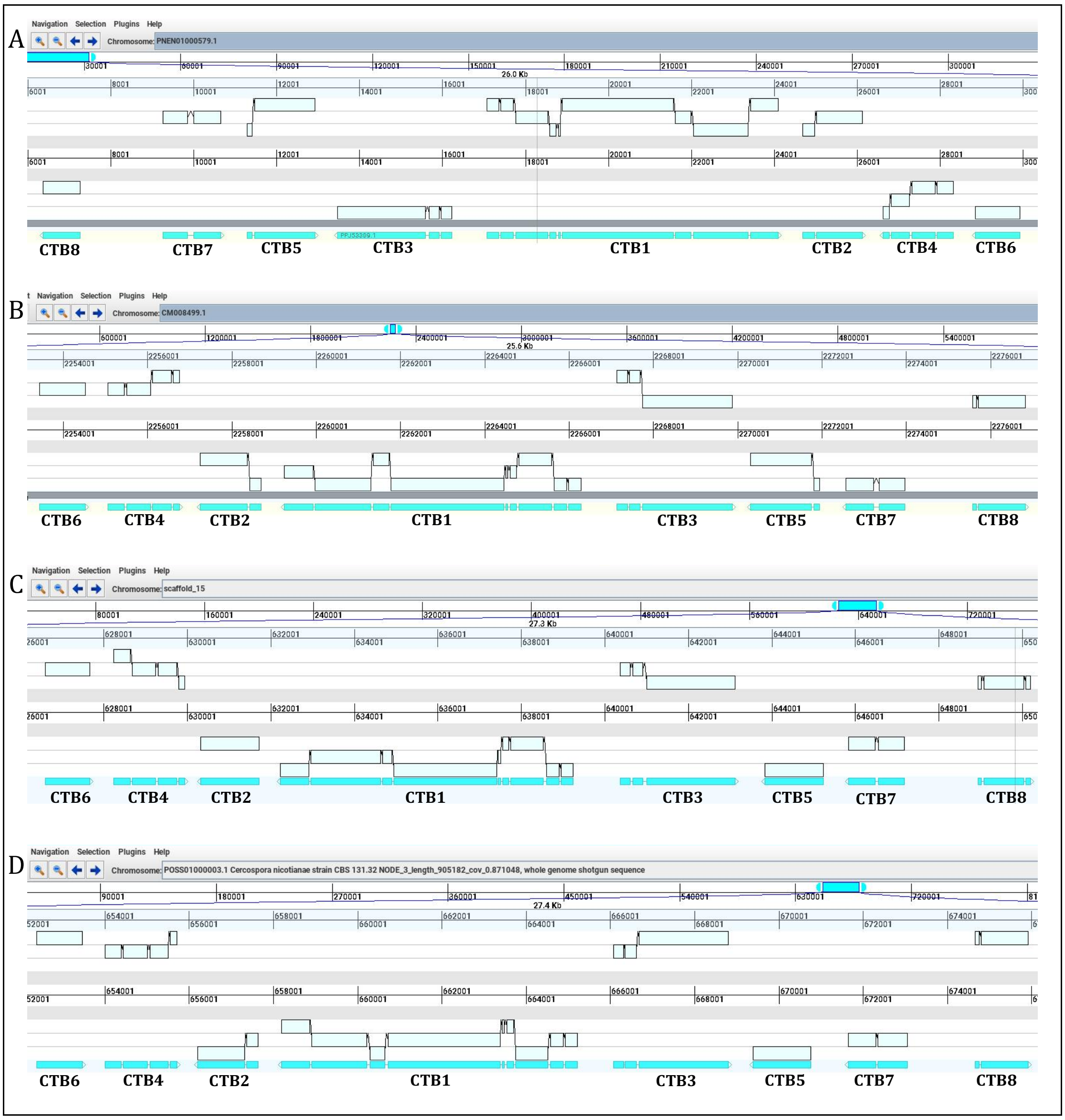

A
Navigation Selection Plugins Help
Chromosome: PNEN01000579.1
26.0 Kb
CTB8
CTB7
CTB5
CTB3
CTB1
CTB2
CTB4
CTB6
B
Navigation Selection Plugins Help
Chromosome: CM008499.1
25.6 Kb
CTB6
CTB4
CTB2
CTB1
CTB3
CTB5
CTB7
CTB8
C
Navigation Selection Plugins Help
Chromosome: scaffold_15
27.3 Kb
CTB6
CTB4
CTB2
CTB1
CTB3
CTB5
CTB7
CTB8
D
Navigation Selection Plugins Help
Chromosome: POSS01000003.1 Cercospora nicotianae strain CBS 131.32 NODE_3_length_905182_cov_0.871048, whole genome shotgun sequence
27.4 Kb
CTB6
CTB4
CTB2
CTB1
CTB3
CTB5
CTB7
CTB8

Figure 3.3 GenomeView gene structures of *Cercospora* species cercosporin biosynthesis cluster genes. *The classification is shown for* C. berteroae *(**A**),* C. beticola *(**B**),* C. zeae-maydis *(**C**), and* C. nicotianae *(**D**). The gene order between each species' cluster differs due to the orientation of the respective contig sequences.*

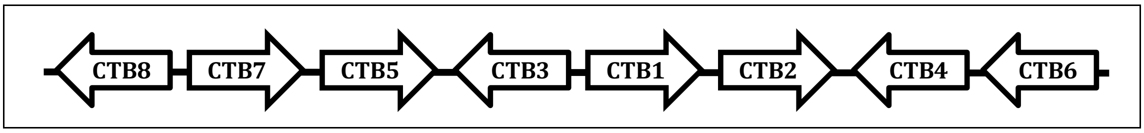


Figure 3.4 Synteny of the genes of the cercosporin toxin biosynthesis cluster. *The gene directions for all clusters were inferred from the GenomeView gene structures (Figure 3.4.2 and Figure 3.4.3).*

3.4.9 Orthologs between *Cercospora* species

The Orthofinder ortholog inference analysis provided a list of Orthogroups with the relevant orthologs from each species per Orthogroup. In total there were 13,578 Orthogroups predicted, with 7698 Orthogroups represented by at least one ortholog from each species (Figure 3.5). For the four *Cercospora* species a total of 6,641 single copy orthologs were predicted, being Orthogroups containing one ortholog per species. The maize-infecting species shared 263 unique Orthogroups, while the sugar beet-infecting species shared 760 unique Orthogroups. There were very few Orthogroups present in only one of the species.

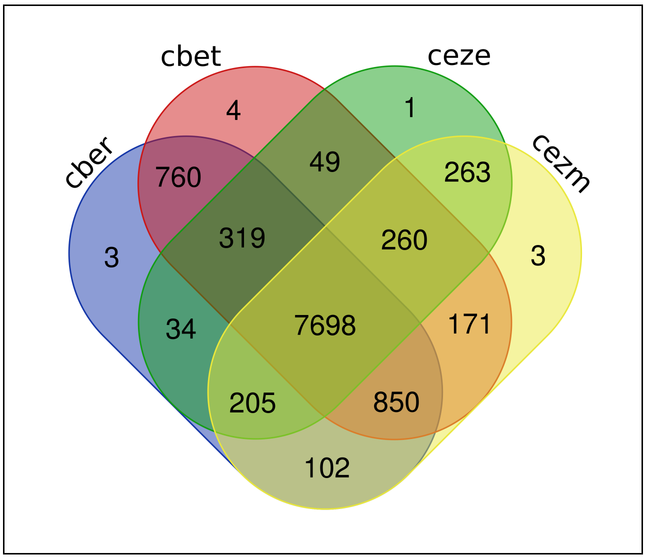


Figure 3.5 Venn-diagram of Orthofinder-inferred Orthogroup content compared between the *Cercospora* species. *Total Orthogroups per species are represented. The classification is shown for* C. berteroae *(**cber**),* C. beticola *(**cbet**),* C. zeina *(**ceze**), and* C. zeae-maydis *(**cezm**).*

3.4.10 Functional gene annotation

The predicted functional annotation of each protein of the four *Cercopora* species were separately predicted using multiple prediction algorithms. The InterProScan results were used for many of the annotation predictions.

3.4.10.1 Small secreted proteins

Small secreted proteins have been shown to be important for plant pathogenesis. The proteins in each species that contain a predicted signal peptide in the first 20 amino acids and has a length smaller than 200 amino acids were classified as small secreted proteins. For *C. zeina* a total of 147 small secreted proteins were predicted, while the numbers for the other species were 233 for *C. berteroae*, 238 for *C. beticola* and 98 for *C. zeae-maydis*.

Extracting the Pfam domains present in the SSPs for each species from the InterProScan output, it was found that only 11 Pfam domains are conserved among all species, with 48 unique domains being present in total. Each genome showed some domains unique to the species protein set, with no common domains present only in the maize-infecting species, while there are 3 unique domains present in the sugar beet-infecting species (Figure 3.6). These domains were the PF00313 ('Cold-shock' DNA-binding domain [DUF2237)], PF03330 (Lytic transglycolase), and PF09996 (Uncharacterized protein conserved in bacteria). The *C. zeina*-specific domains were PF00166 (Chaperonin 10 Kd subunit), PF07452 (CHRD domain) and PF00491 (Arginase family).

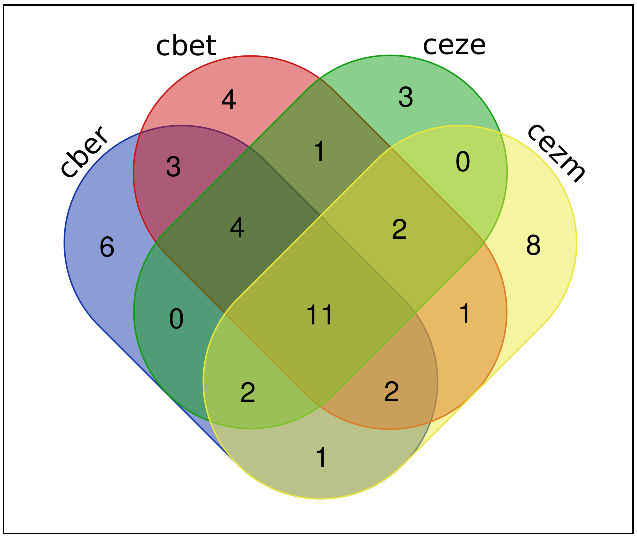


Figure 3.6 Venn diagram of the Pfam domains present in the small secreted proteins for the *Cercospora* species.

3.4.10.2 Eukaryotic Orthologous Groups

The eggNOG-mapper was used to predict KOG groups for the proteins of each *Cercospora* species. The output shows that all KOG classifications are present for all species (Figure 3.7) There were differences in the numbers of KOG classifiers predicted for the respective species, and the differences were normalized relative to the total number of genes predicted for each species (Table 3.8). Similar trends and ranges are apparent for all species.

Table 3.8 KOG classifications of the *Cercospora* species. *Percentages in parentheses indicate the total number of KOG classifications as percentage of the total gene content of the species.*

KOG classification	*C. zeina*	*C. berteroae*	*C. beticola*	*C. zeae-maydis*
A	263 (2.58%)	275 (2.31%)	288 (2.31%)	268 (2.23%)
B	86 (0.84%)	92 (0.77%)	97 (0.78%)	93 (0.77%)
C	354 (3.47%)	397 (3.34%)	425 (3.41%)	399 (3.32%)
D	143 (1.4%)	146 (1.23%)	150 (1.2%)	143 (1.19%)
E	370 (3.63%)	418 (3.51%)	431 (3.46%)	389 (3.24%)
F	106 (1.04%)	116 (0.97%)	127 (1.02%)	114 (0.95%)
G	623 (6.11%)	768 (6.45%)	767 (6.15%)	662 (5.51%)
H	135 (1.32%)	147 (1.23%)	151 (1.21%)	138 (1.15%)
I	360 (3.53%)	429 (3.6%)	457 (3.67%)	402 (3.34%)
J	351 (3.44%)	371 (3.12%)	378 (3.03%)	368 (3.06%)
K	373 (3.66%)	395 (3.32%)	419 (3.36%)	401 (3.34%)
L	220 (2.16%)	237 (1.99%)	240 (1.92%)	218 (1.81%)
M	100 (0.98%)	119 (1%)	129 (1.03%)	100 (0.83%)
N	2 (0.02%)	1 (0.01%)	1 (0.01%)	2 (0.02%)
O	564 (5.53%)	607 (5.1%)	637 (5.11%)	608 (5.06%)
P	203 (1.99%)	233 (1.96%)	232 (1.86%)	209 (1.74%)
Q	464 (4.55%)	550 (4.62%)	610 (4.89%)	488 (4.06%)
S	2259 (22.16%)	2714 (22.8%)	2845 (22.82%)	2501 (20.81%)
T	416 (4.08%)	436 (3.66%)	444 (3.56%)	417 (3.47%)
U	364 (3.57%)	389 (3.27%)	395 (3.17%)	381 (3.17%)
V	47 (0.46%)	62 (0.52%)	66 (0.53%)	56 (0.47%)
W	5 (0.05%)	9 (0.08%)	9 (0.07%)	7 (0.06%)
Y	24 (0.24%)	26 (0.22%)	27 (0.22%)	25 (0.21%)

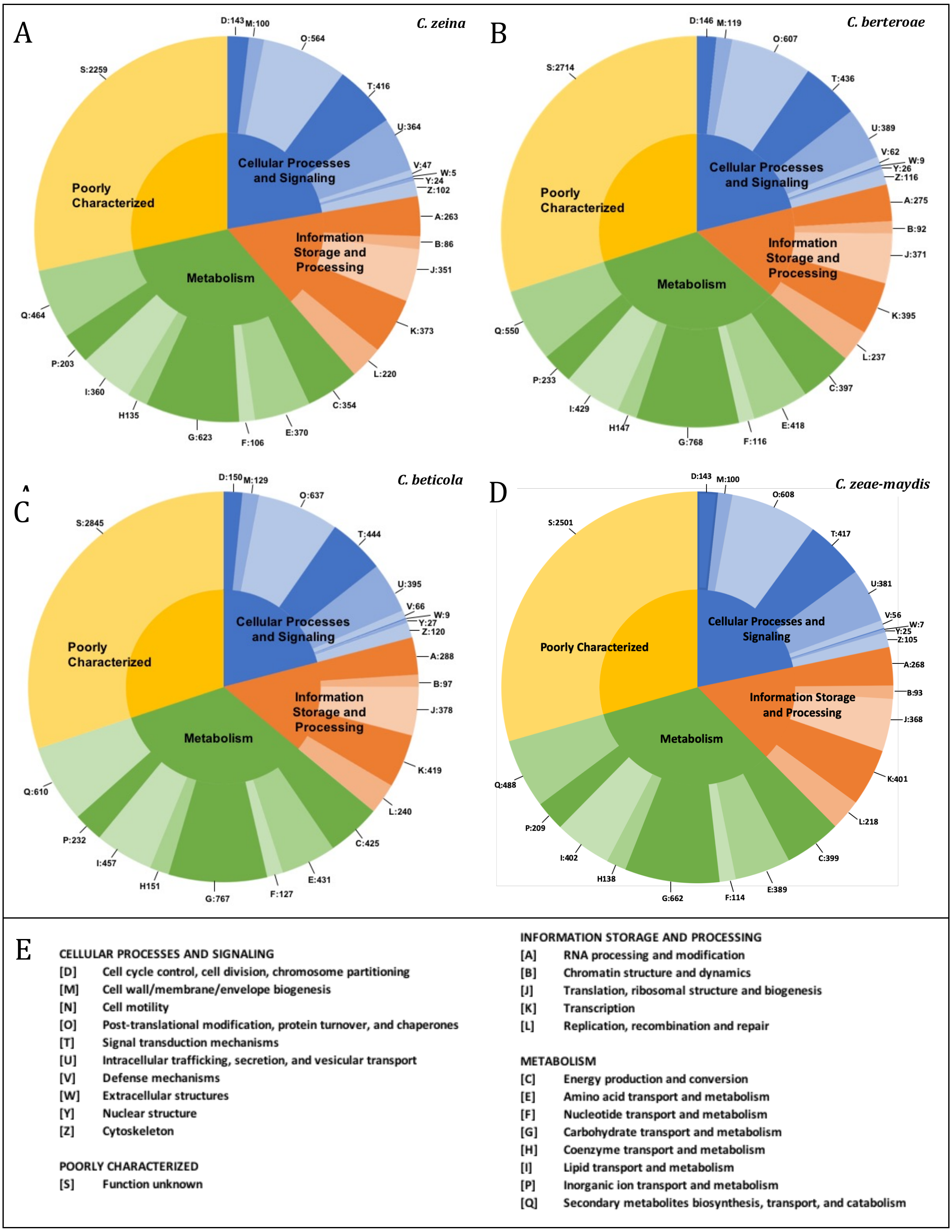


Figure 3.7 Classification of *Cercospora* species proteins into KOG database groups. *Proteins are classified into the main groups of metabolism, information storage and processing, cellular processes and signaling, or poorly characterized proteins. The classification is shown for* C. zeina *(**A**),* C. berteroae *(**B**),* C. beticola *(**C**), and* C. zeae-maydis *(**D**), with the number of the subgroups, indicated in (**E**) shown on the outer rim of the respective charts.*

3.4.10.3 Gene ontology

Gene ontology descriptions were predicted for each protein during the InterProScan analysis (Section 3.4.10.1). A total of 5,576 (55% of genome total) proteins from *C. zeina* were assigned with at least one GO term, with 6,306 (53% of genome total) assigned for *C. berteroae*, 6,686 (54% of genome total) assigned for *C. beticola* and 5,914 (49% of genome total) assigned for *C. zeae-maydis* respectively (Table 3.9).

Table 3.9 Numbers of GO namespace element predictions for the *Cercospora* species. *Numbers represent all GO-terms predicted for each species for the respective namespace elements, and include proteins which have multiple GO-terms assigned. Percentage values indicate the ratio of GO-terms relative to the total number of proteins predicted for each species*

Main GO namespace elements	*C. zeina*	*C. zeae-maydis*	*C. berteroae*	*C. beticola*
Biological process	3,616 (35%)	3,894 (32%)	4,393 (37%)	5,069 (41%)
Cellular component	1,671 (16%)	1,774 (15%)	1,983 (17%)	2,168 (17%)
Molecular function	5,725 (56%)	6,010 (50%)	6,882 (58%)	7,778 (62%)

A total of 1,182 GO-terms are common to all the species, while each species had multiple unique GO-terms. In addition there are 10 GO-terms shared by the maize-infecting species, while there are 83 unique GO-terms shared in the sugar beet-infecting species (Figure 3.8).

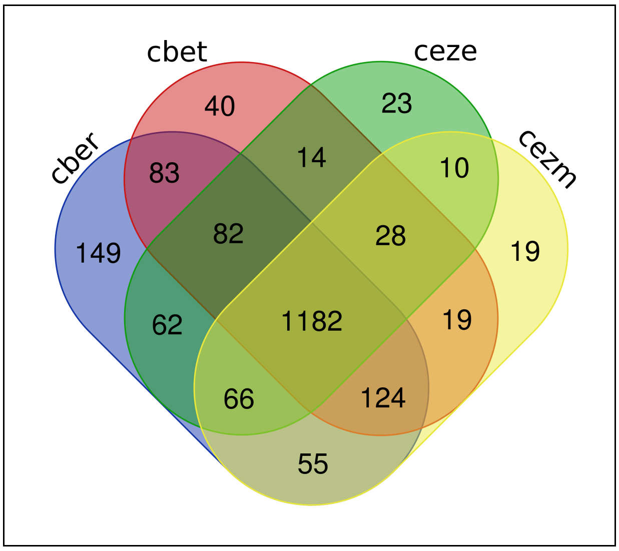


Figure 3.8 Venn-diagram of Gene Ontology category number comparison between the *Cercospora* species. *All predicted GO ontologies are included, with only unique GO numbers within species represented. The classification is shown for* C. berteroae *(**cber**),* C. beticola *(**cbet**),* C. zeina *(**ceze**), and* C. zeae-maydis *(**cezm**).*

3.4.10.4 Carbohydrate-Active Enzymes

The CAZyme prediction indicated a total of 437 CAZyme classes predicted for *C. zeina*, with 531 predicted for *C. berteroae*, 545 predicted for *C. beticola* and 469 predicted for *C. zeae-maydis* respectively (Table 3.10). Some 114 CAZyme classes were common to all species, with the sugar beet-infecting species sharing 10 unique classes, while there were no unique classes shared by the maize-infecting species (Figure 3.9). The total numbers of genes per subclass for each species is listed in the Appendix (Table A4).

Table 3.10 Numbers per class of CAZyme predicted for *Cercospora* species. *Percentages in parentheses indicate the total number of CAZyme class numbers as percentage of the total gene content of the species.*

CAZyme class	***C. zeina***	***C. zeae-maydis***	***C. berteroae***	***C. beticola***
Carbohydrate-Binding Modules	25 (0.24%)	30 (0.24%)	37 (0.31%)	38 (0.3%)
Carbohydrate esterases	96 (0.94%)	96 (0.77%)	110 (0.92%)	119 (0.95%)
Glycoside hydrolases	208 (2.04%)	230 (1.84%)	249 (2.09%)	253 (2.03%)
Glycosyl transferases	102 (1%)	108 (0.87%)	126 (1.06%)	128 (1.03%)
Polysaccharide lyases	6 (0.06%)	5 (0.04%)	9 (0.08%)	7 (0.06%)

The output of the CAZyme prediction showed some significant differences between the respective species, with 8 CAZyme classes being present in the sugar beet-infecting species while being absent in the maize-infecting species (Table 3.11). The missing classes might represent a difference in infection strategy between the four species, though none of these enzyme classes are enriched for functions related to cell wall components more prevalent in dicots. *C. zeae-maydis* appears to contain fewer CAZymes relative to the total gene content that the other species.

Table 3.11 CAZyme class member numbers only present in sugar beet-infecting *Cercospora* species.

CAZyme class	***C. berteroae***	***C. beticola***	***C. zeina***	***C. zeae-maydis***
Carbohydrate-Binding Modules CBM4	1	1	0	0
Carbohydrate esterase CE7	1	2	0	0
Glycoside hydrolase GH106	1	2	0	0
Glycoside hydrolase GH33	1	1	0	0
Glycoside hydrolase GH42	1	1	0	0
Glycoside hydrolase GH88	1	2	0	0
Glycosyl transferase GT91	2	2	0	0
Polysaccharide lyase PL22	2	1	0	0

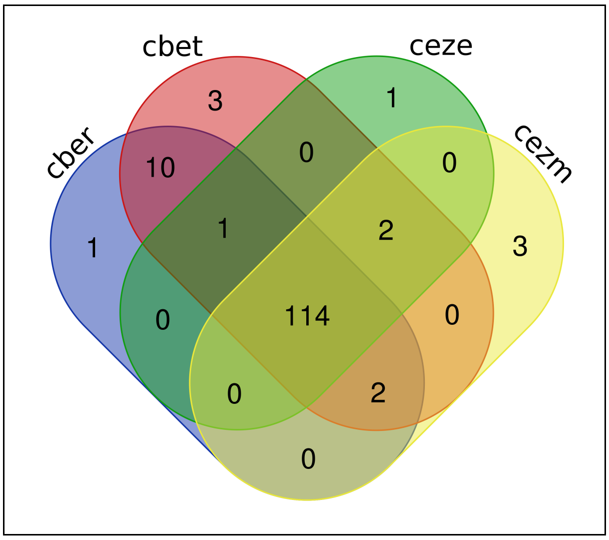


Figure 3.9 Venn-diagram of predicted carbohydrate-active enzyme categories in the *Cercospora* species. *All predicted CAZyme families and sub-families are included, with only unique designations within species represented. The classification is shown for* C. berteroae *(**cber**),* C. beticola *(**cbet**),* C. zeina *(**ceze**), and* C. zeae-maydis *(**cezm**).*

3.4.10.5 Secondary metabolite production genes

Polyketide synthases (PKSs) and non-ribosomal peptide synthetases (NRPSs) can function as the initial steps of the synthesis of secondary metabolites, and were predicted using the SMURF webserver. Outputs grouped the predicted proteins into one of four functional classes, i.e. NRPS, NRPS-like, PKS and PKS-like. *C. beticola* contains the largest number of predicted proteins, while the maize-infecting species have similar numbers of proteins (Table 3.12)

Table 3.12 Numbers of Polyketide synthases and Non-ribosomal peptide synthetases predicted in the *Cercospora* species.

	C. zeina	***C. zeae-maydis***	***C. berteroae***	***C. beticola***
PKS	9	11	15	16
PKS-Like	2	2	1	2
NRPS	9	7	11	16
NRPS-Like	11	8	13	19

The InterProScan results (Section 3.4.10.1) were used to obtain the Pfam domains present in the proteins for the predicted classes. NRPS and NRPS-like classes shared 4 Pfam domains, while PKS and PKS-like classes shared 11 Pfam domains (Table 3.13).

Table 3.13 Predicted Pfam domains present in predicted Polyketide synthases and Non-ribosomal peptide synthetases of the *Cercospora* species. *Grey shaded blocks indicate the presence of the domain in the species, and white blocks denote the absence of the domain.*

Pfam	Domain description	*cber*	*cbet*	*ceze*	*cezm*
	Non-ribosomal peptide synthetases				
PF13193	AMP-binding enzyme C-terminal domain	■	■	■	■
PF00550	Phosphopantetheine attachment site	■	■	■	■
PF00668	Condensation domain	■	■	■	■
PF00501	AMP-binding enzyme	■	■	■	■
	Non-ribosomal peptide synthetase-like				
PF13193	AMP-binding enzyme C-terminal domain	■	■	■	■
PF00550	Phosphopantetheine attachment site	■	■	■	■
PF07993	Male sterility protein	■	■	■	■
PF00668	Condensation domain	■	■	■	■
PF00501	AMP-binding enzyme	■	■	■	■
PF00106	Short chain dehydrogenase	■	■	■	■
PF07690	Major Facilitator Superfamily		■		
PF13641	Glycosyltransferase like family 2		■		
	Polyketide synthases				
PF08659	KR domain	■	■	■	■
PF00668	Condensation domain	■	■	■	■
PF14765	Polyketide synthase dehydratase	■	■	■	■
PF16073	Starter unit:ACP transacylase in aflatoxin biosynthesis	■	■	■	■
PF00109	Beta-ketoacyl synthase, N-terminal domain	■	■	■	■
PF00975	Thioesterase domain	■	■	■	■
PF00550	Phosphopantetheine attachment site	■	■	■	■
PF08242	Methyltransferase domain	■	■	■	■
PF08240	Alcohol dehydrogenase GroES-like domain	■	■	■	■
PF16197	Ketoacyl-synthetase C-terminal extension	■	■	■	■
PF02801	Beta-ketoacyl synthase, C-terminal domain	■	■	■	■
PF00107	Zinc-binding dehydrogenase	■	■	■	■
PF13602	Zinc-binding dehydrogenase	■	■	■	■
PF00698	Acyl transferase domain	■	■	■	■
PF00501	AMP-binding enzyme	■		■	■
PF13193	AMP-binding enzyme C-terminal domain			■	■
	Polyketide synthase-like				
PF02801	Beta-ketoacyl synthase, C-terminal domain	■	■	■	■
PF00109	Beta-ketoacyl synthase, N-terminal domain	■	■	■	■
PF14765	Polyketide synthase dehydratase		■	■	■
PF00698	Acyl transferase domain		■	■	■
PF16197	Ketoacyl-synthetase C-terminal extension		■	■	■
PF08659	KR domain		■		■
PF08242	Methyltransferase domain		■		■
PF00550	Phosphopantetheine attachment site		■		■
PF13602	Zinc-binding dehydrogenase		■		
PF03824	High-affinity nickel-transport protein			■	
PF00107	Zinc-binding dehydrogenase				■

Terpene synthases are more difficult to predict since the enzymes involved are highly variable, and intermediate products can be modified by various enzymes to yield a variety of secondary metabolites. Additionally the enzymes are usually part of a gene cluster (Khaldi *et al.*, 2010), and this is the rationale behind the antiSMASH prediction HMM. For *C. zeina* 27 terpene synthases were predicted, with 27 predicted for *C. berteroae*, 26 predicted for *C. beticola* and 24 predicted for *C. zeae-maydis* respectively. In order to show the functional elements in the proteins involved in these pathways, the Pfam domain content of the genes were extracted, and the presence/absence reported (Table 3.14). The domains with unknown functions were removed from the table.

Table 3.14 Predicted Pfam domains present in predicted terpene synthases of the *Cercospora* species. *Grey shaded blocks indicate the presence of the domain in the species, and white blocks denote the absence of the domain. Results are shown for C. berteroae (cber), C. beticola (cbet), C. zeina (ceze) and C. zeae-,maydis (cezm).*

Pfam	Description	*cber*	*cbet*	*ceze*	*cezm*
PF00067	Cytochrome P450	■	■	■	■
PF00144	Beta-lactamase	■	■	■	■
PF00155	Aminotransferase class I and II	■	■	■	■
PF00168	C2 domain				■
PF00225	Kinesin motor domain	■	■	■	■
PF00296	Luciferase-like monooxygenase	■	■	■	■
PF00327	Ribosomal protein L30p/L7e	■		■	
PF00348	Polyprenyl synthetase	■	■	■	■
PF00397	WW domain				■
PF00494	Squalene/phytoene synthase	■	■	■	■
PF00632	HECT-domain (ubiquitin-transferase)	■	■	■	■
PF01036	Bacteriorhodopsin-like protein	■	■	■	■
PF01170	Putative RNA methylase family UPF0020	■	■		■
PF01248	Ribosomal protein L7Ae/L30e/S12e/Gadd45 family	■	■	■	■
PF01593	Flavin containing amine oxidoreductase	■	■	■	■
PF03055	Retinal pigment epithelial membrane protein	■	■	■	■
PF03676	Uncharacterised protein family (UPF0183)	■	■	■	■
PF05183	RNA dependent RNA polymerase	■	■	■	■
PF05653	Magnesium transporter NIPA	■	■	■	■
PF05978	Ion channel regulatory protein UNC-93	■	■	■	
PF06046	Exocyst complex component Sec6	■	■		■
PF07690	Major Facilitator Superfamily	■	■		■
PF07885	Ion channel		■	■	
PF08318	COG4 transport protein	■	■	■	■
PF12861	Anaphase-promoting complex subunit 11 RING-H2 finger	■	■	■	
PF13931	Kinesin-associated microtubule-binding	■	■	■	■

3.4.10.6 Secreted lipases

Secreted lipases have been shown to be implicated in plant pathogenicity. The different classes of lipases were predicted using BLAST similarity searches to the different classes of the Lipase Engineering Database. The resultant dataset was screened for the presence of a signal peptide which might indicate secretion. The secreted lipase content of the *C. zeina* protein set is predicted to be 96, with 107 predicted for *C. berteroae*, 133 predicted for *C. beticola* and 73 predicted for *C. zeae-maydis* respectively (Table 3.15). There was one class specific for the sugar beet-infecting species, classified as deacetylases. There were no maize-infecting species-specific classes predicted.

Table 3.15 Number and superfamilies of secreted lipases predicted for *Cercospora* species.

Lipase superfamilies	*C. zeina*	*C. zeae-maydis*	*C. berteroae*	*C. beticola*
Acyl-transferases	13	9	13	17
Carboxylesterases	14	10	14	17
Cutinase	10	5	10	12
Cytosolic hydrolases	14	10	14	17
Deacetylases	0	0	1	1
Dipeptidyl peptidase IV-like	0	3	3	3
Filamentous fungi lipases	3	2	3	3
Hormone sensitive lipases	15	10	14	21
Hydroxynitrile lyases	0	2	2	2
Lipoprotein lipases	0	0	0	1
Lysophospholipase	13	10	16	18
Microsomal hydrolases	13	9	13	16
Prolyl endopeptidases	1	2	3	4
Thioesterases	0	1	1	1

3.4.10.7 Secreted protease prediction

Secreted proteases may play roles in the degradation of host plant tissues and the digestion of proteins involved in the plant response against pathogens (Ohm *et al.*, 2012), and as such are important in studying the fungal host response. For *C. zeina* a total of 131 small secreted proteases were predicted, while the numbers for the other species were 155 for *C. berteroae*, 96 for *C. beticola* and 162 for *C. zeae-maydis* (Table 3.16). The number of secreted proteases per class for each species is listed in the Appendix (Table A5).

Table 3.16 Number of secreted protease families predicted for *Cercospora* species.

Protease families	*C. zeina*	*C. zeae-maydis*	*C. berteroae*	*C. beticola*
Aspartic proteases	12	13	14	14
Cysteine proteases	24	3	3	2
Metallo proteases	41	77	61	34
Serine proteases	50	65	72	45
Threonine proteases	4	4	5	1

From Table 3.16 it is apparent that *C. zeina* contains an excess of secreted cysteine proteases compared to the other *Cercospora* species, and if functionally confirmed this could indicate an important role in plant pathogenesis. The proteome of *C. beticola* appeared to be depleted of metallo and serine proteases when compared to *C. beteroae* which shares its host.

3.5 Discussion

The study yielded a gene annotation that is 95.4% complete in terms of conserved core gene content, with a predicted 10,193 genes present in *C. zeina*. The annotation has been used in a published functional study on the proteins involved in the cercosporin biosynthesis cluster (Swart *et al.*, 2017). The comparison with the gene content of the related *C. zeae-maydis*, *C. berteroae* and *C. beticola* gene annotations show a range of total predicted gene numbers, with the numbers for *C. zeina* being lower than the other species, although in a similar size order. The functional content of the genes is similar in terms of carbohydrate active enzymes, secondary metabolite producing enzymes, secreted proteases, secreted lipases and secreted small proteins. Differences in classes and numbers of genes might reflect differences in lifestyle among the four species.

Ab initio gene prediction algorithms require training on organism-specific gene subsets to provide the most accurate gene model prediction. For closely related species the prediction models can be transferred, but training on accurate genes from the species of interest will provide the highest accuracy. Therefore the availability of a high-quality set of genes from an organism is crucial, and usually entails the initial step in the genome annotation process. For the identification of possible gene coordinates the mapping of high-confidence data to the genome is critical. In this study the proteins and transcripts of the related species, *C. zeae-maydis* were used due to the close phylogenetic relationship, and the assumption that the gene models would be similar in number and structure. To supplement the *C. zeae-maydis* data, the UniProt Swiss-Prot database was also mapped to provide high-confidence, manually curated protein information from multiple organisms. The mapping of RNAseq data is important for the identification of intron positions, and the *in vitro* cultured *C. zeina* RNAseq data was used. Finally the *C. zeina* Trinity transcriptome assembly mapping was also used to provide additional information regarding exon-intron boundaries.

Graphical user interface genome browsers with gene coordinate editing capabilities are very important in the manual gene prediction process. In this case the GenomeView software package (Abeel *et al.*, 2012) was used (Figure 3.1), providing the flexibility to upload and visualize all the data formats and perform coordinate and open reading frame editing, but also to perform BLAST similarity searches directly from the interface. The mapped data provided approximate gene position coordinates, but often excluded intron positions or did not provide positions for all exons of a specific gene. Gene structures were identified by manually changing each open reading frame to conform to the correct intron splice/donor sites and start and stop sites. BLAST similarity searches to know proteins provided additional information on missed exons, as well as the correct positioning of start and stop codons and introns. Unfortunately the excessive reliance on the similarity of gene structures to known information has the potential to force the resultant gene structures to possibly incorrect models, especially if closely related species' information is not available, or was generated with a similar method (Danchin *et al.*, 2018). Since protein sequence similarity is enriched for active centers and regions crucial for folding, there are risks in obtaining mapping to incomplete genes. Arbitrarily extending the 5’ and 3’ regions to identify the correct start and stop codons could artificially force the gene models larger, and exclude functional units such as signal peptides (Haridas *et al.*, 2018). For the approach in the study it was decided to exclude genes for which there were no BLAST similarity to proteins with known function (hypothetical and unknown proteins), therefore excluding possible incorrect gene structures from decreasing the quality of the training data set and the efficiency and accuracy of the gene predictors.

Ab initio gene predictors require high quality training datasets to optimize the mathematical descriptions of their HMM since they usually function without additional evidence such as EST data. The efficiency of the SNAP *ab initio* gene predictor was evaluated with the first MAKER prediction analysis. The first 145 manually predicted genes served as training set, and the resultant MAKER output, relying solely on the SNAP predictions as well as the *C. zeae-maydis* transcript information yielded 10,447 gene models. As rule of thumb, an annotation with >95% genes with a MAKER AED score of <0.5 is indicative of a good annotation (Campbell *et al.*, 2014). In this case the annotation showed 96% of genes with AED <0.5, so in theory the annotation completeness was satisfactory, though the SNAP training was performed with a limited dataset and had to be expanded. The SNAP prediction output was manually evaluated on a further 50 genes on genome assembly contigs not previously manually annotated. The resultant gene coordinates were added to the SNAP gene training file to update the training HMM. Training the SNAP HMM more than twice does not improve the accuracy, and runs the risk of overtraining the HMM which can lead to a decrease in the prediction accuracy (Campbell *et al.*, 2014). Since it is recommended to use multiple gene predictors it was decided to also use AUGUSTUS in the MAKER prediction pipeline. The training of the AUGUSTUS *ab initio* gene predictor requires several hundred high quality gene models. In this study the initial 195 manually curated gene models were used as training set, but the result was a 100% false-positive quality metric output, indicative of an insufficient number of training genes. The use of gene predictions from the initial MAKER analysis for training AUGUSTUS was attempted, using genes with AED scored between 0 - 0.2, since these genes have the highest prediction confidence (Campbell *et al.*, 2014). The second training set consisted of 3,506 such high confidence genes. The subsequent training process yielded a quality metric output with a sensitivity of 0.45 which was accurate for further gene prediction analyses (Stanke *et al.*, 2006). The AUGUSTUS HMM was therefore not re-trained or updated subsequently.

The Genemark-ES *ab initio* gene predictor is different from AUGUSTUS and SNAP since it does not need a training data set, but only requires the genome of the organism. The program was developed on fungal data, and is equipped to predict gene models in fungal genomes (Borodovsky & Lomsadze, 2011). The training of Genemark-ES in this study completed successfully and was included in the second MAKER analysis. The output of the analysis yielded no predicted gene models, and logfiles indicated conflict with GeneMark-ES. The predictor was subsequently removed for the analysis workflow, with improved results. Genemark-ES was not used in the MAKER pipeline as replacement for either of the other predictors due to the confidence in the gene prediction accuracies of SNAP and AUGUSTUS, and their lack of conflict during the MAKER analysis.

The final MAKER analysis, using SNAP and AUGUSTUS as gene predictors and the Trinity transcriptome assembly provided an annotation with 10,339 gene models. The output showed 98.4% of predicted genes with an AED score <0.5, and this was therefore an improvement in the accuracy of gene prediction compared to the initial analysis. Since the number of gene models was in a similar range to related species, i.e. *S. musiva* (10,233) and *C. zeae-maydis* (12,020), the automated gene prediction was concluded. During the preparation of the annotation data for Genbank upload some problems were identified with the gene predictions which were manually corrected. Due to fungi having overlapping UTR regions (Haridas *et al.*, 2018), MAKER artificially extended the majority of UTRs to overlap genes on the opposite strand, and these were removed from the annotation. In addition multiple genes were incorrectly concatenated to form artificially large genes, while several large genes were incorrectly split into closely adjacent small genes, while some predicted gene models spanned gaps in the genome assembly. Following the correction, the number of gene models were reduced to 10,193, which was still in the same size range as related species.

Similar to its use in evaluating genome assemblies, the BUSCO completeness analysis tool can be used to evaluate the completeness of an annotated protein data set (Simao *et al.*, 2015). BUSCO completeness analyses were performed for *C. zeina*, *C. zeae-maydis*, *C. berteroae* and *C. beticola* to compare the genome annotations for the related species. The four genomes were assembled to difference sequence depths and completeness levels and annotated with different methods and with different levels of manual curation. The gene content of the genomes would therefore be different in regards to the gene content (Table 3.3). All the species showed a BUSCO completeness >95% (Table 3.5), and for *C. zeina* this was identical to the genome assembly completeness analysis. The annotation did have the largest percentage of missing genes, which would proportionately explain the smaller number of predicted genes in the genome, and might be the result of an incomplete genome assembly.

The cercosporin biosynthesis cluster of genes have been studied in *C. zeina* and *C. zeae-maydis* due to the hypothesized importance of the toxin in plant pathogenesis (Swart, 2017). The presence of the cluster in the genome of *C. zeina* was confirmed using the homology of *C. nicotianae* CTB genes to obtain gene coordinates, followed by manual curation and cDNA sequencing. The gene cluster was also identified in the other species (including *C. nicotianae*), with manual gene structure confirmation (Figure 3.2 and Figure 3.3). The synteny of the genes in all species are identical (Figure 3.4). The curation of the genes in *C. zeina* indicated that CTB7 is a pseudogene, supporting previous studies confirming the absence of the toxin *in vitro*. The role of the toxin in *C. zeina* plant pathogenesis is still being studied. One hypothesis involves a paralog of CTB7 taking part in the synthesis pathway, though no paralog has been found during this study. Alternatively cercosporin might not be produced by *C. zeina* at all, while another toxin or pathogenesis agent is produced to enable plant infection (Swart *et al.*, 2017).

The ortholog inference analysis provided a list of genes which have orthologs in the respective species, as well as paralogs. Orthologs were clustered into orthologous groups based on sequence similarity, and the majority of these orthogroups contain at least one ortholog from each species, with 6,641 groups containing a single gene from each species (Figure 3.5). These single gene orthogroups can possibly be useful for refining phylogenetic relationships. Interesting to note is the very low single-species orthologs, therefore the species share orthologs for most genes. There are, however some groups only shared between the maize-infecting species and others only between the sugar beet-infecting species. The gene content of these orthogroups might be analyzed in detail to explain lifestyle differences between the two groups.

Functional annotation of the predicted proteins is difficult and normally incomplete due to the lack of data for all proteins. The use of prediction pipelines to infer functional units in proteins based on experimental evidence is one of the first types of analyses performed, and the InterProScan pipeline is a common tool. Multiple predictors are used to infer these functional units, which could be signal peptides, trans-membrane regions, protein domains, orthologous gene clusters, etc. The functional annotation of all four species was performed in tandem to allow direct comparison of the functional content of the species. Multiple classifications can be obtained for each protein depending on the domains and other functional units present. The InterProScan data was parsed and mined for various types of data. Most relevant was the SignalP prediction output which predicted the presence of signal peptides, suggesting the secretion of the relevant protein. Of equal importance was the predicted Pfam domain content, which indicates the presence of functional domains which can be used to classify proteins into specific functional classes, e.g. proteases. The classification of each protein into functional Gene Ontology classes is of secondary importance for this study due to the reliance on additional functional prediction methods. However, a total of 1,182 GO-terms were found to be common for all the species, with each species having multiple unique GO-terms. In terms of host-specific differences, the maize-infecting species shared 10 unique GO-terms, while the sugar beet-infecting species shared 83 unique GO-terms (Table 3.8 and Figure 3.8). Further classification of these shared GO-terms might provide a functional basis for the differences in lifestyle for these species. In a similar trend, the prediction of Eukaryotic Orthologous Groups for each species provided a general classification of proteins into different functional classes, but the numbers for the species are very similar and very few obvious differences can be observed. These differences become even less when taking the different numbers of predicted proteins for each species into account and presenting the KOG numbers as percentages of the total predicted proteins for each species (Table 3.8). Of more insightful use is the classification of proteins into more specialized functional classes which might play a role in plant pathogenesis such as small secreted proteins, carbohydrate active enzymes, secondary metabolite producing proteins, secreted lipases and secreted proteases (Ohm *et al.*, 2012) (Figure 3.10).

Small secreted proteins have been studied for their importance in fungal plant pathogenesis (Martin *et al.*, 2008 ; Stergiopoulos & de Wit, 2009). These proteins were predicted in all species, with the numbers in the sugar beet-infecting species being higher. This might be an indication that these proteins are more important for pathogenesis of the dicotyledonous host relative to the monocotyledonous maize host. The small secreted proteins are also not functionally well characterized due to the lower Pfam domain content of these proteins (from 17% - 26%) relative to the rest of the genomes (63% - 69%). The SSPs specific for the sugar beet-infecting species include Pfam domains present in bacteria, and might therefore be prediction artifacts. There is a similar situation when looking at the *C. zeina*-specific Pfam domains, with the only domain of interest being part of the arginase family which play important roles in arginine/agmatine metabolism, the urea cycle, histidine degradation, and other pathways. It is not clear whether members of this family are typically secreted, and might be a false positive prediction as well.

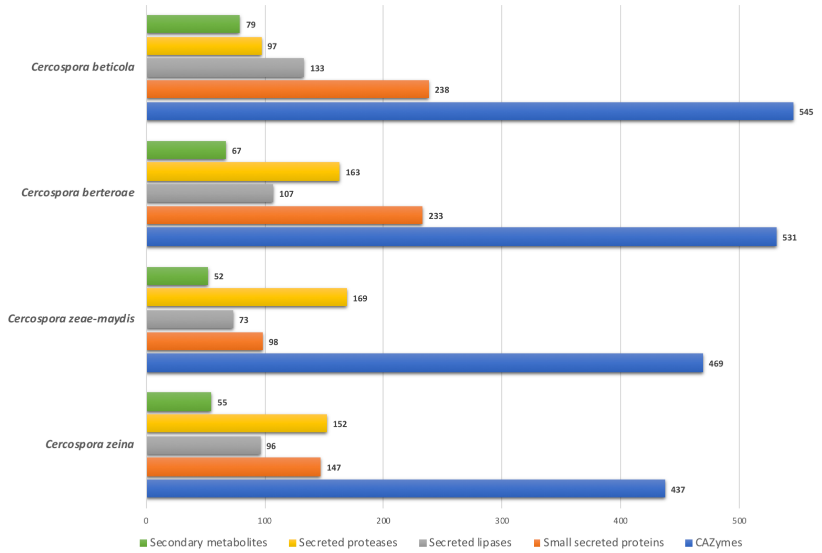


Figure 3.10 Important functional classes of proteins predicted for the *Cercospora* species. *The X-axis indicates the number of genes. The number of genes for each category are indicated next to the respective bars.*

Carbohydrate-active enzymes are involved with all facets of carbohydrate metabolism, catabolism and transport. They are especially important in fungal plant pathogens due to the role for this class of proteins in enabling pathogenesis due to the destabilization of the plant cell wall. Additionally these proteins are involved with extracting energy from the carbohydrate complement of the host cells. There are five functional classes of CAZymes, including glycoside hydrolases, glycosyltransferases, polysaccharide lyases, carbohydrate esterases and carbohydrate-binding families. The last class is considered to be non-catalytic, but is usually included due to the association with catalytic classes. A large variety of CAZyme classes were predicted for each species (Table 3.10) and Appendix Table A4), with the numbers per class also provided as a percentage of the total gene content. *C. zeae-maydis* appears to contain fewer CAZymes than the other species. The obvious CAZyme class to consider for pathogenesis is the cellulose-degrading classes, i.e. GH61, GH6, GH7, GH45 and CMB1, and the predicted datasets are severely depleted for these enzymes, with only one enzyme for each species in the GH7 class. This suggests either that these species don't rely on cellulose degradation for pathogenesis, or that they use different strategies to degrade cellulose. This is not an isolated result, since it has been shown that other members of the *Capnodiales* also show depletion for these classes of enzymes in their gene set (Ohm *et al.*, 2012). Regarding the degradation of xylan, the two main xylanase families GH10 and GH11 are represented in all species to identical numbers, while the two acetylxylan esterase families CE1 and CE3 show broadly similar numbers with no trend difference between the maize-infecting and sugar beet-infecting species. There is a similar trend for pectin degradation in the pectate lyases families PL1 and PL3 and the pectin methylesterases family CE8, with similar or identical numbers for all species. These similar profiles for these important degradation mechanisms might not be important to explain some of the host-specific requirements for pathogenesis between the species. There are eight classes with components only present in the sugar beet-infecting species (Table 3.11 and Table 3.17), and a functional study with these proteins might explain the specific requirements for infecting dicotyledonous plants.

Table 3.17 Descriptions of CAZyme classes unique to sugar beet-infecting *Cercospora* species.

CAZyme class	Activity Description
Carbohydrate-Binding Modules CBM4	Binding xylan, β-1,3-glucan, β-1,3-1,4-glucan, β-1,6-glucan
Carbohydrate esterase CE7	Acetyl xylan esterase, cephalosporin-C deacetylase
Glycoside hydrolase GH106	α-L-rhamnosidase
Glycoside hydrolase GH33	Sialidase or Neuraminidase
Glycoside hydrolase GH42	β-galactosidase, α-L-arabinopyranosidase
Glycoside hydrolase GH88	d-4,5-unsaturated β-glucuronyl hydrolase
Glycosyl transferase GT91	β-1,2-mannosyltransferase
Polysaccharide lyase PL22	Oligogalacturonate lyase / oligogalacturonide lyase

There was only one CAZyme class unique to *C. zeina*, i.e. Polysaccharide lyase class PL6 with alginate lyase and chondroitinase B activity. Since alginate is not produced in maize and chondroitin is present in most cells, the presence of this predicted functional unit in *C. zeina* is most probably not pathogenesis-related.

Fungal secondary metabolite producing genes are very diverse, and are often localized in clusters transcriptionally co-regulated by the same genetic mechanisms and factors (Sarikaya-Bayram *et al.*, 2015). The prediction of especially terpene synthases challenging due to the diversity in the gene sequences and structures, but the antiSMASH and SMURF prediction webservers provide comprehensive results based on analysis of genome assemblies and proteomes respectively. The SMURF webserver provided predictions of polyketide synthases and non-ribosomal peptide synthases content of the species, and the sugar beet-infecting species contained the most predicted genes, while the maize-infecting species contained broadly similar numbers of these genes (Table 3.12). The Pfam domain content of these genes in the species are also very conserved (Table 3.13), with only *C. beticola* containing two unique domains for NRPSs, i.e. the Major Facilitator Superfamily and Glycosyltransferase like family 2. The comparative Pfam domain content in the PKSs are more varied (Table 3.14) with the sugar beet-infecting species being deficient in AMP binding domains. *C. zeina* contains a unique PKS-like domain, the high-affinity nickel-transport protein which is involved in the incorporation of nickel into the H_2-uptake hydrogenase enzymes, and is essential for the expression of catalytically active hydrogenase which could also play a possible role in terpene synthase (Fu *et al.*, 1994). The terpene synthases are much more diverse that the PKSs and

NRPSs and therefore a prediction HMM was developed to find these gene clusters from the genome sequences. The prediction of complete biosynthesis clusters are rare, and depend on the presence of the genes in the species of interest, as well as the predictive information provided by closely related species. For the *Cercospora* species there were no complete clusters predicted, with only selected components of these clusters present. For *C. beticola* there were five TPS clusters predicted and for *C. berteroa* there were four, while *C. zeina* contained three and *C. zeae-maydis* four clusters. The numbers of predicted genes in these clusters were reported and the numbers were very similar for all the species, though there were differences in Pfam domain content for all species. *C. zeina* had more absent domains compared to the other species, possibly indicating that these TPS genes were enriched for a number of similar genes.

The breakdown and penetration of the plant cuticle is a crucial process in fungal pathogenesis (Rogers *et al.*, 1994 ; Voigt *et al.*, 2005 ; Ohm *et al.*, 2012) and is accomplished by the secretion of cutinases and lipases which break down the hydrophobic lipid-based polymers coating plant leaves (Yeats & Rose, 2013). The Lipase Engineering Database contains sequence data for a wide range of lipase classes, and 14 of these classes are conserved in the secreted protein sets of the *Cercospora* species. Though there is variation in the numbers for each species (Table 3.15), the two most relevant classes are the cutinases and deacetylases. All the *Cercospora* species contained multiple secreted cutinase orthologs, suggesting that the class of enzymes is important for pathogenesis, hypothetically through cuticle degradation for conidial attachment, though this has not been confirmed for *C. zeina*. The importance of the deacetylase class lies in its absence in the secreted protein set of the maize-infecting species. A study on endophytic fungi showed that chitin deacetylases play a role in hiding these fungi from the plant immune system (Cord-Landwehr *et al.*, 2016). This result might reflect a difference in the infection and plant immune system evasion strategy for pathogenic fungi of dicotyledonous hosts.

The role of secreted proteases in pathogenesis are widespread, and include the degradation of host tissue for infection and propagation, as well as the digestion of host immune system proteins. The five main classes, i.e. Aspartic, Cysteine, Metallo, Serine and Threonine proteases are all present in all the secreted protein set of the *Cercospora* species (Table 3.16 and Appendix Table A5), though with variation in numbers. *C. beticola* appears to be depleted for secreted proteases based on the numbers of members of each class present, though this does not appear to be a sugar beet-infection phenomenon, since *C. berteroae* contains a large contingent of secreted protease proteins. Of interest is the enrichment of *C. zeae-maydis* for metalloproteases, with 77 members of the class present. The enrichment in the carboxypeptidase A1 (M14) and leucine aminopeptidase-1 (M28) classes explain this result, though these classes are not enriched in *C. zeina*. In addition *C. zeina* is enriched for the carboxypeptidase (S10) class, with more members than the other species combined. There are also two aspartic proteases (A11 and A28) unique to the *C. zeina* secreted protein set, although it is unclear whether these differences are essential for plant pathogenesis.

The *C. zeina* genome annotation process yielded a total of 10,193 gene models which were functionally classified. The gene content was compared to other *Cercospora* species and there were similarities in the type and numbers of functional units present in the gene set. The orthology between the different species will be expanded in Chapter 4 to include other *Dothideomycete* species to find genes which can be used to refine the phylogenetic classification of the *Cercospora* species complex.

3.6 References

Abeel, T., Van Parys, T., Saeys, Y., Galagan, J., and Van de Peer, Y. (2012) GenomeView: a next-generation genome browser. *Nucleic Acids Research* **40(2)**:e12

Adachi, N., and Lieber, M. R. (2002) Bidirectional gene organization: a common architectural feature of the human genome. *Cell* **109(7)**:807-809

Akiva, E., Brown, S., Almonacid, D. E., Barber, A. E., 2nd, Custer, A. F., Hicks, M. A., Huang, C. C., Lauck, F., Mashiyama, S. T., Meng, E. C., Mischel, D., Morris, J. H., Ojha, S., Schnoes, A. M., Stryke, D., Yunes, J. M., Ferrin, T. E., Holliday, G. L., and Babbitt, P. C. (2014) The structure-function linkage database. *Nucleic Acids Research* **42(Database issue)**:D521-D530

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215(3)**:403-410

antiSMASH-bacterial. https://antismash.secondarymetabolites.org/.

antiSMASH-fungal. https://fungismash.secondarymetabolites.org.

Aravin, A. A., Sachidanandam, R., Bourc'his, D., Schaefer, C., Pezic, D., Toth, K. F., Bestor, T., and Hannon, G. J. (2008) A piRNA pathway primed by individual transposons is linked to *de novo* DNA methylation in mice. *Molecular Cell* **31(6)**:785-799

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet* **25(1)**:25-29

Attwood, T. K., Coletta, A., Muirhead, G., Pavlopoulou, A., Philippou, P. B., Popov, I., Roma-Mateo, C., Theodosiou, A., and Mitchell, A. L. (2012) The PRINTS database: a fine-grained protein sequence annotation and analysis resource - its status in 2012. *Database (Oxford)* **2012**:bas019

Biscotti, M. A., Olmo, E., and Heslop-Harrison, J. S. (2015) Repetitive DNA in eukaryotic genomes. *Chromosome Research* **23(3)**:415-420

Biscotti, M. A., Canapa, A., Forconi, M., Olmo, E., and Barucca, M. (2015) Transcription of tandemly repetitive DNA: functional roles. *Chromosome Research* **23(3)**:463-477

BLAST. https://blast.ncbi.nlm.nih.gov.

Blobel, G., and Dobberstein, B. (1975) Transfer of proteins across membranes. I. Presence of proteolytically processed and unprocessed nascent immunoglobulin light chains on membrane-bound ribosomes of murine myeloma. *Journal of Cell Biology* **67(3)**:835-851

Borodovsky, M., and Lomsadze, A. (2011) Eukaryotic gene prediction using GeneMark.hmm-E and GeneMark-ES. *Current Protocols in Bioinformatics* **Chapter 4**:Unit 4.6.1-4.6.10

Buels, R., Yao, E., Diesh, C. M., Hayes, R. D., Munoz-Torres, M., Helt, G., Goodstein, D. M., Elsik, C. G., Lewis, S. E., Stein, L., and Holmes, I. H. (2016) JBrowse: a dynamic web platform for genome visualization and analysis. *Genome Biology* **17**:66

Burge, C., and Karlin, S. (1997) Prediction of complete gene structures in human genomic DNA. *Journal of Molecular Biology* **268(1)**:78-94

Cacho, R. A., Tang, Y., and Chooi, Y. H. (2014) Next-generation sequencing approach for connecting secondary metabolites to biosynthetic gene clusters in fungi. *Frontiers in Microbiology* **5**:774

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014) Genome annotation and curation using MAKER and MAKER-P. *Current Protocols in Bioinformatics* **48**:4.11.1-4.11.39

Campbell, M. S., Law, M., Holt, C., Stein, J. C., Moghe, G. D., Hufnagel, D. E., Lei, J., Achawanantakun, R., Jiao, D., Lawrence, C. J., Ware, D., Shiu, S. H., Childs, K. L., Sun, Y., Jiang, N., and Yandell, M. (2014) MAKER-P: a tool kit for the rapid creation, management, and quality control of plant genome annotations. *Plant Physiology* **164(2)**:513-524

Cantarel, B. L., Korf, I., Robb, S. M., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., and Yandell, M. (2008) MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. *Genome Research* **18(1)**:188-196

CAZy. http://www.cazy.org/

Cord-Landwehr, S., Melcher, R. L., Kolkenbrock, S., and Moerschbacher, B. M. (2016) A chitin deacetylase from the endophytic fungus *Pestalotiopsis* sp. efficiently inactivates the elicitor activity of chitin oligomers in rice cells. *Scientific Reports* **6**:38018

Curwen, V., Eyras, E., Andrews, T. D., Clarke, L., Mongin, E., Searle, S. M., and Clamp, M. (2004) The Ensembl automatic gene annotation system. *Genome Research* **14(5)**:942-950

Cyverse. http://www.cyverse.org/.

Danchin, A., Ouzounis, C., Tokuyasu, T., and Zucker, J. D. (2018) No wisdom in the crowd: genome annotation in the era of big data - current status and future prospects. *Microbial Biotechnology* **11(4)**:588-605

Dawson, N. L., Lewis, T. E., Das, S., Lees, J. G., Lee, D., Ashford, P., Orengo, C. A., and Sillitoe, I. (2017) CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research* **45(D1)**:D289-D295

dbCAN. https://cys.bios.niu.edu/dbCAN2/.

Dolfini, D., Zambelli, F., Pavesi, G., and Mantovani, R. (2009) A perspective of promoter architecture from the CCAAT box. *Cell Cycle* **8(24)**:4127-4137

Eddy, S. R. (2011) Accelerated profile HMM searches. *PLoS Computational Biology* **7(10)**:e1002195

emapper. https://eggnogdb.embl.de/app/emapper.

Emms, D. M., and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**:157

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J., and Bateman, A. (2016) The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44(D1)**:D279-D285

Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H. Y., Dosztanyi, Z., El-Gebali, S., Fraser, M., Gough, J., Haft, D., Holliday, G. L., Huang, H., Huang, X., Letunic, I., Lopez, R., Lu, S., Marchler-Bauer, A., Mi, H., Mistry, J., Natale, D. A., Necci, M., Nuka, G., Orengo, C. A., Park, Y., Pesseat, S., Piovesan, D., Potter, S. C., Rawlings, N. D., Redaschi, N., Richardson, L., Rivoire, C., Sangrador-Vegas, A., Sigrist, C., Sillitoe, I., Smithers, B., Squizzato, S., Sutton, G., Thanki, N., Thomas, P. D., Tosatto, S. C., Wu, C. H., Xenarios, I., Yeh, L. S., Young, S. Y., and Mitchell, A. L. (2017) InterPro in 2017-beyond protein family and domain annotations. *Nucleic Acids Research* **45(D1)**:D190-D199

Fischer, M., and Pleiss, J. (2003) The Lipase Engineering Database: a navigation and analysis tool for protein families. *Nucleic Acids Research* **31(1)**:319-321

Fourel, G., Magdinier, F., and Gilson, E. (2004) Insulator dynamics and the setting of chromatin domains. *Bioessays* **26(5)**:523-532

Fu, C., Javedan, S., Moshiri, F., and Maier, R. J. (1994) Bacterial genes involved in incorporation of nickel into a hydrogenase enzyme. *Proceedings of the National Academy of Sciences of the United States of America* **91(11)**:5099-5103

Gagniuc, P., and Ionescu-Tirgoviste, C. (2012) Eukaryotic genomes may exhibit up to 10 generic classes of gene promoters. *BMC Genomics* **13**:512

Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talon, M., Dopazo, J., and Conesa, A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite. *Nucleic Acids Research* **36(10)**:3420-3435

Gough, J., Karplus, K., Hughey, R., and Chothia, C. (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *Journal of Molecular Biology* **313(4)**:903-919

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith, R. K., Jr., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L., and White, O. (2003) Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research* **31(19)**:5654-5666

Haft, D. H., Selengut, J. D., and White, O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Research* **31(1)**:371-373

Haridas, S., Salamov, A., and Grigoriev, I. V. (2018) Fungal genome annotation. *Methods in Molecular Biology* **1775**:171-184

Harish, A., and Caetano-Anolles, G. (2012) Ribosomal history reveals origins of modern protein synthesis. *PLoS ONE* **7(3)**:e32776

Heslop-Harrison, J. S., and Schmidt, T. 2001. Plant nuclear genome composition. Pages 1-8. in: Encyclopedia of Life Sciences. John Wiley and Sons.

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016) BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32(5)**:767-769

Hood, L., and Rowen, L. (2013) The Human Genome Project: big science transforms biology and medicine. *Genome Medicine* **5(9)**:79

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., and Bork, P. (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Molecular Biology and Evolution* **34(8)**:2115-2122

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C., and Bork, P. (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research* **44(D1)**:D286-D293

International Human Genome Sequencing, C. (2004) Finishing the euchromatic sequence of the human genome. *Nature* **431(7011)**:931-945

InterPro. http://www.ebi.ac.uk/interpro.

Jady, B. E., Richard, P., Bertrand, E., and Kiss, T. (2006) Cell cycle-dependent recruitment of telomerase RNA and Cajal bodies to human telomeres. *Molecular Biology of the Cell* **17(2)**:944-954

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S. Y., Lopez, R., and Hunter, S. (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30(9)**:1236-1240

Kalia, R. K., Rai, M. K., Kalia, S., Singh, R., and Dhawan, A. K. (2011) Microsatellite markers: an overview of the recent progress in plants. *Euphytica* **177(3)**:309-334

Kall, L., Krogh, A., and Sonnhammer, E. L. L. (2004) A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology* **338(5)**:1027-1036

Kanehisa, M., and Goto, S. (2000) KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Research* **28(1)**:27-30

Kapustin, Y., Souvorov, A., Tatusova, T., and Lipman, D. (2008) Splign: algorithms for computing spliced alignments with identification of paralogs. *Biology Direct* **3**:20

Kejnovsky, E., Tokan, V., and Lexa, M. (2015) Transposable elements and G-quadruplexes. *Chromosome Research* **23(3)**:615-623

Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423(6937)**:241-254

Kellis, M., Wold, B., Snyder, M. P., Bernstein, B. E., Kundaje, A., Marinov, G. K., Ward, L. D., Birney, E., Crawford, G. E., Dekker, J., Dunham, I., Elnitski, L. L., Farnham, P. J., Feingold, E. A., Gerstein, M., Giddings, M. C., Gilbert, D. M., Gingeras, T. R., Green, E. D., Guigo, R., Hubbard, T., Kent, J., Lieb, J. D., Myers, R. M., Pazin, M. J., Ren, B., Stamatoyannopoulos, J. A., Weng, Z., White, K. P., and Hardison, R. C. (2014) Defining functional DNA elements in the human genome. *Proceedings of the National Academy of Sciences of the United States of America* **111(17)**:6131-6138

Khaldi, N., Seifuddin, F. T., Turner, G., Haft, D., Nierman, W. C., Wolfe, K. H., and Fedorova, N. D. (2010) SMURF: Genomic mapping of fungal secondary metabolite clusters. *Fungal Genetics and Biology* **47(9)**:736-741

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* **14(4)**:R36

Kishore, S., and Stamm, S. (2006) The snoRNA HBII-52 regulates alternative splicing of the serotonin receptor 2C. *Science* **311(5758)**:230-232

Kitts, P. (2003) The NCBI Handbook. National Center for Biotechnology Information.

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

Korf, I., Flicek, P., Duan, D., and Brent, M. R. (2001) Integrating genomic homology into gene structure prediction. *Bioinformatics* **17 Suppl 1**:S140-S148

Koyanagi, K. O., Hagiwara, M., Itoh, T., Gojobori, T., and Imanishi, T. (2005) Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. *Gene* **353(2)**:169-176

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* **305(3)**:567-580

Langdon, W. B. (2015) Performance of genetic programming optimised Bowtie2 on genome comparison and analytic testing (GCAT) benchmarks. *BioData Mining* **8(1)**:1

Lawson, M. J., and Zhang, L. (2006) Distinct patterns of SSR distribution in the *Arabidopsis thaliana* and rice genomes. *Genome Biology* **7(2)**:R14

LED. http://www.led.uni-stuttgart.de/.

Letunic, I., and Bork, P. (2017) 20 years of the SMART protein domain annotation resource. *Nucleic Acids Research* **46(D1)**: D493-D496

Lewis, S. E., Searle, S. M., Harris, N., Gibson, M., Lyer, V., Richter, J., Wiel, C., Bayraktaroglu, L., Birney, E., Crosby, M. A., Kaminker, J. S., Matthews, B. B., Prochnik, S. E., Smithy, C. D., Tupy, J. L., Rubin, G. M., Misra, S., Mungall, C. J., and Clamp, M. E. (2002) Apollo: a sequence annotation editor. *Genome Biology* **3(12)**: research0082.1–82.14

Lodish, H. F. (2000) Molecular cell biology. 4th ed. W.H. Freeman, New York.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014) The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research* **42(Database issue)**:D490-D495

Lundin, M., Nehlin, J. O., and Ronne, H. (1994) Importance of a flanking AT-rich region in target site recognition by the GC box-binding zinc finger protein MIG1. *Molecular Cell Biology* **14(3)**:1979-1985

Lupas, A., Vandyke, M., and Stock, J. (1991) Predicting coiled coils from protein sequences. *Science* **252(5009)**:1162-1164

Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., Gwadz, M., Hurwitz, D. I., Lu, F., Marchler, G. H., Song, J. S., Thanki, N., Wang, Z., Yamashita, R. A., Zhang, D., Zheng, C., Geer, L. Y., and Bryant, S. H. (2017) CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. *Nucleic Acids Research* **45(D1)**:D200-D203

Martin, F., Aerts, A., Ahren, D., Brun, A., Danchin, E. G., Duchaussoy, F., Gibon, J., Kohler, A., Lindquist, E., Pereda, V., Salamov, A., Shapiro, H. J., Wuyts, J., Blaudez, D., Buee, M., Brokstein, P., Canback, B., Cohen, D., Courty, P. E., Coutinho, P. M., Delaruelle, C., Detter, J. C., Deveau, A., DiFazio, S., Duplessis, S., Fraissinet-Tachet, L., Lucic, E., Frey-Klett, P., Fourrey, C., Feussner, I., Gay, G., Grimwood, J., Hoegger, P. J., Jain, P., Kilaru, S., Labbe, J., Lin, Y. C., Legue, V., Le Tacon, F., Marmeisse, R., Melayah, D., Montanini, B., Muratet, M., Nehls, U., Niculita-Hirzel, H., Oudot-Le Secq, M. P., Peter, M., Quesneville, H., Rajashekar, B., Reich, M., Rouhier, N., Schmutz, J., Yin, T., Chalot, M., Henrissat, B., Kues, U., Lucas, S., Van de Peer, Y., Podila, G. K., Polle, A., Pukkila, P. J., Richardson, P. M., Rouze, P., Sanders, I. R., Stajich, J. E., Tunlid, A., Tuskan, G., and Grigoriev, I. V. (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature.* **452(7183)**:88-92

Martoglio, B. (2003) Intramembrane proteolysis and post-targeting functions of signal peptides. *Biochemical Society Transactions* **31(Pt 6)**:1243-1247

Maston, G. A., Evans, S. K., and Green, M. R. (2006) Transcriptional regulatory elements in the human genome. *Annual Review of Genomics and Human Genetics* **7**:29-59

Mehta, G. D., Agarwal, M. P., and Ghosh, S. K. (2010) Centromere identity: a challenge to be faced. *Molecular Genetics and Genomics* **284(2)**:75-94

MEROPS. http://www.ebi.ac.uk/merops.

Mi, H., Muruganujan, A., and Thomas, P. D. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research* **41(Database issue)**:D377-D386

NCBI. http://www.ncbi.nlm.nih.gov.

Nielsen, H. (2017) Predicting secretory proteins with SignalP. *Methods in Molecular Biology* **1611**:59-73

Ohm, R. A., Feau, N., Henrissat, B., Schoch, C. L., Horwitz, B. A., Barry, K. W., Condon, B. J., Copeland, A. C., Dhillon, B., Glaser, F., Hesse, C. N., Kosti, I., LaButti, K., Lindquist, E. A., Lucas, S., Salamov, A. A., Bradshaw, R. E., Ciuffetti, L., Hamelin, R. C., Kema, G. H., Lawrence, C., Scott, J. A., Spatafora, J. W., Turgeon, B. G., de Wit, P. J., Zhong, S., Goodwin, S. B., and Grigoriev, I. V. (2012) Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PLoS Pathogens* **8(12)**:e1003037

Pan, Q., Shai, O., Lee, L. J., Frey, B. J., and Blencowe, B. J. (2008) Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. *Nature Genetics* **40(12)**:1413-1415

Panaccione, D. G., Scottcraig, J. S., Pocard, J. A., and Walton, J. D. (1992) A cyclic peptide synthetase gene required for pathogenicity of the fungus *Cochliobolus carbonum* on maize. *Proceedings of the National Academy of Sciences of the United States of America* **89(14)**:6590-6594

Pedruzzi, I., Rivoire, C., Auchincloss, A. H., Coudert, E., Keller, G., de Castro, E., Baratin, D., Cuche, B. A., Bougueleret, L., Poux, S., Redaschi, N., Xenarios, I., and Bridge, A. (2015) HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Research* **43(Database issue)**:D1064-D1070

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011) SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* **8(10)**:785-786

Piovesan, D., Tabaro, F., Paladin, L., Necci, M., Micetic, I., Camilloni, C., Davey, N., Dosztanyi, Z., Meszaros, B., Monzon, A. M., Parisi, G., Schad, E., Sormanni, P., Tompa, P., Vendruscolo, M., Vranken, W. F., and Tosatto, S. C. E. (2018) MobiDB 3.0: more annotations for intrinsic disorder, conformational diversity and interactions in proteins. *Nucleic Acids Research* **46(D1)**:D471-D476

Polyak, K., and Meyerson, M. (2003) Cancer medicine. 6 ed. BC Decker Inc.

Ratsch, G., Sonnenburg, S., Srinivasan, J., Witte, H., Muller, K. R., Sommer, R. J., and Scholkopf, B. (2007) Improving the *Caenorhabditis elegans* genome annotation using machine learning. *PLoS Computational Biology* **3(2)**:e20

Rawlings, N. D., Barrett, A. J., Thomas, P. D., Huang, X., Bateman, A., and Finn, R. D. (2018) The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Research* **46(D1)**:D624-D632

Reiner, R., Ben-Asouli, Y., Krilovetzky, I., and Jarrous, N. (2006) A role for the catalytic ribonucleoprotein RNase P in RNA polymerase III transcription. *Genes & Development* **20(12)**:1621-1635

Robinson, J. T., Thorvaldsdottir, H., Winckler, W., Guttman, M., Lander, E. S., Getz, G., and Mesirov, J. P. (2011) Integrative genomics viewer. *Nature Biotechnology* **29(1)**:24-26

Rogers, L. M., Flaishman, M. A., and Kolattukudy, P. E. (1994) Cutinase gene disruption in *Fusarium solani* f. sp. *pisi* decreases its virulence on pea. *Plant Cell* **6(7)**:935-945

Rutherford, K., Parkhill, J., Crook, J., Horsnell, T., Rice, P., Rajandream, M. A., and Barrell, B. (2000) Artemis: sequence visualization and annotation. *Bioinformatics* **16(10)**:944-945

Salamov, A. A., and Solovyev, V. V. (2000) *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Research* **10(4)**:516-522

Salzberg, S. L., Pertea, M., Delcher, A. L., Gardner, M. J., and Tettelin, H. (1999) Interpolated Markov models for eukaryotic gene finding. *Genomics* **59(1)**:24-31

Santos, F. C., Guyot, R., do Valle, C. B., Chiari, L., Techio, V. H., Heslop-Harrison, P., and Vanzela, A. L. (2015) Chromosomal distribution and evolution of abundant retrotransposons in plants: gypsy elements in diploid and polyploid *Brachiaria* forage grasses. *Chromosome Research* **23(3)**:571-582

Sarikaya-Bayram, O., Palmer, J. M., Keller, N., Braus, G. H., and Bayram, O. (2015) One Juliet and four Romeos: VeA and its methyltransferases. *Frontiers in Microbiology* **6**:1

Scarpato, M., Angelini, C., Cocca, E., Pallotta, M. M., Morescalchi, M. A., and Capriglione, T. (2015) Short interspersed DNA elements and miRNAs: a novel hidden gene regulation layer in zebrafish? *Chromosome Research* **23(3)**:533-544

Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D., and Kahn, D. (2002) ProDom: automated clustering of homologous domains. *Briefings in Bioinformatics* **3(3)**:246-251

Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., and Bucher, P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Briefings in Bioinformatics* **3(3)**:265-274

Simao, F. A., Waterhouse, R. M., Ioannidis, P., Kriventseva, E. V., and Zdobnov, E. M. (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics* **31(19)**:3210-3212

Slater, G. S., and Birney, E. (2005) Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics.* **6**:31

Sleator, R. D. (2010) An overview of the current status of eukaryote gene prediction strategies. *Gene* **461(1-2)**:1-4

Smale, S. T., and Kadonaga, J. T. (2003) The RNA polymerase II core promoter. *Annual Reviews in Biochemistry* **72**:449-479

Smit, A. F. A., Hubley, R., and Green, P. (2013–2015) RepeatMasker Open-4.0. http://www.repeatmasker.org

SMURF. http://www.jcvi.org/smurf.

Souvorov, A. (2010) Gnomon - the NCBI eukaryotic gene prediction tool. National Center for Biotechnology Information.

Stanke, M., Keller, O., Gunduz, I., Hayes, A., Waack, S., and Morgenstern, B. (2006) AUGUSTUS: *ab initio* prediction of alternative transcripts. *Nucleic Acids Research* **34(Web Server issue)**:W435-W439

Stein, L. (2001) Genome annotation: from sequence to biology. *Nature Reviews Genetics* **2(7)**:493-503

Stein, L. D. (2013) Using GBrowse 2.0 to visualize and share next-generation sequence data. *Briefings in Bioinformatics* **14(2)**:162-171

Sterck, L., Billiau, K., Abeel, T., Rouze, P., and Van de Peer, Y. (2012) ORCAE: online resource for community annotation of eukaryotes. *Nature Methods* **9(11)**:1041

Stergiopoulos, I., and de Wit, P. J. (2009) Fungal effector proteins. *Annual Reviews in Phytopathology* **47**:233-263

Swart, V. (2017) Functional genomics of the cercosporin biosynthetic gene cluster in the maize pathogen *Cercospora zeina*. *PhD Thesis, University of Pretoria, Pretoria, South Africa*

Swart, V., Crampton, B. G., Ridenour, J. B., Bluhm, B. H., Olivier, N. A., Meyer, J. J. M., and Berger, D. K. (2017) Complementation of CTB7 in the maize pathogen *Cercospora zeina* overcomes the lack of *in vitro* cercosporin production. *Molecular Plant-Microbe Interactions* **30(9)**:710-724

Tatusov, R. L., Galperin, M. Y., Natale, D. A., and Koonin, E. V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* **28(1)**:33-36

tbl2asn2. http://www.ncbi.nlm.nih.gov/genbank/tbl2asn2.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., and Borodovsky, M. (2008) Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. *Genome Research* **18(12)**:1979-1990

The Gene Ontology, C. (2017) Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Research* **45(D1)**:D331-D338

Treangen, T. J., and Salzberg, S. L. (2011) Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews Genetics* **13(1)**:36-46

Trinklein, N. D., Aldred, S. F., Hartman, S. J., Schroeder, D. I., Otillar, R. P., and Myers, R. M. (2004) An abundance of bidirectional promoters in the human genome. *Genome Research* **14(1)**:62-66

UniProt Consortium, T. (2018) UniProt: the universal protein knowledgebase. *Nucleic Acids Research* **46(5)**:2699

Voigt, C. A., Schafer, W., and Salomon, S. (2005) A secreted lipase of *Fusarium graminearum* is a virulence factor required for infection of cereals. *The Plant Journal* **42(3)**:364-375

Weber, T., Blin, K., Duddela, S., Krug, D., Kim, H. U., Bruccoleri, R., Lee, S. Y., Fischbach, M. A., Muller, R., Wohlleben, W., Breitling, R., Takano, E., and Medema, M. H. (2015) antiSMASH 3.0-a comprehensive resource for the genome mining of biosynthetic gene clusters. *Nucleic Acids Research* **43(W1)**:W237-W243

Wei, W., Pelechano, V., Jarvelin, A. I., and Steinmetz, L. M. (2011) Functional consequences of bidirectional promoters. *Trends in Genetics* **27(7)**:267-276

Wheelan, S. J., Church, D. M., and Ostell, J. M. (2001) Spidey: a tool for mRNA-to-genomic alignments. *Genome Research* **11(11)**:1952-1957

Wingfield, B. D., Berger, D. K., Steenkamp, E. T., Lim, H. J., Duong, T. A., Bluhm, B. H., de Beer, Z. W., De Vos, L., Fourie, G., Naidoo, K., Olivier, N., Lin, Y. C., Van de Peer, Y., Joubert, F., Crampton, B. G., Swart, V., Soal, N., Tatham, C., van der Nest, M. A., van der Merwe, N. A., van Wyk, S., Wilken, P. M., and Wingfield, M. J. (2017) IMA Genome-F 8: Draft genome of *Cercospora zeina*, *Fusarium pininemorale*, *Hawksworthiomyces lignivorus*, *Huntiella decipiens* and *Ophiostoma* ips. *IMA Fungus* **8(2)**:385-396

Wu, C. H., Nikolskaya, A., Huang, H., Yeh, L. S., Natale, D. A., Vinayaka, C. R., Hu, Z. Z., Mazumder, R., Kumar, S., Kourtesis, P., Ledley, R. S., Suzek, B. E., Arminski, L., Chen, Y., Zhang, J., Cardenas, J. L., Chung, S., Castro-Alvear, J., Dinkov, G., and Barker, W. C. (2004) PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Research* **32(Database issue)**:D112-D114

Yandell, M., and Ence, D. (2012) A beginner's guide to eukaryotic genome annotation. *Nature Reviews Genetics* **13(5)**:329-342

Yang, G., Rose, M. S., Turgeon, B. G., and Yoder, O. C. (1996) A polyketide synthase is required for fungal virulence and production of the polyketide T-toxin. *Plant Cell* **8(11)**:2139-2150

Yeats, T. H., and Rose, J. K. (2013) The formation and function of plant cuticles. *Plant Physiology* **163(1)**:5-20

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., and Xu, Y. (2012) dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research* **40(Web Server issue)**:W445-W451

Zhang, A. T., Langley, A. R., Christov, C. P., Kheir, E., Shafee, T., Gardiner, T. J., and Krude, T. (2011) Dynamic interaction of Y RNAs with chromatin and initiation proteins during human DNA replication. *Journal of Cell Science* **124(Pt 12)**:2058-2069

Chapter 4

Selection of *Cercospora*-specific genes for phylogenetic species classification

N. A. Olivier, Y-C. Lin, Y. Van de Peer, R. de Jonge, O. Reva, D. K. Berger

Department of Plant and Soil Sciences, Faculty of Natural and Soil Sciences, Forestry and Agricultural Research Institute, University of Pretoria, Pretoria, 0002, South Africa

All analyses and molecular biology procedures were performed by N. A. Olivier. Dr Y-C. Lin from the laboratory of Prof Y. Van de Peer provided invaluable assistance during the planning of the experimental approach. Dr R. de Jonge assisted with the implementation and interpretation of the ortholog analysis. Prof O. Reva provided assistance during the phylogenetic analysis of the orthologs to find informative sequences, assisted with the primer design criteria and also the critical revision of the manuscript. Prof D. K. Berger assisted with the primer design criteria and primer design quality evaluation, interpreting the molecular biology results and critical revision of the manuscript. He also assisted with funding for a research visit to the laboratory of Prof. Y. Van de Peer where the planning and conception of the ortholog analysis were implemented.

4.1 Abstract

The *Cercospora* genus in the Class *Dothideomycetes* contains >3,000 species, the majority of which are plant pathogens. In the genus there are 19 species for which no definitive classification can be made based on either morphological characteristics or the standard phylogenetic marker sequences (Groenewald *et al.*, 2013). The need for new phylogenetic marker genes prompted the analysis of 25 fungal proteomes to detect orthologous proteins unique to the *Cercospora* species. The orthologs were analyzed for their phylogenetic information content, and based on functional description criteria eight genes were selected for further analysis. Degenerate primers amplifying regions with amino acid identity in four Cercospora species were designed for four of the genes, and the primers evaluated by PCR and sequencing on *C. zeina* and *C. zeae-maydis* genomic DNA. The primers of the Glucoamylase 1 and Alkaline protease 1 genes were found to amplify the correct gene regions in *C. zeina* and *C. zeae-maydis*, and are ready for use in the analysis of additional *Cercospora* species to attempt the resolution of the species classification challenges.

4.2. Introduction

Species phylogenies which are derived from single-gene comparisons show significant inconsistencies due the presence of horizontal gene transfer, paralogous sequences and a variable rate of evolution for different organisms and genes (Snel *et al.*, 1999). The use of whole genomes for the construction of these phylogenies have proven successful, especially since the large number of genes used in the analysis ensure robust phylogenetic relationships (Aguileta *et al.*, 2008). The large number of genes is also valuable since the number of genes two organisms have in common depends on their evolutionary distance (Snel *et al.*, 1999). Unfortunately there are too few genome sequences available to make these whole genome phylogenies the standard method of phylogenetic construction. The importance of informative genomic loci which can be used for phylogenies, and for which the sequences can be obtained through single step amplifications and sequencing, can therefore not be over-emphasized.

4.2.1 DNA barcoding

DNA barcoding is an approach that uses short genetic markers to identify an organism as belonging to a specific species (Hebert *et al.*, 2003). The approach does not seek to infer evolutionary relationships as is the case with molecular phylogeny (Kress *et al.*, 2005). Different gene regions are generally used for different Kingdoms as selected by the respective committees of the Consortium for the Barcode of Life (CBOL) due to their applicability in the respective biological systems. For animals and many other eukaryotes the mitochondrial Cytochrome C oxidase subunit I (COI) gene is used (Pentinsaari *et al.*, 2016), while for fungi the internally transcribed spacer (ITS) region is more informative (Schoch *et al.*, 2012). Plants currently require the concatenation of the chloroplast loci *rbcL* and *matK*, although additional regions, such as ITS, are being evaluated to provide more complete resolution for land plants (China Plant BOL Group *et al.*, 2011). In practice, there are four criteria a gene region needs to satisfy to be considered as a barcode, i.e. 1) the region should contain sufficient sequence variability between species, 2) should have low intra-species variability, 3) should be short enough to be sequenced during a single sequencing reaction, and 4) should contain a conserved sequence region amongst species for the design of universal primers (Savolainen *et al.*, 2005 ; Stielow *et al.*, 2015). Barcoding studies have shown that morphological classifications are not sufficient, since new species have been uncovered while analyzing known morphologically classified specimens. The concept of the “barcoding gap” was subsequently introduced to define new species based on the barcoding sequence, and was defined as the mean interspecific variation should by 10 times exceed the mean intraspecific variation for the group under study (Hebert, P. D. *et al.*, 2004 ; Candek & Kuntner, 2015).

4.2.2 Fungal barcoding and classification genes

A number of fungal genes have been used for the phylogenetic classification of fungal species as well as their evolutionary relationships. The ITS region, though sufficient for separating a large number of fungal species, does not provide the required resolution for genera consisting of a large number of species due to the insufficient sequence variation in the genera ITS region. Other loci considered for barcoding included the small (SSU) and large (LSU) ribosomal subunits, DNA-directed RNA polymerase II subunit rpb1 (RPB1), DNA-directed RNA polymerase II subunit rpb2 (RPB2) and the mini-chromosome maintenance complex component 7 (MCM7). The protein coding genes are important in fungal classification since they are selected as single copy genes in the fungal genomes, and also accumulate less mutations in their exonic regions, thus maintaining their lengths. In addition they also contain introns which sometimes have a faster evolution rate and are useful for resolving higher taxonomic levels. Additional protein coding genes used in past classifications include the translation elongation factor 1-alpha (TEF1), beta-tubulin (tub2/BenA), DNA topoisomerase 1 (TOPI), Phosphoglycerate kinase (PGK) and the LNS2 protein (Raja *et al.*, 2017).

4.2.3 Phylogenetic informativeness

The informativeness of characters used in phylogenetic is crucial for resolving several phylogenetic controversies, i.e. which types of characters are more informative; are increased taxonomic or character sampling more informative; and can historical polytomies resulting from rapid radiations be accurately identified (Townsend, 2007). Different parts of the same tree might exhibit differing evolutionary rates, thereby increasing uncertainty of the actual tree branch lengths and increasing uncertainty of the accuracy of the species inference from these trees (Aguileta *et al.*, 2008). Since some genes have higher phylogenetic information content, the use of only a few of these can be used to construct representative phylogenies, while minimizing the number of genes that have to be sequenced (Aguileta *et al.*, 2008). Several measures have been proposed to characterize informativeness, including the tree-length skew (Huelsenbeck, 1991), consistency index (Farris, 1989) and profiling informativeness for explicit historical epochs (Townsend, 2007) amongst others.

4.2.3.1 Tree-length skew

A single sequence might contain bases that are invariate and might thus provide an unambiguous alignment, while highly variable bases in the same sequence might become saturated with evolutionary changes and the signal essentially becomes random. The total phylogenetic signal in the sequence will thus be masked by the random noise, and the best tree will not be a good indication of the true phylogeny. This problem of distinguishing between data containing phylogenetic structure and data with excess sequence changes masking true phylogenetic signal has, amongst others, been examined by assessing the amount of skew in the distribution of tree lengths. For random data, or data that support multiple hypotheses, the tree length distribution has been found to be symmetrical and the optimal tree was to be found anywhere in the distribution. The optimal tree could even be close to the most non-parsimonious tree. For data supporting only one hypothesis the tree length distribution was found to be skewed to the left, and the real tree was close to the most parsimonious tree. The G1 statistic was defined as a measure of the skewness of a distribution, with a perfectly symmetrical distribution having a G1 statistic value of 0, while a left-skewed distribution has a negative G1 statistic value (Hillis, 1991 ; Huelsenbeck, 1991 ; Hillis & Huelsenbeck, 1992).

4.2.3.2 Consistency index

The consistency index was introduced to measure the consistency of a tree to the data, a measurement of the amount of homoplasy in a system. The index also reflects the number of taxa in a dataset, as well as the degree to which each base carries phylogenetic information. The index is calculated by counting the minimum number of changes in a dataset, and dividing it by the actual number of sequences required for a tree, or even a single base (Farris, 1989).

4.2.3.3 Profiling phylogenetic informativeness

The approach attempts to quantify the informativeness of a base over a historic timescale, since characters can show different rates of evolution at different times in history (Lopez-Giraldez *et al.*, 2013). It is imperative to use genes which evolve at a pace that is appropriate to resolve ancient branching. In addition, if a polytomy is identified, the required data should include sequence data for taxa that branch close to the polytomy, as well as new sequence data that are informative for the relevant time period. The profiling of phylogenetic informativeness requires rates of evolution for each site of a locus of interest. The prior information can be obtained from 1) a well-studied subset of the taxa of interest, 2) data from a well-studied sister-clade, or 3) comparative genomic data. Studies show that the metric can be used to assess whether all bases in a good candidate gene should be included in the phylogenetic analyses, or whether a smaller subset, which is easier to sequence, would provide the information required for constructing the correct trees. In addition, several smaller genes/gene regions, again at smaller sequencing cost, might be more informative than a long, established sequence (Townsend, 2007 ; Lopez-Giraldez *et al.*, 2013).

4.2.4 Classification gene mining approach

Classically genes or genomic regions used for species classification were selected due to practical or historical criteria. The use of a limited or unsuitable gene-set might be problematic since classification based on gene-trees are not always congruent with species-trees, and might not reflect true evolutionary relationships (Aguileta *et al.*, 2008). The availability of additional, informative genes is therefore critical. Current strategies for mining new classification genes require the availability of total proteomes for the genus under study, obtained from whole genome sequencing projects. From these proteomes single-copy orthologs are inferred across all species, and subsequent filter and selection criteria used to select a subset suitable for use. It is crucial for only single-copy orthologs to be used in the analysis, since the presence of paralogs will hinder the correct phylogenetic reconstruction of the genus (Koonin, 2005). The use of single copy orthologs in searching for new phylogenetic marker genes is well documented (Aguileta *et al.*, 2008 ; Lopez-Giraldez *et al.*, 2013 ; Stielow *et al.*, 2015 ; Raja *et al.*, 2017).

Orthologs are defined in terms of the evolutionary history of genes, with two genes being orthologs if they arose from a single gene in the last common ancestor (LCA) of the species in question. For instances where the LCA genome was shown to contain at least two paralogs, the concept of orthology is considered to be incorrect. Although it is common for orthologs to perform similar functions, the concept of "functional orthologs" is not applicable according to the evolutionary definition of orthology. A case in point is where similar functions are performed by proteins which have no orthologous relationship. Therefore the concept of orthology must be strictly defined in terms of evolutionary relationships, and inferring these relationships starts at a sequence similarity basis, but must include a phylogenetic analysis step, usually followed by a tree reconciliation for final orthology assessment (Koonin, 2005 ; Aguileta *et al.*, 2008). A consequence of the definition of orthology is that the phylogenetic tree of a set of orthologs has by definition the same topology as the corresponding species tree, which is useful for species classification based on orthology (Altenhoff & Dessimoz, 2009). In addition, inferring orthology using proteins sequences is more efficient due to the lack of confidence when aligning divergent DNA sequences (Aguileta *et al.*, 2008).

4.2.5 Ortholog inference software

A variety of software packages have been developed to infer orthology, all with some measure of phylogenetic correction or validation. Two main approaches followed by software are apparent, the first being inferring the pairwise relationships between genes in two species, and subsequently extending the orthology to multiple organisms by identifying genes spanning these orthogroup pairs. A problem with this approach is that gene duplications confound the orthology relationships which are then not transitive. These methods have high precision, but low rates of complete orthogroup detection. The second approach attempts the identification of complete orthogroups, thereby including paralogs (Emms & Kelly, 2015).

Only a subset of the available software packages will be discussed, while the COG approach was discussed in Section 3.2.7.5.

4.2.5.1 OrthoDB

The OrthoDB resource provides a catalog of orthologous protein-coding genes across vertebrates, arthropods, fungi, plants, and bacteria. For implementation, best reciprocal hits of genes between genomes are initially identified using MMseqs2 (Steinegger & Soding, 2017). Subsequently genes with higher matched than between genomes are identified as co-orthologous, and finally best reciprocal hits and in-paralogs are clustered into groups of orthologous genes. Only the longest isoform per gene was retained. In addition, the resource attempts to map functional categories and ascribe tentative functional annotations to the orthologous clusters (Kriventseva *et al.*, 2015 ; Kriventseva *et al.*, 2018). Currently the online resource contains a total of 37 million genes, from 1,271 eukaryotes and 6,013 prokaryotes (www.orthodb.org).

4.2.5.2 OrthoFinder

OrthoFinder was developed to address a previously undetected gene length bias in orthogroup inference. When using BLAST similarity analyses, longer sequences tend to have better bit scores and lower e-values than those of short best-hit sequences. This decreases the sensitivity of this approach and leads to a significant false negative rate. Following a correction/normalization for gene length in the pairwise BLAST analyses, the orthogroup graph was subjected to Markov clustering to yield the final orthogroup set. The algorithm yields orthogroups which are well balanced between precision and sensitivity, while it is also robust to missing data (Emms & Kelly, 2015).

4.2.5.3 OrthoMCL

OrthoMCL has been the most widely used method for orthology inference (Emms & Kelly, 2015), and the algorithm also utilizes BLAST pairwise analyses for identifying similarity scores between multiple species, followed by Markov clustering of the orthologs and paralogs (Li *et al.*, 2003). The OrthoMCL algorithm does appear to result in significant false-positive orthogroups (Altenhoff & Dessimoz, 2009).

4.2.5.4 PANTHER

The basis of PANTHER (Protein ANalysis THrough Evolutionary Relationships) is a library of phylogenetic trees for protein families which is then used for orthology inference. A HMM is created for each protein-family to identify new orthologs. In addition, functional inferences are also made for the protein families via manual curation (Mi *et al.*, 2013).

4.2.5.5 OMA

The OMA (Orthologous MAtrix) algorithm compares genes based on evolutionary distance with reciprocal Smith-Waterman alignments, and takes gene losses and uncertainty in distance inference into account. Splice variants have been considered carefully, and instead of using only the longest transcript for each gene, OMA selects the variant with the highest number of significant matches in all other genomes during the reciprocal similarity analyses (Dessimoz *et al.*, 2005 ; Altenhoff *et al.*, 2011). The OMA database (omabrowser.org) automatically identifies orthologs between publicly available complete genomes, and currently contains eleven million protein sequences from 1,617 bacteria, 141 archaea and 327 eukaryotes (Altenhoff *et al.*, 2018).

4.2.5.6 InParanoid

The InParanoid (In-paralog and ortholog identification) algorithm operates on genome pairs, where ortholog clusters are created with a reciprocal pairwise similarity match, after which high-confidence in-paralogs are added (Remm *et al.*, 2001). The InParanoid database (inparanoid.sbc.su.se) provides ortholog data for 246 eukaryotes, 20 bacteria and 7 archaea (Sonnhammer & Ostlund, 2015).

4.2.5.7 EggNOG

The EggNOG pipeline (eggnogdb.embl.de) utilizes the SIMAP (Similarity Matrix of Proteins) database, which identifies protein similarity and domains using the FASTA algorithm and HMMs (Rattei *et al.*, 2006), identifies reciprocal best matches and subsequently uses triangular linkage clustering to identify orthologs. The graph-based clustering algorithm is executed at different pre-defined taxonomic levels, both to cover evolutionary relevant groups and well-studied model organisms. The data is therefore presented in a hierarchical structure and is freely available (Huerta-Cepas *et al.*, 2016 ; Huerta-Cepas *et al.*, 2017).

In this study we analyzed the proteomes of 18 *Dothideomycete* and 7 outgroup fungal species for protein orthologs using the OrthoFinder software package. Both BLAST (Altschul *et al.*, 1990) and DIAMOND (Buchfink *et al.*, 2015) similarity algorithms were evaluated for specificity, and the analyses yielded 61 *Cercospora*-specific single-copy orthologs in *C. zeina*, *C. zeae-maydis*, *C. berteroae* and *C. beticola*. The orthologs were analyzed for phylogenetic information content, with 60 genes found to be suitable for phylogenetic analyses. Functional description criteria yielded eight genes suitable for primer design analysis. For four of the genes suitable regions for primer design could not be identified, while degenerate primers for the remaining four genes were designed. Additional *C. zeina*-specific primer pairs were designed up- and downstream from the degenerate primer region for each gene to serve as positive control for degenerate primer binding. The *C. zeina*-specific primers amplified the correct size products from *C. zeina* genomic DNA, while the degenerate primers of only two of the genes were suitable for the amplification of the respective gene regions from *C. zeina* and *C. zeae-maydis* genomic DNA. Sanger sequencing of PCR products confirmed the amplification of the correct gene regions for both the Glucoamylase 1 and Alkaline protease 1 degenerate primer pairs.

4.3 Materials and Methods

4.3.1 Proteome data

The proteome data for 25 fungal species were used during the study, and where required were downloaded from the relevant sequence databases (Table 4.1).

Table 4.1 Fungal species used in ortholog inference

Species	Strain	Sequence Database	Reference(s)
Aspergillus nidulans	FGSC A4	*Aspergillus* Genome Database	(Galagan *et al.*, 2005 ; Arnaud *et al.*, 2012)
Aspergillus niger	CBS 513.88	*Aspergillus* Genome Database	(Pel *et al.*, 2007 ; Arnaud *et al.*, 2012)
Baudoinia compniacensis	UAMH 10762	JGI Project 402532	(Ohm *et al.*, 2012)
Botrytis cinerea	B05.10	NCBI BioProject PRJNA15632	(Amselem *et al.*, 2011 ; Staats & van Kan, 2012)
Cercospora berteroae	CBS538.71	NCBI BioProject PRJNA270309	(de Jonge *et al.*, 2018)
Cercospora beticola	09-40	NCBI BioProject PRJNA270309	(de Jonge *et al.*, 2018)
Cercospora zeae-maydis	SCOH1-5	JGI Project 401984	-
Cercospora zeina	CBS142763	NCBI BioProject PRJNA355276	(Wingfield *et al.*, 2017)
Cladosporium fulvum	CBS131901	NCBI BioProject PRJNA86753	(de Wit *et al.*, 2012 ; Ohm *et al.*, 2012)
Cochliobolus heterostrophus	C5	JGI Project 52344	(Ohm *et al.*, 2012 ; Condon *et al.*, 2013)
Cochliobolus sativus	ND90Pr	JGI Project 401995	(Ohm *et al.*, 2012 ; Condon *et al.*, 2013)
Leptosphaeria maculans	v23.1.3	NCBI BioProject PRJEB24469	(Rouxel *et al.*, 2011)
Pseudocercospora (Mycosphaerella) fijiensis	CIRAD86	JGI Project 16189	(Arango Isaza *et al.*, 2016)
Mycosphaerella graminicola	CBS 115943	JGI Project 16205	(Goodwin *et al.*, 2011)
Neurospora crassa	FGSC 73	JGI Project 1019194	(Baker *et al.*, 2015)
Pyrenophora teres	NFNB isolate 0-1	NCBI BioProject PRJNA50389	(Ellwood *et al.*, 2010)
Pyrenophora tritici-repentis	Pt-1C BFP	NCBI BioProject PRJNA18815	(Manning *et al.*, 2013)
Rhytidhysteron rufulum	CBS 306.38	NCBI BioProject PRJNA81799	(Ohm *et al.*, 2012)
Saccharomyces cerevisiae	M3707	NCBI BioProject PRJNA18815	(Amselem *et al.*, 2011)
Sclerotinia sclerotiorum	1980	NCBI BioProject PRJNA18815	(Amselem *et al.*, 2011)
Septoria musiva	SO2202	JGI Project 401987	(Ohm *et al.*, 2012 ; Dhillon *et al.*, 2015)
Septoria populicola	P02.02b	NCBI BioProject PRJNA81737	(Ohm *et al.*, 2012 ; Dhillon *et al.*, 2015)
Setosphaeria turcica	Et28A	JGI Project 401988	(Ohm *et al.*, 2012 ; Condon *et al.*, 2013)
Stagonospora nodorum	SN15	NCBI BioProject PRJNA21049	(Hane *et al.*, 2007)
Verticillium dahlia	VdLs.17	NCBI BioProject PRJNA225532	(Klosterman *et al.*, 2011)

4.3.2 Additional *Cercospora* genome sequences

The genome assembly files for *Cercospora sojina* strain N1 (NCBI BioProject PRJNA183604), *Cercospora canescens* strain BHU (NCBI BioProject PRJNA371568) and *C. nicotianae* strain CBS 131.32 (NCBI BioProject PRJNA270309) were downloaded from Genbank.

4.3.3 Ortholog inference

Orthologous genes were inferred in the species (Section 4.3.1) using the OrthoFinder 2.2.6 package (Emms & Kelly, 2015). The BLAST (Altschul *et al.*, 1990) and DIAMOND (Buchfink *et al.*, 2015) similarity search algorithms were used separately, with e-value cut-offs set at 1x10^{-3}. Proteome-specific protein identifiers and sequences for each orthogroup were extracted from the `Orthogroups.txt` results file using command-line tools, and the number of genes in each orthogroup for the respective species were extracted from the `Orthogroups.GeneCount.csv` output file. Only orthogroups containing single copy orthologs specific for the four *Cercospora* species were further analyzed, while only genes common to both the BLAST and DIAMOND mediated analyses were retained.

4.3.4 Phylogenetic informativeness

The *Cercospora*-specific protein ortholog sequences (Section 4.3.3) were extracted from the relevant proteome sequence files (Section 4.3.1). Amino acid sequences for each orthogroup were grouped and converted to NEXUS file format (Maddison *et al.*, 1997). The NEXUS files for each orthogroup were imported in the PAUP* (Phylogenetic Analysis Using PAUP) 4.0a (build 163) software (Wilgenbusch & Swofford, 2003). A separate, exhaustive Maximum Likelihood search was performed for each orthogroup, and the value of the G1 statistic recorded.

4.3.5 Gene selection for primer design

The orthogroups with positive G1 statistics (Section 4.3.4) were removed from consideration. The remaining orthogroups were analyzed for putative functional annotations information using BLASTP on the Genbank nr database (BLAST). Genes with putative descriptions of 'Hypothetical' were removed from consideration. The remaining genes were selected for primer design.

4.3.6 Primer design

The amino acid sequence data for each orthogroup with a functional description (Section 4.3.5) were aligned using the CLUSTALW algorithm (Larkin *et al.*, 2007) in the BioEdit software v7.2.5 (BioEdit). The alignment was saved to `.aln` format and imported in CLC Main Workbench v8.0.1. (CLC). Regions in each protein alignment with 100% amino acid identity between the *C. zeina*, *C. zeae-maydis*, *C. berteroae* and *C. beticola* species of at least 15 amino acids in length and with separation of 130 – 200 amino acids were selected for primer design. The size selection ensured the amplification of fragments with suitable size for sequencing, while the primers also flanked intronic regions. Multiple sites conforming to these conditions were selected for each gene. The consensus amino acid sequence for each region was used to extract the nucleic acid sequence co-ordinates in the genome assembly files for *C. zeina*, *C. zeae-maydis*, *C. berteroae*, *C. beticola*, *C. sojina, C. canescens* and *C. nicotianae* using the TBLASTN similarity search algorithm. Output co-ordinate files were converted to `.bed` files and the nucleic acid sequences extracted using the *getfasta* command of the BEDtools software package (Quinlan, 2014).

The nucleic acid sequences for each region were aligned using CLUSTALW in BioEdit. Selected sequences for forward and reverse regions (in forward orientation) were concatenated with 100 No-call (N) bases separating the regions and imported in Primer Designer v4.20 (Sci-Ed Software). Primers were automatically designed with minimum product length of 100 bp, primer length of 18 bp and additional criteria described in Table 4.2.

Table 4.2 Primer design criteria for Primer Designer software

Criteria description	**Values**	
GC and Tm values		
GC% Range	Min 50%	Max 60%
Tm range	Min 55°C	Max 80°C
Match pairs	GC ±5%	Tm ±10°C
Stability values		
Annealing temperature	55	
Stability 5' vs 3'	1.2	
Dimers and runs		
Matches at 3'	<3	
Adjacent homology	<7	
Repeated base runs	<3	
Repeats dinucleotide pairs	<3	
Within 6 bp of 3'	<8	

The suggested primer pairs were evaluated for homodimer and heterodimer formation stability.

4.3.6.1 Degenerate primers

Suggested primer pairs from the Primer Designer v4.20 software (Sci-Ed Software) for degenerate regions were evaluated for the level of degeneracy. Where polymorphisms were indicated in the aligned primer region the degeneracy was recorded in the primer sequence, while regions where a polymorphism was only present in one species, the polymorphism was disregarded. The regions with the least degeneracies were selected for final primer selection.

4.3.6.2 *C. zeina*-specific primers

Regions upstream of forward degenerate primers and downstream of reverse primers were identified in the *C. zeina* genome sequence and primers designed using the Primer Designer v4.20 software (Software). The primers were selected to yield maximum product lengths of 700 bp, and products had to include the degenerate primer binding sites.

4.3.7 Genomic DNA isolation

C. zeina and genomic DNA was isolated as previously (Section 2.3.5). *C. zeae-maydis* genomic DNA was isolated using a similar protocol.

4.38 Polymerase Chain Reaction

The primers were used in polymerase chain reactions (PCR) to validate the primer binding and amplification efficiency in *C. zeina* and *C. zeae-maydis* genomes. Primers were manufactured by Inqaba Biotec (Pretoria, South Africa). For all PCRs a no-template control was included, containing dd-H_2O instead of template DNA.

4.3.8.1 *C. zeina*-specific primer PCRs

The Ampliqon Taq DNA Polymerase Master Mix Red (Odense C, Denmark) was used in all PCRs. The general PCR setup for the *C. zeina*-specific primer reactions using the *ceze-4*, *ceze-5*, *ceze-7* and *ceze-8* primer sets are provided in Table 4.3. A negative control reaction containing no template DNA was also included.

Table 4.3 Setup for *C. zeina*-specific PCRs

Description	Amounts
Genomic DNA (*C. zeina*)	3ng
Forward primer	5pmol
Reverse primer	5pmol
2X Taq mastermix	10µl
dd-H_2O	Volume up to total 20µl

PCRs were performed with the cycling conditions provided in Table 4.4.

Table 4.4 Cycling conditions for *C. zeina*-specific PCRs

Description	Temperature	Time	Cycles
Initial denaturing	94°C	7:00	1
Cycle denaturing	94°C	0:45	35
Annealing	60°C	0:45	
Extension	72°C	0:45	
Final extension	72°C	7:00	1

4.3.8.2 *Ceze-4* degenerate primer PCR

The optimized PCR setup for the *ceze-4* degenerate primer set reactions on the *C. zeina*-specific PCR products are provided in Table 4.5. The *C. zeina*-specific PCR products were neutralized of dNTPs and the unincorporated primers, diluted step-wise to 1/1,000, 1/5,000 and 1/10,000 in dd-H_2O, and used as input template for PCRs using the respective degenerate primers. A negative control reaction containing no template DNA was also included.

Table 4.5 Setup for *ceze-4* primer set PCRs on *C. zeina*-specific PCR products

Description	Amounts
C. zeina-specific PCR products (1/1000)	1µl
Forward primer	5pmol
Reverse primer	5pmol
2X Taq mastermix	10µl
dd-H_2O	Volume up to total 20µl

PCRs were performed with the cycling conditions provided in Table 4.6.

Table 4.6 Cycling conditions for *ceze-4* primer set PCRs on *C. zeina*-specific PCR products

Description	Temperature	Time	Cycles
Initial denaturing	94°C	7:00	1
Cycle denaturing	94°C	0:45	30
Annealing	60°C	0:10	
Extension	72°C	0:45	
Final extension	72°C	7:00	1

The optimized PCR setup for the *ceze-4* degenerate primer set reactions on *C. zeina* genomic DNA are provided in Table 4.7. A negative control reaction containing no template DNA was also included.

Table 4.7 Setup for *ceze-4* primer set PCRs on *C. zeina* DNA

Description	Amounts
Genomic DNA (*C. zeina*)	20ng
Forward primer	10pmol
Reverse primer	10pmol
2X Taq mastermix	10µl
dd-H_2O	Volume up to total 20µl

PCRs were performed with the cycling conditions provided in Table 4.8.

Table 4.8 Cycling conditions for *ceze-4* primer set PCRs on *C. zeina* DNA

Description	Temperature	Time	Cycles
Initial denaturing	94°C	7:00	1
Cycle denaturing	94°C	0:45	30
Annealing	60°C	0:45	
Extension	72°C	0:45	
Final extension	72°C	7:00	1

The PCR setup for the *ceze-4* degenerate primer set reactions on *C. zeae-maydis* genomic DNA are provided in Table 4.9. A negative control reaction containing no template DNA was also included.

Table 4.9 Setup for *ceze-4* primer set PCRs on *C. zeae-maydis* DNA

Description	Amounts
Genomic DNA (*C. zeae-maydis*)	50ng
Forward primer	40pmol
Reverse primer	40pmol
2X Taq mastermix	10µl
dd-H_2O	Volume up to total 20µl

PCRs were performed with the cycling conditions provided in Table 4.10.

Table 4.10 Cycling conditions for *C. zeina*-specific PCRs

Description	Temperature	Time	Cycles
Initial denaturing	94°C	7:00	1
Cycle denaturing	94°C	0:45	30
Annealing	60°C	0:45	
Extension	72°C	0:45	
Final extension	72°C	7:00	1

4.3.8.3 *Ceze-5* degenerate primer PCR

The optimized PCR setup for the *ceze-5* degenerate primer set reactions on the *C. zeina*-specific PCR products are provided in Table 4.11. The *C. zeina*-specific PCR products were neutralized of dNTPs and the unincorporated primers, diluted step-wise to 1/100, 1/500 and 1/1000 in dd-H_2O, and used as input template for PCRs using the respective degenerate primers. A negative control reaction containing no template DNA was also included.

Table 4.11 Setup for *ceze-5* primer set PCRs on *C. zeina*-specific PCR products

Description	Amounts
C. zeina-specific PCR products (1/100)	1µl
Forward primer	5pmol
Reverse primer	5pmol
2X Taq mastermix	10µl
dd-H_2O	Volume up to total 20µl

PCRs were performed with the cycling conditions provided in Table 4.12.

Table 4.12 Cycling conditions for *ceze-5* primer set PCRs on *C. zeina*-specific PCR products

Description	Temperature	Time	Cycles
Initial denaturing	94°C	7:00	1
Cycle denaturing	94°C	0:45	35
Annealing	58°C	0:10	
Extension	72°C	0:45	
Final extension	72°C	7:00	1

The PCR setup for the *ceze-5* degenerate primer set reactions on *C. zeina* genomic DNA are provided in Table 4.13. A negative control reaction containing no template DNA was also included.

Table 4.13 Setup for *ceze-5* primer set PCRs on *C. zeina* DNA

Description	Amounts
Genomic DNA (*C. zeina*)	40ng
Forward primer	100pmol
Reverse primer	100pmol
2X Taq mastermix	10μl
dd-H_2O	Volume up to total 20μl

PCRs were performed with the cycling conditions provided in Table 4.14.

Table 4.14 Cycling conditions for *ceze-5* primer set PCRs on *C. zeina* DNA

Description	Temperature	Time	Cycles
Initial denaturing	94°C	7:00	1
Cycle denaturing	94°C	0:45	35
Annealing	58°C	0:10	
Extension	72°C	0:45	
Final extension	72°C	7:00	1

Subsequent PCR conditions were adapted with cycling annealing temperatures varying from 54°C-60°C on a S1000 gradient PCR machine (Bio-Rad, Hercules, California, USA), step-up PCR cycling with increasing annealing temperatures, as well as the setup of a Taguchi PCR optimization with varying concentrations of *C. zeina* genomic DNA and primers.

4.3.8.4 *Ceze-7* degenerate primer PCR

The optimized PCR setup for the *ceze-7* degenerate primer set reactions on the *C. zeina*-specific PCR products are provided in Table 4.15. The *C. zeina*-specific PCR products were neutralized of dNTPs and the unincorporated primers, diluted step-wise to 1/1,000, 1/5,000 and 1/10,000 in sterile dd-H_2O, and used as input template for PCRs using the respective degenerate primers. A negative control reaction containing no template DNA was also included.

Table 4.15 Setup for *ceze-7* primer set PCRs on *C. zeina*-specific PCR products

Description	Amounts
C. zeina-specific PCR products (1/10,000)	1μl
Forward primer	5pmol
Reverse primer	5pmol
2X Taq mastermix	10μl
dd-H_2O	Volume up to total 20μl

PCRs were performed with the cycling conditions provided in Table 4.16.

Table 4.16 Cycling conditions for *ceze-7* primer set PCRs on *C. zeina*-specific PCR products

Description	Temperature	Time	Cycles
Initial denaturing	94°C	7:00	1
Cycle denaturing	94°C	0:45	30
Annealing	60°C	0:10	
Extension	72°C	0:45	
Final extension	72°C	7:00	1

The optimized PCR setup for the *ceze-7* degenerate primer set reactions on *C. zeina* genomic DNA are provided in Table 4.17. A negative control reaction containing no template DNA was also included.

Table 4.17 Setup for *ceze-7* primer set PCRs on *C. zeina* DNA

Description	Amounts
Genomic DNA (*C. zeina*)	3ng
PCR products (positive control reactions)	1μl
Forward primer	10pmol
Reverse primer	10pmol
2X Taq mastermix	10μl
dd-H_2O	Volume up to total 20μl

PCRs were performed with the cycling conditions provided in Table 4.18.

Table 4.18 Cycling conditions for *ceze-7* primer set PCRs on *C. zeina* DNA

Description	Temperature	Time	Cycles
Initial denaturing	94°C	7:00	1
Cycle denaturing	94°C	0:45	35
Annealing	58°C	0:45	
Extension	72°C	0:45	
Final extension	72°C	7:00	1

The PCR setup for the *ceze-7* degenerate primer set reactions on *C. zeae-maydis* genomic DNA are provided in Table 4.19. A negative control reaction containing no template DNA was also included.

Table 4.19 Setup for *ceze-7* primer set PCRs on *C. zeae-maydis* DNA

Description	Amounts
Genomic DNA (*C. zeae-maydis*)	1ng
Forward primer	10pmol
Reverse primer	10pmol
2X Taq mastermix	10μl
dd-H_2O	Volume up to total 20μl

PCRs were performed with the cycling conditions provided in Table 4.20.

Table 4.20 Cycling conditions for *ceze-7* primer set PCRs on *C. zeae-maydis* DNA

Description	Temperature	Time	Cycles
Initial denaturing	94°C	7:00	1
Cycle denaturing	94°C	0:45	35
Annealing	58°C	0:45	
Extension	72°C	0:45	
Final extension	72°C	7:00	1

4.3.8.5 *Ceze-8* degenerate primer PCR

The optimized PCR setup for the *ceze-8* degenerate primer set reactions on the *C. zeina*-specific PCR products are provided in Table 4.21. The *C. zeina*-specific PCR products were neutralized of dNTPs and the unincorporated primers, diluted step-wise to 1/100, 1/500 and 1/1000 in dd-H_2O, and used as input template for PCRs using the respective degenerate primers. A negative control reaction containing no template DNA was also included.

Table 4.21 Setup for *ceze-8* primer set PCRs on *C. zeina*-specific PCR products

Description	Amounts
C. zeina-specific PCR products (1/100)	1µl
Forward primer	5pmol
Reverse primer	5pmol
2X Taq mastermix	10µl
dd-H_2O	Volume up to total 20µl

PCRs were performed with the cycling conditions provided in Table 4.22.

Table 4.22 Cycling conditions for *ceze-8* primer set PCRs on *C. zeina*-specific PCR products

Description	Temperature	Time	Cycles
Initial denaturing	94°C	7:00	1
Cycle denaturing	94°C	0:45	35
Annealing	58°C	0:10	
Extension	72°C	0:45	
Final extension	72°C	7:00	1

The PCR setup for the *ceze-8* degenerate primer set reactions on *C. zeina* genomic DNA are provided in Table 4.23. A negative control reaction containing no template DNA was also included.

Table 4.23 Setup for *ceze-8* primer set PCRs on *C. zeina* DNA

Description	Amounts
Genomic DNA (*C. zeina*)	40 ng
Forward primer	100 pmol
Reverse primer	100 pmol
2X Taq mastermix	10 µl
dd-H_2O	Volume up to total 20 µl

PCRs were performed with the cycling conditions provided in Table 4.24.

Table 4.24 Cycling conditions for *ceze-8* primer set PCRs on *C. zeina* DNA

Description	Temperature	Time	Cycles
Initial denaturing	94°C	7:00	1
Cycle denaturing	94°C	0:45	35
Annealing	58°C	0:45	
Extension	72°C	0:45	
Final extension	72°C	7:00	1

Subsequent PCR conditions were adapted with cycling annealing temperatures varying from 54°C-60°C on a S1000 gradient PCR machine (Bio-Rad, Hercules, California, USA), step-up PCR cycling with increasing annealing temperatures, as well as the setup of a Taguchi PCR optimization with varying concentrations of *C. zeina* genomic DNA and primers.

4.3.8.6 Primer and dNTP neutralization

Following thermal cycling the remaining primers and dNTPs in the PCR mix were neutralized using the ExoSAP-IT™ PCR Product Cleanup Reagent (ThermoFisher Scientific, Waltham, Massachusetts, USA) according to manufacturer's instructions. Briefly, 10µl PCR products were incubated with 4µl ExoSAP-IT™ Reagent at 37°C for 15 minutes, followed by incubation at 80°C for 15 minutes to deactivate the reagents. Neutralized PCR products were used for sequencing.

4.3.9 Gel electrophoresis

PCR products were visualized using agarose gel electrophoresis. Agarose (2% (w/v), SeaKem® LE Agarose, Lonza, Switzerland) was melted in Borax (5mM, Dischem Pharmacists Choice, South Africa) with Ethidium Bromide (2µl of a 10 mg/ml (w/v) solution, Sigma-Aldrich, St. Louis, USA) added to the final gel solution. Electrophoresis was conducted in Borax (5mM) at 180V, and a 100 bp DNA ladder (New England Biolabs, Ipswich, Massachusetts, USA) was used on each gel to qualify PCR product sizes (Appendix Figure A9). Gels were visualized using a Gel Doc™ XR+ Gel Documentation System (Bio-Rad, Hercules, California, USA).

4.3.10 Sanger sequencing

PCR product sequences were determined using Sanger sequencing using the BigDye™ Terminator v3.1 Cycle Sequencing Kit (ThermoFisher, Waltham, Massachusetts, United States) according to the manufacturers specifications. The sequencing mix conditions are provided in Table 4.25.

Table 4.25 Sequencing setup for PCR products

Description	Amounts
Template / PCR products	100ng
5x Sequencing buffer	1µl
Primer	5pmol
BigDye mix	2µl
dd-H_2O	Volume up to total 10µl

The cycling conditions for the sequencing PCR is provided in Table 4.26.

Table 4.26 Cycling conditions for sequencing PCR

Description	Temperature	Time	Cycles
Initial denaturing	96°C	1:00	1
Cycle denaturing	96°C	0:10	25
Annealing	56°C	0:05	
Extension	60°C	4:00	
Final hold	4°C	∞	-

Sequencing reaction products were precipitated with 2 µl 3M sodium acetate (pH 5.2 ; Merck-Millipore, Burlington, Massachusetts, United States), 10µl sterile dd-H_2O and 50µl absolute ethanol (Merck-Millipore, Burlington, Massachusetts, United States). Following 10 min incubation on ice, the samples were centrifuged for 30 min at RCF 10,000 at room temperature. The precipitated samples were washed with 70% ethanol/dd-H_2O and centrifuged at RCF 10,000 for 10 min at room temperature. After three washes the samples were dried at room temperature to remove residual ethanol.

Sequencing trace files were analyzed using 4Peaks on Mac OS X and base-calling for ambiguous reads corrected. Sequences were extracted in FASTA-format and consensus sequences created using BioEdit. Sequences were compared with the expected sequences to verify primer target specificity. ClustalW multiple sequence alignments were performed against the respective *C. zeina* gene-specific genomic sequences in BioEdit to verify sequence identities.

4.4 Results

4.4.1 Ortholog analysis

The OrthoFinder analysis using the DIAMOND similarity search algorithm yielded a total of 120,590 orthogroups. There were no orthogroups containing single copies of each species. A total of 85 orthogroups contained single copy proteins specific for the four *Cercospora* species. The protein IDs were extracted from the `Orthogroups.txt` file, the `SequenceIDs.txt`, the `SpeciesIDs.txt` with reference to the original proteome FASTA file for each species.

The OrthoFinder analysis using the BLAST similarity search algorithm yielded a total of 89,195 orthogroups. There were no orthogroups containing single copies of each species. A total of 85 orthogroups contained single copy proteins specific for the four *Cercospora* species. The protein IDs were extracted from the `Orthogroups.txt` file, the `SequenceIDs.txt`, the `SpeciesIDs.txt` with reference to the original proteome FASTA file for each species.

The group of proteins present in both the DIAMOND and BLAST orthogroup sets were extracted, and it was confirmed that the same proteins were present in each orthogroup in comparison between the two datasets. A total of 61 orthogroups were present in both the DIAMOND and BLAST similarity search datasets.

4.4.2 Phylogenetic informativeness

The sequences for the proteins in each orthogroup were converted to NEXUS format, and each group separately uploaded to PAUP*. The exhaustive maximum likelihood analysis was completed for each orthogroup, and the G1 score values recorded. Of the 61 orthogroups, 60 orthogroups all showed G1 score values ranging from -0.707107 to -0.605855, and were all considered phylogenetically informative, while 1 group was found to have a G1 score of 0.539874, and was thus not phylogenetically informative and removed from consideration.

4.4.3 Genes for phylogenetics

To exlude proteins with Genbank functional descriptions listed as "Hypothetical", BLASTP similarity analysis were performed on the 60 orthogroups. The analysis yielded 8 proteins which show similarity with fungal organisms which all list similar functional descriptions (Table 4.27).

Table 4.27 Details of proteins considered for primer design

Name	**Genbank/JGI accession**				**Genbank functional description**
	C. zeina	*C. zeae-maydis*	*C. beticola*	*C. berteroae*	
ceze-1	PKR99811.1	jgi|Cerzm1|81198	PPJ56580.1	PIA88808.1	Heterokaryon incompatibility protein 6
ceze-2	PKR94824.1	jgi|Cerzm1|45618	PPJ51197.1	PIA97815.1	Carboxylesterase
ceze-3	PKR98996.1	jgi|Cerzm1|28847	PPJ59958.1	PIA94520.1	Acetylxylan esterase
ceze-4	PKS02172.1	jgi|Cerzm1|110710	PPJ57837.1	PIA97577.1	Glucoamylase 1
ceze-5	PKS00848.1	jgi|Cerzm1|97921	PPJ53786.1	PIA91697.1	Putative agmatine deiminase
ceze-6	PKS01758.1	jgi|Cerzm1|41101	PPJ51566.1	PIA90479.1	Choline transport protein
ceze-7	PKR99720.1	jgi|Cerzm1|108840	PPJ56611.1	PIA89053.1	Alkaline protease 1
ceze-8	PKS04314.1	jgi|Cerzm1|33152	PPJ58717.1	PIA99785.1	Putative steroid dehydrogenase 4

4.4.4 Primer design

Regions for possible primer design were identified in protein multiple sequence alignments, and subsequently analyzed for possible primer site selection.

4.4.4.1 *Ceze-1* (Heterokaryon incompatibility protein 6)

There were no regions in the multiple alignment of the *ceze-1* heterokaryon incompatibility protein 6 amino acid sequences of the *Cercospora* species which were suitable for primer design (Figure 4.1).

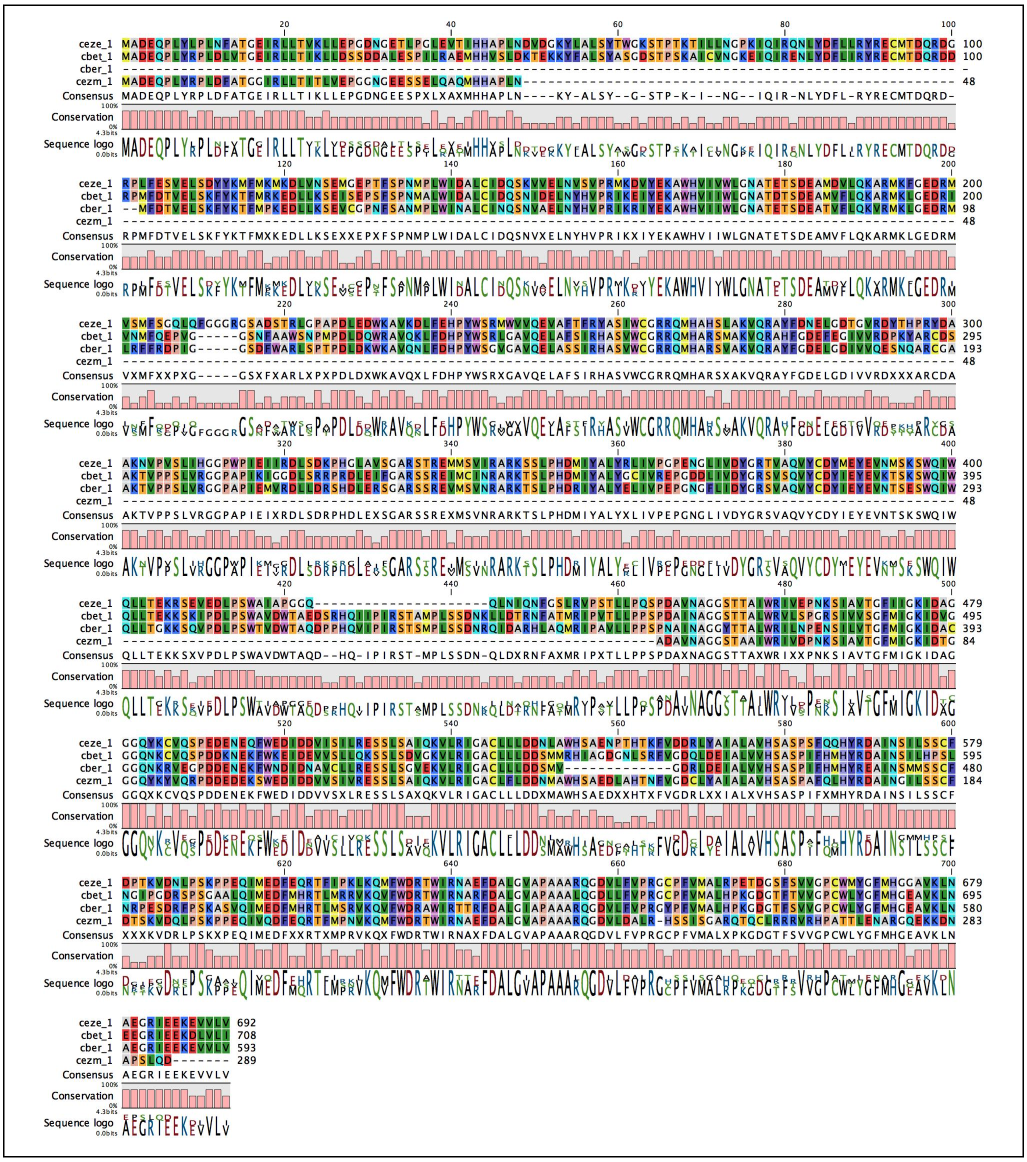


Figure 4.1 Multiple amino acid sequence alignment for *ceze-1* (Heterokaryon incompatibility protein 6) proteins. *Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**) and* C. zeae-maydis *(**cezm**). The Conservation bars are a visual representation of the number of aligned amino acids similar to the consensus amino acid for each position.*

4.4.4.2 *Ceze-2* (Carboxylesterase)

Conserved regions for primer design were identified in the multiple sequence alignment of the *ceze-2* carboxylesterase amino acid sequences of the *Cercospora* species (Figure 4.2).

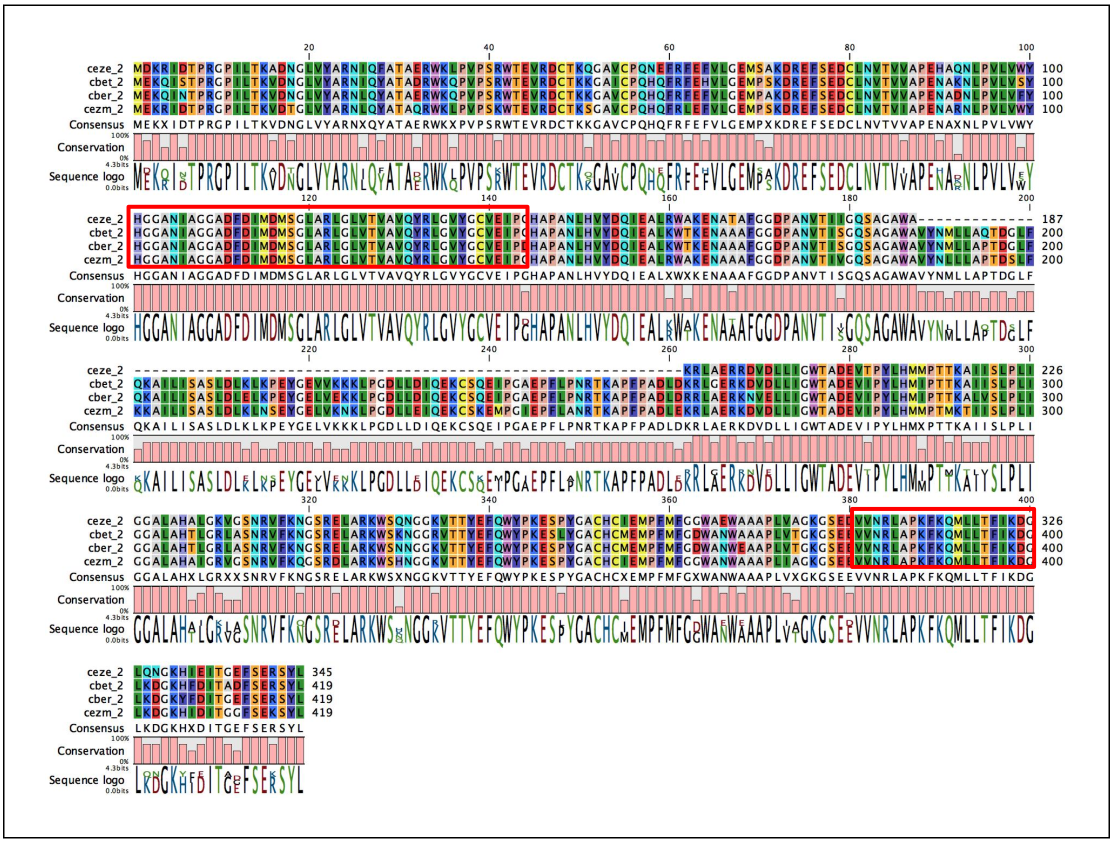


Figure 4.2 Multiple amino acid sequence alignment for *ceze-2* (Carboxylesterase) proteins. *The red blocks indicate the regions selected for forward and reverse primer design. Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**) and* C. zeae-maydis *(**cezm**).*

The region was deemed to not be suitable for primer design due to No-call (N) bases present in the forward primer design region of the *C. zeina* genome assembly sequence. These bases were introduced during the genome assembly scaffolding process.

4.4.4.3 *Ceze-3* (Acetylxylan esterase)

Conserved regions for primer design were identified in the multiple sequence alignment of the *ceze-3* acetylxylan esterase amino acid sequences of the *Cercospora* species (Figure 4.3).

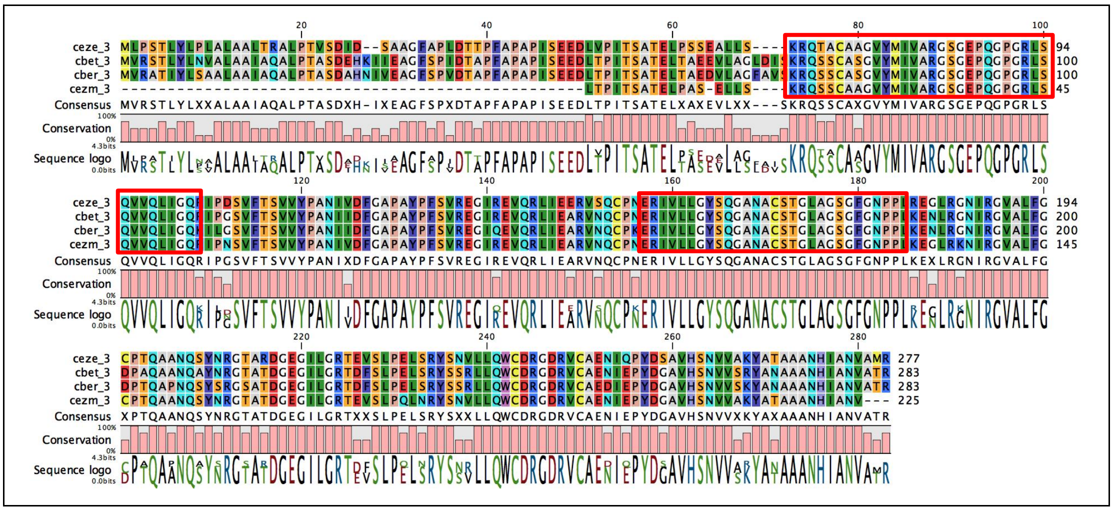

20
40
60
80
100
ceze_3 MLPSTLYLPLALAALTRALPTVSDID--SAAGFAPLDTTPFAPAPISEEDLVPITSATELPSSEALLS----KRQTACAAGVYMIVARGSGEPQGPGRLS 94
cbet_3 MVRSTLYLNVALAAIAQALPTASDEHKIIEAGFSPIDTAPFAPAPISEEDLTPITSATELTAEEVLAGLDISKRQSSCASGVYMIVARGSGEPQGPGRLS 100
cber_3 MVRATIYLSAALAAIAQALPTASDAHNIVEAGFSPVDTAPFAPAPISEEDLTPITSATELTAEDVLAGFAVSKRQSSCASGVYMIVARGSGEPQGPGRLS 100
cezm_3 --LTPITSATELPAS-ELLS----KRQSSCAAGVYMIVARGSGEPQGPGRLS 45
Consensus MVRSTLYLXXALAAIAQALPTASDXH-IXEAGFSPXDTAPFAPAPISEEDLTPITSATELXAXEVLXX---SKRQSSCAXGVYMIVARGSGEPQGPGRLS
100%
Conservation
0%
4.3bits
Sequence logo
0.0bits
120
140
160
180
200
ceze_3 QVVQLIGQRIPDSVFTSVVYPANIVDFGAPAYPFSVREGIREVQRLIEERVSQCPNERIVLLGYSQGANACSTGLAGSGFGNPPLREGLRGNIRGVALFG 194
cbet_3 QVVQLIGQRIPGSVFTSVVYPANIIDFGAPAYPFSVREGIREVQRLIEARVNQCPNERIVLLGYSQGANACSTGLAGSGFGNPPLKENLRGNIRGVALFG 200
cber_3 QVVQLIGQKILGSVFTSVVYPANIIDFGAPAYPFSVREGIQEVQRLIEARVNQCPKERIVLLGYSQGANACSTGLAGSGFGNPPLKENLRGNIRGVALFG 200
cezm_3 QVVQLIGQRIPNSVFTSVVYPANIVDFGAPAYPFSVREGIREVQRLIEARVNQCPNERIVLLGYSQGANACSTGLAGSGFGNPPLKEGLRKNIRGVALFG 145
Consensus QVVQLIGQRIPGSVFTSVVYPANIXDFGAPAYPFSVREGIREVQRLIEARVNQCPNERIVLLGYSQGANACSTGLAGSGFGNPPLKEXLRGNIRGVALFG
100%
Conservation
0%
4.3bits
Sequence logo
0.0bits
220
240
260
280
ceze_3 CPTQAANQSYNRGTARDGEGILGRTEVSLPELSRYSNVLLQWCDRGDRVCAENIQPYDSAVHSNVVAKYATAAANHIANVAMR 277
cbet_3 DPAQAANQAYNRGTATDGEGILGRTDFSLPELSRYSSRLLQWCDRGDRVCAENIEPYDGAVHSNVVSRYANAAANHIANVATR 283
cber_3 DPTQAPNQSYSRGSATDGEGILGRTDFSLPELSRYSSRLLQWCDRGDRVCAEDIEPYDGAVHSNVVSKYANAAANHIANVATR 283
cezm_3 CPTQAANQSYNRGTATDGEGILGRTEVSLPQLNRYSNVLLQWCDRGDRVCAENIEPYDGAVHSNVVAKYATAAANHIANV--- 225
Consensus XPTQAANQSYNRGTATDGEGILGRTXXSLPELSRYSXXLLQWCDRGDRVCAENIEPYDGAVHSNVVXKYAXAAANHIANVATR
100%
Conservation
0%
4.3bits
Sequence logo
0.0bits

Figure 4.3 Multiple amino acid sequence alignment for *ceze-3* (Acetylxylan esterase) proteins. *The red blocks indicate the regions selected for forward and reverse primer design.. Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**) and* C. zeae-maydis *(**cezm**).*

No primer pairs conforming to the primer design parameters could be designed in the indicated regions.

4.4.4.4 *Ceze-4* (Glucoamylase 1)

Conserved regions for primer design were identified in the multiple sequence alignment of the *ceze-4* glucoamylase 1 amino acid sequences of the *Cercospora* species (Figure 4.4).

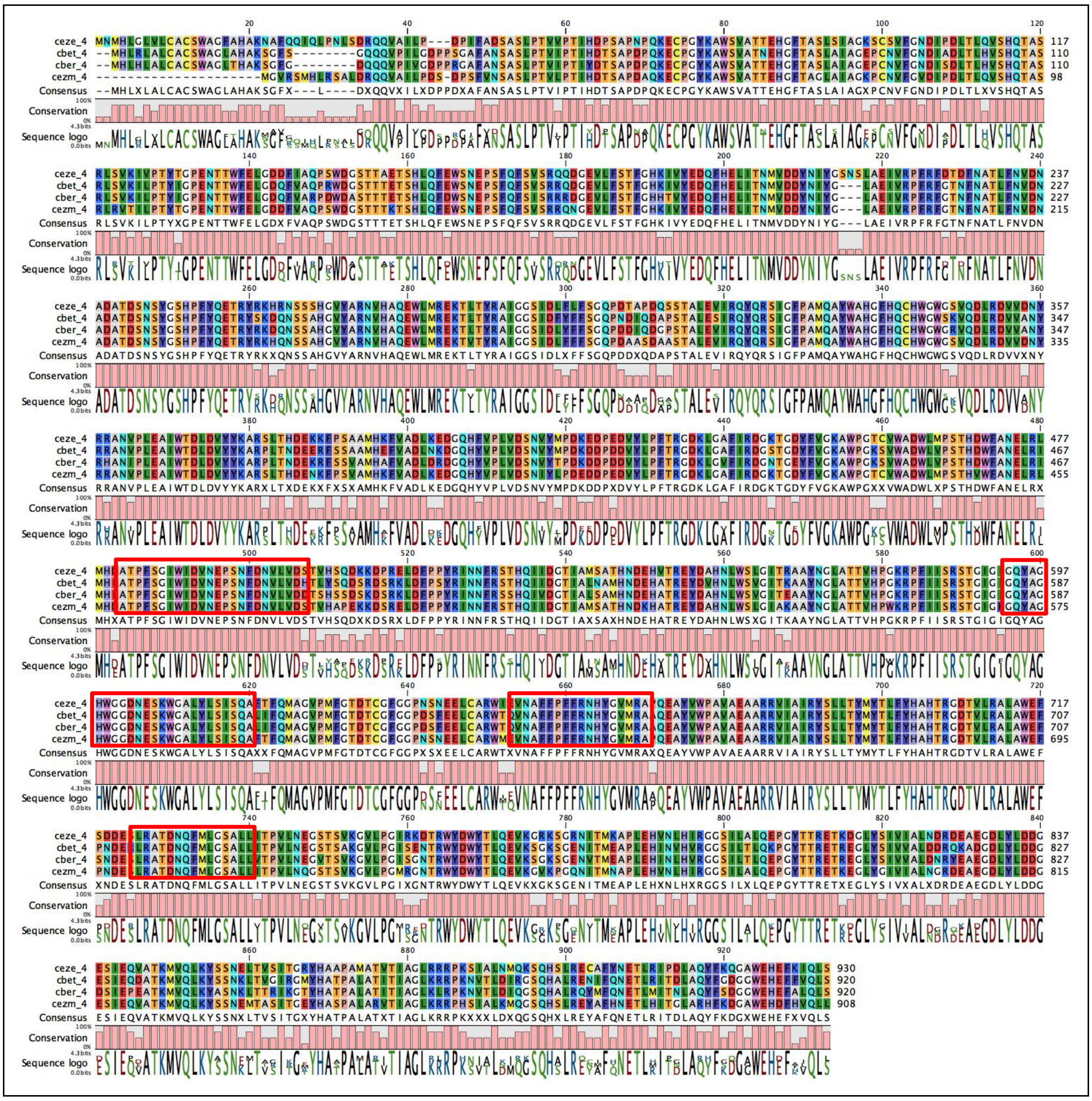


Figure 4.4 Multiple amino acid sequence alignment for *ceze-4* (Glucoamylase 1) proteins. *The red blocks indicate the regions selected for forward and reverse primer design. Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**) and* C. zeae-maydis *(**cezm**).*

The conserved regions were analyzed for possible primer sites, and only one suitable primer pair was possible in the regions indicated in Figure 4.4, with the primer details provided in Table 4.28. Sequence alignments of the forward and reverse sequence regions for the relevant *Cercospora* species are shown in Figure A1 and Figure A2 of the Appendix.

Table 4.28 *Ceze-4* degenerate primer details

Name	Sequence	Length	Tm	Degeneracy	Product size
CEZE-4-DF	GGCCAATATGCWGGACAY	18	54°C	4	466 bp
CEZE-4-DR	GATTRTCVGTAGCTCGYA	18	55°C	12	

Regions upstream of the *CEZE-4-DF* and downstream of the *CEZE-4-DR* primer sites were analyzed for *C. zeina* specific primer design. Primers were designed with details provided in Table 4.29.

Table 4.29 *Ceze-4 C. zeina*-specific primer details

Name	Sequence	Length	Tm	Product size
CEZE-4-PF	GCGATGTGGTCGATAACT	18	58°C	602 bp
CEZE-4-PR	AGTGCTATCCACGAGAAC	18	58°C	

4.4.4.5 *Ceze-5* (Putative agmatine deiminase)

Conserved regions for primer design were identified in the multiple sequence alignment of the *ceze-5* putative agmatine deiminase amino acid sequences of the *Cercospora* species (Figure 4.5).

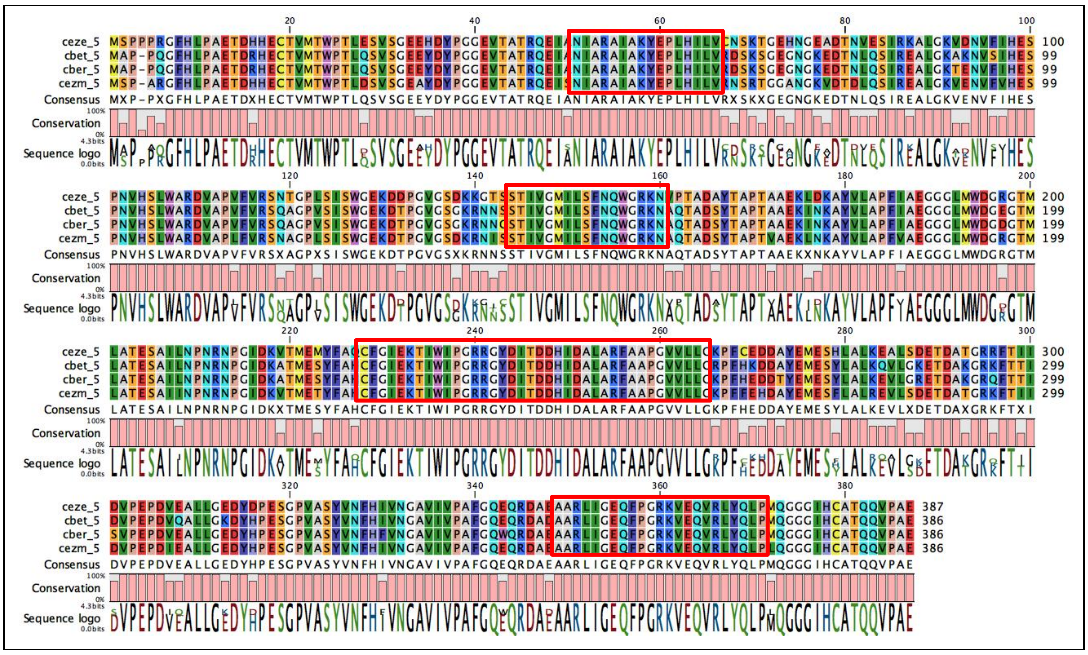

20
40
60
80
100
ceze_5 MSPPPRGFHLPAETDHHECTVMTWPTLESVSGEEHDYPGGEVTATRQEIANIARAIAKYEPLHILVCNSKTGEHNGEADTNVESIRKALGKVDNVFIHES 100
cbet_5 MAP-PQGFHLPAETDRHECTVMTWPTLQSVSGEEYDYPGGEVTATRQEIANIARAIAKYEPLHILVRDSKSGEGNGKEDTNLQSIREALGKAKNVSIHES 99
cber_5 MAP-PQGFHLPAETDRHECTVMTWPTLQSVSGEEYDYPGGEVTATRQEIANIARAIAKYEPLHILVRDSKSGEGNGKEDTNLQSIREALGKTENVFIHES 99
cezm_5 MSP-ARGFHLPAETDHHECTVMTWPTLDSVSGEAYDYPGGEVTATRQEISNIARAIAKYEPLHILVRNSRTGGANGKVDTDLQSIREALGKVENVFVHES 99
Consensus MXP-PXGFHLPAETDXHECTVMTWPTLQSVSGEEYDYPGGEVTATRQEIANIARAIAKYEPLHILVRXSKXGEGNGKEDTNLQSIREALGKVENVFIHES
100%
Conservation
0%
4.3bits
Sequence logo
0.0bits
120
140
160
180
200
ceze_5 PNVHSLWARDVAPVFVRSNTGPLSISWGEKDDPGVGSDKKGTSSTIVGMILSFNQWGRKNVPTADAYTAPTAAEKLDKAYVLAPFIAEGGGLMWDGRGTM 200
cbet_5 PNVHSLWARDVAPVFVRSQAGPVSISWGEKDTPGVGSGKRNNSSTIVGMILSFNQWGRKNAQTADSYTAPTAAEKINKAYVLAPFIAEGGGLMWDGEGTM 199
cber_5 PNVHSLWARDVAPVFVRSQAGPVSISWGEKDTPGVGSGKRNNCSTIVGMILSFNQWGRKNAQTADSYTAPTAAEKINKAYVLAPFIAEGGGLMWDGDGTM 199
cezm_5 PNVHSLWARDVAPLFVRSNAGPLSISWGEKDTPGVGSDKRNISSTIVGMILSFNQWGRKNAQTADSYTAPTVAEKLNKAYVLAPFVAEGGGLMWDGRGTM 199
Consensus PNVHSLWARDVAPVFVRSXAGPXSISWGEKDTPGVGSXKRNNSSTIVGMILSFNQWGRKNAQTADSYTAPTAAEKXNKAYVLAPFIAEGGGLMWDGRGTM
100%
Conservation
0%
4.3bits
Sequence logo
0.0bits
220
240
260
280
300
ceze_5 LATESAILNPNRNPGIDKVTMEMYFAQCFGIEKTIWIPGRRGYDITDDHIDALARFAAPGVVLLCKPFCEDDAYEMESHLALKEALSDETDATGRRFTII 300
cbet_5 LATESAILNPNRNPGIDKATMESYFAHCFGIEKTIWIPGRRGYDITDDHIDALARFAAPGVVLLCRPFHKDDAYEMESYLALKQVLGKETDAKGRKFTTI 299
cber_5 LATESAILNPNRNPGIDKATMESYFAHCFGIEKTIWIPGRRGYDITDDHIDALARFAAPGVVLLCKPFHEDDTYEMESYLALKEVLGRETDAKGRQFTTI 299
cezm_5 LATESAIINPNRNPGIDKVTMETYFAHCFGIEKTIWIPGRRGYDITDDHIDALARFAAPGVVLLCKPFFEHDAYEMESFLALREVLSDETDATGRKFTII 299
Consensus LATESAILNPNRNPGIDKXTMESYFAHCFGIEKTIWIPGRRGYDITDDHIDALARFAAPGVVLLGKPFHEDDAYEMESYLALKEVLXDETDAXGRKFTXI
100%
Conservation
0%
4.3bits
Sequence logo
0.0bits
320
340
360
380
ceze_5 DVPEPDVEALLGEDYDPESGPVASYVNFHIVNGAVIVPAFGQEQRDAEAARLIGEQFPGRKVEQVRLYQLPMQGGGIHCATQQVPAE 387
cbet_5 DVPEPDVQALLGKDYHPESGPVASYVNFHIVNGAVIVPAFGQEQRDADAARLIGEQFPGRKVEQVRLYQLPMQGGGIHCATQQVPAE 386
cber_5 SVPEPDVEALLGEDYHPESGPVASYVNFHFVNGAVIVPAFGQWQRDAEAARLIGEQFPGRKVEQVRLYQLPMQGGGIHCATQQVPAE 386
cezm_5 DVPEPDIEALLGEDYHPESGPVASYVNFHIVNGAVIVPAFGQEQRDAEAARLIGEQFPGRKVEQVRLYQLPLQGGGIHCATQQVPAE 386
Consensus DVPEPDVEALLGEDYHPESGPVASYVNFHIVNGAVIVPAFGQEQRDAEAARLIGEQFPGRKVEQVRLYQLPMQGGGIHCATQQVPAE
100%
Conservation
0%
4.3bits
Sequence logo
0.0bits

Figure 4.5 Multiple amino acid sequence alignment for *ceze-5* (Putative agmatine deiminase) proteins. *The red blocks indicate the regions selected for forward and reverse primer design. Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**) and* C. zeae-maydis *(**cezm**).*

The conserved regions were analyzed for possible primer sites, and only one suitable primer pair was possible in the regions indicated in Figure 4.5, with the primer details provided in Table 4.30. Sequence alignments of the forward and reverse sequence regions for the relevant *Cercospora* species are shown in Figure A3 and Figure A3 of the Appendix.

Table 4.30 *Ceze-5* **degenerate primer details**

Name	Sequence	Length	Tm	Degeneracy	Product size
CEZE-5-DF	ATYGTCGGCATGATCTTG	18	49°C	2	664 bp
CEZE-5-DR	AYCTCACYTGCTCCACYT	18	53°C	8	

Regions upstream of the *CEZE-5-DF* and downstream of the *CEZE-5-DR* primer sites were analyzed for *C. zeina* specific primer design. Primers were designed with details provided in Table 4.31.

Table 4.31 *Ceze-5 C. zeina***-specific primer details**

Name	Sequence	Length	Tm	Product size
CEZE-5-PF	AGGAGTCGGCTCGGATA	18	55°C	747 bp
CEZE-5-PR	TGCGTCGCGCAATGGAT	18	58°C	

4.4.4.6 *Ceze-6* (Choline transport protein)

There were no regions in the multiple alignment of the *ceze-6* choline transport protein amino acid sequences of the *Cercospora* species which were suitable for primer design (Figure 4.6).

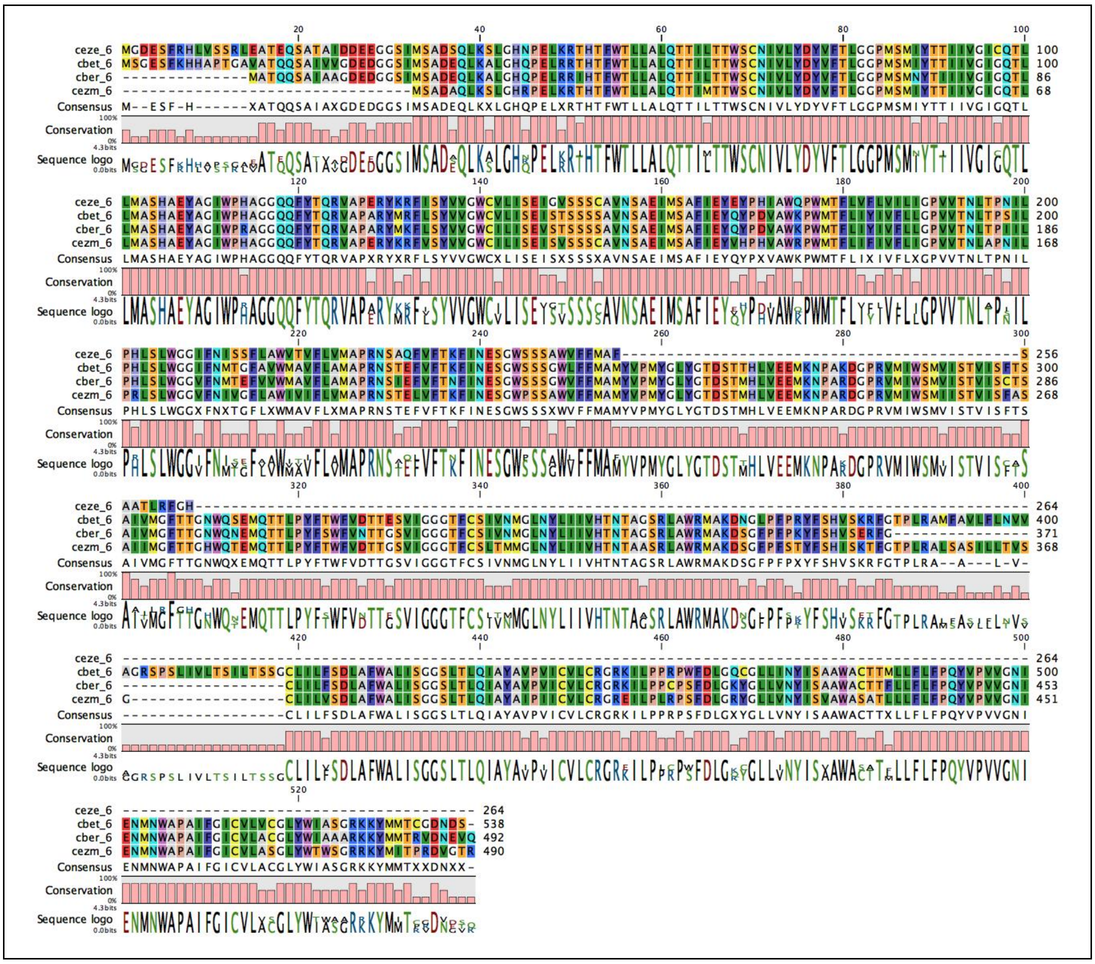


Figure 4.6 Multiple amino acid sequence alignment for *ceze-6* (Choline transport protein) proteins. *Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**) and* C. zeae-maydis *(**cezm**).*

4.4.4.7 *Ceze-7* (Alkaline protease 1)

Conserved regions for primer design were identified in the multiple sequence alignment of the *ceze-7* alkaline protease 1 amino acid sequences of the *Cercospora* species (Figure 4.7).

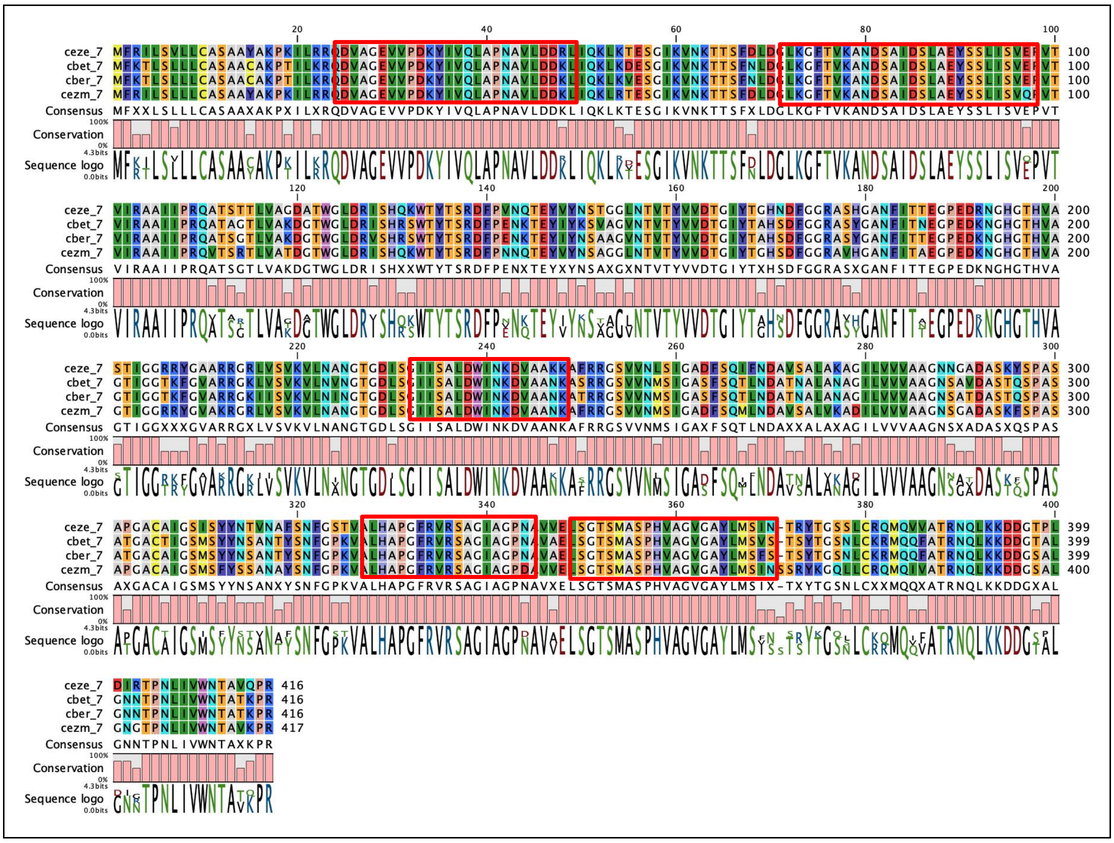


Figure 4.7 Multiple amino acid sequence alignment for *ceze-7* (Alkaline protease 1) proteins. *The red blocks indicate the regions selected for forward and reverse primer design. Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**) and* C. zeae-maydis *(**cezm**).*

The conserved regions were analyzed for possible primer sites, and only one suitable primer pair was possible in the regions indicated in Figure 4.7, with the primer details provided in Table 4.32. Sequence alignments of the forward and reverse sequence regions for the relevant *Cercospora* species are shown in Figure A5 and Figure A6 of the Appendix.

Table 4.32 *Ceze-7* degenerate primer details

Name	Sequence	Length	Tm	Degeneracy	Product size
CEZE-7-DF	YCTCGACTGGATCAACAA	18	51°C	2	358 bp
CEZE-7-DR	RAAGCCTGGTGCRTGTAA	18	55°C	4	

Regions upstream of the *CEZE-7-DF* and downstream of the *CEZE-7-DR* primer sites were analyzed for *C. zeina* specific primer design. Primers were designed with details provided in Table 4.33.

Table 4.33 *Ceze-7 C. zeina*-specific primer details

Name	Sequence	Length	Tm	Product size
CEZE-7-PF	TAGCCGGACAGAAGCTAA	18	59°C	559 bp
CEZE-7-PR	CCGGACAGTTCAACTACA	18	58°C	

4.4.4.8 *Ceze-8* (Putative steroid dehydrogenase 4)

Conserved regions for primer design were identified in the multiple sequence alignment of the *ceze-8* putative steroid dehydrogenase 4 amino acid sequences of the *Cercospora* species (Figure 4.8).

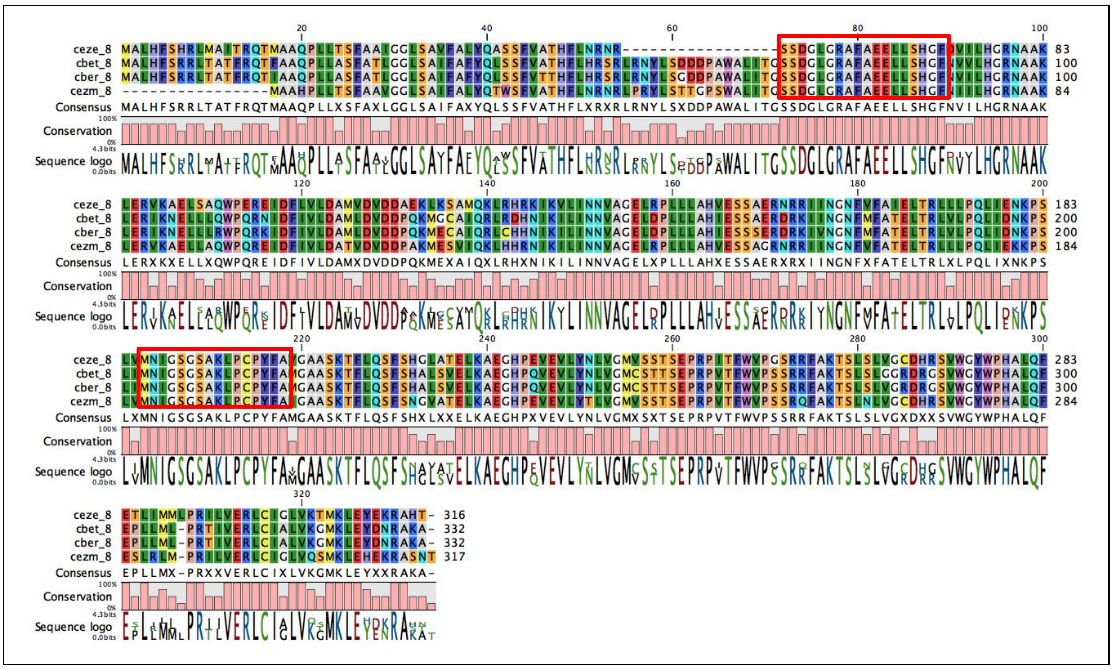


Figure 4.8 Multiple amino acid sequence alignment for *ceze-8* (Putative steroid dehydrogenase 4) proteins. *The red blocks indicate the regions selected for forward and reverse primer design. Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**) and* C. zeae-maydis *(**cezm**).*

The conserved regions were analyzed for possible primer sites, and only one suitable primer pair was possible in the regions indicated in Figure 4.8, with the primer details provided in Table 4.34. Sequence alignments of the forward and reverse sequence regions for the relevant *Cercospora* species are shown in Figure A7 and Figure A8 of the Appendix.

Table 4.34 *Ceze-8* degenerate primer details

Name	Sequence	Length	Tm	Degeneracy	Product size
CEZE-8-DF	CRCYAGCTTCGCGRCYBT	18	60°C	54	499 bp
CEZE-8-DR	VAGCCTRGTCAGCTCTAT	18	54°C	6	

Regions upstream of the *CEZE-8-DF* and downstream of the *CEZE-8-DR* primer sites were analyzed for *C. zeina* specific primer design. Primers were designed with details provided in Table 4.35.

Table 4.35 *Ceze-8 C. zeina*-specific primer details

Name	Sequence	Length	Tm	Product size
CEZE-8-PF	GGCCATCACTCGACAGA	17	60°C	639 bp
CEZE-8-PR	AAGCCGCTCCAACAGCA	17	64°C	

4.4.5 Primer validation

Degenerate and *C. zeina*-specific primers were used in PCRs to validate primer binding, specificity and utility in the amplification and sequencing of the genes of interest.

4.4.5.1 *Ceze-4* (Glucoamylase 1) PCR

The *ceze-4 C. zeina*-specific primer pair was used in PCRs on *C. zeina* genomic DNA (product 602 bp), while the degenerate primer pair was used on positive control PCR products, *C. zeina* and *C. zeae-maydis* genomic DNA (products 466 bp). The PCR reactions were optimized to provide the required bands at the expected sizes (Figure 4.9).

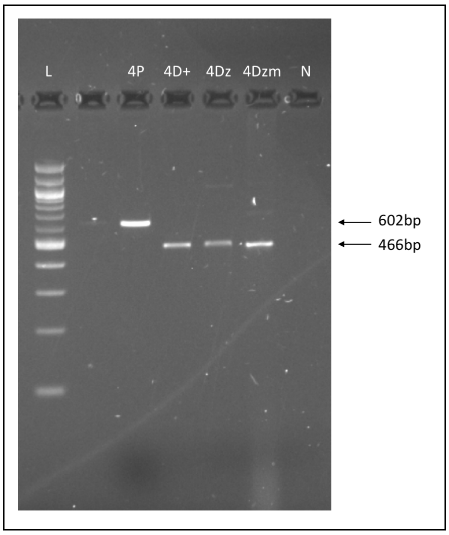


Figure 4.9 Agarose gel electrophoresis gel image of *ceze-4* PCRs. *The lane labels are 100 bp DNA ladder (**L**),* C. zeina-*specific primer products (**4P**),* C. zeina-*specific positive control (**4D+**),* C. zeina *genomic DNA with degenerate primers (**4Dz**),* C. zeae-maydis *genomic DNA with degenerate primers (**4Dzm**) and no-template control (**N**).*

The *ceze-4 C. zeina*-specific primer pair PCR product is specific for the Glucoamylase 1 gene in *C. zeina* according to the sequence alignment of PCR product sequences with *C. zeina* genomic DNA, as well as a BLASTX analysis on the Genbank non-redundant database (Figure 4.10). The sequencing of the *ceze-4* positive control and the *C. zeina* and *C. zeae-maydis* genome products was not successful due to inconsistent base-calling caused by the degenerate primers during sequencing (data not shown). An overflow band is evident in the empty lane between the ladder and 4P lanes, most probably from the 4P lane.

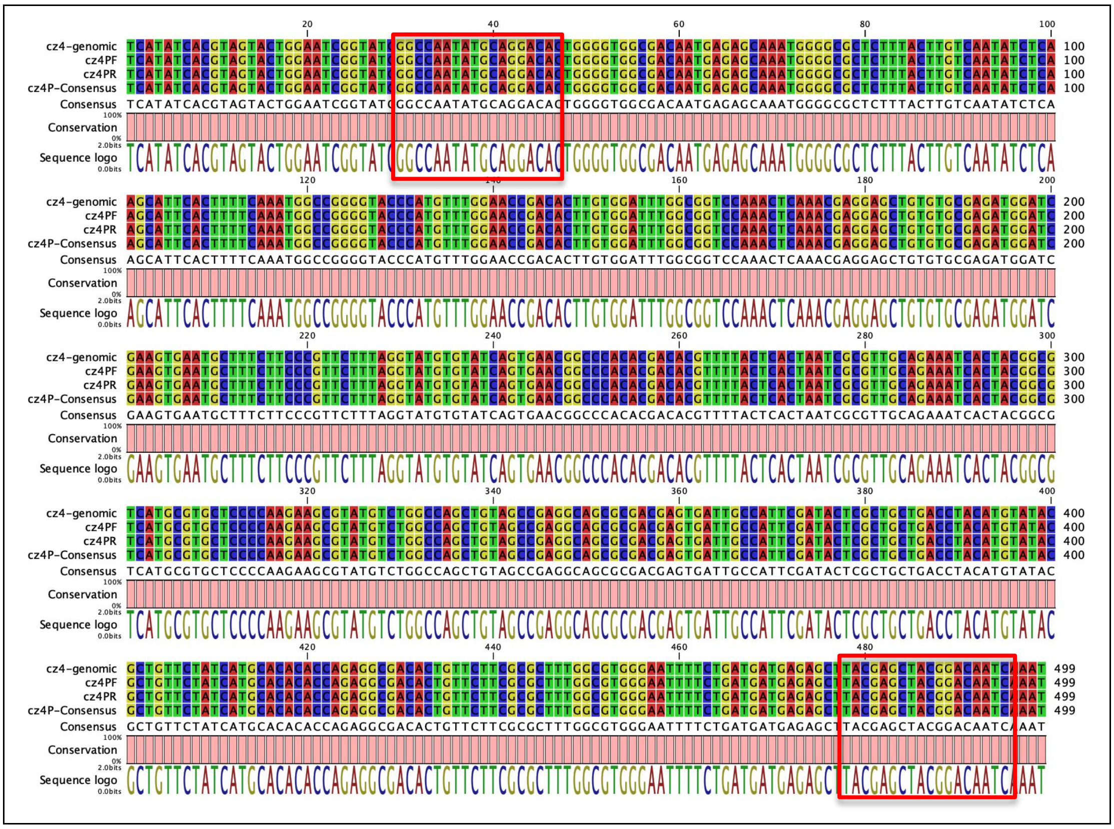


Figure 4.10 Sequence alignment of *ceze-4 C. zeina*-specific PCR product sequences with *C. zeina* genomic DNA. *The sequence labels are* C. zeina *genome sequence (**cz4-genomic**),* C. zeina*-specific forward (**cz4PF**) and reverse (**cz4PR**) primer products and the consensus sequence (**cz4P-Consensus**). The red boxes indicate the forward and reverse degenerate primer binding sequences.*

4.4.5.2 *Ceze-5* (Putative agmatine deiminase) PCR

The *ceze-5 C. zeina*-specific primer pair was used in PCRs on *C. zeina* genomic DNA (product 747 bp), while the degenerate primer pair was used on positive control PCR products and *C. zeina* genomic DNA (expected products sizes 664 bp). The degenerate PCR reactions could not be optimized to amplify the expected regions from the *C. zeina* genomic DNA. Amplification of the expected regions from the positive controls using the degenerate primers were successful, although not the expected size products, nor specific enough to ensure sequencing success. The PCR on the *C. zeae-maydis* genomic DNA was not performed. The primers did therefore recognize the correct sequence regions, but the specific genomic PCRs were not energetically favoured to amplify the regions of interest (Figure 4.11).

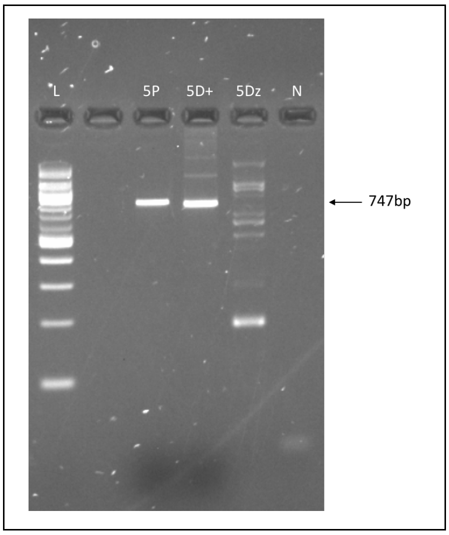


Figure 4.11 Agarose gel electrophoresis gel image of *ceze-5* PCRs. *The lane labels are 100 bp DNA ladder (**L**),* C. zeina*-specific primer products (**5P**),* C. zeina*-specific positive control (**5D+**),* C. zeina *genomic DNA with degenerate primers (**5Dz**) and no-template control (**N**).*

The *ceze-5 C. zeina*-specific primer pair PCR product is specific for a putative Agmatine deiminase gene in *C. zeina* according to the sequence alignment of PCR product sequences with *C. zeina* genomic DNA, as well as a BLASTX analysis on the Genbank non-redundant database (Figure 4.12).

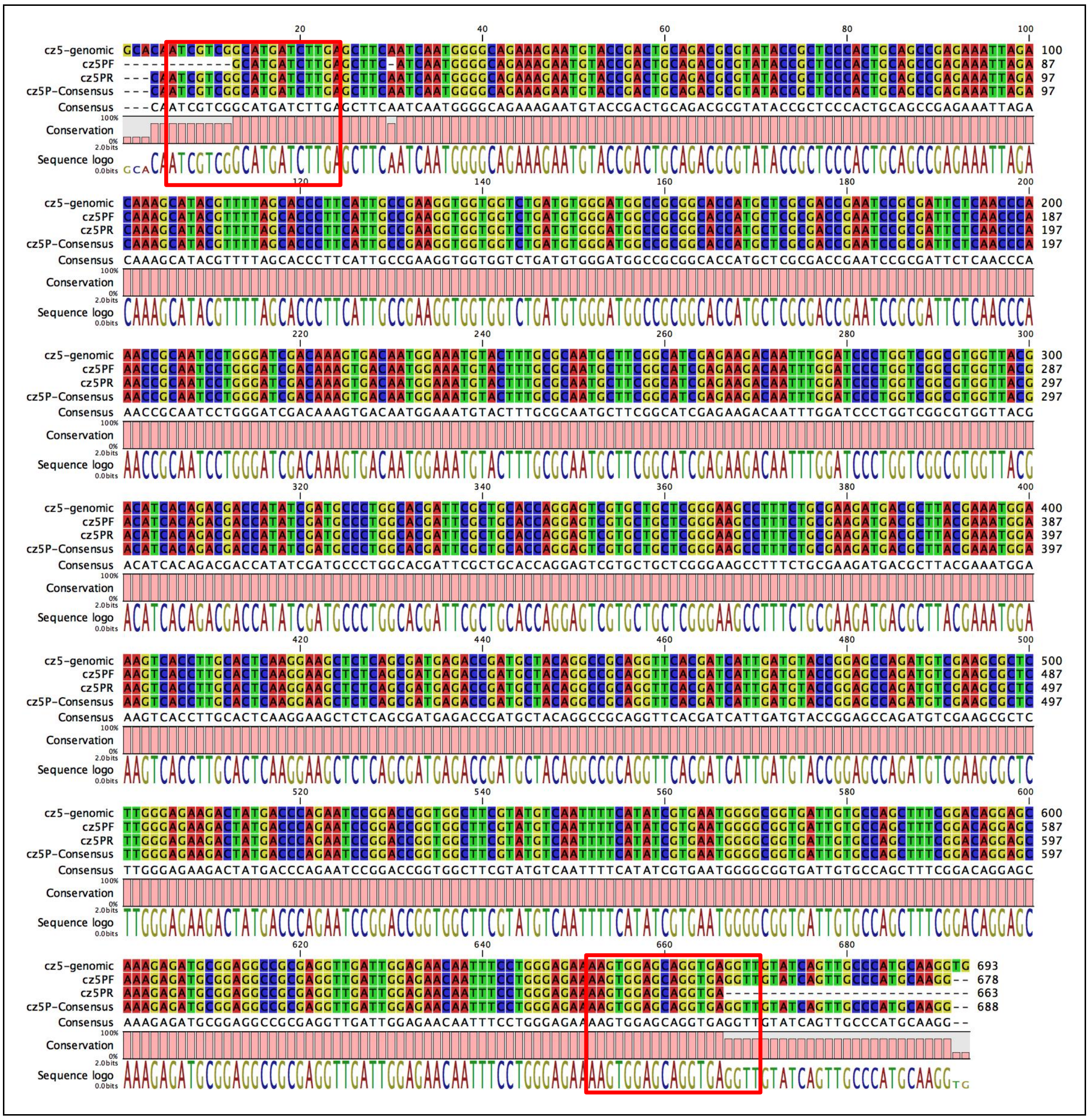


Figure 4.12 Sequence alignment of *ceze-5 C. zeina*-specific PCR product sequences with *C. zeina* genomic DNA. *The sequence labels are* C. zeina *genome sequence (**cz5-genomic**),* C. zeina-*specific forward (**cz5PF**) and reverse (**cz5PR**) primer products and the consensus sequence (**cz5P-Consensus**). The red boxes indicate the forward and reverse degenerate primer binding sequences.*

4.4.5.3 *Ceze-7* (Alkaline protease 1) PCR

The *ceze-7 C. zeina*-specific primer pair was used in PCRs on *C. zeina* genomic DNA (product 559 bp), while the degenerate primer pair was used on positive control PCR products, *C. zeina* and *C. zeae-maydis* genomic DNA (product 358 bp). The PCR reactions were optimized to provide the required bands at the expected sizes (Figure 4.13).

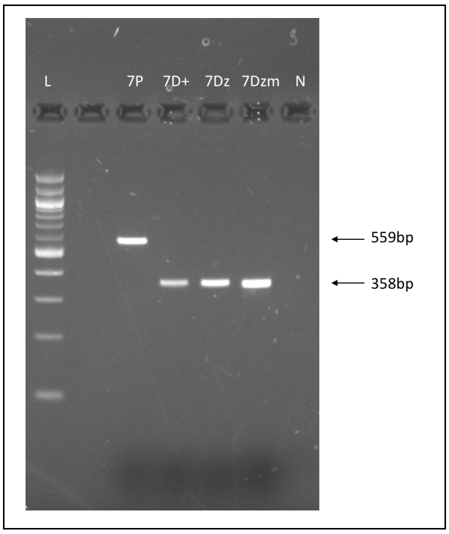


Figure 4.13 Agarose gel electrophoresis gel image of *ceze-7* PCRs. *The lane labels are 100 bp DNA ladder (**L**),* C. zeina*-specific primer products (**7P**),* C. zeina*-specific positive control (**7D+**),* C. zeina *genomic DNA with degenerate primers (**7Dz**),* C. zeae-maydis *genomic DNA with degenerate primers (**7Dzm**) and no-template control (**N**).*

The *ceze-7 C. zeina*-specific primer pair PCR product is specific for the Alkaline protease 1 gene in *C. zeina* according to the sequence alignment of PCR product sequences with *C. zeina* genomic DNA, as well as a BLASTX analysis on the Genbank non-redundant database (Figure 4.14). The sequencing of the *ceze-7* positive control PCR products and the *C. zeina* genomic DNA using the degenerate primers was successful, although sequencing of the *C. zeae-maydis* genome products was not successful due to inconsistent base-calling caused by the degenerate primers during sequencing (data not shown).

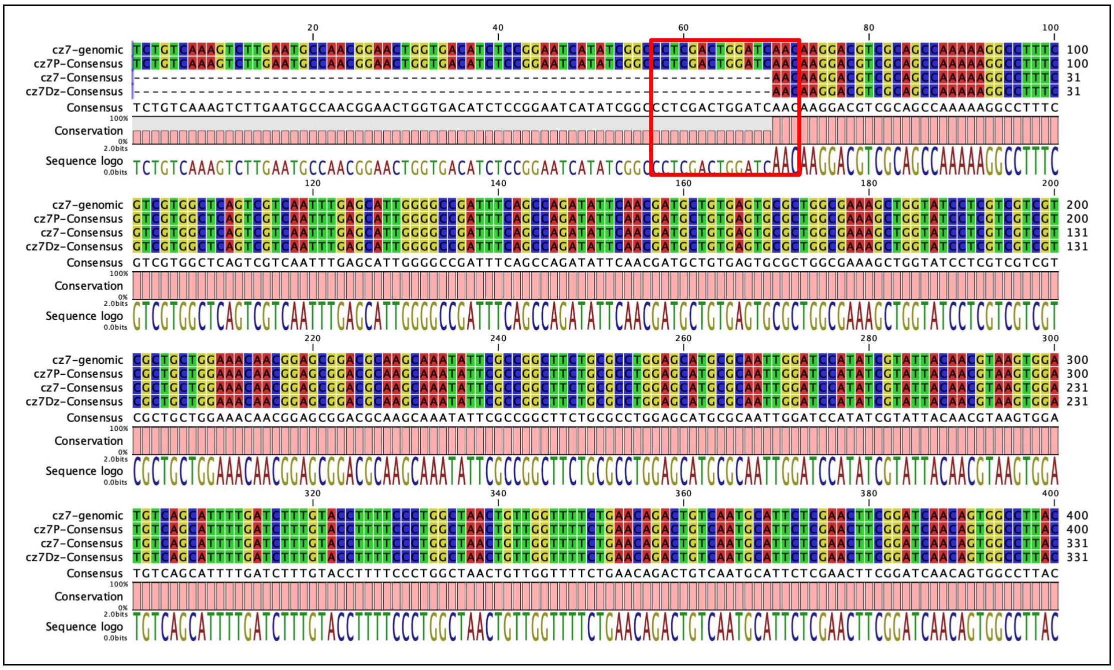


Figure 4.14 Sequence alignment of *ceze-7 C. zeina*-specific PCR product and *C. zeina* genomic sequences with *C. zeina* genomic DNA. *The sequence labels are* C. zeina *genome sequence (**cz7-genomic**),* C. zeina*-specific (**cz7P-Consensus**),* C. zeina *positive control (**cz7-Consensus**) and* C. zeina *degenerate-primer (**cz7Dz-Consensus**) sequences. The red box indicates the forward degenerate primer binding sequence.*

4.4.5.4 *Ceze-8* (Putative steroid dehydrogenase 4) PCR

The *ceze-8 C. zeina*-specific primer pair was used in PCRs on *C. zeina* genomic DNA (product 639 bp), while the degenerate primer pair was used on positive control PCR products and *C. zeina* genomic DNA (products 499 bp). The PCR reactions for the *C. zeina* genomic DNA could not be optimized to yield one single band, and the PCR on the *C. zeae-maydis* genomic DNA was not performed (Figure 4.15).

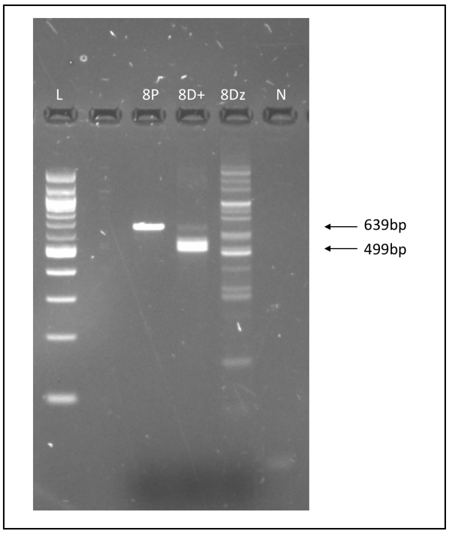


Figure 4.15 Agarose gel electrophoresis gel image of *ceze-8* PCRs. *The lane labels are 100 bp DNA ladder (**L**),* C. zeina*-specific primer products (**8P**),* C. zeina*-specific positive control (**8D+**),* C. zeina *genomic DNA with degenerate primers (**8Dz**) and no-template control (**N**).*

The *ceze-8 C. zeina*-specific primer pair PCR product is specific for a putative Steroid dehydrogenase 4 gene in *C. zeina* according to the sequence alignment of PCR product sequences with *C. zeina* genomic DNA, as well as a BLASTX analysis on the Genbank non-redundant database (Figure 4.16).

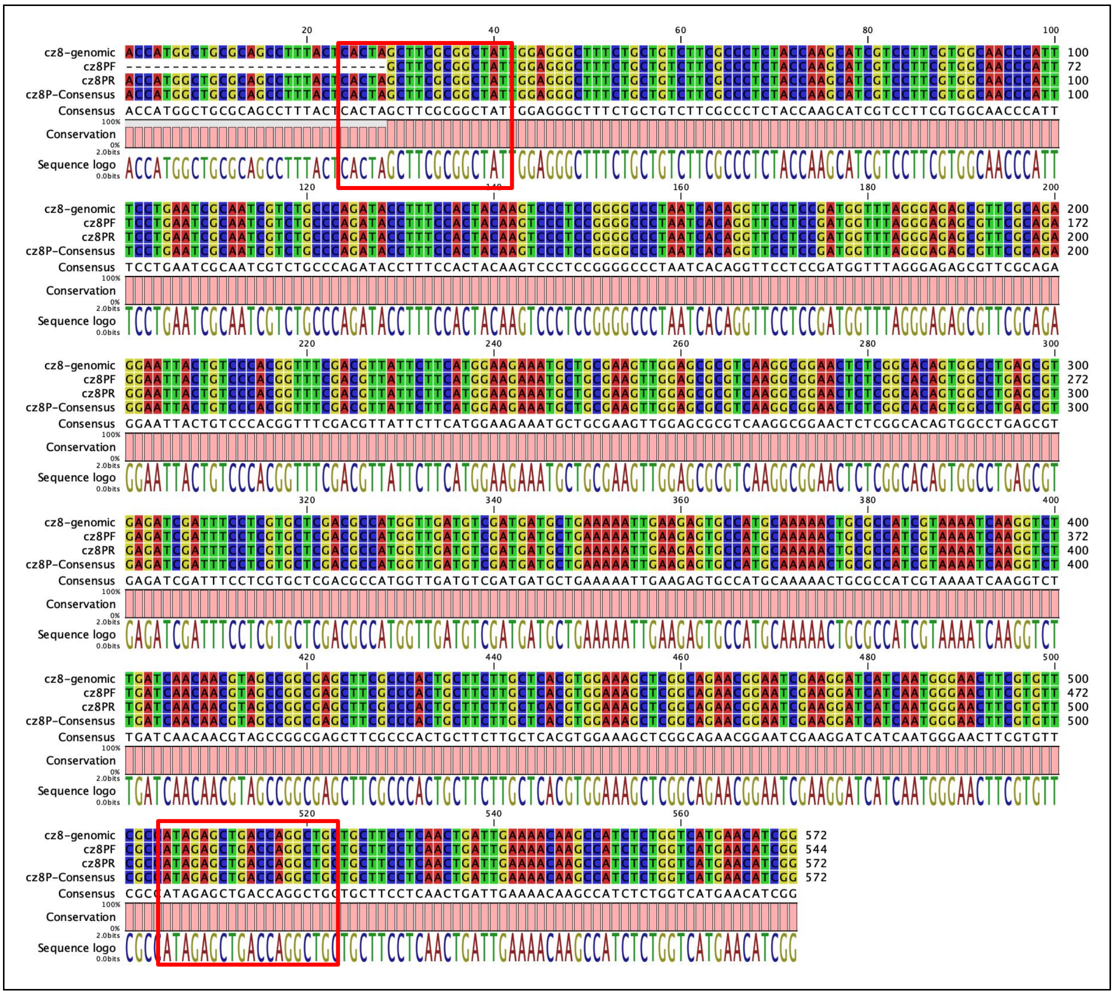


Figure 4.16 Sequence alignment of *ceze-8 C. zeina*-specific PCR product sequences with *C. zeina* genomic DNA. *The sequence labels are* C. zeina *genome sequence (**cz8-genomic**),* C. zeina*-specific forward (**cz8PF**) and reverse (**cz8PR**) primer products and the consensus sequence (**cz8P-Consensus**). The red boxes indicate the forward and reverse degenerate primer binding sequences.*

4.5 Discussion

The study yielded a total of 61 protein orthologs specific for *Cercospora* species which could be used for phylogenetic analyses. A subset of 8 genes were selected for primer design, yielding degenerate primer pairs for two genes which were used to amplify and sequence PCR products in *C. zeina* and *C. zeae-maydis* genomic DNA. The primers are suitable for use in the amplification and sequencing of the genomic DNA of these gene regions in un-classified *Cercospora* species. The additional gene sequences, in combination with sequences from standard phylogenetic genes, should be able to phylogenetically resolve unknown *Cercospora* species to species-level identity.

The non-*Cercospora* fungal species used for the ortholog analysis were selected primarily based on the analysis performed in Ohm *et. al.* (2012). All the species were from the Division Ascomycota, with the majority from the Class *Dothideomycetes* (Table 4.36), similar to the *Cercospora* species.

Table 4.36 Orders and Classes of the fungal species used in the ortholog analyses.

Species	Order	Class
Baudoinia compniacensis	*Capnodiales*	*Dothideomycetes*
Cercospora berteroae		
Cercospora beticola		
Cercospora zeae-maydis		
Cercospora zeina		
Cladosporium fulvum		
Pseudocercospora (Mycosphaerella) fijiensis		
Septoria musiva		
Septoria populicola		
Zymoseptoria tritici		
Rhytidhysteron rufulum	*Hysteriales*	
Cochliobolus heterostrophus	*Pleosporales*	
Cochliobolus sativus		
Leptosphaeria maculans		
Pyrenophora teres		
Pyrenophora tritici-repentis		
Setosphaeria turcica		
Stagonospora nodorum		
Aspergillus nidulans	*Eurotiales*	*Eurotiomycetes*
Aspergillus niger		
Botrytis cinerea	*Helotiales*	*Leotiomycetes*
Sclerotinia sclerotiorum		
Saccharomyces cerevisiae	*Saccharomycetales*	*Saccharomycetes*
Verticillium dahlia	*Hypocreales*	*Sordariomycetes*
Neurospora crassa	*Sordariales*	

Two secondary considerations in species selection were the availability of the genome and proteome sequences, with an added advantage of the inclusion of organisms not in the *Dothideomycete* class to get an expanded protein set for ortholog inference. There were no additional *Cercospora* proteome datasets available, with only three additional genome sequences, i.e. *C. nicotianae*, *C. sojina* and *C. canescens* available, although without annotation of proteome data. These genomes were therefore not useful for confirming or expanding the *Cercospora*-specific ortholog proteins with the other fungal species, but were crucial for the refinement of degenerate primer design criteria and the identification of degenerate base sequences. The future availability of additional *Cercospora* proteomes could be useful for expansion or refinement of *Cercospora*-specific orthologs. In addition, additional *Cercospora* genomes could assist in degenerate primer design and the identification of representative degenerate bases of interest.

Several software packages are available to identify orthologs between organisms. The OrthoFinder package was used due to the fast performance, non-reliance on a relational database, compatibility with the current hardware setup, as well as the output files which could easily be parsed for downstream analyses without additional processing. It was decided to use both BLAST and DIAMOND as similarity search algorithms to benchmark the performance and sensitivity of the respective algorithms. Anecdotal evidence suggests that many current ortholog analyses are primarily performed using DIAMOND due to the algorithm's fast processing and smaller computational burden and cost. The dataset produced during the study produced a 1.36-fold increase in the number of orthologs between the *Cercospora* species than the BLAST analysis. Numerically the BLAST and DIAMOND datasets intersect set comprised of only 71% of the DIAMOND dataset, while it comprised of 91% of the BLAST dataset. This suggests that the DIAMOND algorithm is less stringent using the default analysis settings, and that the stringency of the analysis should be therefore be increased to account for this characteristic. The stringency increase evaluation was not performed during this study, but the results validated the decision to perform both similarity searches and use the intersect dataset for the downstream analyses.

Phylogenetic informativeness is a measure of the variation in the sequences of orthologs which could be used to discriminate species. The G1 statistic is a numerical measure of this function, and a negative G1 statistic following an exhaustive likelihood search is an indication of the suitability of the protein set for phylogenetic analysis. An example of the usefulness is the positive G1 statistic obtained for Calmodulin, Actin and Histone H3, standard genes which are used for the discrimination between different Classes and Orders, but not between species due to the high sequence identity at species-level. For the genes selected following ortholog analysis there was a very high proportion of informative genes, with only one gene failing the analysis. This might be an indication of the ortholog analysis pre-selecting for genes which are useful. A more stringent BLAST/DIAMOND similarity analysis during the OrthoFinder process might select for much more similar orthologs with less variation in sequences, and might thus select for genes which might not be suitable for species-level discrimination. It might, however, be a way to select for genes used in the discrimination at higher levels of organization. The dataset of four sequences per analysis is most probably too small to provide a definitive analysis outcome, although the Calmodulin, Actin and Histone H3 analyses were also performed using four sequences each and did not yield the same negative G1 statistic value. The use of more sequences would be a recommendation, which should give more confidence in the obtained values.

The selected genes all have a functional description, with only two of the genes with descriptions that include the 'Putative' label. When a BLASTP similarity search against the Genbank non-redundant database is performed, most of the genes have descriptions that suggest a similarity in function. Amino acid sequence-alignments of the genes of the four species show high sequence similarity, having regions of perfect identity with some sequence variation in between. The regions of perfect identity were the targets sites to be used for primer design. Even though the selected genes were identified as orthologous, there were still significant sequence variation in some of the genes which disqualified these genes for primer design, specifically the Heterokaryon incompatibility protein 6 (*Ceze-1*) and the choline transport protein (*Ceze-6*). Even though the Carboxylesterase (*Ceze-2*) and Acetylxylan esterase (*Ceze-3*) genes contained regions of perfect identity for primer design, no primers conforming to the design criteria could be found. In addition, the sequence region for *Ceze-2* in *C. zeina* contained no-call bases which were inserted during the scaffolding process. Even though the amino acid sequence conservation is very high, the current genomic DNA sequence data contains a gap in the primer area which cannot be accommodated or overcome.

The four remaining genes all contained regions with sequence identity that could be used for degenerate primer design, and all regions were analyzed for suitable primer sites in the forward and reverse strand (where applicable). The respective primer degeneracies were mostly 12 or below, while only one primer (*CEZE-8-DF*) has a higher degeneracy (56), which was 5-fold higher than the next highest primer (*CEZE-4-DR*). The *C. zeina*-specific primers were designed in the regions directly up- and downstream from the degenerate primer sites, and for all the primer sites there were only one acceptable primer sequence possible. For the putative steroid dehydrogenase 4 gene there were no *C. zeina*-specific primer sites available with the specified parameters due to the close proximity of the start codon om the five-prime end of the sequence. The relaxation of the design parameters to accept 17-mer primers led to the successful design of the *CEZE-8PF* and *CEZE-8PR* primers. As expected, the great majority of the degenerate bases in the degenerate primer regions were situated on the wobble position of codons.

The *C. zeina*-specific primes for all four of the genes successfully amplified the correct regions on the *C. zeina* genomic DNA without additional optimization of the PCR parameters required. The correct size bands were amplified as evident from the agarose gel images (Figures 4.10, 4.12, 4.14 and 4.16). Following Sanger sequencing of the respective PCR products, the correct sequences were obtained after alignments with the gene region sequences from the *C. zeina* genome assembly (Figures 4.11, 4.13, 4.15 and 4.17). For all the genes the degenerate primer binding sequences could be identified form the sequence alignments, except for the reverse degenerate primer for the Alkaline protease 1 gene region (*CEZE-7DR*).

The *C. zeina*-specific primer products were used as positive control for the binding of the degenerate primers to the *C. zeina* sequences. During the optimization of the PCR reactions for amplifying the PCR products, the templates were diluted to decrease the template/primer ratio, since multiple products were initially amplified. The primer annealing times were also reduced, as were the extension times. The optimization yielded amplification of the correct regions for all the genes, except for the putative agmatine deiminase gene primers which amplified a band too large in size, and the PCR could not be optimized satisfactorily.

The amplification of the respective gene regions from the *C. zeina* genomic DNA was successful for only the Glucoamylase 1 and Alkaline protease 1 genes. These PCR products were the correct sizes (Figures 4.10 and 4.15), while sequencing of the product of the Alkaline protease 1 degenerate primers indicated the correct sequence following sequence alignment. Sequencing of the Glucoamylase 1 PCR product was not successful, with a mixed sequence obtained which was not useable. PCR optimization for the putative agmatine deiminase 5 and steroid dehydrogenase 4 genes were attempted without success. Initially the annealing temperatures were decreased while the primer concentrations were increased. Subsequently a step-up reaction was performed, from an annealing temperature of 5°C degrees below Tm for 5 cycles followed by 30 cycles at 2°C below Tm. The third optimization involved reactions on a Taguchi PCR optimization reaction where the primer and template concentrations were adjusted in parallel. None of the PCRs showed amplifications of the bands of interest, and for each reaction multiple bands were amplified. The PCR optimization for these genes were abandoned, and the genes were removed from consideration due to the unavailability of additional primer binding sites.

Due to the success in amplifying the gene regions from the *C. zeina* genomic DNA, only the degenerate primers for the Glucoamylase 1 and Alkaline protease 1 genes were used in the amplification of gene regions from *C. zeae-maydis* gene regions. The primers for the Alkaline protease 1 gene was successful at low template input and did not need optimization. The primers for the Glucoamylase 1 gene required additional optimization to amplify the product size of interest. Unfortunately the sequencing reactions for both these regions failed, but can be deemed successful due to the correct sizes of products amplified.

The successful amplification and sequencing of the two gene regions served as initial proof of concept for the gene selection and primer design process. The final proof of the procedure would be the successful implementation of new Glucoamylase 1 and Alkaline protease 1 gene regions from the unclassified *Cercospora* species in an extended phylogenetic study which resolves the unclassified species into their correct clades. The failure of the sequencing reactions on the *C. zeae-maydis* genomic DNA was not unforeseen due to the difficulty in sequencing using degenerate primers, as previously studied (Elbrecht *et al.*, 2018). The most appropriate solution for this problem is the use of the degenerate primers to amplify the regions of choice from the genomic DNA of the unclassified species. The PCR products can be cloned into plasmid vectors. Following the cloning of these vectors into the appropriate bacteria, the primer regions on the cloning vectors can be used to successfully sequence the gene regions. This approach should provide high-quality sequence data for the relevant genes. Alternatively the amplified PCR products could be subjected to selective restriction enzyme digestion to identify sequence differences via differences in digestion fragment size patterns.

The *Cercospora* genus is predicted to have split from the most recent common ancestor ~19 – 47 million years ago (Zaccaron *et al.*, 2015). It is presumably following this event that the *Cercospora*-specific proteins started to diverge from those of related genera. The two *Cercospora*-specific enzymes successfully amplified have functions that were found to play important roles in fungal pathogenesis and nutrition. The Glucoamylase 1 enzyme has been found to be important for fungal nutrition. Glucoamylases are classified as 1,4-α-d-glucan glucohydrolases (Zhao *et al.*, 2000), and play important roles in degrading starch, one of the major carbohydrate storage molecules in plants (Zeeman *et al.*, 2010). In Basidiomycetous fungi the similarity between glucoamylases is between 45 – 61%, and a similar trend is possible in Ascomycetous fungi (Zhao *et al.*, 2000). Such trends might explain the lower similarity found during the orthologue analysis. Due to the importance of starch for nutrition the similarity between these enzymes in the *Cercospora* genus is to be expected, especially if these species employ similar nutrition acquisition mechanisms. The Alkaline protease 1 enzyme is not well characterized in terms of specific function, though proteases have been found to be important for both nutrition and pathogenesis (Jashni *et al.*, 2015 ; Hou *et al.*, 2018). Similar pathogenesis mechanisms in *Cercospora* species might require a similar proteases which would explain the similarity of the enzyme in species in the *Cercospora* genus.

4.6 References

Aguileta, G., Marthey, S., Chiapello, H., Lebrun, M. H., Rodolphe, F., Fournier, E., Gendrault-Jacquemard, A., and Giraud, T. (2008) Assessing the performance of single-copy genes for recovering robust phylogenies. *Systematic Biology* **57(4)**:613-627

Altenhoff, A. M., and Dessimoz, C. (2009) Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Computational Biology* **5(1)**:e1000262

Altenhoff, A. M., Schneider, A., Gonnet, G. H., and Dessimoz, C. (2011) OMA 2011: orthology inference among 1000 complete genomes. *Nucleic Acids Research* **39(Database issue)**:D289-D294

Altenhoff, A. M., Glover, N. M., Train, C. M., Kaleb, K., Warwick Vesztrocy, A., Dylus, D., de Farias, T. M., Zile, K., Stevenson, C., Long, J., Redestig, H., Gonnet, G. H., and Dessimoz, C. (2018) The OMA orthology database in 2018: retrieving evolutionary relationships among all domains of life through richer web and programmatic interfaces. *Nucleic Acids Research* **46(D1)**:D477-D485

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215(3)**:403-410

Amselem, J., Cuomo, C. A., van Kan, J. A., Viaud, M., Benito, E. P., Couloux, A., Coutinho, P. M., de Vries, R. P., Dyer, P. S., Fillinger, S., Fournier, E., Gout, L., Hahn, M., Kohn, L., Lapalu, N., Plummer, K. M., Pradier, J. M., Quevillon, E., Sharon, A., Simon, A., ten Have, A., Tudzynski, B., Tudzynski, P., Wincker, P., Andrew, M., Anthouard, V., Beever, R. E., Beffa, R., Benoit, I., Bouzid, O., Brault, B., Chen, Z., Choquer, M., Collemare, J., Cotton, P., Danchin, E. G., Da Silva, C., Gautier, A., Giraud, C., Giraud, T., Gonzalez, C., Grossetete, S., Guldener, U., Henrissat, B., Howlett, B. J., Kodira, C., Kretschmer, M., Lappartient, A., Leroch, M., Levis, C., Mauceli, E., Neuveglise, C., Oeser, B., Pearson, M., Poulain, J., Poussereau, N., Quesneville, H., Rascle, C., Schumacher, J., Segurens, B., Sexton, A., Silva, E., Sirven, C., Soanes, D. M., Talbot, N. J., Templeton, M., Yandava, C., Yarden, O., Zeng, Q., Rollins, J. A., Lebrun, M. H., and Dickman, M. (2011) Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *PLoS Genetics* **7(8)**:e1002230

Arango Isaza, R. E., Diaz-Trujillo, C., Dhillon, B., Aerts, A., Carlier, J., Crane, C. F., T, V. d. J., de Vries, I., Dietrich, R., Farmer, A. D., Fortes Fereira, C., Garcia, S., Guzman, M., Hamelin, R. C., Lindquist, E. A., Mehrabi, R., Quiros, O., Schmutz, J., Shapiro, H., Reynolds, E., Scalliet, G., Souza, M., Jr., Stergiopoulos, I., Van der Lee, T. A., De Wit, P. J., Zapater, M. F., Zwiers, L. H., Grigoriev, I. V., Goodwin, S. B., and Kema, G. H. (2016) Combating a global threat to a clonal crop: banana black Sigatoka pathogen *Pseudocercospora fijiensis* (Synonym *Mycosphaerella fijiensis*) genomes reveal clues for disease control. *PLoS Genetics* **12(8)**:e1005876

Arnaud, M. B., Cerqueira, G. C., Inglis, D. O., Skrzypek, M. S., Binkley, J., Chibucos, M. C., Crabtree, J., Howarth, C., Orvis, J., Shah, P., Wymore, F., Binkley, G., Miyasato, S. R., Simison, M., Sherlock, G., and Wortman, J. R. (2012) The *Aspergillus* Genome Database (AspGD): recent developments in comprehensive multispecies curation, comparative genomics and community resources. *Nucleic Acids Research* **40(Database issue)**:D653-D659

Baker, S. E., Schackwitz, W., Lipzen, A., Martin, J., Haridas, S., LaButti, K., Grigoriev, I. V., Simmons, B. A., and McCluskey, K. (2015) Draft genome sequence of *Neurospora crassa* strain FGSC 73. *Genome Announcements* **3(2)**: e00074-15

BioEdit. http://www.mbio.ncsu.edu/BioEdit/bioedit.html.

BLAST. https://blast.ncbi.nlm.nih.gov.

Buchfink, B., Xie, C., and Huson, D. H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12(1)**:59-60

Candek, K., and Kuntner, M. (2015) DNA barcoding gap: reliable species identification over morphological and geographical scales. *Molecular Ecology Resources* **15(2)**:268-77

CBOL. http://www.barcodeoflife.org/.

China Plant BOL Group, Li, D., Gao, L., Li, H., Wang, H., Ge, X., Liu, J., Chen, Z., Zhou, S., Chen, S., Yang, J., Fu, C., Zeng, C., Yan, H., Zhu, Y., Sun, Y., Chen, S., Zhao, L., Wang, K., Yang, T., and Duana, G. (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. *Proceedings of the National Academy of Sciences of the United States of America* **108(49)**:19641-19646

CLC. http://www.qiagenbioinformatics.com/products/clc-main-workbench/.

Condon, B. J., Leng, Y., Wu, D., Bushley, K. E., Ohm, R. A., Otillar, R., Martin, J., Schackwitz, W., Grimwood, J., MohdZainudin, N., Xue, C., Wang, R., Manning, V. A., Dhillon, B., Tu, Z. J., Steffenson, B. J., Salamov, A., Sun, H., Lowry, S., LaButti, K., Han, J., Copeland, A., Lindquist, E., Barry, K., Schmutz, J., Baker, S. E., Ciuffetti, L. M., Grigoriev, I. V., Zhong, S., and Turgeon, B. G. (2013) Comparative genome structure, secondary metabolite, and effector coding capacity across *Cochliobolus* pathogens. *PLoS Genetics* **9(1)**:e1003233

de Jonge, R., Ebert, M. K., Huitt-Roehl, C. R., Pal, P., Suttle, J. C., Spanner, R. E., Neubauer, J. D., Jurick, W. M., Stott, K. A., Secor, G. A., Thomma, B. P. H. J., Van de Peer, Y., Townsend, C. A., and Bolton, M. D. (2018) Gene cluster conservation provides insight into cercosporin biosynthesis and extends production to the genus *Colletotrichum*. *Proceedings of the National Academy of Sciences of the United States of America* **115(24)**:E5459-E5466

de Wit, P. J., van der Burgt, A., Okmen, B., Stergiopoulos, I., Abd-Elsalam, K. A., Aerts, A. L., Bahkali, A. H., Beenen, H. G., Chettri, P., Cox, M. P., Datema, E., de Vries, R. P., Dhillon, B., Ganley, A. R., Griffiths, S. A., Guo, Y., Hamelin, R. C., Henrissat, B., Kabir, M. S., Jashni, M. K., Kema, G., Klaubauf, S., Lapidus, A., Levasseur, A., Lindquist, E., Mehrabi, R., Ohm, R. A., Owen, T. J., Salamov, A., Schwelm, A., Schijlen, E., Sun, H., van den Burg, H. A., van Ham, R. C., Zhang, S., Goodwin, S. B., Grigoriev, I. V., Collemare, J., and Bradshaw, R. E. (2012) The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. *PLoS Genetics* **8(11)**:e1003088

Dessimoz, C., Cannarozzi, G., Gil, M., Margadant, D., Roth, A., Schneider, A., and Gonnet, G. H. (2005) OMA, A Comprehensive, automated project for the identification of orthologs from complete genome data: Introduction and first achievements. Edited by A. McLysaght, and D. H. Huson, *Comparative Genomics*. Springer, Berlin, Heidelberg.

Dhillon, B., Feau, N., Aerts, A. L., Beauseigle, S., Bernier, L., Copeland, A., Foster, A., Gill, N., Henrissat, B., Herath, P., LaButti, K. M., Levasseur, A., Lindquist, E. A., Majoor, E., Ohm, R. A., Pangilinan, J. L., Pribowo, A., Saddler, J. N., Sakalidis, M. L., de Vries, R. P., Grigoriev, I. V., Goodwin, S. B., Tanguay, P., and Hamelin, R. C. (2015) Horizontal gene transfer and gene dosage drives adaptation to wood colonization in a tree pathogen. *Proceedings of the National Academy of Sciences of the United States of America* **112(11)**:3451-3456

Elbrecht, V., Hebert, P. D. N., and Steinke, D. (2018) Slippage of degenerate primers can cause variation in amplicon length. *Scientific Reports* **8(1)**:10999

Ellwood, S. R., Liu, Z., Syme, R. A., Lai, Z., Hane, J. K., Keiper, F., Moffat, C. S., Oliver, R. P., and Friesen, T. L. (2010) A first genome assembly of the barley fungal pathogen *Pyrenophora teres* f. teres. *Genome Biology* **11(11)**:R109

Emms, D. M., and Kelly, S. (2015) OrthoFinder: solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology* **16**:157

Farris, J. S. (1989) The retention index and the rescaled consistency index. *Cladistics* **5**:417-419

Galagan, J. E., Henn, M. R., Ma, L. J., Cuomo, C. A., and Birren, B. (2005) Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Research* **15(12)**:1620-1631

Goodwin, S. B., M'Barek S, B., Dhillon, B., Wittenberg, A. H., Crane, C. F., Hane, J. K., Foster, A. J., Van der Lee, T. A., Grimwood, J., Aerts, A., Antoniw, J., Bailey, A., Bluhm, B., Bowler, J., Bristow, J., van der Burgt, A., Canto-Canche, B., Churchill, A. C., Conde-Ferraez, L., Cools, H. J., Coutinho, P. M., Csukai, M., Dehal, P., De Wit, P., Donzelli, B., van de Geest, H. C., van Ham, R. C., Hammond-Kosack, K. E., Henrissat, B., Kilian, A., Kobayashi, A. K., Koopmann, E., Kourmpetis, Y., Kuzniar, A., Lindquist, E., Lombard, V., Maliepaard, C., Martins, N., Mehrabi, R., Nap, J. P., Ponomarenko, A., Rudd, J. J., Salamov, A., Schmutz, J., Schouten, H. J., Shapiro, H., Stergiopoulos, I., Torriani, S. F., Tu, H., de Vries, R. P., Waalwijk, C., Ware, S. B., Wiebenga, A., Zwiers, L. H., Oliver, R. P., Grigoriev, I. V., and Kema, G. H. (2011) Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genetics* **7(6)**:e1002070

Groenewald, J. Z., Nakashima, C., Nishikawa, J., Shin, H. D., Park, J. H., Jama, A. N., Groenewald, M., Braun, U., and Crous, P. W. (2013) Species concepts in *Cercospora*: spotting the weeds among the roses. *Studies in Mycology* **75(1)**:115-170

Hane, J. K., Lowe, R. G., Solomon, P. S., Tan, K. C., Schoch, C. L., Spatafora, J. W., Crous, P. W., Kodira, C., Birren, B. W., Galagan, J. E., Torriani, S. F., McDonald, B. A., and Oliver, R. P. (2007) Dothideomycete plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. *Plant Cell* **19(11)**:3347-3368

Hebert, P. D., Stoeckle, M. Y., Zemlak, T. S., and Francis, C. M. (2004) Identification of birds through DNA barcodes. *PLoS Biology* **2(10)**:e312

Hebert, P. D. N., Cywinska, A., Ball, S. L., and deWaard, J. R. (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society B: Biological Sciences* **270(1512)**:313–321

Hillis, D. M. (1991) Discriminating between phylogenetic signal and random noise in DNA sequences, *Phylogenetic analysis of DNA sequences*. Oxford University Press, New York.

Hillis, D. M., and Huelsenbeck, J. P. (1992) Signal, noise, and reliability in molecular phylogenetic analyses. *Journal of Heredity* **83(3)**:189-195

Hou, S., Jamieson, P., and He, P. (2018) The cloak, dagger, and shield: proteases in plant-pathogen interactions. *Biochemical Journal* **475(15)**:2491-2509

Huelsenbeck, J. P. (1991) Tree-length distribution skewness: an indicator of phylogenetic information. *Systematic Zoology* **40(3)**:257-270

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., and Bork, P. (2017) Fast genome-wide functional annotation through orthology assignment by eggNOG-Mapper. *Molecular Biology and Evolution* **34(8)**:2115-2122

Huerta-Cepas, J., Szklarczyk, D., Forslund, K., Cook, H., Heller, D., Walter, M. C., Rattei, T., Mende, D. R., Sunagawa, S., Kuhn, M., Jensen, L. J., von Mering, C., and Bork, P. (2016) eggNOG 4.5: a hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences. *Nucleic Acids Research* **44(D1)**:D286-D293

Jashni, M. K., Mehrabi, R., Collemare, J., Mesarich, C. H., and de Wit, P. J. (2015) The battle in the apoplast: further insights into the roles of proteases and their inhibitors in plant-pathogen interactions. *Frontiers in Plant Science* **6**:584

Klosterman, S. J., Subbarao, K. V., Kang, S., Veronese, P., Gold, S. E., Thomma, B. P., Chen, Z., Henrissat, B., Lee, Y. H., Park, J., Garcia-Pedrajas, M. D., Barbara, D. J., Anchieta, A., de Jonge, R., Santhanam, P., Maruthachalam, K., Atallah, Z., Amyotte, S. G., Paz, Z., Inderbitzin, P., Hayes, R. J., Heiman, D. I., Young, S., Zeng, Q., Engels, R., Galagan, J., Cuomo, C. A., Dobinson, K. F., and Ma, L. J. (2011) Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. *PLoS Pathogens* **7(7)**:e1002137

Koonin, E. V. (2005) Orthologs, paralogs, and evolutionary genomics. *Annual Reviews in Genetics* **39**:309-338

Kress, W. J., Wurdack, K. J., Zimmer, E. A., Weigt, L. A., and Janzen, D. H. (2005) Use of DNA barcodes to identify flowering plants. *Proceedings of the National Academy of Sciences of the United States of America* **102(23)**:8369-8374

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simao, F. A., and Zdobnov, E. M. (2018) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Research* **47(D1)**:D807-D811

Kriventseva, E. V., Tegenfeldt, F., Petty, T. J., Waterhouse, R. M., Simao, F. A., Pozdnyakov, I. A., Ioannidis, P., and Zdobnov, E. M. (2015) OrthoDB v8: update of the hierarchical catalog of orthologs and the underlying free software. *Nucleic Acids Research* **43(Database issue)**:D250-D256

Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics* **23(21)**:2947-2948

Li, L., Stoeckert, C. J., Jr., and Roos, D. S. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Research* **13(9)**:2178-2189

Lopez-Giraldez, F., Moeller, A. H., and Townsend, J. P. (2013) Evaluating phylogenetic informativeness as a predictor of phylogenetic signal for metazoan, fungal, and mammalian phylogenomic data sets. *BioMed Research International* **2013**:621604

Maddison, D. R., Swofford, D. L., and Maddison, W. P. (1997) NEXUS: an extensible file format for systematic information. *Systems Biology* **46(4)**:590-621

Manning, V. A., Pandelova, I., Dhillon, B., Wilhelm, L. J., Goodwin, S. B., Berlin, A. M., Figueroa, M., Freitag, M., Hane, J. K., Henrissat, B., Holman, W. H., Kodira, C. D., Martin, J., Oliver, R. P., Robbertse, B., Schackwitz, W., Schwartz, D. C., Spatafora, J. W., Turgeon, B. G., Yandava, C., Young, S., Zhou, S., Zeng, Q., Grigoriev, I. V., Ma, L. J., and Ciuffetti, L. M. (2013) Comparative genomics of a plant-pathogenic fungus, *Pyrenophora tritici-repentis*, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. *G3 (Bethesda)* **3(1)**:41-63

Mi, H., Muruganujan, A., and Thomas, P. D. (2013) PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. *Nucleic Acids Research* **41(Database issue)**:D377-86

Ohm, R. A., Feau, N., Henrissat, B., Schoch, C. L., Horwitz, B. A., Barry, K. W., Condon, B. J., Copeland, A. C., Dhillon, B., Glaser, F., Hesse, C. N., Kosti, I., LaButti, K., Lindquist, E. A., Lucas, S., Salamov, A. A., Bradshaw, R. E., Ciuffetti, L., Hamelin, R. C., Kema, G. H., Lawrence, C., Scott, J. A., Spatafora, J. W., Turgeon, B. G., de Wit, P. J., Zhong, S., Goodwin, S. B., and Grigoriev, I. V. (2012) Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PLoS Pathogens* **8(12)**:e1003037

Pel, H. J., de Winde, J. H., Archer, D. B., Dyer, P. S., Hofmann, G., Schaap, P. J., Turner, G., de Vries, R. P., Albang, R., Albermann, K., Andersen, M. R., Bendtsen, J. D., Benen, J. A., van den Berg, M., Breestraat, S., Caddick, M. X., Contreras, R., Cornell, M., Coutinho, P. M., Danchin, E. G., Debets, A. J., Dekker, P., van Dijck, P. W., van Dijk, A., Dijkhuizen, L., Driessen, A. J., d'Enfert, C., Geysens, S., Goosen, C., Groot, G. S., de Groot, P. W., Guillemette, T., Henrissat, B., Herweijer, M., van den Hombergh, J. P., van den Hondel, C. A., van der Heijden, R. T., van der Kaaij, R. M., Klis, F. M., Kools, H. J., Kubicek, C. P., van Kuyk, P. A., Lauber, J., Lu, X., van der Maarel, M. J., Meulenberg, R., Menke, H., Mortimer, M. A., Nielsen, J., Oliver, S. G., Olsthoorn, M., Pal, K., van Peij, N. N., Ram, A. F., Rinas, U., Roubos, J. A., Sagt, C. M., Schmoll, M., Sun, J., Ussery, D., Varga, J., Vervecken, W., van de Vondervoort, P. J., Wedler, H., Wosten, H. A., Zeng, A. P., van Ooyen, A. J., Visser, J., and Stam, H. (2007) Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. *Nature Biotechnology* **25(2)**:221-231

Pentinsaari, M., Salmela, H., Mutanen, M., and Roslin, T. (2016) Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. *Scientific Reports* **6**:35275

Quinlan, A. R. (2014) BEDTools: The Swiss-army tool for genome feature analysis. *Current Protocols in Bioinformatics* **47**:11.12.1-11.12.34

Raja, H. A., Miller, A. N., Pearce, C. J., and Oberlies, N. H. (2017) Fungal identification using molecular tools: a primer for the natural products research community. *Journal of Natural Products* **80(3)**:756-770

Rattei, T., Arnold, R., Tischler, P., Lindner, D., Stumpflen, V., and Mewes, H. W. (2006) SIMAP: the similarity matrix of proteins. *Nucleic Acids Research* **34(Database issue)**:D252-D256

Remm, M., Storm, C. E., and Sonnhammer, E. L. (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology* **314(5)**:1041-1052

Rouxel, T., Grandaubert, J., Hane, J. K., Hoede, C., van de Wouw, A. P., Couloux, A., Dominguez, V., Anthouard, V., Bally, P., Bourras, S., Cozijnsen, A. J., Ciuffetti, L. M., Degrave, A., Dilmaghani, A., Duret, L., Fudal, I., Goodwin, S. B., Gout, L., Glaser, N., Linglin, J., Kema, G. H., Lapalu, N., Lawrence, C. B., May, K., Meyer, M., Ollivier, B., Poulain, J., Schoch, C. L., Simon, A., Spatafora, J. W., Stachowiak, A., Turgeon, B. G., Tyler, B. M., Vincent, D., Weissenbach, J., Amselem, J., Quesneville, H., Oliver, R. P., Wincker, P., Balesdent, M. H., and Howlett, B. J. (2011) Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. *Nature Communications* **2**:202

Savolainen, V., Cowan, R. S., Vogler, A. P., Roderick, G. K., and Lane, R. (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding. *Philosophical Transactions of the Royal Society B: Biological Sciences* **360(1462)**:1805–1811

Schoch, C. L., Seifert, K. A., Huhndorf, S., Robert, V., Spouge, J. L., Levesque, C. A., Chen, W., Fungal Barcoding, C., and Fungal Barcoding Consortium Author, L. (2012) Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proceedings of the National Academy of Sciences of the United States of America* **109(16)**:6241-6246

Snel, B., Bork, P., and Huynen, M. A. (1999) Genome phylogeny based on gene content. *Nature Genetics* **21(1)**:108-110

Sci-Ed Software. http://www.scied.com.

Sonnhammer, E. L., and Ostlund, G. (2015) InParanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research* **43(Database issue)**:D234-D239

Staats, M., and van Kan, J. A. (2012) Genome update of *Botrytis cinerea* strains B05.10 and T4. *Eukaryot Cell* **11(11)**:1413-1414

Steinegger, M., and Soding, J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology* **35(11)**:1026-1028

Stielow, J. B., Levesque, C. A., Seifert, K. A., Meyer, W., Iriny, L., Smits, D., Renfurm, R., Verkley, G. J., Groenewald, M., Chaduli, D., Lomascolo, A., Welti, S., Lesage-Meessen, L., Favel, A., Al-Hatmi, A. M., Damm, U., Yilmaz, N., Houbraken, J., Lombard, L., Quaedvlieg, W., Binder, M., Vaas, L. A., Vu, D., Yurkov, A., Begerow, D., Roehl, O., Guerreiro, M., Fonseca, A., Samerpitak, K., van Diepeningen, A. D., Dolatabadi, S., Moreno, L. F., Casaregola, S., Mallet, S., Jacques, N., Roscini, L., Egidi, E., Bizet, C., Garcia-Hermoso, D., Martin, M. P., Deng, S., Groenewald, J. Z., Boekhout, T., de Beer, Z. W., Barnes, I., Duong, T. A., Wingfield, M. J., de Hoog, G. S., Crous, P. W., Lewis, C. T., Hambleton, S., Moussa, T. A., Al-Zahrani, H. S., Almaghrabi, O. A., Louis-Seize, G., Assabgui, R., McCormick, W., Omer, G., Dukik, K., Cardinali, G., Eberhardt, U., de Vries, M., and Robert, V. (2015) One fungus, which genes? Development and assessment of universal primers for potential secondary fungal DNA barcodes. *Persoonia* **35**:242-263

Townsend, J. P. (2007) Profiling phylogenetic informativeness. *Systematic Biology* **56(2)**:222-231

Wilgenbusch, J. C., and Swofford, D. (2003) Inferring evolutionary trees with PAUP*. *Curr Protoc Bioinformatics* **Chapter 6**:Unit 6.4

Wingfield, B. D., Berger, D. K., Steenkamp, E. T., Lim, H. J., Duong, T. A., Bluhm, B. H., de Beer, Z. W., De Vos, L., Fourie, G., Naidoo, K., Olivier, N., Lin, Y. C., Van de Peer, Y., Joubert, F., Crampton, B. G., Swart, V., Soal, N., Tatham, C., van der Nest, M. A., van der Merwe, N. A., van Wyk, S., Wilken, P. M., and Wingfield, M. J. (2017) IMA Genome-F 8: Draft genome of *Cercospora zeina*, *Fusarium pininemorale*, *Hawksworthiomyces lignivorus*, *Huntiella decipiens* and *Ophiostoma* ips. *IMA Fungus* **8(2)**:385-396

Zaccaron, A., Ridenour, J., Smith, J., Sharma, S., Lawson, N., Zaccaron, M. L., Fakhoury, A., and Bluhm, B. H. (2015) Molecular mechanisms of *Cercospora* pathogenicity revealed through comparative genomics. APS Annual Meeting. Pasadena, California, USA.

Zeeman, S. C., Kossmann, J., and Smith, A. M. (2010) Starch: its metabolism, evolution, and biotechnological modification in plants. *Annual Reviews in Plant Biology* **61**:209-234

Zhao, J., Chen, Y. H., and Kwan, H. S. (2000) Molecular cloning, characterization, and differential expression of a glucoamylase gene from the basidiomycetous fungus *Lentinula edodes*. *Applied Environmental Microbiology* **66(6)**:2531-2535

Chapter 5

Conclusions and Future Prospects

Conclusions

Maize is an important staple food, especially in sub-Saharan Africa, and understanding the biology of fungal maize pathogens such as *Cercospora zeina* could help identify strategies to reduce yield losses and improve food security in the region. Generating information on pathogen biology, infection strategies and distribution is therefore critical to control their spread and proliferation. Genome sequences not only provide details on the gene and regulatory components of organisms, but can also yield marker sequences important for investigating population structure and diversity. In this study, the genome of *C. zeina* was assembled using high-throughput sequencing reads, and the genome was also shown to be transcriptionally functional. Gene prediction and functional annotation of the genome identified pathogenesis strategies of *C. zeina*, and also provided sequence information useful for the phylogenetic grouping of unclassified *Cercospora* species.

The infection and nutrient acquisition strategies utilized by pathogenic Ascomycete fungi have been well studied as described in **Chapter 1**. The use of model pathogenesis systems have revealed the complexity of the interactions between fungi and hosts, although each pathogenesis system involves different components due to the adaptation of hosts and the nutrient acquisition requirements and strategies of invading pathogens. These were described in the context of Grey Leaf Spot disease of maize caused by *C. zeina* / *C. zeae-maydis*, being responsible for severe yield losses in affected regions. The differences between the biology of the related *C. zeina* and *C. zeae-maydis* species were also discussed, while the geographic distribution of the two species was highlighted. Finally the genus *Cercospora* was discussed, emphasizing the lack of a general barcoding gene useful for resolving unclassified species in the genus.

Assembly of the genome sequence in **Chapter 2** yielded a functional genome with more than half of the bases in scaffolds larger than 160 Kbp. The GC-content of the assembly (49.7%) was similar to other Dothideomycete species, such as *Parastagonospora nodorum* (50.52%) and *Cochliobolus heterostrophus* (49.81%) (Ohm *et al.*, 2012). However, the assembly has a smaller total size (37 Mbp) than expected when compared with other *Mycosphaerellaceae* species, e.g. *C. zeae-maydis* (46.61 Mbp) and *Zymoseptoria tritici* (39.69 Mbp) (JGI, 2011 ; Ohm *et al.*, 2012). The high degree of RNAseq reads mapping to the genome indicated that at minimum the coding regions of the assembly were correctly assembled to yield functional gene regions. The *C. zeina* cercosporin biosynthesis gene cluster was annotated to support this conclusion, and the synteny of genes in and surrounding the cluster were found to be identical to several other *Cercospora* species.

The genome sequence developed in this study has enabled several outputs from other postgraduate studies in the Molecular Plant-Microbe Interactions research group, FABI, Department of Plant and Soil Sciences, University of Pretoria, Pretoria, South Africa. In a PhD study (Swart, 2017), further analysis of the CTB cluster led to identification of the CTB7 mutation affecting cercosporin production. Differences in the size of the CTB7 gene between *C. zeina* and *C. zeae-maydis* were exploited to develop a PCR-diagnostic to distinguish the species (Swart *et al.*, 2017). In an MSc study (Muller, 2015), the non-coding regions of the assembly were mined for repetitive sequences, specifically microsatellites. These were used to successfully study the population structure of *C. zeina* in both a South African (Muller *et al.*, 2016) and African context (Nsibo *et al.*, 2019). Furthermore, the genome sequence data allowed the development of a PCR assay to identify the mating type gene present in each isolate from South African populations of the fungus, which provided evidence for sexual reproduction (Muller *et al.*, 2016 ; Kunene, 2016 ; Nsibo *et al.*, 2019). These studies indicated that, firstly, the non-coding regions of the genome were adequately assembled to a functional degree to enable the study of related organisms. Secondly, the genome assembly was useful in addressing questions relating to both coding and non-coding regions and was therefore functionally complete.

The annotation of the *C. zeina* genome assembly in **Chapter 3** yielded 10,193 gene models. The accuracy of the coordinates of ~30% of these was manually confirmed using the GenomeView manual annotation software. BUSCO completeness assessment of the gene models showed 95.4% completeness which is comparable to that of related species, while the number of genes was similar to genome annotations of other species in the Dothideomycete class. Functional annotation of the gene products yielded valuable information on possible host infection strategies, nutrition acquisition and host defense regulation. To correlate and compare the proteome of *C. zeina* with that of related *Cercospora* species, the same analyses were performed on all proteomes regardless of the availability of prior functional gene composition information. This approach enabled the comparison of class-specific gene products involved specifically in infection, nutrient acquisition and secondary metabolite production. The functional content of the proteomes were found to be very similar, with some protein classes possibly showing areas of interest for further studies on the differences in monocot-vs-dicot infection strategies.

The annotated genome of *C. zeina* was used to find alternative phylogenetic marker genes in **Chapter 4**. Due to the large number of unclassified *Cercospora* species and the failure of standard phylogenetic marker genes to resolve all these species (Groenewald *et al.*, 2005 ; Groenewald *et al.*, 2013 ; Bakhshi *et al.*, 2018), an alternative method of phylogenetic gene marker identification was applied. *Cercospora*-specific single-copy orthologs were identified from 18 Dothideomycete and 7 related genomes, and 8 phylogenetically informative genes selected for primer design. The primers from two genes were successful in amplifying the correct fragments and these sequences will be added to the standard phylogenetic analysis to enrich *Cercospora* species classification.

The primers used in the amplification of these gene fragments flanked both exonic and intronic regions. The inclusion of intronic sequence regions when attempting to phylogenetically classify closely related species is a tested approach providing adequate resolution (Creer, 2007). Due to the high selective pressure on the standard phylogenetic marker genes and exonic sequence regions the level of diversity could alter too slowly to be useful when attempting the classification of closely related species in the same genus. In contrast, the accelerated mutation rate of intronic sequences compared to exonic sequence regions provides the required additional informative sites for classification. This is especially important when working with a genus such as *Cercospora* where so many species have not been confirmed to have a sexual stage of procreation where recombination can introduce diversity in exonic regions.

This study provides a valuable genome resource for functional and diversity analysis of the *C. zeina*-maize pathosystem as evident from the number of studies utilizing the data to date. Additional future analyses would overcome challenges encountered, and would improve the quality of the resource and provide even greater value for the fungal research community at large.

Future prospects

The completeness and functionality of the genome assembly was sufficient for enabling a number of diverse studies into the population structure, physiology and pathogenicity of *C. zeina*. There were challenges encountered with the quality of the short reads generated with the Illumina paired-end and shorter mate-pair sequencing technology.

This project had three Illumina short read libraries, i.e. 400 bp paired-end, 3 Kbp mate-pair, and 8 Kbp mate-pair. Genome assembly projects in which data is available from different Illumina libraries generally start with the assembly of the paired-end reads into contigs, followed by scaffolding using information from the longer mate-pair libraries. However, the first attempt at assembly using the paired-end and 3 Kbp mate-pair libraries yielded an assembly to which very few RNAseq reads mapped, and it was therefore not functional. When comparing the GC-content profiles for these two libraries with the 8 Kbp mate-pair library it was evident that these two libraries contained anomalous base compositions, most likely as a result of poor library preparation. In addition, analysis of the scaffolding efficiency of the 3 Kbp mate-pair library indicated poor size-selection during library preparation since the read pairs for the library were only 200 – 300 bp apart instead of the required 3 Kbp. The assembly using only reads from the 8 Kbp mate-pair library was shorter than the original, non-functional assembly (37,099,429 bp vs 41,702,458 bp), though the high number of mapping RNAseq reads indicated that it was functional. The 8 Kbp mate-pair library assembly was also more fragmented than the original assembly (N50 of 160,632 bp vs 720,376 bp), though the improved functionality of the more fragmented assembly indicated that only relying on the assembly statistics for judging assembly quality is not always a reliable approach.

The anomalous base compositions of the paired-end and 3 Kbp mate-pair sequencing libraries possibly affected the assembly process negatively, since the poor sequencing quality of these libraries contributed less sequence data to the assembly process. The paired-end reads were discarded completely from the assembly process, while the assembly was probably more fragmented due to the poor insert size correlation of the 3 Kbp mate-pair library, thereby not allowing for a more effective linking of contigs to create larger scaffolds. In addition, the use of the larger 8 Kbp mate-pair library sequences for the main assembly most possibly resulted in neighbouring paralogs either being collapsed into one sequence, or skipped altogether. The large number of 8 Kbp mate-pair library sequences did allow for a significantly more complete assembly than would be the case when relying solely on the paired-end library sequences.

A comparison of the number of genes predicted for the different *Cercospora* species indicated that the *C. zeina* assembly codes for ~16% fewer genes, and an incomplete assembly is the most probable explanation. When trying to identify neighbouring paralogs, future studies could use larger insert paired-end libraries (800 bp – 1 Kbp) as main bulk of reads to create contigs instead of the 400 bp insert reads planned for this project. Alternatively, the use of long-read sequencing technologies, such as PacBio or MinION could be used to identify larger regions with high degrees of certainty. Both sequencing platforms have been used in fungal genome sequencing (Derbyshire *et al.*, 2017 ; Luo *et al.*, 2017), though the MinION data would mainly be used to scaffold contigs since the high error profile for this technology precludes its use for building contigs. These technologies could also greatly decrease the fragmentation of the assembly, since short read technologies cannot sequence through large repeat regions and assembly algorithms fail to correctly assemble these regions, thereby contributing to the fragmentation of genome assemblies. More complete genome assemblies could greatly aid studies on the repetitive sequence content of genomes, being especially valuable in pathogenic fungal genomes where effector genes are often localized in close proximity to repeat regions (de Jonge *et al.*, 2013). These genes tend to benefit from accelerated mutations creating altered protein products for successful infection. Finally, the genome comparison of closely related species using more complete genomes could illuminate evolutionary adaptations lost in fragmented assemblies, especially in the case of dispensable chromosomes mis-assembled with core chromosomes, or chromosomal region inversion or gene order shuffling. In short, the standardization of genome sequencing and assembly procedures could greatly enhance comparative genomic studies.

Future studies will investigate whether the differences in gene numbers and completeness of the respective proteomes are a result of the varying depth of sequencing and genome assembly accuracies. The use of different gene prediction algorithms also provide varying degrees of confidence in gene model predictions. The investigation into the annotation accuracy suggests that the *C. zeina* assembly is not complete, therefore the number of predicted genes will increase with a more complete assembly, while standardization of the gene prediction algorithm between the four species would most likely result in a much more similar number of genes predicted for the species. The gene predictors applied for gene identification in *C. zeina* used an intrinsic evidence approach instead of a more general Kingdom or Division specific prediction model. The use of Genemark-ES for fungal gene prediction might be a viable alternative (Campbell *et al.*, 2014), though it was deemed to be less informative due to compatibility problems with other prediction tools, and the availability of RNA sequencing data and related species sequences were of more relevance for the *Cercospora*-specific gene pool than the much more general and probably less specific model incorporated in Genemark-ES.

The functional annotation of the proteomes of the four species showed broadly similar numbers of functional representatives relative to the respective proteome sizes. There were exceptions, especially in the orthogroup analysis where 760 orthogroups were specific to the sugar beet infecting species, while only 263 orthogroups were specific to the maize infecting species. It is possible that these discrepancies result from the more complete sequencing of the sugar beet infecting species' genomes, as well as the intensive manual annotation process yielding a larger number of gene models. With more complete *C. zeina* and *C. zeae-maydis* genome sequences and annotations, these discrepancies could be reduced, and would then assist in the detection of genes involved in the monocot or dicot specific infection strategies. For the *C. zeina* annotation the repetitive regions were masked during gene prediction, therefore most likely underestimating the number of genes in the assembly. Including genes predicted in the repetitive regions would also improve the likelihood of more fungal effector genes shown to occur in these regions as discussed by De Jonge *et al.*, (2013). The use of newer gene prediction algorithms, such as BRAKER (Hoff *et al.*, 2016), might also result in a more accurate and complete gene model complement.

A problem endemic to Genbank and accelerating due to the ever increasing number of genome assemblies is the propagation of unsubstantiated functional gene descriptions (Koskinen *et al.*, 2015). Performing BLAST similarity searches to obtain general functional information on the proteome of a species is ideal in theory, but biased manual editing and less strict automatic annotation pipelines might favour functional descriptions not based on empirical functional evidence over the much less preferred hypothetical functional description. Accurate functional classification is only valid once proteins have been studied in isolation and their *in vivo* and *in vitro* functions verified using biochemical and molecular biology tools. Due to the large increase in genome sequencing projects and the financial and time investments required for the confirmation of protein function, the dependence on database classifications will only increase, along with the errors introduced during these procedures.

The fungal genome sequence resources provided by the Joint Genome Institute are valuable, especially for researchers without the financial resources to sequence their own genomes of interest. However, the JGI automated assembly and annotation pipelines do create challenges when measured against manually assembled, annotated and curated genomes. Use of these automated genome annotations decreased the efficiency of the ortholog-finding procedure in this

Table 5.1 **Number of genes in Orthogroups (OGs) for analyzed Ascomycete species (adapted from Sections 4.3.1, 4.3.3 and Table 4.3.1).** *Rows highlighted in grey indicate genome assemblies not optimally annotated with the JGI automatic annotation pipeline.*

Organism	Database	Total number of genes	Genes in OGs	Unassigned genes	Single copy genes	Organism-specific OGs	Genes in organism-specific OGs
Aspergillus nidulans	AGD	10,680	9,834	846	6,760	3	2
Aspergillus niger	AGD	14,097	10,665	3,432	6,377	7	20
Baudoinia compniacensis	JGI	10,513	8,504	2,009	6,937	5	6
Botrytis cinerea	NCBI	16,447	11,093	5,354	7,477	9	5
Cercospora berteroae	NCBI	11,972	11,767	205	8,392	2	2
Cercospora beticola	NCBI	12,495	12,132	363	8,268	0	0
Cercospora zeae-maydis	JGI	12,020	11,006	1,014	8,384	1	2
Cercospora zeina	NCBI	10,193	9,872	321	7,807	1	3
Cladosporium fulvum	NCBI	14,127	12,022	2,105	7,466	9	15
Cochliobolus heterostrophus	JGI	215,529	206,036	9,493	327	61	349
Cochliobolus sativus	JGI	179,628	174,736	4,892	313	44	88
Leptosphaeria maculans	NCBI	12,469	9,292	3,177	7,739	0	0
Zymoseptoria tritici	JGI	10,952	9,430	1,522	6,929	18	31
Neurospora crassa	JGI	10,785	8,884	1,901	5,858	24	23
Pseudocercospora fijiensis	JGI	115,028	104,671	10,357	526	173	355
Pyrenophora teres	NCBI	11,799	11,268	531	8,809	9	15
Pyrenophora tritici-repentis	NCBI	12,169	11,088	1,081	8,946	5	44
Rhytidhysteron rufulum	NCBI	12,117	10,884	1,233	7,257	15	13
Saccharomyces cerevisiae	NCBI	5,974	4,494	1,480	3,089	23	15
Sclerotinia sclerotiorum	NCBI	14,503	10,308	4,195	7,705	3	7
Septoria musiva	JGI	10,233	9,552	681	7,870	2	23
Septoria populicola	NCBI	9,739	9,210	529	7,696	4	21
Setosphaeria turcica	JGI	159,577	154,337	5,240	320	62	112
Stagonospora nodorum	NCBI	12,380	10,537	1,843	8,188	2	8
Verticillium dahliae	NCBI	10,535	9,283	1,252	6,749	8	16

study. Analysis of the genomes annotated using the JGI automated pipeline (Table 5.1) indicated that some of the genomes suffer from an over-estimation of the number of gene models for these organisms, especially *C. heterostrophus*, *C. sativus*, *P. fijiensis* and *S. turcica* (more than 115,028 gene counts). Also indicative of the poor annotation quality is the very low number of single copy genes in these annotations, on average being only between 4% and 7% of the value of the other annotations. This could be indicative of poor gene prediction, and also multiple predictions per gene model site which were not collated to provide only the single most likely gene model for each site. This is also indicated by the larger-than-average number of organism-specific orthogroups found for these annotations.

In addition, other inconsistencies were seen for some JGI-annotated genome datasets from Table 5.1 even though the gene prediction counts were within the range predicted from manually annotated genomes from other phytopathogenic fungi. For example, the *C. zeae-maydis* protein functional prediction results in Chapter 3 indicated multiple instances where the number of predicted functional elements was not in the expected range when compared to the other genomes. Another example was the incomplete annotation of the *C. zeae-maydis* CTB cluster genes (data not shown) where the CTB8 gene was not present in the annotation file and had to be manually annotated for the purposes of this study, even though the gene was clearly present with a consistent sequence and position compared to related species. When these automated genome annotations are used it is prudent to re-annotate using similar tools as the annotations under study, thereby removing false-negative or false-positive results from the analysis. This might be especially important where accurate predictions are required to make conclusions, for example regarding host infection strategies or host immune modulation based on a few number of genes being present or absent.

References

Bakhshi, M., Arzanlou, M., Babai-Ahari, A., Groenewald, J. Z., and Crous, P. W. (2018) Novel primers improve species delimitation in *Cercospora*. *IMA Fungus* **9**:299-332

Campbell, M. S., Holt, C., Moore, B., and Yandell, M. (2014) Genome annotation and curation using MAKER and MAKER-P. *Current Protocols in Bioinformatics* **48**:4.11.1–4.11.39

Creer, S. (2007) Choosing and using introns in molecular phylogenetics. *Evolutionary Bioinformatics Online* **3**:99–108

De Jonge, R., Bolton, M. D., Kombrink, A., van den Berg, G. C., Yadeta, K. A., and Thomma, B. P. (2013) Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome Research* **23(8)**:1271-1282

Derbyshire, M., Denton-Giles, M., Hegedus, D., Seifbarghy, S., Rollins, J., van Kan, J., Seidl, M. F., Faino, L., Mbengue, M., Navaud, O., Raffaele, S., Hammond-Kosack, K., Heard, S., and Oliver, R. (2017) The complete genome sequence of the phytopathogenic fungus *Sclerotinia sclerotiorum* reveals insights into the genome architecture of broad host range pathogens. *Genome Biology and Evolution* **9(3)**:593–618

Groenewald, J. Z., Nakashima, C., Nishikawa, J., Shin, H. D., Park, J. H., Jama, A. N., Groenewald, M., Braun, U., and Crous, P. W. (2013) Species concepts in *Cercospora*: spotting the weeds among the roses. *Studies in Mycology* **75(1)**:115-170

Groenewald, M., Groenewald, J. Z., and Crous, P. W. (2005) Distinct species exist within the *Cercospora apii* morphotype. *Phytopathology* **95(8)**:951-959

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M., and Stanke, M. (2016) BRAKER1: Unsupervised RNA-Seq-based genome annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32(5)**:767-769

JGI. 2011. https://genome.jgi.doe.gov/Cerzm1/Cerzm1.home.html.

Koskinen, P., Toronen, P., Nokso-Koivisto, J., and Holm, L. (2015) PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics* **31(10)**:1544-1552

Luo, R., Zimin, A., Workman, R., Fan, Y., Pertea, G., Grossman, N., Wear, M. P., Jia, B., Miller, H., Casadevall, A., Timp, W., Zhang, S. X., and Salzberg, S. L. (2017) First draft genome sequence of the pathogenic fungus *Lomentospora prolificans* (formerly *Scedosporium prolificans*). *G3 (Bethesda)* **7(11)**:3831-3836

Muller, M. F. (2015) Molecular diversity of the maize pathogen *Cercospora zeina* in South Africa. MSc Thesis, University of Pretoria, Pretoria, South Africa

Muller, M. F., Barnes, I., Kunene, N. T., Crampton, B. G., Bluhm, B. H., Phillips, S. M., Olivier, N. A., and Berger, D. K. (2016) *Cercospora zeina* from maize in South Africa exhibits high genetic diversity and lack of regional population differentiation. *Phytopathology* **106(10)**:1194-1205

Kunene, N. T. (2016) Mating-type gene analysis of *Cercospora zeina* in commercial and small-holder maize farms in South Africa. MSc Thesis, University of Pretoria, Pretoria, South Africa

Nsibo, D. L., Barnes, I., Kunene, N. T., and Berger, D. K. (2019) Influence of farming practices on the population genetics of the maize pathogen *Cercospora zeina* in South Africa. *Fungal Genetics and Biology* **125**:36-44

Ohm, R. A., Feau, N., Henrissat, B., Schoch, C. L., Horwitz, B. A., Barry, K. W., Condon, B. J., Copeland, A. C., Dhillon, B., Glaser, F., Hesse, C. N., Kosti, I., LaButti, K., Lindquist, E. A., Lucas, S., Salamov, A. A., Bradshaw, R. E., Ciuffetti, L., Hamelin, R. C., Kema, G. H., Lawrence, C., Scott, J. A., Spatafora,

J. W., Turgeon, B. G., de Wit, P. J., Zhong, S., Goodwin, S. B., and Grigoriev, I. V. (2012) Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. *PLoS Pathogens* **8(12)**:e1003037

Swart, V. (2017) Functional genomics of the cercosporin biosynthetic gene cluster in the maize pathogen *Cercospora zeina*. PhD Thesis, University of Pretoria, Pretoria, South Africa

Swart, V., Crampton, B. G., Ridenour, J. B., Bluhm, B. H., Olivier, N. A., Meyer, J. J. M., and Berger, D. K. (2017) Complementation of CTB7 in the maize pathogen *Cercospora zeina* overcomes the lack of *in vitro* cercosporin production. *Molecular Plant-Microbe Interactions* **30(9)**:710-724

Summary

The fungal pathogen *Cercospora zeina* causes Grey leaf spot (GLS), a devastating yield-limiting foliar disease of maize. GLS negatively impacts food security, especially in sub-Saharan Africa where maize is a staple food source. In this study the genomic DNA of *C. zeina* was sequenced using next-generation sequencing, and the genome assembled to a 95.4% completeness based on the presence of core genes. The functionality of the genome was confirmed by transcriptome sequencing data mapping to the genome. Phylogenetics analysis confirmed the genome to cluster with other *C. zeina* isolates. The functional elements and gene regions were predicted using the MAKER genome annotation pipeline. The predicted proteins were compared with the closely related species *Cercospora zeae-maydis*, *Cercospora beticola* and *Cercospora berteroae*. Functional annotation of proteins of specific classes were performed to identify differences in secreted proteins, carbohydrate-active enzymes, lipases, proteases and components of secondary metabolite biosynthesis clusters. The synteny of the genes in the cercosporin toxin biosynthesis cluster was also confirmed in all four species. To enhance the accuracy of the phylogenetic classification of *Cercospora* species the orthologous relationship between proteins of a number of Dothideomycete species were predicted. The single-copy orthologs specific to the *Cercospora* genus were analyzed for phylogenetic information content, and eight genes selected for primer design in regions of protein identity. Primers for four genes were synthesized and tested for specificity during amplification of *C. zeina* and *C. zeae-maydis* genomic DNA. Degenerate primer pairs for two genes were selected for further analysis, due to sequencing confirming the correct identity of the amplification products.

Appendix

Table A1 MAKER-P behavior settings provided by the maker_opts.ctl file. *These settings were provided for the first annotation prediction.*

Descriptor name	Value	Description
Genome file and description		
genome	genome.fasta	Genome sequence
organism_type	eukaryotic	Eukaryotic or prokaryotic
Re-annotation Using MAKER Derived GFF3		
maker_gff		MAKER derived GFF3 file
est_pass	0	Use ESTs in maker_gff: 1
altest_pass	0	Use alternate organism ESTs in maker_gff: 1
protein_pass	0	Use protein alignments in maker_gff: 1
rm_pass	0	Use repeats in maker_gff: 1
model_pass	0	Use gene models in maker_gff: 1
pred_pass	0	Use ab-initio predictions in maker_gff: 1
other_pass	0	Pass through anything else in maker_gff: 1
EST Evidence		
est		Set of ESTs or assembled RNAseq
altest		EST/cDNA sequence file from an alternate organism
est_gff		Aligned ESTs or RNAseq from an external GFF3 file
altest_gff		Aligned ESTs from a closely relate species in GFF3 format
Protein homology evidence		
Protein		Protein sequence file
protein_gff		Aligned protein homology evidence from an external GFF3 file
Repeat Masking		
model_org	all	Model organism for RepBase masking
rmlib		Organism specific repeat library
repeat_protein	TE_proteins.fasta	Transposable element proteins
rm_gff		Pre-identified repeat elements from an external GFF3 file
prok_rm	0	Forces MAKER to repeatmask prokaryotes
softmask	1	Use soft-masking rather than hard-masking
Gene Prediction		
snaphmm	czeina.hmm	SNAP HMM file
gmhmm		GeneMark HMM file
augustus_species		Augustus gene prediction species model
fgenesh_par_file		FGENESH parameter file
pred_gff		Ab-initio predictions from an external GFF3
model_gff		Annotated gene models from an external GFF3
est2genome	0	Infer gene predictions directly from ESTs, 1
protein2genome	0	Infer predictions from protein homology, 1
unmask	0	Run ab-initio prediction programs on unmasked sequence, 1
Other Annotation Feature Types		
other_gff		Extra features to pass-through to final MAKER generated GFF3
External Application Behavior Options		

alt_peptide	C	Amino acid used to replace non-standard amino acids in BLAST databases
cpus	1	Max number of cpus to use
MAKER Behavior Options		
max_dna_len	100,000	Length for dividing up contigs into chunks
min_contig	1	Skip genome contigs below this length
pred_flank	200	Flank for extending evidence clusters sent to gene predictors
pred_stats	0	Report AED and QI statistics for all predictions as well as models
AED_threshold	1	Maximum Annotation Edit Distance allowed
min_protein	0	Require at least this many amino acids in predicted proteins
alt_splice	0	Take extra steps to try and find alternative splicing, 1
always_complete	0	Extra steps to force start and stop codons, 1
map_forward	0	Map names and attributes forward from old GFF3 genes, 1
keep_preds	0	Concordance threshold to add unsupported gene prediction
split_hit	10,000	Length for the splitting of hits (expected max intron size for evidence alignments)
single_exon	0	Consider single exon EST evidence when generating annotations, 1
single_length	250	Min length required for single exon ESTs if 'single_exon is enabled'
correct_est_fusion	0	Limits use of ESTs in annotation to avoid fusion genes
tries	2	Number of times to try a contig if there is a failure
clean_try	0	Remove all data from previous run before retrying, 1
clean_up	0	Removes theVoid directory with individual analysis files, 1
TMP		Specify a directory for temporary files

Table A2 BLAST and Exonerate statistics thresholds provided to MAKER-P by the maker_bopts.ctl file. *These settings were unchanged for all annotation predictions.*

Factor	Value	Description
blast_type	NCBI+	
pcov_blastn	0.8	Blastn percent coverage threshold: EST-Genome alignments
pid_blastn	0.85	Blastn percent identity threshold: EST-Genome alignments
eval_blastn	1e-10	Blastn e-value cutoff
bit_blastn	40	Blastn bit cutoff
depth_blastn	0	Blastn depth cutoff (0 to disable cutoff)
pcov_blastx	0.5	Blastx percent coverage threshold: Protein-Genome alignments
pid_blastx	0.4	Blastx percent identity threshold: Protein-Genome alignments
eval_blastx	1e-06	Blastx e-value cutoff
bit_blastx	30	Blastx bit cutoff
depth_blastx	0	Blastx depth cutoff (0 to disable cutoff)
pcov_tblastx	0.8	tBlastx percent coverage threshold alt-EST-Genome alignments
pid_tblastx	0.85	tBlastx percent identity threshold alt-EST-Genome alignments
eval_tblastx	1e-10	tBlastx e-value cutoff
bit_tblastx	40	tBlastx bit cutoff
depth_tblastx	0	tBlastx depth cutoff (0 to disable cutoff)
pcov_rm_blastx	0.5	Blastx percent identity threshold for transposable element masking
pid_rm_blastx	0.4	Blastx percent identity threshold for transposable element masking
eval_rm_blastx	1e-06	Blastx e-value cutoff for transposable element masking
bit_rm_blastx	30	Blastx bit cutoff for transposable element masking
ep_score_limit	20	Exonerate protein percent of maximal score threshold
en_score_limit	20	Exonerate nucleotide percent of maximal score threshold

Table A3 Applications and algorithms required for MAKER-P gene prediction provided to MAKER-P by the maker_exe.ctl file. *These settings were unchanged for all annotation predictions.*

Mapping and similarity applications	Algorithms for gene prediction
NCBI+ makeblastdb	SNAP
NCBI+ blastn	GeneMark eukaryotic
NCBI+ blastx	GeneMark prokaryotic
NCBI+ tblastx	Augustus
NCBI+ formatdb	FGENESH (External license required)
NCBI+ blastall	Probuild (required for GeneMark)
WUBLAST xdformat	
WUBLAST blasta	
RepeatMasker	
Exonerate	

Table A4 CAZyme classes predicted for the Cercospora species, with the number of proteins in each class indicated.

CAZY	*C. berteroae*	*C. beticola*	*C. zeina*	*C. zeae-maydis*
Carbohydrate-Binding Module CBM4	1	1	0	0
Carbohydrate-Binding Module CBM13	2	2	1	1
Carbohydrate-Binding Module CBM14	1	1	1	1
Carbohydrate-Binding Module CBM18	4	5	4	7
Carbohydrate-Binding Module CBM19	2	2	0	1
Carbohydrate-Binding Module CBM20	5	4	4	4
Carbohydrate-Binding Module CBM21	1	2	1	1
Carbohydrate-Binding Module CBM32	2	2	2	1
Carbohydrate-Binding Module CBM35	1	1	1	1
Carbohydrate-Binding Module CBM37	0	0	0	1
Carbohydrate-Binding Module CBM42	1	1	1	1
Carbohydrate-Binding Module CBM43	2	2	2	2
Carbohydrate-Binding Module CBM48	1	1	1	1
Carbohydrate-Binding Module CBM50	14	13	7	8
Carbohydrate-Binding Module CBM52	0	1	0	0
Carbohydrate esterase CE1	24	23	19	20
Carbohydrate esterase CE3	6	7	3	5
Carbohydrate esterase CE4	5	5	5	5
Carbohydrate esterase CE5	9	10	9	9
Carbohydrate esterase CE7	1	2	0	0
Carbohydrate esterase CE8	1	1	1	1
Carbohydrate esterase CE9	2	2	2	3
Carbohydrate esterase CE10	56	63	51	48
Carbohydrate esterase CE12	3	3	2	2
Carbohydrate esterase CE14	1	1	2	1
Carbohydrate esterase CE16	2	2	2	2
Glycoside hydrolase GH1	2	3	3	3
Glycoside hydrolase GH2	6	6	5	5
Glycoside hydrolase GH3	18	16	14	16
Glycoside hydrolase GH5	14	14	12	11
Glycoside hydrolase GH7	1	1	1	1
Glycoside hydrolase GH9	1	1	1	1
Glycoside hydrolase GH10	4	4	3	2
Glycoside hydrolase GH11	4	4	3	2
Glycoside hydrolase GH12	3	3	2	2
Glycoside hydrolase GH13	18	20	16	8
Glycoside hydrolase GH15	2	2	1	1
Glycoside hydrolase GH16	20	18	16	18

Glycoside hydrolase GH17	7	7	6	6
CAZY	***C. berteroae***	***C. beticola***	***C. zeina***	***C. zeae-maydis***
Glycoside hydrolase GH18	8	8	8	10
Glycoside hydrolase GH20	1	1	1	1
Glycoside hydrolase GH23	2	2	1	1
Glycoside hydrolase GH24	1	1	1	1
Glycoside hydrolase GH27	3	3	2	2
Glycoside hydrolase GH28	5	5	4	5
Glycoside hydrolase GH29	2	3	2	2
Glycoside hydrolase GH30	2	2	1	1
Glycoside hydrolase GH31	8	8	8	8
Glycoside hydrolase GH32	3	3	2	3
Glycoside hydrolase GH33	1	1	0	0
Glycoside hydrolase GH35	1	2	1	1
Glycoside hydrolase GH36	1	1	1	1
Glycoside hydrolase GH37	2	2	2	2
Glycoside hydrolase GH38	1	1	1	1
Glycoside hydrolase GH39	0	1	1	1
Glycoside hydrolase GH42	1	1	0	0
Glycoside hydrolase GH43	16	17	12	34
Glycoside hydrolase GH47	9	9	7	8
Glycoside hydrolase GH51	2	2	2	2
Glycoside hydrolase GH53	1	1	1	1
Glycoside hydrolase GH54	1	1	1	1
Glycoside hydrolase GH55	5	5	4	4
Glycoside hydrolase GH62	1	1	1	1
Glycoside hydrolase GH63	2	2	2	2
Glycoside hydrolase GH64	4	4	4	4
Glycoside hydrolase GH65	1	1	1	1
Glycoside hydrolase GH67	1	1	1	1
Glycoside hydrolase GH71	2	2	2	1
Glycoside hydrolase GH72	8	8	7	7
Glycoside hydrolase GH74	1	1	1	0
Glycoside hydrolase GH76	10	10	10	9
Glycoside hydrolase GH78	2	2	1	2
Glycoside hydrolase GH79	3	3	2	3
Glycoside hydrolase GH81	1	1	1	1
Glycoside hydrolase GH85	0	1	1	1
Glycoside hydrolase GH88	1	2	0	0
Glycoside hydrolase GH92	8	7	6	6
Glycoside hydrolase GH95	0	1	0	0

Glycoside hydrolase GH105	6	6	5	5
Glycoside hydrolase GH106	1	2	0	0
CAZY	***C. berteroae***	***C. beticola***	***C. zeina***	***C. zeae-maydis***
Glycoside hydrolase GH109	15	13	12	14
Glycoside hydrolase GH114	2	2	1	2
Glycoside hydrolase GH115	1	1	1	1
Glycoside hydrolase GH125	3	3	3	3
Glycosyl transferase GT1	7	6	3	3
Glycosyl transferase GT2	20	19	16	17
Glycosyl transferase GT3	1	1	1	1
Glycosyl transferase GT4	5	5	4	5
Glycosyl transferase GT5	0	0	0	1
Glycosyl transferase GT8	11	14	8	8
Glycosyl transferase GT15	4	4	3	3
Glycosyl transferase GT17	2	2	3	2
Glycosyl transferase GT20	3	3	3	3
Glycosyl transferase GT21	2	2	1	1
Glycosyl transferase GT22	4	4	4	4
Glycosyl transferase GT24	1	1	1	1
Glycosyl transferase GT25	4	6	7	5
Glycosyl transferase GT28	1	1	1	1
Glycosyl transferase GT31	6	5	4	5
Glycosyl transferase GT32	5	6	4	4
Glycosyl transferase GT33	1	1	1	1
Glycosyl transferase GT34	9	9	7	8
Glycosyl transferase GT35	1	1	1	1
Glycosyl transferase GT39	3	3	3	3
Glycosyl transferase GT41	1	1	1	1
Glycosyl transferase GT48	1	1	1	1
Glycosyl transferase GT50	1	1	1	1
Glycosyl transferase GT57	3	3	3	3
Glycosyl transferase GT58	1	1	1	1
Glycosyl transferase GT59	1	1	1	1
Glycosyl transferase GT62	3	3	3	3
Glycosyl transferase GT64	1	1	1	1
Glycosyl transferase GT66	1	1	1	1
Glycosyl transferase GT69	2	2	2	2
Glycosyl transferase GT71	4	4	2	2
Glycosyl transferase GT76	1	1	1	1
Glycosyl transferase GT90	14	13	9	13
Glycosyl transferase GT91	2	2	0	0

Polysaccharide lyase PL1	3	3	3	3
Polysaccharide lyase PL3	3	2	2	1
Polysaccharide lyase PL4	1	1	0	1
CAZY	***C. berteroae***	***C. beticola***	***C. zeina***	***C. zeae-maydis***
Polysaccharide lyase PL6	0	0	1	0
Polysaccharide lyase PL22	2	1	0	0

Table A5 Number and classes of secreted proteases predicted for *Cercospora* species.

Protease class	*C. berteroae*	*C. beticola*	*C. zeina*	*C. zeae-maydis*
Aspartic proteases				
A01	14	14	6	13
A11	0	0	5	0
A28	0	0	1	0
Cysteine proteases				
C01	1	0	0	1
C02	1	0	3	0
C03	0	0	1	0
C13	0	0	0	1
C14	0	0	5	0
C19	0	1	5	0
C26	0	0	1	0
C48	0	0	1	0
C56	0	0	6	0
C60	0	0	0	1
C69	1	1	0	0
C82	0	0	1	0
Metallo proteases				
M01	1	0	0	0
M02	0	0	1	0
M03	1	0	2	1
M04	2	1	0	2
M10	7	2	7	3
M12	9	3	3	6
M13	0	0	1	0
M14	5	15	4	19
M16	0	4	2	0
M19	0	0	2	0
M20	1	0	2	7
M24	0	2	2	2
M28	10	1	1	31
M35	3	2	0	2
M36	0	1	0	0
M38	0	0	2	0
M41	0	0	4	0
M43	5	0	0	2
M48	0	0	1	0
M50	0	0	1	0
M54	15	1	1	2
M56	0	0	1	0
M57	1	1	0	0
M67	0	0	4	0
M84	1	1	0	0
Serine proteases				
S01	2	4	5	4
S08	16	2	8	23
S09	3	13	19	10
S10	29	13	0	8
S12	5	1	4	5
S13	0	0	1	0
S15	6	5	1	4

S16	0	0	1	0
S26	0	0	2	0
S28	2	4	2	2
S33	2	2	5	2
S51	1	0	0	0
S53	5	0	0	6
S54	1	1	2	1
Threonine proteases				
T01	1	0	2	2
T02	3	0	0	1
T03	1	1	1	1
T05	0	0	1	0

Table A6 KOG categories and functional classifications for the four main functional groups.

Category	KOG functional classifications
CELLULAR PROCESSES AND SIGNALING	
D	Cell cycle control, cell division, chromosome partitioning
M	Cell wall/membrane/envelope biogenesis
N	Cell motility
O	Post-translational modification, protein turnover, and chaperones
T	Signal transduction mechanisms
U	Intracellular trafficking, secretion, and vesicular transport
V	Defense mechanisms
W	Extracellular structures
Y	Nuclear structure
Z	Cytoskeleton
INFORMATION STORAGE AND PROCESSING	
A	RNA processing and modification
B	Chromatin structure and dynamics
J	Translation, ribosomal structure and biogenesis
K	Transcription
L	Replication, recombination and repair
METABOLISM	
C	Energy production and conversion
E	Amino acid transport and metabolism
F	Nucleotide transport and metabolism
G	Carbohydrate transport and metabolism
H	Coenzyme transport and metabolism
I	Lipid transport and metabolism
P	Inorganic ion transport and metabolism
Q	Secondary metabolites biosynthesis, transport, and catabolism
POORLY CHARACTERIZED	
S	Function unknown
R	General Functional Prediction only

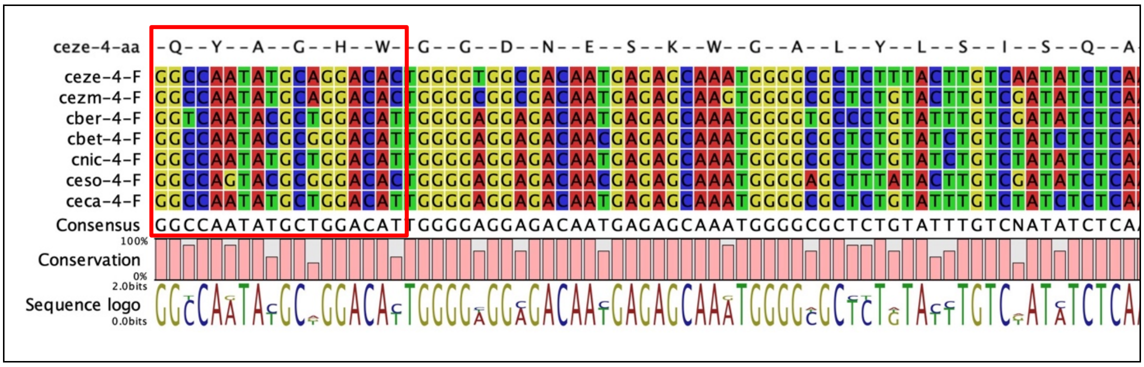

ceze-4-aa -Q--Y--A--G--H--W--G--G--D--N--E--S--K--W--G--A--L--Y--L--S--I--S--Q--A
ceze-4-F GGCCAATATGCAGGACACTGGGGTGGCGACAATGAGAGCAAATGGGGCGCTCTTTACTTGTCAATATCTCA
cezm-4-F GGCCAATATGCAGGACACTGGGGCGGCGACAATGAGAGCAAGTGGGGCGCTCTGTACTTGTCGATATCTCA
cber-4-F GGTCAATACGCTGGACATTGGGGAGGAGACAATGAGAGCAAATGGGGTGCCCTGTATTTGTCGATATCTCA
cbet-4-F GGCCAATACGCGGGACATTGGGGAGGAGACAACGAGAGCAAATGGGGCGCTCTGTATCTGTCTATCTCTCA
cnic-4-F GGCCAATATGCTGGACATTGGGGAGGAGACAATGAGAGCAAATGGGGCGCTCTGTATCTGTCTATATCTCA
ceso-4-F GGCCAGTACGCGGGACACTGGGGAGGAGACAACGAGAGCAAATGGGGAGCTTTATACTTGTCGATATCTCA
ceca-4-F GGCCAATATGCTGGACATTGGGGAGGAGACAATGAGAGCAAATGGGGCGCTCTGTATTTGTCTATCTCTCA
Consensus GGCCAATATGCTGGACATTGGGGAGGAGACAATGAGAGCAAATGGGGCGCTCTGTATTTGTCNATATCTCA
100%
Conservation
0%
2.0bits
Sequence logo
0.0bits

Figure A1 **Multiple sequence alignment for design of *ceze-4* forward degenerate primer.** *The red blocks indicate the regions selected for forward degenerate primer design. Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**),* C. zeae-maydis *(**cezm**)* C. nicotianae *(**cnic**),* C. sojina *(**ceso**) and* C. canescens *(**ceca**). The* C. zeina *amino acid sequence for the selected region is indicated (**ceze-4-aa**).*

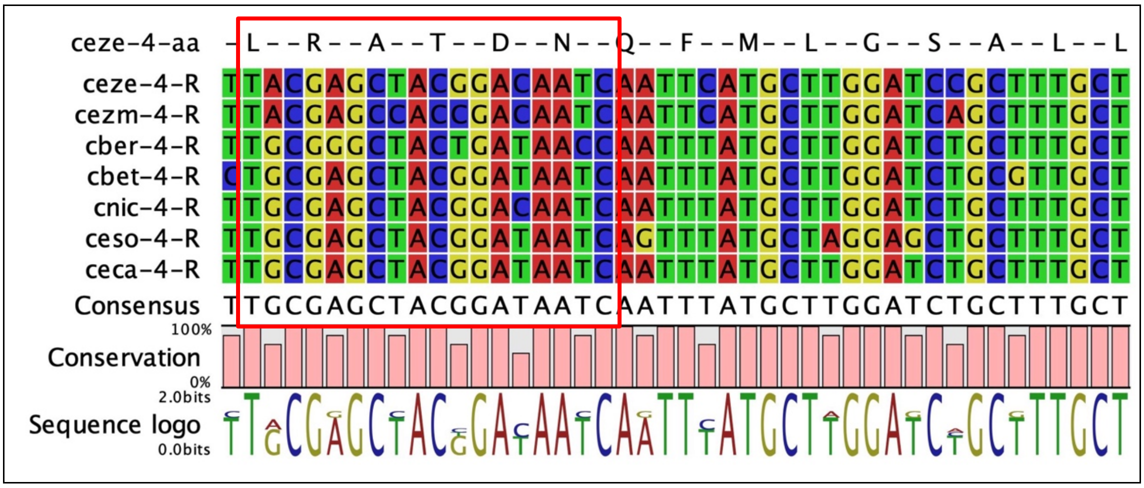

ceze-4-aa -L--R--A--T--D--N--Q--F--M--L--G--S--A--L--L
ceze-4-R TTACGAGCTACGGACAATCAATTCATGCTTGGATCCGCTTTGCT
cezm-4-R TTACGAGCCACCGACAATCAATTCATGCTTGGATCAGCTTTGCT
cber-4-R TTGCGGGCTACTGATAACCAATTTATGCTTGGATCTGCTTTGCT
cbet-4-R CTGCGAGCTACGGATAATCAATTTATGCTTGGATCTGCGTTGCT
cnic-4-R TTGCGAGCTACGGACAATCAATTTATGCTTGGATCTGCTTTGCT
ceso-4-R TTGCGAGCTACGGATAATCAGTTTATGCTAGGAGCTGCTTTGCT
ceca-4-R TTGCGAGCTACGGATAATCAATTTATGCTTGGATCTGCTTTGCT
Consensus TTGCGAGCTACGGATAATCAATTTATGCTTGGATCTGCTTTGCT
100%
Conservation
0%
2.0bits
Sequence logo
0.0bits

Figure A2 **Multiple sequence alignment for design of *ceze-4* reverse degenerate primer.** *The red blocks indicate the regions selected for reverse degenerate primer design. Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**),* C. zeae-maydis *(**cezm**)* C. nicotianae *(**cnic**),* C. sojina *(**ceso**) and* C. canescens *(**ceca**). The* C. zeina *amino acid sequence for the selected region is indicated (**ceze-4-aa**).*

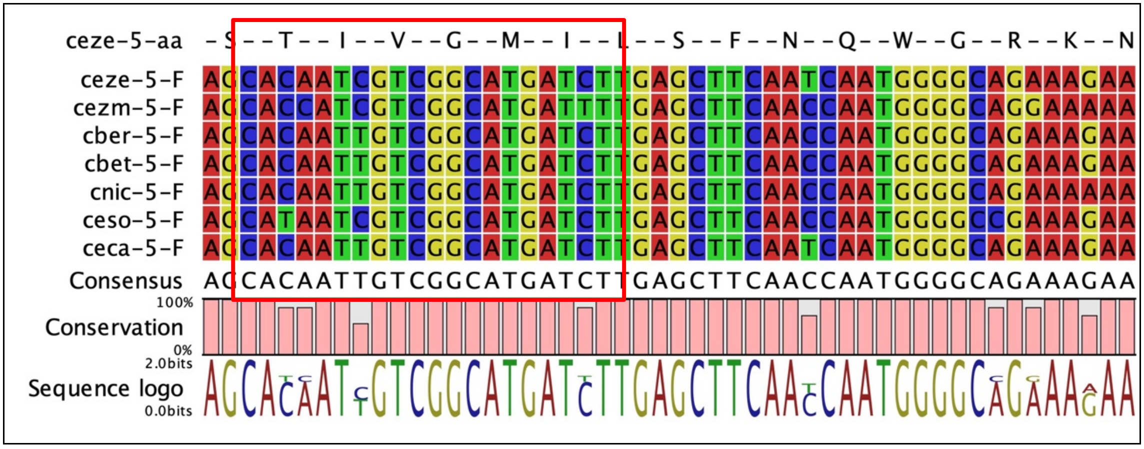

ceze-5-aa -S--T--I--V--G--M--I--L--S--F--N--Q--W--G--R--K--N
ceze-5-F AGCACAATCGTCGGCATGATCTTGAGCTTCAATCAATGGGGCAGAAAGAA
cezm-5-F AGCACCATCGTCGGCATGATTTTGAGCTTCAACCAATGGGGCAGGAAAAA
cber-5-F AGCACAATTGTCGGCATGATCTTGAGCTTCAACCAATGGGGCAGAAAGAA
cbet-5-F AGCACAATTGTCGGCATGATCTTGAGCTTCAACCAATGGGGCAGAAAGAA
cnic-5-F AGCACAATTGTCGGCATGATCTTGAGCTTCAACCAATGGGGCAGAAAAAA
ceso-5-F AGCATAATCGTCGGCATGATCTTGAGCTTCAACCAATGGGGCCGAAAGAA
ceca-5-F AGCACAATTGTCGGCATGATCTTGAGCTTCAATCAATGGGGCAGAAAGAA
Consensus AGCACAATTGTCGGCATGATCTTGAGCTTCAACCAATGGGGCAGAAAGAA
100%
Conservation
0%
2.0bits
Sequence logo
0.0bits

Figure A3 Multiple sequence alignment for design of *ceze-5* forward degenerate primer. *The red blocks indicate the regions selected for forward degenerate primer design. Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**),* C. zeae-maydis *(**cezm**)* C. nicotianae *(**cnic**),* C. sojina *(**ceso**) and* C. canescens *(**ceca**). The* C. zeina *amino acid sequence for the selected region is indicated (**ceze-5-aa**).*

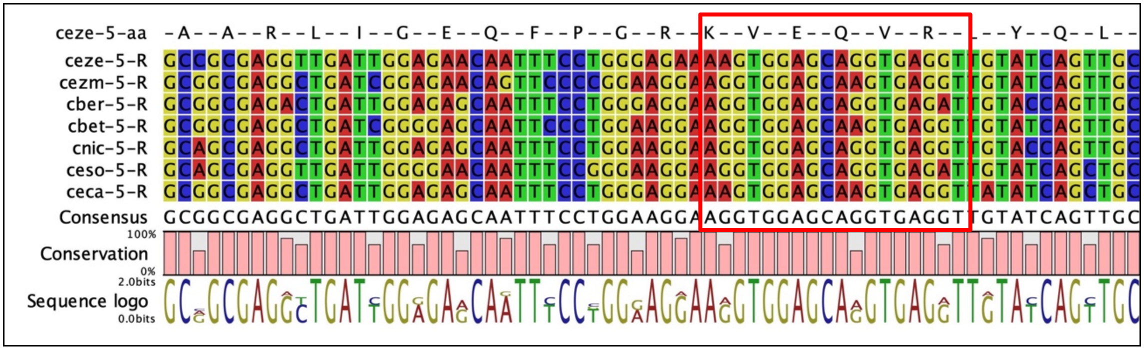
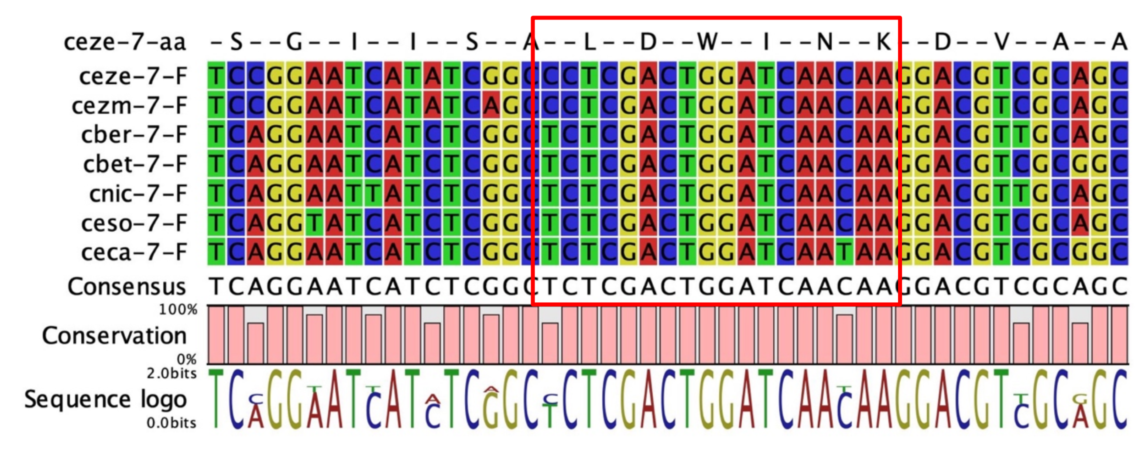
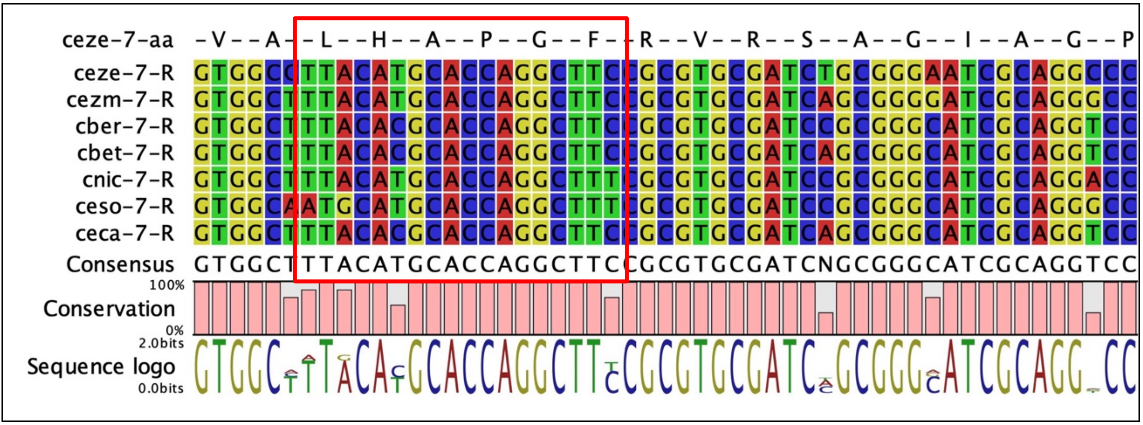

ceze-5-aa -A--A--R--L--I--G--E--Q--F--P--G--R--K--V--E--Q--V--R--L--Y--Q--L--
ceze-5-R GCCGCGAGGTTGATTGGAGAACAATTTCCTGGGAGAAAAGTGGAGCAGGTGAGGTTGTATCAGTTGC
cezm-5-R GCGGCGAGGCTGATCGGAGAACAGTTCCCCGGAAGGAAGGTGGAGCAAGTGAGGTTGTATCAGTTGC
cber-5-R GCGGCGAGACTGATTGGAGAGCAATTTCCTGGGAGGAAGGTGGAGCAGGTGAGATTGTACCAGTTGC
cbet-5-R GCGGCGAGGCTGATCGGGGAGCAATTCCCTGGAAGGAAGGTGGAGCAAGTGAGGTTGTATCAGTTGC
cnic-5-R GCAGCGAGGCTGATTGGAGAGCAATTTCCTGGAAGGAAGGTGGAGCAGGTGAGGTTGTACCAGTTGC
ceso-5-R GCAGCGAGGTTGATTGGGGAACAATTTCCGGGAAGGAAGGTGGAGCAGGTGAGATTGTATCAGCTGC
ceca-5-R GCGGCGAGGCTGATTGGAGAGCAATTTCCTGGGAGGAAAGTGGAGCAAGTGAGGTTATATCAGCTGC
Consensus GCGGCGAGGCTGATTGGAGAGCAATTTCCTGGAAGGAAGGTGGAGCAGGTGAGGTTGTATCAGTTGC
100%
Conservation
0%
2.0bits
Sequence logo
0.0bits

Figure A4 Multiple sequence alignment for design of *ceze-5* reverse degenerate primer. *The red blocks indicate the regions selected for reverse degenerate primer design. Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**),* C. zeae-maydis *(**cezm**)* C. nicotianae *(**cnic**),* C. sojina *(**ceso**) and* C. canescens *(**ceca**). The* C. zeina *amino acid sequence for the selected region is indicated (**ceze-5-aa**).*



```
ceze-7-aa   -S--G--I--I--S--A--L--D--W--I--N--K--D--V--A--A
ceze-7-F    TCCGGAATCATATCGGCCCTCGACTGGATCAACAAGGACGTCGCAGC
cezm-7-F    TCCGGAATCATATCAGCCCTCGACTGGATCAACAAGGACGTCGCAGC
cber-7-F    TCAGGAATCATCTCGGCTCTCGACTGGATCAACAAGGACGTTGCAGC
cbet-7-F    TCAGGAATCATCTCGGCTCTCGACTGGATCAACAAGGACGTCGCGGC
cnic-7-F    TCAGGAATTATCTCGGCTCTCGACTGGATCAACAAGGACGTTGCAGC
ceso-7-F    TCAGGTATCATCTCGGCTCTCGACTGGATCAACAAGGACGTCGCAGC
ceca-7-F    TCAGGAATCATCTCGGCTCTCGACTGGATCAATAAGGACGTCGCGGC
Consensus   TCAGGAATCATCTCGGCTCTCGACTGGATCAACAAGGACGTCGCAGC
```

Figure A5 **Multiple sequence alignment for design of *ceze-7* forward degenerate primer.** *The red blocks indicate the regions selected for forward degenerate primer design. Sequences are shown for* C. zeina ***(ceze)***, C. beticola ***(cbet)***, C. berteroae ***(cber)***, C. zeae-maydis ***(cezm)*** C. nicotianae ***(cnic)***, C. sojina ***(ceso)*** *and* C. canescens ***(ceca)***. *The* C. zeina *amino acid sequence for the selected region is indicated (**ceze-7-aa**).*



```
ceze-7-aa   -V--A--L--H--A--P--G--F--R--V--R--S--A--G--I--A--G--P
ceze-7-R    GTGGCCTTACATGCACCAGGCTTCCGCGTGCGATCTGCGGGAATCGCAGGCCC
cezm-7-R    GTGGCTTTACATGCACCAGGCTTCCGCGTGCGATCAGCGGGGATCGCAGGGCC
cber-7-R    GTGGCTTTACACGCACCAGGCTTCCGCGTGCGATCCGCGGGCATCGCAGGTCC
cbet-7-R    GTGGCTTTACACGCACCAGGCTTCCGCGTGCGATCAGCGGGCATCGCAGGTCC
cnic-7-R    GTGGCTTTACATGCACCAGGCTTTCGCGTGCGATCCGCGGGCATCGCAGGACC
ceso-7-R    GTGGCAATGCATGCACCAGGCTTTCGCGTGCGATCCGCGGGCATCGCAGGGCC
ceca-7-R    GTGGCTTTACACGCACCAGGCTTCCGCGTGCGATCAGCGGGCATCGCAGGTCC
Consensus   GTGGCTTTACATGCACCAGGCTTCCGCGTGCGATCNGCGGGCATCGCAGGTCC
```

Figure A6 **Multiple sequence alignment for design of *ceze-7* reverse degenerate primer.** *The red blocks indicate the regions selected for reverse degenerate primer design. Sequences are shown for* C. zeina ***(ceze)***, C. beticola ***(cbet)***, C. berteroae ***(cber)***, C. zeae-maydis ***(cezm)*** C. nicotianae ***(cnic)***, C. sojina ***(ceso)*** *and* C. canescens ***(ceca)***. *The* C. zeina *amino acid sequence for the selected region is indicated (**ceze-7-aa**).*

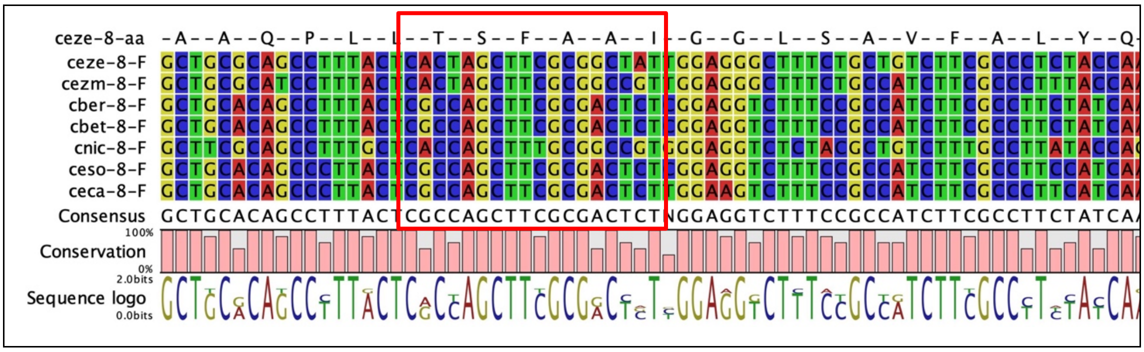

ceze-8-aa -A--A--Q--P--L--L--T--S--F--A--A--I--G--G--L--S--A--V--F--A--L--Y--Q-
ceze-8-F
cezm-8-F
cber-8-F
cbet-8-F
cnic-8-F
ceso-8-F
ceca-8-F
Consensus GCTGCACAGCCTTTACTCGCCAGCTTCGCGACTCTNGGAGGTCTTTCCGCCATCTTCGCCTTCTATCA
100%
Conservation
0%
2.0bits
Sequence logo
0.0bits

Figure A7 **Multiple sequence alignment for design of *ceze-8* forward degenerate primer.** *The red blocks indicate the regions selected for forward degenerate primer design. Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**),* C. zeae-maydis *(**cezm**)* C. nicotianae *(**cnic**),* C. sojina *(**ceso**) and* C. canescens *(**ceca**). The* C. zeina *amino acid sequence for the selected region is indicated (**ceze-8-aa**).*

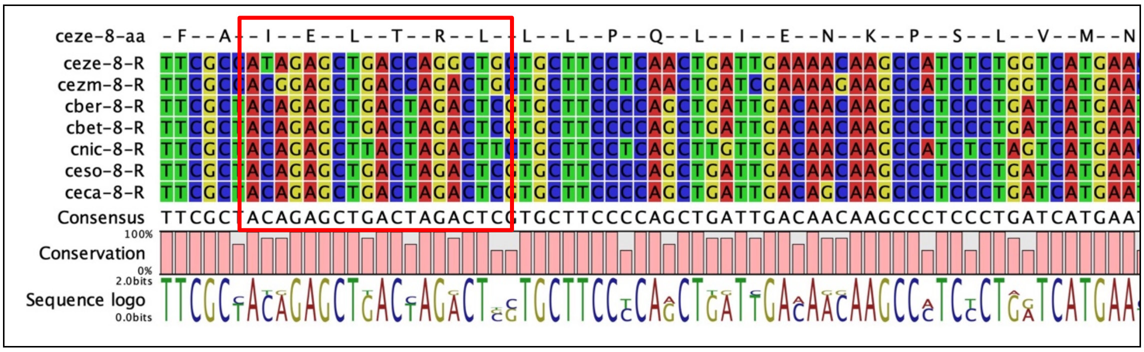

ceze-8-aa -F--A--I--E--L--T--R--L--L--L--P--Q--L--I--E--N--K--P--S--L--V--M--N
ceze-8-R
cezm-8-R
cber-8-R
cbet-8-R
cnic-8-R
ceso-8-R
ceca-8-R
Consensus TTCGCTACAGAGCTGACTAGACTCGTGCTTCCCCAGCTGATTGACAACAAGCCCTCCCTGATCATGAA
100%
Conservation
0%
2.0bits
Sequence logo
0.0bits

Figure A8 **Multiple sequence alignment for design of *ceze-8* reverse degenerate primer.** *The red blocks indicate the regions selected for reverse degenerate primer design. Sequences are shown for* C. zeina *(**ceze**),* C. beticola *(**cbet**),* C. berteroae *(**cber**),* C. zeae-maydis *(**cezm**)* C. nicotianae *(**cnic**),* C. sojina *(**ceso**) and* C. canescens *(**ceca**). The* C. zeina *amino acid sequence for the selected region is indicated (**ceze-8-aa**).*

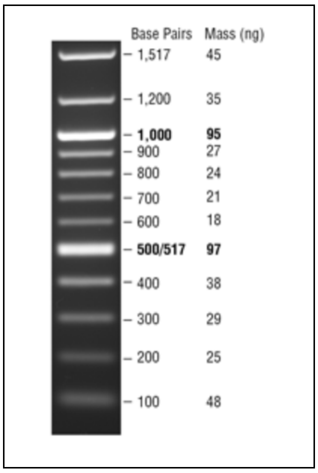


Figure A9 **DNA 100 bp ladder size separation for agarose gel electrophoresis (New England Biolabs).**